Introduction

- Question answering and reading comprehension are crucial areas of research in Natural Language Processing, with many practical applications including digital assistants (Siri, Alexa) and web search.
- The Stanford Question Answering Dataset (SQuAD) 2.0 task tests both a system’s ability to answer reading comprehension questions and to determine when a question cannot be answered given the provided passage.
- A Bidirectional Attention Flow (BiDAF) baseline model (Figure 1) without the original model’s character embedding layer was used as our baseline.
- We made several modifications to the baseline, including augmentations to the word embedding layer, addition of a character embedding layer, and replacing the LSTMs with GRUs. We ensembled several of our models together to form our ‘ORCHESTRA’ model, which achieved a max F1 score of 67.00 and EM score of 63.86 on the test set.

Approach

1. Augmenting the representations for the input layer by adding a character embedding layer (CharEmb) and by concatenating the input GLoVe word embeddings with a token’s word2vec embeddings (w2v+GLoVe).
2. Using GRUs in place of LSTMs in the BiDAF Model. Three such models were produced (1GRU, 2GRU, 3GRU), with LSTMs and GRUs mixed together (see Table 1 for which layers used LSTM vs. GRU).
3. Ensembling the models from 1. and 2. together (see Table 3 for permutations of models used for ensembling).

Experiments and Results

- F1 and EM metrics were used for evaluation. For examples that lack an answer, F1 and EM are defined as 1 if the model correctly predicts no answer, and 0 if the model predicts there to be an answer.
- We trained all models for 30 epochs with a fixed learning rate of 0.50 Batch gradient descent with batch size of 64 was employed, and dropout probability was set at 0.2 for all experiments.
- We trained all models for 30 epochs with a fixed learning rate of 0.50.

Dataset

- SQuAD 2.0 Sample Task
  - Context: However, some computational problems are easier to analyze in terms of more unusual resources. For example, a non-deterministic Turing machine is a computational model that is allowed to branch out to check many different possibilities at once.
  - Question: What type of Turing machine can be characterized by checking multiple possibilities at the same time?
  - Answer: Non-Deterministic

- Training: 129,941 labeled training examples (question/context/answer) from the official SQuAD 2.0 dataset.
- Development: 5,951 Examples (about half) of the official SQuAD 2.0 development dataset.
- Testing: The remaining official SQuAD 2.0 dataset examples, with additional examples from the course teaching staff (5,915 examples).

Analysis and Discussion

- Common Types of Errors in ORCHESTRA based off of analysis of a subset of ORCHESTRA’s outputs for the CS224N SQuAD 2.0 development set:
  - Answering Unanswerable ‘When’ Questions
  - Answer Content Correct, span does not precisely match provided answer
  - Inclusion of Adjectives Irrelevant to the Question
- Broadly, F1 and EM scores increased with the number of ensembled models

Conclusions and Next Steps

- Combining multiple word embedding models, adding a character embedding layer, and replacing LSTMs with GRUs improved the baseline’s performance.
- As our best ensemble model, ORCHESTRA performed similarly on the test and dev sets (less than 0.5 difference in F1 and EM scores), indicating that significant overfitting to the dev set did not occur.
- Further work may involve tuning the hyperparameters of the models we used in our ensemble. Our experiments with decaying learning rates were inconclusive: using a cyclic learning rate that decays over time but rises once performance starts to plateau may be beneficial.

References