Introduction

Following the release of the Stanford Question Answering Dataset (SQuAD), rapid innovations have been made in the realm of question answering.

Question: Why was Tesla returned to Gospic?
Context paragraphs: On 31 March 1879, Tesla was returned to Gospic under police guard for not having a residence permit. On 17 April 1879, Milutin Tesla died at the age of 60 after contracting an unspecified illness (although some sources say that he died of a stroke). During that year, Tesla taught a large class of students in his old school, Higher Real Gymnasium, in Gospic.
Answer: not having a residence permit

The dataset provides a passage and an associated question. The purpose of the dataset is to train deep learning models that can comprehend the provided passage and answer the associated question solely using information found within the passage.

Approach

To achieve better results than the baseline model, we will augmented the algorithm in three main ways:

- **Character-Level Embeddings**: to help handle out-of-vocabulary and infrequently used words.
- **Switching to a Self-Attention Layer**: to effectively aggregate evidence from the whole passage to infer the answer.
- **Using a gated recurrent unit (GRU) instead of LSTM**: to help solve vanishing gradient problems.

Related Work

Character Embeddings

R-NET (Self-Attention Layer)

Figure 1. Bidirectional Attention Flow Model from Sori's University of Washington paper [3]. This inspired our character-level embedding implementation (circled in red).

Figure 2. Flow model of Microsoft Research Asia's R-NET [2]. This inspired our self-attention implementation (circled in red).

Results

<table>
<thead>
<tr>
<th></th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>57.08</td>
<td>60.42</td>
</tr>
<tr>
<td>Milestone (Character Embeddings only)</td>
<td>59.02</td>
<td>62.62</td>
</tr>
<tr>
<td>Best dev Score</td>
<td>59.14</td>
<td>62.51</td>
</tr>
<tr>
<td>Final test Score</td>
<td>59.17</td>
<td>62.48</td>
</tr>
</tbody>
</table>

Conclusions

After continuously testing our model, we found that:
- Adding character-level embeddings and lowering batch size improved upon a BIDAF model.
- Replacing an LSTM cell type with GRU also improved performance.
- Self-matching attention to significantly slow down the model, and self-attention did not improve results over the BIDAF attention layer.

Our implementation ultimately beat the baseline score, but fell short of improving upon the performance of Microsoft's R-NET model.

Acknowledgements

Thank you to the entire CS 224N teaching team! Was an amazing class and quarter!

References