Lab Report

Advanced Recurrent Neural Network Architectures for Enhanced Phoneme Recognition On TIMIT Corpus

Rahul Gupta

San Diego State University

Abstract

Accurate phoneme recognition is an essential for speech processing tasks. This report presents a comparative analysis of advanced recurrent neural network architectures: Gated Recurrent Unit (GRU), Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM). By evaluating phoneme error rates and phoneme misclassifications patterns, the study aims to find the most optimal network parameters that produce the highest amount of recognition accuracy. The results offer some insights into the architectural features that contribute to the accuracy of such phoneme recognition systems.

Introduction

Automatic speech recognition (ASR) at it's core relies on phoneme recognition tasks on small chunks of sound inputs. The TIMIT dataset serves as an ideal benchmark for advanced RNN architectures due to its rich diversity of phonemes. This report details the initial experiment done with 3 distinct architectures: GRU, LSTM, CNN each known to handle sequential data in its own unique way. The goal was to understand the best architecture for the classification task.

Methodology

The TIMIT corpus was preprocessed and the features were extracted using MFCC to capture the phonetic characteristics. In this experiment, three distinct RNN architectures were created using similar hyper parameters. For training, we used cross validation strategy to prioritize robustness and kept track of phoneme error rates and validation accuracy metrics.

Experiment

Experiment One: GRU-LSTM—Based RNN

- This model consisted of GRU and a LSTM hybrid architecture with batch normalization.
- Dropout layers and L2 regularization was used to mitigate overfitting.

Experiment Two: CNN

- This model consisted of a convolutional neural network architecture with batch normalization.
- L2 regularization was used to mitigate overfitting.

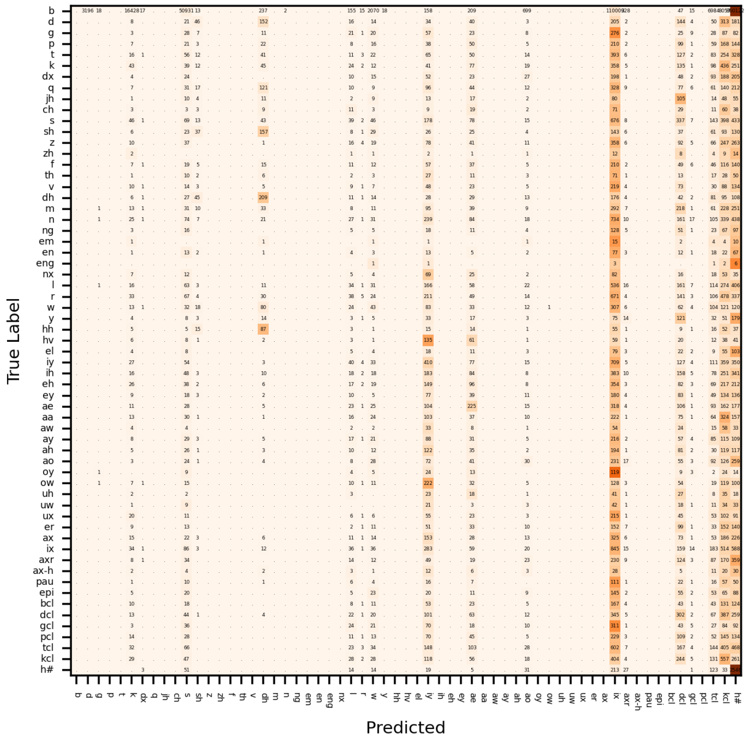
Experiment Three: LSTM-Based RNN

- This model consisted of a pure, multiple LSTM architecture with batch normalization.
- Dropout layers and L2 regularization was used to mitigate overfitting.

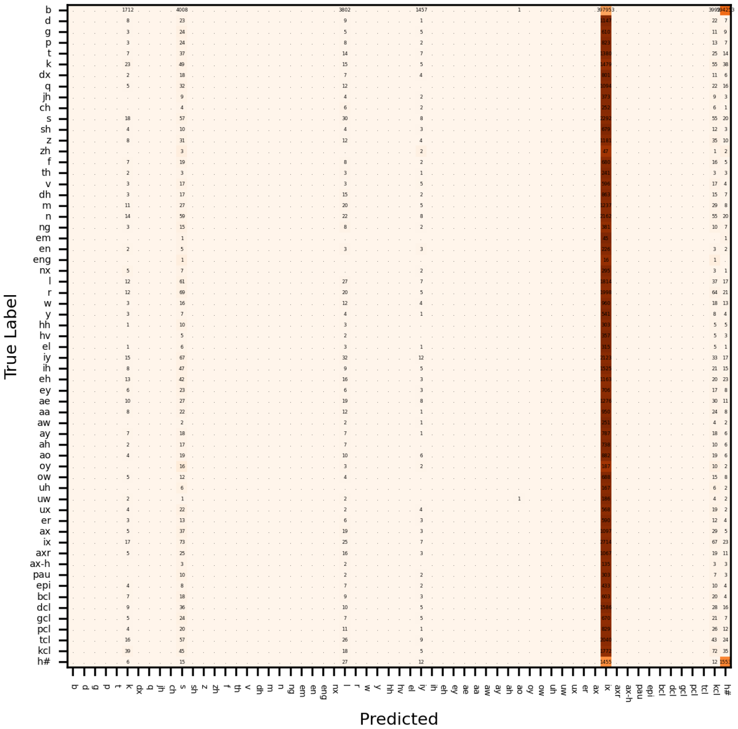
Results

The experiment produced less than satisfactory results. The GRU-LSTM architecture showed moderately successful classifications however had high loss values.CNN, which excels at feature extraction however did not perform well on feature classification. The LSTM model outperformed all the other architectures in both in accuracy and efficiency albeit by a narrow margin. Below are the matrices for each of the architectures and as we can see we don't see any signs of strong diagonals making the experiment observations less than optimal.

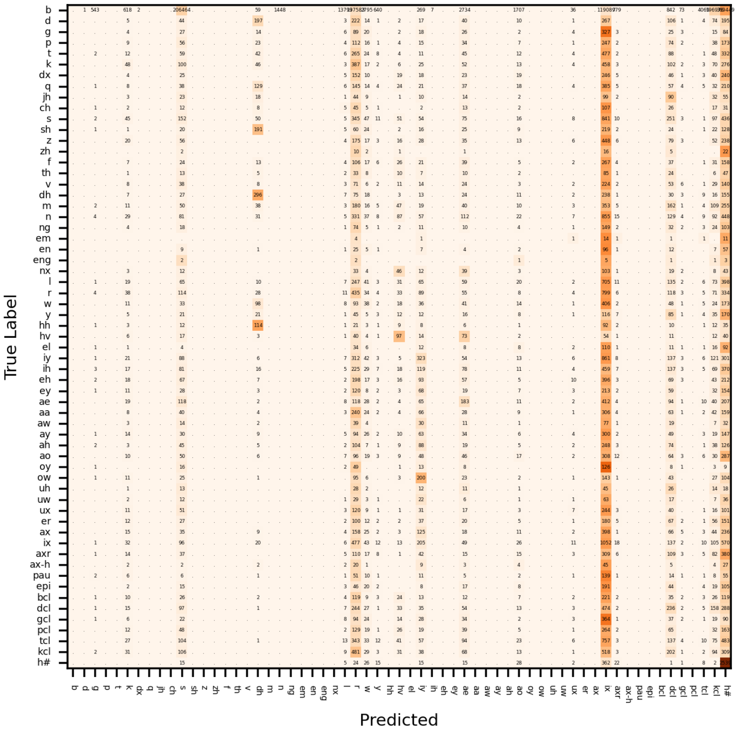
GRU: (loss: 0.4547 - categorical_accuracy: 0.0130)



CNN: (loss: 0.2062 - categorical_accuracy: 0.0043)



GRU: (loss: 0.4508 - categorical_accuracy: 0.0134)



Discussion

The experiment revealed a few findings that were less than optimal however, we could still observe that the LSTM had an edge over the other architectures. Error analysis also reflected the common confusion in recognition of phonemes that had an acoustic similarity. As the writer of this report I hold a slight suspicion that the feature manipulation in the matrices could have been skewed as part of the experimentation. It is clear that GRUs and LSTMs hold an architectural edge over its convolutional counterparts.

Conclusion

The investigation sheds light at the complexities of phoneme classification tasks. LSTMs and GRUs tend to remember more nuanced information for a longer duration where as CNNs excel at robust feature extraction. A future study to exploit the strengths of each of these architectures in a hybrid manner could yield fruitful results.

References

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L. and Zue, V. (1993). Timit acoustic-phonetic continuous speech corpus, pp. 8. Philadelphia, PA: Linguistic Data Consortium, Univ. of Pennsylvania.

I promise that the attached assignment is my own work. I recognize that should this not be the case, I will be subject to penalties as outlined in the course syllabus. [Rahul Gupta]