11737 MNLP Assignment 2 Report

Members:
Greeshma Karanth (grk@andrew.cmu.edu),
Srikumar Subramanian (srikumas@andrew.cmu.edu)

Sub Task 1 - Individual ASR models

For this task, Fairseq was used to train the ASR model on Guarani because Fairseq provides an easy interface to integrate different models and datasets to work on everyday NLP tasks.

Dataset

The data used is sourced from Common Voice. The Guarani dataset contains ~1410 rows in the training split, 352 in the validation, and 811 in the test split.

Preprocessing

We preprocessed the dataset, by removing punctuations and normalizing the input text by removing content within parentheses, converting the text to lowercase, and then splitting it into words with spaces between them.
The sentences were tokenized as unigrams.

Without preprocessing the WER was ~3000 but dropped to ~100 with preprocessing thus demonstrating the impact of data cleaning.

Experimental setup

The feature extractor used was wav2vec and a transformer-based sequence-to-sequence model called s2t_transformer was used for generating transcripts. CTC loss was used to train the model.

Sub Task 2 - Multilingual Joint Training

To further improve performance on the task of Automatic Speech Recognition, we tuned hyperparameters and employed multilingual joint training by additionally training it with Portuguese data.

For this task, we moved away from using fairseq for experimentation to a setup that was easier to use, as mentioned in this blog https://huggingface.co/blog/fine-tune-xlsr-wav2vec2.

Dataset

We used Guarani and Portuguese from Common Voice. The objective was to choose a language that is similar to Guarani. While Italian and Spanish are also languages that are closely related to Guarani, we

chose Portuguese because the Spanish and Italian datasets were larger in size and would need more compute resources to use for experimentation.

Experimental Setup

For our experiments, Wav2vec was used as the feature extractor and XLS-R to generate transcripts. We performed the following experiments

1. Training on Guarani (30 epochs)
   - In this experiment, a speech recognition system was trained exclusively on data in the Guarani language for 30 epochs.
   - The Word Error Rate (WER), which measures the accuracy of the system, was approximately 0.52. A lower WER indicates higher accuracy in recognizing words in speech.

2. Training with Further Modified Hyperparameters (adding attention dropout 0.1, hidden dropout 0.4)
   - In the second experiment, the same Guarani language model was trained, but with a modification to the hyperparameters. Specifically, both attention dropout (0.1) and hidden dropout (0.4) were incorporated into the training process. This was done to enhance accuracy by encouraging the model to focus on more relevant information, reducing the risk of overfitting and improving generalization.
   - The resulting WER was around 0.56. This suggests that the introduction of dropout had a slight negative impact on the model's accuracy, which indicates that we might be excessively suppressing important information during training.

3. Training with Modified Hyperparameters (adding attention dropout 0.1)
   - In the third experiment, the Guarani language model's hyperparameters were further adjusted. Only the attention dropout rate of 0.1 was retained to see if reducing dropout helps improve performance.
   - Surprisingly, the WER remained approximately the same at 0.56. This indicates that while changes to the hyperparameters were made, they did not significantly affect the WER.

4. Multilingual Training on Portuguese + Guarani (30 epochs)
   - This experiment involved training a multilingual model that incorporated both Portuguese and Guarani data. This approach aimed to create a system capable of recognizing speech in both languages.
   - However, the WER for this multilingual model was notably higher at around 0.98. This suggests that training a single model on multiple languages might have posed challenges in maintaining high accuracy, possibly due to language diversity and the model's ability to differentiate between them.
- In summary, joint training with Portuguese made the model perform worse. Thus we made the decision to not experiment with other languages with much larger corpora such as Spanish or Italian, since this indicated that multilingual training was not super effective.

5. Training on Quechua and testing the Guarani model on Quechua (unsuccessful)
- The final experiment involved training the model on a Quechua dataset with an 80-20 train test split. This was unsuccessful however due to issues with compute resources. Reducing the batch size and changing gradient accumulation steps still did not help the data fit inside the GPU.
- Another experiment was testing the model trained on Guarani on the Quechua dataset but it faced the same issue. With further time and computational resources, we would have liked to explore this thread of effort further.

Issue description -

```
Unset
OutOfMemoryError: CUDA out of memory. Tried to allocate 1.67 GiB (GPU 0; 14.75 GiB total capacity; 10.96 GiB already allocated; 1.34 GiB free; 12.39 GiB reserved in total by PyTorch) If reserved memory is >> allocated memory try setting max_split_size_mb to avoid fragmentation. See documentation for Memory Management and PYTORCH_CUDA_ALLOC_CONF
```

Links and References

WandB: https://wandb.ai/idlis/mnlp/workspace?workspace=user-grkgrk
HuggingFace Blog: https://huggingface.co/blog/fine-tune-xlsr-wav2vec2
Github: https://github.com/GreeshmaKaranth/AutomaticSpeechRecognition_MNLP