Automated Image Captioning: A Study In Deep Learning For Generating Accurate And Descriptive Captions

Upendra Mishra,[a] Sagar Srivastava,[b] Tushar Pant,[c] and Anshika[d]

KIET Group of Institutions, Delhi-NCR, Ghaziabad, India

[a]upendra.mishra@kiet.edu
[b]srivastavasagar2001@gmail.com
[c]tusharpant2001@gmail.com
[d]info2anshika@gmail.com

Abstract. It is challenging to label Earth from space with air conditions, many types of land cover, and land use. To help the international community comprehend where, how, and why deforestation occurs all across the world, we proposed an algorithm. Upcoming developments in satellite image science are anticipated to create new opportunities for a more precise examination of both significant and minute changes, including deforestation, that are taking place on Earth. Around 5% of the Amazon rainforest has been destroyed in the past 40 years. The forest is being analyzed and estimated by this application. In order to train deep convolutional neural networks (CNNs) on visual information from satellite images, a variety of classification frameworks, such gate recurrent unit label captioning, are used. In this study, we introduce a machine learning model that uses captions to explain images. This study focuses on finding various objects in an image, figuring out how they relate to one another, and creating captions for those objects. The suggested experiment will be demonstrated using the Flickr 8K dataset, Python 3, and a machine learning technique known as Transfer Learning. The potential for using image caption generators to assist persons who are blind is enormous.
Keywords: ML, AI, Image, Caption, CNN, RNN, LSTM, Neural Networks

INTRODUCTION

A. Overview

On a daily basis, we come across numerous images from diverse origins such as the internet, news articles, document diagrams, and advertisements. These images often lack descriptions, leaving the viewers to interpret them on their own. While humans can typically understand the meaning behind images without detailed captions, machines require some form of interpretation to generate automatic image captions that are understandable to humans. Image captioning holds significant importance due to its ability to aid in accessibility for individuals with visual impairments, enhance search engine optimization, assist in e-commerce by providing detailed product information, suggest ideas for content creation, and facilitate the creation of educational materials for students. Providing captions for every image on the internet could potentially result in faster and more accurate image searches and indexing. Initially, researchers focused on identifying objects within images, but it soon became apparent that simply providing the names of recognized objects was insufficient. A more comprehensive and human-like description of the image was necessary to make a lasting impression. Until machines can replicate human thought processes, communication, and behavior, generating natural language descriptions will continue to pose a challenge. Nonetheless, image captioning has numerous practical applications in diverse fields, including biomedicine, e-commerce, web search, and military operations. Social media platforms such as Instagram and Facebook can automatically generate captions from images. [13]

1) Process: The process of generating a textual description of a picture using computer vision and natural language processing methods is known as image captioning. It is a well-known area of study within Artificial Intelligence (AI) that involves visual comprehension and the creation of a corresponding linguistic description. To understand an image, object recognition and detection play a crucial role. Understanding, the type of scene or location that is being shown in the image as well as the characteristics of the things there and how they relate to one another is also necessary for image comprehension. Both syntax and semantics must be well-understood in order to construct intelligible and grammatically accurate phrases. Identifying pertinent picture elements is essential for understanding an image. Automatic picture indexing using image captioning technology is essential for content-based image retrieval (CBIR) and has practical applications in various industries such as biomedicine, business, the military, education, digital libraries, and web search. Additionally, prominent social media sites like Facebook and Twitter enable the automatic generation of image descriptions, which might contain details about the scene's setting (such as a beach or cafe), clothing, and going-on.

2) Techniques: There are two primary categories of techniques used for this purpose, namely conventional machine learning and deep machine learning approaches. Traditional machine learning techniques rely on manually created features, such as Local Binary Patterns (LBP), Scale-Invariant Feature Transforms (SIFT), Histograms of Oriented Gradients (HOG), and combinations thereof. In these methods, characteristics are taken from the input data and used to categorize an item using a classifier like Support Vector Machines (SVM). [11] The drawback of employing handcrafted features is that they are customized to particular activities, and it is impractical to extract features from a large variety of diverse data. Real-world data, including photographs and videos, are very complicated and have a wide range of semantic interpretations. Deep machine learning approaches, in contrast, can handle a wide variety of photos and videos and automatically discover features from the training data. While classifiers like soft max are used for classification, Convolutional Neural Networks (CNN) are a classic example of deep machine learning approaches that excel at feature learning. To create subtitles, CNNs are frequently combined with recurrent neural networks (RNN) or long short-term memory networks (LSTM). Deep learning [11] algorithms are skilled at navigating the difficulties and complexity of creating image captions.

B. Research Problem

With the air quality and various descriptions of the land cover or land use, it is challenging to label the satellite image. The results of the algorithms will assist the international community in determining the best course of action as well as what, how, and why deforestation is happening all across the world. Furthermore, it is usually difficult for present methods to discern between human and natural causes of forest degradation. Although it has already been shown that using higher resolution photographs in this way is very beneficial, stable methods for imaging planets have not yet been developed. In order to address this issue, our plan is to develop an encoder-decoder architecture based on the CNN and RNN algorithms. The advent of deep learning has revolutionized the field of computer vision, allowing machines to recognize and classify images with remarkable accuracy. One of the challenging tasks in this area is the automatic generation of image captions, which involves not only identifying the objects and actions depicted in an image, but also understanding the context and generating a coherent and semantically meaningful description. The creation of image captions has several useful uses, including helping those who are blind, raising search engine rankings, and improving social networking platform user experiences. The purpose of this research study is to create and assess a deep learning-based picture caption generator that can provide accurate and grammatically correct descriptions of images automatically. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs), two cutting-edge deep learning approaches, will be used by the proposed system to extract pertinent visual elements from the input image and produce a related caption. The research will explore different architectures and training strategies, as well as evaluate the system's performance on standard benchmarks and real-world datasets. [1]

C. Motivation

Creating descriptions for images is a significant task that pertains to the fields of Computer Vision and Natural Language Processing. It is a remarkable achievement in the realm of Artificial Intelligence to emulate the human capacity of generating descriptions for images using a machine. The key difficulty in generating text descriptions for images is to effectively capture the relationships between objects within the image and express them in a natural language such as English. In the past, computer systems have relied on predefined templates to generate text descriptions for images. However, this approach lacks the necessary diversity needed to produce linguistically sophisticated text descriptions. The deficiency in text description generation for images has been mitigated through the advancement of neural networks. Modern models employ neural networks to generate captions by taking an image as input and predicting the subsequent lexical units in the output sentence. Labeling satellite images with diverse captions that accurately depict land cover or land use, including air conditions, can be a challenging task. However, the output produced by advanced algorithms can provide valuable insights to the global community about the causes and effects of deforestation worldwide, as well as suggest the best course of action. Unfortunately, existing techniques often fail to differentiate between human-induced and natural causes of forest decline. Utilizing high-resolution images has proven to be a highly effective method for this task. However, reliable techniques for processing Planet imaging have yet to be developed. [5] To address this challenge, our aim is to develop an encoder-decoder architecture that leverages Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) to generate captions for these satellite photos.

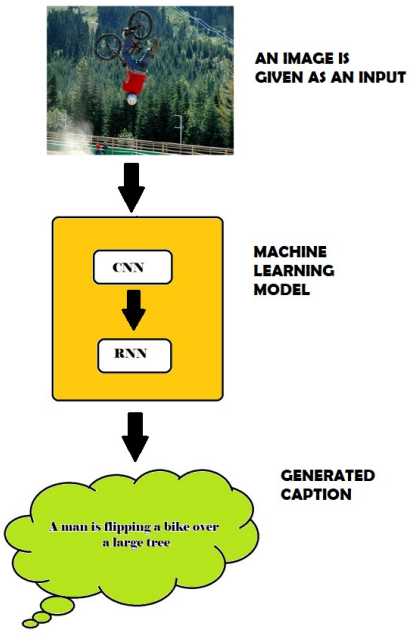

FIGURE 1. Pictorial Representation of the Model

RELATED WORK

Over the years, the problem of generating captions for images on the Internet has been addressed by various researchers using different algorithms and techniques. One such technique was proposed by Krizhevsky and team, who developed a neural network based on non-saturating neurons and an efficient GPU convolution function. To tackle overfitting, they utilized a regularization method called dropout. Their network comprised max pooling layers and a final 1000-way soft max. Another significant contribution to this field was made by Deng and colleagues, who introduced a large collection of images called ImageNet, categorized into a semantic hierarchy with numerous nodes. Karpathy and FeiFei utilized this dataset, along with its sentence descriptions, to explore the connections between visual data and language. They developed a Multimodal Recurrent Neural Network architecture that leveraged the colinear arrangement of features to generate creative image descriptions. Yang and colleagues have put forward a technique that can enhance the comprehension of images by automatically generating natural language descriptions for them. [9] This approach employs a multi-model neural network that incorporates object detection and localization modules, and its similarity to the human visual system is noteworthy. Another study by Aneja and team has proposed a convolutional network model for both machine translation and conditional picture generation. This model is fundamentally sequential across time and addresses the complexity associated with LSTM units. Extensive experimentation with network design was conducted by Pan and colleagues on large datasets containing diverse content, resulting in a novel model that outperformed previous captioning models in terms of accuracy. Meanwhile, Vinyals and team developed a generative model that utilizes computer vision and machine translation to generate natural image descriptions with a high level of accuracy. This model ensures that the generated sentences closely describe the target image. Xu and co-workers proposed an attention-based model that can automatically identify and characterize visual regions. This model was trained using traditional back propagation techniques, maximizing a lower variable.

PROPOSED WORK AND METHODOLOGY

There are numerous datasets of annotated images available for image captioning tasks, including MS COCO, Flickr 8K, and Pascal VOC. In this study, we utilized the Flickr 8K Image Captioning dataset, which comprises 8,092 images sourced from Kaggle.com. This particular dataset features a range of everyday activities, each accompanied

by a corresponding caption. The images in the dataset are labeled based on the objects they contain, and a description is generated accordingly. [6] In this study, we split the dataset of 8,000 images into three groups: a development set and a test set, each containing 1,000 images, and a training set containing 6,000 images. The potential of the image-caption pairs to join previously undiscovered images and captions with each another must be assessed in order to evaluate the image-caption pairs. [10] The BLEU (Bilingual Evaluation Understudy) Score can be used to evaluate models that produce natural language sentences. It compares a natural statement to one that was created by a human. It is frequently used to assess the effectiveness of machine translation. [8] For the purpose of calculating a BLEU score, sentences are compared using a modified n-gram precision approach, where accuracy is determined using the following equation:

$$BLEU = BP * exp(\sum_{n=1}^{N} W_n loglogP_n) \tag{1}$$

$$Count_c = min(Count, MaxRefCount) \tag{2}$$

where c = clip

Our image captioning model is built on the backbone of multimodal recurrent network and convolutional neural networks (CNNs). Initially, the image features are extracted with the help of a CNN, and later these features and captions are passed through a recurrent neural network (RNN). The model consists of 3 phases:

1) Image Feature Extraction: Because of the VGG 16 model's superior object identification performance, it is used to extract photo features from the Flickr 8K dataset. A convolutional neural network (CNN) of 16 layers, including 2 convolutional layers, 1 dropout layer, and a fully connected layer at the end, is called the VGG 16. [7] The VGG 16 model, which can learn quickly, has dropout layers to prevent the model from overfitting on the training data. The dense layer produces a 4096-dimensional vector representation of the image as its output, which is then transferred to the LSTM layer for additional processing.

2) Sequence Processor: To deal with the input text, there is a sequence processor which serves as a word embedding layer. The embedded layer consists of a mask which is used to disregard padding data and algorithms to extract out the necessary text features. A LSTM is subsequently joined to the network to complete the image captioning process.

3) Decoder: The model's ultimate stage mixes the input from the Image with other data. Utilizing an extractor phase and the sequence processor phase extra action was then transmitted to a layer of 256 neurons, followed by a dense layer that generates a soft-max prediction as the final output over the full lexicon, of the following word in the caption, which was created from the analyzed text data in the processor phase sequence. [3]

A. Training Phase

During the training phase, we provide the image captioning model with two input images and their accompanying captions. Every potential object in a photo must be recognized for the VGG model to function properly. After seeing the image and all of the words before it, the LSTM portion of the model is trained to predict each word in the phrase. To denote the start and finish of the sequence, we add two additional symbols to each caption. The sentence generator halts and marks the string's end when a stop word is encountered. [12] The loss function for the model is calculated as, where 'I' stands for the input image and 'S' for the output caption. The length of the resulting sentence is N.

Log Loss

$$L(x - \hat{x}) = -\sum_{i=1}^{n} x_i log(\hat{x}_i) \tag{3}$$

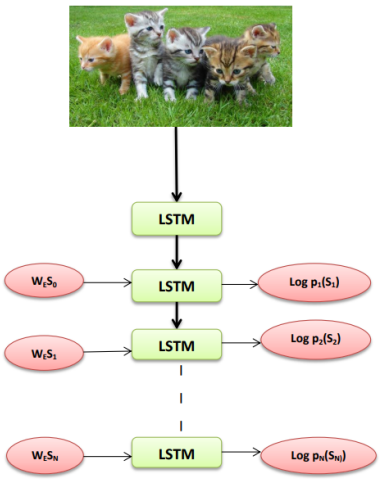

FIGURE 2. Structure of the Model

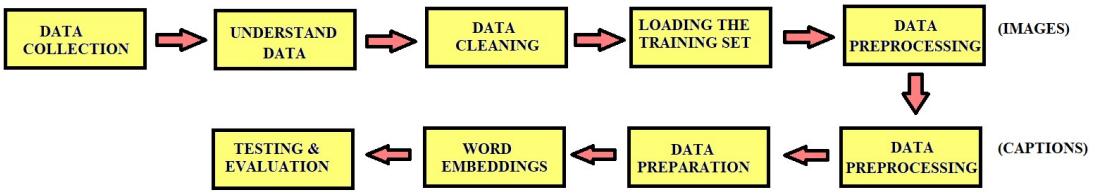

FIGURE 3. Flowchart of the Model

B. Implementation

- The Python SciPy environment was used for the model's implementation. Due to the presence of the VGG net, which was utilized for object recognition, keras 2.0 was utilized to create the deep learning model. The keras framework which is used for creating and training deep neural networks has the TensorFlow library installed as a backend in it. A deep learning library called TensorFlow was created by Google. It offers a platform that is heterogeneous, allowing algorithms to be run on both large-scale distributed systems with thousands of GPUs and low-power devices like mobile.
- Each compatible device can run a graph once it has been defined. The photo features are computed and saved using the pretrained model. These attributes are then input into our model as the interpretation of a single photo in the dataset, saving us the trouble of having to run each photo through the network each time we want to test a different language model configuration.
- For the real-time implementation of the image captioning model, the picture features are additionally preloaded. [4]

RESULTS & CONCLUSION

After implementing the picture captioning model, we were able to produce captions that were somewhat comparable to those produced by humans. To begin with, the VGG net model assigns probabilities to each object that could

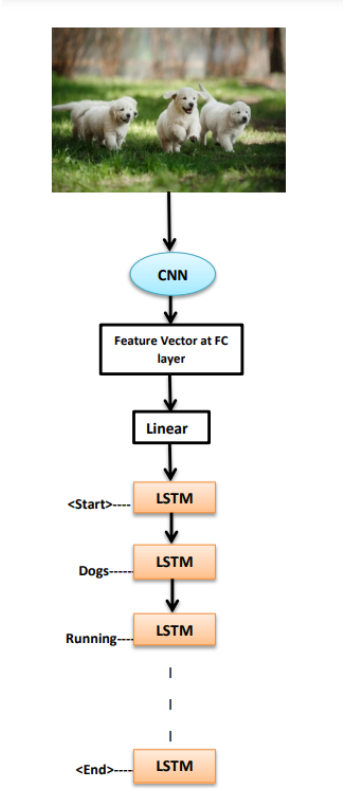

FIGURE 4. Implementation of the Model

potentially be present in the image. The image is transformed into word vector by the model. The LSTM cells receive this word vector as input and use it to construct a sentence. For example, see Fig.5, This gives this image a BLEU score of 57. The BLEU score is a metric used to compare a generated sentence with a reference sentence. A perfect match between the two sentences results in a score of 1.00, whereas an incorrect mismatch leads to a score of 0.00. The authors of a particular study are currently conducting experiments to improve the model's feature extraction in the future. To achieve this, they are exploring the use of alternating pre-trained photo models. Additionally, the scientists intend to use word vectors on a much bigger corpus of data, such as news stories and other internet sources of data, in order to increase performance. The model's configuration was fine-tuned, however different configurations can be trained to see whether the performance of the picture captioning model can be enhanced.

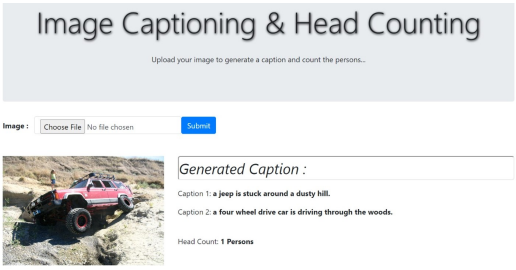

FIGURE 5. Conversion of an uploaded image into caption

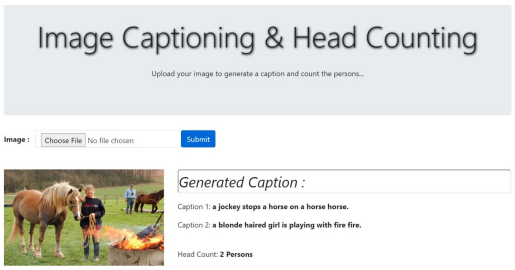

FIGURE 6. Conversion of an uploaded image into caption

FUTURE SCOPE

Image captioning has emerged as a critical problem in recent times due to the rapid proliferation of images across social media and the internet. This report delves into past research on image retrieval, highlighting the various techniques and methodologies employed in such research. Given the difficulties associated with feature extraction and similarity calculation in this domain, there exists significant scope for future research in this field. Presently, image retrieval systems rely on similarity calculations that utilize features like color, tags, histograms, and image retrieval through image captioning. However, these methodologies may not yield entirely accurate results since they do not take into account the context of the image. Therefore, extensive research in image retrieval that leverages the context of images, such as image captioning, could help to address this issue in the future. This project has the potential to be improved by incorporating more image captioning datasets to enhance the identification of classes with lower precision. Additionally, it can be beneficial to combine this methodology with previous image retrieval techniques, such as histograms and shapes, to evaluate whether it leads to improved image retrieval results. [2]

REFERENCES

1. Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 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*, pages 6077–6086, 2018.
2. Himanshi Chaudhary, Upendra Mishra, Arpita Gupta, and Amisha Singh. Comparative analysis of rainfall prediction using machine learning and deep learning techniques. In *2022 3rd International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*, pages 1–6. IEEE, 2022.
3. Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. Attend to you: Personalized image captioning with context sequence memory networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 895–903, 2017.
4. Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. Attend to you: Personalized image captioning with context sequence memory networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 895–903, 2017.
5. Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Show, control and tell: A framework for generating controllable and grounded captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8307–8316, 2019.
6. Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482, 2015.
7. Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. Semantic compositional networks for visual captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5630–5639, 2017.
8. Yan Huang, Qi Wu, Chunfeng Song, and Liang Wang. Learning semantic concepts and order for image and sentence matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6163–6171, 2018.
9. Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.
10. Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. *Advances in neural information processing systems*, 27, 2014.
11. Upendra Mishra, Deepak Gupta, and Barenya Bikash Hazarika. An intuitionistic fuzzy random vector functional link classifier. *Neural Processing Letters*, pages 1–22, 2022.
12. Vignesh Ramanathan, Armand Joulin, Percy Liang, and Li Fei-Fei. Linking people in videos with "their" names using coreference resolution. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 95–110. Springer, 2014.
13. Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.