International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 05 | May - 2023 Impact Factor: 8.176 ISSN: 2582-3930

AUTOMATED IMAGE CAPTION GENERATOR

Dr.N.Lakshmi Priya
Assistant Professor
Department of Computer Science and Engineering

Nalla Malla Reddy Engineering College
Narapally, Divyanagar, Hyderabad
lakshmipriya.cse@nmrec.edu.in

Utheja,Nikhila,Nava Kumar Reddy,Sai Ram
B.Tech Final Year
Department of Computer Science and Engineering
Nalla Malla Reddy Engineering College
Narapally,Divyanagar,Hyderabad
19B61A0523@nmrec.edu.in

Abstract - Image captioning is the process of generating a sentence description for an image using automated techniques. Our model aims to develop a model that takes an image as input and generates an English sentence as output, describing the contents of the image. To achieve this, we employ the concepts of Convolutional Neural Networks (CNNs) and Long Short Term Memory (LSTM) algorithms. The CNN acts as an encoder, extracting features from the image, while the LSTM works as a decoder, generating words that describe the image. Once the image is processed and the caption is generated, we evaluate the effectiveness of our model. The aim is to create a system that provides accurate and descriptive captions for the given input image. By implementing this image caption generator, we hope to assist users in understanding and interpreting the contents of the image, making it easier for them to comprehend and communicate the visual information.

Keywords–Image Caption, Web application, Accuracy, Less error rate, CNN, LSTM.

I. INTRODUCTION

Image caption generation is a challenging task that combines computer vision and natural language processing techniques. The objective is to enable a computer to understand the contents of an image and describe it in a human-like language. Our Python-based project implements an image caption generator using Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models. To generate captions, we first extract image features using an exceptional CNN model that was trained on the Flickr 8k dataset. These features are then fed into the LSTM model, which generates the image captions. This approach allows us to create accurate and detailed descriptions of images automatically, making it useful in a wide range of applications, from social media to image tagging and search engines. Overall, our project aims to leverage the latest advancements in AI to create better tools for image analysis and understanding.

II. LITERATURE SURVEY

This system enables authorized access for effective use of an organization through proper login. College placement officers can manage placement information, and students with access can add their information to use as a resume with only one-time registration required. The system promotes an eco-friendly approach with a focus on reducing water, paper, space, and time to preserve the environment. The author suggests implementing different rounds, typically 3-5, during placement drives to reduce the number of candidates for better job allotments. The paper explores the benefits of online versus offline media for various purposes, and argues that online media is preferable for modern activities due to fast and easy access to data. A study on the aesthetics and design of e-commerce web pages found that users are influenced by the design, simplicity, and amenities offered. The Android operating system was developed by the Open Handset Alliance, a coalition of over 50 mobile technology companies. Several papers were presented on the development of campus employment information networks, including the use of SQLite for data storage, SMS integration for instant notifications, and the utilization of data mining algorithms for understanding data. Other papers explored features such as attendance monitoring, progress tracking, scheduling, discussion forums, and alumni contact for students. In summary, the papers presented various approaches to improving campus employment networks through the use of technology, including online platforms, database management, and communication tools.

III. METHODOLOGY

The caption generation system comprises of three models that work together to streamline the process of describing an image: (a) a Feature Extraction Model, (b) an Encoder Model, and (c) a Decoder Model.

IV. DEVELOPMENT

GUI frameworks and Machine Learning models have made the process of writing GUI and building ML models much simpler. These tools enable developers to focus on their application's unique features without worrying about common development issues such as user management. Most of these frameworks are open-source software, which makes them easily accessible and cost-effective. Although there is a potential for developing applications on Internet operating systems, there are currently limited options available. These frameworks and models promote rapid application development, which can be beneficial for businesses looking to create applications in a shorter amount of time.

V. EXISTING SYSTEM

Image Caption generation is commonly achieved by,

Query expansion method:

The approach used in this method involves retrieving images similar to a given input image from a large dataset, and then utilizing the distribution associated with these retrieved images to create an extended query. The candidate descriptions are then reordered by estimating the cosine similarity between the distributed representation and the extended query vector. Finally, the closest description is selected as a description for the input image.

Feature Extraction Methods:

The approach utilized in this method involves two key components - a statistical probability language model for generating handcrafted features and a neural network model based on an encoder-decoder language model for extracting deep features.

- The first component involves generating handcrafted features using a statistical probability language model. This is achieved by utilizing the model to analyze the statistical probability of certain features in the dataset. These features are then used to create a set of handcrafted features that are specific to the dataset.
- The second component involves utilizing a neural network model based on an encoder-decoder language model to extract deep features. This involves training the model to learn the underlying patterns and relationships in the dataset and using this knowledge to generate deep features that capture the essence of the data.

By combining these two components, the method is able to generate both handcrafted and deep features that can be used to accurately analyze and classify the data.

Attention Mechanism:

The attention mechanism is a cognitive ability observed in human beings that stems from the study of human vision in cognitive neurology. People have the ability to consciously ignore certain information while focusing on other important information when receiving information. This selective ability is known as attention. The attention mechanism was first proposed to be utilized in image classification in the field of visual images by applying it to the RNN model.

VI. LIMITATIONS

A combination of Convolutional neural network and Recurrent neural network has been widely used for generating captions. However, these models often encounter problems such as gradient vanishing, inaccurate identification of objects and their relationships, and the generation of captions only for seen images. As a result, the existing systems have a high error rate in generating captions. Although various methods have been proposed to address these issues, they all share the common disadvantage of not making intuitive feature observations on objects or actions in the image, nor do they provide an end-to-end general model to solve this problem. Despite these challenges, the efficiency and popularity of neural networks have led to breakthroughs in the field of image description, and with the emergence of big data and the outbreak of deep learning methods, new hopes have emerged in this .

VII. PROPOSED SYSTEM

The process of image caption generation involves utilizing deep learning and computer vision techniques to recognize the content of an image and provide it with relevant captions. During model training, datasets are used to label images with English keywords. The flickr 8k dataset is commonly utilized to train the CNN model called

Xception, which is responsible for feature extraction from the image. The extracted features are then fed into the LSTM model to generate the final image caption. The ultimate goal of this project is to develop a model that can accurately generate captions for images of varying clarity, reducing the error rate and producing precise captions.

VIII. ARCHITECTURE

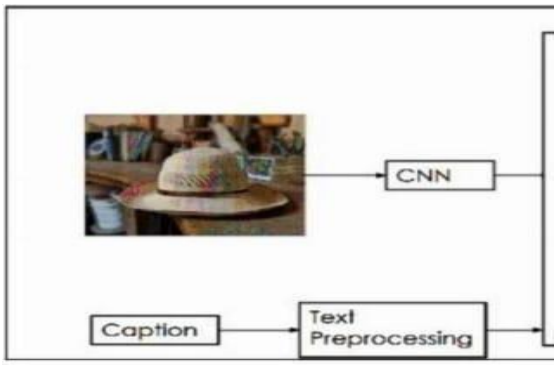

Figure 1. A

IX. MODULES

Image Pre-Processing:

The image processing system is a crucial step for machine learning models that work with images. Since the machines cannot understand images, the input image is first converted into a fixed-sized pixel matrix, typically of size 224x224x3. This matrix represents the color code of each pixel in the respective location.

Then, the image undergoes a pre-processing stage where the noise is removed to obtain a clear image. In this stage, the image is first converted into grayscale, and then the threshold value is set to divide the image into foreground and background. Edge detection is performed to identify the boundaries of every object in the image.

The final output of the image pre-processing module is a pixel matrix that can be fed into the next module of the machine learning pipeline for further processing.

Image based model:

The Image Based Module is a modified version of CNN, where Convolutional and Pooling layers extract features from the input matrix of pixels. This module comes after the Image Pre-processing Module, and its output is a feature vector containing important features such as objects, verbs (action of the object), color of the object, and their relationships. The feature vector is then linearly transformed to match the input size of the LSTM network, which is used in the next module.

Language based model:

The Language Based Module takes in the encoded feature vector from the Image Based Module and aims to convert it into simple language that can be understood by users. This is achieved using the Long Short-Term Memory (LSTM) algorithm, which overcomes the gradient issue of the RNN algorithm and can store long sequences of data without forgetting the order. The LSTM algorithm uses its memory cells to store the data. Before training the Language Based Model (LB), the labels and target text need to be pre-defined. The label stores the data in a sequence starting with a start token, and the target stores the sequence of data with an end token at the end, so that the algorithm knows when to stop. For example, consider the caption "X and Z are playing basketball". The label would be "[start, X, and, Z, are, playing, basketball, .]" and the target would be "[X, and, Z, are, playing, basketball, end]".

Caption Generation:

The Caption Generation module is the final module of the image caption generator, which takes as input the output from the previous Language Based Module. The goal of this module is to generate a caption in a linear sequence based on the encoded image features and the pre-defined label and target texts. The generated caption is in a human-readable form and provides a concise and accurate description of the input image. The module uses the previously trained LSTM algorithm to predict the most likely sequence of words that form the caption. Once the prediction is made, the words are decoded and transformed into a simple language to generate the final caption.

X. RESULT ANALYSIS AND CONCLUSION

The image caption generator in this Python-based project uses CNN (Convolutional Neural Networks) and LSTM (Long short term memory) to provide accurate and relevant captions for the given image input. The CNN model used for feature extraction is Xception, trained on the flickr 8k dataset. These features are then passed to the LSTM model, responsible for generating the image captions, resulting in precise captions and reduced error rates. The model can also generate captions for unseen or untrained images. Additionally, EfficientNet-B4 is used in the proposed model as it offers similar FLOPS to ResNet-50 while improving the top-1 accuracy from 76.3% to 92.6%.

XI. FUTURE SCOPE

The project currently focuses on the LSTM approach for generating image captions, but there is potential to expand its scope and make the system even more efficient. One possible extension could be to add a feature that converts text messages into speech, which would be particularly useful for people with vision impairments. The goal of this proposed extension is to create a system that is accessible and can help people with vision loss achieve their full potential.

XII. REFERENCES

1. Base Paper: Katiyar, S., &Borgohain, S. K. (2021). Comparative evaluation of CNN architectures for image caption generation. *arXiv preprint arXiv:2102.11506*.
2. Kalena, P., Malde, N., Nair, A., Parkar, S., & Sharma, G. (2019). Visual Image Caption Generator Using Deep Learning. In *2nd International Conference on Advances in Science & Technology*.
3. Xu, K., CA, U., Ba, J. L., CA, U., Kiros, R., EDU, T., ... &Bengio, Y. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention (Supplementary Material).
4. Jia, X., Gavves, E., Fernando, B., &Tuytelaars, T. (2015). Guiding the long-short term memory model for image caption generation. In *Proceedings of the IEEE international conference on computer vision* (pp. 2407-2415).
5. Yagcioglu, S., Erdem, E., Erdem, A., &Cakıcı, R. A Distributed Representation Based Query Expansion Approach for Image Captioning (Supplementary Material).
6. Kinghorn, P., Zhang, L., & Shao, L. (2018). A region-based image caption generator with refined descriptions. *Neurocomputing*, *272*, 416-424.