B. TECH. (CSE)

VI SEMESTER

UE19CS343-Topics in Deep Learning

Final Report of Project Implementation

Image Captioning

Submitted by : Batch 5

Rahul S Bhat: PES2UG19CS315

Rishika Satheesh :PES2UG19CS330

Rithika Reddy Kara :PES2UG19CS331

January-May 2022

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
ELECTRONIC CITY CAMPUS, BANGALORE

TABLE OF CONTENTS

1.Introduction

Problem Statement

The process of text generation for a given set of visuals/images. The describing of the picture by applying machine learning and deep learning concepts and models to create suitable text which corresponds to the images respectively.

Applications

The applications of image captioning are simple and yet profound and have a great impact. The use of image captioning in the field of social media is tremendous. The visually impaired get a great benefit with the knowledge of the world around them. To capture a photo of their surroundings, and having the captions generated accurately will help the visually impaired self reliant. Virtual assistants have it easy with the image captioning helping them around speaking of the text or a document where images are present. Image indexing is another great application of image captioning.

2.Literature Survey

Paper 1:
Image Captioning - A deep learning ApproachPublished: 2018
Published: 2018

Authors: Lakshminarasimhan Srinivasan, Dinesh Sreekanthan, Amutha A.L

Proposed Methodology: VGG 16 for identification of objects, LSTM for captioning, BLEU as evaluation metric

Dataset Used: Flickr 8K

Result: Effective BLEU score of 0.683 was obtained

The CNN and LSTM models were used with the Flickr 8K dataset. The more you train, the accuracy is better. The training phase of the model consisted of three parts, The Image feature extraction. The images were fed into the VGG 16 model. After the feature extraction, the images are converted into a vector element representation as the output. Those vector element representation is fed into the LSTM layer. The second part is the sequence processor. It will extract features for the text which is needed. Next, comes the training part. In the training phase, inputs are provided in two parts. The image and it's appropriate captions.

Paper 2:

Automatic Image Captioning using Neural Networks

Published: March 2020

Authors: Subash Pandey, Rabin Kumar Dhamala, Bikram Karki, Saroj Dahal and Rama Bastola

Proposed methodology: CNN,LSTM

Dataset Used: MSCOCO

In this model, the top down and bottom up paradigms are combined to come up with the recurrent neural network and building upon both of them to generate captions. KNN approach is used to find the nearest images and deep features are utilized to come up wit a set of possible captions out of which one will be chosen. Out of the set of captions, where will be a consensus caption which will have the highest score to be chosen as the final caption. The process of chosing the highest scoring caption as mentioned is known as CIDEr(Consensus-based Image Description Evaluation)/ BLEU:bilingual evaluation understudy metric. The architecture consists of encoder-decoder as well as multimodal parts. With CNN for feature extraction.

Paper 3:

An overview of Image Caption Generation

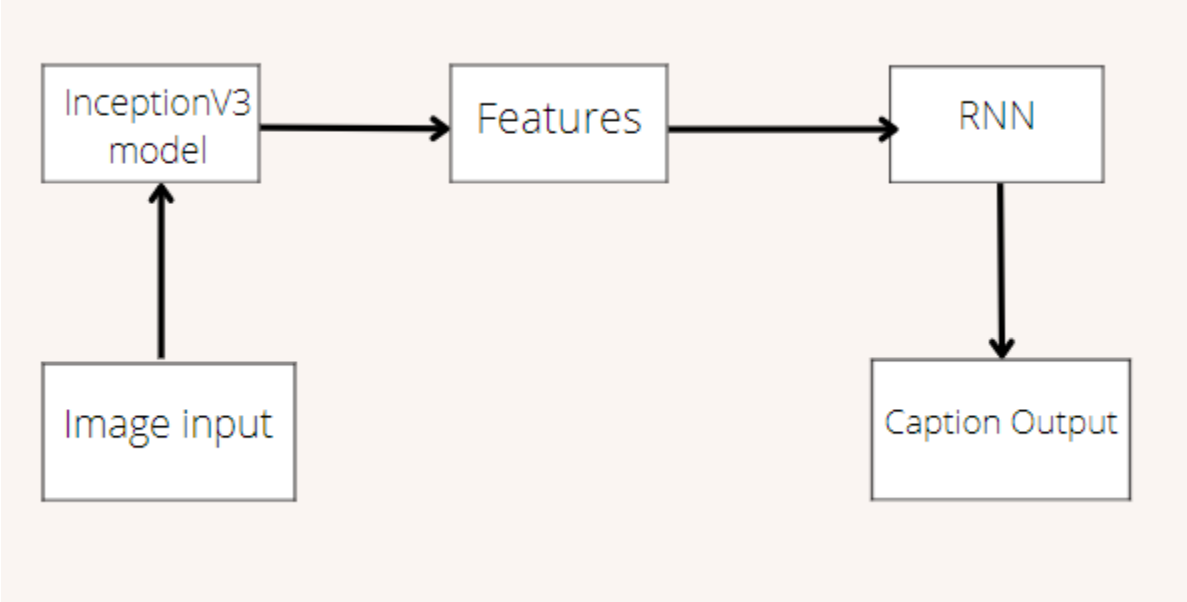
Published: 11 Dec 2019

Authors: Haoran Wang, Yue Zhang, Xiaosheng Yu

The basic approach for this project is by taking the parts of speech and identifying which category each word is taken into account. The

prediction of the image will happen with the most likely each part of the speech utilized. The 3D visual analysis system is utilized to make a correct a sentence from the most used part of speech obtained which are combined with the semantic trees and formed with the grammar. The maximum likelihood estimation is used for te feature extraction. The words are detected, adn made into a sentence. The rerank of the sentences is done, to fit it into an order which makes the most sense. THe statistical language model is used. Topics like RNN, LSTM, attention mechanism are utilized as the deep learning topics to integrate into this project. The metrics BLEU,METEOR,ROUGE,CIDEr,SPICE are used in their respective fields as a metric for moving forward with the best.

3. Design



Two models were utilized to obtain the respective captions. The inceptionV3 model and the RNN, to obtain the output. The inceptionV3 model, is a image recognition model from GoogLeNet.

It is trained on the dataset ImageNet dataset. A convolutional neural network model which is utilized for the image feature extraction. The convolution of images takes place with the help of kernels. The sizes of kernels will be different, with each pixel and it's local neighbors included. The pooling which is required for reducing the dimensions of the feature maps, would do both max pooling as well as pooling average. With regularization being done by auxiliary classifiers, inceptionV3 is very unique. As usually, auxiliary classifiers are used to have a deeper network, not for being used as regularizers. Recurrent Nueral Netwroks, is the other model which was utlilized to generate captions from the extracted features from the images.The images are fed into the inceptionv3 model as the input and the features as the output. The features fed as input into the RNN to obtain the captions as the output.

4. Implementation

For each image, there are captions provided. The images are separated by hashes and captions by tags respectively. It's all stored in a dictionary. With keys as a particular image, and the values as corresponding list of captions. For the training part, the list of images and list of captions are kept separate. The vector's are produced which are fed into the CNN. For the training part, fetch the corresponding captions from the respective images provided and add them to the training captions list. The same goes for testing set as well. Use the encoder to get the features which are fed into the layers of RNN. The MSCOCO dataset was used for the project.
Size of training dataset - 40k
Size of testing dataset - 8k

The descriptions is needed, load_descriptions function is needed to to make a list of all of the descriptions. Descriptions is a dictionary, maintained with key as the specific image id provided in the dataset,

and the values as the description of the specific image. The next step is to clean out the descriptions and store them separately and have them in clean_descriptions function.

The loaded descriptions are converted into a vocabulary of words saved into a file, all the descriptions one per line. A list of all of the training images with their full path are noted into a list. The same occurs with the testing set. The images with the image path are sent into the preprocessing phase where all the images are turned into the target size of 299,299.

The inception v3 model is loaded. A new model is created by removing the output layer. The images are all encoded inot a vector of size 2048. Next, all the images are encoded after which, the bottle neck of the train features are saved to the disk. The function to encode all the test images. A list of all of the training captions are put in a list. From the dictionary of clean discriptions, they will be converted into a list of descriptions. The embedding of the layers is done, and the initial weights for the model are set into it.

The adam optimizer and loss function cross entropy is used.

5. The Testing and results

The results all noted down though with the random images brought in and fed into the system. Training loss was found to be around 2.07. The BLEU score was used as a similarity measure between the actual and the predicted captions.The result should display the image with the correct caption.

6. Conclusion

The InceptionV3 is pre-trained on the Imagenet dataset and is a module of GoogLeNet.This model was trained for 10 epochs on 40k training samples. The loss was found to saturate around 2.07. The highest obtained BLEU score was 0.75.

7. References

[1]https://www.ripublication.com/ijaer18/ijaerv13n9_102.pdf

[2]https://www.academia.edu/69976768/Automatic_Image_Captioning_Using_Neural_Networks

[3]https://www.hindawi.com/journals/cin/2020/3062706/

[4]https://www.youtube.com/watch?v=3e4hsHLDHZA&feature=youtu.be

[5]https://youtube.com/playlist?list=PL12YWfULs0pnL_9Pj6udM6PE2-SWZbTlX