Image Retrieval Using Image Captioning

By Ruthu GS,Varsha C and Namana under the guidance of Prof.Dinesh Singh

Abstract— ***Due to the advance technologies in social media platform, there has been tremendous increase in the usage of digital images. Hence the complexity of searching a specific image and retrieving the particular data associated with is becoming a tedious task. In this paper, we concern about the problem of complex image retrieval by reasoning image dense captions, which is similar to the way of human perception for searching images. Specifically, we transform the problem of complex image retrieval into a captioning issue by using structured language descriptions for retrieval. In this current work we are using MSCOCO 2014 dataset. Furthermore, a generative merge model based on Convolutional Neural Network (CNN) and Long-Short Term Memory (LSTM) is applied especially for image captioning.Caption -Matching is quite critical for ranking images,Therefore jaccard similarity is used as a matching strategy.***

Keywords—Index Terms—Image Retrieval, Dense Caption Reasoning,Captioning, Deep Learning

I. INTRODUCTION

Retrieving images by visual query is one of the most attracting vision problem, which aims to search for images by reasoning about the visual elements of query image. It is a very challenging problem since an ideal retriever should be able to not only understand the whole scene but also the describing contents in details. Plenty of previous arts exist for addressing this task.An image consists of several information such as the objects, attributes, scenes and activities. Humans are capable of generating captions for images with much less difficulty. However, automatic caption generation for a given image is a very challenging task for machine (Yang et al., 2018).

Automatic image caption generation involves two tasks: 1) recognizing and understanding significant objects in an image and 2) describing the proper relationship between these objects. To perform these two tasks, image captioning uses a combination of two subnetworks, CNN for salient object detection in images and LSTM for understanding relationship objects and decoding into sentences (Shiru et al., 2017).

With the availability of extremely large numbers of images in internet nowadays, image captioning becomes more and more popular for retrieving images by Google search engines or newspaper companies (Huda et al., 2018). In addition, image captioning is useful for description of images for visually impaired persons, teaching concepts for children and social media network like Facebook and Twitter can directly generate captions from images (Zakir et al., 2018).

Caption matching is quite critical for ranking images given the produced query and candidate captions. It is a text matching problem in the field of NLP(Natural Language Processing), and traditional method for text matching involves string-based method[9], corpus-based method[10] and knowledge-based method[10]. However, these methods are not designed for image caption matching, which concentrates on matching the structured visual elements in images, i.e., objects, interactions between objects and the attributes of objects. Therefore, Jaccard similarity is presented to handle this problem.

II. LITERATURE STUDY

In this section we'll brief about the papers referred.

In [1] the researchers have focused on showing the way to minimize the similarity gap in the retrieving method. Captioning of images is done using Neural Network like CNN and RNN. In brief Object method of classification is done using the methods like CNN and the obtained output from CNN is sent to RNN it is considered as its input. Later RNN will make use of LSTM to persevere the necessary information in the storage.

In [2]The proposed approach focuses on generating the dense captions targeting the regions and then using scene graph parser for structuring the generated caption and matching to the other images in the dataset, calculating their distances. The results prove, this approach is effective over several baseline models.

In [3] The dense captioning is introduced which needs a model to parallelly limit and describe regions of an image. Later the FCLN architecture was developed. FCLN architecture is mixture of CNN-RNN models. Experiments in both creation and extraction settings which will reveal the power and as well as regulation of the model.

In [4] They have explained about their model which outputs the human annotated descriptions for an image based on it's regions which are built with very less assumptions. They have explained about Multimodel Recurrent Neural Networks which helps in generating descriptions of visual data.

In [5] The captions generated by traditional methods tend to be repetitive making use of self retrieval module and reinforcement learning(backpropagation)the proposed approach improves distinctness of captions. By the results we also conclude that state-of- the-art performance can be achieved on two widely used captioning datasets

In [6],they have used scene graphs as a novel representation for detailed semantics in visual scenes, and introduced a novel dataset of scene graphs grounded to real-world images. Later they have used this representation and dataset to construct a CRF model for semantic image retrieval using scene graphs as queries. And as well as they have shown that this model out-performs methods based on object detection and low-level visual features.

III. DATASET

We are using *MSCOCO 2014 dataset* which is created by collecting images of everyday whereabouts

which has objects that are common in their natural context. . Each image is associated with 5 human-annotated captions as shown in the fig_1 below. With total of 91 object categories and 2.5 million instances that are labeled in 330k images, it is used for object detection, image classification and image captioning. There are 82,783 images for training, 40,504 images for validation and 40,775 images for testing.The system can recommend images from COCO training and validation datasets and our own dataset scraped from Google Images from which the images are recommended. The system can also retrieve similar images from Google images.

A man is skate boarding down a path and a dog is running by his side.
A man on a skateboard with a dog outside.
A person riding a skate board with a dog following beside.
This man is riding a skateboard behind a dog.
A man walking his dog on a quiet country road.

Fig_1

IV. METHOD

In this chapter we will be discussing about our approach, implementation and methodologies used.

Overview :

This project can be divided into two divisions:

Image Captioning System :

The first part is image captioning, where given an image the system must be able to generate captions. This system should generate captions for the images in the dataset as well as for an unknown image that is given as query.

Image Retrieval System :

This is the second part where in the similarity calculation between the generated captions of the input image and the captions of the images in the dataset is calculated and based on that score most similar images to the query is returned.

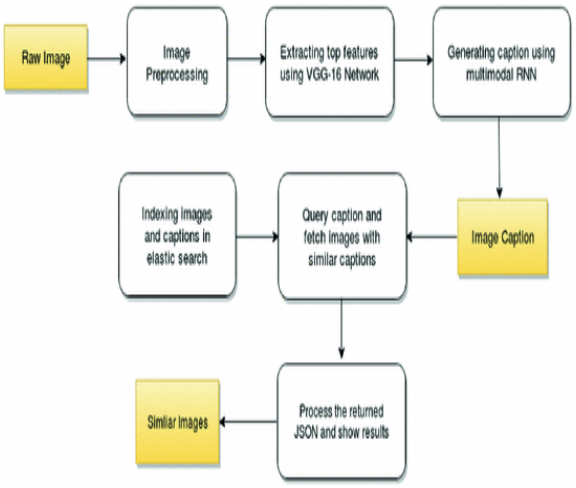


Fig.2

Data Preparation : Feature extraction from the image is an important step in image processing using deep learning. In this the images are converted to features in order to train the deep learning model. Thus, CNN is used for image feature extraction because this neural network was successful in classifying the image into one of the 1000 classes in ImageNet dataset.

Data Preparation for Caption generation

The dataset we have used in this project is MSCOCO dataset. In this dataset, each image has multiple captions associated. So, we are taking image id as the key and the values are its associated captions.

The raw text of the captions are converted to a format to feed to neural networks by:

1.Removing punctuation

2.Removing numbers

3.Removing stopwords

4.And converting uppercase letters to lowercase.

Original Captions	Captions after Data cleaning
Two people are at the edge of a lake, facing the water and the city skyline.	two people are at the edge of lake facing the water and the city skyline
A little girl rides in a child 's swing.	little girl rides in child swing
Two boys posing in blue shirts and khaki shorts.	two boys posing in blue shirts and khaki shorts

Fig.3

MSCOCO 2014 dataset has training images in train2014 and validation dataset in val2014.The training dataset are fed into the model and evaluation of performance is done on the results obtained for the validation dataset. Images and their features are fed into the model.

Text Encoding of the captions

The captions are processed by converting each word into vector as is done in all deep learning projects as numeric features are required for computations on the model. Along with the associated image each word is passed to model one at a time. For example, at first, the first word along with the image is passed and mapped to generate the second word. Next, the first two words are passed along with image and third word is generated. This process is repeated for all the words in the caption and for all the captions.

Input Sequence	Next word
firstword	brown
firstword, brown	dog
firstword, brown, dog	plays
firstword, brown, dog, plays	with
firstword, brown, dog, plays, with	the
firstword, brown, dog, plays, with, the	hose
firstword, brown, dog, plays, with, the, hose	lastword

Fig.4.Caption Processing

To the model, two arrays are passed : One array for image features and the other for encoded text data of the captions. The output is the next word in encoded format.

The tag “firstword” symbolizes the start of the caption and “lastword” symbolizes the ending of caption.

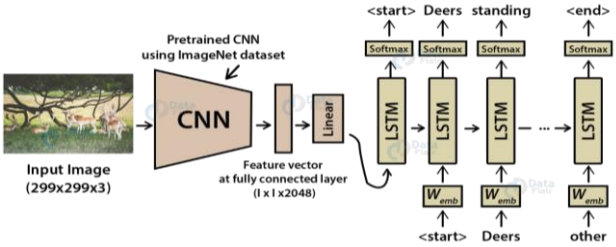
Prediction of words

The encoded words into numbers when passed to a word embedding layer, along with features of the image gives probability of 0 or 1. This layer will cluster words of similar context in the vocabulary. The correct next word is set to 1 while other words are set to 0 which corresponds to inputted image.

Architecture :

Adopted encoder-decoder architecture for image captions and captions are compared using Jaccard similarity.Encoder is convolutional-neural network and decoder is long-short term memory.

Here, RNN is used for encoding text data, and do not include any image features. The multimodal layer which comes after this is fed with both image and text features. This has enabled to feed preprocessed text instead of the raw one.



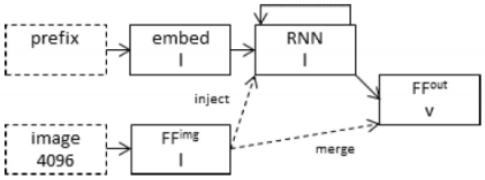


Fig 5. Inner architecture

Basically, three major divisions:

1. Feature extractor from images
2. Processor of text
3. Predicting output

Feature Extractor from Image

We have trained 5 models with this encoder-decoder architecture on COCO train2014 and the system can recommend images from train2014, val2014 and scrapped image dataset from google as well. The system can recommend images with image as the input and as well only captions as the input

Model 1:

Encoder – ResNet 152

Decoder – single-layer LSTM with 512 dimensional hidden states

Each word represented using 256 dimensional learnable embedding.

Model 2:

Encoder – Resnet 101

Decoder – single layer LSTM with 512 dimensional hidden states.

Each word dimension same as Model 1

Model 3 :

Encoder- Resnet 152

Decoder – single layer LSTM with 256-dimensional hidden states.

Each word is represented using 128 dimensional learnable.

Model 4:

Encoder – ResNet 101

Decoder – double-layer LSTM with 512 dimensional hidden states

Each word represented using 256 dimensional learnable embedding.

Model 5:

Encoder – Resnet 101

Decoder – single layer LSTM with 512 dimensional hidden states for both layers.

Each word represented with 300 dimensional Glove vectors kept fixed throughout training.

Processor of Text

A word embedding layer is used to encode text of the captions. 256 of memory units is added to LSTM with RNN and output is also of length 256. As in image feature vector, 50% dropout rate is used in order to avoid overfitting of the model.

Predicting Output

The outputted vector from both images and captions are of same length. These vectors are merged in decoder by using an addition operation. This is then passed to two layers of which the first layer is of 256 length, second layer predicts the most probable word by using softmax layer.

Image Retrieval

The images from the validation set is sent as query. The above model is applied on it generate captions. Then using Jaccard similarity as the metric, the generated captions of the inputted image are compared with the captions in the dataset and based on that similarity score, most similar images to the query are retrieved.

Training phase

The encoder extracts feature vectors from a given input image. Then feature vector is transformed to have same dimension as the input dimension of LSTM network.The source and target sequence are already known, so LSTM is trained as language model using these sequences and feature vector.

Testing Phase

Encoder part remains same, but only difference is batchnorm layer here uses moving average and variance instead of mini-batch statistics.Also, in test phase LSTM decoder cannot see the image description, so LSTM feeds back the previously generated word to next input.

V. Results and Conclusion

Image and natural language respectively are given as input query.

Image as a query :
We randomly choose an images from our proposed dataset as queries.For which the captions will be generated. Using Jaccard similarity as the metrics, the similarities between the captions were calculated and based on that the all the 5 models retrieves top 4 images as shown below. The images are recommended from MSCOCO training and validation dataset and also from the dataset which scraped from google.The system can also retrieve similar images from Google images.

Fig 6.Query Image

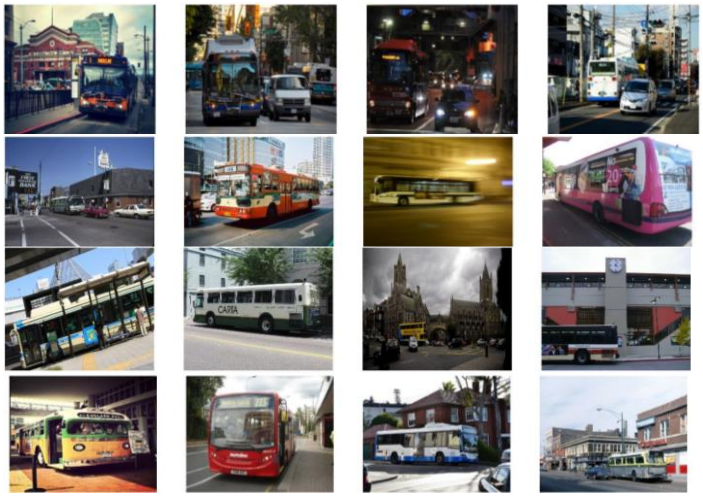


Fig.7 Images retrieved

Natural language as query :
Our method also could be used for searching images by natural language descriptions. The process of experiment is almost the same as above, instead of using image query, human generated text is passed as query to retrieve relevant images.

Fig.8 Image Retrieved when description is given as a query

VI. Conclusion

We propose a multimodal Recurrent Neural Network (m-RNN) framework that performs at the state-of-the-art in three tasks: caption generation, image retrieval given an image and image retrieval given query sentence.Our implementation uses elastic search for indexing of the available images in the server and intermediate captioning mechanism for both search and retrieval process. The system is capable of interpreting the user query and automatically extracting the semantic feature that can make the retrieval more effective and accurate.The image captioning has been carried out using Resnet (Convolutional Neural Network).The retrieved image quality demonstrated promising performance and suggests that an intermediate captioning-based image search could be an alternative to metadata-based search engines.

VII. FUTURE WORK

Due to the exponential growth of digital images in the recent days,image retrieval has become an important problem of discussion.The model when given an image as query,generates captions using the captioning models.Considering the Jaccard's coefficient/similarity score the similar images are retrieved. Even though the proposed image retrieval system has given importance to

semantic concept however,the retrieval by abstract attributes was still not satisfied to human perception. Therefore, there is a need to provide maximum support towards bridging the semantic gap between low level visual features and high level concepts for better image understanding between human and machine and also contribute to have more intelligent, user friendly besides accuracy and efficiency image retrieval that confirm to human perception without involving human interference

REFERENCES

[1] Recurrent Neural Network for Content Based Image Retrieval Using Image Captioning Model-S Sindhu,R Kousalya

[2] Image Retrieval by Dense Caption Reasoning by *Xinru Wei, Yonggang Qi, Jun Liu, Fang Liu* J.

[3] 3 Dense Cap: Fully Convolutional Localization Networks for Dense Captioning by Justin Johnson , Andrej Karpathy , Li Fei-Fei

[4] Deep Visual-Semantic Alignments for Generating Image Descriptions By Andrej Karpathy , Li Fei-Fei

[5] 3.7 Image Captoning by Self-Retrieval with Partially labelled data Xihui Liu, Hongsheng Li, Jing, Shao, Dapeng Chen, Xiaogang Wang

[6]] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in CVPR 2015. IEEE Computer Society Conference on, 2015, pp. 2625–2634.