# Evaluation of Deep Convolutional Neural Network architectures for Emotion Recognition in the Wild

**Conference Paper** · June 2019

DOI: 10.1109/ISCE.2019.8900994

**4 authors:**

Abudukaiyoumu Talipu
Marche Polytechnic University
**7** PUBLICATIONS **75** CITATIONS
SEE PROFILE

Andrea Generosi
Pegaso University
**38** PUBLICATIONS **342** CITATIONS
SEE PROFILE

Maura Mengoni
Marche Polytechnic University
**176** PUBLICATIONS **1,543** CITATIONS
SEE PROFILE

Luca Giraldi
University of Turin
**42** PUBLICATIONS **286** CITATIONS
SEE PROFILE

# Evaluation of Deep Convolutional Neural Network architectures for Emotion Recognition in the Wild

A. Talipu , A. Generosi, M. Mengoni
Dipartimento di Ingegneria Industriale e Scienze Matematiche,
Università Politecnica delle Marche, Ancona, Italy
Email: {t.abudukaiyoumu, a.generosi}@pm.univpm.it, m.mengoni@univpm.it

L. Giraldi
Emoj s.r.l.
Ancona, Italy
l.giraldi@emojlab.com

***Abstract*— This paper presents a software based on an innovative Convolutional Neural Network model to recognize the six Ekman's universal emotions from the photos of human faces captured in the wild. The CNN was trained using three different datasets already labeled and merged after making them homogeneous. A comparison among different types of CNN architectures using the Keras framework for Python language is proposed and the evaluation results are presented.**

***Keywords—deep learning; emotion recognition; convolutional neural network***

## I. INTRODUCTION

Today, emotion recognition is one of the most challenging issues for numerous applicative sectors: from automotive where it is used for self-driving, to industry for supporting decision-making, till robotics for realizing empathetic interaction. Currently, the best way to recognize human emotions is processing the video and pictures captured by a video camera without introducing bias due to the use of intrusive technologies such as some wearable technologies (e.g. helmets, bracelets) or distributed sensors. To this aim, the most known methodology was devised by Paul Ekman, that with the FACS [1] method, tried to recognize the six universal emotions (joy, surprise, sadness, anger, fear and disgust) from face frames. Several researches exist to achieve this goal in an automated way, like in [2], [3], [4], and most use Convolutional Neural Networks.

Usually these researches refer to trained models using datasets built in controlled environments, where it is possible to obtain the best accuracy scores, or, trained with data obtained by crawlers on the web, with low accuracy but mostly reflecting real contexts. To the best of our knowledge, approaches that allow to obtain good accuracy using data obtained "in the wild", have not yet been evaluated.

In this context, a hybrid approach that try to obtain good accuracy to recognize human emotions in many contexts as possible, has been performed: the main purpose for the implemented software will be to recognize human emotions beyond the context in which the user is located, therefore in a retail activity using security cameras, as well as in front of a smartphone.

Applications could be different: User Experience analysis to improve the usability of web and mobile applications [5], entertainment for home automation [6] and Customer Experience analysis for retail contexts [7][8].

The aim of this research is to implement a software capable of recognizing Ekman's universal emotions by training a convolutional neural network based on the Keras and Tensowflow frameworks, merging three different public datasets and trying and testing the most known CNN architectures, like VGG13, VGG16 [9] and so on.

Conclusions discuss the results obtained from confusion matrices in order to evaluate the architectures with the best accuracies.

## II. THE PROPOSED SYSTEM

Discriminating emotion based on appearance is essentially an image classification problem. Therefore, a state-of-the-art Deep Convolutional Neural Network model that performs well in image classification should also perform well in facial expression recognition [3].

All the facial expression recognition models have reached different accuracies according to the datasets they have trained on. The [2] listed the different model accuracies that various models have reached. It has been observed that all the models with high accuracies are trained on the lab generated datasets such as MMI, CK+. However, the models trained with the datasets with in the wild properties (usually web crawled face images) have lower accuracies, that's because most of the datasets collected from the world wild web have inaccurate labels [3]. Moreover, the level of exposure of a human face on an image could also lead to inaccurate recognitions.

In the presented work, we conducted experiments with a dataset constructed by merging the lab generated CK+, the re-tagged FER and AffectNet. The assumption is that by combining the lab generated highly accurate dataset with the "in the wild datasets", it may result a better accuracy model for the in the wild benchmarks.

The implementation of the proposed system mainly uses Python scripts that import the Dlib library for face recognition and the Keras and Tensorflow frameworks for the Training and Deploy phases, so to train the model and using it to predict

Thank you to our sponsor Emoj for the contribution

emotions. To train Convolutional Neural Networks there exist different types of architectures to refer to, the above framework for Python, Keras, provides a frontend layer to the below Tensorflow framework, making available to the developers different interfaces that simplify the implementation of these architectures; among them we can find: VGG13, VGG16, VGG19, Inception. For this research has been implemented and tested a Python training script for each of the aforementioned architectures, so to see which one gives the best results for emotion recognition.

## III. Experimental Assessment

### A. Dataset

Datasets play a crucial role in supervised learning, the neural network models depend greatly on them. There are many public datasets for facial expression recognition, since most of them are prepared by web crawled face images with emotion related keywords, the label accuracy is not very high. The lab generated datasets like CK+ [10] on the other hand, has a high label accuracy but the dataset size is small. FER+ dataset is the re-tagged version of original FER dataset with crowd sourcing. It has a label accuracy over 90% but it contains only about 35k images [3]. AffectNet dataset has over 1 million web crawled face images, it also contains 450k categorically annotated images by expert human labelers. For our study we have examined all the images, and implemented a script to discard all the photos without faces or with multiple faces.

A new dataset is constructed by merging filtered AffectNet, CK+ and FER+ images tagged with one of the happy, surprise, sad, anger, disgust, fear and neutral tags. The dataset includes over 260k images, data distribution over different categories is showed on Fig. 1.

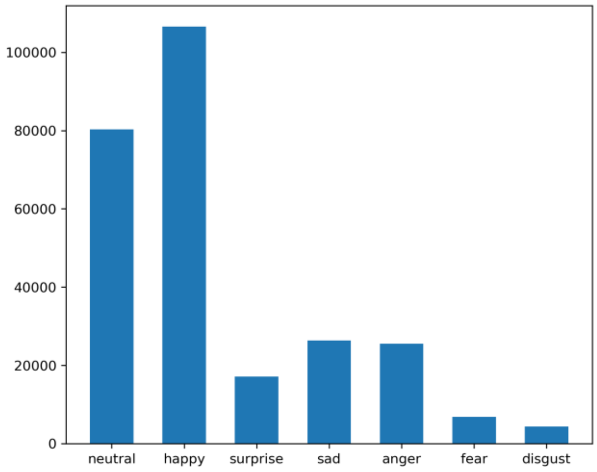

Fig. 1. Data distribution

Facial alignment techniques are used during the construction of the dataset to improve the accuracy of the dataset during construction. Given a set of facial landmarks, the facial aligner wraps and transforms the image to an output coordinate space. All faces across entire dataset are centered in the image. The images are rotated so that eyes lie on a horizontal line and the facial images are scaled to 64x64 pixels such that the size of the faces are approximately identical.

### B. Experimental results

Different model architectures are tested such as VGG13, VGG16, VGG19 [9] InceptionV2 and InceptionV3 [11] to compare the performances on emotion recognition task. The test accuracy of the models trained are listed on the TABLE I.

Network hyperparameters are initialized as it is stated on [3], then the other variations are also experimented such as validation split 0.1, 0.2, number of epochs 30, 50 and 100 and dynamic learning rate defined as

$$lr = lr \times (1 - \frac{\text{epoch}}{\max epoch})$$

Learning rate ($lr$) is initialized with 0.025, and updated on each epoch accordingly.

TABLE I. Performance of different models

| Architectures | Accuracy (%) |
|---|---|
| VGG13 | **75.48** |
| VGG16 | 74.48 |
| VGG19 | 73.14 |
| InceptionV2 | 75.26 |
| InceptionV3 | 67.20 |

The best performance gain is achieved by the VGG13 architecture. The training and validation accuracy of the best model VGG13 is plotted as a reference to the conducted experiments on Fig. 2.

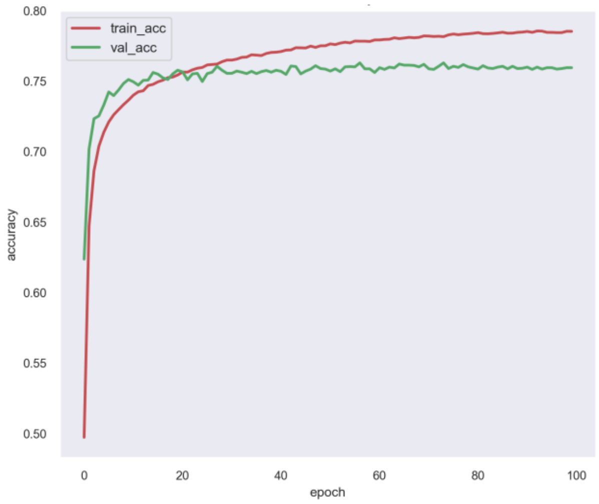

Fig. 2. Training and validation accuracy of VGG13

The test accuracy of each emotion category on the confusion matrix plotted as heatmap on Fig. 3, shows some interesting results. The images with fear tag misclassified as surprise and disgust tag misclassified as anger has over 20% rate.

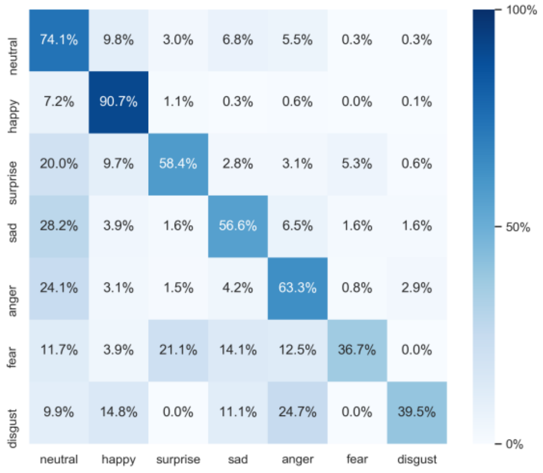

Fig. 3. Accuracy of each emotion category

### C. Evalution

The model VGG13 we trained is also evaluated on EmotioNet [4] 2018 challenge dataset: the Ohio State University, on their website [12], has in fact made available their dataset to give anyone the opportunity to compare their results with those of the challenges of 2017 and 2018. The table shows the evaluation result.

| Categories | Accuracy | F1 |
|---|---|---|
| happy | **0.9770** | **0.9799** |
| anger | 0.75 | 0.1198 |
| disgust | 0.0128 | 0.0099 |
| sad | 0.5955 | 0.2888 |
| surprise | 0.7059 | 0.3944 |

### D. Discussion

In this experiment, the performances of different Deep Convolutional Neural Network architectures are compared. As it can be seen from the confusion matrix, the emotion categories have more training samples resulted in higher accuracy compared to the ones with fewer training samples.

It is also believed that a more evenly distributed dataset may improve the model accuracy further.

Overfitting occurs during training using other deep neural network architectures, except for the VGG13, when epochs number is bigger than 30.

## IV. Conclusion and Future Work

Experimental results show that the accuracies of facial expression categories such as fear and disgust are low mainly because of the small dataset of the corresponding category. The other facial expression categories however have reached relatively high classification accuracies. The evaluation results on EmotioNet dataset also supports the results of our experiment. Future work will be focused on the collection of more data on the small dataset categories to improve the classification accuracies of those corresponding categories and on the exploration of generative adversarial networks to increase recognition accuracy.

## Acknowledgment

All the work provided by Research Fellow Giacomo Turri on design and development has been greatly appreciated.

## References

[1] P. Ekman and W. V. Friesen, 1978, “Manual for the Facial Action Coding System,” Consulting Psychologists Press.

[2] S. Li and W. Deng, “Deep Facial Expression Recognition: A Survey,” pp. 1–25, 2018.

[3] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, “Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution,” 2016.

[4] C. F. Benitez-Quiroz, R. Srinivasan, Q. Feng, Y. Wang, and A. M. Martinez, “EmotioNet Challenge: Recognition of facial expressions of emotion in the wild,” 2017.

[5] A. Generosi et al., “MoBeTrack: A Toolkit to Analyze User Experience of Mobile Apps in the Wild” *2019 IEEE International Conference on Consumer Electronics,* January 2019

[6] A. Altieri et al., “An Adaptive System to Manage Playlists and Lighting Scenarios Based on the User’s Emotions” *2019 IEEE International Conference on Consumer Electronics,* January 2019

[7] S. Ceccacci, A.Generosi, L.Giraldi and M.Mengoni, “Tool to Make Shopping Experience Responsive to Customer Emotions” *International Journal of Automation Technology,* vol. 12, pp. 319-326, May 2018

[8] A. Generosi, S.Ceccacci and M.Mengoni, “A deep learning-based system to track and analyze customer behavior in retail store” *2018 IEEE 8th International Conference on Consumer Electronics-Berlin,* September 2018

[9] K. Simonyan and A. Zisserman, “VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION,” in *American Journal of Health-System Pharmacy*, 2015.

[10] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, “The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression,” *2010 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. - Work. CVPRW 2010*, no. July, pp. 94–101, 2010.

[11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the Inception Architecture for Computer Vision,” 2015.

[12] EmotioNet Challenge, Retrieved April 15, 2019 from https://cbcsl.ece.ohio-state.edu/EmotionNetChallenge/index.html