Facial expression recognition application using CNNs

Peagno Eleonora
1237035

Peron Giovanni
1237783

Rossi Daniel
1211017

Abstract

The facial expression recognition is a challenging task in machine learning field, and there is an active research on this topic. Being able to make a machine understand the human emotions is a fascinating goal. The purpose of this report is to exploit all the topics studied during the Vision and Cognitive Services course and others like Machine Learning and Deep Learning in order to implement a system able to predict human emotions. We will describe how we realized this system specifying all the steps performed, from the first CNN we tried to the final model we obtained. We will illustrate all the procedure we used trying to achieve better results. Our target was not very high since this is a challenging task as we said, however with our final model we arrived to a predition accuracy of **...** *that is better than the human accuracy. Finally using this model we reached our goals implementing a nice application.*

1. Introduction

Facial expressions recognition (FER) is an interesting and a challenging problem in machine learning field. It is also a task that can be applied in many important applications. Facial expressions have an important role in every human interaction so having a machine able to recognize and understand human expressions automatically can be very useful in many existing and novel fields [5].
One of these fields is behaviomedics that are systems which exploit automatic analysis of affective and social signals to aid diagnosis, monitoring and treating medical condition that alter behavior. Facial expression recognition can be also use in data analytics field for example to understand emotions of people that are looking at ads or political debate and make statistics related to people's preferences. Another application field for Facial expression recognition is human-computer interaction, understanding human emotions would make the attitude of systems like vocal assistants or robots much closer to the way that humans interact with each other. Recognizing expressions could also be useful to improve the identification of micro facial expressions which can be used in deceit detection applications. Due to all these possible applications, facial expressions recognition is widely studied also because recognizing human expressions in natural condition environment is a very challenging task. With this project we aim to build a facial expressions classifier able to reach and overtake the human accuracy on this task. The main idea was focusing mostly on study different types of model in order to understand a good way to achieve valid outcomes. For this reason all preprocessing techniques that can be applied to the input data for improving classification results were not be consider.

The chosen dataset is the FER2013 it was selected after some researches, above some examples of images of this dataset are reported. FER2013 was the desired dataset, the most suited for our purpose. Training a model with a dataset like that was the challenge with the right level of difficulty we wanted. Moreover FER2013 provides a very large set of examples well differentiated in terms of person age, face pose, illumination and other realistic conditions. So using this dataset we realize an ensemble model using many types of Convolutional Neural Networks (CNNs), in order to achieve a test accuracy greater than 65.5%, that is the human accuracy measured on FER 2013 dataset [1]. Our target was reached using the final model described in the next sections, which achieve test accuracy of 71.52% on FER2013.

2. Related Work

The first step for reaching our targets it has been a research of all the scientific papers related to the facial expression recognition. We found many different works related to the topic all reviewed in *"Deep Facial Expression Recognition: A Survey"* by S. Li and W. Deng [4], thank to this paper we could find a lot of different works about the same kind of classification we were looking for. We searched for all the works reported that were using the FER2013 dataset, in order to understand a way to realize a model aligned with the best results achieved in the last years. We found six different papers related to the FER2013 dataset, they were classified by accuracy reached and type of neural networks used. After inspecting all these works we chose to follow the paper that achieved the best accuracy, moreover it used an ensemble method and we were interested in understanding better this approach. So we decided to follow the methods used in *"Facial Expression Recognition using Convolutional Neural Networks: State of the Art"* by C. Pramerdorfer and M. Kampel [6], that consists on an ensemble method made up eight different CNNs of three different types, the accuracy reported for this method was 75.2%. This paper has been enticing for us because in it are explained the power and the bottlenecks of the CNNs method, moreover the final part of that paper was destined to explain how is possible to overcome the problems of CNN. We found also other papers that we used as support to realize many different types of Convolution Neural Networks:

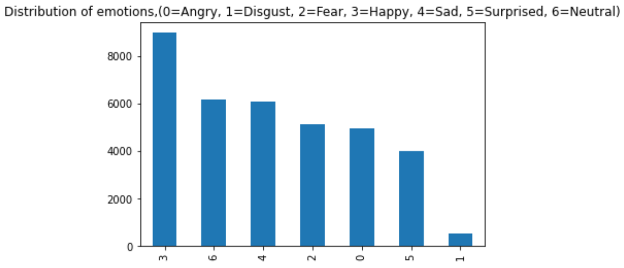
- "Very Deep Convolutional Networks for Large-Scale Image Recognition" by K. Simonyan and A. Zisserman [7];
- "Going Deeper with Convolutions" by C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich [8];
- "Deep Residual Learning for Image Recognition" by K. He, X. Zhang, S. Ren, and J. Sun [2].

Following we report a table that summarizes three of the works executing facial expressions recognition task on the FER2013 dataset, we considered from [4].

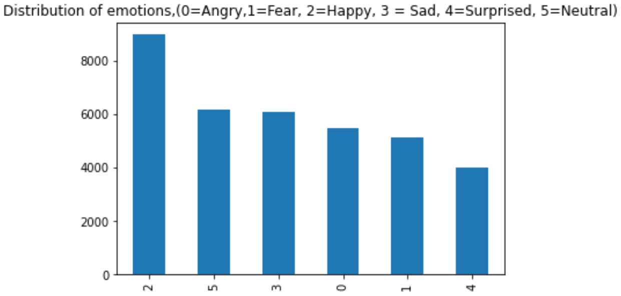
Papers	Network type	Accuracy reached
Zhang et al. [9]	CNN Multitask Network	Test: 75.10
Kim et al. [3]	CNN Network Ensemble	Test: 73.73
Pramerdorfer et al. [6]	CNN Network Ensemble	Test: 75.2

3. Dataset

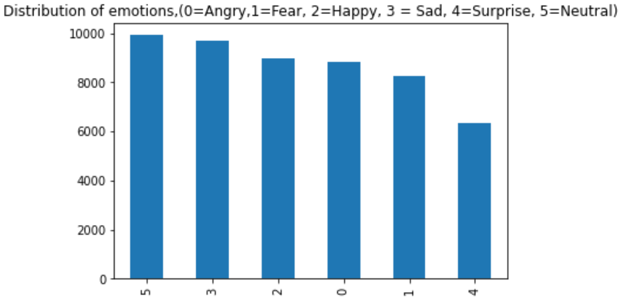
The FER2013 database was created by Pierre Luc Carrier and Aaron Courville. It was introduced during the ICML 2013 Challenges in Representation Learning, it is a large-scale and unconstrained database. This dataset was builted collecting images in an automatic way using the Google image search API. In order to find useful faces images a research has been carried out combining a set of emotion related keywords with other words associated to gender, age or ethnicity. In this way about 600 strings were obtained and they were used to query the search API. Then all the images collected with this system have been cropped and resized to 48*48 pixels and they have been also converted to grayscale. We choose this dataset for many reasons, for example we find it is cited in many papers and we consider it well formed dataset for the reason that it contains images which have different illumination, subjects with different age, pose and expression intensity, moreover some images have also occlusions. In general the images contained in the FER2013 dataset represent a good sampling under realistic conditions. The most important reason that pushed us to choose it is the huge number of images that it provides. Precisely it contains 28709 training images, 3589 validation images and 3589, so in total 35887 images for each of them there is an associated class that represent seven different expression labels: Angry, Disgust, Fear, Happy, Sad, Surprised, Neutral. Below we report an histogram showing the distribution of all the FER2013's labels.



We tried to apply as few preprocessing techniques as possible because we want to focus on the model performance providing to it a dataset not too optimized. We do this choice in order to emphasize the power of the model built, in the sense we want a model able to reach interesting performance independently the goodness of the dataset provided to it. For this reasons we operate only some minimal preprocessing modification over the dataset, in order to prepare the data to be learned. First of all we notice that the class disgust is not so big with respect to the others, so we decide to merge it with the angry class. Following a graph showing the merged classes result.



After this operation the classes are not so much leveled, in particular compared to happy class. In fact we have a lot of images for the happy class instead the ones for the surprised emotion are very few. To solve this issue we decide to replicate the 60% of the images of each class, excluding the happy class because as we said it already has enough images. In this way we make the images amount for each class more balanced and we reach the following distribution with a total amount of 52025 images.



Finally we performed a normalization operation over the dataset in order to provide to the model a numeric common scale. After all these needed preprocessing steps we obtain a training set formed by 44847 images and a test set and a validation set each one formed by 3589 images.
An evaluation of other dataset has been done, we considered also CK+ (CohnKanade) dataset that is also commonly used for evaluating FER systems. The first point that brought us to discard CK+ were the huge amount of data provided by FER2013 we decided to have more images as possible in order to obtain a model that could well generalize, moreover the aim of reaching an interesting accuracy with FER2013 was more challenging. We decided to prefer FER2013 also because it appeared to us more easier to work with, in that FER2013 was provided in a csv file. A more important reason that led us to prefer FER2013 is the fact this last provides specified training, validation and test sets, CK+ instead does not, so in order to compare our results with other works it was the most appropriate one.

4. Method

5. Experiments

6. Conclusion

We overtakes the accuracy reached from three papers which describe models that perform the same task with the same dataset used by us cited in [4]

References

[1] Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, Radu Ionescu, Marius Popescu, Cristian Grozea, James Bergstra, Jingjing Xie, Lukasz Romaszko, Bing Xu, Zhang Chuang, and Yoshua Bengio. Challenges in representation learning: A report on three machine learning contests. 2013.

[2] K. He, X. Zhang, haoqing Ren, , and J. Sun. Deep residual learning for image recognition. 2015.

[3] B.-K. Kim, S.-Y. Dong, J. Roh G. Kim, and S.-Y. Lee. Fusing aligned and non-aligned face information for automatic affect recognition in the wild: A deep learning approach. 2016.

[4] S. Li and W. Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 2020.

[5] M. V. B. Martinez. Advances, challenges, and opportunities in automatic facial expression recognition. *Advances in Face Detection and Facial Image Analysis*, 2016.

[6] C. Pramerdorfer and M. Kampel. Facial expression recognition using convolutional neural networks: State of the art. 2016.

[7] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. 2014.

[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, , and A. Rabinovich. Going deeper with convolutions. 2015.

[9] C. C. Loy Z. Zhang, P. Luo and X. Tang. Learning social relation traits from face images. 2015.