Machine Translation Systems for English Captions to Hindi Language Using Deep Learning

Arvinder Singh, *Ninad* Bhase, *Manav* Jain, *and Tushar* Ghorpade

Department of Computer Engineering
Ramrao Adik Institute of Technology
D.Y. Patil Deemed to be University
Nerul, Navi Mumbai, India.

Abstract. Machine Translation is the process of translating text from one language to another which helps to reduce the conversation gap among people from different cultural backgrounds. The task performed by the Machine Translation System is to automatically translate between pairs of different natural languages, where Neural Machine Translation System stands out from all because it provides fluent translation along with reasonable translation accuracy. The Convolution Neural Network encoder is used to find patterns in the images and encode it into a vector that is passed to the Long Short Term Memory decoder which finds the caption word-by-word to best describe the image. Upon reaching the end-line token, the entire description of the image in English is generated and that is our output for that particular image. Automatically creating the description of an image in English using any natural language sentences and then translating it using Neural Machine Translation to Hindi is a very challenging task. It requires expertise in both image processing as well as natural language processing. In this paper, the aim is to compare the two Machine Translation Systems: Google Translation System and the proposed Neural Machine Translation System to convert the text obtained from an image in English to Hindi language.
Keywords - Image captioning, English language, Hindi language, encoder-decoder framework, Neural network, Machine translation.

1. Introduction

English is the universally used language and Hindi is the national language of India, as per the analysis, English to Hindi machine translation system is very important to enhance the knowledge of Indians without any language barrier. Machine translation [1] is an application of Natural Language Processing (NLP) which aims to fill the communication gap between different parts of society. Machine Translation works with large amounts of source and target languages that are compared and matched against each other by amachine translation engine. The benefit of Machine Translation is that it is possible to translate large amounts of text in a very short time.

Further the paper is divided as follows: - in the section 2 we discuss about the Literature Survey, section 3 includes the problem definition of proposed system, section 4 contains the details of the Proposed System, section 5 shows the Results and Analysis for the proposed system and in the section 6 the paper is concluded and discusses about the future work.

2. Literature Survey

In this section, many research papers were reviewed that are revolving around the topic. From these papers, the main focus was on the methodologies used and their brief descriptions along with their limitations.

The research paper by Lucia Benkova and Lubomir Benko on Neural Machine Translation as a Novel Approach to Machine Translation aims to introduce a novel approach i.e., Neural Machine Translation. The Neural machine translation structure is developed on the coding frame. The encoder converts the sentence in the source language into a continuous space representation with a recurring neural network. The results from this paper reveal that neural machine translation system provides with a high accuracy result but still needs to be assessed in the future.

*Corresponding Author : arvindersingh0912@gmail.com

By referring to the research paper written by Ankit Rathi on Deep Learning Approach with Image Captioning in Hindi [2], it is seen that the model used in this paper is trained to predict the image caption using the image feature vector and the previous word. This paper has tested their model on four databases and concluded that image definition quality increases after training the model with a pure database. This study gives an idea of how to generate captions in Hindi using a coding model. It is clear from their experiment that a model with a single definition defined for each image produces high quality captions.

The Literature Survey involves studying about many machine translation systems like Rule-Based Machine Translation (RBMT) [3], Corpus-Based Machine Translation (CBMT) [4], Hybrid-Based Machine Translation (HBMT) [5], Neural Machine Translation (NMT) [6] along with Google Translation System (GTS) [7]. The paper published by Sabine Hunsicker, talks about HBMT being a combination of two or more machine translation techniques where the advantages of the individual techniques are combined to achieve an overall better translation. This paper points out that HBMT has a drawback that it needs extensive editing and human translators are also required sometimes. To learn about CBMT, a research paper published by B. Premjith, M. Anand Kumar and K.P. Soman was very helpful. According to this paper, Statistical Machine Translation (SMT) comes under CBMT and the advantage of SMT is that translation can be done without linguistic language. Still, creation of parallel corpus is one of the drawbacks of SMT. Daniel Torregrosa and Nivranshu Pasricha in their research paper suggest that RBMT can achieve high accuracy within narrow subsets of language. Although RBMT does not require much data, it needs expert human involvement along with many rules in order to improve the quality which results in a very complex system. In order to grasp the concept of Google Translation System, research paper published by Santosh Kumar Mishra and Rijul Dhir was very informative. This paper talks about using a dataset in English Language and then translating it into Hindi Language. Accuracy of GTS is mainly dependent on the source and language pairs. However, Google translation sometimes creates translations which are grammatically incorrect as a result the meaning of these translations are lost.

3. Problem Definition

This proposal aims to define, or rather in appropriate terminology, “formulate” the problem statement as follows: The concept behind machine translation is to generate grammatical rules for the source and target language. Machine Translation works on those set of rules and acts as a kind of translation between languages. The problem that arises is the addition of new content, language pairs because maintaining and extending such a set of rules was too tedious and expensive. Aim is to work on a comparative study between two Machine Translation Systems: - Neural Machine Translation and Google Translation System (GTS).

4. Proposed Methodology

4.1 Proposed Work:

The main motive is to look into generating captions in English and then converting it to Hindi with the help of Machine Translation System Viz. Neural Machine Translation System and Google Translation System. Further on the basis of different BLEU scores analyzing the standard of the existing Machine Translation output.
In addition, natural language should define semantic knowledge i.e., to develop language models in terms of visual comprehension.

4.2 Implementation Details:

For Image Captioning, we host ResNet50 [8], which is a deep neural network with 50 layers. The deeper the neural network, the more difficult it is to train. The ResNet50 architecture facilitates network training and allows them to be more in-depth, leading to increased performance in various tasks.

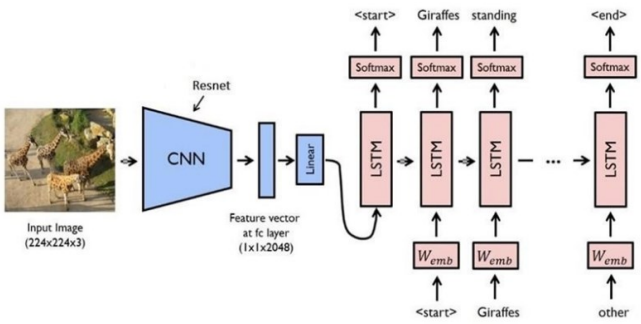

Fig. 1. Image Captioning Architecture [9]

To find the captions for the images, the encoder decoder framework model is used as shown in Fig.1. Here the system uses the ResNet50 model as an encoder to produce a vector containing the image features. After that, LSTM is adopted as a language decoder model to decode the vector into a sentence. Here, image is taken using OpenCV [6]. It gives image in BGR format but ResNet50 only accepts in RGB format. Also, ResNet50 only accepts image of size 224x224. After this pre-processing of image, it is loaded into the ResNet50 model. ResNet50 model produces 3D vector of 2048 which contains all the image features. Typically, a pre-trained CNN removes features from our installation image. The feature vector is converted linearly to the same size as the input LSTM network input. This network is trained as a language model in the vector tool. LSTMs have become an innovative solution for problems related to sequencing and time series, but LSTMs are difficult to train. Even the simplest models require a lot of time and

system resources. Due to limitations of SMT, translation with the help of Neural Network was introduced to deal with problems of accuracy and ability to analyse context. Here a Feed-Forward Neural Network is used to calculate the result of phrase pairs by considering fixed size phrases. However, this is not always the case as the length of phrase may change depending on the scenario. It is observed that small phrases perform well when Recurrent Neural Networks (RNN) is used for neural machine translation. For lengthy sentences Statistical Machine Translation is preferred. It is tough to control complicated context-dependent cases by RNN. Hence, RNN combined with LSTM has potential to recollect long-term features. The other aspects which enhance the effectiveness of NMT system are stacked RNNs, test-time decoding using beam search, input feeding using attention mechanism.

The implementation details are as follows:

• Data Analysis Model: It is implemented by the use of the environment provided by Google Colab for data analysis and machine learning, which is a perfect online platform to analyse/train/share data of system with the bindings as per required. The model is implemented in this environment using key Python libraries.

• Model Analysis: It is implemented by the use of the environment provided by Python IDLE- 3.7/Anaconda/Spyder for data analysis and machine learning, which is a perfect platform to analyse/train/share data of system with the models as per required. The model is implemented in this environment using key Python libraries.

• GUI Web Application: It is implemented by the use of the environment provided by Flask for Python HTML5 for web designing and ML based API construction, which is a perfect online platform to present/showcase the system with an interactive interface for the User. The model is implemented in this environment using key aspects of Flask, and an online form is constructed using HTML5 with the form connected to the project.

4.3 System Design

4.3.1 Image Captioning

The image captioning is divided into two logical modules-one image based-model-it is used for feature selection of the images, and two language model-it converts those features of image into natural sentence. CNN is used for image feature extraction and RNN is used for caption generation. ResNet stands for Residual Network. It is the new terminology that is introduced in residual learning. There is sequence of breakthrough Deep Convolution Neural Networks in the field of image captioning. ResNet model has a different model called as ResNet50 having 50 layers which contains 48 Convolution layers, 1 MaxPool and 1 Average Pool layer. Basically, in a deep convolutional neural network, various layers are stacked together and trained to perform a task. The network learns various features at end of its layers. In residual learning, instead of trying to learn certain aspects, it tries to learn some residuals. Long Short Term Memory (LSTM) [4] is an advanced RNN model that was designed to read a long list of sentences that did not occur in normal RNN. The LSTM model will learn a function that creates a sequence of previous recognition as input detection output. Thus, the sequence of views must be converted into multiple examples where LSTM can learn. The NMT system is trained using sentence pairs directed at the English and Hindi source. The parallel corpus contains 40,455 cases, including data sets from a various source.

• This technique mainly uses in-depth reading and deep learning to produce new captions. In this method the image features are first extracted using a CNN model and then it creates captions for the images using the LSTM model.
• Use CNN as an image editing encoder and LSTM as a decoder to generate descriptive sentences.
• This method uses image data sets and their meaning in the native language and requires a word-for-word connection between the description and the visual data.
• The first model aligns pieces of sentences with visual cues, and then creates a single meaning with multiple embedding. This definition is treated as learning data for a second model of a repetitive neural network that has learned to detect captions. The proposed Machine Translation Systems adopted the concept of Open NMT in NMT-1 and GTS in NMT-2 system respectively –

• The system has trained the Neural Machine Translation System 1 & 2 with the help of English to Hindi parallel training corpus to obtain the required models. So, for training purpose of the models, two test cases were taken under consideration.

• Both the systems were then re-trained with the help of the same corpus. Then with increasing the number of iterations and testing the model, it was saved, thus acquiring the predicted results at thirteen and fourteen for machine 1 and 2 respectively.

• The last step was to validate the results obtained from both the systems using the sample images to check the confluence of the training process.

4.3.2 Machine Translation

Machine Translation (MT) is the process of converting a source language into target language without any human entanglement such that the meaning of the source language is preserved in the translated text.
With a large amount of research available in the present-day, several types of NMT are already being looked into and being established in the industry. The Encoder-Decoder model is the most popular kind of the NMT model. The NMT model aims to take an individual sentence as an input and convert this sentence

to a different language which is the final output as shown in Fig.2.

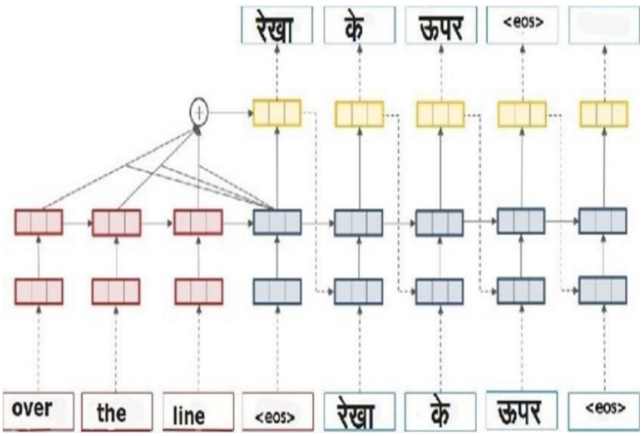

Fig. 2. System Design [3]

The encoder takes a phrase from the input language and develops a vector using this phrase. This vector then holds on to the meaning of the phrase and dispatches it to the decoder which manifests the translation into the target language. The decoder issues predictive sentences of variable sizes, where the decoder continues to predict words till it reaches the <EOS> tag.

Once the <EOS> tag is detected, the work of decoder is done and the sentence is completely translated. The translation is more exact when Attention mechanism concept is used in NMT.

In this concept basically at every step, the decoder creates a memory vector by recalling all the encrypted vectors of the encoder. For prediction of the following word in the translated sentence, it uses the memory vector along with hidden vector in the decoder. In this process, decoder only uses crucial details from the encoder which would otherwise cause damage.

5. Results And Analysis

5.1 Dataset

A cluster of picture-based captions, consisting of more than 8,000 images as shown in Fig 3. each paired with five different captions that provide clear descriptions of important businesses and events.

Images are selected from six different Flickr groups, that usually doesn't contain any celebrities or locations, but are hand-picked to reflect different scenarios.

In this proposed system, the dataset has been prepared by combining the Flickr8k English and Hindi Caption dataset [10]. This dataset consists of three columns: source, english_sentence and hindi_sentence.

```
images_path = '/content/drive/MyDrive/Images/'
images = glob(images_path+'*.jpg')
len(images)
```

```
8091
```

Fig. 3. Total number of images in the dataset

The source column consists of all the image names, english_sentence column contains five different descriptions of the source image in English and the hindi_sentence column consists of the corresponding descriptions of these images in Hindi as shown in Fig.4.

```
eng_hin=pd.read_csv('/content/drive/MyDrive/english_hindi_dataset.csv')
eng_hin.head()
```

	source	english_sentence	hindi_sentence
0	1000268201_693b08cb0e	A child in a pink dress is climbing up a set o...	गुलाबी पोशाक में एक बच्चा प्रवेश के रास्ते में...
1	1000268201_693b08cb0e	A girl going into a wooden building .	एक लड़की लकड़ी की इमारत में जा रही है।
2	1000268201_693b08cb0e	A little girl climbing into a wooden playhouse .	एक छोटी लड़की लकड़ी के प्लेहाउस में चढ़ गई। ...
3	1000268201_693b08cb0e	A little girl climbing the stairs to her playh...	एक छोटी सी लड़की अपने प्लेहाउस में सीढ़ियाँ चढ़...
4	1000268201_693b08cb0e	A little girl in a pink dress going into a woo...	एक गुलाबी पोशाक में एक छोटी लड़की एक लकड़ी के ...

Fig. 4. English-Hindi Dataset

```
eng_hin.dropna(inplace=True)
eng_hin=eng_hin[:50000]
eng_hin.drop(['source'],axis=1,inplace=True)
eng_hin.shape
```

```
(40455, 2)
```

Fig. 5. Total number of captions in the dataset

Fig 5. shows that there are a total of 40,455 captions for both English and Hindi languages in the dataset for 8091 images. The accuracy of the model is shown in Fig 6.

```
Epoch 47/50
189/189 [==============================] - 33s 173ms/step - loss: 1.0620 - accuracy: 0.7308
Epoch 48/50
189/189 [==============================] - 33s 174ms/step - loss: 1.0243 - accuracy: 0.7409
Epoch 49/50
189/189 [==============================] - 33s 173ms/step - loss: 1.0024 - accuracy: 0.7451
Epoch 50/50
189/189 [==============================] - 33s 173ms/step - loss: 0.9732 - accuracy: 0.7518
<keras.callbacks.History at 0x7f1b5b24c0d0>
```

Fig. 6. Corresponding Accuracy and Loss of the trained model at Epoch value 50

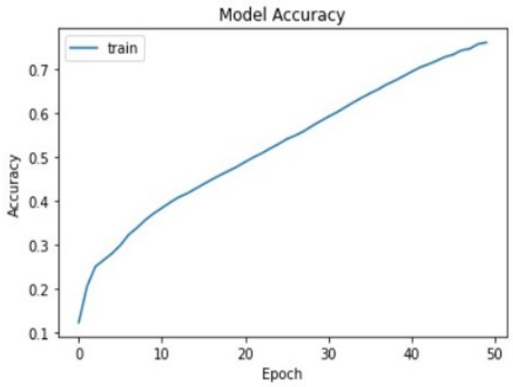

Fig. 7. Epoch vs Accuracy Graph

From Figure 7 it's seen that, as the Epoch value increases, the accuracy of the model increases simultaneously.

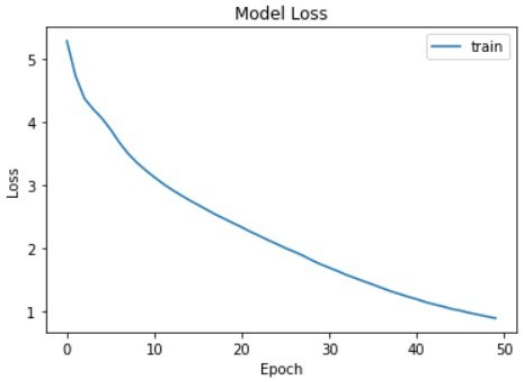

Fig. 8. Epoch vs Loss Graph

From Figure 8 it's seen that, as the Epoch value increases, the loss of the model decreases simultaneously.

5.2 Results on comparison of NMT and GTS

Table 1. Comparison of the proposed NMT Model with GTS

Input Image	Neural Machine Translation (NMT)	Google Translation System (GTS)
	एक पर्वतारोही एक बर्फ से ढंके पहाड़ पर चढ़ते हुए एक पेय लेने के लिए रुकता है।	एक पर्वतारोही एक स्नोसूट गियर पहनता है जो एक पहाड़ के सामने है।
	एक लाल ट्रक एक चट्टानी सतह पर चल रहा है।	एक लाल जीप एक चट्टानी सतह के निशान क्षेत्र पर लटकी हुई है।
	एक व्यक्ति एक चट्टानी पहाड़ी के नीचे एक बाइक की सवारी करता है।	एक आदमी अपनी पथरीली राह पर चलता है।

Table 1 shows a few examples of the input image along with its caption in English and the translated text in Hindi Language.

5.3 About BLEU Score

The result which are being expected, are based on evaluation metrics of BLEU score i.e., Bilingual Evaluation Understudy Score. In BLEU score, it does the comparison between reference caption and generated caption. So, the evaluation of the Machine Translation can be done on the entire dataset, however here evaluation is done on generated captions i.e., sentences, which means that comparing calculations will be done phrase by phrase. To calculate the adjusted accuracy of the test corpus, the total n-gram count is added for all candidates, and it is divided by the n-gram number of test corpus. A BLEU score value of 0 means that the translated output and the reference translation are not overlapping i.e., the translated text is of low quality. A BLEU score value of 1 means that there is perfect overlap with the reference translation i.e., the translated text is of high quality. It is observed that even human translators cannot achieve a perfect score of 1.0.

Table 2 shows the three sentences from our training dataset which we took under consideration for now to compare the BLEU Score of both, the proposed NMT and GTS models.

Table 2. BLEU Score of the proposed NMT and GTS model

Neural Machine Translation (NMT)	Google Translation System (GTS)	BLEU Score
एक पर्वतारोही एक बर्फ से ढंके पहाड़ पर चढ़ते हुए एक पेय लेने के लिए रुकता है।	एक पर्वतारोही एक स्नोसूट गियर पहनता है जो एक पहाड़ के सामने है।	0.242
एक लाल ट्रक एक चट्टानी सतह पर चल रहा है।	एक लाल जीप एक चट्टानी सतह के निशान क्षेत्र पर लटकी हुई है।	0.306
एक व्यक्ति एक चट्टानी पहाड़ी के नीचे एक बाइक की सवारी करता है।	एक आदमी अपनी पथरीली राह पर चलता है।	0.626

6. Conclusion

The proposed Neural Machine Translation system will be used for English to Hindi translation and this system will be compared with the Google Translation System with respect to the BLEU score. After exploring the results, it can be concluded that Machine Translation based on neural network needed improvements when it comes to recognizing unspecified words, repetition of words and varied translation of the generated phrase.

In addition, it is seen how total loss changes during training. As the value of epoch increases i.e., the model is being trained a greater number of times, it is seen that the total loss of model decreases. As seen from the results, the accuracy of the model increases over the time.

References

1. Benková, Lucia & Benko, Ľubomír. (2020). Neural Machine Translation as a Novel Approach to Machine Translation, (2020).

2. A. Rathi, "Deep learning apporach for image captioning in Hindi language," 2020 International Conference on Computer, Electrical & Communication Engineering (ICCECE), 2020, pp. 1-8, doi: 10.1109/ICCECE48148.2020.9223087, (2020).

3. Aspects of Terminological and Named Entity Knowledge within Rule-Based Machine Translation Models for Under-Resourced Neural Machine Translation Scenarios. (2020)

4. B., Premjith & Kumar, M. & Kp, Soman. (2019). Neural Machine Translation System for English to Indian Language Translation Using MTIL Parallel Corpus: Special Issue on Natural Language Processing. Journal of Intelligent Systems. 28. 10.1515/jisys-2019-2510. (2019).

5. Nair, Jayashree & Krishnan, K & Deetha, R. (2016). An efficient English to Hindi machine translation system using hybrid mechanism. 2109-2113. 10.1109/ICACCI.2016.7732363. (2016).

6. P. Anderson, X. He, C. Buehler et al., "Bottom-up and top-down attention for image captioning," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, June (2018).

7. Laskar, Sahinur Rahman et al. "Neural Machine Translation: English to Hindi." 2019 IEEE Conference on Information and Communication Technology (2019): 1-6. (2019).

8. JalFaizy Shaikh. "Automatic Image Captioning using Deep Learning (CNN and LSTM) in PyTorch", (2018).

9. K. Loganathan, R. Sarath Kumar, V. Nagaraj, Tegil J. John, CNN & LSTM using python for automatic image captioning, Materials Today: Proceedings, 2020, ISSN 2214-7853, doi.org/10.1016/j.matpr.2020.10.624. (2020).

10. https://github.com/manavjain179/Machine-Translation. (2022)