Streamlining Speech-to-Text: A Novel Approach to Audio Transcription Using Deep Learning

Authors: Avadhoot Tavhare, Prerna Joshi, Sunita Barve, Diptee Chikmurge(Ghuse)

MIT Academy of Engineering
(An Autonomous Institute affiliated to Savitribai Phule Pune University)

2022-23 : Deep Learning

Abstract

Audio-to-text technology has experienced substantial growth, transforming spoken words into written text. Advancements in transcription services, such as Google Cloud Speech-to-Text and Microsoft Azure Speech, leverage automatic speech recognition systems. This paper presents an innovative approach to audio file transcription using deep learning models like Recurrent Neural Network (RNN) and Gated Recurrent Unit (GRU). The LJ Speech dataset is utilized, employing a sliding window mechanism for transcription. Spectrograms, offering a visual representation of frequency content over time, are used for audio analysis. The model integrates convolutional layers, bidirectional GRU layers, and a Connectionist Temporal Classification (CTC) loss for speech-to-text tasks. The research successfully predicts text outputs from audio inputs, facilitating transcription and signal analysis.

Objectives

- Create a novel method for audio transcription.
- Use the LJ Speech dataset for assessment and training.
- Use a sliding window mechanism to efficiently transcribe audio.
- Add bidirectional and convolutional GRU layers to the model's architecture.
- Apply CTC loss (Connectionist Temporal Classification) to optimize.

Dataset

TThe LJ Speech Dataset, a public domain speech dataset, is utilized in our research. It comprises 13,100 short audio clips of a single speaker reading passages from 7 non-fiction books, with each clip having a corresponding transcription. The clips, ranging from 1 to 10 seconds, sum up to approximately 24 hours of audio. The dataset provides metadata in a ‘transcripts.csv’ file, which includes the ID, transcription, and normalized transcription for each audio file.

Model Analysis

The proposed model leverages a sophisticated architecture designed for the accurate transcription of audio files, with a specific emphasis on utilizing Spectrograms as the primary feature representation. The architecture is tailored to address the challenges of capturing temporal dependencies inherent in audio data.

1. RNN: The core of the model incorporates a Bidirectional Recurrent Neural Network (RNN). This specific architecture is selected due to its proficiency in identifying sequential patterns from both forward and reverse directions.
2. GRU: Within the RNN architecture, a bidirectional Gated Recurrent Unit (GRU) layer is employed. The GRU layer is well-suited for sequence-to-sequence tasks, making it particularly relevant for transcribing audio data. Its gating mechanism allows the model to selectively retain and update information, facilitating the learning of long-range dependencies

Methodology

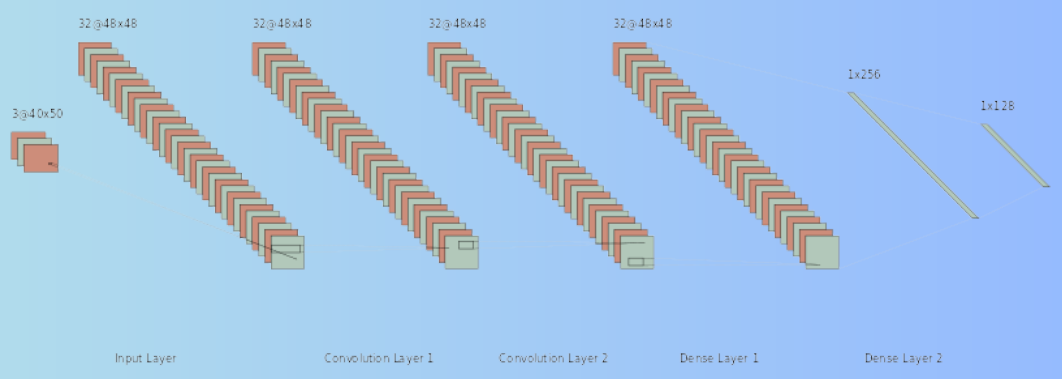

Figure 1.1 Methodlogy

Result and Analysis

TOur audio transcription system, utilizing deep learning algorithms like Bidirectional RNN, GRU, and CNN, is evaluated for performance using NLP techniques, as indicated by the outcomes in \cite{b13}. Sample results for each algorithm affirm the model's effectiveness in accurately transcribing diverse audio data, validating its practical utility.

Metric	Value
Word Error Rate (WER)	0.3673
Estimated Character Error Rate (CER)	0.15
Estimated Model Accuracy	0.85

Figure 1.2 Result Analysis

Future Scope

- Explore advanced deep learning structures for enhanced transcription accuracy.
- Investigate attention mechanisms to improve model focus on relevant audio features.
- Enhance reaction time through optimization techniques for real-time processing.
- Address domain-specific challenges to broaden the system's applicability.
- Investigate strategies to mitigate accuracy issues in specific conditions.

Conclusion

The paper explores Autonomous Speech Recognition (ASR) technology, focusing on audio-to-text transcription with RNNs, GRUs, and CNN. Using a dataset of 13,100 audio clips and spectrograms for model training, it achieves an 85% accuracy evaluated through CTC loss, WER, and CER. Future directions include advanced structures and addressing challenges like domain specificity for improved accessibility. Ongoing research aims to refine speech-to-text technology, enhancing user satisfaction across applications.