Training Report

Base vs finetuned wav2vec2 model

After finetuning the wav2vec2 model on the Common Voice train dataset (cv-valid-train), we evaluated the performance of the base and finetuned models on the Common Voice dev dataset (cv-valid-dev). The Word Error Rate (WER) was computed by comparing the generated transcriptions with the ground truth transcriptions.

The finetuned model has a WER of 0.065 on the cv-valid-dev dataset, as compared to 0.11 for the base model. This shows that the finetuned model is more accurate than the base model on the cv-valid-dev dataset.

The chart below shows the WER of the base and finetuned models on the `cv-valid-dev` dataset.

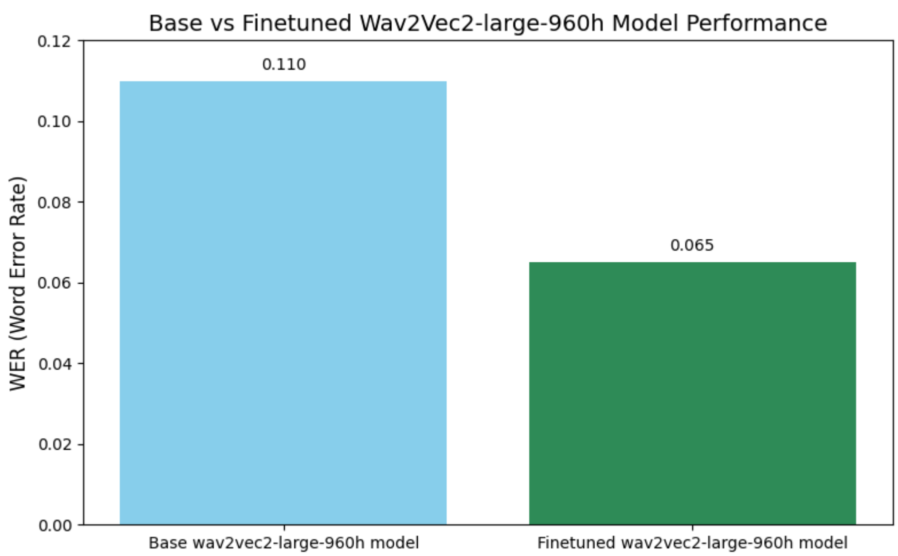

Figure 1: WER of base and finetuned models on cv-valid-dev

Error Analysis

In order to better understand the kind of errors made by the finetuned model, we use `jiwer` to list all the substitutions, insertions and deletions along with their frequencies. The error analysis is shown below.

Noisy Common Voice Dataset

My immediate observation is that the Common Voice dev dataset contains a lot of mislabelled data, and many of the errors are due to incorrect labels rather than model errors.

For example, we can see the following substitution errors:

Actual	Predicted	Frequency
dont	don't	11x
voicesand	and	9x
'everyone	everyone	3x
doesnt	doesn't	2x
'eat	eat	2x
'm	i'm	2x

These errors can be attributed to the misspelled labels. In these cases, the model predicted the correct word, but WER penalizes the model due to the misspelled labels.

Contractions and Expansions

Another source of model errors comes from incorrect contractions and expansions.

Actual	Predicted	Frequency
you're	are	4x
we	we're	4x
there's	is	4x
i've	have	3x

It appears that the dataset consists of a mixture of contractions and expansions, which can be a challenge for the model to know when to use which.

Phonetically Similar Errors

Another source of error comes from phonetically similar words.

Actual	Predicted	Frequency
men	man	3x
four	for	3x
horsemen	horseman	3x
saddened	sadden	2x

In these cases, the model is unable to disambiguate between phonetically similar words, leading to incorrect predictions.

Accent related errors

In some cases, the model struggles to recognize words spoken with different accents, which leads to some very odd errors.

Actual	Predicted	File ID
the lightbulbs need changing again	THE LIGHTBIRBS NEED CHANGING AGAIN	382
the model has effectively three fully connected layers	THE MARTEL HAS EFFECTEELY THREE FULLY CONNECTED LAIRS	218
you can't desert now	YOU CALV DESERT DO	124

Suggested Improvements

Training Dataset

Based on the error analysis, we can see that the finetuned model sometimes struggle with utterances with different accents. To improve the model's performance, we can consider finetuning the model further on a dataset with a diverse range of accents. We can use the `Voxpopuli` dataset, which is large scale dataset collected from European Parliament recordings. The recordings were made by speakers with different accents, which can help to improve the model's ability to generalize across different accents.

Other Experiments

Data Augmentation Strategies During training, we can also apply data augmentation to improve the model's performance. In particular, we can use additive noise methods, speed perturbations, volume perturbations or random cropping to increase the diversity of the training data. Additive noise methods add random background noises, speed perturbations changes the speed of the audio, volume perturbations changes the volume of the audio, and random cropping removes random segments of the audio. These data augmentation strategies can help the model to generalize better to different types of recordings.

Shallow Fusion with Language Model Although Wav2Vec2 predicts phonetic sequences, it does not model the probability of the resulting word sequences in natural language. As a result, it can output results that are linguistically invalid (e.g. "YOU CALV DESERT DO" or "LIGHTBIRBS"). Incorporating a

Language Model helps constrain decoding to more probable sequences of words based on prior linguistic knowledge, reducing errors and improving accuracy.

One common method for combining a Langauge Model with Wav2Vec2 is shallow fusion. In this method, Wav2Vec2 and a Language Model are trained independently and combined only during inference using a weighted sum. We can conduct further experiments with a lightweight Language Model such as KenLM to see if it can improve the model's performance.