Experiments with Dynamic Memory Network Models for QA tasks

Apeksha Singhal(112045146)[1], **Gagandeep Kaur(#112101150)**[2]
{apeksha.singhal, gagandeep.kaur.3}@stonybrook.edu

Abstract

Experiment with DMN model to handle Squad 1.1 and FB bAbI Q&A tasks and compare their performance with an understanding of reasons behind the difference in performance.

1 Introduction

Q&A is a topic of broad research in NLP. The problem, while it seems straightforward, exacts good reasoning and understanding of the inherent relationships between different components of the text. Several models emerge every day from various research teams that try to do better on QA tasks, yet, no universal approach/model has come about that can handle all types of Q&A seamlessly.
Q&A is integral part of human speech in general and it's the primal way of testing knowledge comprehension. So, this is also a way to test the machine's understanding of the text we're feeding it. Hence, basic Q&A is a very practical task that we'd want to integrate into an NLP model. The applications of this are many folds. e.g. In customer care, in basic fact retrieval from a document etc.
Main challenges posed by QA tasks:

1. Scalability : The time taken in the process to find answer to a single question can vary significantly as it might require going through a large text database.

2. Representation: The model needs to understand the relationship between different representations of expressions in question and paraphrase. e.g. A question that asks for a relationship between 2 people should be able to draw similarities with the one that asks about an act that defines that relationship.

3. Reasoning : The model has to be able to combine pieces of information together. As the number of relevant pieces grows, the search space grows exponentially making it a hard search problem. Since reasoning is closely coupled with how the text meaning is stored, an optimal representation should be learned end-to-end in the process of knowledge storing and reasoning.

2 Literature Review

There have been multiple attempts to solve the Q&A tasks. While the architecture and efficiency of these approaches vary according to the structure of the problem at hand, we are listing down the milestone models that have been written for generic QA handling:

1. LSTM , (Sutskever et al., 2014) : They work by reading the story until the point they reach a question and then have to output an answer

2. MemNNs (Weston et al., 2014) : They work by a controller neural network performing inference over the stored memories that consist of the previous statements in the story. The original proposed model performs 2 hops of inference: finding the first supporting fact with the maximum match score with the question, and then the second supporting fact with the maximum match score with both the question and the first fact that was found.

3. End To End Memory Networks The objective of the paper is introduce a novel RNN with a large external memory i.e. a memory from where the RNN reads multiple times before giving output. Lots of experiment were being conducted with Memory Networks at the time. This paper acted upon the idea of having the need to train the memory networks

with strong supervision and their application to general tasks.

4. Dynamic Memory Networks (Kumar et al., 2015): The DMN first computes a representation for all inputs and the question. The question representation then triggers an iterative attention process that searches the inputs and retrieves relevant facts. The DMN memory module then reasons over retrieved facts and provides a vector representation of all relevant information to an answer module which generates the answer.

5. N-Gram Classifiers , Richardson et al. (2013) [1]: construct a bag-of-N-grams for all sentences in the story that share at least one word with the question, and then learn a linear classifier to predict the answer using those features.

MNN and DMNN are the state of the art on bAbi dataset. However, they both perform poorly on a couple of tasks where they have to deal with the positions and directions in the data. This might be due to the fact that the tasks that need advanced forms of induction and deduction require a general search algorithm to be built into the inference procedure which memory networks lack.

Also, through our search we haven't found an experiment of DMN on squad dataset which is much closer to real world comprehension and Q&A problem, basically a generic DMN that works on multiple Q&A datasets. We're planning to run a bunch of experiments on FB bAbi and Squad 1.1 datasets to come up with a smart enough model to handle both with reasonable accuracy. We'll test Squad with DMN, making a generic model and infrastructure to seamlessly train and test both the datasets.We'll make some architecture changes in our starter model to accommodate both datasets and achieve a good accuracy.
Human comprehension works on retaining facts and knowledge in order to answer questions and this is a universal thing. So, the idea is that the memory networks should work irrespective of the type of task given that we can achieve a correct input and output mechanism. We will be running our model on both SQUAD and bABI datests.

The evaluation measure for bABI QA success will be accuracy.

Task
Two Supporting Facts
Three supporting facts
Three Argument Relations
Yes/No Questions
Counting
Lists/Sets
Simple negation
Basic coreference
Time reasoning
Basic induction
Positional reasoning
Size Reasoning
Path finding

Table 1: bABI Task Description

3 DataSets

3.1 bABI dataset

The bABI dataset is a synthetic dataset which is used to evaluate a model on a set of 20 different types of tasks relating to different fields like inference, reasoning, counting, deduction etc. The tasks are defined in Table 1. An Example:

```
1 Mary moved to the bathroom.
2 John went to the hallway.
3 Where is Mary?        bathroom        1
4 Daniel went back to the hallway.
5 Sandra moved to the garden.
6 Where is Daniel?      hallway 4
7 John moved to the office.
8 Sandra journeyed to the bathroom.
9 Where is Daniel?      hallway 4
10 Mary moved to the hallway.
11 Daniel travelled to the office.
12 Where is Daniel?     office  11
13 John went back to the garden.
14 John moved to the bedroom.
15 Where is Sandra?     bathroom        8
1 Sandra travelled to the office.
2 Sandra went to the bathroom.
3 Where is Sandra?      bathroom        2
```

Figure 1: Example bAbi dataset

3.2 SQUAD1.1

Squad is a QA dataset prepared on a set of Wikipedia articles where the answer to every question is a segment of text, or span, from the corresponding reading passage, or the question might be unanswerable. This dataset has two versions 1.1 and 2.0. We will be covering our experiments on SQUAD 1.1 . It has 10,000+ QA pairs derived from 500+ articles. The evaluation measured for SQUAD datasets are : Exact Match and F1 score. The state-of the art accuracy of bABI is 100% for

majority of the tasks and best F1 score for SQUAD is 91.121. If our model gives an accuracy between 91 to 100 for bABI dataset and an F1 score of 80-90, we can safely assume our model to be a good model and bad otherwise.

4 Baseline Architecture

We will first explain the architecture of Dynamic Memory Network and the Neural Reasoner and then detail the modifications done to perform the experiments.

4.1 Dynamic Memory Network

The network consists of four modules namely: Input, Question, Memory and Answer module.

- **Input Module**: The input module runs a GRU over the sequence of the words and encodes the input.
- **Question Module**: This module runs the GRU as in the input module and provides an encoding over the questions.
- **Memory Module**: This module takes the facts and the questions extracted from the input and the question module. It is composed of two nested GRUs. These GRUs goes over all the facts and uses an attention function to score the facts in order to create an episode. The output from first GRU is passed to the outer GRU which again generates episodes from this input and the attention function. The final state of this module is the passed to the answer module.
- **Answer Module**: It takes the last memory as input and generates an answer by doing a softmax over the memory.

Stochastic Gradient Descent is then applied on the whole system to find the loss and improve over it.

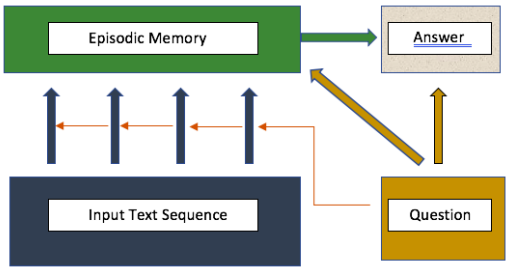

Figure 2: DMN Architecture

4.2 Neural Reasoner

This architecture was tested only on the tasks that do not do well on bAbi dataset e.g. the positional reasoning and inference tasks. It applied multiple hops of DNN on the inputs and the questions and tries to capture the interactions between the question and the facts. Based on these interactions, the representation of the question is modified. This modified representation is passed on to the next hop. Even though this was tested on only two tasks of bAbi dataset, it improved over the state of art results on these tasks.

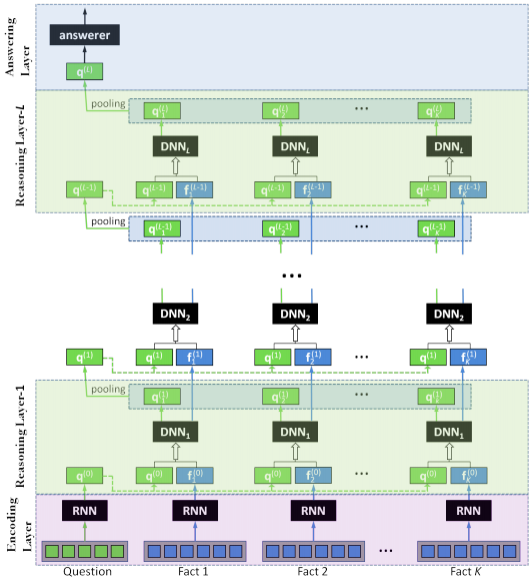

Figure 3: Neural Reasoner

- **Encoding Layer:** This layers encodes the input facts and questions to semantic representations. RNN is employed to provide representations. The equations are as follows:

$$z_t = \sigma(W_{xz}E_{x_t} + W_{hz}h_{t-1})$$

$$r_t = \sigma(W_{xr}E_{x_t} + W_{hr}h_{t-1})$$

$$\hat{h_t} = tanh(W_{xh}E_{x_t} + U_{hh}(r_t \cdot h_{t-1}))$$

$$h_t = (1 - z_t) \cdot h_{t-1} + z_t \cdot \hat{h_t}$$

- **Reasoning Layer** : The reasoning layer performs a question fact interaction. Ater the interaction, the representation of the question is updated by max pooling and is then an answering later.

Question Fact Interaction: The representation of a question at layer l is updated after interaction with each fact k. Therefore,

$$q_k^{(l)} = \sigma(W_l^T[(q^(l-1))^T, (f_k^{(l-1)})^T] + b_l)$$

Pooling: Pooling fuses the updated representation of the question over all the facts.
Answering Layer: After reaching the last reasoning layer, the representation is sent to a softmax to predict the final answer.

$$y = softmax(W_{softmax}^T q^L)$$

5 Experiments

For our experiments, we picked the starter code available on github.[9] This code is implemented in theano and does a pretty good job on accurately handling most of the task types (Table 2).However, there's "positional Reasoning" and "Path Finding" tasks where it doesn't do that well.

5.1 bAbifying Squad dataset

5.1.1 Motivation

We wanted to be able to test DMN model on squad dataset and see if it can perform fairly there. If not, why . The main challenge before running that experiment was to reduce squad data set in a similar structure like bAbi.

5.1.2 Details

Squad data set is vastly different from the bAbi data set in that,

- it comes from a different, most complex and vast distribution, wikipedia.
- It is in a less structured format that bAbi.
- each fact sequence has multiple questions corresponding to it.
- Answer is a sentence, not a single word like the bAbi dataset.

We implemented a parser to perform above transformations. In addition to that, we did some manual modification at conflicting parts.

5.1.3 Results

The following picture represents the modified SQUAD dataset in the babified form. Table 2 captures the result of evaluation.

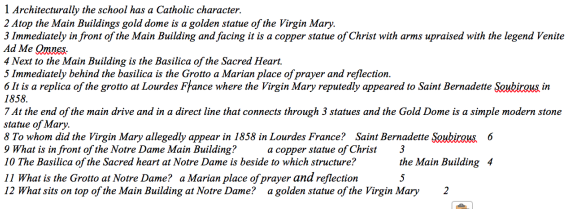
1 Architecturally the school has a Catholic character.
2 Atop the Main Buildings gold dome is a golden statue of the Virgin Mary.
3 Immediately in front of the Main Building and facing it is a copper statue of Christ with arms upraised with the legend Venite Ad Me Omnes.
4 Next to the Main Building is the Basilica of the Sacred Heart.
5 Immediately behind the basilica is the Grotto a Marian place of prayer and reflection.
6 It is a replica of the grotto at Lourdes Fﬂance where the Virgin Mary reputedly appeared to Saint Bernadette Soubirous in 1858.
7 At the end of the main drive and in a direct line that connects through 3 statues and the Gold Dome is a simple modern stone statue of Mary.
8 To whom did the Virgin Mary allegedly appear in 1858 in Lourdes France? Saint Bernadette Soubirous 6
9 What is in front of the Notre Dame Main Building? a copper statue of Christ 3
10 The Basilica of the Sacred heart at Notre Dame is beside to which structure? the Main Building 4
11 What is the Grotto at Notre Dame? a Marian place of prayer and reflection 5
12 What sits on top of the Main Building at Notre Dame? a golden statue of the Virgin Mary 2

Figure 4: bAbified squad

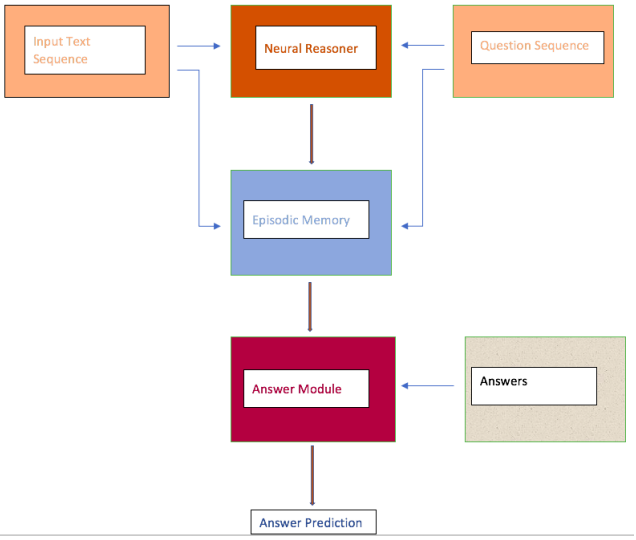

Figure 5: Neural reasoner and DMN fusion architecture

5.2 Fusing Neural reasoner into DMN architecture

5.2.1 Motivation

Neural reasoner performs quite well on tasks where DMN fails i.e. 17&19. We tried to integrate it's "reasoning" part with our DMN architecture and come up with a fusion model that does well on all tasks.

5.2.2 Details

We integrated the neural reasoning into DMN by adding code to include neural network layers between input fusion module and episodic memory module.The modified architecture is depicted in Figure 4.
We performed multiplication of facts encodings and question encodings in order to capture the interaction between them. These went through NN and max pooling to result into an enhanced question form that captures relationship between question and facts. For our experiments we used 2 layers of reasoning neural networks.

Task	DMN	Fusion Model
Task 1	100	45.70
Task 17	**49.50 - 58**	**61**

Table 2: Evaluation Accuracy on bABI with DMN + DNN

5.2.3 Results and Analysis

We ran 50 epochs of training and testing to evaluate our new model on task 17 where the DMN base line was 58%. Our model gives around 3-4% more accuracy than this on many occasions but this doesn't remain consistent.

We suspect that this performance can be improved by performing auxiliary tasks mentioned in the original neural reasoner paper.[10] We tried some additional modifications in the neural reasoner module to improve it such as, trying Relu activation, capturing interactions by using addition instead of multiplication, updating common weights for NN layers etc. **Best accuracy was achieved by the original experiment although with a slight margin.**

The accuracy for Task1 decreased with the combination. The neural reasoner performed well for the two reasoning tasks because the answer for those tasks either a Yes/No or from a limited set of words and thus a softmax over the last question representation was able to give good results. But for the other tasks on mixing with DMN, the new representation of question might be a distortion as the answer is not from a limited set of numbers.

5.3 Generating sentence answers through DMN

5.3.1 Motivation

Squad data set expects answers in variable word lengths. Original DMN is designed to give just single word answers owing to the structure of bAbi data sets. In order to try squad on DMN, we modified the architecture to accommodate multiple word answering.

5.3.2 Details

We tweaked the answer module to work with sentences. The changes included:

- Modifying answer module to use a recurrent NN so it can generate sequences as answers.
- modifying loss calculation and other subsequent steps to accommodate these changes.
- Evaluating SQUAD using these generated sequences.
- We sought out to consider modify the task 19 problem as a multi-word answer and instead of using one letter to represent a direction we gave one word. e.g. "south" instead of "s". Therefore the answer to the previous question will be "south east". We then conducted an experiment for this dataset with the Dynamic Memory network to generate multi-word sequences

5.3.3 Results

Our model is able to generate multiple words outputs. Even though we are now able to generate such sequences, but even after running a lot of epochs, the model is unable to learn and improve the responses and is thus not giving accurate predictions.

5.4 Testing DMN starter code with custom embeddings

5.4.1 Motivation

The aim behind this experiments is to see the effect of embedding size and scope on the model performance. Also, the squad data was not very clean and glove failed to retrieve a substantial part of the corpus which we felt affected the accuracy.

5.4.2 Details

In order to generate our custom embeddings, we used gensim libraries and saved the results. We fed this to our model, instead of precooked glove embeddings. For this experiment we tried:

- Custom trained embeddings for both squad and bAbi over default glove embeddings.
- Glove embeddings of different lengths.(50, 100, 200).

5.4.3 Results

Refer - Table 3 We observed almost same accuracy across all variation of embeddings except for a 2% improvement with 200 dimensional one. This can be attributed to the fact that the vocab for this particular data set is very limited, thus not commanding a rich set of features in embeddings. However, it should be vital for squad data set.

Embedding Variation	Accuracy(Task 17)
Glove-50	58
Glove-100	58
Glove-200	60
Custom Embeddings	58

Table 3: Experiment Results with word embeddings

6 Conclusion

Although we acheive some improved accuracy over DMN with neural reasoner integration, it hugely deteriorates the accuracy for other tasks. So, we think it's not really helpful to fuse the 2 architectures together. Neural reasoner is an exclusive model for Path finding and positional reference tasks.
Our experiments on squad data set are not yet fully finished. We would continue the work to make the DMN episodic memory module compatible with multiple word answers.

7 Code

The modifications over the baseline model for this project are located in this github repository: https://github.com/ghundal93/QA-Experiments-with-DMN

8 Future Work

- Currently, we've just modified the model to produce multiple word anwsers but the memory module is not really compatible with this change hence it's not generating comparable predictions. We'll endeavour to improve it and get some decent results for squad dataset.
- Perform more experiments with squad tasks.
- Test DMN on squad with BERT embeddings.[11]

References

[1] Richardson, Matthew, Burges, Christopher JC, and Renshaw, Erin. Mctest: A challenge dataset for the open-domain machine comprehension of text. In EMNLP, pp. 193203, 2013.

[2] Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart van Merrienboer, Armand Joulin Tomas Mikolov: TOWARDS AI-COMPLETE QUESTION ANSWERING :A SET OF PREREQUISITE TOY TASKS

[3] Caiming Xiong, Stephen Merity, Richard Socher:Dynamic Memory Networks for Visual and Textual Question Answering

[4] Sutskever, Ilya, Vinyals, Oriol, and Le, Quoc VV. Sequence to sequence learning with neural networks. In Advances in Neural Information Processing Systems, pp. 31043112, 2014.

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

[6] Kumar, Ankit, Irsoy, Ozan, Su, Jonathan, Bradbury, James, English, Robert, Pierce, Brian, Ondruska, Peter, Gulrajani, Ishaan, and Socher, Richard. Ask me anything: Dynamic memory networks for natural language processing. http://arxiv.org/abs/1506.07285, 2015.

[7] https://rajpurkar.github.io/SQuAD-explorer/explore/1.1/dev/

[8] https://research.fb.com/downloads/babi/

[9] Towards Neural Network-based Reasoning, Baolin Peng1, Zhengdong Lu2 Hang Li2, Kam-Fai Wong https://arxiv.org/pdf/1508.05508.pdf

[10] https://github.com/YerevaNN/Dynamic-memory-networks-in-Theano

[11] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova https://arxiv.org/abs/1810.04805