Emotion Facial Classification With Masks

Cheng Qiu
Sung-Lin Hsieh
Christopher Jang
Walker Larivee

Link: https://github.com/chengq220/EmotionDetection

Introduction

Emotion recognition involves developing models that can recognize human emotions from various inputs like facial expressions, voice intonations, body movements, and even text. The goal is to enable machines to interpret and respond to human emotions in a way that mimics human-like understanding. This technology has broad applications, including in customer service to enhance interactions, in healthcare for monitoring patient moods, and in automotive industries for safer driving by assessing the driver's emotional state. One unique feature that we have added to this is a mask. Inspired by the COVID-19 pandemic, we decided to investigate the effect of masks on images in classifying emotions using the FER2013 dataset which contains thousands of images of human faces. The main question for our project is whether models can accurately classify emotions with a mask as opposed to without a mask. We hypothesized that the model will not be able to accurately predict the emotion because important features used to understand emotions are covered by a mask.

Literature Review

The state of the art in emotion classification, particularly in the context of facial emotion recognition (FER), has seen significant advancements with the utilization of deep learning models, especially Convolutional Neural Networks (CNNs) [1]. CNNs are frequently used for FERs due to their powerful automatic feature extraction capabilities and computational efficiency. These models have shown great potential in accurately and robustly recognizing facial emotions, despite challenges such as the heterogeneity of human faces and variations in images due to different facial poses and lighting [2]. To further understand these classification models and their performance compared to ground truth, the most used metrics to do this include the accuracy, F1-Score, ROC-AUC, and etc.

A notable model in emotion recognition is the VGG which achieved the highest single-network classification accuracy on the FER2013. In the study, the VGG's hyperparameters were rigorously fine-tuned and further experimented with various optimization methods, resulting in SOTA accuracy of 73.28% without using extra training data [3]. In another study, to tackle the bottleneck of basic and shallow CNN architecture for emotion detection, an ensemble method was proposed which is an ensemble of up to eight CNN networks and this method achieved an accuracy of 75.23% on the Fer2013 [5]. However, the main challenge in facial recognition still remains and that is the fact that there is no publicly available dataset that is large enough for current deep learning standards.

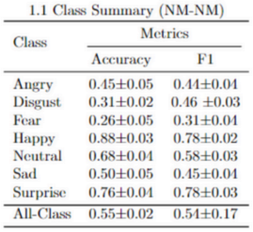
Task of emotion classification with masks involves challenges distinct from standard facial emotion recognition. Research in this area often employs CNNs to identify masks and then estimating emotions based on the visible facial features. One such study proposed a method to enhance low-light images and analyze upper facial features using CNN. By covering the lower part of the face with a synthetic mask, the study leveraged boundary and regional representation methods to highlight the head and upper facial features. Feature extraction was then performed based on the detected facial landmarks of the partially covered face. This method achieved an accuracy of 69.3% on the AffectNet dataset which is another facial emotion dataset [4].

Method

The baseline method that we chose was SVM. We used Scikit's SVC class to implement our SVM model and the first step was to reduce the dimensionality and complexity of our dataset using principal component analysis. Since each picture was 48x48 pixels, the images were broken down into arrays of 2304 floats, with each float representing the greyscale value for a pixel in the image. After experimenting with the model, we decided to utilize 150 principal components (PC), which accounted for ~90% of the variance in our dataset. Any more PCs would cost exponentially more processing power and diminishing returns on accuracy. The SVM parameters were chosen using cross-validation, and the model was then trained using the provided training dataset. After training, the SVM model was used to classify each image in the test set, and these predictions were compared to the true classifications for each test image.

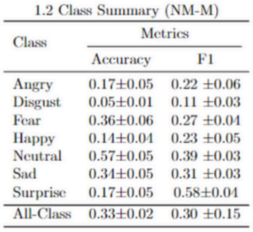
Our second method uses the MobileNetV2 Convolutional Neural Network architecture as the main model. MobileNetV2 consists of 53 layers and some notable features of this network include depth-wise separable convolutions, inverted residuals with linear bottlenecks, width multiplier, and resolution multiplier. The preprocessing step involves normalizing and resizing the images, and the labels are encoded using one-hot-encoding. The MobileNetV2 is a pre-trained model with weights from ImageNet loaded without the top classification layers using Tensorflow Keras with the output layer having 7 units with softmax. In addition, some layers of the model are unfrozen (allowing the neural network to update some weights) for fine tuning. The model is trained with Adam optimizer and categorical cross-entropy loss, along with data augmentation techniques such as rotation, shifting, etc. to diversify the training data.

Result



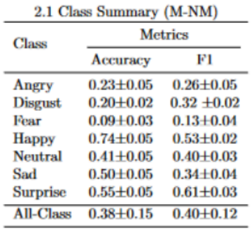
1.1 Class Summary (NM-NM)

Class	Metrics	
	Accuracy	F1
Angry	0.45±0.05	0.44±0.04
Disgust	0.31±0.02	0.46 ±0.03
Fear	0.26±0.05	0.31±0.04
Happy	0.88±0.03	0.78±0.02
Neutral	0.68±0.04	0.58±0.03
Sad	0.50±0.05	0.45±0.04
Surprise	0.76±0.04	0.78±0.03
All-Class	0.55±0.02	0.54±0.17



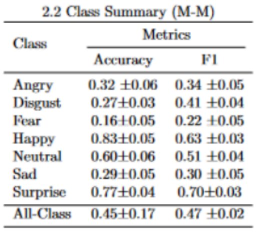
1.2 Class Summary (NM-M)

Class	Metrics	
	Accuracy	F1
Angry	0.17±0.05	0.22 ±0.06
Disgust	0.05±0.01	0.11 ±0.03
Fear	0.36±0.06	0.27 ±0.04
Happy	0.14±0.04	0.23 ±0.05
Neutral	0.57±0.05	0.39 ±0.03
Sad	0.34±0.05	0.31 ±0.03
Surprise	0.17±0.05	0.58±0.04
All-Class	0.33±0.02	0.30 ±0.15



2.1 Class Summary (M-NM)

Class	Metrics	
	Accuracy	F1
Angry	0.23±0.05	0.26±0.05
Disgust	0.20±0.02	0.32 ±0.02
Fear	0.09±0.03	0.13±0.04
Happy	0.74±0.05	0.53±0.02
Neutral	0.41±0.05	0.40±0.03
Sad	0.50±0.05	0.34±0.04
Surprise	0.55±0.05	0.61±0.03
All-Class	0.38±0.15	0.40±0.12



2.2 Class Summary (M-M)

Class	Metrics	
	Accuracy	F1
Angry	0.32 ±0.06	0.34 ±0.05
Disgust	0.27±0.03	0.41 ±0.04
Fear	0.16±0.05	0.22 ±0.05
Happy	0.83±0.05	0.63 ±0.03
Neutral	0.60±0.06	0.51 ±0.04
Sad	0.29±0.05	0.30 ±0.05
Surprise	0.77±0.04	0.70±0.03
All-Class	0.45±0.17	0.47 ±0.02

Figure 1. (Upper) Performance of CNN trained on unmask dataset and (Lower) Performance of CNN trained on mask dataset on unmasked test data (NM) and masked test data (M) respectively

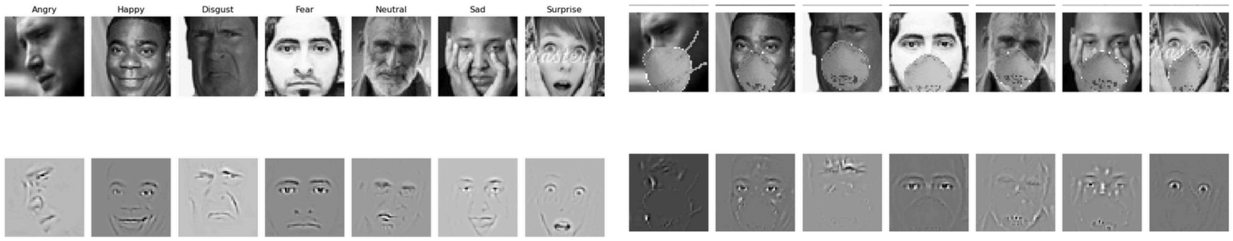


Figure 2. Saliency map of the (Left) model trained on unmasked data and (Right) model trained on masked data

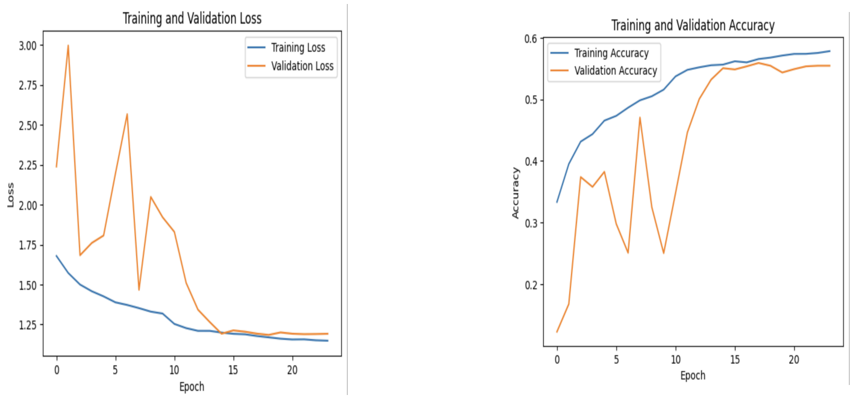


Figure 3. (Left)Training/Validation Loss. (Right) Training/Validation Accuracy for model MobileNetV2

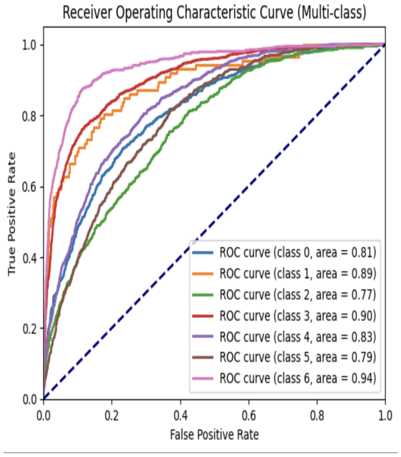
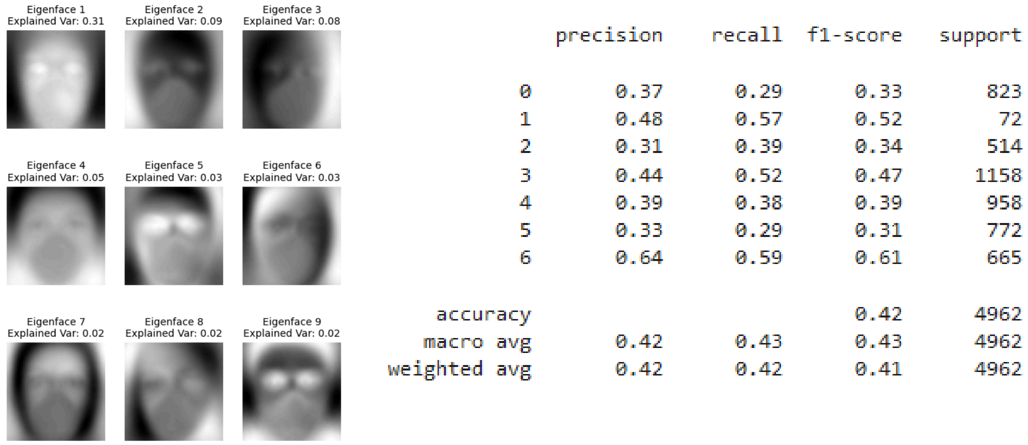


Figure 4. ROC curves for each emotion class from MobileNetV2



```
              precision    recall  f1-score   support

           0       0.37      0.29      0.33       823
           1       0.48      0.57      0.52        72
           2       0.31      0.39      0.34       514
           3       0.44      0.52      0.47      1158
           4       0.39      0.38      0.39       958
           5       0.33      0.29      0.31       772
           6       0.64      0.59      0.61       665

    accuracy                           0.42      4962
   macro avg       0.42      0.43      0.43      4962
weighted avg       0.42      0.42      0.41      4962
```

Figure 5: (Left) 9 most significant principal components for SVM. (Right) Performance metric for SVM

Discussion

The SVM model achieved an accuracy of 42% with a recall of 0.43 and an F1 score of 0.43. The model performed best on the "surprised" emotion with an accuracy of 64%, and performed worst on the "fear" emotion with an accuracy of 31% shown in Figure 5. We did not have high hopes for the SVM model primarily because SVM does not perform well when the data is noisy and has overlapping target classes. To improve our model, more principal components could be used at the cost of runtime, and accuracy could potentially be improved by providing the model with more features. This would allow the model to identify more detailed characteristics for each emotion, which is especially useful when the mouth of each face is covered and the model is heavily relying on changes observed in and around the eyes.

As for the MobileNetV2 model, it achieved an accuracy of around 60% with F1 score of 0.51 and recall of 0.54. We experimented with different depths for each of the models and it seems to be the case that there's minimal gains with deeper networks compared to models that are not as deep. Figure 3 shows that the training stopped after validation loss starts to increase after 20 epochs which means that the model did not overfit. In Figure 3, the validation accuracy/loss curves start out very shaky but smooth out towards the end of training, indicating that the model is slowly improving over time. Figure 4 shows the ROC curves for each emotion class and unexpectedly, happy emotion, which we had the most data for, came in second while surprise emotion did the best.

Overall, the models described above seem to do alright on masked data, but how does it compare to models trained on unmasked data? Figure 1, which shows the performance metric for models trained using mask and unmasked data, showed some interesting results. Generally, the accuracy of model trained on unmasked data is higher than model trained on masked data which aligns with our hypothesis. However, when trying to inference on masked data using an unmasked-trained model, there were significant drops of up to 80% in accuracy while inference on unmask data using masked-trained models experienced up to 20% drop in accuracy. Through visualization techniques such as Guided-Backpropagation shown in Figure 2, we discovered that models trained with unmasked images focus mainly on eyes and mouth which is especially for classes such as surprise and happiness. On the other hand, models trained with masked images were able to extract more refined features such as the contours around the face that are indicative of certain emotions. By adding in a mask, the model trained with unmasked images loses that information thus resulting in such a significant loss in performance.

Conclusion/Future Steps

One major setback when working on this project is that the dataset is heavily imbalanced which led to much difficulty developing models that work decently for those minority classes. A few of the workarounds that we've taken is using image augmentation techniques to diversify the dataset as well as testing out other loss functions that seek to mitigate the effect of imbalance of the dataset. All of these workarounds led to incremental improvement, so one avenue for future works is in order to get better performance, we might have to balance out the dataset with more examples for the minority class through means such as finding more images online or image generation. Another avenue for future works is since the dataset is imbalanced, we can experiment with models that's trained on each of one of those classes independently. Then, by pooling the prediction for the different class model together, we can experiment to see if it will achieve higher accuracy.

Contributions

We as a group contributed equally. Chris performed the literature review and assisted with presentation creation. Cheng conducted the experiment to understand how the models behave with different datasets. Sung-Lin worked specifically on the deep neural network on Fer2013 masked dataset and performed analysis on the performance of the model. Walker focused on implementing the SVM model for this dataset and performed the necessary analysis to understand the model's performance. Everyone contributed to the final project writeup.

References

[1] Papers with Code - Emotion Classification. paperswithcodecom. https://paperswithcode.com/task/emotion-classification.

[2] Naziha Hmidi, Rim Afdhal, Hamdi M, Ridha Ejbali, Mourad Zaied. 2024. Emotion estimation of people wearing masks using machine learning. International Journal of Computers, Communications & Control (Print). 19(1). doi:https://doi.org/10.15837/ijccc.2024.1.5363.

[3] Khaireddin Y, Chen Z. 2021 May 8. Facial Emotion Recognition: State of the Art Performance on FER2013. arXiv:210503588 [cs]. https://arxiv.org/abs/2105.03588.

[4] Mukhiddinov M, Djuraev O, Akhmedov F, Mukhamadiyev A, Cho J. 2023. Masked Face Emotion Recognition Based on Facial Landmarks and Deep Learning Approaches for Visually Impaired People. Sensors. 23(3):1080. doi:https://doi.org/10.3390/s23031080.080

[5] Pramerdorfer C, Kampel M. 2016 Dec 8. Facial Expression Recognition using Convolutional Neural Networks: State of the Art. arXiv:161202903 [cs]. https://arxiv.org/abs/1612.02903.