Image Captioning

Shyamkrishna Menon, Siddharth Nair, Atharva Ranade, Omkar Thavai

Guide- Prof. Sofiya Mujawar

Abstract

This project aims at generating automated captions by learning the contents of the image. At present images are annotated with human intervention and it becomes nearly impossible task for huge commercial databases. The image database is given as input to a deep neural network (Convolutional Neural Network (CNN)) encoder for generating "thought vector" which extracts the features and nuances out of our image and RNN (Recurrent Neural Network) decoder is used to translate the features and objects given by our image to obtain sequential, meaningful description of the image. In this project, we systematically analyse different deep neural network-based image caption generation approaches and pretrained models to conclude on the most efficient model with fine-tuning.

Keywords: image caption, image processing, convolution neural network, CNN, long short term model, LSTM, recurrent neural network, RNN, VGG16, embedding, Flickr30k

Introduction

A large amount of information is stored in an image. Everyday huge image data is generated on social media and observatories. Deep learning can be used to automatically annotate these images, thus replacing the manual annotations done. This will greatly reduce the human error as well as the efforts by removing the need for human intervention The generation of captions from images has various practical benefits, ranging from aiding the visually impaired, to enabling the automatic, cost-saving labelling of the millions of images uploaded to the Internet every day, recommendations in editing applications, beneficial in virtual assistants, for indexing of images, for visually challenged people, for social media, and several other natural language processing applications. The field brings together state-of-the-art models in Natural Language Processing and Computer Vision, two of the major fields in Artificial Intelligence. One of the challenges is availability of large number of images with their associated text ever expanding internet.

Generating captions automatically from images is a complex task as it entails the model to extract features from the images and then form a meaningful sentence from the available features. Basically, the feature extraction is done by training a CNN (Convolutional Neural Network) with huge number of images and the correct weights are identified by multiple forward and backward iterations. With the help of RNN (Recurrent Neural Network) and the extracted features, a sentence is generated.

Literature Survey

1. Deep Learning based Automatic Image Caption Generation

The aim of the paper[1] is to generate captions to the image which is normally, manually annotated by data annotators. It first creates feature vectors with the help of CNN and later uses RNN for creation of sentences with the help of features gained before. For the purpose of automated captioning, a pre-trained model called VGG16 model is being used. This model makes use of a recurrent neural network which encodes the variable length input into a fixed dimensional vector and uses this representation to "decode" it to the desired output sentence. An encoder is a process of extracting vectors which describe contents of an image. A decoder reverses the process of encoding. Decoder process uses layers like tokenizer, embedding, GRU and dense layer. The paper also points few previous works done on image captioning. The paper[1] uses 2 approaches for obtaining image captioning with the same dataset i.e. MS-COCO, one without using Attention Model and one using Attention Model. Finally, the paper concludes with important points like different epochs used for different models, deeper network constitutes to easier image captioning, etc.

2. Image Annotation via deep neural network

This paper[2] proposes a novel framework of multimodal deep learning where the convolutional neural networks (CNN) with unlabeled data is utilized to pre-train the multimodal deep neural network to learn intermediate representations and provide a good initialization for the network then use backpropagation to optimize the distance metric functions on individual modality.

3. Automatic image annotation using DL representation

In this paper[3], the last layer of CaffeNet of the CNN based model is replaced with a projection layer to perform regression and the resulting network is trained for mapping images to semantically meaningful word embedding vectors. Advantage of this modelling is: firstly, it does not require dozens of handcrafted features and secondly, the approach is simpler to formulate than any other generative or discriminative models.

4. Show and Tell: A Neural Image Caption Generator

This paper[4] proposes a network of the same name. In this network, deep convolutional network is used for image classification and sentence generation is done by a powerful Recurrent Neural Network which is trained with the visual input so that RNN can keep track of the objects explained by the text.

5. An Empirical Study of Language CNN for Image Captioning

In this paper[5], we introduce a language CNN model which is suitable for statistical language modelling tasks and shows competitive performance in image captioning.

The primary contribution lies in incorporating a language CNN, which is capable of capturing long-range dependencies in sequences, with RNNs for image captioning. Our model yields comparable performance with the state-of-the-art approaches on Flickr30k and MS COCO which validate the proposal and analysis of the experiments conducted. Performance improvements are clearly observed when compared with other image captioning methods.

6. Image Captioning - A Deep Learning Approach

This paper[6] proposes a hybrid system employing the use of multilayer Convolutional Neural Network (CNN) to generate vocabulary describing the images and a Long Short Term Memory (LSTM) to accurately structure meaningful sentences using the generated keywords. The efficiency of the proposed model is checked using Flickr30k and Flickr8k datasets and also by utilising Bleu metric gives superior results in comparison to other state-of-the-art models.

Experiment set up

We use Python as our programming language as it is a popular language when it comes to using deep learning approaches and image processing. We use Deep learning for training the model using Convolutional Neural Networks and Recurrent Neural Networks (deep learning model) to detect features from image and predict the captions respectively. There are few python libraries that we will be using. We use pandas for data manipulation and analysis, opencv for loading images, numpy for mathematical operations, Keras Framework(Using Tenserflow Backend) is used for building our model architecture for Image Captioning and also used for importing VGG-16 for Transfer Learning. All these are implemented in Jupyter Notebook enabling Python 3 langauage.

Proposed system

1. Explanation of proposed system
    a. First we will import Flickr30k[7] dataset and process. Flicker datasets are used for image captioning. 30k stands for 30,000 images of various instances.
    b. We use VGG16 model for image captioning. VGG16 is used for embedding of features within the image like identifying a person, thing, etc and LSTM is used for encapsulating all features and describing it as a sentence.
2. Block diagram- Fig 1 depicts the block diagram

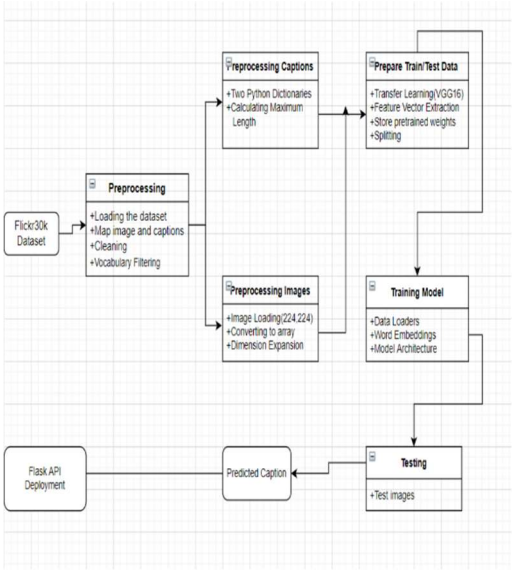

Fig 1: Block diagram

3. Model algorithm diagram- Fig 2 depicts Model algorithm diagram

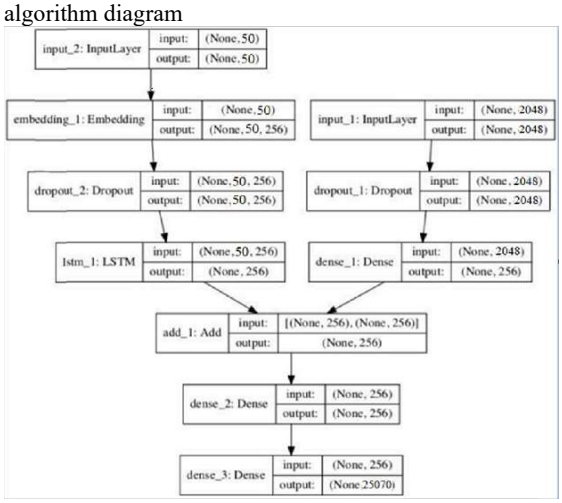

Fig 2: Model algorithm diagram

4. Pseudocode
    a. Read dataset- read the Flickr30k dataset.
    b. Processing of data- firstly, create a dictionary of imageID and descriptions, then create a vocabulary and finally filter out words which are more frequent.
    c. Transfer learning- use VGG16 for getting vectors for every image.
    d. Word embeddings- preprocess captions and put it into a fixed length using glove.
    e. Training- combine image and caption as input (obtained from step c and d) and train the model.
    f. Testing- obtain caption from learned weights during training of model.
5. Expected output
    a. Expected output would be apt sentence formation of given input image.

b. With the help of different features learnt, the model will provide words relevant to the image.

References

[1] Shahar Banu , Seemakousar B , Sanchita S M , Nivedita A, Arun Joshi, Rajeshwari S.G, 2021, Deep Learning based Automatic Image Caption Generation, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 10, Issue 08 (August 2021)

[2] S. Chengjian, S. Zhu and Z. Shi, "Image annotation via deep neural network," 2015 14th IAPR International Conference on Machine Vision Applications (MVA), 2015, pp. 518-521, doi: 10.1109/MVA.2015.7153244

[3] Venkatesh N.Murty et al, Automatic image annotation using DL representation, ICMR '15: Proceedings of the 5th ACM on International Conference on Multimedia RetrievalJune 2015 Pages 603–606,

[4] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, "Show and tell: A neural image caption generator," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3156-3164, doi: 10.1109/CVPR.2015.7298935.

[5] J. Gu, G. Wang, J. Cai and T. Chen, "An Empirical Study of Language CNN for Image Captioning," 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1231-1240, doi: 10.1109/ICCV.2017.138.

[6] Srinivasan, Lakshminarasimhan and Dinesh Sreekanthan. “Image Captioning-A Deep Learning Approach.” (2018).

[7] Kaggle dataset for flickr30k- https://www.kaggle.com/adityajn105/flickr30k?select=Images