Journal of Artificial Intelligence and Capsule Networks (2020)
Vol.02/ No. 01
Pages: 53-63
http://irojournals.com/aicn/
DOI: https://doi.org/10.36548/jaicn.2020.1.006

Analysis of Neural Network Based Language Modeling

Dr. P. Karuppusamy,
Professor, Department of EEE,
Shree Venkateshwara Hi-Tech Engineering College,
Erode, India.
Email id: pkarrupusamyphd@gmail.com

Abstract: The fundamental and core process of the natural language processing is the language modelling usually referred as the statistical language modelling. The language modelling is also considered to be vital in the processing the natural languages as the other chores such as the completion of sentences, recognition of speech automatically, translations of the statistical machines, and generation of text and so on. The success of the viable natural language processing totally relies on the quality of the modelling of the language. In the previous spans the research field such as the linguistics, psychology, speech recognition, data compression, neuroscience, machine translation etc. As the neural network are the very good choices for having a quality language modelling the paper presents the analysis of neural networks in the modelling of the language. Utilizing some of the dataset such as the Penn Tree bank, Billion Word Benchmark and the Wiki Test the neural network models are evaluated on the basis of the word error rate, perplexity and the bilingual evaluation under study scores to identify the optimal model.

Keywords: Neural Networks, Natural Language Processing, Language Modelling, Word Error Rate, Perplexity and the Bilingual Evaluation under Study Scores

1. Introduction

The modelling of the languages mostly indicates the abstract structure development for the contents of the language. This is done in order to comprehend the hints in the languages, automatic generation or transformation of the language contents or into other languages respectively. The modelling in language has received a great deal of interest from different research groups over the past three to four decades, ranging from linguistics, psychology, speech recognition, data compression, neuroscience, machine translation etc. This is highly prominent in the processing of natural language not excluding the field of semantic and the syntactic computing, correction of spelling, handwriting recognizing, recognizing speech automatically, generating of texts, computer translations and optical character recognition etc.

ISSN: 2582-2012 (online)

Journal of Artificial Intelligence and Capsule Networks (2020)
Vol.02/ No. 01
Pages: 53-63
http://irojournals.com/aicn/
DOI: https://doi.org/10.36548/jaicn.2020.1.006

Statistical approximation procedures such as “N-gram structures” have become common because of the ease of design, capable of being used across a variety of languages and the ever-increasing quantities of online language tools that such information that are disadvantaged but are searching for optimal data techniques. The success story of N-gram models of impeded work relies on linguistic knowledge-based approaches. The solution for the problem of sparseness in the “N-gram structures” was developed with the help of the artificial neural networks. The neural network structures apart from predicting were also capable of getting trained with the representations of the real value in the vector form. From the beginning the “N-gram structures” and the neural network structures almost showed similar results and were used as complementary whenever the structures of Ngram were utilized together along the variant of the N-Gram.

The neural-network work in the beginning centered on developing improved training methods, time taken to train and the parameters required to train while retaining performance of the model as before. As the neural networks had high level of processing capability, and linguistic resources availability most of the researcher’s preferred the neural networks in their research process. Further they even utilized the Deep learning models to develop a certain level of confidence in the research for carry out trial with the designs used for language modelling using neural networks.

The inputs in the textual form are usually sequential in nature and the appropriate networks found for the modelling of such inputs as the recurrent-neural networks, so the Recurrent based neural network for the modelling of the languages were developed. Though were very advantageous for the encoding on long term dependencies certain pitfalls in this recurrent model neural network were the difficulties caused in training utilizing the back propagation techniques. To handle these problems the modifications were introduced in the recurrent model, such as the LSTM-Long Short Term Memory and GRU- Gated Recurrent Units, to manage the contents of the languages in natural with the long term dependencies.

Conventionally these language modeling were considered to be a prediction methodologies that were used in predicting the phrase possibility founded on words as linguistic units which restricted these models to fixed vocabulary. So the paper presents the analysis of the neural network based language modelling and evaluates the same on three different data sets on the basis of the “Word Error Rate, Perplexity and the Bilingual Evaluation under Study Scores” to identify the competencies of each method.

The analysis of the neural network founded modelling of language for assisting the processing’s in the natural languages are planned with the related works in its instant part next to introduction and the description of the neural network structures in its next section, the analysis on the basis of the “Word Error

ISSN: 2582-2012 (online)

Journal of Artificial Intelligence and Capsule Networks (2020)
Vol.02/ No. 01
Pages: 53-63
http://irojournals.com/aicn/
DOI: https://doi.org/10.36548/jaicn.2020.1.006

Rate, Perplexity and the Bilingual Evaluation under Study Scores" in the following section and the conclusion in the final section.

2. Related Works

As prior to the analysis the section details the prevailing works that wer done to have a quality language modeling as well the particulars of the techniques used in it. As a preliminary of the analysis the work of Rumelhart, David E et al [1] who presented the "Learning representations by back-propagating errors." I his paper was studied along with the McClelland. E et al [2] "Parallel distributed processing." In the modelling of language Jelinek, E et al [3] presented the "A dynamic language model for speech recognition."

Lau, E et al [4] performed the "Trigger-based language models a maximum entropy approach." To have a perfect language modelling, providing a successful processing of natural languages. Goodman E et al [5] achieved "A bit of progress in language modeling." Bengio E et al [6] presented the "A neural probabilistic language model."

Mikolov E et al [7] put forth the "Recurrent neural network based language model." Whereas Tomáš. E et al [8] conducted the "Statistical language models based on neural networks." Arisoy E et al [9] performed the study on the "Deep neural network language models. "To identify the optimal models that well suited for performing the language modelling using reduced resources, time without sacrificing the performance. Merity, E et al [10] performed the "Regularizing and optimizing LSTM language models."

Raj, Jennifer S E et al [11] conducts "A Comprehensive Survey on the Computational Intelligence Techniques and Its Applications." Shakya, Subarna E et al [12] elaborates the "Machine Learning Based Nonlinearity Determination for Optical Fiber Communication-Review." Joseph, S. E et al [13] describes the "Survey of data mining algorithms for intelligent computing system." J. Vijitha Ananthi E et al [14] has conducted the "Recurrent Neural Networks and Nonlinear Prediction in Support Vector Machines."

ISSN: 2582-2012 (online)

Journal of Artificial Intelligence and Capsule Networks (2020)
Vol.02/ No. 01
Pages: 53-63
http://irojournals.com/aicn/
DOI: https://doi.org/10.36548/jaicn.2020.1.006

3. The Neural Network Models Based Modelling For Languages

The estimation of joint possible happenings of the sequences of words is termed as the modelling of the language. Every word in the sentences are considered as the tokens and the possible dispersion of the sentences in the training quantity. The probability of the sentences are represented as shown in equation 1 below.

$$Prob(Sentence1,..Sentence\ n) = \prod_{x=1}^{n} Prob(\ sentence\ x\ |Sentence1,..Sentence\ x-1) \quad (1)$$

But as this type of processing is bit costly, if 'x' be likely to have a value that is larger, to minimize the cost simple processing using the assumptions of Markov, that says the possibility of the 'x' depends only on the 'x-2' and 'x-1' this is represented in equation 2

$$Prob(\ sentence\ x\ |Sentence1,..Sentence\ x-1) \approx Prob(\ sentence\ x\ |Sentencex-2,Sentence\ x-1) \quad (2)$$

This is referred as the "3-gram" structure and to deal with problems that has a large context it is generalized as "N-gram" and the estimation is done using the 'n-1'word length so the determination of the possibility is done using

$$Prob(\ sentence\ x\ |Sentence1,..Sentence\ x-1) \approx Prob(\ sentence\ x\ |Sentencex-n+1,..Sentence\ x-1) \quad (3)$$

The aforementioned were the some of the earlier mentioned structures in modelling the languages. The following section presents the different types of neural-network frame works in modeling the language. The probabilistic modeling of the language [6] using the feed forward networks was capable of simultaneously learning the representations that are dispersed for every word in the form of vectors as well as possible dispersion for each sequences of word that are stated as symbols. The following flow chart in figure.1 explains the process.

ISSN: 2582-2012 (online)

Journal of Artificial Intelligence and Capsule Networks (2020)
Vol.02/ No. 01
Pages: 53-63
http://irojournals.com/aicn/
DOI: https://doi.org/10.36548/jaicn.2020.1.006

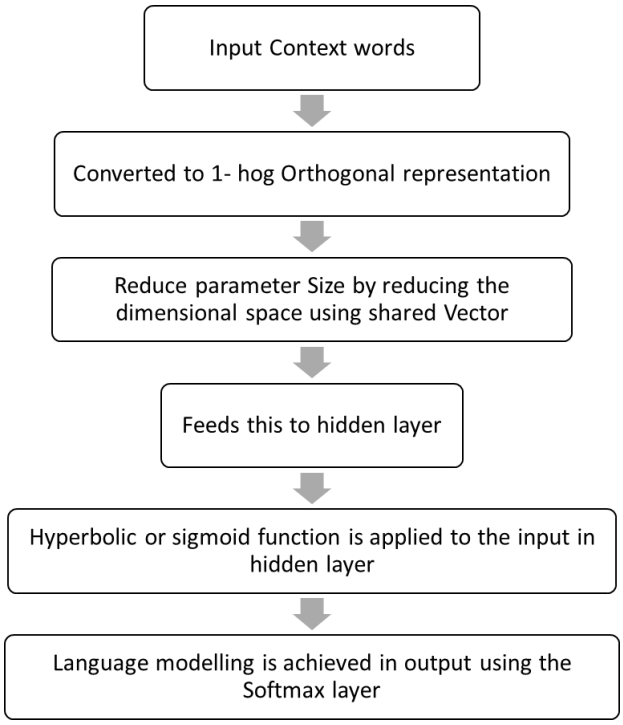

Figure.1 NN Flowchart

To further improve the process of training, the clustering methods pertained model were used, this also ensures minimized computation complexities in the output layer and also affected the overall performance of the network. The concept of sampling, noise constrictive estimation was recommended as an alternative to the softmax operations but this increased the complexities in computation. To further reduce the time in training process the parallel implementation was also suggested.

The deep neural network were framed increasing the number of hidden layers to improvise the modeling quality. Without increasing the time of training and the cost incurred.

But as the Recurrent model provided a better performance for sequential encoding, the Feed forward were replaced by the recurrent based neural networks for modeling the languages. The recurrent networks are usually trained applying the “error back propagation algorithm” or the “error back propagation through

ISSN: 2582-2012 (online)

Journal of Artificial Intelligence and Capsule Networks (2020)
Vol.02/ No. 01
Pages: 53-63
http://irojournals.com/aicn/
DOI: https://doi.org/10.36548/jaicn.2020.1.006

time" the diagram below in figure.2 explains the operation of normal recurrent –NN and the recurrent-NN with classing.

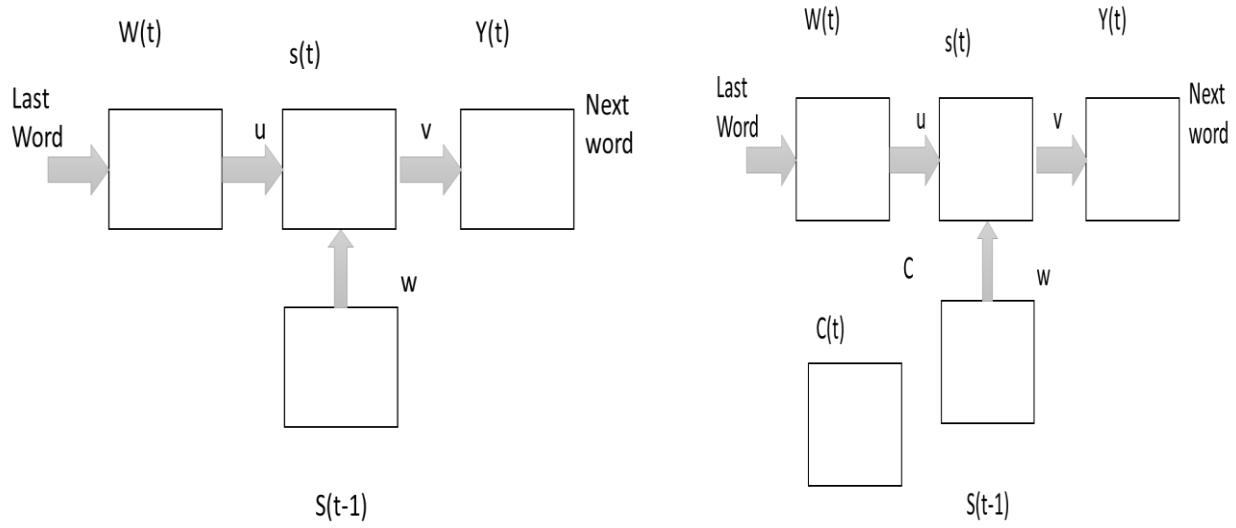

Figure.2 Normal RNN (Left) and RNN with Classing (Right)

But the difficulties caused in training using the back propagation led to the emergence of modifications in the RNN leading to modeling of languages using the "LSTM and the GRU"

In LSTM's the subnets or the memory cells are recurrently connected to enhance the Recurrent-NN context dimensions. It is an extension of RNN substituting the memory blocks for the hidden layers, every layer in the LSTM holds a more than one or a single self-linked cells for memory and triple multiplicative units basically resembling the operations of the memory cell, reset, read and write. The next input is taken in only when the "activation value ≠ 0" and the cells are written with new data. "So that it is made available to the net much later in the sequence, by opening the output gate." The figure3 is the diagram of LSTM based modeling of language.

ISSN: 2582-2012 (online)

Journal of Artificial Intelligence and Capsule Networks (2020)
Vol.02/ No. 01
Pages: 53-63
http://irojournals.com/aicn/
DOI: https://doi.org/10.36548/jaicn.2020.1.006

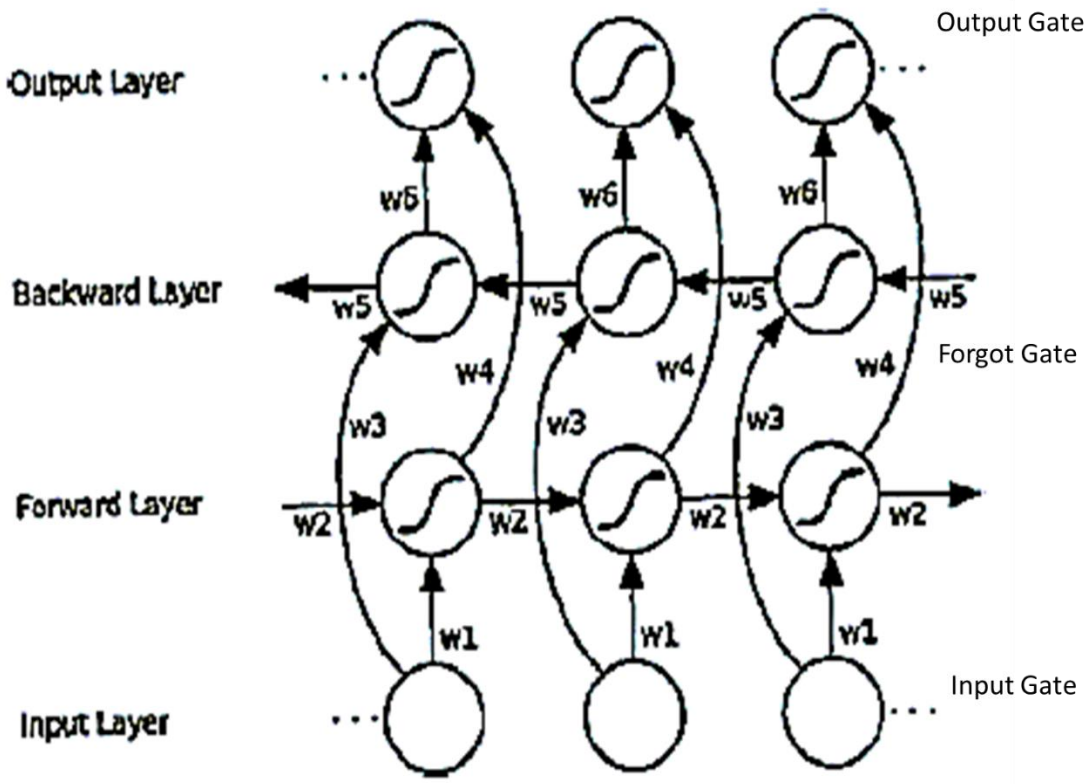

Figure.3 LSTM based Language Modeling [15]

Gated recurrent unit is also a modification of the Recurrent-NN and operates similar to the LSTM, but uses hidden state instead of memory state unlike LSTM. It is equipped with dual gates, “update” and “Reset” while “update” performs similar action of the input gate in LSTM, the “Reset” decided the count of the past information’s that are to be unremembered.

Further Bidirectional Recurrent –NN was introduced to operate the information’s of the NLP in dual directions. This model includes a dual hidden layers that are separate but feeds the same output layer. The computation in forward and reverse sequence takes place in this model in the hidden layers and then generates the output sequences iterating through the layer form back ward and then forward resulting in the output layer update. The LSTM could be used in its hidden layer to have BI-LSTM model for modeling of language. The BI-LSTM was very successful than all the other RNN models.

ISSN: 2582-2012 (online)

Journal of Artificial Intelligence and Capsule Networks (2020)
Vol.02/ No. 01
Pages: 53-63
http://irojournals.com/aicn/
DOI: https://doi.org/10.36548/jaicn.2020.1.006

4. Analysis of Performance

The various Recurrent –NN based language models discussed were evaluated using three different data sets (a) Penn Tree bank that holds a vocabulary dimension of 10000 and has eight hundred and ninety thousand training set, (b) Billion Word with eight hundred million words with vocabulary size of eight hundred thousand and (c) Wiki-Test-2 that has two million datasets of 30000 vocabulary size. These were evaluated on the basis of the (i) perplexity of the sequence of the word, (ii) error rate in word and (iii) Bilingual evaluation.

The "perplexity reflects the quality of the model learned compared to quality of the true model that is under consideration. The all the models mentioned above were trained using 70 % of the three different dataset and the 30 % was used in evaluating. The table.1 below presents the perplexity of the NN, RNN, RNN with Classing (RNN-C), LSTM, GRU, BI-RNN and BI-LSTM

Neural Network Models	**Penn Treebank**	**Billion Word**	**WikiText-2**
NN	60.8	62.4	63.4
RNN	60.4	61.2	59.3
RNN-C	55.3	55.23	55.18
LSTM	54.2	24.3	29.8
GRU	54.19	24.32	29.76
BI-RNN	55	47	34
BI-LSTM	52	24.06	29.0

Table.1 Perplexity

The perplexity is measured as shown in the equation 4 below.

ISSN: 2582-2012 (online)

Journal of Artificial Intelligence and Capsule Networks (2020)
Vol.02/ No. 01
Pages: 53-63
http://irojournals.com/aicn/
DOI: https://doi.org/10.36548/jaicn.2020.1.006

$$Perplexity = \sqrt[R]{\prod_{x=1}^{n} 1 / Prob(\ sentence\ x\ |Sentence1,..Sentence\ x-1\ )} \quad (4)$$

So according to equation lesser perplexity leads to better quality. The table.2 below is the rate of error that is estimated using the equation 5

$$Error\ \%_{Word} = Subsitutions + Deletions + insertions / Number\ of\ Words\ in\ sentence \quad (5)$$

Neural Network Models	**Penn Treebank**	**Billion Word**	**WikiText-2**
NN	44.56	45.89	48.5
RNN	42.34	43.56	42.78
RNN-C	40.84	40.55	40.43
LSTM	22.889	22.789	22.345
GRU	22.43	22.41	22.40
BI-RNN	34..45	34.65	34.7
BI-LSTM	22.03	22.01	22.0

Table.2 Error Rate in the Word

The Bilingual Assessment are usually done for automatic computer translations, though not accurate it allows to have fast and less expensive estimate, and are very easy to comprehend as it correlate highly with the human assessment and are independent of language. This is most probably used in the summarizing the text, translation, image captioning, speech recognition and language generation.

ISSN: 2582-2012 (online)

Journal of Artificial Intelligence and Capsule Networks (2020)
Vol.02/ No. 01
Pages: 53-63
http://irojournals.com/aicn/
DOI: https://doi.org/10.36548/jaicn.2020.1.006

5. Conclusion

To prove the competencies of the neural network models over the n- gram model used in structuring the languages. The paper has presented the analyses of the recurrent neural networks and its further modifications to reduce the time and difficulties in the training process without compromising the performance. Some of the modified models analyzed were LSTM, GRU and BI-RNN (BI-LSTM), these models along with the models of neural networks that were initially used and the RNN with and without classing were trained and tested utilizing three different datasets on the basis of the perplexity and the word error rate the LSTM and the GRU based models proved to provide a better quality of language modeling compared to the other models analyzed.

References

[1] Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors." *nature* 323, no. 6088 (1986): 533-536.

[2] McClelland, James L., David E. Rumelhart, and PDP Research Group. "Parallel distributed processing." *Explorations in the Microstructure of Cognition* 2 (1986): 216-271.

[3] Jelinek, Frederick, Bernard Merialdo, Salim Roukos, and Martin Strauss. "A dynamic language model for speech recognition." In *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991*. 1991.

[4] Lau, Raymond, Ronald Rosenfeld, and Salim Roukos. "Trigger-based language models: A maximum entropy approach." In *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 45-48. IEEE, 1993.

[5] Goodman, Joshua. "A bit of progress in language modeling." *arXiv preprint cs/0108005* (2001).

[6] Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. "A neural probabilistic language model." *Journal of machine learning research* 3, no. Feb (2003): 1137-1155.

[7] Mikolov, Tomáš, Martin Karafiát, Lukáš Burget, Jan Černocký, and Sanjeev Khudanpur. "Recurrent neural network based language model." In *Eleventh annual conference of the international speech communication association*. 2010.

[8] Mikolov, Tomáš. "Statistical language models based on neural networks." *Presentation at Google, Mountain View, 2nd April* 80 (2012).

[9] Arisoy, Ebru, Tara N. Sainath, Brian Kingsbury, and Bhuvana Ramabhadran. "Deep neural network language models." In *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever*

ISSN: 2582-2012 (online)

Journal of Artificial Intelligence and Capsule Networks (2020)
Vol.02/ No. 01
Pages: 53-63
http://irojournals.com/aicn/
DOI: https://doi.org/10.36548/jaicn.2020.1.006

Really Replace the N-gram Model? On the Future of Language Modeling for HLT, pp. 20-28. Association for Computational Linguistics, 2012.

[10] Merity, Stephen, Nitish Shirish Keskar, and Richard Socher. "Regularizing and optimizing LSTM language models." *arXiv preprint arXiv:1708.02182* (2017).

[11] Raj, Jennifer S. "A Comprehensive Survey On The Computational Intelligence Techniques And Its Applications." *Journal of ISMAC* 1, no. 03 (2019): 147-159.

[12] Shakya, Subarna. "Machine Learning Based Nonlinearity Determination For Optical Fiber Communication-Review." *Journal of Ubiquitous Computing and Communication Technologies (UCCT)* 1, no. 02 (2019): 121-127.

[13] Joseph, S. Iwin Thanakumar, and Iwin Thanakumar. "Survey of data mining algorithm's for intelligent computing system." *Journal of trends in Computer Science and Smart technology (TCSST)* 1, no. 01 (2019): 14-24.

[14] Raj, Jennifer S., and J. Vijitha Ananthi. "Recurrent Neural Networks And Nonlinear Prediction In Support Vector Machines." *Journal of Soft Computing Paradigm (JSCP)* 1, no. 01 (2019): 33-40.

[15] Graves, Alex. "Supervised sequence labelling." In *Supervised sequence labelling with recurrent neural networks*, pp. 5-13. Springer, Berlin, Heidelberg, 2012.

ISSN: 2582-2012 (online)