Melodic Phrase Attention Network for Symbolic Data-based Music Genre Classification (Student Abstract)

Li Li, Rui Zhang, Zhenyu Wang*
Department of Software Engineering, South China University of Technology, Guangzhou, Guangdong, PR China
lili961028@gmail.com, zhang1rui4@outlook.com, wangzy@scut.edu.cn

Abstract

Compared with audio data-based music genre classification, researches on symbolic data-based music are scarce. Existing methods generally utilize manually extracted features, which is very time-consuming and laborious, and use traditional classifiers for label prediction without considering specific music features. To tackle this issue, we propose the Melodic Phrase Attention Network (MPAN) for symbolic data-based music genre classification. Our model is trained in three steps: First, we adopt representation learning, instead of the traditional musical feature extraction method, to obtain a vectorized representation of the music pieces. Second, the music pieces are divided into several melodic phrases through melody segmentation. Finally, the Melodic Phrase Attention Network is designed according to music characteristics, to identify the reflection of each melodic phrase on the music genre, thereby generating more accurate predictions. Experimental results show that our proposed method is superior to baseline symbolic data-based music genre classification approaches, and has achieved significant performance improvements on two large datasets.

Introduction

Existing music genre classification methods require manual extraction of features for different datasets, and need to extract a large number of music features because of the complexity of music, which results in costly feature engineering. In addition, these methods typically use traditional algorithms (such as SVM) for classification, so they are not able to make full use of complex musical features.

In this paper, we propose a melodic phrase attention network for symbolic data-based music genre classification to improve the classification performance.

Proposed Approach

Melody Segmentation Representation

The melody segmentation method we use is refer as the musical energy feature vector-based segmentation(Bingjie 2014), which defines four variables,namely *PitchArea*, *VolumeArea*, *MelodyArea* and *SoundArea*. A music piece is divided into several melodic phrases using the above method.

*Corresponding author.

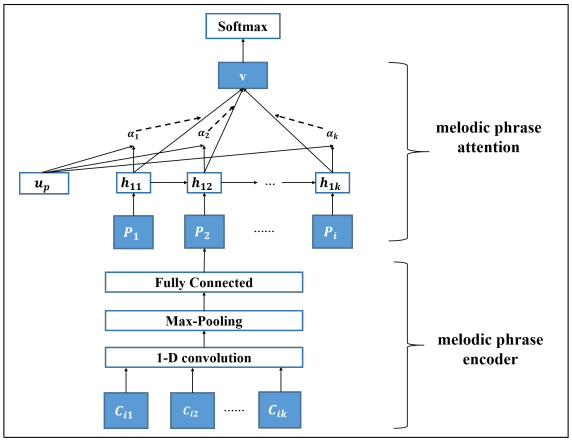

Figure 1: Overall framework of the Melodic Phrase Attention Network.

We use the alphabetic letters with an octave indication(e.g.C4 for note C in the fourth octave—middle C) to represent pitch of a note. After representing a music slice as a word, we apply a representation learning approach similar to word2vec, to get the vectorized representation: For each music slice encoded as a word d_t in a corpus of size T, the model tries to predict the surrounding music slice in a window c.

Melodic Phrase Attention Network

Figure 1 illustrates the overall architecture of our Melodic Phrase Attention Network, which consists of a melodic phrase encoder and melodic phrase attention network.

Melodic Phrase Encoder In this module, we obtain the vectorized representation c_{it} of each music slice in the i-th melodic phrase, and the input of the encoder is represented as $X_i = \{c_{i1}, c_{i2}, \cdots, c_{ik}\}$. We apply a CNN model as the Melodic Phrase Encoder, which is composed of a 1D convolutional layer, a max pooling layer and a fully-connected layer. We obtain vectorized representation P_i of the i-th melodic phrase through the Melodic Phrase Encoder.

Melodic Phrase Attention In order to identify melodic phrase that contribute to the classification task, we apply the

Model	TAGT	MASD
Musical Feature-SVM	0.409	0.176
Representation Learning-CNN	0.494	0.245
Representation Learning-GRU	0.449	0.218
Sia-2 (Ferraro and Lemström 2018)	—	0.358
Melodic Phrase Attention Network	0.796	0.508

Table 1: Comparison of our models to baseline model and state-of-the-art.

attention mechanism (Yang et al. 2016) to obtain the weight of each melodic phrase. Given the melodic phrase vectors $\{P_1, P_2, \cdots, P_k\}$, we first use a GRU layer to encode these phrases, then a attention network is applied for weight calculation, defined as:

$$u_i = tanh(W_s h_i + b_s) \quad (1)$$

$$\alpha_i = \frac{exp(u_i^T u_p)}{\sum_i exp(u_i^T u_p)} \quad (2)$$

$$v = \sum_i \alpha_i h_i \quad (3)$$

where u_p to evaluate the importance of the melodic phrases in the music piece. Finally, a softmax layer is used as classifier:

$$p = Softmax(W_c v + b_c) \quad (4)$$

The training loss of the network is cross entropy error of the correct genre label:

$$L = -log(p_c) \quad (5)$$

where c is the label of predicted music piece.

Experiments

Dataset and Benchmarks

We conduct experiments on two large-scale symbolic music datasets: (1) the **Tagtraum Genre Annotations** dataset (Schreiber 2015) (TAGA) consists of 21,353 MIDI files and is divided into 15 genres; (2) the **MSD Allmusic Top Genre** dataset (MASD) (Schindler, Mayer, and Rauber 2012) contains 24,623 MIDI files encompassing 25 genres.

We compare our model with the following benchmarks: (1) **Musical Feature-SVM**; (2) **Representation Learning-CNN**; (3) **Representation Learning-GRU**; and (4) **Sia-2** (Ferraro and Lemström 2018).

Experimental Results

Table 1 demonstrates the experimental results of our proposed method and the state-of-the-art approaches. Overall, our proposed method outperforms other benchmarks. Compared with the best baseline, our approach improve the accuracy of 0.302 on the TAGT dataset and 0.150 on the MASD dataset. Figure 2 illustrates accuracy result of part of the genres on TAGT and MASD. In both datasets, Melodic Phrase Attention Network improves the accuracy of almost every genre labels compared with the baseline model. Therefore, our proposed model performs better on TAGT and MASD.

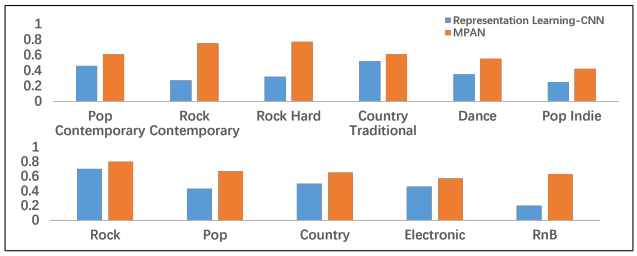

Figure 2: Accuracy results of part of the genres on MASD and TAGT.

Conclusion

In this paper, we presented Melodic Phrase Attention Network(MPAN), which captures the music genre feature of each melody phrase through the attention mechanism, for music classification. Experiments show that our proposed architecture outperform baseline models on two datasets and achieves the best results on TAGT. As future work, we plan to research other widely used features for music genre classification and improve our architecture according to other music characteristics.

Acknowledge

This work was supported by the Natural Science Foundation of Guangdong Province (No. 2019A1515011792), Science and Technology Program of Guangzhou, China (No. 201802010025), University Innovation and Entrepreneurship Education Fund Project of Guangzhou (No. 2019PT103), Guangdong Province Major Field Research and Development Program Project (No. 2019B010154004). The authors also thank the editors and reviewers for their constructive editing and reviewing.

References

Bingjie, H. 2014. *A Research on Music Emotion Analysis Based on Feature Vector*. Master's thesis, Xidian University.

Ferraro, A.; and Lemström, K. 2018. On large-scale genre classification in symbolically encoded music by automatic identification of repeating patterns. In *Proceedings of the 5th International Conference on Digital Libraries for Musicology*, 34–37.

Schindler, A.; Mayer, R.; and Rauber, A. 2012. Facilitating Comprehensive Benchmarking Experiments on the Million Song Dataset. In *ISMIR*, 469–474.

Schreiber, H. 2015. Improving Genre Annotations for the Million Song Dataset. In *ISMIR*, 241–247.

Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; and Hovy, E. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 1480–1489.