CASCADED FINE-TUNING OF DEEP CONVOLUTIONAL NEURAL NETWORKS FOR AGE ESTIMATION FROM UNCONSTRAINED FACIAL IMAGERY

SUBMITTED IN PARTIAL FULFILLMENT FOR THE DEGREE OF MASTER OF SCIENCE

HENDRIK MATTHIAS TILMAN KÖNIG
s2104903

MASTER MEDIA TECHNOLOGY
FACULTY OF SCIENCE
LEIDEN INSTITUTE OF ADVANCED COMPUTER SCIENCE
LEIDEN UNIVERSITY

2019-11-05

	Primary Advisor	**Secondary Advisor**
Title, Name	Prof. Dr. Holger Hoos	Dr. Jan van Rijn
Affiliation	Universiteit Leiden, FNWI, LIACS	Universiteit Leiden, FNWI, LIACS
Email	h.h.hoos@liacs.leidenuniv.nl	j.n.van.rijn@liacs.leidenuniv.nl

Universiteit
Leiden
Leiden Institute of
Advanced Computer Science

ABSTRACT

Age estimation from facial imagery has been an active research field in the domain of computer vision for many years and various methods have been proposed to encode facial features and map them to age. Those facial feature encodings can be hand-crafted or deep-learned, where the latter relies on Convolutional Neural Networks (CNN) which are able to automatically learn image descriptors from labeled training data. In this work, we present a comparison of different CNN architectures that have previously been applied to the task of age estimation. Furthermore, we propose a multi-step fine-tuning procedure during which we train the models on a large number of training examples, while overcoming the commonly faced issue of label noise in large-scale aging datasets. Using this method, we achieve competitive performance on the FG-NET-AD and Adience benchmark datasets.

1 INTRODUCTION

The human face can be considered the main biometric feature used by humans to identify a person [25]. Besides that, several important attributes may be instantly derived from the human face, such as gender, ethnicity or age.

At the same time, more and more systems rely on the automatic extraction of facial information. One could think of face recognition and identification systems in modern smartphones as one of the most prominent examples. These systems encode facial features of their users into a unique identifier used for authentication purposes. Such facial features, however, can provide further information about a person. More precisely, they can be linked to demographic attributes. The amount of wrinkles on a face, for example, can give information about a person's age [28].

The present work presents a system for automatic age estimation from facial imagery using Deep Convolutional Neural Networks. Estimating age from facial imagery has been a prominent task in the field of computer vision for decades [8, 13, 17, 28, 39] and remains challenging. This is not only because of individual differences in the aging process, but also due to the complex computational tasks required to perform such analysis on facial imagery. That is, given an image as an input, the system has to 1) detect and locate a face, 2) construct a representation of the facial features and 3) make the age prediction based on these encoded features.

Representing facial imagery has been approached in various ways and spans from using facial measurements [11] or micro-patterns [1] to more complex methods such as using deep-learned facial image descriptors [29]. The latter has proven especially useful for dealing with unconstrained image data. This includes images that are taken under uncontrolled conditions, also referred to as in-the-wild imagery, and stands in contrast to constrained imagery, where face images have usually been taken under optimal conditions and in frontal view [32]. Especially for real-world applications, it stands to reason that systems need to be capable to appropriately deal with noise and variation in the input data. In this light, we evaluate our system on constrained, but also on unconstrained face image datasets.

Moreover, we make the following contributions:

- We compare multiple Convolutional Neural Network architectures for their ability to represent facial aging features.
- We propose a novel, multi-step network training procedure using most recent face image datasets.
- We test different classification and regression models for their effectiveness on mapping facial image features to human age.
- Benchmarking against several state-of-the-art methods, we achieve second-best performance on the FG-NET Aging Database [34].
- We investigate the learning behavior of the Convolutional Neural Network models and present a qualitative assessment of the facial features relevant for age estimation.

Possible applications for an automatic age estimation system are manifold. Conceivable in this regard are for instance access control mechanisms; i.e. checking a person's age before giving access to age-restricted areas, either in the virtual or physical space. Another application could be in the field of internet governance, including for example the detection of underage persons in digital image or video content. Especially in such a context it is important that the age estimation system can efficiently deal with unconstrained imagery as previously explained. Furthermore, it has to be able to estimate age across all age groups, and show a high recall when it comes to detecting minors.

The paper is organized as follows. Firstly, we give an overview of historical and current methods for representing face imagery, thereby setting the context for our proposed system (Chapter 2). Next, we provide a review of publicly available aging datasets and the datasets used in this work for model training and evaluation (Chapter 3). Chapter 4 provides information on feature encoding and the age estimation model. In the remaining chapters, experimental results are presented and discussed.

2 RELATED WORK

In general, age estimation models are based on either hand-crafted or automatically learned feature representations. That means the face image has to be processed in such way that relevant facial features are extracted, while this extraction can be done manually, i.e. based on a predefined set of rules, or automatically, i.e. based on a deep learning algorithm. Subsequently, these features are used to fit statistical models for age group classification or ordinal age estimation, respectively. Furthermore, facial features can be divided into global and local features, where the former describe facial texture and shape and the latter refer to partial face regions such as wrinkles, hair or eyes. Both global and local features need to be considered for age estimation since they both give information about a person's age. That is, facial shape mainly changes during childhood and youth and only slightly during adulthood [36]. This stage is instead characterized by texture change, such as changes in skin color or elasticity [13]. An effective age estimation system therefore needs to consider both global and local features and their respective relevance at different ages stages [3]. In the following sections, selected methods for feature extraction and representation will be presented, with special emphasis on deep learning-based methods.

Hand-crafted Feature Representations

The first approach to representing face images in the context of computational age estimation has been presented in the form of Anthropometric Models [28]. These models are mainly concerned with measurements or proportions of the human face [11]. Kwon & Lobo [28] implemented such a model by detecting facial features such as eyes, nose, mouth, chin, top and sides of the face. Based on distance ratios between these landmarks, they were able to distinguish child faces from those of adults and seniors. To refine the classification, they further constructed a wrinkle model in which they detected and measured wrinkle patterns, enabling them to discriminate senior from child and adult faces. Fusing these models resulted in a system capable of classifying facial images into three age classes, that is, as long as these images are taken in frontal view, due to the sensitivity in their geometrical computation method.

Another shortcoming of Anthropometric Models is their focus on facial shape, while at the same time missing important information gained from texture. In order to efficiently detect texture variation, further descriptors have been introduced, most notably Local Binary Patterns [1] and Biologically Inspired Features [19]. The idea of using Local Binary Patterns for face description is based on the fact that faces can be seen as a composition of micro-patterns. Thus, these micro-patterns and their frequency of appearance can be used to construct a global description of a facial image and can furthermore be linked to age [18]. Biologically Inspired Features, on the other hand, present a more sophisticated representation of a facial image by first feeding it through a layer of Gabor filters and then through a pooling layer, where the pooling operation results in the model being more robust to small variation in rotation or scale [19].

Another commonly used model for representing face images is the Active Appearance Model which represents a combination of models of both face shape and (gray-level) texture variation, that is the intensities of all pixels within the target object or face [8]. While Active Appearance Models consider both texture and geometric information, they also rely on a large number of training examples in the form of face images annotated with facial landmarks. Furthermore, dimensionality reduction leads to the loss of sensitive information, as fine-grained facial features might be discarded in that step [20].

With the availability of large scale age datasets, Age Manifold has been introduced as another representation method for face imagery [14]. Using this method, sequential patterns in facial aging are represented in a low-dimensional manifold. Age Manifolds can be considered an extension of the Aging Pattern Subspace [17], in that they aim at finding a common aging pattern rather than a specific pattern for an individual person.

Besides method-specific shortcomings, there is another, general disadvantage to manual feature extraction from facial imagery, as they all rely on images taken under controlled conditions and therefore do not apply well to unconstrained, in-the-wild face image data [47]. Furthermore, making manual choices always requires domain-specific expert knowledge, i.e., on facial aging processes. Lastly, each method tends to extract a specific kind of information from the images at the expense of others. This might explain the relatively strong performance of highly complex models that use multiple feature representations in a combined manner [30, 31].

Deep-learned Feature Representations

Given the restrictions of hand-crafted feature representations, it has been proposed to use Convolutional Neural Networks (CNN) for facial image analysis. A CNN is a type of artificial neural network that can learn representations of the spatial structure of multi-channel images. Their very large number of network parameters must be learned from training examples, equivalent to the learning procedure as understood in the context of Multilayer Perceptrons [41]. Historically, the lack of sufficient amounts of training data as well as the absence of readily available computational power has formed a barrier in creating models capable of complex computing tasks such as object recognition. However, this

Figure 1: Example of a convolutional operation.

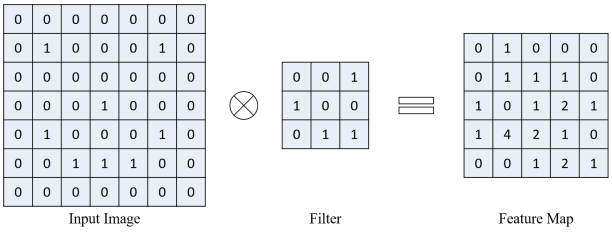
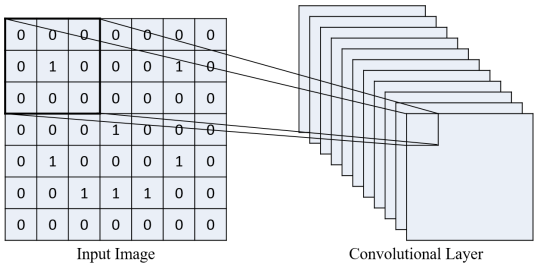


Figure 2: Illustration of a convolutional layer.



has changed with the availability of Graphics Processing Units and the introduction of a CNN capable of classifying the 1.2 million high-resolution images in the 2012 ILSVRC contest, while reaching superior performance to every previous method [27].

Convolutional Neural Networks perform convolutional operations. This means, instead of learning separate weights per input feature, they learn weights which are shared over a local region. Hence, if the network is presented with an image as an input matrix, it keeps the image's spatial information by not considering each pixel separately, but by taking into account their spatial neighborhood. This operation, which is illustrated in Figure 1, is carried out by kernels that move over the input image and it is repeated for each pixel location. The kernels can be viewed as feature detectors and are learned by the network. Each kernel performs a different kind of operation on the input image and thereby extracts different information from the input image. Ultimately, a set of feature maps is created in each network layer, and their output is maximized when corresponding structures are found in the input; see Figure 2 for a visual representation of a convolutional layer.

As the number of convolutional layers increases, their output is passed through nonlinear transfer functions. Furthermore, output maps are downsized using pooling techniques before they are finally transformed into a one-dimensional feature vector. This feature vector is fed into fully connected layers which map the learned features to the desired output.

In the field of age estimation from facial imagery, CNN models have shown significant improvement over previous methods. For instance, the authors of [29] proposed a relatively simple architecture that performed very efficiently on age group classification from an unconstrained dataset. Xing et al. [44] employed a deep multi-task which estimates age based on simultaneously classified race and gender attributes of the subject. In both cases, CNN models were trained from scratch on the target task and dataset. However, deep learning-based models usually require a large dataset to learn appropriate feature representations and they usually suffer from insufficient training data.

Against this background, it is not surprising that more recent approaches applied transfer learning, where the CNN learns image representations from large-scale datasets such as Imagenet [33]. These image descriptors were shown to have sufficient representational power to be applied in other areas and even to fine-grained visual classification tasks, where distinctive class features might be difficult to analyze.

In that context, Rothe et al. [40] proposed a solution using a VGG-16 CNN pre-trained on a generic object recognition task and deployed by the authors of [43] as a base model which was then trained on a specialized age dataset. They furthermore showed that the age can be efficiently estimated as the expected value of all output activations.

Their approach has been extended and improved by Antipov et al. [2] who pre-trained their CNN model on a face recognition and thus on a more related source task than generic object recognition. They furthermore used label distributions as age encodings, as initially proposed by the authors of [16], thereby further enhancing estimation accuracy.

Duan et al. [9] propose a rather complex three-level system including feature extraction via different CNN models, followed by feature fusion, age grouping via an extreme learning machine (ELM) classifier to achieve a more narrow age range and, lastly, age estimation via an ELM regressor. Their CNN models were trained using different targets, namely age, gender and race class, respectively. The main idea is that age estimation improves if features related to age are merged with those related to gender and race. All CNN models were initialized with pre-learned parameters, similar to those presented in the work of Rothe et al. [40].

Most recently, the idea of using pre-trained CNN models as global feature extractors for facial age estimation has been taken up on by the Zhang et al. [47], who combined Residual Network (ResNet) [22] as well as Residual Network of Residual Network (RoR) [46] models trained on generic object recognition tasks with a Long Short-Term Memory unit to enhance the model's ability to pick up fine-grained visual cues related to age. Using this method, they achieved state-of-the-art results on several benchmark datasets and outperformed all previously proposed methods. However,

their results show that implementing the attention mechanism does not significantly improve the performance over the same models without the Long Short-Term Memory unit, shedding light onto the potential of transferring image descriptors from pre-trained Residual Networks onto the age estimation task.

Extending Previous Research

In the present work, we compare different CNN model architectures and we furthermore present a novel, multi-step training procedure for age estimation from facial images. CNN models have been chosen to encode the image features, as those, especially in combination with transfer learning, have evidently demonstrated their effectiveness in the domain of age estimation from face imagery [2, 40, 46, 47]. While borrowing some parts of previously deployed systems, we propose a refined methodology, as we employ models pre-trained on recently introduced large-scale datasets from the related field of face recognition and adapt them to the age estimation task via a cascaded fine-tuning process.

More precisely, we employ ResNet-50, ResNet-50 with Squeeze-and-Excitation blocks (SENet-50) as well as VGG-16 models [23]. These models are pre-trained on a face recognition or identification task using the large-scale face dataset VGGFaces-2 which is the largest dataset of its kind and shows a lot of variation in terms of pose, age or ethnicity, while keeping label noise at a minimum [4]. The reported performance on the face recognition task indicates sufficient representational power and generalization ability, supporting our assumption that these feature representations translate well into the related domain of facial age estimation.

To the best of our knowledge, only Antipov et al. [2] and Rodriguez et al. [38] have followed a comparable strategy, as they used a VGG-16 CNN model pre-trained on the VGGFaces face recognition dataset [35]. This dataset is, however, significantly smaller than the recently introduced VGGFaces-2, with 2.6M (VGGFaces) vs. 3.3M (VGGFaces-2) face images and less variant in terms of identity, age or pose. Besides that, Antipov et al. [2] fine-tune their model on the IMDB-WIKI dataset [39] which, even after semi-manual cleaning, can be subject to incorrect annotations. We account for this not only by relying on multiple data cleaning steps, but also by applying a second fine-tuning step on a high-quality aging dataset with noise-free labels. On the other hand, Rodriguez et al. [38] fine-tune directly on the target dataset, Adience, MORPH-II or IoG, while employing an attention mechanism to pick up relevant aging features; however, they do not evaluate their method on small-scale datasets like FG-NET-AD and it is therefore not apparent whether their approach generalizes well to small datasets in the target domain.

While previous work mostly applied the VGG-16 model to the task at hand, the authors of [47] achieve state-of-the-art results on multiple benchmarks using RoR and ResNet models pre-trained on generic object recognition tasks and fine-tuned on a manually cleaned version of the IMDB-WIKI dataset. In that light, we assume that a residual network can potentially outperform VGG-16 when initialized with parameters learned from a face recognition task and fine-tuned on both high-quantity and high-quality aging datasets. We therefore include residual networks into our analysis and compare VGG-16 to the ResNet-50 as well as SENet-50 CNN models. The SENet-50 can be seen as an extension of the ResNet-50, as it shows a similar base architecture, but contains additional Squeeze-and-Excitation blocks [23]. Squeeze-and-Excitation networks have won the 2017 ILSVRC and have also been applied to face recognition, achieving state-of-the-art results on various benchmarks [4].

3 DATA

We use multiple open-source datasets in the present work. These can be categorized into training data, which is the data we use to adapt our models to the age estimation task, and benchmark data, which is the data we use to evaluate model performance and compare our results to those from other works. Before choosing the datasets to work with, we conducted a review of available aging datasets and present an overview in Table 1. Decision criteria for selecting training datasets are: (i) Quantity of image samples, (ii) quality of image labels and (iii) variance in the label space, which, in our case, refers to a possibly large age range containing children as well as adults and seniors. Benchmark datasets were chosen based on (i) their appearance in related work, (ii) variance in the label space and (iii) image conditions, that is, whether they have been taken in a controlled or uncontrolled (i.e., unconstrained) environment. The latter bears significantly more challenges, as the model has to deal with greater variance in terms of image properties, such as lighting, pose, etc. and therefore allows us to draw better conclusions on the generalization abilities of the model being evaluated.

Training data

IMDB-Wiki. The IMDB-Wiki dataset is, to our best knowledge, the largest, publicly available dataset used in the age estimation domain [39]. Besides age labels, images also have gender annotations. Originally, it contains 523,051 images from more than 20,000 subjects, collected from the IMDB and Wikipedia websites. From this joint dataset, we use the larger IMDB subset containing 460,723 images. However, due to the semi-automatic data collection process, the dataset suffers from a lot of noise in the label space. In order to tackle this problem, we apply multiple cleaning steps and end up

Table 1: Comparison of open-source face aging datasets, sorted by year of publication.

Dataset	Year	Images	Subjects	Age range	Age labels	Unconstrained	Noise-free labels
FG-NET-AD [34]	2002	1,002	82	0-69	Ordinal	No	Yes
MORPH-II [37]	2006	55,134	13,618	16-77	Ordinal	No	Yes
IoG [15]	2009	5,080	23,231	0-66+	Group labels	Yes	No
Adience [10]	2014	163,446	2,000	16-62	Group labels	Yes	No
CACD [5]	2014	163,446	2,000	16-62	Ordinal	Yes	No
IMDB-Wiki [39]	2015	523,051	20,284	0-100	Ordinal	Yes	No
AgeDB [32]	2017	16,488	568	1-101	Ordinal	Yes	Yes

with a final dataset size of 171,858 images. Those cleaning measures are: Removing samples with age labels outside the 0-100 range, removing samples with a face score below 1.0 (as determined by the face detector used when collecting the images), removing samples with more than one face per image and lastly, removing samples with missing gender label. The images are cropped around a face location with a 40% margin.

AgeDB. Age-DB represents the “first manually collected, in-the-wild age database” and contains 16,488 facial images with accurate and noise-free age labels [32]. As opposed to other datasets, such as for instance CACD [5] or IMDB-Wiki [39], Age-DB is a manually collected database and supposedly the only database available providing images that are both unconstrained and accurately labeled with ordinal, real age values. In the present work, this dataset serves as a second fine-tuning set which is motivated by the aforementioned qualities as well as its relatively large sample size. The images are cropped around the face location with a 40% margin.

Benchmark data

FG-NET-AD. The FG-NET Aging Database is a relatively small dataset that contains 1002 color and gray-scale images of 82 subjects with ordinal age annotations [34]. Per individual there are on average 12 images, showing them at different ages. Those ages range from 0 to 69 years. The dataset is constrained, as the images are not taken "in the wild" but under controlled conditions. That means, images only show frontal portraits taken in neutral environments. Some images in the database were collected from digital archives, while others were collected by scanning paper photographs. Despite the controlled environment, diversity in head pose and facial expression require a system with sufficient ability to generalize over this variation. For evaluating age estimation models on FG-NET-AD, the Leave One Person Out protocol is common practice. This protocol represents a 82-fold cross validation which we will also adapt to compare our method to previous work. Example images can be viewed in Figure 3.

Adience. While FG-NET-AD uses numerical values as image labels, Adience [10] provides 26,580 images together with age group labels. More specifically, it divides subjects into 8 age groups ranging from 0 to 60 years and older (0-2, 4-6, 8-12, 15-20, 25-32, 38-43, 48-53, 60-100). The images show a total of 2,284 subjects and are, in contrast to FG-NET-AD, collected "in the wild". This leads to great variation in resolution, head pose and facial expression as well as the occurrence of blur and obscured faces. The evaluation protocol for this dataset suggests a five-fold cross-validation, where splits have been made in such way that images of the same subject do not occur in both training and test sets of the same fold. Example images are displayed in Figure 4.

It should be noted that we end up with only 17,417 images after cleaning the dataset and removing those entries with missing or non-usable, i.e. ordinal and out-of-range age labels. It remains unclear to us how this problem has been tackled in previous works, as those mostly report a dataset size of 26,580 images with exactly 8 age group labels [38, 40, 47]. This also stands in contrast to the number of images which is reported in the original paper, namely 17,643 images distributed over the 8 age groups [10]. The authors note that not all faces could be labeled for age; however, this issue is not explicitly addressed in previous works. Besides that, the creators of the Adience dataset provide the images in two versions: in the first, images have only been cropped around a face location. In the second, they have not only been cropped, but also in-plane aligned. While this version is reportedly used by Levi & Hassner [29], we could not find information on the dataset versions used in other work we reviewed. In the present work, the latter dataset version is used.

Overall, we find these uncertainties important to report, as discrepancies between dataset versions as well as data preparation methods might affect the results and therefore restrict comparability between our and other presented methods.

Label Distribution. It should be noted that ideally we would like to find similar distributions of age labels across all datasets. However, as seen in Figure 5, age in IMDB tends

Figure 3: Example images from FG-NET-AD for one subject. Corresponding age labels are provided below.

Figure 4: Example images from Adience for multiple subjects. Corresponding age group labels are provided below.

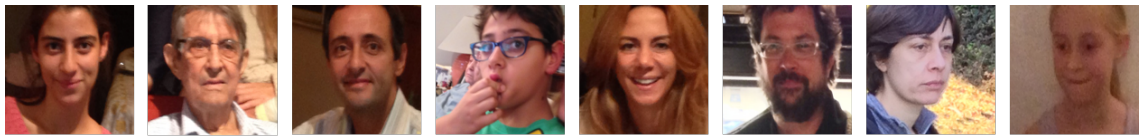


Figure 5: Cumulative Distribution of face images with respect to age for IMDB, AgeDB and FG-NET-AD.

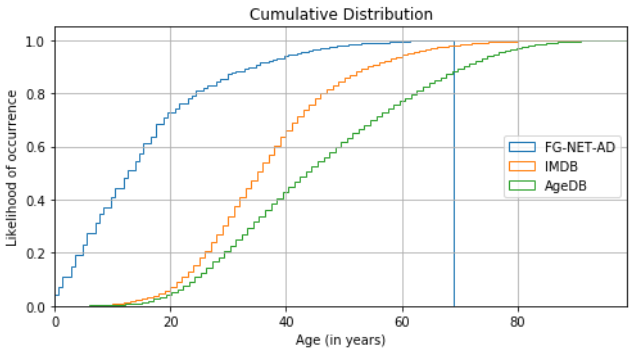
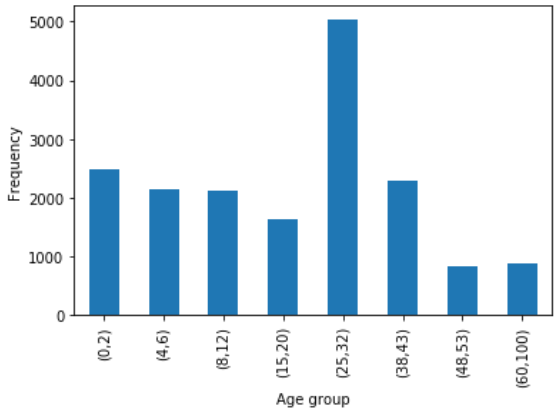


Figure 6: Distribution of face images with respect to age group for Adience.



to be slightly smaller than in AgeDB. FG-NET-AD centers around the age of 20 and therefore gravitates towards an even smaller average age than the aforementioned datasets. Adience is not included in this Figure due to the class labeling; though, the age labels center around age group 25-32 and therefore falls close to the distribution in FG-NET-AD. Moreover, distribution in Adience is not equally balanced among age classes; see Figure 6 for more details.

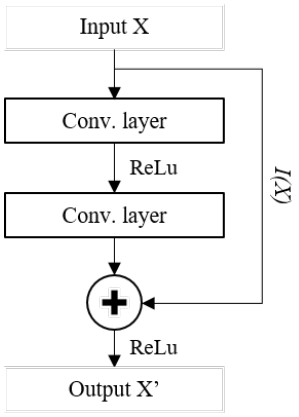
4 METHODOLOGY

Image Feature Representation

We choose three CNN models pre-trained on a face recognition task to encode information from the facial imagery. These models are VGG-16, ResNet-50 and SENet-50 and have been pre-trained on the VGGFaces-2 face dataset [4]. In the following, we will provide an overview of the selected models and the method used for applying them to the age estimation task.

VGG-16. The VGG-16 model has been released in 2014, when it was the runner-up at the ILSVRC [43]. In this network architecture, input is passed through a stack of convolutional layers with filter size set to 3×3. After each convolutional layer block, the input size is reduced by applying spatial pooling. This operation effectively halves the input size, thereby counterbalancing the increase in model complexity caused by a rising number of filters after each convolutional layer block. Output of the final block is fed into a set of fully connected layers which we set to have size 512 and 256, respectively. Lastly, the prediction is made by a softmax layer which we change to have 91, 63, 8 or 3 output neurons,

Figure 7: Schematic of the residual skip connection.



depending on the dataset (IMDB, FG-NET-AD, Adience or modified Adience) the network is being trained on. VGG-16 is 16 layers deep and has 27,717,184 parameters.

ResNet-50. The ResNet-50 is a CNN model from the family of Residual Networks which won the 2015 ILSVRC competition [22]. Prior to their introduction, increasing network depth has frequently led to issues related to network training and often resulted in performance drops as well as high computational expenses. To tackle these problems, Residual Networks rely on skip connections, where the original input to a convolutional layer block is added to its output, before it is passed through a nonlinear transfer function. Thereby, the output of early layers is largely maintained throughout the network, which concurrently learns a residual mapping between the inputs and outputs. Furthermore, it enables the network to skip stacks of layers where weights gravitate towards zero. In these cases, their output equals the input, which allows the network to skip those layers during error propagation. Using these skip connections, together with performing dimensionality down- and up-sampling before and after each residual block, arguably leads to reduced computational expenses and more efficient network training. A residual block is depicted in Figure 7. We leave the model architecture unchanged, except for the final softmax layer which we adapt to the label space of our target dataset, as explained in the previous section. ResNet-50 is 50 layers deep and has 23,867,786 parameters.

SENet-50. The SENet-50 is a ResNet-50 model equipped with Squeeze-and-Excitation blocks [23]. The main idea behind these blocks is to provide the network with access to global information, as opposed to local information gathered by each filter individually, and recalibrate filter responses based on the importance assigned to each channel. To learn the importance of each filter, they are first globally averaged, resulting in a vector with size 1x1xC, where C is the number of filters. Then, in order to enable the model to learn (possibly nonlinear) interaction between those scalar channel descriptors, the input is passed through a ReLu function, followed by a Sigmoid layer. Sigmoid is chosen here, as channel relationships are considered to be non-mutually exclusive. Lastly, fully connected layers are wrapped around the nonlinear transfer function. These layers perform dimensionality reduction (based on a reduction ratio r) before and dimensionality increase after activation and thereby limit model complexity while arguably improving generalization abilities. Multiplying the output weights of the Sigmoid layer with the corresponding feature map finally restores the original input dimensions and moreover results in convolutional layers that embody information on the importance of individual channels in a global context. Figure 9 shows a visual representation of the SE-ResNet module (i.e., a Squeeze-and-Excitation module integrated into the ResNet architecture). Again, we do not change the model architecture, except for the final softmax layer, as explained earlier. SENet-50 is 50 layers deep and contains 26,296,944 parameters.

Transfer Learning. Given the large number of parameters learned by these models and the relatively small size of the dataset at hand, we do not train them from scratch, but apply transfer learning, where the CNN learns image representations from large-scale datasets such as Imagenet [33].

More specifically, we encode image features using pretrained ResNet-50, SENet-50 and VGG models and use those to train classifiers on different tasks within our target domain. As Yosinski et al. [45] have shown, transferability of image features is strongly dependent on the layer from which they are taken and on whether the network is re-trained or fine-tuned on data from the target domain. Most notably, they have demonstrated that image descriptors tend to become more specific with increasing network depth. Thus, the less target and source domain are related, the more the representational power of the final layers decreases. This problem can be worked against by fine-tuning the network, thereby adapting it to the target domain.

We build on those findings by integrating multiple fine-tuning steps during which we re-train the face recognition models on the large-scale IMDB [39] (independent subset of IMDB-WIKI) as well as the high-quality AgeDB aging dataset [32] before applying them to the final benchmarks. Intuitively, we thereby ensure that the CNN models are exposed to (i) a large number of training examples, allowing the model to pick up general feature representations and (ii) a medium number of training examples with fully accurate labeling, ensuring that model training is not constrained

Figure 8: Illustration of the technical framework and different training modes.

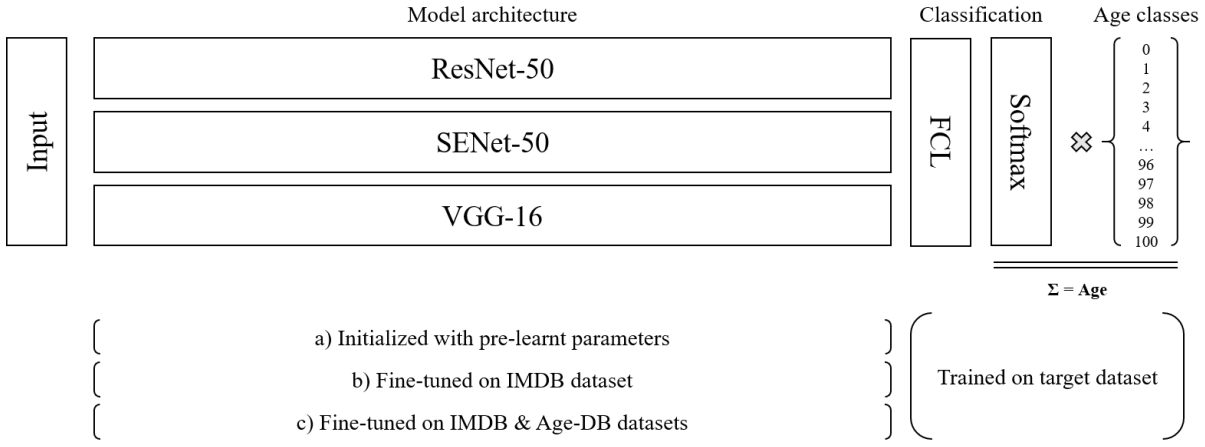
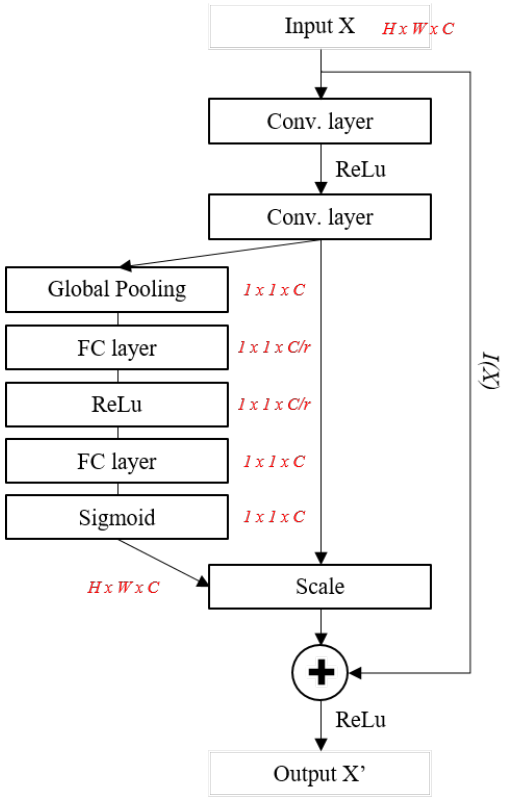


Figure 9: Schematic of the SE-ResNet module. Input dimensions are highlighted in red.



by noise in the data. Dataset details and specifications can be found in section 3.

Training Procedures

We distinguish between three training modes, that is a) using the model with pre-learned parameters and only re-train classification layers, b) fine-tuning the full model (i.e., all parameters unlocked) on the IDMB dataset and then on the FG-NET-AD or Adience target dataset and lastly, c) fine-tuning the full model on the IDMB dataset as well as the AgeDB dataset and then on the target dataset. The general setup as well as the different training modes are visualized in Figure 8

For all training modes, the classification layer weights are initialized using the Gaussian He initialization as proposed by the authors of [21]. Furthermore, we use the ADAM optimizer [26] as our default optimizer. Loss is calculated via the categorical cross-entropy loss function, as the task at hand is a multi-class, single label categorization. On the other hand, batch size and learning rate settings differ between training modes. That is, for each experiment we perform a grid search across learning rates 0.1, 0.01, 0.001 and 0.0001 as well as batch sizes 12, 32, 48 and 64. We test for each combination of those hyper-parameters by training the model over 1 epoch on the training set of IMDB, the training set of AgeDB, the training sets of the first 6 folds of FG-NET-AD or the training set of the first fold of Adience, respectively, and select the setup that achieves the best performance in terms of validation loss.

In mode a), this method leads to a final learning rate of 0.001, with batch size fixed at 48. Classification layers are trained over 20 epochs with $n_{samples}/batchsize$ iterations.

In training mode b), we randomly split IMDB and train on 80% of the dataset over 60 epochs with a learning rate of 0.01 for ResNet-50 and SENet-50 and 0.0001 for VGG-16. Batch size is set to 32 for each model. We apply a learning rate

schedule, halving the learning rate after 5 epochs without improvement in terms of validation loss. To adapt the models to the FG-NET Aging Database, we train ResNet-50 over 20 epochs with learning rate 0.001, SENet over 20 epochs with learning rate 0.01 and VGG-16 over 20 epochs with learning rate 0.0001. Batch size is set to 16 for each model. Here we also apply learning rate schedule, but reduce patience to 3 epochs without improvement.

In training mode c), we only consider the best performing model from the previous round and fine-tune the VGG-16 on AgeDB over 30 epochs with learning rate 0.0001, using the learning rate schedule with patience fixed at 3. Again, we randomly split the dataset using a 80:20 ratio for training and validation. Batch size is set to 32. The resulting model is then applied to the FG-NET-AG as well as the Adience benchmark. For FG-NET-AD, the same hyper-parameter setting applies as in training mode b). On the other hand, we adjust batch size to 32 when finally evaluating on Adience.

In order to account for potential learning biases and to enhance the robustness of the model, we augment our training data by randomly rotating, shifting, cropping and flipping each image in the given training set. Furthermore, we normalize each image and set the input size to 224x224 in the RGB color space.

Age Estimation Regression Model

Besides predicting age via an end-to-end system approach, we also use the best performing CNN model as a feature extractor, where image features are fed into different regressors. We choose this CNN model as the feature extractor, as we assume it to return the most informative image descriptors. In this experiment, we want to see if a regression model can possibly outperform the softmax classification and expected value method.

Features are extracted from the last convolutional layer, whereas we only consider the global maximum of each filter map in order to reduce model complexity. We further compare these features to (i) features extracted from the first convolutional layer and (ii) features that have not been globally pooled, but flattened instead. That is to see, whether generic image descriptors might also yield useful information with regards to facial age and if we loose information when pooling the feature maps.

We choose an automated machine learning approach, and more specifically auto-sklearn [12], where the system is to find and apply the optimal pre-processing methods as well as find and fit the optimal estimators for a new dataset at hand. The output of this system is an ensemble model, where predictions are made by multiple estimators and then weighted based on each estimator's performance.

Implementation Details

All implementations are made in Python 3.6 using Keras [7] running on a Tensorflow back-end for training the CNN models. Besides that, we use the auto-sklearn library [12] for regression analysis. We configure auto-sklearn to run for a total time limit of 12 hours and a maximum of 60 minutes for a single call to the machine learning model; further settings are fixed at default settings. Experiments run on a cluster with 34 nodes, of which 26 are equipped with two Intel Xeon E5-2683 CPUs and 94 GB RAM. Remaining nodes have additional NVIDIA GeForce GTX 1080 Ti GPU with 11 GB memory.

We use the CPU nodes for fitting the regression models and the GPU nodes for CNN model training. Fine-tuning the networks on the IMDB dataset took approximately 2 days (with slight variation between networks), whereas training on AgeDB could be done within around 4 hours. Finally, time consumption for training and testing the models on FG-NET-AD and Adience benchmarks could be drastically reduced by running experiments on every fold in parallel.

5 RESULTS

Metrics

Model performance is measured by the Mean Absolute Error (MAE) they produce. The MAE represents the sum of absolute values of the residuals divided by the number of data points and thus provides a natural measure of average estimation error. Moreover, it was found suitable for the purpose of this research which is an evaluation and inter-comparison of different models and their performance errors. For each experiment, we construct 95% confidence intervals around the mean based on the standard deviation of the errors and use these intervals to determine statistical significance. If not explicitly stated differently, it means that we could not find any statistical significance.

Since our CNN models are trained as classifiers, we cannot straightforwardly calculate the MAE for those. Instead, we calculate the expected value of all output activations, as proposed by Rothe et al. [40], and thus arrive at the MAE. This expected value is the dot-product of the softmax-normalized probabilities in the output layer and the values of the output neurons and can be expressed as:

$$age = \sum_{i=1}^{n} i * a_i \tag{1}$$

where n is the number of neurons in the output layer and a_i is the activation (here, class probability in the softmax layer) of output neuron i.

Table 2: Previous results obtained on the FG-NET dataset.

Dataset	Method	MAE
FG-NET-AD	Pre-trained (Imagenet) ResNet/RoR with LSTM attention mechanism [47]	2.39
FG-NET-AD	Pre-trained (VGGFaces) VGG-16 with Label Distribution Age Encoding [2]	2.84
FG-NET-AD	Pre-trained (Imagenet) VGG-16 with Expected Value [40]	3.09

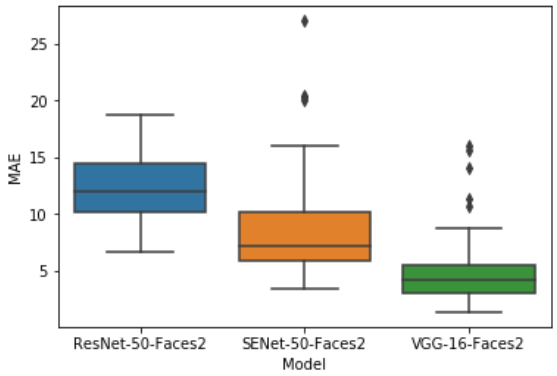
Table 3: Previous results obtained on the Adience dataset.

Dataset	Method	Accuracy
Adience	Pre-trained (Imagenet) ResNet/RoR with LSTM attention mechanism [47]	67.83%
Adience	Pre-trained (Imagenet) VGG-16 with Expected Value [40]	64.00%
Adience	Pre-trained (VGGFaces) VGG-16 with attention mechanism [38]	61.78%
Adience	Shallow CNNs [29]	50.70%

Baselines

Several works have used FG-NET-AD and Adience as benchmark datasets. Their results are reported in form of the MAE in case of FG-NET-AD and Accuracy for Adience and can be found in Table 2 and 3. We use these results as baselines to compare the efficiency of our method against. Though more papers than the ones listed have benchmarked on FG-NET-AD, such as the works of Chen et al. [6] or Hu et al. [24], we only include those into our comparison that reportedly followed the Leave One Person Out protocol. Besides that, those are considered the closest to our approach in methodological terms. Consequently, papers that used Adience for evaluation purposes, such as the works of Chen et al. [6] or Duan et al. [9], have also been discarded if the standard 5-fold cross-validation protocol was not followed.

Figure 10: MAE on FG-NET-AD for CNN models pre-trained on face recognition.



CNN Model Results

In the following section, experimental results are being presented. We do all experiments on FG-NET-AD and hold out Adience until we reach the final setup. This way we ensure to not overfit the model on the benchmark datasets and gain more certainty over the general model performance across different targets.

Age Estimation using pre-trained CNN models. In the first setup, we only train a classifier on top of the pre-trained face recognition model. Experimental results in form of the MAE are shown in Table 4. Most interestingly, VGG-16 produces a lower average error than ResNet-50 and SENet-50, with ResNet-50 showing the weakest performance. However, SENet-50 and VGG-16 produce more error outliers, as depicted in Figure 10. This indicates that they might be more sensitive to age-related facial features, but at the same time do not generalize well across all instances of the test data, i.e. subjects in the dataset.

Age Classification using fine-tuned CNN models. Next, models are fine-tuned on the IMDB dataset and then evaluated on FG-NET-AD. Results can be found in Table 4. Overall, we see the error decreasing for each model, with ResNet-50 showing the largest relative decrease of around 65%, thereby achieving almost similar performance as SENet-50. Remarkably, ResNet-50 and SENet-50 now show more outliers, while the variance in errors produced by VGG-16 has reduced; see Figure 11 for more details. VGG-16 again achieves the lowest error from all three models. It is for this reason that we perform all further experimentation on VGG-16 only, as we consider it to be the most effective for the task (and data) at hand.

The Effect of Cascaded Fine-tuning. In this setup, we further fine-tune the VGG-16-Faces2 model on AgeDB. This additional fine-tuning step reduces the error by 0.09 years and thus achieves the best performance of all experimental

Table 4: Experimental results obtained on FG-NET-AD using CNN models (i) pre-trained on face recognition (ii) fine-tuned on IMDB only and (iii) fine-tuned on IMDB and AgeDB. Values printed in bold represent the lowest error achieved from all experiments. Previous methods are attached below for direct comparison.

Model	MAE	MedianAE	Upper Quartile	Lower Quartile
ResNet-50-Faces2	12.28	12.02	14.43	10.18
SENet-50-Faces2	8.61	7.21	10.13	5.95
VGG-16-Faces2	4.78	4.24	5.54	3.06
ResNet-50-IMDB	4.33	3.46	5.12	2.71
SENet-50-IMDB	4.24	3.20	5.17	2.28
VGG-16-IMDB	2.77	2.34	3.47	1.54
VGG-16-IMDB-AgeDB	**2.68**	**2.24**	**3.35**	**1.58**
ResNet/RoR with LSTM attention mechanism [47]	2.39	-	-	-
VGG-16 with Label Distribution Age Encoding [2]	2.84	-	-	-
VGG-16 with Expected Value [40]	3.09	-	-	-

Figure 11: MAE on the 82 folds of FG-NET-AD for CNN models fine-tuned on (i) IMDB only and (ii) IMDB and AgeDB. Dotted lines represent baseline results from [47] (Red), [38] (Green) and [40] (Blue).

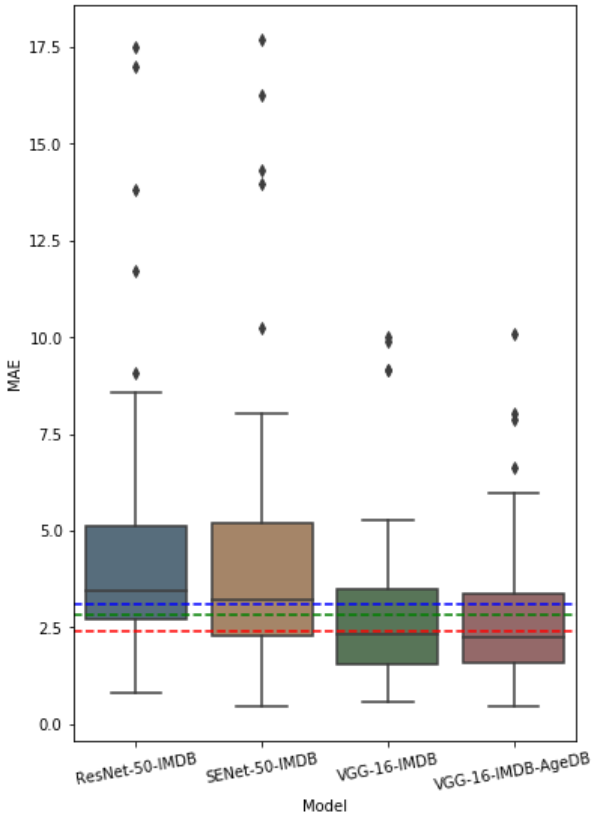


Figure 12: MAE on the 82 folds of FG-NET-AD for regression models using different feature extractors.

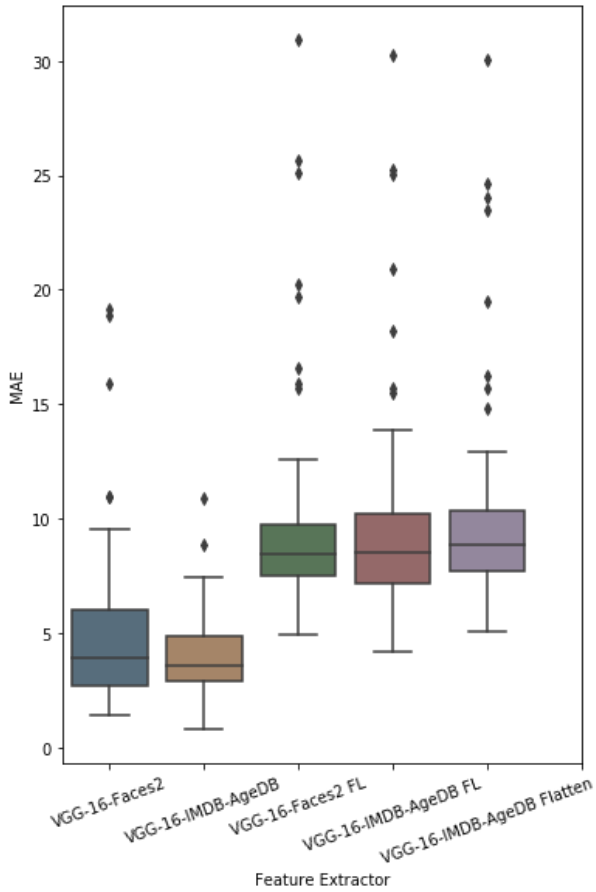


Table 5: Experimental results on FG-NET-AD obtained from regression models. Columns represent the models used as feature extractors.

	MAE	R^2
VGG-16-Faces2	5.05	0.40
VGG-16-IMDB-AgeDB	**3.98**	**0.45**
VGG-16-Faces2 First Layer	9.68	-1.43
VGG-16-IMDB-AgeDB First Layer	9.53	-1.40
VGG-16-IMDB-AgeDB Flattened	9.98	-1.52

rounds, that is an MAE of 2.68. It furthermore achieves competitive, i.e. second-best, results compared to the baselines found in Table 2. See Table 4 and Figure 11 for comparison with previous modes. Interestingly, variance in error reduces after fine-tuning on AgeDB which is mainly due to improvements in the top error margin.

The CNN model fine-tuned on IMDB and AgeDB is lastly applied to the Adience dataset to ensure we did not optimize the models specifically towards FG-NET-AD. On this benchmark, the model achieves an accuracy of **61.41**±4.37% and thus competitive results to the baselines presented in Table 3. It furthermore achieves a 1-off-accuracy of 87.80±2.22% which represents the proportion of predictions in which the model selects a neighboring class of the true label. We again point at the discrepancies in reported dataset statistics as outlined in section 3 and therefore explicitly state that the results reported in this work cannot be directly compared to previous ones. They consequently only apply to the version of the Adience dataset that we ended up with after applying the previously explained pre-processing steps.

Regression Model Results

Lastly, we present the results, in this case the MAE as well as R^2 values, from fitting regression models on each of the 82 folds within FG-NET-AD in Table 5 and Figure 12 and we make several interesting observations. First of all, features extracted from the fine-tuned VGG-16 model seem to provide the most information to the estimators, though they do not outperform the softmax classification/expected value method. Secondly, the results obtained when extracting features from the first convolutional layers of both the face recognition and age estimation model are almost identical. Possibly, this is because both networks rely on the same general features in early layers. This finding is furthermore in line with findings from Yosinksi et al. [45], who showed that even random weights in early layers do not strongly affect model performance, demonstrating that these representations are entirely generic and not related to any specific domain, even after extensive model training.

Interestingly, the worst performance is achieved when extracting features without globally averaging the filter maps. This might be explained by the large number of features that we end up with after flattening the convolutional layer. In fact, this feature extraction mode results in a dataset with the number of dimensions exceeding the number of observations by a multiple.

6 DISCUSSION

Experimental results indicate that the proposed method applies well to the age estimation task. At the same time, they raise several questions, such as:

- Which age group is the most difficult for the network to make predictions about?
- What causes the outliers in prediction errors on the FG-NET Aging Database?
- Which visual features are the most relevant for Age Estimation and how do these differ from those relevant for Face Recognition?
- How do activations differ across the different CNN models we experimented on?

Child vs. Teenager vs. Adult Classification

To investigate the first question, that is, which age group is the most challenging for the age estimation model, we further experiment on a modified version in the Adience dataset. More precisely, we merge age groups in such a way that we end up with 3 (instead of 8) groups; these are: 0-12, 15-20 and 25-100. Hereby, we want to see (i) how well the model can distinguish between the more general age groups children, teenagers and adults and (ii) if there there is an age group that poses a particular challenge to the model.

Overall, this model achieves an accuracy of **92.59%**, that is on fold 1 of the Adience dataset. A confusion matrix can be found in Table 6. Most interestingly, we achieve the highest precision and recall (0.95 and 0.98, respectively) for the "child class", that is age group 0-12. In contrast, the model shows the lowest performance for the "teenager class", spanning over the age range from 15 to 20. However, this class is also significantly underrepresented in the dataset, compared to the merged age groups above and below this range, which might impact results and model performance. Lastly, precision is found to be equally high for the "adult group" (25-100), whereas recall slightly drops to 0.92.

Variance in Prediction Error

When looking at the distribution of error values per held out subject in FG-NET-AD, we find that the model produces significantly large errors for some of them, even though their images do not show any suspicious characteristics that might account for these deviations, at least to the naked eye.

Table 6: Confusion matrix for 3-class age group prediction on the Adience subset. Columns represent the predicted age group, whereas rows show the true age group label.

	0-12	**15-20**	**25-100**
0-12	1646	12	9
15-20	12	36	104
25-100	66	143	2362

Against this background, we hypothesize the source of error to be found not in the image content, but in the distribution of age labels.

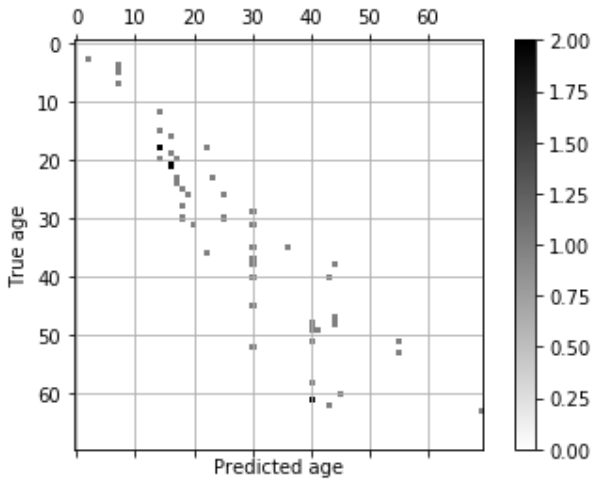
In FG-NET-AD, we independently train the network on 81 subjects and evaluate on the remaining one. Since there are multiple images per subject, showing them at different ages, the model can perform most optimally if the distribution of age labels in this holdout set matches the distribution of the overall dataset, given that image descriptors are capturing all relevant visual information.

However, we find that for some subjects, age label distribution strongly differs from the overall dataset. As shown in Figure 5, more than half of the images in FG-NET-AD are labeled with ages below 20, while the upper regions of the label space are only sparsely sampled. More precisely, for age labels higher than 50 years, an average of 1.75 images is provided per label. Hence, when holding out a fold that contains samples drawn from this region, it is likely that there are only very little or even no training examples for the given class. In consequence, the model is then to make a prediction for a class it has never been presented with during training.

To further investigate this assumption, we select the 4 subjects (2,3,4 and 5) producing the largest prediction error and jointly use those as the test set, while training on all remaining subjects. It should be noted that this experimental setup results in a training set with only 10% of the age labels being larger than 30, whereas more than 50% of the samples in the test set are drawn from this region.

As seen in Figure 13, prediction errors spread out with increasing age, confirming our hypothesis that large prediction errors are mostly found within sparsely sampled label space regions. In order to circumvent this issue, one could introduce weighting into the model; however, we refrain from this as the present work does not aim at optimizing the model towards the FG-NET Aging Database and its unique characteristics, but rather at finding the optimal feature representation for facial age estimation in general. Moreover, we assume the error to be reduced drastically, if all regions in the label space are well sampled.

Figure 13: Confusion matrix for predictions of FG-NET-AD subset containing the top-4 subjects producing the largest prediction errors.



Class Activation Mapping

Lastly, we investigate the learning behavior of the CNN models. That is, we are interested in (i) differences between ResNet-50, SENet-50 and VGG-16 in terms of image (or facial) feature importance and (ii) the effects of fine-tuning the networks away from face recognition and towards age estimation.

To that end, we create Class Activation Maps as proposed by the authors of [42]. These maps can be understood as a visualization tool highlighting which regions of an image are important for class discrimination and used by the CNN to identify a certain class. In detail, this is done by average-pooling the gradients of the softmax layer with respect to the output of the last convolutional layer and multiplying these pooled gradients with each channel in the feature map. Intuitively, we hereby amplify channel output values according to their importance in making the class prediction. Feature maps are then averaged per image region and their output activations are plotted in a heat map. Lastly, this heat map is projected on top of the original input image. It is due to the CNN model's black box character that we choose this rather qualitative approach to investigate learning behavior and feature importance. Generally, more systematic methods of model diagnostics would be preferred.

We randomly select different images from the FG-NET Aging Database and create Class Activation Maps for those. An example can be found in Figure 14. Two phenomena can be observed in the Class Activation Maps. Firstly, when comparing CNN models pre-trained on the VGGFaces-2 dataset, we find their receptive fields to differ quite strongly in size. That

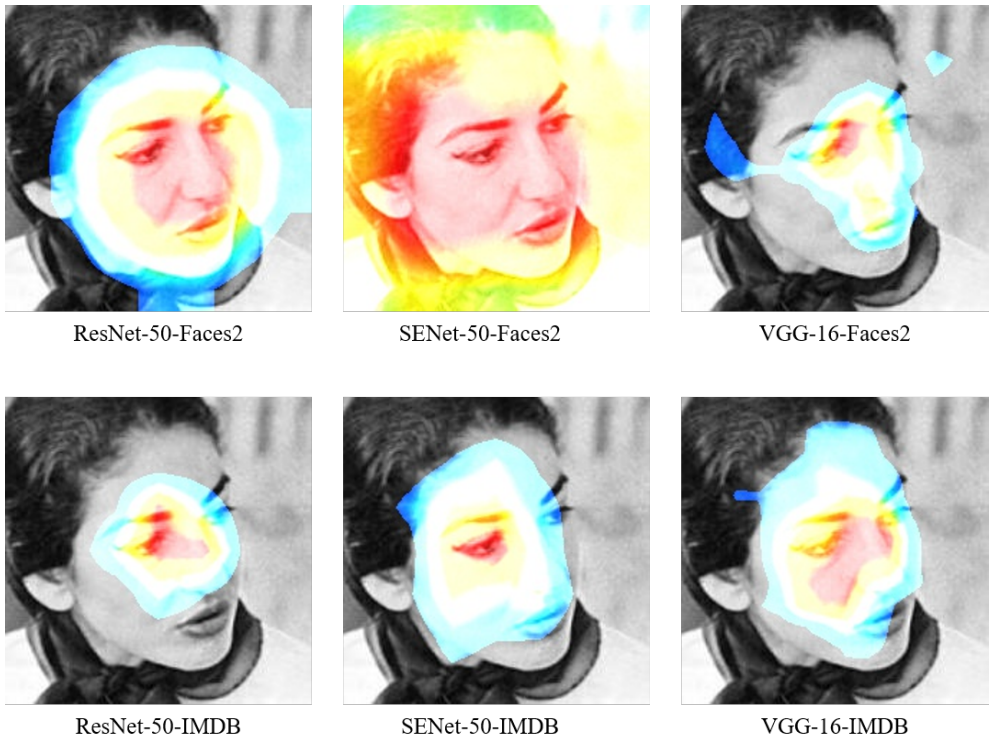


Figure 14: Class Activation Maps for different CNN architectures and model tasks. ***Top:*** **Comparison between ResNet-50, SENet-50 and VGG-16 only trained on face recognition task.** ***Bottom:*** **Comparison between ResNet-50, SENet-50 and VGG-16 fine-tuned on age estimation task.**

is, for SENet-50, activations span the whole image, while they focus on a much smaller section for VGG-16. This degree of specialization might partially explain the relatively good performance of VGG-16 in the first experimental round; see Table 4 for reference.

Receptive fields of each model align after fine-tuning on the age estimation task. That is, activations now center around the eye/nose region and then diminish into immediate surroundings, indicating the importance of this region for age estimation from facial images. Curiously, ResNet-50, and especially SENet-50, very much focus on the eye region, whereas VGG-16 also considers the nose and cheek areas. This might possibly represent a learning bias in the ResNet-50 and SENet-50 models which, given their very large number of parameters and the still relatively small size of the used training data dataset, might be more likely to suffer from overfitting to certain facial characteristics.

Study Limitations

In general, this study and its results are based upon very specific data, as they only consider two aging datasets. This leads to various consequences. First of all, findings are restricted to these datasets and therefore cannot be generalized to the age estimation task in general. Each of these datasets has their own characteristics (sometimes unseen to the eye) and it is not said that the model would generalize well to any face image it gets to see. In fact, we performed further experiments, in which we presented the model with FG-NET-AD images without prior re-training on this dataset and got results that were barely better than random predictions.

The generalization ability of the proposed system is furthermore limited due to the fact that we cannot ensure that all image data has been collected in a proper manner, as there might have been systematic flaws in the data collection process; for example, images of subjects at young ages in FG-NET-AD tend to be in gray-scale, as those are mostly scanned versions of old analog photographs. In that light, all reported results or errors cannot be interpreted as general prediction errors, but statistical errors specific to the dataset at hand.

This also applies to the reported model performances. As we drew from fixed sets of hyper-parameters, we cannot view their errors as general performance measures for those models, but as statistical errors produced under a specific set of circumstances which do not only include model hyper-parameters, but also the way in which the data was prepared;

e.g., how face images were cropped and aligned. One could only draw general conclusions after comparing additional models under a greater range of different conditions. With regard to hyper-parameters, this means that a more thorough analysis of the different settings and their combinations is needed before drawing general conclusions. In that light, hyper-parameters do not only include dataset-specific model configurations, e.g. learning rate or batch size, but also superordinate settings dealing with inter-dependent hyper-parameters, for example, the epoch ratio between different fine-tuning steps.

7 CONCLUSION

Overall, we have shown that CNN models pre-trained on face recognition transfer well onto the related age estimation task. We furthermore have demonstrated that multiple fine-tuning steps on both high-quantity and high-quality datasets positively affects model performance. For that purpose, we have initialized three CNN models, VGG-16, ResNet-50 and SENet-50, pre-trained on a large-scale face recognition dataset and have gradually fine-tuned them towards the age estimation task. This method yields competitive results against various state-of-the-art benchmarks. However, further experimentation is required in order to see if and how this approach can be applied to other datasets or single images.

REFERENCES

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (12):2037–2041, 2006.

[2] G. Antipov, M. Baccouche, S.-A. Berrani, and J.-L. Dugelay. Effective training of convolutional neural networks for face-based gender and age prediction. *Pattern Recognition*, 72:15–26, 2017.

[3] C. Belver, I. Arganda-Carreras, and F. Dornaika. Evaluating age estimation using deep convolutional neural nets. *Electronic Imaging*, 2017(17):100–105, 2017.

[4] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[5] B.-C. Chen, C.-S. Chen, and W. H. Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *European conference on computer vision*, pages 768–783. Springer, 2014.

[6] S. Chen, C. Zhang, and M. Dong. Deep age estimation: From classification to ranking. *IEEE Transactions on Multimedia*, 20(8):2209–2222, 2017.

[7] F. Chollet et al. Keras, 2015.

[8] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):681–685, 2001.

[9] M. Duan, K. Li, C. Yang, and K. Li. A hybrid deep learning cnn–elm for age and gender classification. *Neurocomputing*, 275:448–461, 2018.

[10] E. Eidinger, R. Enbar, and T. Hassner. Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security*, 9(12):2170–2179, 2014.

[11] L. G. Farkas and S. A. Schendel. Anthropometry of the head and face. *American Journal of Orthodontics and Dentofacial Orthopedics*, 107(1):112–112, 1995.

[12] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, and F. Hutter. Efficient and robust automated machine learning. In *Advances in neural information processing systems*, pages 2962–2970, 2015.

[13] Y. Fu, G. Guo, and T. S. Huang. Age synthesis and estimation via faces: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 32(11):1955–1976, 2010.

[14] Y. Fu, Y. Xu, and T. S. Huang. Estimating human age by manifold analysis of face pictures and regression on aging features. In *2007 IEEE International Conference on Multimedia and Expo*, pages 1383–1386. IEEE, 2007.

[15] A. C. Gallagher and T. Chen. Understanding images of groups of people. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 256–263. IEEE, 2009.

[16] X. Geng, C. Yin, and Z.-H. Zhou. Facial age estimation by learning from label distributions. *IEEE transactions on pattern analysis and machine intelligence*, 35(10):2401–2412, 2013.

[17] X. Geng, Z.-H. Zhou, and K. Smith-Miles. Automatic age estimation based on facial aging patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 29(12):2234–2240, 2007.

[18] A. Gunay and V. V. Nabiyev. Automatic age classification with lbp. In *2008 23rd International Symposium on Computer and Information Sciences*, pages 1–4. IEEE, 2008.

[19] G. Guo, G. Mu, Y. Fu, and T. S. Huang. Human age estimation using bio-inspired features. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 112–119. IEEE, 2009.

[20] G. Guo and X. Wang. A study on human age estimation under facial expression changes. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2547–2553. IEEE, 2012.

[21] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[22] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[23] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[24] Z. Hu, Y. Wen, J. Wang, M. Wang, R. Hong, and S. Yan. Facial age estimation with age difference. *IEEE Transactions on Image Processing*, 26(7):3087–3097, 2016.

[25] A. Jain, L. Hong, and S. Pankanti. Biometric identification. *Communications of the ACM*, 43(2):90–98, 2000.

[26] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[28] Y. H. Kwon and N. da Vitoria Lobo. Age classification from facial images. *Computer vision and image understanding*, 74(1):1–21, 1999.

[29] G. Levi and T. Hassner. Age and gender classification using convolutional neural networks. In *Proceedings of the iEEE conference on computer vision and pattern recognition workshops*, pages 34–42, 2015.

[30] C. Li, Q. Liu, W. Dong, X. Zhu, J. Liu, and H. Lu. Human age estimation based on locality and ordinal information. *IEEE transactions on cybernetics*, 45(11):2522–2534, 2014.

[31] K.-H. Liu, S. Yan, and C.-C. J. Kuo. Age estimation via grouping and decision fusion. *IEEE TRANSACTIONS on information forensics and security*, 10(11):2408–2423, 2015.

[32] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou. Agedb: the first manually collected, in-the-wild age

database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 51–59, 2017.

[33] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724, 2014.

[34] G. Panis, A. Lanitis, N. Tsapatsoulis, and T. F. Cootes. Overview of research on facial ageing using the fg-net ageing database. *Iet Biometrics*, 5(2):37–46, 2016.

[35] O. M. Parkhi, A. Vedaldi, A. Zisserman, et al. Deep face recognition. In *bmvc*, volume 1, page 6, 2015.

[36] N. Ramanathan and R. Chellappa. Face verification across age progression. *IEEE Transactions on Image Processing*, 15(11):3349–3361, 2006.

[37] K. Ricanek and T. Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 341–345. IEEE, 2006.

[38] P. Rodríguez, G. Cucurull, J. M. Gonfaus, F. X. Roca, and J. Gonzalez. Age and gender recognition in the wild with deep attention. *Pattern Recognition*, 72:563–571, 2017.

[39] R. Rothe, R. Timofte, and L. Van Gool. Dex: Deep expectation of apparent age from a single image. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 10–15, 2015.

[40] R. Rothe, R. Timofte, and L. Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2-4):144–157, 2016.

[41] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[42] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[43] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[44] J. Xing, K. Li, W. Hu, C. Yuan, and H. Ling. Diagnosing deep learning models for high accuracy age estimation from a single image. *Pattern Recognition*, 66:106–116, 2017.

[45] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014.

[46] A. K. Zhang, B. L. Guo, C. C. Gao, D. Z. Zhao, E. M. Sun, and F. X. Yuan. Age group classification in the wild with deep ror architecture. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1587–1591. IEEE, 2017.

[47] K. Zhang, N. Liu, X. Yuan, X. Guo, C. Gao, Z. Zhao, and Z. Ma. Fine-grained age estimation in the wild with attention lstm networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.