DEEP LEARNING FOR CLASSROOM ACTIVITY DETECTION FROM AUDIO

Robin Cosbey[1] *Allison Wusterbarth*[*3] *Brian Hutchinson*[1,2]

[1]Computer Science Department, Western Washington University, Bellingham, WA, USA
[2]Computing and Analytics Division, Pacific Northwest National Laboratory, Seattle, WA, USA
[3]Conversica, Bellingham, WA, USA

ABSTRACT

Increasingly, post-secondary instructors are incorporating innovative teaching practices into their classrooms to improve student learning outcomes. In order to assess the effect of these techniques, it is helpful to quantify the types of activity being conducted in the classroom. Unfortunately, self-reporting is unreliable and manual annotation is tedious and scales poorly. We introduce a set of deep learning classifiers for automatic activity annotation, evaluating them on a collection of classroom recordings, with frames labeled as "single-voice" (primarily lecture), "multi-voice" (primarily group discussion), "no-voice" (primarily silent work) or "other." We find that our best approach obtains a 7.1% frame error rate (92.7% weighted F-measure) on held out class sessions from previously seen instructors (a 7 hr test set), and 10.1% error (89.1% weighted F-measure) on previously unseen instructors (a separate 18 hr test set). These represent 32% and 45% relative reductions in error over the existing state-of-the-art for this task. We also show that our estimates of how much classroom time spent per task are better correlated with actual time spent than existing systems.

Index Terms— classroom, activity detection, deep neural network, recurrent neural network, education

1. INTRODUCTION

Having a strong science, technology, engineering and math (STEM) workforce is critical for meeting the needs of modern society. Currently, college-level STEM education is primarily lecture-based. Many studies have shown that effectiveness of instruction can be improved by incorporating student-centered active learning strategies into the classroom [1, 2]. Further, innovative student-centered teaching practices have been shown to improve student retention rates, particularly for under-represented students [3]; both are essential to produce a vibrant STEM workforce. Examples of student-centered activities include *think-pair-share*, in which students reflect on a question, discuss it in small groups, and share out to the class, and *polling*, in which students vote on a prompt via polling device (e.g. clicker), often followed by a discussion of the results. Automatically and accurately quantifying classroom activity offers many potential benefits. For example, it makes it easier for individual instructors to track their adoption of these techniques so that they can identify where to further integrate them. Used at scale, it would allow STEM education researchers to track trends and measure the impact of student-centered techniques on student learning outcomes, enabling them to offer better recommendations for improving STEM education. Unfortunately, until recently, large scale quantification of student-centered technique use has been impractical.

The Decibel Analysis for Research in Teaching (DART) tool[1] was introduced by Owens et al. [4] to enable large scale, automatic annotation of classroom activity from classroom recordings (e.g. as collected by a handheld audio recorder). In contrast to more complicated schemes designed for manual annotation (e.g. [5, 6, 7, 8]), they use a streamlined scheme:

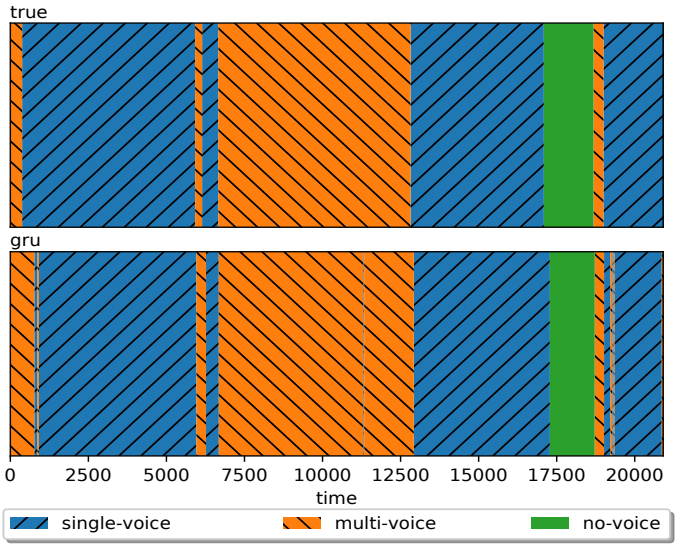

Fig. 1. Illustration of activity in a sample class session (top is ground truth, bottom is GRU prediction). The x-axis denotes time within the class. To reduce clutter, all detections less than 5s long (a handful of detections) were filtered out.

[*]Work conducted while a student at Western Washington University.

[1]DART is publicly available as a web service at https://dart.sfsu.edu.

each frame is labeled as "single-voice," (e.g. lecture, student question or student answer), "multi-voice" (e.g. group discussion), "no-voice" (e.g. silent reflection or writing) and "other." An example annotation trace for a classroom session is shown in Fig. 1, with the top row being ground truth annotation and the bottom row a prediction from an approach introduced in this paper. Owens et al. note that various instructional techniques form distinctive patterns; for example, think-pair-share generally follows a "no-voice, multi-voice, single voice" pattern. The DART system is a decision tree, where the features are statistics of the energy over a local window of the audio signal. To train their system, Owens et al. collected 1,720 hours of classroom audio from 67 distinct courses, and hand-annotated roughly 85 hours at a 0.5s granularity, achieving $\sim$90% frame level accuracy.

While the DART system has very good accuracy overall, there is room for improvement, particularly in the confusability between the dominant class ("single voice") and the less common classes, which tend to have relatively low recall. This paper contributes three new systems for automatic classroom activity annotation from audio using deep and recurrent neural network architectures and a thorough experimental evaluation of each system's generalization to new classes and instructors. Using DART data, we report substantial improvements in performance over DART, logistic regression and majority class baselines with minimal increase in computational cost, enabling higher quality, scalable and automatic annotation of classroom activity.

Researchers have explored several related tasks involving the analysis of classroom audio. Wang et al. [9] used the LENA system [10] in K-12 classrooms to automatically detect characteristics of classroom speech such as speaker age (adult or child), distance from microphone (near or far), presence of overlapping speech, noise or broadcast speech. They trained a random forest to map from these features to the classroom activities "teacher lecturing," "whole class discussion" and "group work" at a 30 second temporal granularity. In contrast, our system does not require the proprietary LENA system, learning instead to predict classroom activity from acoustic features directly. Donnelly et al. [11] automatically detect a set of classroom activities using instructor-worn wireless microphone audio in a middle school environment, with a Naïve Bayes classifier on assorted prosodic, NLP and acoustic features. Our approach instead requires virtually no feature engineering, opting to learn discriminative features directly via expressive deep learning models. We report substantially higher F-measure values, admittedly on a different dataset with different classroom activity labels. Blanchard et al. [12] use machine learning models to detect teacher questions in a middle school classroom environment, finding a Naïve Bayes model to work best for their task. Finally, classroom activity detection has also been explored from a video analysis perspective [13], enabling the system to pick up on non-verbal cues.

2. METHODS

Deep learning has been shown to produce impressive, often state-of-the-art, performance on a range of tasks. Here we introduce deep learning approaches to frame level classroom activity classification, along with a set of baseline methods. Each of our neural network methods was implemented in Tensorflow [14] and trained using stochastic gradient descent to minimize frame level cross-entropy loss.

2.1. Deep Learning Approaches

Deep neural networks (DNNs) model non-linear input-output relationships through a series of non-linear transformations, from input through a series of hidden representations to the output. For each frame, our DNN takes as input x either the feature vector for that frame, or a window of k frames (the concatenation of the feature vector for the frame with the $(k-1)/2$ previous and $(k-1)/2$ following frames). For our task, the output y is in $\mathbb{R}^4$ and uses softmax activation, giving posterior probabilities over the four activity labels.

Recurrent neural networks (RNNs), whose hidden representation at time t is not only a function of x_t but also $x_1, \ldots, x_{t-1}$, have been widely employed for sequence modeling tasks that arise in language processing. We first consider standard (Elman) RNNs [15], feeding at each timestep the feature vector a single frame, and predicting the activity label at that frame. Because RNNs are known to struggle with propagating information over long time spans [16], we also consider Gated Recurrent Unit (GRU) networks [17] which, like LSTM networks [16], are effective at storing and utilizing long-term history. No post-processing of the labels was done (e.g. to enforce smoothness in labels over time) for any of our models.

2.2. Baselines

We consider three baseline systems. First, we compare against DART, described in Section 1, representing the current state-of-the-art on this task. The DART predictions were provided by the authors of [4]. Additionally, to assess the effect of model depth, we also compare to a logistic regression classifier (LR) baseline, which has the exact same setup as the DNN, except the number of hidden layers is set to zero. Finally, given the imbalance of classes in the data, we report results for the majority class baseline, which predicts all frames as single-voice.

3. EXPERIMENTS

3.1. Experimental Setup

We evaluate the effectiveness of our models on an annotated classroom audio corpus collected by Owens et al. for [4]. The

corpus contains 85 hours of audio split between 54 class sessions collected from seven instructors. Manual ground truth annotation is provided every 0.5s using the four-way annotation scheme described in Section 1: single-voice, multi-voice, no-voice and other. We split the data into four sets to include train, development and two test sets. Train, development and test1 are an 80%-10%-10% split of the class sessions from the first five instructors. Test1 is designed to estimate generalization performance to new class sessions for already seen instructors. Test2 contains all of the class sessions for the remaining two instructors and is designed to test generalization to unseen instructors. While it would be interesting to explicitly measure generalization to unseen classrooms or students, classroom and student metadata is not available.

The DART corpus audio was collected using Sony ICD-PX333 handheld audio recorders placed at the front of the classroom and stored in a compressed (mp3) format. From each frame we extract 40 log mel-filterbank features plus energy using HTK [18]. We consider and report results on two frame sizes: 0.5s with a 0.25s offset, and 1s with a 0.5s offset. We also tried the traditional 25 ms frame with a 10 ms offset used for speech processing and found worse performance for this task (in some cases on par with the majority class baseline). To provide greater temporal context to the DNN and logistic regression models, we window the frames, with total window sizes up to 31. (DART uses a rolling 15s window.)

DNN, RNN and GRU hyperparameters, including learning rate and number of hidden units, were tuned on the development set. For the DNN we additionally tuned the number of layers; we did not stack the RNN or GRU. Trained models were evaluated on both test sets with frame error rate, recall, precision, F-measure per label and weighted F-measure ($\sum_{c=1}^{C} \alpha_c F_c$, where F_c is the F-measure for label c and $\alpha_c \in [0, 1]$ is the fraction of points with true label c).

3.2. Results and Analysis

First, in Table 1 we compare method and frame size, reporting frame error rate and weighted F-measure on both test sets. The best overall model is the GRU using 0.5s frames, and performance begins to stagnate at the longer frame size.

Table 2 shows the effect of window size on the LR and DNN models using windowed features. It shows a clear trend favoring larger window sizes. While we could likely see performance improvements with even larger window sizes, given the comparison to recurrent networks in Table 1, we find use of recurrent models to be the more promising approach. Contrasting the LR and DNN results in Table 2 also highlights the advantage of model depth, particularly when generalizing to previously unseen instructors as shown in test2.

Table 3 breaks down the performance of each method by precision (P), recall (R) and F-measure, for each of single-voice, multi-voice and no-voice using 1s frames with 0.5s offsets (other was very infrequent in the training data and not predicted by most models, so we omit it from this analysis). It shows that for for both test sets, deep learning approaches offer the best performance. It also suggests that single-voice appears to be the easiest category to distinguish, with average F-measures across methods on test1 of 0.954, followed by multi-voice (0.736) and no-voice (0.721). The trend persists on test2, where the same classes have average F-measures of 0.899, 0.538 and 0.422, respectively. The table also highlights how much more challenging the methods found detecting multi- and no-voice for unseen instructors.

While Tables 1, 2 and 3 analyze frame level performance, one may not always need accurate frame level predictions. For example, one plausible use case is to simply estimate the fraction of time spent on single-voice, multi-voice, no-voice and other activities. To assess performance for this use case, for each class session we totaled up the number of seconds

		test1		*test2*	
Frame Size	**Method**	**Err**	**F**	**Err**	**F**
	MC	0.200	—	0.222	—
	DART	0.104	0.883	0.184	0.773
0.5s/0.25s	**LR**	0.097	0.899	0.225	0.742
	DNN	0.077	0.919	0.155	0.836
	RNN	0.076	0.918	0.140	0.850
	GRU	***0.071***	***0.927***	***0.101***	***0.891***
1s/0.5s	**LR**	0.090	0.907	0.227	0.751
	DNN	**0.072**	**0.926**	0.177	0.821
	RNN	0.077	0.919	0.154	0.838
	GRU	0.083	0.914	**0.108**	**0.883**

Table 1. Experimental results (frame error rate and weighted F-measure) on the two test sets contrasting frame size and method: majority class (MC), DART, logistic regression (LR), DNN, RNN and GRU. The DNN and LR models used a window size of 31 frames. Best model of its kind is bolded; best overall also italicized.

		test1		*test2*	
Model	**W Sz**	**Err**	**F**	**Err**	**F**
LR	**1**	0.158	0.826	0.235	0.711
LR	**3**	0.131	0.720	**0.225**	0.728
LR	**11**	0.105	0.892	0.227	0.742
LR	**17**	0.095	0.901	**0.225**	0.745
LR	**31**	**0.090**	**0.907**	0.227	**0.751**
DNN	**1**	0.120	0.876	0.215	0.777
DNN	**3**	0.093	0.903	0.171	0.819
DNN	**11**	0.080	0.916	0.155	0.832
DNN	**17**	0.076	0.921	***0.142***	***0.846***
DNN	**31**	***0.072***	***0.926***	0.177	0.821

Table 2. Effect of window size (# frames) on logistic regression (LR) and DNN, fixing frame size to 1s with 0.5s offsets, measured with frame error rate and weighted F-measure. Best model of its kind is bolded; best overall also italicized.

test1	Single			Multi			No		
Method	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
DART	0.892	**0.995**	0.941	**0.921**	0.476	0.628	**0.954**	0.577	0.719
LR 31	0.935	0.968	0.951	0.764	0.691	0.726	0.887	0.644	**0.746**
DNN 31	**0.948**	0.980	**0.964**	0.852	0.730	**0.786**	0.801	**0.688**	0.740
RNN	0.939	0.983	0.960	0.830	0.723	0.773	0.890	0.573	0.697
GRU	0.944	0.968	0.956	0.780	**0.751**	0.765	0.842	0.601	0.702
test2	Single			Multi			No		
DART	0.818	**0.982**	0.892	0.773	0.250	0.378	**1.000**	0.132	0.233
LR 31	0.820	0.915	0.864	0.468	0.277	0.348	0.887	0.360	**0.512**
DNN 31	0.879	0.914	0.896	0.636	0.533	0.580	0.750	**0.365**	0.491
RNN	0.882	0.931	0.906	0.678	0.584	0.627	0.735	0.360	0.484
GRU	**0.908**	0.963	**0.935**	**0.818**	**0.703**	**0.756**	0.881	0.250	0.389

Table 3. An analysis of the types of errors made by each method, with per-activity precision (P), recall (R) and F-metrics reported on both test1 and test2. The frame size is fixed to 1s with 0.5s offsets.

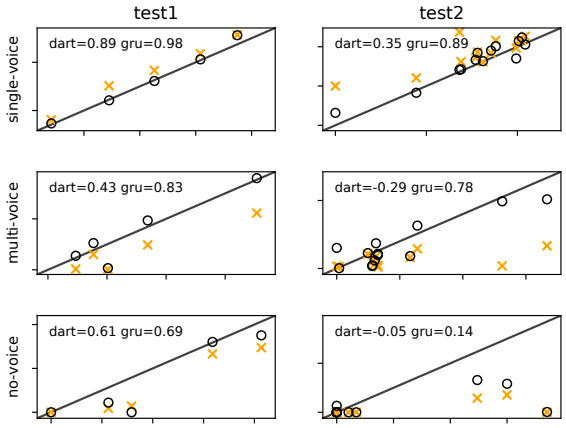

Fig. 2. Analysis of the correlation between the predicted amount of time spent on single-voice, multi-voice and no-voice for DART (orange x's) and GRU (black circles) relative to the ground truth. DART and GRU R^2 listed in each subfigure. For brevity, "other" activity plots not shown.

predicted for each activity type for DART and GRU and measured the correlation between those estimates and the actual time spent on each activity type. The results are shown in Fig. 2. Six scatter plots compare DART (orange Xs) and GRU (black circles) to ground truth for three activity types and both test sets. Coefficient of determination (R^2) values are listed for each case. Consistent with the findings in [4], the DART system errs on the side of over-predicting single-voice and under-predicting multi- and no-voice. This makes sense from the perspective of minimizing false detections of student-centered activities, which typically manifest as multi- and no-voice. Fig. 2 shows that the GRU can avoid excessive false detections, while providing estimates that correlate more strongly with the true proportion of time spent on activities, particularly for multi-voice.

4. CONCLUSIONS

This paper introduced a set of deep and recurrent neural network approaches for identifying college classroom activity from audio recordings. Evaluating on two test sets from the DART corpus, we show substantial improvements in frame error rate and F-measure over baseline systems. Notably, we observe 32% and 45% relative reductions in frame error rate relative to the current state-of-the-art for the task when generalizing to new class sessions from previously seen instructors and class sessions from previously unseen instructors, respectively. Additionally, we show high performance on estimating the total time spent per class session on single-voice, multi-voice and no-voice activity, including an average R^2 across the two test sets of above 0.9 for single-voice and above 0.8 for multi-voice. We find that a GRU-based system performs the best overall, although standard RNNs perform well, as do DNNs with windowed input.

One way this work could be extended would be to use more sophisticated deep learning models, such as bidirectional and stacked GRU or LSTM networks; we expect that this would further improve performance. Another direction would be to integrate the supervised activity classification models with unsupervised or semisupervised diarization or speaker change models, leveraging the unannotated data collected in [4] to more accurately segment the classroom audio.

Acknowledgements: The authors thank Kimberly Tanner, Melinda Owens and the DART pilot study instructors for sharing their data and DART predictions with us, and for their insights and feedback on this work. The authors also thank Nvidia for their donations of Titan X and Titan Xp GPUs used in this research.

5. REFERENCES

[1] President's Council of Advisors on Science and Technology, "Engage to excel: Producing one million additional college graduates with degrees in science, technology, engineering, and mathematics," 2012.

[2] S Freeman, SL Eddy, M McDonough, MK Smith, N Okoroafor, H Jordt, and MP Wenderoth, "Active learning increases student performance in science, engineering, and mathematics," *Proceedings of the National Academy of Sciences*, vol. 111, no. 23, pp. 8410–8415, 2014.

[3] MJ Graham, J Frederick, A Byars-Winston, A-B Hunter, and J Handelsman, "Increasing persistence of college students in STEM," *Science*, vol. 341, no. 6153, pp. 1455–1456, 2013.

[4] MT Owens, SB Seidel, M Wong, TE Bejines, S Lietz, JR Perez, S Sit, Z-S Subedar, GN Acker, SF Akana, B Balukjian, HP Benton, JR Blair, SM Boaz, KE Boyer, JB Bram, LW Burrus, DT Byrd, N Caporale, EJ Carpenter, Y-HM Chan, L Chen, A Chovnick, DS Chu, BK Clarkson, SE Cooper, C Creech, KD Crow, JR de la Torre, WF Denetclaw, KE Duncan, AS Edwards, KL Erickson, M Fuse, JJ Gorga, B Govindan, LJ Green, PZ Hankamp, HE Harris, Z-H He, S Ingalls, PD Ingmire, JR Jacobs, M Kamakea, RR Kimpo, JD Knight, SK Krause, LE Krueger, TL Light, L Lund, Leticia M M-M, BK McCarthy, LJ McPheron, VC Miller-Sims, CA Moffatt, PC Muick, PH Nagami, GL Nusse, KM Okimura, SG Pasion, R Patterson, PS Pennings, B Riggs, J Romeo, SW Roy, T Russo-Tait, LM Schultheis, L Sengupta, R Small, GS Spicer, JH Stillman, A Swei, JM Wade, SB Waters, SL Weinstein, JK Willsie, DW Wright, CD Harrison, LA Kelley, G Trujillo, CR Domingo, JN Schinske, and KD Tanner, "Classroom sound can be used to classify teaching practices in college science courses," *Proceedings of the National Academy of Sciences*, vol. 114, no. 12, pp. 3085–3090, 2017.

[5] D Sawada, MD Piburn, E Judson, J Turley, K Falconer, R Benford, and I Bloom, "Measuring reform practices in science and mathematics classrooms: The reformed teaching observation protocol," *School Science and Mathematics*, vol. 102, no. 6, pp. 245–253, 2002.

[6] MK Smith, FHM Jones, SL Gilbert, CE Wieman, and EL Dolan, "The classroom observation protocol for undergraduate stem (COPUS): A new instrument to characterize university stem classroom practices," *CBELife Sciences Education*, vol. 12, no. 4, pp. 618–627, 2013.

[7] JB Velasco, A Knedeisen, D Xue, TL Vickrey, M Abebe, and M Stains, "Characterizing instructional practices in the laboratory: The laboratory observation protocol for undergraduate STEM," *Journal of Chemical Education*, vol. 93, no. 7, pp. 1191–1203, 2016.

[8] SL Eddy, M Converse, MP Wenderoth, and J Schinske, "PORTAAL: A classroom observation tool assessing evidence-based teaching practices for active learning in large science, technology, engineering, and mathematics classes," *CBELife Sciences Education*, vol. 14, no. 2, pp. ar23, 2015, PMID: 26033871.

[9] Z Wang, X Pan, KF Miller, and KS Cortina, "Automatic classification of activities in classroom discourse," *Computers & Education*, vol. 78, pp. 115–123, 2014.

[10] M Ford, C Baer, D Xu, U Yapanel, and S Gray, "The LENA language environment analysis system.," Tech. Rep. LTR-03-02, LENA Foundation, 2008.

[11] PJ Donnelly, N Blanchard, B Samei, AM Olney, X Sun, B Ward, S Kelly, M Nystran, and SK D'Mello, "Automatic teacher modeling from live classroom audio," in *Proc. UMAP*, 2016, pp. 45–53.

[12] N Blanchard, PJ Donnelly, AM Olney, B Samei, B Ward, X Sun, S Kelly, Nystrand M, and SK DMello, "Identifying teacher questions using automatic speech recognition in live classrooms," in *Proc. SIGdial*, 2016, pp. 191–201.

[13] N Bosch, C Mills, JD Wammes, and D Smilek, "Quantifying classroom instructor dynamics with computer vision," in *Artificial Intelligence in Education*, Cham, 2018, pp. 30–42, Springer International Publishing.

[14] M Abadi, P Barham, J Chen, Z Chen, A Davis, J Dean, M Devin, S Ghemawat, G Irving, M Isard, M Kudlur, J Levenberg, R Monga, S Moore, DG Murray, B Steiner, P Tucker, V Vasudevan, P Warden, M Wicke, Y Yu, and X Zheng, "TensorFlow: A system for large-scale machine learning," in *Proc. OSDI*, 2016, pp. 265–283.

[15] J L Elman, "Finding structure in time," *COGNITIVE SCIENCE*, vol. 14, no. 2, pp. 179–211, 1990.

[16] S Hochreiter and J Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] K Cho, B van Merrienboer, Ç Gülçehre, F Bougares, H Schwenk, and Y Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. EMNLP*, 2014.

[18] SJ Young, "The HTK hidden markov model toolkit: Design and philosophy," Tech. Rep. TR.152, Cambridge University Engineering Department, 1993.