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**
I use the HotpotQA dataset to test my developed model. HotpotQA is designed 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 document. 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).

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

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).

**Results**
I evaluated 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].

**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 explore 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**