Visual Emotion Detection

Technical Task

Luong Phat Nguyen

April 28, 2024

Contents

1 Introduction

Facial expression recognition has been active in the computer vision field for a long time. Several existing methods were used for solving the problem such as hand-crafted features based or deep learning methods. Recently, convolutional neural networks (CNN) have shown its efficiency in many computer vision applications.

In this technical report, we are interested in investigating the efficacy of CNN in recognizing facial emotion. More interestingly, Facial Emotion Recognition Plus (FER+) dataset is used in order to validate this. The FER+ dataset contains around 35000 images of 8 facial emotions: 'neutral', 'happiness', 'surprise', 'sadness', 'anger', 'disgust', 'fear', 'contempt'. In this dataset, the multiple label are used for describing one sample. For example, one facial emotion image can be both happy and surprised at the same time (Figure 1). Beside FER+, I have also conducted the same experiments on FER original dataset to see how the models can deal with the single label classification with only 7 classes.

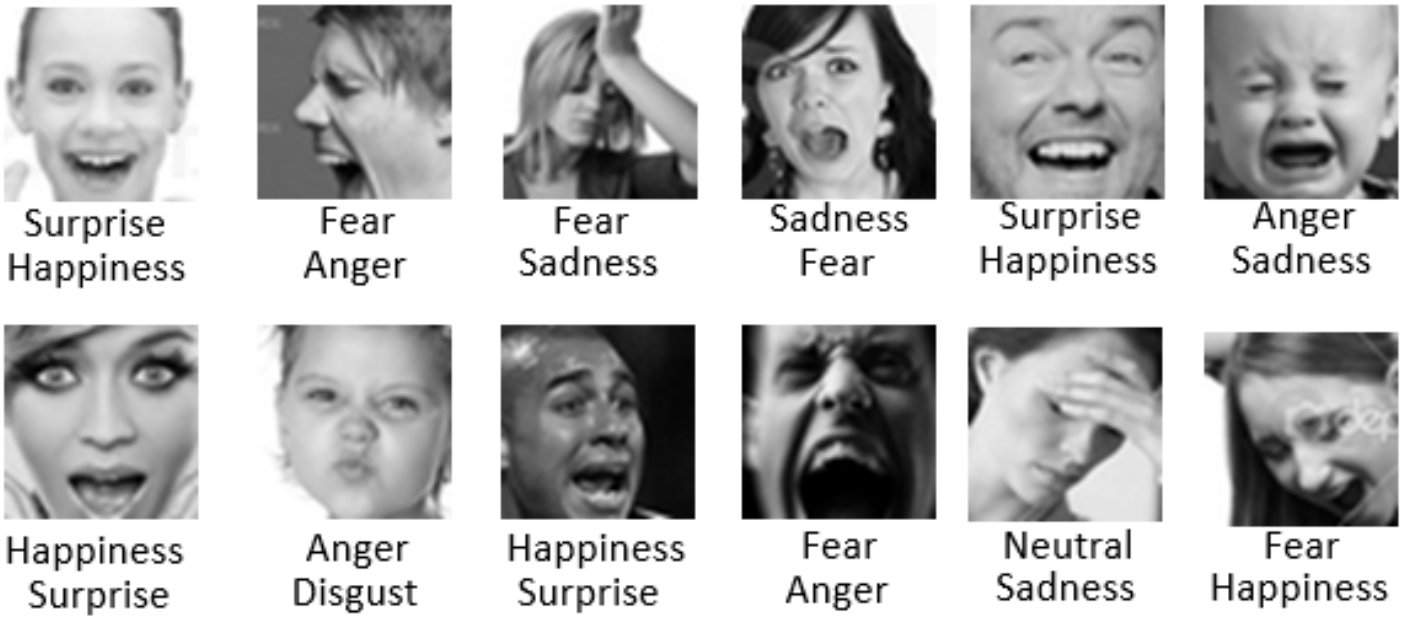


Figure 1: Some examples of FER+ dataset.

2 Model selection and implementation detail

In this section, we are going to walk through some of the implementations of convolutional neural network. It is interesting to see how LeNet works out for this application as it performs well for written numbers datasets (eg. MNIST). Furthermore, ResNet, one of the famous architecture can be used for image classification. In our case, it can be applied in the case of facial emotion detection. Two of the famous architectures of ResNet are ResNet 18 and ResNet 34. They are often known for small size models and can be applied in real-time application. Beside these models, MobileNet is also used in this work. The following subsections will walk through the details of each of these models. We will also take a look at the overall architecture of the proposed network. Furthermore, the implementations of these networks are done using PyTorch as the main framework.

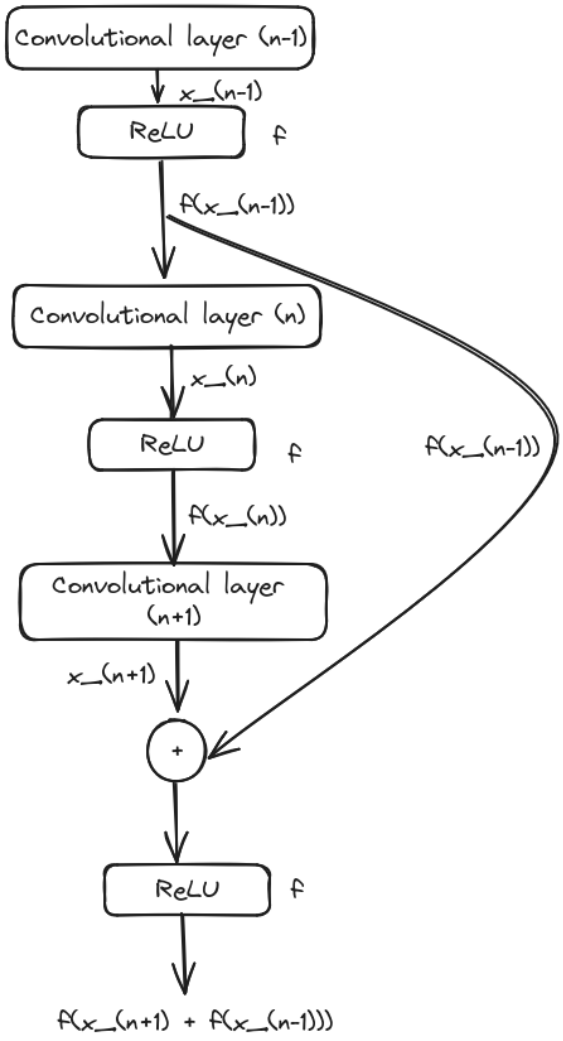


Figure 2: Skip connection in ResNet block.

2.1 Implementation of LeNet

LeNet was used for recognizing hand written digits from 0 to 9. It is the first convolutional neural network used for computer vision application. In this implmentation, the classic LeNet is used. Due to size of the image, the detail of the architectur is shown in the README.md file in the Github repository.

2.2 Implementation of ResNet 18 and ResNet 34

ResNet 18 and ResNet 34 were introduced to solve image classification for ImageNet dataset. It has shown its efficiency in many benchmark datasets and are still in used for many recognition applications. The skip-connection in ResNet architecture is demonstrated in Figure 2. Why does skip connection work? The method actually helps CNN to perform well because it reduces overfitting, solves the gradient vanishing problem and preserves spatial information. The ResNet 18 and 34 are the group of successive ResNet blocks. Depending on the depth of the model, we have the ResNet 18 and ResNet 34 (deeper). For visual detail of the ResNet 18 and ResNet 34 applied to the case of FER and FER+ dataset, the images are

shown in the README.md file in the Github repository.

2.3 Overall architecture

In this subsection, we present the overall architecture using different model. Figure 3 demonstrate the overall architecture. Look at the figure, we can see that different loss methods are used for each dataset FER and FER+. For FER+, Binary cross entropy is used. Meanwhile, cross entropy is used for the FER dataset as experimenting on any other classification dataset. The reason why BCE Loss is used in the case is that we want to have the difference between each of the class in each sample.

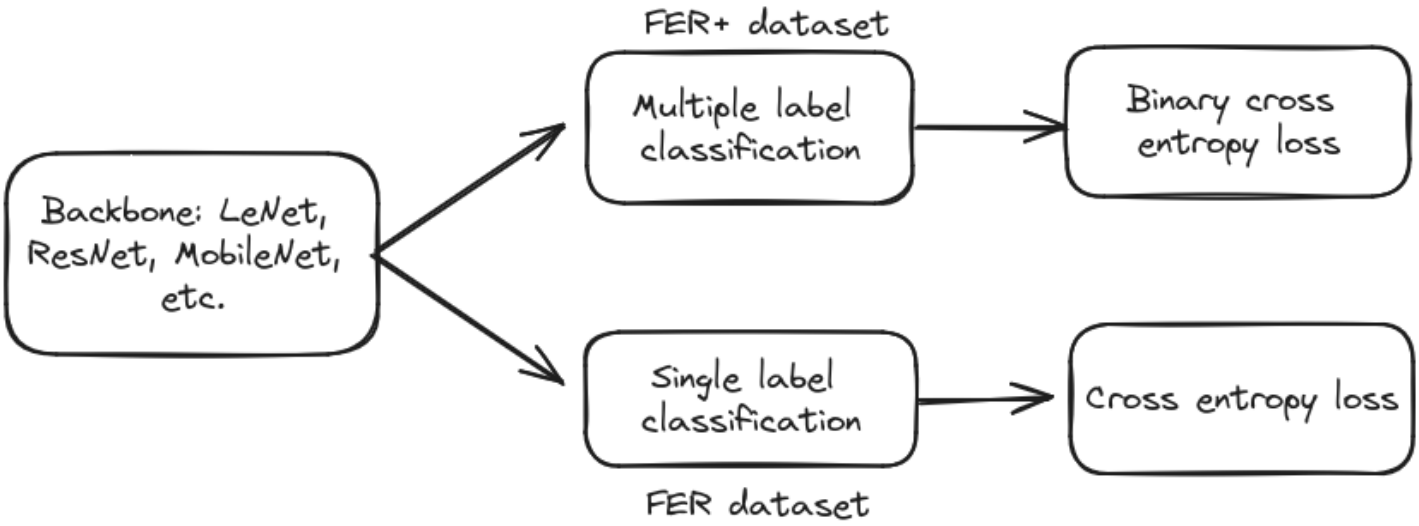


Figure 3: The overall architecture used for our experiments.

For the case of FER+ dataset, using the BCE Loss is reasonable in this case rather than usual cross entropy loss (which is often used for single label classification). In the FER+ dataset, a sample can be represented by multiple classes (as shown in Figure 1). Therefore, a binary cross entropy with logit loss seems to be more suitable in training neural network in this case.

3 Experimental results and discussions

3.1 Data augmentation

From the perspective of the model, it is felt that we may only want the random horizontal flip that flips the view points of the images from left to right (or vice versa). We can also add other type of augmentations. Apart from horizontally random flipping, I have also added the RandAugment method in order to increase the training data.

3.2 Experimental results

In the experimental result, we have also added the results getting from mobileNet v2 architecture but it is not explained in the Implementation details. However, we still want to see how it performs on both

datasets FER and FER+. As FER and FER+ datasets have 3 subsets: "Training", "PublicTest" and "PrivateTest", we use them as training, validating and testing subset respectively. We apply augmentation methods (Random horizontal flip and RandAugment) only for the training dataset to increase the data used for training model. We then validate the model by choosing the best model among all the epochs that produces best result on the validation set (in other words, the "PublicTest"). After that, we produce the test result in the test set ("PrivateTest" subset). For the optimizer of the model, we choose Adam as a stochastic gradient descent method to optimize our models with a learning rate of 0.001. Furthermore, we train each model for 50 epochs.

The following tables show the results we obtain for the Fer and FER+ dataset with RandomHorizontalFlip and RandAugment method for data augmentation.

Table 1: Classification Report for LeNet tested in FER+ dataset with an accuracy of 56.72%

Emotion	**Precision**	**Recall**	**F1-Score**	**Support**
Neutral	0.57	0.63	0.60	822
Happiness	0.77	0.81	0.79	833
Surprise	0.65	0.72	0.68	303
Sadness	0.51	0.10	0.17	229
Anger	0.52	0.35	0.42	199
Disgust	0.00	0.00	0.00	8
Fear	0.80	0.09	0.17	43
Contempt	0.00	0.00	0.00	9
Micro Avg	0.66	0.62	0.64	2446
Macro Avg	0.48	0.34	0.35	2446
Weighted Avg	0.64	0.62	0.61	2446
Samples Avg	0.43	0.43	0.43	2446

Table 2: Classification Report for ResNet 18 tested in FER+ dataset with an accuracy of 67.05%

Emotion	**Precision**	**Recall**	**F1-Score**	**Support**
Neutral	0.63	0.89	0.74	822
Happiness	0.92	0.90	0.91	833
Surprise	0.59	0.96	0.73	303
Sadness	0.57	0.66	0.61	229
Anger	0.59	0.88	0.71	199
Disgust	0.43	0.38	0.40	8
Fear	0.41	0.63	0.50	43
Contempt	0.15	0.22	0.18	9
Micro Avg	0.68	0.87	0.77	2446
Macro Avg	0.54	0.69	0.60	2446
Weighted Avg	0.71	0.87	0.77	2446
Samples Avg	0.61	0.61	0.61	2446

Table 3: Classification Report for ResNet 34 tested in FER+ dataset with an accuracy of 67.28%

Emotion	**Precision**	**Recall**	**F1-Score**	**Support**
Neutral	0.66	0.86	0.75	822
Happiness	0.93	0.90	0.91	833
Surprise	0.70	0.89	0.78	303
Sadness	0.44	0.80	0.57	229
Anger	0.58	0.85	0.69	199
Disgust	0.14	0.12	0.13	8
Fear	0.45	0.53	0.49	43
Contempt	0.33	0.11	0.17	9
Micro Avg	0.69	0.86	0.77	2446
Macro Avg	0.53	0.63	0.56	2446
Weighted Avg	0.72	0.86	0.78	2446
Samples Avg	0.60	0.60	0.60	2446

Table 4: Classification Report for MobileNet v2 tested on FER+ dataset with an accuracy of 62.62%

Emotion	**Precision**	**Recall**	**F1-Score**	**Support**
Neutral	0.56	0.85	0.68	822
Happiness	0.88	0.87	0.88	833
Surprise	0.65	0.86	0.74	303
Sadness	0.47	0.47	0.47	229
Anger	0.61	0.64	0.63	199
Disgust	0.33	0.12	0.18	8
Fear	0.68	0.40	0.50	43
Contempt	0.00	0.00	0.00	9
Micro Avg	0.66	0.79	0.72	2446
Macro Avg	0.52	0.53	0.51	2446
Weighted Avg	0.68	0.79	0.72	2446
Samples Avg	0.55	0.55	0.55	2446

Table 5: Classification Report for LeNet tested on FER dataset

Emotion	**Precision**	**Recall**	**F1-Score**	**Support**
Surprise	0.44	0.47	0.45	489
Fear	0.41	0.16	0.23	55
Angry	0.38	0.21	0.27	527
Neutral	0.67	0.85	0.75	876
Sad	0.43	0.38	0.40	589
Disgust	0.65	0.66	0.66	414
Happy	0.50	0.55	0.53	624
Accuracy			0.54	3574
Macro Avg	0.50	0.47	0.47	3574
Weighted Avg	0.52	0.54	0.52	3574

Table 6: Classification Report for ResNet 18 tested on FER dataset

Emotion	**Precision**	**Recall**	**F1-Score**	**Support**
Surprise	0.58	0.61	0.59	489
Fear	0.78	0.65	0.71	55
Angry	0.51	0.57	0.54	527
Neutral	0.89	0.85	0.87	876
Sad	0.52	0.54	0.53	589
Disgust	0.83	0.76	0.79	414
Happy	0.64	0.62	0.63	624
Accuracy			0.67	3574
Macro Avg	0.68	0.66	0.67	3574
Weighted Avg	0.68	0.67	0.67	3574

Table 7: Classification Report for ResNet 34 tested on FER dataset

Emotion	**Precision**	**Recall**	**F1-Score**	**Support**
Surprise	0.67	0.53	0.59	489
Fear	0.78	0.58	0.67	55
Angry	0.57	0.43	0.49	527
Neutral	0.87	0.89	0.88	876
Sad	0.52	0.56	0.54	589
Disgust	0.79	0.82	0.80	414
Happy	0.58	0.73	0.64	624
Accuracy			0.68	3574
Macro Avg	0.68	0.65	0.66	3574
Weighted Avg	0.68	0.68	0.67	3574

Table 8: Classification Report for MobileNet v2 tested on FER dataset

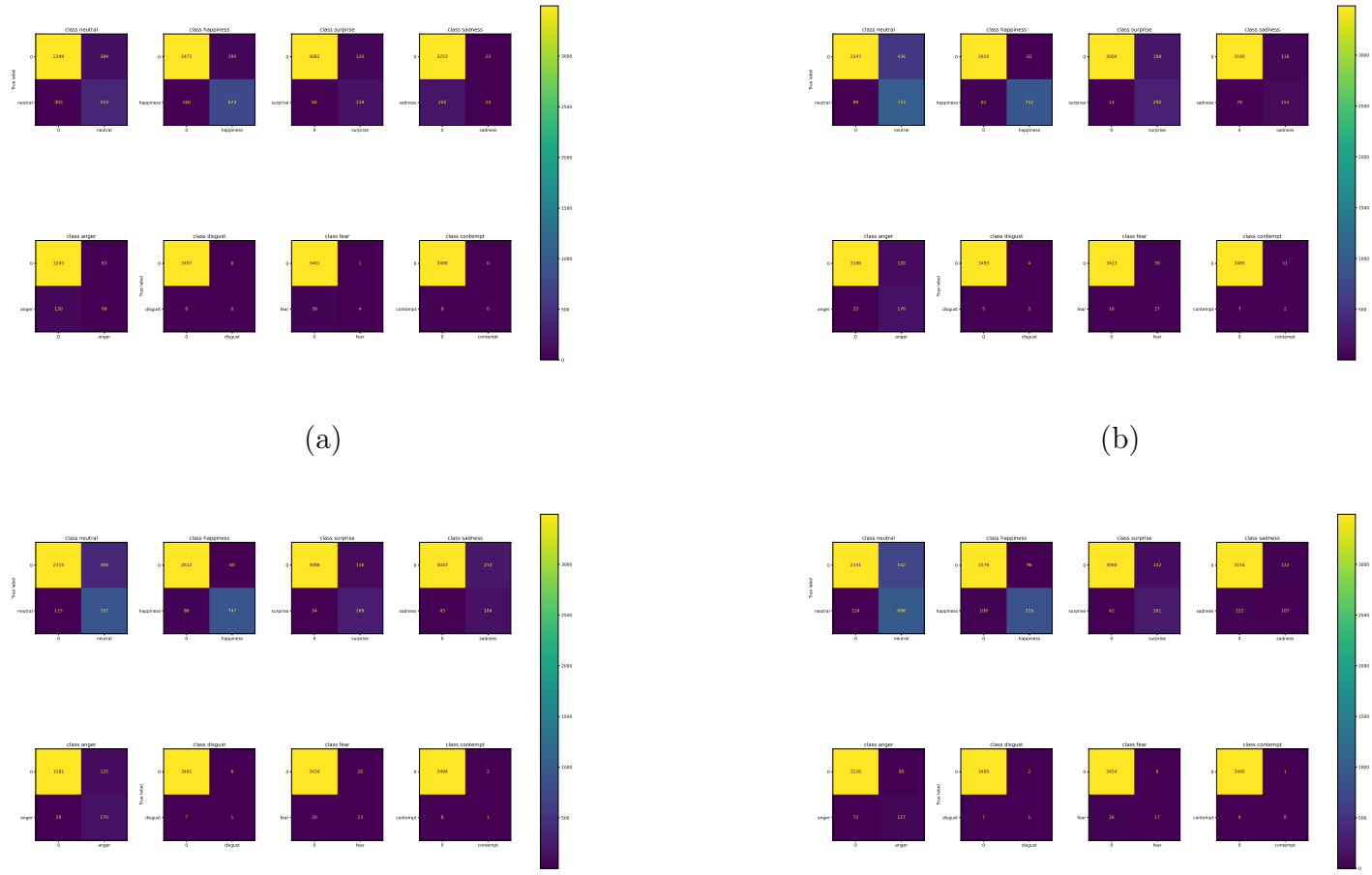
Emotion	**Precision**	**Recall**	**F1-Score**	**Support**
Surprise	0.56	0.46	0.51	489
Fear	0.57	0.56	0.57	55
Angry	0.47	0.28	0.35	527
Neutral	0.82	0.80	0.81	876
Sad	0.45	0.56	0.50	589
Disgust	0.60	0.80	0.68	414
Happy	0.54	0.59	0.56	624
Accuracy			0.60	3574
Macro Avg	0.57	0.58	0.57	3574
Weighted Avg	0.59	0.60	0.59	3574

As we can see from tables 1, 2, 3 and 4 show the results on the FER+ dataset. First of all, it can be seen that both LeNet and mobileNet v2 fail to recognize any test samples for contempt emotion. Furthermore, LeNet also fails to recognize the disgust emotion. For ResNet architectures, they recognize some of the samples but the number of correct predictions for these two classes are not so good as the precision, recall and f1-score are very low. For other classes, the ResNet models do pretty well especially for Happiness class. It can also be seen that the scores are effected by the number of training samples for each class in the training dataset. Tables 5, 6, 7 and 8 also follow the same tendency for FER dataset as ResNet 34 performs best.

In the above results, we can see that ResNet 34 outperforms other networks in terms of accuracy for both FER and FER+ datasets. The least performant architecture is LeNet as it seems to be shallow to capture enough important features of facial emotions. Furthermore, ResNet 18, despite of its smaller model sizecomparing to ResNet 34, it performs really well on the FER and FER+ dataset and its results on both dataset are very close to the results produced by ResNet 34. Hence, the deeper the network is, the more important features to get good classification rates. However, from ResNet 18 and ResNet 34, we can see that the results are very close on both datasets. It can be said that both have captured meaningful features but ResNet 34 also generates other unnecessary features that do not help improving a lot the results.

Figure 4 illustrates the confusion matrices of the 4 models trained on the FER+ dataset. We plot a confusion matrix for each class to compare with the rest of the dataset. It can be seen that the neutral and happiness classes are the one that have less false positives and false negatives for noth ResNet architectures as their number of samples in the dataset seems to be high in terms of proportion. Therefore, it seems to be reasonable that they produced better recognition. For LeNet, the results are much lower for each class compared with the other three models. For MobileNet v2, it performs really well for the happiness class but it has a lot of false positives for the neutral class.

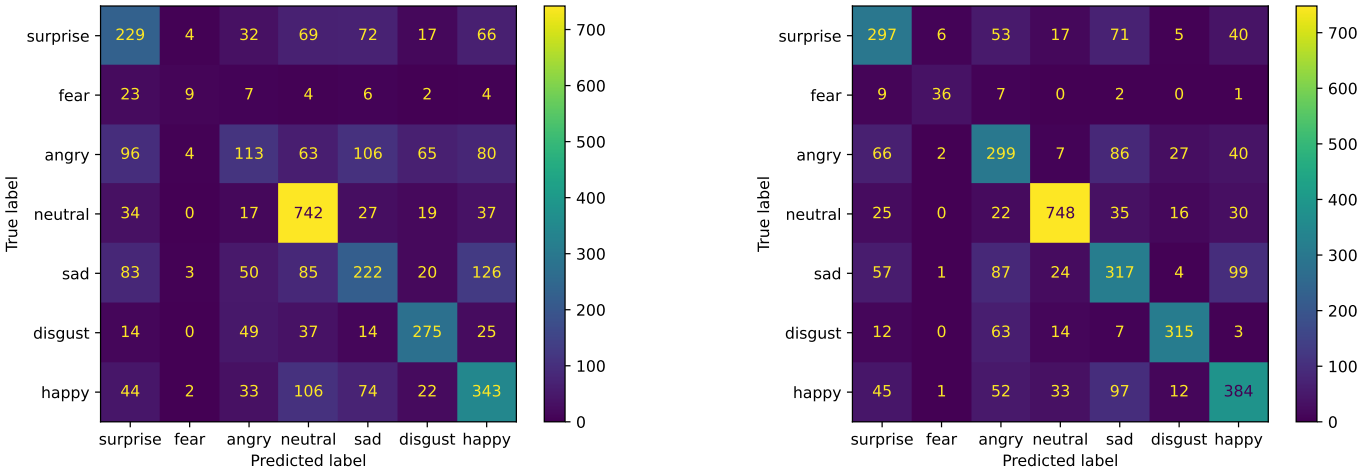
Meanwhile, Figure 5 demonstrates the confusion matrices for the FER dataset. Again, the matrices performed on FER also follow the same tendency as the FER+ dataset. Again, the classification results in both ResNet architectures are a lot better than the other 2 models. It can be seen from the matrices that a lot of angry samples are confused with the sad emotion. Beside, the happy and sad emotion are also confused between each of them too. These confusions happen to be in all the four trained models.



(a) (b)

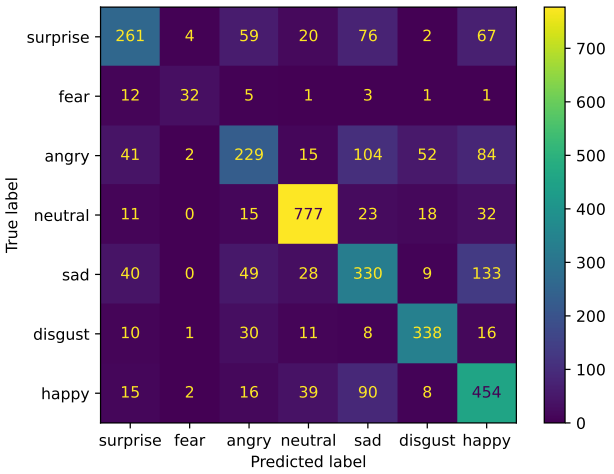
(c) (d)

Figure 4: Confusion matrices for the models trained on FER+ dataset: LeNet (a), ResNet 18 (b), ResNet 34 (c) and MobileNet v2 (d).

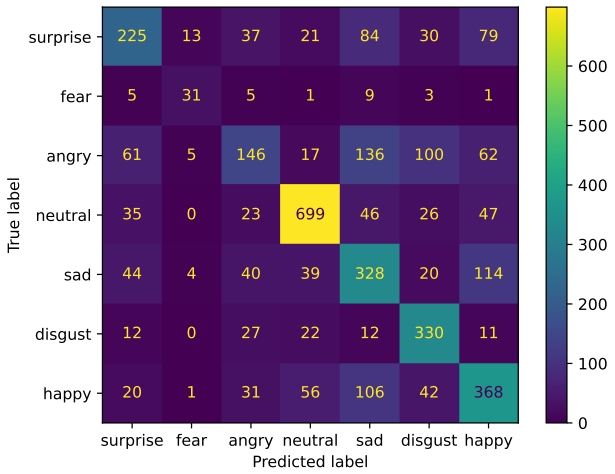


(a)

(b)



(c)



(d)

Figure 5: Confusion matrices for the models trained on FER dataset: LeNet (a), ResNet 18 (b), ResNet 34 (c) and MobileNet v2 (d).

Perhaps, some of the samples of sadness and happiness are difficult to distinguish the differences.

4 Conclusions and perspectives

In conclusion, the experimental results show that ResNet 34 seems to outperform other CNNs, although ResNet 18's accuracy is not so far behind. The experimental results are below the expected numbers. One of the reason to explain this is because of the imbalance of the dataset where some classes have a lot more samples than others. This can lead the trained models tend to better classify the samples from dominant classes. The experimental results have shown the consequences of such dataset.

Despite these results, further investigations and perspectives can be done to improve such problems produced by the datasets. These can be:

- Augmentation: augmentation for dominated classes can help us to balance the dataset where the number of samples in each class are equivalent to others.
- Redesign the models: with the recent advance in the deep learning community, vision transformer models can be viewed as an alternative to CNNs although the complexity is higher. However, optimized vision transformer models can play a role in improving the results.
- Exploring the loss surfaces to help generalising the model, helping the model to perform well for every classes. Some of the existing methods include sharpness-awareness minimization.