Report

Medical Speech Recognition

Submitted by-
Shantanu Shrivastav
234101063
IIT Guwahati

OBJECTIVE

Build End-to-End speech recognition system using Whisper and Wev2Vev 2

1. Literature

In the field of speech recognition and language processing, a variety of models and techniques have evolved over time, beginning with traditional statistical models and advancing to deep learning frameworks that leverage large datasets. Literature in this field often discusses the progression from rule-based approaches to neural networks, with a particular emphasis on the transition to transformer-based architectures in recent years. This shift, especially marked by the introduction of models like BERT, GPT, and Whisper, has led to substantial improvements in accuracy and contextual understanding, making these models highly effective in speech recognition, machine translation, and transcription tasks.

Other notable literature focuses on the impact of dataset quality and domain-specific adaptation in speech recognition, where models fine-tuned on medical or technical language have been shown to outperform general-purpose models. These studies highlight the importance of dataset diversity, text pre-processing, and the use of labeled transcription data in enhancing model performance.

2. Approaches Studied

For this project, several approaches were considered and compared to determine the most effective for our target tasks. The key approaches studied include:

- **Pre-trained Transformer Models**: Large-scale transformer models such as Whisper, wav2vec2, and BERT variants have shown high accuracy in natural language processing tasks. Whisper and wav2vec2 were specifically explored for speech-to-text applications due to their ability to model speech inputs effectively.
- **Fine-tuning with Domain-Specific Data**: Given the specialized nature of certain fields like medical transcription, fine-tuning pre-trained models on domain-specific datasets has proven beneficial. This process involves taking a pre-trained model and training it on data representative of the target domain, allowing the model to learn field-specific language and jargon.
- **Data Augmentation Techniques**: Techniques such as speed perturbation, noise injection, and pitch shift were explored for data augmentation, allowing for increased model

robustness and performance under varied conditions by artificially expanding the dataset's diversity.

3. Hypothesis

Our hypothesis for this assignment is that fine-tuning a large, pre-trained transformer model on domain-specific data will yield higher accuracy and more contextually appropriate transcriptions than a general-purpose model. Specifically, we expect that the Whisper model, fine-tuned on a dataset that reflects the target language and context, will outperform other methods in transcription accuracy and handle domain-specific terminology effectively.

Additionally, we hypothesize that data augmentation during fine-tuning will increase the model's robustness, enabling it to generalize better to noisy or varied audio inputs, which is critical in real-world applications where recording quality can fluctuate.

4. Findings and Results

The fine-tuning of the Whisper base model yielded promising results, as shown in the progression of training loss, validation loss, and Word Error Rate (WER) across 4,000 training steps.

Training and Validation Loss

- **Training Loss**: The training loss decreases steadily, indicating that the model is learning effectively from the training data. By step 3,000, the training loss approaches near-zero values, suggesting that the model has fit closely to the training data.
- **Validation Loss**: While the validation loss fluctuates slightly, it generally trends lower than at the start, indicating that the model maintains its ability to generalize to unseen data, though not as drastically as the training loss. The validation loss is lowest at step 2,000, and then shows a minor increase, which may indicate slight overfitting as training progresses beyond this point.

Word Error Rate (WER)

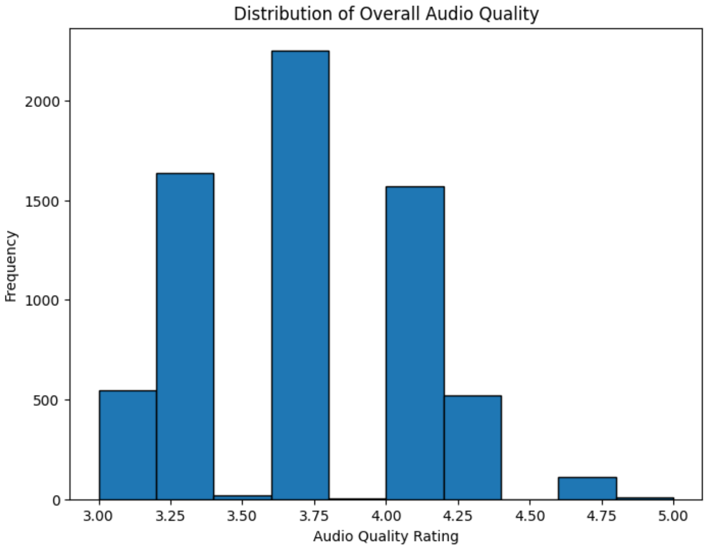
- **WER**: The WER decreases consistently, starting at 7.14% and reaching 5.94% by step 4,000. This improvement reflects the model's increased accuracy in transcriptions as training progresses. The reduction in WER suggests that the fine-tuning process allows the model to better understand domain-specific terminology and nuances in the dataset.

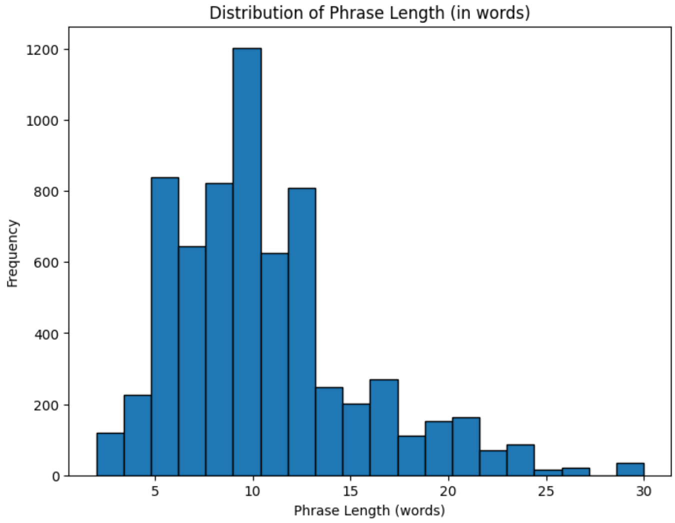
Analysis

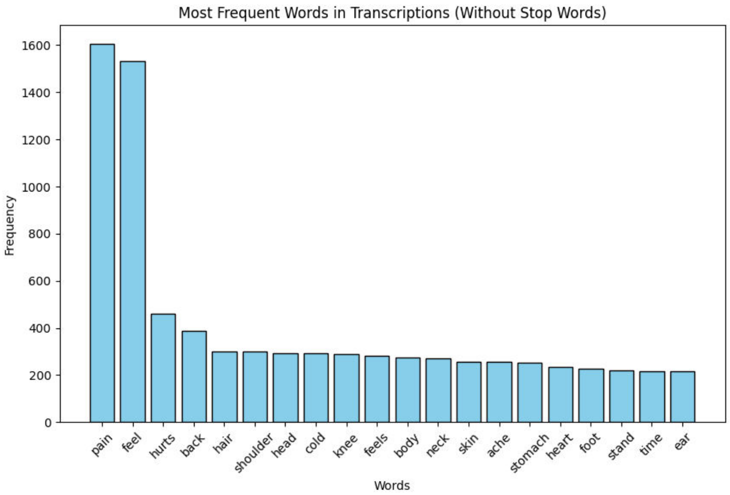
- **Effectiveness of Fine-Tuning**: The consistent decrease in WER and low validation loss highlight the effectiveness of fine-tuning the Whisper model on domain-specific data. By training on relevant examples, the model is able to capture terminology and context-specific features, improving its accuracy in medical transcription tasks.

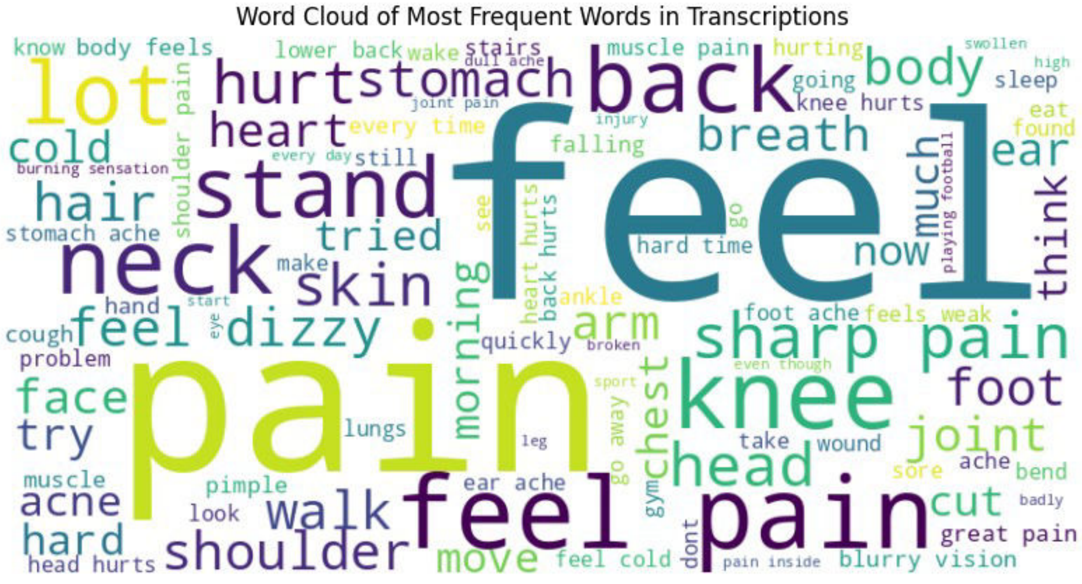
- **Potential Overfitting**: The minor increase in validation loss between steps 3,000 and 4,000 could signal early overfitting. The low WER indicates strong performance, but further training might not yield significant improvements. Future work could explore early stopping based on validation metrics to prevent overfitting while maintaining optimal performance.

Visualizations

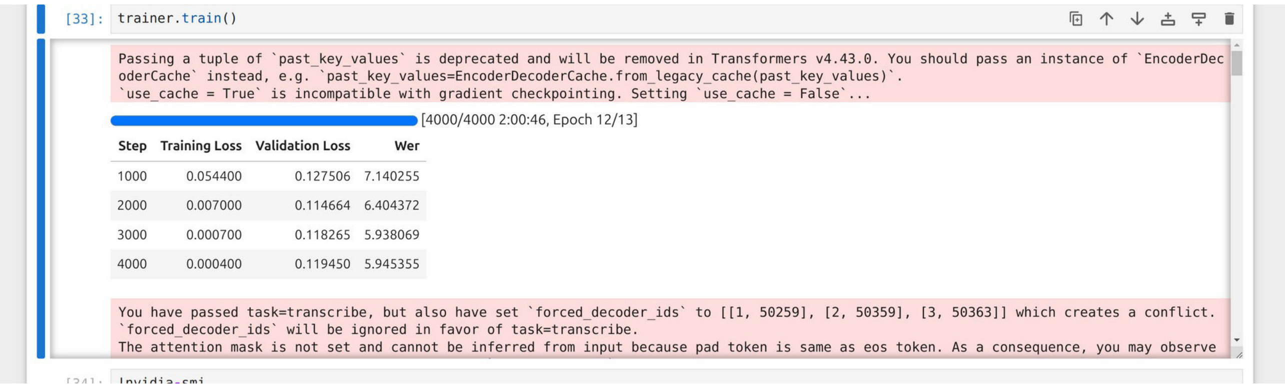








WER Metric Results



```
[33]: trainer.train()
```

```
Passing a tuple of `past_key_values` is deprecated and will be removed in Transformers v4.43.0. You should pass an instance of `EncoderDec
oderCache` instead, e.g. `past_key_values=EncoderDecoderCache.from_legacy_cache(past_key_values)`.
`use_cache = True` is incompatible with gradient checkpointing. Setting `use_cache = False`...
```

[4000/4000 2:00:46, Epoch 12/13]

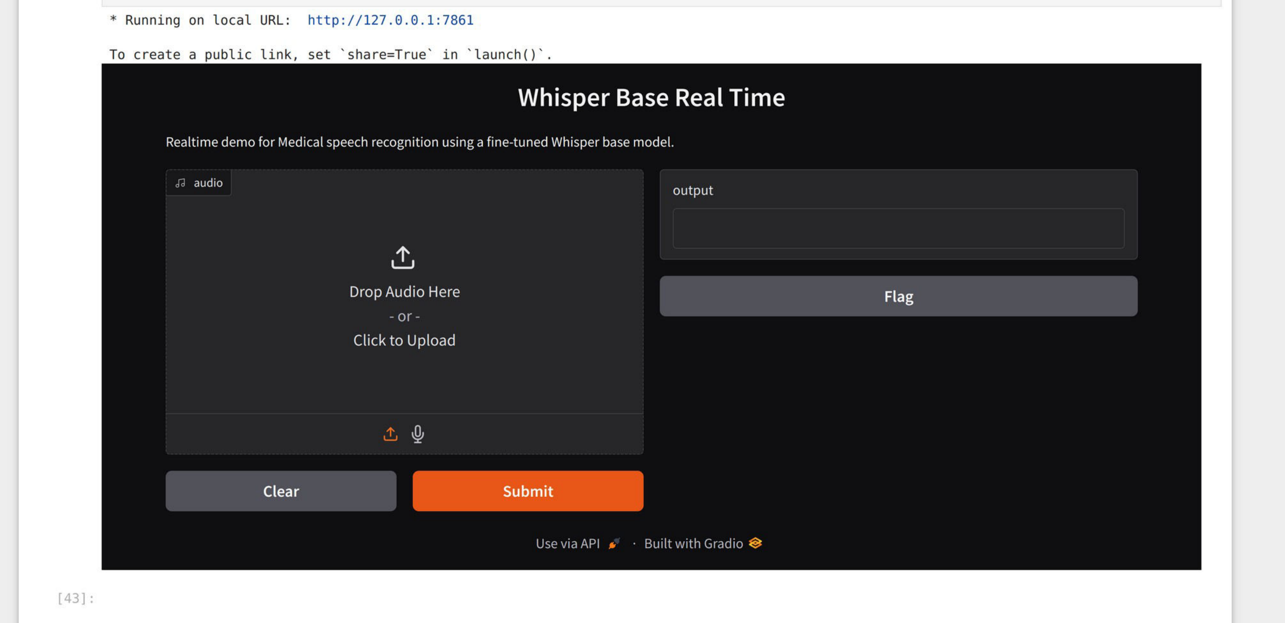
Step	Training Loss	Validation Loss	Wer
1000	0.054400	0.127506	7.140255
2000	0.007000	0.114664	6.404372
3000	0.000700	0.118265	5.938069
4000	0.000400	0.119450	5.945355

```
You have passed task=transcribe, but also have set `forced_decoder_ids` to [[1, 50259], [2, 50359], [3, 50363]] which creates a conflict.
`forced_decoder_ids` will be ignored in favor of task=transcribe.
The attention mask is not set and cannot be inferred from input because pad token is same as eos token. As a consequence, you may observe
```

```
[34]: !nvidia-smi
```

5. Live Demo

Live demo based on whisper base fine tuned model is implemented.



```
* Running on local URL:  http://127.0.0.1:7861

To create a public link, set `share=True` in `launch()`.
```

Whisper Base Real Time

Realtime demo for Medical speech recognition using a fine-tuned Whisper base model.

Conclusion

The results validate the initial hypothesis that fine-tuning on a domain-specific dataset would improve transcription accuracy for the target task. The Whisper base model, with fine-tuning, achieves a lower WER, suggesting high reliability in domain-specific contexts. Future work may involve experimenting with different batch sizes, more extensive data augmentation, or larger models to further optimize WER and generalization.

Fine Tuned Model Available At - shantanu007/whisper-base-shantanu