ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 10, Issue 5,May 2024

Enhanced Image Caption Generation Using VGG16 and LSTM Networks

Akhila Alaparthi[1], V.Ramyasri [2], M.Jaswanth[3], G.Surendra[4], N.Janaki Ram[5]
Assistant Professor [1], Student [2][3][4][5]
SRK Institute of Technology, Department of Information Technology [1][2][3][4][5]
Vijayawada, India
Mail Id: akhilaalaparthy@gmail.com[1], ramyasri93900@gmail.com[2],
mangamurijaswanth1224@gmail.com[3] ,
gayamsurendra21@gmail.com[4], janaki.namburi@gmail.com[5]

Abstract: The primary objective of this endeavor is to improve the generation of image captions by combining the visual feature extraction capabilities of the VGG16 architecture with the sequential data processing strength of LSTM networks. By utilizing a benchmark dataset consisting of 8,000 images, each accompanied by five unique captions that identify significant entities and events, the system strives to generate accurate text across a wide range of scenarios obtained from various Flickr groups. By attempting to improve the coherence and relevance of captions, this methodology seeks to expand the capabilities of automated image description and retrieval systems. The efficient extraction of high-level visual features from images is made possible through the use of VGG16. On the other hand, LSTM networks demonstrate exceptional performance in modeling sequential dependencies present in textual data, thereby guaranteeing the generation of accurate and contextually rich captions. The robust training of the model is facilitated by the inclusion of multiple captions per image in the dataset, which allows for the capture of diverse perspectives and subtleties that are inherent in the interpretation of images. Through the integration of these methodologies, the system endeavors to surmount obstacles including uncertainty and fluctuation in the interpretation of images, thus augmenting the caliber of captions produced. This development has substantial ramifications for accessibility, content comprehension, and image search, as it facilitates more accurate and informative depictions of visual content, thereby augmenting user experience and practicality.

Keywords: *LSTM, visual features, VGG16, integration methodology, interpretation of images.*

I. Introduction

One of the most recent developments in the field of computer vision and machine learning is the process of training computers how to provide captions for images automatically. This endeavour includes a number of components, including the comprehension of image scenes, the extraction of features, and the translation of visual representations into regular languages. This effort holds a great deal of promise in a number of areas, including the creation of different kinds of assistive devices for persons who are blind and the provision of aid with captioning responsibilities. The purpose of this project is to come up with captions that are suitable for the image that has been provided. The captions will be selected in a manner that is reflective of the context being offered by the photos. The methods that are currently in use make use of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) or their modifications in order to generate captions that are deemed acceptable. Through the utilization of recurrent neural networks (RNNs) as decoders to supply linguistic descriptions and VGGs to encode the image into feature vectors, these networks offer an encoder-decoder technique that can be utilized to accomplish this objective. The architectures of the models could be very different. When it comes to the encoder component, features from an image are retrieved with the help of a CNN that has been pre-trained. Following the incorporation of the image into the CNN, these distinctive characteristics are extracted from particular layers. After that, the feature vector is incorporated into the architecture for the purpose of picture captioning. CNNs such as AlexNet, VGGNet, ResNet, and GoogleNet (Inception models) are the ones that are utilized for decoding the message the most commonly. In order to complete the decoder part, multiple RNNs are utilized. Long-short-term memory, often known as LSTM, are particularly useful for managing sequences that have significant long-term dependencies. The process of embedding involves the incorporation of words from specified dictionaries into LSTMs. The transformation of words into vectors is accomplished by the utilization of statistical methods and algorithms. Images and captions taken from big data sets that have been labeled, such as those found in Microsoft COCO and Flickr, provide information about the events and things that have occurred. In order to compose meaningful explanations for brand-new photographs, you can make use of these captions. There are metrics such as BLUE that may be utilized to evaluate the appropriateness and relevance of the generated captions in relation to the captions of the data set. A BLEU score is determined by the length of the caption as well as the general resemblance between the two. Over the course of the last few decades, a number of relevant research papers have sought to carry out this task; nevertheless, they have encountered a number of obstacles as a result of their efforts. These obstacles include linguistic problems, cognitive absurdity, and irrelevant content. As a means of resolving those challenges, we have devised this method, in which we make use of computer vision and natural language processing techniques to obtain pertinent content that is accompanied by appropriate sentence formulation. This model is designed to assist individuals who are visually impaired in using it.

II. RELATED WORK

The integration of computer vision (CV) and natural language processing (NLP) techniques was employed to achieve image captioning prior to the advent of deep learning. Encoder-decoder architectures and other deep recurrent model implementations have significantly improved the performance of image captioning. The encoder of these models extracts feature vectors from input images using convolutional neural networks (CNNs), while the decoder generates descriptive sentences using recurrent neural networks (RNNs). In recent times, the integration of object detection algorithms has enabled the provision of more extensive captions that pertain to specific areas within an image. Karpathy and Fei-Fei introduced a deep visual-semantic alignment model in their research, which assigns scores to words and regions via object detection. They trained a multi-modal recurrent neural network (m-RNN) with image caption data and these scores in order to generate region-based phrases. Johnson et al. proposed DenseCap, a system that generates dense captioning for particular regions by employing fully convolutional localization networks and an RNN language model trained on extracted features. Attention mechanisms, which have been the subject of extensive investigation in computational neuroscience, have gained prominence in deep learning applications across various domains, such as natural language processing and speech recognition. For the purpose of image captioning, encoder-decoder models have integrated attention. To partition the image, feature vectors are generated by the encoder for each grid region. The decoder employs the weights allocated to these vectors by the attention model in order to generate context vectors designed for captioning. By incorporating attention, models demonstrate improved performance, exemplifying how attention guides itself towards specific areas of an image in order to produce captions. Chen et al. devised multiple attention models for spatial, activation, and object features that exhibited superior performance than single attention models. The developments mentioned above underscore the importance of attention mechanisms and deep learning in the improvement of captioning systems for images. This enhances their ability to generate cogent and contextually relevant descriptions, thereby advancing the domain of automated image comprehension and interpretation.

III. PROPOSED MODEL

The model being evaluated is the "explainable image caption generator," which is specifically engineered to generate captions that faithfully depict the objects identified within an image. Furthermore, it strives to offer a justification for the production of every caption. The architecture is composed of two fundamental modules, namely generation and explication. The generation module is designed with an encoder-decoder architecture. Through the utilization of the VGG-16 model, the encoder produces a feature vector that incorporates the entirety of the input image, ensuring that the size conversion remains consistent across all images. Following this, the feature vector is transmitted to the decoder, which employs a long short-term memory (LSTM) network to produce words in a sequential fashion. The ultimate layer of the LSTM utilizes a softmax function to produce individual words, while the training process is facilitated by a negative log likelihood loss function. Throughout the training process, the encoder and decoder parameters are optimized in tandem to improve the precision of the captioning. The capability to identify particular image components, however, is not present in this module.

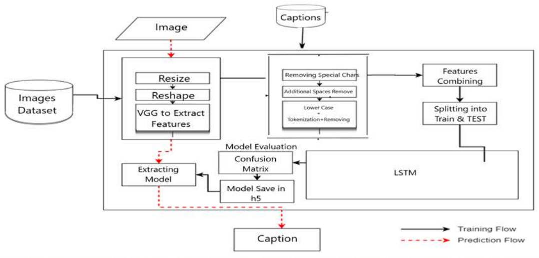


Fig.1 Architecture

To surmount this limitation, the explanation module is integrated, functioning as a dual-purpose component during both the training and testing stages. During the training process, the algorithm produces the image-sentence relevance loss (Losse), a metric that assesses the extent to which the generated caption considers the objects that were detected. Enhanced object-related caption generation is made possible by loss during the training procedure. During testing, the explanation module produces a weight matrix that illustrates the correlation between regions extracted from the input image and words generated by the generation module. The ultimate outcome selects the most substantial weight values with the intention of emphasizing critical associations. The interpretability enhancement (IE) model and the region-word attention model comprise the explanation module. The region-word attention model considers both words extracted from the generated caption and regions acquired from object detection when constructing the weight matrix. During this time, the IE model computes Losse, a metric that evaluates the accuracy of the generated caption in representing the identified objects, by utilizing the weight matrix. In summary, the model being evaluated integrates explanation and generation modules to produce captions that not only provide an accurate depiction of detected objects but also provide rationale by employing graphical correlations between words and image regions. By enhancing understanding of the reasoning behind the creation of specific captions, this methodology improves the interpretability and usability of image captioning initiatives. [4] discussed that Liver tumor division in restorative pictures has been generally considered as of late, of which the Level set models show an uncommon potential with the advantage of overall optima and functional effectiveness. The Gaussian mixture model (GMM) and Expected Maximization for liver tumor division are introduced.

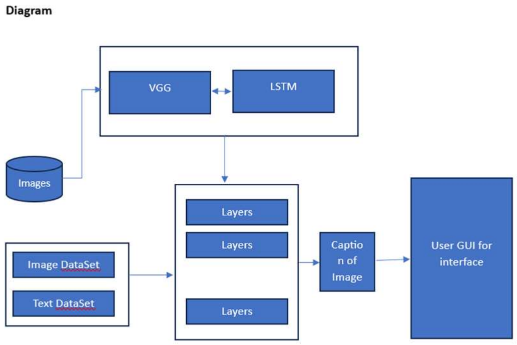


Fig.2 Block Diagram

IV. EXPERIMENTS AND RESULTS

The investigations were conducted utilizing the three datasets listed below: MSCOCO, Flickr8K, and Flickr30K. MSCOCO is composed of a cumulative sum of 40,775 test images, 40,783 training images, and 40,504 validation images. Flickr8K comprises a total of 8,000 images, of which 1,000 are designated for training purposes and 1,000 are intended for assessment. Flickr30K, which was created as an extension of Flickr8K, is comprised of 31,783 images with 158,915 captions. In order to promote fair and unbiased comparisons with prior works, the allocation of captions is as follows: 29,000 are designated for training purposes, 1,000 are designated for validation, and 1,000 are designated for testing. Each image in the datasets is accompanied by five descriptive captions. The evaluation of the outcomes is focused on the explanation module, which illustrates the region-word attention model's qualitative performance. The model extracts regions containing detected objects and captions produced by the generation module from the input image. Subsequently, a weight matrix is produced to depict the correlation between the significance of objects and words. This procedure is illustrated through the use of a test image, its predicted caption, and the corresponding weight matrix. Distinct regions are delineated by a variety of colored borders, while the anticipated caption is presented in the text beneath the image. The relevance evaluations between objects and words are illustrated in the weight matrix. The regions annotated with the word "person" carry heavier weights for terms such as "people" and "group," which are frequently associated within the training dataset (see Figure 3). When regions labeled "screen" and "laptop" are paired with relevant terms to indicate their semantic proximity, a comparable pattern is discernible. These associations facilitate understanding of the reasoning that underpins the choice of specific words in the generated captions.

The performance of the entire model is assessed by the colored words that are used to depict regions of the image in the generated caption. A weight value is assigned to each word in the caption, which is obtained from the attention model. One word is present in each detected region due to the training procedure. Further training diversification, as depicted in Figure 5, yields intriguing results when two region-word pairs are selected for every detected region. The generated caption for specific detected regions contains no more than two words, as illustrated in Figure 5. This constraint is implemented to provide clarification regarding both the object and contextual concepts. In the caption "A man is standing on a skateboard in front of a car," for instance, the terms "man" and "standing" are employed to illustrate how the phrase "person region" captures both the object and its state. Under specific conditions, however, the allocation of two words to a single region may not be the most efficient course of action, as evidenced by the low weight values of certain words. [6] discussed that The study of viruses and their genetics has been an opportunity as well as a challenge for the scientific community. The recent ongoing SARSCov2 (Severe Acute Respiratory Syndrome) pandemic proved the unpreparedness for these situations. Not only the countermeasures for the effect caused by virus need to be tackled but the mutation taking place in the very genome of the virus is needed to be kept in check frequently. One major way to find out more information about such pathogens is by extracting the genetic data of such viruses.

combinations.

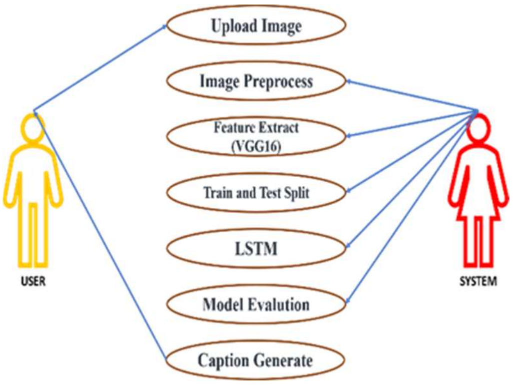
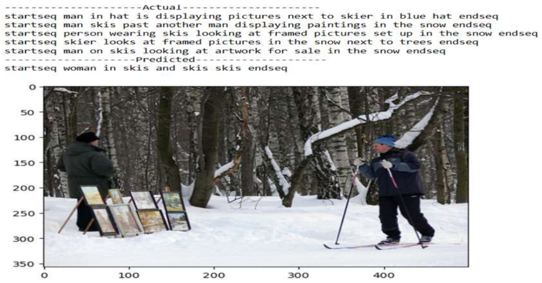


Fig.3 Use Case Diagram



```
----------------------Actual----------------------
startseq man in hat is displaying pictures next to skier in blue hat endseq
startseq man skis past another man displaying paintings in the snow endseq
startseq person wearing skis looking at framed pictures set up in the snow endseq
startseq skier looks at framed pictures in the snow next to trees endseq
startseq man on skis looking at artwork for sale in the snow endseq
---------------------Predicted--------------------
startseq woman in skis and skis skis endseq
```

Fig.4 Sample Out Put

The effective association of regions and words by the model provides evidence for the selection of words in the generated captions. Although the incorporation of contextual concepts enhances the capacity to elucidate, extreme care must be taken to optimize the correlations between terms and regions to ensure captions possess significance.

V. CONCLUSION

The objective of this research is to develop an explainable image caption generator that produces captions while considering the objects within the image and offering justifications for the process of word selection. The explanation module of our model computes a relevance loss between images and sentences, which serves as a reference point for the training process of the generation module. Moreover, to enhance the graphical representation of these correlations, the explanation module produces a weight matrix illustrating the associations between particular areas of the input image and the terms that are included in the generated caption. The findings of our inquiries revealed that our model generated descriptive captions and insightful explanations for the output in an effective manner. We demonstrated the merits of our methodology through qualitative analysis, with an emphasis on its ability to generate captions that are contextually relevant and derived from the objects detected. Moving forward, our aim is to enhance our model by addressing any deficiencies that have been identified. The aim of our research is to develop a semantic attention module capable of discerning more complex attributes that are associated with individual regions. By implementing this enhancement, our model will be capable of assimilating semantic information at a finer level, leading to the production of captions that are more sophisticated and precise. It is anticipated that the integration of a semantic attention mechanism will result in further improvements to the caliber of captions and the lucidity of explanations. The aforementioned advancement aligns with our aim of enhancing the existing standard of image captioning by incorporating sophisticated attention mechanisms and semantic comprehension. In essence, the explainable image caption generator we have proposed represents a significant progression in the field of image comprehension and captioning. Our primary goal is to greatly enhance the interpretability and descriptive power of our model through continuous innovation and refinement. This will allow us to make a significant contribution to the advancement of automated image captioning systems.

VII. References

[1] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 2961-2969).

[2] Waqas Zamir, S., Arora, A., Gupta, A., Khan, S., Sun, G., Shahbaz Khan, F., ... & Bai, X. (2019). isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 28-37).

[3] Tian, Z., Shen, C., Wang, X., & Chen, H. (2021). Boxinst: High-performance instance segmentation with box annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5443-5452).

[4] Christo Ananth, M Kameswari, Densy John Vadakkan, Dr. Niha.K., "Enhancing Segmentation Approaches from Fuzzy-MPSO Based Liver Tumor Segmentation to Gaussian Mixture Model and Expected Maximization", Journal Of Algebraic Statistics, Volume 13, Issue 2, June 2022,pp. 788-797.

[5] Shen, X., Yang, J., Wei, C., Deng, B., Huang, J., Hua, X. S., ... & Liang, K. (2021). Dct-mask: Discrete cosine transform mask representation for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8720-8729).

[6] Christo Ananth, Pranav Pushkar, Preeti Nagrath, Jehad F. Al-Amri, Vividha, Anand Nayyar, "Mutation Prediction for Coronaviruses using Genome Sequence and Recurrent Neural Networks", Computers, Materials & Continua, Tech Science Press, Volume 73, Issue 1, April 2022,pp. 1601-1619.

[7] Wang, X., Zhang, R., Kong, T., Li, L., & Shen, C. (2020). Solov2: Dynamic and fast instance segmentation. *Advances in Neural information processing systems*, *33*, 17721-17732.

[8] Xie, E., Sun, P., Song, X., Wang, W., Liu, X., Liang, D., ... & Luo, P. (2020).

Polarmask: Single shot instance segmentation with polar representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12193-12202).

[9] Cheng, T., Wang, X., Chen, S., Zhang, W., Zhang, Q., Huang, C., ... & Liu, W. (2022). Sparse instance activation for real-time instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4433-4442).

[10] Chen, H., Sun, K., Tian, Z.,Shen, C., Huang, Y., & Yan, Y. (2020). Bendmask: Top-down meets bottom-up for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8573-8581).

[11] Zimmermann, R. S., & Siems, J. N. (2019). Faster training of Mask R-CNN by focusing on instance boundaries. *Computer Vision and Image Understanding*, *188*, 102795.

[12] Zhang, G., Lu, X., Tan, J., Li, J., Zhang, Z., Li, Q., & Hu, X. (2021). Refinemask: Towards high-quality instance segmentation with fine-grained features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6861-6869).

[13] Zhang, Y., Chu, J., Leng, L., & Miao, J. (2020). Mask-refined R-CNN: A network for refining object details in instance segmentation. *Sensors*, *20*(4), 1010.

[14] Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., ... & Lin, D. (2019). Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4974-4983).

[15] Kong, S., & Fowlkes, C. C. (2018). Recurrent pixel embedding for instance grouping. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 9018-9028).

[16] Wang, W., Feiszli, M., Wang, H., Malik, J., & Tran, D. (2022). Open-world instance segmentation: Exploiting pseudo ground truth from learned pairwise affinity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4422-4432).

[17] Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018). Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8759-8768).

[18] Han, Y., Han, Z., Wu, J., Yu, Y., Gao, S., Hua, D. and Yang, A., 2020. Artificial intelligence recommendation system of cancer rehabilitation scheme based on IoT technology. Ieee Access, 8, pp.44924-44935.

[19] Peng, S., Jiang, W., Pi, H., Li, X., Bao, H., & Zhou, X. (2020). Deep snake for real-time instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8533-8542).