Neural Computing and Applications
https://doi.org/10.1007/s00521-021-05981-0

S.I. : NCACVIP

Effective training of convolutional neural networks for age estimation based on knowledge distillation

Antonio Greco[1] · Alessia Saggese[1] · Mario Vento[1] · Vincenzo Vigilante[1]

Received: 20 November 2020 / Accepted: 25 March 2021

Abstract
Age estimation from face images can be profitably employed in several applications, ranging from digital signage to social robotics, from business intelligence to access control. Only in recent years, the advent of deep learning allowed for the design of extremely accurate methods based on convolutional neural networks (CNNs) that achieve a remarkable performance in various face analysis tasks. However, these networks are not always applicable in real scenarios, due to both time and resource constraints that the most accurate approaches often do not meet. Moreover, in case of age estimation, there is the lack of a large and reliably annotated dataset for training deep neural networks. Within this context, we propose in this paper an effective training procedure of CNNs for age estimation based on knowledge distillation, able to allow smaller and simpler "student" models to be trained to match the predictions of a larger "teacher" model. We experimentally show that such student models are able to almost reach the performance of the teacher, obtaining high accuracy over the LFW+, LAP 2016 and Adience datasets, but being up to 15 times faster. Furthermore, we evaluate the performance of the student models in the presence of image corruptions, and we demonstrate that some of them are even more resilient to these corruptions than the teacher model.

Keywords Age estimation · Knowledge distillation · Facial age dataset · Facial age benchmark

1 Introduction

Age is nowadays a relevant demographic attribute in several applications: digital signage, for example, leverages the estimated age of people to offer appropriate advertising for maximizing the effect of the advertisement campaign and then to increase conversion rates [17]; social robotics applications take advantage of multiple soft biometrics, including age, in order to enhance empathy [51]. In the above-mentioned applications, it is required a reliable age estimation even if the face image is acquired in unconstrained conditions, with strong variations of lighting and pose, and with non-collaborative people, namely with individuals that are not aware of the presence of the camera and then do not stop and look towards the camera. The above constraints make the age estimation task, already complicated for humans, even harder for automatic computer vision algorithms. Considering that such analysis must be performed in "real time", the time constraint may be a few hundreds milliseconds [25] for interactive applications such as digital signage and social robotics; consequently, there is a strong need for age estimation methods that are both accurate and fast.

The first efforts to tackle the task of age estimation relied on handcrafted features [15]; however, these techniques were able to achieve reasonable accuracy only in controlled conditions (e.g., frontal pose, high quality, high resolution), while their accuracy dropped when exposed to the variations in lighting and pose happening in real environments [6]. The advent of deep learning greatly allowed for age estimation methods that are significantly

✉ Alessia Saggese
asaggese@unisa.it

Antonio Greco
agreco@unisa.it

Mario Vento
mvento@unisa.it

Vincenzo Vigilante
vvigilante@unisa.it

[1] University of Salerno, Via Giovanni Paolo II 132, Fisciano, SA, Italy

Published online: 20 April 2021

more reliable in simple scenarios, and vastly superior in challenging conditions, making possible the design of algorithms sufficiently accurate for real applications [6, 42, 45].

Although very effective, the methodologies based on convolutional neural networks are often slow and resource demanding. Efficient network architectures targeted at biometric analysis exist [18], but they often require a sacrifice in accuracy, while the most accurate methodologies can be extremely bulky and slow [2], namely unusable for practical applications despite their reliability.

A second problem that hinders researchers in the field of age estimation is the absence of a large, reliably annotated dataset. This problem is due to the cost and difficulty of annotating a wide dataset. The one proposed during the ChaLearn Looking at People challenge 2016 [14], also known as LAP 2016 or APPA-REAL, is very reliable, since each face has been annotated with apparent age by multiple people; the process is accurate but costly and, in fact, the dataset includes only 7,591 images; for this reason, it is typically adopted for fine-tuning after a pre-training on a wider dataset [6]. Similar considerations apply to other well-known datasets in the literature (see Table 1), namely the Adience Dataset (26,580 samples) [12], the LFW+ dataset (15,699 samples) [47] and others, as shown in [6]; their size makes them perfectly suitable as benchmarks, while it poses a challenge when trying to train large convolutional neural networks.

The largest dataset for age estimation available in the literature is IMDB-Wiki [48], whose authors adopted a different approach for age labeling. They tapped into the public image databases of Wikipedia and the Internet Movie Database; from the page of famous people, they took and automatically annotated more than 500, 000 images, obtaining the age from the birth date of the person and the date of the picture; of course, this procedure does not ensure the reliability of the annotations, so much that the authors themselves recommend using the dataset with caution as there are several errors. Similar considerations apply for Cross-Age Celebrity Dataset (CACD) [7], which include around 163, 000 images annotated with the same protocol adopted for IMDB-Wiki.

Table 1 Absolute distribution of the samples in VMAGE, IMDB-Wiki, LFW+, LAP 2016 and Adience within the age groups 0–15, 16–25, 26–35, 36–45, 46-60 and 61-100

Age	# of samples				
	VMAGE	IMDB-Wiki	LFW+	LAP	Adience
0–15	18,864	7813	52	887	6983
16–25	451,999	50,216	372	1829	1655
26–35	1,342,493	103,240	1855	2376	4950
36–45	702,677	86,688	1822	2350	2350
46–60	589,558	58,087	3661	943	830
61–100	131,401	21,566	2385	387	875

In the absence of alternatives, IMDB-Wiki is anyway the current standard for pre-training convolutional neural networks for age estimation. However, to obtain state-of-the-art performance, it is necessary to carefully "clean" the dataset in order to consider only the correctly labeled images. The authors of [2], winner of the LAP 2016 competition, applied a combination of automatic and manual filtering strategies to discard almost half of the images and to obtain a cleaner version they call *IMDB-Wiki-cleaned*, unfortunately not publicly available and, thus, not easily reproducible. Therefore, a significant effort is required to design and implement an effective training procedure of convolutional neural networks for age estimation.

In this paper, we propose the use of knowledge distillation to overcome these limitations [54]. Knowledge distillation [24] is a technique used to train small, efficient convolutional neural networks with reduced need of resources (i.e., processing time, memory, and so on) transferring the knowledge learned by a more complex model. The method, in its general form, consists in the extraction of the class probability vectors produced by a large model, also called *teacher*, and the adoption of these vectors as a target for training the smaller model, known as *student*. An alternative, naive approach, would be to train the small network directly on the same dataset that was used to train the large model; however, it has been demonstrated that for complex problems the student network can achieve higher accuracy when trained with knowledge distillation than if it is directly trained with the labels of the original dataset [3]. The intuition behind distillation, i.e., the supposed advantage, is that the large *teacher* model is able to better fit the dataset and encode its peculiarity due to its higher representative power, in a way that the smaller model just could not; the *student* model may be, however, able to leverage the knowledge that has been pre-digested and encoded into a simpler annotation, namely the output probability vectors of the teacher.

Recent literature demonstrated the effectiveness of knowledge distillation in various pattern recognition tasks, even related to face analysis. In [24] Hinton et al. showed that knowledge distillation allows a 2% accuracy improvement of a student model for speech recognition with respect to one trained using the original labels of the dataset; with this technique the simple student model performs similarly to the much more elaborate teacher model. In [43] the authors demonstrated that a network trained with the distillation approach makes a CNN more robust to

perturbations by a considerable amount; this effect is explained considering that the *student* network sees its training input in a clearer way and is able to organize its weights around a more representative manifold. In [16] Low et al. applied distillation on selected, most informative faces, to train a face recognition network that achieves good performance on images with low resolution. In [10] the authors trained different convolutional neural networks (CNNs) for facial expression recognition with incomplete labeling; they find that the *student* model often outperforms the *teacher* on the considered task.

The method proposed in this paper is a variant of the standard knowledge distillation technique. We apply it to the problem of age estimation, to address its peculiar limitations, namely the absence of a large dataset with reliable annotation and the lack of a handy procedure allowing to train effective and efficient CNNs for age estimation applicable in real scenarios. We take the popular large-scale dataset VGGFace2 [5], not annotated with age labels, and we run the most accurate method in literature, winner of the LAP 2016 competition; this method consists of a large and complex ensemble of 14 CNNs that analyze 8 versions of each input image [2], and is trained on *IMDB-Wiki-cleaned*. We use the resulting predictions as target labels to train a variety of different CNN architectures, requiring about 15 times less operations. Therefore, we obtain the twofold advantage of having a large dataset annotated for age estimation enabling the possibility to perform a standard (and fast) training procedure of smaller student models.

We show that our approach allows to achieve state-of-the-art results on the LAP dataset with much simpler methods only composed of a single CNN, outperforming other complex methods, except the one used as teacher. We show that using our own cleaned version of IMDB-WIKI as training dataset, the accuracies reached by the same CNNs are much lower, thus proving the effectiveness of the proposed approach with respect to the traditional procedure. We also show that the student models are even able to outperform the teacher in the presence of strong image corruptions; to this aim, we evaluate the networks on the LFW+C dataset [19], which includes blur, noise, occlusions, variations in brightness and contrast and jpeg compression.

To summarize, the contribution of this paper is threefold:

1. We recognize the limitations of the state of the art, i.e., the absence of a large and reliably annotated dataset for age recognition and the absence of a handy procedure for training effective and efficient methods for age estimation, and we propose the application of a tailored knowledge distillation approach to overcome those limitations.
2. We benchmark the most popular and modern CNN architectures proposed for image recognition and fine-tuned on age estimation, reporting the results achieved on the most popular datasets with our approach. We make publicly available with the name of *VMAGE dataset* the labels used for training, along with the code to allow other researchers to replicate our findings and to train new models with our approach.
3. We evaluate the effect of different categories of image corruptions on the accuracy of the student models trained with the proposed approach and demonstrate that some of them are more resilient to image corruptions than the teacher model. To the best of our knowledge, this is the first attempt of evaluating the impact of image corruptions on the performance of CNNs for age estimation; to make the experiment reproducible, we also distribute the code for performing this benchmark.

1.1 State of the art

Before the deep learning revolution, the methods for age estimation were based on handcrafted features extracted from face images. Biologically inspired features (BIF) [21] and histogram of oriented gradients (HOG) [53] were the most common choices, but the best accuracy was obtained through the use of texture-based features, namely local binary pattern (LBP) histograms [39] or their variants as directional age primitive pattern (DAPP) [32]. The methods based on handcrafted features have been gradually abandoned in recent years, since even on datasets acquired in controlled laboratory conditions as MORPH-II, FACES and LIFESPAN, the average mean absolute error (MAE) is 2–4 years higher than the corresponding value obtained by more recent methods based on convolutional neural networks; the gap on more challenging datasets as Adience becomes even higher, namely 7.5% less in terms of age group classification accuracy.

The age estimation methods based on deep learning are extensively described in a recent survey on the topic [6]. We notice that all the approaches achieving the best ranking within the ChaLearn LAP competitions are indeed deep learning based [13, 14], and so are most of the new methods obtaining the best results that were proposed after the competitions. Since the datasets available for training were quite small (4691 images in 2015, 7591 in 2016, but only 50% of the samples for training), the biggest challenge was the preparation of a dataset sufficiently large and representative for age estimation pre-training. Rothe et al. [49, 50] won the competition in 2015, proposing for the

first time the IMDB-Wiki dataset and adopting a cleaned version to pre-train a VGG-16 network for age estimation. Then, the authors applied a bagging procedure to fine-tune 20 versions of the pre-trained one on different splits of the LAP 2015 training set, augmented 10 times with random rotations and translations. The final age prediction is the average of the outputs of 20 CNNs. An even more complex method based on the use of 10 structured output SVMs reached the third place in the competition of 2016 [58]. The method which achieved the second top rank in 2015 [38] is an ensemble of 4 classifiers and 4 regressors based on GoogLeNet, pre-trained for face recognition by using the CASIA-WebFace dataset and fine-tuned for real age estimation over MORPH-II, CACD and WebFace. Then, starting from the learned weights, 4 classifiers and 4 regressors are trained with different versions of the training samples of LAP 2015. The approach holding the best performance over this dataset is the one proposed by Tan et al. [56], which also claims the second top result over the dataset used in the competition of 2016. The method is based on a version of VGG-16, modified with an output layer including K+7 neurons, where K is the number of age labels; each neuron acts as a binary classifier which decides whether the sample belongs to an age group of 7 years. The predictions of the classifiers are then analyzed by an age decoding algorithm which provides the final estimation. The network is pre-trained on a cleaned version of IMDB-Wiki and fine-tuned over a 36 times augmented (random flip, rotation and noise addition) training set of LAP 2016. Dehghan et al. [9] achieve a similar performance with a private CNN pre-trained for face recognition with a private dataset composed of 4 million of images and 40,000 identities and for age estimation with around 600,000 samples manually labeled by human annotators. Another ensemble, composed of 4 VGG-Face models trained on different folds of IMDB-Wiki and LAP 2015 datasets, achieved the second top rank in 2016. The winner of the competition of 2016 is the approach that we chose as teacher network [2], which we will describe in detail in the following. It won the challenge with a substantial gap (0.069 ϵ-error) and holds also the best performance over MORPH-II, thanks to an ensemble of 14 CNNs (3 dedicated to individuals under 12 years old) but especially to the contribution of 26 people which carefully cleaned the samples and the age annotations of IMDB-Wiki and extended the training set with a private dataset of children.

The analysis of the state of the art about age estimation points out that the approaches achieving the best performance are complex ensembles of classifiers trained with large private datasets, prepared and annotated with costly procedures, or by adopting strong data augmentation techniques. This evidence justifies the contribution of this paper, which aims at providing the scientific community with a standard, fast and publicly available training protocol for effective and not necessarily complex convolutional neural networks for age estimation.

2 Proposed knowledge distillation approach

Our aim is to train simple and efficient CNNs that are able to perform accurate age estimation. As mentioned in Sect. 1, reliable datasets for this task are not big enough to effectively train a deep neural network, while large datasets such as *IMDB-Wiki* (contain spurious annotation that will cause inefficiencies in the training process [48] leading to suboptimal results).

In our approach, we train simple CNN architectures using knowledge distillation. The procedure consists of two steps: as a first step, we automatically annotate VGG-Face2 Dataset by means of a teacher method, an ensemble of CNNs trained for age estimation; we call this dataset *VGG-Face2 Mivia Age* (VMAGE). As a subsequent step, we use VMAGE to train a variety of simpler student models, effectively distilling the knowledge of the teacher into these students.

As student models, we experiment standard CNN architectures that have been previously used for the task, namely VGG [55], SENet [29], DenseNet [30] and MobileNet [52]. More detail about those architectures will be discussed in Sect. 3.1, but a crucial point is that, according to the published benchmarks [4], each of them has an inference time of 100 ms or less, even on a low power embedded device such as the NVIDIA Jetson TX1. The authors of the teacher method instead report an average execution time of 6.3 s per image (and our experiments confirmed that figure). Therefore, by design, the student models will have at least one advantage over the teacher, the running time, so that the student models could be used in the applications described in the introduction, while the teacher model could not.

The proposed procedure allows the simple student models to achieve state-of-the-art results surpassing the many complex methods, thanks to the fact that they exploit a very large dataset and rely on the labels generated by the teacher method. The same network architectures, trained with the naive approach (i.e., without the VMAGE dataset), yield much lower results, as it will be shown in the experimental analysis of Sect. 4.

The VMAGE dataset provides an estimation of the apparent age for each face. The teacher method [1] that we used is the best performing method in the state of the art according to the ChaLearn LAP 2016 benchmark. We believe that this benchmark provides an accurate estimation of the performance of different methods in realistic

scenarios due to two main reasons. Firstly, the annotation is obtained by crowdsourcing, so that an accurate estimation of the apparent age is used as target rather than the real age: this is arguably an advantage for developing systems that aim to replicate the human ability to estimate age from the appearance. Secondly, the benchmark uses a metric which weights the errors to match the human perception. We give more detail about the LAP benchmark and its metrics in Sect. 3.3.

2.1 Teacher method

In this section, we describe how the teacher method works. More details can be found on the original paper [1]. The "Head Hunter" face detector [41] is applied to the input image to determine the position of the face; it is then aligned using a similarity transform based on the Multi-view Facial Landmark Detector [57] and resized to 224x224 pixels. A total of 8 variants are obtained from each input sample: the original, the horizontal mirror, two rotated versions ($\pm 5^{\circ}$), two horizontally shifted versions ($\pm 5\%$) and two scaled versions ($\pm 5\%$). Each of these images is processed by the classification core of the method.

The classification core is composed of an ensemble of 14 CNN models; each model is based on the VGG-16 architecture, 3 are trained to recognize age in children (0–12 years old), while the others are trained for general age estimation (0–99). Each model outputs a vector of 100 age probabilities and a soft voting rule is used to determine the consensus between the 11 generic models executed on the 8 variants of the image. If the age is higher than 12, the result is determined by the 11 generic models; otherwise, the 3 children models are executed on the 8 variants and the 3×8 vectors of 13 age probabilities are aggregated to produce the final apparent age estimation, expressed as a floating point value.

Each model in the classification core was trained by the authors in multiple steps using the well-known fine-tuning technique. The VGG-16 architecture is pre-trained on the VGG-Face dataset for the task of identity recognition, and then, it is fine-tuned on the *IMDB-Wiki* dataset; each of the 11 general age estimation models is then fine-tuned again using a different 11-fold partition of the LAP training set using distribution label encoding as loss function. The children-specialized models are trained starting from the *IMDB-Wiki* model described above, then fine-tuned again on a children-only private dataset and finally fine-tuned on the children images from the LAP training and validation set using 0/1 classification encoding. Three different checkpoints are chosen to be the 3 members of the children-specialized ensemble. All CNNs are optimized using gradient descent with momentum 0.9 and batch size of 32. Each image is repeated 5 times in a data-augmentation fashion, using horizontal mirroring, random rotation, random shift and random scale.

The teacher method achieved an impressive 0.2433 ϵ-error in the ChaLearn LAP 2016 competition, winning by a large margin on the second classified. However, its accuracy is paid in terms of processing time, which is 6.3 seconds for each face image.

2.2 The VMAGE dataset

The VMAGE dataset is the intermediate product of the proposed knowledge distillation process. We create it in the first step of our procedure in order to transfer the knowledge of the teacher to the student models in the training phase.

The dataset is built upon the image data collected for the task of image identification in the VGG-Face2 dataset [5]. It includes 9116 identities among actors, athletes and other public figures with a total count of 3.3 million faces.

In Table 1 we give details about the composition of the dataset, while the histogram in Fig. 1 allows to note that the distribution of labels by age group is similar to other datasets from literature. In particular ages in the range 25–35 are the most represented, while there are few elders (60–100) and even fewer children (ages 0–15). LAP 2016 and (especially) Adience are exceptions to this rule since they focus on those less represented classes; LFW+ is skewed towards older ages, with most faces in the 45–60 and 60–100 ranges.

We notice that, in absolute numbers, the proposed VMAGE dataset is larger than every other dataset, and this is still true across every considered age group, even the least represented one. Due to the class imbalance, we observe that the VMAGE dataset is most useful as a pre-training tool, while countermeasures can be taken if the target application includes children; fine-tuning on a more balanced dataset is a suitable strategy for fixing the imbalance problem, as we show in our experimentation.

The VMAGE dataset includes age labels for the images that were artificially annotated by the teacher ensemble. The implementation of the teacher method that we used is based on the Caffe framework [33] and was kindly provided by the authors under the GNU GPL-3.0 license. The execution took 962 GPU hours for 3.3 million images and was performed over 2 weeks using 3 NVIDIA Titan X GPUs. We make the labels for the VMAGE dataset available for research purposes[1].

For each face the exact predicted age is given, which is the most versatile approach. Datasets that are annotated with an age class (e.g., child, adult, elder) are useful only if

[1] https://mivia.unisa.it/datasets/vmage.

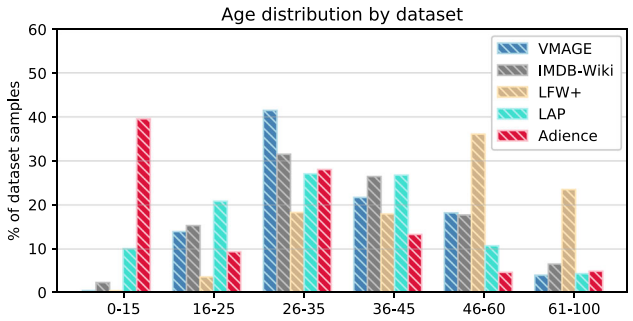


Fig. 1 Relative distribution of the samples in VMAGE, IMDB-Wiki, LFW+, LAP 2016 and Adience within the age groups 0–15, 16–25, 26–35, 36–45, 46–60 and 61–100

the problem is posed as a classification task, with the exact same class boundaries. We argue that this position does not fit all the possible applications and so we report the exact predicted age, allowing for the reuse of the dataset in different types of contexts as shown in Sect. 3.

We retained the intermediate confidence vectors predicted by the teacher ensemble, before voting happens; the agreement between ensemble members may be used a metric of difficulty of each sample, reproducing a multi-annotator labeling scheme like the one used for the construction of the LAP dataset, allowing for additional considerations. Although we did not use those in this work, we believe that this information will be useful for future work on the topic, and then, we decided to make them publicly available.

3 Experimental framework

In order to evaluate the effectiveness of the proposed approach, we train a number of well-known CNN architectures, representative of different families among the most commonly used for face analysis. As a baseline for comparison, we train the same architectures using a prominent strategy from the state of the art, namely we apply a data-cleaning procedure to the large-scale dataset IMDB-Wiki and use the resulting corpus. In Sect. 3.1 we describe the considered architectures, while in Sect. 4 we show how the CNNs trained with our knowledge distillation methodology consistently outperform the corresponding neural networks trained directly on the IMDB-Wiki cleaned corpus.

In order to compare our results with the ones published in the literature and the teacher network, we evaluate our method on the LAP 2016 (a.k.a. APPA-REAL) dataset [14]. In addition, we evaluate the accuracy of the considered CNNs over LFW+ [47] and Adience [12]. The datasets have different characteristics in terms of age distribution, face appearance and a different evaluation protocol. We will describe all these datasets and protocols in detail in Sect. 3.3.

Finally, we will evaluate the robustness of our method to corruptions of the input images: it has been shown that images acquired in real operating environments exhibit a significant amount of diverse types of corruptions, such as Gaussian noise, motion blur, compression artifacts and so on. We will discuss these corruptions in Sect. 3.4 and we will show in Sect. 4 that our training procedure allows the student networks to overcome the accuracy of the teacher in such challenging conditions.

3.1 CNN architectures

In our analysis, we selected 4 different convolutional neural networks, widely adopted in several face analysis tasks: VGG, SENet, DenseNet and MobileNet, each with different characteristics.

VGG, introduced in [55], is the family of CNNs most widely used for face analysis tasks, especially for the availability of a version of VGG-16, namely VGG-Face [44], pre-trained for face recognition by using the VGG-Face2 dataset. Such network, fine-tuned on specific datasets, achieved state-of-the-art performance in gender, ethnicity and emotion recognition. The peculiarity of this CNN architecture is the adoption of 3×3 filters to build larger filters (5×5) in order to obtain a more effective receptive field while reducing the number of weights and the cost of adding convolutional layers. This choice demonstrated to give VGG the capability to achieve good generalization even when the dataset is quite small. In this paper, we use the VGG-16 version, which consists of 13 convolutional and 3 fully connected layers, resulting in 138M weights and more than 15G of operations with 224×224 input size.

SENet, proposed in [29], is based on the well-known ResNet-50 architecture [22], with the addition of the *squeeze and excitation* modules. The ResNet architecture has been designed with the idea to increase the number of layers for achieving higher accuracy. Therefore, a shortcut module learns the residual mapping to solve the problem of vanishing gradients happening in very deep networks (especially in the earlier layers during backpropagation). In addition, it adopts the bottleneck approach by using 1×1 filters to capture cross-channel correlation and reduce the number of weights. The original ResNet-50 consists of 1 convolutional layer, 16 shortcut modules and 1 fully connected layer, resulting in 25.5M weights and 3.9G operations with 224×224 input size. The addition of the modern *squeeze and excitation* modules, namely a particular type of depthwise convolution with dynamic weights, allows to learn a function for giving more importance to specific channels of the input feature map by reducing the

magnitude of the activations in the other channels. This choice demonstrated to increase the accuracy in various computer vision tasks [29].

DenseNet, proposed in [30], is a family of CNNs designed according to the experimental evidence that a CNN can be more accurate and efficient to train if it contains direct connections between input and output layers. In DenseNet, each layer is connected to every other layer (dense blocks) to favor the propagation and the reuse of the feature maps; this concept, widely investigated in recent years, is also known as *feature map aggregation*. To solve the problem that feature maps with different spatial resolution cannot be aggregated, DenseNet complements the use of dense blocks with the adoption of transition layers, which normalize the size of the feature maps computed by the different layers through specific pooling operations. In this paper, we use the DenseNet-121 version, resulting in 7M weights and about 3G operations with 224×224 input size.

MobileNet, described in [52], is a family of CNNs among the most efficient available in the literature, designed for running on board of mobile and embedded devices. It includes the more modern devices for reducing the number of weights and operations while holding a high accuracy, namely residual blocks, depthwise convolutions followed by pointwise convolutions and bottleneck layers. In this paper, we use the newest MobileNet V3 Large and Small versions [28], which also include squeeze and excitation modules, swish nonlinearities and hard sigmoid and are globally optimized through the NetAdapt algorithm. MobileNet-Large requires 5.4M weights and around 219M operations with 224×224 input size, while MobileNet-Small 2.5M weights and about 54M operations with 96×96 input size. Hereinafter, we will refer to these CNNs with the names *MN3-Large* and *MN3-Small*.

3.2 Training

In our experiments, we train all the architectures starting from the ImageNet pre-trained weights. Using pre-trained weights from a large-scale generic dataset is a common strategy in many applications of deep learning, since it allows to alleviate overfitting and improve convergence [6].

In our training pipeline, as a first step the bounding rectangle of the face is localized; for face detection and localization we use a lightweight face detector based on the SSD framework [37]. The face rectangle is expanded to have a square aspect ratio and the image is cropped and resampled with the bilinear algorithm to match the input size of the network. As a final step, from each image we subtract the average value computed separately for each channel on the VGG-Face dataset by the authors [44]; this step allows for the input distribution to be 0-centered on average, allowing to take full advantage of the ReLu nonlinearity and achieve faster convergence.

During the training process, every sample image is perturbed using one of more random augmentation policies [59]. The policies include random crop and horizontal flip, rotation, skew, brightness and contrast. The parameters for these transformations are chosen randomly according to the distributions reported in Table 2; we chose the parameters empirically, ensuring that the augmented images are representative for the dataset.

The training is carried out for 70 epochs and the SGD optimizer is used. The learning rate is initialized to 0.005 and reduced with a factor of 0.2 every 20 epochs. For the VGG-16 network, we use 0.00005 as initial learning rate, since it needs lower learning rates for ensuring convergence; this is due to the architectural peculiarities of this network, namely the absence of batch normalization.

When needed, the CNNs are possibly fine-tuned according to the official evaluation protocol for each considered benchmark, as explained in the following Sect. 3.3.

3.3 Datasets

LFW+ [47] is the dataset that we chose for testing the performance of the student networks in the task of real age estimation. It consists of 15, 699 face images belonging to 8, 000 different subjects. The dataset is not partitioned in training and test set, so we decided to use the whole dataset for our experiments without fine-tuning. This procedure of testing without fine-tuning has been used on the same LFW+ dataset in different tasks such as gender recognition [2, 18]; it is called *cross-dataset evaluation* and allows to assess the generalizability of the features that can be learned through the training dataset.

Table 2 Augmentation policies and parameters used for training. Parameters are randomly computed using the bounded normal distribution $\overline{\mathcal{N}}$, defined as follows $\overline{\mathcal{N}}(\mu,\sigma) = \min(\mu + 2\sigma, \max(\mu - 2\sigma, \mathcal{N}(\mu,\sigma)))$

Policy	Parameter	Value
Crop	Δ_x, Δ_y, Δ_w, Δ_h	$\Delta_x, \Delta_y \sim \overline{\mathcal{N}}(-\frac{\sigma}{10}, \sigma)$
		$\Delta_w, \Delta_h \sim \overline{\mathcal{N}}(\frac{\sigma}{4}, 2\sigma)$
Horiz. flip	Probability P	0.5
Rotation	Degrees q	$\sim \overline{\mathcal{N}}(0, 5)$
Skew	s_x, s_y	$\sim \lvert\overline{\mathcal{N}}(0, 0.05)\rvert$
Brightness	b	$\sim \overline{\mathcal{N}}(0, 24)$
Contrast	c	$\sim \overline{\mathcal{N}}(1, 0.5)$

The evaluation metric we adopt for this dataset is the mean absolute error (MAE). Let us denote with a_i the age predicted on the i-th sample and with r_i the corresponding real label, the MAE is the average error over the K test samples. Being $e_i = |a_i - r_i|$ the error on the i-th sample:

$$\text{MAE} = \frac{\sum_{i=1}^{K} e_i}{K} \tag{1}$$

Testing without fine-tuning allows us to investigate the cross-dataset generalization capability of the networks.

LAP 2016 a.k.a. APPA-REAL [14] is a dataset for estimating the apparent age of people, whose age annotations have been collected through crowdsourcing. It contains 7591 samples, already divided in training (4113), validation (1, 500) and test (1, 978) sets. The experimental protocol requires a standard training or fine-tuning of the neural networks by using the proposed partition. This dataset contains a small number of samples, but it is considered one of the most challenging in terms of face variations and reliable regarding the age annotations. To weight differently the errors done by the neural networks on images annotated with difficulties also by humans, the organizers of the ChaLearn Looking at People challenge [13, 14] designed a specific metric for apparent age estimation, namely the ϵ-*error*. Being m_i and v_i^2 the mean and the variance of the distribution of the predictions a_i done by the annotators for the i-th sample, the estimation error ϵ_i is computed as:

$$\epsilon_i = 1 - e^{-\frac{(a_i - m_i)^2}{2v_i^2}} \tag{2}$$

According to this metric, the error on the i-th sample is normalized by the corresponding variance, in order to penalize less the errors done on samples with high variance. The ϵ-error is finally computed as the mean of the ϵ_i over the K samples of the test set.

Being the dataset already divided in training, validation and test set, we perform the fine-tuning of our CNNs with the same procedure described in Sect. 3.2, by starting from the weights pre-trained on VMAGE of IMDB-Wiki.

Adience [12] is a dataset that we use for age group classification. It is very challenging, produced by automatically extracting images from about 200 Flickr albums, thus collected in uncontrolled conditions and including variations in pose, lighting and image quality. The whole dataset is composed of 26, 580 face images, of which only about one half are almost frontal. A subset of the face images (17, 643) is annotated with 8 unbalanced age categories: 0–2, 4–6, 8–13, 15–20, 25–32, 38–43, 48–53, 60+. Adience is probably the dataset containing more children images in percentage than the other benchmarks publicly available. The standard experimental protocol is a fivefold cross-validation, with the folds already provided by the authors. Being a classification problem, the performance of the neural networks tested on this dataset is evaluated in terms of accuracy, namely the ratio between the number of correct classifications and the total number of samples. Since the dataset is very challenging, the protocol requires the computation of two variants: the *top-1* and the *1-off*. For computing the accuracy top-1, a classification is considered correct whether it corresponds to the true age group; as for accuracy 1-off, the evaluation metric considers correct also the predictions for age groups which are adjacent to the one in groundtruth.

Since the benchmark protocol recommends fine-tuning on predefined folds, we fine-tune our networks using the procedure explained in Sect. 3.2, except that the starting learning rates are 10 times smaller than the ones used for pre-training. To choose the parameters, we ran a first experiment in which we trained on 3 folds and use the 4th for validation for 70 epochs, while the fifth was never used in the training procedure and was saved for testing; with this procedure we established that the optimal number of epochs was about 35 for all the models. Following the approach taken by our predecessors [34], we train our final fine-tuned models on 4 folds for 35 epochs and test on the fifth. Intuitively, given the small size of the Adience dataset we may assume that training on 4 folds will be significantly advantageous over training on 3 folds and using the fourth for validation. Experimental results confirm this intuition, so we report in Sect. 4 the results achieved by the models trained on 4 folds.

Since our networks are pre-trained as regressors, we need a small architectural adjustment for our fine-tuned networks: we remove the last fully connected layer with its one neuron that predict the age and replace it with a fully connected layer with 8 neurons (one for each age group) and add softmax activation. This means that we explicitly convert the network into a classifier and optimize that specifically. All the layers of the network are fine-tuned, since we have empirically found this approach to be more effective with respect to training only the topmost layers.

3.4 Corruptions

Recent studies [23] demonstrate that the modern convolutional neural networks suffer a drop of the accuracy when the input images are affected by strong corruptions, which are common in real environments. Applications of age estimation such as digital signage, access control and social robotics require the use of a network that is robust to these perturbations. In [43] it was shown that a student network trained with knowledge distillation was more robust to image corruptions than the teacher; therefore, we aim to

evaluate the performance drop of the CNNs trained with the proposed approach when applied on corrupted images.

In particular, we reproduce the experimental framework described in [23] and apply 13 different types of corruptions with 5 levels of severity on the LFW+ dataset. The resulting test set, hereinafter LFW+C, is composed of 1, 020, 435 samples. Examples of images extracted from the dataset are depicted in Fig. 2, while more detailed information about the implementation of the image corruptions and the parameters for each severity value are reported in "Appendix 1". In the following we describe the considered blur, noise and digital corruptions.

Blur corruptions Various types of blur can affect the images acquired for real applications, especially in social robotics. *Gaussian blur* may be artificially applied by modern cameras to reduce the negative effect of the acquisition noise. *Defocus blur* can happen when the environment is characterized by a depth of field larger than the limit of the camera. *Zoom blur* appears whether a person moves towards the camera; this corruption can happen in access control applications. *Motion blur* occurs when a person suddenly changes the pose of the face; this category of blur is very common in digital signage and social robotics applications.

Noise corruptions Cameras used for surveillance or on board of a social robot are subjected to overheating, due to 24 hours working or to the external temperature, and may be installed in places characterized by high exposure. These environmental issues cause the presence of random speckles on the acquired images, which can be categorized as two categories of noise. *Gaussian noise* happens when the temperature of the sensor increases over a certain threshold, while *shot noise* occurs in case of high exposure.

Digital corruptions This category incorporates all the digital modifications that can appear on the acquired image due to contrast, brightness, occlusions, compression and rescaling. In particular, *contrast increase*, *contrast decrease*, *brightness increase* and *brightness decrease* happen when the modern cameras apply image corrections such as dynamic contrast and automatic gain control to improve the quality of the acquired images. *Spatter* is instead a corruption introduced to reproduce partial occlusions of the face, which can be due to scarves, glasses, sunglasses, masks, parts of the body or other people; this effect is obtained by adding bright random patterns on the image for low corruption severity and dark elements for higher corruption severity. *JPEG compression* is often applied in real applications running server side to reduce the bandwidth consumption; this effect is reproduced by reducing the compression quality with a value inversely proportional to the severity of the corruption. Finally, *pixelation* is the corruption introduced to reproduce the effect of upscaling, which is typically necessary when the input size of the neural network is higher than the size in pixels of the face image. Considering that the input size of the adopted convolutional neural network is 224×224, this corruption can happen very often when the person is not very close to the camera.

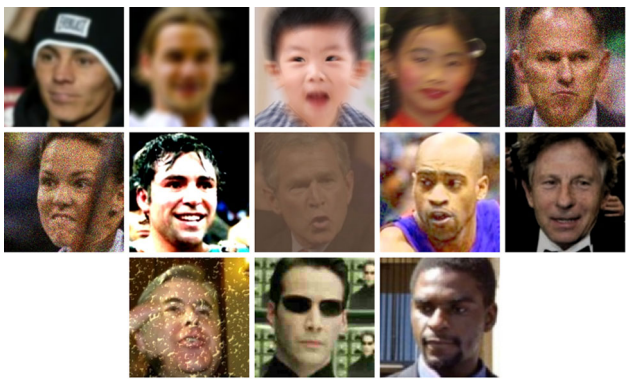


Fig. 2 A collection of 13 samples from the LFW+C dataset, each of them perturbed with a different kind of corruption. More details about the corruption categories, their severity and their mathematical definition are reported in "Appendix 1"

4 Experimental results

4.1 Results on LFW+

As a first experiment, we compare the MAE achieved by each architecture when trained on the distilled VMAGE dataset and on the previously described IMDB-Wiki dataset. In Table 3 we present those results sorted in ascending order of the MAE over the LFW+ dataset that has not been used for training, thus being a fair benchmark. The results show the higher generalization capability obtained by the networks trained with VMAGE; in fact, they achieve a MAE around 1 year lower than the corresponding CNNs trained with IMDB-Wiki.

In the same table, we also report the results on the VMAGE test set and on the IMDB-Wiki test set; as expected [6], the trend is that the performance on the test portion of the dataset used for training is better than the one obtained on an external independent dataset. This is the main reason why most of the competitions allow for fine-tuning on the target dataset. On the VMAGE test set, the models trained on VMAGE achieve significantly lower MAE (up to 3 years) than their IMDB-Wiki-trained rivals, as expected. On the other hand, the advantage of IMDB-Wiki-trained architectures is negligible over the IMDB-Wiki test set (less than 0.1 year in every case). This proves the superior representativeness of the VMAGE dataset with

Table 3 MAE achieved by the considered convolutional neural networks over the test set of VMAGE, IMDB-Wiki and LFW+

Method	Training set	MAE (years)		
		VMAGE	IMDB-Wiki	LFW+
SENet	VMAGE	**1.75**	7.20	**5.58**
VGG	VMAGE	1.82	7.20	**5.58**
MN3-Large	VMAGE	1.84	7.23	5.65
MN3-Small	VMAGE	2.02	7.27	5.69
DenseNet	VMAGE	1.90	7.44	5.89
VGG	IMDB-Wiki	5.56	7.14	6.20
MN3-Small	IMDB-Wiki	4.84	7.17	6.45
DenseNet	IMDB-Wiki	4.82	7.16	6.48
SENet	IMDB-Wiki	5.17	7.23	6.88
MN3-Large	IMDB-Wiki	5.40	**7.11**	7.27

The best results for each dataset are highlighted in bold. The methods are sorted in ascending order of the MAE over LFW+, which can be considered an impartial benchmark, since it was not used for training

respect to the IMDB-Wiki dataset: the networks trained with the former are able to provide better performance on all the datasets, while the networks trained on the latter are comparable only when tested on the IMDB-Wiki but do not generalize as well.

This comparison confirms the effectiveness of the proposed knowledge distillation technique over the naive approach of training with the standard procedure over the IMDB-Wiki dataset.

Among the different architectures, SENet and VGG are the CNNs achieving the best performance over LFW+ (5.58*y*), even if the former obtains a slightly smaller MAE on VMAGE (1.75*y* vs 1.82*y*). The two versions of MN3, Large and Small, achieve a similar MAE (5.65*y* and 5.69*y*), while DenseNet is at the 5th place (5.89*y*).

4.2 Results on LAP 2016

The results achieved in terms of ϵ-error over the LAP 2016 dataset are reported in Table 4. The student model based on SENet obtains a notable 0.3033, which is the best performance on this dataset except for the one obtained by the teacher network during the competition [1]. This result is even more relevant if we consider that this CNN overcomes the performance achieved by complex and bulky CNNs or ensembles of them, such as the ones described in [9, 31, 56, 58]. Examples of face images analyzed by this model are reported in Fig. 3.

In general, all the student CNNs trained with the proposed knowledge distillation technique achieve result very close to the performance obtained by substantially more computationally expensive deep neural networks. VGG and MN3-Large (0.3362 and 0.3404) achieve a performance higher than DenseNet and MN3-Small (0.3589 and 0.3601), but the gap with respect to SENet is significant.

Table 4 ϵ-error achieved by the considered convolutional neural networks over LAP 2016

Method	Pre-training	ϵ-error
Antipov et al. [1]	IMDB-Wiki+Private	0.2411
SENet	VMAGE	0.3033
Tan et al. [56]	Augmented IMDB-Wiki	0.3100
Dehghan et al. [9]	Private	0.3190
Huo et al. [31]	IMDB-Wiki	0.3214
Uricar et al. [58]	IMDB-Wiki	0.3361
VGG	VMAGE	0.3362
MN3-Large	VMAGE	0.3404
DenseNet	VMAGE	0.3589
MN3-Small	VMAGE	0.3601
Malli et al. [40]	IMDB-Wiki	0.3668
Duan et al. [11]	IMDB-Wiki	0.3679
Gurpinar et al. [20]	VGG-Face	0.3740
MN3-Large	IMDB-Wiki	0.3944
DenseNet	IMDB-Wiki	0.4029
MN3-Small	IMDB-Wiki	0.4284
SENet	IMDB-Wiki	0.4351
VGG	IMDB-Wiki	0.4543

The methods are sorted in descending order of the ϵ-error, so that the best result is at the top

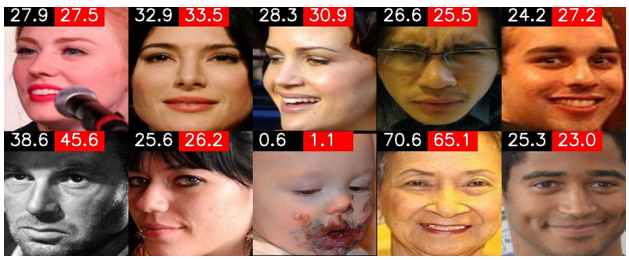


Fig. 3 Examples of LAP 2016 images analyzed by the proposed student model based on SENet. The apparent age in groundtruth is reported in the black box, while the age estimated by the CNN is annotated in the red box

The same architectures trained with IMDB-Wiki achieve a performance which is substantially lower. The highest gap can be noted over SENet (0.3033 vs 0.4351) and VGG (0.3362 vs 0.4543), but also on MN3-Large (0.3404 vs 0.3944), DenseNet (0.3589 vs 0.4029) and MN3-Small (0.3601 vs 0.4284) and it is substantial. It is interesting to note that all the CNNs trained with this procedure are not able to achieve performance comparable with the ones obtained by the methods who participated in the competition; this experimental evidence demonstrates

that state-of-the-art performance is not easily achievable with standard CNNs through the standard pre-training procedure with IMDB-Wiki and further confirms the effectiveness of the proposed technique.

4.3 Results on Adience

The results analyzed so far demonstrated the capability of the proposed training procedure to produce effective CNNs for real and apparent age estimation. In this experiment, whose results are reported in Table 5, we show that the procedure also allows to achieve remarkable performance in age group classification.

In fact, the proposed student model based on SENet holds the third top rank (top-1 65.0%, 1-off 97.1%), followed closely by MN3-Large (top-1 64.1%, 1-off 97.0%), VGG (top-1 64.0%, 1-off 96.9%), DenseNet (top-1 63.5%, 1-off 96.2%) and MN3-Small (top-1 62.5%, 1-off 96.6%). The high accuracy 1-off for all the CNNs pre-trained with VMAGE demonstrates that these models make a negligible mistake, confusing the exact age group with an adjacent one in most cases.

The significant superiority with respect to the corresponding CNNs pre-trained with IMDB-Wiki is a further confirmation of the effectiveness of the proposed technique compared to that typically used in the literature. Indeed, it allows to achieve an accuracy higher or very close to the ones obtained by CNNs more complex, as the residual of residual network (RoR) with 152 layers adopted by Zhang et al. [60] or the already described ensemble of 20 CNNs used by Rothe et al. [50], or architectures tailored for the purpose, such as the VGG-16 modified by Hou et al. [26] with smoothed adaptive activation functions (SAAF) for reducing the regression bias.

Table 5 Accuracy top-1 and 1-off achieved by the considered CNNs over Adience

Method	Pre-training	Top-1	1-Off
Zhang et al. [60]	IMDB-Wiki	67.3	97.5
Hou et al. [26]	IMDB-Wiki	67.3	97.4
SENet	VMAGE	65.0	97.1
MN3-large	VMAGE	64.1	97.0
VGG	VMAGE	64.0	96.9
Rothe et al. [50]	IMDB-Wiki	64.0	96.6
DenseNet	VMAGE	63.5	96.2
Lapuschkin et al. [34]	IMDB-Wiki	62.8	95.8
MN3-small	VMAGE	62.5	96.6
DenseNet	IMDB-Wiki	61.1	95.5
Hou et al. [27]	ImageNet	61.1	94.0
VGG	IMDB-Wiki	60.7	94.5
MN3-large	IMDB-Wiki	60.6	94.3
Liu et al. [36]	ImageNet	60.2	93.7
SENet	IMDB-Wiki	59.9	94.4
Qawaqneh et al. [46]	VGG-Face	59.9	90.6
MN3-Small	IMDB-Wiki	57.6	92.8
Chen et al. [8]	Mixed	52.9	88.4
Levi et al. [35]	No	50.7	84.7
Eidinger et al. [12]	No	45.1	80.7

The methods are sorted in descending order of the accuracy top-1, so that the best result is at the top

4.4 Robustness to image corruptions

In our last experiment, we evaluate the robustness of the considered models to generalize to the image corruptions described in Sect. 3.4. This experiment allows to estimate the performance of these models on images acquired in real scenarios and to compare the robustness of the student models with the one of the teacher.

The results summarized in Fig. 4 confirm the experimental findings reported in [43]. In fact, we notice that three of the student models, namely SENet, MN3-Small and VGG (MAE of 7.87, 7.96 and 8.03 years), are more robust to corruptions than the teacher (MAE = 8.07y). In particular, SENet achieves a MAE 0.2 years lower than the teacher, which in turn obtained a MAE 0.2 years lower on the original LFW+ dataset; this result demonstrates that the proposed distillation technique allows to provide some of the student models with a higher generalization capability than the teacher over corrupted images. In all the cases, the CNNs trained with VMAGE achieve lower MAE (around 0.5 years for SENet and MN3-Large, 1 year for MN3-Small and 0.1 years for VGG) over LFW+C than the corresponding ones adopting IMDB-Wiki, except for

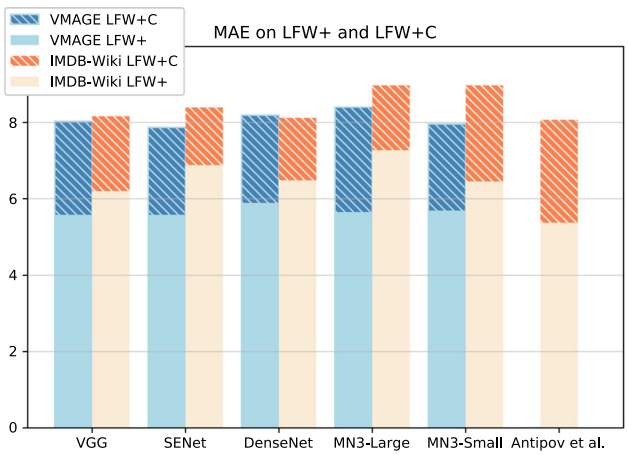


Fig. 4 MAE achieved by the considered CNNs on the LFW+ dataset (light bar) and its corrupted version LFW+C (dark bar). We compare the results achieved when the networks are pre-trained using IMDB-Wiki (orange) and VMAGE (blue) (color figure online)

DenseNet (8.19*y* vs 8.12*y*). A noteworthy result is that obtained by MN3, whose Small version achieves a better performance than the Large one; this result can be explained by the fact that a smaller CNN may generalize better over images very different from the ones used for training.

The analysis can be further extended by evaluating the MAE achieved by the considered CNNs over images perturbed with specific corruption categories; the detailed results of this experiment are reported in Table 6. We can note that the teacher is more robust to Gaussian blur, Gaussian and shot noise, JPEG compression and pixelation, while it suffers in case of brightness and contrast variations, spatter (11.28*y*) and, significantly, in case of defocus blur (18.53*y*, almost 2 years more than the average of the student models). SENet achieves very balanced performance over the different categories and the best MAE when dealing with zoom blur (6.42*y*, as well as VGG), contrast increase (6.20*y*) and decrease (5.91*y*), brightness increase (6.23*y*) and decrease (6.21*y*). MN3-Small is substantially more resilient than other CNNs to spatter (8.91*y*), while VGG obtains the best MAE over motion blur (11.27*y*) and DenseNet over defocus blur (16.19*y*).

In general, we can note that spatter, motion blur and defocus blur are the corruptions causing more problems to the considered CNNs. This evidence can be explained by the fact that these corruptions strongly reduce the facial details, substantially more than the other categories.

Table 6 MAE achieved by the considered CNNs on the corruption categories in LFW+C

		Blur				Noise		Digital					
Method	LFW+C	Gaussian	Defocus	Zoom	Motion	Gaussian	Shot	Cont.Inc.	Cont.Dec.	Brig.Inc.	Brig.Dec.	Spatter	JPEG Comp.	Pixelation
SENet	7.87	7.05	16.63	**6.42**	11.29	6.83	7.08	**6.20**	**5.91**	**6.23**	**6.21**	10.83	5.76	5.82
MN3-small	7.96	7.23	16.54	6.44	11.42	7.48	7.69	6.55	6.36	6.67	6.41	**8.91**	5.89	5.92
VGG	8.03	7.04	16.58	**6.42**	**11.27**	7.05	7.45	6.23	5.97	6.38	6.22	11.93	5.91	5.95
Antipov et al. [1]	8.07	**7.01**	18.53	6.47	11.41	**6.40**	**6.49**	6.97	6.31	6.66	6.31	11.28	**5.50**	**5.59**
DenseNet	8.19	7.66	**16.19**	7.02	12.46	7.11	7.54	6.58	6.76	6.61	6.71	9.70	6.00	6.15
MN3-large	8.97	7.69	19.24	6.72	11.92	7.62	8.27	6.42	6.51	6.65	6.30	10.29	5.80	5.81

The columns are divided in three blocks, one for each corruption category (blur, noise, digital). The methods are sorted in ascending order of the MAE over LFW+C, so that the best result is at the top, while the best MAE for each corruption category is highlighted in bold

5 Conclusions

In this paper we demonstrated that our knowledge distillation approach allows to train effective convolutional neural networks for age estimation. With this method, relying on the use of a teacher model based on an ensemble of 14 CNNs, we produced and made publicly available the age annotations for a new dataset (VMAGE) and designed a handy procedure for training student models based on standard architectures, namely VGG, SENet, DenseNet and MobileNet. All the CNNs trained with this procedure outperformed the corresponding models trained with IMDB-Wiki, the dataset traditionally adopted so far by the scientific community, and achieved remarkable performance over LFW+, LAP and Adience. In particular, the student model based on SENet was able to obtain the best performance over LAP 2016 except for the teacher model, and the third top rank over Adience. The last experiment over LFW+ demonstrated also that the student models based on SENet, MN3-Small and VGG are more robust to image corruptions than the teacher. All these results are obtained with a substantially reduced training effort and

with an inference time about 15 times lower than the teacher model. Possible future directions may include the extension of the training set with more images of children and elders, in order to reduce the imbalance of the dataset; in addition, for age group classification the fine-tuning procedure may exploit the ordinal ranking of the age groups. These devices may further improve the performance of the student models over LAP 2016 and Adience without complicating the training procedure.

Appendix 1: Image corruptions

In the following table, we report the considered corruptions and parameter values adopted to implement the various levels of severity (s). $I(x, y, d)$ refers to the original image and $I_c(x, y, d)$ to the corrupted image, while $d \in \{R, G, B\}$ represents the channel in the RGB color space.

The following figure shows a sample for each type of corruption. Each column represents a type of corruption, with increasing value of severity s in each row.

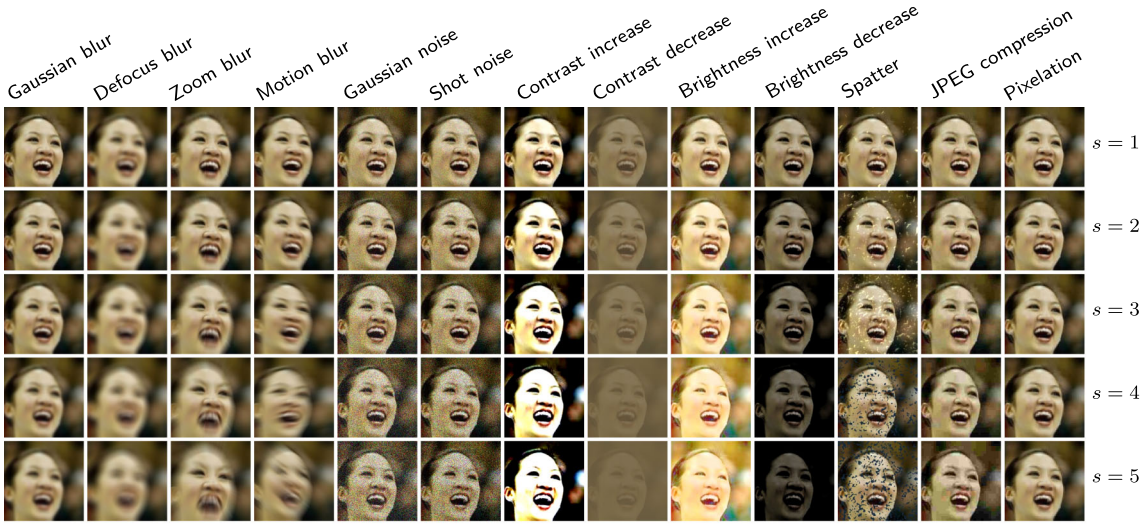
Corruption	Parameter	Values					Description
		$s=1$	$s=2$	$s=3$	$s=4$	$s=5$	
Gaussian blur	σ	1	1.8	2.6	3.4	4	$I_c(x,y,d) = I(x,y,d) * G_\sigma(x,y), \forall d \in \{R,G,B\} G_\sigma(x,y) = e^{(x^2+y^2)/2\sigma^2}$
Defocus blur	r	1.5	2	2	2.5	3	$I_c(x,y,d) = I(x,y,d) * (D_r(x,y) * G_\sigma(x,y))$
	σ	0.1	0.2	0.3	0.4	0.4	being D_r a disc shaped kernel[2] with radius r
Zoom blur	z	1.11	1.18	1.26	1.32	1.4	$I_c(x,y,d) = \frac{1}{\lvert T\rvert}\sum_t zoom_t(I(x,y,d)), t \in T = \{t = 1 + n\epsilon, t <= z, \forall n \in \mathbb{N}\}$
	ϵ	0.01	0.01	0.02	0.02	0.03	$zoom_t(I)$ enlarges the image I by a factor t using linear interpolation[3]
Motion blur	r	3.3	5	5	5	6.7	Implementation from the ImageMagick library[4], with random angle[2]
	σ	1	1.7	2.7	4	5	
Gaussian noise	σ	0.08	0.12	0.18	0.24	0.3	$I_c(x,y,d) = I(x,y,d) + \mathcal{N}(0, \sigma^2)$
Shot noise	q	60	29	15	8	5	$I_c(x,y,d) = Poisson(I(x,y,d) * q)/q$
Contrast inc.	q	1.5	1.9	2.6	3.3	5	$I_c(x,y,d) = (I(x,y,d) - \mu_d) * q + \mu_d$,
Contrast dec.	q	0.4	0.33	0.24	0.16	0.1	where μ_d is the average value of the original image I over the channel d
Brightness inc.	q	0.1	0.2	0.3	0.4	0.5	$I_c(x,y,v) = I(x,y,v) + q, \quad \forall x,y,$
Brightness dec.	q	-0.1	-0.2	-0.3	-0.4	-0.5	where v is the v channel in hsv image representation
Spatter	c_0	0.65	0.65	0.65	0.65	0.67	Implementation by [23]
	c_1	0.3	0.3	0.3	0.3	0.4	
	c_2	4	3	2	1	1	
	c_3	0.69	0.68	0.68	0.65	0.65	
	c_4	0	0	0	1	1	
JPEG compr.	Quality	25	18	15	10	7	Implementation from the Pillow library[5]
Pixelation	q	0.6	0.5	0.41	0.3	0.25	$I_c(x,y,d) = I(\lfloor xf \rfloor/q, \lfloor yf \rfloor/q, d)$,

[2] The r and σ parameter value are expressed in pixels and referred to a 48×48 image. The value of the parameter is scaled proportionally with larger images

[3] Implementation from the SciPy library: https://docs.scipy.org/doc/scipy/reference/generated/scipy.ndimage.zoom.html.

[4] https://imagemagick.org/api/effect.php#MotionBlurImage

[5] https://pillow.readthedocs.io/en/3.1.x/handbook/image-file-formats.html#jpeg



Funding Open access funding provided by Università degli Studi di Salerno within the CRUI-CARE Agreement. This research was partially supported by the Italian MIUR within PRIN 2017 grants, Projects Grant 20172BH297 002CUP D44I17000200005 and by A.I. Tech srl (www.aitech.vision).

Availability of data and material The authors do not provide supplementary data and material.

Code availability The code is available at: https://github.com/MiviaLab/AgeEstimationFramework.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

1. Antipov G, Baccouche M, Berrani SA, Dugelay JL (2016) Apparent age estimation from face images combining general and children-specialized deep learning models. In: Proceedings of IEEE conference on CVPR workshops, pp 96–104
2. Antipov G, Baccouche M, Berrani SA, Dugelay JL (2017) Effective training of convolutional neural networks for face-based gender and age prediction. Elsevier, pp 15–26
3. Ba J, Caruana R (2014) Do deep nets really need to be deep? In: Advances in neural information processing systems, pp 2654–2662
4. Bianco S, Cadene R, Celona L, Napoletano P (2018) Benchmark analysis of representative deep neural network architectures. IEEE Access 6:64270–64277
5. Cao Q, Shen L, Xie W, Parkhi OM, Zisserman A (2018) Vggface2: A dataset for recognising faces across pose and age. In: Proceedings of IEEE international conference on automatic face and gesture recognition, pp 67–74
6. Carletti V, Greco A, Percannella G, Vento M (2020) Age from faces in the deep learning revolution. IEEE Trans Pattern Anal Mach Intell 42(9):2113–2132
7. Chen BC, Chen CS, Hsu WH (2014) Cross-age reference coding for age-invariant face recognition and retrieval. In: Proceedings of Springer ECCV
8. Chen JC, Kumar A, Ranjan R, Patel VM, Alavi A, Chellappa R (2016) A cascaded convolutional neural network for age estimation of unconstrained faces. In: Proceedings of IEEE international conference on BTAS, pp 1–8
9. Dehghan A, Ortiz EG, Shu G, Masood SZ (2017) Dager: deep age, gender and emotion recognition using convolutional neural network. arXiv:1702.04280
10. Deng D, Chen Z, Shi BE (2020) Multitask emotion recognition with incomplete labels. In: 2020 15th IEEE international conference on automatic face and gesture recognition (FG 2020) (FG), pp 828–835
11. Duan M, Li K, Li K (2017) An ensemble CNN2ELM for age estimation. IEEE Trans Inf Forensics Secur 13(3):758–772
12. Eidinger E, Enbar R, Hassner T (2014) Age and gender estimation of unfiltered faces. IEEE Trans Inf Forensics Secur 9(12):2170–2179
13. Escalera S, Fabian J, Pardo P, Baró X, Gonzalez J, Escalante HJ, Misevic D, Steiner U, Guyon I (2015) Chalearn looking at people 2015: apparent age and cultural event recognition datasets and results. In: Proceedings of IEEE ICCV, pp 1–9
14. Escalera S, Torres Torres M, Martinez B, Baró X, Jair Escalante H, Guyon I, Tzimiropoulos G, Corneou C, Oliu M, Ali Bagheri M, et al. (2016) Chalearn looking at people and faces of the world: face analysis workshop and challenge 2016. In: Proceedings of IEEE conference on CVPR workshops, pp 1–8

15. Fu Y, Guo G, Huang TS (2010) Age synthesis and estimation via faces: a survey. IEEE Trans Pattern Anal Mach Intell 32(11):1955–1976
16. Ge S, Zhao S, Li C, Li J (2018) Low-resolution face recognition in the wild via selective knowledge distillation. IEEE Trans Image Process 28(4):2051–2062
17. Greco A, Saggese A, Vento M (2020) Digital signage by real-time gender recognition from face images. In: 2020 IEEE international workshop on metrology for industry 4.0 IoT, pp 309–313
18. Greco A, Saggese A, Vento M, Vigilante V (2020) A convolutional neural network for gender recognition optimizing the accuracy/speed tradeoff. IEEE Access 8:130771–130781
19. Greco A, Saggese A, Vento M et al (2020) Gender recognition in the wild: a robustness evaluation over corrupted images. J Ambient Intell Hum Comput. https://doi.org/10.1007/s12652-020-02750-0
20. Gurpinar F, Kaya H, Dibeklioglu H, Salah A (2016) Kernel ELM and CNN based facial age estimation. In: Proceedings of IEEE conference on CVPR workshops, pp 80–86
21. Han H, Otto C, Liu X, Jain AK (2015) Demographic estimation from face images: human vs. machine performance. IEEE Trans Pattern Anal Mach Intell 37(6):1148–1161
22. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
23. Hendrycks D, Dietterich T (2019) Benchmarking neural network robustness to common corruptions and perturbations. In: International conference on learning representations (ICLR)
24. Hinton G, Vinyals O, Dean J (2015) Distilling the knowledge in a neural network. arXiv:1503.02531
25. Holler J, Casillas M, H Kendrick K, C Levinson S (2016) Turn-taking in human communicative interaction. Frontiers Media SA
26. Hou L, Samaras D, Kurc T, Gao Y, Saltz J (2017) Convnets with smooth adaptive activation functions for regression. In: International conference on artificial intelligence and statistics, pp 430–439
27. Hou L, Yu CP, Samaras D (2016) Squared earth mover's distance-based loss for training deep neural networks. arXiv:1611.05916
28. Howard A, Sandler M, Chu G, Chen LC, Chen B, Tan M, Wang W, Zhu Y, Pang R, Vasudevan V, Le QV, Adam H (2019) Searching for mobilenetv3
29. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7132–7141
30. Huang G, Liu Z. Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708
31. Huo Z, Yang X, Xing C, Zhou Y, Hou P, Lv J, Geng X (2016) Deep age distribution learning for apparent age estimation. In: Proceedings of IEEE conference on CVPR workshops, pp 722–729
32. Iqbal MTB, Shoyaib M, Ryu B, Abdullah-Al-Wadud M, Chae O (2017) Directional age-primitive pattern (DAPP) for human age group recognition and age estimation. IEEE Trans Inf Forensics Secur 12(11):2505–2517
33. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T (2014) Caffe: convolutional architecture for fast feature embedding. arXiv:1408.5093
34. Lapuschkin S, Binder A, Müller KR, Samek W (2017) Understanding and comparing deep neural networks for age and gender classification. In: Proceedings of IEEE ICCV
35. Levi G, Hassner T (2015) Age and gender classification using convolutional neural networks. In: Proceedings of CVPR workshops, pp 34–42
36. Liu H, Lu J, Feng J, Zhou J (2018) Label-sensitive deep metric learning for facial age estimation. IEEE Trans Inf Forensics Secur 13:292–305
37. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC (2016) SSD: Single shot multibox detector. In: European conference on computer vision, pp 21–37. Springer
38. Liu X, Li S, Kan M, Zhang J, Wu S, Liu W, Han H, Shan S, Chen X (2015) Agenet: deeply learned regressor and classifier for robust apparent age estimation. In: Proceedings of IEEE ICCV workshops, pp 16–24
39. Lou Z, Alnajar F, Alvarez JM, Hu N, Gevers T (2018) Expression-invariant age estimation using structured learning. IEEE Trans PAMI 40:365–375
40. Malli RC, Aygun M, Ekenel HK (2016) Apparent age estimation using ensemble of deep learning models. In: Proceedings of IEEE conference on CVPR workshops, pp 714–721
41. Mathias M, Benenson R, Pedersoli M, Van Gool L (2014) Face detection without bells and whistles. In: European conference on computer vision, pp 720–735. Springer
42. Othmani A, Taleb AR, Abdelkawy H, Hadid A (2020) Age estimation from faces using deep learning: a comparative analysis. Comput Vis Image Underst 196:102961
43. Papernot N, McDaniel P, Wu X, Jha S, Swami A (2016) Distillation as a defense to adversarial perturbations against deep neural networks. In: 2016 IEEE symposium on security and privacy (SP). IEEE, pp 582–597
44. Parkhi OM, Vedaldi A, Zisserman A et al (2015) Deep face recognition. In: British machine vision conference (BMVC), vol 1, p 6
45. Punyani P, Gupta R, Kumar A (2020) Neural networks for facial age estimation: a survey on recent advances. Artif Intell Rev 53(5):3299–3347
46. Qawaqneh Z, Mallouh AA, Barkana BD (2017) Deep convolutional neural network for age estimation based on VGG-face model. arXiv:1709.01664
47. Rafique I, Hamid A, Naseer S, Asad M, Awais M, Yasir T (2019) Age and gender prediction using deep convolutional neural networks. In: 2019 International conference on innovative computing (ICIC), pp 1–6
48. Rothe R, Timofte R, Gool LV (2018) Deep expectation of real and apparent age from a single image without facial landmarks. Int J Comput Vis 126(2–4):144–157
49. Rothe R, Timofte R, Van Gool L (2015) Dex: deep expectation of apparent age from a single image. In: 2015 IEEE international conference on computer vision workshop (ICCVW), pp 252–257
50. Rothe R, Timofte R, Van Gool L (2016) Deep expectation of real and apparent age from a single image without facial landmarks. Int J Comput Vis. https://doi.org/10.1007/s11263-016-0940-3
51. Saggese A, Vento M, Vigilante V (2019) Miviabot: a cognitive robot for smart museum. In: Vento M, Percannella G (eds) Comput Anal Images Patterns. Springer, Cham, pp 15–25
52. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC (2018) Inverted residuals and linear bottlenecks: mobile networks for classification, detection and segmentation. arXiv
53. Sawant MM, Bhurchandi K (2019) Hierarchical facial age estimation using Gaussian process regression. IEEE Access 7:9142–9152
54. Schorn C, Elsken T, Vogel S, Runge A, Guntoro A, Ascheid G (2020) Automated design of error-resilient and hardware-efficient deep neural networks. Neural Comput Appl 32(24):18327–18345
55. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
56. Tan Z, Wan J, Lei Z, Zhi R, Guo G, Li SZ (2017) Efficient group-n encoding and decoding for facial age estimation. IEEE Trans Pattern Anal Mach Intell 40(11):2610–2623

57. Uřičář M, Franc V, Thomas D, Sugimoto A, Hlaváč V (2015) Real-time multi-view facial landmark detector learned by the structured output SVM. In: 2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG), vol 2, pp 1–8. IEEE
58. Uricar M, Timofte R, Rothe R, Matas J, Gool LV (2016) Structured output SVM prediction of apparent age, gender and smile from deep features. In: Proceedings of IEEE conference on CVPR workshops, pp 730–738
59. Wang X, Wang K, Lian S (2020) A survey on face data augmentation for the training of deep neural networks. Neural Comput Appl 32:1–29
60. Zhang K, Gao C, Guo L, Sun M, Yuan X, Han TX, Zhao Z, Li B (2017) Age group and gender estimation in the wild with deep RoR architecture. IEEE Access 5:22492–22503