IN3060/INM460 Computer Vision Coursework report

- **Student name, ID and cohort:** Sam Clastine Jesumuthu (220038747) - PG
- **Google Drive folder:** https://drive.google.com/drive/folders/1Vg2PJ2o6Fn3Q5m0y_sESoCozkVMgGn-L?usp=sharing

Data

The data consist of 2394 training and 458 testing images, and labelled as 0, 1 and 2 which are no mask is worn, mask is worn and worn improperly respectively. When analysing the label distribution, an imbalance of classes was revealed, most of the classes were belonged to the masked faces class (1) showed in below images for both train and test set.

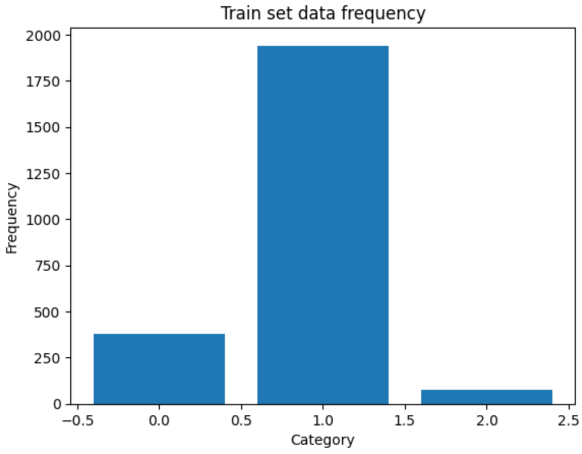


Figure: 01

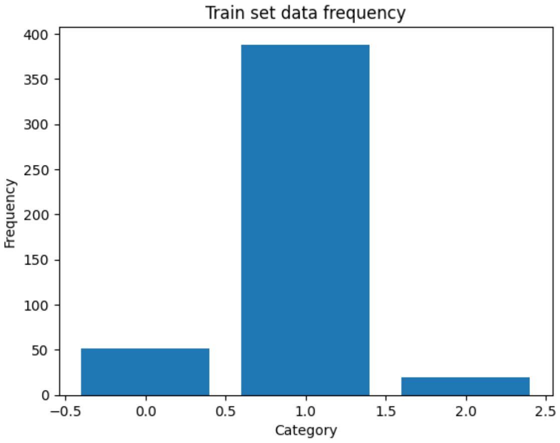


Figure: 02

Several strategies, such as data augmentation, resampling, or the use of multiple assessment criteria, can be used to overcome this issue. Data augmentation techniques, such as flipping, rotating, or resizing the photos, can assist balance the classes by producing new examples from existing data. Resampling approaches can either oversample minorities by duplicating existing data or undersample majorities by deleting part of the data. Oversampling, on the other hand, can result in overfitting, while undersampling can result in the loss of valuable information from the data.

Overall, addressing class imbalance in the dataset is critical to ensuring that the model is not biased towards any single class and can generalise successfully on unknown data.

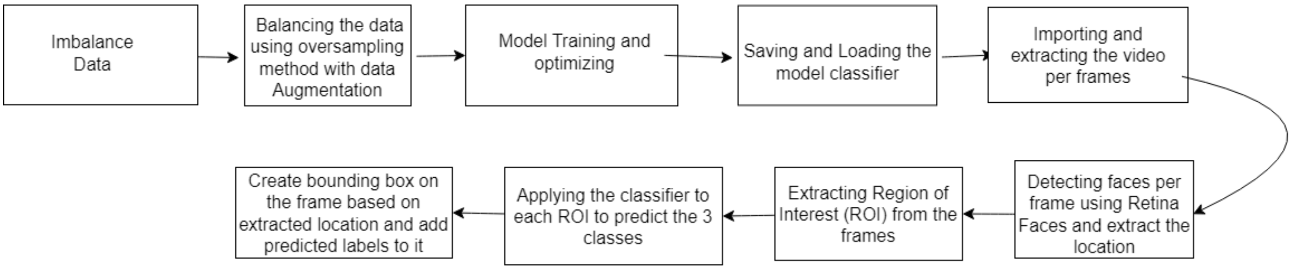
Implemented methods

On the given dataset, at the first step we applied oversampling method with data augmentation on the selective data to balance it. The augmentation method included in this study are rotate the image by 10 degrees, shift the image horizontally and vertically, shear and zoom the image by 20% and flipped the image horizontally, the image is resized (64x64) and normalized based on min max scalar. Splitting the dataset for training (80%) and validation (20%) is done for both balanced and imbalance dataset to validate the performance, so we had trained the models with both the dataset.

The algorithms used for training the mask dataset are HOG + MLP, Baseline CNN and RESNET-16. A *Multilayer Perceptron (MLP)* is a neural network made up of several layers of linked neurons. In our analysis we had use 7 hidden layers, activation function as sigmoid and a dropout after 3rd layer to avoid overfitting. Backpropagation and optimisation techniques like as gradient descent can be used to train the MLP. Implementing Histogram of oriented gradients (HOG) a feature descriptor in MLP helps the classifier to perform good [1]. *Baseline Convolutional Neural Network (CNN)* with one convolutional layer, followed by a max pooling layer, a flatten layer and dense layer for the output. The input size is (64x64x64x3) and output size of the neuron in dense layer is 3. *ResNet-50* contains 50 layers, making it more complex than many existing CNN architectures. This enables it to learn more complicated characteristics from pictures, making it more successful for image classification, object recognition, and segmentation tasks. Furthermore, because of its architecture, which allows for greater gradient flow and faster convergence, it can be taught faster than other deep neural networks.

Face Detection is a crucial stage, so we had used some pretrained models for quick generalization and stable detection. The detectors we had selected for study is Retina face which performed well when compared to Face Cascader and MTCNN. After detecting the face from each from of the video it is then cropped based on extracted positional values (x1,y1,x2,y2) and given to our classifier and predicts the values using Argmax we extracted the highest value among it and mapped it to the labels and embed it to the video. The below Figure 3 illustrates the stages of the study.

Figure 3: System Architecture


Imbalance Data
Balancing the data using oversampling method with data Augmentation
Model Training and optimizing
Saving and Loading the model classifier
Importing and extracting the video per frames
Detecting faces per frame using Retina Faces and extract the location
Extracting Region of Interest (ROI) from the frames
Applying the classifier to each ROI to predict the 3 classes
Create bounding box on the frame based on extracted location and add predicted labels to it

Results

For the mask classification task, the following models were investigated in this study: HOG + MLP, baseline CNN, and ResNet 50. The neural net's input size was 64x64x3. The pretrained model's classification layers were not changed. The training lasted 100 epochs, with early stopping and patience set to 10. HOG+ MPL was used to optimise the hyperparameters by varying the learning rate, number of input neurons, and dropout value. ADAM was chosen as the optimizer for all of the models because it is regarded as one of the best optimizers. We trained the model with both datasets to compare the performance of the balanced and imbalance datasets in order to identify the optimal model for the job.

Table 01 displays the sensitivity of each class as well as the model accuracy on the train and test sets. When compared to other models, the HOG feature descriptor with MLP with imbalanced dataset performed the worst, with test accuracy of 18% and the most class predicted by this model being people with no mask (88%), whereas the model with balanced dataset performed well in test set with an accuracy of 79%. Baseline CNN with one convolution layer and no hyperparameter optimisation produced good evaluation metrics in both balanced and unbalanced datasets, with 37% and 94% as true positive rates for minority classes, respectively, compared to 0% for MLP. The sensitivity of projected minority class for Balanced dataset is pretty strong, especially for inappropriate mask, which has 74% sensitivity, which is greater than the model trained on unbalance set. ResNet 50 has the greatest true positive rate of 42% and 96% for both minority classes trained on unbalanced datasets, making it the best model to use for the final classification job.

Models	Classes	Imbalanced Dataset			Balanced Dataset		
		Sensitivity	Train Accuracy	Test Accuracy	Sensitivity	Train Accuracy	Test Accuracy
HOG+MLP	No Mask	0.88	0.9713	0.18	0.16	0.9232	0.79
	With Mask	0.11			0.92		
	Improper Mask	0			0		
Baseline CNN	No Mask	0.94	1	0.9475	0.84	0.9615	0.9323
	With Mask	0.98			0.95		
	Improper Mask	0.37			0.74		
ResNet 50	No Mask	0.96	1	0.96288	0.96	0.9909	0.93013
	With Mask	0.99			0.94		
	Improper Mask	0.42			0.74		
Best Model		ResNet50 in imbalanced Dataset with 96% Test Acc.					

Table 01 Model Performance

The Below figure 04 shows the original class and predicted class by ResNet 50 model, the class with mask and No mask has high positive rate compared to the improper mask, the below sample shows the false predicted class.

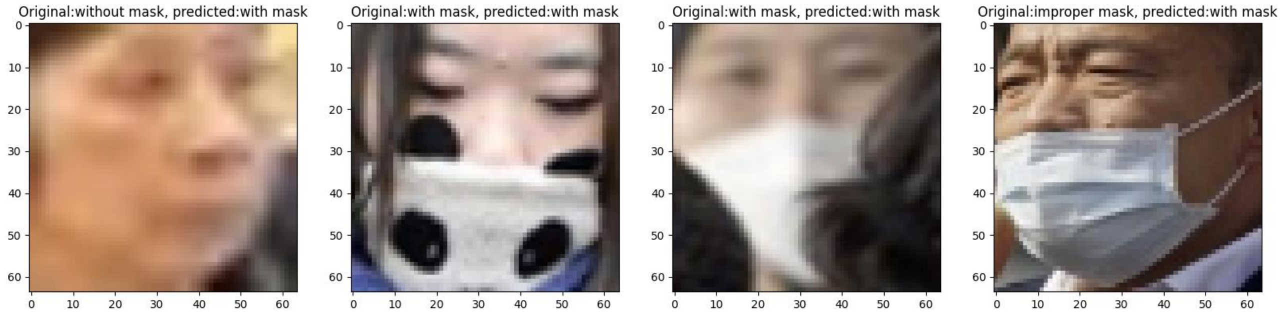


Figure: 04 Class Prediction using ResNet 50 model

We acquired an extraordinarily precise and robust face mask detection after including all phases of our design. As RetinaFace was chosen as our face detector, ResNet50 with imbalanced dataset was picked as the best Face Mask classifier. And while it identified all three groups mainly correctly, the architecture struggled to recognise many faces in a single shot. However, in real-world circumstances, motion blur, dynamic focus, and frame transition make it challenging to distinguish faces. In low light, the classifier correctly spotted the mask-wearing face, indicating that the model performs well in low-light photos.

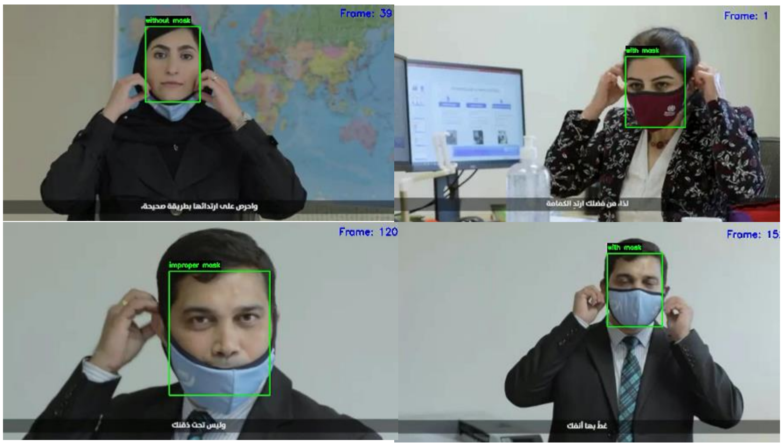


Figure: 05 Video Output

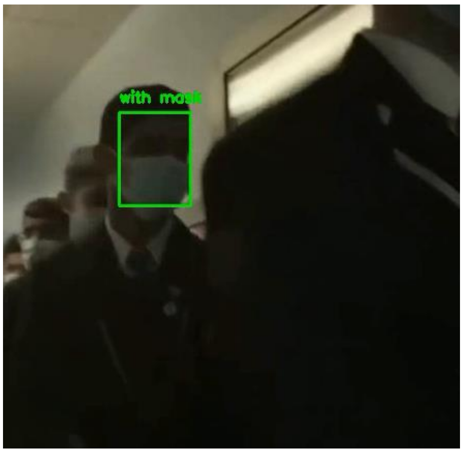


Figure: 06 Low Light Video Output

References

[1] Kar, Nikunja & Babu, Korra & Jena, Sanjay. (2017). Face Expression Recognition Using Histograms of Oriented Gradients with Reduced Features. 10.1007/978-981-10-2107-7_19.