Zachodniopomorski Uniwersytet Technologiczny w Szczecinie

Wydział Informatyki

Szczecin, 20.02.2024

dr hab. inż. Paweł Forczmański, prof. ZUT

Zachodniopomorski Uniwersytet Technologiczny w Szczecinie

Wydział Informatyki

pforczmański@zut.edu.pl

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

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

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

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

1. Zakres, cel, teza i charakter rozprawy

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

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

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

2. Układ rozprawy i jej składowe

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

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

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

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

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

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

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

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

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

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

3. Analiza źródeł

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

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

4. Metodyka badań

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

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

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

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

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

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

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

5. Oryginalność rozwiązania postawionego problemu badawczego

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

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

Najważniejsze oryginalne osiągniecia Autora przedstawione w pracy to:

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

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

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

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

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

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

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

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

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

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

7. Znaczenie uzyskanych wyników i ich praktyczne wykorzystanie

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

8. Konkluzja

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

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

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

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

Paweł Forczmański