Emotions Classification using a Convolutional Neural Network

Raffaele Francesco Barbagallo RAFFAELE.BARBAGALL92@GMAIL.COM

1. Model Description

FER_KD is a deep model that consists of 5 convolutional layers and 1 full-connected layer to perform multiclass classification (see Fig. 1). Input pictures (48x48 faces pictures) flows from the beginning to the end of the model. The model ultimately classify the emotion of the picture in one of the 7 classes (angry, disgust, fear, happy, sad, surprise and neutral).

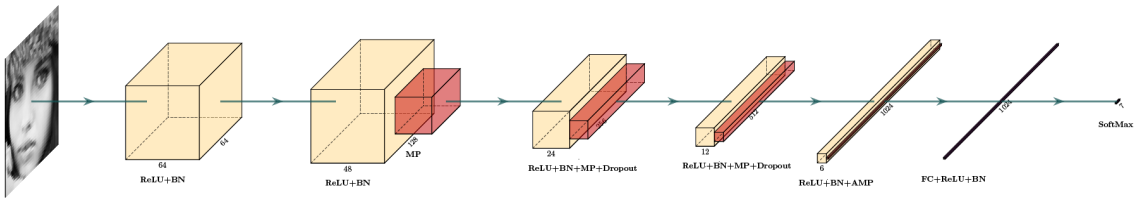


Figure 1: A simple representation of the proposed model.

The model have two initial convolutional layers (64x48x48 and 128x48x48 out channels) followed by a max pooling (128x24x24). The third (256x24x24) and the fourth (512x12x12) convolutional layers are followed by max pooling (256x12x12 and 512x12x12) and each one has a dropout of 0.25. The final convolutional layer (1024 channels) is followed by an Adaptive Max Pooling. The output of convolutional layers is then flattened in 1024 neurons and flows into the final fully-connected layer. All of the convolutional layers have 3x3 kernels with 1 of padding and 1 of stride. Every layer, convolutional or fully connected, has a ReLU activation function and a Batch Normalization. The output classification layer consists in a softmax function that returns the class with the highest probability. Even if dropout is usually not advisable in convolutional layers, in this particular model it has lead to an improved final accuracy. The model has 10,479,111 parameters in total. The final softmax function is defined as:

$$\sigma(\overrightarrow{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{1}$$

where σ is the softmax function, $\overrightarrow{z}$ is the input vector, e^{z_i} is the standard exponential function for the input vector, K the number of classes in the multi-class classifier and e^{z_j} the standard exponential function for the output vector.

2. Dataset

The dataset is FER2013. It consists of 35887 grayscale face images of size 48x48 proposed for a kaggle challenge. Every one of these images is labeled with one of seven expressions/emotions (angry, disgust, fear, happy, sad, surprise and neutral). The dataset is very noisy and mislabels are common. This is why it is hard to have consistent results and, even when models are properly trained, it is very difficult to obtain an optimal accuracy. The dataset is divided in three subsets: Training, PublicTest and PrivateTest which are made of 28709, 3589 and 3589 images respectively. It has been decided to use Training as the training set, PublicTest as the test set (which was not given during the original contest) and PrivateTest as the validation set. During the training phase, in order to improve the model transforms have been applied to do data augmentation. In particular, images have been horizontal flipped and normalized. The horizontal flipping has been introduced to improve the training phase even though it is not important considering that pictures of the faces have been taken from different angles and positions. After the flipping transformation, a normalization with 0.5 mean and 0.5 std has been applied.

3. Training procedure

With the final model, an ablation study has been performed in order to understand how each adding or removing layers improved or worsened the results. During the ablation study, various number and types of layers have been added or removed to the model.

A cross-entropy loss has been employed to train, but also to validate, the model:

$$CE(\boldsymbol{y}, \sigma(\overrightarrow{z})_i) = -\sum_{i=1}^{\mathrm{K}} y_i \log(\sigma(\overrightarrow{z})_i) \tag{2}$$

where $\boldsymbol{y}$ is a one-hot label vector containing the correct label and K is the number of classes.

All the models have been trained for 15 epochs on Google Colab. The optimizer is Adam with 0.001 of weight decay and 0.0005 of learning rate. The batch size has been set to 64 and the training set has been shuffled by the data loader while validation and test sets have not.

4. Experimental Results

Several experiments were conducted to test the final model. Models were trained as described in the previous section. Table 1 shows test results for the proposed architectures as well as of ablation studies. eeeeeee

The final model is the best performing in the test dataset. It is important to notice, though, that the model with 2 layers more than the chosen model (1 fully-connected and 1 convolutional) performs better on the validation set. The difference between the two accuracies is not large enough to choose the more complex model though. It is also clear that convolutional layers seem to improve the model better than the fully-connected ones. When adding one fully-connected layer, the model shows less improvements than when adding a convolutional layer. At each decrease in the number of convolutional layers, the model shows a noticeable decrease in performance.

Model	Validation	Test
+ 1 FC + 1 Conv Layers	**64.73**%	63.42%
+ 1 Conv Layer	64.42%	63.14%
+ 1 FC Layer	64.04%	62.23%
FER_KD	64.42%	**63.68**%
- 1 Conv Layer	61.90%	60.04%
- 2 Conv Layers	57.05%	55.58%
- 3 Conv Layers	54.23%	52.94%
- 4 Conv Layers	48.77%	47.30%

Table 1: Accuracy for validation and test datasets during the ablation studies. Best font indicates the best accuracies.

5. Results

From the confusion matrix (Fig. 2) and the final metrics it is possible to draw some conclusions. First of all, it can be seen that some classes are over-represented like happiness, while others have a small number of observations (disgust has only 56).

Considering that happiness is the most-represented expression, it is not strange that the model finds it easier to recognise it them compared to, for example, disgust.

When the model predicts disgust, it rarely is a different emotion (its precision is 93% (Tab. 2), but it has a recall of 46% which is low. Fear has the poorest performance compared to other emotions, followed by sadness and neutral.

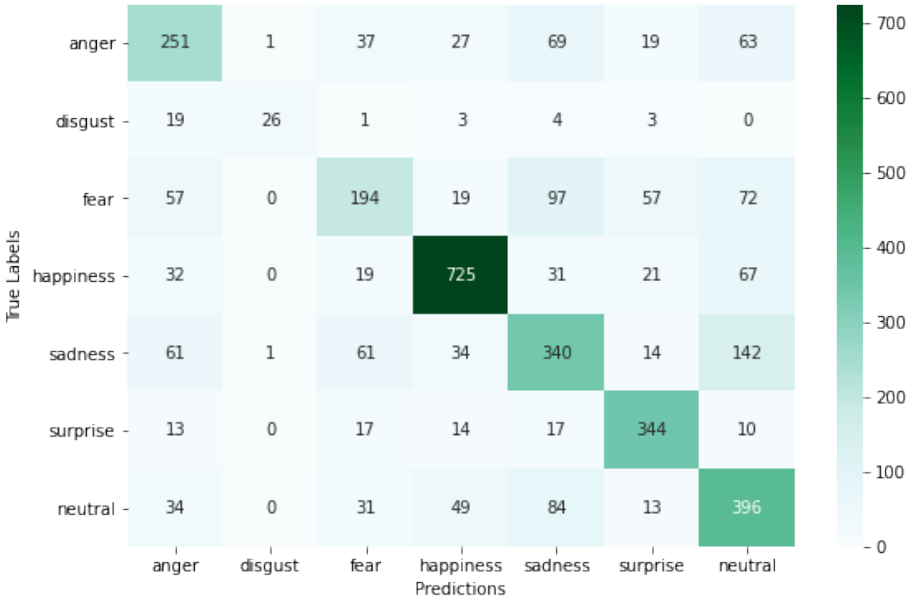


Figure 2: The confusion matrix of the final model..

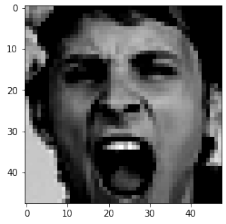
	Precision	Recall	F1-score	Support
Class Results				
Anger	0.54	0.54	0.54	467
Disgust	0.93	0.46	0.62	56
Fear	0.54	0.39	0.45	496
Happiness	0.83	0.81	0.82	895
Sadness	0.53	0.52	0.53	653
Surprise	0.73	0.83	0.78	415
Neutral	0.53	0.65	0.58	607
Metrics				
Accuracy			0.63	3589
Macro AVG	0.66	0.60	0.62	3589
Weighted AVG	0.64	0.63	0.63	3589

Table 2: Metrics for the results of the final model.

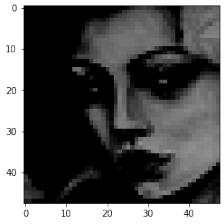
The task of understanding emotions, could be difficult even for humans and as it has been said before and there are a lot of mislabeled faces.

Fig. 3, contains an example of a correct prediction and an example of a wrong prediction. It is clear that the mislabeled picture, is difficult to label even for humans and it also seems to be more similar to the predicted label (neutral) than to anger. (Barsoum et al., 2016) tried to solve this issue by creating FER+, which is a set of new labels for the FER2013 dataset. Every image has multiple possible labels and this allow to solve the noisy label problem by also using probability distributions to determine the category of each image. FER+ allowed for enormous improvements in facial expressions recognition (e.g.the model proposed by (Khaliluzzaman et al., 2019) obtained 69.10% of accuracy on the FER2013 and 86.54% of accuracy on the FER+).

It could be reasonable to test the proposed model also on better datasets, like the previously cited FER+, to see how it is good and how much the obtained results depend on the quality of the dataset.



(*a*) Label:Anger
Predicted:Anger



(*b*) Label:Anger
Predicted:Neutral

Figure 3: Examples of model's predictions

References

Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *ACM International Conference on Multimodal Interaction (ICMI)*, 2016.

M. Khaliluzzaman, S. Pervin, M. R. Islam, and M. M. Hassan. Automatic facial expression recognition using shallow convolutional neural network. In *2019 IEEE International Conference on Robotics, Automation, Artificial-intelligence and Internet-of-Things (RAAICON)*, pages 98–103, 2019. doi: 10.1109/RAAICON48939.2019.42.