Dr hab. inż. Maciej Piasecki
Politechnika Wrocławska
maciej.piasecki@pwr.edu.pl
ul. Wybrzeże Wyspiańskigo 27
50-370 Wrocław

17 sierpnia 2023

Recenzja rozprawy doktorskiej

Jakuba Pokrywki,

zatytułowanej:

Optimization and Evaluation in Machine Learning Challenges

1. Problem badawczy i jego znaczenie

Przedstawiona rozprawa składa się z 7 publikacji, których spis został przedstawiony na stronie 3. Wszystkie prace zostały opublikowane w recenzowanych materiałach konferencyjnych umieszczonych na liście ministerialnej, aczkolwiek dwie z tych prac (nr 2 i 4 na liście) zostały opublikowane w materiałach konferencji FedCSIS, która nie jest indeksowana w międzynarodowym rankingu CORE. Ponadto, praca numer 1 została opublikowana nie na samej konferencji NAACL (CORE A), tylko w „Findings of the Association for Computational Linguistics: NAACL 2022", czyli wśród prac dobrze ocenionych, ale jednak nie przyjętych do prezentacji podczas samej konferencji.
Uwaga marginalna, ponieważ podana została punktacja prac, to należy zauważyć, że przedstawia ona górne maksimum, jako że konferencja FedCSIS (prace 2 i 4) ma 20 pkt na liście ministerialnej (wydaje się, że w rozprawie potraktowano obie prace jako rozdziały w monografii), a w przypadku pracy nr 1 zwyczajowo przyjmuje się, że prace opublikowane w „Findings" mają punktację zgodną z daną konferencją, jednak nie jest to ta sama ranga. Pozostałe cztery prace (nr 3, 5-7) mają właściwą punktację, zostały opublikowane w materiałach konferencji o randze B w indeksie CORE (praca nr 3, opisana jako w druku, została w międzyczasie już opublikowana).
W cyklu są 3 prace jednoautorskie (5-7), pozostałych Doktorant występuje jako współautor. Jest to jak najbardziej zrozumiałe w tej dziedzinie, w której złożoność prac badawczo-eksperymentalnych wymaga współdziałania zespołowego (we wszystkich praca wieloautorskich, współautorem jest Promotor).
Tytuł rozprawy jest bardzo szeroki i zestawia ze sobą terminy odnoszące się do bardzo rozległych dziedzin badań, uprawianych od dziesięcioleci. Określenie „konkursów uczenia maszynowego" jako obszaru zainteresowania również nie pomaga w intuicyjnym zrozumieniu głównej osi tematu rozprawy, bo nadal mamy do czynienia z wielością i różnorodnością konkursów.
Przyglądają się cyklowi prac i ich treści, widzimy, że możemy je podzielić na trzy grupy:

- prace prezentujące rozwiązania dla wybranych zadań konkursowych w konkursach z dziedziny przetwarzania obrazów (prace nr 5-7),
- prace dotyczące przygotowania multimodalnego (obraz i tekst) zbioru testowego (praca nr 1 i w małym stopniu praca nr 4),
- prace prezentujące zastosowania technik uczenia maszynowego dla wybranych problemów: praca nr 2 dotyczy konstrukcji i trenowania modeli językowych, praca nr 3 estymacji czasu zapominania, m.in. przy użyciu metod uczenia maszynowego (w oparciu o inne dane niż tekst i obraz) oraz wspomniana już praca nr 4 dotyczy rozpoznania zamaskowanego rodzaju gramatycznego i jest oparta na zmodyfikowanych danych z konkursu.

Co łączy te prace? Z pewnością wykorzystanie metod uczenia maszynowego (w każdej z nich), osoba Doktoranta (warto podkreślić jego dużą aktywność – prace pochodzą z dwóch lat) oraz aspekt

testowania rozwiązań na opublikowanych zbiorach danych, co jednak jest prawie standardem w sztucznej inteligencji. Tytułowy aspekt „optymalizacji" jest w bardzo podstawowym stopniu obecny we wszystkich pracach, ale w większości z nich w tle. Konkretne rozwiązania zostały przedstawione i przetestowane, ale ich optymalizacji, analizie wariantów i porównaniu do stanu badań poświęcono wszędzie niewiele uwagi. Zagadnienie to będzie rozwinięte w dalszej części niniejszej recenzji. Ocena jest obecna w każdej pracy, ale każde rozwiązanie musi być ocenione. Ocena w sensie organizacji procesu testowania jest obecna najbardziej w pracy nr 1, ale praca ta przedstawia adaptację wcześniejszej metody zaproponowanej przed bieżącym cyklem prac.
Podsumowując, różnorodność prac świadczy dobrze o szerokich zainteresowaniach i umiejętnościach Doktoranta, ale znakomicie utrudnia postrzeganie przedstawionego cyklu prac jako „powiązanych tematycznie artykułów naukowych", zgodnie z wymogami ustawy. Przede wszystkim, bardzo trudno dojrzeć w tym cyklu prac konkretny cel badań i realizację planu badawczego w postaci spójnej sekwencji zadań badawczych dotyczących problemów szczegółowych. Osiągnięte wyniki, ciekawe w większości przypadków, trudno ułożyć w jedną całość prowadzącą do określonych wniosków z tej całości.
Dodatkowo, choć to tylko drugorzędna kwestia edycyjna, ułożenie prac w odwrotnym porządku chronologicznym nie sprzyja zrozumieniu jak plan badawczy autora był realizowany w ramach cyklu publikacji.
Problemu z postrzeganiem cyklu prac jako powiązanych tematycznie nie poprawiają niestety ani rozdział 1 – wprowadzający, ani też rozdział 2 – zawierający przegląd. Ten drugi przybrał postać bardzo suchej prezentacji każdej pracy z osobna w postaci rodzaju jednostronicowych rekordów metadanych, zawierających tylko krótkie opisy wyzwań, wkładu Doktoranta i abstrakt. Ponieważ każda praca cyklu jest opisana na osobnej stronie, te syntetyczne opisy nie pomagają w zrozumieniu spoiwa tematycznego cyklu.

Jaki jest najważniejszy problem rozważany w rozprawie?

Ze względu na brak wyraźniej spójności tematycznej prac, należy przyjrzeć się każdej prac z osobna i podejmowanym tam problemom. Z braku innego lepszego kryterium, omówione zostaną prace w porządku ich wymienienia w opisie cyklu.
W pracy nr 1 („Challenging America: Modeling language in longer time scales") przedstawiony został pomysł na wygenerowanie zestawu wzorcowych testów (ang. benchmarks) na podstawie historycznego korpusu języka angielskiego „Chronicling America". Zaproponowane zostały trzy wzorcowe zadania testowe: klasyfikacji temporalnej tekstów, geolokalizacji tekstów (z dokładnością do miejscowości) oraz uzupełniania luk. Opracowano również kilka wzorcowych modeli językowych znanych z literatury. Główną przesłanką pracy jest wykorzystanie bardzo dużego zbioru danych jakim jest korpus „Chronicling America", natomiast koncepcja testów opiera się na wcześniejszych pracach części współautorów. Szereg interesujących problemów naukowych, które się wyłaniają podczas lektury pracy, nie zostało przeanalizowanych dostatecznie uważnie lub wręcz pominiętych (praca ma ograniczoną objętość) w tym wpływ błędów OCR na wzorcowy test temporalny i wyniki modeli językowychosiąganych w nim oraz charakter zadania geolokalizacji, w którym miejsce wydawania jest brane jako metadana wzorcowa.
Praca nr 2 „Temporal Language Modeling for Short Text Document Classification with Transformers" dotyczy konstruowania modeli językowych, które łączą reprezentację tekstu z wymiarem temporalnym i są budowane na korpusach tekstowych z metadanymi temporalnymi. Temat ciekawy i ważny, ale uwaga koncentruje się na kilku zaproponowanych technikach.
Praca nr 3 („Modeling Spaced Repetition with LSTMs") w odróżnieniu od pozostałych nie dotyczy ani danych tekstowych, ani obrazowych, tylko danych numerycznych zebranych od użytkowników aplikacji wspomagających uczenie się i zapamiętywanie wiedzy. W pracy nr 3 zaproponowano

modyfikacje algorytmów uczenia maszynowego do estymacji czasu zapominania materiału przez ucznia dostosowane do specyfiki danych.
Praca nr 4 („Using Transformer models for gender attribution in Polish") wbrew swojemu tytułowi koncentruje się nie tyle na rozpoznaniu (w lit. atrybucji) płci autora tekstu (rodzaju mu właściwemu wg metadanych), tylko na zadaniu znacznie rzadziej poruszanemu w literaturze i trochę sztucznemu, czyli rozpoznaniu pierwotnego rodzaju gramatycznego dominującego w tekście. Tekst jest wpierw poddany swoistej normalizacji do gramatycznego rodzaju męskoosobowego, polegającej na zamianie oryginalnych form wyrazowych w pierwszej osobie (np. w rodzaju żeńskim) na ich analogi w rodzaju męskoosobowym. Sama procedura normalizacji jest deterministyczna i oparta na algorytmie i zasobach językowych. Należy tu podkreślić, że płeć (rodzaj z poziomu pozajęzykowego) autora bywa skorelowany z rodzajem gramatycznym dominującym w formach pierwszej osoby, ale z bardzo wielu powodów tak nie musi być (np. narracja, świadomy zabieg, podszywanie się itd.). Podsumowując, problem jest specyficzny, choć ciekawy, natomiast jego prezentacja w pracy jest wysoce myląca.
Prace nr 5-7 stanowią podzbiór związany z rozwiązywaniem różnych wzorcowych zadań testowych z dziedziny zastosowań przetwarzania obrazów.
W pracy nr 5 („YOLO with High Dataset Augmentation for Vehicle Class and Orientation Detection") zaproponowano rozwiązanie dla problemu wykrywania występowania na obrazie pojazdu, jego typu i orientacji. Doktorant wykorzystał podejście oparte o kombinację kilku znanych algorytmów.
W pracy nr 6 („Efficient GPU Training of a Diversified Model Ensemble for the Crowdsensing-based Road Damage Detection Challenge") są zaprezentowane nie tyle algorytmy poprawiające efektywność modeli neuronowych do klasyfikacji obrazów, co propozycja heurystyki sterującej procesem stopniowego douczania modeli neuronowych.
W pracy nr 7 („Gradient Boosted Trees for Privacy-Preserving Matching of Encrypted Images"). Doktorant koncentruje się na inżynierii cech na potrzeby reprezentacji obrazów w zadaniu identyfikacji obrazu po jego zakodowaniu. Wykorzystywane są znane algorytmy maszynowego uczenia. Zaproponowana reprezentacja jest silnie ukierunkowana na wybrane konkretne wzorcowe zadanie testowe (ang. benchmark).

Czy ma on charakter naukowy?

Z uwagi na duże zróżnicowanie problemów poruszanych w cyklu, a raczej ograniczonym czasowo zbiorze prac, konieczne jest odniesienie się do każdej z prac z osobna. Można jednak dostrzec ich kilka wspólnych cech. Są to prace krótkie (wszystkie to artykuły konferencyjne) i poświęcają bardzo mało uwagi porównaniu proponowanych rozwiązań do prac z literatury – wnikliwemu porównaniu z aktualnie najlepszymi rozwiązaniami w dziedzinie. Ponadto w pracach dotyczących algorytmów analiza błędów i analiza ablacyjna zaproponowanych rozwiązań jest w większości przykładów co najwyżej tylko dotknięta.
Głównym rezultatem pracy 1 jest ciekawy zasób naukowy pochodny z dostępnego wcześniej dużego korpusu tekstowego „Chronicling America". Zasób został zbudowany w oparciu o sprawdzoną procedurę wyboru części testowej, która umożliwia jej ciągłe rozszerzanie wraz z przyrostem materiału w oryginalnym źródle. Głównym ograniczeniem zaproponowanego zasobu są niestety błędy OCR zawarte w tekstach źródła i brak analizy tego problemu w pracy. Oczywiście jest to dalej kwestia odpowiedzialności osób stosujących zasób i testy wzorcowe, ale praktyka naukowa pokazuje, że zasoby tego typu zaczynają być używane jak czarne skrzynki.
W pracy nr 2 zaproponowano i przebadano kilka sposobów wprowadzania informacji temporalnej do modeli językowych. Opis większości z nich jest dość podstawowy, co utrudnia powtarzalność badań, ale szczęśliwie dla czytelnika najlepsze rezultaty wykazało podejście najbardziej bezpośrednie polegające na wprowadzeniu dat jako tekstowych prefiksów do danych treningowych.

W pracy 3 zaproponowano rozszerzenia kilku algorytmów w celu dopasowania ich do rozważanego zadania i zbioru danych.
W pracy nr 4 przebadano kilka modeli językowych w przyjętym specyficznym zadaniu rozpoznawania pierwotnego rodzaju gramatycznego przed normalizacją tekstu. Pomimo, iż praca nie wnika głęboko zarówno w cechy zbioru danych jak i własności metod klasyfikacji, poruszony problem i osiągnięte wyniki mają charakter naukowy.
Prace nr 5-7 to prace konkursowe – przedstawiają rozwiązania konkretnych wzorcowych zadań testowych. Koncentrują się na osiągnięciu jak najlepszego wyniku z sukcesem (2 i 3 miejsce). Z pewnością mają charakter godny uwagi i stanowią prace naukowe. Ponieważ bardzo brakuje w nich porównania ze stanem badań i głębszej perspektywy, to trudno powiedzieć na ile można uogólnić zaproponowane podejścia, ale prace są intrygujące. Szkoda, że któraś z nich, albo i wszystkie trzy, nie stały się punktem wyjścia do dalszych badań o pogłębionej refleksji.

Czy ma on znaczenie praktyczne?

Aspekt praktyczny jest silną stroną wszystkich prac z cyklu. Warto podkreślić, że we wszystkich pracach zaprezentowano rozwiązania otwarte, zarówno dane jak i kod. Pomimo ograniczeń wynikających z ograniczonej analizy i porównania, wszystkie prace cyklu wnoszą cenny wkład praktyczno-techniczny w rozwój poszczególnych obszarów badań.

2. Wkład autora

Wkład Autora rozprawy w poszczególne prace został opisany w informacjach podanych dla poszczególnych prac na str. 8-14. Opisy te są zgodne z deklaracjami podpisanymi przez wszystkich współautorów i dołączonymi do rozprawy.
Prace nr 5-7 są jednoautorskie i wkład Doktoranta jest równoważny z całością pracy.
W przypadku pracy nr 1, wkład Doktoranta miał charakter głównie techniczny i sprowadzał się do implementacji algorytmów wybranych przez pozostałych współautorów. Własny wkład naukowy ograniczył się do algorytmu selekcji tekstów i skanów z korpusu źródłowego oraz propozycji metryki do zadania geolokalizacji. Wkład ten można określić jako pomocniczy w stosunku do głównej wartości naukowej pracy.
W pracy nr 2 wkład Doktoranta był kluczowy i objął zaproponowane metody reprezentacji informacji temporalnej oraz ich analizę eksperymentalną.
W pracy nr 3 Doktorant zaproponował rozszerzenie jednej z przebadanych metod, czyli wniósł wkład w głównych aspekt kreatywny pracy. Ponadto był odpowiedzialny za analizę wyników badań eksperymentalnych (niestety dość ograniczoną, podobnie jak w przypadku pozostałych prac).
W pracy nr 4 Doktorant był odpowiedzialny głównie za prace implementacyjne oraz brał aktywny udział w napisaniu pracy.
Podsumowując, w ramach przedłożonego cyklu, Doktorant miał kluczowy lub znaczący udział w naukowej warstwie 5 z 7 prac, tj. 2, 3 i 5-7.

3. Poprawność

Czy stwierdzenia zawarte w rozprawie są godne zaufania?

Wyniki wszystkich prac włączonych do cyklu zostały poddane podstawowej weryfikacji naukowej. W każdym przypadku zadbano o spełnienie podstawowych wymogów testowania i wyniki przeprowadzonych badań są wiarygodne.
W przypadku pracy nr 1 i w pewnym stopniu też pracy nr 4 zwrócono uwagę na problem selekcji danych źródłowy oraz zadbano o odpowiednie wydzielenie zbiorów treningowego i testowego. Niestety nie

spojrzano wnikliwiej i szerzej na kwestię jakości zbiorów danych i wynikających z tego ich obciążeń. W przypadku zbioru z pracy nr 1 więcej uwagi wymaga kwestia błędów OCR i ich różnorodne korelacje, np. z okresem i miejscem wydania. Nie zwrócono też uwagi na semantyczne obciążenie wzorcowego zadania testowego dotyczącego temporalnego wymiaru tekstów. Możliwe, że przyjęto to jako część definicji zadania, ale nie jest to jawnie rozważone.
W przypadku pracy nr 4 całe zadanie zostało myląco nazwane. Jak już to było wcześniej wspomniane, w opisie zadania nie podkreślono jego specyficznego charakteru i nie odróżniono go od typowego zadania rozpoznania płci autora. Co więcej nie przeanalizowano konsekwencji deterministycznego charakteru algorytmu normalizacji danych dla zaproponowanych metod klasyfikacji tekstu.

Czy uzasadnienia są poprawne?

Przedłożone prace mają dobrą strukturę i większość zagadnień jest przedstawiona w jasny sposób. Niemniej, można dostrzec szereg kwestii, które nie są wyjaśnione w sposób całkowicie przekonywujący. W dużej mierze wynika to z ograniczonego rozmiaru prac i ich wąskiego ukierunkowania na zaprezentowanie rozwiązań własnych z mniejszą uwagą na ich kontekst i odniesienie do stanu badań.
W pracy nr 1 pada stwierdzenie (str. 31 wg numeracji w pliku PDF zawierającego rozprawę) dość niejasne a istotne: "but limited only to the 50% of texts that would be assigned to the training set". Nie wynika z tych kilku zdań, czy do zbioru treningowego trafiło tylko 50% wszystkich tekstów, czy też potencjalne przecięcie zbioru testowego i treningowego nie będzie większe niż 50%.
Ponadto w pracy mało uwagi jest poświęcone kwestii reprezentacji oryginalnego rozkładu danych w zbudowanych zbiorach treningowych i testowych.
Zarówno w pracy nr 1 jak i we wszystkich pozostałych brakuje próby analizy statystycznej istotności zaobserwowanych różnic pomiędzy wynikami poszczególnych metod czy modeli.
W pracy nr 2 nie zostało dostatecznie jasno wyjaśnione jak zostały zdefiniowane i wytrenowane osadzenia (ang. embeddings) dla wprowadzanych do modelu indeksów temporalnych. Osiągnięte rezultaty są dość rozczarowujące, ponieważ najprostsza metoda wprowadzenia do modelu językowego informacji temporalnej poprzez datę zapisaną jako tekst okazała się najlepsza. Brakuje jednak bardzo w pracy chociaż trochę bardziej wnikliwej analizy stanu badań. Prace nad różnymi metodami adaptacji i rozszerzania modeli językowych są praktycznie niezauważone. Możliwe, że słaby wynik osadzania wymiarów temporalnych bierze się z nieadekwatnej struktury modelu neuronowego.
Zarówno w tej pracy, jak i wszystkich pozostałych, praktycznie nie ma odpowiedniego porównania z metodami reprezentującymi aktualny stan badań. Oczywiście najlepiej byłoby je samemu uruchomić (często po odpowiednim dostosowaniu), sprawdzić wynik, przebadać porównawczo w replikowalny sposób. Praktycznie we wszystkich pracach cyklu brakuje wnikliwej analizy błędów i analizy ablacyjnej pokazującej, które czynniki decydują o osiągniętym wyniku oraz czy jest za to odpowiedzialna postulowana nowość podejścia.
W pracy nr 3, gdzie zaproponowano dość proste, ale ciekawe rozszerzania znanych algorytmów, powtarzają się wszystkie niedostatki wspomniane wcześniej (brak analizy statystycznej istotności, brak porównania, analizy błędów i analizy ablacyjnej). Ponadto rezultat, że "RNN with exponential decay" jest najlepszą metodą nie zaskakuje, a nie jest to dobrze skomentowane. Wszystkie badania przeprowadzono na jednym konkretnym zbiorze danych. Byłoby ciekawe zobaczyć jak poczynione obserwacje się generalizują na inne podobne problemy predykcji w czasie.
W pracy nr 4, poza problematycznym charakterem samego zbioru i zadania, nie jest jasne czy zbiory treningowe i testowe zostały rozdzielone odpowiednio na etapie pobierania materiałów źródłowych (np. kwestia duplikatów).

"We only used the data that was available in the HSSS challenge to avoid any data leaks in the other data sets." – czy oznacza to, że całość danych została użyta do zbudowania lub dotrenowania modeli językowych?
"We extracted the last layer tokens and averaged them. " – czy chodzi o tokeny czy też wektory tokenów?
Ciekawą inicjatywą było przeprowadzenie studiów nad zanieczyszczeniem modeli ("the contamination study").
Całość badań została przeprowadzona tylko na jednym zbiorze danych i dla jednego języka. Dobrze byłoby rozszerzyć badania na inne podobne zbiory, w tym na właściwe zadanie rozpoznania płci autora.
Całość pracy jest interesującym studium nad testowaniem modeli, ale bez oczekiwanego pogłębienia analizy pod względem zbioru danych i stosowanych metod.
W pracy nr 5 przede wszystkim nie jest wyjaśnione na czym polega główna nowość naukowa, metody kombinowania klasyfikatorów (ang. ensemble) są znane od dziesiątek lat. W pracy nie odniesiono się też dobrze do tej bogatej tradycji badań i wypracowanej głębokiej wiedzy.
W sumie praca sprowadza się do uważnej kombinacji znanych metod – bardzo udanej, co zaowocowało zajęciem 3 miejsca. Niemniej, całość prezentacji koncentruje się na własnej propozycji autora (brak analizy statystycznej istotności, brak porównania, analizy błędów i analizy ablacyjnej). Budzi wątpliwości co do generalizacji wyników poza konkretny zbiór i zadanie.
W pracy nr 6 nie zostało na wstępie dobrze określone jak będzie rozumiane pojęcie wydajności („efficient"). Jest to temat bardzo intensywnych badań od kilku lat i to wyłącznie ograniczając pole do samych modeli opartych na głębokich sieciach neuronowych: ich konstrukcji, redukcji, reprezentacji, wydajnego i efektywnego dostrajania modeli itd. Niestety, praca nr 6 nie odnosi się do tego zaawansowanego stanu badań w odpowiedni sposób. Za to cytowane są dość przypadkowe pozycje literaturowe, np. (str. 75) "[16] improves the Nepali language model tested on the Nepali test dataset using additional English and Hindi training datasets".
W pracy skupiono się na stopniowym dostrajaniu modelu, gdzie wydajność jest rozumiana jako ograniczenie rozmiaru wykorzystywanych danych. Tego typu podejście również pojawiało się już w literaturze, a brakuje do tego odniesień w pracy, np.

- Haoran Xu, Seth Ebner, Mahsa Yarmohammadi, Aaron Steven White, Benjamin Van Durme, Kenton Murray. Gradual Fine-Tuning for Low-Resource Domain Adaptation., https://aclanthology.org/2021.adaptnlp-1.22.pdf

Zaproponowaną metaheurystykę organizacji procesu trenowania przetestowano tylko w jednej wersji (przynajmniej tylko tyle widać w pracy). Podobnie jak w innych pracach: brak analizy statystycznej istotności, brak porównania, analizy błędów i analizy ablacyjnej. Rozwiązanie jest oparte na kombinacji kilku wersji modelu (ang. ensemble) – brak odniesienia do wiedzy z kombinowania klasyfikatorów, podobnie jak w przypadku pracy nr 5. Nie rozważono nawet podstawowych warunków jakie muszą spełnić klasyfikatory, aby ich kombinacja przyniosła pozytywny skutek. Przy nacisku w pracy na wydajność, umknęła kwestia narzutu obliczeniowego jaki wprowadza kombinacja klasyfikatorów.
Sformułowanie (str. 74) „The described approach focuses on efficient GPU utilization during model training, but without sacrificing the use of large object detection architectures" — jest niejasne ponieważ 'efficient GPU utilization' nie stoi w sprzeczności z 'large object detection architectures'.
(str. 75) "The improved version with corrected annotations and augmentations with a Generative Adversarial Network" — GAN jest rodzajem sieci neuronowej, nie jest jasne dlaczego jest wspominane w kontekście własności zbioru danych.
Przy testach na wszystkich dziedzinach vs jedna dziedzina sformułowanie "was further fine-tuned solely on a specific country's dataset, the performance on the test dataset was not better but slightly worse." rodzi obawy, że dane dla określonego kraju pojawiają się dwa razy w różnych wariantach rozszerzenia

danych (ang. augmentation), co nadawałoby im dodatkową ukrytą wagę i wprowadzało niejawne obciążenie do modelu.
Praca nr 7 przyniosła sukces w postaci 2 miejsca w konkursie, co należy odnotować i docenić. Niestety podobnie jak prace nr 5 i 6, jest bardzo skupiona na prezentacji własnego rozwiązania dla jednego konkretnego zadania i dwóch zbiorów danych. Przegląd literatury jest bardzo ograniczony i płytki. Doktorant wiele decyzji tylko komunikuje, np. „I chose Gradient Boosted Trees [12], which usually achieves the best results, as confirmed in many machine learning competitions" lub (str. 84) "I extracted the following features:" nie wchodząc w głębszą dyskusję nad intuicjami i odniesieniem do literatury. Główną nowością pracy jest zaproponowany zestaw cech do reprezentacji obrazów. Pojawia się tu kilka problemów:

- brak dyskusji intuicji i zbioru innych możliwych cech, porównania do literatury,
- duża liczba pochodnych złożonych cech,
- brakuje analizy korelacji pomiędzy cechami,
- nie ma porównania do dobrej metody nieopartej na inżynierii cech, np. jakiejś głębokiej sieci neuronowej,
- brakuje dyskusji do jakiego stopnia proponowane rozwiązanie eksploruje akcydentalne cechy konkretnych zbiorów danych z konkursu.

Biorąc pod uwagę charakter proponowanego rozwiązania, to brak analizy ablacyjnej jest uderzający.
"For negative pairs, I took all of the encrypted images paired with randomly selected original images." – jeżeli zakodowane obrazy pochodzą z negatywnych próbek, to wydaje się, że istnieje szansa, że zbudowane zostaną pozytywne pary.
Praca dotyczy w dużej mierze rozszerzania danych (ang. data augmentation), które jest obszarem badań o bardzo bogatej tradycji. Nie ma do tego odniesienia w pracy, ani porównania z wybranymi metodami z literatury.

4. Wiedza kandydata

Które z rozdziałów (lub sekcji w rozdziałach) rozprawy omawiają istniejący stan wiedzy i dzięki temu potwierdzają ogólny stan wiedzy kandydata w zakresie Informatyki?

Biorąc pod uwagę zaproponowane rozwiązania Doktorant wykazał się dobrą wiedzą na temat dostępnych metod z dziedziny uczenia maszynowego, jak i też z przetwarzania zbiorów danych. Natomiast mają na względzie niedostatki przedstawionych prac w zakresie prezentowania stanu badań, trudno ocenić na ile jest to wiedza dobrze ugruntowana w relewantnych obszarach.

Jakie obszary tych dyscyplin zostały omówione w tych rozdziałach/sekcjach? Jaka jest opinia recenzenta o jakości tych rozdziałów sekcji?

Zarówno same publikacje jak i ich prezentacja w rozprawie są bardzo skrótowe w odniesieniu do prezentacji stanu wiedzy w dyscyplinie.

Jaka jest opinia recenzenta o bibliografii? Na ile bibliografia jest kompletna?

Każda z prac wykazuje duże braki w bibliografii i posługuje się bibliografią w sposób wybiórczy.

5. Podsumowanie

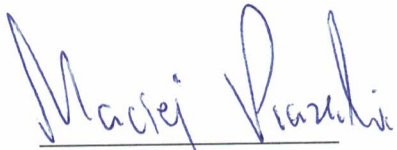
Biorąc pod uwagę opinie zaprezentowane w poprzednich punktach i wymagania zdefiniowane przez art. 187 Ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (z późniejszymi zmianami)[1] moja ocena rozprawy pod względem trzech podstawowych kryteriów jest następująca:

- rozprawa w postaci cyklu publikacji porusza szereg problemów naukowych, ale nie jest spójna tematycznie i przez to nie proponuje oryginalnego rozwiązania problemu naukowego, które wykazywałoby odpowiedni poziom dojrzałości i dogłębności,
- kandydat posiada ogólną wiedzę teoretyczną w dyscyplinie, o czym świadczy dobór rozwiązań dla podjętych problemów szczegółowych,
- oraz kandydat posiada umiejętność samodzielnego prowadzenia badań eksperymentalno-rozwojowych, ale nie wykazał się w przedstawionych pracach umiejętnością właściwego ugruntowania prowadzonych badań w oparciu o stan wiedzy, a także umiejętnościami analizy wyników badań.

Podsumowując, Doktorant zaprezentował szereg zainteresowań badawczych, ciekawych rozwiązań szczegółowych problemów i ogólnie dobry potencjał do dalszego prowadzenia pracy naukowej. W przedstawionej rozprawie brakuje koncentracji i pogłębionej pracy nad wybranymi tematami badawczymi, która pozwoliłaby na osiągnięcie bardziej dojrzałych wyników naukowych. Wydaje się, że po doprecyzowaniu tematu lub tematów, przeprowadzenie dalszych badań, przedstawienie wyników badań w postaci rozprawy pozwoliłoby na osiągniecie bardzo dobrego wyniku naukowego spełniającego zwyczajowe wymagania stawiane rozprawom doktorskim.

Przedstawiona rozprawa w obecnym kształcie cyklu publikacji nie spełnia wszystkich wymogów merytorycznych i formalnych stawianych przed pracami doktorskimi.

Wnoszę o skierowanie pracy do zasadniczej poprawy, sugeruję zmianę jej formy na rozprawę i niezbędnego uzupełnienie w zakresie przeprowadzonych badań i ich prezentacji.



Podpis

[1] http://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20190000276