CharBiDAF with Self-Attention on SQuAD

Zahra Abdullah (zahraab), Darrith Phan (darrithp)
CS 224N (Natural Language Processing with Deep Learning), Stanford University

Introduction

- The SQuAD challenge is a question answering task. It provides a measure for how well a system "understands" a piece of text. Question answering systems can help humans to quickly extract pertinent information from complex documents.
- We implemented a bi-directional attention flow (BiDAF) model with character-level word embeddings and self-attention.
- We also experimented with using GRUs in place of LSTMs and various hyperparameter adjustments.

Problem Statement

- Input: (C, Q) where the context C and the query Q are some lengths of text.
- Output: N/A if question is not answerable (start, end) where start and end are indexes into the context. The context slice from start to end is then the predicted answer.

Data

Dataset: SQuAD v2.0 Dataset
- SQuAD 2.0 is a reading comprehension dataset of context paragraphs (from Wikipedia), questions, and answers (crowdsourced using AMT).
- There are around 150k questions in total.
- About half the questions cannot be answered from the context.
- The answer for an answerable question is a span of text directly from the context.
- Each answerable question has 3 answers provided (from different AMT responders).

Experiments

- Ran BiDAF default model.
- Next, we ran the BiDAF model with character-level word embeddings.
- Finally, we combined the previous model with the custom self-attention encoder block.

Training Parameters

- 129,941 examples in the training set, 6078 examples in the dev set, and 5915 in the test set.
- Number of Epochs = 30, Batch Size = 64.
- Varied learning rate = 0.3, 0.5, 0.7, 0.9.
- Varied dropout = 0.1, 0.2, 0.3.
- Experimented with Adadelta Optimizer and Adam Optimizer.

Model

- Embedding Layer: Converts each word in the context and query to a character-level and a word-level word embedding, which are concatenated and fed to a highway network.
- Encoder Layer: Applies a bi-directional LSTM to the output of the embedding layer.
- Self-Attention Block: Based on the QANet Encoder Block (without the position encoding layer). The self-attention layer uses Multi-Head Attention with 8 heads.
- Modeling Layer: Applies a bi-directional LSTM to the output of the embedding layer.
- Self-Attention Block: v3 Again.
- Output Layer: Produces two vectors of probabilities (start and end probabilities) corresponding to each position in the context.

Results

- Incorporating character-level word embeddings gives a large improvement on the baseline model.
- Implementing self-attention caused a small drop in performance from CharBiDAF, but this model was still well above the baseline.
- It’s possible that further exploration of the hyperparameter space could yield a self-attention model that is better than CharBiDAF.

Conclusions

- We would like to thank Chris Manning and the rest of the teaching staff for CS 224N for their advice and mentorship.

Acknowledgements

- The SQuAD challenge is a question answering task. It provides a measure for how well a system "understands" a piece of text. Question answering systems can help humans to quickly extract pertinent information from complex documents.
- We implemented a bi-directional attention flow (BiDAF) model with character-level word embeddings and self-attention.
- We also experimented with using GRUs in place of LSTMs and various hyperparameter adjustments.

References

- The SQuAD challenge is a question answering task. It provides a measure for how well a system "understands" a piece of text. Question answering systems can help humans to quickly extract pertinent information from complex documents.
- We implemented a bi-directional attention flow (BiDAF) model with character-level word embeddings and self-attention.
- We also experimented with using GRUs in place of LSTMs and various hyperparameter adjustments.

Figure 1: Example of a question and context paragraph taken from the default project handout.

Figure 2: Self-Attention Block

Figure 3: Example Input/Output

Figure 4: Quantitative Evaluation Plots

Figure 5: Evaluation on Test Set
These are the results on the test set of the best-performing versions of our two models.

Figure 6: Example Input/Output
CharBiDAF with self-attention was the only model to predict the output because of its better understanding of context.

Dataset: SQuAD v2.0 Dataset
- SQuAD 2.0 is a reading comprehension dataset of context paragraphs (from Wikipedia), questions, and answers (crowdsourced using AMT).
- There are around 150k questions in total.
- About half the questions cannot be answered from the context.
- The answer for an answerable question is a span of text directly from the context.
- Each answerable question has 3 answers provided (from different AMT responders).

Figure 2: Self-Attention Block

Model

- Embedding Layer: Converts each word in the context and query to a character-level and a word-level word embedding, which are concatenated and fed to a highway network.
- Encoder Layer: Applies a bi-directional LSTM to the output of the embedding layer.
- Self-Attention Block: Based on the QANet Encoder Block (without the position encoding layer). The self-attention layer uses Multi-Head Attention with 8 heads.
- Modeling Layer: Applies a bi-directional LSTM to the output of the embedding layer.
- Self-Attention Block: v3 Again.
- Output Layer: Produces two vectors of probabilities (start and end probabilities) corresponding to each position in the context.

Results

- Incorporating character-level word embeddings gives a large improvement on the baseline model.
- Implementing self-attention caused a small drop in performance from CharBiDAF, but this model was still well above the baseline.
- It’s possible that further exploration of the hyperparameter space could yield a self-attention model that is better than CharBiDAF.

Conclusions

- We would like to thank Chris Manning and the rest of the teaching staff for CS 224N for their advice and mentorship.

Acknowledgements

- The SQuAD challenge is a question answering task. It provides a measure for how well a system "understands" a piece of text. Question answering systems can help humans to quickly extract pertinent information from complex documents.
- We implemented a bi-directional attention flow (BiDAF) model with character-level word embeddings and self-attention.
- We also experimented with using GRUs in place of LSTMs and various hyperparameter adjustments.

References

- The SQuAD challenge is a question answering task. It provides a measure for how well a system "understands" a piece of text. Question answering systems can help humans to quickly extract pertinent information from complex documents.
- We implemented a bi-directional attention flow (BiDAF) model with character-level word embeddings and self-attention.
- We also experimented with using GRUs in place of LSTMs and various hyperparameter adjustments.