**Introduction & Overview**

- Reading comprehension has always been a challenging task in machine learning and natural language processing. In reading comprehension task, we are given a context and a query, and our goal is to predict a segment of the context that represents the answer to the query.
- In this paper, we built upon the baseline BiDAF model with the addition of character-level embedding and self-attention. We also adapted some state-of-the-art model architecture such as DRQA to our baseline model.

**Dataset**

- We used Stanford Question Answering Dataset (SQuAD 2.0) to train and test our model, where the dataset comes from the questions posed by crowdworkers on Wikipedia articles and answers by a span, quoting a segment of the text. The dataset consists of 100,000 questions from SQuAD 1.1 with over 50,000 new, unanswerable questions in order for the model to be adaptive to problems with no answers. For this project, we are only given training set and dev set for SQuAD 2.0 because the test set is kept secret and we split the dev set into dev set and test set. For each question in the dataset, we are given three gold answers which we will make comparison with our predicted result.
- **Some examples of the dataset:**
  - Context: “Both B cells and T cells carry receptor molecules that recognize specific targets. T cells recognize a “non-self” target, such as a pathogen, only after antigens (small fragments of the pathogen) have been processed and presented in combination with a “self” receptor called a major histocompatibility complex (MHC) molecule. There are two major subtypes of T cells: the killer T cell and the helper T cell. In addition there are regulatory T cells which have a role in modulating immune response. Killer T cells only recognize antigens coupled to class I MHC molecules, while helper T cells and regulatory T cells only recognize antigens coupled to class II MHC molecules. These two mechanisms of antigen presentation reflect the different roles of the two types of T cell. A third, minor subtype are the T cells that recognize intact antigens that are not bound to MHC receptors.”
  - Question: “What kind of T cells have the purpose of modulating the immune response?”
  - Answer: “regulatory T cells”

**Baseline Models**

- The model is provided with a BiDAF model with the following highlighted layers:
  - 1. Word embedding layer
    - We used GloVe pre-trained word vectors to retrieve vector representation of each word in context and query.
  - 2. Contextual embedding layer
    - The contextual embedding layer will further extract embeddings of the words based on the surrounding words in the context. The baseline model used bi-directional LSTM network to refine 300 dimensional pre-trained GloVe word vector into context based word vectors.
  - 3. BiDAF attention flow layer
    - The BiDAF attention flow layer is mainly responsible for drawing connections between the context word vector and query word vector obtained from contextual embedding layer

**Our improvements**

- **Character-level embedding**
  - We adapted the character embedding from assignment 5 that allows our baseline model to both generate embedding vectors for both word-level and character-level. In order to feed both vectors into our encoder. We concatenate the word and character vectors and feed them to a highway layer. Highway layer is inspired by the gates of an LSTM. It adaptively carry some dimension of the input directly to the output. Because we concatenated the word and character vectors into one vector, the hidden size of the resulting output doubled; we therefore added a linear projection layer to halve the hidden size.
  - **Self-attention layer**
    - Self-attention can connect distant words through shorter network paths. Attention defines what portion of the input that the model focuses on, and it is largely used on perception in its early days. Self-attention refers to the attention that words pay to each other within a sentence. At every time step of an RNN, a weighted average of all the previous states are used as inputs to compute the next state.
  - **POS embedding and NE embedding**
    - In addition to self-attention layer, another improvement is to add POS(part of speech) and NE(name entity) tag as additional feature to the word embedding according to. We used n-gram to generate POS and NE tags for each word in sentence and use zero index as word padding. We also treat OOV as padding index for both POS and NE tags. This addition improves about 1.5 in both F1 and EM scores.
  - **No-answer layer**
    - We have also added no-answer layer. It predicts "no-answer" when both start index and end index equal to zero that yields the highest probability. In our approach, we add another output layer which outputs score that gives us the probability whether the question is answerable. Therefore, our revised objective function becomes Equation. However, in our experiment, this modification does not produce much difference compared with baseline model.

\[
NLL = -\log \left( \sum_{i=1}^{n} \frac{e^{x_i + b_i}}{\sum_{j=1}^{n} e^{x_j + b_j}} \right) \quad [1]
\]
\[
Loss = -\log \left( \frac{(1 - \delta)^2 e^{x_1 + b_1} + \delta e^{x_i + b_i}}{e^{x_1 + b_1} + \sum_{i=1}^{n} e^{x_i + b_i}} \right) \quad [2]
\]

**Analysis**

We have also discovered some wrong answers when the model falsely aligned the attention between the context and the question.

**Context:** "Advances in polynomial algebra were made by mathematicians during the Yuan era. The mathematician Zhu Shijie (1249–1314) solved simultaneous equations with up to four unknowns using a rectangular array of coefficients, equivalent to modern matrices. Zhu used a method of elimination to reduce the simultaneous equations to a single equation with only one unknown. His method is described in the Jade Mirror of the Four Unknowns, written in 1303. The opening pages contain a diagram of Pascal’s triangle. The summation of a finite arithmetic series is also covered in the book.”

**Question:** “What modern math concept did Zhu Shijie do work similar to?” Ground truth: “matrices”

**Predicted:** "Jade Mirror of the Four Unknowns" The model fails to align "work similar to" in the question with "equivalent to" in the context. Consequently, the model predicts the wrong answer. One possible improvements to overcome this problem is to try to use convolution-based attention layer to replace our attention layer in baseline model.

**Conclusion**

In this paper, we have made several modifications on the baseline model. Specifically, we have added character-level embedding, self-attention layer, POS embedding and NE embedding, and no-answer layer to our baseline model. The results showed improvements on each modification that we made, with DrQA being the best improvement. The final EM and F1 scores we have on test non-PCE leaderboard are 60.152 and 63.612, respectively. Throughout this project, we have learned the ability to improve the model performance by researching on state-of-the-art methods and analyzing the advantages and disadvantages of different modifications.

**Reference**


**Results**

<table>
<thead>
<tr>
<th>Model Variations</th>
<th>EM score</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline model</td>
<td>56.298</td>
<td>59.920</td>
</tr>
<tr>
<td>Self-attention</td>
<td>57.16</td>
<td>60.25</td>
</tr>
<tr>
<td>DrQA</td>
<td>59.02</td>
<td>62.38</td>
</tr>
<tr>
<td>No-answer layer</td>
<td>57.03</td>
<td>60.08</td>
</tr>
<tr>
<td>Self-attention+DrQA</td>
<td>59.08</td>
<td>62.55</td>
</tr>
<tr>
<td>Ensemble</td>
<td>60.21</td>
<td>63.40</td>
</tr>
</tbody>
</table>

We have fine-tuned the model on dev set into test set. For each question in the dataset, we are given three gold answers which we will make comparison with our predicted result.

**CS224N Final Project**

Taiming Zhang (tzhang55), Liuming Zhao (lxz299)