Emotion detection with Masked data

Team Information

1. 김상욱 (20150145)
2. 송소정 (20160317)
3. 오유경 (20180392)
4. 이수로 (20160830)

GitHub link

https://github.com/Oh-yk/CS470_EmotionDetection

Introduction

Coronavirus has changed our lives a lot. Exchange student programs have been canceled and online classes are being conducted worldwide. Moreover, people are always required to wear masks outdoors, otherwise unable to use various facilities including public transportation. This made it hard for people to recognize other people's facial expressions in everyday life. The paper aims to build a model that allows us to recognize the emotions of mask-wearing people in the midst of COVID by using various artificial intelligence techniques.

We used 'Deep Emotion' [1] as the baseline model which is a facial expression recognition model that uses an attentional convolutional network. In this project we first replicate the existing paper. Then, we build an artificially masked dataset and create our own models that effectively guess the emotions of mask-wearing people.

Method

1. Datasets

The datasets that we use are labeled with 7 emotional states : angry, disgusted, frightened, happy, sad, surprised, and neutral.

	JAFFE	CK+	FER 2013
amount of data	213 images	981 images	2938 images
people	10 Japanese Female Models	123 college students	-
feature	aligned face	aligned face	unaligned face

Table 1: Description of the three datasets used

2. Data Preprocessing

We needed a masked dataset to train our model, but we could not get an emotion-labelled dataset with masked people. Therefore, we generated our own dataset. In this process, MTCNN [2] was used to detect the landmarks of the eyes, nose, and mouth of the face. The mask area was set by dividing the distance between the eyes and the nose by 8:2 ratio, and then we zeroed out the pixels in the mask area. Moreover, MTCNN accurately detects landmarks on images of people wearing masks. Thus during deployment in real-life scenarios, the masked area of mask-wearing people can be calculated in the same manner, and we can zero out the pixels of the area..

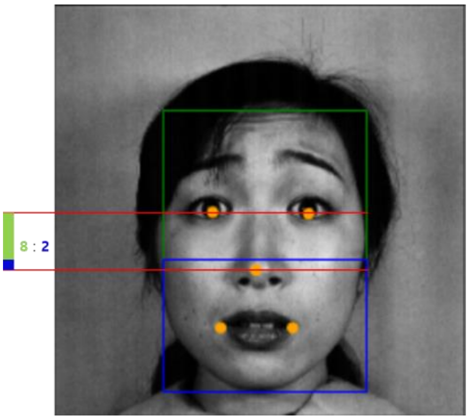


Figure 1: Landmark detection for masking

3. Model

The main model is based on an attentional convolutional network, and uses spatial transformations as an attention mechanism. The STN(spatial transformer Network) consists of two convolution layers. The feature extraction part consists of four convolutional layers, each two followed by a max-pooling layer and rectified linear unit (ReLU) activation function. They are then followed by a dropout layer and two fully-connected layers. We trained this model with our masked datasets. Also, we made some modifications to the model to improve accuracy. To improve the baseline model, we modified the model by removing the spatial transformer network, increasing the number of intermediate convolutional layers from 10 to 50, and adding L1 regularization.

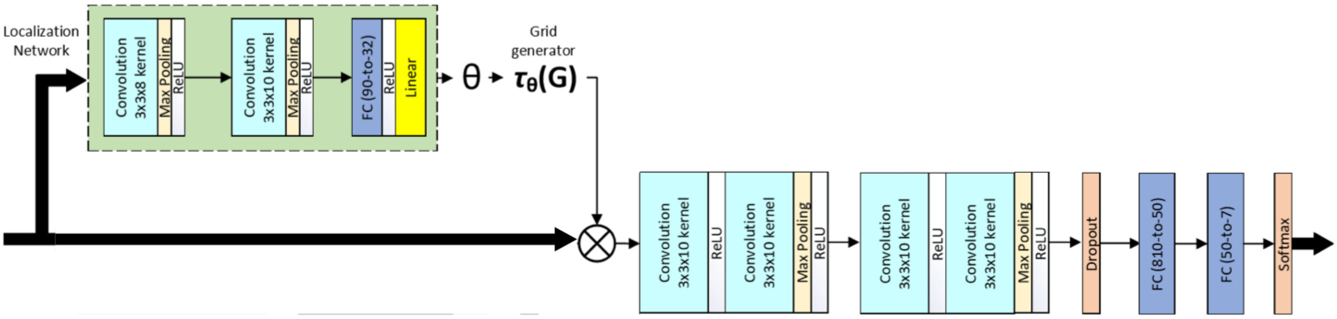


Figure 2: Baseline model for facial expression detection

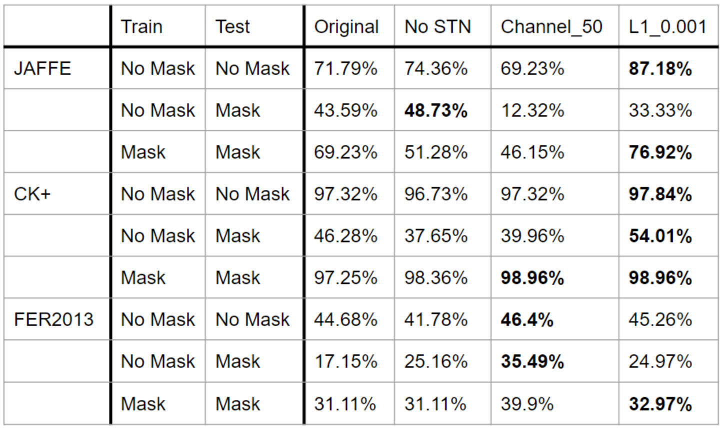
Experimental Results

1. Training Procedure

We trained eight models per dataset, four of which were trained with the original unmasked dataset, and four of which were trained with the artificially masked dataset. Out of the four models trained with one type of dataset, one was identical to the Deep Emotion model, one had the Spatial Transformer Network removed, one had 50 channels instead of 10, and one had L1 regularization with weight 0.001 added. All models were trained for around 100 to 150 epochs, with a learning rate of 0.004, and a batch size of 64.

2. Classification Accuracy

We performed experiments on three popular facial expression datasets: FER2013, the extended Cohn-Kanade (CK+) dataset, and the Japanese Female Facial Expression (JAFFE) dataset. Specifically, we measure the improvements achieved in masked emotion detection by training on an artificially masked dataset and measure the improvements achieved by modifying the original Deep Emotion model.



	Train	Test	Original	No STN	Channel_50	L1_0.001
JAFFE	No Mask	No Mask	71.79%	74.36%	69.23%	**87.18%**
	No Mask	Mask	43.59%	**48.73%**	12.32%	33.33%
	Mask	Mask	69.23%	51.28%	46.15%	**76.92%**
CK+	No Mask	No Mask	97.32%	96.73%	97.32%	**97.84%**
	No Mask	Mask	46.28%	37.65%	39.96%	**54.01%**
	Mask	Mask	97.25%	98.36%	**98.96%**	**98.96%**
FER2013	No Mask	No Mask	44.68%	41.78%	**46.4%**	45.26%
	No Mask	Mask	17.15%	25.16%	**35.49%**	24.97%
	Mask	Mask	31.11%	31.11%	39.9%	**32.97%**

Table 2: Test accuracies of each of the models. The train and test columns refer to the type of dataset (unmasked/masked) used for training and the last four columns refer to the modifications to the model we made. ‘No STN ‘ is the model without the spatial transformer, ‘Channel_50’ is the model with 50 channels, and ‘L1_0.001’ is the model with L1 regularization of weight 0.001.

From Table 2, one can see that models trained with unmasked faces perform poorly when predicting the emotions of masked individuals, typically resulting in large test accuracy drops compared to prediction of unmasked individuals. An example of one extreme case is the CK+ dataset, where the original Deep Emotion model results in a test accuracy drop of 51.04% when attempting to predict the emotions of masked individuals. By training with masked individuals, however, we improve test accuracies significantly, practically achieving the same performance as the unmasked model predicting unmasked individuals. For instance, the original Deep Emotion model trained on the masked CK+ dataset results in a 97.25% accuracy, almost perfectly reverting the performance drop mentioned above. Thus, it is clear that training with masked individuals significantly boosts the performance of emotion prediction on masked individuals. This result is also meaningful because it largely implies that not all facial features are necessary to correctly predict the emotion of an individual.

Furthermore in the rows of Table 2, we can see that the original Deep Emotion model is not optimal regardless of training method and dataset. We tried to improve performance by trying three empirically determined methods. First, we remove the Spatial Transformer Network (STN), which did not significantly impact the performance in the JAFFE and CK+ datasets. However, this is because these two datasets are already aligned datasets that do not require the STN to align or scale the images for them. In the FER2013 dataset, which is unaligned, removing the STN slightly lowers the accuracy of prediction. The abnormal accuracy improvement present in masked prediction with an unmasked-trained model seems to be due to the fact that the model is not well trained for masked prediction to begin with, resulting in luckier guesses without STNs.

Performance improvement occurs in unaligned datasets when STNs are added because the STN better aligns and scales the faces prior to emotion detection, making classification a bit more spatially invariant. VIsualization of images transformed by the STN are in Figure 3.

Then, we tried adding channels and L1 regularization, both of which performed better than the original Deep Emotion model. Overall, the original Deep Emotion model with L1 regularization of weight 0.001 performs the best compared to other types of models.

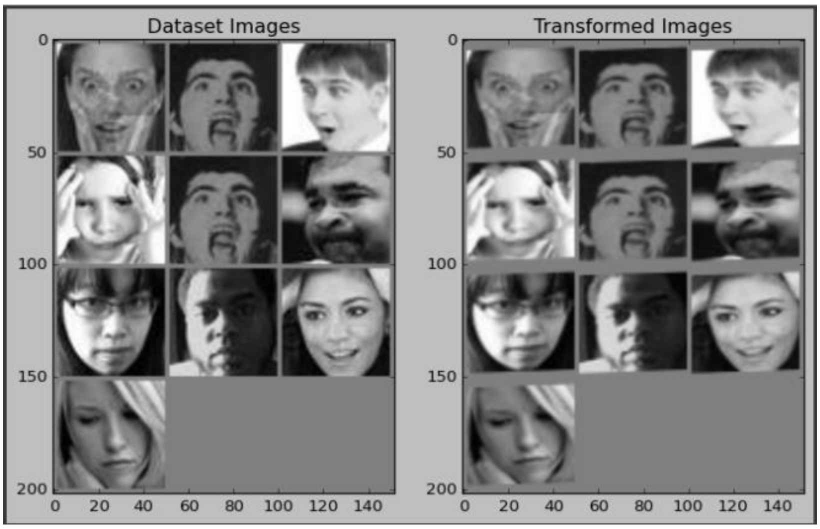


Figure 3: Images from the FER2013 datasets before and after transformations by the STN. *Left*: Original images. *Right*: Images transformed by the STN. Faces are more circular and centered.

3. Confusion Matrix

We now present confusion matrices to better visualize results. The rows of the confusion matrix represent the prediction and the columns represent the ground truth label. In Figure 4, we can see that the model trained with unmasked individuals results in many misclassifications, frequently misclassifying disgust and happiness for anger. Such errors are removed once we train with the artificially masked dataset, resulting in an almost diagonal confusion matrix. As shown in Figure 5, a similar improvement can be seen when increasing the number of channels or when adding L1 regularization.

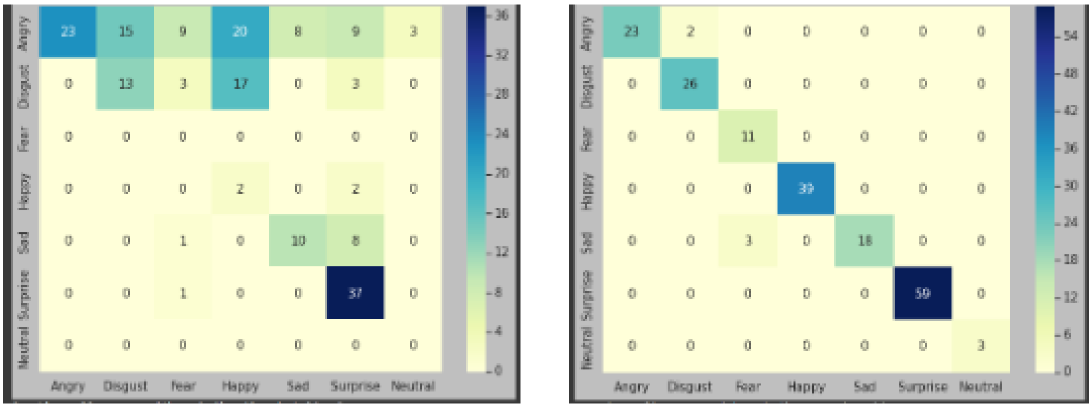


Figure 4: Confusion matrix results using the CK+ dataset. *Left*: Confusion matrix for masked classification using a model trained with unmasked individuals. *Right*: Confusion matrix for masked classification using a model trained with artificially masked individuals.

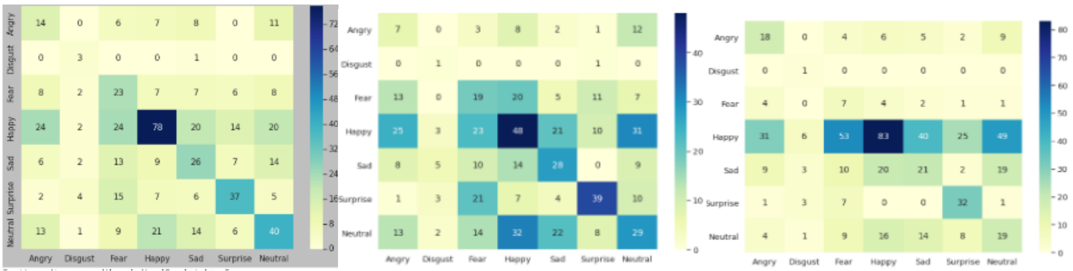


Figure 5: Confusion matrix results using the FER2013 dataset. *Left*: Confusion matrix for masked classification using the original Deep Emotion model trained with masked individuals. *Middle*: Confusion matrix for masked classification using the 50 channel model trained with masked individuals. *Right*: Confusion matrix for masked classification using the L1 regularization model trained with masked individuals.

4. Saliency Map

In order to visualize the salient facial regions for emotion classification, we create saliency maps similar to [1]. First, we zero out an Nx N region at the top-left corner of the image. If occluding the region results in a misclassification, we consider that region to be important in classification. By using a sliding window and shifting using a stride s, we repeat this procedure for the entire image.

In Figure 6-8, we present the saliency maps for the 'surprise' emotion and the changes in saliency as we add L1 regularization and then train with the artificially masked dataset. Initially in Figure 6, the saliency map for the baseline Deep Emotion model seems to have no interpretable meaning. We believe that this is due to overfitting caused by the lack of regularization. However in Figure 7, adding L1 regularization results in a saliency map that clearly indicates that the eyes and mouth are the important regions for correctly classifying the 'surprise' emotion. This is because the model has generalized well to unseen data, understanding that wide open eyes and a slightly open mouth are the characteristics of 'surprise'. Once we mask the image, however, the mouth becomes occluded. Thus in Figure 8, the salient region around the mouth disappears and the salient regions around the eyes become larger, even extending to the eyebrows. As expected, the model learns to focus on only the eyes to classify 'surprise' and learns to utilize extra given information (ex. eyebrows) to better classify emotion.

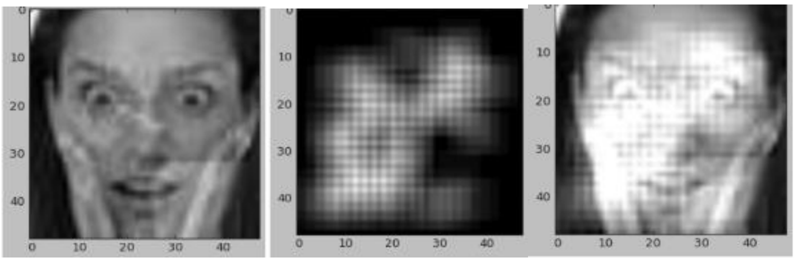


Figure 6: Saliency map of the 'surprise' emotion. The original Deep Emotion model trained with the unmasked FER2013 dataset. *Left*: Original image. *Middle*: Saliency map. *Right*: Saliency map on top of the original image.

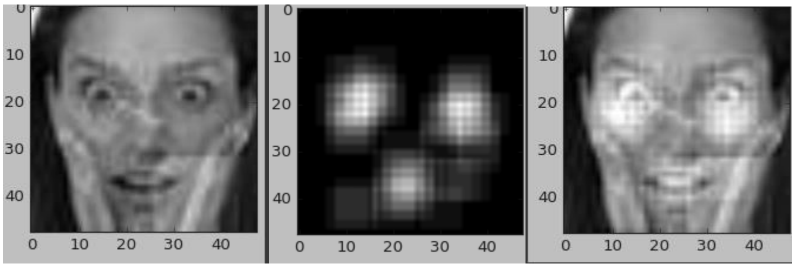


Figure 7: Saliency map of the 'surprise' emotion. The L1 regularization-added model trained with the unmasked FER2013 dataset. *Left*: Original image. *Middle*: Saliency map. *Right*: Saliency map on top of the original image.

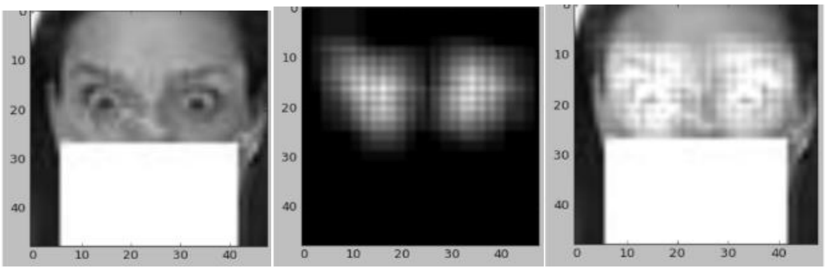


Figure 8: Saliency map of the 'surprise' emotion. The L1 regularization-added model trained with the masked FER2013 dataset. *Left*: Masked image. *Middle*: Saliency map. *Right*: Saliency map on top of the masked image.

Conclusion

The paper proposes a training method that makes facial expression detection robust to face masks. In addition, the model presented in the paper outperforms the baseline model by a noticeable margin. We provide an extensive analysis of our results through various quantitative and qualitative metrics, which show the effectiveness of our approach. Our work is significant in that we show models can learn to focus on only the unoccluded parts of a masked image, resulting in robustness to masking.