



e-ISSN: 2319-8753 | p-ISSN: 2347-6710

IJIRSET

International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 11, Issue 5, May 2022

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 8.118

English-French Neural Machine Translation using LSTM

Sudha Narang¹, Anirudh Garg², Ayush Devrani³, Aviral Gautam⁴

Assistant Professor, Department of Computer Science & Engineering, M.A.I.T Rohini, Delhi, India¹

U.G. Student, Department of Computer Science & Engineering, M.A.I.T, Rohini, Delhi, India²⁻⁴

ABSTRACT: Language translation brings individuals from all over the world together to cooperate, share knowledge, and develop bonds. Compared to rule-based and statistical machine translation, Neural Machine Translation has shown substantial improvement. Neural machine translation has many applications in our day to day life be it in language translation, Sign language translation. An encoder and a decoder are frequently used in neural machine translation models. From a variable-length input text, the encoder extracts a fixed-length representation, from which the decoder creates an accurate translation. The rise of neural machine translation is both practically and theoretically relevant. Traditional statistical machine translation (SMT) models use far more memory than neural machine translation models. Because of their capacity to learn long-term dependencies, long short-term memory (LSTM) units on sequence-based models are employed in translation, question-answering systems, and classification tasks. LSTM networks are producing excellent results on text generation models by learning language models with grammatically stable syntaxes in natural language creation.

KEYWORDS: Neural Machine Translation, LSTM, Natural Language.

I. INTRODUCTION

Machine Translation is a crucial domain in Natural Language Processing because it allows individuals who speak various languages to communicate with one another and share information, therefore eliminating the linguistic barrier. There should be more machine translation so that translated knowledge can be affordable to humans discovered in a flash. The most widely spoken language is English, which is extremely important in the world today. As a result, it is critical to translate information from English to languages other than English.

RNNs (recurrent neural networks) are connectionist models that may transfer information selectively across sequence stages while processing sequential data one element at a time. As a result, they may simulate input and/or output that consists of non-independent sequences of items. Recurrent neural networks can describe sequence and temporal dependencies on different scales at the same time.

However, it has been shown that as the time range of dependence grows, gradient descent of error criterion becomes progressively inefficient. Short sequences perform well with RNNs. However, the gradients cannot be transmitted through many cells in longer sequences. It is called the vanishing gradient problem.

The network is attempting to transmit all of the information across longer sequences, according to one interpretation of the vanishing gradient problem. During BPTT, the gradients increase smaller as the sequences get longer. Long short term memory network (LSTM) is used to solve this problem. By propagating just the relevant information back, LSTMs are able to maintain long-term information. LSTMs are a kind of RNN that has many gates instead of a single activation function in the hidden layer they are the cell states, input gate, forget gate, and output gate. The gates will assess which historical knowledge is relevant to keep and which information must be discarded at each time point. [1]

The input gate limits how much information from the previous state must be forgotten, while the output gate filters the information to be sent on to the next layer. Because of this, LSTMs can maintain the history of long phrases and are widely employed in NLP applications ranging from question-answering systems to machine translation.

II. RELATED WORK

The most widely used NMT model is Ilya et al., 2014[2], which is an encoder-decoder technique with both encoders and decoders built utilising LSTMs. The encoder reads one word at a time from the input sequence, to generate a hidden context vector. This vector is used by the decoder to store the number of words viewed so far, and generate the output sequence.

The model can learn from the past by feeding the input sequences in the order of words. Bidirectional RNN was proposed by Schuster et al.[3], in which the input sequences are trained not only from first to last word, but also in reverse order. This worked well for machine translation tasks, since the grammatical structure of certain destination languages differs from the source language, and training in both directions allowed the model to account for future input terms.

In the Research Paper by Kyunghyun et. al[4] titled "On the Properties of Neural Machine Translation: Encoder-Decoder Approaches", The features of neural machine translation are examined in research utilising two models: RNN Encoder-Decoder and a newly designed gated recursive convolutional neural network. They Demonstrated that The neural machine translation works effectively on short phrases quite effectively without unfamiliar terms, but as the phrase lengthens, its performance suffers and the quantity of unfamiliar terms grows. They also discovered that the gated recursive convolutional network can learn the grammatical structure of a sentence.

In another Paper on the topic of improving the performance of RNN based Encoder Decoder Model by Yan Liu et.al[5] implemented attention mechanism which is used imaginatively to examine each layer of the encoder. Many components of the sequence might be learnt this way. The linguistic properties and semantics information of a word in a phrase, for example, might be clearly recorded and ultimately impact the development of the current translation term. Experiments have demonstrated that their approach is successful, with BLEU scores increasing by an average of 5.67 points.

III. METHODOLOGY

LSTM works by encapsulating meaning and mapping phrases with similar meanings to one another. The LSTM sequence to sequence is aware of the order of words and is extremely consistent with both active and passive meanings. An encoder and a decoder are part of the architecture. The layers of the encoder are as follows:

1. **Embedding layer:** Takes an English word and turns it into a fixed-dimension vector. Layers for embedding can be utilised in a variety of ways. It may be trained independently to learn word embeddings and then utilised with the model, or it can be used as part of the model to learn embeddings with the word.
2. **LSTM layers:** These layers process input word-by-word, resulting in the construction of a fixed-size vector representation of the words viewed thus far. The number of LSTM layers might be more or lower than one.

The decoder, like the encoder, is made up of a stack of LSTM units that decode the output sequence from the context vector and generate the target language phrase. At a given time, it constructs a vector representation of the output word. The LSTM employed on the decoder side is unique in that it is influenced by the input. LSTM becomes a viable alternative for the model due to a considerable delay between the input and output.[6]

Supervised learning is machine learning with parallel corpora. This is where our learning style fits in. The source language is the language that will be translated given the parallel corpus in which sentences will be grouped as pairs. The target language is another language. To begin, both languages' phrases must first be tokenized into tokens. A source phrase would look like this: $X = x_1, x_2, \dots, x_n$, where x_i represents each token.

Dataset and Preprocessing

The corpus is used to extract the training dataset, validation dataset, and test dataset. The dataset file was pretty enormous, thus we cut it into 25000 subsets owing to our environment limits in Google Colab Pro. The Configuration used in Google Colab Pro is Google Compute Engine backend (GPU) with 25.46 GB RAM.

The dataset was divided in 90% and 10 % ratio for training and testing hence 22500 samples were used for training and remaining 2500 samples were used for testing Sentences were splitted into lines. Then in data cleaning Phase We did many steps in order to get better trained model. Some of the steps were Removing Punctuations, converting characters to lowercase, removing tokens having numbers in them, normalizing the unicode characters. The English Vocabulary Size is 3976 with Max Length: 5 and for French Sentences Vocabulary Size is 8050 with Max Length as 12. We also did One hot encoding to max phrase length. The activation function used is Softmax Total params are 10,363,762 having Trainable params as 10,363,762 and Non-trainable params as 0. The patience is set to 10 and the minimum delta is set at 0.01. The model is fitted in batches of 25 and epochs is set to 50. The model terminated training after 18 epochs since the accuracy didn't increase any further based on the parameters supplied into the function.

IV. EXPERIMENTAL RESULTS

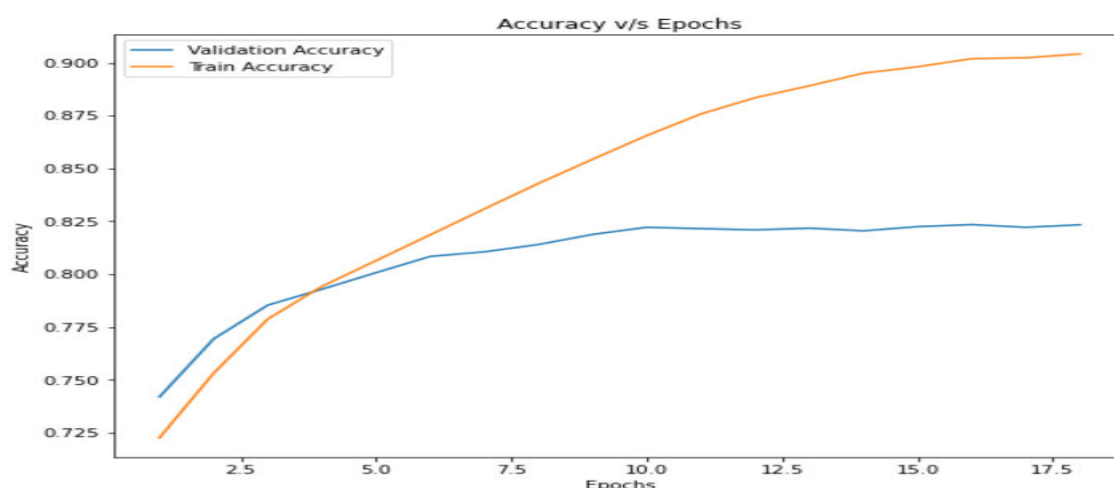


Fig 1 Accuracy v/s Epochs

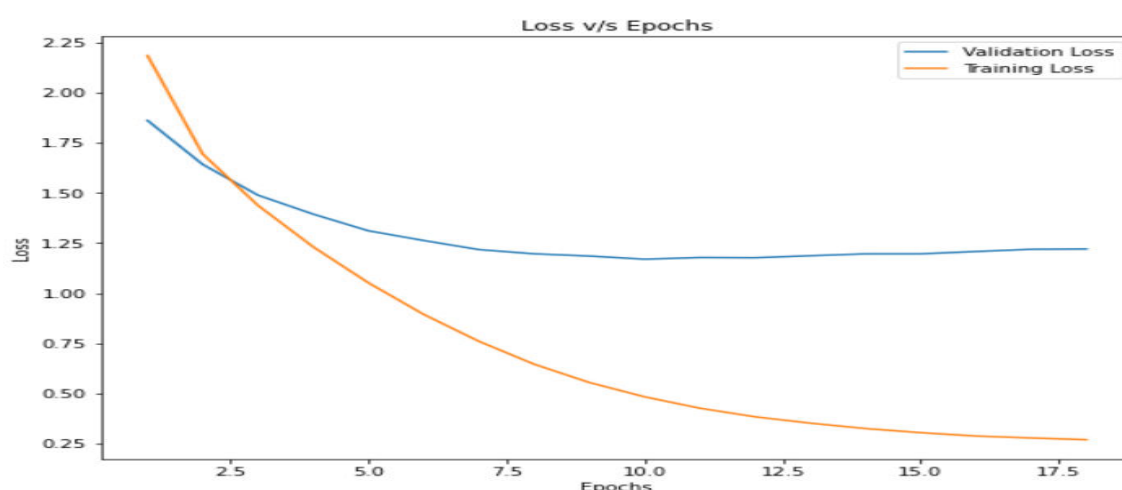


Fig 2 Loss v/s Epochs

BLEU, F-measure, WER, accuracy, perplexity, and other assessment techniques are applied in this field of study. Metrics like BLEU, NIST, and F-measure might be utilised to assess against a reference data in downstream activities like machine translation or question-answering. We utilised the Bilingual Evaluation Understudy (BLEU) score metric for evaluation and comparison. The BLEU score range is [0, 1], Although the BLEU score is widely accepted as a de

facto benchmark for measuring machine translation system performance, it is not perfect. By looking at the translations we can say with confidence that it did a good job translating, particularly short lines. When the source phrases are lengthy, however, the neural machine translation models' performance degrades.

BLEU SCORES	Training	Testing
BLEU 1-gram	0.79	0.50
BLEU 2-gram	0.73	0.38
BLEU 3-gram	0.69	0.31
BLEU 4-gram	0.54	0.19

TABLE 1. COMPARISON OF TRAINING AND TESTING DATA**V. CONCLUSION AND FUTURE WORK**

We have made our LSTM model and got good BLEU score especially for shorter sentences. Although on the longer sentences accuracy decreases drastically which can be improved in future work by using more advanced models such as

Attention based models like Transformers can be used

We can also do a Research on Comparing our model with Transformer

Future research can also be done on the syntactic, semantic, and morphological aspects of both languages in order to increase the accuracy. We'd also look at how well this model performed in other Indian languages, because the grammar and syntactic structures of many Indian languages are similar. We also anticipate using gated convolutional neural networks into cutting-edge transformer topologies.

REFERENCES

- [1] . S. Santhanam, "Context Based Text-Generation Using LSTM Networks," 4 2020. [Online]. Available: <https://arxiv.org/abs/2005.00048>.
- [2] . I. V. a. Q. V. L. Sutskever, "Sequence to Sequence Learning with Neural Networks," Advances in neural information processing system, 2014.
- [3] . M. S. a. K. k. Paliwal, "Bidirectional recurrent neural networks," Signal Processing, IEEE Transactions, vol. 45, no. 11, pp. 2673-2681, 1995.
- [4] . B. v. M. . D. B. . Y. Kyunghyun Cho, "arxiv," 10 2014. [Online]. Available: <https://arxiv.org/abs/1409.1259>. [Accessed 14 05 2022].
- [5] . D. Z. G. q. Yan Lie, "A simple but effective way to improve the performance of RNN based Encoder In Neural Machine Translation Task," 2019 IEEE fourth international conference on Data Science in CyberSpace, pp. 416-421, 2019.
- [6] . A. S. A. . S. . R. K. Gaurav Tiwari, "English-Hindi Neural Machine Translation-LSTM Seq2Seq and ConvS2S," International Conference on Communication and Signal Processing, pp. 871-875, 2020.



INNO  SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.118



ISSN  INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 9940 572 462  6381 907 438  ijirset@gmail.com



www.ijirset.com

Scan to save the contact details