G. B. Pant Engineering College, New Delhi

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

SYNOPSIS
OF
IMAGE CAPTION GENERATOR

B. Tech (CSE - 7th Semester)

SUBMITTED BY: -
Priya Chawla(41420902717)
Pritika Rana(02420902717)
Mayank Singh(00420907218)

SUBMITTED TO: -
Dr. Sunita Tiwari

Statement of the problem:-

To use machine program for providing description to image provided by user, by extracting information from image using CNN and LSTM and API.

Motivation :-

- Convolution Neural Networks and Long Short Term Memory (RNN) are the best techniques available today. Through this project we will try to explore the possibilities deep learning offer, and implement it to make computer system describe an image.
- Easing the human work of exploring through a large chunk of images for finding a particular image. Also, it can provide description to image, easing human labour of naming them. Because of their description it can be used to categories the images too.
- By Further Development, it can help visually impaired peoples by giving them audio description.
- The strings generated can be used for social media sharing, for determining the actions, locations, sentiments etc.
- Automatic description generationfor video frames would help security authorities manage more efficiently andutilize large volumes of monitoring data.
- Image search engines could potentially benefitfrom image description in supporting more accurate and targeted queries for endusers.

Overview of the project :-

- Image Caption generator is a task that involves computer vision and natural language processing concepts to recognize the context of an image and describe them in natural language like English.

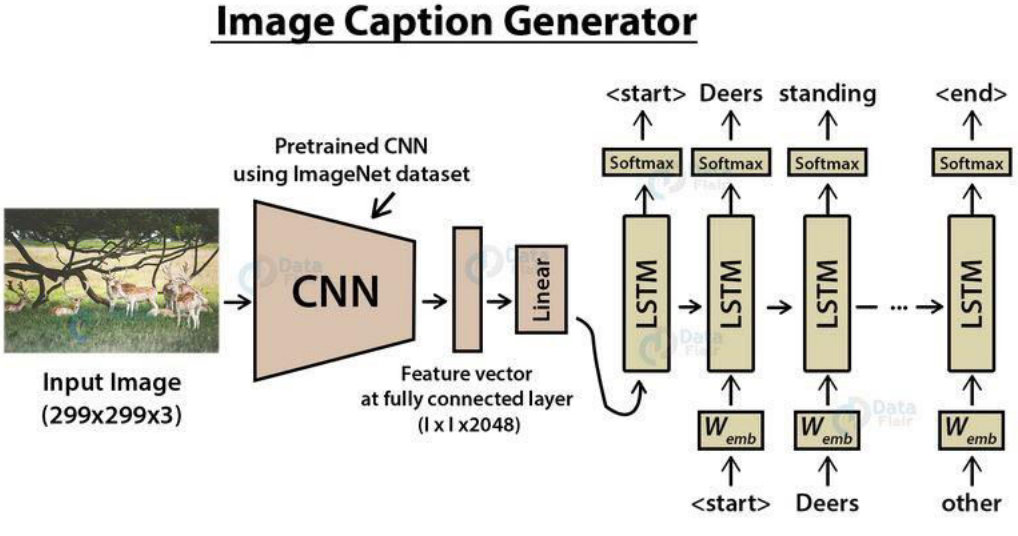

Figure:- Flow diagram of the image caption generation process Credits:- DataFlair

Objective and Scope of the project:-

- To provide a description (caption) for the provided image by implementing the revolutionary feature of CNN and LSTM along with API for better caption results.

- To create an interface where a user can provide one, or multiple images and the program will provide the description for the images.

Literature Survey :-

1) Automatic Caption Generation for News images - YansongFeng[3]

Data resources where images and their textual description co-occur naturally are used for generating captions for the images.

- Lavrenko et al.,2003 existing generative model is used to learn visual and textual correspondence under noisy conditions of image and text. Using Latent Dirichlet Allocation (Blei and Jordan, 2003),visual andtextual modalities are representedjointly as a probability distribution over a set of topics.
- The availability of news documents in our dataset allows to perform the captiongeneration task in a fashion akin to text summarization; save one important differencethat our model is not solely based on text but uses the image in order to select contentfrom the document that should be present in the caption. The backbone for both approaches is topic-based image annotationmodel. Extractive models examine how to best select sentences that overlap incontent with image annotation model. It is an existing abstractive headlinegeneration model incorporating visual information. Here model operates over image description keywords and document phrases by taking dependencyand word order constraints into account.

2) Tag Clustering algorithm LMMSK - Jing Yang and Jun Wang[4]

- Through extensive testing and comparison on random and personalized user tags, it is found the LMMSK method is the best among the traditional K-means clustering algorithm, MMSK-means clustering algorithm combined with latent semantic analysis (LSA). The dataset used in the project is obtained from DeliciousSocial Bookmarking System from 2004 to 2009.

3) Web 2.0 and Folksonomy - MohmedhanifNashipudi[5]

- The emergence of web 2.0 technologieshas given rise to Tagging - adding keywords to content in order to categorize it. This is similar to subject indexing but without a controlled vocabulary.
- Folksonomy is the result of personal free tagging of information and objects (anything with a URL) for one's own retrieval. The tagging is done in a social environment (usually shared and open to others). Folksonomy is created from the act of tagging by the person consuming the information.
- Some of the Folksonomy-based systems are:
  - Del.icio.us (www.del.icio.us)
  - CiteULike (www.citeulike.org)
  - Connotea (www.connotea.org)
  - Flickr (www.flickr.com)
  - Furl (www.furl.net)

- Folksonomy represents some of the good and bad aspects organization of content. Its uncontrolled nature is disorganized, suffers from problems of imprecision, ambiguity, etc when compared to a well-developed controlled vocabulary. Conversely, systems employing free-form tagging encourages users to organize content in their own ways. These systems are highly responsive to user needs and vocabularies.

4) Neural Caption Generation for news images - Vishwash Batra, Yulan He, George Vogiatis

- It proposed a methodology for automatically generating captions for news paper articles consisting of a text paragraph and an image. It uses deep neural network architectures built upon Recurrent Neural Networks. Results on a BBC News dataset show that our proposed

approach outperforms a traditional method based on Latent Dirichlet Allocation using both automatic evaluation based on BLEU scores and human evaluation.

5) Automatic Caption Generation for News Images - Priyanka Jadhav, Sayali Joag, Rohini Chaure, Sarika Koli

- First stage consists of content selection which identifies what the image and accompanying article are about, whereas second stage surface realization determines how to put the chosen content in a proper grammatical caption.
- For content selection, the project is using probabilistic image annotation model that suggests keywords for an image.

Methodology:-

Our projects concern with providing a meaningful description to the provided image in a natural language, here English. For this we are going to implement two broad methodologies: -

1) Recognizing the image and its details by using Convolution Neural Network technique.
2) Then our trainedmodel will provide a meaningful description to the image by using Long Short Term Memory technique of RNN.
3) The description string is then passed to out API which will suggest few caption options to the user for selecting there preferred one.

Hardware and Software Requirement:-

Software Interface :-

1) Python - Jupyterlab
2) Libraries :- tensorflow, keras, pillow, numpy, tqdm
3) Dataset :- Flickr8k[1]
4) Suggestion API

Hardware Interface :- NA

References:-

1) Dataset:-

a) https://github.com/jbrownlee/Datasets/releases/download/Flickr8k/Flickr8k_text.zip
b) https://github.com/jbrownlee/Datasets/releases/download/Flickr8k/Flickr8k_Dataset.zip

2) Books:-

a) Foundations of Statistical Natural Language Processing - By Christopher Manning, Hinrich Schutze
b) Natural Language Processing with Python - By Steven Bird, Ewan Klein, and Edward Loper

3) Yansong Feng - Automatic Caption Generation for News Images - University of Edinburgh - [Feng 2011]
4) Jing Yang and Jun Wang - Tag clustering algorithm LMMSK: improved K-means algorithmbased on latent semantic analysis - Journal of Systems Engineering and Electronics -Vol. 28, No. 2, April 2017, PG.374 – 384 - School of Economics and Management, Beihang University, Beijing 100191, China - [Yang&Wang 2017]
5) http://www.ijodls.in/uploads/3/6/0/3/3603729/3_mohmedhanif__29-35_.pdf