Question Answering on the SQuAD Dataset

Jonas Shomorony
Codalab username: jshom
jshom@stanford.edu
CS 224N Winter 2017-2018

Abstract

After reading a paragraph and being given a question about it, humans can rather easily point out and highlight where in the paragraph the answer can be found. This task, however, is not as simple for computer programs. This task is a challenging one for machine learning, in general, because of the difficulty in relating the question to the paragraph. In this paper, we look at one approach to solve this problem, mainly focusing on the BiDAF paper, as a source for inspiration.

1 Introduction

The task of reading comprehension is one that is fairly natural to people, and is being more and more popularized as the popularity of machine learning and artificial intelligence rise. The idea of the problem is as follows: our model is given a paragraph, as well as a question about the paragraph, as input, and the goal is for our model to specify the start and end indices of the words of the paragraph where the answer is found. Of course, for this to work, the answer must be found continuously (we can't use multiple disjoint parts of the paragraph to answer the question), and the answer must be able to be found in that paragraph alone.

My approach to solve this problem was to try to implement various layers as done in the BiDAF paper, to try to maximize F1 score. The model is an improvement of the baseline model, which by itself, contained three components: an RNN encoder layer, that encodes the question and context paragraphs into hidden states, followed by an attention layer, combining the representations of the question and context, and an output layer, that applies a fully connected layer and two softmax layers to then determine the most likely start and end word indices of the context where the answer to the question can be found. My main improvements to the baseline model were an improvement to the attention layer, using bidirectional attention, and the introduction of a new modeling layer, inbetween the attention layer and the ReLU nonlinearity of the output layer.

2 Background and Related Work

As previously suggested, this question answering task has been studied in a lot of detail by NLP researchers of top universities and tech companies and the most successful models have been using Deep Learning approaches to tackle this problem. A very common mechanism utilized in the most successful models is some sort of attention mechanism, which serves as inspiration to my implementation of BiDAF. This attention mechanism was introduced in the Bi-Directional Attention Flow for Machine Comprehension paper, which is, as the title suggests, a Bi-Directional Attention Flow network, which, other than this attention mechanism, uses a multi-stage hierarchical process that represents the context paragraphs in different levels of granularity.

Another well-performing question answering model is the Dynamic Coattention Network, which uses a coattention mechanism to allow the model to focus on important parts of both the question and context, as well as a dynamic pointing decoder that helps the model to not be stuck in local maxima

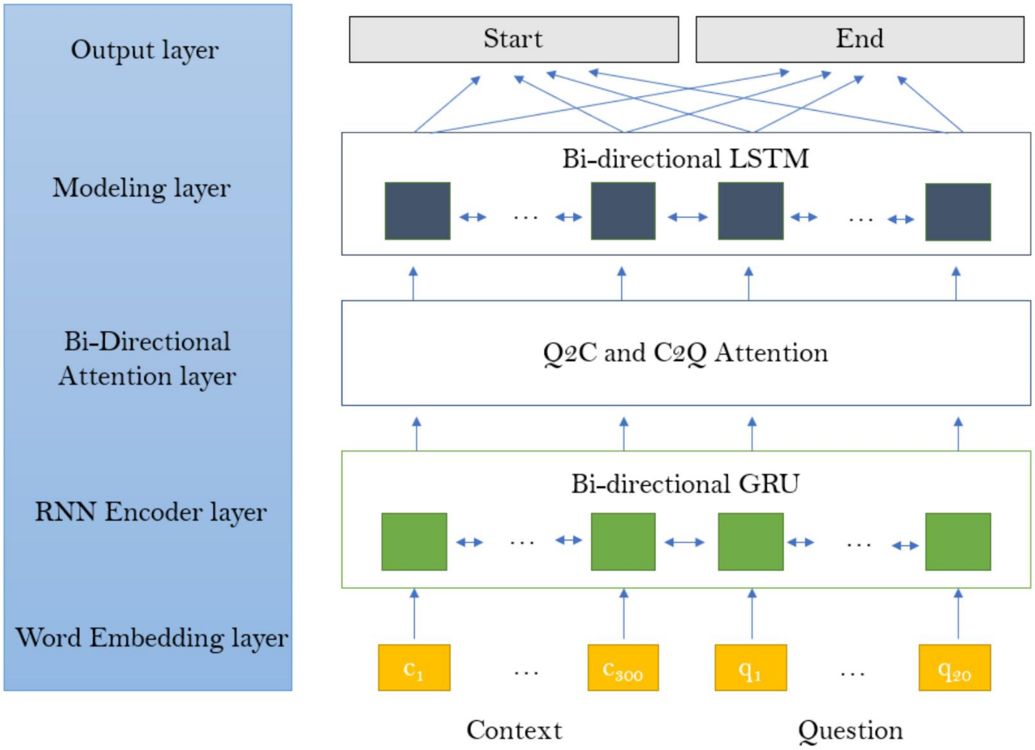


Figure 1: Model workflow

that correspond to incorrect answers. The dynamic pointer decoder makes use of a highway maxout network that computes the probabilities of the starting and ending word indeces of the context. Both of these models (BiDAF and DCN), fully trained and ensembled, were able to reach F1 scores of approximately 80%.

There are many more other models that have been constructed for the question answering task that have been very successful, such as the R-Net, Match-LSTM, Answer Pointer, Dynamic Chunk Reader, DrQA, Stochastic Answer Network, and Gated Attention Reader.

3 Approach

My model used the initial baseline model, and made improvements using inspiration from the Bi-directional Attention Flow Network. The model starts off by using a word embedding layer that maps individual words to vectors in a high-dimensional space, followed by an RNN encoder layer to encode the question and context into hidden states, followed by a bi-directional attention layer that is able to combine information from both the context and the question, followed by a modeling layer is able to capture relationships between the context and the question, and finally followed by the output layer. Now that we've introduced the model, we spend some time going over the details:

1. **Word Embedding layer**

   We use 100-dimensional pre-trained GloVe embeddings to represent words, which do not change during training. The contexts and questions are represented by a concatenation of the word embeddings corresponding to the words in each text. These embeddings are then passed into the RNN encoder layer.

2. **RNN Encoder layer**

   We use a 1-layer bidirectional GRU, which produces a sequence of forward and backward hidden states for both the context and question. We then concatenate forward and backward

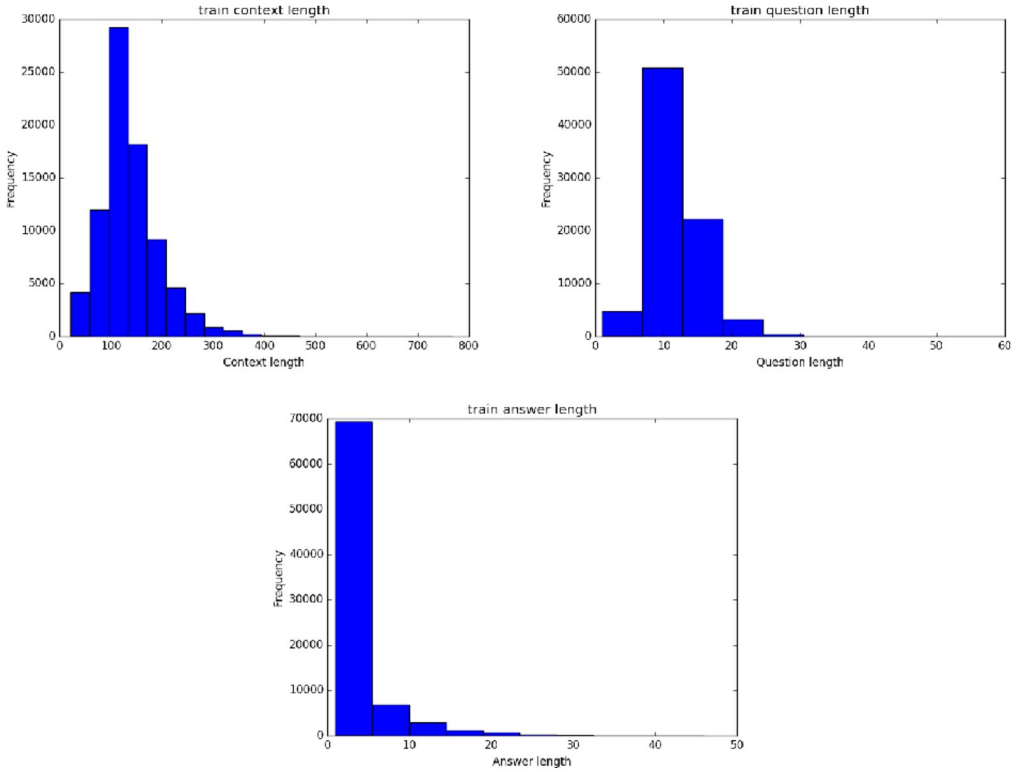


Figure 2: SQuAD training data context, question, and answer lengths. Source: Budianto, FNU. ”Reading Comprehension on the SQuAD Dataset.” 2017.

hidden states, producing context hidden states and question hidden states. These are then used for the bi-directional attention layer.

3. **Bi-Directional Attention layer**

   The attention layer receives the context and question hidden states ($c_1, ..., c_N$ and $q_1, ..., q_M$, respectively), and first computes a similarity matrix S, where each element $S_{i,j}$ represents the similarity score for the pair (c_i, q_j).

   Then, we calculate the Context-to-Question Attention by taking the softmax of each row of S and obtain attention distributions α^i, which we then use to take weighted sums of the question hidden states, giving us the context-to-question attention outputs a_i.

   Next, we calculate Question-to-Context Attention. For each context location i, we find the max of the corresponding row of S, m_i. After this, we find the softmax over m, which gives us an attention distribution β over context locations, which we then use to take a weighted sum of the context hidden states, giving us c', the question-to-context attention output.

   Finally, we combine each c_i and a_i for each index i of the context, with c', giving us the output B of the bi-directional attention layer. We then use B to proceed with the Modeling layer.

4. **Modeling layer**

   The modeling layer uses the output of the attention layer B to be passed through a bi-directional LSTM. From this, we get a matrix M which is then passed into the output layer. The idea is that M is able to capture the interaction among the context words based on the question, so that it can contain information about each word with respect to the whole context and question.

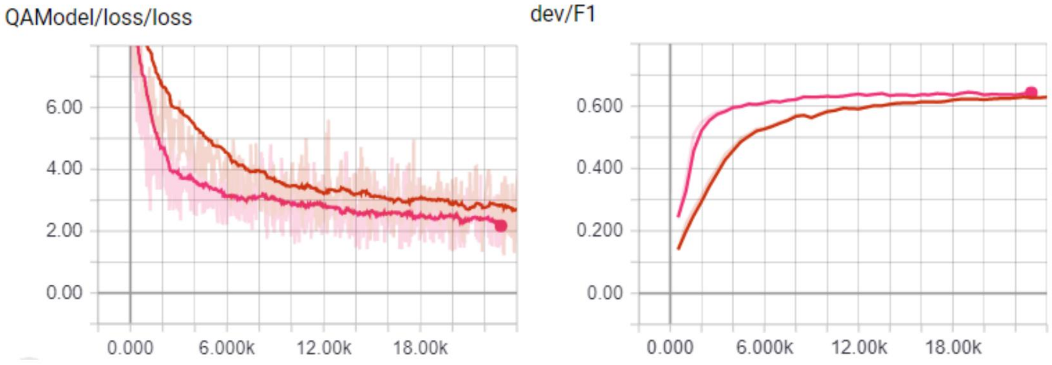


Figure 3: Differences between model performance using different hyper-parameters and optimizers.

Table 1: Score Comparisons of Models

Model	F1 score	EM score
Humans	91.2	82.3
BiDAF (single)	77.3	67.7
My model	69.8	59.6
Baseline	44.2	34.8

5. **Output layer**
   Finally, the output layer is used to make sense of the modeling layer's output. It passes the input through a ReLU nonlinearity, which we then use to assign a start score and an end score to each one of the context locations. Lastly, we use softmax to transform the scores into probability distributions. We simply take the index with highest start probability and the index with highest end probability, and return these two indices as start and end of the answer.

Figure 1 provides a visual representation of the workflow of this model.

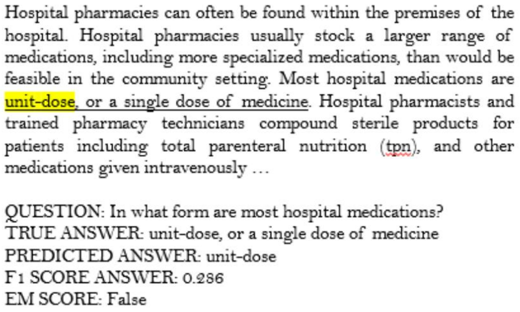
4 Experiments

4.1 Dataset

The dataset I used was the Stanford Question Answering Dataset (SQuAD). It is a reading comprehension dataset, that consists of 100,000+ question-answer pairs with corresponding context texts, from over 500 Wikipedia articles. Every question in the dataset is answerable by just the context paragraph alone. For the purposes of our experimentation, the data is divided into a training set, a dev set, and a test set. Interesting and useful analysis has been done on the SQuAD dataset, which allows us to better focus our experimentation. Figure 2 shows us that the vast majority of context lengths are under 300 words in size, and the vast majority of questions are under 20 words in size. We use these numbers as limits to our maximum context and question lengths, respectively, to ensure faster training without much loss in accuracy.

4.2 Model Details

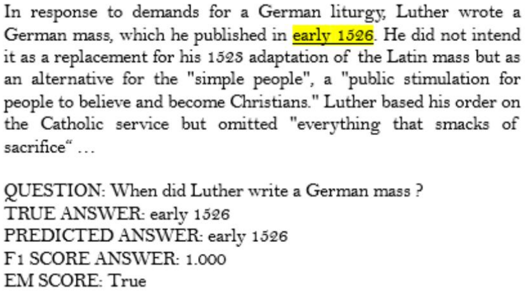
I experimented with many different hyper-parameters, and different optimizers until I was able to have good results, that would not take too long to train. For my final model, I used a context length of 300, a question length of 20, a batch size of 30, a hidden state size of 200, and a 0.15 probability of dropout. I used the Adam optimizer with a learning rate of 0.001. I trained my model on a standard NV6 Azure virtual machine, for about 18k iterations to achieve my final working results. I



Hospital pharmacies can often be found within the premises of the hospital. Hospital pharmacies usually stock a larger range of medications, including more specialized medications, than would be feasible in the community setting. Most hospital medications are unit-dose, or a single dose of medicine. Hospital pharmacists and trained pharmacy technicians compound sterile products for patients including total parenteral nutrition (tpn), and other medications given intravenously …

QUESTION: In what form are most hospital medications?
TRUE ANSWER: unit-dose, or a single dose of medicine
PREDICTED ANSWER: unit-dose
F1 SCORE ANSWER: 0.286
EM SCORE: False

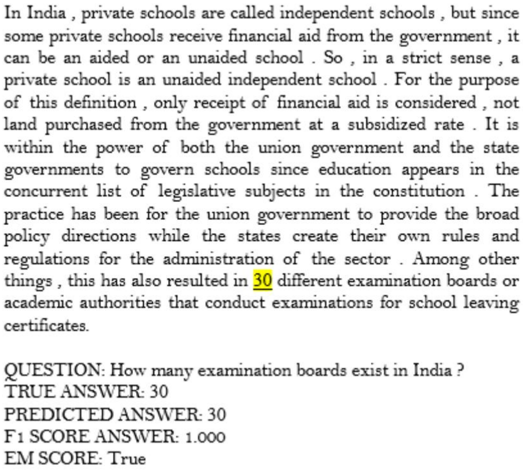
Figure 4



In response to demands for a German liturgy, Luther wrote a German mass, which he published in early 1526. He did not intend it as a replacement for his 1523 adaptation of the Latin mass but as an alternative for the "simple people", a "public stimulation for people to believe and become Christians." Luther based his order on the Catholic service but omitted "everything that smacks of sacrifice" …

QUESTION: When did Luther write a German mass ?
TRUE ANSWER: early 1526
PREDICTED ANSWER: early 1526
F1 SCORE ANSWER: 1.000
EM SCORE: True

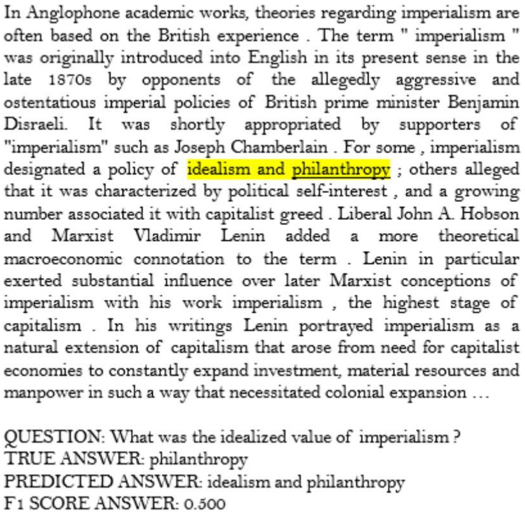
Figure 5



In India , private schools are called independent schools , but since some private schools receive financial aid from the government , it can be an aided or an unaided school . So , in a strict sense , a private school is an unaided independent school . For the purpose of this definition , only receipt of financial aid is considered , not land purchased from the government at a subsidized rate . It is within the power of both the union government and the state governments to govern schools since education appears in the concurrent list of legislative subjects in the constitution . The practice has been for the union government to provide the broad policy directions while the states create their own rules and regulations for the administration of the sector . Among other things , this has also resulted in 30 different examination boards or academic authorities that conduct examinations for school leaving certificates.

QUESTION: How many examination boards exist in India ?
TRUE ANSWER: 30
PREDICTED ANSWER: 30
F1 SCORE ANSWER: 1.000
EM SCORE: True

Figure 6



In Anglophone academic works, theories regarding imperialism are often based on the British experience . The term " imperialism " was originally introduced into English in its present sense in the late 1870s by opponents of the allegedly aggressive and ostentatious imperial policies of British prime minister Benjamin Disraeli. It was shortly appropriated by supporters of "imperialism" such as Joseph Chamberlain . For some , imperialism designated a policy of idealism and philanthropy ; others alleged that it was characterized by political self-interest , and a growing number associated it with capitalist greed . Liberal John A. Hobson and Marxist Vladimir Lenin added a more theoretical macroeconomic connotation to the term . Lenin in particular exerted substantial influence over later Marxist conceptions of imperialism with his work imperialism , the highest stage of capitalism . In his writings Lenin portrayed imperialism as a natural extension of capitalism that arose from need for capitalist economies to constantly expand investment, material resources and manpower in such a way that necessitated colonial expansion …

QUESTION: What was the idealized value of imperialism ?
TRUE ANSWER: philanthropy
PREDICTED ANSWER: idealism and philanthropy
F1 SCORE ANSWER: 0.500

Figure 7

also attempted using similar hyper-parameters as the original BiDAF paper, but the resulting model took many more iterations to approach comparable F1 scores to my final model, as shown in Figure 3.

4.3 Evaluation Metrics

The main evaluation metric used was F1, followed by Exact Match. The Exact Match measure determines whether for each example in the testing set, the model is able to predict exactly the starting and ending indeces for the answer, and then the total EM score is a percentage of correct answers. The F1 evaluation metric is more forgiving – it is calculated by the average of precision and recall, so that the model is not fully penalized for having partially correct answers.

4.4 Results

My final model was able to achieve a F1 score of 69.827 and an EM score of 59.614 on the test set, approximately a 25% improvement over the baseline model. It is a more basic model than the original BiDAF paper's model, so a lower score than that one achieved was expected. Table 1 shows the comparison in scores achieved on the test set between some different models.

Figures 4, 5, 6, and 7 show examples of contexts and questions, with the model's predicted answers. In general, the model seems to do well with questions that ask for a number or an exact time – questions that begin with ”how many” and ”when,” for example, as seen in Figures 5 and 6. This is probably because these sorts of questions are very direct, so our neural network is able to learn the types of responses wanted easily. However, questions that require a little more thought for humans, such as the ”what,” ”how,” and ”why” questions are also harder for our model to answer. An example is seen in Figure 7.

An interesting observation is that in some cases, our model chooses an answer that is arguably better than the ground truth answer, indicating that the F1 and EM scores are not necessarily great representations of how well the model actually works. For example, in Figure 4, the predicted answer ”unit-dose” makes more sense than the full ground truth answer ”unit-dose, or a single dose of medicine” since the part after the comma is only a description of what a ”unit-dose” means.

5 Conclusion

In this project, I implemented a simplified version of the Bi-directional Attention Flow Network. My final model uses an encoder layer, followed by a bi-directional attention flow layer, which allows the model to focus on and combine information from both the context and the question texts, a modeling layer to capture interactions between words in the context based on the question, finally followed by an output layer that determines the starting and ending word indices of the context where the answer is found. After just the implementation of the bidirectional attention layer, the model received a F1 score of 45%, a small improvement over the baseline, but after the modeling layer was implemented, then our model was really able to start doing well, with about a 25% increase in F1 score. This suggests that attention is not necessarily ”all you need,” but with a little help from other layers, it can be vital for a good question-answering neural network.

The most obvious future idea to improve the model is to integrate more parts of the original BiDAF network in my model, such as a Character-level CNN layer, as well as a modification in the Modeling layer to include a sequence of more than one LSTMs. Additionally, having more than one well-working model, would allow us to ensemble them to have even better results than their individual scores, so implementing another completely different model like the R-NET using self-matching attention and pointer networks would be extremely useful.

Acknowledgments

I would like to thank the CS 224N instructor, Richard Socher, for teaching this great class, as well as all the TAs for being so attentive and helpful online and during office hours.

References

[1] CS 224N Default Final Project: Question Answering. Stanford University, Winter 2018.

[2] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. CoRR, abs/1606.05250, 2016.

[3] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. arXiv preprint arXiv:1611.01603, 2016.

[4] Budianto, FNU. Reading Comprehension on the SQuAD Dataset. Stanford University. CS 224N Winter 2017.

[5] Shuohang Wang and Jing Jiang. Machine comprehension using match-lstm and answer pointer. arXiv preprint arXiv:1608.07905, 2016.

[6] Yang Yu, Wei Zhang, Kazi Hasan, Mo Yu, Bing Xiang, and Bowen Zhou. End-to-end answer chunk extraction and ranking for reading comprehension. arXiv preprint arXiv:1610.09996, 2016.

[7] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading wikipedia to answer open-domain questions. arXiv preprint arXiv:1704.00051, 2017.

[8] Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William W Cohen, and Ruslan Salakhutdinov. Gated-attention readers for text comprehension. arXiv preprint arXiv:1606.01549, 2016.

[9] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. arXiv preprint arXiv:1707.07328, 2017.