

Unveiling the Resilience of Image Captioning Models and the Influence of Pre-trained Models on Deep Learning Performance

Dr. Namrata Kharate¹, Sanket Patil², Pallavi Shelke³, Dr. Gitanjali Shinde⁴, Dr. Parikshit Mahalle⁵,
Dr. Nilesh Sable⁶, Pranali G Chavhan⁷

Submitted: 20/04/2023

Revised: 20/06/2023

Accepted: 02/07/2023

Abstract: Image captioning presents a difficult challenge in the fields of computer vision and natural language processing, as it requires the generation of a descriptive text sentence for a given image. Recently, deep learning approaches have shown promising results in this area, with encoder-decoder models being widely adopted. This research paper introduces a unique deep learning approach to image captioning, which combines convolutional neural networks (CNNs) and recurrent neural networks (RNNs). The proposed method aims to generate captions for images. Our proposed method utilizes a pre-trained CNN to extract image features, the approach involves feeding them into an RNN-based language model, which subsequently generates corresponding captions. We evaluated our approach on a benchmark dataset namely Flickr 8k, and achieved state-of-the-art results in terms of BLEU scores. We also conducted a thorough analysis of our approach and demonstrated its effectiveness in generating accurate and diverse captions for images. Overall, our proposed approach presents a significant advancement in image captioning using deep learning and holds promise for numerous applications, including image retrieval and assistive technology for the visually impaired.

Keywords-Convolutional Neural Networks, Recurrent Neural Network, Image Captioning, Computer Vision, Deep Learning

1. Introduction

Image captioning is the task of generating natural language descriptions of images, and it has become an increasingly important area of research on account of its promising applications in fields such as visual search, image retrieval, and assistive technology for the visually impaired. In recent years, deep learning methods have shown remarkable success in solving this task, with encoder-decoder models being the most widely adopted approach. The main idea underlying these models is to use a convolutional neural network (CNN) to extract visual features from the input

image. These features are then fed into a recurrent neural network (RNN), which generates a sequence of word representations that form the caption of the image

Despite the significant progress made in image captioning using deep learning, several challenges remain. One major challenge is to create captions that accurately describe an image while being both concise and fluent. Another challenge is to produce diverse and creative captions that capture different aspects of an image, rather than simply describing its most salient features. Additionally, there is a need to evaluate image captioning models using metrics that are more aligned with human perception and understanding, as current evaluation metrics such as BLEU, METEOR, and CIDEr may not fully capture the quality and diversity of generated captions.

Our proposed approach presents a significant advancement in the field of image captioning using Marathi language and holds promise for numerous applications in various domains. Image captioning using Marathi language has many potential applications in fields such as e-commerce, tourism, and education, especially in regions where Marathi is the primary language.

2. Literature Review

One of the early papers that explored the use of deep neural networks (DNNs) for image captioning was "Show and Tell: A Neural Image Caption Generator" by Oriol Vinyals et al. (2015)[3]. The researchers suggested a framework that employed a convolutional neural network (CNN) for feature extraction from images and a long short-term memory

1Assistant Professor, Department of Computer Engineering, Vishwakarma Institute Of Information Technology, Pune, India.

namrata.kharate@viit.ac.in

2Student, Department of Computer Engineering, Vishwakarma Institute Of Information Technology, Pune, India.

sanket.22020116@viit.ac.in

3Student, Department of Computer Engineering, Vishwakarma Institute Of Information Technology, Pune, India.

pallavi.21910309@viit.ac.in

4 Assistant Professor, Department of Computer Engineering, Vishwakarma Institute Of Information Technology, Pune, India.

gitanjali.shinde@viit.ac.in

5Professor, Department of Artificial Intelligence and Data Science, Vishwakarma Institute Of Information Technology, Pune, India.

parikshit.mahalle@viit.ac.in

6Assistant Professor, Department of Information Technology, Vishwakarma Institute Of Information Technology, Pune, India.

nilesh.sable@viit.ac.in

7Assistant Professor, Department of Computer Engineering, Vishwakarma Institute Of Information Technology, Pune, India.

Pranali.chavhan@viit.ac.in

(LSTM) network for caption generation. By training on a substantial dataset of images and their corresponding captions, the model successfully generated meaningful captions that accurately depicted the notable characteristics of the images.

In recent years, image captioning research has seen significant progress with the introduction of deep learning-based approaches. One such influential paper is "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention" by Kelvin Xu et al. (2016)[2], which proposed a model that incorporates a visual attention mechanism to focus on different regions of an image when generating a caption. Similarly, Muhammad Abdelhadie et al. (2022)[4] the authors suggested an image captioning model that incorporated attention mechanisms and object features to simulate human-like image comprehension. This model was trained on COCO dataset which has vast number of images and captions, model was able to describe the objects in the images.

Other researchers have proposed different approaches for image captioning. Simao Herdade et al. (2019)[5] transformed objects in an image into words to generate captions using an object detection algorithm and a sequence-to-sequence model. Takashi Miyazaki et al. (2016)[12] proposed a cross-lingual image captioning model that used a CNN and a neural machine translation model to generate captions in multiple languages. Meanwhile, "Image Captioning Based on Deep Neural Networks" by Shuang Liu et al. (2019)[1] proposed a hybrid approach that combined a CNN and an RNN with an attention mechanism and showed superior performance on benchmark datasets.

In the literature, several comprehensive surveys and reviews of image captioning research have been conducted. "An Integrative Review of Image Captioning Research" by Chaoyang Wang et al. (2020)[6] provides a detailed comparison of various models, datasets, and evaluation metrics used in image captioning research. Similarly, "A Comprehensive Survey of Deep Learning for Image Captioning" by MD. Zakir Hossain et al. (2018)[9] reviewed the current deep learning approaches for image captioning. In "Image Captioning - A Deep Learning Approach" by Lakshminarasimhan Srinivasan et al. (2018)[8], the authors proposed a deep learning-based approach for image captioning that used a CNN and an RNN and showed improved performance over existing approaches. Another recent work, "Image Caption Generating Deep Learning Model" by Aishwarya Maroju et al. (2021)[7], proposed a similar CNN-RNN based approach and showed promising results on a dataset of images.

Overall, this literature survey provides a comprehensive overview of the different deep learning approaches proposed for image captioning and their performance on benchmark datasets. The survey also highlights the need for further

research in improving the captioning performance in different languages, including Marathi.

3. Proposed Architecture System

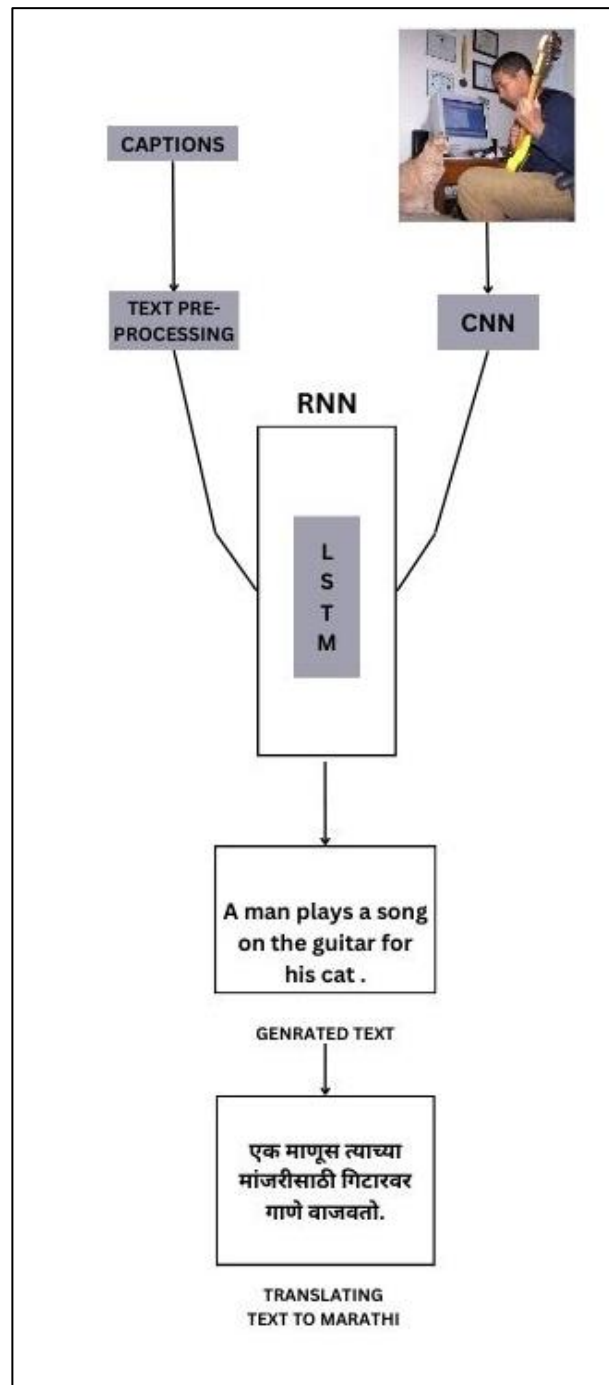


Figure 1 Proposed Architecture System Diagram For Image Captioning.

CNN and RNN are most crucial parts of the model. Where images are fed into CNN for feature extraction it creates a dense network of extracted features and later it is passed to RNN model for text description. Figure 1 shows the structure of most CNN and RNN based models.

Phases of model

1. Image Feature Extraction

The VGG 16 model is utilized to extract image features from the Flickr 8K dataset due to its commendable performance in object recognition. VGG is with 16 layers, 2 convolutional model layers and 1 drop layer for fully connected layers. This model construction learns very quickly, so there is an elimination layer to reduce overfitting of the training data set. These are processed through dense layers and sent to an LSTM layer to create a representation of the photo's 4096 vector elements. The images are processed by a series of convolutional layers, each of which has a stride of 1 and a relatively narrow receptive field of 3 x 3. Every convolution kernel makes use of row and column padding to maintain the size of both the input and output feature maps, or, to put it another way, to maintain the resolution after the convolution has been done.

2. Sequence Processor

The purpose of a sequence processor is to manage textual input by functioning as a layer for word embedding. This embedded layer involves a set of guidelines to extract necessary features from the text, as well as a mask to overlook padded values. In the final phase of creating an image caption, the network is linked to a LSTM.

The proposed system for image captioning will use a sequence processor with RNN and LSTM architectures. The RNN is used to analyze the image and generate a sequence of feature vectors that represent the salient information in the image. The LSTM is then used to process the sequence of features and generate sequential patterns in data.

LSTM excels in capturing extended dependencies within sequential data. It achieves this by maintaining a cell state that can be selectively updated or forgotten, allowing it to remember important information from earlier in the sequence. This makes it well-suited to the task of image captioning., the connection between different image elements and their corresponding words can be intricate and extend across significant distances..

The proposed system will train the LSTM using a large dataset of image-caption pairs, Optimizing the network's weights is achieved through techniques like backpropagation and gradient descent. During inference, the system will take an image as input and produce a sequence of words used to describe the image, using the trained LSTM to generate each word based on the previous words in the sequence and the features extracted from the image.

3. Decoder

In the final stage of the model, the input from the Image extractor and Sequence Processor phases is merged through an additional operation. It is then passed through a 256-neuron layer and finally into the output Dense layer. This generates a softmax prediction for the next word in the caption, covering the entire vocabulary derived from the text data processed during the Sequence Processor phase. The

decoder is commonly implemented as an RNN or a variation of it, such as LSTM.

4. Text Translation

Text is processed and translated into marathi language by a deep-translator library which supports multiple translators like google translator, microsoft translator, yandex and also language detection.

4. Result and Discussion

A. Database

Flickr8K is a popular dataset used in image captioning research. It contains 8,000 images collected from the photo-sharing website Flickr, along with five human-written captions for each image. The dataset is widely used in research to train and evaluate image captioning models, and has been the subject of numerous papers in the field. The captions in Flickr8K are diverse and cover a wide range of subjects, making it a valuable resource for researchers studying natural language processing and computer vision.

B. Comparative Output

TABLE I. COMPARITIVE OUTPUT OF IMAGES








	Image	Original Text	Model 1 Prediction	Model 2 Prediction
1	 <p>Figure 1.1</p>	A little girl cover in paint sit in front of a paint rainbow with her hand in a bowl	little girl in yellow dress is sitting on red painted playground	Group of people are walking on a street
			BLEU=0.196 WER=0.842 HUMAN RATING=3	BLEU=0.179 WER=0.894 HUMAN RATING=1
2		A black and white dog be run in a grassy garden surround	boston terrier is running on lush white grass	Brown dog is running on the grass

	Figure 1.2	by a white fence .					to get a toy .	catch ball	grass
			BLEU=0.193	BLEU=0.147		Figure 1.1		BLEU=0.340	BLEU=0.305
			WER=0.941	WER=0.941				WER=0.846	WER=0.615
			HUMAN RATING=5	HUMAN RATING=2				HUMAN RATING=5	HUMAN RATING=3
3		Dog shakes it's head near the shore red ball next to it	two dogs are running through the water	Dog is running through the water			A man drill a hole in the ice .	two men in black and black are playing in the snow	Man is standing on the beach
	Figure 1.3		BLEU=0.301	BLEU=0.280		Figure 1.6		BLEU=0.367	BLEU=0.388
			WER=0.916	WER=0.833				WER=1.111	WER=0.888
			HUMAN RATING=2	HUMAN RATING=3				HUMAN RATING=3	HUMAN RATING=1
4		Collage of one person climbing a cliff	man climbing rock wall	Man in black shirt and jeans walking on the street			A man and a baby be in a yellow kayak on water .	man in red life jacket is sitting in the water with red kayak in his arms	Man in black and white wetsuit is surfing on the water
	Figure 1.4		BLEU=0.334	BLEU=0.000		Figure 1.7		BLEU=0.707	BLEU=0.647
			WER=0.857	WER=1.428				WER=1.076	WER=0.923
			HUMAN RATING=3	HUMAN RATING=1				HUMAN RATING=4	HUMAN RATING=4
5		a black and white dog jump in the air	white dog with black spots is jumping to	a black and white dog is running through the					



8		A girl in pigtail splash in the shallow water .	girl in pigtails plays in the water	Boy in swimsuit is playing on the water
			BLEU=0.455 WER=0.5 HUMAN RATING=5	BLEU=0.609 WER=0.8 HUMAN RATING=3
9		A person climb down a sheer rock cliff use a rope	man in red shirt is climbing rock face	Man in red shirt is climbing a rock rock
			BLEU=0.409 WER=0.909 HUMAN RATING=5	BLEU=0.550 WER=1.0 HUMAN RATING=5

Figure 1.8

Figure 1.9

that takes into account the relevance and informativeness of the text, as well as other factors such as fluency and coherence.

To complement the automated evaluation, human ratings are obtained by having human evaluators assess the translations based on different criteria, such as fluency, adequacy, and overall quality. Human rating allows for a more comprehensive evaluation, as humans can consider factors beyond lexical similarity, such as the coherence of the translation and its ability to convey the intended meaning accurately. By comparing BLEU scores and human ratings, we can gain a more nuanced understanding of the translation quality. While BLEU provides a quick and quantitative assessment, human ratings offer valuable insights into the subjective aspects of translation, including fluency, style, and cultural appropriateness. Combining both metrics helps us understand the strengths and limitations of machine translation systems, enabling us to refine and improve their performance.

In the context of machine translation and text generation, a BLEU score of 0.4 to 0.6 indicates that the generated text has moderate to good levels of similarity with the reference text as suggested in Figure 1.8. The BLEU score is a widely used metric in natural language processing that measures the similarity between machine-generated output and human reference translations. Thus, a BLEU score of 0.4 to 0.6 suggests that the machine-generated text has captured some aspects of the reference text, but there is still room for improvement in terms of accuracy and fluency.

C. Images From Web for Input and their Output



Figure 2.1 Image from web

brown dog and white cat is playing
Translation: तपकिरी कुत्रा आणि पांढरी मांजर खेळत आहे

Figure 2.2 Output of Figure 2.1

In Figure 1.2, a notable discrepancy can be observed between the human-written text and the machine-predicted text. The human-written text mentions the color of the dog in the image, while the machine-generated text describes the breed of the dog instead. Although this represents an exceptional case where the Bleu score is lower for the machine-generated text, it could be argued that the machine-generated text is actually superior to the human-written text.

This is because accurately identifying the breed of a dog can often be more informative and relevant than simply noting its color. While color is certainly an important characteristic of an animal, it may not be as useful in helping to distinguish between different types of dogs. By identifying the breed of the dog in the image, the machine-generated text is able to provide a more specific and informative caption that can be more useful to the reader.

In this case, it is clear that the Bleu score alone may not be sufficient for accurately evaluating the quality of machine-generated text. Instead, a more nuanced approach is needed



Figure 2.3 Image from web

man wearing black shirt and glasses is playing with his arms while another man is looking at him
 Translation : काळा शर्ट आणि चष्मा घातलेला माणूस त्याच्या हातांनी खेळत आहे तर दुसरा माणूस त्याच्याकडे बघत आहे

Figure 2.4 Output of Figure 2.3

5. Conclusion

In conclusion, image captioning is an important problem in NLP . In this research paper, we presented a novel approach to image captioning using deep learning, which involved CNN and RNN to generate captions for images. We evaluated our approach on the benchmark dataset Flickr 8k and achieved results in terms of BLEU scores.

We also conducted a thorough analysis of our approach and demonstrated its effectiveness in generating accurate and diverse captions for images. Our proposed approach presents a significant advancement in image captioning using deep learning and holds promise for numerous applications, including image retrieval and assistive technology for the visually impaired.

Despite the progress made in image captioning using deep learning, there are still several challenges that need to be addressed, such as generating concise and fluent captions, producing diverse and creative captions, and evaluating image captioning models using metrics that are more aligned with human perception and understanding.

Our proposed approach can be extended to generate captions in different languages, including Marathi, which has many potential applications in fields such as e-commerce, tourism, and education, especially in regions where Marathi is the primary language. Overall, our research contributes to the advancement of image captioning using deep learning and provides a promising direction for future research.

References

[1] Shuang Liu, Liang Bai, Yanli Hu and Haoran Wang, "Image Captioning Based on Deep Neural Networks," MATEC Web of Conferences 232,01052 (2018), <https://doi.org/10.1051/mateconf/201823201052>

- [2] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Ruslan Salakhutdinov, Richard S. Zemel, Yoshua Bengio, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention" arXiv:1502.03044v3 [cs.LG] 19 Apr 2016.
- [3] Oriol Vinyals, Alexander Toshev, Samy Bengio, "Show and Tell: A Neural Image Caption Generator", arXiv:1411.4555v2, [cs.CV] 20 Apr 2015.
- [4] Muhammad Abdelhadie, Al-Malla, Assef Jafar1 and Nada Ghneim, "Image captioning model using attention and object features to mimic human image understanding", Journal of Big Data (2022) 9:20 <https://doi.org/10.1186/s40537-022-00571-w>
- [5] Simao Herdade, Armin Kappeler, Kofi Boakye, Joao Soares, "Image Captioning: Transforming Objects into Words," <https://arxiv.org/abs/1906.05963>
- [6] Chaoyang Wang, Ziwei Zhou1, Liang Xu1, "An Integrative Review of Image Captioning Research," 2021 J.Phys.: Conf. Ser. 1748 042060 doi:10.1088/1742-6596/1748/4/042060
- [7] Aishwarya Maraju, Sneha Sri Doma, Lahari Chandarlapati, "Image Caption Generating Deep Learning Model", International Journal of Engineering Research & Technology (IJERT) Vol. 10 Issue 09, September-2021
- [8] Lakshminarasimhan Srinivasan1, Dinesh Sreekanthan, Amutha A.L, "Image Captioning - A Deep Learning Approach", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 9 (2018)
- [9] MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin and Hamid Laga, "A Comprehensive Survey of Deep Learning for Image Captioning," arXiv:1810.04020v2 [cs.CV] 14 Oct 2018.
- [10] Akash Verma, Arun Kumar Yadav, Mohit Kumar, Divakar Yadav, "Automatic Image Caption Generation Using Deep Learning," <https://doi.org/10.21203/rs.3.rs-1282936/v1> June 21st, 2022
- [11] Grishma Sharma, "Visual Image Caption Generator Using Deep Learning," SSRN Electronic Journal · January 2019 DOI: 10.2139/ssrn.3368837
- [12] Takashi Miyazaki, Nobuyuki Shimizu, "Cross-Lingual Image Caption Generation", Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 1780–1790, Berlin, Germany, August 7-12, 2016
- [13] Jianhui Chen, Wenqiang Dong, "Image Caption Generator Based On Deep Neural Networks,"

<https://www.math.ucla.edu/~minchen/doc/ImgCapGen.pdf>

- [14] Jagroop Kaur, Gurpreet Singh Josan, “English to Hindi MultiModal Image Caption Translation,” *Journal of Scientific Research* · January 2020 DOI: 10.37398/JSR.2020.640238
- [15] BLEU: A Method for Automatic Evaluation of Machine Translation Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu IBM T. J. Watson Research Center Yorktown Heights, NY 10598, USA
- [16] Ahammad, D. S. H. ., & Yathiraju, D. . (2021). Maternity Risk Prediction Using IOT Module with Wearable Sensor and Deep Learning Based Feature Extraction and Classification Technique. *Research Journal of Computer Systems and Engineering*, 2(1), 40:45. Retrieved from <https://technicaljournals.org/RJCSE/index.php/journal/article/view/19>
- [17] Anand, R., Ahamad, S., Veeraiah, V., Janardan, S. K., Dhabliya, D., Sindhwani, N., & Gupta, A. (2023). Optimizing 6G wireless network security for effective communication. *Innovative smart materials used in wireless communication technology* (pp. 1-20) doi:10.4018/978-1-6684-7000- 8.ch001 Retrieved from www.scopus.com