



St. Francis Institute of Technology
Department of Computer Engineering
Mini Project – Sem VI

Automatic Image Captioning

Guide:

Mrs. Safa Hamdare

Asst. Professor

Group Members

Name of Student	Class-Roll No.
Aishwarya Sreenivasan	TE CMPN A-72
Ankit Jaiswal	TE CMPN A-73
Richie Jacob	TE CMPN A-74

Content

- Introduction
- Literature
- Problem Statement
 - Proposed Solution
- (Workflow) of the system(Block Diagram)
- Algorithm with Implementation details (if any)
- Data set description (if any)
- Working demo along with validation (GUI Design, tests etc.)
- Results/Analysis



Introduction



Tap Tap
See
For Android & iOS

- **Specify the need of the system**
 - **Societal Impact**
 - **Social Media.** *Platforms like facebook can infer directly from the image, where you are (beach, cafe etc), what you wear (color) and more importantly what you're doing also (in a way).*
 - **Identifying and Locating people** *from live videos or cctv footages.*
 - **Applications like TapTapSee and Seeing AI** *are both camera-based programmes, which identify objects using artificial intelligence. It can be used to describe images to people who are blind or have low vision and who rely on sounds and texts to describe a scene.*



Introduction

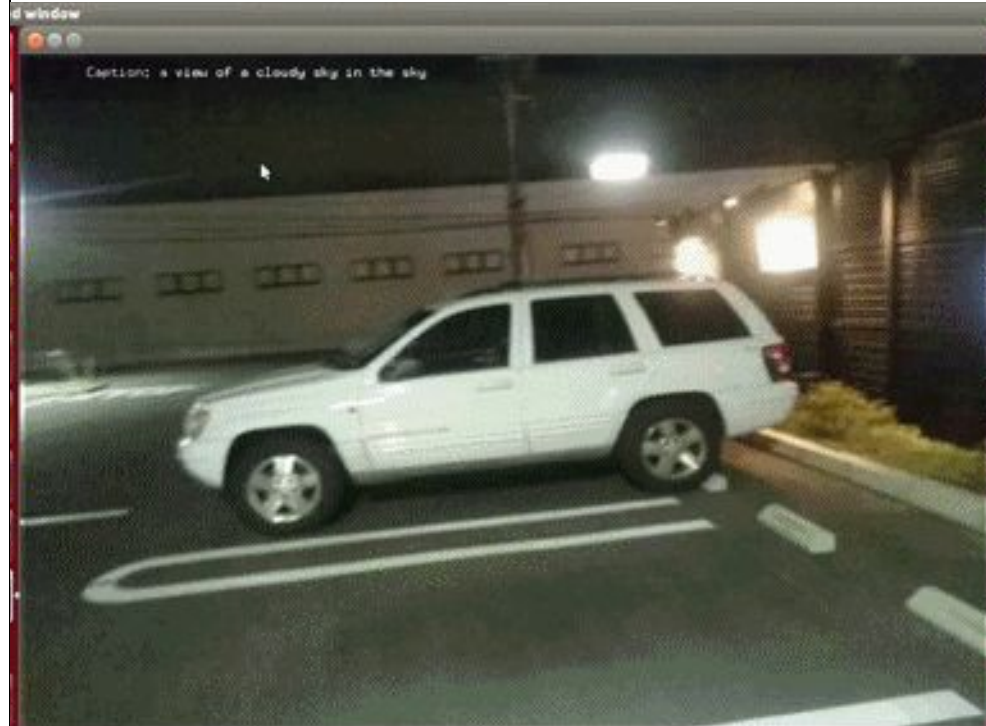
– Application

– **Sorting images based on content.**

– In **web development**, it's good practice to provide a description for any image that appears on the page so that an image can be read or heard as opposed to just seen. This makes web content accessible.

– Caption Bot by Microsoft

– Video captioning



Caption: a view of a cloudy sky in the sky

(This gif is for presentation purpose only. It has no resemblance to the project under consideration.)



Introduction

- Research

- Convolutional Neural Network (CNN), which is used to train the images as well as to detect the objects in the image with the help of various pre-trained model like VGG.
- The second neural network used is Recurrent Neural Network (RNN) based Long Short Term Memory (LSTM), which is used to generate captions from the generated object keywords.
- GPU based computing is required to perform the Deep Learning tasks more effectively.
- The dataset used for this model like Flickr 8k (containing 8k images).



Literature

- **Work done before**

- One of the early non-neural approaches on describing images was done by Farhadi et al who proposed a method based on multi-label Markov random field. The proposed approach works in two phases — mapping the image to a meaning space in the format of <object, action, scene>, and mapping the meaning space to a sentence using some predefined templates.
- Vinyals et al. proposed Neural Image Caption (NIC) — a generative model based on deep recurrent architecture that maximizes the likelihood of generating the target caption given an input image.

- **Objective**

- Take image as an input, process uploaded image using Convolutional Neural Network (CNN) for salient feature detection and on top of that a Recurrent Neural Network (RNN) that generates sequential words to construct image captions. This caption is displayed as the output to the user.



Literature

- **Scope**

- Our model will be able to produce grammatically and logically correct sentences for the image uploaded by the user. This can be used to automatically caption images uploaded on webpages or social media sites instead of manually typing them.
- This project can further be extended to do tasks such as live captioning or video captioning as well as it would reduce the tedious task to manually typing out the description.
- This project can be extended to predict Sign Languages by making use of appropriate dataset.
- Model can further trained to read banners from images as well.

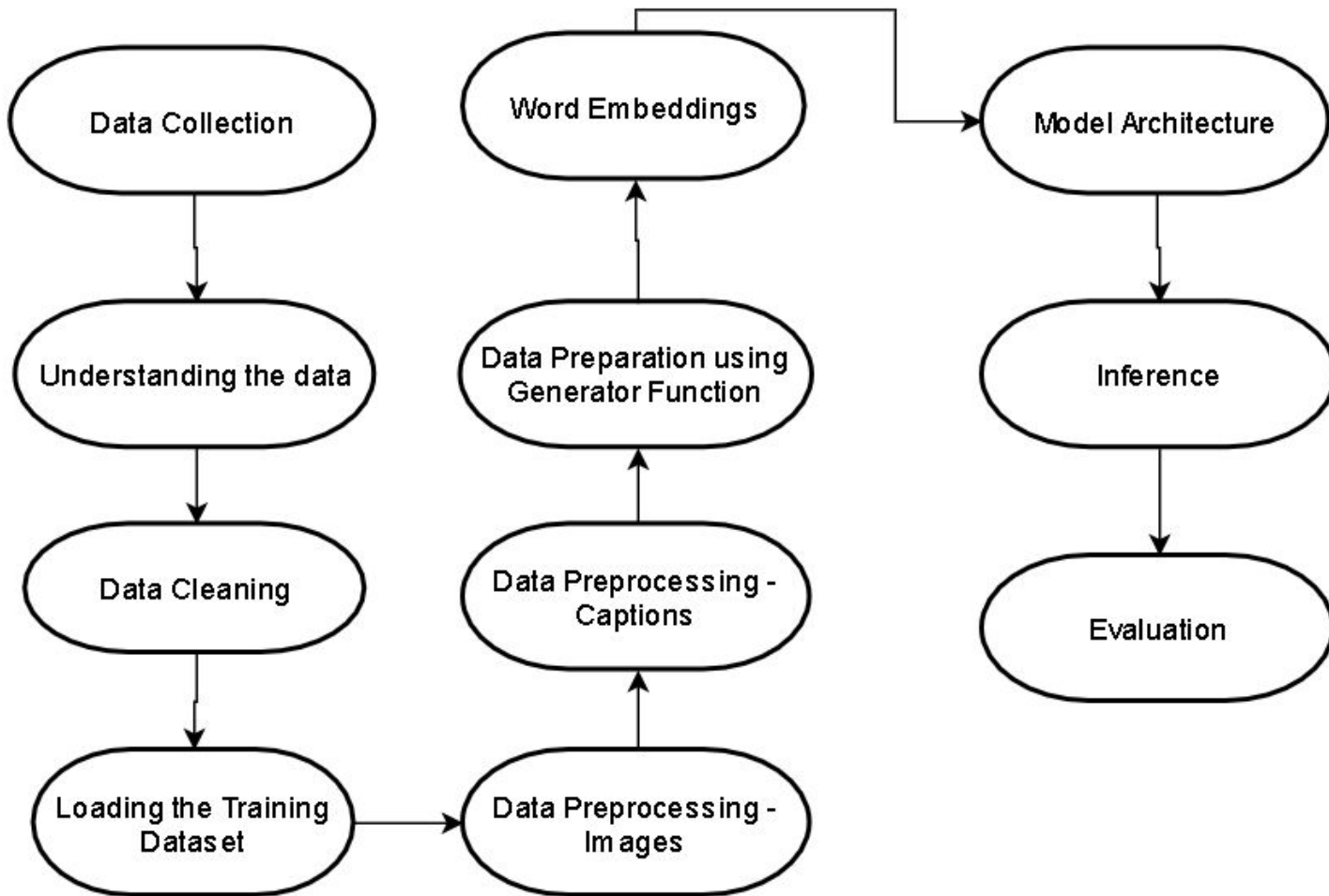


Problem Statement

- The problem introduces a captioning task, which requires a computer vision system to both localize and describe salient regions in images in natural language.
- The image captioning task generalizes object detection where the descriptions consist of a single word.
- Given a set of images and prior knowledge about the content find the correct semantic label for the entire image(s).
- Input: An image.
- Expected Output: Natural language description of the input image.



Work Flow of the system



Description of data set

- **Dataset used: Flickr 8k** (containing 8k images)
- This dataset contains 8000 images each with 5 captions (an image can have multiple captions).
- These images are bifurcated as follows:
 - a) Training Set — 6000 images
 - b) Development Set — 1000 images
 - c) Testing Set — 1000 images
- Didn't use Flickr30k because training a model with large number of images may not be feasible on a system which is not a very high end PC/Laptop.
- **Input to the model:** Image
- **Output given by the model:** A grammatically correct sentence describing the objects, people or background in the uploaded image.

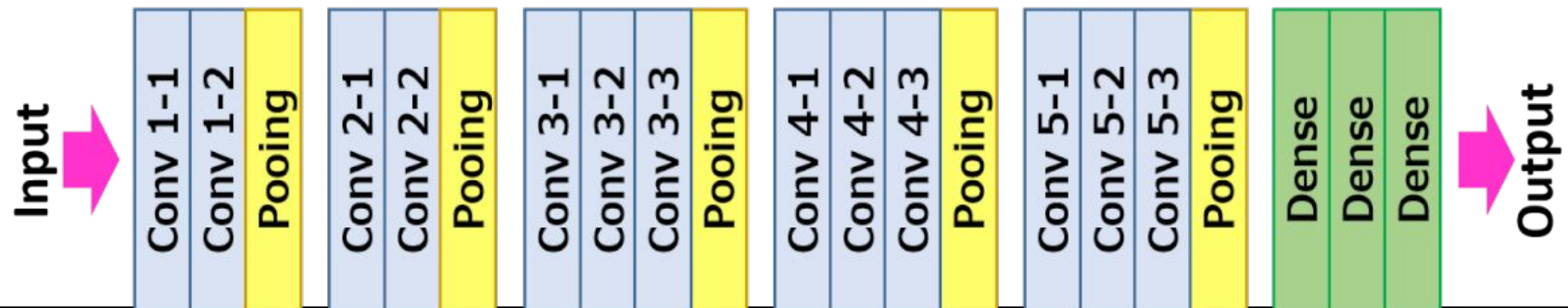


Algorithm used

• CNN: VGG16

- Convolutional Neural networks are specialized deep neural networks which can process the data that has input shape like a 2D matrix. Images are easily represented as a 2D matrix and CNN is very useful in working with images.
- CNN is basically used for image classifications and identifying if an image is a bird, a plane or Superman, etc.
- VGG-16 is a convolutional neural network architecture, it's name VGG-16 comes from the fact that it has 16 layers.

VGG-16



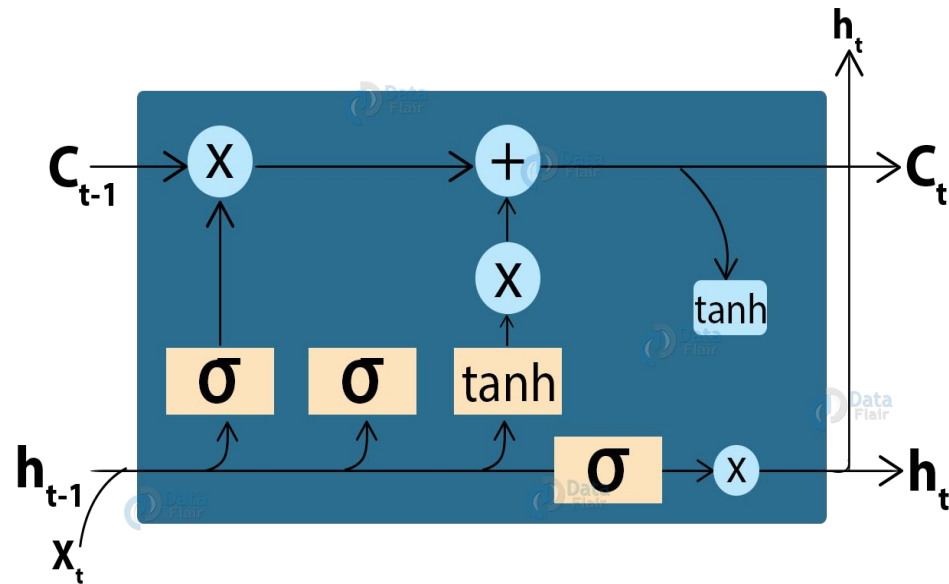
Algorithm used

● RNN: LSTM

- LSTM is a type of RNN.
- Is well suited for sequence
- prediction problems.
- Based on the previous text, we can predict the next word.
- Overcomes the limitations of traditional RNN which had short term memory.
- Can carry out relevant information throughout the processing of inputs and with a forget gate, it discards non-relevant information.



LSTM Cell Structure



Working demo along with validation

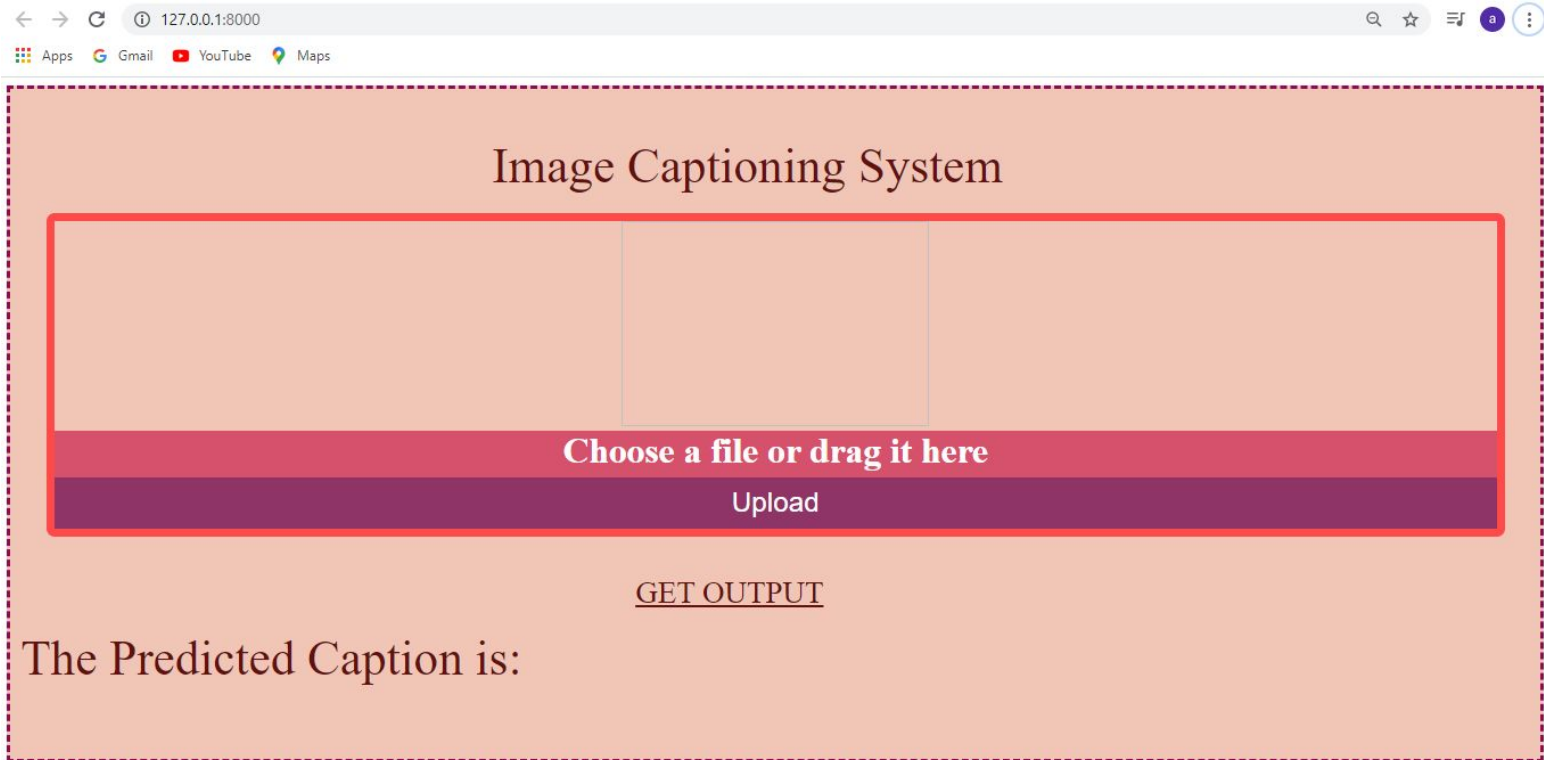


Image Captioning System

Choose a file or drag it here

Upload

GET OUTPUT

The Predicted Caption is:

Before uploading

Working demo along with validation



After Uploading



Predicted Caption



Conclusion

- Image captioning has many applications including helping visually impaired people, image captioning on social media, websites and can be extended to video captioning. This is made possible with the help of pre trained models and powerful deep learning frameworks like Tensorflow and Keras.
- This is a Deep Learning project, which makes use of multiple Neural Networks like CNN and LSTM to detect objects and caption the images. To deploy our model as a web application, we have made use of Django Framework.
- Our present model generates captions only for the image, which itself is a complex task and captioning live video frames is much complex to create.



References

1. A. Karpathy and F. Li. Deep visual-semantic alignments for generating image descriptions. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, pages 3128–3137, 2015.
2. <https://towardsdatascience.com/image-captioning-with-keras-teaching-computers-to-describe-pictures-c88a46a311b8>
3. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(4):652–663, April 2017.
4. <https://data-flair.training/blogs/python-based-project-image-caption-generator-cnn/>

