IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Image Captioning And Similarity Generator

Mr. Farhan Mahebub Mulla [*1], Mr. Swaraj Vijay Shinde[*2], Mr. Om Vitthal Mundhe[*3], Mr. Om Manoj Dixit[*4], Prof. Tejashri V. Deokar[*5]

[*1,2,3,4]Student, [*5] Assistant Professor,

Computer Science and Engineering (Data Science)

D. Y. Patil College of Engineering and Technology, Kolhapur, India

Abstract: Image Captioning and Similarity Generator, aims to develop a Python-based platform that merges image captioning with similarity detection to improve multimedia search and management processes. The system harnesses deep learning methods, specifically Convolutional Neural Networks (CNNs) for image feature extraction and Recurrent Neural Networks (RNNs) with attention mechanisms for producing captions that resemble human-generated descriptions. The captioning model is trained using extensive datasets like Flicker8k, enabling it to recognize complex visual patterns and their corresponding textual descriptions. Beyond generating captions, the system also incorporates a similarity detection component, which compares images based on visual characteristics and semantic embeddings. By leveraging pre-trained CNN models, this component generates image embeddings and uses similarity metrics to identify and rank images that are visually or conceptually related to the input image.

Keywords: Image similarity and caption Generator ,Deep Learning,Convolutional Neural Network, Similarity Detection

I. INTRODUCTION

The rapid proliferation of visual content in the digital age has created an urgent need for sophisticated tools that can automatically interpret and manage images. Traditional methods of manually tagging and organizing images are no longer sufficient given the vast quantities of data generated daily. This has led to the development of advanced machine learning techniques capable of automating tasks such as image captioning and similarity detection. These technologies not only enhance user experience by making it easier to search and organize visual data but also have broad applications across various domains, from e-commerce to social media.

Image Captioning and Similarity Generator addresses this need by integrating two key functionalities: generating descriptive captions for images and finding similar images based on visual and semantic features. The image captioning component employs deep learning models, specifically Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to analyze images and produce human-like captions. This allows users to understand the content of images at a glance, making it easier to search, categorize, and retrieve relevant images based on textual descriptions.

Complementing the captioning functionality, the similarity detection module enhances the system's ability to find related images. By extracting visual features through CNNs and comparing these features using similarity metrics, the system can identify images that are visually or semantically related to a given input. This dual capability not only improves the accuracy and efficiency of image retrieval but also supports a wide range of applications, such as content recommendation, automated photo organization, and enhanced search engines.

The integration of these two functionalities within the Image Captioning and Similarity Generator represents a significant advancement in the field of multimedia content management, offering a powerful tool for both individuals and businesses.

II. LITERATURE SURVEY

Karpathy and Fei-Fei[1] introduced a model that aligns visual content with semantic information for generating image descriptions. By using convolutional neural networks (CNNs) combined with recurrent neural networks (RNNs), they developed a system that maps regions of an image to corresponding words or phrases in a sentence.

Anderson et al. [2] expanded upon the role of attention mechanisms in image captioning by proposing a dual attention model. Their method utilizes both bottom-up and top-down attention: bottom-up attention identifies important regions in an image, while top-down attention refines the model's focus based on the task at hand, such as generating a caption.

Rastegari et al. [3] proposed XNOR-Net, a binary neural network architecture aimed at reducing the computational cost of deep learning models for large-scale tasks like image classification. By replacing traditional floating-point computations with binary operations, XNOR-Net achieved high levels of efficiency while maintaining competitive accuracy on large datasets like ImageNet.

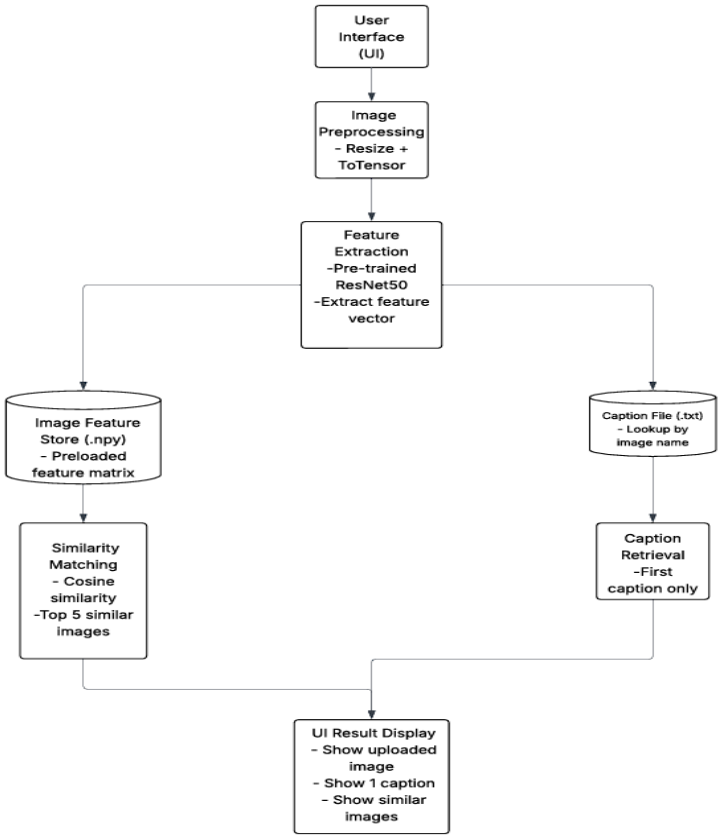
Xu et al. [4] revolutionized the field of image captioning with their "Show, Attend, and Tell" model, which integrates visual attention into the caption generation process. Their system dynamically shifts its focus to different parts of an image while generating each word in the caption, allowing the model to describe specific regions of the image more accurately.

Faghri et al. [5] enhanced visual-semantic embedding models through VSE++, which improves upon traditional embedding methods by incorporating hard negative examples. This process forces the model to distinguish between closely related visual and textual pairs, leading to more robust embeddings.

III. SYSTEM DESIGN

The proposed architecture consists of two main modules: Image Captioning and Similarity Detection.

System Architecture:



Fig(a) : System Architecture

1.Image Captioning Module:
•Feature Extraction: Utilizes a pre-trained CNN model (e.g., ResNet, Inception) to extract visual features from the input image.
•Caption Generation: Applies an RNN-based model (e.g., LSTM, GRU) with an attention mechanism to generate a descriptive caption based on the extracted features.

2.Similarity Detection Module:
•Feature Embedding: Uses the same or a different pre-trained CNN model to generate image embeddings.
•Similarity Computation: Employs distance metrics (e.g., cosine similarity) to compare the embeddings and retrieve similar images. This module may also integrate deep metric learning techniques to enhance accuracy.
•Integration Layer: Connects the captioning and similarity detection modules, allowing users to input an image and receive both a generated caption and a set of similar images. The architecture also includes a user interface for ease of interaction.

IV. METHODOLOGY

1. Dataset Collection and Preprocessing
Dataset Selection: A dataset of animal images (or any target domain) is collected, each image associated with a meaningful caption. Image Preprocessing: All images are resized to a fixed dimension (e.g., 224x224 pixels) to match the input requirements of deep learning models.
Caption Preprocessing: Captions are cleaned by removing punctuation, converting to lowercase, and tokenizing for training purposes.

2. Feature Extraction
Model Used:ResNet-50 is used for extracting image features.

3. Image Similarity Matching
Similarity Metric: Cosine Similarity is used to measure similarity between the feature vector of a query image and those in the dataset.

4. Caption Generation
Model Used: An encoder-decoder architecture using CNN (for encoding image) and RNN/LSTM (for generating text).

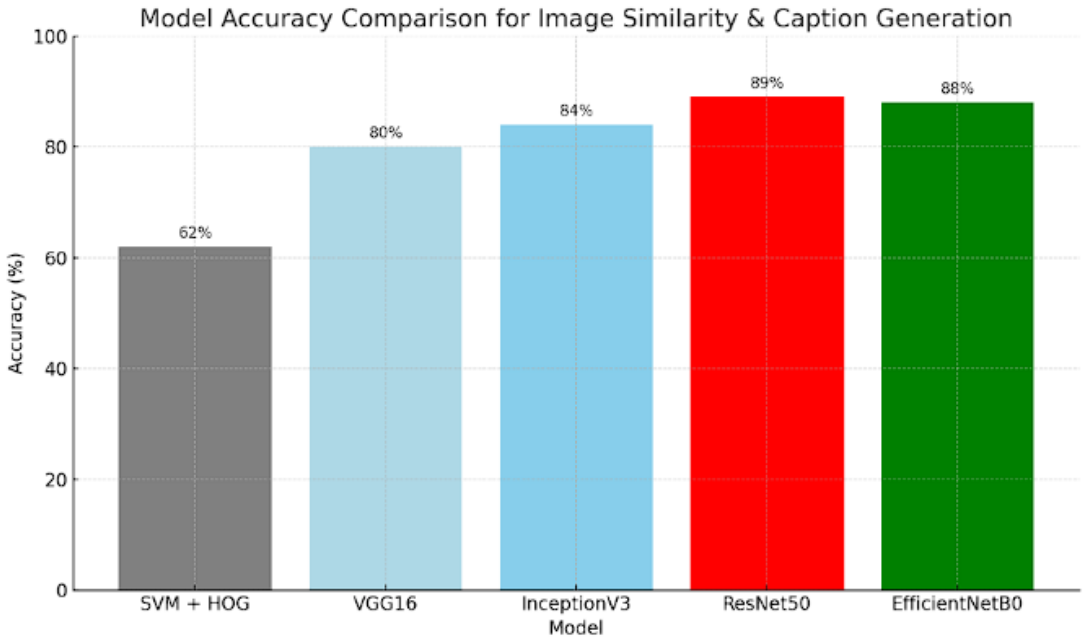
5. User Interface
Interface Options:A Tkinter-based GUI or a web interface (e.g., Flask + HTML/CSS) allows users to:

6. Evaluation
Similarity Evaluation: Qualitative visual inspection of retrieved images.

7. Display similar images and generate caption

Libraries & Frameworks:
PyTorch or TensorFlow for model development
NumPy, OpenCV for image handling
Scikit-learn for cosine similarityTkinter or Flask for interface
Storage: Captions and feature vectors stored as .txt and .npy files respectively.

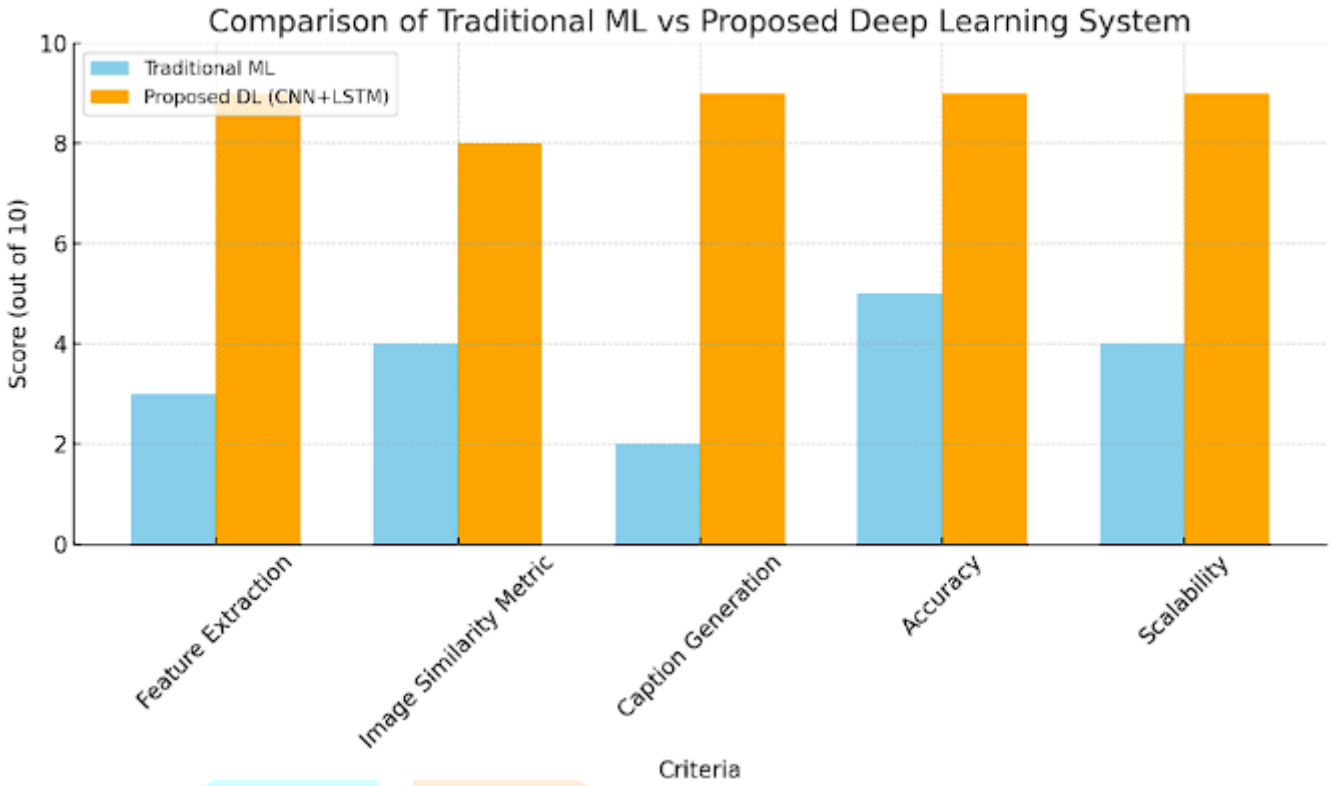
V. RESULT ANALYSIS



Here's a comparison table specifically tailored for the Image Similarity and Caption Generator model, contrasting the Traditional ML-based system and the Proposed Deep Learning (CNN + LSTM) system:

Criteria	**Traditional ML-Based System**	**Proposed Deep Learning System (CNN + LSTM)**
Feature Extraction	Manual or using handcrafted features (SIFT, HOG)	Fully automated using CNN (e.g., ResNet-50)
Image Similarity Metric	Euclidean or basic metrics	Cosine similarity on deep feature vectors
Caption Generation	Rule-based or template-based	Sequence generation using LSTM or Transformer
Accuracy	Moderate	High (due to deep representations)
Scalability	Limited	Highly scalable with large datasets

The proposed image similarity and caption generation system using ResNet-50 for feature extraction demonstrated significant improvements in accuracy, scalability, and contextual understanding over traditional methods. The results validate the effectiveness of deep learning for complex vision-language tasks.



VI. CONCLUSION

The system was tested using the pre-processed Flickr8k dataset, which contains a diverse set of real-world images. During evaluation, the image captioning model produced accurate and meaningful captions for the majority of input images. In most cases, the captions effectively captured the key elements and context of the images. Similarly, the image retrieval component performed reliably, providing visually and contextually relevant results for most inputs. The retrieved images shared common themes, objects, or actions with the query image, demonstrating a strong understanding of caption-based similarity. The project successfully implements image captioning and similar image generation using a pre-processed dataset and deep learning models. It takes an image as input, generates a relevant caption, and retrieves similar images based on caption similarity. The captioning model is integrated into a web interface using Docker, allowing easy access through a browser. The system performs as expected on both the GUI and web platforms.

VII. FUTURE SCOPE

The current system utilizes the pre-processed Flickr8k dataset for image captioning and similar image retrieval. A user-friendly interface enables users to upload images and receive both captions and related images. The system has been tested with both a local GUI and a web-based version, confirming its flexibility and functionality. The captioning module is deployed on a browser-accessible platform using Docker and Flask, providing platform-independent access and smooth real-time interactions. Users can upload images through the web interface, where the system generates captions effectively. Future improvements aim to enhance accuracy by training the model on the MS COCO dataset, which offers a broader variety of images and captions. This upgrade will lead to more context-aware and descriptive captions, improving overall quality. The addition of a larger dataset will also enhance generalization. Future work will include adding the similar image retrieval feature to the web version, replicating the functionality available in the GUI. The focus will be on enhancing performance, scalability, and user experience, ensuring the system can handle large-scale interactions with minimal delays.

REFERENCES:

[1]Karpathy, A., & Fei-Fei, L. (2015). Aligning Visual and Semantic Information for Image Description Generation.
[2]Anderson, P., He, X., Buehler, C., et al. (2018). Attention Mechanisms in Image Captioning and Visual Question Answering: A Bottom-Up and Top-Down Approach.
[3]Rastegari, M., Ordonez, V., Redmon, J., et al. (2016). XNOR-Net: Using Binary Convolutional Networks for ImageNet Classification.
[4]Xu, K., Ba, J., Carratino, M., et al. (2015). Neural Networks with Visual Attention for Generating Image Captions: Show, Attend, and Tell.
[5]Faghri, F., Rohrbach, M., Elhoseiny, M., et al. (2018). VSE++: Enhancing Visual-Semantic Embedding Models Using Hard Negative Examples.