MENTION INTELLIGENCE ARTIFICELLE
Deep Learning Project

Speech recognition

Hervé Le Borgne

Kiyoshi Araki
Gabriel Souza
Lucas Tramonte

Décembre 2024

1 The problem addressed

Speech recognition is the process of converting spoken language into written text, aiming to transcribe words from audio recordings accurately. This task, whether in real-time or from pre-recorded audio, faces challenges such as variations in accents, speaking speeds, and background noise. Modern advancements leverage cutting-edge technologies like Large Language Models (LLMs) and deep learning frameworks to enhance transcription accuracy. Architectures such as Transformers, known for their attention mechanisms, have transformed the processing of sequential data like speech, enabling systems to capture long-range dependencies and produce more precise results.

Key techniques like word embeddings, such as Word2Vec (W2V), further improve semantic understanding by offering contextually accurate transcription. Pre-trained models on large datasets allow systems to manage linguistic diversity and ambiguities effectively. However, the complexity of human speech, influenced by accents, colloquial expressions, and environmental noise, necessitates robust neural architectures and extensive domain-specific datasets like LibriSpeech. These resources are critical for training and evaluating speech recognition systems in diverse real-world scenarios.

2 The available data

The LibriSpeech dataset , introduced by Vassil Panayotov et al. in Librispeech: An ASR corpus based on public domain audio books, is a foundational resource for Automatic Speech Recognition (ASR) research. It comprises approximately 1,000 hours of high-quality audiobook recordings sourced predominantly from the LibriVox project, which relies on public domain texts from Project Gutenberg. This diverse dataset captures a wide range of linguistic expressions, making it an invaluable benchmark for developing and testing ASR systems.

LibriSpeech is systematically divided into three training partitions of 100 hours, 360 hours, and 500 hours, designed to support experiments at different scales. The development (dev) and test sets are further classified into 'clean' and 'other' categories, representing varying levels of transcription difficulty based on the quality of the recordings and potential background noise. Each dev and test set includes approximately 5 hours of audio, facilitating controlled and comparative evaluation across systems.

3 Experiments under consideration

Our project will be based on two key papers: **Conformer: Convolution-augmented Transformer for Speech Recognition** and **Listen, Attend and Spell**. The goal is to understand these models, reproduce their methodologies, and compare the results to evaluate their respective performances in the field of Automatic Speech Recognition (ASR).

3.1 Conformer: Convolution-augmented Transformer for Speech Recognition

This paper introduces the **Conformer**, a novel architecture that combines Convolutional Neural Networks (CNNs) and Transformers to achieve state-of-the-art performance in end-to-end Automatic Speech Recognition (ASR). The Conformer aims to model both local and global dependencies of audio sequences in a parameter-efficient manner, leveraging the strengths of CNNs for local feature extraction and Transformers for capturing long-range global interactions [1].

Conformer Model Architecture

- **Conformer Block**: Combines:
  - Two **Macaron-style Feed-Forward Modules** sandwiching other components.
  - **Multi-Head Self-Attention (MHSA)** with relative positional embeddings for robust sequence modeling.
  - **Convolution Module** for efficient local feature extraction.
- The model processes input audio using a convolutional subsampling layer before feeding it into the Conformer blocks.

3.2 Listen, Attend and Spell (LAS)

This paper presents the **Listen, Attend and Spell (LAS)** model, a neural network architecture for end-to-end speech recognition that transcribes speech utterances into character sequences. Unlike traditional DNN-HMM systems, LAS jointly learns all components of the speech recognition pipeline without relying on phonemes or pronunciation dictionaries [2].

LAS Model Architecture

- Consists of two main components:
  - **Listener**: A pyramidal recurrent neural network (RNN) encoder that converts speech signals into high-level features.
  - **Speller**: An attention-based RNN decoder that generates character sequences from the features provided by the Listener.
- Uses content-based attention to align audio features with output characters dynamically.
- Overcomes limitations of Connectionist Temporal Classification (CTC) by modeling dependencies between characters.

4 Means of calculation

For our speech recognition experiments, we rely on the computational resources provided by the Data Centre d'Enseignement (DCE) . These resources are crucial for handling the intensive calculations required for training and testing deep learning models on large datasets, such as the LibriSpeech corpus.

References

[1] William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *ICASSP*, pages 4960–4964. IEEE, 2016.

[2] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech 2020*, pages 5036–5040, 2020.