Objectives

The goal of this project it to build reading comprehension systems for the Stanford Question Answering Dataset (SQuAD) 2.0 without pre-trained models. We have three contributions:
1. Improve the baseline BIDAF[1] model.
3. Improve QANet performance with using an input embedding refine layer and a condition output layer

BIDAF

We improved the embedding layer of BIDAF model by introducing learnable character level embedding. We also introduced a fusion function after the co-attention layer to better fuse different attention components. Fused attention is computed as:
\[ A_{\text{fuse}} = \text{ReLU}(W[c, a \odot c; a \odot b) + b] \]

QANet

QANet is a feed-forward model that consists of only convolutions and self-attention. The core building block of QANet is encoder block. It consists of a sinusoidal positional encoding layer, followed by \( \times \) convolution layers, a self multi-head attention layer and a feed-forward layer (Figure 1). QANet adopts the core bi-directional attention idea from BIDA. It is computed as:
Compute similarity matrix \( S \) and normalize over each row and column to get \( \tilde{S} \) and \( \tilde{S} \) respectively
the context-to-query attention is: \( A = \tilde{S} \cdot Q^T \), and the the query to context attention is: \( B = \tilde{S} \cdot \tilde{S}^T \cdot C^T \)

Improved QANet Layers

Input embedding layer with convolution and linear projections
We added two additional 1D convolutions while adopting the linear projection. The first convolution refines character embedding into 128-D hidden size and the second convolution further refines the 256-D concatenated embedding into the final 128-D representation. (Figure 2)

Output layer with conditioning end prediction on start prediction
It’s helpful to know where the answer starts when predicting the end of the answer. We designed the new output layer with this conditioning (Figure 3). The \( P_{\text{start}} \) is computed the same as before, the calculations for \( P_{\text{end}} \) are as follows:
\[ A = [M_0, M_1], B = [M_0, M_2] \]
\[ A_{\text{weights}} = W_A A_2 = \text{ReLU}(W_B B) \]
\[ A_{\text{end}} = A \odot A_{\text{weights}} \]
\[ A_2 = \text{PositionEncoding}(A_2) \]
\[ A_3 = \text{ReLU}(W_A A_3) \]
\[ P_{\text{end}} = \text{softmax}(W_A A_3) \]

In \( A_{\text{weights}} \) words with higher probability of being the answer start will be more activated. The \( A_{\text{end}} \) is then sent through a position encoding function for position information hardening. Finally, the output is used as additional information when predicting the end position.

Experiment Results

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Dev FI</th>
<th>Dev EM</th>
<th>Dev AQA</th>
<th>Test FI</th>
<th>Test EM</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIDAF</td>
<td>60.00</td>
<td>57.07</td>
<td>184.14</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BIDAF + Char Embedding</td>
<td>62.32</td>
<td>59.07</td>
<td>190.67</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BIDAF + Char Embedding + Fusion</td>
<td>62.26</td>
<td>60.52</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>QANet (small)</td>
<td>64.50</td>
<td>61.25</td>
<td>72.13</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>QANet medium (5 blocks, 4 heads)</td>
<td>65.78</td>
<td>61.5</td>
<td>72.09</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>QANet (7 blocks, 3 heads)</td>
<td>66.35</td>
<td>62.07</td>
<td>71.18</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>QANet large (9 blocks, 3 heads)</td>
<td>65.15</td>
<td>61.06</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>QANet large (510 hidden size)</td>
<td>66.10</td>
<td>64.75</td>
<td>71.70</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Improved QANet

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Dev FI</th>
<th>Dev EM</th>
<th>Dev AQA</th>
<th>Test FI</th>
<th>Test EM</th>
</tr>
</thead>
<tbody>
<tr>
<td>QANet (small)</td>
<td>64.10</td>
<td>65.94</td>
<td>72.90</td>
<td>67.81</td>
<td>68.82</td>
</tr>
</tbody>
</table>

Analysis

We break down questions by common question words. We observed relatively consistent performance across all categories (61.9-79.1 EM, 66.2-79.17 F1). The top 3 model performing categories are "Whose", "When" and "Who"

The 3 least performing categories are "Where", "How" and "Why". These are naturally more difficult to answer because these categories require reasoning.

Reference