Graph Convolutional Network Based Semi-Supervised Learning on Multi-Speaker Meeting Data

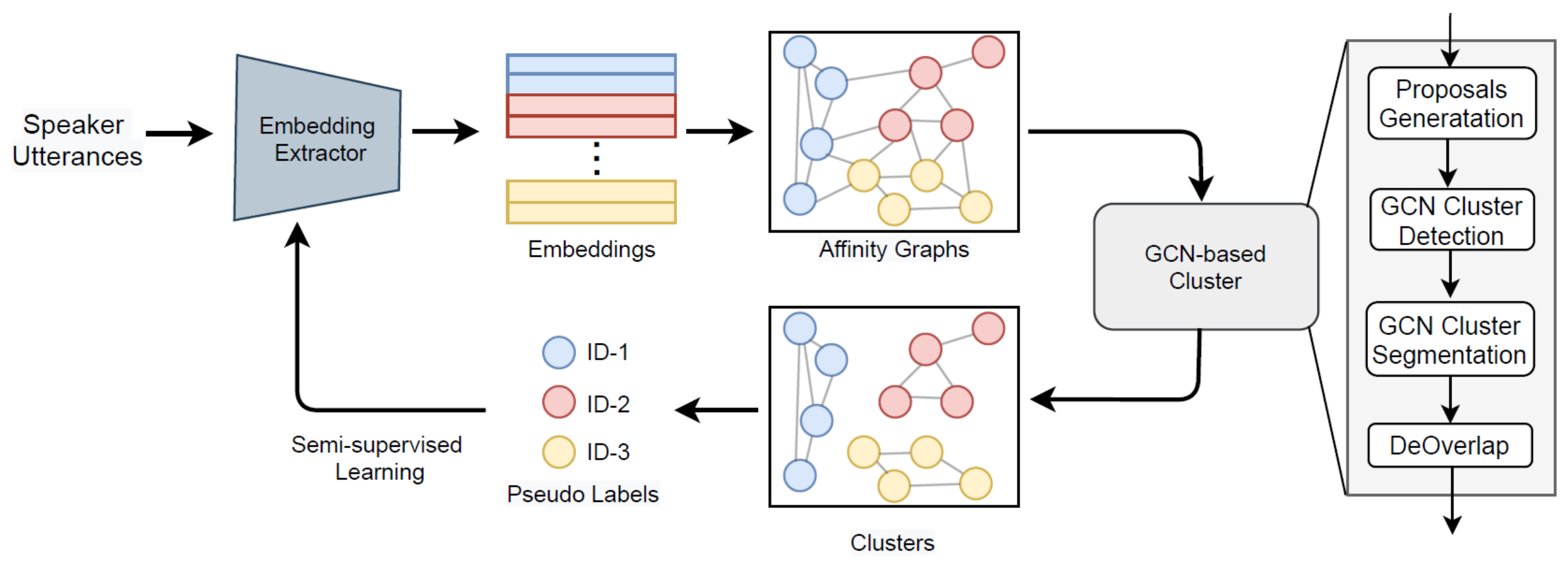
Fuchuan Tong, Siqi Zheng, Min Zhang, Yafeng Chen,
Hongbin Suo, Qingyang Hong, Lin Li

3490

Abstract

Unsupervised clustering on speakers is becoming increasingly important for its potential uses in semi-supervised learning. In reality, we are often presented with enormous amounts of unlabeled data from multi-party meetings and discussions. An effective unsupervised clustering approach would allow us to significantly increase the amount of training data without additional costs for annotations. Recently, methods based on graph convolutional networks (GCN) have received growing attention for unsupervised clustering, as these methods exploit the connectivity patterns between nodes to improve learning performance. In this work, we present a GCN-based approach for semi-supervised learning. Given a pre-trained embedding extractor, a graph convolutional network is trained on the labeled data and clusters unlabeled data with “pseudo-labels”. We present a self-correcting training mechanism that iteratively runs the cluster-train-correct process on pseudo-labels. We show that this proposed approach effectively uses unlabeled data and improves speaker recognition accuracy.

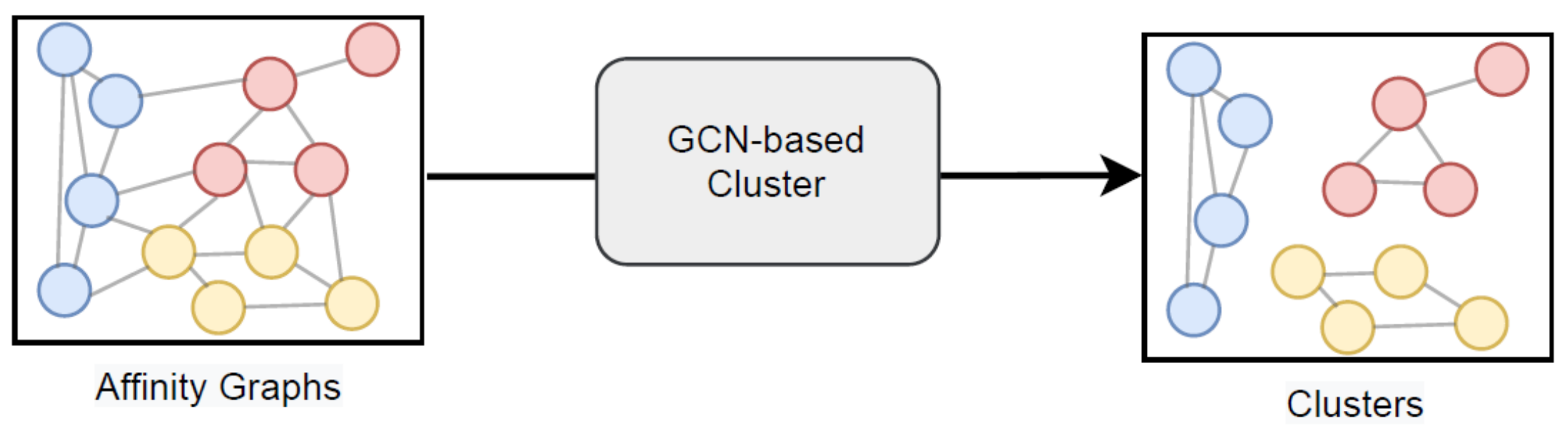
The Semi-Supervised Speaker Recognition Pipeline



- ➢ Utterances are first fed into feature extractors to obtain speaker embeddings.
- ➢ An affinity graph is constructed to perform clustering.
- ➢ The cluster results with pseudo-labels are applied to re-training the deep speaker embedding extractor.

GCN-based Semi-Supervised Learning

● GCN-based Clustering



- ➢ STEP1: Affinity Graph Construction
- ➢ STEP2: Cluster Proposal Generation
- ➢ STEP3: Cluster Detection
- ➢ STEP4: Cluster Segmentation
- ➢ STEP5: De-Overlapping

● Label Noise Optimization

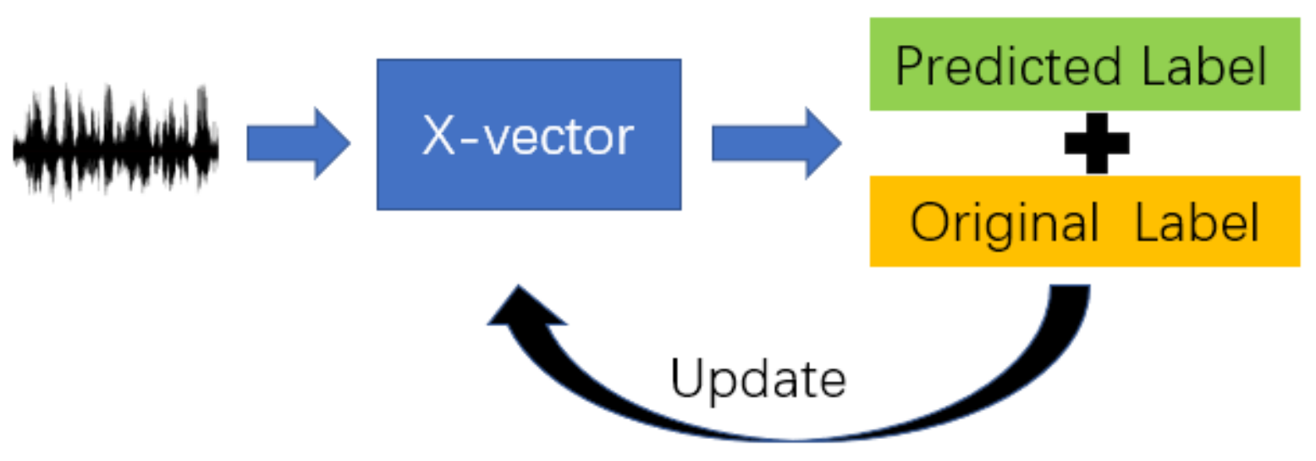
1. In the retraining processing, we apply the Label Noise Correction Loss:

$$L' = -\frac{1}{B}\sum_{i=1}^{B}\{(1-\alpha_t)\log(P_{i,y_i}) + \alpha_t\log(P_{i,\hat{y}_i})\}$$

2. We introduce "sub-class" centers [3] into the loss function:

```
AM Softmax ==> sub center AM-Softmax(Sub-AM)
```

3. During training, noisy labels are corrected on-the-fly based on the network’s predictions.



Datasets and Experimental Results

- We use VoxCeleb1 to represented a small amount of labeled data and use VoxCeleb2 to represented large amounts of unlabeled meeting data.
- We train the speaker network on a labeled VoxCeleb1 set to extract speaker embedding.
- In speaker clustering experiments, we randomly select three test groups from the VoxCeleb1 test set with three, six, and nine speakers.
- In semi-supervised speaker recognition experiments, the VoxCeleb2 development set are randomly shuffled and sampled into 666 meetings without overlapped identities. The number of speakers in the meeting data ranges from two to ten.

Comparison of speaker clustering

Table 2: Comparison of speaker clustering when the number of clusters is 3, 6, and 9. The results are the average of the clustering results on 10 different sets of testing data.

#	Methods	Precision	Recall	F-score
3	K-means	0.80	0.52	0.63
	SC	0.76	0.68	0.71
	AHC	0.75	0.77	0.75
	GCN	**0.82**	**0.79**	**0.80**
6	K-means	0.78	0.56	0.65
	SC	0.71	0.65	0.67
	AHC	0.77	**0.79**	0.78
	GCN	**0.84**	0.78	**0.81**
9	K-means	0.77	0.53	0.63
	SC	0.73	0.66	0.69
	AHC	0.82	0.76	0.78
	GCN	**0.85**	**0.80**	**0.82**

Performance comparisons of semi-supervised learning-based speaker recognition

Table 3: Performance comparisons of clustering and speaker recognition results using models trained with different clustering pseudo-labels. The * symbol indicates that label de-noising was employed.

Model	Precision	Recall	F-score	EER	minDCF
Baseline	-	-	-	3.34	0.384
+ K-means	0.78	0.54	0.64	2.04	0.255
+ SC	0.74	0.67	0.70	1.73	0.213
+ AHC	0.79	0.77	0.77	1.51	0.186
+ GCN	0.83	0.79	0.81	1.43	0.174
+ GCN*	0.83	0.79	0.81	**1.30**	**0.152**
Oracle	-	-	-	1.28	0.165

Summary

- We provide a novel approach to improve speaker recognition by leveraging large amounts of unlabeled data. Pseudo-labels are generated from cluster predictions on unlabeled data.
- Experimental results show that GCN-based clustering outperforms the existing clustering methods, and the results demonstrate its effectiveness in semi-supervised learning.
- When combining the clustering method with label de-noising processing, this system achieves comparable results compared to fully-supervised training on the Voxceleb1 and Voxceleb2 datasets.
- We conclude that the GCN-based clustering method is an effective method to provide insights into the practice of speaker recognition with unlabeled data.

Acknowledgement

This research is funded by the National Natural Science Foundation of China (Grant No. 61876160 and No. 62001405) and Fundamental Research Funds for the Central Universities (No. 20720210087).

Contact

Lin Li Email: lilin@xmu.edu.cn
Qingyang Hong Email: qyhong@xmu.edu.cn
Website: https://speech.xmu.edu.cn

References

1. L. Yang, X. Zhan, D. Chen, J. Yan, C. C. Loy, and D. Lin, “Learning to cluster faces on an affinity graph,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2298–2306.

2. F. Tong, Y. Liu, S. Li, J.Wang, L. Li, and Q. Hong, “Automatic error correction for speaker embedding learning with noisy labels,” in Proc. INTERSPEECH, 2021, pp. 4628–4632