Classification of Indians into North and South Indians

A Report Submitted
in Partial Fulfillment of the Requirements
for the Degree of
Bachelor of Technology
in
Computer Science & Engineering

by

Tuhin Subhra Patra (20164142)
Upamanyu Jamwal (20164169)
Rajat Dipta Biswas (20164114)
S Pranav Ganesh (20164098)

to the

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
MOTILAL NEHRU NATIONAL INSTITUTE OF TECHNOLOGY
ALLAHABAD
May, 2019

UNDERTAKING

I declare that the work presented in this report titled "*Classification of Indians into North and South Indians*", submitted to the Computer Science and Engineering Department, Motilal Nehru National Institute of Technology, Allahabad, for the award of the ***Bachelor of Technology*** degree in ***Computer Science & Engineering***, is my original work. I have not plagiarized or submitted the same work for the award of any other degree. In case this undertaking is found incorrect, I accept that my degree may be unconditionally withdrawn.

May, 2019
Allahabad

(Tuhin Subhra Patra)

(Upamanyu Jamwal)

(Rajat Dipta Biswas)

(S Pranav Ganesh)

CERTIFICATE

Certified that the work contained in the report titled "*Classification of Indians into North and South Indians*", by
Tuhin Subhra Patra (20164142)
Upamanyu Jamwal (20164169)
Rajat Dipta Biswas (20164114)
S Pranav Ganesh (20164098)
has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.

(Prof. Suneeta Agarwal)
Computer Science and Engineering Dept.
M.N.N.I.T, Allahabad

May, 2019

Preface

Human facial images provide demographic information, such as ethnicity and gender. Conversely, ethnicity and gender also play an important role in face-related applications. Image-based ethnicity identification problems are addressed in machine learning framework. The human face is a highly rich stimulus that provides diverse information for adaptive social interaction with people. Humans are able to process a face in a variety of ways to categorize it by its identity, along with a number of other demographic characteristics, including ethnicity (or race), gender and age. Over the past few decades, a lot of effort has been devoted in the biological, psychological, and cognitive sciences areas, to discover how the human brain perceives, represents and remembers faces. Computational models have also been developed to gain some insight into this problem. In this paper, we have reduced the ethnicity classification into a two-category classification problem, North Indian and South Indian, each of which have relatively same anthropometrical features. We have used AlexNet, VGGs, ResNets and their variants to for this task and compared the performance of these models on a particular dataset.

Acknowledgments

We would like to thank everyone who contributed in some way in helping us achieve our goals with regards to our project. We are especially grateful to our mentor Prof. Suneeta Agarwal for assisting us in our project. Without her expertise and suggestions, we would never have been able to complete our work.

We would also like to thank teachers of the Computer Science and Engineering Department for the formation of groups for the project which enabled us to work on this project.

Contents

Chapter 1

Introduction

Ethnicity identification is the process of recognizing the ethnic group of an individual from a facial image. The human face provides a wealth of information including identity, gender, age, race, expression etc. Among the demographic attributes, ethnicity remains invariant through all lifetime and greatly supports face recognition systems, therefore, automatic facial ethnicity classification has been receiving increasing attention in recent years and several methods have been proposed. However, accurate and swift classification of different races based on human face in an uncontrolled environment is challenging. For efficient classification, one has to find race sensitive features from face images. These discriminative features can be differentiated in three categories namely, chromatic/skin tone, local features and global features. Due to similar skin colour for different races and extreme variation in illumination conditions for real world scenarios, skin tone alone cannot classify. However, combined with local or global descriptors classification accuracy can be boosted. Deep learning methods have produced state-of-the-art accuracy on many different classification tasks especially for image classification. CNN architectures have been compared on the basis of coarse grained classification tasks like Cats vs Dogs [1], Dog Breed Classification [2] in the recent past.

Face conveys the most direct and quick impression of an individual. We can often guess a person's ethnicity by the way he or she looks. However in a multi-cultural and complex society like India where people have very similar facial anatomy partially due to their close geographical relationships, it's more challenging to classify the population as North and South Indians. Gleaned the power of the recent advances in computer vision and machine learning, we took the challenge to investigate whether or not it is possible to

classify North and South Indians whose faces are highly resembling and whether or not there might be a mathematical rationale behind these nuances that are not so obvious to human eyes.

1.1 Objective and Motivations

The goal of the study is to predict if a person is a North or South Indian given his or her facial image. In order to make such predictions, a set of facial images categorized into North and South Indians was collected. We randomly selected 80% of the dataset as the training data to train the machine learning algorithm, 10% as validation set and 10% as test set. The facial images were first cropped and augmented and important features were extracted and fed into different neural network learning approaches for which learning curves and results were analysed.

Chapter 2

Related Work

There are a number of algorithms that have been devised over the years for the ethnicity identification of humans. Work done by P. Viola and M. Jones [3] has provided the efficient and rapid method of detecting face in input image. This is a novel approach which uses Adaboost classifier. This has high detection rate with very less computation time on the data set consisting of images under varying condition like illumination, pose, color, camera variation; etc. This algorithm was used on our data set to detect face in the image which will be later processed further. Lu et al. [4] has proposed ethnicity classification algorithm in which image of the faces were examined at multiple scales. The Linear Discriminant Analysis (LDA) scheme is used for input face images to improve the classification result. The accuracy of the performance of this approach is 96.3% on the database of 2,630 sample images of 263 subjects. However, the dataset considered in this work consisted only of two classes i.e. Asian and non-Asian. Hosoi et al.[5] have integrated the Gabor wavelet features and retina sampling for their work. These features were then used with the Support Vector Machines (SVM) classifier. This approach has used three categories: Asian, African and European. And the accuracy achieved for each category is: 96%, 94% and 93% respectively. However their approach seemed to have issues when considering other ethnicities. In [6], ethnicity classification under the varying age and gender was performed on the very large scale dataset for the first time. The MORPHII dataset was used for this work which had 55,000 images. Guo and Mu has used Gabor features for classification problem of five ethnicities: Black, White, Hispanic, Asian and Indian. The prediction results for Black and White were good: 98.3% and 97.1% respectively. But due to insufficient dataset for other three races, prediction results

deteriorated to 74.2% for Hispanic, 59.5% for Asian and 6.9% for Indian. S. Md. Mansoor has used Viola Jones [3] algorithm for face detection problem. After the detection of face, various features namely skin color; lip color and normalized forehead area were extracted from the image. This classification problem has used the Yale, FERET [8] dataset of Mongolian, Caucasian and Negroid images. The overall accuracy achieved in this work with these features was 81%. It was evident from the above mentioned works that none had considered geometric features for their solution. Also the scope of pre-trained Convolution Neural Network has yet not been explored for this problem so far. Hence in this problem, we have considered geometric features for training the ANN and have also attempted to use convolutional neural networks for solving this problem

Chapter 3

Proposed Work

Majority of previous researches have been focused on coarse grained classifications like Asian and non-Asian; Asian, African and European; Chinese, Korean and Japanese [10] of people based on their faces. We intend to characterise the performance of computers on a fine-grained classification task. In recent years numerous Convolutional Neural Network (CNN) architectures have been developed for various image recognition and classification tasks. Many sophisticated techniques have been developed to go deeper given the computational powers of modern day computers.

CNN architectures have been compared on the basis of coarse grained classification tasks like Cats vs Dogs, Dog Breed Classification in the recent past. We take the challenge of comparing some state-of-the-art CNN architectures by using transfer learning. We also evaluate same architectures with different configurations and weigh the practical benefits of using more computationally heavier versions of a given architecture.

Chapter 4

Data Set and Pre-Processing

4.1 Data Set Description

Despite that there is a wealth of large-scare facial image databases available, such as Chicago Face Database[11] and the CAS PEAL database[12], nevertheless, these databases are not applicable to our objective, as the images from these data sets are labeled by ethnicity i.e. Indian, Sri Lankan, Pakistani etc. instead of locality. Only few face data sets labelled on Indian localities exist. Owing to the dearth of the data set, we used the Data set used in the paper, *"Are you from North or South India? A hard race classification task reveals systematic representational differences between humans and machines"* by Harish Katti and S.P. Arun, [9] which contains images labelled as North and South Indians, since it contains ample number of images in both classes as is needed by deep neural architectures that we intended to benchmark. The data set is harvested utilizing the Google Custom Search API to collect a total of 1647 photos. Out of these, 459 were the pictures which he clicked themselves whereas the remaining 1188 were taken from the Centre for Neuroscience Indian Face Dataset(CNSIFD). This dataset was first converted into CSV format to be used by us.

Since Indian names are strongly determined by their ethnicity, a total of 128 typical first and 325 last names from each region were identified. Example first names were Birender, Payal for North, Jayamma, Thendral for South. Example last names were Khushwaha, Yadav for North and Reddy, Iyer for South. Google Image search APIs was used to search for face photographs associated with combinations of these typical first and last names. All images were directly downloaded from the internet. The number of

Face set	Total	Male	Female	North	South	Other
Set 1	459	260	199	140	209	110
Set 2	1188	710	478	636	552	0
Total	1647	970	677	776	761	110

Figure 1: Summary of face dataset. Set 1 consisted of face photographs taken with consent from volunteers who declared their own race. Set 2 consisted of face images downloaded from the web.

facial images for each subgroup is listed in Figure 1.

The age distribution of the data set for males and females separately is shown in Figure 2. The gender wise north and south Indian sample distribution is shown in Figure 3 and 4.

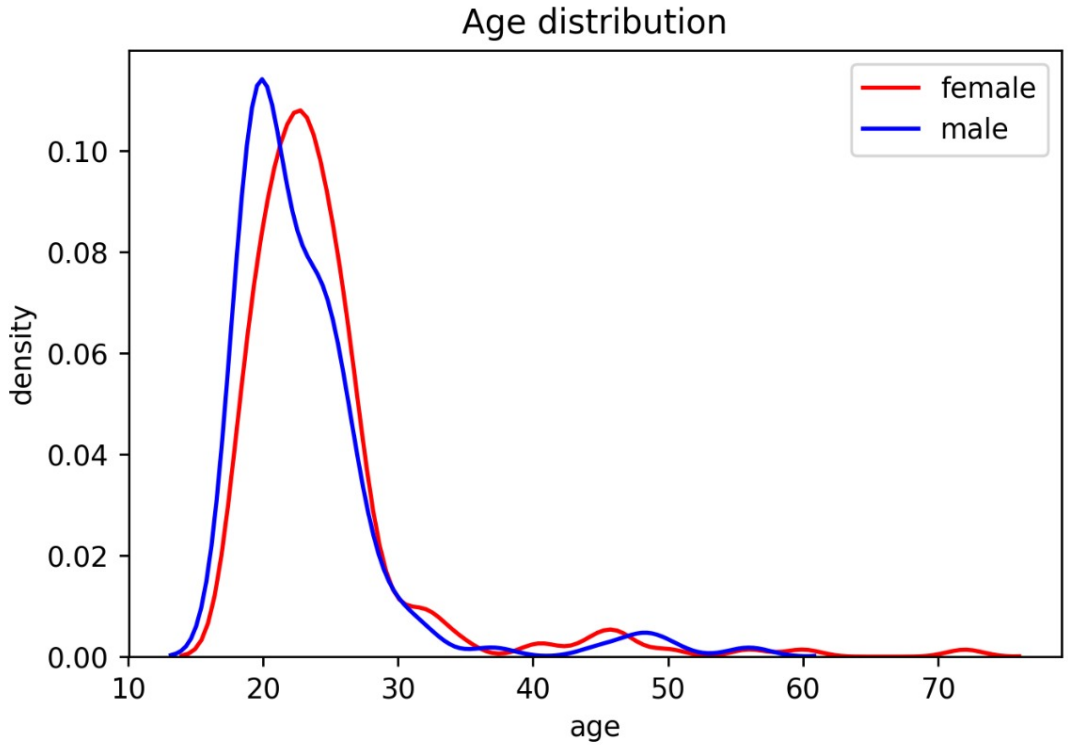

Figure 2: Age distribution of samples for males and females.

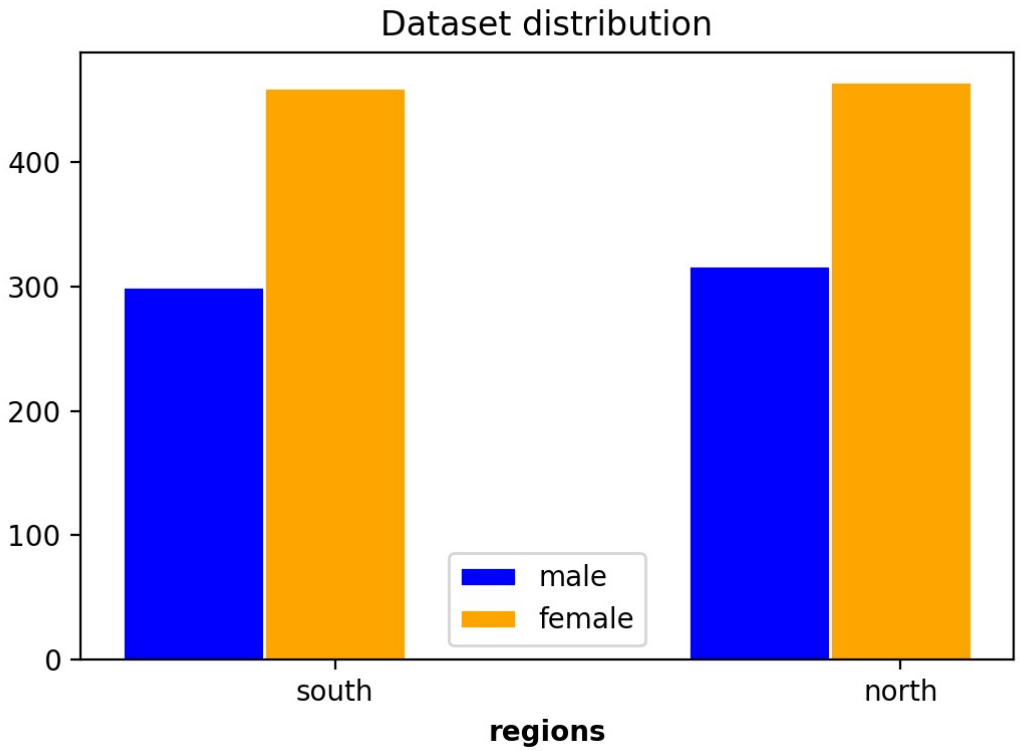

Figure 3: Gender distribution of samples.

Region	Gender	Total
North	Male	317
	Female	465
South	Male	300
	Female	460

Figure 4: Gender distribution of samples.

4.2 Image Preprocessing

The images were preprocessed in the CNSIFD dataset and packed as MATLAB arrays. We extracted the grayscale images from the MATLAB format and converted it to CSVs and JPEGs. The input facial images being of varying sizes needed to be normalized so that each training data-point would have the same dimension. Each faced was normalised by rotation and scaling such that the mid- point between the eyes coincided across faces and the vertical distance from chin to eyebrow became 250 pixels without altering the aspect ratio because a larger image significantly bogged down the computational speed for neural network approaches. Afterwards different preprocessing methods were applied, including converting to gray scale and mean subtraction. Converting images to gray scale is a powerful way of eliminating the differences of brightness and contrast in the input. Converting the input images to gray scale, reduced the input dimension by 3 times. However this was not helpful for us as pretrained weights available used in transfer learning are trained on colorful images only. Since pretrained models in transfer learning are trained on colourful images only, we converted the grayscale image to JPEG format which automatically assigned RGB values to the image.

Chapter 5

Methods

5.1 AlexNet

AlexNet is a convolutional neural network that is trained on more than a million images from the ImageNet database. The network is 8 layers deep and can classify images into 1000 object categories, such as keyboard, mouse, pencil and several different species of animals. It is composed of 5 convolutional layers followed by 3 fully connected layers, as depicted in Figure 5. AlexNet, proposed by Alex Krizhevsky, uses ReLu (Rectified Linear Unit) for the non-linear part, instead of a tanh or sigmoid function which was the earlier standard for traditional neural networks. The advantage of ReLu over the sigmoid function is that it trains much faster than the latter. This is due to the fact that the derivative of sigmoid becomes very small in the saturating region and therefore the updates to the weights almost vanish. Another problem that this architecture solved was that it reduced overfitting by using a Dropout layer after every FC layer.

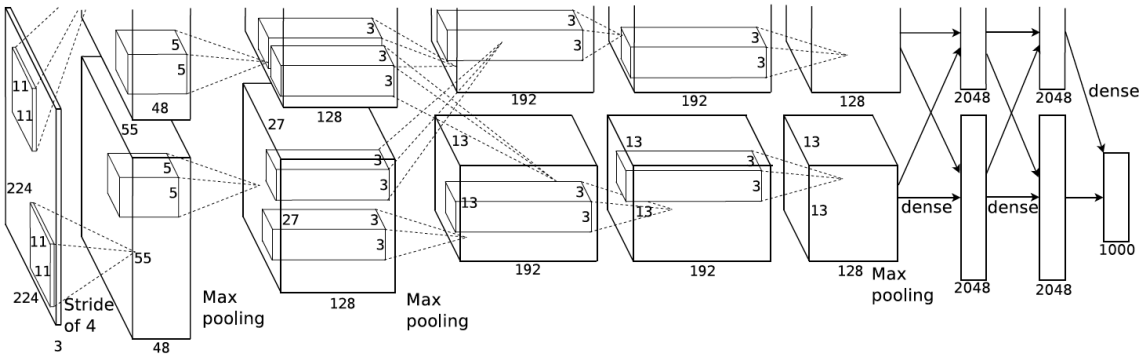

Figure 5: AlexNet neural network structure (source: `http://web.ics.ei.tum.de/~karinne/Pdfs/MasterarbeitICS-FINALVERSION-Niklas.pdf`)

5.2 VGG16 and VGG19

VGG16 is a convolutional neural network model proposed by K. Simonyan and A.Zisserman from the University of Oxford in the paper "Very Deep Convolutional Networks for Large-Scale Image Recognition"[13]. The input to Conv 1 layer is of fixed size 224 x 224 RGB image. The image is passed through a stack of convolutional (conv.) layers, where the filters were used with a very small receptive field: 33. In one of the configurations, it also utilizes 11 convolution filters, which can be seen as a linear transformation of the input channels. The convolution stride is fixed to 1 pixel; the spatial padding of conv. layer input is such that the spatial resolution is preserved after convolution, i.e. the padding is 1-pixel for 33 conv. layers. Spatial pooling is carried out by the max-pooling layers, which follow some of the conv. layers. Max-pooling is performed over a 22 pixel window, with a stride of 2. Three Fully-Connected (FC) layers follow a stack of convolutional layers (which has a different depth in different architectures): the first two have 4096 channels each, the third performs 1000-way ILSVRC classification and thus contains 1000 channels (one for each class). The final layer is the soft-max layer. The configuration of the fully connected layers is the same in all networks.

VGG19 has a similar model architecture as VGG16 with three additional convolutional layers, it consists of a total of 16 Convolution layers and 3 dense layers. Figure 8 shows the architecture of VGG19 model. In VGG networks, the use of 3 x 3 convolutions with stride 1 gives an effective receptive field equivalent to 7 x 7.

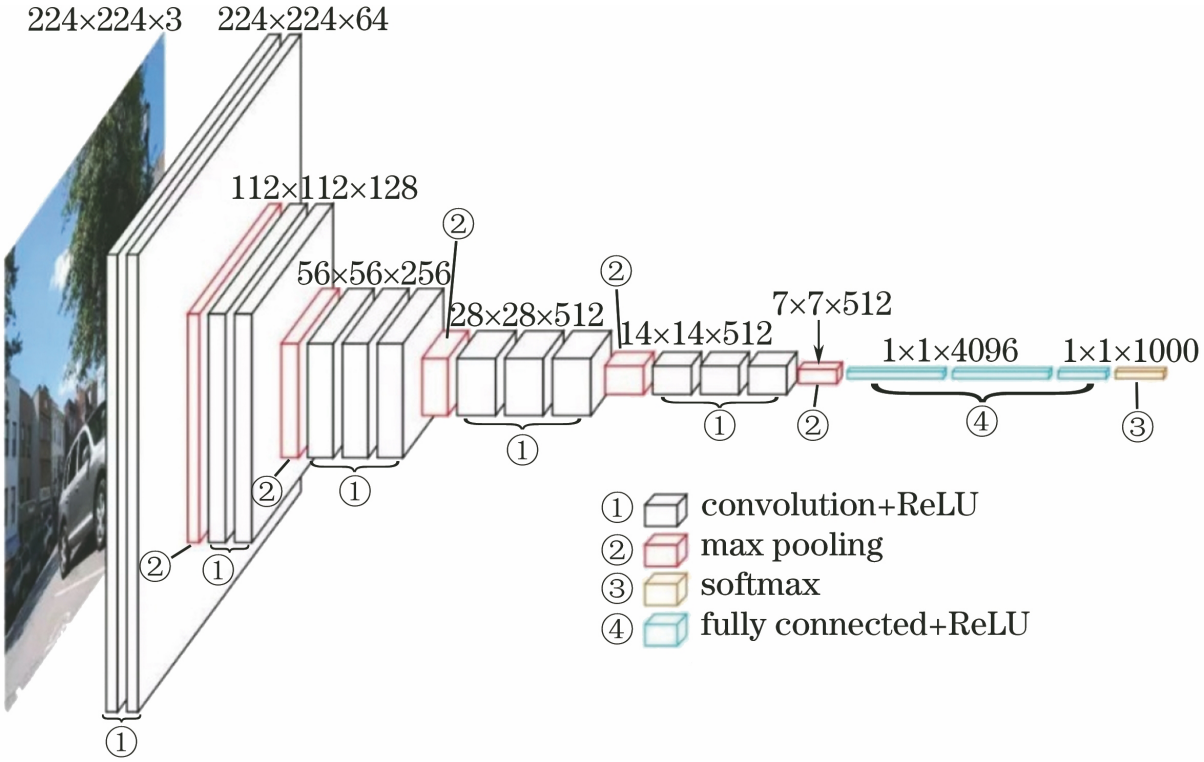

Figure 6: VGG16 neural network structure (source: `https://neurohive.io/en/popular-networks/vgg16/`)

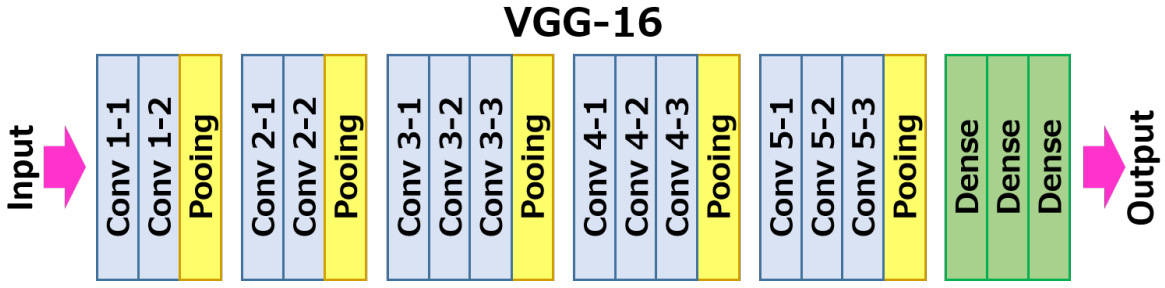

Figure 7: VGG16 architecture (source: `https://www.sciencepubco.com/index.php/ijet/article/download/18588/8470`)

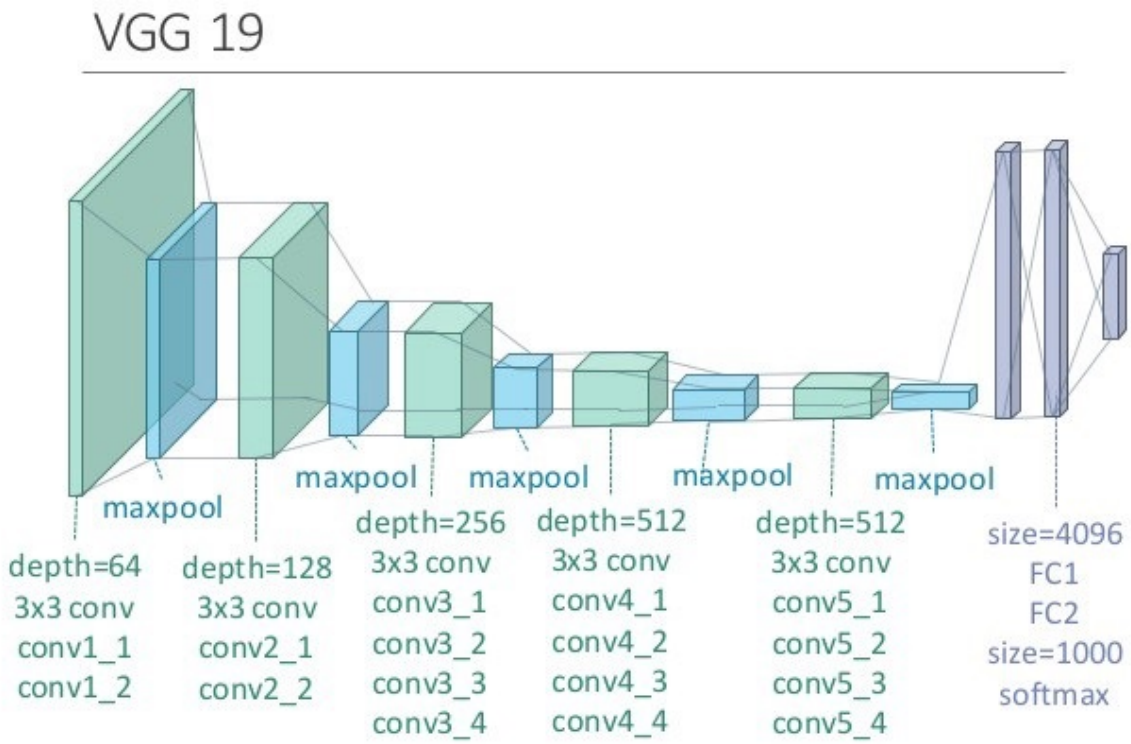

Figure 8: VGG19 architecture (source: `https://medium.com/machine-learning-algorithms/image-style-transfer-740d08f8c1bd`)

5.3 ResNet

All the previous models used deep neural networks in which they stacked many convolution layers one after the other. It was learnt that deeper networks are performing better. However, it turned out that this was not really true.
Following are some of the problems associated with deeper networks:

1. The network becomes difficult to optimize
2. Vanishing/Exploding Gradients
3. Degradation Problem (accuracy first saturates and then degrades)

So to address these problems, authors of the ResNet architecture came up with the idea of skip connections with the hypothesis that the deeper layers should be able to learn something as equal as shallower layers. A possible solution is copying the activations from shallower layers and setting additional layers to identity mapping. These connections are enabled by skip connections which are shown in Figure 9. So the role of these connections is to perform identity function over the activation of shallower layer, which in-turn produces the same activation. This output is then added with the activation of the next layer. To enable these connections or essentially enable these addition operation, one need to ensure the same dimensions of convolutions throughout the network. This is why ResNets have the same 3 by 3 convolutions throughout.

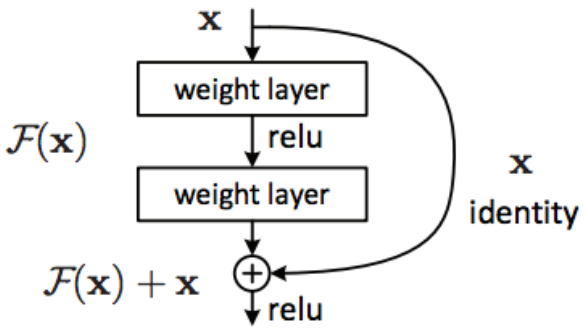

Figure 9: ResNet identity block (source: `https://neurohive.io/en/popular-networks/resnet/`)

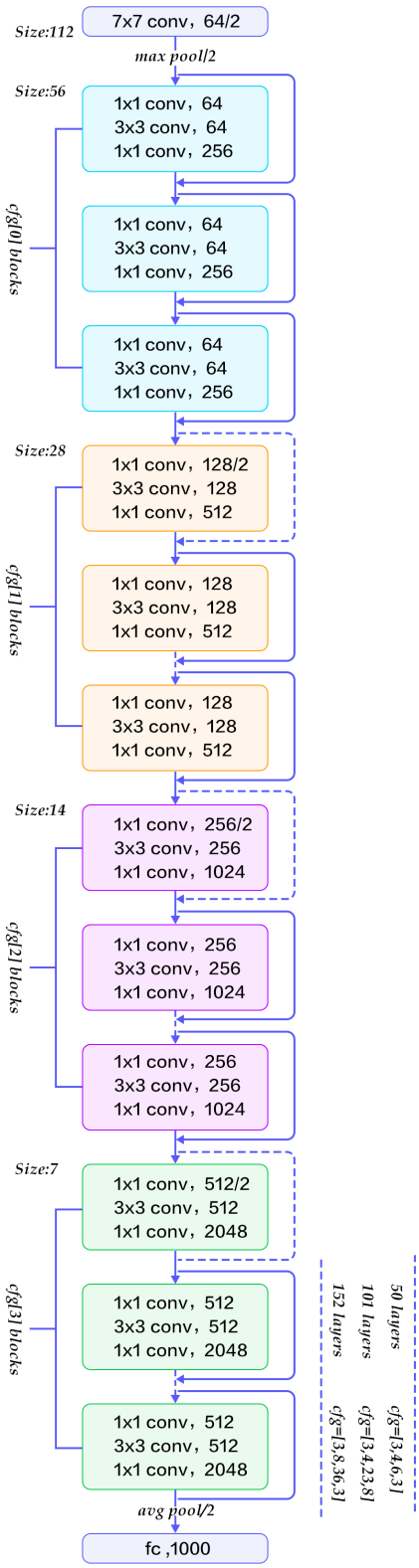

Figure 10: ResNet full architecture (source: `http://www.cse.ohio-state.edu/~panda/5194/slides/3.a-3.b.caffe_caffe2.pdf`)

Chapter 6

Results and Analysis

The training process of ResNet50 for 12 epochs with manually specified learning rates has been summarised in Figure 11. The top losses for the same has been shown in Figure 12.

The confusion matrices for the different models - AlexNet, VGG16, VGG19, ResNet50 and ResNet152 - obtained for the classification task are summarised below.

Figure 18 shows the comparison of the losses in the training dataset and validation dataset. The blue bar shows the loss in the training set and the red bar shows the loss in the validation set. The chart concludes that the losses are almost equal.

Figure 19 shows the comparison of all the accuracy values obtained from the different models used. It can be observed that ResNet50 and VGG16 give the best accuracy values.

Figure 20 visualises data in the confusion matrices in an intuitive way. Half of the graph is for the images that are labeled as north and the second half is for the images that are labeled as south. The blue and yellow portions show how many images the models got the correct prediction for and the red and green portions show how many mistakes the models made.

```
[105] learn.fit_one_cycle(12, max_lr=slice(1e-5,1e-4))
```

epoch	train_loss	valid_loss	accuracy	time
0	0.779558	0.667272	0.593103	00:27
1	0.794279	0.643220	0.641379	00:27
2	0.802565	0.612118	0.700000	00:28
3	0.765868	0.610731	0.679310	00:27
4	0.731630	0.602753	0.693103	00:26
5	0.698876	0.621028	0.682759	00:27
6	0.681923	0.612765	0.710345	00:26
7	0.646857	0.611278	0.717241	00:28
8	0.612747	0.601574	0.727586	00:27
9	0.589422	0.584658	0.724138	00:26
10	0.587082	0.588168	0.724138	00:26
11	0.565178	0.591550	0.720690	00:27

Figure 11: Training Process of ResNet50 with 12 epochs

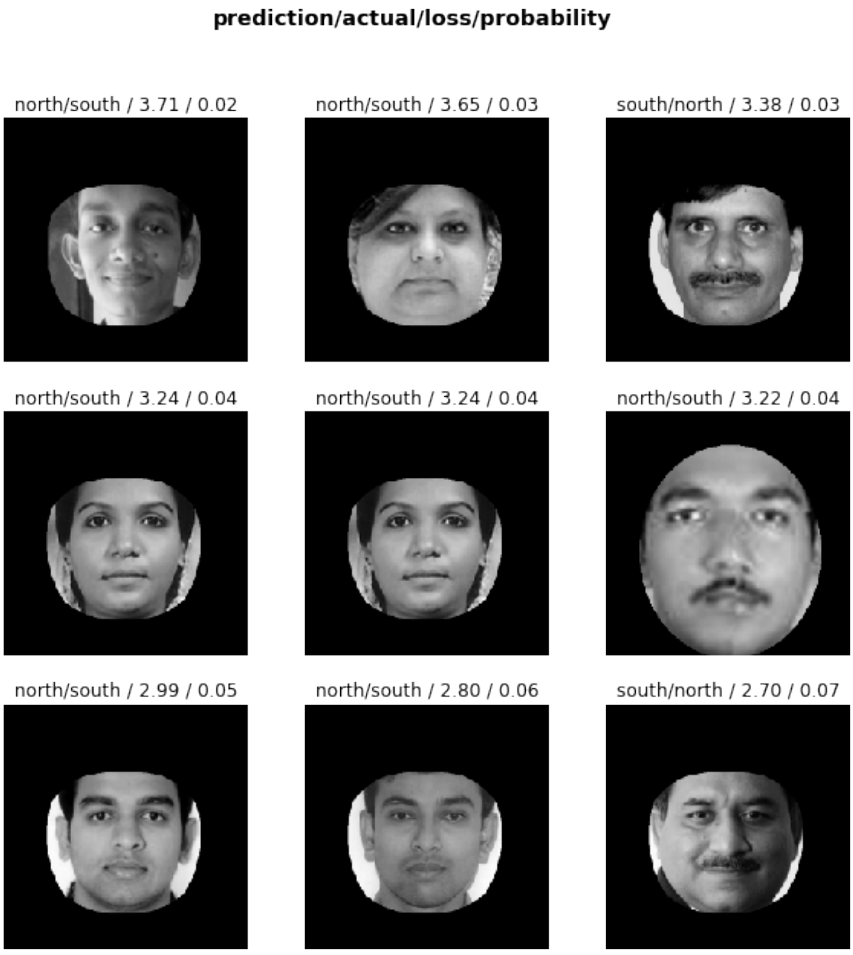

Figure 12: The top losses for ResNet50

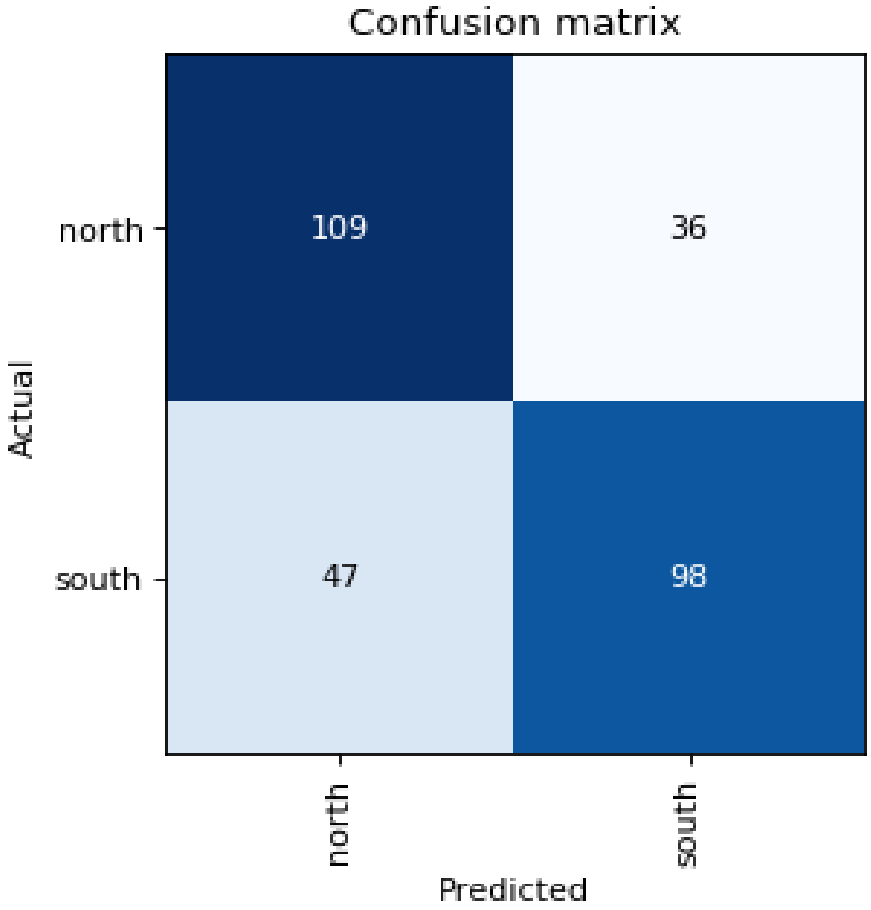

Figure 13: AlexNet Confusion Matrix

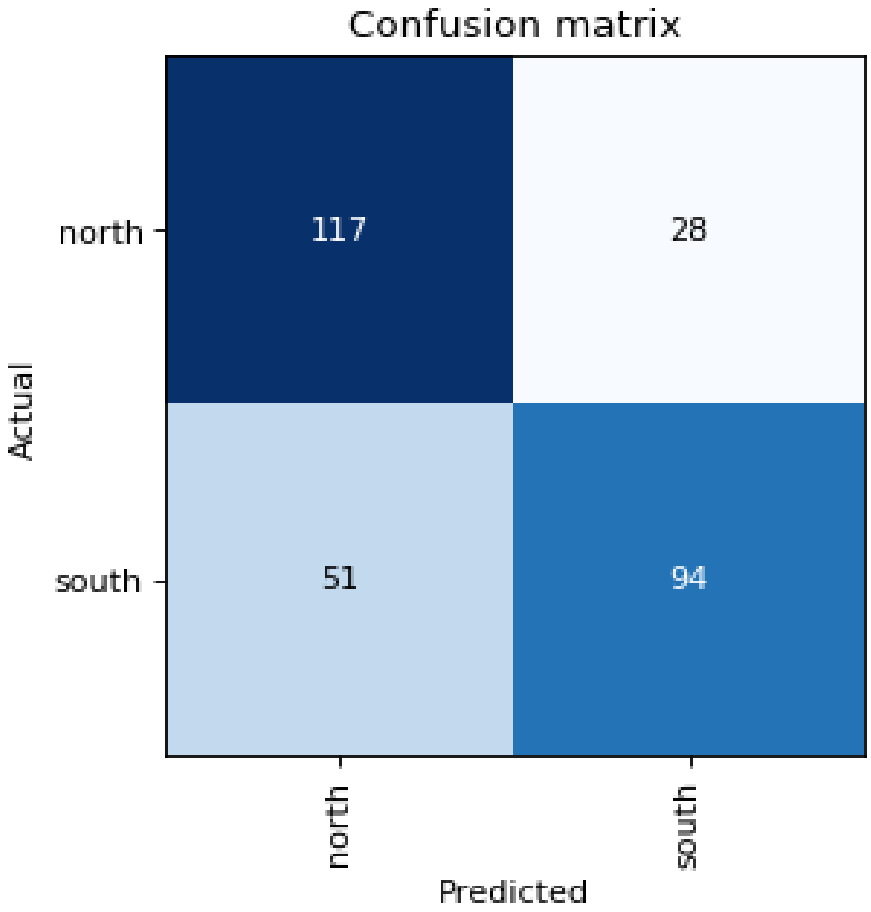

Figure 14: VGG16 Confusion Matrix

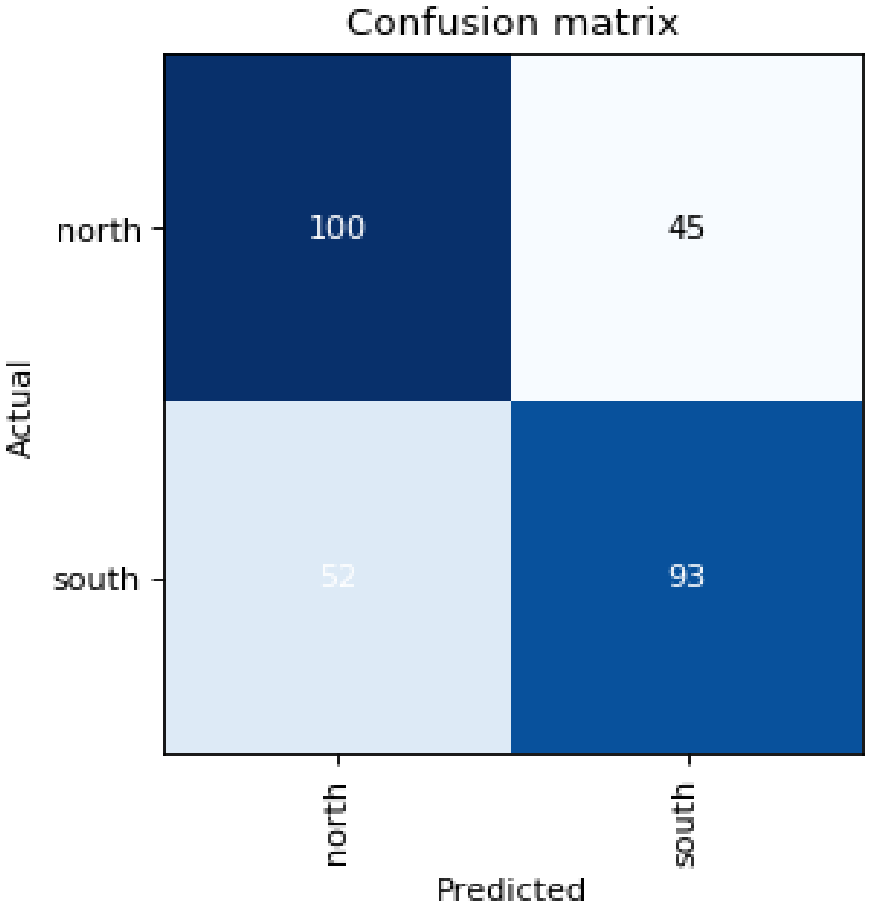

Figure 15: VGG19 Confusion Matrix

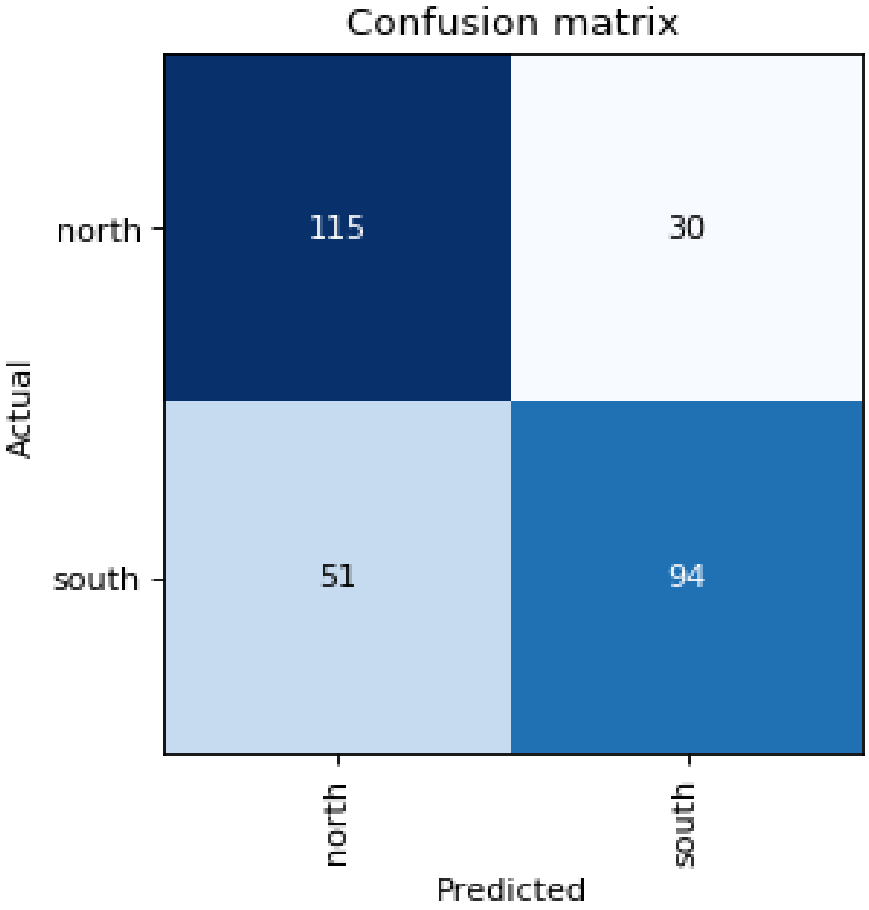

Figure 16: ResNet50 Confusion Matrix

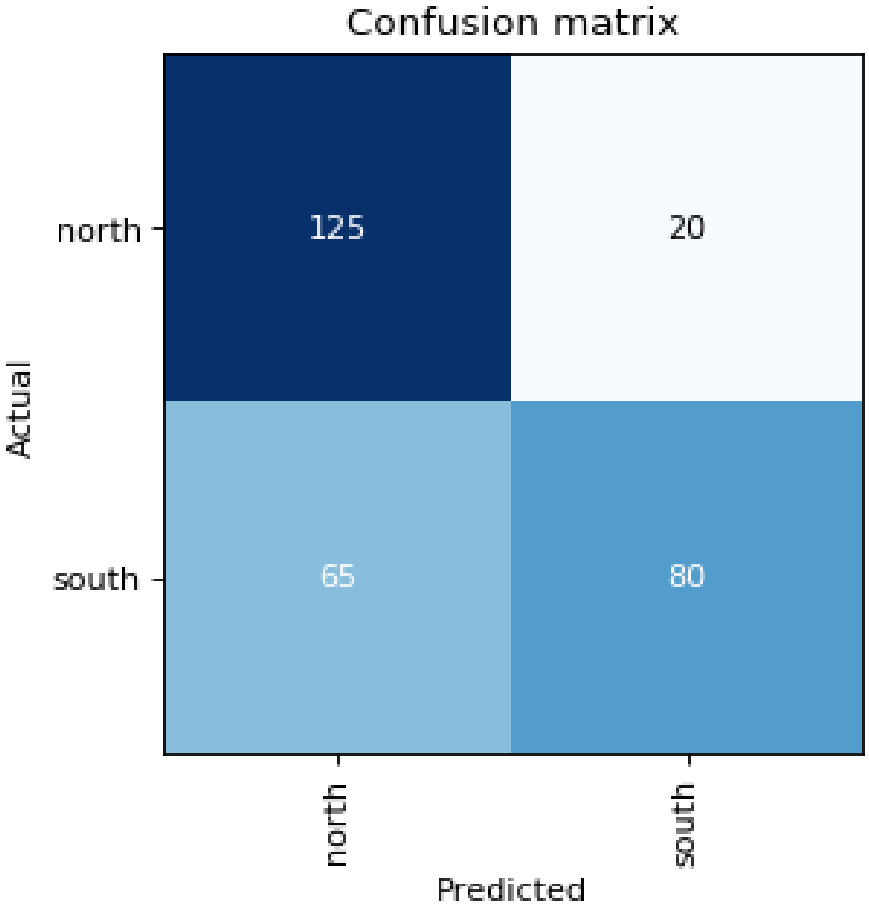

Figure 17: ResNet152 Confusion Matrix

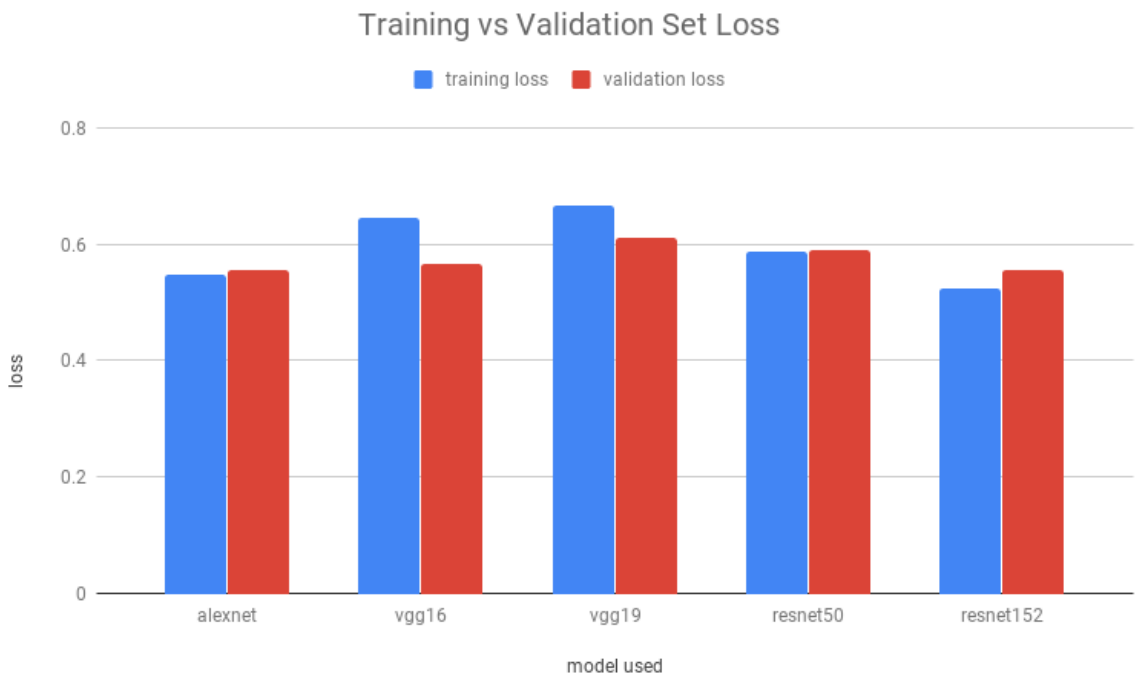

Figure 18: Training set and validation set loss comparison

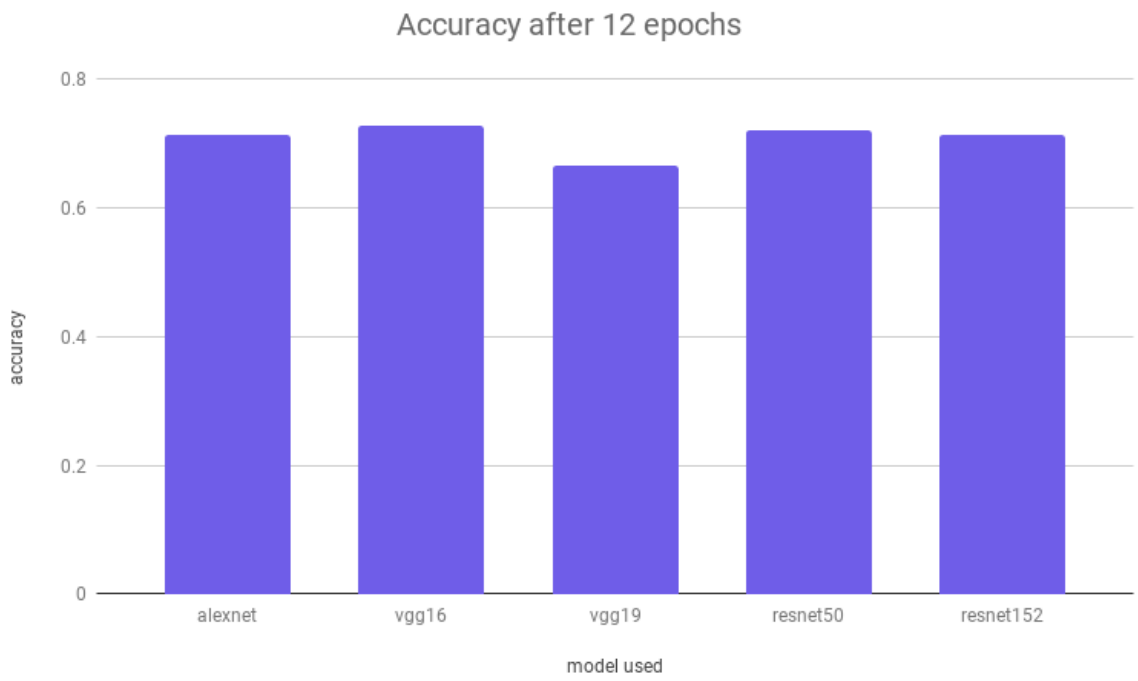

Figure 19: Accuracy comparison by model

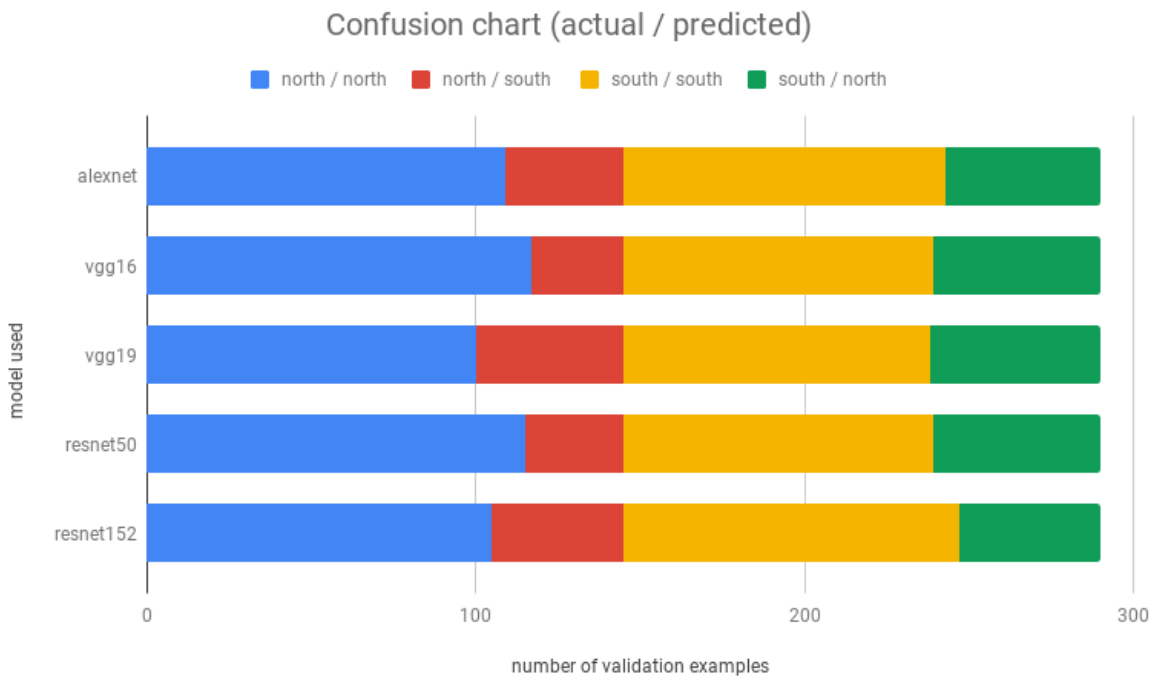

Figure 20: Confusion chart comparison by model

Chapter 7

Conclusion and Future Work

7.1 Conclusion

We have characterized machine performance on a hard race classification task using VGGs, ResNets and AlexNet. Our main finding is that many different computational models can achieve human levels of performance (64% [9]) or even better. Assuming that the kind of faces in our dataset are similar to the training examples experienced by humans, this raises the interesting question of what features humans extract from faces and how they learn it.

7.2 Future Work

Fine grained race classification can be improved with more advanced architectures that permit us to build more deeper neural nets. This classification can help us reveal the extent to which our face structures depend on our geolocations. Also extending this study to finer grained classification can reveal how much intermingling of facial features is present as a function of distance. Studying finer-grained classification can greatly improve coarse grained classifications such as Asian, African, Hispanic, etc. These classifications might help in revealing similarities between ethnic groups and thus giving more clues to the spatial evolution of ethnic groups over time.

References

[1] Create an algorithm to distinguish dogs from cats, https://www.kaggle.com/c/dogs-vs-cats

[2] Determine the breed of a dog in an image, https://www.kaggle.com/c/dog-breed-identification

[3] P. Viola and M. Jones, "Robust Real-Time Face Detection", In Proceedings of International Journal of Computer Vision, May 2004, Volume 57, Issue 2, pp 137154.

[4] Xiaoguang Lu and Anil K. Jain, "Ethnicity identification from face images", Department of Computer Science & Engineering, Michigan State University. In Proceedings of SPIE 5404, Biometric Technology for Human Identification, 2004.

[5] S. Hosoi, E. Takikawa and M. Kawade, "Ethnicity estimation with facial images, Automatic Face and Gesture Recognition", 2004. Proceedings. Sixth IEEE International Conference on, 2004, pp. 195-200.

[6] G. Guo and G. Mu, "A study of large-scale ethnicity estimation with gender and age variations", 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, San Francisco, CA, 2010, pp. 79-86.

[7] S. M. M. Roomi, S. L. Virasundarii, S. Selvamegala, S. Jeevanandham and D. Hariharasudhan, "Race Classification Based on Facial Features, Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG)", 2011 Third National Conference on, Hubli, Karnataka, 2011, pp. 54-57.

[8] FERET Database for facial images. http://user18808:Gz1Bv4Ft@nigos.nist.gov:8080/colorferet/

[9] Centre for Neuroscience Indian Face Dataset, https://github.com/harish2006/CNSIFD

[10] Yu Yan, "Is he Chinese, Korean or Japanese?", Stanford University, 450 Serra Mall, Stanford, CA 94305, 2015.

[11] The Chicago Face Database, https://chicagofaces.org/default/

[12] The CAS-PEAL face database, http://www.jdl.ac.cn/peal/Home.htm

[13] Karen Simonyan, Andrew Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", 4 Sep 2014, https://arxiv.org/abs/1409.1556