SAPIENZA
UNIVERSITÀ DI ROMA

FACOLTÀ DI
INGEGNERIA DELL'INFORMAZIONE, INFORMATICA E STATISTICA

DIPARTIMENTO
DI INGEGNERIA INFORMATICA, AUTOMATICA E GESTIONALE

Neural Networks course
Final project

A.Y. 2020/21

Report of C3AE Project

Professors:
Aurelio Uncini
Simone Scardapane

Students:
Francesco Petri 1797147
Giovanni Pecorelli 1799865

Contents

Chapter 1

Introduction

In the past decade soft biometrics has emerged out to be a new area of interest for the researchers due to its growing real-world applications. This includes a classic learning problem in computer vision: the estimation of demographic traits, such as the age. Researchers are trying to develop models which can accurately estimate the age or the age group of a person using different biometric traits. Currently, neural networks give the best classification results for age estimation using human faces. Many CNNs (convolutional neural networks) such as AlexNet, VggNet, GoogLeNet and ResNet are able to accomplish this task with promising performance.

However, to obtain more precise accuracy these networks have grown deeper and larger. This trend has resulted in increasingly higher computational costs in either training or deploying. In particular, deploying the previously mentioned models on mobile phones, cars and robots is next to impossible due to the model size and computational cost.
Recently other models have been proposed with the aim to reduce the number of parameters, thus yielding lightweight models without weakening their efficience.

In this work we want to investigate the limits of compact models for small-scale images and focus on one the most compact models for age classification, implementing it in practice to evaluate its performance.
The following report presents the development of the final project for the Neural Networks course at Università degli studi di Roma "La Sapienza", A.Y. 2020/21.

1.1 Related works

Our work is based on the study made by Chao Zhang, Shuaicheng Liu, Xun Xu and Ce Zhu in the paper *"C3AE: Exploring the Limits of Compact Model for Age Estimation"* [1] in which they propose a **C**ompact basic model, **C**ascaded training and multi-scale **C**ontext, aiming to tackle small-scale image **A**ge **E**stimation. The model is called **C3AE**.

The proposed model is able to achieve a state-of-the-art performance compared with alternative compact models and even outperforms many bulky models. With an extremely compact size of 0.25 MB for the full model, which is possibly the smallest that has been obtained so far on the facial recognition, C3AE is suitable to be deployed even on low-end mobiles and embedded platforms.

1.2 Report organization

This report is structured as follows:

- In Chapter 2 we describe the datasets used for the training and the evaluation of the model, as well as the preprocessing and the augmentation techniques used on them.
- In Chapter 3 we present the theory behind C3AE, explaining its model composition and our implementation.
- ...

Chapter 2

Datasets

Our C3AE implementation was trained and tested on the following datasets:

- **Wiki**: a large dataset containing 62,328 labelled images[1] collected from Wikipedia [2]. Despite the size, it lacks samples of very young or very old people, and it is quite noisy. The cropped and aligned version of the dataset was used in order to ensure each picture has a single face in it. This dataset was used as a pre-training set for ((one experiment)) and as the training set for the whole ablation study.
- **UTKFace**: a dataset containing over 20,000 labelled images[2]. It covers a wider range of ages compared to Wiki, therefore this dataset was used as the main training set for ((that same experiment)) and as a fine-tuning set after the pre-training with Wiki.
- **FG-NET**: a dataset containing 1000 labelled images[3] [3]. It is significantly higher-quality and better-curated than the other datasets listed here, but it is extremely small as well. This dataset was used as the test set for all experiments.

2.1 Preprocessing

Each dataset initially came in the form of a set of image files each with the corresponding ground-truth age encoded in the file name in some way. In order to make the datasets usable by our model, the following preprocessing procedure was applied to each of them.

First of all, the age information is extracted from the file name of each image through a dataset-dependent regular expression.
Second, the face in each image is detected with the MTCNN face-detection network[4] [4], and the resulting information is used as a base to position the three bounding boxes to be used later to generate the different crops that C3AE uses as multi-scale context (see chapter 3: The C3AE Model).
Then, faces with associated age outside the [0, 120] range are filtered out of the dataset. Images where less or more than one face is detected are also discarded.

Finally, images, age labels and bounding boxes are all organized into a `pandas` table which is saved to disk in the `pickle` format, in order to be loaded later by the other parts of the code.

[1] Collected and avaliable at https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/
[2] Collected and avaliable at https://susanqq.github.io/UTKFace/
[3] Collected and avaliable at https://yanweifu.github.io/FG_NET_data/
[4] Imported as external code from https://github.com/ipazc/mtcnn

2.2 Data generation

Training, validation and test data are generated in real time during the respective phase from the preprocessed dataset.
This requires a small amount of additional processing, which normally consists of applying a reflect-type padding to handle bounding boxes that are partially outside the border of the image, then generating three cropped images, one per bounding box, and finally resizing each of those crops to 64×64 pixels, the accepted input size of the C3AE model.

2.2.1 Training data augmentation

Training data undergo additional transformations during this generation process in order to augment the dataset and make the trained model more robust:

- Random erasing: before adding the reflect padding, an arbitrary portion of the image is deleted and replaced with random noise [5].
- Random shift: Before cropping, each bounding box is independently shifted by a random amount. If any part of a box would move past the border of the padded image, it stops at the border instead.
- Random contrast, brightness and color temperature change[5]
- Random rotation
- Vertical flipping

Each operation has a random probability to be applied and is independent of the others, and the intensity of each transformation (except flipping) is also random, but limited.

[5] The temperature change code was imported as external from https://www.askaswiss.com/2016/02/how-to-manipulate-color-temperature-opencv-python.html and adapted to randomize the intensity of the change.

Chapter 3

The C3AE Model

3.1 Model details

The C3AE plain model takes as input a source image and tries to output a plausible estimation for the age of the portrayed person.

Since the objective is to obtain a lightweight model the total number of parameters needs to be as low as possible, as long as it doesn't affect the overall performance. For this reason the input image size is limited (64×64×3) and the other channels sizes are also small.
Standard convolutional layers are adequate for the trade-off between accuracy and compactness (as opposed to the separable convolution block used in the bigger models like MobileNets and ShuffleNets), followed by batch normalization, Relu and average pooling (BRA).
The model is composed of five of these standard convolutions and two fully connected layers as shown in Table 3.1.

Layer	Kernel	Stride	Output size	Parameters
Image	-	1	64×64×3	-
Conv. 1	3×3×32	1	62×62×32	896
BRA	-	1	31×31×32	128
Conv. 2	3×3×32	1	29×29×32	9248
BRA	-	1	14×14×32	128
Conv. 3	3×3×32	1	12×12×32	9248
BRA	-	1	6×6×32	128
Conv. 4	3×3×32	1	4×4×32	9248
BN + ReLu	-	1	4×4×32	128
Conv5	1×1×32	1	4×4×32	1056
Feat.	1×1×32	1	12	6156
Predict.	1×1×1	1	1	13

Table 3.1: Architecture of the model

To estimate people's age C3AE considers two objectives simultaneously: the first one minimizes the Kullback-Leibler loss between distributions, and the second one optimizes the squared loss between discrete ages.

In the following sections are mentioned this and other techniques used in C3AE.

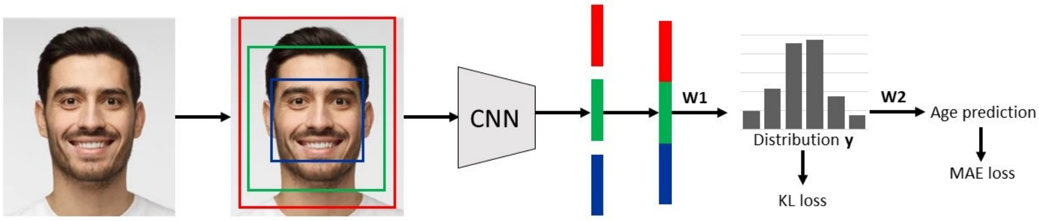

Figure 3.1: Overview of the model age estimation process

Context-based Regression

The resolution and the size of small-scale image is limited, so the idea is to exploit facial information at different granularity levels. After the face recognition task we identify three different crops of different sizes of the subject's face, like in Figure 3.1. Each cropped image has a special view on the face. The smaller image contains rich local information; in return the bigger one may contain global and scene information. The three crops are then fed into the shared CNN network, and finally the bottlenecks of the three-scale facial images are aggregated by concatenation.

Two-point Age Representation

With C3AE we don't predict directly the age as a number. Instead, this model uses a two-point age representation through a distribution over two discrete adjacent bins.
For example let's consider the corresponding representation of an age of 22 with 10 bins. In this case the set of bins is [10, 20, 30, 40, 50, 60, 70, 80, 90, 100] and the corresponding vector representation of the age is [0, 0.8, 0.2, 0, 0, 0, 0, 0, 0, 0].

Cascade Training

From the above section, the age value can be represented as a distribution vector. The mapping from this vector to age value is decomposed into two steps, where we define two different losses for the two cascade tasks.
The first one (L_{kl}) measures the discrepancy between ground-truth label and predicted age distribution. We adopt KL-Divergence as the measurement.
The second loss (L_{reg}) controls the prediction of the final age and is implemented as an L1 distance or mean absolute error (MAE loss).
In the training process the two loss functions are considered in the cascade style as shown in Figure 3.1 but they are still trained jointly, and the total loss is given as $L_{total} = \alpha L_{kl} + L_{reg}$, where α is the hyperparameter to balance the two losses.

3.2 Implementation

TODO

Chapter 4

Experiments

In this chapter we detail the experiments performed on the model described in chapter 3: The C3AE Model and discuss the respective results.

4.1 First experiment

An initial experiment was performed by training the C3AE model for 400 epochs on the Wiki dataset as a means to test the `tensorflow` environment and the functioning of the implemented C3AE model within it.

The evolution of training and validation loss is shown in Figure 4.1. The experiment ran to completition with no runtime errors, and it can be observed that the model was able to decrease its loss, and that such loss reached an asymptote around epoch 50.

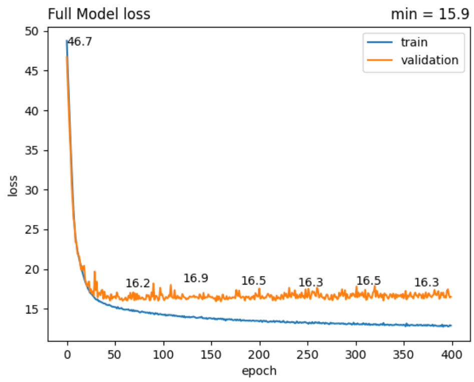

Figure 4.1: Initial experiment loss (400 epochs on Wiki)

For this reason, we concluded that a training time of 400 epochs is too much for this model and decided to set the epoch limit to 100 for all following experiments, assuming that they would have a similar evolution to this one and therefore all significant improvement would happen much before the 100$^{\text{th}}$ epoch, in order to reduce the experiments' computation time.

4.2 Performance evaluation over multiple datasets

????????????????????????

(One on Wiki, one on UTK, one on Wiki+UTK, all tested w/ FGNET)

????????????????????????

4.3 Ablation Study

A separate set of experiments was performed to study the impact on performance of the following components of the model and the training process: the context module and the cascade module of the C3AE model, and the training data augmentation.

We trained the following variants of the full C3AE model:

- *Full model*: the standard model with no changes, to serve as a benchmark against the other variants.
- *No augmentation*: the data augmentation transformations on the training data are disabled.
- *No context*: the context module of C3AE is excluded. Therefore, only one crop, the outermost one, is given as input to the model, and obviously there is no concatenation phase.
- *No cascade*: the cascade module of C3AE is excluded. Consequently, the intermediate layer between the concatenated feature vector and the age output loses its meaning of two-point representation of age and becomes a plain hidden layer. This also means that this variant does not compute any KL Divergence and outputs only the final age estimation; the loss depends only on the age MAE as well.
- *No context and no cascade*: both modules are disabled.

Each variant was trained for 100 epochs on the Wiki dataset. ((Use of test set needs some debugging))

((plot and discuss results))

Chapter 5

Project structure

Probabilmente non ci serve, ma non voglio scordarmi che esiste

Chapter 6

Conclusions

Model	Dataset	Augmentation	Epochs	Validation MAE	Test MAE
Full	Wiki	$\times$	100	8.36	24.41
Full	Wiki	Some	400	6.82	22.72
Full	Wiki	✓	100	6.79	18.11
Full	UTK	✓	100	8.67	9.79
Full	Wiki + UTK	✓	100+100	8.23	8.64
Full Ablation	Wiki	✓	100	12.60	-
No Cascade	Wiki	✓	100	12.60	-
No Context	Wiki	✓	100	6.88	-

Bibliography

[1] Chao Zhang, Shuaicheng Liu, Xun Xu, Ce Zhu. *C3AE: Exploring the Limits of Compact Model for Age Estimation.* CoRR, 2019.
URL: `https://arxiv.org/pdf/1904.05059v2.pdf`

[2] Rasmus Rothe and Radu Timofte and Luc Van Gool. *Deep expectation of real and apparent age from a single image without facial landmarks.* IJCV 126(2-4):144–157, 2016.

[3] Yun Fu, Guodong Guo, and Thomas S Huang. *Age synthesis and estimation via faces: A survey.* IEEE Trans. on Pattern Analysis and Machine Intelligence 32(11):1955–1976, 2010.

[4] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, Yu Qiao. *Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks.* IEEE Signal Processing Letters 23(10):1499–1503, 2016.

[5] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. *Random erasing data augmentation.* arXiv preprint arXiv:1708.04896, 2017.