

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.



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 strategii.

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 *Leveraging Complementary Visual Foundation Model* wprowadza metodę nazwaną CLIP-DINO, która dodaje do reprezentacji CLIP umiejętność lokalizacji obiektów z reprezentacji uczonych z samonadzorem, np. DINO, z wykorzystaniem prostego modułu adaptacyjnego. Rozdział ten oparto 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 dobrane prawidłowo biorąc pod uwagę ich istotność i aktualność. Zostały one poprawnie przeanalizowane i wkomponowane w treść rozprawy.

7

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 obydwie 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 zadań 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.

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 tematycznie 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, opatrzyć 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

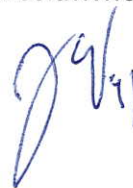
Rozprawa stanowi oryginalne rozwiązanie problemu naukowego i wskazuje na wysoki poziom wiedzy teoretycznej i praktycznej Kandydatki w dyscyplinie naukowej Informatyka Techniczna i Telekomunikacja.

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,



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

A handwritten signature in blue ink, consisting of stylized, cursive letters, likely representing the name Jarosław Wąs.