**Introduction**

Reading and comprehending the human languages is a challenging task for machines, which requires understanding of natural languages and the ability to do reasoning over various clues. Question answering (QA) is one of popular problems in this field and has been actively researched in Natural Language Processing. QA has gained great significance and popularity since it has a wide range of applications, such as web search, e-learning and interactive voice response. In this project, we focus on designing a question answering system that has good performance on SQuAD 2.0.

**Data**

- SQuAD 2.0 dataset
- 150,000 questions in the format of question, context, answer
- 50% of the questions that can be answered
- 50% of the questions are not answerable using the given paragraph
- paragraphs are from Wikipedia
- 50% of the questions in the format of <question, context, answer>
- 150,000 questions in the format of <question, context, answer>

**Approach**

- BiDAF
- Character Embeddings
- Self-Attention Layers
- From BiDAF++
- From RNet
- BERTs
- Ensemble method

**Experiments**

- The baseline BiDAF model with a character-level embedding layer performs a little better than the one without.
- The BERT-Base-Uncased model performs significantly better than BiDAF and BERT-Large-Uncased model performs better than BERT base model. When we increase the maximum sequence length, the result gets better.
- Self-attention layers do help model represent the question-related-context vector more suitably.

**Models**

![BiDirectional Attention Flow Model](image1)

![Differences in pre-training model architectures](image2)

![High-level structure of the final ensemble model](image3)

**Results**

We chose the ensemble model with the highest dev scores as our final model and tested it on the test set. This final ensemble model with 77.312 EM and 79.984 F1 on the test set is a reasonably good system.

**Analysis**

If the question has keywords that are able to found in context and the answer is around the keywords, the model will correctly answer the question with high probability. If the answer needs not only the context but also some logic, the model will return N/A at most of the time.

**Conclusion**

We introduce an ensemble of self-attention BiDAF models with character embedding and three fine-tuned BERT models. We use 2 types of self-attention layers in BiDAF in different locations. The experimental quantitative evaluations show that our model achieves the state-of-the-art results in SQuAD2.0. Therefore, our model is able to answer non-trivial questions by attending correct locations in the given context.

**Reference**


