Context Aware Text Summarizer

A Project Report

submitted by

ANASHKA NAINA U , MUHAMMED SHIRAZ
ROHIT P, SARATH SREEDHAR

in partial fulfilment of the requirements
for the award of the degree of

BACHELOR OF TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
COLLEGE OF ENGINEERING,TRIVANDRUM
May 2019

CERTIFICATE

This is to certify that the project report entitled **Context Aware Text Summarizer**, submitted by **Anashka Naina U, Muhammed Shiraz, Rohit P, Sarath Sreedhar** to the College of Engineering Trivandrum, for the award of the degree of **Bachelors of Technology**, is a bona fide record of the project work carried out by them under my supervision. The contents of this report, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr. Ajeesh Ramanujan	**Prof. Vipin Vasu A V**	**Dr. Salim A**
Assistant Professor	Associate Professor	Professor
(Project Guide)	(Project Coordinator)	(Head of Department)

Place: Trivandrum

Date: May 3, 2019

DECLARATION

We undersigned hereby declare that the Project report **CONTEXT AWARE TEXT SUMMARIZER** , submitted for partial fulfillment of the requirements for the award of degree of Bachelor of Technology of the APJ Abdul Kalam Technological University, Kerala is a bonafide work done by us under supervision of **Dr. Ajeesh Ramanujan**. This submission represents our ideas in our own words and where ideas or words of others have been included, we have adequately and accurately cited and referenced the original sources. We also declare that we have adhered to ethics of academic honesty and integrity and have not misrepresented or fabricated any data or idea or fact or source in our submission. We understand that any violation of the above will be a cause for disciplinary action by the institute and/or the University and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been obtained. This report has not been previously formed the basis for the award of any degree, diploma or similar title of any other University

Place: Trivandrum **Team**
Date:

ACKNOWLEDGEMENTS

Firstly, We would like to thank the almighty for giving us the wisdom and grace for completing our project.

With a profound sense of gratitude, We would like to express our heartfelt thanks to our guide **Dr. Ajeesh Ramanujan**, Assistant Professor, Department of Computer Science and Engineering for his expert guidance, co-operation and immense encouragement in pursuing this project.

I am very much thankful to **Dr. Salim A**, Head of Department and **Prof. Vipin Vasu A V**, Associate Professor of Department of Computer Science and Engineering, our guide for providing necessary facilities and his sincere co-operation.

My sincere thanks is extended to all the teachers of the department of Computer Science and Engineering and to all my friends for their help and support.

Team

ABSTRACT

Keywords: Text summarization, Semantic representation, Abstractive model, Neural networks, Word embeddings, GloVe, seq2seq RNN model, Attention layer

Text Summarization has always been a difficult task since the earlier days it started, the extent to which the computers can mimic the human has always been limited in scenarios where abstractive analysis is required. In this project, we would try to address this issue and present a solution to this in the domain of text summarization. There exist two fundamental approaches for text summarization, which are extractive and abstractive summarization. In an extractive summarizer model, important words or phrases are identified from a document and the summary retains this set of words or phrases. Humans summarize text by creating a semantic representation of the content in the brain. The way humans forms this semantic representation is an abstractive model which can be simulated with the help of neural networks. The model that we examine in this project, makes use of word embeddings, initially, word embeddings are created using pre-trained GloVe model [6]. This word embeddings are then used to train a seq2seq RNN model [7] which is used for generating a summary. An attention layer is used which acts as a memory layer for the model to keep track of the content to be tracked. The dataset used for the initial training process is The Signal Media One-Million News Articles Dataset. Further training would be proceeded based on the availability of standard datasets.

TABLE OF CONTENTS

DECLARATION i

ACKNOWLEDGEMENTS ii

ABSTRACT iii

LIST OF TABLES vi

LIST OF FIGURES vii

ABBREVIATIONS viii

1 INTRODUCTION 1

1.1 Report Outline 2

1.2 Motivation 2

1.3 Contribution 3

1.4 Problem Definition 3

2 BACKGROUND AND RELATED WORKS 4

2.1 Generating News Head lines with Recurrent Neural Networks [4] 4

2.2 Sequence to Sequence learning with neural networks [7] 5

2.3 Effective Approaches Attention-based neural machine translation [5] 6

3 CONCEPTS 7

3.1 Encoder Decoder Architecture 7

3.2 Attention [4] 9

4 SYSTEM ARCHITECTURE 12

5 EVALUATION **16**

6 RESULT **17**

7 CONCLUSION AND FUTURE WORKS **19**

LIST OF TABLES

6.1 Example predictions 18

LIST OF FIGURES

3.1 Encoder-decoder neural network architecture 7
3.2 Complex Attention 10
3.3 Simple Attention 11

4.1 Encoder Network 13
4.2 Decoder Network 14
4.3 Decoder Network-Attention Layer 15

ABBREVIATIONS

GloVe	Global Vector
seq2seq	Sequence to Sequence
NLP	Natural Language Processing
NLU	Natural Language Understanding
LSTM	Long Short Term Memory
RNN	Recurrent Neural Network
WMT	Workshop on Machine Translation
BLEU	Bilingual Evaluation Understudy Score
NMT	Neural Machine Translation
EOS	End Of Sequence

CHAPTER 1

INTRODUCTION

Today we are living in a busy world which is overwhelmed by information from different sources like blogs, news article, etc. It is so hard to absorb only necessary information from long articles because it is time-consuming. In such situations, it will be a blessing to get these long articles summarized. While summarizing the text the meaning and important information of the text should be preserved.

Text summarization can be extractive or abstractive. The words phrases or sentences from the document will be selected to generate a summary in extractive summarization. The summaries generated through extractive summarization usually be as perfect as human-written summaries. Abstractive text summarization follows another way. After thoroughly understanding the text, rephrase it into a shorter version. The summary may contain new words. The method of abstractive text summarization is somewhat complex than extractive. It may associate the capabilities of generalization, paraphrasing and may use some external knowledge.

In our model, we use GloVe which is a word-word co-occurrence count that have the potential for encoding some form meaning. Seq2seq model simulates the working brain and the attention model helps to store knowledge.

In the case of GloVe, the counts matrix is preprocessed by normalizing the counts and log-smoothing them. Compared to word2vec, GloVe allows for parallel implementation, which means that its easier to train over more data. It is believed (GloVe) to combine the benefits of the word2vec skip-gram model in the

word analogy tasks, with those of matrix factorization methods exploiting global statistical information. GloVe is essentially a log-bilinear model with a weighted least-squares objective. The model rests on a rather simple idea that ratios of word-word co-occurrence probabilities have the potential for encoding some form of meaning which can be encoded as vector differences.

1.1 Report Outline

This project report will proceed as follows, the rest of this chapter discusses the motivation for the project and contribution from our side towards this project. The Chapter 2 discusses the background and related works, the Chapter 3 discusses the underlying concepts governing the working of this project. The remaining chapters discusses the system architecture and discusses the current working sate of the work and its evaluation.

1.2 Motivation

We know that for any application, the reasoning based on the knowledge and contextual information produce better results. When we are processing a Natural language text, it needs to be processed in the right context in order to get the correct meaning. Current methods of NLP are largely driven by computational statistics. These methods dont attempt to understand the text, instead they convert the text into data and then attempt to learn from the pattern in that data. Research attempts to identify textual context is too less. Context identification is highly depended on natural language understanding and ability to reason. Our project tries to consider NLU, context and reasoning together and we aim at general public due to large booming of textual data in internet.

1.3 Contribution

Our approach is to summarize text using GloVe based seq2seq encoder-decoder model which imitates the working of brain.The existing models are largely statistical driven and abstractive model does not account for context which an important criteria in analyzing languages.In our model we use GloVe which is a word-word co-occurrence count that have the potential for encoding some form meaning. Seq2seq model simulates the working brain and the attention model helps to store knowledge.

1.4 Problem Definition

To design and develop an abstractive summarizer which is Context Aware using deep learning concepts.

CHAPTER 2

BACKGROUND AND RELATED WORKS

Our project planning started with the tag ”Thought Vector/Improved Text Summarizer”, over the period of time our project didn't just focus on to a single topic, as a result, the tag of the project varied from Context Analyzing to Context Modeling and finally to Context Identification.

As thought vector predicts the next thought, the input and output associated must be similar. Two sentences are said to be similar if they have a set of common properties and feature values which are similar to each other (weights in the case of the neural network). If we can identify these features and define a common framework to express these features then probably we can consider it as a context thus the term context modeling.

The summary of some papers referred during the different phases of project are added below:

2.1 Generating News Head lines with Recurrent Neural Networks [4]

The paper describes generating news headlines from news articles. The paper describes an encoder-decoder recurrent neural network with LSTM units and an attention layer. The encoder-decoder networks are responsible for the generation of phrases and the attention layer is responsible for mimicking the human brain working by giving weights to the phrases to signify the importance of phrases towards the summary of the article. Thus the paraphrasing of news articles from

the given news description is done effectively done by the discussed model. The paper also explores the possibility of neural networks in the field and identifying the function of the different neurons in a simplied attention mechanism.

2.2 Sequence to Sequence learning with neural networks [7]

The paper discusses on the limitation of the deep neural network's inability to map sequence to sequence and introduces a general end-to-end model/approach that makes a minimal amount of assumption on the sequence structure compared to the prior models. The encoder-decoder model function as follows where the encoder maps the sequence to a vector and in the decoder section the sequence is decoded back from the vector generated during the encoder phase. The translation from English to French is considered as the main outcome of this project. The project was done using a standard dataset WMT-14 and achieved a BLEU score of 34.8. To map the input sequence to a fixed dimensionality vector, a multilayered LSTM RNN is used. And to decode the vector to sequence another deep LSTM is used, which forms the encoder-decoder pair. The end-outcome of the project discusses the importance of word order in a sentence, that is the model was inclined towards sensible sentences and word order. To strengthen the claim they retried the experiment by reversing the order of words in the given input and the performance was similar, thus claiming that the short word dependencies formed during such action have a larger effect on the result, thus making the optimization easier.

2.3 Effective Approaches Attention-based neural machine translation [5]

The main disadvantages of the earlier model were that giving weight-age to a particular word during the translation was missing. Although this is a known fact to all the amount of research and works exploring a mechanism to overcome this has been minimal. The attention mechanism that is discussed in this paper introduces a layer that improves the NMT (Neural Machine Translation) by giving weight-age to parts of source sentence during the translation activity so that the model can selectively focus on parts of sources that have a higher influence on the output sequence generated. The attention model discussed in the paper has mainly two classes of focusing (attention mechanism): a global approach and a local approach. In the global approach focus/attention is on all the words present in source and in the local approach focus/attention is on a particular subset of words at a particular instance. This model was able to achieve a gain of 5 BLEU score over the conventional non-attentional model. The ensemble model this team developed was able to achieve a new state-of-the-art result in the WMT's translation of English-to-German using neural networks.

In the next chapter we would be discussing the underlying concepts that governs the working of our project.

CHAPTER 3

CONCEPTS

3.1 Encoder Decoder Architecture

The encoder-decoder architecture[7; 5] comprises of an encoder and a decoder components. Both these components/parts are recurrent neural networks which make use of neural networks to improve the performance of the sequence to sequence mapping of a translation process.

The input given to the encoder is the text of the document that we want to summarize and the text of the document is fed one word at a time. Each word is converted into the corresponding distributed representation [1] by passing the word through an embedding layer, thus the name word embedding. The distributed representation refers to a form which makes the conversion of sequence to vectors easier with the help of sub-sampling of frequent words like 'a', 'the',etc.. and Softmax layer. Instead of a softmax layer, there is scope for in-cooperating a negative sampling which tries to reduce the probability of occurrence of two words close to each other, this process is a less complex and computationally easier work.

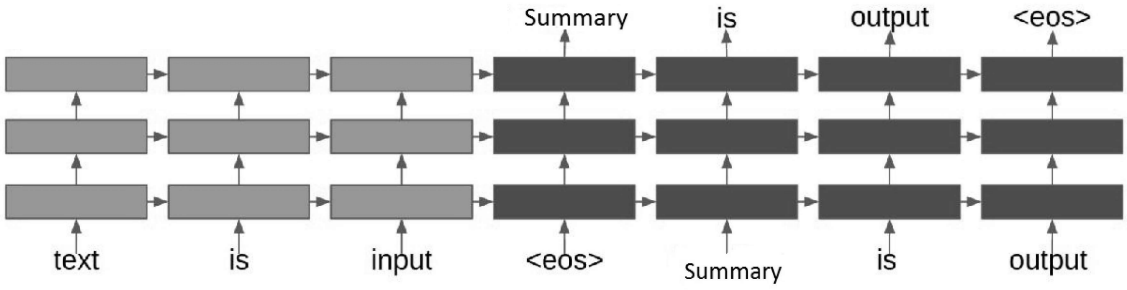

Figure 3.1: Encoder-decoder neural network architecture

A multi-layer neural network is then used to combine the distributed representation that is generated by the embedded layer. The multi-layer neural network comprises of hidden layers that are generated after supplying the previously generated words to the encoder, if the current word is the first word of the source document then all zeros are fed to the encoder. The hidden layers generated after feeding in the final word of the source text of the document to the encoder is fed as input to the decoder. End of sequence(EOS) symbol is fed as input first to the decoder and then uses the embedding layer to create distributed representation of the result of the encoder. The decoder then creates each word of the headline ending with the end-of-sequence symbol by using a softmax layer and the attention mechanism. After the generation of each word, this word is fed as input to the encoder-decoder network and creates the next word and so-on.

The loss function we use is the log loss function

$$-logP(y_1, \ldots, y_{T'} \mid x_1, \ldots, x^T) = -\sum_{t=1}^{T'} logP(y_t \mid y_1, \ldots, y_{t-1}, x_1, \ldots, x_T) \tag{3.1}$$

Here x and y represent input and output words respectively.

To stand unique in terms of efficiency compared to the prior models, the model discussed here include 'teacher forcing'[3] during the training phase. Teacher forcing is an optimization procedure done during the training phase that helps in training the recurrent neural network faster and more efficiently. This is done by utilizing the output from a previous time step as the input for the current layer. Instead of generating a new word at each run and feeding it in as input for generating the next word, an expected word from the actual headline is fed to the network. However, during the testing, this practice is not followed, and the

previously generated word from the previous layer is introduced when generating the next word. This gives rise to a break between training and testing. To overcome this break, during training, the network is fed randomly with a generated word, instead of the expected word as in the model discussed earlier [1]. Specifically, we do this 10% of the time, as also done in [2].

3.2 Attention [4]

The attention mechanism is a mechanism that helps the network to remember/focus certain parts of the source document better, such as names, numbers, facts, etc... The attention mechanism is used after each word is output in the decoder. For each output word, the weighing mechanism calculates the weight for each input word that determines how much attention/focus should be paid to that input word. The total weights adds up to one and are used to calculate the weighted average of the last layers created after processing each input word. This weighted average, referred to as context, is then entered into the softmax layer along with another hidden layer from the current decoding step. There are two different attention mechanism we looked into : The first attention mechanism, which we refer to as complex attention. This mechanism is shown in Fig: 3.2. The attention weight for the input word at position t, computed when outputting the $t^{'}$ th word is:

$$a_{y_{t'}}(t) = \frac{exp(h_{x_t}^T h_{y_{-}t'})}{\sum_t^T exp(h_{x_{t-}}^T h_{y_{t'}})} \tag{3.2}$$

where $h_{x_t}^T$ represents the last hidden layer generated after processing the t-th input word, and $h_{y_{t'}}$ represents the last hidden layer from the current step of decoding. The main highlight of this mechanism is that the same hidden units are used for computing the attention weight as for computing the context of the source.

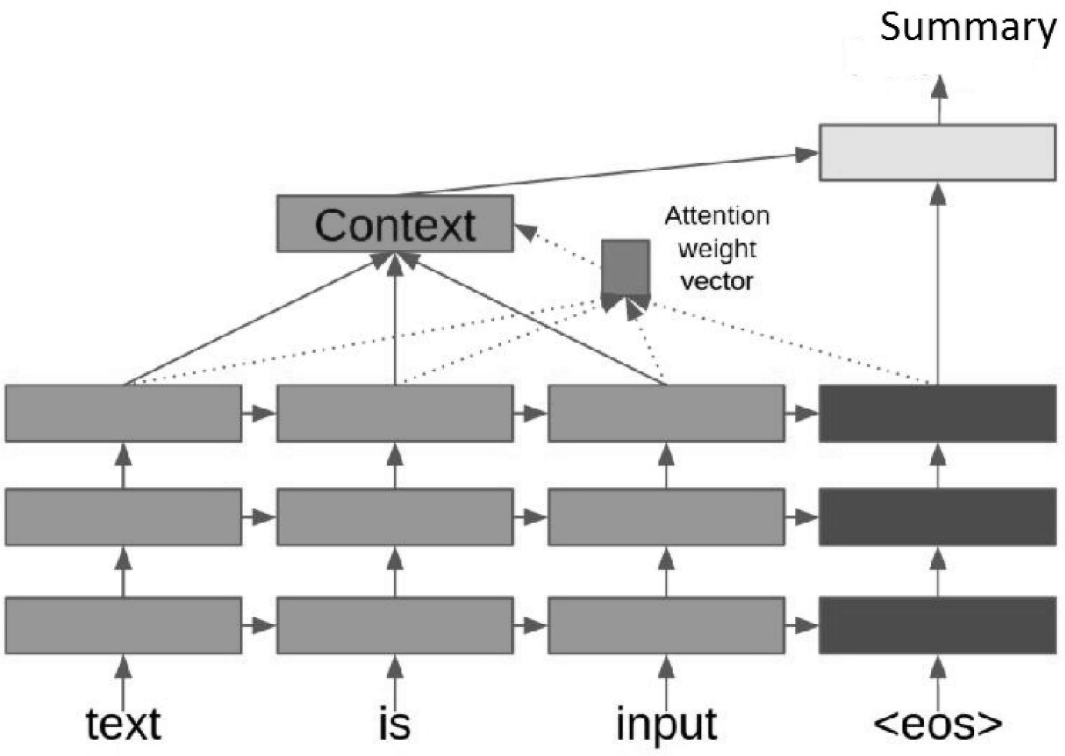

Figure 3.2: Complex Attention

The second attention mechanism, which is referred to as simple attention, is a slight variation of the complex mechanism that makes it easier to analyze how the neural network learns to compute the attention weights. This mechanism is shown in Fig: 3.3. Here, the hidden units of the last layer generated after processing each of the input words are split into 2 sets: one set of size 50 used for computing the attention weight, and the other of size 550 used for computing the context. Analogously, the hidden units of the last layer from the current step of decoding are split into 2 sets: one set of size 50 used for computing the attention weight, and the other of size 550 fed into the softmax layer. Aside from these changes the formula for computing the attention weights, given the corresponding hidden units, and the formula for computing the context is kept the same.

In the next chapter we would be discussing how we implemented these concepts into code and a snap of the overall network is included.

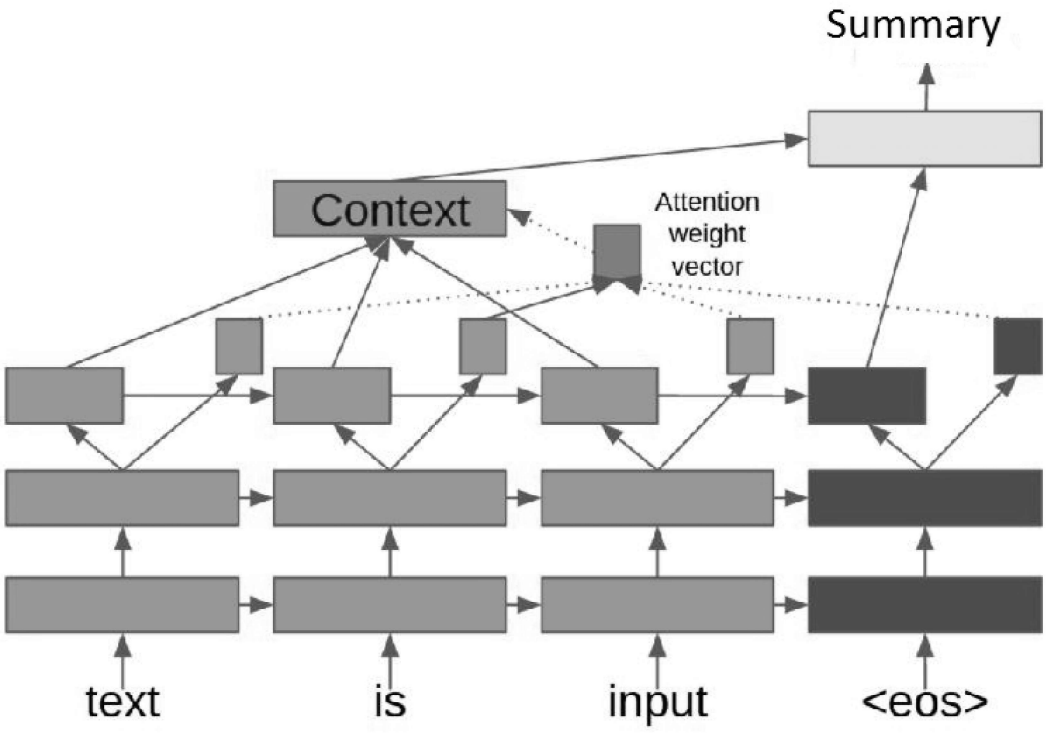

Figure 3.3: Simple Attention

CHAPTER 4

SYSTEM ARCHITECTURE

The system comprises of two recurrent neural networks . The first one is an encoder network which takes in an input sequence and creates an encoded representation of it. The second one is a decoder network which takes in the earlier encoded representation as to its input and it will generate an output sequence decoding it. The model comprises of three stacked LSTM networks shown in Fig: 4.1. The weight of the first layer is initialized with pre-trained GloVe embeddings which are computationally feasible. The embedding layer is meant to turn input into fixed-size vectors. The cross-entropy losses are minimized using rms-prop [8]. The decoder network has the same LSTM architecture of the encoder network shown in Fig: 4.2. The model is initialized with same pre-trained GloVe embedding weights. The input is the vector representation generated after feeding in the last word of the input text. It will generate its own representation using its embedding layer. The next step is to convert this representation into a word. The decoder will generate a word as its output and that same word will be fed in as input when generating the next word until it produces a summary. An attention mechanism is used when outputting each word in the decoder depicted in Fig : 4.3. For each output word, it computes a weight over each of the input words that determines how much attention should be given to that input word. All the weight sum up to 1 and are used to compute a weighted average of the last hidden layers generated after processing each of the inputted words. The weighted average is fed into the softmax layer along with the last hidden layer from the current step of the decoder[4].

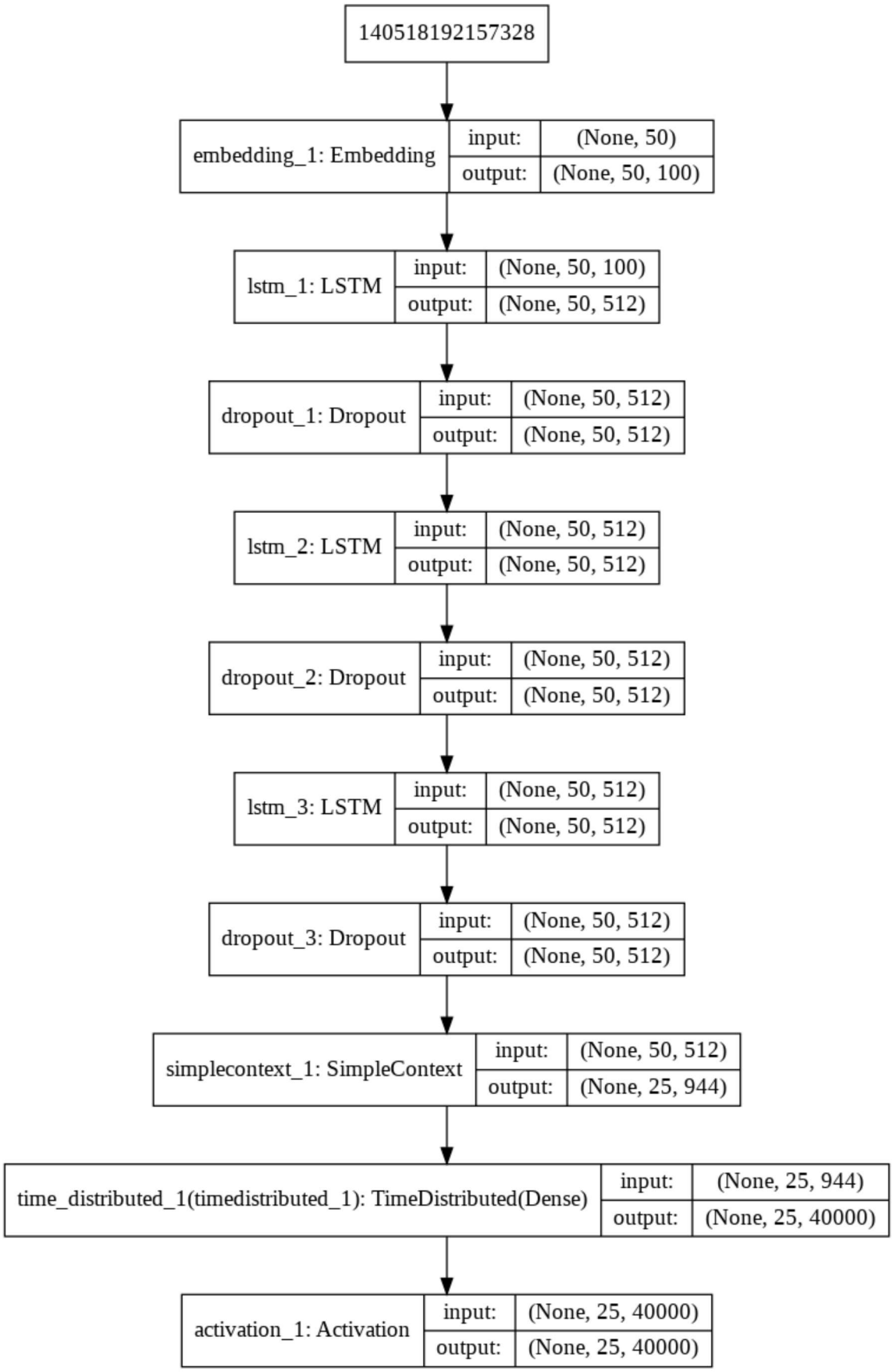

Figure 4.1: Encoder Network

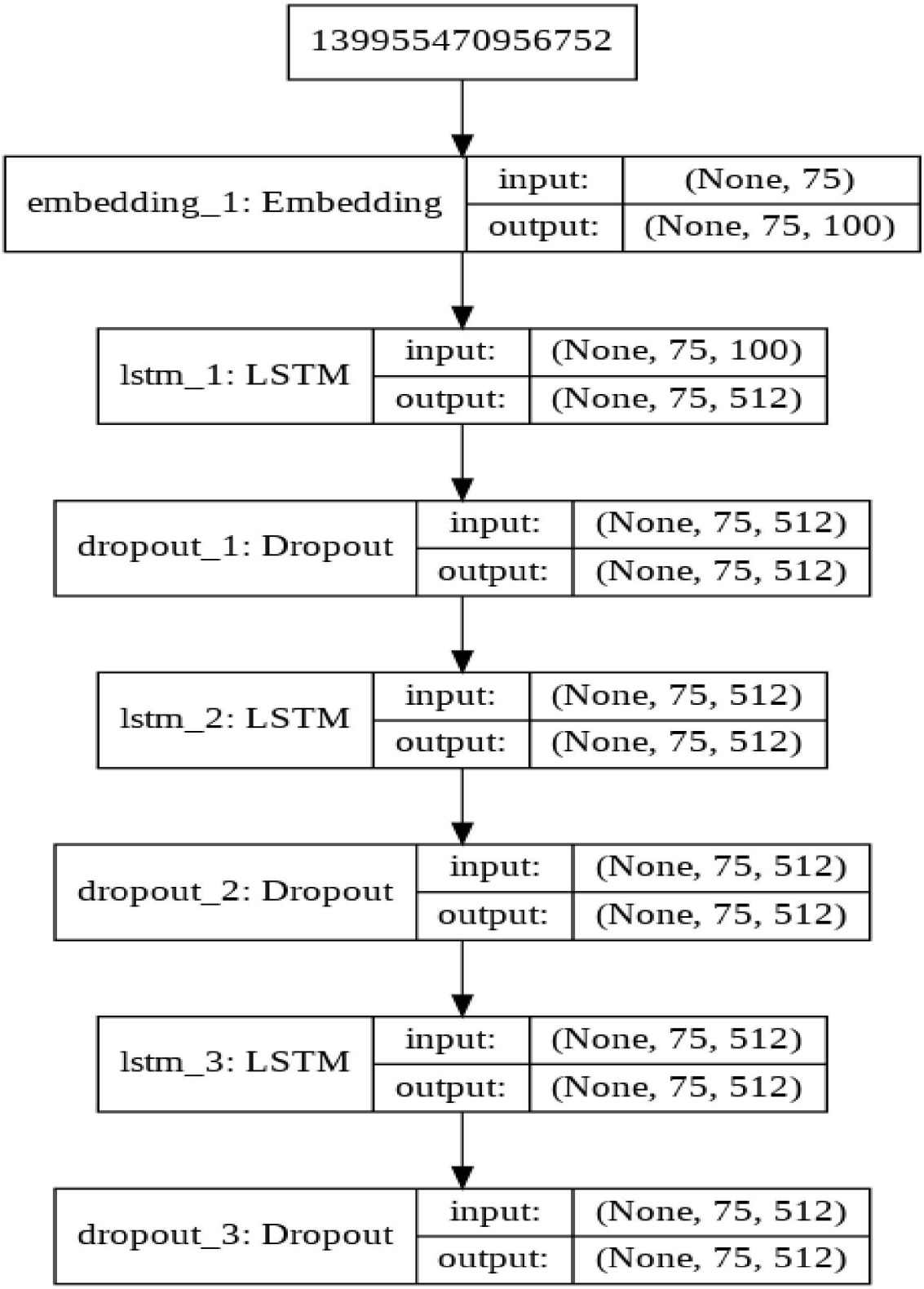

Figure 4.2: Decoder Network

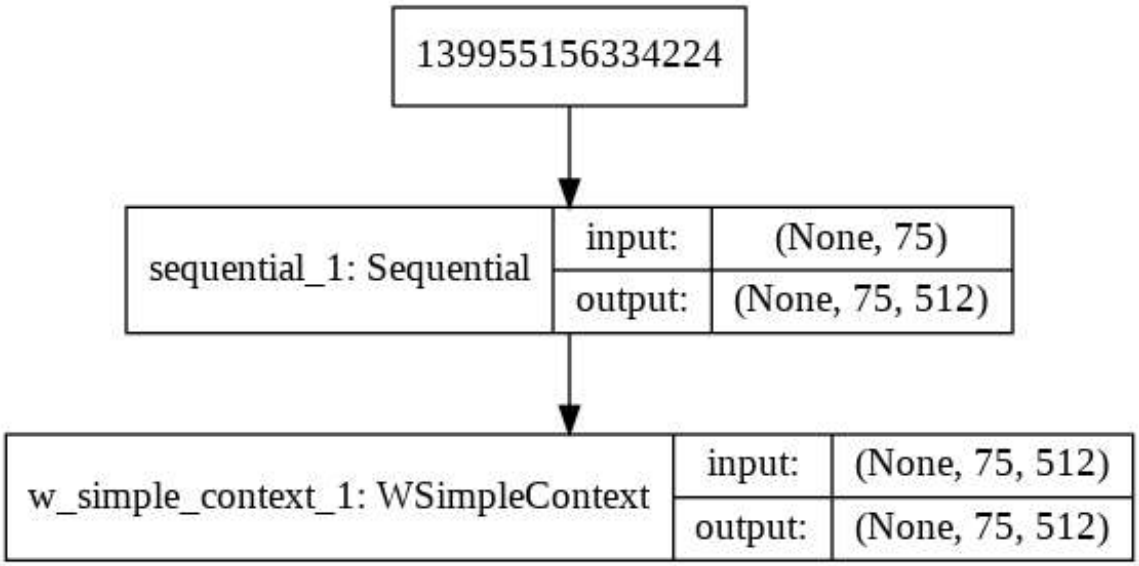

Figure 4.3: Decoder Network-Attention Layer

CHAPTER 5

EVALUATION

BLEU metric was used for the evaluation of generated summary. BLEU uses precision for measurement. It computes the number of overlapping n gram. Uni-gram based BLEU is used for evaluation. The actual summary generated is compared with target summary. The word-word occurrences is computed using BLEU score implemented in nltk package. BLEU is a metric used widely for machine based translation evaluation. The summaries are tokenized and fed to the function. The average of BLEU score is computed over hundred text sample. The max score obtained is one using this metric which resembles an exact match between generated and actual summaries. The score reduces when there is no word match between actual and generated summaries. For an abstractive summarization process the metrics as some shortcomings, An abstractive summary tends to replace word used in text to some alternative for enhanced understanding, so a word co-occurrences metric doesnt actually model an exact scenario of use-case. Human based evaluation are more advisable than system based metrics. The performance of BLEU metric also depends on the quality of actual summary. This intrinsic evaluation doesnt account for acceptability or further usage of summary.

The next chapter we would discuss about the performance of the model discussed.

CHAPTER 6

RESULT

The given model was evaluated using the BLEU metric mentioned in the evaluation section. The average BLEU score generated for around 700 examples was **0.19**. Sample predictions are shown in Table: 6.1. The model has certain drawbacks. Some of them are listed here. Words which were not present in the word embedding is shown error while prediction. These words were marked using the label _UNK. Example 5 shows this case, The word Kleinheinz was not present in the vocabulary embedding. The second one is that the model reproduced some factual data incorrectly. This error was present in example 4, Norwegian was replaced by Swiss even though it was not present in the text sample. The third one was that the model repeated a word and ended sentences abruptly. The issue was raised in example 6, the sentence ended with the phrase government of government.

In the next chapter we would summarize the report and discuss the future works we intend to do.

Text	Actual Headline	Predicted Headline	BLEU Score
1. India will invest US$3 billion (euro2.3 billion) in developing oil and gas fields in Russia to build a partnership in the energy sector, a news report said Sunday.	India to invest in Russian oil, gas projects	India to invest in oil pipeline deal in	0.63338
2. The government will sell its last direct stake in the country's largest bank, the National Bank of Greece, the finance ministry said.	Greece to sell final stake in largest bank	Government to sell stake in commercial bank	0.61919
3. The Ulster Defense Association, Northern Ireland's largest outlawed group, announced Sunday it will stop fighting, saying it wants to rejoin peacemaking efforts.	Outlawed Protestant group seeks to rejoin Northern Ireland's peace process	Northern Ireland ' s Protestant group to sue government over	0.57311
4. The government on Friday refused to back a bid to host the 2014 Winter Olympics by the Arctic city of Tromsoe, saying it was too costly and too soon after the 1994 Lillehammer Games.	Norwegian government refuses to back 2014 Winter Olympics bid	Swiss government refuses to bid off bid for Olympics	0.55555
5. Austria's Markus Kleinheinz won a men's singles luge World Cup race Sunday, handily beating Russia's Albert Demtchenko.	Kleinheinz wins men's World Cup luge singles	Austria ' s _UNK win World Cup singles	0.55156
6. Iranian Nobel Peace Prize winner Shirin Ebadi, praised by President George W. Bush and honored at universities for her work on behalf of democracy and human rights, is suing the U.S. government for standing in the way of the publishing of her memoirs.	Nobel Laureate sues U.S. for rights to publish memoirs	Nobel Nobel laureate for Bush apology on U . S . rights program	0.53846
7. Rebels accused Ivory Coast's government Monday of preparing a fresh offensive against them in the north of the country.	Rebels accuse Ivory Coast government of preparing fresh offensive	Ivory Coast rebels accuse government of government	0.53676

Table 6.1: Example predictions

CHAPTER 7

CONCLUSION AND FUTURE WORKS

The project describes the development of a text summarizer. Apart from the conventional text summarizers which summarize based on the word count/importance, the model we discussed in this report summarizes a textual document in an abstractive way. The abstractive manner is similar to the semantic representation of the human perceives while summarizing a text. As the number of works done in this field is minimal the development of the work faced quite a number of difficulties. The major difficulty was the absence of a standard dataset and depreciation of certain packages by the tensorflow.

The model starts with pre-processing of the textual document, for that purpose we make use of GloVe which is an unsupervised learning algorithm which takes the word-to-word co-occurrence count statistics from a corpus to generate a vector representation for words. This representation has the potential for encoding some form of meaning for the representation. The Seq2seq encoder-decoder model simulates the working of the human brain and the attention model helps to store knowledge which tries to mimic the memory capacity of the human brain. The possible future works to be done as a part of the project are:

- To create a dataset for analyzing context in textual data.
- To formulate an appropriate data structure for context representation.

REFERENCES

[1] **Bengio, S.**, **O. Vinyals**, **N. Jaitly**, and **N. Shazeer**, Scheduled sampling for sequence prediction with recurrent neural networks. *In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15. MIT Press, Cambridge, MA, USA, 2015. URL `http://dl.acm.org/citation.cfm?id=2969239.2969370`.

[2] **Chan, W.**, **N. Jaitly**, **Q. V. Le**, and **O. Vinyals** (2015). Listen, attend and spell. *CoRR*, **abs/1508.01211**. URL `http://arxiv.org/abs/1508.01211`.

[3] **Goodfellow, I.**, **Y. Bengio**, and **A. Courville**, *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[4] **Lopyrev, K.** (2015). Generating news headlines with recurrent neural networks. *CoRR*, **abs/1512.01712**. URL `http://arxiv.org/abs/1512.01712`.

[5] **Luong, T.**, **H. Pham**, and **C. D. Manning**, Effective approaches to attention-based neural machine translation. *In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 2015. URL `https://www.aclweb.org/anthology/D15-1166`.

[6] **Pennington, J.**, **R. Socher**, and **C. D. Manning**, Glove: Global vectors for word representation. *In Empirical Methods in Natural Language Processing (EMNLP)*. 2014. URL `http://www.aclweb.org/anthology/D14-1162`.

[7] **Sutskever, I.**, **O. Vinyals**, and **Q. V. Le** (2014). Sequence to sequence learning with neural networks. *CoRR*, **abs/1409.3215**. URL `http://arxiv.org/abs/1409.3215`.

[8] **Zaremba, W.**, **I. Sutskever**, and **O. Vinyals** (2014). Recurrent neural network regularization. *CoRR*, **abs/1409.2329**. URL `http://arxiv.org/abs/1409.2329`.