dr hab. inż. Jakub Nalepa, prof. PŚ
Katedra Algorytmiki i Oprogramowania
Politechnika Śląska
Jakub.Nalepa@polsl.pl

Gliwice, 27 sierpnia 2024 r.

Recenzja rozprawy doktorskiej

Tytuł rozprawy: Strategie postępowania w przypadku małej liczby obserwacji uczących w problemie klasyfikacji danych nieustrukturyzowanych

Autor: Mgr Dominik Lewy

Promotor: Prof. dr hab. inż. Jacek Mańdziuk

1 Tematyka rozprawy

Recenzowana rozprawa doktorska Pana mgr. Dominika Lewego dotyczy rozwiązywania istotnego problemu związanego z ograniczoną dostępnością (lub z całkowitym brakiem) wysokiej jakości danych wzorcowych, które mogłyby zostać wykorzystane do treningu nadzorowanego metod uczenia maszynowego. Obecnie generowane są ogromne ilości danych, zarówno ustrukturyzowanych jak i nieustrukturyzowanych, w praktycznie każdej dziedzinie nauki i przemysłu. Ich ręczne opisywanie, tj. nadawanie odpowiednich etykiet danym surowym, niezbędnych do przeprowadzenia uczenia nadzorowanego, jest jednak bardzo kosztowne, czasochłonne oraz często wymaga specjalistycznej wiedzy dziedzinowej, a wypracowanie spójnych etykiet w praktyce wymaga wdrożenia odpowiednich procedur zapewnienia jakości danych wzorcowych, zwłaszcza jeśli dane opisywane są przez kilku ekspertów[1]. W literaturze istnieją metody syntetycznego wzbogacania istniejących zbiorów danych wzorcowych (zawierających dane o różnych modalnościach, np. obrazów barwnych, medycznych, serii czasowych czy danych tekstowych), poprzez syntetyczne generowanie danych treningowych [2–5].

W niniejszej rozprawie Doktorant podjął się opracowania metod, które mogą pomóc we wzbogacaniu istniejących zbiorów danych nieustrukturyzowanych (obrazów i tekstów). W ramach pracy zweryfikowano również, czy możliwe jest wytrenowanie sieci głębokich z wykorzystaniem danych o zróżnicowanej jakości (i często zaszumionych), które zostały pobrane z Internetu, bez wdrażania dodatkowych procedur zapewniania jakości takich danych. Warto podkreślić, że Doktorant skupił się dodatkowo na projektowaniu metod wzbogacania danych dla uczenia federacyjnego, w którym współdzielenie danych pomiędzy węzłami obliczeniowymi może być utrudnione lub niemożliwe, ze względu na aspekty związane z prywatnością i ochroną danych lokalnych.

Uważam, że tematyka rozprawy jest aktualna, zgodna z obecnym nurtem badań i dotyczy istotnych aspektów praktycznego wykorzystywania technik i modeli uczenia maszynowego, zwłaszcza tych o dużej pojemności, w rzeczywistych problemach, w których ilość danych wzorcowych jest ograniczona, a ich dodatkowe pozyskanie jest kosztowne lub nawet niemożliwe.

2 Ocena strony merytorycznej rozprawy doktorskiej

Rozprawa doktorska jest napisana w języku polskim i składa się z ośmiu rozdziałów oraz bibliografii.

[1] Warto też zauważyć, że nawet jeden ekspert może opisać te same dane w różny sposób, np. w przypadku segmentacji semantycznej obrazów medycznych, jeśli procedura opisu danych zostanie powtórzona kilkukrotnie [1].

Rozdział pierwszy jest krótkim wprowadzeniem do tematyki rozprawy, w którym Autor omawia cele i hipotezy badawcze. Cele te skupiają się wokół tworzenia nowych metod wzbogacania zbiorów danych treningowych, zwłaszcza opartych na technikach mieszania danych, oraz wykorzystania ich do wzbogacania zbiorów danych obrazowych i tekstowych, w przypadku których można wykorzystać ich specyficzne cechy i własności. Doktorant podkreśla również istotność zweryfikowania, czy opracowany algorytm wzbogacania zbiorów danych obrazowych może być użyty w treningu federacyjnym, w którym dane lokalne nie powinny (lub nie mogą) być przesyłane pomiędzy węzłami obliczeniowymi lub pomiędzy węzłami obliczeniowymi i serwerem głównym. Ostatnim z wymienionych celów pracy jest zweryfikowanie, czy dane obrazowe pozyskane z Internetu (zaszumione, o niepewnej jakości etykiet, ale objętościowo duże) mogą być z powodzeniem wykorzystane do wytrenowania wysokiej jakości modeli uczenia maszynowego.

W rozdziale pierwszym Doktorant omawia artykuły, w których opublikowane zostały niektóre z wyników umieszczonych w rozprawie doktorskiej. Moim zdaniem, na szczególną uwagę zasługują dwie publikacje – przegląd algorytmów wzbogacania danych treningowych opartych na technikach mieszania danych opublikowany w prestiżowym czasopiśmie *Artificial Intelligence Review* [6] (*Impact Factor* (2023): 10,7, 140 punktów ministerialnych), oraz artykuł [7] opisujący autorski algorytm *StatMix*, będący istotnym elementem niniejszej rozprawy (artykuł konferencyjny na 29^{th} *International Conference on Neural Information Processing*, ICONIP 2022, 70 punktów ministerialnych, CORE B). We wszystkich wymienionych artykułach Doktorant jest wiodącym (pierwszym) autorem.

Rozdział drugi jest zbiorem najistotniejszych definicji oraz pojęć, które są pomocne w zrozumieniu kolejnych części rozprawy. Autor krótko omawia modalności danych, którymi będzie zajmował się w pracy (podkreślając, że są to dane nieustrukturyzowane), wraz z kategoriami zadań rozważanych w literaturze i związanych z analizą takich danych (obrazowych i tekstowych). Doktorant podkreśla najistotniejsze różnice pomiędzy metodami opracowanymi w ramach pracy, przedstawia prostą architekturę systemu rozproszonego, w którym może być realizowane uczenie federacyjne, a także krótko omawia reprezentacje danych oraz architektury sieci głębokich, które będą wykorzystane w ramach eksperymentów. W ostatniej części rozdziału drugiego omówione zostały modele generatywne, które również mogą być z powodzeniem wykorzystywane do wzbogacania zbiorów danych treningowych.

W **rozdziale 3.** omówione zostały strategie rozszerzania zbiorów danych, które najogólniej można podzielić na ręczne (lub zautomatyzowane) tworzenie etykiet dla nowych danych nieoetykietowanych, wzbogacanie istniejących zbiorów oraz syntetyczne generowanie danych sztucznych o zadanych etykietach i charakterystyce. Jako jedną ze strategii rozszerzania danych treningowych Doktorant wymienia również „pozyskiwanie ogólnodostępnych etykietowanych obserwacji z różnych platform w Internecie". W **rozdziale 4.** Doktorant przedstawia i krótko omawia istniejące rozwiązania literaturowe dla każdej ze strategii omówionej w poprzednim rozdziale (dla danych obrazowych i tekstowych), szczególną uwagę poświęcając strategiom opartym na mieszaniu danych. Autor zapezentował również krótki przegląd technik wzbogacania danych treningowych, które mogą znaleźć zastosowanie w uczeniu federacyjnym oraz szereg strategii treningu modeli uczenia maszynowego, takich jak uczenie z przeniesieniem wiedzy czy uczenie stopniowe, które są pomocne w przypadku uczenia modeli na podstawie danych o zróżnicowanej jakości czy niewielkiej objętości.

W **rozdziałach 5.–7.**, Doktorant szczegółowo omawia trzy techniki opracowane w ramach niniejszej rozprawy, stanowiące jej trzon. Każdy z rozdziałów podzielony jest na trzy główne podrozdziały, w których przedstawiona jest (1) motywacja, która stała za opracowaniem każdej z technik, (2) opracowana metoda oraz (3) przeprowadzone eksperymenty i otrzymane wyniki eksperymentalne. W **rozdziale 5.**, Autor przedstawia pierwszą z opracowanych metod wzbogacania zbiorów danych treningowych, która została zaprojektowana dla danych tekstowych (dla problemu klasyfikacji tekstu). W algorytmie *AttentionMix*, Doktorant zaproponował wykorzystanie mechanizmów atencji w sieciach głębokich do kierowania procedurą mieszania przykładów treningowych. Wpływ użycia mechanizmu *AttentionMix* w procesie wzbogacania treningowych zbiorów danych na jakość wytrenowanych modeli głębokich został zweryfikowany z wykorzystaniem trzech benchmarkowych

zbiorów danych. W kolejnym **rozdziale szóstym** przedstawiona została metoda *StatMix*, będąca techniką wzbogacania zbiorów obrazowych danych treningowych, którą Doktorant opracował z myślą o jej wykorzystaniu w architekturze uczenia federacyjnego. Metoda *StatMix* nie tylko nie wymaga wysyłania danych lokalnych do innych węzłów obliczeniowych w takiej architekturze, ale także pozwala na istotne zredukowanie ilości danych przesyłanych pomiędzy węzłami (lub pomiędzy węzłami obliczeniowymi i serwerem głównym). Wpływ wykorzystania algorytmu *StatMix* na jakość działania sieci głębokich zweryfikowano dla dwóch popularnych architektur sieci splotowych, z wykorzystaniem dwóch benchmarkowych zbiorów danych (CIFAR-10 i CIFAR-100). W **rozdziale siódmym**, Doktorant przedstawił swoje badania dotyczące trenowania sieci głębokich z wykorzystaniem danych obrazowych, które mogą zostać pobrane z Internetu. Stworzenie takich zbiorów danych jest nieskomplikowane i może być wygodnie zautomatyzowane, ale takie zbiory (zwłaszcza bez wdrożenia dodatkowych procedur zapewnienia ich wysokiej jakości) mogą zawierać dane zróżnicowanej jakości, np. z niepoprawnymi etykietami. Doktorant, oprócz opracowania procedury przygotowania takich danych i treningu algorytmów uczenia maszynowego na ich podstawie, przygotował dwa zbiory danych: *InstaFood1M* oraz *InstaPascal2M*. Pierwszy z nich zawiera 1 milion obrazów barwnych pobranych z serwisu Instagram i przypisanych do odpowiedniej klasy (z 10 klas, odpowiadających najpopularniejszym potrawom w Stanach Zjednoczonych). Drugi z wymienionych zbiorów (zawierający 2,1 miliona obrazów barwnych) jest odpowiednikiem popularnego zbioru PascalVOC2007 [8]. Oba zbiory zostały podzielone na podzbiory treningowe, walidacyjne i testowe. W ramach badań eksperymentalnych Doktorant zweryfikował czy klasyfikator oparty na sieci głębokiej, która została wytrenowana z wykorzystaniem danych pobranych z serwisu Instagram może zapewnić klasyfikację wysokiej jakości pomimo tego, że takie dane mogą być zaszumione. Ponadto, w ramach dodatkowych eksperymentów, Autor zweryfikował trzy dodatkowo postawione hipotezy badawcze, w ramach których Doktorant skupił się na analizie wpływu wykorzystania wag klasyfikatora głębokiego pretrenowanego na zbiorze *ImageNet* na zbieżność treningu na nowych danych (oraz na jego możliwości generalizacji), na analizie wpływu rozszerzania zbiorów danych treningowych zawierających dane pozyskane z serwisu Instagram na jakość wypracowywanych modeli, a także na porównaniu jakości modeli trenowanych na mniejszych (reprezentatywnych) zbiorach danych oraz tych, które zostały wytrenowane z wykorzystaniem danych pozyskanych z serwisu Instagram.

W ostatnim **rozdziale 8.**, Autor podsumowuje rozprawę – rozdział rozpoczyna się od omówienia zalet i wad metod zaproponowanych przez Doktoranta. W kolejnych podrozdziałach Autor odnosi się do postawionych hipotez badawczych oraz omawia swój wkład w rozwój dyscypliny naukowej. Rozdział jest zwieńczony krótkim opisem możliwości dalszego rozwoju opracowanych technik.

Rozprawa doktorska mgr. Dominika Lewego prezentuje zarówno wiedzę teoretyczną jak i praktyczną Doktoranta w dyscyplinie *Informatyka Techniczna i Telekomunikacja*, a jej stronę merytoryczną oceniam zdecydowanie pozytywnie. W ramach pracy zaproponowano dwie metody wzbogacania treningowych zbiorów danych, dedykowane do danych tekstowych i obrazowych, a także zaimplementowano procedurę akwizycji, przygotowania i wykorzystania (w procesie treningu) objętościowo dużych zbiorów danych obrazowych, których jakość (zarówno obrazów jak i odpowiadających im etykiet) może być – i w praktyce jest – zróżnicowana. Na uwagę zasługuje fakt, że Autor opracował dwa zbiory danych obrazowych, które mogą być wykorzystane do tworzenia i oceny eksperymentalnej nowych algorytmów uczenia maszynowego trenowanych na rzeczywistych danych. W ramach badań eksperymentalnych Doktorant zweryfikował poprawnie postawione hipotezy badawcze i w konsekwencji osiągnął cele badawcze, które dotyczą istotnych problemów bezpośrednio związanych z opracowywaniem modeli uczenia maszynowego w problemach, w których wysokiej jakości, reprezentatywne i heterogenne zbiory treningowe są trudne, kosztowne lub niemożliwe do pozyskania.

Jestem przekonany, że Doktorant jest dobrze przygotowany do tego, żeby prowadzić dalsze badania związane z projektowaniem, implementowaniem i weryfikacją systemów wykorzystujących metody uczenia maszynowego, zwłaszcza dla rzeczywistych problemów dotyczących przetwarzania i analizy danych obrazowych i tekstowych, dla których pozyskanie dodatkowych oetykietowanych danych wzorcowych jest trudne lub niemożliwe, i które wymagają wdrożenia dodatkowych technik wzboga-

cania danych treningowych w przypadku uczenia nadzorowanego.

2.1 Pytania i uwagi

W trakcie lektury rozprawy nasunęły mi się następujące pytania i uwagi:

1. W *Streszczeniu* (na str. V) możemy przeczytać, że „Nieadresowane mogą prowadzić do błędów w wynikach (...)" – jakie błędy w wynikach miał tutaj Doktorant na myśli? Czy chodzi o brak możliwości generalizacji niepoprawnie wytrenowanego modelu uczenia maszynowego i o jego błędy, np. klasyfikacji, danych testowych, które nie były wykorzystywane w trakcie treningu?

2. W podrozdziale 1.1 (*Tematyka rozprawy*) czytamy, że niedobór danych uczących jest powszechnym problemem w kontekście zastosowań uczenia *głębokiego* w obszarze problemów biznesowych – moim zdaniem ten problem tak samo dotyczy wykorzystania metod klasycznego uczenia maszynowego, nie tylko uczenia głębokiego.

3. Dlaczego Doktorant uważa, że problem niedoboru danych uczących dotyczy „głównie danych nieustrukturyzowanych" (str. 1)? Czy w przypadku danych ustrukturyzowanych ten problem nie występuje?

4. Na str. 13 czytamy: „*(...) dane ręcznie anotowane (charakteryzujące się wysoką jakością i niewielką liczbą błędów)*" – uważam, że jest to duże „uproszczenie". W niektórych dziedzinach, np. w radiologii i w analizie obrazów medycznych [9], czy w analizie danych satelitarnych – ręczne etykietowanie danych, np. obrysowywanie interesujących obszarów, często prowadzi do etykiet o zróżnicowanej jakości (niekoniecznie wysokiej), zapewnienie powtarzalności i obiektywności procesu opisu danych są często trudne, a sam proces jest zależny od wielu czynników, takich jak czas etykietowania, praktyczne doświadczenie eksperta w analizie konkretnego rodzaju danych i wielu innych.

5. Co Doktorant miał na myśli pisząc (w *Definicji 6*), że „*Proces kodowania, prowadzony przez koder, pozwala na przekształcenie złożonych i abstrakcyjnych danych w bardziej zrozumiałą i kompaktową postać*" (str. 15)? Czy reprezentacje danych wejściowych, wypracowywane przez kodery są „bardziej zrozumiałe" niż np. wysokowymiarowe dane ustrukturyzowane lub obrazy, które mogą być danymi wejściowymi w takiej architekturze, np. typu autoenkoder?

6. Dlaczego Autor uważa, że ResNet jest „*zaawansowaną architekturą głębokiej sieci neuronowej*" (*Definicja 8*)? Kiedy architektura sieci neuronowej staje się „zaawansowana"?

7. W *Definicji 10* Doktorant zauważa, że „*Wprowadzenie DLA pozwala na skuteczniejsze i efektywniejsze modelowanie złożonych danych obrazowych oraz redukuje ilość wymaganych parametrów, co przekłada się na mniejszą złożoność obliczeniową i mniejsze ryzyko przeuczenia modelu.*" – korzystnie byłoby syntetycznie podsumować, np. w formie tabeli, wszystkie architektury sieci neuronowych, które zostały wykorzystane w rozprawie i uwypuklić ich własności (np. liczbę parametrów trenowalnych, liczbę warstw itd.).

8. Pewien niedosyt pozostawiają Rysunki 2.8 oraz 2.9 – każdy z rysunków, który jest umieszczony w rozprawie (i w dowolnym artykule naukowym) powinien być możliwie łatwy w interpretacji i zrozumieniu bez wgłębiania się w tekst (np. z Rys. 2.9 nie wiemy w jaki sposób powstaje macierz uwagi).

9. Co Autor miał na myśli przez „dane czyste" na str. 24 („*np. wiedzę można przenosić z danych czystych/z platform internetowych*")? Czy chodziło o dane bez etykiet, o dane niezaszumione czy o jeszcze inny „rodzaj" danych?

10. W podrozdziale 3.3 (*Generowanie danych syntetycznych*), Doktorant podkreśla, że syntetyczne tworzenie nowych danych, które nie bazuje na istniejących obserwacjach, może być realizowane na dwa sposoby, i że historycznie polegało na opracowaniu reguł, które pozwalały na stworzenie danych syntetycznych np. poprzez wklejenie spreparowanego logo na jednolity obszar losowego zdjęcia ze zbioru treningowego. Czy taka procedura nie bazuje na istniejących obserwacjach?

11. Na str. 30 czytamy: „*Przykładem takiej metody jest [2], gdzie niepewność liczona jest na podstawie gradientów*" – w jaki dokładnie sposób wyznaczana jest ta niepewność?

12. Doktorant dość pobieżnie opisał klasyczne metody wzbogacania zbiorów treningowych zawierających dane obrazowe – np. czy w przypadku kontrolowanego wstrzykiwania szumu to musi być „biały szum" (warto byłoby pogłębić przegląd istniejących metod, np. wstrzykiwania szumu [10])?

13. Bardzo dobrym pomysłem było umieszczenie w rozprawie przykładów wizualnych przedstawiających syntetycznie wygenerowane obrazy (np. Rysunek 4.1) – takie przykłady można byłoby przedstawić dla wszystkich omawianych metod. Warto byłoby także przedstawić wartości hiperparametrów dla wszystkich operacji, które zostały wykorzystane do wygenerowania przedstawionych obrazów. Podobnie, na Rysunku 4.5, Autor powinien dokładniej opisać w jaki sposób powstały „wzbogacone zdjęcia".

14. W rodziale 4. zabrakło mi syntetycznego i zwięzłego podsumowania omawianych metod wzbogacania danych treningowych, w którym Doktorant mógłby podsumować zalety i wady istniejących metod i ich najistotniejsze własności.

15. W jaki sposób wyznaczono zbiór walidacyjny w przypadku zbioru IMDB (str. 52)? Czy jeśli jest to 10% losowych przykładów ze zbioru treningowego, to czy rozkład liczby przykładów należących do każdej klasy ze zbioru treningowego został zachowany w zbiorze walidacyjnym?

16. Umieszczenie w rozprawie schematów blokowych w sekcjach dotyczących opisów eksperymentów znacznie ułatwiłoby ich zrozumienie.

17. Na str. 55, Doktorant podkreśla, że „*głównym modyfikowanym czynnikiem było dostosowanie liczby epok, co wynikało z konieczności przedłużenia procesu trenowania ze względu na wprowadzone wzbogacanie danych*" – dlaczego nie wprowadzono tzw. szybkiego warunku stopu (z ang. *early stopping*), który pozwoliłby na uniknięcie ręcznego dostosowywania liczby epok treningowych? Autor wskazuje również, że „*wartość zwiększana była do momentu wypłaszczenia się krzywych uczenia dla wszystkich eksperymentów*" – czy chodzi o krzywą (krzywe) wyznaczoną dla zbioru treningowego czy walidacyjnego?

18. W rozdziale 5. zabrakło mi umieszczenia i omówienia rzeczywistych syntetycznych przykładów otrzymanych za pomocą opracowanej metody *AttentionMix* dla każdego z analizowanych zbiorów danych (warto byłoby zaprezentować przykłady, w odpowiednim rozdziale, dla każdej z opracowanych metod wzbogacania danych, tj. *AttentionMix* i *StatMix*, dla każdego z analizowanych zbiorów danych).

19. W sekcji *Dodatkowe badania* (w podrozdziale *5.3.4 Wyniki*) Doktorant zaprezentował bardzo ciekawe wyniki i analizę badań eksperymentalnych – warto byłoby rozszerzyć tę analizę, która ułatwia interpretację otrzymywanych wyników.

20. Czy Doktorant rozważał wykorzystanie kilku metod augmentacji danych tekstowych do wzbogacenia danych treningowych jednocześnie? Czy moglibyśmy oczekiwać dalszego polepszenia jakości wyników otrzymywanych dla zbiorów testowych w przypadku wdrożenia kilku technik wzbogacania danych jednocześnie?

21. Dlaczego w niektórych eksperymentach Doktorant wykorzystał optymalizator Adam, a w niektórych SGD?

22. W opisie „*Części serwerowej*" metody *StatMix* (str. 70) możemy przeczytać, że w każdym z N węzłów znajduje się $\boldsymbol{K}$ lokalnych zdjęć. Dlaczego zakładamy, że w każdym węźle znajduje się dokładnie tyle samo zdjęć? Moim zdaniem takie założenie jest zbyt mocne i raczej niepraktyczne.

23. Czy Doktorant rozważał porównanie metody *StatMix* z innymi technikami wzbogacania zbiorów danych treningowych, które mogłyby łatwo zostać wykorzystane w treningu federacyjnym, np. metodami opartymi na wstrzykiwaniu szumu do danych lokalnych?

24. Czy Doktorant rozważał wykorzystanie opracowanej metody wzbogacania zbiorów danych obrazowych do próby polepszenia najlepszych obecnie znanych modeli uczenia głębokiego dla zbiorów CIFAR-10/CIFAR-100? Taki eksperyment mógłby pomóc w zweryfikowaniu, czy opracowane techniki augmentacji danych mogą pomóc w polepszeniu jakości modeli, które charakteryzują się bardzo dobrą jakością i możliwościami generalizacji.

25. W rozprawie pojawiają się nieprecyzyjne sformułowania, np. w pierwszej hipotezie, przedstawionej na str. 93, Autor pisze, że *„klasyfikator inicjalizowany losowymi wagami będzie zbiegał wolniej, niż ten inicjalizowany wagami z ImageNet, ale jego skuteczność na danych reprezentatywnych nie będzie dużo niższa"* – kiedy, według Doktoranta, skuteczność staje się „dużo niższa"? Czy w tym wypadku możemy zdefiniować jakąś konkretną wartość „progu"?

26. Czy Doktorant rozważał wdrożenie metod wzbogacania zbiorów danych treningowych do wzbogacenia zbiorów danych pozyskanych z serwisu Instagram?

27. Pomimo tego, że zbiory testowe wykorzystane w rozprawie są zbalansowane, niedosyt pozostawia wykorzystanie wyłącznie dokładności jako miary jakości analizowanych modeli uczenia maszynowego.

28. W pracy zabrakło mi przeprowadzenia odpowiednich testów statystycznych – czy różnice pomiędzy opracowanymi technikami wzbogacania zbiorów treningowych i tymi znanymi z literatury są statystycznie istotne?

29. Na str. 88, Doktorant podał link do repozytorium, w którym możemy znaleźć kod wykorzystany do badań eksperymentalnych (w ramach ostatniej grupy eksperymentów omówionej w rozprawie). Zapewnienie powtarzalności eksperymentów i ich reprodukowalności jest obecnie szczególnie istotne, zwłaszcza w obliczu „kryzysu reprodukowalności", z którym musimy się mierzyć [11] – korzystnie byłoby, gdyby Autor zebrał wszystkie implementacje, które zostały wykorzystane w ramach prac w jednym repozytorium.

30. Warto byłoby zaproponować jasny podział na podzbiory treningowe, walidacyjne i testowe opracowanych zbiorów danych oraz zestawy metryk, które zawsze powinny być wyznaczane w trakcie porównywania algorytmów weryfikowanych z wykorzystaniem opracowanych zbiorów. Tak przygotowane zbiory danych (często nazywane „AI-ready") mogłyby zostać wykorzystane do dalszych badań, prowadzonych również przez inne grupy badawcze. Być może warto byłoby zorganizować konkurs na podstawie zebranych danych, którego celem mogłoby być stworzenie jak najlepszych modeli uczenia maszynowego trenowanych na zaszumionych (rzeczywistych) danych. Czy według Doktoranta zorganizowanie takiego konkursu/hackathonu miałoby sens? Jeśli tak, jak Doktorant zaprojektowałby taki konkurs?

3 Ocena strony formalnej rozprawy doktorskiej

3.1 Ocena układu pracy

Układ rozprawy doktorskiej jest poprawny i logiczny, a kolejne rozdziały wprowadzają czytelnika w najistotniejsze aspekty prac Doktoranta.

3.2 Ocena zastosowanego piśmiennictwa

Bibliografia zawiera *około* 119 pozycji, które zostały uporządkowane alfabetycznie. Dobór prac jest poprawny i jest jednym z dowodów na to, że Autor bardzo dobrze orientuje się w aktualnym nurcie badań związanych z tematyką rozprawy. W trakcie analizy wykorzystanego piśmiennictwa zauważyłem niewielkie uchybienia:

- W niektórych wpisach w bibliografii możemy zauważyć niepoprawne wykorzystanie wielkich i małych liter, np. zamiast „*Polish AI*" czytamy „*polish ai*" (w [17]); w [12] Autor używa zapisu „BERT" i „bert".
- W bibliografii możemy znaleźć niekompletne wpisy, w których brakuje np. stron (np. [4], [5], [13]). Brakuje też spójności zapisu nazw konferencji – w niektórych wpisach w bibliografii nazwy konferencji pojawiają się zapisane wyłącznie skrótem (np. CVPR w [13]), a w innych – pełną nazwą (np. w [3]).
- W bibliografii zauważyłem powtórzony wpis ([46] i [47]), dlatego napisałem, że bibliografia zawiera *około* 119 pozycji.
- Unikałbym umieszczania odnośników do artykułów na początku zdania, np. „[87] proponuje miarę niepewności (...)" (str. 29).

3.3 Uwagi redakcyjne

Rozprawa jest napisana starannie, a tekst jest poprawny pod względem językowym, stylistycznym i interpunkcyjnym (zauważyłem drobne błędy interpunkcyjne, np. nadmiarowe przecinki oraz brakujące spacje, np. „CutMix[98] na str. 34). W pracy można dostrzec pewne uchybienia:

1. W rozprawie możemy zauważyć niepotrzebne kalki z j. angielskiego, np. „nieadresowane [wyzwania]" w *Streszczeniu*, „Metoda ta adresuje (...)" (str. 36).
2. Autor używa dywizu zamiast (pół)pauzy (myślnika).
3. Raczej unikałbym „personifikowania" algorytmów i metod, np. „Metoda *AttentionMax* pokazała..." (w *Streszczeniu*).
4. W rozprawie pojawiają się niefortunne sformułowania, np. „Dodatkowo wykazano, że modele wysokiej jakości można skutecznie uczyć na podstawie zaszumionych danych, pobranych z internetu" – w tym wypadku takie modele stają się dopiero „wysokiej jakości" (tj. posiadają możliwość generalizacji) po odpowiednim przeprowadzeniu treningu.
5. W rozprawie zauważyłem nieliczne błędy typograficzne, np. „sieroty" (str. V).
6. W streszczeniu w j. angielskim Doktorant używa zarówno brytyjskiego jak i amerykańskiego angielskiego – powinno to zostać uspójnione.
7. Cudzysłów powinien być zapisany z wykorzystaniem pisowni polskiej, np. „State of Polish AI" zamiast "State of Polish AI" (str. 1).
8. W rozprawie możemy natknąć się na literówki, zarówno w tekście, np. „generatywna sieć adversarialna" (str. 20), „ulegaja zmniejszeniu" (str. 23), „kmneas++" (str. 30), „stworzonego do mierzenie skuteczności" (str. 85), jak i w rysunkach, np. „dane lokalna" (Rysunek 2.2), „Zdjcie" (Rysunek 4.4), „Zjęcie" (Rysunek 4.9), „Średnia waga atencji poszczególnych części mowy na tle inny" (Rysunek 5.7); zamiast „3.4 miliarda" powinno być „3,4 miliarda" (str. 46), zamiast „51.17%" powinno być „51,17%" (str. 56).
9. Podpis Tabeli 7.1 powinien być umieszczony nad nią.
10. W rozprawie zauważyłem niezdefiniowane skróty (np. VGG, SGD itd.) – każdy skrót powinien być zawsze zdefiniowany przy jego pierwszym użyciu.
11. Autor miejscami niepoprawnie używa „ilość" i „liczba", np. zamiast „ilość tokenów w zdaniu" powinno być „liczba tokenów w zdaniu" (Rysunek 2.5), zamiast „ilością kroków dodawania szumu" powinno być „liczbą kroków dodawania szumu" (Rysunek 2.12), zamiast „ilość pikseli w zdjęciu" powinna być „liczba pikseli w zdjęciu" (str. 28), zamiast „ilość zdjęć w ramach węzła" powinna być „liczba zdjęć w ramach węzła" (Algorytm 2, str. 69).

12. W pracy możemy znaleźć rysunki, w których pojawia się tekst w języku polskim (np. Rysunek 2.2), tekst w języku angielskim (Rysunek 4.10), lub tekst w języku polskim i angielskim (Rysunek 2.4) – powinno to zostać uspójnione (sugerowałbym j. polski, ponieważ rozprawa została napisana w j. polskim). Zauważyłem również, że w niektórych rysunkach tekst jest niepoprawnie sformatowany (np. „Reprezentacja wektorowa" na Rysunku 2.9), a czcionka jest czasami zbyt mała (np. na Rysunku 5.1).

13. Dlaczego w *Definicji 9* „warstwy Normalizacji partii danych" są zapisane z wielkiej litery?

14. Po równaniach, np. po Równaniu 5.1, powinien pojawić się przecinek. Usunąłbym też wcięcie przed „gdzie n jest liczbą tokenów (...)". Inne równania, np. 5.2 i 5.3, mogłyby zostać „wplecione" w tekst.

15. Niektóre tytuły sekcji są dość niefortunne i mogłyby być bardziej informatywne, np. *„2.2 Kontekst metod"* czy *„2.3 Kontekst eksperymentów"*.

16. Czy podrozdział *„4.1 Efektywne strategie rozszerzania danych"* nie powinien nazywać się *„4.1 Strategie wyboru danych do stworzenia etykiet/etykietyzacji"*?

17. W *Definicji 7* czytamy „redukcji wymiarowości (angielski termin max pooling)" – max pooling jest jednym z rodzajów redukcji wymiarowości (angielski termin dla wyrażenia „redukcja wymiarowości" brzmiałby „dimensionality reduction").

18. W pracy możemy zauważyć brak spójności zapisu niektórych terminów (np. spektrum vs. spectrum).

19. Najlepsze wyniki w każdej z tabel powinny być pogrubione, żeby ułatwić ich analizę.

20. W Tabeli 5.1 Autor używa j. polskiego i angielskiego (warto byłoby również przemyśleć nagłówki kolumn w tej tabeli).

Powyższe uchybienia nie wpływają na mój pozytywny odbiór rozprawy doktorskiej.

4 Konkluzja

Z pełnym przekonaniem stwierdzam, że recenzowana dysertacja Pana mgr. Dominika Lewego **spełnia** wymagania stawiane rozprawom doktorskim przez Ustawę – praca prezentuje ogólną wiedzę teoretyczną i praktyczną Doktoranta w dyscyplinie *Informatyka Techniczna i Telekomunikacja* oraz umiejętność samodzielnego prowadzenia pracy naukowej, a przedmiotem rozprawy doktorskiej jest oryginalne rozwiązanie precyzyjnie zdefiniowanego problemu naukowego.

W związku z powyższym, **wnioskuję o przyjęcie rozprawy doktorskiej oraz o dopuszczenie mgr. Dominika Lewego do publicznej obrony**.

Jakub Nalepa

Literatura

[1] D. Gut, M. Trombini, I. Kucybała, K. Krupa, M. Rozynek, S. Dellepiane, Z. Tabor, W. Wojciechowski, Use of superpixels for improvement of inter-rater and intra-rater reliability during annotation of medical images, Medical Image Analysis 94 (2024) 103141. doi:https://doi.org/10.1016/j.media.2024.103141.
URL https://www.sciencedirect.com/science/article/pii/S1361841524000665

[2] C. Shorten, T. M. Khoshgoftaar, A survey on image data augmentation for deep learning, Journal of Big Data 6 (1) (2019) 60. doi:10.1186/s40537-019-0197-0.
URL https://doi.org/10.1186/s40537-019-0197-0

[3] C. Shorten, T. M. Khoshgoftaar, B. Furht, Text data augmentation for deep learning, Journal of Big Data 8 (1) (2021) 101. doi:10.1186/s40537-021-00492-0.
URL https://doi.org/10.1186/s40537-021-00492-0

[4] A. Mumuni, F. Mumuni, Data augmentation: A comprehensive survey of modern approaches, Array 16 (2022) 100258. doi:https://doi.org/10.1016/j.array.2022.100258.
URL https://www.sciencedirect.com/science/article/pii/S2590005622000911

[5] G. Iglesias, E. Talavera, Á. González-Prieto, A. Mozo, S. Gómez-Canaval, Data augmentation techniques in time series domain: a survey and taxonomy, Neural Computing and Applications 35 (14) (2023) 10123–10145. doi:10.1007/s00521-023-08459-3.
URL https://doi.org/10.1007/s00521-023-08459-3

[6] D. Lewy, J. Mańdziuk, An overview of mixing augmentation methods and augmentation strategies, Artificial Intelligence Review 56 (3) (2023) 2111–2169. doi:10.1007/s10462-022-10227-z.
URL https://doi.org/10.1007/s10462-022-10227-z

[7] D. Lewy, J. Mańdziuk, M. Ganzha, M. Paprzycki, Statmix: Data augmentation method that relies on image statistics in federated learning, in: M. Tanveer, S. Agarwal, S. Ozawa, A. Ekbal, A. Jatowt (Eds.), Neural Information Processing, Springer Nature Singapore, Singapore, 2023, pp. 574–585.

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman, The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results, http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[9] M. Benchoufi, E. Matzner-Lober, N. Molinari, A.-S. Jannot, P. Soyer, Interobserver agreement issues in radiology, Diagnostic and Interventional Imaging 101 (10) (2020) 639–641. doi:https://doi.org/10.1016/j.diii.2020.09.001.
URL https://www.sciencedirect.com/science/article/pii/S2211568420302175

[10] M. E. Akbiyik, Data Augmentation in Training CNNs: Injecting Noise to Images (2023). arXiv:2307.06855.
URL https://arxiv.org/abs/2307.06855

[11] S. Kapoor, A. Narayanan, Leakage and the reproducibility crisis in machine-learning-based science, Patterns (Aug. 2023). doi:10.1016/j.patter.2023.100804.
URL https://www.cell.com/patterns/abstract/S2666-3899(23)00159-9