MT-GCN FOR MULTI-LABEL AUDIO TAGGING WITH NOISY LABELS

Harsh Shrivastava[⋆†] *Yifang Yin*[⋆] *Rajiv Ratn Shah*[†] *Roger Zimmermann*[⋆]

[⋆] National University of Singapore [†] MIDAS Lab IIIT-Delhi

ABSTRACT

Multi-label audio tagging is the task of predicting the types of sounds occurring in an audio clip. Recently, large-scale audio datasets such as Google's AudioSet, have allowed researchers to use deep learning techniques for this task but this comes at the cost of label noise in the datasets. Audio datasets such as the AudioSet are usually built following a hierarchical structure known as ontology which captures the relationships between different sound events with domain knowledge. However, existing methods for audio tagging failed to utilize this domain knowledge about label relationships in their models, resulting in models being sensitive to label noise. We therefore present MT-GCN, a Multi-task Learning based Graph Convolutional Network that learns domain knowledge from ontology. The relationships between sound events in our proposed method are described by a graph. We propose two ontology-based graph construction methods, and conduct extensive experiments on the FSDKaggle2019 dataset. The experimental results show that our approach outperforms the baseline methods by a significant margin.

Index Terms— Audio Tagging, Graph Convolutional Networks, Multi-task Learning

1. INTRODUCTION

Multi-label audio tagging is the task of labelling the sound recordings with the types of sounds present in them [1]. It can be applied to many tasks such as music tagging [2], information retrieval [3] and acoustic monitoring [4]. Due to the release of large-scale audio datasets, research in this area has gained momentum because it allowed researchers to use complex deep neural networks [5]. Since it is difficult to manually label these large-scale datasets, the creators mostly use some heuristics on associated meta-data to infer the labels, which inevitably results in label noise.

Label noise harms the performance of deep neural networks for audio tagging as they tend to overfit to it. Coping with label-noise has been attempted quite extensively and intensively in computer vision [6, 7] but in audio recognition, it has not been paid as much of the attention. To foster research in this direction, Fonseca *et al.* [8] proposed a dataset called FSDKaggle2019 which has been featured in Detection and Classification of Acoustic Scenes and Events 2019 (DCASE 2019) Challenge. They proposed to investigate the scenario where a very small manually-labelled dataset is available, along with a large noisy-labelled dataset in multi-label audio tagging setting based on a vocabulary of 80 labels. These 80 labels are chosen from the Google's AudioSet ontology [9]. An ontology is a hierarchical structure representing the relationships among the categories of sounds. Despite being a rich source of domain knowledge, existing audio tagging systems failed to make use of it. Almost all of the submissions to the DCASE 2019 Task 2 were some kind of formulation of deep neural networks and all of them could not utilize this information. Research has shown that domain knowledge can be used to regularize the machine learning models [10, 11].

We consider this question, *How can we utilize the ontology to improve the performance of deep networks for audio tagging with noisy labels?* In this paper, we propose, Multi-task Graph Convolution Network (MT-GCN), to incorporate ontology-based domain knowledge and use it as a regularization in a multi-task learning setup to deal with label noise. The system overview is illustrated in Figure 1. The approach is to train two network modules together i.e. *multi-task learning module* which trains on two tasks of learning on clean data and learning on noisy data respectively and *Graph Convolution module* which trains on the ontology-based graph and is shared between the two tasks. The key idea is to learn semantic-enriched representations for each class that encode the label relations derived from the ontology. A shared-weight GCN is adopted among different tasks as both the clean and noisy datasets should follow the domain knowledge derived from the ontology. The introduction of the ontology-based GCN in our proposed method can be considered as an effective regularization to jointly learn representative acoustic features from both the curated and noisy labeled data. The main contributions of this paper are summarized as follows:

- We present a novel approach for building general audio tagging systems with noisy labels by utilizing the ontology-based domain knowledge.

This research has been supported in part by Singapores Ministry of Education (MOE) Academic Research Fund Tier 1, grant number T1 251RES1713. We gratefully acknowledge the support of NVIDIA Corporation with the donation of a Titan Xp GPU used for this research. Rajiv Ratn Shah is partly supported by the Infosys Center for AI, IIIT Delhi.

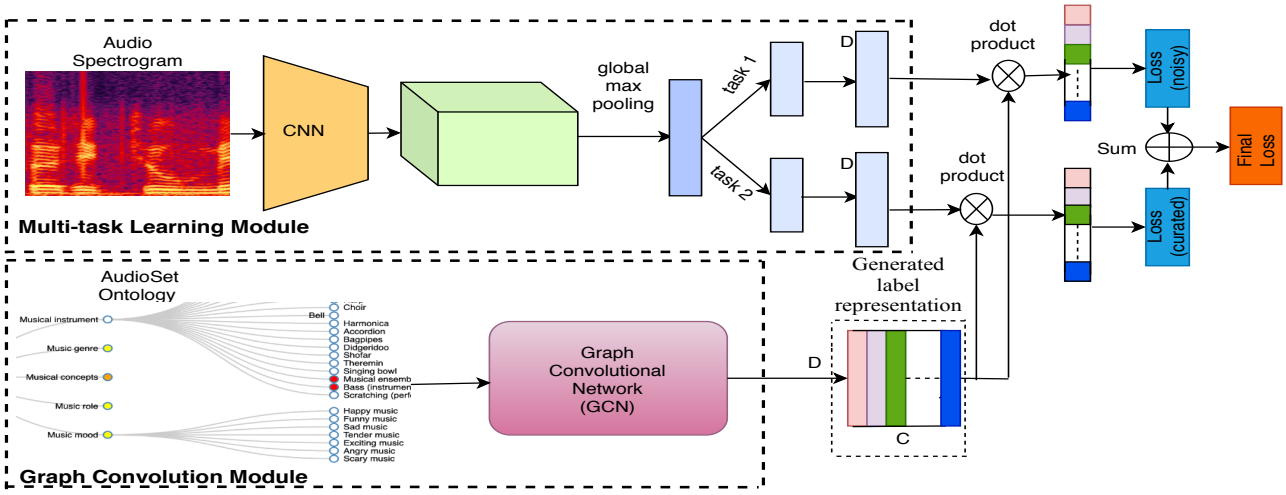

Fig. 1. Block diagram of MT-GCN. A stacked GCN is learnt over the ontology-based graph to map these label representations into a set of label representations capturing the domain knowledge i.e. $\mathbf{W} \in \mathbf{R}^{C \times D}$, which are applied to audio spectrogram representation extracted from noisy-labelled data and curated data.

- We propose two effective methods to build the label correlation graph based on the ontology. To the best of our knowledge, we are the first to design and apply an ontology-based regularization to noisy labeled audio tagging.
- We perform extensive experiments on the FSDKaggle2019 dataset. Experimental results show that our proposed methods for building the label correlation graph for the GCN outperforms all the previous methods by a significant margin.

2. METHODOLOGY

2.1. Problem Statement

Given a training set $D = \{(x_1, y_1), ..., (x_n, y_n)\}$, where x_i denotes the i^{th} audio clip and $y_i \in \{0, 1\}^C$ a one-hot vector representing the clip-level label over C classes, our goal is to learn a neural network parameterized by θ trained on D to predict the occurrence of C labels in an input unseen audio clip. We formulate this problem as multi-label audio classification problem. The training set D consists of two datasets i.e. a small set of manually-labeled data, and a larger set of noisy-labeled data. The dataset is labeled using a large vocabulary of labels from an ontology.

2.2. MT-GCN for Audio-tagging

We now introduce our method for multi-label audio classification with noisy labels, MT-GCN. The goal of this appproach is to use information about label relationship as a regularization for learning generalized representation of the audio from the noisy labeled data. To accomplish this goal, we train two network modules jointly: a *multi-task learning module,* which trains on two tasks of same goals as audio tagging but one with noisy-labeled dataset and other with manually-labelled dataset, in a standard multi-task learning setting and a *graph convolution module,* which trains on the ontology.

2.2.1. *Multi-task Learning Module*

We denote the multi-task network as a function $f_\theta(x)$ with parameters θ which takes input audio sample x. In this network, we use hard parameter sharing approach, in which we predict for both of the noisy-labeled data and curated data using the shared set of parameters θ. We use ResNet-101 [12] as the shared network in the experiments. We use this multi-task network to learn the audio spectrogram representations for each kind of the data. Therefore at the last shared layer, we further add data-specific layers to get the corresponding representations of dimension D (2048) for curated and noisy data, respectively. We denote noisy-labelled data representations by x^{noisy} and curated data representations by $x^{curated}$. We denote the true noisy labels by y^{noisy} and true curated labels by $y^{curated}$.

2.2.2. *Graph Convolution Module*

Following Chen *et al.* [13], we use GCN to learn label representations for our task. Given a Graph G, A GCN learns a function $f(.)$ that takes as input : a feature matrix $\mathbf{H}^l \in \mathbf{R}^{n \times d}$ and the corresponding correlation matrix $\mathbf{A} \in \mathbf{R}^{n \times n}$, (where n is the number of nodes and d is the dimensionality of node features), and outputs updated node features as $\mathbf{H}^{l+1} \in \mathbf{R}^{n \times d'}$. A GCN layer can be written as $\mathbf{H}^{l+1} = f(\mathbf{X}^l, \mathbf{A})$. A correlation matrix represents the graphical structure among

the labels. Now, we can represent f by applying convolutional operation as

$$\mathbf{H}^{l+1} = h(\hat{\mathbf{A}}\mathbf{H}^{l}\mathbf{W}^{l}), \tag{1}$$

where $\mathbf{W}^l$ is a parameter matrix to be learned and $\hat{\mathbf{A}} \in \mathbf{R}^{n \times n}$ is a normalized correlation matrix $\mathbf{A}$, and $h(.)$ is a non-linear transformation.

2.2.3. MT-GCN based learning

The input to the first layer of GCN is $\mathbf{Z} \in \mathbf{R}^{C \times d}$ matrix where C is the number of classes and d is the dimensionality of the label representations. To represent the labels, we use the one-hot encoding scheme. Other word embeddings can be used but as being pointed out [13], it should not make a huge difference in the results. The last layer of GCN outputs $\mathbf{M} \in \mathbf{R}^{C \times D}$ matrix, where D is the dimensionality of the image representations produced by the multitask network. We obtain final audio predictions by applying the learned label representations to the image representations as follows:

$$\hat{y}^{noisy} = \mathbf{M}x^{noisy} \tag{2}$$

$$\hat{y}^{curated} = \mathbf{M}x^{curated} \tag{3}$$

We assume that the ground truth label of an audio is $y \in \mathbf{R}^C$, where $y^i = 0, 1$ denotes if the label i appears in the audio recording. We finally compute the multi-label classification loss as follows:

$$L = \sum_{c=1}^{C} y^c \log(\sigma(\hat{y}^c)) + (1 - y^c) \log(1 - \sigma(\hat{y}^c)) \tag{4}$$

where σ is the sigmoid function. We calculate this for noisy labelled data as well as for curated data and sum them together to get the final loss function of the whole network as:

$$Loss = L^{noisy} + L^{curated} \tag{5}$$

2.3. Correlation Matrix for MT-GCN

GCN learns useful node representations by propagation of information between nodes based on the correlation matrix. Thus, building a correlation matrix A is a crucial challenge for GCN.

2.3.1. Labels Co-occurrence based Correlation Matrix

In this work [13], authors proposed a co-occurrence based correlation matrix. They construct the matrix by counting the occurrence of label pairs in the training set and normalizing the matrix by individual label count. They further binarize this matrix for improving its generalization capacity. For more details we refer interested readers to [13]. As this formulation of correlation matrix for GCN is not studied in the context of Audio Tagging, we experiment with this in our experiments for comparing with our proposed matrix construction methods. We created two co-occurrence based matrices: (1) With curated dataset only and (2) with curated+noisy dataset. We use noisy dataset labels for increasing the size of the label set to get denser correlation matrix.

2.3.2. Ontology Based Correlation Matrix

The problem with the co-occurrence based matrix is that they are dataset-specific. They may capture label relationships which are reflected in our dataset as evident from the improvements in the testset performance (see Table 1). We propose to use ontology which is a universal description of the label relations. In our paper, we consider the AudioSet ontology but our method can be extended to any other ontology. The AudioSet ontology is introduced in [9] which covers over 600 audio classes from human sounds to animal sounds, environmental sounds *etc.* FSDKaggle2019 dataset is labeled with a subset of the labels from this ontology. Let N be the total number of nodes (labels) in the AudioSet ontology out of which n number of labels are used to create such a dataset. To utilize ontological information for audio tagging, we propose the following approaches for creation of ontology-based correlation matrix:

Ontology-based Method One: In the first approach, we propose to create the correlation matrix $\mathbf{A}$ by the following procedure : Look for the labels at index of each row and each column of $\mathbf{A}$. Find the parent nodes of those labels from the ontology. If the two parents are the same, then $\mathbf{A}_{i,j} = 1$ else $\mathbf{A}_{i,j} = 0$, where i, j are row and column index of $\mathbf{A}$.

The motivation behind this approach is that sounds of similar classes tend to occur together and more dependent to each other. For example, musical sounds like flute sound and harmonium sound have greater probability of occurring together than sounds from different categories like sound of flute and the sound of car racing.

Ontology-based Method Two: In the second approach, We wish to utilize the entire ontology Graph. Thus we initialize the feature matrix $\mathbf{Z}$ for all the N labels (all the nodes in the ontology) and similarly the correlation matrix $\mathbf{A}$ is created. If two labels at indices i and j are connected by an edge in the Ontology, then $\mathbf{A}_{i,j} = 1$ else $\mathbf{A}_{i,j} = 0$. Since we only have ground truth for n labels of the dataset, we slices out only those n nodes representation from the N nodes representations we obtain after applying GCN.

3. EXPERIMENTS AND RESULTS

In our experiments, we used the FSDKaggle2019 Dataset introduced in [8] for the Task 2 of DCASE 2019 Challenge. The dataset is composed of two subsets: Curated subset, which is a small set of 4970 audio clips and Noisy subset, a larger set of 19815 audio clips. We divided this dataset into train-set for

Table 1. Comparisons of Per Class Lwlrap for top 20 classes (samples/class) in the dataset

Methods	M.fan	Bark	Ch.bell	Sciss.	Strum	Accord.	Ch.speech	Crowd	W.F.	Fill	Stream	Chirp_tweet	Knock	Waves_surf	Cutlery	Micro_oven	Hiss	Ele_guitar	Cupboard	Clapping
MTN	0.52	0.71	0.57	0.88	0.72	0.89	0.55	**0.76**	0.55	0.87	0.90	0.70	0.82	0.69	0.74	0.56	0.81	0.54	0.60	0.42
MT-GCN_1	0.53	0.77	0.60	0.90	0.76	0.85	**0.79**	0.62	0.38	**0.95**	0.94	0.66	0.76	0.58	0.71	0.63	0.75	0.50	0.49	**0.69**
MT-GCN_2	0.58	0.77	0.58	0.92	0.75	0.92	0.73	0.74	0.57	0.94	**1.**	**0.79**	0.80	0.66	0.68	0.56	0.81	0.49	**0.61**	0.62
MT-GCN_3	0.63	0.78	0.60	0.88	0.77	0.92	0.75	0.71	0.48	0.93	**1.**	0.70	0.80	0.71	0.68	0.55	**0.86**	0.44	0.53	0.67
MT-GCN_4	**0.66**	**0.83**	**0.61**	**0.94**	**0.79**	**0.93**	0.78	0.70	**0.66**	0.89	**1.**	0.75	**0.86**	**0.72**	**0.92**	**0.68**	0.82	**0.59**	0.40	0.68

Table 2. Comparison of Overall Lwlrap

Methods	Overall Lwlrap
MTN	0.6794
MT-GCN_1	0.6941
MT-GCN_2	0.7178
MT-GCN_3	0.7244
MT-GCN_4	**0.7405**

system development, val-set for hyper-parameter tuning and test-set for evaluation. We use *label-weighted label-ranking average precision* (lwlrap) as the evaluation metric for our method comparison [8]. This metric is introduced in the Task 2 of DCASE Challenge 2019.

3.1. Experimental Settings

We experimented our proposed framework in five Experimental settings:
1. Multitask Network (MTN). This works as a baseline method.
2. Co-occurrence based MT-GCN Method 1 (MT-GCN_1). The correlation matrix is derived from the curated dataset as described in Section 3.3.1.
3. Co-occurrence based MT-GCN Method 2 (MT-GCN_2). The correlation matrix is derived from the curated and the noisy dataset combined as described in Section 3.3.1
4. Ontology based MT-GCN Method 1 (MT-GCN_3). See method 1 in Section 3.3.2 .
5. Ontology based MT-GCN Method 2 (MT-GCN_4). See method 2 in Section 3.3.2

3.2. Results and Discussion

In the following paragraphs, we present and discuss our research findings. We apply model ensembles to reduce the randomness by training each model three times and taking the average of the Softmax probabilities as the final predictions on the test set. We then calculate the per class Lwlrap and overall Lwlrap for each model which is presented in Table 1 and Table 2, respectively.

As can be seen in Table 2, MT-GCN_1 and MT-GCN_2 which utilized labels co-occurrence based matrix derived from curated dataset and curated + noisy dataset outperformed the baseline method. It is expected because these methods use additional domain knowledge learned based on the label co-occurrences in the training dataset. Since the correlation matrix built on the curated dataset is sparser than the one built on the noisy dataset, it cannot capture as many co-occurrence patterns as captured by the correlation matrix built on the curated+noisy dataset. This is what our intuition is for the reason why method MT-GCN_2 performs better than method MT-GCN_1.

MT-GCN_3 and MT-GCN_4 which utilized the explicit domain knowledge representation of the audio dataset in the form of the ontology graph, outperformed all the previous approaches. Since MT-GCN_3 incorporates only a part of the ontology for finding the labels with similar parents rather than the entire ontology which is incorporated in MT-GCN_4, it brings less improvement over previous methods than that of the MT-GCN_4. The entire ontology captures the universal domain knowledge for the audio dataset while the co-occurrence based matrix could only capture the local label relationships. By local label relationships we mean that the label dependencies as mined from the available training dataset. This is the reason why MT-GCN_4 emerges as the top performing method for audio tagging with noisy labels.

A quite similar story is reflected in per class lwlrap scores (see Table 1) of the methods for the top 20 classes (according to samples/class). MT-GCN_4 outperformed the previous methods in the majority of the classes.

4. RELATED WORK

State-of-the-art multi-label audio tagging systems are based on deep learning approaches such as Convolutional Neural Networks (CNN) [14]. However, all the top performing submissions to DCASE 2018 [15] and DCASE 2019 [16] ignored the label relationship (ontology) associated with the dataset in their model design. A few efforts have been made on capturing the label relations for regularizing the network architectures for multi-label image recognition. For example, Chen *et al.* [13] proposed to learn label relations by learning a GCN from the occurrence patterns of the objects (labels) present in the image. To the best of our knowledge, our approach is the first attempt in the direction of using ontology to cop with label noise for audio tagging.

5. CONCLUSION

We presented our approach for audio tagging with noisy labels called MT-GCN which utilizes domain knowledge from the AudioSet ontology. We proposed two methods for building the ontology-based correlation matrix for GCN. Through experiments we have shown that our proposed method outperforms the baseline approaches in this task.

6. REFERENCES

[1] Yifang Yin, Meng-Jiun Chiou, Zhenguang Liu, Harsh Shrivastava, Rajiv Ratn Shah, and Roger Zimmermann, "Multi-level fusion based class-aware attention model for weakly labeled audio tagging," in *ACM International Conference on Multimedia*, 2019, pp. 1304–1312.

[2] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, "A survey of audio-based music classification and annotation," *IEEE Transactions on Multimedia*, vol. 13, no. 2, pp. 303–319, April 2011.

[3] Rainer Typke, Frans Wiering, and Remco Veltkamp, "A survey of music information retrieval systems," 01 2005, pp. 153–160.

[4] Ella Browning, Rory Gibb, Paul Glover-Kapfer, and Kate Jones, "Passive acoustic monitoring in ecology and conservation," 10 2017.

[5] Yifang Yin, Rajiv Ratn Shah, and Roger Zimmermann, "Learning and fusing multimodal deep features for acoustic scene categorization," in *ACM International Conference on Multimedia*, 2018, pp. 1892–1900.

[6] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *Advances in neural information processing systems*, 2018, pp. 8527–8537.

[7] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei, "Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels," *arXiv preprint arXiv:1712.05055*, 2017.

[8] Eduardo Fonseca, Manoj Plakal, Daniel P. W. Ellis, Frederic Font, Xavier Favory, and Xavier Serra, "Learning sound event classifiers from web audio with noisy labels," in *Proc. IEEE ICASSP 2019*, Brighton, UK, 2019.

[9] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[10] N. Muralidhar, M. R. Islam, M. Marwah, A. Karpatne, and N. Ramakrishnan, "Incorporating prior domain knowledge into deep neural networks," in *2018 IEEE International Conference on Big Data (Big Data)*, Dec 2018, pp. 36–45.

[11] Sandro Radovanović, Boris Delibašić, Miloš Jovanović, Milan Vukićević, and Milija Suknović, "Framework for integration of domain knowledge into logistic regression," in *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*, New York, NY, USA, 2018, WIMS '18, pp. 24:1–24:8, ACM.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[13] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo, "Multi-label image recognition with graph convolutional networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[14] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin W. Wilson, "CNN architectures for large-scale audio classification," *CoRR*, vol. abs/1609.09430, 2016.

[15] Il-Young Jeong and Hyungui Lim, "Audio tagging system for dcase 2018: Focusing on label noise, data augmentation and its efficient learning," Tech. Rep., DCASE2018 Challenge, September 2018.

[16] Osamu Akiyama and Junya Sato, "Multitask learning and semi-supervised learning with noisy data for audio tagging," Tech. Rep., DCASE2019 Challenge, June 2019.