Speech Recognition System

Project Report
CSE445.6

Submission Date: 01/12/2024

Mehrab-E-Rashid Khan 1931071642
Nowshin Nawal Mim 1911054042

Introduction

Our project recognises the importance of speech recognition systems in the revolutionization of human-computer interaction. In this project we take an attempt at utilizing the latest available tools of AI models to improve on speech recognition systems. We can see the possibility of several implementations of this project in various different fields, such as: Voice assistant, Accessibility solution, unsupervised human interaction between client and hosts.

Objective

The primary objective of this project is to design and implement a speech recognition system that can efficiently process audio input and generate accurate text output. To achieve this, we will utilize the Keithio LJ Speech Dataset for training and fine-tuning a DeepSpeech2 model. The model will be trained on Google Colab, a cloud-based platform, taking advantage of its powerful GPU resources.

Methodology

The methodology involves the following steps:

1. **Data Preparation:** The Keithio LJ Speech Dataset will be pre-processed to extract audio samples and corresponding transcriptions. The dataset will be split into training, validation, and test sets in an 80:10:10 ratio. Data augmentation techniques will be applied to the training set to increase its diversity and improve model generalization.
2. **Model Selection and Training:** The DeepSpeech2 model, a state-of-the-art speech recognition model, will be chosen as the backbone of our system. The model will be trained using a combination of techniques, including Connectionist Temporal Classification (CTC) loss and beam search decoding. CTC loss will be used to track the model's progress during training.
3. **Model Evaluation:** The performance of the trained model will be evaluated on the held-out test set using standard metrics such as Word Error Rate (WER) and Character Error Rate (CER).
4. **Deployment:** The final model will be deployed to a suitable platform, enabling real-time speech recognition and transcription..

Impact of this Project

This project has the potential to significantly impact various fields. In the healthcare industry, it can aid in medical transcription and patient record documentation. In the education sector, it can facilitate language learning and accessibility for individuals with disabilities. Additionally, it can enhance voice-controlled devices and virtual

assistants, making them more intuitive and user-friendly. By contributing to the advancement of speech recognition technology, this project aims to improve human-computer interaction and overall user experience.

Results

Due to hardware limitations imposed by the free Google Colab tier (16GB RAM, 4-hour runtime), we were constrained to a batch size of 32 to prevent system crashes. The model was trained for 14 epochs, resulting in a steady decrease in both training and validation loss. While further training and optimization could potentially improve performance, the achieved results demonstrate the effectiveness of the DeepSpeech2 model in speech recognition tasks, even under resource constraints.

References

Model Building

- Amodei, D., Beyer, J., Gomero, A., & Ng, A. Y. (2016). Deep Speech 2: End-to-End Speech Recognition in English and Mandarin. arXiv preprint arXiv:1512.02595.
- Mozilla. (2020). DeepSpeech. [Software]. GitHub repository. https://github.com/mozilla/DeepSpeech
- TensorFlow. (n.d.). Audio classification and speech recognition: Using TensorFlow and Keras. TensorFlow.org. Retrieved from https://www.tensorflow.org/tutorials/audio

Dataset

- Yang, K., Snyder, S., Jernigan, J., Nam, H., Moniatte, D., & Strom, J. (2018, May 15). LJ Speech Dataset. Retrieved from https://keithito.com/LJ-Speech-Dataset/
- Keith Ito. (2017). LJ Speech Dataset. GitHub repository. Retrieved from https://github.com/keithito/tacotron

Conclusion

This project successfully implemented a speech recognition system using DeepSpeech2. Future work could focus on improving accuracy, reducing latency, and exploring multilingual capabilities. Additionally, integrating the model into real-world applications, such as voice assistants and transcription tools, would further enhance its impact.