Project Milestone

Abhishek Goswami
Microsoft
Redmond, WA 98052
agoswami@microsoft.com

Abstract

This document details the progress made towards the project milestone of Course CS224n, Winter 2019.

1 Introduction

Machine comprehension (MC) and question answering (QA) tasks have gained significant interest in the past few years, with several end-to-end models showing promising results for multiple datasets. A key factor in recent advancements has been the use of neural attention mechanisms. They fundamental idea behind attention is extract useful signal by exploiting the notion of *matching*.

Several attention approaches have been proposed in literature. Chen et al [1] propose a *uni-directional* attention mechanism whereby the query is used to attend to the context paragraph. In BiDAF, Seo at al [5] introduce *bi-directional* attention flow to obtain a query-aware context representation. This provides complimentary information from both the context and query. Wang et al [6] note that question-aware passage representations have limited knowledge of the context itself. There exists some sort of lexical and syntactic difference between the question and the passage. This motivates them to directly match the question-aware passage representation against itself as a form of *self-matching* attention.

Another key tenet of the proposed techniques is to use a model to process sequential inputs. This is typically done in the form of an embedding encoder layer. While recurrent neural networks have been the model of choice for this, recent work by Yu et al [7] propose using a convolution and self-attention mechanism instead.

In this project we explore two novel extensions. First, we observe that most existing models dive straight into the encoding layer, given the sequential inputs. Attention is an afterthought. One problem of such a representation is that it strains the embedding encoder layer, since the rest of the modeling layers are all stacked on top of it. We propose adding a 'Base Attention Layer' as a form of self-attention over the raw word and character input embeddings and explore whether that improves model performance. Second, we explore whether we can have a contextual embed layer that uses a combination of recurrent layers and convolution layers. The motivation behind this is to bring the best of both worlds in the contextual embedding space.

2 Our Models

In this section, we first formulate the machine comprehension problem and then describe the models we explored.

2.1 Problem Statement

The machine comprehension task considered in this paper is as follows. Given a context paragraph with N words C = c1, c2, ..., cN and a query sentence with M words Q = q1, q2, ... qM output a span S = ci, ci+1,...ci+j from the original paragraph C. In the following we use x to denote both the original word and its embedding vector for any x in C, Q.

Table 1: An example of a machine comprehension task.

Question	Economy, Energy and Tourism is one of the what?
Context	Subject Committees are established at the beginning of each parliamentary session, and again the members on each committee reflect the balance of parties across Parliament. Typically each committee corresponds with one (or more) of the departments (or ministries) of the Scottish Government. The current Subject Committees in the fourth Session are: Economy, Energy and Tourism; Education and Culture; Health and Sport; Justice; Local Government and Regeneration; Rural Affairs, Climate Change and Environment; Welfare Reform; and Infrastructure and Capital Investment
Answer	current Subject Committees

2.2 Models

In this section we describe the following model types we explore (1) baseline model (2) base attentional model (3) rnn-conv contextual embedding model

2.2.1 Baseline Model

Our baseline model is based on BiDAF [5]. The default project did not include a character-level embedding. Character-level embeddings allow us to extract signal from the internal structure of words. As per the original BiDAF model, we include a character-level embedding layer using character-level convnets and consider this to be our baseline model.

2.2.2 Base Attentional Model

This model extends the baseline model by adding a 'Base Attention Layer' as a form of self-attention over the raw word and character input embeddings. The motivation for adding this layer is two fold (a) the encoder layers could have missed some signal, hence so just having post-attention layers may be insufficient (b) feed in more inputs from the input embed layers to the contextual embedding layers.

2.2.3 RNN-Conv Contextual Embedding Model

This model extends the baseline model by having a contextual embedding layer that uses both RNN and Conv nets. The recurrent layers helps in processing sequential input, while convolution captures local structure of the text.

3 Experiment

In this section, we conduct experiments to study the performance of our models. We will benchmark our models on the Stanford Question Answering Dataset (SQuAD) 2.0 [4], considered to be one of the most competitive datasets in QA tasks. We also provide some implementation details for our models and present the main results.

3.1 Dataset

We consider the Stanford Question Answering Dataset (SQuAD) 2.0 [4] for machine comprehension.

3.2 Main Results

Table 2 details the preliminary results we have till this point.

Table 2: Preliminary Results

	Dev Set		**Test Set**	
	EM	F1	EM	F1
Baseline Model	59.25	62.28	59.469	62.464
Baseline Model Variants (with self-similarity before attention)	58.83	62.05	57.469	60.873
Base Attentional Model				
RNN-Conv Contextual Embedding Model				

4 Related Work

Machine comprehension (MC) and question answering (QA) tasks have gained significant interest in the past few years. Overall, the models and techniques that work best of these tasks fall into two categories.

One, techniques that leverage pre-trained contextual embeddings (PCE). Examples of such PCE-based techniques are ELMo [3] and BERT [2]. The core idea of such techniques is that in order to represent a piece of text, we should use word embeddings that depend on the context in which the word appears in the text. This is typically achieved by pretraining the weights on a large-scale language modeling dataset, and using the pre-trained weights for the initial model layers.

Secondly, there are the several end-to-end, non-PCE models which have shown promising results. Examples of such techniques are BiDAF [5], R-NET [6] and QANet [7].

5 Conclusion

So far we have seen that character-level word embeddings help improve the baseline model. As next step we plan to implement the Base Attentional Model (Section 2.2.2) and the RNN-Conv Contextual Embedding Model (Section 2.2.3)

References

[1] D. Chen, J. Bolton, and C. D. Manning. A thorough examination of the CNN/Daily Mail reading comprehension task. In *Association for Computational Linguistics (ACL)*, 2016.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[4] P. Rajpurkar, R. Jia, and P. Liang. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*, 2018.

[5] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*, 2016.

[6] W. Wang, N. Yang, F. Wei, B. Chang, and M. Zhou. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 189–198, 2017.

[7] A. W. Yu, D. Dohan, M.-T. Luong, R. Zhao, K. Chen, M. Norouzi, and Q. V. Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*, 2018.