Politechnika Warszawska
Wydział Elektroniki i Technik Informacyjnych

Warszawa, 23 kwietnia 2019 r.

D z i e k a n a t

Uprzejmie informuję, że na Wydziale Elektroniki i Technik Informacyjnych Politechniki Warszawskiej odbędzie się w dniu 14 maja 2019 r. publiczna obrona rozprawy doktorskiej

Mgr. Karola Piczaka

temat: „Klasyfikacja dźwięku za pomocą splotowych sieci neuronowych"

promotor	dr hab. inż. Jarosław Arabas, prof. Politechniki Warszawskiej Wydziału Elektroniki Technik Informacyjnych
recenzenci:	prof. dr. hab. inż. Bożena Kostek z Wydziału Elektroniki, Telekomunikacji i Informatyki Politechniki Gdańskiej
	prof. dr. hab. inż. Krzysztof Krawiec z Wydziału Informatyki Politechniki Poznańskiej

Obrona odbędzie się w dniu 14 maja 2019 r. w sali 116 na Wydziale Elektroniki i Technik Informacyjnych – Gmach im. Janusza Groszkowskiego, Warszawa, ul. Nowowiejska 15/19; początek godz. 11.00

Po adresem: **www.elka.pw.edu.pl/Wydzial/Rada-Wydzialu/Harmonogram-obron-doktorskich-streszczenia-i-recenzje** zapewniony jest na stronie Wydziału dostęp do tekstów streszczenia rozprawy i recenzji, jak również do tekstu rozprawy umieszczonej w Bazie Wiedzy Politechniki Warszawskiej.

Dziekan

prof. dr hab. inż. Krzysztof Zaremba

Mgr Karol Piczak

Promotor – dr hab. inż. Jarosław Arabas, prof. PW
Tytuł rozprawy:
Klasyfikacja dźwięku za pomocą splotowych sieci neuronowych

Streszczenie:
Niniejsza rozprawa skupia się na temacie wykorzystania splotowych sieci neuronowych do klasyfikacji dźwięku. Jej celem jest wykazanie, że modele tego typu, których efektywność została wcześniej potwierdzona w licznych zagadnieniach rozpoznawania obrazów, można z powodzeniem zastosować również w zadaniach klasyfikacji dźwięków o ogólnym charakterze i to pomimo występującej między tymi obszarami dysproporcji w dostępności etykietowanych zbiorów danych.
Rozprawa prezentuje jedne z pierwszych opublikowanych w literaturze przykładów użycia splotowych sieci neuronowych do klasyfikacji dźwięków środowiskowych i rozpoznawania gatunków ptaków śpiewających. Zaproponowana w tym celu metoda opiera się na przetwarzaniu spektrogramów wyrażonych w skali melowej za pomocą sieci wykorzystujących wertykalne filtry w pierwszej warstwie splotowej. Podejście takie zapewnia połączenie dobrej dokładności klasyfikacji z korzystnymi parametrami wydajnościowymi w porównaniu do architektur splotowych typowych dla przetwarzania obrazów. Zarówno wyniki przeprowadzonych eksperymentów, jak i pozytywny odbiór koncepcji przez szerszą społeczność zajmującą się tą tematyką, potwierdzają, że splotowe sieci neuronowe są obiecującym narzędziem w obszarze rozumienia dźwięku.
Poza wyczerpującym omówieniem literaturowym tematyki splotowych sieci neuronowych i klasyfikacji dźwięku, rozprawa zawiera również szczegółową analizę wrażliwości zaproponowanych modeli na zmiany wartości hiperparametrów. Zestawienie to jest jednym z najszerszych porównań tego typu przeprowadzonych dotychczas w literaturze przedmiotu.
Integralnym efektem prac badawczych podjętych w ramach rozprawy jest także utworzenie zbioru nagrań środowiskowych „ESC-50”, mające na celu poprawę sytuacji ograniczonej publicznej dostępności zasobów z etykietowanymi danymi tego typu. Znaczenie tej inicjatywy potwierdzają liczne publikacje innych autorów wykorzystujące ten zbiór jako punkt odniesienia w przeprowadzanych eksperymentach.

Słowa kluczowe:
splotowe sieci neuronowe, klasyfikacja dźwięku, spektrogram, ESC-50

Prof. dr hab. inż. Krzysztof Krawiec
Instytut Informatyki
Politechnika Poznańska
ul. Piotrowo 2
60-965 Poznań

Poznań, 23.02.2019

Recenzja rozprawy doktorskiej mgr Karola Jerzego Piczaka pt. "Klasyfikacja dźwięku za pomocą splotowych sieci neuronowych"

1 Tematyka rozprawy

Tematem rozprawy mgr Karola J. Piczaka są algorytmy analizy dźwięku, a dokładniej klasyfikacji nagrań dźwiękowych nie będących mową ani muzyką, czyli tzw. dźwięków środowiskowych. Doktorant skupił się na badaniu i rozwijaniu szerokiej gamy głębokich splotowych sieci neuronowych, które stanowią obecnie najlepiej sprawdzającą się w tym obszarze zastosowań i najintensywniej studiowaną klasę modeli uczenia maszynowego.

W związku z tym tematykę pracy mgr Piczaka uważam za aktualną, i zdecydowanie umiejscowioną w dyscyplinie informatyka, a dokładniej w nośnym obecnie nurcie uczenia maszynowego, analizy danych i odkrywania wiedzy.

2 Ocena treści rozprawy i wkładu oryginalnego

2.1 Ocena treści rozprawy

Rozprawa jest obszerna: składa się z 8 rozdziałów, jednego załącznika, bibliografii, i ma w sumie 211 stron.

Rozdział 1 motywuje zakres prac zrealizowany w rozprawie, wskazując m.in. na szeroką gamę potencjalnych zastosowań systemów klasyfikacji dźwięków środowiskowych. Autor zdefiniował tu też szczegółowe cele rozprawy, które w większości adekwatnie adresują wyzwania związane z tym obszarem badań.

Rozdział 2 stanowi przystępne i aktualne wprowadzenie do aktualnego stanu prac naukowych i praktyki głębokich sieci neuronowych, ze szczególnym naciskiem na sieci splotowe. Autor przystępnie i zwięźle wprowadza pojęcia, w tym te bardziej wymagające

(np. model przesunięcia, s. 50). Zwraca uwagę kompetentne oddanie perspektywy historycznej, np. w sięganiu do najstarszych architektur splotowych jak neocognitron. Jakość prezentacji materiału jest tak wysoka, że nie zawahałbym się rekomendować go np. studentom jako zwięzłego tutorialu czy krótkiego skryptu. Z drugiej strony można zadać pytanie czy rozdział ten nie został nadmiernie rozbudowany (liczy aż 53 strony).

Prezentacja problemu klasyfikacji dźwięku jako zadania uczenia maszynowego, dokonana w rozdziale 3, przeprowadzona jest kompetentnie i przystępnie. Dobrze przemyślane diagramy znakomicie ilustrują pojęcia. Zakres prezentowanych zagadnień (np. prezentacja nie tylko cech spektralnych, ale też cepstralnych i MFCC) potwierdza bardzo dobrą orientację Autora w temacie. Przejawia się to szczególnie w sekcji 3.3, gdzie Autor porusza nie tylko czysto naukowe aspekty rozwoju tego obszaru badań (oraz w konsekwencji jego obecnego stanu), ale także aspekty środowiskowe (sekcja 3.3.2). Co prawda uważam że pewne podstawowe pojęcia (np. typy błędów na s. 78, niektóre podstawowe miary prezentowane na kolejnych stronach) można było wprowadzić zwięźlej, lub nawet pominąć, zakładając ich znajomość u czytelnika, ale przyznaję że ich prezentacja czyni pracę bardziej zamkniętą (w sensie *self-contained*).

Decyzja Autora o przygotowaniu nowego zbioru danych, opisanego w Rozdziale 4, wydaje się dobrze uzasadniona i przemyślana, stanowiąc istotne uzupełnienie niedoskonałości dotąd dostępnych zbiorów. Zbiór jest stosunkowo obszerny zarówno w sensie liczby klas, jak i przykładów. Zostało to potwierdzone popularnością tej kolekcji, przekonująco zaprezentowaną przez Autora w sekcji 4.5 (ponad 50 cytowań od 2016 roku!). Dodatkowym atutem zbioru jest przeprowadzenie jego weryfikacji na respondentach-ludziach (sekcja 4.4.1) i podstawowych klasyfikatorach (sekcja 4.4.2).

Metodyka eksperymentalna obrana w rozdziale 5 jest poprawna, a nawet wychodzi ponad standardy (np. w stosowaniu walidacji krzyżowej (sekcja 5.3), którą często zarzuca się w przypadku głębokich sieci neuronowych z racji wysokich kosztów obliczeniowych procesu uczenia, czy strojenia parametrów na osobnym podzbiorze danych). Proponowane architektury zawierają interesujące i nietrywialne rozszerzenia, np. podawanie sieci spektrogramu różnicowego w drugim kanale, czy różne typy agregacji sugestii sieci (probability voting i majority voting). Wyniki zostały zwizualizowane na wiele sposobów, zestawione z wynikami eksperymentu crowdsourcingowego, wnikliwie przeanalizowane, z wyraźnym uwzględnieniem wiedzy dziedzinowej (w tym m.in. charakterystyki percepcyjnej klasyfikowanych dźwięków). Autor wykazał się umiejętnością projektowania nowoczesnych architektur neuronowych, i w szczególności pokazał jak zastosowanie usprawnień opracowanych w ostatnich latach pozwala na polepszenie lub utrzymanie (w zależności od kontekstu i zbioru danych) trafności klasyfikowania nowszych modeli (rodzina E2018) względem tych starszych (rodzina modeli E2015), przy jednoczesnej redukcji liczby parametrów o rząd wielkości.

Wyniki zaprezentowane w rozdziale 5 potwierdzają główną tezę rozprawy, a w szczególności możliwość realizowania wymagających zadań klasyfikacji dźwięków środowiskowych przy pomocy sieci zawierających w praktyce jedynie warstwy splotowe. Inny interesujący wynik to obserwacja że wnioskowanie na podstawie części nagrania może być bardziej skuteczne niż na podstawie całości (s. 127). Dodatkowym przyczynkiem

jest wnikliwa analiza działania klasyfikatora, przeprowadzona na dwa sposoby: z wykorzystaniem przykładów syntetycznych (optymalizowanych gradientowo) i rzeczywistych (pochodzących ze zbioru danych), oraz analiza przyczyn pomyłek klasyfikatora poprzez symulowanie niedostępności części danych, co pozwoliło autorowi na wskazanie części spektrogramu które mogą być 'zwodnicze' dla klasyfikatora. Z drobnych uwag krytycznych, zabrakło mi nieco opisu jakie jeszcze architektury Autor przetestował na drodze do E2018, a spodziewam się że proces dobierania architektur, hyperparametrów, etc. był czasochłonny – niemniej jak okazuje się dalej tego typu studium zaprezentowane jest w rozdziale 7.

Doktorant wykorzystał pozyskaną wiedzę w studium przypadku dotyczącym klasyfikowania głosów dużej (999) grupy ptaków w konkursie BirdCLEF 2016, opisanym w rozdziale 6. Choć modele wypracowane przez autora nie uplasowały się w ścisłej czołówce rankingu konkursowego, to studium to można potraktować jako dodatkową weryfikację zebranej wiedzy i potwierdzenie tezy że w pełni konwolucyjne głębokie sieci neuronowe (modele B i C w Tabeli 6.1) umożliwiają skuteczne rozwiązywanie tego zadania, oraz że tworzenie komitetów klasyfikatorów jest pomocne także w tej klasie zastosowań.

Część empiryczną pracy zamyka rozdział 7, opisujący eksperyment porównawczy, którego celem było wypracowanie rekomendacji odnośnie pożądanych właściwości architektur i hiperparametrów architektur neuronowych. Punktem wyjścia jest model E2018 używany w rozdziale 5; z którym porównywane są modele z filtrami kwadratowymi, połączeniami rezydualnymi, oraz architektury intensywnie wykorzystujące sploty 1x1 (tzw. squeeze nets, s. 158). Analiza przeprowadzona została dla 3 zbiorów danych, i dotyczyła aż 12 aspektów wymienionych w sekcji 7.1.3. Należy nadmienić że eksperyment ten jest bardzo obszerny i wymagał zapewne znacznych zasobów obliczeniowych. Większość wyników potwierdza zasadność ustawień standardowych (np. przydatność normalizacji partiami), choć zdarzają się też wyjątki (np. systematyczna przewaga LReLU na ReLU). Interesujące jest także to że wszystkie te eksperymenty porównawcze potwierdziły przydatność architektur dwukanałowych (tj. z drugim kanałem będącym kanałem różnicowym). Studium to jest o tyle wartościowe, że tego typu masywne eksperymenty porównawcze prezentowane są stosunkowo rzadko w innych pracach (m.in. z racji znacznych nakładów obliczeniowych wymaganych do ich przeprowadzenia). Pewne obserwacje dokonane przez Autora w efekcie tych analiz mogą stanowić cenne wskazówki dla autorów analogicznych prac.

Pracę zamyka podsumowujący rozdział 8, w którym Autor wymienia też kilka interesujących obszarów kontynuacji badań (np. analiza przyczyn przydatności kanału różnicowego). Zwraca też uwagę sumienna analiza dyskusyjnych (zdaniem Autora) elementów pracy.

2.2 Ocena redakcji pracy

Praca zredagowana jest nadzwyczaj starannie – nie przypominam sobie abym kiedykolwiek czytał bardziej sumiennie zredagowany doktorat. Nie dopatrzyłem się zasadniczo żadnych błędów językowych czy literówek. Uwagę zwracają liczne, przemyślane i bardzo estetyczne ilustracje. Tekst podany jest przystępnie i ma logiczną strukturę. Formalizmy

stosowane są z umiarem i skutecznie wspomagają zrozumienie tekstu. Autor umiejętnie stosuje odnośniki do poszczególnych części pracy, zarówno wsteczne i wyprzedzające.

2.3 Ocena wkładu oryginalnego

Za oryginalne elementy pracy uważam przede wszystkim:

- Zebrane w pracy pionierskie (na moment ich publikacji w powiązanych artykułach) architektury splotowych sieci neuronowych w zastosowaniu do klasyfikacji dźwięków środowiskowych.
- Systematyczne przebadanie wielu nietrywialnych architektur splotowych i ich czułości na poszczególne hiperparametry, z wykorzystaniem kilku zbiorów danych o zróżnicowanej charakterystyce.
- Przygotowanie nowego zbioru danych ESC, który znalazł zastosowanie nie tylko w recenzowanej rozprawie, ale także w przynajmniej kilkudziesięciu pracach innych autorów.
- Próby interpretacji charakterystyki nauczonych modeli. sieci.
- Zastosowanie proponowanych architektur do wymagającego zadania klasyfikacji dźwięków ptaków.
- Zebranie i przystępne całościowe przedstawienie aktualnej wiedzy o algorytmach klasyfikacji dźwięków środowiskowych metodami uczenia maszynowego.

2.4 Uwagi polemiczne

Pewne wątpliwości budzi we mnie główna hipoteza pracy, tj. „Czy splotowe sieci neuronowe mogą być z powodzeniem zastosowane w zadaniach klasyfikacji dźwięków niebędących mową?". Mowa stanowi specyficzną, raczej wąską klasę przebiegów dźwiękowych, m.in. dlatego że tor wokalny człowieka nie ma ograniczone możliwości i nie nadaje się do generowania wielu typów dźwięków. Można zatem argumentować że rozważenie szerszej (jak rozumiem w domyśle) klasy dźwięków powinno być zatem w ogólności zadaniem łatwiejszym.

Nasuwa się pytanie dlaczego w pracy nie wykorzystano architektur rekurencyjnych. Autor wspomina o nich na s. 144. Potencjalnie pozwoliłyby one skuteczniej odrywać i modelować zależności czasowe w badanych nagraniach.

Uważam że można było całą pracę opatrzyć bardziej informatywnym tytułem – jego obecna wersja może zniechęcać potencjalnych czytelników swoją ogólnością.

Z drobniejszych uwag, podział na 'zbiór danych' $\mathbf{X}$ i 'etykiety $\mathbf{y}$' w formule 2.39 jest zbyteczny i niepotrzebnie zawężający, bo przecież regularyzację da się stosować niezależnie od tego czy rozważanym zadaniem uczenia jest klasyfikacja (co sugeruje termin 'etykiety'), a nawet szerzej - niezależnie od tego czy zmienne zostały jawnie podzielone

na wejściowe (niezależne) i wyjściowe (zależne), jak ma to np. miejsce w zadaniach autoasocjacji. Zatrzymywanie procesu uczenia po z góry ustalonej licznie iteracji trudno jest podciągać pod 'wcześniejsze zatrzymanie procesu uczenia' (s. 53). Nie jest dla mnie jasne dlaczego w omawiając architekturę GAN, Autor nawiązuje do rozkładu brzegowego (s. 72) - wg mojej wiedzy celem generatora w tych architekturach jest możliwie pełne 'symulowanie' rozkładu *łącznego* danych. Na s. 82 autor słusznie wskazuje na ułomność trafności klasyfikowania polegającą na jej ograniczonej zawartości informacyjnej, ale słabość ta dotyczy przecież praktycznie wszystkich skalarnych miar oceny wymienionych wcześniej. W formułach w sekcji 3.2.1 pojawia się symbol kropki na oznaczenie mnożenia, choć wcześniej nie był używany. Wykorzystanie klasyfikatora kNN do klasyfikowania zapisów przy użyciu 72 lub większej liczby atrybutów (sekcja 4.4.2) jest dyskusyjne z racji specyficznych właściwości metryki Euklidesowej w wysokowymiarowych przestrzeniach.

Odnośnie nomenklatury, termin 'funkcja straty' stosowany jest zamiennie z 'funkcja kosztu', co może być momentami mylące. Termin 'niezmienność' na str. 34 (i gdzie indziej) powinien być moim zdaniem zastąpiony terminem 'niezmienniczość'. Fraza 'wprowadzenie metod adaptacyjnych' na s. 38 jest niejasna w tamtejszym kontekście, bo praktycznie wszystko co dotyczy SSN angażuje adaptację (domyślam się że chodzi o metody adaptujące hiperparametry, np. prędkość uczenia). Termin 'tempo uczenia' jest nieco niestandardowy; wydaje się że 'prędkość uczenia' jest bardziej ugruntowany w Polsce. Można też się zastanawiać czy w rozwinięciu skrótu GAN warto stosować kalkę językową z angielskiego ('adwersarz') w miejsce wydaje się bardziej oczywistego słowa 'przeciwnik'. Dokładność klasyfikowania (s. 81) jest chyba jednak częściej nazywana w polskiej literaturze trafnością klasyfikowania.

W dorobku Doktoranta powiązanym z rozprawą przydałaby się choć jedna publikacja czasopismowa. Zakładam że próby opublikowania wybranych wyników w czasopismach została przez niego podjęta, i być może praca(e) są obecnie w recenzji. Z drugiej strony, przynajmniej 2 z publikacji wymienionych w sekcji 1.4 ukazały się i były prezentowane na wartościowych wydarzeniach organizowanych przez renomowane instytucje (IEEE i ACM), co pozwala mi sądzić że znalazły uznanie tych środowisk.

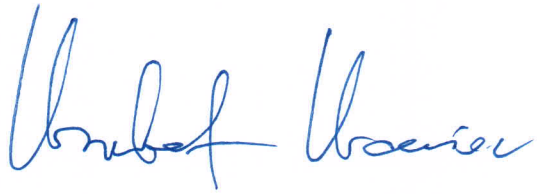
3 Konkluzja końcowa

Przedstawiona do oceny rozprawa doktorska mgr Piczaka zawiera oryginalne i wartościowe osiągnięcia, mocno podparte wynikami empirycznymi uzyskanymi na wymagających danych rzeczywistych. Wymienione powyżej uwagi polemiczne odnośnie treści i prezentacji pracy nie podważają głównych konkluzji rozprawy i mojej pozytywnej jej oceny. Uważam że cele postawione przez Autora pracy zostały osiągnięte.

Istotnym atutem wyników prezentowanych w pracy a wcześniej w publikowanych przez Autora artykułach jest ich pionierski charakter – prace te były, obok dwóch równolegle ukazujących się artykułów, pierwszymi w skali globalnej głębokimi splotowymi architekturami neuronowymi zaprojektowanymi z myślą o klasyfikacji dźwięków środowiskowych.

Wobec powyższego stwierdzam, że **rozprawa doktorska mgr Karola Jerzego Pi-**

czaka spełnia z nawiązką warunki stawiane przez ustawę o tytule naukowym i stopniach naukowych w odniesieniu do rozpraw doktorskich, a zatem powinna być dopuszczona do publicznej obrony, o co wnoszę do Rady Wydziału Elektroniki i Technik Informacyjnych Politechniki Warszawskiej. Ponadto, z racji pionierskiego charakteru publikowanych prac, dogłębnego charakteru przeprowadzonych w pracy badań, oraz jej szczególnie starannej redakcji, wnioskuję o jej wyróżnienie.



Prof. dr hab. inż. Bożena Kostek, prof. zw. PG
Politechnika Gdańska, Wydział Elektroniki,
Telekomunikacji i Informatyki
Lab. Akustyki Fonicznej

5. 03. 2019 r.

Opinia nt. rozprawy doktorskiej mgra Karola J. Piczaka pt.: „Klasyfikacja dźwięku za pomocą splotowych sieci neuronowych".

Cel, zawartość rozprawy, tezy, metodyka rozwiązywania problemów

Przedmiotem recenzji jest rozprawa doktorska mgra Karola J. Piczaka przygotowana pod kierunkiem prof. Jarosława Arabasa. Recenzowana rozprawa doktorska ma charakter eksperymentalno-implementacyjny, obejmuje 211 stron tekstu i składa się z ośmiu rozdziałów, Bibliografii, wykazu stosowanych oznaczeń oraz dodatku A pt.: „Zestawienie błędów standardowych dla analizy wrażliwości.

Problematyka i główny cel rozprawy doktorskiej dotyczą rozumienia maszynowego otoczenia dźwiękowego w postaci sygnału fonicznego, a w szczególności zastosowania splotowych sieci neuronowych w zadaniach klasyfikacji dźwięków, które nie są mową. Należy zwrócić uwagę, że dopiero w ostatnich latach wysiłek badawczy został skierowany na tę tematykę i dopiero obecnie zaczyna przynosić rezultaty, które można uznać za możliwe w zastosowaniach praktycznych. **W tym miejscu chciałabym dodać, że zarówno problematykę, jak i cel rozprawy należy uznać za aktualne. Problematyka automatycznej klasyfikacji dźwięków środowiskowych ma duże znaczenie** również w innych obszarach, np. monitoring przestrzeni publicznych w kontekście zagrożeń, monitoring hałasu poprzez analizę dźwięków środowiskowych (w tym rekreacyjnych), systemy zarządzania ruchem drogowym (informacja dźwiękowa o zdarzeniach), itd. Obszary te przenikają się ze względu na stosowane metody analiz oraz metody uczenia maszynowego, dlatego postęp w jednej dziedzinie przyspiesza rozwój innych obszarów.

Zawartość rozprawy:

We Wprowadzeniu, w podrozdziale zatytułowanym „Motywacja" doktorant przedstawia krótki, ogólny przegląd dotyczący rozwoju badań w dziedzinie przetwarzania sygnałów oraz treści multimedialnych, a także odnosi się do głównych motorów innowacji (m.in. możliwość tworzenia modeli o zwiększonej złożoności obliczeniowej ze względu na dostępne obecnie zasoby sprzętowe), jak i kamieni milowych (modele wykorzystujące sieci splotowe, które przyczyniły się do powstania wielu praktycznych rozwiązań z dokładnością przewyższającą zdolności percepcyjne człowieka). W końcowej części tego podrozdziału doktorant zawarł krótką dyskusję,

która stanowi jednocześnie motywację prowadzenia badań w obszarze rozumienia maszynowego otoczenia dźwiękowego.

W kolejnym podrozdziale doktorant przedstawia w sposób szczegółowy cel rozprawy w formie pytania, a mianowicie: „czy splotowe sieci neuronowe mogą być z powodzeniem zastosowane w zadaniach klasyfikacji dźwięków niebędących mową, w szczególności mając na uwadze ograniczoność etykietowanych zbiorów danych w tej dziedzinie", a także wskazuje kilku wyznaczników badań:

- zaprezentowanie działania systemów opartych na metodzie sieci neuronowych w formie studium przypadku dla przykładowych zadań klasyfikacji dźwięków środowiskowych i śpiewu ptaków,

- przeanalizowanie wpływu zmian w architekturze i procesie uczenia sieci na skuteczność klasyfikacji;

- udostępnienie etykietowanego zbioru nagrań środowiskowych adekwatnych dla zadań klasyfikacji wraz z poziomem odniesienia w postaci oceny skuteczności rozpoznawania osiąganej przez ludzi,

- pokazanie możliwości głębszego rozumienia działania splotowych sieci neuronowych w zastosowaniach dźwiękowych poprzez wizualizację efektów uczenia sieci bazujących na przetwarzaniu spektrogramów.

Pewien niedosyt budzi brak podania w tej części Wprowadzenia tez/hipotez rozprawy, które doktorant następnie by weryfikował poprzez zastosowaną metodologię i eksperymenty. Dopiero w Podsumowaniu (rozdział 8) pojawia się odniesienie do hipotezy badawczej, a mianowicie, że do celu klasyfikacji dźwięków występujących w środowisku naturalnym (niebędących muzyką i mową) można zaadaptować splotowe sieci neuronowe - modele, które sprawdziły się z zastosowaniu do rozpoznawania obrazów (pomimo ograniczonych rozmiarów zbiorów uczących dostępnych w tej dziedzinie). Ta hipoteza obecnie nie nosi już znamion nowości, o czym autor rozprawy wspomina m.in. we Wprowadzeniu czy Podsumowaniu. Sądzę, że łatwo można by - w oparciu o przedstawioną do recenzji rozprawę - sformułować tezy, które odnosiłyby się w sposób bezpośredni do przeprowadzonych eksperymentów i które odzwierciedlają oryginalny wkład doktoranta w obszar badawczy, np.:

1. Możliwe jest przy pomocy sieci splotowych uzyskanie podobnej lub większej skuteczności klasyfikacji dźwięków środowiskowych (statystycznie istotne różnice w przypadku większej skuteczności) jak w przypadku typowych algorytmów uczących się.
2. Wykorzystanie kanału "delta" w postaci spektrogramu różnicowego po czasie pozwala na uzyskanie większej dokładności klasyfikacji.

itd. ...

Ostatnie podrozdziały we Wprowadzeniu odnoszą się do przedstawienia układu rozprawy oraz publikacji, na bazie których powstały rozdziały autorskie 4-7.

Rozdział 2 zawiera przegląd zagadnień związanych z sieciami neuronowymi i głębokim uczeniem z uwzględnieniem sieci splotowych. Obok założeń uczenia modeli głębokich, główny nacisk położony został na możliwe architektury i koncepcje tworzenia sztucznych sieci neuronowych. Doktorant przedstawił te zagadnienia w formie spojrzenia poprzez pryzmat rozwoju architektur splotowych.

W rozdziale 3. doktorant omawia problematykę klasyfikacji dźwięku, metody reprezentacji sygnałów fonicznych oraz stosowane w literaturze podejścia rozwiązywania problemów związanych z klasyfikacją dźwięku. Przywołane deskryptory są przykładami parametrów stosowanych w rozpoznawaniu dźwięków czy sygnałów fonicznych.

Kolejne rozdziały (4-7) stanowią oryginalny wkład doktoranta w badania dotyczące klasyfikacji dźwięków środowiskowych. W pierwszej kolejności (rozdział 4) autor rozprawy przedstawia prace związane ze stworzeniem zbioru nagrań środowiskowych (ESC), które zostały udostępnione na zasadzie licencji niekomercyjnej, co w rezultacie pozwoliło na odniesienie się do wyników uzyskanych przez innych badaczy na tym samym zbiorze w postaci zestawienia modeli poddawanych walidacji na zbiorze ESC. Ważny jest też wątek badawczy, w którym autor przytacza wyniki eksperymentu *crowdsourcingowego*, w którym słuchacze mieli za zadanie przypisanie właściwej klasy do odsłuchiwanego dźwięku. Pozwoliło to na odniesienie wyników uzyskanych w automatycznej klasyfikacji do ludzkich zdolności w zakresie rozpoznawania dźwięków środowiskowych.

W rozdziale 5. doktorant zawarł wyniki prowadzonych eksperymentów z zastosowaniem splotowych sieci do klasyfikacji dźwięków środowiskowych. Bardzo cennym elementem tego rozdziału jest przedstawienie wizualizacji efektów uczenia splotowych sieci neuronowych przetwarzających spektrogramy. Wizualizacja tego typu pozwala na dogłębną ocenę zasady działania splotowych sieci neuronowych w zastosowaniach klasyfikacji dźwięków. Można zauważyć, że w ten sposób uzyskuje się częściowo możliwość interpretacji wyników i wpływu np. zwiększenia siły regularyzacji czy kanału różnicowego, co pozwala na stwierdzenie, że sieć neuronowa przestaje być w pełni zamkniętą „czarną skrzynką".

Rozdział 6 odnosi się do zagadnienia rozpoznawania gatunków ptaków śpiewających z wykorzystaniem splotowych sieci na przykładzie zadania konkursowego *Bird*CLEF 2016. Ostatni z rozdziałów eksperymentalnych zawiera pogłębioną analizę zachowania zastosowanych modeli splotowych sieci neuronowych w zadaniach klasyfikacji sygnałów fonicznych. W szczególności doktorant zestawia czynniki mające wpływ na dokładność klasyfikacji dla czterech typów architektur splotowych. Ze względu na fakt, że baza *Bird*CLEF 2016 jest wymagająca w kontekście automatycznego rozpoznawania dźwięku, dlatego uzyskane wyniki (6, 8, 9

i 10 miejsce w konkursie) mogą służyć jako punkt wyjścia do uzyskania bardziej obiecujących wyników w przyszłości.

Rozdział 7 stanowi szczegółową analizę modeli splotowych (i ich architektur) w odniesieniu do zbiorów danych wykorzystanych w klasyfikacji dźwięku. Szczególnie interesujące jest przedstawienie wyników analizy wrażliwości dla poszczególnych czynników procesu klasyfikacji, tj. kształtu filtrów splotowych w modelach, szerokości modelu, głębokości modelu, prawdopodobieństwa *dropoutu*, wykorzystanej normalizacji, zastosowanej funkcji aktywacji neuronów, ustawienia procesu uczenia, metody inicjalizacji wag, rozmiaru partii uczącej, itd. W podsumowaniu tego rozdziału doktorant formułuje wnioski, które stanowią podsumowanie poszczególnych elementów tej pogłębionej analizy.

Rozdziały 4-7 stanowią rozszerzenie bądź bazują na publikacjach przygotowanych i opublikowanych przez autora rozprawy, a także na upublicznionej bazie dźwięków środowiskowych. W rezultacie przyniosło to możliwość odniesienia się do wyników, które uzyskali inni badacze, co jest bardzo cennym osiągnięciem autora rozprawy.

Ostatnim rozdziałem jest Podsumowanie, w którym doktorant podaje główne rezultaty rozprawy odnoszące się do głównych osiągnięć autora, wnioski szczegółowe wynikające z przeprowadzonych badań i analiz oraz dyskusję - w tym komentarz odnoszący się do słabszych elementów rozprawy i omówienia otwartych problemów i kierunków możliwych dalszych badań. Przytoczony w Podsumowaniu: „Komentarz odnośnie słabych elementów rozprawy" stanowi raczej dojrzałą dyskusję uzyskanych wyników i rozwoju planów badawczych, a nie opis „słabych elementów" rozprawy.

Ocena zawartości rozprawy i metod rozwiązywania postawionych problemów

W pierwszej kolejności należy podkreślić bardzo dużą dojrzałość doktoranta w umiejętności syntetycznego przedstawienia zagadnień związanych z automatyczną analizą dźwięków środowiskowych, przeglądu literatury w tym zakresie oraz wyników oryginalnych prac. Uznanie budzi również warsztat prowadzenia eksperymentu badawczego przez doktoranta, ale chyba przede wszystkim umiejętność analizy (w tym wizualizacji etapów wyników, bowiem ten etap posłużył do pogłębionych analiz), jak również syntetycznego przedstawienia uzyskanych wyników i przedstawienia ich na tle badań innych autorów. Ważnym aspektem było przygotowanie przez autora zbioru nagrań dźwięków środowiskowych, które udostępnił na licencji *Creative Commons*, co dało możliwość bezpośredniego porównania wyników własnych i innych autorów. Należy też zwrócić uwagę na charakter implementacyjny przeprowadzonych badań. Rozprawa ma wyraźnie charakter praktyczny i może stanowić *benchmark* w badaniach dźwięków środowiskowych.

Doktorant rozwiązuje dwa problemy, tj. klasyfikuje za pomocą sieci splotowych dźwięki środowiskowe oraz gatunki ptaków, wykorzystując do tego bazy nagrań (w tym bazy własne). W rozwiązywaniu tych problemów, a także w celach porównawczych doktorant wykorzystał zarówno powszechnie stosowane algorytmy uczące się (k-NN (k-Najbliższych Sąsiadów), SVM (*Support Vector Machine*), RF (*Random Forest*)), sieci splotowe, jak również eksperyment *crowdsourcingowy,* w którym uczestnicy testów przypisywali etykiety słyszanym dźwiękom.

Doktorant wykazał szereg autorskich rozwiązań, które jednocześnie stanowią główne osiągnięcia rozprawy oraz wkład w rozwój dziedziny. Przywołam je za autorem poniżej:

- sporządzenie i upublicznienie etykietowanego zbioru 2000 nagrań środowiskowych "ESC-50" wraz z wynikami klasyfikacji za pomocą podstawowych metod automatycznych i osiąganych przez uczestników eksperymentu *crowdsourcingowego*,
- przeprowadzenie szczegółowej analizy wpływu zmian wartości hiperparametrów na dokładność klasyfikacji dźwięku dla modeli będących przedstawicielami czterech typów architektur splotowych,
- wykorzystanie reprezentacji różnicowej podawanej na wejście sieci splotowej zwiększyło dokładność klasyfikacji,
- zastosowanie wertykalnych filtrów splotowych w pierwszej warstwie, motywowane osiągnięciami obszaru rozpoznawania mowy, co zapewniło uzyskanie w przeprowadzonych eksperymentach wyników porównywalnych, a nawet lepszych niż w przypadku małych filtrów kwadratowych. Zaletami rozwiązania zaprezentowanego w rozprawie jest większa wydajność uczenia i generowania predykcji dla tego typu modeli, skalowalność przy zwiększaniu rozdzielczości częstotliwościowej spektrogramu, a także bogatsza wartość informacyjna w przypadku pobieżnej wizualizacji wag pierwszej warstwy,
- wprowadzenie wizualizacji efektu uczenia jako jednego z etapów analizy i interpretacji wyników
- zaprezentowanie jednego z pierwszych, opublikowanych zastosowań splotowych sieci neuronowych do klasyfikacji dźwięków środowiskowych,
- zaprezentowanie jednego z pierwszych opublikowanych zastosowań splotowych sieci neuronowych do rozpoznawania gatunków ptaków śpiewających,

Rozdziały autorskie oraz dyskusja zawarta w rozdziale końcowym pokazały, że doktorant osiągnął wyniki pozwalające na stwierdzenie, że cel rozprawy został w pełni osiągnięty. Ponadto, uzyskane wyniki świadczą o jakości

prowadzonych eksperymentów, w szczególności rozpoznawania dźwięków środowiskowych.

Poniżej w Uwagach zawarto kilka bardziej szczegółowych punktów dotyczących oceny rozprawy i uwag do ewentualnej dyskusji.

Uwagi ogólne:

Czy autor rozprawy mógłby udzielić odpowiedzi na poniższe pytania i sugestie?

1. Jak wspomniano wcześniej pewien niedosyt budzi brak podania hipotez badawczych we Wprowadzeniu do rozprawy. Sądzę, że w prezentacji rozprawy na obronie doktoratu (oraz w odpowiedzi na recenzję) warto by podać takie hipotezy.

2. Doktorant odniósł się do złożoności obliczeniowej omawianych problemów, zarysowuje też aspekt skalowalności problemu, ale warto by podać też odniesienie do całościowego czasu przetwarzania i klasyfikacji (w kontekście rozwiązań sprzętowo-programowych) dźwięków środowiskowych i rozpoznawania gatunków ptaków na podstawie dźwiękowych reprezentacji. Informacje te można znaleźć w rozprawie, ale są one podawane cząstkowo dla poszczególnych zadań.

3. W przeglądzie literatury zabrakło odniesienia do publikacji, które odnoszą się do szerzej rozumianego *środowiska*, jakim jest monitorowanie przestrzeni publicznej w kontekście niebezpiecznych dźwięków. Poniżej podane zostały przykłady takich publikacji, autorzy niektórych prac również korzystali z przywołanych przez doktoranta baz dźwiękowych. Warto zauważyć, że w pracach tych w szerszym zakresie analizowane są parametry MPEG 7 (również parametry autorskie) w kontekście automatycznej analizy środowiska niż w rozprawie mgra Piczaka.

4. Ponieważ uzyskiwane wyniki z wykorzystaniem typowych algorytmów uczących się i wektora cech są porównywalne z wynikami, które otrzymuje się w oparciu o sieci splotowe, może warto byłoby „dostroić" tę pierwszą metodologię (o ile nie pojawia się problem skalowalności), wykorzystując w większym zakresie przetwarzanie sygnałów i bogatsza parametryzację?

5. Zapewne ze względu na spójność pracy nie zostały szerzej przywołane również prace, które odnoszą się do aktualnych analiz sygnału mowy (w różnych kontekstach) z wykorzystaniem uczenia głębokiego przy **zastosowaniu różnych reprezentacji dwuwymiarowych** sygnału mowy. Sądzę, że warto taki przegląd wykonać przy okazji tworzenia programu dalszych badań.

6. Edycja i wydanie rozprawy zajęły zapewne autorowi kilka miesięcy i w związku z tym nie było możliwe odniesienie się w niej do aktualnych wyników dotyczących - w szczególności - klasyfikacji gatunków ptaków. Czy doktorant mógłby w prezentacji odnieść się do aktualnych wyników badań (zdaję sobie sprawę, że mogą się one odnosić do nowszej wersji bazy *bird*CLEF2017/2018). Autor rozprawy

odnosi się tylko bardzo ogólnie do tych nowszych prac. Przy czym szczególnie istotne byłoby podanie wykorzystanych reprezentacji sygnałowych (w tym metody wstępnego przetwarzania) oraz charakterystyk algorytmów uczących się. Pytanie to może odpowiedzieć na pytanie czy niska efektywność klasyfikacji uzyskiwana dla bazy *bird*CLEF2016 wynika z właściwości samej bazy czy stosowanej metodologii.

Przykłady prac w kontekście monitorowania przestrzeni publicznej:

Ntalampiras S, Potamitis I, Fakotakis N (2011) Probabilistic novelty detection for acoustic surveillance under real-world conditions. IEEE Trans Multimed 13(4):713–719.
M. Pleva, E. Vozarikova, L. Dobos, and A. Cizmar, The Joint Database of Audio Events and Backgrounds for Monitoring of Urban Areas, J. Electrical and Electronics Eng'g, vol. 4, pp. 185–188, 2011 May.
Ntalampiras, Stavros; Potamitis, Ilyas; Fakotakis, Nikos, Acoustic Detection of Human Activities in Natural Environments, J. Audio Eng. Soc., vol. 60, No. 9, 2012.
Kotus J., Łopatka K., Czyżewski A., Detection and localization of selected acoustic events in acoustic field for smart surveillance applications; Multimedia Tools and Applications, 68: 5. https://doi.org/10.1007/s11042-012-1183-02014.
Ntalampiras, Stavros, Audio Pattern Recognition of Baby Crying Sound Events, J. Audio Eng. Soc., Vol. 63, no. 5, pp. 358-369, May 2015.
Bitzer, Joerg; Kissner, Sven; Holube, Inga, Privacy-Aware Acoustic Assessments of Everyday Life, J. Audio Eng. Soc., vol. 64, no. 6, pp. 395-404; June 2016.
Li, Weihong; Zhao, Bingxin; Peng, Shuyong; Gong, Weiguo, Improved Local Mean Decomposition Based on the T-distribution for Feature Extraction of Abnormal Sounds in Public Places, J. Audio Eng. Soc., vol. 65, no. 10, pp. 806-816; October 2017.
Hrabina, Martin; Sigmund, Milan, Audio Event Database Collected for Gunshot Detection in Open Nature (GUDEON), J. Audio Eng. Soc., vol. 67, no. ½, pp. 54-59; January 2019. DOI: https://doi.org/10.17743/jaes.2018.0075
A. Mitilineos, N.-A. Tatlas, S. M. Potirakis and M. Rangoussi, Neural Network Fusion for Noise-Efficient Sound Classification, J. Audio Eng. Soc., vol. 67, no. 1/2, pp. 27–37, (2019 January/February). DOI: https://doi.org/10.17743/jaes.2018.0071

Uwagi redakcyjne

Rozprawa doktorska jest przygotowana **bardzo starannie od strony edycyjnej**, język rozprawy jest również poprawny, chociaż w tekście pojawiają się drobne usterki (wyrażenia kolokwialne czy drobne potknięcia interpunkcyjne), jednak są one na tyle nieliczne, że nie widać potrzeby ich przytaczania. Świadczy to o jakości przygotowania i edycji rozprawy.

Podsumowanie

Tematyka rozprawy mgra K. J. Piczaka jest bardzo istotna w kontekście możliwości automatycznej klasyfikacji dźwięków środowiskowych, pozwalającej w przyszłości na automatyczne monitorowanie przestrzeni publicznych.

Do osiągnięć rozprawy zaliczam niewątpliwie bardzo systematyczne i rzetelne przeprowadzenie analiz i ich zestawienie z wynikami prac innych autorów. Na uwagę zasługuje umiejętność bardzo jasnego przedstawiania analizowanego problemu,

(zarówno w formie opisu, jak i wizualizacji problemu), właściwego wyciągnięcia wniosków oraz ogólnego podsumowania uzyskanych wyników. Należy też docenić zarysowanie kierunków rozwoju prowadzonych prac, co - ze względu na zdobyte - doświadczenie badawcze może stanowić w przyszłości ukoronowanie osiągniętych przez doktoranta wyników. W szczególności ważny jest aspekt skalowalności omawianych zagadnień, o którym wspomina autor rozprawy, jak również zastosowanie metod, które pozwolą na lepszą separację sygnałów użytecznych (klasyfikowanych) od szumów i tła akustycznego.

Przygotowanie i upublicznienie etykietowanego zbioru 2000 nagrań środowiskowych "ESC-50" stanowi z wraz z wynikami klasyfikacji za pomocą podstawowych metod automatycznych i osiąganych przez uczestników eksperymentu *crowdsourcingowego* niewątpliwie jedno z większych osiągnięć rozprawy. Natomiast **za najbardziej oryginalny wątek badawczy rozprawy uważam wprowadzenie dla danych wejściowych dodatkowego kanału "delta" w postaci spektrogramu różnicowego analizowanego po czasie** w eksperymentach. Reasumując, mogę stwierdzić, że autor rozprawy wykazał należyty ogólny poziom wiedzy w zakresie będącym przedmiotem rozprawy oraz biegłość systematykę w prowadzeniu eksperymentów badawczych. Doktorant osiągnął postawiony cel rozprawy.

Z kolei do osobistych i ważnych osiągnięć doktoranta chciałabym zaliczyć cytowania trzech referatów konferencyjnych (**w sumie 166 cytowań w bazie SCOPUS**; indeks H: 3). W szczególności doktorant uzyskał liczne cytowania w odniesieniu do dwóch prac: Piczak, K.J., *Environmental sound classification with convolutional neural networks*, IEEE International Workshop on Machine Learning for Signal Processing, MLSP, 2015-November,7324337, 2015 (**100 cytowań**) oraz praca: Piczak, K.J., *ESC: Dataset for environmental sound classification*, MM 2015 - Proceedings of the 2015 ACM Multimedia Conference, pp. 1015-1018 (**62 cytowania;** warto zauważyć, że konferencja ACM Multimedia jest notowana w bazie **CORE** z kategorią **A***).

W podsumowaniu stwierdzam, że przedłożona mi do recenzji rozprawa p. mgra Karola Jerzego Piczaka spełnia wymagania stawiane w Ustawie rozprawom doktorskim i wnoszę o jej dopuszczenie do publicznej obrony.

Ze względu na walory implementacyjne rozprawy, opublikowanie wyników rozprawy na czołowych konferencjach stanowiących *benchmark* w obszarze klasyfikacji dźwięków środowiskowych oraz uzyskanie licznych cytowań do tych prac wnoszę o wyróżnienie rozprawy doktorskiej.