Siamese Neural Networks for Class Activity Detection

Hang Li[1], Zhiwei Wang[2], Jiliang Tang[2], Wenbiao Ding[1], and Zitao Liu[1(✉)]

[1] TAL Education Group, Beijing, China
{lihang4,dingwenbiao,liuzitao}@100tal.com
[2] Data Science and Engineering Lab, Michigan State University, East Lansing, USA
{wangzh65,tangjili}@msu.edu

Abstract. Classroom activity detection (CAD) aims at accurately recognizing speaker roles (either teacher or student) in classrooms. A CAD solution helps teachers get instant feedback on their pedagogical instructions. However, CAD is very challenging because (1) classroom conversations contain many conversational turn-taking overlaps between teachers and students; (2) the CAD model needs to be generalized well enough for different teachers and students; and (3) classroom recordings may be very noisy and low-quality. In this work, we address the above challenges by building a Siamese neural framework to automatically identify teacher and student utterances from classroom recordings. The proposed model is evaluated on real-world educational datasets. The results demonstrate that (1) our approach is superior on the prediction tasks for both online and offline classroom environments; and (2) our framework exhibits robustness and generalization ability on new teachers (i.e., teachers never appear in training data).

Keywords: Multimodal learning · Neural networks · Class activity detection

1 Introduction

It is essential to equip instructor training with informative dialogic feedback on their classroom activities, which allows teachers to adjust and refine their teaching instructions [1,4,10,17,24]. Prior researches have been demonstrated that pedagogical teaching styles and instructions may significantly influence students' engagements and academic achievements [18,22,26]. Traditionally, providing such feedback is very logistically complex and expensive, as it heavily relies on human annotations [3,14,20,21]. This makes it inapplicable in real-world education scenarios. Thus, in this work, we focus on building an automatic AI driven solution to solve this fundamental class activity detection (CAD) problem. More specifically, we aim at automatically annotating classroom audio recordings by

Z. Wang—Work was done when the authors did internship in TAL Education Group.

I. I. Bittencourt et al. (Eds.): AIED 2020, LNAI 12164, pp. 162–167, 2020.
https://doi.org/10.1007/978-3-030-52240-7_30

recognizing different speakers' roles, i.e., student or teacher. CAD solutions produce basic information about the quantities and distributions of classroom conversations, which are one of the essential steps for deep classroom analysis [16].

A large spectrum of models have been developed to solving the CAD problem [2,6,8,22]. Owens et al. proposed a machine learning algorithm that captures distinctive patterns in different instructional techniques and classifies the classroom sound into different class activities [22]. Cosbey et al. targeted on the same classroom sound classification problem as in [22] and adopted deep recurrent neural networks to extract meaningful features from audio frames [6]. Wang et al. conducted CAD by using LENA system [11] and identified three discourse activities of teacher lecturing, class discussion and student group work [30].

However, CAD in real-world scenarios is still extremely difficult because of three challenges: (1) *conversational turn-taking overlap*: Classroom conversations usually contain many frequent talk exchanges between teachers and students, which leads to a number of inextricable speech overlaps; (2) *vocal variability and uniqueness*: Every person's voice is different and unique, which poses a difficult question on the generalization ability of the CAD solution; and (3) *classroom noise*: Both online and offline classrooms in reality are dynamic, complex and noisy. In the attempt to solve the aforementioned challenges, we develop the Siamese neural framework to precisely detect teacher and student activities from classroom audio recordings. The contributions of this work are summarized as follows: (1) It presents a pioneer research on the CAD problem and proposes a novel Siamese neural framework to tackle this problem; and (2) we comprehensively evaluate our framework with different realizations and their benefits on both online and offline real-world, large-scale classroom datasets.

2 The Siamese Neural Framework

In this section, we describe our end-to-end Siamese neural framework for the CAD problem in details. Our framework consists of three key components: (1) feature extraction module that extracts window-level raw embeddings from a pre-train large-scale audio encoding neural network; (2) the representation learning module, which extracts semantic representations from each classroom audio segment; and (3) an attentional prediction module that predicts the activity type for each window. The overall framework architecture is shown in Fig. 1.

Feature Extraction. We first utilize a well-studied voice activity detection (VAD) system to segment audio streams into pieces of utterances and filter out the noisy and silent ones [23,25,27]. Then we transform each segment into frames of pre-defined width and step, and log-mel-filterbank energies of dimension 40 are extracted from each frame. After that, we obtain windows by using non-overlapping sliding windows of a fixed length on these frames. Once we create these audio windows from both teachers' vocal sample segments and classroom recording segments, we extract windows' corresponding low-dimensional dense vocal representations from a pre-trained acoustic neural network.

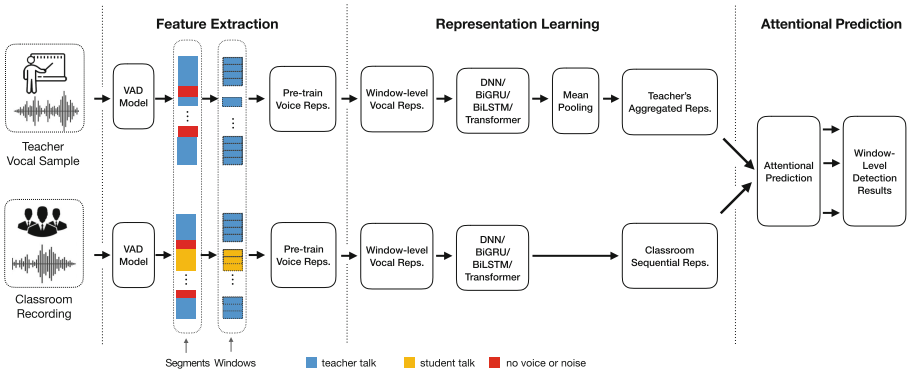

Fig. 1. The overview of our Siamese neural framework. VAD is short for voice activity detection.

Representation Learning. We learn a refined vocal representation for each window by utilizing the contextual dependencies within each segment (either from teachers' vocal samples or classroom recordings). In our framework, any existing sequential modeling function such as long short-term memory (LSTM), gated recurrent unit (GRU), etc. can be used [7,13,29]. By considering the contextual windows across entire segment, we are able to model the changes of tones and pitches in the audio stream smoothly and reduce the noises and outliers in the raw feature extraction component.

Attentional Prediction. We design an attentional prediction module focusing on the window-level class activity detection tasks. Our attentional prediction module is inspired by the intuition that all the audio windows spoken by the teacher share common attributes that are very different from those shared from student's audio windows. Thus, we use teachers' vocal samples as an aggregated query and compute an attention score with each individual window from classroom recordings. The higher the attentional score is, the more likely the audio window is spoken by the teacher. Based on this idea, we first add a mean pooling layer to aggregate all the teacher's vocal sample representations. This yields a robust and representative query embedding of the teacher's voice signals. The obtained vector is used as a voice biometrics query to compute attention scores with each individual window representation. In order to effectively train our framework, we design a cross-entropy loss function as the optimization objective. We use mini-batch stochastic gradient decent algorithm to minimize the objective and update the our model parameters.

3 Experiments

We evaluate our framework with two real-world K-12 education datasets: (1) the *online* dataset, which includes 400 classroom recordings and 300 distinct teachers

from a third-party online education platform[1]; and (2) the *offline* dataset that includes 100 recordings and 36 distinct teachers from physical offline classrooms. We randomly select 100 and 10 recordings from *online* and *offline* dataset respectively as our test sets. The prediction results are denoted as "Main". Moreover, in order to evaluate the model generalization ability to new teachers, we further filter out teachers from above test set if the teachers appear in the training set and the prediction results are denoted as "Generalization". We choose to use area under curve (AUC) score to evaluate the model performance [9].

We choose the following approaches as our baselines: (1) *Average*: Vocal representations from feature extraction component are directly used for attentional prediction; (2) *DNN/GRU/LSTM*: A single layer fully connected neural network/a bidirectional GRU/a bidirectional LSTM is used in the representation learning component [5,12,19]. We use 128 neurons and ReLU as the activation function; and (3) *Transformer*: A transformer is used in the representation learning component [28]. We choose to use 2 layers in the transformer and set 4 heads for each layer. We set the dimension of each head to 16.

Experimental Results: The results are shown in Table 1. For the main task, we find that (1) the *Average* performs much worse than any other method. This suggests that the fine-tuned representation learning plays an important role in the final prediction; (2) compared to *GRU*, *LSTM*, and *Transformer*, *DNN* has achieve a lower detection accuracy. This is expected as it is not able to capture the contextual information of windows within each segment; (3) the performance of all methods on *online* dataset is generally better than results on *offline* dataset. We argue that this is because the signal to noise ratio of offline recordings is much higher than the ratio in online recordings [16]; and (4) both *GRU* and *Transformer* have comparable performance, which is consistent with the previous findings [15]. For the generalization task, we have similar observations. The high accuracy achieved by Transformer and LSTM demonstrates the generalization ability of the proposed framework.

Table 1. Experimental results on the *online* and *offline* datasets.

Task	Dataset	Average	DNN	GRU	LSTM	Transformer
Main	Online	0.895	0.926	0.936	0.933	0.942
	Offline	0.713	0.810	0.881	0.858	0.858
Generalization	Online	0.895	0.922	0.932	0.931	0.937
	Offline	0.749	0.840	0.880	0.805	0.882

[1] https://www.xes1v1.com/.

4 Conclusion

We present a Siamese framework to tackle the CAD problem. Experiments demonstrate both detection performance and generalization ability of our framework. In the future, we would like to design models that can combine both audio and video data to generate more comprehensive classroom activity feedback.

Acknowledgements. Zhiwei Wang and Jiliang Tang are supported by the National Science Foundation of United States under IIS1714741, IIS1715940, IIS1715940, IIS1845081 and IIS1907704.

References

1. Akalin, S., Sucuoglu, B.: Effects of classroom management intervention based on teacher training and performance feedback on outcomes of teacher-student dyads in inclusive classrooms. Educ. Sci. Theory Pract. **15**(3), 739–758 (2015)
2. Bergman, D.: Comparing the effects of classroom audio-recording and video-recording on preservice teachers' reflection of practice. Teacher Educator **50**(2), 127–144 (2015)
3. Brinko, K.T.: The practice of giving feedback to improve teaching: what is effective? J. Higher Educ. **64**(5), 574–593 (1993)
4. Chen, J., Li, H., Wang, W., Ding, W., Huang, G.Y., Liu, Z.: A multimodal alerting system for online class quality assurance. In: Isotani, S., Millan, E., Ogan, A., Hastings, P., McLaren, B., Luckin, R. (eds) International Conference on Artificial Intelligence in Education, pp. 381–385. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-23207-8_70
5. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
6. Cosbey, R., Wusterbarth, A., Hutchinson, B.: Deep learning for classroom activity detection from audio. In: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3727–3731. IEEE (2019)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
8. Donnelly, P.J., et al.: Automatic teacher modeling from live classroom audio. In: Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization, pp. 45–53 (2016)
9. Fawcett, T.: An introduction to ROC analysis. Pattern Recogn. Lett. **27**(8), 861–874 (2006)
10. Freiberg, H.J., Waxman, H.C.: Alternative feedback approaches for improving student teachers' classroom instruction. J. Teacher Educ. **39**(4), 8–14 (1988)
11. Ganek, H., Eriks-Brophy, A.: The language environment analysis (LENA) system: a literature review. In: Proceedings of the Joint Workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC, Umeå, 16th November 2016, pp. 24–32. No. 130, Linköping University Electronic Press (2016)
12. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)

13. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
14. Kane, T.J., Staiger, D.O.: Gathering feedback for teaching: combining high-quality observations with student surveys and achievement gains. research paper. met project. Bill & Melinda Gates Foundation (2012)
15. Karita, S., et al.: A comparative study on transformer vs RNN in speech applications, pp. 449–456 (2019). https://doi.org/10.1109/ASRU46091.2019.9003750
16. Li, H., et al.: Multimodal learning for classroom activity detection. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 9234–9238. IEEE (2020)
17. Liu, Z., et al.: Dolphin: a spoken language proficiency assessment system for elementary education. In: Proceedings of The Web Conference 2020. p. 2641–2647. ACM (2020)
18. Lockheed, M.E.: School and classroom effects on student learning gain: the case of Thailand. World Bank (1987)
19. Murtagh, F.: Multilayer perceptrons for classification and regression. Neurocomputing **2**(5–6), 183–197 (1991)
20. Nystrand, M.: CLASS: A Windows Laptop Computer System for the In-Class Analysis of Classroom Discourse. https://dept.english.wisc.edu/nystrand/class.html. Accessed 10 Oct 2019
21. Nystrand, M.: Research on the role of classroom discourse as it affects reading comprehension. Res. Teach. English **40**, 392–412 (2006)
22. Owens, M.T., et al.: Classroom sound can be used to classify teaching practices in college science courses. Proc. Nat. Acad. Sci. **114**(12), 3085–3090 (2017)
23. Ramırez, J., Segura, J.C., Benıtez, C., De La Torre, A., Rubio, A.: Efficient voice activity detection algorithms using long-term speech information. Speech Commun. **42**(3–4), 271–287 (2004)
24. Scheeler, M.C., Ruhl, K.L., McAfee, J.K.: Providing performance feedback to teachers: a review. Teacher Educ. Special Educ. **27**(4), 396–407 (2004)
25. Sohn, J., Kim, N.S., Sung, W.: A statistical model-based voice activity detection. IEEE Signal Process. Lett. **6**(1), 1–3 (1999)
26. Tanner, K.D.: Structure matters: twenty-one teaching strategies to promote student engagement and cultivate classroom equity. CBE'Life Sci. Educ. **12**(3), 322–331 (2013)
27. Tanyer, S.G., Ozer, H.: Voice activity detection in nonstationary noise. IEEE Trans. Speech Audio Process. **8**(4), 478–482 (2000)
28. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
29. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. arXiv preprint arXiv:1710.10903 (2017)
30. Wang, Z., Pan, X., Miller, K.F., Cortina, K.S.: Automatic classification of activities in classroom discourse. Comput. Educ. **78**, 115–123 (2014)