Prof. Michał Woźniak
Politechnika Wrocławska
michal.wozniak@pwr.edu.pl

Wrocław, 17 września 2023 r.

Recenzja rozprawy doktorskiej

mgra inż. Dawida Rymarczyka

zatytułowanej:

Interpretable Deep Learning with Prototypical Parts for Supervised and Weakly-Supervised Learning

1. Problem badawczy i jego znaczenie

W pracy rozważane są zagadnienia związane z wyjaśnialnymi metodami uczenia głębokiego. Praca ma charakter naukowy - doktorant proponuje w pracy autorskie metody wykorzystujące propotypy obiektów obrazu oraz poolingu atencyjnego, a także dokonuje ich ewaluacji na drodze eksperymentu komputerowego. Opracowane metody mogą znaleźć szerokie zastosowanie praktyczne, głównie w naukach przyrodniczych i medycznych.

2. Wkład autora

Celem pracy było zaprojektowanie modyfikacje dwóch wybranych podejść do interpretacji modeli głębokich. O ile zakres oraz zawartość rozprawy oceniam pozytywnie, to zwrócić należy uwagę, że teza rozprawy nie została sformułowana, a zamiast niej przedstawiono ogólny cel pracy, który określono nieprecyzyjnie. Autor formułuje go następująco (tłumaczenie własne) “Możliwe jest zaproponowanie bardziej przejrzystych i godnych zaufania systemów opartych na uczeniu maszynowym w oparciu o prototypy i mechanizm uwagi”. Tak sformułowany cel jest zawsze spełniony. Należałoby w nim zawrzeć jakimi cechami ilościowymi będą się odznaczały zaproponowane metody, np. w stosunku do metod znanych z literatury.
Do najważniejszych osiągnięć rozprawy zaliczam:

- Zaproponowanie trzech modyfikacji architektury ProtoPNet:
    - ProtoPShare współdzieląca część prototypową pomiędzy klasami poprzez łączenie w wytrenowanym modelu z wykorzystaniem podobieństw semantycznych pomiędzy prototypami.
    - ProtoPool wykorzystujący techniki regularyzacyjne oparte o Gumbel-Softmax oraz *focal similarity*, do wykrycia charakterystycznych części prototypowych.

    - ProGReST, która uogólnia powyższe metody do zadania regresji.
- Rozwinięcie techniki pooolingu atencyjnego w uczeniu wieloinstancyjnym:
    - Zaproponowanie SA-AbMILP, która korzysta z mechanizmu self-attention do nauki zależności pomiędzy instancjami oraz pozwalana ocenę wpływu poszczególnych cech na decyzję modelu.
    - Zaproponowanie połączenia części prototypowej oraz poolingu atencyjnego w formie architektury ProtoMIL, co pozwala na lokalną i globalną interpretowalność uczenia wieloinstancyjnego.
- Wnikliwą ocenę zaproponowanej metod na drodze eksperymentu komputerowego.
- Wskazanie zastosowanie praktycznego w obszarze nauk chemicznych.

Wyniki uzyskiwane w trakcie pracy nad rozprawą zostały zawarte w materiałach konferencyjnych bardzo dobrych konferencji naukowych, w tym KDD (CORE A*), ECCV (CORE A*), SIAM SDM (CORE A), ECML/PKDD (CORE A), IEEE WACV (CORE A), które odbyły się w latach 2021-2023.

3. Poprawność

Praca bazuje na wspomnianych artkułach autora, opublikowanych na renomowanych konferencjach i może wydawać się, że stanowi swojego rodzaju przewodnich po wspomnianych publikacjach. W mojej opinii, zdecydowanie wykracza ona poza tzw. przewodnik i stanowi pełnoprawną rozprawę ze wspomnianymi artykułami, jako miejscami, w których raczej można znaleźć bardziej szczegółowy opis zaprezentowanych rozwiązań oraz wyników. Rozprawa jest zredagowana bardzo starannie. Autor dużą wagę przykłada do dogłębnego wyjaśnienia zagadnień oraz przyjmuje jasny standard opisu własnych dokonać, tj. przedstawia jaki problem i dlaczego wymaga rozwiązania, jak go proponuje rozwiązać oraz omawia szczegółowe rozwiązanie. Takie podejście bardzo ułatwia lekturę pracy i jest poprawne, z punktu widzenia warsztatu badawczego, gdyż wszystkie sformułowane propozycje bazują na wnikliwej analizie literatury, dostrzegając możliwości poprawy zawartych tam metod. Jako narzędzie ewaluacji zaproponowanego algorytmu przyjęto głównie podejście oparte na eksperymencie komputerowym oraz często z wykorzystaniem oceny propozycji przez użytkowników (*user study*). Niestety, szczegóły user study nie zawsze były dokładnie opisane (np. w pracy *Interpretable Image Clasification with Differentiable Prototype Assigment* – autorzy odsyłają do tzw. *supplementary materials*, który nie znalazł się w rozprawie, choć udało mi się z nim zapoznać w wersji online https://arxiv.org/pdf/2112.02902.pdf).
Lektura rozprawy prowadzi do sformułowania poniższych uwag.

Uwagi ogólne:

- Ja wspomniano wcześniej, w pracy nie została sformułowana poprawnie teza pracy, a cel jest dość ogólnikowy.
- Wśród cytowanych prac brakuje mi dość ważnych prac z tego zakresu, szczególnie, że są one związane z wykorzystaniem prototypów do wyjaśniania, m.in. pracy Cynthia Rudin, *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*, a także prac z zespołów Klausa-Roberta Müllera oraz Wojciecha Samka.
- Jednym z problemów oceny metod eksplanacyjnych jest ich obiektywna ocena za pomocą wiarygodnych metryk. W pracy brak jest takiego wątku ani próby oceny wspomnianych metod za pomocą metryk zobiektywizowanych.
- W pracach z cyklu nie jest z reguły jasny protokół eksperymentalny, a także przesłanki jakim kierował się autor w przypadku wyboru metryk jakości klasyfikacji (różnią się między pracami). Dla części eksperymentów wskazano wartości odchylenia standardowego, stąd, dlaczego nie pokuszono się do wykorzystanie testów statystycznych do analizy wyników?

4. Wiedza kandydata

Na podstawie lektury uważam, że doktorant posiada ugruntowaną wiedzę z zakresu informatyki, w szczególności w zakresie wyjaśnianej sztucznej inteligenci, ze szczególnym uwzględnieniem metod bazujących na propotypach obiektów obrazu oraz poolingu atencyjnego. Doktorant posługuje się prawnie zaawansowanym aparatem matematycznym, a także potrafi zaplanować i przeprowadzić eksperyment komputerowy w celu oceny jakości zaproponowanych metod.
Przegląd literaturowy dotyczący zagadnień przedstawionych w rozprawie pozwala stwierdzić, że doktorant posiada aktualną wiedzę z zakresu tematyki rozprawy, a także potrafi dokonać krytycznego przeglądu źródeł w celu wskazania ciekawych kierunków badań. Zawarty w dysertacji spis źródeł literaturowych, zawierających 82 pozycje, jest aktualny i kompletny oraz uzupełniony o pozycje cytowane w dołączonych do rozprawy artykułach.

5. Podsumowanie

Doktorant wykazał się w recenzowanej rozprawie właściwie stosowanym podejściem analitycznym i eksperymentalnym oraz dobrą znajomością aktualnej problematyki związanej metodami wyjaśnialnej sztucznej inteligencji, głównie w zakresie wyjaśniani aprzez podobieństwo do prototypów obrazów, czy z wykorzystaneim metod poolingu atencyjnego w zadaniach klasyfikacji wieloinstancyjnej. Zostało to poparte bardzo dobrymi studiami

literaturowymi, obejmującymi aktualne piśmiennictwo związane z problematyką rozprawy, co świadczy o bardzo dobrej wiedzy doktoranta z tego zakresu. Dla poruszanych problemów doktorant sformułował ciekawe i użyteczne modyfikację modelu ProtoPNet oraz dwie propozycje wykorzystania poolingu atencyjnego SA-AbMILP oraz ProtoMIL. Doktorant przeanalizował wyniki przeprowadzonych badań eksperymentalnych ocenił jakość zaproponowanych metod na tle algorytmów znanych z literatury oraz wskazał możliwości dalszego rozwoju w obszarze związanym z rozprawą.

Recenzowana dysertacja przedstawia rozwiązanie ważnego i oryginalnego problemu, wzbogacając naszą wiedzę dotyczącą wyjaśnialności metod bazujących na sieciach neuronowych. Zawarte w niej wyniki badań eksperymentalnych wskazują również na możliwość wykorzystania otrzymanych metod w praktyce, co pokazano min. Na przykładzie zastosowania w problemach interpretacji z zakresu problemów chemicznych, dotyczących wyjaśnialności w zakresie oceny przydatności cząsteczek w zadaniu projektowaniu leków.

Przedstawione w punkcie 3 recenzji uwagi mają jedynie charakter dyskusyjny, w żaden sposób nie deprecjonując osiągniętych przez doktoranta rezultatów oraz nie wpływają na wyjątkowo pozytywne wrażenie o przedłożonej rozprawie. Jestem przekonany, że doktorant zdaje sobie sprawę z możliwości kontynuowania rozpoczętej w pracy tematyki, czemu dał wyraz prezentując skrótowo plan dalszych kierunków badań.

Reasumując, biorąc pod uwagę powyższe opinie i wymagania stawiane pracom stwierdzam, że rozprawa mgra inż. Dawida Rymarczyka pt. *Interpretable Deep Learning with Prototypical Parts for Supervised and Weakly-Supervised Learning* spełnia wymagania stawiane pracom doktorskim, w szczególności:

- Rozprawa zawiera oryginalne rozwiązanie problemu naukowego.
- Kandydat posiada ugruntowana, głęboką wiedzę w dyscyplinie Informatyka techniczna i telekomunikacja.
- Doktorant posiada umiejętność samodzielnego prowadzenia pracy naukowej.

Wnoszę o jej przyjęcie i dopuszczenie mgr inż. Dawida Rymarczyka do publicznej obrony, a biorąc pod uwagę wysoki poziom rozprawy, znaczenie dla dyscypliny informatyka techniczna i telekomunikacja oraz miejsca publikacji wyników, wnioskuje o jej wyróżnienie.

Podpis