A streamlined model for fine tuning automatic speech recognition for specific users on sparse data

CS221 Final Project

Shubh Khanna

Anushree Aggarwal

Abel John

Introduction:

In this paper, we introduce a streamlined model for fine tuning automatic speech recognition (on the Whisper Model) for specific users on sparse data and show that the WER significantly decreases. Leveraging large-scale annotated speech to improve the word error rate on automatic speech recognition (ASR) performance for specific speakers has been a longstanding research problem. ASR systems are typically trained on large amounts of annotated speech data, where each example in the training data has been transcribed and labeled by a human. However, annotated speech data is expensive and time-consuming to collect, so having access to a large amount of unannotated speech data can be very useful for training ASR systems. We propose a novel pre-training framework that aims to decrease the word error rate (WER) on the Automatic Speech Recognition Task by fine-tuning limited amounts of domain-specific (speaker-specific) data. Our pretraining framework can be particularly useful for ASR on specific speakers who have limited annotated data available.

Literature Review:

W2V-BERT: Combining Contrastive Learning and Masked Language Modeling for
Self-Supervised Speech Pre-Training.

This paper proposes a w2v-BERT model that explores masked language modeling (MLM) for self-supervised speech representation learning. This model directly optimizes a contrastive loss and a masked prediction loss simultaneously for end-to-end self-supervised speech representation learning. It works on the LibriSpeech task and improves the real-world recognition task over conformer-based wav2vec 2.0. First was the pertaining phase, where they used unlabeled speech data to train the contrastive module along with the quantizer. They use a feature encoder for extracting latent speech representations from raw acoustic inputs. They also use a contrastive module that utilizes a quantization mechanism to discretize the feature encoder output into a finite set of representative speech units. Finally, the masked prediction model directly takes in the context vector and extracts high-level contextualized speech representations. Second, they fine-tuned using the labeled data. They performed two tasks in this phase: LibriSpeech and voice search. They took the pre-trained checkpoints at 400k steps for w2v-BERT XL and w2v-BERT-XXL and fine-tuned them on the supervised data. In this phase, in addition to fine-tuning, they also employed practical techniques to further improve the models' performance on ASR, such as data augmentation and language model fusion for decoding. The w2v-BERT model underscores the inevitability of contrastive learning for enabling masked prediction. Their results further highlight that mask prediction is useful for alleviating the problem of "easy negative samples" in contrastive learning.

Robust Speech Recognition via Large-Scale Weak Supervision

The standard approach to speech recognition is using unsupervised learning. While this is an excellent approach to audio encoding, audio decoders suffer from overfitting on datasets and poor performance in translating speech-to-text. What has been shown is that supervised training of ASR systems yields much better results and more generalizable models. Notably, weakly supervised datasets "improve the robustness and generalization of models". However, there just isn't enough of this training data to compare these models to those that are built using unsupervised training. As a result, OpenAI's approach to building Whisper is to find different ways to dramatically increase the labeled data being fed to the model. The paper also discusses using weakly-supervised training data to be both multilingual and multitask. Whisper's data processing dataset was built with audio matched with transcripts from across the Internet. These transcripts are both machine-generated and human-generated, which has been shown to help with model generalization. For language classification, VoxLingua107 was used, also to ensure that the spoken language matched the language of the transcripts (if it didn't, that example was rejected). The model involves off-the-shelf architecture to avoid confounding conclusions and utilizes an encoder-decoder Transformer. When compared to professional transcribers operating on the Kincaid46 dataset, the "pure-human performance is only a fraction of a percentage point better than Whisper's."

Dataset and Baseline:

Whisper Model (before fine tuning), Wav2Vec 2.0, fine tuned Wav2Vec 2.0

We will be making our own custom data set which ensures that no audio set was used for training any of the aforementioned models. After that, we'll test our data set on the three models for error rate and finetune it further to decrease the Word Error Rate (WER)

Description of data you are using - the size of the dataset, distribution of classes, any preprocessing you needed to do

We are training the model on a dataset consisting of short form audio snippets that capture a variety of accents and verbiage. This dataset consists of 3 different classes, and each contains a distinct accent indicative of a geographical or cultural vernacular. Each class has 10 minutes of recorded data, subdivided into 20 clips containing 30 seconds of audio. Class A contains the audio of a speaker with a traditional American accent. Class B is the audio of a speaker with an Indian accent, and Class C is the audio of a speaker with a stereotypical Southern (Texas) accent. The text that is transcribed is poetry written by authors such as Robert Frost and Paul Graham.

The intention of using poetry to train the model is to make use of the wide variety of vocabulary presented. This not only improves the model's capability on test data, it also reduces the risk of overfitting should the language present in the training and test set be too similar.

Main approach:

Our research problem is based on the whisper model, which is a type of neural network architecture designed to process natural language inputs. It can operate on very low-power devices. This language model is based on a recurrent neural network architecture, which allows it to process input sequences and maintain context over time. This makes it well-suited for applications like language translation and speech recognition, where understanding the meaning of a sentence or phrase requires taking into account the words that came before it. Overall, the whisper model is a compact and efficient way of processing natural language inputs.

Input: Audio of speech. Output: Text transcript of the speech.

We first calculated the domain-specific error rate on the baseline generic model. Then we explored different finetuning techniques to maximize information use of sparse data which shows the maximum decrease in word error rate.

Evaluation Metric:

Word Error Rate on Whisper Model; before fine tuning, and after fine tuning as a function of amount of data.

Final Contrastive Loss: $L_c = L_w + \alpha \cdot L_d$

Where c_t is a context vector corresponding to a masked time step t. q_t is the true quantized vector from a set of K distractors. We denote the loss as L_w and further augment it with a codebook diversity loss L_d to encourage a uniform usage of codes. The α is 0.1.

Masked Prediction Loss: $L_p = \beta \cdot L_c + \gamma \cdot L_m$

The context vectors produced by the contrastive module are passed to the masked prediction module, which produces the final context vectors to be used to complete a masked prediction task. L_m denotes the cross-entropy loss for the masked prediction task. β and γ are 1.

In this paper, we quantify the success of our predictive algorithm based on word error rate, a standard metric that quantifies differences between sentences. Word error rate is defined as the sum of substitutions, deletions, and insertions divided by the total number of words in the reference. It is the measure of the accuracy of an automatic speech recognition (ASR) system. A lower word error rate indicates that the ASR system is performing better and is able to recognize speech more accurately. To calculate the word error rate, the number of words in the test set is first determined. The ASR system is then run on the test set and the number of incorrectly recognized words is counted. The word error rate is then calculated as the number of incorrectly recognized words divided by the total number of words in the test set, multiplied by 100. For example, suppose that an ASR system is tested on a test set containing 1000 words. If the system incorrectly recognizes 50 of those words, the word error rate would be 5% (50/1000 * 100). The word error rate is an important metric for evaluating the performance of ASR systems because it provides a direct and intuitive measure of

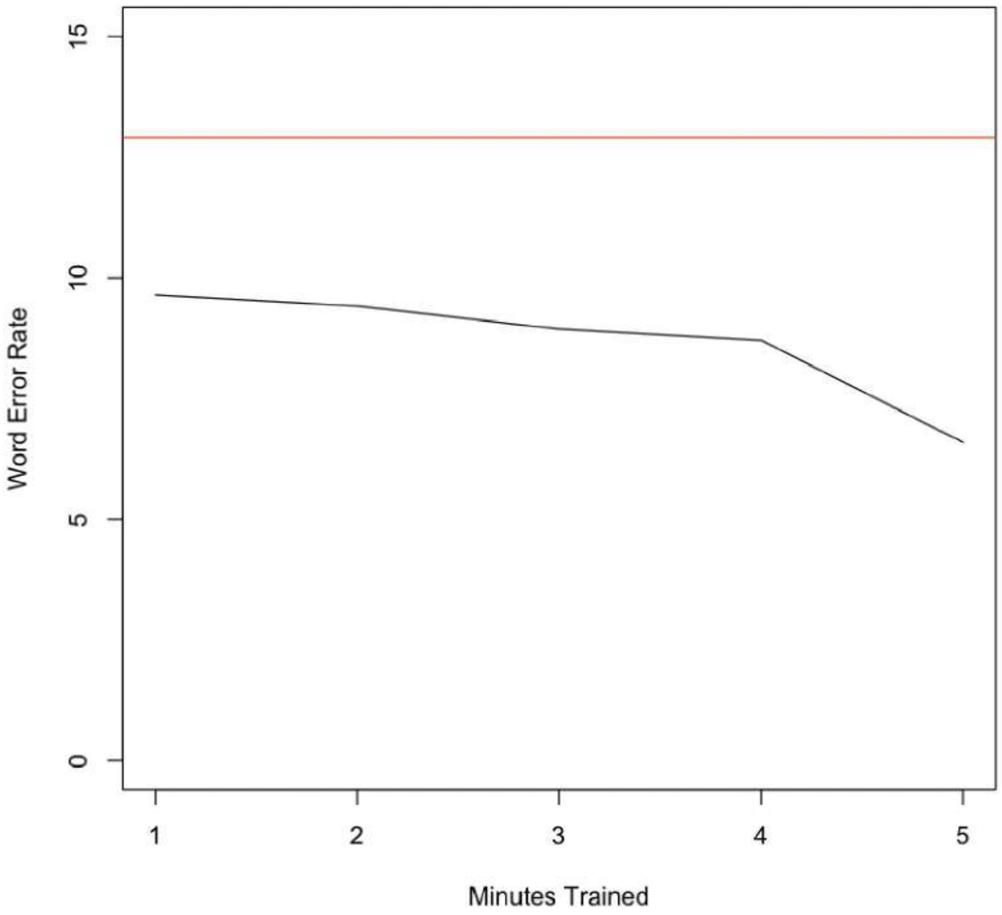
$$\mathbf{WER} = \frac{\mathbf{S + D + I}}{\mathbf{N}}$$

where...
S = number of substitutions
D = number of deletions
I = number of insertions
N = number of words in the reference

how well the system is able to recognize speech. By comparing the word error rates of different ASR systems, it is possible to determine which system is the most accurate and reliable.

Results & Analysis:

The following image depicts the average WER across all 3 classes:



With five minutes of training, we saw that the WER was: 6.588235

With four minutes of training, we saw that the WER was: 8.705882

With three minutes of training, we saw that the WER was: 8.941176

With two minutes of training, we saw that the WER was: 9.411765

With one minute of training, we saw that the WER was: 9.647059

We calculated the baseline word error rate of the sparse data to be 12.9.

Analysis of the results:

These results show us that with training on sparse datasets we can greatly improve the accuracy of the model relative to the word error rate with no training whatsoever. Considering Whisper's performance on the Kinkaid-46 dataset was only a fraction better than that of a human, we can generalize that not training a Whisper model is akin to a human transcriber unfamiliar with a particular accent being asked to transcribe the spoken language, and this shows that the Whisper model is capable of improving upon the baseline word-error rate with minimal training.

How these results will help us fine-tune the main approach:

With these results and the knowledge that WER is greatly improved with five minutes of training, it would make sense to continue to train on more training data that is of a similar length.

Error Analysis -

To understand the efficacy of our system, we compared with the standard baseline performance of the Whisper model with no additional personalized training snippets. Unsurprisingly, we were able to outperform this—with only a few minutes of specific training data, we were able to adjust our model to be more accurate in transcribing for an individual person. This is ultimately best, as the use case for our model is in person-specific speech transcription.

Future Work

As we move forward, our priority is to finetune the Whisper model and decrease the word error rate on our unique dataset.

One of the biggest challenges in the way is collecting large chunks of data. Our objective is to create a more streamlined process that can allow us to collect data at large scale and be able to more effectively train and personalize our algorithm. Constructing the dataset is a major challenge as we have to ensure that it hasn't been previously trained on the Whisper model, which has many hours of training data. Furthermore, we have to standardize data collection so that we can migrate towards having effective input sources for our model to be trained on individuals.

Moving ahead, for the future we hope to finetune the model to the point where we only need to label the first fifteen minutes of the data and significantly decrease the word error rate on specific speakers. Currently, annotated speech data is expensive and time-consuming to collect, so having access to a large amount of unannotated speech data can be very useful for training ASR systems. By leveraging large-scale annotated speech to improve the word error rate on automatic speech recognition (ASR), we would make the task inexpensive and less time-consuming with a lower word error rate.

References

Chung, Yu-An, et al. “W2V-Bert: Combining Contrastive Learning and Masked Language Modeling for Self-Supervised Speech Pre-Training.” *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, https://doi.org/10.1109/asru51503.2021.9688253.

Radford, Kim, et al. “Robust Speech Recognition via Large-Scale Weak Supervision.” *2021 OpenAI*, 2021, https://cdn.openai.com/papers/whisper.pdf.