Introduction

In the task of reading comprehension or question answering, a model will be given a paragraph and a question about that paragraph, as input. The goal is to answer the question correctly. This is an interesting task as it could be viewed as how well a model can "understand" text. Current end-to-end machine reading and question answering models such as BiDAF[1] can achieve relatively good results. However, given how powerful transformer models are for tasks such as translation and text summarization, we want to further improve the QA system performance by borrowing ideas from transformers.

Methods

In an attempt to improve the QA system performance, we propose the following workflow:

1. Improve on the given BiDAF-based model by adding character level embedding, as it allows us to condition on the internal structure of words (morphology), and better handle out-of-vocabulary words.
2. Use data augmentation to further improve the BiDAF-based model. We used a T-5 based model that could generate questions by simply passing the text.
3. Implement QANet(2) and compare its performance with the BiDAF model.
4. Explore other methods such as ensembling to improve the model performance.

Experiments & Results

Dataset

We run all of our experiments on the Stanford Question Answering Dataset (SQuAD) 2.0 [3] dataset. There are around 150k questions in total, and roughly half of the questions cannot be answered using the provided paragraph. If the question is answerable, the answer is a chunk of text taken directly from the paragraph.

Question: Why was Tesla returned to Gasicp? Context paragraph: ...Tesla was returned to Gasicp under police guard for not having a residence permit. On 17 April 1879 Mumin Tesla died at the age...

Answer: not having a residence permit

Evaluation Metrics

Evaluation is based on Exact Match (EM) score and F1 score, comparing the prediction with three provided answers by the crowd worker, where the maximum score will be considered. We also report AvNa.

- Exact Match is a binary measure of whether the output matches the ground truth answer exactly.
- F1 score is less strict, it is calculated with 2 * prediction + recall / (precision + recall)
- AvNa measures the accuracy when only considering the answer vs. no-answer predictions.

Result

<table>
<thead>
<tr>
<th>Model</th>
<th>EM</th>
<th>AvNa</th>
</tr>
</thead>
<tbody>
<tr>
<td>BiDAF</td>
<td>70.7</td>
<td></td>
</tr>
<tr>
<td>BiDAF + Char Emb</td>
<td>70.7</td>
<td></td>
</tr>
<tr>
<td>BiDAF + Char Emb + Data Augmentation</td>
<td>70.7</td>
<td></td>
</tr>
<tr>
<td>QANet(1) + Char Emb</td>
<td>68.46</td>
<td>74.47</td>
</tr>
<tr>
<td>QANet(2) + Char Emb</td>
<td>64.76</td>
<td>68.78</td>
</tr>
<tr>
<td>QANet(3) + Char Emb</td>
<td>61.5</td>
<td>59.22</td>
</tr>
<tr>
<td>QANet(4) + Char Emb</td>
<td>66.13</td>
<td>72.25</td>
</tr>
<tr>
<td>Ensemble (Average)</td>
<td>70.09</td>
<td>74.36</td>
</tr>
<tr>
<td>Ensemble (Manual)</td>
<td>69.96</td>
<td>67.36</td>
</tr>
</tbody>
</table>

Analysis

1. Dataset analysis

   - Bias between datasets
   - Potential reason why model with data augmentation doesn't work well on the dev set: it decreased the proportion of unanswerable questions

2. Comparing QANet and BiDAF

   - Answer vs. No Answer predictions

   *QANet in this section refers to QANet(3)*

   - BiDAF + Char Emb Prediction
     - Answerable: 37.05% (TP)
     - Unanswerable: 37.05% (FN)
   - QANet(3) + Char Emb Prediction
     - Answerable: 60.20% (TP)
     - Unanswerable: 60.20% (FN)

3. Performance by Question Type

   - Though QANet achieves higher performance on most of question types, BiDAF model perform better on "Who" questions.

4. Performance by Answer Length

   - When answer length is short, both models perform worse.

3. Manual Classification Ensemble Model

   - Use BiDAF when the question is "who" type
   - Use QANet when the question is "why" type
   - Prefer using shorter answers
   - Use "Unanswerable" when either model outputs "Unanswerable"

References

3. https://nlp.stanford.edu/squad