iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.53389

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 V May 2023- Available at www.ijraset.com

Captioning Image Using Deep Learning Approach

Arpan Sen[1], Aman Kumar Jaiswal[2], Atul Singh[3], Bidrohi Pradhan[4], Subhasis Dey[5], Aniket Mondal[6], Moloy Dhar[7], Pallabi Das[8]
Guru Nanak Institute of Technology, Kolkata, India
[1]rony990353@gmail.com, [2]pratiksehjpal2000@gmail.com, [3]atulsingh7213@gmail.com, [4]pradhanbidrohi3@gmail.com, [5]deysubhasis59@gmail.com, [6]aniketmondal478@gmail.com, [7]moloy.dhar@gnit.ac.in, [8]pallabi.das@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. The process of image captioning involves two main components: an image encoder and a language decoder. The image encoder is basically a Convolutional Neural Network (CNN) that processes the input image extracting high-level features and representations. These features capture the visual content and generating meaningful captions. The language decoder on the other hand is usually a Recurrent Neural Network (RNN), such as a long short-term memory (LSTM) network. It takes the encoded image features as input and generates a sequence of words to form the caption.***
Keywords: ML, CNN, NLP, LSTM, RNN, METEOR

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. Image description models can analyze the content of an image and generate coherent and meaningful textual descriptions that accurately capture the visual scene. Image description using deep learning is a task that involves generating a textual description or caption for an input image. Deep learning models, particularly convolution neural networks (CNNs) and recurrent neural networks (RNNs), are commonly used for image captioning tasks. Captioning image using deep learning has seen significant advancements in recent years, and various techniques have been proposed to improve the quality and accuracy of generated captions. These include attention mechanisms, reinforcement learning, and combining multiple modalities such as visual and textual information. Overall, deep learning-based image captioning provides a powerful approach to automatically generate descriptive captions for images, with potential applications in areas such as content indexing, image retrieval, and assisting visually impaired individuals.

II. METHODOLOGY

Captioning image using deep learning involves the following methodology they are:

1) *Data Collection And Preprocessing:* A large dataset of images with corresponding captions is collected. Each image is paired with one or more captions describing its content. The images are preprocessed, usually by resizing them to a fixed size and normalizing the pixel values.
2) *CNN Feature Extraction:* A pre-trained CNN model, such as VGGNet, ResNet, or Inception, is used to select visual looks from the input image. The CNN model is typically trained on a large-scale image classification task and can capture meaningful representations of the image content.
3) *Sequence Modeling with RNN:* The visual features extracted by the CNN are then glutted into an RNN, such as a Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU), to produce the textual description. The RNN processes the visual features along with an initial input (usually a start token) and generates words one at a time. The generated word is then fed back as input to predict the next word in the sequence, forming an auto regressive process.
4) *Training the Model:* The image-caption pairs are recycled to train the deep learning model. The model is trained to minimize the discrepancy between the predicted captions and the ground truth captions using a loss function such as cross-entropy. This involves adjusting the weights and parameters of the CNN and RNN through back propagation and gradient descent optimization.
5) *Inference and Caption Generation:* Once the model is trained, it can be recycled for produce captions for new, unseen images. The process involves feeding an image through the trained CNN to extract visual features, which are then passed to the RNN

© IJRASET: All Rights are Reserved |

for caption generation. The RNN generates words sequentially until it predicts an end token or reaches a maximum caption length.

III. RELATED WORK

1) *"Deep Visual-Semantic Alignments for Generating Image Descriptions" (2015) by Karpathy and Fei-Fei:* This work proposed a novel approach to align images with their corresponding captions by learning a joint embedding space for images and text. The model utilized a CNN for image encoding and an LSTM for generating captions. It also introduced a dataset Flickr8K for evaluation.
2) *"Bottom-Up and Top-Down Attention for Image Captioning" by Anderson et al.(2017):* This work introduced an attention mechanism that allows the model to focus on specific regions of the image while generating captions. It employed a bottom-up approach by first generating a set of region proposals using a Faster CNN object detector, and then used an LSTM with a top-down attention mechanism for caption generation.
3) *"Image Captioning with Semantic Attention" by you et al. (2017):* This work proposed a semantic attention mechanism that incorporates semantic information from the image into the caption generation process. It introduced a semantic attention module that uses pre-trained word embeddings and a semantic attention weight matrix to attend to relevant image regions.

IV. DATASET

The Flickr8k dataset is a widely used dataset in the field of computer vision and Neural Language Processing (NLP). It was created to facilitate research in image captioning, which involves generating textual descriptions for images. The dataset contains a total of 8000 images, hence the name "Flickr8k." Each image is accompanied by five different captions, resulting in a total of 40,000 captions.

V. 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. One popular approach is the encoder-decoder architecture, where a CNN is recycled 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.

Fig. 1

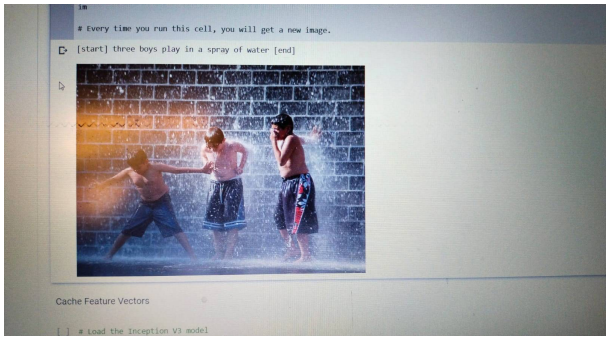


Fig. 2

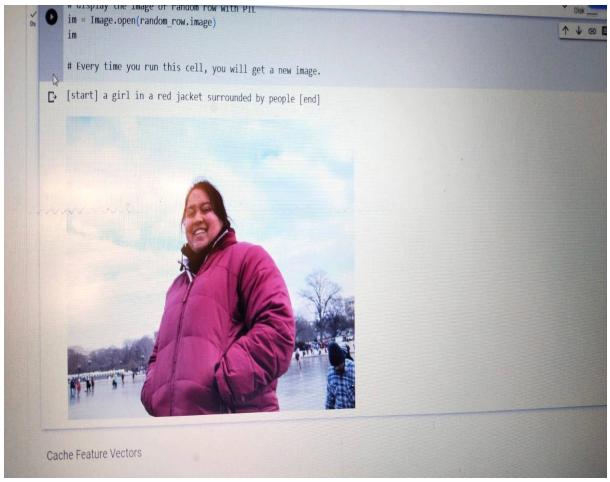


Fig. 3

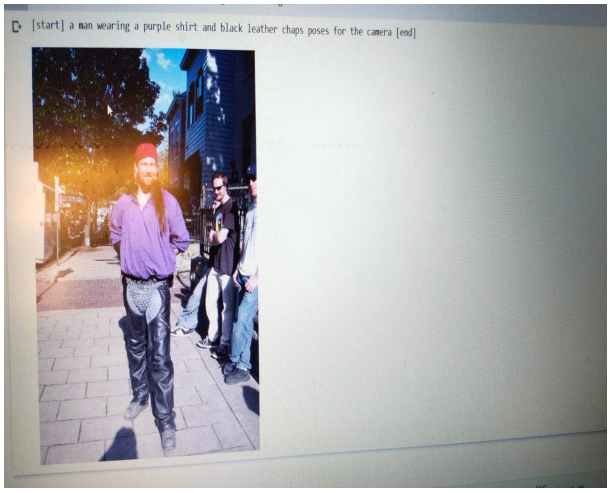


Fig. 4

OUTPUT

Large group of happy friends is having fun on mountain top and looks at mountain's valley.

Fig. 5

Fig. 1-5 shows the output generated by our system.

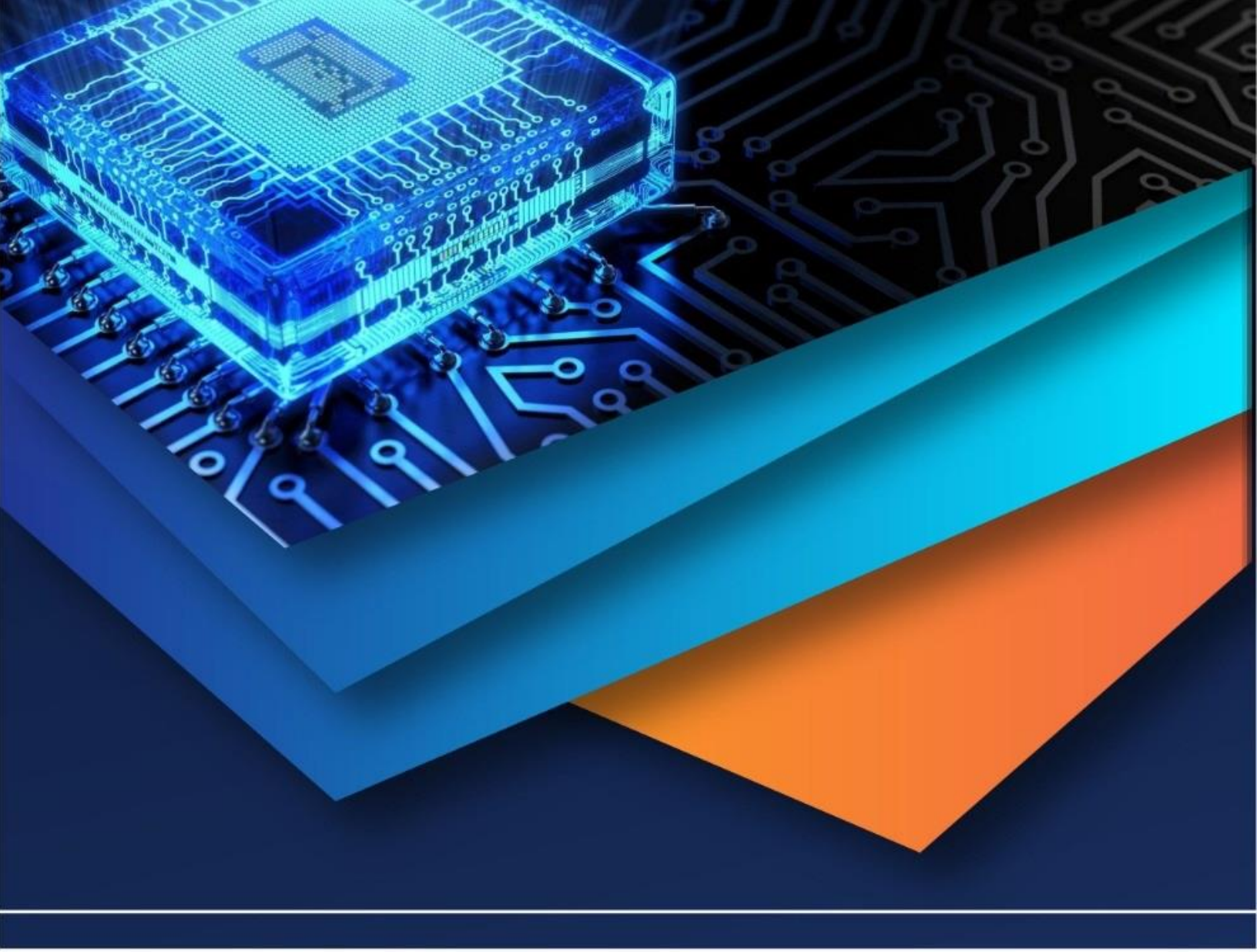
VI. CONCLUSION

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. A lot of research are going, on to gather more accuracy .Here in this paper we have successfully completed what we mentioned in this paper proposal using Flicke8K dataset.

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] Wu, J., & Hu, H. (2017). Cascade recurrent neural network for image caption generation. Electronics Letters
[6] Tan, J. H., Chan, C. S., & Chuah, J. H. (2019). COMIC: Toward A Compact Image Captioning Model With Attention. IEEE Transactions on Multimedia.
[7] W.-Y. Lan, X.,-X. Wang, G, Yang, X.-R Li (2019) Improving Chinese Image Captioning by Tag Prediction
[8] Li, X., & Jiang, S. (2019). Know more say less: Image captioning based on scene graphs. IEEE Transactions on Multimedia, 21(8)
[9] Chen, X., Zhang, M., Wang, Z., Zuo, L., Li, B., & Yang, Y. (2018). Leveraging unpaired out-of-domain data for image captioning. Pattern Recognition Letters.
[10] Fang, F., Wang, H., Chen, Y., & Tang, P. (2018). Looking deeper and transferring attention for image captioning. Multimedia Tools and Applications, 77(23)

© IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)