SDNet for SQuAD
Derek Huang, Eli Pugh, Taresh Sethi

Introduction
As contextized embeddings from BERT-centered models dominate the leaderboards in Question-Answering, there is a lot of discussion about what the direction of future models will be. Quite recently, one model named SDNet achieved state-of-the-art on the CoQA dataset using many layers of both self- and inter-attention sequential mechanisms and BERT contextualized embeddings as features. We are very intrigued by this approach, especially since BERT has found to be such a difficult model to expand upon and adapt to receive better results than simpler methods. In this paper, we explore utilizing BERT as a feature, and attention mechanisms that take advantage of a more complex feature space.

Data
SQuAD
• Each example has a context, question, and 3 reference answers or marked unanswerable.
• Train/Dev/Test Split : 129,941 / 6,078 / 5,915
• Adapted from the official SQuAD 2.0 dataset by CS224n staff
CoQA (Additional data for multi-task experiment)
• Each example has a context and question. The example can be marked with reference answers, unanswerable, or answerable with yes/no.
• Past question and answer history within context is utilized
• Taken from the official CoQA challenge datasets

Results

<table>
<thead>
<tr>
<th></th>
<th>FM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Total</td>
</tr>
<tr>
<td>BDAF Baseline</td>
<td>56.296</td>
<td>59.920</td>
</tr>
<tr>
<td>SDNet Baseline</td>
<td>63.968</td>
<td>65.488</td>
</tr>
<tr>
<td>+ Conv Layer</td>
<td>57.963</td>
<td>59.270</td>
</tr>
<tr>
<td>BERT Fine Tuning</td>
<td>56.416</td>
<td>60.040</td>
</tr>
<tr>
<td>Multitask Training</td>
<td>59.175</td>
<td>68.852</td>
</tr>
</tbody>
</table>

• Results are lower than expected, especially given the complexity of the attention layers in SDNet
• F1 scores for all experiments continued to steadily increase; more training time would likely improve performance
• SDNet has a much higher answerable F1 score than baseline

Analysis
Hand-picked examples for qualitative analysis:
• Errors where the predicted response was a sensical alternative to the target response were common
  Context: "... The Islamist groups like Hezbollah in Lebanon and Hamas in Palestine participate in democratic and political process as well as armed attacks, seeking to abolish the state of Israel ...
  Question: "What is the goal of Islamist groups like Hezbollah and Hamas?"
  Target Response: "abolish the state of Israel"
  Predicted Response: "democratic and political process as well as armed attacks"
• Our model seems to favor answering more often, which may be attributed to its complexity
  Context: "Like many cities in Central and Eastern Europe, infrastructure in Warsaw suffered considerably during its time as an Eastern Bloc economy – though it is worth mentioning that the initial three-Year Plan to rebuild Poland (especially Warsaw) was a major success, but what followed was very much the opposite ...
  Question: "What was a major failure, especially in the building of Warsaw?"
  Target Response: "no answer"
  Predicted Response: "Three-Year Plan to rebuild Poland"

Conclusion
While our model scored higher than our BiDAF baseline, we learned that adding lots of complexity over BERT doesn’t improve over more simple methods. While SDNet may fare better on more complex tasks, it was not the most performant on SQuAD.
In the future, we would like to work more on our multitask experiments. We found this area of focus to be very interesting and, and we hope to try to adapt our model for decaNLP tasks. Finally, it would be interesting to replace all the RNNs with CNNs to speed up computation, or test blending different types of CNN architectures into SDNet to achieve higher scores.

References