Machine reading and question answering (Q&A) is essential for evaluating how well computer systems understand human languages.

We investigated an improved version of BiDAF model. We combined character embedding, self-attention and average-attention layers to a BiDAF model using GRU network to improve the accuracy of the baseline model. We experimentally proved that adding our learnable weighted average-attention layer is beneficial based on the significant improvement of model performance and negligible extra computational cost.

SQuAD 2.0

- More than 100,000 question-answer pairs on more than 500 articles, and more than 50,000 unanswerable questions.
- The original SQuAD dataset has three splits, train, dev, and test, with the first two publicly accessible and the last one held privately.
- The original dev set was divided into two, one for dev and another for test.
- We also analyzed the amount of questions that start with different key words (“how”, “what”, “why”, “which”, “who”, “where”, “when”), as shown in Figure 1.
- “what” dominates both training set and dev set.

Figure 1: Dataset composition on different question types

Methods and Results

- Adding character-level embedding to the baseline, an increase in both F1 and EM score has been observed.
- Learnable weighted average-attention was then added to the embedding layer to further boost the performance.
- To decrease the training time, we leveraged GRU to accelerate the training process.
- After adding self-attention layer, the training time increased to 18 hours
- Adam were chosen to stabilize the training
- The performance of two-layer RNN model is better than the one-layer model during the first few iterations, and then becomes worse for the rest of the time. (overfitting)

Figure 2: An overview of our model architecture. CE stands for char-embedding. WE stands for word-embedding.

Figure 3: The experimental results of two models with one-layer and two-layer of embeddings.

Figure 4: F1 score of QANet for comparison

Ablation Study

<table>
<thead>
<tr>
<th>Model</th>
<th>F1</th>
<th>EM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline + Char Embedding</td>
<td>60.758</td>
<td>57.469</td>
</tr>
<tr>
<td>Baseline + Char Embedding + Weighted average-attention</td>
<td>64.734</td>
<td>61.049</td>
</tr>
<tr>
<td>Baseline + Char Embedding + Weighted average-attention + Self-attention</td>
<td>66.241</td>
<td>62.679</td>
</tr>
<tr>
<td>QANet</td>
<td>64.013</td>
<td>63.235</td>
</tr>
</tbody>
</table>

Table 1: Model results at each implementation level

The results of the BiDAF models at every improvement level is shown in Table 1. The last row of the table is the results of QANet.

QANet serves as a comparison guideline, we did not submit QANet result to the leaderboard, the result is from our own evaluation.

Conclusion

- Adding character-embedding, weighted average-attention, and self-attention can boost the performance.
- The best F1 (66%) and EM (62%) scores were from our self-attention model.
- Adding extra layers might not help but make neural network more difficult to train.

Future Work

- Try different optimizer methods, such as AdaBound
- Add Elmo pre-trained embeddings
- Implement Transformer and Transformer-XL

Reference