http://www.jctjournals.com ISSN(ONLINE):2278 – 3814

Volume-15, Issue-01, January 2026
JOURNAL OF COMPUTING TECHNOLOGIES (JCT)
International Journal
Page Number: 01-10

An Image Similarity Computation Model Using Convolution Artificial Neural Networks

Aryan Rawat[1], Dr Vivek Richhariya[2]
[1]M.Tech Scholar, [2] Professor,
[1,2]Department of Computer Science Engineering *(CSE)*
[1,2] Lakshmi Narain College of Technology, Bhopal, INDIA

***Abstract*— The classification of objects into the right classes has long been one of the most crucial objectives of machine learning or deep learning. Due to the similarities between various things, their textures, colors, and other physical properties, object recognition still presents significant challenges despite the importance of categorizing specific groups of images. In computer vision, object detection has a wide range of uses, such as face and vehicle detection, video surveillance, and plant leaf detection. When conducting studies on flowers, using flowers as medicine, analyzing floral patents, etc., automatic classification of flowers is crucial. Traditionally, low-level characteristics like color, shape, texture, and geometry are used to classify flowers. The feature description has a significant impact on how accurately and robustly flowers are classified. In recent years, deep features have demonstrated good performance on high-resolution photos, but they are unable to extract precise global features from low-resolution images. Deep neural networks have been widely used in computer vision applications because they are effective at identifying picture patterns. An advanced deep-learning model that can precisely calculate the similarity of floral photos is required in the field of flower image analysis. In comparison to the given model mentioned in the exhibit results, the proposed model performs adequately. The proposed network can still be improved in terms of learning parameters, validation accuracy, loss, and training time. Fine-tuned deep learning models for similarity computation on flower datasets have a bright and broad future ahead of them, with lots of room for development and use. In this research work, the future potential of optimized deep learning models for the calculation of similarity on floral datasets is promising for growth and innovation. These models have the power to fundamentally alter the way we see, value, and engage with the world of flowers. The proposed model is supported by augmentation so the distance method is used against two augmentation views of the trained model being calculated.**

Keywords— Machine learning, Deep learning, object recognition, Convolution neural network, Similarity computation.

I. INTRODUCTION

The ability to identify the objects present in an image or scene is one of the most basic requirements when it comes to interacting with one‟s environment. While it seems completely effortless with humans and in fact most animals, trying to teach computers to see and also understand" what they are seeing has proven extremely difficult. The key to understanding visual scenes are three closely related sub-problems. The easiest one will be called classification in the following. For classification, the one dominant object in a given image should be determined and labelled. The next more demanding task is object localization: In addition to labeling the dominant object, it also needs to be localized in the image, usually by determining a bounding box around the image region that is occupied by the object. The difficulty of this task again increases if not only one but all objects in an image need to be labeled and multiple objects of the same category can appear in one image, figure 1 showing the process of object detection.

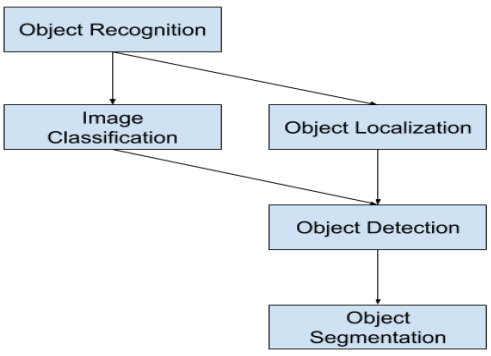

Fig. 1: Object detection process

Image degradation is inevitable during the transmission and alteration of images. For example, the quality of an image shot by a camera is sometimes low due to the distortion of camera's optics scheme, the relative motion of the photographed object and the camera, the ecological change and the arbitrary disturbance [1]. The image

enhancement is an important technique that can improve the quality of the degraded image and offer some interesting image features selectively. Image enhancement is basically improving the interpretability or perception of information in images for human viewers and providing better input for other automated image processing techniques. The main objective of image enhancement is to modify attributes of an image to make it more suitable for a given task and a specific spectator. For the duration of this process, one or more characteristic of the image are customized [2]. The alternative of attributes and the way they are customized are specific to a given problem. Moreover, observer-specific factor, such as the person visual system and the observer's experience, will bring in a great deal of subjectivity into the choice of image enhancement methods.

A. Images Features

Image features refer to the information collected from images that can uniquely identify the image or can be used for further processing. Broadly, image features can be classified into general features and domain-specific features [5]. General features, such as color and texture are applicable to all image data and do not depend on the application being considered. Domain-specific features on the other hand, are specific to the application at hand, such as, minutiae in fingerprints. Figure 2 showing the different features of an image. Based on the locality of features, image features can be categorized into [6]:

(i) Local features: Local features are the patterns in images that differ from its immediate neighborhood. These features are extracted from a patch in the image and are useful in applications such as object recognition. Some examples of local features are Shape Invariant Feature Transform (SIFT), Local Binary Pattern (LBP), and Speeded up Robust Features (SURF).

(ii) Global features: Global features represent the whole image. These features are extracted considering the whole image as one patch/object and are useful in applications such as image retrieval and image classification, where a rough segmentation of objects is available. Some examples of global features are Histogram Oriented Gradient (HOG) and Shape Matrices.

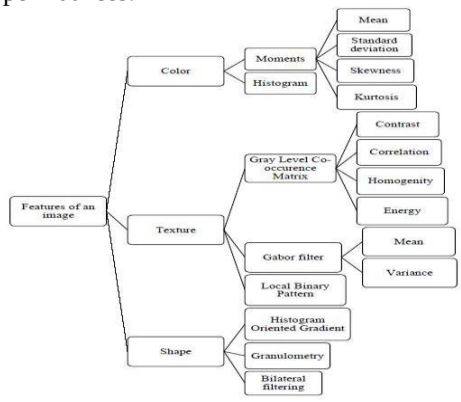

Fig 2: Features of an image

B. Classification

Once a network is trained against a subset of valid objects, one of the tasks that the network can be used for is classification. In the simplest form, a classification problem can be stated as such: given an object and a set, is this object in the set. One of the underlying principles of the deep learning architecture is the reconstruction of valid objects into their original pattern. Of course, if a random image is sent into such a network, it will not resemble itself very well. However, if the object is a close relation to the objects trained, it should be reconstructed with a high fidelity. Figure 3 present image classification based on image features. Using this, a simple threshold can be established, and if the reconstruction is in error beyond this threshold, it can be declared not in the set

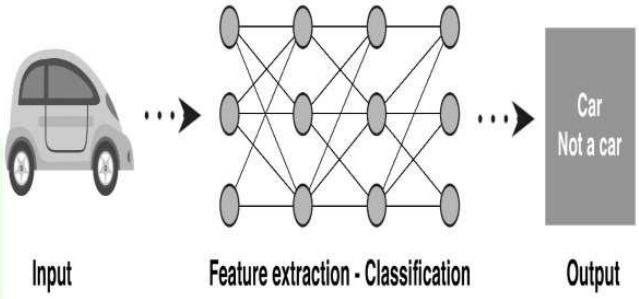

Fig 3: Image feature based classification process

C. Application Cases Of Classification

Application cases of image classification are classified into scene classification, object detection and object extraction. Scene classification is the process of determining the type of a remote sensing image based on its content. Object detection is the process of determining the locations and types of the targets to be detected in a remote sensing image and labeling their locations and types with bounding boxes. Object extraction is the process of determining the accurate boundaries of the objects to be extracted in a remote sensing image. In this section, we summarize these application cases.

Scene classification Scene classification is a mapping process of learning and discovering the semantic content tags of image scenes. Generally, an image scene is a collection of multiple independent geographic objects. These objects have different structures and contain different texture information, and they form different types of scenes through different combinations and spatial locations. For scene classification studies in the remote sensing field, the UC Merced land use dataset is commonly viewed as the reference dataset.

Object detection Object detection from remote sensing images detects the locations and the types of objects. The object detection application cases from remote sensing images use the candidate region-based object detection method. The method involves three steps: the generation of candidate regions, feature extraction by the image classification techniques and classification of candidate regions. Candidate regions are a series of locations in which the objects may appear in the pre-generated image. All of these locations will be used as the input for the

image classification techniques for feature extraction and classification.

Object segmentation To extract objects from a remote sensing image, it is necessary to segment the objects of interest in the image and to produce a pixel-level image classification map. Two types of methods are primarily used in the existing CNN-based studies on object segmentation from remote sensing images, namely patch-based CNN methods and end-to-end CNN methods. A patch-based CNN method generally first obtains a prediction model by training a CNN on a training dataset, and then, based on the prediction model, it generates image patches using a sliding window pixel by pixel and predicts the type of each pixel of the image.

D. Computer Vision

Computer vision has been revolutionized by high capacity Convolution Neural Networks (ConvNets) and large-scale labeled data. Recently weakly-supervised training on hundreds of millions of images and thousands of labels has achieved state-of-the-art results on various benchmarks. Interestingly, even at that scale, performance increases only log linearly with the amount of labeled data. Thus, sadly, what has worked for computer vision in the last five years has now become a bottleneck: the size, quality, and availability of supervised data. Unsupervised representation learning is highly successful in natural language processing. But supervised pre-training is still dominant in computer vision, where unsupervised methods generally lag behind. The reason may stem from differences in their respective signal spaces. Language tasks have discrete signal spaces (words, sub-word units, etc.) for building tokenized dictionaries, on which unsupervised learning can be based. Computer vision, in contrast, further concerns dictionary building, as the raw signal is in a continuous, high-dimensional space and is not structured for human communication (e.g., unlike words). Several recent studies present promising results on unsupervised visual representation learning using approaches related to the contrastive loss. Though driven by various motivations, these methods can be thought of as building dynamic dictionaries. The “keys” (tokens) in the dictionary are sampled from data (e.g., images or patches) and are represented by an encoder network. Unsupervised learning trains encoders to perform dictionary look-up: an encoded “query” should be similar to its matching key and dissimilar to others. Learning is formulated as minimizing a contrastive loss.

Unsupervised learning has been widely studied in the Machine Learning community [16], and algorithms for clustering, dimensionality reduction or density estimation are regularly used in computer vision applications. For example, the \bag of features" model uses clustering on handcrafted local descriptors to produce good image-level features [14]. A key reason for their success is that they can be applied on any specific domain or dataset, like satellite or medical images, or on images captured with a new modality, like depth, where annotations are not always available in quantity.

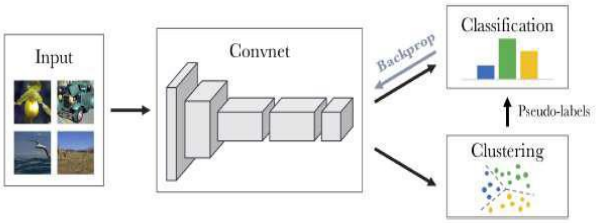

Fig 4: Illustration of the convent method with clustering and classification

II. LITERATURE REVIEW

A. Previous Work Done

Siamese networks have become a common structure in various recent models for unsupervised visual representation learning. These models maximize the similarity between two augmentations of one image, subject to certain conditions for avoiding collapsing solutions

[1]. In this paper, they report surprising empirical results that simple Siamese networks can learn meaningful representations even using none of the following: (i) negative sample pairs, (ii) large batches, (iii) momentum encoders. Their experiments show that collapsing solutions do exist for the loss and structure, but a stop-gradient operation plays an essential role in preventing collapsing. They also provide a hypothesis on the implication of stop-gradient, and further show proof-of-concept experiments verifying it. A main purpose of unsupervised learning is to pre-train representations (i.e., features) that can be transferred to downstream tasks by fine-tuning

[2]. Here author present Momentum Contrast (MoCo) for unsupervised visual representation learning. From a perspective on contrastive learning as dictionary look-up, they build a dynamic dictionary with a queue and a moving-averaged encoder. This enables building a large and consistent dictionary on-the-fly that facilitates contrastive unsupervised learning. Human observers can learn to recognize new categories of images from a handful of examples; yet doing so with artificial ones remains an open challenge.

They [3] hypothesize that data-efficient recognition is enabled by representations which make the variability in natural signals more predictable. They therefore revisit and improve Contrastive Predictive Coding, an unsupervised objective for learning such representations. This new implementation produces features which support state-of-the art linear classification accuracy on the ImageNet dataset. When used as input for non-linear classification with deep neural networks, this representation allows us to use 2–5x less labels than classifiers trained directly on image pixels. Clustering is a class of unsupervised learning methods that has been extensively applied and studied in computer vision. Little work has been done to adapt it to the end-to-end training of visual features on large scale datasets. In this research work, author

[4] present Deep Cluster, a clustering method that jointly learns the parameters of a neural network and the cluster assignments of the resulting features. Deep Cluster iteratively groups the features with a standard clustering

algorithm, k- means, and uses the subsequent assignments as supervision to update the weights of the network. Unsupervised image representations have significantly reduced the gap with supervised pretraining, notably with the recent achievements of contrastive learning methods. These contrastive methods typically work online and rely on a large number of explicit pair wise feature comparisons, which is computationally challenging. In this research work

[5] they propose an online algorithm, SwAV, that takes advantage of contrastive methods without requiring computing pair wise comparisons. Specifically, our method simultaneously clusters the data while enforcing consistency between cluster assignments produced for different augmentations (or "views") of the same image, instead of comparing features directly as in contrastive learning. Simply put, they use a "swapped" prediction mechanism where they predict the code of a view from the representation of another view. Their method can be trained with large and small batches and can scale to unlimited amounts of data. Compared to previous contrastive methods, our method is more memory efficient since it does not require a large memory bank or a special momentum network. One core objective of deep learning is to discover useful representations, and the simple idea explored here is to train a representation-learning function, i.e. an encoder, to maximize the mutual information (MI) between its inputs and outputs. This work investigates unsupervised learning of representations by maximizing mutual information between an input and the output of a deep neural network encoder. Importantly,

[6] they show that structure matters: incorporating knowledge about locality in the input into the objective can significantly improve a representation‟s suitability for downstream tasks. They further control characteristics of the representation by matching to a prior distribution adversarial. The main challenge of unsupervised embedding learning is to discover visual similarity or weak category information from unlabeled samples.

This research works [7] studies the unsupervised embedding learning problem, which requires an effective similarity measurement between samples in low dimensional embedding space. Motivated by the positive concentrated and negative separated properties observed from category-wise supervised learning, they propose to utilize the instance wise supervision to approximate these properties, which aims at learning data augmentation invariant and instance spread out features. To achieve this goal, they propose a novel instance based softmax embedding method, which directly optimizes the „real‟ instance features on top of the softmax function. It achieves significantly faster learning speed and higher accuracy than all existing methods. The proposed method performs well for both seen and unseen testing categories with cosine similarity. It also achieves competitive performance even without pretrained network over samples from fine-grained categories. Pre-training general-purpose visual features with convolution neural networks without relying on annotations is a challenging and important task. Most recent efforts in unsupervised feature learning have focused on either small or highly curated datasets like ImageNet, whereas using non-curated raw datasets was found to decrease the feature quality when evaluated on a transfer task.

A. Problem Statement

The classification of objects into the right classes has long been one of the most crucial objectives of machine learning or deep learning. Due to the similarities between various things, their textures, colors, and other physical properties, object recognition still presents significant challenges despite the importance of categorizing specific groups of images. In computer vision, object detection has a wide range of uses, such as face and vehicle detection, video surveillance, and plant leaf detection. When conducting studies on flowers, using flowers as medicine, analyzing floral patents, etc., automatic classification of flowers is crucial. Traditionally, lowlevel characteristics like color, shape, texture, and geometry are used to classify flowers. The requirement for extensive and varied datasets is one important challenge. The acquisition of comprehensive flower image datasets that include a wide range of species and environmental conditions is frequently resource-intensive and time-consuming. The implemented work will create a fine-tuned deep-learning model specifically created for computing the similarity between flower images in a large-scale flower dataset in light of the scope and gap in this research area. The diversity of flower species and their visual characteristics make manual classification time-consuming and prone to error. The research work is therefore directed towards addressing the following problem such as handling variations within the image dataset, intra-class variability in image dataset, the semantic gap between query image and classified image, and feature extraction complexity.

III. IMAGE CLASSIFICATION OVERVIEW

A. Feature Extraction Techniques

Feature extraction is an important technique used in image classification, pattern recognition and object recognition. In order to have effective classification of plant species researchers should decide to extract efficient features Curvature Descriptors Curvature Scale Space (CSS) is a technique used to measure the contours of shapes, extracts the concavity and convexity of curvature. It is invariant to translation and rotation in a viewpoint direction but not in scale, because it varies with the Gaussian kernel (a) and cannot easily fix the value of the Gaussian kernel. It leads to misclassification of serrated and lobe-shaped leaves. Curvature is a vital property of leaves and curvatures are computed using differential techniques. However, it produces more noise, is sensitive to rotation, and generates different feature vectors with different scales. It is impossible to sustain all the curvature features combined together in one feature vector. Aligning them all in one particular point is a difficult task, because the features differ for each scale.

Multi-scale Descriptors :- The multi-scale descriptors furnish much more information about leaf contours. Derived from the scale space and image pyramid structure, it extracts image features at various levels by capturing

local and global features from low- to high-resolution scales. It provides the maximum discriminating power and is robust to noise depending on the boundaries of leaves and not the regions of an image. As a result, it works well on feature space rather than image space. Multiscale Triangular Area Representation (MTAR) which is affine invariant, robust to noise and provides the features of images concavity and convexity. Multi-scale Arc Height Descriptor (MARH) which is invariant to translation, It enumerates a local normalization technique for each scale to employ rotation and scaling, because the local normalization rendered for each scale is based on the maximum value of arch height descriptors. It leads to shape dissimilarity at each different scale, so is invariant to translation and scaling. It measures the arch height of palmate shaped and lobe-shaped leaves but is unsuitable for overlapped leaves. In this method, the local normalization scheme is applied for scaling and rotation. It takes longer execution time, compared to other invariant descriptors.

B. Point and Edge-based Feature Descriptors

A new descriptor called the shape context was introduced to dissociate shape information from different shapes. It is a technique used to extract point information from a shape"s contours, measure similarity differences between feature vectors of various points in an image, and isolate information from the neighboring pixels of an image. Figure 5 present image feature extraction process, The transformation of an object does not affect shape context information. It is invariant to rotation since it performs log polar operations while computing shape context information. It is invariant to small affine transformations, occlusions, the presence of outliers, and is applicable to clear images. Shape context is used to calculate the local and spatial information of an image. An advanced shape context method was introduced to reduce computational cost. In this method they used two sets: a voting set and a computing set. While the voting set was used to build the histogram information of the shape, the computing set was used to compute the shape context information of various shapes.

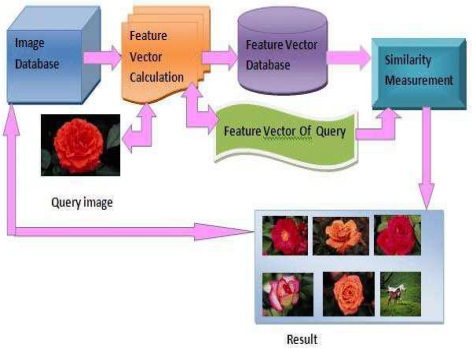

Fig 5 Query image and similarity image extraction based on their features

C. Invariant Feature Detectors

Image transforms convert sets of images into a series of orthogonal images in the form of unitary matrices. The primary aim of transformation is to represent a unit image into a set of linear combination basic images, extract features like the edges and corners of images, and determine shift invariant rotations and scaling invariant images. Pyramid Histogram-oriented Gradient (PHOG) computes the local shape and global spatial information in leaf images. It extracts edge contour information and calculates histogram bins on each local bin. It operates on dense grid cells and is invariant to geometric and photometric variations, except object orientation. The power spectrum with harmonic analysis has applied TSO invariance, such as translation, rotation, scaling and mirroring based on Fourier descriptors. They introduced affine invariant harmonic analysis of radii spectrum for an affine invariant transform. It is calculated based on image moments. A redundant discrete wavelet transform identifies orthogonal moments. Unlike other wavelet transforms, it does not consider all the input pixel values of images. It considers only odd pixels for scaling, including pixels for wavelet coefficients and reduces computational complexity. A polar Fourier transform (PFT) converts an original image into polar space so it is translation invariant, and as phase information is neglected, it is rotation invariant as well. The first magnitude value is normalized into scaling invariants, compared to other moment-based Zernike polynomials. They classified leaves using probabilistic neural network by incorporating shape, vein, color and texture features with it and achieved 93.2% of classification accuracy compared to geometric and moment invariant features in their own databases. A log polar transform used with rotation and scale invariant features to classify different texture patterns. It follows point singularities and converts images into concentric circles. They stated that ridgelet transform was useful for texture classification and these features are rotational and scale invariants. They demonstrated that it provided 100% accuracy, an excellent result compared to the result produced by log polar transform on a rotational and scale invariant database of images. It is optimal to find only lines of the size of the image.

D. SVM Classification

Despite the increasing popularity and high effectiveness of CNN classification techniques, the direct deployment of CNN techniques requires large training datasets [14] that are potentially difficult to obtain when the underlying data is privacy sensitive. In addition, black-box transformation of CNN-based methods to their privacy-preserving equivalents will result in classifiers that are computationally prohibitive to use. Thus using a light-weight classification method such as SVMs can be beneficial in privacy sensitive environments, and their evaluation can be done (as we show) in a secure manner. With the CNN features, an SVM can learn quickly from very few positive examples, which shows that they are useful to perform one-shot learning [15]. Thus, we opted for the design of a private SVM classifier, while using the techniques of CNN-based transfer learning in the context of feature extraction, which does not raise privacy concerns.

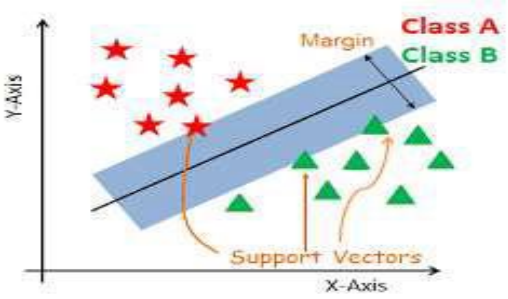

Fig 6: An example of image classification with SVM**.**

*E. Convolutional Neural Network **(CNN)***

CNNs, as one type of deep leaning networks, have the following advantages over shallow structure models:

- CNNs directly apply a convolution operation to the pixels of an image to extract abstract data features. This feature extraction can be applied to various scenarios and has more powerful generalization ability.
- CNNs are able to represent image information in a distributed manner and rapidly acquire image information from massive volumes of data. The structure of CNNs can effectively solve complex nonlinear problems (e.g., the rotation and translation of an image).
- CNNs are characterized by sparse connections, weight-sharing and spatial sub-sampling, which result in a simpler network structure that is more adaptable to image structures. In order to better understand CNN-based image classification, this section will briefly introduce the structure of CNNs and its training method, followed by several popular CNN models in the computer vision field.

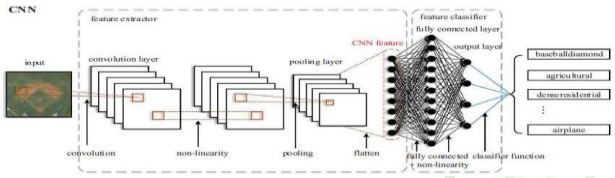

Fig 7: Structure of CNN

*F. **Deep Metric Learning***

Traditional machine learning techniques are limited by their ability to process data on raw data. Therefore, they need feature engineering, such as preprocessing and feature extraction steps before classification or clustering tasks. All of these steps require expertise and they are not directly within the classification structure. However, deep learning learns the higher level of data directly in the classification structure. This perspective shows the fundamental difference between traditional machine learning methods and deep learning. Unlike traditional machine learning methods, deep learning needs a high size of data to achieve successful results, since it is not successful enough in low data size. Besides, deep learning algorithms require a lot of time to train data because of the high data size and a large number of parameters of the algorithm. However, these pre-defined metrics have limited capabilities in data classification. Hence, an approach based on the Mahalanobis metric was proposed to classify the data into traditional metric learning to address this problem. In this approach, the data is transformed into a new feature space with higher discrimination power. Usually, metric learning approaches are related to the linear transformation of the data without any kernel function. However, these approaches are not successful enough to reveal nonlinear knowledge of the provided to overcome this problem, there is no obvious success due to some issues such as scaling. Unlike traditional metric learning methods, deep learning solves this problem using activation functions that have nonlinear structure.

IV. PROPOSED MODEL

A. Implemented Work

Visual object recognition is an important machine learning application, deployed in numerous real-life settings. Machine Learning as a Service (MLaaS) is becoming increasingly popular in the era of cloud computing, data mining, and knowledge extraction. Object recognition is such a machine learning task that can be provided as a cloud service. However, in most application scenarios, straightforward outsourcing of the object recognition task is not possible due to privacy concerns. Generally, the image holder who wishes to perform the image classification process, requires their input images to remain confidential. On the other hand, the classification algorithm provider wishes to commercially exploit their algorithm; hence, requires the algorithm parameters to remain confidential. Evolutionary computation was developed with the idea that it could be used as a tool for optimization and solutions to problems could be evolved using operators of natural selection. In Classification is play a vital role in image processing task, the classification task done with the various classifier such as the neural network classifier. Feature extraction is the process we used for the classification or retrieval of an image through various features like shape, texture and color. In this work we used the texture feature and improve the classification ratio of mage retrieval than previous techniques. The proposed method used the genetic algorithm for the best fitness function and evaluation with the number of image as used as an input and targeted output.

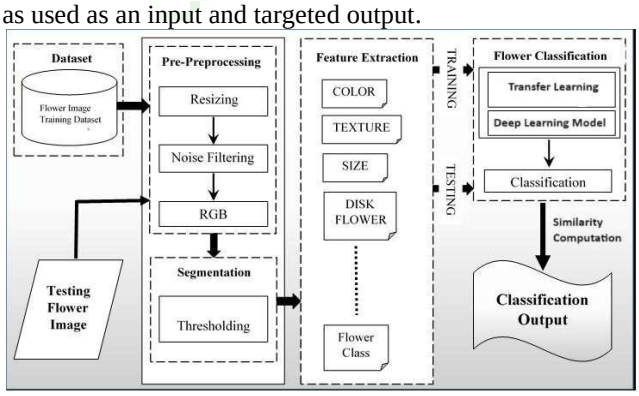

Fig 8: Implemented work steps

B. Dataset Preparation

It is a crucial and fundamental phase in the creation of machine learning models is dataset preparation, which entails gathering, organizing, and preparing the data that will be used to train, validate, and test your models. This step includes collection, labeling, cleaning, and splitting tasks in the proposed work. This step is important for the quality of results, model performance, and generalization of the prepared dataset for other‟s use.

Pre-processing: Data processing includes operations like data augmentation, normalization, embedding extraction for non-image data, and data resizing. Data processing improves the model's ability to recognize minute similarities and differences between cases in addition to making the input data more efficient and standardized. It makes sure the model has the greatest information possible, paving the way for accurate and trustworthy similarity computations that support a variety of applications, from image retrieval to recommendation systems. In order to improve data quality, reduce irrelevant information, and guarantee consistent input dimensions, noise filtering, segmentation, and resizing are essential in fine-tuning deep learning models.

V. EXPERIMENTAL RESULT ANALYSIS

A. Dataset Description

There are many species of flowers in the world. Some species have many colors, such as roses, and tulips. It is hard to remember the names of all the flowers and their information. Furthermore, there are some flowers from similar species that can confuse anybody. For example, white Champaka and Champak have similar names and petal shapes, but they have different colors and petal lengths. This data set contains 1000 images of flowers for training and testing under 5 different classes. These classes are : daisy, dandelion, rose, sunflower, and tulip. Sample images of the given dataset have been shown in figure 5.1 with mentioned caption image class to which it belongs.

Fig 9: represent different sample images in a dataset

B. Experimental Result Analysis

Deep Learning Model Implementation: Implementation of carried work is done in 3 steps where the comparative study of accuracy and loss is done and predicted similarity computation of query image was calculated in the proposed model. In these 3 steps, 3 types of models have been implemented for the completion of the task. All models work on deep learning ideas that impart a vital role in machine learning. All implemented deep learning mode was implemented in the cloud platform in which the first two was implemented on google collab and other were implemented in Kaggle with the same dataset. Our proposed model was implemented in two variations and the results of all implemented models are compared in a comparison table on attributes of accuracy, loss, trainable param, non-trainable params, and time taken / step for execution. As the proposed model is complex and deep enough compared to others, so time is taken with some other models but it also has more than no trainable parameters due to augmentation. Transfer learning plays an important role in a proposed model where CNN as a base model is used. MobileNet, Inception V3, and VGG16 net are the important models used in this implementation work.

Model-1 [MobileNet - Linear]: In the first phase of model building a sequential model has been built for identifying the similarity of query images which can be randomly matched with the trained model. The trained model identified whether the predicted level of the flower is right or wrong. This is due to the machine learning model implemented under a single linear layer model having a batch size of 10, train steps of 100 images batch, and evaluated for every 10 iterations. In the Background this model uses the concept of calculating feature vectors and finding the distance between them. This model confusion matrix has been shown below in figure-5.2. Confusion matrix and results of correctly and incorrectly predicted images are shown below. This model is a linear model so specific to the similar kind of images in the dataset will give an accurate result. If a random sample of more no of the image with more classes is mixed up it will not be suitable for that scenario.

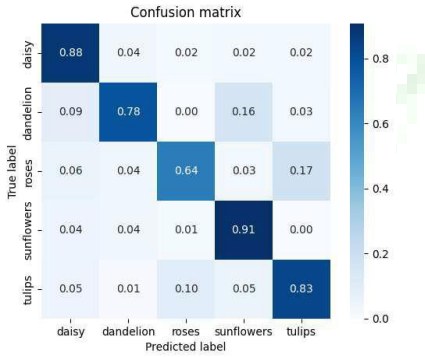

Fig 10: Represent confusion matrix overview using liner mobilenet model for the experimental work

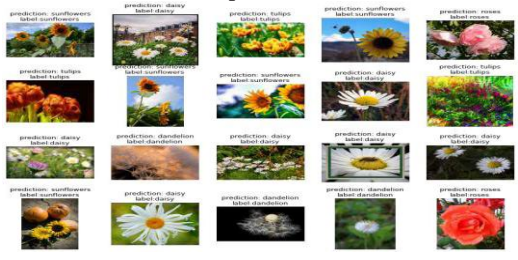

Fig 11: Represent correctly classified predicted image using liner mobilenet model for the experimental work

Fig 12: Represent incorrectly classified predicted image using liner mobilenet model for the experimental work

Proposed Model: [Inception V3]: Inception V3 is a convolutional neural network for assisting in the recognition and classification of images. It is a type of transfer learning where a model trained on a more general problem can be used to solve a related but different problem. Inception V3 was trained for the ImageNet Large Visual Recognition Challenge using the data from 2012. This network worked on the configuration of layers. The model achieves state-of-the-art results on the ImageNet dataset. The model consists of a deep convolutional net using the Inception module that has been trained on the ImageNet Large Visual Recognition. This inception v3 model uses for feature vector calculation and distance calculation is done on sequential deep learning model.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
inception_v3 (Functional)    (None, 8, 8, 2048)        21802784
_________________________________________________________________
global_average_pooling2d (Gl (None, 2048)              0
_________________________________________________________________
dropout (Dropout)            (None, 2048)              0
_________________________________________________________________
dense (Dense)                (None, 1024)              2098176
_________________________________________________________________
dense_1 (Dense)              (None, 5)                 5125
=================================================================
Total params: 23,906,085
Trainable params: 2,103,301
Non-trainable params: 21,802,784
_________________________________________________________________
```

Fig 13: Represent inception v3 model summary for image classification using this model for the experimental work

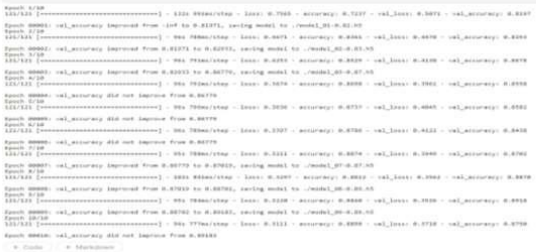

Fig 14: Inception V3 model simulation

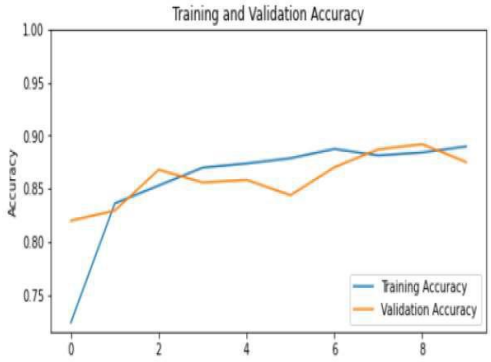

Fig 15: Model accuracy for training and validation

Figure 15 represent test data accuracy with number of epochs for training and validation data based on inception v3 model. Here two axes are defined (x & y), where x axis represent different number of epoch‟s value for the training and validation with respective accuracy value of the model which is mentioned in y axis.

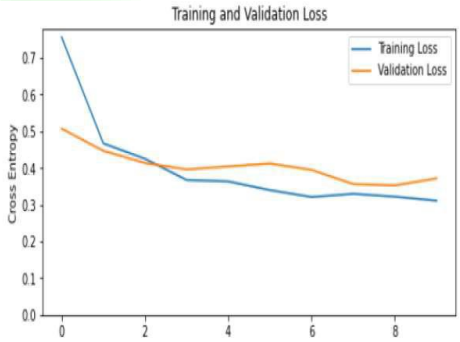

Fig 16: Accuracy and loss graphs in inception V3 model

Figure 16 represent test data loss with number of epochs for training and validation data based on inception v3 model. Here two axes are defined (x & y), where x axis represent different number of epochs value for the train and test dataset with respective loss value which is mentioned in y axis. The trainable parameters improved here and good accuracy is reported by this model as augmentation improves the quantity of the dataset. Convolution neural network works in the background of the stated implemented model. This showcases that more training iteration does not improve the accuracy and loss parameters in results so the nature of the model is not suitable for random dataset similarity computation. Implemented model so far in this experimental work oriented towards findding sufficient accuracy and then the calculation of similarity in prediction percentage. The discussed implemented model results are not sufficient in accuracy so they are not included in progressive work of aimed research for similarity computation in query images. The last model which is implemented as a target is given below.

Model 3 [Proposed model with augmentation and transfer learning]:

This second proposed model implemented the concept of transfer learning of custom network VGG 16. When the training data is insufficient, deep neural networks may be unable to learn properly. Transfer learning can help to

solve this problem and adjust the model to fit the new task. When applying transfer learning, we can look for a few things. In the background part, 16 layer model works for improving the accuracy of the network. There are 5 Block of layers has been created in VGG for feature extraction first then these features are used to identify the exact class of the query image and similarity computation. This proposed model has reached to the optimal enough accuracy to find the predict the similarity of query image from other class of similar images. The comparison table for all the model has been given in table 1. The proposed model is a composite in terms of customization of the recent deep learning model and transfers learning and pre-trained one.

Table 1: Implemented model performance comparison

Model Name	Accuracy	Loss
Model 1 (mobile Net) Linear Layer	80.76	35.37
Model 2 (Inception V3) without augmentation	88.29	32.46
A Proposed model with transfer learning & augmentation	91.16	26.06

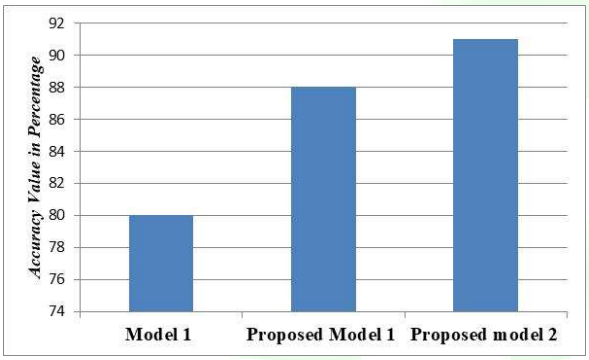

Fig 17: Comparative study of between existing model and current model for image classification and performance parameter is accuracy

Figure 17 represent comparative experimental study of between existing model and current model for image classification and performance parameter is accuracy, Here two axes are defined (x & y), where x axis represent used approach with respective accuracy value of the model which is mentioned in y axis.

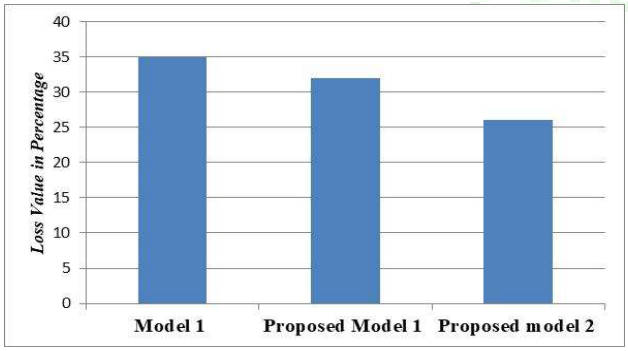

Fig18 : Comparative study of between existing model and current model for image classification and performance parameter is loss

Figure 18 represent comparative experimental study of between existing model and current model for image classification and performance parameter is loss, Here two axes are defined (x & y), where x axis represent used approach with respective loss value of the model which is mentioned in y axis.

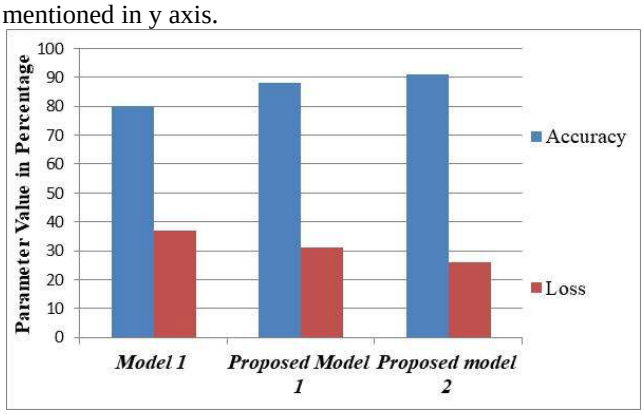

Fig 19: Comparative study of between existing model and current model for image classification and performance parameter is accuracy and loss.

Figure 19 represent comparative experimental study of between existing model and current model for image classification and performance parameter is accuracy and loss, Here two axes are defined (x & y), where x axis represent used approach with respective accuracy and loss value of the model which is mentioned in y axis.

VI. Conclusion

The classification of objects into the right classes has long been one of the most crucial objectives of machine learning or deep learning. Due to the similarities between various things, their textures, colors, and other physical properties, object recognition still presents significant challenges despite the importance of categorizing specific groups of images. In computer vision, object detection has a wide range of uses, such as face and vehicle detection, video surveillance, and plant leaf detection. When conducting studies on flowers, using flowers as medicine, analyzing floral patents, etc., automatic classification of flowers is crucial. Traditionally, lowlevel characteristics like color, shape, texture, and geometry are used to classify flowers. Between floral classes, there are significant intra-class diversity as well as interclass similarities. Because they are dependent on visual search, search engine-based flower identification and categorization methods are not effective and reliable. The feature description has a significant impact on how accurately and robustly flowers are classified. In recent years, deep features have demonstrated good performance on high-resolution photos, but they are unable to extract precise global features from low-resolution images. Deep neural networks have been widely used in computer vision applications because they are effective at identifying picture patterns. An advanced deep-learning model that can precisely calculate the similarity of floral photos is required in the field of flower image analysis. Computer vision and botanical research have significantly advanced with the development of fine-tuned deep-learning models for similarity computation on flower datasets. These proposed models inception V3 and VGG 16 have been applied to tackle the challenge of manually classifying and comparing flower species due to

their diversity and complexity. Similarity computation models have revolutionized the field of botanical research by achieving remarkable accuracy and efficiency in flower image analysis. These models, fine-tuned deep learning models, can handle existing datasets and accommodate new flower species and improved models. This research work states that different similarity computation model; proposed model with augmentation gives better results than existing models.

Future research in this field is expected to focus on enhancing robustness, accuracy, interpretability, and scalability. By improving their capacity to recognize minute visual characteristics, these models will enable them to distinguish precisely between closely related flower species. Additionally, they will be made clearer to promote trust and comprehension of the underlying botanical traits that affect the model's conclusions. A comprehensive understanding of flowers will be given through the integration of multimodal data sources, such as textual descriptions, fragrance profiles, and genetic data, enhancing similarity computations.

References

[1]. Sarah K. Alhabeeb, Amal A. Al-Shargabi, "Text-to-Image Synthesis With Generative Models: Methods, Datasets, Performance Metrics, Challenges, and Future Direction", IEEE Access, 2024, pp. 24412-24427.

[2]. Wenbin Yang, Xueluan Gong, "SwiftTheft: A Time-Efficient Model Extraction Attack Framework Against Cloud-Based Deep Neural Networks", Chinese Journal of Electronics, vol. 33, 2024, pp. 90-100.

[3]. Toan Khac Nguyen, Minh Dang, "Utilizing Deep Neural Networks for Chrysanthemum Leaf and Flower Feature Recognition", AgriEngineering 2024, pp. 1133–1149.

[4]. Dipanjali Kundu, Mahbubur Rahman, "Federated Deep Learning for Monkeypox Disease Detection on GAN-Augmented Dataset", IEEE Access, 2024, pp. 32819-32830.

[5]. Giriraj Gautama, Anita Khanna, "Content Based Image Retrieval System Using CNN based Deep Learning Models", International Conference on Machine Learning and Data Engineering, 2023, pp. 3131-3141.

[6]. Q. X. Zhang, W. C. Ma, Y. J. Wang, et al., "Backdoor Attacks on Image Classification Models in Deep Neural Networks," Chinese Journal of Electronics, 2022, pp. 199-212.

[7]. S. Hong and J. Chae, "Active learning with multiple kernels," IEEE Transactions on Neural Networks and Learning Systems, 2022, pp. 2980–2994.

[8]. Vasileios C. Pezoulas, Grigoris I. Grigoriadis, George Gkois, Nikolaos S. Tachos, Tim Smole, "A computational pipeline for data augmentation towards the improvement of disease classification and risk stratification models: A case study in two clinical domains", Computers in Biology and Medicine, 2021.

[9]. Jun Sun and Qiao Sun, "Bearing Prognostics: An Instance-Based Learning Approach with Feature Engineering, Data Augmentation, and Similarity Evaluation", Signals 2021, pp 662–687.

[10]. Thananya Phreeraphattanakarn, Boonserm Kijsirikul, "Text data-augmentation using Text Similarity with Manhattan Siamese long short-term memory for Thai Language", 2021, pp. 1-7.

[11]. Luca Bertinetto, Jack Valmadre, Joao F. Henriques, Andrea Vedaldi, Philip H. S. Tor, "Fully-Convolutional Siamese Networks for Object Tracking", 2021, pp. 1-16.

[12]. Mathilde Caron, Ishan Misra, Julien Mairal, "Unsupervised Learning of Visual Features by Contrasting Cluster Assignments", 2021, pp. 1-23.

[13]. Xinlei Chen, Kaiming He, "Exploring Simple Siamese Representation Learning", IEEE 2020, pp. 1-10.

[14]. Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie,Ross Girshick, "Momentum Contrast for Unsupervised Visual Representation Learning", IEEE 2020, pp. 1-12.

[15]. Olivier J. Henaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, Aaron van den Oord, "Data-Efficient Image Recognition with Contrastive Predictive Coding", 2020, pp. 1-13.

[16]. Mathilde Caron, Piotr Bojanowski, Armand Joulin, Matthijs Douze, "Deep Clustering for Unsupervised Learning of Visual Features", 2019, pp. 1-30.

[17]. Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, Armand Joulin, "Unsupervised Learning of Visual Features by Contrasting Cluster Assignments", 34th Conference on Neural Information Processing Systems, 2020, pp. 1-23.

[18]. R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, Yoshua Bengio, "Learning Deep Representations By Mutual Information Estimation And Maximization", Published as a conference paper at ICLR 2019, pp. 1-24.

[19]. Mang Ye, Xu Zhang, Pong C. Yuen, Shih-Fu Chang, "Unsupervised Embedding Learning via Invariant and Spreading Instance Feature", 2019, pp. 1-11.

[20]. Mathilde Caron, Piotr Bojanowski, Julien Mairal, Armand Joulin, "Unsupervised Pre-Training of Image Features on Non-Curated Data", 2019, pp. 1-14.

