Understanding and Comparing Deep Neural Networks for Age and Gender Classification

Sebastian Lapuschkin
Fraunhofer Heinrich Hertz Institute
10587 Berlin, Germany
sebastian.lapuschkin@hhi.fraunhofer.de

Alexander Binder
Singapore University of Technology and Design
Singapore 487372, Singapore
alexander_binder@sutd.edu.sg

Klaus-Robert Müller
Berlin Institute of Technology
10623 Berlin, Germany
klaus-robert.mueller@tu-berlin.de

Wojciech Samek
Fraunhofer Heinrich Hertz Institute
10587 Berlin, Germany
wojciech.samek@hhi.fraunhofer.de

Abstract

Recently, deep neural networks have demonstrated excellent performances in recognizing the age and gender on human face images. However, these models were applied in a black-box manner with no information provided about which facial features are actually used for prediction and how these features depend on image preprocessing, model initialization and architecture choice. We present a study investigating these different effects.

In detail, our work compares four popular neural network architectures, studies the effect of pretraining, evaluates the robustness of the considered alignment preprocessings via cross-method test set swapping and intuitively visualizes the model's prediction strategies in given preprocessing conditions using the recent Layer-wise Relevance Propagation (LRP) algorithm. Our evaluations on the challenging Adience benchmark show that suitable parameter initialization leads to a holistic perception of the input, compensating artefactual data representations. With a combination of simple preprocessing steps, we reach state of the art performance in gender recognition.

1. Introduction

Since SuperVision [20] entered the ImageNet [33] challenge in 2012 and won by a large margin, much progress has been made in the field of computer vision with the help of Deep Neural Networks (DNN). Improvements in network architecture and model performance have been steady and fast-paced since then [44, 39, 42, 41]. The use of artificial neural networks also has revolutionized learning-based approaches in other research directions beyond classical computer vision tasks, *e.g.* by learning to read subway plans [15], understanding quantum many-body systems [36], decoding human movement from EEG signals [40, 35] and matching or even exceeding human performance in playing games such as Go [37], Texas hold'em poker [29], various Atari 2600 games [25] or Super Smash Bros. [10].

Automated facial recognition and estimation of gender and age using machine learning models has held a high level of attention for more than two decades [21, 30, 6, 16, 13] and has become ever more relevant due to the abundance of face images on the web, and especially on social media platforms. The introduction of DNN models to this domain has largely replaced the need for hand crafted facial descriptors and data preprocessing considerably increased possible prediction performances at an incredible rate. DNN models have been not only successfully applied for age and gender recognition, but also for the classification of emotional states [2]. In the previous three years alone, age recognition rates increased from 45.1% [8] to 64% [32] and gender recognition rates from 77.8% to reportedly 91% [7] on the recent and challenging Adience benchmark [8], mirroring the overall progress on other available benchmarks such as the Images of Groups data set [12], the LFW data set [18] or the Ghallagher Collection Person data set [11].

Next to the indisputable performance gains across the board, the probably most important factor for the popularity of DNN architectures is the low entry barrier provided by intuitive and generic (layer) building blocks, the one-fits-all applicability to many learning problems and most importantly the availability of highly performing and accessible software for training, testing and deployment, *e.g.* Caffe [19], Theano [43], and Tensorflow [1], to name a few, supported by powerful GPU-Hardware.

However, until recently, DNNs and other complex, non-linear learning machines have been used in a black-box manner, providing little information about which aspect of an input causes the actual prediction. Efforts to *explaining* such complex models in the near past have resulted in several approaches and methods [44, 45, 31, 14, 5, 38, 3] allowing for insights beyond the performance ratings obtainable on common benchmarks. This is a welcome development, as in critical applications such as autonomous driving or in the medical domain, it is often of special importance to know why a model decides the way it does, given a certain input, and whether it can be trusted outside laboratory settings [22].

In this paper, we compare the influence of model initialization with weights pretrained on two real world data sets to random initialization and analyze the impact of (artefactual) image preprocessing steps to model performance on the Adience benchmark dataset for different recent DNN architectures. We can show that suitable pretraining can yield a robust set of starting model weights, compensating artefactual representation of the data, via cross-method test set swapping. Using Layer-wise Relevance Propagation [3], we visualize how those choices made prior to training affect how the classifier interacts with the input on pixel level, *i.e*. how the provided input is used to make a decision, and what parts of it. We rectified the performance of [32] on gender recognition referred to in [7] with a more likely result and report our own result, slightly exceeding that baseline. Via a combination of simple preprocessing steps, we can reach state of the art performance on gender recognition from human face images on the Adience benchmark dataset.

2. Related Work

One of the more recent face image data sets is the Adience benchmark [8], which has been published in 2014, containing 26,580 photos across 2,284 subjects with a binary gender label and one label from eight different age groups[1], partitioned into five splits. The key principle of the data set is to capture the images as close to real world conditions as possible, including all variations in appearance, pose, lighting condition and image quality, to name a few. These conditions provide for an unconstrained and challenging learning problem: The first results on the Adience benchmark achieved 45.1% accuracy for age classification and 77.8% accuracy for gender classification using a pipeline including a robust, (un)certainty based in-plane facial alignment step, Local Binary Pattern (LBP) descriptors, Four Patch LBP descriptors and a dropout-SVM classifier [8]. For reference, the same classification pipeline achieves 66.6% accuracy for age classification and 88.6% accuracy for gender classification on the Ghallagher data set. The authors of [17] introduce a 3D landmark-based alignment preprocessing step, which computes frontalized versions of the unconstrained face images from [8], which slightly increases gender classification accuracy to 79.3% on the Adience data set, otherwise using the same classification pipeline from [8].

The first time a DNN model was applied to Adience benchmark was with [24]. The authors did resort to an end-to-end training regime, *e.g*. the face frontalization preprocessing from [17] was omitted and the model was completely trained from scratch, in order to demonstrate the feature learning capabilities of the neural network type classifier. The architecture used in [24] is very similar to the BVLC Caffe Reference Model [19], with the fourth and fifth convolution layers being removed. The best reported accuracy ratings increased to 50.7% for age classification and 86.6% for gender classification, using an over-sampling prediction scheme with 10 crops taken from a sample (4 from the corners and the center crop, plus mirrored versions) instead of only the sample by itself [24].

To the best of our knowledge, the current state of the art results for age and gender predictions are reported in [32] and [7] with 64% and 91% accuracy respectively. The model from [32] was the winner of the ChaLearn Looking at People 2015 challenge [9] and uses the VGG-16 layer architecture [39], which has been pretrained on the IMDB-WIKI face data set. This data set was also introduced in [32] and is comprised of 523,051 labelled face images collected from IMDb and Wikipedia. Prior to pretraining on the IMDB-WIKI data, the model was initialized with the weights learned for the ImageNet 2014 challenge [33]. The authors attribute the success of their model to large amounts of (pre)training data, a simple yet robust face alignment preprocessing step (rotation only), and an appropriate choice of network architecture.

The 91% accuracy achieved by the commercial system from [7] is supposedly backed by 4,000,000 carefully labelled but non-public training images. The authors identify their use of landmark-based facial alignment preprocessing as a critical factor to achieve the reported results. Unfortunately no details are given about the model architecture in use. The authors of [7] compare their results to [32] and other systems, yet only selectively list the age estimation of competing methods, such as [32]. The authors of [7] also report the gender recognition performance of [32] as only 88.75%, which is rather low given the early results from [24], the performance of [32] on age recognition and our own attempts to replicate the models of referenced studies.

Recapitulating, we can identify three major factors contributing to the performance improvements among the models listed in Table 1: (1) Changes in architecture. (2) Prior knowledge via pretraining. (3) Optional dataset preparation via alignment preprocessing.

[1] (0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, 60-)

		gender	age	age (1-off)
[8]	(2014)	77.8	45.1	79.5
[17]	(2015)	79.3	–	–
[24]	(2015)	86.8	50.7	84.7
[32]	(2016)	–	64.0	96.6
[7]	(2017)	91.0	61.3	–

Table 1. An overview over the developments for age and gender recognition results on the Adience benchmark in recent years. Accuracy values are reported in percent.

In the following sections, this paper will briefly describe a selection of DNN architectures and investigate the influence of random weight initialization against pretraining on generic (ImageNet) or task-specific (IMDB-WIKI) real world data sets, as well as the impact of data preprocessing by comparing affine reference frame based alignment techniques to coarse rotation-based alignment. Due to its size and the unconstrained nature of the data and the availability of previous results, we use the Adience benchmark data set as an evaluation sandbox. The dataset is available as only rotation aligned version, and as a version with images preprocessed using the affine in-plane alignment [8], putting the shown faces closer to a reference frame of facial features. We then use Layer-wise Relevance Propagation (LRP) [3] to give a glimpse into the model's prediction strategy, visualizing the facial features used for prediction on a per-sample basis in order to explain major performance differences.

3. Architectures, Preprocessing and Model Initialization

This section provides an overview about the evaluated DNN architectures, data preprocessing techniques and weight initialization choices. All models are trained using the Caffe Deep Learning Framework [19], with code based on `https://github.com/GilLevi/AgeGenderDeepLearning`, containing the configurations to reproduce the results from [24].

3.1. Evaluated Models

We compare the architectures of the model used in [24] (in the following referred to as AdienceNet), the BVLC Caffe Reference Model [19] (or short: CaffeNet), the GoogleNet [42] and the VGG-16 [39], on which state of the art performance on age classification has been reported in [32]. The AdienceNet is structurally similar to the CaffeNet, with the main difference lying in smaller convolution masks learned in the input layer (7×7 vs 11×11) and two less convolution layers being present. The number of hidden units composing the fully connected layers preceding the output layer is considerably lower (512 vs 4096) for AdienceNet. The VGG-16 consists of 13 convolution layers of very small kernel sizes of 2 and 3, which are interleaved with similarly small pooling operations, followed by two fully connected layers with 4096 hidden units each, and a fully connected output layer. The fourth model we use and evaluate is the GoogleNet, which connects a series of inception layers. Each inception layer realizes multiple convolution/pooling sequences of different kernel sizes (sizes 3×3 to 7×7 in the input inception module) in parallel, feeding from the same input tensor, of which the outputs are then concatenated along the channel axis. Compared to the VGG-16 architecture, the GoogleNet is fast to train and evaluate, while slightly outperforming the VGG-16 model on the ImageNet 2014 Challenge with 6.6% vs 7.3% top-5 error in the classification task [33].

3.2. Data Preprocessing

One choice to be made for training and classification is regarding data preprocessing. The SVM-based system from [17] improves upon [8] by introducing a 3D face frontalization preprocessing step, with the goal of rendering the inputs to the pipeline invariant to changes in pose. Landmark-based preprocessing also is identified in [7] as an important step for obtaining the reported model performances. Both [24] and [32] only employ simple rotation based preprocessing, which roughly aligns the input faces horizontally, trusting the learning capabilities of neural networks to profit from the increased variation in the data and learn suitable data representations.

The Adience benchmark data set provides both a version of the data set with images roughly rotated to horizontally aligned faces, as well as an affine 2D in-plane aligned version for download. We prepare training and test sets from both versions using and adapting the original splits and data preprocessing code for [24] available for download on github. We also create a mixed data set from a union of both previous data sets, which has double the number of training samples and allows the models to be trained on both provided alignment techniques simultaneously.

3.3. Weight Initialization

An invaluable benefit of DNN architectures is the option to use pretrained models as a starting point for further training. Compared to random weight initialization, using a pretrained models as starting points often results in faster convergence and overall better model results, due to initializing the model with meaningful filters.

In this paper, we compare models initialized with random weights to models starting with weights trained on other data sets, namely the ImageNet data set and the IMDB-WIKI data sets, whenever model weights are readily available. That is, we try to replicate the results from [24] and train an AdienceModel only from scratch, since no

weights for either pretraining data set are available. Instead, we use the comparable CaffeNet to estimate the results obtainable when initialzing the model with ImageNet weights. We also train the GoogleNet from scratch and initialized with ImageNet weights. Due to the excessive training time required for the VGG-16 model, we only try to replicate the results from [32] and train models both initialized with available ImageNet and IMDB-WIKI weights.

4. Visualizing Model Perception

We complement our quantitative analysis in Section 5 with qualitative insights on the perception and reasoning of the models by explaining the predictions made via the importance of features for or against a decision at input level. Following the success of DNNs, the desire to understand the inner workings of those black box models has vitalized research efforts dedicated to increasing the transparency of complex models. Several methods for explaining individual predictions have emerged since then, with robust yet computationally expensive occlusion-based [44] and sampling-based analysis [31, 45], (gradient-based) sensitivity analysis [14, 5, 38] and backpropagation-type approaches [3, 44, 26] among them. In an intensive study [34], Layer-wise Relevance Propagation (LRP) was found to outperform considered competing approaches in computing meaningful explanations for decisions made by DNN classifiers. Further, the method is in contrast to sampling or occlusion-based approaches computationally inexpensive and applicable to a wide range of architectures and classifier types [3, 22]. We therefore use LRP to supportively complement the quantitative results shown in Section 5 and visualize the perception of the model and its interaction with the input under the evaluated training conditions. For our experiments, we use the current version[2] of the toolbox [23] provided by the authors.

We refer the interested reader to [28] for a tutorial on methods for understanding and interpreting deep neural networks.

4.1. Layer-wise Relevance Propagation for DNNs

LRP is a principled and general approach to decompose the output of a decision function f, given an input x, into so-called *relevance values* R_p for each component p of x such that $\sum_p R_p = f(x)$. The method operates iteratively from the model output to its inputs layer-by-layer in a backpropagation-style algorithm, computing relevance scores R_i for hidden units in the interim. Each R_i corresponds to the contribution an input or hidden variable x_i has had to the final prediction, such that $f(x) = \sum_i R_i$ is true for all layers. The method assumes that the decision function of a model can be decomposed as a feed-forward graph of neurons, *e.g.*

$$x_j = \sigma\left(\sum_i x_i w_{ij} + b_j\right), \tag{1}$$

where σ is some monotonically increasing nonlinear function (*e.g.* a ReLU), x_i are the neuron inputs, x_j is the neuron output and w_{ij} and b_j are the learned weight and bias parameters. The behaviour of LRP can be described by taking as example a single neuron j: That neuron receives a relevance quantity R_j from neurons of the upper layer, which is to be redistributed to its input neurons i in the lower layer, proportionally to the contribution of i in the forward pass:

$$R_{i\leftarrow j} = \frac{z_{ij}}{z_j} R_j \tag{2}$$

Here, z_{ij} is a quantity measuring the contribution of neuron i to the activation of neuron j and z_j is the aggregation of all forward messages z_{ij} over i at j. The relevance score R_i at neuron i is then consequently obtained by pooling all incoming relevance quantities $R_{i\leftarrow j}$ from neurons j to which i contributes:

$$R_i = \sum_j R_{i\leftarrow j} \tag{3}$$

Both the above relevance decomposition and pooling steps satisfy a local conservation property, *i.e.*

$$R_i = \sum_j R_{i\leftarrow j} \quad \text{and} \quad \sum_i R_{i\leftarrow j} = R_j \tag{4}$$

ensuring $f(x) = \sum_i R_i$ for i iterating over the neurons of any layer of the network.

The relevance redistribution obtained from Equations 2 and 3 is a very general one, with exact definitions depending on a neuron or input's type and position in the pipeline [22]. All DNN models considered in this paper consist in one part of ReLU-activated (convolutional) feature extraction layers towards the bottom, followed by inner product layers serving as classifiers [27]. We therefore apply to inner product layers the ϵ-decomposition

$$R_{i\leftarrow j} = \frac{x_i w_{ij}}{b_j + \sum_i x_i w_{ij}} R_j \tag{5}$$

with small epsilon ($\epsilon = 0.01$) of matching sign added to the denominator for numeric stability, to truthfully represent the decisions made via the layers' linear mappings consistently. Since the ReLU activations of the convolutional layers below serve as a gate to filter out weak activations, we apply the $\alpha\beta$ decomposition formula with $\beta = -1$ [3]

$$R_{i\leftarrow j} = \left(\alpha \frac{z_{ij}^+}{\sum_i z_{ij}^+} + \beta \frac{z_{ij}^-}{\sum_i z_{ij}^-}\right) R_j, \tag{6}$$

[2] https://github.com/sebastian-lapuschkin/lrp_toolbox/tree/caffe-wip

which handles the activating and inhibiting parts of z_{ij} separately as z_{ij}^+ and z_{ij}^- and weights them with α and β respectively [3]. Since $z_{ij} = z_{ij}^+ + z_{ij}^-$, enforcing $\alpha + \beta = 1$ ensures the conservation property from Equation 4. Theoretical insights into above decomposition types can be found in [26].

Once relevance scores are obtained on (sub)pixel level, we sum-pool the relevance values over the color channel axis. This leaves us with only one value R_p per pixel p. We visualize the results using a color map centered at zero, since $R_p \approx 0$ indicates neutral or no contribution of input component p to $f(x)$ and $R_p > 0$ and $R_p < 0$ identify components locally speaking for or against the global prediction. All models use vastly different filter sizes (from 2 to 11) in the bottom layers. We follow [4] in distributing R_j for all neurons of some of the lower layers uniformly across their respective inputs, such that the granularity of the visualizations for all models are comparable.

5. Evaluation and Results

We score all trained models using the oversampling evaluation scheme [24], by using the average prediction from ten crops (four corner and one center crop, plus mirrored versions) per sample. Results for age and gender prediction are shown in Tables 2 and 3 respectively. The columns of both tables correspond to the described models; the **A**dienceNet, **C**affeNet, **G**ooglenet and **V**GG-16. Following previous work we also report 1-off accuracy results – the accuracy obtained when predicting at least the age label adjacent to the correct one – for the age prediction task.

The row headers describe the training and evaluation setting: A first value of [i] signifies the use of [i]n-plane face alignment from [8] as a preprocessing step for training and testing, [r] stands for [r]otation based alignment and [m] describes results obtained when both rotation aligned and in-plane aligned images have been [m]ixed for training and images from the [r] test set have been used for evaluation. Second values [n] or [w] describe weight initialization using Image[n]et and IMDB-[w]IKI respectively. No second value means the model has been trained from scratch with random weight initialization.

The results in above tables list the measured performance after a fixed amount of training steps. Intermediate models which might have shown slightly better performance are ignored in favour of comparability. With our attempt to replicate the results from [24] based on the code provided by the authors, we managed to exceed the reported results in both accuracy by ($+1.2\%$) and 1-off accuracy ($+2.7\%$) for age prediction and accuracy ($+1.5\%$) for gender prediction. As expected, the structurally comparable CaffeNet architecture obtains relatable results for both learning problems with random model weight initialization. We then further compared the relatively fast to train CaffeNet model to the GoogleNet model in all data preprocessing configurations when trained from scratch and fine-tuned based on the ImageNet weights. We try to replicate the measurements from [32] to verify the observations made based on the other models. Here, we did not fully manage to reach the reported results, despite using the model pre-trained on the IMDB-WIKI data as provided by the authors. However, we closely scrape by the reported results with slight differences in both accuracy (-1.2%) and 1-off accuracy (-0.8%), averaged over all five splits of the data with a model trained on the mixed training set. In all evaluated settings shown in Figure 1 we can observe overall trends in choices for architecture, dataset composition and preprocessing and model initialization.

	A	**C**	**G**	**V**
[i]	51.4 87.0	52.1 87.9	54.3 89.1	–
[r]	51.9 87.4	52.3 88.9	53.3 89.9	–
[m]	53.6 88.4	54.3 89.7	56.2 90.7	–
[i,n]	–	51.6 87.4	56.2 90.9	53.6 88.2
[r,n]	–	52.1 87.0	57.4 91.9	–
[m,n]	–	52.8 88.3	58.5 92.6	56.5 90.0
[i,w]	–	–	–	59.7 94.2
[r,w]	–	–	–	–
[m,w]	–	–	–	62.8 95.8

Table 2. Result for **age** classification in accuracy in percent, using oversampling for prediction. Small numbers next to the accuracy score show 1-off, e.g the accuracy with which at least an adjacent age group has been predicted.

	A	**C**	**G**	**V**
[i]	88.1	87.4	87.9	–
[r]	88.3	87.8	88.9	–
[m]	89.0	88.8	89.7	–
[i,n]	–	89.9	**91.0**	**92.0**
[r,n]	–	90.6	**91.6**	–
[m,n]	–	90.6	**91.7**	**92.6**
[i,w]	–	–	–	90.5
[r,w]	–	–	–	–
[m,w]	–	–	–	**92.2**

Table 3. Results for **gender** classification in accuracy, using oversampling for prediction. Bold values match or exceed the currently reported state of the art results from [7] on the Adience benchmark.

5.1. Remarks on Model Architecture

In all settings, the CaffeNet architecture is outperformed by the more complex and deep GoogleNet and VGG-16 models. For gender classification under comparable settings, the best VGG-16 models outperform the best GoogleNet models. Figure 2 visualizes the different characteristics of input faces as used by the classifiers to predict

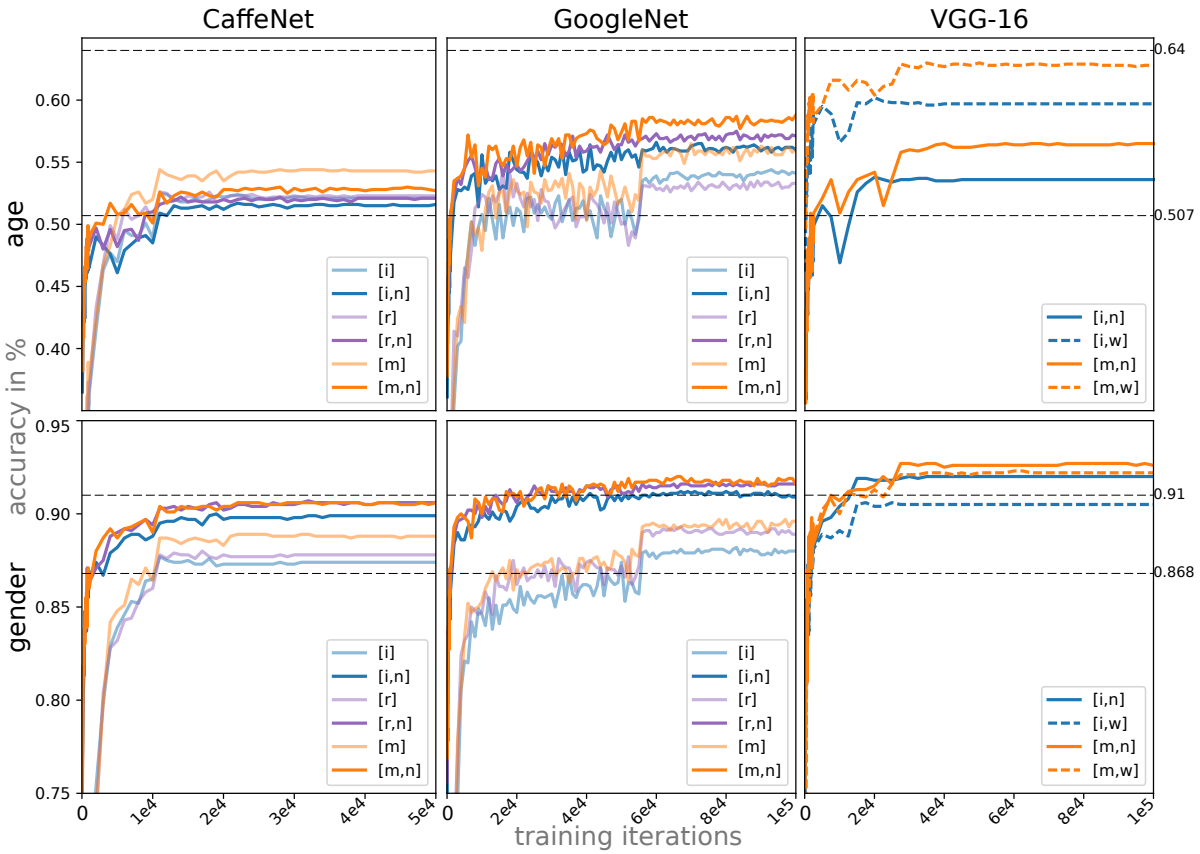

Figure 1. The plots are ordered column-wise over model architectures and row-wise according to prediction problem, showing model performance over training time given different initializations and data preprocessing settings. The top and bottom dashed lines in each plot show worst and best reference accuracy results from [24, 32] and [7], with the horizontal axis increasing with training iterations. Thick lines show results taken by us. Color coding corresponds to data preprocessing and shading to model initialization: Blue color stands for affine [i]n-plane alignment. Violet lines correspond to [r]otation alignment. Orange lines show the model performances for training on the [m]ixed training set. Translucent line color stands for training with random model initialization, fully opaque and solid lines show performance for finetuning on ImageNet weights and dashed lines correspond to model initialization using IMDB-WIKI weights, only applied to the VGG-16 model. All results are averaged over the five splits of the Adience data set.

either male or female gender.

We observe that model performance correlates with network depth, which in turn correlates with the structure observable in the heatmaps computed with LRP. For instance, all models recognize female faces dominantly via hair line and eyes, and males based on the bottom half of the face. The CaffeNet model tends to contentrate more on isolated aspects of a given input compared to the other two, especially for men, while being less certain in its prediction, reflected by the stronger negative relevance.

5.2. Observations on Preprocessing

For all three models, we can observe the overall trend for both prediction problems, that the in-plane alignment preprocessing step is not beneficial to classifier performance, compared to rotation alignment. The only exception to this trend is the randomly initialized GoogleNet model, which loses one percent accuracy for age prediction under rotation alignment albeit still gaining performance in measured 1-off prediction. We reason the better performance on only rotation aligned images to be justified in the potential of and for DNNs to learn for the domain of face images canonically meaningful sets of features. For the face images aligned using the technique presented in [8], this is more difficult. Especially for images of children, the faces aligned to reference frames suitable for adults result in head shapes of uncharacteristic aspect ratios for the age group or even faulty alignments. Figure 3 demonstrates the nature of this artefactual noise introduced to the data by unsuitable alignment.

All models benefit the most from combining both the rotation aligned and the landmark aligned data sets for training. For one, this effectively doubles the training set sizes, but also – perhaps more importantly – allows the learning of a more robust feature set: The models trained on a com-

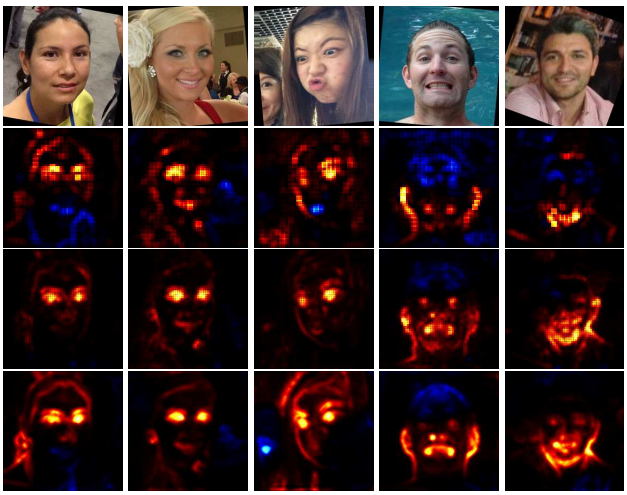

Figure 2. From top to bottom: Input image, followed by relevance maps for the best performing **CaffeNet**, **GoogleNet** and the **VGG-16** model for **gender** prediction. Hot colors identify parts of the image contributing to the predicted class. Cold hues show evidence contradicting the predicted class, as perceived by the model. Smoother heatmaps are a consequence of smaller filters and stride in the bottom layers.

bination of both the landmark aligned and rotation aligned images perform well on test sets resulting from both preprocessing techniques. Tables 2 and 3 show results for models trained on the combined set which were evaluated on the rotation aligned test set. Performance measurements on in-plane aligned data are with $< 1\%$ only insignificantly lower.

In order to underline the effect of increased robustness of the models trained on the more diverse [r]oration aligned training set we evaluated models trained on [i]n-plane aligned images with [r]otation aligned test images and vice versa. Corresponding model performances are listed in Tables 4 and 5. Some models trained on data prepared with one alignment technique evaluated against the test set of the other perform even worse than the early SVM-based models from [8], despite their competitive results from the combined training set. The models trained on the in-plane aligned images have more difficulty predicting on the unseen setting than the models trained on the only rotated images, where the original facial pose and the proportions of the face image are mostly preserved.

For the VGG-16 model, we compared the in-plane alignment to the mixed training set – the worst to the best expected results. Here again, the mixed training data results in a better model than when only in-plane alignment is used. Figure 1 shows an overview of all results over training time.

5.3. Observations on Initialization

We find that the GoogleNet model responds well to fine-tuning on the weights pre-trained on ImageNet and responds with an increase in performance for both classification problems and in all dataset configurations. The CaffeNet, however, slightly loses performance when fine tuned for age group prediction, while benefiting in gender prediction. The better response of the GoogleNet compared to the CaffeNet, when initialized with their respective ImageNet weights might be caused by the quality of the initial parameters: While the GoogleNet achieves a 6.6% top-5 error on ImageNet, the CaffeNet only reaches 19.6%. Evaluating on the *incorrect* test data (Tables 4 and 5), both fine tuned models trained on rotation aligned images manage to recover their respective performance ratings compared to models trained from scratch and being evaluated on the *correct* data. The GoogleNet model even exceeds the performance of the same architecture initialized randomly but both trained and evaluated on the rotated images. The measurable beneficial effect of appropriate pretraining is visualized in Figures 4 and 5. ImageNet pretraining leads to the use of larger and meaningful parts of the face for prediction for the GoogleNet, while the randomly initialized model picks out single characteristics during training which correlate the most with the target class. This includes eyebrows and lips defining female faces and nose, chin and uncovered ears for men for gender recognition. We see comparable results for the VGG-16 on age group estimation when comparing pretraining on ImageNet and IMDB-WIKI. The model initialized with IMDB-WIKI weights, with the pretraining task being age estimation on 101 age categories, concentrates more on the facial features themselves, while the ImageNet-initialized one is more prone to distraction from background elements and clothing items. Facial features seen in examples of opposing classes of the respectively weaker models in both figures – independent of the ensemble of facial features – leads to less certain, noisy decisions. For the problem of gender recognition, the VGG-16 is affected less from weight initialization than from the quality of data preprocessing. Here, IMDB-WIKI pretraining might have an only diminished effect due to firstly the ImageNet weights providing an already good set of starting weights and secondly, the pretraining objective (age recognition) being orthogonal to the task of gender recognition. In fact, other than for age recognition, the VGG-16 models initialized with ImageNet weights converged to better parameters than their counterparts.

Figure 1 reports the prediction performances of the CaffeNet, the GoogleNet and the VGG-16 model in all evaluated settings, averaged over the five splits of the Adience data set. The recorded model scores over time illustrate that suitably initializing a model largely outweighs the problems introduced with artefactual data in our experiments. Next to the overall better model obtained after convergence, we also observe a considerably faster increase in the learning progress early in training.

Figure 3. Top: Samples taken from the only rotation aligned variant of the Adience data set. Bottom: In-plane aligned samples. The left five image pairs show faces taken from the age group of (0-2) which are classified correctly under rotation alignment and are placed at least one age group above by the predictor under landmark-based alignment, with the middle image to the left being predicted as age group (8-13) by the GoogleNet. The in-plane alignment technique applied to one variant of the Adience data set tends to elongate faces vertically. The remaining image show misclassified and misaligned samples picked at random.

Figure 4. Heatmaps for **GoogleNet** models and **gender** recognition. Input images are shown above heatmaps for a DNN pretrained on Imagenet, which are shown above heatmaps for a DNN initialized randomly. The finetuned model predicts based on an ensemble of facial features, whereas the model starting with random weights has overfit on an isolated set of features characteristic to the target classes.

Figure 5. Heatmaps for **VGG-16** and **age** prediction. Input images are shown above heatmaps for a DNN pretrained on IMDB-WIKI, which are shown above heatmaps for a DNN pretrained on ImageNet.

	A	**C**	**G**
[i]	40.8 $_{75.4}$	40.3 $_{76.3}$	44.6 $_{80.8}$
[r]	46.9 $_{82.8}$	46.1 $_{82.5}$	46.4 $_{83.2}$
[i,n]	–	45.2 $_{82.02}$	49.4 $_{87.2}$
[r,n]	–	48.8 $_{84.9}$	53.6 $_{89.9}$

Table 4. Test set swapping results for **age** prediction. Performance is considerably worse when the *incorrect* preprocessing is used for testing, due to overfit feature sets. Pretraining can yield robust model parameters, compensating for the deviating test statistics.

	A	**C**	**G**
[i]	81.1	80.5	83.5
[r]	81.3	84.6	86.0
[i,n]	–	84.5	89.6
[r,n]	–	88.5	90.0

Table 5. Test set swapping results for **gender** prediction.

6. Conclusion

Recent deep neural network models are able to accurately analyze human face images, in particular recognize the persons' age, gender and emotional state. Due to their complex non-linear structure, however, these models often operate as black-boxes and until very recently it was unclear *why* they arrived at their predictions. In this paper we opened the black-box classifier using Layer-wise Relevance Propagation and investigated which facial features are actually used for age and gender prediction. We compared different image preprocessing, model initialization and architecture choices on the challenging Adience dataset and discussed how they affect performance. By using LRP to visualize the models' interactions with the given input samples, we demonstrate that appropriate model initialization via pretraining counteracts overfitting, leading to a holistic perception of the input. With a combination of simple preprocessing steps, we achieve state of the art performance for gender classification on the Adience benchmark data set.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] F. Arbabzadeh, G. Montavon, K.-R. Müller, and W. Samek. Identifying individual facial expressions by deconstructing a neural network. In *Pattern Recognition - 38th German Conference, GCPR 2016*, volume 9796 of *LNCS*, pages 344–354. Springer, 2016.

[3] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):e0130140, 2015.

[4] S. Bach, A. Binder, K.-R. Müller, and W. Samek. Controlling explanatory heatmap resolution and semantics via decomposition depth. In *Proc. of the IEEE International Conference on Image Processing (ICIP)*, pages 2271–2275, 2016.

[5] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller. How to explain individual classification decisions. *Journal of Machine Learning Research*, 11:1803–1831, 2010.

[6] S. Baluja, H. A. Rowley, et al. Boosting sex identification performance. *International Journal of Computer Vision*, 71(1):111–119, 2007.

[7] A. Dehghan, E. G. Ortiz, G. Shu, and S. Z. Masood. Dager: Deep age, gender and emotion recognition using convolutional neural network. *arXiv:1702.04280*, 2017.

[8] 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.

[9] S. Escalera, J. Fabian, P. Pardo, X. Baró, J. Gonzalez, H. J. Escalante, et al. Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results. In *Proc. of the IEEE International Conference on Computer Vision Workshops*, pages 1–9, 2015.

[10] V. Firoiu, W. F. Whitney, and J. B. Tenenbaum. Beating the world's best at super smash bros. with deep reinforcement learning. *arXiv:1702.06230*, 2017.

[11] A. Gallagher and T. Chen. Clothing cosegmentation for recognizing people. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[12] A. Gallagher and T. Chen. Understanding images of groups of people. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[13] F. Gao and H. Ai. Face age classification on consumer images with gabor feature and fuzzy lda method. *Advances in Biometrics*, pages 132–141, 2009.

[14] M. Gevrey, I. Dimopoulos, and S. Lek. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological Modelling*, 160(3):249–264, 2003.

[15] A. Graves, G. Wayne, M. Reynolds, T. Harley, I. Danihelka, A. Grabska-Barwińska, S. G. Colmenarejo, E. Grefenstette, T. Ramalho, J. Agapiou, et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476, 2016.

[16] G. Guo, G. Mu, Y. Fu, C. Dyer, and T. Huang. A study on automatic age estimation using a large database. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pages 1986–1991. IEEE, 2009.

[17] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4295–4304, 2015.

[18] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[19] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv:1408.5093*, 2014.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105, 2012.

[21] Y. H. Kwon et al. Age classification from facial images. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 762–767. IEEE, 1994.

[22] S. Lapuschkin, A. Binder, G. Montavon, K.-R. Müller, and W. Samek. Analyzing classifiers: Fisher vectors and deep neural networks. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2912–2920, 2016.

[23] S. Lapuschkin, A. Binder, G. Montavon, K.-R. Müller, and W. Samek. The lrp toolbox for artificial neural networks. *Journal of Machine Learning Research*, 17(1):3938–3942, 2016.

[24] G. Levi and T. Hassner. Age and gender classification using convolutional neural networks. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 34–42, 2015.

[25] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[26] G. Montavon, S. Bach, A. Binder, W. Samek, and K.-R. Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 2017.

[27] G. Montavon, K.-R. Müller, and M. L. Braun. Layer-wise analysis of deep networks with gaussian kernels. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1678–1686, 2010.

[28] G. Montavon, W. Samek, and K.-R. Müller. Methods for interpreting and understanding deep neural networks. *arXiv:1706.07979*, 2017.

[29] M. Moravčík, M. Schmid, N. Burch, V. Lisỳ, D. Morrill, N. Bard, et al. Deepstack: Expert-level artificial intelligence in no-limit poker. *arXiv:1701.01724*, 2017.

[30] A. J. O'Toole, T. Vetter, N. F. Troje, and H. H. Bülthoff. Sex classification is better with three-dimensional head structure than with image intensity information. *Perception*, 26(1):75–84, 1997.

[31] M. T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[32] 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*, pages 1–14, 2016.

[33] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[34] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 2017. in press.

[35] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, et al. Deep learning with convolutional neural networks for brain mapping and decoding of movement-related information from the human eeg. *arXiv:1703.05051*, 2017.

[36] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko. Quantum-chemical insights from deep tensor neural networks. *Nature Communications*, 8:13890, 2017.

[37] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[38] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv:1312.6034*, 2013.

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

[40] I. Sturm, S. Lapuschkin, W. Samek, and K.-R. Müller. Interpretable deep neural networks for single-trial eeg classification. *Journal of Neuroscience Methods*, 274:141–145, 2016.

[41] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv:1602.07261*, 2016.

[42] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[43] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv:1605.02688*, 2016.

[44] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Proc. of European Conference on Computer Vision (ECCV)*, pages 818–833. Springer, 2014.

[45] L. M. Zintgraf, T. S. Cohen, and M. Welling. A new method to visualize deep neural networks. *arXiv:1603.02518*, 2016.