Introduction

Attending in BiDAF and QANet
Lucas Orts, Yanal Qushair, Sophia Sanchez
1 Computer Science Department, Stanford University

Improving QA results in not only better QA services, but better understanding of natural language semantics.

DATASET

Stanford Question Answering Dataset (SQuAD) 2.0.
- A set of (question, context, answer) triples based on Wikipedia text excerpts.
- For each question, the QA model attempts to return an answer to the question that is similar to the human-produced answer based on the context.
- Not all the questions can be answered from the context.

EVALUATION METRICS

The EM and F1 Score as defined in project handout.

RELATED WORK


Model Implementations

BIDAF MODEL

Character Embedding. We implement character-level embeddings to condition on words' internal structure to better handle out-of-vocabulary words. For each word, we concatenate an additional character-level embedding onto the GloVe vectors.

Co-Attention Layer. Based on [3], we implement two-way attention between the context and the question. This involves a second-level attention computation, which attends over representations that are themselves attention outputs.

Self-Attention Layer. Inspired by [4], we implement a self-attention layer, which directly matches the question-answering passage representation against itself using a similarity matrix similar to that of the BiDAF attention layer.

QANet

We implement a QANet model from scratch, based in [2]. The architecture has five layers: (1) Input Embedding Layer, (2) Embedding Encoder Layer, (3) Context-Query Attention Layer, (4) Model Encoder Layer, (5) Output Layer.

Results and Evaluation

EXPERIMENTS

Baseline. Default BiDAF implementation.

<table>
<thead>
<tr>
<th>Model</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>BiDAF</td>
<td>58.01</td>
<td>61.25</td>
</tr>
<tr>
<td>Character Embedding (CE)</td>
<td>60.39</td>
<td>63.71</td>
</tr>
<tr>
<td>CE, Self- and Co-attention</td>
<td>61.25</td>
<td>64.65</td>
</tr>
<tr>
<td>QANet (hidden 128, att. heads 1)</td>
<td>53.77</td>
<td>57.23</td>
</tr>
<tr>
<td>QANet (hidden 64, att. heads 8)</td>
<td>61.27</td>
<td>64.32</td>
</tr>
</tbody>
</table>

Discussion

- Best performing models are BiDAF with CE, Self- and Co-Attention with EM-61, F1-64. More attention helps for BiDAF. Hyperparameter tuning is highly impactful for QANet.
- All models, but especially the BiDAF models, tend to fail most often by giving an answer when there was none. Less obvious for QANet.
- Naturally, more abstract questions like “How” and “Why” are harder for the model. However, the best QANet model does significantly better than this BiDAF.
- Basic BiDAF models (i.e. baseline and CE) do better the larger the context, but CE with Self- and Co-Attention does much better for smaller context windows. QANet model does not seem as impacted by context length.

Plots

Challenges and Future Work

- Limited hyperparameter tuning due to time constraints. Limited model size due to hardware constraints.
- Could ensemble BiDAF and QANet model to leverage their respective strengths.
- Could improve both models with e.g. data augmentation techniques.
- More in-depth analysis of different attention mechanisms using heat maps.