International Journal for Modern Trends in Science and Technology, 6(12): 181-185, 2020

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST061235
Available online at: http://www.ijmtst.com/vol6issue12.html

Image to Caption Generator

Jitesh Gupta[1] | Mohd Zeeshan[1]| Karan Garg[1]| Ms. Kavita Saxena[2]

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

To Cite this Article
Jitesh Gupta, Mohd Zeeshan, Karan Garg and Kavita Saxena, "Image to Caption Generator", *International Journal for Modern Trends in Science and Technology*, 6(12): 181-185, 2020.

Article Info
Received on 08-November-2020, Revised on 28-November-2020, Accepted on 02-December-2020, Published on 05-December-2020.

ABSTRACT

Deep Learning is generally another field and it has caught a ton of eye since it gives more elevated level of precision in perceiving objects than at any other time prior. NLP is additionally one field that has made an immense effect in our life. NLP has made considerable progress from creating a lucid synopsis of the writings to investigation of psychological sickness, it shows the effect of NLP. Image Captioning task consolidates both NLP and Deep Learning. Depicting pictures in an important manner should be possible utilizing Image subtitling. Depicting a picture don't simply mean perceiving objects, to portray a picture appropriately we first need to recognize objects present in the picture and afterward the connection between those articles. In this investigation we have utilized CNN-LSTM based system. CNN will be utilized to remove highlights of the picture while with the assistance of LSTM we will attempt to produce important sentences. This assessment furthermore discusses uses of Image Captioning and huge challenges glanced in achieving this endeavour.

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

I. INTRODUCTION

Image Captioning, basically, is a programmed picture caption generator that encourages clients to auto-create the depiction of the given image. This task model means to take an information picture and create a sentence depiction of the fundamental substance of the picture. Portraying the substance of a picture in straightforward language is one of the intricate and principal undertakings. With the assistance of trend setting innovation and the accessibility of datasets, building models has now become a potential errand.

People, with the assistance of their sight vision, can characterize and precisely tell the portrayal of any picture introduced to them. Much the same as people, Machines have been developing at a quick rate and can perceive the essential activities classified by an object, perceive its state and highlights. Despite the fact that, characterizing a picture with exactness in basic and plain language which is effectively understandable by people has become a generally new and intriguing errand.

Auto Image Captioning plays out its capacity in a succession of undertakings. The initial move towards understanding a picture starts with the extraction of the picture with its relative encompassing the following stage, the connection between the distinguished articles has been recognized for additional assessment for example for objects book and table, the connection between two to be characterized as "the book is on the table".

When the items and their associations with one another have been characterized, further valuation happens in the content portrayal. Arrangements of words must be placed as it were so when once

framed it will bode well and legitimize the genuine relationship of articles set in the picture.

For the main undertaking for example for separating the highlights out of the picture we have utilized Convolutional Neural Network (CNN) in this task. It is essential to take note of that 'extracting feature' alludes to eliminating the last softmax layer in most cases. For the subsequent part, which is to produce a printed depiction we will utilize 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.

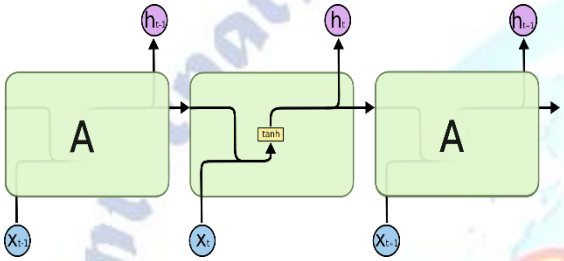

Fig1. Overview of the LSTM

II. RELATED WORK

In Work[1] it was obviously indicated that the Deep learning has begun getting a great deal of consideration in most recent couple of years and a ton of headways have additionally been made in this field. This is very obvious too when we take a gander at the details. In 2015 just four effective articles were distributed however after that the ubiquity of the field developed dramatically and it tends to be seen from the way that 57 articles were distributed in 2017–2018.

Work[2] done by Di Lu and Spencer Whitehead proposed that another assignment can be made for which picture depictions will be given as contribution to the framework. The paper likewise referenced that the Image Captioning which has been being used presently does not have the particular inspiration of elements that frames the fundamental structure of picture. In this paper, they additionally proposed the answer for this issue. The paper recommended that CNN-LSTM model should be prepared so it will be fit for producing inscription dependent on pictures spoke 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 fed into 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 paper considers all the previous research that has been done in this field as of now and is likewise influenced from those research. The majority of the works that we have contemplated utilizes CNN and RNN based design. An intriguing finding that we have from the past exploration done on this point is that "adding more layers to the model doesn't really implies that we will get more precision".

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

III. METHODOLOGY AND IMPLEMENTATION

As of now examined in the theoretical, the fundamental point of this paper is to give subtitles to the picture in the real time. The dataset that is utilized for building this undertaking is Flickr8k dataset. In Flickr8k dataset each picture has 5 subtitles comparing to it. The dataset gives 6000 pictures to preparing purpose,1000 pictures for approval reason and the rest 1000 pictures for the testing reason. The undertaking has been separated into five errands essentially:

Data Cleaning

Getting the Image id from the Dataset and making a dictionary that will map pictures with the

captions. The token.txt document has picture id and captions as sections and from this token.txt record just we will plan each picture with their individual inscriptions. The all out words that we have in our dataset is near around 37000. Presently we need to diminish this words as this will impact our calculation and furthermore on the off chance that a word is coming for less time, at that point it doesn't bode well to utilize it. Presently, we have put a limit estimation of 10 so in the event that a word has recurrence under 10, at that point we're not mulling over that word. Subsequent to separating the words based on edge recurrence we have just 1845 words and this comprises our jargon word reference.

Encoding the Images

Presently we can give a image to our model anyway as opposed to people, machines can't comprehend the image by observing them. So we need to change the photo into an encoding all together that the machine can perceive the examples in it. For this, We have utilized the transfer learning for example we utilize a pre-trained rendition that has been now taught on huge datasets and concentrate the capacities from these examples and use them for our photographs. For this examination, we have used InceptionV3 pretrained model which has been already trained on Imagenet. We can without much of a stretch import this model from keras.

```
def preprocess(path,img_to_captions):

    preprocessed_images = {} ## map img_id with feature vector of corresponding img
    cnt = 0

    for img_id in img_to_captions.keys():

        image_path = path + "\\" + img_id + '.jpg'
        try :
            img = image.load_img(image_path,target_size=(224,224))
        except :
            continue
        img = image.img_to_array(img)
        img = np.expand_dims(img,axis = 0) ## to make img look like batch of size 1 = (1,224,224,3)

        # Normalising
        img = preprocess_input(img)
        feature_vector = model.predict(img) # shape (1,2048)
        feature_vector = feature_vector.reshape((-1,)) # shape = 2048

        preprocessed_images[img_id] = feature_vector

        cnt += 1
        if cnt%100 == 0:
            print('|',end = ' ')

    print("Preprocessing Completed")
    return preprocessed_images
```

```
start_t = time.time()
train_images = preprocess(path, train_captions)
end_t = time.time()
print('Time taken: ', end_t - start_t)
```

```
| | | | | | | | | | | | | | | | | | | | | | | | | | 
| | | Preprocessing Completed
Time taken:  3016.9878885746
```

Vectorizing 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.

```
def create_mapping(vocab):

    word_to_index = {}
    index_to_word = {}

    for index,word in enumerate(vocab):
        word_to_index[word] = index + 1
        index_to_word[index + 1] = word

    index = len(word_to_index)
    word_to_index['<start>'] = index + 1
    index_to_word[index + 1] = '<start>'

    word_to_index['<end>'] = index + 2
    index_to_word[index + 2] = '<end>'

    return word_to_index,index_to_word
```

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.

```
def build_model(embedding_matrix,max_length,vocab_size,embedding_dims):

    ## smaller model 1
    image_input = Input(shape = (2048,)) ## NOTE : input given in batch but shape of one sample given
    L1 = Dropout(0.3)(image_input)
    image_encoding = Dense(256,activation = 'relu')(L1)

    ## smaller model 2
    caption_input = Input(shape = (max_length,))
    L2 = Embedding(input_dim = vocab_size + 3, output_dim = embedding_dims , mask_zero = True)(caption_input)
    L3 = Dropout(0.3)(L2)
    caption_encoding = LSTM(256)(L3)

    ##DECODE
    decoder = add([image_encoding,caption_encoding])
    L4 = Dense(256,activation='relu')(decoder)
    output = Dense(vocab_size + 3,activation='softmax')(L4)

    ## Combined Model
    model = Model(inputs = [image_input,caption_input], outputs = output)

    ## setting the embedding layer bu ourself and not training it
    model.layers[2].set_weights([embedding_matrix])
    model.layers[2].trainable = False

    return model
```

IV. DISCUSSION

4.1 Challenges confronted

4.1.1 Detecting Multiple Objects

The models that we have these days can identify different articles yet models can't generally decipher the connections present between those items. Subsequently the model can't generally give exact portrayals of the picture. Likewise, the dataset which we have utilized for example Flickr8k dataset has just 8k pictures. Presently on the off chance that we need our model to precisely depicts the picture and that too in linguistically right structure then we need to prepare our model on a lot bigger dataset. Discussing huge datasets, enormous datasets likewise set aside tremendous effort to prepare in this manner speed of preparing, testing additionally stays a major issue that is should have been tended to.

4.1.2 Availability of Datasets

The most widely recognized datasets that are commonly utilized for Image Captioning are Flickr8k, Flickr30K and MS-COCO. Presently these datasets are generally in English. As referenced in work[] starting at now we have a ton of datasets which we can use to prepare our model yet the vast majority of the preparation tests are either in English or Chinese. This is a significant issue to deliver in the event that we need to utilize the image captioning model for useful applications, at that point accessibility of cross language preparing tests are a lot of required.

4.2 Applications

To help visually impaired individuals in the society, image subtitling can come out to be a useful gadget whenever grew precisely. The advancement of a programmed picture captioning framework that gives exact picture captions as a free framework can be an extreme assignment. Here, Images that have been taken can be utilized as a contribution for auto picture subtitling. Accordingly, the yield can be furnished with the assistance of noisy clamor, which can enable visually impaired individuals to more readily comprehend their environmental factors/surroundings.

V. RESULTS

Actual caption: a black and white dog jumps over a hurdle
Predicted caption: a black and white dog jumps over a hurdle

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

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

VI.CONCLUSION

Deep learning can give astounding changes in the general public and lately image captioning has made significant advances. Image Captioning can give a ton of uses in different areas like agribusiness, savvy checking of the frameworks. It is very stunning to see that image captioning isn't utilized in spaces like traffic examination which could be profited a great deal by it. This examination depends on different articles and past explores done in the field. The examination searched for different explicit models and systems utilized for picture subtitling and we found that the for separating highlights and substance CNN is the most appropriate model and is generally utilized also. For producing portrayal, the models which are regularly utilized are RNN and LSTM(special sort 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 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 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