Project Proposal

Satya Lakshmi Tejaswini Gunnapaneni 700754037

Title: Comparative Analysis of Deep Learning Architectures for Automatic Speech Recognition (ASR)

1. Introduction:

Automatic Speech Recognition (ASR) plays a pivotal role in facilitating seamless natural language interaction between humans and machines, powering a diverse range of applications from virtual assistants to dictation software and voice-controlled systems. Its significance lies in its capability to transcribe spoken language into text or commands, effectively bridging the communication gap between human users and computational systems.

This project aims to compare deep learning architectures such as recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformer-based models for ASR. The objective is to discern the respective strengths and weaknesses of these architectures across various ASR scenarios. Specifically, the models will be implemented and evaluated on a benchmark ASR dataset to compare accuracy and efficiency. Additionally, considerations will be made regarding the scalability and adaptability of each model to different data volumes and computational resources. The findings from this analysis will offer valuable insights into selecting the most suitable model for real-world ASR applications requiring high performance and scalability.

2. Related Works:

Previous research in Automatic Speech Recognition (ASR) has been marked by a surge in exploring diverse deep-learning architectures and rigorously evaluating their performance across benchmark datasets. Notably, studies have demonstrated the effectiveness of several prominent models, including Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and transformer-based architectures, in addressing the complexities of ASR tasks. Particularly, transformer-based models have garnered considerable attention due to their ability to capture long-range dependencies and achieve state-of-the-art results in recent years. Despite advancements, a comprehensive comparison of diverse deep learning architectures for ASR is lacking.

3. Execution Plan:

- Dataset Selection: Identify and procure suitable ASR datasets representing a range of speech characteristics.
- Model Implementation: Implement deep learning models, including RNNs, CNNs, and transformers, using TensorFlow/Pytorch/Keras, ensuring consistency in architecture.
- Training and Evaluation: Train the models on the selected datasets and evaluate their performance using standard ASR metrics such as word error rate (WER), accuracy, and computational efficiency.
- Comparison Analysis: Assess the efficacy of various models by comparing their accuracy, resilience to speed variations, and scalability across different datasets.

- Documentation: Record the experimental configuration, outcomes, and insights acquired during the project to ensure reproducibility and serve as a valuable resource for future reference.

4. Expected Learning/Contribution:

- Acquire an in-depth comprehension of diverse deep learning architectures commonly employed in ASR, including recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformer-based models.
- Cultivate hands-on expertise in implementing and training deep learning models using popular frameworks like TensorFlow/Pytorch, and Keras.
- Contribute to the existing body of knowledge in the field of ASR by conducting a systematic comparative analysis of different ASR models and providing valuable insights into the comparative performance of various ASR models and their suitability for practical applications.

5. Evaluation Plan:

This project will employ both quantitative and qualitative evaluation methods to assess the performance of each ASR model on the selected datasets. Quantitatively, standard ASR metrics such as accuracy, and computational efficiency will be used to measure the effectiveness of each model in transcribing speech accurately and efficiently. Additionally, qualitative evaluation will involve a detailed analysis of the strengths and weaknesses of each model based on experimental results and observations. This qualitative assessment will provide deeper insights into the underlying mechanisms and behaviors of the models, elucidating their performance nuances beyond numerical metrics. Finally, a comprehensive comparison analysis will be conducted to identify the most effective model(s) for different ASR scenarios. By synthesizing both quantitative and qualitative findings, this comparison will inform decision-making regarding the selection and deployment of ASR models in practical applications.

6. Conclusion:

This project endeavors to enhance our comprehension of deep learning models in the context of Automatic Speech Recognition (ASR) by undertaking a meticulous and systematic comparison of various architectural approaches. Through the evaluation of recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformer-based models across diverse datasets, the aim is to uncover nuanced insights into their respective advantages and limitations. By discerning the relative strengths and weaknesses of each architecture, this endeavor seeks to propel the current state-of-the-art in ASR technology forward. Ultimately, this comparative analysis will contribute to the refinement and optimization of ASR systems, paving the way for more accurate, efficient, and adaptable speech recognition solutions.

7. References:

1. Mallol-Ragolta, Adria, and Björn Schuller. "Coupling Sentiment and Arousal Analysis Towards an Affective Dialogue Manager." *IEEE Access* (2024).
2. Kabore, Moumini, et al. "Voice Interaction in Moore Language Study on Isolated Word Recognition in Audio Samples." (2024).
3. Feng, Siyuan, et al. "Towards inclusive automatic speech recognition." *Computer Speech & Language* 84 (2024): 101567.

4. Stan, Adriana, and Beáta Lőrincz. "Generating the Voice of the Interactive Virtual Assistant." *Virtual Assistant*. IntechOpen, 2021.