ABSTRACT

The combination of computer vision and natural language processing in Artificial intelligence has sparked a lot of interest in research in recent years, thanks to the advent of deep learning. The context of a photograph is automatically described in English. When a picture is captioned, the computer learns to interpret the visual information of the image using one or more phrases. The ability to analyze the state, properties, and relationship between these objects is required for the meaningful description generation process of high-level picture semantics. Using CNN -LSTM architectural models on the captioning of a graphical image, we hope to detect things and inform people via text messages in this research. To correctly identify the items, the input image is first reduced to grayscale and then processed by a Convolution Neural Network (CNN). The COCO Dataset 2017 was used. The proposed method for blind individuals is intended to be expanded to include persons with vision loss to speech messages to help them reach their full potential and to track their intellect. In this project, we follow a variety of important concepts of image captioning and its standard processes, as this work develops a generative CNN-LSTM model that outperforms human baselines

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABBREVIATION	EXPANSION
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
NLP	Natural Language Processing
DL	Deep Learning
FC	Fully Connected
LRCN	Long-term Recurrent Convolutional Network

CHAPTER 1

INTRODUCTION

1.1 Introduction

Every day, we are bombarded with photos in our surroundings, on social media, and in the news. Only humans are capable of recognizing photos. We humans can recognize photographs without their assigned captions, but machines require images to be taught first. The encoder-decoder architecture of Image Caption Generator models uses input vectors to generate valid and acceptable captions. This paradigm connects the worlds of natural language processing and computer vision. It's a job of recognizing and evaluating the image's context before describing everything in a natural language like English.

Our approach is based on two basic models: CNN (Convolutional Neural Network) and LSTM (Long Short-Term Memory). CNN is utilized as an encoder in the derived application to extract features from the snapshot or image, and LSTM is used as a decoder to organize the words and generate captions. Image captioning can help with a variety of things, such as assisting the visionless with text-to-speech through real-time input about the scenario over a camera feed, and increasing social medical leisure by restructuring captions for photos in social feeds as well as spoken messages.

Assisting children in recognizing chemicals is a step toward learning the language. Captions for every photograph on the internet can result in faster and more accurate authentic photograph exploration and indexing. Image captioning is used in a variety of sectors, including biology, business, the internet, and in applications such as self-driving cars wherein it could describe the scene around the car, and CCTV cameras where the alarms could be raised if any malicious activity is observed. The main purpose of this research article is to gain a basic understanding of deep learning methodologies.

1.1.2 Objective

1. The project aims to work on one of the ways to context a photograph in simple English sentences using Deep Learning (DL).
2. The need to use CNN and LSTM instead of working with RNN

1.1.3 Scope

Our project extends and is being used in any large-scale business industry and also small-scale business industry.

1.2 Statement Problem

In our world, information is considered valuable and some humans face a serious problem regarding visualizing an image. We hence dig into this matter, considering blindness as a major factor, and generate a sentence by allowing users to upload or scan a visual image.

Advantage

- Recommendations in Editing Applications
- Assistance for visually impaired
- Social Media posts
- Self-Driving cars
- Robotics
- Easy to implement and connect to new data sources

Disadvantages

- Do not make intuitive feature observations on objects or actions in the image
- Nor do they give an end-to-end mature general model to solve this problem

CHAPTER 2

LITERATURE SURVEY

Literature survey is the most important step in the software development process. Before developing the tool, it is necessary to determine the time factor, economy, and company strength. Once these things are satisfied, then the next step is to determine which operating system and language can be used for developing the tool. Once the programmers start building the tool, they programmers need a lot of external support. This support can be obtained from senior programmers, books, or websites. Before building the system, the above considerations are taken into account for developing the proposed system.

The major part of the project development sector considers and fully surveys all the required needs for developing the project. For every project, a Literature survey is the most important sector in the software development process. Before developing the tools and the associated designing it is necessary to determine and survey the time factor, resource requirement, manpower, economy, and company strength.

To improve and tailor the user experience on its products, photos use image classification. Intraclass variation, occlusion, deformation, size variation, perspective variation, and lighting are all frequent issues in computer vision that are represented by the picture classification problem.

Methods that work well for picture classification are likely to work well for other important computer vision tasks like detection, localization, and segmentation as well.

Image captioning is a great illustration of this. Given an image, the image captioning challenge is to generate a sentence description of the image. The picture captioning problem is comparable to the image classification problem in that it expects more detail and has a bigger universe of possibilities. Image classification is used as a black box system in modern picture captioning systems, therefore greater image classification leads to better captioned.

The image captioning problem is intriguing in and of itself because it brings together two significant AI fields: computer vision and natural language processing. An image captioning system demonstrates that it understands both image semantics and natural language.

Once these things are satisfied and fully surveyed, then the next step is to determine about the software specifications in the respective system such as what type of operating system the project would require, and what all the necessary software is needed to proceed with the next step such as developing the tools, and the associated operations.

To construct an image sentence, image classification is a key stage in the object recognition and picture analysis process. The final output of the image categorization phase might be a statement.

To date, a variety of image captioning techniques have been presented. Several studies have been carried out in attempt to determine the best image captioning technique. It's difficult to pick one approach as the finest of them all because the results and accuracy are dependent on a variety of circumstances.

In order to achieve the most accurate results, traditional approaches have been constantly modified as well as new image captioning techniques invented during the previous few decades.

Each caption generator technique has its own set of benefits and drawbacks. The focus of the research today is on combining the desired qualities of various techniques in order to boost efficiency.

Many high-level tasks, such as image classification, object detection, and, more recently, semantic segmentation, have recently been proven to obtain outstanding results using convolutional neural networks with many layers. A two-stage technique is frequently used, especially for semantic segmentation. Convolutional networks are trained in this way to offer good local pixel-wise data for the second stage, which is often a more global graphical analysis model.

We will use Long short-term memory (LSTM), which is a subset of RNNs, to tackle the problem of Vanishing Gradient. The main goal of LSTM is to solve the problem of Vanishing Gradients. The unique feature of LSTM is that it can keep data values for long periods, allowing it to address the vanishing gradient problem.

When compared to applying RNN, the results revealed that using a mixture of LSTM generated better outcomes.

CNNs employ multilayer convolution to accomplish feature engineering and integrate these features internally, unlike traditional image recognition algorithms. It also employs the pooling and fully connected (FC) layers, as well as SoftMax.

CHAPTER 4

PROJECT IMPLEMENTATION, ALGORITHMS, AND METHODOLOGY

4.1 Introduction

This project is loaded with CNN and LSTM which act as the platform to generate the sentences from a simple image. This can be worked on all applications.

4.2 Hardware Requirements

- System: i3 Processor
- Hard Disk: 500 GB.
- Monitor: 15”LED
- Input Devices: Keyboard, Mouse
- Ram: 4GB.

4.3 Software Requirements

- Platform: Google Colab
- Coding Language: Python

4.4 Working Explanation

1. A user uploads an image that they want to generate a caption for.
2. A gray-scale image is processed through CNN to identify the objects.
3. A gray-scale image is processed through CNN to identify the objects.
4. CNN scans images left-right, and top-bottom, and extracts important image features.
5. By applying various layers like Convolutional, Pooling, Fully Connected, and thus using activation function, we successfully extracted features of every image.
6. It is then converted to LSTM.
7. Using the LSTM layer, we try to predict what the next word could be.