RADA NAUKOWA DYSCYPLINY

INFORMATYKA TECHNICZNA I TELEKOMUNIKACJA POLITECHNIKI WARSZAWSKIEJ

zaprasza na

OBRONĘ ROZPRAWY DOKTORSKIEJ

mgr inż. Moniki Kai Wysoczańskiej

która odbędzie się w dniu **19 września 2025 roku,** o godzinie 12:00 w trybie hybrydowym

Temat rozprawy:

„Task Adaptation Strategies for Vision-Language Models"

Promotor: prof. dr hab. inż. Tomasz Trzciński – Politechnika Warszawska

Promotor pomocniczy: dr inż. Jacek Komorowski – Politechnika Warszawska

Recenzenci: Piotr Bojanowski, PhD – Meta Facebook AI Research, Francja

prof. Hilde Kuehne, PhD - Uniwersytet w Tubingen, Niemcy

prof. dr hab. inż. Jarosław Wąs - Akademia Górniczo-Hutnicza

Obrona odbędzie się w Audytorium Centralnym Wydziału Elektroniki i Technik Informacyjnych Politechniki Warszawskiej oraz jednocześnie zdalnie na platformie MS Teams (tryb hybrydowy). Osoby zainteresowane uczestnictwem w obronie proszone są o zgłoszenie chęci uczestnictwa w formie elektronicznej na adres sekretarza komisji: dr hab. inż. Tomasza Gambin, prof. uczelni email: tomasz.gambin@pw.edu.pl, do dnia 18 września 2025r. godz. 17:00.

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.-Monika-Kaja-Wysoczanska

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

Strategie Adaptacji Modeli Wizualno-Językowych do Zadań Docelowych

Niniejsza praca doktorska bada adaptacyjność modeli fundamentalnych do reprezentacji obrazu, do złożonych zadań docelowych, ze szczególnym naciskiem na podejścia wykorzystujące wewnętrzne możliwości modeli fundamentalnych przy minimalizacji kosztów ich modyfikacji, w tym dodatkowego treningu.

Nasze badania koncentrują się głównie na modelach wizualno-językowych (VLM), analizując, jak ich multimodalne reprezentacje obrazu i tekstu mogą być rozszerzone do zadań wymagające szczegółowej reprezentacji obrazu. Poprzez pięć powiązanych ze sobą prac, odpowiadamy na centralne pytanie: W jakim stopniu reprezentacje wizualne z modeli fundamentalnych mogą być wykorzystane do różnych zadań docelowych przy minimalnym nadzorze?

W pierwszej części doktoratu omawiamy metody adaptacji modelu CLIP do zadania semantycznej segmentacji z nieograniczonym słownikiem klas. Nasza pierwsza metoda, CLIP-DIY, demonstruje, możliwości adaptacji CLIP poprzez modyfikacje sposobu inferencji tym samym bez konieczności trenowania modelu. Następnie prezentujemy CLIP-DINOiser, metodę, która wzbogaca reprezentacje CLIP na poziomie pikseli poprzez wykorzystanie komplemetarnych umiejętności lokalizacji obiektów z reprezentacji uczonych z samonadzorem, takich jak DINO, za pomocą prostego modułu adaptacyjnego. W tej części pracy pokazujemy również jak poprawić skuteczność segmentacji poprzez starannie dobrane prompty tekstowe, które pozyskujemy poprzez analizę dużego korpusu danych wykorzystanych do trenowania VLM.

Wykraczając poza adaptację modelu, kolejna część doktoratu prezentuje metodę do ewaluacji reprezentacji wizualnych w złożonym zadaniu Visual Question Answering (VQA). Tak skonstruowany sposób ewaluacji pozwala dogłębnie zrozumieć skuteczność poszczególnych reprezenetacji wizualnych do zadań rozumowania na podstawie obrazu. W ostatniej części pracy pokazujemy praktyczne zastosowanie adaptacji VLM do zadania personalizacji wizualnych podsumowań na przykładzie problemu na platformie Booking.com, łącząc analizę reprezentacji wizualnej z analizową tekstowych recenzji użytkowników platformy.

Podsumowując, niniejsza praca pokazuje, że modele fundamentalne do reprezentacji obrazu mogą być efektywnie adaptowane do złożonych zadań wymagających szczegółowego rozumienia obrazu przy minimalnych wymaganiach obliczeniowych i anotacyjnych. Nasze obserwacje pogłębiają wiedzą dotyczącą rożnych sposobów reprezentacji obrazu jednocześnie dostarczając praktycznych rozwiązań do adaptacji modeli fundamentalnych w różnorodnych zastosowaniach.

Słowa kluczowe: metody adaptacji do zadań złożonych, reprezentacje wizualno-tekstowe,

segmentacja semantyczna z nieograniczonym słownikiem, uczenie nienadzorowane/ samonadzorowane

Meta

June 9, 2025

Re: Dissertation review for Monika Wysoczańska's PhD thesis "Task Adaptation Strategies for Vision-Language Models"

This thesis is coping with the problem of vision foundation model adaptation. The manuscript explores the different strategies that can be applied to solve new problems or improve current systems by leveraging these large models trained on gigantic data corpora. The most significant part of the manuscript copes with the problem of open vocabulary semantic segmentation. In open vocabulary segmentation the set of classes that are used to densely label the image is not known a priori, and the same model should therefore work across many datasets with largely different class ontologies. This is a very hard problem, because the problem setup does not allow much model tuning, and the set of categories on which the model is applied can be arbitrarily finegrained. The use of off-the-shelf foundation models is hard, as the nature of the inference is very different. This thesis presents several key state-of-the-art contributions in that space.
An additional part of the manuscript focuses on the adaptation of vision foundation models in two other setups: visual question answering and automatic curation of photo collections.

Summary

The thesis is composed of an introduction and related work chapters (chapters 1 and 2), followed by five technical chapters that correspond to previously published papers (chapters 3 through 7), and a conclusion (chapter 8).

Chapter 1 serves as an introduction. It sets the context, describing the conceptual difference between siloed custom-tailored solutions, and foundation models. The candidate then describes how foundation models could be used in specialized scenarios and what adaptation they need to undergo. The introduction finishes with stating five research questions, grouped into three buckets, that cover the content of the five technical chapters.

- The introduction states that dense annotations are nearly impossible to obtain, especially in expert domains such as medical imaging. Can foundation models like CLIP reliably address this problem? What about the scale of weakly aligned text-image data that is needed to train CLIP (or DFN, SigLIP, PE)?
- Regarding the research question 4: have the findings from VQA evals really changed anything in the way foundation models are designed? What could be easily actionable and scalable improvements?
- In the introduction, the contribution from Chapter 7 is presented as a "real world" case, seemingly in opposition to the other chapters. What makes it more "real"? The

fact that the data and problem was specific to a company?

Chapter 2 describes previous work and focuses on two research themes: foundation models for vision, and model adaptation.

- The categorization of foundation models, and structure of Sec. 2.1 seems right to me. However, it feels like the coverage could be larger, as many influential works in that space have been omitted. For supervised learning alone, the list of NN architectures is a bit expeditive (VGG), not to mention the early hassle of ViT training (DeiT). For CLIP-like models, one could dive deeper in older work, and most importantly be more exhaustive about cutting edge models (EVA, AIMv2).
- The same applies to adaptation strategies in Sec. 2.2. While the broad categories are right, the amount of details per category could be improved. The “dataset adaptation” represents by itself an immense field of research around transfer learning. On the other hand, the subsequent chapters all have their own related work, and the bibliography contains 259 entries, so I am confident that the candidate properly acknowledged all relevant papers. This does not invalidate the scientific soundness of the presented work, only aims at improving the added value of this chapter.

Chapter 3 describes a training-free method for obtaining segmentation masks for any textual prompt coined CLIP-DIY. The method works by computing text-patch similarities at several scales using an off-the-shelf CLIP model (Eq. (3.1)), and further gating this score using an unsupervised object discovery model like FOUND or CutLER (Eq. (3.4)). The proposed algorithm outperforms (or nearly matches) the state-of-the-art on Pascal VOC and COCO. This is particularly strong, given that the method did not have any trainable parameters!

- From Table 3.4.2 it seems that the method simply does not work without the gating using “objectness”. However, without many scales the numbers are already quite strong. Have you considered this problem the other way around? Could you formulate it as a zero-shot classification of segments given by FOUND or CutLER? In that case, could you leverage the CLS token of CLIP?

Chapter 4 describes an improvement upon CLIP-DIY. While the aforementioned model leveraged two off-the-shelf foundation models (CLIP and FOUND), in this contribution the different models are fused together. A lightweight model is trained atop CLIP to mimic the self-similarities of DINO. Such cleaned-up self-similarities are used to pool clip features from larger regions, effectively denoising even further the CLIP feature map. Finally, a simple foreground background model is trained atop CLIP features by distilling the scores of FOUND. The proposed method gives strong performance across many datasets, setting a new standard in that space.

- The work on registers by Timothee Darcet (ICLR 2024) suggests that CLIP models could be trained at scale and obtain much better attention maps with some care. One of the two advantages of using DINO here is to clean up noisy local features from CLIP. Do you think that most modern CLIP-derivatives would still benefit from a DINO model?
- A lot of energy seems spent on dealing with the “background”. This is a quite ill-defined concept. Could the evaluation protocol be changed, in order to circumvent

this problem?

Chapter 5 describes a method for improving the way that open-world, open-vocabulary semantic segmentation is performed. Because open-vocabulary semantic segmentation models leverage contrastively-trained foundation models like CLIP, for a given class q, in order to classify each “patch” as positive or negative, the practitioner needs to properly define the “background”. This chapter describes a clean evaluation protocol which modifies multi-class semantic segmentation into a series of binary segmentation problems, and evaluates two strategies for defining background classes for a given class query: either by mining statistics from large image-caption datasets, or prompting an LLM for relevant negative prompts.

- The proposed strategies allow defining a set of “negatives” for a given query q, staying with the multinomial logistic model of CLIP. If we transform this problem into a sequence of binary problems, what would be the performance of a binary classifier with a threshold optimized on the training set of the given dataset? How many annotated images do you need to make that work? This baseline is mentioned at the beginning of Sec. 5.3.1 but only reported in the appendix (Fig. 5.6.6).
- With a slight abuse of notation, the proposed procedure feels like populating the training set of a non-parametric classifier like kNN. The training set is composed of one positive (the query), and we are trying to generate negative examples, in order to define the best decision boundary for that class (a convex polytope). Could that be generalized to potentially leverage the non-linear nature of non-parametric models?

Chapter 6 describes a benchmark for evaluating different image representations through the lens of visual question answering (VQA). The proposed evaluation is simple and clear, but limited to toyish data (CLEVR). It was somewhat pioneering, appearing before the consensus on architectures and datasets to evaluate foundation models with VQA (efforts like Cambrian).

- The best performing methods according to the experiments in this chapter are either Slot-Attention or DTI-Sprites. This is probably an artifact of the type of dataset that this evaluation was run on. My main question around this work is: how can we translate the learnings from this work into recommendation for future foundation model development? How would DINOSAUR perform in this benchmark (it appeared after the paper was presented)?

Chapter 7 describes a simple method for summarizing a photo collection in a personalized way, with application to Booking.com data. The model takes as input some user context (reviews), and a pool of images from a property, and proposes the most relevant pictures for that user. This is a creative adaptation of a visual foundation model like CLIP to the problem of personalization. In user studies, the proposed multimodal algorithm performs much better than the unimodal baseline.

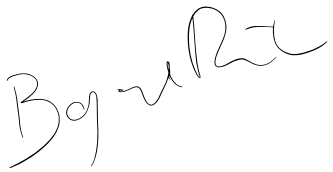
Chapter 8 provides some final remarks and discusses open problems and future work.

Appraisal

This thesis presents a very large body of work, spanning several themes. The writing is clear and the document is properly structured. The end of the introductions provides a clear overview of the following chapters, and connects them to publications. The very coherent set of contributions around open-vocabulary semantic segmentation is of prime quality and has already had a lot of impact on the research community. The manuscript illustrates both the technical mastery and deep knowledge of previous literature of the candidate.

I am confident that there is sufficient material and novelty in this manuscript for an oral defense before a jury.

Piotr Bojanowski
Research Director
bojanowski@meta.com



Tübingen AI Center
Universität Tübingen

Universität Tübingen · Tübingen AI Center
Maria-von-Linden-Straße 6 · 72076 Tübingen · Germany

Tuebingen, 14.07.2025

Gutachten zur Dissertation "Task Adaptation Strategies for Vision-Language Models" eingereicht von Ms. Monika Wysoczańska.

Prof. Dr.-Ing. Hilde Kühne
Professor for Multimodal Learning
Paul-Ehrlich-Str. 5
phone +49 7071 29-70867
h.kuehne@uni-tuebingen.de

https://tuebingen.ai

The thesis submitted by Ms. Wysoczańska targets the problem of adapting pretrained vision-language foundation models to various downstream tasks.

The thesis is, besides introducation and realted work broken down into five different aspects task-adaptation: First, the adaptaton of visual foundation models for fine-grained localization tasks, realized by a framework called CLIP-DIY (published at WACV 2024); second the problem of combining different backbones, namely CLIP and DINO for better performance (published at ECCV 2024); this idea is then further extended towards test-time adadptaon (publication accpeted to TMLR on July 9th[1]); next, the idea is proposed that the process of task adaptation can serve as an evaluation for visual representations (published at NeurIPS Workshop 2022 and IEEE Access 2024); finally, those findings are used to improve the adaptation of VLMs for personalizations such as summarizing large image collections (published at AAAI 2024).

The first part, "Task Adaptation through Modified Inference" introduces the problem of adapting visual foundation models open-vocabulary semantic segmentation (OVSS) together with an approach called CLIP-DIY. This method extends the capabilities of CLIP to OVSS through modified inference strategies. employs a multi-scale architecture that leverages CLIP's classification abilities at various spatial resolutions and enhances segmentation accuracy using foreground/background separation scores derived from unsupervised object localization techniques. Notably, CLIP-DIY achieves competitive results on standard semantic segmentation benchmarks without additional training or manual annotations. The approach combines CLIP with an unsupervised saliency detection method based on DINO, a self-supervised learning model known for its strong object localization capabilities. This integration of vision-language modeling and self-supervised learning paradigms shows promising potential for further exploration. The work was

[1] https://openreview.net/forum?id=wyOv4kGkbU

presented at the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) 2024.

The second part, “Leveraging Complementary Visual Foundation Model,” discusses the adaptation of CLIP through complementary visual foundation models such as DINO. The proposed method ‚CLIP-DINOiser, combines the complementary representations of image-text-aligned and self-supervised visual representations to enhance performance in open-vocabulary semantic segmentation. To this end, DINO’s good localization priors are integrated into CLIP’s representation space via a lightweight adaptation module trained with DINO’s supervision while preserving CLIP’s original representations. CLIP-DINOiser has shown to reach state-of-the-art performance at that time while only requiring a single CLIP forward pass and two lightweight modules during inference. The work was presented at ECCV 2024.

The thrid part, “Leveraging Statistics from Pre-training Dataset” deals with the role of prompts for OVSS proposing, first, automated approaches for generating contrasting concepts (CC) at inference time based on an LLM as well as on a statistical analyis of the VLM’s pre-training dataset, and, second, an evaluation framework to cature real-world challenges in OVSS. It shows that integrating the pretraining distribution can positively influnce the perfromance of CLIP-based semantic segmentation methods. The work is accepted for TMRL.

The fourth part, “Task Adaptation for Foundation Model Selection and Analysis” proposes an evaluation protocol for visual representations in the context of VQA through downstream task adaptation. To enable comparison of differrent visual representations, a model is designed that accommodates varying input representation dimensions to allow for direct comparative studies of different visual representations and their task-specific suitability. In this context, VQA is leveraged as an exemplary task. The framework allows to derive implications for the development of more robust visual representations. This work was presented at NeurIPS Workshop 2022, and published by IEEE Access in 2024.

Finally, the work discusses the “Task Adaptation in Real-world Scenarios”. To this end, findings with respect to the adaptability of visual foundation models across different downstream tasks, are considered with respect to their practical implications in real-world contexts, namely adaptation of a Vision-Language for personalized image collection summarization. To this end, rating information for a travel protal website are leveraged to personalize summaries for future users of the platform. The approach employs a proprietary VLM to establish connections between visual content and textual user feedback, resulting in a more relevant and personalized visual summaries without requiring additional manual annotations. This work was published at AAAI 2024.

Overall, Ms. Wysoczańskas work features a high-quality contribution to the field of vision-langauge learning. The thesis is highly streamlined, ranging from basic representation learning problems to questions of how to best evaluate such methods and well as how to make use of their capabilities in real-life. It provides a cohesive structure, with each part of the work building up on the previous ones. This structural approach allows Ms. Wysoczańska to create high-performing frameworks and models, which she shows to be useful in their specific task. At the same time, all topics targeted in the thesis are innovative and well-structured and executed.

With five first-author publications at top venues and good journals (ECCV, AAAI, WACV, TMLR, IEEE access) and several successful collaborations, Ms. Wysoczańska has demonstrated her ability to develop, implement, and evaluate new ideas and concepts in the field of vision-langauge understanding specifically and computer vision in general. With this work, Ms. Wysoczańska made a significant contribution to current research in the field of computer vision and multimodal learning.

I, therefore, I consider the thesis as pass and assess the outcome of the submitted PhD Thesis with:

Summa cum laude

H. Kuehne

Prof. Dr. Hildegard Kuehne

AGH UNIVERSITY OF KRAKOW
FACULTY OF ELECTRICAL ENGINEERING, AUTOMATICS, COMPUTER SCIENCE AND BIOMEDICAL ENGINEERING
INSTITUTE OF APPLIED COMPUTER SCIENCE

Kraków 12 czerwca 2025

Prof. dr. hab. inż. Jarosław Wąs,
Wydział Elektrotechniki, Automatyki, Informatyki i Inżynierii Biomedycznej
Akademia Górniczo Hutnicza im. Stanisław Staszica w Krakowie

Recenzja rozprawy doktorskiej

Recenzję pracy doktorskiej mgr inż. Moniki Wysoczańskiej zatytułowanej *„Task Adaptation Strategies for Vision-Language Models"*, opracowano na prośbę prof. Jarosława Arabasa Przewodniczącego Rady Dyscypliny Informatyka Techniczna i Telekomunikacja (na podstawie pisma z dnia 9 kwietnia 2025).

Tematyka rozprawy i tytuł pracy

Tematyka rozprawy doktorskiej mgr inż. Moniki pt. ,, *Task Adaptation Strategies for Vision-Language Models*'', dotyczy problematyki wykorzystania modeli wizualno-językowych (ang. VLM) do zadań wymagających szczegółowej reprezentacji obrazu.

Główny problem badawczy jest związany z odpowiedzią na pytanie dotyczące możliwości wykorzystania reprezentacji wizualnych z modeli fundamentalnych do rozwiązywania konkretnych zadań przy założeniu minimalnego kosztu. W mojej ocenie temat pracy poprawnie koresponduje z zawartymi treściami.

Charakterystyka rozprawy i ocena układu rozprawy

Praca doktorska mgr inż. Moniki Wysoczańskiej składa się łącznie z ośmiu rozdziałów, obejmujących: wstęp (rozdział 1), zasadniczą część pracy obejmującą rozdziały 2-7, podsumowanie (rozdział 8) oraz bibliografię. Tekst pracy zajmuje 171 stron wraz z bibliografią.

1

Pierwszy rozdział pracy zatytułowany „Introduction" poświęcony jest określeniu celów ogólnych i szczegółowych pracy, wskazaniu problemu naukowego oraz określeniu metodyki pracy z uwzględnieniem: badań literaturowych, wypracowaniu generycznego podejścia i opisem szeregu szczegółowych rozwiązań w zakresie adaptacji zadań związanej z minimalizacją kosztu (sekcja 1.2.1.), wyborem modelu (sekcja 1.2.2) oraz zastosowań do realnych problemów (sekcja 1.2.3).

Rozdział 2 zatytułowany *„Background"* przedstawia tło prowadzonych badań z podziałem na Visual Foundation Models (sekcja 2.1.) obejmujących klasyfikację tychże modeli oraz Downstream task adaptation strategies (sekcja 2.2.) obejmująca dyskusję poszczególnych stategii.

Z kolei w rozdziale 3 opisano pierwszą z metod zaproponowanych przez Autorkę, służącą do semantycznej segmentacji z nieograniczonym słownikiem klas nazwaną CLIP-DIY Rozdział ten napisano na bazie artykułu Autorki opublikowanego w ramach publikacji konferencyjnej IEEE/CVF WACV 2024.

Rozdział 4 *Lavereging Complementery Visual Foundation Model* wprowadza metodę nazwaną CLIP-DINO, która dodaje do reprezentacji CLIP umiejętności lokalizacji obiektów z reprezentacji uczonych z samonadzorem, np. DINO, z wykorzystaniem prostego modułu adaptacyjnego. Rozdział ten oaprto na publikacji Autorki z konferencji ECCV 2024.

Rozdział 5 *Leveraging Statistics from Pre-training Dataset* przedstawia analizę częstości dystrybucji w docelowym zbiorze danych z użyciem pre-trenowanych danych użytych w CLIP.

Rozdział 6 *Task adaptation for Foundation Model Selection and Analysis* koncentruje się na ewaluacji reprezentacji wizualnych w zadaniu Visual Question Answering (VQA).

Rozdział 7 *Task Adaptation in real-world Scenarios* ukazuje praktyczne zastosowanie adaptacji VLM do połączenia reprezentacji wizualnej oraz analizy tekstowej recenzji umieszczanych w serwisie booking.com.

Ostatni z rozdziałów 8 *Final Remarks* stanowi podsumowanie odnoszące się do uzyskanych wyników. Odniesiono się do realizacji celów ogólnych i szczegółowych pracy, które zostały osiągnięte.

W mojej ocenie układ redakcyjny pracy jest poprawny. Autorka posługuje się poprawną terminologią w zakresie zagadnienia modeli wizualno-językowych i, patrząc szerzej, stosowanych metod sztucznej inteligencji.

Należy stwierdzić, że oceniana praca jest napisana poprawnym językiem. Bibliografia jest generalnie aktualna. Odwołania do pozycji literaturowych są prawidłowe.

Tytuły i kolejność rozdziałów jest również prawidłowa, treść rozdziałów odpowiada, w mojej ocenie, wymaganiom stawianym dla prac doktorskich.

Ocena zastosowanego piśmiennictwa

Bibliografia składa się z 259 pozycji związanych z tematyką pracy. Pozycje literaturowe zostały dobrany prawidłowo biorąc pod uwagę ich istotność i aktualność. Zostały one poprawnie przeanalizowane i wkomponowane w treść rozprawy.

Wskazanie oraz ocena celu pracy

Można wskazać dwa główne cele pracy: pierwszym celem była odpowiedź na pytanie jak reprezentacje modeli VLM mogą być rozszerzane do zadań wymagających szczegółowej reprezentacji obrazu.
Drugim celem była odpowiedź na pytanie w jakim stopniu reprezentacje wizualne z modeli fundamentalnych mogą być wykorzystywane do różnych zadań przy założeniu minimalnego kosztu.

W mojej ocenie obydwa te cele zostały zrealizowane w prawidłowy sposób.

Ocena czy rozprawa doktorska stanowi oryginalne rozwiązanie problemu naukowego

W pracy rozwiązano szereg zagadnień związanych z reprezentacjami modeli wizualno-językowych do zadań szczegółowej reprezentacji obrazu oraz analizę wykorzystania reprezentacji wizualnych z modeli fundamentalnych do konkretnych zdań przy założeniu minimalnego kosztu (w postaci np. minimalnego dodatkowego nadzoru).
Po pierwsze zaproponowano rozszerzenie metody CLIP do CLIP-DIY przez modyfikację sposobu inferencji, co zredukowało konieczność czasochłonnego trenowania modelu.
Po drugie skonstruowano metodę CLIP-DINOiser, której zadaniem jest wzbogacenie reprezentacji CLIP poprzez wykorzystanie umiejętności lokalizacji obiektów z relatywnie prostych reprezentacji typu DINO za pomocą modułu adaptacyjnego.

Po trzecie zaproponowano metodę poprawy skuteczności segmentacji poprzez analizę dużego korpusu danych wykorzystanych do trenowania VLM

Po czwarte zaprezentowano metodę do ewaluacji reprezentacji wizualnych w klasie zadań typu VQA -Visual Question Answering.
Po piąte zaprezentowano praktyczne zastosowanie metody adaptacji VLM do połączenia reprezentacji wizualnej z reprezentacją tekstową recenzji użytkowników booking.com.

Biorąc pod uwagę wszystkie składowe uważam, że oceniana rozprawa zawiera oryginalne rozwiązanie problemu naukowego.

W mojej ocenie praca bardzo dobrze wpisuje się w dyscyplinę informatyka techniczna i telekomunikacja i napisana jest poprawnie z punktu widzenia metodyki badawczej. Na uwagę zasługuje fakt, że Autorka wykonała szereg prac we współpracy z zespołami międzynarodowymi, co pozwoliło na szersze spojrzenie na omawiane problemy naukowe.

1

Uwagi:

Jak to zostało napisane wcześniej, praca została napisana poprawnym językiem i złożona starannie ze strony edytorskiej.

Poniżej przedstawiono uwagi dotyczące ocenianej pracy:

- W sekcji 1. Autorka przedstawia wprowadzenie do pracy. Jedną z głównych koncepcji są reprezentacje wizualne z modeli fundamentalnych przy minimalnym koszcie. Brakuje mi tu jednak precyzyjnego opisu jak, konkretnie, rozumiany jest minimalny koszt. Owszem w sekcji 1.2.1 i sekcjach sąsiednich znajdujemy pewne wyjaśnienia, ale przydałoby się je opisać w zwartej formie w jednym miejscu.
- Przydałoby się już we wstępie doprecyzować opis do pytania badawczego 4, co oznacza, że dany model fundamentalny jest najlepszy (w kontekście specyficznych zadań) - ze wskazaniem kryteriów optymalizacji.
- Autorka w kilku miejscach referuje do metod typu zero-shot. Warto byłoby zebrać argumenty za i przeciw stosowaniu takiego podejścia w jednym miejscu, w kontekście prac przeprowadzonych przez Autorkę.
- Ponieważ praca opiera się na kilku powiązanych tematyczne publikacjach, moja generalna uwaga dotyczy wprowadzenia wspólnego aparatu pojęciowego dla poszczególnych prac gdzieś na początku pracy. Owszem we wstępie można znaleźć pewne wyjaśnienia i wskazówki, ale bardziej systematyczny przegląd w pierwszej części ułatwiłby lekturę pracy.
- Rozdział 6 pracy obejmuje dosyć istotną tematykę związaną z nienadzorowanym wizyjnym rozumowaniem i pojęciem cech off-the - shelf. Został on opublikowany w ,,skromniejszym'' miejscu niż sąsiednie rozdziały – tzn. w IEEE Access. Przydałoby się to opisać raz jeszcze, opatrzeć tą ideę dobrymi przykładami i spróbować lepiej rozpropagować w środowisku naukowym.

Opisane powyżej uwagi nie wpływają jednak negatywnie na mój bardzo pozytywny odbiór ocenianej pracy doktorskiej.

Wniosek końcowy

Rozprawa stanowi oryginalne rozwiązanie problemu naukowego i wskazuje na wysoki poziom wiedzy teoretycznej i praktycznej Kandydatki w dyscyplinie naukowej Informatyka Techniczna i Telekomunikacja.
Przedstawiona praca doktorska spełnia w pełni warunki określone w art. 187 ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (j.t. Dz.U. z 2020 r. poz. 85 z późn. zm.). W szczególności zgodnie z wymaganiami ww. ustawy oceniana rozprawa pozytywnie prezentuje wiedzę teoretyczną Kandydatki oraz umiejętność samodzielnego prowadzenia pracy naukowej,

1

przedmiotem rozprawy jest oryginalne rozwiązanie problemu naukowego. W opinii recenzenta praca może więc być dopuszczona do publicznej obrony.
Po dokładnym zapoznaniu się z rozprawą i publikacjami Autorki wnioskuję o wyróżnienie rozprawy.

Prof. dr hab. inż. Jarosław Wąs
jaroslaw.was@agh.edu.pl