Enhancing facebook:wav2vec2-large-960h model

The facebook:wav2vec2-large-960h model is finetuned and presented in the "cv-train-2a.ipynb" notebook. The finetuning approach taken is straightforward and consists simply of freezing all but the head and final 6 encoder layers of the network. Comparing the performance of the finetuned model and its pre-finetuned counterpart, we get the following Word Error Rate (WER) scores.

WER (%) for pre-finetuned and finetuned model

Pre-finetuned	Finetuned
10.8	7.6

As shown, the WER has fallen after finetuning, but only modestly. Building powerful deep learning models often require experimentation as the ideal training approach depends on the application, availability of useful datasets, required performance, etc. The following steps are proposed for this experimentation process for further model improvement.

1. **Model Architecture**
   The model architecture obviously plays an important role. The largest improvements can probably be attained through the following means.
   a. **Unfreezing more layers** – due to hardware limitations in this preliminary study, only six encoder layers have been unfrozen for finetuning. Assuming adequate CPU and GPU provisions are available, more (or all) layers can be unfrozen for further training, and the improvement in model performance tracked.
   b. **Use language model** – the current model does not include a language model, as it is focused on speech representation. A language model can be:
      i. added to the decoder to aid in transcription;
      ii. included as a separate head so that the model can be completely embedded in the model. This approach may require greater effort and should be re-visited if the required accuracy cannot be reached using the other enhancement approaches suggested here.
2. **Data**
   If the approaches above (structural changes to model) is inadequate or requires too much effort, the dataset can be enhanced before further training.
   a. **Labeled data acquisition** – If step 1 above does not produce appreciable improvement in model performance, the dataset at hand for training may be inadequate. The dataset can be grown by engaging participants to read prepared scripts. This step should be assessed to ascertain its feasibility, given the high costs involved.
   b. **Unlabeled data acquisition** – If gathering more labeled data (2a) is not viable, unlabeled data can be collected for Self-Supervised Learning (SSL) of the wav2vec2 encoder. Assuming the model is meant for application in a noisier environment (since common voice is used for finetuning), recordings of everyday speech (therefore noisy with ambient noise) can be taken from participants without labelling them. Powerful pre-processing techniques like Audio Activity Detection (AED) and Voice Activity Detection (VED) can be used to extract only relevant speech recordings.

c. **Data augmentation** – Distortions can be introduced to the Common Voice dataset to create more data. Ambient noise can be artificially introduced into clean, low-noise datasets like the LibriSpeech dataset.