Aging Face Verification Using Deep Learning

by

Fairooz Nawar Bushra
16241010
Farhat Lamia Elma
16201058
Ramisa Sadeque Khan
16241004
Shiana Shahba
16241008

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
October 2020

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

FAIROOZ NAWAR BUSHRA

Fairooz Nawar Bushra
16241010

Shiana Shahba

Shiana Shahba
16241008

Farhat Lamia Elma

Farhat Lamia Elma
16201058

Ramisa Sadeque Khan
16241004

Approval

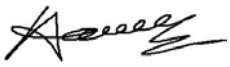
The thesis/project titled "Aging Face Verification Using Deep Learning" submitted by

1. Fairooz Nawar Bushra (16241010)
2. Farhat Lamia Elma (16201058)
3. Ramisa Sadeque Khan (16241004)
4. Shiana Shahba (16241008)

Of Summer, 2020 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on October 5, 2020.

Examining Committee:

Supervisor:
(Member)



Amitabha Chakrabarty, Phd
Associate Professor
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Md. Golam Rabiul Alam, PhD
Associate Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Mahbubul Alam Majumdar, PhD
Professor
Department of Computer Science and Engineering
Brac University

Ethics Statement

We, hereby declare that this thesis is based on the results we obtained from our work. Due acknowledgement has been made in the text to all other material used. This thesis, neither in whole nor in part, has been previously submitted by anyone to any other university or institute for the award of any degree.

Abstract

The era of technological security has grown more attention and interest than ever in the past few years. From wired video surveillance and passcodes, to wireless cameras and facial recognition. Over the years Deep learning has seen a high rate of improvement and its approaches for facial recognition and verification have been observed to have the most optimistic results. Our research focuses on the analysis of different Convolutional Neural Networks (CNNs) that have been developed in recent years. We carry out an extensive analysis of the differences in the performances of the VGG-19 architecture, the ResNet-50 architecture, the InceptionResNet v2 architecture and the Xception architecture while verifying images of the same or different identities with a large age gap on the two widely used datasets namely the MORPH-II dataset and the FG-NET dataset. Our results show that the VGG-19 model has an accuracy rate of 58.005%, InceptionResNet v2 has 44.26%, ResNet-50 has 35.26% and lastly, VGG-19 has 24.74%.

Keywords: Convolutional Neural Network (CNN); Face Recognition; Deep Learning; Aging Face Recognition; Face Verification; Deep Neural Networks.

Acknowledgement

To begin with, we want to thank our Almighty for empowering us to direct our research by giving us the capacity to invest our best amounts of energy and complete it. Secondly, we would like to express our sincere admiration to our supervisor Dr. Amitabha Chakrabarty sir for his criticism, support, direction and commitment in leading the exploration and arrangement of our report. He offered us his valuable time by directing us consistently and being available to offer any kind of assistance we requested from him. We additionally prefer to stretch out our appreciation to our family and friends, who guided us with thoughtfulness and gave us the motivation we needed with their support. Lastly, we express gratitude toward BRAC University for giving us this wonderful opportunity to conduct this research and confer a Bachelor's degree.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

ϵ Epsilon

AdaGrad Adaptive Gradient Algorithm

ATT American Telephone and Telegram

AUC Area Under a Curve

CACD Cross-Age Celebrity Dataset

CACD − VS Cross-Age Celebrity Dataset Verification Subset

CMYK Cyan Magenta Yellow Black

CNN Convolutional Neural Network

DBN Deep Belief Network

DEX Deep Expectation

EDA Emperical Data Anylitics

EMR Electronic Medical Records

FC Fully-Connected

FMR False Match Rate

FN False Negative

FNMR False Non Match Rate

FP False Positive

FSN Fuzzy Swarm Net

GAN Generative Adversarial Network

HSV Hue, Saturation, and Value

HTER Half Total Error Rate

IARPA Intelligence Advanced Research Projects Activity

IFD Itelligent Fault Diagnosis

$IJB-C$ IARPA Janus Benchmark C

$ILSVRC$ ImageNet Large Scale Visual Recognition Challenge

$IMDB-WIKI$ Internet Movie Database-Wikepedia

IRP Iterative Regional Proposals

$L-NMS$ Landmark-based Non Maximum Supression

MLP Multi-Layer Perceptron

MTL Multi-task Learning

NMS Non Maximal Suppression

NN Neural Network

PCA Principal Component Analysis

$R-CNN$ Region based Convolutional Neural Network

$ReLU$ Rectified Linear Unit

RGB Red, Green and Blue

RLN Representation Learning Sub-Net

$RMSE$ Root Mean Square Error

ROC Receiver Operating Characteristic

SRM Structural Risk Minimization

SSN Social Security Number

SVM Support Vector Machine

SVR Support Vector Regression

TN True Negative

TP True Positive

VGG Visual Geometric Group

Chapter 1

Introduction

In this section we will explain our research objective, research contribution and how important image verification is in the real world today. We will state its many advantages and usefulness in the commercial world.

1.1 Research Objective

The purpose behind this research is to explore the discrete architectures that are implemented in the areas of computer perception for face recognition and verification and make a comparative analysis of which architecture provides the optimal result. Primarily our research objectives comprise of :

- Through this research we are hoping to figure out how the CNN architecture helps in boosting the verification process in images of individuals with large age gaps.
- Devising an experimentation to implement the architectures that are used in modern face recognition systems.
- Accustoming two distinct datasets for each architecture for an impartial evaluation.
- Through the above combinations of architectures and datasets paired up with similar algorithms, we will be hoping to find comprehensive insight to each architecture.
- The results provided by each architecture will then be compared and analysed to which approach ensures the best performance in accuracy levels.
- Considering the current market trends in this field and providing the best approach to scoring a better result for the market.

1.2 Research Contribution

In today's progressive world, verification of an aging face has become very important resulting in a variety of use ranging from surveillance to identity verification. Researchers have conducted several researches and different approaches have been applied to overcome the different obstacles faced while verifying an aging face using deep learning. In our research, we have conducted a comparison between the different CNN architectures, hence providing the accuracy rate of aging face verification using each of the architectures.

1.3 Problem Statement

Over the past few years, face recognition in real time with enhanced accurate results in a system has been a concern with several methodologies, algorithms and architecture. Thus it is crucial for innovations to be created which would present the best results within the areas of computer perception. In spite of the fact that there's solid inquiry about exertion in this region, face recognition systems are not optimal enough to execute adequately in various types of circumstances within the actual world. Applications such as Picture database examinations, looking through picture databases of convicts, help in finding the probe in news photographs or surveillance cameras at the same time in the social networking websites. It the modern era, incase of the airplane-boarding gate where the face recognition implementation is crucial, to screen passengers to identify and verify whether or not they are the probes who are required for further investigation and also in case of the casinos, where strategic cameras at face height are used to check faces for distinguishing proof functions and capture images of the probe to construct a exhibition for future watch-list, distinguishing proof and authentication work. On the other hand, it could also be the case of a missing child, where the present face of the child needs to be compared and verified, which has been captured by means of a recent picture taken or surveillance camera footage. Very few technologies are able to analyze the pictures that are being screened in order to match and verify whether or not it is the identified individual but with secondary preciseness. It is decisive to figure out the architectures along with the methodologies that pair up to reinforce the certainty of the results.

1.4 Thoughts behind working in the marketing landscape

In a world where we have a lot of security breaches, it is extremely necessary to have an accurate aging face verification system. By using this verification system, we can strengthen our security-law enforcement, bring advancements in health and medicine, and also bring improvements in the marketing and retail sectors. As of now, this system is used mostly for security purposes such as identifying and tracking criminals, identifying terrorists at airports and train stations, as well as finding exploited children. Aging face verification, which is an automated process, can be used for serving various identification processes starting from wired video surveillance and pass codes to patient authentication. It is a hot topic in the current technological and socio-economic landscape.

Aging face verification is an innovation that examines an individual's facial features to affirm identity, grant access to data, and help in diagnosing uncommon hereditary infections as referenced. Facial acknowledgement depends on the bio metric strategy, which particularly distinguishes an individual by assessing the patient's stored information of their physical or conduct qualities depending on that individual's present status. As the medical care industry pushes ahead with paperless frameworks, and with the expanding utilization of EMR arrangements, facial recognition frameworks are starting to be the favored decision of confirmation to address the requirement for more tight security for persistent verification. Moreover, Healthcare applications with incorporated facial acknowledgment consider quicker analysis of uncommon hereditary issues that are regularly conducted with blood tests, giving specialists admittance to patients who are distant and don't have appropriate medical services accessible to them.

Recently, for example, Russia is using facial recognition to tackle Coronavirus, as tens of thousands of cameras are being fitted with facial recognition software on the streets of Moscow. This works in real time, scanning people's faces and sending an instant alert to the police who can detect anyone who might be breaking quarantine laws. So, if a patient affected with coronavirus leaves their house, the cameras in the streets would recognize them instantly. One of the biggest advantages of this system is that it can be done in real time. The quick accuracy of the system is very beneficial for such epidemic protection. It's not only quick but the facial detection is also possible while a person is wearing the face mask, as our module detects the periocular region and extracts them in many steps to match with the patient's face data.

The outburst of facial acknowledgment innovation in our cell phones – for example in iPhone X, constantly quickening the adoption of this biometric innovation. The innovation isn't just going to reshape the business activities however will change the manner in which we live and shop. Antithesis Research predicts that more than 1 billion cell phones will join facial recognition innovation. It implies 64 percent of cell phones in 2020 will have this innovation when contrasted with 5 percent in 2017. Facial verification has been around for a couple of years now. Nonetheless, the progressions in man-made reasoning and profound learning calculations have presented new transformations in this innovation. Hence making it as accurate as possible which enhances the retailer and consumer marketing. So, we can understand that the impact of facial recognition in business sectors is impossible to overlook.

Chapter 2

Background Analyisis

This section provides a detailed study of the framework that is Neural Network and Convolutional Neural Network. Its development process is studied in great details starting from the core roots to the four CNN architectures that we are going to be using for conducting our research.

2.1 Neural Network

A neural network is a progression of algorithms that attempts to perceive fundamental connections in a lot of information through a cycle that copies the manner in which the human mind works. In this manner, a neural network refers to a system of neurons, either natural or artificial in nature. Neural networks can adjust to evolving input, so the network creates the most ideal result without redesigning the output criteria. A neural network works comparatively to the human cerebrum's neural organization. A "neuron" in a neural network is a numerical capacity that gathers and groups data as indicated by a particular design. The network looks somewhat like statistical methods, for example, curve fitting and regression analysis.

There are connections between the neurons from one layer to the layers on either side. A neural network has interconnected nodes where each node is called a Perceptron. It is a single neuron. These neurons are interconnected in different layers as shown on the figure above. The perceptron feeds the signal produced by a multiple linear regression into an activation function that may be nonlinear. In a MLP, perceptrons are orchestrated in interconnected layers. The first layer gathers input patterns. The output layer has classifications or output signals to which input patterns may map. The feedforward network estimates a function f^*. So when

$$y = f^*(x) \tag{2.1}$$

is used it is considered as a classifier which maps input x to a category of y. The feedforward network describes a mapping

$$y = f(x; \theta) \tag{2.2}$$

and studies the values of parameters θ [2]. The parameter is to be studied to provide the best function estimation.

2.2 Convolutional Neural Network

Convolutional Neural Networks are algorithms used in Facial Recognition with deep learning and these algorithms have been improved with the advancement of time by using hierarchical architecture which lets the network study facial representations with great detail. The usual drill is to first take an input image and then allocate important weights and biases to certain differentiable properties in the image. It has been seen through research that the pre-processing in a Convolutional Network has a relatively low requirement when compared to other classification algorithms.

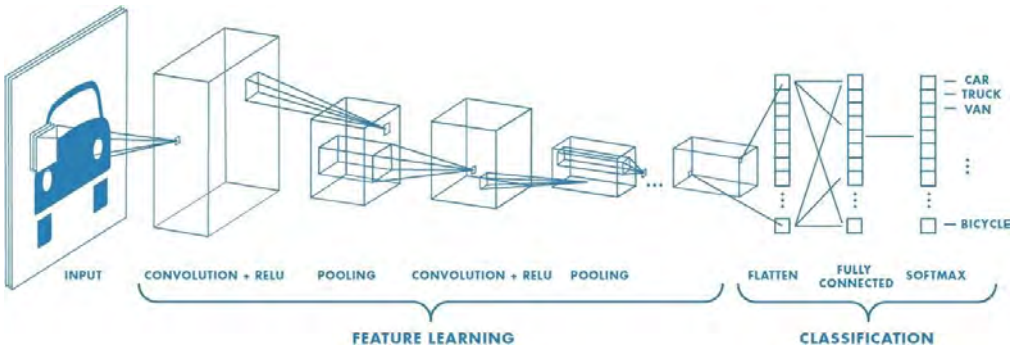


Figure 2.1: Sequence of CNN network in image recognition

Figure 2.1 [36] shows a typical sequence of a CNN network in image recognition. However, parallel to the increasing number of researchers that use this algorithm in face recognition due to its high success rate there is a birth of various CNN architectures that have such complex designs that it is hard to even visualize the model as a whole now. These architectures have evolved with time starting from 5 convolutional layers to 50, from plain ones to modules, from 2 towers to 32 and so on. These changes have made it possible for us to use these models with a very high accuracy rate for cases where face recognition and face verification can bring positive changes. As a result, it is safe to say that today the actual implementation of these models to help real world problems is exceptionally high. This is why it is very important to carefully analyze and keep track of all the changes that are happening in this area of technological development so that the best use can be made out of them.

2.3 How Convolutional Neural Networks work

In this section we have explained how each neuron in a convolutional neural network layer receives an input from different locations of the previous layer and ultimately gives an output.

2.3.1 Input Layer

An image is input in the input layer of a convolutional neural network. The image is then converted to a matrix which depends on the image dimensions and type. This matrix is an array of pixels. Images can exist in a wide number of color spaces such as gray-scale, RGB, HSV, CMYK, etc. A gray-scale image has only one

pixel, whereas an RGB image has 3 pixels due to its 3 different color planes – Red, Green and Blue. A convolutional layer is required to reduce the matrix size in order to produce a form of the image in which it is easier to process and where there is no loss of features [36].

2.3.2 Convolutional Layer

The matrix will go through the convolutional layer in the beginning. A smaller filter is present here that acts as a feature detector that moves from the top left of the input matrix and around it. The reason why this filter is used is because it multiplies the value with image pixel values. After the multiplication is done, the product value is then added for generation of a single numeric value in the end. This whole process is done while the filter moves to the right by 1 unit when stride is 1. Once the filter is done passing through all the positions, a feature map is derived which is a smaller matrix.

The input matrix has a dimension of $W_1 \times H_1 \times D_1$ where W_1 denotes the width, H_1 denotes the height and D_1 denotes the depth of the matrix. The final output from the convolution layer is $W_2 \times H_2 \times D_2$ and the calculation is as follows:

$$W_2 = \frac{W_1 - F + 2P}{S + 1} \tag{2.3}$$

$$H_2 = \frac{H_1 - F + 2P}{S + 1} \tag{2.4}$$

$$D_2 = K \tag{2.5}$$

where K denotes the number of filters, F denotes the size of filter, S denotes the size of the stride and P denotes the amount of zero padding [32].

This procedure causes the loss of a big amount of features. Important features aren't left rather features that are supposed to be used in upcoming layers are going to be kept. The obtained feature map from this procedure contains important features that we need to recognize and classify different types of images in the Convolutional Neural Network(CNN).

2.3.3 Non-Linearity

In a fully convolutional neural network, a non-linearity layer is composed of an activation function that maps the convolutional layer-generated feature map and produces the activation map as its result. The activation function is also an element-wise function from over range of the source, so the input and output parameters are similar. Where layer l is a non-linearity layer, the function density $Y_i^{(l-1)}$ is taken from a convolutional layer $(l-1)$ and the activation density $Y_i^{(l)}$ is generated [33]:

$$Y_i^{(l)} = f(Y_i^{(l-1)}) \tag{2.6}$$

with [33]

$$Y_i^{(l)} \, \epsilon \, \mathbb{R}^{m_1^{(l)} \times m_2^{(l)} \times m_2^{(l)}} \tag{2.7}$$

$$Y_i^{(l-1)} \epsilon \mathbb{R}^{m_1^{(l-1)} \times m_2^{(l-1)} \times m_2^{(l-1)}} \tag{2.8}$$

$$m_1^{(l)} = m_1^{(l-1)} \wedge m_2^{(l)} = m_2^{(l-1)} \wedge m_3^{(l)} = m_3^{(l-1)} \tag{2.9}$$

All this can be represented as $m_1^{(l-1)}$ number of 2-dimensional characteristic map data (generated by $m_1^{(l-1)}$ number of convolutional layer filters $(l-1)$) for each of the $m_2^{(l-1)} \times m_3^{(l-1)}$ size.

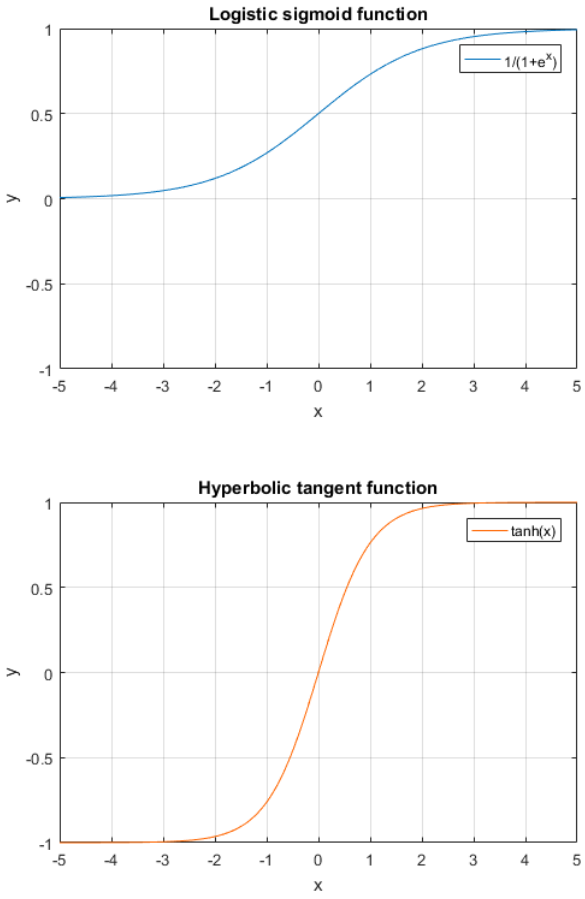


Figure 2.2: Logistic Sigmoid and Hyperbolic Tangent

The activation functionality is usually implemented as logistic (sigmoid) or hyperbolic tangent functions, analogous to multilayer perceptrons. More new studies, however, indicates that rectified linear units (ReLUs) are beneficial over conventional activation functions , especially in convolutional neural networks. Figure 2.3 [33] shows a logistic sigmoid function and a hyperbolic tangent function which both demonstrate an S shape.

2.3.4 Rectification Layer

In a convolutional neural network, a rectification layer defines an element-wise fractal dimension function on the source amount usually the amount of activation). If layer l is a rectification layer, the activation volume $Y_i^{(l-1)}$ is taken from a non-linearity layer $(l-1)$ and the rectified activation density $Y_i^{(l)}$ is created [33]:

$$Y_i^{(l)} = \left|Y_i^{(l-1)}\right| \tag{2.10}$$

It serves a significant role in the development of its convolutional neural network by removing the consequences of redundancy in successive layers. Especially whenever an ordinary pooling system is being used, the negative values inside the activation density are likely to reject the positive activations, greatly undermining the network's precision.

2.3.5 Pooling Layer

After the nonlinearity layer comes the pooling layer where pooling is done to images to obtain the most relevant and important features of the image and leave out irrelevant and unnecessary features. This helps in receiving a small sized feature map which has easier processing. This is like picking an activation in randomly based on a multinomial distribution during the time of training. There are many types of pooling like Max Pooling, Sum Pooling and Average Pooling. As a result, we can say that this layer basically takes the input of an image's feature maps that is received from the convolutional layer and generates a pooled feature map as an output [8].

2.3.6 Rectified Linear Units

A special deployment that incorporates non-linearity and rectification layers in convolutional neural networks is the rectified linear units. A linear unit rectified (i.e. lower limit at zero) is a regression coefficients piecewise specified as [33]:

$$Y_i^{(l)} = max(0, Y_i^{(l-1)}) \tag{2.11}$$

In convolutional neural networks, the rectified linear units have three key benefits relative to conventional logistic or hyperbolic tangent activation functions:

1. Rectified linear units essentially spread the gradient and thereby reduce the risk of a gradient vanishing problem that is widespread in deep convolutional neural architectures.

2. Rectified linear units limit low traits to zero, thus fixing the issue of redundancy and resulting in a much smaller amount of activation at its outcome. For various factors, sparsity is beneficial, but primarily offers robustness for minor input variations including such interference.

3. Rectified linear units comprises only specific computational operations (primarily equivalences) and are thus far more accurate in integrating them in convolutional neural networks.

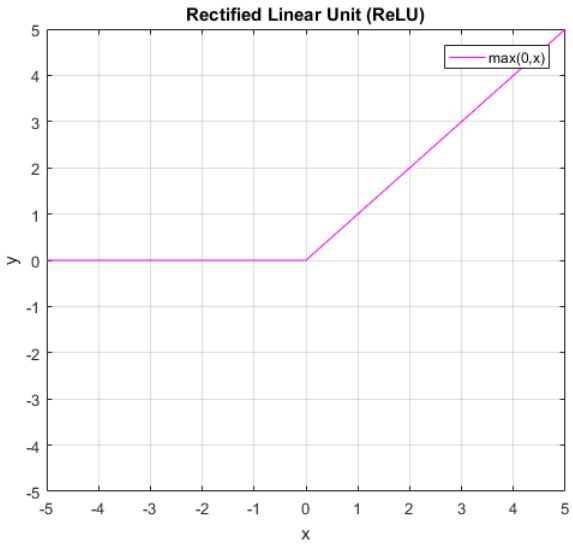


Figure 2.3: ReLU

Figure 2.3 [33] shows a graph of a ReLU. Almost all of the prospective architectures of convolutional neural networks use rectified linear unit layers (or their variants such as noisy or leaky ReLUs) as their non-linearity layers rather than just conventional non-linearity and rectification layers due to several benefits and efficiency.

2.3.7 Fully Connected Layer

In a convolutional network, the fully connected layers are basically a multilayer perceptron (usually a two or three layer MLP) that attempts to translate the activation density of $m_1^{(l-1)} \times m_2^{(l-1)} \times m_3^{(l-1)}$ from the composition of multiple former levels into a spectrum of class probability [33]. Thus, the multilayer perceptron output unit may have $m_1^{(l-i)}$ outputs, i.e. output neurons where i represent the number of levels in the perceptron multilayer.
Where $l-1$ is a fully connected layer, $y_i^{(l)}$ and $z_i^{(l)}$ [33] are generated:

$$y_i^{(l)} = f(z_i^{(l)}) \quad with \quad z_i^{(l)} = \sum_{j=1}^{m_1^{l-1}} w^(l)_{i,j} y_i^{(l-1)} \tag{2.12}$$

The purpose of the fully connected structure is just to modify the synaptic weights $w_{i,j}^{(l)}$ to establish a deterministic portrayal of the probability of each group based on the activation maps created by the computation of layers of convolution, nonlinearity, reassessment and pooling. Whereas the only difference being its input layer, individual completely connected layers work indistinguishably mostly with layers of the multilayer perceptron. Figure 2.4 [33] shows the structure of a Fully Connected Layer.

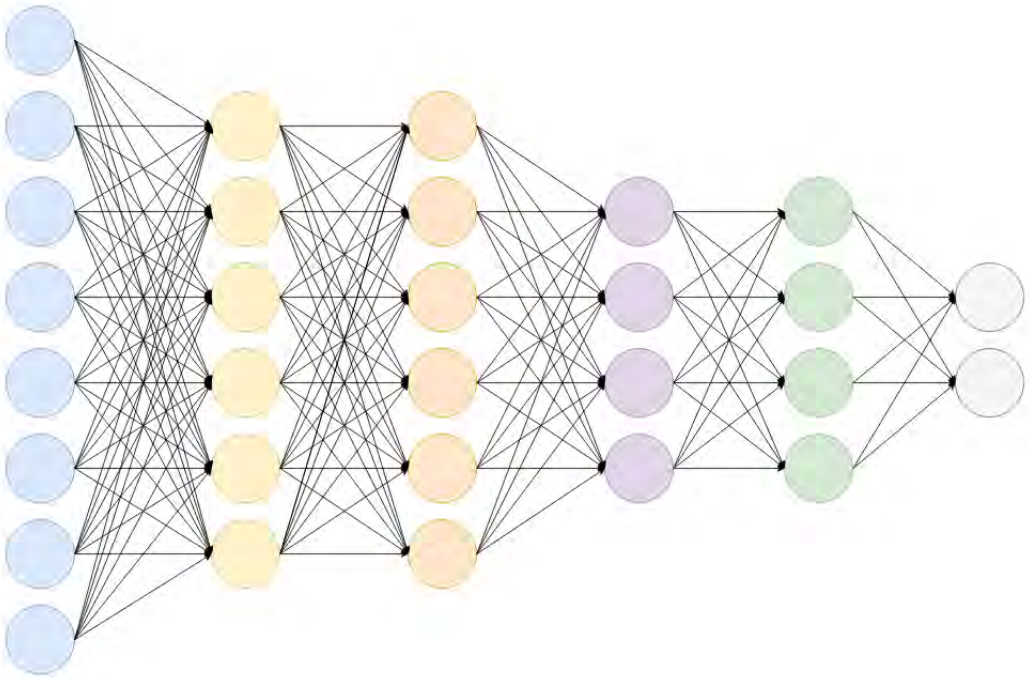


Figure 2.4: Fully Connected Layers in CNN

2.4 Related Works

In recent years, it has been discovered that deep learning techniques have shown to have the best results in face recognition, identification, verification as well as age estimation. CNN is an extraordinary sort of Neural System which has appeared commendable execution on a few events associated with the area of Computer Perception and dealing with Pictures. It's application areas range from Image Classification Segmentation, Object Detection, Video Processing to Natural Language Processing, and Speech Recognition [29]. The natural scientific categorization of CNN architectures which are divided into seven different categories equal to the spatial exploitation, depth, multi-path, width, feature-map exploitation, channel boosting, and attention are further exploited into understanding their factors, challenges and utilizations. This study too gives an knowledge into the essential construction of CNN as well as its authentic point of view, showing distinctive periods of CNN that follow back from its beginning to its most recent improvements and accomplishments [29]. An experimentation was done to bring about enhanced age classification scores with ResNet-18 and ResNet-34 which resulted in proving that ResNet-18 and ResNet-34 are impeccable frameworks to deal with intricate classification issue [5]. The techniques include using CNN for facial attribute extraction from images in the various datasets that are available online. Using images from MORPH II database, a dataset is created and a four step approach is used such as preprocessing image, eliciting feature using PCA, devaluating feature using PCA and finally Deep Belief Network based classification process for face detection and age estimation [21]. The network is assembled in the initial layer by unsupervised training and the plot considered for the work is first-class lineaments that can be studied through highlights of minor classes whereas the correct ones can be forged for pattern classification at the conclusion and appropriately, profound models have the potential to contain more abstracts and complex highlights in higher layers and eventually forms a DBN [21].

Another method proposed provides an architecture that jointly operates cross-age amalgamation of the face and acknowledgment in a shared boosting way and accomplishes nonstop face maturing with photo-prudent and identity-sustaining properties [1]. The experiments are conducted on multiple demographic groups where subjects may belong to one or more demographic groups and images of equal or similar proportions of demographics were used for each class and preprocessing, feature extraction, dimensionality reduction, and classification are performed on each of the groups and the results were compared with COTS and eventually this performance of the method was evaluated using L1 distances measuring the similarities between pairs of feature vectors and PCA was used to improve performance [1]. Additionally some methods propose a modern extensive Cross-Age Face Recognition criterion dataset to improve the age invariable confront acknowledgement investigation of the most well known unconstrained datasets IJB-C [27]. AIM is constructed from an auto-encoder based GAN and incorporates a detached RLN and a FSN for age-invariable face acceptance [27]. RLN comprises of an encoder, which encodes distinguishable highlights with the deviation from multi-age realms to handling cross-entropy regularization with a label steady approach to restrain cross-age portrayal with inconclusive detachability, and a discriminator, which cover dual agents that uniformly distribute the features to smooth the age transformation while storing the different

facial features [27]. Another method suggests a DEX pipeline which includes sturdy face calibration, VGG-16 architecture and allotment for determining the real and apparent age [22]. The paper proposes a deep learning clarification to age assessment from an image containing a single face, using the IMDB-WIKI dataset [22]. The important consideration in the elucidation gives profound studied models from vast data, sturdy face calibration and conventional result arrangement for age regression and the approaches are validated on specified criterions and attains results for both legitimate and possible age estimation [22]. The paper uses deep learning and extraction using CNN characteristics of deep learning [22], [8].

Regularization methods include quadratic regression, Gaussian process and SVR which can be also used to detect age. But rank-based methods are practiced for age detection where the age-axis strategy is applied to determine age instead of classification and regression methods [19]. The paper uses aspects of CNN field of image application through deep learning method to revive attributes and to choose a factor analysis model to takeout robust attributes [19]. After consecutive analysis of rank based age estimation learning methods, a divide and rule face estimator is put forward and its performance is proved better than SVM and SRM [19]. A solution for this consists of a CNN that is collectively trained in a MTL structure where the criteria of minor layers of CNN are shared with all the functions, can deal with face detection, face verification and recognition as well as age estimation [13], [14]. The paper exhibits multi-use algorithms for face detection, face alignment, pose, gender, smile and age detection at the same time using single deep CNN [13]. The method as mentioned builds a structure that assigns the common specifications of CNN making different domains and tasks work together thus showing that it accomplishes advanced results [13]. VGG-16 model is used due to its high performance in most deep learning programming frameworks and the patch CNN calculates the higher resolution patches based on their relevance foreseen by the consideration grid to reduce computational requirements for its architecture they reused the first convolutional layers of the consideration CNN, then a MLP that assimilates the knowledge collected from both CNNs and operates the final allotment and tested their model by using the Adience dataset that abides of 26.5K images dispersed in eight age groups with corresponding genders [14]. Most importantly, some researches use the periocular region as a subject to aging by using extraction features of CNN, a deep learning method and classification characteristics of deep learning [20]. The extractor module is categorized in three layers, convolutional layers, pooling layers and at last fully connected layers and the fully connected layer which is output layer gives the recognition result is the last layer of their CNN architecture, thus the layer with number of nodes represent the output as probabilistic prediction to the class labels [20]. Current methods for age recognition include generative and discriminative procedures. But demographic method concludes the development of faces at diverse ages, a pretrained multi-layer CNN can be used to automatically select facial attributes that are particular at different ages and also by using ReLu, pooling and FC layers through a set-established approach to find the apparent and real age of a subject [8]. The images for a subject, with pictures of various ages, as a individual set are as to the set of images that belongs to new subjects thus a set-established approach to the face recognition susceptible to aging and about 2.6 million images from the Web as a dataset are used to train the VGG-Face [8]. The set-based identification experiment is done across time lapse and half of the experi-

ments conducted are comparable to the younger images in the gallery and other half to the older images in the gallery for the first experiment, thus the image pairs are built that contain a subject photograph of a test dataset and another photograph of the gallery and so when all the images are matched to one corresponding subject, the pairs are labelled as positive otherwise negative [8]. The gender and the age of a person can also be identified from video records with the implementation of CNN to process the single video frames through binary classification and regression problems using algorithms like Arithmetic Mean or the Sum Rule, Product Rule and Simple voting [18].

Since CNN consumes a lot of memory and has computational complexity, a mobile system will not be able to withstand the load, thus a specially trained MobileNet-based CNN to identify concurrent gender and age [18]. The aggregation decision for each video frame with Dempster-Shafer theory is used to generate a set of classifiers by firstly making decision patterns for each class with the help of auxiliary training set and then using the Dempster-Shafer belief degree to evaluate the final decision [18]. New progresses in profound learning have appeared that CNNs are competent of evaluating an self-assertive compound work, thus an independent blended CNN to fuse the hyper features is constructed which includes AlexNet model and ResNet-101 model, two narrated CNN architectures that perform confront discovery, landmark localization, posture assessment and gender identification by blending the intermediate layers of the networks and IRP and L-NMS, two post-processive methods which use the multi-task data gotten from the CNN to move forward the in general execution and finally execution of R-CNN-based ways for individual tasks and as well as the multi-task approach without intermediate layer blend is studied [12]. The features are fused using hierarchical element wise addition and combining halfway layers progresses the execution for structure subordinate assignments of posture estimation and landmark localization, as the highlights ended up invariant to geometry in more profound layers of CNN [12].

The researchers contemplated utilizing Convolutional Neural Networks in the recognition of the age and the gender of a subject via videos of the said subjects face and this is to be used in future mobile video analytics systems [25]. A multi yield expansion of the MobileNet is uniquely developed and it is pre-skilled to perform confront acknowledgment utilizing the VGGFace2 dataset. Extra layers of the proposed arrangement are re-tuned for age and gender perception on Adience and IMDB-Wiki datasets to determine an individual's age and gender [25]. This paper proposes 2 different models of CNN architectures that compares its output with another model of the LeNet-5 architecture and after each convolutional layer batch normalization is performed on the images along with max-pooling at each pooling layer to extract important facial features which are sent to the fully-connected layers that gives out the final output [17]. The dataset is partitioned into preparing and testing data and the CNN model is adjusted for efficient classification of the facial skin images by implementing RMSE method, for both the preparing and authorization set, by utilizing the adaptive Adam optimizer (Adaptive Moment Estimation) to alter the parameters amid the preparing stage and AdaGrad and PMSProp are also implemented in this method and the initialization in this paper of parameters is consummated by Gaussian distribution [17].

A deep CNN exemplary is prepared for confront acknowledgment on an awfully huge database and the trained CNN extracts the facial features from the input

images and performs age estimation using the extracted features [11]. It uses VGG-Face model, which consists of 11 layers, where 8 of them are convolutional layers and 3 are fully-connected layers and a rectification layer comes after each convolutional layer and a max pool layer is operated at the end of each convolutional block which then is passed on to the fully-connected layers where the to begin with two fully-connected layers are taken after by a dropout layer and stochastic gradient descent method with mini-batches is used to optimize the variables of the fully-connected layers to estimate the age of an individual [11]. Another paper outlines a survey of face recognition and age estimation is provided along with the few challenges that are faced in this area and also gives a scene mapping based on coordination into a basic and coherent scientific categorization which is emphasized by comparing objectives of numerous approaches [16]. It provides some information about recognition approaches like stating that the performance of these face techniques vary from one data set to another and displays the gaps for future reference [16]. This paper recognizes face recognition as a very important application in several areas like for security purposes, law authorization operations and attendance systems and also mention that the main usage of facial features is in looking for a lost child, thus outlining a overview of confront acknowledgment and age evaluation is provided along with the few challenges that are faced in this area [16]. On the other hand, classifying gender and age from an image for the autonomous detection of unusual human activity is done using transfer learning which is a new deep learning method and this method is for image classification is based on a pre-skilled profound demonstration of component eradication and depiction followed by a Support Vector Machine classifier [15]. AlexNet was used as a component eradication to separate the convenient components of the image and then used the SVM classifier for image perception and for anomaly detection they have used the EDA which is a method that uses recursive calculation which helped them with the efficiency of their model because it doesn't keep huge sums of information within the recollection, rather it reuses and updates existing values only and they also compare the classification process of using Haar-like features and pre prepared profound learning organize highlights [15]. New techniques have been introduced for age estimation known as ranking technique and this paper discusses about the CNN-based ranking approach which are pre-trained and fine-tuned and thus the binary outputs generated give the final age estimation proving that the approach presented here gives a smaller estimation error because of the ordinal relations between age is considered [6]. The paper introduces a CNN based framework that involves a sequence of basic CNNs with age figures and their binary outputs are combined to give a final estimated age [6]. The procedure also shows that ranking CNN gives a smaller estimation error compared to other approaches on benchmark datasets [6].

It is analysed that distinctive methods appear with diverse precision rates. Here in Figure 4, the blue bar appears the exactness percentage of the ATT database which is superior than the exactness of the IFD database defined by the red bar for distinctive methods [9].The essence of photos in the database is one of the reasons. In this case SVM classifiers can be considered the strongest one. So we can evaluate that there are many methods for facial recognition and feature extraction techniques. Facial features require the detection and localization of certain points which are addressed as the most striving facial features along with distortion posture, ageing, lighting and partial occlusion. Research indicates that facial expressions

alter with regard to age; thus, in face recognition, they can not be permanently modeled. But Neural Network classifier has proven to be the most suitable way of recognition as it provides with features that include Versatile learning, Self- organization, Blame-resilience, Huge Execution and precision, capacity to infer meaning from loose information arrangement to genuine world issues with comparatively less space and time complexity, etc [9].

All methods using deep convolutional neural networks suggest image pre-processing through CNN layers and extraction of the features which is then classified and predicted through various algorithms as suggested and also verified under different benchmarks. It is additionally demonstrated that profound learning strategies have outflanked conventional machine learning approaches in each conceivable way. Training of such major networks, nevertheless, increases the computational effort and comparatively becomes challenging as the networks grow [23]. Thus extensive experiments are done to show that Gabor CNNs are significantly improved computation when using Gabor in comparison with normal CNNs as they exhibited progress in energy demand during training, and also reduction of training time, storing and shared memory energy for inconsistent loss in classification accuracy [23]. But by making alterations in the weights concurring to the target, CNN learns through backpropagation calculation during training. The convoluted, progressive design of the profound CNN, gives it the capacity to extricate features ranging from to high-level features [29]. Subsequently, in acknowledging the comprising tons of picture groups, deep CNNs have appeared to have considerable execution enhancement over routine vision-based models. This proves that CNN architectures can really boost the representational potential of a CNN to increase further use of CNN in tasks of entity classification and segmentation.In expansion to its utilization as a guided study of component, the capability of profound CNNs can moreover be overworked to extricate valuable manifestations from a vast scale of unidentifiable information [29].

2.5 Supervised Learning

Supervised learning is the type of learning in which a machine is trained using a training set of labelled data. Once the training of the machine is over, it is given some new data that it doesn't know and it conducts an analysis of the training set, producing the correct outcome of the labelled data. This technique helps us collect, produce and optimize data output based on previous learnings. Once the system is trained using the training data, it will be tested using the testing set. Since the system has learned features of the object from the image, it will be able to classify the testing data based on what it has previously learned. Supervised learning can be categorized in two ways:

1. Classification: This means that the output of a system will be grouped inside a particular class. It is called a binary classification when the algorithm labels the input into two distinct classes and when it is selecting more than two classes it is known as multiclass classification.

2. Regression: This technique is used when the output of the system is a real value like "dollars" or "weight". It is used in probabilistic interpretations as it detects continuous data points and can be used to prevent overfitting by regularization. However, it can't be used for complex problems as it is not that flexible.

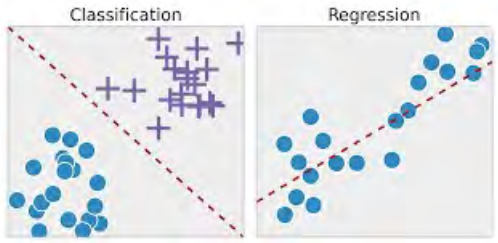


Figure 2.5: Classification and Regression

Figure 2.5 [37] displays how classification and Regression works. For face verification, we will have to use classification supervised learning because our system has to define a set of target classes and be trained with a set of labelled images that already has the correct answer labelled on them. We will be using this approach for our research as we will need to confirm if the given input image matches the second and third input image or not. For example, if input image 1 matches either input image 2 or 3 and if it matches both, the system will show true/false to indicate which is the case. Moreover, classification trees are known to perform very well in practice.

2.6 Market Research

The discoverers of facial-recognition software, permits advertisers to use facial acknowledgment inside their promoting methodology. They can also use marketing research to learn about their company's brands, reputation. Market research through the process of collecting, dissecting, and understanding the underlying information about a market, product or service could give the organization an edge over their peers.

Market research helps companies collect, record present and past data on the performances by targeting selective age groups for their advertisement purposes to fulfill their marketing goals respectively.

2.7 Architectures

In this section we have provided a detailed description of the four pre-trained Convolutional Neural Network models that we have used to conduct our experiment and analysis. It includes information about the layers that are present in these architectures, their loss functions and activation functions. We also provide information about how they were developed starting from one version and progressing into another more recent one.

2.7.1 VGG-16 and VGG-19

VGG network brings and retains certain concepts from all its previous efforts of Convolutional Neural Networks and builds all over them and incorporates profound neural layers of convolution to enhance performance. The VGG network is defined by its efficiency, through the use of a growing extent of just 3×3 convolution layers built on top of one another. Density size devaluation is managed by max pooling. A series of convolutional layers, which has various depth in various configurations, is preceded by three Fully-Connected layers in which the first two abide 4096 channels each [8], the third operates 1000-way ILSVRC identification and thus includes 1000 channels for each class. The soft-max layer is the outer layer. On all networks, the layout of the fully linked layers is just the same.

VGG 16 is a 16-layer layout including a set of convolutional filters, a pooling layer and a completely connected layer at the end. The overall number of convolutionary and completely associated layers is 16 in the VGG-16 architecture stacked with 3×3 and 2×2 max pooling layers. The relu-activation operation is implemented among these levels as all hidden layers are configured with non-linearity in rectification [8],[29], as follows:

$$f(x) = \max(0,\mathrm{x}) \tag{2.13}$$

$$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases} \tag{2.14}$$

VGG-19 is a 19-layer layout similar to the VGG-16 shown in the figure below along with a couple of more convolutional layers with the same 3×3 convolution layers built on top of one another anticipated by three FC layers and a softmax layer. Rectified linear unit is also implemented in VGG-19 to add non-linearity to enhance the categorization of the structure and to boost computational time as tanh

or sigmoid functions were used by older designs, which turned much superior than those [29].

On each step of multiplying through each stack of convolution layers, the amount of filter used is nearly doubled and that is another basic concept used to build this network architecture. Using limited filters provides an added advantage of low computational complexity by limiting the number of parameters [8], [29]. These results of the study set a new pattern for CNN to work with narrower filters in analysis. By putting 1×1 convolutions in between fully connected layers, VGG determines the configuration of a system, which also discovers a linear combination of the extracted features-maps. For both picture categorization and interpretation issues, VGG demonstrated better performance [29]. Figure 2.6 [31] demonstrates the structures of VGG16 and VGG19 architectures.

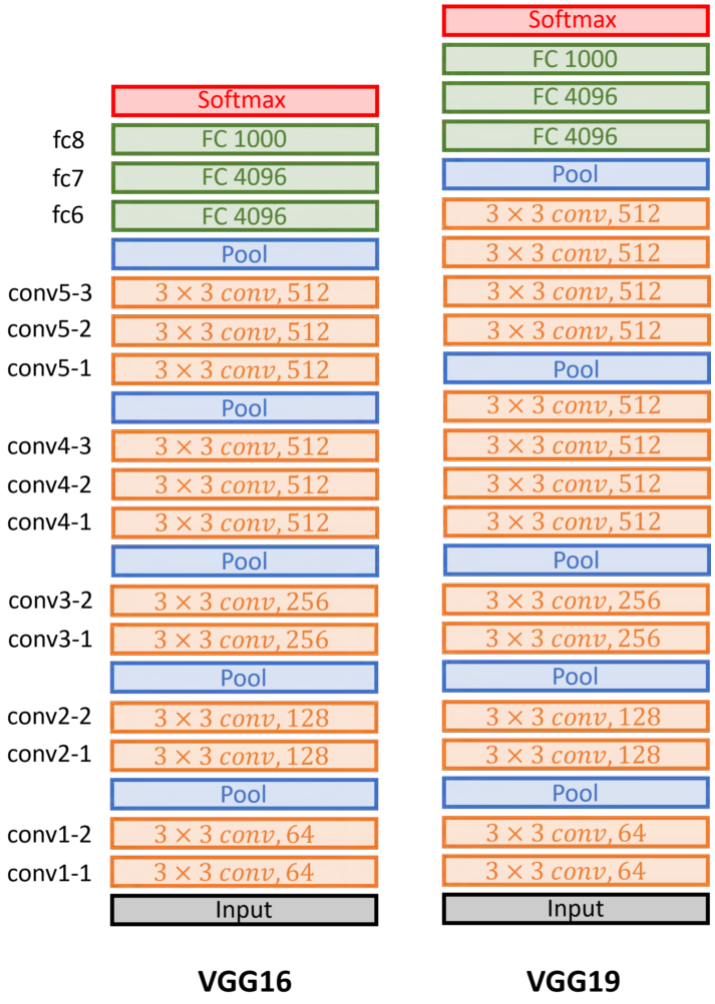


Figure 2.6: Layers of VGG-16 and VGG-19

2.7.2 InceptionNets

InceptionNet also known as GoogleNet was the winner of "ImageNet Large Scale Visual Recognition Challenge" back in 2014 [29], [4]. This architecture was composed of 22 layers [4] with the main intention of designing a network within a network topology. This topology specifically increases the computational power of a neural network. In the InceptionNet architecture there are convolution layers parallel to each other using different filters. After the convolution there is concatenation followed by capturing features at 1×1, 3×3 and 5×5 [4]. Finally, the clustering is done on the features. An addition of 1×1 convolutional layer followed by a rectified linear activation [3], [4] satisfied their goal of increasing the computational power of the architecture. The rectified linear activation function reduces the depth by preserving the spatial dimensions so that the overall network's dimensions don't increase exponentially. This additional 1×1 convolution reduces the number of operations greatly and the reason for this is that the input matrix would first be convoluted by the 1×1 convolutional layer with a filter and then be convoluted once more by the 5×5 convolution layer [4]. This causes the number of operations to be highly reduced. GoogleNet (Inception v1) was built using the concept of this dimension reduced inception module. It has 9 inception modules that are linearly stacked. It is 22 layers deep and has 5 pooling layers which makes it 27 layers in total. Figure 2.7 [34] shows the structure of the Inception module. Figure 2.8 [34] shows the structure of the Inception v2 module.

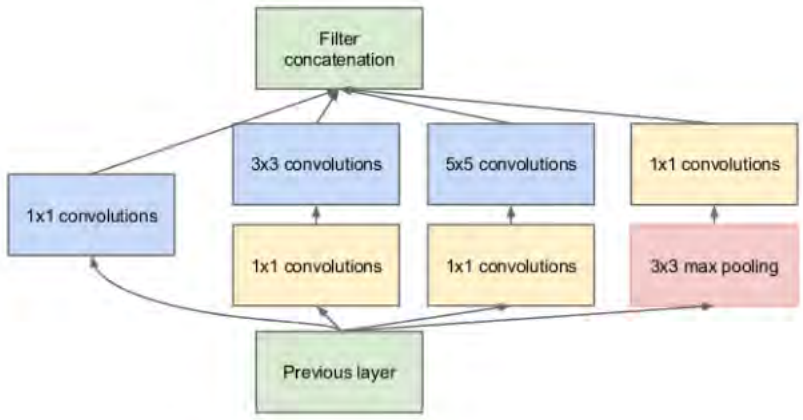


Figure 2.7: The Inception Module

In this architecture, an input image goes through different sizes of convolutions and max-pooling. Afterwards different types of features are extracted. At the end of its last inception module, a global average pooling is used and in the fully connected layers the most favourable network topology can be created by examining the activation correlation statistics in the last layer and clustering neurons that have the highest correlated output. Finally, the input for the next module is prepared by concatenating all feature maps at different paths. However, this architecture has some lackings because its 5×5 convolution decreases the input dimensions [4] largely which in turn reduces the accuracy of the neural network. Moreover, a decrease in complexity was seen as well when bigger convolutions like 5×5 were used compared to 3×3. It was claimed earlier that the auxiliary classifier could help reduce the

vanishing gradient problem in deep networks but was later confirmed that this was not the case.

Inception v2 was introduced in 2015 with necessary changes that covered the failings of Inception v1. A few architectural changes were made like the replacement of the 5×5 convolution with two 3×3 convolutions. This change decreased the computational time which meant that the computational speed was increased. Moreover, a 5×5 convolution costs more than a 3×3 convolution so it was better to use two 3×3 convolutions as it increased the performance of the architecture greatly.

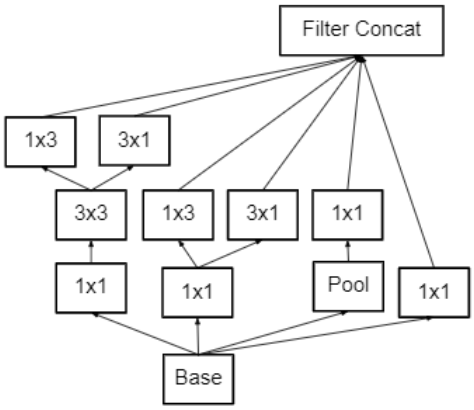


Figure 2.8: Inception v2 Module

Batch normalization was introduced in this version [3]. Changes also included the conversion of $N \times N$ factorization into $1 \times N$ and $N \times 1$ factorizations. This further reduces the cost in terms of computational complexity in comparison to 3x3. To remove the representational bottleneck [4], this module was made wider instead of making it deeper so the excessive reduction in dimensions and loss of information was avoided.

Unfortunately, this version of the architecture wasn't as efficient either and it was found that the auxiliary classifiers weren't that helpful. At least not until the end of the training process was it reached when the accuracy rates were near the saturation point.

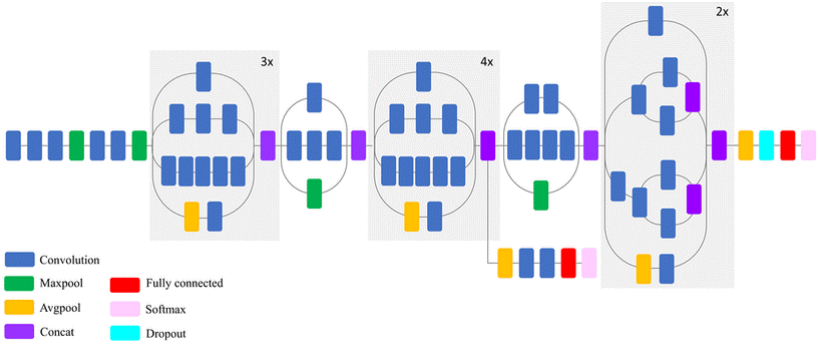


Figure 2.9: Schematic diagram of Inception v3

It wasn't long before Inception v3 came into light with further improvements like additional factorization [3]. It is a 42-layer deep learning network with fewer parameters. On top of the last 17×17 layer, the number of auxiliary classifiers was decreased down to 1 from 2 classifiers in order for it to act as a regularization so that the module generalizes better. It further improves performance by factorizing the traditional 7×7 convolution into three 3×3 convolutions. Figure 2.9 [38] shows the structure od Inception v3. This reduces the overfitting problem. Furthermore, the label smoothing regularization method is used to prevent a very large logit value [3].

$$new\_labels = (1 - \epsilon) * one\_hot\_labels + \epsilon/K \tag{2.15}$$

where ϵ, a hyperparameter is 0.1 and K, the number of classes is 1000.

Finally, we have the Inception v4 architecture. It is a simpler and a uniform architecture and more inception modules when compared to Inception v3 [3]. This architecture doesn't have any residual connections and it is a purely Inception variant. After Inception v3, it was noticed that some modules could be more simplified and making them more uniform could make the network's performance even better. Figure 2.10 [34] demonstrates the evolution of Inception v4 from GoogleNet.

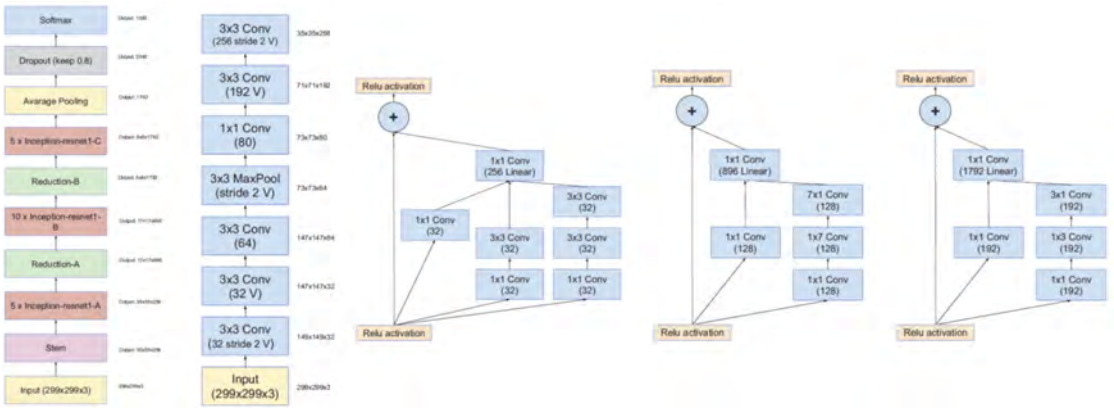


Figure 2.10: Inception v4 evolved from GoogleNet

The layers towards the beginning are modified to make it more uniform. "Reduction Blocks" are used in this version which changes both the width and height of the grid.

It is discussed that for training deep learning models, residual connections are very essential and as inception architectures are very deep, it is normal to use residual connections in the place of the filter concatenation stage of the Inception architecture [3]. A hybrid inception module was created by using the architecture of ResNet due to it's high performance and Inception which was named the Inception-Resnet. It has two sub versions which are Inception Resnet v1 and Inception Resnet v2. The difference between these two versions is very small and this is:

- The computational cost of Inception Resnet v1 is similar to Inception v3.
- The computational cost of Inception Resnet v2 is similar to Inception v4.

The modules and the reduction blocks of these two sub versions are the same but their hyper-parameter and stem structure is different. The main idea was to add the output from convolution layers to the input by using residual connections.

Figure 2.11 [34] shows how pooling layers are replaced by residual connection and a convolutional layer is added.

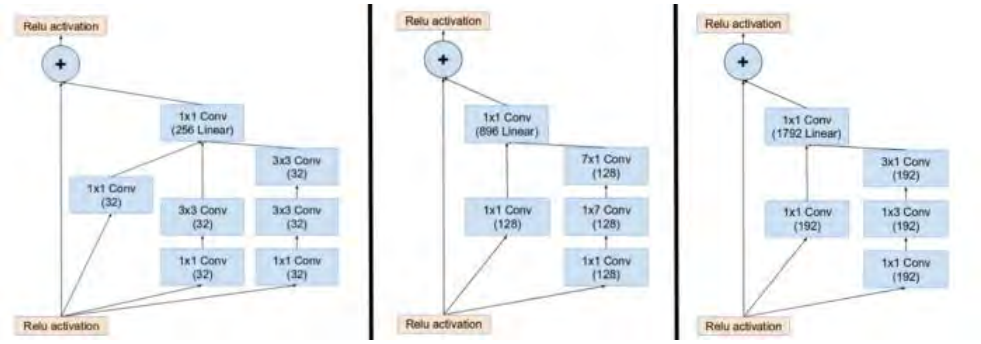


Figure 2.11: Replacement of pooling layers by residual connection and addition of 1×1 convolution

This was specifically done first by adding a 1×1 convolution following the original convolutions for matching depth sizes as the input and output after convolution needed to be of the same dimensions otherwise the residual connection would not work [3]. Secondly, by replacing the pooling layers of the main inception modules with residual layers. However, the pooling operations can still be seen in the reduction blocks.

Furthermore, the residual activation value was scaled from 0.1 to 0.3 to increase the stability [3] and prevent the network from "dying" when the number of filters became more than 1000. It was seen after implementation that Inception-ResNet architecture could achieve higher accuracies at a lower epoch. Figure 2.12 [34] shows the final network layout for Inception-Resnet.

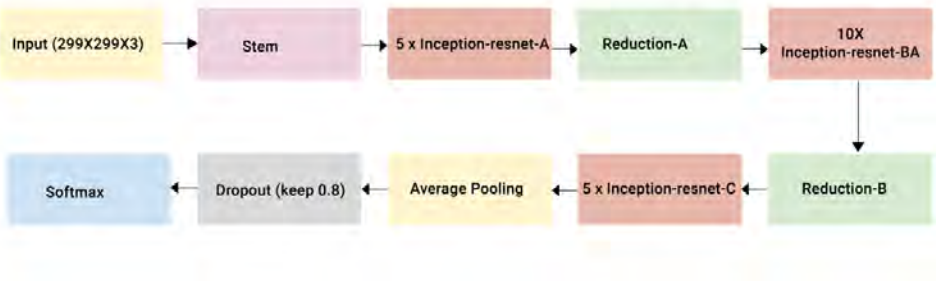


Figure 2.12: The layout of Inception-ResNet

2.7.3 ResNet-50

ResNet-50 is a residual learning framework for convolutional neural networks that is 50 layers deep. This framework comprises 5 stages, each with a convolution and Identity block where 3 convolution layers are in each convolution and Identity block. In the field of computer vision, such as object localisation and detection, image classification, this layout is implemented and uses skip relation to apply the output from the previous layer to the next layer, which eases the issue of the vanishing gradient.

By presenting the concept of enduring learning in CNNs, ResNet recast the CNN architecture and articulated a productive strategy for adapting profound systems that can be correlated to Interstate Systems and are also set under the CNNs focused on Multi-Path. ResNet, which was 20 and 8 times more profound than AlexNet and VGG, individually, appeared to have less computational complexity than already proposed systems [29]. With regards to the plain network, the residual networks are more safe to the debasement which resulted as one of the advantages of the residual network [5].

When a generalized link is cut short by a transient link, the total data measured before a current connection is passed on to the rest of the network. In this case, the alternative connections provide the chance for the network to utilize the new time saving connections rather than the regular ones [5]. The following equations shows the mathematical representation of ResNet [29] :

$$F^k_{m+1} = g_c(F^k_{l\to m}, k_{l\to m}) + F^k_l \qquad , m \geq l \tag{2.16}$$

$$F^k_{m+1} = g_a(F^k_{m+1}) \tag{2.17}$$

$$g_c(F^k_{l\to m}, k_{l\to m}) = F^k_{m+1} - F^k_l \tag{2.18}$$

Here, $g_c(F^k_{l\to m}, k_{l\to m})$ is a reconstructed signal, F^k_l is an input of the l^{th} layer. The $k_{l\to m})$ in equation(1) demonstrates the k^{th} processing unit (kernel), $l \to m$ represents the residual block that contains one or more hidden layers. The actual F^k_l is added to $g_c(F^k_{l\to m}, k_{l\to m})$ which is the transformed signal to provide the output F^k_{m+1} and this is later designated to the upcoming layer after applying the activation function $g_a(F^k_{m+1})$. $F^k_{m+1} - F^k_l$ on the other hand , gives the residual value, which is then used to refine weights dependent on context. This exemplifies that the leftover functions are convenient to configure and can gain exactness with significantly expanded depth [29]. Figure 2.13 [29] shows the residual construct of ResNet simple function.

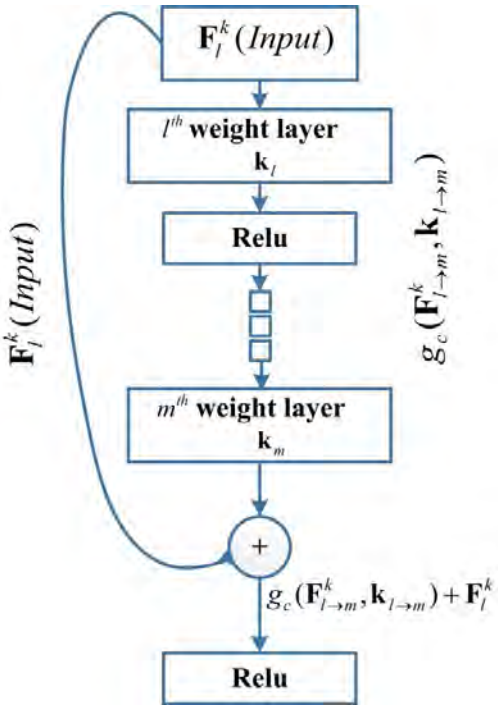


Figure 2.13: Residual construct here as ResNet simple functional element

ResNet presented easy route associations inside layers to empower cross-layer networks; in any case, these associations are content-independent with no parameter relative to the doors of Interstate Systems. Be that as it may, in ResNet, leftover data is continuously passed, and character alternate routes are never closed. Thus the enduring links, shortcut connections, accelerate the meeting of deep networks which eventually gives ResNet the capacity to avert the vanishing gradient problems [29].

2.7.4 Xception

Xception is an architecture of CNN which is built upon depthwise separable convolutions. This architecture has 36 convolutional layers that have been arranged into 14 modules, where each module has a linear stack of layers. This type of layering makes the module convenient for defining and modifying.

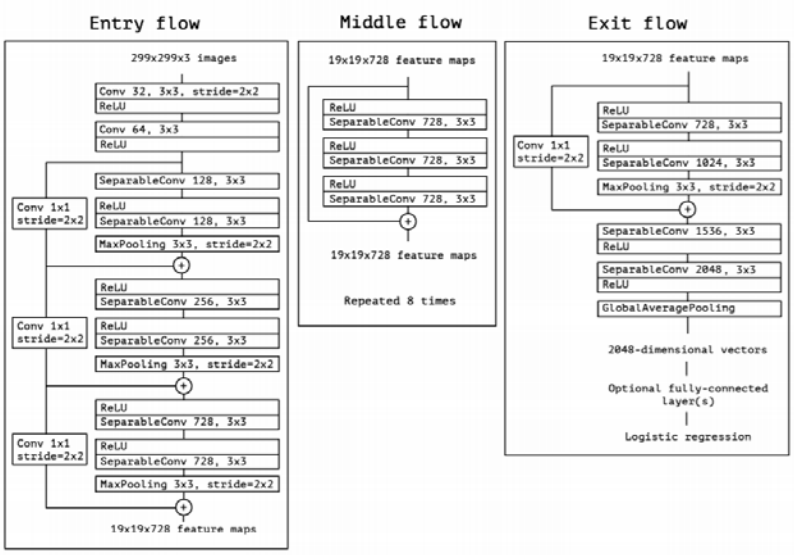


Figure 2.14: Overview of the Xception architecture

In Figure 2.14 [7] it shows that at first, the data enters the Entry flow, after which it goes through the Middle flow where the process is repeated 8 times. Finally, it exits through the Exit flow. Here, all the separable convolutional layers have a depth multiplier of 1. As per the figure, SeparableConv resembles a depthwise separable convolution. All the layers mentioned are followed by a process called batch normalization [7].

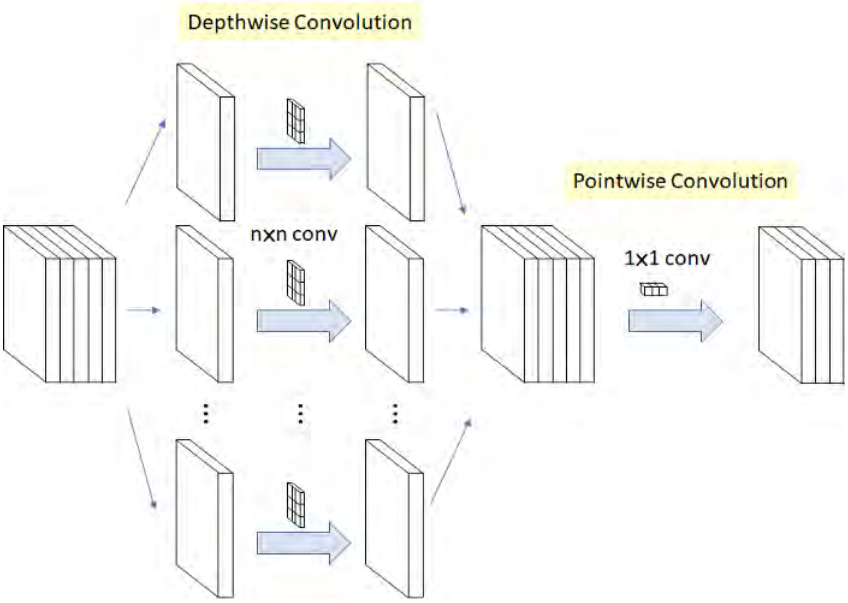


Figure 2.15: Depthwise Convolution

Normally the depthwise separable convolution is followed by a pointwise convolution, as shown in Figure 2.15 [35] Here, when compared to normal convolution, there is no for convolution to be performed on all channels, resulting in the total number of connections being lower and the model being a lot lighter in the end

1. Depthwise convolution is the channel-wise nn spatial convolution. In the figure above, there are 5 channels, then there is a 5 nn spatial convolution.

2. Pointwise convolution is the 1×1 convolution to change the dimension.

The mathematical formulations are as follows [10]:

$$Conv(W, y)_{(i,j)} = \sum_{k,l,m}^{K,L,M} W_{(k,l,m)} \cdot Y_{(i+k,j+l,m)} \tag{2.19}$$

$$PointwiseConv(W, y)_{(i,j)} = \sum_{m}^{M} W_m \cdot Y_{(i,j,m)} \tag{2.20}$$

$$DepthwiseConv(W, y)_{(i,j)} = \sum_{k,l}^{K,L} W_{(k,l)} \odot Y_{(i+k,j+l)} \tag{2.21}$$

$$SepConv(W_p, W_{d,Y})_{(i,j)} = PointwiseConv_{(i,j)}(W_p, DepthwiseConv_{(i,j)}(W_{d,Y})) \tag{2.22}$$

Accordingly, the actual thought behind depthwise divisible convolutions is to replace the element learning worked by ordinary convolutions over a joint "space-cross-channels domain" into two simpler steps.

However, In a modified depthwise separable convolution, depthwise separable convolution occurs after a pointwise convolution, as shown in Figure 2.16 [35]

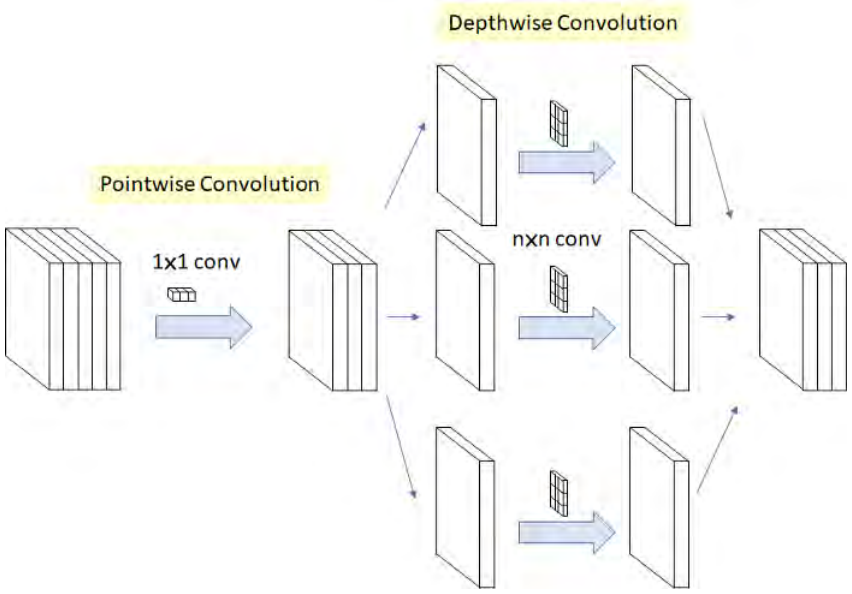


Figure 2.16: Modified Depthwise Convolution

1. Perfoms 1×1 convolution first then channel-wise spatial convolution. This is professed to be insignificant in light of the fact that when it is utilized in stacked settings, there are just little contrasts shown up toward the start and toward the end of all the chained inception modules.

2. The modified depthwise separable convolution has no intermediate ReLU non linearity.

These modified depthwise layers make the Xception architecture faster than Inception. As the name suggests that Xception is Extreme Inception. Now Inception and Xception uses similar parameters but Xception is easier to use and faster to edit as regular convolution layers.

Chapter 3

Model Training

MORPH-II, FG-NET and CACD datasets are collected and undergone Image Pre-processing such as Face Detection and ALignment, Landmark Localization, Resizing, etc. Afterwards through Mini-Batch Selection, Data Augmentation, Fine-tuning and Training, the architectures are trained on CACD dataset.

3.1 Dataset Collection

We used three datasets which are the MORPH-II dataset, the CACD dataset and the FG-NET dataset for evaluating the performance of the four state-of-the-art CNN architectures.

Figure 3.1: The MORPH-II Dataset

The MORPH-II dataset is a cross age dataset that has 55,134 images of 13,618 persons. The ages range from 16 years to 77 years. It has 4 images per person on average. Figure 3.1 [30] shows a part of the MORPH-II dataset. We divided the dataset for train and test purposes. For training we divided 20000 images of 10000 subjects and each of these subjects have two paired images with a large age gap i.e youngest age to oldest age. Similarly, the test set has a set of probe images and a set of gallery images using the rest of the 3000 images. The probe set has images of the oldest age of each subject and the gallery set has images of the youngest age of those subjects. The gallery set has only one image of each subject.

The CACD dataset is another cross age dataset that has 163,446 images of 2000 subjects with ages ranging from 14 years to 62 years old. A new database was created using this dataset for verification systems which is called the CACD-VS dataset and it has 2000 positive pairs and 2000 negative pairs of images. The positive pairs are made by using two images of the same person with a large age gap and the negative

pairs are made using two image pairs of different persons with a large age gap. We used this dataset to fine-tune our models.

Figure 3.2: The FG-NET dataset

Lastly, we have the FG-NET dataset which is slightly smaller compared to the other two. The number of images per person in this dataset is higher but the total number of subjects is comparatively lower. Figure 3.2 [30] shows a part of the FG-NET dataset. For this reason, to induce a fair comparison we proposed to use the MORPH-II dataset during our evaluation process. This dataset contains 1002 images of 82 subjects and the ages range from 0 years to 69 years old. All the images are named using 6 characters and numbers where the last two numbers denote the person's age and the first four characters with a combination of a variable and a number denotes the identity of the person.

Summary of datasets used			
Dataset	No. of photos	No. of subjects	No. of photos per subject
CACD	163,446	2000	81.7
MORPHE-II	55,134	13,618	4.1
FG-NET	1002	82	12.2

Table 3.1: Dataset Summarized

A summarized view of the number of photos, number of subjects and number of photos per subject in the three datasets that we have used can be seen in the table 3.1.

3.2 Image pre-processing

Image pre-processing is used to improve the image data and terminate undesired distortions or strengthen some image features that might be needed for further processing.

3.2.1 Face Detection and Alignment

It's the identification of human faces from digital images. It uses biometrics to map facial features from an image. The reason for face identification is to discover all the face positions in the picture and mark them with rectangular boxes. The output of the algorithm is the directions of the rectangular face limits in the picture.

Face alignment is a Computer vision innovation for recognizing the mathematical structure of human faces in computerized pictures. Given the area and size of a face, it naturally decides the state of the face parts, for example, eyes and nose. A face arrangement program regularly works by iteratively modifying a deformable model, which encodes the earlier information on face shape or appearance, to consider the low-level picture confirmations and discover the face that is available in the picture as shown in Figure 3.3. A spatial transformation of images is done in this face alignment stage, called the geometric transformation. Which is responsible for:

- Aligning images taken in different times with different sensors.
- Correcting the images for lens distortion and camera orientation.
- Image morphing to other special effects.

Figure 3.3: Face detection using MTCNN

Other geometric transformations can also be applied as required. They are, affine transformation, point matching,interpolation, inverse projection, image mapping and lens distortion correction.

3.2.2 Facial Landmark Localization

During landmark localization locates the position of landmarks, we locate the facial landmarks such as the corner of the mouth, eyes and the nose. It might likewise incorporate size, tilt edge and other data varying. In order to apply multi-view face detection and landmark localization in a complex environment, we employed the MTCNN method. We train both frontal and non frontal faces based on the MTCNN method in order to get a better multiview detection [4]. The output can be seen in Figure 3.4.

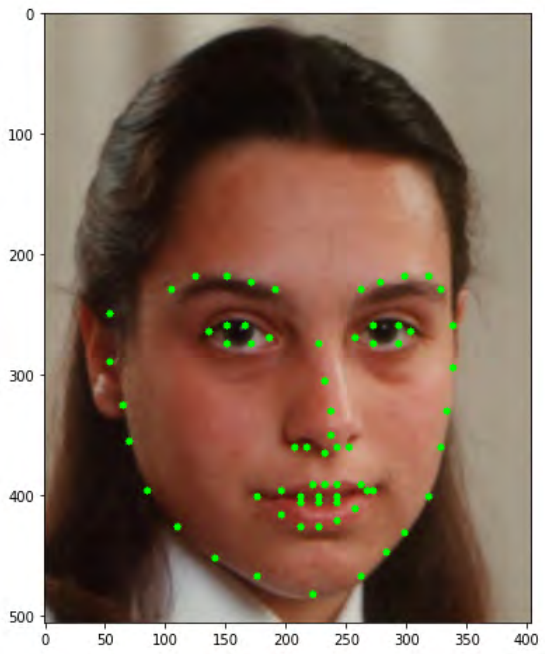


Figure 3.4: Face landmark detection using MTCNN

Like the bounding box regression task, facial landmark identification is detailed as a regression issue and we limit the Euclidean misfortune as.

$$L_I^{Landmark} = \left\| \widehat{y}_i^{Landmark} - y_i^{Landmark} \right\|_2^2 \tag{3.1}$$

Here $\widehat{y}_i^{Landmark}$ the facial landmark coordinates obtained from the network, $y_i^{Landmark}$ is the ground truth coordinate for the ith sample. For the five facial landmarks (left eye, right eye, nose, left mouth corner, right mouth corner) we use $y_i^{Landmark} \epsilon \mathbb{R}^{10}$

3.2.3 Resizing

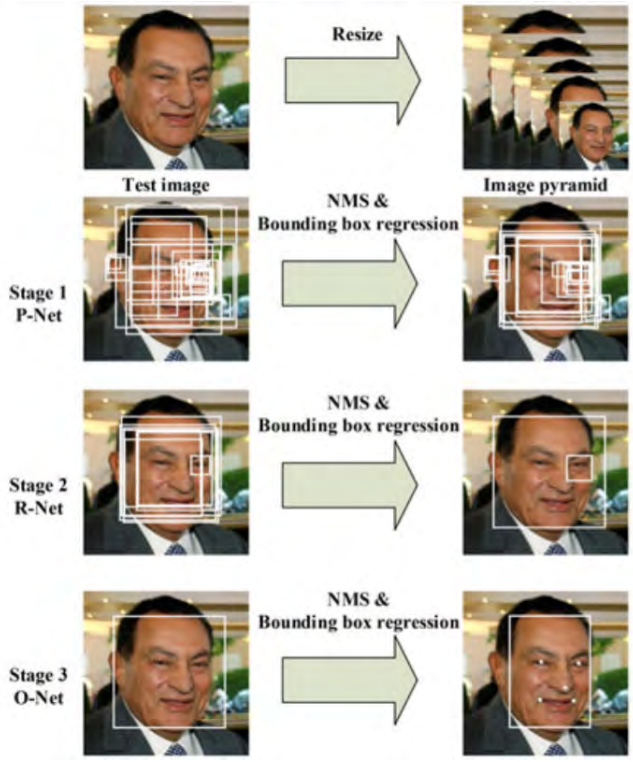


Figure 3.5: Resizing into various scales

When provided an image, the first thing to do is to resize it to various scales, with the objective being to create an image pyramid. This will be the input of the three stage cascaded framework:

- Stage 1: Taking advantage of Proposals network, which is a convolutional network, to retrieve candidate facial windows and bounding box regression vectors. They are then calibrated based on agency estimated bounding box vectors. Non maximum suppression is then used to merge overlapped candidates.

- Stage 2: Candidates are then put into Refine Network, which is another CNN. Here more false candidates are rejected, calibration of the bounding box regression is done, and NMS is conducted.

- Stage 3: This is similar to stage 2, but the difference that face regions are identified with higher supervision. This network will give us an output for five facial landmark positions.

The overall pipeline of our approach is shown in the Figure 3.5 [4].

3.3 Training Strategies

Mini-Batch Selection is used to make classes and pick images for every class and the images undergo Data Augmentation and Fine-tuning before the Training procedure.

3.3.1 Mini-Batch Selection

Subsetting through the entire dataset and utilizing all the pictures consecutively is the better approach because of inadequate memory. However for a relatively small dataset, it is difficult to access all training images in memory. Thus, a subset of pairs must then be picked for the batch system. The technique for mini-batch sampling utilized here is attributed to as "Classes-Then-Images" [26]. A few classes are encompassed in the present mini-batch which can be done based on random or a concentrated class prospecting technique as shown in Figure 3.6 [26]. Multiple images are picked for every chosen class to be used in the mini batch which can also be done based on random or with some form of concentrated class prospecting technique [26]. Through this method of sampling, exactly a certain amount of times every class emerges in the training can be determined. If the classes are arbitrarily collected, it is less probable that images from image-rich classes are collected into the mini-batch than images from several, smaller classes. The sampling procedure could be well-competent to differentiate amongst closely looking faces of distinctive individuals, such as Doppelganger Mining, for datasets with a greater number of classes [26]. The technique involves and uses classes and photos from customized training data as a training dataset and ultimately collects a limited number of distinct training data in every mini-batch, indicating that the specific problem is learned by the network.

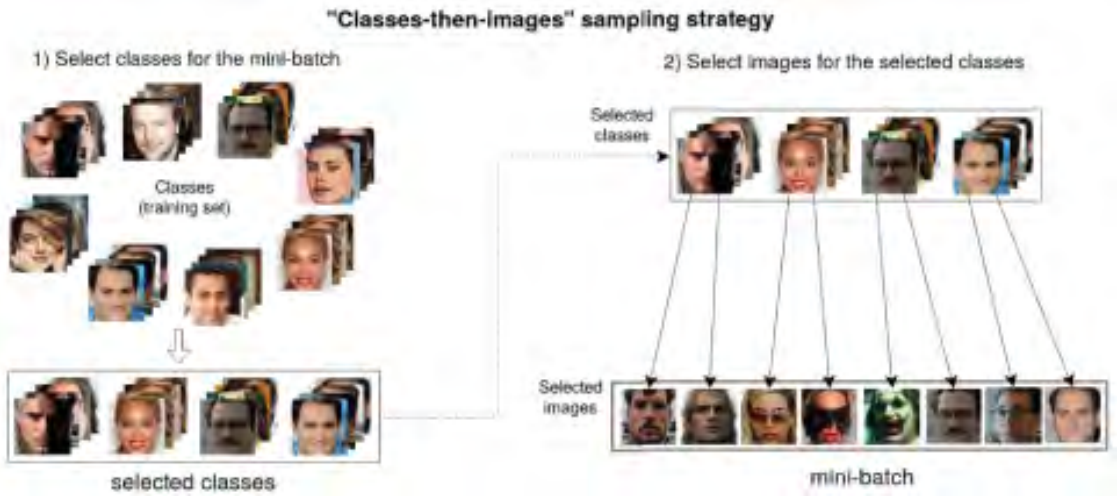


Figure 3.6: "Classes-then-images" technique for sample selection

3.3.2 Data Augmentation

Augmentation of data is an optimal solution to substitute for inadequate evidence on facial preparation. To make assumptions well in unrestrained contexts, deep learning depends heavily on massive and complicated training sets. There is also a shortage of differences in current datasets relative to the studies throughout the real world. Random Cropping, Photometric Distortion: including unpredictable modifications to images such as colour, contrast, and intensity are used in the method of data augmentation for face detection. Whereas the Facial Recognition data augmentation procedure requires geometric transformation, such as shifting, rotating, zooming, scaling, cropping, angle transformation, elastic distortion, etc [4]. Therefore in this manner, from basic information, we can ensure greater use of the face feature extraction process. In order to refine the preparation, we then utilize our dataset so that the framework adapts to our real implementation simulations.

3.3.3 Fine-tuning

To train the VGG16, VGG19, Inception-Resnet, ResNet-50, Xception deep CNN framework using CACD face images, the preferred fine-tuning methodology was taken which is a transition computing methodology. It demonstrates that only the data collected to resolve one challenge is used to resolve another problem [24]. Thousands of RGB face images learn the pre-trained RGB construct of CACD, and hence the model's parameters are obtained. These parameters were determined as the deep CNN parameters. By precisely modifying the preliminary model parameters, training of the deep CNN models are performed. The fine-tuning method described mitigates the necessity of specific instruction. Direct training is a strategy in which a deep CNN model 's initial variables are defined arbitrarily before training [24]. The suggested finetuning technique helps the deep CNN models to be adequately learned. This is because of the resemblances between the deep CNN models parameters and the pre-trained RGB construct of CACD parameters. Through direct training, the two approaches have been trained and the convolution filter product of the two approaches use the same properties, through greater levels in facial structures, and are equally trained to quickly recognize the facial structures [24].

3.3.4 Training

Aging face reconstruction entails identifying certain faces matching the sample face in the database, where there is a significant increasing age gap between the sample facial images and the relevant facial images in the database. Through the use of the nonlinear function process, weight deterioration λ and learning rate α are calculated [28]. First, the learning rate is defined, $\alpha = 0.001$. Then prepare a series of models on a tiny proportion of CACD (only images of five individuals) as well as pick the attribute that contains the highest results on the CACD-VS approval range, despite multiple choices of λ. λ is equivalent to α in the choice cycle [28]. We resolve the weight deterioration by increasing the value of λ gradually from 0.0001 to 0.001, as well as alter the values of α. Afterwards we pick the function that provides the better outcome on the CACD-VS testing collection. With a progressively increasing number of images of individuals taken from the CACD dataset, this entire procedure is repeated. In order to perform testing on CACD-VS, we only use pre-trained

models on CACD [28]. In fact, we do some tests to examine if this image pre-processing such as face alignment can improve the efficiency. We first trace the five feature points for learning images that are 2 different eye centers, nose tip as well as the two mouth edges. Subsequent implementation demonstrates that the use of aligned test data and data augmented endorses the reliability that renders the system further adaptive. For the verification tests on FG-NET, and MORPH-II dataset, we train the architectures solely on CACD dataset.

Chapter 4

Result and Analysis

This section has the outline of our experimentation methods and the results from these experimentations with a detailed analysis of the obtained results. It gives a thorough analysis about the different error rates and performance measures we used to calculate the accuracy and error rates which helped us in evaluating the performance of each architecture.

4.1 Experimentation

This paper selects the four Xception, Inception-ResNet V2, ResNet-50 and VGG-19 CNN architectures as shown in Figure 4.1, to train the models using the CACD dataset. In order to analyze the image preprocessing as well as data augmentation, etc, it is then evaluated on CACD-VS testing data. Finally, the four architectures will be tested on the MORPH-II and FG-NET databases. The results obtained after testing provide us the values of the FMR, FNMR and HTER. The values are then used on both the MORPH-II and FG-NET datasets to map the ROC curves of all four architectures. We calculate the AUC value from the ROC curve, which is eventually used in the verification process to evaluate the error rate. Finally, we can estimate which architecture has the optimum performance from the AUC values with the AUC value being closer to 1 as well as having a lower error rate whereas closer 0.5 having a higher error rate.

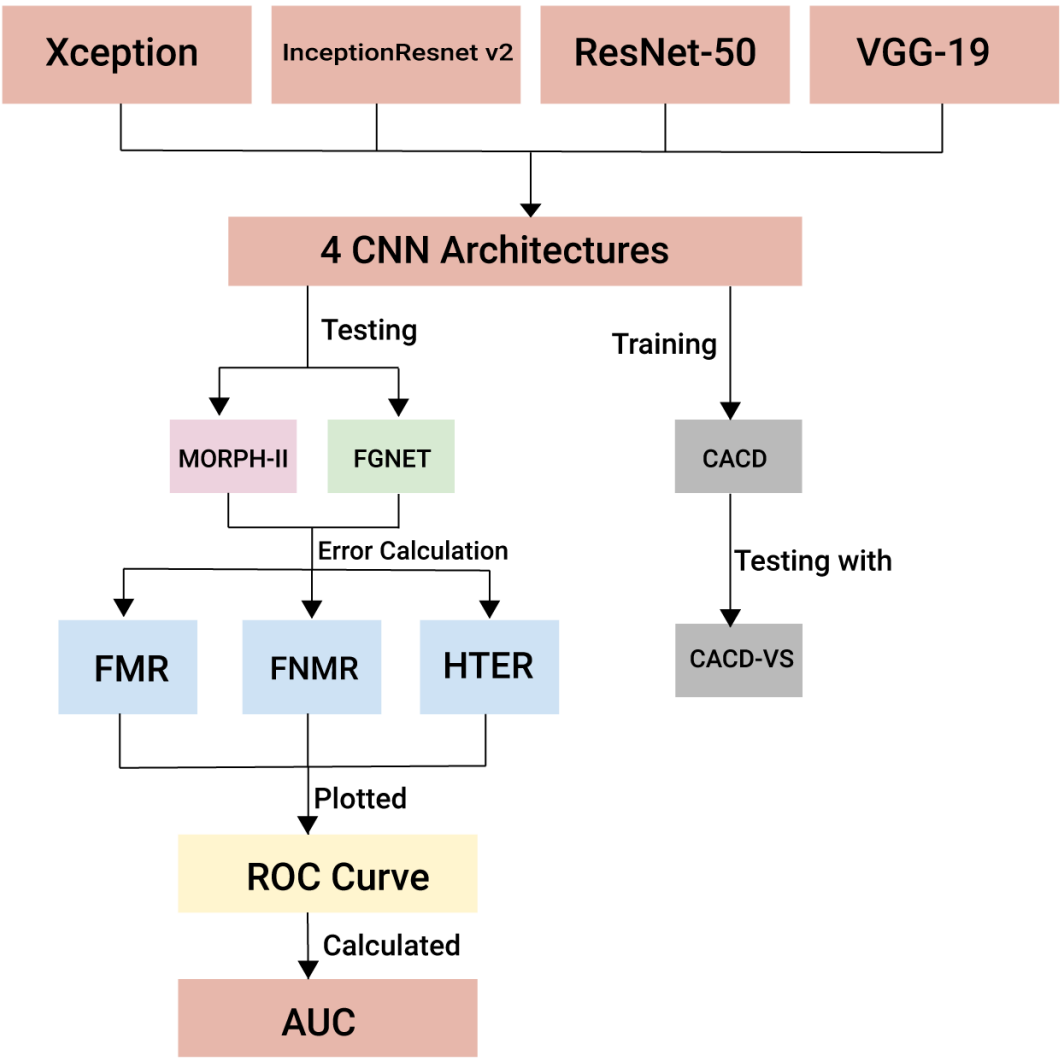


Figure 4.1: Workflow of experimentation process

4.2 Evaluation

The performance of face recognition systems are done to evaluate which ones have the optimum architecture for best results. Our research focuses on face recognition systems that verifies the identity of a probe when it is filtered against a reference identity. In order to do this, we have to teach the system each of the identities during the training. When a probe matches a given reference identity, it is going to be called "genuine" and on the other hand, if it is a different identity then it will be known as an "imposter".

In order to execute this process a part of the dataset is taken for teaching the model the identities of all the subjects and the other part is used as a probe. For generating results, each probe is brought against the known identities for similarity scores or distance scores.

	Predicted positive	Predicted negative
Actual positive	TP	TN
Actual negative	FP	FN

Table 4.1: Confusion matrix

From table 4.1, four situations can be derived. These are [28],

1. True Positive **(TP)**: These are the number of cases where the system estimates that the images of the probe and reference matched.

2. True Negative **(TN)**: These are the number of cases where the system estimates that the images of the probe and reference is different.

3. False Positive **(FP)**: These are the number of cases where the system estimates that the images of the probe and reference matched but actually they are different.

4. False Negative **(FN)**: These are the number of cases where the system estimates that the images of the probe and reference are different but actually they are the same.

4.3 Error Calculation

After training and testing our models, we have calculated the errors using FMR, FNMR, HTER and plotted ROC and AUC curves to get the accuracy.

4.3.1 FMR

The False Match Rate is used to measure the amount of error by wrong matches by a system. FMR represents type 1 error. The following equation is used to calculate FMR [28],

$$FMR = \frac{FP}{FP + TN} \tag{4.1}$$

In this paper, the CNN architecture with the lower FMR value has a less error rate of the verification of faces than the architectures with a higher FMR value.

4.3.2 FNMR

The False Non Match Rate is used to measure the amount of type 2 error where correct matches are identified as wrong matches. FNMR can be calculated using the following equation [28],

$$FNMR = \frac{FN}{FN + TP} \tag{4.2}$$

The architecture that has a lower FNMR value has less error rate of false mismatch of faces than the architectures with a higher FNMR value.

4.3.3 HTER

The Half Total Error Rate is used to measure the performance of a system by calculating the error rate. It can be calculated using the FMR and FNMR values. The following equation is used to calculate HTER [28],

$$HTER = \frac{FMR + FNMR}{2} \times 100 \tag{4.3}$$

The architecture with a lower HTER value has a lower error rate of face verification and it is better performed than the architectures with a higher HTER value.

4.3.4 ROC-AUC

The Receiver Operating Characteristics Curve is used to measure the performance of a system by analyzing the error rates of the system. The ROC curve is plot using the False Match Rate on the x-axis and the True Match Rate on the y-axis, with coordinates from 0 to 1. The architecture with a ROC curve which has a True Match Rate closer to 1 and a FMR closer to 0 has a lower error rate than the other architectures [28].

The Area Under a Curve is used to evaluate how well a system performs. The AUC value of the ROC curve is used to evaluate which architecture has a lower error in verifying the faces. The AUC value for an ROC curve ranges from 0.5 to 1, where 0.5 is depicted as the worst and 1 is depicted as the best. The architecture with an AUC value closer to 1 has less error while the verification of faces than the architectures which have an AUC value closer to 0.5 [28].

4.3.5 Accuracy

Accuracy is used in order to measure how efficiently a model is performing. It is a measure of positive face verifications amongst the total number of verifications. Accuracy is measured using the following equation [28],

$$\text{A}ccuracy = \frac{TN + TP}{TN + TP + FN + FP} \times 100 \tag{4.4}$$

In this paper, the CNN architecture with the higher accuracy has a higher rate of positive verification of ageing faces compared to the other CNN architecture with the lower accuracy.

4.4 Results

This section shows the results we have obtained from performing the evaluation processes on the four CNN architectures.

4.4.1 Verification Outcome of Architectures

Face verification evaluation is done corresponding to the various ages of gallery images and the image of the probe. The X-axis in the gallery photo indicates the age of an individual and the age of the Y-axis indicates the age of the probe. The colors in the graph reflect the precision of identity from 0 which is in the shade blue exhibits that none of the gallery and probe pairs had been matched whereas 1 being the shade in red which shows all probable probe and gallery image variations that were matched. The graphs contain two separate datasets for every architecture. Better results along the diagonal suggest that approaches are effective at verifying minimal age gaps and lowest scores demonstrate poor output in testing the ageing face. The findings suggest the architectures work better along the gradient when the age gap between the gallery image and the probe image is minimal.

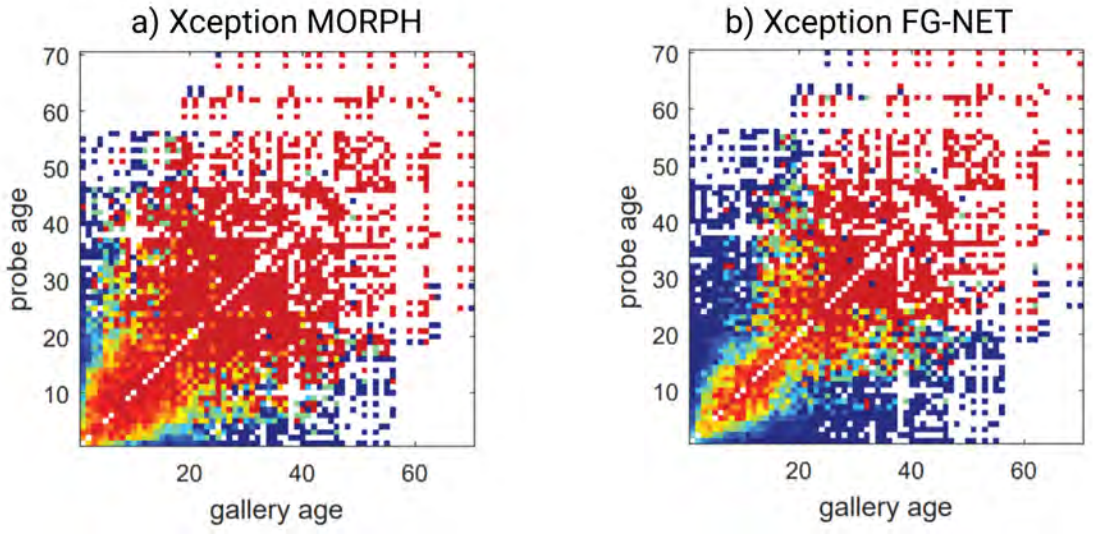


Figure 4.2: A Study of face verification of Xception with MORPH-II and FG-NET database

In Figure 4.2a, the higher amount of red color shows accuracy of Xception architecture in verifying probe images and gallery images on MORPH-II dataset. Although Figure 4.2b shows a comparatively lower verification rate on FG-NET due to the subordinate number of red colors rather than on MORPH-II dataset.

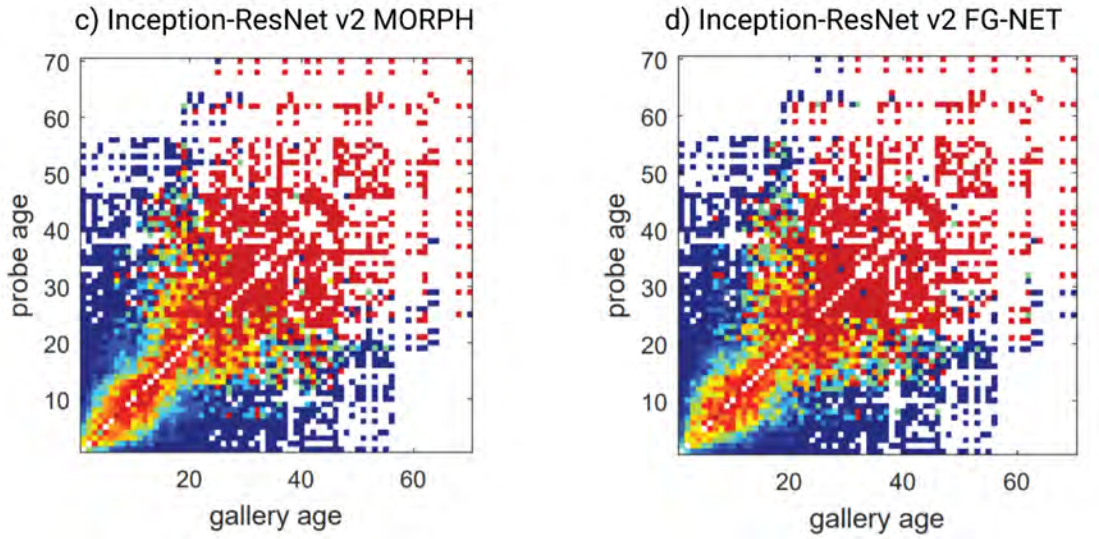


Figure 4.3: A Study of face verification of Inception-ResNet v2 with MORPH-II and FG-NET database

The performance of Inception-ResNet v2 architecture is comparable for both MORPH-II and FG-NET databases. The slight difference in the number of red colored representations in Figure 4.3c and 4.3d, show that it is achieving a better verification in MORPH-II than the FG-NET database.

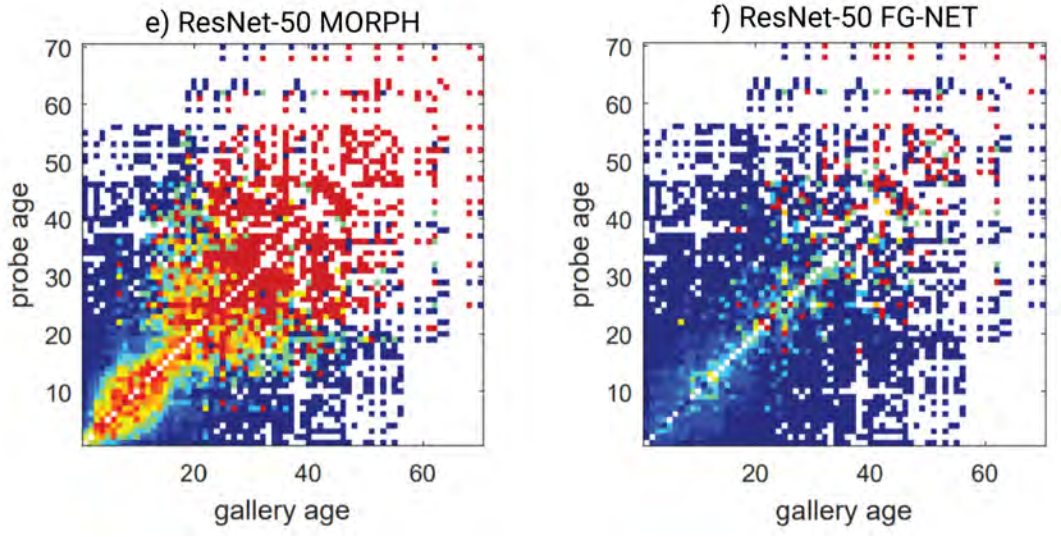


Figure 4.4: A Study of face verification of ResNet-50 with MORPH-II and FG-NET database

The blue colored representation in Figure 4.4f is much greater than that of Figure 4.4e. This exhibits that the architecture of ResNet-50 performs limited in the case of FG-NET whereas shows surpassing verification incase of MORPH-II database with probe and gallery image.

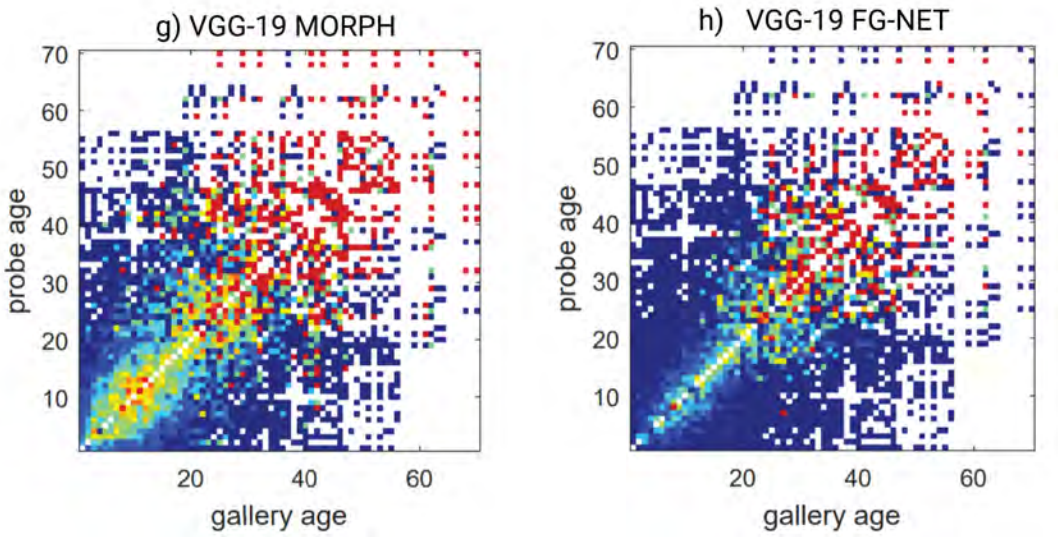


Figure 4.5: A Study of face verification of VGG-19 with MORPH-II and FG-NET database

Figure 4.5g shows VGG-19 architecture has a higher accuracy rate with MORPH-II database as it displays greater representation of red color in matching probe and gallery images. Whereas Figure 4.5h demonstrates the deterioration of the performance of ResNet-50 on FG-NET compared to all other architectures on the FG-NET database.

4.4.2 ROC Curves and AUC Values

We have obtained the ROC curve as shown in Figure 4.6 for the MORPH-II dataset after implementing the VGG-19, ResNet-50, InceptionResnet v2 and Xception pretrained models. The graph shows that the best performance was given by the Xception model as shown in Figure 4.2. The AUC values were calculated using these ROC curves for further evaluation.

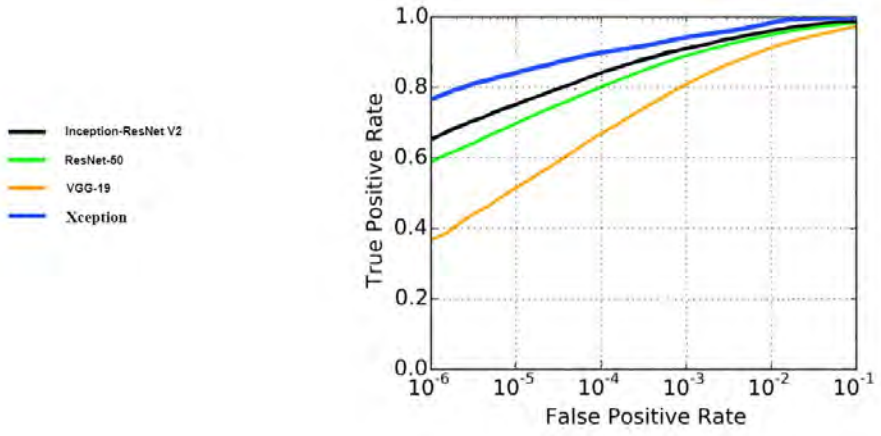


Figure 4.6: ROC curve for MORPH-II

Table 4.2 shows the AUC values that were derived from the ROC curves of the four architectures when they were tested using the MORPH-II dataset. The performance is showed in the table according to a pattern from highest to lowest where highest performance is given by the Xception architecture and lowest performance was given by the VGG-19 model.

Architecture	AUC value
Xception	0.733
InceptionResnet v2	0.5936
ResNet-50	0.582
VGG-19	0.337

Table 4.2: AUC values of MORPH-II

Similar to the MORPH-II dataset, we also implemented the four pretrained models on the FG-NET dataset for fair comparison analysis. This is because the MORPH-II dataset doesn't have that many pictures per subject but it has a large number of subjects. On the other hand, the FG-NET dataset has a large range of pictures for each person but a smaller number of subjects. For the FG-NET dataset we obtained the ROC curve shown in Figure 4.7 and calculated the AUC values from this ROC curve. The values show us a similar pattern as the MORPH-II dataset with Xception having the highest performance and VGG-19 having the lowest performance.

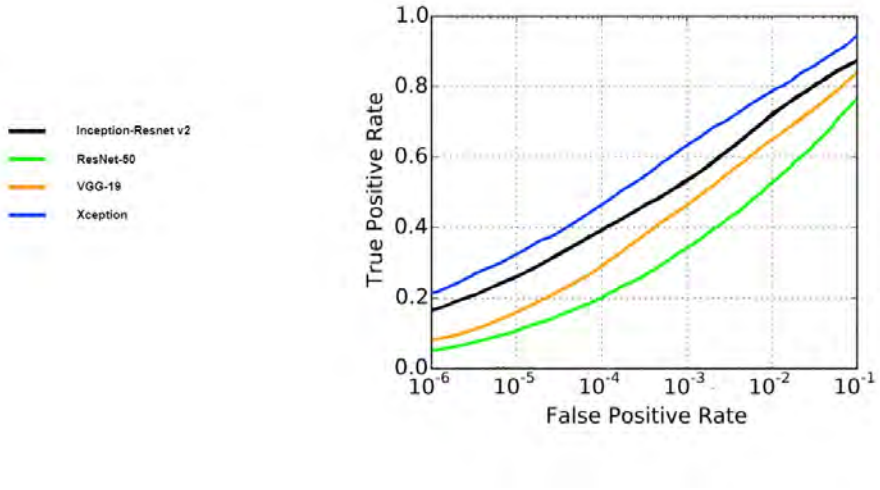


Figure 4.7: ROC curve for FG-NET

Table 4.3 shows the AUC values that were derived from the ROC curves of the four architectures when they were tested using the FG-NET dataset. The performance is showed in the table according to a pattern from highest to lowest where highest performance is given by the Xception architecture and lowest performance was given by the VGG-19 model.

Architecture	AUC value
Xception	0.4271
InceptionResnet v2	0.2916
ResNet-50	0.1231
VGG-19	0.1577

Table 4.3: AUC values of FG-NET

4.5 Analysis

The Xception architecture shows a higher performance in the MORPH-II dataset as it gives a higher AUC value. It gives an AUC value of 0.733 on the MORPH-II dataset. On the other hand, it has a comparatively lower AUC and performance on the FG-NET dataset and gives a value of 0.4271 on the FG-NET dataset. Similarly, the InceptionResnet v2 architecture also shows a better performance in the MORPH-II dataset with an AUC value of 0.5936 compared to the FG-NET dataset which has a AUC value of 0.2916 whereas ResNet-50 architecture shows a relatively poor performance for FG-NET with an AUC value of 0.1231 as to the MORPH-II database and all other architectures. Besides, VGG-19 too gives coinciding performance in both MORPH-II with an AUC value of 0.337 and FG-NET databases with an AUC value of 0.1577. One possible reason for poor performance in the FG-NET database could be the size of the dataset. It does not have enough pictures for proper evaluation.

Figure 4.8 shows the ROC curves of the four architectures on both the datasets. It can be derived from the graphs that all the architectures showed a similar pattern of performance when they were tested using both datasets. However, the performance given during testing using the FG-NET dataset was much lower compared to the MORPH-II dataset.

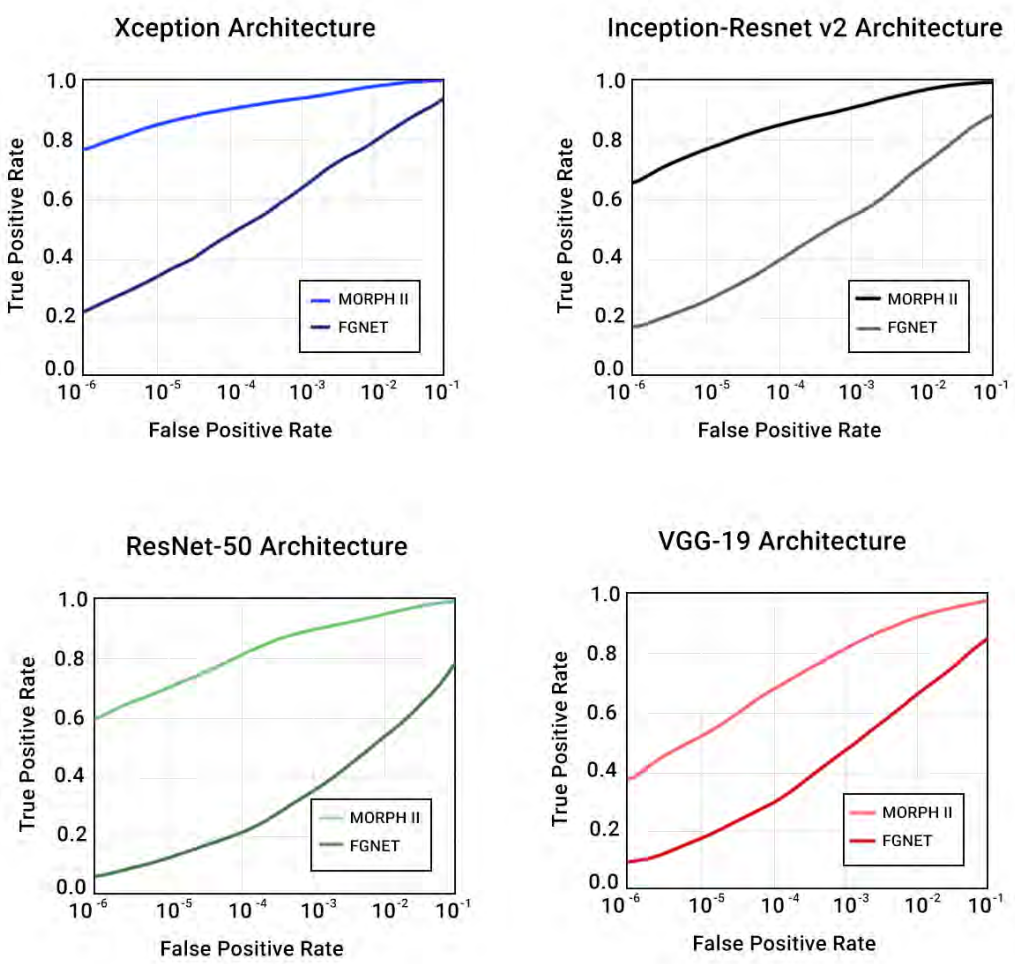


Figure 4.8: ROC curve for 4 CNN architectures on MORPH II and FGNET

The pre-trained models that we have used to conduct our result and analysis made it easier for us as it gave us good results by just using 10 epochs. Table 4.4 shows us the different FMR, FNMR and HTER values that we have derived from the four architectures by using their TP, TN, FP and FN values. It can be seen from the table that the highest FMR, FNMR and HTER values were obtained from the VGG-19 model and the lowest ones were obtained from the Xception model.

Architectures	FMR(%)	FNMR(%)	HTER(%)
Xception	10.49875	31.49625	20.9975
Inception-Resnet V2	13.772	32.018	22.87
ResNet-50	22.659	42.081	32.37
VGG-19	30.104	45.156	37.63

Table 4.4: Average FMR, FNMR and HTER values for 4 different CNN architectures

Table 4.5 shows us the different accuracy rates that was finally calculated using the TP, TN, FP and FN rates from both the datasets and then they were averaged to get a single accuracy value. It shows that Xception has the highest accuracy rate of 58.005%, InceptionResNet v2 has the seconnd highest accuracy rate of 44.26%, ResNet-50 has the third highest accuracy rate of 35.26% and finally VGG-19 has the lowest accuracy rate of 24.74%.

Architecture	Accuracy
Xception	58.005%
InceptionResNet v2	44.26%
ResNet-50	35.26%
VGG-19	24.74%

Table 4.5: Average verification accuracy

VGG-19 took a longer time compared to other models for during the fine-tuning process which can be a serious issue when larger datasets are going to be used. Since we have used comparatively smaller datasets, we did not face much problem. This is because the model has a lot of weight parameters. It uses very small 3×3 convolutional filters and increases the number of feature maps after the 2×2 pooling layer. The depth of the network is 19 weight layers. Hence, it is a heavier model and requires more training time. However, the verification accuracy rate that was given by this model was 24.74%. Figure 4.9 is a bar chart diagram that shows that the VGG-19 architecture has a FMR value of 30.104%, a FNMR value of 45.156% and a HTER value of 37.63%. It gives a low performance rate because the loss

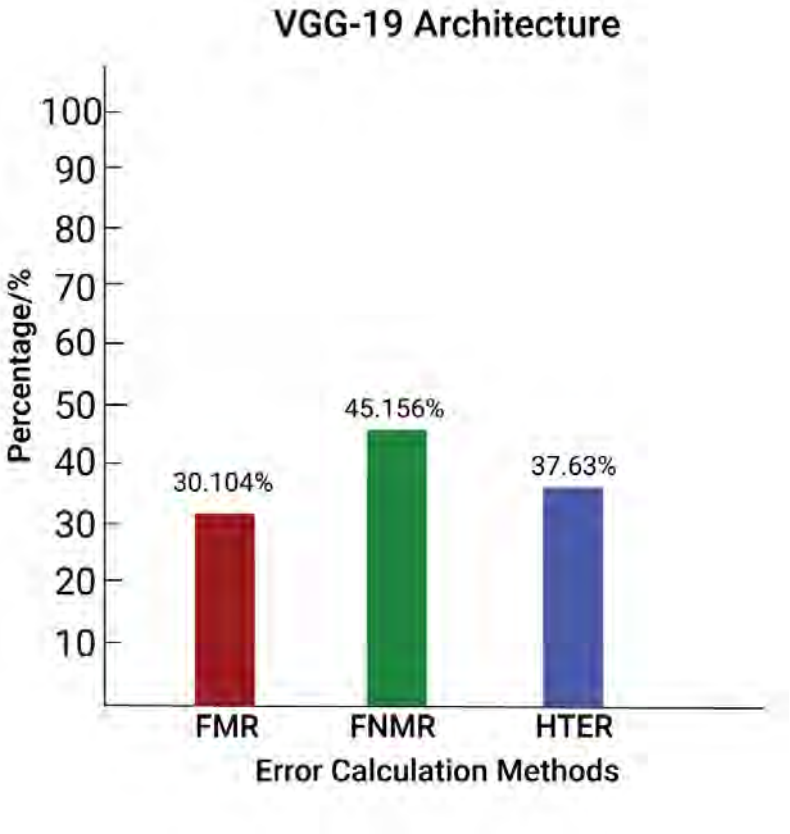


Figure 4.9: Bar Chart for VGG-19 Architecture

function used in this model is not highly efficient. The VGG loss is a Euclidean-distance-based loss which is a metric learning method that reduces intra-variance and enlarges inter-variance by embedding images into the euclidean space. It is based on the ReLU activation layers of the pre-trained network. To show the feature map derived from the j^{th} convolution after the activation and before the i^{th} maxpooling layers of the network, ϕ, j^{th} is used. Finally, the VGG loss is defined as the euclidean distance between the feature representations of a reconstructed image $G_{\theta_G}(I^{LR})$ and the reference image I^{HR}:

$$L^{SR}_{\underset{i,j}{VGG}} = \frac{1}{W_{i,j}H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} \left(\phi_{i,j}(I^{HR})_{x,y} - \phi_{i,j}\left(G_{\theta_G}(I^{LR})\right)_{x,y} \right)^2 \tag{4.5}$$

Where, $W_{i,j}$ and $H_{i,j}$ defines the dimensions of the feature maps in the network.

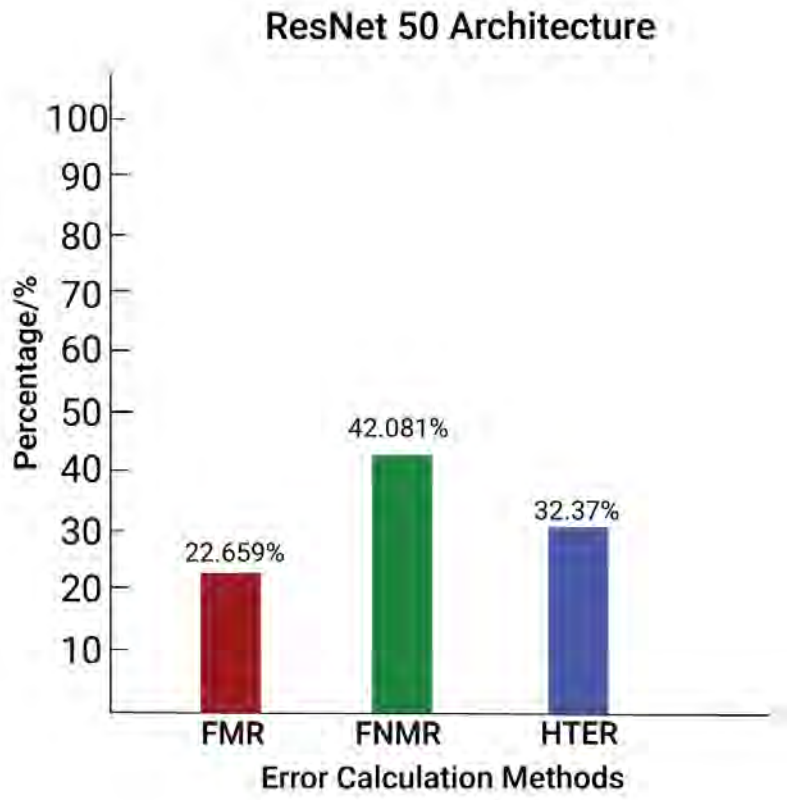


Figure 4.10: Bar Chart for ResNet-50 Architecture

ResNet-50 is much deeper with 50 layers than VGG-19 and its primary goal was to increase accuracy rate and we have observed that it increases the accuracy rate. The accuracy increased to a 35.26%. Figure 4.10 is a bar chart diagram that shows that the ResNet-50 architecture has a FMR value of 22.659%, a FNMR value of 42.081% and a HTER value of 32.37%. Even though it is much deeper but the model size is substantially smaller as instead of using fully-connected layers it uses global average pooling. This network shows an improvement in the performance as it was initially introduced as a solution to the vanishing gradient problem. This problem is caused as a result of using more layers with activation functions which causes the gradients of the loss functions to become zero making it hard to train the network. For shallow networks this isn't that big of a problem but as a network increases its number of layers the gradient can become too small for efficient training. This problem is solved by the ResNet-50 architecture as it provides the gradients with a clear pathway in the network to pass them through by using its identity mapping using layer inputs

$$F(x) := H(x) - x \tag{4.6}$$

where $H(x)$ is the desired underlying mapping.

InceptionResnet v2 was much faster than VGG-19 and ResNet-50. Each epoch in this model took 1/4th of the time taken by the VGG-19 model. This is because the InceptionResnet v2 is much deeper than VGG-19 which is a relatively shallow network. The InceptionResnet v2 model gives a verification accuracy of 44.26%. The further improved performance involves the contribution of its adoption of a triplet loss function based on triplets of roughly aligned matching/non-matching

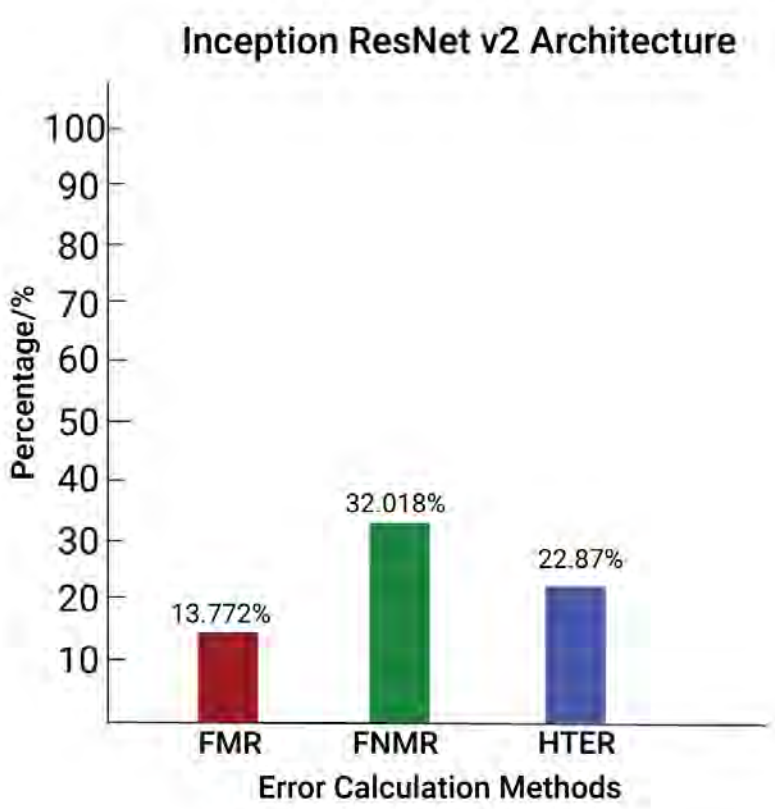


Figure 4.11: Bar Chart for InceptionResnet v2 Architecture

face patches that were generated by an online novel triplet mining method. This loss function needs the face triplets. It minimizes the anchor and positive sample distance of the same identity and maximizes the anchor and negative sample distance of a different identity. In simple words, the anchor image is supposed to be closer to the positive image and farther from the negative image in the euclidean space which is calculated using:

$$\left\| f(A) - f(P) \right\|^2 + \alpha \leq \left\| f(A) - f(N) \right\|^2 \tag{4.7}$$

Where,
$\left\| f(A) - f(P) \right\|^2$ is the distance between anchor and positive;
$\left\| f(A) - f(N) \right\|^2$ is the distance between anchor and negative;
Margin α is added to the positive to keep the positives further apart from the negatives set. Figure 4.11 is a bar chart diagram that shows that the InceptionResNet v2 architecture has a FMR value of 13.772%, a FNMR value of 32.018% and a HTER value of 22.87%.

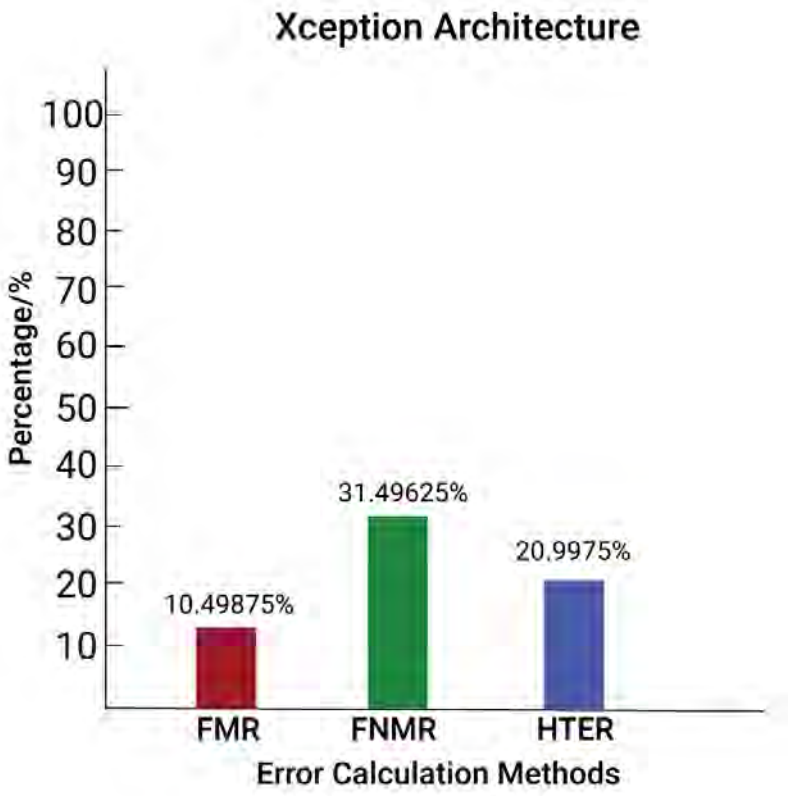


Figure 4.12: Bar Chart for Xception Architecture

Xception displays the smallest weight serialization at just 91MB. Meanwhile it also outperforms the Inception module with depthwise seperable convolution for better element learning worked by ordinary convolutions over a joint "space-cross-channels domain" into two simpler step. Xception shows a further improvement in the accuracy rate due to the replacement of standard inception modules with depthwise separable convolutions. The accuracy rate was improved to a 58.005%. In figure 4.12 we can see a bar chart diagram that shows that the Xception architecture has a FMR value of 10.49875%, a FNMR value of 31.49625% and a HTER value of 20.9975%.

Chapter 5

Conclusion and Future Work

Aging face recognition to this day remains a difficult task but this proposed method is aimed at minimizing complexity and providing an efficient procedure. In this paper, we analyzed CNN's four distinct architectures and demonstrated that CNN architecture Xception reliably outperforms all the other CNN architectures InceptionResnet v2, ResNet-50, VGG-19 tested here, in the case of verifying an ageing face. The key concept was to evaluate their performance on the basis of the ROC, FMR, FNMR, HTER to have the best performing architecture with the lowest error rate, the AUC value. This outcome is reliable nevertheless, for an increased gap in the age of an individual and because of fewer quantities of images for each individual, there is a less consistent impact, remaining around the same or perhaps even rising marginally in one instance.

Authenticating an aging face is a perplexing issue which might involve ongoing efforts for added increase in the efficacy of identification. In this research, we concentrated on verifying the aging face, but the techniques can potentially be protracted to classify other features, such as race classification, invariance of the expression, etc. In future work, with a larger number of datasets we can determine the CNN architecture with the best performance in researching age estimation from facial images through deformation and various lighting conditions to further enhance the accuracy of age-related problems.

Bibliography

[1] H. El Khiyari and H. Wechsler, "Face verification subject to varying (age, ethnicity, and gender) demographics using deep learning", *Journal of Biometrics and Biostatistics*, vol. 7, no. 323, p. 11, 2016.

[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[3] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning", *arXiv:1602.07261*, 2016.

[4] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks", *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[5] M. F. Aydogdu, V. Celik, and M. F. Demirci, "Comparison of three different cnn architectures for age classification", in *2017 IEEE 11th international conference on semantic computing (ICSC)*, IEEE, 2017, pp. 372–377.

[6] S. Chen, C. Zhang, M. Dong, J. Le, and M. Rao, "Using ranking-cnn for age estimation", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5183–5192.

[7] F. Chollet, "Xception: Deep learning with depthwise separable convolutions", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[8] H. El Khiyari, H. Wechsler, *et al.*, "Age invariant face recognition using convolutional neural networks and set distances", *Journal of Information Security*, vol. 8, no. 03, p. 174, 2017.

[9] N. K. Gondhi and E. N. Kour, "A comparative analysis on various face recognition techniques", in *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, IEEE, 2017, pp. 8–13.

[10] L. Kaiser, A. N. Gomez, and F. Chollet, "Depthwise separable convolutions for neural machine translation", *arXiv preprint arXiv:1706.03059*, 2017.

[11] Z. Qawaqneh, A. A. Mallouh, and B. D. Barkana, "Deep convolutional neural network for age estimation based on vgg-face model", *arXiv:1709.01664*, 2017.

[12] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 121–135, 2017.

[13] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa, "An all-in-one convolutional neural network for face analysis", in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, IEEE, 2017, pp. 17–24.

[14] P. Rodrıguez, G. Cucurull, J. M. Gonfaus, F. X. Roca, and J. Gonzàlez, "Age and gender recognition in the wild with deep attention", *Pattern Recognition*, vol. 72, pp. 563–571, 2017.

[15] X. Wang, A. M. Ali, and P. Angelov, "Gender and age classification of human faces for automatic detection of anomalous human behaviour", in *2017 3rd IEEE International Conference on Cybernetics (CYBCONF)*, IEEE, 2017, pp. 1–6.

[16] R. R. Atallah, A. Kamsin, M. A. Ismail, S. A. Abdelrahman, and S. Zerdoumi, "Face recognition and age estimation implications of changes in facial features: A critical review study", *IEEE Access*, vol. 6, pp. 28 290–28 304, 2018.

[17] C.-L. Chin, M.-C. Chin, T.-Y. Tsai, and W.-E. Chen, "Facial skin image classification system using convolutional neural networks deep learning algorithm", in *2018 9th International Conference on Awareness Science and Technology (iCAST)*, IEEE, 2018, pp. 51–55.

[18] A. Kharchevnikova and A. V. Savchenko, "Neural networks in video-based age and gender recognition on mobile platforms", *Optical Memory and Neural Networks*, vol. 27, no. 4, pp. 246–259, 2018.

[19] H. Liao, Y. Yan, W. Dai, and P. Fan, "Age estimation of face images based on cnn and divide-and-rule strategy", *Mathematical Problems in Engineering*, vol. 2018, 2018.

[20] Y. Mahajan and S. Sondur, "Aging face recognition using deep learning", *International Journal of Engineering and Applied Sciences*, vol. 33, 2018.

[21] N. Mualla, E. H. Houssein, and H. H. Zayed, "Face age estimation approach based on deep learning and principle component analysis", *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 2, 2018. DOI: 10.14569/IJACSA.2018.090222. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2018.090222.

[22] R. Rothe, R. Timofte, and L. Van Gool, "Deep expectation of real and apparent age from a single image without facial landmarks", *International Journal of Computer Vision*, vol. 126, no. 2-4, pp. 144–157, 2018.

[23] A. Alekseev and A. Bobe, "Gabornet: Gabor filters with learnable parameters in deep convolutional neural network", in *2019 International Conference on Engineering and Telecommunication (EnT)*, IEEE, 2019, pp. 1–4.

[24] J. Kim, H. Jo, M. Ra, and W.-Y. Kim, "Fine-tuning approach to nir face recognition", in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 2337–2341.

[25] A. V. Savchenko, "Efficient facial representations for age, gender and identity recognition in organizing photo albums using multi-output convnet", *PeerJ Computer Science*, vol. 5, e197, 2019.

[26] E. Smirnov, A. Oleinik, A. Lavrentev, E. Shulga, V. Galyuk, N. Garaev, M. Zakuanova, and A. Melnikov, "Face representation learning using composite mini-batches", in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[27] J. Zhao, Y. Cheng, Y. Cheng, Y. Yang, F. Zhao, J. Li, H. Liu, S. Yan, and J. Feng, "Look across elapse: Disentangled representation learning and photorealistic cross-age face synthesis for age-invariant face recognition", in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, 2019, pp. 9251–9258.

[28] Dayer, "Face recognition systems: Performance evaluation and bias analysis", 2020.

[29] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks", *Artificial Intelligence Review*, pp. 1–62, 2020.

[30] X. Liu, Y. Zou, H. Kuang, and X. Ma, "Face image age estimation based on data augmentation and lightweight convolutional neural network", *Symmetry*, vol. 12, no. 1, p. 146, 2020.

[31] *013 cnn vgg 16 and vgg 19 — master data science*, http://datahacker.rs/deep-learning-vgg-16-vs-vgg-19/.

[32] *Cs231n convolutional neural networks for visual recognition*, https://cs231n.github.io/convolutional-networks.

[33] *Layers of a convolutional neural network - convolutional neural networks for image and video processing - tum wiki*, https://wiki.tum.de/display/lfdv/Layers+of+a+Convolutional+Neural+Network/.

[34] B. Raj, *A simple guide to the versions of the inception network*, https://towardsdatascience.com/a-simple-guide-to-the-versions-of-the-inception-network-7fc52b863202.

[35] *Review: Xception — with depthwise separable convolution, better than inception-v3*, https://towardsdatascience.com/review-xception-with-depthwise-separable-convolution-better-than-inception-v3-image-dc967dd42568.

[36] S. Saha, *A comprehensive guide to convolutional neural networks—the eli5 way*, https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53.

[37] D. Soni, *Supervised vs. unsupervised learning*, https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d.

[38] S. H. Tsang, *Review: Inception-v4 — evolved from googlenet, merged with resnet idea (image classification)*, https://towardsdatascience.com/review-inception-v4-evolved-from-googlenet-merged-with-resnet-idea-image-classification-5e8c339d18bc.