INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Few-Shot Audio Classification with Attentional Graph Neural Networks

Shilei Zhang, Yong Qin, Kewei Sun, Yonghua Lin

IBM Research - China, Beijing 100193

{slzhang, qinyong, sunkewei, linyh}@cn.ibm.com

Abstract

Few-shot learning is a very promising and challenging field of machine learning as it aims to understand new concepts from very few labeled examples. In this paper, we propose attentional framework to extend recently proposed few-shot learning with graph neural network [1] in audio classification scenario. The objective of proposed attentional framework is to introduce a flexible framework to implement selectively concentration procedure on support examples for each query process. we also present an empirical study on confidence measure for few-shot learning application by combining posterior probability with normalized entropy of the network's probability output. The efficiency of the proposed method is demonstrated with experiments on balanced training set of Audio set for training and a 5-way test set composed of about 5-hour audio data for testing.

Index Terms: Few-shot learning, audio classification, attentional framework, confidence measures.

1. Introduction

The ability of machine learning approaches to learn effective features from vast datasets is proven, however, it is still very challenging to train a conventional end-to-end supervised model, such as deep learning models, from limited data or even very few examples. Few-shot classification aims to learn a classifier to recognize unseen classes during training with limited labeled examples, which is of great significance both academically and industrially. Most research work on few-shot learning focus on image process, natural language processing, etc. [1, 2, 3, 4, 5, 6, 7]. In this work, we introduce few-shot learning to audio classification application. Analysis of environmental audio sounds has been a popular topic which has the potential to be used in many applications, such as public security surveillance, smart homes, smart cars and healthcare monitoring, etc. In real application of audio classification [8, 9], there are in practice usually a large number of different categories but very few examples per category, few-shot classification models would help alleviate expensive data collection and labeling as they only require very limited training data to achieve reasonable performance. In contrast to standard classification problem, in few-shot learning, the classes can be split into training classes and test classes, while the training classes have sufficient classes and data for metric representation learning, and in test process, the novel or unseen test classes have only a few labeled data. A q -shot K -way classification in few-shot learning means that we have K novel classes and q examples in each novel class. For either training classes or test classes, the few-shot model is trained by support and query sets, where a support set is formed by sampling q examples from each of K classes and a query set is formed by sampling from the rest of the K classes' samples. The 5-shot 5-way audio classification process is illustrated in Figure 1.

Generally speaking, few-shot learning is an approach to classification that works with only a few human labeled examples. In standard few-shot learning, examples in support set are treated equally, as all examples equally contribute to the prediction result of query sample. Actually, for different test sample cases, different support examples maybe more important than others. For example, examples similar with the query case may be specified to have a greater weight and an increased contribution to the prediction in comparison to the irrelevant examples in support set. To achieve the above goal, we propose attentional mechanism to extend recently proposed few-shot learning with graph neural network [1] in audio classification scenario. First, we employ the Soundnet convolutional neural network [14] pre-trained on unsupervised video data to extract low-level acoustic features. Then we apply few-shot audio classification learning with attentional graph neural network, which predicts an attention vector to weight different examples according to their importance. Graph neural networks (GNNs) [10, 12, 13] are capable of processing arbitrary graphs, which have been used to model a variety of structural data. In [1], GNNs have also been used to model relationships between images for few-shot classification, which defined a graph neural network architecture that generalizes several of the typical few-shot learning models. Finally, confidence measure is computed for the real application by combining posterior probability with normalized entropy of the network's probability output.

The rest of the paper is organized as follows: In section 2 GNNs based few-shot learning as baseline method is briefly introduced. Section 3 describes the proposed few-shot audio classification with attentional graph neural networks. In section 4 experiments are presented and the results will be discussed. We will draw some conclusions in section 5.

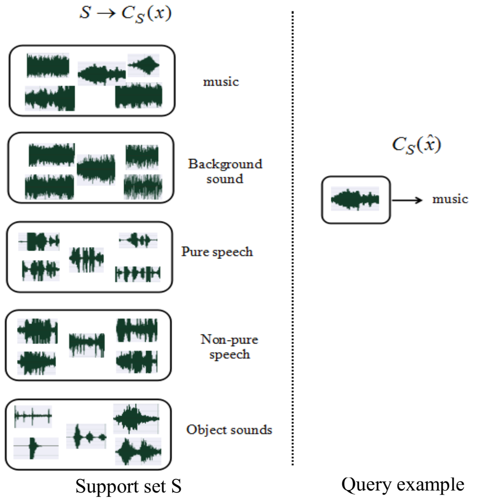

Figure 1: *5-way 5-shot audio classification*

http://dx.doi.org/10.21437/Interspeech.2019-1532

2. Few-shot learning with graph neural networks

We first formulate the few-shot learning problem following the definitions in previous works [1, 11]. In contrast to standard classification problem, the classes are split into two types in few-shot learning: the training classes that have sufficient classes and data for few-shot learning, while the novel or unseen test classes that have only a few labeled data. The success of few-shot learning approaches relies on effectively transferring the representation learned in the training classes to the novel test classes. In other words, the few-shot learning problem can be formulated as learning a metric function ϕ from training set, which can generalize to novel classes in the inference process. For q -shot, K -way classification learning, we are given a support set $S=\{(x_1,y_1),\cdots,(x_N,y_N)\}$, where N is the number of labeled samples, and the corresponding class label of the sample x_i is denoted by $y_i \in \{1,\cdots,K\}$, K is the number of classes. S_k denotes the set of examples labeled with the class k , where each label appears exactly q times. The ability of an algorithm to perform q -shot K -way few-shot learning is run as follows:
a) a few-shot classification model is given a query sample belonging to a new class and a support set consisting of q examples each from K different unseen new classes;
b) The algorithm determines which of the support set classes the query sample belongs to.

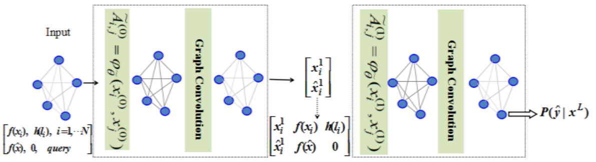
Input
$\begin{bmatrix} f(x_i), h(l_i), i=1,\cdots N \\ f(\hat{x}), 0, \ query \end{bmatrix}$
$\tilde{A}^{(0)}_{i,j} = \varphi_{\tilde{\theta}}(x_i^{(0)}, x_j^{(0)})$
Graph Convolution
$\begin{bmatrix} x_i^1 \\ \hat{x}_i^1 \end{bmatrix}$
$\begin{bmatrix} x_i^1 & f(x_i) & h(l_i) \\ \hat{x}_i^1 & f(\hat{x}) & 0 \end{bmatrix}$
$\tilde{A}^{(1)}_{i,j} = \varphi_{\tilde{\theta}}(x_i^{(1)}, x_j^{(1)})$
Graph Convolution
$P(\hat{y} \mid x^L)$

Figure 2: *Graph Neural Network illustration with 2 blocks*

For graph neural networks based few-shot learning, the architecture associate S with a fully-connected graph $G_S=(V,E)$ where nodes $v_a \in V$ correspond to the samples present in S , and E is the set of undirected edges. The algorithm is formally implemented by the following iterative procedure:
1) For each sample $x_i \in S$ with known label y_i , the one-hot encoding of the label is concatenated with the features of the sample at the input of the GNN, where the initial node feature vector is constructed by: $x_i^{(0)}=(f(x_i),h(l_i)))$, where $f(x_i)$ is the initial feature vector, and $h(l)$ is a one-hot encoding of the corresponding label. For testing sample $\hat{x}$ with unknown label, we set the label variable 0 for full uncertainty.
2) A graph is constructed by taking each sample as a node, and the adjacency matrix $\widetilde{A}^{(k)}$ is learned by:

$$\widetilde{A}^{(k)}_{i,j}=\varphi_{\widetilde{\theta}}(x_i^{(k)},x_j^{(k)})=MLP_{\widetilde{\theta}}(abs(x_i^{(k)},x_j^{(k)})) \qquad (1)$$

The trainable adjacency is then normalized by using a softmax along each row, then the generator family is updated by $A=\{\widetilde{A}^{(k)},1\}$.
3) The new features $x^{(k+1)} \in R^{V\times d_{k+1}}$ can be obtained by GNN projection layer $G_c(\cdot)$ with the signal $x^{(k)} \in R^{V\times d_k}$ and generator family as equation (2).

$$X^{(k+1)}=G_c(X^{(k)})=\sigma(\sum_{B\in A} BX^{(k)}\theta_B^{(k)}) \qquad (2)$$

where A is adjacency operator; $\theta_B^{(k)} \in R^{d_k \times d_{k+1}}$ are trainable parameters and $\sigma(\cdot)$ is active function, like leaky ReLU. X^{k+1} is the embedding matrix after $k+1$ layer.
4) Repeat steps 2) and 3) and get the final embedding feature.
5) The final layer of the GNN is thus a softmax mapping the node features to the the probabilities of the corresponding K classes. The cross-entropy loss is used as metric function for the graph training.

3. Few-shot audio classification with attentional graph neural networks

3.1. Motivation

Few-shot learning aims to solve sample efficiency issue imitating human intelligence and perception, such as a child could recognize an object after seeing one or several pictures. Sometimes, human makes classification decision through selective input samples, i.e. focus on part of the effective samples in the recognition process.

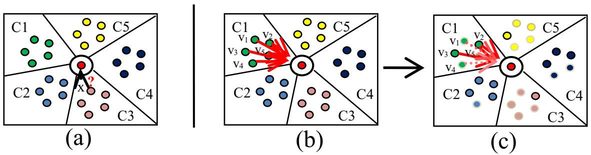

Figure 3: *motivation and conceptual illustration of our method. (a) Examples of the 5-way 5-shot learning (b) the elements of the support set equally contribute to predict the class label of test sample $\hat{x}$; (c) For few-shot learning problem, we developed end-to-end architecture to actively weight those support sets as input via attention mechanisms.*

In this work, we apply attention mechanism to actively select the samples of support set for effective few-shot learning. Attention mechanism is an effective means of guiding the model to focus on a partial set of most relevant features for each input instance. An illustration of the motivation for attentional network relates to the few-shot procedure is shown in Figure 1. It works by generating weight coefficients for the given features in an input-adaptive manner, to allocate more weights to the features that are found to be relevant for the given input. As shown in b) and c) in Figure 1, for the specific class $C1$ and the corresponding examples $\{v1,\cdots,v5\}$,the elements of the support set equally contribute to predict the class label of test sample in the standard few-shot learning process, while our proposed method actively weight those support sets as input via attention mechanisms, such as high weight for $v2,v3$, relatively low weight for $v1,v4,v5$ for the testing case.

3.2. Few-shot audio classification

3.2.1. Acoustic features extraction

We explored Soundnet convolutional neural networks with 8 convolution layers [14] to directly extract the acoustic representation from the raw audio waveforms. As a 1D convolution network, Soundnet can extract local audio features effectively. It consists of full convolutional layers and pooling layers as shown in Figure 4, so it could deal with variable length audio waves. The features extracted from middle layers of the Soundnet achieve significant improvement over the traditional audio features on the audio event recognition task [14]. Since convolution networks have great generalization ability, in this work, we extract features from the conv5 layer of Soundnet directly as the short-time acoustic features. The mean of all conv5 features are used as the feature representation vector of each segment example.

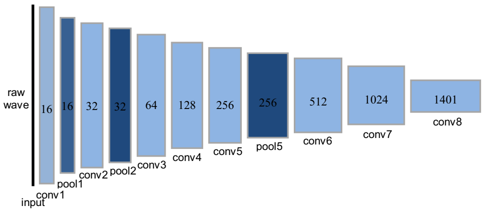

Figure 4: *The structure of the Soundnet. The number inside of the layer is the number of filters.*

3.2.2. *Attentional graph neural networks learning*

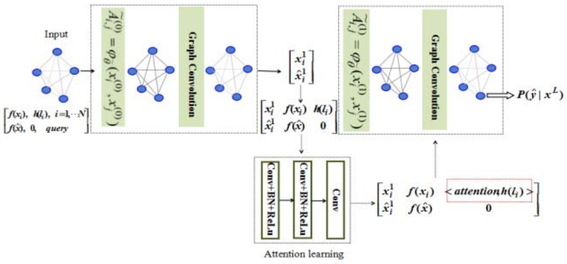
Input
$\tilde{A}^{(0)}_{i,j} = \varphi_{\tilde{\theta}}(x^{(0)}_i, x^{(0)}_j)$
Graph Convolution
$\begin{bmatrix} f(x_i), & h(l_i), & i=1,\cdots N \\ f(\hat{x}), & 0, & query \end{bmatrix}$
$\begin{bmatrix} x^1_i \\ \hat{x}^1_i \end{bmatrix}$
$\begin{bmatrix} x^1_i & f(x_i) & h(l_i) \\ \hat{x}^1_i & f(\hat{x}) & 0 \end{bmatrix}$
$\tilde{A}^{(1)}_{i,j} = \varphi_{\tilde{\theta}}(x^{(1)}_i, x^{(1)}_j)$
Graph Convolution
$P(\hat{y} \mid x^L)$
Conv+BN+ReLu
Conv+BN+ReLu
Conv
$\begin{bmatrix} x^1_i & f(x_i) & < attention, h(l_i) > \\ \hat{x}^1_i & f(\hat{x}) & 0 \end{bmatrix}$
Attention learning

Figure 5: *Attentional graph neural network illustration with 2 blocks*

The standard graph neural networks for few-shot learning assume all instances in the support set contribute equally and independently to the query sample's label. The attentional graph neural network is a generalization, where a different weight scalar value is allowed for each example. The architecture asserts that each example contributes independently but not necessarily equally to the prediction of query example. Similar to human perception process, the attention mechanism in our model highly weight the input samples which are more contribute to classification while ignoring the unrelated samples. As shown in Figure 5, The attentional weight is computed after the first layer of the GNN by using a softmax attention over all nodes of the graph called global attention or the nodes from each separate class called intra-class attention. Here, the attention mechanism weights the input samples, then update the feature representation. Proposed global attention and intra-class attention are described as follows:

a) Global attention: For attentional learning, similar with active learning formulation in [1], a function $g(x_i)$ parametrized by a three layer neural network is applied to project each vector node to a scalar value, then attention weights can be obtained after applying softmax over these scalar values:

$$Attention = Soft\max(g(x^{(1)}_{\{1,\cdots,N\}})) \qquad (3)$$

b) Intra-class attention: In contrast to global attention, we can also apply the attentional learning into every class S_k ,

$$Attention_k = Soft\max(g((x^{(1)}_{\{1,\cdots,q\}}, y_i = k))) \quad k \in \{1,\cdots,K\} \quad (4)$$

Then attentional class label values are given by multiplying the above attention weights by the one-hot label vectors, and the current representation $x_i^{(k+1)}$ is further updated by

$$x_i^{(k+1)} = [G_c(x_i^k), x_i^{(k)}] = [G_c(x_i^k), (f(x_i), < attention, h(l_i) >)] \quad (5)$$

Next, the attentional information will be fed into the following layer as input feature, and the attentional GNNs can be trained end-to-end by back propagation to more effectively learn few-shot classification metric.

3.2.3. *Confidence measures*

The field of few-shot learning is making fast progress but there is still a gap between performance and practice application requirement on the challenging task. Confidence measures (CM) [15, 16, 17, 18, 25] for few-shot classification are used to evaluate reliability of classification results which does not directly try to improve the classification accuracy, but detects incorrect classifications to increase the usability of practical user-friendly systems. It is straight-forward that the posterior probability output corresponding to prediction class from graph neural network is a good candidate for CM since it is an absolute measure of how well the classification result is. As the limited classification performance in few-shot classification scenario, we present an empirical study and analysis on further confidence score improvement in *section 4.2.2* to satisfy the requirement of real application. From the experiment, we can see another confidence score that takes the probability distribution into account is the normalized entropy of the network's probability output:

$$H(\hat{x}) = (-\sum_{i=1}^{K} p_i \log p_i) / \log K \qquad (6)$$

Implicit to this measure is the belief that as approximated posterior class probabilities get closer, mis-classifications are more probable. Since this confidence score can provide complementary information, better performance will be achieved if linearly combining normalized entropy into posterior probability as equation (7), which is adequate for achieving robust estimations of CM.

$$CM = \alpha CM_{pp} + (1-\alpha)(1-H(\hat{x})) \qquad (7)$$

where α is the hybrid CM weight factor, and it is set to 0.8 in our experiment. Initially, we also evaluate the feasibility of entropy of attention weight as confidence measure as shown in *section 4.2.2*, but did not achieve the desired effect.

4. Experimental results

4.1. Dataset

Two datasets were used in the experiments: a) balanced training set of Audio Set dataset [19] for training; Audio Set contains over 2 million 10 seconds audio clips extracted from YouTube videos, which consists of 527 classes of audio with a hierarchy structure. In the balanced evaluation training sets, 22,176 segments from distinct videos chosen each class to have the same number of examples. Audio Set is a large scale weakly labeled dataset of sound clips, and is defined for tasks such as audio tagging to perform multi-label classification on fixed-length audio chunks. Because single segments can have multiple labels (on average 2.7 labels per segment), we only choose the first label as class tag for few-shot learning training. As the multiple labels for one segment, Audio set is not feasible to be partly selected as test set for few-shot learning experiment. b) a 5-way test set [20] collected from real-life scenario for testing, which are about 5 hours in total and 1 hour of each class including: pure speech, speech with various background noises, music sound, environment noise and animal sounds. For testing process, the audio signal of each class is uniformly segmented into nonoverlapping 10s long clips.

In our experiment, the few-shot classification model are trained from available large amount of audio data and classes, and transfer the knowledge and metric to the classification of 5-way new unseen classes, which is to mimic the requirement in real application scenarios.

4.2. Evaluation

4.2.1. Few-shot audio classification performance

Table 1: *test set classification accuracy with 95% confidence intervals.*

accuracy(%)	*5-way*		
	1-shot	*5-shot*	*10-shot*
GNN	67.3±0.92	73.4±0.73	80.6±0.75
+global attentional GNN	69.4±0.66	78.3±0.46	83.6±0.98
+intra-class attentional GNN	/	76.0±0.64	81.6±0.86

We evaluate our models by performing different typical q-shot, 5-way experiments on the dataset. For every few-shot task, we sample 5 random classes from the training set during training and employ 5-way classes from test set during testing, and from each class we sample 1, 5, 10 random samples, respectively. An extra sample to classify is chosen from one of that 5 classes. We train at most 6000 iterations for each model and the model that achieves the best performance using 10-fold cross validation is used as the final training model. 15000 times random tests on test set are performed for accuracy performance evaluation.

In this study, we employ the open source software by [21] to provide baseline setup, where three blocks are used in this work with parameter *nf*=96. For attentional learning, a network is used consisting of three one-dimensional convolutions layers followed by a softmax layer resulting in a N dimensional attention weight, where N is the number of support samples. The experimental results are presented at Table 1, we can see graph neural networks can obtain reasonable accuracy in few-shot audio classification scenario; and our proposed attentional framework can further improve the performance by extending the model more power on data sample weighing. As the more flexible attention capability, global attentional GNN can achieve the better performance compared with intra-class attentional GNN. Further, experimental results demonstrate the proposed framework's robustness on different q -shot cases, on the other hand, 5-shot case get better relative accuracy improvement than that of 1-shot and 10-shot cases, showing moderate size of support examples is an important factor to explore capability of the proposed framework.

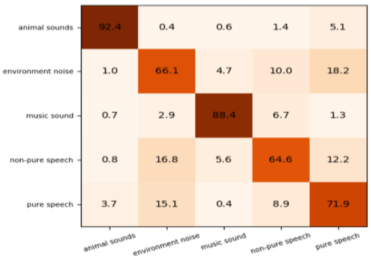

Figure 6: *Confusion matrix of 5-way 5-shot test-set prediction based on intra-class attentional GNN.*

We also evaluated intra-class attentional GNN on the test set and computed the fraction of labels of a certain class with respect to their predictions shown in this color-coded confusion matrix representation. As shown in Figure 6, animal sounds and music sounds were predicted to be from the correct class with probabilities larger than 88%, while environment noise, non-pure speech and pure speech are easily miss-classified from each other. Generally speaking, balanced accuracy can be achieved via the proposed framework for few-shot audio classification application.

4.2.2. Confidence measures

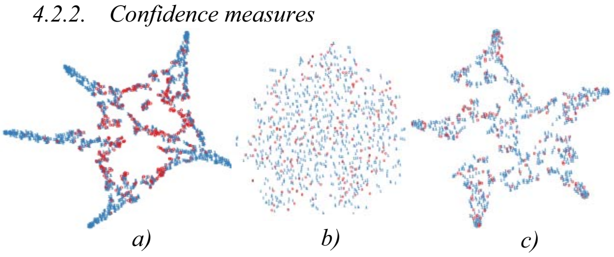

Figure 7: *t-SNE projections of a) probability output vector; b) global attention weight; c) intra-class attention weight of predicted class; on 5-way 5-shot classification, labeled by "1" for correct classification, and "0" for wrong classification.*

In this section, we further use t-SNE [22] to project probability output vector representations and attention weights onto a 2-d space for visual comparisons and CM analysis. We employ the scikit-learn [23] t-SNE implementation with all parameters kept default. As shown in Figure 7 a), the probability output distributions differ between correctly and incorrectly classified examples, and dominant class of distribution is inclined to correct classification result, while uniform distribution usually lead to incorrect classification case. So we can see entropy of probability output is a good candidate of CM score. As it is difficult to compute multi-class AUC score alone applying entropy based CM score, we compute and compare the baseline posterior probability CM and proposed hybrid CM in this paper, respectively. By our experimental result as shown in Table 2, evaluated by the balanced average across all classes of AUC [24], entropy score derived from probability output distributions provide an additional information to alone softmax classifier posterior probabilities, and hybrid CM are effective score to be translated to reliable confidence measures.

Table 2: *Performance comparison in confidence measures*

AUC	*5-way 5shot*
CM$_{pp}$	0.9146
CM_hybrid	0.9388

We also analyze the distribution of the attention weight, however, from b) and c) in Figure 7, either global attention weight or intra-class attention weight can not provide valuable knowledge for confidence measures to distinguish between correctly and incorrectly classified examples.

5. Conclusions

Few-shot learning is a very important field of machine learning for many real applications. In this paper, we proposed attentional graph-based models, and the experimental results demonstrate the ability of the framework to operate well on audio classification tasks. The proposed few-shot audio classification with attentional GNNs effectively transfers the metric representation learned in the training class to the novel classes. Further, the suggested hybrid confidence measures can improve the AUC performance and helps us to develop user-friendly and robust few-shot classification application system.

6. References

[1] Garcia, Victor and Bruna, Joan, "Few-shot learning with graph neural networks," *In Proceedings of the International Conference on Learning Representations*, 2018.

[2] Mo Yu, Xiaoxiao Guo, Jinfeng Yi, Shiyu Chang, Saloni Potdar, Yu Cheng, Gerald Tesauro, Haoyu Wang, and Bowen Zhou, "Diverse few-shot text classification with multiple metrics," *arXiv:1805.07513*, 2018.

[3] Eli Schwartz, Leonid Karlinsky, et al., "Delta-encoder: an effective sample synthesis method for few-shot object recognition", *NIPS,* 2018.

[4] Jake Snell, Kevin Swersky and Richard S. Zemel, "Prototypical Networks for Few Shot Learning," *arXiv: 1703.05175*, Jun 2017.

[5] Oriol Vinyals, Charles Blundell, et al., "Matching Networks for One Shot Learning," *arXiv:1606.04080v2*, Dec 2017.

[6] Hang Gao, Zheng Shou, et al., "Low-shot Learning via Covariance-Preserving. Adversarial Augmentation Networks," *arXiv:1606.04080v2*, Dec 2017.

[7] YongWang, Xiao-Ming Wu and Qimai Li, "Large Margin Few-Shot Learning," *arXiv:1807.02872v2*, Sep 2018.

[8] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold et al., "CNN architectures for large-scale audio classification," *arXiv preprint arXiv:1609.09430*, 2016.

[9] Qiuqiang Kong, Yong Xu, Wenwu Wang and Mark D. Plumbley, "Audio Set classification with attention model: A probabilistic perspective," *arXiv:1711.00927v2*, Feb 2018.

[10] P.W. Battaglia, J.B. Hamrick, et al., "Relational inductive biases, deep learning, and graph networks," *arXiv:1806.01261*, June 2018.

[11] M. Guo, E. Chou, D. -A. Huang, S. Song, S. Yeung, L. Fei-Fei, "Neural Graph Matching Networks for Fewshot 3D Action Recognition," *ECCV*, 2018.

[12] Zhou Jie, et al., "Graph neural networks: a review of methods and applications," *arXiv:1812.08434v2*, 2019.

[13] Thomas N. Kipf and Max Welling, "Semi-supervised classification with graph convolutional networks," *arXiv: 1609.02907v4*, 2017.

[14] Yusuf Aytar, Carl Vondrick, and Antonio Torralba, "Soundnet: Learning sound representations from unlabeled video," *In Advances in Neural Information Processing Systems*, pp. 892-900, 2016.

[15] Gethin Williams and Steve Renals, "Confidence measures from local posterior probability estimates," *Computer Speech & Language*, 1999.

[16] Hui Jiang, "Confidence measures for speech recognition: A survey," *Speech Communication*, pp. 455–470, 2005.

[17] Changhao Shan, Junbo Zhang, Yujun Wang, Lei Xie, "Attention-based End-to-End Models for Small-Footprint Keyword Spotting," *INTERSPEECH*, 2018

[18] Shilei Zhang, Danning Jiang, Yong Qin, "Utterance verification using improved confidence measures based on alignment confusion rate in Chinese digits recognition," *ICASSP*, pp.1309-1312, 2009.

[19] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," *ICASSP*, pp. 776–780, 2017.

[20] Shilei Zhang, Hongchen Jiang, Shuwu Zhang, Bo Xu, "Fast SVM training based on the choice of effective samples for audio classification," *INTERSPEECH*, pp. 1654-1657, 2006.

[21] *https://github.com/vgsatorras/few-shot-gnn.*

[22] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, pp. 2579–2605, 2008.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, pp. 2825–2830, 2011.

[24] T. Fawcett, "Roc graphs: Notes and practical considerations for researchers," *Machine learning*, vol. 31, no. 1, pp. 1–38, 2004.

[25] Shilei Zhang, Zhiwei Shuang, Qin Shi, Yong Qin, "Improved Mandarin Keyword Spotting Using Confusion Garbage Model," *ICPR*, pp. 3700-3703, 2010.