Reasoning on Multi-Hop Questions with HotpotQA

Jamil Dhanani, Suma Kasa, Suprita Shankar; *Mentor: Peng Qi*

Introduction

- HotpotQA [1] is a question-answering (QA) dataset focusing on multi-hop reasoning; the answers to the questions come from aggregating parts of the context
- Also requires explanation in the form of "supporting facts" to produce an answer
- We augmented the baseline model for HotpotQA, by proposing improvements in **learning and optimization**, **attention**, **reasoning**, and **representation**

Example Question

Question: Who is older, Annie Morton or Terry Richardson?

Gold Answer: Terry Richardson

Gold Supporting Facts: Annie Morton, Terry Richardson

Baseline Answer: Annie Morton

Baseline Supporting Facts: Annie Morton, Kenton Richardson

Modifications

- For **learning and optimization**, we adapted the learning rate decay, added regularization, and tried different optimizers (such as Adam)
- For **attention**, we added self-attention on the query and context, with shared weights between the two, with multiple architectures
- We tried architectures adding bidirectional attention at various points in the network, to assist with **multi-hop reasoning**
- Replaced GRUs with Gated CNNs for selected architectures; some work has shown improvement with CNNs for NLP tasks
- Improved **representation** with architectural changes, adding hidden layers and mitigating bottlenecks by ensuring that inputs to each layer are sufficiently large to represent the information required.

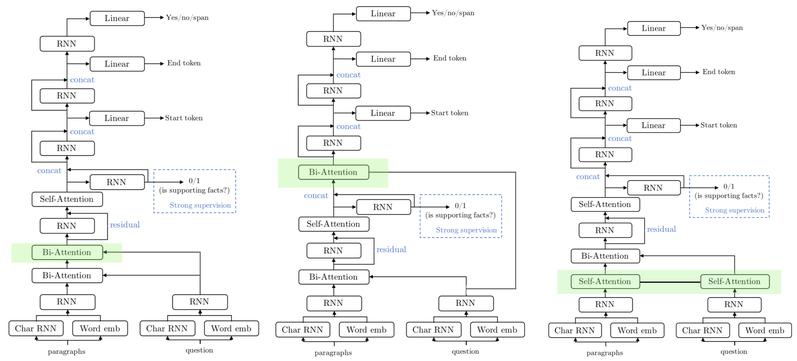
Refined Model Answer

Question: Who is older, Annie Morton or Terry Richardson?

Model Answer: Terry Richardson

Model Supporting Facts: Annie Morton, Terry Richardson

Models improving reasoning and attention; Rsn-A, Rsn-B, Att-C



Results Analysis

- On best-best-performing model, improvement of +**6.70** points in Ans F1, +**14.49** points in Sup F1
- Our **Rsn-A** model makes significant gains in answers which match in the EM metric; answering correctly an additional **1781** questions, and regressing in only **385**
- The number of correct answers with correct supporting facts increases from only **4.93%** to **19.75%**
- An example question is shown on the left, of a question from the dataset, the baseline answer, and our best model (Rsn-A) answer

Table 4: Results from Best-Performing Models

	Model	Ans F1	Ans EM	Sup F1	Sup EM	SP Prec	SP Rec	Joint F1	Joint EM
Base	Baseline	58.28	44.44	66.66	21.95	65.55	70.00	40.86	11.56
Opt-C	adam +dropout	60.25	45.66	66.12	20.42	66.83	71.11	42.31	10.88
Att-C	self-att,q+c	60.03	46.33	69.11	23.52	68.28	76.25	44.09	12.80
Rep-A	no linear, +hidden	64.47	49.95	75.41	33.12	75.22	81.04	50.92	19.00
GCNN-D	+bi-att post-self-att	65.52	49.36	70.78	31.22	73.29	81.94	50.41	20.43
Rsn-A	Rep-A +bi-att start-token	**65.72**	**50.91**	**78.99**	**37.20**	**77.20**	85.27	**53.77**	**21.24**
Rsn-B	+2 bi-att ques_output	64.88	50.37	78.47	35.62	75.87	**85.80**	52.93	19.89

Table 6: Comparison of Baseline and Best-Performing Model EM

	Baseline Correct	Baseline Wrong
Model Correct	1808	1781
Model Wrong	385	3402

Table 7: Comparison of Baseline and Best-Performing Model F1

Baseline F1 > Model F1	Model F1 > Baseline F1	Baseline F1 = Model F1
606	2786	3984

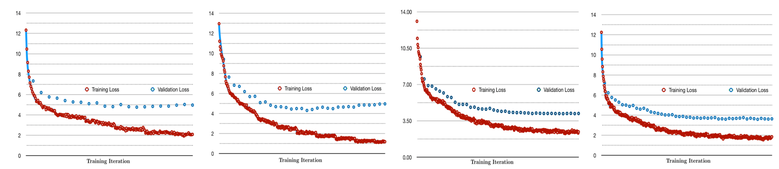


Figure 4: **Baseline** `lr /= 2.0`

Figure 5: **Opt-A** `lr /= 1.5`

Figure 6: **Opt-B** `lr /= 1.5` + Dropout

Figure 7: **Opt-C** Adam + Dropout

Loss curves when exploring learning and optimization improvements

Analysis — Context Attention

- As seen below, in the baseline model, the attention distribution in the self-attention layer is spread out across many different entities in the context paragraphs
- In model **Rsn-A**, the attention is focused on the correct answer ("Terry Richardson")

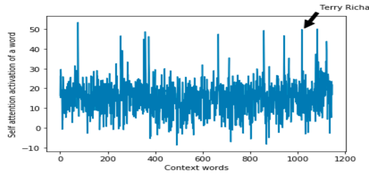


Figure 8: Self-attention Activation, Baseline

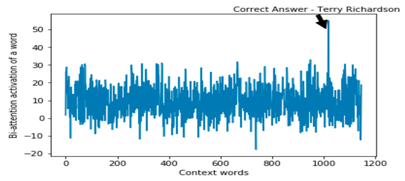
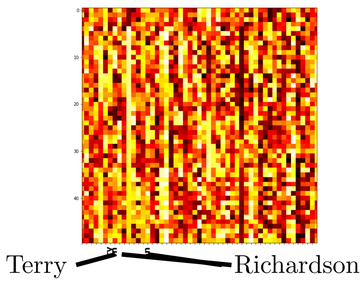


Figure 9: Bi-attention Activation, **Rsn-A**

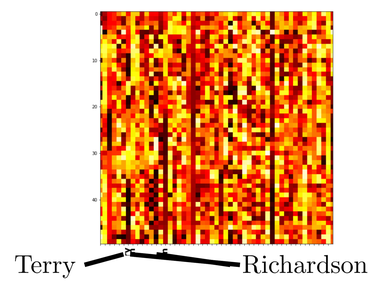
Analysis — Query Attention

- In the Bidirectional Attention Layer, the baseline model activations show little prominence for the correct answer in the query, across all context words
- In model **Rsn-A**, again the attention is focused on the correct answer, in the query

Baseline



Model Rsn-A



Conclusion & Future Work

- Our changes to the baseline model show significant improvements, bringing the F1 scores from 59.02 to 65.72 with our best model, Rsn-A.
- There is still room for additional hyperparameter optimization to get marginal gains from the reported F1 scores
- We would like to explore memory networks and other models which aid in more complex reasoning

[1] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov, and C. D. Manning, "HotpotQA: A dataset for diverse, explainable multi-hop question answering," 2018