Music Genre Classification Using Multiple Musical Features

Sung-Hyun Cho, Jaesung Lee
Department of Artificial Intelligence, Chung-Ang University, 221, Heukseok-Dong, Dongjak-Gu, Seoul 06974, Republic of Korea
e-mail : saintcho94 at gmail.com, jslee.cau at gmail.com

Abstract

Various studies in deep learning have accelerated the development of many classification tasks, including music genre classification. F Retrieval researchers have proposed various structures based on Convolutional Neural Networks that mainly achieve state-of-the-art results in the music genre classification tasks. Using multiple musical features as model inputs can improve classification accuracy. Therefore, this study proposes a new Convolutional Neural Network model using three musical features for music genre classification: Short-Time Fourier Transform, Mel-Spectrogram, and Mel-Frequency Cepstral Coefficient. In addition, to compensate for the small amount of learning data in music datasets, we increase learning efficiency by stacking 30-second datasets with 10-second overlapping by 5-second.

Key Terms – Acoustic Features, Convolutional Neural Network, Deep Learning, Mel-Frequency Cepstral Coefficient, Mel-Spectrogram, Music Genre Classification, Music Information Retrieval, Short-Time Fourier Transform

I. Introduction

As algorithms improve and high-quality data and computing power increase, research in deep learning has accelerated growth in various classification tasks, including music genre classification, urban sound classification[1], music mood classification, and music instrument classification.

The Convolutional Neural Network (CNN), which utilizes images, allowed music, which was considered a representative of voice signals, to be converted into images through preprocessing to achieve high classification accuracy. Typical converted images include Short-Time Fourier Transform (STFT), mel-spectrogram, Mel-Frequency Cepstral Coefficient (MFCC), rolloff, and chromagram[2]. Music Information Retrieval (MIR) researchers continued to study the classification of genres, moods, and instruments with high accuracy using these musical features. This music classification can help music listeners choose music that suits their tastes.

However, many conventional studies have tried to improve classification accuracy by using only one musical feature, even though the various musical features mentioned above can be used. In addition,

most of the studies showed good performance through the results of one or two music datasets. This study proposes a novel algorithm that extracts three musical features from seven music datasets: STFT, mel-spectrogram, and MFCC, and uses them all as input in the proposed method. Seven popular music genre datasets undergo fair evaluation regarding classification accuracy, and the proposed model shows superior accuracy to the comparison algorithms. This study shows that higher classification accuracy can be obtained if multiple musical features are simultaneously used as input to a neural network.

II. Related Work

Ghildiyal et al. built a four-layer neural network model to extract musical features from the GTZAN dataset[3][4]. Mel-spectrogram was used to compare the result with Artificial Neural Network, Support Vector Machine, Multi Layer Perceptron, and Decision Tree and achieved 91% classification accuracy. Palanisamy et al. conducted a study using three music datasets(ESC-50[5], UrbanSound8k[6], and GTZAN)[7]. By experimenting with the pretrained CNN model from ImageNet, Palanisamy found that there are various performance differences in the classification result in the same model even with the same pretrained weight. The two reasons that the paper suggested are because of random initialization of linear classification and random mini-batch orderings. Also, they argued that the GTZAN dataset could not achieve a classification accuracy of more than 94.5% due to its noise. Bidirectional Recurrent Neural Network(BRNN) was proposed by Yu et al.[8]. They compared serial and parallelized attention models in GTZAN and Ballroom-Extended[9] datasets. Due to the lack of the number music dataset, they cut each excerpt into 18 smaller clips which last 3 seconds respectively, with 50% overlaps. In the study, the BRNN model with parallelized CNN attention achieves 92.7% of classification accuracy. In the study of [10], three music datasets (MagnaTagATune[11], Million Song Dataset[12], and MTG-Jamendo[13]) were transformed into mel-spectrogram with a 50% overlap. The study showed that the model on short audio chunks outperforms the model on full songs. That is because an audio excerpt can have a tag guitar if a guitar appears in the song even though the selected excerpt does not include guitar sound in it. Bisharad and Laskar mainly compared similarities and differences between genres from the GTZAN dataset [14]. The mel-spectrogram was computed using a sliding Hamming window of 20 milliseconds with 50% overlap. For example, disco was mainly confused with blues and pop with disco. However, hip-hop and disco are least confused with any other genres. Zhang et al. proved that when using Recurrent Neural Network, the classification accuracy can achieve up to 64% only by using a short excerpt of 0.5 seconds [15]. They insist that it is because MFCC does not change much in time.

Ⅲ. Materials and Methods

3.1 Materials

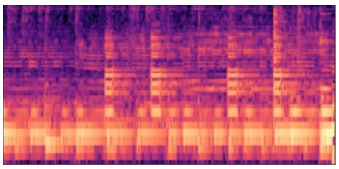
Table 1. Statistics of music datasets

Datasets	|W|	|L|	Avg. Size of Tags
Ballroom	698	8	87.25 ± 17.56
Ballroom-Extended	4,180	13	321.54 ± 195.70
GTZAN	1,000	10	100.00 ± 0.00
HOMBURG	1,886	9	209.56 ± 134.87
ISMIR04	2,187	6	121.17 ± 95.60
MICM	2,187	6	162.43 ± 119.72
Seyerlehner-Unique	3,115	14	311.50 ± 248.35

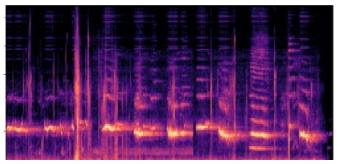
Table 1 briefly provides descriptions of seven music genre domain datasets used in this study. According to Table 1, the first column shows the name of seven music genre datasets: Ballroom[16], Ballroom-Extended, GTZAN, HOMBURG[17], ISMIR04[18], MICM[19], and Seyerlehner-Unique[20]. The terms |W| and |L| indicates the number of

patterns of each dataset and the number of labels of each dataset. The last column represents the average number of patterns per label and its standard deviation. Datasets such as GTZAN and ISMIR04 are popularly used in the studies of MIR and add credibility to this study.

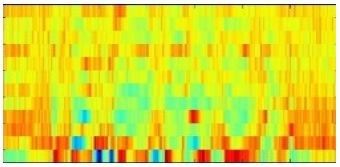
Figure 1. Extracted musical features



(a) STFT



(b) Mel-Spectrogram



(c) MFCC

3.2 Preprocessing

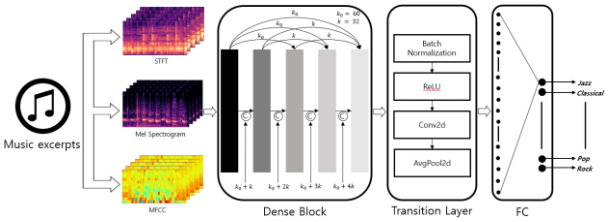
The excerpts of the datasets are loaded and converted into each musical features using *Librosa* library[21]. The example of extracted features are in Figure 1. In order to unify all the excerpts to a length of 30 seconds, an excerpt shorter than 30 seconds went through zero padding, and an excerpt longer than 30 seconds was cut off after 30 seconds. Than the features were splitted into five 10-second slices with 50% overlap: 0-10 sec., 5-15 sec., 10-20 sec., 15-25 sed., and 20-30 sec., respectively. Through this preprocessing, the amount of small music datasets could be easily increased for training and testing.

3.3 Methods

As in Figure 2, our proposed model has been adjusted so that the existing DenseNet[22] can receive STFT, mel-spectrogram, and MFCC as inputs at the same time. DenseNet attaches feature maps of all layers. Unlike ResNet[23], concatenations are performed instead of adding. Therefore, when concatenating, the feature map size must be the same. This allows to use all the information from previous layers. As a result, less number of parameters are required.

Training each of the three features makes it easier to capture general patterns in each genre. Capturing general patterns can yield better results when classifying music genres. Therefore, we can gather internal data representations for each item separately. The gathered internal data were combined and mapped into the same vector space. Finally, these data are used as input to the final classification layer.

Figure 2. Proposed Convolutional Neural Network Architecture



Our experiment was implemented using the Pytorch framework[24]. Adam optimizer[25] and cross-entropy[26] was used for loss function. All the experiments were fairly done with the same epochs, iterations, learning rate, batch size, and the ratio of the train-test set. The models were trained for 60 epochs, 10 iterations, 0.001 learning rate, batch size of 32, and the performance was measured by dividing the training and testing set by 8:2. The accuracy is the average and standard deviation of the 10 iteration results, and the accuracy ranking for each dataset was obtained, and the

final ranking was displayed in Table 2.

Ⅳ. Experimental Results

Proposed model outperforms four comparative models in Table 2. In five datasets out of seven datasets, the classification accuracy of proposed model was higher than other conventional algorithms. We proved that inputting various musical features can get better results.

Table 2: Comparison results of four classification methods in terms of accuracy ± standard deviation

Dataset	Proposed	[1]	[2]	[3]
Ballroom	57.36±3.96	48.86±4.41	39.29±3.82	**60.36±2.84**
Ballroom−Extended	**85.32±0.61**	75.89±3.09	39.07±1.49	−
GTZAN	**87.70±3.01**	74.50±3.78	58.85±3.79	78.20±1.62
HOMBURG	54.13±1.91	54.50±2.93	47.14±1.89	**55.00±3.51**
ISMIR04	**80.07±1.64**	68.36±1.61	62.28±2.42	70.84±1.83
MICM	**49.39±1.79**	39.56±2.53	39.12±2.78	**40.66 ± 3.50**
Seyerlehner−Unique	**71.03±1.49**	63.88±1.53	55.59±3.45	−
Avg. Rank	1.43	2.43	3.57	2.14

Ⅴ. Conclusions

In this work, we experimented using seven popular music genre datasets. Three musical features (STFT, mel−spectrogram, MFCC) were extracted from each dataset. The three datasets supplement the sparse part of the input values in training, improving classification accuracy. Experimental results showed superiority in classification accuracy of five datasets out of seven compared to other CNN classification models.

We can think of two things in future work. First, it will be possible to devise a model that can learn by better selecting the characteristics of the three inputs. Second, creating a model that reduces learning time will be possible.

References

[1] M. Massoudi, S. Verma and R. Jain, "Urban Sound Classification using CNN," 2021 6th International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2021, pp. 583−589.

[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] Ghildiyal, A., Singh, K., & Sharma, S. (2020, November). Music genre classification using machine learning. In 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 1368−1372). IEEE.

[5] PICZAK, Karol J. ESC: Dataset for environmental sound classification. In: Proceedings of the 23rd ACM international conference on Multimedia. 2015. p. 1015−1018.

[6] SALAMON, Justin; JACOBY, Christopher; BELLO, Juan Pablo. A dataset and taxonomy for urban sound research. In: Proceedings of the 22nd ACM international conference on Multimedia. 2014. p. 1041−1044.

[7] Palanisamy, K., Singhania, D., & Yao, A. (2020). Rethinking cnnmodels for audio classification. arXivpreprint arXiv:2007.11154.

[8] Yu, Y., Luo, S., Liu, S., Qiao, H., Liu, Y., & Feng, L. (2020). Deep attention basedmusic genre classification. Neurocomputing, 372, 84-91.

[9] U. Marchand and G. Peeters, "The extended ballroom dataset," 2016.

[10] Won, M., Ferraro, A., Bogdanov, D., & Serra, X. (2020). Evaluation of cnn-based automatic music tagging models. arXivpreprint arXiv:2006.00751.

[11] LAW, Edith, et al. Evaluation of algorithms using games: The case of music tagging. In: ISMIR. 2009. p. 387-392.

[12] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The Million Song Dataset. In Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011), 2011.

[13] Bogdanov, D., Won M., Tovstogan P., Porter A., & Serra X. (2019). The MTG-Jamendo Dataset for Automatic Music Tagging. Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML 2019).

[14] Bisharad, D., & Laskar, R. H. (2019, October). Music Genre Recognition Using Residual Neural Networks. In TENCON 2019-2019 IEEE Region 10 Conference (TENCON) (pp. 2063-2068). IEEE.

[15] Zhang, S., Gu, H., & Li, R. (2019). MUSIC GENRE CLASSIFICATION: NEAR-REALTIME VS SEQUENTIAL APPROACH.

[16] 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.

[17] H. Homburg, I. Mierswa, B. Möller, K. Morik, and M. Wurst, "A benchmark dataset for audio classification and clustering." in ISMIR, vol. 2005, 2005, pp. 528-31.

[18] P. Cano, E. Gómez, F. Gouyon, P. Herrera, M. Koppenberger, B. Ong, X. Serra, S. Streich, and N. Wack, "Ismir 2004 audio description contest," Music Technology Group of the Universitat Pompeu Fabra, Tech. Rep, 2006.

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

[20] K. Seyerlehner, G. Widmer, and T. Pohle, "Fusing block-level features for music similarity estimation," in Proc. of the 13th Int. Conference on Digital Audio Effects (DAFx-10), 2010, pp. 225-232.

[21] McFee, Brian, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. "librosa: Audio and music signal analysis in python." In Proceedings of the 14th python in science conference, pp. 18-25. 2015.

[22] HUANG, Gao, et al. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p. 4700-4708.

[23] HE, Kaiming, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 770-778.

[24] PASZKE, Adam, et al. Automatic differentiation in pytorch. 2017.

[25] KINGMA, Diederik P.; BA, Jimmy. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[26] D. R. Cox. "The Regression Analysis of Binary Sequences" Journal of the Royal Statistical Society. Series B (Methodological) Vol. 20, No. 2 (1958), pp. 215-242.