Università degli studi di
Milano-Bicocca

Advanced Machine Learning

Final Project

CAOS
Continuous Angle Orientation System

Authors:
Davide Brinati - 771458 - d.brinati@campus.unimib.it
Davide Meloni - 834333 - d.meloni5@campus.unimib.it
Alberto Raimondi - 771357 - a.raimondi21@campus.unimib.it

June 21, 2019


UNIVERSITA' DEGLI STUDI DI MILANO
BICOCCA

Abstract

The task of image orientation consists in the prediction of the most natural angle that a human would use to take a picture. In this paper we propose CAOS, a new system to allow a common convolutional neural network to deal with the cyclical nature of predicting the natural angle of an image. The solution is not straightforward due to the periodic behavior of angles. The model proposed is a supervised neural network that doesn't need feature points or landmarks as inputs but uses a standard convolutional neural network with a particular kind of loss function and output pair. The model reaches state of the art results in a specific domain without signs of overfitting.

1 Introduction

The current state of the art model for dealing with image orientation is the RotNet model (Gidaris et al., 2018) which uses a convolutional network to predict the discrete rotation applied to and image. The predicting of a continuous angle is still an open problem where the state of the art results still show a considerably large error in their predictions (Fischer et al., 2015).

This paper explores different methods to solve the problem of predicting the natural angle of human faces using a standard convolutional neural network trained using a novel of loss suitable for the task.

2 Datasets

The dataset used in this work is VGGFaces2 (Cao et al., 2018), 30000 images are used as a training set while 3000 are used as validation, the test set used is a sample of 300 images coming from the UTKface dataset (Zhang and Qi, 2017). The resulting images are given to a generator that randomly applies an horizontal flip and color jittering to reduce overfitting. The images fed to the model have a dimension of 64x64 pixels. Before being fed to a network each image is rotated by a random angle that is then given as a label for the supervised tasks. The random rotation makes it very hard for the model to overfit for the angle and limits the possible overfitting to specific characteristics of the training images.

3 The Methodological Approach

The task of this project is to predict the original angle of the photo using a standard convolutional network. The approach taken for this work is a supervised one (Goodfellow et al., 2016), this makes relevant the problem of the ground-truth being an unreliable source of information on the natural angle of the images. The images are not always taken with a straight angle and so the assumption that the original angle is a good target induces a significant source of error in the training process. By using the original angle of the image the model is biased due to the fact that a lot of images of faces are taken with a non-neutral angle. Another relevant challenge for this task is to adapt a neural network to deal with the cyclical nature of the rotations, the standard loss function used are not fitted to this task because they deal with probabilities or euclidean distances. The network needs a way to understand that a rotation of 360° is equal to no rotation.

The loss function used for training [1] was engineered as a way to deal with the periodicity of the angles. It represent the normalized length of the chord strung on a unit circle by the angle $(\gamma - \omega)$ where γ is the prediction of the neural network and ω is the target angle.

$$\text{Chord-loss} = \frac{1 - \cos(\gamma - \omega)}{2} \qquad [1]$$

The parity property of the cosine function allow us to produce a stable gradient for each output irrespective of the magnitude of the output of the network but only dependent on the internal angle between the prediction and the target value, in fact

$$\lim_{\gamma \to \omega} < \frac{1 - \cos(\gamma - \omega)}{2} > = < \frac{(\gamma - \omega)^2}{4} > \propto MSE(\gamma, \omega), \qquad [2]$$

where $< \cdot >$ is the mean operation on all γ and ω pairs in the set.

The network has two linear units as outputs (a, b), their ratio is fed to the arctangent2 function to convert it to an angle γ [3] that can be used by the loss function.

$$\gamma = \arctan 2(\frac{b}{a}) = \begin{cases} \arctan(\frac{b}{a}) & \text{if } x > 0 \\ \arctan(\frac{b}{a}) + \pi & \text{if } x < 0 \wedge y \geq 0 \\ \arctan(\frac{b}{a}) - \pi & \text{if } x < 0 \wedge y < 0 \\ +\frac{\pi}{2} & \text{if } x = 0 \wedge y > 0 \\ -\frac{\pi}{2} & \text{if } x = 0 \wedge y < 0 \\ \text{not defined} & \text{if } x = 0 \wedge y = 0 \end{cases} \quad [3]$$

The behaviour of the arctangent2 function is well represented by the graphic shown in Fig.1

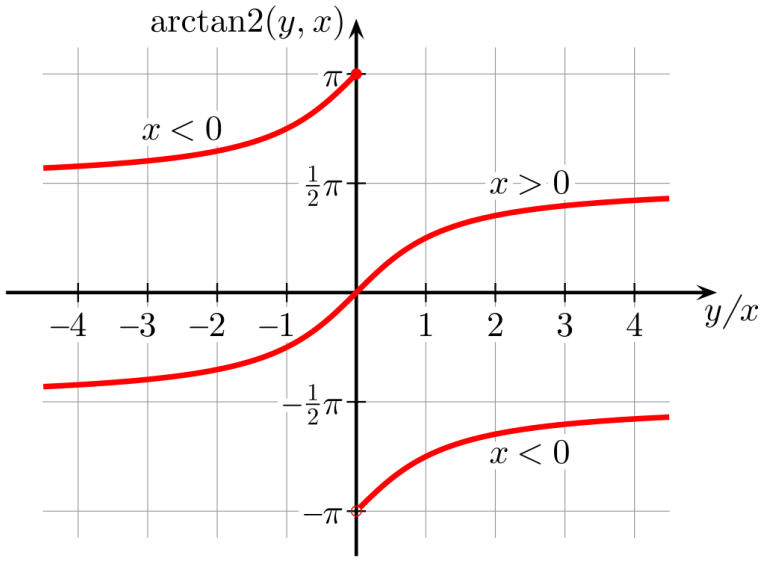

Figure 1: The arctan2 function as defined in Eq.3.

3.1 Training

The images have been resized to 64x64x3 and normalized; at every epoch a random continuous angle from -180 to 180 degrees has been picked out for each image and the picture is rotated accordingly. Some data augmentation techniques have been applied to reduce overfitting and make the model more prone to generalize on new data distributions and be robust for live

usage (Bishop, 2006). ColorJittering and RandomCrop together with the previously explained RandomRotation have done the job.

The images processed in this way are fed to a SqueezeNet model (Iandola et al., 2016), this type of neural networks are very effective even with 10x fewer parameters than the SOTA models. The network outputs two linear values as explained before, which are the arguments of the arctan2 function.

The network is trained through stochastic gradient descent using the Adam optimization method (Kingma and Ba, 2014), the momentum parameters are to $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate used is $\eta = 1e^{-3}$ and the training is done for 80 epochs.

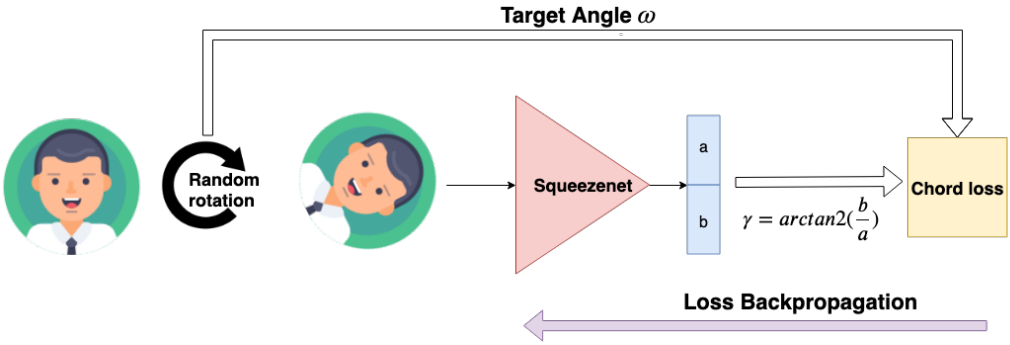

Figure 2: Neural net architecture.

4 Results and Evaluation

The graphic in Fig.3 shows the loss curves for training and validation sets during the training process. The curves start from a value very close to 0.5, indeed the baseline Mean Absolute error should be 90 degrees and the Chord-Loss for that value is

$$\text{Chord-Loss}(|\gamma - \omega| = 90°) = \frac{1 - \cos\left(\frac{\pi}{2}\right)}{2} = \frac{1}{2}. \quad [4]$$

The trends of the two curves appear very similar and the model does not show any sign of overfitting. On the other hand this was a task intrinsically hard to overfit, the images was never the same: at every epoch the generator rotates the original images of a continuous random angle from -180 to 180

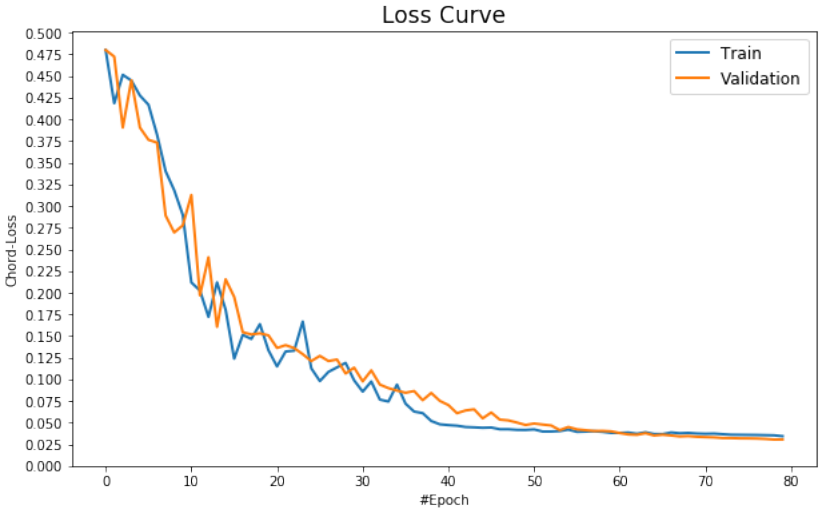

Figure 3: Loss curves of train set(blue) and validation set(orange) during the training process of 80 epochs.

degrees. Then, although the subjects of the images are always the same, the input of the model is different.

After the training process is time to evaluate the performances of the model. The Tab.1 shows the results in terms of error and loss values; it's possible to see that the model have a very similar Median Absolute Error on all the sets and even the Mean Absolute Error is the same, but it is about three times higher than the median error. This is due to the presence of a considerable number of outlier images in each of the three datasets.

	Train set	Validation set	Test set
Median Error	8°	7°	7°
MAE	21°	20°	21
Chord-Loss	0.034	0.030	0.032

Table 1: Results obtained with CAOS on the three datasets used. While the train and validations set come from the same source dataset, the test set has a completely different data distribution.

Another concern is to verify the actual homoscedasticity of error: the the absolute error should not depend from the rotation angle. It might be that images rotated with angles greater than 90 degrees produce errors

significantly bigger than the ones rotated with smaller angles. The graphic in Fig.4 shows how the Absolute error is distributed according to the angle of rotation.

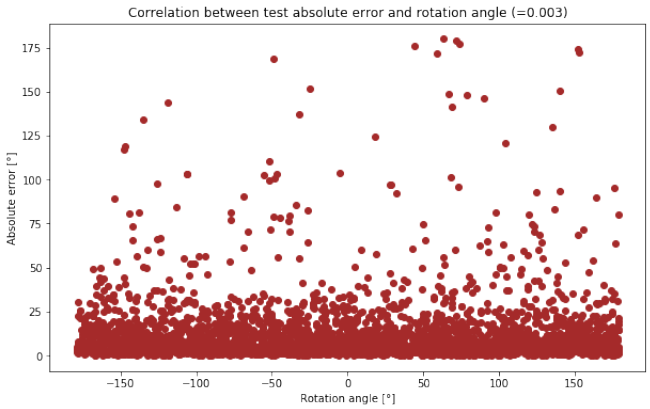

Figure 4: The graphic shows how the absolute error on the predicted angle depends from the angle each test image has been rotated with.

5 Discussion

The results obtained are encouraging and show that the model has surpasses the previous state of the art models on the task of face orientation. it's speculated however that the results will become worse and realign with the state of the art if the model is trained on the more general domain of all natural images. Specifically the proposed model improves on previous state of the art word by about 1 degree of accuracy on images taken from the same distribution that produced the training images, and has reached the same performance on a test set taken from a different dataset with a different data generating distribution. This results are interesting because of the ability of the model to generalize pretty well to all image faces and work on live application.

A relevant problem of the implemented method is that the labeling of the original dataset was implicit and not explicit, infact the images have been taken with common sense and this assumption does not hold true for the whole dataset. Specifically lots of images of babies are taken with an angle that is straight relative to their crawling body and not to their faces,

or lots of images with more than two subjects have only one of the faces with a straight angle. it's speculated that trusting the original images to have been taken with a natural angle induces a significant source of noise and error in the model and that a manually labeled dataset would probably significantly improve performance in this task. It's interesting to notice that the images with the highest error, like the ones in Fig.5, are confusing even for a human observer: the body pose often is not well defined or faces appear only in the images background. When dealing with standard image faces like the ones representing some of the most famous ML scientists, shown in Fig.6, the results are pretty more accurate, often indistinguishable from the ground-truth by the human eye.

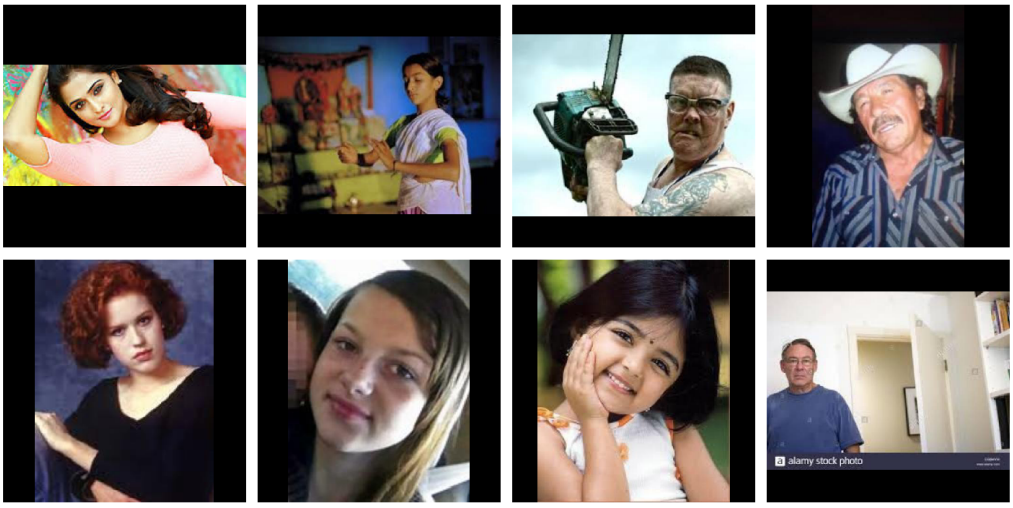

Figure 5: Some of the test images with the highest absolute error. There are images where the body pose is not so clear, in some pictures the face is inclined with respect to the body or it appears only in the background, in other images still the photo was taken just crooked.

The loss curve on the training and validation sets are very similar and indicate that the overfitting of the model is non-existent, this is probably due to the low number of weights used in the network and the randomization of the angle of the data generator. Thanks to the high number of images seen and the low number of available weight the network is forced to generalize the task and reduce overfitting to a minimum.

Figure 6: Above: randomly rotated inputs. Middle: outputs predicted by the neural network. Below: ground truth

6 Conclusions

This work considers the task of natural angle orientation and produces a working model that can be applied to various downstream tasks that need a vertically oriented face as an input.

The method applied in order to avoid the issues due to the periodical nature of angles has proved to be robust and has lead to interesting results.

The CAOS model is really effective but future work must be done to study the influence of dataset noise in the results and the generalization performances on a broader image domain.

7 Acknowledgements

We want to say thank you to everyone who gave us a feedback on this project, especially to Simone Bianco, who had the patience to listen to our difficulties always giving precious advice.

To run the whole CAOS project we used a Google Cloud instance with 24 GB of RAM and one Nvidia K80 GPU to train the model, not a penny was spent thanks to the first year free trial kindly offered by Google. All the code has been written in Python and the ML model has been developed leveraging the Pytorch framework.

References

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and Zisserman, A. (2018). Vggface2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*.

Fischer, P., Dosovitskiy, A., and Brox, T. (2015). Image orientation estimation with convolutional networks. 9358:368–378.

Gidaris, S., Singh, P., and Komodakis, N. (2018). Unsupervised representation learning by predicting image rotations. *CoRR*, abs/1803.07728.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Iandola, F. N., Moskewicz, M. W., Ashraf, K., Han, S., Dally, W. J., and Keutzer, K. (2016). Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <1mb model size. *CoRR*, abs/1602.07360.

Kingma, D. P. and Ba, J. (2014). Adam: A Method for Stochastic Optimization. *arXiv e-prints*, page arXiv:1412.6980.

Zhang, Zhifei, S. Y. and Qi, H. (2017). Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.