Karol J. Piczak
Politechnika Warszawska
Instytut Informatyki
ul. Nowowiejska 15/19
00-665 Warszawa
karol.piczak@pw.edu.pl

KKRRi 2020
Łódź, 17-18 września 2020

ANALIZA MODELI SPLOTOWYCH W ZADANIACH KLASYFIKACJI DŹWIĘKÓW ŚRODOWISKOWYCH

ANALYSIS OF CONVOLUTIONAL MODELS IN ENVIRONMENTAL SOUND CLASSIFICATION TASKS

Streszczenie: **Splotowe sieci neuronowe są obecnie popularnym narzędziem wykorzystywanym w rozpoznawaniu dźwięków środowiskowych. Na skuteczność ich działania wpływa wiele potencjalnych czynników. Niniejszy referat przedstawia podsumowanie wyników uzyskanych w rozprawie doktorskiej autora w zakresie analizy wrażliwości modeli splotowych na dobierane wartości hiperparametrów. W szczególności zastosowanie techniki *dropout* okazuje się mieć znaczący wpływ na funkcjonowanie tego typu modeli.**

Abstract: **Convolutional neural networks are a popular tool used in environmental sound recognition tasks. Their performance depends on multiple factors. This paper presents a summarized extract from author's PhD dissertation on analyzing the sensitivity of convolutional models to hyperparameter values. In particular, *dropout* happens to play an important role in these kinds of models.**

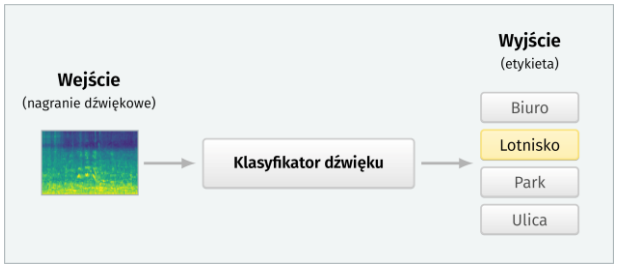
Słowa kluczowe: **klasyfikacja dźwięków środowiskowych, spektrogram, splotowe sieci neuronowe**

Keywords: **convolutional neural networks, environmental sound classification, spectrogram**

1. WSTĘP

Rozumienie dźwięków środowiskowych stało się w ostatnich latach coraz bardziej docenianym problemem badawczym. Widać to najwyraźniej w dynamicznie rosnącej liczbie publikacji tematycznych i zgłoszeń, w tym pochodzących od podmiotów komercyjnych, do cyklicznych konkursów związanych z detekcją i klasyfikacją scen i zdarzeń akustycznych[1]. Oprócz zastosowań bezpośrednich, związanych z monitoringiem, indeksacją treści multimedialnych czy diagnostyką akustyczną, efekty dobrego rozumienia kontekstu dźwiękowego na podstawie nagrania wspomagają również pośrednio takie kwestie, jak np. dobieranie odpowiedniego profilu działania aparatów słuchowych [1].

Obecnie najczęściej spotykanym podejściem w obszarze rozumienia dźwięków środowiskowych jest wykorzystanie różnych wariantów głębokich sieci neuronowych zawierających warstwy splotowe. Koncepcja ta została zaproponowana w zadaniu klasyfikacji dźwięków, przedstawionym schematycznie na rysunku 1., po raz pierwszy w roku 2015, równolegle w pracach autora [3], Zhanga et al. [8] oraz Espiego et. al. [2]. W kolejnych latach publikowane były różne propozycje usprawnień tych koncepcji.



Rys. 1. Schemat działania klasyfikatora dźwiękowego, przypisującego pojedynczą etykietę do nagrania.

Charakterystyczną cechą wykorzystania modeli opierających się na głębokich sieciach neuronowych jest bardzo duża liczba decyzji, które projektant takiego rozwiązania musi podjąć przy wyborze architektury modelu i wartości hiperparametrów. Względnie długi czas uczenia głębokich sieci neuronowych powoduje, że szersze porównania w tym zakresie są jednak dosyć rzadko spotykane w literaturze.

Celem niniejszego referatu jest zaprezentowanie najważniejszych wniosków z rozdziału rozprawy doktorskiej autora [5] poświęconego analizie wrażliwości tego typu modeli na ustalenia podejmowane przy doborze ich architektury i wartości hiperparametrów. Zgodnie z wiedzą autora jest to najszersze zestawienie tego typu przeprowadzone dotychczas w obszarze rozumienia dźwięków środowiskowych i z tego względu w ramach referatu przedstawione zostaną tylko wybrane podsumowania wyników przeprowadzonych eksperymentów.

2. METODYKA EKSPERYMENTÓW

2.1. Rozważane architektury splotowe

Pierwszym przyjętym założeniem w eksperymentach było ograniczenie analizy do czterech przykładów architektur splotowych będących reprezentantami pewnych szerszych rodzin modeli.

[1] DCASE (*Detection and Classification of Acoustic Scenes and Events*) Challenge: http://dcase.community/

Architektura bazowa (B) to sieć splotowa o 7 warstwach wykorzystująca w pierwszej warstwie filtry wertykalne, które ograniczają się do przetwarzania krótkich wycinków czasowych spektrogramu, natomiast poddają analizie jednocześnie cały zakres częstotliwości.

Architektura z filtrami kwadratowymi (K) wykorzystuje filtry kwadratowe o rozmiarze 3 × 3 z dodatkowym wypełnieniem (*padding*) pozwalającym zachować niezmienione rozmiary wynikowych map aktywacji. Sieć taka jest dosyć typowym przykładem rozwiązań stosowanych w obszarze klasyfikowania obrazów (tzw. *2D CNN*).

Architektura z połączeniami rezydualnymi (R) jest minimalną modyfikacją modelu *ResNet-18* popularnego w rozpoznawaniu obrazów. Bazuje na 17 warstwach splotowych zgrupowanych w bloki uczące się reprezentacji rezydualnej.

Architektura typu SqueezeNet (S) wzorująca się na modelu pozwalającym utrzymać poziom dokładności sieci *AlexNet* klasyfikującej zbiór *ImageNet* przy 50-krotnej redukcji liczby parametrów.

2.2. Zbiory danych

Do oceny dokładności klasyfikacji uzyskiwanej przez poszczególne warianty modeli splotowych wykorzystane zostały trzy zbiory danych: *ESC-50* [4], *UrbanSound8K* [7] i część zbioru nagrań udostępnionych w ramach *BirdCLEF 2016*, dalej określana jako *Minibirds*. Zawierają one nagrania zdarzeń dźwiękowych występujących w środowisku naturalnym, miejskim oraz wokalizacje ptaków. *ESC-50* i *UrbanSound8K* stanowią uznane punkty odniesienia dla pomiaru dokładności modeli klasyfikacji dźwięków środowiskowych. Szerszy opis tych zbiorów danych zawierają prace wymienione w odesłaniach literaturowych.

2.3. Format danych wejściowych i domyślne wartości hiperparametrów

Nagrania przetwarzane były do spektrogramów generowanych według ujednoliconych ustawień: częstotliwość próbkowania – 44.1 kHz, długość okna FFT – 50 ms, przeskok – 20 ms, liczba pasm melowych – 60, limit górnej częstotliwości – 16 kHz. Długość segmentów uczących wynosiła 125 ramek (2,5 s) dla zbiorów *ESC-50* i *UrbanSound8K* oraz 500 ramek (10 s) dla przykładów z *Minibirds*.

Uczenie przeprowadzane było przez minimalizację entropii krzyżowej za pomocą metody spadku stochastycznego gradientu (tempo uczenia $\eta = 0{,}01$, momentum Nesterova o wartości 0,9, regularyzacja L2 o sile $\alpha = 0{,}001$) przy rozmiarze partii uczącej wynoszącym 128 przykładów. Inicjalizacja wag domyślnie dokonywana była według reguły *LeCun uniform*. Uczenie prowadzone było przez 20 epok odpowiadających w przybliżeniu 5000 partii uczących dla zbiorów *ESC-50* i *UrbanSound8K* oraz 50 epok równoważnych około 8500 krokom aktualizującym wagi dla *Minibirds.*

Miarą wykorzystywaną przy raportowaniu wyników jest dokładność uzyskiwana na zbiorze testowym uśredniona dla dziesięciu końcowych epok uczenia.

3. WYNIKI ANALIZY WRAŻLIWOŚCI

3.1. Wykorzystanie kanału różnicowego

We wszystkich eksperymentach modele uczone z użyciem dodatkowych spektrogramów różnicowych po czasie uzyskiwały wyniki lepsze od swoich odpowiedników opierających się tylko na spektrogramach bazowych.

3.2. Kształt filtrów splotowych

W przypadku filtrów wertykalnych architektury bazowej zmniejszenie wysokości zastosowanych filtrów poprawiało dokładność klasyfikacji. Pokazuje to, że ważnym elementem polepszania zdolności generalizacyjnych modelu jest jego uodpornienie na przesunięcia w dziedzinie częstotliwości. Efekt ten jest o tyle intrygujący, że nie były to tylko niewielkie przesunięcia. Oznacza to, że dla klasyfikowania dźwięków środowiskowych ważniejsze okazują się być konkretne wzorce w dziedzinie częstotliwości (struktury harmoniczne) niż ich dokładne umiejscowienie na osi częstotliwości.

Dla typowej architektury zaadaptowanej z obszaru przetwarzania obrazów (*2D CNN*), wykorzystującej na wejściu małe filtry kwadratowe (lub bliskie im prostokątne), kwestia doboru konkretnego rozmiaru tych filtrów okazała się być względnie mało istotna. Różnice wyników w większości przypadków nie wykraczały poza wahania z tytułu losowej inicjalizacji parametrów.

Regulowanie szerokości filtra splotowego nie było istotnym kryterium doboru modelu.

3.3. Szerokość modelu

Szczególnie w przypadku architektury typu *K* zwiększanie liczby filtrów pozwalało na zbiorze *ESC-50* na uzyskanie wyniku istotnie lepszego od pozostałych modeli. Proces ten był jednak bardzo kosztowny, jeśli chodzi o przyrost wymagań czasowych i pamięciowych uczenia.

3.4. Głębokość modelu

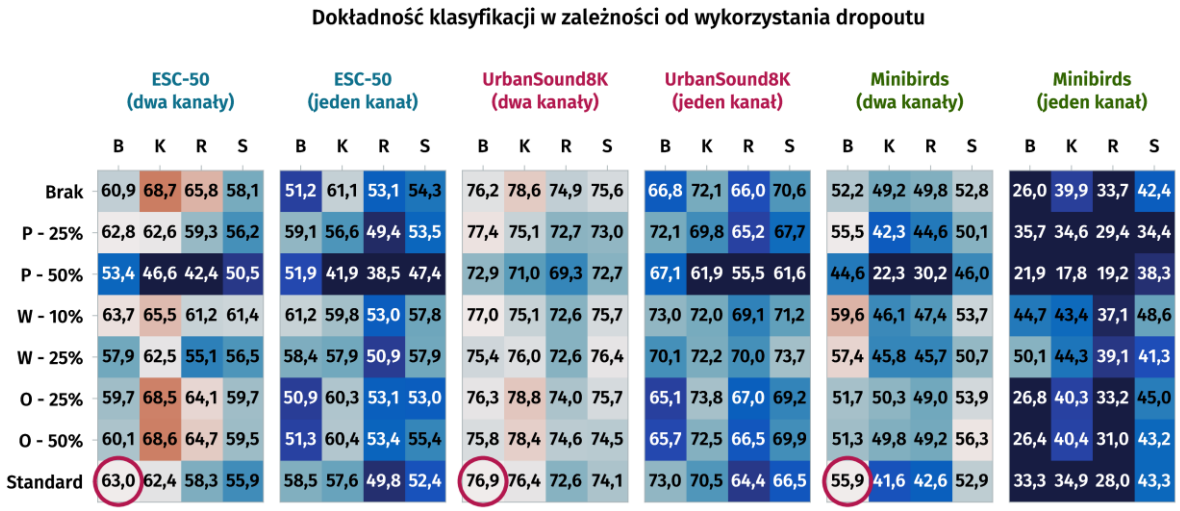
Dla *ESC-50* ogólną tendencją było uzyskiwanie lepszych wyników przez modele płytsze. Podobny kierunek można zauważyć dla *UrbanSound8K*, chociaż efekt ten był w jego przypadku nieco mniej wyraźny.

Osobliwie zachowywał się natomiast zbiór *Minibirds*. O ile liczba warstw splotowych nie odgrywała kluczowej roli w poprawie zdolności klasyfikacyjnych, to istotnym elementem okazało się być dołożenie dodatkowych warstw w pełni połączonych przed warstwą wyjścia typu *softmax*. Pokazuje to, że odmienny charakter nagrań tego zbioru domaga się dodatkowych możliwości przetwarzania, których nie zapewniają modele czysto splotowe.

3.5. Prawdopodobieństwo *dropoutu*

Analiza zestawienia na rysunku 2. pokazuje, że kwestia doboru odpowiednich ustawień *dropoutu* jest bardzo ważnym obszarem decyzyjnym, mogącym w drastyczny sposób wpłynąć na poziom uzyskiwanych wyników. Co więcej, najlepsze podejście w tym obszarze zależy w dużej mierze od typu architektury.

W przypadku architektury bazowej wariant domyślny (*dropout* 25% bezpośrednio po pierwszej warstwie splotowej i 50% przed warstwą wyjścia) okazał się



Rys. 2. Wyniki klasyfikacji dla rozważanych architektur (kolumny „B”, „K”, „R”, „S”) w zależności od ustawienia dropoutu. Wiersz „Standard” zakłada dropout z prawdopodobieństwem 25% po pierwszej warstwie i 50% przed warstwą wyjścia. Wiersz „Brak” przewiduje całkowitą rezygnację z dropoutu. W pozostałych wierszach oceniane były modele z dropoutem o określonym prawdopodobieństwie zastosowanym tylko po pierwszej warstwie (P), po wszystkich warstwach modelu (W) lub tylko przed ostatnią (O). Komórki macierzy przedstawiają dokładność klasyfikacji wyrażoną w procentach. Wyniki dla modelu bazowego omawianego w rozprawie zakreślone zostały kolorem ciemnoróżowym.

być skuteczny. Rezygnacja z *dropoutu* po pierwszej warstwie („O – 50%”) wiązała się z jednoznacznym pogorszeniem wyników, więc jest to element ważny dla modeli z filtrami wertykalnymi. W przypadku pozostałych typów architektur decyzja o wprowadzeniu *dropoutu* na tak wczesnym etapie okazała się mieć zgoła odwrotny efekt. Model oparty na małych filtrach kwadratowych po pozbawieniu *dropoutu* stawał się bezkonkurencyjny w klasyfikowaniu zbiorów *ESC-50* i *UrbanSound8K*.

Różnicę w zachowaniu między architekturami można interpretować dwuaspektowo. Po pierwsze, wykorzystanie filtrów wertykalnych o bardzo dużych rozmiarach (macierz kilkuset wag) może zwiększać skłonność sieci do skupiania się już w pierwszej warstwie na bardzo specyficznych wzorcach, co pogarsza jej zdolność generalizacji. Zastosowanie częściowego *dropoutu* na tym etapie pozwala zniwelować ten efekt. W przypadku pozostałych architektur opierających się na małych filtrach wejściowych (zwykle 3×3, co najwyżej 7 × 7) ryzyko takie jest dużo mniejsze. Z drugiej strony, zwłaszcza po uwzględnieniu lepszego funkcjonowania modeli z większą liczbą filtrów, można domniemywać, że dla tych architektur użycie tak wczesnego *dropoutu* zbyt mocno ogranicza ich efektywną szerokość.

Ciekawym zachowaniem cechują się modele bazowe (B) uczone na zbiorze Minibirds. W ich przypadku wyraźnie pojawiał się problem przeuczenia, któremu udawało się zaradzić przez wprowadzenie *dropoutu* na poziomie wszystkich warstw modelu („W – 10%”, „W – 25%”). Poprawa z tego tytułu była zauważalna nie tylko dla architektury dwukanałowej, ale też w przypadku ograniczenia się wyłącznie do bazowych spektrogramów. Model typu „B” w wariancie „W – 25%” był jedną z nielicznych kombinacji hiperparametrów, dla której udało się przekroczyć poziom 50% przy klasyfikacji jednokanałowego Minibirds.

3.6. Wykorzystanie normalizacji partiami

Wprowadzenie normalizacji partiami (*batch normalization*) jest ważnym aspektem uczenia modeli o dużej głębokości, co potwierdzają przeprowadzone eksperymenty. Dla architektury *SqueezeNet* brak tego rozwiązania kompletnie uniemożliwiał efektywne uczenie, dla pozostałych modeli dysproporcje nie były tak drastyczne, ale spadek dokładności był zdecydowanie zauważalny.

3.7. Zastosowana funkcja aktywacji neuronów

W zakresie zestawienia ze sobą alternatywnych form aktywacji typu *ReLU* (*Leaky ReLU, Parametric ReLU, Exponential Linear Unit, Scaled Exponential Linear Unit)* trudno było wskazać na różnice o charakterze sytematycznym. Ewidentnie nieskuteczne okazało się natomiast wykorzystanie funkcji logistycznej (sigmoidalnej). Chociaż tangens hiperboliczny został w obecnie stosowanych modelach praktycznie całkowicie wyparty przez warianty *ReLU*, to w eksperymencie nie ustępował on dokładnością aż w tak znaczącym stopniu, jak można by oczekiwać po jego małej popularności w obecnie stosowanych modelach.

3.8. Ustawienia procesu uczenia (optymalizatora)

Zakres ustawień przynoszących akceptowalne rezultaty był dosyć szeroki, a wykorzystanie *momentum* pozwalało w dużej części uniknąć negatywnych konsekwencji zbyt małego tempa uczenia. Mimo wszystko, dostrojenie tej wartości było nadal ważne. Co więcej, nawet w przypadku metod o charakterze adaptacyjnym, które w obiegowej opinii nie wymagają takich zabiegów, była to kwestia istotnie wpływająca na wyniki.

3.9. Siła regularyzacji

Poza ustawieniem ze skrajnie wysoką wartością, wpływ regularyzacji na dokładność klasyfikacji można

uznać za kosmetyczny. Korzyścią z utrzymania niewielkiego stopnia regularyzacji jest możliwość otrzymania filtrów splotowych lepiej nadających się do późniejszej interpretacji wizualnej.

3.10. Metoda inicjalizacji wag

Niewielkie różnice w uzyskiwanych wynikach pokazują, że nie jest to obecnie obszar kluczowy przy uczeniu modeli splotowych. O ile sytuacja ta wyglądała zgoła inaczej jeszcze w niedalekiej przeszłości, to wprowadzenie procedury normalizacji partiami pozwoliło właściwie zapomnieć o tym aspekcie doboru hiperparametrów. Z tego powodu poprzestanie na domyślnych sposobach inicjalizacji proponowanych przez najpopularniejsze biblioteki wydaje się całkowicie wystarczające.

3.11. Rozmiar partii uczącej

Ustalenie liczby próbek wykorzystywanych w każdej iteracji uczenia wpływa wieloaspektowo na cały proces. Wartość domyślna przyjęta w eksperymentach (128) jest bezpiecznym punktem wyjścia. Małe liczby przykładów (16, 32) powodowały pogorszenie wyników najprawdopodobniej związane z niestabilnością wyznaczania gradientu w czasie uczenia.

3.12. Długość segmentu

W przypadku *Minibirds* modyfikacje tego aspektu uczenia nie powodowały znaczących zmian w uzyskiwanych wynikach. Dla *UrbanSound8K* zwiększenie długości segmentu uczącego wpływało pozytywnie na architekturę typu *K*. Z kolei dla *ESC-50* nieznaczną poprawę dla architektury typu *B* przynosiło skrócenie segmentów, czego dodatkową korzyścią było zmniejszenie wymagań obliczeniowo-pamięciowych i szybsze uczenie modelu.

3.13. Rozdzielczość spektrogramu

Zwiększenie liczby pasm czy skrócenie przeskoku drastycznie wydłużają czas uczenia standardowych modeli z obszaru rozpoznawania obrazów. Przy wartościach skrajnych (200 pasm, 10 ms) ich wymagania względem pamięci karty graficznej wykraczają ponad obecne możliwości sprzętowe.

Pod tym względem osobliwym zachowaniem odznaczają się modele typu *1D CNN*, dla których zwiększenie rozdzielczości wertykalnej (liczby pasm częstotliwości) powoduje tylko niewielki przyrost złożoności obliczeniowej zamykający się na poziomie pierwszej warstwy splotowej. Na ich przykładzie widoczny jest potencjał zastosowania reprezentacji o zwiększonej rozdzielczości do zadań klasyfikacji dźwięków środowiskowych [6]. Najlepszą kombinacją w eksperymentach okazało się połączenie 100 pasm częstotliwości z ramką o długości 10 ms.

4. WNIOSKI

Wnioskiem podsumowującym płynącym z analizy uzyskanych wyników eksperymentalnych jest podkreślenie istotności doboru odpowiednich nastaw hiperparamterów. Porównanie najlepszych modeli przy ustawieniach początkowych i dostrojonych wykazuje różnice nawet do 7 punktów procentowych bez zmiany architektury modelu.

W kwestii bardziej szczegółowych spostrzeżeń, w eksperymentach potwierdzona została tendencja dokładniejszego klasyfikowania przez modele dwukanałowe. Kluczową rolę okazuje się odgrywać *dropout* dobierany odpowiednio do danej architektury. Nieodzownym elementem w sieciach głębokich okazuje się też być technika *batch normalization*. Dla modeli wykorzystujących filtry wertykalne (*1D CNN*) ważny jest aspekt uodpornienia modelu na przesunięcia w dziedzinie częstotliwości zapewniany przez filtry pierwszej warstwy o ograniczonej wysokości.

Chociaż standardowe modele z obszaru przetwarzania obrazów radzą sobie bardzo dobrze z klasyfikacją nagrań na podstawie spektrogramów, to zastosowanie w pierwszej warstwie wertykalnych filtrów splotowych pozwala na uzyskanie porównywalnych wyników. Istotną zaletą takiego rozwiązania jest natomiast większa wydajność uczenia i generowania predykcji, skalowalność przy zwiększaniu rozdzielczości częstotliwościowej spektrogramu, a także bogatsza wartość informacyjna w przypadku pobieżnej wizualizacji wag pierwszej warstwy, do której niestety często ogranicza się diagnostyka modeli splotowych. Z tego powodu rozwiązania te są atrakcyjną alternatywą dla typowych modeli przetwarzania obrazów.

LITERATURA

[1] Büchler, M. et al. 2005. „Sound Classification in Hearing Aids Inspired by Auditory Scene Analysis". *EURASIP Journal on Applied Signal Processing*, 18: 2991–3002.

[2] Espi, M. et al. 2015. „Exploiting Spectro-Temporal Locality in Deep Learning Based Acoustic Event Detection". *EURASIP Journal on Audio, Speech, and Music Processing*, 26: 1–12.

[3] Piczak, K. J. 2015. „Environmental Sound Classification with Convolutional Neural Networks". *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Boston, USA: 1–6.

[4] Piczak, K. J. 2015. „ESC: Dataset for Environmental Sound Classification". *Proceedings of the ACM International Conference on Multimedia (ACMMM)*, Brisbane, Australia: 1015–1018.

[5] Piczak, K. J. „Klasyfikacja dźwięku za pomocą splotowych sieci neuronowych". Rozprawa doktorska, Politechnika Warszawska, 2018.

[6] Piczak, K. J. 2017. „The details that matter: Frequency resolution of spectrograms in acoustic scene classification". *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, Monachium, Niemcy: 103–107.

[7] Salamon, J. et al. 2014. „A Dataset and Taxonomy for Urban Sound Research". *Proceedings of the ACM International Conference on Multimedia (ACMMM)*, Orlando, USA: 1041–1044.

[8] Zhang, H. et al. 2015. „Robust Sound Event Recognition Using Convolutional Neural Networks". *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia: 559–563.