Problem

NLP task: question-and-answering problem

Input:

- Context paragraph with \( n \) words \( C = \{c_1, \ldots, c_n\} \)
- Input question with \( m \) words \( Q = \{q_1, \ldots, q_m\} \)

Output: Span prediction (if no answer available in the context paragraph, the model abstains from answering)

Context: “The 6- and 10-county definitions are not used for the greater Southern California Megaregion, one of the 11 megaregions of the United States. The megaregion's area is more expansive, extending east into Las Vegas, Nevada, and south across the Mexican border into Tijuana.”

Question: “What is the name of the region that is not defined by the eight or 10-county definitions?”

Answer: “Southern California Megaregion”

Methods: QANet

QANet

Methods: BiDAF

BiDAF (baseline)

Data

Data Source: SQuAD 2.0 (Stanford Question Answering Dataset)

- Sourced from Wikipedia articles
- Contains 150,000 questions: 100,000 answerable questions and 50,000 adversarial unanswerable questions
- Split into 3 separate datasets: train, dev, and test.
- Train set has 129,941 examples
- Dev set has 6078 examples
- Test set has 5915 examples

Analysis

QANet v1 did worse than the baseline because we had not yet implemented the encoder block layers

QANet v2: Mis-implemented the separable depthwise convolutions and forgot the ReLU activation and masked the attention maps incorrectly

Self-attention is a key component of the QANet model

- Allows every word in the input attend to every other word
- Places greater emphasis on non-word elements which lead to worse performance.

Ground Truth

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Answer (No Answer)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>2410</td>
<td>1203</td>
</tr>
<tr>
<td>Negative</td>
<td>438</td>
<td>1960</td>
</tr>
<tr>
<td>Total</td>
<td>2848</td>
<td>3103</td>
</tr>
<tr>
<td>Precision</td>
<td>96.13</td>
<td>95.61</td>
</tr>
<tr>
<td>Recall</td>
<td>94.62</td>
<td>61.23</td>
</tr>
<tr>
<td>F1 Score</td>
<td>84.62</td>
<td>58.77</td>
</tr>
</tbody>
</table>

Conclusion

- Achieved good results on the SQuAD 2.0 dataset (dev F1-score of 64.5 and test F1-score of 62.57)
- Dev set has more unanswerable questions than answerable ones, but model predicted an answer more often than no-answer as indicated by the TPR being higher than the TNR
- Limitations: Only tested our model on the SQuAD 2.0 dataset
- Don’t know if QANet model will produce promising results on other Q&A datasets or perform well on other tasks

Future Work

- Incorporate ensemble methods by combining multiple ‘weaker’ learners to build a stronger learner for the task
- Could potentially average start/end probabilities that each model outputs
- Could also imply majority voting whether the most popular start and end tokens amongst the models will be the answer for the predicted span

Experiments

1. All models trained for 2.5 million iterations
2. Note QANet-final used a warmup scheme learning rate where first 1000 steps LR exponentially increases from 0 to 0.001

<table>
<thead>
<tr>
<th>Model</th>
<th>Optimizer</th>
<th>LR</th>
<th>F1-score/dev</th>
<th>F1-score/test</th>
<th>EM-score/dev</th>
<th>EM-score/test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>AdaDelta</td>
<td>0.5</td>
<td>N/A</td>
<td>N/A</td>
<td>55</td>
<td>N/A</td>
</tr>
<tr>
<td>QANet v1</td>
<td>Adam</td>
<td>0.001</td>
<td>50.294</td>
<td>N/A</td>
<td>50.294</td>
<td>N/A</td>
</tr>
<tr>
<td>QANet v2</td>
<td>Adam</td>
<td>0.001</td>
<td>52.193</td>
<td>N/A</td>
<td>52.193</td>
<td>N/A</td>
</tr>
<tr>
<td>QANet final</td>
<td>Adam</td>
<td>0.0001</td>
<td>64.5</td>
<td>62.570</td>
<td>61.334</td>
<td>58.969</td>
</tr>
</tbody>
</table>

References