Question Answering with Neural Networks

Ye Tian, Nicholas Huang, Tianlun Li

Motivation

We tackle the Question Answering problem in this project with Neural Network models.

Most if not all **Natural Language Understanding** questions can be cast as Question Answering problem. It is also a fundamental question in building towards artificial intelligence, combining Natural Language Understanding, Information Retrieval, and even higher-order reasoning.

The attention mechanism originally was applied on Machine Translation problem, but it has been shown to work on Questions Answering problems as well. Traditional LSTM networks have problem with long sentences, and even worse for question based on contexts with multiple sentences, given a fixed monolithic representation length. The attention mechanism can solve this problem by "attend to" some portion of the contexts while paying less attention to others. Intuitively this puts a lighter burden on the vector to represent all the semantic information.

End-to-End Memory Network with Attention (MemN2N) has received academic interest in recent years, including a NIPS workshop in 2015. End-to-End Neural Networks have the advantage over traditional machine learning methods that they don't require feature engineering or supervision. Memory Network is unique in the sense that in additional to the hidden states, it relies on an external memory representation that we can analyze during the learning process.

Dataset

We used *bAbI* dataset designed by Facebook researchers. The dataset consists of 3-tuples of

Context-Question-Answer (Supporting evidence index). Contexts are typically 2~10 short sentences. Each answer is a single word, and the supporting evidence indices refer to context sentences that contribute to the answer.

The Questions and Answers are carefully constructed to remove real world bias. For example, in the "Basic deduction" task set, there was this question

1 Wolves are afraid of mice.
2 Sheep are afraid of mice.
...
8 Gertrude is a wolf.
...
11 What is gertrude afraid of? mouse 8 1

Which apparently makes no sense if World Knowledge is involved, but also means the system cannot rely on world knowledge embedded in some representation to answer this kind of problem.

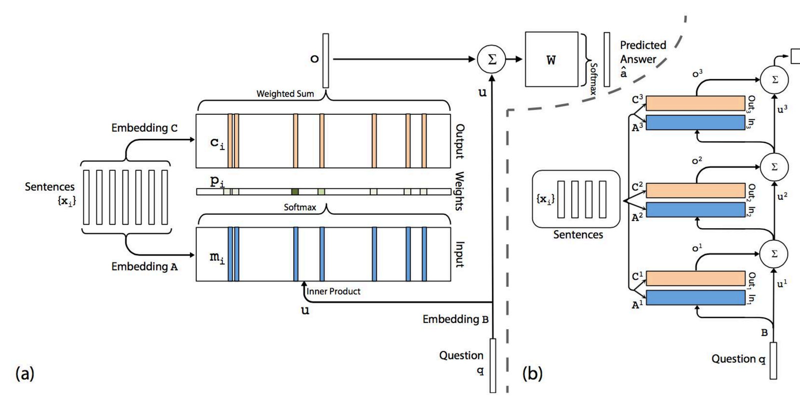
The overall vocabulary size of the dataset is trivial, measured at around 100, which can be a weak point of the dataset.

1- Single Supporting Fact
2 - Two Supporting Facts
3 - Three Supporting Facts
4 - Two Arg. Relations
5 - Three Arg. Relations
6 - Yes/No Questions
7 - Counting
8 - Lists/Sets
9 - Simple Negation
10 - Indefinite Knowledge
11 - Basic Coreference
12 - Conjunction
13 - Compound Coref.
14 - Time Reasoning
15 - Basic Deduction
16 - Basic Induction
17 - Positional Reasoning
18 - Size Reasoning
19 - Path Finding
20 - Agent's Motivation

Figure 1: The tasks in the *bAbI* dataset

Model

We attempted the End-to-End Memory Network and compare it with three variants of LSTMs: LSTM with attention mechanism, Pyramid LSTM, and Pyramid LSTM with attention mechanism.



"End-To-End Memory Networks", Sukhbaatar el al., 2015. arXiv:1503.08895.

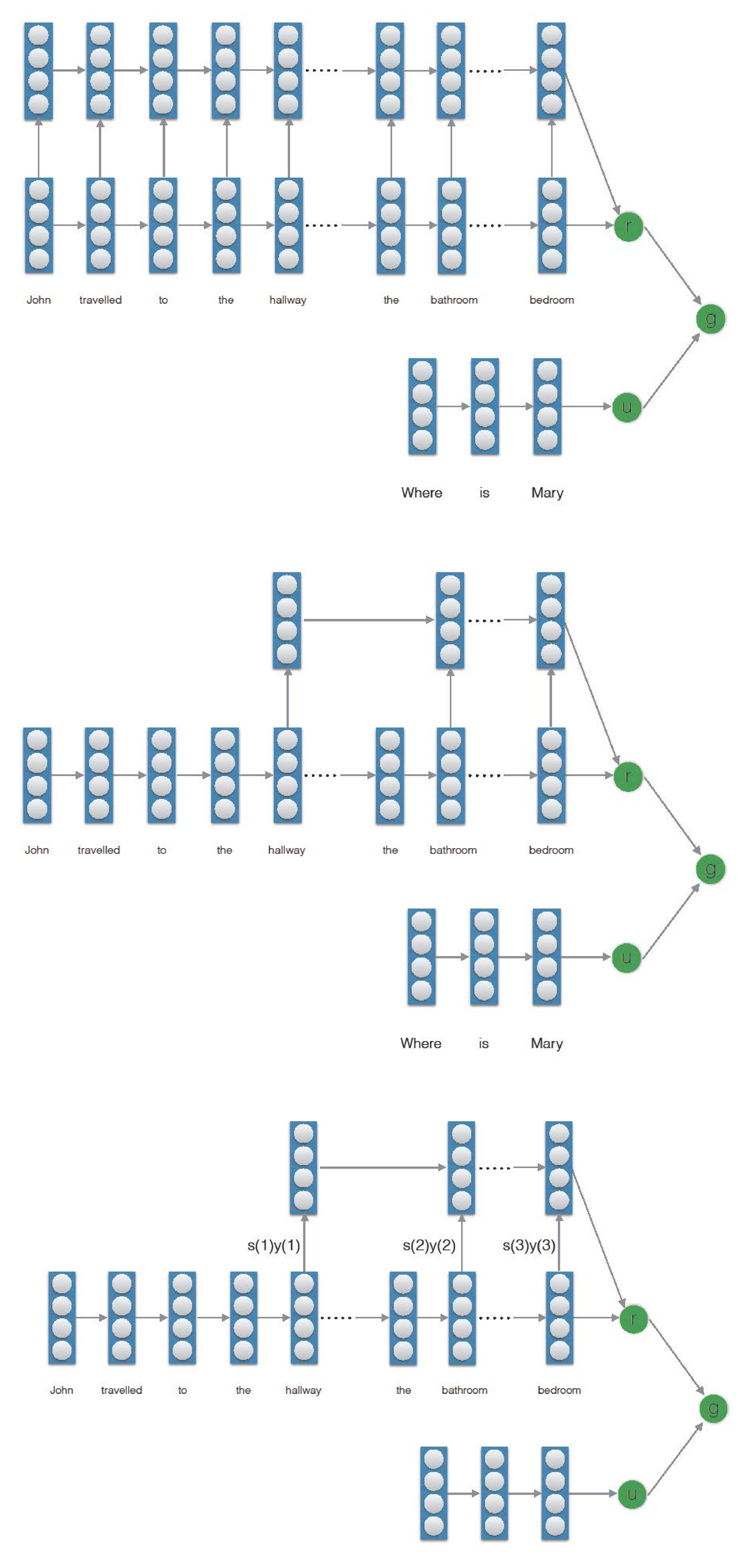
As is shown in the architecture graph, Memory Network layers rely on a common external memory "Sentences".

End-to-End Memory Network is a smooth version of the Memory Network designed by Weston et al. (2015). The main improvement is to make the max selection of u a softmax function, thus the whole system can be differentiated and optimized end-to-end with Stochastic Gradient Descent or RmsProp.

Long Short-Term Memory has long been regarded as useful tool for sequence-based learning. In this project we implement modifications to the classical LSTM structure.

LSTM was introduced to solve problem of RNN with long term dependency. But even with LSTM structure long term dependency is still problematic. **Attention mechanism** can let us visualize the whole model. For example, in language translation machine we can understand the process of translation by visualize the weight matrix.

Pyramid LSTM: instead of combining word representation by simple addition, a Pyramid LSTM feeds context sentences into a second LSTM, and so on, to get the context representation r. This has the potential benefit that the representation will shift less, as shown below.



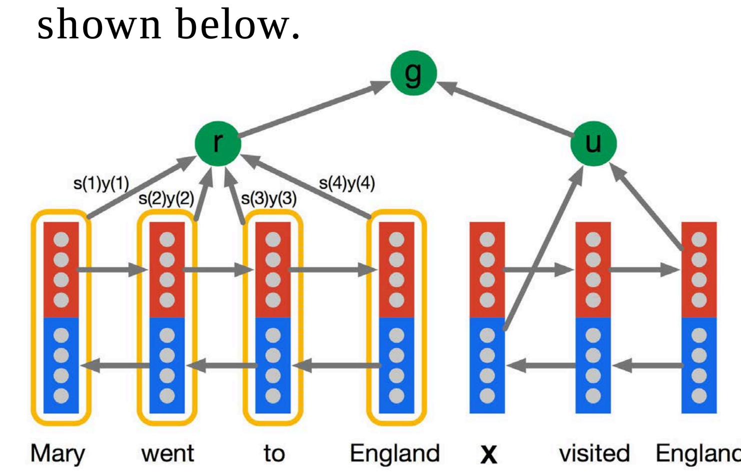


Figure 2: From Top to Bottom: MemN2N, LSTM, LSTM with Attention, Pyramid LSTM with Attention, Pyramid LSTM

Results

As shown below is the result of running five models on the 20 tasks in the *bAbI* dataset.

Among the results we see **MemN2N** is outperforming LSTMs on most tasks except task 18 (size reasoning), but after **Position Encoding** is implemented it is performing as well, or outperforming LSTMs. Among the tasks, Task 2, 3 (Two facts, Three facts), Task 19 (spatial reasoning) are particularly hard for all models.

We suspect Task 19 is due to different representation of direction words ("South") in the context and in the answer ("S").

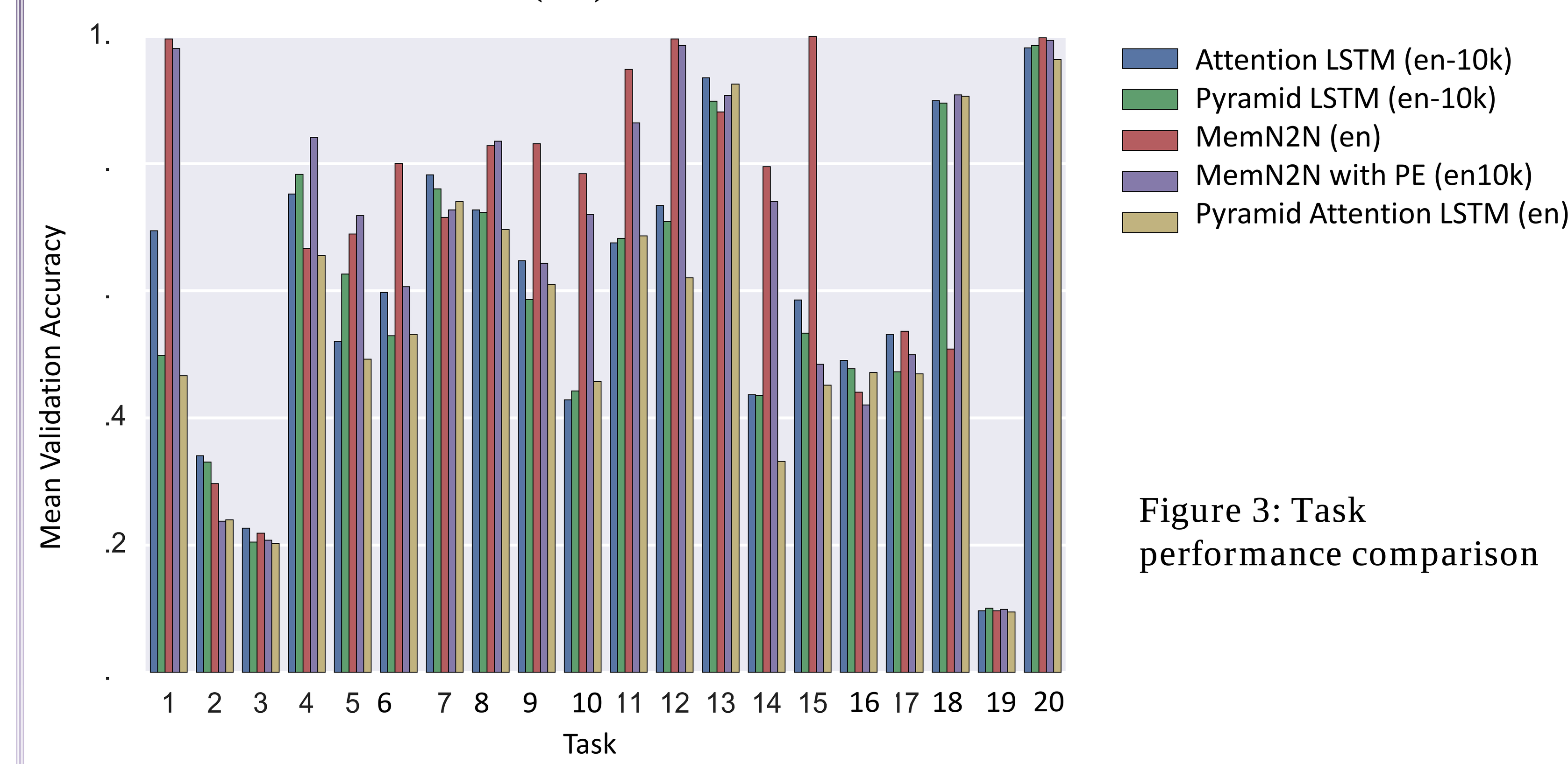


Figure 3: Task performance comparison

Future Work

1. Tune hyperparameters to match the paper performance. Also we should train on en-10k dataset, and with more iterations
2. Error Analysis. What errors do our models make systematically?
3. Visualize Attention in the sentences over iterations. This will provide a lot of insight to the working of attention mechanism.
4. Generalize to the Children's Book Test dataset, which is also developed by Facebook Research, and Deepmind Daily Mail dataset. Both dataset follow similar pattern, which is asking a single fact-based question after a short (usually 15-sentences-long paragraph). The answer will also be a single word so it is easy to adapt to for neural networks. The difference between them and *bAbI* is in the v
   Vocabulary size
   - bAbI ~100
   - CBT ~53,628
   - Daily Mail ~ 208,045
5. Theoretically we understand, that the number of hops in a neural network should be at least greater than the steps of logical deduction in a reasoning problem. We could examine the hypothesis, that given enough hops, the neural networks can eventually overcome reasoning problems of any complexity.
6. Memory Network which learns to control its own number of hops
7. Maybe we can formulate and implement LSTM with explicit Memory Representation.

References

1. Neural Machine Translation by Jointly Learning to Align and Translate Bahdanau et al., 2014. arXiv:1409.0473
2. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention Xu et al., 2015. arXiv:1502.03044
3. Memory Networks, Wetson et al., 2014. arXiv:1410.3916
4. End-To-End Memory Networks, Sukhbaatar el al., 2015. arXiv:1503.08895.
5. Towards AI Complete Question Answering: A Set of Prerequisite Toy Tasks, Wetson et al., 2015. arXiv:1502.05698.
6. The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations, Felix Hill, Antoine Bordes, Sumit Chopra, JasonWeston. 2015. arXiv:1511.02301
7. Teaching Machines to Read and Comprehend, Hermann et al., 2015. arXiv:1506.03340
8. Ask Me Anything { Dynamic Memory Networks for Natural Language Processing, Kumar et al., 2015. arXiv:1506.07285
9. Neural turing machines, Graves, Alex, Greg Wayne, and Ivo Danihelka, 2014. arXiv:1410.5401