IMPROVING CHARACTER EMBEDDINGS OF THE BIDAF ARCHITECTURE

A PREPRINT

Tomas Sykora
xsykor25@stud.fit.vutbr.cz
tms.sykora@gmail.com
Brno University of Technology

January 15, 2020

ABSTRACT

Bidirectional attention flow architecture showed new state-of-the-art results on the various machine comprehension tasks like question answering is. By increasing the convolutional neural network layers for character level embeddings creation to two, adding dropout and especially batch normalisation layers, significant improvements to the original bidirectional attention flow architecture were achieved.

***K*eywords** question answering · bidaf improvement · character embeddings · attention · convolutional neural network · batch normalisation

1 Introduction

The machine comprehension area witnessed many improvements over the last years. Since the deep learning boom a few years ago, the models achieved many state-of-the-art results in image recognition, speech recognition, robotics, natural language processing (NLP) or other machine learning areas. Machine comprehension represents a subset of NLP problems in which understanding of its inputs (e.g. text) is necessary to solve the task. The task may be a text summarization, language modelling or question answering. In question answering, a text containing specific information is given as input together with a question. The question may or may not be answered in the given text. The goal of a question answering system is to either answer the question or state that the answer is not present in the given text. One the approaches of solving the question answering task is a bidirectional attention flow (BiDAF) architecture presented in [1]. This work presents improvements to the original BiDAF architecture. By performing a two layer convolution and a batch normalisation of the input character embeddings, a XX percentage points improvement in the F1 score was achieved.

2 Bidirectional attention flow architecture and proposed improvements

The original work in the [1] presents a machine comprehension architecture taking both word and character embeddings on its input (context and query) as depicted in the figure 1. The embeddings are then run by a bidirectional long short-term memory (LSTM) units [2], an attention layer, two more LSTM layers, and some a feedforward layer with softmax in the end. The word embeddings used in the original paper were the GLOVE embeddings. Character embeddings were based on one convolutional layer over the input characters.

3 Proposed enhancements of the original architecture

The baseline code of this project provided in the course contained implementation of the word embeddings only. The character information was not provided although the SQUAD dataset [3] contained this information. The models showed in this work at first completed the missing implementation part from the original paper which was the character level embeddings taken as input to the bidirectional LSTM layer.

This work experimented with the character embedding part of the architecture. At first, multi layer convolutional networks (CNN) were tried (1-3 layers). During the following experiments, different kernel sizes of the CNN layers and different dropout probabilities between the CNN layers were used.

To further enhance the achieved results, the following architecture modification was implemented. Talking only about the character embedding part of the attention layer input, it consisted of two parts. One was the output of the first convolutional layer with a kernel of size 3. The other one was the output of the second convolutional layer with a kernel of size 5. Both layers were followed by a dropout with a probability set to 0.3.

Different combinations of the all of the above mentioned BiDAF architecture modifications resulted in very similar F1 score improvements between the 2-3 percent points on average in comparison to the course baseline system. However, one major improvement on top of all of the others, batch normalisation, brought a huge improvement of 2.89 percentage points over the second best model and 5.11 percentage points over the course baseline architecture.

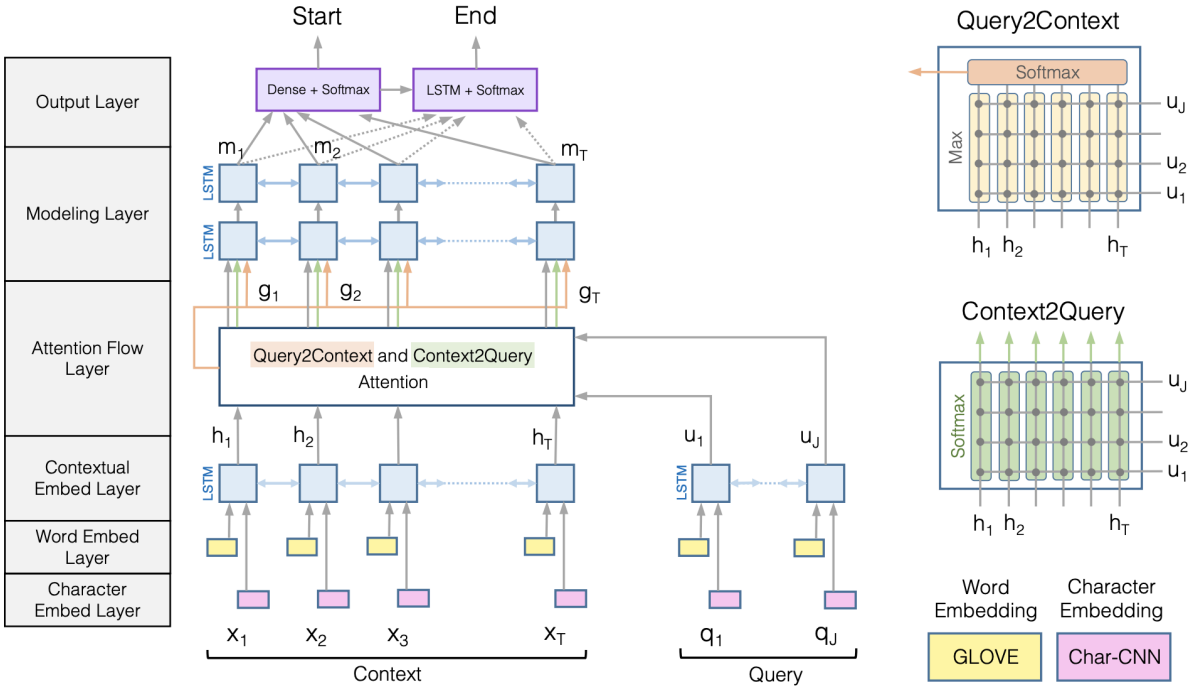

Figure 1: The original bidirectional attention flow architecture (better viewed in color) [1]

4 Experiments

The experiments showed that all of the proposed improvements to the original architecture brought similar results. However, a batch normalisation layer itself improved the results the most, which is clearly demonstrated by the figure visualising the F1 score during the model training 2 (the top dark blue line is the one with batch normalisation).

It has to be mentioned, that only one experiment with the batch normalisation layer was run as I discovered this improvements shortly before the project deadline. Further experiments should be tested in this topic.

Model	EM	F1
baseline	58.23	61.44
1-layer cnn emb	59.57	62.90
2-layer cnn emb + residual	60.46	63.65
2-layer cnn emb + residual + batch norm	63.17	66.55

Comparison of different improvements of the character embedding layer.

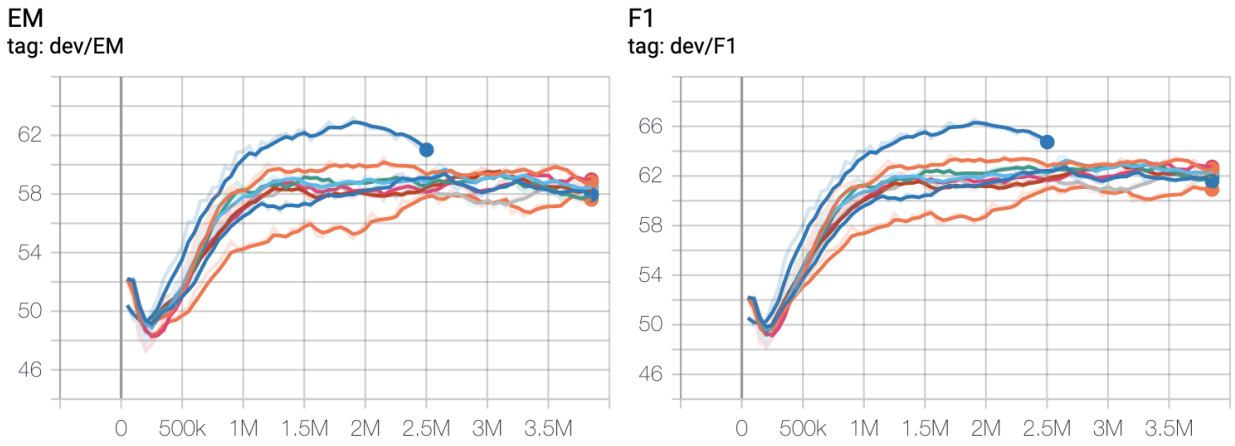

Figure 2: Training of different BiDAF architectures. The top grey line represents the setup with a batch normalisation layer.

5 Conclusion

This work presented improvements of the BiDAF architecture for the question answering task. The presented improvements were 4.94 and 5.11 percentage points in the EM and F1 scores respectively.

References

References

[1] Min Joon Seo and Aniruddha Kembhavi and Ali Farhadi and Hannaneh Hajishirzi. Bidirectional Attention Flow for Machine Comprehension. *arXiv preprint arXiv:1611.01603*, 2018.

[2] Hochreiter, Sepp and Schmidhuber, Jürgen. Long Short-term Memory. *Neural computation*, 1997

[3] Rajpurkar, Pranav and Zhang, Jian and Lopyrev, Konstantin and Liang, Percy. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016