Inter IIT Tech Meet 10.0

Bosch Age and Gender detection

By

Team-13

21 March, 2022

Contents

List of Figures

Chapter 1

Introduction

1.1 Overview of Problem

Built a solution to estimate the age and gender of people detected from a surveillance footage feed obtained from places like malls, retail stores, hospitals, etc. Considering low-resolution cameras, camera quality, camera position, light condition, and occlusions.

1.2 Motivation of the problem

Nowadays surveillance systems are deployed at many places (like malls, retail stores, hospitals, etc.) for security purposes. Over the recent decade, detecting human beings in a video scene of a surveillance system is attracting topic of research due to its wide range of applications in human gait characterization, human-computer interaction, abnormal event detection, surveillance monitoring, person counting in a dense crowd, person identification, gender and age classification, fall detection for elderly people, video content analysis, biometrics, targeted advertising and entertainment, etc.

The task of person retrieval in the video is challenging due to camera quality, camera position, light condition, and occlusions. Most of the scenes captured are from a static camera with minimal background change. Frames obtained are with low resolution. Often objects are detected as far-field in outdoor

surveillance systems.

Commonly, person tracking is done by manually searching through videos. However, there are limitations in the human capacity to monitor simultaneous events in surveillance displays. Hence, In the computer vision and pattern recognition area, Automatic analysis of human motion from CCTV footage has become one of the hottest and most attractive research topics.

Chapter 2

Background Research

2.1 Literature Survey

In order to implement an optimal pipeline, we needed to look at existing literature and works. Going through the various available literature enabled us to explore many unique datasets and models.
In order to develop the optimal pipeline, we started to look into various neural networks primarily for identifying the face/body from the CCTV footage, super resolving the image thus obtained, and then CNN models to classify the age/gender of the person identified previously. First we needed to find a suitable model/literature for the detection of a human in the footage.
We went through a lot of research papers but we weren't able to find any models good enough for detection of pedestrians initially. We then came across Zhao et al. which uses Faster RCNN based systems in order to detect objects(in this case pedestrians) with a high level of accuracy.

We then needed to super resolve the image to enhance the image and its features for better identification of the person. First, we went through the original paper provided in the event description; it showed us the internal working of the SRGAN and how it applies a deep network in combination with an adversary network to produce higher resolution images. We then looked for a pre-trained model of SRGAN that we could apply to our footage. We found several of them, like BasicSR, GFPGAN, and Real-ESRGAN, variations of SRGAN. They used architecture similar to SRGAN and had much better results adapted to real-life datasets like that of enhancing CCTV im-

ages. Real-ESRGAN especially had exceptional results compared to the other three when adapted to the given scenario.

For the third part of the pipeline, i.e. identifying the age and gender of the person shown in the footage, we went through the literature on identifying the person from face, body, and gait. In current literature and research, it was found that face and gait gave the highest accuracy while classifying age/gender, as mentioned in Geelen et al. Geelen et al. used conventional machine learning models (Support Vector Machine, Random Forest)rather than neural networks and reported an excellent accuracy of 89 percent.
Upon further investigation into the literature, we went through many papers on classification using the person's gait or through various other methods such as estimating the gender from their motion (Jang et al.). Then, we came upon a fascinating method that separated the various parts of the body (i.e., face, upper body, midsection, lower body) and trained different models on each part independently and then took the average score and classified the gender of the person.
We also came across a pre-trained state-of-the-art model (Tae et al.), which used a multi-layer perceptron with various other modifications like skip-connections batch normalization. This model was very well built for the classification of age and gender. The Adience dataset was used to train it is oriented more towards East Asian faces like those belonging to China, the Philippines, and Korea, and hence, it gave terrible results when tested on Indian faces/datasets. We also found many models using pre-built models like Resnet ( ).
After going through the literature, we concluded that when it came to gender classification, a majority of the literature/works used models like GoogleNet, VGG, Resnet, FaceNet, and models with similar architecture in order to classify and did not give much difference in results when compared with state-of-the-art models like that from Tae et al. The results of Tae et al.

Chapter 3

Design and Implementation

3.1 Methodology

The approach is made to do three tasks, body and face detection, super-resolving the bounded images, and followed by age and gender prediction. We divide the task into a pipeline of 3 models for the following tasks:

Model 1: This deals with the bounding box detection of the body in the video frames or the image. It uses a Faster RCNN approach and body bounded image ahead.

OpenCV Caffemodel: OpenCV CaffeModel is the improved version of the OpenCV Haarcasscade Model. It isolates the faces from the body image found from the frames of the video. If the face is not detected due to the face visibility of the person or the person's resolution, then we upload the face as a zero tensor ahead which means it will not have an effect ahead. Pretrained weights of Caffemodel are taken from the official documentation.

Model 2: This deals with the super-resolving and upscaling of the image. The ESRGAN model, the improved variant of SRGAN algorithms, is used. Both Face and Body images are passed forward in M2 separately to get the super-resolved face and body images. Globally available implementation of the real ESRGAN is used.

Model 3: This deals with age and gender prediction through the face images, body images, and age/gender labels. M3 is the combination of two

models, one is the Age/Gender Prediction CNN trained on the face images only, and the second is also an Age/Gender Prediction CNN but trained on the body images only. Both models are used as feature extractors only (by removing the last classification layers), and the features of both models are concatenated. Finally, two sequential layers of Age and Gender prediction are connected to do the age and gender classification using the concept collaborative learning approach(explained ahead). All the models are trained from scratch.

3.2 Model Selection

Model 1:

For Model 1, We tried out various models and Fast R-CNN not only had the highest accuracy but also was the fastest compared to all the other models we tried. The comparison between various models and the performance metrics are reported later on in this documentation giving deeper insights into our choices.

Model 2:

For Model 2, We went through many SRGAN models and variants. Of those, two of them primarily stuck us to very relevant. The first one was Real-ESRGAN. This model extends the powerful ESRGAN to a practical restoration application (namely, Real-ESRGAN), which is trained with pure synthetic data. And the second one was GFPGAN, which aims at developing a Practical Algorithm for Real-world Face Restoration. It leverages rich and diverse priors encapsulated in a pre-trained face GAN (e.g., StyleGAN2) for blind face restoration. We initially inclined towards GFPGAN, which was specifically developed to restore faces in real-world applications. However, after a few trials on the PETA dataset and other CCTV footage, we realized that GFPGAN performed significantly worse than Real-ESRGAN when we tried to super resolve the person's body. We also found that Real-ESRGAN performed just as well at restoring people's faces from CCTV footage compared to GFPGAN as the latter requires a particular resolution to perform significantly better than Real-ESRGAN, which unfortunately could not be provided by the input footage. For these very reasons, we decided to choose the Real-ESRGAN over GFPGAN.

Model 3:

Coming to the final Model of the pipeline, for this we had 2 train 2 different models for the body and the face respectively. We tried out various models including multiple variants of Resnet like Resnet 18 and Resnet 50 along with models from Scratch and FaceNet for the face classifier while we tried variants of Resnet, VGG-16 and a few other models made scratch in order to boost the accuracy.

3.3 Dataset Selection

Model 1:

COCO dataset is used for training the M1 model. Pretrained weights of COCO dataset applied on Faster RCNN is used. COCO is large-scale object detection, segmentation, and captioning dataset. COCO has several features: Object segmentation, Recognition in context, Superpixel stuff segmentation, 330K images (¿200K labeled), 1.5 million object instances, 80 object categories, 91 stuff categories, 5 captions per image, 250,000 people with key points.

Model 2:

ESRGAN is trained on the 3 dataframes, DIV2K, Flickr2K, and OST dataset.

DIV2K: The DIV2K dataset is divided into:

- Train data: starting from 800 high definition high-resolution images we obtain corresponding low-resolution images and provide both high and low-resolution images for 2, 3, and 4 downscaling factors.
- Validation data: 100 high definition high-resolution images are used for generating low resolution corresponding images, the low res is provided from the beginning of the challenge and are meant for the participants to get online feedback from the validation server; the high-resolution images will be released when the final phase of the challenge starts.
- Test data: 100 diverse images are used to generate low resolution corresponding images; the participants will receive the low-resolution images

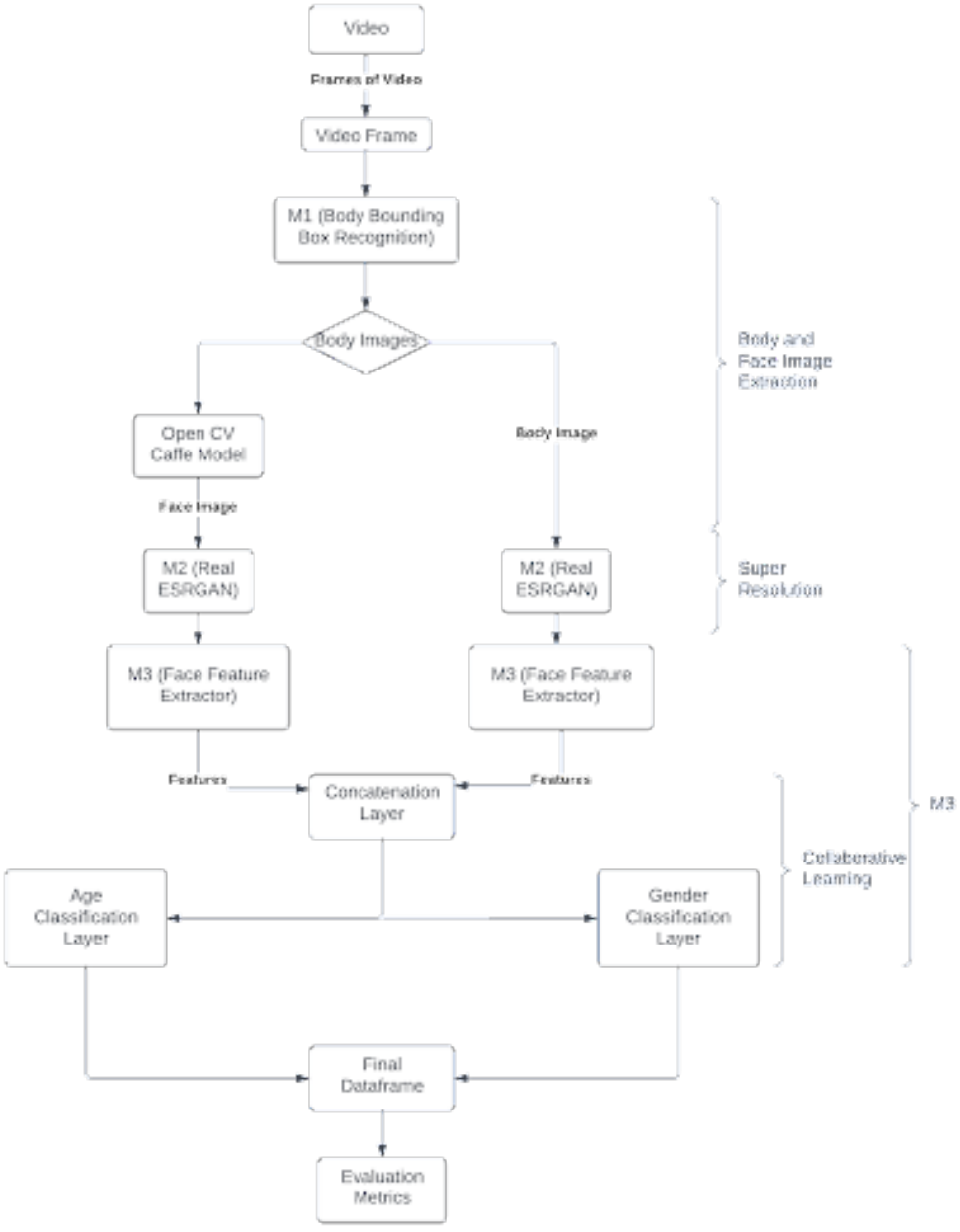

Figure 3.1: Bird's Eye view of the pipeline

when the final evaluation phase starts and the results will be announced after the challenge is over and the winners are decided.

FLICKR2K: The Flickr30k dataset contains 31,000 images collected from Flickr, together with 5 reference sentences provided by human annotators.

QST: OST (Outdoor Scenes), OST Training,7 categories images with rich textures. OST300 300 test images of outdoor scenes.

Model 3: In order to train the age/gender classifier we went through quite a number of datasets.

ADIENCE: The CAVE dataset/Adience dataset was developed by Open university of Israel. Total number of photos are 26,580.Total number of subjects: 2,284.Number of age groups / labels: 8 (0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, 60-). Gender labels: Yes. In the wild: Yes. Subject labels: Yes. This dataset has a relatively higher number of faces which have a fair complexion in stark contrast to Indian conditions which tend to have a darker skin tone. The dataset was used by Tae et. al. and through that model we realised that this dataset was unsuitable for Indian conditions. Also in Adience there are a lot of age gaps for example there isn't a single photo from the age 32-38 or 43-48 and there exist many such age gaps.

PEdesTrian Attribute dataset: PEdesTrian Attribute dataset, popularly known as PETA, is a mixture of many pedestrian photos from various databases. It has images from 10 different datasets, and it contains CCTV images with annotations of age and gender. It has five classes: Less15, Less30, Less45, Less60, and Larger60. This is the perfect dataset for the problem statement as the photos in this are from various angles and are taken from CCTV footage. This enables us to replicate the conditions specified in the problem statement. We used this dataset to train the body image part of our pipeline mainly, and this was the only dataset available such that the dataset had photos of the body visible. Hence, it was perfect to use to train our body pipeline.

UTK Face dataset: UTKFace dataset is a large-scale face dataset with a long age span (from 0 to 116 years old). The dataset consists of over 20,000 face images with age, gender, and ethnicity annotations. UTKFace dataset has two significant advantages over the Adience dataset. Firstly, the main problem with the Adience dataset was that it lacked diversity while the UTKFace dataset was built considering ethnic diversity, and it is a reasonably

balanced dataset that mitigates class imbalance to a great extent (Karkkainen et al.). Secondly, the UTKFace dataset has images of people ranging from age 0-116 which gives us an option to decide the classes ourselves and also opt for regression if needed, whereas, in Adience, there were many age gaps; for example, there is not a single photo from the age 32-38 or 43-48. Hence, we chose UTKFace to train the Face pipeline of our model in comparison to Adience.

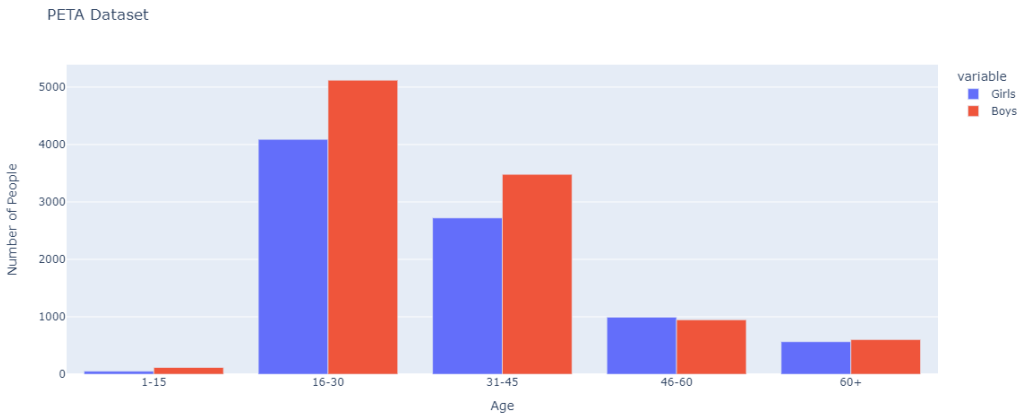

Figure 3.2: This the distribution of the PETA dataset. Here on the Y axis is the count of the people and on the Y axis is the Age

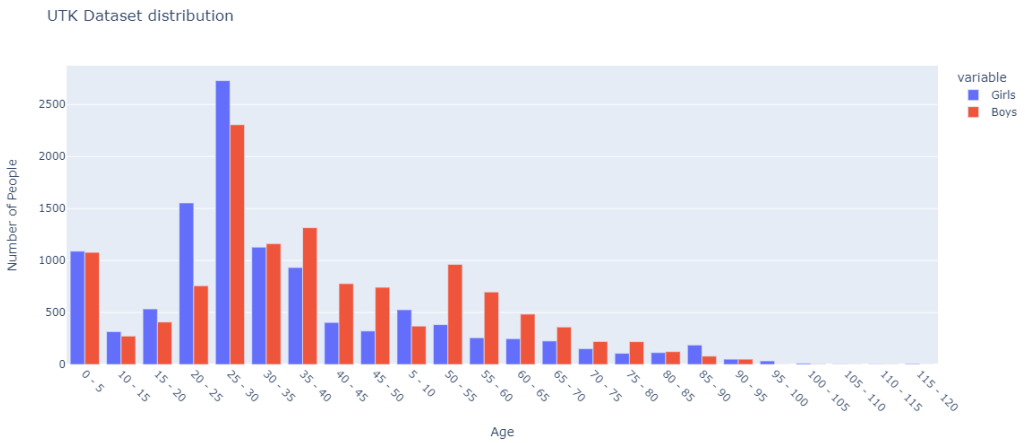

Figure 3.3: This the distribution of the UTK dataset. Here on the Y axis is the count of the people and on the Y axis is the Age

Chapter 4

Results

Note: We were unable to pass it through model 2 due to time constraint and hence the accuracy was considerably less than what we expected

Accuracy for VGG-16: 17 %

Chapter 5

Detailed analysis of all Models

5.1 Detailed Analysis of M1 (Body Detection Model)

For the task of person detection in the camera feed we are using the pretrained model Faste RCNN of pytorch which is pretrained on COCO2015 dataset. Fast R-CNN overcomes several issues in R-CNN. As its name suggests, one advantage of the Fast R-CNN over R-CNN is its speed. Faster-RCNN is faster than usual RCNN models because it shares computations across multiple proposals, thus making it a great choice for human subject detection is video feed where accuracy with low inference time matters a lot.

However, when using the same pre-trained VGG-16 network as a starting point and bounding-box regression, Fast R-CNN achieves 68.1 percent AP, which is a fine improvement over the 63.1 percent mAP obtained by SPPnet, both trained using PASCAL VOC 2007 without "difficult" examples. Not only does Fast R-CNN produce higher accuracy, it is also notably faster. When performing detection using VGG-16, Fast R-CNN can be

3× faster than SPPnet and 9× faster than R-CNN.

It is also worth mentioning that Fast R-CNN is faster to train when compared to R-CNN and SPPnet. When using other feature extractors with better average accuracies such as ResNet-10 and Inception ResNet V2, Faster R-CNN still makes the most accurate model. Later we separate face and body from the person detected to get their separate features and predict age and gender according to both.

One drawback we saw of Faster R-CNN is that for the RPN, all anchors in the mini-batch are extracted from a single image. Because all samples from a single image may be correlated (i.e. their features are similar), the network may take a lot of time until reaching convergence.

5.2 Overview of OpenCV Caffemodel (Face Recognition Model)

To detect faces from the extracted body images from M1, we used pretrained- DNN Face Detector. It is a Caffe model which is based on the Single Shot-Multibox Detector (SSD) and uses ResNet-10 architecture as its backbone.

5.3 Detailed Analysis of M2 (Super Resolving Algorithm)

The body and face images obtained after M1 and OpenCV Caffemodel are separately inputted into M2 to get Super Resolved images. For super resolving and upscaling of images, we used

Real-Enhanced SRGAN (Real-ESRGAN), which is trained with pure synthetic data. The Real-ESRGAN is an extension to ESRGAN, which is a more practical algorithm for real-world image restoration. ESRGAN is an improved version of SRGAN.

5.4 Motivation behind M3 Structure

The inputs given to us would be surveillance video feeds obtained from low resolution cameras placed at a height. Faces of people would not necessarily be visible in the scenes obtained from the videos and the features of the human body extracted from the video feed would not be enough for predicting the age of people. Thus, to be able to predict the age and gender of maximum people from the surveillance video feed, we consider both the face and body images extracted from the video. We use a combination of two feature extractors for age and gender prediction, i.e., a face feature extractor and a body feature extractor, and do collaborative learning to decide on the best parameters for prediction.

5.5 Detailed Analysis of M3 Face Only Model

The M3 face only model will take the super resolved face images from the M2(ESRGAN) and predicts age and gender based on the image along with the output feature from the M3 body only model. We concatenated the output feature vector for the face image from the M3 face only mode with the output feature vector for the body image from the M3 body only model to create a single feature vector for final prediction.

Several pre-existing research works have focused on estimating the age and gender of people from face images. Some of the open-sourced datasets used by them are UTK Face Dataset, Adience dataset, AFAD dataset, FairFace dataset, CelebA dataset, IMFDB dataset, Cave dataset and All Age Face dataset. We used the UTK Face dataset which is for training our Face-Only age and gender prediction model. The original UTK Face dataset treated age prediction as a regression task, however, we are converting it into classification problem with following age classes : 1-5, 6-10, 11-15, 16-20 and so on till 116 years of age. Following are the models we tried for age and gender estimation using face images with the trian validation split ratio as 0.2 :

ResNet-18 pretrained: 25 Percent on Validation ResNet-18 trained from Scratch: 31.93 Percent on Validation ResNet-50 pretrained: 30.87 Percent on Validation FaceNet pretrained (With Augmentation): 35.4195 Percent on Validation

5.6 Detailed Analysis of M3 Body Only Model

M3 body only model will take the super resolved body image from the M2(ESRGAN) model and give the feature vector that we later concatenate with the M3 face only models output. Hardly any research has been done on age and gender prediction using only body images hence there are just one or two open-sourced full-body annotated datasets. We used the PETA dataset to train the body-only age and gender prediction model. The open-sourced PETA dataset has the following age class distributions : 1-15, 16-30, 31-45, 46-60 and 60+. The results obtained from pre-trained models were not satisfactory and we implemented the scratch model architecture as described in the paper Chowdhury et al. . Following are the models implemented

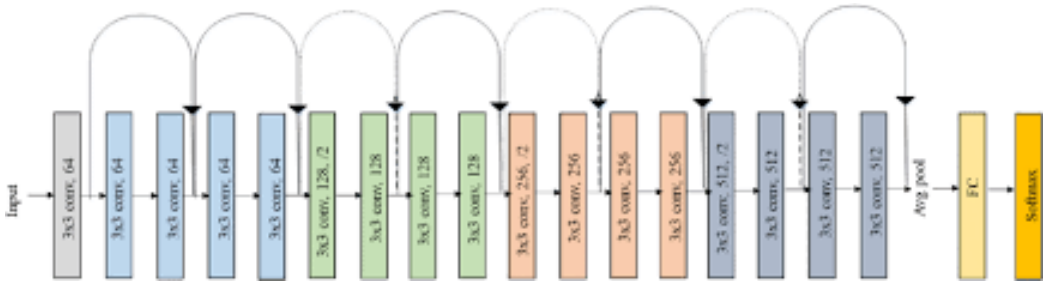

Figure 5.1: Resnet 18 Architecture (Source : Google)

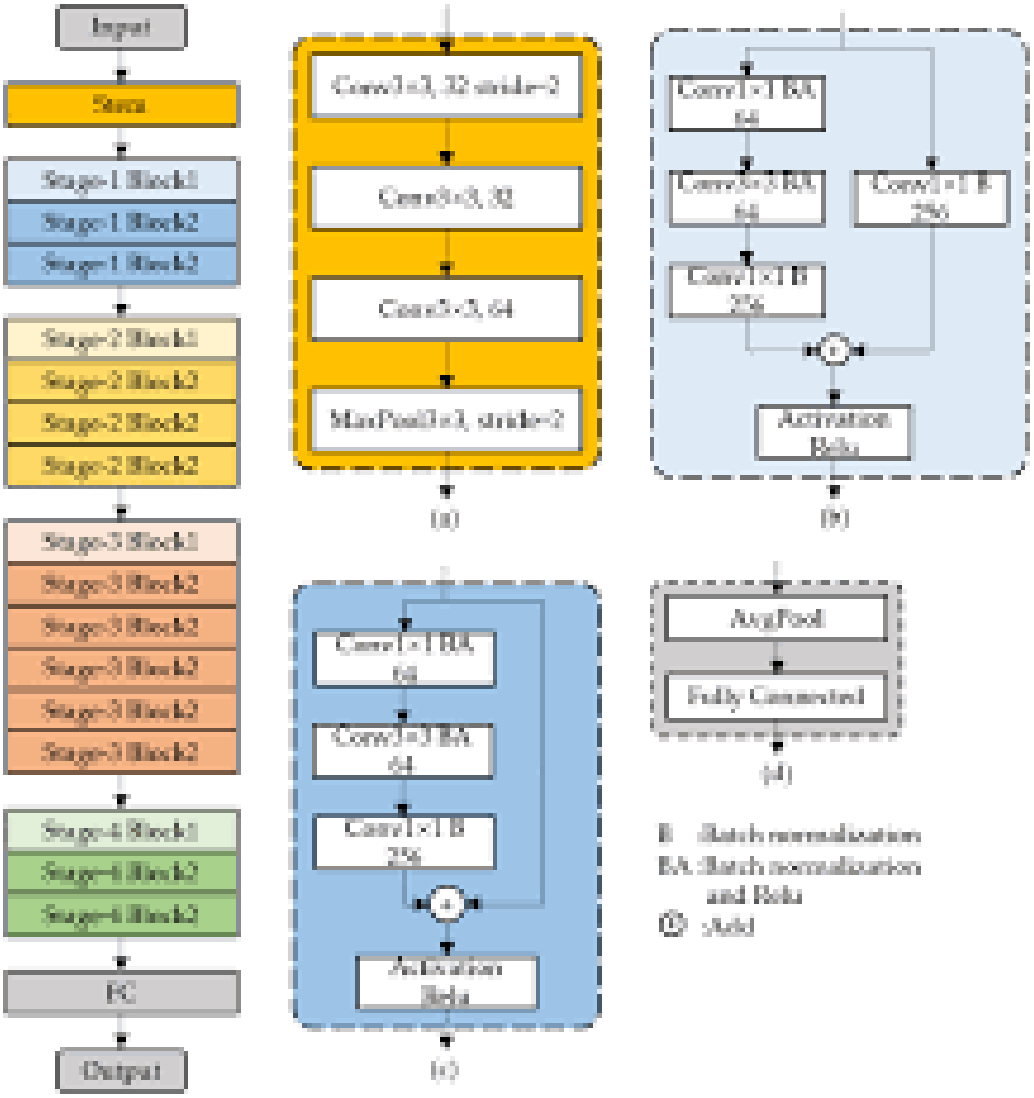

Figure 5.2: Resnet 50 Architecture (Source : Google)

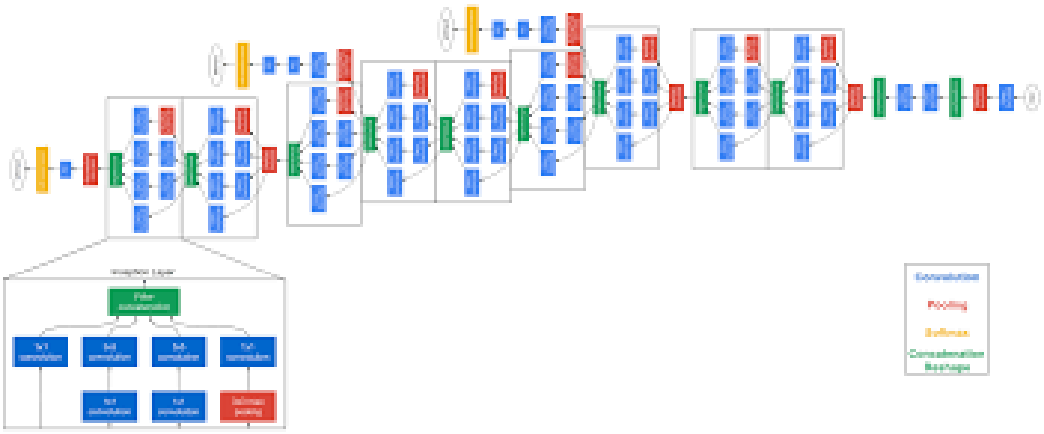

Figure 5.3: Facenet Architecture (Source : Google)

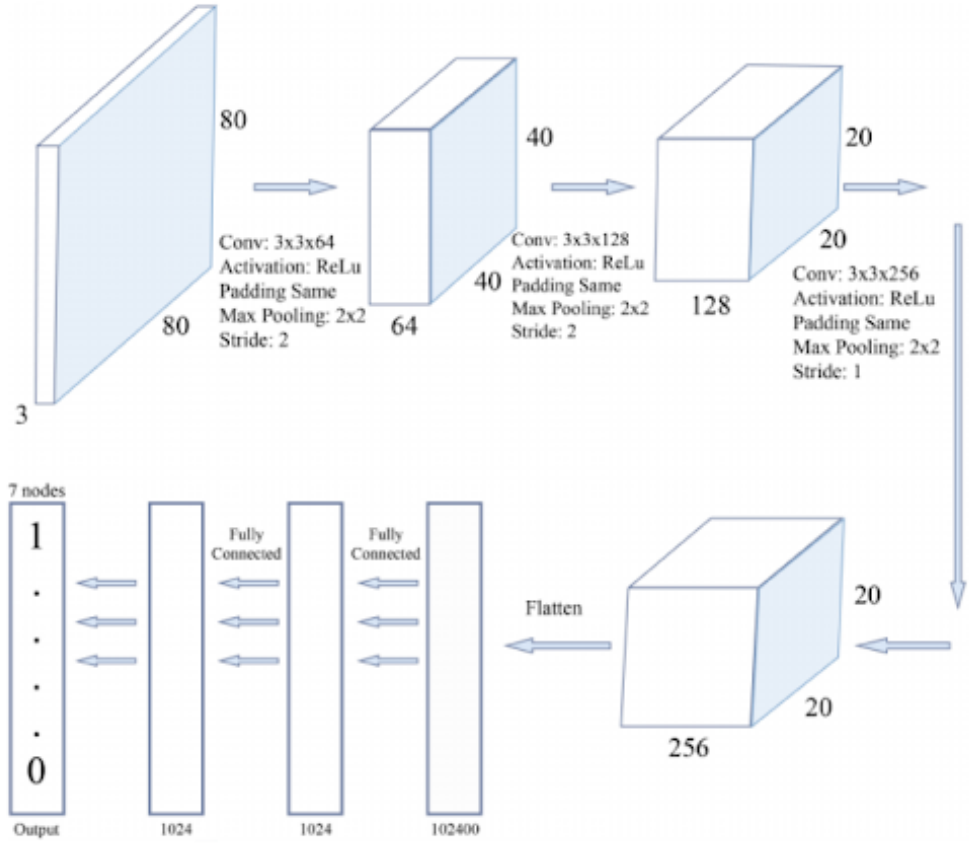

Figure 5.4: Scratch Architecture (Source : Citation 2)

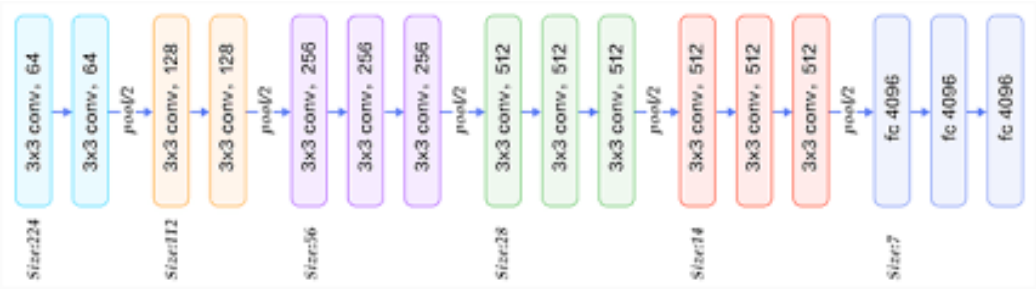

Figure 5.5: VGG-16 Architecture (Source : Google)

:

ResNet-18 pretrained VGG-16 pretrained Scratch model

5.7 Detailed Analysis of Combined M3 (Concatenation of M3 Face only and M3 Body Only Models)

As mentioned previously, to predict the age and gender of maximum people in the surveillance video feed and improve upon the prediction accuracy, we have used a combination of M3 face only and M3 body only feature extractors. We concatenate the feature vectors obtained from the two feature extractors and pass the concatenated vector through linear layers for age and gender prediction. There are two losses, the loss from age classifier and the loss from gender classifier. Through collaborative learning, we initially give equal weight to the two losses for backpropagation. However, if the individual accuracy of age prediction or the individual accuracy of gender prediction is greater than a threshold of 70 percent, we alter the weights of the two losses in 0.9 : 0.1 ratio, i.e. 0.1 is the weight for the loss of gender if the gender prediction accuracy reaches greater than 70 percent. This way, the model is able to learn the best weights for

prediction of both age and gender.

5.8 Advantages of Collaborative learning over Sequential Method on each model

The advantages of Collaborative learning is primarily that we can learn a huge diverse selection of features while using the same backpropogation simultaneously while in sequential learning we can only learn a particular feature from the same backpropogation at a time. This reduces our learning time drastically and also allows us to discover new patterns between various features which would give us a better chance at classification.

Chapter 6

Limitations and Future Improvements:

6.1 Limitations in Pipeline

Model 1:

The model 1 is for pedestrian detection implemented using Faster RCNN. The model gives a good prediction of bounding boxes around the humans which are closer with a very good confidence whereas if the human is at a far distance (after a certain distance in meters) it is unable to detect the persons properly with a good confidence. Thus, objects placed quite far are detected with very less confidence and the bounding boxes are very small in dimensions. So the cropped out images are with very resolution, as a result the quality of the face thus obtained is also not good.

M1 is confidence sensitive. The threshold value for the same is upto user's choice. If user choose the value to be quite high, then number of persons detected will be less on the other hand if the threshold value is quite low then number of false detection

will be less. So, user need to choose the threshold quite wisely. In our case, we have chosen a threshold value that reduces the false detection.

If the background gets camouflage with the person, then the confidence of the bounding box is very poor. It might give several bounding boxes.

In the second part of the M1, we extracted the faces from the body images that were extracted earlier. We used the caffee model for extracting the faces from the human image. But while training model 3, to get the face images, we have used the fact that the face must lie in one third of the top portion of the human detected. We used this concept for training because the peta dataset had very small dimension images that the caffee model was unable to identify.
Model 2:

The result of M1, body images as well as face images, are passed through model 2, that is, Real ESRGAN for enhancing it's quality. If the quality of the images are very less, then the model tends to smoothen the image. The body as well as the face but the face is more vulnerable. The facial features like lips, eyebrows, eyes, nostrils tends to get smoothen. In other way if the image feed had a good quality, then the model is doing a good work by resoluting the face as well as body, though sometime it might smoothen the body sometimes but most of the features remain identifiable.

6.2 Limitations of Collaborative Learning

Collaborative learning is a machine learning architecture for training applied classifier. The primary problem with Collab-

orative learning is the loss function. Sometimes even if one part of the classification is working much better than the other, and the other has a very high loss, this would cause the model to believe that the both the parts have improper weights causing it to back propagate wrong information at times.
We tried to tackle this by changing the loss function such that if the accuracy of either part is greater than a certain threshold we would increase the weight-age given to the loss of the other parameter which would prevent changing of the weights on the former part.

6.3 Future Improvements

For our future improvements, we would primarily like to tackle problems and disadvantages of the various parts of the pipeline. For Model 1 we would like to improve the detection of people far away and prevent the false boxes due to the background. For Model 2, we would like to customise and train it to enable better feature enhancement of the input images from Model 1. This would in turn impact the performance of Model 3 drastically as it would give more definite features to Model 3 which would enable it to learn the class distinctions better. For Model 3, we would like to tackle the problem caused by the loss function and arrive at optimal solution which would capture the prime essence of collaborative learning. This would in-turn drastically change the accuracy and improve it to a great extent. We firmly believe that the pipeline we proposed has great potential.

Bibliography

1. https://ieeexplore.ieee.org/document/9308296
2. https://www.researchgate.net/publication/338171619_Pedestrian_Age_and_Gender_Identification_from_Far_View_Images_Using_Convolutional_Neural_Network
3. https://link.springer.com/article/10.1007/s11042-020-09964-6
4. https://www.researchgate.net/publication/339993494_A_multi-branch_separable_convolution_neural_network_for_pedestrian_attribute_recognition
5. https://arxiv.org/abs/1506.01497
6. https://github.com/Chang-Chia-Chi/Pedestrian-Detection/blob/master/Pedestrian_Detection.ipynb.
7. https://blog.paperspace.com/faster-r-cnn-explained-object-detec#:~:text=Faster%20R%2DCNN%20is%20a,the%20locations%20of%20different%20objects.
8. https://paperswithcode.com/paper/photo-realistic-single-image-s
9. http://mmlab.ie.cuhk.edu.hk/projects/PETA.html
10. https://github.com/Chang-Chia-Chi/Pedestrian-Detection
11. https://github.com/AryaHassanli/Gendage#using-the-model
12. https://susanqq.github.io/UTKFace/

13. https://github.com/xinntao/Real-ESRGAN

14. https://github.com/TencentARC/GFPGAN

15. https://talhassner.github.io/home/projects/Adience/Adience-data.html

16.

17.

18.

19.

20.

21.