한국지능정보시스템학회 [2021 추계학술대회] 논문 투고

논문 전문(Full Paper) 제출 희망 여부 (√표시)	예 [ ]	아니오 [√]
심사용 전문 제출 용도 (√표시) – 복수 선택 가능	우수논문 심사 [ ] 학술지 Fast Track 심사 [ ]	**해당사항 없음**

Music Genre Classification using CNN architecture with feature concatenation

문아성 (중앙대학교 소프트웨어대학 AI 학과 석사 과정, 주저자 aseong002@gmail.com)

조성현 (중앙대학교 소프트웨어대학 AI 학과 석사 과정 saintcho94@gmail.com)

이재성 (중앙대학교 소프트웨어대학 AI 학과 부교수, 교신저자 curseor@cau.ac.kr)

[저자 연락처]

◎ 문아성 (010-4544-1320)

주소: 서울특별시 동작구 흑석동 흑석로 84 중앙대학교 소프트웨어대학 AI 학과

◎ 조성현 (010-3121-2331)

주소: 서울특별시 동작구 흑석동 흑석로 84 중앙대학교 소프트웨어대학 AI 학과

◎ 이재성 (010-2511-0100)

주소: 서울특별시 동작구 흑석동 흑석로 84 중앙대학교 소프트웨어대학 AI 학과

Music Genre Classification using CNN architecture with feature concatenation

문아성
중앙대학교 AI 학과
aseong002@gmail.com

조성현
중앙대학교 AI 학과
saintcho94@gmail.com

이재성
중앙대학교 AI 학과
curseor@cau.ac.kr

Abstract – Music Genre Classification has received increasing attention in recent years. In this field, the convolutional neural network is the most used algorithm as its basis. In most previous studies that conducted experiments by converting the existing music signal into an image, only one Mel Spectrogram and Mel-frequency Cepstral Coefficient image were used as input data. However, two types of images can be used as inputs to improve classification accuracy. Therefore, this study proposes a new deep learning architecture with heterogeneous images for music genre classification. The proposed model and comparison algorithms were tested using six popular datasets of music genres. Our proposed method can classify genres more effectively than other widely used convolutional neural network models.

Key Terms – Convolutional Neural Network, Music Genre Classification, Mel Spectrogram, Mel-frequency Cepstral Coefficient

Acknowledgment

This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (2021-0-01341, Artificial Intelligence Graduate School Program(Chung-Ang University))

I. Introduction

Deep learning, a field of machine learning, has recently received much attention as it has achieved breakthrough results in various fields. In particular, deep learning models using Convolutional Neural Networks (CNN) have realized performance beyond humans in computer vision, and deep learning research continues in various other fields [1]. As various devices have become common, music information retrieval technology is also in the spotlight. This technology has various detailed research topics such as sound source separation, automatic tagging, and melody extraction.

As the input of music genre classification significantly influences the learning algorithm's accuracy, various types of input can be considered according to the application. For example, there is an audio spectrogram using Music Information Retrieval (Dieleman et al., 2011; Choi et al., 2016a), signal data using time information of music, and Short-Time Fourier Transform [2], Mel Spectrogram[2], and Mel-Frequency Cepstral Coefficient (MFCC)[2] created by extracting unique features of voice. Previous studies performed music genre classification using only one type of function. Therefore, we propose a music genre classification model using heterogeneous two images in this study.

In this study, a deep learning model was designed by generating Mel Spectrogram and MFCC images with different characteristics as input data using music data. In addition, the proposed method compared the classification performance with four popular CNN models (DenseNet, MobileNet, ResNet, ShuffleNet-v2) and improved classification accuracy.

II. Proposed Method

This study proposes a deep learning model using Mel Spectrogram and MFCC images, with different characteristics as inputs. In this study, the spectrogram images of music were used as input data of CNN. The CNN

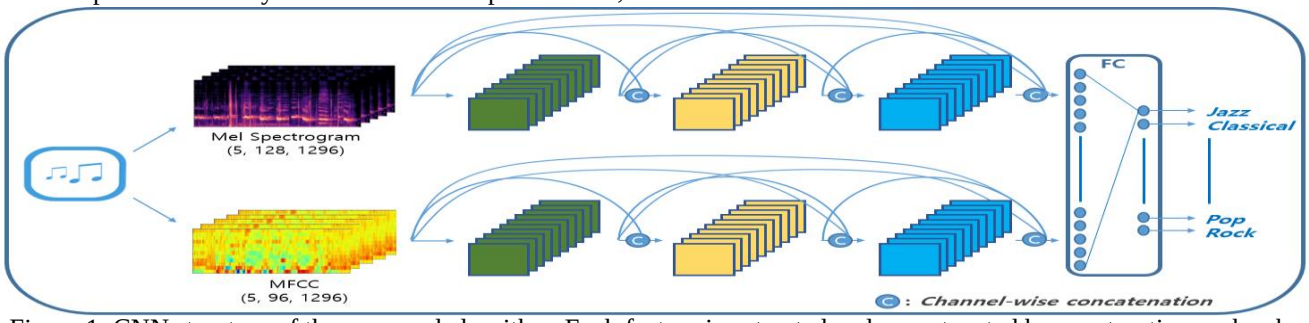

Figure 1: CNN structure of the proposed algorithm. Each feature is extracted and concatenated by constructing each sub-network with Mel spectrogram and MFCC images to classify music genres using a fully connected layer.

of the proposed model uses the concept of dense blocks of DenseNet. In the dense block, all feature maps have the same size, and they are all added to the output value of the dense block. This allows the learning to continue while keeping the initial information at a meaningful level. In addition, the spectrogram image has a grayscale value, and the depthwise convolution is performed by dividing the time axis and superimposing it in the channel-wise. Therefore, it is possible to extract features while maintaining the initial information in each time section.

After passing through each CNN model, we aim to combine two images, Mel Spectrogram and MFCC. Learning the two types of images separately can capture common patterns, which have good data representations for classification. Therefore, we separately obtain an internal data representation for every item after training them on the genre classification task from each sub-network described in Figure 1. The internal data obtained in this way were combined and mapped into the same vector space. Then these are then used as input to the final classification layer.

III. Experiment Results

We conducted experiments using six music genre datasets. Table 1 indicates statistics for the six genre domain datasets used in our experiments. It includes the number of patterns |W|, number of labels |L|, and Avg. Size of Tags represents the average number of patterns per tag and its standard deviation, respectively.

Table 1: Statistics of music datasets

Datasets	|W|	|L|	Avg. Size of Tags
GTZAN [3]	1,000	10	100.00 ± 0.00
Ballroom [4]	698	8	87.25 ± 17.56
Ballroom-Extended [5]	4,180	13	321.54 ± 195.70
Tropical Genres [6]	1,500	5	300.00 ± 0.00
FMA-SMALL [7]	7,996	8	999.50 ± 0.50
MICM [8]	2,187	6	121.17 ± 95.60

First, a 30-second long Mel Spectrogram and MFCC array were generated for six datasets. Then, the generated array is converted into input image data of five channels by dividing them into 0-10 sec, 5-15 sec, 10-20 sec, 15-25 sec, and 20-30 sec, respectively. Zero padding is used at the end of music under 30 seconds to match length.

Our experiment was implemented using the Pytorch framework. Adam algorithm was used to optimize the weights of the network while minimizing the binary cross-entropy loss. The learning rate is initially set to 0.001. Epoch was set to 60, and the epoch with the least loss was selected as the final model, and the performance was measured by dividing train and test set by 8:2. The experiment was conducted 30 times in total, and The Accuracy is used as the average of 30 times as the experimental result metric.

Table 2 shows the experimental results of the proposed model and four comparative models for six datasets. We proved that the model using two types of inputs, Mel Spectrogram and MFCC, showed better results than the comparative model using only Mel Spectrogram.

Table 2: Comparison results of five classification methods in terms of Accuracy ± standard deviation

Datasets	Proposed	ResNet	MobileNet	ShuffleNet
GTZAN	$\mathbf{86.4 \pm 2.14}$	$\mathbf{70.4 \pm 3.02}$	$\mathbf{63.6 \pm 2.58}$	$\mathbf{63.8 \pm 2.53}$
Ballroom	$\mathbf{70.0 \pm 3.70}$	$\mathbf{57.5 \pm 4.07}$	$\mathbf{53.1 \pm 3.89}$	$\mathbf{48.7 \pm 3.40}$
Ballroom-Extended	$\mathbf{81.7 \pm 1.55}$	$\mathbf{68.3 \pm 3.51}$	$\mathbf{77.9 \pm 1.51}$	$\mathbf{61.1 \pm 2.13}$
Tropical Genres	$\mathbf{97.9 \pm 0.82}$	$\mathbf{94.8 \pm 2.80}$	$\mathbf{89.3 \pm 1.49}$	$\mathbf{87.3 \pm 2.49}$
FMA-SMALL	$\mathbf{56.3 \pm 0.99}$	$\mathbf{50.3 \pm 1.79}$	$\mathbf{49.7 \pm 0.95}$	$\mathbf{47.6 \pm 1.12}$
MICM	$\mathbf{40.0 \pm 2.85}$	$\mathbf{35.6 \pm 3.54}$	$\mathbf{38.4 \pm 3.46}$	$\mathbf{38.0 \pm 3.40}$

VI. References

[1] Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017, August). Understanding of a convolutional neural network. In 2017 International Conference on Engineering and Technology (ICET) (pp. 1-6). Ieee.

[2] Bahuleyan, Hareesh. "Music genre classification using machine learning techniques." arXiv preprint arXiv:1804.01149 (2018).

[3] G. Tzanetakis and P. Cook, “Musical genre classification of audio signals,” IEEE Transactions on speech and audio processing, vol. 10, no. 5, pp. 293–302, 2002.

[4] F. Gouyon, S. Dixon, E. Pampalk, and G. Widmer, “Evaluating rhythmic descriptors for musical genre classification,” in Proceedings of the AES 25th International Conference, vol. 196, 2004, p. 204.

[5] U. Marchand and G. Peeters, “The extended ballroom dataset,” 2016.

[6] C. Salazar, “Tropical genres dataset.” Kaggle, 2020.

[7] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, “Fma: A dataset for music analysis,” arXiv preprint arXiv:1612.01840, 2016.

[8] ——, “Micm music dataset.” Kaggle, 2018. [Online]. Available: https://www.kaggle.com/dsv/193325