RADA NAUKOWA DYSCYPLINY
INFORMATYKA TECHNICZNA I TELEKOMUNIKACJA POLITECHNIKI WARSZAWSKIEJ
zaprasza na
OBRONĘ ROZPRAWY DOKTORSKIEJ

mgr inż. Witolda Oleszkiewicza
która odbędzie się w dniu **6 czerwca 2024 roku,** o godzinie **12:00** w trybie stacjonarnym
Temat rozprawy:

„ Wyjaśnialne uczenie maszynowe z zastosowaniem konceptów zrozumiałych dla człowieka”

Promotor: dr hab. inż. Robert Nowak, prof. uczelni – Politechnika Warszawska

Recenzenci: dr hab. inż. Paweł Forczmański – Zachodniopomorski Uniwersytet Technologiczny w Szczecinie
prof. dr hab. Anna Gambin – Uniwersytet Warszawski
dr hab. Marek Sikora – Politechnika Śląska

Obrona odbędzie się w Sali nr 116 Gmach Wydziału Elektroniki i Technik Informacyjnych Politechniki Warszawskiej, ul. Nowowiejska 15/19, 00-665 Warszawa.
Z rozprawą doktorską i recenzjami można zapoznać się w Czytelni Biblioteki Głównej Politechniki Warszawskiej, Warszawa, Plac Politechniki 1.

Streszczenie rozprawy doktorskiej i recenzje są zamieszczone na stronie internetowej: https://www.bip.pw.edu.pl/Postepowania-w-sprawie-nadania-stopnia-naukowego/Doktoraty/Wszczete-po-30-kwietnia-2019-r/Rada-Naukowa-Dyscypliny-Informatyka-Techniczna-i-Telekomunikacja/mgr-inz.-Witold-Oleszkiewicz

Przewodniczący Rady Naukowej Dyscypliny
Informatyka Techniczna i Telekomunikacja

Przewodniczący Rady Naukowej Dyscypliny
Informatyka Techniczna i Telekomunikacja
Politechniki Warszawskiej
prof. dr hab. inż. Jarosław Arabas

Wyjaśnialne uczenie maszynowe z zastosowaniem konceptów zrozumiałych dla człowieka

W niniejszej pracy przedstawiam serię pięciu publikacji poświęconych zagadnieniom wyjaśnialności modeli uczenia głębokiego.

Traktowanie sztucznych sieci neuronowych jako czarnych skrzynek, powoduje to, że nie ma pewności, czy decyzje modeli są podjęte na podstawie właściwych przesłanek. W moich pracach przedstawiam nowe statyczne metody wyjaśniające modele uczenia głębokiego, gdzie wyjaśnienie globalne jest generowane po wytrenowaniu modelu. Trzy prace dotyczą metody klasyfikatorów diagnostycznych, które badają informacje zawarte w reprezentacjach modeli. Jest to metoda powszechnie stosowana w przetwarzaniu języka naturalnego, jednak do tej pory nie miała ona swojego odpowiednika w widzeniu maszynowym. W moich pracach wprowadzam intuicyjną taksonomię wizualną, która zawiera znaki, słowa i zdania wizualne, analogicznie do liter, słów i zdań języka naturalnego. Dzięki temu definiuję szereg klasyfikatorów diagnostycznych, które pozwalają na badanie różnych cech reprezentacji modeli. Pokazuję przydatność metody klasyfikatorów diagnostycznych na przykładzie wyjaśniania reprezentacji samonadzorowanych. Metoda ta opiera się na obliczeniowej teorii widzenia Marra, dzięki czemu analizujemy reprezentacje za pomocą zrozumiałych dla człowieka cech wizualnych, takich jak tekstury, kolory, kształty i linie. Moje badania pokazują, że relacje między językiem a obrazem są skutecznymi i intuicyjnymi narzędziami do wyjaśniania modeli uczenia głębokiego.

W dwóch pozostałych pracach przedstawiam nową metodę do anonimizacji zbiorów danych oraz metodę wyjaśniającą działanie modeli uczenia głębokiego w diagnostyce raka piersi. Metoda do anonimizacji obrazów działa z wykorzystaniem syjamskich generatywno-przeciwstawnych sieci neuronowych i pozwala na zbadanie, czy reprezentacje modeli uczenia głębokiego zawierają informacje o tożsamości osób na obrazie. Metoda wyjaśniająca w diagnostyce medycznej bada wpływ perturbacji obrazu na decyzję lekarza oraz maszyny, dzięki czemu stwierdzamy, że modele uczenie głębokiego w dużej mierze korzystają z informacji zawartej w składowych obrazu o wysokiej częstotliwości w przestrzeni Fouriera, które to informacje są niedostrzegane przez lekarzy.

Podsumowując, wszystkie powyższe zaproponowane przeze mnie nowe metody wyjaśniające pomagają lepiej zrozumieć modele sztucznej inteligencji. Dzięki tym metodom jesteśmy w stanie zbadać obciążenie modeli, określić ich silne i słabe strony, a także wskazać które pojęcia są dla nich istotne podczas podejmowania decyzji.

Słowa kluczowe: Wyjaśnialna Sztuczna Inteligencja, Klasyfikatory Diagnostyczne, Widzenie Maszynowe, Uczenie Głębokie

UNIWERSYTET WARSZAWSKI

Instytut Informatyki
ul. Banacha 2
02–097 Warsaw
POLAND

prof. dr hab. Anna Gambin
Phone: +(48 22) 5544 212
Fax: +(48 22) 5544 400
e-mail: a.gambin@uw.edu.pl

3 kwietnia 2024

KWESTIONARIUSZ-RECENZJA ROZPRAWY DOKTORSKIEJ DLA RADY WYDZIAŁU ELEKTRONIKI I TECHNIK INFORMACYJNYCH POLITECHNIKI WARSZAWSKIEJ

Tytuł: WYJAŚNIALNE UCZENIE MASZYNOWE Z ZASTOSOWANIEM KONCEPTÓW ZROZUMIAŁYCH DLA CZŁOWIEKA

Autor: MGR WITOLD OLESZKIEWICZ

1. Jakie zagadnienie naukowe/badawcze jest rozpatrywane w pracy (cel i teza rozprawy) i czy zostało ono dostatecznie jasno sformułowane przez autora?

Rozprawa porusza szerokie spektrum zagadnień związanych z wyjaśnialnością modeli uczenia głębokiego. Autor wykorzystuje obliczeniową teorię widzenia Marra jako podstawę do opracowania intuicyjnej taksonomii wizualnej, która służy jako narzędzie do analizy i interpretacji wyników modeli uczenia maszynowego. Umożliwia one lepsze zrozumienie sposobu, w jaki ludzki mózg przetwarza informacje wizualne, co może być użyteczne w kontekście analizy wyników obrazowych.

Ponadto, Autor eksploruje zastosowanie klasyfikatorów diagnostycznych, które pierwotnie zostały opracowane dla przetwarzania języka naturalnego, do problemów widzenia maszynowego. Wykorzystanie tych klasyfikatorów pozwala na adaptację zaawansowanych technik przetwarzania języka naturalnego do analizy obrazów (również medycznych).

Podsumowując, teza pracy doktorskiej wydaje się być adekwatnie sformułowana. Opracowane metody mogą mieć znaczenie w dziedzinie przetwarzania obrazów, wprowadzając

nowe spojrzenie na interpretowalność modeli uczenia maszynowego. Dodatkowo zaproponowano nową metodę anonimizacji danych oraz uzyskano ciekawe wnioski w badaniu różnic między interpretacją obrazów medycznych przez radiologów a modelami uczenia maszynowego.

2. Czy w rozprawie przeprowadzono w sposób właściwy analizę źródeł, w tym literatury światowej, stanu wiedzy i zastosowań w przemyśle?

Rozprawa składa się ze wstępu, który zawiera syntetyczny opis uzyskanych wyników oraz cyklu artykułów. W każdym z tych artykułów Autor przedstawia aktualny stan wiedzy dotyczący konkretnego zagadnienia. Choć ta struktura pozwala na systematyczne podejście do prezentacji badań i ich wyników, może ograniczać możliwość zdobycia szerszego oglądu danego obszaru. Jednak, poprzez analizę każdego artykułu czytelnik może uzyskać adekwatne zrozumienie ewolucji wiedzy w danej dziedzinie oraz wkładu Autora w badaną problematykę.

3. Czy autor rozwiązał postawione zagadnienia, czy użył właściwej metody i czy przyjęte założenia są uzasadnione?

Autor zaprezentował w swojej pracy kilka zagadnień o istotnym znaczeniu w kontekście interpretowalności i wyjaśnialności metod głębokiego uczenia. Te zagadnienia obejmują interpretowalność modeli widzenia maszynowego, anonimizację danych w uczeniu maszynowym oraz metody wyjaśniające działania modeli głębokiego uczenia w diagnostyce raka piersi.

Podejście Autora do tych problemów jest twórcze i zróżnicowane. Zademonstrowano umiejętność zastosowania różnorodnych metod a Autor przyjął adekwatne założenia, co pozwala na odpowiednie rozwiązanie postawionych zagadnień.

Należy podkreślić, że pełne rozwiązanie postawionych zagadnień jest niezwykle trudne. Jednak Autor rozprawy podchodzi do postawionych problemów z odpowiednią starannością i kreatywnością i z tego powodu zaproponowane rozwiązania mają szansę przyczynić się do istotnego postępu w dziedzinie.

4. Na czym polega oryginalność rozprawy, co stanowi samodzielny i oryginalny dorobek autora, jaka jest pozycja rozprawy w stosunku do stanu wiedzy czy poziomu techniki reprezentowanych przez literaturę światową?

Oryginalność rozprawy wynika z wieloaspektowego wkładu Autora oraz kreatywnego podejścia do rozwiązywania problemów (np. wykorzystanie klasyfikatorów diagnostycznych dla wyjaśnialności metod widzenia maszynowego). Artykuły składające się na rozprawę są wieloautorskie, ale wkład Autora wynosi między 15 % a 75%, co świadczy o jego istotnym zaangażowaniu i wkładzie w proces badawczy.

Muszę zaznaczyć, że publikacja A3 stanowi rozszerzoną wersję publikacji A1 i niestety duża część tekstu jest w obydwu pracach identyczna. Wykazany wkład Autora jest dużo wyższy dla publikacji A3, co wskazuje na jego zaangażowanie w eksperymenty i rozwijanie idei prezentowanych w pełnej wersji pracy. Doceniam też wykonanie przez Autora rozprawy nietrywialnych analiz obrazów radiologicznych w publikacji A5, którą oceniam jako bardzo wartościową.

Pozycja rozprawy w stosunku do stanu wiedzy oraz poziomu techniki reprezentowanych przez literaturę światową jest znacząca. Publikacje składające się na rozprawę ukazały się na renomowanych konferencjach, takich jak *International Joint Conferences on Artificial Intelligence*, oraz w cenionych czasopismach, takich jak *IEEE Access i Scientific Reports*, co świadczy o ich wysokiej jakości i znaczeniu w dziedzinie

5. Czy autor wykazał umiejętność poprawnego i przekonywującego przedstawienia uzyskanych przez siebie wyników (zwięzłość, jasność, poprawność redakcyjna rozprawy)?

Rozprawa jest złożona z kilku oddzielnych wieloautorskich artykułów, co utrudnia ocenę wkładu Autora w prezentację poszczególnych wyników. Pomimo precyzyjnego określenia procentowego wkładu autora w każdy z artykułów, ocena jego wpływu na sposób prezentacji wyników w poszczególnych fragmentach pracy jest trudna.

W kontekście poprawności redakcyjnej i prezentacji wyników można ocenić jedynie krótki wstęp do rozprawy. Został on napisany w sposób klarowny, co pozwala czytelnikowi zrozumieć kontekst i cele badawcze. Pomimo jego krótkiej formy, efektywnie omawia tematykę i zakres pracy, co korzystnie wpływa na zrozumienie treści. Dodatkowo, wyniki uzyskane w ramach poszczególnych projektów oraz wnioski z przeprowadzonych badań są też krótko zaprezentowane we wstępie, co ułatwia zrozumienie i ocenę osiągniętych rezultatów.

6. Jaka jest przydatność rozprawy dla nauk inżynieryjno-technicznych?

Modele uczenia głębokiego znajdują zastosowanie w większości obszarów działania człowieka. Interpretowalność i wyjaśnialność predykcji jest kluczowa w każdym zastosowaniu, a zwłaszcza w diagnostyce medycznej. Zaproponowane przez Autora rozprawy podejścia pozwalają na eksplorację przestrzeni reprezentacji modeli, co jest niezwykle kluczowe, zwłaszcza w kontekście metod uczenia bez nadzoru. Podobnie opracowana nowa metoda anonimizacji danych powinna znaleźć szerokie zastosowania, a zrozumienie różnic pomiędzy decyzjami radiologów i modeli uczących w stawianiu diagnozy pozwoli na wzrost zaufania do tych ostatnich.

7. Do której z następujących kategorii recenzent zalicza rozprawę:

a) **nie spełniająca wymagań stawianych rozprawom doktorskim przez obowiązujące przepisy**

b) **wymagająca wprowadzenia poprawek i ponownego recenzowania**

√ c) **spełniająca wymagania**

d) **spełniająca wymagania z wyraźnym nadmiarem**

e) **wybitnie dobra, zasługująca na wyróżnienie**

Podsumowując stwierdzam, że recenzowana przeze mnie praca spełnia wymagania stawiane rozprawom doktorskim przez obowiązujące przepisy i wnoszę o dopuszczenie magistra Witolda Oleszkiewicza do dalszych etapów przewodu doktorskiego.

Katowice, 03.03.2024

Dr hab. Marek Sikora
Katedra Sieci i Systemów Komputerowych
Politechnika Śląska
ul. Akademicka 16
44-100 Gliwice
Email: marek.sikora@polsl.pl

Recenzja rozprawy doktorskiej

Tytuł rozprawy: Wyjaśnialne uczenie maszynowe z zastosowaniem konceptów zrozumiałych dla człowieka

Autor rozprawy: mgr inż. Witold Oleszkiewicz

Promotor rozprawy: dr hab. inż. Robert Marek Nowak

Dziedzina: nauki inżynieryjno-techniczne

Dyscyplina: informatyka techniczna i telekomunikacja

1. Temat i cel rozprawy

Tematyka rozprawy obejmuje zagadnienia globalnej wyjaśnialności modeli głębokiego uczenia poprzez zastosowanie tzw. klasyfikatorów diagnostycznych badających, czy w reprezentacjach wytrenowanych modeli występują pewne elementy – cechy wizualne. Prace przedstawione do oceny koncentrują się na aspektach wyjaśnialności modeli dedykowanych do rozpoznawania i klasyfikacji obrazów – wizji komputerowej. Nadrzędnym celem przedstawionych prac jest lepsze niż dotychczas zrozumienie podstaw podejmowania decyzji przez modele maszynowego uczenia. Autor proponuje szereg intersujących metod, których osnową jest wyjaśnialność bazująca na tzw. taksonomii wizualnej zawierającej znaki, słowa i zdania wizualne, a więc pojęcia, które mogą być zrozumiałe dla człowieka.

W pracy nie zdefiniowano tezy głównej ani tez pomocniczych. Jest to zrozumiałe, gdyż przedmiotem oceny jest cykl publikacji, natomiast w autoreferacie jasno przedstawiono i uzasadniono cel i zakres badań. Zawartość wszystkich z przedstawionych do oceny publikacji jest zgodna ze zdefiniowanym celem i zakresem badań, a prace te rozważane jako całość stanową spójną logicznie całość.

Uzasadnienie wyboru tematu nie budzi żadnych wątpliwości, Autor bardzo dobrze i trafnie uzasadnia celowość podjęcia badań opisanych w rozprawie. Tematyka badań jest bardzo istotna, wyjaśnialność systemów sztucznej inteligencji jest aktualnym tematem naukowym, zwłaszcza w kontekście nadchodzących regulacji prawnych dotyczących zaufana do tego typu systemów.

2. Zawartość i charakter publikacji/rozprawy

Do oceny przedstawiono cykl pięciu publikacji, z których dwie wydano w czasopismach naukowych, a trzy stanowią doniesienia konferencyjne. Dokładne dane bibliograficzne publikacji przedstawiono poniżej:

1. D. Basaj, W. Oleszkiewicz i inni: Explaining Self-Supervised Image Representation with Visual Probing. IJCAI 2021 (Core A*, 200 pkt. MEiN).
2. W. Oleszkewicz i inni: Which Visual Features Impact the Performance of Target Task in Self-supervised earning? CCS 2022 (Core A, 140 pkt. MEiN).
3. W. Oleszkiewicz i inni: Visual Probing: Cognitive Framework for Explaining Slef-Supervised Image Representations. IEEE Access 11, 2023 (IF 3.476; 100 pkt. MEiN).
4. W. Oleszkiewicz i inni: Siamese Generative Adversarial Privatizer for Biometric Data. ACCV 2018 (Core B; 70 pkt. MEiN).
5. T. Makino, S. Jastrzębsk, W. Oleszkiewicz i inni: Differences between human and machine perception in medical diagnosis. Scientific Reports 12, 2022 (IF 4.996; 140 pkt. MEiN).

Doktorant deklaruje, że jego wkład w postanie publikacji był następujący:

1. Współudział w zdefiniowaniu problemu badawczego, opracowanie metod – adaptacja z domeny przetwarzania języka naturalnego do analizy obrazów. Zdefiniowane kluczowych zadań diagnostycznych. Doktorant zaplanował i wykonał również większą cześć eksperymentów. Wkład Doktoranta oceniam jako dominujący.
3. Praca 3 jest rozszerzeniem pracy 1. Praca 1 jest pracą konferencyjną, a 3 jej rozszerzoną wersją. Doktorant, poza zadaniami wymienionym w pkt. 1, realizował również zadania związane z przygotowaniem i opracowaniem badań ankietowych. Wkład doktoranta w powstanie tej pracy jest zdecydowanie dominujący.
2. Doktorant brał udział w zdefiniowaniu zadania badawczego, miał kluczowy wkład w opracowanie algorytmu do tworzenia amnezyjnych zadań diagnostycznych. Przeprowadził pełną implementację metod oraz wykonał zdecydowaną większość eksperymentów. Wkład Doktoranta w powstanie publikacji był moim zdaniem zdecydowanie dominujący.
4. Doktorant brał udział w zdefiniowaniu problemu badawczego. Przeprowadził wszystkie eksperymenty wraz z analizą uzyskanych wyników. Wkład Doktoranta oceniam na więcej niż proporcjonalny do liczby autorów publikacji.
5. W ostatniej z prac przedstawionych do oceny wkład doktorant był mniejszościowy, ale również istotny, gdyż Doktorant brał udział w zdefiniowaniu problemu badawczego, był współautorem implementacji, a także wykonał część eksperymentów.

Publikacje 1 i 3 traktują o zastosowaniu metody zadań diagnostycznych w wizji komputerowej. Autorzy, poprzez analogię do wyjaśniania reprezentacji w dziedzinie przetwarzania języka naturalnego, definiują zadania diagnostyczne dla reprezentacji obrazów uzyskanych za pomocą metod ML (koncentrują się na reprezentacjach tworzonych przez metody samonadzorujące się). W pracy 1 zdefiniowano przybliżoną taksonomię zawierającą pojęcia będące wizualnymi odpowiednikami słów i zdań. W pracy 3 rozszerzono tę koncepcję, wprowadzając semantykę słów/pojęć wizualnych przy wykorzystaniu kogniwistycznej teorii percepcji wzrokowej Marra. Proponowane podejścia pozwalają w sposób bardziej intuicyjnych – w odniesieniu do aktualnych metod – wyjaśniać reprezentacje obrazów za pomocą konceptów zrozumiałych dla użytkownika (człowieka).

Praca 2 stanowi dalsze rozszerzenie metod wyjaśniania poprzez wprowadzenie tzw. amnezyjnych zadań diagnostycznych. W metodzie tej autorzy badają nie tylko występowanie określonych pojęć wizualnych w reprezentacji obrazów, ale sprawdzają także, czy ich usunięcie wpływa na decyzje wytrenowanego modelu.

Praca 4 przedstawia metodę anonimizacji zborów danych poprzez modyfikację nie tylko obrazów, ale również ich reprezentacji odzwierciedlonych w modelu. Celem metody jest usunięcie informacji pozwalającej zidentyfikować tożsamość osób widocznych na obrazie.

Ostatnia z prac, pozycja 5, podejmuje temat badania różnic pomiędzy przesłankami leżącymi u podstaw decyzji diagnostycznych (analiza obrazów rentgenowskich, diagnostyka raka piersi)

podejmowanych przez lekarzy i głębokie sztuczne sieci neuronowe. Metoda zaproponowana przez autorów polega na analizie wpływu zakłóceń wprowadzanych do oryginalnych obrazów na decyzje lekarzy i modeli ML.

6. Analiza źródeł, zastany stan wiedzy, dorobek publikacyjny autora

Bibliografia przywoływana w publikacjach cyklu zawiera odpowiednio 28, 21, 62, 36 i 47 pozycji literaturowych. Autor cytuje je w odpowiednim kontekście. Źródła te dobrze przedstawiają bieżący stan wiedzy w zakresie zagadnień podejmowanych w pracy. W szczególności, Autor przedstawia propozycje metod wyjaśnialność stosowanych w objaśnianiu działania metod wizji komputerowej. Z jednej strony, recenzent odczuwa pewien niedosyt związany z brakiem szerszego przeglądu metod wyjaśnialności, w szczególność pokazania szerszego kontekstu tego zagadnienia w odniesieniu do złożonych modeli ML stosowanych nie tylko w analizie obrazów, ale również w analizie danych tabelarycznych, szeregów czasowych, etc. Z drugiej strony, recenzent jest świadomy, że w przypadku publikacji naukowej nie mającej charakteru rozprawy doktorskiej jako takiej, sekcja related work (lub jej odpowiednik) musi być ukierunkowana stricte na zagadnienia poruszane w publikacji, w szczególność w przypadku gdy jest to praca konferencyjna. Przywoływana literatura jest aktualna – duża część cytowanych prac została wydana po roku 2019. Wyjątek stanowią referencje zawarte w pracy wydanej w 2019, jednak i w tej pracy autorzy cytują aktualne w tamtym czasie pozycje literatury.

Prace przedstawione do oceny publikowane były w dobrych czasopismach (EEE Access, Scientific Reports) oraz w materiałach prestiżowych konferencji naukowych. Dorobek publikacyjny Doktoranta oceniam jako bardzo dobry.

7. Oryginalne wyniki i ich znaczenie

Doktorant podejmuje ważny problem definiowania zrozumiałych dla użytkownika wyjaśnień działania złożonych systemów rozpoznawania i klasyfikacji obrazów. W swojej pracy podejmuje również tematykę anonimizacji obrazów, ukierunkowanej na eliminacje z obrazów cech charakterystycznych – np. osobniczych – umożliwiających np. ustalenie tożsamość osób. Przy czym eliminacja ta nie wpływa w sposób istotny na użyteczność zanonimizowanych przykładów jako źródła danych treningowych dla systemów maszynowego uczenia.
Przedstawiony do recenzji cykl publikacji prezentuje nowatorskie podejście zarówno do zagadnień wyjaśnialności, jak również anonimizacji danych obrazowych.

Za najbardziej wartościowe wyniki uzyskane przez Doktoranta uważam:

- Wprowadzenie taksonomii konceptów graficznych zawierającą „znaki", „słowa" i „zdania" wizualne. W początkowej fazie badań koncepty wizualne reprezentowane są jako super-piksele składające się niepodzielnych pikseli mających wspólne cechy lub tworzących wyodrębnione obiekty. W dalszej części badań, bazując na obliczeniowej

teorii widzenia Marra, wybrano sześć cech: jasność, kolor, tekstura, linia, kształt, forma w celu bardziej zrozumiałego dla człowieka opisu słów wizualnych. Zdefiniowanie konceptów wizualnych pozwoliło w dalszej części badań na analizę reprezentacji obrazów za pomocą klasyfikatorów diagnostycznych. Autor zdefiniował pięć takich klasyfikatorów mających na celu m.in. wykrywanie słów diagnostycznych na obrazie, identyfikację liczby unikalnych słów wizualnych na obrazie, wykrywanie modyfikacji obrazu, czy też – w pewny sensie – podobieństwa ich reprezentacji.

Klasyfikatory te pozwalają na szeroką i wszechstronną diagnostykę reprezentacji obrazów uzyskiwanych przez metody samonadzorujące się.

- Opracowanie metody klasyfikatorów amnezyjnych. Metoda ta bada czy usunięcie z reprezentacji informacji o występowaniu określonych pojęć wizualnych wpływa na decyzje modelu. Dokładniej, po usunięciu z reprezentacji informacji o danym koncepcie mierzona jest jakość klasyfikatora i porównywana jest z jakością reprezentacji zawierającej usunięty koncept. Metoda znacząco poszerza funkcjonalność wyjaśnień oferowanych przez klasyfikatory diagnostyczne oraz rozwija taksonomię pojęć wizualnych, umożliwiając badanie i porównywanie preferencji oraz "uprzedzeń" różnych metod trenowania modeli.
- Opracowanie metody anonimizacji obrazów w celu eliminacji informacji pozwalających na odkrycie tożsamości widocznych na nich osób. Zastosowanie sieci neuronowej do wyrywania cech identyfikujących osobę, a następnie podejścia generatywnego, które ukrywa te informacje przy minimalnym zmniejszeniu użyteczność (rozumianej jako możliwość użycia obrazu jako przykładu treningowego) przekształconego obrazu. Rozwiązanie to uważam za bardzo ciekawe i nowatorskie.

8. Redakcja publikacji będących podstawą do ubiegania się o stopień doktora, ocena sposobu prezentacji wyników

Publikacje przedstawione do oceny zredagowane są w sposób dobry. W dużej mierze układ prac determinowany jest wymaganiami czasopism i konferencji.

Wyniki prezentowane są zarówno w postaci zestawień ilościowych odnoszących się np. do mary AUC (ang. Area Under the ROC Curve) w przypadku klasyfikatorów diagnostycznych, jak również jakościowy polegający na prezentacji na przykładowych obrazach zidentyfikowanych elementów graficznych.

We wszystkich pracach brakuje mi jednak case study(ies) ilustrujących, jak rzeczywiście mogłoby wyglądać objaśnienie generowane dla końcowego odbiorcy wyników.

Zarówno sam autoreferat, jak i publikacje przedstawione do oceny czyta się bardzo dobrze, stosukowo łatwo jest zrozumieć intencje i wkład autora.

9. Słabe strony i uwagi krytyczne/dyskusyjne

Recenzent nie wnosi zasadniczych uwag – w szczególności uwag negatywnych – do przedstawionych do oceny publikacji.

W przedstawionych do recenzji publikacjach brakuje spójnej metodyki wyjaśniania bazującej na propozycjach autora. Nie do końca jest dla mnie jasne, jak wyglądałaby wyjaśnialność dla końcowego użytkownika nie będącego specjalistą z zakresu maszynowego uczenia. Czy propozycje Autora są adresowane jedynie do użytkowników zaawansowanych? A celem wyjaśnialności jest lepsza diagnostyka – i w dalszej perspektywie poprawa – trenowanych przez nich modeli?

Brak mi również szerszej pespektywy dotyczącej efektywność działania metody wyjaśnialności w kontekście typów analizowanych obrazów. Czy wpływ na wyniki ma ich rozdzielczość, występowanie kolorów lub ich brak etc.?

Uwag i pytania szczegółowe:

1. Z czego wynika mała liczba zbiorów danych, jakie Autor analizuje w przywoływanych pracach? Czy zdaniem Autora może to ograniczać zaufanie do efektywności metody? Czy też chodzi o to, że w analizowanych zbiorach liczba obrazów (przykładów) jest duża, zatem zdaniem Autora wystarczająca do weryfikacji przedstawianych propozycji?
2. Nie zauważyłem, aby Autor udostępniał implementacje opracowanych przez siebie metod, będzie to zdecydowanie utrudniać odtworzenie wyników innym badaczom. Czy metody opracowane w ramach doktoratu – ich implementacje – są dostępne dla szerszego grana badawczy i użytkowników?
3. W przypadku ostatniego artykułu z cyklu brakuje mi dyskusji dotyczącej medycznych podstaw – przyczyn i różnic – w jaki eksperci i metody ML analizują rozważaną w tej pracy grupę obrazów. Rozumiem, że autor nie ma wykształcenia medycznego, ale nasuwa się pytanie, jakie wnioski dla ekspertów dziedzinowych (lekarzy), a jakie dla twórców systemów automatycznej diagnostyki obrazowej mogą wypływać z przeprowadzonych badań?
4. Czy metody przedstawione przez Doktoranta maja zastosowanie w wyjaśnianiu metod nadzorowanych?

10. Podsumowanie i wniosek końcowy

Po analizie rozprawy mogę stwierdzić, że zamieszczone w niej rezultaty badań uzyskano w sposób rzetelny, a wyniki stanowią nowy wkład w dyscyplinę informatyka techniczna i telekomunikacja – w szczególności wnoszą wkład do metodyk wyjaśniania decyzji podejmowanych przez złożone systemy rozpoznawania i klasyfikacji obrazów. Rozprawa potwierdza zdolność Doktoranta do dalszej pracy naukowej. Uwagi krytyczne nie umniejszają mojej jednoznacznie pozytywnej oceny rozprawy.

Stwierdzam, że recenzowana rozprawa pt. „Wyjaśnialne uczenie maszynowe z zastosowaniem konceptów zrozumiałych dla człowieka", będąca cyklem publikacji naukowych, przygotowana przez mgr. inż. Witolda Oleszkiewicza spełnia wymagania

i warunki określone w ustawie z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (jednolity tekst Dz. U. z 2023 r. z późn. zm.) i wnoszę o jej przyjęcie, dopuszczenie jej do publicznej obrony i dalszych etapów postępowania doktorskiego.

Zachodniopomorski Uniwersytet Technologiczny w Szczecinie

Wydział Informatyki

Szczecin, 20.02.2024

dr hab. inż. Paweł Forczmański, prof. ZUT

Zachodniopomorski Uniwersytet Technologiczny w Szczecinie

Wydział Informatyki

pforczmański@zut.edu.pl

RECENZJA ROZPRAWY DOKTORSKIEJ DLA RADY DYSCYPLINY INFORMATYKA TECHNICZNA I TELEKOMUNIKACJA POLITECHNIKI WARSZAWSKIEJ

Autor rozprawy doktorskiej: **mgr inż. Witold Oleszkiewicz**

Tytuł rozprawy: **Wyjaśnialne uczenie maszynowe z zastosowaniem konceptów zrozumiałych dla człowieka**

Promotor: **dr hab. inż. Robert Nowak, prof. uczelni**

1. Zakres, cel, teza i charakter rozprawy

Tematyka recenzowanej pracy doktorskiej dotyczy problematyki uczenia maszynowego (głębokiego) metod umożliwiających skuteczniejszą interpretację sposobu ich działania. Problematyka wyjaśnianej sztucznej inteligencji nie jest zagadnieniem nowym i znanych jest wiele prac na ten temat. Nowatorstwo recenzowanej pracy polega na umiejętnym połączeniu wybranych metod sztucznej inteligencji (konkretnie głębokiego uczenia) z metodami wzorowanymi na przetwarzaniu języka naturalnego, co pozwala na wsparcie użytkownika komputerowego na etapie podejmowania decyzji. Rozprawa doktorska ma formę cyklu pięciu powiązanych tematycznie artykułów naukowych opublikowanych w międzynarodowych czasopismach oraz materiałach konferencji naukowych w latach 2019-2023. Przedmiotem badań były statyczne metody wyjaśniające zastosowane do wybranych modeli uczenia głębokiego ukierunkowanych na zadania widzenia komputerowego. Tematyka ta jest bardzo aktualna a zapotrzebowanie na wyjaśnianą sztuczną inteligencję stale rośnie. Analizując istniejące rozwiązania Autor zauważył, że w dotychczasowych badaniach wiele miejsca poświęca się na tworzenie modeli, które zwyczajowo określa się jako „czarne skrzynki", które pomimo wysokiej skuteczności działania (niezależnie, czy jest to klasyfikacja, czy predykcja), nie są akceptowalne jako rozwiązana mogące zastąpić lub też przynajmniej wspierać człowieka w typowych zadaniach, np. wyszukiwaniu i rozpoznawaniu obrazów lub diagnostyce medycznej. Stąd, zgodnie z obecnymi trendami, pojawiła się koncepcja opracowania nowych metod wyjaśniających tworzone modele uczenia maszynowego i ujawniające przesłanki, na podstawie których podejmują swoje decyzje. Istota tej koncepcji została sformułowana w postaci następującej tezy rozprawy, zaprezentowanej w sposób jawny na str. 10: „Główna teza badawczą jest stwierdzenie, że relacje między językiem a obrazem są skutecznymi i intuicyjnymi narzędziami do wyjaśniania modeli uczenia głębokiego". Sformułowana teza jest wystarczająco precyzyjna, choć wspomniane w niej „relacje"

mogą być nie do końca jednoznaczne, bez wspomnienia, że chodzi tu o opisowe traktowanie cech wizualnych ekstrahowanych lub tworzonych na etapie działania modelu.

W tym samy miejscu Autor deklaruje cel pracy, którym jest „[...] opracowanie narzędzi uczenia maszynowego, które dostarczają wyjaśnień predykcji sztucznych sieci neuronowych za pomocą konceptów zrozumiałych dla człowieka". Jest to w pewnym sensie przeformułowana teza pracy. Tak postawiony cel jest czytelny i dość oczywisty. Jego skuteczna realizacja wynika bezpośrednio z cyklu publikacji Autora.

2. Układ rozprawy i jej składowe

Przedstawiona do recenzji rozprawa ma formę autoreferatu wydanego na 103 stronach. Główna część to krótki przegląd literatury, opis uzyskanych wyników, kopie publikacji wraz ze wskazaniem autorskiego wkładu Doktoranta oraz lista dodatkowych dokonań naukowych, organizacyjnych i dydaktycznych. Właściwe osiągnięcie naukowe, jak już wspomniałem wcześniej, składa się z cyklu pięciu publikacji, do których należą:

[a1] Dominika Basaj, Witold Oleszkiewicz, Igor Sieradzki, Michal Górszczak, Barbara Rychalska, Tomasz Trzcinski, Bartosz Zielinski: Explaining Self-Supervised Image Representations with Visual Probing. IJCAI 2021: 592-598

[a2] Witold Oleszkiewicz, Dominika Basaj, Tomasz Trzcinski, Bartosz Zielinski: Which Visual Features Impact the Performance of Target Task in Self-supervised Learning? ICCS (1) 2022: 331-344

[a3] Witold Oleszkiewicz, Dominika Basaj, Igor Sieradzki, Michal Górszczak, Barbara Rychalska, Koryna Lewandowska, Tomasz Trzcinski, Bartosz Zielinski: Visual Probing: Cognitive Framework for Explaining Self-Supervised Image Representations. IEEE Access 11: 13028-13043 (2023)

[a4] Witold Oleszkiewicz, Peter Kairouz, Karol J. Piczak, Ram Rajagopal, Tomasz Trzcinski: Siamese Generative Adversarial Privatizer for Biometric Data. ACCV (5) 2018: 482-497

[a5] Taro Makino, Stanisław Jastrzębski, Witold Oleszkiewicz, et al.: Differences between human and machine perception in medical diagnosis. Scientific Reports 12(1):1-13 (2022)

Wszystkie wymienione publikacje są wieloautorskie i zostały opracowane w zespołach międzynarodowych, reprezentujących uczelnie polskie i zagraniczne oraz przedsiębiorstwa. W trzech z nich Autor jest wymieniony na pierwszym miejscu listy autorów. Wszystkie prace realizowane były w ramach projektów naukowych finansowanych przez instytucje zewnętrzne. Doktorant zadeklarował dominujący (50% i więcej) wkład w trzech z prezentowanych publikacji. W pozostałych dwóch Jego wkład jest istotny lecz można go traktować jako uzupełnienie dorobku.

Prace są ze sobą powiązane poprzez odniesienie do zagadnienia wyjaśnialnej sztucznej inteligencji i dotyczą ogólnego zadania rozpoznawania obrazów [a1-a3], anonimizacji danych biometrycznej [a4] i radiologicznej diagnostyki medycznej [a5].

Rozprawa uzupełniona jest spisem pozostałych publikacji naukowych Autora oraz innych, ważnych z punktu widzenia rozwoju naukowego, osiągnięć. Są one znaczące i w sposób istotny wpływają na pozytywny odbiór dorobku Autora.

Rozprawa stoi na dobrym poziomie językowym, stylistycznym i edycyjnym. Jej struktura jest prawidłowa. Doktorant w skondensowanej formie oraz w logiczny i czytelny sposób pokazał wszystkie istotne zagadnienia zawiązane z przeprowadzonymi badaniami.

3. Analiza źródeł

Bibliografia przedstawiona w autoreferacie zawiera 49 pozycje obejmujące głównie artykuły publikowane w czasopismach zagranicznych (m.in. IEEE TPAMI, IEEE TIP, ACM Computing Surveys) referaty prezentowane na konferencjach międzynarodowych (m.in. NeurIPS, ICCCN, ECML PKDD, CVPR, ICCV, ECCV, ICML) oraz monografie publikowane w latach 1982 – 2022, z czego zdecydowana ich większość to publikacje z ostatnich lat.

Wśród omówionych prac naukowych znajdują się najważniejsze publikacje związane z tematyką poruszaną w rozprawie, w szczególności z wyjaśnialną sztuczną inteligencją, widzeniem komputerowym, uczeniem głębokim, uczeniem nadzorowanym, nienadzorowanym i samonadzorowanym. Dodatkowo, biorąc pod uwagę szeroki zakres literatury cytowanej w poszczególnych publikacjach, można stwierdzić, że Doktorant ma obszerną i aktualną wiedzę dziedzinową.

4. Metodyka badań

Zaprezentowana w pracy metodyka badań obejmowała wstępną analizę problemu wyjaśnialnej sztucznej inteligencji oraz opracowanie algorytmów ukierunkowanych na zwiększenie możliwości interpretacji sposobu podejmowania decyzji w systemach bazujących na uczeniu maszynowych (w szczególności uczeniu głębokim). Autor skupił się na jednej z klas metod wyjaśniających, mianowicie globalnych metodach statycznych, stosowanych post-hoc, czyli po wytrenowaniu modelu. Pominął w ten sposób cały zestaw metod wyjaśniających wpływ doboru danych na trening modelu, co spowodowane było prawdopodobnie chęcią skupienia się na rozwiązaniach niezależnych od typu i charakteru danych. Przyjęta strategia wydaje się być słuszna, gdyż opisane metody są dość uniwersalne i mogą być stosowane w wielu obszarach badawczych, co pokazały publikacje Autora. Eksperymentalne wykazanie skuteczności opracowanych rozwiązań doprowadziło do udowodnienia tezy postawionej w rozprawie.

Do najważniejszych osiągnięć Autora należy zaliczyć metodę zadań diagnostycznych ukierunkowanych na obszar widzenia komputerowego (zwane *visual probing tasks*), które w sposób naturalny łączą zdania w języku naturalnym z ukrytą reprezentacją tworzoną przez modele uczenia głębokiego. Autor pokazał, że można w ten sposób łączyć obszary przetwarzania języka naturalnego (NLP) z teorią percepcji wzrokowej Marra i uczeniem się reprezentacji przez modele głębokie. Powstały w ten sposób tzw. słowa wizualne [a1] i bardziej złożona hierarchia kognitywno-wizualna [a3], które zostały przyrównane do koncepcji znaków, słów i zdań w języku naturalnym. Opracowane w publikacji [a3] klasyfikatory diagnostyczne w ciekawy sposób łączą elementy NLP z widzeniem komputerowym, szczególnie w kontekście metod uczenia samonadzorowanego.

Uzupełnieniem opisanych powyżej metod są wprowadzone w pracy [a2] tzw. amnezyjne klasyfikatory diagnostyczne, których celem jest odpowiedź na pytanie, które koncepty percepcyjne znajdują odzwierciedlenie w reprezentacjach wytworzonych na drodze uczenia samonadzorowanego. Autor wykorzystał tutaj prostą obserwację, której kwintesencją jest to, że usuniecie odpowiednich reprezentacji istotnych konceptów percepcyjnych z przestrzeni ukrytych cech spowoduje obniżenie skuteczności realizacji docelowego zadania klasyfikującego. Potwierdziły to badania eksperymentalne. Ta sama metodyka weryfikacji została również wykorzystana w pozostałych publikacjach z cyklu.

Zastosowanie wymienionych wcześniej koncepcji zostało zaprezentowane w publikacji [a4], która prezentuje sposób anonimizacji biometrycznych danych obrazowych a weryfikacja skuteczności tego procesu następuje za pomocą oryginalnej syjamskiej sieci neuronowej o architekturze typu GAN. Opracowany filtr ma za zadanie usunięcie z obrazu elementów, które mogłyby być wykorzystane do skojarzenia go z obrazami, które są uzupełnione danymi identyfikującymi np. tożsamością prezentowanej osoby. Co ważne, w pracy uwzględniony został fakt, że anonimizacja może powodować znaczące zniekształcenia, dlatego na etapie przetwarzania są one identyfikowane za pomocą klasycznej metryki SSIM. W omawianym przypadku wyjaśnienie działania modelu i istotności cech wykorzystywanych w zadaniu docelowym przeprowadzono w sposób niebezpośredni, wykonując procedurę porównywania (weryfikacji) tożsamości. Badania zrealizowano na dwóch zbiorach danych graficznych: zbiorze rysunkowych twarzy i rzeczywistych odcisków palców.

Ostatnia z uwzględnionych w cyklu publikacji [a5] dotyczy ważnego problemu oceny porównawczej skuteczności diagnostyki radiologicznej pomiędzy specjalistami z tej dziedziny a algorytmami komputerowymi w obecności typowej niedoskonałości obrazu - rozmycia. Przedstawione badania wykorzystywały radiogramy prezentujące zmiany chorobowe związane z rakiem piersi. Obrazy były rozmywane w taki sposób, aby wykryć poziom wpływu wysokoczęstotliwościowych komponentów na końcową diagnozę. Okazało się, że po takiej operacji skuteczność klasyfikatora bazującego na sieci głębokiej spadała, gdyż był on zbyt silnie ukierunkowany na te właśnie charakterystyki obrazu. Metoda wyjaśniająca została w tym wypadku również wykorzystana nie wprost ale na zasadzie analizy skuteczności modelu docelowego dokonującego predykcji dla danych oryginalnych i zniekształconych.

Obserwacje wynikające z analizy ww. problemów potwierdziły, że metody wyjaśniające pozwalają na zidentyfikowanie elementów modelu głębokiego uczenia odpowiedzialnego za skuteczną realizację postawionego zadania. Uniwersalność opracowanego podejścia polega na tym, że nie zależy ono od docelowego zadania, tj. klasyfikacji czy rozpoznawania.

W odniesieniu do aktualnego stanu wiedzy wyniki badan uzyskane przez Autora są oryginalne i innowacyjne. Biorąc pod uwagę zadeklarowany wkład (zarówno procentowy, jak i szczegółowy) w poszczególnych publikacjach, przedstawione wyniki stanowią samodzielny i oryginalny dorobek Autora.

5. Oryginalność rozwiązania postawionego problemu badawczego

Przedstawiona do recenzji praca stoi na wysokim poziomie naukowym i inżynierskim a aktualność problemu, czyli potrzeba wytłumaczenia, choćby poprzez lepszą identyfikację i wizualizację informacji

pośredniej, decyzji podejmowanych przez tzw. *black-box*, stanowi jej dużą zaletę. Autor precyzyjnie zdiagnozował kwestie wynikające z niejednoznacznej interpretacji i wyjaśnianiem modeli tworzonych przez algorytmy głębokiego uczenia. Zaproponował rozwiązanie, które za pomocą konceptów zrozumiałych dla człowieka pozwala na lepsze zrozumienie działania modelu. Dużą zaletą prowadzonych badań jest szeroki zakres wykorzystywanych danych, tj. dane medyczne, dane biometryczne oraz różnego rodzaju dane obrazowe.

Najważniejsze oryginalne osiągniecia Autora przedstawione w pracy to:

- Opracowanie podstaw teoretycznych sposobu połączenia elementów NLP z mechanizmem percepcji wzrokowej i stworzenie metodyki jego praktycznej weryfikacji;
- Przeniesienie koncepcji zadań diagnostycznych do dziedziny analizy danych graficznych, czyli wprowadzenie tzw. *visual probing tasks*;
- Sprawdzenie opracowanych koncepcji na przykładach pochodzących z obszaru widzenia komputerowego, przetwarzania danych biometrycznych i diagnostyki radiologicznej.

6. Główne wady rozprawy, słabe stron wraz z krytycznymi uwagami szczegółowymi

Publikacje, które wchodzą w skład ocenianego osiągnięcia zostały opublikowane w renomowanych czasopismach i recenzowanych materiałach konferencji naukowych, co gwarantuje, że prezentują odpowiedni poziom merytoryczny i techniczny. Dlatego też nie jest łatwo wskazać ich konkretne wady, czy też uchybienia. Poniżej wymienię jedynie kilka wybranych uwag o charakterze dyskusyjnym, które mogłyby być przyczynkiem do dyskusji w czasie obrony pracy:

- Z oczywistych względów opublikowane artykuły prezentują jedynie wybrane wycinki większego problemu naukowego i z tego powodu trudno oczekiwać w nich szerszego przeglądu literaturowego. Wydaje się, że w autoreferacie można by oczekiwać pogłębionego odniesienia do istniejących metod, poza lapidarnym spisem istniejących metod wyjaśnialnej sztucznej inteligencji. Pytanie dotyczy, jak w prezentowanych przypadkach użycia opracowanych metod zachowywałyby się metody typu SHAP czy LIME?
- Prace będące składnikami osiągnięcia są wieloautorskie a udział Doktoranta jest określony w dużej mierze przez zdefiniowanie problemów badawczych, przegląd literatury, realizację oprogramowania i prowadzenie oraz obróbkę wyników eksperymentów. W związku z tym, jak należy interpretować znacząco mniejszy udział procentowy (25%) w publikacji [a1]?
- Metody opisane w pracach [a1-a3] są ze sobą silniej związane, niż metody będące tematami prac [a4] i [a5], które wydają się być dodane jedynie jako uzupełnienie dorobku, istotne, ale jednak tylko uzupełnienie. Prosiłbym o komentarz i wyjaśnienie.
- *Visual probing* jest ciekawą koncepcją, ale uwzględnia jedynie obecność/nieobecność pewnych konceptów semantycznych, bez tworzenia ich hierarchii lub też analizy ich wzajemnych relacji (co ma miejsce np. w analizie logicznej zdania). Wydaje się, że warto by było rozważyć ten aspekt, np. poprzez prostą analizę relacji geometrycznych lub też rachunku zbiorów). Prosiłbym o odniesienie się do tej kwestii.

- Wydaje się, że metoda opisana w [a4] powinna zostać porównana z rozwiązaniami tzw. transferu stylu (np. modele CycleGAN, Pix2Pix), gdyż oba podejścia mogą prowadzić do podobnych efektów.

- W mojej ocenie słabość metody opisanej w [a5] polega na tym, że trzeba dysponować zmodyfikowanym zbiorem danych lub znać charakter różnic jakościowych aby móc ocenić sposób podejmowania decyzji przez model. Proszę o wyjaśnienie.

- W rozprawie brakuje próby generalizacji uzyskanych wyników i dyskusji słabych stron opracowanych metod – są one umieszczone w każdym z artykułów, jednak przydałaby się odpowiednia sekcja autoreferatu dotycząca tej kwestii.

- W podsumowaniu (którego *defacto* nie ma) nie przedstawiono propozycji dalszych prac badawczych stanowiących rozszerzenie osiągnięć uzyskanych w rozprawie – tak jak powyżej, informacje takie są umieszczone w podsumowaniach prac wchodzących w skład cyklu, jednak, ponownie, przydałaby się odpowiednia sekcja w autoreferacie.

- Z punktu widzenia naukowego, problem wyjaśnialności może być rozpatrywany na poziomie „podglądania" czy też wizualizacji wybranych danych tworzonych/wykorzystywanych przez model, natomiast będzie to zależało nie tylko od samego modelu ale również od danych, jakimi by „karmiony" na etapie treningu. Dlatego wydaje się, że koncepty zrozumiałe dla człowieka powinny również dotyczyć tego typu zagadnienia. Może warto rozważyć tworzenie całej struktury słów wizualnych również na etapie budowy zbiorów treningowych? Pewną inspiracją mogą być metody typu bag-of-visual-words czy też algorytmy z grupy zero-shot classification.

- Po stronie edycyjnej można zarzucić Autorowi mało staranne zredagowanie literatury na str.. 29-33. W kilku przypadkach brakuje informacji o czasopiśmie, konferencji itp. (poz. 6, 33, 45) a pozycje 34 i 35 to ten sam tekst źródłowy.

7. Znaczenie uzyskanych wyników i ich praktyczne wykorzystanie

Koncepcja badań i otrzymane wyniki, zarówno teoretyczne, jak i praktyczne są bardzo interesujące i o dużym potencjale zastosowań praktycznych. Dzięki nowym metodom wyjaśnialnej sztucznej inteligencji znacząco zwiększa się świadomość społeczna dotycząca stosowania i wiarygodności modeli tworzonych za pomocą metod AI. Moim zdaniem w każdym z obszarów badawczych (połączenie NLP i CV, anonimizacja danych biometrycznych, diagnostyką medyczna) istnieją duże możliwości aplikacyjne a opracowane metody mogłyby być z powodzeniem podstawą realizacji praktycznych.

8. Konkluzja

Recenzowana rozprawa stanowi oryginalne rozwiązanie jednoznacznie sformułowanego zagadnienia naukowego. Autor rozprawy mgr inż. Witold Oleszkiewicz w przekonujący sposób wykazał umiejętność samodzielnego prowadzenia badań naukowych, a także ich prawidłowej i wnikliwej interpretacji. Wymienione powyżej uwagi ogólne, polemiczne oraz szczegółowe nie mają znaczącego wpływu na jednoznacznie pozytywną ocenę rozprawy. Dodatkowy dorobek naukowy, niezwiązany z

realizowaną dysertacją oraz inne istotne osiągnięcia w obszarze popularyzacji nauki zaprezentowane przez Doktoranta w autoreferacie świadczą o dojrzałości naukowej i tylko podnoszą końcową ocenę.

W związku z powyższym uważam, iż przedstawiona mi do recenzji rozprawa doktorska mgr inż. Witolda Oleszkiewicza spełnia wymogi stawiane rozprawom doktorskim przedstawione w Ustawie z dnia 10 marca 2023 r. w sprawie ogłoszenia jednolitego tekstu ustawy - Prawo o szkolnictwie wyższym i nauce (Dz.U. 2023 poz. 742), art. 186 i 187 i niniejszym wnoszę o dopuszczenie jej do publicznej obrony.

Jednocześnie, biorąc pod uwagę uzyskane wyniki, fakt publikacji w wysokopunktowanych czasopismach i materiałach dziedzinowych konferencji międzynarodowych oraz ogólny wysoki poziom naukowy rozprawy, wnoszę o jej wyróżnienie.

Paweł Forczmański

POLITECHNIKA WARSZAWSKA

DYSCYPLINA NAUKOWA INFORMATYKA TECHNICZNA
I TELEKOMUNIKACJA
DZIEDZINA NAUK INŻYNIERYJNO-TECHNICZNYCH

Rozprawa doktorska

mgr inż. Witold Oleszkiewicz

Wyjaśnialne uczenie maszynowe z zastosowaniem konceptów zrozumiałych dla człowieka

Promotor
dr hab. inż. Robert Marek Nowak, prof. PW

WARSZAWA 2023

Podziękowania

Jestem niezmiernie wdzięczny mojemu promotorowi, Robertowi Nowakowi, za jego ogromne wsparcie i nieustającą wiarę w moje możliwości.

Moje dokonania nie byłyby możliwe bez nieocenionego i szczodrego wsparcia Tomasza Trzcińskiego, któremu zawdzięczam wiele wspaniałych możliwości naukowych.

Szczególnie podziękowania kieruję do Bartosza Zielińskiego, który jak nikt inny motywował mnie do pracy, dawał cenne wskazówki i razem ze mną dopracowywał publikacje w noc przed terminem.

Moje serdeczne podziękowania kieruję do wszystkich moich współautorów. W szczególności dziękuję Dominice Basaj, Igorowi Sieradzkiemu i Karolowi Piczakowi za ich zaangażowanie, wytrwałą pracę i inspirujące pomysły.

Dziękuję Krzysztofowi Gerasowi, który gościł mnie w New York University oraz Peterowi Karouz, u którego spędziłem trzy miesiące na Stanfordzie. Dzięki Wam miałem niesamowitą możliwość spojrzeć na świat nauki z nieprawdopodobnie inspirującej perspektywy.

Jestem także głęboko wdzięczny mojej żonie, dzieciom oraz rodzicom za ich cierpliwość i nieustającą miłość.

Wyjaśnialne uczenie maszynowe z zastosowaniem konceptów zrozumiałych dla człowieka

W niniejszej pracy przedstawiam serię pięciu publikacji poświęconych zagadnieniom wyjaśnialności modeli uczenia głębokiego.

Traktowanie sztucznych sieci neuronowych jako czarnych skrzynek, powoduje to, że nie ma pewności, czy decyzje modeli są podjęte na podstawie właściwych przesłanek. W moich pracach przedstawiam nowe statyczne metody wyjaśniające modele uczenia głębokiego, gdzie wyjaśnienie globalne jest generowane po wytrenowaniu modelu. Trzy prace dotyczą metody klasyfikatorów diagnostycznych, które badają informacje zawarte w reprezentacjach modeli. Jest to metoda powszechnie stosowana w przetwarzaniu języka naturalnego, jednak do tej pory nie miała ona swojego odpowiednika w widzeniu maszynowym. W moich pracach wprowadzam intuicyjną taksonomię wizualną, która zawiera znaki, słowa i zdania wizualne, analogicznie do liter, słów i zdań języka naturalnego. Dzięki temu definiuję szereg klasyfikatorów diagnostycznych, które pozwalają na badanie różnych cech reprezentacji modeli. Pokazuję przydatność metody klasyfikatorów diagnostycznych na przykładzie wyjaśniania reprezentacji samonadzorowanych. Metoda ta opiera się na obliczeniowej teorii widzenia Marra, dzięki czemu analizujemy reprezentacje za pomocą zrozumiałych dla człowieka cech wizualnych, takich jak tekstury, kolory, kształty i linie. Moje badania pokazują, że relacje między językiem a obrazem są skutecznymi i intuicyjnymi narzędziami do wyjaśniania modeli uczenia głębokiego.

W dwóch pozostałych pracach przedstawiam nową metodę do anonimizacji zbiorów danych oraz metodę wyjaśniającą działanie modeli uczenia głębokiego w diagnostyce raka piersi. Metoda do anonimizacji obrazów działa z wykorzystaniem syjamskich generatywno-przeciwstawnych sieci neuronowych i pozwala na zbadanie, czy reprezentacje modeli uczenia głębokiego zawierają informacje o tożsamości osób na obrazie. Metoda wyjaśniająca w diagnostyce medycznej bada wpływ perturbacji obrazu na decyzję lekarza oraz maszyny, dzięki czemu stwierdzamy, że modele uczenie głębokiego w dużej mierze korzystają z informacji zawartej w składowych obrazu o wysokiej częstotliwości w przestrzeni Fouriera, które to informacje są niedostrzegane przez lekarzy.

Podsumowując, wszystkie powyższe zaproponowane przeze mnie nowe metody wyjaśniające pomagają lepiej zrozumieć modele sztucznej inteligencji. Dzięki tym metodom jesteśmy w stanie zbadać obciążenie modeli, określić ich silne i słabe strony, a także wskazać które pojęcia są dla nich istotne podczas podejmowania decyzji.

Słowa kluczowe: Wyjaśnialna Sztuczna Inteligencja, Klasyfikatory Diagnostyczne, Widzenie Maszynowe, Uczenie Głębokie

Explainable machine learning using concepts understandable to humans

In this work, I present a series of five publications concerning the explainability of deep learning models.

Treating artificial neural networks as black boxes makes it impossible to determine whether the models' decisions are based on the proper premises. I present new static, post-hoc methods generating global explanations of deep learning models. Three publications concern the method of probing classifiers that examine the information encoded in model representations. This method is commonly used in natural language processing, but until now, it has not been applied in computer vision. In my works, I introduce an intuitive visual taxonomy that includes visual characters, words, and sentences analogous to characters, words, and sentences in natural language. Thanks to this, I define several probing tasks that examine various features of model representations. I show the usefulness of the diagnostic classifier method in the example of explaining self-supervised representations. This method is grounded in Marr's computational theory of vision, and it concerns visual features understandable to humans, like textures, colors, shapes, and lines. My research shows that relations between language and vision can be an effective yet intuitive tool for discovering how machine learning models work.

In the other two works, I present a new method for anonymizing datasets and a method explaining the decisions of deep learning models in breast cancer diagnosis. The image anonymization method works using Siamese generative-adversarial neural networks. It allows us to examine whether the representations of deep learning models contain information about the identity of people in the image. The explanatory method in medical diagnostics examines the impact of image perturbations on the decision of the doctor and the machine, thanks to which we conclude that deep learning models broadly use the information contained in the high-frequency image components in Fourier space, which information is invisible to doctors.

All the new explanation methods I have proposed above help us better understand deep learning models. Thanks to these methods, we can examine the biases of models, determine their strengths and weaknesses, and indicate which concepts are essential for them when making decisions.

Keywords: Explainable Artificial Intelligence, Probing Classifiers, Machine Vision, Deep Learning

Spis treści

1 Wskazanie osiągnięcia doktorskiego

Niniejszy autoreferat przedstawia osiągnięcie doktorskie zatytułowane:

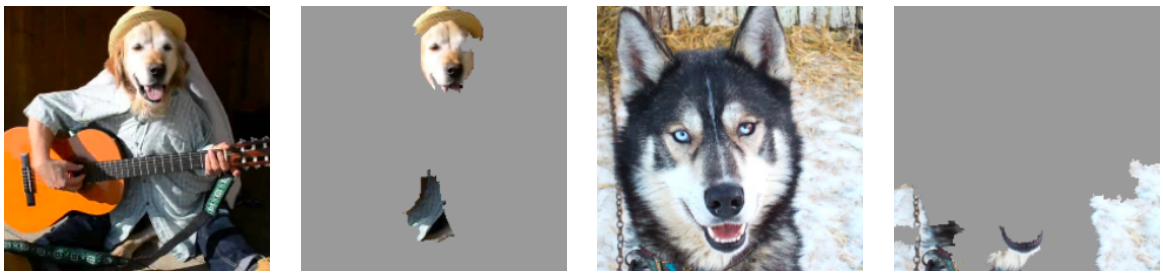
Wyjaśnialne uczenie maszynowe z zastosowaniem konceptów zrozumiałych dla człowieka.

Na osiągnięcie składa się cykl pięciu artykułów powiązanych tematycznie. Poniżej przedstawiono motywację do przeprowadzenia badań, a także listę artykułów naukowych wchodzących w skład osiągnięcia doktorskiego wraz z ich omówieniem. W rozdziale 2 zamieszczono szczegółowy opis prac badawczych wraz z dyskusją uzyskanych wyników, zaś w rozdziale 3 zamieszczone są kopie moich artykułów.

1.1 Wprowadzenie

Rozwój metod sztucznej inteligencji, w szczególności metod uczenia głębokiego, jest obecnie powszechnie dostrzegalnym zjawiskiem, mającym duży wpływ na wiele dziedzin życia. Metody te są stosowane w obszarach tak kluczowych, jak medycyna, wojskowość, bezpieczeństwo, finanse lub prawo. Efektywność metod uczenia głębokiego wynika w dużym stopniu z ogromnego rozmiaru modeli, które są parametryzowane milionami, a nawet miliardami współczynników liczbowych. Ekspertom, a tym bardziej użytkownikom takich głębokich sieci neuronowych trudno jest zrozumieć tak duże i skomplikowane modele.

Traktowanie wyuczonych sztucznych sieci neuronowych jako czarnych skrzynek, których decyzje nie są wytłumaczone, niesie ze sobą szereg zagrożeń. Nawet jeżeli model osiąga dobre wyniki, to nie ma pewności, czy jego decyzje są podjęte na podstawie właściwych przesłanek. Ten problem został dobrze opisany w literaturze. Na rysunku 1 jest przykład z pracy [36], gdzie model niepoprawnie sklasyfikował psa rasy husky jako wilka na podstawie pokrywy śnieżnej widocznej na obrazie.



(a) Golden retriever sklasyfikowany poprawnie. (b) Wyjaśnienie. (c) Pies husky sklasyfikowany jako wilk. (d) Wyjaśnienie.

Rysunek 1: Ilustracja wyniku działania metody LIME [41] do wyjaśniania decyzji sztucznej sieci neuronowej. Przykładowe obrazy sklasyfikowane poprawnie (a) i niepoprawnie (c) wraz z wyjaśnieniami (b) i (d). W przypadku zdjęcia psa rasy golden retriever model podejmuje decyzje na podstawie istotnych fragmentów obrazu, które są charakterystyczne dla danej rasy. Jednakże w przypadku psa rasy husky, decyzja klasyfikatora jest błędna, co wynika z tego, że model skupia się na pokrywie śnieżnej widocznej w tle. Rysunek pochodzi z pracy [41].

Wyjaśnialność modeli sztucznej inteligencji nie ma jednej definicji. Jest to pojęcie wieloaspektowe, zawierające w sobie wiele wątków związanych ze zrozumieniem uzasadnienia decyzji lub przewidywań dokonanych przez sztuczną inteligencję. Jedną z potrzeb ludzkich, którą ma zaspokoić wyjaśnialna sztuczna inteligencja to zaufanie. Jeżeli ludzie mają opierać swoje wybory na decyzjach modeli, to muszą im zaufać. To co może zwiększyć zaufanie użytkownika końcowego do modelu to informatywność wyjaśnienia, czyli dostarczenie szeregu dodatkowych informacji, które wzbogacą zrozumienie i wskażą czynniki stojące za predykcją. Czasem takie wyjaśnienie może mieć charakter przyczynowo-skutkowy, gdzie przyczyna stojąca za decyzją będzie podana w sposób zrozumiały dla człowieka. Szczególnym aspektem związanym z wyjaśnialną sztuczną inteligencją są kwestie etyczne. Występuje oczekiwanie, zarówno społeczne jak i prawne, że wytrenowany model będzie dostarczał predykcje zgodnie z określonym systemem wartości. W szczególności to może dotyczyć kontrolowania i eliminowania obciążeń modelu, które dyskryminują osoby ze względu na płeć, wiek czy pochodzenie. Wyjaśnialna sztuczna inteligencja ma sprawić, że system działa zgodnie ze standardami etycznymi i prawnymi, a jego decyzje są odpowiedzialne i nie powodują niebezpieczeństwa.

Zaniedbanie wyjaśnialności modeli sztucznej inteligencji może prowadzić do postępującej degradacji użyteczności takich modeli w trakcie cyklu ich życia i do podejmowania błędnych decyzji, które mogą wiązać się z dużymi kosztami ekonomicznymi i społecznymi. W związku z tym wyjaśnialna sztuczna inteligencja zdobywa obecnie coraz większą popularność w badaniach i zastosowaniach praktycznych [5, 23, 40], a jej stosowanie staje się wymogiem prawnym w wielu krajach [21].

Celem opisywanych poniżej badań było opracowanie narzędzi uczenia maszynowego, które dostarczają wyjaśnień predykcji sztucznych sieci neuronowych za pomocą konceptów zrozumiałych dla człowieka.

Poniżej przedstawiono listę prac, stanowiących osiągnięcie naukowe. Dla każdej z prac podano aktualną liczbę punktów i wskaźnik IF, a także liczbę cytowań na podstawie Google Scholar.

[A1] Dominika Basaj*, Witold Oleszkiewicz*, Igor Sieradzki, Michał Górszczak, Barbara Rychalska, Tomasz Trzciński, Bartosz Zieliński.
Explaining Self-Supervised Image Representations with Visual Probing.
International Joint Conference on Artificial Intelligence (IJCAI 2021), DOI:10.24963/ijcai.2021/82, p. 592–598, 2021.
Punkty MEiN: 200, Core rank: A*.
Cytowania: 15 (Google Scholar)

Wkład: Brałem istotny udział w definiowaniu problemu badawczego. Dokonałem przeglądu literatury w zakresie zadań diagnostycznych w widzeniu maszynowym oraz metod samonadzorowanych uczenia maszynowego. Miałem kluczowy udział w zaproponowaniu i zaadaptowaniu zadań diagnostycznych do wyjaśniania reprezentacji modeli uczenia głębokiego, w tym miałem istotny udział w zdefiniowaniu kluczowych zadań diagnostycznych: *Word Content*, które bada zawartość semantyczną reprezentacji oraz *Sentence Length*, które bada złożoność semantyczną reprezentacji. Miałem istotny udział w zdefiniowaniu mapowania pomiędzy widzeniem komputerowym a przetwarzaniem ję-

zyka naturalnego. Zaprojektowałem większą część eksperymentów, w tym eksperymenty związane z efektywnym przygotowaniem reprezentacji modeli samonadzorowanych oraz etykiet do zadań diagnostycznych oraz walidacją zadań diagnostycznych. Zaimplementowałem większą część algorytmów, w tym algorytmy zadania diagnostycznego *Word Content*, *Sentence Length* oraz dużą część algorytmów zadania diagnostycznego *Character Bin*. Przeprowadziłem zdecydowaną większość eksperymentów, wraz z analizą i opracowaniem wyników. Przygotowałem dane do wszystkich eksperymentów badających dokładność zadań diagnostycznych, przeprowadziłem badania zadań diagnostycznych *Word Content* oraz *Sentence Length*. Miałem istotny udział w zaprojektowaniu, przeprowadzeniu i analizie badań ankietowych. Przygotowałem pytania oraz ilustracje do ankiety, nadzorowałem proces przeprowadzenia badań ankietowych, opracowałem sposób interpretacji odpowiedzi uczestników ankiety. Miałem istotny wkład w redagowanie pracy, opisałem szczegóły przeprowadzanych eksperymentów, przygotowałem ilustracje oraz tabele, opisałem wnioski z eksperymentów. **Mój wkład ogólny szacuję na 25%.**

[A2] Witold Oleszkiewicz, Dominika Basaj, Tomasz Trzciński, Bartosz Zieliński.
Which Visual Features Impact the Performance of Target Task in Self-supervised Learning?
International Conference on Computational Science (ICCS 2022), DOI:10.1007/978-3-031-08751-6_24, p. 331–344, 2022.
Punkty MEiN: 140, Core rank: A.
Cytowania: 0 (Google Scholar)

Wkład: Brałem kluczowy udział w definiowaniu problemu badawczego. Dokonałem przeglądu literatury. Miałem kluczowy udział w zaproponowaniu algorytmu do tworzenia amnezyjnych zadań diagnostycznych. Zaprojektowałem wszystkie eksperymenty, w tym eksperymenty związane z usuwaniem informacji semantycznej o słowach wizualnych z reprezentacji. Zaimplementowałem albo zaadaptowałem wszystkie algorytmy, w tym algorytmy badające zmiany związane z usuwaniem informacji semantycznej o słowach wizualnych z reprezentacji. Przeprowadziłem wszystkie eksperymenty, wraz z analizą i opracowaniem wyników. Miałem bardzo istotny wkład w redagowanie pracy, przygotowałem wszystkie ilustracje, tabele oraz wykresy.**Mój wkład ogólny szacuję na 75%.**

[A3] Witold Oleszkiewicz, Dominika Basaj, Igor Sieradzki, Michał Górszczak, Barbara Rychalska, Koryna Lewandowska, Tomasz Trzciński, Bartosz Zieliński.
Visual Probing: Cognitive Framework for Explaining Self-Supervised Image Representations.
IEEE Access, vol. 11, pp. 13028-13043, 2023, DOI: 10.1109/ACCESS.2023.3242982.
Punkty MEiN: 100, IF: 3.476
Cytowania: 2 (Google Scholar)

Wkład: Brałem istotny udział w definiowaniu problemu badawczego. Dokonałem przeglądu literatury w zakresie zadań diagnostycznych w widzeniu maszynowym oraz metod samonadzorowanych uczenia maszynowego. Miałem kluczowy udział w zaproponowaniu i zaadaptowaniu nowych zadań diagnostycznych do wyjaśniania reprezentacji modeli

uczenia głębokiego, a przede wszystkim zdefiniowałem i zaimplementowałem nowe zadanie *Mutual Word Content*, które porównuje semantyczną zawartość pary reprezentacji. Miałem istotny udział w adaptowaniu teorii widzenia Marra do problemu badania reprezentacji modeli uczenia głębokiego, w tym weryfikowałem wiele początkowych hipotez, sprawdzających w jaki sposób można wykorzystać teorię widzenia Marra do optymalnej klasteryzacji słów wizualnych. Zaprojektowałem i przeprowadziłem zdecydowaną większość eksperymentów: przygotowałem modele uczenia samonadzorowanego, wygenerowałem reprezentacje samonadzorowane oraz etykiety do zbadania dokładności zadań diagnostycznych. Zaimplementowałem większą część algorytmów, w tym w całości algorytmy do zadań diagnostycznych: *Word Content*, *Mutual Word Content*. Przeprowadziłem zdecydowaną większość analiz wyników eksperymentów, wraz z opracowaniem wyników i przygotowaniem tabel, wykresów i ilustracji do publikacji. Miałem kluczowy udział w zaprojektowaniu, przeprowadzeniu i analizie badań ankietowych. Miałem istotny wkład w redagowanie pracy. **Mój wkład ogólny szacuję na 65%.**

[A4] Witold Oleszkiewicz, Peter Kairouz, Karol Jerzy Piczak, Ram Rajagopal, Tomasz Trzciński.
Siamese Generative Adversarial Privatizer for Biometric Data.
Asian Conference on Computer Vision (ACCV 2018), DOI:10.1007/978-3-030-20873-8_31, p. 482–497, 2019.
Punkty MEiN: 70, Core rank: B.
Cytowania: 22 (Google Scholar)

Wkład: Brałem istotny udział w szczegółowym rozwinięciu problemu badawczego. Zaprojektowałem architekturę rozwiązania układu sieci do anonimizacji obrazów. Zaprojektowałem większą część eksperymentów, w tym eksperymenty do modyfikacji obrazów, tak aby usunąć z nich informacje o tożsamości osoby. Zaimplementowałem wszystkie algorytmy w pracy. Przeprowadziłem wszystkie eksperymenty, wraz z analizą i opracowaniem wyników. Miałem istotny wkład w redagowanie pracy. **Mój wkład ogólny szacuję na 50%.**

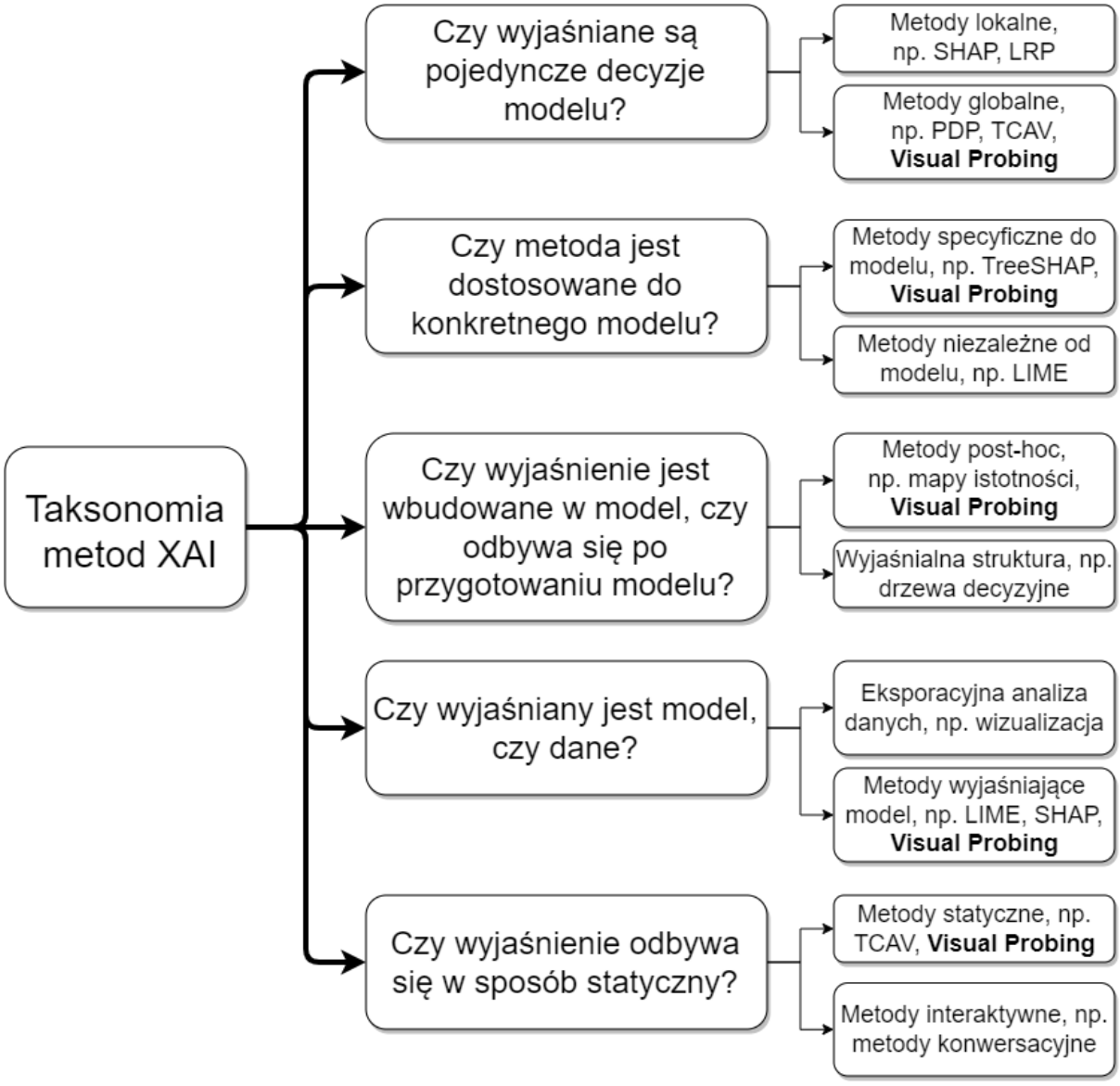
[A5] Taro Makino, Stanisław Jastrzębski, Witold Oleszkiewicz, Celin Chacko, Robin Ehrenpreis, Naziya Samreen, Chloe Chhor, Eric Kim, Jiyon Lee, Kristine Pysarenko, Beatriu Reig, Hildegard Toth, Divya Awal, Linda Du, Alice Kim, J. Park, Daniel K. Sodickson, Laura Heacock, Linda Moy, Kyunghyun Cho, Krzysztof J. Geras.
Differences between human and machine perception in medical diagnosis.
Scientific Reports (12), ISSN 2045-2322, DOI:10.1038/s41598-022-10526-z, p. 1–13, 2022.
Punkty MEiN: 140, IF: 4.996
Cytowania: 17 (Google Scholar)

Wkład: Brałem udział w definiowaniu problemu badawczego. Zaprojektowałem część eksperymentów związanych z filtrowaniem obrazów w domenie częstotliwości. Zaimplementowałem część algorytmów związanych z filtrowaniem obrazów w domenie częstotli-

wości. Przeprowadziłem część eksperymentów, wraz z analizą i opracowaniem wyników. Pomagałem z redagowaniem części pracy. **Mój wkład ogólny szacuję na 15%.**

1.2 Zakres badań

Wyjaśnialne uczenie maszynowe jest obecnie szybko rozwijającą się dziedziną, w związku z czym należy na początek zakreślić taksonomię metod wyjaśniających.



Rysunek 2: Proponowana taksonomia wyjaśnialnego uczenia maszynowego oparta na pytaniach o poziom (lokalny lub globalny), przedmiot (dane lub model), sposób (wbudowany w strukturę lub post-hoc oraz statyczny lub interaktywny) i aplikowalność (specyficzny dla konkretnego modelu lub niezależny od modelu) wyjaśnień. Moja metody rozwiniętej w ramach cyklu publikacji [A1, A2, A3] jest zaznaczona za pomocą pogrubienia na powyższym wykresie.

Jak przedstawiono na Rysunku 2, taksonomia wyszczególnia metody: wyjaśniające dane lub

model, metody statyczne oraz metody interaktywne, metody specyficzne dla konkretnego typu modelu lub niezależne od typu modelu, metody bazujące na interpretowalnej strukturze modelu lub metody wyjaśnialne post-hoc, a także metody wyjaśniające lokalnie lub globalnie.

Metody wyjaśniające dane koncentrują się wyłącznie na analizie eksploracyjnej danych, które są używane do trenowania modeli i nie analizują samych modeli. Takie metody wykorzystują przede wszystkim techniki wizualizacji. Metody interaktywne, w przeciwieństwie do statycznych, korzystają z informacji zwrotnej uzyskiwanej od użytkownika w trakcie korzystania z metody. Przegląd podejść interaktywnych w wyjaśnialnej sztucznej inteligencji znajduje się w pracy [45]. Metody lokalne stosuje się do wyjaśniania pojedynczą decyzję modelu, zaś metody globalne tłumaczą zachowania całego modelu. Przykładem metod lokalnych są np. metody SHAP [33], LIME [41], LRP [7], natomiast do popularnych metod globalnych można zaliczyć metody TCAV [29], PDP [17].

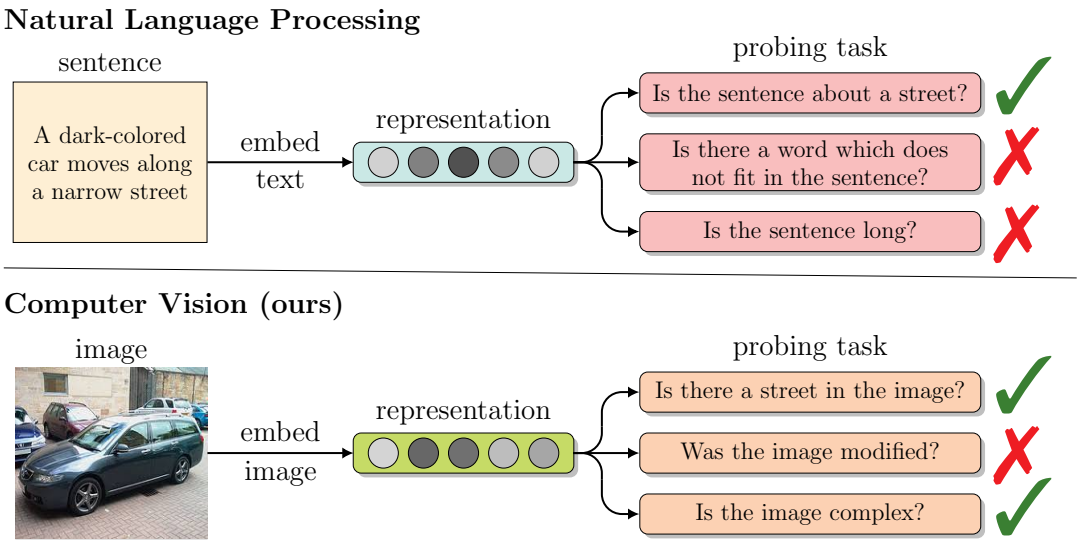
Wszystkie prace przedstawione jako osiągnięcie naukowe prezentują metody statyczne wyjaśniające modele uczenia głębokiego, gdzie wyjaśnienie jest generowane po wytrenowaniu modelu i jest to wyjaśnienie globalne.

W pracy [A1] wprowadzam zadania diagnostyczne (ang. probing tasks) do analizy reprezentacji obrazów, uzyskanych za pomocą metod samonadzorujących się (ang. self-supervised). Jest to jedna z pierwszych prac, która podejmuje próbę wyjaśnienia popularnych obecnie metod samonadzorujących się. Jest to też jedna z pierwszych prób zastosowania metody zadań diagnostycznych w dziedzinie wizji komputerowej, która to metoda często jest stosowana do objaśniania reprezentacji w dziedzinie przetwarzania języka naturalnego [12] (jak pokazano na Rysunku 3). W ramach pracy [A1] zdefiniowałem przybliżoną taksonomię, która czerpiąc inspiracje z dziedziny przetwarzania języka naturalnego, definiuje nowe pojęcia – wizualne odpowiedniki słów i zdań, w celu skutecznego ich zastosowania do wyjaśniania reprezentacji obrazów. W wyniku przeprowadzonych eksperymentów stwierdziliśmy, że reprezentacje metod samonadzorujących się zawierają informacje semantyczne opisujące zawartość, złożoność i spójność obrazu.

Praca [A3] rozszerza badania z pracy [A1], wprowadzając systematykę słów wizualnych korzystając z kognitywistycznej teorii percepcji wzrokowej Marra [35]. Systematyka Marra wprowadza sześć cech wizualnych: jasność, kolor, tekstura, linie, kształt i forma. Dzięki tej systematyce możemy opisać najistotniejsze elementy obrazu dla analizowanych modeli w sposób bardziej zrozumiały dla człowieka. W rezultacie przeprowadzonych eksperymentów stwierdziliśmy, że występowanie na obrazie słów wizualnych związanych z liniami i formami ma największy wpływ na decyzje modeli samonadzorujących się.

W pracy [A2] rozszerzam powyższe metody wyjaśniające, wprowadzając amnezyjne zadania diagnostyczne. W ten sposób badamy nie samą obecność wprowadzonych wcześniej słów wizualnych w reprezentacji, lecz wpływ obecności tych słów na decyzje modeli. W procesie stosowania amnezyjnych zadań diagnostycznych przeprowadzamy interwencje, które usuwają informacje o konkretnych słowach wizualnych z reprezentacji obrazu, a następnie mierzymy, jak to wpływa na jakość klasyfikacji badanego modelu. Z przeprowadzonych eksperymentów dla różnych modeli samonadzorujących się wynika, że usunięcie z reprezentacji informacji dotyczącej słów wizualnych związanych z formą zmniejszają dokładność klasyfikacji bardziej, niż usuwanie słów wizualnych związanych z teksturą.

W pracy [A4] przedstawiam metodę do anonimizacji zbiorów danych za pomocą syjamskich



Rysunek 3: Klasyfikatory diagnostyczne stosowane w dziedzinie przetwarzaniu języka naturalnego sprawdzają, czy w wytrenowanej reprezentacji są zawarte informacje o konceptach zrozumiałych dla człowieka. W pracy [A1] wprowadzam taksonomię, która pozwala tworzyć zadania diagnostyczne do badania zawartości reprezentacji obrazów. Rysunek pochodzi z pracy [A1].

generatywno-przeciwstawnych sieci neuronowych. Wyniki eksperymentalne wykazują na to, że możliwa jest taka modyfikacja reprezentacji obrazów oraz samych obrazów, która zapewni równowagę pomiędzy anonimizacją obrazów a zachowaniem użyteczności danych do trenowania modeli uczenia maszynowego. Dzięki temu uzyskaliśmy zrozumienie reprezentacji modeli, w szczególności uzyskaliśmy odpowiedź na pytanie: czy reprezentacje modeli uczenia głębokiego mogą zawierać informacje o tożsamości osób widocznych na obrazie oraz czy możliwa jest modyfikacja tych informacji.

W pracy [A5] przeprowadziliśmy analizę wyjaśniającą działanie modeli uczenia głębokiego w diagnostyce medycznej. Skoncentrowaliśmy się na próbie zrozumienia, czy sieci neuronowe oraz lekarze podejmują decyzje diagnostyczne na podstawie tych samych przesłanek. W tym celu zaproponowaliśmy metodę, w której zbadano wpływ perturbacji obrazu na decyzję człowieka oraz maszyny w kontekście zadania wykrywania raka piersi na podstawie obrazów mammograficznych. Po przeprowadzeniu eksperymentów stwierdziliśmy, że modele uczenie głębokiego w dużej mierze korzystają z informacji zawartej w składowych obrazu o wysokiej częstotliwości w przestrzeni Fouriera, które to informacje są niedostrzegane przez radiologów.

Wnioski z przeprowadzonych przeze mnie badań, wskazują, że wszystkie powyższe nowe metody pomagają lepiej zrozumieć modele sztucznej inteligencji. Dzięki tym metodom jesteśmy w stanie zbadać obciążenie modeli, określić ich silne i słabe strony, a także wskazać które koncepty są dla nich istotne podczas podejmowania decyzji.

2 Szczegółowy opis wyników

2.1 Wyjaśnialne uczenie maszynowe do analizy samonadzorujących się reprezentacji obrazu za pomocą klasyfikatorów diagnostycznych [A1, A3]

Reprezentacje wizualne mają kluczowe znaczenie w wielu współczesnych zastosowaniach uczenia maszynowego, takich jak wyszukiwanie wizualne [43], klasyfikacja obrazów [31] oraz odpowiadanie na pytania na podstawie obrazu [4]. Jednakże uczenie się reprezentacji w sposób nadzorowany, z wykorzystaniem dużych zbiorów danych jest problematyczne, przede wszystkim ze względu na konieczność etykietowania danych. Jest to czynność bardzo pracochłonna, kosztowna oraz podatna na błędy. W związku z tym w uczeniu się reprezentacji obecnie popularność zdobywają metody samonadzorowane, które są w stanie osiągnąć podobną skuteczność przy znacznie mniejszym zapotrzebowaniu na dane etykietowane [10, 11, 22, 25]. W przeciwieństwie do metod uczenia z nadzorem, metody samonadzorowane nie potrzebują informacji o etykiecie próbki na początkowym etapie uczenia się. Zamiast tego te metody korzystają z kontrastowej funkcji kosztu, która mierzy podobieństwo pomiędzy próbkami w przestrzeni reprezentacji i odróżnia pary reprezentujące zmodyfikowane wersje tej samej próbki od pary, której elementy pochodzą z różnych próbek.

Istnieje wiele prac badających wyjaśnialność reprezentacji obrazów [15, 27, 46]. Jednakże większość z tych prac dotyczy wyjaśnialności reprezentacji powstałych w procesie uczenia z nadzorem [49], a nie dotyczy reprezentacji samonadzorowanych. Ponadto większość proponowanych podejść polega na analizie wpływu poszczególnych pikseli obrazu na decyzję końcową modelu [2, 42], zaś występujące na obrazie koncepty semantyczne, które są zrozumiałe dla człowieka, nie są brane pod uwagę podczas wyjaśniania.

Opisane poniżej moje publikacje mają na celu przezwyciężyć te ograniczenia. W pracy [A1] oraz w jej rozszerzeniu [A3] zaczerpnęliśmy inspirację z prostej obserwacji, że to właśnie język jest używany przez ludzi jako naturalne narzędzia do wyjaśniania tego, czego dowiadujemy się o świecie za pomocą widzenia [32]. Dlatego, biorąc pod uwagę że te same algorytmy uczenia maszynowego mogą być z powodzeniem stosowane do rozwiązywania zarówno zadań związanych z przetwarzaniem obrazu, jak i z językiem naturalnym [14, 9], postulujemy, że metody używane do analizy reprezentacji tekstowej mogą być również wykorzystywane do badania reprezentacji wizualnych.

Bardzo popularnymi narzędziami do wyjaśniania reprezentacji tekstowych są klasyfikatory diagnostyczne (ang. probing classifiers) [12]. Klasyfikator diagnostyczny w dziedzinie przetwarzania języka naturalnego (ang. Natural Language Processing, NLP) to prosty klasyfikator, sprawdzający czy dana reprezentacja tekstowa koduje określoną właściwość, taką jak np. treść zdania, jego długość albo spójność semantyczną, nawet jeżeli te cechy nie są bezpośrednio brane pod uwagę podczas definiowania funkcji kosztu w trakcie uczenia się modelu. Analizując dokładność klasyfikatorów diagnostycznych, można zweryfikować, czy badana reprezentacja zawiera określone informacje. Podczas gdy klasyfikatory diagnostyczne są prostymi, intuicyjnymi i szeroko stosowanymi narzędziami w NLP, ich zastosowanie w dziedzinie wizji maszynowej

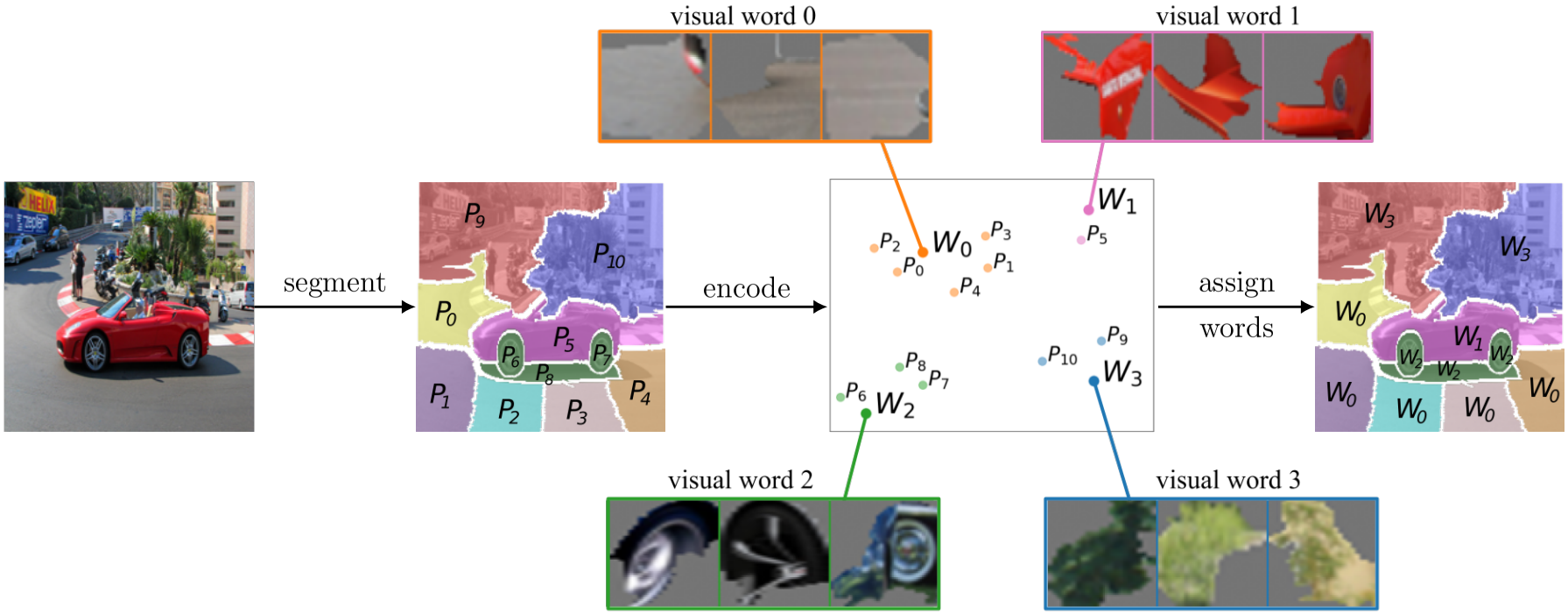
(ang. computer vision, CV) jest ograniczone [3], głównie ze względu na brak odpowiednich analogii między modalnościami tekstowymi a wizualnymi.

2.1.1 Odwzorowanie pomiędzy językiem naturalnym a widzeniem maszynowym

W pracy [A1] wprowadziłem intuicyjne odwzorowanie pomiędzy językiem naturalnym a obrazem. Należy zaznaczyć, że nie jest to ścisłe odwzorowanie, lecz wyłącznie inspiracja, dzięki której możliwe jest zastosowanie klasyfikatorów diagnostycznych z dziedziny NLP w domenie widzenia maszynowego. W celu zaadaptowania klasyfikatorów diagnostycznych do badania reprezentacji obrazów zaproponowałem wizualną taksonomię, która zawiera znaki, słowa i zdania wizualne, korzystając z analogii do liter, słów i zdań języka naturalnego. W danej taksonomii obrazy są traktowane analogicznie do zdań w języku naturalnym. Tak jak zdanie jest uporządkowaną grupą słów, które zawiera zrozumiałą dla człowieka treść, tak obraz zawiera wyodrębnione elementy, które mogą być odbierane i rozumiane przez człowieka.

Elementami obrazu w danym odwzorowaniu są nienakładające się na siebie superpiksele, które składają się z grupy niepodzielnych pikseli mających wspólne cechy lub tworzących wyodrębnione dla człowieka obiekty. Podział obrazu na superpiksele może odbywać się np. za pomocą algorytmu SLIC [1]. W efekcie otrzymujemy obraz zbudowany z superpikseli o określonej pozycji i znaczeniu, analogicznie do zdania zbudowanego ze słów. Każdy superpiksel zawiera określoną liczbę pikseli, podobnie jak słowa składają się ze znaków. Tym samym uzyskujemy intuicyjne mapowanie pomiędzy domeną wizualną a domeną tekstową.

Do powyższego odwzorowania należy dodać jeszcze jedno uszczegółowienie. Superpiksele różnią się koncepcyjnie od swoich językowych odpowiedników, jakimi są słowa, w jednym istotnym aspekcie: superpiksele nie powtarzają się pomiędzy różnymi obrazami, podczas gdy w tekście słowa często powtarzają się w różnych zdaniach. Z tego powodu słowo wizualne zostało zdefiniowane nie jako pojedynczy superpiksel, lecz jako klaster grupujący podobne superpiksele w przestrzeni reprezentacji. W ten sposób każdemu superpikselowi można przypisać jedno słowo wizualne, wyznaczając najbliższy temu superpikselowi środek takiego klastra. Tworzenie słownika słów wizualnych (czyli klasteryzacja superpikseli w przestrzeni reprezentacji) może być dokonane np. za pomocą metod TCAV (ang. Testing with Concept Activation Vectors) [29] oraz ACE (ang. Automatic Concept-based Explanations) [20]. Metody te generują wysokopoziomowe koncepty, które są zrozumiałe dla człowieka. Takie podejście wymaga dodatkowej sztucznej sieci neuronowej wytrenowanej w sposób nadzorowany, która wygeneruje przestrzeń reprezentacji, dzięki czemu uzyskamy miarę odległości pomiędzy superpikselami. Uzyskane w ten sposób słowa wizualne nie są w żaden sposób zależne od reprezentacji metod samonadzorujących się, które są przedmiotem badań w pracach [A1, A3]. Podsumowując, proces dzielenia obrazu na słowa wizualne składa się z trzech etapów: segmentacji obrazu na superpiksele, uzyskania reprezentacji superpikseli oraz przypisania superpikseli do słów wizualnych, co jest pokazane na Rysunku 4.



Rysunek 4: Podziału obrazu na słowa wizualne. Najpierw segmentujemy obraz na superpiksele $P_0, P_1, .., P_{10}$. Następnie, dla każdego superpiksela wyznaczana jest reprezentacja. Reprezentacje przypisujemy do najbliższego środka klastra reprezentacji wcześniej wygenerowanych słów wizualnych W_0, W_1, W_2, W_3. Rysunek pochodzi z pracy [A1].

2.1.2 Obliczeniowa teoria widzenia Marra

W przeciwieństwie do słów z języka naturalnego, słowa wizualne wprowadzone w pracy [A1] nie mają dobrze zdefiniowanego znaczenia, które jest wymagane do przeprowadzenia dogłębnej analizy reprezentacji. Dlatego w pracy [A3] wprowadzam systematykę kognitywno-wizualną, czerpiąc inspiracje z założenia, że tworzenie słów wizualnych może być podobne do procesu formowania pojęć. Pojęcia w kognitywistyce są rozumiane jako konstrukcje myślowe odzwierciedlające zbiór podobnych rzeczy, zjawisk, itp. Innymi słowy, pojęcia mogą być tworzone w odniesieniu do cech, które stanowią podobieństwo między badanymi obiektami. W pracy [A3] cechy, które mogłyby stanowić podstawę do powstania słów wizualnych, zdefiniowano na podstawie obliczeniowej teorii widzenia Marra [30, 34]. David Marr w ramach tej teorii założył, że w procesie percepcji człowiek dostrzega charakterystyczne cechy strukturalne obiektów, które są następnie porządkowane w szereg reprezentacji wizualnych. Trzy główne reprezentacje to: „szkic pierwotny", „szkic 2.5D" i „model 3D" [34]. Szkic pierwotny to dwuwymiarowy obraz, który zawiera informacje o zmianach natężenia światła, krawędziach, kolorach i teksturach [16, 37]. Szkic 2.5D przedstawia głównie dwuwymiarowe kształty i ich orientację względem obserwatora. Na tym etapie uzyskuje się poczucie głębi obrazu [30]. Wreszcie model 3D jest reprezentacją odpowiedzialną za wyobrażenie obiektu z różnych perspektyw. Obejmuje to również powierzchnie, które są niewidoczne dla obserwatora.

W pracy [A3] zdecydowaliśmy się na wykorzystanie sześciu cech z obliczeniowej teorii widzenia Marra: jasności, koloru, tekstury, linii, kształtu oraz formy, w celu bardziej zrozumiałego dla człowieka opisu słów wizualnych. Przykładowe słowa wizualne przedstawiam na Rysunku 5. Wstępna analiza poszczególnych słów wizualnych pokazała, że wyżej wymienione sześć cech z teorii Marra bardzo trafnie opisują poszczególne typy wygenerowanych słów wizualnych z prac [A1, A3]. W celu potwierdzenia tych wstępnych obserwacji przeprowadziliśmy badania ankietowe. Wyniki tych badań potwierdzają nasze początkowe obserwacje oraz pozwalają na skategoryzowanie wszystkich słów wizualnych według cech Marra. Dzięki temu możemy lepiej ustalić znaczenie słów wizualnych, co pomaga w lepszym zrozumieniu reprezentacji modeli uczenia głębokiego.

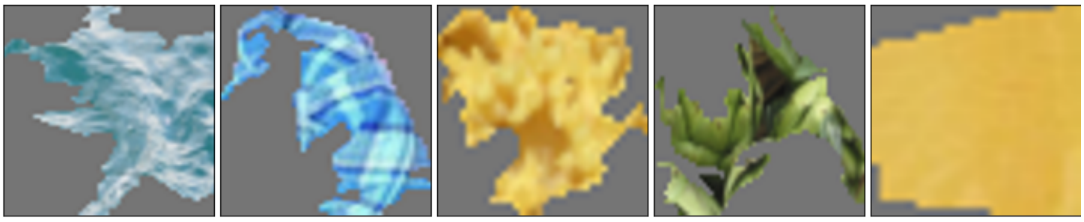
2.1.3 Klasyfikatory diagnostyczne

Możliwość opisu obrazu za pomocą słów wizualnych pozwala nam na zbadanie reprezentacji tych obrazów za pomocą klasyfikatorów diagnostycznych. W pracy [A3] przedstawiłem pięć różnych klasyfikatorów diagnostycznych, część z których została zaadaptowana z dziedziny NLP [12, 15].

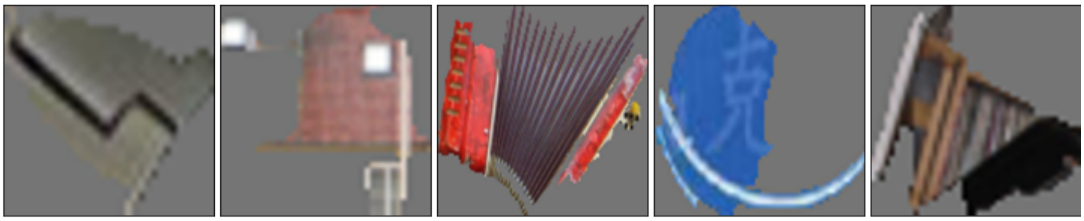
- *Word Content* (WC) to klasyfikator diagnostyczny, który ma na celu wykrywanie obecności słów wizualnych na obrazie. Wejściem klasyfikatora jest reprezentacja obrazu, a wyjściem są etykiety binarne, które określają obecność słowa wizualnego na obrazie. Tym samym dany klasyfikator diagnostyczny Word Content pozwala porównać różne metody uczenia głębokiego pod kątem występowania konceptów wizualnych w ich reprezentacjach.
- *Sentence Length* (SL) ma na celu rozróżnienie pomiędzy prostymi i skomplikowanymi obrazami. Wejściem klasyfikatora jest reprezentacja obrazu, a wyjściem jest liczba unikalnych

(a) Jasność.

(b) Kolor.



(c) Tekstura.



(d) Linie.

(e) Kształt.

(f) Forma.

Rysunek 5: Przykładowe superpiksele reprezentujące sześć różnych cech wizualnych z obliczeniowej teorii percepcji wzrokowej Marra.

słów wizualnych występujących na obrazie. Tym samym dany klasyfikator pozwala ocenić, czy reprezentacja zawiera informacje o tym jak bardzo skomplikowany jest obraz pod względem liczby różnych słów wizualnych.

- *Character Bin* (CB) to klasyfikator diagnostyczny który bada nie cały obraz, lecz poszczególne superpiksele, odpowiadające poszczególnym słowom wizualnym. Klasyfikator ten bada, czy reprezentacja superpiksela zawiera informacje o złożoności słowa wizualnego. Wejściem klasyfikatora jest reprezentacja superpiksela, a wyjściem są *a)* kompaktowość superpiksela S, zdefiniowana jako pole powierzchni superpiksela $A(S)$ podzielone przez pole $A(C)$ koła C o takim samym obwodzie jak superpiksel S; lub *b)* zmienność (ang. Inter-Cluster Variance [6]) superpiksela S, zdefiniowana jako średnie odchylenie standardowe rozkładu wartości dla poszczególnych kanałów superpiksela w przestrzeni RGB. Tym samym dany klasyfikator pozwala ocenić, czy reprezentacja zawiera informacje o kształcie i zróżnicowaniu kolorystycznym słów wizualnych.
- *Semantic Odd Man Out* (SOMO) to klasyfikator diagnostyczny, którego celem jest rozpoznanie, czy obraz został zmodyfikowany. Analizowane są modyfikacje polegająca na zastąpieniu jednego superpiksela innym superpikselem o podobnym kształcie. Podstawiany superpiksel pochodzi z innego obrazu i musi odpowiadać innemu słowu wizualnemu. Klasyfikator przyjmuje na wejście reprezentację obrazu oryginalnego lub zmodyfikowanego i zwraca etykietę binarną informującą o tym, czy dany obraz był modyfikowany. Tym samym dany klasyfikator bada, czy reprezentacja obrazu jest wrażliwa na zaburzenia w jego spójności.
- *Mutual Word Content* (MWC) to klasyfikator, który identyfikuje, które słowa wizualne zbliżają do siebie reprezentacje różnych obrazów. Wejściem klasyfikatora jest para reprezentacji dwóch różnych obrazów, a wyjściem są etykiety binarne, które oznaczają, że dane słowo wizualne występuje jednocześnie na dwóch badanych obrazach. Jeżeli dokładność tego klasyfikatora spada wraz ze wzrostem odległości pomiędzy parą reprezentacji, oznacza to, że współwystępowanie danego słowa wizualnego w obu reprezentacjach jest skorelowane ze zbliżaniem się tych reprezentacji do siebie.

2.1.4 Wyniki i wnioski

Wyniki zbiorcze z pracy [A3] są w Tabeli 1. Wszystkie analizowane reprezentacje samonadzorowane (MoCo v1, SimCLR v2, BYOL i SwAV) zachowują informacje o informacji semantyce (słowach wizualnych), złożoności i spójności obrazu. Co więcej, dokładność klasyfikatorów diagnostycznych nie jest skorelowana z dokładnością klasyfikacji zadania docelowego na podstawie tych reprezentacji.

Wyniki z Tabeli 1 dla klasyfikatora *Word Content* pokazują, że reprezentacje badanych metod samonadzorowanych zawierają informacje semantyczne o słowach wizualnych. Chociaż analizowane metody samonadzorowane mają różną dokładność klasyfikacji w zadaniu docelowym, to wszystkie charakteryzują się podobnym poziomem wydobywania informacji semantycznej o słowach wizualnych. Odkrycie to potwierdza wniosek z pracy [18], że wiedza semantyczna tylko częściowo przyczynia się do skuteczności w klasyfikacji. Okazuje się jednak, że obecność niektórych słów wizualnych jest lepiej rozpoznawana przez klasyfikatory diagnostyczne *WC*. W ogólności reprezentacje samonadzorowane mają większą wiedzę o wizualnych słowach

	Zadanie docelowe	Klasyfikatory diagnostyczne					
		WC	**MWC**	**SL**	**CB shape**	**CB color**	**SOMO**
MoCo v1	0.606	0.793	0.763	0.771	0.797	0.872	0.850
SimCLR v2	0.717	**0.811**	**0.777**	**0.775**	**0.850**	0.876	**0.878**
BYOL	0.723	0.803	0.775	0.770	0.844	**0.893**	0.845
SwAV	**0.753**	0.802	0.776	0.769	0.842	0.879	0.856

Tabela 1: Wynik metryki AUC dla klasyfikatorów diagnostycznych oraz dokładność klasyfikacji w zadaniu docelowym dla reprezentacji samonadzorowanych. Wyniki pochodzą z pracy [A3].

zawierających formy i linie niż o tych słowach, które zawierają kształty i tekstury.

Wyniki dla klasyfikatora diagnostycznego *MWC* z kolei pokazują, że to samo wizualne słowo współwystępujące w parze obrazów zwykle jest związane ze zbliżaniem się tych reprezentacji. Dotyczy to prawie wszystkich badanych słów wizualnych (45 z 50), a szczególnie tych, które zawierają skomplikowane formy i linie oraz zieleń. Wyniki dla klasyfikatorów diagnostycznych *SL*, *CB* oraz *SOMO* pokazują, że samonadzorowane reprezentacje zawierają informacje o złożoności semantycznej obrazu oraz poszczególnych słów wizualnych, a także zawierają informacje o spójności semantycznej obrazu.

W ogólności metody zaproponowane w pracach [A1, A3] pozwalają zrozumieć, które koncepty wyższego poziomu zostały wyuczone przez model. Dzięki temu możemy korzystać z wytrenowanych modeli, mając większą świadomość preferencji charakterystycznych dla każdej z metod. Wyniki eksperymentów potwierdzają skuteczność i przydatność tych metod w lepszym zrozumieniu reprezentacji metod samonadzorowanych. W ramach badań zweryfikowaliśmy, czy reprezentacje te zawierają informacje o semantyce, złożoności oraz spójności obrazów. Ponadto szczegółowa analiza każdego klasyfikatora diagnostycznego pozwoliła ujawnić różnice w reprezentacjach generowanych różnymi metodami.

Wyjaśnienia dostarczane przez te metody wykraczają poza wyjaśnienia dotyczące poszczególnych próbek. Pozwalają one zidentyfikować zrozumiałe dla człowieka słowa wizualne, odpowiadające konceptom wyższego poziomu. Zaletą metody klasyfikatorów diagnostycznych jest mierzalność, która pozwala porównać, w jakim stopniu poszczególne metody kodują informacje o różnych pojęciach w swoich reprezentacjach. Potencjalnie metoda klasyfikatorów diagnostycznych może mieć szersze zastosowanie, nie tylko do wyjaśniania metod uczenia się samonadzorowanego, ale do wyjaśniania dowolnych reprezentacji. Zaletą tego podejścia jest to, że jest ono wystarczająco ogólne, przez co możliwe jest np. użycie różnych algorytmów segmentacji, co doprowadzi do powstanie innego słownika słów wizualnych.

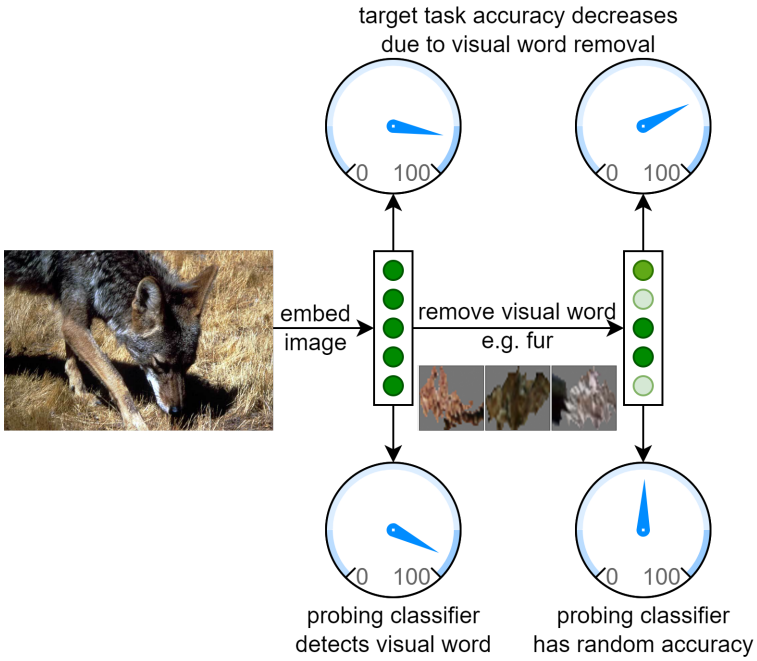
2.2 Amnezyjne klasyfikatory diagnostyczne do wyjaśniania decyzji klasyfikatorów [A2]

Metody zaproponowane w pracach [A1, A3] pozwalają na zmierzenie w jakim stopniu informacje o poszczególnych konceptach są zawarte w reprezentacjach wyuczonych za pomocą samonadzorowanych metod uczenia maszynowego. Nie daje to jednak wyjaśnienia, czy badane

koncepty są istotne i brane pod uwagę podczas klasyfikowania nowych próbek. W związku z tym praca [A2] uzupełnia tę lukę, wprowadzając metodę amnezyjnych klasyfikatorów diagnostycznych (ang. amnesic visual probing). W ramach tej metody z reprezentacji usuwane są informacje o określonych konceptach wizualnych, a następnie mierzy się, w jaki sposób to wpływa na dokładność klasyfikacji.

2.2.1 Metoda usuwania informacji o słowach wizualnych z reprezentacji

W celu zbadania, czy zawarta w reprezentacji informacja o obecności słów wizualnych rzeczywiście ma wpływ na dokładność klasyfikacji, usuwamy z reprezentacji informacje o poszczególnych słowach wizualnych. Usunięcie słowa wizualnego z reprezentacji definiujemy jako modyfikację reprezentacji, po której klasyfikator diagnostyczny *Word Content* nie będzie mógł wykryć obecności tego słowa. Dokonujemy tego za pomocą iteracyjnej procedury rzutowania na przestrzeń stosując algorytm INLP (ang. Iterative Null-space Projection) [39]. Algorytm INLP liniowo przekształca reprezentację, w taki sposób aby zminimalizować wyjście klasyfikatora diagnostycznego *Word Content* wykrywającego obecność usuwanego słowa wizualnego na całym zbiorze danych. Tym samym po wielu powtórzeniach otrzymujemy zmodyfikowaną reprezentację, która nie zawiera informacji o usuwanym słowie wizualnym. Po usunięciu słowa wizualnego mierzymy spadek dokładności klasyfikacji w zadaniu docelowym.



Rysunek 6: Metoda amnezyjnych klasyfikatorów diagnostycznych usuwa określone słowo wizualne (futro) z samonadzorowanej reprezentacji obrazu (wilk). Po usunięciu słowa wizualnego klasyfikator diagnostyczny nie może wykryć obecności futra na podstawie reprezentacji, natomiast dokładność klasyfikacji zadania docelowego spada. Wielkość spadku oznacza istotność rozważanego konceptu w klasyfikacji. Rysunek pochodzi z pracy [A2].

2.2.2 Wyniki i wnioski

Opisana w pracach [A1, A3] metoda klasyfikatorów diagnostycznych pozwala na lepsze zrozumienie reprezentacji w samonadzorowanych metodach uczenia maszynowego. Jednak sama informacja o obecności w reprezentacjach konceptów zrozumiałych dla człowieka nie mówi o tym, czy te koncepty są przydatne, np. do zadania klasyfikacji obrazów. Dlatego metoda amnezyjnych klasyfikatorów diagnostycznych lepiej pozwala zrozumieć, które słowa wizualne mają wpływ na działanie modelu. Dodatkowo dzięki zachowaniu taksonomii semantycznej słów wizualnych z wykorzystaniem cech wizualnych z obliczeniowej teorii widzenia Marra, możemy zbadać i porównać preferencje oraz uprzedzenia poszczególnych metod.

Wyniki przedstawione w pracy [A2] potwierdzają, że usunięcie słów wizualnych z samonadzorowanych reprezentacji zmniejsza dokładność klasyfikacji. Jest to wynik zgodny z założeniem przedstawionym w pracy [A1], o tym, że samonadzorowane reprezentacje zawierają informacje o semantyce obrazu. Co więcej, okazuje się, że w zależności od metody i rodzaju słów wizualnych, spadek dokładność klasyfikacji różni się. W przypadku metody SimCLR v2 słowa wizualne związane z kształtem i formą mają największy wpływ na decyzje klasyfikatora, podczas gdy dla metody BYOL największy wpływ ma jasność, a dla metody SwAV kolor.

2.3 Anonimizacja obrazów [A4]

W pracy [A4] zbadałem, czy reprezentacje uczenia maszynowego zawierają informacje o tożsamości osoby widocznej na obrazie oraz czy da się precyzyjnie usunąć takie informacje. W tym celu zaproponowaliśmy metodę SGAP (ang. Siamese Generative Adversarial Privatizer), która wykorzystuje właściwości syjamskiej sieci neuronowej do znalezienia cech identyfikujących osobę. W połączeniu z podejściem generatywnym ta metoda jest w stanie poprawnie zlokalizować, a następnie ukryć informacje identyfikujące na obrazie, przy minimalnym zmniejszeniu użyteczności zanonimizowanego zbioru danych do innych zadań, np. klasyfikacji.

Dotychczas zapewnienie prywatności w zbiorach danych odbywało się poprzez usunięcie wszystkich danych osobowych (np. nazwisk lub dat urodzenia). Ten sposób jednak nie jest niezawodny, gdyż pokazano skuteczne ataki korzystające z korelacji danych z danymi z innych źródeł [24, 38]. W pracy [A4] zaproponowaliśmy nowe podejście, które umożliwia publikowanie zbiorów danych (patrz Rysunek 7). Nasza metoda traktuje anonimizację zbioru danych jako grę pomiędzy dwoma stronami: jedna strona próbuje ukryć informacje o tożsamości, a druga próbuje te dane odgadnąć. Wykorzystanie syjamskiej sieci neuronowej pozwala na identyfikację tych części obrazu, które są najważniejsze w rozpoznaniu tożsamości, po to by te cechy skutecznie zaburzyć i tym samym wymusić anonimizację. Dodatkowo w pracy [A4] zdefiniowaliśmy metryki, które pozwalają ocenić kompromis pomiędzy anonimizacją danych a użytecznością zanonimizowanego zbioru danych.

2.3.1 Metoda anonimizacji obrazów

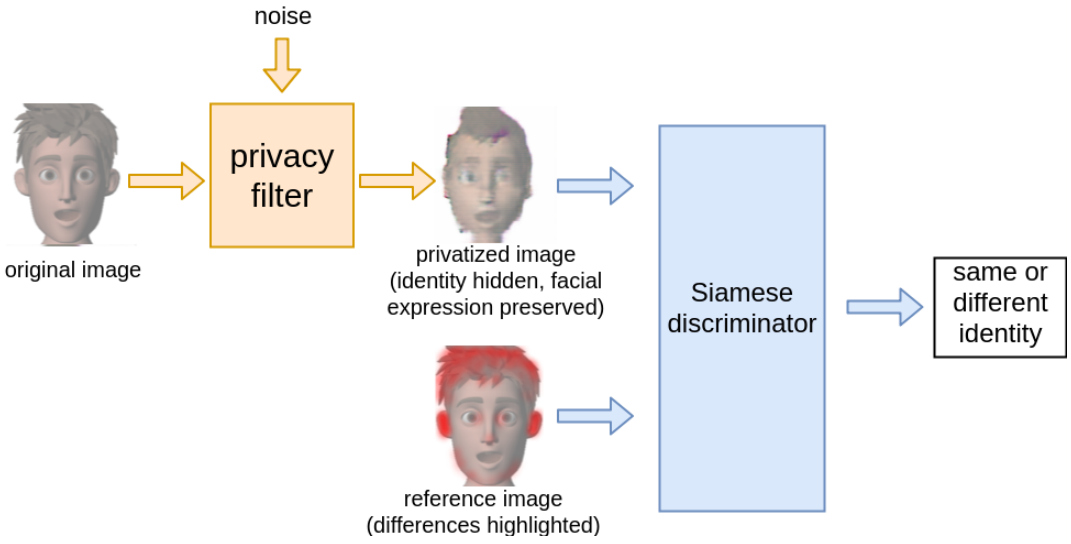
Założenia metody opisanej w pracy [A4] są takie, że: z jednej strony należy zachować prywatność osoby widocznej na przetwarzanym obrazie poprzez upewnienie się, że cechy identyfikujące osobę nie są w żaden sposób dostępne, a z drugiej strony należy zachować użyteczność

zbioru danych po anonimizacji, tak aby można było na nim wykonać inne zadania uczenia maszynowego, np. klasyfikację. Jednocześnie w tej metodzie możemy płynnie sterować poziomami prywatności i użyteczności zbioru danych. W celu zapewnienia powyższych założeń użyto niestandardowej architektury sieci neuronowej, która składa się z dwóch konkurujących ze sobą sieci neuronowych: generatora oraz dyskryminatora. Dyskryminator próbuje przewidzieć tożsamość osoby na obrazie, natomiast generator próbuje wygenerować taki obraz, który oszuka dyskryminator i tym samym ukryje tożsamość osoby. Jako sieci dyskryminującej użyto architektury syjamskiej [8], która jest trenowana na podstawie par obrazów, gdzie celem dyskryminatora syjamskiego jest sklasyfikowanie, czy na obrazach jest ta sama osoba.

Dla opisanej powyżej architektury należy dodać dodatkowy warunek na ograniczenie zniekształceń, co pozwoli zapewnić, że zanonimizowane obrazy nie będą zbyt mocno różnić się od oryginalnych obrazów, a jedyne wprowadzone zmiany będą dotyczyć ukrycia tożsamości, dzięki czemu użyteczność zbioru danych będzie zachowana. W tym celu posłużyłem się metryką SSIM (indeks podobieństwa strukturalnego, ang. structural similarity index measure) [47], która uwzględnia informacje o widocznej na obrazie scenie oraz strukturze obiektów.

2.3.2 Wyniki i wnioski

W celu weryfikacji, czy metoda SGAP skutecznie anonimizuje obrazy, skorzystano z oszacowania informacji wzajemnej pomiędzy obrazami a etykietami związanymi z tożsamością. Jeżeli informacja wzajemna jest wystarczająco mała, to nie da się wiarygodnie dowiedzieć się o tożsamości osoby na obrazie. W celu zmierzenia użyteczność zbioru danych po anonimizacji używamy zadania klasyfikacji wyrazu twarzy. Przeprowadziliśmy eksperymenty dla różnych wartości ograniczenia na maksymalne zniekształcenie obrazu według metryki SSIM, w celu



Rysunek 7: Zarys metody SGAP (ang. Siamese Generative Adversarial Privatizer). Filtr prywatności generuje zanonimizowany obraz. Tożsamość osoby jest ukryta, jednocześnie inne cechy obrazu, m.in. wyraz twarzy, są zachowane. Dyskryminator syjamski identyfikuje cechy dyskryminacyjne obrazów, które są podświetlone na czerwono. Rysunek pochodzi z pracy [A4].

dostosowania poziomu prywatności. Uzyskane w pracy [A4] wyniki pokazują, że da się dobrać odpowiedni poziom zniekształcenia przy którym zachowana jest zarówno użyteczność zbioru danych, jak i anonimowość osób.

2.4 Zrozumienie różnic pomiędzy radiologami i modelami uczenia głębokiego w diagnozowaniu raka piersi [A5]

W pracy [A5] zbadaliśmy, jakie są różnice pomiędzy radiologami i modelami uczenia głębokiego w diagnozowaniu raka piersi. Ponieważ głębokie sieci neuronowe są stosowane w diagnostyce obrazów medycznych, istotnym jest zrozumienie przesłanek stojących za decyzjami sztucznych sieci neuronowych. Dlatego zaproponowaliśmy metodę do porównania decyzji radiologów i modeli uczenia głębokiego, z uwzględnieniem analizy różnych podgrup pacjentów.

W przeciwieństwie do błędów diagnostycznych sztucznych sieci neuronowych, błędy lekarzy są często lepiej zrozumiałe. Radiolodzy są częściej narażeni na błąd z powodu trudności zadania, ale bardzo rzadko są narażeni na błąd z powodu nadmiernego skupienia się na mało istotnym i mało widocznym elemencie obrazu. Wynika to z tego, że ludzie wykorzystują wiedzę medyczną i związki przyczynowo-skutkowe. Zmiana na obrazie nie musi być tylko skorelowana z obecnością choroby, ale zazwyczaj istnieje wyraźny fizjologiczny powód, który lekarz jest w stanie zrozumieć i wytłumaczyć. Dlatego, aby zbudować zaufanie do diagnostyki z wykorzystaniem uczenia maszynowego, ważne jest, aby wiedzieć, czy sztuczna sieć neuronowa używa tego samego zestawu przesłanek co lekarz.

Metoda porównująca decyzje lekarzy i komputerów jest inspirowana badaniami związanymi z odpornością modeli wizyjnych na zakłócenia [13, 19, 26, 28, 48, 44]. Usuwając pewne informacje z danych wejściowych i analizując wynikającą z tego zmianę predykcji, możemy wnioskować, w jakim stopniu informacje te zostały wykorzystane. Aby porównać decyzje radiologów i sieci neuronowych pod względem odporności na zakłócenia zastosowaliśmy różne filtry dolnoprzepustowe na zbiorze danych mammograficznych. Następnie przeprowadziliśmy badania z udziałem radiologów, którzy mieli oceniać te same mammogramy, które oceniały modele uczenia głębokiego. Oceniliśmy wpływ filtrowania dolnoprzepustowego na pewność decyzji, a także porównaliśmy, czy obszary wykorzystywane przez algorytm są podobne do tych, które radiolodzy uznali za najistotniejsze.

2.4.1 Wyniki i wnioski

Wynik badań z pracy [A5] pokazują, że decyzje ludzi i maszyn różnią się w zależności od rodzaju choroby widocznej na obrazie. W przypadku wystąpienia na obrazie mammograficznym mikrozwapnień, zarówno radiolodzy jak i sieci neuronowe okazali się być wrażliwi na filtrowanie dolnoprzepustowe. Natomiast w przypadku uszkodzeń tkanek miękkich odkryliśmy, że filtrowanie dolnoprzepustowe nie wpływa na decyzje radiologów, podczas gdy sieci neuronowe są wrażliwe na te zmiany. Stąd wyciągamy wniosek, że w tym przypadku sztuczne sieci neuronowe korzystają z fałszywych przesłanek podczas diagnozowania choroby i nadmiernie skupiają się na wysokoczęstotliwościowych artefaktach, które nie są klinicznie istotne w chorobie nowotworowej.

2.5 Literatura

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012.

[2] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *International Conference on Neural Information Processing Systems*, pages 9525–9536, 2018.

[3] G. Alain and Y. Bengio. Understanding intermediate layers using linear classifier probes. In *ICLR Workshop*, 2017.

[4] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual Question Answering. In *ICCV*, 2015.

[5] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv:1909.03012*, 2019.

[6] Wanda Benesova and Michael Andrew Kottman. Fast superpixel segmentation using morphological processing. 2014.

[7] Alexander Binder, Grégoire Montavon, Sebastian Bach, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers, 2016.

[8] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. In *Advances in Neural Information Processing Systems 6*, pages 737–744. Morgan-Kaufmann, 1994.

[9] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, pages 213–229, 2020.

[10] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. 33:9912–9924, 2020.

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, pages 1597–1607, 2020.

[12] Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Annual Meeting of the Association for Computational Linguistics*, pages 2126–2136, 2018.

[13] Samuel Dodge and Lina Karam. A study and comparison of human and deep learning recognition performance under visual distortions. In *ICCCN*, pages 1–7. IEEE, 2017.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2020.

[15] Max Eichler, Gözde Gül Şahin, and Iryna Gurevych. Linspector web: A multilingual probing suite for word representations. In *Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 127–132, 2019.

[16] Cheng en Guo, Song-Chun Zhu, and Ying Nian Wu. Primal sketch: Integrating structure and texture. *Computer Vision and Image Understanding*, 106(1):5–19, 2007. Special issue on Generative Model Based Vision.

[17] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2000.

[18] Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. On the surprising similarities between supervised and self-supervised models. In *NeurIPS Workshop SVRHM*, 2020.

[19] Robert Geirhos, Carlos R. Medina Temme, Jonas Rauber, Heiko H. Schütt, Matthias Bethge, and Felix A. Wichmann. Generalisation in humans and deep neural networks. In *NeurIPS*, pages 7549–7561, 2018.

[20] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems (NeurIPS)*, 32:9277–9286, 2019.

[21] Leilani H Gilpin, Cecilia Testart, Nathaniel Fruchter, and Julius Adebayo. Explaining explanations to society. *arXiv:1901.06560*, 2019.

[22] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 21271–21284, 2020.

[23] Riccardo Guidotti, Anna Monreale, Stan Matwin, and Dino Pedreschi. Black box explanation by learning image exemplars in the latent feature space. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pages 189–205. Springer, 2019.

[24] Arif Harmanci and Mark Gerstein. Quantification of private information leakage from phenotype-genotype data: linking attacks. *Nat Meth*, 13(3):251–256, 2016.

[25] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020.

[26] Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019.

[27] Zixuan Huang and Yin Li. Interpretable and accurate fine-grained recognition via region grouping. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8662–8672, 2020.

[28] Jason Jo and Yoshua Bengio. Measuring the tendency of CNNs to learn surface statistical regularities. *arXiv*, 1711.11561, 2017.

[29] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

[30] Patricia Kitcher. Marr's computational theory of vision. *Philosophy of Science*, 55(1):1–24, 1988.

[31] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 25:1097–1105, 2012.

[32] Sawan Kumar and Partha Talukdar. NILE : Natural language inference with faithful natural language explanations. In *ACL*, 2020.

[33] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017.

[34] David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Co., Inc., New York, NY, USA, 1982.

[35] David Marr. Vision: A computational investigation into the human representation and processing of visual information, henry holt and co. *New York*, 1982.

[36] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[37] Michael J. Morgan. Features and the 'primal sketch'. *Vision Research*, 51(7):738–753, 2011. Vision Research 50th Anniversary Issue: Part 1.

[38] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pages 111–125. IEEE, 2008.

[39] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online, July 2020. Association for Computational Linguistics.

[40] Sylvestre-Alvise Rebuffi, Ruth Fong, Xu Ji, and Andrea Vedaldi. There and back again: Revisiting backpropagation saliency methods. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8839–8848, 2020.

[41] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 1135–1144, 2016.

[42] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR Workshop*, 2014.

[43] Josef Sivic and Andrew Zisserman. Video google: Efficient visual search of videos. In *Toward Category-Level Object Recognition*. Springer, 2006.

[44] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014.

[45] Stefano Teso, Öznur Alkan, Wolfang Stammer, and Elizabeth Daly. Leveraging explanations in interactive machine learning: An overview, 2022.

[46] Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. Probing pretrained language models for lexical semantics. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, 2020.

[47] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004.

[48] Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. In *NeurIPS*, pages 13255–13265, 2019.

[49] Quan-shi Zhang and Song-chun Zhu. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering*, 19(1):27–39, 2018.

3 Publikacje z cyklu

Explaining Self-Supervised Image Representations with Visual Probing

Dominika Basaj[1,2*], **Witold Oleszkiewicz**[1*], **Igor Sieradzki**[3], **Michał Górszczak**[3], **Barbara Rychalska**[1,4], **Tomasz Trzciński**[1,2,3] and **Bartosz Zieliński**[3,5]
[1]Warsaw University of Technology
[2]Tooploox
[3]Faculty of Mathematics and Computer Science, Jagiellonian University
[4]Synerise
[5]Ardigen
dominika.basaj@tooploox.com, witold.oleszkiewicz@pw.edu.pl, igor.sieradzki@uj.edu.pl, michal.gorszczak@student.uj.edu.pl, b.rychalska@mini.pw.edu.pl, tomasz.trzcinski@pw.edu.pl, bartosz.zielinski@uj.edu.pl

Abstract

Recently introduced self-supervised methods for image representation learning provide on par or superior results to their fully supervised competitors, yet the corresponding efforts to explain the self-supervised approaches lag behind. Motivated by this observation, we introduce a novel visual probing framework for explaining the self-supervised models by leveraging probing tasks employed previously in natural language processing. The probing tasks require knowledge about semantic relationships between image parts. Hence, we propose a systematic approach to obtain analogs of natural language in vision, such as visual words, context, and taxonomy. We show the effectiveness and applicability of those analogs in the context of explaining self-supervised representations. Our key findings emphasize that relations between language and vision can serve as an effective yet intuitive tool for discovering how machine learning models work, independently of data modality. Our work opens a plethora of research pathways towards more explainable and transparent AI.

1 Introduction

Visual representations are cornerstones of a multitude of contemporary computer vision and machine learning applications, ranging from visual search [Sivic and Zisserman, 2006] to image classification [Krizhevsky *et al.*, 2012] and visual question answering (VQA) [Antol *et al.*, 2015]. However, learning representations from data typically requires tedious annotation. Therefore, recently introduced self-supervised representation learning methods concentrate on decreasing the need for data labeling without reducing their performance [Chen *et al.*, 2020b; Grill *et al.*, 2020; Caron *et al.*, 2020]. Because of the fundamental role representations play in real-life applications, a lot of research focuses on explaining these embeddings [Vulić *et al.*, 2020; Eichler *et al.*, 2019; Huang and Li, 2020]. Nevertheless, most of them concentrate on fully supervised embeddings [Zhang and Zhu, 2018] and not on their self-supervised counterparts. Moreover, the majority of the proposed approaches rely on pixel-wise image analysis [Simonyan *et al.*, 2014; Adebayo *et al.*, 2018], while general semantic concepts present in the images are often ignored.

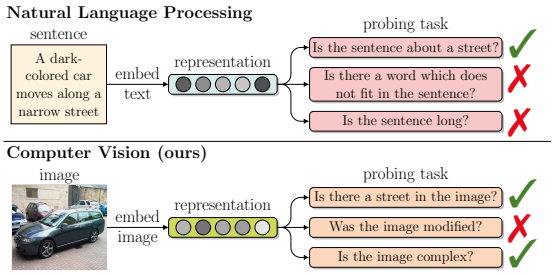


Figure 1: *Probing tasks*, widely used in natural language processing, validate if a *representation* implicitly encodes a given property, *e.g.*, sentence topic or length. We introduce a visual taxonomy along with the corresponding probing framework that allow to build analogous *visual probing tasks* and explain the self-supervised image representations. As a result, we *e.g.* discover that even though all analysed models build similar semantic knowledge, some of them focus more on texture and therefore achieve better accuracy on target tasks.

Here, we attempt to overcome these shortcomings and draw inspiration from a simple yet often overlooked observation that humans use language as a natural tool to explain what they learn about the world through their eyes [Kumar and Talukdar, 2020]. Therefore, considering that the very same machine learning algorithms can be successfully applied to solve both vision and natural language processing (NLP) tasks [Dosovitskiy *et al.*, 2020; Carion *et al.*, 2020], we postulate that the methods used to analyze text representation can also be employed to investigate visual inputs.

Very popular tools for explaining textual embeddings are *probing tasks* [Conneau *et al.*, 2018]. As shown in the up-

*Equal contribution
The code is at: github.com/BioNN-InfoTech/visual-probes

per part of Fig. 1, a probing task in NLP is a simple classifier that asks if a given textual representation encodes a particular property, such as a sentence length or its semantic consistency, even though this property was not a direct training objective. For instance, we can create a textual probing task by substituting a word in a sentence and checking if a simple classifier that takes the representation of the original and altered sentence can detect this change. By analyzing the accuracy of a probing task, one can verify if the investigated representation contains certain information and understand the rationale behind embedding creation. However, while probing tasks are straightforward, intuitive, and widely used tools in NLP, their computer vision application is limited [Alain and Bengio, 2017], mainly due to the lack of appropriate analogs between textual and visual modalities.

In this paper, we address this limitation by introducing a mapping between vision and language that enables applying the NLP probing tools in the computer vision (CV) domain. For this purpose, in Sec. 3, we propose a taxonomy of visual units that includes *visual sentences*, *words*, and *characters*. We then employ these units as building blocks for a more general visual probing framework that contains a variety of NLP-inspired probing tasks, such as *word content*, *sentence length*, *character bin*, and *Semantic Odd Man Out* [Conneau *et al.*, 2018; Eichler *et al.*, 2019]. The results we obtain provide us with unprecedented insights into semantic knowledge, complexity, and consistency of self-supervised image representations, *e.g.* we discover that semantics of the image only partially contribute to target task accuracy. Our framework also allows us to compare the existing self-supervised representations from a novel perspective, as we show in Sec. 5.

Our contributions can be therefore summarized as follows:

- We propose a mapping between visual and textual modalities that constructs a visual taxonomy.
- We introduce novel visual probing tasks for comparing self-supervised image representations inspired by similar methods used in NLP.
- We show that leveraging the relationship between language and vision serves as an effective yet intuitive tool for discovering how self-supervised models work.

2 Related Works

The visual probing framework aims to explain image representations obtained from self-supervised methods. It is inspired by probing tasks used in NLP. Therefore, we consider related works from three research areas: self-supervised computer vision models, probing tasks in natural language processing, and explainability methods in computer vision.

Self-supervised computer vision models. Recently published self-supervised methods provide state-of-the-art results across computer vision tasks. They usually base on contrastive loss [Hadsell *et al.*, 2006] that measures the similarities of patches in representation space and aims to discriminate between positive and negative pairs. The positive pair contains modified versions of the same image, while the negative pairs correspond to two images in the same dataset. One of the methods, MoCo v1 [He *et al.*, 2019] trains a slowly progressing visual representation encoder, driven by a momentum update. This encoder plays a role of a memory bank of past representations and delivers negative examples. SimCLR v2 [Chen *et al.*, 2020b], unlike MoCo v1, proposes a different way of generating negative pairs. Instead of a memory bank, they propose to use a large batch size of up to 4096 examples. Other improvements proposed by SimCLR v2 are a projection head and carefully tuned data augmentation. The projection head maps representations to space where contrastive loss is applied, which is important due to the loss of information. BYOL [Grill *et al.*, 2020] also uses the projection head, but unlike MoCo v1 and SimCLR v2, it achieves a state-of-the-art performance without the explicitly defined contrastive loss function, so it does not need negative examples. On the other hand, SwAV [Caron *et al.*, 2020] takes advantage of contrastive methods without pairwise comparisons. Instead, it learns the representations by clustering them and predicting the labels of their clusters. Our paper provides a framework to analyze the representation generated by those methods in terms of the semantic knowledge they encode.

Probing tasks in NLP. One of the classic examples of the NLP probing task aims to probe sentence embeddings for interesting linguistic features such as the depth of the parse tree or whether the sentence contains a specific word [Conneau *et al.*, 2018]. Others propose to focus on lexical knowledge concerning the qualities of individual words more than the whole sentences [Vulić *et al.*, 2020; Eichler *et al.*, 2019]. We consider both these objectives in our approach, i.e. we study probing tasks on specific concepts and their compositions. Moreover, while most works on probing tasks focus on one selected language, the others [Eichler *et al.*, 2019] are designed with multilingual settings in mind. This paper reflects the latter because it can be applied to various image domains.

Explainability methods in CV representation learning. The existing methods for explaining image representations either verify the relevance of hidden layers of supervised classification networks [Alain and Bengio, 2017] or highlight individual pixels that are essential for the model [Simonyan *et al.*, 2014; Adebayo *et al.*, 2018]. Moreover, they usually generate the important regions as pixel clouds, which are not understood as concrete semantic concepts. In contrast, approaches such as [Huang and Li, 2020; Ghorbani *et al.*, 2019] aim to detect important image segments but are often difficult to understand in practice, even though they are crucial for the model objective. In this work, we extend the existing methods by analyzing the semantic information stored in the self-supervised representation.

3 Visual Probing

This section introduces a novel visual probing framework that analyzes the information stored in self-supervised image representations. For this purpose, in Sec. 3.1, we propose a mapping between visual and textual modalities that constructs a visual taxonomy. As a result, the image becomes a "visual sentence" and can be analyzed with visual probing tasks inspired by similar methods used in NLP (see Sec. 3.2).

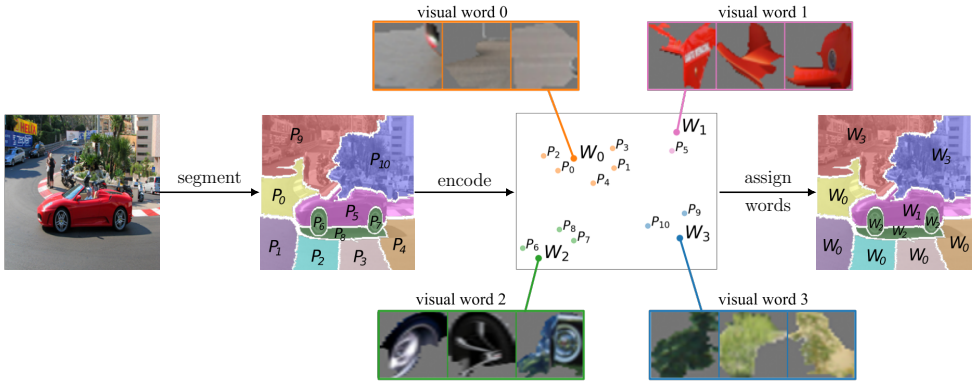


Figure 2: The process of dividing an image into visual words. First, an image is segmented into multiple superpixels: P_0, P_1, ..., P_{10}. Then, each superpixel is embedded in the latent space previously used to generate the dictionary of visual words: W_0, W_1, W_2, W_3. Finally, each superpixel is assigned to the closest word in the visual word dictionary. This results in mapping between vision and language and enables using the visual probing framework that includes a variety of NLP-inspired probing tasks.

3.1 Mapping Between Vision and NLP

While an image can be considered a sentence equivalent in a probing task, the question remains, what is the equivalent to words and characters? There are multiple possible answers to this question. One of the intuitive ones is to divide an image into non-overlapping superpixels that group pixels into perceptually meaningful atomic regions [Achanta *et al.*, 2012]. As a result, we obtain an image built from superpixels as an analogy of a sentence built from the words. The superpixels, similarly like words, have order and meaning. Moreover, each superpixel contains a specific number of pixels, like the number of word's characters. As a consequence, we obtain an intuitive mapping between visual and textual domains.

However, superpixels treated as visual words would significantly differ from their linguistic counterparts because they do not repeat between images, while in text, the words often repeat between sentences. Therefore, we propose to define visual words as the clusters of superpixels in representation space and assign each superpixel to the closest centroid. For this purpose, we could use the original definition of visual words from [Leung and Malik, 2001]. However, it does not take into account the importance of those words for a model's prediction. Therefore, instead of that, we use TCAV methodology [Kim *et al.*, 2018; Ghorbani *et al.*, 2019] that generates high-level concepts, which are important for prediction and easily understandable by humans. Such an approach requires a supervisory training network but generates visual words independent of any compared self-supervised techniques, which is crucial for a fair comparison. Therefore, the process of dividing an image into visual words consists of three steps: segmentation into superpixels, their encoding, and assignment to visual words (see Fig. 2).

3.2 Visual Probing Tasks

After dividing an image into visual words, it can be analyzed by the visual probing framework, which can adapt almost any NLP probing task. Here, we describe the four that are well known by the NLP community [Conneau *et al.*, 2018; Eichler *et al.*, 2019]. Moreover, except for defining visual probing tasks, we provide their original NLP definitions to make the paper self-contained.

Word Content (WC). The word content probing task aims to identify which visual words are present in the image. The *input* of this probing task is a self-supervised representation of the image. The *target labels* represent the presence of a particular visual word. As we describe in Sec. 4, we select 100 representative visual words. Hence, there are 100 binary *target labels*. Fig. 2 illustrates the process of determining which visual words are present in the image. The NLP inspiration of the task probes for surface information, the type of information that does not require any linguistic knowledge [Conneau *et al.*, 2018]. In contrast, its adaptation requires ***semantic knowledge*** to understand which concept is represented by a superpixel.

Sentence Length (SL). The aim of the sentence length probing task is to distinguish between simple and complex images, as presented in Fig. 3. The *input* of this probing task is a self-supervised representation of the image. The *target label* is the number of unique visual words in the image, which can be determined based on the WC labels. The original NLP probing task predicts the number of words (or tokens) and retains only surface information [Conneau *et al.*, 2018]. At CV, it serves as a proxy for ***semantic complexity***, requiring the semantic understanding of the image.

Character Bin (CB). The aim of the character bin probing task is to check whether the representation stores information about the complexity of the image. The *input* of this probing task is a self-supervised representation of the image's superpixel. The *target label* is the size of the superpixel defined as the number of non-grey pixels, as presented in Fig. 4. The original NLP probing task is defined as a classifier of the number of characters in a single word [Eichler *et al.*, 2019]. From this perspective, the character bin retains only surface

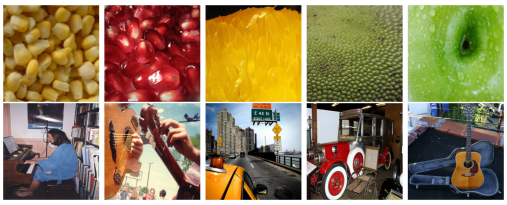


Figure 3: The SL probing task measures how well the representation encodes the information about the number of unique visual words in the image. *Top row:* a low number of unique visual words (<13). *Bottom row*: a high number of unique visual words (>42).

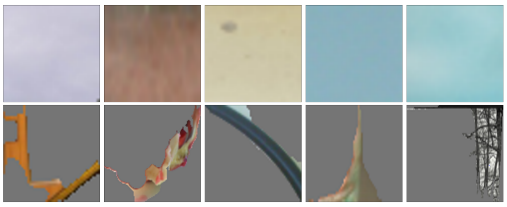


Figure 4: The CB probing task measures if the representation retains information about the superpixel's size. *Top row:* examples of large superpixels. *Bottom row:* example of small superpixels.

information in both domains.

Semantic Odd Man Out (SOMO). The objective of the SOMO probing task is to predict whether the image was modified by replacing a random superpixel in the image with a similarly shaped superpixel from another image, and corresponding to a different visual word, as presented in Fig. 5. The *input* of this probing task is a self-supervised representation of the image. The *target label* is binary, i.e. the image was modified or not. The original NLP task predicts if the sentence was altered by replacing a random noun or verb [Conneau *et al.*, 2018]. In both domains, it requires the ability to detect alterations in ***semantic consistency***.

4 Experimental Setup

In this section, we describe the procedure of generating visual words and training probing tasks.

Generating visual words. We use the original settings of the ACE algorithm described in [Ghorbani *et al.*, 2019] that first divides images into superpixels using SLIC algorithm [Achanta *et al.*, 2012] with three resolutions of 15, 50, and 80 segments for each image. It then computes representations of these superpixels as an output of a *mixed4c* layer of GoogLeNet [Szegedy *et al.*, 2015] trained on the ImageNet dataset. Finally, representations are clustered using the k-means algorithm, resulting in clusters that correspond to the visual words (see Fig. 6). As there are over a dozen visual words generated for each of the classes, the dictionary's size grows significantly with the size of the analyzed dataset. Therefore, in this paper, we decided to analyze its subset containing 55 classes grouped into 5 categories: animals, vehicles, musical instruments, buildings, fruits. Moreover, to further limit the dictionary size, we only keep 100 of the most relevant visual words (according to TCAV score [Kim *et al.*, 2018]), while ensuring that each class is represented by at least one of them. These 100 visual words form our visual language.

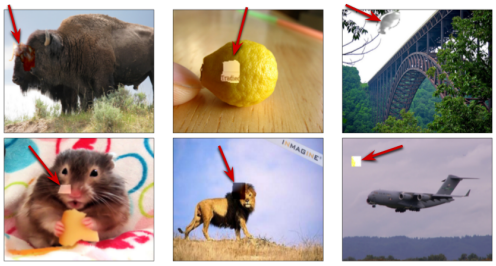


Figure 5: The SOMO probing task predicts if the image was altered by replacing a random superpixel. *Top row:* examples of images for which SimCLR v2 correctly recognizes the modification, while SwAV fails. *Bottom row:* examples of images where both SimCLR v2 and SwAV do not recognize superpixel modification.

Generating a self-supervised representation. We examine four self-supervised methods: MoCo v1 [He *et al.*, 2019], SimCLR v2 [Chen *et al.*, 2020b], BYOL [Grill *et al.*, 2020], and SwAV [Caron *et al.*, 2020]. For all of them, we use publicly available models trained with ImageNet. Although they all use the penultimate layer of ResNet-50 to generate representations, their training hyperparameters differ, which is described in Supplementary Materials (SM in the following)[1].

Assigning visual words. To assign a superpixel to a visual word, we pass it through the GoogLeNet to generate a representation from the *mixed4c* layer (similarly to generating visual words). We can then determine the visual word closest to a superpixel, as both are embedded in the same space.

Training probing tasks. We use a logistic regression classifier with a maximum of 1000 iterations and the LBFGS solver to train all diagnostic classifiers. As an input, we use representations generated by the self-supervised methods. The output depends on the probing task. In the case of the WC, we train 100 classifiers corresponding to 100 visual words. We expect an image to be assigned to a particular visual word if at least one of its superpixels is assigned to it. Finally, we report the average AUC scores over 100 classifiers (see Tab. 1). To obtain classification setup in the sentence length probing task, we group the possible output into 5 equally-wide bins, resulting in one-vs-rest OVR AUC, which is resistant to class imbalance. A similar procedure is applied to the character bin probing task, except that we use 6 bins in this case. SOMO is formulated as a binary classification task, in which we predict whether the image was modified or not. The training and validation datasets are balanced. We conduct all of our experiments on the ImageNet dataset [Deng *et al.*, 2009], keeping its standard train/validation split. More-

[1]Supplementary materials: http://www.ii.uj.edu.pl/~zielinsb/papers/visual_probing_ijcai_supplement.pdf

	Target	Probing tasks (ours) WC	SL	CB	SOMO
MoCo v1	0.606	0.790	0.868	0.937	0.559
SimCLR v2	0.717	**0.800**	**0.877**	**0.964**	**0.625**
BYOL	0.723	0.795	0.876	0.961	0.615
SwAV	**0.753**	0.761	0.838	0.956	0.530

Table 1: AUC score for our probing tasks and accuracy on the linear evaluation (Target). Like the linear evaluation, our probing tasks are also trained on top of the frozen base network, and test accuracy is used as a proxy for representation quality. Hence, they provide complementary knowledge about the representation.

over, we apply the random over-sampling if needed to deal with the imbalanced classes. The details on the experimental setup are presented in SM.

5 Results and Discussion

Tab. 1 summarizes the results obtained in our experiments. It presents the performance of our probing tasks and the target task accuracy for reference. The reported target task performance is the classification accuracy calculated for the whole ImageNet validation set. The first conclusion is that self-supervised representations retain information about semantic knowledge and semantic complexity, but they do not code much information about image consistency. Secondly, the performance on probing tasks do not correlate with accuracy on the target task. Finally, SimCLR v2 overpasses other methods in all probing tasks. In the following, we analyze those aspects in greater detail.

Self-supervised representations contain strong semantic knowledge. As outlined in 3.2, we treat the results of the word content probing as an approximation of semantic knowledge present in a representation. The AUC scores for this probing task reported in Tab. 1 vary from 0.76 for SwAV to 0.8 for SimCLR v2. This shows ability to predict which visual words are present in the image. Based on this we can say that semantic knowledge is encoded in the examined self-supervised representations.

The level of semantic knowledge does not correlate with target task accuracy. It is surprising that although examined self-supervised methods have diverse target task accuracy, they all have a similar level of semantic knowledge. E.g. MoCo v1 obtains the worst target task accuracy (61%), but the results of the WC probing task is on par with stronger self-supervised methods. Even more surprising is that SwAV, despite its highest accuracy on the target task, is below the scores of other tested methods in terms of semantic knowledge measured by the WC probing task. This finding supports the view that semantic knowledge only partially contributes to the target task accuracy [Geirhos *et al.*, 2020].

Certain types of semantic knowledge are represented better than others. The probing task's ability to predict which visual words are present in the image varies, as some words are better predicted than others. We conducted a user study to understand the difference between best and worst predicted visual words presented in Fig. 6. According to the results, the five best recognizable visual words are perceived to have distinct, non-uniform textures. In contrast, the five worst recognizable visual words have more uniform textures. This may indicate that self-supervised representations are pattern-biased. This sheds new light on this problem, as previous results [Geirhos *et al.*, 2020] suggest the opposite. See SM for details on the user study.

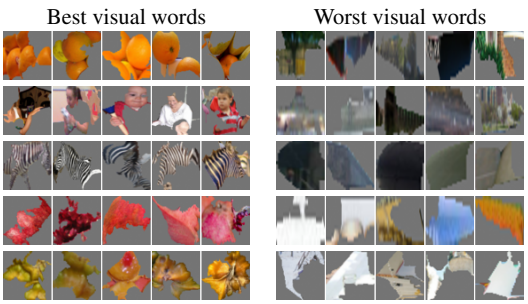
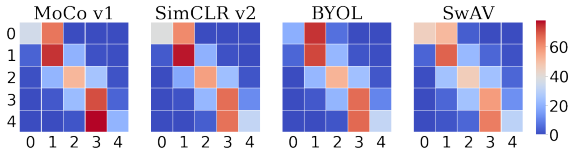


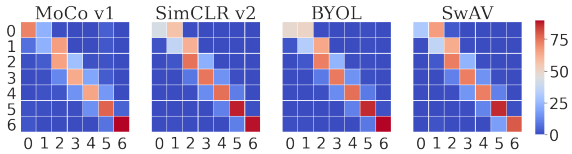
Figure 6: Visualization of the best and the worst predicted visual words, according to the results of the WC probing task (on average by all self-supervised methods). Our user study shows that the best recognizable visual words are perceived to have distinct non-uniform textures contrary to the worst recognizable ones which have more uniform textures. This may indicate that self-supervised representations better encode information about patterns.

There are visible differences in semantic knowledge retained by different self-supervised methods. Our user study shows variability by comparing the semantic content of representations on individual visual words. We take a closer look at the visual words that some self-supervised methods encode better or worse than others. The examples of these visual words are in SM. Looking at the top five visual words that MoCo v1 encodes better than the other representations, we can see that these words have distinct patterns. Moreover, the user study shows that MoCo v1 is better than the others at recognizing non-uniform textures. On the other hand, SimCLR v2, BYOL, and SwAV are above average in recognizing uniform textures.

Self-supervised representations contain information about semantic complexity. We design two probing tasks - sentence length and character bin - which validate the complexity of an image. Based on the results in Tab. 1, we observe that representations reflect the level of semantic complexity to a high degree. Information about the number of unique visual words (SL) is equally well predicted by probing classifiers for all self-supervised representations. These results are consistent with the results for semantic knowledge. For both probing tasks, SwAV's performance is slightly below the scores of other tested methods. This demonstrates the link between semantic complexity and semantic knowledge. AUCs are even higher for predicting the size of a visual word, which indicates that representation encodes the approximation of its shape (although technically, we predict the number of pixels).



(a) Sentence length (SL).



(b) Character bin (CB).

Figure 7: Confusion matrices for SL and CB (results in %). The results indicate that the ability of self-supervised representations to retain information about complexity differs depending on the level of image complexity. Moreover, even though the final AUCs are similar, their confusion matrices vary.

The ability of self-supervised representations to retain information about semantic complexity differs. There are no substantial differences in the ability to encode the complexity of the images between self-supervised methods. However, some preferences can be observed once we do not aggregate predictions into one AUC number. A closer look at the confusion matrices for the SL probing task in Fig. 7 shows that BYOL does worse when it comes to recognizing less complex images, but it performs well in comparison to other self-supervised methods. That is in contrast to SwAV, which overall has the lowest AUC metric, but it stands out when it comes to passing on information about simple images.

Self-supervised representations struggle to retain information about semantic consistency. Contrary to what we observe for semantic knowledge and complexity, self-supervised representations do not encode well the information about semantic consistency. The ability to distinguish altered images differs between methods, with the smallest AUC metric (53%) for representations extracted by SwAV and the highest for SimCLR v2 (63%). Manual inspection of examples from the top and bottom performers classified as true positive and false negative with high (>80%) certainty indicates differences in decision making of probing classifiers. Firstly, we observe that SimCLR v2 does relatively well with examples that people easily recognize as modified. However, it performs worse on more blended alterations (Fig. 5), which do not disturb the huge chunks of textures or colors. At the same time, in most cases, information encoded in SwAV's representation does not reflect well enough even such visible alterations. Fig. 5 shows correct predictions for SimCLR v2 which SwAV predicted as not changed. Analysis of visual words for which we replaced the original ones across true positive and false negative for both SimCLR v2 and SwAV does not indicate any substantial differences between them. Hence, we conclude that the performance of the SOMO does not depend on the visual word we use as a replacement, but rather to what extent the semantic sentence is altered.

Self-supervised representations are resistant to modifications. Even though the replacements of visual words do not disturb the substantial part of the image, this lack of ability to distinguish alterations is interesting in the light of [Hendrycks *et al.*, 2019], which claims that self-supervised methods improve out-of-distribution detection. We do not contradict this conclusion, but our results show that in particular setups, self-supervised representations do not exhibit enough ability to distinguish between corrupted and not corrupted images. Considering that various, even minor and not visible, alterations might lead to a change in the outcome of the prediction, we postulate that the tendency of self-supervised representations not to retain information about consistency might pose a risk. When it comes to the differences in the AUC for the examined representations, they might be partially explained by differences in the architecture. E.g. SimCLR v2 and BYOL are trained with projection head, whereas SwAV and MoCo v1 are not. The projection allows retaining information about the transformation of the input image [Chen *et al.*, 2020a]. Therefore, we hypothesize that this information may cause differences in the AUC score.

6 Conclusions

In this work, we introduce a novel visual probing framework that analyzes the information stored in self-supervised image representations. It is inspired by probing tasks employed in NLP and requires similar taxonomy. Hence, we propose a set of mappings between visual and textual modalities to construct visual sentences, words, and characters. The results of the experiments confirm the effectiveness and applicability of this framework in understanding self-supervised representations. We verify that the representations contain information about semantic knowledge and complexity of the images, although they struggle to retain information about image consistency. Moreover, a detailed analysis of each probing task reveals differences in the representations encoded by various methods. This provides knowledge about representation complementary to the accuracy of linear evaluation.

Finally, we show that the relations between language and vision can serve as an effective yet intuitive tool for explainable AI. Hence, we believe that our work will open new research directions in this domain.

Acknowledgments

This research was supported by: grant no POIR.04.04.00-00-14DE/18-00 carried out within the Team-Net program of the Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund; grant no 2018/31/N/ST6/02273 funded by National Science Centre, Poland; Priority Research Area Digiworld under the program Excellence Initiative – Research University at the Jagiellonian University in Kraków; and POB Research Centre for Artificial Intelligence and Robotics of Warsaw University of Technology within the Excellence Initiative Program - Research University (ID-UB).

References

[Achanta *et al.*, 2012] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11), 2012.

[Adebayo *et al.*, 2018] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. Sanity checks for saliency maps. In *NeurIPS*, 2018.

[Alain and Bengio, 2017] G. Alain and Y. Bengio. Understanding intermediate layers using linear classifier probes. In *ICLR Workshop*, 2017.

[Antol *et al.*, 2015] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual Question Answering. In *ICCV*, 2015.

[Carion *et al.*, 2020] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.

[Caron *et al.*, 2020] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint 2006.09882*, 2020.

[Chen *et al.*, 2020a] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

[Chen *et al.*, 2020b] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. Hinton. Big self-supervised models are strong semi-supervised learners. In *NeurIPS*, 2020.

[Conneau *et al.*, 2018] A. Conneau, G. Kruszewski, G. Lample, L. Barrault, and M. Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *ACL*, 2018.

[Deng *et al.*, 2009] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

[Dosovitskiy *et al.*, 2020] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint 2010.11929*, 2020.

[Eichler *et al.*, 2019] M. Eichler, G. G. Şahin, and Ir Gurevych. LINSPECTOR WEB: A multilingual probing suite for word representations. In *EMNLP-IJCNLP: System Demonstrations*, 2019.

[Geirhos *et al.*, 2020] R. Geirhos, K. Narayanappa, B. Mitzkus, M. Bethge, F. A. Wichmann, and W. Brendel. On the surprising similarities between supervised and self-supervised models. In *NeurIPS Workshop*, 2020.

[Ghorbani *et al.*, 2019] A. Ghorbani, James Wexler, J. Zou, and Been Kim. Towards automatic concept-based explanations. In *NeurIPS*, 2019.

[Grill *et al.*, 2020] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint 2006.07733*, 2020.

[Hadsell *et al.*, 2006] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006.

[He *et al.*, 2019] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint 1911.05722*, 2019.

[Hendrycks *et al.*, 2019] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song. Using self-supervised learning can improve model robustness and uncertainty. In *NeurIPS*, 2019.

[Huang and Li, 2020] Zixuan Huang and Yin Li. Interpretable and accurate fine-grained recognition via region grouping. In *CVPR*, 2020.

[Kim *et al.*, 2018] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *ICML*, 2018.

[Krizhevsky *et al.*, 2012] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012.

[Kumar and Talukdar, 2020] Sawan Kumar and Partha Talukdar. NILE : Natural language inference with faithful natural language explanations. In *ACL*, 2020.

[Leung and Malik, 2001] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. In *International journal of computer vision*. Springer, 2001.

[Simonyan *et al.*, 2014] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR Workshop*, 2014.

[Sivic and Zisserman, 2006] Josef Sivic and Andrew Zisserman. Video google: Efficient visual search of videos. In *Toward Category-Level Object Recognition*. Springer, 2006.

[Szegedy *et al.*, 2015] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[Vulić *et al.*, 2020] I. Vulić, E. M. Ponti, R. Litschko, G. Glavaš, and A. Korhonen. Probing pretrained language models for lexical semantics. *arXiv preprint arXiv:2010.05731*, 2020.

[Zhang and Zhu, 2018] Quan-shi Zhang and Song-chun Zhu. Visual interpretability for deep learning: A survey. In *Frontiers of Information Technology & Electronic Engineering*, 2018.

Which Visual Features Impact the Performance of Target Task in Self-supervised Learning?

Witold Oleszkiewicz[1][0000−0002−7234−393X], Dominika Basaj[2][0000−0002−9377−3517], Tomasz Trzciński[1,2,3][0000−0002−1486−8906], and Bartosz Zieliński[3,4][0000−0002−3063−3621]

[1] Warsaw University of Technology, plac Politechniki 1, Warszawa, Poland
witold.oleszkiewicz@pw.edu.pl
[2] Tooploox, Teczowa 7, Wrocław, Poland
[3] Faculty of Mathematics and Computer Science, Jagiellonian University, Łojasiewicza 6, Kraków, Poland
[4] Ardigen, Podole 76, Kraków, Poland

Abstract. Self-supervised methods gain popularity by achieving results on par with supervised methods using fewer labels. However, their explaining techniques ignore the general semantic concepts present in the picture, limiting to local features at a pixel level. An exception is the visual probing framework that analyzes the vision concepts of an image using probing tasks. However, it does not explain if analyzed concepts are critical for target task performance. This work fills this gap by introducing amnesic visual probing that removes information about particular visual concepts from image representations and measures how it affects the target task accuracy. Moreover, it applies Marr's computational theory of vision to examine the biases in visual representations. As a result of experiments and user studies conducted for multiple self-supervised methods, we conclude, among others, that removing information about 3D forms from the representation decrease classification accuracy much more significantly than removing textures.

Keywords: Explainability · Self-supervision · Probing tasks

1 Introduction

Visual representations are critical in many computer vision and machine learning applications. The spectrum of these applications is broad, starting with visual search [21] to image classification [16] and visual question answering [3]. However, supervised representation learning requires a large amount of labeled data, usually time-consuming and expensive. Hence, self-supervised methods gain popularity, achieving results on par with supervised methods using fewer labels [6, 8, 13].

Along with the increasing proliferation of self-supervised methods for representation learning, there is a growing interest in developing methods that allow the interpretation of the resulting representation space and draw conclusions regarding the information it conveys. However, most of them focus on supervised

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_24

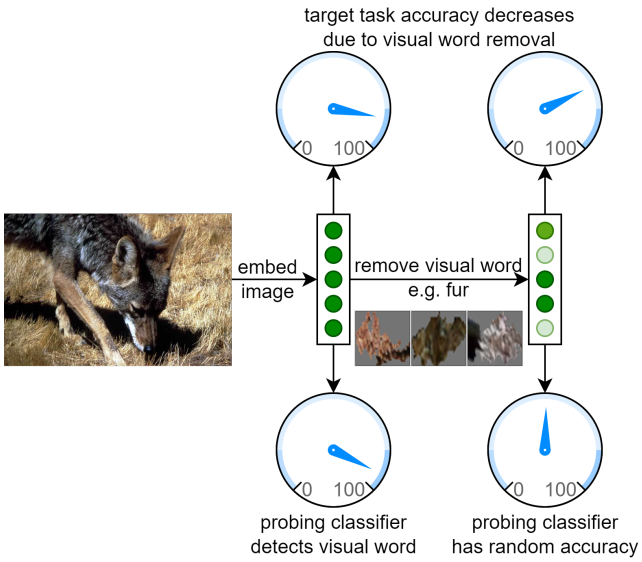


Fig. 1: Amnesic visual probing removes a specific visual concept (here corresponding to fur) from the self-supervised representation of an image (here corresponding to a wolf). As a result, the probing classifier cannot detect the presence of fur in the representation, and the target task accuracy decreases. The level of decrease represents the importance of the considered concept.

approaches and study local features at a pixel level [2, 20]. At the same time, the general semantic concepts present in the image are often overlooked, and their influence on model decisions is unknown. From this perspective, an exception is visual probing [4] that analyzes the vision concepts of an image using probing tasks. The probing tasks provide information about the presence of visual concepts in the representations but do not explain if they are critical for target task performance.

In this work, we overcome this limitation, providing a method that investigates the importance of visual features in the context of target task performance, referring to the amnesic probing [10] used in natural language processing (NLP). We remove information about particular visual concepts from image representations using the Iterative Nullspace Projection [19] and measure how it affects the target task accuracy. In addition, we conduct user studies to describe the visual concepts using Marr's computational theory of vision [17]. As a consequence, we can examine the biases in image representations.

Our contributions can be summarized as follows:

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_24

– We propose amnesic visual probing, a method for analyzing which visual features impact the performance of a target task.
– We apply Marr's computational theory of vision to examine the biases in visual representations.
– We conduct a complete user study and assign automatically generated visual concepts to one of six visual features from Marr's computational theory of vision.

2 Related Works

Our work corresponds to two research areas: self-supervised learning and probing tasks. We briefly cover the latest achievements in these two topics in the following paragraphs.

Self-supervised image representations Image representations obtained in a self-supervised manner are increasingly popular due to the competitive performance compared to supervised approaches. It is because they leverage the power of datasets without label annotations. One of the methods, called MoCo v1 [14], is based on a dictionary treated as a queue of data samples. It contains two encoders for query and keys, which are matched by contrastive loss. This queue enables to use of a large dictionary of examples previously limited to the batch size. SimCLR v2 [8] is another powerful method, which builds upon its predecessor, SimCLR [7] that maximizes the agreement between two views of the same sample by contrastive loss. In [8], the authors use a deeper and thinner backbone (ResNet-152 3x), deepen the projection head, which is not removed after contrastive training, and adapt memory mechanism from MoCo to increase the pool of negative examples. SwAV [6] takes advantage of contrastive methods. However, it compares clusters of data instead of single examples. The consistency between clusters, which can be seen as views of the same data sample, is enforced by learning to predict one view from another. In contrast to the above methods, BYOL [13] does not use the explicitly defined contrastive loss function, so it does not need negative samples. Instead, it uses two neural networks, referred to as online and target networks, that interact and learn the representation of the same image from each other.

Probing tasks The probing tasks originally come from Natural Language Processing (NLP). Their objective is to discover the characteristics interpretable by humans, which are encoded in the representation obtained by neural networks [5]. Probing is usually a simple classifier applied to trained representations like word embeddings. The probing classifier predicts whether the linguistic phenomenon that we want to verify exists or not. The probing classifiers in the NLP research community are popular tools for inspecting the internals of representations. However, some recent work extends the usability of probing tasks by introducing the concept of amnesic probing [10] to measure the influence of the phenomenons on the target task performance.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_24

Although probing tasks are popular in NLP, they only recently have been adapted to the Computer Vision (CV) domain in [4] based on the mapping defined between NLP and CV domains. These visual probing tasks allow one to gain intuition about the knowledge conveyed in the representation by the various self-supervised methods. However, there is no clear consensus on their impact on the target task performance.

3 Methods

This section introduces amnesic visual probing (AVP), a tool for explaining visual representations. It analyzes how important are particular visual concepts for a target task. Therefore, to define AVP, we first provide visual concepts (here called Visual Words, VW) and then obtain their meaning. Finally, we remove information about VW from the representation and analyze how it influences a target task.

Generating visual words To generate visual words, we use the established ACE algorithm [12]. It starts by dividing the image into superpixels using the SLIC algorithm [1]. Because different superpixel sizes are preferred, we run the algorithm three times with different parameters and obtain three sets with 15, 50, and 80 superpixels for each image. Then, we pass all the superpixels through the network trained on ImageNet to obtain their representations. These representations are clustered separately for each class using the k-means algorithm with $k = 25$ (infrequent and unpopular clusters are removed as described in [12]). Clusters obtained this way could be directly used as visual words. However, so many visual words would be impractical due to the similarity between concepts of ImageNet classes. Therefore, to obtain a credible dictionary with visual words shared between different classes, we filter out concepts with the smallest TCAV score [15] and cluster the remaining 6,000 ones using the k-means algorithm into $N = 50$ new clusters. These N clusters are visual words that form our visual language (see Fig. 2).

Cognitive vision systematic To obtain the meaning of the generated visual words, we use cognitive visual systematic [18] based on Marr's computational theory of vision [17]. According to Marr's theory, three levels of visual representations play an essential role in perception and discovering essential features of visible objects. These are the primal sketch, the 2.5D sketch, and the 3D model representation. The primal sketch is a two-dimensional image representation that uses light intensity changes, edges, colors, and textures. The 2.5D sketch represents mostly two-dimensional shapes, and the 3D model representation allows an observer to imagine the spatial object features based on its two-dimensional image. We will analyze six visual features from Marr's theory: brightness, color, texture, and lines (all primal sketch), shape (2.5D sketch), and form (3D model representation). We conduct user studies to establish the relationship between these features and individual visual words (see Fig. 3).

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_24

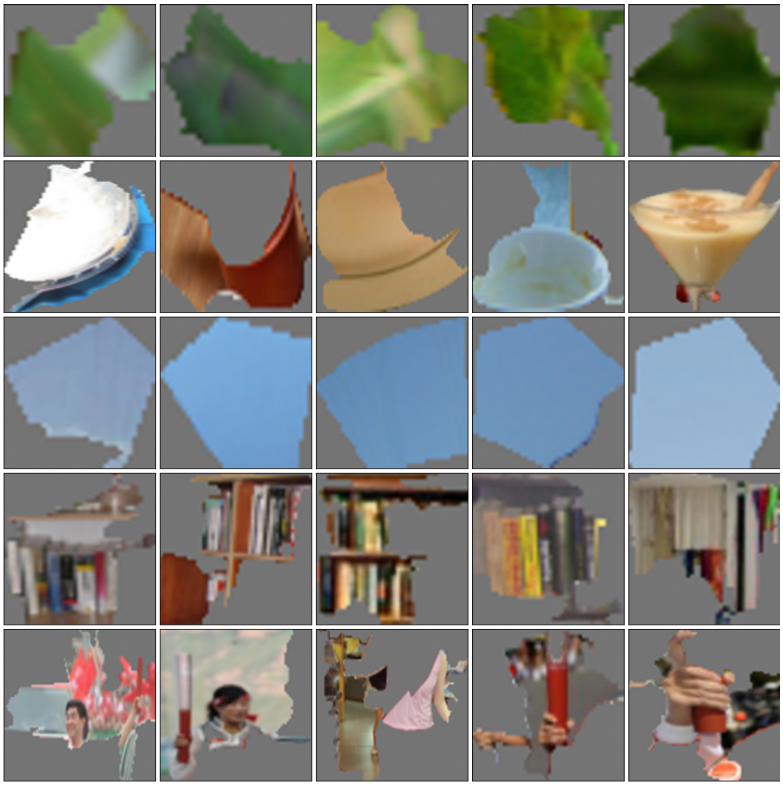


Fig. 2: Sample visual words, each represented by a row of 5 superpixels.

Amnesic visual probing We want to remove the information about a visual word from the representations and analyze how they differ from the original ones. For this purpose, we divide an image into superpixels, pass them through the network to obtain their representations, and assign them to the closest visual word. Then, we *define Word Content labels* $z_i \in \{0,1\}^N$ for representations $x_i \in \mathbb{R}^d$, where $z_i[j] = 1$ means that at least one superpixel of i-th image is assigned to j-th visual word.

Then, we *remove information about j-th visual word from a representation* x_i. For this purpose, we adapt an algorithm called Iterative Nullspace Projection (INLP) [19]. The probing classifier for $z_i[j]$ is parameterized by the matrix W_0. We first construct a projection matrix P_0 such that $W_0(P_0 x_i) = 0$ for all representations x_i (using method from [19]). Then, we iteratively train additional classifiers W_1 and perform the same procedure until no linear information re-

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_24

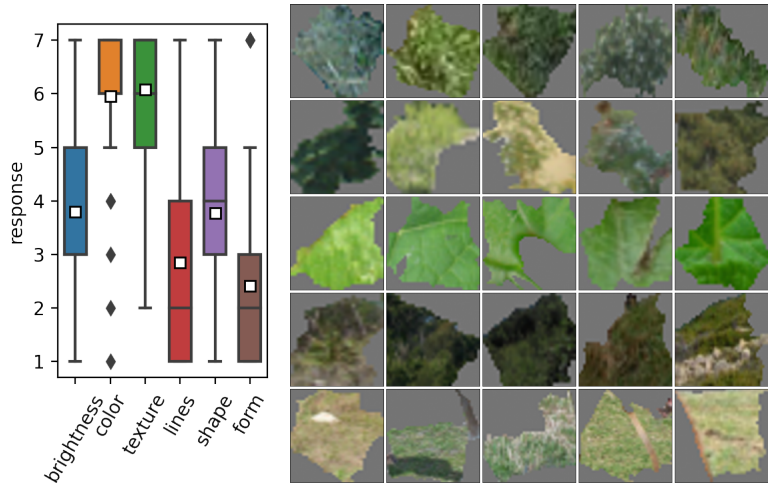


Fig. 3: Sample visual word (corresponding to grass) and its distributions of Likert scores obtained from user studies. One can observe that users mostly decided to assign this word to color and texture from the Marr's computational theory of vision.

garding $z_i[j]$ remains in x_i, i.e., until the chance of predicting the presence of a j-th visual word by the linear model is random. As a result we obtain a matrix $P_n \cdot P_{n-1} \cdot \ldots \cdot P_0$ which, when applied to representation, removes information about visual word j.

Finally, one can analyze changes in target task performance after removing information about a particular visual word. In this case, a target task is defined as multi-class classification with labels $y_i \in \{1, \ldots, k\}$, where $k = 1000$ is the number of ImageNet's classes. It is trained and tested for two types of representations, original and with removed visual word information.

4 User studies

To understand the meaning of visual words, we conduct user studies with 97 volunteers (64 males, 32 females, and 2 others aged 25 ± 7 years), including 71.1% students or graduates of computer science and related fields. Users completed an online survey with the number of questions corresponding to the number of visual words. We presented 12 typical (randomly chosen) superpixels for each visual word, and we asked to what extent a particular visual feature was essential for its creation. In reference to Marr's computational theory of vision [17] (see Section 3), six features were taken into consideration: brightness, color, texture,

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_24

Algorithm 1 Amnesic visual probing (AVP)

Require: X – set of image representations, Y – set of target labels, Z – set of visual words labels, C – codebook of visual words,
getNullSpaceProj(X, Z) – returns projection matrix that removes information about a visual word from representations,
trainValProb(X, Z) – trains model on probing task and returns validation accuracy,
trainValTarget(X, Y) – trains model on target task and returns validation accuracy
for each: $c \in$C
$X_{proj} \leftarrow X$
repeat
$P \leftarrow$getNullSpaceProj(X_{proj}, Z)
$X_{proj} \leftarrow PX_{proj}$
$acc_{prob} \leftarrow$trainValProb(X_{proj}, Z)
until $acc_{prob} \geq \frac{1}{2}$
$acc_{target} \leftarrow$trainValTarget(X, Y)
$acc^{\neg c}_{target} \leftarrow$trainValTarget$(X_{proj}, Y)$
$influence^c = acc^{\neg c}_{target} - acc_{target}$

lines (all primal sketch), shape (2.5D sketch) and form (3D model representation). We use the Likert scale with seven numerical responses from 1 to 7, corresponding to insignificant and key features, respectively.

Before completing the survey, users got familiarized with the examples of visual words with particular features selected by a trained cognitivist. They also completed two training trials to become familiar with the main task. Moreover, completing the task was not limited in time. Finally, due to the high number of visual words, assessing all 50 visual words would be tedious for the users. Therefore, we have prepared four questionnaire versions (one with twenty visual words and three with ten visual words) and assigned them to users randomly.

Based on the user studies results, we ranked the most representative visual words for each of the six features: brightness, color, texture, lines, shape, and form. We used those rankings to obtain detailed results of the amnesic visual probing.

5 Experimental Setup

Models We examine four self-supervised methods (MoCo v1 [14], SimCLR v2 [8], BYOL [13], and SwAV [6]), with a publicly available implementation based on the ResNet-50 (1x) architecture, trained on the entire ImageNet dataset[5]. We use the penultimate layer of ResNet-50 to generate representations with a length of 2048.

[5] We use the following implementations of the self-supervised methods: https://github.com/{google-research/simclr, yaox12/BYOL-PyTorch, facebookresearch/swav, facebookresearch/moco}.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_24

Data and target task We consider ImageNet [9] classification as the target task, but our approach could also be applied to other tasks. In order to get the classification model, we freeze the self-supervised trained model and fine-tune an ultimate fully-connected layer for 100 epochs. We conduct our experiments with a standard train/validation split.

Removing visual words Interventions that remove visual words are parametrized by 2048×2048 matrices applied to self-supervised representations. We obtain these matrices with our adaptation of the INLP algorithm, where we iterate until the probing classifier (detecting a visual word) achieves random accuracy.

Metric We consider the difference in top-5 classification accuracy before and after the intervention. For each self-supervised method, we carry out a series of interventions, removing the information about successive visual words from the ranking obtained based on the user studies (see Section 4). For each of the six features, we start with visual words considered as crucial for a given feature.

6 Results

As shown in Table 1, *removing visual words from self-supervised representations reduces the top-5 accuracy* of the target task. It is expected because, as presented in [4], image representation contains semantic knowledge. However, depending on a self-supervised model and a type of visual word, the level of degradation significantly differs. In the case of SimCLR v2, visual words related to the shape and form have the most significant influence on the classifier decisions. For BYOL, brightness and form have the greatest influence. Results for SimCLR and BYOL are also similar because they are least sensitive to texture removal from the representations. In contrast, MoCo and SwAV are the least sensitive to removing shape. In the case of MoCo, we also observe the most significant decrease in classification accuracy when removing forms, while the performance of SwAV is the most sensitive to color removal.

In Fig. 4, we present the most important visual words (according to our user studies) for each of the six visual concepts from Marr's computational theory of vision. These are visual words that we first remove from the representation.

In general, except for MoCo v1, representations are the least sensitive to removing textures from representations, which is inconsistent with what is found in [11]. Also, the two-dimensional shape is the most influential feature only for the classifier using the SimCLR v2 model. On the other hand, on average, removing visual words corresponding to the three-dimensional form and color from the self-supervised representation causes the most significant drop in the classification accuracy.

In Fig. 5, we present the change of target task accuracy when removing the successive most important visual words of the considered Marr's visual features (obtained with user studies). In general, the classification accuracy decreases as we remove the successive visual words. There are only a few exceptions to this,

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_24

Table 1: Removing visual words from the self-supervised representations influences the top-5 accuracy. The results are presented for six visual concepts from Marr's computational theory of vision. For each visual feature we remove five visual words according to the ranking obtained based on the user studies. The colors denote higher (dark blue) or lower (light blue) accuracy drop (in percentage points). These results demonstrate the biases in the self-supervised representations.

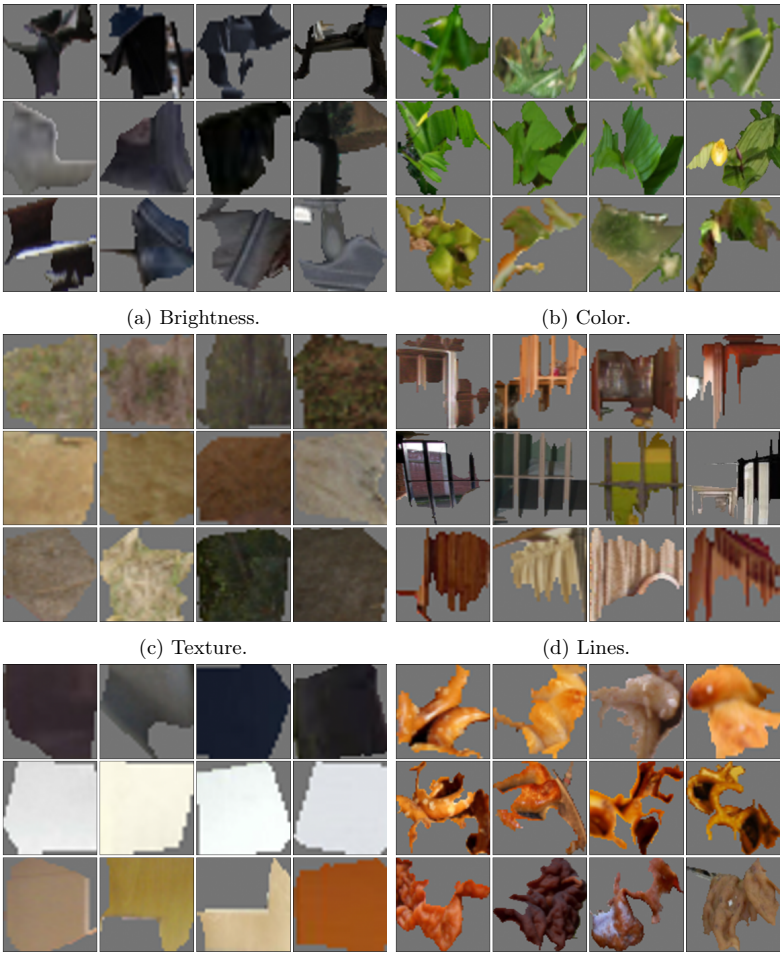
	top-5 acc.	**decrease in top-5 acc.**					
	no interv.	**remove visual words**					
		bright.	**color**	**texture**	**lines**	**shape**	**form**
MoCo v1	82.5	−3.09	−4.27	−3.84	−4.04	−2.98	−4.73
SimCLR v2	86.0	−2.00	−2.44	−1.60	−1.68	−2.51	−2.51
BYOL	86.5	−3.99	−3.37	−2.35	−2.75	−2.36	−3.49
SwAV	92.4	−2.20	−2.94	−1.56	−1.85	−1.00	−2.09

most notable in the case of SimCLR v2. We notice that in some cases, after removing two or three visual words from a given category, deleting the next ones causes only a slight further decrease in accuracy. It happens, for example, when removing visual words related to shape from SwAV representations or texture from SimCLR v2 representation. We also notice that in the case of three models (except MoCo v1), initially, when removing a small number of visual words, the most significant loss of accuracy occurs when removing the simplest visual features such as brightness (BYOL and SwAV) and color (SwAV and SimCLR v2). However, as we remove more visual words, the impact of removing more complex visual words corresponding to three-dimensional forms increases. This result may be because three-dimensional forms are more diverse and heterogeneous than colors and brightness.

Amnesic visual probing vs. Word Content probing task The correlation between the results of amnesic visual probing and the Word Content (WC) probing task is relatively weak, as presented in Fig. 6. The Pearson correlation coefficient ranges from 0.14 for SimCLR v2 to 0.52 for MoCo v1. In Fig. 6 we can see that although the WC probing task shows that there is a similar level of information about the visual words corresponding to lines and forms in SimCLR's representation, removing forms from this representation causes a much more significant decrease of target task accuracy than removing lines. The same relationship regarding lines and forms is also valid for BYOL, in which case the correlation between target task accuracy and WC results is the largest among the examined methods, even though it is still weak.

In general, this weak correlation supports the thesis that the WC probing task focuses on what visual words are encoded in the representation, but it does not assess how this information is used. Therefore, we conclude that the *Word Content probing task cannot be directly used to evaluate target task accuracy,*

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_24



(a) Brightness. (b) Color.

(c) Texture. (d) Lines.

(e) Shape (2D). (f) Form (3D).

Fig. 4: The most important visual words (according to our user studies) for each of the six visual concepts from Marr's computational theory of vision.

which justifies the introduction of amnesic visual probing. Nevertheless, WC is still needed for amnesic visual probing to analyze the representation and should be considered as a complementary tool.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_24

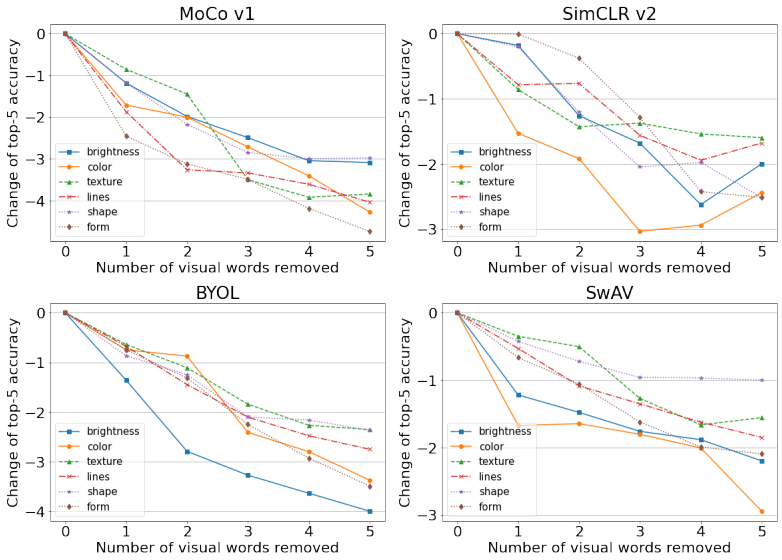


Fig. 5: Decrease in top-5 accuracy (in percentage points) when removing the information about successive visual words according to the ranking obtained based on the user studies, presented for six visual concepts from Marr's computational theory of vision.

7 Conclusions

The visual probing framework provides interesting insight into the self-supervised representations. However, this insight does not correspond to the performance of the target task. Hence, we propose Amnesic Visual Probing (AVP) to analyze the visual concepts that influence the target task. Thanks to preserving the semantic taxonomy of visual words from the visual probing framework, we can use AVP to examine and compare the biases of individual self-supervised methods. Finally, the user studies allow us to describe those biases using six visual features from Marr's computational theory of vision.

Acknowledgments

This research was funded by Foundation for Polish Science (grant no POIR.04.04.00-00-14DE/18-00 carried out within the Team-Net program co-financed by the European Union under the European Regional Development Fund), National Science Centre, Poland (grant no 2020/39/B/ST6/01511). The authors have applied a CC BY license to any Author Accepted Manuscript (AAM) version aris-

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_24

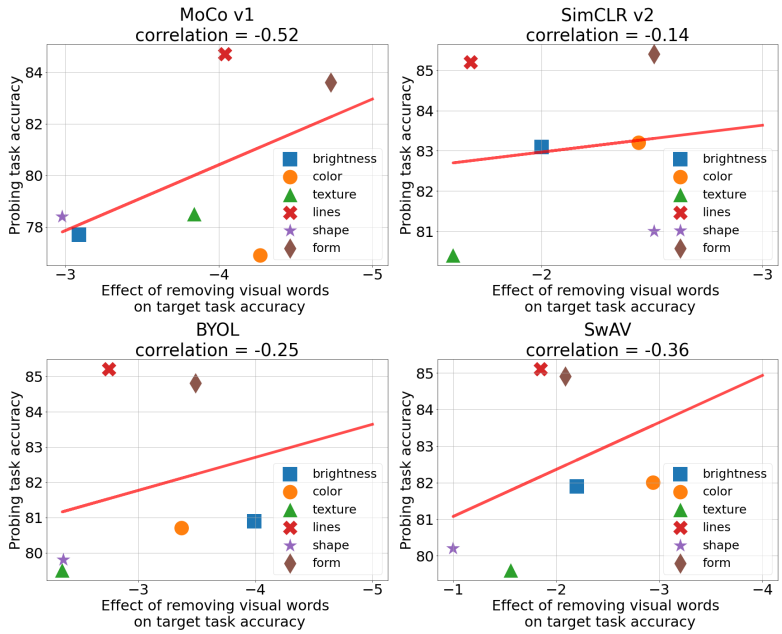


Fig. 6: There is a weak correlation between the results of amnesic visual probing (in percentage points) and the Word Content (WC) probing task (in percents). It means that WC cannot be directly used to evaluate target task accuracy. Hence introducing the amnesic visual probing is justified.

ing from this submission, in accordance with the grants' open access conditions. Dominika Basaj was financially supported by grant no 2018/31/N/ST6/02273 funded by National Science Centre, Poland.

References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE transactions on pattern analysis and machine intelligence **34**(11), 2274–2282 (2012)
2. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. arXiv preprint arXiv:1810.03292 (2018)
3. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
4. Basaj, D., Oleszkiewicz, W., Sieradzki, I., Górszczak, M., Rychalska, B., Trzcinski, T., Zieliński, B.: Explaining self-supervised image representations with visual probing. In: IJCAI-21. pp. 592–598 (8 2021). https://doi.org/10.24963/ijcai.2021/82

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_24

5. Belinkov, Y., Glass, J.: Analysis methods in neural language processing: A survey. Transactions of the Association for Computational Linguistics **7**, 49–72 (2019)
6. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Proceedings of Advances in Neural Information Processing Systems (NeurIPS) (2020)
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 1597–1607. PMLR (13–18 Jul 2020)
8. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E.: Big self-supervised models are strong semi-supervised learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 22243–22255. Curran Associates, Inc. (2020)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
10. Elazar, Y., Ravfogel, S., Jacovi, A., Goldberg, Y.: Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. Transactions of the Association for Computational Linguistics **9**, 160–175 (03 2021)
11. Geirhos, R., Narayanappa, K., Mitzkus, B., Bethge, M., Wichmann, F.A., Brendel, W.: On the surprising similarities between supervised and self-supervised models. arXiv preprint arXiv:2010.08377 (2020)
12. Ghorbani, A., Wexler, J., Zou, J., Kim, B.: Towards automatic concept-based explanations. arXiv preprint arXiv:1902.03129 (2019)
13. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., Piot, B., kavukcuoglu, k., Munos, R., Valko, M.: Bootstrap your own latent - a new approach to self-supervised learning. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 21271–21284. Curran Associates, Inc. (2020)
14. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9726–9735 (2020). https://doi.org/10.1109/CVPR42600.2020.00975
15. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In: International conference on machine learning. pp. 2668–2677. PMLR (2018)
16. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25**, 1097–1105 (2012)
17. Marr, D.: Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. Henry Holt and Co., Inc., New York, NY, USA (1982)
18. Oleszkiewicz, W., Basaj, D., Sieradzki, I., Górszczak, M., Rychalska, B., Lewandowska, K., Trzcinski, T., Zielinski, B.: Visual probing: Cognitive framework for explaining self-supervised image representations. CoRR **abs/2106.11054** (2021), https://arxiv.org/abs/2106.11054
19. Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., Goldberg, Y.: Null it out: Guarding protected attributes by iterative nullspace projection. In: Proceedings

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_24

of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7237–7256. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.647
20. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
21. Sivic, J., Zisserman, A.: Video google: Efficient visual search of videos. In: Toward category-level object recognition, pp. 127–144. Springer (2006)

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_24

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 5 January 2023, accepted 2 February 2023, date of publication 6 February 2023, date of current version 10 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3242982

RESEARCH ARTICLE

Visual Probing: Cognitive Framework for Explaining Self-Supervised Image Representations

WITOLD OLESZKIEWICZ[1], DOMINIKA BASAJ[1,5], IGOR SIERADZKI[2], MICHAŁ GÓRSZCZAK[2], BARBARA RYCHALSKA[1,4], KORYNA LEWANDOWSKA[3], TOMASZ TRZCINSKI[1,2,5], (Senior Member, IEEE), AND BARTOSZ ZIELIŃSKI[2,6]

[1]Warsaw University of Technology, 00-661 Warszawa, Poland
[2]Faculty of Mathematics and Computer Science, Jagiellonian University, 31-007 Kraków, Poland
[3]Cognitive Neuroscience and Neuroergonomics, Institute of Applied Psychology, 30-060 Kraków, Poland
[4]Synerise, 30-383 Kraków, Poland
[5]Tooploox, 53-601 Wrocław, Poland
[6]Ardigen, 30-394 Kraków, Poland

Corresponding author: Witold Oleszkiewicz (witold.oleszkiewicz@pw.edu.pl)

This work was supported in part by the Foundation for Polish Science (carried out within the Team-Net Program co-financed by the European Union under the European Regional Development Fund) under Grant POIR.04.04.00-00-14DE/18-00; and in part by the National Science Centre, Poland, under Grant 2020/39/B/ST6/01511. The work of Dominika Basaj and Barbara Rychalska was supported by the National Science Centre, Poland, under Grant 2018/31/N/ST6/02273.

ABSTRACT Recently introduced self-supervised methods for image representation learning provide on par or superior results to their fully supervised competitors, yet the corresponding efforts to explain the self-supervised approaches lag behind. Motivated by this observation, we introduce a novel visual probing framework for explaining the self-supervised models by leveraging probing tasks employed previously in natural language processing. The probing tasks require knowledge about semantic relationships between image parts. Hence, we propose a systematic approach to obtain analogs of natural language in vision, such as visual words, context, and taxonomy. Our proposal is grounded in Marr's computational theory of vision and concerns features like textures, shapes, and lines. We show the effectiveness and applicability of those analogs in the context of explaining self-supervised representations. Our key findings emphasize that relations between language and vision can serve as an effective yet intuitive tool for discovering how machine learning models work, independently of data modality. Our work opens a plethora of research pathways towards more explainable and transparent AI.

INDEX TERMS Computer vision, explainability, probing tasks self-supervised representation.

I. INTRODUCTION

Visual representations are cornerstones of a multitude of contemporary computer vision and machine learning applications, ranging from visual search [8] to image classification [9] and Visual Question Answering, VQA [10]. However, learning representations from data typically requires tedious annotation. Therefore, recently introduced self-supervised representation learning methods concentrate on decreasing the need for data labeling without reducing their performance [1], [20], [21]. Because of the fundamental role representations play in real-life applications, much research focuses on explaining these embeddings [6], [15], [17]. Nevertheless, most of them concentrate on fully supervised embeddings [11] rather than on their self-supervised counterparts. Moreover, the majority of the proposed approaches rely on pixel-wise image analysis [13], [14], while general semantic concepts present in the images are often ignored.

Here, we attempt to overcome these shortcomings and draw inspiration from a simple yet often overlooked

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Shorif Uddin.

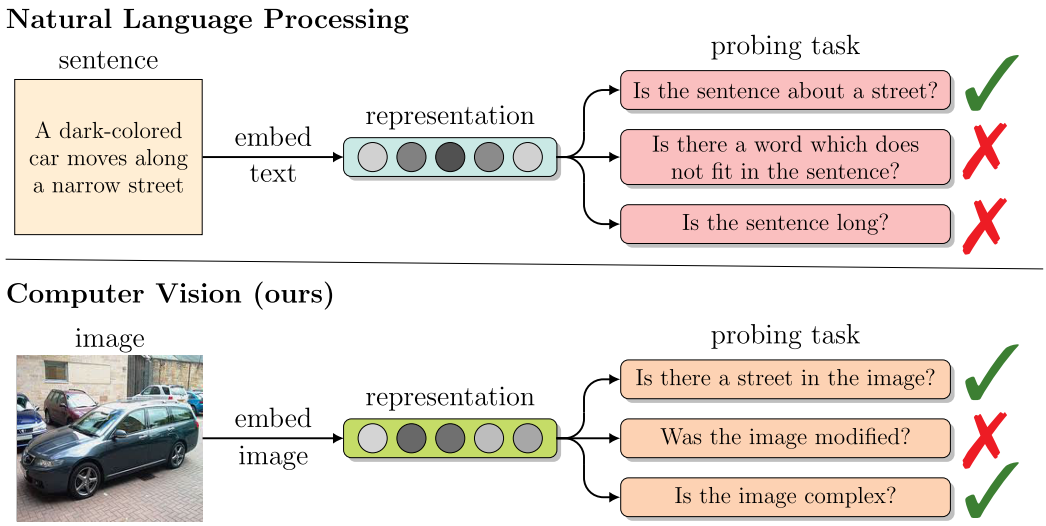


FIGURE 1. ***Probing tasks*, widely used in natural language processing, validate if a *representation* implicitly encodes a given property, e.g., a sentence topic or its length. We introduce a visual taxonomy along with the corresponding probing framework that allows building analogous *visual probing tasks* and explain the self-supervised image representations. As a result, we e.g., discover that the information stored by self-supervised representations is more biased towards lines and forms than textures.**

observation that humans use language as a natural tool to explain what they learn about the world through their eyes [16]. Therefore, considering that the very same machine learning algorithms can be successfully applied to solve both vision and Natural Language Processing (NLP) tasks [23], [24], we postulate that the methods used to analyze text representation can also be employed to investigate visual inputs.

Very popular tools for explaining textual embeddings are *probing tasks* [18]. As shown in the upper part of Figure 1, a probing task in NLP is a simple classifier that asks if a given textual representation encodes a particular property, such as a sentence length or its semantic consistency, even though this property was not a direct training objective. For instance, we can create a textual probing task by substituting a word in a sentence and checking if a simple classifier that takes the representation of the original and altered sentence can detect this change. Furthermore, by analyzing the accuracy of a probing task, one can verify if the investigated representation contains certain information and understand the rationale behind embedding creation. However, while probing tasks are straightforward, intuitive, and widely used tools in NLP, their computer vision application is limited [12], mainly due to the lack of appropriate analogs between textual and visual modalities.

In this paper, we address this limitation by introducing an intuitive mapping between vision and language that enables applying the NLP probing tools in the computer vision (CV) domain. For this purpose, in Section III, we propose a taxonomy of visual units that includes *visual sentences*, *words*, and *characters*. We describe them using visual features presented in Marr's computational theory of vision [36], such as texture, shapes, and lines. Finally, we employ them as building blocks for a more general visual probing framework that contains a variety of NLP-inspired probing tasks, such as *Word Content*, *Sentence Length*, *Character Bin*, and *Semantic Odd Man Out* [17], [18]. The results we obtain provide us with unprecedented insights into semantic knowledge, complexity, and consistency of self-supervised image representations, e.g. we discover that semantics of the image only partially contribute to target task accuracy. One of our key findings is that the information stored by self-supervised representations is much more influenced by lines and forms than textures. What confirms the design choices behind hand-crafted visual representations such as SIFT [52] or BRIEF [53]. Our framework also allows us to compare the existing self-supervised representations from a novel perspective, as shown in Section VI.

Our contributions can be therefore summarized as follows:

- We propose an intuitive mapping between visual and textual modalities that constructs a visual taxonomy.
- We introduce novel visual probing tasks for comparing self-supervised image representations inspired by similar methods used in NLP.
- We show that leveraging the relationship between language and vision serves as an effective yet intuitive tool for discovering how self-supervised models work.

II. RELATED WORKS

The visual probing framework aims to explain image representations obtained from self-supervised methods. Moreover, it is inspired by probing tasks used in NLP. Therefore, in this

section, we consider related works from three research areas: self-supervised computer vision models, probing tasks in natural language processing, and explainability methods in computer vision.

A. SELF-SUPERVISED COMPUTER VISION MODELS

Earliest self-supervised methods were based on a pretext task, for example, image colorization [45], or rotation prediction [46] using cross-entropy loss. However, recently published state-of-the-art methods usually base on contrastive loss [30], which measures the similarities of patches in representation space and aims to discriminate between positive and negative pairs. The positive pair contains modified versions of the same image, while the negative pairs correspond to two images in the same dataset. One of the methods, called MoCo v1 [27] trains a slowly progressing encoder, driven by a momentum update. This encoder plays the role of a large memory bank of past representations and delivers information about negative examples. Another method, called SimCLR v2 [1], proposes a different way of generating negative pairs, using a large batch size of up to 4096 examples. Other important improvements proposed by SimCLR v2 are the projection head and carefully tuned data augmentation. The projection head maps representations into space where contrastive loss is applied to prevent the loss of information. On the other hand, BYOL [20] also uses the projection head, but unlike MoCo v1 and SimCLR v2, it achieves a state-of-the-art performance without the explicitly defined contrastive loss function, so it does not need negative examples. Finally, SwAV [21] first obtains ''codes'' by assigning features to prototype vectors and then solves a ''swapped'' prediction problem wherein the codes obtained from one data augmented view are predicted using the other view. Our paper provides a framework for analyzing the representations generated by those methods regarding the semantic knowledge they encode.

B. PROBING TASKS IN NLP

NLP probing tasks aim to probe word or sentence representations for interesting linguistic features to discover whether they contain linguistic knowledge [48]. Probing is usually achieved with a binary or multi-class classifier, which takes one or two-word embeddings as input, and predicts the existence or absence of a chosen linguistic phenomenon in the input representation(s) [18]. The qualities of a good probing classifier are the subject of debate, as too expressive probes could learn important features on their own, even if the information is not present in the representations [5]. Thus, probing is usually achieved with simple classifiers.

Classic probing literature considers various linguistic aspects, from the simplest to very complex ones. In [18], the probed linguistic features are, for example, the depth of the sentence parse tree or whether the sentence contains a specific word. Other works propose to focus on lexical knowledge concerning the qualities of individual words more than the whole sentences [6], [17], probing token embeddings for qualities such as gender, case, and tense, or differentiation between real words and pseudowords [17]. Other approaches focus on certain kinds of words, e.g., function words, such as *wh*-words and propositions [4]. We consider all these objectives in our approach, i.e., we study probing tasks on individual concepts and their compositions. Moreover, while most works on probing tasks focus on one selected language, the others [17] are designed with multilingual settings in mind. It has been shown that it is possible to create NLP probing tasks that are transferable across languages, even if the languages vary considerably in their structure, which means that probing tasks can touch upon more universal cognitive phenomena [2]. This paper also aims at the flexibility and universality of our probing tasks, as our approach can be applied to various image domains.

C. EXPLAINABILITY METHODS IN CV REPRESENTATION LEARNING

eXplainable Artificial Intelligence (XAI) gains popularity fuelled by the black-box character of today's deep neural networks [19], [39]. Popular explainability approaches for model explanations are saliency or attention maps, which provide the importance of weights to pixels based on the first order derivatives [13], [14], [40], [41] but do not fully explain the reasoning behind the actual decision [54] and do not describe the concrete semantic concepts. Moreover, some of the methods are even agnostic of the model itself [13] and thus are not able to explain it. Another common local approach is perturbation-based interpretability, which applies changes to either data [55] or features [56] and observes the influence on the output.

Some methods verify the relevance of network hidden layers. For example, [12] uses linear classifiers trained on representations from these layers to measure how suitable they are for the classification. Subsequent efforts focused on understanding the function of hidden layers led to the introduction of network dissection [42], [43], which enables quantifying the interpretability of latent representations by evaluating the alignment between their hidden units and a set of visual semantic concepts obtained from human annotators.

More recent methods are inspired by the human brain and how it explains its visual judgments by pointing to prototypical features that an object possesses [57]. I.e., a certain object is a car because it has tires, a roof, headlights, and a horn. For example, prototypical part network [44] applies this paradigm by focusing on parts of an image and comparing them with prototypical parts of a given class. At the same time, the extension proposed in [58] uses data-dependent merge-pruning of the prototypes to allow sharing them among the classes. Another promising approach is Concept Activation Vector (CAV), defined in the feature space to quantify the degree to which a predefined concept is vital for a prediction [26]. This approach has recently been extended to automatically discovered concepts [19] and to interactive

techniques used by pathologists to indicate what characteristics are essential when searching for similar images [59].

We propose to continue and extend this line of research by introducing visual word probing, which systematically explains the self-supervised representations. Our work presents a framework that focuses on model analysis. It interprets the internal representation of the deep learning model using visual probing tasks, e.g., it shows which semantic concepts are included in the representation and to what extent.

III. VISUAL PROBING

This section introduces a novel visual probing framework that analyzes the information stored in self-supervised image representations. For this purpose, in Section III-A, we propose a mapping between visual and textual modalities that constructs a visual taxonomy. As a result, the image becomes a ''visual sentence'' constructed from ''visual words'' and can be analyzed with visual probing tasks inspired by similar methods used in NLP (see Section III-C). Moreover, for in-depth analysis of the concepts trained by self-supervised methods, in Section III-B, we provide a cognitive visual systematic that identifies a visual word with structural features from Marr's computational theory [36].

A. MAPPING BETWEEN VISION AND NLP

After defining the images as analogous to sentences within our framework, the question remains which parts of an image should be considered equivalent to individual words and characters? There are multiple possible answers to this question. One of the intuitive ones is to divide an image into non-overlapping superpixels that group pixels into perceptually meaningful atomic regions, e.g., using SLIC algorithm [28]. As a result, we obtain an image built from superpixels, an analogy of a sentence built from words. The superpixels, similarly to words, have their order and meaning (see Section III-B). Moreover, each superpixel contains a specific number of pixels, like the number of characters in a word. As a consequence, we obtain an intuitive mapping between visual and textual domains.

However, superpixels differ conceptually from their linguistic counterparts in one important aspect: they do not repeat between different images, while in text, the words often repeat between sentences. Therefore, we propose to define visual words as the clusters of all training superpixels in representation space and assign each superpixel to the closest centroid from such a dictionary. For this purpose, we could use the original definition of visual words from [25]. However, it does not take into consideration the importance of those words for a model's prediction. Therefore, we use TCAV methodology [19], [26] that generates high-level concepts, which are important for prediction and easily understandable by humans. Such an approach requires a supervisory training network but generates visual words independent of the analyzed self-supervised techniques, which is crucial for a fair comparison. To summarize, the process of dividing an image into visual words consists of three steps: segmentation into superpixels, their encoding, and assignment to visual words (see Figure 2).

B. COGNITIVE VISUAL SYSTEMATIC

In contrast to words in NLP, visual words do not have a well-defined meaning required for in-depth analysis of self-supervised representations. Hence, in this section, we introduce cognitive visual systematic, considering that generating visual words is similar to the process of concept formation. This process, described in psychology and cognitive science, is traditionally understood as an internal cognitive representation of a set of similar objects, i.e., ''an idea that includes all that is characteristically associated with it'' [37]. In other words, concepts are created in relation to features that constitute similarity amongst included objects.

What features could then be the basis for the formation of visual words? Reference to Marr's computational theory of vision [35], [36] seems to be an appropriate aid in answering this question. Marr assumed that perception is achieved by detecting an object's specific structural features, which are then organized in a series of visual representations. Among those, three constitute the major representations: the ''primal sketch'', the ''2.5D sketch'' and the 3D model representation'' [36]. The primal sketch is a two-dimensional image that uses information on light intensity changes, featuring blobs, edges, lines, boundaries, bars, and terminations. Colors and textures are also thought to be detected on this level [34], [38]. The 2.5D sketch represents mostly two-dimensional shapes and their orientation towards a viewer-centered location (the sense of image depth is achieved in this stage [35]). Finally, the 3D model is a representation suitable for object recognition. In this stage, the observer can imagine the object from different views. This includes surfaces that are currently invisible to the observer [35], [36].

To simplify visual word description in terms of Marr's theory, we decided to use concepts of light intensity (brightness), color, texture, and lines in relation to primal sketch, shape in relation to 2.5D sketch, and form in relation to 3D model (examples are depicted in Figure 3). Our initial analysis of individual visual words shows that these six categories from Marr's theory describe very well the particular types of our visual words. In the process of creating visual words (see Figure 2), similar superpixels cluster together. As a result, we generate different visual words consisting of similar lines, similar shapes, similar colors, etc. To confirm that our observations are not accidental, we conducted user studies that confirmed our assumptions and categorized visual words into specific categories from Marr's theory, such as brightness, color, texture, lines, forms, and shapes. This user study helps us to establish the meaning of the visual words we use.

C. VISUAL PROBING TASKS

After dividing an image into visual words, its representation can be analyzed by the visual probing framework that can adapt most NLP probing tasks. Here, we describe adaptations

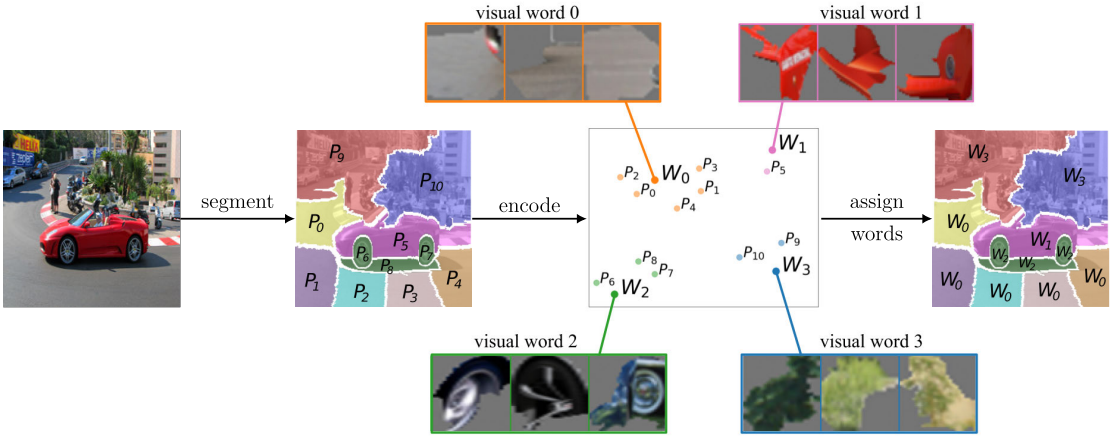


FIGURE 2. The process of dividing an image into visual words. First, an image is segmented into multiple superpixels: P_0, P_1, ..., P_{10}. Then, each superpixel is embedded in the latent space previously used to generate the dictionary of visual words: W_0, W_1, W_2, W_3. Finally, each superpixel is assigned to the closest word in the visual word dictionary. This results in mapping between vision and language and enables using the visual probing framework that includes a variety of NLP-inspired probing tasks.

of five of them, including those well known in the NLP community [17], [18] together with their original NLP definitions to make the paper self-contained.

1) WORD CONTENT (WC)

The *Word Content* probing task aims to identify which visual words are present in an image (see Figure 4). The *input* of this probing task is a self-supervised representation of the image. The *target labels* represent the presence of a particular visual word. As we describe in Section IV, all visual words are clustered into 50 clusters. Hence, there are 50 binary *target labels*. Figure 2 illustrates the process of determining which visual words are present in the image. This is similar to the bag of words representation.

The NLP inspiration of this task probes for surface information, i.e., the type of information that does not require any linguistic knowledge [18]. In contrast, its adaptation requires *semantic knowledge* to understand which concept is represented by a superpixel.

2) SENTENCE LENGTH (SL)

The aim of the *Sentence Length* probing task is to distinguish between simple and complex images, as presented in Figure 5. The *input* of this probing task is a self-supervised representation of the image. The *target label* is the number of unique visual words in the image, which can be determined based on the WC labels. The original NLP probing task predicts the number of words (or tokens) and retains only surface information [18]. In CV, it serves as a proxy for *semantic complexity*, requiring the semantic understanding of the image.

3) CHARACTER BIN (CB)

The aim of the *Character Bin* probing task is to check whether the representation stores information about the complexity of the visual word represented by a superpixel. The *input* of this probing task is a self-supervised representation of the image's superpixel, and we define two *target labels* that are commonly used in CV literature to describe superpixels. The first target label is the compactness (CO) [49] of the superpixel S defined as the area of the superpixel $A(S)$ divided by the area $A(C)$ of a circle C with the same perimeter as S:

$$CO(S) = \frac{A(S)}{A(C)}.$$

Sample superpixels with various ranges of CO are presented in Figure 6a. The second target label is Intra-Cluster Variation (ICV) [50] defined as the average standard deviation $\sigma_c(S)$ of channels $c \in C$ for superpixel S:

$$ICV(S) = \frac{1}{|C|} \sum_{c \epsilon C} \sigma_c(S).$$

Sample superpixels with various ranges of ICV are in Figure 6b. The original NLP probing task is defined as a classifier of the number of characters in a single word [17]. From this perspective, the *Character Bin* retains only surface information in both domains.

4) SEMANTIC ODD MAN OUT (SOMO)

The objective of the SOMO probing task is to predict whether the image was modified. We replace a center-biased superpixel in the image with a similarly shaped superpixel from another image that corresponds to different visual words. We pick a superpixel using a two-dimensional Gaussian distribution center in the middle of the image. Regarding replacement, we consider two setups, SOMO close and far, depending on how often two visual words co-occur in the training images. In SOMO close, we replace a center-biased superpixel with visual words that often co-occur with the replaced visual word. In SOMO far, we replace superpixel

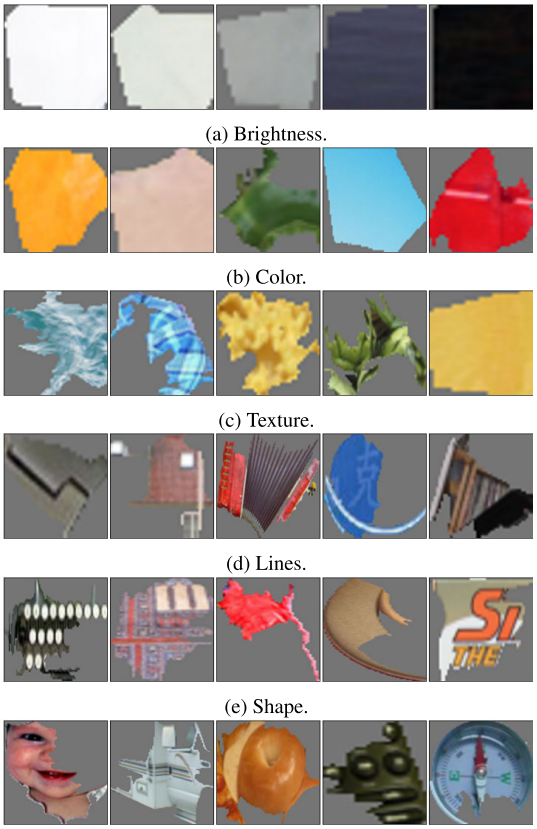


FIGURE 3. Sample superpixels illustrate six visual concepts from the Marr's computational theory of vision.

FIGURE 4. Sample visual words, each represented by one row of five superpixels.

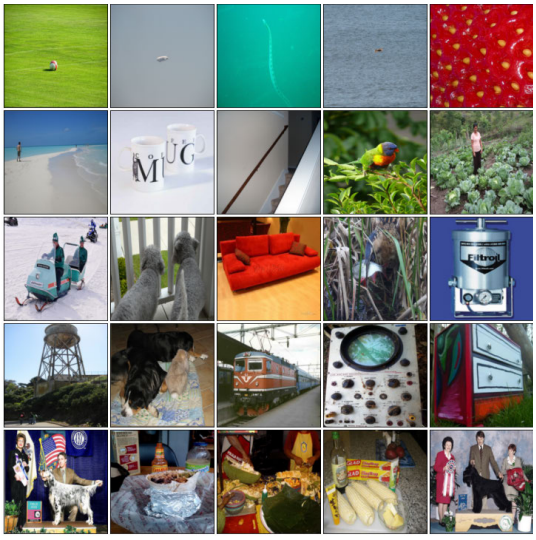
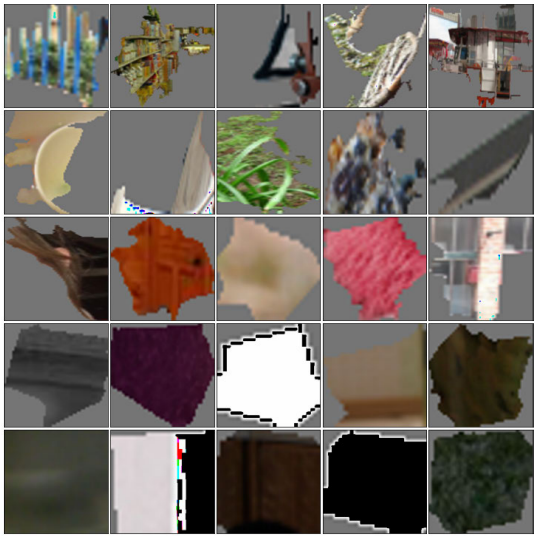


FIGURE 5. Sample images grouped into rows with increasing value of *Sentence Length* from top to bottom. One can observe that SL correlates with the semantic complexity of the image.

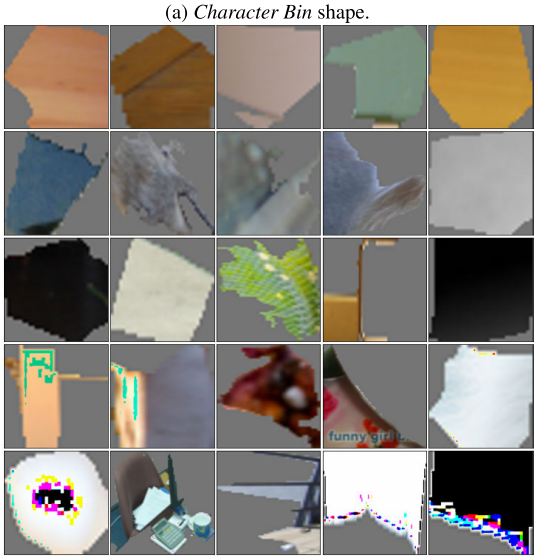
with the rarely co-occurring visual word (see Figure 7). In both cases, the *input* of the probing task is a self-supervised representation of the image. The *target label* is binary, i.e., whether the image was modified or not. The original NLP task predicts if replacing a random noun or verb alters the sentence [18]. In both domains, it requires the ability to detect alterations in *semantic consistency*.

5) MUTUAL WORD CONTENT (MWC)

The *Mutual Word Content* (MWC) probing task aims to discover which visual words bring two self-supervised representations close to each other and which ones push them farther away (see Figure 8). The *input* of this probing task is a pair of self-supervised representations of two images. The *target labels* represent the presence of a particular visual word in both images. The probing task classifier is validated on equally-sized subsets $\{S^i_{val}\}$ corresponding to the increasing cosine distance between pairs of representations. Discrepancies in the classifier accuracy show the impact of visual words on the representations' distance. More precisely, if the probing task performance drops with increasing representations' distance, the visual word information in both representations brings them closer. To quantify this relationship, we introduce the attraction coefficient. To calculate this coefficient, we use the Linear Regression fit to the points (i, AUC_i), where i is the index of the subset and AUC_i is the MWC probing task performance on this subset. Thus, the attraction coefficient is the first derivative of the fitted model.



(a) *Character Bin* shape.



(b) *Character Bin* color.

FIGURE 6. Sample superpixels grouped into rows with increasing values of *CO* (a) and *ICV* (b) from top to bottom. One can observe that bottom rows contain superpixels of rounder shape (a) and higher contrast (b).

This probing task does not have a direct counterpart in the NLP domain.

IV. EXPERIMENTAL SETUP

This section describes how we generate visual words and self-supervised representations, assign visual words to images, and train the probing tasks.[1]

[1]The code available at github.com/BioNN-InfoTech/visual-probes

(a) SOMO far.

(b) SOMO close.

FIGURE 7. Sample images from SOMO far (a) and close (b) setup. Replacements in SOMO close come from a set of visual words that often co-occur with the replaced visual word. In SOMO far, we replace superpixel with rarely co-occurring visual words. One can observe that in the case of SOMO close, the differences are less visible. Notice that red arrows indicate alterations to the image.

A. GENERATING VISUAL WORDS

This paper presents a general framework that can be used with various methods of generating visual words. However, choosing a high-quality method is crucial to draw meaningful conclusions from the probing tasks. That is why we use the established ACE algorithm [19]. It first divides images into superpixels using SLIC (Simple Linear Iterative Clustering) algorithm [28]. This algorithm clusters pixels in the combined five-dimensional color and position space to generate compact, nearly uniform superpixels. This approach has only one parameter that specifies the number of output superpixels. We use the SLIC algorithm with three resolutions of 15, 50, and 80 segments for each image. Next, it generates representations of these superpixels as an output of the *mixed4c* layer of GoogLeNet [7] trained on the ImageNet dataset. Then, for each class separately, corresponding representations are clustered using the k-means algorithm with $k = 25$ and filtered to remove infrequent and unpopular clusters (as described in [19]). This results in around 18 concepts per class and approximately 18, 000 concepts for the whole ImageNet dataset. They could be directly used as visual words. However, such words would be exclusive for particular classes, and some of them would be ambiguous due to the small TCAV score [26]. Hence, to obtain a reliable dictionary with visual words shared between classes, we filter out 12, 000 concepts with the smallest TCAV score and cluster the remaining 6, 000 concepts using the k-means algorithm into 50 clusters treated as visual words (see Figures 4 and 9). We do not treat the number of clusters as a tunable hyperparameter. Instead, we set a fixed number of clusters, ensuring various concepts and making user studies feasible.

B. GENERATING A SELF-SUPERVISED REPRESENTATION

We examine four self-supervised methods: MoCo v1 [27], SimCLR v2 [1], BYOL [20], and SwAV [21]. For all of them,

(a) Superpixels that attract representations.

(b) Superpixels that push representations away.

FIGURE 8. Sample pairs of images with visual words (represented by marked superpixels) that attract (a) or push away (b) the representations. Visual words that attract representations include words with complicated forms and "green" words. On the other hand, visual words that push representations away contain fine textures.

we use publicly available models trained on ImageNet.[2] Although they all use the penultimate layer of ResNet-50 to generate representations, their training hyperparameters differ, which is presented in Table 5.

[2]We use the following implementations of self-supervised methods: https://github.com/{google-research/simclr, yaox12/BYOL-PyTorch, facebookresearch/swav, facebookresearch/moco}. We use ResNet-50 (1x) variant for each self-supervised method.

TABLE 1. Bins ranges corresponding to the classes in *Sentence Length* (SL) and *Character Bin* (CB) probing tasks.

bin	SL	CB shape	CB color
0	< 18	< 0.153	< 0.063
1	$[18, 21)$	$[0.153, 0.207)$	$[0.063, 0.085)$
2	$[21, 23)$	$[0.207, 0.263)$	$[0.085, 0.104)$
3	$[23, 26)$	$[0.263, 0.336)$	$[0.104, 0.125)$
4	$[26, 28)$	$[0.336, 0.462)$	$[0.125, 0.155)$
5	$28 \leq$	$0.462 \leq$	$0.155 \leq$

C. ASSIGNING VISUAL WORDS

To assign a superpixel to a visual word, we first pass it through the GoogLeNet to generate a representation from the *mixed4c* layer (similarly to generating visual words). Since all concepts considered in Section IV-A are grouped into 50 clusters (visual words), we use a two-stage assignment. First, we find the closest concept and then assign the superpixel to the visual word containing this concept.

D. TRAINING PROBING TASKS

We use a logistic regression classifier with the LBFGS solver [61] to train all diagnostic classifiers. As input, we use representations generated by the self-supervised methods. The output depends on the probing task. In the case of *Word Content*, we train 50 classifiers corresponding to 50 visual words. Furthermore, we expect an image to be assigned to a particular visual word if at least one of its superpixels is assigned to it. Finally, we report the average AUC scores over 50 classifiers (see Table 2). To formulate a classification setup in the *Sentence Length* probing task, we group the possible output into six equally-sized bins (see Table 1), resulting in one-vs-one OVO AUC, which is resistant to class imbalance. A similar procedure is applied to the *Character Bin* probing tasks. SOMO is formulated as a binary classification task in which we predict whether the image was modified. We train two separate classifiers for two use cases, SOMO far and SOMO close, with balanced training and validation sets.

We conduct all of our experiments on the ImageNet dataset [29] with standard train/validation split. Moreover, we apply random over-sampling if needed to deal with the imbalanced classes.

V. USER STUDIES

While the cognitive visual systematic introduced in Section III-B presents the possible way of obtaining the meaning of visual words, it requires human observers to reliably decide which visual features should be assigned to particular visual words. Hence, in this section, we describe user studies conducted to establish this assignment.

Overall, 40 volunteers participated in the study (30 males and 10 females aged 29 ± 10 years) recruited online. 62.5% of the participants were students/graduates of computer science and related fields, and the remaining attendees represented various backgrounds.

The description and questions of the study were in English and Polish. Participants ranged from 18 to 66 years of age.

The average age of the participant is 29, and 68% of the participants are between 20 and 38 years old. 75% of the participants declared themselves as male, 25% as female and 0% chose other options. Participants were recruited online. 62% of the participants were students or graduates of computer science and related fields, and the remaining attendees represented various backgrounds, e.g., medicine, law, and psychology. 35% of participants have at least a bachelor's degree.

Users completed an online questionnaire. Their task was to assess the similarity of superpixels representing a visual word and provide key features associated with this visual word. To this end, users were presented with 20 visual words consisting of 12 representative superpixels (close to the visual word center) each. Participants were instructed to use Likert scales with seven numerical responses with only endpoints labeled (1 and 7) for clarity. First, they were asked to evaluate the homogeneity of a given set (scale endpoints: great variety; great homogeneity; see Figure 14). Next, they evaluated to what extent a given feature was essential for visual word creation. In reference to Marr's computational theory of vision [36] (see Section III-B), six features were taken into consideration: light intensity (brightness), color, texture, lines, shape (Marr's 2.5D sketch) and form (Marr's 3D model representation). Scale endpoints were labeled as a not significant feature and a key feature (see Figure 14).

Before the main task, users obtained an instruction that included sample visual words with particular features (selected by a cognitivist). They also underwent two training trials to familiarize themselves with the task. There were no time constraints for trial or task completion. The order of visual words and on-screen localization of superpixels were semi-randomized for each participant.

Due to the high number of visual words, the assessment of all 50 visual words would be tedious for the users. That is why we decided to limit our user study to the twenty most reliable visual words. They were chosen based on the results of *Word Content* probing task by selecting best and worst-performing clusters, as well as the ones with the largest performance difference between considered self-supervised models.

Based on the results of the user studies, we select the most representative visual words for each of the six features: brightness, color, texture, lines, shape, and form. Those words are then used to obtain detailed results of the *Word Content* probing task presented in Table 3.

VI. RESULTS AND DISCUSSION

As we show in Table 2, all self-supervised representations retain information about semantic knowledge, complexity, and image consistency. However, SimCLR v2 surpasses other methods in all probing task except CB color. Moreover, the performance on probing tasks does not correlate with the accuracy of the target task. In the following, we analyze those aspects in greater detail.

A. SELF-SUPERVISED REPRESENTATIONS CONTAIN SEMANTIC KNOWLEDGE WHICH DOES NOT CORRELATE WITH THE TARGET TASK

As reported in Table 2, the AUC scores for *Word Content* probing task vary from 0.793 for MoCo v1 to 0.811 for SimCLR v2. This shows that considered self-supervised methods can predict which visual words are present in the image, i.e., they code the semantic knowledge in the generated representations.

Surprisingly, although the examined self-supervised methods have diverse target task accuracy, they all have a similar level of semantic knowledge. For instance, MoCo v1 obtains the worst target task accuracy (60.6%), but its results for the WC probing task are on par with more accurate self-supervised methods. Moreover, although SwAV has the highest accuracy on the target task, it does not provide the best performance in terms of semantic knowledge. This finding supports the conclusion from [31] that semantic knowledge only partially contributes to the target task accuracy.

B. CERTAIN TYPES OF VISUAL WORDS ARE REPRESENTED BETTER THAN THE OTHERS, DEPENDING ON THE METHOD

According to the results presented in Table 3 and Figure 9, self-supervised representations have more knowledge about visual words containing forms and lines than about those containing shapes and textures. This may indicate that the representations are lines- and form-biased, which sheds new light on this problem, considering that according to [31], self-supervised representations are texture-biased. Moreover, the information encoded by various self-supervised methods differs. It is especially visible for brightness and color, where the MoCo v1 works significantly worse than the remaining methods. We assume that it is caused by the lack of projection head in the former, which is important due to the loss of information induced by the contrastive loss [33].

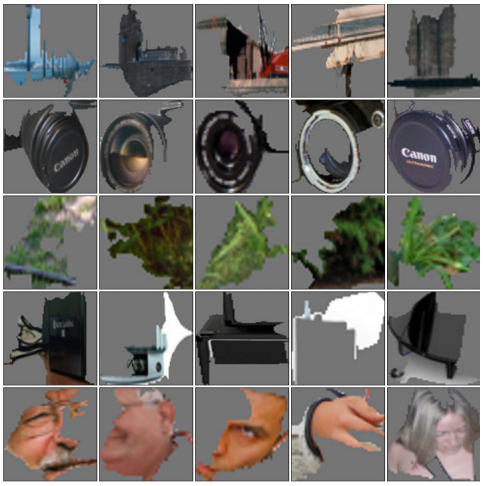
C. THE SAME VISUAL WORD IN A PAIR OF IMAGES USUALLY BRINGS THEIR REPRESENTATIONS CLOSER

The results of the MWC probing task presented in Table 4 show that the same visual word in a pair of images usually brings their representations closer. This is true for almost all visual words (45 out of 50), and especially for those presented in Figure 10a that contain complicated forms and lines or green areas. The remaining five visual words, usually corresponding to fine textures (see Figure 10c), are neutral or pushing representations away.

Interestingly, SwAV differs from the other methods in the case of lines and shapes, and BYOL differs in the case of shape. As in both cases, the presence of those features usually does not bring representations learned by those methods closer together. Differences in the training procedures of these methods might partially explain these results. SwAV and BYOL are trained without negative pairs, whereas

TABLE 2. AUC score for all our probing tasks (WC, MWC, SL, CB, and SOMO) and accuracy on the linear evaluation (Target) for the considered self-supervised methods.

	Target	Probing tasks (ours)						
		WC	MWC	SL	CB shape	CB color	SOMO far	SOMO close
MoCo v1	0.606	0.793	0.763	0.771	0.797	0.872	0.850	0.830
SimCLR v2	0.717	**0.811**	**0.777**	**0.775**	**0.850**	0.876	**0.878**	**0.857**
BYOL	0.723	0.803	0.775	0.770	0.844	**0.893**	0.845	0.817
SwAV	**0.753**	0.802	0.776	0.769	0.842	0.879	0.856	0.839

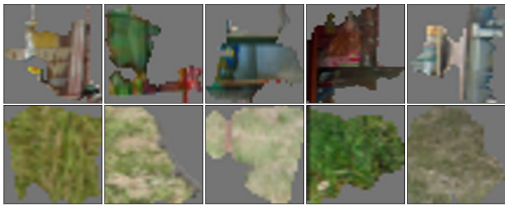


(a) Best predicted visual words.

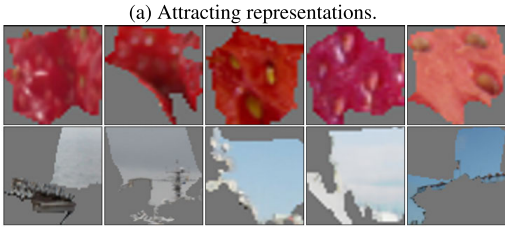
(b) Worst predicted visual words.

FIGURE 9. Visualization of the best (a) and the worst (b) predicted visual words according to the results of the WC probing task. It supports the results from Table 3 that self-supervised representations contain more information about lines and forms than textures.

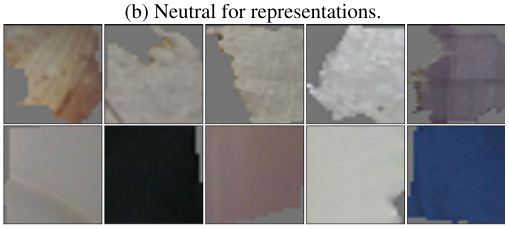
SimCLR v2 and MoCo v1 use both positive and negative pairs during the training. Therefore, we hypothesize that this may cause differences in the MWC probing task result.



(a) Attracting representations.



(b) Neutral for representations.

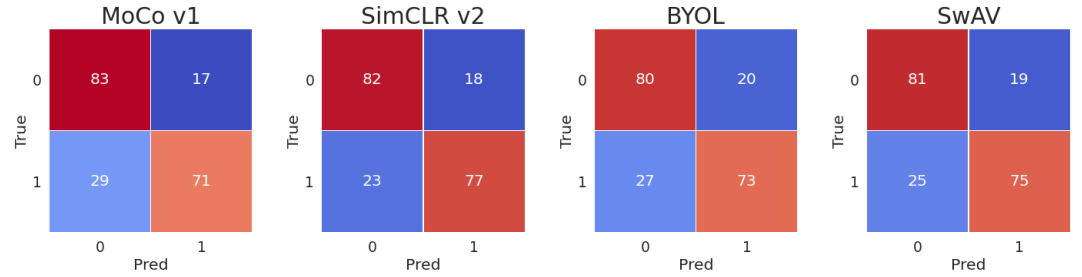


(c) Pushing representations away.

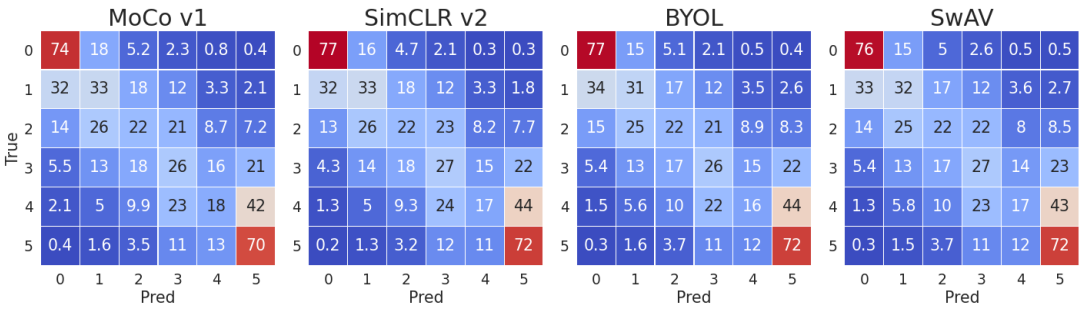
FIGURE 10. Sample visual words that attract (a), are neutral (b), or push away (c) the representations of two images. One can observe that the visual words that attract representations include words with complicated forms or green areas. On the other hand, visual words that push representations away contain fine textures.

D. SELF-SUPERVISED REPRESENTATIONS CONTAIN INFORMATION ABOUT SEMANTIC COMPLEXITY THAT DIFFERS BETWEEN METHODS

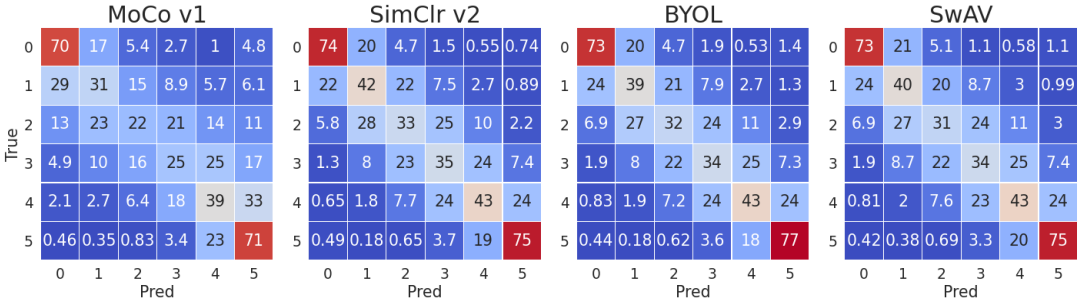
Based on the results in Table 2 and Figure 11, we observe that considered self-supervised methods code the level of semantic complexity, as they all obtain approximately 0.77 AUC for *Sentence Length*, and even higher AUCs are observed for CB shape and color (from 0.797 to 0.893 AUC). Moreover, when it comes to recognizing variance in superpixel color, BYOL works best, in contrast to all other probing tasks, where SimCLR v2 has the highest AUC. The potential reason for this behavior is the fact that a positive pair with similar color histograms provide more information in BYOL than in



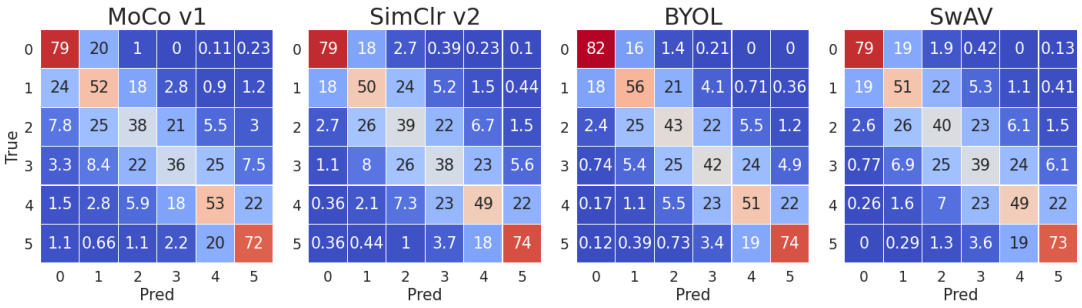
(a) SOMO.



(b) Sentence length.



(c) *Character Bin* shape.



(d) *Character Bin* color.

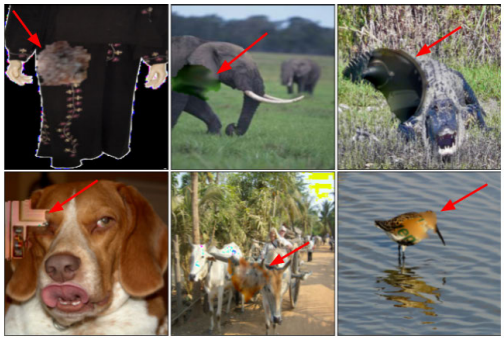
FIGURE 11. Confusion matrices for ***Sentence Length*** and ***Character Bin*** probings (results in %). The results indicate that the ability of self-supervised representations to retain information about complexity differs depending on the level of image complexity. Moreover, even though the final AUC of SL and CB for self-supervised methods are similar, their confusion matrices differ.

TABLE 3. Biases of the representations measured by WC probing tasks. The colors indicate a higher (orange) or lower (blue) AUC score compared to the overall performance for all visual words.

	all visual words	Types of visual words					
		brightness	color	texture	lines	shape	form
MoCo v1	0.793	0.777	0.769	0.785	0.847	0.784	0.836
SimCLR v2	0.811	0.831	0.832	0.804	0.852	0.810	0.854
BYOL	0.803	0.809	0.807	0.795	0.852	0.798	0.848
SwAV	0.802	0.819	0.820	0.796	0.851	0.802	0.849

TABLE 4. The attraction coefficient of MWC probing task. The attractions of representations containing the same visual word are marked in orange.

	Types of visual words					
	brightness	color	texture	lines	shape	form
MoCo v1	0.872	0.875	0.937	1.839	0.677	1.928
SimCLR v2	0.409	0.424	0.442	0.803	0.500	0.593
BYOL	0.396	0.379	0.502	0.336	−0.007	0.566
SwAV	0.382	0.445	0.150	−0.119	−0.118	0.248



(a) Superpixel substitution correctly classified.

(b) Superpixel substitution incorrectly classified.

FIGURE 12. Sample images from the SOMO probing task. Images for which the superpixel substitution was correctly classified based on the representations by all methods (a) and by none of them (b). One can observe that probing struggles with more subtle changes, which are still visible to the human eye. Notice that red arrows indicate alterations to the image.

SimCLR, as presented in Section V of [20]. Therefore, BYOL puts more attention on the color characteristic.

E. SELF-SUPERVISED REPRESENTATIONS CONTAIN INFORMATION ABOUT SEMANTIC CONSISTENCY THAT DIFFERS BETWEEN METHODS

The results of the SOMO probing task in Table 2 and Figure 11 show that self-supervised representations reflect

(a) Correctly classified only by MoCo v1.

(b) Correctly classified only by SimCLR v2.

(c) Correctly classified only by BYOL.

(d) Correctly classified only by SwAV.

FIGURE 13. Sample images from SOMO probing task, correctly classified when embedded by MoCov1 (a), SimCLR v2 (b), BYOL (c), or SwAV (d) only. One can observe no clear differences between the types of inconsistencies classified correctly and incorrectly by the particular methods. Notice that red arrows indicate alterations to the image.

changes in the center of the image. However, as presented in Figure 12, the probing classifier struggles with more subtle changes, which are still visible to the human eye. Moreover, SimCLR v2 has the highest ability to recognize altered images, but surprisingly, BYOL has the lowest performance. However, as shown in Figure 12 and 13, there are no visible reasons for this result. Overall, our results are in line with [32], which claims that self-supervised methods improve out-of-distribution detection. However, they contradict our previous results [51], where the replaced superpixel is selected entirely randomly (without center bias). Nevertheless, we decided to change replacement to center-biased

TABLE 5. Differences between architecture and training of the considered self-supervised methods.

		MoCo v1	SimCLR v2	BYOL	SwAV
Architecture	InfoNCE	yes	yes	no	no
	Positive pairs	yes	yes	yes	yes
	Negative pairs	yes, minibatches queue	yes, large batches	no	no
	Online to target network	copied with momentum	same	copied with momentum	same
	Size of patches	224x224	224x224	224x224	224x114 and 96x96
	Augmentations	resize, crop, color jittering, horizontal flip, grayscale conv.	crop, resize, horizontal flip, color distortion, grayscale conv., Gaussian blur, solarization	like in SimCLR v2	two types of crops, small and original, the rest like in SimCLR v2
	Projection	no	yes	yes	yes
Training	epochs	200	600	300	800
	Batch size	256	2048	4096	4096
	Time of training	53	170	not mentioned	49

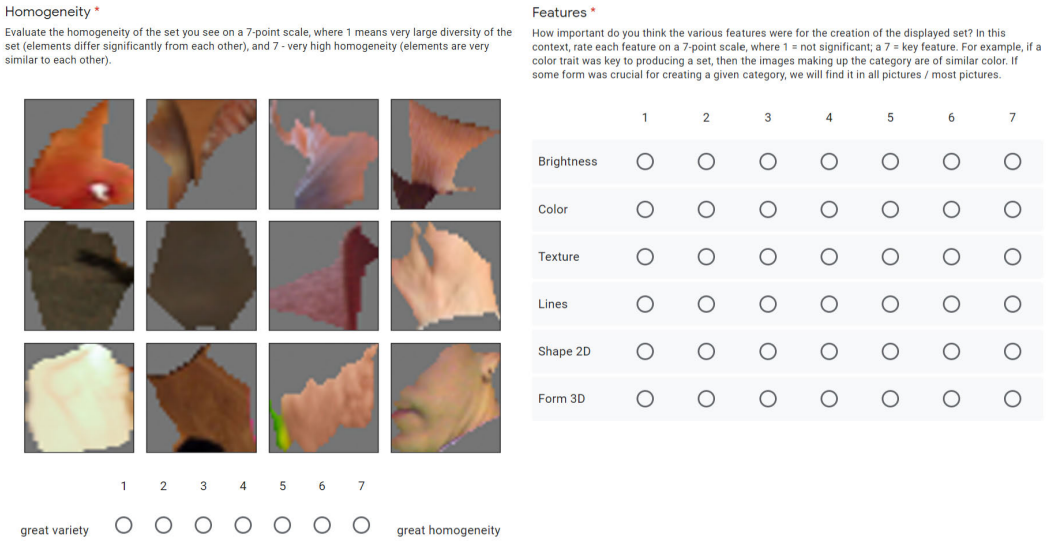


FIGURE 14. Sample question from our user study that allows in-depth analysis of the Word Content probing task using Marr's computational theory of vision.

in this work because they better correspond to semantic inconsistency.

F. THE ABILITY TO DISTINGUISH ALTERED IMAGES DEPENDS ON HOW OFTEN THE REMOVED VISUAL WORD CO-OCCURS WITH THE REPLACEMENT

As presented in Table 2, SOMO far has higher performance than SOMO close. It is expected because recognizing alterations obtained by replacing a visual word with a non-fitting one is simpler. However, this difference in performance for all self-supervised representations leads us to believe that there is a family of alterations that might not be reflected well enough in a self-supervised representation. Hence, considering that even minor alterations might lead to a change in the prediction [60], this disability might pose a risk to the stability of the classification results.

VII. CONCLUSION

In this work, we introduce a novel visual probing framework that analyzes the information stored in self-supervised image representations. It is inspired by probing tasks employed in NLP and requires similar taxonomy. Hence, we propose a set of intuitive mappings between visual and textual modalities to construct visual sentences, words, and characters. Moreover, we provide a cognitive visual systematic that identifies a visual word with structural features from Marr's computational theory [36] and provide the meaning of the words.

Our cognitive framework reveals insights into high-level concepts that the model has learned. Such insights can be applied to promote the safer use of self-supervised learning, being aware of the biases encoded in the representations. The results of the provided experiments confirm the

effectiveness and applicability of this framework in understanding self-supervised representations. We verify that the representations contain information about semantic knowledge, complexity, and consistency of the images. Moreover, a detailed analysis of each probing task reveals differences in the representations encoded by various methods, providing complementary knowledge to the accuracy of linear evaluation.

Our framework goes beyond per-sample explanations to identify higher-level human-understandable visual concepts that apply across the entire dataset. The existing work, the closest to ours, is a work [19] in which high-order concepts are automatically determined for images from each class. We build upon this work and propose a new approach using probing tasks. The advantage of our method is measurability, which enables us to compare to what extent individual self-supervised models encode information about concepts in their representations.

We note a couple of limitations of our framework. Our framework only applies to concepts in the form of groups of pixels. This assumption gives us plenty of insight into the model, but more complex and abstract concepts might be difficult to be noticed. In addition, the success of our approach depends on the quality of the generated labels for probing tasks. In our work, we presented five probing tasks inspired by NLP. However, in future work, one can consider a more generic approach to creating desirable probing tasks. One possible way to do this is to use a model that generates images based on a given text description, e.g., DALL-E 2 [62]. Thanks to this, we can create training pairs of image and text, generate probing labels based on text and use a probing classifier to image representation. In this case, creating new probing tasks would be equivalent to formulating appropriate queries for the image-generating model. These queries describe the human-understandable concept influencing the model's trait we want to investigate.

Potentially, our method may have a wider application, not only for self-supervised learning but for any learning methods for which individual layers of representation can be explained using our cognitive framework. The advantage of our framework is that it is generic. For example, using a different segmentation algorithm will lead to a different visual vocabulary. Therefore, conducting additional user studies to evaluate visual words generated using different methods and parameters would be worthwhile.

Finally, we show that the relations between language and vision can serve as an effective yet intuitive tool for explainable AI. Hence, we believe that our work will open new research directions in this domain.

ACKNOWLEDGMENT

The authors have applied a CC BY license to any Author Accepted Manuscript (AAM) version arising from this submission, in accordance with the grants' open access conditions.

REFERENCES

[1] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. Hinton, "Big self-supervised models are strong semi-supervised learners," 2020, *arXiv:2006.10029*.

[2] K. Krasnowska-Kieraś and A. Wróblewska, "Empirical linguistic study of sentence embeddings," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 5729–5739.

[3] J. Hewitt and C. Manning, "A structural probe for finding syntax in word representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Language Technol.*, vol. 1, Jun. 2019, pp. 4129–4138.

[4] N. Kim, R. Patel, A. Poliak, P. Xia, A. Wang, T. McCoy, I. Tenney, A. Ross, T. Linzen, B. Van Durme, S. R. Bowman, and E. Pavlick, "Probing what different NLP tasks teach machines about function word comprehension," in *Proc. 8th Joint Conf. Lexical Comput. Semantics (SEM)*, 2019, pp. 235–249.

[5] J. Hewitt and P. Liang, "Designing and interpreting probes with control tasks," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 1–11.

[6] I. Vulić, E. M. Ponti, R. Litschko, G. Glavaš, and A. Korhonen, "Probing pretrained language models for lexical semantics," 2020, *arXiv:2010.05731*.

[7] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[8] J. Sivic and A. Zisserman, "Video Google: Efficient visual search of videos," in *Toward Category-Level Object Recognition*. Berlin, Germany: Springer, 2006, pp. 127–144.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1097–1105.

[10] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2425–2433.

[11] Q.-S. Zhang and S.-C. Zhu, "Visual interpretability for deep learning: A survey," *Frontiers Inf. Technol. Electron. Eng.*, vol. 19, no. 1, pp. 27–39, Jan. 2018.

[12] G. Alain and Y. Bengio, "Understanding intermediate layers using linear classifier probes," 2016, *arXiv:1610.01644*.

[13] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," 2018, *arXiv:1810.03292*.

[14] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2013, *arXiv:1312.6034*.

[15] Z. Huang and Y. Li, "Interpretable and accurate fine-grained recognition via region grouping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8662–8672.

[16] S. Kumar and P. Talukdar, "NILE: Natural language inference with faithful natural language explanations," 2020, *arXiv:2005.12116*.

[17] M. Eichler, G. G. Şahin, and I. Gurevych, "LINSPECTOR web: A multilingual probing suite for word representations," 2019, *arXiv:1907.11438*.

[18] A. Conneau, G. Kruszewski, G. Lample, L. Barrault, and M. Baroni, "What you can cram into a single vector: Probing sentence embeddings for linguistic properties," 2018, *arXiv:1805.01070*.

[19] A. Ghorbani, J. Wexler, J. Zou, and B. Kim, "Towards automatic concept-based explanations," 2019, *arXiv:1902.03129*.

[20] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent: A new approach to self-supervised learning," 2020, *arXiv:2006.07733*.

[21] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," 2020, *arXiv:2006.09882*.

[22] L. R. Sipe, "How picture books work: A semiotically framed theory of text-picture relationships," *Children's Literature in Educ.*, vol. 29, pp. 97–108, Jun. 1998.

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[24] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[25] T. Leung and J. Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons," *Int. J. Comput. Vis.*, vol. 43, pp. 29–44, Jun. 2001.
[26] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 2668–2677.
[27] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9729–9738.
[28] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.
[29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[30] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2006, pp. 1735–1742.
[31] R. Geirhos, K. Narayanappa, B. Mitzkus, M. Bethge, F. A. Wichmann, and W. Brendel, "On the surprising similarities between supervised and self-supervised models," 2020, *arXiv:2010.08377*.
[32] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song, "Using self-supervised learning can improve model robustness and uncertainty," 2019, *arXiv:1906.12340*.
[33] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.
[34] C.-E. Guo, S.-C. Zhu, and Y. N. Wu, "Primal sketch: Integrating structure and texture," *Comput. Vis. Image Understand.*, vol. 106, pp. 5–19, Apr. 2007.
[35] P. Kitcher, "Marr's computational theory of vision," *Philosophy Sci.*, vol. 55, no. 1, pp. 1–24, Apr. 1988.
[36] D. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York, NY, USA: Henry Holt, 1982.
[37] D. Medin, "Concepts and conceptual structure," *Amer. Psychologist*, vol. 44, pp. 81–1469, Dec. 1989.
[38] M. J. Morgan, "Features and the 'primal sketch,'" *Vis. Res.*, vol. 51, no. 7, pp. 738–753, Apr. 2011.
[39] M. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you': Explaining the predictions of any classifier," 2016, *arXiv:1602.04938*.
[40] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," 2014, *arXiv:1412.0035*.
[41] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," 2017, *arXiv:1710.11063*.
[42] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network dissection: Quantifying interpretability of deep visual representations," 2017, *arXiv:1704.05796*.
[43] B. Zhou, D. Bau, A. Oliva, and A. Torralba, "Interpreting deep visual representations via network dissection," 2017, *arXiv:1711.05611*.
[44] C. Chen, O. Li, A. Barnett, J. Su, and C. Rudin, "This looks like that: Deep learning for interpretable image recognition," 2018, *arXiv:1806.10574*.
[45] R. Zhang, P. Isola, and A. Efros, "Colorful image colorization," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 649–666.
[46] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," 2018, *arXiv:1803.07728*.
[47] P. Goyal, Q. Duval, J. Reizenstein, M. Leavitt, M. Xu, B. Lefaudeux, M. Singh, V. Reis, M. Caron, P. Bojanowski, A. Joulin, and I. Misra. (2021). *VISSL*. [Online]. Available: https://github.com/facebookresearch/vissl
[48] Y. Belinkov and J. Glass, "Analysis methods in neural language processing: A survey," *Trans. Assoc. Comput. Linguistics*, vol. 7, pp. 49–72, Apr. 2019.
[49] A. Schick, M. Fischer, and R. Stiefelhagen, "Measuring and evaluating the compactness of superpixels," in *Proc. 21st Int. Conf. Pattern Recognit. (ICPR)*, 2012, pp. 930–934.
[50] W. Benesova and M. Kottman, "Fast superpixel segmentation using morphological processing," in *Proc. Int. Conf. Mach. Vis. Mach. Learn.*, 2014, pp. 1–9.
[51] D. Basaj, W. Oleszkiewicz, I. Sieradzki, M. Górszczak, B. Rychalska, T. Trzcinski, and B. Zieliński, "Explaining self-supervised image representations with visual probing," in *Proc. 13th Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 592–598.
[52] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, pp. 91–110 Nov. 2004.
[53] M. Calonder, V. Lepetit, M. Ozuysal, T. Trzcinski, C. Strecha, and P. Fua, "BRIEF: Computing a local binary descriptor very fast," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1281–1298, Jul. 2012.
[54] L. Sixt, M. Granz, and T. Landgraf, "When explanations lie: Why many modified BP attributions fail," 2019, *arXiv:1912.09818*.
[55] J. Bernard, M. Hutter, C. Ritter, M. Lehmann, M. Sedlmair, and M. Zeppelzauer, "Visual analysis of degree-of-interest functions to support selection strategies for instance labeling," in *Proc. EuroVA, Int. Workshop Vis. Anal.*, 2019.
[56] M. Ribeiro, "'Why should i trust you?': Explaining the predictions of any classifier," in *Proc. SIGKDD*, 2016, pp. 97–101.
[57] R. Salakhutdinov, "One-shot learning with a hierarchical nonparametric Bayesian model," in *Proc. ICML UTL Workshop*, 2012, pp. 1–13.
[58] D. Rymarczyk, L. Struski, J. Tabor, and B. Zielinski, "ProtoPShare: Prototypical parts sharing for similarity discovery in interpretable image classification," in *Proc. SIGKDD*, 2021, pp. 1420–1430.
[59] C. J. Cai, E. Reif, N. Hegde, J. Hipp, and B. Kim, "Human-centered tools for coping with imperfect algorithms during medical decision-making," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2019, pp. 1–14.
[60] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–11.
[61] H. Matthies and G. Strang, "The solution of nonlinear finite element equations," *Int. J. Numer. Methods Eng.*, vol. 14, no. 11, pp. 1613–1626, 1979.
[62] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with CLIP latents," 2022, *arXiv:2204.06125*.

WITOLD OLESZKIEWICZ received the M.Sc. degree in computer science from the Institute of Computer Science, Warsaw University of Technology, in 2017, where he is currently pursuing the Ph.D. degree with the Division of Artificial Intelligence, Institute of Computer Science. He is currently an Assistant with the Division of Artificial Intelligence, Institute of Computer Science, Warsaw University of Technology. His professional appointments include work with Samsung, in 2013, and Braster, from 2015 to 2018, where he worked on the use of machine learning in breast cancer detection. He was a Visiting Scholar at Stanford University, in 2018, where he worked on privacy-preserving generative models, and New York University, in 2019, where he worked on understanding the robustness of deep learning for breast cancer screening.

DOMINIKA BASAJ received the master's degree in quantitative methods in economics and information systems from the Warsaw School of Economics, in 2016. She was developing machine learning models in financial institutions. In 2019, she was a Visiting Researcher at the Nanyang University of Technology, where she worked on discourse-aware neural machine translation, and at the University of California at Davis, where she worked on the prediction of protein structure. She is currently a Senior AI Engineer with Toooploox. Her research interests include the interpretability and robustness of neural networks.

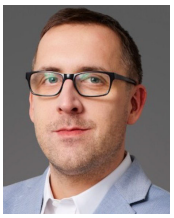
IGOR SIERADZKI received the M.Sc. degree in computer science on active learning in computer-aided drug design from Jagiellonian University, in 2016, where he is currently pursuing the Ph.D. degree with the Faculty of Mathematics and Computer Science, Institute of Computer Science and Computer Mathematics. He is currently an Assistant with the Faculty of Mathematics and Computer Science, Institute of Computer Science and Computer Mathematics, Jagiellonian University, Kraków, since 2019. Before the academic position, he worked with Applica.ai on the modern use of deep learning in natural language processing. His research internships include a stay at the University of Edinburgh, in 2015.

MICHAŁ GÓRSZCZAK recevied the B.Eng. degree in applied computer science from the University of Science and Technology, Kraków, in 2019. He is currently pursuing the master's degree with the Faculty of Mathematics and Computer Science, Jagiellonian University, Kraków.

BARBARA RYCHALSKA received the master's degree in computer science from the Warsaw University of Technology in 2016, also studied applied linguistics at the Warsaw University, where she is currently pursuing the Ph.D. degree with the Faculty of Mathematics and Information Science. She is currently an AI Research Scientist with Synerise, where she works on topics ranging from natural language processing to recommender systems. Previously, she worked at Samsung Research and Development Research Institute Warsaw and Findwise AB, as an AI Researcher. She was a Visiting Scientist at the Nanyang Technological University, Singapore, in 2019.

KORYNA LEWANDOWSKA received the M.A. degree in psychology and the Ph.D. degree in psychology on the influence of decision bias on visual recognition memory from Jagiellonian University, Kraków, in 2011 and 2019, respectively. She is currently an Assistant with the Department of Cognitive Neuroscience and Neuroergonomics, Faculty of Management and Social Communication, Institute of Applied Psychology, Jagiellonian University. She is also a Lecturer with the College of Economics and Computer Science. Her research interests include the realization of projects concerning issues from the fields of cognitive psychology, cognitive neuroscience, chronopsychology, and consumer neuroscience. She is a member of the Polish Association for Cognitive and Behavioral Therapy.

TOMASZ TRZCINSKI (Senior Member, IEEE) received the M.Sc. degree in research on information and communication technologies from the Universitat Politècnica de Catalunya, the M.Sc. degree in electronics engineering from the Politecnico di Torino, in 2010, the Ph.D. degree in computer vision from the École Polytechnique Fédérale de Lausanne, in 2014, and the D.Sc. degree (Habilitation) from the Warsaw University of Technology, in 2020. He has been an Assistant Professor with the Division of Computer Graphics, Institute of Computer Science, Warsaw University of Technology, since 2015. His professional appointments include work with Google, in 2013; Qualcomm Corporate Research and Development, in 2012; and Telefónica Research and Development, in 2010. He was a Visiting Scholar at Stanford University, in 2017, and Nanyang Technological University, in 2019. He is a Co-Organizer of warsaw.ai a member of Computer Vision Foundation, an Expert of the National Science Centre and Foundation for Polish Science, as well as a member of the Scientific Board for PLinML and Data Science Summit conferences. He is a Chief Scientist and a Partner at Tooploox, where he leads a team of machine learning researchers and engineers. He has co-founded Comixify, a technology startup focused on using machine learning algorithms for editing videos. He is currently an Associate Editor of IEEE Access and frequently serves as a reviewer in major computer vision conferences (CVPR, ICCV, ECCV, ACCV, BMVC, ICML, MICCAI) and international journals (IEEE Transactions on Pattern Analysis and Machine Intelligence, *IJCV*, *CVIU*, IEEE Transactions on Image Processing, IEEE Transactions on Multimedia).



BARTOSZ ZIELIŃSKI received the M.Sc. degree in computer science from Jagiellonian University, in 2007, and the Ph.D. degree in computer science with the Institute of Fundamental Technological Research, Polish Academy of Science, in 2012. He is currently an Assistant Professor with the Faculty of Mathematics and Computer Science, Institute of Computer Science and Computer Mathematics, Jagiellonian University, Kraków, since 2012. His professional appointments include work with Volantis Systems Ltd., in 2009, and Samsung, in 2018. He was a Visiting Scholar at the Vienna University of Technology, in 2015, and the Instituto Superior Técnico, Lisbon, in 2019. He is a Co-Organizer of the Cracow Cognitive Science Conference and Theoretical Foundations of Machine Learning. He is a Lead Data Scientist at Ardigen, where he leads a team of medical image analysis researchers and engineers. He frequently serves as a Reviewer in international journals on machine learning and medical image analysis (*AIR*, *CSBJ*, *CBM*, IEEE Transactions on Biomedical Engineering, *Trends in Microbiology*).

Siamese Generative Adversarial Privatizer for Biometric Data

Witold Oleszkiewicz[1(✉)], Peter Kairouz[2], Karol Piczak[1], Ram Rajagopal[2], and Tomasz Trzciński[1,3]

[1] Warsaw University of Technology, Warsaw, Poland
witold.oleszkiewicz@pw.edu.pl
[2] Stanford University, Stanford, USA
[3] Tooploox, Wrocław, Poland

Abstract. State-of-the-art machine learning algorithms can be fooled by carefully crafted adversarial examples. As such, adversarial examples present a concrete problem in AI safety. In this work we turn the tables and ask the following question: can we harness the power of adversarial examples to *prevent malicious adversaries from learning identifying information* from data while allowing non-malicious entities to *benefit from the utility* of the same data? For instance, can we use adversarial examples to anonymize biometric dataset of faces while retaining usefulness of this data for other purposes, such as emotion recognition? To address this question, we propose a simple yet effective method, called *Siamese Generative Adversarial Privatizer* (SGAP), that exploits the properties of a Siamese neural network to find discriminative features that convey identifying information. When coupled with a generative model, our approach is able to correctly locate and disguise identifying information, while minimally reducing the utility of the privatized dataset. Extensive evaluation on a biometric dataset of fingerprints and cartoon faces confirms usefulness of our simple yet effective method.

1 Introduction

Large-scale datasets enable researchers to design and apply state-of-the-art machine learning algorithms that can solve progressively challenging problems. Unfortunately, most organizations release datasets rather reluctantly due to the excessive amounts of sensitive information about participating individuals.

Ensuring the privacy of subjects is done by removing all personally identifiable information (e.g. names or birthdates) – this process, however, is not foolproof. Correlation and linkage attacks [15,25] often identify an individual by combining anonymized data with personal information obtained from other sources. Several such cases have been presented in the past, e.g. deanonymization of users' viewing history that was published in the Netflix Prize competition [25], identifying subjects in medical studies based on fMRI imaging data [9], and linking DNA profiles of anonymized participants with data from the Personal Genome Project [32].

C. V. Jawahar et al. (Eds.): ACCV 2018, LNCS 11365, pp. 482–497, 2019.
https://doi.org/10.1007/978-3-030-20873-8_31

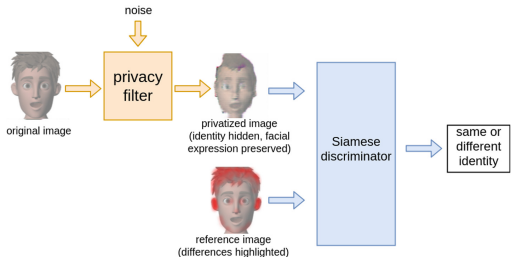


Fig. 1. Basic functionality of the proposed Siamese Generative Adversarial Privatizer: given an original face image, the privacy filter generates a privatized image. The original identity is hidden, at the same time other useful features, e.g. facial expression, are preserved. Siamese discriminator identifies the discriminative features of the images.

Typical approaches to countering the shortcomings of anonymization techniques leverage data randomization. While randomizing datasets with differential privacy [7] provides much stronger privacy guarantees, the utility of machine learning models trained on such randomized data is often significantly impaired [16,18,30]. We therefore believe that there is an ever increasing need for new privatization methods that preserve the value of the data while protecting the privacy of individuals.

The above privacy problem becomes critical when dealing with sensitive biometric and medical images. Several breakthrough applications of computer vision have been proposed in this domain: [12] used machine learning algorithms to parcellate human cerebral cortex, [29] utilized convolutional networks to detect arrhythmia, and [8] used machine learning to realize a precision medicine system. These applications, though critical for the advancement of the domain, rely on the access to highly sensitive data. This calls for novel privatization schemes that allow for the publication of images containing medical and biometric information without sacrificing the utility of the applications discussed above.

1.1 Our Contributions

In this work, we take a new approach towards enabling private data publishing. Instead of adopting worst case, context-free notions of statistical data privacy (such as differential privacy), we present a novel framework that allows the publisher to privatize images in a context-aware manner (Fig. 1). Our framework builds up on the recent work [17] where they propose a Generative Adversarial Privacy (GAP) method that casts the privatization as a constrained minimax game between a privatizer and an adversary that tries to infer private data. The approach we propose here is focused on biometric images and exploits a Siamese neural network architecture to identify image parts that bear the highest discriminative power and perturb them to enforce privatization. Contrary to other works that quantify privacy in a subjective manner using user surveys [28], we define here empirical conditions our privatizer needs to fulfill and propose

metrics that allow to evaluate the privacy-utility trade-off we aim to explore. Finally, we present the results of our experiments on datasets of fingerprints and cartoon faces. Our results show that the proposed framework prevents an attacker from re-identifying privatized data while leaving other important image features intact. We call this approach *Siamese Generative Adversarial Privatizer* (SGAP).

To summarize our contributions are twofold:

- a novel *privatization method* that uses a Siamese architecture to identify identity-discriminative image parts and perturbates them to protect privacy, while preserving the utility of the resulting data for other machine learning tasks, and
- an empirical data-driven *privacy metric* (c.f. Sect. 4.2) based on mutual information that allows to quantify the privatization effects on biometric images.

1.2 Paper Outline

The remainder of this paper is organized as follows. In Sect. 2, we provide a brief survey of recent relevant works. In Sect. 3, we present the architectural details of our SGAP model. The main results of our paper are presented in Sect. 4. We conclude our paper in Sect. 5.

2 Related Work

Privatization of data has been an active area of research with multiple works touching on this subject [1,16,18,30]. Our approach extends the concept of context-independent data privatization by incorporating context-dependent information as an input to the privatization algorithm. More precisely, it identifies the discriminative characteristics of the data and distorts them to enforce privacy. Although standard methods of protecting privacy based on erasing personal information have been widely used, correlation and linkage attacks allow to re-identify the users, even when explicitly identifying information is not present in the released datasets [25].

Those kinds of attacks pose an even greater threat to individual privacy when used against publicly available medical databases [14]. [15] show that using publicly available genotype-phenotype correlations, an attacker can statistically relate genotype to phenotype and therefore re-identify individuals. Publicly available profiles in the Personal Genome Project can be linked with names by using demographic data [32]. Also, when considering fRMI imaging data, individual variability across individuals is both robust and reliable, thus can be used to identify single subjects [9].

Although numerous works are focused on finding discriminative patterns within the data [10,34], we use a Siamese neural network architecture [4] since it allows us to learn a discriminant data embedding in an end-to-end fashion. Contrary to the typical goal of a Siamese architecture, i.e. learning similarity,

we use it to identify discriminant parts of a pair of images and alter those parts with minimal impact on other useful features. When both examples come from the same individual, this setup allows us to learn a perturbation that carefully disguises the individual's identity, hence protecting their privacy.

One can consider the problem of data anonymization to be conceptually similar to the idea of adversarial examples in neural network architectures [3,19–21,33]. In the case of adversarial examples, the adversary wants to trick the neural network into misclassifying a slightly perturbed input of a given class. Similarly, our goal is to modify the data points in such a way that the identity of the individual corresponding to the data cannot be correctly classified. The most relevant work is [20], where they use a Generative Adversarial Network (GAN) [13] framework to create adversarial examples and use them in training to increase the robustness of the classifier.

Similar to us, [28] analyses the trade-off between data privacy and utility. In their work, however, privacy and utility metrics are defined based on a user-study, where the users were asked to assess the usefulness of the anonymized images in the context of social media distribution. The privacy, on the other hand, was measured by first enlisting a number of attributes linked to privacy (*e.g.* passport number or registration plates) and then asking the users to validate if a given privacy attribute is visible in the photo or not. We argue that this way of measuring both privacy and utility is limited to a very specific subset of applications. In our work we propose fundamentally different metrics for both privacy and utility that have backing in information theory and machine learning.

Another relevant and recent works [33], [5] address the privatization problem using a generative adversarial approach while providing theoretical privacy-utility trade-offs. The work of [5], which is the most similar to our work, proposes an architecture combining Variational Autoencoder (VAE) and GAN to create an identity-invariant representation of a face image. Their approach differs from ours as they use an additional discriminator, which explicitly controls which useful features of the images are to be preserved, whereas in our approach the model has no information about other features of the images, except that it knows whether a pair of images belongs to the same person or different people. This is a significant contribution because in practice, one cannot expect to know all potential applications of the privatized images. Therefore our approach proves to be more robust towards real-life applications.

[27] presents a similar game-theoretic perspective on image anonymization. However, the difference is that it focuses on adversarial image perturbations (carefully crafted perturbations invisible to human), while our privatizer introduces structural changes to the image. In [31], a head inpainting obfuscation technique is proposed by generating a realistic head inpainting using facial landmarks. On contrary, our goal is to hide the identity of a person without knowing which part of the image is responsible for identity. Thanks to this, our framework is more universal and has a much wider field of application, not only to hide face identity, but also hide identity in cases where there is no prior knowl-

edge of which part of the image should be obfuscated. [23,24] are relevant to our work and deal with a problem similar to ours. However, the formulation of the problem is different from ours. [23,24] transform an input face image in a way such that the transformed image can be successfully used for face recognition (so the identity is preserved) but not for gender classification. Our goal is the opposite, we want to hide identity while maintaining as much other features as possible, without explicitly modeling the non-malicious classification tasks. Another difference is that our model requires only identity labels. The architecture of the models presented in [23] and our work are similar, however we use Siamese discriminator what makes our approach advantageous when applied to large datasets with thousands or even millions of people, since this architecture reduces the output of the discriminator to a binary output rather than create a long list of individual class predictions.

3 Method

The goal of our approach is to develop a privatizer that converts an input image into its privatized version in such a way that: (1) the privacy of the subject is preserved by making sure that the identifying features are hidden, (2) the utility of the original image is maintained by preserving the non-identifying features that are vital for other machine learning tasks, and (3) the privacy-utility trade-off can be adjusted.

3.1 Proposed Approach

To enforce the above conditions, we will use a custom neural network architecture, dubbed *Siamese Generative Adversarial Privatizer*, that consists of two tightly coupled models: a generator $G(\theta_g)$ and a discriminator $D(\theta_d)$. This coupling is inspired by Generative Adversarial Networks (GANs) [13]. Two neural networks compete with each other: the discriminator tries to predict the identity of the person in the image, while the generator tries to generate such an image which fools the discriminator and thus hides the identity of the person.

We use a Siamese architecture [4] for the discriminator. This allows us to extract discriminative and identifying features from images. More importantly, this architecture reduces the output of the discriminator to a single value (from 0 to 1) rather than create a long list of individual class predictions, an approach which would be prohibitive when applied to large datasets with thousands or even millions of people. In this case, we use pairs of images (instead of single images) to train the neural network, and the goal of the Siamese discriminator is to classify whether the two images belong to the same person or to different people.

Furthermore, the above problem is subjected to a distortion constraint, which ensures that the privatized images are not too different from the original images.

We did not use L_2 since it is sensitive to small changes (e.g. shift, rotation, etc.) which do not significantly affect the content of the image. Instead

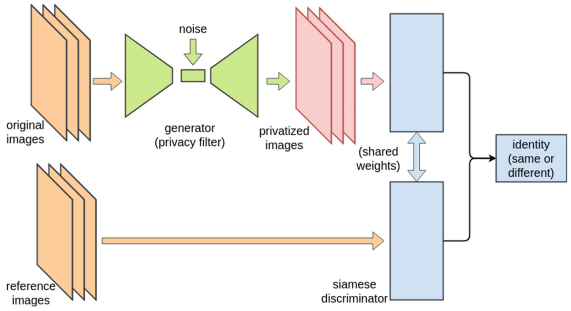


Fig. 2. Overview of our Siamese Generative Adversarial Privatizer model. The generator acts as a privacy filter, which hides the identity of the person in the original images. The Siamese discriminator recognizes whether the person in the privatized image is the same person as in the reference image.

we chose SSIM (structural similarity index) [35] which is sensitive to the structural changes of images, not pixel-by-pixel differences like L_2 [36]. We enforce a constraint on SSIM which allows us to control the level of distortion added to protect identity, and thus ensure that the quality of privatized images is not substantially degraded. The architecture overview can be seen in Fig. 2.

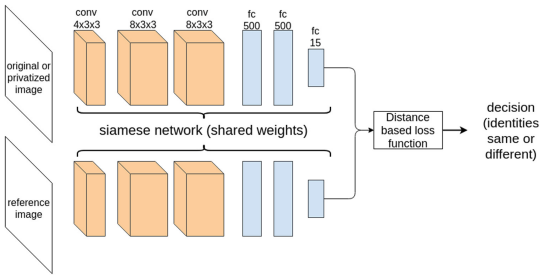


Fig. 3. Discriminator's architecture. We use a Siamese neural network to verify the identities of people in the images. The discriminator classifies whether a pair of images belongs to the same person or to different people. We get the output from the range between 0 and 1 applying distance-based loss function to the output of the last fully connected layer of the Siamese discriminator.

3.2 Architecture

Our discriminator is a Siamese convolutional neural network, which consists of two identical branches with shared weights, as shown in Fig. 3. Each branch consists of 3 blocks of the following form: (1) Convolutional layer (mask 3×3, stride $= 1$, padding $= 0$), (2) Leaky rectified linear unit ($\alpha = 0.1$), (3) Batch normalization, (4) Dropout ($p = 0.2$). The blocks are followed by 2 dense layers

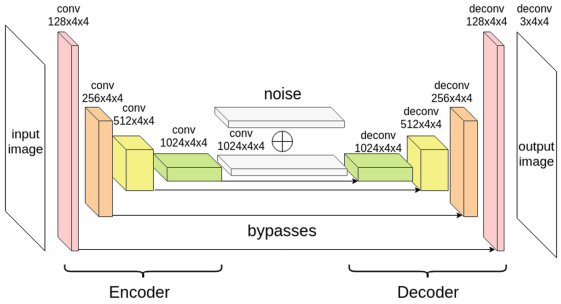


Fig. 4. Generator's architecture. We use Variational Autoencoder-like architecture to generate a privatized image in a context-aware manner based on the original image. At the bottleneck of the generator we get a compressed representation of the image without identity features, and thanks to the bypasses between the layers we preserve other useful features of the original image.

(500 neurons, leaky rectified linear unit, $\alpha = 0.1$) and an output layer (15 neurons). A discriminator network converts two input images to two output representations (embeddings) $D(\boldsymbol{X}_1, \boldsymbol{X}_2) \rightarrow (\boldsymbol{o}_1, \boldsymbol{o}_2)$.

The generator network, as presented in Fig. 4, consists of two parts: the encoding part and the decoding part. The encoder follows the typical architecture of a convolutional neural network. It consists of 5 blocks of the following form: (1) Convolutional layer (mask 4×4, stride $= 2$, padding $= 1$), (2) Leaky rectified linear unit ($\alpha = 0.1$), (3) Batch normalization. At each downsampling step we double the number of feature channels.

The decoder consists of 5 blocks of the following form: (1) Transpose convolutional layer (mask 4×4, stride $= 2$, padding $= 1$), (2) Leaky rectified linear unit ($\alpha = 0.1$), (3) Batch normalization, (4) Dropout ($p = 0.5$). At each upsampling step we halve the number of feature channels. Also we concatenate the feature maps of the decoder part with the corresponding feature map from the encoder part (these are bypasses). Last deconvolutional layer is followed by a hyperbolic tangent activation function.

A noise matrix $\mathbf{Z}$ is added to the bottleneck part of the generator, *i.e.* to the latent space variable representing input image in a low-dimensional space. We use a noise matrix instead of a vector, as we do not use a standard fully-connected layers in our generator and retain convolutional layers instead. The output of generator network is a privatized version of original image: $G(\mathbf{Z}, \boldsymbol{I}) \rightarrow \tilde{\boldsymbol{I}}$.

3.3 Training

When iterating over training dataset we get tuples: $(\boldsymbol{I}_i, \boldsymbol{I}_i^{'}, l_i)$, where $\boldsymbol{I}_i$ and $\boldsymbol{I}_i^{'}$ is a pair of images and l_i is a binary label where $l_i = 0$ if the images have the same identity and $l_i = 1$ for different identities. There are two types of pairs in the training set. Firstly, when the generator is turned off, $\boldsymbol{I}_i, \boldsymbol{I}_i^{'}$ are images from the original training set. Secondly, when the generator is turned on, $\tilde{\boldsymbol{I}}_i = G(\mathbf{Z}_i, \boldsymbol{I}_i)$

is the privatized version of the image $\boldsymbol{I}_i$ from the original training set. $\boldsymbol{I}_i^{'}$ is the reference image, also from the original training set. In both cases mentioned above we use stratified random sampling in order to balance two classes: $l = 0$ and $l = 1$.

The discriminator D takes a pair of images $\boldsymbol{I}, \boldsymbol{I}^{'}$ and outputs a probability that both images come from the same person, i.e. $l = 0$, based on a distance-based metric:

$$D(\boldsymbol{I}, \boldsymbol{I}^{'}) \rightarrow \frac{1 + e^{-m}}{1 + e^{d(\boldsymbol{o}, \boldsymbol{o}^{'})^2 - m}} = P(\boldsymbol{I} \overset{\text{sim.}}{\sim} \boldsymbol{I}^{'})$$

where m is a predefined margin and $d(\boldsymbol{o}, \boldsymbol{o}^{'})$ is an Euclidean distance between embeddings $\boldsymbol{o}$ and $\boldsymbol{o}^{'}$ in the last fully connected layer of the discriminator. Given this formulation of the discriminator we use a cross entropy loss for training:

$$\mathcal{L}(l, D(\boldsymbol{I}, \boldsymbol{I}^{'})) = -(1 - l) \log D(\boldsymbol{I}, \boldsymbol{I}^{'}) - l \log \Big(1 - D(\boldsymbol{I}, \boldsymbol{I}^{'})\Big)$$

We train our model similarly to GAN. When the generator training is frozen, our goal is to train the discriminator to recognize whether a pair of images belongs to the same person or to different people. When the generator is trained, there is a minmax game between the generator and the discriminator in which the generator is trying to fool the discriminator and generate an image that hides the identity of the subject. The training equation for our privatization task is:

$$\min_D \max_G \frac{1}{N} \sum_{i=0}^{N-1} \mathcal{L}(l_i, D(\boldsymbol{I}_i, \boldsymbol{I}_i^{'})) + \frac{1}{N} \sum_{i=0}^{N-1} \mathcal{L}(0, D(\boldsymbol{I}_i^{'}, G(\mathbf{Z}_i, \boldsymbol{I}_i)))$$

Furthermore, the above minimax optimization problem is subject to the following critical constraint: $\frac{1}{N} \sum_{i=0}^{N-1} d(\boldsymbol{I}_i, G(\mathbf{Z}_i, \boldsymbol{I}_i)) < \delta$, where $d(x, y)$ is a distortion metric and δ is a distortion threshold. The distortion constraint is used to limit all the other image changes except for hiding identity and therefore the utility of the images is preserved. We use Structural Similarity Index as the distortion metric. The above constraint can be incorporated into the main minimax objective function as follows:

$$\min_D \max_G \sum_{i=0}^{N-1} \mathcal{L}(l_i, D(\boldsymbol{I}_i, \boldsymbol{I}_i^{'})) + \sum_{i=0}^{N-1} \mathcal{L}(0, D(\boldsymbol{I}_i^{'}, G(\mathbf{Z}_i, \boldsymbol{I}_i))) + \lambda \sum_{i=0}^{N-1} d(\boldsymbol{I}_i, G(\mathbf{Z}_i, \boldsymbol{I}_i)) \tag{1}$$

Our Siamese Generative Adversarial Privatizer network is trained for 100 epochs using ADAM optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

4 Results

In this section, we present the results of evaluation of our method. We first present the datasets and evaluation metrics. Then we show qualitative and quantitative results of our evaluation that confirm usefulness of our approach in the context of data privatization.

4.1 Datasets

Fingerprints. To validate how well our method performs in terms of identity privatization, we evaluate it on a dataset of fingerprints. Although the main purpose of fingerprint datasets is to identify people and therefore their privatization may not be needed in their real-life use cases, we treat this dataset as our toy example and evaluate how well we can hide the privacy of the fingerprint owner. Since there exists a trade-off between the privatization and the utility of the resulting data, we refer to a proxy task of finger type classification to validate how useful our privatization method is. In other words, we try to classify the type of the finger (*e.g.* middle finger, index finger, ring finger) while gradually increasing the privacy of the dataset. Section 4.4 presents the results of this experiment.

We use NIST 8-Bit Gray Scale Images of Fingerprint Image Groups [26]. This database contains 4000 8-bit grayscale fingerprint images paired in couples. Each image is 512-by-512 pixels with 32 rows of white space at the bottom. We use only one image of each pair in our experiments. The dataset contains images for 2000 individuals. For each person there are two different fingerprint shots of the same finger (denoted as: f, s). Our method requires pairs of images as input. In each epoch the dataset is iterated over 4000 pairs of images.

For the first half of the pairs when index of a pair is $i < 2000$ we return a label $l = 0$ and a pair of images (f, s) belonging to the person with $ID = i$.

For the second half of the pairs when index $i >= 2000$ we return a label $l = 1$ and two images. First image is image f of person with $ID = i - 2000$. Second image is an image (f or s) of a different person (selected at random).

This way we have a 50%/50% split over similar/dissimilar pairs and the dataset loader is quasi-deterministic (for a given index i the first image is guaranteed to be constant).

Animated Faces. The second dataset that we use is FERG dataset [2]. FERG is a dataset of cartoon characters with annotated facial expressions. It contains 55769 annotated face images of six characters. The images for each identity are grouped into 7 types of facial expressions, such as: anger, disgust, fear, joy, neutral, sadness and surprise.

In each epoch the dataset is iterated over 10000 pairs of images. For the first half of the pairs we use different randomly selected images of the same person. In this case $l = 0$. For the second half of the pairs we use randomly selected images of different people. In this case $l = 1$. This way we have a 50%/50% split over similar/dissimilar pairs and the dataset loader is quasi-deterministic.

4.2 Evaluation Metrics

To evaluate the performance of our SGAP model and show that it learns privacy schemes that are capable of hiding biometric information even from computationally unbounded adversaries, we propose computing the mutual information between: (a) $X = (X_1, X_2)$ where X_1 is a privatized image and X_2 is an original

image, and (b) Y where $Y = 0$ when X_1 and X_2 belong to the same person and $Y = 1$ when they belong to different people. X_1 is privatized using the scheme that is learned in a data-driven fashion using SGAP. By Fano's inequality, if $I(X;Y)$ is low then Y cannot be learned from X reliably (even under computationally infinite adversaries) [6]. In other words, if $I(X;Y)$ is sufficiently small, there's no way we can reliably learn whether or not a privatized image belongs to the same person in another non-privatized image. This ensures that privacy is guaranteed in a strong sense.

In practice, we do not have access to the joint distribution $P(X,Y)$. We instead have access to a dataset of i.i.d observations $\mathcal{D} = \{(X_i, Y_i\}_{i=1}^n\}$. Here, the X_i's are computed after the SGAP training phase is over by applying the learned privacy scheme on a separate test set. We are thus interested in empirically estimating $I(X;Y)$ from $\mathcal{D}$. We will call this estimate "empirical mutual information" and denote it by $\hat{I}(X;Y)$. To compute $\hat{I}(X;Y)$, we can use the following formula:

$$\hat{I}(X;Y) = \hat{H}(X) - \hat{H}(X|Y)$$

where $\hat{H}(X)$ and $\hat{H}(X|Y)$ are the empirical entropies of X and X given Y. To compute these empirical entropies, we use the Kozachenko-Leonenko entropy estimator [11] which we briefly explain next. Letting $R_i = \min_{j,j\neq i} \|X_i - X_j\|$, for $j = 1, \ldots, n$, we get

$$\begin{aligned}\hat{H}(X) &= \frac{1}{n}\sum_{i=1}^{n} \log\left((n-1)R_i^d\right) + constant \\ &= \frac{d}{n}\sum_{i=1}^{n} \log R_i + \frac{1}{n}\sum_{i=1}^{n} \log(n-1) + constant\end{aligned}$$

where d is the dimension of X, i.e. $X_i \in \mathbb{R}^d$. Assuming we have a two-class problem ($Y = 0$ for same identities, $Y = 1$ for different identities), the conditional entropy is given by

$$\hat{H}(X|Y) = \hat{H}(X|Y=0)\hat{P}(Y=0) + \hat{H}(X|Y=1)\hat{P}(Y=1)$$

Notice that $\hat{P}(Y=0) = \frac{n_0}{n}$, $\hat{P}(Y=1) = \frac{n_1}{n}$, where n_0 and n_1 are the counts of samples with label Y equals 0 and 1 respectively. We divide sample X into two partitions. Letting $i_1, i_2, \ldots, i_{n_0}$ be the indices corresponding to $Y_i = 0$, we have a set $\mathcal{X}_0 = \{X_{i_1}, X_{i_2}, \ldots, X_{i_{n_0}}\}$. Automatically we have $i'_1, i'_2, \ldots, i'_{n_0}$, the indices of samples associated with $Y_i = 1$. Thus, we get $\mathcal{X}_1 = \{X_{i'_1}, X_{i'_2}, \ldots, X_{i'_{n_1}}\}$. Therefore we calculate the nearest neighbor distance for each sample within the particular set as follows:

$$R_{i_k} = \min_{l\neq k, l=1,\ldots,n_0} \|X_{i_k} - X_{i_l}\| \qquad R_{i'_k} = \min_{l\neq k, l=1,\ldots,n_1} \|X_{i'_k} - X_{i'_l}\|$$

$$\hat{H}(X|Y=0) = \frac{1}{n_0}\sum_{k=1}^{n_0}\log\left((n_0-1)R^d_{i_k}\right) + constant$$
$$\hat{H}(X|Y=1) = \frac{1}{n_1}\sum_{k=1}^{n_1}\log\left((n_1-1)R^d_{i'_k}\right) + constant$$

Then the empirical mutual information can be expressed as

$$\hat{I}(X,Y) = \hat{H}(X) - \left(\hat{H}(X|Y=0)\hat{P}(Y=0) + \hat{H}(X|Y=1)\hat{P}(Y=1)\right)$$
$$= \frac{1}{n}\sum_{i=1}^{n}\log\left((n-1)R^d_i\right) +$$
$$- \left(\left(\frac{1}{n_0}\sum_{k=1}^{n_0}\log\left((n_0-1)R^d_{i_k}\right)\right)\frac{n_0}{n} + \left(\frac{1}{n_1}\sum_{k=1}^{n_1}\log\left((n_1-1)R^d_{i'_k}\right)\right)\frac{n_1}{n}\right)$$

To estimate values of R_{i_k} and $R_{i'_k}$ we use L_2 norm between image pixels projected to a 3-dimensional space via t-SNE [22]. We reduce the dimensionality to increase the efficiency of computation, but our metric remains agnostic to image distance calculation and other methods can also be used here.

The second approach to quantify privacy is by measuring an identity misclassification rate. We measure what percentage of privatized images effectively fool our Siamese discriminator.

To quantify utility of privatized dataset we measure accuracy of the proxy classification task (finger type classification for fingerprint dataset and facial expression classification for faces dataset). More precisely, we evaluate how good in terms of accuracy a separate independent method can be trained for using a privatized dataset. We use fine-tuned ResNet architecture, pre-trained on ImageNet without freezing. In addition we split the dataset into training and validation. The accuracy is measured using k-fold validation ($k = 4$).

4.3 Qualitative Results

In this section, we present the qualitative results of our evaluation, demonstrating the ability of our network to increase the privacy of input data.

Figures 5 and 6 show sample results obtained as an output of our privatization. In Fig. 6 we see that the identities of people have been hidden, while other useful features, in this case facial expressions, have been preserved. Figures 7, 8 and 9 illustrate the trade-off between utility and privacy while tuning λ distortion metric constraint. We see that by tuning the λ parameter we can adjust the level of privacy and utility, finally finding the optimal value for both privacy and utility.

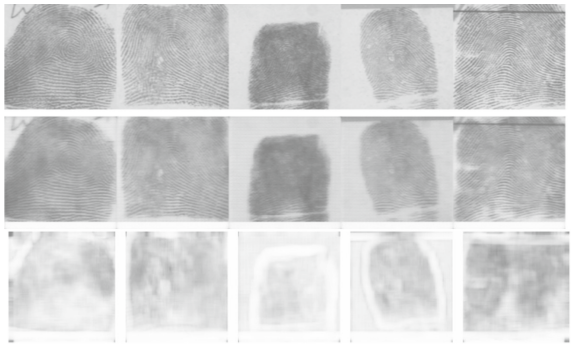


Fig. 5. A toy example of how our privatization method can hide identities of the fingerprint owners. Original fingerprints in the upper row. Fingerprints with added artifacts that fool identity discriminator in the middle row. Structural Similarity difference [35] of the original and privatized images is presented in the bottom row. Our Siamese Generative Adversarial Privatizer learns to locate discriminant image features, such as fingerprint minutiae, and substitutes them with anonymizing artifacts. Although in practice fingerprints are used for person identification, we validate if privatized images can be useful (*i.e.* if they can retain utility) for a proxy task of finger type classification. Since our method does not add noise arbitrarily across the image, but only focuses on hiding sensitive personal information, the resulting dataset can be published and used by machine learning for other tasks, e.g. finger type classification or skin disease detection.

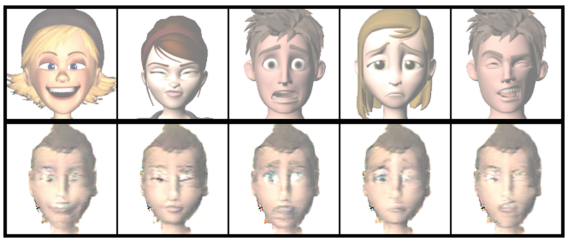


Fig. 6. Original cartoon faces in the upper row. Privatized versions of cartoon faces in the bottom row. Our Siamese Generative Adversarial Privatizer learns to hide the identity of the people, while other important image features, such as facial expression remain intact.

4.4 Quantitative Results

To obtain quantitative results we train our SGAP model with different values of maximal distortion constraint λ (see Eq. 1) in order to adjust the privacy level of the dataset. The goal of our generator is to add such noise to the latent space that privatized image fools the discriminator, which the discriminator in turn has to verify if the pair of images comes from the same person. After SGAP is trained, the generator part can be used to privatize datasets.

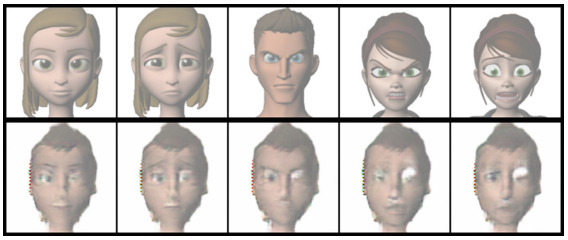


Fig. 7. Too much privacy, utility is not preserved. Original cartoon faces in the upper row. Privatized versions of cartoon faces in the bottom row. Our model has been tuned too much towards ensuring privacy, so that the utility of the images has not been preserved, facial expressions are hard to recognize.

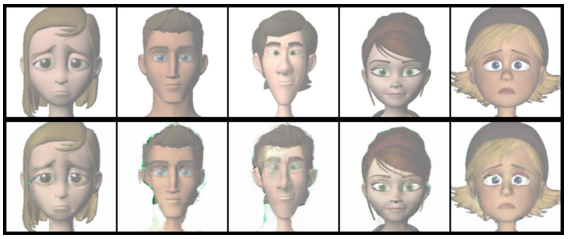


Fig. 8. Not enough privacy, utility is preserved. Original cartoon faces in the upper row. Privatized versions of cartoon faces in the bottom row. Our model has been tuned too much towards preserving utility, so that the identities of the people in the images are not hidden, only minor changes have been added to the images.

Fig. 9. Images in the first column are the original ones, next there are privatized images generated for different values of distortion constraint $\lambda \in \{10, 8, 6, 4, 2, 1, 0.7\}$. Original images of different identities collapse into an anonymous identity with the expression preserved from the original image.

To measure the utility of the privatized fingerprints dataset, we refer to a proxy task of finger type classification. Although in fingerprints are typically used to identify the identity of an individual, in our case we use the proposed privatization method to hide the identity and anonymize the dataset. The objective of this experiment is to evaluate how increasing data privacy effects the utility of the resulting dataset when used as training data for a machine learning algorithm. Hence, we use a proxy machine learning task, finger type classification. To measure the utility of the privatized cartoon faces dataset, we use

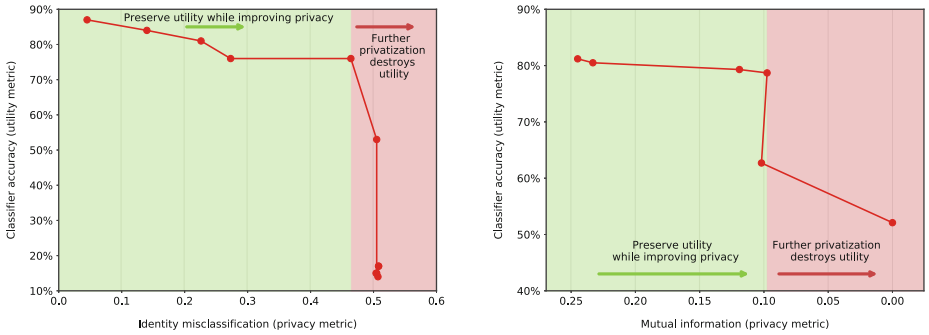


Fig. 10. Left: Graph of identity misclassification rate and the accuracy of a classifier trained with cartoon faces dataset privatized with different maximal constraint distortion thresholds. In green the region where the utility of dataset is preserved while the likelihood of classifying the privatized version of the image as belonging to a given person is reduced. This result proves that by using our privatization method we are able to significantly increase the privacy of the biometric dataset, while not reducing its utility for a task of facial expression classification. Right: Graph of mutual information estimation and the accuracy of a classifier trained with fingerprint dataset privatized with different maximal constraint distortion thresholds. In green the region where the utility of dataset is preserved while the likelihood of classifying the privatized version of the image as belonging to a given person is reduced. This result proves that by using our privatization method we are able to significantly increase the privacy of the biometric dataset, while not reducing its utility for a proxy task of finger type classification. (Color figure online)

facial expression classification as machine learning task. As a classifier, trained on privatized datasets, we use fine-tuned ResNet architecture, pre-trained on ImageNet without freezing. For each dataset generated using different maximal distortion constraint, we calculate classification accuracy and quantify the privacy by estimation of mutual information (fingerprint dataset) or using identity misclassification rate (faces dataset).

Figure 10 shows the results. In both cases we see a significant drop in privacy metric, while for the same range of parameters, the accuracy of the classifier remains stable, indicating that the utility of the dataset is not decreased.

5 Conclusions

We presented the *Siamese Generative Adversarial Privatizer* (SGAP) model for privacy-preserving of biometric data. We proposed a novel architecture combining Siamese neural network, autoencoder, and Generative Adversarial Network to create a context-aware privatizer. Experimental results on two public datasets demonstrate that our approach strikes a balance between privacy preservation and dataset utility.

Acknowledgment. The work was partially supported as RENOIR Project by the European Union Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 691152 (project RENOIR) and by Ministry of Science and Higher Education (Poland), grant No. W34/H2020/2016. We thank NVIDIA Corporation for donating Titan Xp GPU that was used for this research.

References

1. Abadi, M., et al.: On the protection of private information in machine learning systems: two recent approaches. CoRR abs/1708.08022 (2017)
2. Aneja, D., Colburn, A., Faigin, G., Shapiro, L., Mones, B.: Modeling stylized character expressions via deep learning. In: Lai, S.-H., Lepetit, V., Nishino, K., Sato, Y. (eds.) ACCV 2016. LNCS, vol. 10112, pp. 136–153. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-54184-6_9
3. Baluja, S., Fischer, I.: Adversarial transformation networks: learning to generate adversarial examples. CoRR abs/1703.09387 (2017)
4. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature verification using a "siamese" time delay neural network. In: Advances in Neural Information Processing Systems, vol. 6, pp. 737–744. Morgan-Kaufmann (1994)
5. Chen, J., Konrad, J., Ishwar, P.: VGAN-based image representation learning for privacy-preserving facial expression recognition. CoRR abs/1803.07100 (2018). http://arxiv.org/abs/1803.07100
6. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley Series in Telecommunications and Signal Processing. Wiley, New York (2006)
7. Dwork, C.: Differential privacy: a survey of results. In: International Conference on Theory and Applications of Models of Computation, pp. 1–19 (2008)
8. Famm, K., Litt, B., Tracey, K.J., Boyden, E.S., Slaoui, M.: Drug discovery: a jump-start for electroceuticals. Nature **496**(7444), 159–161 (2013)
9. Finn, E.S., et al.: Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. Nat. Neurosci. **18**(11), 1664–1671 (2015)
10. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Ann. Eugen. **7**(7), 179–188 (1936)
11. Fournier, N., Delattre, S.: On the Kozachenko-Leonenko entropy estimator. ArXiv e-prints, February 2016
12. Glasser, M.F., et al.: A multi-modal parcellation of human cerebral cortex. Nature **536**(7615), 171–178 (2016)
13. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, vol. 27, pp. 2672–2680 (2014)
14. Gymrek, M., McGuire, A.L., Golan, D., Halperin, E., Erlich, Y.: Identifying personal genomes by surname inference. Science **339**(6117), 321–324 (2013)
15. Harmanci, A., Gerstein, M.: Quantification of private information leakage from phenotype-genotype data: linking attacks. Nat. Methods **13**(3), 251–256 (2016)
16. Hayes, J., Melis, L., Danezis, G., De Cristofaro, E.: LOGAN: evaluating privacy leakage of generative models using generative adversarial networks. ArXiv e-prints (2017)
17. Huang, C., Kairouz, P., Chen, X., Sankar, L., Rajagopal, R.: Context-aware generative adversarial privacy. CoRR abs/1710.09549 (2017)
18. Kairouz, P., Bonawitz, K., Ramage, D.: Discrete distribution estimation under local privacy. CoRR abs/1602.07387 (2016)

19. Kos, J., Fischer, I., Song, D.: Adversarial examples for generative models. CoRR abs/1702.06832 (2017)
20. Lee, H., Han, S., Lee, J.: Generative adversarial trainer: defense to adversarial perturbations with GAN. CoRR abs/1705.03387 (2017)
21. Liang, B., Li, H., Su, M., Li, X., Shi, W., Wang, X.: Detecting adversarial examples in deep networks with adaptive noise reduction. CoRR abs/1705.08378 (2017)
22. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**, 2579–2605 (2008). http://www.jmlr.org/papers/v9/vandermaaten08a.html
23. Mirjalili, V., Raschka, S., Namboodiri, A.M., Ross, A.: Semi-adversarial networks: convolutional autoencoders for imparting privacy to face images. CoRR abs/1712.00321 (2017)
24. Mirjalili, V., Ross, A.: Soft biometric privacy: retaining biometric utility of face images while perturbing gender. In: IJCB, pp. 564–573 (2017)
25. Narayanan, A., Shmatikov, V.: Robust de-anonymization of large sparse datasets. In: 2008 IEEE Symposium on Security and Privacy, SP 2008, pp. 111–125. IEEE (2008)
26. NIST: NIST 8-bit gray scale images of fingerprint image groups (FIGS)
27. Oh, S.J., Fritz, M., Schiele, B.: Adversarial image perturbation for privacy protection - a game theory perspective. CoRR abs/1703.09471 (2017)
28. Orekondy, T., Fritz, M., Schiele, B.: Connecting pixels to privacy and utility: automatic redaction of private information in images. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018
29. Rajpurkar, P., Hannun, A.Y., Haghpanahi, M., Bourn, C., Ng, A.Y.: Cardiologist-level arrhythmia detection with convolutional neural networks. ArXiv e-prints (2017)
30. Raval, N., Machanavajjhala, A., Cox, L.P.: Protecting visual secrets using adversarial nets. In: CVPR Workshop Proceedings (2017)
31. Sun, Q., Ma, L., Oh, S.J., Gool, L.V., Schiele, B., Fritz, M.: Natural and effective obfuscation by head inpainting. CoRR abs/1711.09001 (2017)
32. Sweeney, L., Abu, A., Winn, J.: Identifying participants in the personal genome project by name (a re-identification experiment). CoRR abs/1304.7605 (2013)
33. Tripathy, A., Wang, Y., Ishwar, P.: Privacy-preserving adversarial networks. CoRR abs/1712.07008 (2017)
34. Trzcinski, T., Lepetit, V.: Efficient discriminative projections for compact binary descriptors. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7572, pp. 228–242. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33718-5_17
35. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
36. Zhao, H., Gallo, O., Frosio, I., Kautz, J.: Loss functions for neural networks for image processing. CoRR abs/1511.08861 (2015). http://arxiv.org/abs/1511.08861

scientific reports

Check for updates

OPEN

Differences between human and machine perception in medical diagnosis

Taro Makino[1,2]✉, Stanisław Jastrzębski[1,2,3], Witold Oleszkiewicz[7], Celin Chacko[2], Robin Ehrenpreis[2], Naziya Samreen[2], Chloe Chhor[2], Eric Kim[2], Jiyon Lee[2], Kristine Pysarenko[2], Beatriu Reig[2,6], Hildegard Toth[2,6], Divya Awal[2], Linda Du[2], Alice Kim[2], James Park[2], Daniel K. Sodickson[2,3,5,6], Laura Heacock[2,6], Linda Moy[2,3,5,6], Kyunghyun Cho[1,4] & Krzysztof J. Geras[1,2,3,5]✉

Deep neural networks (DNNs) show promise in image-based medical diagnosis, but cannot be fully trusted since they can fail for reasons unrelated to underlying pathology. Humans are less likely to make such superficial mistakes, since they use features that are grounded on medical science. It is therefore important to know whether DNNs use different features than humans. Towards this end, we propose a framework for comparing human and machine perception in medical diagnosis. We frame the comparison in terms of perturbation robustness, and mitigate Simpson's paradox by performing a subgroup analysis. The framework is demonstrated with a case study in breast cancer screening, where we separately analyze microcalcifications and soft tissue lesions. While it is inconclusive whether humans and DNNs use different features to detect microcalcifications, we find that for soft tissue lesions, DNNs rely on high frequency components ignored by radiologists. Moreover, these features are located outside of the region of the images found most suspicious by radiologists. This difference between humans and machines was only visible through subgroup analysis, which highlights the importance of incorporating medical domain knowledge into the comparison.

Following their success in the natural image domain[1–7], deep neural networks (DNNs) have achieved human-level performance in various image-based medical diagnosis tasks[8–20]. DNNs have a number of additional benefits: they can diagnose quickly, do not suffer from fatigue, and can be deployed anywhere in the world. However, they currently possess a weakness which severely limits their clinical applicability. They cannot be fully trusted, given their tendency to fail for reasons unrelated to underlying pathology. For example, a dermatologist-level skin cancer classifier, approved for use as a medical device in Europe, learned to associate surgical skin markings with malignant melanoma[21]. As a result, the classifier's false positive rate increased by 40% in an external validation. Also, a pneumonia classifier was found to exploit differences in disease prevalence between sites, and was not making predictions solely based on underlying pathology[22].

In contrast, humans are more likely to fail because of the difficulty of the task, rather than for a superficial reason. This is partly because humans use features rigorously developed in their respective medical fields. Instead of being merely correlated with the presence of disease, there is a physiological reason such features are predictive. Therefore, in order to establish trust in machine-based diagnosis, it is important to know whether machines use different features than humans. Drawing inspiration from the natural image domain, we perform this comparison with respect to perturbation robustness[23–28]. By removing certain information from the input and analyzing the resulting change in prediction, we can infer the degree to which that information was utilized. We extend this line of work, taking into account a critically important consideration for medical diagnosis.

We argue that subgroup analysis is necessary in order to draw correct conclusions regarding medical diagnosis. In terms of predictive performance, different types of predictive errors can vary greatly in clinical

[1]Center for Data Science, New York University, New York, NY, USA. [2]Department of Radiology, NYU Langone Health, New York, NY, USA. [3]Center for Advanced Imaging Innovation and Research, NYU Langone Health, New York, NY, USA. [4]Department of Computer Science, Courant Institute, New York University, New York, NY, USA. [5]Vilcek Institute of Graduate Biomedical Sciences, NYU Grossman School of Medicine, New York, NY, USA. [6]Perlmutter Cancer Center, NYU Langone Health, New York, NY, USA. [7]Faculty of Electronics and Information Technology, Warsaw University of Technology, Warszawa, Poland. ✉email: taro@nyu.edu; k.j.geras@nyu.edu

significance[29]. All errors are treated equally when using empirical risk minimization, which can lead to a large disparity in predictive performance across subgroups. Robust optimization addresses this issue by considering the performance of various subgroups, and optimizing the worst-case subgroup performance[30,31]. Additionally, a failure to incorporate subgroups can lead to incorrect conclusions about perception due to Simpson's paradox[32]. The paradox is that subgroup-specific relationships can disappear or even reverse when the subgroups are aggregated. We therefore propose a framework which uses subgroup-specific perturbation robustness to compare human and machine perception in medical diagnosis. We demonstrate our framework with a case study in breast cancer screening, and show that failure to account for subgroups would indeed result in incorrect conclusions. It is important to note that while we analyze perturbation robustness, our purpose here is not to improve the robustness of machine-based diagnosis. Instead, we use perturbation robustness as a means of comparing human and machine perception.

In this framework, we identify subgroups that are diagnosed by humans in a significantly different manner, and specify an input perturbation that removes clearly-characterizable information from each of these subgroups. See Fig. 1 for an illustration of this applied to breast cancer screening. Predictions are collected from humans and machines on medical images perturbed with varying severity. We then apply probabilistic modeling to these predictions to capture the isolated effect of the perturbation, while factoring out individual idiosyncrasies. The resulting model is used to compare the perturbation robustness of humans and machines in terms of two criteria that are important for diagnosis: predictive confidence and class separability. Predictive confidence measures the strength of predictions, and is independent of correctness. Class separability represents correctness, and is quantified as the distance between the distributions of predictions for positive and negative cases. If humans and machines exhibit a different sensitivity to this perturbation, it implies that they are using different features. Next, we investigated the degree to which humans and machines agree on the most suspicious regions of an image. Radiologists annotated up to three regions of interest (ROIs) that they found most suspicious, and we analyzed the robustness of DNNs when low-pass filtering is applied to the interiors and exteriors of the ROIs, and to the entire image. See Fig. 2 for a visualization of this procedure for comparing humans and machines in the setting of breast cancer screening.

In our case study, we examined the sensitivity of radiologists and DNNs to Gaussian low-pass filtering, drawing separate conclusions for microcalcifications and soft tissue lesions. Low-pass filtering removes clearly-characterizable information in the form of high frequency components. In order to draw precise conclusions, it is more important for the removed information to be clearly-characterizable than clinically realistic. For example, a change in medical institutions is clinically realistic, but it is unclear what information is being removed. For microcalcifications, we found that radiologists and DNNs are both sensitive to low-pass filtering. Therefore, we could not conclude that humans and DNNs use different features to detect microcalcifications. Meanwhile, for soft tissue lesions, we found that humans are invariant to low-pass filtering, while DNNs are sensitive. This divergence suggests that humans and DNNs use different features to detect soft tissue lesions. Furthermore, using the ROIs annotated by radiologists, we found that a significant proportion of the high frequency components in soft tissue lesions used exclusively by DNNs lie outside of the regions found most suspicious by radiologists. Crucially, we show that without subgroup analysis, we would fail to observe this difference in behavior on soft tissue lesions, thus artificially inflating the similarity of radiologists and DNNs.

Results

Experimental setup. We experimented with the NYU Breast Cancer Screening Dataset[33] developed by our research team and used in a number of prior studies[12–14,34,35], and applied the same training, validation, and test set data split as previously reported. This dataset consists of 229,426 screening mammography exams from 141,473 patients. Each exam contains at least four images, with one or more images for each of the four standard views of screening mammography: left craniocaudal (L-CC), left mediolateral oblique (L-MLO), right craniocaudal (R-CC), and right mediolateral oblique (R-MLO). Each exam is paired with labels indicating whether there is a malignant or benign finding in each breast. See Supplementary Information Fig. 1 for an example of a screening mammogram. We used a subset of the test set for our reader study, which is also the same subset used in the reader study of[12]. For our DNN experiments, we used two architectures in order to draw general conclusions: the deep multi-view classifier (DMV)[12], and the globally-aware multiple instance classifier (GMIC)[13,14]. We primarily report results for GMIC, since it is the more recent and better-performing model. The corresponding results for DMV, which support the generality of our findings, are provided in the Supplementary Information section.

Perturbation reader study. In order to compare the perception of radiologists and DNNs, we applied Gaussian low-pass filtering to mammograms, and analyzed the resulting effect on their predictions. We selected nine filter severities ranging from unperturbed to severe, where severity was represented as a wavelength in units of millimeters on the physical breast. Details regarding the calculation of the filter severity are provided in the Methods section. Figure 1 demonstrates how low-pass filtering affects the appearance of malignant breast lesions.

We conducted a reader study in order to collect predictions for low-pass filtered images from radiologists. This reader study was designed to be identical to that of[12], except that the mammograms were randomly low-pass filtered in our case. We assigned the same set of 720 exams to ten radiologists with varying levels of experience. The images were presented to the radiologists in a conventional format, and an example is shown in Supplementary Information Fig. 1. Each radiologist read each exam once, and for each exam, we uniformly sampled one severity level out of our set of nine, and applied it to all images in the exam. The radiologists made binary predictions indicating the presence of a malignant lesion in each breast. We describe the details of the reader study in the Methods section.

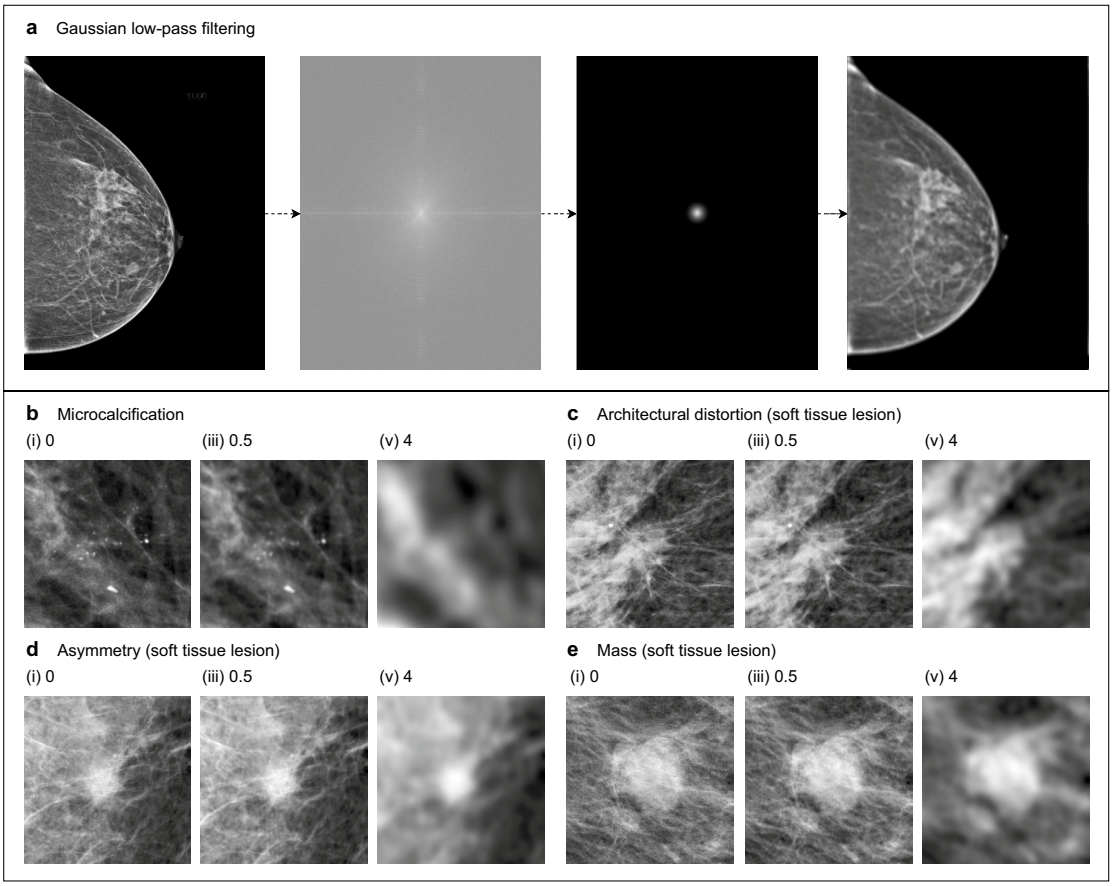


Figure 1. Identification of subgroups and an input perturbation. In our breast cancer screening case study, we separately analyzed two subgroups: microcalcifications and soft tissue lesions, using Gaussian low-pass filtering as the input perturbation. **(a)** Gaussian low-pass filtering is composed of three operations. The unperturbed image is transformed to the frequency domain via the two-dimensional discrete Fourier transform (DFT). A Gaussian filter is applied, attenuating high frequencies. The image is then transformed back to the spatial domain with the inverse DFT. **(b–e)** Gaussian low-pass filtering applied to various types of malignant breast lesions. Subfigures (i–iii) show the effects of low-pass filtering of increasing severity. **(b)** Microcalcifications are tiny calcium deposits in breast tissue that appear as white specks. Radiologists must often zoom in significantly in order to see these features clearly. Since these microcalcifications have a strong high frequency component, their visibility is severely degraded by low-pass filtering. **(c)** Architectural distortions indicate a tethering or indentation in the breast parenchyma. One of their identifying features are radiating thin straight lines, which become difficult to see after filtering. **(d)** Asymmetries are unilateral fibroglandular densities that do not meet the criteria for a mass. Low-pass filtering blurs their borders, making them blend into the background. **(e)** Masses are areas of dense breast tissue. Like asymmetries, masses generally become less visible after low-pass filtering, since their borders become less distinct. In our subgroup analysis, we aggregated architectural distortions, asymmetries, and masses into a single subgroup called "soft tissue lesions." This grouping was designed to distinguish between localized and nonlocalized lesions. Soft tissue lesions on the whole are far less localized than microcalcifications, and they require radiologists to consider larger regions of the image during the process of diagnosis. Figure created with drawio v13.9.0 https://github.com/jgraph/drawio.

We then trained five DNNs from random weight initializations, and made predictions on the same set of 720 exams. We repeated this nine times, where the set of exams was low-pass filtered with each of the nine filter severities. We note that for each DNN, we made a prediction for every pair of exam and filter severity. In contrast, for each radiologist, we only had predictions for a subset of the possible combinations of exam and filter severity. This means that if we arrange the predictions in a matrix where each row represents a filter severity and each column an exam, the matrix of predictions is sparse for each radiologist, and dense for each DNN. This fact is visualized in Fig. 2b, c. The sparsity of the radiologist predictions is by design; we were careful to ensure that each radiologist only read each exam once, since if they were to have seen the same exam perturbed with multiple filter severities, their predictions would have been unlikely to be independent. However, the sparsity prevents

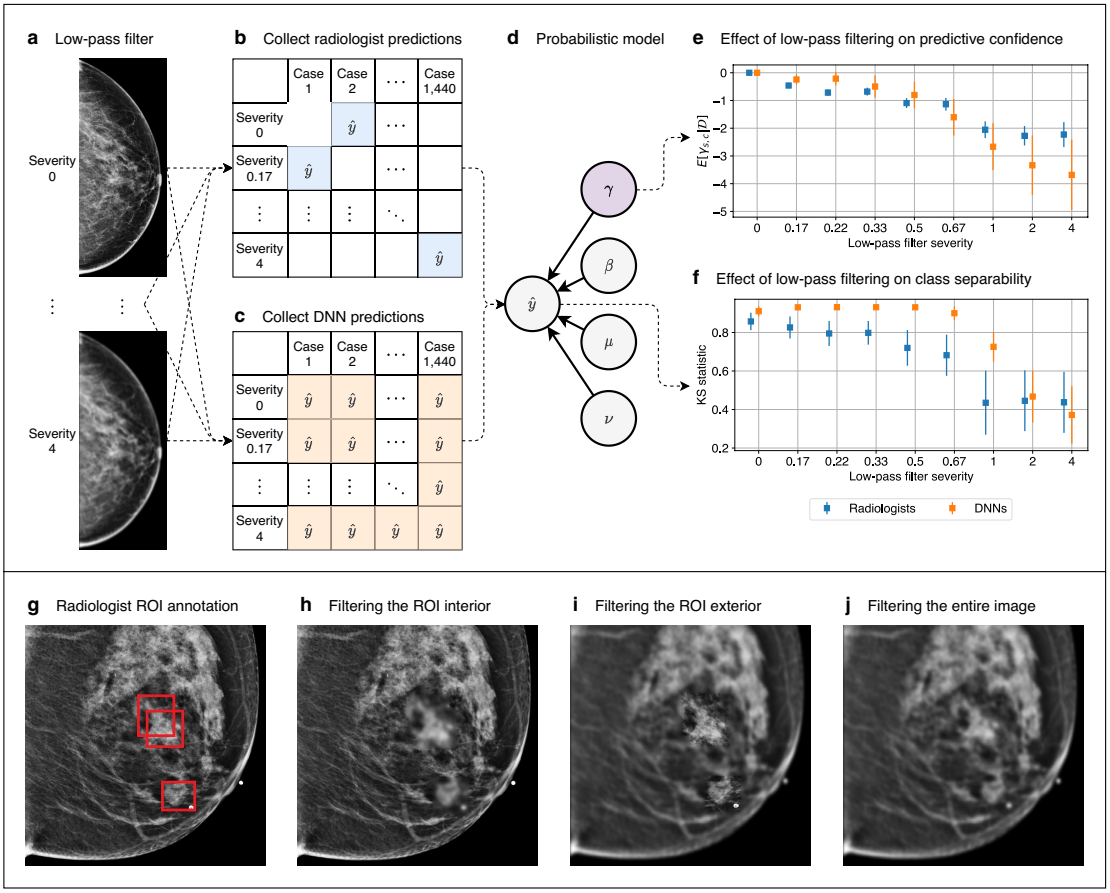


Figure 2. Our framework applied to breast cancer screening. **(a–f)** Comparing radiologists and DNNs with respect to their perturbation robustness. **(a)** We applied low-pass filtering to a set of mammograms using a wide range of filter severities. **(b)** We conducted a reader study in which each reader was provided with the same set of mammograms. Each reader saw each exam once, and each exam was filtered with a random severity. Thus, each radiologist's predictions populate a sparse matrix. **(c)** Predictions were collected from DNNs on the same set of exams. Unlike radiologists, DNNs made a prediction for all pairs of filter severities and cases, so their predictions form a dense matrix. **(d)** Probabilistic modeling was applied to the predictions, where a latent variable γ measures the effect of low-pass filtering, and a separate variable η factors out individual idiosyncrasies. **(e)** We examined the posterior expectation of γ to evaluate the effect of low-pass filtering on predictive confidence. **(f)** We sampled from the posterior predictive distribution and computed the distance between the distributions of predictions for malignant and nonmalignant cases. This represents the effect that low-pass filtering has on class separability. **(g–j)** Comparison of radiologists and DNNs with respect to the regions of an image they find most suspicious. **(g)** Radiologists annotated up to three regions of interest (ROIs) that they found most suspicious. We then applied low-pass filtering to: **(h)** the ROI interior, **(i)** the ROI exterior, and **(j)** the entire image. We analyzed the robustness of DNNs to these three filtering schemes in order to understand the degree to which the DNNs utilize information in the interiors and exteriors of the ROIs. Figure created with drawio v13.9.0 https://github.com/jgraph/drawio.

us from comparing radiologists and DNNs using evaluation metrics that use predictions for the complete set of exams. We therefore utilized probabilistic modeling to use the available radiologist predictions to infer values for the missing predictions.

A probabilistic model of predictions. We applied probabilistic modeling to achieve two purposes. The first is to study the effect of low-pass filtering on specific subgroups of lesions in isolation, after factoring out various confounding effects such as the idiosyncracies of individual radiologists and DNNs. The second is to infer the radiologists' predictions for each pair of exam and filter severity, since some pairs were missing by design. We modeled the radiologists' and DNNs' predictions as i.i.d. Bernoulli random variables. Let us denote radiologist (DNN) r's prediction on case n filtered with severity s as $\hat{y}_{r,s}^{(n)}$. We parameterized our model as

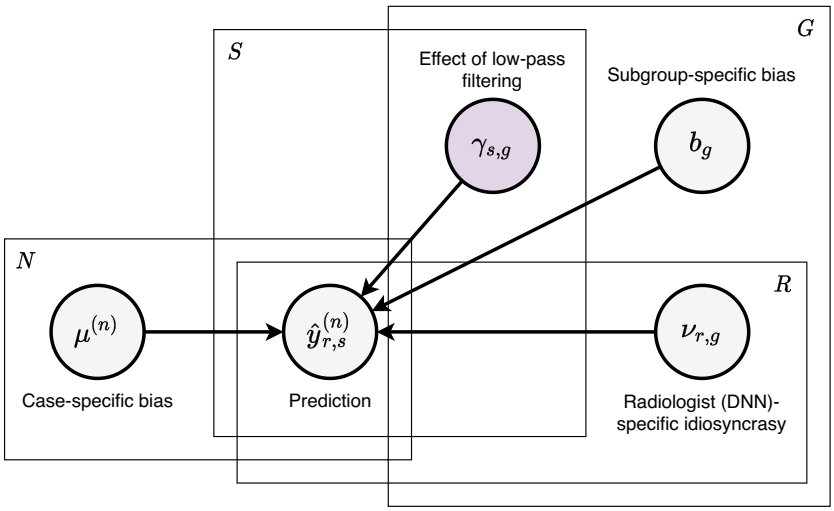


Figure 3. Probabilistic model. Our modeling assumption is that each prediction of radiologists and DNNs is influenced by four latent variables. $\hat{y}^{(n)}_{r,s}$ is radiologist (DNN) r's prediction on case n filtered with severity s. As for the latent variables, b_g represents the bias for subgroup g, $\mu^{(n)}$ is the bias for case n, $\gamma_{s,g}$ is the effect that low-pass filtering with severity s has on lesions in subgroup g, and $\nu_{r,g}$ is the idiosyncrasy of radiologist (DNN) r on lesions in subgroup g. Our analysis relies on the posterior distribution of $\gamma_{s,g}$, as well as the posterior predictive distribution of $\hat{y}^{(n)}_{r,s}$. The other latent variables factor out potential confounding effects. Figure created with drawio v13.9.0 https://github.com/jgraph/drawio.

$$\hat{y}^{(n)}_{r,s} \sim \text{Bernoulli}\left(\sigma\left(b_g + \mu^{(n)} + \gamma_{s,g} + \nu_{r,g}\right)\right), \tag{1}$$

where σ is the logistic function. There are four latent variables with the following interpretation: b_g represents the bias of exams in subgroup g, $\mu^{(n)}$ is the bias of exam n, $\gamma_{s,g}$ is the effect that low-pass filtering with severity s has on exams in subgroup g, and $\nu_{r,g}$ is the idiosyncrasy of radiologist (DNN) r on exams in subgroup g. See Fig. 3 for a graphical representation of our model. We considered several parameterizations of varying complexity, and selected the one with the maximum marginal likelihood. See the Methods section for details regarding our probabilistic model.

Comparing humans and machines with respect to their perturbation robustness. Using the probabilistic model, we compared how low-pass filtering affects the predictions of radiologists and DNNs, separately analyzing microcalcifications and soft tissue lesions. We performed each comparison with respect to two metrics: predictive confidence and class separability. Since the latent variable $\gamma_{s,g}$ represents the effect of low-pass filtering on each prediction, we examined its posterior distribution in order to measure the effect on predictive confidence. We sampled values of $\hat{y}^{(n)}_{r,s}$ from the posterior predictive distribution in order to quantify how low-pass filtering affects class separability. We computed the Kolmogorov-Smirnov (KS) statistic between the sampled predictions for the positive and negative class. This represents the distance between the two distributions of predictions, or how separated the two classes are. Sampling from the posterior predictive distribution was necessary for radiologists, since we did not have a complete set of predictions from them. Although such sampling was not strictly necessary for DNNs given the full set of available predictions, we performed the same posterior sampling for DNNs in order to ensure a fair comparison.

Figure 4a presents the results for microcalcifications. We only consider DNNs that are trained with unperturbed data in this section. The results for training DNNs on low-pass filtered data are discussed in the next section. The left subfigure represents predictive confidence, as measured by the posterior expectation of $\gamma_{s,g}$. Since low-pass filtering removes the visual cues of malignant lesions, we hypothesized that it should decrease predictive confidence. In other words, we expected to see $\mathbb{E}[\gamma_{s,g} \mid \mathcal{D}] \leq 0$. Above the left subfigure, we report $P(\gamma_{s,g} > 0 \mid \mathcal{D})$ in order to quantify how much the posterior distributions $P(\gamma_{s,g} \mid \mathcal{D})$ align with this hypothesis. Small values indicate a significant negative effect on predictive confidence. We note that these values are not intended to be interpreted as the p-values of a statistical test. Instead, they quantify the degree to which each $\gamma_{s,g}$ is negative. We observe that for microcalcifications, low-pass filtering decreases the predictive confidence of both radiologists and DNNs. There is, however, a nuanced difference in that for the range of most severe filters, the effect is constant for radiologists, while DNNs continue to become less confident.

The right subfigure of Fig. 4a depicts the effect of low-pass filtering on class separability. This is quantified by the KS statistic between the predictions for the positive and negative class, where the positive class is restricted to malignant microcalcifications. Similar to our hypothesis that low-pass filtering decreases predictive confidence, we hypothesized that it should also reduce class separability. That is, we expected the KS statistics for severity

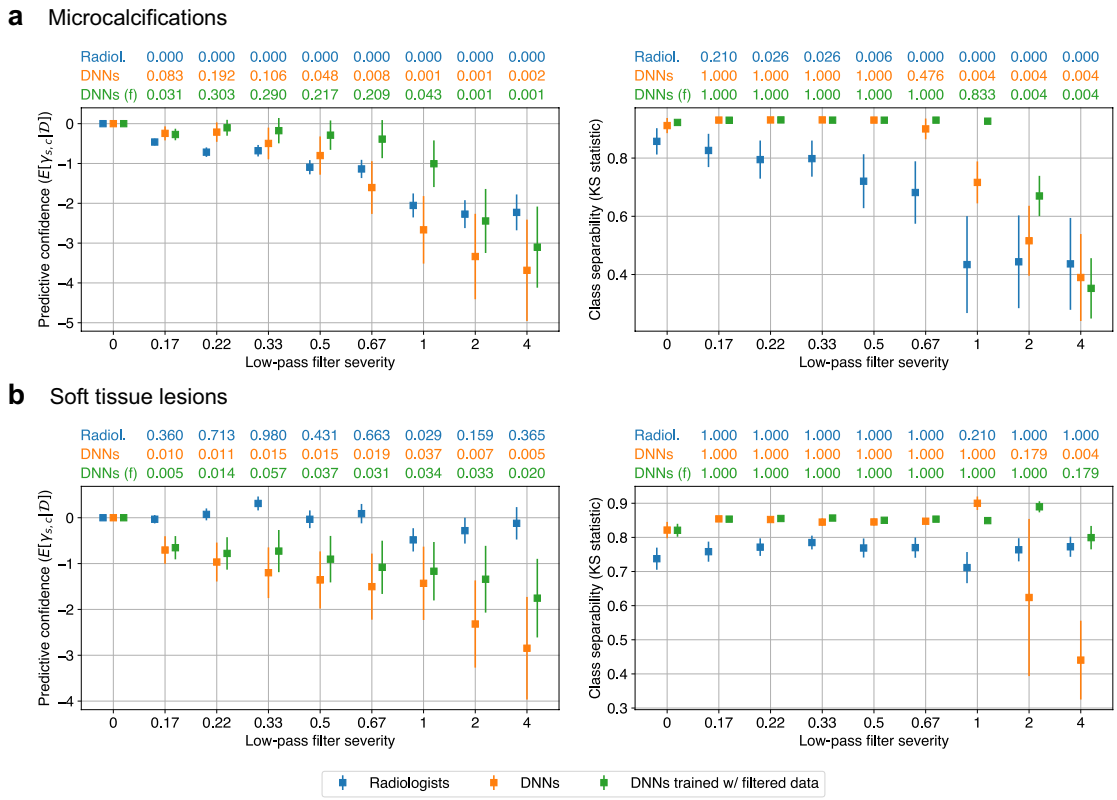


Figure 4. Comparing human and machines with respect to their perturbation robustness. The left subfigures represent the effect on predictive confidence, measured as the posterior expectation of $\gamma_{s,g}$ for severity s and subgroup g. The values at the top of each subfigure represent the probability that the predictive confidence for each severity is greater than zero. Smaller values for a given severity indicate a more significant downward effect on predictive confidence. The right subfigures correspond to the effect on class separability, quantified by the two-sample Kolmogorov–Smirnov (KS) statistic between the predictions for the positive and negative classes. The values at the top of each subfigure are the p-values of a one-tailed KS test between the KS statistics for a given severity and severity zero. Smaller values indicate a more significant downward effect on class separability for that severity. **(a)** For microcalcifications, low-pass filtering degrades predictive confidence and class separability for both radiologists and DNNs. When DNNs are trained with filtered data, the effects on predictive confidence and class separability are reduced, but not significantly. **(b)** For soft tissue lesions, filtering degrades predictive confidence and class separability for DNNs, but has no effect on radiologists. When DNNs are trained with filtered data, the effect on predictive confidence is reduced, and DNN-derived class separability becomes invariant to filtering. Figure created with drawio v13.9.0 https://github.com/jgraph/drawio.

$s > 0$ to be smaller than those for $s = 0$. This is because removing the visual cues for malignant lesions should make it more difficult to distinguish between malignant and nonmalignant cases. We tested this hypothesis using the one-tailed KS test between the KS statistics for $s = 0$ and $s > 0$. The p-values for this test are reported above the right subfigures, where small values mean that filtering significantly decreases class separability. We found that low-pass filtering decreases class separability for both radiologists and DNNs, but in different ways. The radiologists' class separability steadily declines for the range of less severe filters, while it is constant for DNNs. Meanwhile, similar to what we observed for predictive confidence, the radiologists' class separability is constant for the range of most severe filters, while it continues to decline for DNNs. While there are some nuanced differences, speaking generally, the radiologists and DNNs are both sensitive to low-pass filtering on microcalcifications. Therefore, we cannot conclude that humans and DNNs use different features to detect microcalcifications.

Next, we compared how low-pass filtering affects the predictions of radiologists and DNNs on soft tissue lesions (Fig. 4b). The results show that low-pass filtering degrades the predictive confidence and class separability of DNNs, while having almost no effect on radiologists. Since DNNs are sensitive to a perturbation that radiologists are invariant to, we conclude that humans and DNNs use different features to detect soft tissue lesions. This is a significant difference between human and machine perception, and may be attributable to certain inductive

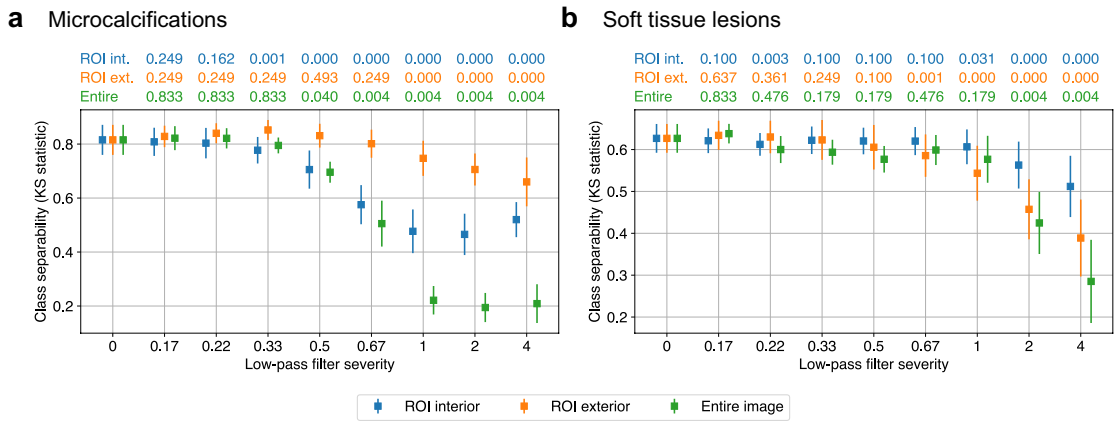


Figure 5. Comparing humans and machines with respect to the regions of an image deemed most suspicious. The performance of DNNs trained on unfiltered images was evaluated on images with selective perturbations in regions of interest (ROIs) identified as suspicious by human radiologists. **(a)** For microcalcifications, filtering the ROI interior decreases predictive confidence, but not as much as filtering the entire image. Filtering the ROI exterior decreases predictive confidence as well, meaning that DNNs utilize high frequency components in both the interior and the exterior of the ROIs, whereas humans focus more selectively on those ROIs. **(b)** For soft tissue lesions, filtering the ROI interior has very little effect on class separability. Meanwhile, filtering the ROI exterior has a similar effect to filtering the entire image. This implies that the high frequency components used by DNNs in these lesion subgroups are not localized in the areas that radiologists consider suspicious. Figure created with drawio v13.9.0 https://github.com/jgraph/drawio.

biases possessed by DNNs, such as their tendency to look at texture over shape[36]. Such differences must be better understood and reconciled in order to establish trust in machine-based diagnosis.

Training DNNs with low-pass filtered data. We observed that low-pass filtering decreases the predictive confidence and class separability of DNNs for all lesion subgroups. However, since the DNNs only encountered low-pass filtering during testing, it is possible that this effect is solely due to the dataset shift between training and testing. We therefore repeated the previous experiments for DNNs, where the same filtering was applied during both training and testing. We then examined whether the effects of low-pass filtering on the DNNs' perception could be attributed to information loss rather than solely to dataset shift.

For microcalcifications (Fig. 4a), training on filtered data slightly reduced the effect of low-pass filtering on predictive confidence and class separability, but the effect was still present, particularly for the most severe filters. This is evident from comparing the p-values for the two types of DNNs. The DNNs trained on unperturbed data generally have smaller p-values, except for the two most severe filters. This implies that the effect of filtering on microcalcifications can be attributed to information loss, and not solely to dataset shift. In other words, high frequency components in microcalcifications contain information that is important to the perception of DNNs.

Meanwhile, for soft tissue lesions (Fig. 4b), training on low-pass-filtered data significantly reduces the effect on predictive confidence and class separability, even for severe filters. This suggests that the effect of low-pass filtering on soft tissue lesions can primarily be attributed to dataset shift rather than information loss. In fact, DNNs trained with low-pass-filtered data maintain a similar level of class separability compared to networks trained on the unperturbed data. This confirms what we observed for radiologists, which is that high frequency components in soft tissue lesions are largely dispensable, and that more robust features exist.

Annotation reader study. Our analysis thus far has purely been in the frequency domain. Here, we extend our comparison to the spatial domain by examining the degree to which radiologists and DNNs agree on the most suspicious regions of an image. We conducted a reader study in which seven radiologists annotated up to three regions of interest (ROIs) containing the most suspicious features of each image. 120 exams were used in this study, which is a subset of the 720 exams in the perturbation reader study. See the Methods section for details regarding this reader study. We then applied low-pass filtering to the interior and exterior of the ROIs, as well as to the entire image. Examples of the annotation and the low-pass filtering schemes are shown in Fig. 2g–j. We made predictions using DNNs trained with the unperturbed data in order to understand the relationship between the high frequency components utilized by DNNs, and the regions of mammograms that are most suspicious to the radiologists.

Comparing humans and machines with respect to the regions of an image deemed most suspicious. We began by comparing the effect of the three low-pass filtering schemes on the DNNs' predictions for microcalcifications (Fig. 5a). We observed that filtering the ROI interior has a similar effect to filtering the entire image for mild filter severities. This suggests that for these frequencies, DNNs primarily rely on the same regions

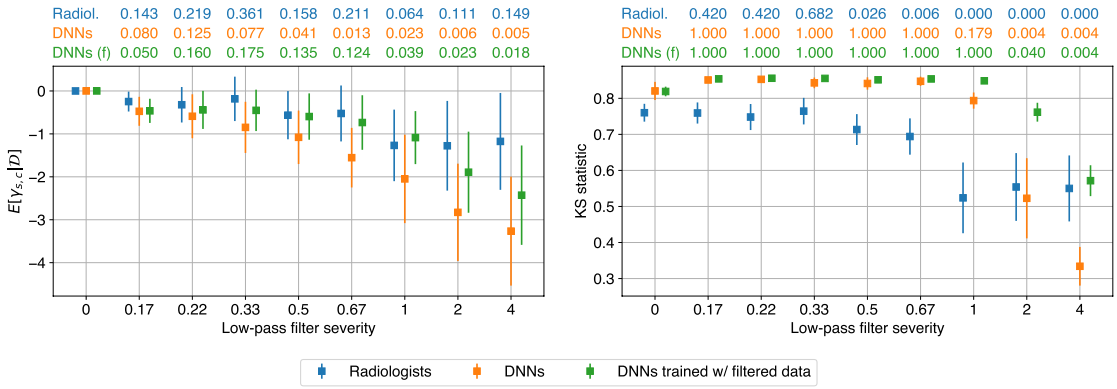


Figure 6. Simpson's paradox leads to incorrect conclusions. If we merged microcalcifications and soft tissue lesions into a single subgroup, we would incorrectly conclude that radiologists and DNNs exhibit similar perturbation robustness both for predictive confidence (left) and for class separability (right). This highlights the importance of performing subgroup analysis when comparing human and machine perception. Figure created with drawio v13.9.0 https://github.com/jgraph/drawio.

that radiologists consider suspicious. Meanwhile, class separability for the two ROI-based filtering schemes diverge significantly for high severities: filtering the ROI interior ceases to further decrease class separability at some threshold filter severity, whereas exterior filtering continues to degrade class separability beyond this threshold. The implication is that a range of high frequency components utilized by DNNs exist in the exterior of the ROIs deemed important by radiologists.

For soft tissue lesions (Fig. 5b), filtering the ROI interior decreases class separability, but to a lesser degree compared to filtering the entire image. This means that DNNs do utilize high frequency components in regions that radiologists find suspicious, but only to a limited degree. Meanwhile, filtering the ROI exterior has a similar effect on class separability as filtering the entire image. These observations suggest that the high frequency components that DNNs use for soft tissue lesions may be scattered across the image, rather than being localized in the areas that radiologists consider suspicious. While it is already established that DNNs utilize global information in screening mammograms[37], this effect appears to be more pronounced for soft tissue lesions compared to microcalcifications.

Discussion

Our framework draws inspiration from perturbation robustness studies in the adjacent domain of natural images, where there is also an ongoing crisis regarding the trustworthiness of machine perception. One key innovation in our work is to incorporate subgroup analysis to draw precise conclusions regarding perception. Not accounting for subgroups can be very dangerous, as it can lead to drawing erroneous conclusions due to Simpson's paradox. In our case study, our conclusions would change significantly if we treated microcalcifications and soft tissue lesions as a single subgroup. As shown in Fig. 6, we would incorrectly conclude that radiologists and DNNs have comparable perturbation robustness in terms of both predictive confidence and class separability, thus artificially inflating the similarity between human and machine perception.

The identification of subgroups with clinically meaningful differences is a crucially important component of our framework, as it strongly influences the conclusions. It requires domain knowledge, and is not as simple as enumerating all possible subgroups. The reason is that, due to the rarity of some subgroups, there is a balance to strike between the number of specified subgroups and the amount of available data. In our case study, we combined architectural distortions, asymmetries, and masses into the soft tissue lesion subgroup because there are only 30 cases of malignant soft tissue lesions in our reader study dataset. By doing this, we addressed data scarcity while accounting for the fact that soft tissue lesions as a whole are much less localized than microcalcifications, and thus require a significantly different diagnostic approach.

The choice of input perturbation is another important consideration. For the purpose of understanding perception, it is more important for the removed information to be clearly-characterizable than for the perturbation to be clinically realistic. For example, analyzing robustness in cross-institutional settings is clinically realistic, but it does not allow us to draw precise conclusions regarding perception, since it is unclear what information may have changed between institutions. Having said that, if the perturbation removes clinically relevant information and is additionally clinically realistic, this is beneficial because it allows us to reason about robustness in a plausible scenario. Our choice of Gaussian low-pass filtering is clinically relevant, as another type of low-pass filtering called motion blurring does occur in practice. Mammograms can be blurred by motion caused by patients or imaging devices[38], and[39] demonstrated that it can degrade the ability of radiologists to detect malignant lesions. While there exist differences between Gaussian low-pass filtering and motion blurring, we expect that robustness to the former will translate to the latter. This is because DNNs have been shown to exhibit similar robustness characteristics to various types of blurring[27]. We noticed that when comparing the class separability between

radiologists and DNNs trained with low-pass-filtered data (Fig. 4), DNNs are more robust to low-pass filtering for microcalcifications, while both are largely invariant for soft tissue lesions. This may be a significant advantage of DNNs in clinical practice.

It is nontrivial to ensure a fair comparison between humans and machines, and there is a growing body of work contemplating this issue. In[40], the authors draw inspiration from cognitive science, and argue that we should compare humans and machines with respect to competence, rather than performance. One suggestion they make is to make humans similar to machines, or vice versa. As an example, consider the fact that the human visual system is foveated, meaning that incoming light is sampled with spatially-varying resolution. In contrast, machines typically perceive images at a uniform resolution. In[41], the authors encode this inductive bias into a DNN architecture to demonstrate several advantages in generalization and robustness. The authors of[42] also advocate for aligning experimental conditions between humans and machines in order to avoid drawing conclusions which are incorrect due to flawed experimental design. With these related works in mind, we made several considerations to ensure a fair comparison between humans and machines.

First, given the infeasibility of perfectly aligning experimental conditions, we did not directly compare radiologists and DNNs with respect to any evaluation metrics. In other words, we avoided drawing conclusions such as "radiologists are more robust than DNNs to low-pass filtering because their predictive confidence is higher." Instead, we compared radiologists and DNNs to themselves in the unperturbed setting, and drew conclusions by contrasting how they changed when exposed to low-pass filtering.

Second, we made sure that none of the reported differences between radiologists and DNNs was due to the low-pass filtering being imperceptible to humans. The main difference we observe between radiologists and DNNs is that for soft tissue lesions, DNNs are sensitive to a range of high frequencies that humans are invariant to. This observed difference occurs at a range of high frequencies which, according to previous work, is perceptible to humans. We measure the severity of low-pass filtering as a wavelength in units of millimeters on the physical breast. Previous work[39] showed that simulated image blurring is visible to radiologists from 0.4 mm of movement. We experimented with eight severities ranging from 0.17 to 4 mm. Of these, {0.17 mm, 0.2 mm, 0.33 mm} are below the visible threshold of 0.4 mm reported in[39]. Our conclusions remain in tact even if we were to remove these mild severities from our analysis.

Third, it is unclear to what degree low-pass filtering affects radiologists by removing salient information, versus making the images look different from what they are used to seeing. We minimized the latter effect by choosing a Gaussian filter which does not leave visible artifacts. Since it is impractical to retrain radiologists using low-pass filtered images, it is arguable which is a fairer comparison: DNNs trained on unperturbed images, or DNNs trained on low-pass filtered images. We therefore included results for both, and verified that our main conclusions do not depend on this choice.

Finally, we accounted for the fact that radiologists routinely zoom into suspicious regions of an image, and also simultaneously look at multiple images within an exam. Similar to[41], we experimented with DNN architectures designed to mimic these behaviors. The GMIC architecture[13,14] exhibits the zooming behavior, but only processes one image at a time. On the other hand, the DMV architecture[12] does not zoom in, but processes multiple images simultaneously. All of our conclusions hold for both architectures, which suggests that our conclusions are not sensitive to either of these radiologists' behaviors.

In summary, we proposed a framework for comparing human and machine perception in medical diagnosis, which we expect to be applicable to a variety of clinical tasks and imaging technologies. The framework uses perturbation robustness to compare human and machine perception. To avert Simpson's paradox, we draw separate conclusions for subgroups that differ significantly in their diagnostic approach. We demonstrated the efficacy of this framework with a case study in breast cancer screening, and revealed significant differences between radiologists and DNNs. For microcalcifications, radiologists and DNNs are both sensitive to low-pass filtering, so we were unable to conclude whether they use different features. In contrast, radiologists are invariant and DNNs are sensitive to low-pass filtering on soft tissue lesions, which suggests that they use different features for this subgroup. From our annotation reader study, we found that these DNN-specific features in soft tissue lesions predominantly exist outside of the regions of an image found most suspicious by radiologists. We also showed that we would have missed this stark divergence between radiologists and DNNs in soft tissue lesions if we failed to perform subgroup analysis. This is evidence that future studies comparing human and machine perception in other medical domains should separately analyze subgroups with clinically meaningful differences. By utilizing appropriate subgroup analysis driven by clinical domain knowledge, we can draw precise conclusions regarding machine perception, and potentially accelerate the widespread adoption of DNNs in clinical practice.

Methods

All methods were carried out in accordance with relevant guidelines and regulations, and consistent with the Declaration of Helsinki. The NYU Breast Cancer Screening Dataset[33] was obtained under the NYU Langone Health IRB protocol ID#i18-00712_CR3. Informed consent was waived by the IRB. This dataset was extracted from the NYU Langone Health private database, and is not publicly available.

DNN training methodology. We conducted our DNN experiments using two architectures: the Deep Multi-View Classifer[12], and the Globally-Aware Multiple Instance Classifier[13,14]. With both architectures, we trained an ensemble of five models. A subset of each model's weights was initialized using weights pretrained on the BI-RADS label optimization task[43], while the remaining weights were randomly initialized. For each architecture, we adopted the same training methodology used by the corresponding authors.

Probability calibration. We applied Dirichlet calibration[44] to the predictions of DNNs used in our probabilistic modeling. This amounts to using logistic regression to fit the log predictions to the targets. We trained the logistic regression model using the validation set, and applied it to the log predictions on the test set to obtain the predictions used in our analysis. We used L2 regularization when fitting the logistic regression model, and tuned the regularization hyperparameter via an internal 10-fold cross-validation where we further split the validation set into "training" and "validation" sets. In the cross-validation, we minimized the classwise expected calibration error[44].

Gaussian low-pass filtering. Low-pass filtering is a method for removing information from images that allows us to interpolate between the original image and, in the most extreme case, an image where every pixel has the value of the mean pixel value of the original image. We experimented with nine filter severities selected to span a large range of the frequency spectrum. We implemented the Gaussian low-pass filter by first applying the shifted two-dimensional discrete Fourier transform to transform images to the frequency domain. The images were multiplied element-wise by a mask with values in [0, 1]. The values of this mask are given by the Gaussian function

$$M(u, v) = \exp\left(\frac{-D^2(u, v)}{2D_0^2}\right), \tag{2}$$

where u and v are horizontal and vertical coordinates, $D(u, v)$ is the Euclidian distance from the origin, and D_0 is the cutoff frequency. D_0 represents the severity of the filter, where frequencies are reduced to 0.607 of their original values when $D(u, v) = D_0$. Since the mammograms in our dataset vary in terms of spatial resolution as well as the physical length represented by each pixel, we expressed the filter severity D_0 in terms of a normalized unit of cycles per millimeter on the breast. Let $\alpha = \min(H, W)$ where H and W are the height and width of the image, and let β denote the physical length in millimeters represented by each pixel. Then we can convert cycles per millimeter D_0 to cycles per frame length of the image D_0^{img} using

$$D_0^{\text{img}} = D_0 \cdot \alpha \cdot \beta. \tag{3}$$

Perturbation reader study . In order to compare humans and machines with respect to their perturbation robustness, we conducted a reader study in which ten radiologists read 720 exams selected from the test set. While all radiologists read the same set of exams, each exam was low-pass filtered with a different severity for each radiologist. Except for the low-pass filtering, the setup of this reader study is identical to that of[12], and it was up to the radiologists to decide what equipment and techniques to use. Each exam consists of at least four images, with one or more images for each of the four views of mammography: L-CC, R-CC, L-MLO, and R-MLO. All images in the exam were concatenated into a single image such that the right breast faces left and is presented on the left, and the left breast faces right and is displayed on the right. Additionally, the craniocaudal (CC) views are on the top row, while the mediolateral oblique (MLO) views are on the bottom row. An example of this is shown in Supplementary Information Fig. 1. Among the 1440 breasts, 62 are malignant, 356 are benign, and the remaining 1022 are nonbiopsied. Among the malignant breasts, there are 26 microcalcifications, 21 masses, 12 asymmetries, and 4 architectural distortions, while in the benign breasts, the corresponding counts are: 102, 87, 36, and 6. For each exam, radiologists make a binary prediction for each breast, indicating their diagnosis of malignancy.

Probabilistic modeling. We modeled the radiologists' and DNNs' binary malignancy predictions with the Bernoulli distribution

$$\hat{y}_{r,s}^{(n)} \sim \text{Bernoulli}\,(p_{r,s}^{(n)}), \tag{4}$$

where $n \in \{1, 2, \ldots, 1440\}$ indexes the breast, $r \in \{1, 2, \ldots, 10\}$ the reader, and $s \in \{1, 2, \ldots, 9\}$ the low-pass filter severity. Each distribution's parameter $p_{r,s}^{(n)}$ is a function of four latent variables

$$p_{r,s}^{(n)} = \sigma(b_g + \mu^{(n)} + \gamma_{s,g} + \nu_{r,g}),$$

where $c \in \{1, \ldots, 5\}$ indexes the subgroup of the lesion. We included the following subgroups: unambiguous microcalcifications, unambiguous soft tissue lesions, ambiguous microcalcifications and soft tissue lesions, mammographically occult, and nonbiopsied. We considered these five subgroups in order to make use of all of our data, but only used the first two in our analysis. We assigned the generic weakly informative prior $\mathcal{N}(0, 1)$ to each latent variable. We chose the Bernoulli distribution because it has a single parameter, and thus makes the latent variables interpretable. Additionally, radiologists are accustomed to making discrete predictions in clinical practice. The posterior distribution of the latent variables is given by

$$p(\boldsymbol{b}, \boldsymbol{\mu}, \boldsymbol{\gamma}, \boldsymbol{\nu} \mid \hat{\boldsymbol{y}}) = \frac{p(\hat{\boldsymbol{y}} \mid \boldsymbol{b}, \boldsymbol{\mu}, \boldsymbol{\gamma}, \boldsymbol{\nu})}{p(\hat{\boldsymbol{y}})}.$$

The exact computation of the posterior is intractable, since the marginal likelihood $p(\hat{\boldsymbol{y}})$ involves a four-dimensional integral. We therefore applied automatic differentiation variational inference (ADVI)[45] in order to approximate the posterior. ADVI, and variational inference in general, optimizes over a class of tractable

distributions in order to find the closest match to the posterior. For our choice of tractable distributions, we used the mean-field approximation, meaning that we optimized over multivariate Gaussians with diagonal covariance matrices.

In practice, while radiologists made binary predictions, DNNs made continuous predictions in [0, 1] that we then calibrated. Despite the DNN predictions not being binary, we used equivalent procedures to specify the probabilistic model for radiologists and DNNs. To see how, let $\hat{y}^{(n)} \in \{0, 1\}$ denote a DNN's unobserved binary prediction for case n, and let $\hat{z}^{(n)} \in [0, 1]$ denote its observed real-valued prediction for the same case. In order to obtain $\hat{y}^{(n)} \in \{0, 1\}$, we could treat it as a random variable $\hat{y}^{(n)} \sim \text{Bernoulli}(\hat{z}^{(n)})$ and obtain values for it through sampling. We instead used $\hat{z}^{(n)}$ directly, specifying the log joint density as

$$\begin{aligned}
\log p(\hat{\boldsymbol{y}} \mid \theta) &= \sum_{n=1}^{N} \log p(\hat{y}^{(n)} \mid \theta^{(n)}) \\
&\approx \sum_{n=1}^{N} \mathbb{E}_{\hat{y}^{(n)}}[\log p(\hat{y}^{(n)} \mid \theta^{(n)})] \\
&= \sum_{n=1}^{N} \mathbb{E}_{\hat{y}^{(n)}}\left[\hat{y}^{(n)} \log(\theta^{(n)}) + (1 - \hat{y}^{(n)}) \log(1 - \theta^{(n)})\right] \\
&= \sum_{n=1}^{N} \Pr(\hat{y}^{(n)} = 1) \log(\theta^{(n)}) + \Pr(\hat{y}^{(n)} = 0) \log(1 - \theta^{(n)}) \\
&= \sum_{n=1}^{N} \hat{z}^{(n)} \log(\theta^{(n)}) + (1 - \hat{z}^{(n)}) \log(1 - \theta^{(n)}).
\end{aligned}$$

Annotation reader study. In order to compare humans and machines with respect to the regions of an image deemed most suspicious, we conducted a reader study in which seven radiologists read the same set of 120 unperturbed exams. The exams in this study were a subset of the 720 exams from the perturbation reader study, and also included all malignant exams from the test set. This study had two stages. In the first stage, the radiologists were presented with all views of the mammogram, and they made a malignancy diagnosis for each breast. This stage was identical to the reader study in[12]. In the second stage, for breasts that were diagnosed as malignant, the radiologists annotated up to three ROIs around the regions they found most suspicious. Overlapping ROIs were permitted. The radiologists annotated each view individually, and the limit of three ROIs applied separately to each view. For exams that contained multiple images per view, the radiologists annotated the image where the malignancy was most visible. The radiologists annotated the images using Paintbrush on MacOS, or Microsoft Paint on Windows. In order to constrain the maximum area that is annotated for each image, we included a 250×250 pixel blue ROI template in the bottom corner of each image to serve as a reference. The radiologists then drew up to three red ROIs such that each box approximately matched the dimensions of the reference blue ROI template. In our subsequent analysis, we computed results individually using each radiologist's ROIs, and then reported the mean and standard deviation across radiologists

Data availability

The radiologist and DNN predictions used in our analysis are available at https://github.com/nyukat/perception_comparison under the GNU AGPLv3 license.

Code availability

The code used in this research is available at https://github.com/nyukat/perception_comparison under the GNU AGPLv3 license. This, combined with the predictions which we also open-sourced, makes our results fully reproducible. We used several open-source libraries to conduct our experiments. The DNN experiments were performed using PyTorch[46], and the probabilistic modeling was done with PyStan (https://github.com/stan-dev/pystan), the Python interface to Stan[47]. The code for the DNNs used in our experiments is also open-source, where GMIC is available at https://github.com/nyukat/GMIC, and DMV at https://github.com/nyukat/breast_cancer_classifier.

Received: 30 September 2021; Accepted: 6 April 2022
Published online: 27 April 2022

References

1. Krizhevsky, A., Sutskever, I., & Hinton, G. E. ImageNet classification with deep convolutional neural networks. In *NIPS* 1106–1114 (2012).
2. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *ICLR* (2015).
3. Ren, S., He, K., Girshick, R. B., & Sun, J. Faster R-CNN: towards real-time object detection with region proposal networks. In *NIPS* 91–99 (2015).
4. Redmon, J., Divvala, S. K., Girshick, R. B., & Farhadi, A. You only look once: unified, real-time object detection. In *CVPR* 779–788 (2016).
5. He, K., Zhang, X., Ren, S., & Sun, J. Deep residual learning for image recognition. In *CVPR* 770–778 (2016).

6. Huang, G., Liu, Z., Maaten, L. van der, & Weinberger, K. Q. Densely connected convolutional networks. In *CVPR* 2261–2269 (2017).
7. He, K., Gkioxari, G., Dollár, P., & Girshick, R. B. Mask R-CNN. In *ICCV* 2980–2988 (2017).
8. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**(7639), 115–118 (2017).
9. Lindsey, R. V. *et al.* Deep neural network improves fracture detection by clinicians. *Proc. Natl. Acad. Sci. USA* **115**(45), 11591–11596 (2018).
10. Coudray, N. *et al.* Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**(10), 1559–1567 (2018).
11. Liu, X. *et al.* A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Dig. Health* **1**(6), e271–e297 (2019).
12. Wu, N. *et al.* Deep neural networks improve radiologists performance in breast cancer screening. *IEEE Trans. Med. Imag.* **39**(4), 1184–1194 (2019).
13. Shen, Y. *et al.* Globally-aware multiple instance classifier for breast cancer screening. In *International workshop on machine learning in medical imaging* 18–26 (Springer, New York, 2019).
14. Shen, Y. *et al.* An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. arXiv:2002.07613 (2020).
15. Rodriguez-Ruiz, A. *et al.* Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *JNCI* **111**(9), 916–922 (2019).
16. Ardila, D. *et al.* End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat. Med.* **25**(6), 954–961 (2019).
17. McKinney, S. M. *et al.* International evaluation of an AI system for breast cancer screening. *Nature* **577**(7788), 89–94 (2020).
18. Kim, H.-E. *et al.* Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *Lancet Dig. Health* **2**(3), e138–e148 (2020).
19. Schaffter, T. *et al.* Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. *JAMA Netw. Open* **3**(3), e200265–e200265 (2020).
20. Liu, Y. *et al.* A deep learning system for differential diagnosis of skin diseases. *Nat. Med.* **26**, 1–9 (2020).
21. Winkler, J. K. *et al.* Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatol.* **155**(10), 1135–1141 (2019).
22. Zech, J. R. *et al.* Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med.* **15**(11), e1002683 (2018).
23. Szegedy, C. *et al.* Intriguing properties of neural networks. In *ICLR* (2014).
24. Jo, J., & Bengio, Y. Measuring the tendency of CNNs to learn surface statistical regularities. arXiv:1711.11561 (2017).
25. Dodge, S. & Karam, L. A study and comparison of human and deep learning recognition performance under visual distortions. In *ICCCN*. IEEE. 1–7. (2017)
26. Geirhos, R. *et al.* Generalisation in humans and deep neural networks. *NeurIPS* **31**, 7549–7561 (2018).
27. Hendrycks, D., & Dietterich, T. G. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR* (2019).
28. Yin, D., Lopes, R. G., Shlens, J., Cubuk, E. D. & Gilmer, J. A fourier perspective on model robustness in computer vision. *NeurIPS.* **32**, 13255–13265 (2019).
29. Oakden-Rayner, L., Dunnmon, J., Carneiro, G., & Ré, C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *CHIL*. Ed. by M. Ghassemi. ACM 151–159. (2020)
30. Sagawa, S., Koh, P. W., Hashimoto, T. B., & Liang, P. Distributionally robust neural networks for group shifts: on the importance of regularization for worst-case generalization. In *ICLR* (2020).
31. Goel, K., Gu, A., Li, Y., & Ré, C. Model patching: closing the subgroup performance gap with data augmentation. arXiv:2008.06775 (2020).
32. Pearl, J. Comment: understanding Simpsons paradox. *Am. Stat.* **68**(1), 8–13 (2014).
33. Wu, N. *et al. The NYU breast cancer screening dataset v1* (Tech. rep, NYU, 2019).
34. Févry, T. *et al.* Improving localization-based approaches for breast cancer screening exam classification. arXiv:1908.00615 (2019).
35. Wu, N., Jastrzębski, S., Park, J., Moy, L., Cho, K., & Geras, K. J. Improving the ability of deep neural networks to use information from multiple views in breast cancer screening. In *Medical Imaging with Deep Learning*. PMLR. 827–842 (2020).
36. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*. (2019).
37. Wu, N. *et al.* Reducing false-positive biopsies with deep neural networks that utilize local and global information in screening mammograms. arXiv:2009.09282 (2020).
38. Choi, J. J. *et al.* Mammographic artifacts on full-field digital mammography. *JDI* **27**(2), 231–236 (2014).
39. Abdullah, A. K. *et al.* The impact of simulated motion blur on lesion detection performance in full-field digital mammography. *Brit. J. Radiol.* **90**(1075), 20160871 (2017).
40. Firestone, C. Performance vs. competence in human-machine comparisons. *Proc. Natl. Acad. Sci.* **117**(43), 26562–26571 (2020).
41. Deza, A. & Konkle, T. Emergent properties of foveated perceptual systems. arXiv:2006.07991 (2020).
42. Funke, C. M. *et al.* Five points to check when comparing visual perception in humans and machines. *J. Vis.* **21**(3), 16 (2021).
43. Geras, K. J. *et al.* High-resolution breast cancer screening with multi-view deep convolutional neural networks. arXiv:1703.07047 (2017).
44. Kull, M., Perelló-Nieto, M., Kängsepp, M., Menezes e Silva Filho, T. de, Song, H., & Flach, P. A. Beyond temperature scaling: obtaining well-calibrated multi-class probabilities with Dirichlet calibration. In *NeurIPS*. 12295–12305. (2019)
45. Kucukelbir, A., Ranganath, R., Gelman, A., & Blei, D. Automatic variational inference in Stan. In *NIPS*. 568–576. (2015).
46. Paszke, A. *et al.* PyTorch: an imperative style, high-performance deep learning library. In *NeurIPS*. 8024–8035. (2019).
47. Carpenter, B. *et al.* Stan: a probabilistic programming language. *J. Stat. Softw.* **76**, 1 (2017).

Acknowledgements

The authors would like to thank Mario Videna, Abdul Khaja and Michael Costantino for supporting our computing environment and Eric K. Oermann for helpful comments on the draft of this paper. We also gratefully acknowledge the support of NVIDIA Corporation, which donated some of the GPUs used in this research. This work was supported in part by grants from the National Institutes of Health (P41EB017183 and R21CA225175), the National Science Foundation (HDR-1922658), and the Gordon and Betty Moore Foundation (9683).

Author contributions

S.J. conceived the initial idea and designed the first set of experiments. T.M., S.J., W.O., K.C. and K.J.G. designed the final version of the experiments. T.M. and W.O. conducted the experiments with neural networks. T.M. and W.O. preprocessed the screening mammograms. T.M., S.J., L.M. and L.H. conducted the reader study. CeC, N.S.,

R.E., D.A., Ch.C., L.D., E.K., A.K., J.L., J.P., K.P., B.R. and H.T. collected the data. T.M., S.J. and W.O. conducted literature search. T.M., K.C. and K.J.G. designed the probabilistic model. T.M. performed the probabilistic inference. L.M., L.H. and B.R. analyzed the results from a clinical perspective. D.S. contributed to framing the paper. S.J., K.C. and K.J.G. supervised the execution of all elements of the project. D.S., L.M., L.H., K.C. and K.J.G. acquired funding for this research. All authors provided critical feedback and helped shape the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-10526-z.

Correspondence and requests for materials should be addressed to T.M. or K.J.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

4 Informacja o pozostałych osiągnięciach naukowych, dydaktycznych, organizacyjnych oraz popularyzujących naukę

4.1 Pozostałe publikacje naukowe

[B1] Cichosz, P., Jagodziński, D., Matysiewicz, M., Neumann, Ł., Nowak, R., Okuniewski, R. & **Oleszkiewicz, W.** Novelty detection for breast cancer image classification. *Photonics Applications In Astronomy, Communications, Industry, And High-Energy Physics Experiments 2016*. https://doi.org/10.1117/12.2249183

[B2] Jagodziński, D., Matysiewicz, M., Neumann, Ł., Nowak, R., Okuniewski, R., **Oleszkiewicz, W.** & Cichosz, P. Feature selection and definition for contours classification of thermograms in breast cancer detection. *Photonics Applications In Astronomy, Communications, Industry, And High-Energy Physics Experiments 2016*. https://doi.org/10.1117/12.2249064

[B3] Matysiewicz, M., Neumann, Ł., Nowak, R., Okuniewski, R., **Oleszkiewicz, W.**, Cichosz, P. & Jagodziński, D. Automatic recognition of thermographic examinations for early detection of breast cancer. *Photonics Applications In Astronomy, Communications, Industry, And High-Energy Physics Experiments 2016*. https://doi.org/10.1117/12.2249067

[B4] Neumann, Ł., Nowak, R., Okuniewski, R., **Oleszkiewicz, W.**, Cichosz, P., Jagodziński, D. & Matysiewicz, M. Preprocessing for classification of thermograms in breast cancer detection. *Photonics Applications In Astronomy, Communications, Industry, And High-Energy Physics Experiments 2016*. https://doi.org/10.1117/12.2249307

[B5] Nowak, R., Okuniewski, R., **Oleszkiewicz, W.**, Cichosz, P., Jagodziński, D., Matysiewicz, M. & Neumann, Ł. Asymmetry features for classification of thermograms in breast cancer detection. *Photonics Applications In Astronomy, Communications, Industry, And High-Energy Physics Experiments 2016*. https://doi.org/10.1117/12.2249066

[B6] Okuniewski, R., Nowak, R., Cichosz, P., Jagodziński, D., Matysiewicz, M., Neumann, Ł. & **Oleszkiewicz, W.** Contour classification in thermographic images for detection of breast cancer. *Photonics Applications In Astronomy, Communications, Industry, And High-Energy Physics Experiments 2016*. https://doi.org/10.1117/12.2249065

[B7] **Oleszkiewicz, W.**, Cichosz, P., Jagodziński, D., Matysiewicz, M., Neumann, Ł., Nowak, R. & Okuniewski, R. Application of SVM classifier in thermographic image classification for early detection of breast cancer. *Photonics Applications In Astronomy, Communications, Industry, And High-Energy Physics Experiments 2016*. https://doi.org/10.1117/12.2249063

4.2 Wystąpienia na warsztatach przy konferencjach

- Wystąpienie na warsztacie The Bright and Dark Sides of Computer Vision: Challenges and Opportunities for Privacy and Security przy konferencji CVPR 2018.

- Wystąpienie na warsztacie AI for Affordable Healthcare przy konferencji ICLR 2020.

4.3 Udział w grantach badawczych

- Stypendium w grancie Bio-inspired artificial neural networks Fundacji Nauki Polskiej (TeamNet), 01.04.2020r. – 31.03.2023r. Projekt realizowany na Uniwersytecie Jagielońskim w Krakowie, kierownik projektu: prof. Jacek Tabor.
- Udział w grancie Reverse EngiNeering of sOcial Information pRocessing (Horizon 2020), 08.06.2018r. – 08.09.2018r. Projekt realizowany na Politechnice Warszawskiej i Uniwersytecie Stanforda, kierownik projektu: prof. Janusz Hołyst.

4.4 Staże naukowe

- Staż naukowy u dr. Peter Kairouz na Uniwersytecie Stanforda, USA, 08.06.2018r. – 08.09.2018r.
- Staż naukowy u dr. Krzystofa Gerasa na Uniwersytecie Nowego Jorku, USA 23.02.2019r. – 16.03.2019r.

4.5 Praca dydaktyczna

Prowadzenie przedmiotów na studiach inżynierskich i magisterskich w języku polskim i angielskim dla studentów Wydziału Elektroniki i Technik Informacyjnych Politechniki Warszawskiej. Prowadzone przedmioty:

- Podstawy Informatyki i Programowania (360 godzin laboratorium, 170 godzin projekt),
- Podstawy Programowania (30 godzin laboratorium, 30 godzin projekt),
- Programowanie obiektowe (60 godzin laboratorium),
- Systemy Operacyjne (330 godzin laboratorium),
- Distributed Computing and Systems (215 godzin projekt),
- Systemy Rozproszone (30 godzin projekt),
- Podstawy Sztucznej Inteligencji (30 godzin projekt).

4.6 Nagrody

- Nagroda zespołowa dydaktyczna III stopnia Rektora Politechniki Warszawskiej za opracowanie nowych materiałów dydaktycznych do przedmiotu Podstawy Informatyki i Programowania, 09.2022.
- Stypendium Rektora Politechniki Warszawskiej dla najlepszych doktorantów, 2017-2018r.
- Stypendium programu wymiany zagranicznej do USA Fundacji Kościuszkowskiej, 03.2020r.

4.7 Recenzowanie prac naukowych

Recenzowałem pięć prac na ICML Workshop on Dynamic Neural Networks, 2022.

4.8 Udział w wydarzeniach popularyzujących naukę i innych konferencjach

- Wystąpienie na konferencji Prezesa Urzędu Ochrony Danych Osobowych, Sztuczna Inteligencja złodziejem Twoich danych osobowych, 30.11.2018r., Warszawa.
- Plakat na konferencji MLinPL, 23.11.2019r., Warszawa.
- Wystąpienie na konferencji Horizon Europe Weeks, 8.10.2021r., Kraków.
- Wystąpienie na konferencji AI & NLP Day, Warszawa, 25.10.2021r.
- Wystąpienia na seminariach Koła Naukowego Sztucznej Inteligencji Golem, Politechnika Warszawska.

4.9 Opieka naukowa nad studentami

- Promotor pracy inżynierskiej pt. Aplikacja mobilna dobierająca na podstawie jednego ubrania pozostałą część garderoby zgodnie z aktualnymi trendami. Praca obroniona w 02.2023r.
- Promotor pomocniczy dwóch prac magisterskich pt. Wspomaganie diagnozowania zapalenia płuc za pomocą analizy obrazów algorytmami sztucznej inteligencji oraz Breast cancer diagnosis from mammography screenings employing convolutional neural networks. Prace obronione w 10.2019r. oraz 07.2020r.
- Promotor pomocniczy dwóch prac inżynierskich w języku angielskim pt. Extending a dataset using Genarative Adversarial Networks (GANs) oraz Data anonymization techniques using generative adversarial networks (GANs) to increase prediction accuracy of deep neural networks. Prace obronione w 02.2019r.

4.10 Wykonane ekspertyzy lub inne opracowania na zamówienie

- Przygotowanie materiałów i przeprowadzenie szkoleń (32 godziny dla 16-stu osób) dotyczących zastosowania metod uczenia maszynowego dla pracowników Ministerstwa Finansów, 11.2018r., Politechnika Warszawska.
- Przygotowanie prototypu systemu telemedycznego do zdalnej diagnostyki pediatrycznej oparty na innowacyjnym urządzeniu multi-sensorycznym i algorytmach automatycznej interpretacji dla firmy Higosense, 01.–06.2018r, kierownik projektu: dr hab. Robert Nowak.
- Przygotowanie systemu wykrywanie zmian nowotworowych na podstawie obrazów termograficznych piersi dla firmy Braster, 01.2015r.–01.2018r., kierownik projektu dr hab. Robert Nowak.