Project Synopsis

# Image Captioning Based on Deep Neural Networks


**Prepared by**

Mrunal S Jadhav

Aditya G Joshi


**Under the guidance of**

Prof. Sachin Malave

**Bachelor of Engineering (BE)**

**Department of Computer Engineering**

**A. P. Shah Institute of Technology, Thane**

**University of Mumbai**

**2019-2020**

# Image Captioning Based on Deep Neural Networks


## Abstract

The visual world is populated with a vast number of objects, the most appropriate labelling of which is often ambiguous, task specific, or admits multiple equally correct answers. A quick glance is sufficient for a human to understand and describe what is happening in the picture. The quest of connecting computer vision and natural language processing is a long way of touching the holy grail in artificial intelligence. The task is to transform a sentence S written in its source language, into its translation T in the target language, by maximising the probability P(T|S).

A combination of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) which seeks to progress directly from image features to text can be progressed to define a single end-to-end model to maximize the likelihood of the target description sentence, given an image, instead of requiring sophisticated data preparation or a pipeline of specifically designed models. Thus we can develop a generative model, a probabilistic framework, based on deep recurrent architecture that combines advances in computer vision and machine translation to generate natural sentences describing an image.


## 1.Introduction

Image captioning is a task that a machine learns to generate natural language sentences to describe the salient parts of an image. Being able to automatically describe the content of an image using properly formed English sentences is a very challenging task. This task is significantly harder, than the well-studied image classification or object recognition tasks, which have been a main focus in the computer vision community. Indeed, a description must capture not only the objects contained in an image, but it also must express how these objects relate to each other as well as their attributes and the activities they are involved in.

Moreover, the above semantic knowledge has to be expressed in a natural language like English, which means that a language model is needed in addition to visual understanding. Generating complete and natural image descriptions automatically has large potential effects, such as titles attached to news images, descriptions associated with medical images, text-based image retrieval, information accessed for blind users, human-robot interaction. These applications in image captioning have important theoretical and practical research value. Therefore, image captioning is a more complicated but meaningful task in the age of artificial intelligence. The challenge of image captioning is to design a model that can fully use image information to generate more human-like rich image descriptions. The meaningful description generation process of high level image semantics requires not only the understanding of objects or scene recognition in the image, but also the ability to analyse their states, understand the relationship among them and generate a semantically and syntactically correct sentence.

## 2.Literature Review

In this project we aim to incorporate all the best methods in each stage of creating efficient deep learning model for Image captioning. Recently, a great progress in image captioning has been achieved by using semantic concepts detected from the image, which is very similar to the cognition process of humans. Researchers have proposed a multimodal Recurrent Neural Network model that creatively combines the CNN and RNN model to solve the image captioning problem. Because of the gradient disappearance and the limited memory problem of ordinary RNN, the LSTM model is a special type of structure of the RNN model that can solve the above problems.

## 3.Problem Statement

For the image captioning task, humans can easily understand the image content and express it in the form of natural language sentences according to specific needs; however, for computers, it requires the integrated use of image processing, computer vision, natural language processing and other major areas of research results. The challenge of image captioning is to design a model that can fully use image information to generate more human-like rich image descriptions.

## 4.Objectives

To Generate an Encoder-Decoder Network Architecture with :

a. CNNs to produce a rich representation of the input image by embedding it into a fixed-length vector, such that this representation can be used for a variety of vision tasks.
b. RNN network that obtains historical information through continuous circulation of the hidden layer, which has better training capabilities and can perform better than mining deeper linguistic knowledge such as semantics and syntax information implicit in the word sequence

## 5.Scope

Translation work is achieved by using an “encoder” RNN that reads the source sentence and transforms it into a rich fixed-length vector representation, which in turn is used as the hidden state of a “decoder” RNN that generates the target sentence. We propose to follow this elegant receipe by replacing the encoder RNN by a deep CNN which can produce a rich representation of input by embedding it in a fixed-length vector, such that this representation can be used for variety of tasks.

Thus, we develop an CNN model by pretraining it for an image classification task and using the last hidden layer as an input to the RNN decoder which uses the fixed dimensional vector representation to “decode” it to the desired output sentence. We intend to develop a single end to end network to

develop more accurate feature extraction and efficiently generate textual description which can provide detailed information about the given image.

## 6.Benefits for environment

- Generate captions for images which can promote safety and protection of environment.
- Determine various pollutants present in a given image and caption them so as to reduce its generation and manage it.
- Image search tools can help in finding environmental problems and its solutions.

## 7.Benefits for society

- Helps visually impaired to understand the image by converting the captioned text into speech.
- Helps colour blind and other vision problem patients to understand image more effectively.
- Providing more accurate captions for images which can be easily misunderstood.
- Finding hidden features from keywords generated by captions.

## 8.Applications

The web is filled with billions of images, helping to entertain and inform the world on a countless variety of subjects. The existing NLP applications that benefit which extract insights/summary from given text data or an essay etc can be extended to people who would benefit from automated insights from images. While automatic image captioning can help solve this problem, accurate image captioning is a challenging task that requires advancing the state of the art of both computer vision and natural language processing. Elaborate applications in image captioning are:

- Help Visually Impaired:
  Much of the visual information is not accessible to those with visual impairments. A rich automated description of the image would benefit them.
- E -commerce assistant:
  Image captions, manually added by website authors using Alt-text HTML, is one way to make this content more accessible, so that a natural-language description for images that can be presented using text-to-speech systems. However, existing human-curated Alt-text HTML fields are added for only a very small fraction of web images.
- Mapping images to natural language and vice versa would also help in medical image understanding that conveys the clinical physician that the algorithm has found something fishy

in the patient by mapping the physiological parameters and images which may require further investigations.

- Platforms like Facebook can infer directly from the image, where you are ( beach, cafe etc), what you wear (colour) and more importantly what you're doing also (in a way).
- These applications can be extended to explaining what is happening in a video frame by frame.

## 9.Technology stack

Colab : Colaboratory is a Google research project created to help disseminate machine learning education and research. It's a Jupyter notebook environment that requires no setup to use and runs entirely in the cloud.

Pytorch : PyTorch is an open source machine learning library based on the Torch library, used for applications such as computer vision and natural language processing. It is primarily developed by Facebook's artificial intelligence research group

Numpy : NumPy is the fundamental package for scientific computing with Python. It contains among other things:

- a powerful N-dimensional array object
- sophisticated (broadcasting) functions
- tools for integrating C/C++ and Fortran code
- useful linear algebra, Fourier transform, and random number capabilities

Pandas : In computer programming, pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

Keras : Keras is an Open Source Neural Network library written in Python and a high-level API wrapper for the low-level API that runs on top of Theano or Tensorflow. Keras doesn't handle Low-Level API such as making the computational graph, making tensors or other variables because it has been handled by another library called the "Backend". Keras High-Level API handles the way we make models, defining layers, or set up multiple input-output models. In this level, Keras also compiles our model with loss and optimizer functions, training process with fit function.

Sklearn : Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

Tensorflow : TensorFlow is an end-to-end open source platform for machine learning. It has a comprehensive, flexible ecosystem of tools, libraries and community resources that lets researchers push the state-of-the-art in ML and developers easily build and deploy ML powered applications.

**References**

[1] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutionalnetworksforvisualrecognitionanddescription,"in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2015.

[2] K.Cho,B.vanMerrienboer,C.Gulcehre,F.Bougares,H.Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in Proc. Conf. Empirical Methods in Natural Language Processing, 2014.

[3] A. Farhadi "Every picture tells a story: Generating sentences from images," in Proc. 11th Eur. Conf. Comput. Vis.: Part IV, 2010, pp. 15–29.

[4] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge.

[5] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in Int. Conf. Learn. Representations, 2013.

Mrunal S Jadhav

Aditya G Joshi