ZERO-SHOT LEARNING FOR AUDIO-BASED MUSIC CLASSIFICATION AND TAGGING

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

[1] Graduate School of Culture Technology, KAIST

[2] NAVER Corp.

`{jeong.choi, richter, juhannam}@kaist.ac.kr, j.y.park@navercorp.com`

* Equally contributing authors.

ABSTRACT

Audio-based music classification and tagging is typically based on categorical supervised learning with a fixed set of labels. This intrinsically cannot handle unseen labels such as newly added music genres or semantic words that users arbitrarily choose for music retrieval. Zero-shot learning can address this problem by leveraging an additional semantic space of labels where side information about the labels is used to unveil the relationship between each other. In this work, we investigate the zero-shot learning in the music domain and organize two different setups of side information. One is using human-labeled attribute information based on Free Music Archive and OpenMIC-2018 datasets. The other is using general word semantic information based on Million Song Dataset and Last.fm tag annotations. Considering a music track is usually multi-labeled in music classification and tagging datasets, we also propose a data split scheme and associated evaluation settings for the multi-label zero-shot learning. Finally, we report experimental results and discuss the effectiveness and new possibilities of zero-shot learning in the music domain.

1. INTRODUCTION

Audio-based music classification and tagging is a task that predicts musical categories or attributes such as genre, mood, instruments and other song quality from music tracks. Current state-of-the-arts algorithms are based on supervised learning of deep convolutional neural networks that directly predict the labels in the output layer [18]. That is, the neural networks are trained to minimize the prediction errors with regards to the labels. The prediction results can be used to automatically annotate music tracks or retrieve music tracks using the labels as a query word [32]. By the nature of the setting in the supervised learning, however, the approach allows only a fixed set of word labels in the annotation and retrieval.

Zero-shot learning is a learning paradigm that can overcome this limitation and enables the trained model to predict unseen labels [12, 20]. For example, it allows the model to predict newly added music genres after the training or retrieve songs using a query word that users arbitrarily choose. This is possible by utilizing side information that derives a separate semantic space from labels. For example, the side information can be musical instrument annotation vectors of music genres or word embedding learned from sentences. The zero-shot learning approach conducts supervised learning between the semantic space and audio feature space. Once the mapping between two embedding spaces is learned, the model can predict unseen labels. Figure 1 illustrates the concept of zero-shot learning applied to music classification and tagging.

The zero-shot learning approach was previously applied to music data [28]. However, they focused on evaluating a specific semantic embedding method that works for general multimedia data rather than delving into zero-shot learning in the music domain. In this work, we carefully investigate how the concept of zero-shot learning can be properly applied to audio-based music classification and tagging. Specifically, we designed two settings of side information. One is using human-labeled attribute information and the other is using general word semantic information. Also, considering a music track is usually multi-labeled in music classification and tagging datasets, we propose a data split scheme that yields a comprehensive list of combinations for seen or unseen audio and labels, and evaluate them in the various settings. Through the experiments, we show the effectiveness and new possibilities of zero-shot learning in the music domain.

2. BACKGROUND: ZERO-SHOT LEARNING

Zero-shot learning has been studied mainly for object recognition in the field of computer vision [34]. They have attempted to build a model that can recognize novel visual categories without any associated training samples by employing a joint embedding space of both images and their class labels [34]. It was originally inspired by human's ability to recognize objects without seeing training examples or even create new categories dynamically based on semantic analysis [9]. What enables this semantic exploitation of unseen images is "information transfer" that
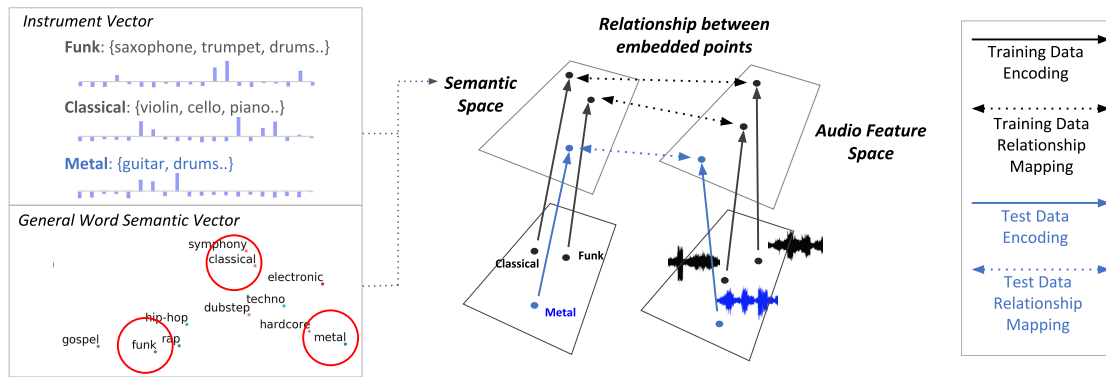
Figure 1. Overview of zero-shot learning concept applied to music classification and tagging.

applies the knowledge learned in an auxiliary domain to the main targeted task. Thus, it aims to explore how seen and unseen images are semantically related based on the side information.

The types of side information can be largely divided into two categories [9, 34, 37]. One is human-annotated attributes of the labels. In the computer vision community, there are publicly available human-annotated attribute datasets such as AWA (Animals With Attributes) [12] and CUB (Caltech-UCSD Birds-200) [35]. For example, AWA offers images of 50 animals annotated on 85 attributes associated with animals characteristics such as "furry" or "has tail". These examples may be equivalent to music genre classification where corresponding attributes are musical instruments. These attributes can be used as binary output to train a classifier and infer unseen class based on similarity measure [2, 12] or to learn their relative strength [21, 30]. When such explicit attributes data are not available, the semantic attributes can be learned with hierarchy of classes or other relationships [14, 25, 27]. In general, this attribute-based approach has advantages in interpretability but requires well-defined attributes and annotation data.

The other type of side information is a general word semantic space learned from different resources. A commonly used choice is neural language models such as Word2Vec [1, 8, 15, 16, 19, 36] and GloVe [1, 23, 36]. Another choice is learning relational semantic embeddings using pre-defined lexical hierarchy such as WordNet [17, 27]. These general semantic spaces have an advantage in that they have a large set of words in the vocabulary to predict unseen labels. However, if the target labels have a specific context, the general semantic space may fail to capture it [28].

3. DATA SPLIT SCHEME FOR MULTI-LABEL ZERO-SHOT LEARNING

Many of music classification and tagging datasets have multiple labels to annotate music tracks. However, zero-shot learning in the image domain has been treated primarily as a single-label problem, although a few studies have attempted to address it in multi-label classification [10,14]. An important difference between single-label and multi-

label zero-shot learning is data split scheme for training and test. In general, in zero-shot learning, both data and labels are split into seen and unseen sets. In the single-label setup, the label split can automatically divide the dataset into training and test sets (Figure 2 (a)). In the multi-label setup, however, specifying reasonable instance or label splits is not straightforward.

3.1 Previous Approaches

Most of previous works on multi-label zero-shot learning are conducted the instance-first split [26]. They first split instances into train and test, and only used seen labels for training as shown in the left of Figure 2 (b). In this case, some of the instances in the train set can have positive annotations for unseen labels. As an alternative, the label-first split was proposed in [33]. They first split labels into seen and unseen groups, and select training instances to have no positive annotation for unseen labels and select test instances to have at least positive annotation on unseen labels as shown in the right of Figure 2 (b). However, in this case, due to the nature of multi-label data, too many instances can be assigned to the test set.

Meanwhile, it is an important issue to determine which split in instance (train and test) or label (seen and unseen) should be evaluated in measuring zero-shot learning performance. A generalized zero-shot learning evaluation setting is proposed in [37]. It includes both seen and unseen labels at test time to examine more natural annotation performance compared to use only unseen labels at test time. In multi-label zero-shot learning, however, the data split and evaluation settings are still not clear and there is no agreed consensus yet.

3.2 Proposed Data Split Scheme

We propose a data split scheme and evaluation settings for multi-label zero-shot learning to measure the performance in more refined and various settings. The proposed data split is shown in Figure 2 (c). We first divide labels into seen (X) and unseen (Y) groups and then split instances into three groups. The first subset (A) of instances are labeled with at least one from seen labels and not labeled with any of unseen labels. The second subset (B) of in-
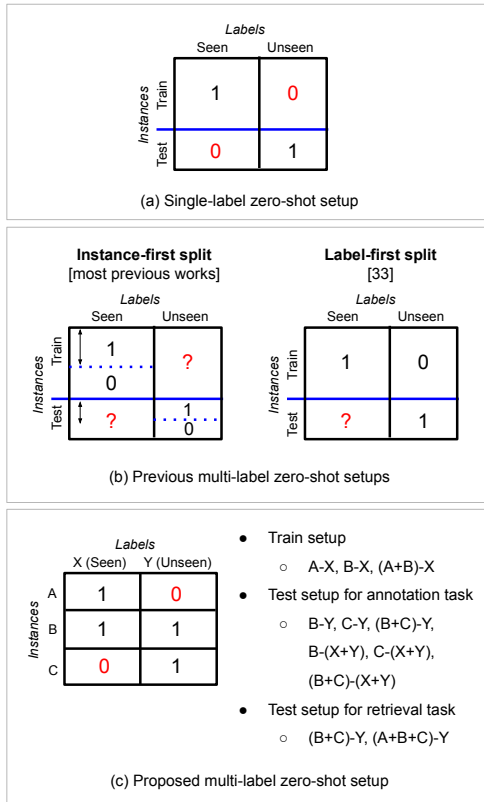
Figure 2. Data split for zero-shot learning

stances are labeled with at least one from each of seen and unseen labels. Lastly, the third subset (C) of instances are only labeled with at least one from unseen labels. Then we create the three setups (including train setup, test setup for annotation task, and test setup for retrieval task) with a combination of instance subsets (A, B, and C) and label subsets (X and Y) as described in Figure 2 (c).

In this configuration, the label-first split can be obtained when the training set is A-X (a subset where instances are from A and labels are from X) and the test set is (B+C)-Y (a subset where instances are from (B+C) and labels are from Y). Also, the generalized zero-shot learning evaluation setting can be expressed to include not only the Y area but also the (X+Y) area at test time. We therefore extend the split and evaluation settings more comprehensively, allowing for many aspects of multi-label zero-shot learning to be considered.

We take the train setup from A-X, B-X, and (A+B)-X. The instances in A-X only contain annotations on seen labels and the instances in B-X contain annotations on both seen and unseen labels. By distinguishing these two areas, we expect to see the difference in model learning by using instances with and without annotations in the unseen labels in multi-label zero-shot learning.

The test setup for the annotation task is made up of combinations of B, C, or (B+C) with Y. In addition, combinations of B, C, or (B+C) with (X+Y) can be also considered to measure generalized zero-shot learning performance on the annotation task as explained in the Section 3. The test setup for retrieval task is composed of (B+C)-Y

and (A+B+C)-Y. We can regard (A+B+C)-Y as a case of generalized zero-shot learning evaluation setting because the retrieval is performed not only on the instances of unseen labels ((B+C) split) but also on instances of seen labels (A split). The C-Y area cannot be formed in the retrieval evaluation. The reason for this is that once we split labels into seen and unseen and then assign overlapping instances to the B area, we may not guarantee that all the unseen labels have at least one positive activation on the instances in the C area.

4. MODEL

4.1 Deep Embedding Model

Zero-shot learning is primarily performed by a compatibility function that maps multimedia embedding and semantic embedding. The joint embedding methods can be categorized into learning linear compatibility, nonlinear compatibility, intermediate attribute classifier, and their hybrid [37]. In this work, we focus on learning nonlinear compatibility. The model takes audio mel-spectrogram as input rather than audio embedding extracted from pre-trained model. A convolutional neural network (CNN) module for audio is learned directly with semantic embedding from ground truth annotations. Figure 3 illustrates the model architecture. The model takes audio from one module and randomly selected a positive word and a negative word from the ground truth annotations of the audio via the semantic vector lookup table. Following the previous works [8, 22], the loss function is chosen as a max-margin hinge loss as below:

$$L(A, W) = \max \left[0, \Delta - Rel \left(A, W^+ \right) + Rel \left(A, W^- \right) \right]$$

where Δ is the margin, W^+ denotes the label with positive annotation for the audio input, and W^- denotes the label with negative annotation. The cosine similarity of the last hidden layer of the audio module and semantic module is used as a relevance score [22]:

$$Rel(A, W) = Similarity_{cosine} \left(y_A, y_W \right) = \frac{y_A^T y_W}{|y_A| \|y_W\|}$$

where y_A and y_W denote the output of the last hidden layer for the audio module and the semantic module, respectively.

The mel-spectrogram based audio CNN module is constructed with four 1D convolutional layers with a 2D filter [5, 7, 13, 24]. Each layer is followed by a rectified linear unit (ReLU) activation and a max pooling layer. We added a convolutional layer and an average pooling layer on top of them to construct a fixed-size audio embedding vector compatible to the semantic module. The semantic module is constructed by adding a fully connected layer over a semantic embedding (the output of semantic vector lookup table). In this case, the semantic embedding (or semantic vector lookup table) can be composed with both human-annotated attributes data or general word semantic space.
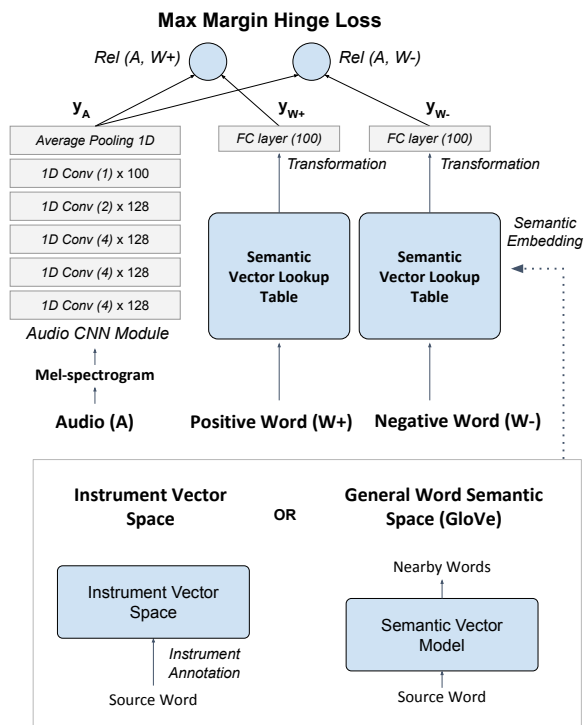
Figure 3. Deep embedding model for zero-shot learning using instrument vector space or general word semantic space.

4.2 Classification Model

The deep embedding model can be used for not only the zero-shot learning setting but also the conventional multi-label classification and tagging problem where we train and evaluate the model with only seen labels. Therefore, we construct a classification model and compare it with the deep embedding model to evaluate the learning capability of the deep embedding model. The classification model is basically the same as the audio module of the deep embedding model but the output is the binary representation of the labels. To this end, we added a fully-connected output layer with the size of seen labels on top of the average pooling layer of the audio module. We used the sigmoid activation and binary cross-entropy loss for the multi-label classification.

4.3 Training Details

We extract mel-spectrogram from audio with 128 mel bins, 1024 size FFT (Hanning window), and 512 size hop in 22,050 Hz sampling rate. We standardized the mel-spectrogram across all training data to have zero mean and unit variance. We randomly selected three seconds of audio chunk (130 frames) as an input size to the CNN module. We optimized the model using stochastic gradient descent with 0.9 Nesterov momentum, 0.001 learning rate, and $1e^{-6}$ learning rate decay for all models and datasets. Our system is implemented in Python 3.5.2, Keras 2.2.2, and Tensorflow-gpu 1.6.0 for the back-end of Keras.[1]

---

[1] The source code is available at `https://github.com/kunimi00/ZSL_music_tagging`

Dataset	Label Split			Instance Split			
	X (seen)	Y (unseen)	Total	A	B	C	Total
FMA	125	32	157	11606	6935	925	19466
MSD	900	226	1126	199385	188967	18057	406409

Table 1. Data split statistics.

In the test phase, we took the average of the fixed-size audio embedding vectors over a single music track to obtain a track-level embedding, and made the predictions by the distance between track-level audio embedding and tag embedding.

5. EXPERIMENTAL SETTINGS

We apply the proposed data split scheme and the deep embedding model to publicly available music classification and tagging datasets. We experiment with this in two settings of side information that we discussed in Section 2. For all the splits, we reserve 10% instances of train set as a validation set randomly.[2]

5.1 Experiment 1: Genre with Instrument Attributes

Musical instrument is one of the most important elements that determine music genre. Thus, recognizing instruments in a music track can be a strong cue to predict an unseen (or unheard) genre (assuming that the predictor learned the genre from some literature). For this experiment, we use Free Music Archive (FMA) [6] and OpenMIC-2018 datasets [11]. FMA contains audio files and genre annotations. OpenMIC-2018, which was originally designed for multiple instrument recognition, has 20 different instrument annotations to the audio files in FMA. We filtered the audio files to have both genre and instrument annotations. As a result, 19,466 audio with 157 genre labels and 20 instrument annotations are left. Following the proposed data split scheme, we randomly split 157 labels into 125 seen and 32 unseen ones. Then, three groups of audio instances (A, B, C) are created naturally. The statistics of audio and labels is described in Table 1.

The annotations on 20 different instruments in the OpenMIC-2018 dataset are labeled as a likelihood measure summarized from crowd-sourced annotations. They are regarded as positive if the likelihood value of an instrument annotation is larger than 0.5 and, otherwise, negative. The exact value of 0.5 means that it's unannotated. So, we treated positive and negative information independently and created 40 dimensional instrument vectors that can represent both positive and negative information. The genre to instrument attribute relationship is then constructed by accumulating 40 dimensional instrument vectors of the songs according to the genre labels. Finally, we standardized the instrument vectors to have zero mean and unit variance. This was used as the semantic vector lookup table in the learning model.

---

[2] All the data splits are available along with the source code.

Annotation Task		FMA					MSD				
Train	Test	AUC-i	MAP-i	P@1	P@5	P@10	AUC-i	MAP-i	P@1	P@5	P@10
A-X	B-Y	0.8736	0.5118	0.3465	0.4810	0.5008	0.8559	0.3391	0.3895	0.2944	0.3109
	C-Y	0.7917	0.3887	0.1563	0.0813	0.0643	0.8956	0.4536	0.3867	0.4278	0.4403
	(B+C)-Y	0.8640	0.4973	0.3333	0.4648	0.4855	0.8593	0.3491	0.3893	0.3060	0.3221
	B-(X+Y)	0.8834	0.3524	0.3725	0.2604	0.3056	0.9107	0.2664	0.4653	0.2735	0.2265
	C-(X+Y)	0.7958	0.1316	0.0162	0.0841	0.1089	0.9019	0.1594	0.0522	0.1292	0.1449
	(B+C)-(X+Y)	0.9207	0.4264	0.4018	0.3540	0.3932	0.9315	0.2842	0.3762	0.2645	0.2504
B-X	B-Y	0.8907	0.5633	0.4099	0.5371	0.5545	0.8679	0.3904	0.4803	0.3481	0.3623
	C-Y	0.8144	0.4008	0.2281	0.3556	0.3834	0.8926	0.4814	0.4379	0.4573	0.4674
	(B+C)-Y	0.8817	0.5442	0.3885	0.5158	0.5344	0.8700	0.3983	0.4766	0.3576	0.3715
	B-(X+Y)	0.9373	0.5142	0.5740	0.4206	0.4778	0.9217	0.2733	0.4462	0.2651	0.2283
	C-(X+Y)	0.8414	0.1886	0.0389	0.1494	0.1694	0.8941	0.1389	0.0666	0.1045	0.1231
	(B+C)-(X+Y)	0.8872	0.3606	0.3713	0.2881	0.3263	0.9276	0.2509	0.3272	0.2232	0.2147
(A+B)-X	B-Y	0.8864	0.5036	0.3090	0.4748	0.4951	0.8632	0.3563	0.4349	0.3124	0.3278
	C-Y	0.8286	0.4224	0.2530	0.3803	0.4064	0.8971	0.4836	0.4405	0.4591	0.4699
	(B+C)-Y	0.8796	0.4940	0.3024	0.4636	0.4846	0.8662	0.3674	0.4354	0.3252	0.3402
	B-(X+Y)	0.9118	0.4275	0.4784	0.3350	0.3813	0.9231	0.2977	0.5111	0.3094	0.2605
	C-(X+Y)	0.8276	0.1923	0.0530	0.1507	0.1738	0.9044	0.1920	0.1016	0.1629	0.1780
	(B+C)-(X+Y)	0.9073	0.3812	0.3770	0.3064	0.3445	0.9370	0.2984	0.3987	0.2808	0.2660

Table 2. Zero-Shot learning results for annotation task.

5.2 Experiment 2: General Word Semantic Space

The other type of side information is word embedding learned from a large-scale text dataset separately from music datasets. It can represent words as vectors in a semantic space. We adopted GloVe as a word embedding model [23]. Instead training it from scratch, we used a publicly available pre-trained GloVe model[3]. It consists of 300-dimensional vectors of 19 million vocabularies trained from documents in Common Crawl data. Since this can cover a large vocabulary of words, we used a music dataset with rich annotations, which is Million Song Dataset (MSD) with the Last.fm tag annotations [3]. From the full set of 498,035 tags in the Last.fm annotations, we filtered the tags that correspond to 2,000 genre/sub-genre classes contained in Tagtraum genre ontology [29]. We filtered the result (1800 tags) again into 1,126 tags after eliminating missing words in the dictionary of the pre-trained GloVe model. We used 406,409 audio instances annotated with the refined 1,126 tags. Following the proposed data split scheme, we randomly split 1,126 tags into 900 seen and 226 unseen ones, and organized three groups of audio instances (A, B and C). They are summarized in Table 1.

5.3 Evaluation Metrics

We used the area under the ROC curve averaged over instance (AUC-i), mean average precision over instance (MAP-i), and precision at K (P@K) as evaluation metrics for the annotation task. The retrieval task is evaluated using the area under the ROC curve averaged over label (AUC-l) and mean average precision over label (MAP-l).

6. RESULTS

6.1 Multi-label Zero-Shot Annotation

We compare the results of the combination of the proposed multi-label zero-shot learning split in the annotation task.

[3] The Common Crawl model was trained with 42B tokens containing 1.9M vocabulary. https://nlp.stanford.edu/projects/glove/

Data Split		FMA		MSD	
Train	Test	AUC-l	MAP-l	AUC-l	MAP-l
A-X	(B+C)-Y	0.6793	0.0904	0.6740	0.0279
	(A+B+C)-Y	0.6771	0.0392	0.6673	0.0149
B-X	(B+C)-Y	0.7194	0.1280	0.6907	0.0295
	(A+B+C)-Y	0.7236	0.0662	0.6843	0.0158
(A+B)-X	(B+C)-Y	0.7314	0.1170	0.6864	0.0310
	(A+B+C)-Y	0.7377	0.0518	0.6789	0.0172

Table 3. Zero-Shot learning results for retrieval task.

From Table 2, we can find that training with (A+B) or B instance set shows better performance than that with A in general. This indicates that the instances in B give better supervision over the entire tag set. In the case of test on C-(X+Y), the MAP-i and P@K scores are very low. This is because the case is generalized zero-shot learning evaluation setting [37] which makes predictions of seen label even when the ground truth of seen labels have only negatives.

We also see that some results have different trends between the two datasets. For example, test on B gives better results than C on FMA, but the results are opposite to those on MSD. We suspect that this may be due to difference in label cardinality, the average number of labels per instance [31], which can significantly affect performance in multi-label classification [4]. Specifically, FMA tracks have cardinality of 1.18 for B-Y and 1.15 for C-Y, whereas MSD tracks have cardinality of 2.04 in B-Y and only 1.11 in C-Y. This lower cardinality may cause better performance in C-Y than B-Y for MSD. However, we need further investigation considering differences in datasets and side information.

6.2 Multi-label Zero-Shot Retrieval

The results of the retrieval task are reported in Table 3. As in the annotation task, the overall performance is high when training with (A+B)-X. Also, the test results on

Data Split			FMA	MSD
Train	Test	Model	AUC-l	AUC-l
A-X	B-X	Deep Embedding Model	0.7381	0.7161
		Classification Model	0.7250	0.6980

Table 4. Results on retrieval task that compares deep embedding model to classification model. The detail of classification model is described in Section 4.2.

Query	Top 5 Retrieved Tracks (Title / Artist)	Original Last.fm Annotation
guitar	Iron Acton / Beak	psychedelic, experimental, krautrock, english, bass
	Drink Whiskey And Shut up / Brian Setzer	rock
	Thar She Blows / The Halibuts	party, surf, surfrock
	All Quiet On 34th Street / Eric Burdon And The New Animals	rock, rocknroll, hardrock, screamo, 00s
	Gimme Some Lovin' / Traffic	60s, british, classicrock, rock
lovely	Eddie My Love / The Chordettes	50s, doowop, oldies, pop, vocal
	Vaya Con Dios / Les Paul & Mary Ford	50s, jazz, oldies
	(I Can't Help You) I'm Falling Too / Skeeter Davis	country, oldies
	Mr. Blue / The Fleetwoods	oldies, 50s, pop, doowop, ballad
	I'm Blue Again / Patsy Cline	blue, country

Table 5. Top 5 retrieved tracks for a query word from unseen tag subset ('guitar') and an arbitrary word ('lovely').

(A+B+C)-Y are lower than that on (B+C)-Y. We can regard the test on (A+B+C)-Y as a generalized zero-shot learning evaluation setting for the retrieval task. This means that even for instances that do not have a positive annotation on unseen label according to the split (instances that were denoted as A), the evaluation is performed including this instances so to consider whether the model actually make a negative prediction on them. Thus, this is a more strict evaluation setting.

6.3 Deep Embedding Model vs. Classification Model

We also conducted an additional experiment to verify the performance of the deep embedding model in the conventional multi-label classification and tagging task [4]. Table 4 shows results in the retrieval scenario following previous work [5, 7, 13, 24]. We can see that the deep embedding model outperforms the classification model on both datasets. This indicates that associating audio with labels via the side information is a powerful approach even in the conventional multi-label classification and tagging task.

---

[4] In this evaluation, some labels were excluded because there are some labels that have negative annotations for all instances according to the split.

Smells like teen spirit Nirvana	Superstition Stevie Wonder	Theme to Grace / Lament George Winston
classicrock (unseen)	funk (unseen)	folk
punk	soul	instrumental
rock (unseen)	pop (unseen)	jazz
80s (unseen)	jazz	piano (unseen)
alternative	80s (unseen)	singersongwriter
punkrock	blues (unseen)	chillout
90s	classicrock	acoustic
metal	90s (unseen)	blues
vintage	disco	mellow
alternativerock	dance	chill

Table 6. Top 10 auto-tagging results for examples of well-known songs including unseen tags during training.

Query	General Semantic Space	Zero-shot Embedding Space
guitar	bass, acoustic, piano, vocals, violin, percussion, strings, vocal, music, jazz	instrumental, minimal, rock, acidrock, progressiverock, alternative, psychedelic, folkrock, classicrock, band
lovely	awesome, love, cool, romantic, relaxing, summer, christmas, holiday, vintage, soft	relaxing, relax, lovesongs, easylistening, baby, country, romantic, easy, americana, ballad

Table 7. Comparison of top 10 nearest word vectors (out of 1,126 tags) on general semantic space and the trained zero-shot embedding space.

6.4 Case Study

We conducted case studies to better understand the performance of the zero-shot learning model. Table 6 shows the results of annotation on several famous music tracks. They show that the predictions are reasonable for both seen or unseen tags. Table 5 lists retrieved tracks given a query word and their original labels. We can see that the results are reasonable even if we did not use seen tags. Furthermore, we compared the general semantic space of GloVe model to our trained deep embedding space in Table 7. The examples show the deep embedding space learns the relationship between words in a more musically meaningful way.

7. CONCLUSIONS

In this paper, we showed that zero-shot learning is capable of associating music audio with unseen labels using side information. This allow to use a rich vocabulary of words to describe music, thereby enhancing the experience of music retrieval or recommendation in a more human-friendly way. There is a large room to explore for future work. For example, lyrics can be used as side information which can be obtained without manual human annotation. The neural language models can be also trained to contain more musical context, for example, using text descriptions of playlists or music articles.

8. ACKNOWLEDGMENT

9. REFERENCES

[1] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proc. of The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2927–2936, 2015.

[2] Ziad Al-Halah, Makarand Tapaswi, and Rainer Stiefelhagen. Recovering the missing link: Predicting class-attribute associations for unsupervised zero-shot learning. In *Proc. of The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5975–5984, 2016.

[3] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proc. of International Society for Music Information Retrieval Conference (ISMIR)*, 2011.

[4] Francisco Charte, Antonio Rivera, María José del Jesus, and Francisco Herrera. Improving multi-label classifiers via label reduction with association rules. In *International Conference on Hybrid Artificial Intelligence Systems*, pages 188–199. Springer, 2012.

[5] Keunwoo Choi, György Fazekas, Mark Sandler, and Kyunghyun Cho. Convolutional recurrent neural networks for music classification. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2392–2396, 2017.

[6] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. Fma: A dataset for music analysis. In *Proc. of International Society for Music Information Retrieval Conference (ISMIR)*, 2017.

[7] Sander Dieleman and Benjamin Schrauwen. End-to-end learning for music audio. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6964–6968, 2014.

[8] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2121–2129. Curran Associates, Inc., 2013.

[9] Yanwei Fu, Tao Xiang, Yu-Gang Jiang, Xiangyang Xue, Leonid Sigal, and Shaogang Gong. Recent advances in zero-shot recognition: Toward data-efficient understanding of visual content. *IEEE Signal Processing Magazine*, 35(1):112–125, 2018.

[10] Yanwei Fu, Yongxin Yang, Tim Hospedales, Tao Xiang, and Shaogang Gong. Transductive multi-label zero-shot learning. *British Machine Vision Association*, 2014.

[11] Eric Humphrey, Simon Durand, and Brian McFee. Openmic-2018: an open dataset for multiple instrument recognition. In *Proc. of International Society for Music Information Retrieval Conference (ISMIR)*, 2018.

[12] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2014.

[13] 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.

[14] Thomas Mensink, Efstratios Gavves, and Cees GM Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *Proc. of The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2441–2448, 2014.

[15] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proc. of International conference on Learning Representations (ICLR)*, 2013.

[16] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[17] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[18] Juhan Nam, Keunwoo Choi, Jongpil Lee, Szu-Yu Chou, and Yi-Hsuan Yang. Deep learning for audio-based music classification and tagging: Teaching computers to distinguish rock from bach. *IEEE Signal Processing Magazine*, pages 41–51, 2019.

[19] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. In *Proc. of International conference on Learning Representations (ICLR)*, 2013.

[20] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. Zero-shot learning with semantic output codes. In *Advances in neural information processing systems*, pages 1410–1418, 2009.

[21] Devi Parikh and Kristen Grauman. Relative attributes. In *International Conference on Computer Vision (ICCV)*, pages 503–510. IEEE, 2011.

[22] Jiyoung Park, Jongpil Lee, Jangyeon Park, Jung-Woo Ha, and Juhan Nam. Representation learning of music using artist labels. In *Proc. of International Society for Music Information Retrieval Conference (ISMIR)*, pages 717–724, 2018.

[23] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proc. of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[24] Jordi Pons, Thomas Lidy, and Xavier Serra. Experimenting with musically motivated convolutional neural networks. In *Proc. of the International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 1–6, 2016.

[25] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *Proc. of The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 49–58, 2016.

[26] Zhou Ren, Hailin Jin, Zhe Lin, Chen Fang, and Alan Yuille. Multiple instance visual-semantic embedding. In *Proc. of the British Machine Vision Conference (BMVC)*, volume 1, page 3, 2017.

[27] Marcus Rohrbach, Michael Stark, György Szarvas, Iryna Gurevych, and Bernt Schiele. What helps where– and why? semantic relatedness for knowledge transfer. In *Proc. of The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 910–917. IEEE, 2010.

[28] Ubai Sandouk and Ke Chen. Multi-label zero-shot learning via concept embedding. *arXiv preprint arXiv:1606.00282*, 2016.

[29] Hendrik Schreiber. Genre ontology learning: Comparing curated with crowd-sourced ontologies. In *Proc. of International Society for Music Information Retrieval Conference (ISMIR)*, pages 400–406, 2016.

[30] Krishna Kumar Singh and Yong Jae Lee. End-to-end localization and ranking for relative attributes. In *European Conference on Computer Vision*, pages 753–769. Springer, 2016.

[31] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007.

[32] Douglas Turnbull, Luke Barrington, David Torres, and Gert Lanckriet. Semantic annotation and retrieval of music and sound effects. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 16(2):467–476, 2008.

[33] Qian Wang and Ke Chen. Multi-label zero-shot human action recognition via joint latent embedding. *arXiv preprint arXiv:1709.05107*, 2017.

[34] Wei Wang, Vincent W Zheng, Han Yu, and Chunyan Miao. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):13, 2019.

[35] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-UCSD birds 200. 2010.

[36] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *Proc. of The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 69–77, 2016.

[37] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning - the good, the bad and the ugly. In *Proc. of The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.