IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.67

Volume 5, Issue 4, February 2025

# Comparing CNN Architectures and Image Transformer for Hindi Image Captioning

**Aparna Paliwal**
Department of Computer Science
Banasthali Vidyapith, Banasthali, Rajasthan, India

**Abstract:** *In this paper, we have shown the development of some image captioning that provides the textual description of an image. We have shown how Image captions in Hindi can be generated using an English Image captioning dataset. The Flickr dataset was used for training the model for image captioning and English-Hindi parallel corpus was used across domains for training the Neural Based Encoder-Decoder model. For our study, we developed a model based on sequence-to-sequence architecture. We trained two image captioning models based on VGG16 and RESNET50 architectures which generated English captions and were then translated into Hindi using seq2seq NMT model. The systems were evaluated using the standard BLEU MT evaluation metric. It was found that Image Transformers performed better then RESNET50 and VGG16 based models.*

**Keywords:** Multimodal Data, Image Captioning, Machine Translation, Encoder-Decoder Model

## I. INTRODUCTION

The capacity to perceive picture highlights for producing literary depictions is for different assignments in picture handling and Computer Vision. Henceforth, picture inscribing is an interdisciplinary research issue that stands between Computer vision and natural language processing (NLP). Its objective is to generate a textual description of the image. The typical model has been used for captioning an image i.e. is an en- coder-decoder deep network, where the encoder captures the essence of an image while the decoder is responsible for generating a sentence describing the image. The signifi- cance of the research is to develop an image caption generation model for image de- scription and its text in Hindi. Image Captioning is interesting to us because it is con- cerned with what we understand about perception concerning machines.

Over the decades, the world's Hindi-speaking population has increased from 250 million to 400 million. Concurrently, the population of English speakers has risen from 300 million to 470 million. Hindi with 260 million speakers is found to be the fifth most widely spoken language on the planet (after Chinese, Spanish, English, and Arabic). Hindi is a morphologically rich language (MRL), e.g. the gender category is reflected in the forms of nouns, verbs, and also adjectives (Sreelekha S and Bhattacharyya, 2017). It is a simple task for a human, yet very testing assignment for a machine since it not only includes understanding the content of a picture but also, how to decipher this com- prehension of context into regular language. Subsequently, we have to discover how Deep Neural Network (DNN) models can be utilized to straightforwardly create de- scriptions for pictures, for example, pictorial articles, and photos (Fig.1).

The rest of the paper is as follows: section 2 discusses the work done in this area. Section 3 explains the experiment setup. Section 4 explains the proposed methodology. Section 5 evaluates the results of the captions produced. Section 6 concludes the paper.

DOI: 10.48175/568

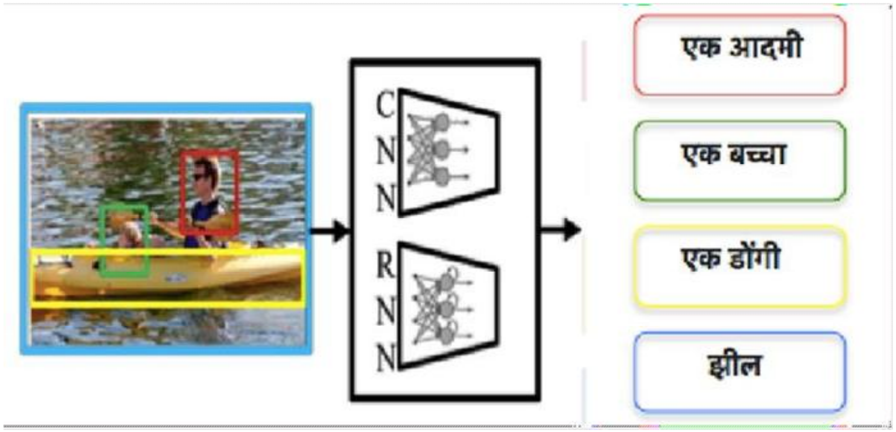

Figure 1: Image Captioning Example

## II. LITERATURE SURVEY

Li & Sun (2005) developed automatic image notation with absolute techniques that provided semantic image retrieval using textual descriptions. A maximum entropy based model approach has been proposed for this task. Automatic image notation is a process where automatically image labeling of contents is done with a set of arranged keywords that are further used to have image semantics. Compared to other traditional methods, Automatic image notation outperformed.

Hsieh et al. (2005) presented a model named 'PCFG Parser' which was used to check proficiency of the grammar. Grammar binarization approach was introduced to evaluate the efficiency of the documentations. To maintain precision level of grammar linguis- tically driven feature was added to the grammar rules book. Further, while testing it was observed how to deal with the error that had arisen due to word segmentation and part of speech labeling. Hence, "Grammar Blending" technique was introduced to deal with similar grammatical errors.

Li & Su (2006) targeted at the challenge faced earlier, where results of annotation performance were unsuccessful due to the lost meaning between the visual content and the text. Hence, they proposed an idea of including lexical semantics into image-anno- tation process. Data retrieval was being used for semantically constrained K-means with classified corps. At certain phase they have described image without label, key- words that are most commonly used are forecasted by the resultant probability of each keyword. Every keyword is given visual term at the first-level classifier ensemble, and refining of annotation quality for the words that are extracted from the keyword anno- tated in the training set of images is done at the second-level language model.

Caicedo et al. (2012) stated that the massive collection of images could be collected from the Web. Thereafter, the collection of images is incorporated to data without im- age such as text, ratings given by users, comments, and tags. Surplus modalities of data like these provided a framework to the visual content that improvised various contents of image analysis cases. Hence, they proposed a method with a new approach of using non-negative matrix factorization for designing multimodal image representation. This approach is used to generate latent semantic factors where data from different modali- ties can be correlated. They involved two data models in their work: one for visual features and the other for text terms. Their objective was to reduce the existing semantic ambiguity of the visual content. As per their further analysis, they examined that strat- egies of matrix factorization could be best for multi-modal latent factors also, multi- modal image representations could be useful to perform multiple image collection anal- ysis tasks.

Li et al. (2012) presented "Content Based Image Retrieval" (CBIR), which was stud- ied and researched for past many years. It was observed that the performance of CBIR was not appropriate, due to the unusual semantic gap. CBIR was used to extract images as per its visual effects, whereas users often prefer to use in-built text based image searching. Hence, they proposed probabilistic latent semantic analysis (PLSA). On ad- dition to this, to determine the model parameters they designed Expectation-Maximi- zation (EM) algorithm. They had also presented a hybrid framework, for PLSA to model features of images in a propagative learning stage. This was used to set combi- nation of classifiers that chain to classify the multi-label data in a discriminative stage of learning. As the framework was designed to take generative and discriminative learn- ing in a combination, that could forecast semantic annotation for unknown images with precision. The outcomes of experiment showed that the approach used outperforms many modern approaches.

Gong et al. (2014) said Multi-label image annotation was found to be the most chal- lenging computer vision task having various real-world applications. As per their ex- isting techniques, for multi-label annotation they used conventional visual features, and Deep Neural Networks (DNN), which have shown potentially a significant boost performance. Their study had proposed ranking and training deep CNNs for multi-label image annotation problems. They investigated several different loss functions based on ranking for training CNN, and found that the weighted approximated-ranking loss works well for problems with multi-label annotation. They also performed experiments on the open source multi-label image dataset NUS-WIDE. They demonstrated that the

significance of using top k-ranking network training. In the future, they would like to use very large amount of noisy multi-label images from internet network training.

Socher et al. (2014) represented that while working on RNNs (Recursive Neural Networks) models can be used to generate compositional feature vectors to represent and classify sentences or images accurately. However, the sentence vectors of previous models were not able to accurately represent visually authentic meaning. Hence, they proposed DT-RNN model. This model utilizes dependency trees to collaborate sen- tences in a vector space for image retrieval that are described by those sentences. DT- RNNs normally focus on the action and agent composition of a sentence. This model outperformed as compared to other types of neural networks.

Kiros et al. (2014) proposed “Multimodal Neural Language Models”. This model described a first step towards tackling the earlier described modeling challenges. They had described two novel language models in the context of image-text learning, word representations and image features. Their models were designed to obtain improved Bleu scores to existing approaches for sentence generation; this outperformed a strong baseline of bag-of-words model.

Vedantam (2015) has designed a new automatic evaluation protocol which is con- sensus for evaluating image description (CIDEr). This practice has been used to enable an objective where machine generation approaches are compared based on their “hu- man-likeness”. This task was done without making random calls on weighing content, grammar, saliency, etc. comparison to each other. They have developed an annotation model that is used for measuring consensus, also designed a metric CIDEr that is capa- ble of computing consensus automatically with two datasets PASCAL-50S and ABSTRACT-50S with 50 sentences per image. It was demonstrated that CIDEr had improved accuracy as compared to existing metrics for measuring consensus.

Mao et al. (2015) presented a “Multimodal Recurrent Neural Network” (m-RNN) model for creating image captions. This model was straight-forward and was used to represent the probability distribution to generate a new word from the previous words given along with the image. Image captioning was performed in accordance with given distribution. The model had two sub-networks: a deep recurrent neural network (Deep RNN) for sentences and a deep convolutional network (Deep CNN) for images. Deep RNN and Deep CNN form whole m-RNN model when they interact in a multimodal layer. Results showed that m-RNN connected images and sentences in a powerful way. It proved to be flexible to more complex image representations as well as sophisticated language models.

Vinayls et al. (2015) presented a model, which was based on a deep recurrent archi- tecture. This had lead to the joint study of machine translation and computer vision, which was used for image description to normal sentences. The model introduced NIC (Neural Image Caption), which was based on end-to-end neural network. This consisted of a Convolutional Neural Networks based computer vision, which followed language generation, RNN (Recurrent Neural Networks). Complete sentences were generated from a given image in the natural language.

Chen &Zitnick (2015) explored the possibility of bi-directional mapping between images and their sentence-based descriptions. Their approach faced a difficulty in Re- current Neural Network (RNN). It built a visual representation of the scene dynamically

that was being created or read as a caption. The representation was to learn and remem- ber long-term visual concepts. The model generated novel captions of given images, and recreated visual features from a given image description. Evaluation was performed on various tasks, which included generation of sentences and retrieval of sentences and images. The results were declared for generating unique image descriptions. When compared to human generated captions, the automatically generated captions were equal to or preferred by humans 21.0% of the time. Outcome was far better than the image and sentence retrieval tasks. These tasks were applied for using similar visual features.

Mao et al. (2015) illustrated multimodal Recurrent Neural Network (m-RNN) model to generate appropriate image captions. This framework performed 3 tasks: sentence generation, sentence obtained from given image and image predicted from given query sentence. The model is a combination of a deep RNN, a deep CNN and 2 sub-networks that interact with each other in a multimodal layer. Results generated by them have proved that both predicting image and sentence can be easily done in complex image and advance language models.

Caglyan et al. (2016) represented the system developed by LIUM and CVC who contributed in two proposed tasks for WMT 2016 Multimodal machine translation and multimodal image description. Here, translation of an English sentence was being de- scribed from an image into German (where both text and image is given). Model used for performing experiment is phrase-based system and Neural MT system.

Calixto et al. (2016) proposed a Doubly-Attentive MMT model. This model was used to decode image descriptions. It used the source language descriptions as well as the images. As compared to the previous results, multimodal attention was more com- plex. Hence, they integrated Attention-based neural MT which would translate and gen- erate image description in a combined model. In this, two separate mechanisms of at- tention were operated on the language and visual modalities. It was an ethical method of learning source words and image regions followed by generation of words in the target description. It was found that when data volume in multimodal translation da- tasets was increased, neural models were more competitive.

Shah et al. (2016) stated submission of The University of Sheffield's for a new WMT16 Multimodal Machine Translation shared task. Here, the model was trained to generate description of images in an expected target language. An image and one or more descriptions are given as inputs in source language. When participated in Model 1, which was taken from a description in source language and further translated in the target language with the help of image information. They trained some methods that were baseline phrase-based SMT systems. The objective of the evaluation was to find a way to use image features to re-rank the n-best list produced by the SMT model, to improve the performance by grounding the translations on images. They were capable of providing successful text-only baseline system for both directions.

Elliott et al. (2016) introduced the "Multi30K" dataset to stimulate multilingual mul- timodal research. Multi30K was an extension of the Flickr30K dataset. This study did two tasks, first, German translations were created by professional translators over a subset of the English descriptions, and second, descriptions were crowd sourced inde- pendently of the original English descriptions.

Hitschler et al. (2016) demonstrated that multimodal pivots into a target side retrieval model to improve SMT of image descriptions by multimodal pivots. These were de- fined in visual space. Their idea was to incorporate image retrieval as compared to col- lection of images and generate the caption in the target language, and also use the cap- tions of the most similar images for cross lingual re-ranking of translation outputs.

You et al. (2016) have proposed a new model. In previous experiments either top- down approach or bottom-up approach is used. Top-down approach extracts objects of image and converts into words, where bottom-up approach describes words along with image and its concluded accordingly. Here they have proposed new algorithm which uses both approaches through a training model with semantic attention. Fusion of these strategies in extracting various information from image and then selectively joined with RNN that computes results with both local and global information for appropriate cap- tion.

Caglayan eta al. (2017) described two models: Mono and Multi-modal NMT systems that were developed by LIUM and CVC for WMT17 Shared Task on Multimodal Translation (MT). Here baseline Neural Machine Translation NMT was a variant of attentive encoder-decoder with a Conditional GRU (CGRU) decoder. The LIUM-CVC system was used for Multimodal Machine Translation (MMT) evaluation campaign of English to German and English to French. The final systems designed were ranked for above language pairs with respect to the automatic evaluation metrics METEOR and BLEU.

Wang and Chan (2017) designed a framework that uses CNN to generate captions. The CNN+CNN model was used for image captioning. The objective was that CNN extracts feature from the image and language, also CNN is applied to model the sen- tence. The attention module and prediction module fuse the information from image and word contexts. The convolutions in the language CNN use causal filters (depending only on the present and past words), to enable feed-forward sentence generation. Re- sults have explored the influence of the kernel width and the layer depth of the language CNN. They have shown that the ability of the CNN-based framework is competitive to LSTM-based models, but can be trained faster. This also visualizes the learned attention maps to show that the model is able to learn concepts and pay attention to the corre- sponding areas in the images in a meaningful way.

Zang et al. (2017) described the NICT-NAIST system in WMT 2017 that had shared MMT task for two language pairs, i.e. English-German and English-French.

Ren et al. (2017) had come up with decision making framework for image caption- ing. They are using two networks for identifying exact word for an image, i.e. 'policy network' and 'value network'. Algorithm is based visual-semantic embedding. In this model is trained using deep reinforcement learning. Policy network is a combination of CNN and RNN, where value network is a combination of a CNN, a RNN and a MLP. Policy network prediction is considered as local guidance and value network. They both are used to generate captions. Best results are evaluated using reinforcement learning approach with visual-semantic embedding.

Pedersoli et al. (2017) introduced an attention based model for image captioning. three ways are implemented to select parts of image i.e areas of attention. New model is designed by three pairwise interactions that implies RNN state, describing image
region, and words embedding vectors. Further they have applied spatial transformer networks to the trained model. Results are evaluated in three ways CNN activation grids, object proposals, and our region proposal network. Region based network obtains best results using the spatial transformer. They basically determine image specific at- tention areas and are easy to give best results with other areas of network.

Srinivasan et al. (2018) proposed a hybrid system employing the use of multilayer Convolutional Neural Network to generate vocabulary describing the images and a Long Short Term Memory to accurately structure meaningful sentences using the gen- erated keywords. The convolutional neural network compared the target image to a large dataset of training images, and then generated an accurate description using the trained captions. CNN was used to extract the object from an image, which were then fed into a RNN along with the captions.

Helcl et al. (2018) applied a self-attentive network in place of a RNN. They evalu- ated on two methods of incorporating the visual features in the model- first included the image representation as input to the network; and second, predict the visual features by including training model and used it as an auxiliary objective. Their study has been based on the Transformer.

This component of training had been introduced in RNNs by Elliott and Kádár (2017) and they called it the "imagination". The best translation quality was achieved when training data was extended with additional image captioning data and parallel textual data. In this unconstrained setup, the best scoring model employs the previously introduced in RNN-based sequence-to-sequence models imagination component.

Elliott et al. (2018) werre basically trying to figure out the best output from the type of multilingual grounded language learning. They have shown multilingual training is better than bilingual training. They have designed a multilingual model that can work well on either multiple translations or comparable sentence pairs. They have considered four languages and shown their performance results multilingual over bilingual train- ing.

Aneja et al. (2018) had worked on convolutional image captioning technique demon- strating its efficiency keeping LSTM in baseline. Their main two features described are: convolutional technique that is CNN based depicting contrary results to LSTM base approach and Improved version of CNN that uses attention mechanism with spa- tial image features. They have achieved best results to demonstrate qualitatively objects of image.

Anderson et al. (2018) presented top-down and bottom-up together collectively for VQA and image captioning. Bottom-up approach uses faster R-CNN method for iden- tifying each object vector from image and top-down approach uses attention mecha- nism over the image and then evaluating averaged weighted feature of overall images. Their results have shown the good outcome of this approach both animate captioning and visual question answering. Their task has also involved visual and linguistic feature in predicting objects.

Hewitt et al. (2018) used the largest available dataset, with images for approximately 10,000 words in each of 100 languages. They have worked on 12 high-resource and 20 low- resource languages. Evaluating translation quality with the help of parts-of-speech and introduced a method for predicting word concreteness from image features which gives similar output that can be easily predicted by human perception. They have demonstrated that concrete subsets of words to translate gives best results and their accuracy is high than existing methods.

Guo et al. (2019) introduced informative way of captioning images. They are send- ing structured words to LSTM and then describing image in a descriptive manner, giv- ing proper attributes (like color, size, activities depicted in the image.) They had pro- posed mulitasking learning to produce structured caption. They had trained LSTM en- coder-decoder MT model, where the sequence of words is given as input to generate structured output images description.

Zhaoa et al. (2019) had come up with new approach other than traditional methods. Retrieval-based and Template-based methods are restricted to describe text of an im- age. They had combined the image attributes and sentence feature with a multimodal layer. The performance of the proposed model is evaluated on the Flickr8k, Flickr30k and MSCOCO datasets. Best results are obtained from combination of ATTssd and CNNm, that means combining image high-level attributes and sentence described through temporal convolutional with the recurrent neural network for visual text.

Wang et al. (2019) have been working on accuracy and diversity of the captions. Hence, they had proposed endogenous–exogenous attention architecture to apprehend both endogenous attention, and the exogenous attention. Endogenous attention implies to stochastic visual orienting, , where exogenous attention implies to deterministic visual orienting. At every step, model generates endogenous–exogenous attention heat map and then combines into hidden state of LSTM for generating words sequentially. Accuracy of captions have been captured with good results from this model. Performance has been tested from Flickr30K and MSCOCO datasets.

Cao et al. (2019) introduced a new approach for image captioning of a model, i.e. bidirectional semantic attention-based guiding of long short-term memory (Bag- LSTM). They have integrated semantic attention mechanism and bidirectional LSTMs for their work. Model has been designed in a manner where CNN is used for visual representation, Bi-LSTM for generating captions and Visual features and Semantics associated to it are combined together using an attention mechanism. Their performance results of generating caption have considered both historical and future context.

## III. EXPERIMENTAL SETUP

For our experiments, we have taken the Flickr30K dataset which provides English descriptions (captions) for images. It is a collection of around thirty thousand images that are collected from Flickr. Each image in the dataset has five human-annotated ref- erence sentences. To make this dataset work in Hindi, we manually translated each image's two annotated English sentences into Hindi. Since manual translation of Eng- lish sentences into Hindi is a time-consuming task. We were only able to translate two sentences. For the generation of captions, we used two pre-trained image recognition convolutional neural network (CNN) architectures viz VGG16 and ResNet50.

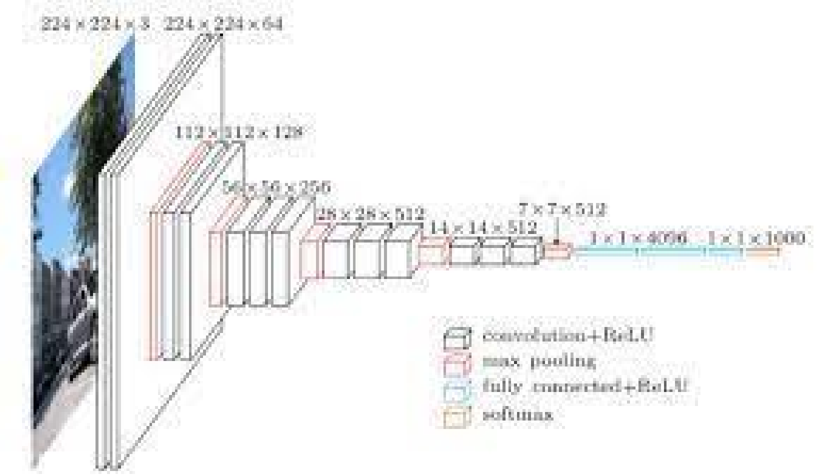

Fig.2:VGG16 Architecture

VGG16 is a CNN architecture that is widely used for image recognition tasks. It was introduced by Karen Simonyan and Andrew Zisserman of the Visual Geometry Group Lab of Oxford University in 2014. The architecture consists of 16 convolutional layers and is very appealing because of its uniformity. The input to the network is an image of dimensions (224, 224, 3). The first two layers have 64 channels of a 3x3 filter size and the same padding. Then after a max pool layer of stride (2, 2), two layers have convolution layers of 128 filter size and filter size (3, 3). This is followed by a max- pooling layer of stride (2, 2) which is the same as the previous layer. Then there are 2 convolution layers of filter size (3, 3) and 256 filters. After that, there are 2 sets of 3 convolution layers and a max pool layer. Each has 512 filters of (3, 3) size with the same padding. This image is then passed to the stack of two convolution layers 12. Figure 2 shows the schematic diagram of this architecture.

ResNet50 is a convolutional neural network (CNN) architecture that was developed by Microsoft Research in 2015. It is a variant of the popular ResNet architecture, which stands for "Residual Network". The "50" in the name refers to the number of layers in the network, which is 50 layers deep. The architecture is widely used for image recog- nition and other computer vision tasks.

The ResNet50 architecture consists of 48 convolutional layers, 1 MaxPool layer, and 1 Average Pool layer. It uses a bottleneck design for the building block, which reduces the number of parameters and matrix multiplications. The architecture follows two basic design rules: first, the number of filters in each layer is the same depending on the size of the output feature map; second, if the feature map's size is halved, it has double the number of filters to maintain the time complexity of each layer. ResNet50 is widely used for image recognition and other computer vision tasks, and it compares well with other models such as VGG-19 and 18-layer ResNet. Figure 3 shows the sche- matic diagram of this architecture.

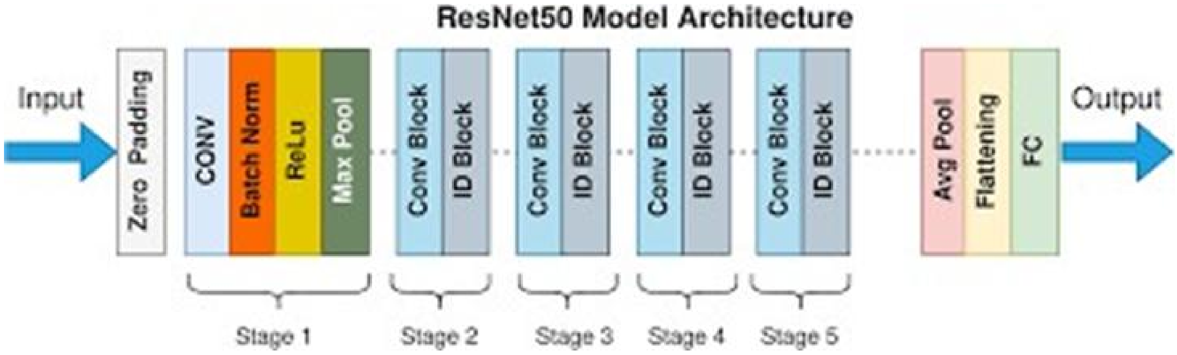

Fig.3: RESNET50 Architecture

Since years Computer Vision Domain has been dominantly ruled by CNNs. Then came Transformers which became the SOTA in NLP domain. Researchers at Google first introduced transformers for NLP tasks. Similarly in 2021, for Computer Vision tasks, researchers at Google introduced transformers for images. In CNN architectures, CNNs starts by looking locally and with time their receptive field becomes more global in each layer. Vision transformers inherently have a large field of view from the start. A vision transformer looks at the image by taking the input image and splitting it into patches of 16x16 pixels. These patches are then flattened by a linear transformation (to convert them into vectors). Form here everything works like our Text Transformer. Figure 4 shows the diagram of a basic vision transformer.

DOI: 10.48175/568

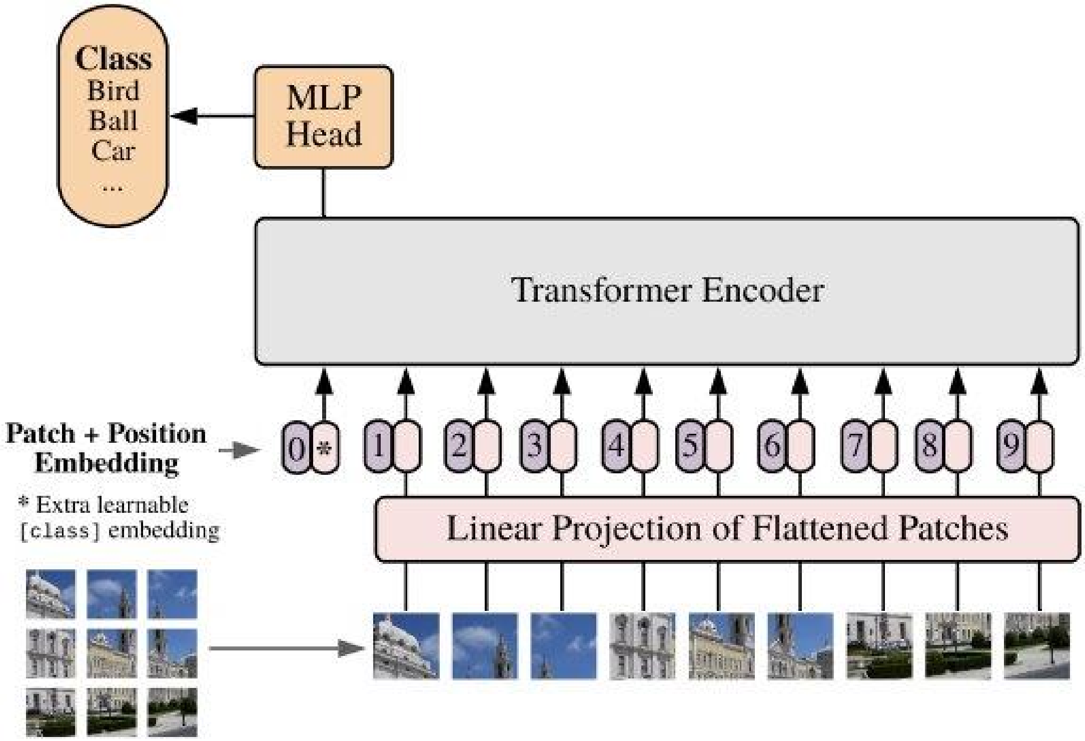

Fig. 4: Vision Transformer Architecture

## IV. METHODOLOGY

The process of image captioning using VGG16/RESNET50 was done using the fol- lowing steps and the entire working is shown in figure 4.

1. Preprocessing: The input image was preprocessed to a fixed size of (128, 128, 128, 3).
2. Feature extraction: The image architecture was then used to extract features from the preprocessed image.
3. Sequence generation: The LSTM was used to generate a sequence of words based on the extracted features.
4. Caption generation: The generated sequence of words was used to generate a textual description of the image.

Once this was done, the English captions were translated into Hindi using the en- coder-decoder NMT system. For the development of the machine translation system, 5 lac parallel sentences were collected from articles in art, culture, and tourism do- mains. Next text preprocessing was performed where the text was first tokenized using a sentence piece tokenizer. This generated the subword tokens of the text. The process was performed on English as well as Hindi sentences. This tokenized text was then sent for training of the model which was done in two phases. The first phase encoded the English sentences into word vectors where each word of the English sentence was con- verted to a word vector and was then sent to an LSTM cell for processing. Once all the English input words were processed, the output of the final LSTM cell was processed as a count vector which was then sent for translation (decoding). Here again while the training the model, each Hindi word was again converted into the word vector. Each vector (Hindi word) was sent to the LSTM cell for decoding. Once all the words were processed. The output of each cell was given to the Softmax activation function which selected the appropriate word. Since the production NMT system would be aware of the next word in the decoding sequence. Teacher forcing was applied. It involved feed- ing observed ground-truth samples back into the LSTM cell at each step. Figure 5 shows the working of the NMT system.

DOI: 10.48175/568

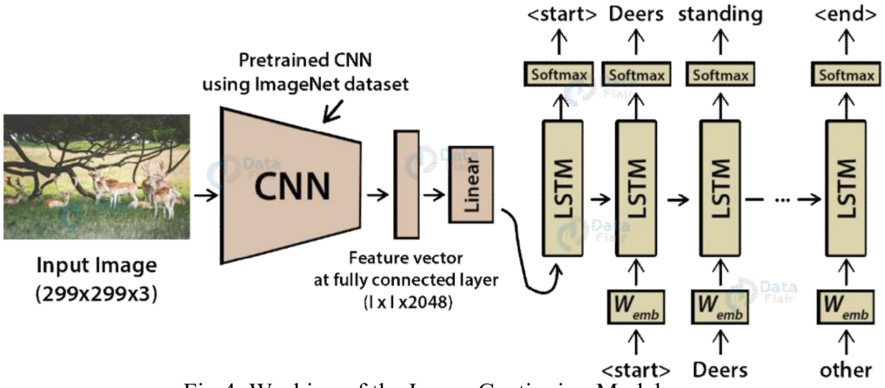

Fig.4: Working of the Image Captioning Model

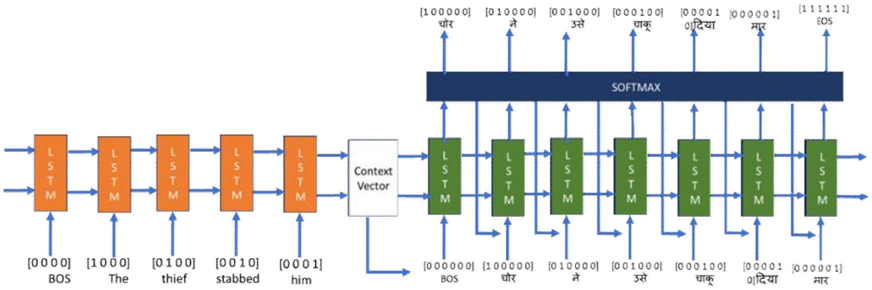

Fig.5: Working of Encoder-Decoder NMT Model

## V. EVALUATION

The trained machine translation system was tested on 2 thousand sentences. BLEU MT evaluation metric (Papineni et al., 2002) was used for this. The metric compared translated sentences produced the MT system with human-annotated Hindi translations that were developed earlier. While translating the text generated through VGG16 image captioning architecture, it was found that most of the sentences were not performing well. Thus, the BLEU score produced was 12.35. On testing with the RESNET50 image captioning model, it was found that the system produced better results than the VGG16 model. The BLEU score produced was 18.02. Further Image Transformer model produced even better results and produced a score of 24.37. Table 1. shows the results of this evaluation.

**Table 1.** BLEU-Based Evaluation of Multimodal Machine Translation System.

| **Model** | **BLEUScore** |
|---|---|
| VGG16Architecture | 12.35 |
| RESNET50Architecture | 18.02 |
| Image Transformer | 27.37 |

## VI. CONCLUSION

In this paper, we have shown the development of a multimodal machine translation system that generated image captions for input images. Two systems were developed. The first system was trained using VGG16 architecture and the second system was trained using RESNET50 architecture. The third system was trained using Image Transformers. These image captioning models generated English descriptions which were then translated into Hindi using a neural machine translation system. The generated Hindi captions were then compared with the human-generated descriptions of the images. This was done using the BLEU metric. The VGG16 model produced a score of 12.35 while the RESNET50 model produced a BLEU score of 18.02 which was slightly better. Image Transformer model produced a BLEU score of 24.37 which was best among the three.

DOI: 10.48175/568

## REFERENCES

**[1].** Caicedo, J. C., BenAbdallah, J., González, F. A., &Nasraoui, O. (2012). Multimodal repre- sentation, indexing, automated annotation and retrieval of image collections via non-negative matrix factorization. Neurocomputing, 76(1), 50- 60.

**[2].** Feng, Y., & Lapata, M. (2013). Automatic caption generation for news images. IEEE trans- actions on pattern analysis and machine intelligence, 35(4), 797- 812.

**[3].** Gong, Y., Jia, Y., Leung, T., Toshev, A., & Ioffe, S. (2013). Deep convolutional ranking for multilabel image annotation. arXiv preprint arXiv:1312.4894.

**[4].** Mao, J., Xu, W., Yang, Y., Wang, J., & Yuille, A. L. (2014). Explain images with multimodal recurrent neural networks. arXiv preprint arXiv:1410.1090

**[5].** Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., & Yuille, A. (2014). Deep captioning with multimodal recurrent neural networks (m-rnn). arXiv preprint arXiv:1412.6632.

**[6].** Socher, R., Karpathy, A., Le, Q. V., Manning, C. D., & Ng, A. Y. (2014). Grounded compo- sitional semantics for finding and describing images with sentences. Transactions of the As- sociation for Computational Linguistics, 2, 207-218.

**[7].** Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., & Yuille, A. (2014). Deep captioning with multimodal recurrent neural networks (m-rnn). arXiv preprint arXiv:1412.6632.

**[8].** Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4566-4575).

**[9].** Elliott, D., Frank, S., & Hasler, E. (2015). Multilingual image description with neural se- quence models. arXiv preprint arXiv:1510.04709.

**[10].** Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... & Bengio, Y. (2015, June). Show, attend and tell: Neural image caption generation with visual attention. In Inter- national conference on machine learning (pp. 2048-2057).

**[11].** Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image cap- tion generator. In Proceedings of the IEEE conference on computer vision and pattern recog- nition (pp. 3156-3164).

**[12].** Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE conference on computer vision and pattern recog- nition (pp. 2625-2634).

**[13].** Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE conference on computer vision and pattern recogni- tion (pp. 3128-3137).

**[14].** Chen, X., & Lawrence Zitnick, C. (2015). Mind's eye: A recurrent visual representation for image caption generation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2422-2431).

**[15].** Caglayan, O., Aransa, W., Wang, Y., Masana, M., García-Martínez, M., Bougares, F.,... & Van de Weijer, J. (2016). Does multimodality help human and machine for translation and image captioning? arXiv preprint arXiv:1605.09186.

**[16].** Calixto, I., Elliott, D., & Frank, S. (2016). Dcu-uva multimodal mt system report. In Proceed- ings of the First Conference on Machine Translation: Volume 2, Shared Task Papers (Vol. 2, pp. 634-638).

**[17].** Huang, P. Y., Liu, F., Shiang, S. R., Oh, J., & Dyer, C. (2016). Attention-based multimodal neural machine translation. In Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers (Vol. 2, pp. 639-645).

**[18].** Libovický, J., Helcl, J., Tlustý, M., Pecina, P., & Bojar, O. (2016). CUNI system for WMT16 automatic post-editing and multimodal translation tasks. arXiv preprint arXiv:1606.07481. Guasch, S. R., & Costa-Jussà, M. R. (2016). WMT 2016 Multimodal translation system de- scription based on bidirectional recurrent neural networks with double-embeddings. In Pro- ceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers (Vol. 2, pp. 655-659).

**[19].** Caglayan, O., Barrault, L., &Bougares, F. (2016). Multimodal attention for neural machine translation. arXiv preprint arXiv:1609.03976.

**[20].** Elliott, D., Frank, S., Sima'an, K., & Specia, L. (2016). Multi30k: Multilingual english- ger- man image descriptions. arXiv preprint arXiv:1605.00459.

**[21].** Elamri, C., & Planque, T.D. (2016). Automated Neural Image Caption Generator for Visually Impaired People.

**[22].** Hitschler, J., Schamoni, S., &Riezler, S. (2016). Multimodal pivots for image caption trans- lation. arXiv preprint arXiv:1601.03916.

**[23].** Liu, Y., Fu, J., Mei, T., & Chen, C. W. (2016). Storytelling of photo stream with bidirectional multi-thread recurrent neural network. arXiv preprint arXiv:1606.00625.

**[24].** You, Q., Jin, H., Wang, Z., Fang, C., & Luo, J. (2016). Image captioning with semantic atten- tion. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4651-4659).

**[25].** Caglayan, O., Aransa, W., Bardet, A., García-Martínez, M., Bougares, F., Barrault, L., ... & Van de Weijer, J. (2017). Lium-cvc submissions for wmt17 multimodal translation task. arXiv preprint arXiv:1707.04481.

**[26].** Calixto, I., Liu, Q., & Campbell, N. (2017). Doubly-attentive decoder for multi-modal neural machine translation. arXiv preprint arXiv:1702.01287.

**[27].** Zhang, J., Utiyama, M., Sumita, E., Neubig, G., & Nakamura, S. (2017). NICT-NAIST Sys- tem for WMT17 Multimodal Translation Task. In Proceedings of the Second Conference on Machine Translation (pp. 477-482).

**[28].** Duselis, J., Hutt, M., Gwinnup, J., Davis, J., & Sandvick, J. (2017). The AFRL-OSU WMT17 Multimodal Translation System: An Image Processing Approach. In Proceedings of the Sec- ond Conference on Machine Translation (pp. 445-449).

**[29].** Pedersoli, M., Lucas, T., Schmid, C., & Verbeek, J. (2017). Areas of attention for image cap- tioning. In Proceedings of the IEEE international conference on computer vision (pp. 1242- 1250).

**[30].** Ren, Z., Wang, X., Zhang, N., Lv, X., & Li, L. J. (2017). Deep reinforcement learning-based image captioning with embedding reward. In Proceedings of the IEEE conference on com- puter vision and pattern recognition (pp. 290-298).

**[31].** Liu, C., Sun, F., Wang, C., Wang, F., & Yuille, A. L. MAT: A Multimodal Attentive Trans- lator for Image Captioning. arXiv 2017. arXiv preprint arXiv:1702.05658.

**[32].** Lakshminarasimhan Srinivasan, D. S., & Amutha, A. L. (2018). Image Captioning- A Deep Learning Approach. International Journal of Applied Engineering Research, 13(9), 7239- 7242.

**[33].** Wang, Q., & Chan, A. B. (2018). CNN+ CNN: Convolutional decoders for image captioning. arXiv preprint arXiv:1805.09019.

**[34].** Helcl, J., Libovický, J., &Variš, D. (2018). CUNI System for the WMT18 Multimodal Trans- lation Task. arXiv preprint arXiv:1811.0469.

**[35].** Hewitt, J., Ippolito, D., Callahan, B., Kriz, R., Wijaya, D. T., & Callison-Burch, C. (2018, July). Learning translations via images with a massively multilingual image dataset. In Pro- ceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Vol- ume 1: Long Papers) (pp. 2566-2576).

**[36].** Kádár, Á., Elliott, D., Côté, M. A., Chrupała, G., &Alishahi, A. (2018). Lessons learned in multilingual grounded language learning. arXiv preprint arXiv:1809.07615.

**[37].** Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6077- 6086).

**[38].** Aneja, J., Deshpande, A., & Schwing, A. G. (2018). Convolutional image captioning. In Pro- ceedings of the IEEE conference on computer vision and pattern recognition (pp. 5561-5570). Laskar, S. R., Singh, R. P., Pakray, P., & Bandyopadhyay, S. (2019, November). English to Hindi Multi-modal Neural Machine Translation and Hindi Image Captioning. In Proceedings of the 6th Workshop on Asian Translation (pp. 62-67).

DOI: 10.48175/568

**[39].** Zhao, D., Chang, Z., & Guo, S. (2019). A multimodal fusion approach for image captioning. Neurocomputing, 329, 476-485.

**[40].** Guo, R., Ma, S., & Han, Y. (2019). Image captioning: from structural tetrad to translated sentences. Multimedia Tools and Applications, 78(17), 24321-24346.

**[41].** Wang, T., Hu, H., & He, C. (2019). Image Caption with Endogenous–Exogenous Attention. Neural Processing Letters, 50(1), 431-443.

**[42].** Cao, P., Yang, Z., Sun, L., Liang, Y., Yang, M. Q., & Guan, R. (2019). Image Captioning with Bidirectional Semantic Attention-Based Guiding of Long Short-Term Memory. Neural Processing Letters, 50(1), 103-119.

**[43].** Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Associa- tion for Computational Linguistics (pp. 311-318).

**[44].** Ballard, D. H., & Yu, C. (2003, April). A multimodal learning interface for word acquisition. In 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). (Vol. 5, pp. V-784). IEEE.

**[45].** Li, W., & Sun, M. (2005, October). Automatic image annotation using maximum entropy model. In International Conference on Natural Language Processing (pp. 34-45). Springer, Berlin, Heidelberg.

**[46].** Hsieh, Y. M., Yang, D. C., & Chen, K. J. (2005, October). Linguistically-motivated grammar extraction, generalization and adaptation. In International Conference on Natural Language Processing (pp. 177-187). Springer, Berlin, Heidelberg.

**[47].** Li, W., & Sun, M. (2006, February). Automatic image annotation based on wordnet and hier- archical ensembles. In International Conference on Intelligent Text Processing and Computa- tional Linguistics (pp. 417-428). Springer, Berlin, Heidelberg.

**[48].** Li, Z., Zhao, W., Li, Z., & Shi, Z. (2012, October). Automatic Image Annotation and Retrieval Using Hybrid Approach. In International Conference on Intelligent Information Processing (pp. 347-356). Springer, Berlin, Heidelberg.

**[49].** Kiros, R., Salakhutdinov, R., & Zemel, R. (2014, January). Multimodal neural language mod- els. In International Conference on Machine Learning (pp. 595- 603).

**[50].** Shah, K., Wang, J. K., & Specia, L. (2016, August). Shef-multimodal: Grounding machine translation on images. In Proceedings of the First Conference on Machine Translation (Vol. 2, pp. 660-665). ACL.

**[51].** Wang, C., Yang, H., Bartz, C., & Meinel, C. (2016, October). Image captioning with deep bidirectional LSTMs. In Proceedings of the 24th ACM international conference on Multime- dia (pp. 988-997). ACM.

DOI: 10.48175/568