End-to-End Memory Networks on bAbI Dataset

Syed Muhammad Hasan Rizvi
hasan@hawk.iit.edu
A20374805

Ekrem Guzelyel
eguzelyel@hawk.iit.edu
A20384767

Abstract

Attention has gained a lot of popularity due to its performance on text and image data. It is able to extract only the words or pixels that are related to the task. Applying this idea to memory has a great potential and variable use cases. In this report we will inquire how the original author of the paper "End-to-End Memory Networks" (Sukhbaatar, et al.) implemented memory; and we will mention different approaches for the model. We will explain the dataset we used for the tasks; then we will compare how our results stand out against the paper's result.

1. Introduction

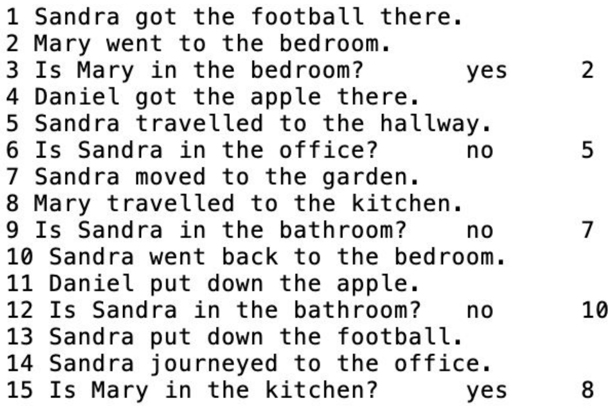
Memory networks is the extension of the concept of attention as being used in terms of memory. In the end presumably the model 'attends' only the historic events that are relative to the problem. In the paper "Memory Networks" Watson's team introduces the term memory network by combining inference components with long-term memory. The same team puts the network end-to-end afterwards, and finds out that with the increased computation power, the results are improved.

In this report, we try to reproduce similar results as the author by following his methods. We try his different approaches and note them down. In section 2, we explain the problem and the dataset. Then, we explain different approaches in section 3. Section 4 is about the detailed end-to-end memory network model. In section 5, we publish our results comparing the original paper; and finally conclude in section 6.

2. Problem and Dataset

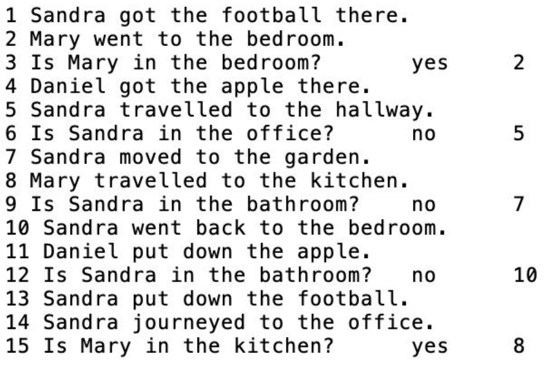
Building models that can make multiple computational steps to answer questions, and models that can hold long term dependency matrices in sequential data are big challenges in state-of-the-art artificial intelligence. The paper uses the memory network on text data to process question-answering (QA) tasks. The tasks are provided by Facebook Research, and are "organized towards the goal of automatic text understanding and reasoning."(Weston, et al.) The 20 tasks promote different aspect of understanding. Some example tasks are, starting from the simpler tasks, single supporting fact (Figure 1), yes/no question (Figure 2), basic induction (Figure 3), and size reasoning (Figure 4).

Figure 1: Single Supporting Fact



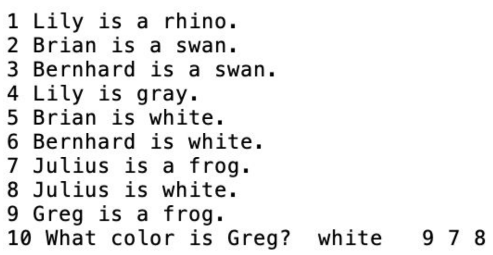
```
1 Sandra got the football there.
2 Mary went to the bedroom.
3 Is Mary in the bedroom?       yes     2
4 Daniel got the apple there.
5 Sandra travelled to the hallway.
6 Is Sandra in the office?      no      5
7 Sandra moved to the garden.
8 Mary travelled to the kitchen.
9 Is Sandra in the bathroom?    no      7
10 Sandra went back to the bedroom.
11 Daniel put down the apple.
12 Is Sandra in the bathroom?   no      10
13 Sandra put down the football.
14 Sandra journeyed to the office.
15 Is Mary in the kitchen?      yes     8
```

Figure 2: Yes/No Question



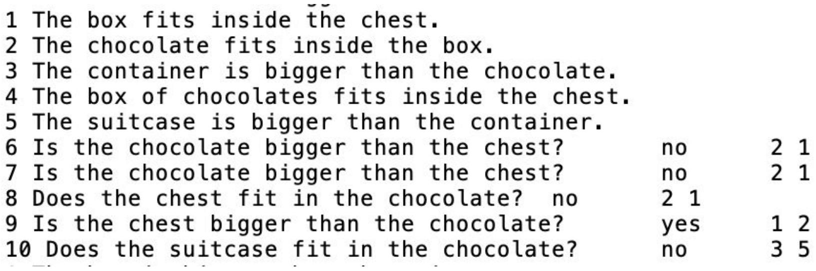
```
1 Sandra got the football there.
2 Mary went to the bedroom.
3 Is Mary in the bedroom?       yes     2
4 Daniel got the apple there.
5 Sandra travelled to the hallway.
6 Is Sandra in the office?      no      5
7 Sandra moved to the garden.
8 Mary travelled to the kitchen.
9 Is Sandra in the bathroom?    no      7
10 Sandra went back to the bedroom.
11 Daniel put down the apple.
12 Is Sandra in the bathroom?   no      10
13 Sandra put down the football.
14 Sandra journeyed to the office.
15 Is Mary in the kitchen?      yes     8
```

Figure 3: Basic Induction



```
1 Lily is a rhino.
2 Brian is a swan.
3 Bernhard is a swan.
4 Lily is gray.
5 Brian is white.
6 Bernhard is white.
7 Julius is a frog.
8 Julius is white.
9 Greg is a frog.
10 What color is Greg?  white   9 7 8
```

Figure 4: Size Reasoning



```
1 The box fits inside the chest.
2 The chocolate fits inside the box.
3 The container is bigger than the chocolate.
4 The box of chocolates fits inside the chest.
5 The suitcase is bigger than the container.
6 Is the chocolate bigger than the chest?       no      2 1
7 Is the chocolate bigger than the chest?       no      2 1
8 Does the chest fit in the chocolate?  no      2 1
9 Is the chest bigger than the chocolate?       yes     1 2
10 Does the suitcase fit in the chocolate?      no      3 5
```

3. Approaches

The paper introduces two approaches for the weight tying: Adjacent and layer-wise (RNN-like) model. In adjacent approach the output of the first input embedding layer is the input of the next embedding layer. Which in mathematical terms $A^{k+1}=C^{k}$. Although presumably this approach would give better results, it increases the amount it takes to train the model significantly, which is not feasible for our task. The second approach is layer-wise embedding, meaning that all embedding layers set to be the same. Mathematically $A^{1}=A^{2}=...=A^{k}$, and $C^{1}=C^{2}=...=C^{k}$. In order to update u between hops a linear mapping, $\boldsymbol{H}$, is added. $u^{k+1}=\boldsymbol{H}u^{k}+o^{k}$.

Also, adding and LSTM layer in the end, instead of multiplying with a W matrix is suggested by Keras team. In order to have this one compatible with the structure of the model, a hidden layer should be added.

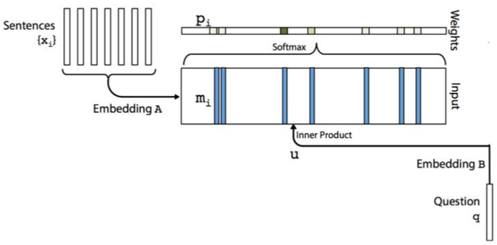
4. Detailed Model

The model that is introduced in the paper has five steps. In the first step we embed both question (embedding layer B) and input sequences (embedding layer A). The layer A has the dimension of (dxV), V being the input sequence max_length, and d being the custom given embedding dimension.

$$m_i = \text{Embedding}(x_i) = A$$

Figure 5: Input, Question Embedding

$$p_i = softmax(u^T m_i).$$



Layer B, the question embedding has a similar structure with (dxV) dimension, but this time V representing the question max_length.

$$u = \text{Embedding}(q) = B$$

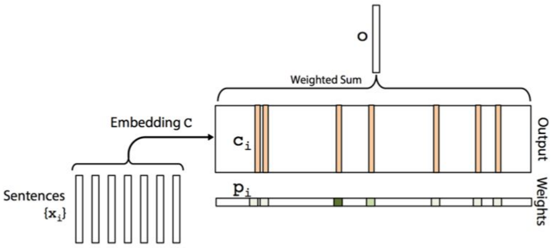
As seen in the Figure 5, then, attention layer is applied, where the paper calls it as the match matrix. It is computed by taking inner product of the output of layer B, u, and layer A; and taking the softmax of the answer.

$$p_i = \mathit{softmax}(u^T m_i)$$

To get the "response" from the input side a similar embedding is applied to input; and match (p) and embedding C is added. This time the dimension of this layer is defined in the paper as (input_lenth, query_length), however this approach makes the response (o) of the input side and output (u) of the question side mismatch. We solve this issue by flattening the layer. Another approach is to add another layer here to get the dimensions back again.

Figure 6: Match

$$o = \sum_i p_i c_i.$$

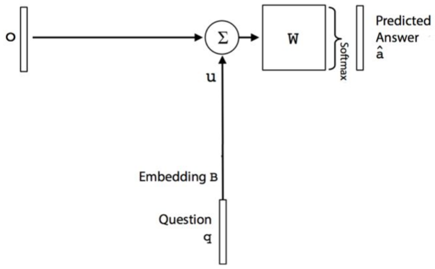


$$c_i = \text{Embedding}(x_i) = C$$
$$o = \Sigma(p_i c_i)$$

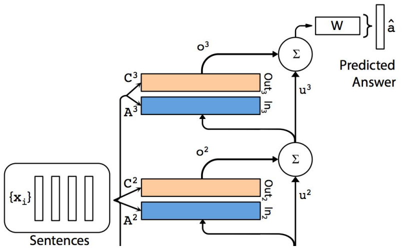
Finally, the result from o and u is multiplied with a W matrix and taken softmax to find the answer, â.

$$\hat{a} = \mathit{softmax}(W(o+u))$$

$$\hat{a} = softmax(W(o + u))$$



However, this explained approach is the only Memory Network part. In order to make this end-to-end, the paper suggests adding them back to back, which Sukhbaatar's team call it as "hops."
As suggested in the Section 3, two different computation can be made in order to get the output from individual hops. We used RNN-like (layer-wise) approach, since it was not as time consuming as the adjacent approach, and it gave a good result. We



tried several number of hops, three being the most dominantly used one. See Figure 8.

Also, as mentioned above, implementing the hidden layer approach instead of the flatten layer in the end, we use a Dense layer that has the same number of nodes as the desired output dimension of query_len*embedding_dim. After permuting it back to (embedding_dim, query_len), we put it into an LSTM layer.

5. Results

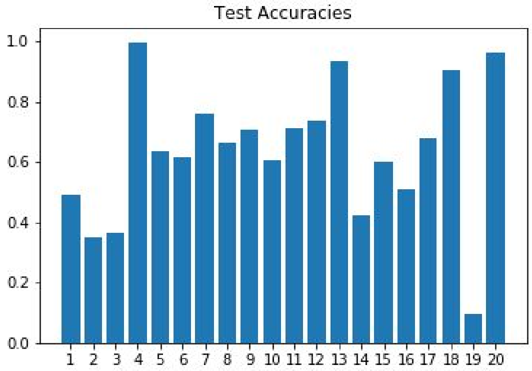
Wide variety of models we tried gave wide variety of results. Implementing the MemNN model only gave similar results with the paper for the single supporting fact, however not good enough in the other tasks. We got 94% accuracy in task 1, while the paper suggests 98%. For the rest of the tasks we got the half of the accuracies that the paper claimed they did.

Though, when we look at the MemN2N model, which has the concept of hops, we see a close correlation. Figure 8 and 9 (see appendix) are the loss and accuracy charts that we got from 20 task.

Some select comparison with the original paper is tasks numbered 1,6,16,18, and 19. In task 1, single supporting fact, the paper claimed that they got 99% accuracy with 3 hops, while we got 64% with the same structure. One thing to note here is that we only have the capacity to use bag-of-words, while the paper looks at the problem from different angles. This is why the chart that's provided by the paper has many columns. In task 6, counting, we get 75% accuracy, and the paper got 82%. In task 16, basic induction, we have 45% accuracy, while the paper gets similar results with 1,2 hops, but an enormously good result of 96% with 3 hops. In task 18, size reasoning, we get a really good result of 92%, whereas the paper got 91%. The task number

19, path finding, is the worst performed model in both of our implementations, which is no more than 10%

Finally, the paper also comments on how the models behave when they use all the tasks jointly. If we look at our results, we again, see a good correlation in the accuracy matrix.



6. Conclusion and Future Work

Overall, the results we got are not far away from the suggested results from the paper. Memory Networks, and using them End-to-End has a good potential of being useful in many cases with question-answering (QA). Some things to improve the model would be tuning the number of nodes for embedding layers; adding new hidden layers; and using adjacent layer approach. A computer with better computational power would also give the flexibility to increase number of parameters, which may result in a better accuracy.

Appendice

Figure 10: Losses of Individual Tasks

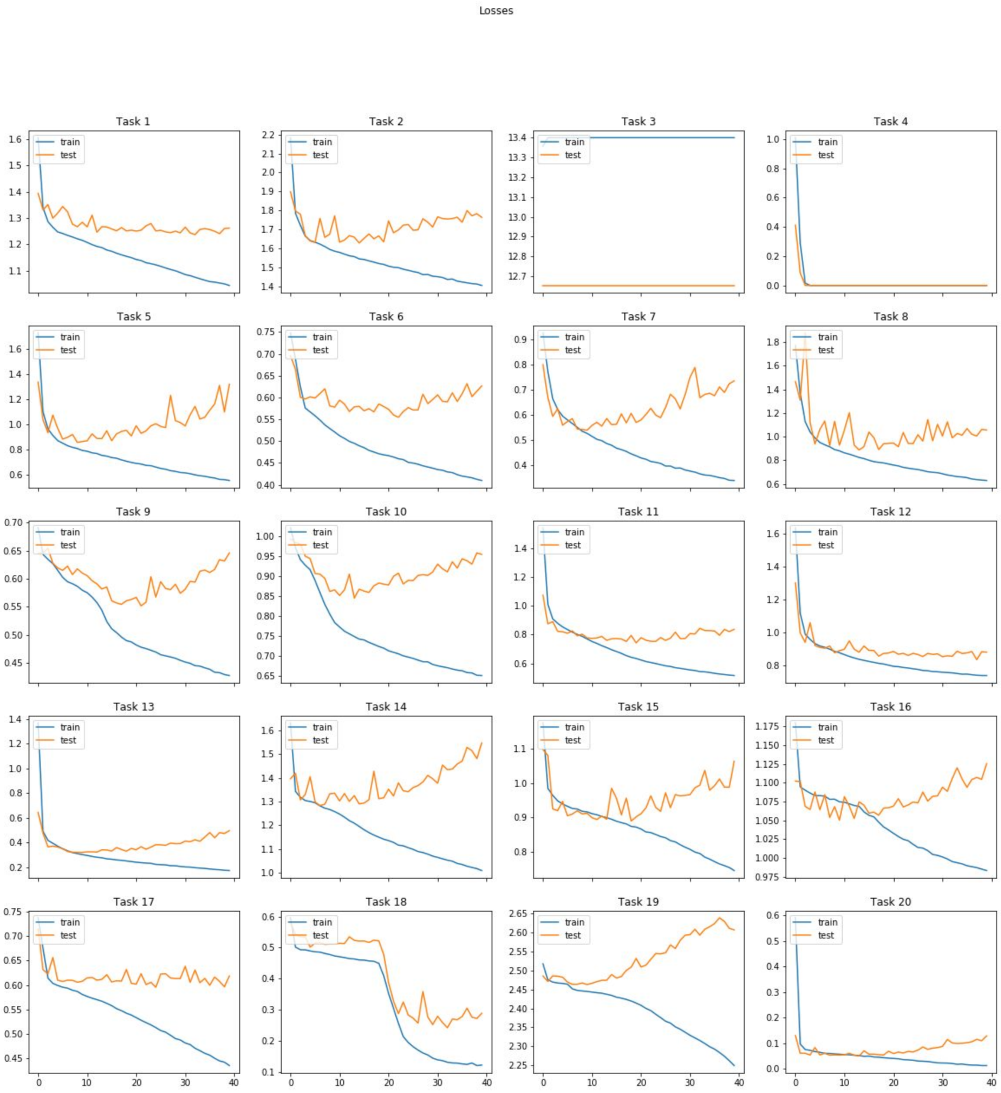
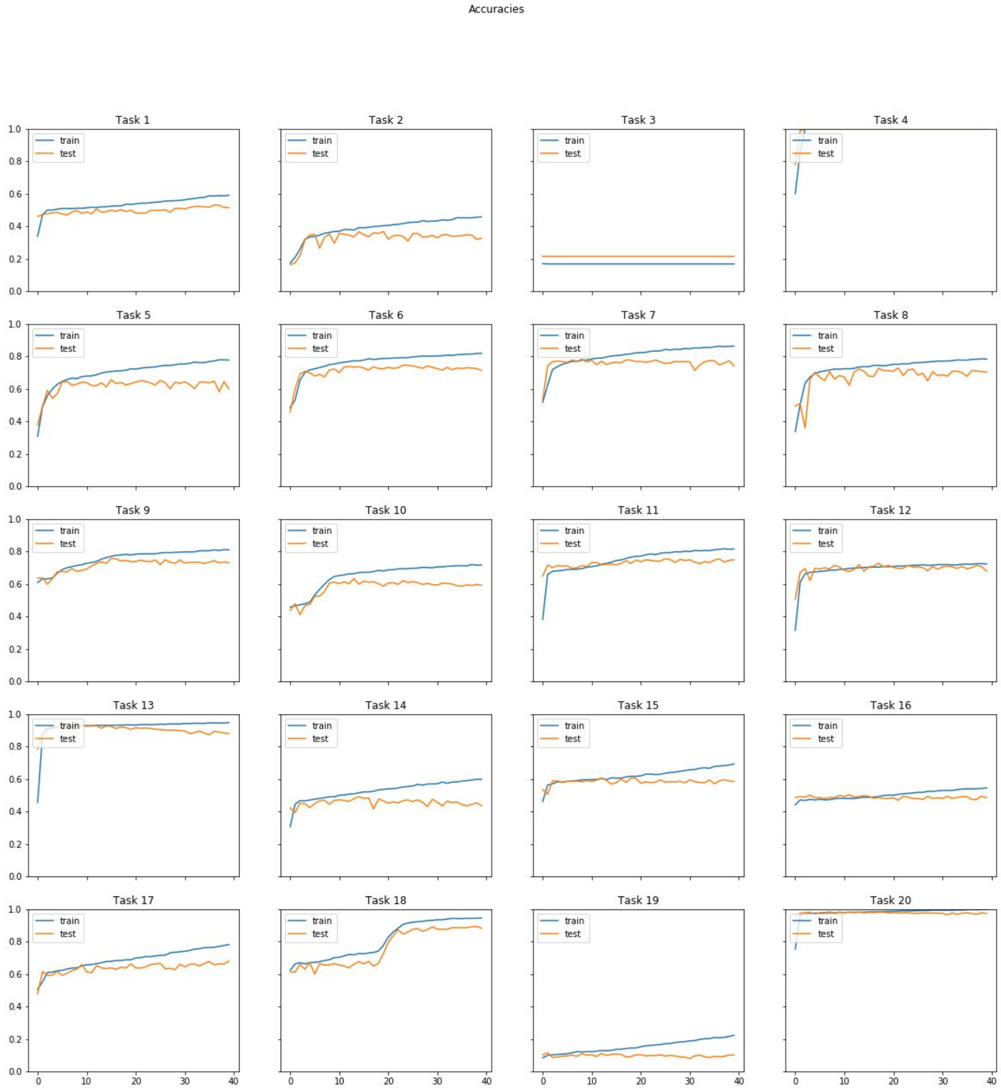


Figure 11: Accuracies of Individual Tasks



References

Hui, J. Memory Network. *Jonathan hui blog.*Retrieved from jhui.github.io/2017/03/15/Memory-network/

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart van Merriënboer, Armand Joulin and Tomas Mikolov. Towards AI Complete Question Answering: A Set of Prerequisite Toy Tasks, *arXiv*:1502.05698.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In Advances in neural information processing systems, pages 2440—2448.

Weng, L. Attention? Attention! *Lil-log*. Retrieved from lilianweng.github.io/lil-log/2018/06/24/attention-attention.html

Weston, Jason, Chopra, Sumit, and Bordes, Antoine. Memory networks. *CoRR*, abs/1410.3916, 2014.