Representation Learning of Music Using Artist, Album, and Track Information

Jongpil Lee [1] Jiyoung Park [2] Juhan Nam [1]

Abstract

Supervised music representation learning has been performed mainly using semantic labels such as music genres. However, annotating music with semantic labels requires time and cost. In this work, we investigate the use of factual metadata such as artist, album, and track information, which are naturally annotated to songs, for supervised music representation learning. The results show that each of the metadata has individual concept characteristics, and using them jointly improves overall performance.

1. Introduction

Representation learning of music has been recently performed by supervised deep learning using semantic labels such as genres, moods and instruments (Choi et al., 2017; Lee & Nam, 2017). However, annotating music with such semantic labels requires significant time and cost and the labels are often ambiguous, resulting in disagreement among annotators (Kim et al., 2017). Meanwhile, metadata such as artist labels require no cost and they are factual information with no ambiguity. We recently investigated the possibility of using artist information for representation learning of music and evaluated it in transfer learning settings (Park et al., 2018). The results showed that the learned representation is comparable to those using the semantic labels. In this work, we extend the use of music metadata to album and track information, which are more specific levels than the artist information. We use a similarity-based learning model following the previous work and also report the effects of the number of negative samples and training samples.

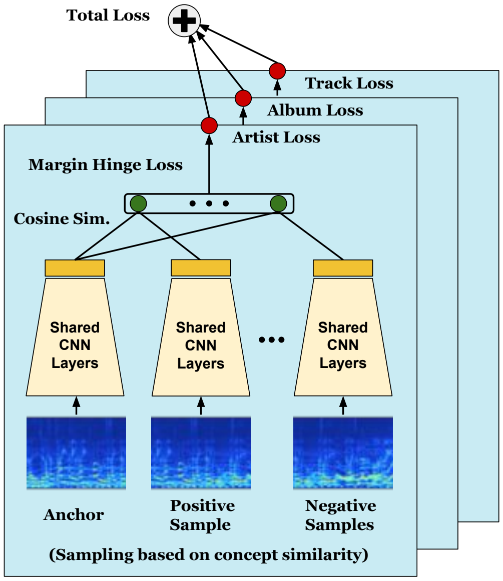

Figure 1. Joint learning model using artist, album, and track information.

2. Models

Figure 1 illustrates the overview of representation learning model using artist, album, and track information. Following the previous work, we use a Siamese-style Convolutional Neural Network (CNN) with multiple negative samples[1]. We build one large model that jointly learns artist, album, and track information and three single models that learns each of artist, album, and track information separately for comparison. The single model basically takes anchor sample, positive sample, and negative samples based on the similarity notion. For example, in the artist similarity concept, positive and negative samples are selected based on whether the sample is from the same artist as the anchor sample. We should note that the model takes a segment of audio (e.g. 3 second long), not the whole chunk of the song audio. Thus, in the track similarity concept, positive and negative samples are chosen based on whether the sample segment is from the same track as the anchor segment. Finally, we construct a joint learning model by simply adding three loss functions from the three similarity concepts, and share model parameters for all of them.

[1] Graduate School of Culture Technology, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea [2] NAVER Corp., South Korea. Correspondence to: Juhan Nam <juhannam@kaist.ac.kr>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

[1] In this work, we used twice the number of filters for all layers.

Table 1. Hold-out positive and negative sample prediction.

Learned concept	Artist	Album	Track	Artist +Album +Track
Artist	0.680	0.634	0.539	0.686
Album	0.732	0.822	0.653	0.763
Track	0.922	0.958	0.971	0.945

Table 2. Transfer learning experiment. Baseline results are generated by performing genre classification directly without transfer learning.

Learned concept	genres (baseline)	Artist	Album	Track	Artist +Album +Track
GTZAN	0.547	0.724	0.652	0.564	0.745
FMA small	0.533	0.598	0.560	0.463	0.593
NAVER Korean	0.720	0.662	0.641	0.549	0.663

Table 3. The effect of the number of negative samples. The model is trained with 1000 artists, 2000 albums, and its related track concepts.

Number of Negative Samples	1	2	4	8	16
GTZAN	0.665	0.663	0.681	0.702	0.711
FMA small	0.544	0.535	0.568	0.573	0.578
NAVER Korean	0.643	0.634	0.658	0.676	0.673

Table 4. The effect of the number of training samples. The model is trained with 4 negative samples with artist, album, and track concepts. The number of albums used is twice the number of artists.

Number of Training Artists	500	1000	2000	5000	10000
GTZAN	0.638	0.681	0.706	0.745	0.755
FMA small	0.517	0.568	0.588	0.593	0.603
NAVER Korean	0.636	0.658	0.668	0.663	0.686

3. Experiments and Evaluations

The four models are trained with Million Song Dataset (MSD) and its artist and album metadata (Bertin-Mahieux et al., 2011). We first build two splits based on each artist and album information. The artist split is the same as the previous work, which has 20 songs for each artist. For the album split, we selected 10 songs for each album and used twice as many albums to match the number of training samples of artist. Then, 10 songs of one album are divided into 8 songs, 1 song, and 1 song for training, validation and testing. The artist split is twice these numbers. For the validation sampling of artist or album concept, the positive sample is selected from the training set and the negative samples are chosen from the validation set based on the validation anchor's concept. For the track concept, it basically follows the artist split, and the positive sample for the validation sampling is chosen from the other part of the anchor song.

The evaluation is conducted in two ways: 1) hold-out positive and negative sample prediction and 2) transfer learning experiment. The hold-out positive and negative sample prediction was designed to see how well the models distinguish each concept. The evaluation is conducted on the test set of the above splits. For the artist and album concept, the positive sample is selected from the validation set and the negative samples are from test set based on the anchor's concept. In this evaluation, the random guess is 20% when the model uses 4 negative samples. The transfer learning experiment is performed on three external genre classification datasets including GTZAN (a fault-filtered version) (Tzanetakis & Cook, 2002; Kereliuk et al., 2015), FMA small (Defferrard et al., 2017), and NAVER Korean (Park et al., 2018). In this experiment, the learned representation is extracted and injected into a linear softmax classifier. This experiment was designed to see the generalization ability of the learned representations. For both evaluations, we used a model trained with 5000 artists (or/and 10000 albums) with 4 negative samples. After grid search, the margin values of loss function were set to 0.4, 0.25, and 0.1 for artist, album, and track concepts, respectively.

4. Results

The result of hold-out positive and negative sample prediction is shown in Table 1. We can see that each of the models performs best when the concept matches between the training set and test set. Also, the jointly learned model achieves good performance for all concepts.

The transfer learning experiment result is shown in Table 2. The artist model shows the best performance among the three single concept models, followed by the album model. This is probably because the genre classification task is more similar to the artist concept discrimination than album or track. The jointly learned model slightly outperforms the artist model. Finally, we included the baseline results obtained by performing genre classification directly without transfer learning. The results show that transfer learning using large music corpora with the factual metadata is highly effective in the GTZAN and FMA datasets, but not in NAVER dataset. This was due to the cross-cultural differences between the source and target datasets when looking closely at class-wise performances.

The effects of the number of negative samples and the number of training samples are shown in Table 3 and Table 4, respectively. We can see that increasing the number of negative samples and the number of training songs improves the model performance as expected.

References

Bertin-Mahieux, T., Ellis, D. P., Whitman, B., and Lamere, P. The million song dataset. In *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, volume 2, pp. 591–596, 2011.

Choi, K., Fazekas, G., Sandler, M., and Cho, K. Transfer learning for music classification and regression tasks. In *Proc. International Society for Music Information Retrieval Conf.*, pp. 141–149, 2017.

Defferrard, M., Benzi, K., Vandergheynst, P., and Bresson, X. Fma: A dataset for music analysis. In *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 316–323, 2017.

Kereliuk, C., Sturm, B. L., and Larsen, J. Deep learning and music adversaries. *IEEE Transactions on Multimedia*, 17 (11):2059–2071, 2015.

Kim, K. L., Kum, S., Park, C. L., Lee, J., Park, J., and Nam, J. Building k-pop singing voice tag dataset: A progress report. In *Late Breaking Demo in the International Society for Music Information Retrieval Conf.*, 2017.

Lee, J. and Nam, J. 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.

Park, J., Lee, J., Park, J., Ha, J., and Nam, J. Representation learning of music using artist labels. In *Proc. of International Society for Music Information Retrieval Conference (ISMIR)*, pp. 717–724, 2018.

Tzanetakis, G. and Cook, P. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5):293–302, 2002.