Machine Comprehension using Improved Bi-Directional Attention Flow Network

Sijun He
Stanford University
sijunhe@stanford.edu

Jiajun Sun
Stanford University
jiajuns@stanford.edu

Mingxiang Chen
Stanford University
ming1993@stanford.edu

Abstract

1 Introduction

Machine Comprehension (MC) and Question Answering (QA) were increasingly popular in the past few years. Almost all human knowledge are recorded as text. Just like what we did in school doing reading comprehension questions, extracting information from a specific piece of context would enable artificial intelligence to get to a higher level. Different kinds of natural language processing structures such as Gated Recurrent Unit (GRU) and Long Short Term Memory (LSTM) has been introduced.

These years, there are several QA database that has been released. In this paper, we describe a way using an improved version of bi-directional attention flow for the task of question answering using recently published Stanford Question Answering Dataset (SQuAD) which consisted of approximately 100K question-answer pairs with the context.

The objective of the study was to extract the answer for a given question from a certain context. The model was built based on the previous hierarchical multi-stage bi-directional attention flow (BiDAF)) model (Minjoon Seo et al. 2017) while evaluating the correctness using corresponding F1 score and exact-match (EM) score.

2 Models

Our machine learning model (Figure 1) is hierachical multi-stage with 5 layers.

2.1 Word Embedding Layer

The word embedding layer maps each individual word into a high-dimension vector space. It is a set of pretrained GloVe word vectors using datasets provided by Wikipedia 2014 and Gigaword 5. The dimensionality of the vectors are 100.

2.2 Contextual Embedding Layer

In this layer, we introduced a Bi-directional Long Short-Term Memory (BiLSTM) network as our encoding layer. Contexts and questions will be fed into the network seperately, while the output from

each state will all be recorded. Note that the output of which will be 2d-dimension vectors (where d is 100 in our model) due to the concatenation of two output for each direction of the network.

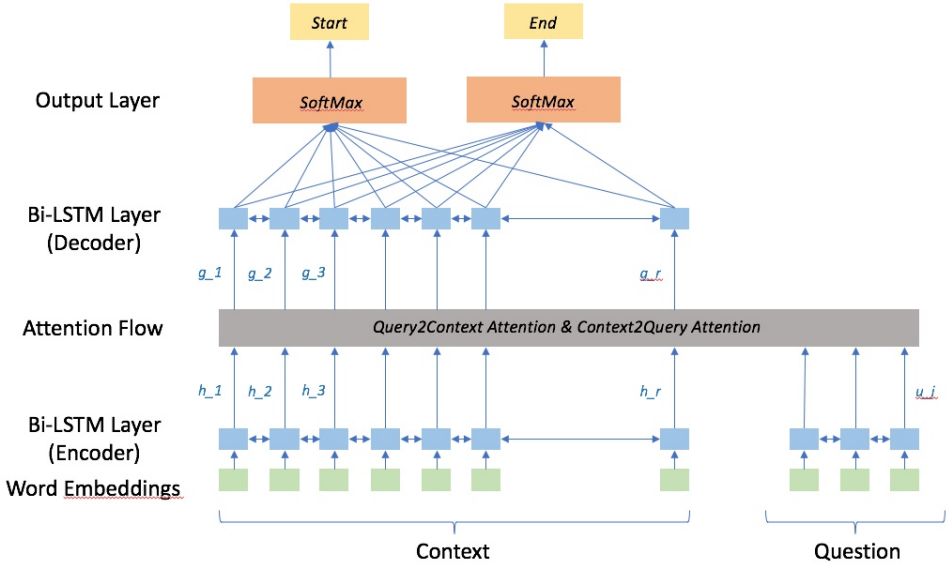

Figure 1: Bi-Directional Attention Flow Model

2.3 Attention Flow Layer

The attention layer is introduced to link together the information from the question and the information given by the context. It tells us how important it is that we should pay attention to each word vector in the text (both contexts and questions). The input of this layer is contextual vector of the context H and the question U.

In this layer, the attention is calculated from the similarity matrix of H and U as follow:

$$S_{ij} = \alpha(H_{:,t}, U(:, j))$$

where S_ij indicates the similarity of the word i in the context and the word j in the query, and α is the bilinear factor which is defined as

$$\alpha(h, u) = h^T W_{bi} u$$

W is a trainable matrix of $R^{2d \times 2d}$ (Chen et al. 2016).

2.3.1 Context-to-query Attention

The context-to-query attention indicates the weight of each word in the query considering the given context. The attention weight is computed as follow

$$weight = SoftMax(S_{i:})$$

so that the query vectors turn out to be

$$\hat{U} = \sum_j a_{ij} U_{:j}$$

2.3.2 Query-to-context Attention

The query-to-context attention indicates the weight of each context word considering the given query. Similar to what we did in context-to-query attention, the attention weights are calculated from

$$b = SoftMax(max_{col}(S)$$

and

$$\hat{H} = \sum_t b_t H_{:t}$$

Then we combined three vectors as the yield G, where β is a trainable vector of $R^{8d \times T}$

$$G = \beta(H, \hat{U}, \hat{H})$$

2.4 Modeling Layer

The input of the modeling layer (decoding layer) encrypts the information of the context together with some information from the query. So that when we pass the tensor to a bi-LSTM network, similar to what we did in the contextual embedding layer, the output would be a matrix of $R^{2d \times T}$, where T indicates the number of context in the dataset.

2.5 Output Layer

In this layer, we provides the answer to questions by calculating the probability for each position in the context as the start and the end of the answer with SoftMax. However, in some situations, though very rare, we would flip over the start and ending labels if the ending predicted maybe even appears earlier than the start in the context. It would be discussed more in detail in section 4.

3 Experiment

3.1 Dataset

In this experiment, we take Stanford Question Answering Dataset (SQuAD) to train our model. The dataset is split into training and validation set. The training set consists of 81381 samples. Each sample comprises of a context paragraph, question and answer. The context paragraphs have length smaller than 1000. Questions have length smaller than 100.

3.2 Baseline

The baseline model processes contextual embeddings and word embeddings through bi-LSTM layer respectively. The hidden state vector from each bi-LSTM layer then go through the attention flow layer. The tensor comes out from attention flow layer has information from both context and question. This tensor will then pass to decoder. The decoder is another bi-LSTM layer; and two softmax layers one for predicting start position and the other for predicting end position.

3.3 Modified Model

In addition to the baseline model, we also modify the model in several different ways to compare the result. As shown in figure xx, a filter layer is added between the contextual embedding layer and Bi-LSTM layer. This filter layer will compute the similarity between context and question. As a result, the filter layer will give an emphasis on the context embeddings. Another modification we made is on the glove word vector dimension.

3.4 Model Details

4 Result

4.1 Discussion

4.2 Error Analysis

5 Conclusion

6 Headings: first level

First level headings are lower case (except for first word and proper nouns), flush left, bold and in point size 12. One line space before the first level heading and 1/2 line space after the first level heading.

6.1 Headings: second level

Second level headings are lower case (except for first word and proper nouns), flush left, bold and in point size 10. One line space before the second level heading and 1/2 line space after the second level heading.

6.1.1 Headings: third level

Third level headings are lower case (except for first word and proper nouns), flush left, bold and in point size 10. One line space before the third level heading and 1/2 line space after the third level heading.

7 Citations, figures, tables, references

These instructions apply to everyone, regardless of the formatter being used.

7.1 Citations within the text

Citations within the text should be numbered consecutively. The corresponding number is to appear enclosed in square brackets, such as [1] or [2]-[5]. The corresponding references are to be listed in the same order at the end of the paper, in the **References** section. (Note: the standard BibTeX style `unsrt` produces this.) As to the format of the references themselves, any style is acceptable as long as it is used consistently.

As submission is double blind, refer to your own published work in the third person. That is, use "In the previous work of Jones et al. [4]", not "In our previous work [4]". If you cite your other papers that are not widely available (e.g. a journal paper under review), use anonymous author names in the citation, e.g. an author of the form "A. Anonymous".

7.2 Footnotes

Indicate footnotes with a number[1] in the text. Place the footnotes at the bottom of the page on which they appear. Precede the footnote with a horizontal rule of 2 inches (12 picas).[2]

7.3 Figures

All artwork must be neat, clean, and legible. Lines should be dark enough for purposes of reproduction; art work should not be hand-drawn. The figure number and caption always appear after the

[1] Sample of the first footnote

[2] Sample of the second footnote

Table 1: Sample table title

PART	DESCRIPTION
Dendrite	Input terminal
Axon	Output terminal
Soma	Cell body (contains cell nucleus)

figure. Place one line space before the figure caption, and one line space after the figure. The figure caption is lower case (except for first word and proper nouns); figures are numbered consecutively.

Make sure the figure caption does not get separated from the figure. Leave sufficient space to avoid splitting the figure and figure caption.

You may use color figures. However, it is best for the figure captions and the paper body to make sense if the paper is printed either in black/white or in color.

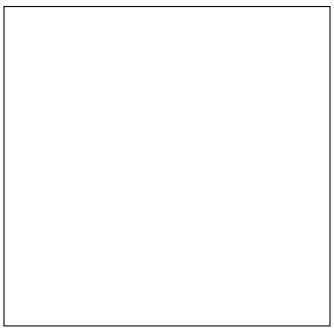

Figure 2: Sample figure caption.

7.4 Tables

All tables must be centered, neat, clean and legible. Do not use hand-drawn tables. The table number and title always appear before the table. See Table 1.

Place one line space before the table title, one line space after the table title, and one line space after the table. The table title must be lower case (except for first word and proper nouns); tables are numbered consecutively.

8 Final instructions

Do not change any aspects of the formatting parameters in the style files. In particular, do not modify the width or length of the rectangle the text should fit into, and do not change font sizes (except perhaps in the **References** section; see below). Please note that pages should be numbered.

Acknowledgments

Use unnumbered third level headings for the acknowledgments. All acknowledgments go at the end of the paper. Do not include acknowledgments in the anonymized submission, only in the final paper.

References

References follow the acknowledgments. Use unnumbered third level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font size to 'small' (9-point) when listing the references. **Remember that this year you can use a ninth page as long as it contains *only* cited references.**

[1] Danqi Chen., Jason Bolton. & Christopher D. Manning. (2016) A thorough examination of the cnn/daily mail reading comprehension task. *arXiv:1606.02858*

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D. S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609-616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer-Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.