Image Captioning: Teaching Computers To Describe Pictures

Review-I Document

Group 15

Nakul Jadeja (19BCE0660)

Ishi Yadav (19BCE0482)

Sayan Saha (19BCE510)

Vardhan Khara (19BCE0833)

Submitted to

Prof. Gladys Gnana Kiruba B, SCOPE

School of Computer Science and Engineering

VIT®

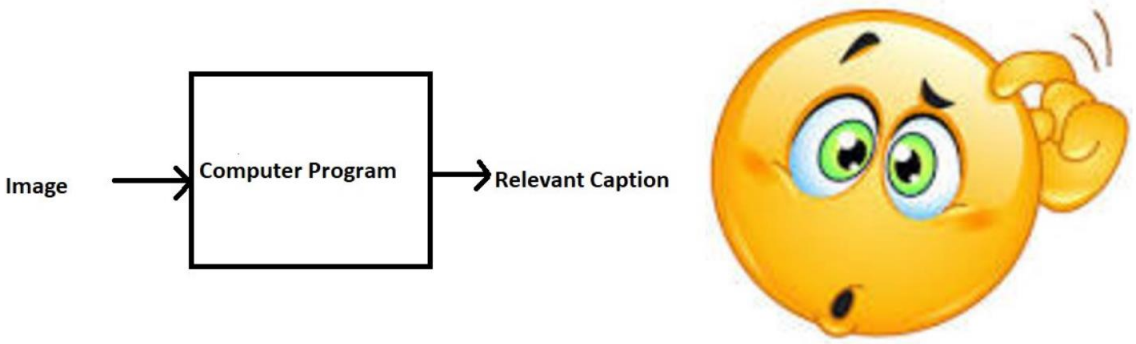
Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

Abstract:

Caption generation is a challenging artificial intelligence problem where a textual description must be generated for a given photograph.

Our project consists of a computer program that takes an image as input and automatically produces a relevant caption as an output.



The Problem

This project also makes use of basic Deep Learning concepts like Multi-layered Perceptrons, Convolution Neural Networks, Recurrent Neural Networks, Transfer Learning, Gradient Descent, Backpropagation, Overfitting, Probability, Text Processing, Python syntax and Data Structures, use of Keras library, etc.

Image captioning is a task to generate a new caption using the training data of the image and caption. Since existing Deep Learning is a black-box model, it is crucial to analyse the influence on each module for understanding the model. In this project, we plan to explore the impact of the various modules and do a comparative analysis according to three losses and two optimisations using two different datasets. From extensive experiments, the best component of each module will be identified as an improved method.

Keywords:

Text Processing, Image Captioning, Neural Networks, Backpropagation, Python Language, Data structures, Keras Library, Multi-Layered Perceptrons, Gradient Descent, Convolution Neural Networks, Overfitting

Introduction:

Image captioning such as sports commentary, video storytelling, and video captioning is a training method for models using image and caption data describing the images. Image captioning is a relatively difficult problem. It needs multi-modal processing with two different data types, natural language processing for caption data and computer vision for efficient information extraction from images. In image captioning, research that uses the attention concept will mainly be studied for the scope of our project. For example, semantic attention and text-guided attention improved the accuracy of the attention. In addition, to represent the relationship between objects in image captioning, we aim to extract features from the object detector and acquire a boundary box on the object portion.

References:

1. https://cs.stanford.edu/people/karpathy/cvpr2015.pdf
2. https://arxiv.org/abs/1411.4555
3. https://arxiv.org/abs/1703.09137
4. https://arxiv.org/abs/1708.02043
5. https://machinelearningmastery.com/develop-a-deep-learning-caption-generation-model-in-python/
6. https://www.youtube.com/watch?v=yk6XDFm3J2c
7. https://www.appliedaicourse.com/