Positioning Self in SQuAD

Stanford Computer Science

CS 224N

Ankit Baghel, Neel Yerneni {abaghel, nyerneni}@stanford.edu

Task and Overview

- **Background**: Current question-answering model frameworks entail highly complex architectures
  - Most use expensive RNNs for encoding inputs
  - A number of modularized techniques have been introduced to improve question-answering performance
- **Task**: Create a question-answering system for SQuAD that improves on the baseline accuracy and efficiency of the BiDAF model
- **Proposed Solution:** Incorporate various state-of-the-art techniques into the existing BiDAF framework
  - Character Embeddings, Self-Attention, GRUs instead of LSTMs, and Positional Encodings
  - Explore the use of QANet Encoder blocks over RNNs

Background/Related Work

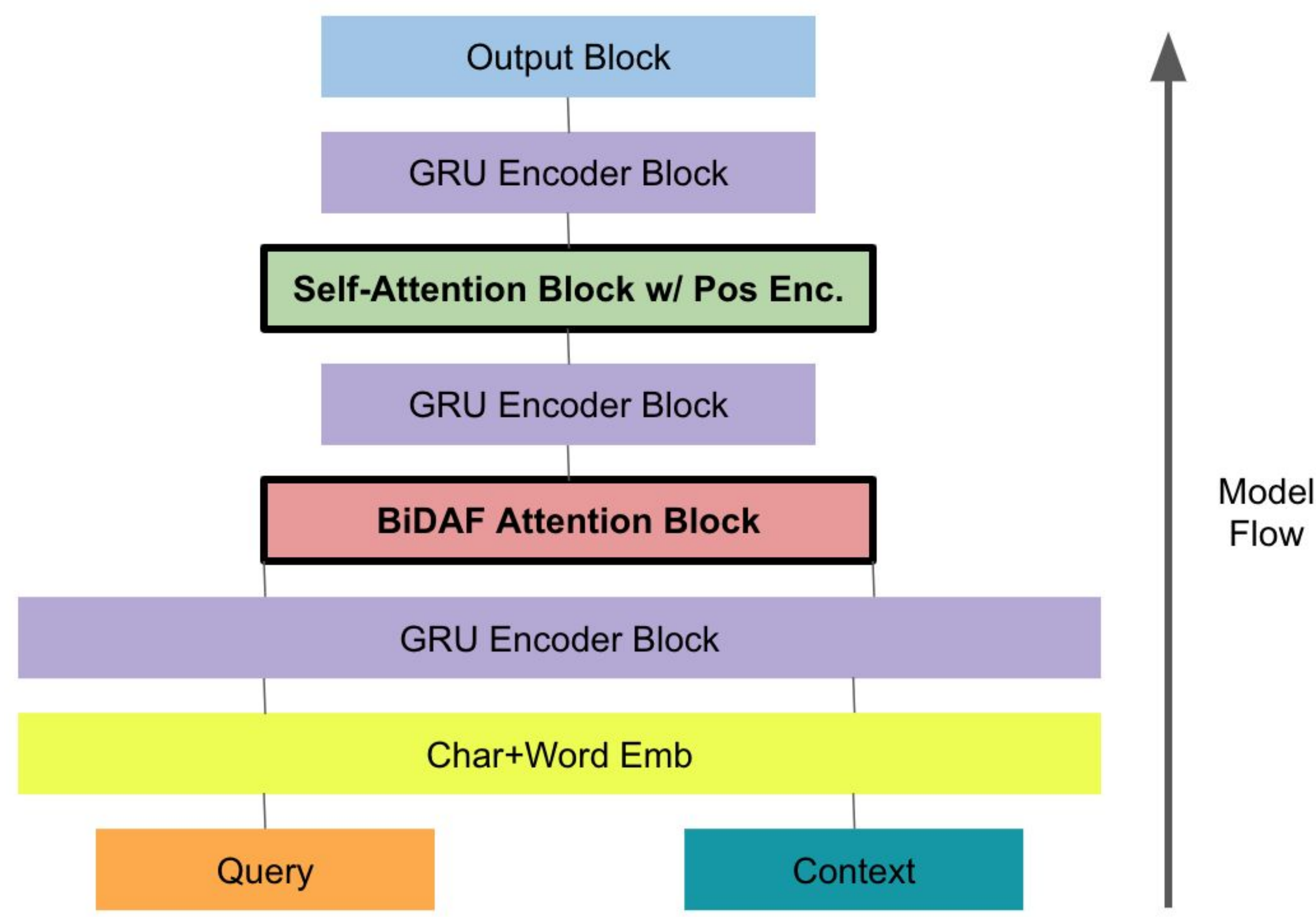
- **BiDAF:** Bidirectional Attention Flow model used as the baseline as introduced in Seo et al.
  - Computes bidirectional attention between the question and context while computing encodings of these attention-weighted inputs via RNNs
- **Self-Attention:** More richly capture the relationships between words in the passage, as introduced in the R-Net paper
  - Computed as follows for some input p:

$$s = p^T(W_1 p + W_2 p)$$
$$a = SoftMax(s)$$
$$q = ap$$

- **Character Embeddings:** Encode pretrained embeddings with 1D Convolutions and feed-forward+highway layer
- **GRUs:** Replace LSTMs with GRUs for faster training time
- **QANet Encoder:** Alternative to RNN encoder in the BiDAF framework
  - Convolutional, multi-self-attention, and feed-forward layers that are stacked and lined with residual connections

Approach

- **Model:** Add a Self-Attention block that also uses positional embeddings while testing various replacements for the intermediate encoder blocks
  - Tested both GRUs and Convolutional Encoders



- **Training**
  - Trained for 30 epochs while utilizing early stopping
  - GRU version was trained with a learning rate of 0.5, hidden dimension of 100, and dropout prob of .2
  - QANet version trained with a learning rate of 0.001/0.0005, hidden size of 96, and multi-self-attention head size of 1 or 4
- **Evaluation:**
  - **F1**: harmonic mean between precision and recall with respect to overlapping words between answers
  - **EM:** Percentage of words that exactly match
  - **AvNA:** Accuracy only with respect to whether the existence of an answer was predicted correctly or not

Results

- Initial Results:

Model	Epoch Training Time	Dev F1	Dev EM	Dev AvNA	Dev NLL
Baseline	18 min	61.29	57.79	68.38	3.1
Baseline + Char Level Embeddings	24 min	63.78	60.19	70.69	3.13
Self_Attention_Before_BiDAF	49 min	63.62	59.74	70.43	3.01
Self_Attention_After_BiDAF	41 min	65.07	61.75	71.4	2.79

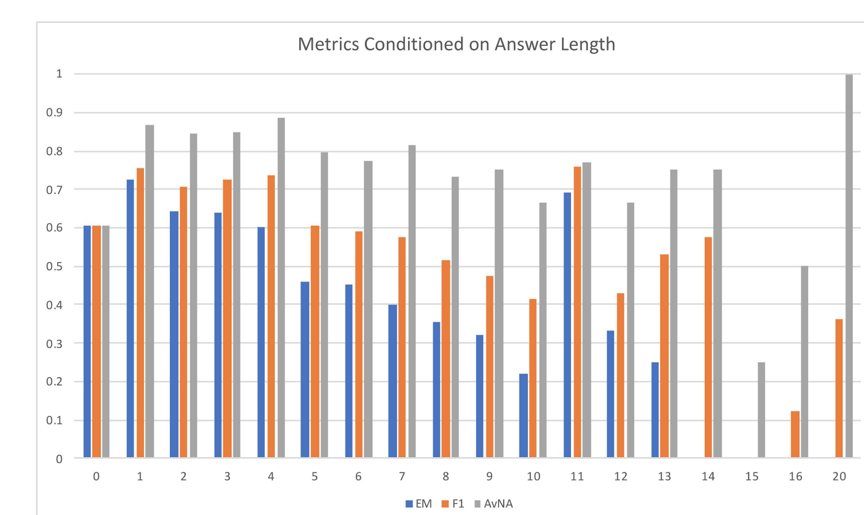
Table 1: Phase 1 Results for Dev Set. (Submitted to the Dev Non-PCE Leaderboard)

- Final Model Results:

Model	Epoch Training Time	Dev F1/EM	Test F1/EM
Self_Attention_Pos_Encoding_BiDAF	41 min	65.528/62.040	64.074/60.575

Table 3: Final Model Results on Non-PCE Leaderboards

- For QANet, we were unable to fully implement the architecture to obtain promising results
  - However, gained insight on effect of learning rate and number of attention heads on performance
- Also explored the effect of answer length on various metrics



Discussion

- Able to make major improvements in BiDAF architecture while minimizing computational costs through the use of GRUs
- QANet, though unsuccessful, was a learning experience and allowed us to modularize techniques like positional encodings to incorporate in other parts of our model
- In the future, we would hope to not only fully implement QANet and measure its performance and efficiency gains, but also fully implement transformers similar to Google's BERT