2024 IL-IN Section Conference
University of Illinois Chicago | April 13, 2024

Less is More: Efficiently Fine-tuning Whisper for Educational Content

Christian Classen, University of Illinois Urbana-Champaign

Christian Classen is a student at UIUC. His primary research interest is machine learning. He is interested in tackling problems in the intersection of AI and major social/ethical issues. By understanding the capabilities and flaws of current ML technology, we can ensure that emerging AI systems are used for the greater good while preventing harm.

Less is More: Efficiently Fine-tuning Whisper for Educational Content

Abstract

Recent advancements in deep learning have improved the performance of automatic speech recognition (ASR) technology. For educators, this has made the use of speech-to-text models for automated captioning an attractive option to improve the accessibility of instructional content. However, the accuracy of these models drops significantly when exposed to stuttering, accents, and technical language, limiting their effectiveness in realistic classroom or lecture hall scenarios.

In this work, we analyze how effectively Open AI's Whisper model can be adapted to a specific lecturer and acoustic environment by fine-tuning on limited subsets of manually-captioned lecture audio. Furthermore, we will discuss the implications of our results on the potential for further refinement of existing speech transcription models.

Introduction

The development of neural network technologies has led to their rapid adoption by both consumers and companies. While this has been most prominent for text-based language models such as OpenAI's ChatGPT, automatic speech recognition (ASR) systems have also seen dramatic increases in performance by replacing traditional architectures with deep end-to-end (E2E) models[1], [2]. These E2E models streamline the separate acoustic, lexicon, and language models of traditional ASR systems into a single fully-neural model, making them compact and easier to use [3], [4].

In the context of engineering education, existing work has created and evaluated sophisticated web platforms that use automatic speech recognition for undergraduate engineering classes. Examples include live captioning tool, ScribeAR [5], [6], and ClassTranscribe. The latter utilizes automatic speech to text (and crowd-sourced editing) to provide accurate captions, transcriptions ([7], [8]), and the ability to digital books in epub, pdf, and html format from video content [9]. Thus, improving the accuracy of speech to text of engineering content will have a direct and positive effect on the quality, accessibility, and inclusivity of engineering education.

Although state of the art E2E models can reach accuracy rates of 95% or higher on standard performance benchmarks [10]–[12], past research has demonstrated that the presence of rare words, non-native speakers, and disfluencies (e.g., stuttering) can cause transcription accuracy to drop [11], [13], [14]. This is particularly detrimental when trying to transcribe educational content, as it requires accurate recognition of technical language to be effective.

The most common way to improve ASR accuracy in such situations is to use domain or speaker adaptation techniques, namely fine-tuning a large pre-trained model with domain or speaker-specific data [11]. However, since speaker-specific training data generally requires audio to be manually transcribed, the practicality of obtaining a sufficiently large fine-tuning dataset is uncertain. Therefore, to further the understanding of the viability of E2E model adaptation for educators, we investigate the effects of fine-tuning a state of the art E2E model with a limited, speaker-specific dataset. Specifically, we examine the effectiveness of fine-tuning OpenAI's Whisper model [15] using manually-transcribed lecture audio on the scale of 5-10 hours.

Methods

Data collection and processing:

The audio and transcript data used in our first experiment was collected from lecture recordings of the CS 361 course at the University of Illinois at Urbana-Champaign (UIUC). Machine generated captions from ClassTranscribe were manually corrected to ensure accuracy. In total, approximately 5.5 hours of audio data was collected.

To standardize the text data, transcriptions were processed to remove symbols, punctuation, and capitalization. Audio data was segmented into 656 chunks, each roughly 30 seconds in length. These chunks of audio, along with their corresponding transcripts, were randomly assigned to training, validation, or test sets with probabilities of 70%, 15%, and 15% respectively. The resulting dataset consisted of 452 chunks for training, 97 for testing, and 94 for validation.

For the second experiment, we collected audio and transcripts from the University of Michigan's MICASE dataset [16]. In particular, we used approximately 4 hours of audio with significant amounts of non-native English speech from native speakers of Chinese, Korean, or Japanese. As in experiment 1, transcripts were processed to standardize text, with an additional step to remove text enclosed in brackets or parenthesis, as those represent simultaneous speech or descriptions of sounds. After segmenting the audio into 522 roughly 30 second chunks, we used the same random assignment method to get 338 chunks for training, 103 chunks for testing, and 81 chunks for validation.

Finetuning and hyperparameters

We used the 8-bit version of the ADAM optimizer to fine-tune Whisper's large model. Before training, 500 steps of warm-up were performed to determine an appropriate learning rate. In experiment 2, this was decreased to only 250 steps of warm-up. During fine-tuning, we used a batch size of 1 for training and a batch size of 16 for validation. Both prompts and timestamps were independently provided to the model for 50% of training batches. Gradient accumulation was also performed, updating the model's parameters every 64 steps, and gradient clipping was used with a max norm of 1.0.

Fine-tuning was performed for 1000 steps in experiment 1 and 500 steps in experiment 2. Every 250 steps, the current state of the model was saved, and its loss on the validation set was computed. Overall, using an NVIDIA A100 SXM4 80GB GPU, fine-tuning took 6 hours and 49 minutes for experiment 1 and 4 hours and 56 minutes for experiment 2.

Evaluation

To evaluate the performance of each version of the fine-tuned model on the testing set, we compare their unweighted and weighted word error rates (WER) to that of Whisper's large model. Weighted WER gives 0.5 weight to insertion and deletion errors while giving 1.0 weight to substitution errors. Before calculating WER, transcripts were processed using OpenAI's Whisper normalizer [15] to prevent penalization for superficial differences.

Results

Model	WER (Unweighted)	WER (Weighted)
Whisper Large Model	0.230	0.149
250 Step Model	0.187	0.146
500 Step Model	0.183	0.155
750 Step Model	0.371	0.341
1000 Step Model	0.188	0.178

Figure 1: Weighted and Unweighted WER Calculated for Experiment 1

Model	WER (Unweighted)	WER (Weighted)
Whisper Large Model	0.292	0.239
250 Step Model	0.309	0.262
500 Step Model	0.353	0.316

Figure 2: Weighted and Unweighted WER Calculated for Experiment 2

Experiment 1

The weighted and unweighted WER of each of our model checkpoints are shown in 1%. During fine-tuning, we reach the minimum weighted WER in the middle of the first epoch, and reach the minimum unweighted error rate around its end. Despite the WER decreasing relative to Whisper's large model, the validation error, as shown in 3, consistently increased as the number of fine-tuning steps increased.

Experiment 2

The WER in this experiment, as shown in 2, slightly increases in the first 250 steps of fine-tuning before significantly worsening over the following 250 steps. This is likely an issue with the MICASE dataset itself, which often alternates between several speakers and contains many instances of overlapping speech. Even though these transcriptions are easy to follow for a human, they contain non-standard tokens such as words cut off with a hyphen to represent stuttering, which may be difficult for ASR models to understand.

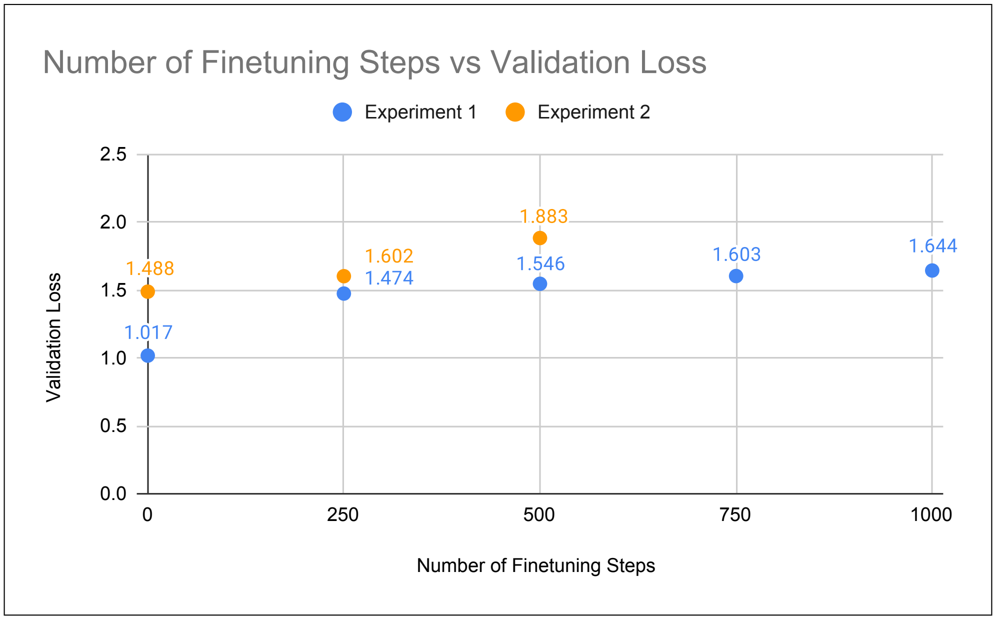

Figure 3: Validation Loss of the Model During Fine-tuning in Both Experiments

Findings

The results of these experiments demonstrate Whisper's ability to improve its performance on a specific speaker with a small amount of training data. While experiment 2 suggests that ASR models such as Whisper struggle to adapt to multi-speaker data, further investigation of this case is required.

We also see a pattern of increasing validation error during fine-tuning, regardless of whether the model's accuracy increased or decreased. This suggests that validation error may be too affected by minor discrepancies and stylistic differences to represent the effectiveness of training.

Limitations

Although our work attempts to accurately assess the effectiveness of fine-tuning Whisper efficiently on audio from a specific lecturer and acoustic environment, several factors limit the scope of our results.

Our experiments were performed with a limited set of data. The first experiment required several hours of educational content from the same speaker along with time-aligned transcriptions, so we were limited to using transcriptions that were manually corrected. Our . preventing us from performing our experiments with a large variety of different speakers and environments.

Additionally, some level of error in our transcripts was caused by a lack of a consistency when transcribing common disfluencies such as repeated phrases or blocking [17]. While we did attempt to mitigate this by manually correcting the transcripts and cleaning the data, it likely still negatively affected the reported performance of our fine-tuned model when evaluated using WER.

Conclusion

In this work, we evaluated the Whisper model's ability to improve its accuracy via limited fine-tuning on a speaker's voice. Although our results suggest that fine-tuning ASR models with a highly limited dataset, even for a single epoch, can significantly improve accuracy for speaker-specific transcription, it appears that more sophisticated methods of model adaptation may be necessary to reach human level performance.

Acknowledgments

This research was supported in part by the Illinois Computes project which is supported by the University of Illinois Urbana-Champaign and the University of Illinois System.

References

[1] G. Hinton, L. Deng, D. Yu, *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups", *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012. DOI: 10.1109/MSP.2012.2205597.

[2] A. Graves and N. Jaitly, “Towards end-to-end speech recognition with recurrent neural networks”, in *Proceedings of the 31st International Conference on Machine Learning*, E. P. Xing and T. Jebara, Eds., ser. Proceedings of Machine Learning Research, vol. 32, Bejing, China: PMLR, Jun. 2014, pp. 1764–1772.

[3] J. Li, *Recent advances in end-to-end automatic speech recognition*, 2022. arXiv: 2111.01690 [eess.AS].

[4] A. Hannun, C. Case, J. Casper, *et al.*, “Deep speech: Scaling up end-to-end speech recognition”, 2014. arXiv: 1412.5567 [cs.CL].

[5] L. Angrave, C. P. Lualdi, M. Jawad, and T. Javid, “Scribear: A new take on augmented-reality captioning for inclusive education access”, in *2021 Illinois-Indiana Regional Conference*, 2021.

[6] Y. Wang, C. P. Lualdi, L. Angrave, and G. N. Purushotam, “Using deep learning and augmented reality to improve accessibility: Inclusive conversations using diarization, captions, and visualization”, in *2023 ASEE Annual Conference & Exposition*, 2023.

[7] C. Mahipal, L. Angrave, Y. Xie, B. Chatterjee, H. Wang, and Z. Qian, ““What did I just miss?!” Presenting ClassTranscribe, an Automated Live-captioning and Text-searchable Lecture Video System, and Related Pedagogical Best Practices”, in *2019 ASEE Annual Conference & Exposition*, 2019.

[8] L. Angrave, K. Jensen, Z. Zhang, *et al.*, “Improving Student Accessibility, Equity, Course Performance, and Lab Skills: How Introduction of ClassTranscribe Is Changing Engineering Education at the University of Illinois”, *Grantee Submission*, 2020.

[9] H. Liu, L. Angrave, D. Dalpiaz, *et al.*, “A Digital Book Based Pedagogy to Improve Course Content Accessibility for Students with and without Disabilities in Engineering or other STEM courses (WIP)”, in *2022 ASEE Annual Conference & Exposition*, 2022.

[10] D. S. Park, W. Chan, Y. Zhang, *et al.*, “Specaugment: A simple data augmentation method for automatic speech recognition”, in *Interspeech 2019*, ISCA, Sep. 2019. DOI: 10.21437/interspeech.2019-2680. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2680.

[11] G. Synnaeve, Q. Xu, J. Kahn, *et al.*, *End-to-end asr: From supervised to semi-supervised learning with modern architectures*, 2020. arXiv: 1911.08460 [cs.CL].

[12] K. Kim, F. Wu, Y. Peng, *et al.*, *E-branchformer: Branchformer with enhanced merging for speech recognition*, 2022. arXiv: 2210.00077 [eess.AS].

[13] L. Clark, B. R. Cowan, A. Roper, S. Lindsay, and O. Sheers, “Speech diversity and speech interfaces: Considering an inclusive future through stammering”, in *Proceedings of the 2nd Conference on Conversational User Interfaces*, ser. CUI ’20, Bilbao, Spain: Association for Computing Machinery, 2020, ISBN: 9781450375443. DOI: 10.1145/3405755.3406139. [Online]. Available: https://doi.org/10.1145/3405755.3406139.

[14] A. DiChristofano, H. Shuster, S. Chandra, and N. Patwari, *Global performance disparities between english-language accents in automatic speech recognition*, 2023. arXiv: 2208.01157 [cs.CL].

[15] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, *Robust speech recognition via large-scale weak supervision*, 2022. arXiv: 2212.04356 [eess.AS].

[16] R. C. Simpson, J. O. S. L. Briggs, and J. M. Swales, *The michigan corpus of academic spoken english*, https://quod.lib.umich.edu/m/micase/, The Regents of the University of Michigan, Ann Arbor, MI, 2002.

[17] J. Prasse and G. Kikano, "Stuttering: An overview", *American family physician*, vol. 77, pp. 1271–6, Jun. 2008.