iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VI **Month of publication:** June 2023

DOI: https://doi.org/10.22214/ijraset.2023.54297

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VI Jun 2023- Available at www.ijraset.com

Captioning Image Using Deep Learning: A Novel Approach

Sunanda Chakraborty[1], Shibam Chakraborty[2], Moloy Dhar[3], Nirupam Saha[4], Bidyutmala Saha[5], Pallabi Das[6], Rafiqul Islam[7], Sutapa Sarkar[8]

[1, 2, 3, 4, 5,6,7,8] *Department of Computer Science & Engineering, Guru Nanak Institute of Technology, Kolkata, India*

[1]sunandachakraborty270@gmail.com, [2]chakrabortyshibam184@gmail.com, [3]moloy.dhar@gnit.ac.in, [4]nirupam.saha@gnit.ac.in, [5]bidyutmala.saha@gnit.ac.in, [6]pallabi.das@gnit.ac.in, [7]rafiqul.islam@gnit.ac.in, [8]sutapa.sarkar@gnit.ac.in

Abstract: ***Captioning image using deep learning is a technology that aims to generate descriptive and accurate textual descriptions for images. By using the power of deep neural networks, this approach enables computers to understand and interpret visual content bridging the gap between the visual and textual domains. This paper focuses on developing an image captioning system using deep learning techniques. The paper aims to generate descriptive textual captions for images, enabling machines to understand and communicate the content of visual data. The methodology involves leveraging convolutional neural networks (CNNs) for image feature extraction and recurrent neural networks (RNNs) for sequential language generation. Captioning images is a challenging task in computer vision that involves describing the content of an image in natural language. In recent years, deep learning techniques have shown remarkable success in various computer vision tasks, including image captioning.***

Keywords: CNN, VGG-19, LSTM, NLP, RNN

I. INTRODUCTION

Captioning image using deep learning is an innovative procedure which combines computer vision and natural language processing to automatically generate descriptive captions for images. It addresses the challenging task of overcoming the gap between visual perception and language understanding. VGG-19 has been pre trained on more than a million images from ImageNet database so the output comes out quite accurate. Image captioning is an exciting field at the intersection of computer vision and natural language processing (NLP). It involves generating descriptive textual captions for images, enabling machines to understand and communicate the content of visual data. Deep learning techniques, such as (CNNs) and recurrent neural networks (RNNs), have shown remarkable success in image captioning tasks.

The paper will leverage a pre-trained CNN to extract meaningful features from the images. These extracted features will serve as input to the RNN-based captioning model. The RNN, equipped with recurrent cells such as LSTM or GRU, will learn to generate descriptive captions based on the extracted image features. Training the model will involve optimizing the parameters to minimize the captioning loss. In this paper, we are using MS-COCO dataset which is used for large-scale object detection, segmentation and captioning and are used in various computer vision papers. It contains 328,000 images of everyday objects and humans. It contains feature-rich annotations including object detection, captioning, person-keypoints and more. Accurate image captioning has numerous practical applications, including assisting visually impaired individuals in understanding images, enhancing image search engines, and enabling better image indexing and retrieval. Our approach consists of two main components: an image encoder and a language decoder. The image encoder employs a pre-trained CNN, such as VGG or ResNet, to extract high-level visual features from the input image. These features capture the semantic information present in the image and serve as the input to the language decoder. The language decoder is implemented using an RNN, specifically a long short-term memory (LSTM) network, which generates captions based on the visual features obtained from the image encoder.

II. METHODOLOGY

The result of image captioning is the generation of descriptive and informative captions for images. By using advanced techniques such as deep learning and natural language processing, image captioning models can analyze the content of an image and generate textual descriptions that accurately represent the visual elements.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VI Jun 2023- Available at www.ijraset.com

Before using Tokenizer to convert texts to integers, we mark the start and end of the strings with a unique marker.

start_marker = "aaa"
end_marker = "zzz"

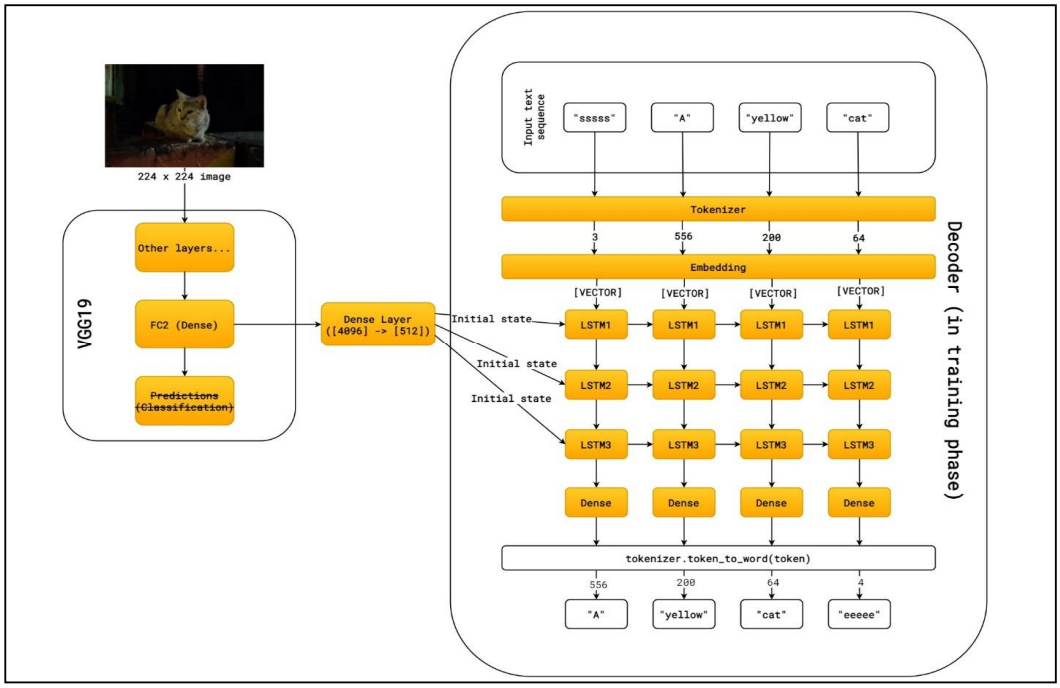


Fig. 1

We are limiting the maximum number of words in our vocabulary. We will only use the 10500 most frequent words in the captions from the training data.

num_words = 10500

III. RESULT

Captioning image using deep learning has got a significant advancement in recent years. Deep learning models, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been developing highly effective image captioning systems.

The decoder consists of 3 LSTM layers whose internal state-sizes are 512

state_size = 512

The embedding layer converts integer tokens into vectors of length 128

embedding_size = 128

Image captioning using deep learning has got a significant advancement in recent years. Deep learning models, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been develop highly effective image captioning systems. One popular approach is the encoder-decoder architecture, where a CNN is used as the encoder to extract image features, and an RNN is employed as the decoder to generate captions. This architecture has demonstrated impressive results in generating coherent and descriptive captions for images.

In summary, image captioning offers valuable solutions for accessibility, search ability, and user engagement. With continued advancements in technology and research, we can expect image captioning systems to become even more sophisticated, enabling better understanding and interaction with visual content. Here, we have used RMSprop optimizer.

A. We Have used RMSprop Optimizer

MODEL SUMMARY: VGG-19

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	[(None, 224, 224, 3)]	0
block1_conv1 (Conv2D)	(None, 224, 224, 64)	1792
block1_conv2 (Conv2D)	(None, 224, 224, 64)	36928
block1_pool (MaxPooling2D)	(None, 112, 112, 64)	0
block2_conv1 (Conv2D)	(None, 112, 112, 128)	73856
block2_conv2 (Conv2D)	(None, 112, 112, 128)	147584
block2_pool (MaxPooling2D)	(None, 56, 56, 128)	0
block3_conv1 (Conv2D)	(None, 56, 56, 256)	295168
block3_conv2 (Conv2D)	(None, 56, 56, 256)	590080
block3_conv3 (Conv2D)	(None, 56, 56, 256)	590080
block3_conv4 (Conv2D)	(None, 56, 56, 256)	590080
block3_pool (MaxPooling2D)	(None, 28, 28, 256)	0
block4_conv1 (Conv2D)	(None, 28, 28, 512)	1180160
block4_conv2 (Conv2D)	(None, 28, 28, 512)	2359808
block4_conv3 (Conv2D)	(None, 28, 28, 512)	2359808
block4_conv4 (Conv2D)	(None, 28, 28, 512)	2359808
block4_pool (MaxPooling2D)	(None, 14, 14, 512)	0
block5_conv1 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv2 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv3 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv4 (Conv2D)	(None, 14, 14, 512)	2359808
block5_pool (MaxPooling2D)	(None, 7, 7, 512)	0

flatten (Flatten)	(None, 25088)	0
fc1 (Dense)	(None, 4096)	102764544
fc2 (Dense)	(None, 4096)	16781312

==========

Total params: 139,570,240
Trainable params: 139,570,240
Non-trainable params: 0

Fig. 2

We use this model differently in training and testing.
However, challenges still exist in image captioning, such as accurately capturing fine-grained details, handling complex scenes, and generating captions that capture context and semantic meaning. Ongoing research and development efforts aim to address these challenges and improve the accuracy and contextual understanding of image captions.

For example, if the sentence is "A dog is running on the ground", then if we input sitting to an LSTM layer, we will train it to output "on". This is called the **"Teacher Forcing"** method.

Whereas, during testing phase, we use the output token of the previous layer as an input to the next layer.

B. Sample outputs

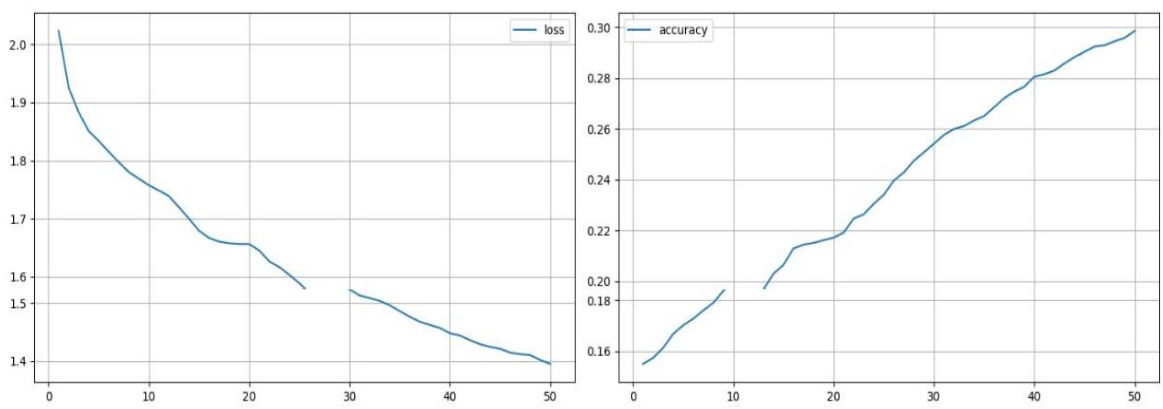


Fig. 3

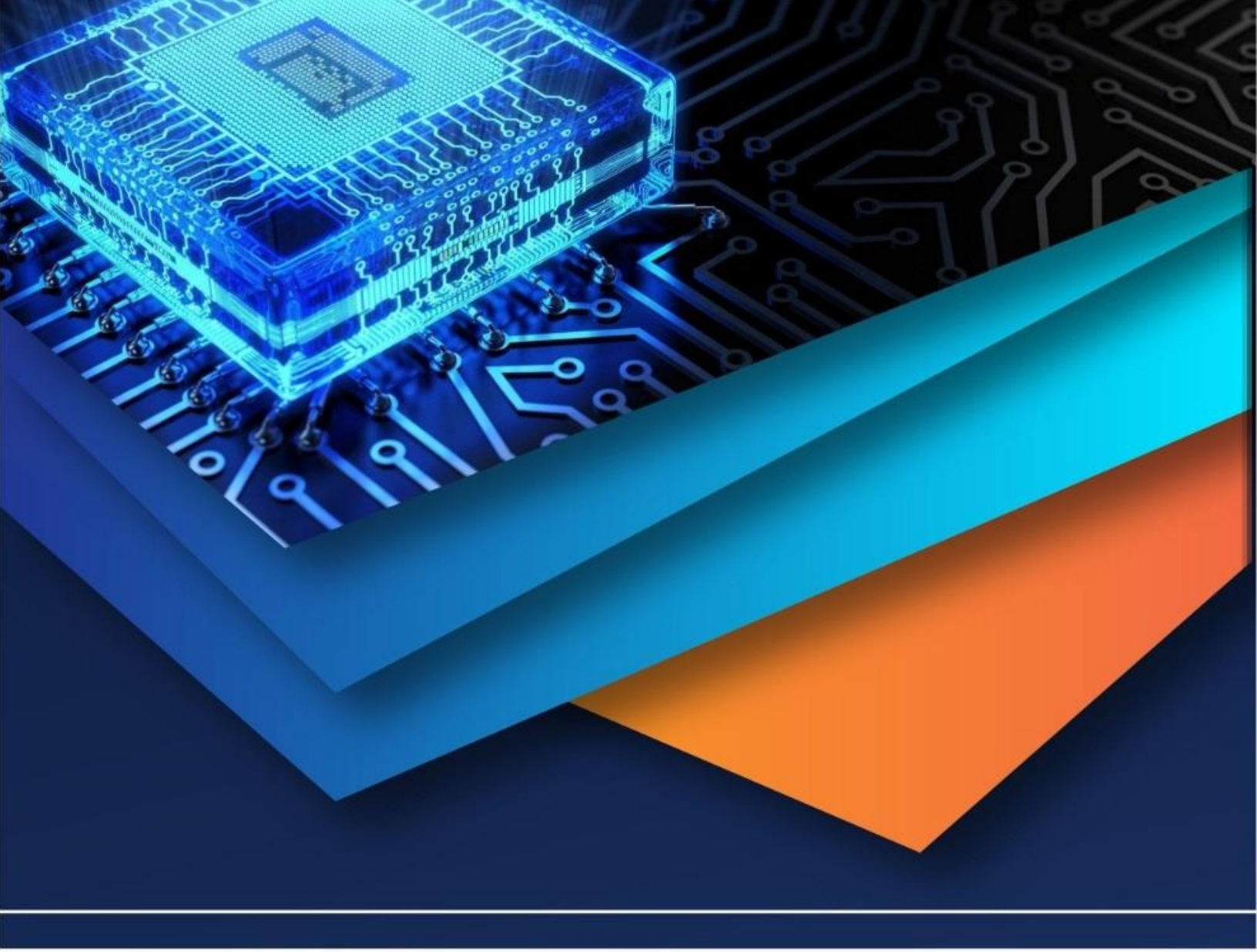
Above picture represents our result.
This enables more effective content retrieval and indexing, facilitating the organization and retrieval of visual data in various applications.

IV. CONCLUSION

Captioning image using deep learning has emerged as a highly promising and effective approach for generating textual descriptions for images using CNN, RNN and LSTM model along with advancements such as attention mechanisms and transfer learning, has led to significant improvements in the quality and accuracy of generated captions.This paper successfully implemented and trained the models on a suitable dataset, evaluated their performance using quantitative metrics, and discussed the obtained results. We compare our model against state-of-the-art methods and demonstrate its superiority in terms of caption quality, fluency, and relevance.
This paper showcased the potential of deep learning in addressing the challenging task of generating accurate and contextually relevant captions for images and videos. By leveraging the power of CNNs for visual feature extraction and RNNs for language modeling, the developed system demonstrated the ability to understand the visual content and generate descriptive captions.

REFERENCES

[1] Yang, L., & Hu, H. (2019). Adaptive syncretic attention for constrained image captioning. Neural Processing Letters
[2] Fu, K., Jin, J., Cui, R., Sha, F., & Zhang, C. (2016). Aligning where to see and what to tell: Image captioning with region-based attention and scene-specific contexts. IEEE transactions on pattern analysis and machine intelligence
[3] Li, J., Yao, P., Guo, L., & Zhang, W. (2019). Boosted Transformer for Image Captioning. Applied Sciences(19)
[4] Oluwasanmi, A., Aftab, M. U., Alabdulkreem, E., Kumeda, B., Baagyere, E. Y., & Qin, Z. (2019). CaptionNet: Automatic end-to-end siamese difference captioning model with attention.
[5] Farhadi, M. Hejrati, M. A. Sadeghi, P. Young,C.Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: ―Generating sentences from images.‖7. Peter Anderson, Xiaodong He, Chris Buehler, DamienTeney, Mark Johnson, Stephen Gould, and LeiZhang. 2017.‖ Bottom-up and top-down attention for image captioning and vqa‖.
[6] J. Donahue,L. Anne Hendricks,S. Guadarrama,M. Rohrbach, S. Venugopalan, K. Saenko, and T.Darrell‖Long-term recurrent convolutional networks for and description‖. InCVPR, 2015.
[7] Oriol Vinyals, Alexander Toshev, Samy Bengio, andDumitru Erhan. 2014. Show and tell: A neural im-age caption generator.
[8] YonghuiWu,MikeSchuster,ZhifengChen,QuocVLe,MohammdNorouzi,WolfgangMacherey,MaximKrikun,YuanCao,Qin Ga0,KlausMacherrey,et at.2016.Google'sneural machine translation system:"Bridging the gap between human and machine translation".
[9] Fang, F., Wang, H., Chen, Y., & Tang, P. (2018). Looking deeper and transferring attention for image captioning. Multimedia Tools and Applications, 77(23).
[10] Oriol Vinyals, Alexander Toshev, Samy Bengio, andDumitru Erhan. 2014. Show and tell: A neural image captiongenerator.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY