

Training report

Note that I am using the base model rather than the large model due to the lack of compute resources on my local machine.

The base original model obtained a Word Error Rate (WER) of 0.14385, while the base fine-tuned version obtained a WER of 0.10821, which is considerably better.

We utilized most of the previous configurations: processor, tokenizer, and architecture. We simply fine-tuned the model on the dataset over 1 epoch to avoid overfitting.

To improve the model, we can fine-tune the model with more data; LibriSpeech, Common Voice (current), TED-LIUM, GlgaSpeech, Lirbi-Light, Multilingual LibriSpeech, Switchboard, TIMIT, LJ Speech, FLERUS, CHILDES, Buckeye Corpus.