Processamento de Linguagem Natural com Transformers

Arquitetura Wav2vec

Wav2Vec é uma arquitetura de modelo de aprendizado de máquina desenvolvida pelo Facebook AI que é projetada para o reconhecimento de fala. O objetivo do Wav2Vec é converter diretamente as ondas sonoras em representações úteis para tarefas de Processamento de Linguagem Natural (PLN), como transcrição de fala em texto.

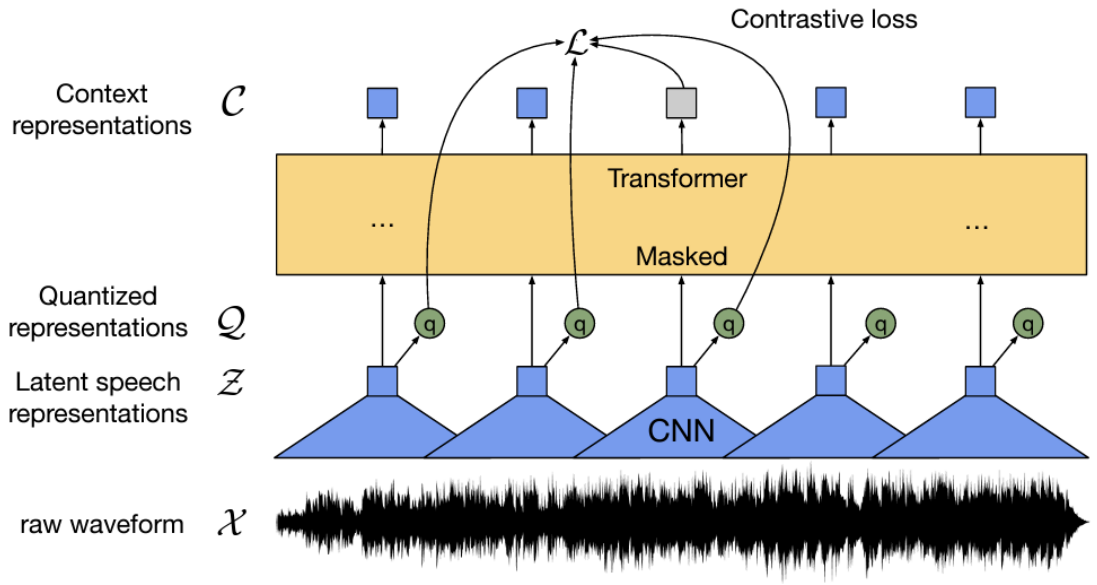
A arquitetura Wav2Vec é composta por duas partes principais:

Codificador de Áudio (Encoder): Esta parte do modelo recebe a entrada de áudio bruto e a transforma em uma sequência de vetores de recursos latentes. O codificador é uma rede convolucional que opera diretamente nas ondas sonoras.

Modelo de Contexto (Context Network): Esta parte do modelo pega a sequência de vetores de recursos latentes e aprende a representação contextualizada desses vetores. O modelo de contexto é uma rede Transformer, que usa mecanismos de atenção para capturar as dependências de longo alcance entre as partes da entrada.

O Wav2Vec é treinado em duas etapas. Primeiro, o modelo é pré-treinado em uma grande quantidade de dados de áudio não rotulados. Durante esta fase, o modelo aprende a representar o áudio de uma maneira útil para tarefas de PLN. Em seguida, o modelo é afinado (fine-tuned) em uma tarefa específica, como o reconhecimento de fala, usando uma quantidade menor de dados rotulados.

Uma das principais vantagens do Wav2Vec é que ele pode ser treinado com menos dados rotulados do que muitos outros modelos de reconhecimento de fala, graças à sua fase de pré-treinamento. Além disso, o Wav2Vec tem demonstrado um desempenho de ponta em várias tarefas de reconhecimento de fala. A figura abaixo ilustra essa arquitetura:



Referência:

wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations
https://arxiv.org/abs/2006.11477