MATCHING PRODUCTS RECOMMENDATIONS ON UPLOADING AN IMAGE ON AN E-COMMERCE PLATFORM USING NEURAL NETWORK MODELS

HEMANT KUMAR KHURANA

Final Thesis Report

NOVEMBER 2025

Abstract

Online shopping platforms are one of the most used methods for buying products these days across the world. Most of the platforms are making use of text-based search for making product recommendations to the customers. Whereas there is a great scope of image-based search on these platforms which will allow the user to get the recommendations of the products which are visually matching with the product which the buyer is looking for. The buyer can just provide the image of the product which he or she is look for and similar products would appear for purchase. The current study works on to find a model for getting the products recommendations for a fashion industry dataset. To achieve this objective, various methods have been developed over time with currently the use of Convolution Neural Network (CNN) for Reverse Image Search or Content Based Image Retrieval (CBIR) being the most widely used and effective approach. Various models have been applied under the study including pre-trained models for image retrieval on ImageNet dataset including MobielNetV2 model and ResNet50 model. These models have been tested with various similarity measures. Apart from pretrained modes, models were finetuned on classification and image augmentation for contrastive learning. And, finally the ResNet50 model was trained from scratch on the fashion industry dataset. The objective of this study was to test multiple models including pretrained, finetuned and custom build models and identify the best model which can provide the most effective recommendations for a given image input for the fashion industry data for an e-commerce website. And, also to measure the impact of pre-processing and similarity measures. It was concluded that the similarity measures do not have major impact on the model results. The study resulted in ResNet50 pretrained model with image padding and Cosine Similarity as the best model which gave Mean Average Precision of .708221, Silhouette Score of -0,078, Intra-Class embedding distance score of 0.168644 and Inter Class embedding distance score of 0.3596 and resulted in giving the best model for matching image retrieval for fashion industry dataset. The second best model was the ResNet50 model finetuned on articleType class. It gave the mean Average Precision of 0.541431, Silhouette Score of -0,5156, Intra-Class embedding distance score of 0.0548 and Inter Class embedding distance score of 0.2038. The results from this model were very competitive and the model can be further improved by using hybrid approach of using both label and contrastive learning for finetuning.

- **TABLE OF CONTENTS**

- **LIST OF TABLES**

- **LIST OF FIGURES**

- **LIST OF ABBREVIATIONS**

CNN……. Convolutional Neural Network

CBIR…… Content Based Image Retrieval

EDA……. Exploratory Data Analysis

mAP……. Mean Average Precision

MoARR…… Multi Objective Optimization on Adaptive Reverse Recommendation

ORB……..... Oriented Fast Rotated Brief

PCA…….. Principal Component Analysis

RISiC…… Reverse Image Searching in Collage

SSD…..... Single Shot Detector

SIFT…….. Scale Invariant Feature Transform

SURF…… Speeded Up Robust Features

• CHAPTER 1: INTRODUCTION

1.1 Background

E-commerce industry is growing day by day. The buyer while making the decision on an e-commerce website makes search for the products which he or she intends to buy. The most commonly used method is the text-based search under which the buyer mentions the name and/or features of the product in the search box and gets the recommendations for the same. This method is being used on the e-commerce website for a long time. One another method which is used is of voice search where the consumer gives the voice input instead of text and the website understands the voice input and gives the related output. The other method which can be of great use is image-based search recommendation. Under this method the customer just has to upload the image of the product which he or she intends to buy and the search engine automatically recognizes the product and make product recommendations from the database which best match the queried image. In order to achieve the reverse image search, several methods have been tried including Compact Composite Descriptor, Colour and Directivity Descriptor, Fuzzy Colour and Texture histogram to achieve image retrieval task with the best being the application of Convolutional Neural Networks also known as CNN (one type of Neural Network Model) (Diyasa et al., 2020).

There are various CNN models such as Single Image Models including AlexNet, VGGNet, GoogleNet, ResidualNet, Region-Based models such RCNN, Fast RCNN and Faster RCNN and One-Shot models such as Yolo and SSD (Ali et al., 2022; Eswaran and Varshini, 2022; Pate et al., 2023). The CNN models extract features of the image and these features are matched with the images in the database and recommendations are made based on the closest match. For doing this similarity search various methods are used such as Euclidean distance, Manhattan distance, Cosine similarity and Annoy Indexing. The one which gives the highest accuracy and efficiency is selected (Mawoneke et al., 2020).

As reverse image search gives the matching images and therefore, it has found varied applications and researchers across the world have tried making use of the same in solving various problem statement. The reverse image technique is also used for object detection, face detection, speech recognition, license plate recognition or disease detection etc. Some of the problem statements which have been solved using the applications of the reverse image search concept are as follows:

a) At present search engines like Google and Bing have started giving the option to the users to make the search based on input of an image. One research paper has also tested the accuracy of search on these search engines (Bitirim, 2022). It used Average precision and Average Normalised Recalls as the accuracy measures at various cut off points.
b) Some Chinese e-commerce websites have started making use of this method and have also found it useful (Mawoneke et al., 2020). One such website, which is the largest Chinese online shopping company named Taobao have made use of the same.
c) This method has also been used to create a search engine on NASA satellite images (Sodani et al., 2021). Under this paper a trained CNN was used to convert the images in list of integers before fitting into the model. The other techniques which can be used for pre-processing are hashing and vectorisation approach.
d) The other application which has been found is in having web page recommendations based on user data related to searches and clicks (Zhao et al., 2024). Based on the user data the model is trained and then used to make website recommendations on future user searches.
e) One research paper used this method in order to detect the Monkeypox Skin disease. The paper used the opensource data and tried various models which could predict the disease well in advance (Ali et al., 2022).
f) The method has been used to lower the product returns by using circular reverse logistics framework for handling e-commerce returns (Nanayakkara et al., 2022).
g) The method has also found application in cross-border e-commerce to curtail the supply of fake products by developing a counterfeiting Scalable Detection Image system (Onesim et al., 2020).
h) Once such paper has found the application of neural network in forecasting of sales and have also discussed the theory of image recognition in e-commerce (Zhang, 2021).

The current study applies this method to an e-commerce website dealing in fashion products. The best part of reverse image search in the application of product matching is that once a product is seen anytime and anywhere in real life, its picture can be clicked and can be uploaded on the related website or tool which supports the particular reverse image search and get the matching recommendations. For example: if someone sees a person

wearing a specific shirt or pant and he or she want to buy a similar design product; the person can just click the picture from mobile phone and upload it on the website and buy the similar product. Many times, the buyer likes a particular design in a retail shop but are unable to get the proper size of that shirt or pant. The customer can just click the picture of the shirt and upload on the website and see if the required size of that shirt is available online.

Other applications are if someone likes an artwork or design of bag, wallet, purse, jewellery, shoes etc. at someplace, he or she can see the products online by just taking a picture of it. Similarly, if someone like some electronic product, car, utensil; the matching product can be bought online by just click of a picture.

This has great benefits, it saves the time of the buyer to search relevant product, helps in product marketing, increase the sales for the seller, efficient buying for the buyers, product satisfaction and increase the economy. As when the customer can get the product immediate when he or she like it, it is more probable that the product would be bought.

Under current study, image retrieval method is applied to a fashion industry dataset with the objective to make recommendations of matching products for a queried image on an e-commerce website. The fashion industry dataset has been referred from a research paper dealing with the similar topic (Mawoneke et al., 2020).

The study has made use of the pre-trained models including MobileNetV2 and ResNet50 and also finetuned ResNet50 with the dataset using classification and contrastive learning methods. Apart from them, the base model ResNet50 is trained for image retrieval. It illustrates the accuracy and efficiency of each model for matching products recommendation dataset and suggests the best model for use on the e-commerce platform for fashion industry dataset.

1.2. Problem Statement

Till date multiple studies have been conducted on getting the matching products recommendations using neural network methods. In these studies, multiple approaches have been used to achieve the most accurate and/or efficient result for a certain scenario or dataset. This field has been heavily explored since early 1990s (Mawoneke et al., 2020). One of the first implementation were in websites such as TinEye. The other popular examples are Ditto, Snap Fashion, ViSenze, Cortica, Bing Image Search, Google Image

Search, Flickr etc. In the 1990s, the development of e-commerce received government support (Rui, 2021). The American e-commerce company named Amazon at first started the online bookstore. Then improved its logistics and distribution channels and then turned into a global company. They used a Dynamo system, a distributed key-value storage system. In 1992, Golberg Nichols, Oki and Terry created and marketed the first consumer-based recommendation system. Tapestry an electronic messaging platform was designed so that the users can rate messages. In 2000, ecommerce retailers began employing fashion suggestion systems. Until 2007-2008 implementation was primarily in development (Thoiba Singh et al., 2023).

As discussed earlier various methods were tested to get the reverse image search and the best method was found to be the use of CNN models. In 2012, Hinton's research team build the CNN network (Rui, 2021). They build the CNN model named AlexNet which won the championship of ImageNet image recognition competition with an absolute advantage. Since then, various CNN models have been developed which has been the most prominent advancement in reverse image search method.

Multiple papers have tried different methods on different datasets for get the matching product recommendations. Under one such research paper, from which the data has been referred for doing our research, used the CNN model with three layers and six epochs and used Euclidean distance for calculating the similarity score (Mawoneke et al., 2020). This method gave an accuracy of 0.60 based on macro average (each class is given equal weight) and 0.93 based on weighted average (with weights based on occurrences in each class). This problem can be solved more efficiently by testing multiple other CNN models both pre-trained and custom build which are suitable for fashion industry dataset for an e-commerce website. The paper also suggested that Alternative Loss functions such as hamming distance and k nearest neighbours can be used to get the recommendations and the one which provides the most suited can be finally selected. The various methods which can be used to solve the above problem statement and which have been applied in case of similar problems are discussed ahead.

Transfer learning can be applied by making use of pre-trained model on a similar dataset e.g. MobileNetV2 or pre-process the images into frequencies and scalars (Diyasa et al., 2020). This paper compared the method of perceptual hashing of images and then vectorizing them with the pre-trained CNN model. It used the ImageNet 2012 data which

had 1000 classes and 14 million images. It was found that the perceptual hashing method is faster but CNN method is more accurate. The models used in this paper can be adjusted to find the variation in results.

Principal Component Analysis method may be used to filter the image features in order to reduce size for faster images search (Singh and Gowdar, 2021). This method is also suitable in cases where there is some noise and redundant information in the images. This paper used Cosine similarity for loss assessment, ReLU to fit the non-linearity and cosine similarity is used for loss assessment. The image is normalized by dividing by 255 and Conv2d model is applied.

Similar to the fashion industry problem, this technique has been applied to the garment dataset (Eswaran and Varshini, 2022). This paper tests the pre-trained model like VGG16[8], ResNet50[7] and InceptionV3[9] on the dataset and finds that the ResNet50 performs the tasks with the highest accuracy and efficiency. These pre-trained models are based on ImageNet dataset. To speed up the process the nearest neighbour technique is used to convert the feature vector into an efficient indexing format. The paper proposed finetuned ResNet50 model on the custom dataset.

Another better method for image retrieval could be using colour, text and shape features of the images (Bansal et al., 2021). This paper discusses that just colour is not enough for the retrieval process and we should be taking in to account the text and shape features as wel. This paper highlights that different classifiers have different strength and various classifiers can be used to get the image retrieval such as Scale Invariant Feature Transform (SIFT), Maximally Stable Extremal Regions (MSER), Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Speed up Robust Features (SURF). Under the study it was found that the MSER performed better in terms of scaling, PCA is suitable when the images are blur and SURF in terms of saving time.

There are others method for model fitting such as to take into account both Gray and RCB features. Using hash algorithm for expansion of modes and virtual nodes are used to make the data evenly distributed (Rui, 2021).

Hybrid recommendation algorithm can be devised (Li et al., 2022). This paper talks about the methods including indexing of images, using metadata, colour difference histogram, SIFT, spatial direction tree, local patch extraction, vocabulary tree etc. for reverse image

search. This paper used the bag of words model and the first fitted the model through back propagation.

Another paper creates an e-commerce website using AI and ML in the website for both the buyer and seller, introduces chatbot for customers, using Power BI tool for business analysis for seller and applies various methods such as VGG16, VGG19, Exception etc. on the dataset (Pate et al., 2023). It also talked about the use of 5 CNN models such as VGG16, VGG19, ResNet50, Inceptoin50 and Exception on the movie dataset. It tested the models on an e-commerce website created using JavaScript and recommended the use of ResNet50 model.

Under current study, the impact of different methods would be identified including pre-trained models, newly trained models, using different loss functions, different methods for pre-processing of images on the fashion industry dataset, as are used in various research papers on similar problems which are discussed above. A model has been found among various selected models which is best suited for e-commerce website dealing in fashion industry products.

1.3. Aim and Objectives

The aim of the study is to test various models on the fashion industry dataset and to find a model for reverse image search for an e-commerce platform which gives the best matching product recommendations.

The objectives are as below:

- To access the effectiveness of the pre trained models in terms of accuracy for reverse image search on an ecommerce platform for fashion industry.
- To build new custom-made model and compare the results with pre-trained models.
- To measure the effect of pre-processing of data and of different similarity measures on results.

1.4. Scope of the study

The current study focuses on the fashion industry dataset. It has various categories for each image such as gender, colour, type of wear, season for wearing, usage, etc. The scope is to create a website which can take the input and to test various pre-trained and custom based

models which can present the products to the user which are best matching the product in the image which the user has uploaded. For example, if the image input is of shoes, then the user will get the matching shoes from the database which best match the features of the product in the uploaded image. Here the scope is limited to recommending the products which are matching- the features of the product of the uploaded images. The study can be further extended to say get the products which are from the same category such as of same colour, same season wear, same usage etc.

1.5. Significance of the study

The study provides the best model, data pre-processing method and similarity measure which can be used to get most accurate product matches for the fashion industry dataset for an e-commerce website. The better the accuracy of the models, the better is the usability for the customers as they can get the products which are matching their requirements. This method can be used in implementing image-based search on ecommerce platforms along with the text-based search.

The method assists the buyer to not just try to find a product by explaining it in words but by just pulling out the mobile phone and click the image of the product which he/she is seeing and can buy the same or similar product online.

The method improves the economy in e-commerce as there would be less returns for the purchase if the customer is buying the product which is actually matching the requirements. It also improves the overall economy as the buyer is more likely to give positive reviews of the products and drive higher customer satisfaction when the product is actually of one's liking.

And the overall sales also tend to increase as when the customer is able to get the relevant products at the time when the need actually arise, the buyer is more likely to make the purchase in comparison to a scenario when he/she is likely to make a search at a later time period.

1.5 Structure of the Study

The structured in total of six chapters. Chapter 1 gives the background of the problem of reverse image retrieval in fashion industry, gives some idea on what work has been done

around this area, lists the aims and objectives along with scope and significance of this study. Chapter 2 discusses in detail the research which has been done in this area and find the research gaps which are covered in the thesis. Chapter 3 discusses the methodologies available to accomplish the reverse image search including data selection, data pre-processing, transformation and visualization techniques which are available and which are selected for this study. Chapter 4 details the implementation of methods shortlisted in Chapter 3 and provides the analysis. Chapter 5 summarizes the results received from implementation of methods in Chapter 4 along with comparison and details the performance of the models. And finally, Chapter 6 gives the conclusion of the study including which model, pre-preprocessing and similarity measure is the best for reverse image retrieval for fashion industry data and also gives the future recommendations.

- **CHAPTER 2: LITERATURE REVIEW**

2.1 Introduction

In order to achieve the ask of image retrieval various studies have been conducted around the Content Based Image Retrieval (CBIR). These studies cover various area of application including the fashion industry, ecommerce in general, medical field, satellite image data, advertisement, search engine image search, library science and digital forensics etc. There are various methods used including the traditional methods of using the hashed data, metadata including text or description, using histogram of colours to compare the colour of pixel data and the Content Based Image Retrieval method. This is achieved with the manual feature extraction methods including Scale Invariant Feature Transform (SIFT) or Speed Up Robust Features (SURF) and the latest method of Convolutional Neural Network methods. The following section discusses various paper which has looked into these application areas including fashion industry, followed by the methods used to achieve the task the task to image retrieval.

2.2 Application areas of Reverse Image Retrieval

The Content Based Image Retrieval technique has been applied across research problems covering various industries and areas. Several research papers have done thorough studies on this. One such study has been done on the fashion industry using CNN (Mawoneke et al., 2020). Under this paper the research was done on the fashion industry data. Similarly, another paper looked into the fashion recommendation system (M Vinitha et al., 2024).

Along with fashion industry application, the image retrieval method is also used specifically for the garment dataset (Eswaran and Varshini, 2022). The idea is that the customer on the website can use image as input for searching the products.

Another application area is of fashion forecasting (Thoiba Singh et al., 2023). Many people are involving in fashion but the industry involves change over the period. There are many sorts of fashion lifecycles. Fashion picture analysis is helpful for a variety of fashion related applications. Many website predicts the fashion trends including Heritech, Wgsn and Trenzoom among others. There are two type of forecasting short term and long term forecasting which is typically of 5 to 10 years.

Till now the application of the method is tested on single images. One paper tested the application of the technique of collage which is a single shot which captures multiple information from where images and present in a single view. Apart from fashion or garment industry application. The image retrieval also has importance for the e-commerce industry in general.

One another paper used deep learning architecture on E-commerce products recognition (Zhang, 2021). On ecommerce application, another research was done on framework and techniques for Image Based Search application with an Ecommerce website (Bansal et al., 2021). Another paper did research on classification of Cross-Border e-Commerce Products based on Image Recognition and Deep Learning (Rui, 2021). One more paper discussed the circular logistics framework for handling the e-commerce returns (Nanayakkara et al., 2022). It discussed how the data is collected on variables to reduce the returns on e-commerce and how the variables are fit in the neural network to train the model and reduce returns. Similarly, one more paper worked on the e-commerce webpage recommendations scheme based on semantic Mining and Neural Networks (Zhao et al., 2024). Under this paper, neural networks are applied to get the webpages recommendations. The user data of search and click is used to train the model and then is used to give the best webpage recommendations. The machine learning as a tool for enhancing eCommerce solutions was also done by another paper (Stoica and Pelican, 2025). It highlights the importance of the need to explore the AI techniques within the ecommerce domain in order to cop up with the competition. The machine learning is used for recommendation system, fraud detection, analyse customer reviews, ratings and detect fake reviews. One more paper tested an image recommender system for E-commerce (Addagarla and Amalanathan, 2021). They used a e-SimNet network with deep learning techniques and index tree using ANNOY algorithm.

Another application of this method is in advertisement on the websites (Patil et al., 2022). Videos platforms use metadata but it is very time consuming and effect the performance of the app and they do not use object detection. The paper worked on to identify the approach to recommend products with mean Average precision of 51.3%.

The other application is to solve the counterfeiting problem in e-commerce. To achieve this objective information sharing is required so that the original products can be differentiated from the fake ones. The method assists to identify whether product is original or not (Onesim et al., 2020). The counterfeiting also is part of terrorism funding and therefore to

solve this is issue is of great importance. Counterfeiting takes away the profits of the original sellers. They exploit the technical hurdles in advertising. Due to lack of knowledge among the consumers they are able to take advantage of the same. To solve this issue information sharing is required. The methods used is to develop a specialized large scale image search engine callsed e-CoS. The E-CoS has two steps, first was indexing and second was searching. Indexing is creating an indexed information of each original photo and then the difference is calculated from the original product photo. If the different is below the threshold then the uploaded product is a fake. If the counterfeiter makes any changes in the product, this technique helps the ecommerce website to catch it. This technique is required to be adopted by China as there is a lot of counterfeiting possibilities while ordering products from China by other countries.

Apart from the fashion, garment, ecommerce industry the image retrival methods has been of parament importance in various other areas. The method is used by google in order to retrieve the results by Google (Bitirim, 2022). In fact, the method is used to get the imagery data from NASA satellite database (Sodani et al., 2021). The research paper created a search engine on the NASA worldview satellite images which would reduce the search time from weeks to minutes.

The method is also used to detect the diseases. One such study used the method to conduct the diagnosis of monkey pox skin disease (Ali et al., 2022). The monkeypox fatality rate is 3% to 6%. Under the study many pre-trained models were tested and a webtool was developed do the online screening. This study proved the usability of CBIR in early diagnosis of the disease.

Another application area is in the Library and Information Science (Adrakatti and Mulla, 2023). The paper discusses the research gap in Multimedia information retrieval (MIR) where Library and Information Science (LIS) can contribute. The main challenge to provide the retrieval by content i.e. to provide results not only based on metadata, textual descriptions but also on the content of the objects. From 1975 to 2021 around 9345 articles were published on Multimedia Information retrieval according to Scopus report and 70% of the articles were published in computer science and engineering domain. The research paper does critical evaluation of MIR in the Library and Information science space. The information retrieval models include classical of Boolean, Vector Space, probabilistic and non-classical of cluster and alternative of Fuzzy logic, Neural Network models. The open-

source application of library automation and institutional repository still use classical Boolean techniques. Search of metadata. They are not keen on retrieval of document stored but are in storage and organising. The various domain which includes image retrieval are media libraires, e-commerce, remote sensing, art & architecture and investigation. Image retrieval is of two types a) text based i.e. description, keyword and b) content based i.e. shape and texture. It is high time for LIS to use content-based techniques to find the image on colour, shape, texture and not limited to metadata. The application area is audio retrieval where audio is retrieved in linear manner as audio signal is one dimensional and on time. Audio retrieval for library, e-media, broadcasting channel and musical. Video retrieval refers to search capability for the entire archive of digital video content. It is used in education, media, security industry and entertainment. It is high time to adopt MIS in LIS. Search is currently limited to metadata. The search engine for audio platforms is restricted to bibliographic details only. User find it difficult to search from large database as they do not support content-based audio retrieval techniques.

The method is also used in digital forensics (Salman and Hasan, 2023). Without the precise data gathering and analysis cyber security has grown more challenging. The science of forensics oversees this. Methods, applications, difficulties and tools are reviewed by this paper. Electronically Stored Information (ESI) is term used to describe this collection of electronic or digital papers or other evidence. Internet of things, mobile devise, communication networks, cloud-based services and cyber physical systems are the places where the digital information exists. There were estimated 3.6 million fraud cases and 2million computer noise violations every as per office of national statistics. Digital forensics refers to the use of scientific techniques from the field of computer technology. The related research has been conducted around extracting emails from the hard images, especially with increasing volume of data. An autopsy software module for distributed identification of emails files which is results in outperformance by 52%. Digital forensic were applied to image using steganography techniques. Another report looked into WhatsApp web to use it as evidence. Another was to identify the fake images while using the data as evidence of eyewitness. The process of forensics includes preservation of the crime scene, prohibition of action which can harm digital information including deletion of data or damage of hardware. Next is to identify of what happened and how the digital data is misused or harmed. Both hardware and software skills are used. Then is the Data Collection where data is gathered from all available sources as proof and then the original data is duplicated and

examined. Lastly, the analysis is done and the reports are provided. It is important to find the erased, obscured or transfigured data is important. Now there are different types of forensics which are done, one is involving network and communication channels. It is done by capturing packet data. Both in and out is monitoring to find the origin of attack. Other one is mobile forensic done on data on the device, here the focus is on user privacy. Then comes the disk and storage devices involving classical methods. Cloud data forensic also requires grasp of network, communication protocols and cloud architecture. The challenges in this are the encryption is used while transferring the data, size of the data, location of the data, sometimes proxies are used and the data is erased after the crime scene.

2.3 Methods for reverse Image retrieval

There are various methods to achieve the task of reverse image research including the old methods of using metadata, histogram, text data, colour difference histogram and the latest methods of extracting feature using manual approach or automated machine learning using the CNN methods.

2.3.1 Initial Methods

Various papers discussed the use of old methods of image retrieval. One such paper discussed various methods including text to text, text to image and finally the image-to-image method on garment dataset (Eswaran and Varshini, 2022).

Another research was done on framework and techniques for Image Based Search application with an Ecommerce website (Bansal et al., 2021). The paper discussed the traditional techniques including the text-based search and then currently the image-based search which is being used.

One such paper looked into the methods of Deep learning for improvement in reverse image search (Singh and Gowdar, 2021). The methods of Content Based image retrieval discusses were comparing patches of images, identifying plagiarism and calculating histogram of RCB values and comparing.

Another paper compared the perceptual hashing and pre-trained CNNs. There are various method including Compact Composite Descriptor e.g. Colour and Directivity Descriptor and Fuzzy Colour and Texture Histograms and the other is CNNs (Diyasa et al.5, 2020).

For feature extraction RGB colour extraction can be used. Under RGB, there is a histogram for each colour and each graph shows the pixel intensities for each colour (Yadav et al., 2024).

2.3.2 Manual Feature Extraction

The features extraction from the images can either be done via automated machine learning techniques using CNN or via the manual methods. Many papers have discussed these methods. Under one such research paper the feature extraction methods were applied to get the collage. Collage is a condense visual data in a single view. The features of the images are extracted using SIFT, SURF and ORB methods. The features are localized for region of interest by binning technique. Manhattan distance is used as similarity index and it achieved the accuracy of 90.96% (Zubair et al., 2023). The content-based image retrieval can be usage in two ways, first is to get the exact duplicates of query image or variants and the second is to get the similar images. The features are extracted and then matching images are found. Features extraction method used are HOG, local binary pattern, Scale Invariant Feature Transform (SIFT), Speeded Up Robust Features (SURF), Oriented Fast Rotated Brief (ORB). The paper proposed SIFT, SURF and ORB. Reverse Image Searching in Collage (RISiC) is not currently being addressed in proposed works. Google, Yandex and Bing have the technique but it is not public. Current paper uses SIFT, SURF and ORB for feature extraction, to find the region of interest Median of Mode method is used. Finally, to validate the updated region of interest, a Euclidean distance is used on the colour ratio set. Some papers have proposed colour-based approach and some use multiple colours encoding to achieve this task. Under the selected method, once the features are extracted, the matching of features is done between vectors. To avoid the false match, region of interest is found. Median of mode is used after the frequency distribution for finding the dense region. After this similarity check is done in the collage. To evaluate the models, 9164 images with 101 categories was used. From this data 372 images were selected for the construction of 30 collages. With image flips in preprocessing and using SURF, it resulted in accuracy of 53.56%. This was the highest accuracy achieved amongst other models including SIFT and ORB methods.

Another research was done on framework and techniques for Image Based Search application with an Ecommerce website (Bansal et al., 2021). It used the Content Based

Image Retrieval technique. The image contains so much of information that it is equivalent to 1000 words. There are various methods of image retrieval. The images with higher resolution would give better results due to the information which is contained therein. In order to retrieve the image, the image features are extracted, distance is calculate using the similarity measure. The various classifiers which the paper has discussed included Scale Invariant Feature Transform (SIFT), Maximally Stable Extremal regions (MSER), Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Speed Up Robust Features (SURF). The image has different features and just colour is not enough for retrieval process. Different classifiers have different strength. The method used under this research paper had used colour, texture and shape features. It used different similarity measures to test the model. Under the paper MSER model better in terms of scaling, PCA performed better when the image is blur and SURF in terms of time efficiency.

2.3.3 CNN method for image retrieval

The latest methods used for image retrieval is the the Convolutional Neural Network (CNN) which is widely used method for achieving the reverse image search. It generally have a structure which includes selection of the data, pre-processing of the data which may include image resize or normalization, after the data is pre-processes it is processed through a selected CNN model. The feature embedding are output of the model. Under some architecture. The features are localised using various techniques so that the retrieval accuracy and application can be improved.

One such paper looks into the methods of Deep learning for improvement in reverse image search (Singh and Gowdar, 2021). CNN in python is applied using Scikit-Learn, TensorFlow and OpenCV. Steps involved are preprocessing, image segmentation, feature extraction and classification result for matching. The related works done in this area include corner detector, local binary patters and geometric transformation method. The pre-trained models are also used for the CBIR. The paper used PCA to filter the features from 2048 to 256 pixels for faster image search. ReLU is used to fit the non-linearity. Cosine similarity is used for loss assessment. The pre-processing techniques include taking Null as zero and the image is normalized by dividing by 255. The model achieved the accuracy of 91.26% and sensitivity of 92.56%.

Another paper looks into the Reverse Image Search based on pre-trained CNN models (Diyasa et al.5, 2020). The reverse image search is one type of Content Based Image Retrieval method. The methodology used by the paper included using ImageNet 2012 data which has 14 million images with 1000 classes. Method 1 is converting the image into hashing and then vectorising it and other method is using the pre-trained CNN model. Perceptual hashing is much faster and CNN is more accurate. Even in CNN there is high difference in Euclidean distance and therefore more research needs to be done for better fit of the model.

The CNN method has been applied on the fashion industry dataset (Mawoneke et al., 2020). Under this paper the research was done on the fashion industry data. It used fashion industry database and applied a 32 features detectors of size 3*3 pooling of 2*2, first layer of 128 and second layer of 64 and third of 45 neurons with 6 epochs. The data is split in 38,000 and 6,446 images. It achieved the accuracy of 0.60 as macro average where each class is given as equal weightage regardless of how many occurrences are within it and achieved the accuracy of 0.93 with weighted average. In order to achieve this result, the similarity measures were used including Euclidean distance and cosine distance and both gave the similar results. It concluded that the features extraction has more role to play in accuracy rather the similarity measure. This research can be further extended using other similarity measures including hamming distance and k nearest neighbours.

Another paper looks into the techniques of reverse image retrieval (Yadav et al., 2024). The reverse image search is introduced on a dataset of 50,000 photos in 25 classes. Method of auto encoder and CNN is used. The framework is accurate and speedy. The paper discusses on deep learning and indexing for reverse image search. The methods used are the EfffficientNetV2 model along with KD-Tree where EffcicientV2 generates the feature vector and KD Tree acts as an organizer which arranges feature vectors and allows quick searches. KD Tree quickly and effectively locates the closes neighbours in the feature space. After the feature vector is extracted, Principal Component Analysis (PCA) is used to reduce the dimensions. Then the reference point matrix which is the similar size of vector of other images is compares with the query images feature vector. For feature extraction pre-trained CNNs can be sued or RGB colour extraction can be used. Indexing involves the organizing and structuring the feature vectors and finally the similarity modelling is done. The paper tested four architectures including

EfficinetNetV2, ResNet, VGGNet and Contrastive Language Image training by Open AI. ResNet gave the accuracy of 98.73% which outperformed other models.

Another paper tested the CNN method on garment dataset (Eswaran and Varshini, 2022). The paper referred to the Yahoo shoes dataset and kitchen appliances dataset which used Inception V3 model for feature extraction and Euclidean distance as the similarity measure for feature matching. In this deep leaning, the paper proposed the ResNet50 model which was finetuned to the custom dataset. They stopped the CNN at the layer which generates the features vector. The paper also referred to the VGG16, ResNet50 and InceptionV3 which are based on ImageNet database which can be used for feature extraction. To speed up the process the nearest neighbour technique is used to convert the feature vector into an efficient indexing format. In order to conduct the study, the researcher converted the images in to RGB from and then converted to numpy representation. The pre-trained model is finetuned to the garment dataset which is created by collecting the images. VGG 16 model was eliminated as it was giving the highest average Euclidean distance and had higher size which limit the portability. ResNet50 models gave more accurate results than InceptionV3. They added the layers to the ResNet50 model and finetuned to get the lower processing time. The similarity score matrix is selected based on which takes the lowest time.

Another paper looked into the fashion recommendation system (M Vinitha et al., 2024). The paper tries to find a novel approach to fashion recommendation. It uses collaborating filtering matrix factorization methods and neural networks. It demonstrates the system effectiveness and user engagement. The goal is to provide a tool that combined deep learning and advanced indexing method to provide the accurate and efficient image retrieval. CNN were used along with KD Trees to facilitate the nearest neighbours search. The system retrieves images from the database in response to a image submission. The system is valuable and is a game changer.

The CNN method is also tested for fashion forecasting (Thoiba Singh et al., 2023). The related works include the scanning method based on lasers and projection system based on moire, laser technique create a 3D imaging. There is potential for improvement in the 3D imaging data ability to extract feature size for garment design. Another researcher was able to drive the feature size needed for clothing by integrating 3D model of human figure. Mesh facet approach is used on 3D physical modelling and uses points,

edges and surfaces to product 3D surface models. Surface points are easy to collect. There have been models around geometric equations but to convert the model from 2D space to 3D space is complicated. It requires to take care of the dimension of human body while designing the product. Multiple models have tried using converting the garment into particle system, energy technique which characterizes the fabric into energy equations. The paper used ResNet to get the matching images from the database. It only builds the system using this model and did not do any quality or accuracy testing.

Another paper used deep learning architecture on E-commerce products recognition (Zhang, 2021). The paper used time series forecasting, multivariable timeseries and LSTM framework and theory of image recognition. Traditional time series model has difficulty in handling the seasonality, dynamism and periodicity of data. Therefore, LSTM time recurrent neural network is suitable for forecasting and processing important events. Deep learning is used in face recognition, speech recognition, license plate recognition and object detection. In order for model to train efficiently the database need to have images from different angles. The paper used LSTM in prediction of sales. The sales data of luggage company is used and 400 is taken as training and 50 is taken as test data.

Another paper did research on classification of Cross-Border e-Commerce Products based on Image Recognition and Deep Learning (Rui, 2021). It looks into the new product classification technology for cross border features and takes into account both Gray and RCB features. It referred different methods of image segmentation including AlexNet, LeNet-5 and used the Resent50 and InceptionV3 model in the research paper. The results were stabilized around 100 epochs.

One more paper discussed the circular logistics framework for handling the e-commerce returns (Nanayakkara et al., 2022). It discussed how the data is collected on variables to reduce the returns on e-commerce and how the variables are fit in the neural network to train the model and reduce returns.

Similarly, one more paper worked on the e-commerce webpage recommendations scheme based on semantic Mining and Neural Networks (Zhao et al., 2024). Under this paper, neural networks are applied to get the webpages recommendations. The user data of search and click is used to train the model and then is used to give the best webpage recommendations.

The transfer learning-based recommendation was tested by building a website with product search, reverse image search, sales analytics and chatbot for user experience (Pate et al., 2023). It referred the papers which used VGG16, VGG19, ResNet50, Inception V3 and Exception models, created search applications, did testing on movie datasets. The research paper worked on the ecommerce website which offers machine learning, Artificial Intelligence and Data Analytics for both buyers and sellers. It used the JavaScript for building interfaces, price-based recommendation, used ResNet50 model, Power BI for data analytics and Land Bot application for Chatbot.

The machine learning as a tool for enhancing eCommerce solutions was also done by another paper (Stoica and Pelican, 2025). It used .NET 8 for backend, angular frontend, python for API and machine learning logic and SQL server. The security was prioritized using hashing, encryption technique and by using clean code. Python based API was responsible for voice to text, text to image, reverse image searching, use based and content-based recommendation system, stripe integration which processes transaction and communicates the result back to the application, created table in SQL database. .NET web API handled the independent module design, business logic, data access and user interface is handles separately. The data used is Mobius 2024 about books and reviewers. It uses the recommendation based on collaborating filtering which make the recommendations based on the aligned preferences using average rating. The metrices used are Prediction Coverage which equals Number of unique recommendations divided by Total number of items, another metric used are Catalog Coverage, Precision which equals Number of Correct recommendations/Number of Recommendations, Recall which is Number of Correct Recommendations divided by Total correct relevant items and F1 score. The paper developed the prototype which can be enhanced for production.

Another area where CNN is used is in advertisement on the websites (Patil et al., 2022). The paper worked on to identify the approach to recommend products with mean Average precision of 51.3%. In this research it is discussed to use Yolo object detection algorithm to detect the objects and providing the links for selected object. Face detection, shape detection, colour detection, skin detection and target detection have been implemented for object detection. The researches which have already been done include the use of ‘Local detection-based Background Subtraction’, target recognition detects the desired object from the video (LIBS). But it was found to be less accurate

for little things. Another model recommended is use of light weight network model called MobileNet. It used the TensorFlow object detection API but the model was too slow. Another method which is used is fast indexing and frame interval, using HOG shape context and Hu moments to extract a shape-based representation from the image. Another paper conducted comparative study of different object detection techniques such as AlexNet. VGG16 and InceptionNet, where AlexNet outperformed with accuracy of 78%. Another method was Recurrent convolution neural network modification. One of the authors employed Yolo in conjunction with GoogleNet model resulting in reducing parameters and faster detection. Another author used TensorFlow to object detection of drowsiness using YoloV3 and Single Shot Detector (SSD), built boundary boxes to build the discovery model such as height, breadth and class name of each photo. YoloV3 is fast and there are not many services available to find the object in the video apart from tags a title which are cumbersome. For the study YoloV3 model was chosen along with the MS CoCo dataset which had 80 classes and 300,000 photos. New model is difficult to build due to complexity of image and video processing. Content Based Image Retrieval (CBIR) is used for reverse image search. The research paper first prepared the framework for API call, the backend server and feeds the frame to the object detection query. It does the pre-processing and object detection. On detection the server uses bounding boxes to isolate all the objects from the original pre-processed frame, allowing individual object to be processed separately. The system ensured the object in focus are detected. Then each object is subjected to Content Based Image Retrieval search. With the results stored by best search match. This allows the user to click on the object and the ecommerce link. The module was able to detect the mean Average Precision of 51.3%.

The CNN method is also used in an image recommender system for E-commerce (Addagarla and Amalanathan, 2021). They used a e-SimNet network with deep learning techniques and index tree using ANNOY algorithm. Fetched N similar items using distance measure and achieved the accuracy of 96.22%. The people have started buying online and the traditional ways do not work properly for products recommendations. The performance of the method is checked using accuracy, error rate, silhouette coefficient and Clinski-Harabasz score. The research paper proposes the CNN method. In other related papers, local feature region method is proposed, bag of words using SIFT is proposed, deep learning is proposed, deep hash is applied to the feature

embedding. Manual feature extraction is done using SURF, indexing is done using VP Trees for computing distance. The research paper used e-SimNet model which uses SqueezeNet as CNN. It used three strategies, first to convert the 3*3 filter to 1*1, second keeping the lower channels to 1*1 filter and down sampling at later stages. These strategies reduced the 1,248,424 parameters to 421,098. ANNOY is implemented to find the nearest neighbour and is implemented to find the top N images. Further Euclidean Distance is used to get the top N recommendations. The paper used the fashion dataset with 10 categories and pre-processed it. SqueezeNet outperformed with accuracy of 96.2 accuracy and 0.0378% which was better than ResNet, VGG, EfficientNet and DesneNet.

The method is used by Google in order to retrieve the results by Google (Bitirim, 2022). The paper tested the effectiveness of the method in Google search results. It tested the results by putting 25 images in google search and 100 matching images are extracted for each sample. The accuracy was measure on Average Precision (AP) and Average Normalized Recalls (ANR) at various cut off points. It used google chrome, google images, search by image and the url of the image of input to test the results.

The method is also used to get the imagery data from NASA satellite database (Sodani et al., 2021). The research paper created a search engine on the NASA worldview satellite images which would reduce the search time from weeks to minutes. The images were converted to features and scaled. The image size was reduced from 2048 to 128 and a snipping tool (UX) was developed for the queries which enhanced the worldview tool efficiency. The paper talked about the related work including the decision trees. They used two approaches first was using metadata and another was using image features. The limitation of the metadata approach is that, the generation of metadata would have to be done by hand. In second approach, a trained CNN is used for converting the image in list of integers. The ResNet model was used and a dense layer was added to ResNet for increasing the accuracy. In order to conduct the study 52k images were taken from NASA and features were stored in 128 features vectors. Vectors are stored in .ann annoy files which assist in search. Annoy search gives indexes of images which matches the query image the most. The indices were converted to url and are returned to the user. The issue here was only that the CNN was trained on the classes of labelled images it has been trained on. Goring forward self-supervised model can be developed to find images without specific classification and labelling.

The method is also used to detect the diseases. One such study used the method to conduct the diagnosis of monkey pox skin disease (Ali et al., 2022). Under the study many pre-trained models were tested and a webtool was developed do the online screening. The GPU and TPU resolved the problem of computing time. It introduces the open-source dataset from web-scrapping of images. The related work has used various models including VGG16, ResNet50 and Inception V3. The images were cropped to 224*224 pixels and 228 images were augmented to 14 fold. The paper conducted various experiments including VGG16 model which contains 3*3 convolution filters, based on residual learning concept, convolution was followed by batch normalization and ReLU non-linearity. Architecture had varied depths; experimentation was done with 4/6/8/12 of bottom layers for better homogeneity and generalization and finally 8 layers were chosen. After flattening, 3 fully connected layers and dropout layers was part of the model. Fully connected layers had 4096, 1072 and 256 nodes and dropout factors were 0.3, 0.2 and 0.15. Finally binary classification was used. The implementation was accelerated using K80 GPUs. Batch size was se to 16, Adam optimizer was used and initial learning rate of 10^-5 and binary cross-entropy loss function was employed for learning. The performance measure used were accuracy, precision and F1 score. In this paper, ResNet50 gave the best accuracy. The issue in this study was that the data was limited. The pre-trained model weights can further be updated using the multi dataset to get the better results. For achieve this objective international collaboration is required to extend the dataset. This study proved the usability of CBIR in early diagnosis of the disease.

2.3.4 Advanced Methods

Another paper looked at the optimization of multi-objectives for Adaptive Reverse recommendation (Dcosta et al., 2022). It is a technique which provides the optimum Neural network considering various objective such as filters and parameters optimization, accuracy, reduction in time and less resource usage. Neural Network architecture enables to search for the best architecture for the given problem. This consumes a lot of resources and time as it goes through all kind of architectures. The paper found an architecture which is small in size, memory efficient, faster with minimal loss of accuracy and with optimization the parameters are decreased. Automated machine learning strives to deliver effective pre-processing, optimizing the model parameters, transfer learning, feature extraction and feature selection. The Neural

Architecture Search (NAS) has three components, a) Search Space which has different layers combinations, search strategy which is a sample from the search space and evaluation strategy which is the accuracy, reduction in time and less resource usage. NAS method includes ENAS, DARTS, NSGA-Net. Multi objective optimization is based on adaptive reverse recommendation (MoARR) which is a technique that aids the creation of light weight architecture. Then a unique multi objective strategy is developed to effectively evaluate historical evaluation information (HEI). The paper applied Soft Filter Pruning Strategy to the architecture before HEI and obtain the pareto optimum architecture. The datasets which are compare are CIFAR-10, CIFAR-100, STL-10 and Food-101 and analyse how MoARR method performs on these. Multi objective optimization (MOO) has objective function to maximize with variables and objective functions. Pruning helps optimizing the Neural Network. They are of two types filter pruning and weight pruning. Filter based approach trims the filters. NAS looks for all types or architectures to find the optimized architecture. Therefore, there is a need of improvement. MoARR excludes and avoid searching for the worst architecture and therefore saves time and resources. It studies the relationship between parameter quality and accuracy. The pruning was tested on various datasets and resulted in saving of parameters and increased accuracy.

One paper looks into the hybrid recommendation algorithm of cross-border e-commerce items based on artificial intelligence and Multiview collaborative fusion (Li et al., 2022). In the paper Content Based Image Retrieval is analysed and an algorithm is proposed. One of the researchers proposed a new image indexing and retrieval algorithm that uses four local patterns. Another proposed colour different histogram. One proposed the mean-shift algorithm on the scale feature transform which is suitable for video analysis in real time version systems. Other proposal is to combine colour histogram and spatial direction tree. SIFT method extraction location, scale and direction aspects, each point has a centre of 8*8 vector along with bag of words and the model is fit through backpropagation. For experiment the data is divided into 5k and 7k images along with using indexed weights. Different thresholds are used for taking the feature point in bag of words. Difference in speed and accuracy was found.

One paper looked into the clustering techniques after the features are extracted from the model for image recommendations (Laamouri and Sael, 2025). It is important for the image recommendation to match the aesthetic preference and visual taste. The

application areas include social networks, art, design and tourism. It involves grouping of products, movies and articles. In this paper the clustering techniques of visual recommendation are explored. The challenges faced in image recommendations are around the accuracy and segmentation partitions. Clustering comes into rescue as it reduces the noise and use clusters for suggestions. The related work around this area referred in the paper include the L mean algorithm, on dataset of garments which resulted in specificity of 0.97. For food recipe dataset with CNN for feature extraction and K means for clustering the recall achieved was 0.795, precision of 0.775 and FI of 0.785. Another paper did latent profile analysis and categorized the products into clusters and took the top 10 recommendations, which gave the recall of 0.0688, Fi 0.0713, precision of 0.0732 on the data of recipes and recall of 0.0637, precision of 0.0654 on data of food dataset. There are new clustering techniques, RCB images, content filtering, clustering models such as Fuzzy C-means, RGB components, texture values and Euclidean distance. There is also a comparative study done across K means, hierarchical clustering, portioning around medoid, in which the K means outperformed. Other method is the online binary k means methos which is tested on the content filtering on the ILSVRC2012 dataset, ILSVRC2012 dataset which is a subset of ImageNet dataset. The models used in this comparative study were ANN and hashing for feature extraction and for clustering model it used K means, BK means, KDK means and K medoids. According to the literature review it was found that the most common technique used for clustering are K means, K medoids, MiniBatch k means and Birch. In the paper, on the fashion industry dataset preprocessing was done including resizing and normalisation. Feature extraction is done using MobileNetV2 model. Clustering is done using three methods and visualisation of clustering is done using t-SNE. Images are resized to 128,, 128, 3. Normalization from 0 to 1 is done. K means cluster is done based on extracted features and each image is assigned to the cluster whose centroid is the closest. K-medoids minimises the sum of dissimilarities between points and their medoids and therefore more to robust to outlier. Mini batch K means handle data in min batches to update the centroids for large datasets. Optimizing of the cluster is done using silhouette coefficient and the k value that maximizes the score is taken. In the study, k means performed the best and therefore it is chosen. K equals to 9 were chosen as the variety in the dataset was high. After that the top 10 images are extracted from the system. The system first finds the cluster to which the images belong and then find thee matching images.

One paper specifically looked in to the Image preprocessing. The paper does the comprehensive analysis on the image search engines (Singh et al., 2024). The focus is one the image qualifies and fixing blurriness, low resolution and noises in the images. The paper evaluated colour histogram, wavelet transforms, neural network architecture, on INRIA holidays and Corel datasets. Grey wolf optimization for feature extraction and sophisticated deblurring for image improvement are also features in the survey. It has applications in ecommerce, health and security. The paper specifically looked into the techniques to improve the image quality. The other referred papers have studied the application of 3D histograms in the HSV colour space, wavelet transform and local binary pattern features, optimized through grey wolf algorithm. The research is conducted using multiple datasets. It used methods including 3D colour histogram, CNN and deep learning. Along with this wavelet transform, local binary pattern and grey wolf optimization is also tested. The models used included the ResNet, Inception V3 and DenseNet and precision, recall were used as accuracy measures. It resulted in the retrieval rate of 97%.

Another important concept is of vector database management systems(Taipalus, 2024). The unstructured data is stored in vectors where each number in the vector represents the features of the image. It has gain popularity due to reverse image search, recommendation systems and chatbots. The vector database management system is not strictly required. But it makes it feasible for other business resources to access control, automat database scalability and query optimization. Vector similarity search is complemented by metadata filters and searching with multiple query vectors. The use case of vector database management includes the closest geolocation from current location which also uses the similarity measure. The higher dimensions such as image, audio and video features are used for testing. The application is not limited to querying and include manipulation of data, collection of metadata, indexing of access control, backups, support of scalability, interface with other systems and offering support for data models. It also has application in Natural language processing. To test the effectiveness similarity is used as the measure. The different techniques for vectorization include the a) products quantization which divides the vectors into smaller parts, b) locality sensitive hashing which hashes similar vectors to same buckets, c) hierarchical navigable small world which creates a hierarchical graph, R-Trees which creates hierarchical structure with bounding boxes, KD Trees which includes binary tree

portioning along dimensions. The vector data management system searches for the nearest neighbour vectors. Various vector DBMS available are Pinecone, chroma, Milvus, Weaviate, Vald, Qdrant and Deep Lake. Vectors are also part of various other data management systems including PostreSQL, MongoDB, Cassandra Redis and SinglStore. The use cases include text, video and chatbots. In generative models memory is not there, so they are linked to Vector database where the queries are stored in vectorized format and are used as long term memories. The challenges in this method are that although it reduces the storage requirement in some cases but computation cost are there. Also, there is a trade off between accuracy and response time. With higher dimensional vectors there may be need for more storage. The other challenge is that the data is spread over multiple vectors.

Another paper explores the fine-grained method for fashion image classification and recommendation (Nayak et al., 2021) . The transfer learning method is proposed. Both vendor and customer face challenge in online purchase and sale. The vendor has to manually upload products, tag them and add description. Customer face problem as it has to enter keywords to find the desired products. This can be solved by Fine-Grained image classification which focuses on tiny differences between subcategories (low level-class variance) as well as variable appearances such as object posture within the same category. The data used for research was iMaterialist2020 dataset which consists of 45,600 training and 3200 test files of the fashion clothing images. It contains 46 clothing objects and 294 associated minor fine-grained attributes. Three convolutional neural networks are discussed SpineNet143, Mask R-CNN and feature pyramid Network. The paper uses a combined model which is the Mask CNN model consisting of Feature Pyramid Architecture (FPA) and SpineNet. At first the feature are extracted using any CNN model and then SpineNet143 model to create a feature pyramid, followed by Mask on the features, bounding boxes and calculation of the loss. SpineNet143 from google proposed a new meta-architecture for recognition. The ResNet did not work well for simultaneous recognition and localization as they made used of scale decreased model which threw away spatial information. SpineNet make use of Neural Architecture Search (NAS). Mask recurrent convolutional neural network is designed to handle instance segmentation. It is divided into two phases first suggest location where item could exist. Second predicts the class of item and refined the bounding box and creates a mask at the pixel level of the object. Backbone is a deep

neural network in the feature Pyramid Architecture. It is made up of bottom pathway and top-down pathway and the lateral connections. The top down produces a feature pyramid map that is similar in size of the bottom-up pathway. In the research paper SpineNet143 model was used. The first phase used a feature extractor and SpineNet model makes up the second stage. For each recommendation region produced in the first step, the network predicts the bounding boxes, segmentation-mask and object class. To do the testing the data was divided into 80:20 for training the Mask RCNN, SpineNet143 plus FPA models. The batch size of 16 is used with 956 epochs. Mean Average precision of 71.9 and accuracy of 90 with SpineNet143 plus FPA vs 71.4 under SpineNet96 plus FPA was achieved.

2.4 Research Gap

Although there have been plenty of research around the reverse image retrieval area including the application in fashion, garment, ecommerce, medical, Satellite images, crime etc. There are several important gaps remain specifically in the context of fashion oriented reverse image retrieval.

1. Lack of systematic comparison across models' types for fashion datasets. There are study which have individually applied some of the methods including VGG16, ResNet50, EfficientNet, InceptionV3 and some custom models but there is gap where study compare different pre-trained, finetuned and build from scratch custom models with different pre-processing techniques and various similarity measure and then are comparison gives the best model which is best suited for image retrieval for the fashion industry dataset.
2. Insufficient evaluation of pre-processing techniques. Although there has been basic mention of using the pre-processing techniques but there are very few studies where the results from various techniques re compared. Along with this very few study use multiple pre-processing and transformation techniques together on the dataset including the similarity image resizing, image padding, normalization, image crop, image slip, image rotation etc and then study the results.
3. Although the study mention the use of similarity measure. Very few study compares the results of similarity measures and find if there is any material different in the results. There are very limited studies which studies this along with various pre-processing methods and multiple models.

4. There have been studies across classification and recommendation but very few studies measure the retrieval performance across recall, precision, MAP, Top k all together on the dataset. This gap remains in retrieval centric evaluation.
5. And there is lack of study which considers all together. Some studies focus on CNN models, other on pre-processing techniques, similarity measure, may be clustering but there are very few studies which sees the holistic view and bring all together and test the results.
6. The current study applies various techniques on different dataset. There is very less study done which uses all the techniques for the fashion industry dataset problem.

2.5 Summary

In the section, it is understood that various result has been conducted around this area including in the areas of fashion industry where many models has been tried and recommendation systems has been developed. Along with this it has been applied to cross border trade to identify the fake products. It has also found its application to the garment industry. The initial disease diagnosis can be performed using the method of reverse image retrieval. The satellite images can be used to track the location. Various methods have been applied to achieve this task including the text to text, text to image and now image to image. In order to extract the features from the image methods including colour histogram, SURF, SIFT or ORB can be used or the latest method including various CNN model including MobileNet, AlexNet, GoogleNet, ResNet models or the customer models can be used. These models can be used along with the hybrid and modern techniques includes pre-processing, vector database management, fine grained method and various clustering methods.

- **CHAPTER 3: RESEARCH METHODOLOGY**

3.1 Introduction

In order to complete the task of finding the model for getting matching images for an image input, there are many methods available to complete this task. The first step involved in any data modelling exercise is cleaning the data. In this research, the different steps involved in data preparation are data selection, data pre-processing and data transformation. First the data is selected for modelling which is the fashion industry dataset which has data of more than 44.4 k images across different items. This image dataset is pre-processed using techniques which include image padding and image resizing. Based on the selected model, the image data is transformed wherein images are augmented for contrastive learning. After this, various techniques are used to visualize the data which include Exploratory Data Analysis and descriptive analysis and the data is interpreted. Once the data understand is made and the data is ready to model, the data is run through the various selected models. There are many models available in the market including Single Image Models, Region Based Models, One Shot object Models and One-shot Learning Models. After understanding all the models, selection is made for the models which are implemented in this study. After selection of the models, the feature database is created using the models. And finally, the models are used for feature extraction and mapping and techniques including Euclidean distance, Cosine Similarity and Manhattan distance are used to find the matching images. In order to implement and test the models a custom website is build. The process is shown in figure 3.1.1

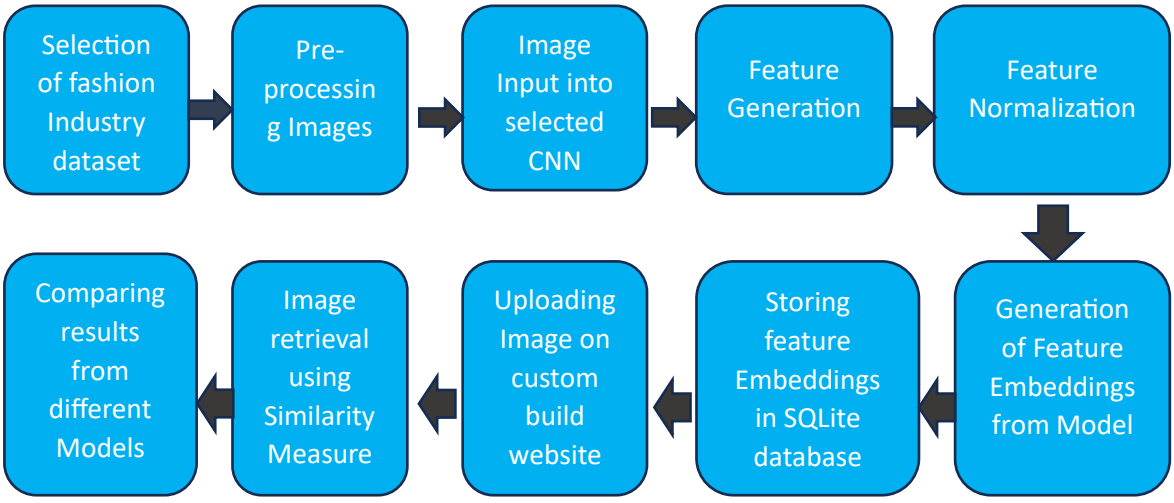


Figure 3.1.1 Process of Reverse Image Search for fashion Industry Dataset

3.2 Data Selection

Under this research study, fashion industry dataset is being used to find the model which gives the best matching recommendations either from the pre-trained or custom model and different loss functions.

The fashion industry dataset is taken from Kaggle. The data is available open source on Kaggle in both the sizes. The same data is available in two data sizes on the platform; one is in small data size of 572.13 MB and other is in large data size of 15.71 GB. For this study considering the limitations of resources, the study is conducted using the small data size of 572.13 MB. The dataset includes 44.4k images along with a csv file. The images are in .jpg format in a folder. The small dataset has the image resolution of 60*80 as against the large dataset which has the data resolution of 2400*1600. The csv file contains the details of each image which is identified by image id which is equal to the image name. Some basic information about the dataset is already given on the dataset. There are 10 columns in csv file which includes id, gender, masterCategory, subcategory, articleType, baseColour, season, year, usage and productDisplayName. The author of the dataset is shown as Param Aggarwal who is B. Tech from Indian Institute of Information Technology, Allahabad. The data has been released under MIT license.

On checking the further details of the dataset, the dataset includes a total of 143 types of articleType including 'Shirts', 'Jeans', 'Watches', 'Track Pants', 'Tshirts', 'Socks', etc. This is an important category type as it bifurcates the data into types of categories which are supposed to have similar features internally and which can be used to fit the model to do the feature extraction and finally doing the reverse image search.

3.3 Pre-processing

The dataset of images 44.4 k images is in the resolution of 60*80 pixel and all the images have the specific articleType given under the csv file. The pre-processing of the data depends on the model which is being used for feature extraction. If the model can take the image input in the size of 60*80 pixel and give the required results, in that case no pre-processing is being done.

In case the model requires the image in different pixel format, the pre-processing techniques which are being used include the padding method and resizing of images. Under the padding

method, the blank padding of (0,0,0) around the images to convert the image to a desired pixel size. This ensures that the image original pixel information is not lost due to stretching of the image and also accomplish the requirement of the model for the input pixel size.

The other method is to resize the image in order to get the structure of the image which is compatible with the model input. This may result in stretching of the pixels but this is one of the methods which has also been used. The images are first converted to numerical tensor before they are given as input to the model and are then normalized to bring it to the values between 0 to1.

3.4 Transformation

The transformation of the dataset would include the generation of images in order to perform either the contrastive learning or triplet loss method. Under the contrastive learning approach, it works on the pairs of images. Each pair is either a positive pair which include the set of images which are two views of the same image (e.g image x_i and image x_j) and other set is of negative pair which includes two different images. The model is trained in a manner which promotes the positive pairs to be close in the embeddings and negative pairs to be far from each other. The advantage of this approach is the image augmentations can be used to create the positive and negative images.

The other approach is to use triplet loss method under which three images are considered together which includes the main image, positive image which has the similar embedding and negative image which is a completely different image and has very varied embedding. Under this method the Triplet Margin Loss is calculated with the following formula which depicts this calculation:

$$\text{Loss} = \max(0, D(a,p) - D(a,n) + m);$$

a -> anchor

p-> positive

n-> negative

m-> margin

This ensures that the distance with the positive image embeddings is smaller than the distance with the negative image embedding.

3.5 Visualization

The dataset includes 44.4 k images along with a csv file which includes the information each image across 9 categories in different columns. Before the reverse image search is applied to the dataset. The visualization is done for the dataset information in csv file to under the width and depth of the type of image dataset which is available to be applied the models. This ensures that the dataset is sufficient and adequate for the exercise.

The methods which have been used are EDA (Exploratory Data Analysis) and descriptive analytics under which data set format is visualized. Bar graph has been used to visualize the frequency as per different column category, and along with that the grid map has been used in order to analyse two variables at a time. descriptive methods for the variables via which it is found that which of the column variables has how many different numbers of categories.

3.6 Interpretation

From the data and visualization section, it is understood that which Gender has the highest number of products, which masterCategory has the highest number of products in the dataset. What is the relationship between different category in the data. Whether the data is evenly distributed along with season. Which is the most important variable which can be used for fitted the model using classification technique. Whether the data is biased or each category has sufficient number of items for the dataset to fit to the model. Whether the dataset represents the real-life scenario.

3.7 Machine Learning Models

In order to make use of the dataset to test and also train different models for reverse image search, it involves features extraction of all the images and store it in a database. Then the same model is used for feature extraction of the image for which the matching images are required and the features of this image are matching with the features of all images in the database using a similarity measures.

Neural Network Models are used in order to get the matching images recommendations from the Fashion Industry dataset using CNNs as part of reverse image search methodology. The structure of CNN model is given in figure 3.7.1. The property of CNN to extract

features is used for getting the matching images by matching the features of the input image with the features of database of stored images.

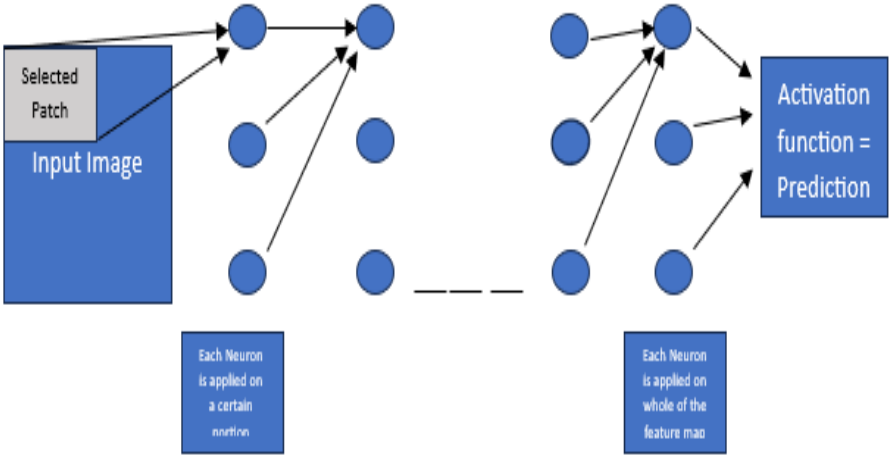


Figure 3.7.1: The structure of CNN Model

The CNN models are used for various real-life applications such as image classification, object detection, face recognition etc. For performing these tasks during the training process the CNN models extract features from the images and train itself using a loss function for the making the most accurate and efficient prediction. The characteristic of feature extraction of CNNs makes them best suited for using in the application of Reverse Image Search. Under this approach the model compares the feature of the queried image with the features of images in the database and presents the images which have the closest features match with the queried image. There are various methods which are used to get the similarity score such as Euclidean distance, Manhattan distance, Cosine similarity and Annoy Indexing.

The challenge in performing the image-based search lies in the size and relevance of data. The bigger the size of the data, the higher is the processing time and cost. Therefore, the selection of model is also based on the time the model takes to present the result. Various techniques are used for indexing the data so that the processing time is reduced. These techniques include hashing, vectorization, mean features so that the size of the data is reduced.

The various CNN models available are Single Image Models including AlexNet, VGGNet, GoogleNet, ResidualNet, Region-Based models such RCNN, Fast RCNN and Faster RCNN and One-Shot models such as Yolo and SSD etc. The pre-trained versions for these models are available on various other available datasets. The same are discussed below:

a) **Single Image Models**: The pre-trained CNN single image models include MobileNet V2 or V3, EfficientNet, InceptionV3, VGG16/VGG19, ResNet50. The user of each of these differ. MobileNet model is used when light features extraction is required. Inception V3 is used when strong features extraction is required and ResNet50 has use case of Universal feature extraction. These models are available with pre-training done on ImageNet dataset. These models are designed for the image classification roles. But as discussed, their features extraction phenomenon allows them to be used for reverse images search. The model runs on each image and extract features including the texture, shape, object, colour, edges of each image and store the final feature layer as embeddings for reverse image search. This makes it easy to implement for the task as the models just needs to be called in python file and can directly be applied on the dataset to get the feature database of the images and the same model can be used for the uploaded image to get the image features. Then the image features are matching with the feature database using one of the similarity function Cosine, Euclidean or Manhattan. The architecture of one of these models i.e. VGG 16 is given below in figure 3.7.2. The drawback of these models is just that they are not actually trained for similarity task and rather trained for classification tasks. For similarity tasks one shots learning models are properly designed to handle such tasks which are discussed after the region based and one-shot object models below.

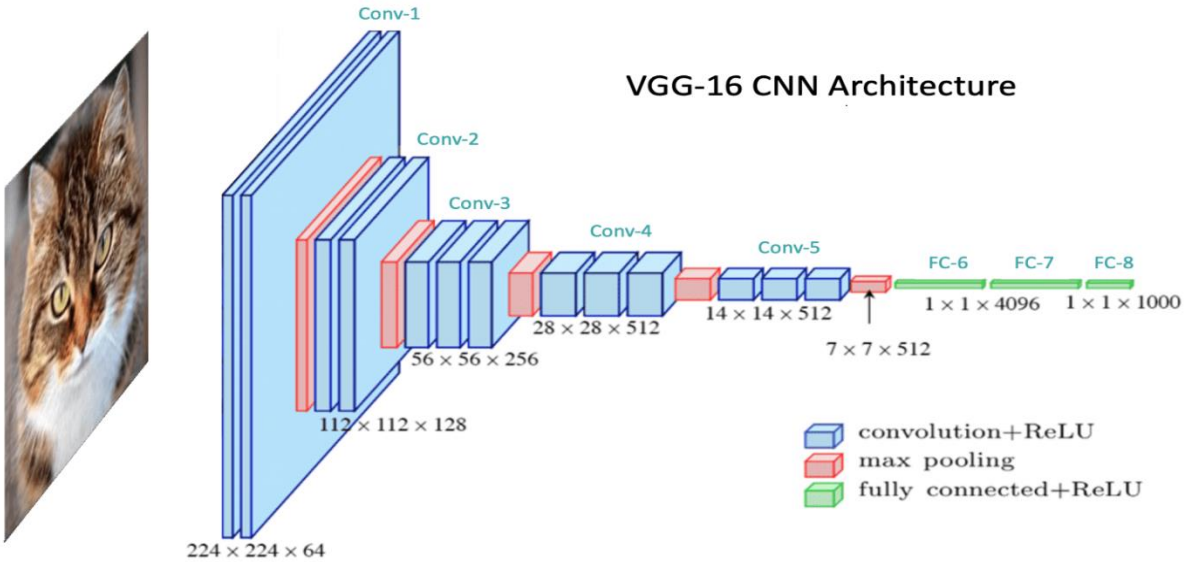


Figure 3.7.2: VGG -16 CNN Architecture

b) **Region Based Models**: Apart from the single image models the other set of models are region-based models. The main objective of the region-based models is to locate

different objects in the image. The model does this task by first detecting different objects in the image and then allocating bounding boxes to each object. The region-based models can be useful in the context that the embedding from different region can be used to comparison. The fashion images can be cropped if there are multiple objects present in the single image. The method can be used to focus on particular object in the images and reduce the background noise. The region-based models include including RCNN, Fast RCNN and Faster RCNN. But the drawback is these are not specifically trained for image similarity task. The image in figure 3.7.3 depicts the output of the region-based model which tell the different objects in the image and also plots the bounding boxes around these objects.

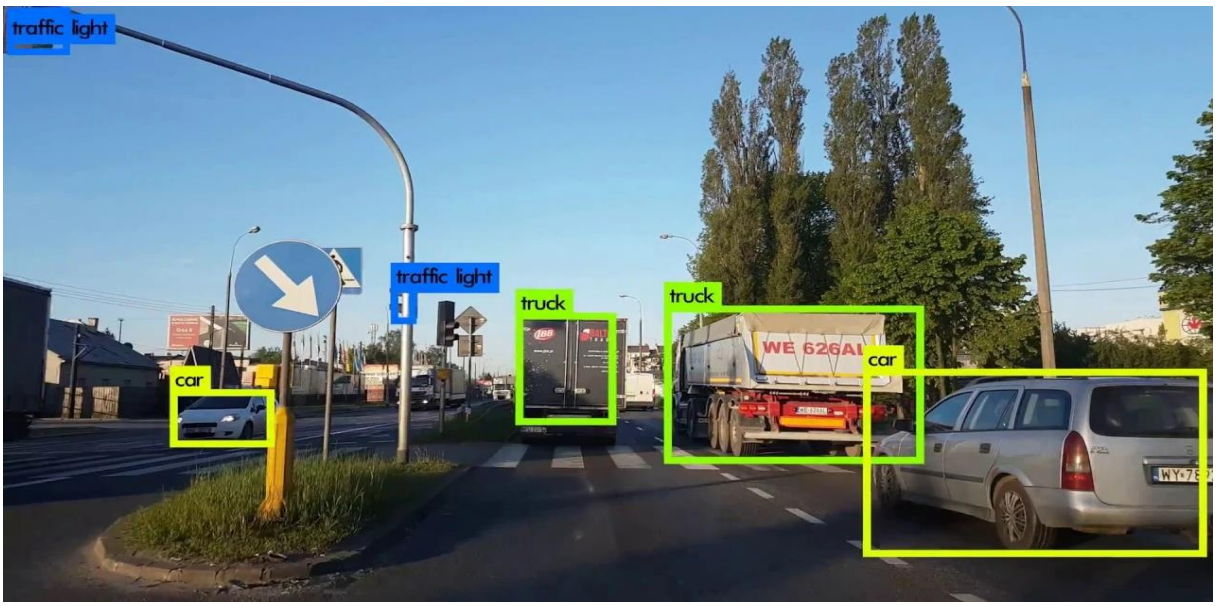


Figure: 3.7.3: Region-based model object detection with bounding boxes

(Image source: https://cdn.analyticsvidhya.com/wp-content/uploads/2018/06/maxresdefault.jpg)

The drawback of the region-based models is that as it involves two step process to first identify the image and then allocate the bounding boxes, it takes more time. The more efficient models are one shot object detection models which does these two tasks together.

c) **One Shot Object Detection Models**: The next type of models are one shot object detection models; these models are faster version of region-based models. They do the object detection and bounding boxes tasks in one step resulting in faster processing. The one-shot object detection models include Yolo and SSD (Single Shot Detector).

d) **One Shot Learning Models**: The final types of models are one shot learning models which are specifically designed in a way which trains the model to do similarity tasks. In these models rather than training the models against the classification, the model compares the embeddings to two similar images and trains the model to reduce the embedding loss between the images. The model generates the embeddings for one image and generates the embedding for the other image and finally compare the embeddings of these two images and work on reducing the contrastive or triplet loss to train the model to get the matching images. The models which do this task are Triplet Networks and Contrastive learning-based models. The CNN backbone model can be from the single image models including the VGG, MobileNet, EfficientNet, ResNet. These single image models are used to generate the embeddings for each image. Using this method the pre-trained model can either be refined on the fashion industry dataset or a new base model can be trained from ground using this method and the results of that model can be compared with the results of pre-trained model. The advantage of this model is that it trains directly on the image similarity rather than the classification method and therefore, is expected to give better reverse image results. The drawback of this method is it requires to first generate the triplet data for training the model and takes too much time for the model to train which may be 4 to 6 times in comparison to the classification models.

Selected Models: In this research at first the pre-trained modes including MobileNetV2 and ResNet50 are applied on the dataset and their results are compared with each other. After that although, there are some self-trained models available under this category but for this study we are limiting this method to finetuning the one-shot learning method on the pre-trained ResNet50 Model. In order to finetune the model on articleType, the overall data is divided into train and test dataset in the ratio of 80:20 for fitting the model. After that the baseline ResNet50 model have been trained the using one shot learning method on the fashion industry database. Then the results from both the models are considered.

Reason for choosing MobileNetV2 is that it is a lightweight model which is fast and have low computational cost. This is used as the first model as to test the working of the whole structure plus the various similarity measure before moving on to heavy models. ResNet50 on the other hand is a deeper and more complex model which allows to check the what is the most images which are being extracted from the dataset using a pre-trained model.

Since, ResNet50 is model used from the pre-trained models, it is best to test the finetuned and the base model which is fully trained using ResNet50 only so that the results can be compared with each other and it can be decided whether pre-trained model is working better or the custom model. The check is also done for the scenario where the feature database is from the ResNet50 Model and the features of the uploaded image extracted using the MobileNetV2 model.

3.8 Creation of Feature Database

In order to test a model accuracy for the suitability of an applicability on fashion industry dataset, at first the features are extracted for all the images in the dataset using the selected model and stored in a SQL database. Each image is loaded using the image path then the selected model is applied on the image and feature are extracted. While extracting the features the images are resized and normalized. This process is applied to all the images in the folder and the extracted features are stored in a database. The database have two columns one for image name and other containing image features array.

Figure 3.8.1: Process for creating the feature database for database of all images

This process is first tried by storing the images on the computer and running the model on the spyder app. In case of space issue, the data is uploaded on google drive and the model is run on google collab with mounting of google drive to access the images and the final database is downloaded on the computer.

This process is applied separately for all the models which are being tested. In other scenarios where the features extracted by one model are compatible with other models, in such cases the results are tested with cross models as well where the features stored are from one model but the model used to extract feature of the input image is from different model.

For the feature extraction there are two options, either to use the pre-trained model such as MobileNetV2 and extract the features and use the same model for input image and the matching the features. The other approach is to first fit the model to the dataset. While fitting the model, the images are uploaded in batches to the model. The complete batch first goes

through the forward pass and then loss is computed and then weights are updated through the back propagation. This process happens until all the batches images are processed through the model. Then there is an option to how many times the complete set of images would go through model, for that the parameter used in modelling is number of epochs. Once the model it fitted to the image dataset. The next steps are same as are used in the pre-trained model.

3.9 Application of different Models and Similarity Measures

After the features are extracted for the image database either using pre-trained or custom build models, the features are stored in the SQLite database and this database is saved in the working directory. Now on upload of the image, the features are extracted using one of the models and then the similarity check is made using similarity measures. Under this study different models are tried and the results are compared in between them.

In order to find the matching images, the similarity measures such as Euclidean distance, Manhattan distance and cosine similarity have been used and results from each one of them are compared in order to find the one which provides the best matching results.

Euclidean Distance: It is the most fundamental measure to quantify distance between two points in an n-dimensional space.

The formula for Euclidean Distance between vector A and B is expressed as:

$$\text{Euclidean Distance} = \sqrt{\sum_1^n (Ai - Bi)\text{^}2}$$

Ai and Bi, denotes the corresponding features vectors of the uploaded image and the database image, respectively. The smaller the distance, the more similar the two images are. The distance is squared and summed to get the total squared deviation.

The value of Euclidean distance ranges from 0 to infinity.

The formula for cosine similarity between two vector A and B is expressed as:

$$\text{Cosine Similarity} = (A \times B)/(||A|| \times ||B||)$$

A and B denote the two vectors and ||A|| and ||B|| represent their respective Euclidean magnitudes. The similarity value for this measure ranges from -1 to 1, where 1 stands for the perfect match. And -1 stands for completely different. The magnitude of each vector is calculated as the square root of the sum of squared feature values.

Manhattan Distance:

The formula for the Manhattan Distance between two vector A and B is expressed as:

$$\text{Manhattan Distance} = \sum_{1}^{n} |Ai - Bi|$$

Ai and Bi are the corresponding feature vectors, n is the number of features. This sums the absolute distance between the corresponding vector elements. The smaller the distance, the greater is the similarity between the images. It ranges between 0 to infinity.

The effect of using different similarity measure is tested with the MobileNetV2 model and based on the results the selection is made for the other models.

In order to test the model, a custom build website is used which have the feature of uploading the image and presenting the matching images on the webpages after processing. This website would be hosted on the localhost server on the laptop using XAMPP Control Panel.

The results of different models both pre-trained and newly trained along with impact of data pre-processing and similarity measures on the final results are summarized and the best combination are found.

The flow chart of the working is as follows:

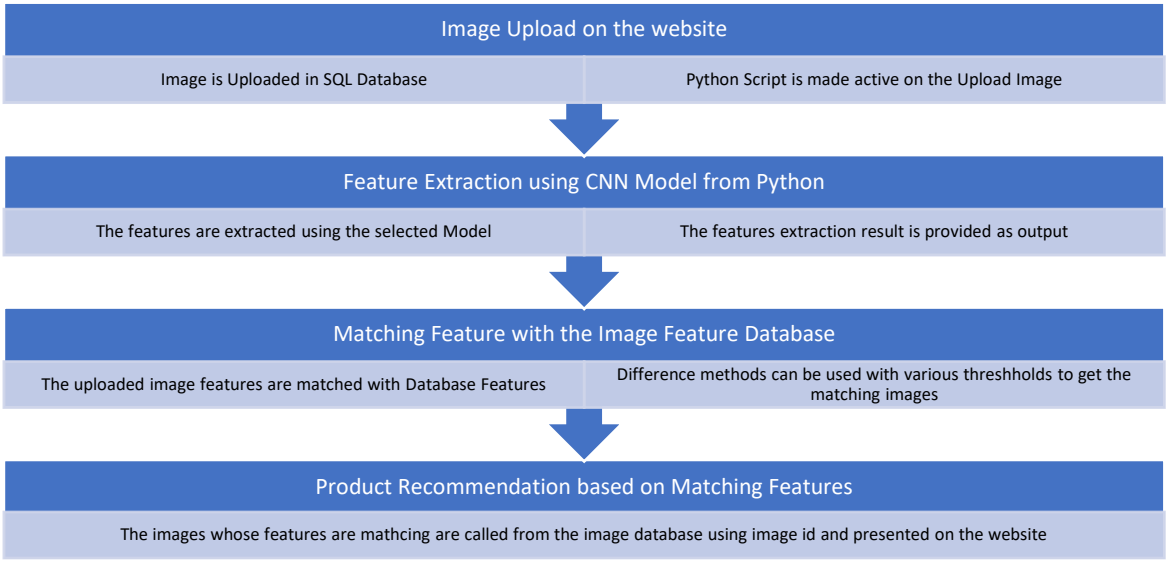


Figure 3.9.1: Process for getting matching images on e-commerce website

The accuracy of the models is tested based on the number of maximum matching images fetching by the model (either precision and/or recall method). Where in precision is

calculated using the relevant images produced by total images produced and recall is calculated using the relevant images produced by the total relevant images.

3.10 Creation of e-commerce website for testing

A website is created using php language to test the models for the reverse image search. The website has the option to upload an image using the upload button. The website after uploading the image stores the image in its database and then recalls the image for pre-processing and feature extraction. Once the image feature extraction is done, the features of this image are compared with already uploaded features table for the whole dataset on the website database. The similarity measurement is done by the specified similarity measure on the website for that particular model. This all is one in the backend by the website. Based on the similarity measurement, the website presents the results of the matching images. The figure 3.10.1 shows the custom website interface which has been created for the testing task.

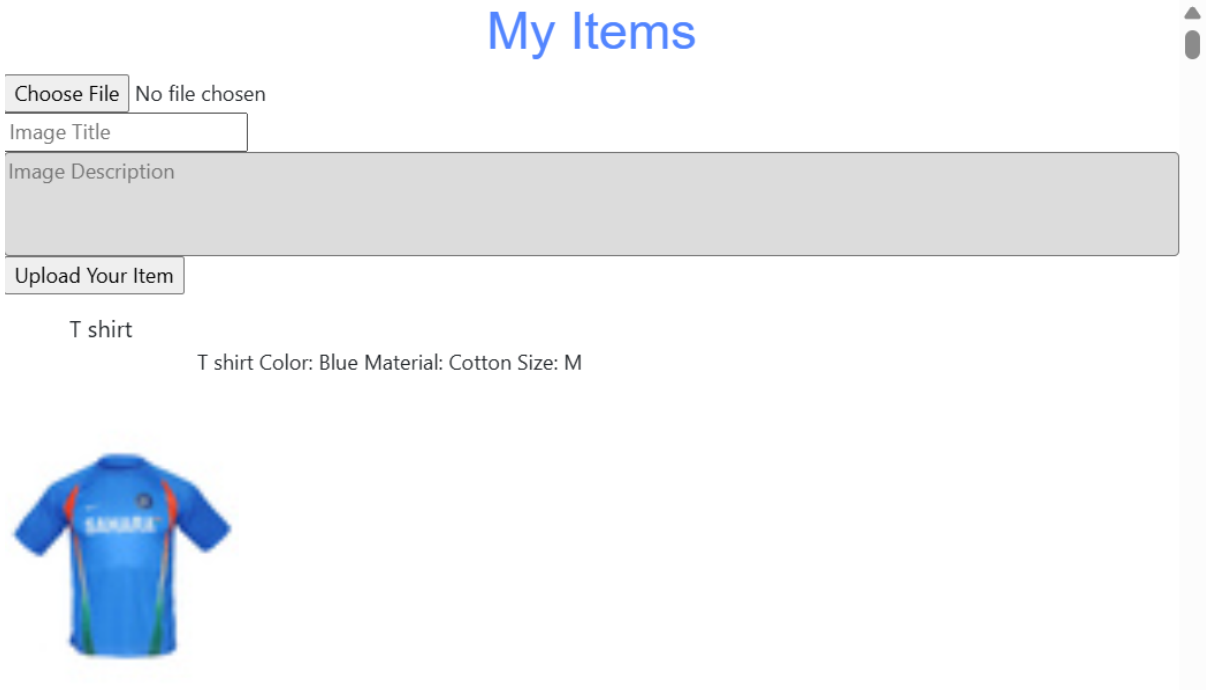


Figure 3.10.1: Custom Website Interface for uploading the image

After the uploaded image is processed in the backend the output of the matching images is presented by the website as given in figure 3.10.2

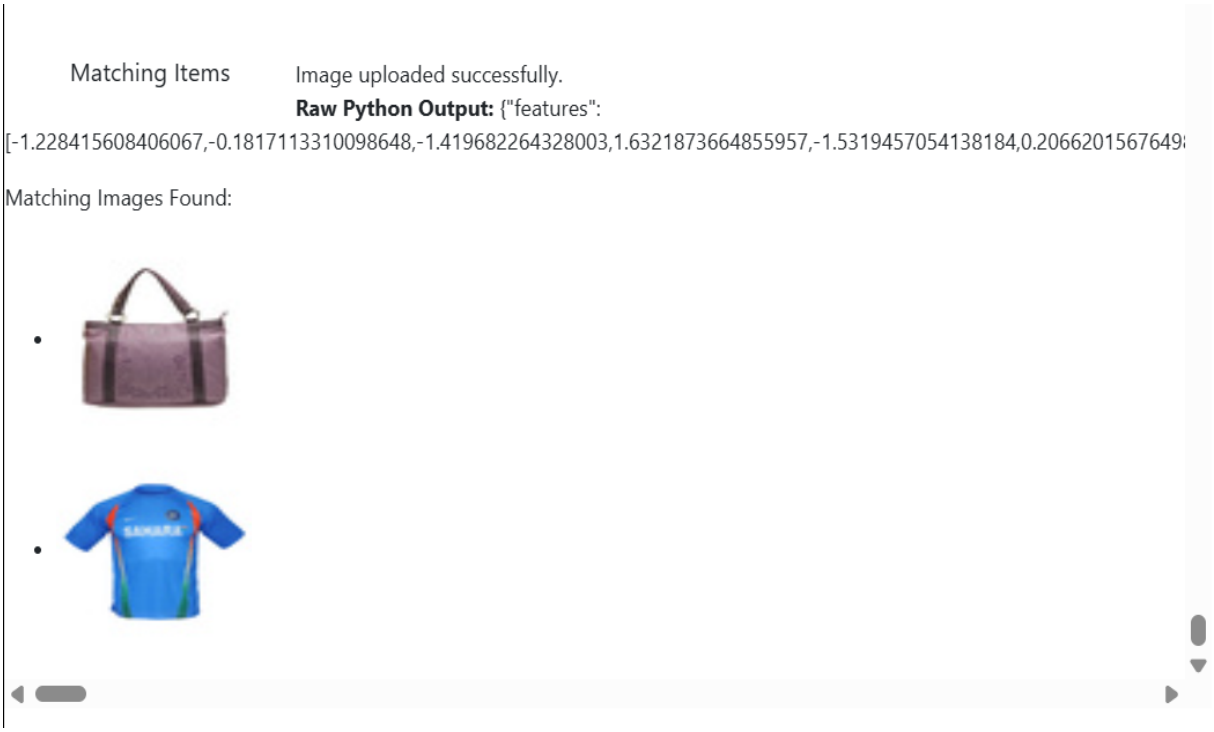


Figure 3.10.2: Custom website output for the matching images

3.11 Summary

In this chapter, at first the details of the selected data were discussed such that it has more than 44.4k images along with a csv file. There are 9 parameter details available for each image in the csv file which include important details related to masterCategory and articleType. The data pre-processing techniques which are to be used in this study includes padding and image resizing. The image transformation technique which is to be used in case of models requiring contrastive learning. Under the image transformation images are augmented from the original image to create positive and negatives of the same image. The data visualization is done using techniques to be used are EDA which gives the understanding of the data. The models which are selected for modelling including MobileNetV2 Model and ResNet50 Model. The Euclidean distance, Cosine similarity and Manhattan distance, similarity measures are to be used in the next section. The features database is created using each model and the same model is used for the uploaded image for feature extraction and similarity measures are used to find the matching images. The models are tested on the custom build website which has the functionality to upload and submit the images and the matching images are presented on the website as output after completion of processing at the backend.

- **CHAPTER 4 IMPLEMENTATION AND ANALYSIS**

4.1 Introduction

In previous chapter, after the data selection, different methodologies have been discussed to pre-process and transform the data, visualize the dataset and finally applying different models to the dataset to get the matching images. In this chapter, those methods are applied. The dataset exhibits variety of products across different category and is sufficiently split across different blocks which makes it fit to use for this study. The data is presented with figures across different variables which gives the highlight of the data. The data required pre-processing as the input required by ResNet50 model is in 224*224 pixel whereas the data is in 60*80 pixel. This is achieved either by resizing or by padding. The models which are used for testing include MobileNetV2 with all three similarity measures. The data was resized to 128*128 pixel for applying this method. The results across the similarity measures are compared. The ResNet50 Model is applied with and without padding and also with finetuning with classification and contrastive learning. The ResNet50 model showed some improvement in results over the MobileNet50 model, whereas finetuning instead of improving, decreased the efficiently of the model. Finally, a custom build model is tested on the dataset. Each method is described in detail along with its output. A sample image is taken from the dataset which used in across the models so that the results can be compared with each other and model can be identified which is giving the best result.

4.2 Exploratory Data Analysis and Descriptive Analysis

The dataset includes more than 44.4k images and a csv file which includes fashion dataset for products across various genders, master categories, sub categories, article type and season. In order to understand the data, the visual description of the data has been done in the upcoming tables and figures to have an insight into the composition of the data.

The table 4.2.1 shows the first 10 rows from the data csv files which has the data information:

Table 4.2.1: Top 10 rows of the data csv file:

ex	Id	gender	masterCategory	subCategory	articleType	baseColour	Season	Year	usage	productDisplayName
0	15970	Men	Apparel	Topwear	Shirts	Navy Blue	Fall	2011.0	Casual	Turtle Check Men Navy Blue Shirt

1	39386	Men	Apparel	Bottomwear	Jeans	Blue	Summer	2012.0	Casual	Peter England Men Party Blue Jeans
2	59263	Women	Accessories	Watches	Watches	Silver	Winter	2016.0	Casual	Titan Women Silver Watch
3	21379	Men	Apparel	Bottomwear	Track Pants	Black	Fall	2011.0	Casual	Manchester United Men Solid Black Track Pants
4	53759	Men	Apparel	Topwear	Tshirts	Grey	Summer	2012.0	Casual	Puma Men Grey T-shirt
5	1855	Men	Apparel	Topwear	Tshirts	Grey	Summer	2011.0	Casual	Inkfruit Mens Chain Reaction T-shirt
6	30805	Men	Apparel	Topwear	Shirts	Green	Summer	2012.0	Ethnic	Fabindia Men Striped Green Shirt
7	26960	Women	Apparel	Topwear	Shirts	Purple	Summer	2012.0	Casual	Jealous 21 Women Purple Shirt
8	29114	Men	Accessories	Socks	Socks	Navy Blue	Summer	2012.0	Casual	Puma Men Pack of 3 Socks

The datafile contains 10 variables. The first variable is id, which is nothing but the Image name in the image folder. In cases any of the variable is to be used for modelling, in that case the id can be used as the reference point for that image and respective variables can be used for the modelling exercise.

The second variable was gender under which there are 5 categories with the maximum number of products for the Men's Category, followed by Women Category, then Unisex, Boys and lastly Girls. This is consistent with the real life scenario as the more products are bought by Men and Women for themselves on online platforms. This Gender wise details are shown in figure 4.2.1

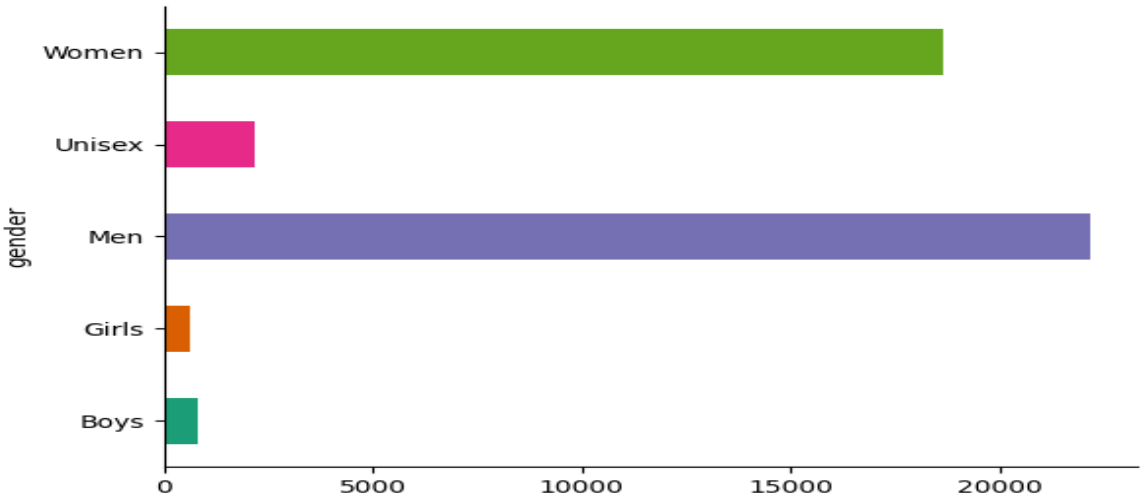


Figure 4.2.1: Gender wise data frequency

The masterCategory had 7 categories including Apparel which had the highest number of datapoints, followed by Accessories, Footwear and Personal Care. There are very few items under the MasterCategory of Free items, Sporting goods and Home items. Again, most of the shopping done is around the Apparels and accessories and then footwears. This shows the data very well represents the real-life scenario of an e-commerce shopping site. The masterCategory wise depiction of data is shown in figure 4.2.2.

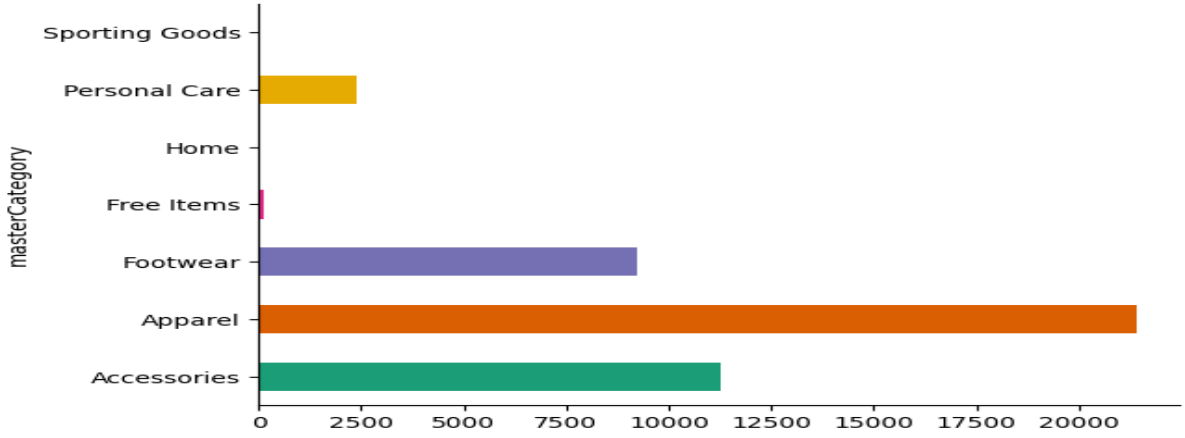


Figure 4.2.2 masterCategory wise data frequency

The third visualization was done as per subcategory. There were 45 subcategories present in the data. The highest number of products are from the topwear subcategory. This was followed by shoes, bottomwear, bags and watches. Again, this data is consistent with the demand in the market for various products and according to that the stocks maintained by an e-commerce firm. This is visible in figure 4.2.3.

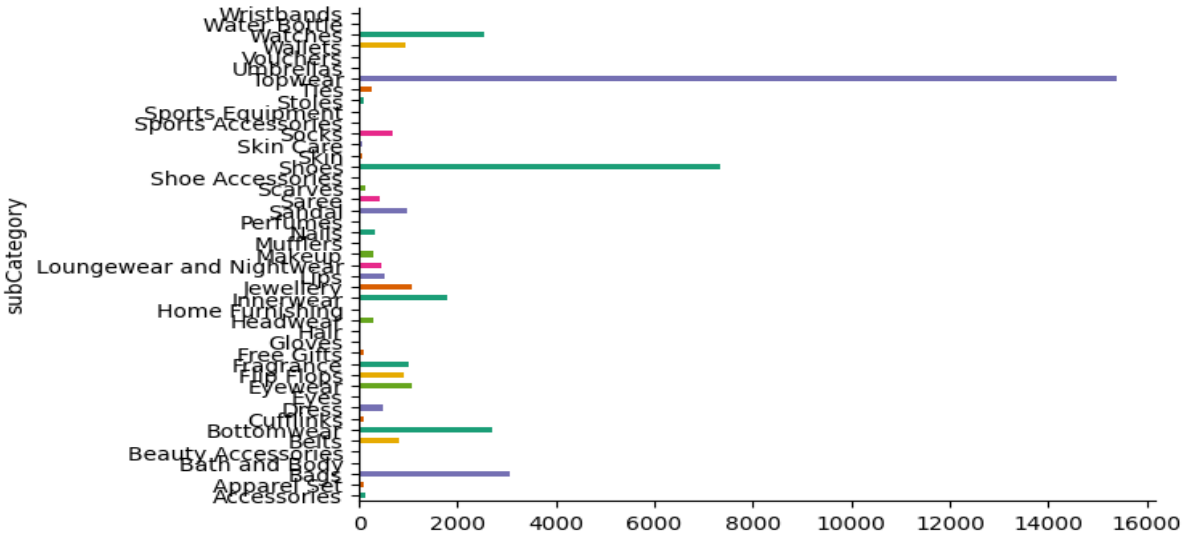


Figure 4.2.3 subCategory wise Data distribution

The fourth category in the data is for the articleType. This variable had 143 categories in it. The visualization was done for the top 10 articleTypes. Tshirts were highest in number followed by shirts, casual shoes, watches, sports shoes, kurtas, tops, handbags, heels and sunglasses. This is visible in figure 4.2.4.

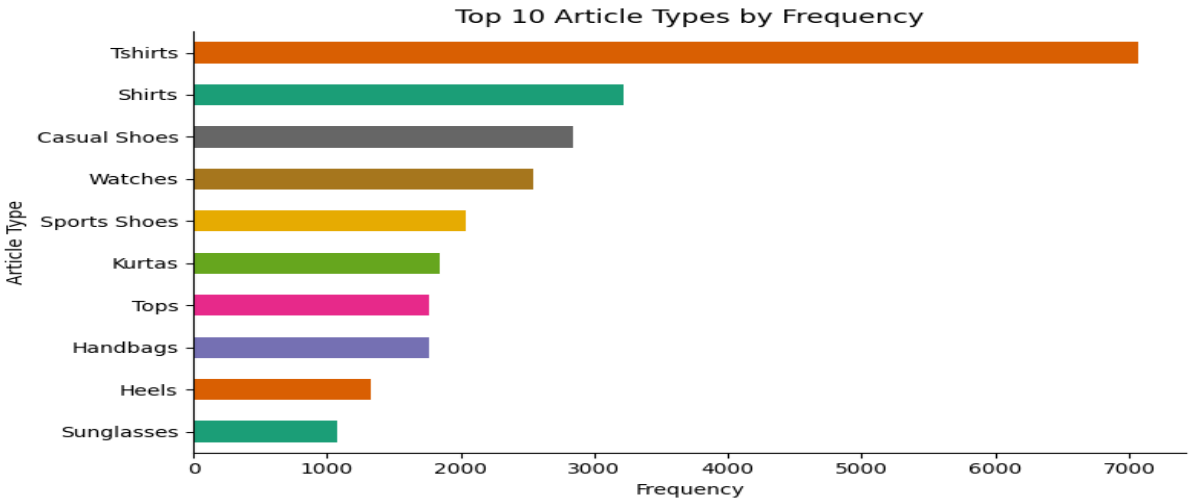


Figure 4.2.4 Top 10 articleType wise Data distribution

The fifth category in the data was basecolour which defines the colour of the product. Again, this category had 47 categories. The top colour with most number was black, followed by white, blue, brown, grey, red, pink, navy blue, silver and yellow. This is visible in figure 4.2.5.

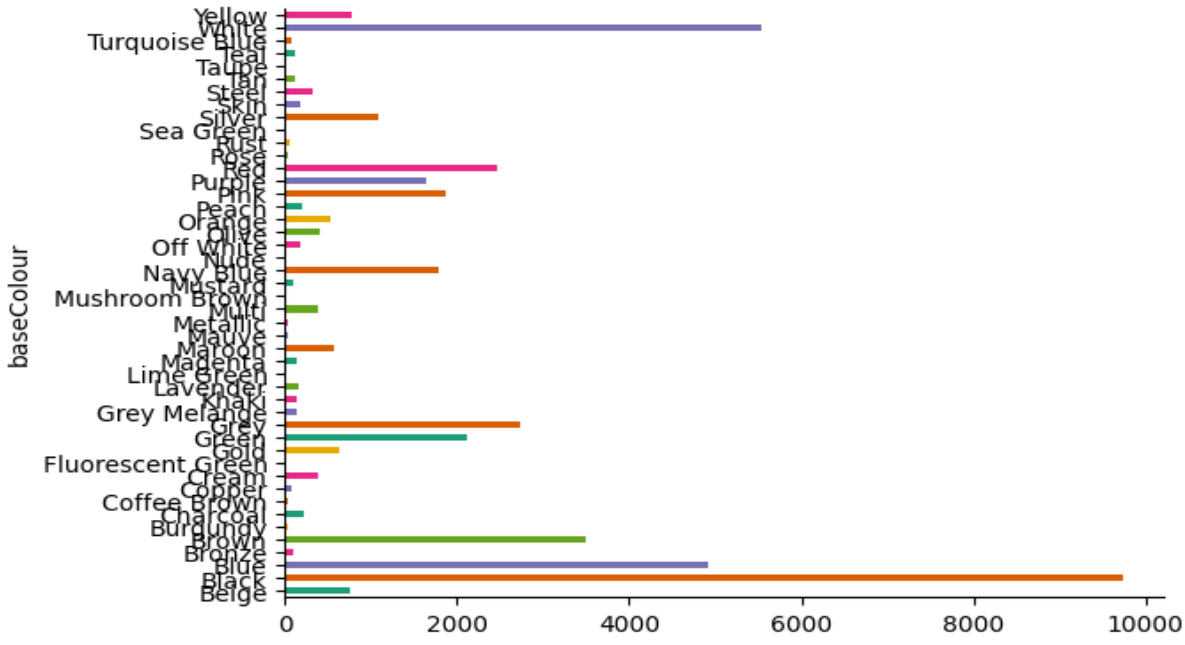


Figure 4.2.5 baseColour wise Data distribution

The next metadata available in the CSV file for the images is season. This data point tells for which season the item is suitable. The highest number of items are for the summer season, followed by fall, winter and spring. This seems reasonable with the market trend. This is visible in figure 4.2.6.

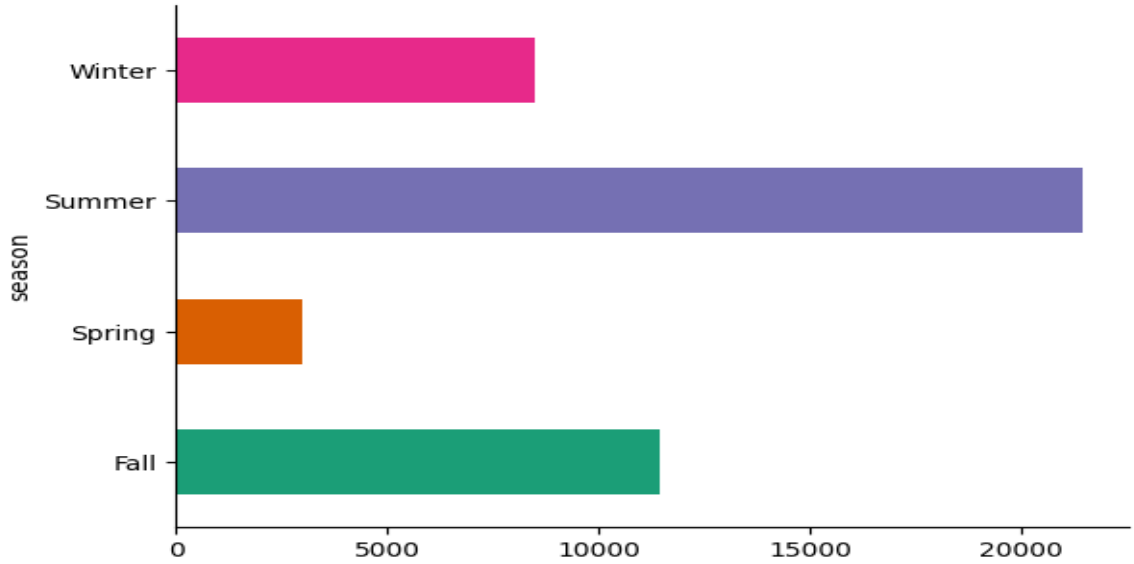


Figure 4.2.6 season wise data frequency

The next meta data tells the year to which the product belongs. The products are from the year 2008 to 2019. Most of the products are from 2011 and 2012 year. This is visible in figure 4.2.7.

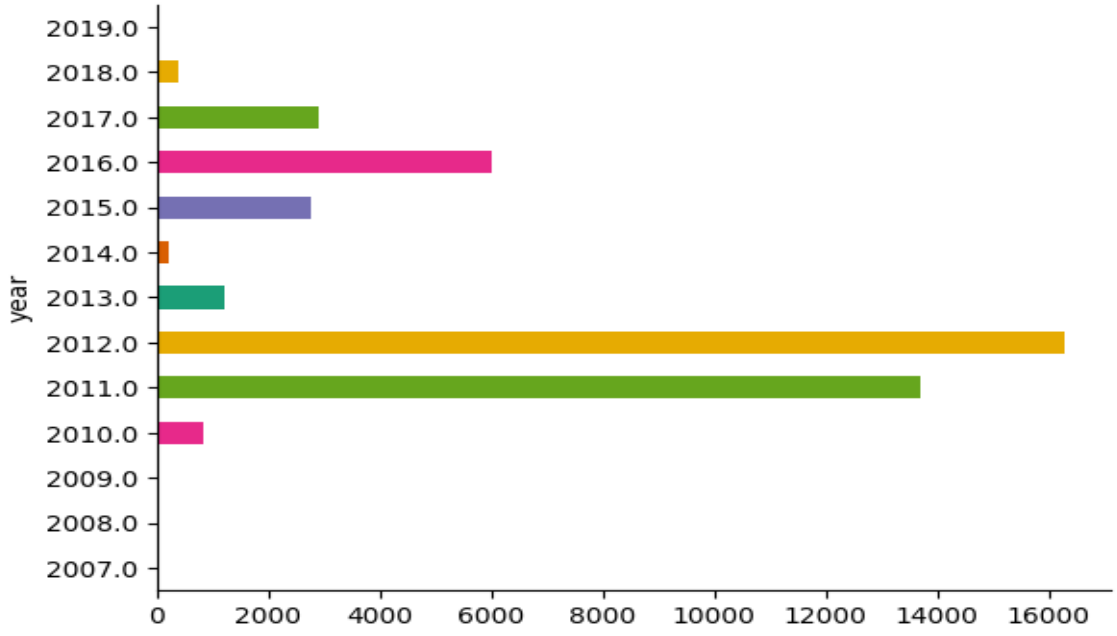


Figure 4.2.7 Year wise Data distribution

The second last data point is product usage which tells the purpose for which is product is used. The most number of products are associated with casual use which is followed by sports ethnic and formal. This is visible in figure 4.2.8.

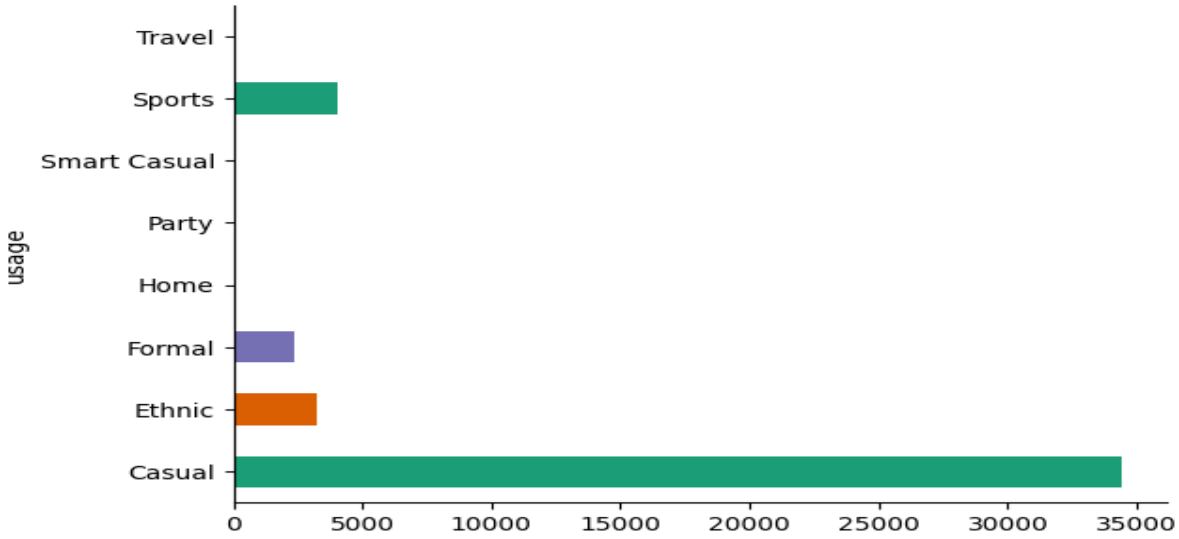


Figure 4.2.8 Usage wise Data distribution

The last metadata has the product displayname.

A cross-category visualization was done for Gender and masterCategory which showed that the most of the images are for products which are for Apparel for Men, followed by Apparel for women, accessories and footwear for man and women. This is visible in figure 4.2.9.

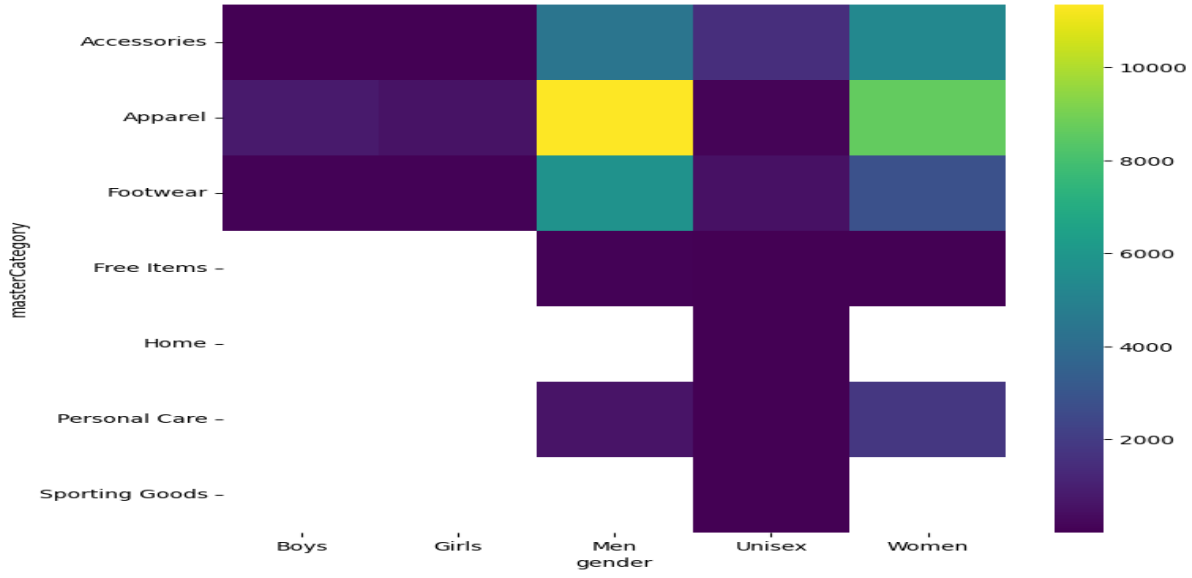


Figure: 4.2.9 Genderwise masterCategory

The other cross category visualization was done for masterCategory and season which tells that the most products are from apparel for summer season. This is visible in figure 4.2.10.

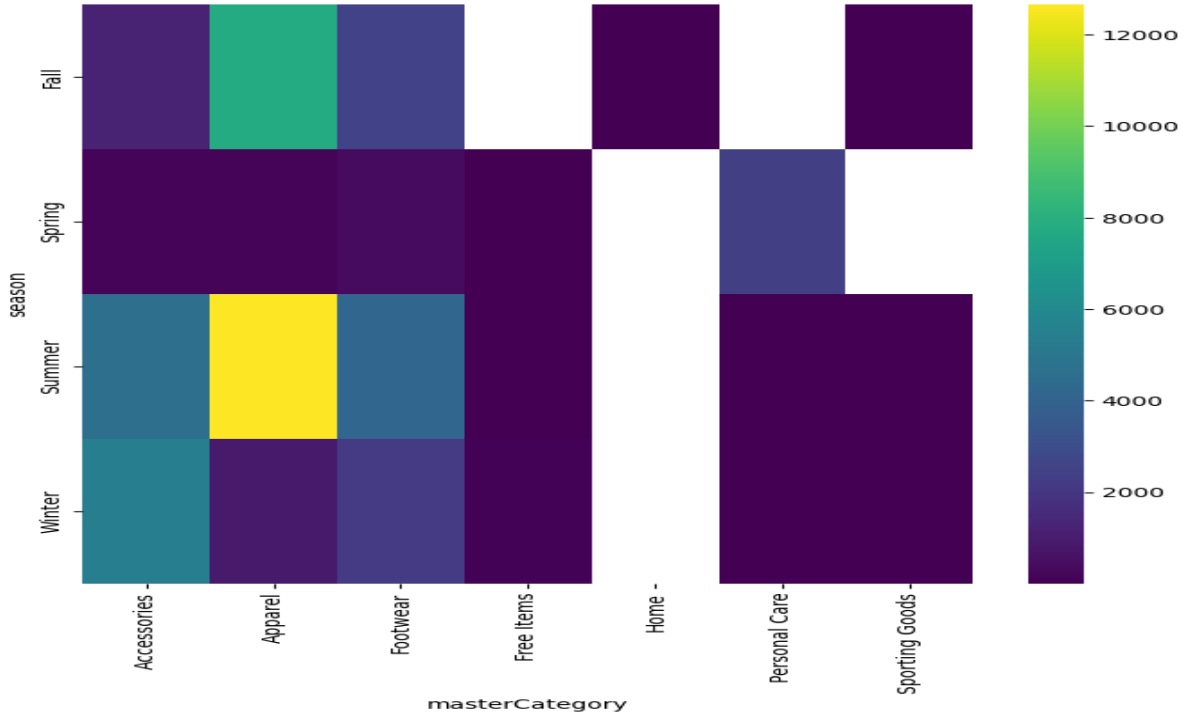


Figure 4.2.10: Season wise frequency in each masterCategory

From the data and visualization, it is understood that the highest number of products are marked to Gender 'Men', followed by 'Women'. MasterCategory 'Apparel' has the highest number of products followed by 'Accessories' and 'Footwear'. The data has products across all season with highest number of products are associated with the summer season. We also know that Men 'Apparels' has the highest number of products followed by Women's 'Apparels'. Apparels summer covers majority of the products. The data shows that it has sufficient number of data across each category and it goes with how is the demand of products in the real world and the data would be enough to test and fit a model to the data and finally find the model which provides the best reverse image search output.

4.3 Predictive Analysis and Model Testing

It is now understood that the dataset has sufficient number of images across various categories which can be used to build a model for the reverse image search for the fashion industry. The models which have been shortlisted include the following:'

a) Pre-trained MobileNetV2 model trained on ImageNet dataset with Euclidean distance

b) Pre-trained MobileNetV2 model trained on ImageNet dataset with Cosine Similarity
c) Pre-trained MobileNetV2 model trained on ImageNet dataset with Manhattan distance
d) Pre-trained ResNet50 model trained on ImageNet dataset for feature database along with MobileNetV2 model for image feature extraction with Manhattan distance
e) Pre-trained ResNet50 model trained on ImageNet dataset with Manhattan distance
f) Pre-trained ResNet50 model trained on ImageNet dataset with image padding with Manhattan distance
g) Fine-tuned (on articleType classification) pre-trained ResNet50 model trained on ImageNet dataset with cosine similarity
h) Contrastive Learning Fine-tuned pre-trained ResNet50 model trained on ImageNet dataset with Cosine Similarity
i) Custom Build Contrastive Learning ResNet50 model trained on fashion industry dataset with Cosine Similarity

a) Pre-trained MobileNetV2 model trained on ImageNet dataset with Euclidean distance

The first model which has been applied on the dataset is the pre-trained MobileNetV2 model. The model is available with its weights trained on the ImageNet dataset. Each image from the dataset is first pre-processed to ensure its compatibility with the model. The dataset has image in 60*80 size whereas the MobileNetV2 models accepts inputs in square shape and above minimum of 32 pixels. At first the images have been resized first to size of 128*128 and are converted into numerical arrays. The arrays are normalized using the preprocess_input() function from keras MobileNetV2 module. After that the preprocessed images are passed through the pre-trained MobileNetV2 model with top classification removed (include_top =False). Global average pooling was applied to the final convolution feature map (pooling = ‘avg’). The feature vectors extracted the high level information about the textures, shapes, colour patterns. Following the extraction of features vector, they were flattened and are stored in the local SQLite database. A database table named image_features was created with two columns named ‘image_name’ for storing the image filename and ‘features’ for storing the comma separated numerical feature vector.

After creation of the feature database, the same are uploaded on the database of the custom build website for testing the model. For testing the model for the suitability of

reverse image search, one image is uploaded on the website which has the build with integrated design to first resize the image to 128*128 pixels size which corresponds to the input dimension required by the MobileNetV2 model and also is consistent with the image features database. After resizing the uploaded image, it is converted into numerical tensor using Tensorflow imge_to_array() function. Then it is reshaped with np.expand_dims() and pixel values are normalized and adjusted using the mobilenetv2.preprocess_input() function. This makes the images feature vector ready for comparison with the features available for all the images in the image database.

In this model, the similarity measure which is used is Euclidean Similarity. When the features are roughly in the range of 0-1 the Euclidean distance is generally in the range of 10-20.

The testing is done starting with the distance of <18 and the image which is used for testing is given as figure 4.3.1. The site presented the results which are depicted in the figure 4.3.2 (a). As the output resulted in two matching images, the similarity distance is relaxed to 22. The output after changing the similarity distance to 22 is given in figure 4.3.2(b). Since with the distance of <22, the output included so many outputs, the distance was reduced to 20. The results with the distance of 20 are shown in figure 4.3.2 (c). Since, the output had one extra output which was not matching. The distance was reduced to 19 and this was the final output which was received from this model. The same is shown in figure 4.3.2 (d).

Figure 4.3.1: The sample image
for uploading for testing

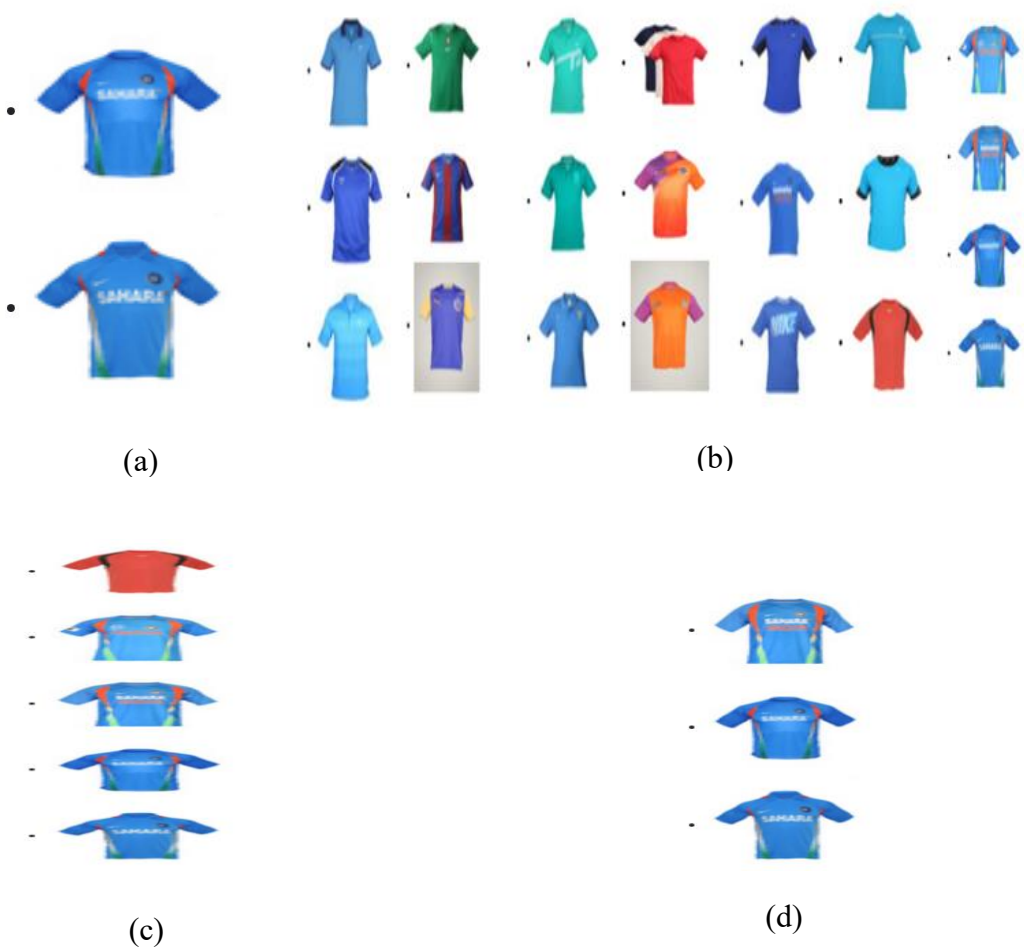


Figure 4.3.2: Results with Pre-trained MobileNetV2 model and Euclidean Distance

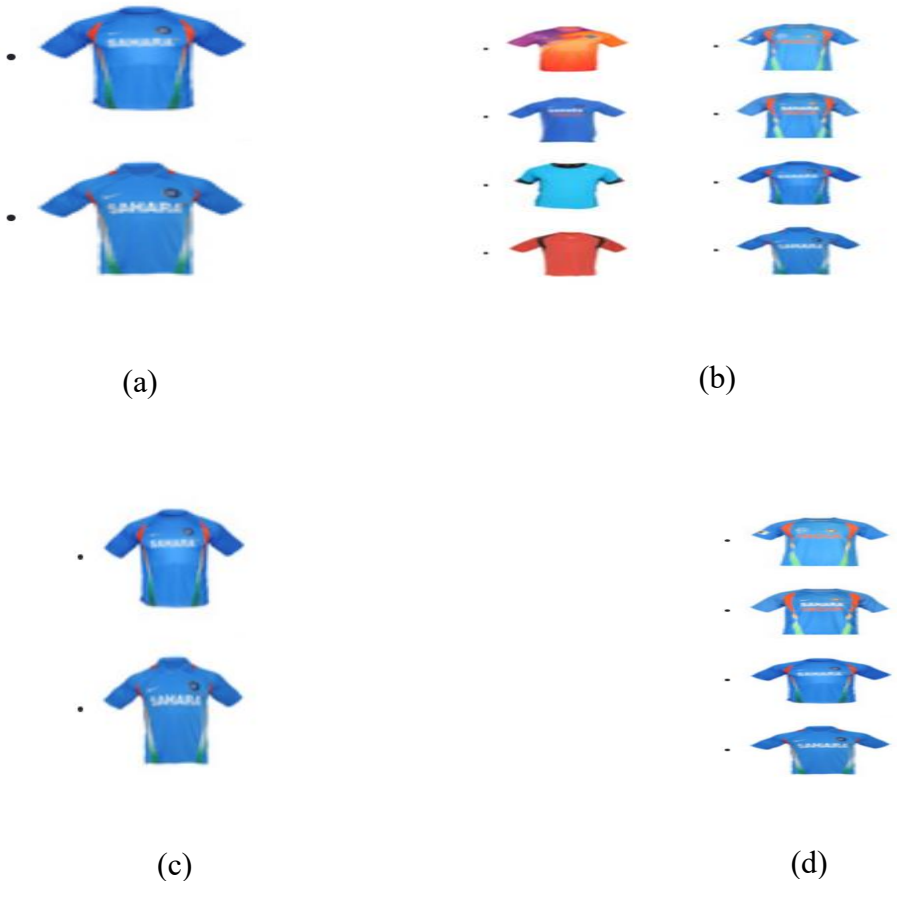
b) Pre-trained MobileNetV2 model trained on ImageNet dataset with Cosine Similarity

In the previous method Euclidean distance was used as the similarity measure. In this method, Cosine similarity is used as the similarity measure with the MobileNetV2 Model. The feature database for the image database is same as is created and saved in the method of Euclidean distance. The only change is in the similarity measure used in this method. As is in the formula for cosine similarity. The dot product of vector A and B is divided by the product of magnitude of A and B. The cosine similarity runs between -1 to 1, where the 1 stands for the perfect match.

The starting value was chosen as Similarity >0.85. The site presented the results given in figure 4.3.3 (a). Since, the results are matching with the input image. To get more

similar images, the threshold is relaxed to >0.75. The output received on the site is depicted in figure 4.3.3 (b). The results with threshold >0.75 resulted in some output which is not very similar to the input image provided and therefore, the threshold was increased to >0.80. The results with this threshold are presented in figure 4.3.3 (c). Since, this output is exactly same as the output received with similarity >0.85. The ideal number exists between 0.75 to 0.80. And after few tests, the final similarity threshold which gave the most matching images without and image which is not matching was achieved at 0.78. The results with similarity >0.78 are depicted in figure 4.3.3 (d).

The output gives 4 matching images for the input image. This is one more than which were found with Euclidean where 3 matching images were found using the rounded off thresholds.



(a) (b)

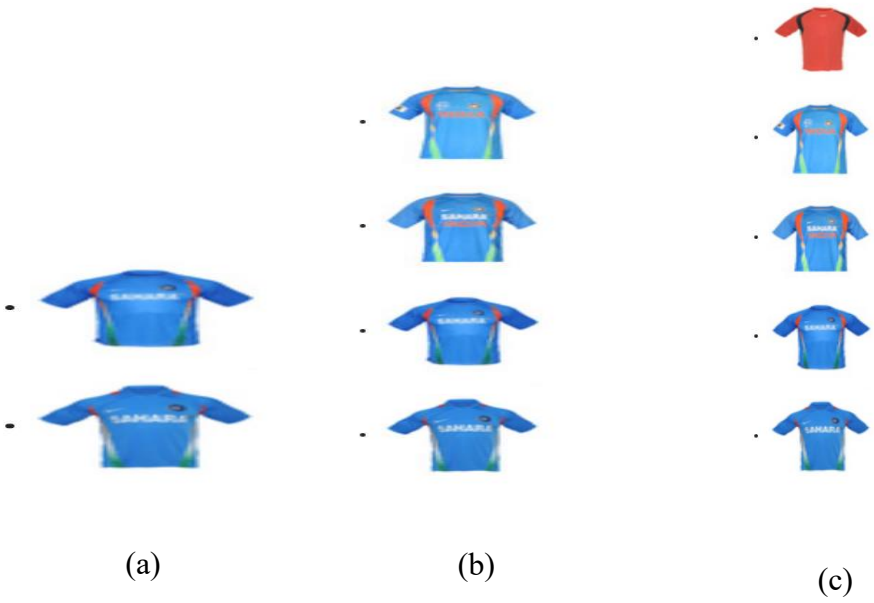
(c) (d)

Figure 4.3.3 Results with Pre-trained MobileNetV2 model and Cosine Similarity

c) Pre-trained MobileNetV2 model trained on ImageNet dataset with Manhattan Distance

Next model which was tested on the dataset was MobileNetV2 but with Manhattan distance. Again, the feature database which was already extracted using the MobileNetV2 model is used and the only change required is in the php file of the website which does the similarity check of the uploaded image features with the features of all images in the dataset.

The code takes the absolute difference in the vector elements. If the features are normalized between 0-1, the Manhattan distance ranges between 200 to 400. We started testing the mode with distance of <300. The output of this model is given in the figure 4.3.4 (a). Since only two matching images were received as the output, the distance was increased to < 350 to get more images. The output of the mode with this threshold is given figure 4.3.4 (b). Since, the output received was equal to the highest matching images received with Cosine distance. One another test was done to check if the model gives any further matching images. Therefore, the distance was increased to <360. The results of the model with this threshold are given in figure 4.3.3 (c). This output resulted in an image which was not matching with the input image. Therefore, the best match with this model is given with the threshold of <350. This result is matching with the best match which was received using the Cosine distance.



(a) (b) (c)

Figure 4.3.4: Results with Pre-trained MobileNetV2 model and Manhattan Distance

Since, the results with different similarity measures are similar with Cosine and Manhattan given same output when the absolute threshold values are inserted. The next model is tested with Manhattan distance similarity measure only.

d) Pre-trained ResNet50 model trained on ImageNet dataset for feature database along with MobileNetV2 model for image feature extraction with Manhattan distance

The next model which is tested on the database was the ResNet50 pretrained model trained on the ImageNet dataset. This model was tested with the Manhattan distance similarity index.

ResNet 50 is a 50 layers deep network which uses the residual connections to avoid the vanishing gradient problem. At first, the feature extraction of the whole image database is done using the ResNet50 model. For this the top classification layer was removed using (include_top=False) and the global average pooling method was applied (pooling='avg'). Before given the images it as input to the model, the images are resized to 224*224 size as this the input size expected by the ResNet50 model. The images are converted to numerical tensor and pre-processed to normalize the pixels. This is done using tf.keras.applications.resnet50.preprocess_input() method. The pre-processed image tensor processed in the CNN and gives the feature embedding of 2048 dimensional vector after applying the global average. These features vectors are saved in the SQLite database which is then uploaded to the custom website database for usage.

Under this model, the uploaded image is using the same MobileNetV2 model which is used in the previous mode. The objective is to check whether the model renders any matching images. The test was conducted but No images were rendered by the model as there was compatibility between the features generated by MobileNetV2 model and the feature database created using the ResNet50 model.

e) Pre-trained ResNet50 model trained on ImageNet dataset with Manhattan distance

The next model used was to use the ResNet50 feature database along with ResNet50 pretrained model for the uploaded image.

The uploaded image on the website is saved in the website database and then the uploaded image is first resized to 224*224 pixel and the feature extraction is done using the pre-trained ResNet50 model. It uses the same flow of first converted the image to numerical tensor and then the pixels are standardised under pre-processing. The ResNet50 model is used without the top layer and feature vector is generated for the uploaded image.

Once the feature vector of the uploaded image is available. The same is compared with the features of all the images saved in the feature database and the matching images are found using the Manhattan distance as similarity measure.

The output received on the website using the distance of <650 (the higher threshold is put to see the most matching images which the model can give) is given in figure 4.3.5 (a). The output resulted in the 7 matching images and 3 on similar images. The next threshold was tested with the distance of 625 and the output of this threshold is given in figure 4.3.5 (b).

(a) (b)

Figure: 4.3.5 Results with Pre-trained ResNet50 model and Manhattan Distance

From the output, it can be seen that even when 7 images were present in the database. The model is given only 5 matching images and the 6th image is dissimilar one. This could be due the image resizing as that reduces the resolution of the image. The next model was tried using the image padding instead of image resizing to make the image resolution of 224*224 pixels.

f) Pre-trained ResNet50 model trained on ImageNet dataset with image padding and with Manhattan distance

In this model, pretrained ResNet50 model trained on ImageNet dataset is applied on image dataset which is pre-processed with image padding instead of image resizing to make it compatible with ResNet50 model. This makes sure that the original aspect ratio of the image is preserved. This ensures that the image is not distorted by stretching or compression. The shapes, texture and spatial feature of the original image are maintained. After resizing each image is converted into numerical array and pre-processed to normalize the pixels. Once this done, the pre-processed image array is processed through the ResNet50 model without the top classification layer. The final feature embeddings of all the images are saved in the SQLite database which is later saved on the website database for usage.

After the feature database is created and saved in the website database. The image is uploaded on the website and saved in the database. After that the image is recalled and unlike direct resizing which can distort the image, the pre-processing of the image is done using the function ImageOps.pad() to resize the image to 224*224 pixel by adding the black borders padding around the original image. The technique makes sure that the original is preserved as it is. After padding, the image is converted into numerical array and is standardized using preprocess_input() function form Keras. Then the image is put to pretrained ResNet50 model without the top classification layer, so as to get the feature embeddings which represent the texture, shape and spatial relationship of the image.

The first testing of the model is done using the Manhattan distance of <325 which gave the output as given in the figure 4.4.6 (a). Since, this threshold resulted in only one image output. The threshold was revised to 425. The output received using this threshold is given in figure 4.3.6 (b). The threshold 425 gave 4 correct matching images. In order to get more images, the threshold is relaxed to < 450. The output with this threshold is figure 4.3.6 (c). This resulted in somewhat 5 very similar images and 2-3 less similar

images. To get more accurate results the threshold is revised to 430 and then 432 and finally at 434 most number of matching images were found which are presented in the figure 4.3.6 (d). This output gave 8 matching images which are the highest till now given by any model.

(a) (b) (c) (d)

Figure 4.4.6 Results with Pre-trained ResNet50 model with image padding and Manhattan Distance

g) Fine-tuned (on articleType classification) pre-trained ResNet50 model trained on ImageNet dataset with cosine similarity

Till now the pre-trained models were tested without any modification. The next model test was done by finetuning the pretrained model on the image classification given in the dataset. The most granular classification of the images is given by the articleType and therefore, the finetuning of the model was done using this parameter. The pretrained model selected for this exercise was ResNet50 as till the now two pre-trained models

were used in ResNet50 was more deep structure and therefore, ResNet50 was selected so that the comparison of results could be done.

The image dataset had their corresponding metadata in the CSV file. In the CSV file each image is identified by ID which is nothing else but the image name for which that row stores the metadata. The article type category has 143 different item including Shirts, Jeans, T Shirt, Shoes etc. For this model the dataset was divided into 80:20 ratio where 80 % was used as training data and 20% as testing data.

For training the images were first rescaled to 224*224 pixel so that it becomes compatible with ResNet50 model. The last layer of the model which is a classification layer was removed and the custom classification head was added. The new head included global average pooling layer, a dense layer of 512 neurons with ReLU activation to lean the discriminative representatives, dropout layer to reduce overfitting and a softmax output layer whose size corresponded to the total number of unique article types in the dataset.

The model was compiled using the Adam optimizer a categorical cross entropy loss function with accuracy as the performance metric. In order to ensure in case of interruption of the code run, the progress is not lost, BackupAndRestore and ModelCheckpoint callbacks were used. The model was trained for 5 epochs with the batch size of 32. After the completion of finetuning the model was saved in H5 file for the use of feature extraction.

Then the finetuned model which had the penultimate output layer of 512 dimension was used to extract the features of the image dataset. The images were resized to 224*224 pixels as is required by ResNet50 model. Then normalization of the pixel is done by scaling the value from 0 to 1. Each image after processing generated a 512 dimensional embedding vector which was then saved into the SQLite database table with column image_name and features and was finally uploaded to the custom website database for model usage.

It took comparatively 5 to 6 times more time to run the model in comparison to using the pre-trained model as this included first training a new model from the dataset, that too which required too many transformations and pre-processing and then extraction of features.

Now, when the image is uploaded on the custom website, the image is first stored on the database and then the image passes though the same finetuned model which was used to get the feature database. The image is first resized to 224*224 pixel and then is converted to numpy array and normalized by scaling between 0 to 1. Then the image is passed through the model and a 512-dimension feature embedding is received as the output which is then compares with the features database of all images. The similarity measure used in the case is cosine similarity.

The reason to use cosine similarity in finetune model is that Manhattan and Euclidean distance measures are scale sensitive. And during the finetuning resulting vector can have varying magnitudes due to layer weights finetuning, activation function, dropout randomness. On the other hand, cosine similarity focus on direction and not magnitude. It is scale invariant and it is tend to capture the sematic closeness much better. With high dimensions, cosine similarity tends to be more effective.

The model was tested multiple times but every time results were very random. One of the results are shown in figure 4.3.7 which uses the threshold >0.95.

Figure 4.3.7 Results with Finetuned ResNet50 model on classification and Cosine Similarity

The output was checked even with 0.99 accuracy and with other similarity measure as well but the model was unable to produce any meaningful results.

h) Contrastive Learning Fine-tuned pre-trained ResNet50 model trained on ImageNet dataset with Cosine Similarity

Since in the last model the finetuning was done with respect to the articleType and then the results were completely random. This time the finetuning was done using the contrastive learning technique. This method is expected to give better results as it takes

an image for training and produces both positives and negative contrasts of that image and train the model to increase the embedding closeness of image with the positive copy and reduce the closeness of embedding with the negative copy of the image.

The ResNet50 pre-trained model is taken as a backbone and a contrastive loss function is used to finetune the model on fashion industry dataset. To create the distinct augmentations of the image different transformation were used including random horizontal flipping, random rotation (up to +-20%) and random zooming (up to 20%) and resizing to a fixed resolution of 224*224 pixel.

To finetune the model, the last classification layer is removed and the average pooling layer output is used as a general-purpose feature extractor. The extracted features are then passes through two fully connected layers i.e. Dense (512, ReLU activation) and Dense (128). The first of these provides non-linear transformation and enables the feature adaption and the second produces a 128-dimension feature vector. All the layers were made trainable on which the fine tuning was done using contrastive training method.

The images are converted into 128 vector and are fed into the model in pairs. The output of the images is normalized and then the model calculates the cosine similarity between every pair of embeddings in the batch both positive and negative. If there are say 32 images in a batch then it will have 64 embeddings. Then the contrastive loss function is applied to make the model minimize the distance between positive pairs while maximising the distance between negatives. And this how the model is finetuned.

After finetuning the model is saved and then the same model is used to extract features of all the images in the database. The features of all the images are produced in 128 dimension vector and are saved in SQLite database.

Then the image is uploaded on the website and saved. Then it is resized to 224*224 pixel and normalized. The images are passed through the model and a 128 dimension vector is generated as output and the embeddings are compared with the feature database embeddings of all the images.

The model is tested using the cosine similarity with the threshold of 0.95. The output is given in figure 4.3.8 (a). Since in the output only two products are visible out of which

only one was matching. Another run was made with similarity of >0.94. The output with 0.94 threshold is given in figure 4.3.8 (b).

(a) (b)

Figure 4.3.8 Results with Contrastive Learning Finetuned ResNet50 model and cosine Similarity

i) Custom Build Contrastive Learning ResNet50 model trained on fashion industry dataset with Cosine Similarity

The last model which has been tried on the dataset was a custom build ResNet50 model. The weights of the model are trained from ground up rather than finetuning of already trained weights using the ImageNet dataset.

The model uses contrastive learning technique and learns to first create augmentation of the image in order to train the model weights. Each image passes through the augmentation process which creates two similar images which are then put into the model in pairs. While image augmentation is done the image is resized to 224*224 acceptable size for the Resnet50 model, the image augmentation is done using the random flip, random rotation 20%, random zoom 20% method. The final classification layer of the ResNet50 model is removed and a global average pooling layer converts the convolution feature in vector. A small head is appended to create a 128-dimension vector. All images are appended to 128-dimension vector and then normalized. The contrastive loss function is used to bring close the embeddings of the positive image to anchor image and takes far the negative image embedding to the main image. The

images are put into the model in pairs and in batches. The training of the model is done using 5 epochs.

Once the fitting of the model is done it is downloaded and the same model is used to extract the features of all the images in the database. The images are resized to 224*224 size, normalized and the output embedding of 128-dimension vector is saved in the SQLite database.

The uploaded image on the website goes through the same process of resize to 224*224 pixels, normalization and the production of 128 dimension embedding. The embedding of the uploaded image is compared with that of the feature database and cosine similarity is used to find the matching images.

This model took at least 10 times more time for run to fit and providing the feature database in comparison to the pretrained models.

The output from the custom build ResNet50 Model based on contrastive learning with cosine learning threshold of 0.997 is given in figure 4.3.9 (a). Since the output with similarity >0.997 gave only one matching image and rest all were non-matching images. The similarity threshold was increased to 0.998 for checking the results. The results are given in figure 4.3.9 (b). Still, the output is only one matching image and one non-matching image.

(a) (b)

Figure 4.3.9 Results with Contrastive Learning custom ResNet50 model and cosine Similarity

4.4 Summary

In this chapter, the analysis was conducted on the fashion industry dataset using various models. At first the fashion industry dataset was selected which had 44.4 k images across various products. The visual techniques were used to analyse the metadata of the images. The analysis was done on gender, masterCategory, subCategory, articlyeType, basecolour, season, year and usage. After having understanding of the data and its suitability for the study, the data was used to test/build the models. The first model which was used was the pretrained MobileNetV2 model on ImageNet dataset. The images were resized to 128*128 pixel to make them compatible with the model. Then the images were converted to numerical tensor and normalised to make the value between 0 to 1. The data was first pre-processed using either resizing the images by stretching or padding as per the model requirements. After pre-processing the model was used to testing. The MobileNetV2 model was used with three different similarity measures. The output gave 3 matching images with Euclidean distance and 4 matching images with Cosine Similarity and Manhattan distance. The next model which was used was pre-trained ResNetV2 model. For this mode the images were resized to 224*224 and then after converting to numerical tensor were normalized between values 0 to 1. The output from this model gave 5 matching images. Then the pre-trained model was used with image padding for resizing the image, this gave 8 matching images. The other models which were tested were the finetuned ResNet50 model on articleType and finetuned ResNet50 model using contrastive learning on fashion industry dataset. The classification finetuning gave completely random results and contrastive finetuning gave 1 matching image only. The final custom-made model again gave only 1 matching output.

- **CHAPTER 5 RESULTS AND DISCUSSIONS**

5.1 Introduction

In order to achieve the task of finding the best model for reverse image search for fashion industry dataset, at first the dataset was visualized, data was pre-processed, transformed and finally various CNN models were used on the dataset including pretrained MobileNetV2 model and ResNet50 model on ImageNet dataset, finetuned ResNet50 model on articleType classification of the fashion industry dataset, finetuned ResNet50 model using contrastive learning by generating positive and negative images for each image under the dataset and finally a customer build ResNet50 model using the contrastive learning technique. The MobileNetV2 model was tested to check the impact of various similarity measures including Euclidean Distance, Cosine similarity and Manhattan Distance on the results of the model. Along with this it was also tested whether two models i.e. MobileNetV2 and ResNet50 can be combined to complete the reverse image search task.

5.2 Modelling Methods and Results

After having the understanding of the data various methods including MobileNetV2 pretrained model on ImageNet dataset, ResNet50 pretrained model on ImageNet dataset, ResNet50 model with image padding, ResNet50 model finetuned on articleType, ResNet50 model finetuned using contrasting learning and finally the custom build ResNet50 model trained on the fashion industry dataset using contrastive learning. The MobileNetV2 model is implemented with three Similarity measures including Euclidean Distance, Cosine Similarity and Manhattan distance. Based on the results experience in MobileNetV2, the pretrained ResNet50 model and ResNet50 model with image padding is used with Manhattan distance only. The finetuned models and customer build model uses Cosine similarity as the similarity measure.

The images in the dataset are present in 60*80 pixels format. In order to use the images as input to the models, while using the MobileNetV2 mode the images are pre-processed to 128*128 pixels as the image input is required to be in square format for the model. For the ResNet50 Model the images are resized to 224*224 pixels. The images are first converted to numerical data and are then standardized before the same are sent as input into the model.

The sample image which was used as input for uploading against which the matching images were extracted is given in Figure 5.2.1.

Figure 5.2.1 The sample image used for testing

The MobileNetV2 model was used to extract the matching images with three similarities i.e. Euclidian, Cosine and Manhattan and the results were more or less similar in all three cases. But it was noticed that with absolute thresholds, the cosine similarity and Manhattan distance was able to produce more matching outputs which was 4 in number, without even giving any mismatching outputs.

The final outputs of all these three models respectively are depicted in figures 5.2.2, 5.2.3 and 5.2.4.

Figure 5.2.2 Final output of MobileNetV2 with Euclidean Distance

Figure 5.2.3 Final output of MobileNetV2 with Cosine Similarity

Figure 5.2.4 Final output of MobileNetV2 with Manhattan Distance

The ResNet50 Model with the pretrained weights from ImageNet dataset was able to give better results with 5 matching images. This model used Manhattan distance as the similarity measure. The final results from this model are shown in figure 5.2.5

Figure 5.2.5 Final output of Pre-trained ResNet50 model with Manhattan Distance

The next model which was used, was the pretrained ResNet50 model on ImageNet dataset but this time image padding was used to resize the images to 224*224 pixels. This method kept the original texture and exposure of the image and therefore the images did not loose the information. This method in its final test was able to give output to 8 matching images which is the best output till now from among all the models. This output is visible in the figure 5.2.6

Figure 5.2.6 Final output of Pre-trained ResNet50 model with Image padding and Manhattan Distance

Now that the testing was done using the pretrained models, the next test was done to check whether this pre-trained model on Imagenet dataset can give better results if it is finetuned on the specific fashion industry dataset under consideration. To conduct this test, the model was finetuned using two methods, first was finetuning the model by fitting the weights so that the images predict the correct articleType of the image. The finetuned model could not give better results instead the model was became completely distorted. The results of this test are given in figure 5.2.7.

Figure 5.2.7 Final output of Finetuned ResNet50 model on classification with cosine similarity

The finetune model with contrastive learning at least did not distort this much but again it lost the accuracy which was inbuild in the trained mode. The results from this mode are given in figure 5.2.8

Figure 5.2.8 Final output of Contrastive finetuned ResNet50 model with cosine similarity

Finally, the custom build model was applied which was trained on contrastive learning but his time the training was done from ground up. There were no pretrained weights which were in built into the model. But again the results were not any better but instead there are was only one matching image. The output from the custom build ResNet50 Model based on contrastive learning with cosine learning threshold of .998 is given in figure 5.2.9.

Figure 5.2.9 Final output of Contrastive custom ResNet50 model with cosine similarity

Therefore, from among the models which were tested on the fashion industry dataset, the ResNet pretrained model on ImageNet dataset, with preprocessing of images using padding method to make the pixel size to 224*224 and with cosine similarity measure gave the best results with 8 matching images.

The models are evaluated based on the following.

1. Recall@1: How often the top most results is from the same category as the query.
2. Recall@5/@10: How often the at least 1 relevant image appears in the top 5 or top 10 retrieved images.
3. mAP (Mean Average Precision): Overall retrieval precision quality. How well the ranking aligns with the correct labels across the dataset.
4. T-SNE plot: 2d visualization of embedding clusters – similar article Types should group together.
5. Intra-class vs Inter-class distance: Average cosine distance between images of same vs different categories.
6. Cluster purity/Silhouette score: Quantitative measure of how tightly items in a class cluster together.
7. Retrieval of Top 5 image results

Evaluation Summary for the self-trained ResNet50 model from scratch, fine-tuned models and ResNet50 with Image padding:

a) Recall and Precision:

Table 5.2.1 Recall and Precision of Models

	Model	Recall@1	Recall@5	Recall@10	mAP
0	ResNet50 (Scratch)	1.0	1.0	1.0	0.378605
1	SimCLR ResNet50	1.0	1.0	1.0	0.459474
2	ResNet50 Fine-tuned	1.0	1.0	1.0	0.541431
3	ResNet50 (Pretrained + Padding)	1.0	1.0	1.0	0.708221

The recall @1 for all four models which means that the top most result is from the same category as the query. Recall@5 and Recall @10 is 1 which means everytime at least one relevant image appears in top 5 and top 10 results. Mean Average Precision for ResNet50 model built from scratch is the lowest at 0.378605, followed by Contrastive finetuned ResNet50 model at 0.459474, then ResNet50 label finetuned model at 0.541431 and maximum for ResNet50 (Pretrained and Image padding model) at .708221. Which means that the ResNet50 with image padding models is able to retrieve images from the correct labels class with maximum precision among these four models. Figure 5.2.10 depicts this comparison among these four models.

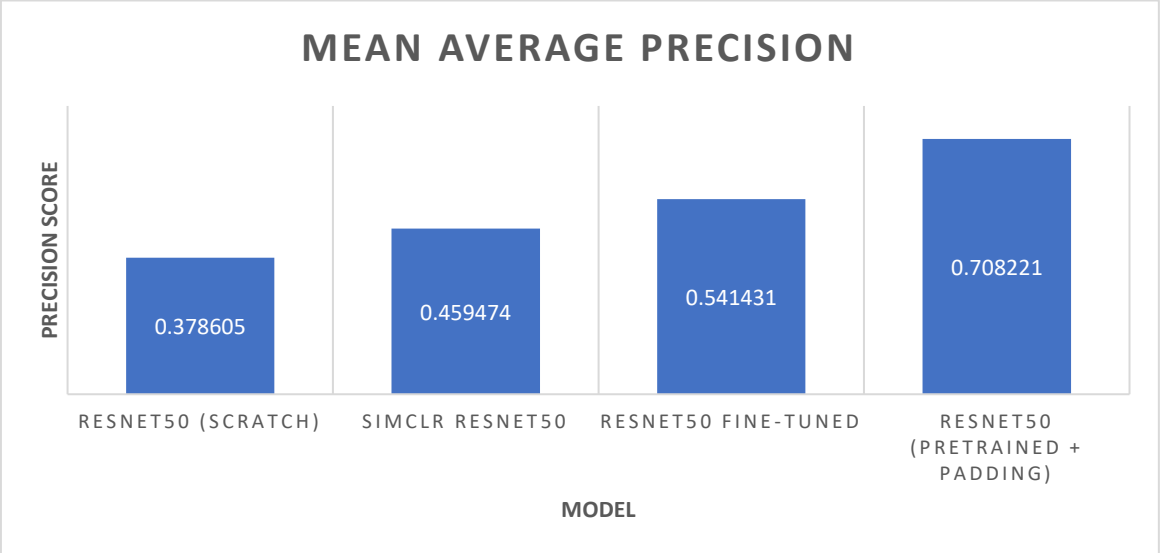


Figure 5.2.10 Precision score of Models

b) t-SNE plots:

In the plot each dot represents an image embedding from the dataset. Dots are coloured by their class label. The algorithm arranges the points so that the image with similar embeddings appear close and dissimilar images appear farther apart. The figure 5.2.11 shows that the embeddings are scattered which shows that the model is not fitted properly, secondly model in Figure 5.2.12 shows a slight improvement in clusters from Figure 5.2.11 but still is scattered. The label finetuned model in Figure 5.2.13 well defined clusters which means the models has fitted good to the data and the images from the same class have close embeddings and the last model in Figure 5.2.14 shows even better clusters meaning that the outputs would be more accurate.

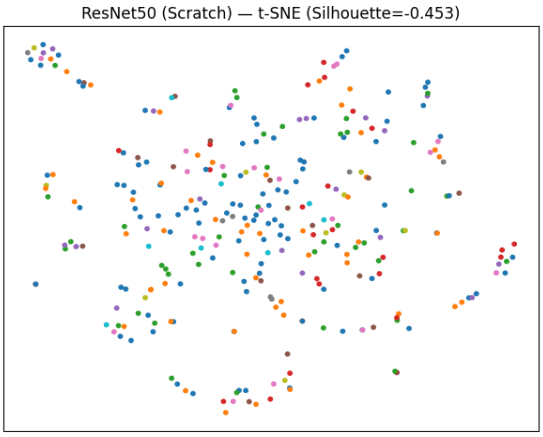


Figure 5.2.11 t-SNE for ResNet50 from Scratch Model

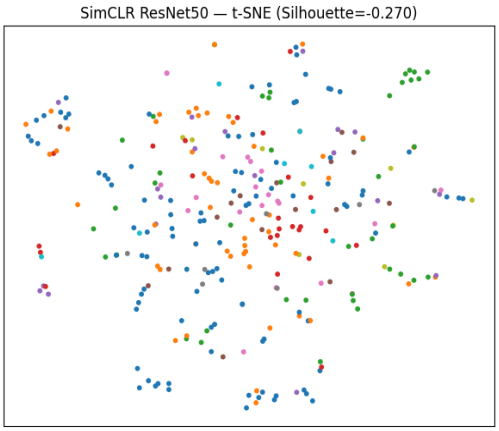


Figure 5.2.12 t-SNE for ResNet50 finetuned with Contrastive Learning

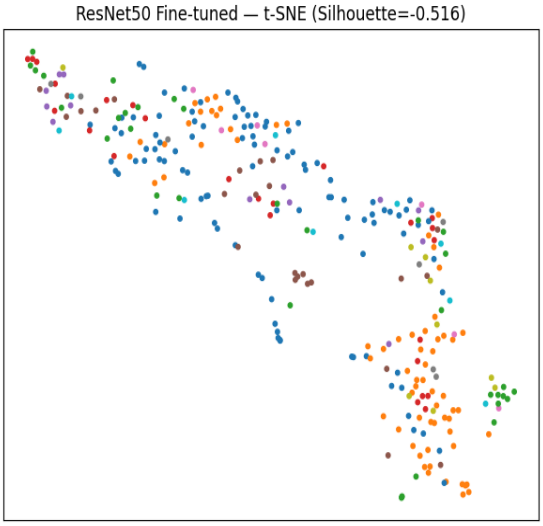


Figure 5.2.13 t-SNE for ResNet50 finetuned with label Model

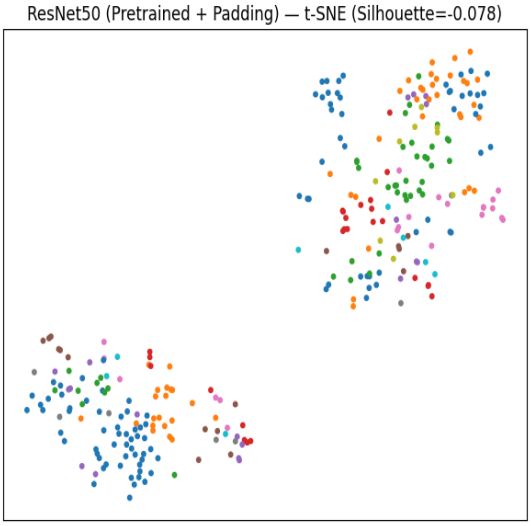


Figure 5.2.14 t-SNE for ResNet50 with Image Padding Model

Table 5.2.2 shows that the Silhouette Score which shows how well the samples are clustered with values closer to 1 shows separation and negative value means overlapping clusters. ResNet50 model with image padding has the best score with -0.078 means having the best clustering. The intra class similarity shows how close embeddings from the class are. Here model finetuned on label gives best score with the lowest distance of 0.054785. The inter class similarity represents how close embedding are from different classes, here the higher value is better which is given by Contrastive Learning Model. Separation ratio is InterClass divided by Intra class value. For Separation ratio, higher is better. The issue with ResNet Model trained from Scratch and finetuned Model with Contrastive learning is it didn't bring close the embedding of images from the same classes.

Table 5.2.2: Embedding Separation Summary

Sr. No.	Model	Silhouette	Intra-Class	Inter-Class	Separation Ratio
0	ResNet50 (Scratch)	-0.453485	0.767261	0.990459	1.290903
1	SimCLR ResNet50	-0.270317	0.709757	0.996486	1.403981
2	ResNet50 Fine-tuned	-0.515556	0.054785	0.203822	3.720384
3	ResNet50 (Pretrained + Padding)	-0.078462	0.168644	0.359602	2.132321

c) Image retrieval using the models:

Figure 5.2.15 gives the results from ResNet50 model built from scratch. It can be seen that the results are not matching as the embeddings of the images are far for the same class images which was noticed through the Silhouette score and Intra Class distance.

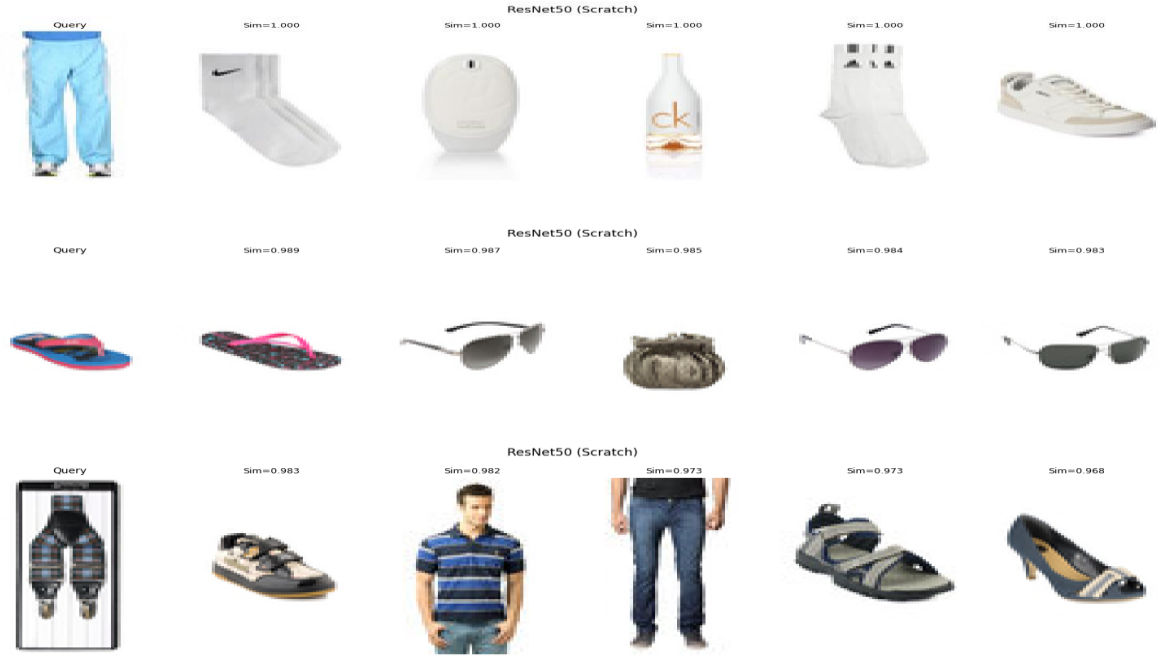


Figure 5.2.15 Image Retrieval results from the ResNet50 model trained from scratch

Figure 5.2.16 gives the results from ResNet50 model finetuned with Contrastive learning. It can be seen that the results are not matching in some cases as the embeddings of the images are still far for the same class images which was noticed through the Silhouette score and Intra Class distance.

Figure 5.2.16 Image Retrieval results from the ResNet50 model finetune with Contrastive Learning

Figure 5.2.17 gives the results from ResNet50 model finetuned with label. It can be seen that the results are matching but in some cases it is giving results from different gender class. The results are consistent with the Silhouette score and Intra Class distance.

Figure 5.2.17 Image Retrieval results from the ResNet50 model finetuned with Label

Figure 5.2.18 gives the results from ResNet50 model with image padding. It can be seen that the results are matching and are best among the tested models. But in some cases, it is giving results from different gender class. The results are consistent with the Silhouette score and Intra Class distance.

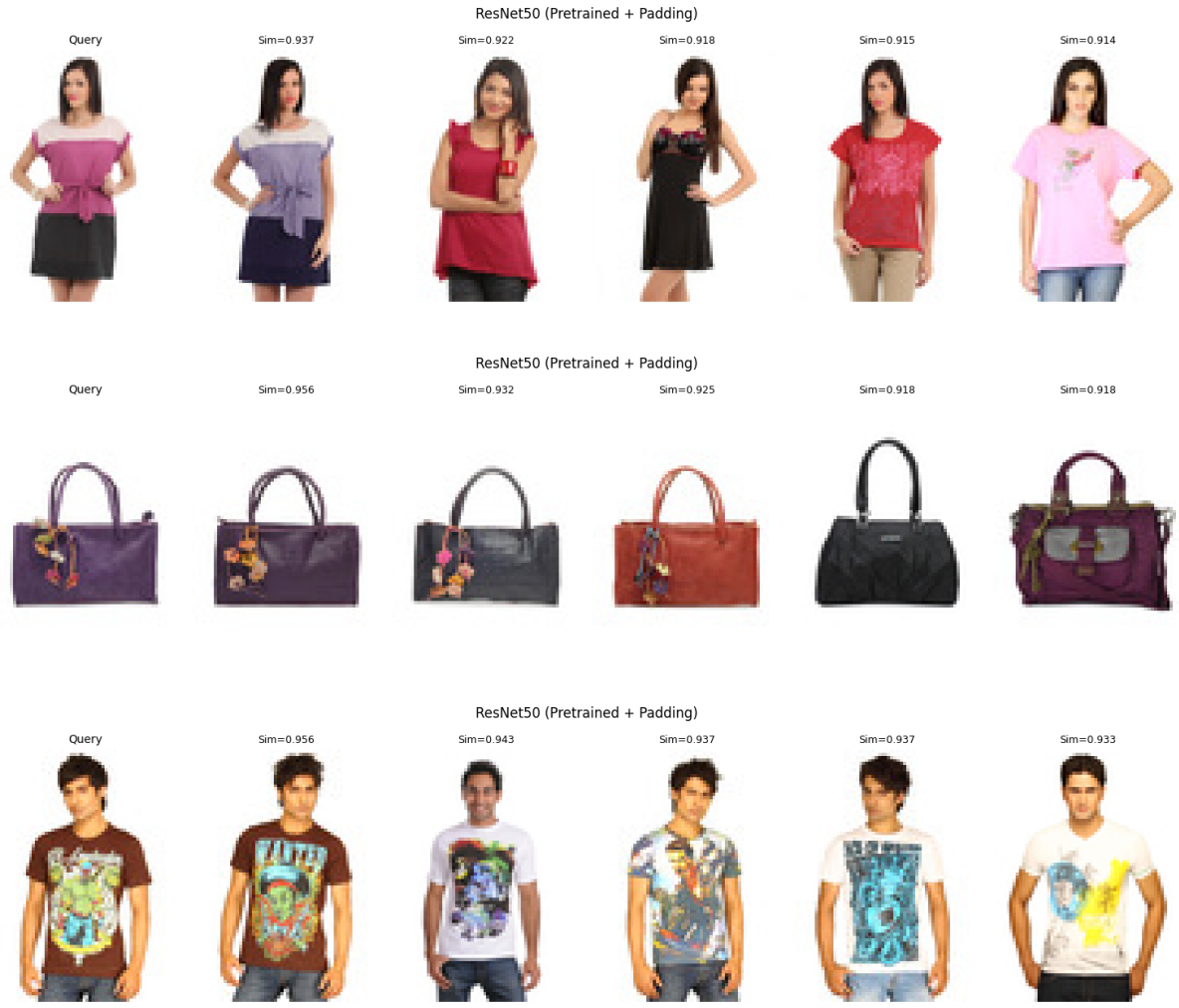


Figure 5.2.18 Image Retrieval results from the pre-trained ResNet50 model with Image Padding

5.4 Summary

Different selected models were tested on the fashion industry dataset in order to get the best model for getting the reverse image search. In order to conduct this, the dataset of images was first pre-processed to make it fit for the models. For the MobileNetV2 model the images were pre-processed to 128*128 and for ResNetV2 to 224*224 pixel. The images were converted to numerical tensors and then normalized before using them as input to the models. The pretrained MobileNet2 model gave the maximum of 4 matching images with Cosine and Manhattan similarity measure. In which comparison the pretrained ResNet50 model performed better both with and without padding. With padding it gave 4 matching images and without padding it gave 5 matching images. The last test was done with

finetuning of models and then customer build model. The finetuned model on classification was a complete misfit. Whereas the finetuned model with contrastive learning and custom build ResNet50 Model using contrastive learning both gave only one matching image. Therefore, the best model was pretrained ResNetV2 model on ImageNet dataset with image preprocessing using padding method which gave 8 matching images from the fashion industry dataset.

- **CHAPTER 6 CONCLUSION AND RECOMMENDATIONS**

6.1 Conclusion

In this study, various models both pretrained including MobileNetV2 and ResNet50 and custom made with ResNet50 as baseline, preprocessing methods for resizing the images and standardization, similarity measures including Euclidean, Cosine and Manhattan were tested for reverse image search on the fashion industry dataset. The best result which was received for the sample image was using the ResNet50 model with image padding which gave maximum 8 matching images for the sample test image. Without image padding the same model gave 5 matching images. Therefore, the impact and importance of image pre-processing is noticed in the study. Whereas the pretrained MobileNetV2 model also fairly performed well with output of 4 maximum matching images but the results from ResNet50 model were better. Apart from the models, the similarity measures although performed equal only that Manhattan did not require to go to decimals to get the best results as compared to Euclidean. The custom fitted model and the finetuned model were not able to match the outputs given by the pre-trained ResNet50 model with padding. If results of the finetuned and custom build model are considered they both gave maximum one matching image and then the non-matching images. This is likely due to variables and methods which were used in the models.

Looking at the evaluation measures and other image retrievals it was noticed that the ResNet50 model fined tuned with labels performed relatively well and the results were very close to the best performed pre-trained model with image padding. Table 6.1 depicts the comparison of these two models.

Table 6.1.1 Comparison of Evaluation scores of ResNet50 with Padding and finetuned ResNet50 with labels

Model	mAP	Intra Class Score	Inter Class Score	Separation Score
ResNet50 (Pre-trained Padding)	0.707221	0.054785	0.203822	3.720384
ResNet50 Finetuned on Labels	0.541431	0.168644	0.359602	2.132321

It is noticed that the best Mean Average Precision for ResNet50 with image padding and cosine similarity model is best which is followed by ResNet50 finetuned model with labels. In fact, Intra Class embedding distance score is better for the finetuned ResNet50 model on labels. This shows that the finetuning worked although the Inter Class score is lower compared to pretrained model with image padding. Looking at the various image retrieval from these two models it was noticed that the finetune models did give results which were very competitive and had the potential. It can be concluded that amongst the tested models pretrained ResNet50 model with image padding performed best on the dataset.

6.2 Recommendations

Since the model finetuned with class gave the very close results to the pre-trained mode with image padding. It is recommended that if the finetuned model with labels is upgraded to also consider the contrastive learning along with the label while finetuning. The model can be significantly improved. This would improve the model inter class score and the finally the retrieval results.

For the standalone contrastive learning model, it gave a clear idea that the contrastive learning model needs improvement. Since the models was only doing contrastive learning instead of considering both the category of the image and the contrast. This led to model considering a scenario where each image has its own class. It only trains on the low-level details i.e. texture, shadow and colour tone but the information like articleType, category etc. are ignored. This particular issue was also confirmed by checking the model output from a different sample images. The output from new sample tests is given in Figure 6.1 which is the sample image and Figure 6.2 which is the output, Figure 6.3 which is the second sample image and Figure 6.4 which is the output for this sample image.

Figure 6.1 New Sample Image

Figure 6.2 Output from the new sample image

It can be seen from Image 6.1 and 6.2 that the output is able to detect some shape and colour texture of the product but is not giving the output from the same articlyType. Also, if more training may be done, some improvement may come with respect to shape and colour.

Figure 6.3 New second Sample Image

Figure 6.4 Output from New Second Sample Image

Here, also it is visible from image 6.3 and 6.4 that the colour is matching and the shape is matching if the sample image is flipped anticlockwise. So is clear that if the model considers the articleType as well while training a better model can be found.

The other improvement could be that the batch size taken was very small of 16 images only. A higher batch size if trained together might have allowed to bring in more information while the losses were minimized. Also, the augmentation also used only flip, zoom and rotation, may be augmentation using cropping, brightness, contrast, etc. could make a significant difference. At last, the training was done only on 5 epochs that may be one of the reasons.

Considering the above the following can be tested:

a) Use of category wise contrastive learning technique instead of only contrastive learning can produce better results. Image of same articleType can act as positives.
b) The batch size can be increased to 128 or greater.
c) The epochs can be increased to 30 to 50. Although this would require a lot higher resources but for production ready model this can be done.
d) The augmented images are generated using all techniques i.e. rotation, flip, zoom, crop, brightness adjustment, contrast adjustment, etc.

These changes should give a model which would produce more accurate results and may be equal or better than the pre-trained model.

This discussion has been around the ResNet50 model only. There are many other model options which can be tested for the given situation. Form among the Single Image Models we have used only MobileNetV2 and ResNet50 models. The other models which are MobileNetV3, Efficient Net, InceptionnV3 and VGG16/19. All these models are also available with pre-trained versions.

Apart from these, under region-based models RCNN, Fast RCNN and Faster RCNN model performance can be checked on the task. This could be useful if the uploaded image has multiple objects in the picture. The object detection task can be faster handled by the Yolo and SSD models. And lastly the single image models can be used along with the contrastive learning and triplet loss techniques which would come under one shot learning models. All these experiments can be done in order to find a better model. Although, the ImageNet model performed decently well and can be used after testing it with multiple images.

The final point which is also very important is that due to limited resources, the fitting was done on small fashion industry dataset which was only 60*80 pixel image dataset. In order for models to fit better, the large size dataset would be a better option which gives the image data in 2400*1600 pixels. This would result in model to capture the texture, contrast, colour, shape and size etc. more accurately which would result in a better model fit for reverse image search for the fashion industry dataset.

Finally, it can be concluded that among the tested models pretrained ResNet50 model with image padding is best recommended for ecommerce fashion industry data reverse image search. It is further concluded that an improved model is possible when the finetuning is done both on contrastive learning and labels along with image padding. Along with this,

other factors including batch size, epochs, image augmentation, other models, improved image quality can be further be tested.

References

Addagarla, S.K. and Amalanathan, A., (2021) e-SimNet: A visual similar product recommender system for E-commerce. *Indonesian Journal of Electrical Engineering and Computer Science*, 221, pp.563–570.

Adrakatti, A.F. and Mulla, K.R., (2023) LIS Perspective on Multimedia Information Retrieval Techniques A Critical Analysis. *International Journal of Library Science*, 21.

Ali, S.N., Ahmed, Md.T., Paul, J., Jahan, T., Sani, S.M.S., Noor, N. and Hasan, T., (2022) Monkeypox Skin Lesion Detection Using Deep Learning Models: A Feasibility Study. [online] Available at: http://arxiv.org/abs/2207.03342.

Bansal, I., Dhar, S., Yuvraj, Gitanjali and Nikam, (2021) A Framework and Techniques for Image-based Search Application with an E-commerce Domain. In: *Proceedings of the 5th International Conference on Trends in Electronics and Informatics, ICOEI 2021*. Institute of Electrical and Electronics Engineers Inc., pp.1102–1109.

Bitirim, Y., (2022) Retrieval Effectiveness of Google on Reverse Image Search. *Journal of Imaging Science and Technology*, 661.

Dcosta, W., Meena, S.M., Gurlahosur, S. V. and Kulkarni, U., (2022) Optimization of multi-objectives for Adaptive Reverse Recommendation method on edge devices. In: *2022 13th International Conference on Computing Communication and Networking Technologies, ICCCNT 2022*. Institute of Electrical and Electronics Engineers Inc.

Diyasa, I.G.S.M., Alhajir, A.D., Hakim, A.M. and Rohman, M.F., (2020) Reverse image search analysis based on pre-trained convolutional neural network model. In: *Proceeding - 6th Information Technology International Seminar, ITIS 2020*. Institute of Electrical and Electronics Engineers Inc., pp.1–6.

Eswaran, A. and Varshini, E., (2022) Reverse Image Search Engine for Garment Industry. In: *8th International Conference on Advanced Computing and Communication Systems, ICACCS 2022*. Institute of Electrical and Electronics Engineers Inc., pp.414–418.

Laamouri, A. and Sael, N., (2025) Image Recommendation Using Clustering Techniques: A Comparative Study. In: *Lecture Notes in Networks and Systems*. Springer Science and Business Media Deutschland GmbH, pp.236–245.

Li, B., Li, J. and Ou, X., (2022) Hybrid recommendation algorithm of cross-border e-commerce items based on artificial intelligence and multiview collaborative fusion. *Neural Computing and Applications*, 349, pp.6753–6762.

M Vinitha, Dr.B. Nagarajanaik, Mallikarjuna Nandi, C Naga Sri Charan and K Priyanka, (2024) Fashion Recommendation System. *International Research Journal on Advanced Engineering Hub (IRJAEH)*, 205, pp.1243–1247.

Mawoneke, K.F., Luo, X., Shi, Y. and Kita, K., (2020) Reverse Image Search for the Fashion Industry Using Convolutional Neural Networks. In: *2020 IEEE 5th International Conference on Signal and Image Processing, ICSIP 2020*. Institute of Electrical and Electronics Engineers Inc., pp.483–489.

Nanayakkara, P.R., Jayalath, M.M., Thibbotuwawa, A. and Perera, H.N., (2022) A circular reverse logistics framework for handling e-commerce returns. *Cleaner Logistics and Supply Chain*, 5.

Nayak, A., Shah, J., Kuruvilla, A., Akshaya, J. and Sandesh, J.B., (2021) Fine-grained Fashion Clothing Image Classification and Recommendation. In: *Proceedings - 2021 2nd International Conference on Electronics, Communications and Information Technology, CECIT 2021*. Institute of Electrical and Electronics Engineers Inc., pp.600–606.

Onesim, R.I., Alboaie, L., Pricop, A. and Panu, A., (2020) Counterfeiting scalable detection image based system for e-commerce. In: *Proceedings of the ACM Symposium on Applied Computing*. Association for Computing Machinery, pp.1914–1919.

Pate, A., Francis, B., Sathe, V., Kalantri, R. and Rajguru, S., (2023) Transfer Learning-Based Recommendation System. In: *2023 6th IEEE International Conference on Advances in Science and Technology, ICAST 2023*. Institute of Electrical and Electronics Engineers Inc., pp.383–388.

Patil, S., Deshmukh, S., Yadav, P. and Chaudhari, S., (2022) Smart Adware and Product Recommendation using Object Detection. In: *IEEE International Conference on Data Science and Information System, ICDSIS 2022*. Institute of Electrical and Electronics Engineers Inc.

Rui, C., (2021) Research on Classification of Cross-Border E-Commerce Products Based on Image Recognition and Deep Learning. *IEEE Access*, 9, pp.108083–108090.

Salman, A.D. and Hasan, E.H., (2023) Survey Study of Digital Forensics: Challenges, Applications and Tools. In: *Proceedings - International Conference on Developments in eSystems Engineering, DeSE*. Institute of Electrical and Electronics Engineers Inc., pp.788–793.

Singh, M.K., Chakraverti, A. and Gupta, S., (2024) Comprehensive Analysis on Image Search Engines. In: *Proceedings - 4th International Conference on Technological Advancements in Computational Sciences, ICTACS 2024*. Institute of Electrical and Electronics Engineers Inc., pp.1438–1442.

Singh, P.N. and Gowdar, T.P., (2021) Reverse Image Search Improved by Deep Learning. In: *2021 IEEE Mysore Sub Section International Conference, MysuruCon 2021*. Institute of Electrical and Electronics Engineers Inc., pp.596–600.

Sodani, A., Levy, M., Koul, A., Kasam, M.A. and Ganju, S., (2021) Scalable Reverse Image Search Engine for NASAWorldview. In: *COSPAR 2021 Cross-Disciplinary Workshop on Machine Learning for Space Sciences, Sydney, Australia*. [online] Available at: http://arxiv.org/abs/2108.04479.

Stoica, F. and Pelican, E., (2025) A machine learning approach of enhancing eCommerce solutions. *Expert Systems with Applications*, 274.

Taipalus, T., (2024) Vector database management systems: Fundamental concepts, use-cases, and current challenges. *Cognitive Systems Research*, 85.

Thoiba Singh, N., Chhikara, L., Raj, P. and Kumar, S., (2023) Fashion Forecasting using Machine Learning Techniques. In: *2023 IEEE International Conference on Integrated Circuits and Communication Systems, ICICACS 2023*. Institute of Electrical and Electronics Engineers Inc.

Yadav, S.M., Joshi, S.K., Mandavia, D.T. and Puri, M.P., (2024) Retroflex: Uncovering Visual Equivalences through Reverse Image Recon. In: *11th International Conference on Computing for Sustainable Global Development (INDIACom)*. IEEE.

Zhang, P., (2021) E-commerce products recognition based on a deep learning architecture: Theory and implementation. *Future Generation Computer Systems*, 125, pp.672–676.

Zhao, W., Liu, X., Xu, R., Xiao, L. and Li, M., (2024) E-commerce Webpage Recommendation Scheme Base on Semantic Mining and Neural Networks. *Journal of Theory and Practice of Engineering Science*, 403, pp.207–215.

Zubair, M., Alim, M.A., Naseem, I., Alam, M.M. and Su'Ud, M.M., (2023) Reverse Image Search for Collage: A Novel Local Feature-Based Framework. *IEEE Access*, 11, pp.78182–78191.

APPENDIX A: RESEARCH PROPOSAL

MATCHING PRODUCTS RECOMMENDATIONS ON UPLOADING AN IMAGE ON AN E-COMMERCE PLATFORM USING NEURAL NETWORK MODELS

HEMANT KUMAR KHURANA

Research Proposal

MAY 2025

Abstract

Online shopping platforms are one of the most used methods for buying products these days across the world. Most of the platforms are making use of text-based search for making product recommendations to the customers. Whereas there is a great scope of image-based search on these platforms which will allow the user to get the recommendations of the products which are visually matching with the product which the buyer is looking for. The buyer can just provide the image of the product which he or she is look for and similar products would appear for purchase. This objective has been met using various methods over the time and the most effective method which is used currently being used is of Reverse Image Search or Content Based Image Retrieval (CBIR). This technique helps getting products recommendations for the consumer who is trying to buy a product on an e-commerce platform. This method is also used for various applications such as identifying the disease by pictures, face recognition, auto driving cars, traffic lights etc. In order to achieve the Image retrieval, the sub-method which is best used is of Convolutional Neural Network Method. The current study tries to get the products recommendations for a fashion industry dataset. Various models would be applied including pre-trained models for image retrieval and also fitting the standard models such as AlexNet, VGGNet, ResNet, CNN, RCNN and Faster RCNN, One Shot CNN, Yolo and SSD etc. to the database and then using them to get the results. Pre-trained models with fine-tuning and also new models would be tested from ground up using base CNN model and compare the results with varied parameters. The objective of this study is to test multiple models and identify the best model which can provide the most effective recommendations for a given image input for the fashion industry dataset on an e-commerce website.

Table of Contents

1. Background

E-commerce industry is growing day by day. The buyer while making the decision on an e-commerce website makes search for the products which he or she intends to buy. The most commonly used method is the text-based search under which the buyer mentions the name and/or features of the product in the search box and gets the recommendations for the same. This method is being used on the e-commerce website for a long time. One another method which is used is of voice search where the consumer gives the voice input instead of text and the website understands the voice input and gives the related output. The other method which can be of great use is image-based search recommendation. Under this method the customer just has to upload the image of the product which he or she intends to buy and the search engine automatically recognizes the product and make product recommendations from the database which best match the queried image. In order to achieve the reverse image search several methods have been tried including Compact Composite Descriptor, Color and Directivity Descriptor, Fuzzy Color and Texture histogram to achieve image retrieval task with the best being the application of Convolutional Neural Networks also known as CNN (one type of Neural Network Model) (Diyasa et al., 2020).

There are various CNN models such as Single Image Models including AlexNet, VGGNet, GoogleNet, ResidualNet, Region-Based models such RCNN, Fast RCNN and Faster RCNN and One-Shot models such as Yolo and SSD (Ali et al., 2022; Eswaran and Varshini, 2022; Pate et al., 2023). The CNN models extract features of the image and these features are matched with the images in the database and recommendations are made based on the closest match. For doing this similarity search various methods are used such as Euclidean distance, Manhattan distance, Cosine similarity and Annoy Indexing. The one which gives the highest accuracy and efficiency is selected (Mawoneke et al., 2020).

As reverse image search gives the matching images and therefore, it has found varied applications and researchers across the world have tried making use of the same in solving various problem statement. The reverse image technique is also used for object detection, face detection, speech recognition, license plate recognition or disease detection etc. Some of the problem statements which have be solved using the applications of the reverse image search concept are as follows:

i) At present search engines like Google and Bing have started giving the option to the users to make the search based on input of an image. One research paper has also tested the accuracy of search on these search engines (Bitirim, 2022). It used Average precision and Average Normalised Recalls as the accuracy measures at various cut off points.
j) Some Chinese e-commerce websites have started making use of this method and have also found it useful (Mawoneke et al., 2020). One such website, which is the largest Chinese online shopping company named Taobao have made use of the same.
k) This method has also been used to create a search engine on NASA satellite images (Sodani et al., 2021). Under this paper a trained CNN was used to convert the images in list of integers before fitting into the model. The other techniques which can be used for pre-processing are hashing and vectorisation approach.

l) The other application which has been found is in having web page recommendations based on user data related to searches and clicks (Zhao et al., 2024). Based on the user data the model is trained and then used to make website recommendations on future user searches.
m) One research paper used this method in order to detect the Monkeypox Skin disease. The paper used the opensource data and tried various models which could predict the disease well in advance (Ali et al., 2022).
n) The method has been used to lower the product returns by using circular reverse logistics framework for handling e-commerce returns (Nanayakkara et al., 2022).
o) The method has also found application in cross-border e-commerce to curtail the supply of fake products by developing a counterfeiting Scalable Detection Image system (Onesim et al., 2020).
p) Once such paper has found the application of neural network in forecasting of sales and have also discussed the theory of image recognition in e-commerce (Zhang, 2021).

The current study tries to apply this method to an e-commerce website dealing in fashion products. The best part of reverse image search in the application of product matching is that once you see a product anytime and anywhere in real life, you can just click the picture of it and upload it on the related website or tool which supports the particular reverse image search and get the matching recommendations. Say, if someone sees a person wearing a specific shirt or pant and he or she want to buy a similar design product; the person can just click the picture from mobile phone and upload it on the website and buy the similar product. Many times, the buyer likes a particular design in a retail shop but are unable to get the proper size of that shirt or pant. The customer can just click the picture of the shirt and upload on the website and see if the required size of that shirt is available online.

Other applications are if someone likes an artwork or design of bag, wallet, purse, jewellery, shoes etc. at someplace, he or she can see the products online by just taking a picture of it. Similarly, if someone like some electronic product, car, utensil; the matching product can be bought online by just click of a picture.

This has great benefits, it saves the time of the buyer to search relevant product, helps in product marketing, increase the sales for the seller, efficient buying for the buyers, product satisfaction and increase the economy. As when the customer can get the product immediate when he or she like it, it is more probable that the product would be bought.

Under current study, image retrieval method is applied to a fashion industry dataset with the objective to make recommendations of matching products for a queried image on an e-commerce website. The fashion industry dataset has been referred from a research paper dealing with the similar topic (Mawoneke et al., 2020).

The study will make use of the pre-trained models and also finetune them with the dataset. Apart from them, the base models such as ResNet-18, ResNet-34, ResNet-50, ResNet-101, ResNet152, VGG-16, VGG-19, GoogleNet, Inception-V3 etc. would be trained for image retrieval. It would illustrate the accuracy and efficiency of each model for matching products recommendation dataset and would suggest the best model for use on the e-commerce platform for fashion industry dataset.

2. Problem Statement or Related Research or Related Work

Till date multiple studies have been conducted on getting the matching products recommendations using neural network methods. In these studies, multiple approaches have been used to achieve the most accurate and/or efficient result for a certain scenario or dataset. This field has been heavily explored since early 1990s (Mawoneke et al., 2020). One of the first implementation were in websites such as TinEye. The other popular examples are Ditto, Snap Fashion, ViSenze, Cortica, Bing Image Search, Google Image Search, Flickr etc. In the 1990s, the development of e-commerce received government support (Rui, 2021). The American e-commerce company named Amazon at first started the online bookstore. Then improved its logistics and distribution channels and then turned into a global company. They used a Dynamo system, a distributed key-value storage system.

As discussed earlier various methods were tested to get the reverse image search and the best method was found to be the use of CNN models. In 2012, Hinton's research team build the CNN network (Rui, 2021). They build the CNN model names AlexNet which won the championship of ImageNet image recognition competition with an absolute advantage. Since then, various CNN models have been developed which has been the most prominent advancement in reverse image search method.

Multiple papers have tried different methods on different datasets for get the matching product recommendations. Under one such research paper, from which the data has been referred for doing our research, used the CNN model with three layers and six epochs and used Euclidean distance for calculating the similarity score (Mawoneke et al., 2020). This method gave an accuracy of 0.60 based on macro average (each class is given equal weight) and 0.93 based on weighted average (with weights based on occurrences in each class). This problem can be solved more efficiently by testing multiple other CNN models both pre-trained and custom build which are suitable for fashion industry dataset for an e-commerce website. The paper also suggested that Alternative Loss functions such as hamming distance and k nearest neighbours can be used to get the recommendations and the one which provides the most suited can be finally selected.

The various methods which can be used to solve the above problem statement and which have been applied in case of similar problems are discussed ahead.

Transfer learning can be applied by making use of pre-trained model on a similar dataset e.g. MobileNetV2 or pre-process the images into frequencies and scalars (Diyasa et al., 2020). This paper compared the method of perceptual hashing of images and then vectorizing them with the pre-trained CNN model. It used the ImageNet 2012 data which had 1000 classes and 14 million images. It was found that the perceptual hashing method is faster but CNN method is more accurate. The models used in this paper can be adjusted to find the variation in results.

Principal Component Analysis method may be used to filter the image features in order to reduce size for faster images search (Singh and Gowdar, 2021). This method is also suitable in cases where there is some noise and redundant information in the images. This paper used Cosine similarity for loss assessment, ReLU to fit the non-linearity and cosine similarity is used for loss assessment. The image is normalized by dividing by 255 and Conv2d model is applied.

Similar to the fashion industry problem, this technique has been applied to the garment dataset (Eswaran and Varshini, 2022). This paper tests the pre-trained model like VGG16[8], ResNet50[7] and InceptionV3[9] on the dataset and finds that the ResNet50 performs the tasks with the highest accuracy and efficiency. These pre-trained models are based on ImageNet dataset. To speed up the process the nearest neighbour technique is used to convert the feature vector into an efficient indexing format. The paper proposed finetuned ResNet50 model on the custom dataset.

Another better method for image retrieval could be using colour, text and shape features of the images (Bansal et al., 2021). This paper discusses that just colour is not enough for the retrieval process and we should be taking in to account the text and shape features as wel. This paper highlights that different classifiers have different strength and various classifiers can be used to get the image retrieval such as Scale Invariant Feature Transform (SIFT), Maximally Stable Extremal Regions (MSER), Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Speed up Robust Features (SURF). Under the study it was found that the MSER performed better in terms of scaling, PCA is suitable when the images are blur and SURF in terms of saving time.

There are others method for model fitting such as to take into account both Gray and RCB features. Using hash algorithm for expansion of modes and virtual nodes are used to make the data evenly distributed (Rui, 2021).

Hybrid recommendation algorithm can be devised (Li et al., 2022). This paper talks about the methods including indexing of images, using metadata, colour difference histogram, SIFT, spatial direction tree, local patch extraction, vocabulary tree etc. for reverse image search. This paper used the bag of words model and the first fitted the model through back propagation.

Another paper creates an e-commerce website using AI and ML in the website for both the buyer and seller, introduces chatbot for customers, using Power BI tool for business analysis for seller and applies various methods such as VGG16, VGG19, Exception etc. on the dataset (Pate et al., 2023). It also talked about the use of 5 CNN models such as VGG16, VGG19, ResNet50, Inceptoin50 and Exception on the movie dataset. It tested the models on an e-commerce website created using JavaScript and recommended the use of ResNet50 model.

Under current study, the impact of different methods would be identified including pre-trained models, newly trained models, using different loss functions, different methods for pre-processing of images on the fashion industry dataset, as are used in various research papers on similar problems which are discussed above. The model would be found which is best suited for e-commerce website dealing in fashion industry products.

3. Aim and Objectives

The aim of the study is to test various models on the fashion industry dataset and to find a model for reverse image search for an e-commerce platform which gives the best matching product recommendations.

The objectives are as below:

- To access the effectiveness of the pre trained models in terms of accuracy for reverse image search on an ecommerce platform for fashion industry.
- To build new custom-made models and compare the results with pre-trained models.
- To measure the effect of pre-processing of data and of different loss functions on results.

4. Significance of the study

The study would provide the best model, data pre-processing method and loss function which can be used to get most accurate product matches for the fashion industry dataset for an e-commerce website. The better the accuracy of the models, the better would be the usability for the customers as they can get the products which are matching their requirements. This method can be used in implementing image-based search on ecommerce platforms along with the text-based search.

The method would assist the buyer to not just try to find a product by explaining it in words but by just pulling out the mobile phone and click the image of the product which he/she is seeing and can buy the same or similar product online.

The method would result in economy in e-commerce as there would be less returns for the purchase if the customer is buying the product which is actually matching the requirements.

It would also improve the overall economy as the buyer is more likely to give positive reviews of the products and drive higher customer satisfaction when the product is actually of one's liking.

And the overall sales would also increase as when the customer is able to get the relevant products at the time when the need actually arise, the buyer is more likely to make the purchase in comparison to a scenario when he/she is likely to make a search at a later time period.

5. Scope of the study

The current study focuses on the fashion industry dataset. It has various categories for each image such as gender, colour, type of wear, season for wearing, usage, etc. The scope is to create a website which can take the input and to test various pre-trained and custom based models which can present the products to the user which are best matching the product in the image which the user has uploaded. For example, if the image input is of shoes, then the user will get the matching shoes from the database which best match the features of the product in the uploaded image. Here the scope is limited to recommending the products which are matching the features of the product of the uploaded images. The study can be

further extended to say get the products which are from the same category such as of same colour, same season wear, same usage etc.

6. Research Methodology

Under this study Neural Network Models would be used to get the matching images recommendations from the Fashion Industry dataset using CNNs as part of reverse image search methodology. The various CNN models available are Single Image Models including AlexNet, VGGNet, GoogleNet, ResidualNet, Region-Based models such RCNN, Fast RCNN and Faster RCNN and One-Shot models such as Yolo and SSD etc.

The CNN models divide the input image into convolutions and process one convolution independent of the other. This makes the process faster, efficient and also helps in reducing the over-fitting of the model. Each CNN model has hidden layers, fully connected layers and the activation function. Each hidden layers have multiple kernels which are used to extract features from each convolution which is applied at a certain part of the input image at one time and rolled over across the image with the decide step size. In case of grey scale images i.e. black and white, the kernels used are of one channel and in case of coloured images the kernel are of three channels for taking care of colours i.e. RGB. For getting values out of each kernel certain pooling method is used to get single value which is used as an input for the next layer. Each neuron in a hidden layer thus produces a feature map and the number of feature map produced in a layer is equal to the number of neurons in a layer. For layer 2 each neuron is applied on a certain area of all the feature maps produced in layer one and thus produces a new feature map. After application of all the hidden layers the final input goes through a fully connected layer to present a vector to be applied on a activation function. Under the fully connected layer the neuron is applied to whole of the features maps rather than a certain patch and thus produces a final vector to be passed through the activation function. The activation function finally gives the prediction. The property of CNN to extract features is used for getting the matching images by matching the features of the input image with the features of database of stored images.

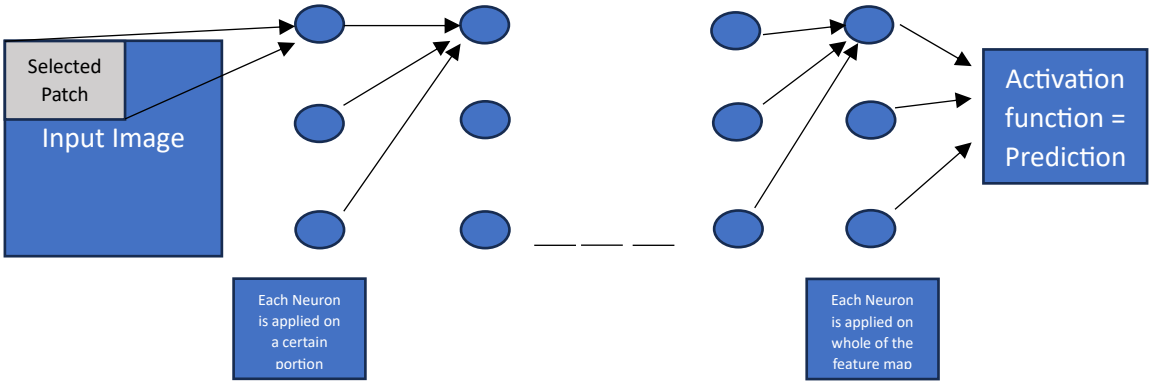


Figure 1: The structure of CNN Model

The CNN models are used for various real-life applications such as image classification, object detection, face recognition etc. For performing these tasks during the training process the CNN models extract features from the images and train itself using a loss function for the making the most accurate and efficient prediction. The characteristic of feature extraction of CNNs makes them best suited for using in the application of Reverse Image Search. Under this approach the model compares the feature of the queried image with the features of images in the database and presents the images which have the closest features match with the queried image. There are various methods which are used to get the similarity score such as Euclidean distance, Manhattan distance, Cosine similarity and Annoy Indexing.

The challenge in performing the image-based search lies in the size and relevance of data. The bigger the size of the data, the higher is the processing time and cost. Therefore, the selection of model is also based on the time the model takes to present the result. Various techniques are used for indexing the data so that the processing time is reduced. These techniques include hashing, vectorization, mean features so that the size of the data is reduced.

a) Introduction to the dataset

This study uses the fashion industry dataset. It includes 44,000 images with different product categories such as clothes, gift sets, socks, sports items, jewellery, bags, belts, shoes, hair colours, hand bands etc. The dataset is available in 2400*1600 image size as well as in smaller version. The size of big data is 25 GB and of the smaller version is 593 MB. The study would test the effect of image size in the models whether it effects the accuracy of the results. The data is available open source on Kaggle in both the sizes. It has one folder named images which contains all the images with the with the image names as numbers and an excel file which mentions the information about each image such as gender, category, article type, colour, season, usage and product name. In order to make use of the dataset of reverse image search we would need to extract features of all the images and store it in a database.

b) Creation of Feature Database

In order to test a model accuracy for the suitability of an applicability on fashion industry dataset, at first the features would be extracted for all the images in the dataset using the selected model and stored in a SQL database. Each image would be loaded using the image path then the selected model would be applied on the image and feature would be extracted. While extracting the features the images would be resized and normalized. This process is applied to all the images in the folder and the extracted features are stored in a database. The database would have two columns one for image name and other containing image features array.

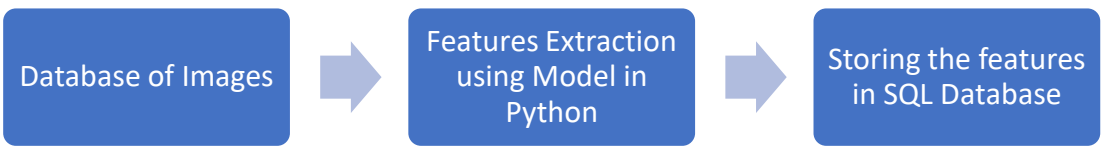


Figure 2: Process for creating the feature database for database of all images

This process would be first tried by storing the images on the computer and running the model on the spyder app. In case there is space issue, in such a case the data would be uploaded on google drive and the model would be run on google collab with mounting of google drive to access the images and the final database would be downloaded on the computer.

This process would be applied separately for all the models which would be tested. In other scenarios where the features extracted by one model are compatible with other models, in such cases the results would be tested with cross models as well where the features stored are from one model but the model used to extract feature of the input image is from different model.

For the feature extraction there are two options, either to use the pre-trained model such as MobileNetV2 and extract the features and use the same model for input image and the matching the features. The other approach would be to first fit the model to the dataset. While fitting the model, the images are uploaded in batches to the model. The complete batch first goes through the forward pass and then loss is computed and then weights are updated through the back propagation. This process happens until all the batches images are processed through the model. Then there is an option to how many times the complete set of images would go through model, for that the parameter used in modelling is number of epochs. Once the model it fitted to the image dataset. The next steps would be same as are used in the pre-trained model.

c) Application of different Models and Loss functions

In order to find the matching images, the loss functions such as Euclidean distance, Manhattan distance, cosine similarity and annoy indexing would be used in order to find the one which provides the best results.

Both pre-trained and custom based models would be used for feature extraction. In order to build the custom based model, the overall data would be divided into train and test dataset in the ratio of 70:30 for fitting the model. The models which would be fitted on the dataset would include AlexNet, VGGNet, GoogleNet, ResidualNet, RCNN, Fast RCNN and Faster RCNN, One-shot CNN models such as Yolo and SSD in order to find the model which achieves the highest accuracy for reverse image search on the fashion industry dataset for an e-commerce website. There is variety available for each of the model with different layer structures and feature extractor. Under this study different models would be tried and the results would be compared in between them.

In order to test the model, a custom build website would be used which would have the feature of uploading the image and presenting the matching images on the webpages after processing. This website would be hosted on the localhost server on the laptop using XAMPP Control Panel.

The results of different models both pre-trained and newly trained along with impact of data pre-processing and loss function on the final results would be summarized and the best combination would be found.

The flow chart of the working would be as follows:

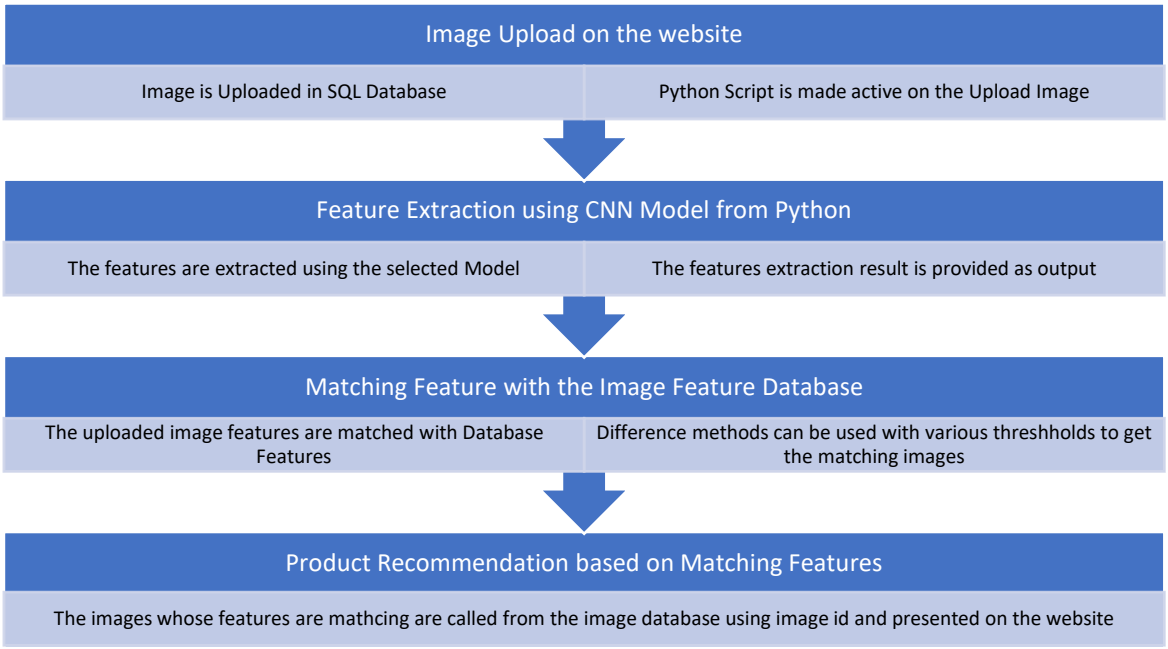


Figure 3: Process for getting matching images on e-commerce website

The accuracy of the model would be tested based on either precision and/or recall method. Where in precision would be calculated using the relevant images produced by total images produced and recall would be calculated using the relevant images produced by the total relevant images.

d) Creation of e-commerce website for testing

A website would be created using php language which would provide the option to upload the image by user and would clicking the upload button would present the matching images. For getting the matching images, python script would be created which would extract the features of the uploaded image using a specific model and then the feature of this image would be matched with the features of various images stored in the database and the images whose feature are closest that of the uploaded image would be presented on the website.

7. Required resources

In order to complete the study, the following resources would be required:

- A laptop with Windows 11 or latest, 16 GB RAM, 512 SSD or above, 100 GB available disk space and access to the internet
- Google Drive accessibility for storage of image data
- Google Collab for features extraction from the image data
- Jupyter notebook for running the python scripts
- SPYDER software for python script testing
- Sublime text editor for php coding
- XAMPP Software for hosting the website and SQL database

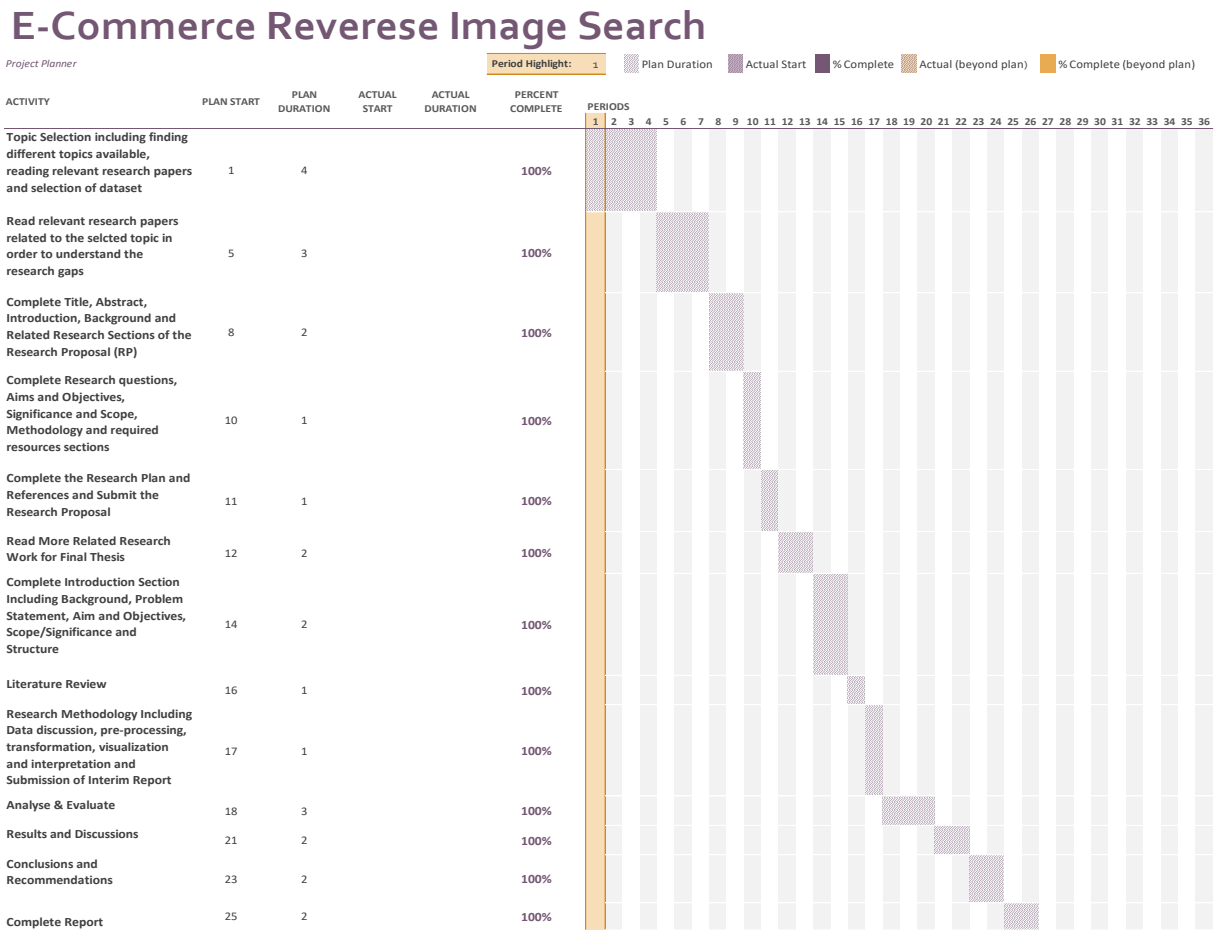
The models would be first tested in Jupyter notebook using the Anaconda distribution. In case an increase in speed is required to fit the models the same would be done in google collab using a higher GPU capacity to get the desired results.

8. Research Plan

The research plan is to finalize the research topic and dataset in four weeks. Then start reading the research papers on the selected topic and find the research gap in the next three weeks. Complete all the chapters of research proposal in next four weeks and submit the research proposal.

The planned timeline for the final thesis is to complete the interim report in six weeks and to complete the final thesis in another nine weeks' time.

The plan for the interim report is to read the research papers in the first two weeks. To complete the Introduction section of the thesis which would include the Background, Problem Statement, Aim and Objectives, Scope/ Significance and Structure of the study in another two weeks. This will be followed by Literature review in the next week and research methodology including dataset discussion, pre-processing, transformation, visualization and interpretation in another week. After the interim report, to the plan is to complete the fourth chapter on analysis in three weeks, to complete the chapter on results from models and discussions in the two weeks, followed by conclusions and recommendations in two weeks. Complete the full report including references and annexures and finally, to make the presentation and submit the files in another two weeks.



References

Ali, S.N., Ahmed, Md.T., Paul, J., Jahan, T., Sani, S.M.S., Noor, N. and Hasan, T., (2022) Monkeypox Skin Lesion Detection Using Deep Learning Models: A Feasibility Study. [online] Available at: http://arxiv.org/abs/2207.03342.

Bansal, I., Dhar, S., Yuvraj, Gitanjali and Nikam, (2021) A Framework and Techniques for Image-based Search Application with an E-commerce Domain. In: Proceedings of the 5th International Conference on Trends in Electronics and Informatics, ICOEI 2021. Institute of Electrical and Electronics Engineers Inc., pp.1102–1109.

Bitirim, Y., (2022) Retrieval Effectiveness of Google on Reverse Image Search. Journal of Imaging Science and Technology, 661.

Diyasa, I.G.S.M., Alhajir, A.D., Hakim, A.M. and Rohman, M.F., (2020) Reverse image search analysis based on pre-trained convolutional neural network model. In: Proceeding - 6th Information Technology International Seminar, ITIS 2020. Institute of Electrical and Electronics Engineers Inc., pp.1–6.

Eswaran, A. and Varshini, E., (2022) Reverse Image Search Engine for Garment Industry. In: 8th International Conference on Advanced Computing and Communication Systems, ICACCS 2022. Institute of Electrical and Electronics Engineers Inc., pp.414–418.

Li, B., Li, J. and Ou, X., (2022) Hybrid recommendation algorithm of cross-border e-commerce items based on artificial intelligence and multiview collaborative fusion. Neural Computing and Applications, 349, pp.6753–6762.

Mawoneke, K.F., Luo, X., Shi, Y. and Kita, K., (2020) Reverse Image Search for the Fashion Industry Using Convolutional Neural Networks. In: 2020 IEEE 5th International Conference on Signal and Image Processing, ICSIP 2020. Institute of Electrical and Electronics Engineers Inc., pp.483–489.

Nanayakkara, P.R., Jayalath, M.M., Thibbotuwawa, A. and Perera, H.N., (2022) A circular reverse logistics framework for handling e-commerce returns. Cleaner Logistics and Supply Chain, 5.

Onesim, R.I., Alboaie, L., Pricop, A. and Panu, A., (2020) Counterfeiting scalable detection image based system for e-commerce. In: Proceedings of the ACM Symposium on Applied Computing. Association for Computing Machinery, pp.1914–1919.

Pate, A., Francis, B., Sathe, V., Kalantri, R. and Rajguru, S., (2023) Transfer Learning-Based Recommendation System. In: 2023 6th IEEE International Conference on Advances in Science and Technology, ICAST 2023. Institute of Electrical and Electronics Engineers Inc., pp.383–388.

Rui, C., (2021) Research on Classification of Cross-Border E-Commerce Products Based on Image Recognition and Deep Learning. IEEE Access, 9, pp.108083–108090.

Singh, P.N. and Gowdar, T.P., (2021) Reverse Image Search Improved by Deep Learning. In: 2021 IEEE Mysore Sub Section International Conference, MysuruCon 2021. Institute of Electrical and Electronics Engineers Inc., pp.596–600.

Sodani, A., Levy, M., Koul, A., Kasam, M.A. and Ganju, S., (2021) Scalable Reverse Image Search Engine for NASAWorldview. [online] Available at: http://arxiv.org/abs/2108.04479.

Zhang, P., (2021) E-commerce products recognition based on a deep learning architecture: Theory and implementation. Future Generation Computer Systems, 125, pp.672–676.

Zhao, W., Liu, X., Xu, R., Xiao, L. and Li, M., (2024) E-commerce Webpage Recommendation Scheme Base on Semantic Mining and Neural Networks. Journal of Theory and Practice of Engineering Science, 403, pp.207–215.