Grenze International Journal of Engineering and Technology, Jan Issue

Multilingual Image Caption Generator

Milind Rane[1], Piyush Waghulde[2], Srushti Yadav[3], Pranamya Vemula[4], Om Wagh[5] and Abhijeet Waikar[6]
[1-6]Vishwakarma Institute of Technology, Department of Multidisciplinary Engineering, Pune, India
Email: milind.rane@vit.edu, {piyush.waghulde21, srushti.yadav21, pranamya.vemula21, om.wagh21, abhijeet.waikar21}@vit.edu

***Abstract*—The rise of social media and global communication requires multilingual image captioning for diverse participants, enabling cross-cultural understanding and engagement. By automatically generating captions in multiple languages, this technology enables seamless communication and enriches the user experience in an increasingly interconnected world. Existing image captioning models leverage alternative methodologies, such as transformer-based models, attention mechanisms, and pre-trained language models, to generate descriptive captions for images without relying on the CNN-LSTM architecture. These approaches provide effective solutions for image captioning, showcasing advancements in natural language processing and machine learning. However, the integration of CNN-LSTM architecture with the Flickr8k dataset and translation enables more accurate and contextually relevant multilingual image captions, enhancing versatility and adaptability for real-world applications. This approach combines deep learning techniques, diverse training data, and machine translation capabilities for superior multilingual image caption generation. The model's accuracy is assessed using accepted evaluation metrics and the model generates accurate captions in multiple languages for the images provided.**

***Index Terms*— Image captioning, Long Term Short Term(LSTM), Convolution Neural Network(CNN), Deep Learning, Flickr8k dataset, Googletrans.**

I. INTRODUCTION

An image contains a great deal of information. On social media and in observatories, enormous amounts of image data are produced daily. In this project, we experiment with different Deep Learning architectures and methodologies such as CNN [12] (Convolution Neural Network) and Long Short Term Memory (LSTM) for feature extraction and to determine which strategy is best for producing written captions from image data automatically. Instead of training on captions datasets in various languages, Google's Ajax API is used to convert the generated text in multiple languages.

A vital and active area of artificial intelligence research, image captioning offers a wide range of applications. Many social media sites, content recommendation systems, and virtual assistants use such technology. Another use for automatic picture captioning that significantly affects society is helping those who are blind or visually handicapped.

This study uses deep learning to recognize, identify, and produce captions for images.

II. LITERATURE REVIEW

Generating an image caption is a complex undertaking that will automatically create a caption for any image that

Grenze ID: 01.GIJET.10.1.562_1

is given as input. To carry out this project we consulted several research articles that summarize the made by numerous experts in this field.

Manish Raypurkar et.al [1] developed an image caption generator with the vision of helping visually impaired people to better understand the context of the images on the web. Their framework consists of a convolution neural network (CNN), which is followed by a recurrent neural network (RNN). The approach can produce image captions that are typically grammatically sound and semantically descriptive by learning from image and caption pairs.

Akash Verma et al. [2] aimed to explore the innate human tendency to extract information from both living and non-living objects. Building upon this idea, they proposed a caption generator project using a Long Short-Term Memory (LSTM) based Recurrent Neural Network (RNN) architecture to generate meaningful captions. To train their model, the researchers utilized the Flickr8k dataset.

Aghasi Poghosyan et al. [3] observed in their study on image caption generation, that the accuracy of the generated captions was primarily influenced by the preceding predicted word rather than the image itself. To address this limitation, the researchers employed a recurrent neural network (RNN) with a modified LSTM cell. This modified LSTM cell included an additional gate specifically designed to incorporate image features into the caption generation process. The authors conducted training and testing of their model using the MSCOCO image dataset, exclusively utilizing images and their corresponding captions.

Aditya Bhattacharya et.al [4] aimed at using excellent algorithms for generating image captions. They have constructed and tested a variety of multi-modal image captioning network architectures, including CNN encoders based on ResNet101, DenseNet121, and VGG19 as well as attention-based LSTM decoders. Their objective was to evaluate the effectiveness of each strategy using different assessment metrics, such as BLEU, CIDEr, ROUGE, and METEOR.

Varsha Kesavan et al. [5] addressed in their research the challenge of annotating images with captions, which can be a daunting task, especially for large commercial databases. To overcome this challenge, the researchers employed deep learning techniques, specifically convolutional neural networks (CNN) and recurrent neural networks (RNN), to automatically extract captions from images. They extensively analyzed and compared different methodologies and pre-trained models based on deep neural networks to identify the most effective approach. Through this rigorous analysis and fine-tuning process, the authors aimed to develop a highly efficient model for generating picture captions.

In their research, Oriol Vinyals et al. [6] identified a significant challenge in the field of artificial intelligence, which involves automatically describing the content of an image by combining computer vision and natural language processing. To address this challenge, the researchers put forth a generative model based on a deep recurrent architecture. Their model incorporated recent advancements in computer vision and machine translation, enabling the generation of meaningful descriptions for real-world images. Through numerical and qualitative analysis, the authors demonstrated that their model consistently achieved high levels of accuracy in generating accurate image descriptions.

Seung-Ho Han et.al [7] spotted that deep learning models are unable to identify things in a picture that are more significant than others or to provide an explanation for the choices made while creating captions. To solve this issue they designed an image caption generator that generates captions for specific objects in specific images and provide them with a visual description.

Grishma Sharma et.al [8] had a good interest in the field of artificial intelligence and so studied and developed an image caption generator. Focusing on various RNN strategies and examining their impact on phrase production, they have presented various models for the generation of captions for images based on deep learning and neural networks. Additionally, they have created captions for test images to compare other feature extraction and encoder model approaches. This is done to determine which model gives better results.

III. Working Methodology Demonstrating the Model

The flowchart shows the essential working which is based on simple encoder-decoder model[9].

Data Preparation: We have a dataset stored in the flickr8k directory, which consists of images in the Images subdirectory and captions in the captions.txt file. The images and captions are loaded and a mapping between image IDs and their corresponding captions is created.

Image Feature Extraction: The VGG16 model, a pre-trained CNN, is used to extract image features. The VGG16 model is loaded, and it removes the last classification layer, and creates a new model that outputs the features from the second-to-last layer. The images are then iterated, preprocessed, and fed through the VGG16 model. They are then stored as the extracted features in a dictionary with the image IDs as keys.

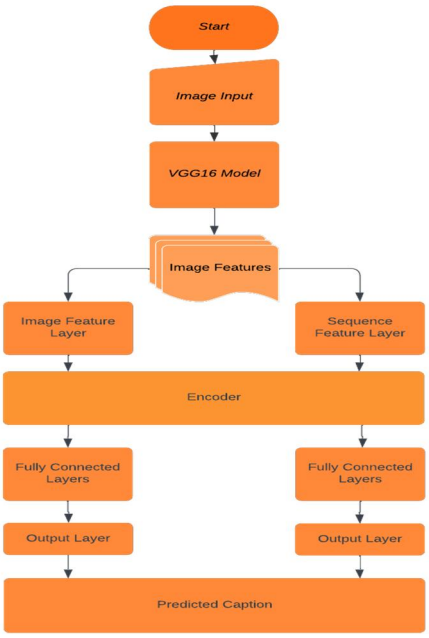


Figure 1. Flow chart of overall working demonstrating simple Encoder Decoder

TABLE I: DATA SPLITTING

Training	Testing
0.9	0.1

Caption Preprocessing: The captions in the mapping are preprocessed to prepare them for training. Several text preprocessing steps such as converting to lowercase, removing digits and special characters, and adding start and end tags are applied to the captions.

Tokenization and Vocabulary Building: The captions are tokenized using the Tokenizer class from Keras, which converts words into integers and builds vocabulary mapping. The vocabulary size is determined based on the tokenizer's word index.

Model Architecture: The architecture for the image captioning model involves two main components: an image feature layer and a sequence feature layer. The detailed view is shown in Fig. 1. The image feature layer takes the image features extracted from VGG16[11] as input, while the sequence feature layer takes the preprocessed captions as input. The outputs of both layers are combined and passed through a decoder consisting of fully connected layers to generate the predicted caption. The overview of working of the entire study is shown in Fig.2.

Model Training: The image captioning mode is trained using a data generator. The generator yields batches of training data, where each batch consists of image features, input sequences, and output sequences. The categorical cross-entropy loss function and the ADAM optimizer are used for model compilation and then trained for a specified number of epochs.

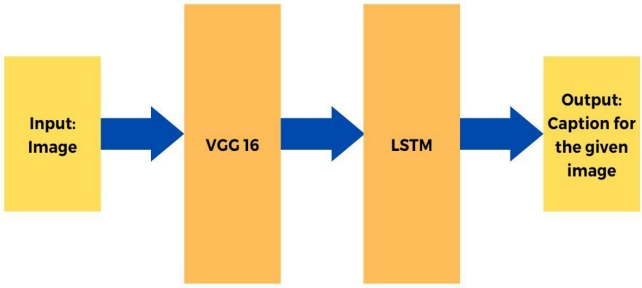


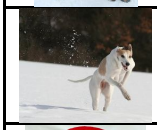
Figure 2. Block Diagram of Overview of Working Architecture

Model Evaluation: After training, the model is evaluated. Captions are predicted of the images and BLEU-1 and BLEU-2 scores are calculated, which measure the similarity between the predicted and actual captions.
Caption Generation: Given an image and the trained model, the function generates a caption by iteratively predicting the next word based on the image features and the previously generated words.
Multilingual Translation: The user is presented with choices of languages for multilingual translation. Googletrans is used here which is a library in Python which utilizes the Google Translate API to provide machine translation capabilities. It sends HTTP requests to the Google Translate service, passing the source language and target language along with the text to be translated, and then retrieves the translated text as the response, enabling seamless translation between multiple languages.

IV. Results and Discussions

The model is trained for 23 epochs after defining and fitting the model. At first, the accuracy is low which causes irrelevancy in the resulting captions which are generated based on the images in the testing dataset. But after training the model for at least 20 epochs it is observed that the generated captions are more related to the images present in the testing dataset. Therefore, we trained it for 20+ epochs since they generated better results. The generated results in multiple languages are shown in Table II.

Table II - Results of Captions Predicted by our Model in Different Languages

Images	English	French	German	Spanish	Japanese
	two horses pull driven in the lead	deux chevaux tirés conduits en tête	zwei Pferden, die an der Spitze getrieben werden, zwei Pferden	dos caballos tiran conducidos a la cabeza	開始シーケンス 2 頭の馬がリードで牽引 終了シーケンス
	dog running in the snow	chien qui court dans la neige	Hund läuft im Schnee	perro corriendo en la nieve	雪の中を走る犬
	girl in red dress is laying on the ground in front of the rainbow painted	fille en robe rouge est allongée sur le sol devant l'arc-en-ciel peint	Mädchen in rotem Kleid liegt auf dem Boden vor dem Regenbogen, bemalt	chica con vestido rojo está tendida en el suelo frente al arco iris pintado	赤いドレスを着た女の子が虹の前で地面に横たわっているペイントされた
	motorcycle rider on motorcycle	motocycliste sur moto	Motorradfahrer auf Motorrad	motociclista en motocicleta	オートバイのライダー

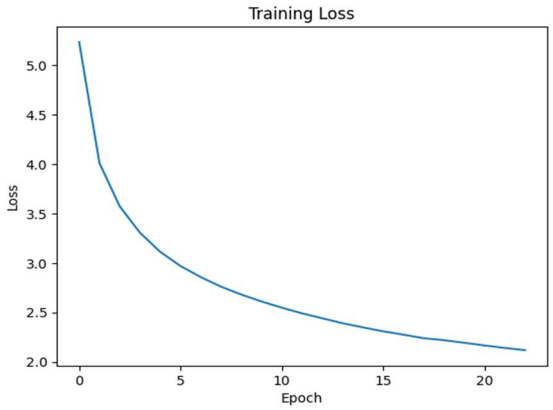


Figure 4. Evaluation of Epochs/Loss Ratio demonstrating overall Training Loss

Performance Evaluation

BLEU Score - BLEU score is chosen as a suitable parameter to measure the performance because it considers precision, n-gram matching, language consistency, aggregation, and provides a standardized evaluation, enabling

TABLE III: BLEU METRICS

BLEU - 1	BLEU - 2
0.525189	0.301326

a comprehensive and quantitative assessment of the quality and similarity of the generated captions to the reference captions.
BLEU-1 refers to the unigram precision, which calculates the percentage of overlapping unigrams between the generated translation and the reference translations.
BLEU-2 refers to the bigram precision, which calculates the percentage of overlapping bigrams between the generated translation and the reference translations.
The BLEU scores of this model is shown in Table III.

V. LIMITATIONS

Considering the limitations of this project, the performance of image caption generators is dependent on the quality, resolution, and clarity of images, so blurry or distorted images may generate inaccurate captions. Also, the model's caption generation is constrained by the vocabulary learned during training. It also has some real-time processing problems as its working completely depends on the model complexity and hardware resources. Furthermore, it lacks semantic understanding which means it may not capture complex or abstract relationships. It may not accurately describe scenes with multiple interacting objects or occlusions as it lacks human-like understanding. But these limitations could be improved to an extent by training it on more improved and diverse data.

VII. CONCLUSION

This paper summarizes the CNN and LSTM-based approach for extracting features from the images and generating captions for the same. The model is trained to increase the sentence's likelihood given the image. The captions are generated in various languages.
The proposed model successfully generates descriptive captions for images, showcasing its ability to bridge the gap between visual and textual information. Through extensive experimentation, we have demonstrated the effectiveness of our approach in terms of generating coherent and contextually relevant captions. The results obtained indicate the potential of our image caption generator in various applications such as image understanding, content indexing, and assistive technologies. While there is room for future improvements, our work contributes to the advancement of the field by providing a robust framework for automatic image captioning and inspires further research in this domain.

REFERENCES

[1] M. Raypurkar, A. Supe, P. Bhumkar, P. Borse, and S. Sayyad, "Deep learning-based image caption generator", International Research Journal of Engineering and Technology (IRJET), vol. 8, no. 03, 2021.
[2] Verma, Akash, Harshit Saxena, Mugdha Jaiswal, and Poonam Tanwar. "Intelligence Embedded Image Caption Generator using LSTM based RNN Model". In 2021 6th International Conference on Communication and Electronics Systems (ICCES), pp. 963-967. IEEE, 2021.
[3] Poghosyan, Aghasi, and Hakob Sarukhanyan. "Long Short-Term Memory with Read-only Unit in Neural Image Caption Generator." IEEE Comput. Sci. Inf. Technol (2017).
[4] Bhattacharya, Aditya, Eshwar Shamanna Girishekar, and Padmakar Anil Deshpande. "Empirical Analysis of Image Caption Generation using Deep Learning". arXiv preprint arXiv:2105.09906 (2021).
[5] Kesavan, Varsha, Vaidehi Muley, and Megha Kolhekar. "Deep learning based automatic image caption generation." In 2019 Global Conference for Advancement in Technology (GCAT), pp. 1-6. IEEE, 2019.
[6] Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. "Show and tell: A neural image caption generator." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3156-3164. 2015.
[7] Han, Seung-Ho, and Ho-Jin Choi. "Domain-specific image caption generator with semantic ontology." In 2020 IEEE International Conference on Big Data and Smart Computing (BigComp), pp. 526-530. IEEE, 2020.
[8] Sharma, Grishma, Priyanka Kalena, Nishi Malde, Aromal Nair, and Saurabh Parkar. "Visual image caption generator using deep learning." In 2nd international conference on advances in Science & Technology (ICAST). 2019.
[9] Mathur, Pranay, Aman Gill, Aayush Yadav, Anurag Mishra, and Nand Kumar Bansode. "Camera2Caption: a real-time image caption generator." In 2017 international conference on computational intelligence in data science (ICCIDS), pp. 1-6. IEEE, 2017.

[10] Kumar, N. Komal, D. Vigneswari, A. Mohan, K. Laxman, and J. Yuvaraj. "Detection and recognition of objects in image caption generator system: a deep learning approach." In 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), pp. 107-109. IEEE, 2019.
[11] Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. ArXiv. /abs/1409.1556
[12] Alzubaidi, L., Zhang, J., Humaidi, A. J., Duan, Y., Santamaría, J., Fadhel, M. A., & Farhan, L. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. Journal of Big Data, 8(1), 1-74. https://doi.org/10.1186/s40537-021-00444-8
[13] Panicker, Megha J., Vikas Upadhayay, Gunjan Sethi, and Vrinda Mathur. "Image caption generator." International Journal of Innovative Technology and Exploring Engineering (IJITEE) 10, no. 3 (2021).
[14] Hossain, MD Zakir, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. "A comprehensive survey of deep learning for image captioning." ACM Computing Surveys (CsUR) 51, no. 6 (2019): 1-36.
[15] Rennie, Steven J., Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. "Self-critical sequence training for image captioning." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7008-7024. 2017.
[16] Pan, Jia-Yu, Hyung-Jeong Yang, Pinar Duygulu, and Christos Faloutsos. "Automatic image captioning." In 2004 IEEE International Conference on Multimedia and Expo (ICME)(IEEE Cat. No. 04TH8763), vol. 3, pp. 1987-1990. IEEE, 2004.
[17] Cui, Yin, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie. "Learning to evaluate image captioning." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5804-5812. 2018.
[18] Jmail, Anas, Sanjana Dwivedi, and Mr Vijayakumara YM. "IMAGE CAPTIONING: TRANSFORMING SIGHT INTO SCENE."
[19] Zhao, Beigeng. "A systematic survey of remote sensing image captioning." IEEE Access 9 (2021): 154086-154111.
[20] Chen, Jianhui, Wenqiang Dong, and Minchen Li. "Image caption generator based on deep neural networks." ttps://www. math. ucla. edu/minchen/doc/ImgCapGen. pdf (2014).