Generative Multi-Hop Question Answering with Compositional Attention Networks

Ammar Alqatari (ammarq@stanford.edu), Mentor: Drew Hudson

Problem and Motivation

A new frontier in machine reading comprehension is multi-hop question answering, which requires composing multiple pieces of evidence from a long context document to arrive at the correct answer. Current state-of-the-art models, such as BiDAF [1], which exceed human performance on extractive QA, fail to achieve comparable results in multi-hop QA.

The recently developed compositional attention network (MAC) [2] uses iterative reasoning steps to make predictions, and has succeeded at visual question answering tasks. The network architecture shows promise of performing well at multi-hop QA. I adapt the model and test it on a multi-hop QA dataset.

Method

Data

I use the HotpotQA dataset to test my developed model. HotpotQA is desigend for "diverse, explainable, multi-hop question answering" [3]. The dataset consists of a quality-controlled collection of crowd-sourced questions and answers based on passages from related Wikipedia articles. The main task is to predict the answer given a question and a context passage composed of 10 paragraphs. An additional task measures the justification ability of the model by asking it to provide supporting sentences as part of the output.


Paragraph A:

Return to Olympus is the only album by the alternative rock band Malfunkshun. It was released after the band had broken up and after lead singer Andrew Wood (later of Mother Love Bone) had died of a drug overdose in 1990. Stone Gossard, of Pearl Jam, had compiled the songs and released the album on his label, Loosegroove Records.

Paragraph B:

Mother Love Bone was an American rock band that formed in Seattle, Washington in 1987. The band was active from 1987 to 1990. Frontman Andrew Wood's personality and compositions helped to catapult the group to the top of the burgeoning late 1980s/early 1990s Seattle music scene. Wood died only days before the scheduled release of the band's debut album, "Apple", thus ending the group's hopes of success. The album was finally released a few months later.

Q: What was the former band of the member of Mother Love Bone who died just before the release of "Apple"?

A: Malfunkshun

Example instance from HotpotQA

The data consists of around 90k training examples, which the authors classify into easy single-hop questions (18k), medium multi-hop questions (56k), and hard (15k) multi-hop questions. The dev and test sets consist of 7k hard multi-hop questions each.

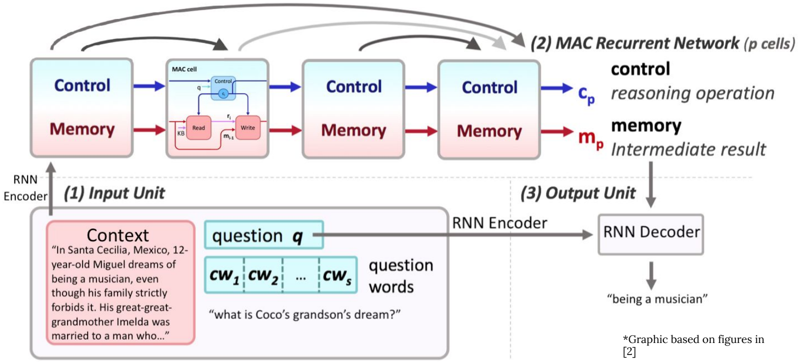
Task and Evaluation

In this project, I focus on the answer prediction task of the dataset. Namely, given a question **q** and context document **c**, the model should produce a variable length answer **a.**

I evaluate the model's performance based on the average exact-match (EM) and F1 score of all predictions on the dev set.

I compare the performance of the model to the baseline developed by the authors of HotpotQA. The baseline uses a BiDAF-based model adapted to processing multiple paragraphs [4]

Model



*Graphic based on figures in [2]

The MAC model is composed of an input unit, a MAC recurrent network, and an output unit. The input unit transforms the raw question string and knowledge base into vector representations which can be fed to the recurrent component. The recurrent component consists of a string of p MAC cells, each consisting of hidden control and memory states. The control and memory states interact through 3 operational units: a control unit, a read unit, and a write unit, based on attention mechanisms over the inputs and the previous hidden states. Finally, the output unit transforms the distributed vector representation outputted by the recurrent component into the desired output form.

Results

I trained several models with a grid search over the learning rate, number of MAC cells, hidden encoding dimension, and number of RNN layers.

The best-performing model achieved an **EM score of 51.5% and F1 score of 61.26** on the dev set.

The model already exceeds the performance of the baseline model developed by the authors of HotpotQA, which achieves an EM score of 45.60% and an F1 score of 59.02 (albeit on the test set).

	Model	Code	Ans		Sup		Joint	
			EM	F_1	EM	F_1	EM	F_1
1 Nov 21, 2018	QFE (single model) *NTT Media Intelligence Laboratories*		53.86	68.06	**57.75**	**84.49**	**34.63**	**59.61**
2 Mar 4, 2019	GRN (single model) *Anonymous*		52.92	66.71	52.37	84.11	31.77	58.47
3 Mar 1, 2019	DFGN + BERT (single model) *Anonymous*		55.17	68.49	49.85	81.06	31.87	58.23
4 Mar 4, 2019	BERT Plus (single model) *CIS Lab*		**55.84**	**69.76**	42.88	80.74	27.13	58.23
5 Oct 10, 2018	Baseline Model (single model) *Carnegie Mellon University, Stanford University, & Universite de Montreal* (Yang, Qi, Zhang, et al. 2018)		45.60	59.02	20.32	64.49	10.83	40.16
- Feb 27, 2019	DecompRC (single model) *Anonymous*		55.20	69.63	N/A	N/A	N/A	N/A

Discussion

- Among all trained models, the best performing model had the largest number of MAC cells, the largest encoding dimension, and largest number of RNN layers in the encoder. Models with larger encoding dimensions lead to out-of-memory errors, while models with a larger number of RNN layers were much slower to train. This result suggests that with more computational power, a model with larger parameters could achieve an even better performance.
- The top-performing leaderboard models make use of BERT. Since my developed model makes use of pre-trained word embeddings but not contextual embeddings, I expect that incorporating contextual embeddings will improve the model.
- The success of MAC on the HotpotQA dataset suggests promise to exploring variants of memory-augmented networks and their effectiveness in various MRC tasks.
- It also calls for testing MAC on other MRC tasks which require compositional reasoning, such as conversational QA to further show the network's robustness and versatility

Future Work

- Evaluating the network's selection of supplementary facts
- Incorporation of the supplementary fact data into the model to directly learn the attention mappings through strong supervision during training.
- Using BERT or ELMO contextual embeddings rather than a randomly-initialized RNN in the input unit
- Using a pointer-generator decoder model rather than an RNN decoder in the output unit
- Adding modules which can extend the network to perform on the fullwiki setting of HotpotQA
- Testing the network architecture on other multi-hop QA datasets, and submitting it to the HotpotQA leaderboard

References

[1] Seo, M., Kembhavi, A., Farhadi, A., Hajishirzi, H. (2016). Bidirectional attention flow for machine comprehension. arXiv preprint arXiv:1611.01603.

[2] Hudson, D. A., & Manning, C. D. (2018). Compositional attention networks for machine reasoning. arXiv preprint arXiv:1803.03067.

[3] Yang, Zhilin, et al. "Hotpotqa: A dataset for diverse, explainable multi-hop question answering." arXiv preprint arXiv:1809.09600 (2018).

[4] Christopher Clark and Matt Gardner. 2017. Simple and effective multi-paragraph reading comprehen- sion. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics.*