Iterative reasoning with bi-directional attention flow for machine comprehension

Anand Dhoot, Anchit Gupta

Stanford University

Introduction

- End to end deep model for machine comprehension combining features of multiple SOTA architectures.
- Iterative self attention mechanism combined with highway networks.
- Single model F1 **78.63**, ensemble F1 **81.12**

Contextual Embedding

- Pre-trained fastText word vectors which incorporate char level sub-word information.
- Cross cosine similarity feature appended to the context, query embeddings
- Shared Bi-LSTM layer

Bi-directional attention

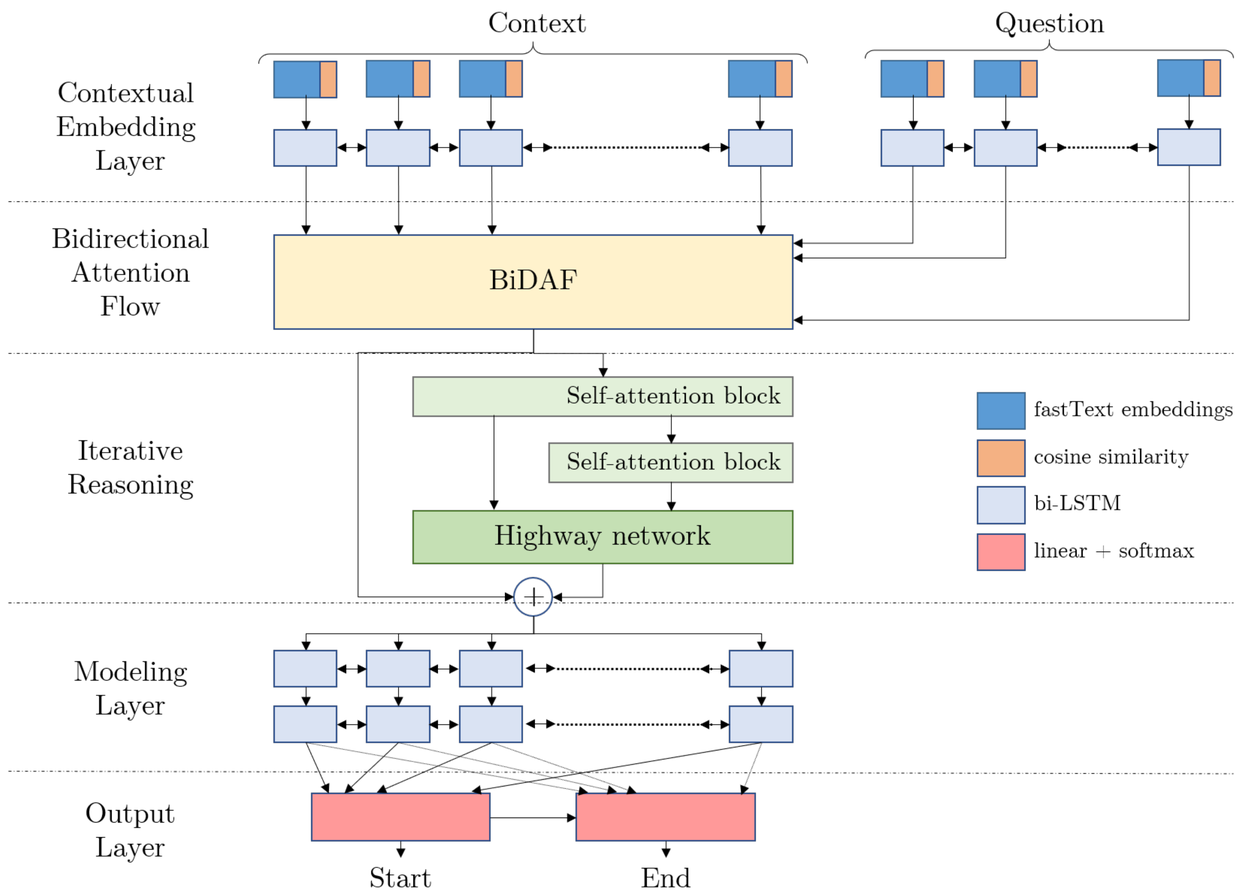
Standard bidaf layer [2]. Computes C2Q (Context to Query) and Q2C (Query to Context) attention

Iterative Reasoning

- Iteratively refine the representation just as humans do while rechecking to confirm their answer.
- Multiple self attention layers combined using highway networks.

Output Layer

- Span start probability distribution is computed from the output of the modeling layer by passing through a linear layer followed by a softmax.
- Conditioned end probability by running a LSTM on the modeling layer output concatenated with the logits of p^{start}. As before passed through a linear layer followed by a softmax to give p^{end}



Optimizations

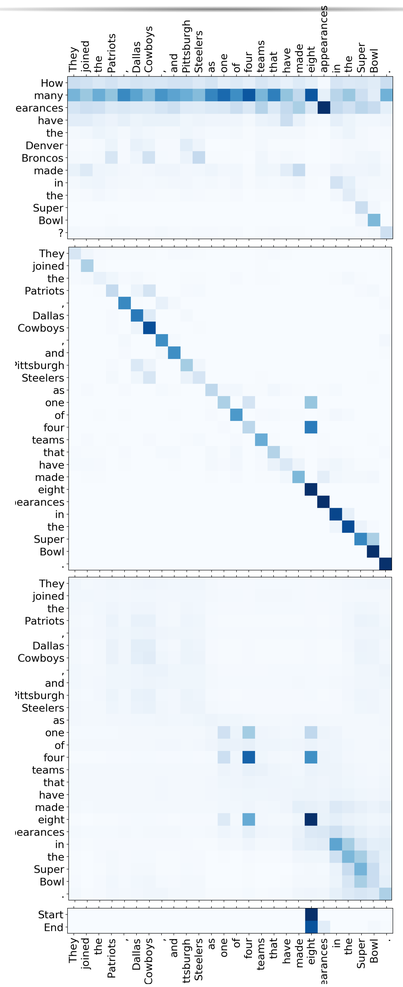
- Dynamic programming based span prediction on p^{start}, p^{end} predicted by the model.
- Using cased embeddings & data because word case often has extra information.
- Use CUDNN LSTMs to give a 4X reduction in run time and better GPU memory usage.
- Embedding lookup on CPU, conserves GPU memory allowing us to have embeddings with a vocabulary size of 2 million words.

Results

Model (Single unless specified)	F1	EM
Multi perspective matching [4]	75.1	65.5
Dynamic Coattention [5]	75.9	66.2
Bidaf [2]	77.3	68.0
R-net [3]	80.7	72.3
Reinforced Mnemonic [1]	81.8	73.2
IR-Bidaf Single Model (Ours)	78.6	69.4
IR-Bidaf Ensemble (Ours)	81.1	73.2

Table: Comparison with published models on test set

Attention Analysis



References

[1] M. Hu, Y. Peng, and X. Qiu. Mnemonic reader for machine comprehension. *CoRR*, abs/1705.02798, 2017.

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

[3] W. Wang, N. Yang, F. Wei, B. Chang, and M. Zhou. Gated self-matching networks for reading comprehension and question answering. pages 189–198, 2017.

[4] Z. Wang, H. Mi, W. Hamza, and R. Florian. Multi-perspective context matching for machine comprehension. *CoRR*, abs/1612.04211, 2016.

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