NLP Assignment 4: Research

Name: Alireza Dastmalchi Saei

Stu No.: 993613026

Question No. 1

Explain the architecture and pre-training process of wav2vec 2.0.

Paper 3 (wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations)

Architecture: Wav2vec 2.0 consists of a convolutional feature encoder followed by a transformer. The feature encoder processes raw audio into latent speech representations, while the transformer context network captures long-range dependencies. Pre-Training Process: The model is pre-trained using a contrastive task where it learns to distinguish true future audio samples from distractors. The training involves masking parts of the latent speech representations and predicting them based on the context provided by the unmasked parts.

Question No. 2

What is the difference between wav2vec 2.0 and wav2vec XLSR-53?

Paper 4 (Unsupervised Cross-lingual Representation Learning for Speech Recognition)

wav2vec 2.0: Focuses on learning representations from raw audio data in a self-supervised manner primarily for a single language. **wav2vec XLSR-53**: Extends wav2vec 2.0 by using multilingual data, enabling cross-lingual transfer. It is trained on speech from 53 languages, allowing it to generalize better across different languages and low-resource settings.

Question No. 3

How is decoding performed in the wav2vec 2.0 model? Explain the method used.

Paper 3

Decoding Method: Decoding in wav2vec 2.0 typically involves using a Connectionist Temporal Classification (CTC) decoder. During inference, the model outputs probability distributions over characters for each time step, and the CTC decoder converts these into the most likely sequence of characters, handling the alignment between input speech frames and output text.

Question No. 4

What method or technique is used to handle the alignment between input speech frames and output text in wav2vec 2.0?

Paper 3

Alignment Technique: The alignment between input speech frames and output text is handled by the CTC loss function. The CTC loss allows the model to learn the alignment implicitly by considering all possible alignments during training and summing their probabilities, enabling end-to-end training without the need for pre-segmented data.