International Journal for Modern Trends in Science and Technology, 6(12): 40-44, 2020
Copyright © 2020 International Journal for Modern Trends in Science and Technology
ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST061208
Available online at: http://www.ijmtst.com/vol6issue12.html

Image Captioning System

Gaurav Joshi[1] | Dr. Amita Goel[2]| Vasudha Bahl[3]| Nidhi Sengar[4]

[1]B-tech scholar, Department of IT Maharaja Agrasen Institute of Technology, Delhi, India.
[2] Professor, Department of IT Maharaja Agrasen Institute of Technology, Delhi, India.
[3] Assistant Professor, Department of IT Maharaja Agrasen Institute of Technology, Delhi, India.
[4] Assistant Professor, Department of IT Maharaja Agrasen Institute of Technology, Delhi, India.

To Cite this Article
Gaurav Joshi, Dr. Amita Goel, Vasudha Bahl and Nidhi Sengar, "Image Captioning System", *International Journal for Modern Trends in Science and Technology*, 6(12): 40-44, 2020.

Article Info
Received on 06-November-2020, Revised on 18-November-2020, Accepted on 25-November-2020, Published on 30-November-2020.

ABSTRACT

Deep Learning is relatively a new field and it has grabbed a lot of attention because it provides higher level of accuracy in recognizing objects than ever earlier. NLP is also one field that has created a huge impact in our life. NLP has come a long way from producing a readable summary of the texts to analysis of mental illness, it shows the impact of NLP. Image captioning task combines both NLP and Deep Learning. Describing images in a meaningful way can be done using Image captioning. Describing an image don't just mean recognizing objects, to describe an image properly we first need to identify objects present in the image and then the relationship between those objects. In this study we have used CNN-LSTM based framework. CNN will be used to extract features of the image while with the help of LSTM we will try to generate meaningful sentences. This study also discusses applications of Image captioning and major challenges faced in achieving this task.

KEYWORDS: *Deep Learning, Convolutional Neural Networks, Image Captioning, LSTM*

I. INTRODUCTION

Image Captioning, to simply put, is an automatic image description generator that helps users to auto-generate the description of the image presented. This project model aims to take an input image and generate a sentence description of the basic content of the image. Describing the content of an image in simple and easy to understand language is one of the complex and fundamental tasks. With the help of advanced technology and the availability of datasets, building models has now become a possible task.

Humans, with the help of their sight vision, can define and accurately tell the description of any image presented to them. Just like humans, computers havebeen growing at a rapid rateand can recognize the basic actions classified by an object, recognize its state and features. Although, defining an image with accuracy in simple and plain language which is easily comprehensible by humans has become a relatively new and thought-provoking task.

Auto image captioning performs its function in a sequence of tasks. The first step towards understanding an image begins with the extraction of the image with its relative surrounding i.e. if the objects are "book" and "table". In the next stage, the relationship between the detected objectshas

been identified for further evaluation i.e. for objects book and table, the relationship between two to be defined as “the book is on the table”.

Once the objects and their relationships with each other have been defined, further valuation takes place in the text description. Sequences of words have to be put in a way so that when once formed it will make sense and justify the actual relationship of objects placed in the image.

For the first task i.e. for extracting the features out of the image we have used Convolutional Neural Network(CNN) in this project. It is very important to note that ‘extracting feature’ refers to removing the last softmax layer in most cases. For the second part, which is to generate a textual description we are going to use Long Short Term Memory(LSTM). LSTMs are a special type of RNN which are used to avoid the long term dependency problems which often occurs in case of RNNs.

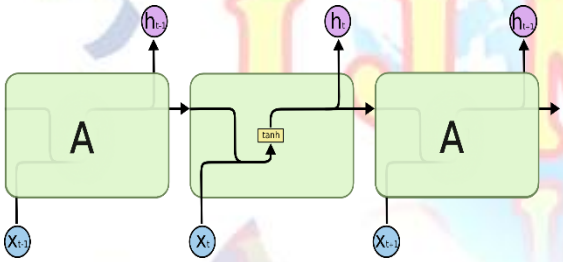

Fig1. Overview of the LSTM

II. Related Work

In Work[1] it was clearly shown that the deep learning has started getting a lot of attention in last few years and a lot of advancements have also been made in this field. This is quite evident as well when we look at the stats. In 2015 only four successful articles were published but after that the popularity of the field grew exponentially and it can be seen from the fact that 57 articles were published in 2017–2018.

Work[2] done by Di Lu and Spencer Whitehead suggested that a new task can be created for which image descriptions will be given as input to the system. The paper also mentioned that the Image Captioning which has been in use currently lacks the specific motivation of entities that forms the basic structure of image. In this paper, they also proposed the solution for this issue.

The paper suggested that CNN-LSTM model should be trained so that it will be capable of generating caption based on images represented to it.

Elamri [3] also proposed a solution based on CNN-LSTM based architecture only. The model uses the CNN to extract the features of a given image, which later is fedinto the RNN or LSTM model. Later the RNN or LSTM model describes the image in grammatically correct form that can describe what is going in the image.The paper also discussed the advantage of Image captioning model to visually impaired person. To help visually impaired people in society, image captioning can come out to be a helpful device if developed accurately.

This project takes into account all the past research that has been done in this field already and is also influenced from those research. Most of the works that we have studied uses CNN and RNN based architecture. An interesting finding that we have got from the past research done on this topic is that “adding more layers to the model doesn’t necessarily means that we will get more accuracy”.

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

Fig2. Table showing the accuracy of different CNN architectures

II. Methodology and Implementation

As already discussed in the abstract, the basic aim of this project is to provide captions for the image in the real time. The dataset that is used for building this project is Flickr8k dataset. In Flickr8k dataset every image has 5 captions corresponding to it. The dataset provides 6000 images for training purpose,1000 images for validation purpose and the rest 1000 images for

the testing purpose. The project has been divided into five tasks mainly:

Data Cleaning

1. Fetching the Image id from the Dataset and creating a dictionary that will map images with the captions. The token.txt file has image id and captions as entries and from this token.txt file only we are going tomap every images with their respective captions.The total words that we have in our dataset is close to around 37000. Now we have to reduce this words as this will effect our computation and also ifa word is coming for very less time then it doesn't make sense to use it. Now

we have put a threshold value of 10 so if a word has frequency less than 10 then we're not taking that word into consideration. After filteringthe words on the basis of threshold frequency we have only 1845 words and this constitutes our vocabulary dictionary.

Encoding the Images

2. Now we are able to give an photo as an input to our model however in contrast tohuman beings, machines can't understand the picture by seeing them. So we need to transform the photograph into an encoding in order that the machine can recognize the patterns in it. For this, I have used the transfer studying i.e. we use a pre-skilled version that has been already educated on large datasets and extract the functions from these patterns and use them for our photos. For this research, I have used Resnet50 version which has been already trained on Imagenet. We can easily import this model from keras.Programs module.

```
In [27]: def encode_image(img):
             img = preprocess_img(img)
             feature_vector = model_new.predict(img)

             feature_vector = feature_vector.reshape((-1,))

             return feature_vector
```

```
In [28]: encode_image(IMG_PATH+"1000268201_693b08cb0e.jpg")
         encoding_train = {}
```

```
In [117]: start = time()


          for ix,img_id in enumerate(train):
              img_path = IMG_PATH+"/"+img_id+".jpg"
              encoding_train[img_id] = encode_image(img_path)

              if ix%50==0:
                  print("Encoding in Progress Time step %d "%ix)

          end_t = time()
          print("Total Time Taken :",end_t-start)
```

Tokenizing the vocabulary

In this step, we need to tokenize all the words present in our vocabulary. Alternatively, we can use tokenizer in Keras to do this task.

Defining the Model

For outlining the shape of our version, we will be the usage of the keras model from functional api. It has three primary steps:

- processing the collection from the textual content
- extracting the characteristic vector from the photograph
- interpreting the output with the aid of concatenating the above layers.

```
In [44]: input_img_features = Input(shape=(2048,))
         inp_img1 = Dropout(0.3)(input_img_features)
         inp_img2 = Dense(256,activation='relu')(inp_img1)
```

```
# Captions as Input
input_captions = Input(shape=(max_len,))
inp_cap1 = Embedding(input_dim=vocab_size,output_dim=50,mask_zero=True)(input_captions)
inp_cap2 = Dropout(0.3)(inp_cap1)
inp_cap3 = LSTM(256)(inp_cap2)
```

```
decoder1 = add([inp_img2,inp_cap3])
decoder2 = Dense(256,activation='relu')(decoder1)
outputs = Dense(vocab_size,activation='softmax')(decoder2)

# Combined Model
model = Model(inputs=[input_img_features,input_captions],outputs=outputs)
```

IV.DISCUSSION

4.1 Challenges faced

4.1.1 Detecting Multiple Objects

The models that we have nowadays have the ability to detect multiple objects but models can't always interpret the relationships present between those objects. Thus the model can't always give accurate descriptions of the image. Also, the dataset which we have used i.e. Flickr8k dataset has only 8k images. Now if we want our model to accurately describes the image and that too in grammatically correct form then we need to train our model on much larger datasets. Talking about large datasets, large datasets also take huge time to train thus speed of training, testing also remains a very big problem that is needed to be addressed.

4.1.2 Availability of Datasets

The most common datasets that are generally used for Image Captioning are Flickr8k, Flickr30K and MS-COCO. Now these datasets are mostly in English. As mentioned in work[] as of now we have a lot of datasets which we can use to train our model but most of the training samples are either in English or Chinese. This is a very important Issue to address if we want to use the image captioning model for practical applications then availability of cross language training samples are very much required.

4.2 Applications

To help visually impaired people in society, image captioning can come out to be a helpful device if developed accurately The development of an automatic image captioning system that provides accurate image descriptions as an independent system can be a tough task. Here, Images that have been taken can be used as an input for auto image captioning. As a result, the output can be provided with the help of loud noise, which can help visually impaired people can better understand their surroundings.

V. RESULTS

Actual caption: the two white dogs are playing in field
Predicted caption: two dogs are playing in the grass

Actual caption: couple is photographed in front of large outdoor fountain
Predicted caption: two people stand outside and pose for picture

Actual caption: man rides his blue bike high in the air over park
Predicted caption: man on bike doing trick on his bike

VI. CONCLUSION

Deep learning has the ability to provide remarkable changes in the society and in recent years image captioning has made major advances. Image captioning can provide a lot of applications in various domains like agriculture, smart monitoring of the systems. It is quite shocking to see that image captioning isn't used in domains like traffic analysis which could be benefitted a lot by it. This research relies on various articles and past researchesdone in the field. The research looked for various specific models and strategies used for image captioning and wefound that the forextracting features and content CNN is the best suited model and is widely used as well. For generating description, the models which are frequently used are RNN and LSTM(special type of RNN).

REFERENCES

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

[2] S. Hochreiter, "LONG SHORT-TERM MEMORY," Neural Computation, December 1997.

[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 CaptionGeneration,"inEmpiricalMethods 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 Processing IEEE*, 8430-8434. (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] JeelSukhadiya, Harsh Pandya, Vedant Singh Comparison of Image Captioning Methods