Processamento de Linguagem Natural com Transformers

Arquitetura Speech2Text

Speech2Text é uma arquitetura de modelo de aprendizado de máquina que é projetada para converter fala em texto, também conhecido como reconhecimento automático de fala (ASR). A arquitetura Speech2Text é usada em uma variedade de aplicações, incluindo assistentes virtuais, transcrição de áudio e sistemas de ditado.

A arquitetura Speech2Text geralmente consiste em três componentes principais:

Codificador de Áudio (Encoder): Este componente recebe a entrada de áudio bruto e a transforma em uma sequência de vetores de recursos. O codificador pode ser uma rede neural convolucional (CNN), uma rede neural recorrente (RNN), ou uma combinação de ambas.

Modelo de Atenção (Attention Model): Este componente é responsável por determinar quais partes da entrada de áudio são relevantes para a previsão atual. O modelo de atenção pode ser implementado usando mecanismos de atenção, como o mecanismo de atenção usado na arquitetura Transformer.

Decodificador (Decoder): Este componente recebe a sequência de vetores de recursos e a sequência de atenção e gera a transcrição de texto. O decodificador é geralmente uma RNN.

A arquitetura Speech2Text é treinada usando pares de áudio e texto correspondente. Durante o treinamento, o modelo aprende a mapear a entrada de áudio para a transcrição de texto correspondente. Uma vez treinado, o modelo pode ser usado para transcrever áudio em texto.

Vale ressaltar que existem diversas variantes da arquitetura Speech2Text, cada uma com suas próprias nuances e melhorias. Alguns exemplos notáveis incluem modelos como Listen, Attend and Spell (LAS), DeepSpeech2 e Wav2Letter.

Referência:

fairseq S2T: Fast Speech-to-Text Modeling with fairseq
https://arxiv.org/abs/2010.05171