Problem Statement

Question answering systems are an exciting but challenging application of Natural Language Processing. While much work has been done on general QA, there is a lack of work in the realm of QA requiring multi-hop reasoning, where the QA system has to reason over information from multiple documents to generate an answer. We aimed to create a multi-hop QA model that utilized novel architecture building blocks to improve upon the publicly available HotpotQA baseline.

Goal:
Train a model that takes in a question requiring multi-hop reasoning, and context paragraphs, and outputs an answer + the supporting facts.

Data & Evaluation

Dataset: HotpotQA Dataset
Statistics: 89,8K train, 7,4K dev
Evaluation: F1, EM

Qualitative analysis (2D CNN, v1)
- Rarely, the model correctly identified both the correct answer and all supporting facts
- Often, the model found the correct answer without identifying any supporting facts

CNN classification module trained in isolation performs comparable to baseline (though lower), lower performance when trained as part of model

Bi-attention processing approach with sigmoid performs reasonably on QA task, but poorly on sup fact classification. Though F1 improves slightly over time, loss also diverges.

2D CNN, v1 best model (applies 2D CNN prior to SP classification rather than after)

Approach

Existing HotpotQA Baseline

CNN Classification module
Independent module for supporting fact classification with 1D CNN

Processed Bi-Attention Model
Bi-attention post-processing + sigmoid/softmax for SP prediction and Q answering

2D CNN Model
Architecture leveraging word-level 2D CNN on self-attention output.

Experiment and Results

Table 2: Score comparison

<table>
<thead>
<tr>
<th>Model</th>
<th>Split</th>
<th>Answer</th>
<th>Fact</th>
<th>EM</th>
<th>F1</th>
<th>EM</th>
<th>F1</th>
<th>Overall</th>
<th>Sup-Fact</th>
</tr>
</thead>
<tbody>
<tr>
<td>HotpotQA baseline</td>
<td>dev</td>
<td>44.44</td>
<td>58.28</td>
<td>21.95</td>
<td>66.66</td>
<td>11.56</td>
<td>40.86</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>train</td>
<td>45.46</td>
<td>58.95</td>
<td>22.24</td>
<td>66.62</td>
<td>11.04</td>
<td>41.37</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CNN classified model</td>
<td>dev</td>
<td>-</td>
<td>18.60</td>
<td>60.24</td>
<td>15.85</td>
<td>56.36</td>
<td>0.092</td>
<td>0.102</td>
<td>30.95</td>
</tr>
<tr>
<td>2D CNN, v1</td>
<td>train</td>
<td>67.71</td>
<td>74.57</td>
<td>9.31</td>
<td>0</td>
<td>6.99</td>
<td>35.38</td>
<td>39.53</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>dev</td>
<td>20.93</td>
<td>33.43</td>
<td>4.92</td>
<td>0</td>
<td>5.15</td>
<td>39.62</td>
<td>34.54</td>
<td>0.19</td>
</tr>
<tr>
<td>Integrated SP CNN</td>
<td>train</td>
<td>43.34</td>
<td>50.64</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6.51</td>
<td>0.19</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>dev</td>
<td>32.86</td>
<td>43.30</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6.38</td>
<td>0.19</td>
<td>0.19</td>
</tr>
<tr>
<td>2D CNN, v2</td>
<td>train</td>
<td>70.96</td>
<td>85.54</td>
<td>3.49</td>
<td>15.69</td>
<td>2.91</td>
<td>13.87</td>
<td>36.22</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>dev</td>
<td>40.78</td>
<td>53.77</td>
<td>2.59</td>
<td>14.17</td>
<td>1.24</td>
<td>8.50</td>
<td>6.03</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>18.36</td>
<td>41.37</td>
<td>6.94</td>
<td>28.05</td>
<td>1.61</td>
<td>13.23</td>
<td>7.92</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8.64</td>
<td>30.44</td>
<td>3.86</td>
<td>26.22</td>
<td>0.90</td>
<td>9.38</td>
<td>9.32</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Qualitative assessment (2D CNN, v1)
- Rarely, the model correctly identified both the correct answer and all supporting facts
- Often, the model found the correct answer without identifying any supporting facts

Results

- CNN classification module trained in isolation performs comparable to baseline (though lower), lower performance when trained as part of model
- Bi-attention processing approach with sigmoid performs reasonably on QA task, but poorly on sup fact classification. Though F1 improves slightly over time, loss also diverges.
- 2D CNN, v1 best model (applies 2D CNN prior to SP classification rather than after)

Conclusions

CNNs seem to be a reasonable architectural building block for this task
Though we were not able to beat the HotpotQA baseline, our best model (2D CNN, v1) used a 2D CNN rather than an RNN and attained lower but comparable overall Answer F1 / EM scores.

Explicit SP classification is not critical for the ultimate QA task
Despite having a low SP score, some models still had a high Answer F1 score, suggesting they were still able to identify supporting facts implicitly despite falling short in explicit identification.

Difficult to optimize SP classification with standard loss calculation
By utilizing a loss function such as ACE loss, we end up minimizing loss by assigning no value to all of the sentences; this lowers the loss, but at the cost of rarely producing a true positive.

Future Work

- Explore alternative techniques for combining question with context. We currently use bi-attention to accomplish this; future work remains to try alternate methods.
- Experiment with deeper 2D CNN layers. We currently only use 1 layer; however, deep layers have proven effective in visual recognition and may also help with this task.
- Hyperparameter tuning. Experimenting with different hyperparameters for the CNN layers (among others) could improve performance.

References