Problem Statement

Goal: Build a question-answering system on SQuAD 2.0 with
• BiDAF (character-level embedding)
• QANet
• Transformer-XL

Current Challenges:
• Switch LSTMs to the transformers to enable faster training
• Process the long context paragraphs with Transformer-XL

Problem Setup

Dataset: SQuAD 2.0, including unanswerable questions
Size:
• Training size: ~130k
• Eval size: 6078
• Test size: 5915
Input: Paragraph + Question
Output: Start and end positions that define the answer range

Related Work

BiDAF [1]:
• Word-level embedding layer, bidirectional LSTM encoding layer, CQ+QC attention layer, modelling layer, output layer
QANet [2]:
• Replace the recurrent structure with convolutions and attention; Faster in training
Transformer-XL [3]:
• Segment-level recurrence mechanism + relative positional encoding; Address the gradient vanishing and explosion problem; Model long-term dependencies

Methods

BiDAF
• Word embedding layer (the baseline model does not include character embedding)
• Embedding encoder layer (bidirectional LSTM)
• Attention layer (context-query bidirectional attention flow)
• Modeling layer (bidirectional LSTM)
• Output layer (bidirectional LSTM)

QANet
• Substituted RNNs with encoder blocks, both in embedding encoder and modeling layers
• QANet encoder block:
  • Deepwise separable convolution layers (2 conv layers in embedding encoder, 4 conv layers in modeling layer)
  • Multihed self-attention layer + MLP
  • Layer dropout -> applied after each layer, scaling dropout probability based on layer depth

QANet
• Substituted RNNs with encoder blocks, both in embedding encoder and modeling layers

Transformer-XL

Segment-level recurrence:

Relative positional encoding:

![Transformer-XL Diagram]

Experiments

Results from BiDAF, QANet, Transformer-XL.

<table>
<thead>
<tr>
<th>Model</th>
<th>Test Set</th>
<th>Dev Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>EM</td>
<td>F1</td>
<td>EM</td>
</tr>
<tr>
<td>Baseline</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Char embed</td>
<td>55.20</td>
<td>62.74</td>
</tr>
<tr>
<td>QANet</td>
<td>49.77</td>
<td>52.14</td>
</tr>
<tr>
<td>Transformer-XL</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Analysis

Question: Who ordered Loudoun to defend Louissbourg?
Context: Loudoun, a capable administrator but a cautious field commander, ... He was then ordered by William Pitt, the Secretary of State responsible for the colonies, to attack Louisbourg first. ... Predictions: William Pitt Answer: N/A

Question: Thousands of madrassas spawned what organization?
Context: The Taliban were spawned by the thousands of madrassas the Doebrechts movement established for impoverished Afghan refugees and supported by governmental and religious groups in neighboring Pakistan. ... Predictions: Doebrechts Answer: The Taliban

Conclusions

• Quantitative results did not demonstrate absolute advantage of QANet, which could be due to an increasing number of learnable parameters. But we observe the transformers are better than the RNNs at capturing long-term correlations in the sequence.
• For future work, we expect to: 1) Try different hyperparameters as well as more regularization techniques when training deep transformer networks like the QANet and Transformer-XL. 2) Pretrain on the NMT task then finetuning on the question answering dataset, which might allow the model to better capture the grammatical information in language.