Katowice, 03.03.2024

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

Recenzja rozprawy doktorskiej

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

Autor rozprawy: mgr inż. Witold Oleszkiewicz

Promotor rozprawy: dr hab. inż. Robert Marek Nowak

Dziedzina: nauki inżynieryjno-techniczne

Dyscyplina: informatyka techniczna i telekomunikacja

1. Temat i cel rozprawy

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

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

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

2. Zawartość i charakter publikacji/rozprawy

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

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

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

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

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

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

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

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

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

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

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

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

7. Oryginalne wyniki i ich znaczenie

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

Za najbardziej wartościowe wyniki uzyskane przez Doktoranta uważam:

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

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

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

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

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

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

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

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

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

9. Słabe strony i uwagi krytyczne/dyskusyjne

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

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

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

Uwag i pytania szczegółowe:

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

10. Podsumowanie i wniosek końcowy

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

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

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