e-ISSN: 2582-5208

International Research Journal of Modernization in Engineering Technology and Science

(Peer-Reviewed, Open Access, Fully Refereed International Journal)

Volume:05/Issue:04/April-2023 Impact Factor- 7.868 www.irjmets.com

SURVEY PAPER ON IMAGE CAPTION GENERATOR USING NEURAL NETWORK

Neha Nemade*[1], Prathmesh Bonde*[2], Dr. S.T. Gandhe*[3],

Priti Deore*[4], Prof. Viddulata A. Patil*[5]

*[1,2,3,4,5]Dept. Of Electronics And Telecommunication Pune Institute Of Computer Technology, India.

DOI : https://www.doi.org/10.56726/IRJMETS35637

ABSTRACT

Image captioning is becoming an essential task for virtual assistants, editing software, and image indexing. Automatically generating a caption for an image is known as image captioning. It is becoming more popular as a freshly developed scientific field. The objective of image captioning requires the capture and expression of semantic information of images in natural languages. The approach for generating image captions generally consists of two general methods i.e., encoding and decoding. The RNN is used to create phrases, whereas the CNN is used to extract characteristics or features from images. When an image is given as input, the model has been trained such that it generates captions that most accurately describe the input image. For a particular image, some software tools create and provide a caption. Using CNN, it recognizes noteworthy items and their connections. A current study has shown CNN's benefits for automatic image caption synthesis and machine translation to address the issue.

Keywords: CNN, LSTM, Image, Caption, Deep Learning, Neural Networks, RNN, Graphics Processing Units.

I. INTRODUCTION

An image contains various information, such as items, attributes, settings, and actions. Making captions for photos is much easier for humans. We see several images every day from several sources, including the internet, news stories, document diagrams, and commercials. The images on these websites are left up to the viewers' imagination. Most viewers can still understand the photos even though most of them lack descriptions.

Every image on the internet should have a description in order to increase the efficiency and descriptiveness of searches and indexing. It has been evident that simply mentioning the names of identified objects does not provide a sufficient description of the image.

To overcome this problem, there are generally two steps to follow: 1) understanding and recognizing significant objects and 2) Establishing the appropriate relationships between the image's various components. Because there are so many photos available nowadays, image captioning is becoming. more and more popular and essential for retrieving photos. There are several potential difficulties that may arise when attempting to collect visual information from google photos, optical system analysis of self-driving cars, and other sources such as privacy concerns, inaccurate or incomplete metadata, technical issues, and copyright restrictions.

Although deep learning techniques and technology have been available for a while, the usage of GPUs and the expansion of digital data have accelerated recent advancements in the field.

II. LITERATURE REVIEW

By Lahari Chandarlapati, Sneha Sri Doma, and Aishwarya Maroju [1], a deep-learning model for creating image captions. A Resnet-LSTM model was used in this paper's picture captioning process. Based on training caption data, the encoder Resnet decodes images and generates words. LSTM is used to decode the output of the Resnet model. The goal of this method is to boost the efficiency of caption production, which is lower when using the standard CNN-RNN model on the Flick8k dataset. Rakshith Shetty, Hamed Tavakoli, and Jorma Laaksonen [2] caption pictures and videos using augmented neural architectures. The language model has access to CNN features throughout the whole caption production process after being initialized with contextual characteristics in this paper's concept for an image captioning system. The proposed model performed better than the traditional CNN-RNN models on the much larger MS-COCO dataset. Despite having some room for improvement, such as the terminology used by caption generators, it is still unable to fully understand the nuances of human evaluation. The method described by G Geetha, T. Kirthigadevi, and G. Godwin [3] in Image Captioning Using Deep Convolutional Neural Networks used the CNN-LSTM model to caption images. The entire model's workflow, from collecting data sets to creating captions, was detailed. Here, convolutional neural networks were used as the encoder and LSTMs as the decoder to create the captions. Haseeb, Srushti G M, Bhamidi Haripriya, and Mrs. Madhura Prakash [4] so that it can automatically view the photograph and produce an accurate description in straightforward language. The image is compressedly encoded using CNN, and the corresponding text is generated using an RNN-LSTM model. By not initializing the weights of the CNN component's system to the model, this technique avoids the issue of deteriorating 4 Overfitting in this study. As the dataset, or the number of photos and text descriptions that can be trained, grows, so does the possibility that the model will generate a suitable phrase and its per- formance. The concept, Imageability- and Length-Changeable Image Captioning, by Marc A. Kastner, Kazuki Unemura, and Ichiro Ide [5] propose the diverse production of image captions with two controllable parameters, namely imageability, and length. If a statement elicits a distinct mental image, it is said to be 'imageable' in psycholinguistics. Its use in picture captioning has previously been investigated, with successful results using LSTM-based models. The study claims that it may be used to change how detailed the captions are, either making them more abstract or more thorough in their descriptions of the event. According to the article, the other parameter, length, enables an additional level of cap- tion customization for a range of applications. The proposed transformer-based method greatly improves caption creation's performance and naturalness. Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan [6], show and tell: A neural image caption generator, introduced the first deep learning model for image caption generation, which uses a CNN to encode the image and LSTM to generate captions. Andrej Karpathy, Li Fei-Fei, deep visual-semantic alignments for generating image captions [7], proposed an approach that uses a multimodal embedding model that maps images and captions into a common semantic space. Vasili Ramanishka, Abir Das, Jianming Zhang, Kate Saenko [8], top-down visual saliency guided by captions. This paper introduced a model that uses top-down visual saliency guided by captions to selectively attend to the most relevant parts of the image when generating captions. M. D. Zeiler and R. Fergus. [9], used deconvolution with max-pooling layers that project class activations back to the input pixels. Capturing Top-Down Visual Attention With Feedback Convolutional Neural Networks, A. Mahendran and A. Vedaldi [10] Understanding Deep Image Representations by Inverting Them, recent top-down saliency methods recover pixel importance for a given class using isolated object labels, we extend the the idea to linguistic sentences. D. Bahdanau, K. Cho, and Y. Bengio [11], "Neural Machine Translation by jointly learning to Align and Translate.", proposes a neural network architecture for machine translation that uses an attention mechanism to align the sources and target sentences during translation. L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville [12], Instead of treating all image regions equally, soft attention assigns different weights to different regions depending on their content. Similarly, in the video captioning, an LSTM with a soft attention layer attends to specific temporal segments of a video while gener- ating the description. M. T. Luong, H. Pham, C. D. Manning [13]," Effective approaches to attention-based neural machine translation", presents a novel approach to neural machine translation that utilizes an attention mechanism to improve translation quality. Feng, Yansong and Mirella Lapata [14]," How many words is a picture worth?". This paper discusses the challenges of image captioning including the need for semantics and context of images. The authors introduced novel attention mechanisms to improve the quality of generated captions. Lisa Anne Hendricks, Subhashini Venugopalan, M. Rohrbach [15]," Deep Compositional captioning: Describing novel object categories", have some limitations such as object segmentation errors, limited vocabulary, inability to handle multiple objects. One disadvantage of soft attention compared to our top-down saliency model is that it necessitates an additional recurrent layer in addition to the LSTM decoder, requiring additional designing of this extra layer's parameters. This layer's size scales proportionally to the number of weighted items, or the number of spatial areas or frames. Contrarily, our method decoder-decoder models extract the mapping between input pixels and output words without the need for explicit modeling of temporal or spatial attention, and without changing the network.

III. MOTIVATION

A critical task in the fields of computer vision and natural language processing is the creation of captions for photographs. People are capable of giving descriptions. It is a tremendous advancement in the field of artificial intelligence for people to be able to create images using their everyday experiences and for the computer to be used for this purpose. The key issue in this work is to convey the relationships be- tween various items in a natural language while also capturing their relationships in both the real world and in images (like English). Until now, text descriptions for photos have been generated by computers using preset templates. This method, however, does not offer the variation needed to produce accurate visual descriptions.

Neural networks can be used to enhance the effectiveness of image captioning by improving accuracy specifically neural networks can be used to learn the relationship between image features and corresponding captions

IV. METHODOLOGY

The vanishing gradient problem in the standard CNN-RNN model prevents the recurrent neural network from learning and being trained effectively. So, in order to alleviate this gradient descent difficulty, according to a survey in order to boost both the efficiency and the accuracy of the caption generation for the image. A flow diagram is shown in figure 1.

As the traditional CNN-RNN model's vanishing gradient issue makes it difficult for the recurrent neural network. Therefore, the model is for improving both the efficiency and the accuracy of the caption creation for the image in order to ease the gradient descent issue.

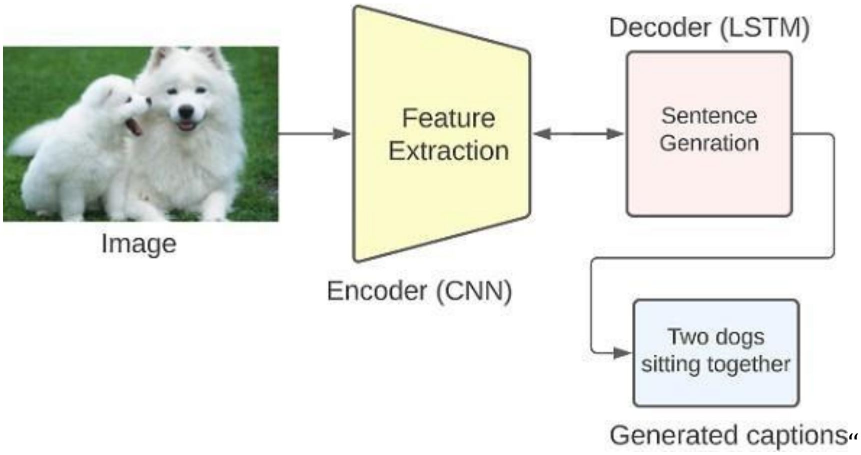

Fig. 1. Data Flow Diagram

V. DATA COLLECTION

There are numerous open-source datasets addressing this issue, including MS COCO (which contains 180k photos), Flickr8k (which contains 8k images), and Flickr30k (which contains 30k images). The dataset considered for this case study is Flick8k dataset. It contains 8000 images, each with five captions for testing purpose. These images are divided into three parts: Training Set has 6000 images, Development Set has 1000 images and the Test Set has 1000 images. The captions in this dataset are manually annotated and designed to be both descriptive and informative. According to literature surveys, this dataset has been widely used in research on image captioning. One of the advantages of this dataset is that it allows researchers to train and test models on a wide range of visual content. It provides a large and diverse set of images.

VI. CONCLUSION

We have looked at methods for captioning images that use deep learning. We've outlined the benefits and drawbacks of each method, offered a taxonomy of picture captioning techniques, and shown broad block diagrams of the main categories. We briefly talked about further directions for this field's research are possible. Although deep learning-based picture captioning algorithms have advanced significantly in recent years, a dependable method that can produce high- quality captions for nearly all photos has not yet been created. With the advent of unique deep-learning network designs, automatic picture captioning will remain a hotly debated subject for a while. The captions for the over 8000 photographs that make up the Flickr 8k dataset that we used are also included in the text file.

We had completed the text cleaning, image preprocessing, and dataset understanding up until this point in the project. The potential for image captioning is very vast in the future because more individuals use social media every day, and the majority of them post images.

VII. FUTURE SCOPE

Image caption generator has a lot of potential for future development and applications. To improve the quality of the image caption, multi-modal learning can be introduced in the future which can combine information from different sources like images, videos, and text. In self-driving cars, captions can be developed in real-time which can enhance the function- ality of applications. Self-driving cars can be beneficial for visually-impaired individuals. Currently, many image caption generation models can create captions only in the English language. But there is growing demand for captions in multiple languages. This can enable communication across different cultures and communities. Today's model can create one line caption for an image but this can be improved so that it can explain the content of the image in natural language.

There is some future scope in the medical field such as disease diagnosis. Image caption generator can be used to automatically generate accurate and informative descriptions of medical images, such as X-rays, and CT scans.

ACKNOWLEDGEMENT

We express our appreciation to the Pune Institute of Com- puter Technology for providing this project-based learning opportunity. We would like to extend our sincere gratitude to Dr. S. T. Gandhe sir, our project guide, and Prof. V.

A. Patil, our project co-guide for his unwavering support and suggestions throughout the project work suggestions that proved to be helpful in the overall progress of this project, for helping us complete stage 1 of the project successfully. The project is titled ”Image description generator using recurrent neural and convolution network.” He taught us a lot, and we learned a lot. Additionally, we appreciate the cooperation of the Director Dr. P. T. Kulkarni and Principal Dr. S. T. Gandhe, and HOED Dr. Mousami Munot. We also value the support and assistance that our friends gave us in finishing the project.

VIII. REFERENCES

[1] Aishwarya Maroju, Sneha Sri Doma, Lahari Chandarlapati, 2021, Image Caption Generating Deep Learning Model, INTERNATIONAL JOUR- NAL OF ENGINEERING RESEARCH TECHNOLOGY (IJERT) Volume 10, Issue 09 (September 2021).

[2] Shetty, R., Rezazadegan Tavakoli, H., Laaksonen, J. (2018). Image and Video Captioning with Augmented Neural Architectures. IEEE Multimedia, 25(2), 34-46.

[3] Geetha,T.Kirthigadevi,G GODWIN Ponsam,T.Karthik,M.Safa,” Image Captioning Using Deep Convolutional Neural Networks(CNNs)” Pub- lished under license by IOP Publishing Ltd in Journal of Physics: Conference Series , Volume 1712, International Conference On Com- putational Physics in Emerging Technologies(ICCPET) 2020 August 2020,Manglore India in 2015

[4] Syed Haseeb, Srushti G M, Bhamidi Haripriya, Mrs. Madhura Prakash, 2019, Image Captioning using Deep Learning, INTERNA- TIONAL JOURNAL OF ENGINEERING RESEARCH TECHNOLOGY (IJERT) Volume 08, Issue 05 (May 2019).

[5] M. A. Kastner et al., ”Imageability- and Length-Controllable Image Captioning,” in IEEE Access, vol. 9, pp. 162951-162961, 2021, doi: 10.1109/ACCESS.2021.3131393.

[6] O. Vinyals et al., “Show and Tell: A Neural Image Caption Generator,” Proc. 2015 IEEE Conf. Computer Vision and Pattern Recognition (CVPR 15), 2015. doi.org/10.1109/CVPR.2015.7298935.

[7] Andrej Karpathy, Li Fei-Fei, ”Deep visual-semantic alignments for generating image captions.”

[8] Ramanishka, V., Das, A., Zhang, J., Saenko, K. (2016). Top-down Visual Saliency Guided by Captions. ArXiv. /abs/1612.07360

[9] M. D. Zeiler and R. Fergus. Visualizing and Understanding Convolu- tional Networks. In European

Conference on Computer vision, pages 818–833. 2014

[10] A. Mahendran and A. Vedaldi. Understanding Deep Image Representa- tions by Inverting Them. In IEEE conference on computer vision and pattern recognition, 2015.

[11] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In Inter- national Conference on Learning Representations (ICLR).

[12] Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., Courville, A. (2015). Describing Videos by Exploiting Temporal Struc- ture. ArXiv. /abs/1502.08029

[13] M.-T. Luong, H. Pham, C. D. Manning, and Q. V. Le, ”Effective Approaches to Attention-based Neural Machine Translation,” in Pro- ceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), Lisbon, Portugal, September 2015

[14] Feng, Yansong, and Mirella Lapata. ”How many words is a picture worth? automatic caption generation for news images.” Proceedings of the 48th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2010

[15] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Ray- mond Mooney, Kate Saenko, Trevor Darrell, Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, et al. 2016. Deep compo- sitional captioning: Describing novel object categories without paired training data. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.