Effective Music Genre Classification using Late Fusion Convolutional Neural Network with Multiple Spectral Features

Sung-Hyun Cho[1], Yechan Park[2], Jaesung Lee[1]
[1]Department of Artificial Intelligence, [2]Department of Public Service,
Chung-Ang University.
Seoul, Republic of Korea
saintcho94@gmail.com, qkrd1285@cau.ac.kr, curseor@cau.ac.kr

Abstract

Music genre classification is getting more and more attention amid the growing content consumption for music. Music Information 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.

Keywords: Convolutional Neural Network, Mel-Frequency Cepstral Coefficient, Mel-Spectrogram, Music Genre Classification, Music Information Retrieval, Short-Time Fourier Transform

1. Introduction

Music Information Retrieval covers three broad categories: music classification, music manipulation, and music creation. Deep learning studies are actively conducted in various classification tasks [1], such as urban sound classification [2], musical instrument classification [3], music mood classification [4], and Music Genre Classification (MGC) [5]. These sound and music classifications are closely related to daily life and are a steadily growing research area. MGC can provide a high-satisfaction music listening environment for users to suit their tastes.

Deep neural networks show approximately 3% performance improvement over traditional classifier techniques by constructing multiple layers of neural networks [6]. Convolutional Neural Network (CNN), one of the representative networks in deep learning, achieved high classification accuracy by converting music, considered a representative of voice signals, into images through preprocessing.

Common transformed images include Short-Term Fourier Transform(STFT) [7], mel-spectrogram [8], Mel-Frequency Cepstral Coefficient(MFCC) [9], tempo and chromagram [10] [11] and so on. Each music feature has its characteristics, some of which play an important role in MGC. The neural network can reduce the training time and obtain meaningful results using appropriate features.

However, many conventional studies have attempted to improve accuracy by using only one musical feature, although the above-mentioned various musical features can be used [12-14]. Using only one feature may cause inaccurate classification due to the insufficient number of samples in the music data. Also, it was not easy to gather a good-quality music dataset in the past. As datasets become easier to collect and their quality improves, various deep learning models such as ResNet [15], DenseNet [16], and AlexNet [17] have emerged. Due to these deep learning models, the performance of MGC has also improved.

2. Related Work

Bisharad and Laskar mainly compared similarities and differences between genres in the GTZAN dataset [32] [18]. According to the experimental results, disco was mainly confused with blues and pop. However, hiphop and disco were the least confused with other genres. Mel-spectrograms were calculated using a sliding Hanning window of 20 milliseconds with 50% overlap. Zhang et al. demonstrated that classification accuracy of up to 64% could be achieved with only a short excerpt of 0.5 seconds when using a recurrent neural network [19]. The reason for achieving compliance classification accuracy with only an abridged excerpt is that MFCC does not change much time.

Ghildial designed a four-layer neural network model to extract musical features from the GTZAN dataset [12]. Results were compared with artificial neural networks, SVMs, multi-layer perceptrons, and DTs using mel-spectrogram and achieved 91% classification accuracy. Palanisamy et al. conducted a study using three musical datasets (ESC-50 [20], UrbanSound8k [21], and GTZAN) [13]. They experimented with ImageNet's pre-trained CNN model and found that even the same pretrained weights had varying performance differences in classification results in the same model. Three music datasets (MagnaTagAtune [22], Million Song Dataset [23], and MTG-Jamendo [24]) were converted to mel-spectrograms with 50% overlap in the study [14]. This study showed that the model using short audio chunks outperforms the entire song. Even though the selected excerpt does not contain a guitar sound, audio excerpts can have guitar tags if the guitar appears in the song.

3. Materials and Methods

We first collected 12 music datasets to demonstrate the performance under various circumstances. The eight datasets are composed initially of genre domains. The remaining four datasets correspond to different domains. However, through the metadata provided by the creator, we converted all datasets into the genre domains, and at the end, we added -G, which means genre.

Table 1 briefly provides descriptions of 12 datasets used in this study. According to Table 1, the first column shows the name of 12 music genre datasets: Ballroom[25], Ballroom-Extended[26], emoMusic-G[27], Emotify-G[28], FMA-SMALL[29], GiantStepsKey-G[30], GMD-G[31], GTZAN, HOMBURG[33], ISMIR04[34], MICM[35], and Seyerlehner-Unique[36]. The terms |W|, |T|, and Avg. Size of Tag indicates the number of patterns, the number of tags, and the average number of patterns per tag and its standard deviation, respectively.

Table 1: The standard characteristics and statistics of employed datasets

Dataset	|W|	|T|	Avg. Size of Tag
Ballroom	698	8	87.25 ± 17.56
Ballroom-Extended	4,180	13	321.54 ± 195.70
emoMusic-G	1,000	8	125.00 ± 0.00
Emotify-G	400	4	100.00 ± 0.00
FMA-SMALL	7,996	8	999.50 ± 0.50
GiantStepsKey-G	604	23	24.9130 ± 28.0274
GMD-G	1,042	18	57.89 ± 70.07
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	1,137	7	162.43 ± 119.72
Seyerlehner:Unique	3,115	14	311.50 ± 248.35

After collecting all the music datasets, we converted the extension to .wav. We used the python library librosa to extract three features for the experiments: STFT, mel-spectrogram, and MFCC. To unify all the excerpts to a length of 30 seconds, audios shorter than 30 seconds went through zero padding, and longer audios were cut off after 30 seconds. First, extracted features went through Rectified Linear Unit (ReLU) [37]. ReLU delivers the input value to the output for numbers greater than zero and outputs zero for numbers less than zero regardless of input. Next, we applied a 2D adaptive average pooling over an input signal composed of several input planes for the output of the ReLU function. Then, we flattened the previous output. After going through all these processes, we concatenated three outcomes. Every end of dense layer has a dropout layer with a probability of 0.3 during training. The dropout layer prevents overfitting. Figure 1 provides the architecture of our proposed model.

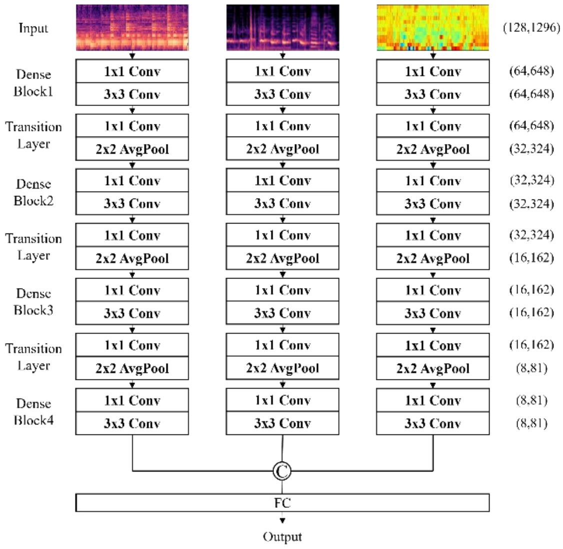

Figure 1: Proposed neural network architecture

4. Experimental Results

For the evaluation metric, we adopted an accuracy measure in the range of [0,100]. The accuracy is calculated as:

$$Accuracy\ (\%) = \frac{TP + TN}{TP + TN + FP + FN} * 100$$

where TP, TN, FP, and FN stand for True Positives, True Negatives, False Positives, and False Negatives.

We compared the proposed method to four conventional MGC models in terms of accuracy. The four comparison models are as follows:

- Li et al. proposed a new loss function to avoid overfitting [38]. They first divided every 30-second long GTZAN dataset excerpt into six samples with a length of five seconds. While importing the conventional ResNet-50, they proposed a new loss function to prevent overfitting by adding two newly defined loss values.
- Pelchat and Gelowitz proposed the CNN model with six convolutional layers [39]. The study transformed three-minute stereo channel songs into mono channel songs. Next, they extracted spectrograms from the mono-channel songs before dividing the original excerpts into 70 slices. Each music slice was only 2.56 seconds long and recorded 85% of test accuracy.

• Cheng et al. designed a new CNN model using dropout on every convolutional layer [40]. They spent a considerable amount of time experimenting. Experimenting for 13.1 hours, they achieved 83.30% accuracy using the GTZAN dataset. The total iteration of the experiment was 2,180, and the epoch was 100.

• Shah et al. compared their model with three conventional classification algorithms [41]. Two different approaches were used for this experiment. The first approach manually extracted the time and frequency domain features and provided them to traditional classification algorithms. The second approach extracts spectrogram images from audio files and gives them to CNNs. Finally, the results of the two methods were evaluated, achieving 74.1% of the proposed model.

Table 2 shows the overall results between the proposed model and comparison models. Of all 12 datasets, our proposed model ranked first with an average ranking of 1.00. Comparing the proposed model to other models, the difference in classification accuracy of the GMD-G dataset was as high as 30.14%. In addition, eight out of 12 datasets show a difference of more than 10% from the model of second place. The proposed model is significantly superior to other comparable models. On the other hand, the MICM dataset showed only a 3.37% difference between the proposed model and the one with the lowest accuracy.

Table 2: The standard characteristics and statistics of employed datasets

	Ballroom	**Ballroom-Extended**	**emoMusic-G**
Proposed	**55.86% ± 3.41**	70.35% ± 3.25	35.77% ± 2.49
[38]	**45.71% ± 2.86**	61.82% ± 3.79	27.72% ± 3.50
[39]	**42.43% ± 2.29**	44.68% ± 1.30	28.99% ± 1.78
[40]	**48.07% ± 4.79**	62.88% ± 4.70	32.95% ± 2.92
[41]	**47.71% ± 4.58**	54.38% ± 5.33	31.34% ± 3.66
	Emotify-G	**FMA-SMALL**	**GiantSteps Key-G**
Proposed	**72.00% ± 4.05**	55.07% ± 0.84	35.29% ± 2.67
[38]	**62.78% ± 2.70**	47.03% ± 1.31	28.43% ± 3.15
[39]	**67.50% ± 4.17**	42.19% ± 1.16	25.21% ± 2.53
[40]	**66.00% ± 3.32**	46.96% ± 1.62	27.19% ± 2.90
[41]	**65.88% ± 5.62**	40.88% ± 1.73	24.30% ± 2.11
	GMD-G	**GTZAN**	**HOMBURG**
Proposed	**65.50% ± 3.55**	79.00% ± 2.94	55.53% ± 2.39
[38]	**61.91% ± 3.04**	77.50% ± 3.81	53.49% ± 1.59
[39]	**63.44% ± 1.71**	63.35% ± 2.08	51.01% ± 1.51
[40]	**39.38% ± 3.36**	75.10% ± 3.84	52.54% ± 2.55
[41]	**35.36% ± 6.35**	72.50% ± 2.13	48.15% ± 2.14
	ISMIR04	**MICM**	**Seyerlehner: Unique**
Proposed	**79.13% ± 2.39**	41.18% ± 2.94	70.85% ± 2.36
[38]	**70.96% ± 2.93**	37.81% ± 3.32	63.62% ± 2.56
[39]	**67.88% ± 1.70**	39.30% ± 2.21	60.72% ± 2.73
[40]	**71.32% ± 1.37**	38.42% ± 3.16	63.50% ± 2.79
[41]	**67.67% ± 1.99**	39.34% ± 2.28	61.62% ± 1.70
	Avg. Rank		
Proposed	**1.00**		
[38]	3.17		
[39]	3.92		
[40]	2.83		
[41]	4.08		

5. Conclusions

MIR is getting more attention in real life and academically. Users want to listen to music that suits their taste, and the importance of MGC increases to solve these problems. In this study, we performed MGC using 12 datasets. Experiments were conducted by extracting STFT, mel-spectrogram, and MFCC for each music clip through preprocessing from 12 datasets. The three musical features can complement each other and achieve higher classification accuracy when training CNN models using one feature. The proposed model ranked first in performance on all 12 datasets, with an overall average accuracy of 1.00.

Acknowledge

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the National Program for Excellence in SW(20170001000061001) supervised by the IITP(Institute of Information & communications Technology Planning&Evaluation) in 2022

References

[1] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, "A survey of audio-based music classification and annotation," IEEE transactions on multimedia, vol. 13, no. 2, pp. 303–319, 2010.

[2] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in 2015 IEEE 25th international workshop on machine learning for signal processing (MLSP), pp. 1–6, IEEE, 2015.

[3] C. Joder, S. Essid, and G. Richard, "Temporal integration for audio classification with application to musical instrument classification," IEEE Transactions on Audio, Speech, and Language Processing, vol. 17, no. 1, pp. 174–186, 2009.

[4] R. Delbouys, R. Hennequin, F. Piccoli, J. Royo-Letelier, and M. Moussallam, " Music mood detection based on audio and lyrics with deep neural net," arXiv preprint arXiv:1809.07276, 2018.

[5] D. S. Lau and R. Ajoodha, "Music genre classification: A comparative study between deep learning and traditional machine learning approaches," in Proceedings of Sixth International Congress on Information and Communication Technology, pp. 239–247, Springer, 2022.

[6] M. Alam, K.-C. Kwon, M. Y. Abbass, S. M. Imtiaz, N. Kim, et al., "Trajectory-based air-writing recognition using deep neural network and depth sensor," Sensors, vol. 20, no. 2, p. 376, 2020.

[7] J. Allen, "Short term spectral analysis, synthesis, and modification by discrete fourier transform," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 25, no. 3, pp. 235–238, 1977.

[8] H. Meng, T. Yan, F. Yuan, and H. Wei, “Speech emotion recognition from 3d log-mel spectrograms with deep learning network,” IEEE access, vol. 7, pp. 125868–125881, 2019.

[9] B. Logan, “Mel frequency cepstral coefficients for music modeling,” in In International Symposium on Music Information Retrieval, Citeseer, 2000.

[10] S. Pauws, “Musical key extraction from audio.,” in ISMIR, 2004.

[11] H. Bahuleyan, “Music genre classification using machine learning techniques,”
arXiv preprint arXiv:1804.01149, 2018.

[12] A. Ghildiyal, K. Singh, and S. Sharma, “Music genre classification using machine learning,” in 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 1368–1372, IEEE, 2020.

[13] K. Palanisamy, D. Singhania, and A. Yao, “Rethinking cnn models for audio classification,” arXiv preprint arXiv:2007.11154, 2020.

[14] M.Won, A. Ferraro, D. Bogdanov, and X. Serra, “Evaluation of cnn-based automatic music tagging models,” arXiv preprint arXiv:2006.00751, 2020.

[15] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778, 2016.

[16] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708, 2017.

[17] A. Krizhevsky, “One weird trick for parallelizing convolutional neural networks,” arXiv preprint arXiv:1404.5997, 2014.

[18] D. Bisharad and R. H. Laskar, “Music genre recognition using residual neural networks,” in TENCON 2019-2019 IEEE Region 10 Conference (TENCON), pp. 2063–2068, IEEE, 2019.

[19] S. Zhang, H. Gu, and R. Li, “Music genre classification: Near-realtime vs sequential approach,” PrePrint, 2019.

[20] K. J. Piczak, “Esc: Dataset for environmental sound classification,” in Proceedings of the 23rd ACM international conference on Multimedia, pp. 1015–1018, 2015.

[21] J. Salamon, C. Jacoby, and J. P. Bello, “A dataset and taxonomy for urban sound research,” in Proceedings of the 22nd ACM international conference on Multimedia, pp. 1041–1044, 2014.

[22] E. Law, K. West, M. I. Mandel, M. Bay, and J. S. Downie, “Evaluation of algorithms using games: The case of music tagging.,” in ISMIR, pp. 387–392, 2009.

[23] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere, “The million song dataset,” 2011.

[24] D. Bogdanov, M. Won, P. Tovstogan, A. Porter, and X. Serra, “The mtgjamendo dataset for automatic music tagging,” 2019.

[25] 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.

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

[27] M. Soleymani, M. N. Caro, E. M. Schmidt, C.-Y. Sha, and Y.-H. Yang, “1000 songs for emotional analysis of music,” in Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia, 2013, pp. 1–6.

[28] A. Aljanaki, F. Wiering, and R. C. Veltkamp, “Studying emotion induced by music through a crowdsourcing game,” Information Processing & Management, vol. 52, no. 1, pp. 115–128, 2016.

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

[30] P. Knees, Á. Faraldo Pérez, H. Boyer, R. Vogl, S. Böck, F. Hörschläger, M. Le Goff et al., “Two data sets for tempo estimation and key detection in electronic dance music annotated from user corrections,” in Proceedings of the 16th International Society for Music Information Retrieval Conference; 2015 Oct 26-30; Málaga, Spain.[Málaga]: International Society for Music Information Retrieval, 2015. p. 364-70. International Society for Music Information Retrieval, 2015.

[31] J. Gillick, A. Roberts, J. Engel, D. Eck, and D. Bamman, “Learning to groove with inverse sequence transformations,” in International Conference on Machine Learning. PMLR, 2019, pp. 2269–2279.

[32] 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.

[33] 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.

[34] 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.

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

[36] 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.

[37] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in Icml, 2010.

[38] J. Li, L. Han, X. Li, J. Zhu, B. Yuan, and Z. Gou, “An evaluation of deep neural network models for music classification using spectrograms,” Multimedia Tools and Applications, vol. 81, no. 4, pp. 4621–4647, 2022.

[39] N. Pelchat and C. M. Gelowitz, “Neural network music genre classification,” Canadian Journal of Electrical and Computer Engineering, vol. 43, no. 3, pp. 170–173, 2020.

[40] Y.-H. Cheng, P.-C. Chang, and C.-N. Kuo, “Convolutional neural networks approach for music genre classification,” in 2020 International Symposium on Computer, Consumer and Control (IS3C), pp. 399–403, IEEE, 2020.

[41] M. Shah, N. Pujara, K. Mangaroliya, L. Gohil, T. Vyas, and S. Degadwala, “Music genre classification using deep learning,” in 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), pp. 974–978, IEEE, 2022.