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.

Architecture

0.0.1 Feature Encoder

The first stage of Wav2Vec 2.0 is the feature encoder, which transforms raw audio waveforms into latent speech representations. This is achieved using a stack of convolutional neural network (CNN) layers. These layers capture local dependencies within the input waveform and produce a sequence of feature vectors.

0.0.2 Quantization Module

The quantization module discretizes the continuous feature vectors into a finite set of discrete codebook entries. This process creates a set of discrete latent representations, often referred to as quantized representations.

During training, the model learns to predict these discrete latent representations. This step is for reducing the complexity of the data and providing a more manageable form for subsequent processing.

0.0.3 Context Network

This network consists of Transformer layers, which are well-suited for capturing long-range dependencies and contextual information across the sequence of quantized representations.

The context network refines the representations by considering the relationships and dependencies between different parts of the input sequence. This step is for enhancing the model's ability to understand and predict sequences of speech features.

0.0.4 Contrastive Task

The model learns to distinguish between the true quantized representation of a masked part of the input and a set of distractor representations.

By predicting the correct quantized representation despite the masking, the model effectively learns to encode useful information about the audio signal into its latent representations.

Pre-Training Process

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.