Natural language Question Answering using Curriculum Learning

Aarti Bagul
Department of Computer Science
Stanford University
aartib@stanford.edu

Abhijeet Shenoi
Department of Computer Science
Stanford University
ashenoi@stanford.edu

Abstract

Question answering is a sub-problem in the larger class of natural language understanding tasks. To effectively answer a question, given a context, a machine must effectively learn to understand the meaning of the context, and must guide its search for an answer based on the question. In this work, we explored various methods to encode the question and context information, different types of attention, different ways to select the answer span and a reinforcement learning approach to directly optmize for the F1/EM score. The curriculum learning approach we used resulted in marginally better performance, as compared to supervised learning, for half the amount of 'supervised' training time. It can further be readily extended to unsupervised tasks. Our single best model had an F1 score of 72.98%. Ensembling our best models resulted in an F1 of 76.8%.

1 Introduction

Reading comprehension is a challenging task in natural language processing that requires the understanding of natural language, and the application of that understanding towards achieving a certain goal. One such goal is to identify the answer to a given question. The SQuAD dataset was introduced by Rajpurkar et. al [1] as a first of its kind large scale high quality dataset to allow for further research in the domain.

The dataset contains over 100,000 question answer pairs on 536 articles. The articles were sourced from Wikipedia, and crowd workers were used to generate the questions and the answers. Each question comes with 3 'gold standard' answers. There are two metrics used to determine the quality of an answer. The F1 score, and the EM score. The F1 score uses the precision and recall of the words returned by the model with respect to the closest gold standard answer. The EM score is a binary measure, which is 1 for an exact match between the returned output and any of the gold standard answers, and 0 if the answer does not match any of the gold standard answers. Note that these metrics are both non-differentiable.

2 Related Work

Since the introduction of the dataset a little over two years ago, substantial progress has been made. The following works have influenced the work in this paper to a large degree.

2.1 BiDAF

This approach by Minjoon Seo et. al [2] uses a hierarchical multi-stage architecture for to model the context at different levels of granularity. BiDAF includes character-level, word-level, and contextual embeddings, and uses bi-directional attention flow to obtain a query-aware context representation.

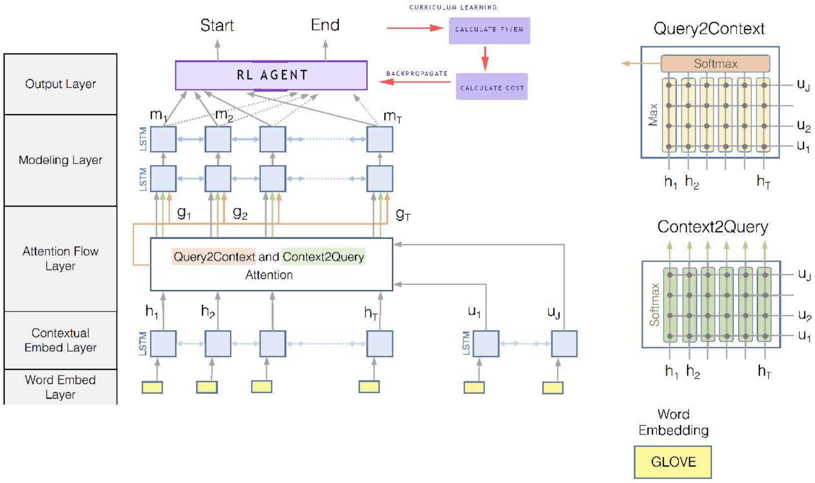


Figure 1: Adapted model from BiDAF

There is a division of labour between the attention layers and the modellings layers. Our adaptation of the network is seen in Figure 1. The attention flow layer is tasked with learning the specific instances within the query that are important, for every context word, whereas the the modelling layer is tasked with understanding the temporal effects these different representations have on each other. Finally, the output layer, which can be swapped out depending on the final task at hand, is used to predict the answer span. The entire model is trained using a softmax cross entropy loss, on the resulting logits across the context length. BiDAF achieved state of the art results at the time of its publication.

2.2 Dynamic Coattention Networks for Question Answering

Similar to BiDAF, the Dynamic Coattention Network [3] first encodes both question and context into embeddings. These embeddings are then used to determine which words in the question are important, and at the same time which words in the context are most similar to the question. The scheme is slightly different from that in BiDAF. Once this is done, a dynamic network iteratively selects the answer span. With $L_{ij} = c_i^T q_j$

$$\texttt{C2Q attention} \quad a_i = \texttt{sofmtax}(L, rows)Q \tag{1}$$

$$\texttt{Q2C attention} \quad s_i = \texttt{sofmtax}(L, rows)\texttt{softmax}(L, columns)C \tag{2}$$

These two attention representations are then used to construct the embedding used in further RNN layers:

$$b_i = [s_i; a_i] \tag{3}$$

The iterative nature of the model allows it to escape from local maxima initially, and then subsequently converge on the correct answer. This dynamic iterative network consists of an LSTm which used to model a state machine. The inputs to this state machine comes from a highway network which takes as inputs the previous state of the LSTM, and the embeddings of the previous estimate of the start and end positions.

2.3 Active Questions Reformulation with Reinforcement Learning [6]

This problem formulates an agent which modifies the question so as to elicit the best performance with a standard question answering module. The resulting augmented module is trained end to end using policy gradients. The model is shown in Figure 2. The approach improved the performance of BiDAF by an F1 of 11%

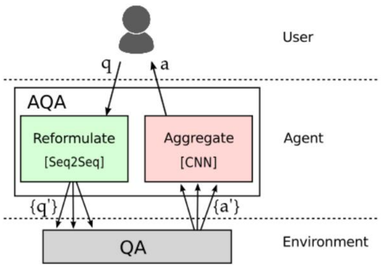


Figure 2: Active Question Reformulation Model

3 Approach

3.1 Neural Baseline Model

The Neural baseline model provided used the following architecture: Embedding Layer: The context and question words are embedded into a space by using pretrained, and constant, GloVE embeddings. These are then passed through bidirectional GRU's to generate context aware question and context embeddings. Attention Layer: The context then attends over the question. An attention representation is constructed and concatenated with the corresponding context embedding. Output Layer: This augmented context representation is then passed through a fully connected layer to obtain another representation. This is then put through two separate softmax layers to obtain a probability distribution over the context words. These distributions are used to predict the start and end positions of the answer, separately. Loss: The loss is a softmax cross entropy loss summed over the the start and end positions.

3.2 Adapted BiDAF

Our BiDAF variant did not use the character level encoding as in the original paper. Instead, we used the following approach: Embedding Layer: Similar to the baseline, the context and question words are embedded into a space by using pre-trained, and constant, GloVE embeddings. These are then passed through bidirectional GRU's to generate context aware question and context embeddings. Attention Layer: The context-question attention and the question-context attention are constructed by first constructing a similarity matrix. This matrix is given by:

$$S_{ij} = w^T [c_i; q_j; c_i \odot q_j]$$

This similarity matrix is then used to construct the attention scores as follows:

$$\begin{aligned} \texttt{C2Q attention} \quad a_i &= \texttt{sofmtax}(S, rows)Q \qquad (4)\\ \texttt{Q2C attention} \quad c' &= \texttt{softmax}(\max(S, rows))C \qquad (5)\\ & \qquad (6) \end{aligned}$$

These two attention representations are then used to augment the context embedding by:

$$b_i = [c_i; a_i; c_i \odot a_i; c_i \odot c'] \tag{7}$$

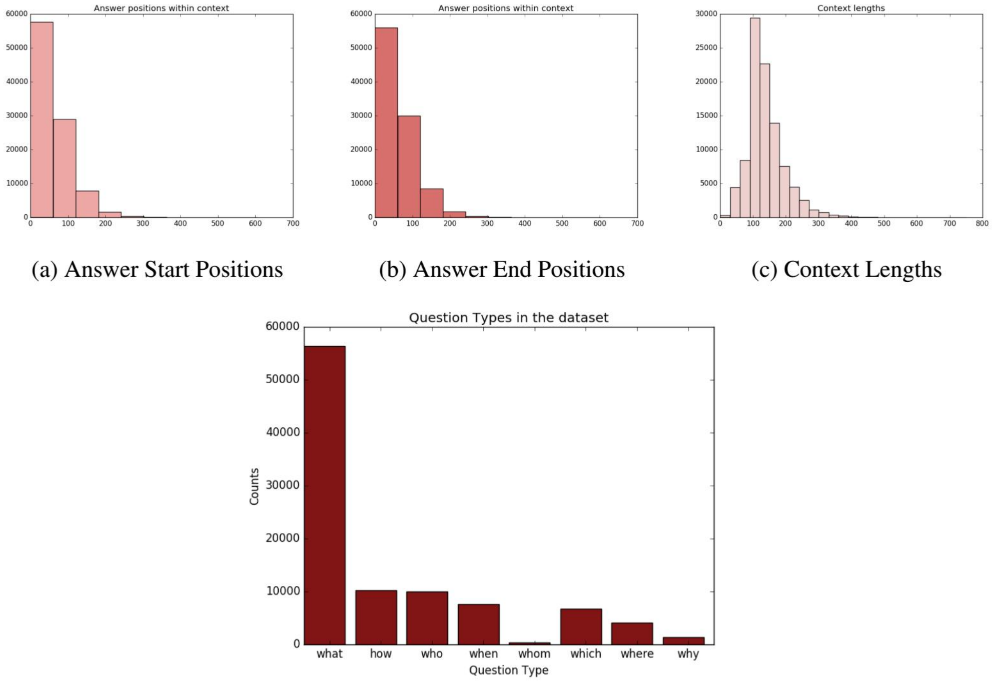
This is then passed through two layers of bidirectional GRU's to obtain the final representation. Output Layer: Two approaches were used. The first was a simple softmax classifier with independent estimation of the start and end positions. The best performing model however was obtained by using a combination of joint prediction and reinforcement learning.
Further, we attempted to model the selection of an answer span as an RL problem, wherein the model must learn a stochastic policy, where each action corresponds to selecting a specific word as the answer start/end position. The reward is then the F1/EM score. This allows us to directly optimize for a non-differentiable metric. We then predict the answer span which has the highest joint probability of selection according to the learned policy, given that the span must be such that the start occurs before the end. We also capped the maximum length of the answer to 30 words.

4 Experiments and Analysis

The following experiments were carried out:

4.1 Data Analysis

We first analyzed the dataset to formulate possible improvements in the pre-processing. We were able to ascertain from Figure 3 that contexts rarely exceed 300 words in length, and answers to most questions are within this 300 word boundary. Hence, we clipped the context lengths to 300, allowing us to nearly halve the memory requirements. Further, we classified questions according to types, as shown in Figure 3. As noticed in the examples, the models more often than not return the correct type of answer (person for ”who” question etc.). Therefore, there is unlikely to be significant difference in performance for a particular type of question. We therefore did not look into performance metrics on a per question-type basis.



(a) Answer Start Positions

(b) Answer End Positions

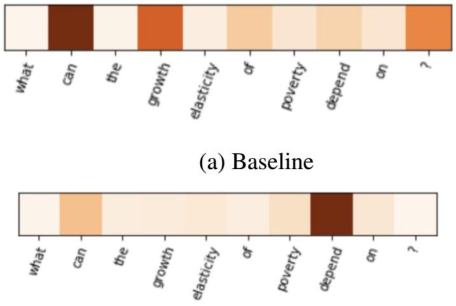
(c) Context Lengths

(d) Question Types

Figure 3: Dataset Statistics

4.2 Attention Visualization

To further expand on the baseline, we saw it fit to analyze the attention mechanism. To do this, we plotted the attention scores along with the question tokens, averaged across the entire context. This gave us an idea of which words in the question were being attended to the most. One particular feature we noticed was the high attention paid to the ’?’ token in the question. As all questions will contain this token, it is not useful to pay attention to it, as it provides no new information. The simple attention scheme is not capable of identifying subtleties within the question. This motivated us to work on improving the attention mechanism. A major difference we saw after the implementation of BiDAF was that the ’?’ token was no longer paid as much attention to, one of the reasons the BiDAF model vastly outperformed the baseline model. Examples can be seen in Figure 4.



(a) Baseline

(b) BiDAF Attention

Figure 4: Attention Visualization over the question ”What can the growth elasticity of poverty depend on?”. Notice how the baseline tends to focus on non salient tokens such as the ’?’ and the word ’can’

Further, the BiDAF has a question to context attention mechanism. We can see in Figure 4b that the context words which are most relevant to the question have high attention scores.

4.3 Error Metrics

On analysis of the errors made by the baseline model, we noticed that a large majority of errors made involved the prediction of the end position before the start position. We thus saw fit to use a new error metric to analyse the the quality of the model: the number of examples wherein the end position was predicted to be before the start position. Further, as a final improvement, we generated the most probable answer spans satisfying the additional constraint that the end position must be after the start position. This correspondence can be seen in Figure 5

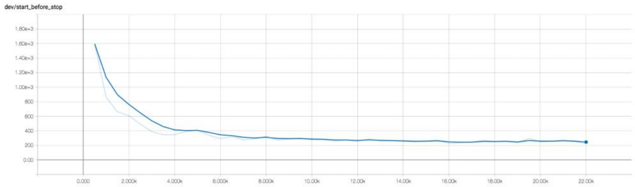
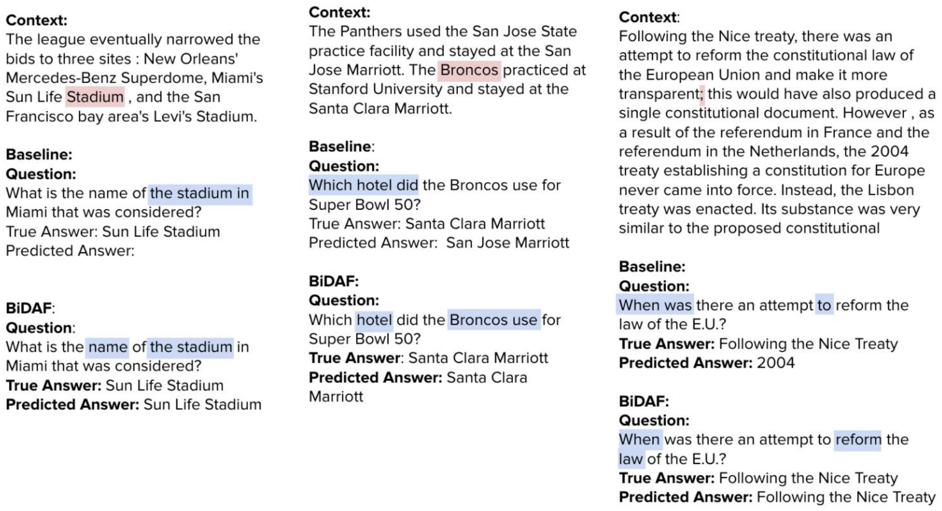


Figure 5: The start-stop positions correlate with the loss plots

Further, the answers almost always belonged to the correct category. ”Who” questions were answered with people. ”Where” questions were responded to with places. These can be seen in Figure 6. It can be noticed that BiDAF tends to attend to questions better, paying attention to more salient tokens. The baseline also exhibits the property of answering with the correct category.



Context:
The league eventually narrowed the bids to three sites : New Orleans' Mercedes-Benz Superdome, Miami's Sun Life Stadium , and the San Francisco bay area's Levi's Stadium.

Baseline:
Question:
What is the name of the stadium in Miami that was considered?
True Answer: Sun Life Stadium
Predicted Answer:

BiDAF:
Question:
What is the name of the stadium in Miami that was considered?
True Answer: Sun Life Stadium
Predicted Answer: Sun Life Stadium

Context:
The Panthers used the San Jose State practice facility and stayed at the San Jose Marriott. The Broncos practiced at Stanford University and stayed at the Santa Clara Marriott.

Baseline:
Question:
Which hotel did the Broncos use for Super Bowl 50?
True Answer: Santa Clara Marriott
Predicted Answer: San Jose Marriott

BiDAF:
Question:
Which hotel did the Broncos use for Super Bowl 50?
True Answer: Santa Clara Marriott
Predicted Answer: Santa Clara Marriott

Context:
Following the Nice treaty, there was an attempt to reform the constitutional law of the European Union and make it more transparent; this would have also produced a single constitutional document. However , as a result of the referendum in France and the referendum in the Netherlands, the 2004 treaty establishing a constitution for Europe never came into force. Instead, the Lisbon treaty was enacted. Its substance was very similar to the proposed constitutional

Baseline:
Question:
When was there an attempt to reform the law of the E.U.?
True Answer: Following the Nice Treaty
Predicted Answer: 2004

BiDAF:
Question:
When was there an attempt to reform the law of the E.U.?
True Answer: Following the Nice Treaty
Predicted Answer: Following the Nice Treaty

Figure 6: Examples: Blue highlights are the most attended to words in the question, and red highlights are the most attended to word in the context (for BiDAF only)

4.4 Curriculum Learning

To model the question answering task as a reinforcement learning problem, we used two approaches. The first was a top down approach, wherein we modelled the state space as the possible final representation to come out of the model. Given this state, we had a 'policy network' which output two distributions over the context length, one for the start position and one for the end position. This allowed us to estimate the probability of a given trajectory. Given this probability, the return for a trajectory (selection of a particular start and end position) was simply the F1 score or the EM score for that prediction. This can be seen in Figure 1. It allows us to optimize over a non-differentiable reward function.

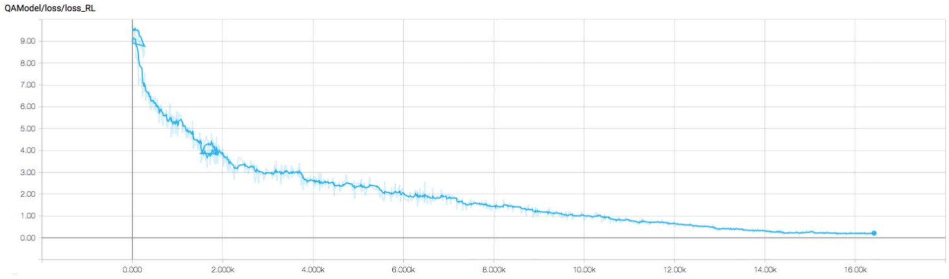


Figure 7: The RL loss with training iterations

In the bottom up approach, the state space was simply the embeddings of the question and context. The entire model built on top of this is considered to be the policy network. The number of parameters is much larger than the top down approach, and hence noise in the gradients causes a severe degradation in performance. This can be combated using a baseline, or a more involved approach such as TRPO (Trust Region Policy Optimization). We did not try these approaches, and instead went with the top down approach which yielded marginally better results.
Training was carried out by using curriculum learning, where the transition to reinforcement learning was made slowly. Over 5000 batches, the loss function was transitioned from the regular loss, to the policy gradient loss, as defined in Equation 8, where the expectation is approximated using a mean and the state is the final hidden state of the network, before the output layer.

$$J(\theta) = -\mathbb{E}[log(\pi_\theta(a_t|S))r_t] = -\mathbb{E}[log\left(P_\theta(a_{start}|S)P_\theta(a_{end}|S)\right)r_t] \tag{8}$$

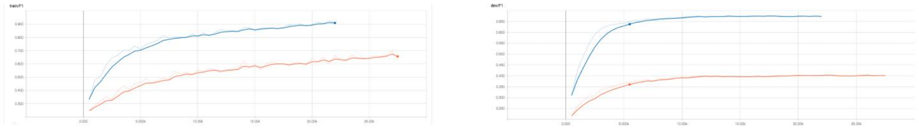
The reward plot (EM plot) plateaus and the RL loss (Figure 7) approaches a minimum, hence we can conclude that the hidden state representation generated by the model is the limiting factor in the learning. Better hidden representations can lead to an improvement in performance. Further, since the RL only updates weights in the output layer, it is less prone to over-fitting than the supervised learning approach. Further, the RL approach achieves slightly better performance with only half the amount of 'supervised' learning. Further, the curriculum learning approach can readily be adapted to deal with unsupervised learning, if there is a way to provide rewards to the model.

4.5 Ensembling

Our best model (without the curriculum learning) was trained with varying initializations to produce an ensemble of Number of models. The probability distributions returned by each of the models was averaged to give an estimate without noise. This new distribution was then used to choose the most probably answer span, again enforcing that the start position must precede the end position and that the length of the answer must be less than 30 words long. This ensemble resulted in the best performance on the test set. We believe that the bottom up approach would do better by using baselines and/or an actor critic RL algorithm. Further, the top down approach can be improved by firstly increasing the size and changing the architecture of the output layer and also further by modelling the span selection as a two step decision process, and then using deep Q learning to optimize the agent.

4.6 Training

Our final model was capable of over-fitting to the training data. In fact, over-fitting was a major problem during training. This was tackled by using a combination of dropout and weight decay. The final model ensemble had models trained with both of these regularization methods. One typical training plot can be seen in Figure 8

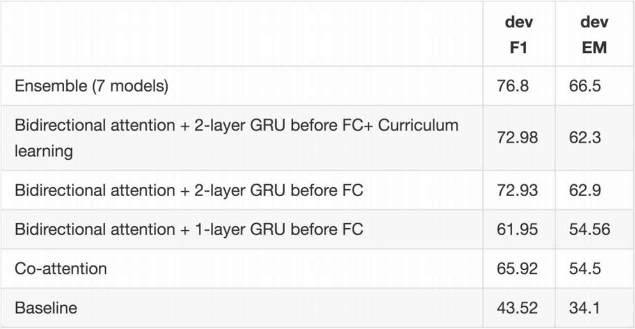


(a) Training F1 score

(b) Dev F1 score

Figure 8: Typical training plots for best model (blue) vs. baseline (orange)

The curriculum learning was implemented by slowly transitioning from supervised learning to RL training. This was done by weighting the losses such that the RL loss slowly increased in proportion. This was done over 5000 batches. Once the transition was complete, pure RL training took place. This essentially froze the first part of the network, allowing the output layer to make best use of the hidden states provided.
Coarse hyper parameter searches were conducted. The learning rate was varied and decayed. Additionally LSTM's were explored instead of GRU's. Hidden layer sizes were changed, and embedding sizes were changed. Additionally iterative output and co-attention were both implemented, but failed to perform as well as our BiDAF model. Self attention was also explored. We believe that these methods have shown to be more effective than BiDAF, however, our implementation was lacking, in that we could not reproduce results.
The final results are shown in Figure 9.



	dev F1	dev EM
Ensemble (7 models)	76.8	66.5
Bidirectional attention + 2-layer GRU before FC+ Curriculum learning	72.98	62.3
Bidirectional attention + 2-layer GRU before FC	72.93	62.9
Bidirectional attention + 1-layer GRU before FC	61.95	54.56
Co-attention	65.92	54.5
Baseline	43.52	34.1

Figure 9: Ablative Analysis of Final Model

5 Conclusions and Further Work

We noticed several shortcomings with the approach we had. However, given more time and computational resources, it is possible to extend on our approach. Further analysis and interesting examples can be found in the **appendix**.

- Iterative reasoning failed to give the desired improvement. We believe that this is due to an incorrect implementation. Re-implementing it correctly on top of our existing BiDAF model is likely to lead to better results.

- The model does not handle out of vocabulary words well. Character level embeddings might help in handling of these out of vocabulary words.

- The RL training is prone to high variance, and does not significantly improve performance. Using the bottom up approach, and modelling the selection of action spans as a two decision process, along with using a more powerful policy gradient approach such as TRPO or an actor-critic method (incorporating a baseline) is likely to do better.

- Since the models are capable of over-fitting, but at the same time have a limited improvement in Dev F1, it shows that they lack the capability to generalize well. Better regularization schemes might help in improving performance on unseen data.

- The example shown here demonstrates one shortcoming of most reading comprehension models:
  Context: Following the success of the 2005 series produced by Russell T Davies, the BBC commissioned Davies to produce a 13-part spin-off series titled Torchwood (an anagram of ”Doctor Who”), set in modern-day Cardiff and investigating alien activities and crime...
  Red coloured words show top 3 most attended words in the question
  Question: What Doctor Who spin-off series was commissioned by the BBC?
  True Answer: Torchwood
  Predicted Answer: Russell T Davies , the BBC commissioned Davies
  The model seems to think that the question is a ”who” question, because of the presence of the words Doctor Who. Lack of context beyond the given paragraph impairs the the ability of the model to do well on certain nuanced questions.
  Further, some questions asked for answers that were informed by data outside of the context.

- The model is confused by double negatives, and negations in general. For example, this synthetic context question and answer pair led to an output which was off the mark:
  Blue coloured words show top 2 most attended words in the context
  Context: ”This” is not the best word ever, unlike ”hello”.
  Red coloured words show top 3 most attended words in the question
  Question: What is not the best word ever? **True Answer:** This **Predicted Answer:** hello
  Augmenting the input with dependency structures could help resolve this issue.

References

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

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

[3] Caiming Xiong, Victor Zhong, and Richard Socher. Dynamic coattention networks for question answering. arXiv preprint arXiv:1611.01604, 2016.

[4] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259, 2014

[5] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In EMNLP, 2014

[6] Buck, C., Bulian, J., Ciaramita, M., Gesmundo, A., Houlsby, N., Gajewski, W., and Wang, W. (2017). Ask the right questions: Active question reformulation with reinforcement learning. arXiv preprint arXiv:1705.07830.