RADA NAUKOWA DYSCYPLINY

INFORMATYKA TECHNICZNA I TELEKOMUNIKACJA POLITECHNIKI WARSZAWSKIEJ

zaprasza na

OBRONĘ ROZPRAWY DOKTORSKIEJ

mgr. inż. Jana Michała Dubińskiego

która odbędzie się w dniu **12 czerwca 2026 roku,** o godzinie **12:00** w trybie hybrydowym

Temat rozprawy:

"***Reliable and Safe Generative Models***"

Promotor: prof. dr hab. inż. Przemysław Rokita – Politechnika Warszawska

Recenzenci: prof. dr hab. inż. Bogusław Cyganek – Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie

prof. dr hab. inż. Rafał Scherer – Politechnika Częstochowska

prof. dr hab. inż. Adam Wojciechowski – Politechnika Łódzka

Obrona odbędzie się stacjonarnie w Sali nr 116 Wydziału Technik Informacyjnych Politechniki Warszawskiej w Warszawie.

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.-Jan-Dubinski

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

Niezawodne i Bezpieczne Modele Generatywne

Niniejsza praca przedstawia zestaw rozwiązań wspierających rozwój wiarygodnych i bezpiecznych modeli generatywnych. Proponujemy nowe podejścia do modelowania generatywnego, ukierunkowane na praktyczne zastosowania, ze szczególnym uwzględnieniem fizyki wysokich energii. Jednocześnie wprowadzamy metody chroniące własność intelektualną zawartą w modelach i danych treningowych. Dzięki temu wzmacniamy bezpieczeństwo i niezawodność generatywnego uczenia maszynowego, wspierając zarówno twórców i użytkowników modeli, jak i właścicieli danych.

W pierwszej części pracy koncentrujemy się na tworzeniu wiarygodnych modeli generatywnych na potrzeby zastosowań naukowych. Szczególny nacisk kładziemy na symulację eksperymentów z zakresu fizyki wysokich energii prowadzonych w Europejskiej Organizacji Badań Jądrowych (CERN). Proponujemy rozwiązania oparte na uczeniu maszynowym jako alternatywę dla tradycyjnych metod symulacji stosowanych w Wielkim Zderzaczu Hadronów. Osiągnięte rezultaty obejmują opracowanie generatywnych sieci adwersarialnych wiernie odwzorowujących różnorodność danych treningowych, a także stworzenie modelu opartego na mieszance ekspertów, umożliwiającego uchwycenie wielomodalnej natury generowanych danych.

W drugiej części pracy skupiamy się na ochronie wartości intelektualnej w modelach uczenia głębokiego. Wraz z rosnącą zdolnością modeli do rozwiązywaniu konkretnych problemów wzrasta również ich wartość. Naturalnie tworzy to potrzebę opracowania metod jej ochrony. W odpowiedzi proponujemy pierwszą metodę obrony przed nieautoryzowanym odtwarzaniem modeli typu enkoder, pozwalającą wykrywać próby eksploracji przestrzeni odpowiedzi modelu przez atakującego.

Ostatnia część pracy rozszerza perspektywę ochrony wartości intelektualnej z modeli na dane. Pokazujemy, że istniejące podejścia do identyfikacji danych używanych do trenowania modeli dyfuzyjnych mają istotne ograniczenia i proponujemy efektywną metodę wykrywania wykorzystania zbiorów chronionych prawem autorskim. Jako pierwsi analizujemy również zagrożenia dla prywatności w autoregresyjnych modelach wizyjnych i skutecznie adaptujemy naszą metodę do tego typu architektur.

Podsumowując, niniejsza praca wnosi wkład w rozwój modeli generatywnych zdolnych w sposób niezawodny odpowiadać na rzeczywiste wyzwania naukowe, a jednocześnie dostarcza mechanizmy ochrony zarówno modeli, jak i danych. Tym samym wspiera tworzenie godnego zaufania, wiarygodnego i odpowiedzialnego ekosystemu uczenia maszynowego opartego na metodach generatywnych.

Słowa kluczowe: Modele Generatywne, Sieci Generatywne Kontradyktoryjne, Modele Dyfuzyjne, Modele Autoregresyjne Obrazów, Fizyka Wysokich Energii, Bezpieczeństwo Uczenia Maszynowego

Politechnika Łódzka
Instytut Informatyki

Łódź, 27 lutego 2026 roku

prof. dr hab. inż. Adam Wojciechowski
Instytut Informatyki
Wydział Fizyki Technicznej, Informatyki i Matematyki Stosowanej
Politechnika Łódzka
Al. Politechniki 8, 93-590 Łódź

RECENZJA ROZPRAWY DOKTORSKIEJ

Tytuł rozprawy: **Reliable and safe generative models**

Autor rozprawy: **mgr inż. Jan Dubiński**

Promotor rozprawy: **prof. dr hab. inż. Przemysław Rokita**

Jakie zagadnienie naukowe jest rozpatrywane w pracy (teza rozprawy) i czy zostało ono dostatecznie jasno sformułowane przez Autora? Czy tematyka rozprawy jest aktualna lub dostatecznie ważna? Jaki charakter ma rozprawa (teoretyczny, doświadczalny, inny)?

Pan mgr inż. Jan Dubiński podjął w swojej pracy doktorskiej, zagadnienie bezpieczeństwa, niezawodności i użyteczności modeli generatywnych. Dysertacja została zredagowana na podstawie sześciu znakomitych publikacji, przyjętych i wygłoszonych na sześciu różnych wybitnych międzynarodowych konferencjach naukowych: NeurIPS, ICML, CVPR, ECAI, WACV, ICONIP. W szczególności Doktorant brał udział w opracowaniu modeli generatywnych na potrzeby symulacji fizycznych wysokich energii, ochrony modeli uczenia maszynowego przed skopiowaniem oraz identyfikacją danych treningowych wykorzystywanych do uczenia generatywnych modeli dyfuzyjnych i autoregresyjnych.

Tematyka rozprawy jest bardzo aktualna i ważna w kontekście współczesnego rozwoju oraz wykorzystania modeli sztucznej inteligencji. Należy podkreślić, że prace przedstawione przez Doktoranta ewidentnie wyznaczają granice współczesnych osiągnieć naukowych w przedmiotowym zakresie rozprawy.

Rozprawa ma charakter eksperymentalny i nosi w sobie istotny potencjał praktyczny. Doktorant w ramach każdego z poruszanych tematów nie tylko proponował oryginalne i wartościowe rozwiązania, ale popierał je bardzo rzetelną analizą porównawczą i szeroko zakrojoną optymalizacją hiperparametrów, uzasadniając merytorycznie podjęte decyzje. Praca z racji podjętej problematyki, zakresu oraz metodologii badawczej jest pełnoprawnym osiągnięciem o charakterze naukowo-badawczym w dziedzinie nauk inżynieryjno-technicznych, w dyscyplinie Informatyka Techniczna i Telekomunikacja.

Czy w rozprawie przeprowadzono w sposób właściwy analizę źródeł (w tym literatury światowej i stanu zagadnień w przemyśle) świadczącą o dostatecznej wiedzy Autora? Czy wnioski z przeglądu źródeł sformułowano w sposób jasny i przekonywujący?

W przedstawionej do oceny rozprawie doktorskiej, przegląd literatury realizowany był niezależnie w każdym z podjętych wątków, gdyż każda z przytoczonych prac, z racji chronologii badań, aktualizowała niezbędny zakres rozwiązań referencyjnych, odzwierciedlający aktualny stan zagadnienia. Za każdym razem Doktorant odnosił się do najnowszych doniesień literaturowych, wskazywał nisze badawcze lub kontestował skutecznie dotychczasowe osiągnięcia innych naukowców,

zarysowując przydatność i wagę swoich badań. Za każdym razem wnioski formułowano w sposób jasny i przekonujący. Sposób prowadzonego przeglądu świadczy dobitnie o doskonałej wiedzy Autora w poruszanej problematyce.

Przegląd aktualnego stanu zagadnienia niejednokrotnie motywował dobór metodologii badawczej dla prowadzonych eksperymentów. W niektórych pracach doprecyzowywano metodologię, wykazując niedoskonałości poprzednich sposobów weryfikacji. Analiza aktualnego stanu zagadnienia dotyczyła również zbiorów danych wykorzystywanych w eksperymentach. Niejednokrotnie Doktorant konstruował własne zbiory danych, gdyż uznawał, że istniejące zbiory i sposób ich wykorzystania są niewystarczające do prowadzenia rzetelnych eksperymentów. Wszystkie te działania potwierdzają świadomą i dogłębną analizę literatury, a podjęte, na podstawie analizy, działania świadczą o dojrzałości naukowej i stanowią bardzo istotne osiągniecie Doktoranta.

Jak Autor rozwiązał postawione zagadnienie, czy użył właściwej do tego metody?

Pierwsza część dokonań, zaraportowanych w ramach rozprawy doktorskiej, odnosi się do nowych metod wykorzystania modeli generatywnych w wizualizacji zjawisk fizycznych wysokich energii. Dwie pierwsze prace powstały w ramach badań zespołu współpracującego z ośrodkiem CERN, który zajmuje się wizualizacjami trajektorii cząstek elementarnych, w ramach eksperymentu ALICE.

W tym kontekście praca zatytułowana *„Selectively increasing the diversity of GAN-generated samples"*, opublikowana na konferencji ICONIP, porusza problem zwiększenia różnorodności wyników modelu generatywnego, którego celem jest wygenerowanie wiarygodnej wizualizacji trajektorii cząstek elementarnych w odniesieniu do trajektorii rejestrowanych przez detektory. Tradycyjne podejście, bazujące na zmiennych warunkowych, generuje ograniczoną różnorodność trajektorii, co limituje użyteczność tak powstałych symulacji. Opracowano zatem metodę *Selective Diversity GAN* (SDI-GAN), która selektywnie regularyzuje generator w sposób stymulujący różnorodność na etapie trenowania modelu danymi. Istotą metody jest wykorzystanie, do trenowania warunkowego modelu generatywnego *cGAN*, odpowiedniej różnorodności danych treningowych oraz monitorowania dedykowanej funkcji straty, która jest regularyzowana poprzez autorski komponent uwzględniający różnorodność (wariancję) próbek uczących. Dodatkowo, poziom niedopasowania wyników jest mierzony w przestrzeni ukrytej, a nie w przestrzeni pikseli obrazu. Eksperymenty prowadzone były zarówno na danych syntetycznych jak i danych pochodzących z kalorymetru zerostopniowego (*Zero Degree Calorimeter*), używanego w eksperymencie ALICE. Badania wykazały realne przełożenie wariancji danych wejściowych, zakodowanych w funkcji straty, na różnorodność wygenerowanych wyników modelem generatywnym (*cGAN*), bez ograniczania wiarygodności i przy zachowaniu atrakcyjnej wydajności procesu. Porównanie z metodami referencyjnymi wykazało zdecydowaną poprawę wyników symulacji względem wyników literaturowych. Doktorant, jako pierwszy i główny autor, zadeklarował swój wkład w postaci opracowania idei rozwiązania, poprzez implementację i przeprowadzenie eksperymentów, a na analizie wyników kończąc.

Kontynuacja badań w przedmiotowym zakresie zaowocowała powstaniem metody *ExpertSim*, opisanej w artykule pt.: *„ExpertSim: Fast Particle Detector Simulation Using Mixture-of-Generative-Experts"*, przedstawionej na renomowanej konferencji ECAI. Praca kontestuje niedoskonałości poprzedniego rozwiązania (*SDI-GAN*) i zwraca uwagę na ograniczenia pojedynczego generatora w symulowaniu złożonych zachowań cząstek fizyki dużych energii. Zaproponowane rozwiązanie bazuje na modularnej architekturze mieszaniny ekspertów (*mixture of experts*), w której każdy ekspert może koncentrować się na innych dystrybucjach odpowiedzi cząstek – złożone zależności pomiędzy własnościami cząstek i wynikami zarejestrowanymi przez detektory. Architektura w pierwszej kolejności wykorzystuje *router* (wielowarstwowa sieć neuronowa z autorską funkcją kosztu) do przekierowania cech cząstki na jeden

z trzech, wyspecjalizowanych modeli generatywnych (*DCGAN – Deep Convolutional Generative Adversarial Network*) pełniących rolę eksperta. Każdy z ekspertów specjalizuje się w modelowaniu wybranych własności cząstek. Autorzy rozbudowali każdego eksperta o odpowiednią funkcję kosztu względem poprzedniej metody *SDI-GAN*, regularyzując różnorodność, intensywność i lokalizację cząstek na obrazie z detektora. Metodę przetestowano na rzeczywistym zbiorze danych, stosując wiarygodną metodologię. Eksperymenty uzupełniono dogłębną analizą hiperparametrów metody. Uzyskane wyniki jednoznacznie wykazują istotną przewagę rozwiązania nad metodami literaturowymi, ale demonstrują również ogromny potencjał skalowania metody na inne eksperymenty fizyki dużych energii. Doktorant, jako drugi autor artykułu, zadeklarował udział w opracowaniu architektury modelu, stworzenie modeli ekspertów, analizę wyników i udoskonalanie metody.

Drugi wątek prac badawczych, opisany w czterech publikacjach, skupiony jest na zabezpieczeniu wybranych modeli uczenia maszynowego przed skopiowaniem oraz próbą odpowiedzi na pytanie, czy zadany zbór danych, lub wręcz konkretny plik obrazu, był wykorzystany do trenowania wybranych modeli generatywnych, co jest związane z zabezpieczeniem własności intelektualnej zbiorów danych, stanowiących współcześnie podstawę uczenia maszynowego.

W pracy zatytułowanej *„Bucks of Buckets (B4B): Active Defenses Against Stealing Encoders"*, opublikowanej na renomowanej konferencji NeurIPS, opisany jest bardzo interesujący i nowatorski sposób aktywnego zapobiegania eksploracji przestrzeni ukrytej enkoderów. Praca wychodzi z założenia, że zwyczajowe wykorzystanie enkoderów odnosi się do wycinka przestrzeni ukrytej, podczas gdy próba eksploracji szerszego zakresu przestrzeni ukrytej jest próbą niepożądanego skopiowania modelu. Szacowanie poziomu pokrycia przestrzeni ukrytej odbywa się za pomocą zaproponowanej funkcji haszującej (ang. *Local Sensitive Hashing*), która haszuje podobne dane z przestrzeni metrycznej w podobne obszary przestrzeni ukrytej (ang. *hash buckets*). W sytuacji podejrzanego użycia modelu, w sposób adaptacyjny wzrasta wartość zaprojektowanej funkcji kosztu, penalizując próby ekstrakcji danych, poprzez wprowadzenie do odpowiedzi szumu Gaussowskiego o różnych charakterystykach, zależnych od wielkości eksploracji przestrzeni ukrytej. Odporność metody na ekstrakcję przestrzeni ukrytej z wielu kont użytkowników uzyskano poprzez losowe transformacje przestrzeni reprezentacji dla każdego użytkownika (m.in. transformacjami: *Affine, Pad, Shuffle, Add, Binary*). Metoda została przetestowana dla dwóch modeli enkoderów (*SimSiam, DINO*) na popularnych zbiorach danych. Niezwykle wartościowa jest optymalizacja hiperparametrów i dyskusja nad konfiguracją metody. Szeroko zdefiniowane testy wykazały uniwersalność funkcji haszującej dla różnych zbiorów danych. Weryfikacja wykazała zarówno odporność metody na skopiowanie modelu przy dużej liczbie zapytań i wiarygodność dla zwykłych użytkowników adresujących rozsądne liczby zapytań. Testy wykazały również uniwersalność metody względem innych zbiorów danych niż te, na których model był uczony. Próba skopiowania modelu z wielu kont również wykazała skuteczność metody *B4B*, obserwowaną poprzez zmniejszenie jakości odpowiedzi. W przeciwieństwie do rozwiązań literaturowych, *B4B* cechuje się wysoką odpornością na kradzieże modelu przy zachowaniu odpowiedniej jakości dla zwyczajnych użytkowników. Doktorant zadeklarował opracowanie mechanizmu monitorowania eksploracji przestrzeni ukrytej, stworzenie pełnego procesu ataku i obrony, implementację eksperymentów oraz analizę wyników.

Kolejna praca, zatytułowana *„Towards more realistic membership inference attacks on large diffusion models"*, opublikowana na konferencji WACV, porusza kwestie wykrywania nieautoryzowanego użycia danych w procesie uczenia modeli generatywnych. Wiąże się to bezsprzecznie z zabezpieczeniem własności intelektualnej, którą często stanowią same dane. Pytanie, na które próbują odpowiedzieć autorzy to, czy konkretny plik został użyty do trenowania modelu dyfuzyjnego (*stable diffusion*). Praca koncentruje się na stworzeniu dedykowanego zbioru danych *LAION-mi*, który w sposób rzetelny

oddziela podzbiory uczące (*members*) i testowe (*non-members*), zapewniając odpowiednią separację zbiorów i adekwatność rozkładów (*deduplication*, *sanitization*). Autorzy przeprowadzają weryfikację skuteczności wykrywania faktu wykorzystania danych w procesie nauki (MIA - ang. *Membership Inference Attack*) metod literaturowych, wykazując szereg niedoskonałości, w tym brak rzetelności i wiarygodności wcześniejszych opracowań. Stosowana metodologia, w tym dobrane miary jakości, są przekonywujące i ciekawe. Eksperymenty są podstawą wykazania, że istniejące audyty wykorzystania danych są wciąż niewystarczająco zbadane i mało wiarygodne. Autorzy wykazali również ciekawy wpływ douczania modelu (*fine-tuning*) zbiorem POKEMON, który istotnie zaburzał wiarygodność wniosków.

Wnioski z pracy badawczej stworzyły odpowiednią kanwę dla dalszych prac badawczych, które w kolejnym artykule, zatytułowanym *„CDI: Copyrighted Data Identification in diffusion models"*, opublikowanym na renomowanej konferencji CVPR, skoncentrowały się na weryfikacji całych zbiorów danych, a nie pojedynczych próbek. Praca bazuje na założeniu, że słabe przesłanki o wykorzystaniu pojedynczych elementów zbioru mogą być zagregowane w mocną i wiarygodną metodę oceny, czy analizowany zbiór był stosowany do nauki wybranego modelu dyfuzyjnego. Metoda bazuje na autorskiej inżynierii cech. Autorzy uzupełniają literaturowe cechy, tj. *Denoising Loss*, *Step-wise Error Comparing Membership Inference score* (*SecMI*), *Proximal Initialization Attack score* (*PIA*, *PIAN*), o cechy analizujące kluczowy obszar ukrytej reprezentacji modelu, w tym maskowanie 20% gradientów o największym wpływie na stratę. Nowe cechy odzwierciedlają rekonstrukcję straty najbardziej istotnych semantycznie regionów przestrzeni ukrytej i intuicyjnie powinny być niewielkie dla danych użytych w uczeniu modelu (*members*). Dla wzmocnienia sygnału/straty predykcji modelu obliczane jest dziesięć kroków dyfuzji. Autorzy przeprowadzają również optymalizację perturbacji zaszumionej reprezentacji ukrytej modelu, która traktowana jest jako kolejna cecha. Dla zaproponowanego zestawu cech obliczana jest regresja logistyczna dla danych z analizowanych podzbiorów. Odpowiednio skonstruowana metodologia sięga do pełnej reprezentacji ukrytej modelu (*white-box*) lub niepełnej reprezentacji (*grey-box*) oraz bazuje na ujawnionej i ukrytej części zbioru, weryfikowanego względem wykorzystania w nauce modelu. Odpowiednio przeprowadzana analiza statystyczna wskaźników regresji (*membership confidence score*) pozwala z dużym prawdopodobieństwem (*Welch t-test*, $p<0.01$) stwierdzić, że próbki ze zbioru testowego były nielegalnie wykorzystane w trenowaniu modelu dyfuzyjnego. Skuteczność i własności metody *CDI* zostały udokumentowane bardzo bogatymi testami i uzupełniającymi analizami, przeprowadzonymi dla ośmiu popularnych modeli dyfuzyjnych. Skoncentrowano się m.in. na minimalnej wielkości zbioru testowego, który warunkuje poprawne wykrycie, różnych konfiguracjach wektora cech, wpływie udziału danych testowych nie wykorzystywanych do nauki na skuteczność weryfikacji, czy odporności metody na weryfikacje fałszywie pozytywne. Doktorant, jako pierwszy autor pracy, zadeklarował istotny udział w opracowaniu metody, implementację modeli, przeprowadzenie eksperymentów oraz analizę wyników.

Kontynuacją badań w zakresie prywatności danych, dla modeli autoregresyjnych, jest praca zatytułowana *„Privacy Attacks on Image AutoRegressive Models"*, która została opublikowana na renomowanej konferencji ICML. Generatywne modele autoregresyjne są interesujące, w kontekście identyfikacji danych źródłowych, z racji na ich rosnące znaczenie spowodowane wydajnością oraz własnościami zapamiętywania danych uczących. Są równocześnie bardziej podatne na wyciek prywatności w porównaniu z modelami dyfuzyjnymi. W identyfikacji danych wykorzystano wektor cech danych wejściowych bazujący na metodzie CLiD, jednak uwzględniający różnicę w odpowiedzi przy różnych warunkach. Dla wybranych modeli autoregresyjnych (MAR) autorzy zaproponowali cenne udoskonalenia, tj. adaptowalna maska binarna, stałe odstępy czasowe odszumiania, czy ograniczona wariancja szumu. Zagregowany sygnał z odpowiedzi MIA, dla danych treningowych, z pomocą testu statystycznego, pozwala sprawdzić czy model był trenowany za pomocą określonego (prywatnego)

zbioru danych. Wnioskowanie w metodzie zostało usprawnione poprzez wyeliminowanie liniowego klasyfikatora cech MIA, co można było zrobić stwierdzając konsekwentne różnice w odpowiedziach dla danych wykorzystywanych do treningu (*members*) i nie wykorzystywanych do treningu (*non-members*). Autorzy zauważyli również możliwość znacznej redukcji próbek identyfikujących dla większych modeli i konieczność większej liczby próbek, zapewniających istotność statystyczną wyników, dla małych modeli autoregresyjnych. Autorzy przeprowadzili też ciekawe eksperymenty ekstrakcji danych uczących z modeli regresyjnych bazując na ich własnościach zapamiętywania oraz stymulowania kontekstu w przestrzeni tokenów. Wyniki pokazują, że zaproponowane rozwiązanie pozwala zrekonstruować znacznie więcej danych (obrazów) uczących niż w rozwiązaniach referencyjnych i wyniki nie są obciążone fałszywie pozytywnie. Doktorant, będąc drugim autorem, zadeklarował istotny udział w tworzeniu mechanizmu audytowania danych i tworzeniu zbioru danych, implementację i przeprowadzenie eksperymentów oraz analizę wyników.

Podsumowując, w obszarze podjętej problematyki, Doktorant rzetelnie, kompleksowo i wieloaspektowo przeprowadził analizę problemów, zaproponował nowatorskie rozwiązania, zaimplementował algorytmy, zaprojektował eksperymenty i przeanalizował wyniki. W każdym z podjętych wątków był autorem przełomowych kontrybucji, wykazując się doskonałą wiedzą, intuicją i rzetelnością badawczą. Przedstawione rozwiązania, eksperymenty oraz wyniki są ciekawe i nowatorskie. Chociaż prace powstawały w zespole badaczy, to w każdej z rozważanych publikacji Doktorant był autorem lub współautorem kluczowych kontrybucji.

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 doktorskiej leży przede wszystkim w zbiorze opracowanych metod, które nie tylko podejmują, analizują, ale również rozwiązują istotne zagadnienia bezpieczeństwa, zastosowania i niezawodności modeli generatywnych. W szczególności najważniejszymi, moim zdaniem, osiągnięciami są:

- opracowanie metody *ExpertSim* zapewniającej odpowiednią dywersyfikację trajektorii cząstek elementarnych, wizualizowanych przez symulator, w eksperymencie ALICE, przeprowadzanym w ośrodku CERN oraz potencjał metody do wizualizowania innych zjawisk fizycznych wysokich energii;

- opracowanie metody *B4B* aktywnego zabezpieczania enkoderów przed skopiowaniem bez zmniejszania użyteczności modelu dla zwykłych użytkowników;

- opracowanie metody CDI do efektywnego audytowania czy zadany zbiór danych był wykorzystany do trenowania modelu dyfuzyjnego;

- opracowanie nowej metody audytowania obrazowych modeli autoregresyjnych, weryfikującej czy dany zbiór danych był wykorzystany do nauki modelu oraz wykazanie, w jakim zakresie możliwe jest zreprodukowanie danych treningowych;

Chociaż wymienione osiągnięcia są, moim zdaniem, najważniejszymi z przedstawionych w dysertacji, to należy podkreślić, że każde z zaprezentowanych rozwiązań stanowi bezsprzecznie oryginalny wkład Doktoranta w rozwój dyscypliny Informatyka Techniczna i Telekomunikacja.

Czy Autor wykazał umiejętność poprawnego i przekonywującego przedstawienia uzyskanych przez siebie wyników (zwięzłość, jasność, poprawność redakcyjna rozprawy)? Jakie są słabe strony rozprawy i jej główne wady?

Praca stanowi bardzo szerokie studium analizy wybranych modeli generatywnych, zarówno w zakresie ich wykorzystania jak i audytu. Z racji przyjętej formy rozprawy, osiągnięcia zostały przedstawione bardzo treściwie, ale wyczerpująco – niezwykle cenne okazały się dodatki do publikacji, które uzasadniały szereg decyzji podjętych podczas opracowywania metod.

Praca jest zredagowana w języku angielskim i oprócz przywołania cyklu sześciu publikacji zawiera ciekawe wprowadzenie do całości jak i do poszczególnych części. Rozprawa zawiera też cenne podsumowania i uzupełnienia opisu badań. Z racji przyjętego kształtu rozprawy doktorskiej uważam, że bardzo trudno jest wskazać jakiekolwiek istotne uchybienia. Uwzględniając zakres prac badawczych ośmielam się stwierdzić, że takich uchybień, w prezentowanym materiale, po prostu nie ma.

Konkluzja

Uważam, że cele rozprawy doktorskiej, pomimo ich różnorodności, zostały w pełni zrealizowane. Doktorant przedstawił w rozprawie nowe badania z zakresu wykorzystania modeli generatywnych w symulacjach fizycznych wysokich energii, mechanizmy zabezpieczenia modeli przed skopiowaniem i metody audytowania modeli względem wykorzystania zbiorów danych w procesie ich nauki. Uzyskane przez Doktoranta wyniki uważam za oryginalne, wartościowe i ciekawe poznawczo. Zakres prowadzonych badań był szeroki, analiza opracowanych rozwiązań wszechstronna, a warsztat badawczy Doktoranta był właściwy. Tym samym rozprawa prezentuje wartościowe osiągnięcia naukowe w obszarze dyscypliny Informatyka Techniczna i Telekomunikacja oraz potwierdza umiejętność prowadzenia przez Doktoranta pracy naukowej.

Bez najmniejszej wątpliwości stwierdzam, że recenzowana rozprawa doktorska Pana mgr inż. Jana Dubińskiego spełnia z wyraźnym nadmiarem wymagania stawiane rozprawom doktorskim, przez obowiązującą ustawę. Wnoszę o jej przyjęcie i dopuszczenie rozprawy do publicznej obrony.

Wnoszę również o wyróżnienie rozprawy, gdyż zakres i jakość prezentowanych, oryginalnych rozwiązań naukowych przekracza zdecydowanie przeciętny poziom osiągnięć naukowych, uzyskiwanych przez doktorantów w większości znanych mi postępowań doktorskich. Na wyróżnienie zasługuje również dorobek publikacyjny, który prezentuje się spektakularnie i bezsprzecznie wymagał ogromnej wiedzy i determinacji Doktoranta.

Adam Wojciechowski

Prof. dr hab. inż. Bogusław Cyganek
Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie
Wydział Informatyki, Elektroniki i Telekomunikacji
Instytut Elektroniki
cyganek@agh.edu.pl

Kraków, 23.03.2026

Recenzja rozprawy doktorskiej
Pana magistra inżyniera Jana Michała Dubińskiego
zatytułowanej:
Reliable and Safe Generative Models

Wstęp

Recenzja rozprawy doktorskiej Pana magistra inżyniera Jana Dubińskiego pt. „*Reliable and Safe Generative Models*[1]", która powstała w roku 2025 na Politechnice Warszawskiej. Promotorem pracy jest Pan prof. dr hab. inż. Przemysław Rokita. Przygotowanie recenzji zostało wykonane na zlecenie Przewodniczącego Rady Dyscypliny Naukowej Informatyka Techniczna i Telekomunikacja, Pana prof. dra hab. inż. Jarosława Arabasa. Rozprawa doktorska została napisana po angielsku. Recenzja została przygotowana w formie odpowiedzi na pytania dotyczące rozprawy doktorskiej Pana mgra inż. Jana Dubińskiego.

1. Problem badawczy oraz jego znaczenie

Jaki jest najważniejszy problem rozważany w rozprawie?

Rozprawa doktorska Pana mgra inż. Jana Dubińskiego wpisuje się w nowoczesną dziedzinę współczesnej nauki związanej z rozwojem sztucznej inteligencji (ang. *artificial intelligence* – AI) i dotyczy bardzo istotnych tzw. modeli generatywnych. Są to systemy AI zdolne do tworzenia nowych treści, takich jak teksty czy obrazy, a jednocześnie spełniających ściśle określone założenia, takie jak statystyczna zgodność z danymi użytymi do ich trenowania. Modele tego typu zyskały wielkie zainteresowanie właśnie ze względu na możliwości twórcze, a nie tylko analityczne, takie jak dla przykładu klasyfikacje, czy detekcje obiektów w obrazach. Dzięki modelom generatywnym możliwe jest więc tworzenie nowych utworów muzycznych, tekstów, obrazów wideo, czy też symulacji zjawisk fizycznych – te ostatnie stanowią jeden z głównych obszarów zainteresowania Doktoranta. Możliwości generatywnego AI otwierają olbrzymie i wcześniej nie spotykane możliwości stosunkowo łatwego i szybkiego w realizacji tworzenia danych multimedialnych. Oczywiście, i jak w przypadku większości nowoczesnych technologii, oprócz niewątpliwych walorów pozytywnych, istnieją też liczne zagrożenia, zarówno dla użytkowników, jak i samych twórców modeli oraz właścicieli danych, za pomocą których modele te są uczone, nie zawsze za zgodą, czy też nawet wiedzą tych ostatnich. Również tego typu

[1] Niezawodne i bezpieczne modele generatywne

zagadnieniom szeroko rozumianej ochronnych wartości intelektualnej w modelach uczenia głębokiego poświęcona jest znaczna część rozprawy doktorskiej Pana mgra inż. Jana Dubińskiego.

Czy praca ma charakter naukowy?

Nie ulega wątpliwości, że praca doktorska Pana mgra inż. Jana Dubińskiego ma charakter naukowy. Tak jak już wspomniano, dotyczy ona głównie najnowszych metod tzw. generatywnej AI, a w szczególności opracowaniu takich modeli, które w najlepszy znany sposób mogą służyć do symulacji eksperymentów z zakresu fizyki wysokich energii, takich jakie są przeprowadzane np. w Wielkim Zderzaczu Hadronów CERN. W rozprawie podjęta została również tematyka problemów związanych z zapewnieniem wartości intelektualnej oraz jej ochrony w tego typu modelach generatywnych. Są to zagadnienia nie tylko będące jednym z pierwszoplanowych przedmiotów współczesnej nauki, ale również dotyczą jej wielu różnych dyscyplin, takich jak informatyka i fizyka.

Czy praca ma znaczenie praktyczne?

Oprócz niewątpliwych walorów czysto naukowych, praca doktorska Pana mgra inż. Jana Dubińskiego ma niebagatelne znaczenie praktyczne. Opracowane przez Doktoranta metody oraz algorytmy miały na celu rozwiązanie istotnych zagadnień praktycznych, takich jak m.in. efektywna symulacja eksperymentów fizyki cząstek wysokich energii. Również zagadnienia dotyczące szeroko rozumianego bezpieczeństwa oraz prywatności danych w kontekście AI mają olbrzymie znaczenie praktyczne, które niewątpliwie będzie tylko rosnąć wraz z dynamicznym rozwojem AI.

2. Wkład Autora

Jaki jest najważniejszy wkład Autora rozprawy?

Najważniejsze osiągnięcia naukowe przedstawione w rozprawie Pana magistra inżyniera Jana Dubińskiego można podzielić na dwa główne zagadnienia dotyczące modeli generatywnych:

A. *Metody rozszerzające możliwości generatywnych metod AI*.

B. *Metody ochrony wartości intelektualnej w modelach uczenia głębokiego*.

Każda z powyższych grup została dodatkowo podzielona na bardziej szczegółowe metody dotyczące rozwiązania specyficznych problemów naukowych, które zostaną przedstawione w poniższym zestawieniu.

A.1. *Metoda selektywnego zwiększania różnorodności próbek generowanych przez GAN*.

Metoda selektywnego zwiększania różnorodności próbek generowanych przez GAN (ang. *generative adversarial networks*) została opracowana głównie do umożliwienia symulacji procesów zachodzących podczas eksperymentów fizyki wysokich energii w Wielkim Zderzaczu Hadronów (LHC) w ośrodku CERN. Dotychczas stosowane metody symulacji detektorów cząstek wykorzystywały metody typu Monte Carlo, które mimo iż o uznanej renomie, to jednak wymagają niebagatelnych nakładów obliczeniowych. W tym kontekście, generatywne metody AI mogą stanowić ich istotną alternatywę. Tym niemniej, wyzwaniem jest tutaj niezawodność tego typu rozwiązań, które muszą uwzględniać zmienne warunkowe, które istotnie wpływają na odpowiedzi detektora padających cząstek. Jak zauważył Doktorant,

problemem jest tu moduł generatora, który wykazuje skłonność do ignorowania szumów wejściowych i koncentruje się wyłącznie na zmiennych warunkujących, generując ograniczony zakres wyników. W rezultacie uniemożliwia to modelowi uchwycenie pełnej zmienności danych zderzeniowych, co zmniejsza jego użyteczność w zadaniach symulacyjnych. Dzięki szczegółowej analizie niedoskonałości istniejących metod generatywnych, które mają tendencję do operowania tak jakby rozkład zmiennych warunkowych był jednorodny, Pan mgr inż. Jan Dubiński wraz ze współautorami zdołał zaproponować oryginalną modyfikację funkcji kosztu, która pozwala obejść te ograniczenia. Jest to metoda nazwana Selective Diversity GAN (SDI-GAN), która umożliwia regularyzację generatora, która z kolei wymusza różnorodność generowanych próbek w sposób zależny od samych danych treningowych. Metoda ta umożliwia więc skalowanie efektu różnorodności generowanych próbek zgodnie z wariancją zmiennych warunkujących, co w efekcie wpływa na generowanie próbek bardziej różnorodnych i realistycznych, a jednocześnie bez utraty dokładności w stosunku do rozkładów danych treningowych. Pozytywne efekty takiego podejścia zostały zaprezentowane w realistycznym przypadku symulacji tzw. kalorymetru zerowego w eksperymencie ALICE.

Metoda ta została opublikowana w roku 2022 na konferencji International Conference on Neural Information Processing (ICONIP), w publikacji pt. „*Selectively increasing the diversity of GAN-generated samples*”, w której Pan mgr inż. Jan Dubiński jest pierwszym z autorów.

A.2. *Metoda symulacji detektora szybkich cząstek z wykorzystaniem mieszanki ekspertów generatywnych – ExpertSim*.

Jest to kontynuacja i udoskonalenie metody zaprezentowanej w poprzednim zadaniu (A.1). Okazało się bowiem, że mimo iż zaprezentowana tam metoda wykazuje świetne rezultaty, to są one ciągle niewystarczające do precyzyjnej symulacji procesów zachodzących pod wpływem multimodalnych danych fizyki wysokich energii. Tak złożone oddziaływania cząstek w detektorach prowadzą do rozkładów, które trudno aproksymować wyłącznie za pomocą tylko jednego modelu generatywnego. To co Pan mgr inż. Jan Dubiński zaproponował wraz z zespołem, to zastąpienie jednego takiego modelu zespołem w postaci złożenia współpracujących modeli. W tym podejściu, zamiast trenować pojedynczy generator obejmujący całą przestrzeń odpowiedzi cząstek, trenuje się kilka generatorów równocześnie, tyle że każdy z nich koncentruje się na innej części rozkładu charakterystycznego dla eksperymentów fizyki wysokich energii. Oryginalnym rozwiązaniem zaproponowanym przez Pana mgra inż. Jana Dubińskiego ze współautorami jest specjalny trenowany moduł rozdzielająco-kontrolny – tzw. ruter – który łączy dane wejściowe z najbardziej odpowiednim dla nich ekspertem. W rezultacie system ten w znacznie lepszy sposób może uchwycić różnorodność odpowiedzi detektorów. Z kolei, każdy z indywidulanych ekspertów w tym systemie, to już wcześniej opracowany SDI-GAN (A.1). Jak zostało to pokazane za pomocą eksperymentów, system ten osiąga wyższą dokładność w odtwarzaniu rozkładów eksperymentalnych, zapewniając jednocześnie znaczne przyspieszenie w porównaniu z tradycyjnymi symulacjami typu Monte Carlo.

Metoda ta została opublikowana na prestiżowej konferencji European Conference on Artificial Intelligence ECAI 2025 w publikacji pt. „*ExpertSim: Fast Particle Detector Simulation Using Mixture-of-Generative-Experts*”, w której Pan mgr inż. Jan Dubiński jest drugim z autorów.

B.1. *Metoda aktywnej ochrony przed kradzieżą enkoderów – Bucks for Buckets* (*B4B*).

Jest to pierwsza z serii metod ochrony wartości intelektualnej systemów powstałych w wyniku kosztownego trenowania modeli AI. Jednym z rosnących zagrożeń jest kradzież modelu – możliwe jest to np. poprzez odpowiednią interakcję z systemem za pośrednictwem dostępnego interfejsu API, za pomocą którego próbuje się zbudować substytut modelu – rodzaj modelu-modelu – który ściśle imituje zachowanie modelu oryginalnego. Pan mgr inż. Jan Dubiński wraz z zespołem zauważyli, że istniejące sposoby obrony przed tego typu kradzieżami są w dużej mierze pasywne, koncentrując się na ograniczaniu dostępu do zapytań lub dodawaniu szumu. Jednakże zawodzą one w konfrontacji z pomysłowymi hakerami, którzy potrafią dynamicznie przystosować i zaadaptować swoje strategie ataku. Dzięki tej obserwacji Pan mgr inż. Jan Dubiński wraz z zespołem zaproponowali metodę Bucks for Buckets (B4B). Jej główna idea to monitorowanie zapytań użytkowników w celu określenia stopnia eksploracji/pokrycia przestrzeni tzw. osadzeń (ang. *embeddings*) danego modelu. Sama idea jest tu dość prosta – jeśli ten stopień „przeszukiwania" jest zbyt rozległy, to może to świadczyć właśnie o próbie wyłudzenia wewnętrznej struktury modelu w celu jego podrobienia. Po zidentyfikowaniu tego typu podejrzanego zachowania, w zaproponowanej metodzie nakładane są dodatkowe koszty adaptacyjne, które penalizują próby ekstrakcji, w rezultacie uniemożliwiając stworzenie repliki modelu. Dodatkowo, aby przeciwdziałać atakom z wykorzystaniem wielu kont, Pan mgr inż. Jan Dubiński zaproponował transformacje przestrzeni reprezentacji dla każdego użytkownika, co uniemożliwia koordynację między kontami, ale jednocześnie jest niezauważalne dla uczciwego użytkownika. Prawidłowe działanie metody zostało wykazane eksperymentalnie z wykorzystaniem takich zbiorów jak FashionMNIST, SVHN, STL10 oraz CIFAR10, jak również enkoderów typu SimSiam oraz DINO.
Metoda ta została opublikowana na prestiżowej konferencji: Conference on Neural Information Processing Systems (NeurIPS) 2023, pt. „*Bucks for Buckets (B4B): Active Defenses Against Stealing Encoders*". Pan mgr inż. Jan Dubiński jest pierwszym z autorów.

B.2. *Metoda ochrony przed atakami wglądu w dane w wielkich modelach dyfuzyjnych*.

Jest to kolejna z metod dotyczących ochrony modeli AI, poszerzona jednak o ochronę danych na podstawie których są one trenowane. Główny problem to stwierdzenie, czy zbiór danych został wykorzystany w treningu modelu, co może się wiązać właśnie z naruszeniem praw autorskich. Jednakże nie jest to zadanie ani łatwe, ani proste w realizacji. Głównym problemem jest tu przeważnie olbrzymia liczba danych używanych do trenowania modeli. W tym przypadku stwierdzenie, czy dana próbka rzeczywiście została użyta w procesie trenowania i w istocie „należy" ona do modelu (jest zapisana w jego wagach) jest bardzo trudne. W tym kontekście Pan mgr inż. Jan Dubiński wraz z zespołem analizowali ataki oparte na ocenie przynależności danych do modeli dyfuzyjnych. Pierwsze osiągnięcie to krytyczne spojrzenie na dotychczasowe metody ewaluacji tego typu metod i zaproponowanie nowych sposobów ewaluacji, które są dostosowane do skali nowoczesnych modeli generatywnych. Kluczowe jest tutaj opracowanie specjalnego zbioru danych do oceny metod wykrycia ataków tego typu. Dzięki opracowaniu tej platformy badawczej Pan mgr inż. Jan Dubiński wraz z zespołem przeprowadzili liczne eksperymenty w celu oceny metod wnioskowania. Wyniki te sugerują, że ataki tego typu są mniej skuteczne niż było to sugerowane w pracach innych autorów.

Metoda została opublikowana na konferencji IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) 2023, w pracy pt. „*Towards More Realistic Membership Inference Attacks on Large Diffusion Models*", w której Pan mgr inż. Jan Dubiński jest pierwszym z autorów.

B.3. *Metoda identyfikacji łamania prawa własności danych w modelach dyfuzyjnych*.

Mimo iż wcześniej opracowane metody detekcji ataków zostały zweryfikowane w wybranych zadaniach, Pan mgr inż. Jan Dubiński wraz z zespołem zauważyli, że nie dostarczają one wiarygodnych wyników w przypadku systemów o większej skali. W tych przypadkach, nie jest jasne w jaki sposób można zweryfikować czy czyjeś zbiory danych zostały użyte, czy też nie, do treningu danego modelu generatywnego. W tego typu zadaniach, ze względu na olbrzymie ilości danych użytych do trenowania bardziej naturalne – jak zauważył Pan mgr inż. Jan Dubiński – jest pytanie o obecność całych zbiorów danych, a nie pojedynczych próbek. Przy tych założeniach Doktorant zaproponował więc nową metodę identyfikacji danych chronionych prawami autorskimi – nazwaną Copyrighted Data Identification (CDI) – która łączy wartości przynależności do modelu nie z jednej, lecz z wielu próbek, a następnie dokonuje ich analizy za pomocą testów statystycznych. CDI została następnie zweryfikowana eksperymentalnie, pokazując że osiąga ona wysoką niezawodność w wykrywaniu zbiorów danych chronionych prawem autorskim, znacznie lepszą od metod już istniejących. Istotne jest tutaj też to, że możliwe jest to bez dostępu do samego modelu, co czyni ją szczególnie atrakcyjną w praktyce, gdzie modele generatywne są często dostępne za pośrednictwem interfejsów API typu black-box.

Metoda ta została opublikowana na prestiżowej konferencji: The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2025, w pracy pt. „*CDI: Copyrighted Data Identification in Diffusion Models*". W publikacji tej Pan mgr inż. Jan Dubiński jest pierwszym z autorów.

B.4. *Metoda ochrony przed atakami na prywatność modeli auto-regresyjnych*.

Jest to kolejna z opracowanych przez Pana mgra inż. Jana Dubińskiego oraz współpracowników metod dotyczących wiarygodnego wykrywania użycia danych, często chronionych prawami autorskimi, do trenowania dużych modeli, nie tylko dyfuzyjnych, ale i nowszych tzw. modeli autoregresyjnych. Te ostatnie nie dokonują procesu odszumiania w celu wykrycia ukrytych danych, ale generuje próbki w sposób sekwencyjny na podstawie tzw. tokenów. Proces ten, przebiegający w następujących po sobie krokach, ułatwia jednak zapamiętywanie danych treningowych. Pan mgr inż. Jan Dubiński wraz z zespołem badali zagrożenia w modelach autoregresyjnych. Dzięki określeniu procesu zapamiętywania przez te modele danych treningowych, zdołali opracować nowe metody, które dokonują ekstrakcji treści wprost z wyjść modeli autoregresyjnych. Badacze zaobserwowali tutaj, że modele autoregresyjne są szczególnie podatne na swoisty wyciek prywatności, często przekraczający podatność modeli dyfuzyjnych. W rezultacie, opracowane metody, które dokonują agregacji wartości przynależności (ang. *membership inference*) pomiędzy wieloma próbkami danych, umożliwiają skuteczną detekcję nieautoryzowanego użycia danych objętych prawami autorskimi. Właściwości metody zostały zbadane eksperymentalnie, które potwierdziły podejrzenia że modele oparte na autoregresji są znacznie bardziej narażone na nadużycie prywatności niż ma to miejsce w przypadku modeli dyfuzyjnych.

Opisana metoda została opublikowana na prestiżowej konferencji: International Conference on Machine Learning (ICML) 2025, pt. „*Privacy Attacks on Image AutoRegressive Models*". Jan Dubiński jest tu drugim autorem.

Wyżej wymienione oryginalne osiągnięcia naukowe Pana mgra inż. Jana Dubińskiego są wysoce nowatorskie i dotyczą najnowszych i często jeszcze dobrze nie zbadanych zagadnień generatywnych modeli AI. Świadczy to o głębokiej wiedzy oraz pomysłowości, jak również dojrzałości naukowej Doktoranta. Opracowanie przez Niego oraz współpracujący zespół badawczy metody dotyczą kluczowych zagadnień naukowych współczesnej informatyki. Metody te zostały opublikowane na wiodących konferencjach naukowych dotyczących AI. Zostały one również dogłębnie zweryfikowane eksperymentalnie. Często też został dołączony kod, który umożliwia weryfikację wyników, jak również dalsze rozszerzenie proponowanych metod przez inne zespoły.

3. Poprawność naukowa

Czy stwierdzenia zawarte w rozprawie są godne zaufania? Czy uzasadnienia są poprawne? Wskaż zauważone słabości i błędy. Wskaż także te aspekty dotyczące poprawności, które są najbardziej wartościowe.

Pan mgr inż. Jan Dubiński podjął uzasadnioną naukowo metodykę prowadzenia badań bazującą na wnikliwej analizie problemów oraz ich rozwiązań opublikowanych przez inne zespoły badawcze, a następnie na analizie ich ograniczeń. Bazując na tej wiedzy Doktorant opracował rozwiązania nowsze i konkurencyjne do już istniejących. Wszystkie w ten sposób opracowane metody, które głównie dotyczyły generatywnych modeli AI, zostały przez Pana mgra inż. Jana Dubińskiego dogłębnie zweryfikowane eksperymentalnie. Podejście takie oceniam jako jak najbardziej poprawne z naukowego punktu widzenia.

W rozprawie Pana mgra inż. Jana Dubińskiego natknąłem się jednak na pewne stwierdzenia oraz zagadnienia, które wymagają albo doprecyzowania, albo też dalszego rozszerzenia.

1. Doktorant wraz z zespołem poświęcili wiele uwagi opracowaniu metod detekcji prób kradzieży, czy też podrobienia wytrenowanego modelu generatywnego, jak również nieautoryzowanego użycia danych do trenowania modeli tego typu. Tymczasem, i jak jest to wielokrotnie raportowane w licznych pracach naukowych, a nawet doniesieniach medialnych, zagrożenie istnieje również ze strony samych generatywnych modeli AI, a być może jest jeszcze bardziej istotne ze społecznego punktu wiedzenia. Chodzi tu głównie o możliwość łatwego i szybkiego tworzenia – generowania – fałszywych treści, które jednak do złudzenia przypominają, czy też udają prawdziwe. Są to tzw. *fake news*, takie jak zmontowane obrazy, czy też treści multimedialne mające właśnie na celu wprowadzenie w błąd osoby, do których są adresowane. Jednakże Autor w ogóle nie wspomina o tym fakcie, mimo iż sama rozprawa jest bardzo obszerna; zawiera też i to wielokrotnie przytaczane przeglądy prac naukowych w tej dziedzinie. Warto więc aby Doktorant przybliżył te właściwości sieci generatywnych, a korzystając ze swojej wiedzy na ten temat, przybliżył też możliwości opracowywania metod identyfikacji tego typu *fake news*, czy też innych falsyfikacji danych.
2. Jak już wielokrotnie podkreśliłem przy ocenie osiągnięć Doktoranta, wszystkie one zostały opracowane w ścisłej współpracy z dość licznym zespołem badawczym. Zaprezentowane w rozprawie metody zostały opublikowane w pracach wielo-autorskich, w których Doktorant

jest pierwszym, a w kilku przypadkach – drugim, z autorów. Rozprawa zawiera również opis oryginalnych osiągnięć Pana mgra inż. Jana Dubińskiego. Tym niemniej, nieco brakuje mi jaśniejszego określenia, które dokładnie metody zostały oryginalnie *wymyślone* przez Doktoranta, a które przez innych członków zespołu.

3. Str 41, wzór (4.3) – nie jest jasne czym X_c różni się od $\mathcal{X}_c$ (tzn. pisanego czcionką kaligrafowaną).
4. Str. 45, rozdz. 4.4.3 – warto by właściwości metody sprawdzić na jakiś innych danych, na których inni badacze sprawdzają jakość GAN, np. MS-GAN itd.
5. Str. 46, rys. 4.4.4 („diverse results") –mimo wszystko są to rezultaty dość różne, przynajmniej wizualnie, od tych "real data" (kształt, położenie itd.), więc chyba trudno tu jedną metodę np. SDI-GAN specjalnie wyróżniać?
6. Str. 47, rozdz. 4.5 – w ewaluacji eksperymentalnej tej metody brakuje mi tzw. *ablation study*; brakuje analizy choćby wpływu parametru λ ze wzoru (4.5).
7. Str. 53, rozdz. 5.2.2 – Autor nie sięga do źródeł metod Mixture-of-Experts, ograniczając się tylko do ery deep learningu. Skoro już pisać rozdział tego typu, to warto by jednak sięgnąć do prac będących u podstaw tego typu metod.
8. Str. 56,– wzór (5.1) chyba jest to samo co (4.4) ale – dla utrudnienia – nieco inaczej zapisane?
9. Str. 57 – wzór (5.5) dobrze by było porównać z (4.5); co tu nowego, czemu?
10. Str. 57 – nie jest jasne jak router jest trenowany; supervised, unsupervised, czy też inaczej? Nie jest też jasne jak jest wyznaczany „most suitable expert" – jeśli był uczony w trybie nadzorowanym, to jak były oznaczane dane?
11. Str. 58 – wzór (5.8), jeśli system był trenowany w trybie end-to-end to już różnych parametrów λ jest kilkanaście, czego jednak Autor nie przedyskutował.
12. Str. 61, rys. 5.5.2 – co to jest „true" zaprezentowane na wykresach? Trudno też coś powiedzieć o samej metodzie ExpertSim z samej obserwacji tych wykresów.
13. Str. 62 – parametry λ są dobierane w procesie, które Doktorant opisał jako „*tuning efforts*", ale może dałoby się i te parametry wytrenować, skoro metoda i tak jest end-to-end? Z drugiej strony może te λ routera można by wpisać "wmnożyć" w wagi modelu i tak trenować?
14. Str. 95 – lepszego wyjaśnienia wymaga odległość FID, dlaczego została użyta ta właśnie?
15. Str. 101, tabela 7.7.1 – jest pewien problem z tymi eksperymentami, bo Autorzy porównują je tylko sami ze sobą; Czy nie ma innych badań w tej dziedzinie, do których można by się porównać?
16. Str.112 – wyjaśnienia wymaga stwierdzenie „*observing that training data representations exhibit a markedly different distribution from test data*", choćby w kontekście trenowania modeli z wykorzystaniem tzw. multi-fold. Czy w tym przypadku, gdy dane wielokrotnie losowo dzielimy na fold treningowy oraz testowy, to rzeczywiście mamy do czynienia z tak różnymi dystrybucjami?
17. Str. 155 i kolejne – Bardzo obszerna bibliografia, która liczy 267 pozycje. Jednakże wiele z nich niekompletnych, np. [8] [24] [50] itd. , bądź też w ogóle bezużytecznych [70] [185].

Należy tutaj jednoznacznie stwierdzić, że powyższe pytania i sugestie poszerzenia dodatkowych zagadnień mają wyłącznie charakter dyskusyjny i w żadnym stopniu nie umniejszają istotnego wkładu merytorycznego przedstawionego w rozprawie Pana mgra inż. Jana Dubińskiego.

4. Wiedza Kandydata

Które z rozdziałów rozprawy omawiają istniejący stan wiedzy i dzięki temu potwierdzają ogólny stan wiedzy kandydata w zakresie informatyki? Jakie obszary tych dyscyplin zostały omówione w tych rozdziałach/sekcjach? Jaka jest opinia recenzenta o bibliografii? Prosimy o podanie innych argumentów za lub przeciw, że kandydat posiada ogólną wiedzę w dyscyplinie ITT.

Rozprawa doktorska Pana mgra inż. Jana Dubińskiego to dzieło bardzo obszerne jak na prace tego typu – liczy aż 250 stron. Składa się ona z 10 rozdziałów głównych, spisu publikacji oraz suplementów. Najważniejsze jednak jest to, że zaprezentowany w rozprawie materiał naukowy jest bardzo obszerny i merytorycznie bogaty. Zaprezentowane metody zostały już opublikowane w 6 publikacjach na czołowych i wysoce punktowanych konferencjach ML/AI, takich jak NeurIPS 2023, ECAI 2025, czy też CVPR 2025 (dokładny spis tych publikacji znajduje się na str. 152 rozprawy). W 4 z tych 6 publikacji Pan mgr inż. Jan Dubiński jest pierwszym autorem. Natomiast w 2 pozostałych, Doktorant jest drugim z autorów. Mimo iż są one wieloautorskie, to najwyższy światowy poziom tych publikacji świadczy też o niekwestionowanej wiedzy Pana mgra inż. Jana Dubińskiego w dyscyplinie informatyka techniczna i telekomunikacja.

Pełniejsza lista istotnych publikacji naukowych, których jednym z autorów jest Pan mgr inż. Jan Dubiński jest bardzo imponująca, uwzględniwszy młody wiek Doktoranta. Baza Web of Science przytacza 11 publikacji tego typu, które są cytowane 26 razy, a tzw. indeks Hirscha Pan mgr inż. Jan Dubiński wynosi 3. Są to wysokie wartości, ale szczególnie istotne są tutaj wysoko punktowane konferencje.

Wszystko powyższe świadczy o ponadprzeciętnej wiedzy Doktoranta Pana magistra inżyniera Jana Dubińskiego w dyscyplinie informatyka techniczna i telekomunikacja, a w szczególności w dziedzinie metod uczenia maszynowego i sztucznej inteligencji.

5. Podsumowanie

Rozprawa doktorska Pana magistra inżyniera Jana Dubińskiego prezentuje ogólną i głęboką wiedzę Kandydata w dyscyplinie informatyka techniczna i telekomunikacja, a w szczególności w zakresie nowoczesnych metod sztucznej inteligencji. Świadczy też o umiejętność samodzielnego prowadzenia badań naukowych na najwyższym światowym poziomie, jak również o dobrym opanowaniu warsztatu badawczego. Zaprezentowane w rozprawie doktorskiej Pana magistra inżyniera Jana Dubińskiego metody i algorytmy są oryginalne i dotyczą rozwiązania istotnych problemów naukowych dotyczących generatywnej sztucznej inteligencji. Opracowane metody stanowią oryginalny i własny wkład Doktoranta Pana magistra inżyniera Jana Dubińskiego w rozwój dyscypliny naukowej informatyka techniczna i telekomunikacja.

Recenzowaną pracę oceniam jako z spełniającą ze znacznym nadmiarem formalne oraz zwyczajowe wymagania stawiane rozprawom doktorskim. Wnioskuję o jej przyjęcie oraz o dopuszczenie Pana magistra inżyniera Jana Dubińskiego do publicznej obrony.

Zważywszy również na bardzo wysoki poziom naukowy opracowanych metod oraz bardzo cenne publikacje, w których Pan mgr inż. Jan Dubiński jest pierwszym z autorów, **wnioskuję o wyróżnienie rozprawy**.

Rafał Scherer
Katedra Sztucznej Inteligencji
Wydział Informatyki i Sztucznej Inteligencji
Politechnika Częstochowska
Wydział Informatyki
Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie

Częstochowa, dn. 2 kwietnia 2026 r.

Recenzja

rozprawy doktorskiej mgr inż. Jana Dubińskiego, pt.: Reliable and Safe Generative Models.

Niniejszą recenzję opracowano zgodnie z uchwałą nr 146/2025 Rady naukowej dyscypliny informatyka techniczna i telekomunikacja PW. Promotorem jest prof. dr hab. inż. Przemysław Rokita.

1. Charakterystyka tematu, celu i tezy badawczej rozprawy

Praca wpisuje się w dynamicznie rozwijający się obszar głębokich modeli generatywnych, łącząc zagadnienia ich wiarygodnego zastosowania w symulacjach fizycznych z problematyką bezpieczeństwa modeli i ochrony danych treningowych. Rozprawa ma charakter monograficzny, oparty na cyklu publikacji i obejmuje zarówno oryginalne propozycje metod generatywnych dla zastosowań w fizyce wysokich energii, jak i nowe rozwiązania z zakresu obrony przed kradzieżą modeli oraz identyfikacji danych użytych w treningu dużych modeli dyfuzyjnych i autoregresyjnych. Znaczenie tej rozprawy wynika z dwóch nakładających się trendów: gwałtownego rozwoju dużych modeli generatywnych oraz ich coraz szerszego użycia w zadaniach od symulacji w fizyce wysokich energii po komercyjne systemy wizyjne. Z jednej strony rośnie więc potrzeba wiarygodnych i wydajnych modeli, które można zastosować np. jako zamiennik kosztownych symulacji; z drugiej, te same modele kumulują ogromną wartość intelektualną i prawną, co rodzi pytania o ich ochronę oraz o prawa właścicieli danych treningowych.

2. Zawartość rozprawy

Recenzowana praca mgr inż. Jana Dubińskiego składa się z dziesięciu rozdziałów, wykazu publikacji, bibliografii oraz załączników. Dokument liczy 250 stron.

Pierwszy rozdział przedstawia ogólny kontekst pracy, w której autor zajmuje się zarówno rozwojem wiarygodnych modeli generatywnych dla zastosowań naukowych, jak i ochroną wartości intelektualnej zawartej w modelach oraz danych treningowych. Na początku pokazuje, że współczesne modele generatywne (GAN, modele dyfuzyjne i autoregresyjne) potrafią bardzo dobrze przybliżać złożone rozkłady danych i są już szeroko stosowane w generowaniu obrazów, wideo, mowy czy muzyki, a także w symulacjach fizycznych, projektowaniu leków, medycynie obliczeniowej i nauce o materiałach. Podkreśla, że ta zdolność generowania realistycznych próbek niesie zarówno potencjał, jak i ograniczenia: wciąż trudno jest budować modele, które wychodzą poza proste benchmarki i naprawdę wiernie odwzorowują złożone rozkłady z rzeczywistych zastosowań, a równocześnie rośnie problem ochrony wartości zakodowanej w samych modelach oraz w danych użytych do treningu.

W dalszej części autor zawęża uwagę do dwóch głównych wątków pracy: budowy niezawodnych modeli generatywnych dla fizyki wysokich energii w CERN oraz ochrony modeli i danych przed nadużyciami. W pierwszym obszarze koncentruje się na symulacji odpowiedzi detektorów w Wielkim Zderzaczu Hadronów i proponuje metody oparte na generatywnych sieciach adwersarialnych, które mają zastępować kosztowne symulacje Monte Carlo. Najpierw przedstawia metodę kontrolowanego zwiększania różnorodności próbek generowanych przez GAN, aby lepiej dopasować się do złożonych, wielomodalnych rozkładów danych z kolizji cząstek. Następnie proponuje generatywny model typu mixture-of-experts, który lepiej opisuje wielomodalność danych z eksperymentów wysokich energii.

Kolejna część wprowadzenia pokazuje, że gdy takie modele zaczynają rozwiązywać realne zadania naukowe, stają się jednocześnie cennym zasobem intelektualnym, bo ucieleśniają duże nakłady pracy, danych i mocy obliczeniowej. Autor argumentuje, że w takiej sytuacji samo projektowanie skutecznych modeli nie wystarcza, gdyż trzeba także zadbać o ich ochronę przed nieautoryzowanym kopiowaniem i wykorzystaniem. Stąd drugi duży wątek rozprawy: aktywna ochrona modeli i danych. W części poświęconej modelom autor zapowiada pierwszą aktywną metodę obrony przed kradzieżą enkoderów, polegającą na wykrywaniu intensywnego sondowania przestrzeni odpowiedzi modelu i blokowaniu prób ekstrakcji. W części dotyczącej danych wskazuje na słabości istniejących technik identyfikacji danych treningowych w dużych modelach dyfuzyjnych i proponuje metodę wiarygodnego wykrywania wykorzystania chronionych prawem autorskim zbiorów, a następnie rozszerza ją na nowe modele autoregresyjne obrazów, analizując ich podatność na wycieki prywatności.

W podrozdziale dotyczącym celów badawczych autor formułuje trzy główne zadania: zaprojektowanie wiarygodnych modeli generatywnych dla konkretnych problemów, takich jak symulacje w fizyce wysokich energii, opracowanie zabezpieczeń chroniących wartość intelektualną modeli przed nieautoryzowaną ekstrakcją oraz ochrona wartości danych treningowych i praw ich właścicieli w erze dużych modeli generatywnych. Następnie krótko opisuje, jak poszczególne części pracy realizują te cele: rozdziały 4-5 odpowiadają za część symulacyjną, rozdział 6 za aktywną obronę modeli, a rozdziały 7-9 za ochronę danych i analizę zagrożeń związanych z prywatnością informacji.

Ostatnia część wprowadzenia szkicuje strukturę i logikę dalszych rozdziałów. Autor wyjaśnia, że najpierw przedstawia tło teoretyczne dotyczące modeli generatywnych, szybkich symulacji w CERN oraz ataków i technik ochrony modeli i danych, a także przegląd literatury. Następnie w pierwszej części omawia zastosowanie modeli generatywnych do szybkich symulacji w ALICE i pokazuje ograniczenia standardowych modeli cGAN, wprowadzając SDI-GAN oraz ExpertSim. W drugiej części przechodzi do obrony enkoderów przed kradzieżą danych i opisuje metodę Bucks for Buckets, która monitoruje pokrycie przestrzeni zanurzeń (embedding) przez zapytania i wprowadza kontrolowane zakłócenia, utrudniając ekstrakcję przy zachowaniu użyteczności dla uczciwych użytkowników. W trzeciej części analizuje ochronę danych treningowych w modelach dyfuzyjnych i autoregresyjnych: krytycznie ocenia skuteczność klasycznych ataków inferencji członkostwa, proponuje ramy oceny bardziej realistycznych ataków, wprowadza metodę CDI do identyfikacji chronionych zbiorów w modelach dyfuzyjnych i pokazuje, że modele autoregresyjne obrazów wykazują silną tendencję do zapamiętywania danych treningowych. Rozdział kończy zapowiedź, że całość pracy zmierza do połączenia wysokiej jakości modeli generatywnych z mechanizmami ochrony modeli i danych, aby budować bardziej godny zaufania ekosystem uczenia maszynowego.

Rozdział 2 wprowadza tło teoretyczne potrzebne do zrozumienia dalszych części pracy: opisuje główne typy współczesnych modeli generatywnych, kontekst szybkich symulacji w CERN oraz podstawowe rodzaje ataków i technik audytu modeli i danych. Najpierw autor omawia autoenkodery i wariacyjne autoenkodery jako metodę uczenia reprezentacji i generowania

danych. Następnie przedstawia ideę GAN, ich zalety i problemy z trenowaniem, w szczególności z mode collapse i brakiem enkodera. Kolejna część dotyczy modeli dyfuzyjnych: krokowego zaszumiania i odszumiania danych, uproszczonej funkcji kosztu opartej na przewidywaniu szumu oraz ich rozszerzenia do Latent Diffusion Models, gdzie proces odbywa się w przestrzeni zakodowanej przez autoenkoder, co przyspiesza uczenie i generowanie. Na końcu tej sekcji autor opisuje modele autoregresyjne obrazów oparte na tokenizacji (np. VQ-VAE) i transformerach, a także nowsze warianty (RAR, VAR, MAR), które zmieniają kolejność generowania lub rezygnują z kwantyzacji, co ma wpływ na własności modeli.

Druga główna część rozdziału dotyczy zastosowań modeli generatywnych do szybkich symulacji w CERN, ze szczególnym naciskiem na eksperyment ALICE i detektor Zero Degree Calorimeter. Autor wyjaśnia, że klasyczne symulacje Monte Carlo są bardzo kosztowne obliczeniowo, podczas gdy modele uczenia maszynowego mogą nauczyć się mapowania z parametrów zderzenia na odpowiedź detektora i generować próbki znacznie szybciej. Opisuje rolę ZDC w wyznaczaniu centralności zderzeń, jego budowę jako siatki włókien światłowodowych dających „obrazy" depozycji energii i podkreśla, że rozkłady odpowiedzi są silnie heterogeniczne i wielomodalne, co stawia wysokie wymagania modelom generatywnym. Ten fragment stanowi bezpośrednią motywację dla późniejszych propozycji SDI-GAN i ExpertSim.

W trzeciej części autor wprowadza pojęcie „ochrony wartości" zawartej w modelach i danych i syntetycznie omawia trzy klasy technik: model stealing, membership inference oraz dataset inference. Model stealing opisuje jako proces trenowania modelu-zastępnika na odpowiedziach modelu ofiary, często z użyciem adaptacyjnych zapytań, co pozwala odtworzyć funkcjonalność bez dostępu do wag czy danych. Membership inference definiuje jako rozstrzyganie, czy pojedynczy przykład był w zbiorze treningowym na podstawie zachowania modelu (np. wartości straty), a dataset inference jako rozszerzenie tej idei na całe zbiory; zamiast jednego punktu testuje się, czy cała kolekcja była użyta do treningu, agregując słabe sygnały w mocniejszy wniosek statystyczny. Autor podkreśla, że te techniki mają charakter dwojaki: są jednocześnie formą ataku i narzędziem audytu, które później wykorzystuje w części poświęconej ochronie danych i praw autorskich.

W rozdziale 3 autor omawia literaturę w trzech obszarach, które odpowiadają trzem częściom pracy. Najpierw omawia dotychczasowe zastosowania modeli generatywnych (głównie GAN) do szybkich symulacji w fizyce wysokich energii, pokazując, że istnieją prace nad zastępowaniem klasycznych symulacji, ale wciąż są problemy z wiernym odwzorowaniem złożonych, wielomodalnych rozkładów i stabilnością trenowania. Następnie przedstawia istniejące metody obrony przed kradzieżą nauczonego modelu (*model stealing*), podkreślając, że skupiają się przeważnie na klasyfikatorach lub modelach generatywnych z inną strukturą i nie są projektowane specjalnie dla enkoderów, co uzasadnia potrzebę nowej aktywnej obrony. W trzeciej części omawiane się prace związane z identyfikacją danych treningowych w modelach generatywnych: klasyczne metody typu membership inference, ich adaptacje do dużych modeli dyfuzyjnych oraz pierwsze próby ataków i audytu na poziomie całych zbiorów, wskazując ograniczenia dotychczasowych podejść i miejsce, w które wpasowuje się zaproponowana później metoda CDI oraz analiza prywatności modeli autoregresyjnych obrazów.

Rozdział 4 przedstawia metodę SDI-GAN, która ma poprawić użyteczność modeli cGAN w symulacji ZDC, poprzez kontrolowane zwiększanie różnorodności próbek tylko tam, gdzie dane rzeczywiście są zróżnicowane. Autor wychodzi od obserwacji, że w klasycznych sieciach cGAN w zadaniach fizycznych często pojawia się zjawisko *mode collapse*: generator ignoruje szum i dla danego warunku zwraca w praktyce jedną „typową" odpowiedź, co jest nieakceptowalne przy symulacji zderzeń cząstek. Omówione są istniejące metody zwiększania

różnorodności (np. MS-GAN, DS-GAN, DivCo) i wskazuje ich słaby punkt: narzucają podobny poziom różnorodności dla wszystkich warunków, podczas gdy w rzeczywistych danych CERN rozrzut wyników bardzo zależy od parametrów zderzenia.

Proponowana metoda dodaje do funkcji kosztu prosty człon regularyzujący, który maksymalizuje stosunek odległości między reprezentacjami dwóch wygenerowanych obrazów do odległości między odpowiadającymi im wektorami ukrytymi, ale skalowany lokalną różnorodnością w danych treningowych dla danego warunku *c*. Odległość między obrazami liczona jest nie w przestrzeni pikselowej, lecz w przestrzeni cech na poziomie warstw dyskryminatora, co ma skoncentrować się na różnicach semantycznych, a nie czysto wizualnym szumie. Całkowita funkcja kosztu to klasyczna strata adversarialna plus ważony człon różnorodności z hiperparametrem λ_{div} wpływającym na siłę regularyzacji.

W części eksperymentalnej metoda jest oceniana najpierw na syntetycznym zbiorze 2D, w którym każda klasa ma dwa warianty („spread = False" z małą wariancją i „spread = True" z dużą), a następnie na danych z GEANT4 dla ZDC (ok. 296 tys. przykładów, obrazy 44×44). Na tym zbiorze, zaproponowany model SDI-GAN lepiej niż bazowe modele cGAN, MS-GAN i DivCo, odtwarza różne poziomy rozrzutu dla poszczególnych warunków, podczas gdy konkurencyjne podejścia mają tendencję do uśredniania zachowania. W zadaniu ZDC autor pokazuje, że SDI-GAN zwiększa różnorodność generowanych odpowiedzi tam, gdzie w danych występuje realna wielomodalność, przy zachowaniu zgodności rozkładów energii i innych metryk z symulacją Monte Carlo. Rozdział kończy się konkluzją, że selektywna regularyzacja różnorodności poprawia przydatność modeli cGAN w zastosowaniach naukowych.

Rozdział 5 wprowadza ExpertSim, model typu mieszanina generatywnych ekspertów dla symulacji odpowiedzi ZDC, który ma rozwiązać problem polegający na tym, że pojedynczy GAN (nawet zaproponowany wcześniej SDI-GAN) nie wystarcza do uchwycenia mocno wielomodalnych rozkładów w danych zderzeń. Autor argumentuje, że odpowiedź detektora zależy od wielu czynników (energia, typ wiązki, geometria zderzenia), a różne obszary przestrzeni wejściowej mają inne parametry fizyczne, więc naturalne jest, aby trenować kilka wyspecjalizowanych generatorów zamiast jednego uniwersalnego.

ExpertSim składa się z kilku ekspertów, modeli GAN (ExpertGAN), z których każdy jest trenowany do modelowania części przestrzeni wejściowej, oraz z sieci „routera", która na podstawie warunków zderzenia wybiera lub miesza ekspertów przy generowaniu próbki. Eksperci mają podobną architekturę do SDI-GAN, ale ich zadanie jest węższe; dzięki temu mogą lepiej odwzorować lokalne rozkłady, natomiast ruter uczy się, który ekspert jest odpowiedni dla danego regionu parametrów. Model jest trenowany tak, aby zarówno eksperci, jak i router minimalizowali łącznie stratę adversarialną i dodatkowe człony zachęcające do specjalizacji (np. ograniczające nakładanie się ekspertów).

W części eksperymentalnej autor pokazuje, że ExpertSim poprawia zgodność rozkładów energii względem cGAN i SDI-GAN, redukuje błędy w ogonach rozkładów i lepiej odtwarza strukturę „plam" energii w obrazach ZDC, przy jednocześnie korzystnym czasie inferencji (szczególnie ważnym z punktu widzenia produkcyjnego użycia w CERN). Analizuje też specjalizację ekspertów: każdy z nich pokrywa inny podzakres energii czy typów zdarzeń, oraz wykonuje badania ablacyjne (liczba ekspertów, różne strategie rutingu). Rozdział kończy się stwierdzeniem, że ExpertSim stanowi jakościowy krok naprzód względem pojedynczego modelu GAN i może zastąpić drogie symulacje Monte Carlo w wybranych częściach łańcucha rekonstrukcji.

Rozdział 6 proponuje aktywną metodę obrony przed kradzieżą enkoderów, nazwaną Bucks for Buckets (B4B), zaprojektowaną tak, aby utrudniać atakującym trenowanie kopii enkodera, przy

minimalnym wpływie na jakość reprezentacji zwracanych uczciwym użytkownikom. Autor przyjmuje scenariusz API (typu OpenAI/Cohere), które udostępnia SSL-owy enkoder wizji (np. SimSiam, DINO) i zwraca wektory zanurzeń; pokazuje, że reprezentacje dla typowych zadań downstream zajmują mały, relatywnie spójny podobszar przestrzeni, podczas gdy skuteczny atakujący musi „pokryć" dużą część przestrzeni zanurzeń, by odtworzyć ogólną funkcjonalność enkodera. To prowadzi do kluczowej intuicji: można odróżnić normalne użycie od ataku, monitorując, jak duży fragment przestrzeni reprezentacji zajmują zapytania danego użytkownika.

Metoda B4B składa się z trzech elementów: (1) modułu szacowania pokrycia przestrzeni embeddingów, (2) funkcji kosztu, która zamienia to pokrycie na „karę", oraz (3) transformacji per-użytkownik, które utrudniają ataki typu Sybil (wiele kont). W proponowanej instancji, pokrycie mierzone jest przez lokalne haszowanie wrażliwe na podobieństwo (LSH): każda reprezentacja wpada do zestawu „buckets", a system zlicza, ile bucketów wypełnił dany użytkownik. Funkcja kosztu jest oparta na użyteczności, tzn. dopóki liczba zajętych bucketów jest mała, szum dodawany do reprezentacji jest znikomy; gdy pokrycie rośnie, obrona zaczyna dodawać coraz silniejszy szum do zanurzeń, pogarszając jakość reprezentacji tylko dla użytkowników „eksplorujących" dużą część przestrzeni. Dodatkowo, na każdą tożsamość użytkownika nakładane są odwracalne transformacje, które zachowują użyteczność, ale powodują, że atakujący nie może łatwo scalać reprezentacji z wielu kont w jednym wspólnym układzie bez dodtkowego kosztu. W części eksperymentalnej autor testuje B4B przeciw skutecznym atakom na enkodery SimSiam i DINO, które wykorzystują kontrastowe uczenie lub MSE i augmentacje do zmniejszenia liczby zapytań.

Rozdział 7 bada, na ile realnie da się przeprowadzić membership inference attacks (MIA) na dużych modelach dyfuzyjnych, ze szczególnym naciskiem na Stable Diffusion v1.4. Autor pokazuje, że część wcześniejszych prac raportuje zbyt optymistyczne wyniki, bo korzysta z nierealistycznych założeń: np. z mocnego fine-tuningu na małych zbiorach (co prowadzi do nadmiernego przeuczenia) albo z nienaturalnego doboru zbioru non-members, który różni się rozkładem od danych treningowych. Proponuje zmodyfikowany przebieg eksperymentu: nie modyfikuje oryginalnego SD-v1.4 i konstruuje nowy zbiór LAION-mi, w którym zbiory „members" i „non-members" są starannie dobrane, duplikowane i „zsynchronizowane" tak, by miały możliwie taki sam rozkład. Następnie testuje różne klasy ataków. Autor stwierdza, że pojedynczo-próbkowe MIA na dużych modelach dyfuzyjnych są w praktyce mało wiarygodne i sensowną drogą jest przejście do obecności całych zbiorów uczących, co proponowane jest w kolejnym rozdziale.

Rozdział 8 proponuje metodę Copyrighted Data Identification (CDI) pozwalającą na sprawdzenie, czy dany zbiór (np. kolekcja zdjęć stockowych, portfolio artysty) był użyty do trenowania dużego modelu dyfuzyjnego, takiego jak Stable Diffusion czy DiT. Autor wychodzi od wyniku z rozdz. 7: membership inference dla pojedynczych obrazów daje zbyt słaby sygnał, by wiarygodnie udowodnić naruszenie praw autorskich. CDI przyjmuje więc jako wejście dwie próbki: zbiór podejrzany P (publicznie dostępne obrazy, co do których właściciel chce sprawdzić użycie w treningu) oraz zbiór kontrolny U z tego samego rozkładu, ale z obrazami niepublikowanymi/nieużytymi, a następnie agreguje wielowymiarowe cechy przynależności dla obu zbiorów i porównuje rozkłady. Architektura składa się z trzech etapów: (1) ekstrakcja wielu sygnałów na poziomie próbek; zarówno z istniejących MIAs (denoising loss, SecMI, PIA/PIAN), jak i nowych ręcznie projektowanych cech; (2) nauka oceny modelu, który łączy te cechy w pojedynczy wynik dla każdej próbki; (3) zastosowanie testu t Studenta (lub pokrewnego testu) między rozkładami ocen P i U, co pozwala z wysoką ufnością orzec, czy P ma „podpis" zbioru treningowego danego DM. W eksperymentach na szeregu architektur (LDM, DiT, U-ViT), różnych rozdzielczościach i trybach warunkowania (bezwarunkowe,

klasowe, tekstowe) autor pokazuje, że CDI potrafi przy próbkach rzędu 70 obrazów osiągnąć ponad 99% ufności przy wykrywaniu wykorzystania zbioru w treningu, pozostając odpornym na fałszywe alarmy i częściowe pokrycie (tylko część P faktycznie w treningu).

Rozdział 9 analizuje prywatność wizyjnych modeli autoregresyjnych: VAR, RAR, MAR i pokrewnych w porównaniu z modelami dyfuzyjnymi (DMs), pokazując, że przy podobnej jakości generacji modele IAR ujawniają znacznie więcej informacji o danych treningowych. Autor proponuje nową metodę membership inference attack dla modeli IAR, łącząc idee z ataków na DMs i LLM-y: wykorzystuje tokenową naturę przewidywań (jak w LLM), a także różnicę między inferencją warunkową i bezwarunkową (jak w DMs). Następnie wykorzystuje MIA jako rdzeń dataset inference (DI) zoptymalizowanego pod IAR: dzięki temu, że wszystkie rozważane MIA dobrze działają, nie jest potrzebny etap wyboru najlepszego MIA dla danego zbioru, a wiarygodne DI można przeprowadzić już na około 6 próbkach – dużo mniej niż ~200 próbek wymagane wcześniej dla DI w DMs. Trzecim filarem jest analiza zapamiętywania (*memorization*).

Rozdział 10 podsumowuje cały wkład rozprawy, podkreślając, że jej głównym celem było połączenie wiarygodnych i bezpiecznych modeli generatywnych: najpierw zaproponowano metody, które umożliwiają wierne i szybkie symulacje w HEP (SDI-GAN i ExpertSim), a następnie zestaw technik chroniących zarówno modele (B4B), jak i dane treningowe (realistyczne MIA dla DMs, CDI, ataki na IAR-y). Autor zbiera to w listę szczegółowych udziałów autora i publikacji w doktoracie oraz pokazuje, że praca przesuwa stan wiedzy od generatywnych modeli, po prostu generujących dane, do bardziej dojrzałego ekosystemu, w którym jakość, wydajność i własność intelektualna mają dużą istotność.

W dalszej części szkicuje przyszły rozwój generatywnej SI: przewiduje dalsze zacieranie granicy między modelami dla zastosowań naukowych i komercyjnych, wzrost skali modeli oraz rosnącą wagę pytań o audyt, odpowiedzialność i regulacje, zwłaszcza w kontekście ochrony danych i praw autorskich.

Dalej następuje spis publikacji oraz załączniki.

3. Ocena rozprawy

W ramach rozprawy doktorskiej Pan mgr inż. Jan Dubiński zaproponował zestaw metod, które w sposób oryginalny łączą projektowanie wydajnych i wiarygodnych modeli generatywnych dla fizyki wysokich energii z nowatorskimi metodami ochrony modeli i danych treningowych, tworząc spójny warsztat dla bezpiecznego wykorzystania dużych modeli generatywnych. Tematyka pracy jest bardzo aktualna i potrzebna, oryginalny dorobek doktoranta polega na:

- stworzeniu nowych metody generatywnych dla fizyki wysokich energii: propozycja SDI-GAN, modyfikacji cGAN selektywnie zwiększającej różnorodność próbek, tak by lepiej odwzorować rozkłady danych z detektorów przy zachowaniu jakości oraz opracowanie ExpertSim, mieszanki modeli generatywnych sterowanej ruterem, pozwalającej na szybkie i wierne symulacje w eksperymencie ALICE.
- opracowaniu metod aktywnej obrony przed kradzieżą enkoderów: zaproponowanie Bucks for Buckets, pierwszej aktywnej metody obrony enkoderów, opartej na monitorowaniu pokrycia przestrzeni zanurzeń, adaptacyjnej funkcji kosztu oraz transformacjach przeciw atakom Sybil.
- Realistyczna ocena membership inference na dużych modelach dyfuzyjnych: krytyczna analiza istniejących MIAs dla Stable Diffusion, pokazująca, że wcześniejsze wyniki są często przeszacowane z powodu nierealistycznych założeń oraz zbudowanie zbioru

LAION-mi i ramy eksperymentalnej, które pokazują, że single-sample MIA na dużych DMs są w praktyce słabe i nie wystarczą jako samodzielne narzędzie audytu.

- Zaprojektowanie metody Copyrighted Data Identification (CDI), która łączy wiele sygnałów przynależności na poziomie próbek ze statystycznym testowaniem na poziomie zbioru oraz empiryczne wykazanie, że metoda ta pozwala z wysoką ufnością (>99%) stwierdzić wykorzystanie danego zbioru w treningu dużego modelu dyfuzyjnego już przy stosunkowo niewielkiej liczbie próbek.
- Analiza prywatności wizyjnych modeli autoregresyjnych (IARs) Opracowanie silnych ataków membership i dataset inference dla IAR, pokazujących, że przy podobnej jakości generacji przeciek prywatności jest u nich znacznie większy niż w dyfuzyjnych oraz empiryczne wykazanie szerokiego i niemal dosłownego zapamiętywania oraz omówienie konsekwencji tego faktu dla projektowania i regulacji przyszłych generatywnych modeli wizyjnych.

Rozprawa doktorska uwidacznia wysoką ogólną wiedzę teoretyczną i praktyczną oraz umiejętność prowadzenia pracy naukowej mgr inż. Jana Dubińskiego. Opracował wprowadzenie do tematyki i przegląd literatury związanej z tematyką pracy. Rozprawa doktorska stanowi oryginalne rozwiązanie problemu naukowego. Zaproponowane metody mają duże znaczenie dla nauk technicznych oraz przemysłu, zarówno teoretyczne, jak i aplikacyjne.

Niezależnie od mojej bardzo dobrej oceny pracy, nasunęły mi się następujące pytania i uwagi:

- Metody SDI-GAN i ExpertSim są pokazane głównie na jednym detektorze (ZDC w ALICE) i ograniczonym zestawie warunków eksperymentalnych. Na ile wyniki przenoszą się na inne detektory/geometrie czy energie zderzeń.
- B4B wymaga doboru kilku progów i parametrów (liczba bucketów, kształt funkcji kosztu, parametry transformacji per-user), co może być trudne w rzeczywistym systemie API. Na ile wyniki są stabilne przy innym doborze hiperparametrów.
- Analiza zakłada konkretny styl ataku (globalne pokrycie przestrzeni embeddingów). Można zapytać, jak B4B zachowa się wobec atakującego, który zna mechanizm obrony i próbuje np. kraść model „po kawałku” (lokalny zakres embeddingów) albo tak projektuje zapytania, by długo nie przekraczać progu kosztu.
- Czy może być sytuacja, że w praktyce właściciel danych może mieć trudność z wygenerowaniem naprawdę reprezentatywnego zbioru U w metodzie CDI.

4. Wnioski końcowe recenzji

Podsumowując recenzję stwierdzam, że rozprawa doktorska Jana Dubińskiego, pt.: Reliable and Safe Generative Models” prezentuje oryginalne rezultaty stanowiące rozwiązanie problemu naukowego oraz wkład w rozwój dyscypliny informatyka techniczna i telekomunikacja. Pan Jan Dubiński wykazał się umiejętnością samodzielnej pracy badawczej, znajomością literatury światowej i wiedzą w zakresie uczenia maszynowego, modeli generatywnych i wizji komputerowej. Rozprawa doktorska prezentuje ogólną wiedzę teoretyczną kandydata. Recenzowana praca spełnia wymagania Ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (Dz.U. z 2022 r. poz. 574 z późn. zm.). w dyscyplinie naukowej informatyka techniczna i telekomunikacja. Wnoszę o jej przyjęcie i dopuszczenie do dalszych etapów postępowania doktorskiego. Ponadto ze względu na ponadprzeciętny poziom rozprawy oraz fakt opublikowania prac związanych bezpośrednio z tematyką rozprawy w materiałach najlepszych konferencji klasy A oraz A* w tym obszarze, wnioskuję o wyróżnienie pracy.

Robert Schaefer