REPRESENTATION LEARNING USING ARTIST LABELS FOR AUDIO CLASSIFICATION TASKS

Jiyoung Park, Jongpil Lee, Juhan Nam
KAIST
Graduate School of Culture Technology
jypark527@kaist.ac.kr

Jangyeon Park, Jung-Woo Ha
NAVER corp.
Seongnam, Korea
jungwoo.ha@navercorp.com

ABSTRACT

In this work, we use a deep convolutional neural network (DCNN) trained with a public dataset, the Million Song Dataset, as a feature extractor. We trained the network from audio mel-spectrogram using artist labels in a discriminative manner. In particular, we used a large number of neurons in the output layer where each neuron represents an artist label. The output of the last hidden layer of the DCNN is regarded as an identity feature of the input data. The DCNN extracts feature vectors by taking 3-second audio segments as input, and summarizes them as an musical feature for the input audio. These extracted features are used for training a Support Vector Machine (SVM) classifier to perform MIREX audio classification tasks such as genre or mood classification. The results show that the proposed approach effectively captures general music audio features.

1. INTRODUCTION

Audio classification systems are often trained with genre, mood or other timbre description labels. However, the process of annotating labels is tedious and sometimes ambiguous to find right ones. Also, high-quality annotation by music experts is time-consuming and expensive. Meanwhile, artist labels are objective information with no disagreement and they are annotated to songs naturally from the album release. Assuming that every artist has his/her own style of music, the artist labels can be regarded as terms that describe diverse styles of music. Thus, the audio features learned with artist labels can be used to explain general music features. With this hypothesis, we trained a DCNN that classifies audio into a large number of labels. We regard the DCNN as a feature extractor and apply it to MIREX audio classification tasks.

In this work, we trained a deep convolutional neural network (DCNN) with 100,000 preview clips of the Million Song Dataset (MSD) [1]. With recent advances in deep learning, CNN has been extensively used in classifying high-dimensional data such as image and audio. Once it is trained, the network can be used as a feature extractor that captures hierarchical representations from the input data [4, 6]. We trained the DCNN to identify different artists from audio mel-spectrogram. In particular, we used a large number of artists in the output layer so that learned features by the DCNN become more general and artist-independent. After that, the extracted features are used for training a Support Vector Machine (SVM) classifier to categorize genre, mood or other classes in MIREX 2017 classification tasks. This transfer learning setting, an approach that applies pre-trained neural networks to other domain datasets or tasks, has proven to be effective in audio classification tasks [2, 4].

2. PROPOSED SYSTEM

The proposed system is based on our recent research in [5]. The system consists of two parts: feature extraction and training/classification. The following sections describe these two parts in detail.

2.1 Feature Extraction

We use a one-dimensional DCNN where the one-dimensional convolution layers slide over only a single temporal dimension as a feature extractor. We used 5,000 artists in the output layer. Twenty songs per artist are used for training and divided into 15, 3 and 2 songs for training, validation, and testing, respectively. We used mel-spectrogram with 128 mel-bands as input. To compute a spectrogram, 1024 samples are used as FFT and hanning window sizes, and 512 samples are used for hop size. We performed magnitude compression with a nonlinear curve as $\log(1 + 10|A|)$.

Table 1 describes the proposed DCNN model configuration. We used categorical cross entropy loss with softmax activation on the prediction layer, batch normalization [3] after every convolution layer, a rectified linear unit (ReLU) activation for every convolution layer and dropout of 0.5 to the output of the last convolution layer. We optimized the loss using stochastic gradient descent with 0.9 Nesterov momentum. We also performed the input data normalization by dividing standard deviation after subtracting mean value across the training data.

We chose 3 seconds as a context size of the DCNN input after a set of experiments to find an optimal length. The song-level feature is built by averaging the segment

Input: 128×129 ($3sec$)	
Layer	Output Dim.
Conv 128×4	128×129
MP 4	128×32
Conv 128×4	128×32
MP 4	128×8
Conv 128×4	128×8
MP 4	128×2
Conv 128×2	128×2
MP 2	128×1
Conv 256×1	256×1
Dropout 0.5	256×1

Table 1. The proposed DCNN model configuration. Batch normalization is used after every convolution layer. MP means max pooling.

features from one input audio clip. Because we used 30-second preview clips of the MSD, the feature vectors of 10 segments are averaged into one feature vector for one audio clip. The same procedure applies to the MIREX training and test set. That is, a 30-second audio clip is divided into 10 segments and 256 feature vectors extracted from the segments are averaged into a single feature vector. As an additional step to improve discriminative power after the averaging, we apply linear discriminant analysis (LDA) to the feature vector. LDA maximizes the between artist variation while reducing the within artist variance. We obtained the LDA transformation matrix with the data used to train DCNN. This reduces the feature dimensions from 256 to 100. These extracted features will be used to classify the classes of the MIREX tasks.

2.2 Training and Classification

We extracted audio features from both training and test sets in the MIREX tasks using the DCNN. Then, we used the extracted features as input to a SVM classifier for the MIREX tasks.

2.3 Implementation

The proposed system is implemented in python using $librosa$ for extracting mel-spectrogram and $keras$ with $theano$ backend for training neural networks.

3. RESULTS AND CONCLUSIONS

Figure 1 shows the MIREX classification task results. Our submission is ranked at the 1st place in Music Mood Classification across all algorithms submitted from 2010 to 2017. Our algorithm also achieved decent results in mixed popular genre classification and most K-pop classification. This indicates that even though our model is trained with artist labels instead of song description labels, the proposed approach effectively captures general music audio features.

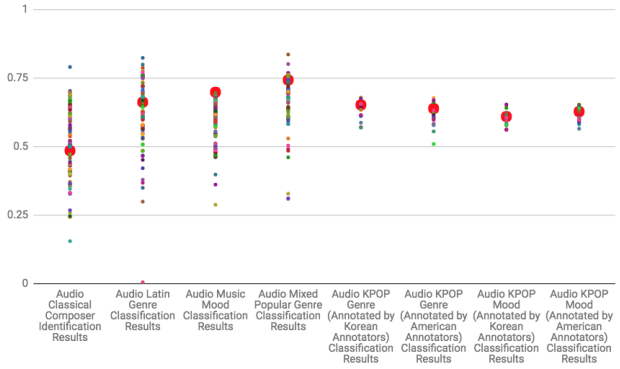

Figure 1. MIREX classification task results. The left four tasks are the results of all algorithms submitted from 2010 to 2017 and the right four tasks (K-pop datasets) are the results of those from 2014 to 2017. Our results are marked with a large red circle.

4. REFERENCES

[1] Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *ISMIR*, volume 2, page 10, 2011.

[2] Keunwoo Choi, György Fazekas, Mark Sandler, and Kyunghyun Cho. Transfer learning for music classification and regression tasks. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 2017.

[3] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

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

[5] Jiyoung Park, Jongpil Lee, Jangyeon Park, Jung-Woo Ha, and Juhan Nam. Representation learning of music using artist labels. *arXiv preprint arXiv:1710.06648*, 2017.

[6] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. CNN features off-the-shelf: An astounding baseline for recognition. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, CVPRW '14, pages 512–519, 2014.