***INDIAN INSTITUTE OF INFORMATION TECHNOLOGY ALLAHABAD***



*VI Semester B.Tech, Department of Information Technology*

*Indian Institute of*  *Information Technology, Allahabad*

# Unsupervised Image Retrieval

DEL Group 13 Members: -

Pushkal Madaan IIT2020005

Ritej Dhamala IIT2020006

Jitesh Kumar IIT2020224

Rahul IIT2020244

## Table of Contents

## 

| **Sl.** | **Topic** | **Page** |
| --- | --- | --- |
| 1 | Abstract | 2 |
| 2 | Introduction | 2 |
| 3 | Literature review | 3 |
| 4 | Dataset description | 6 |
| 5 | Methodology | 7 |
| 6 | Implementation | 10 |
| 7 | Result and Discussion | 13 |
| 8 | Conclusion and Future Work | 15 |
| 9 | References | 16 |

**Abstract**

In this project, we have implemented an unsupervised image retrieval system with the help of the CNN autoencoder, RESNET, and CBIR (Content-Based Image Retrieval) framework. The model design consists of an autoencoder that converts the input data into a new representation in the form of feature vectors from which we evaluate the similarity between images using the KNN (K-Nearest Neighbour) method. Images that are closer to one another in the latent space resemble one another more than those that are further apart. Images then selected are transformed into their original representations from the decoder part of the autoencoder.

**Introduction**

Unsupervised image retrieval is a method for searching and obtaining photos from a sizable database without the need for the images to be manually annotated or labeled. Unsupervised image retrieval makes use of algorithms that can automatically detect patterns and similarities in the photos, as opposed to supervised image retrieval, which trains the system using pre-labeled data. This makes it a more adaptable and scalable method for retrieving photographs, particularly in scenarios where there are significant amounts of unlabelled images. There are several uses for unsupervised image retrieval, including in the fields of computer vision, picture recognition, and content-based image retrieval. In this essay, we will examine the ideas and methods behind unsupervised image retrieval, as well as some of its uses and difficulties. By using visual similarity or other criteria, image retrieval seeks to empower people to locate images that correspond to their interests or requirements. The collection of photos is processed to extract low-level visual features using a feature extraction module. The main Image attributes which are retrieved are color, texture, and shape.

The two main activities in a CBIR system are feature extraction and feature matching.

The primary requirement of CBIR is that a query image be provided as input. It then compares the visual contents of the query image with the photos stored in the archive, and close visual similarity in terms of image feature vectors serves as a foundation for locating images with comparable contents. However, there are still several issues with it, including processing massive datasets, handling variations in images caused by changes in viewpoint, illumination, and occlusion, and developing user-friendly interfaces for defining searches.

Applications for content-based picture retrieval include video surveillance, medical image analysis, and image search engines.

**Literature review:**

| **Title** | **Authors** | **Year** | **Methodology** | **Results** | **Dataset** |
| --- | --- | --- | --- | --- | --- |
| **An unsupervised learning approach to content-based image retrieval** | Yixin Chen,  James Z. Wang, and Robert Krovetz | 2003 | This work offers a unique image retrieval method called CLUster-based retrieval of pictures by unsupervised learning (CLUE) to address the semantic gap issue. The foundation of CLUE is the idea that images with similar semantic properties tend to cluster together. By returning picture clusters based on both the similarities between individual photos and the query's features, it aims to close the semantic gap between them. | The experimental system uses feature extraction and UFM similarity measure. A general-purpose image database (from COREL) with roughly 60,000 images is used to create the system. Using pairwise distances, CLUE divides the images into clusters where within-cluster similarity is the high and between-cluster similarity is low.  The outcomes seem to suggest that unsupervised learning can group together photos with comparable semantic properties to some extent. | general-purpose images |
| **CLUE: cluster-based retrieval of images by unsupervised learning** | Yixin Chen, Member, IEEE, James Z. Wang, Member, IEEE, and Robert Krovetz | 2005 | In order to enhance user involvement with image retrieval systems by fully utilizing similarity information, this work introduces a novel technique called cluster-based retrieval of pictures by unsupervised learning (CLUE).  Using a graph-theoretic clustering algorithm on a set of photographs close to the query, CLUE is able to retrieve image clusters.  Any metric or nonmetric real-valued symmetric similarity measure can be paired with CLUE. As a result, it might be included in a number of current CBIR systems, such as relevance feedback systems. | A 1000-image database numerical tests reveal enhanced retrieval accuracy and good cluster quality.  Additionally, Google's Image Search results for photos point to the possibility of using CLUE to analyze real-world image data and incorporating CLUE into the user interface for keyword-based image retrieval systems. | COREL |
| **Classiﬁcation Using K Nearest Neighbor for Brain Image Retrieval** | [N. Hema Rajini](https://ieeexplore.ieee.org/author/37706098600)  [R. Bhavani](https://ieeexplore.ieee.org/author/37706101900) | 2011 | In this paper, the authors propose the algorithm for the retrieval of the most visually similar images to a given query image from a database of medical images In this approach, shape feature extraction is accomplished using clever Edge detection, and texture feature extraction is accomplished using the Gabor filter. Using the KNN, medical images were categorized based on these features.  The model was implemented using MATLAB | Results of experimentation show that the texture features mentioned in this paper can fully describe the content of the image, and improve the recall and precision rate and classiﬁcation accuracy of medical image retrieval.  From observing the table depicting the retrieval performance of Canny Edge Detection, Gabor Filter, and Combining both Methods, it can be seen that Average precision gradually decreases as the number of retrieved images increases. | Unpaired MR-CT brain dataset |
| **Unsupervised Image Retrieval with Mask-based Prominent Feature Accumulation** | Xinyi Wang, Yajing Xu, Haitao Yang, Si Li | 2019 | The authors use a mask-based feature accumulation method that identifies the most prominent features of an image based on the frequency of appearance of those features across a set of images. The method involves generating a set of masks that segment an image into regions, and then accumulating features (in the form of activations of convolutional neural network layers) within each region across a set of images. The resulting feature vectors for each region are then clustered to identify the most prominent features. | The authors evaluate their method on several benchmark datasets, including CIFAR-10 and Caltech-101, and compare it to several state-of-the-art unsupervised and supervised image retrieval methods. They find that their method achieves competitive results compared to supervised methods, and outperforms other unsupervised methods. | CIFAR10  Caltech101  and Oxford-102 |

**Dataset Description:**

Labelled Faces in the Wild (LFW) Dataset

A database of face images called Labeled Faces in the Wild (LFW) was created to examine the issue of unrestricted face identification. University of Massachusetts, Amherst scholars built and maintain this database (specific references are in the Reference section). The collection of these images was done using the OpenCV implementation of the Viola-Jones face detector which located 5,749 people in 13,233 photos that were gathered from the internet. Each face has the names of the people it represents written on it. There were two or more photos of 1680 of the people who were portrayed in the data set. This is contained in one file out of a total of 11. The remaining 10 files all contain essential metadata that we will use to create the training and test sets for our model.

There are some flaws in this dataset; In LFW, a lot of groups are underrepresented. For instance, there are comparatively few women, few persons above the age of 80, few children, and no babies. A lot of ethnicities are also either completely overseen or barely represented at all. Additionally, circumstances like bad lighting, extreme/odd positions, heavy occlusions, low resolution, and other potentially significant elements do not make up a major portion of the LFW dataset. These are crucial topics for evaluation, particularly for image recognition systems, hence we cannot create a thoroughly vetted system of image recognition and retrieval.

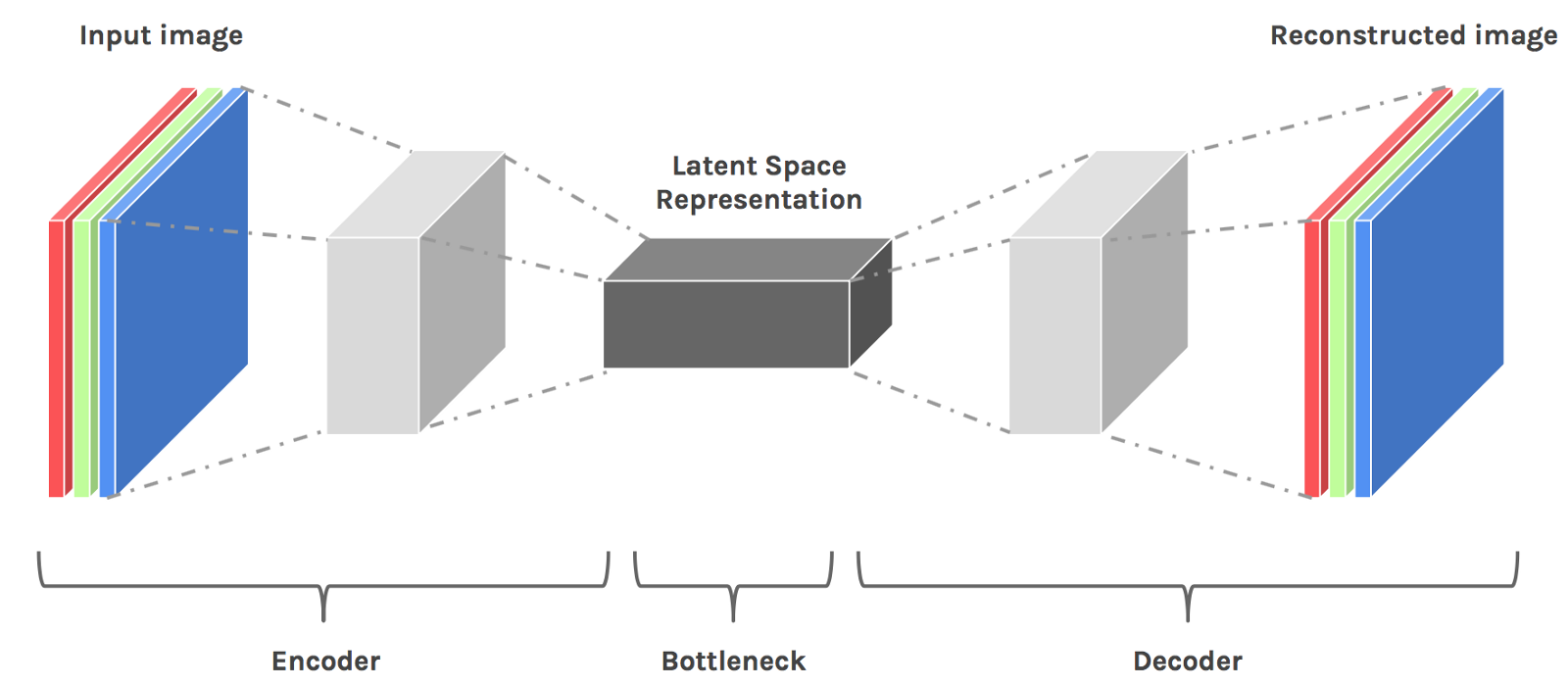
The images present in the dataset are in .jpg format and the image dimensions are 250x250 pixels.

Dataset link: [Labeled Faces in the Wild](http://vis-www.cs.umass.edu/lfw/)

**Methodology**

**Autoencoder**

The autoencoder is made up of two functions, the encoder function, which transforms input data into a new representation, and the decoder function, which returns the representation to the original domain.



**source**: [Towards Representation Learning for Image Retrieval](https://hackernoon.com/hn-images/1*op0VO_QK4vMtCnXtmigDhA.png)

**Encoder**

The encoder portion is rather conventional; to obtain a representation of the desired size, we stack convolutional and pooling layers before finishing with a dense layer. For both convolutional and dense layers, we have used activation='elu'. With kernel size (3, 3), padding=same, and 32, 64, 128, 256 output channels, we must repeat (conv, pool) four times. Finally, flatten the output before adding the last dense layer.

**Decoder**

We utilized "transpose convolution" as a decoder. A patch of an image is used by a traditional convolutional layer to create a number (patch -> number). The goal of "transpose convolution" is to take a number and turn it into an image patch (number -> patch). To "undo" convolutions in the encoder, we require this layer. In order to "undo" the final layer of the encoder, our decoder begins with a dense layer. and modify its output to "undo" the encoder's flattening.

We are now prepared to undo pairs of (conv, pool). This requires stacking four transpose convolution layers with the following output channel counts 128, 64, 32, and three. Each of these layers will develop the ability to "undo" the pair (conv, pool) in the encoder.

For the last layer, we have used removed activation because that is our final image.

**CNN**

CNNs employ a unique architecture created to take advantage of the spatial structure of pictures. Convolutional layers, pooling layers, and completely linked layers are just a few of the layers that they frequently have. Pooling layers minimize the spatial dimensions of the feature maps to improve the network's efficiency whereas convolutional layers utilize filters to extract local information from the input picture. Following the convolutional and pooling layers, fully connected layers are often utilized for classification or regression applications.

**Batch Normalization**

Convolutional neural networks (CNNs) employ batch normalization as a method to increase the model's training speed, stability, and generalization. It entails dividing by the batch's standard deviation and removing the mean to normalize the activations of each layer in the network.

**RESNET50V2**

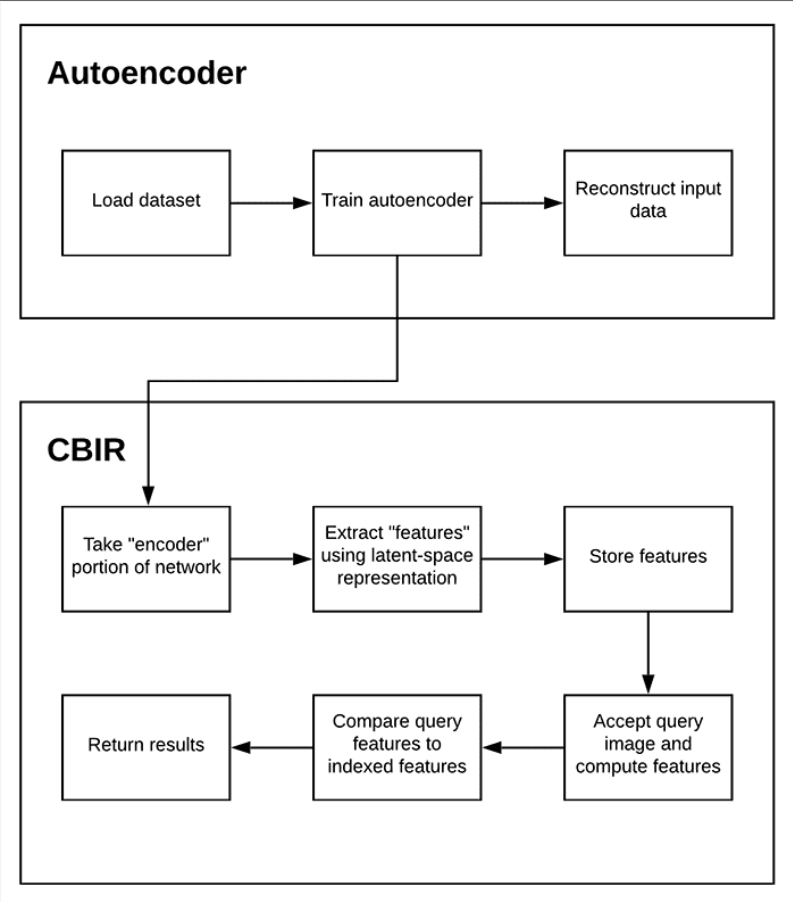
Microsoft Research Asia created the ResNet50V2 convolutional neural network architecture in 2016 as an upgrade to the existing ResNet50 design. The 50-layer ResNet50V2 makes use of residual connections to solve the vanishing gradient issue that frequently affects deep neural networks. The network can learn the residual mapping across layers thanks to the residual connections, which makes it simpler to train deeper networks without suffering performance losses.

**CBIR**

Each image in the database is subjected to a series of low-level characteristics extraction using CBIR algorithms, including color, texture, and form. A high-dimensional feature vector that reflects the visual content of the picture is then created using these features. A similarity metric, such as cosine similarity or Euclidean distance, is used when a query picture is supplied to compare its attributes to those of every image in the database. The retrieval procedure's outcomes are the pictures that look the most like the query picture.

**KNN**

Every picture is changed into a feature vector, which functions as a concise representation of its visual elements. To evaluate the visual similarity between images, we can determine the distance between their corresponding feature vectors. Images that are nearer to each other in the latent space are more similar than those farther away. This feature can be utilized for picture retrieval, where we can retrieve the ‘K’ number (3 in our use case) of most similar images from a collection by searching for their nearest feature vectors in the latent space, given a query image using the K Nearest Neighbour algorithm.

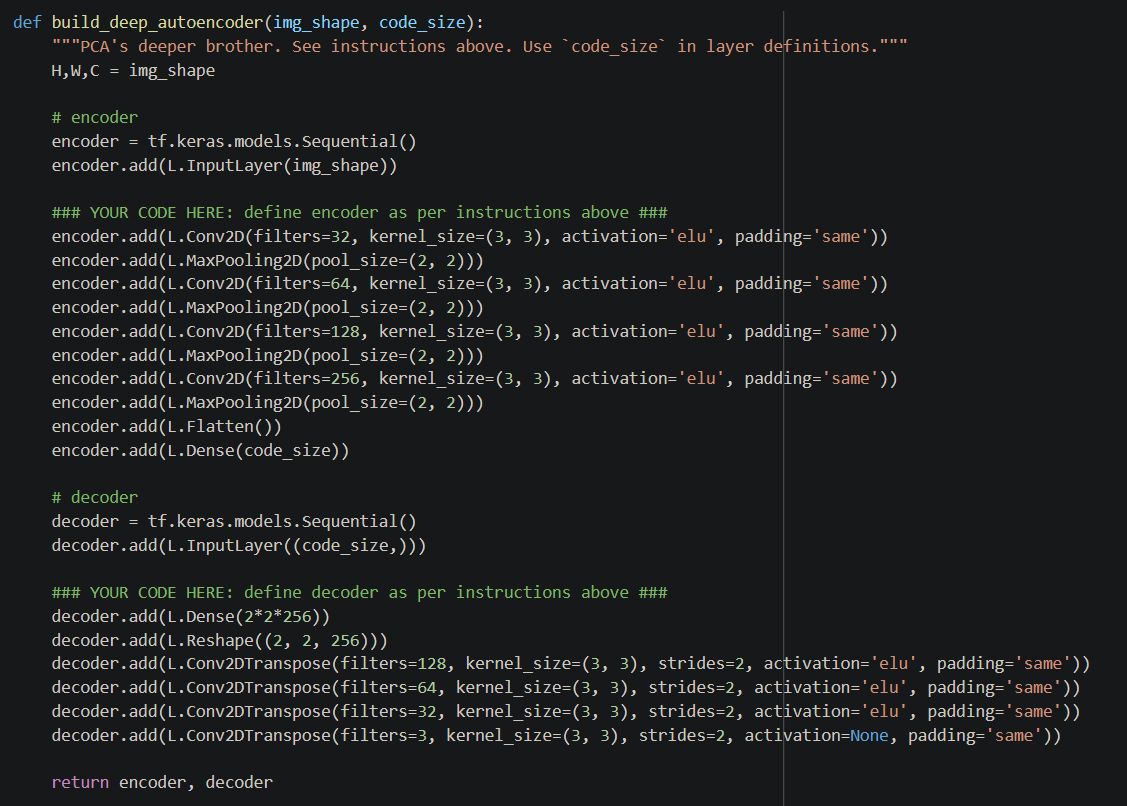


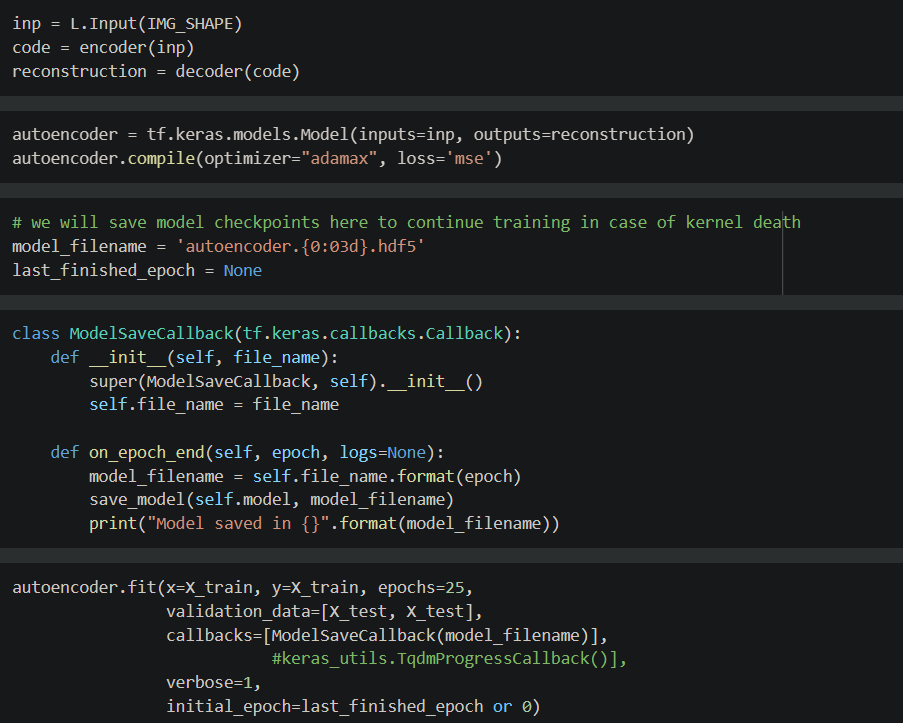
**Source**: [Autoencoders for Content-based Image Retrieval](https://b2633864.smushcdn.com/2633864/wp-content/uploads/2020/03/keras_autoencoder_steps.png?lossy=1&strip=1&webp=1)

**Implementation**



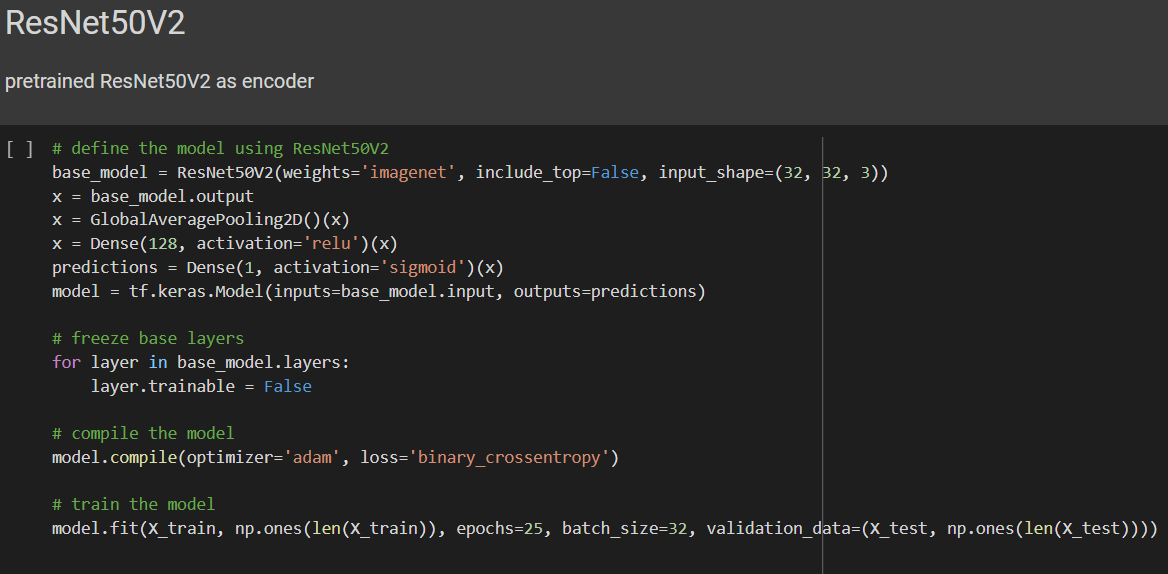
**CNN Autoencoder**



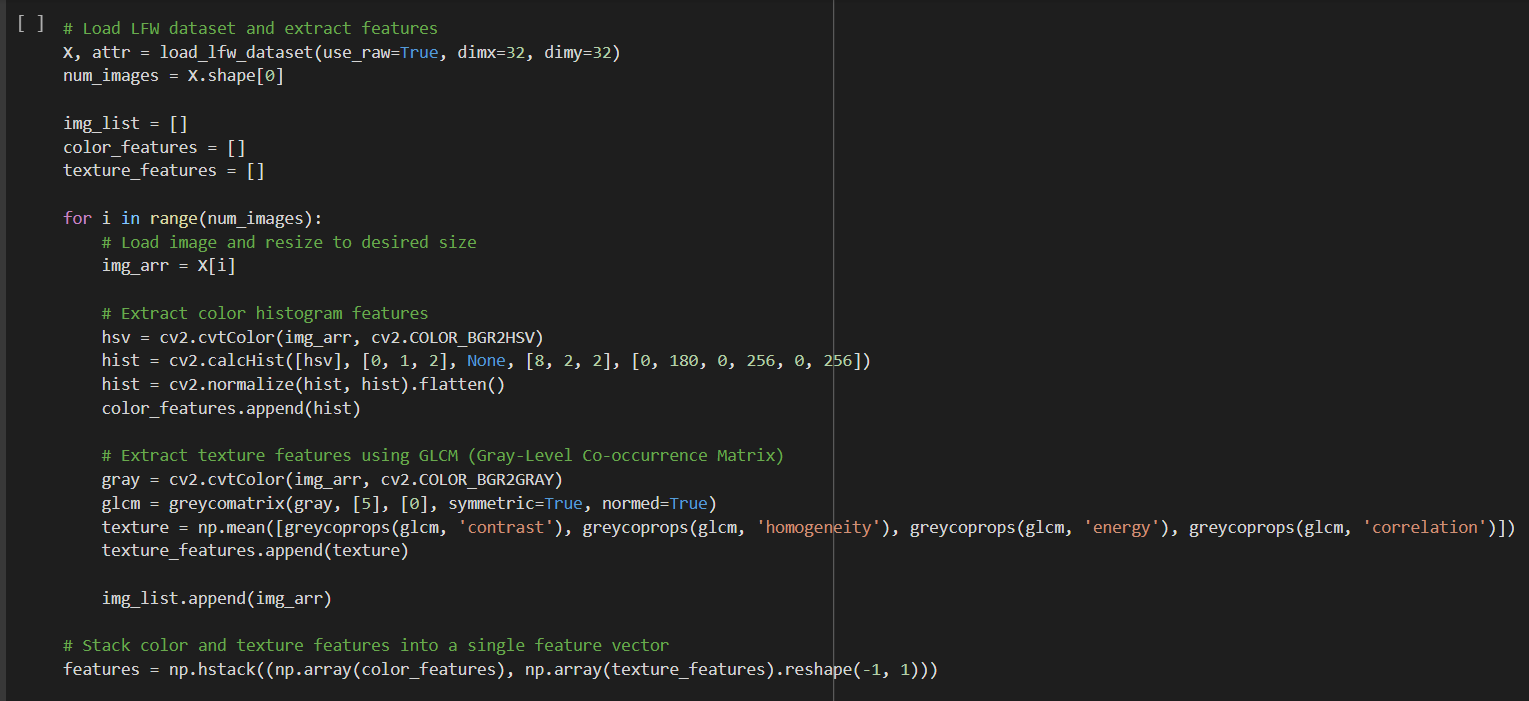




**RESNET50V2**

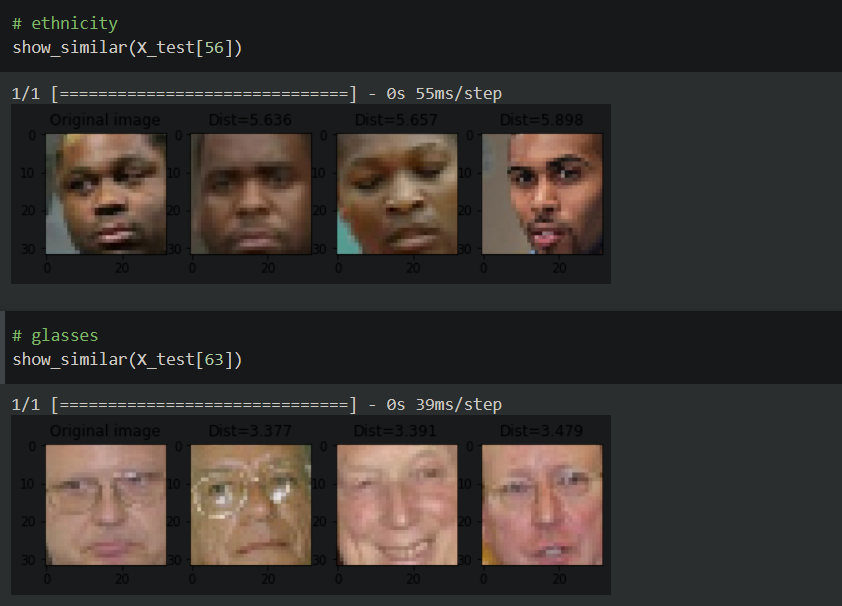
****

**CBIR**

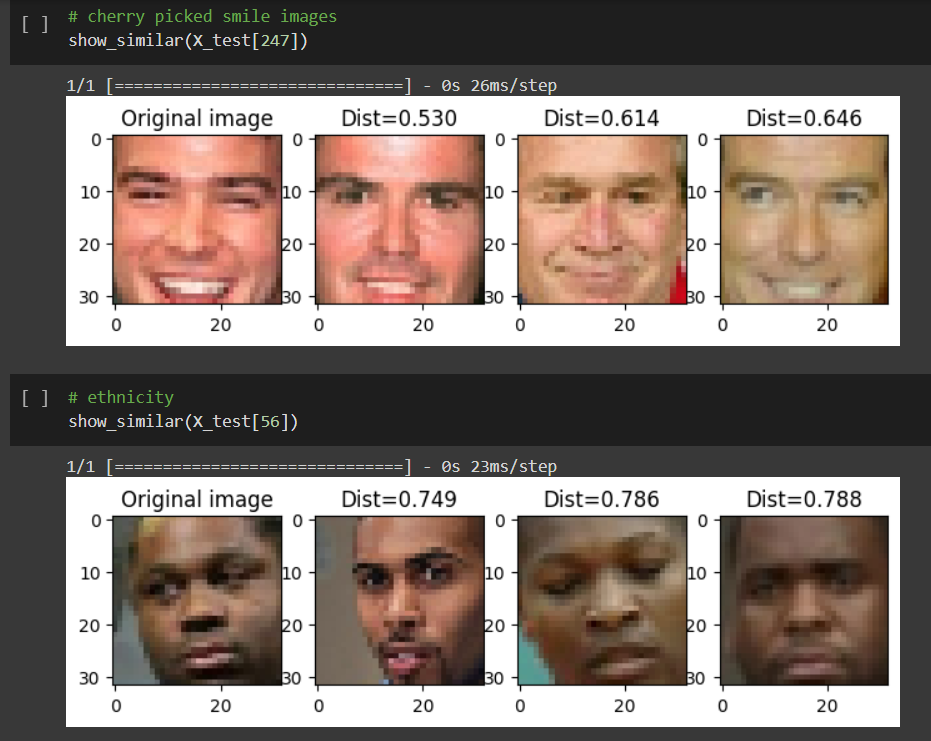
****

**Result**

**CNN Autoencoder**



**CNN with Batchnormalization**

****

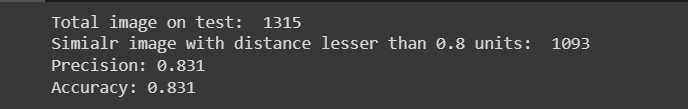
**RESNET50V2**

****

**CBIR**

****

**Performance metric for best performing model (CNN+BN)**



**Future Work**

**Cross-modal retrieval**: Unsupervised image retrieval can be extended to cross-modal retrieval, where the goal is to retrieve images given a text query or retrieve text given an image query. This is a challenging task, as it requires learning a joint embedding space for images and text.

**Transfer learning:** Transfer learning can be used to improve the performance of unsupervised image retrieval models by leveraging pre-trained models on large-scale labeled datasets. This can help to learn more discriminative visual representations and improve the retrieval performance

**Conclusion**

Our implementation of an unsupervised image retrieval system based on the CBIR framework has proven successful. By using an autoencoder to create feature vectors and the KNN method to evaluate similarities, we were able to retrieve images that closely resemble the query image from a large dataset. This system can be useful in many applications where image retrieval is essential, such as in medical image analysis or e-commerce websites. With further optimization and integration with other techniques, this approach can be extended to more complex image retrieval tasks, making it a valuable tool for the research community and industry.

**References**

[1] Yixin Chen, J.Z. Wang and R. Krovetz. CLUE: cluster-based retrieval of images by unsupervised learning. 2005

[2] Y. Chen, J.Z. Wang and R. Krovetz . An unsupervised learning approach to content-based image retrieval. 2003

[3] K Venkataravana Nayak, J Arunalatha and K Venugopal. IR-FF-kNN: Image Retrieval Using Feature Fusion with k-Nearest Neighbour Classifier

[4] P.A.Charde and S.D.Lokhande. Classification Using K Nearest Neighbor for Brain Image Retrieval. 2011

[5] Xinyi Wang, Yajing Xu, Haitao Yang, Si Li. Unsupervised Image Retrieval with Mask-based Prominent Feature Accumulation. 2019

[6] R. Zhang and Z. Zhang, "Effective Image Retrieval Based on Hidden Concept Discovery in Image Database," in IEEE Transactions on Image Processing, vol. 16, no. 2, pp. 562-572, Feb. 2007, doi: 10.1109/TIP.2006.888350.

[7] G. Carneiro, A. B. Chan, P. J. Moreno and N. Vasconcelos, "Supervised Learning of Semantic Classes for Image Annotation and Retrieval," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 3, pp. 394-410, March 2007, doi: 10.1109/TPAMI.2007.61.

[8] B. Wei, R. Jia, S. Fu, X. Lyu, L. Gao and Z. Tang, "MIR: A Benchmark for Molecular Image Retrival with a Cross-modal Pretraining Framework," 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Las Vegas, NV, USA, 2022, pp. 3902-3904, doi: 10.1109/BIBM55620.2022.9995641.

[9] M. Kaur and N. Sohi, "A novel technique for content based image retrieval using color, texture and edge features," 2016 International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2016, pp. 1-7, doi: 10.1109/CESYS.2016.7889955.

[10] E. Cheng, F. Jing, C. Zhang and L. Zhang, "Search Result Clustering Based Relevance Feedback for Web Image Retrival," 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, Honolulu, HI, USA, 2007, pp. I-961-I-964, doi: 10.1109/ICASSP.2007.366069.