International Journal of Recent Advances in Multidisciplinary Topics
Volume 2, Issue 11, November 2021
https://www.ijramt.com | ISSN (Online): 2582-7839

Detecting and Captioning Images Using Deep Neural Networks and Flask

Mohammed Saif[1*], Vaibhav Mohurle[2], Kajal Dhumale[3], Ajay Sonawane[4]
[1,2,3]*Scholars, Department of Information Technology Engineering, SKN Sinhgad Institute of Technology and Science, Lonavala, India*
[4]*Professor, Department of Information Technology Engineering, SKN Sinhgad Institute of Technology and Science, Lonavala, India*

***Abstract*: One of the most important functions of the human visual system is to automatically caption images. There are numerous benefits to having an application that automatically captions the scenes around them and then converts the caption to a plain message. We offer a model based on CNN-LSTM neural networks that recognizes items in photos and creates descriptions for them automatically in this study. It performs the task of object detection using multiple pre-trained models, and the captions are generated using CNN and LSTM. For the job of object detection, it employs Transfer Learning-based pre-trained models. This model is capable of doing two tasks. The first is to recognize objects in the image.**

***Keywords*: RNN, CNN, LSTM, API, Flask, MSCOCO, NLP Model, Flask Rest API, Transfer Learning, VGG Model, Tensorflow, Keras.**

1. Introduction

Caption generation is one of the interesting and focused areas of Artificial Intelligence which has many challenges to pass on. Caption production entails a number of complicated steps, including selecting the dataset, training the model, validating the model, constructing pre-trained models to test the photos, identifying the images, and lastly generating captions. There are various datasets available as open source to train the model like flickr8k, flickr30k and MSCOCO. Every dataset is contained with training and validation images to train the model. The captions are generated by two separate neural networks in our model. The first neural network is the Convolutional Neural Network (CNN), which is used to train images as well as detect objects in photos using pre-trained models such as VGG, Inception, and YOLO. The second neural network employed is a Long Short Term Memory (LSTM) based Recurrent Neural Network (RNN), which is used to produce captions from the generated object keywords.

2. Literature Survey

Since the emergence of deep learning, the challenge of picture captioning has been a challenging and intensively researched academic area. There are numerous proposed solutions to this problem, and new ones are being added every day to replace the old ones. In [1] Karpathy proposed a method that employs multimodal neural networks to produce innovative visual descriptions by giving appropriate descriptions. In [2], Aneja proposed a method that employs multimodal neural networks to produce innovative visual descriptions by giving appropriate descriptions. In [3], Yan proposed a multi-neural network model for generating correct sentence descriptions from images, which has been tested on a wide number of datasets. In [4], Yang proposed a multimodal recurrent neural network-based model that generates image descriptions by identifying objects and converting them to sentences, in a manner that is nearly identical to the human visual system.

3. Dataset

In performing the task of image captioning, we have used flickr8k dataset from flickr.com website. It consists of a total of 8092 images. Those 8092 images were divided into 6000 training images and 1000 images each for development and testing. It is made up of photographs from everyday life with features that cover a wide range of objects. . The images were clear and well-resolved, and the model could easily recognize them throughout training. It's an open source dataset that you may download for free from the internet.

Fig. 1. Flickr 8k dataset

*Corresponding author: mohammedsaif556@gmail.com

4. Phases of Model

1) Creating a pre-trained model using Transfer Learning:

In order to solve difficult machine learning challenges, transfer learning is utilized to develop and utilize pre-trained models. It is a method of storing knowledge gained through the solution of a problem so that it can be applied to a more difficult problem later. In this model, we use a pre-trained model combined with Transfer learning to teach our model new things based on prior knowledge.

2) Object Detection

During this phase, objects in the images were detected. It extracts the features from the image using a Convolutional Neural Network (CNN). For the task of object detection, we will utilize a pre-trained model such as Inception V4 or VGG 16, which is a Convolutional Neural Network. During this phase, each and every object in the image is detected. The names of the objects were also written on them.

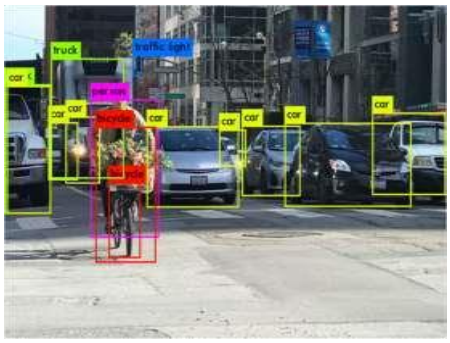

Fig. 2. Object Detection

3) Probabilistic NLP Model

The recognized objects in the image were submitted via this probabilistic NLP (Natural Language Processing) model, which removed the image's superfluous attributes. It only processes the aspects that are relevant and useful in the context of the image, ignoring the unusual and irrelevant ones. It also gets rid of stop words that are repeated and mean the same thing.

4) Caption Generation:

To create captions for the image, this step combines the object detection and probabilistic model phases. To create the captions, the output of the previous phases is fed into a Long Short Term Memory (LSTM), which is a form of Recurrent Neural Network (RNN). Long-term dependencies are stored in LSTMs. It enables RNN to learn over numerous steps by preserving error that can be back propagated through layers and time while maintaining a consistent error. During this stage, captions were created.

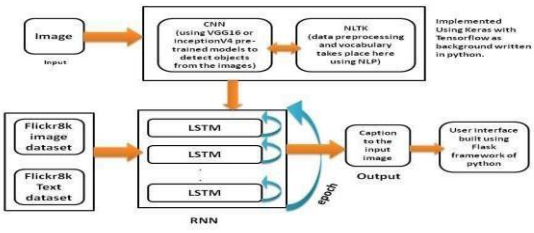

Fig. 3. Architecture of the Model

5) Ranking Based Caption Retrieval

Multiple captions are generated from the image using different layers in LSTMs. The captions generated from the top layers of the LSTMs were prioritized in this phase depending on which captions were supported by the most LSTM layers. The caption with the highest score will be used as the final caption. Different LSTM layers create their own captions. The caption with the greatest number of LSTMs will be selected as the final caption.

6) Deployment to Web Server

The caption generating model will be deployed as a web application in the project's final phase. To deploy the functioning model, we're utilizing Flask Rest API, a Python web framework. Flask is a well-known Python Online development framework for creating and deploying models into web applications. The online application's UI is also designed with html, CSS, and Bootstrap.

7) Training Phase

We give a pair of images in the training phase that can detect all of the possible objects in the dataset as well as the captions for these images in the image. After observing the image, the LSTM component is used to predict each and every word from it. We add starting and ending indicators to each caption so that the sentence can be recognized. When a stop word appears in a sentence, it terminates the construction of the phrase and signals the end of the string. The Loss function must be calculated using the formula below. The input image is represented by 'I', while the generated caption is represented by 'S' in the formula. We must minimize the loss function during the training procedure.

8) Future work

We'll take our effort to the next level by improving our model so that it can create captions even for live video frames. Our current approach only creates captions for images, which is a difficult work in and of itself, and captioning live video frames is even more difficult. This is entirely GPU-based, as captioning live video frames is impossible with standard CPUs. Video captioning is a popular study subject that has the potential to improve people's lives, with application cases that can be found in practically every domain. It automates the most important security duties, such as video surveillance.

5. Conclusion

In essentially every complex area of Artificial Intelligence, image captioning provides numerous advantages. Our model's major application is to assist visually impaired people in comprehending their surroundings and making it simple for them to act in accordance with their surroundings. Because this is a difficult assignment, we were able to complete it with the help of pre-trained models and sophisticated deep learning frameworks such as Tensorflow and Keras. This is a Deep Learning project that uses various Neural Networks to detect objects and caption images, including Convolutional Neural Networks and Long Short Term Memory. We deployed Flask, which is a strong Python web framework, to deploy our model as a web application.

References

[1] Andrej Karpathy, Li Fei-Fei, Deep Visual Semantic Alignments for Generating Image Descriptions, [Online] Available: https://cs.stanford.edu/people/karpathy/cvp r2015.pdf.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li and Li Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database.

[3] Jyoti Aneja, Aditya Deshpande, Alexander Schwing, Convolutional Image Captioning.

[4] Jia-Yu Pan, Hyung-Jeong Yang, Pinar Duygulu, Automatic Image Captioning, Conference: Conference: Multimedia and Expo, 2004. ICME '04. 2004 IEEE International Conference, Volume: 3.

[5] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan, Show and Tell: A Neural Image Caption Generator.