Image Captioning using Deep Learning

Ankit Kumar[1], Abhishek Ranjan[2], Neetu Garg[3]

[1,2]*Maharaja Agrasen Institute of Technology, New Delhi*
[3]*Assistant Professor, CSE Department, Maharaja Agrasen Institute of Technology, New Delhi*

Abstract - *The topic of autonomously producing descriptive words for photographs has piqued interest in natural language processing and computer vision research in recent years. The process of creating a written description of an image is known as image captioning. The captions are generated using both* ***Natural Language Processing*** *and Computer Vision. The authors propose a hybrid system that uses a multi-layer* ***Convolutional Neural Network (CNN)*** *to produce image-descriptive vocabulary and a* ***Long Short Term Memory (LSTM)*** *to accurately form meaningful sentences utilizing the generated keywords in this study. A Convolutional Neural Network (ConvNet/CNN) is a* ***Deep Learning*** *algorithm that uses convolutional neural networks. There are many open source datasets available for this problem, like Flickr8k (containing 8k images), Flickr30k (containing 30k images), MS COCO (containing 180k images), etc.*

1. INTRODUCTION

Caption creation is a fascinating artificial intelligence challenge that involves generating a descriptive text for a given image. It uses two computer vision approaches to comprehend the image's content, as well as a language model from the field of natural language processing to convert the image's comprehension into words in the correct order. Image captioning has a variety of uses, including recommendations in editing software, use in virtual assistants, image indexing, accessibility for visually impaired people, social media, and a variety of other natural language processing applications. Most traditional image captioning systems use an encoder-decoder structure, in which an input image is encoded into an intermediate representation of the information contained within the image, and then decoded into a descriptive text sequence, which is inspired by neural machine translation. This encoding can be a single CNN feature vector output, or numerous visual features extracted from various locations inside the image. The regions can either be uniformly sampled or directed by an object detector, which has been found to increase performance.

1.1 Data Set Collection

For this topic, there are various open source datasets accessible, such as Flickr8k (which contains 8k photos), Flickr30k (which contains 30k images), MS COCO (which has 180k images), and so on.

But for this case study, I've utilized the Flickr 8k dataset, which you can get by filling out this form from the University of Illinois at Urbana-Champaign. Additionally, training a model with a huge number of photos may be impossible on a device that is not a high-end PC or laptop .There are 8000 photos in this collection, each with five captions (as we have already seen in the Introduction section that an image can have multiple captions, all being relevant simultaneously).

1.2 Understanding the Data

We downloaded the data from kaggle, along with the photos, and some text files connected to the photographs. "Flickr8k.token.txt" is one of the files, and it contains the name of each image as well as its five captions.

What do you think you're seeing in the image below?

Some of you could say "A white bird flying ," while others would say "A white bird with black patches" or "A crane is flying over a river." Certainly all of these captions, and maybe more, are likely to be accept able for this photograph. But the point I'm trying to make is t hat looking at an image and describing it in appropriate term s is so easy for us as humans.

This is something that even a 5-year-old could achieve with ease.

Can you, on the other hand, develop a computer program that takes a picture as input and outputs an appropriate caption?

2. Work Related to Image Captioning

Since the invention of the Internet and its broad acceptance as a platform for sharing photographs, the image captioning problem and its suggested solutions have existed. Researchers have proposed a variety of algorithms and strategies from various angles. Krizhevsky et al. created a neural network with non-saturating neurons and a GPU version of the convolution function that is exceedingly efficient and unique. They were able to reduce overfitting by using a regularization process termed dropout. Their neural network has maxpooling layers and a 1000-way softmax at the end. There has also been a family of attention-based techniques to picture captioning suggested [26, 30, 28],

which aim to anchor the words in the anticipated caption to areas in the image. The spatial localisation is restricted and frequently not semantically relevant because visual attention is often obtained from higher convolutional layers of a CNN. Anderson et al. in combined a "bottom-up" attention model with a "top-down" LSTM to solve this issue of traditional attention models, which is most comparable to our work. on was initially developed by Hu et al. for object identification

This systematic literature review is planned, conducted, and reported step by step. First, we highlighted the necessity of performing this study in the planning part. Identifying research topics and developing a search strategy, as well as developing quality evaluation standards and data extraction strategies, are all planned during this stage. We performed our investigation after careful planning.

3. MOTIVATION

We must first comprehend the significance of this issue in real-world circumstances. Let's look at a couple scenarios when a solution to this problem may be quite valuable.

Self-driving automobiles — Automatic driving is one of the most difficult difficulties, and captioning the area around the car can help the self-driving system.

Aid for the blind — We can develop a product for the blind that will lead them on the roads without the need for anyone else's assistance. This may be accomplished by first turning the scene to text, then the text to speech. Both are now well-known Deep Learning applications.

CCTV cameras are already ubiquitous, but if we can provide appropriate captions in addition to watching the world, we can trigger warnings as soon as criminal behaviour is detected someplace. This is likely to help minimise crime and/or accidents.

Automatic captioning might help Google Image Search become as good as Google Search, because every image could be transformed into a caption first, and then searches could be conducted based on the caption.

4. DISCUSSION

In the realm of Deep Learning, it has been a critical and basic job. Captioning images has a wide range of uses. Image captioning may be thought of as an end-to-end Sequence to Sequence challenge since it turns images from pixels to words. If we can do automatic picture annotations, this can be beneficial in both practical and theoretical ways. The huge data that exists on the Internet is the most essential item in the present societal development process. The majority of these data are distinct from traditional data, with media data accounting for a significant share. They're frequently created by Internet services like social networks and news media.

Visual aid to blind- if we can fix the camera on the head of a blind person then that camera will take snaps and from it we can generate relevant captions and that captions can be converted in the form of text and later in audio which can tell the blind person what he is viewing. Self driving cars – Autonomous decision-making systems are what self-driving automobiles are. They can handle data streams from a variety of sensors, including cameras, RADAR, GPS. This information is then modeled using deep learning algorithms, which make judgments based on the car's current surroundings which include generating relevant captions and detecting the right track for the vehicles.

Alarm system- Today, cameras are everywhere, but if we can provide useful captions in addition to watching the world, we can raise alerts as soon as dangerous conduct is detected. This is likely to help minimize crime and/or accidents.

5. Architecture

Image data is mapped to an output variable using Convolutio n Neural Networks. They've proven to be so successful that t hey're now the method of choice for any form of prediction is sue utilising picture data as an input.

Recurrent neural networks, or RNNs, were developed to han dle sequence prediction problems. Sequence prediction chall enges include one-to-many, many-to-one, and many-many.LSTM networks are the most successful RNNs because they can contain a bigger sequence of words or phrases for p rediction. The CNNLSTM Model combines CNN and LSTM int o one model. One of the most exciting and useful neural mod els is created by combining multiple types of networks into h ybrid models

EXAMPLE-
Consider the challenge of creating image captions.

We have an input image and an output sequence, which is th e caption for the input image in this example

Is it possible to model this as a problem of one-to-many sequence prediction?

Yes, but how would an LSTM or other sequence prediction m odel comprehend the input image?

Because they aren't designed to interact with such inputs, we can't directly input the RGB image tensor.

Images and other inputs with spatial organisation are difficul t to represent with the Vanilla LSTM.

Are there any characteristics that we can extract from the su pplied image?
Yes in order to exploit the LSTM architecture for our purpose s, we must do just that. We can extract features from the pict ure using the deep CNN architecture, which are then input in to the LSTM architecture to generate the caption. The CNN LSTM model was created primarily for sequence prediction problems involving spatial inputs, such as pictures or videos. Convolution Neural Network (CNN) layers for feature extraction on input data are paired with LSTMs for sequence prediction on the feature vectors in this design. In a nutshell,

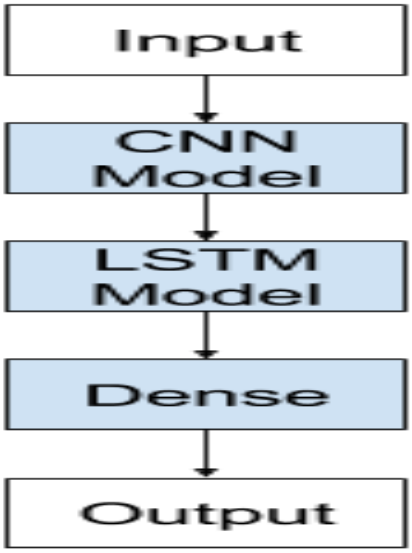
CNN LSTMs are a type of model that is both spatially and temporally deep and is found at the intersection of computer vision and natural language processing. These models have a lot of promise and are increasingly being employed for more complex tasks like text categorization and video conversion. A CNN LSTM Model's as shown below.

Language Model

For picture captioning, we're building an LSTM-based model that uses feature vectors from Resnet to predict caption sequences for total of 100 epochs, the language model is trained.

Other factors can be fiddled with and fine-tuned to your heart's content.

We've included an example of a result we got after training o ur network.



6. Goal

Image captioning's purpose is to turn a given input image int o a natural language description.

In this research, we will use the concepts of CNN and LSTM t o create a Picture Caption Generator model that combines co mputer vision and natural language processing to identify im age context and describe it in natural language such as Englis h

The task of captioning can be divided into two modules

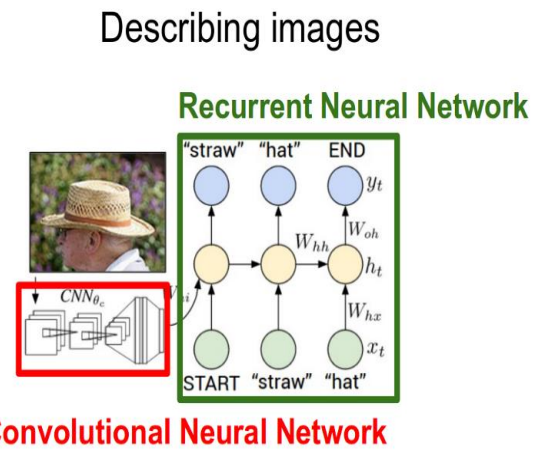
1. Image based model — Extracts the features of image.

2. Language based model which translates the features and objects extracted by our image based model to a natural sentence

We employ CNN for our image-based model and LSTM for our language-based model. The method of Graphic Captioning Generator is summarized in the image below.

We usually use a Convolutional Neural Network model for our image-based models.

Use LSTM for language-based models. The strategy is depicted in the diagram below.



The characteristics from our input image are extracted by a pre-trained CNN. The feature vector is linearly converted such that it has the same dimension as the LSTM network's input dimension. On our feature vector, this network is trained as a language model.

We predefine our label and target text before training our LSTM model. For example, if the description reads "An old guy is wearing a hat," our label and target would be "An old man is wearing a hat."

Label — [<start>,An, old, man, is, wearing, a , hat .]

Target — [An old man is wearing a hat ., <end>]

This is done so that our model understands the start and end of our labeled sequence. The image dataset is divided into 6000 images for training, 1000 images for validation and 1000 images for testing. Here, we will break down the module into following sections for better understanding .

- Preprocessing of Image
- Creating the vocabulary for the image
- Train the set
- Evaluating the model
- Testing on individual images

PREPROCESSING THE IMAGE: To detect images, we use a pre-trained model called Visual Geometry Group (Resnet50). The Keras library already has Resnet50 installed .The image features for feature extraction have a size of 224*224. The image features are extracted just before the final layer of classification because this is the model used to predict a classification for a photo. We aren't interested in image classification, so we skipped the final layer.

CREATING VOCABULARY FOR THE IMAGE:

We cannot fit the raw text into a Machine Learning or Deep Learning model right away. First, we must clean up the text by breaking it down into words and dealing with punctuation and case sensitivity issues. Because computers cannot understand English words, we must represent them numerically and assign a unique index value to each word in the vocabulary, as well as encode each word into a fixed sized vector and represent each word as a number. Only then will the machine be able to read the text and generate image captions.

To achieve the desired vocabulary size, we will clean the text in the following order:

To achieve the aforementioned goals, we define the following five functions:

- The data is being loaded.
- Creating an image-to-description dictionary
- removing punctuation, converting all text to lowercase, and removing numbers from words
- Taking all of the descriptions and separating out the unique words and creating vocabulary.
- Making a descriptions.txt file to save all of the captions.

Training The Model - In our dataset, there is a file called Flickr 8k.trainImages.txt that contains a list of 6000 image names that will be used for training.

We begin by loading the features extracted from the previously described CNN model: This will generate a dictionary with captions for each photo in the list of photos.

Data Generator - We must provide input and output to the model for training in order to make this a supervised learning task. We train our model on 6000 images, and each image contains a 4096-length feature vector as well as the image's caption, which is also represented as a number. Because the large volume of data generated for 6000 images cannot be held in memory, we will use a generator method that will produce batches. The input and output sequences will be generated by the generator

CNN-LSTM MODEL:

For image captioning, we are developing an LSTM-based model that predicts word sequences, known as captions, from feature vectors obtained from the Resnet network.

We will use the 6000 training images to train the model by generating the input and output sequences in batches from the above data generation module and fitting them to the model. We are training the model over a period of ten epochs.

TESTING THE MODEL:

Now that the model has been trained, we can put it to the test against a set of random images. Because the predictions include the maximum length of index values, we will use the same tokenizer. pkl to retrieve the words based on their index values. The generated captions for the images are quite accurate. Certain images are poorly recognised, and there is still room for improvement. Because the model is dependent on data, it cannot predict words that are not in its vocabulary. For better results, we can use a data set with 100,000 images to produce more accurate models.

7. CONCLUSIONS

Image captioning has made large advances in recent years. Recent paintings primarily based totally on deep getting to know strategies has ended in a step forward with inside the accuracy of image captioning .

ACKNOWLEDGEMENT

We would like to express our gratitude to all those who provided us an opportunity on this research. A special thanks to our guide and mentor Mrs Neetu Garg, assistant professor cse department Mait New Delhi whose contribution in mentoring, stimulating suggestions and real-life ideas and for the encouragement of research. Not only she provided the technical knowledge but also life scenarios ideas and how to handle them.

I perceive this opportunity as a big milestone in my career development. I will strive to use gained skills and knowledge in the best possible way, and I will continue to work on their improvement, in order to attain desired career objectives.

8. REFERENCES

[1] Image Captioning: Transforming Objects into Words Lakshminarasimhan Srinivasan1, Dinesh Sreekanthan2, Amutha A.L3

[2] Image Captioning Based on Deep Neural Networks Shuang Liu1, Liang Bai1, a, Yanli Hu1 and Haoran Wang1

[3] Image Captioning with Keras , harsh lamba

https://towardsdatascience.com/image-captioning-with-keras-teaching-computers-to-describe-pictures-c88a46a311b8

[4] IMAGE CAPTION GENERATOR CNN-LSTM Architecture and Image Captioning Arsh Chowdhry

https://blog.clairvoyantsoft.com/image-caption-generator-535b8e9a66ac

[5] Image Captioning Based on Deep Neural Networks Shuang Liu1, Liang Bai1,a, Yanli Hu1 and Haoran Wang1