Reasoning on Multi-Hop Questions with HotpotQA
Jamil Dhanani, Suma Kasa, Suprita Shankar; Mentor: Peng Qi

Introduction
- HotpotQA [1] is a question-answering (QA) dataset focusing on multi-hop reasoning; the answers to the questions come from aggregating parts of the context
- Also requires explanation in the form of “supporting facts” to produce an answer
- We augmented the baseline model for HotpotQA, by proposing improvements in learning and optimization, attention, reasoning, and representation

Example Question
Question: Who is older, Annie Morton or Terry Richardson?
Gold Answer: Terry Richardson
Gold Supporting Facts: Annie Morton, Terry Richardson
Baseline Answer: Annie Morton
Baseline Supporting Facts: Annie Morton, Kenton Richardson

Modifications
- For learning and optimization, we adapted the learning rate decay, added regularization, and tried different optimizers (such as Adam)
- For attention, we added self-attention on the query and context, with shared weights between the two, with multiple architectures
- We tried architectures adding bidirectional attention at various points in the network, to assist with multi-hop reasoning
- Replaced GRUs with Gated CNNs for selected architectures; some work has shown improvement with CNNs for NLP tasks
- Improved representation with architectural changes, adding hidden layers and mitigating bottlenecks by ensuring that inputs to each layer are sufficiently large to represent the information required.

Refined Model Answer
Question: Who is older, Annie Morton or Terry Richardson?
Model Answer: Terry Richardson
Model Supporting Facts: Annie Morton, Terry Richardson

Analysis — Context Attention
- As seen below, in the baseline model, the attention distribution in the self-attention layer is spread out across many different entities in the context paragraphs
- In model Rsn-A, the attention is focused on the correct answer (“Terry Richardson”)

Results Analysis
- On best-best-performing model, improvement of +6.70 points in Ans F1, +14.49 points in Sup F1
- Our Rsn-A model makes significant gains in answers which match in the EM metric; answering correctly an additional 1781 questions, and regressing in only 385
- The number of correct answers with correct supporting facts increases from only 4.93% to 19.75%
- An example question is shown on the left, of a question from the dataset, the baseline answer, and our best model (Rsn-A) answer

Table 4: Results from Best-Performing Models

<table>
<thead>
<tr>
<th>Model</th>
<th>Ans F1</th>
<th>Ans EM</th>
<th>Sup F1</th>
<th>Sup EM</th>
<th>SP Pres</th>
<th>SP Rec</th>
<th>Joint F1</th>
<th>Joint EM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>58.28</td>
<td>44.44</td>
<td>66.06</td>
<td>21.95</td>
<td>65.55</td>
<td>70.00</td>
<td>40.86</td>
<td>11.56</td>
</tr>
<tr>
<td>Opt-C</td>
<td>60.25</td>
<td>45.66</td>
<td>66.12</td>
<td>20.42</td>
<td>66.83</td>
<td>71.11</td>
<td>42.31</td>
<td>10.88</td>
</tr>
<tr>
<td>Att-C</td>
<td>60.03</td>
<td>46.33</td>
<td>69.11</td>
<td>23.52</td>
<td>68.28</td>
<td>76.25</td>
<td>44.09</td>
<td>12.80</td>
</tr>
<tr>
<td>Rep-A</td>
<td>64.47</td>
<td>49.95</td>
<td>75.41</td>
<td>33.12</td>
<td>75.22</td>
<td>81.04</td>
<td>50.92</td>
<td>19.00</td>
</tr>
<tr>
<td>GCNN-B</td>
<td>65.32</td>
<td>49.36</td>
<td>70.78</td>
<td>31.22</td>
<td>73.29</td>
<td>81.94</td>
<td>50.41</td>
<td>20.43</td>
</tr>
<tr>
<td>Rsn-A</td>
<td>65.72</td>
<td>50.91</td>
<td>78.99</td>
<td>37.20</td>
<td>77.26</td>
<td>85.27</td>
<td>53.77</td>
<td>21.24</td>
</tr>
<tr>
<td>Rsn-B</td>
<td>64.88</td>
<td>50.37</td>
<td>78.47</td>
<td>35.62</td>
<td>75.87</td>
<td>85.80</td>
<td>52.93</td>
<td>19.89</td>
</tr>
</tbody>
</table>

Table 5: Comparison of Baseline and Best-Performing Model E

<table>
<thead>
<tr>
<th>Model</th>
<th>Baseline Correct</th>
<th>Baseline Wrong</th>
<th>Rsn-A Correct</th>
<th>Rsn-A Wrong</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>1184</td>
<td>1701</td>
<td>1581</td>
<td>2425</td>
</tr>
<tr>
<td>Rsn-A</td>
<td>1581</td>
<td>2425</td>
<td>1535</td>
<td>2392</td>
</tr>
</tbody>
</table>

Analysis — Query Attention
- In the Bidirectional Attention Layer, the baseline model activations show little prominence for the correct answer in the query, across all context words
- In model Rsn-A, again the attention is focused on the correct answer, in the query

Conclusion & Future Work
- Our changes to the baseline model show significant improvements, bringing the F1 scores from 59.02 to 65.72 with our best model, Rsn-A.
- There is still room for additional hyperparameter optimization to get marginal gains from the reported F1 scores
- We would like to explore memory networks and other models which aid in more complex reasoning

References