Analysis of Content Based Image Retrieval using Deep Feature Extraction and Similarity Matching

Anu Mathews[1], Sejal N[2], Venugopal K R[3]
Dept. of AI&ML, K S Institute of Technology, Bengaluru, India[1]
Dept. of AI&ML, BNM Institute of Technology, Bengaluru, India[2]
Bangalore University, Bengaluru, India[3]

***Abstract*—Image retrieval using a textual query becomes a major challenge mainly due to human perception subjectivity and the impreciseness of image annotations. These drawbacks can be overcome by focusing on the content of images rather than on the textual descriptions of images. Traditional feature extraction techniques demand for expert knowledge to select the limited feature types and are also sensitive to changing imaging conditions. Deep feature extraction using Convolutional Neural Network (CNN) are a solution to these drawbacks as they can learn the feature representations automatically. This work carries out a detailed performance comparison of various pre-trained models of CNN in feature extraction. Features are extracted from men footwear and women clothing datasets using the VGG16, VGG19, InceptionV3, Xception and ResNet50 models. Further, these extracted features are used for classification using SVM, Random Forest and K-Nearest Neighbors classifiers. Results of feature extraction and image retrieval show that VGG19, Inception and Xception features perform well with feature extraction, achieving a good image classification accuracy of 97.5%. These results are further justified by performing a comparison of image retrieval efficiency, with the extracted features and similarity metrics. This work also compares the accuracy obtained by features extracted by the selected pre-trained CNN models with the results obtained using conventional classification techniques on CIFAR 10 dataset. The features extracted using CNN can be used in image-based systems like recommender systems, where images have to be analyzed to generate item profiles.**

Keywords*—Convolutional neural network; deep learning; feature extraction; accuracy; similarity

I. INTRODUCTION

Image retrieval systems browse, search and retrieve visually similar images from large image databases. Traditional image retrieval methods utilize the image annotations for obtaining the metadata that help in finding the similar images, but this is a laborious process and is subjective to human perceptions. Efficient image retrieval is the backbone of most of the search engines and recommender systems. Search engines may not be able to retrieve relevant information according to user's preferences due to imperfect textual query, less knowledge about the search query or due to wrong tagging of the database images. This gap between real intention of a user's search and his understanding of the object is called as semantic gap [1]. With the large amount of image data encountered in social networks, online e-commerce website, medical imaging, etc., it is a challenging problem to search the humongous databases for similar images, especially when it comes to real time image retrieval. Feature extraction techniques help in getting a representation of the attributes of an image, which gives information about the image contents and hence helps in efficient retrieval of visually similar images. Feature extraction transforms data into more informative forms with efficient representation for analysis and classification [2]. Advantages of feature extraction include:

- Reduction of redundancy.
- Reduction in the number of processing resources.
- Helps in avoiding overfitting problem in machine learning models.
- Increase in accuracy.

Motivation: Any image recognition or image retrieval task requires a good feature representation in order to achieve high performance. However, since it is not feasible to define a good feature set manually, feature extraction plays a major role in image retrieval tasks. Features obtained using traditional methods of feature extraction are not capable of expressing the semantic information of the image. Deep Learning techniques make the task of feature extraction automatic and more efficient. Since deep learning techniques require a large amount of labelled training data, transfer learning can be used to reduce this overhead [3]. Fine tuning a CNN allows pretrained models to be used in a new task with a new dataset.

CNN models have two stages, of which, the first is feature extraction and the second is classification. This paper focusses on feature extraction using pretrained CNN and comparing their performance.

The open research questions which are addressed in this work include:

i) How effective are deep learning methods for learning good feature representations from images for Content Based Image Retrieval?

ii) How do the feature extraction and image retrieval time vary with the different models?

A. Contributions:

- This work contributes to the existing knowledge of feature learning by exploring feature extraction and transfer learning techniques using the five selected pre-trained CNN architectures.

- Comparison of the goodness of the features extracted by the selected CNN models by comparing the image retrieval performance using similar image search technique and also by using the extracted features for image classification.

The following sections of this paper are organized as follows. Related work in the domain of feature extraction and image retrieval is discussed in Section II. The fundamental concepts in image retrieval and the methodology adopted are given in Section III. Section IV gives the framework and the algorithm used in Reverse Image Search. Section V discusses the implementation details and results in Feature Visualization and Feature Extraction. Section VI concludes the paper by discussing the open research issues in the field of image retrieval and deriving an outlook for further research.

II. Related Work

The goodness of an effective image retrieval system is attributed to the effectiveness of image feature extraction techniques used by the system. Various techniques have been proposed for feature extraction, starting from extraction of handcrafted features using traditional feature extraction techniques like SIFT, SURF, HOG to recent techniques utilizing Deep Learning. Work proposed in [4] investigated the use of various color, texture and shape features for image retrieval in CBIR (Content Based Image Retrieval) and biometrics systems, by mapping the image content onto low-level features.

An integrated image representation model has been proposed by complementary fusion of SIFT and CNN, to describe contents at multiple levels in images, followed by a series of operations of L2 normalization, PCA and whitening, on the integrated representation for a more compact representation [5].

The various techniques for extraction of texture and color features include color correlogram, color histogram, color co-occurrence matrix, wavelet transform and Tamura feature [6]. Several recent works have focused on the use of machine learning and Deep Learning for feature extraction [7-9]. The convolution layers in CNN can change the shape of the output, thereby enabling the learning of basic object shapes in the primary layers and more complex objects in the deeper layers, with a drastic reduction in the error rate [10].

Studies reveal that high level features are of different behavior from low and middle level features under certain conditions, and hence, CNN features in each of the layers can be utilized for representation of knowledge. The level of noise included in these features have been studied and a thresholding approach is proposed to remove as much noise as possible, thereby generating efficient CNN embedding spaces [11].

Some of the models with combination of CNN features and handcrafted images have also shown good results. The work in [12] automates the extraction of electroencephalogram (EEG) signals the feature extraction methods of Wavelet Packet decomposition and Genetic Algorithm-Based Frequency-Domain Feature Search. A combination of handcrafted and CNN features has been used for liver MR image adequacy assessment yielding a model performance of AUC = 0.94, wherein, HC features of intensity values, topological structure, texture information (GMM, ECC, and GLCM features) are extracted from the images and used for classification using Random Forest classifier.

Many algorithms use Hashing based methods for image retrieval. While some of the algorithms identified approximate regions of the objects in an image using attention sub-network, focusing mainly on the foreground objects [13], image features extracted using deep hashing methods with deep neural networks (DNN) gave better image retrieval performance [14].

Histogram features have also given good results in CBIR. Work proposed in [15] used color spectral histogram for computing the similarity between images, whereas the method proposed in [16] used correlated primary visual texton histogram features.

The use of color and texture features have been proposed in several works. The work proposed in [17], calculates the image similarity by giving equal priority to dominant color Hu moments for CBIR, wherein, the texton template detects and extracts the consistent zone of an image and uses the seed point's selection approach for initiating the clustering process. Study [18] uses both Hu moments and dominant color descriptor on the pixels. Research [19] uses LBP-DBN training algorithm with adaptive momentum learning rate to overcome the issues of longer training time and reduction in classification accuracy. The author in [20] uses Directional Magnitude Local Hexadecimal Patterns, for image retrieval, by reducing the semantic gap problem using a learning- based approach. Quadruplet loss function and feature fusion are also used for clothing retrieval [21]. The features extracted using Resnet-50 are merged with middle-level features to get a combined feature representation. Work proposed in [22] uses a visual re-ranking approach, by using a correlation matrix of an image retrieval list and a CNN model learns the relevance of each of the image pairs simultaneously. The model is further optimized using a weighted MSE loss, which also considers the sparsity of labels. Feature reduction using an improved CNN and PCA quadratic dimensionality reduction has been proposed [23]. Deep Learning techniques for feature extraction are more promising due to their capability to extract and learn features of an image automatically.

Recent advances in image retrieval have started focusing on graph neural networks for CBIR. The work done in [24] proposes graph neural networks to reframe the re-ranking process for CBIR by using a 2-layer GNN to aggregate neighbor information of the entire data. Several new approaches have been proposed for deep feature embedding. In [25] a GNN characterizes and predicts the local correlation structure of images in the feature space, using which, neighboring images collaborate and refine their feature embeddings based on local linear combination. The representation capability of graphs has led to various graph-based image recommendation systems also [26]. Table I gives a performance comparison of some of the results obtained in CBIR.

TABLE I. PPERFORMANCE COMPARISON OF REVIEWED CBIR SYSTEMS

Author	*Method*	*Dataset*	*Performance*
Salamh et al. [7]	Multi-descriptor-local binary patterns	Extended Yale B & Grimace database	Accuracy=99.4%
Alsuwaiket et al.[9]	Wavelet Packet decomposition (WPD), Genetic Algorithm-Based Frequency-Domain Feature Search (GAFDS)	Children's Hospital Boston: EEG recordings of 23 cases	CNN: 97.93% accuracy SVM: 94.49% accuracy RF : 88.03% accuracy
Manjunath et al. [10]	AlexNet inspired CNN	CIFAR-10	KNN:28.2% accuracy SVM:37.4% accuracy CNN: 85.97% accuracy
Y. Zhang et al. [13]	Supervised hashing method which fusions simplex feature similarity & location similarity among multiple objects	VOC 2007 VOC 2012 NUS-WIDE	6% increase on Weighted MAP -3.0% increase on NDCG@1000 -2.9% on ACG@1000
X. Zhang [19]	Classification: Image local feature multi-level clustering and image-class nearest	RPC: Defiance Technology Nanjing Research Institute	Average Classification accuracy=84% as compared to ISDH(Instance Similarity Deep Hashing)

III. FUNDAMENTAL CONCEPTS AND METHODOLOGY

The basic steps involved in image retrieval include extraction of features and computation of image similarity. The features of the collection of database images are extracted and stored. The query image undergoes the feature extraction process and subsequently the similarity matching algorithm computes the similarity between the features of query image and the features of the database images. Distance metrics like Cosine distance and Euclidean distance are used for similarity computation. Finally, top-n similar images are retrieved.

A. CNN Models

Convolutional Neural Networks use supervised learning and train the network by backpropagation. CNNs have the advantage of reduced amount of preprocessing when compared to other deep learning networks. A CNN consists of two basic parts of feature extraction and classification.

The feature extraction part has several convolution layers followed by max pooling layers and an activation function. The classifier part consists of fully connected layers. Normalization layers in CNN help in keeping signals from each layer at suitable levels.

The output size at each of the convolution layer is given by the following formula:

$$\text{Output size} = [(W-F+2P)/S] +1$$

where, W is the size of the input image, F is the size of the receptive field, S is the stride, and P indicates the zero-padding used on the border.

1) ResNet50: The ResNet model was developed by He et al. in the year 2015 with the concept of residual connections [27]. ResNet50 consisting of 50 weighted layers, has five times lesser memory requirement as compared to VGG model because of the usage of Global Average Pooling layer, converting the 2D feature maps of the last layer to an n-classes vector calculating the probability of belonging to each class.

Fig. 1 shows the Residual block which is the basic building block of ResNet model. In a plain CNN network, the input X gets transformed into H(X) by passing through the different layers. In the residual model, the identity connection transforms the input X into F(X)+X, i.e., the original H(x) gets modified into F(X)+X. The ResNet model can learn from much deeper network, as the identity mapping from the input acts as a shortcut for the gradient to pass through, thereby avoiding the vanishing gradient problem.

2) VGG16: VGG 16 architecture proposed by Simonyan and Zisserman in 2014 [28], consists of blocks with increasing number of convolutional layers with 3x3 filters as shown in Fig. 2.

The size of the activation maps reduces by half due to the max-pooling blocks between the convolutional blocks. The classification blocks consist of two layers, each with 4096 neurons, followed by the final output layer of 1000 neurons.

3) VGG19: VGG19 model consists of 16 convolution layers, three Fully connected layers, five Max Pool layers and one SoftMax layer. RGB image of dimension (224 * 224) is given as input and the kernels used are of (3 * 3) size with a stride of 1. Max pooling over a (2 * 2) window with stride 2, followed by Rectified linear unit introduces non-linearity for better classification and improved computation time as compared to the models using sigmoid or tanh functions. Two fully connected layers of size 4096 is followed by a layer with 1000 channels and final layer is softmax function.

4) InceptionV3: Szegedy et al. proposed the Inception V3 architecture [29]. This model uses varying sizes filters and a concatenation of these filters is used to extract features at different scales as shown in Fig. 3. In Inception, the input is compressed using 1x1 convolutions after which, filters of different sizes are used on each of these spaces.

5) Xception: The Xception model is an "extreme" version of the Inception module that involves "Depthwise Separable Convolutions" [30]. The number of computations is high in the case of a basic convolution operation because the operation of applying filters on every input channel and the combining of these values is done in one step.

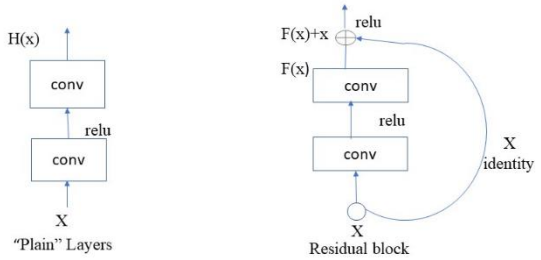

Fig. 1. Residual block in ResNet model.

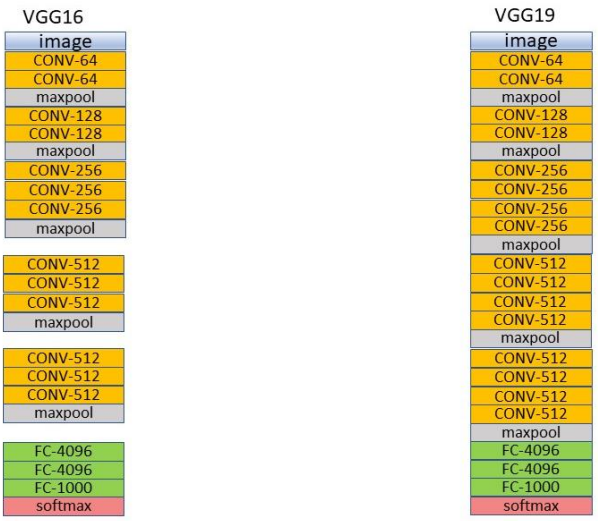

Fig. 2. VGG16 and VGG19 models.

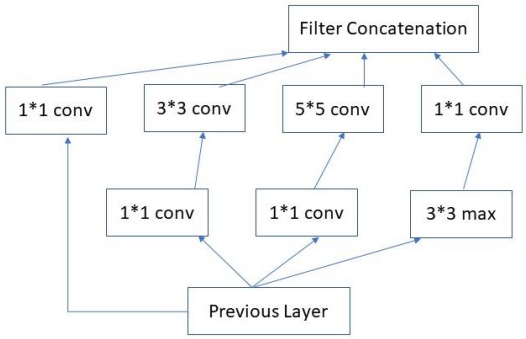

Fig. 3. The inception module.

Depthwise separable convolutions separates these steps into two, the first being the depthwise separation which is the filtering stage, followed by combining stage of pointwise convolution. Depthwise convolution applies convolution to the input image ($D_f * D_f * M$) with filters of depth 1 ($D_k * D_k * 1$), to a single input channel at a time as shown in equation (1) and then pointwise convolution compresses the input by applying N filters of dimension 1*1*M across the depth as shown in equation (2).

$$D_f * D_f * M \rightarrow D_g * D_g \quad (1)$$

$$D_g * D_g * M \rightarrow D_g * D_g * N \quad (2)$$

where D_f is dimension of input image, D_g is dimension of feature map and M is number of channels in input image

Table II gives a summary of the five pre-trained CNN models used in feature extraction.

B. Fine Tuning CNN

Fine tuning is the technique for speeding up the training process by overcoming the problems posed by small size dataset, by taking the weights of a pre-trained neural network and using it to initialize a new model which is being trained on similar domain data. This can be broadly categorized into Transfer Learning and Feature Extraction.

TABLE II. SUMMARY OF CNN MODELS USED IN FEATURE EXTRACTION

Model	*Input Size*	*Length of Feature Vector*	*Parameters*
Inception V3 (2015)	299*299*3	2048	Total parameters: 23,851,784 Trainable parameters: 23,817,352 Non-trainable parameters: 34,432
ResNet50 (2015)	224*224*3	2048	Total parameters: 25,636,712 Trainable parameters 25,583,592 Non-trainable parameters: 53,120 SGD - mini-batch size = 256 Learning rate starts from 0.1 and is divided by 10 when the error plateaus Weight decay = 0.0001 Momentum = 0.9 No Dropout
VGG16 (2014)	224*224*3	4096	Total parameters: 138,357,544 Trainable parameters: 138,357,544 Non-trainable parameters: 0 SGD(lr=0.1,decay=1e-6, momentum=0.9)
VGG19 (2015)	224*224*3	4096	Total parameters: 143,667,240 Trainable parameters 143,667,240 Non-trainable parameters 0
Xception (2015)	299*299*3	2048	Total parameters: 22,910,480 Trainable parameters: 22,855,952 Non-trainable parameters: 54,528

Most of the popular CNN architectures including VGG16, VGG19, ResNet50, AlexNet, InceptionV3, Xception have been pre-trained on ImageNet dataset. ImageNet consists of around 1.2 million training images, 50,000 validation images and 100,000 testing images, belonging to 1000 categories.

1) Transfer learning: Transfer learning works by loading pre-trained weights into a base CNN model, followed by freezing all the base model layers, making them non-trainable. This is followed by creating a new model on top of the output of any of the base model layers and training this newly created model on the new dataset.

2) Feature extraction: Feature Extraction is a lighter approach, wherein, after loading pre-trained weights into a base CNN model, the new dataset is run through the base model to extract the output of the layers of this model, and use this output as the input data for a new model.

C. Classification Algorithms

Classification is a "Supervised Learning" technique for identification of the category of new observations based on the training data. In classification, a program learns from the given observations and then classifies new observations into categories. In this work, classifiers are used for classification of images by utilizing the features extracted by the CNN models on the fashion dataset having men footwear and women clothing.

1) Random forest: Random Forest classifiers use a combination of many classifiers to solve complex problems. The advantages of Random Forest include its capability to maintain high accuracy through cross validation with higher dimensionality dataset, prevention of overfitting and the capability to handle missing data.

2) Support Vector Machine: Support Vector Machine (SVM) is a "supervised" machine learning algorithm that finds a hyperplane in an N-dimensional space that distinctly classifies the data points. SVM is very effective in high dimensional cases with the number of features deciding the dimension of the hyperplane. SVM maximizes the margin between the data points and the hyperplane.

3) K Nearest Neighbors: KNN is a "supervised learning" algorithm for regression and classification. KNN takes into consideration K nearest data points to predict the category or continuous value of the newly observed data. KNN uses all of the training instances for output prediction for new data. Model learning process is performed only at the time when prediction is requested on the new instance.

D. Similarity Metrics

The similarity metrics used for similarity matching are:

- Minkowski Distance: This distance metric takes two vectors and computes the distance between these vectors. The parameter "p" is called the "order", which allows calculation of different distance measures.

$$d(x,y) = (\sum_{i=1}^{n}|x_i - y_i|^p)^{1/p} \quad (3)$$

For Manhattan distance, p = 1.

For Euclidean distance p = 2.

- Manhattan Distance: This distance is computed as the sum of the absolute differences between two vectors.

$$d(x,y) = \sum_{i=1}^{n}|x_i - y_i| \quad (4)$$

- Euclidean Distance:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (5)$$

- Cosine Similarity:

$$sim(a,b) = \frac{a.b}{||a||.||b||} \quad (6)$$

- Jaccard Similarity:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (7)$$

IV. Reverse Image Search Framework and Algorithm

Problem Definition: For a given image query, retrieve the top-n similar images.

The features extracted using the different CNN models are utilized for reverse image search, using the framework shown in Fig. 4. For dimensionality reduction, Principal Component Analysis is used. Manhattan, Euclidean, Cosine distances and Jaccard similarity are used as the similarity measures.

A. Framework

The Reverse Image Search framework for retrieving visually similar images, shown in Fig. 4, consists of the following modules:

1) Data collection and preprocessing module: The Crawler utility fetches the retailer-specific webpages and extracts images. The preprocessing module prepares the images for the feature extraction process according to the input specifications of the selected pre-trained model. The image preprocessing utilities transform the raw image data to dataset objects, that can be further used as inputs to the model.

2) Feature extraction module: Text-based image retrieval methods may fail to retrieve visually similar relevant images because of the absence of query terms in the description of image. The feature extraction module overcomes this disadvantage by incorporating CNN layers for extraction of features. This module consists of the specific CNN model with the last fully connected layers removed, to extract the various levels of features. The extracted feature vectors are of high dimensionality and are therefore computationally expensive, with high memory requirement. Principal component analysis transforms the data into fewer number of dimensions, thereby giving summarized feature vectors. This helps to reduce the complexity in data and at the same time retains the patterns in the images.

3) Similarity Calculation module: This module computes the similarity between the feature vector of the input query image and the feature vectors of the database images extracted using the CNN models and the ranking of images is done according to their similarity. Finally, the top-n ranked images will be retrieved using the similarity metrics as in Equation (3) – Equation (7).

B. Algorithm: Top-n similar Images Retrieval

The algorithm for Top-n similar images retrieval is as follows:

Input : User query Image q, Image database I_{db} with image I.

Output: Top-n similar images

begin

Offline:

Pre-process the images in the image database I_{db} based on the input size of the CNN models

Extract Visual features F using the pre-trained CNN models for each image I in I_{db}

Reduce dimensionality using Principal Component Analysis

Online:

Extract Visual features F_q of the input query image q using the CNN models

for each feature vector Fj in F **do**

Compute the feature similarity *Sim(Fq, Fj)* using Equation (6)

Retrieve the Top-n similar images based on similarity calculation.

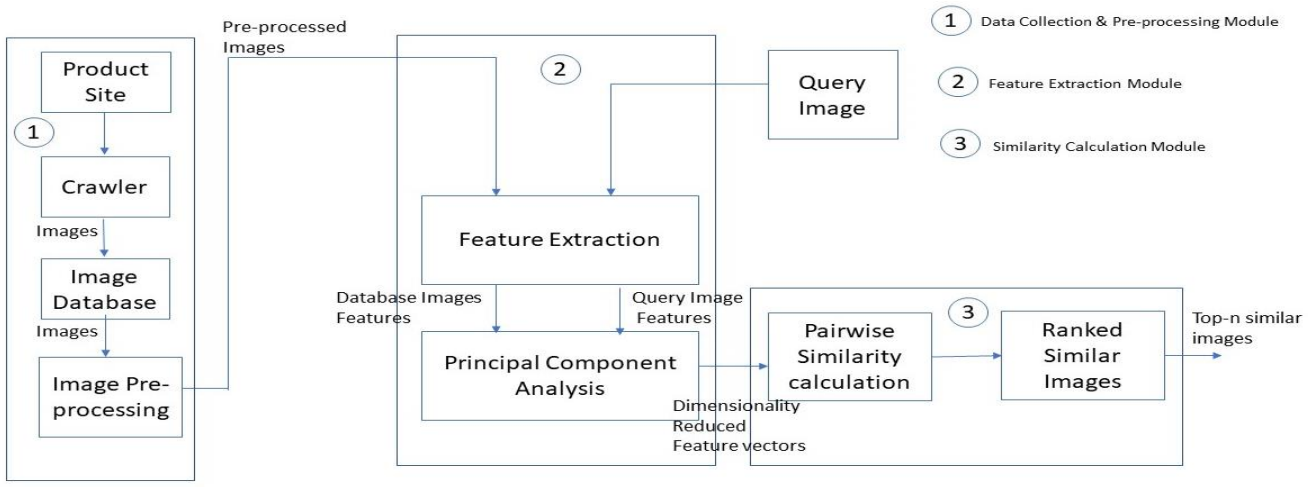

Fig. 4. Reverse image search framework.

V. Implementation and Results

A. Dataset and Experiment Setup

This work is carried out on 11th Gen Intel Core i7 processor with NVIDIA GeForce RTX 3050 Ti GPU. The men footwear dataset consists of a total of 2167 images belonging to three different classes of casual sandals, formal shoes and flip flops. The women clothes dataset consists of 4339 images belonging to five different classes of jeans, salwars, shirts, shorts and tops. The data is split into train and test sets with 80:20 ratio. The work concentrates on feature extraction using the pre-trained models VGG16, VGG19, InceptionV3, Xception and ResNet50 models, followed by evaluating the feature extraction efficiency by giving the extracted features as input to classifiers. Feature Visualization is done to know the depth and granularity of different features extracted at different layers.

B. Feature Extraction and Classification

This work utilizes the Keras Implementation of the pre trained CNN models. After initializing the input image size and applying the corresponding preprocessing functions, the pretrained weights are loaded into the model. The preprocessing function of the ImageDataGenerator module is used for data augmentation to improve generalization. Next, for feature extraction, the training images are loaded after expanding their dimensions according to the input size of the respective CNN models. Random Forest Classifier from Scikit learn library with n estimators' value of 50 is used in this work. SVC from the Scikit learn library is used here with a Linear kernel for SVM classification. KNN classifier with K=20 is used for the nearest neighbour classification.

For the men footwear dataset, as shown in Fig. 5, Random Forest gave an image classification accuracy of 97.5% with VGG19 features, SVM gave an accuracy of 97.5% with Xception features and KNN gave an accuracy of 95.8% with Xception features.

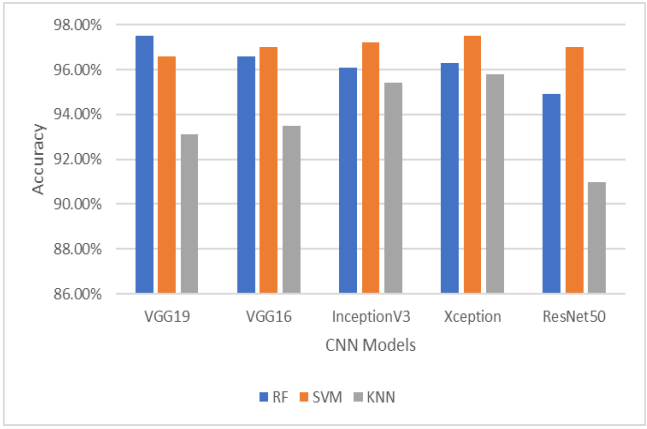

Fig. 5. Classification accuracy with features extracted by the different CNN models on the men footwear dataset.

For the women clothes dataset, as shown in Fig. 6, RF and SVM gave image classification accuracy of 89.1% and 96.8% respectively with InceptionV3 features and KNN gave an accuracy of 85.9% with Xception features.

The accuracy results show the efficiency of our selected pre-trained CNN models in feature extraction as compared to the work proposed in [10], where the authors achieved classification accuracy of 28.2% and 37.4% with KNN and SVM respectively on CIFAR 10 dataset, which consists of objects belonging to 10 classes. The results obtained with pretrained CNN model feature extraction and classification, outperform these with a classification accuracy of 72.4% and 86% with KNN and SVM classifiers respectively, using the features extracted by the Xception model on CIFAR 10 dataset.

Fig. 7 gives the classification accuracy obtained with the features extracted by the pretrained models on the CIFAR 10 dataset, along with the results obtained in [10] given as M1.

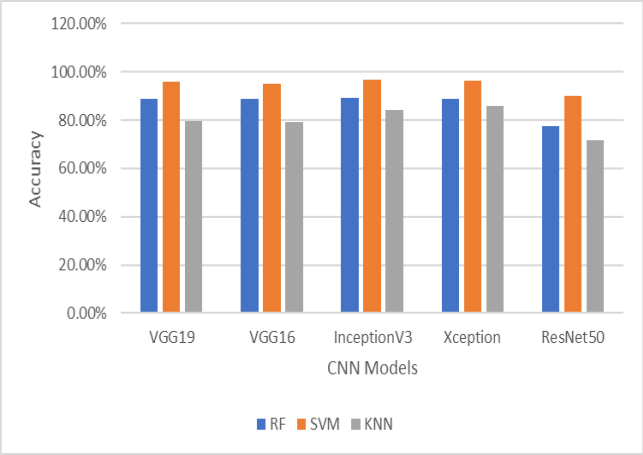

Fig. 6. Classification accuracy with features extracted by the different CNN models on the women clothes dataset.

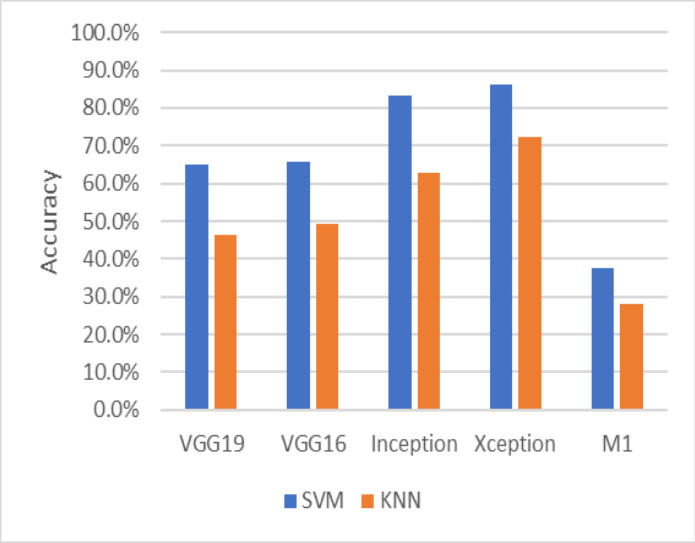

Fig. 7. Classification accuracy on the CIFAR 10 dataset.

C. Feature Visualization

The application of different convolutional filters to the input image, results in different feature maps, wherein, each pixel of each feature map is an output of the convolutional layer. Visualization of feature map for an image helps in understanding the detected features. The feature maps closer to the input layer of the neural network detect details of features like edges. The feature maps closer to the output layer of the model capture much more abstract features like texture.

Fig. 8 shows the feature map visualization of the different blocks of convolution layers of InceptionV3 model on a men footwear input image.

D. Reverse Image Search

The image retrieval results with features extracted by the five CNN models on Men footwear and women clothes dataset using Manhattan, Euclidean, Cosine and Jaccard similarity as the similarity measures, is compared in this section, in terms of the time taken for feature extraction and also the visual similarity of the retrieved images. The utilities for image preprocessing is imported from tf.keras.preprocessing. This transforms the raw image data to a tf.data.Dataset object, that can be further used to train the model. The extracted feature vectors are of high dimensionality and have high memory requirement. For dimensionality reduction, Principal Component Analysis with 300 components is used.

1) Time taken for feature extraction and image retrieval: The Men Footwear train dataset consists of 1732 training images belonging to three classes and the Women Clothes dataset consists of 3469 training images belonging to five classes. On an average, VGG19 takes more time for extracting the features of the training set images followed by VGG16, ResNet50, Inception V3, and Xception.

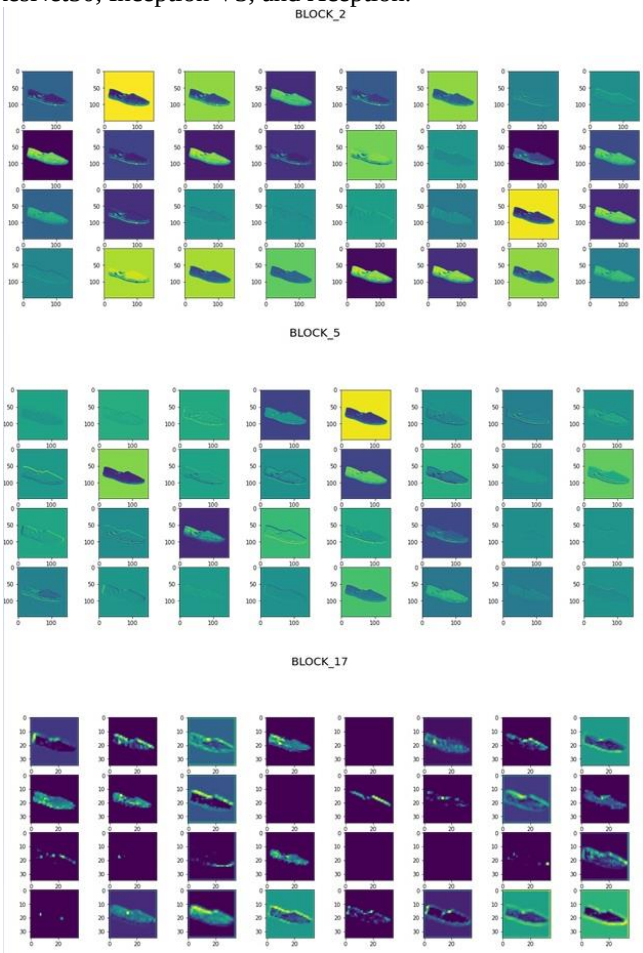

Fig. 8. Feature visualization using Inceptionv3 on men footwear image.

Fig. 9 shows the time taken for extraction of Men Footwear database image features by the different CNN models. Fig. 10 shows the time taken for query image feature extraction by the different CNN models and retrieval of top five similar images.

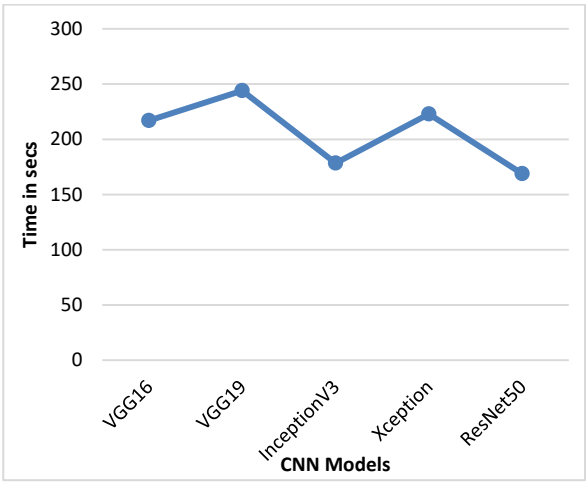

Fig. 9. Time taken in seconds for extraction of men footwear database image features by the different CNN models.

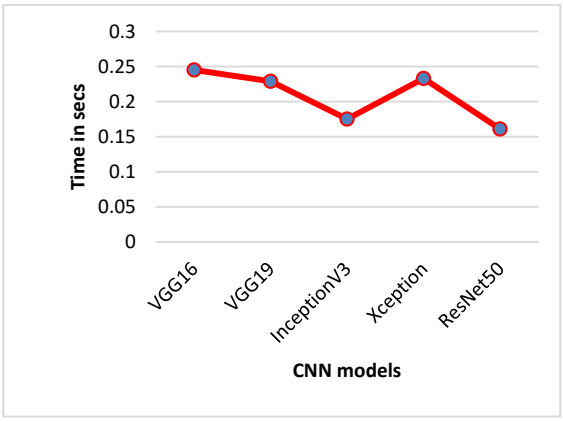

Fig. 10. Time taken in seconds for query image feature extraction by the different CNN models and retrieval of top 5 similar images.

Fig. 11 and Fig. 13 show the image retrieval results with features extracted by the five CNN models on Men footwear and women clothes dataset using Manhattan, Euclidean and Cosine similarity as the similarity measures.

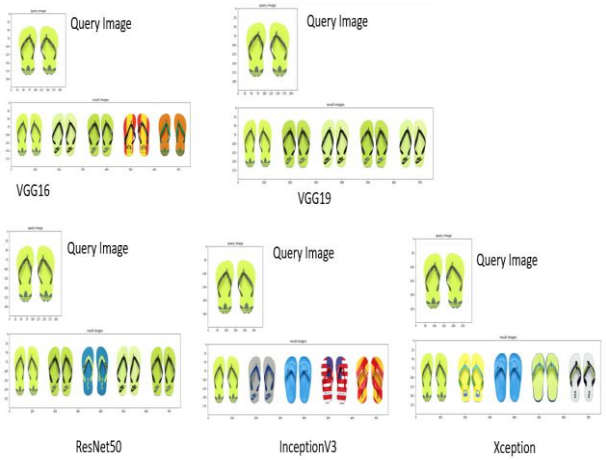

Fig. 11. Men footwear reverse image search with Cosine, Euclidean and Manhattan similarity metrics using the different CNN models for feature extraction (query image on top and top 5 result images below for each CNN model).

Image retrieval results with Jaccard similarity show poor results as compared to Manhattan, Euclidean and Cosine as can be seen from Fig. 12.

Fig. 12. Men footwear reverse image search with Jaccard similarity metrics using VGG19 and InceptionV3 CNN models for feature extraction (query image on top and top 5 result images below).

Fig. 13. Women clothes reverse image search with Cosine, Euclidean and Manhattan similarity metrics using the different CNN models for feature extraction (query image on top and top 5 result images below for each CNN model).

VI. CONCLUSION

The work carried out focuses on feature extraction using VGG19, VGG16, ResNet50, InceptionV3 and Xception models, followed by evaluating the feature extraction efficiency by feeding the extracted features as input to classifiers. Results of feature extraction and image retrieval show that VGG19 features show best results with both men footwear and women clothes dataset. The classification accuracy results show that VGG19 features give more accuracy with men footwear, whereas InceptionV3 and Xception features are better with women clothes. This work also compares the time taken for feature extraction and image retrieval with each of the five CNN models. The features extracted using the deep learning models can be used to analyze images in various image-based systems. One practical application of this would be in image-based recommender systems. The features extracted from images can be used to build item profiles in content-based image recommendations. Recent advances in the field of recommender systems use graph-based methods like Graph Convolutional Network, as they are capable of representing the complex embeddings.

Future research directions include:

- Utilization of Deep Learning models for transfer learning with optimization techniques.
- Using a fusion of different handcrafted local features and CNN features is also one of the challenging future directions for research.

REFERENCES

[1] Enser, Peter, and Christine Sandom. "Towards a comprehensive survey of the semantic gap in visual image retrieval." In International Conference on Image and Video Retrieval, pp. 291-299. Springer, Berlin, Heidelberg, 2003.

[2] Salau, Ayodeji Olalekan, and Shruti Jain. "Feature extraction: a survey of the types, techniques, applications." In 2019 International Conference on Signal Processing and Communication (ICSC), pp. 158-164. IEEE, 2019.

[3] Zhuang, Fuzhen, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. "A comprehensive survey on transfer learning." Proceedings of the IEEE 109, no. 1 (2020): 43-76.

[4] Ryszard S Choras. Image feature extraction techniques and their applications for cbir and biometrics systems. *International journal of biology and biomedical engineering*, 1(1):6–16, 2007.

[5] Ke Yan, Yaowei Wang, Dawei Liang, Tiejun Huang, and Yonghong Tian. Cnn vs. sift for image retrieval: Alternative or complementary? In Proceedings of the 24th *ACM international conference on Multimedia*, pages 407–411, 2016.

[6] Jigisha M Patel and Nikunj C Gamit. A review on feature extraction techniques in content based image retrieval. In 2016 *International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pages 2259–2263. IEEE, 2016.

[7] Salamh, Ahmed B. Salem, and Halil Ibrahim Akyüz. "A Novel Feature Extraction Descriptor for Face Recognition." *Engineering, Technology & Applied Science Research* 12, no. 1: 8033-8038, 2022.

[8] Alasadi, Abdulmalik A., T. H. Aldhayni, Ratnadeep R. Deshmukh, Ahmed H. Alahmadi, and Ali Saleh Alshebami. "Efficient feature extraction algorithms to develop an arabic speech recognition system." *Engineering, Technology & Applied Science Research* 10, no. 2: 5547-5553, 2020.

[9] Alsuwaiket, Mohammed A. "Feature Extraction of EEG Signals for Seizure Detection Using Machine Learning Algorthims." *Engineering, Technology & Applied Science Research* 12, no. 5: 9247-9251, 2022.

[10] Manjunath Jogin, MS Madhulika, GD Divya, RK Meghana, S Apoorva, et al. Feature extraction using convolution neural networks (cnn) and deep learning. In 2018 3rd *IEEE international conference on recent trends in electronics, information & communication technology (RTEICT)*, pages 2319– 2323. IEEE, 2018.

[11] Dario Garcia-Gasulla, Ferran Par´es, Armand Vilalta, Jonatan Moreno, Eduard Ayguad´e, Jesu´s Labarta, Ulises Cort´es, and Toyotaro Suzumura. On the behavior of convolutional nets for feature extraction. *Journal of Artificial Intelligence Research*, 61:563–592, 2018.

[12] Wenyi Lin, Kyle Hasenstab, Guilherme Moura Cunha, and Armin Schwartzman. Comparison of handcrafted features and convolutional neural networks for liver mr image adequacy assessment. *Scientific Reports*, 10(1):1–11, 2020.

[13] Yingqi Zhang, Yong Feng, Jiaxing Shang, Mingliang Zhou, and Baohua Qiang. Attention-aware joint location constraint hashing for multi-label image retrieval. *IEEE Access*, 8:3294–3307, 2019.

[14] Huimin Lu, Ming Zhang, Xing Xu, Yujie Li, and Heng Tao Shen. Deep fuzzy hashing network for efficient image retrieval. *IEEE transactions on fuzzy systems*, 29(1):166–176, 2020.

[15] Maruthamuthu Ramasamy and John Sanjeev Kumar Athisayam. 2d matrix based indexing with color spectral histogram for efficient image retrieval. *Journal of Systems Engineering and Electronics*, 27(5):1122–1134, 2016.

[16] Ahmad Raza, Hassan Dawood, Hussain Dawood, Sidra Shabbir, Rubab Mehboob, and Ameen Banjar. Correlated primary visual texton histogram features for content base image retrieval. *IEEE Access*, 6:46595–46616, 2018.

[17] LK Pavithra and T Sree Sharmila. An improved seed point selection-based unsupervised color clustering for content-based image retrieval application. *The Computer Journal*, 63(3):337–350, 2020.

[18] Guangyi Xie, Baolong Guo, Zhe Huang, Yan Zheng, and Yunyi Yan. Combination of dominant color descriptor and hu moments in consistent zone for content based image retrieval. *IEEE Access*, 8:146284–146299, 2020.

[19] Xiaoli Zhang. Content-based e-commerce image classification research. *IEEE Access*, 8:160213– 160220, 2020.

[20] Ayesha Khan, Ali Javed, Muhammad Tariq Mahmood, Muhammad Hamza Arif Khan, and Ik Hyun Lee. Directional magnitude local hexadecimal patterns: A novel texture feature descriptor for content-based image retrieval. *IEEE Access*, 9:135608–135629, 2021.

[21] Yongwei Miao, Gaoyi Li, Chen Bao, Jiajing Zhang, and Jinrong Wang. Clothingnet: Cross- domain clothing retrieval with feature fusion and quadruplet loss. *IEEE Access*, 8:142669–142679, 2020.

[22] Jianbo Ouyang, Wengang Zhou, Min Wang, Qi Tian, and Houqiang Li. Collaborative image relevance learning for visual re-ranking. *IEEE Transactions on Multimedia*, 23:3646–3656, 2020.

[23] Chen, Rongyu, Lili Pan, Yan Zhou, and Qianhui Lei. "Image retrieval based on deep feature extraction and reduction with improved CNN and PCA." Journal of Information Hiding and Privacy Protection 2, no. 2 (2020): 67.

[24] Zhang, Xuanmeng, Minyue Jiang, Zhedong Zheng, Xiao Tan, Errui Ding, and Yi Yang. "Understanding image retrieval re-ranking: a graph neural network perspective." arXiv preprint arXiv:2012.07620 (2020).

[25] Kan, Shichao, Yigang Cen, Yang Li, Mladenovic Vladimir, and Zhihai He. "Local Semantic Correlation Modeling Over Graph Neural Networks for Deep Feature Embedding and Image Retrieval." IEEE Transactions on Image Processing 31 (2022): 2988-3003.

[26] Wang, Shoujin, Liang Hu, Yan Wang, Xiangnan He, Quan Z. Sheng, Mehmet A. Orgun, Longbing Cao, Francesco Ricci, and Philip S. Yu. "Graph learning based recommender systems: A review." arXiv preprint arXiv:2105.06339 (2021).

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[29] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818-2826, 2016.

[30] Chollet, François. "Xception: Deep learning with depthwise separable convolutions." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251-1258. 2017.