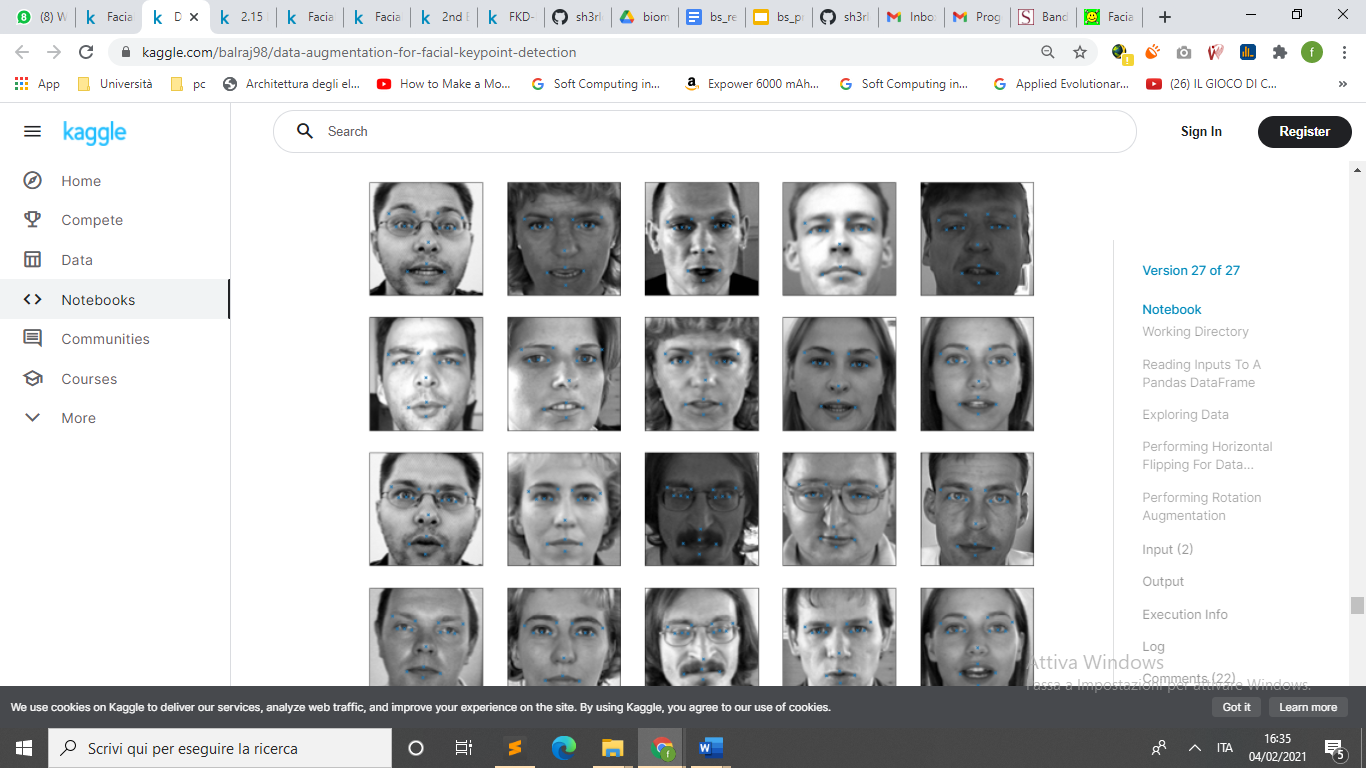
Facial Keypoints Detection

Detect the location of keypoints on face image

Mattia Capparella, 1746513  
Federico Fontana, 1744946

Biometric Systems course project,  
teached by Prof. Maria De Marsico  
AY 2020/21

# Introduction



Facial Key Points (FKPs) Detection

is an important and challenging problem in the fields of computer vision and machine learning. It involves predicting the co-ordinates of the FKPs, e.g. nose tip, center of eyes, etc, for a given face.

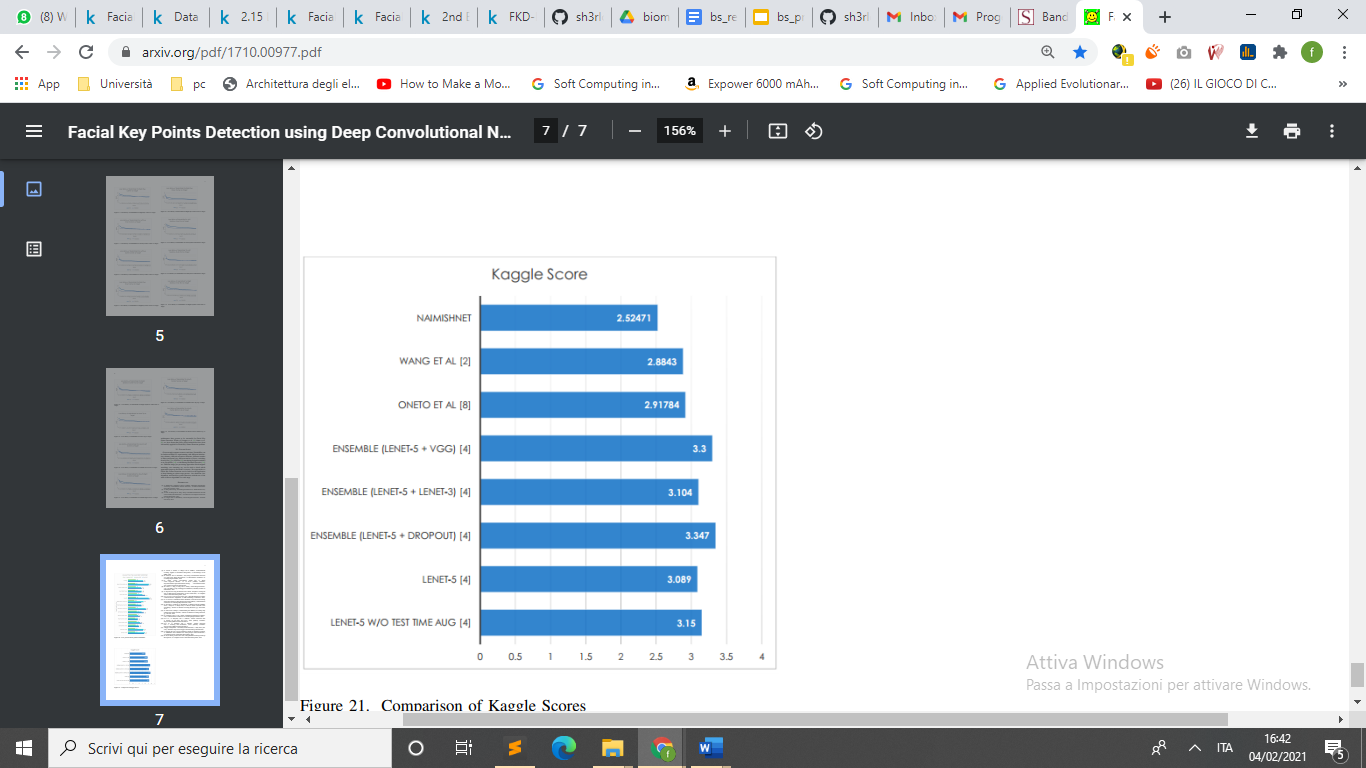
This project is about experimenting different models in FKPs Detection that is a critical element in face recognition.

However, there is difficulty to catch keypoints on the face due to complex influences from original images, and there is no guidance to suitable algorithms.

The **problem is** to predict the (x, y) real-valued co-ordinates in the space of image pixels of the FKPs for a given face image. It finds its application in tracking faces in images and videos, analysis of facial expressions, detection of dysmorphic facial signs for medical diagnosis, **face recognition**, etc.

Facial features **vary greatly from one individual to another**, and even for a single individual there is a large amount of variation due to pose, size, position, etc. The problem becomes even more challenging when the face images are taken under

different illumination conditions, viewing angles, etc.

With this report we want demonstrate that a good preprocessing can increase performance without modifying the model and without slowing down the algorithm too much.

## Datasets

The dataset is taken from the Keggle competition [Facial Keypoints Detection](https://www.kaggle.com/c/facial-keypoints-detection/data):

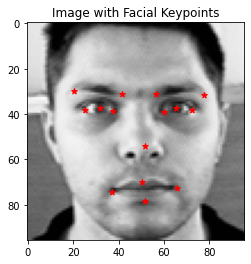
This dataset is formed by 4 files:

* **training.csv**: 7049 “*96x96”* training images. Each row contains the (x,y) coordinates for 15 keypoints, and the image data encoded as a string of pixels.
* **test.csv**: 1783 “*96x96”* test images. Each row contains the *ImageId* and the image encoded a string of pixels.
* **IdLookupTable.csv:** 27124 keypoints to predict. Each row contains a *RowId, ImageId, FeatureName* and finally the *Location,* our target.

In the dataset there are 15 different keypoints, each one specified by a (x,y) pair of coordinates.

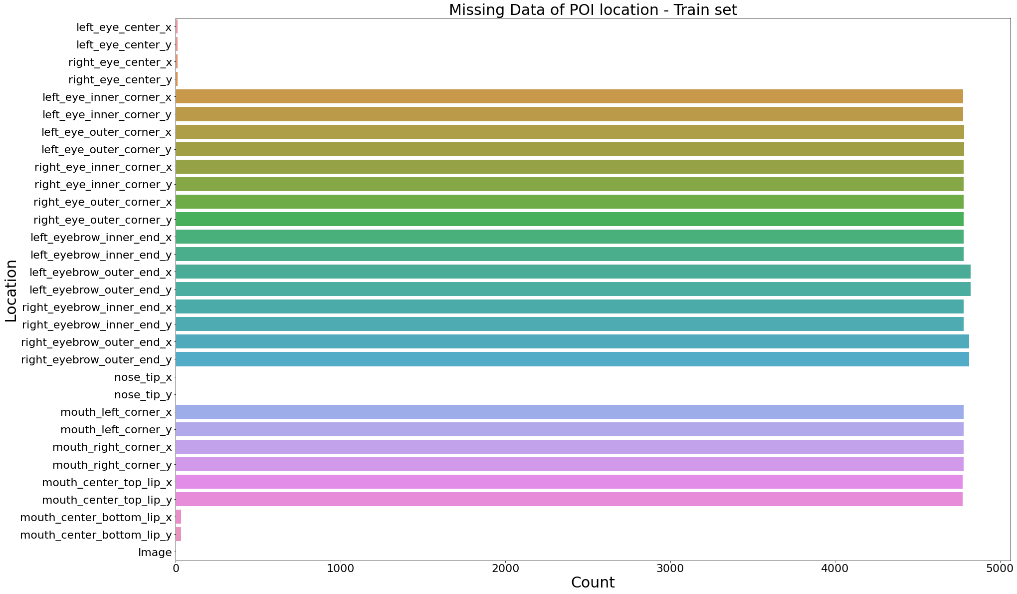
The 15 points represent the following elements:

1. *left\_eye\_center*
2. *right\_eye\_center*
3. *left\_eye\_inner\_corner*
4. *left\_eye\_outer\_corner*
5. *right\_eye\_inner\_corner*
6. *right\_eye\_outer\_corner*
7. *left\_eyebrow\_inner\_end*
8. *left\_eyebrow\_outer\_end*
9. *right\_eyebrow\_inner\_end*
10. *right\_eyebrow\_outer\_end*
11. *nose\_tip*
12. *mouth\_left\_corner*
13. *mouth\_right\_corner*
14. *mouth\_center\_top\_lip*
15. *mouth\_center\_bottom\_lip*



***note:*** *left*&*right* here refers to the point of view of the subject depicted.

It is important to point out that many *keypoints* in the training dataset are missing:



With the removal of images whose (some) keypoints are missing, we would remain with only 2140 images: a loss of possible valuable information of about 68%.

## Tools, libraries, and references

### Google Colab

To carry out the training of our models, we have used [Google Colab](https://colab.research.google.com/) since it offers a free and efficient platform that greatly reduces the *training time,* by providing high-capacity Nvidia GPUs usage.

We’ve made two notebooks: a [training one](https://colab.research.google.com/drive/1V4wo2cJpc9ANQ50VHoUyiMVtLwc_3uV8?usp=sharing) and a [demo one](https://colab.research.google.com/drive/1lPTYewjOPhMs33Tsmx1foCnmXPgT3uF3?usp=sharing) used to test each of the models explained further.

### Pytorch and Torchvision

To develop our model, a *Convolutional* *Neural Network* (CNN), we employed the [Pytorch framework](https://pytorch.org/) for its ease of use and the speedup provided by enabling the computations on GPUs.

Pytorch comes with the [Torchvision library](https://pytorch.org/docs/stable/torchvision/index.html), a collection that contains many functions for image manipulation, used extensively in our preprocessing phases.

## Model selection

To demonstrate valid the idea of applying a *preprocessing* *phase* we looked for a pre-made model that performed well on the very same task, with no preprocessing involved.

Since FKPs detection is a task performed in a vast number of *online scenarios,* it was fundamental that at testing time the application of such new methodology would have improved the results, while not affecting negatively the time performance.

For these reasons we used [LuisOlCo’s model](https://github.com/ruchawaghulde/Facial-Keypoints-Detection/blob/master/Code/FacialKeyPointsDetection_GoogleCollab.ipynb), as published on Github (more precisely: the generic CNN, not the LeNet5 one).

## LuisOlCo’s solution:

The CNN is so formed:

* 4 *convolutional layers*
* 3 *fully connected layers* (note that the last layer has 30 nodes because every FKS is described by a pair of coordinates)

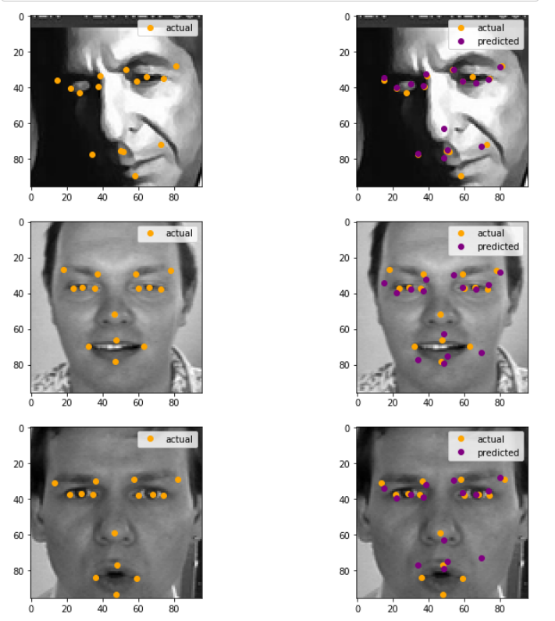
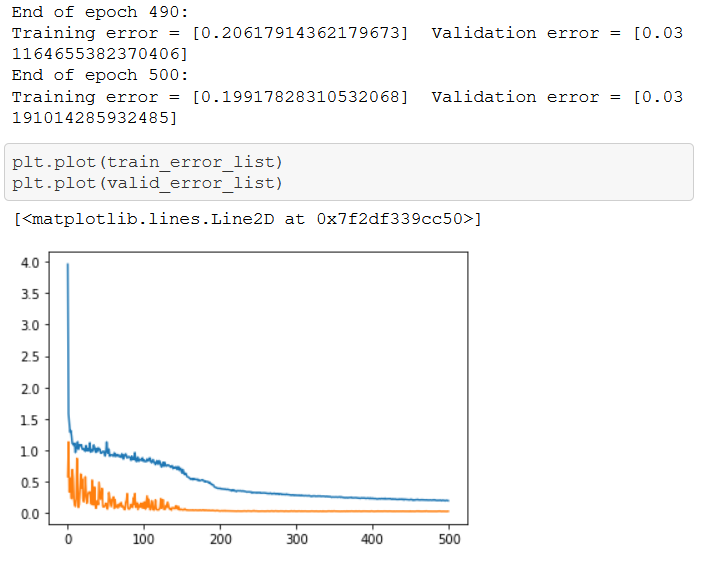
The *error function* being:

* mean squared error (MSE)

and the *optimizer:*

* *Adaptive moment estimation* (Adam)

As shown by the *train/val* loss plot and the *scatter plot,* the results obtained are quite accurate, but the *train loss* has a “*slow”* and the *validation loss* seems suffering from hysteresis and the best results are obtained after a great number of iterations.

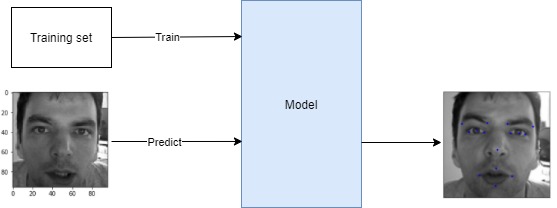


# Our contribution:

We focused on both scenarios of the *preprocessing phase*:

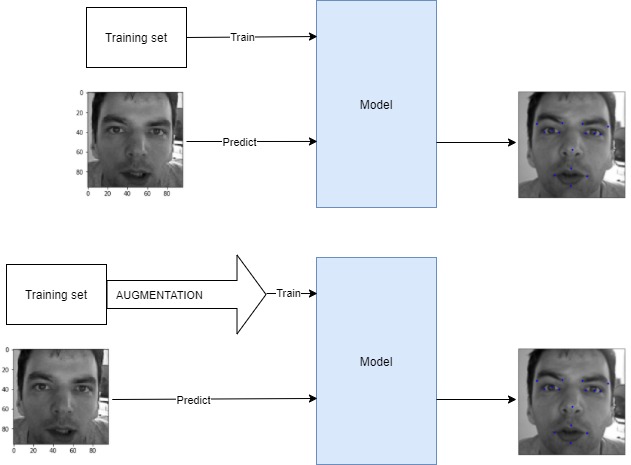
* *Training*: we can use augmentation to increases dataset size and introducing natural image distortion to make the model generalize better.
* *Real*-*time*: the speed is fundamental, so that we can apply our transformations to (a sequence of) images even *online*.

## Baseline Model



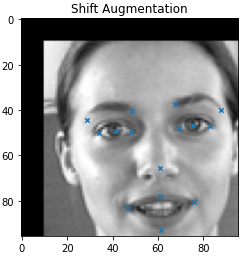
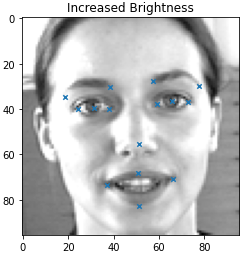
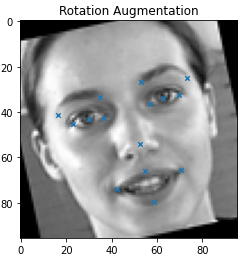
This model trains and makes its predictions without preprocessing. We will use this model to make comparison against ours.

## Augmented Dataset Model



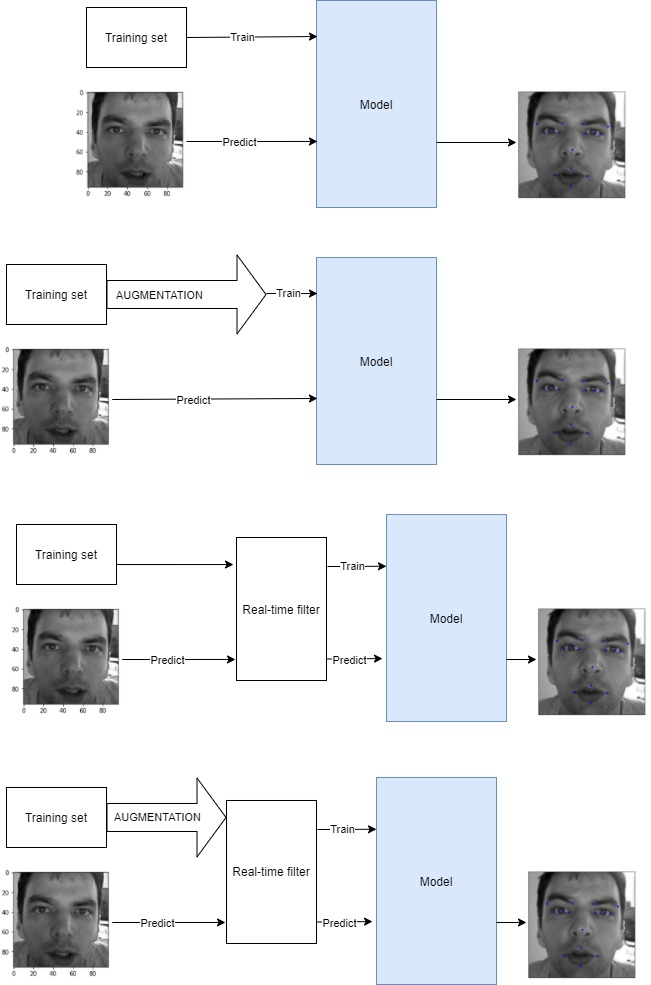
In here, we augment the whole dataset before doing any other step; the modifications we perform are the following:

* Rotations (counter/clockwise)
* Changes in brightness (increase/decrease)
* Shift
* Random noise



*note*: each *pad pixel* of the image is randomly colored with a certain probability independent from the other pixels of the image, set as in our experiments.

## Real-time Augmentation Model



This model does not use pre-augmented datasets but applies instead a filter just before the beginning of every *training epoch* using different probabilities to transform each image in the *batch dataset* “*in a unique way*” to train a model able to generalize better.

The same is applied at *test* time: the incoming (sequence of) images are transformed according to random probabilities.

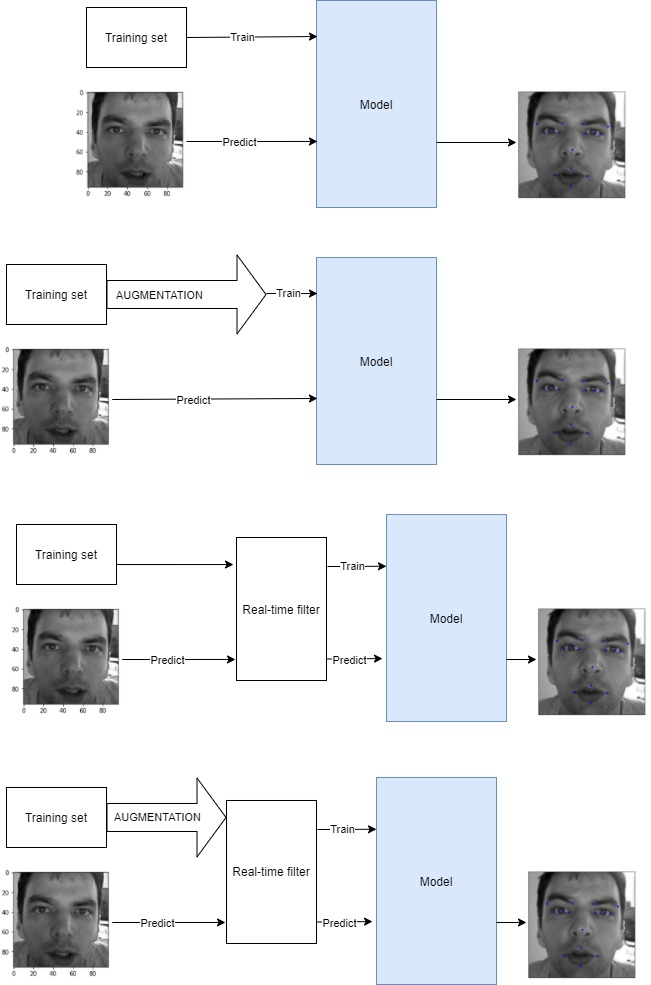
Since the *execution time* is a critical problem, let’s do some math:

we want use this on *real-time scenarios*, so the transformation time must be less than:

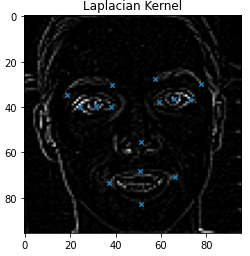
*i.e.* to transform an image as it arrives, we must perform it in less than , but considering that the model takes to predict one image we have:

to transform an image.

## Pre and real-time Augmentation Model



This model exploits both a real-time filter and the augmented dataset too.



(evaluation ot time execution of real-time filter)

(example of real time filter)

(Image of loss on epoch) (image of MSE)

# Performance evaluation

In this section we’ll discuss the performance of the model with different preprocessing above.

## (loss function image 1) (loss function image 3) (loss function image 3)

## (MSE image 1) (MSE image 2) (MSE image 3)

## (time of prediction image 1) (time of prediction image 2) (time of prediction image 3)

## Conclusions

Given these results, we can conclude that:

* **It is possible applying some real-time filter that increase the performance.**
* **The augmentation gets better generalization of the model.**

Below we can see a table that resume the results for all the models, with better scores in bold for each column:

|  |  |  |
| --- | --- | --- |
| **model** | **MSE** | **RMSE** |
| *baseline\_model* |  |  |
| *Augmentation\_model* |  |  |
| *Real\_model* |  |  |
| *Real\_aug\_model* |  |  |

# References

1. Connor Shorten et al., 2019,   
   A survey on Image Data Augmentation for Deep Learning,  
   <https://www.researchgate.net/publication/334279066_A_survey_on_Image_Data_Augmentation_for_Deep_Learning>
2. Naimish et al., 2017,   
   Facial Key Points Detection using Deep Convolutional Neural Network - NaimishNet,  
   <https://arxiv.org/pdf/1710.00977.pdf>
3. Rucha Waghulde, Luis Oliveros Colón and Siddharth Mandgi, 2019,

Facial-Keypoints-Detection,  
<https://github.com/ruchawaghulde/Facial-Keypoints-Detection>

# Resources

* Our Google Colab’s notebook used to train the models
* Our Google Colab’s notebook for demonstration
* Our GitHub repo, home of all the code used