Zero-shot Learning and Knowledge Transfer in Music Classification and Tagging

Jeong Choi [1] **Jongpil Lee** [1] **Jiyoung Park** [2] **Juhan Nam** [1]

Abstract

Music classification and tagging is conducted through categorical supervised learning with a fixed set of labels. In principle, this cannot make predictions on unseen labels. Zero-shot learning is an approach to solve the problem by using side information about the semantic labels. We recently investigated this concept of zero-shot learning in music classification and tagging task by projecting both audio and label space on a single semantic space. In this work, we extend the work to verify the generalization ability of zero-shot learning model by conducting knowledge transfer to different music corpora.

1. Introduction

Zero-shot learning is a paradigm of machine learning to overcome the limitations of categorical supervised learning that can predict only a fixed number of classes. By leveraging additional side information regarding classes, zero-shot learning models can discover relationship among any arbitrary classes with regard to the given task, enabling inference towards classes that are unseen during training. We recently applied this approach to the music domain in attempts to allow prediction of novel music genres or retrieval of songs by a query word that users arbitrarily choose (Choi et al., 2019). By splitting tag labels into seen/unseen groups, we verified that the concept of knowledge transfer is possible within a dataset. In this study, we extend the evaluation of zero-shot embedding model to investigate how learned semantic relationships are transferred not only within a single dataset but also across different music corpora.

[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 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

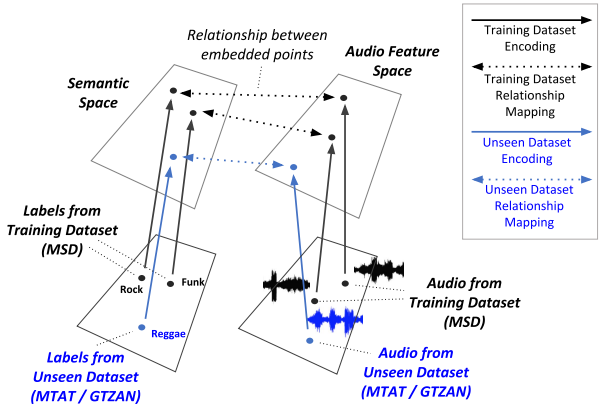

Figure 1. Overview of zero-shot learning and knowledge transfer applied to music domain.

2. Zero-Shot Learning for Music

Figure 1 illustrates the overview of zero-shot learning in the music domain. The semantic space is constructed using side information about the labels such as instrument annotations that describe music genres or a pre-trained word vector space such as GloVe (Pennington et al., 2014). The audio feature space is extracted via a convolutional neural network (CNN) that takes mel-spectrogram as input. The audio feature extraction and the mapping between the semantic space and audio feature space are learned using a Siamese-style network with the triplet loss, following the previous works (Frome et al., 2013; Park et al., 2018). Once the networks are trained, audio examples from the test set can be mapped to a point in the semantic space whose nearest label can be unseen ones during the training phase.

3. Knowledge Transfer

We conducted the evaluation of the zero-shot learning model within the same music corpus in our previous work (Choi et al., 2019). Since each of datasets has different audio characteristics and annotation criteria, evaluating the trained zero-shot model on different datasets can provide an insight into its generalization ability.

To this end, we train the zero-shot model on the Million

Table 1. Classification results. Tag retrieval AUC is reported for the MTAT dataset and genre annotation accuracy is reported for the GTZAN dataset.

MODEL	MTAT (RET. AUC)	GTZAN (ANNO. ACC.)
ZERO-SHOT LEARNING MODEL (MSD-GLOVE)	0.7390	0.7310
CLASSIFICATION MODEL (LEE & NAM, 2017)	0.9021	0.7203

Song Dataset (MSD) (Bertin-Mahieux et al., 2011) with the Last.fm tag annotations and then evaluate the trained model on MagnaTagATune (MTAT) (Law et al., 2009) and GTZAN (Tzanetakis & Cook, 2002) by directly feeding the test set of tracks to the network. We measure the classification performance and compare it with a referenced model.

4. Experiments

The zero-shot learning model is composed of audio and word branches. The audio branch consists of CNN layers and the word branch contains a linear projection layer on top of the side information lookup table built from a pre-trained general word semantic space. The two branches are then jointly trained with a max-margin hinge loss. We train the zero-shot model on MSD using 1,126 tags that are present in the side information lookup table.

We chose the GloVe embedding as our baseline side information for its large vocabulary that facilitates adaptation to unseen datasets (Pennington et al., 2014). We utilized a pre-trained GloVe model available online. It contains 19 million vocabularies with 300 dimensional embedding trained from documents in Common Crawl data. We then evaluated the model on MTAT and GTZAN. For the MTAT dataset, we followed the evaluation setup of using 50 most frequent tags (Dieleman & Schrauwen, 2014). 7 tags that are not present in GloVe vocabulary were omitted, resulting in total of 43 tags. For the GTZAN dataset, we used a fault-filtered version split (Kereliuk et al., 2015) and all 10 genre tags are used as all of them are presented in the GloVe dictionary. The evaluation is conducted on the test tracks of the referenced split for each dataset.

5. Results

The results of knowledge transfer are presented in Table 1. On GTZAN, the zero-shot learning model shows a significant improvement, even surpassing the performance of referenced supervised model which was supervised with the training set of GTZAN. On MTAT, on the contrary, the performance score is relatively low. We speculate that this is because the property of each dataset is different. GTZAN is comparatively more similar to MSD in terms of genre distribution of tracks and their general audio characteristics than MTAT.

Table 2. Comparison of label predictions for MTAT tracks on the original MTAT tags and 1,126 MSD Last.fm tags.

TRACK ID	GROUND TRUTH (MTAT)	PREDICTIONS OUT OF MTAT TAGS (THRESHOLD = 0.6)	PREDICTIONS OUT OF 1,126 MSD TAGS (THRESHOLD = 0.7)
8900	SLOW AMBIENT SYNTH	AMBIENT PIANO ELECTRONIC	INSTRUMENTAL CHILLOUT / AMBIENT SOUNDTRACK / JAZZ / PIANO RELAXING
11308	SLOW / TECHNO ELECTRONIC / BEAT SYNTH / WEIRD	ELECTRONIC TECHNO DANCE	ELECTRONIC / TECHNO ELECTRO / ELECTRONICA DOWNTEMPO / DUB BASSLINE / CHILLOUT EXPERIMENTAL

Table 3. Comparison of label predictions for GTZAN tracks on the original GTZAN genres and 1,126 MSD Last.fm tags.

TRACK ID	GROUND TRUTH (GTZAN)	TOP 1 PREDICTION OUT OF GTZAN GENRES	PREDICTIONS OUT OF 1,126 MSD TAGS (THRESHOLD = 0.6)
560	JAZZ	JAZZ	JAZZ / BLUES INSTRUMENTAL SWING / SMOOTHJAZZ PIANO
600	REGGAE	REGGAE	REGGAE / DUB / SKA ROOTS / DANCEHALL ROOTSREGGAE

We also present case studies of knowledge transfer evaluation. We investigated the predicted tags both from the label candidates of the test dataset and from the tag pool of the training dataset. The predictions on MTAT tracks and GTZAN tracks are shown in Table 2 and Table 3, respectively. From the results, we can see that that the model is able to predict labels that are semantically close to the ground truth labels even though no association between tracks and labels was informed to the model.

6. Conclusion and Future Work

We examined how well the knowledge is transferred across different music classification and tagging datasets via the zero-shot embedding space. For future work, we will explore the side information more thoroughly. For example, a word embedding reflecting more music-specific semantic relationship (e.g., trained with song review articles) would improve the performance of the model.

References

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

Choi, J., Lee, J., Park, J., and Nam, J. Zero-shot learning for audio-based music classification and tagging. In *Proc. of International Society for Music Information Retrieval Conference (ISMIR)*, 2019.

Dieleman, S. and Schrauwen, B. End-to-end learning for music audio. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6964–6968. IEEE, 2014.

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M. A., and Mikolov, T. Devise: A deep visual-semantic embedding model. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 26*, pp. 2121–2129. Curran Associates, Inc., 2013.

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

Law, E., West, K., Mandel, M. I., Bay, M., and Downie, J. S. Evaluation of algorithms using games: The case of music tagging. In *Proc. of International Society for Music Information Retrieval Conference (ISMIR)*, pp. 387–392, 2009.

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.

Pennington, J., Socher, R., and Manning, C. Glove: Global vectors for word representation. In *Proc. of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

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