IN3060/INM460 Computer Vision Coursework report

- **Student name, ID and cohort:** Yumi Heo (230003122) - PG
- **Google Drive folder:** https://drive.google.com/drive/folders/1zuAff22-q73cte_k24hB-WlxOzSRh7vJ?usp=sharing

Data

The number of image datasets given is 2,394 for training and 458 for testing. All images have three colour channels: red, green, and blue. However, as the task is detecting face mask, the classes of those images are imbalanced. Class 1, labelled as wearing a mask, has the largest proportion. Class 0, labelled as not wearing a mask, accounts for the second largest proportion, but there is still a significant gap between classes 1 and 2. Class 2, labelled as wearing a mask incorrectly accounts for the smallest proportion. The smallest image size is 16x11 pixels in training set, and it is 16x13 in the test set. The largest size in the training set is 317x340, while it is 241x281 in the test set. For 'Face Covering Detection in a video' task, a video about the explanation of wearing a mask was used, which is in YouTube, and does not have copyright. The duration of the video is 00:01:27, and it has 2,622 frames, 720 pixels in height and 1280 pixels in width. Additionally, this video has three colour channels.

Implemented methods

A total of four models were created for this task. For feature extraction, HOG (Histogram of Oriented Gradients) or SIFT (Scale-Invariant Feature Transform) was used on image data and applied to the SVM (Support Vector Machines) model. The MLP (Multilayer Perceptrons) model was used with HOG to extract the features from the dataset. The last model is ResNet34 (Residual Network), a pre-trained model loaded from PyTorch. The reason why the two SVM models were built with different feature extractors is to compare the two methods, and MLP was created to check what kind of results it produces in image classification as an early neural network before the advent of CNN. Lastly, as a CNN model, ResNet was selected for this training because it generally shows high performance in image classification [1].

First of all, SVM and MLP models can be trained more effectively when using a feature extractor such as HOG or SIFT rather than using the image data itself. This can improve prediction results when testing. However, ResNet, which is one of the CNN (Convolutional Neural Network) models, does not need feature extraction techniques since the convolutional layers are capable of extracting features from input data during training [2]. Before training the four models, the entire training set was resized to 256x256 square images using inter-linear interpolation for efficient computation. Then, the training set was divided into 80% for the training set and 20% for the validation set for SVM and MLP models. To extract features from the images, SIFT was used firstly. All key features were extracted from the training set and clustered using K-means clustering. The number of clusters was set to 10 times the number of labels, which is 30, and this approach is considered a rule of thumb [3]. These clusters grouped similar features from the training image so that the model could learn relatively low-dimensional data. These grouped features were used to train the SVM model. Regarding HOG, it calculates the gradient at each pixel and creates a histogram in defined cell sizes. By combining all histograms to create features for training, the features from the histograms were used for training another set of SVM and MLP models. The SVM and MLP models were built using the Scikit-learn library. The last model, ResNet34, a pre-trained model from PyTorch, used a different data type for training, as this model requires torch for neural networks from PyTorch. Before changing the data type, the previously resized training images with interpolation were used for data augmentation, employing methods such as flipping or altering the colour of the image data to increase diversity, since the classes are imbalanced. To find the mean and standard deviation of the training set for normalization, the training set was converted to float, and the values of mean and standard deviation were calculated. Normalization was performed at the end of the data augmentation process. The training data was separated again to define a validation set. Each image and label were concatenated using a dataset class. For the final step before training the model, batches were used to generalize the datasets while training and evaluation.

After training the four models, grid search was applied to find the optimal hyperparameters for the SVM models and MLP model. For computational efficiency, the number of parameters for grid search was set differently for each model. For SVM models, it was relatively fast to converge while performing grid search on regularization C, gamma, and kernel. However, for the MLP model, the elective parameters are larger than those of SVM models. The range of parameters for grid search started from a narrow range with selective parameters. First, starting from a single hidden layer, the first hyperparameter grid search was conducted on the hidden units, activation functions, and optimizers. After obtaining results from the first grid search, using the obtained parameters, the second grid search was conducted with regularization alpha, learning rate, and momentum. This approach was applied to the MLP model with 2 hidden layers. For the final model, ResNet34, as it is a pre-trained model, it is difficult to change the core architecture itself. However, fine-tuning, changing some parameters such as learning rate, the number of epochs, and batch size, might be helpful to increase the accuracy in this case. The optimizer Adam has a default learning rate of 0.001. For better training regarding face mask detection on video, lowering the learning rate and setting the number of epochs large would be suggested to make smaller but adjust more updates on weights. Using this idea, the learning rate was set as 0.0001 and the number of epochs as 100. The test set was only changed in size to match the training and validation sets. While training the model, the training set and validation set were used to pick the best model using a batch.

For testing the video, the size of the video was first checked using OpenCV code: VideoCapture. A cascade classifier was then used to detect faces. Initially, the minimum size started from 30x30, but after detecting unnecessary parts other than the face, it was later modified to 300x300. Next, the face images from the video were converted to 224x224 to match the trained image size of ResNet, and the data type was changed to tensor to pass to the model. Here, normalization, unlike during training, used the values suggested by PyTorch to enable normalization to various image sizes. Afterwards, the MaskDetectionVideo function was used to enable the model to detect faces in the video, draw bounding boxes, and play the video with the predictions.

Results

[Qualitative results of performance]

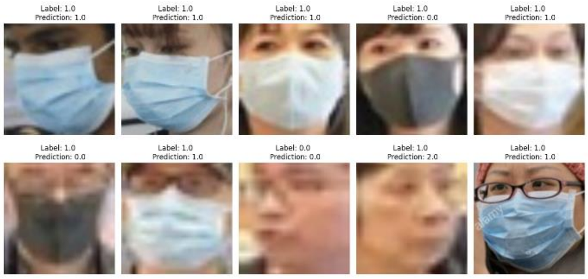

Figure 1. SIFT+SVM Predictions

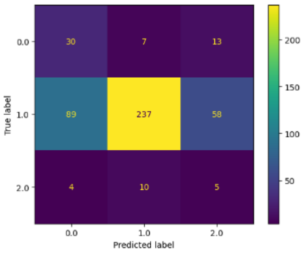

Figure 2. SIFT+SVM Confusion Matrix

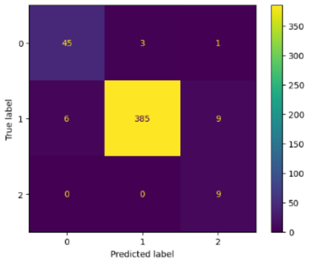

Figure 3. ResNet34 Confusion Matrix

While the highest accuracy is an important indicator to select the best model, the final model was selected considering precision and recall using the evaluation report and confusion matrix, since the class is imbalanced. Therefore, the SVM with SIFT model underwent hyperparameter tuning was selected as the final model as it predicted a few instances of class 0 and 2. Regarding the SVM with HOG and MLP with HOG models, those models predicted only class 1 after hyperparameter tuning. Hence, the baseline of those two models was selected as the final model since those models predicted few instances of classes 0 and 2. Lastly, ResNet34 was selected as the final model itself by setting the learning rate lower than the default value and setting the number of epochs to 100. Even though it made 1 correct prediction for class 2, this model predicted classes 0 and 1 highly accurately compared to the other models.

[Quantitative results of performance]

No.	Model Name	Training Speed (sec)	Test Accuracy	Model Size
1	Final SIFT+SVM	0.42	0.60	374 KB
2	Final HOG+SVM	5.87	0.82	23.1 MB

3	Final HOG+MLP	4.91	0.85	1.6 MB
4	Final ResNet34	4080.90 (Google Colab GPU used)	0.95	81.3 MB

Table 1. Comparison between final models

- **Training Speed: Final ResNet34 > Final HOG+MLP > Final HOG+SVM > Final SIFT+SVM**
  The depth of ResNet34 is much deeper than other models as it consists of 34 layers. Additionally, weights are assigned as data pass through each hidden layer, so the training speed was slow. The second slowest training model is the HOG+MLP model. Even though early stopping was applied to prevent overfitting, it can be assumed that the number of epochs made the training slower than the iteration of HOG SVM. The HOG+SVM model was slower than the SIFT+SVM model. It is inferred that the process of HOG to extract features may take longer than that of SIFT.
- **Test Accuracy: Final ResNet34 > Final HOG+MLP > Final HOG+SVM > Final SIFT+SVM**
  ResNet is a model with few errors in image classification. As a result, it shows the highest accuracy in predicting class 0 and class 2 compared to the other models used in this training. HOG+MLP model has a slight difference in accuracy compared to HOG+SVM. Hence, it is assumed that the accuracy rankings of the two models may change with additional fine hyperparameter tuning.
- **Model Size: Final ResNet34 > Final HOG+SVM > Final HOG+MLP > Final SIFT+SVM**
  As ResNet is a pre-trained model and reapplies the weights from new input data, the model is heavier than the others. The SVM model with HOG is heavier than the MLP model with HOG since SVM operates with a vector matrix. Due to this calculation, the model could become heavier during training. Also, the MLP used here can be lighter than SVM because it has a single hidden layer. Regarding the size of the SVM model with SIFT, it is assumed that the weight of features extracted with SIFT is lighter than that of HOG.

[Qualitative results on test video]

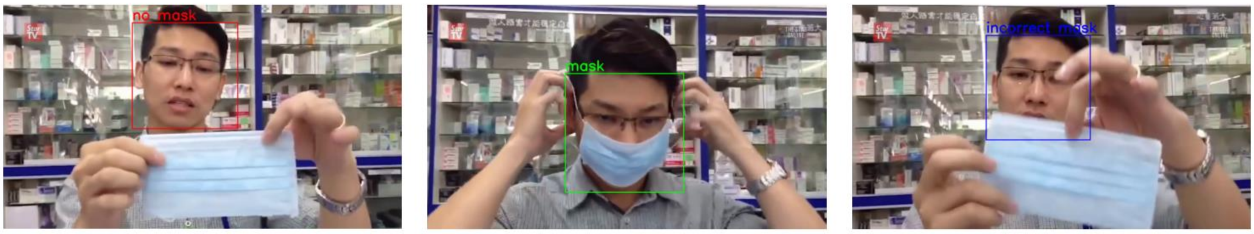

Figure 4. Mask Detection Video

Therefore, ResNet34 was chosen to test the video to detect whether a person wears a mask or not. The video test results confirmed that ResNet34 can be used for image classification tasks with fewer errors. Since this model achieved the highest accuracy on the test set provided as an image, the prediction result for the video was also deemed reliable. Although it incorrectly predicted that a person was wearing a mask incorrectly when the person lifted the mask and part of the face was not visible, it still detected that the mask covered some part of the face, which led the label to be class 2.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, ‘Deep Residual Learning for Image Recognition’, 2015, doi: 10.48550/ARXIV.1512.03385.

[2] S. Alizadeh and A. Fazel, ‘Convolutional Neural Networks for Facial Expression Recognition’, 2017, doi: 10.48550/ARXIV.1704.06756.

[3] G. Tarroni, ‘IN3060/INM460 Computer Vision, Lab tutorial 6’. Accessed: Apr. 16, 2024. [Online]. Available: https://moodle4.city.ac.uk/pluginfile.php/439080/mod_folder/content/0/Lab_06.ipynb?forcedownload=1