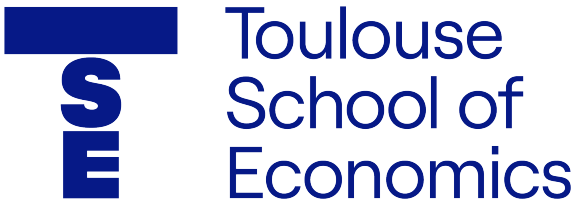
Deep Learning Project

Facial Landmark Point Detection

Johan Santiago RUIZ Khadija HAFSIA

Master 2 Data Science for Social Sciences

Professor : Raphael Sourty



Contents

1 Introduction

The aim of this project is to ameliorate our understanding of a very complex research on the field of facial landmark points detection using knowledge distillation-based neural networks. Tha name of the paper that was the focus of our literature review was Facial Landmark Points Detection Using Knowledge Distillation-Based Neural Networks written by Ali Pourramezan Fard and Mohammad H. Mahoor.

1.1 What is knowledge destilation ?

"It comes from the idea of training a lightweigth network with acceptable accuracy by transfering features and knowledge generated by an ensemble network into the single smaller network." (Fard and Mahoor,2021) In other words, transfer the knowledge of a much more sophisticated and complex network to a smaller one without losing too much efficiency. This lightweight model is called indeed a Student Network.

The authors of this paper formulate a very intersting analogy comparing the state of a butterfly and the concept of Knowledge distilation. At the beginning, the needs of this insect are very different and so is its shape that enables it to fulfill this needs at a very specific stage of its development. However, once it has passed through this stage, its needs are very different, and it does not require anymore to be inside a larva. So that arises the question as for why we should conserve a very cumbersome model once we have found the target probabilites. Now, the question is how to transfer the knowledge of a higly complex model to a simpler one. The solution proposed is to use "Soft Targets". For instance, in a classification problem, we can use the class probabilities yielded by the more complex model as targets for the much smaller model. A more sophisticated way is "to use the logits" in other words, the inputs of the final softmax.

1.2 The concept of coordinate regression

It is a very specific problem where the main objective is to estimate points of interest in an input image. One of the fields where it is most used is face recognition.. Among the main challenges of this problem ther is the so called Spatial Generalization, "which is the ability of a model to generalize the knowledge obtained at one location during the trainingg to another at inference time"(Nibali et al., 2018) . For exemple, a strong and robust algorithm should be able to recognize an eye no matter where it appears in the image, either on the bottom left, on

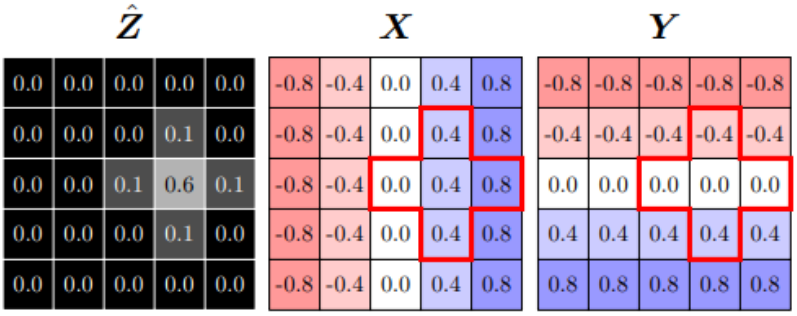


Figure 1: Figure 3 From the paper of Aiden Nibali Zhen He Numerical Coordinnate Regression with Convolutional Neural Networks

the center etc.

Image recognition can be carried out throught heatmaps. Those are a way of decompose an image into pixels and assign a numerical value to it. As in the example used in their paper (Nibali et al., 2018), we can assign a probability to each pixel around a given position to determine whether or not an object is present. The presence or not of the object can then be introduced as a probability distribution, where the pixel where the object is located have a higher probability, and the ones around a lower but still positive probability. As it is shown on the matrix $\hat{Z}$. In the example presented in the paper, they actually used two matrices $X_{i,j}$ and $Y_{i,j}$ where each value of the matrixes x, y is a coordinate such that the top left corner of the image is at (-1,-1) ant the top right corner at (1, 1). We can in this way define the probability of occurence of an object at the coordinate point in a 2-dimensional space c as a discrete bivariate random vector.

$$Pr(c = [X_{i,j}Y_{i,j})$$

The importance of this decomposition of the heatmap, is that it will enable us to compute a transformation from space to numerical. As proposed by (AIDEN) , we can use for instance.

$$DSNT(\hat{Z}) = \mu = [< \hat{Z}, X >_F, < \hat{Z}, Y >_F] =$$

The problem of image interpretation is very complex, and requires a series of mathematical transformations to be able to process the images and insert them into a convulotional network. The way this process is done will highly impact the performance of the network and therefore its final results.

1.3 A reformulation of the problem of face recognition

Interpret an image is a highly complex and daunting task for a machine. The amount of variation makes this task even harder. For instance, the contrast, light orientation of a picture increase the complexity of the problem. One way to deal with this issue is to use Soft-Landmarks. The idea is then to decompose the problem into a simpler one. In the concrete case of emotion interpretation, we can attempt to reconstruct with some points an a human face, and reduce then the amount of parameters that need to be considered in order to decide, for instance, what emotion is expressed in the picture. The idea of modeling has been around for a while (see (Edwards et al., 1998)).

Although some efficient deep-learning methods with good performances have been developed, they relied on networks with many parameters. Greediness in parameters leeds to time-consuming training and inference. However, on the other hand, the lightweight networks are difficult to set up and have a poor accuracy.
Hence, in this paper, the authors try to find a balance between a light network inefficient model and a greedy model with good performances.

1.3.1 The set up of the model

The new method works as follows :
First let's recall that the ultimate objective is to predict the Hard landmark points (meaning the landamarks from the ground truth *i.e.* the image). Their method is split in two phases. The first phase consists in simultaneously and independently train two teacher models (see the graph below) :

- a **tough teacher** that has as inputs hard landmark points, and it is trained through the L2 loss, $L2 = \sum_{i=1}^{n}(y_{true} - y_{pred})^2$.
- a **tolerant teacher** whose inputs are soft landmarks points that are created by *active shape models*, and is similarly optimized with the L2 loss.

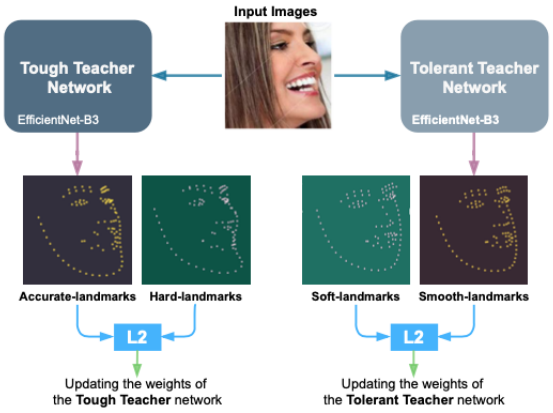


Figure 2: Figure from the paper of Pourramezan Fard and Mahoor : Facial Landmark Points Detection Using Knowledge Distillation-Based Neural Networks

Then in the second phase, in conjuction with the two pretrained teacher models, there is a student model, but this time the loss function is different. They define the KD-loss function that optimizes the student network to predict the hard landmarks points providing the two pretrained teacher models. And the two teachers models are now optimized through a assistive loss ALoss.

Indeed, since the variations of the tolerant teacher model are smaller than the ones of the tough teach, it is easier for the neural networks to learn on soft landmarks (the ones used as inputs in the tolerant teacher model) ; however the accuracy is poorer with the soft landmark points than with the hard landmark points, and for all these reasons the KD-loss is relevant, because it combines the advantages of both methods.

modÃÍle teacher and student.png

Figure 3: Figure from the paper of Pourramezan Fard and Mahoor : Facial Landmark Points Detection Using Knowledge Distillation-Based Neural Networks

1.3.2 How are the landmark points defined ?

The authors on the paper tackled the problem in a very clever way by stating that a given face, can be thought as a vector containing the coordinates of landmark points for each face $f(k+1)$, where k is the number of parameters. Then, they state that any face can be thought as a deviation from the mean, such that any face can be interpreted as.

$$f_{kx1} \approx \overline{f}_{kx1} + V_{kxm} b_{mx1}$$

Where $V = v_1, v_2, ..., v_m$ contains all the eigenvectors of the covariance matrix of all facial landmarks, ans b is given by $b_{mx1} = V^T_{mxk}[f_{kx1} - \overline{f}_{kx1}]$.

And then stating some restricions over b, a new face can be created:

$$f - new_{kx1} = \overline{f}_{kx1} + V_{kxm}\hat{b}_{mx1}$$

1.3.3 How is the Assistive Loss (ALoss) defined ?

The assistive loss guides the student network to predict the ground truth.
The idea behind the assistive loss is to use the prediction landmark points as either a negative or positive assistant. Positive meaning that the loss penalizes the network to predict a point close to the one predicted by the teacher model, and conversly a negative loss penalizes the network to predict a point far from the predicted one.
To do, let's first define :

- P_{Gt} : an arbitrary face landmark points from Hard landmarks set.
- P_{Ac} : The corresponding facelandmark points from the *Accurate-landmark* set (which is the set of landmarks predicted by the Tough teacher network).
- P_{Sm} : The corresponding face landmark points from the *Smooth-landmark* set (which is the set of landmarks predicted by the Tolerant Teacher teacher network).
- P_{Pr} : The corresponding face landmark points predicted by the Student network.

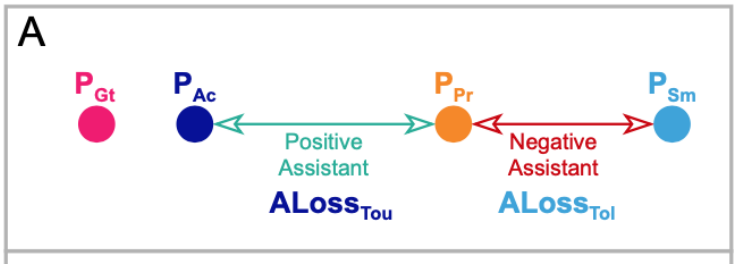


Figure 4: Exemple of the functioning of the ALoss

For example, as one can see in the **A** frame above, since the objective it to minimize the distance between P_{Pr} and P_{Gt}, so we use the the ALoss$_{Tou}$, of the Tough teacher as a Positive assitant and the ALoss$_{Tol}$ of the Tolerant teacher as a Negative assistant.
Furthermore, the authors go beyond, and define a threshold β_{Te}, as well as an assistant weight function w_{Te}.

1.4 What is the innovation of the paper

To propose a Knowledge Destilation Loss function that uses the predictions of a weight heavier model called MobileNetV2. The novelties of this paper are :
Firstly, the application of the KD-loss for a coordinate regression task and not a classification problem ; Secondly, while the original aim of the KD-loss was for the student network to mimic the results of the teacher prediction, the second paper provides two distinct assistive losses associated to two distinct teacher networks ; and accordingly, uses the ALosses and the prediction of the teacher networks to better predict the ground truth.

2 References

[1801.07372] Numerical Coordinate Regression with Convolutional Neural Networks (2022). URL: https://arxiv.org/abs/1801.07372 (visited on 01/08/2022).

Edwards, G. J., Cootes, T. F., and Taylor, C. J. (June 1998). "Face recognition using active appearance models". en. In: *Computer Vision — ECCV'98*. Springer, Berlin, Heidelberg, pp. 581–595. DOI: 10.1007/BFb0054766. URL: https://link.springer.com/chapter/10.1007/BFb0054766 (visited on 01/04/2022).

Fard, Ali Pourramezan and Mahoor, Mohammad H. (Nov. 2021). "Facial Landmark Points Detection Using Knowledge Distillation-Based Neural Networks". en. In: *arXiv:2111.07047 [cs]*. arXiv: 2111.07047. URL: http://arxiv.org/abs/2111.07047 (visited on 01/08/2022).

Nibali, Aiden et al. (May 2018). "Numerical Coordinate Regression with Convolutional Neural Networks". en. In: *arXiv:1801.07372 [cs]*. arXiv: 1801.07372. URL: http://arxiv.org/abs/1801.07372 (visited on 01/08/2022).