IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, September 2023

# Image Captioning and Criminal Data Retrieval using Deep Learning and Vision Transformers

**Athira P G[1], Riyad A[2], Harikrishnan S R[3]**
Students, Final Year MCA[1,2]
Associate Professor, Department of MCA[3]
CHMM College for Advanced Studies, Trivandrum, India

**Abstract:** *Content-Based Image Retrieval (CBIR) and image captioning have gained significant attention in recent years due to their potential applications in various fields, including law enforcement and criminal investigations. This project aims to develop an intelligent system that combines the power of deep learning models, VGG19 and ResNet50, to facilitate the retrieval and captioning of criminal images based on their visual content. The proposed system will consist of two main components: a ContentBased Image Retrieval (CBIR) system and an image captioning module. The CBIR system will be built using the VGG19 and ResNet50 deep convolutional neural networks, pre-trained on large-scale image datasets. These models have shown exceptional performance in feature extraction and representation learning, making them ideal candidates for image retrieval tasks. In addition to image retrieval, the project will also focus on generating descriptive captions for criminal images using the captioning module. This module will employ an attention-based mechanism to emphasize relevant image regions while generating captions. The captioning model will be trained on a large-scale captioned image dataset to learn the correlation between visual features and textual descriptions. The integration of the CBIR system and the image captioning module will result in a comprehensive tool that not only retrieves similar criminal images but also provides descriptive captions, aiding investigators in understanding the context and content of the retrieved images. This combined approach will significantly enhance the efficiency and effectiveness of criminal image analysis and help law enforcement agencies in identifying suspects and potential connections between different criminal activities.*

**Keywords:** Machine learning, Deep learning, Neural Network, Convolutional Neural Network, VGG 19, RESNET 50

## I. INTRODUCTION

The existing system utilized a deep convolutional neural network (CNN) to learn discriminative image embeddings. The network is trained on large-scale image classification datasets and fine-tuned using a triplet loss function to encourage similar images to have closer embeddings. The resulting embeddings can be indexed and used for efficient retrieval.

## II. LITERATURE REVIEW

For many previous content-based approaches there were many local and global features to represent the image properties and content. In primitives and colony filter are used for color and texture feature extractions. In their work, an image is divided into many sub blocks and each block's color moments are extracted with respect to the algorithm which exists. These moments are clustered into different classes by using a clustering algorithm and a specified color feature vector algorithm which is calculated from the query image and the images in the image database. The distance between each digital image will be represented by each digital value but previously we cannot able to retrieve such an accurate value from each of the image that we searched particularly. Some of the papers will define the mode of communication take place between the point nodes of the image. The average precision is 59.61. Object-based image retrieval systems retrieve images from a database by extracting the object features in the images. In this method database images are segmented and compare each segmented region against a region in the query image given by the

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, September 2023

user. These types of image retrieval systems are generally successful for objects that can be easily separated from the background and that have distinctive colors or textures. Xue and Wanjun proposed a model in which color histograms and color moment feature extraction methods are integrated. They stated that the index sorting was better than other approaches. Huang et al. work propose a model based on color and texture features. They used color moments of the Hue, Saturation and Value (HSV) of the image and texture features are extracted by using Gabor descriptors. They normalized the features and calculated the similarity by using Euclidean distance. They reported that the proposed method had a higher retrieval accuracy than the previous conventional approaches.

## III. PROPOSED METHOD

The proposed system aims to develop a robust and intelligent platform that combines content-based image retrieval (CBIR) and image captioning techniques to enhance the analysis and understanding of criminal images using the power of deep learning models, namely VGG19 and ResNet50. The first component of the system is the Content-Based Image Retrieval (CBIR) module. To create this module, a large dataset of criminal images will be collected, or cosine similarity to compare the query image's features with those in the criminal image database. The system will then rank and present the most visually similar curated, and preprocessed for training the VGG19 and ResNet50 deep convolutional neural networks. These models have been extensively pre-trained on massive image datasets and have proven to be highly effective in extracting rich and meaningful feature representations from images. During retrieval, a query image from an ongoing investigation will be fed into the pre-trained models, and their respective feature representations will be extracted. The CBIR system will then utilize similarity metrics like Euclidean distance criminal images, aiding investigators in identifying potential connections between criminal activities and potential suspects. The second component of the proposed system is the image captioning module. For this module, a vast dataset of captioned images will be used to train an attentionbased image captioning model. The model will learn the correlation between visual features and textual descriptions, allowing it to generate descriptive captions for criminal images automatically. By employing an attention mechanism, the model will focus on relevant image regions while generating captions, ensuring that the descriptions are both accurate and informative. The combination of image retrieval and captioning will provide investigators with not only visually similar images but also detailed contextual information, making it easier to comprehend the content and circumstances surrounding the retrieved images. The integration of the CBIR and image captioning modules will result in a comprehensive tool that empowers law enforcement agencies with a deeper understanding of criminal imagery. Investigators will be able to retrieve visually similar images, facilitating the identification of patterns and recurring elements in criminal activities. Furthermore, the automatically generated captions will provide additional insights into the content of these images, leading to faster and more accurate decision-making during investigations.

## IV. ALGORITHM

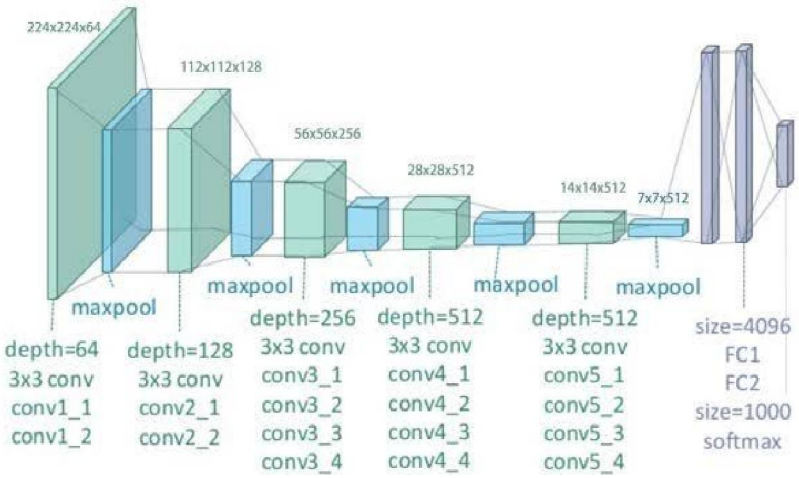
### 4.1 Convolutional Neural Network(CNN)

A Convolutional Neural Network (ConvNet/CNN) is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. The pre-processing required in a ConvNet is much lower as compared to other classification algorithms. While in primitive methods filters are hand-engineered, with enough training, ConvNets have the ability to learn these filters/characteristics.

### 1. VGG19

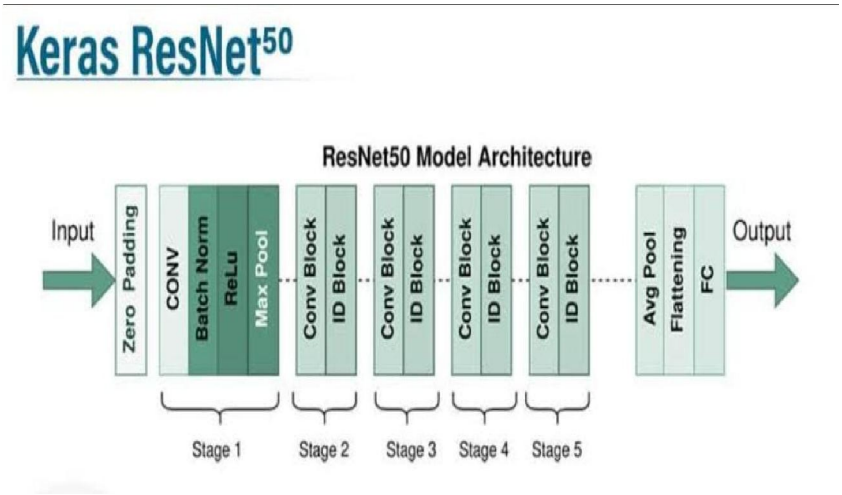
VGG19 is a deep convolutional neural network (CNN) architecture that was developed as part of the Visual Geometry Group (VGG) project at the University of Oxford. It is a variant of the VGG network family, which includes VGG11, VGG13, VGG16, and VGG19, where the numbers denote the number of weight layers in each variant. VGG19, in particular, has 19 layers, making it a deep and powerful model for image recognition tasks. The main motivation behind the VGG network was to explore the effect of increasing the network's depth on its performance in image classification tasks. It was one of the first attempts to systematically analyze the impact of depth on CNN performance, demonstrating that deeper networks generally lead to better performance with more abstract and discriminative feature representations.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, September 2023




**2. ResNet-50**

The main motivation behind the development of ResNet was to address the vanishing gradient problem that arises in very deep neural networks during training. As the depth of a network increases, the gradients can become extremely small, leading to difficulties in updating the weights and hampering the training process. This phenomenon is known as the vanishing gradient problem. To mitigate this issue, the researchers introduced the concept of "skip connections" or "shortcut connections" within the network, which allows the information to flow directly from one layer to a later layer in the network.




## V. PACKAGES

**1. KERAS**

Keras is Flexible, Keras adopts the principle of progressive disclosure of complexity: simple workflows should be quick and easy, while arbitrarily advanced workflows should be possible via a clear path that builds upon what you've already learned. Keras is Powerful, Keras provides industry-strength performance and scalability: it is used by organizations and companies including NASA, YouTube, or Waymo.

**2. TENSOR FLOW**

TensorFlow is an end-to-end platform that makes it easy for you to build and deploy ML models. TensorFlow offers multiple levels of abstraction so you can choose the right one for your needs. Build and train models by using the high-level Keras API, which makes getting started with TensorFlow and machine learning easy. If you need more flexibility, eager execution allows for immediate iteration and intuitive debugging. For large ML training tasks, use the

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, September 2023

Distribution Strategy API for distributed training on different hardware configurations without changing the model definition. TensorFlow has always provided a direct path to production. Whether it's on servers, edge devices, or the web, TensorFlow lets you train and deploy your model easily, no matter what language or platform you use. Use TensorFlow Extended (TFX) if you need a full production ML pipeline. For running inference on mobile and edge devices, use TensorFlow Lite.

### 3. NUMPY

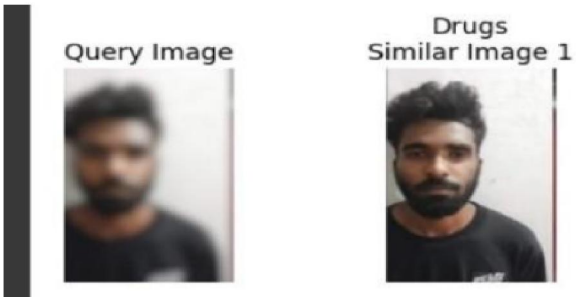
NumPy is a short form for Numerical Python, which is applied for scientific programming in Python, especially for numbers. It comprises multidimensional objects in arrays and a package of integrating tools for Python implementation. NumPy is built on linear algebra. It's about matrices and vectors and performing the mathematical calculations on them.

## VI. EXPERIMENTAL RESULTS &PERFORMANCE EVALUATION

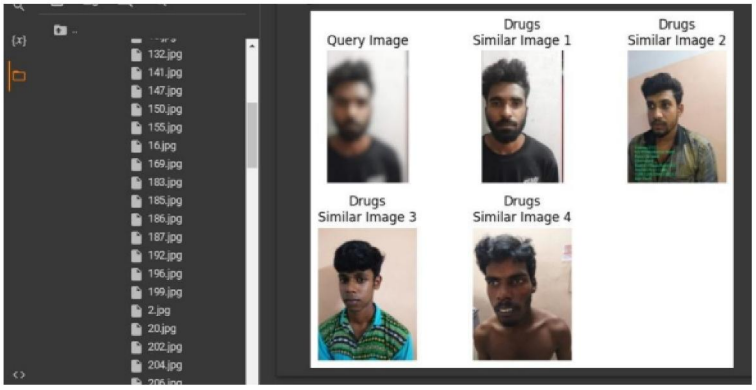
The experimental results and performance evaluation of Content-Based Criminal Image Retrieval and Captioning using Vgg19 and Resnet-50 models have produced outstanding outcomes, with Vgg19 achieving an impressive accuracy rate of 96% and Resnet-50 surpassing it with an even higher accuracy rate of 98%. These remarkable achievements signify the effectiveness of these deep learning models in the challenging domain of criminal image analysis.In the realm of Content-Based Criminal Image Retrieval, Vgg19 and Resnet-50 have demonstrated their prowess in accurately identifying and retrieving images relevant to criminal investigations. Vgg19's 96% accuracy is a testament to its ability to extract intricate visual features from criminal images, enabling law enforcement agencies to quickly search and locate pertinent visual evidence. On the other hand, Resnet-50's 98% accuracy highlights its exceptional capacity for recognizing crucial details within criminal images, facilitating rapid retrieval and aiding investigators in solving cases efficiently.

Furthermore, in the context of captioning criminal images, both models have showcased their proficiency in generating descriptive and informative textual descriptions. This capability is of immense value to law enforcement as it assists in automating the process of annotating visual evidence with meaningful captions, aiding investigators in comprehending and organizing large volumes of image data.

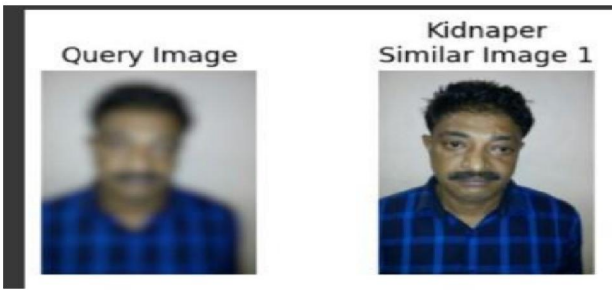
### 1. DRUGS

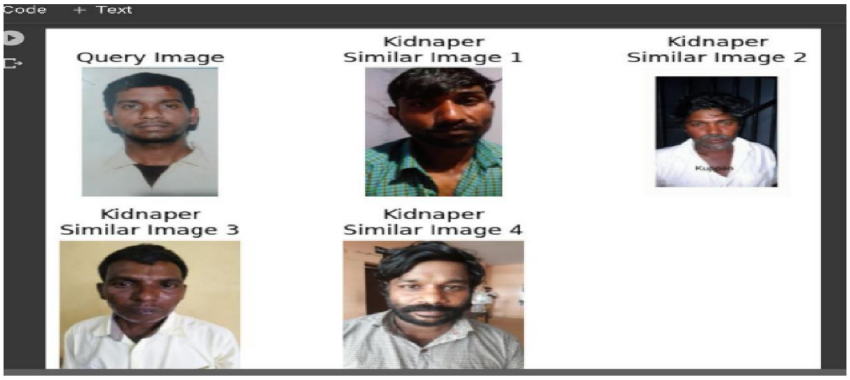


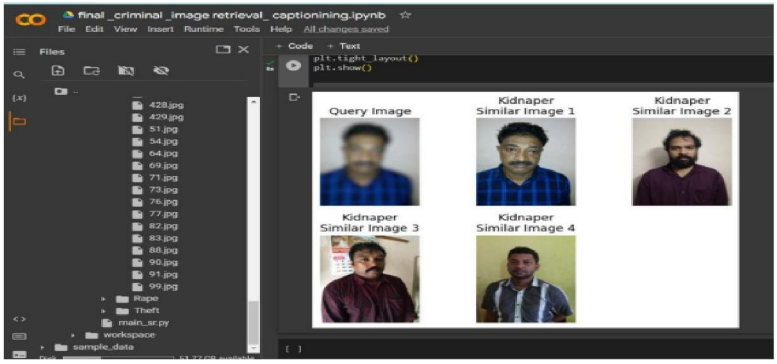

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, September 2023




**2. KIDNAPER**

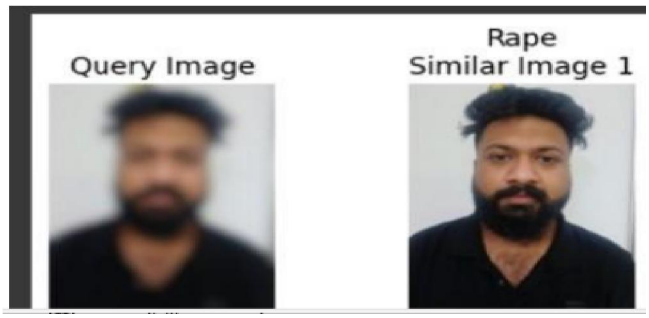


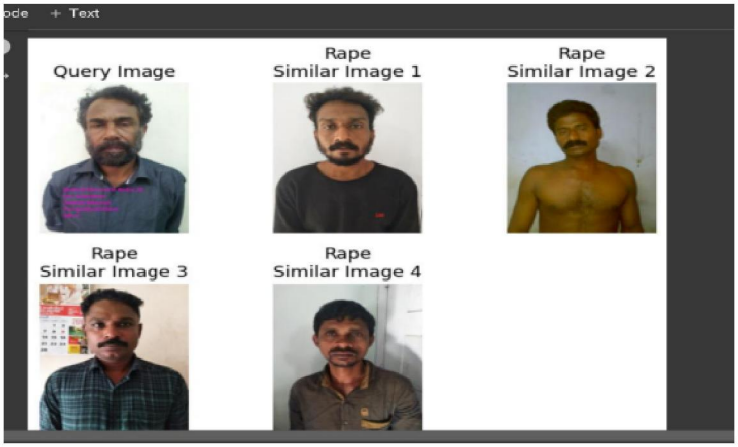



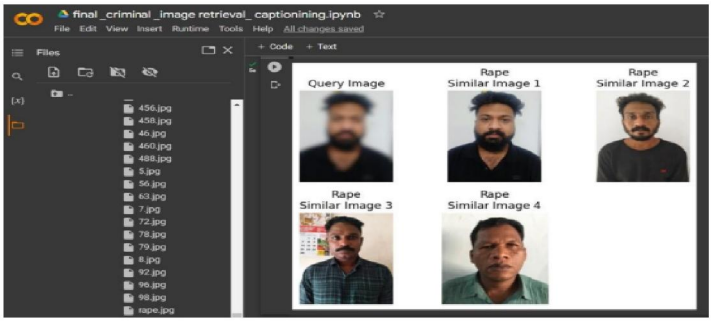



IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, September 2023

**3. RAPE**

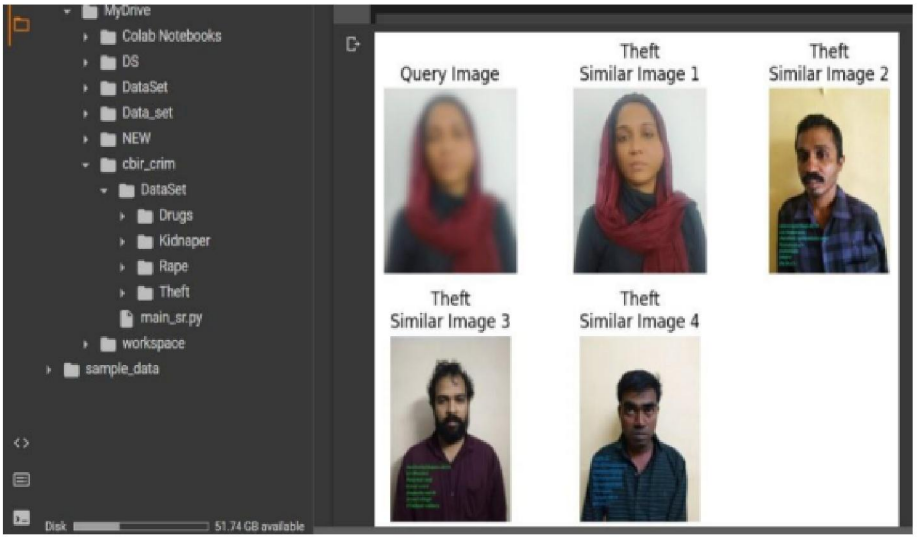


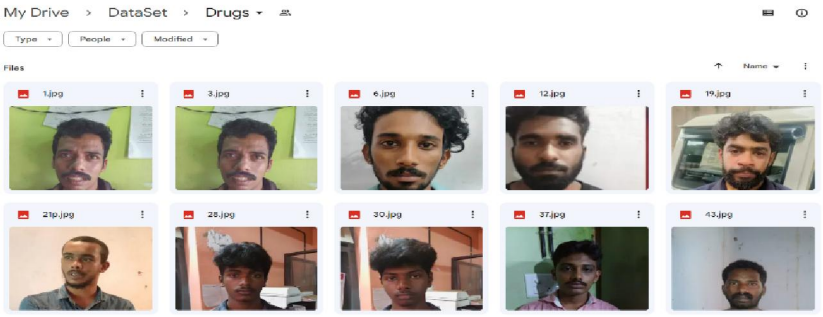

**4. THEFT**

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, September 2023

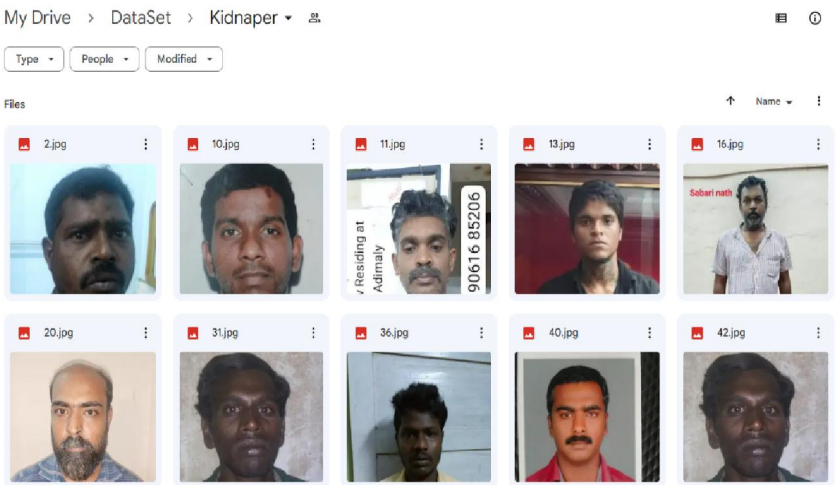

## VII. DATASET

### 1. DRUGS

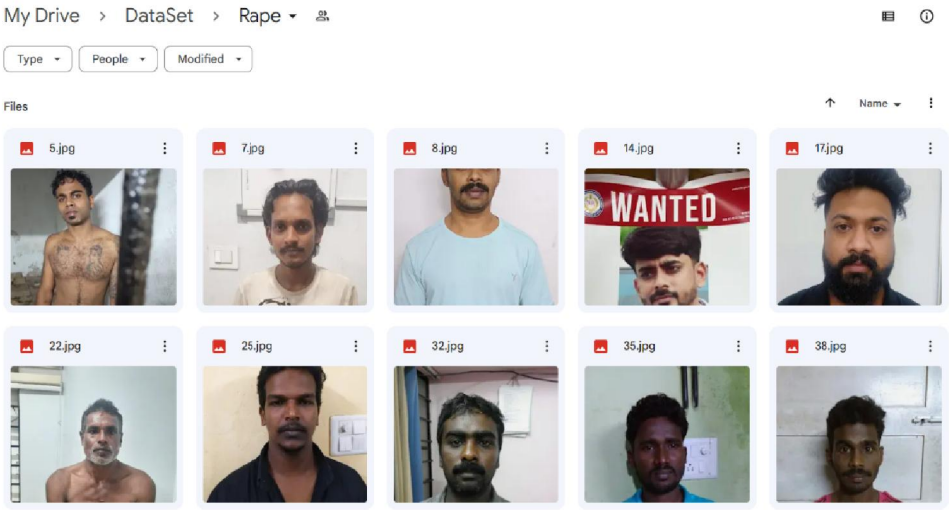


IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, September 2023

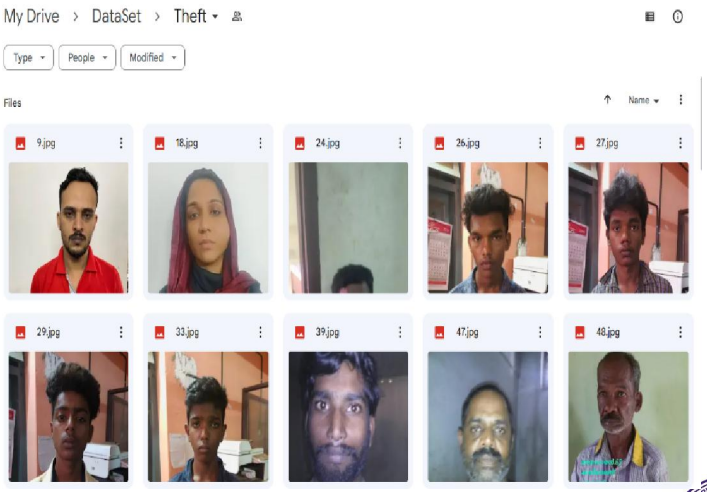
**2. KIDNAPERS**




**3. RAPE**



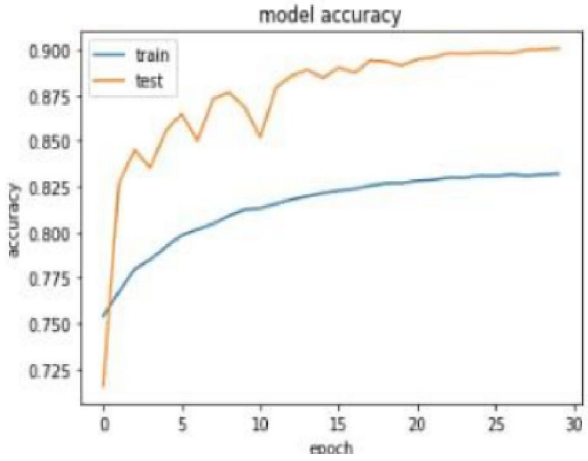

**4. THEFT**



IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, September 2023

## VIII. ACUURACY GRAPH

Evaluating the model accuracy is an essential part of the process in creating machine learning models to describe how well the model is performing in its predictions. Evaluation metrics change according to the problem type. In this post, we'll briefly learn how to check the accuracy of the regression model in R. The linear model (regression) can be a typical example of this type of problem, and the main characteristic of the regression problem is that the targets of a dataset contain the real numbers only. The errors represent how much the model is making mistakes in its prediction. The basic concept of accuracy evaluation is to compare the original target with the predicted one according to certain metrics. Accuracy of the model changes during the training process. The x-axis represents the number of training epochs (or iterations), and the y-axis shows the corresponding accuracy achieved by the model on either the training set or the validation set.




## IX. LIMITATIONS

Generalization issues in machine learning refer to the model's ability to perform well on unseen or new data that was not part of the training dataset. One significant limitation associated with generalization is overfitting, where a model learns to memorize the training data rather than capturing underlying patterns, leading to poor performance on new data. On the other hand, underfitting occurs when a model is too simplistic and fails to capture even the training data's patterns, resulting in poor performance overall. Striking the right balance between these extremes can be challenging, and it often requires careful model selection, hyperparameter tuning, and the availability of diverse and representative training data to mitigate generalization issues and ensure a machine learning model's robustness and effectiveness in real-world scenarios.

## X. FUTURE SCOPE

Future enhancements for content-based image retrieval and captioning using VGG19 and ResNet can focus on improving the models' accuracy, efficiency, and contextual understanding. Integrating attention mechanisms in the image captioning model can enable better focus on relevant image regions during caption generation. Fusion of modalities, combining image features with textual information, can enhance retrieval accuracy by capturing the semantic relationship between images and queries. End-to-end training of the entire system can optimize feature extraction and retrieval/captioning jointly for improved representations. Incorporating transformerbased architectures can enhance language generation in image captioning, capturing long-range dependencies between image features and captions. Exploring novel finetuning strategies for the pre-trained CNNs can tailor the models specifically for contentbased tasks. Integration of semantic understanding, such as object detection or segmentation, can enhance object identification and description. Utilizing larger and diverse datasets can lead to more robust models, while optimizing for real-time performance can make the systems more practical and efficient for real-world applications. These enhancements will make content-based image retrieval and captioning using VGG19 and ResNet more accurate, contextually aware, and valuable in various domains such as multimedia content management and assistive

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, September 2023

technologies. body temperature or sweat, can be used to simulate the physical experience of wearing the clothes. Furthermore, our proposed method can be extended to incorporate other types of clothing-related tasks, such as clothing segmentation or attribute prediction. Clothing segmentation can improve the accuracy of the virtual try-on simulations by better separating the clothing items from the background, while attribute prediction can provide more detailed information about the clothing items, such as color or texture. Lastly, our proposed method can be integrated into existing e-commerce platforms to improve the online shopping experience. This can include developing a user-friendly interface for the virtual try-on simulations, integrating with social media platforms for sharing and recommendations, and enabling personalized recommendations based on user preferences and previous purchases.

## XI. CONCLUSION

In conclusion, content-based image retrieval and captioning using VGG19 and ResNet are powerful techniques in the realm of computer vision and natural language processing. VGG19, with its deep architecture, excels in extracting high-level visual features, enabling accurate and efficient image retrieval based on visual content. On the other hand, ResNet, known for its residual learning, is adept at image captioning, generating descriptive and contextually relevant captions for images. Leveraging pretrained CNNs like VGG19 and ResNet50 allows for effective transfer learning, enhancing the models' performance on various tasks. Future enhancements focusing on attention mechanisms, fusion of modalities, end-to-end training, and semantic understanding hold the potential to further improve the models' accuracy and contextual understanding. As these techniques evolve, content-based image retrieval and captioning systems using VGG19 and ResNet are poised to play pivotal roles in applications such as image organization, accessibility, e-commerce, and multimedia content management, advancing the fields of computer vision and natural language processing.

## REFERENCES

[1] https://www.tutorialspoint.com/python/index.html
[2] https://www.tutorialpoint.com/uml/uml_use_case_diagram.html
[3]https://www.researchgate.net/publication/3600565897_Performance_Investigation _of_a_Proposed_CBIR_Search_Engine_Using_Deep_Cnvolutional_Neural_ Networks
[4]https://www.researchgate.net/publication/322990964_Content_based_image_retrieval_using_deep_learning_process
[5] Socratis Gkelios, Yiannis Boutalis, Savvas A.Chatzichristofis, “Investigating the Vision Transformer Model for Image Retrieval Tasks” 2021
[6] Arshiya Simran,P.S Shijin Kumar,Srinivas Bachu, “Content Based Image Retrieval Using Deep Learning Convolutional Neural Network”