DO USER PREFERENCE DATA BENEFIT MUSIC GENRE CLASSIFICATION TASKS?

Ke Chen **Beici Liang**
QQ Music BU, Tencent Music Entertainment
knutchen@ucsd.edu, beiciliang@tencent.com

ABSTRACT

Using a massive amount of user interaction data with music tracks, we obtain a music embedding characterized by user preference and audio content. In this paper, we propose a transfer learning approach to find out if these music embeddings can benefit music genre classification tasks. Features from two pre-trained models (*musicnn* and our model) are extracted and decomposed via Principal Component Analysis (PCA. Then in the target task, a Support Vector Machine (SVM) is trained to classify the features into genres. Experiment was conducted using *GTZAN* and *FMA-small* datasets. Experimental results not only show that user preference data can benefit genre classification, but also affirm the universally applicable value of our music embeddings.

1. INTRODUCTION

Music genre classification is a widely-practical task in the music information retrieval (MIR) area. This can been done by a variety of data-driven methods using raw waveform [1] or time-frequency representation [2] of the audio data. To efficiently obtain a genre classifier, we adopted transfer learning technique to train a SVM classifier using features extracted from pre-trained models. One kind of feature is from the *musicnn* model which was trained for auto-tagging tasks [2]. The other is from our proposed music embedding model. This paper is focused on genre classification task, but the use of embeddings can provide us some insights on how to extract more general and meaningful features to solve MIR problems.

2. METHOD

An overview schematic of the proposed method is shown in Figure 1. It consists of two feature extractors and one basic classifier. The input audio data is convert to mel-spectrogram and sent to the extractors. After getting two kinds of features, we concatenate them into a single feature vector, which is then compressed into 128 dimensions by PCA [3]. The final prediction is performed using a SVM classifier with the radial kernel (RBF).

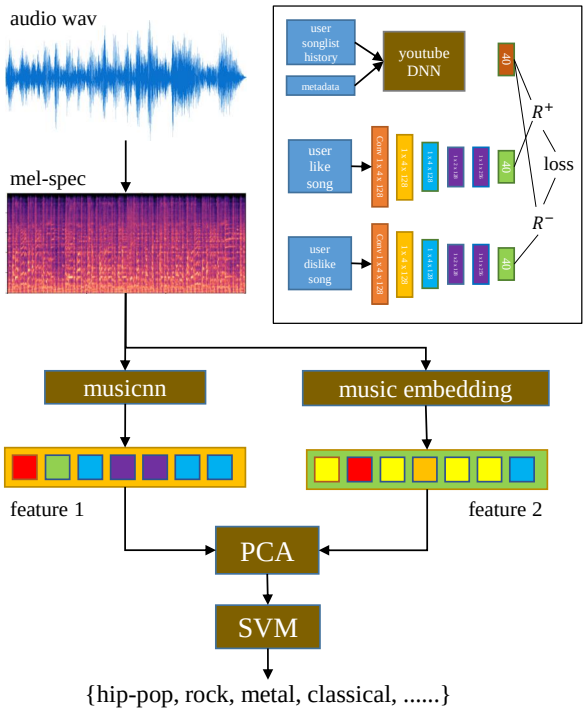

Figure 1. Overview schematic of the proposed method. The target audio is converted to mel-spectrogram, then sent to the pre-trained musicnn model and the proposed music embedding model to obtain features. The upper right corner shows a simplied version of our music embedding model structure.

The first feature extractor is *musicnn*. We directly use the pre-trained model from [4] and extract features from the `max_pool` layer in the musicnn network.

The second feature extractor is our proposed music embedding model, which includes two parts dealing with user and audio data respectively. As shown in the upper right corner in Figure 1, we leverage *YoutubeDNN* [5] to obtain the 40-dimensional user embeddings based on user listening histories from QQ Music. In the audio part, we use mel-spectrogram of audio data as inputs. For each paired training samples, we feed a user-liked track and a user-disliked track into a Siamese convolutional neural networks with shared parameters. The final layer of the audio-part network can output a 40-dimensional embedding for each track. Then we compute the cosine similarity between: 1) the user embedding and the liked-track embedding (R^+), and 2) the user embedding and the disliked-

Model	Accuracy	
	GTZAN	FMA-small
w/o music embedding	0.7653	0.5918
w/ music embedding	**0.8013**	**0.6148**

Table 1. Experimental results showing the accuracy score in two datasets.

track embedding (R^-) using the following equation:

$$R(U,I) = \frac{y_T^U \cdot y_I}{|y_U|\,|y_I|} \tag{1}$$

where y_U and y_I denote the user embedding and the audio embedding from our proposed music embedding model, respectively.

In the training process, we use the metric learning technique to define a loss function like in [6]:

$$loss(U,I) = max[0, \Delta - R^+ + R^-] \tag{2}$$

Here we only use 1 positive sample and 1 negative sample in each data training. We set the hyper-parameter $\Delta = 0.2$.

3. EXPERIMENT

With our proposed method, we conducted the experiment using two datasets: 1) GTZAN dataset with 1000 songs in 10 balanced genres [7], and 2) FMA-small dataset with 8000 songs in 8 balanced genres [8, 9]. According to the split in [10], we split the GTZAN dataset into 443:197:290 for training, validation and testing. In the FMA-small dataset, we split it into 7:3 as training and testing sets.

We compared results without using the proposed music embeddings as features (i.e., only with musicnn features), versus with the music embeddings. The accuracy scores are shown in Table 1. Our proposed method with the music embeddings can improve the performance of music genre classification task in both datasets.

4. REFERENCES

[1] T. Kim, J. Lee, and J. Nam, "Comparison and analysis of samplecnn architectures for audio classification," *IEEE J. Sel. Top. Signal Process.*, vol. 13, no. 2, pp. 285–297, 2019.

[2] J. Pons and X. Serra, "musicnn: Pre-trained convolutional neural networks for music audio tagging," in *Proceedings of the 20th International Society for Music Information Retrieval Conference, Late-Breaking/Demo Session, ISMIR*, Delft, The Netherlands, 2019.

[3] K. Pearson, "On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[4] *Musicnn*. [Online]. Available: https://github.com/jordipons/musicnn

[5] P. Covington, J. Adams, and E. Sargin, "Deep neural networks for youtube recommendations," in *Proceedings of the 10th ACM Conference on Recommender Systems, RecSys*. Boston, MA, USA: ACM, 2016, pp. 191–198.

[6] J. Park, J. Lee, J. Park, J. Ha, and J. Nam, "Representation learning of music using artist labels," in *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR*, Paris, France, 2018, pp. 717–724.

[7] G. Tzanetakis and P. R. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, 2002.

[8] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "FMA: A dataset for music analysis," in *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR*, 2017.

[9] M. Defferrard, S. P. Mohanty, S. F. Carroll, and M. Salathé, "Learning to recognize musical genre from audio," in *The 2018 Web Conference Companion*. ACM Press, 2018.

[10] *Audio Transfer Learning with Scikit-learn and Tensorflow*. [Online]. Available: https://github.com/jordipons/sklearn-audio-transfer-learning