Robust Automatic Speech Recognition and Transcription with Wav2Vec 2.0 and Whisper

Authors: Sivan Ding, Alexandria Guo, Anh-Vu Nguyen, Antonin Vidon, Julia Wang, and Maxwell Zhou
Industry Mentors: Bhushan Jagyasi, Priyanka Pandey, and Surajit Sen

Background and Motivation

Whisper and Wav2Vec2.0 are large open source-based ASR (Automatic Speech Recognition) models developed respectively by OpenAI and Facebook, allowing users to transcribe audio to text. The architectures of both models are summarized in Figure 1.

Figure 1. Wav2Vec 2.0 pre-training (left) and Whisper (right) diagrams

The goal of our capstone project is to evaluate and compare the performance of each model across various complex scenarios, in order to test their robustness and determine the strengths and weaknesses of each model.

Preliminary Results

We examined the ASR frameworks in transcription tasks under more challenging conditions, where we downsampled and added noise to Librispeech audio to compare the robustness of Wav2Vec2.0 and Whisper. The metric we use for evaluation is Word Error Rate (WER) %.

Figure 2. Robustness of Wav2Vec2.0 and Whisper to downsampling (left) and noise (right)

Confidence Level Results

Testing has shown that Whisper overall outperforms Wav2Vec2.0, especially in terms of noise and downsampling robustness. While Whisper gives a better WER over Wav2Vec2.0, fine-tuning has shown consistent improvement over the baseline models.

Table 1. Summary of Training Data used to train and finetune each model

Table: Language | Whisper Pre-trained (hrs) | Wav2Vec2.0-XLSR Pre-trained (hrs) | Fleurs-Train (hrs)
---|---|---|---
Chinese | 23446 | 90 | 10
Korean | 7993 | 61 | 8
Hebrew | 688 | 77 | 10
Telugu | 4 | 62 | 8

Figure 3. Correlation b/w confidence level and error

Figure 4. Sample predictions colored by confidence level of respective token: Wav2Vec2.0 + 4-grams vs. Whisper

Conclusion and Next Steps

Testing has shown that Whisper overall outperforms Wav2Vec2.0, especially in terms of noise and downsampling robustness. While Whisper gives a better WER over Wav2Vec2.0, fine-tuning has shown consistent improvement over the baseline models.

Acknowledgments

We thank Accenture, Bhushan Jagyasi, Surajit Sen, and Priyanka Pandey for advising our team on this project, and providing good insights and guidance on our project. We would also like to thank Meta, OpenAI, and HuggingFace for their ASR open source contributions.

References