УДК 004.89

Автоматичне розпізнавання музичних жанрів глибокими згортковими нейронними мережами

Дорогий Я. Ю., к.т.н., доц., ORCID 0000-0003-3848-9852
e-mail argusyk@gmail.com
Цуркан В. В., к.т.н., ORCID 0000-0003-1352-042X
e-mail v.v.tsurkan@gmail.com
Хапілін О. С., ORCID 0000-0002-3888-584X
e-mail khapilins@yandex.ua
Національний технічний університет України
"Київський політехнічний інститут імені Ігоря Сікорського" kpi.ua
Київ, Україна

***Реферат*—В статті розглядаються алгоритми для автоматичного розпізнавання музичних жанрів та пропо-нується використання глибоких згорткових нейронних мереж для цієї задачі. Спираючись на реальні дані, окреслено архітектуру мережі та оцінено її якість. Робота виконана з використанням дата-сету GTZAN. Було розглянуто задачу класифікації для чотирьох та десяти жанрів з використанням мел-кепстральних коефіцієнтів та аудіо хвилі в якості ознак. Якість запропонованого алгоритму було протестовано на відкладених даних для чотирьох та десяти різних жанрів та порівняно з використанням обмеженої машини Больцмана для чотирьох жанрів.**

Бібл. 11, рис. 3, табл. 2.

Ключові слова — глибокі нейронні мережі; згорткові мережі; пошук музичної інформації; класифікація.

I. ВСТУП

Музичний жанр – характеристика музики, що полягає у певних композиційних та стилістичних ознаках, таких, як інструментальна та ритмічна структура, характеристика вокалу виконавця тощо. Зазвичай жанри використовуються для організації музичних композицій, наприклад, для розміщення різних альбомів в музичних магазинах чи для про-позиції в інтернет-радіо. Із швидким ростом кількості му-зики ручна класифікація композицій стає недоцільною та ресурсоємною. Оскільки існуючих рішень мало, а їх якість ще не досягла достатнього для масового впровадження рівня, дана область досліджень є перспективною. Крім того, все більш широкого поширення набувають такі музичні сервіси, як Spotify, і Tunes та Pandora, що з їх кількістю музичних даних повинні бути зацікавленні в їх автоматичній обробці. Крім того, натреновані на класифікації жанрів мережі можна спробувати використати як основу для інших завдань, наприклад, для пошуку схожих му-зичних композицій. В цій роботі розглянуто метод на основі глибоких згорткових мереж для автоматичної класифікації музичних жанрів.

II. АНАЛІЗ ЛІТЕРАТУРНИХ ДАНИХ І ПОСТАНОВКА ПРОБЛЕМИ

Хоча завдання класифікації музичних жанрів менш поширене, ніж, наприклад, зображень, кількість підходів досить значна. На практиці зазвичай використовують як методи навчання без вчителя [1], [2], так і з учителем [3]. В першому випадку дослідники використовували глибокі згорткові обмежені машини Больцмана [4] для отримання похідних від спектру ознак.

У другому джерелі використовувалися кепстральні мел-коефіцієнти та обмежена машина Больцмана для отримання ознак, що потім відправлялися до нейронної мережі прямого поширення.

Структура машин Больцмана складається з видимого та прихованого шару. В разі глибоких машин Больцмана прихованих шарів декілька, і кожен прихований шар є видимим для наступного шару. Така структура дозволяє створювати складні ієрархічні ознаки без учителя, які можна буде використовувати в подальшому для класифікації. Конфігурація такої системи описується її енергією:

$$E(v,h) = -\sum_i a_i v_i - \sum_j b_j h_j - \sum_i \sum_j v_i w_{i,j},$$

де v, h – вектори видимого та прихованого шарів, w – вагові коефіцієнти між видимим та прихованим шаром, a, b – коефіцієнти зсуву для видимого та прихованого шарів.

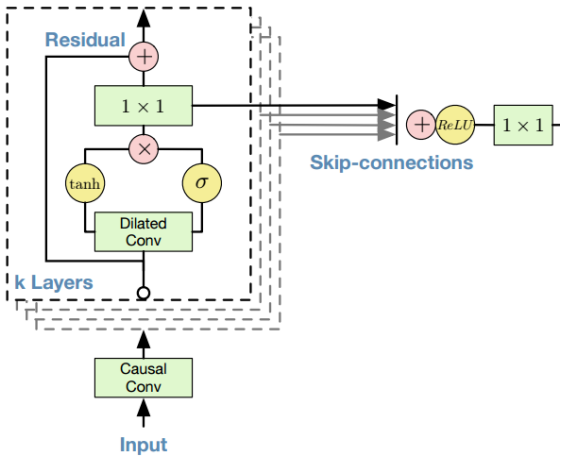

Рис. 1 Залишковий блок

Розподіли ймовірностей в шарах виражаються через функцію енергії:

$$P(v,h)=\frac{1}{Z}e^{-E(v,h)}.$$

В джерелі [3], було використано як алгоритми навчання з учителем, так і без. До алгоритмів, що було використано, належать k-NN, k-means, SVM, нейронні мережі прямого поширення [5]–[8].

В джерелі [3], було використано як алгоритми навчання з учителем, так і без. До алгоритмів, що було використано, належать k-NN, k-means, SVM, нейронні мережі прямого поширення [5]–[8].

У дослідженні [9] увагу було зосереджено на ознаках для здобування інформації з аудіо. В цій роботі разом з описаними нижче мел-кепстральними коефіцієнтами було використано: частоту пересічення нуля, центроїд, rolloff.

Частота пересічення нуля – величина, що опи-сує кількість разів, що амплітуда аудіо хвилі в імпульсному кодуванні змінила знак.

Центроїд —

$$C=\frac{\sum_{1}^{T} fM[f]}{\sum_{1}^{T} M[f]},$$

де – значення швидкого перетворення Фур'єдля проміжку частот , а – кількість проміжків.

Rolloff – таке значення R, що

$$\sum_{1}^{R} M[f]=0.85\sum_{1}^{N} M[f]$$

Далі в роботі використовувалися конкатеновані ознаки, обраховані на всьому аудіо файлі.

III. МЕТА І ЗАВДАННЯ ДОСЛІДЖЕННЯ

Метою роботи є розробка і навчання згорткової нейронної мережі для задачі класифікації музич-них жанрів.

IV. ОПИС МОДЕЛІ ТА ЕКСПЕРИМЕНТАЛЬНІ ДОСЛІ-ДЖЕННЯ

A. Використані ознаки

Було проведено експерименти з двома типами музичних ознак: нормованих значень амплітуд імпульсно-кодової модуляції аудіо хвилі та кепстральних мел-коефіцієнтів. Значення амплітуд зазвичай зчитуються і декодуються із аудіо файлів напряму. Кепстральні коефіцієнти отримуються наступним чином:

- аудіо файл розбивається на відрізки;
- до цих відрізків застосовують Хеммінгове вікно;
- застосовується швидке перетворення Фур'є;
- частоти у герцах переводять у мел-шкалу, яка майже лінійна в області низьких частот і логарифмічна в області високих (це більше відповідає сприйняттю звуків людиною);
- до отриманого спектра застосовується перетворення по косинусу.

В роботі було використано кадри довжиною 23 мс, зміщенням 6 мс та 50 перших коефіцієнтів.

B. Тренувальна вибірка

В якості тренувальної вибірки використовувався дата-сет GTZAN [10]. В ньому міститься 1000 музичних записів 10 жанрів, по 100 пісень в одному жанрі. Частота дискретизації 22050 відліків на секунду. Кожна пісня довжиною 30 секунд. В роботі було використано 700 пісень для навчання, 300 пісень — для тестування нейронної мережі.

C. Модель

За основу було взято нещодавно розроблену мережу WaveNet [11] з кількома ключовими відмінностями. По-перше, в оригінальній мережі усі згорткові шари маскують свій вхідний сигнал таким чином, щоб кожна операція згортки не виконувалася із останнім входом. Це забезпечує можливість моделювати причинні зв'язки і генерувати звук, що і було метою в цієї статті. В даній роботі немає сенсу обмежувати поле зору мережі, тому було використано стандартну операцію розширеної згортки. Завершує мережу global average pooling та стек повнозв'язних шарів.

ТАБЛИЦЯ 1 ТОЧНІСТЬ КЛАСИФІКАЦІЇ ДЛЯ ДЕСЯТИ КЛАСІВ ДЛЯ РІЗНИХ ОЗНАК

	Тренування	**Тестування**
MFCC	99%	46.3%
Аудіохвиля	53.67%	31.7%

Основною структурною одиницею мережі є залишкові блоки (рис. 1). Вони складаються із двох наборів фільтрів, над якими виконується операція згортки із вхідним сигналом. Після чого до них застосовується така функція активації:

$$z = \text{th}\left(W_{f,k} * x\right) \odot \sigma\left(W_{g,k} * x\right),$$

де W – ваги фільтрів, $*$ – операція згортки, $\odot$ – по елементне множення.

До результату функції активації застосовується 1x1 згортка, результат якої додаються з усіма іншими шарами. Також до нього додається залишкове з'єднання.

В роботі було використано 30 шарів, кожен набір фільтрів має ширину 4 та 32 канали.

D. Результати

Точність моделі для класифікації 10 класів, що навчалася на кепстральних коефіцієнтах та на аудіохвилі, можна побачити в табл. 2. Серед використаних класів були блюз, класика, кантрі, диско, хіп-хоп, джаз, метал, поп, реггі та рок. У другій серії експериментів виконувалася класифікація лише 4 жанрів: класика, кантрі, метал та реггі. Використані ознаки були ті ж самі: мел-кепстральні коефіцієнти та аудіо хвиля. Результати другої серії експериментів наведено у табл. 2. Навчальна вибірка була поділена у відношенні 70% для тренування та 30% для тестування. Таким чином для кожного жанру було присутньо 70 композицій в навчальній виборці та 30 в тестовій.

Значення цільової функції можна побачити на рис. 2 та рис. 3. На рис. 2 по осі абсцис позначено ітерацію навчання, по осі ординат значення функції втрат. Рис. 2 показує результати для класифікації 10 жанрів, де верхній графік отриманий з використанням мел-кепстральних коефіцієнтів, а нижній — з використанням аудіо хвилі. Рис. 3 відображає аналогічні значення, за винятком того, що розв'язувалася задача класифікації для 4 жанрів.

ТАБЛИЦЯ 2 ТОЧНІСТЬ КЛАСИФІКАЦІЇ ДЛЯ ЧОТИРЬОХ КЛАСІВ ДЛЯ РІЗНИХ ОЗНАК

	Тренування	**Тестування**
MFCC	99%	61.1%
Аудіо хвиля	95.9%	76.18%
RBM	74.3%	61.15%

Варто зазначити, що модель дуже сильно перенавчається, особливо на кепстральних коефіцієнтах. Перенавчання – явище, за якого модель показує високу якість на тренувальній виборці та набагато гіршу на відкладеній. Для аудіо хвилі така проблема менш помітна, проте час на навчання суттєво більший.

В роботі також проводилася класифікація тільки чотирьох жанрів, так само, як і у [2].

Як видно з табл. 2, проблема перенавчання не зникла, проте тренування на аудіо хвилі дало на тестовій виборці найкращий результат, перевершивши інші описані способи на 15%.

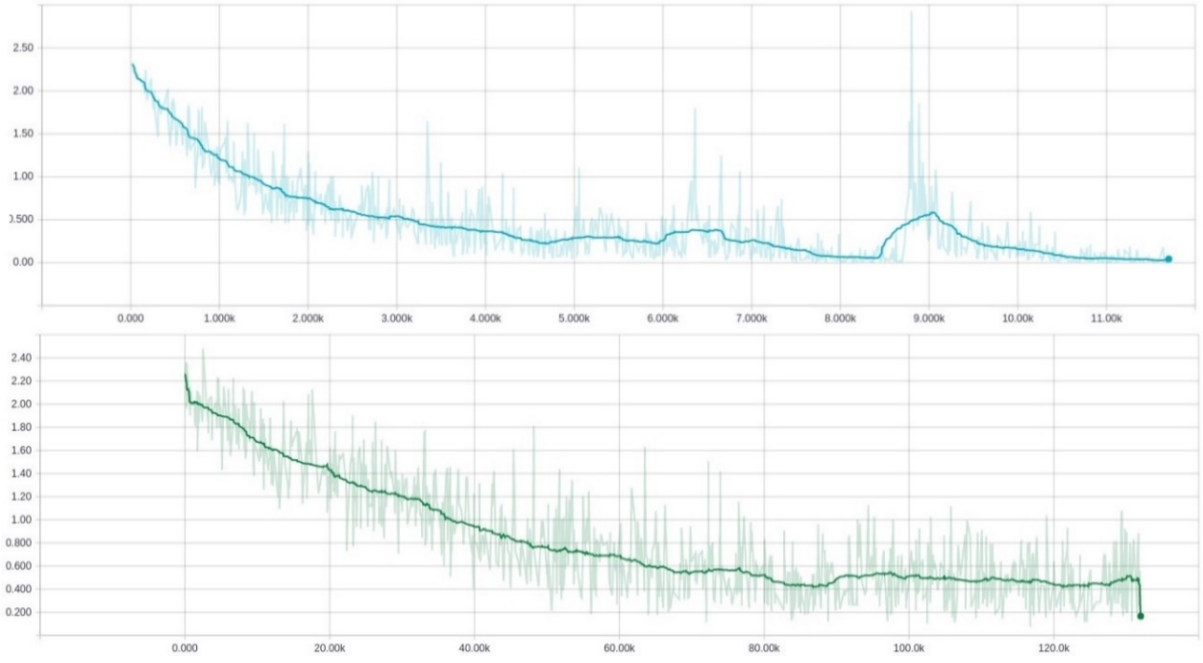

Рис. 2 Падіння цільової функції для ознак MFCC та аудіо хвилі відповідно для 10-ти жанрів

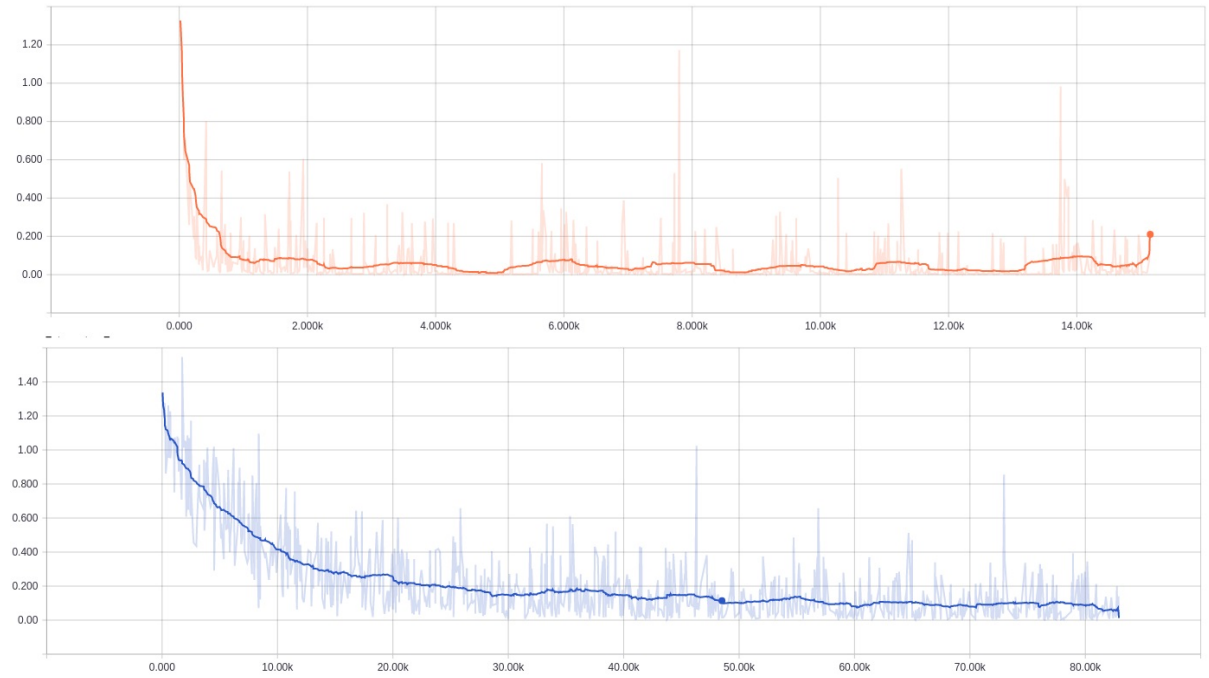

Рис. 3 Падіння цільової функції для ознак MFCC та аудіо хвилі відповідно для 4-ох жанрів

Висновки

Результати показують, що розглянута модель доволі сильно перенавчається на дата-сеті GTANZ. Найвирогідніша причина перенавчання – мала тренувальна вибірка, яка містить усього 1000 уривків композицій по 30 секунд. В той час, коли для більш класичних алгоритмів такої кількості було б достатньо, для глибоких мереж велика кількість даних є критичною для досягнення високої якості. У випадку із більшою кількістю класів перевагу кепстральних коефіцієнтів над аудіо хвилею можна пояснити тим, що вони були розробленні для систем розпізнавання мови і несуть у собі більше важливої інформації про спектр, одночасно нехтуючи менш важливою (наприклад, такою, як обертони) для кращого розрізнення кадрів. У випадку із меншою кількістю класів мережа проявила себе краще, що не дивно, адже зі зменшенням кількості класів задача спрощується. Особливо зменшення класів позитивно вплинуло на навчання із аудіо хвилею, що може бути наслідком більш складної структури даних в такому випадку та наявності більшої кількості даних для навчання, порівняно із кепстральними коефіцієнтами, оскільки під час використання аудіо хвилі значення імпульсно-кодової модуляції використовуються окремо, а не поєднуються в кадри, на відміну від кепстральних коефіцієнтів, що забезпечило менше перенавчання і краще узагальнення. Не зважаючи на перенавчання, нейронна мережа дає точність на 15% вищу, ніж у RBM Тао Фенга [2]. Таким чином, модель з розширеними згортками та активацією з пригніченням навчається не менш складним фільтрам для аналізу звука.

Перелік посилань

[1] H. Lee, P. Pham, Y. Largman, and A. Y. Ng, “Unsupervised feature learning for audio classification using convolutional deep belief networks,” in *Advances in Neural Information Processing Systems 22*, Curran Associates, Inc., 2009, pp. 1096–1104, **URL:** http://papers.nips.cc/paper/3674-unsupervised-feature-learning-for-audio-classification-using-convolutional-deep-belief-networks.pdf.

[2] T. Feng, “Deep learning for music genre classification,” **URL:** https://courses.engr.illinois.edu/ece544na/fa2014/Tao_Feng.pdf.

[3] M. Haggblade, Y. Hong, and K. Kao, “Music Genre Classification,” Stanfor, 2009, **URL:** https://cs229.stanford.edu/Fproj2011/HaggbladeHongKao-MusicGenreClassification.pdf&usg=AOvVaw1Nc3dcWm4P4rW6UsiLu712.

[4] G. E. Hinton, S. Osindero, and Y.-W. Teh, “A Fast Learning Algorithm for Deep Belief Nets,” *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006, **DOI:** 10.1162/neco.2006.18.7.1527.

[5] Y. Cai, D. Ji, and D. Cai, “A KNN Research Paper Classification Method Based on Shared Nearest Neighbor,” in *NTCIR-8 Workshop Meeting*, 2010, pp. 336–340.

[6] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, “An efficient k-means clustering algorithm: analysis and implementation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 881–892, 2002, **DOI:** 10.1109/TPAMI.2002.1017616.

[7] M. Mandel and D. Ellis, “Song-Level Features And Support Vector Machines For Music Classification,” in *6th International Conference on Music Information Retrieval*, 2005, pp. 594–599, **URL:** http://ismir2005.ismir.net/proceedings/1106.pdf.

[8] I. Goodfellow, Y. Bengio, and A. Courville, “Deep feedforward networks,” in *Deep learning*, The MIT Press, 2016, pp. 164–223, **URL:** http://worldcat.org/isbn/9780262035613.

[9] G. Tzanetakis, G. Essl, and P. Cook, “Automatic Musical Genre Classification Of Audio Signals,” in *2nd Annual International Symposium on Music Information Retrieval 2001*, 2001, **URL:** http://ismir2001.ismir.net/pdf/tzanetakis.pdf.

[10] J. Leben, “Data Sets,” *MARSYAS*, 2001. [Online]. Available: http://marsyas.info/downloads/datasets.html.

[11] A. van den Oord *et al.*, “WaveNet: A Generative Model for Raw Audio,” Sep. 2016, **arXiv:** 1609.03499v2.

Надійшла до редакції 24 червня 2017 р.

УДК 004.89

Автоматическое распознавание музыкальных жанров глубокими сверточными нейронными сетями

Дорогой Я. Ю., к.т.н., доц., ORCID 0000-0003-3848-9852
e-mail argusyk@gmail.com
Цуркан В. В., к.т.н., ORCID 0000-0003-1352-042X
e-mail v.v.tsurkan@gmail.com
Хапилин А. С., ORCID 0000-0002-3888-584X
e-mail khapilins@yandex.ua
Национальный технический университет Украины
"Киевский политехнический институт имени Игоря Сикорского" kpi.ua
Киев, Украина

***Реферат*—В статье рассматриваются алгоритмы для автоматического распознавания музыкальных жанров и предлагается использование глубоких сверточных нейронных сетей для этой задачи. Архитектура сети описана и ее качество оценено на реальных данных. Работа выполнена с использованием дата-сета GTANZ. Были рассмотрены задачи классификации для четырех и десяти жанров с использованием мел-кепстральных коэффициентов и аудио волны в качестве признаков. Качество предложенного алгоритма было протестировано на отложенных данных для четырех и десяти разных жанров и сравнено с использованием ограниченной машины Больцмана для четырех жанров.**

Библ. 11, рис. 3, табл. 2.

Ключевые слова — глубокие нейронные сети; сверточные сети; поиск музыкальной информации; классификация.

UDC 004.89

Automatic musical genre recognition using deep convolutional neural networks

Ya. Yu Dorohyi, PhD, Assoc.Prof., ORCID 0000-0003-3848-9852
e-mail argusyk@gmail.com
V. V. Tsurkan, PhD, ORCID 0000-0003-1352-042X
e-mail v.v.tsurkan@gmail.com
O. S. Khapilin, ORCID 0000-0002-3888-584X
e-mail khapilins@yandex.ua
National technical university of Ukraine "Igor Sikorsky Kyiv polytechnic institute" kpi.ua
Kyiv, Ukraine

***Abstract*—For the long time in computer vision and digital signal processing manually developed algorithms and filters were used. With the development of computers technics and constantly growing amount of available data samples, these algorithms became less accurate than modern machine learning approaches. The idea behind them is to construct useful representations based on data itself rather than on expert knowledge. Such approach allows machine learning algorithms**

to choose for themselves which parts of data more important. Today machine learning is successfully applied in such tasks as image recognition in Google image search and speech recognition in Google Now, Siri, Cortana. Nowadays best approaches are built upon different variations of neural network algorithms. One of the fields, where machine learning are successfully applied is music information retrieval, where musical genres classification is one of the main tasks and solving it efficiently can help automatically organize large collections of musical data which are available for now. As music genre aggregates a lot of song information, model for calculating music song similarities based on audio information can possibly be built on proposed model.

In this article, the algorithms for automatic music genre recognition are discussed and usage of deep convolutional neural networks is proposed for this task. The network's architecture is described and its quality evaluated on real-world data. In this work GTZAN dataset is used and classification problem for four and ten genres classification was examined using mel-frequency cepstral coefficients and waveform as features. The quality of proposed algorithm was evaluated on hold-out set for four and ten different genres and compared to using restricted Boltzmann machines for four genres classification. The resulting accuracy for our genres classification task is 76%, which is about 15% better than restricted Boltzmann machine approach. Though model overfits strongly on rather small dataset it can be fixed by using larger amount of data.

The main differences between proposed neural network architecture and traditional convolutional neural networks are gated activations, dilated convolutions and residual connections. Gated activations allow the network to additionally weight and inhibit importance of intermediate features like it is done in recurrent neural networks. Dilated convolutions allow increasing receptive field of network's filters while maintaining small number of trainable parameters. Residual connections are proven to be vital feature for very deep neural networks to prevent gradient degrading and neural networks with residual connections yields best classifications accuracy for image classifications task for now. The proposed neural network is used to classify musical genres, based on pure waveform or mel-frequency cepstral coefficients, which are well known to be good sound representation for speech recognition task.

Ref. 11, fig. 3, tabl. 2.

Keywords — deep neural networks; convolutional networks; music information retrieval; classification.