International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 12, pp: 509-512 www.ijaem.net ISSN: 2395-5252

Image Captioning Bot

Aditya Bhardwaj[1], Tarun Grover[2], Ms. Meenu Garg[3]

[1] *B-tech scholar, Department of IT Maharaja Agrasen Institute of Technology*

[2] *B-tech scholar, Department of IT Maharaja Agrasen Institute of Technology*

[3] *Assistant Professor, Department of IT Maharaja Agrasen Institute of Technology*

Date of Submission: 25-11-2021 Date of Acceptance: 11-12-2021

ABSTRACT –

Deep Learning is a relatively new field that has received a lot of attention since it can recognize objects more precisely than ever before. Another field that has had a big impact on our lives is NLP. The fact that NLP has moved from delivering a comprehensible summary of the texts to analyzing mental diseases demonstrates its impact. NLP and Deep Learning are used to solve the challenge of image captioning. Image captioning can be used to meaningfully explain photos. Describing an image includes more than just detecting objects; in order to successfully describe a picture, we must first identify the objects in the image, then the relationship between those objects. In this study, we used a CNN-LSTM framework. In this study, we used a CNN-LSTM framework. The visual properties will be extracted using CNN, and the relevant words will be constructed using LSTM. This study also looks at how image captioning is used and the problems that it can cause.

KEY WORDS: Image captioning, Webapp, Deep Learning

I. INTRODUCTION

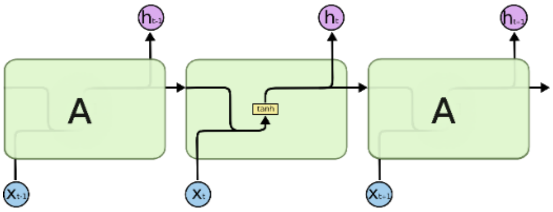
To Simply explained, picture captions are a type of automatic image description generator that allows users to create a description for a given image. The goal of the project model is to get the input image and provide a sentence description of the image's essential information. One of the most difficult and basic challenges is describing the substance of an image in a simple and understandable language. Model creation has become a possibility because to advanced technology and the availability of data sets. Humans can accurately define and explain the description of any image that comes their way thanks to sight perception. Computers, like humans, are rapidly evolving; they can recognize basic tasks such as classifying objects and recognizing their state and attributes.

However, accurately characterizing images in plain, straightforward language that humans can understand is a relatively new and difficult endeavour.

Automatic image captioning is useful for a variety of activities. The extraction of the image and its accompanying context, that is, the objects "books" and "tables," is the first step in interpreting an image. The relationship between the discovered objects has been recognized for further evaluation in the next stage; for example, the link between book and table objects is defined as "book on the table."

After the items and their relationships have been defined, the text description will be evaluated further. The words must be arranged in a way such that they make sense once formed and justify the actual relationship of the things in the image. In this research, we employed a convolutional neural network (CNN) for the first objective, which was to extract information from a picture. It's worth noting that "extract features" usually means eliminating the last softmax layer. We will use short-term memory for the second part, which is the production of written descriptions (LSTM).

LSTMs are a form of RNN that is used to avoid the long-term dependency problem that RNNs frequently encounter.



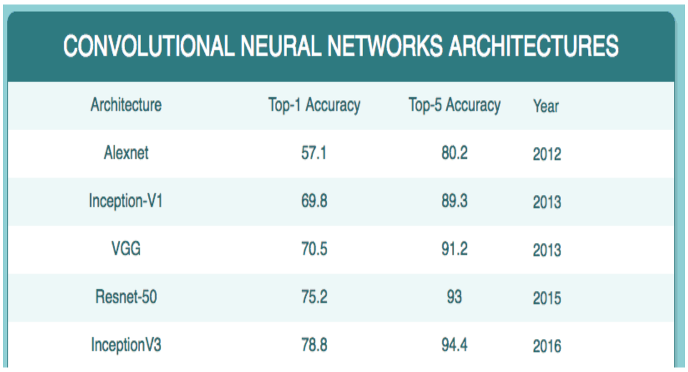
Related Work

Deep learning has gotten a lot of attention in recent years, and a lot of progress has been achieved in this discipline, as evidenced by the study [1]. This is also evident when looking at the numbers. Only four successful articles were published in 2015, but the field's popularity has skyrocketed since then, as evidenced by the 57 articles published in 2017-2018.

Elamri [3] offered a method that was purely based on the CNN-LSTM architecture. The model uses CNN to extract features from an image, which are then fed into an RNN or LSTM model. After that, the RNN or LSTM model may explain what happens in the image in a grammatically acceptable manner. The benefits of the picture caption model for the visually handicapped were also explored in the paper. Image captions may be a beneficial tool for visually impaired people in society if they are correctly produced. This study takes into account all previous studies in the field and is influenced by them. The majority of the studies we looked at used CNN and RNN-based architectures. "Adding more layers to the model does not necessarily guarantee that we will acquire improved accuracy," according to past study on the subject.

According to Di Lu and Spencer Whitehead's [2] research, a new task can be generated, and the system will receive an image description of the task as input. The current application of Image Captioning, according to the text, lacks clear motivation for the elements that make up the image's basic structure. They also provided a remedy to this problem in this paper. The CNN-LSTM model should be trained to generate titles depending on the rendered image, according to the article.

Table-1: Sample Table Format



CONVOLUTIONAL NEURAL NETWORKS ARCHITECTURES

Architecture	Top-1 Accuracy	Top-5 Accuracy	Year
Alexnet	57.1	80.2	2012
Inception-V1	69.8	89.3	2013
VGG	70.5	91.2	2013
Resnet-50	75.2	93	2015
InceptionV3	78.8	94.4	2016

Methodology and Implementation

The project's primary purpose, as stated in the abstract, is to give real-time captions for photographs.

The Flickr8k dataset was utilised to construct this project. Each image in the Flickr8k data collection contains five related titles. The data set includes 6,000 photos for training, 1,000 images for verification, and 1,000 images for testing. Five main tasks have been assigned to the project:

Data Cleaning

Create a dictionary to connect the image to the title using the image id from the data set. Image identification and captions are input into the token.txt file. We will just map each image with its own captions from the token.txt file. Our data collection has over 37,000 words in total. Now we must reduce the number of words because this will have an impact on our calculations, and if a word arrives in a shorter time, it is no longer useful. We've now put the threshold at ten, so if a word's frequency is less than ten, it's not taken into account. After filtering the terms by threshold frequency, we're left with 1,845 words that make up our vocabulary dictionary.

Image encoding

We may now use photos as input to our model, but machines, unlike humans, are unable to comprehend images when they are viewed. As a result, we must transform the photo to a code in order for the machine to recognise the pattern. I used the transfer study for this. We utilized a version that has already been evaluated and trained on a big data set. The functionality of these patterns were extracted and applied to our photos. I utilized a version of Resnet50 that had been trained on Imagenet for this study. This model from Keras may be readily imported into the programme module.

Vocabulary Segmentation/ Tokenization
We must partition all vocabulary words in this phase. To accomplish this goal, we can also use a tokenizer in Keras.

Defining the model
We'll use the keras model in the functional api to sketch out the design of our version. It consists of three basic steps:
Text content collections are processed.
Concatenating the above layers to extract feature vectors from pictures and interpreting the outcome.

II. DISCUSSION

1. Challenges Faced

- **Detecting Multiple Objects**

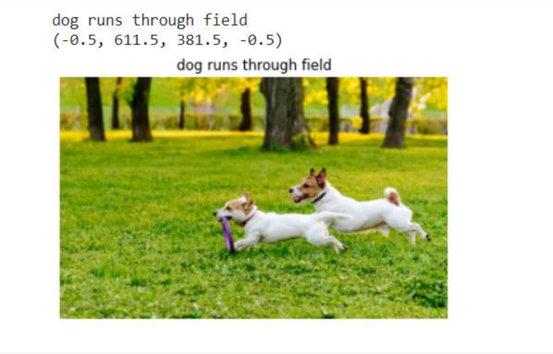
Although today's machine learning models can recognise many items, they can't always interpret the relationships between them. As a result, the model may not always be able to accurately describe the image. Furthermore, for the FlickR8K data set, we only employ an 8k image data set. You must train too many data sets in too large a format in the correct manner if you want to explain the model correctly and grammatically. When it comes to large data sets, large data sets take a long time to train, and testing is still a major issue with which we must contend.

- **Availability of Datasets**

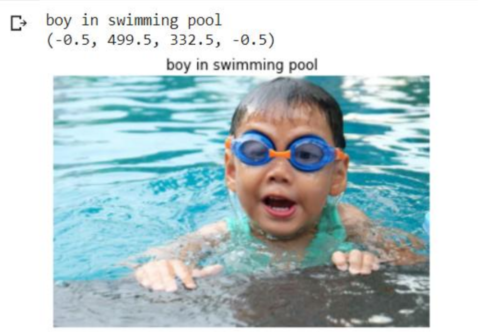
Flickr8k, Flickr30K, and MS-COCO are the most popular databases for image captioning. The majority of these data sets are now available in English. As noted in the literature review, we now have a large number of datasets that can be utilised to train our model, but the majority of the training samples are written in English or Chinese. This is a really essential subject. If we wish to employ image caption templates in real-world applications, we'll require training samples in a variety of languages.

2. Applications

Image captions may be a beneficial tool for visually impaired people in society if they are correctly produced. Developing an automatic image captioning system that can produce an accurate description of the image as a stand-alone system can be tough. The captured image can be utilised as an input for automatic image descriptions in this case. As a result, loud noise can be employed to provide output, allowing the visually blind to better comprehend their surroundings.



```
dog runs through field
(-0.5, 611.5, 381.5, -0.5)
```



```
boy in swimming pool
(-0.5, 499.5, 332.5, -0.5)
```

III. CONCLUSION

Deep learning has the potential to transform civilization, and image captions have come a long way in recent years. Captions for images can be used in a variety of industries, including agricultural and intelligent system monitoring. Surprisingly, image captions are hardly employed in domains like traffic analysis, despite the fact that traffic analysis might greatly benefit from them.

This study draws on past research and articles in the topic. The study explored for a variety of image caption models and methodologies. CNN was proven to be the best suitable model for extracting features and content, and it is also the most extensively utilised. RNN and LSTM are two popular models for producing descriptions (a special type of RNN).

REFERENCES

[1] S. ALBAWI and T. A. MOHAMMED, "Understanding of a Convolutional Neural Network,"in ICET, 2017.

[2] S. ALBAWI and T. A. MOHAMMED,

"Understanding of a Convolutional Neural Network,"in ICET, 2017.

[3] O. Vinyals, A. Toshev, S. Bengio and D. Erhan,"A Neural Image Caption Generator," CVPR 2015 Open Access Repository, vol. Xiv, 17 November 2014.

[4] D. S. Whitehead, L. Huang, H. and S.-F. Chang, "Entityaware Image Caption Generation,"in Empirical Methods in Natural Language Processing, 2018.

[5] C. Elamri and T. Planque, "Automated Neural Image Caption Generator for Visually Impaired People," California, 2016.

[6] Xu, Kelvin, et al. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention" Computer Science,2048-2057,2015.

[7] Papineni, K. "BLEU: a method for automatic evaluation of MT" 2001. Shuang Liu, Liang Bai, Yanli Hu and Haoran Wang "Image captioning based on deep neural networks".

[8] Ranzato, Marc'Aurelio, et al. "Sequence Level Training with Recurrent Neural Networks." *Computer Science* (2015)

[9] Mao, Junhua, et al. "Explain images with Multimodal Recurrent Neural Networks." Computer Science (2014)

[10] Sundermeyer, M., et al. "Comparison of feedforward and recurrent neural network language models. "IEEE International Conference on Acoustics, Speech and Signal(2013)

[11] Simonyan, Karen, and A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." *Computer Science* (2014)

[12] Szegedy, Christian, et al. "Going deeper with convolutions." IEEE Conference on Computer Vision and Pattern Recognition IEEE, 1-9. (2015).

[13] Mao, Junhua, et al. "Explain Images with Multimodal Recurrent Neural Networks." Computer Science (2014).

[14] Aneja, Jyoti, A. Deshpande, and A. Schwing. "Convolutional Image Captioning." (2017)

[15] Jeel Sukhadiya, Harsh Pandya, Vedant Singh Comparison of Image Captioning Methods