INFORMATICS INSTITUTE OF TECHNOLOGY

INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

UNIVERSITY OF WESTMINSTER

CA-VQA: Context Aware - Visual Question Answering

A dissertation by

Mr. Mohamed Nasif Nuha

Supervised by

Mr. Guhanathan Poravi

Submitted in partial fulfilment of the requirements for the BEng Software Engineering degree at the University of Westminster.

May 2021

Abstract

VQA has emerged as a multidisciplinary challenge that integrates both vision and language processing. It has gained a lot of interest from both computer vision and natural language processing communities. Simply put, given an image and a natural language question about the image, a VQA task requires the system to find the correct answer by combining visual features of the image with inferences drawn from the question. A successful system must be able to comprehend an image semantically, understand the textual input, and generate a response based on its visual, textual, and logical interpretation of the inputs. This is typically done by making use of deep learning based techniques that extract the visual features of the image and the textual features of the question. Many researchers are interested in VQA because of its various application. Multiple methods and approaches have been utilized to propose a number of VQA models over time.

The author believes that given the complexity of an image, simply using the image and the question is insufficient for a VQA model to perform well. Considering this, the author proposes CA-VQA, an experiment that aims to validate the use of additional textual information about an image in order to produce better results. A third input is used by CA-VQA to define the image. The accuracy of the proposed method is 59.19%, which is considered acceptable.

Keywords: Deep Learning, Computer Vision, Visual Question Answering, Natural Language Processing