Positioning Self in SQuAD

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:
    \[ s = p^T (W_p + W_p^T) \]
    \[ u = \sigma(f_{\text{Max}}(s)) \]
    \[ q = s^p \]
- **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
- **BiDAF Attention Block**: Model Flow
- **Output Block**: GRU Encoder Block
- **Self-Attention Block w/ Pos Enc**: GRU Encoder Block
- **QANet Encoder Block**: Char+Word Emb
- **Query**: Context

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**:
  
- **Final Model Results**:
  - 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