IMAGE CAPTION GENERATOR

PROJECT SYNOPSIS

OF MAJOR PROJECT

BACHELOR OF TECHNOLOGY

Computer Science and Engineering

SUBMITTED BY

AYUSH SHARMA (2000290100044)

DEEPESH SINGH (2000290100052)

AYUSH SINGH (2000290100045)

KIET Group of Institutions, Delhi-NCR, Ghaziabad (UP)

Department of Computer Science and Engineering

Table of Contents

INTRODUCTION

Every day, we encounter a large number of images from various sources like the net, news articles, document diagrams and advertisements. These sources contain images that viewers would have to interpret themselves. Most images don't have a description, but the human can largely understand them without their detailed captions. However, machines have to interpret some variety of image captions if humans need automatic image captions from it. Image captioning is very important for several reasons. Captions for each image on the web can result in faster and descriptively accurate images searches and indexing.

The goal of this project is to generate appropriate captions for a given image. The captions will be generated in order to capture the contextual information on the images. Current methods uses convolutional neural networks (CNNs) and recurrent neural networks (RNNs) or their variants to generate appropriate captions. These networks provide an encoder- decoder method to do this task, where CNNs encode the image into feature vectors and RNNs are used as decoders to generate language descriptions .

Image captioning finds various applications in various fields such as commerce, biomedicine, web searching and military etc. Social media such as Instagram , Facebook etc can generate captions automatically from images.

RATIONALE

Creating an image caption generator project using the MERN (MongoDB, Express.js, React, Node.js) stack and machine learning has several compelling rationales:

Cross-Domain Skill Development: Developing this project will allow you to gain proficiency in both web development (MERN stack) and machine learning. This combination of skills is highly valuable in the technology industry, as it bridges the gap between front-end development and data science.

Real-World Application: Image caption generation has practical applications in various domains, such as accessibility for visually impaired individuals, content indexing for search engines, and enriching multimedia content with descriptive captions. This project can serve as a portfolio piece to showcase your skills to potential employers or clients.

Machine Learning Integration: Implementing a machine learning model to generate captions for images demonstrates your ability to work with complex algorithms, neural networks, and deep learning techniques. You can leverage pre-trained models like Convolutional Neural Networks (CNNs) for image feature extraction and Recurrent Neural Networks (RNNs) for caption generation.

Data Management: Using MongoDB (or any other database) in the MERN stack allows you to efficiently store and manage image data, captions, and user information. This experience is valuable for understanding how to handle diverse data types in real-world applications.

User Interface: React, a component-based JavaScript library, provides a powerful and interactive user interface for your image caption generator.

OBJECTIVES

The objectives of a project to create an image caption generator using the MERN stack and machine learning are multifaceted, combining both technical and functional goals. Here are the primary objectives:

The project aims to work on one of the ways to context a photograph in simple English sentences using Deep Learning (DL)

Technical Proficiency: Develop proficiency in the MERN stack, including MongoDB, Express.js, React, and Node.js, by actively using these technologies in the project.

Machine Learning Integration: Implement and integrate machine learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to generate descriptive captions for images.

Data Management: Create a robust database architecture to efficiently store and manage image data, captions, user information, and related metadata.

User-Friendly Interface: Design an intuitive and user-friendly web interface (using React) that allows users to upload images and receive generated captions seamlessly.

Image Processing: Develop image processing capabilities to preprocess and prepare images for input into the machine learning model, including resizing, normalization, and feature extraction.

Model Training and Optimization: Train, fine-tune, and optimize the machine learning model(s) for accurate and contextually relevant image caption generation.

LITERATURE REVIEW

A review on image caption generating technologies:

Introduction:

Image captioning involves generating natural language descriptions for images.
Image captioning has various applications, including social media, e-commerce, web search, and more.
The project aims to use deep learning techniques to automatically generate captions for images in a natural language like English.

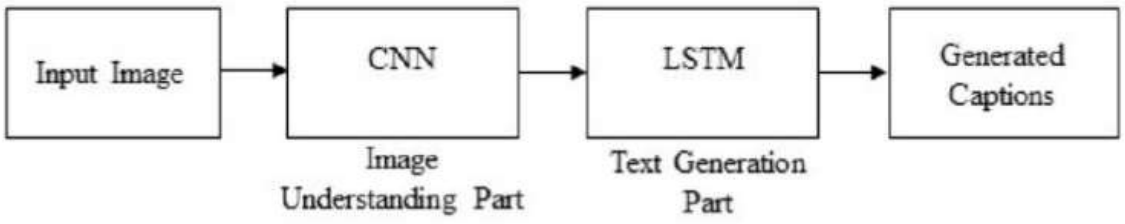
Existing System:

Image positioning and object detection are well-researched aspects of computer vision.
Social media users often post images, and search engines like Google are used to find image descriptions.
Challenges include image quality, complexity, and the time required for processing.

Proposed System:

Deep Neural Networks, specifically CNNs and LSTMs, can be used to generate expressive and fluent image captions.
The system will provide a platform for social media users to upload images and automatically generate captions.
It can handle color and black-and-white images of any size and read out captions in English.
The system uses TensorFlow and algorithms to efficiently calculate automatic metrics and generate captions.
Working Model and Project Architecture: The review mentions the

working model and project architecture without providing specific details.



Algorithms:

Convolutional Neural Network (CNN): CNNs are used to process image data and can analyze images by scanning them from left to right and top to bottom.
Long Short-Term Memory (LSTM): LSTMs are a type of recurrent neural network (RNN) suitable for sequence prediction tasks and can capture relevant information over time.
Flickr8k Dataset:

The project uses the Flickr8k dataset, which contains 8000 photos, each with five captions.
This dataset is suitable because it includes multiple captions for each image, making the model more robust.
Conclusion:

The literature review concludes by summarizing the discussed deep learning-based image captioning approaches.
It acknowledges that while significant progress has been made, creating robust image captioning systems that can generate high-quality captions for any image remains a challenge.
The review suggests that as the use of social media and image sharing grows, the demand for automatic image captioning will also increase.

FEASIBILITY STUDY

Technical Feasibility: The project involves the use of MERN Stack and Machine Learning, both of which are well-established technologies with extensive documentation and community support. The technical skills required for this project include JavaScript proficiency, understanding of MERN stack, and knowledge of Machine Learning algorithms, particularly those related to image processing and natural language processing.

Economic Feasibility: The economic aspects to consider include the cost of development time, hosting the application, and potential expenses related to database storage (MongoDB), server costs (Node.js/Express.js), and potentially Machine Learning computation if not using a free service. However, all the technologies involved are open-source, which significantly reduces costs.

Legal Feasibility: As all the technologies proposed are open-source, there are no licensing fees involved. However, it is important to comply with the terms and conditions of any third-party APIs or datasets used for machine learning model training.

Schedule Feasibility: The timeline for completion would depend on the complexity of the machine learning model and the proficiency of the development team with MERN Stack. A basic version of the project could potentially be completed in a few months.

Market Feasibility: There is a demand for automated image captioning in various sectors such as digital marketing (for SEO), accessibility (for visually impaired users), and content management systems.

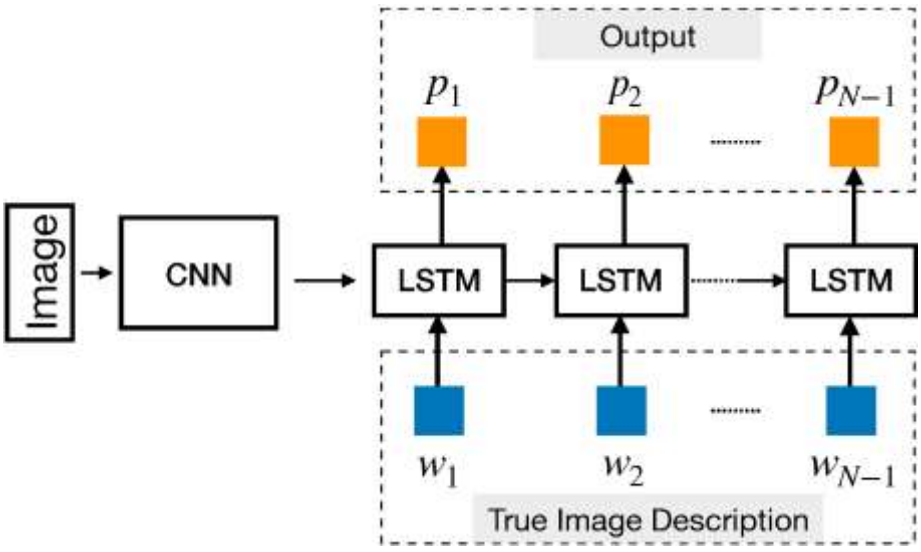
METHODOLOGY/ PLANNING OF WORK

This project needs a dataset that includes both photographs and captions. The image captioning model should be able to be trained using the dataset.

FLICKR 8K DATASET

The Flickr 8k dataset is a publicly available image-to-sentence description benchmark. There are 8091 photos in this collection, each with five captions. These photos were taken from a variety of groups on the Flickr website. Each caption gives a detailed description of the objects and events seen in the photograph. The collection represents a wide range of events and settings and excludes photographs of well-known persons and locations, making it more generic. The dataset has the following characteristics that make it ideal for this project are:

- When many captions are mapped to a single image, the model becomes more general and avoids overfitting.
- Using a variety of training images allows the image captioning model to cope with a variety of image types, making the model more robust.



FACILITIES REQUIRED

Minimum system requirements :

- Operating system: Windows, macOS, or Linux
- Processor: 1 GHz or faster processor
- Memory: 2GB RAM or higher
- Storage: 500 MB free disk space or higher

Technical skills required :

front-end technologies:

- HTML
- CSS
- JAVASCRIPT
- REACTS

Backend Technologies:

- NODE JS
- PYTHON

Database Used:

- MONGO DB

EXPECTED OUTCOMES

The primary expected outcome is a fully functional image caption generator application. Users should be able to upload images, and the system should generate descriptive captions for those images in natural language.

A web application that can generate captions for images uploaded by users or fetched from the internet.

A machine learning model that can understand the context of an image and produce appropriate captions using natural language processing techniques.

A MERN stack that can provide a full-stack framework for developing and deploying the web application, using MongoDB for the database, Express.js for the backend server, React.js for the frontend, and Node.js for the runtime environment.

A user interface that can display the images and their captions, as well as allow users to interact with the application and provide feedback.

Ensure that the application works seamlessly across various web browsers and devices, enhancing its accessibility.

REFERENCES

[1] Wu, Q., Sheen, C., Liu, L., Dick, A., & van den Hengelo, A. (2016). What value do explicit high-level concepts have in vision to language problems? Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 203–212

[2] He, K., Zhang, X., Ran, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

[3]Cho, K., Courville, A., & Bengi, Y. (2015). Describing multimedia content using attention- based encoder-decoder networks. IEEE Transactions on Multimedia, Vol. 17

[4] Saad Alawi, Tareq Abed Mohammed, and Saad Al-Zai, "Understanding of a convolution neural network", IEEE – 2017

[5] Oriol Vinals, Alexander Torshavn, Samy Bengi, and Dumitru Erhan, "Show and Tell: A Neural Image Caption Generator", (CVPR 1, 2- 2015)

[6] Ma, Ningning, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun." Shuffle net v2: Practical guidelines for efficient CNN architecture design." In Proceedings of the European conference on computer vision (ECCV), pp. 116-131. 2018.

[7] Rehab Alahmadi, Chung Hyuk Park, and James Hahn, "Sequence-to sequence image caption generator", (ICMV-2018)

[8] MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shirat Uddin, and Hamid Laga, "A Comprehensive Survey of Deep Learning for Image Captioning", (ACM-2019)