Assignment 2: Multilingual Speech Recognition

Team:

1. Anusha Rao(anushasa@andrew.cmu.edu
2. Jason Rauchwerk (jrauchwe@andrew.cmu.edu)

Github: https://github.com/anushasa-rao/multilingual-speech-processing

Subtask 1:

Build an ASR model for the Guarani Dataset using Fairseq

1. Baseline Model changes to handle CommonVoice Gurani and Quechua:
    a. Preprocessing:
        i. The LIBRISPEECH dataset loader is specialized for Librispeech data, so the class was rewritten (with the same API) to select the correct fields and splits from the CommonVoice tsv files.
        ii. Additionally, the sample rate had to be adjusted for the mp3 files.
        iii. Finally, the train text file (which contains the sentences to be tokenized) had to be made into a permanent file (rather than tmp file) to work in Google Colab.
    b. Model Architecture: S2T Transformer (small) model provided by fairseq
    c. Training Parameters: 10 Epochs with 2E-3 LR
    d. Criterion: Label Smoothed Cross Entropy
    e. Results:
        i. WER - 143.12(Validation), 178.36 (Test)
        ii. Loss - 10.972
2. Custom Model Architecture:
   Starting with the S2T transformer (medium), tried out the S2T Conformer and LSTM architectures.
    a. S2T Conformer: Models that leverage Transformer and Convolutional layers for speech recognition
    b. LSTM: A gated recurrent neural network, capable of handling speech and text data streams
3. Augmenting Training Objectives
    a. Label Smoothing Cross Entropy with CTC: The connectionist temporal classification algorithm is an alignment-free scoring function that doesn't require an alignment between the input and output. CTC works by summing over the probability of all possible alignments.
    b. Label Smoothing Cross Entropy with Alignment: Computes the cross entropy loss based on an alignment between the sequence of token labels and the sequence of token predictions (https://arxiv.org/abs/2004.01655)

4. Advanced Decoding Techniques
    a. Beam Search: A search algorithm that considers a fixed number of probable candidates and selects the sequence with the highest joint probability

All the ablations details with **Results** can be found at:

- Documentation: Ablations
- WandB: https://wandb.ai/anushasa/mnlp-assignment-2

Subtask 2:

Finetuning the pre-trained model XLSR/Wave2Vec2 for ASR for Guarani

1. Baseline:
   XLSR, by Facebook AI, is a multilingual version of Wave2Vec2, a pre-trained model for Automatic Speech Recognition(ASR). The model represents cross-lingual speech representations and highlights the ability of the model to learn these representations that can be useful for multilingual training. While there exist various versions of this model, in our approach we focus on the smallest model, Wav2Vec2-XLS-R-300M
    a. Preprocessing: Basic text processing that involves the removal of special characters and converting text to lowercase. Meanwhile, for processing the audio, the clips are converted to a 1D array using huggingface dataset and sampled at the rate of 16KHz
    b. Tokenizer: The text is tokenized using Wav2Vec2CTCTokenizer
    c. Feature extraction: To sample and process the video we use Wav2Vec2FeatureExtractor
    d. Processor: The tokenizer and feature extraction are wrapped using Wav2Vec2Processor
    e. Training: The model is finetuned using the huggingface trainer. The initial layers of the model are frozen to ensure that the initial representations learned by the model that represent the independent, meaningful features of real-world data are preserved. The model was trained on the train dataset and validated on the validation data.
    f. Results:
        i. WER: 0.92
        ii. Loss: 0.612643
2. Experiments (Gurani)
    a. Increase attention_dropout to 0.05
        i. Results
            1. wer: 0.89220
            2. Loss: 1.88078
    b. Increase hidden_dropout to 0.05
        i. Results
            1. wer: 0.90767
            2. Loss: 2.08210

c. Increase attention_dropout to 0.05 and hidden_dropout to 0.05
    i. Memory error on Colab (no results)
d. Augment data with other language
    i. LangRank
        1. Used langrank [4] to determine best transfer language
        2. Hungarian was predicted to be the best
        3. 1. ted_hun : score=-0.01
            1. Transfer lang dataset size : score=0.45;
            2. Transfer over target size ratio : score=0.34;
            3. Transfer lang TTR : score=0.21
    ii. Results
        1. wer: 0.91571
        2. Loss: 1.16258
e. Increase attention_dropout to 0.05 and augment data with Hungarian
    i. Results
        1. Wer: 0.92483
        2. Loss: 1.10211

References:

[1] https://commonvoice.mozilla.org/en/datasets
[2] https://huggingface.co/docs/transformers/model_doc/wav2vec2
[3] https://huggingface.co/blog/fine-tune-xlsr-wav2vec2
[4] https://github.com/neulab/langrank