ISSN No:-2456-2165

Image Captioning and Image Retrieval

Ankit Kumar
Department of Information Technology
Maharaja Agrasen Institute of Technology (GGSIPU)
Delhi, India

Kinshuk Kumar
Department of Information Technology
Maharaja Agrasen Institute of Technology (GGSIPU)
Delhi, India

Meenu Garg (Assistant Professor)
Department of Information Technology
Maharaja Agrasen Institute of Technology (GGSIPU)
Delhi, India

Abstract:- Most images do not have a description, but the human can largely understand them without their detailed captions, but machine needs to understand some form of description of image. In Image Captioning, textual description of an image is generated. These captions could be used for various purposes like automatic image indexing. Image indexing is an important of Content-Based Image Retrieval (CBIR) and hence, it can be applied to many areas, including biomedicine, commerce, the military, education, digital libraries, and web searching. We have divided the task into two parts- one is image based model which is used to extract the content of the image for that purpose we have used CNN model and other a language model which is used to translate the feature in sentences for that purpose we have used RNN(LSTM).The applicative part of this model is Image Retrieval ,both Text based(TBIR) and Content based(CBIR).In TBIR, a keyword (text) is used to perform image search whereas in CBIR, an image is used as the search object. We used Flickr8k dataset to train our model which consists of 8000 images along with their captions. For evaluation we used BLUE metric. BLEU (Bilingual evaluation understudy) is a metric that is used to measure the quality of machine generated text.

Keywords:- *CNN, RNN, LSTM, TBIR, CBIR, BLEU.*

I. INTRODUCTION

Image captioning means generating a description in form of text for an image. Image captioning needs understanding and recognition of objects. It requires a computer model or an encoder, to understand the content of the image and a language model or a decoder from natural language processing to convert the understanding of the image into words in correct order. What is most impressive about these methods is a single end-to-end model can be defined to predict a caption, given a photo, instead of requiring complicated data preparation. In deep learning based techniques, automatically features are learned from training data and large and diverse set of images are also handled. We have used pre-trained VGG 16 model which won ImageNet competition. Vanilla RNN suffers from the problem of vanishing gradient for that purpose we have used LSTM. LSTM is capable of keeping track of the objects that have already been described using text. While generating captions, image information is included to the initial state of an LSTM. The next words are generated based on the previous hidden state and current time step. This process is repeated until it gets the ending token of the sentence. We have used Flickr8k dataset for our training our model because it is small in size and can be loaded completely in ram. We can load the entire dataset in ram in one go. BLEU (Bilingual evaluation understudy) is a metric that is used to measure the quality of machine generated text. The central idea behind BLUE is to check the closeness in the text generated by machine to a human translator. However, grammatical correctness is not taken into account herein the image retrieval part, images are searched on the basis of caption generated from the image captioning model. In text based image retrieval, user inputs the keyword and search operation is performed on image database whereas in content based image retrieval, an image is used for performing search operation.

II. RELATED WORK

Krizhevsky et al. [1] proposed a neural network that uses non-saturating neurons and a very efficient method GPU implementation of the convolution function. They employed a regularization technique called dropout which reduces the overfitting of the training dataset. Deng et al. [2] proposed a new database which they called ImageNet. ImageNet was an extensive collection of images. ImageNet organized the divergent classes of images in a densely populated semantic pecking. Karpathy and FeiFei [3] made use of datasets of images and their textual descriptions in order to learn about the inner correlation between visual data and language. They described a Multimodal Recurrent Neural Network architecture that employ the inferred co-linear arrangement of features in order to learn how to generate novel descriptions of images. Yang et al. [4] implemented a system for the automatic generation of a natural language interpretation of an image, which will help acutely in image understanding. The proposed multimodel neural network technique, consisting of object detection and localization modules which is very like to the human visual system which is able to learn how to describe an of images automatically. Aneja et al. [5] proposed a convolutional neural network model for machine transferal and conditional image generation in order to solve to problem of LSTM being composite. Xu et al. [6] proposed an attention based model that learned to describe the image regions automatically. The model was trained using backpropagation techniques. The model was able to recognize object boundaries, at the same time was able to induce a precise description. Flickr8k [7] is a widely used dataset that consists of 8000 images taken from Flickr.com. The dataset consists of 8,092 photographs in JPEG format and each image in the dataset has 5 corresponding captions

annotated by humans. The dataset is divided into three disjoint sets: pre-defined training dataset containing 6,000 images, development dataset and test dataset containing 1,000 images each. BLEU (bilingual evaluation understudy) [8] metric is used for evaluating the text that has been generated by a machine from one language to another. The central idea behind BLUE is to check the closeness in the text generated by machine to a human translator. In estimating the overall quality of the generated text, the computed scores are averaged. However, grammatical correctness are not taken into account. BLUE metric ranges from 0 to 1. Sepp Hochreiter and JÃijrgen Schmidhuber describes LSTM [9] neural networks type of recurrent neural network that has hidden units along with standard units .LSTM units consists of 256 hidden states, it is memory cell that can preserve information in memory for longer durations. LSTM solves the vanishing gradient problem of vanilla RNN. Huang and Dai [10] proposed a texture-based image retrieval system which coalesce the wavelet decomposition and the gradient vector. Lin et al. [11] proposed a method using Colour-Texture and Colour-Histogram for Image Retrieval .The model uses three disjoint features namely image colour, image texture and colour distribution. Hiremath and Pujari [12] proposed a content based image retrieval system that utilised colour, texture and shape features and divides the image into matrix. Zhang [13] proposed a method in which he enumerates colour and texture features from the image database. So, for every image, colour and textures features are calculated. For any query image these features are calculated and then the images are classified according to the colour features then the top-ranked images from colour features are again ranked according to the texture features.

III. DATASET

Various datasets are used for training, testing, and evaluation of the image captioning methods. The dataset differ in various perspective such as the number of images, the number of captions per image, format of the captions, and image size. Some of the popularly used datasets for training are: Flickr8k, Flickr30k, MS COCO, Instagram dataset, Stock3M, FlickrStyle10k.

We have used Flickr8k dataset for our training our model because it is small in size and can be loaded completely in ram. We can load the entire dataset in ram in one go. The dataset consists of 8,092 photographs in JPEG format and each image in the dataset has 5 corresponding captions annotated by humans. The dataset is divided into three disjoint sets: pre-defined training dataset containing 6,000 images, development dataset and test dataset containing 1,000 images each.

IV. EVALUATION METRIC

BLEU (bilingual evaluation understudy) metric is used for evaluating the text that has been generated by a machine from one language to another. The central idea behind BLUE is to check the closeness in the text generated by machine to a human translator. In estimating the overall quality of the generated text, the computed scores are averaged. However, grammatical correctness are not taken into account. The performance of the BLEU metric is varied depending on the number of reference translations and the size of the generated text. The BLUE metric ranges from 0 to 1. Score closer to 1 is considered to be perfect i.e. machine generated and human annotated text are similar and score 0 is perfect mismatch. This evaluation method was proposed by Kishore Papineni, et al. Since it is inexpensive to calculate and language independent it is widely accepted and used.

The following equation is used to calculate BLUE score:

$$\text{BLEU} = \min\left(1, \frac{\textit{output-length}}{\textit{reference-length}}\right) \left(\prod_{i=1}^{4} \textit{precision}_i\right)^{\frac{1}{4}}$$

V. ENCODER-DECODER ARCHITECTURE

Our model is neural network-based image captioning which works in end to end manner. Here image features are extracted VGG-16 model which is a 16 layer model and then fed them into an LSTM to generate a sequence of words, which are later used to form meaningful sentences. We described the model in three parts:

A. Photo Feature Extractor

This is a 16-layer VGG model pre-trained on the ImageNet dataset. We have pre-processed the photos with the VGG model. We have removed the last layer as we are using this for feature extraction not for classification .Later the features extracted by the model are fed to the LSTM. These features can be used as interpretation of a given image in the dataset. The input to the feature extractor is given in the form a 4,096 element vector which then are processed by a Dense layer to produce a 256 element representation of the photo.

B. Sequence Processor

This is embedding layer, followed by a fully connected layer with 256 memory units (LSTM) recurrent neural network layer. The function of a sequence processor is for handling the text input by acting as a word embedding layer. The embedded layer consists of rules to extract the required features of the text and consists of a mask to ignore padded values. This is followed by an LSTM layer containing 256 hidden states. The dropout layer is there to reduce overfitting of the training dataset.

C. Decoder

The final phase of the model is to merge the feature extractor and sequence processor together which are later processed by a 256 neuron layer to make a final prediction. The networks are merged simply by adding two vectors of size 256. This is then fed to two Dense layer and a final Dense layer that return the probability that the word is in vocabulary by making a softmax prediction over the entire output vocabulary in order to predict the next word for the sequence.

ISSN No:-2456-2165

VI. IMPLEMENTATION

A. Image Captioning

Firstly the "image features" are pre-computed using the pre-trained captioning model and then save them into file. These features are then loaded and fed into the model as an interpretation of the photo in the dataset. So instead of running the photo through full VGG model, we do it through the pre-computed image features, which are both similar. This helps in faster processing and working of the model, making it more optimized and also consumes less memory. The last layer of the CNN model was removed since classification is not required in image captioning. These are the "features" that the model has extracted from the photo. Following steps for taken for description cleansing.

- ✓ All the words were converted to lowercase.
- ✓ All punctuations were removed.
- ✓ Removed all words that are less than or equal to one character in length.
- ✓ All the words having number in them were removed.

Training of data was done on all the photos and captions present in the dataset. The model we have developed will generate a caption for a given photo. We used the strings 'startseq' and 'endseq' to mark the start and end of the caption. Evaluate the model using BLEU score. And finally generating new captions.

B. Retrieval of Image

➢ *Database and Description Generation*

All the images on which the search queries are to be processed are placed in a separate folder. This folder serves as the Database to the search query. Directory of the folder is stated and captions are generated for each image using image captioning model. These captions are stored in the *description.txt* file which is used for retrieving images as per the search queries. This *description.txt* file contains textual description of the images along with the filename of the image.

➢ *Text based Image Retrieval*

Firstly, the keyword for search is entered by the user, this keyword is then searched in the descriptions generated in the file *description.txt*. If any match is found, all the images corresponding to the keyword are shown as the result by using Matplotlib library.

➢ *Content based Image Retrieval*

In CBIR, instead of using a keyword entered by the user, an image is used for performing search operation. The search image is passed through the Image description model which generated a textual description of the image. All the stopwords from the description generated from the image are removed and only main keywords are used. These keywords are then matched in the *description.txt* file if any match is found then all the images corresponding to the keyword are shown as the result by using Matplotlib library.

VII. RESULTS

We implemented the model and it was able to generate captions successfully. Fig 1 shows the input image along with the predicted caption that is dog is running through the grass. Fig 2 shows a test image given to the model in order to retrieve similar images, it extracts the features and generates keywords like dog and water and returns images on that basis. Fig 3 and Fig 4 are the retrieved images. Fig 5 shows the evaluated BLUE score of our model. Fig 6 shows a graph of accuracy and validation loss.

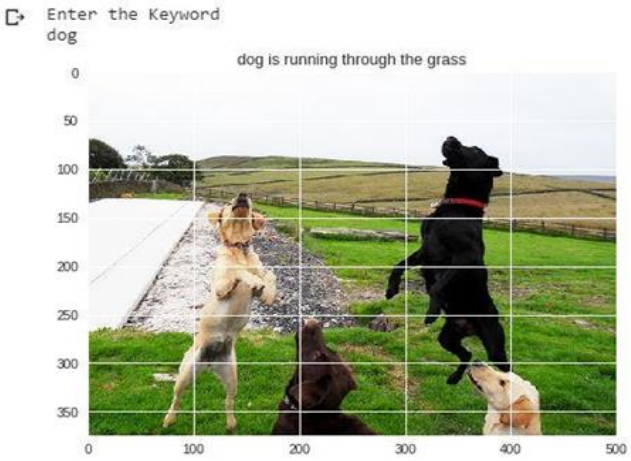

Fig 1:- Input Image

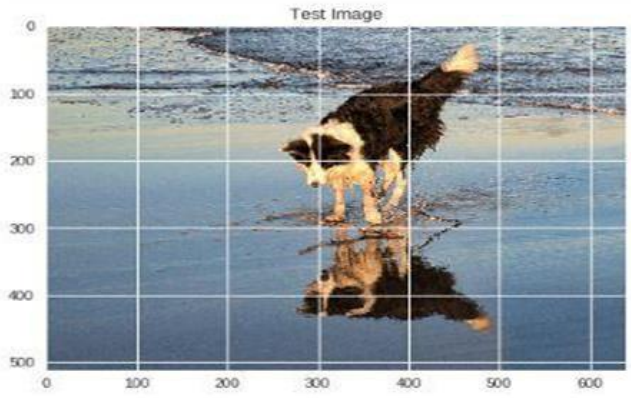

Fig 2:- Test Image

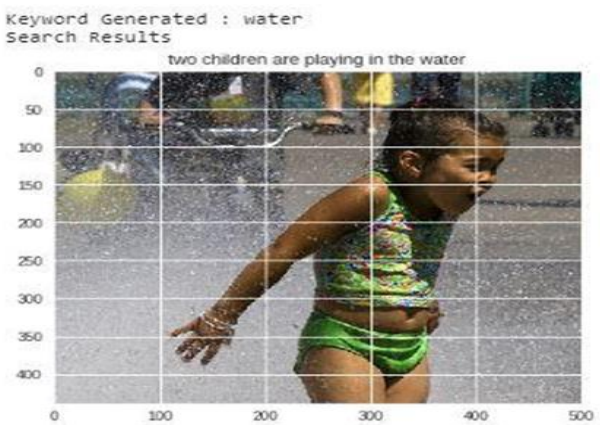

Fig 3:- Retrieved Image

ISSN No:-2456-2165

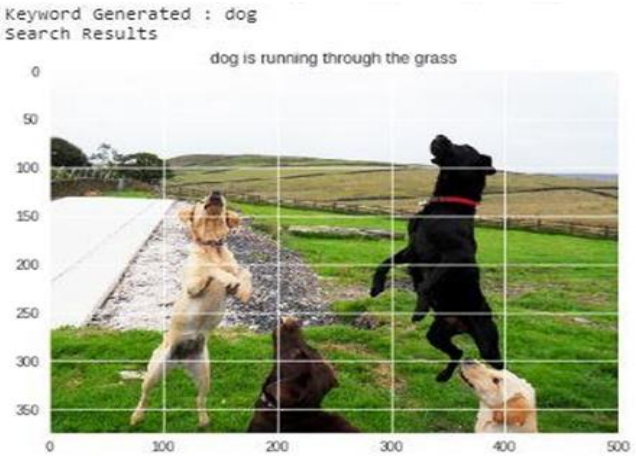

Fig 4:- Retrieved Image

```
# load the model
filename = 'model_19.h5'
model = load_model(filename)
# evaluate model
evaluate_model(model, test_descriptions, test_features, tokenizer, max_length)
```

```
Dataset: 6000
Descriptions: train=6000
Vocabulary Size: 7579
Description Length: 34
Dataset: 1000
Descriptions: test=1000
Photos: test=1000
BLEU-1: 0.504166
BLEU-2: 0.259168
BLEU-3: 0.169212
BLEU-4: 0.072790
```

```
ls
```

```
drive/          model_11.h5  model_15.h5  model_19.h5  model_4.h5  model_8.h5
features.pkl    model_12.h5  model_16.h5  model_1.h5   model_5.h5  model_9.h5
model_0.h5      model_13.h5  model_17.h5  model_2.h5   model_6.h5  model.png
model_10.h5     model_14.h5  model_18.h5  model_3.h5   model_7.h5  sample_data/
```

Fig 5:- BLUE Score

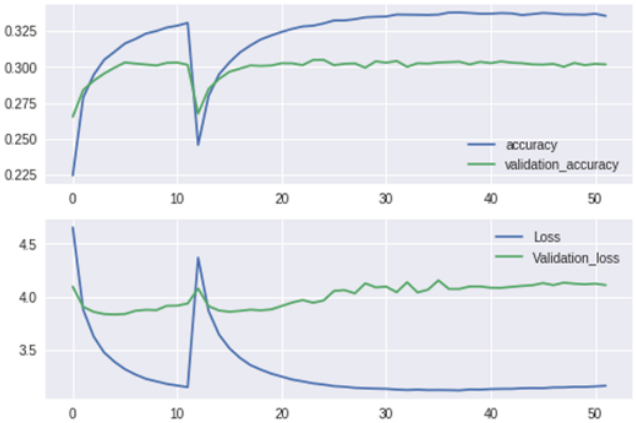

Fig 6:- Accuracy and Validation Loss

VIII. CONCLUSION

In this paper the authors have developed an image captioning model using deep learning approach with VGG-16 model feature extractor and LSTM for making descriptive sentences which are syntactically and semantically correct and we have achieved BLEU score of 0.50.Text based and Content Based Image Retrieval is also performed.

REFERENCES

[1]. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks". Volume 60 Issue 6, Pages 84-90, June 2017.Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li and Li Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database.

[2]. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li and Li Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database". IEEE Conference on Computer Vision and Pattern Recognition 2009.

[3]. Andrej Karpathy, Li Fei-Fei, "Deep Visual Semantic Alignments for Generating Image Descriptions", Volume 39 Issue 4, Page 664-676 IEEE Transactions on Pattern Analysis and Machine Intelligence, April 2017.

[4]. Zhongliang Yang,Yu-Jin Zhang, Sadaqatur Rehman, Yongfeng Huang, "Image Captioning with Object Detection and Localization", arXiv:1706.02430, 2017.

[5]. Jyoti Aneja, Aditya Deshpande, Alexander Schwing, "Convolutional Image Captioning", arXiv: 1711.09151, 2017.

[6]. Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, Yoshua Bengio, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention", International Machine Learning Society, Volume 37, Pages 2048-2057, 2015.

[7]. Micah Hodosh, Peter Young, and Julia Hockenmaier. "Framing image description as a ranking task: Data, models and evaluation metrics". Journal of Artificial Intelligence Research 47 (2013), 853–899, 2013.

[8]. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. "BLEU: a method for automatic evaluation of machine translation". In Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 311–318, 2002.

[9]. Sepp Hochreiter and JÃijrgen Schmidhuber. "Long short-term memory". Neural Computation 9 8 (1997), 1735–1780, 1997.

[10]. P. W. Huang and S. K. Dai, "Image retrieval by texture similarity," Pattern Recognit., vol. 36, no. 3, pp. 665–679, 2003.

[11]. C.-H. Lin, R.-T. Chen, and Y.-K. Chan, "A smart content-based image retrieval system based on color and texture feature," Image Vis. Comput., vol. 27, no. 6, pp. 658–665, 2009.

[12]. P. S. Hiremath and J. Pujari, "Content Based Image Retrieval based on Color, Texture and Shape features using Image and its complement," Int. J. Comput. Sci. Secur., vol. 1, no. 4, pp. 25–35, 2007.

[13]. Zhang, "Improving Image Retrieval Performance by Using Both Color and Texture Features," Third Int. Conf. Image Graph. pp. 172–175, 2004.