Computer Vision - Assignment 2

Gianluca Vico i6183186 - Alexandra Gianzina i629435

1 Introduction

The goal is to implement a classifier in order to predict the emotion of given images with faces. The emotion model used includes the six basic emotions (anger, disgust, fear, happiness, sadness and surprise) plus neutral. In addition, we aim to use our classifier for real time classification on videos.

2 Approach

2.1 Dataset

We are using the Facial Expression Recognition (FER) dataset, which consists of 35887 48x48 gray images of faces and a label for each image. The FER dataset is already split in train (80%), validation (10%) and test (10 %). We are not performing data augmentation or pre-processing the images. but we are using the dataset as it is.

Figure 1 shows the first image of the train set. As we can see, the faces are not all frontal and centred. This could help producing a more robust model.

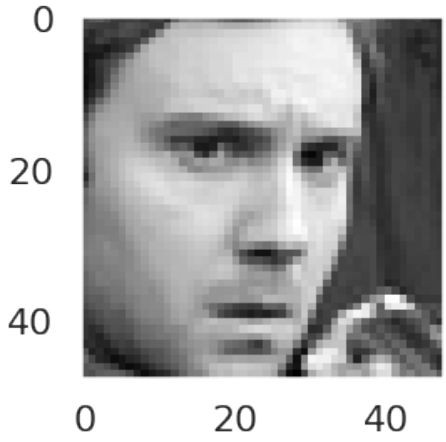

Figure 1: Sample image from the train set. Its label is 0 - "Angry".

Each image is labelled with an emotion: 0 - "Angry", 1 - "Disgust", 2 - "Fear", 3 - "Happy", 4 - "Sad", 5 - "Surprise", and 6 - "Neutral".

Figure 2 show the frequencies of the emotions in the train, validation and test sets. The labels have similar distribution in the three subsets. However, the FER dataset is imbalanced and, for example, 1 - "Disgust" is not well represented, while 3 - "happy" dominates the dataset.

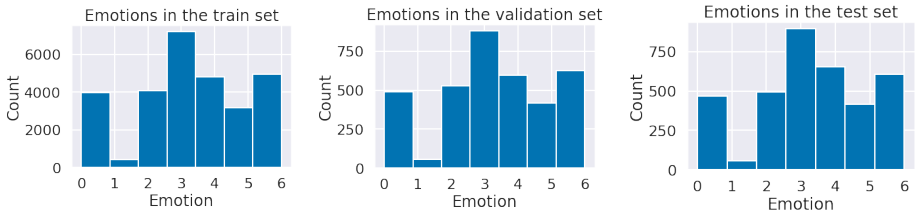

Figure 2: From left to right, the frequency of emotions in the train, validation and test set. The y-axis reports the number of occurrences for each label.

2.2 Models

We compare three different models. The loss function is cross-entropy for all the models. The models are compared on the F1 score: the dataset is unbalanced, hence, the accuracy is not appropriate, and we do not have reason to prefer a model with only high recall or only precision. We report accuracy, precision, recall, F1 score, and the confusion matrix

The first model one is a small convolutional network with convolution and max pooling. Due to this model's size and complexity we assume that this model should not require much data to train. Figure 3 shows a schema of this model.

The second model is deeper and it uses residual blocks. This model requires more data than the previous one since it has more parameters. Figure 4 shows a schema of this model.

The last model connects the first layers with some deeper layers in a way similar to U-net. This might not be that most appropriate model for classification because, after the convolutional part, we have a latent space with a size similar to the input size. Figure 5 shows a schema of this model.

All the models end with a classification head made up of one or more dense layer with ReLU. Softmax is used as activation function for the output.

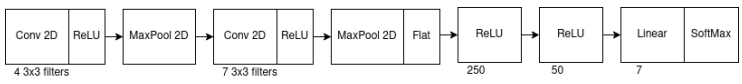

Figure 3: Schema of the first model. We indicate the number and the size of the filters for the convolutional layers and the output size of the dense layers. The stride is always 1.

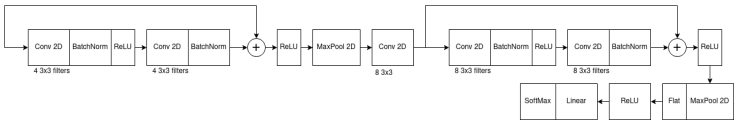

Figure 4: Schema of the second model. We indicate the number and the size of the filters for the convolutional layers and the output size of the dense layers. The stride is always 1. For the residual blocks, we use zero padding to keep the same size before and after the block.

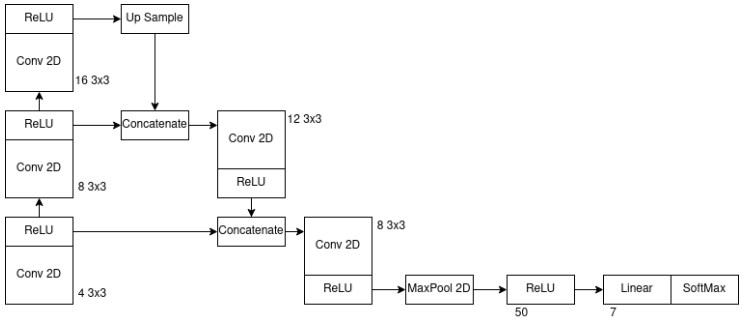

Figure 5: Schema of the third model. We indicate the number and the size of the filters for the convolutional layers and the output size of the dense layers. The stride is always 1. The concatenation is performed on the channel axis.

We use a classifier that predicts only 3 - "Happy" as baseline model (dummy and happy :)).

3 Experiments and results

3.1 Training and testing the models

We train each model for 100 epochs with a learning rate of 0.001. The optimizer is Adam and the loss function is cross-entropy. We plot the training and validation loss and measure the validation accuracy. Figures 6, 7, and 8 shows the loss and accuracy on the validation set for the three models.

We can notice that all three models quickly overfit the training set. The three models achieve the minimum validation loss after respectively 21, 11, and 16 epochs.

Table 1 outputs the various metrics for each model on the test set. The output of the models is a distribution of the labels. For computing the metrics, we used the label with the highest value. We can notice that all three models outperform the baseline, but the second model achieve the best results. The worse model, besides the baseline, is model three. We think this is due to the convolutional architecture used.

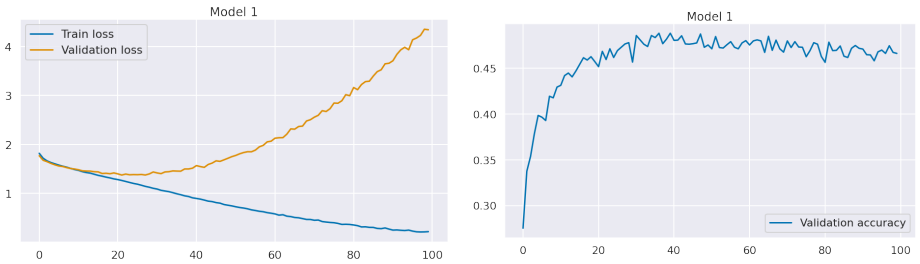

Figure 6: Model 1 losses and validation accuracy

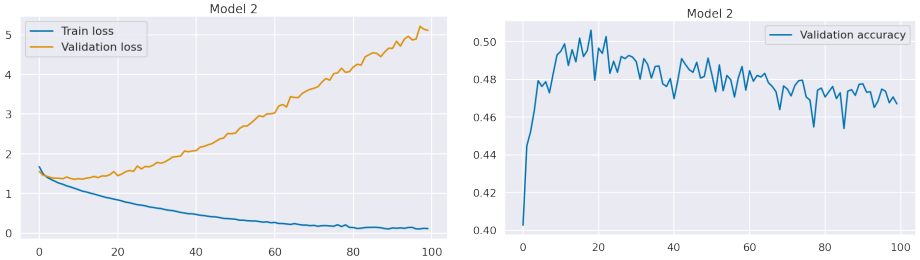

Figure 7: Model 2 losses and validation accuracy

Tables 2, 3, and 4 report the confusion matrices of the models on the test set. We can notice that the first model tends to classify many pictures as 0 - "Angry" although they depict other emotions. Also, it fails to classify 2 - "Fear": in this case the label seems random. 1 - "Disgust" is often classified as 0 - "Angry".

The second model is better at classifying 1 - "Disgust", but most of the time it is still classified as 0 - "Angry". It is also better at classifying 2 - "Fear", although it is sometimes confused 3 - "Sad" or 6 - "Neutral". Also, we can notice than 4 - "Sad" is confused with 6 - "Neutral" or 3 - "Happy".

The third model never predicts 1 - "Disgust". 0 - "Angry" is mostly classified as 0 - "Angry" or 3 - "Happy". Also, 2 - "Fear" and 4 - "Sad" are mostly classified as 3 - "Happy" rather than the correct class. 5 - "Surprise" is classified relatively correctly.

In general, we can see that classifying 1 - "Disgust" is hard for the models. The diagonal tends to have the largest values for all the models, however, many samples are misclassified.

3.2 Real-time classification

For generic videos we cannot assume that the face is centred and that the video is 48x48 gray, therefore we need to detect the faces. Also, we need to run both the face detector and emotion classifier while recording/reproducing the video.

For the face detector, we use an Haar classifier from OpenCV. When we

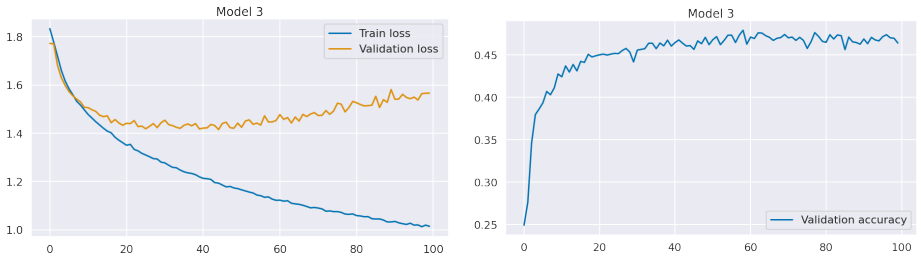

Figure 8: Model 3 losses and validation accuracy

	Accuracy	**Recall**	**Precision**	**F1**
Baseline	0.24	0.14	0.04	0.05
Model 1	0.47	0.40	0.47	0.40
Model 2	**0.51**	**0.46**	**0.50**	**0.47**
Model 3	0.44	0.36	0.36	0.34

Table 1: Results of the models on the test set. The highest scores are in bold.

detect a face, we crop and resize it. Then we use the second model trained on the entire dataset for the prediction of the emotion.

We plot squares with the label on the video to visualize the predictions.

During the video, the subject might show different emotions or the classifier might predict the wrong labels. So, we report the frequency the different labels predicted during the video to have an indication of which is the general emotion in the video. For example, if only one frame is classified as 0 - "Angry" is could a wrong classification; if 40 % of the predictions are 5 - "Surprise" and 40 % are 3 - "Happy", then these two emotions are shown in the video.

Table **??** shows an example of emotions predicted from a video. Figure 9 shows the real-time classification.

3.3 Visualization

We try to visualise the second model by plotting its convolutional filters and weights and by plotting the activation maps.

Figure 10 shows the filters in each layer.

We can observe that the third filter on the first layer might detect corners, while the second and the forth might be able to detect some edges. However, filter in the deeper layers are not directly interpretable. We can only notice that the weights in the first dense layer do not show any particular pattern.

In Figures 11, 12 we display the activation maps for two different images. In particular we focus on the two residual blocks, the convolutional layer in between, and the activation before the softmax.

We can think that first two channels after the residual block detect the surface of the face, while the other two seems to focus on the edges of the face and

	Confusion Matrix for Model 1						
	Angry	**Disgust**	**Fear**	**Happy**	**Sad**	**Surprise**	**Neutral**
Angry	178	2	32	98	66	16	75
Disgust	21	5	4	11	8	1	6
Fear	88	1	97	89	98	53	70
Happy	59	0	20	695	53	20	48
Sad	124	0	54	135	211	28	101
Surprise	31	0	38	40	27	253	26
Neutral	89	1	31	120	91	20	255

Table 2: Confusion Matrix for Model 1. The rows are the true labels, while the columns are the predicted labels.

	Confusion Matrix for Model 2						
	Angry	**Disgust**	**Fear**	**Happy**	**Sad**	**Surprise**	**Neutral**
Angry	178	2	46	76	72	15	78
Disgust	16	15	3	8	7	0	7
Fear	64	5	131	64	92	54	86
Happy	37	0	25	711	32	19	71
Sad	88	1	58	120	223	10	153
Surprise	28	3	36	28	17	268	35
Neutral	63	1	31	100	87	12	313

Table 3: Confusion Matrix for Model 2. The rows are the true labels, while the columns are the predicted labels.

some landmarks on the face. In the activation map of the convolutional layer, the 7^{th} map seems to focus everything but the eyes. In the second residual block, the second map is highlighting in particular the eyes. So, it seems reasonable that the eyes are an important feature to classify emotions.

The output layer is clearly highlighting the predicted label.

4 Discussion and conclusion

We compare three different architecture on the FER dataset. A model with residual blocks gives the best results, however many samples are still misclassified. Moreover, the model can be used for real-time classification. Finally, we attempted to interpret the internal representation of the network.

	Confusion Matrix for Model 3						
	Angry	**Disgust**	**Fear**	**Happy**	**Sad**	**Surprise**	**Neutral**
Angry	137	0	42	131	38	28	91
Disgust	17	0	9	11	6	1	12
Fear	63	0	93	123	38	72	107
Happy	29	0	28	720	32	24	62
Sad	98	0	57	207	105	32	154
Surprise	25	0	31	56	11	257	35
Neutral	64	0	46	150	41	25	281

Table 4: Confusion Matrix for Model 3. The rows are the true labels, while the columns are the predicted labels.

	Frequency	**Relative Frequency**
Neutral	715	0.53
Sad	474	0.35
Happy	100	0.07
Fear	50	0.04
Angry	19	0.01
Surprise	1	0.00

Table 5: distribution of emotions in a video. I was trying to see which expression are classified as 3 - "Happy" and 4 - "Sad".

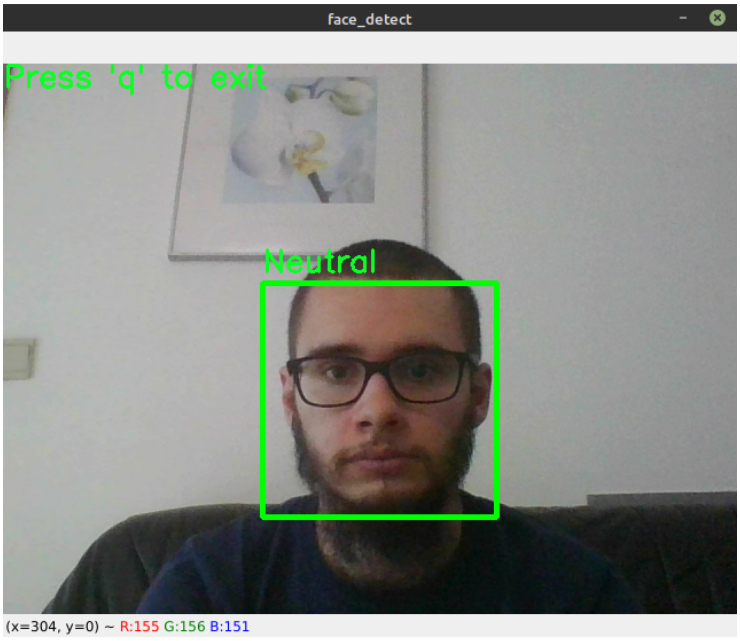

Figure 9: Example of the classifier used on a video recorded from a laptop webcam.

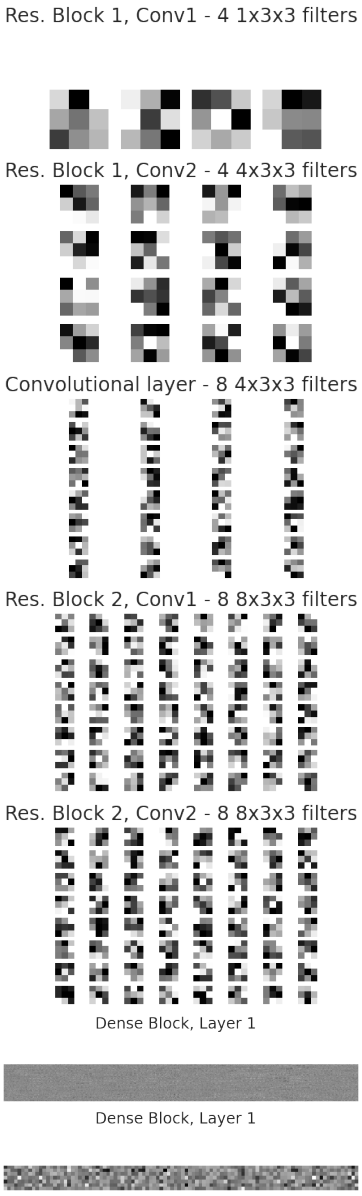

Figure 10: Filter in the convolutional layers and weights in the dense layers of the second model.

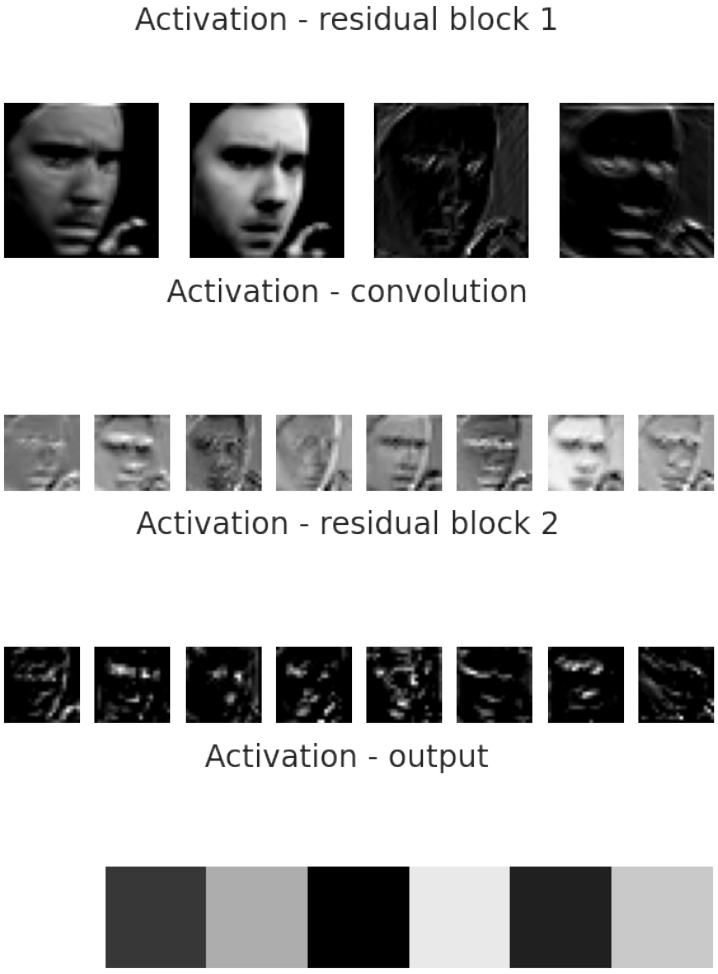

Figure 11: Activation maps of the second model for the first image

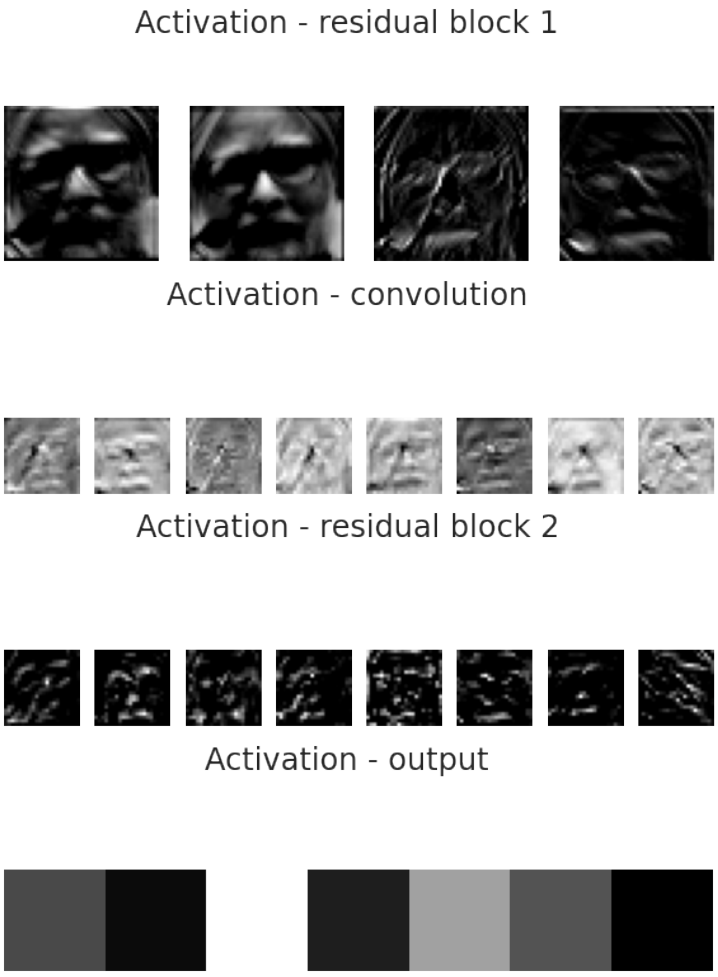

Figure 12: Activation maps of the second model for the second image