JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Automatic Image Captioning

[1]Nihal Kazi, [2]Prof.Priyanka Mane, [3]Anand Dhakane, [4]Dattatray Jadhav, [5]Omkar Mangnale

[1] Student, Dept. of IT, Genba Sopanrao Moze College of Engineering, Pune.

[2] Professor, Dept. of IT, Genba Sopanrao Moze College of Engineering, Pune. [3]Student, Dept. of IT, Genba Sopanrao Moze College of Engineering, Pune. [4]Student, Dept. of IT, Genba Sopanrao Moze College of Engineering, Pune. [5]Student, Dept. of IT, Genba Sopanrao Moze College of Engineering, Pune.

Abstract - The paper 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 paper, we systematically analyze different deep neural network-based image caption generation approaches and pretrained models to conclude on the most efficient model with fine-tuning. The analyzed models contain both with and without 'attention' concept to optimize the caption generating ability ofthe model. All the models are trained on the same dataset for concrete comparison.

Keywords - Automated captions, deep neural network, CNN, RNN, feature extraction, attention.

I. 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. However, most of this data is noisy and hence it cannot be directly used in image captioning model. For training an image caption generation model, a huge dataset with properly available annotated image is required. In this paper, we plan to demonstrate a system that generates contextual description about objects in images. Given an image, break it down to extract the different objects, actions, attributes and generate a meaningful sentence (caption/description) for the image.

II. LITERATURE SURVEY

One of the most striking notice is the Image Net project, where they publicly supported huge number of named pictures and prepared models for the last ten years to perceive objects in the picture. Beginning around 2010, the yearly Image Net Large Scale Visual Recognition Challenge (ILSCRC) holds a contention consistently, to vie for most elevated precision on different visual acknowledgment undertakings. Presently the profound CNN networks have more exactness than people in acknowledgment. Any way Captioning pictures could be a lot testing task, since it includes object acknowledgment and tracking down connections among them. This has been unthinkable as of not long ago, attributable to gigantic improvement in computational power. Despite the fact that there are different scientists taking care of on a similar issue, there are two groups that stood apart with their calculations. One from Google, and the other from Stanford University. Google delivered a paper "Sharing time: A Neural

Image Caption Generator" in 2014. Their model is prepared to expand the probability of the objective portrayal sentence, given the picture. The model is prepared on different datasets like Flickr30K, SBU, and MSCOCO and has accomplished human level execution in creating subtitles. At the point when google originally delivered a paper in 2014, the framework utilized the "Commencement V1" picture characterization model which accomplished 89.6% exactness. The most recent delivery in 2016 utilized "Commencement V3" model, which accomplishes 93.9% precision. Before Google, picture subtitling was conceivable utilizing Disbelief software system.
Then Google delivered TensorFlow execution, which utilizes GPU power and contrasted with before executions, the preparation time is decreased by a variable of 4.

III. PROPOSED MODEL

A) Block Diagram of Proposed model:

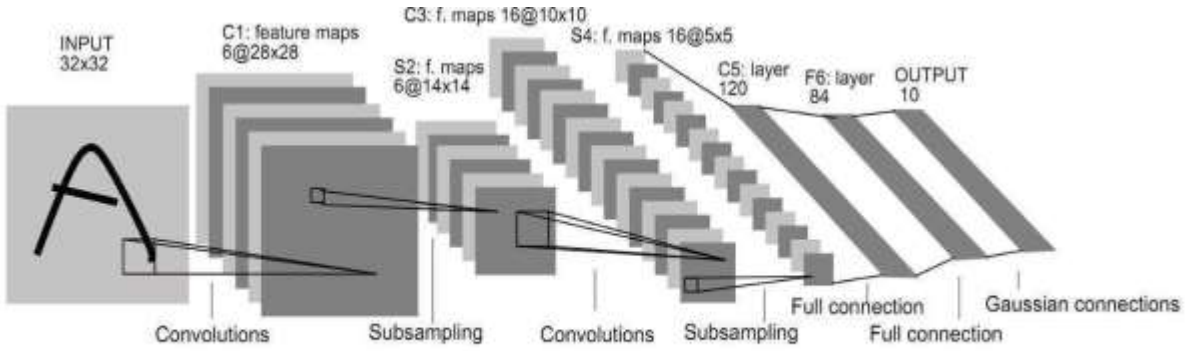


Fig. 2. Architecture of LeNet-5, a Convolutional Neural Network, here for digits recognition. Each plane is a feature map, i.e. a set of units whose weights are constrained to be identical.

Fig.1: An example of CNN architecture

- Convolutional Neural Networks have 2 main components.

  ● **Feature learning**: We have convolution, ReLU, Pooling layer stages here.Edges, shades, lines, curves, in this Feature learning step are get extricated.
  ● **Classification**: There is Fully Connected (FC) layer in this stage. They will relegate alikelihood for the item on the picture being what the calculation predicts it is.

  For our use-case, we are only interested in Feature learning component of CNN.

Feature learning:

ReLU(Rectified Linear Unit):

An additional an activity known as ReLU has been utilized after every single Convolution activity. ReLu is a non-direct activity. ReLU is an issue savvy activity (applied per pixel) and replaces all awful pixel values in the limit map by using zero. The justification for ReLU is to introduce non-linearity in our ConvNet, when you consider that the greater part of this present reality measurements we would favor our ConvNet to look at would be non-straight.

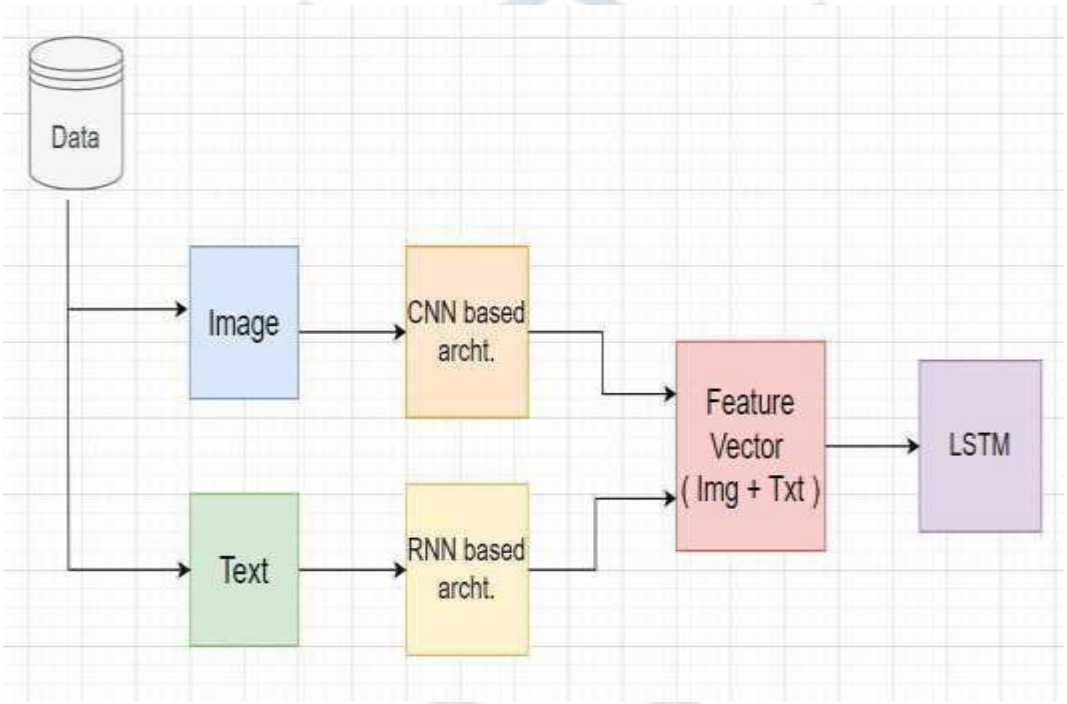
Pooling layer: In this segment the dimensionality of conv layer or trademark map gets diminished protecting the imperative data. From time to time this spatial pooling is furthermore known as down sampling or subsampling. This pooling layers could likewise be Max pooling, Avg pooling, total pooling. Frequently we see Max pooling is utilized most.

B) Hardware and Software Requirement :

Hardware Interfaces:
- 16 GB RAM
- NVIDIA GPU
- M1 core
- 50 GB physical storage

Software Requirements
- MACOS (X)
- Python3
- CUDA 9
- Tensorflow
- Keras
- Numpy
- Matplotlib

C) Flowchart:



IV. Result And Analysis

The Web-App is deployed the models on a local server and ran it using Flask, results of models is shown in below :-Flow of Web :

Fig: Home Page

The generated sentence are shown in Fig : Generated sentences are " several people are standing around in a fish fountain while actual humans read as " several people are standing around fish fountain and watching them".

Fig: shows image of captions generated for the image

V. Conclusion and Future Scope Reference

Image captioning is nonetheless a creating subject and many researches are nonetheless in progress. Recent work primarily based on deep studying methods has resulted in a leap forward in the accuracy of photograph captioning as it have breakdown the complicated fashions to easy structure. The textual content description of the picture can enhance the content-based photograph retrieval efficiency, the increasing utility scope of visible grasp in the fields of medicine, security, navy and different fields, which has a vast utility prospect. At the equal time, the hypothetical system and query procedures of photo inscribing can advance the improvement of the thought and utility of photograph comment and apparent question addressing (VQA), go media recovery, video subtitling and video exchange, which has vital instructive and reasonable programming esteem.

We have presented an end-to-end neural network system that can automatically view an image and generate a reasonable description in plain English. It is based on a convolution neural network that encodes an image into a compact representation, followed by a recurrent neural network that generates a corresponding sentence. The model is trained to maximize the likelihood of the sentence given theimage. We also saw the effect of the encoder-decoder approach combined with attention and made analysis.

VI. Reference

[1] Shuang Liu, Liang Bai,a, Yanli Hu and Haoran Wang. Image Captioning Based on Deep Neural Networks. EITCE (2018)
Available : Link
[2] Lakshminarasimhan Srinivasan, Dinesh Sreekanthan,Amutha A.L. I2T: Image Captioning - A Deep Learning Approach (2018).
Available : Link
[3] Simao Herdade, Armin Kappeler, Kofi Boakye, Joao Soares.Image Captioning: Transforming Objects into Words . San Francisco, CA (2019).
Available : Link
[4] Aishwarya Maroju ,Sneha Sri Doma ,Lahari Chandarlapati , Image Caption Generating Deep Learning Model ,J.N.T.U, Hyderabad , Sreenidhi Institute of Science And Technology (2021).
Available : Link
[5] Zhongliang Yang, Yu-Jin Zhang, Sadaqat ur Rehman, Yongfeng Huang: Image Captioning with Object Detection and Localization, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
Available : Link
[6] Oriol Vinyals , Alexander Toshev , Samy Bengio Dumitu Erhan(2014). Show and Tell : A Neural Image Caption Generator . Google
Available : Link