World Journal of Engineering Research and Technology
WJERT
www.wjert.org
SJIF Impact Factor: 7.029

IMAGE CAPTION GENERATION USING DEEP LEARNING

Krishna Swaroop A., Nithin K., Prarthana K.*[1], Ranjitha M. B.[2], Vidya Y. C.[3], Bhanupriya D. R.[4]

Assistant Professor,

[*1,2,3,4]Department of Information Science & Engineering, Malnad College of Engineering, Hassan.

Article Received on 11/04/2024 Article Revised on 01/05/2024 Article Accepted on 23/05/2024

***Corresponding Author**
Prarthana K.
Department of Information Science & Engineering, Malnad College of Engineering, Hassan.

ABSTRACT

The image caption generator is one of the processes of identifying images and providing similar captions using deep learning and computer vision techniques. It involves labeling an image with English keywords based on datasets used during model training. This process generates descriptions that explain the context of the image or provide descriptions of the image content. It is particularly useful for applications such as analyzing unstable or unlabeled images and identifying hidden patterns for machine learning applications, such as self-driving cars and software for visually impaired individuals. Using deep learning models, image captioning can be achieved efficiently. With advanced techniques in deep learning and natural language processing, generating image captions for given images has become easier. In the field of artificial intelligence, image caption generationis of great interest to researchers. Over the years, it has presented a challenge for those studying artificial intelligence.

KEYWORDS: Image Captioning, Deep Learning, Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Multimodal Learning, Natural Language Processing (NLP).

I. INTRODUCTION

Over the years, generating captions for images has been a challenging task, often requiring relevance between the image and its caption. However, with the advancements in neural networks and natural language processing, many previously daunting tasks in machine learning have become easier, effectively addressing various challenges. These technologies have proven invaluable for tasks such as image recognition, classification, and captioning within artificial intelligence. Image captioning essentiallyinvolves generating a description of a given input image. This process takes the image as input and outputs or generates the caption. With the use of advanced technology, generating image captions has become easier, and efficiency is increasing. For the purpose of image captioning, various models have been created, including object detection models, visual attention-based image captioning, and image captioning using deep learning techniques. Within deep learning, different model types exist, such as the inception model, VGG model, ResNet-LSTM model, and traditional CNN-RNN model. Creating or generating image captions for a given image requires a combination of computer vision methods to understand the image's content and a language model from the field of natural language processing to convert the image into understandable words in the correct order.

II. LITERATURE SURVEY

The automatic generation of a caption describing an image is crucial for facilitating smoother human-machine interaction. One of the earliest successful methods in this domain, which served as a precursor to modern image-captioning techniques, is referenced in.[4] This method employs a typical encoder-decoder architecture, initially creating a representation of the image and then generating text based on this representation, typically word by word. Most contemporary image-captioningarchitectures follow a similar structure.

Two seminal works that have notably enhanced the performance of image-captioning models are cited in.[8,9] The first work proposes utilizing the REINFORCE algorithm to directly optimize the discrete model's quality metric known as CIDEr, thereby enhancing its value directly. The second work suggests employing an object detector in the image as an encoder (the study itself uses Faster R-CNN[24]), and subsequently generating a caption based on the detected objects and their attributes, rather than attempting to compress the entire image into a singlevector and generating text based on such a representation.

Furthermore, beginning with reference[5], the attention mechanism has been actively incorporated into image-captioning tasks. This mechanism enables the model to dynamically allocate attention to objects or regions that are most relevant for generating a textual description of the current image. High-qualityworks such as those referenced in[10,33,34] utilize this approach.

Recently, transformers[14] have gained increasing popularity, emerging as state-of-the-art models for numerous tasks within natural language processing (NLP). Their integration enhances the quality of image-captioning models as well, as explored in.[15,16,18]

Additionally, research in the realm of image captioning is progressing towards the integration of models with other tasks related to vision-language understanding and generation. Studies such as those mentioned in[35,36] leverage networks pretrained onextensive datasets to develop models capable of addressing various vision-language challenges. However, due to the significant computational resources required for such pretraining, researchers encounter challenges in refining these models and conducting comprehensive studies on them.

III. EXISTING SYSTEM

The most extensively studied aspects of computer vision encompass image recognition and object detection. Social media users now have the ability to upload photographs of varying sizes and complexities and utilize search engines like Google to find descriptions. However, there are notable deficiencies in terms of upgradeability, performance, flexibility, and scalability. High-quality images are essential for accurate results, as low-resolution photographs may obscure important features and make complex scenes difficult to analyze. Employing proxies aims to expedite the picture search process. Processing intricate input images can be time-consuming, potentially hindering users from uploading grayscale images or generating captions through speech.

IV. PROPOSED SYSTEM

Deep Neural Networks can effectively address the challenges encountered in both scenarios by generating suitable, expressive, and seamless subtitles, thus expediting the subtitle creation process. With our system, social media users will no longer need to spend hours searching for subtitles on Google. Our technology offers a user-friendly platform for uploading selected photographs, eliminating the need for manual caption input by users. The proposed framework efficiently resolves the picture search issue, accommodating color and black-and-white photos of any size, while also capable of vocalizing the captions in English. Neural networks leverage Tensor Flow and algorithms to tackle various problems and produce suitable, expressive, and fluent subtitles. Automatic metrics can be calculated efficiently, eliminating the need for time-consuming caption searches as subtitles will be generated automatically.

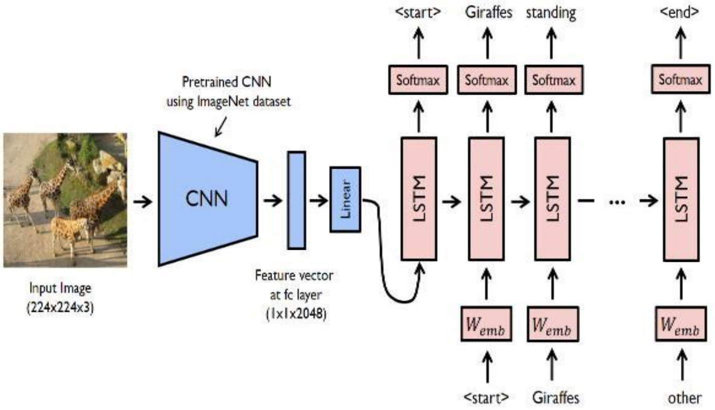
A. Task

The objective is to develop a system that accepts an image in the form of a dimensional array, analyzes itscharacteristics, and generates syntactically and grammatically accurate statements as output.

B. Corpus

The Flickr 8K dataset is utilized for this purpose, comprising 8000 photos, each accompanied by five captions. This comprehensive collection enables the system to cover a wide range of conceivablescenarios. The dataset is divided into training(Flickr8k.trainImages.txt with 6,000 photos), development (Flickr8k.devImages.txt with 1,000 images), and test (Flickr8k.testImages.txt with 1000 images) sets.

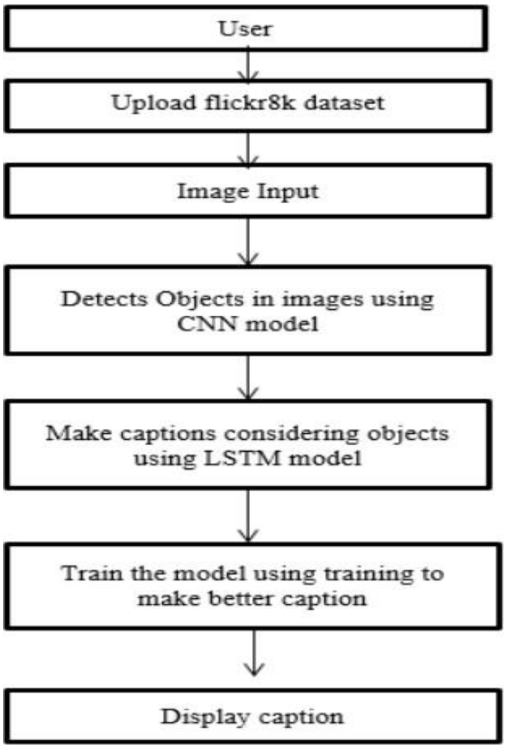
In summary, employing LSTM and CNN on the Flickr 8K dataset for image-caption generation provesto be a robust solution for producing textual descriptions for images. This technology finds applications across various domains, including image search engines, automatic image captioning for the visually impaired, and social media platforms.



WORKING MODEL

A. Convolutional Neural Network

A Convolutional Neural Network (ConvNet/CNN) is a Deep Learning system designed to process images. It takes an image as input and learns to assign importance (learnable weights and biases) to various aspects or objects within the image, enabling it to distinguish between them. Compared to other classification algorithms, ConvNets require significantly less preprocessing. They are specifically designed to handle data in the form of 2D matrices, making them particularly effective for processing images. CNNs scan images from left to right and top to bottom, extracting important features before combining them for classification. This ability allowsCNNs to effectively work with images that have beenaltered, rotated, scaled, or otherwise transformed.


User
Upload flickr8k dataset
Image Input
Detects Objects in images using CNN model
Make captions considering objects using LSTM model
Train the model using training to make better caption
Display caption

PROJECT ARCHITECTURE

B. Long Short-Term Memory

LSTM, or long short-term memory, is a type of recurrent neural network (RNN) specialized in addressing sequence prediction tasks. Its architecture enables it to effectively predict the next phrase basedon preceding context. LSTM has demonstratedsuperiority over traditional RNNs by addressing their limitations in retaining information over long sequences. It achieves this by leveraging its ability tomaintain relevant information over extended periods while discarding irrelevant information through the use of "forget gates." This allows LSTM to perform accurate computations during input processing, leading to enhanced performance in various tasksrequiring sequential analysis.

C. Flickr8k Dataset

The Flickr8k dataset serves as a publicly available benchmark for image-to-text instruction tasks.

Comprising 8000 photos, each accompanied by five captions, this dataset offers a rich and diverse collection of visual and textual data. These photos were sourced from various Flickr groups, capturing awide array of events and settings rather than focusingsolely on renowned individuals or landmarks. The dataset's breadth enhances its applicability to diverse projects.

The dataset is partitioned into three subsets: a training dataset containing 6000 photos, a developmentdataset with 1000 images, and a test dataset alsoconsisting of 1000 images.

Several properties of this dataset make it suitable for the project at hand.

- The presence of multiple captions for each image ensures that the model is exposed to diverse descriptions, thereby reducing the risk of overfitting and promoting generalization.
- The variety of image categories in the trainingdataset enables the image annotation model to work effectively across different picture categories, enhancing its robustness and adaptability to various scenarios.

V. RESULTS

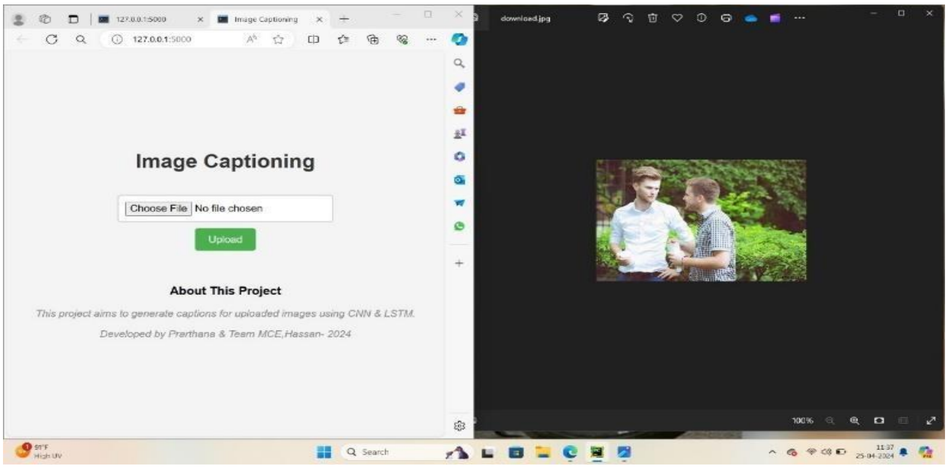


Image Captioning

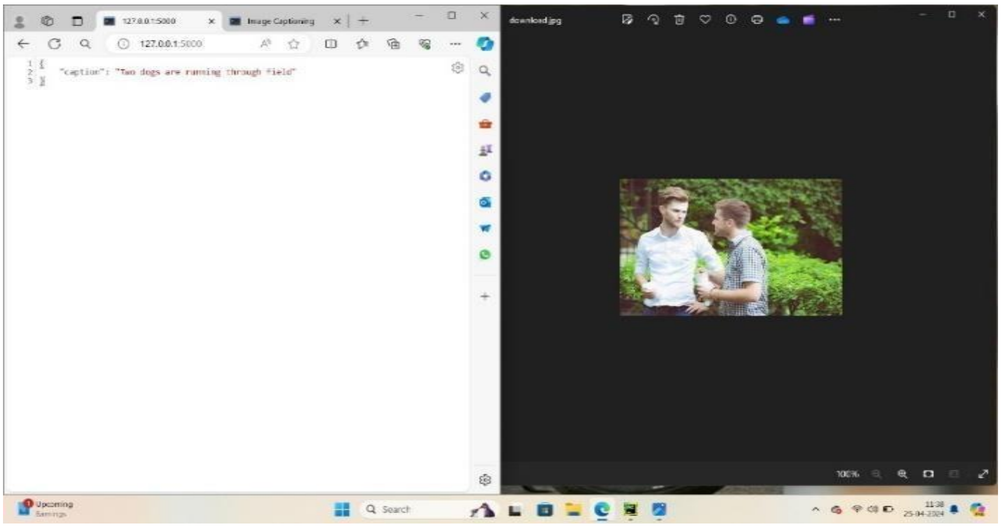
Choose File No file chosen

Upload

About This Project

This project aims to generate captions for uploaded images using CNN & LSTM.

Developed by Prarthana & Team MCE,Hassan- 2024


```
{
    "caption": "Two dogs are running through field"
}
```

The paper presents an Image Caption Generator that stands out due to its incorporation of online photo upload functionality. Unlike previousreference papers, this feature allows users to directly upload images from social media platforms, eliminating the need for manual caption input. By leveraging CNNs for image feature extraction and LSTM networks for natural language sentence generation, the proposed system offers a seamless and user-friendly experience. Through the integration of online photo upload, the Image Caption Generator improves image accessibility for visually impaired individuals by providing automated descriptions for uploaded images. This unique feature distinguishes the proposed system, making it a valuable contribution to the field of image captioning.

VI. CONCLUSION

In this paper, we have introduced and developed a technique for an Image Caption Generator capable of providing users with captions or descriptions based on input images. The model consists of two main components: an Image-Based Model responsible for extracting features from the image, and a Language- Based Model tasked with translating these features and identified objects into natural language sentences. The Image-Based Model utilizes Convolutional Neural Networks (CNN), while the Language-Based Model employs Long Short-Term Memory (LSTM) networks.

The workflow of our approach involves several steps, starting with data gathering, followed by pre- processing, model training, and finally, predictiongeneration.

The primary objective of the Image Caption Generator is to enhance social media platforms by enabling automatic generation of captions or descriptions for images. Additionally, it aims to improve image indexing and accessibility for visually impaired individuals by providing automated descriptions for images.

VII. REFERENCES

1. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3156–3164. [Google Scholar]
2. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 June 2015; pp. 2048–2057. [Google Scholar]

3. Rennie, S.J.; Marcheret, E.; Mroueh, Y.; Ross, J.; Goel, V. Self-critical sequence training for image captioning. In Proceedingsof the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7008–7024. [Google Scholar]
4. Anderson, P.; He, X.; Buehler, C.; Teney, D.;Johnson, M.; Gould, S.; Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6077–6086. [Google Scholar]
5. Huang, L.; Wang, W.; Chen, J.; Wei, X.Y. Attention on attention for image captioning. In Proceedings of the International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 4634–4643. [Google Scholar]
6. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008. [Google Scholar]
7. Cornia, M.; Stefanini, M.; Baraldi, L.; Cucchiara, R. Meshed-Memory Transformer for Image Captioning. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Online, 14–19 June 2020; pp. 10578–10587. [Google Scholar]
8. Li, G.; Zhu, L.; Liu, P.; Yang, Y. Entangled Transformer for Image Captioning. In Proceedings of the International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 8928–8937. [Google Scholar]
9. Zhu, X.; Li, L.; Liu, J.; Peng, H.; Niu, X. Captioning transformer with stacked attention modules. Appl. Sci, 2018; 8: 739. [Google Scholar] [CrossRef] [Green Version]
10. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. Adv. Neural Inf. Process. Syst. 2015, 28, 91–99. [Google Scholar] [CrossRef] [PubMed] [Green Version]
11. Wang, W.; Chen, Z.; Hu, H. Hierarchical attention network for image captioning. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp.8957–8964. [Google Scholar]
12. Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei,F.; et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In European Conference on Computer Vision; Springer: Cham, Switzerland, 2020; pp. 121–137. [Google Scholar]
13. Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; Gao, J. Vinvl:

Revisiting visual representations in vision-language models. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Online, 19–25 June 2021; pp. 5579–5588.[Google Scholar]

14. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv 2015,arXiv:1409.1556. [Google Scholar]