Machine Comprehension

Akhila Josyula , Sreeparna Mukherjee
University of Massachusetts Amherst

Overview

- **Motivation**: Machine Comprehension(MC), answering a question based on a given context, has grown popular in the past few years. Researchers have made significant progress using deep learning techniques, especially variations of recurrent neural networks and attention mechanism. In this project, we experiment with two such implementations called Dynamic Co-attention networks (DCN) and Bidirectional Attention Flow (BiDAF).
- **Problem Statement**: Let $c = \{c_1, c_2, \ldots, c_N\}$ be a sequence of context words of size N and $q = \{q_1, q_2, \ldots, q_M\}$ be a sequence of question words of size M. Our models learns the function $f : (q, c) \rightarrow (start, end)$, where $1 \leq start \leq end \leq N$ defines the span of words in the context that corresponds to the answer to the question.

Context: In meteorology, precipitation is any product of the condensation of atmospheric water vapour that falls under **gravity** . The main forms of precipitation include drizzle, rain, sleet, snow and hail..
Question: What causes precipitation to fall?
Answer: gravity

Dataset

We use Stanford Question Answering Dataset (SQuAD), which consists of 100,000 questions posed by crowd workers on Wikipedia Articles. It is randomly partitioned into a training set(80%), a development set(10%) and a test set (10%).[1]

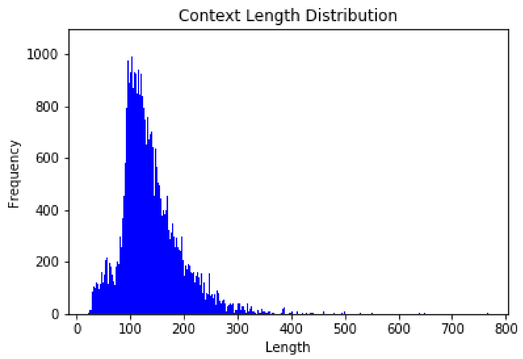
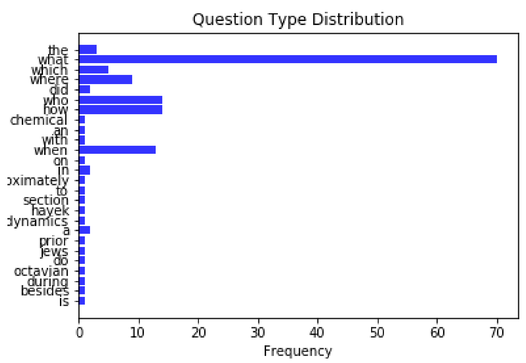


Figure 1: Context Distribution



Architecture

Baseline Model

Our baseline model has three components: a **RNN encoder layer**, that encodes both the context and the question into hidden states,a **context-to-query attention layer**, that combines these representations, and an output layer, which applies a fully connected layer and then two separate **softmax layers** (one to get the start location, and one to get the end location of the answer span).

Dynamic Co-attention Network (DCN)

The Dynamic Co-attention Network (DCN), illustrated in Fig. 3, is an end-to-end neural network for question answering. The model consists of a co-attentive encoder that **captures the interactions between the question and the document**, as well as a dynamic pointing decoder that **alternates between estimating the start and end of the answer span**. This mechanism helps the network to fuse co-dependent representations of the question and the document, so as to focus on relevant parts of both. Following this, the dynamic pointer decoder iterates over potential answer spans which helps the model to recover from initial local maxima corresponding to incorrect answers. [2]

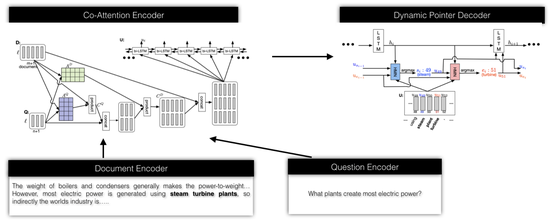


Figure 3: Dynamic Co-Attention Network Architecture

Bidirectional Attention Flow Model(BiDAF)

From our baseline, we incrementally developed the BiDAF model by appending **a contextual layer that encodes the embedded sequences of context and question using bi-directional LSTM**. We then get the representation matrix for the context and the question words. The **attention layer captures the similarity these matrices and uses these to compute the attented question vectors and context vectors**. The attended question vectors are then modeled using bi-directional LSTM and the output span is predicted using Logistic Regression with softmax. Cross-entropy loss is used during training.[3]

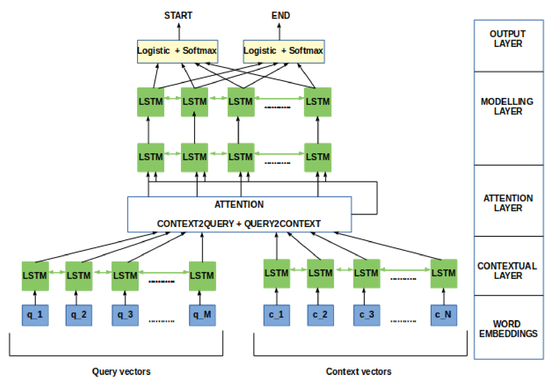


Figure 4: Bidirectional Attention Flow Architecture

Experiment

Evaluation on the SQuAD dataset consists of two metrics. The exact match score (EM) calculates the exact string match between the predicted answer and a ground truth answer. The F1 score calculates the overlap between words in the predicted answer and a ground truth answer. Because a document-question pair may have several ground truth answers, the EM and F1 for a document-question pair is taken to be the maximum value across all ground truth answers.

Baseline Model

The baseline model has been trained and the plots of the EM, F1 score for development set has been recorded for 15k iterations as 40 and 29. After 15k iterations, we notice that the development loss starts increases, at this point we realise that the model has started to over-fit.

Figure 5: F1 and EM for the baseline model

Experiments for DCN and BiDAF

We achieved F1 and EM scores of 71.8 and 59 over the development set for the BiDAF model, as can be seen from the figure below. From the results, we can see that the performance is comparable to state of the art machine comprehension methods. The performance gap with the reference implementation of BiDAF can be explained by lack of character level embeddings

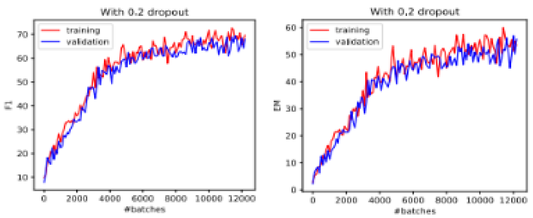


Figure 6: F1 and EM for the BiDAF model

We achieved F1 and EM scores of 55 and 40.1 over the development set for the DCN model. On the decoder side, we have experimented with various pool sizes for the HMN layers. After several analysis, we found the best performance when pool size is equal to 16. The following table illustrates the F1 and EM Scores for each of these experiments.

Model	Dev F1 Score	Dev EM Score
Dynamic Co-Attention Network		
Pool size 16 HMN	55	41
Pool Size 8 HMN	53	39
Pool Size 4 HMN	51	35

Figure 7: F1 and EM for the DCN model

Results

The figure below lists the performance of our implementations of BiDAF and DCN with our baseline model. We have also compared with the baseline performance from the first squad paper using Logistic Regression. We have also mentioned the human performance for reference.

MODEL	F1 score	EM score
Baseline- Logistic Regression (first squad paper)	51.0	40.0
Baseline – RNN with Attention (our implementation)	39.5	28.2
Dynamic Coattention network (our implementation)	55.0	41.0
Bidirectional Attention flow (our implemention)	71.8	59.0
Human Performance	91.2	82.3

Figure 8: Results on Development Set

Conclusion & Future Work

Our implementation has clarified that in the current setting of experimentation, the BiDAF model performs well, despite it's rather simple bi-directional attention mechanism. The DCN model could not surpass the BiDAF model, but have performed significantly better than both the baseline models. In future, we would like to experiment with iterative reasoning techniques as well as test and improve our model to adversarial inputs.

References

[1] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100, 000+ questions for machine comprehension of text," *CoRR*, vol. abs/1606.05250, 2016.

[2] C. Xiong, V. Zhong, and R. Socher, "Dynamic coattention networks for question answering," *CoRR*, vol. abs/1611.01604, 2016.

[3] M. J. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, "Bidirectional attention flow for machine comprehension," *CoRR*, vol. abs/1611.01603, 2016.