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
  - $\{i_{start}, i_{end}\}$ where i_{start} and i_{end} are indexes into the context. The context slice from i_{start} to i_{end} is then the predicted answer.

Data

Question: Why was Tesla returned to Gospic?
Context paragraph: On 24 March 1879, Tesla was returned to Gospic under police guard for not having a residence permit. On 17 April 1879, Milutin Tesla died at the age of 60 after contracting an unspecified illness (although some sources say that he died of a stroke). During that year, Tesla taught a large class of students in his old school, Higher Real Gymnasium, in Gospic.
Answer: not having a residence permit

Figure 1: Sample <Question, Context, Answer> Triple

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

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)

Methods

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

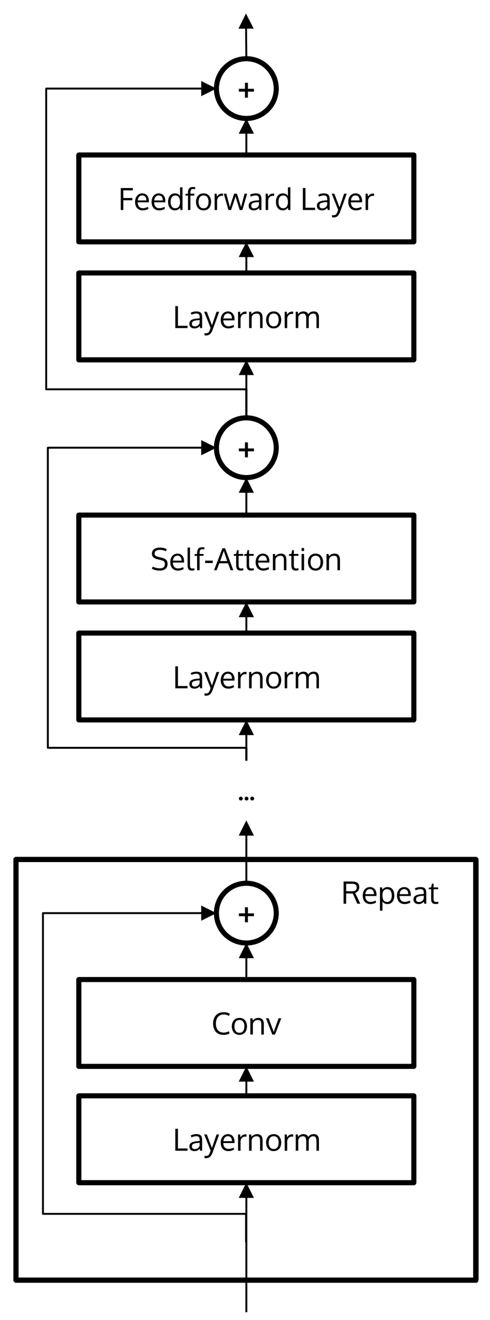
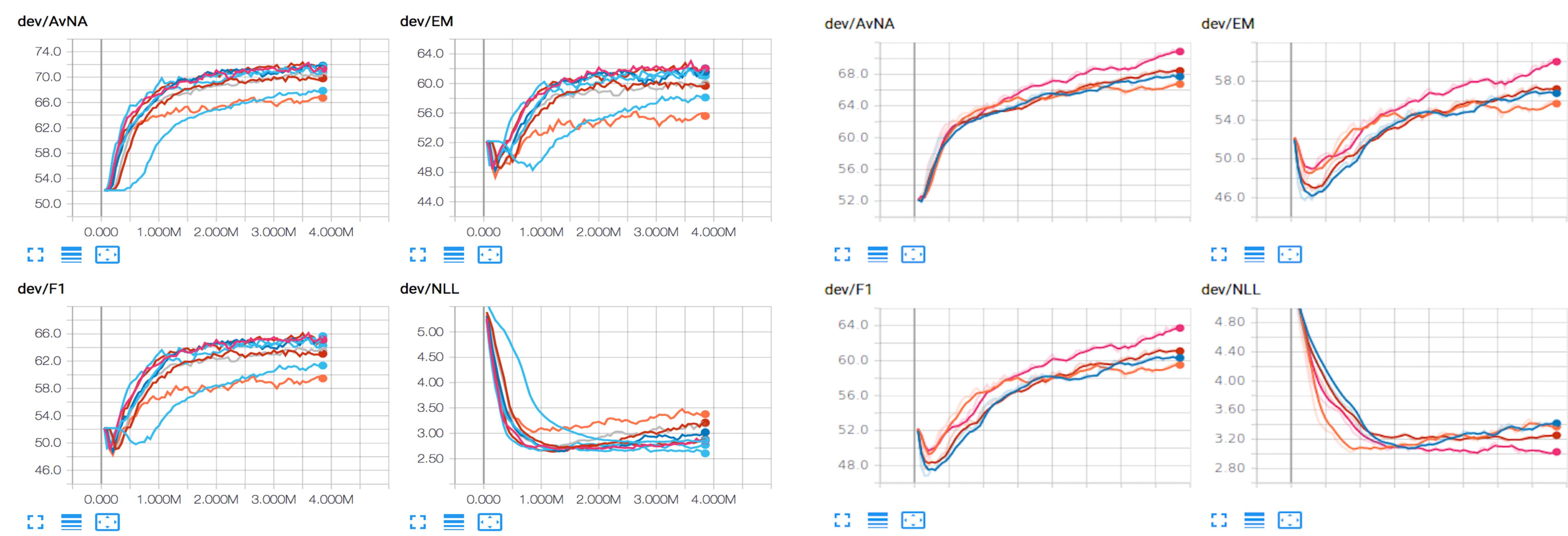


Figure 2: Self-Attention Block

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.
- **Context-Query Attention Layer:** Models context-to-query and query-to-context attention.
- **Modeling Layer:** Applies a bi-directional LSTM to the output of the embedding layer.
- **Self-Attention Block:** x3 Again
- **Output Layer:** Produces two vectors of probabilities (start and end probabilities) corresponding to each position in the context.

Results



CharBiDAF | **CharBiDAF with Self-Attention**

Figure 4: Quantitative Evaluation Plots

Answer vs. No-Answer, Exact Match, F1, and Negative Log Likelihood plots for various versions of the CharBiDAF and CharBiDAF with self-attention models. CharBiDAF outperforms CharBiDAF with self-attention on all these metrics. Both models outperform the baseline.

Model	LR	Dropout	F1	EM
Baseline	0.5	0.2	59.77	56.21
CharBiDAF	0.9	0.3	**64.938**	**61.302**
CharBiDAF + Self-Attention	0.99	0.10	**63.167**	**58.833**

Figure 5: Evaluation on Test Set

These are the results on the test set of the best-performing versions of our two models.

- **Question:** Who designed the garden for the University Library?
- **Context:** Another important library – the University Library, founded in 1816, is home to over two million items. The building was designed by architects Marek Budzyński and Zbigniew Badowski and opened on 15 December 1999. It is surrounded by green. The University Library garden, designed by Irena Bajerska, was opened on 12 June 2002. It is one of the largest and most beautiful roof gardens in Europe with an area of more than 10,000 m2 (107,639.10 sq ft), and plants covering 5,111 m2 (55,014.35 sq ft). As the university garden it is open to the public every day.
- **Answer:** Irena Bajerska
- **Prediction:** Marek Budzyński and Zbigniew Badowski

Figure 6: Example Input/Output

CharBiDAF with self-attention was the only model to predict the output because of its better understanding of context.

Conclusions

- 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.
- There is great leeway in how we incorporate self-attention into the model. Tweaking our implementation could improve results.

Acknowledgements

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

References

- Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. {\it arXiv preprint arXiv:1611.01603}, 2016.
- R-Net: Machine Reading Comprehension with Self-Matching Networks\\ \url{https://www.microsoft.com/en- us/research/wp-content/uploads/2017/05/r-net.pdf}
- Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. Qanet: Combining local convolution with global self-attention for reading comprehension. arXiv preprint arXiv:1804.09541, 2018.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. Attention Is All You Need. arXiv preprint arXiv:1706.03762, 2017.