IMPROVEMENT OF MEMORY IN A NEURAL NETWORK

NEURAL REASONING MODEL FOR QUESTION ANSWERING

18-090

Preliminary Progress Review

(Preliminary Progress Review Documentation submitted in partial fulfilment of the requirement for the Degree of Bachelor of Science Special (Honours) In Information Technology)

H.J.D.A. Karunaratne (IT15047748)

Bachelor of Science (Honours) in Information Technology
(Specialization in Software Engineering)

Department of Information Technology

Sri Lanka Institute of Information Technology

May 2018

TABLE OF CONTENT

1.0. INTRODUCTION

1.1. Purpose

The purpose of this document is to provide the preliminary progress of the “Improvement of Memory” component of “Neural Reasoning Model for Question Answering”.

This document is primarily intended to be proposed to the audience who has already referred to the proposal of “Neural Reasoning Model for Question Answering” and also to anyone who has a knowledge and interest on deep learning. Further, this document will be used as a reference for the progress of the component.

1.2. Scope

The document will illustrate the purpose and the main areas to be focused throughout the component. The literature review section will demonstrate the research gap that this project is attempting to fulfill by analyzing the existing solutions in the QA domain, revealing the advantages and disadvantages of those systems, and how the proposed system can improve on those failings. Furthermore, the research objectives and the proposed methodology will be discussed. Latter part of the document will contain the data sources, anticipated benefits and the anticipated deliverables for this project.

1.3. Overview

The proposed system aims to achieve a remarkable improvement in the Reasoning, Attention and Memory(RAM) Model in Deep Neural Networks, thus achieving effective and accurate answers in the Question Answering system with increased performance. The main goal of the memory component would be to focus on proper control and management of memory and persistence, in the deep neural network.

2. STATEMENT OF WORK

2.1. Background Information and Overview of Previous Work based on Literature Survey

Automated QA systems have been gaining a lot of prominence since the early 60's. QA has mostly been used to develop intricate dialogue systems such as chat-bots and other systems that mimic human interaction [1]. The methodologies used in these systems vary from Information retrieval based statistical approaches to machine learning approaches and then to deep learning based evolving systems.

Traditionally, most of these systems use the tried methods of parsing, part-of-speech tagging, etc that come from the domain of NLP research. While there is absolutely nothing wrong with these techniques, they do have their limitations. [1] W.A. Woods et al. shows how we can use NLP as a front end for extracting information from a given query and then translate that into a logical query which can then then be converted into a database query language that can be passed into the underlying database management system. In addition to that there needs to be a lexicon that functions as an admissible vocabulary of the knowledge base so that it is possible to filter out unnecessary terminology. The knowledge base is processed to an ontology that breaks it down into classes, relations and functions [2]. Natural Language Database Interfaces (NLDBIS) are database systems that allow users to access stored data using natural language requests. Some popular commercial systems are IBM's LanguageAccess and Q&A from Symantec [3].

Information retrieval (IR) is another technique that has been used to address the problem of QA. With IR systems pay attention to the organisation, representation and storage of information artifacts such that when a user makes a query the system is able to to return a document or a collection of artifacts that relate to the query [4]. Recent advances in OCR and other text scanning techniques have meant that it is possible to retrieve passages of text rather than entire documents. However IR is still widely seen as from the document retrieval domain rather than from the QA domain.

Template based question answering is another technique that has been used for QA and is currently being used by the START system which has answered over a million questions since

1993 [5]. START uses natural language annotations to match questions to candidate answers. An annotation will have the structure of 'subject-relationship-object' and when a user asks a question, the question will be matched to all the available annotation entries at the word level (using synonyms, IS-A, etc) and the structure level. When a successful match is found, the annotation will point to an information segment which will be returned as the answer. When new information resources are incorporated into the SMART system, the natural language annotations have to be composed manually [6]. START uses Omnibase as the underlying database system to store information and when the annotation match is found, the database query must be used to retrieve the information. While this system has been relatively successful, it requires a lot of preprocessing which must be done manually.

Deep learning is the state-of-the-art in areas such as speech recognition, natural language understanding, visual object recognition, etc. Convolutional neural networks have brought many breakthroughs in areas such as processing images, pictures and speech, whereas Recurrent networks have been extremely successful in areas such as processing text and speech.

Recurrent Neural Networks are special because of its ability to take the previously perceived information into consideration. While its counterpart Feed-forward neural networks are only concerned with the state of the inputs at a given time, RNN takes the past decisions into consideration. The decision a recurrent net reached at time step t-1 affects the decision it will reach one moment later at time step t. This is because of the feedback loop that enables RNNs to ingest their own outputs moment after moment as input retaining memory in the sequence itself.

The sequential information in an RNN can be preserved in the hidden state of the network and it can be carried forward several steps so that it would affect processing new inputs. This enables RNN to find co-relations between inputs that are separated by many moments. As long as the memory of a recurrent net can persist the past decisions of would influence the current decision making.

One of the major fallbacks of the traditional RNNs is the vanishing gradient problem. These RNNs were not able to form connections between the final outputs and the events that happened

several steps before. Because of this the gradient for that particular period cannot be calculated which resulted in the “vanishing gradient” for that period in the graph. One reason for the vanishing gradient is that the information passing through the network passes through several stages of multiplication. If we were to explain this problem using basic mathematics, if a value is to be multiplied by another value more than 1(even slightly) continuously it can become immeasurably large. In the same way if the multiple is less than 1, the value tends to do the inverse. The first example causes Exploding Gradients but this can be solved easily by truncating. But the vanishing gradient caused by the second example is harder to solve because it can become too small for computers to work with.

To solve this problem a variation of RNN came up with Long Short-Term Memory units, or LSTMs. LSTMs preserve a constant error that can be that can be back-propagated through many time and layers, sometimes as many as 1000. This opens up channels to link cause and effect remotely. LSTM achieves this by storing the information outside the normal flow in a gated cell. The cell makes decisions about what to store, and when to allow reads, writes and erasures, via gates that open and close. These are analog gates implemented with element-wise multiplication of sigmoids. Being analog makes the gates differentiable which is an advantage in backpropagation.

RNNs have been able to make a significant improvement in contrast to other machine learning techniques due to their ability to learn and carry out complicated transformations of data its ability maintaining long term as well as short term dependencies. RNN contains an interplay of Reasoning, Attention and Memory, commonly referred to as the ‘RAM model’ in Deep Neural Networks.

2.2. Identification and Significance of the Problem

Question Answering(QA) is one of the research areas that have gained a lot of interest over time. Out of the various technologies used in QA recently, deep learning stands to be the most prominent technology among the prevailing technologies. This research is addressing the shortcomings of performance and accuracy of QA, by improving and supporting the RAM model. The idea is to build a Question Answering system using Deep learning technique by

Improving on the existing state-of-the-art leveraging optimized Deep Learning techniques by combining and improving Reasoning, Attention and Memory mechanisms to assess the system in a real context.

The significance of the research is that we will be improving question preprocessing, Reasoning, Attention and Memory of our deep neural network which are vital and technically challenging tasks to carry out.

Improvement of question preprocessing: Provide the neural network with the most effective vector representation.

Improvement in Reasoning: Improving reasoning by making it capable to infer over multiple facts in a way insensitive to the number of supporting facts and data.

Improvement in Attention: Improving focus on relevant parts of the input more than the irrelevant parts and improvement in finding correspondence between source and target words.

Improvement in Memory: Improvement in the ability of the neural network to retain the most important data inputted and identify the threshold level to retain memory.

2.3. Technical objectives

According to the explanations given in previous parts “NEURAL REASONING MODEL FOR QUESTION ANSWERING” project is based on DNN (Deep Neural Network) where the preprocessing, attention, reasoning and memory layers of dnn are enriched to produce better results than existing state-of-art techniques. This part of the report will demonstrate the Software and Hardware requirements of the research which will be used for performing technical tasks in order to improve the above mentioned for components.

2.3.1 Specific software requirements

The most renowned and accepted programming language for the machine learning problems is Python. Python is widely used across the world for solving machine learning problems for its readability and wide range of machine learning libraries. Our research project will be using Python as the main programming language.

TensorFlow is an open source artificial intelligence library, using data flow graphs to build models. It allows developers to create large-scale neural networks with many layers and TensorFlow allows distribution of computation across different computers, as well as multiple CPUs and GPUs within a single machine. TensorFlow provides a Python API so we will be using TensorFlow for our deep learning research.

Keras is a high-level neural network API, helping lead the way to the commoditization of deep learning and artificial intelligence. It runs on top of a number of lower-level libraries, used as backend, including TensorFlow and Theano. Keras code is portable, meaning that you can implement a neural network in Keras using Theano as a backend and then specify the backend to subsequently run on TensorFlow, and no further changes would be required to your code. Theano is a Python library for fast numerical computation that can be run on the CPU or GPU. It is a key foundational library for Deep Learning in Python that you can use directly to create Deep Learning models or wrapper libraries that greatly simplify the process. And this research will be using Keras and Theano as key software resources.

2.4. Detail design

Deep learning techniques have shown to outperform humans in computer vision, pattern recognition. [14][15] However, QA has not shown similar achievements or human level performance through deep learning. This shows that deep learning lacks reasoning, attention and memory[7]. The improvements in these areas have led to very promising results [8][9][10] yet they are required to make progress on complex tasks.
The proposed system aims to address each of these areas in order to improve them, thus improving the performance of the Memory based Network.

In order to improve reasoning, the proposed framework would focus on Memory and Attention which are crucial for reasoning.

Memory is another important aspect in DNNs as the cutting edge Deep Learning models continue to push the limits of GPU RAM. This system aims to develop a desirable model to train more data consuming less memory. The dependencies between the operations in the deep network is represented using Computation Graphs. There are mainly two ways: performing back-propagation on the same graph or explicitly representing a backwards path to calculate the required gradients. In addition, mxnet[12] have introduced explicit backward path for gradient calculation. Also, methods such as In-place memory sharing [12], Standard memory sharing[12] are also proven to optimize memory. Furthermore, the proposed system will consider Static and Dynamic Memory Allocation methods and Parallelization to optimize memory.

Next aspect that affects reasoning would be Attention. Before Attention was introduced, translation relied on reading a complete sentence and compressing all information into a fixed-length vector. Thus, a sentence with hundreds of words represented by several words will surely lead to information loss, inadequate translation. Attention allows machine translator to look over all the information the original sentence holds, then generate the proper word according to current word it works on and the context. MAC network developed by Stanford approaches problems by decomposing them into a series of attention-based reasoning steps. [11]

2.5. Sources for test data & analysis

The amount of data that we create is growing at a staggering pace and keeping track of it becomes more difficult. Accordingly, it is important that we use techniques to make the information that we need accessible. A question answer system is concerned with automatically answering questions posed by humans naturally, which would return the most specific or direct answer rather than returning a set of documents relating like in search engines. There are two broad types of question answering systems that could be built upon. One is Closed-domain question answering system, which basically deals with questions under a specific domain (bAbi – a closed domain dataset provided by Facebook). The other type is Open domain question answering system, which is concerned with question answering about almost any kind of question posed, therefore, it is required to use the information provided in the dataset as well as any additional available knowledge(WIKIQA - an open domain dataset). However, for the purpose of the research, we have settled our center of interest towards a open domain question answering system. This does not necessary mean that it will answer every open domain question. We would focus on answering general questions with an enhanced Deep Neural Network than state-of-the-art systems.

We plan to build a QA system using deep learning to help interpret a question posed and provide a suitable answer from a given context. We will facilitate the QA system to be built under any domain, complying with certain conditions to increase performance. The precision and the weight of reliability of the provided answers will majorly depend upon the accuracy of the context. However the goal in this research is to emphasize that by using deep learning techniques by combining Reasoning, Attention and Memory mechanisms, we are able to reduce some of the complexities and barriers that are present at the moment and provide an innovative product that utilizes the platform.

We will be using the TriviaQA dataset to evaluate our model.

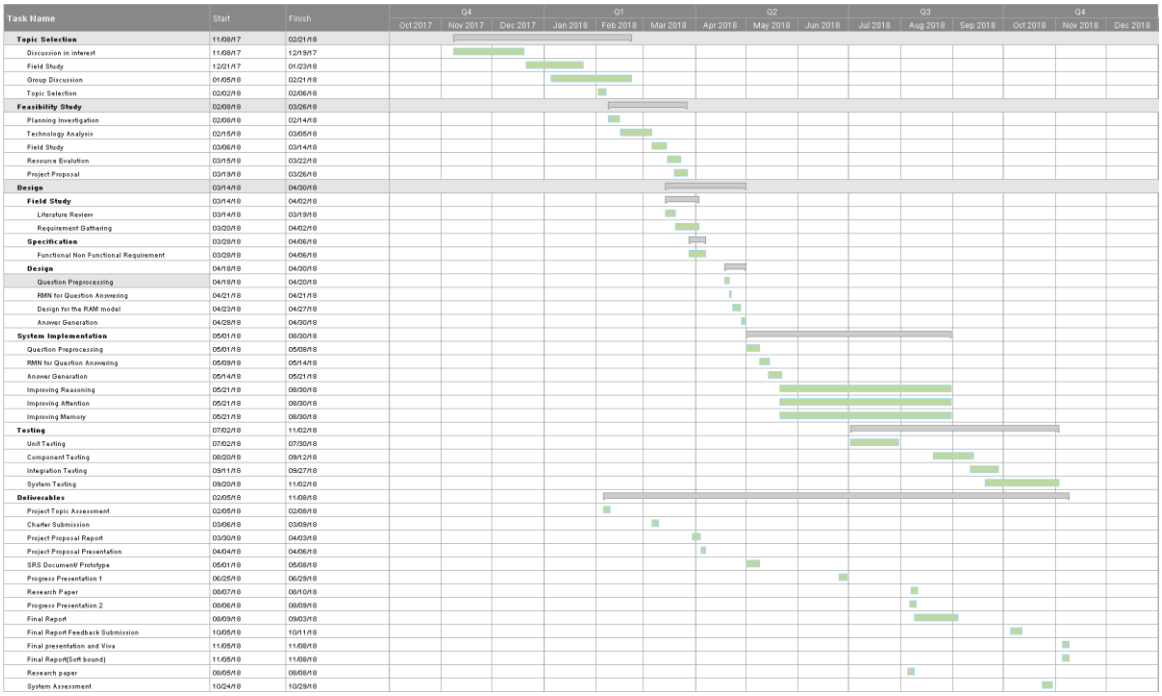
2.6. Anticipated benefits

As we demonstrated above the latest improvements in deep learning methodologies made significant impact in natural language QAS for producing human like answers for a given question. This lead to creation of many QAS in open domain and closed domain with their significant changes and specialties. Through the proposed model we tend to improve four major components of a dnn which are preprocessing, attention, reasoning and memory in order to produce a human like answers for a given question captivating state-of-art techniques. So that the user will be highly benefited with the answer than misleading with an incorrect answer. This will make user more enthusiastic to us our QAS system more and more.

Humans are capable of reasoning about entities and relationships intuitively. However, machines possess this capability up to some extent only. Machines would have to utilize the previously acquired knowledge to do so. This would allow machines to reason about their entities and relations from unstructured data. Solving this would open infinite possibilities for Artificial Intelligence. Through the reasoning the model will gain artificial intelligent which makes answers more human interactive.

3. PROJECT PLAN OR SCHEDULE

This figure below shows the project plan we follow as a team. By using this Gantt chart it will be easier to schedule tasks and workload within the team. This increases the efficiency of the team as well.



	Task Name	Start	Finish
1	**Topic Selection**	11/08/17	02/21/18
2	Discussion in interest	11/08/17	12/19/17
3	Field Study	12/21/17	01/23/18
4	Group Discussion	01/05/18	02/21/18
5	Topic Selection	02/02/18	02/06/18
6	**Feasibility Study**	02/08/18	03/26/18
7	Planning Investigation	02/08/18	02/14/18
8	Technology Analysis	02/15/18	03/05/18
9	Field Study	03/06/18	03/14/18
10	Resource Evalution	03/15/18	03/22/18
11	Project Proposal	03/19/18	03/26/18
12	**Design**	03/14/18	04/30/18
13	**Field Study**	03/14/18	04/02/18
14	Literature Review	03/14/18	03/19/18
15	Requirement Gathering	03/20/18	04/02/18
16	**Specification**	03/28/18	04/06/18
17	Functional Non Functional Requirement	03/28/18	04/06/18
18	**Design**	04/18/18	04/30/18
19	Question Preprocessing	04/18/18	04/20/18
20	RMN for Question Answering	04/21/18	04/21/18
21	Design for the RAM model	04/23/18	04/27/18
22	Answer Generation	04/28/18	04/30/18
23	**System Implementation**	05/01/18	08/30/18
24	Question Preprocessing	05/01/18	05/08/18
25	RMN for Question Answering	05/09/18	05/14/18
26	Answer Generation	05/14/18	05/21/18
27	Improving Reasoning	05/21/18	08/30/18
28	Improving Attention	05/21/18	08/30/18
29	Improving Memory	05/21/18	08/30/18
30	**Testing**	07/02/18	11/02/18
31	Unit Testing	07/02/18	07/30/18
32	Component Testing	08/20/18	09/12/18
33	Integration Testing	09/11/18	09/27/18
34	System Testing	09/20/18	11/02/18
35	**Deliverables**	02/05/18	11/08/18
36	Project Topic Assessment	02/05/18	02/08/18
37	Charter Submission	03/06/18	03/09/18
38	Project Proposal Report	03/30/18	04/03/18
39	Project Proposal Presentation	04/04/18	04/06/18
40	SRS Document/ Prototype	05/01/18	05/08/18
41	Progress Presentation 1	06/25/18	06/29/18
42	Research Paper	08/07/18	08/10/18
43	Progress Presentation 2	08/06/18	08/09/18
44	Final Report	08/09/18	09/03/18
45	Final Report Feedback Submission	10/05/18	10/11/18
46	Final presentation and Viva	11/05/18	11/08/18
47	Final Report(Soft bound)	11/05/18	11/08/18
48	Research paper	08/05/18	08/08/18
49	System Assessment	10/24/18	10/29/18

Figure 3.0 Gantt Chart of the Component

4. RESEARCH CONSTRAINTS

One of the main constraints that we have to face is the suitability of the dataset for our model. Although, we are planning to use the TriviaQA[13] dataset initially, we may have to change the dataset based on Trial and Error.

Another major constraint is that the training process is time consuming and in order to speed up this process it is required to use high end servers with GPU processing capabilities. Using these resources are costly and infeasible at the development stage. Therefore the development iterations might be time consuming.

Another challenge is that there must be continuous monitoring of the training process to avoid overfitting the model to the given dataset. The challenge is to balance overfitting and underfitting when training the model. This can be challenging sometimes with a neural net that is as complicated as RNN's.

The lack of expertise and online help in the context of Information Extraction using DNN is another constraint in this research component. Since we are focusing more on an implementation rather than a highly mathematical study, online help would be highly beneficial. Therefore this is another research constraint that we have to face.

It is clear that like every research there are several research constraints that should be tackled and dealt with appropriately in order to make this research a success.

5. SPECIFIED DELIVERABLES

As the main outcome of the project we aim to deliver a learning domain specific QA system with better performance through enhanced reasoning, memory and attention.

6. REFERENCES

[1] W.A Woods, R.M. Kaplan and B. Nash-Webber, “The lunar sciences natural language information system”, BBN Rep. 2378, Bolt Beranek and Newman, Cambridge, Mass., USA, 1977

[2] R. Dale, H. Moisl and H. Sommers, Handbook of Natural Language Processing, 1st ed. New York: Marcel Dekker AG, 2006, pp. 215 - 250.

[3] L. Hirschman and R. Gaizauskas, “Natural language question answering: the view from here”, Natural Language Engineering, 7 (4), 2001, pp. 275-300.

[4] "The START Natural Language Question Answering System", Start.csail.mit.edu, 2017. [Online]. Available: http://start.csail.mit.edu/index.php. [Accessed: 26- Mar- 2017]

[5] B. Katz, G. Borchardt and S. Felshin, “Natural Language Annotations for Question Answering”, Proceedings of the 19th International FLAIRS Conference (FLAIRS 2006), 2006

[6] Y, Bengio, P. Simard, and P. Frasconi, “ Learning long-term dependencies with gradient descent is difficult.” *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994.

[6] N. Gupta, "Artificial Neural Network", Network and Complex Systems, vol. 3, no. 1,pp. 24-28, 2013.

[7] "RAM Workshop", Thespermwhale.com, 2018. [Online]. Available: http://www.thespermwhale.com/jaseweston/ram/. [Accessed: 01- Apr- 2018].

[8] Memory Networks. Jason Weston, Sumit Chopra, Antoine Bordes. International Conference on Representation Learning, 2015

[9] Teaching Machines to Read and Comprehend. Karl Moritz Hermann et. al. arXiv Pre-Print, 2015.

[10] Large-scale Simple Question Answering with Memory Networks. Antoine Bordes, Nicolas Usunier, Sumit Chopra, Jason Weston. arXiv Pre-Print, 2015.

[11] Arxiv.org, 2018. [Online]. Available: https://arxiv.org/pdf/1803.03067.pdf. [Accessed: 01- Apr- 2018].

[12] "Optimizing Memory Consumption in Deep Learning — mxnet documentation", Mxnet.incubator.apache.org, 2018. [Online]. Available: https://mxnet.incubator.apache.org/architecture/note_memory.html. [Accessed: 01- Apr- 2018].

[13] "TriviaQA", Nlp.cs.washington.edu, 2018. [Online]. Available: http://nlp.cs.washington.edu/triviaqa/. [Accessed: 01- Apr- 2018].

[14] Krizhevsky, A., Sutskever, I. & Hinton, “G. ImageNet classification with deep convolutional neural networks. In Proc. Advances in Neural Information Processing Systems”, 2012. [Accessed: 03- Apr- 2018]

[15] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller and W. Burgard, "Multimodal deep learning for robust RGB-D object recognition," *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Hamburg, 2015, pp. 681-687. [Accessed: 03- Apr- 2018]

APPENDIX I: DESCRIPTION OF PERSONAL AND FACILITIES

The description of the personnel involved in this project is as follows:

Supervisor: Mr. Yashas Mallawarachchi

Co-supervisor: Mr. Anupiya Nugaliyadde

Implementation team:

K.S.D.Ishwari

A.K.R.R.Aneeze

S.Sudheesan

H.J.D.A. Karunaratne