Automated Image Captioning using Artificial Neural Networks

Design Project Report submitted in partial fulfilment of the requirements for the degree of B.Tech COE

by

Pranjali Ajay Parse
(Roll No: COE16B031)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, DESIGN AND MANUFACTURING, KANCHEEPURAM

November 2019

Certificate

This is to certify that the B.Tech Design Project titled, "**Automated Image Captioning using Artificial Neural Networks**" submitted by **Pranjali Ajay Parse**, with Roll No: **COE16B031** to the Indian Institute of Information Technology Design and Manufacturing, Kancheepuram for the award of "**Bachelor of Technology in Computer Engineering**" is a bonafide record of the project work done by her under my supervision. The contents of the project, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Date: Student's Signature

Dr. J UMARANI

Project Guide

Assistant Professor

Computer Engineering

Indian Institute of Information Technology Design and Manufacturing Kancheepuram

Date: Project Guide's Signature

Abstract

Image Captioning refers to automatically generating a description for the content of an image. To achieve the goal of Image Captioning, semantic information of images needs to be captured and expressed in natural languages. Connecting both research communities of Computer Vision and Natural Language Processing, Image captioning is a quite challenging task.

In this project, we aim at building a neural network model architecture that extracts the features and nuances of the image, and translates the extracted features and objects to a natural sentence. There has been a substantial increase in number of proposed models for image captioning task since neural language models and convolutional neural networks(CNN) became popular. This project has its base on one of such works, which uses a variant of recurrent neural network coupled with a CNN. We intend to enhance this model by making subtle changes to the architecture and using phrases as elementary units instead of words, which may lead to better semantic and syntactical captions.

Contents

Chapter 1

Introduction

Automatic image captioning (also known as automatic image reading) is the process by which a system or a model automatically generates content in the form of captioning or keywords to a digital image. This method can be regarded as a type of multi-class image classification and Image-to-Text translation with a very large number of classes - as large as the vocabulary size. Typically, image analysis in the form of extracted feature vectors and the training annotation words are used by machine learning techniques to attempt to automatically apply annotations to new images.

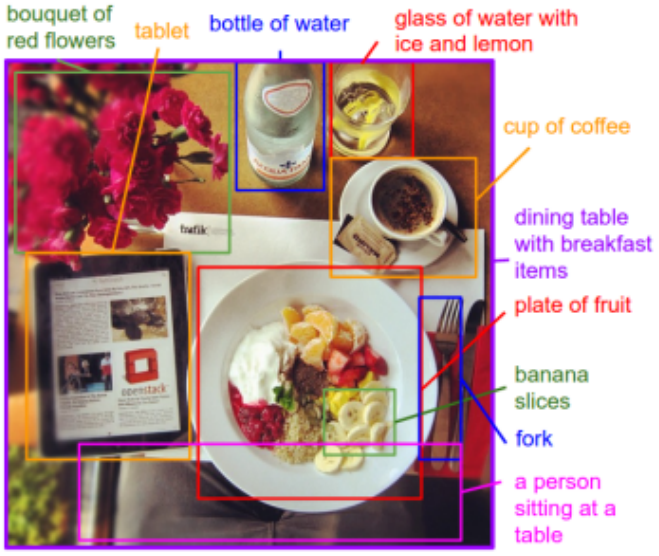


FIGURE 1.1: Motivation/Concept Figure: Generating descriptions of image regions.

A quick glance at an image is sufficient for a human to point out and describe an immense amount of details about the visual scene. However, this remarkable ability has proven to be an elusive task for visual recognition models. The majority of previous work in visual

recognition has focused on labeling images with a fixed set of visual categories and great progress has been achieved in these endeavors.

In this work, we strive to take a step towards the goal of generating dense descriptions of images as shown in Fig. 1.2.

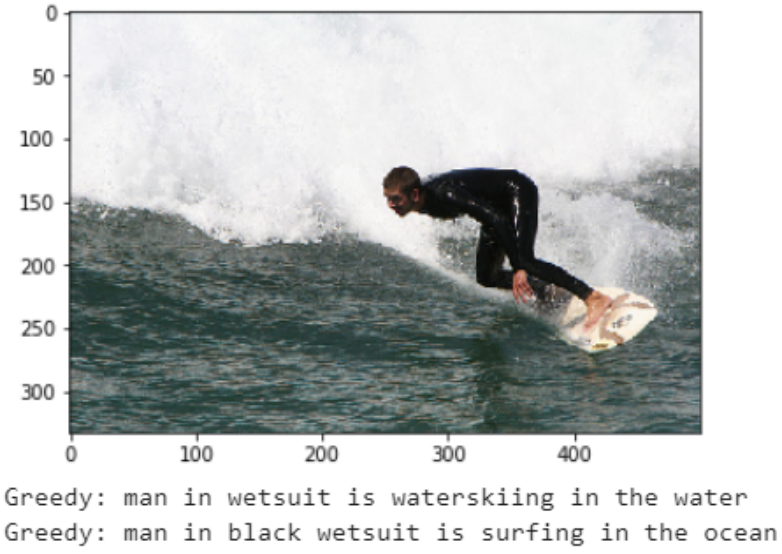


FIGURE 1.2: Automated Captioning of an Image

1.1 Motivation

Generating captions for images is a vital task relevant to the area of both Computer Vision and Natural Language Processing. Mimicking the human ability of providing descriptions for images by a machine is itself a remarkable step along the line of Artificial Intelligence. The main challenge of this task is to capture how objects relate to each other in the image and to express them in a natural language (like English). Traditionally, computer systems have been using pre-defined templates for generating text descriptions for images.

However, this approach does not provide sufficient variety required for generating lexically rich text descriptions. This shortcoming has been suppressed with the increased efficiency of neural networks. Many state of art models use neural networks for generating captions by taking image as input and predicting next lexical unit in the output sentence.

1.2 Background

The majority of previous work in visual recognition has focused on labeling images with a fixed set of visual categories and great progress has been achieved in these endeavors. However, these models often rely on hard-coded visual concepts and sentence templates, which imposes limits on their variety. A few of the challenges faced in the previous work:

1. Building a design of a model that can simultaneously reason about contents of images and their representation in the domain of natural language
2. Relying on hard-coded templates, rules or categories instead of the learning from the training data
3. Mentions of several entities whose locations in the images are unknown

1.3 Problem Formulation

Given an input as an image, we must retrieve a collection of descriptions/captions that perfectly fit to the given input. Image Reader's focus is to achieve extremely fast captioning and learn about the inter-modal correspondences between language and visual data that could be used in various applications.

1.4 Contribution

We aim at developing a deep neural network model that infers the latent alignment between segments of sentences and the region of the image that they describe. Further, we will introduce a multi-modal Recurrent Neural Network architecture that takes an input image and generates its description in text.

Chapter 2

Literature Review

Image Captioning overlaps with a number of other areas. In addition to the connection with Computer Vision and Natural Language Processing, it relies heavily on feature extraction and information retrieval. The majority of previous work in visual recognition has focused on labeling images with a fixed set of visual categories and great progress has been achieved in these endeavors. However, while closed vocabularies of visual concepts constitute a convenient modeling assumption, they are vastly restrictive when compared to the enormous amount of rich descriptions that a human can compose.

2.1 Dense Image Annotations

Several works studied the problem of holistic scene understanding in which the scene type, objects and their spatial support in the image is inferred. However, the focus of these works is on correctly labeling scenes, objects and regions with a fixed set of categories, while our focus is on richer and higher-level descriptions of regions.

2.2 Generating Descriptions

The task of describing images with sentences has also been explored. A number of approaches pose the task as a retrieval problem, where the most compatible annotation in the training set is transferred to a test image, or where training annotations are broken up and stitched together. Several approaches generate image captions based on fixed templates that are filled based on the content of the image or generative grammars, but this approach limits the variety of possible outputs.

2.3 Grounding natural language in images

A number of approaches have been developed for grounding text in the visual domain. More closely related is the work of Karpathy et al. [1], who decompose images and sentences into fragments and infer their inter-modal alignment using a ranking objective. In contrast to their model which is based on grounding dependency tree relations, our model aligns contiguous segments of sentences which are more meaningful, interpretable, and not fixed in length.

2.4 Neural networks in visual and language domains

Multiple approaches have been developed for representing images and words in higher-level representations. On the image side, Convolutional Neural Networks (CNNs) have recently emerged as a powerful class of models for image classification and object detection. On the sentence side, our work takes advantage of pretrained word vectors to obtain low-dimensional representations of words. Finally, Recurrent Neural Networks have been previously used in language modeling, but we additionally condition these models on images.

Chapter 3

Methodology

3.1 Image Reader Dataset

Image Reader uses a dataset Flickr8k consisting of 8000 visual images of everyday objects, scenes and entities. The dataset consists of both images and five-different human descriptions of each image summing upto 40000 text descriptions in total. We use a dataset of 6000 images with 30000 text descriptions to extract the features and nuances out of every training image and translate the features and objects to a natural sentence.

3.2 Model Architecture

3.2.1 Image Based Model

The image based model is used for extracting the features and nuances out of every image. The image is converted into a fixed sized 2048-length vector which can then be fed as input to the neural network. For this purpose, we opt for transfer learning by using the InceptionV3 model (Convolutional Neural Network) created by Google Research.

Inception-v3 Architecture (Batch Norm and ReLU are used after Conv) is 42 layers deep and its computation cost is only about 2.5 higher than that of GoogLeNet, and much more efficient than that of VGGNet. This model was trained on Imagenet dataset to perform image classification on 1000 different classes of images. However, our purpose here is not to classify the image but just get fixed-length informative vector for each image. This process is called automatic feature engineering. Hence, we just remove the last softmax layer from the model and extract a 2048 length vector (bottleneck features) for every image.

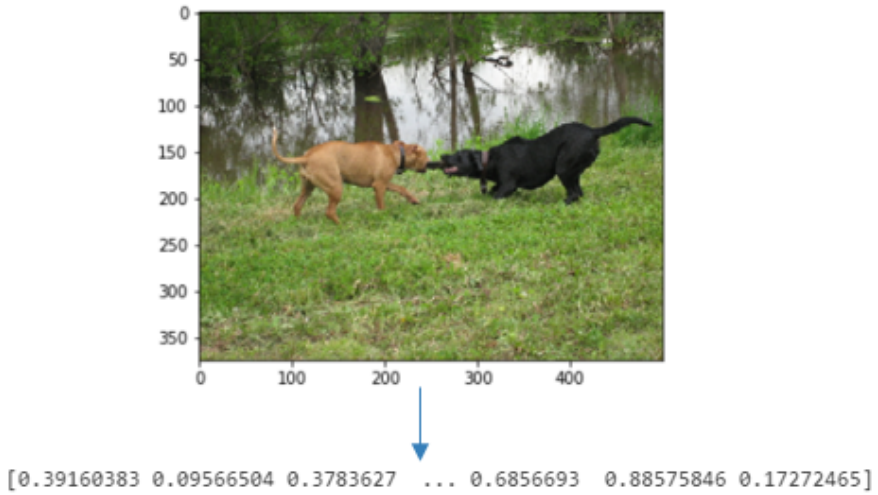


FIGURE 3.1: Encoding of Image to Vector

3.2.2 Language Based Model

The language based model is used for translating the features and objects given by the image based model to a natural sentence.

Firstly, we generate a data matrix for each caption provided for every image as shown in Fig. 3.2.

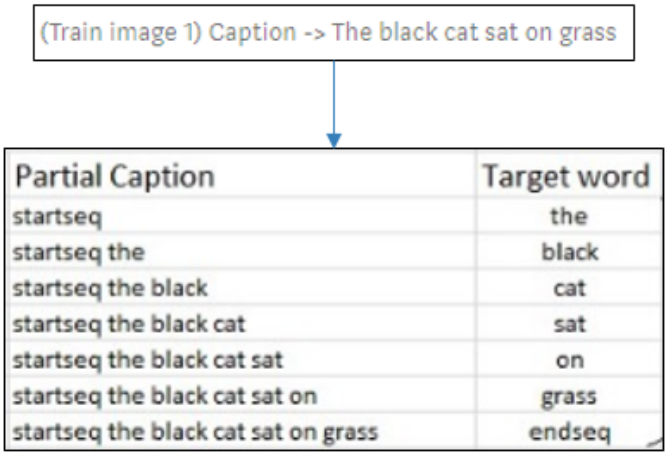


FIGURE 3.2: Encoding of Text to Data Matrix

Later, this data matrix is converted into an indices data matrix as shown in Fig. 3.3.

Partial Caption	Target word
[9, 0, 0 ...., 0]	10
[9, 10, 0, 0 ...., 0]	1
[9, 10, 1, 0, 0 ...., 0]	2
[9, 10, 1, 2, 0, 0 ...., 0]	8
[9, 10, 1, 2, 8, 0, 0 ...., 0]	6
[9, 10, 1, 2, 8, 6, 0, 0 ...., 0]	4
[9, 10, 1, 2, 8, 6, 4, 0, 0 ...., 0]	3

FIGURE 3.3: Generation of index-data matrix

3.2.3 Integrated Model Architecture

Since the input consists of two parts, an image vector and a partial caption, we cannot use the Sequential API provided by the Keras library. For this reason, we use the Functional API which allows us to create Merge Models.

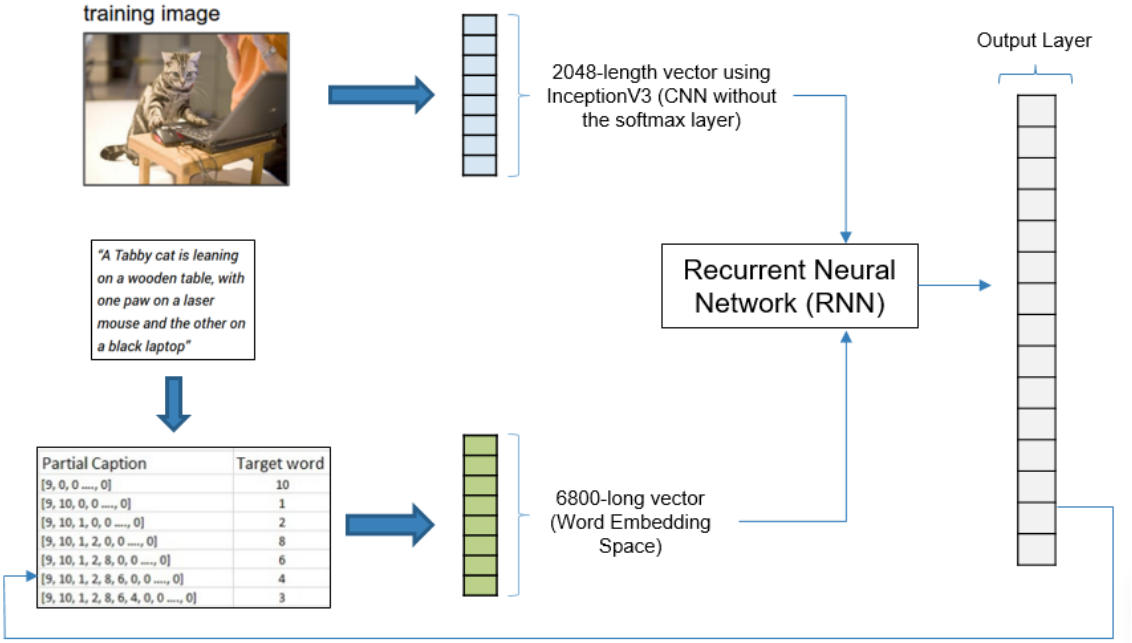


FIGURE 3.4: Integrated Model Architecture

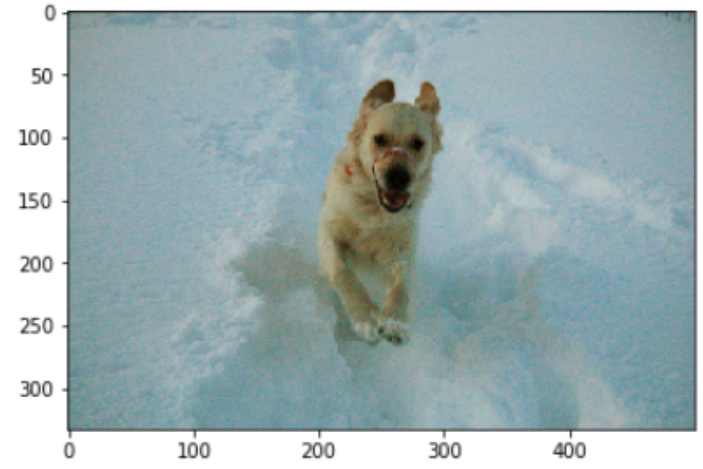
In summary, the output is an appropriate word, next in the sequence of partial caption provided in the input (or in probability terms we say conditioned on image vector and the partial caption)

3.3 Results

We intend to report accuracy using BLEU (Bilingual Evaluation Understudy) which scores machine translation hypotheses by aligning them to one or more reference translations.

In Fig. 3.5,

- The **Description-1** describes about **dog running through the snow** without referring to any physical attributes and hence, the **BLEU score** for that particular description is **0.96**.
- The **Description-2** describes about **a "white" dog running through the snow** and hence, the **BLEU score** is **0.98** due to a more scrutinized description.



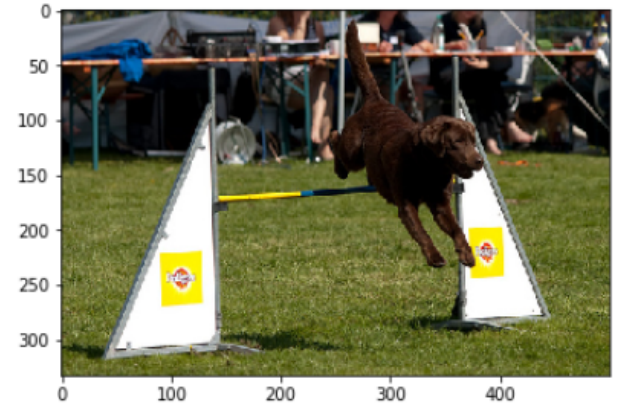

```
Description 1: dog is running through the snow
Description 2: white dog is running through the snow
************************************************************
BLEU Score 1: 0.96325106
BLEU Score 2: 0.98665442
```

FIGURE 3.5: Test Image-1:

Similarly, a few other examples of Image Captioning with their BLEU scores.


```
Description 1: man in blue shirt and helmet is riding
               bike on the street
Description 2: man in blue shirt and jeans is riding
               bike on the street
************************************************************
BLEU Score 1: 0.8739351325046805
BLEU Score 2: 0.8739351325046805
```

FIGURE 3.6: Test Image-2




```
Description 1: dog is running through the grass
Description 2: dog is jumping over hurdle
************************************************************
BLEU Score 1: 0.5623413251903491
BLEU Score 2: 0.9036020036098448
```

FIGURE 3.7: Test Image-3

Chapter 4

Conclusion and Future Work

We generated a model that infers correspondences between objects present in images and generates natural language descriptions of them and their respective regions. The features and nuances out of every image is first extracted using an image based model comprising of Inception-v3 CNN (without the softmax layer). Later, the language based model translates the features and objects given by the image based model to a natural sentence using an integrated model architecture. BLEU metric is used for evaluating a generated sentence to the reference sentences already existing in the dataset.

In future work, we hope to investigate whether the results of this model would remain similar when the experiments are repeated on other applications of conditioned neural language models such as neural machine translation or question answering. Furthermore, by keeping language and image information separate, merge architectures lend themselves to potentially greater portability and ease of training.

For example, it should be possible in principle to take the parameters of the RNN and embedding layers of a general text language model and transfer them to the corresponding layers in a caption generator. This would reduce training time as it would avoid learning the RNN weights and the embedding weights of the caption generator from scratch. As understanding of deep learning architectures evolves in the NLP community, one of our goals should be to maximise the degree of transferability among model components.

Bibliography

[1] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. [Online]. Available: https://cs.stanford.edu/people/karpathy/cvpr2015.pdf

[2] S. Bai and S. An, "A survey on automatic image caption generation," *In Neurocomputing*, pp. 15–16, 2018.

[3] L. Z. D. P. Ramakrishna, VedantamC., "Cider: Consensus-based image description evaluation," in *IEEE Conference On Computer Vision And Pattern Recognition (CVPR)*, ser. CVPR '15. IEEE Computer Society, 2015. [Online]. Available: https://arxiv.org/abs/1411.5726

[4] L. B. Serra, "Paying more attention to saliency: Image captioning with saliency and context a ention," 2018.

[5] H. Lamba, "Image captioning with keras," in *Medium*, November 2019. [Online]. Available: https://towardsdatascience.com/

[6] SUTime, *SUTime Temporal Tagger*. The Stanford Natural Language Processing Group, http://nlp.stanford.edu:8080/sutime/process.