Reading Comprehension using Recurrent Neural Network

Sambartika Guha
MS, Computer Science
Texas A&M University
Email: sambartika.guha@tamu.edu

Abhilash Vallamkonda
MS, Computer Science
Texas A&M University
Email: vrabhilash@tamu.edu

Abstract

Reading comprehension is one of the most challenging tasks of Natural Language Processing and the goal is to move one step closer to enabling machines to understand the world. The problem of reading comprehension has profound implications for Information Retrieval applications, for example : A query like "broken chair" should be able to return listings for shops selling Furniture which is possible only if the system understands that the solution to a broken chair is usually to buy a new one. One way to check if a system actually understands something is to see if the model can reason about the available information and answer questions about it. This is exactly what the Reading Comprehension task entails and a system which does well in this task can be said to be able to understand and reason about the provided information. Recurrent Neural Networks(RNNs) work well on NLP tasks. So, for the problem of reading comprehension where extracting the relevant information is critical, using RNNs with attention seems like a natural choice. In this project, we implement LSTMs with coattention for Reading Comprehension.

1 Introduction

During the 90s and the early 2000s, Internet was like a wonderland to people. For the first time ever, humanity had access to an enormous amount of information. Since then, the Internet has helped us in learning new things, allowed people from around the world to work together and has completely revolutionized the way we access information. Due to the ease of creating websites, every person with access to the Internet is both a creator and consumer of information. This has lead to the "Information Age" in which we live right now. But for the past few years, the Internet has grown to the extent that it is impossible for a human being to efficiently use the enormous amount of information available. So, there is a need to build a system which can understand the information and give us exactly what we need. Reading comprehension tries to tackle this problem. Given a piece of information and a question, the model needs to find the right answer. On a larger scale, such a system can understand the queries and search for the answer over all the information available on the Internet.

The key challenge in Reading Comprehension is that the model has to actually understand the question and the context and then figure out how the information in the context can be used to answer the given question. This directly translates into the fundamental problem of understanding human language. So, in order to answer the question, the model needs to solve the classical problems in NLP like finding all phrases which refer to an entity in a given text (co-reference resolution) and understanding cause-effect relations.

The main idea which allows our model to do well in Reading Comprehension is that of attention[1]. Attention allows the model to focus only on the relevant information while ignoring the rest. GRUs and LSTMs work better than traditional RNNs but still cannot remember the entire context while answering a question. Attention alleviates this problem by allowing the model to focus only on the relevant parts of the question and context.

Our baseline model uses bidirectional GRU with attention mechanism for reading comprehension task. This model achieved a F1 score of 40%. We tried to tune the parameters and only got 42% F1 score after tuning. Our final model is based on co-attention network with tuned parameters and achieves a F1 score of 67.37%.

2 Dataset

One of the major challenges to solving the problem of Reading Comprehension was the lack of a proper dataset. Earlier datasets were human annotated and high quality but too small to train expressive models. The artificially generated datasets were larger but unlike their human annotated counterparts, they lacked questions which required certain types of reasoning[15]. The SQuAD(Stanford Question and Answer Dataset)[2] released in 2016 is both large and also involves a diverse set of reasoning approaches in order to answer the questions.

We are using SQuAD published in 2016 to train and evaluate our model on the reading comprehension task. It contains more than 100,000 question answer pairs on more

than 500 passages. All the answers in SQuAD are contained in the context which allows us to train our model to answer questions by predicting the start and end points in the context.

For better understanding the dataset, we performed some basic analysis on the dataset. We found that most of the context lengths are less than 400 words and question lengths are less than 30 words. Also, among the 87600 answers in the dataset, only 11(0.01%) have ending beyond the 380th word and 45(0.05%) have ending beyond the 300th word. Since the largest context length is 653, this suggests that the first part of the context is more likely to contain the answer and the entire context may not be needed.

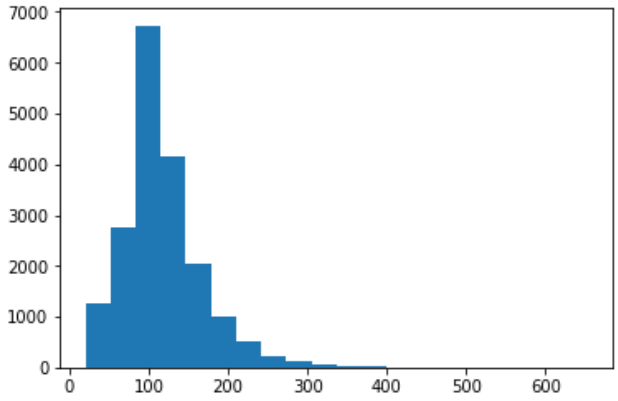

Fig. 1. Context length distribution

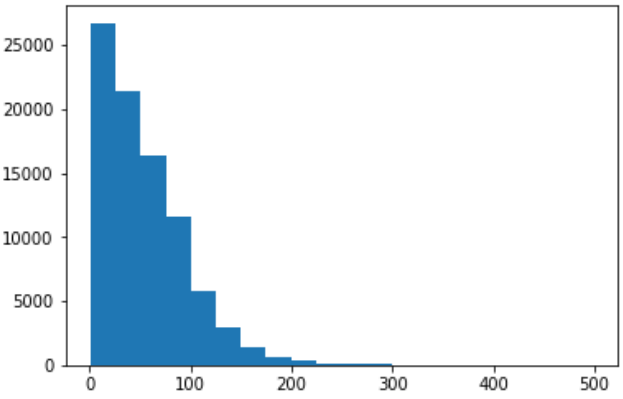

Fig. 2. Distribution of answers in the context

Example:
Architecturally, the school has a Catholic character. Atop the Main Building's gold dome is a golden statue of the Virgin Mary. Immediately in front of the Main Building and facing it, is a copper statue of Christ with arms upraised with the legend "Venite Ad Me Omnes". Next to the Main Building is the Basilica of the Sacred Heart. Immediately behind the basilica is the Grotto, a Marian place of prayer and reflection. It is a replica of the grotto at Lourdes, France where the Virgin Mary reputedly appeared to Saint Bernadette Soubirous in 1858. At the end of the main drive (and in a direct line that connects through 3 statues and the Gold Dome), is a simple, modern stone statue of Mary.
Question: To whom did the Virgin Mary allegedly appear in 1858 in Lourdes France?
Answer: Saint Bernadette Soubirous

3 Prior Work

The basic approach for solving reading comprehension is to use RNN Encoder-Decoder model which encodes the input into a fixed length vector and decodes it to generate the output which is given in [8]. In [3], Bahdanau et al.,2014 suggest that the problem with these architectures is that the network is expected to encode all the input information into a single fixed length vector which makes it difficult to cope with long inputs. They propose an attention mechanism which allows the network to search the input for parts relevant to the current output. In the past few years, LSTMs[11] with attention have shown remarkable success in problems like Machine Translation[3], Image captioning[9] and speech recognition[10].

Machine Comprehension: Since the release of SQuAD in 2016, there has been a lot of work done using the dataset. Ideas like Encoder-Decoder models and attention have been successfully applied to the problem of Reading Comprehension and many problem-specific modifications have been proposed. [4] proposes the Bi-Directional Attention Flow(BiDAF) network which uses a multi-stage hierarchical process to represent the context at different levels of granularity. This model uses both question-to-context and context-to-question attention for extracting the relevant information from both question and context. [6] places another attention mechanism over the document level attention creates an attention over attention to decide the relevant inputs while producing the final predictions.

Another popular approach was Match-LSTM[12] which is an end to end model including preprocessing data using LSTM, creating an attention vector with a bidirectional LSTM, finally feeding the attention vector into a Pointer Net that predicts the start and end locations in the context of the answer.

Our project makes use of co-attention[13] which like BiDAF, involves two-way attention between the question and the context. Co-attention takes things a step further and involves computing a second-level attention where attention is computed over representations which themselves are attention outputs.

4 Data Preprocessing

The first step of the reading comprehension task was data preprocessing. We have used NLTK toolkit to split the

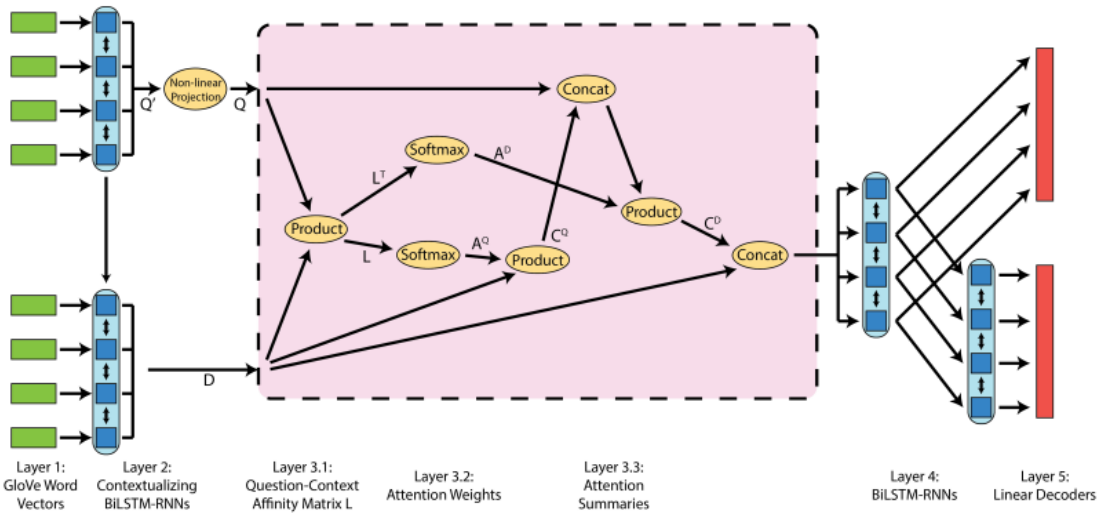

Fig. 3. Coatttention Model Architecture

sentences into tokens. We represent the words using pre-trained GLoVE embeddings. This results in a d-dimensional vector representation for every word.

5 Model Architecture

All the models use the same architecture for encoding the question and context. For each example, the context and question are represented by a sequence of d-dim Glove word embeddings, so context is $x_1, x_2, ..., x_N$ and the question is $y_1, y_2, ..., y_M$. These are then input to a bidirectional GRU or LSTM to generate hidden states, $c_1, c_2, ...c_N$ and $q_1, q_2, ...q_M$ corresponding to the context and the question. These hidden states are used as inputs to the attention layers.

$$\{c_1, c_2, ...c_N\} = biGRU(\{x_1, x_2, ..., x_N\})$$
$$\{q_1, q_2, ...q_M\} = biGRU(\{y_1, y_2, ..., y_M\})$$

The goal of the attention layer is to create a representation of the context along with the parts of the question relevant to the context called b_i. The various b_i vectors are then input to a softmax layer to decide whether or not the current context is the starting or ending point of the answer.

$$logits_i = w^T b_i$$
$$pred_{start} = softmax(logits)$$

The various models differ in the method in which these b_i representations are generated.

5.1 Basic Attention Model(Baseline)

In this model, basic dot product attention is applied with every c_i attending to the question states q_j (C2Q). The scores are then fed into a softmax layer to get the C2Q attention weights.

$$scores_i = [c_i^T q_1, c_i^T q_2, ..., c_i^T q_M]$$
$$\alpha^i = softmax(scores_i)$$

The various q_j are then weighted using these weights to obtain a weighted attention output corresponding to c_i.

$$a_i = \sum_{j=1}^{M} \alpha_j^i q_j$$

The context hidden state is concatenated with the attention output which represents the parts of the question relevant to the context under consideration.

$$b_i = [c_i; a_i]$$

This is then input to the softmax layer which predicts the starting and ending points.

5.2 Coattention

We compute C2Q attention as in the basic attention model but also compute Q2C attention with every q_j attending to the context states.

$$scores_j = [c_1^T q_j, c_2^T q_j, ..., c_N^T q_j]$$
$$\beta^j = softmax(scores_j)$$
$$b_j = \sum_{i=1}^{N} \beta_i^j c_i$$

From C2Q, we have the relevant parts of the question for every c_i and from Q2C we have the parts of the context relevant

to every q_j. By taking a product, we have a question aware context representation. For every c_i, we consider the relevant parts of the question and also the parts of the context relevant to them, context represented using a second level of attention, s_i.

$$s_i = \sum_{j=1}^{M} \alpha_j^i b_j$$

This representation of the context is combined with the parts of the question which are relevant to it and then fed into a softmax layer. The model is described graphically in Fig. 3.

6 Dropout in RNNs

Neural networks are powerful models but are prone to over fitting, hence good regularization is required while training them. Dropout[19] is the most effective regularization technique for feedforward networks. However, naive application of dropout to RNNs does not produce good results. Zaremba et al. show that applying dropout only over the non-recurrent connections can produce significant improvements in performance [14]. It is theorized that the recurrent connection amplifies the noise generated as a result of applying dropout which disturbs the complex dynamics that allow RNNs to memorize information over multiple time steps. This degrades the performance. As suggested in [14], we apply dropout on the non-recurrent connections(dashed lines in Fig. 4) and achieve a significant performance improvement in the coattention model.

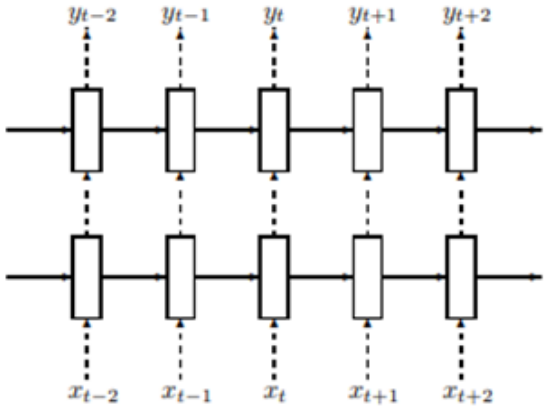

Fig. 4. Dropout

7 Evaluation

Two metrics were mainly used for evaluation. The first one is F1 score. F1 score is a traditional evaluation metric in information retrieval and computes the harmonic mean of precision and recall which measures how well the model is able to return the relevant information while ignoring the irrelevant stuff. Intuitively, F1 score indicates how well the returned answer matches the expected result.
The stricter evaluation metric is exact match(EM). EM score is binary. It gives one if the predicted answer is exactly same as the actual answer and gives 0 otherwise.

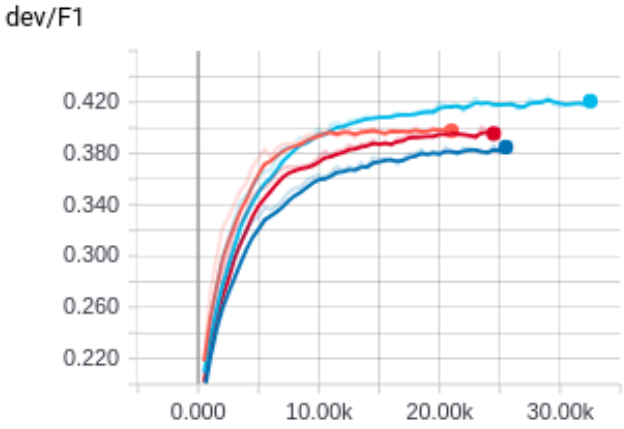

Fig. 5. Baseline model performance

8 Experimental Setup

We adapted the code from the cs224n course offered at Stanford University[17] for our baseline model[16]. Natural Language Processing Toolkit(NLTK) is used to split the context into words. Then, we obtain the embedding vectors for each of the words using Glove embedding[18]. We used LSTMs instead of GRUs and also tuned the hyperparameters which allowed us to improve the F1 score over the baseline. We experimented with 300 dimensional and 100 dimensional Glove vectors and found that the model using 300 dimensional Glove works better. So, in our final model we have used 300 dimensional Glove vectors. The embedding vectors are passed into the RNN encoder where we experimented with the various hidden layer sizes. A hidden state size of 200 worked best and this is the value we chose for the final model.

Most of the context lengths are less than 450 (from Fig. 1). So, we have truncated the context length to 450 for those contexts whose length is greater than 450. Also, we have set the maximum question length to 30.

We have used mini-batch gradient descent with Adam optimization and batch size of 128. The training starts with an initial learning rate of 0.001 and decay rate 0.88 for every 1000 iterations. We also tried with 0.1 and 0.01 learning rate values and found that a learning rate of 0.001 works better.

As we reached the limits of the performance possible with fine-tuning the Baseline model, we implemented the more powerful coattention model as described in Xiong et al. [5]. We started with the same parameters as in the baseline

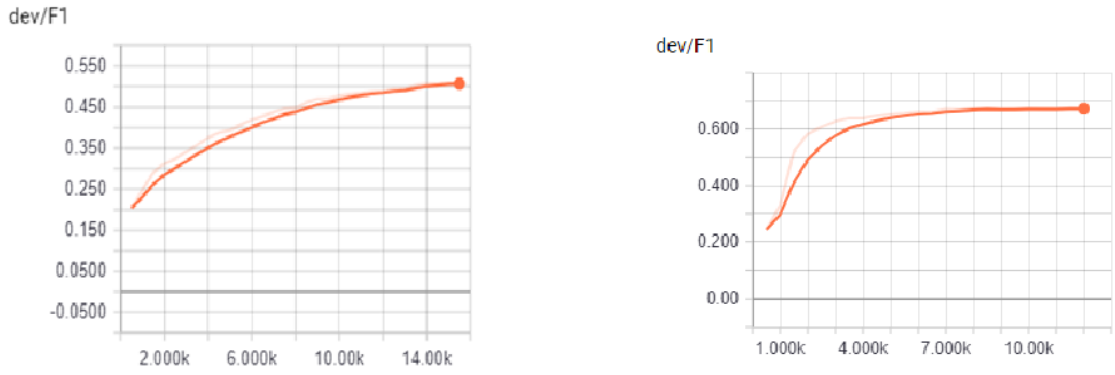

Fig. 6. Coattention without and with tuned parameters

model but found that the was overfitting the dataset. Changing the hidden state size and embedding size did not resolve the issue. Finally we introduced dropout in our model and this alleviated the problem. A fixed dropout of 0.5 was used for all of the LSTM cells. We also tried with dropout rates 0.2 and 0.4 but neither worked as well as 0.5.

9 Results

The performance of baseline model with attention layer on SQuAD dataset is shown in Fig 5. The F1 score on the training set keeps increasing whereas the F1 score on the dev set converges in 15000 iterations. This clearly indicates that the model is overfitting. The F1 score of the model was around 40%.

We tried to improve the performance of the baseline model by tuning of hyperparameters. Tuning the learning rate, embedding vector size, dropout and using LSTMs instead of GRUs only resulted in a modest improvement in F1 score from 40% to 42% with the problem of overfitting still persisting.

The coattention model[5] was more effective and using the same hyperparameters as the baseline we were able to achieve an F1 score of 51%.

Unlike the baseline model, finetuning of hyperparameters significantly improved the performance of coattention model. Using 300 dimensional word vectors instead of 100 dimensional word vectors improved the performance over 5%. The maximum context length in the training set was around 650, however using a smaller context length of 450, the performance of the model and also the time taken per iteration improved significantly. The performance of coattention model is shown in Fig 6.

Another important factor was answer length. Most of the answers were within length 15. When we restricted the answer length to 15, performance of our model improved. Also, from Fig 7 we can see that our model performed well with short answers and it's performance degraded with long answers. Until length 6, the performance was around 60%. But, as the length increased, the performance degraded. That means our model finds it difficult to find long answers and it can easily detect short answers.

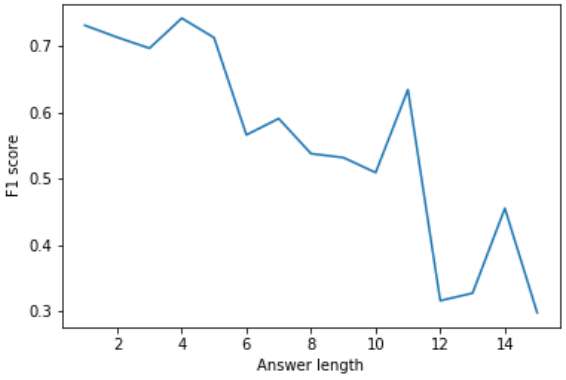

Fig. 7. Performance with answer lengths

Model	F1	EM
Basic attention	40%	29%
Attention + tuned hyperparameters	42%	31%
Co-attention	51%	37.25%
Co-attention + tuned hyperparameters	67.37%	41.21%

Table 1. Table of performances od various models

On the other hand, the performance of the model was almost same for longer contexts. This shows that our attention model was able to identify the important parts of the context. The F1 score of the model was 67.37% which is 27.37% better than the baseline.

The table 1 shows how performance varies from the baseline model to our final co-attention model.

We further analyzed the results of our model on various types of questions like 'why', 'when', 'how', 'where' questions. We found that our model performed well on 'who', 'when', 'which' types of questions whose answers are usually short and specific. But for 'how' and 'what' questions, the F1 score is good but EM is poor. This means that our model is able to understand where the answer is but it is doing a poor job identifying the exact answer.Fig 8 shows how our model performs with various types of questions.

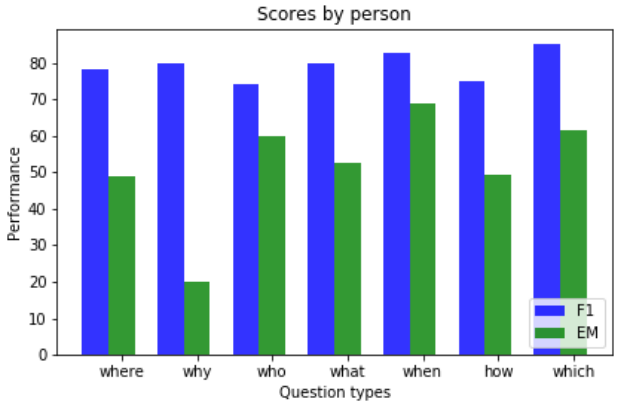

Fig. 8. Performance in various types of answers

10 Error Analysis

We examined the error cases to check the common types of error. One very common error was incorrect answer boundaries. Sometimes, our model identifies the right phrase as the answer but includes one or two additional words. The additional words identified result in a 0 for the EM score. One such example was:

Context:ipcc author richard lindzen has made a number of criticisms of the tar . among his criticisms , lindzen has stated that the wgi summary for policymakers (spm) does not faithfully summarize the full wgi report . for example , lindzen states that the spm understates the uncertainty associated with climate models . john houghton , who was a co-chair of tar wgi , has responded to lindzen 's criticisms of the spm . houghton has stressed that the spm is agreed upon by delegates from many of the world 's governments , and that any changes to the spm must be supported by scientific evidence .

Question: what did houghton say is necessary for any changes to the spm ?

True answer: scientific evidence

Predicted answer: must be supported by scientific evidence

Another common error was, our model gave shorter answer which affected the EM score. For example,

Context:once mutual's appeals against the fcc were rejected , rca decided to sell nbc blue in 1941 , and gave the mandate to do so to mark woods . rca converted the nbc blue network into an independent subsidiary , formally divorcing the operations of nbc red and nbc blue on january 8, 1942, with the blue network being referred to on-air as either "blue" or "blue network" . the newly separated nbc red and nbc blue divided their respective corporate assets . between 1942 and 1943 , woods offered to sell the entire nbc blue network , a package that included leases on landlines , three pending television licenses (wjz-tv in new york city , kgo-tv in san francisco and wenr-tv in chicago) , 60 affiliates , four operations facilities (in new york city , chicago , los angeles and washington d.c.) , contracts with actors , and the brand associated with the blue network . investment firm dillon , read & co. (which was later acquired by the swiss bank corporation in 1997) offered $ 7.5 million to purchase the network , but the offer was rejected by woods and rca president david sarnoff .

Question:what network was converted into an independent subsidiary by rca in 1942 ?

True answer: nbc blue network

Predicted answer: nbc blue

Sometime during incorrect placement of attention, the model gives completely incorrect result. It gives zero F1 and EM score. For example,

Context: in april 1970 , congress passed the public health cigarette smoking act which banned cigarette advertising from all television and radio networks , including abc , when it took effect on january 2 , 1971 . citing limited profitability of its cinemas , abc great states , the central west division of abc theatres , was sold to henry plitt in 1974 . on january 17 , 1972 , elton rule was named president and chief operating officer of abc a few months after goldenson reduced his role in the company after suffering a heart attack .

Question: when did the ban on cigarette advertising take effect for television networks ?

True answer: january 2 , 1971

Predicted answer: april 1970

11 Future Work

Reading Comprehension task is a very challenging task of Natural Language Processing. We have only scratched the surface of the problem. Our model will only be able to answer if the answer completely lies within the context. The future work can be building models which can produce answers even if they do not fully lie within the given context. As for the model, for now we have used simple softmax layer as a decoder to identify the start and end points of the answer. It is likely that more sophisticated decoders like answer pointer decoder will result in better performance. Also, there are many other architectures like match LSTM, Bidirectional Attention Flow and using an ensemble of such diverse models will almost certainly improve the performance over the

reading comprehension task.

12 Work Split

.

Background research on RNNs, prior work
(Abhilash - 50%, Sambartika - 50%)
Training and fine-tuning the basic attention model
(Abhilash - 100%)
Implementation of Co-attention model
(Sambartika - 100%)
Fine-tuning the co-attention model
(Abhilash - 40%, Sambartika - 60%)
Presentation Slides
(Abhilash - 60%, Sambartika - 40%)
Report
(Abhilash - 50%, Sambartika - 50%)

13 References

.

[1] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, Bidirectional Attention Flow for Machine Comprehension, ArXiv e-prints, Nov. 2016.
[2] https://rajpurkar.github.io/SQuAD-explorer/
[3] D. Bahdanau, K. Cho, and Y. Bengio, Neural Machine Translation by Jointly Learning to Align and Translate, ArXiv e-prints, Sep. 2014.
[4] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, Bidirectional Attention Flow for Machine Comprehension, ArXiv e-prints, Nov. 2016.
[5] C. Xiong, V. Zhong, and R. Socher, Dynamic Coattention Networks For Question Answering, ArXiv e-prints, Nov. 2016.
[6] Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, Guoping Hu Attention-over-Attention Neural Networks for Reading Comprehension
[7] https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2762006.pdf
[8] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, Yoshua Bengio Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation
[9] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, Yoshua Bengio Show, Attend and Tell: Neural Image Caption Generation with Visual Attention
[10] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel and Yoshua Bengio End-To-End Attention-based Large Vocabulary Speech Recognition
[11] https://colah.github.io/posts/2015-08-Understanding-LSTMs/
[12] Wang and Jiang, "Machine Comprehension Using Match-LSTM and Answer Pointer.", https://arxiv.org/pdf/1608.07905.pdf.
[13] C. Xiong, V. Zhong, and R. Socher, "Dynamic coattention networks for question answering," https://arxiv.org/pdf/1611.01604.pdf.
[14] Wojciech Zaremba, Ilya Sutskever, Oriol Vinyals, "Recurrent Neural Network Regularization", https://arxiv.org/pdf/1409.2329.pdf.
[15] Danqi Chen, Jason Bolton, and Christopher D. Manning. A thorough examination of the cnn/daily mail reading comprehension task. In Association for Computational Linguistics (ACL), 2016.
[16] http://web.stanford.edu/class/cs224n/default_project/default_project_v2.pdf
[17] http://web.stanford.edu/class/cs224n/
[18] Jeffrey Pennington, Richard Socher, Christopher D. Manning. "GloVe: Global Vectors for Word Representation", https://nlp.stanford.edu/pubs/glove.pdf
[19] Srivastava, Nitish. Improving neural networks with dropout. PhD thesis, University of Toronto, 2013.