IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 6, June 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306073|

Efficient Image Retrieval Using AI

Amar Chouhan, Geetanjali Sahoo, Mahesh Vallakati, Shreyans Chheda, Vidya Gogate

U.G. Student, Electronic & Computer Science Engineering, Shah & Anchor Kutchhi Engineering College, Mumbai, India

Associate Professor, Electronic & Computer Science Engineering, Shah & Anchor Kutchhi Engineering College, Mumbai, India

ABSTRACT: This research describes the creation of an advanced image search system that combines a search system capable of processing text queries in real time with an image captioning model to extract relevant photos.

The system uses advanced semantic search algorithms and reinforcement learning in a large image library to demonstrate accurate and intuitive results while highlighting ethical aspects such as bias detection and mitigation.

It includes a detailed description of the system's architecture, implementation, and performance, focusing on how it can revolutionize interaction with visual data and inspire further AI advances in image search.

KEYWORDS: Image retrieval, Artificial Intelligence, Natural Language Processing, Computer Vision, Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Beam search, Microsoft GIT, Transformers pipeline.

I. INTRODUCTION

The project aims to transform the way users interact with images by combining artificial intelligence, natural language processing, and computer vision, allowing them to retrieve related photos using text queries.

The central component of the system is a high-quality image captioning model trained on the well-known MS COCO dataset, optimized using TensorFlow and Keras frameworks, and built using Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN). By incorporating Microsoft's open-source MS Git model, we leverage advanced machine learning and artificial intelligence technologies to further improve text-to-image captioning.

The image search system uses various NLP techniques to improve search capabilities, identifying and displaying images that exactly match a user-generated text query. This overview provides an in-depth look at the system's architecture, implementation, and performance evaluation, and highlights its potential to transform user interaction with visual data and open the door to further developments in AI-driven image search.

II. RELATED WORK

1. Introduction to Convolution Neural Networks

The importance of Convolutional Neural Networks (CNNs) in image processing is explored in a literature review.

CNNs are an effective technique for removing fine details from images, which has enabled the creation of robust image identification and classification systems. Experts have conducted a detailed analysis of the use of CNNs in various fields and demonstrated their effectiveness in tasks such as feature extraction, image segmentation, and object detection.

The paper highlights the continuous improvements in the performance and design of Convolutional Neural Networks by tracing the evolution of CNN architectures from early models such as LeNet to modern deep learning frameworks such as ResNet and VGG.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306073|

2. Introduction to Recurrent Neural Network

Recurrent Neural Networks (RNNs) are discussed in literature reviews on sequential data analysis and natural language processing. RNNs perform well in sequential data processing tasks, making them ideal for applications such as time series forecasting, speech recognition, and language modelling. Experts have studied many versions of RNNs such as Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) to address issues such as long-term dependencies and vanishing gradients. To tackle complex tasks such as image captioning and language translation, this overview highlights how versatile RNNs are in extracting time dependencies in data and how they can be integrated into various deep learning architectures.

3. Microsoft GIT

This literature review delves into the Microsoft Generative Image-to-text Transformer (GIT) paradigm, designed to revolutionize tasks involving language and vision, such as text recognition, question answering, and image captioning. GIT simplifies the traditional architecture by employing a unified image encoder and text decoder, setting new benchmarks for generating diverse and high-quality captions across various visual contents. The review highlights GIT's superior performance in encoding multimodal information and creating relevant captions for complex visual contexts, surpassing human capabilities on challenging benchmarks like TextCaps. GIT's proficiency in computer vision and NLP is evident through its ability to manage a wide range of scenarios, from interpreting handwritten text to providing detailed scene descriptions. This positions GIT as a groundbreaking advancement in generative models for language and vision tasks.

III. METHODOLOGY

1. Image Captioning

We use the Git-Large Text-to-Image Synthesis (GIT-Large) model to generate descriptive image captions. This mechanism uses a Transformer-based architecture consisting of an encoder and a decoder. The encoder processes the input image and extracts a set of visual features, which are fed into the decoder. The decoder generates a set of tokens that form a coherent descriptive caption for the input image. Specifically, the GIT-Large model uses a vision-language pre-training approach, where the model is pre-trained on a large dataset of image-text pairs, allowing it to learn the relationship between visual and linguistic representations.

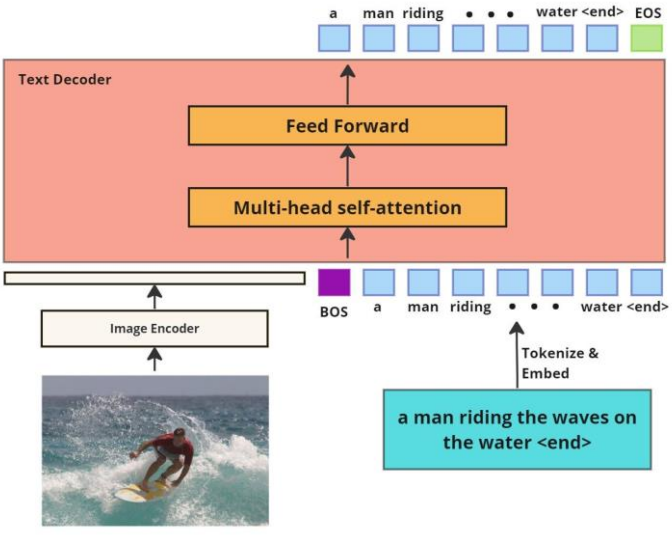

Fig. 1. Pre-trained in MS GIT

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306073|

During inference, the model takes an input image and generates a label by iteratively predicting the next token in a sequence depending on the previous token and visual features extracted from the image. This process is facilitated by the model's ability to focus on different regions of the image, allowing it to focus on the most relevant visual features when generating labels.

2. Search Mechanism

The system also includes a gallery view that displays the retrieved images in an attractive and user-friendly manner and includes features such as scrolling, zooming, and image preview to optimize user interaction with the search results. The system improves the flexibility and adaptability of the search mechanism to various user inputs by enabling various query types, including keyword-based, phrase-based, and semantic-based search, demonstrating a comprehensive approach to image search based on user-generated text queries.

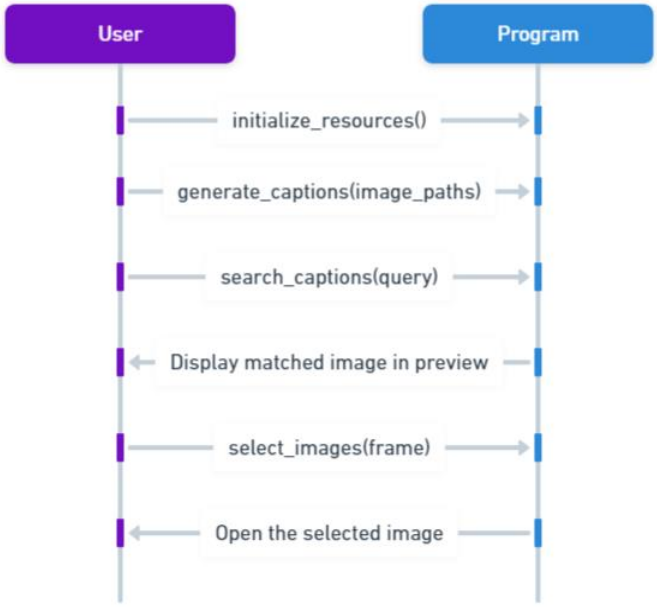

Fig. 2. User-Program Interaction

To maximize user interaction with the search results, the system also includes a gallery view that displays the returned photos graphically in an attractive and user-friendly manner and includes features such as scrolling, zooming, and image preview to optimize user interaction with the search results. The system improves the flexibility and adaptability of the search mechanism to various user inputs by enabling various query types, including keyword-based, phrase-based, and semantic-based search. It presents a comprehensive image retrieval method based on user-generated text queries.

This image captioning and retrieval system achieves its goal using a combination of pre-trained models and custom algorithms. Image captioning is done using a pipeline model from the Transformers library optimized on a dataset of image-caption pairs. The system uses NLP techniques for text processing, tokenization, and speech analysis, while it also uses computer vision techniques for image preprocessing and feature extraction. A graphical user interface developed in Tkinter provides users with an intuitive platform to interact with the system and allows image selection, caption generation, and search functionality.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306073|

IV. EXPERIMENTAL RESULTS

Below are the steps of user interaction with the image captioning system

Step 1: Select Images
The user selects one or more images from their computer or device. The images are uploaded to the model, and the system begins generating captions for each image.

Step 2: Generate Captions
The system generates captions for each selected image using the Microsoft GIT-Large TextCaps model. The captions are generated based on the visual content of the images.

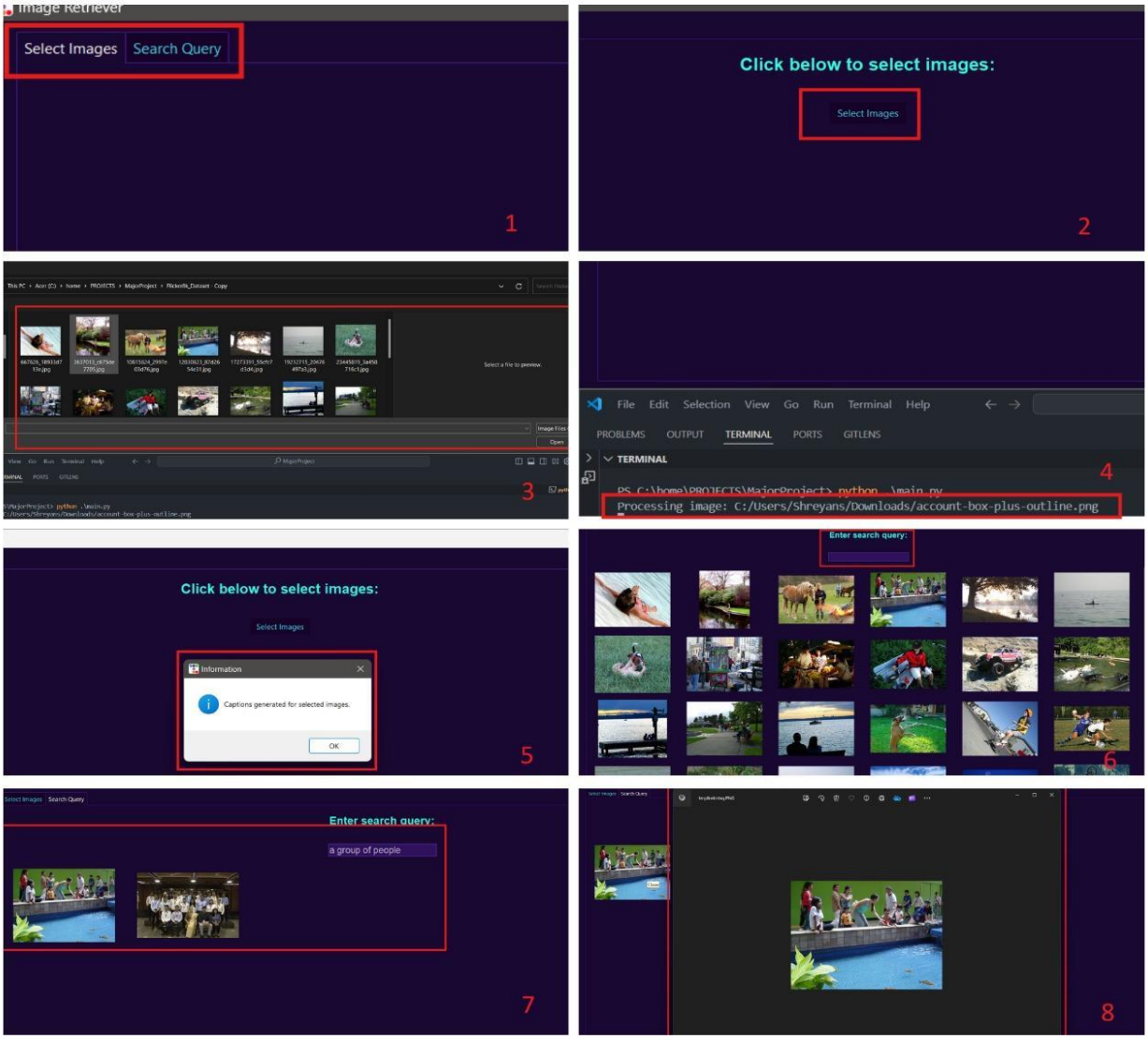

Fig. 3. User interaction with the system

Step 3: Search for Images
The user enters a text query in the search bar. The system searches for images that match the query by comparing the query words with the captions of the images.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306073|

Step 4: Display Search Results
The system displays a list of images that match the search query. Each image is displayed with a caption and can be opened by clicking on it.

V. CONCLUSION

The image captioning and retrieval system demonstrated a promising approach for generating descriptive captions and retrieving images based on user queries. Future research could explore the integration of more complex NLP models, such as transformer-based architectures, to improve the quality and accuracy of image captioning. Furthermore, integrating additional modalities, such as audio and video, could enable more thorough multimedia queries and analysis. Optimizations such as distributed computing, caching, and efficient algorithms should be considered to ensure the responsiveness and scalability of the system. The user interface can be improved with intuitive design elements, and the search functionality can be improved with query expansion and refinement techniques. Developing evaluation metrics allows for accurate evaluation and comparison of system performance. Pursuing these approaches will allow the system to evolve and improve, ultimately resulting in more accurate, efficient, and user-friendly image caption and retrieval capabilities.

REFERENCES

1. S. Liu, L. Bai, Y. Hu, H. Wang, "Image Captioning Based on Deep Neural Networks," MATEC Web of Conferences.
2. P. Anderson, B. Fernando, M. Johnson, S. Gould, "Guided Open Vocabulary Image Captioning with Constrained Beam Search," Thesis, The Australian National University, Canberra, Australia.
3. M. Browne, S. Ghidary, "Convolutional Neural Networks for Image Processing: An Application in Robot Vision," 2003.
4. B. Shuai, Z. Zuo, W. Gang, "Quad Directional 2D-Recurrent Neural Networks for Image Labeling," IEEE Signal Processing Letters, 2015
5. P. M. Ashok Kumar, T. Rao, A. R. Lakshminarayanan, E. Pugazhendi, "An Efficient Text-Based Image Retrieval Using Natural Language Processing (NLP) Techniques," 2021.
6. S. Lai, L. Xu, K. Liu, J. Zhao, "Recurrent Convolutional Neural Networks for Text Classification," Thesis, National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, China.
7. J. A. Bullinaria, "Recurrent Neural Networks," Neural Computation Lecture 12, 2015.
8. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," Thesis, 2015.
9. E. Cohen, C. Beck, "Empirical Analysis of Beam Search Performance Degradation in Neural Sequence Models," Proceedings of the 36th International Conference on Machine, 2019.
10. J. Wang et al., "GIT: A Generative Image-to-text Transformer for Vision and Language," Paper
11. M. Jaderberg, K. Simonyan, A. Vedaldi, A. Zisserman, "Reading text in the wild with convolutional neural networks," IJCV, 2016.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details