CROSS-CULTURAL TRANSFER LEARNING USING SAMPLE-LEVEL DEEP CONVOLUTIONAL NEURAL NETOWRKS

Jongpil Lee, Jiyoung Park, Juhan Nam
KAIST
richter@kaist.ac.kr

Chanju Kim, Adrian Kim, Jangyeon Park, Jung-Woo Ha
NAVER Corp.
jungwoo.ha@navercorp.com

ABSTRACT

This submission aims to show how cultural differences in music affect audio feature learning. Measuring how largely music content differ in each culture is conducted via various MIREX 2017 audio classification tasks in a transfer learning setting. In this submission, two pre-trained deep neural networks are provided. One is trained with Million Song Dataset with the LastFM tag annotations and the other is learned from a music dataset from NAVER with genre annotations. Furthermore, two methods are applied to improve classification accuracy. One is using raw waveform-based sample-level deep convolutional neural networks as a feature extractor. The other is multi-level feature extraction and aggregation from the pre-trained networks to tackle various levels of abstractions in MIREX audio classification tasks. Finally, we put them into a classifier based on a support vector machine and make final predictions for each target task. Our submissions were ranked at the first in 6 out of a total of 8 Train/Test tasks of MIREX 2017.

1. INTRODUCTION

Current music genre or mood research is heavily focused on classical and popular music from western culture [3–5, 9], being limited in handling non-western music genres or moods. To take account of this problem, we compare two features acquired from western pop music and K-pop music. Furthermore, we mix these features to see how they can cooperate with each other. From the following sections, we describe the techniques applied to enhance the classification performance.

1.1 Combination of Multi-level Features

The labels in the MIREX 2017 audio classification tasks have various levels of abstractions in hierarchy and timescales, such as genre, mood and composers. To take account of the diversity, we extracted features from three last hidden layers of pre-trained networks [6]. After the features are extracted, we summarize them into a single feature vector and the final prediction is performed using a SVM classifier.

1.2 Sample-level Deep Convolutional Neural Networks

Learning from raw audio allows the network to learn very low-level features. Generally, in audio classification tasks, raw waveforms are converted to a time-frequency representation and used as input to the system. However, in this preprocessing stage, parameters, such as hop size or window size of Short Time Fourier Transform (STFT), are often ignored even though optimal parameters for each sound class may vary [2, 8]. To take account of this and also to avoid exhausting parameter search, we used a previously proposed network that learns features from raw waveforms using very small sample-level filters [7].

2. DATASETS

We used two pre-trained networks as a feature extractor. The datasets used for training are detailed in this section.

- Million Song Dataset with Last.FM tag annotations (MSD) [1]: 201680/11774 training/validation split, auto-tagging (50 tags)
- NAVER Music with genre annotations: 98998/12312 training/validation split, genre classification (107 genres)

The dataset configuration for MSD is described in our previous work [7, 8]. The transfer-learning experiments in [7] show that the networks learned from MSD can be applied well to different music classification datasets. Inspired by this result, we train a similar networks on a music dataset from NAVER with genre annotations to observe how cultural differences in music affect audio feature learning. When filtering the NAVER Music dataset, songs that have been played more than 30 times by users who have listened to more than 30 times of any songs are only considered. Then, most frequently used 107 genres out of 1660 genres are used as labels. Considering that most users of the NAVER Music service are Korean, the NAVER dataset reflects the unique characteristics of K-pop. In fact, we should note that, in our internal experiments, features learned from the NAVER dataset showed better classification accuracy on K-pop music tagging than those learned from MSD. Therefore, we expect to see interesting findings from the MIREX submission.

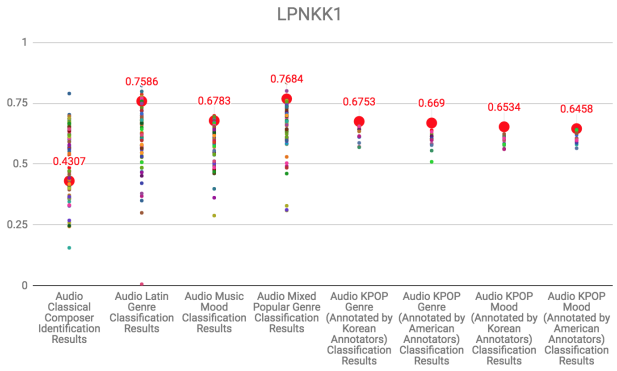

Figure 1. LPNKK1 results (red dots) and 8 years (2010-2017) of statistics for the MIREX audio classification tasks.

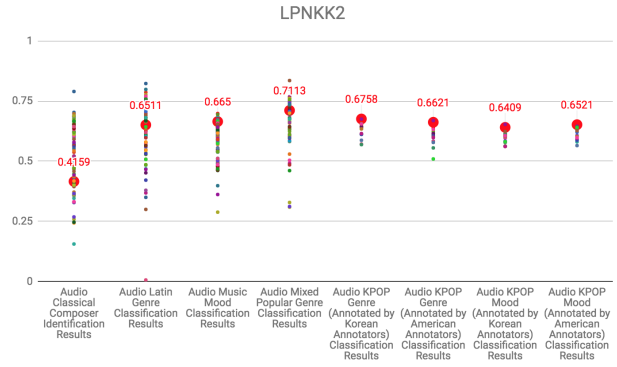

Figure 2. LPNKK2 results (red dots) and 8 years (2010-2017) of statistics for the MIREX audio classification tasks.

3. SUBMISSIONS AND RESULTS

Submission codes and their descriptions are as follows:

- LPNKK1: multi-level features extracted from pre-trained model with MSD
- LPNKK2: multi-level features extracted from pre-trained model with the NAVER dataset
- LPNKK3: concatenated features from both LPNKK1 and LPNKK2

After the results are announced, we summarized them including 8 years of statistics for the MIREX audio classification tasks. For the four K-pop related tasks, we summarized only 4 years (2014-2017) of statistics since they released from 2014. Figure 1, 2 and 3 show that we achieved the best place in the four K-pop tasks and also reached high scores in other tasks except the classical composer identification task. We should note that features from MSD showed better performances in most tasks. For the K-pop tasks, however, the features from the NAVER dataset is comparable to those from MSD. This indicates that cultural difference in music affects the transfer learning. Also, it is encouraging that their combination improves the performance overall. For future work, we need to refine the training with the NAVER dataset by trimming noisy labels and adding more songs.

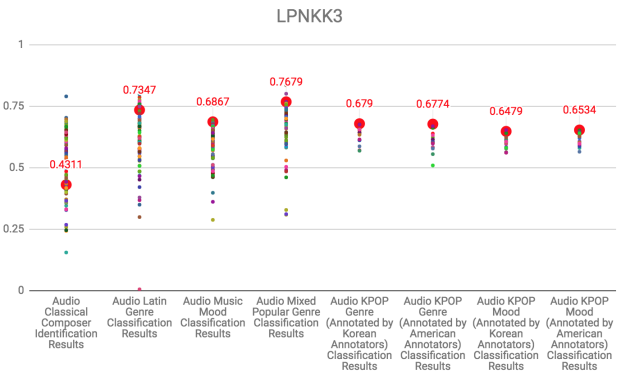

Figure 3. LPNKK3 results (red dots) and 8 years (2010-2017) of statistics for the MIREX audio classification tasks.

4. REFERENCES

[1] Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *ISMIR*, pages 591–596, 2011.

[2] Keunwoo Choi, Deokjin Joo, and Juho Kim. Kapre: On-gpu audio preprocessing layers for a quick implementation of deep neural network models with keras. *International Conference on Machine Learning (ICML), Machine Learning for Music Discovery Workshop*, 2017.

[3] Xiao Hu and Jin Ha Lee. A cross-cultural study of music mood perception between american and chinese listeners. In *ISMIR*, pages 535–540, 2012.

[4] Jin Ha Lee, Kahyun Choi, Xiao Hu, and JH Downie. K-pop genres: A cross-cultural exploration. In *ISMIR*, pages 529–534, 2013.

[5] Jin Ha Lee, J Stephen Downie, and Sally Jo Cunningham. Challenges in cross-cultural/multilingual music information seeking. In *ISMIR*, pages 1–7, 2005.

[6] Jongpil Lee and Juhan Nam. Multi-level and multi-scale feature aggregation using pre-trained convolutional neural networks for music auto-tagging. *IEEE Signal Processing Letters*, 24(8):1208–1212, 2017.

[7] Jongpil Lee and Juhan Nam. Multi-level and multi-scale feature aggregation using sample-level deep convolutional neural networks for music classification. *International Conference on Machine Learning (ICML), Machine Learning for Music Discovery Workshop*, 2017.

[8] Jongpil Lee, Jiyoung Park, Keunhyoung Luke Kim, and Juhan Nam. Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms. *Sound and Music Computing Conference (SMC)*, pages 220–226, 2017.

[9] Yi-Hsuan Yang and Xiao Hu. Cross-cultural music mood classification: A comparison on english and chinese songs. In *ISMIR*, pages 19–24, 2012.