Ukrainian Catholic University

Faculty of Applied Sciences

Data Science Master Programme

Facial Expression Recognition in Natural Images

Machine Learning final project progress report

Authors:
Anastasiia Khaburska
Andrii Yurkiv

27 April 2019

1 Introduction

Facial expression is one or more motions or positions of the muscles beneath the skin of the face. It is one of the most important forms of nonverbal communication and the primal mean of social information exchange between humans.

It is assumed that certain facial expressions and gestures correspond to specific emotions (for instance, happiness is associated with laughter and smiling, sadness with tears, anger with clenched jaw, fear with grimace, surprise with raised eyebrows and wide eyes and disgust with wrinkled nose and squinted eyes) and are recognized by humans regardless of culture, language or time. But in general, this hypothesis has not been scientifically verified and received both critical and supportive reviews.

Both sides of this scientific debate agree that the face expresses emotion. The controversy surrounds the uncertainty about what specific emotional information is read from a facial expression [2].

2 Motivation

It has been proved multiple times that the majority of information human perceives (up to 83%) during message absorption is obtained through eyesight. For this purpose body language in general, and facial expression, in particular, are the things which provide the most essential and specific information about the intentions and emotional state of the message source.

It's not hard to conclude that accurate perception of facial expressions is a key to effective face-to-face communication.

Humans have a great ability to perform this kind of tasks. And since main facial expressions are universal and do not vary with culture or environment, we usually recognise others emotional state pretty well (at least when our interlocutors are not intentionally hiding it).

But for a machine, it is not an easy task. There are few reasons why machine learning model usually perform much worse compared to humans in recognising facial expressions (while outperforming in other classification tasks):

- Humans do not perceive facial expressions separately from other parts of body language. For facial expressions, context is very important, because it provides a great deal of additional information.
- Regardless of their universality, facial expressions are very diverse and vary substantially due to the race, age, gender, nationality and culture.
- People are brilliant at determining how they feel. But in some cultures, it is normal to hide your real emotions behind the neutral or happy facial expression. It

is especially distinguishable in western cultures, where excessive emotionality is not considered as an element of effective communication. Machine learning model doesn't have this prior knowledge and hence use a generalised approach to cases where cultural nuances are crucial.

- Some kinds of facial expressions are very similar (for example, without context, it's hard to distinguish surprise from fear, or sadness from neutrality). And image similarity is a fundamental property when the model is trying to determine class depicted on it. The model fails to generalise well and makes mistakes by focusing not on the wrong features.

In this project, we are going to develop real-time end-to-end machine learning system of facial expression recognition in natural images. It would consist of two parts: a face detector and expression recogniser. Unfortunately, we haven't found a data set to evaluate our end-to-end system (we need a dataset that contains images with faced in different contexts labelled with facial expression classes). So, we aim to achieve state-of-the-art performance on `fer2013` data set, which is equal to 75.2%

3 Data

The project idea was inspired by Kaggle competition "Challenges in Representation Learning: Facial Expression Recognition Challenge". It was organized in 2013, but we think that the topic is still relevant today. The `fer2013` dataset [1] for this competition was prepared by Pierre-Luc Carrier and Aaron Courville.

The `fer2013` data set consists of 48×48 pixel grayscale images of faces. The faces have been automatically registered so that the face is more or less centred and occupies about the same amount of space in each image. The task is to categorise each face based on the emotion shown in the facial expression into one of seven categories (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral).

The data set is divided into train, validation and test splits. The training set consists of $28,709$ examples. The validation set consists of $3,589$ examples. The final test set consists of another $3,589$ examples.

Expression	Train	Validation	Test	% of samples
Angry	3995	467	491	13.8
Disgust	436	56	55	**1.52**
Fear	4097	496	528	14.27
Happy	7215	895	879	**25.05**
Neutral	4965	607	626	17.27
Sad	4830	653	594	16.93
Surprise	3171	415	416	11.15

Table 1: Distribution of classes in the data set.

It is obvious that classes in the data set are not well-balanced. First of all, we have highly underrepresented `Disgust` class, which will for sure need some special treatment such as data augmentation in case if the model doesn't work well on it.

Regarding other classes, we still have a huge difference between some of them. For example, 25.05% of samples correspond to the `Happy` class, which is more than the fraction of `Angry` and `Surprise` altogether (24.95%).

This imbalance issue imposes additional activities in our development process:

- In case if the model performs poorly on underrepresented classes, we would need to perform data augmentation in order to make it more robust. Or remove samples from overrepresented classes.
- We would surely need to use other evaluation metrics, such as the confusion matrix to determine the weaknesses of the model and possible improvement strategies.

Besides the imbalance, `fer2013` data set contains trash samples (which do not contain a face) and several misclassified examples. These imperfections make the classification harder because the model has to generalise well and be robust to incorrect data. At this point, we don't see any solution to this problem that can be performed in a reasonable amount of time.

4 Baseline model

At this point, for face detection we are using pre-trained Haar Cascade model from `cv2` python module. Since we don't have a data set on which we can get asses both fact detection and facial recognition model performance, we would not concentrate on the implementation of this part.

In the next iteration, to make our pipeline more homogeneous, we are going to implement and train our own model for face detection.

Our baseline facial expression recognition model is a 4-layer CNN with 3 dense layers. The model architecture is visualized in figure 1.

On each layer except output one we used ReLU activation function. On the output layer we used Softmax function. Besides it, we applied the Dropout regularisation technique at each layer of our model because otherwise it didn't train properly and stuck in the local minima.

The details of implementation can be viewed in corresponding jupyter notebook: (`classification.ipynb`).

You can find the example of model results in figure 2 in the Appendix.

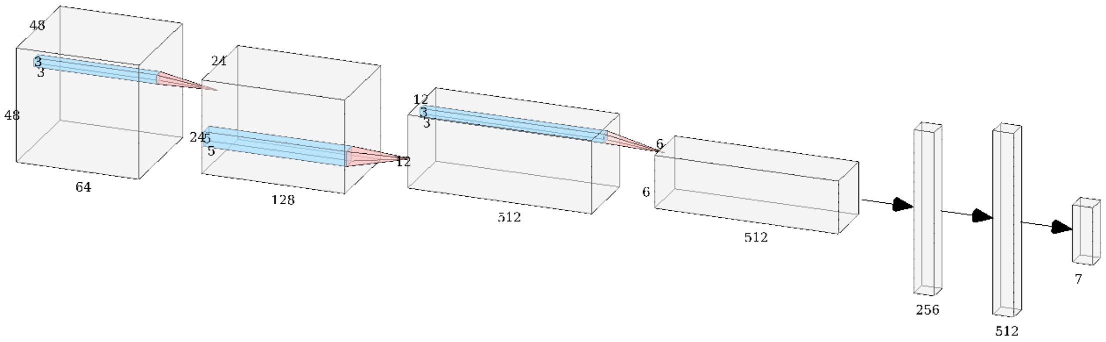


Figure 1: Baseline CNN model

5 Evaluation

Our baseline model gave 63.25% accuracy on the test dataset. In the original competition leaderboard, it would be in the top 10 submissions. We think that it is possible to get even higher accuracy by training the model for larger number of epochs. Still, it is only a baseline model, and in the future iteration we are going to improve this result.

During the data analysis stage of our project, we understood that one of the primary weaknesses of this data set is its imbalance. We haven't tackled this issue yet (plan to do it in the next iteration), but we understood that dealing with it is a major part of this particular problem solution.

The thing is that during the evaluation stage we've got very unexpected results. `Disgust` emotion while being extremely underrepresented in our data set is not the one that suffers most from misclassification. It is `Angry` and `Fear`, that model most often treats as other facial expression.

`Fear` is mostly confused with `Sad`. Maybe because often when a person experiences fear, it doesn't influence her facial expression and has an impact only on some inner feelings. In general, when we are afraid of something, we can have different facial expression, depending on the situation, surrounding and the cause of fear. Fear is a complex emotion no single facial expression corresponds to it.

From the confusion matrix, we can conclude 3 out of 6 facial expressions are often confused with `Sad`. It looks like this is the most obvious way for the model to describe the appearance if it is not sure what it sees.

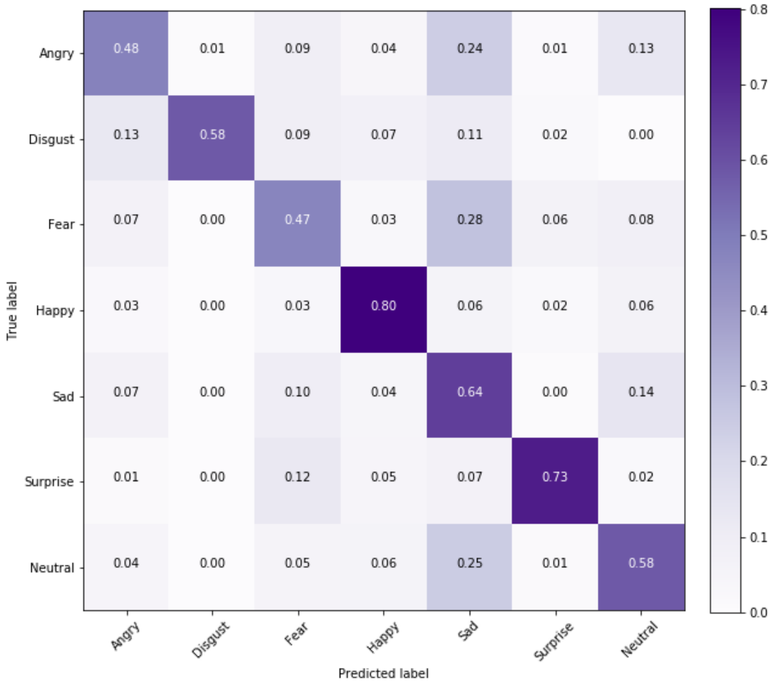


Figure 2: Normalized confusion matrix

One important thing should be noted here: the results produced by our the model does not confirm our primary assumption that accuracy of classification of images will be smaller for underrepresented classes. It proves one more time that the problem of facial expression recognition is not as simple as it seems from the first sight.

6 Future work

We plan to do the following steps during the next iterations:

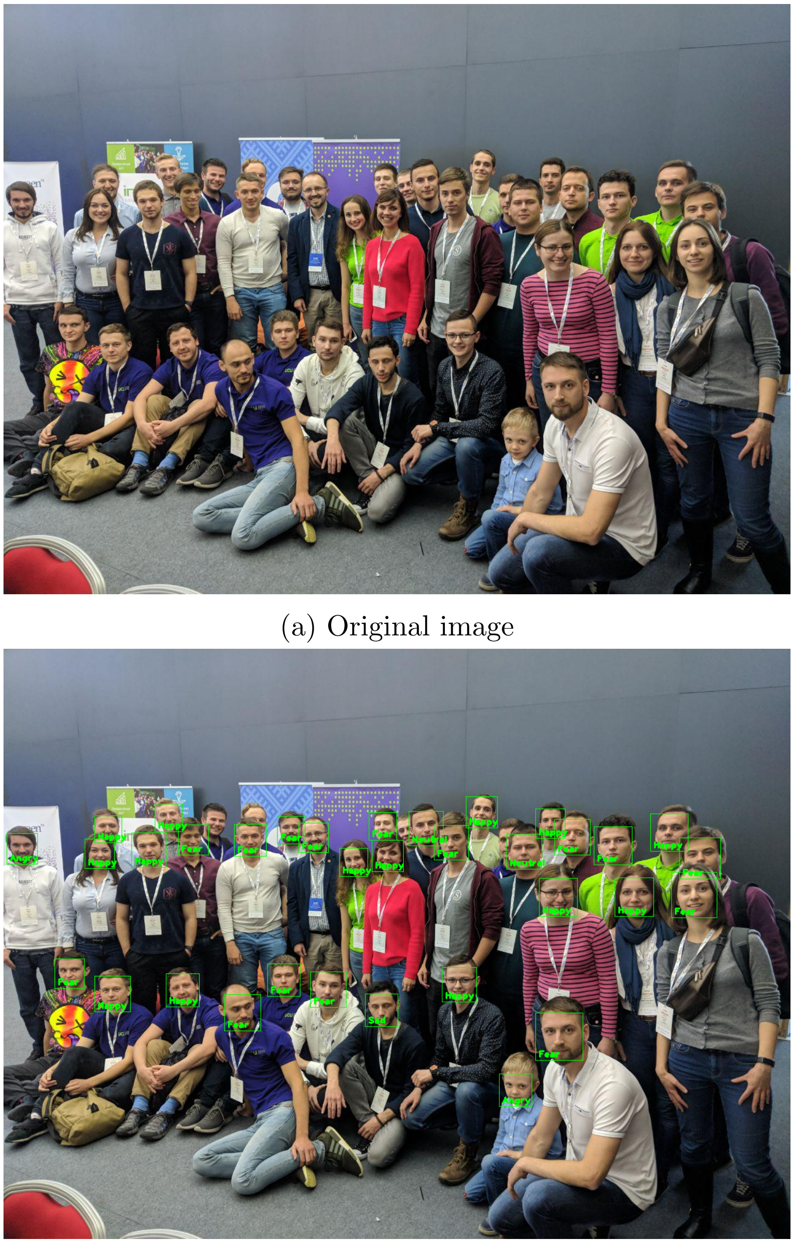
- Perform data augmentation for underrepresented classes and check whether it improves the results.
- Apply other CNN architecture to this problem: VGG, Inception, ResNet.
- Icrease number of features by adding Face Landmarks and HOG.
- Write script for facial expression recognition in videos.
- Perform some basic hyperparameter tuning (model is complex, so it's complicated to iterate fast).
- Try to solve misclassification problem for `Angry` and `Fear` classes.

References

[1] I. Goodfellow et al. *Challenges in Representation Learning: A report on three machine learning contests*. arXiv, 2013

[2] Wikipedia contributors *Facial expression* Facial expression

A Appendix



(a) Original image

(b) Predicted facial expressions

Figure 3: Baseline model results