Research on Architecture for Long-tailed Genre Computer Intelligent Classification with Music Information Retrieval and Deep Learning

Gang Sun*
School of Art, Huazhong University of Technology, Wuhan, China

*Corresponding author: sungang@hust.edu.com

Abstract. In this paper, we propose a Musical Attention Network (MAN) architecture for long-tailed, imbalanced music genre classification which is often ignored and quite prevalent in Music Information Retrieval (MIR). Here, the challenge is to classify the genre of long-tailed music accurately. Inspired by the recent progress in NLP, the proposed model can take advantage of genre correlations to better identify informative segments. Comprehensive experimental results demonstrate our model brings significant improvement comparing with other music genre classification models on a large-scaled benchmark dataset.

Keywords: Music Genre Classification; Music Information Retrieval; Neutral Networks.

1. Introduction
The Music Genre Classification (MGC) [1] is an essential task, which has a variety of applications in the field of Music Information Retrieval. In the previous research, many effects have been devoted to deal with the problem via statistic machine learning methods [2] [3], which required high-quality feature selection. Three kinds of audio features, song-level, frame-level and segment-level features were presented to recognize the music patterns including distribution(s), rhythmic information, pitch(es), tempo, melody, and so on. However, these methods, can’t deal with long-tailed music classification very well, because it’s very difficult to extract the precise feature of the tailed genre with only few music segments given.

Recently, deep learning models have been widely applied in Music Genre Classification tasks. However, their performance heavily depends on the scale and quality of training data. What's more, the current DNN models can’t handle with the problem of long-tail music genre classification: the DNN models’ performances degrade quickly when extracting long-tail genres, because many long-tail music genres suffer from data insufficiency. These difficulties lead the classification of long-tail music genre to be an extremely hard problem.

The long-tail music genre cannot be ignored for there is rich music information contained in them. What's more, the long-tail, imbalance data is quite common in real-world settings. Free Music Archive (FMA) [4] dataset is a widely used corpus for the classification of long-tail music genre. And nearly 82% of the genres in FMA are long-tailed as shown in Figure 1. Hence, the research of long-tail music
genre classification is significant, and the methods can extract long-tail genres with high accuracy are in great needs.

![Label frequency distribution of classes in FMA dataset.](image)

**Fig. 1.** Label frequency distribution of classes in FMA dataset.

Deal with the long-tail problems is extremely difficult for there are few examples for the data-hungry neural networks to leverage. Inspired by the achievements in NLP, we propose a novel musical attention network. Similar to the conventional attention-based method, our method also computes an attention score for each music segment according to its significance of expressing the corresponding genre.

Our experimental results on the FMA dataset show that: (1) our model is effective compared to baselines especially for long-tail music genres; (2) our models could take advantage of genre correlations to better identify informative segments, especially for those long-tail genres.

The rest of the paper is organized as follows. We will discuss the related work from the second section. The third section introduces our method, and the fourth section reports our experimental results. Finally, the fifth part summarizes and prospects the full text.

2. Related Works

The traditional supervised music genre classification models have achieved satisfactory results on balanced datasets. In 2009 Pham et al. [5] did a pioneering job, constructing a genre classifier with convolutional deep belief networks. Siddharth Sigtia et al. [6] adopted a rectifier convolutional neural network to learn the features which can extract relevant information from audio data. Simon Leglaiveve et al. [7] applied a bidirectional long-short term memory (BiLSTM) recurrent neural network (RNN) on singing voice detection. There are also some works mixing these models for better performance. Choi et al. [8] combined CNNs and RNNs, introducing convolutional recurrent neural networks (CRNNs). Suvra et al. proposed the clustering augmented learning method (CALM) [9] which is based on the concept of simultaneous heterogeneous clustering and classification to learn deep feature representations of the features obtained from LSTM autoencoder.

There are only a few studies on long-tail music classification tasks. J. Choi et al. [10] proposed the zero-shot learning to handle unseen labels such as newly added music genres or semantic words that users arbitrarily choose for music retrieval. S. Craw et al. [11] exploited the combined knowledge, from audio and tagging, using a hybrid representation that extends the track’s tag-based representation by adding semantic knowledge extracted from the tags of similar music tracks. However, the performance they can achieve is still not satisfactory. Besides, they are regardless the rich correlation of genres. Here we use attention mechanism to further improve the performance of long-tail music genre classification.
3. Methodology

3.1. Frameworks

The design framework of the long-tail musical classification model built in this paper is shown in Figure 2. The first step is to preprocess the music segment data set, and then encode music audio segments in the neural network. Then, the genre selective attention can output the prediction results.

3.1.1. Musical Segment Encoder. Given an instance and its mentioned entity pair, we employ deep neural networks to encode the music segments into vector.

3.1.2. Genre Selective Attention. Under the guidance of the final audio embeddings, the musical attention mechanism is aimed to select the most informative instance exactly matching relevant genre.

3.2. Musical Segment Encoder

3.2.1. Embedding Layer. The embedding layer is used to map the whole music segment to multiple continuous input embeddings. Multiframe strategy [12] is adopted to embed the raw audio data. We firstly convert original audio signal into log scale mel-spectrogram with 128 mel bands and 16000 sample rate and divide the original spectrogram into small chunks to create 128×128 feature maps. Each chunk stands for approximately 3s audio sequence. In addition to using log mel-spectrogram to preprocess the audio segments, we can also convert original audio signal into MFCC. We thus obtain an embedding sequence ready for the encoding layer. Figure 3 shows the spectrum after processing with log mel-spectrogram.

3.2.2. Encoding Layer

The encoding layer aims to compose the input embeddings of a given instance into its corresponding instance embedding. In this study, we choose the convolutional neural architectures, CNN, CRNN to encode input music segments into instance embeddings. Other neural architectures such as recurrent neural networks can also be used as audio encoders.
The convolutional neural network (CNN) slides a convolution kernel with the window size $m$ over the input sequence $\{x_1, \ldots, x_n\}$ to get the $k_h$-dimensional hidden embeddings. A max-pooling is then applied over these hidden embeddings to output the final instance embedding. The CNN structure is illustrated in Figure 4.

The CNN slides a convolutional kernel with window size $m$ over the input sequence $\{x_1, \ldots, x_n\}$ to get the $k_h$-dimensional hidden embeddings.

$$h_i = \text{CNN} \left( x_{i-\frac{m-1}{2}}, \ldots, x_{i+\frac{m-1}{2}} \right)$$

(1)

A max-pooling is then applied over these hidden embeddings to output the final instance embedding $s$ as follows,

$$[s]_j = \max_{i \leq i \leq n} \{h_i\}_j$$

(2)

The convolutional recurrent neural network (CRNN) is a hybrid model, which combines convolutional layers and GRU layers together. Gated recurrent unit (GRU) is a gating mechanism in RNN, which is similar with LSTM while has fewer parameters by eliminating the output gate. The CRNN structure is illustrated in Figure 5.

### 3.3. Genre Selective Attention

The genre selective attention scheme computes attention for each music segment $s_i$ to indicate how well the segment can match a specific genre. We assign a query vector $q_r$ to each genre $g \in G$, and attention for each sentence $S = \{s_1, s_2, \ldots, s_m\}$ is defined as follows,

$$e_i = q^TW_s s_i, \ldots$$

$$\alpha_i = \frac{\exp(e_i)}{\sum_{j=1}^{m} \exp(e_j)}$$

(3)

(4)

Where $W_s$ is the weight matrix. The attention scores can be calculated as follows:
By taking \( g \) as the music segment representation of the same author, we define the conditional probability \( P(r|h, t, S_{h,t}) \),

\[
P(g|S) = \frac{\exp(o_g)}{\sum_{g \in G} \exp(o_g)}
\]

(6)

Where \( o \) is the scores of all genres, which is defined as follows:

\[
o = Mg
\]

(7)

Where \( M \) is the representation matrix to calculate the genre scores.

4. Experiments

4.1. Experimental Design

In this paper, our experiments evaluate our classifier model on the Free Music Achieve (FMA) which is arranged in a hierarchical taxonomy of 161 genres, which is a presentative benchmark for the long-tail Music Genre Classification task. This dataset is available in different sizes, such as Full, Large, Medium and Small size depending upon the number of samples in it.

We use the medium-FMA subset and predict the 16 top-level genres. And the default split schema proposed by FMA is adopted to make this research reproducible. Other datasets, such as GTZAN, Small-FMA, are also widely used in music genre classification. Because these datasets are class-balanced, so we do not use these datasets. However, the music genres in the real world present the long-tail and imbalanced distribution. Our goal is to classify long tail music genres with high accuracy.

4.2. Softmax Classifier Results

For evaluation, we report classification accuracy for all models. the formula is as follows:

\[
\text{Classification accuracy} = \frac{\text{Verify the correct number of data}}{\text{Total data}}
\]

(8)

4.3. Comparative Analysis of Other Methods

In order to prove the superiority of the musical attention networks in this paper, compared with neural network methods: CNN and CRNN. \(+\text{ATT}\) is our method.

4.4. Results

| Method       | Classification accuracy |
|--------------|-------------------------|
| CNN          | 66.4%                   |
| RNN          | 63.4%                   |
| CNN+ATT      | 68.3%                   |
| CRNN+ATT     | 69.4%                   |

Table 1. Classification accuracy of music genre classification models

Table 1 shows the classification accuracies of different methods, it can be seen that the classification accuracy of CRNN+ATT is the highest. Although our method has achieved obvious progress on the long-tail genres as compared with the plain ATT method, the results of all these methods are still far from satisfactory. We will incorporate more schemes and extra information to solve this problem in the future.
5. Conclusion
In this study, we present a novel musical attention network architecture to classify genres from large-scaled music dataset, especially the imbalanced data. The results showed that our method achieved state of art in long-tail musical genre classification. To the best of our knowledge, this is the first effort to use neural network to classify music genres on large imbalanced data sets. In the future, we will study the methods of NLP to make a greater breakthrough in music classification.

Acknowledgments
The research of this thesis is supported by The Ministry of education of Humanities and Social Science project (No.: 18YJA760045), we would like to extend our sincere gratitude to them.

References
[1] D. C. Corrêa and F. A. Rodrigues, “A survey on symbolic data-based music genre classification,” Expert Syst. Appl., vol. 60, pp. 190–210, 2016.
[2] J. Bergstra, N. Casagrande, D. Erhan, D. Eck, and B. Kégl, “Aggregate features and adaboost for music classification,” Mach. Learn., vol. 65, no. 2-3, pp. 473–484.
[3] T. Li, M. Oghihara, and Q. Li, “A comparative study on content-based music genre classification,” in SIGIR 2003: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 28 - August 1, 2003, Toronto, Canada, C. L. A. Clarke, G. V. Cormack, J. Callan, D. Hawking, and A. F. Smeaton, Eds. ACM, 2003, pp. 282–286.
[4] M. Defferrard, K. Benzi, P. Vanderghynst, and X. Bresson, “FMA: A dataset for music analysis,” in 18th International Society for Music Information Retrieval Conference (ISMIR), 2017.
[5] H. Lee, P. T. 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 22:23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, Eds. Curran Associates, Inc., 2009, pp. 1096–1104.
[6] S. Sigtia and S. Dixon, “Improved music feature learning with deep neural networks,” in IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014. IEEE, 2014, pp. 6959–6964.
[7] S. Leglaive, R. Hennequin, and R. Badeau, “Singing voice detection with deep recurrent neural networks,” in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015. IEEE, 2015, pp. 121–125.
[8] K. Choi, G. Fazekas, M. B. Sandler, and K. Cho, “Convolutional recurrent neural networks for music classification,” in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017. IEEE, 2017, pp. 2392–2399.
[9] T. Li, M. Oghihara, and Q. Li, “A comparative study on content-based music genre classification,” in SIGIR 2003: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 28 - August 1, 2003, Toronto, Canada, C. L. A. Clarke, G. V. Cormack, J. Callan, D. Hawking, and A. F. Smeaton, Eds. ACM, 2003, pp. 282–286.
[10] J. Choi, J. Lee, J. Park, and J. Nam, “Zero-shot learning for audio-based music classification and tagging,” in Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR2019, Delft, The Netherlands, November 4-8, 2019, A. Flexer, G. Peeters, J. Urbano, and A. Volk, Eds., 2019, pp. 6.
[11] S. Craw, B. Horsburgh, and S. Massie, “Music recommendation: Audioneighbourhoods to
discover music in the long tail,” in Case-Based Reasoning Research and Development - 23rd International Conference, ICCBR 2015, Frankfurt am Main, Germany, September 28-30, 2015, Proceedings, ser. Lecture Notes in Computer Science, E. Hüllermeier and M. Minor, Eds., vol. 9343. Springer, 2015, pp. 73–

[12] E. Liebman and P. Stone, “Artificial musical intelligence: A survey,” CoRR, vol. abs/2006.10553, 2020