Інформаційні технології

УДК 004.8

Іванов Олександр Андрійович

студент

Національного технічного університету України

«Київський політехнічний інститут імені Ігоря Сікорського»

АЛГОРИТМ НАВЧАННЯ ЗГОРТКОВИХ НЕЙРОННИХ МЕРЕЖ ДЛЯ РОЗПІЗНАВАННЯ АУДІОПОДІЙ

Анотація. *В роботі описані методи та алгоритми навчання нейронних мереж. Також запропоновано організацію навчання нейронної мережі, що була успішно застосована до загорткової нейронної мережі для вирішення задачі класифікації аудіоподій.*

Ключові слова: *навчання, класифікація аудіоподій, згорткові нейронні мережі, глибокі нейронні мережі.*

Вступ. Глибокі нейронні мережі (DNN) останнім часом досягли великих успіхів у багатьох завданнях класифікації. У цій роботі використовуються DNN для класифікації зображень, та їх застосування до завдання класифікації аудіоподій.

Історично класифікації аудіоподій (AED - Acoustic Event Detection) розглядається з такими ознаками звуку, як MFCC та класифікаторами на основі GMM [1], HMM [2], NMF [3] або SVM [4]. Новіші підходи використовують певну форму DNN, включаючи CNN [5] і RNN [6]. Зокрема, глибокі згорткові нейронні мережі (CNN) дуже добре підходять до проблеми класифікації звуку навколишнього середовища: по-перше, вони здатні захоплювати моделі модуляції енергії через час і частоту при застосуванні до входів спектрограм. Видно, що це є важливою ознакою для

розрізнення різних, часто шумоподібних, звуків, таких як двигуни та відбійні молотки [7]. По-друге, використовуючи згорткові ядра (фільтри) з малим рецептивним полем, мережа повинна вміти успішно вивчати і пізніше виявляти спектро-часові структури, які є репрезентативними для різних класів звуку, навіть якщо частина звуку маскується (у часі або частоті) іншими джерелами (шум), де традиційні аудіо ознаки працюють не коректно [8].

Глибокі нейронні мережі, які мають високу модельну потужність, особливо залежать від наявності великої кількості навчальних даних для того, щоб дізнатися нелінійну функцію від входу до виходу, що добре узагальнює і дає високу точність класифікації за невидимими даними. Можливим поясненням обмеженого використання CNN для завдання класифікації аудіоподій є відносний дефіцит розмічених даних. Хоча в останні роки було випущено кілька нових наборів даних вони все ще є значно меншими, ніж набори даних, доступні для дослідження класифікації зображень [9].

У цій роботі представлено алгоритм навчання загорткових нейронних мереж, який дозволяє досягти значної точності за малий проміжок часу з використанням невеликого набору даних.

Набір даних. У цій роботі було використано набір даних UrbanSound8K [10]. Цей набір даних містить 8732 розмічених аудіо файлів у форматі WAV тривалістю менше 4 секунд. Частота дискретизації, бітова глибина і кількість каналів такі ж, як і у вихідних файлів, завантажених з Freesound (Freesound.org - спільне сховище ліцензованих аудіосемплів), отже можуть змінюватися від файлу до файлу. Кожен запис відноситься до одного з 10 класів: кондиціонер, клаксон автомобіля, діти, гавкання собаки, буріння, двигун автомобіля, постріл, відбійник, сирена, музика. Всі дані взяти з польових записів завантажених з Freesound.

Попередня обробка даних. Аудіо дані проходять попередню обробку: частота дискретизації перетворюється до 22,05 кГц, а також скорочує кількість каналів до 1 (моно), також дані нормалізуються таким чином, щоб значення перебували в діапазоні від -1 до 1. Завдяки цьому вирішено усі проблеми пов'язані з різними частотами дискретизації, кількістю каналів, діапазонами значень у оригінальному наборі даних. Оброблені аудіо дані розкладаються короткочасним перетворенням Фур'є. Далі отримана спектрограма перетворюється у лог-мел спектрограму розміром 224x224, яка формує вхід до всіх класифікаторів.

Навчання моделі. Для експерименту було обрано одну з сучасних загорткових нейронних мереж ResNet50 [11]. Модель навчалася на графічному процесорі Nvidia Tesla K80. При цьому використовувався оптимізатор Adam. В якості функції втрат застосовується крос-ентропія. Після кожного загорткового шару застосовується пакетна нормалізація. Також для уникнення перенавчання використовуються шари Dropout та інші поширені методи регуляризації. Набір даних поділено на пакети розміром 32. Розрахунок функції втрат для набору навчання проходить на кожній ітерації, для набору валідації – наприкінці кожної епохи. Під час навчання прогрес відслідковується за допомогою accuracy, precision, recall та F score.

Для порівняння було використано три підходи до навчання нейронної мережі. Усі варіанти тренуються лише 10 епох та наприкінці порівнюється точність, яку модель встигла досягти за такий короткий проміжок часу.

У якості базового рівня була взята модель ResNet50, яка навчалася без використання підходів до пришвидшення збіжності нейронної мережі. У якості початкових ваг використовуються випадкові значення. Під час навчання значення швидкості навчання та моменту дорівнює 0.003 та 0.9

відповідно та залишається незмінним. Графік функції втрат показано на рисунку 1.

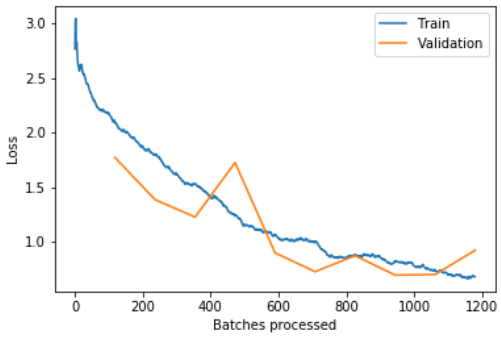

Рис. 1. Графік функції втрат першого підходу

Під час другого підходу використовувалася техніка трансферного навчання. Трансферне навчання - це метод навчання по вже існуючій/навченій моделі, яка була навчена з використанням методу контрольованого/неконтрольованого навчання і володіє характеристиками відмінного розпізнавання ознак. Для цього навчають тільки декілька прихованих шарів поверх вже існуючих нейронних мереж, і підлаштовують класифікатор, використовуючи дуже маленький обсяг даних, все ще досягаючи тієї ж точності в порівнянні з повністю навченими моделями. У якості початкових ваг використовуються ваги моделі, яка була попередньо натренована на наборі даних ImageNet [12]. Тренування поділено на два етапи: перший етап складається з п'яти епох, на яких навчаються тільки останні класифікуючи шари; другий етап fine-tune складається з п'яти епох на яких розморожуються усі ваги та нейронна мережа навчається повністю. Графік функції втрат для цього підходу показано на рисунку 2.

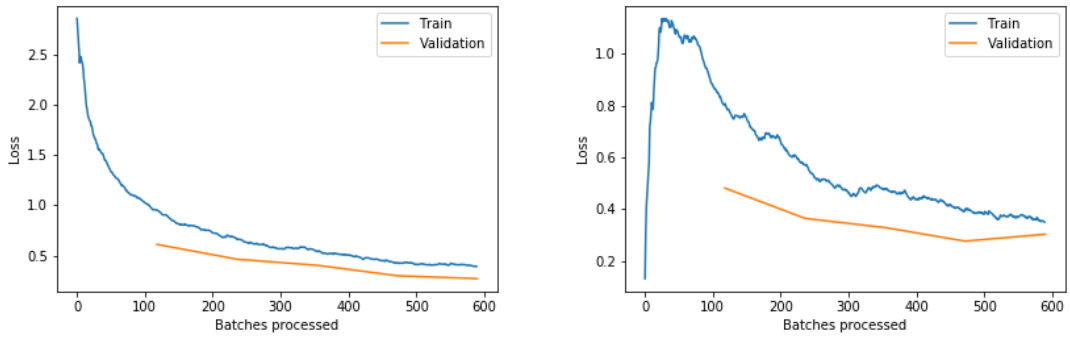

Рис. 2. Графіки функції втрат під час двох етапів другого підходу

У третьому підході на додачу до техніки трансферного навчання використовується політика 1-го циклу Леслі Сміта [13]. Такий підхід дає дуже швидкі результати для навчання складних моделей. Під час навчання мінімальне та максимальне значення моменту становить 0.85 та 0.95 відповідно. На першому етапі максимальна швидкість навчання дорівнює 0.003 та змінюється під час навчання, яка показано на графіку (рис. 3). На етапі fine-tune максимальна швидкість навчання знаходиться в проміжку [1e-4, 1e-3] та змінюється так само.

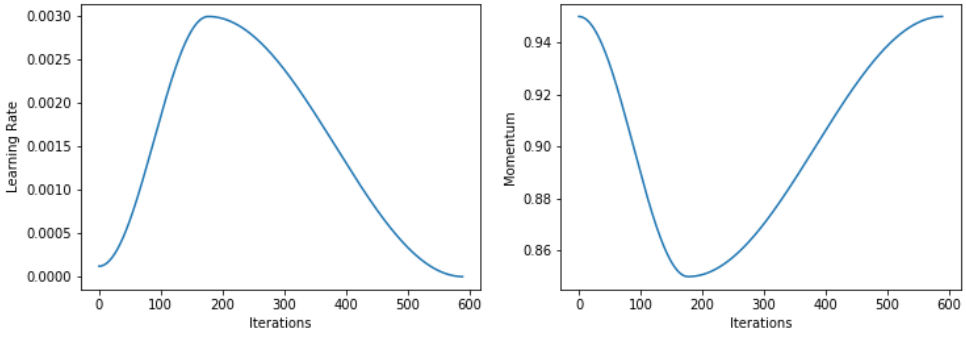

Рис. 3. Зміна швидкості навчання та моменту під час навчання нейронної мережі за політикою 1-го циклу Леслі Сміта

Графік функції втрат для третього підходу показано на рисунку 4.

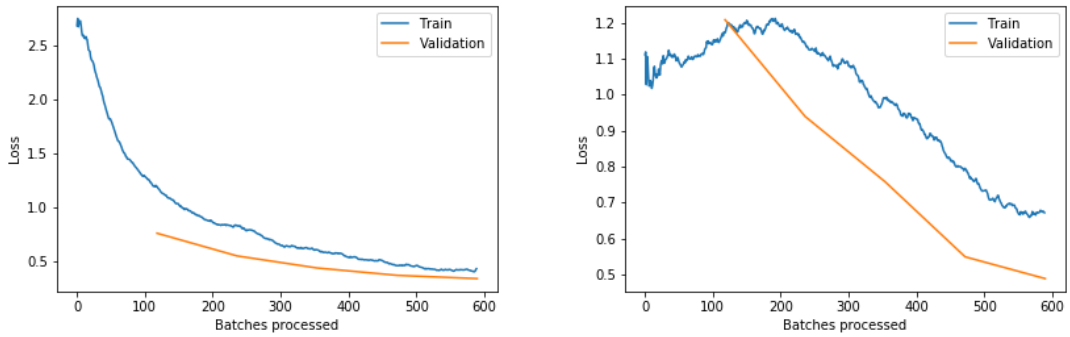

Рис. 4. Графіки функції втрат під час двох етапів третього підходу

Таблиця 1

Результати навчання

Підхід	Train loss	Validation loss	Accuracy
Без використання підходів для пришвидшення навчання	0.678033	0.921396	0.693252
З трансферним навчанням	0.349470	0.302574	0.900613
З трансферним навчанням та політикою 1-го циклу	0.116651	0.095018	0.974233

З таблиці видно, що запропонована комбінація методів підвищення швидкості навчання моделі показує досить високі результати. За короткий проміжок часу модель досягла точності у 97%. У той час, як без використання комбінації методів модель встигла досягти лише 69%.

Висновки. Запропонований алгоритм навчання показує високу швидкість збіжності моделі. Навчені за алгоритмом моделі досягають високої точності класифікації аудіоподій за малий проміжок часу. Після 10 епох навчання точність класифікації вже на 28% більше ніж у моделі, яка тренується без використання комбінації методів.

Література

1. A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings," in Signal Processing Conference, 2010 18th European. IEEE, 2010, pp. 1267–1271.

2. X. Zhuang, X. Zhou, M. A. Hasegawa-Johnson, and T. S. Huang, "Real-world acoustic event detection," Pattern Recognition Letters, vol. 31, no. 12, pp. 1543–1551, 2010.
3. J. F. Gemmeke, L. Vuegen, P. Karsmakers, B. Vanrumste, et al., "An exemplar-based nmf approach to audio event detection," in 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. IEEE, 2013, pp. 1–4.
4. A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, "Clear evaluation of acoustic event detection and classification systems," in International Evaluation Workshop on Classification of Events, Activities and Relationships. Springer, 2006, pp. 311–322.
5. Jordi Pons, Xavier Serra, "Randomly weighted CNNs for (music) audio classification," ICASSP, May, 2018.
6. G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 6440–6444.
7. "Feature learning with deep scattering for urban sound analysis," in 2015 European Signal Processing Conference, Nice, France, Aug. 2015.
8. C. V. Cotton and D. P. W. Ellis, "Spectral vs. spectro-temporal features for acoustic event detection," in IEEE Worksh. on Apps. of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, Oct. 2011, pp. 69–72.
9. A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in Advances in neural information processing systems (NIPS), 2012, pp. 1097–1105.
10. Justin Salamon, Christopher Jacoby, and Juan Pablo Bello, "A dataset and taxonomy for urban sound research," in Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014, pp. 1041–1044.

11. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," arXiv preprint arXiv:1512.03385, 2015.
12. Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, Li Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," arXiv preprint arXiv:1409.0575, 2014.
13. Leslie N. Smith, "A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum and weight decay," arXiv preprint arXiv:1803.09820, 2018.