IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

IMAGE CAPTION GENERATOR

[1]Salma K,[2]Anagha Raj,[3] Midhun M, [4] Akhila S, [5]Chippy T

[1]UG Scholar,[2] UG Scholar,[3]UG Scholar, [4] UG Scholar, [5]Assistant Professor

[1,2,3,4,5] Mahaguru Institute of Technology, Kattachira

Abstract: Image captioning is one of the most frequently requested requirements today. Many applications, including the creation of image search engines with sophisticated natural language queries and assisting those who are blind or visually handicapped in understanding their surroundings, benefit from automatic image captioning. With the use of deep neural network models, all of these tasks are accomplished. Picture captioning is the process of creating a description for a picture. Recognizing the significant things, their characteristics, and the connections between the objects in an image are necessary. In our study, we demonstrate an image captioning technique that creates captions from photos using GAN. Python is what we use for backend development. Our use of an attention-based GAN in this study will aid in better captioning.

Index Terms – GAN, Image Caption Generator, CNN, OpenCV, Kaggle Datasets.

I. INTRODUCTION

Providing a natural language description of an image's content is known as image captioning. It sits at the nexus of natural language processing and computer vision [1]. Many applications, including the creation of image search engines with sophisticated natural language queries and assisting those who are blind or visually handicapped in understanding their surroundings, benefit from automatic image captioning. As a result, research into image captioning has been vigorous. The development of novel object detection architectures and convolutional neural networks has greatly enhanced picture captioning. For precise image caption synthesis, advanced sequential models like attention-based recurrent neural networks have also been proposed.

Most current deep learning-based picture captioning techniques use an encoder-decoder system, which was inspired by neural machine translation. Here, we employ an encoder-decoder framework, with a convolutional neural network (CNN) model serving as an encoder for the extraction of picture features and a long-short-term memory (LSTM) model serving as a language decoder for the creation of captions [2][3]. Generative adversarial networks (GAN) are what we use [4]. GAN is an extremely straightforward approach that uses stochastic gradient descent to attempt to optimize a mathematical equation. We use GAN to produce captions.

What it is capable of:

1. Image identification
2. Finding details in an image
3. Object recognition in an image
4. Image colour detection
5. Use the tools to generate relevant captions.
6. Can produce both text and audio captions.
7. Boost the quality of captions that are generated for actual photos.

An algorithm called GAN, or generative adversarial network, aids in the creation of more effective image descriptions.

II. PROBLEM STATEMENT

To construct an image caption generator with a broad range of meaningful captions with the use of GAN in order to address the issue of limited caption production.

III. MOTIVATION

Social media is being used much more widely. People enjoy using social media to share everything that happens in their lives. They wish to use them to communicate their feelings. It cannot be captured in a single image. If the goal of the image is to convey a certain message and elicit an emotional response, the audience must be aware of what they are looking at. Very few photos can survive without captions, and all classic images have text to support them. Therefore, it could be challenging for people to find captions for those pictures. We used GAN to develop an image caption generator based on this concept.

IV. LITERATURE REVIEW

Automatic image captioning has emerged as a viable study field thanks to developments in deep neural network models. Hossain et al. [5] provide an in-depth analysis of the subject. They categorize the approaches into three groups: novel caption generation, template-based image captioning, and retrieval-based image captioning. Fixed templates with a number of empty slots are used by template-based approaches [6] to generate captions. These techniques start by identifying various objects, properties, and actions, after which the empty spots in the templates are filled. Templates, however, are predetermined and unable to produce captions of variable length. Additionally, visual space and multi-modal space can be used to get captions [7]. Captions are collected from a collection of already-existing captions in retrieval-based systems [8]. These techniques result in captions that are often syntactically sound. They do, however, have limits when it comes to creating syntactically sound captions for individual images [9]. Both visual space and multimodal space can be used to create original captions [10, 11]. A common method in this category is to first evaluate the image's visual information before utilizing a language model to create the image captions. Compared to the previously discussed techniques, these algorithms can produce image captions that are semantically more correct [9].

The majority of techniques in this category generate image captions using an encoder-decoder architecture [10]. These techniques use an LSTM as a decoder to create captions from the visual representations that were extracted using a vanilla CNN acting as the encoder. These techniques, however, struggle to recognize obvious items in the image. Because they selectively focus on the pertinent objects of a picture, attention-based approaches [12] can depict the prominent objects in captions. As a result, we create a description of an image using an attention-based strategy. Three widely used publicly accessible datasets, namely MSCOCO [13], Flickr 30k [14], and Flickr 8k [15], are frequently used by these deep learning-based picture captioning techniques for training and testing the networks. Humans gathered and annotated these datasets. Deep learning-based techniques, however, have certain difficulties using this data.

- To learn the visual representations using these methods, a sizable and varied amount of data is needed.
- The common objects that co-occur in a common environment are overfit by existing models. For instance, if a model is tested in unobserved circumstances like a bed and forest even though it was trained for a situation with a bed and bedroom, The model will have trouble extrapolating from these scenes.Large volumes of data must be manually labelled, which is costly, inaccurate, and time-consuming.

V. EXISTING SYSTEM

Natural language processing and computer vision have both focused on the creation of textual descriptions of images. On this subject, a number of deep learning-based techniques have been put forth. These methods train and evaluate the models using human-annotated photos. For these models to function at their best, a lot of training data is needed. It costs money and takes time to gather human-generated photographs with related captions.

VI. PROPOSED SYSTEM

- Image caption generator based on GAN.
- GAN has been used to create captions from images. For building the backend, we employ Python.
- Perform the feature selection using Mobile Net.
- Analyse each pixel individually to find the features.
- Create pertinent captions based on the features that were seen.

VII. TOOLS AND METHODOLOGY

Hardware specifications: Intel 3 or later processor. Memory (RAM) installed: 4GB or more. 500 GB or more on a hard drive.
Operating system: 64-bit Windows 10 or above.
Display: at least 800 by 500.
Software prerequisites Back end: Python 3.6.5, MySQL, and Django.
Useful tools include XAMPP 1.8.3 and Anaconda 5.2.0.

VIII. PROPOSED METHODOLOGY

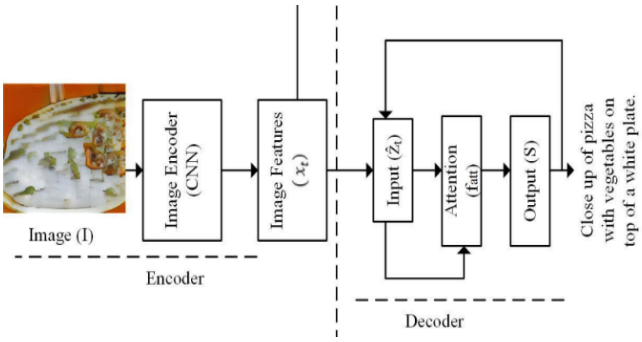


Fig.1 Architecture of our proposed method

- The input data, which typically comprises photographs and their related written descriptions, is pre-processed by the data pre-processing module. Functions for importing and resizing photos, tokenizing, and cleaning textual descriptions, and generating numerical representations of descriptions might all be included in the module.

- The image encoder module converts the input image into a feature vector so that the generator and discriminator modules can use it. A convolutional neural network (CNN) architecture, which can extract features from the image, is often used to create the encoder module.

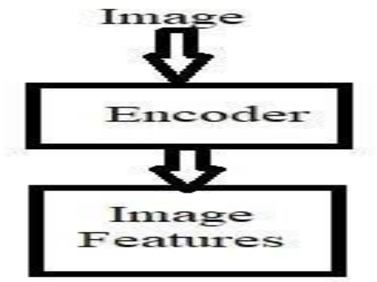


Fig. 2 Encoder Module

- Generator Module: The generator module creates a string of written descriptions that correlate to the input image using the feature vector from the image. Typically, a recurrent neural network (RNN) architecture, such as a long-short-term memory (LSTM) or gated recurrent unit (GRU), is used to create the generator module.

- Discriminator Module: The discriminator module assesses whether a set of textual descriptions produced by the generator module are a good match for the input image based on the feature vector of the input image. Usually, a CNN architecture is used to create the discriminator module, which can categorise both images and textual descriptions.

- Adversarial Training Module: Using the GAN framework, the adversarial training module oversees training the generator and discriminator modules in an adversarial fashion. The module may have tools for calculating loss functions and refining the generator and discriminator modules' settings.

- Evaluation Module: This module assesses how well the picture caption generator performs using metrics like the BLEU score, which gauges how closely the generated captions resemble the human-written captions in the dataset.

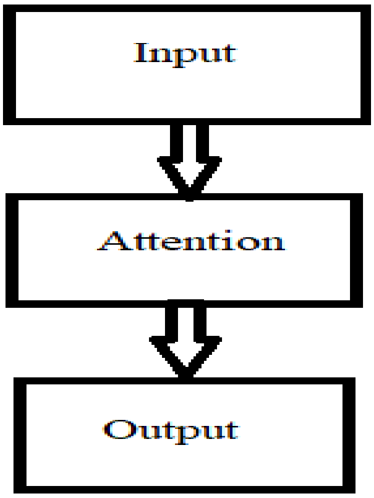
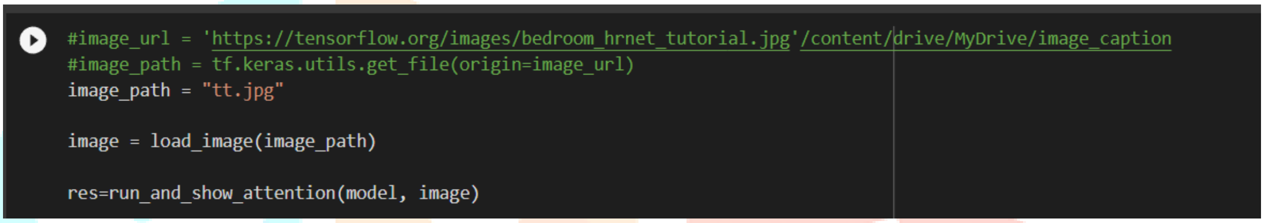


Fig. 3 Decoder Module

IX. RESULTS



```
#image_url = 'https://tensorflow.org/images/bedroom_hrnet_tutorial.jpg'/content/drive/MyDrive/image_caption
#image_path = tf.keras.utils.get_file(origin=image_url)
image_path = "tt.jpg"

image = load_image(image_path)

res=run_and_show_attention(model, image)
```

Fig. 4 Image insertion code.

Image path or URL can be given us input to insert image.

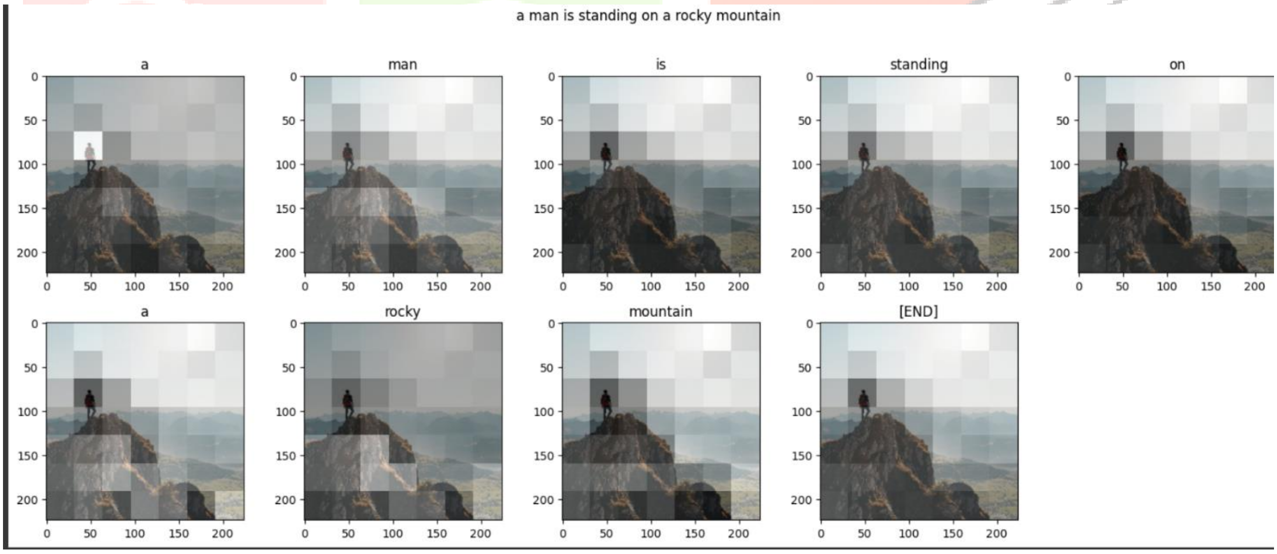


Fig.5 Result.

X. CONCLUSION

An image caption generator can be used to create captions by extracting the image's features. It examines an image pixel by pixel to ascertain its features. With the help of Mobile Net (CNN), we select the features. Based on the features that have been found, it provides captions for a picture that are more powerful. This endeavor will aid in finding solutions to the issue of proper captions.

XI. FUTURE SCOPE

Using the dataset's synthetic photos, we want to increase the project's potential in the future. GANs have been widely used to create artificial images. We can generate captions of higher quality if we can train and test an image captioning module using both these artificial and real-world photos.

XII. ACKNOWLEDGMENT

We would like to express our sincere gratitude to everyone who has assisted us in finishing this project. Due to a lack of knowledge and expertise, we encountered numerous difficulties while working on the project, but these individuals assisted us in overcoming all these obstacles and in the successful realization of our plan to create a sculpture. We would like to express our gratitude to Asst. Prof. Mrs. Chippy T for her guidance and Asst. Prof. Mrs. Namitha T N for coordinating our project, which enabled our entire team to understand the minute details of project work. We would like to express our appreciation to our HOD, Mrs. Suma S. G., for her ongoing assistance and supervision during the project work. Finally, we would like to express our gratitude to the administration of Mahaguru Institute of Technology for giving us this learning opportunity.

XIII. REFERENCES

1. https://colab.research.google.com/

2. https://ieeexplore.ieee.org/abstract/document/9416431

3. A. Karpathy and L. Fei-Fei, ‘‘Deep visual-semantic alignments for generating image descriptions,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2015, pp. 3128–3137.

4. K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, ‘‘Show, attend and tell: Neural image caption generation with visual attention,’’ in Proc. Int. Conf. Mach. Learn., 2015, pp. 2048–2057.

5. S. He, H. R. Tavakoli, A. Borji, and N. Pugeault, ‘‘Human attention in image captioning: Dataset and analysis,’’ in Proc. IEEE Int. Conf. Comput. Vis., Oct. 2019, pp. 8529–8538.

6. https://ieeexplore.ieee.org/document/8615810

7. https://ieeexplore.ieee.org/document/8947893

8. S. Hochreiter and J. Schmidhuber, ‘‘Long short-term memory,’’ Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.

9. https://ieeexplore.ieee.org/document/8728516

10. Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, ‘‘Gradient-based learning applied to document recognition,’’ Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

11. A. Krizhevsky, I. Sutskever, and G. E. Hinton, ‘‘ImageNet classification with deep convolutional neural networks,’’ in Proc. Adv. Neural Inf. Process. Syst., 2012, pp. 1097–1105.

12. T. Qiao, J. Zhang, D. Xu, and D. Tao, ‘‘MirrorGAN: Learning text-toimage generation by redescription,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2019, pp. 1505–1514.

13. B. Dai, S. Fidler, R. Urtasun, and D. Lin, ‘‘To wards diverse and natural image descriptions via a conditional GAN,’’ in Proc. IEEE Int. Conf. Comput. Vis., Oct. 2017, pp. 2970–2979.

14. C. Chen, S. Mu, W. Xiao, Z. Ye, L. Wu, and Q. Ju, ‘‘Improving image captioning with conditional generative adversarial nets,’’ in Proc. AAAI Conf. Artif. Intell., vol. 33, 2019, pp. 8142–8150.

15. W. Jiang, L. Ma, Y.-G. Jiang, W. Liu, and T. Zhang, ‘‘Recurrent fusion network for image captioning,’’ in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 499–515.