Adaptive Variance Based Label Distribution Learning For Facial Age Estimation

Xin Wen[1,2], Biying Li[1,2], Haiyun Guo[1,3], Zhiwei Liu[1,2], Guosheng Hu[5], Ming Tang[1], and Jinqiao Wang[1,2,4]

[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3] ObjectEye Inc., Beijing, China
[4] NEXWISE Co., Ltd, Guangzhou, China
[5] Anyvision
{xin.wen,biying.li,haiyun.guo,zhiwei.liu,tangm,jqwang}@nlpr.ia.ac.cn, huguosheng100@gmail.com

Abstract. Estimating age from a single facial image is a classic and challenging topic in computer vision. One of its most intractable issues is label ambiguity, i.e., face images from adjacent age of the same person are often indistinguishable. Some existing methods adopt distribution learning to tackle this issue by exploiting the semantic correlation between age labels. Actually, most of them set a fixed value to the variance of Gaussian label distribution for all the images. However, the variance is closely related to the correlation between adjacent ages and should vary across ages and identities. To model a sample-specific variance, in this paper, we propose an adaptive variance based distribution learning (AVDL) method for facial age estimation. AVDL introduces the data-driven optimization framework, meta-learning, to achieve this. Specifically, AVDL performs a meta gradient descent step on the variable (i.e. variance) to minimize the loss on a clean unbiased validation set. By adaptively learning proper variance for each sample, our method can approximate the true age probability distribution more effectively. Extensive experiments on FG-NET and MORPH II datasets show the superiority of our proposed approach to the existing state-of-the-art methods.

Keywords: age estimation, distribution learning, meta-learning

1 Introduction

Age estimation is a challenging and hot research topic, which is to predict the person's age from his/her facial image. It has a lot of potential applications, including demographic statistics collection, commercial user management, video security surveillance, etc. However, there are numerous internal or external factors that affect the estimation results, including the race, illumination, image

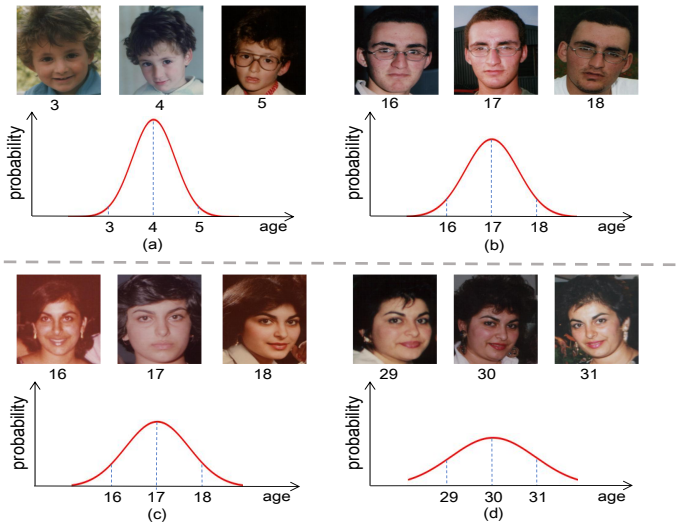

Fig. 1. The motivation of the proposed method. In each subfigure, the age probability distribution in the lower part corresponds to the middle image in the upper. The images above the dotted line belong to the same person and so do the images below the dotted line. On the one hand, by comparing (a) with (b) or (c) with (d), we can see that the facial appearance variation between adjacent ages of the same person varies at different ages. Correspondingly, the variance of the age probability distribution should differ across ages. On the other hand, by comparing (b) with (c), we can see that even at the same age, the aging process between different persons differs, thus the variance also varies across different persons

quality and so on. Besides, facial images from adjacent ages of the same person, especially for adults, usually look similar, resulting in the label ambiguity.

Recently, several deep learning methods have been proposed to improve the performance of facial age estimation. The most common methods model the face age prediction as a classification or a regression problem. The classification based methods treat each age as an independent class, which ignores the adjacent relationship between classes. Considering the continuity of age, regression methods predict age according to the extracted features. However, as presented by previous work [31, 33], the regression methods face the overfitting problem, which is caused by the randomness of the human aging process and the ambiguous mapping between facial appearance and the actual age. In addition, some ranking based methods are proposed to achieve more accurate age estimation. Those approaches make use of individuals' ordinal information and employ multiple binary classifiers to determine the final age of the input image. Furthermore, Geng et al. [13, 8] propose the label distribution learning (LDL) method which assumes that the real age can be represented by a discrete distribution. As their experiments show, it can help improve age estimation using Kullback-Leibler

(K-L) divergence to measure the similarity between the predicted and ground truth distribution.

For the label distribution learning methods, the mean of the distribution is the ground truth age. However, the variance of the distribution is usually unknown for a face image. The previous methods often treat variance as a hyper-parameter and simply set it to a fixed value for all images. We think these methods are suboptimal because the variance is highly related to the correlation between adjacent ages and should vary across different ages and different persons, as illustrated in Fig. 1. The assumption that all the images sharing the same variance potentially degrades the model performance.

To tackle the above issues, in this paper, we propose a novel adaptive variance based distribution learning method (AVDL) for age estimation. Specifically, we introduce meta-learning which utilizes validation set as meta-objective and is applicable to online hyper-parameter adaptation work [28], to model *sample-specific* variance and thus better approximate true age probability distribution. As Fig. 2 shows, we firstly select a small validation set. For each iteration, with a disturbing variable added to variance, we use K-L loss as the training loss to update the training model parameter. Then we share the updated parameter with validation model and use predicted expectation age and ground truth on validation set to get L1 loss as the meta-objective. With this meta-guider, the disturbing variable is updated by gradient descent and adaptively find the proper variance with which model could perform better on validation set. The main contributions of this work can be summarized as follows:

- We propose a novel adaptive variance based distribution learning (AVDL) method for facial age estimation. AVDL can effectively model the correlation between adjacent ages and better approximate the age label distribution.
- Unlike the existing deep models which assume the variance across ages and identities is the same, we introduce a data-driven method, meta-learning, to learn the *sample-specific* variance. To our knowledge, we are the first deep model using meta-learning method to adaptively learn different variances for different samples.
- Extensive experiments on FG-NET and MORPH II datasets show the superiority of our proposed approach to the existing state-of-the-art methods.

2 Related Work

2.1 Facial Age Estimation

In recent years, with rapid development of convolution neural network (CNN) in computer vision tasks, such as facial landmark detection[23], face reconition[38, 3], pedestrian attribute[35], semantic segmentation [46, 45], deep learning methods were also improved the performance of age estimation. Here we briefly review some representative works in the facial age estimation field. Dex et al. [30] regarded the facial age estimation as a classification problem and predicted ages with

the expectation of ages weighted by classification probability. Tan et al. [33] proposed an age group classification method called age group-n-encoding method. However, these classification methods ignored the adjacent relationship between classes or groups. To overcome this, Niu et al. [24] proposed a multiple output CNN learning algorithm which took account of the ordinal information of ages for estimation. Shen et al. [32] proposed Deep Regression Forests by extending differentiable decision trees to deal with regression. Furthermore, Li et al. [22] proposed BridgeNet, which consists of local regressors and gating networks, to effectively explore the continuous relationship between age labels. Tan et al. [34] proposed a complex Deep Hybrid-Aligned Architecture (DHAA) that consists of global, local and global-local branches and jointly optimized the architecture with complementary information. Besides, Xie et al. [39] proposed two ensemble learning methods both utilized ordinal regression modeling for age estimation.

2.2 Distribution Learning

Distribution learning is a learning method proposed to solve the problem of label ambiguity [10], which has been utilized in a number of recognition tasks, such as head pose estimation [12, 8], and age estimation [41, 20]. Geng et al. [13, 11] proposed two adaptive label distribution learning (ALDL) algorithms, i.e. IIS-ALDL and BFGS-ALDL, to iteratively learn the estimation function parameters and the label distributions variance. Though ALDL used an adaptive variance learning, our proposed method is different in three ways. Firstly, ALDL utilized traditional optimization method like BFGS while ours uses deep learning and CNN. Secondly, ALDL chose better samples in current training iteration to estimate new variance while our method uses meta-learning to get adaptive variance. The third point is ALDL updated variance only by estimating the training sample, which may cause overfitting. Our adaptive variance is supervised by validation set to be more general. Distribution learning of label was also used to remedy the shortage of training data with exact ages. Hou et al. [20] proposed a semi-supervised adaptive label distribution learning method. It used unlabeled data to enhance the label distribution adaptation to find a proper variance for each age. However, aging tendencies varies and variances of people at the same age could be different. Gao et al. [9] jointly used LDL and expectation regression to alleviate the inconsistency between training and testing. Moreover, Pan et al. [25] proposed a mean-variance loss for robust age estimation. Li et al. [21] proposed label distribution refinery to adaptively estimate the age distributions without assumptions about the form of label distribution, barely took into account the correlation of adjacent ages. While our method used Gaussian label distribution with adaptively meta-learned variance, which pays more attention to neighboring ages and ordinal information.

2.3 Meta-learning

Our proposed AVDL is an instantiation of meta-learning [36, 1], i.e., learning to learn. According to the type of leveraged meta data, this concept can be

classified to several types [37] including transferring knowledge from empirically similar tasks, transferring trained model parameters between tasks, building meta-models to learn data characteristics and learn purely from model evaluations. Model Agnostic Meta-Learning (MAML) [7] learned a model parameter initialization to perform better on target tasks. With the guidance of meta information, MAML took one gradient descent step on meta-objective to update model parameters [16]. The idea of using validation loss as meta-objective was applied in few-shot learning [27]. With reference to few-shot learning, Ren et al. [28] proposed a reweighting method (L2RW) guided by validation set. This method tried to solve the problem that data imbalance and label noise are both in the training set. The crucial criteria of L2RW is a small unbiased clean validation set which was taken as the supervisor of learning sample weight. As validation set performance measures the quality of hyper-parameters, taking it as meta-objective could not only be applied to sample reweighting but also to any other online hyper-parameter adaptation tasks. Inspired by this, we propose AVDL to incorporate validation set based meta-learning and label distribution learning to adaptively learn the label variance.

3 Methodology

In this section, we firstly give a description of the label distribution learning (LDL) method in age estimation. Then we introduce our adaptive variance based distribution learning(AVDL) method based on meta-learning framework.

3.1 The Label Distribution Learning Problem Revisit

Let X denote an input image with ground truth label y, $y \in \{0, 1, ..., 100\}$. The model is trained to predict a value as close to the ground truth label as possible. For traditional age estimation method, the ground truth is an integer. While in LDL method, to express the ambiguity of labels, Gao et al. [8] transform the real value y to a normal distribution $\mathbf{p}(y, \sigma)$ to denote the new ground truth. Mean value is set to the ground truth label y and σ is the normal distribution variance. Here we adopt the boldface lowercase letters like $\mathbf{p}(y, \sigma)$ to denote vectors, and use $p_k(y, \sigma)$ $(k \in [0, 100])$ to represent the k-th element of $\mathbf{p}(y, \sigma)$:

$$p_k(y, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(k-y)^2}{2\sigma^2}) \tag{1}$$

where p_k is the probability that the true age is k years old. It represents the connection between the class k and y in a normal distribution view.

In the training process, assuming $G(*, \theta)$ as the classification function of the trained estimation model, θ represents the model parameters, $\mathbf{z}(X, \theta) = G(X, \theta)$ transforms the input X to the classification vector $\mathbf{z}(X, \theta)$. A softmax function is utilized to transfer $\mathbf{z}(X, \theta)$ into a probability distribution $\hat{\mathbf{p}}(X, \theta)$, the k-th

element of which can be denoted by:

$$\hat{p}_k(X,\theta) = \frac{\exp(z_k(X,\theta))}{\sum_n \exp(z_n(X,\theta))} \tag{2}$$

where $z_k(X,\theta)$ is the k-th element of $\mathbf{z}(X,\theta)$.

LDL tries to generate the predicted softmax probability distribution as similar to the ground truth distribution as possible. So the Kullback-Leibler (K-L) divergence is employed to measure the difference between the predicted distribution and ground-truth distribution [8]:

$$L_{KL}(X,y,\theta,\sigma) = \sum_k p_k(y,\sigma) ln \frac{p_k(y,\sigma)}{\hat{p}_k(X,\theta)} \tag{3}$$

Then the K-L loss is used to update model parameters with SGD optimizer.

LDL method aims to construct a normal distribution of ground truth to approximate the real distribution, the key of which is the variance σ. For most LDL methods, this hyper-parameter is simply set to a fixed value, 2.0 in most cases. However, in fact, the variances for different people, or people of different ages could not be absolutely the same. So we propose a method to search proper variance for each image.

3.2 Adaptive Distribution Learning Based on Meta-learning

In machine learning, the loss on validation set is one of the guiders to adjust hyper-parameters for generalization. Therefore, using a clean unbiased validation set can help train a more general model. However, traditional training mode usually tunes the hyper-parameter manually. Inspired by the meta-learning work [28], we propose the adaptive variance based distribution learning (AVDL) algorithm guided by validation set, which offers an effective strategy to learn the sample-specific variance.

As we mentioned in Section 3.1, the most important hyper-parameter of LDL is the variance σ. Because our goal is to search for proper σ of each image while training, in this section we use σ to represent the variance vector for a batch of training data. The optimal σ in each iteration depends on the optimal model parameter θ:

$$\theta^*(\sigma) = \arg\min_{\theta} L_{KL}(X_{tr}, y_{tr}, \theta, \sigma) \tag{4}$$

$$\sigma^* = \arg\min_{\sigma,\sigma\geq 0} L_1(X_{val}, y_{val}, \theta^*, \sigma) \tag{5}$$

where $L_1(X_{val}, y_{val}, \theta^*, \sigma)$ denotes the validation loss. X_{tr} is the training input image while y_{tr} is its label. X_{val} is the validation input image while its label is y_{val}. To solve the optimization problem, we divided the training process into several process. Fig. 2 shows the computation graph of our proposed method.

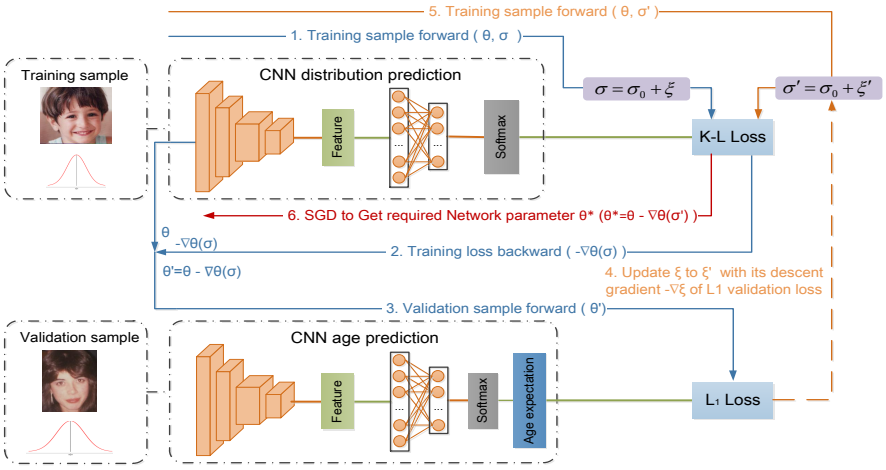

Fig. 2. Computation graph of AVDL in one iteration. The ground truth of each input image is transformed to a normal distribution. The model on top is for training and the other is for validation. The train model and validation model share the network architecture and parameters. The training loss is K-L loss while the validation loss is L1 loss. Process 1,2,3 belongs to traditional training steps. Perturbing variable ξ is added to initial distribution variance to get variance σ. By adding the training gradient descent step $-\bigtriangledown \theta$, the training model parameter θ is updated to θ' and is assigned to the validation model. Process 4 uses the descent gradient of ξ in validation loss to get the modified ξ' and σ'. Process 5,6 shows the improved forward and backward computation with a proper variance σ'.

We choose a fixed number of images with correct labels from each class in the training set n to make a small unbiased validation set with m images, $m \ll n$. We utilize σ_i to denote variance for i-th image image while we set the initial value of variances of all images to a fixed value σ_{i0}. To search a proper variance, we perturb each σ_i by ξ_i:

$$\sigma_i = \sigma_{i0} + \xi_i \tag{6}$$

where ξ_i is the i-th component of perturbing vector ξ which is set to 0 for initialization. Clearly, searching a proper σ is equal to searching a proper ξ.

Firstly, as Fig. 2 process 1, 2 and 3 show, in the t-th iteration, the input training batch calculates K-L loss as described in Section 3.1 with a perturbed σ. Update the model parameter θ_t with SGD to get $\hat{\theta}_{t+1}$:

$$\hat{\theta}_{t+1} = \theta_t - \alpha \bigtriangledown_\theta L_{KL}(X_{tr}, y_{tr}, \theta_t, \sigma) \tag{7}$$

α is the descent step.

The training loss is related to distribution. To compensate the lack of constrain in the final predicted age value, we adopt L1 loss on validation to measure

the distance between expectation age of prediction and the validation ground truth [9]:

$$L_1(X_{val}, y_{val}, \hat{\theta}_{t+1}, \xi) = \left| \hat{y}^*(X_{val}, \hat{\theta}_{t+1}, \xi) - y_{val} \right| \tag{8}$$

$$\hat{y}^*(X_{val}, \hat{\theta}_{t+1}, \xi) = \sum\nolimits_k \hat{p}_k(X_{val}, \hat{\theta}_{t+1}, \xi) l_k \tag{9}$$

where $\hat{p}_k$ is the k-th element in the prediction vector of validation input X_{val} and l_k denotes the age value of the k-th class, i.e. $l_k \in \mathcal{Y}$. The expectation age computing method is also used for estimating test images in Section 4.

Algorithm 1 Adaptive Variance Based Distribution Learning

Input: Training set $\mathcal{S}_{tr} = X_{tr}, y_{tr_n}$; Validation set $\mathcal{S}_{val} = X_{val}, y_{val_m}$, $m \ll n$; Initial model parameter θ_0; Initial variance σ_0

Output: Final model parameter θ_T

1: **for** t = 0,1,2...T-1 **do**
2: Sample training batch $S_{tr,t}$ from S_{tr}
3: Sample validation batch $S_{val,t}$ from S_{val}
4: $\xi \leftarrow 0$
5: $\sigma \leftarrow \sigma_0 + \xi$
6: $L_{KL}(X_{tr}, y_{tr}, \theta_t, \sigma) \leftarrow \text{NetForward}(X_{tr}, y_{tr}, \theta_t, \sigma)$
7: $\hat{\theta}_{t+1}(\sigma) \leftarrow \theta_t - \alpha \bigtriangledown_{\theta_t} L_{KL}(X_{tr}, y_{tr}, \theta_t, \sigma)$ % $\hat{\theta}_{t+1}$ is a function of σ
8: $L_1(X_{val}, y_{val}, \hat{\theta}_{t+1}(\sigma), \sigma) \leftarrow \text{NetForward}(X_{val}, y_{val}, \hat{\theta}_{t+1}(\sigma), \sigma)$
9: $\hat{\xi} \leftarrow \xi - \beta \bigtriangledown_{\xi} L_1(X_{val}, y_{val}, \hat{\theta}_{t+1}(\sigma), \sigma)$ % the gradient of ξ on $L1$ loss equals to the gradient of σ on $L1$ loss
10: $\hat{\sigma} \leftarrow \sigma_0 + \hat{\xi}$ % modify σ adaptively
11: $\hat{L}_{KL}(X_{tr}, y_{tr}, \theta_t, \hat{\sigma}) \leftarrow \text{Forward}(X_{tr}, y_{tr}, \theta_t, \hat{\sigma})$
12: $\theta_{t+1} \leftarrow SGD(\hat{L}_{KL}(X_{tr}, y_{tr}, \theta_t, \hat{\sigma}), \theta_t)$ % update with SGD optimizer
13: **end for**

The better hyper-parameter means better validation performance. In that, we update the perturbation ξ with gradient descent step:

$$\hat{\xi} = \xi - \beta \bigtriangledown_{\xi} L_1(X_{val}, y_{val}, \hat{\theta}_{t+1}, \xi) \tag{10}$$

where β is the descent step size. This step is corresponding to the process 4 in Fig. 2. Due to the non-negativity restriction of σ, we normalize the ξ into the range [-1,1], using the mapping $\xi_i \rightarrow \frac{2\xi_i - \max(\xi) - \min(\xi)}{\max(\xi) - \min(\xi)}$. Then update the variance σ according to Eq.(6). In the third step of training, with the modified variance, we calculate forward K-L loss of the training input, then update model parameter with SGD optimizer, as the process 5,6 in Fig. 2 shows.

We listed step-by-step pseudo code in Algorithm 1. According to step 9 in Algorithm 1, there is a two-stage deviation computation of variable ξ. PyTorch autograd mechanism can achieve this operation handily.

4 Experiments

In this section, we first introduce the datasets used in the experiments, i.e., MORPH II [29], FG-NET [26] and IMDB-WIKI [31]. Then we detail the experiment settings. Next, we validate the superiority of our approach with comparisons to the state-of-the-art facial age estimation methods. Finally, we conduct some ablation studies on our method.

4.1 Datasets

Morph II is the most popular dataset for age estimation. The dataset contains 55,134 color facial images of 13,000 individuals whose ages range from 16 to 77. On this dataset, we employ three typical protocols for evaluation: **Setting I: 80-20 protocol.** We randomly divide this dataset into two non-overlapped parts, i.e., 80% for training and 20% for testing. **Setting II: Partial 80-20 protocol.** Following the experimental setting in [33], we extract a subset of 5,493 facial images from Caucasian descent, these images are splitted into two parts: 80% of facial images for training and 20% for testing. **Setting III: S1-S2-S3 protocol.** Similar to [33, 22], Morph II dataset is splitted into three non-overlapped subsets S1, S2 and S3, and all experiments are repeated twice. Firstly, train on S1 and test on S2+S3. Then, train on S2 and test on S1+S3. The performance of the two experiments and their average MAE are shown respectively.

FG-NET contains 1,002 color or gray facial images of 82 individuals whose ages are ranging from 0 to 69. We follow a widely used leave-one-person-out (LOPO) protocol [25, 4] in our experiments, and the average performance over 82 splits is reported.

IMDB-WIKI is the largest facial image dataset with age labels, which consists of 523,051 images in total. This dataset is constituted of two parts: IMDB (460,723 images) and WIKI (62,328 images). We follow the practice in [22] and use this dataset to pretrain our model. Specifically, We remove non-face images and partial multi-face images. Finally, about 270,000 images are reserved.

4.2 Implementation Details

We use the detection algorithm in [44] to obtain the face detection box and five facial landmark coordinates, which are then used to align the input facial image of the network. We resize the input image to 224×224.

Following the settings in [9], we augment the face images with random horizontal flipping, scaling, rotating and translating during training time. For testing, we input both the image and its flipped version to the network, and then average their predictions as the final results.

We adopt ResNet-18 [19] as our backbone network and pretrain the network on IMDB-WIKI dataset for better initialization. We use the SGD optimizer with

batch size 32 to optimize the network. The weight decay and the momentum are set to 0.0005 and 0.9. The initial learning rate is set to 0.01 and decays by 0.1 for every 20 epochs. we set the initial value of variances of all images to 1, and train the deep convolution neural network with PyTorch on 4 GTX TITAN X GPUs.

Table 1. The comparisons between the proposed method and other state-of-the-art methods on MORPH II under Setting I. Bold indicates the best (* indicates the model was pre-trained on the IMDB-WIKI dataset; † indicates the model was pre-trained on the MS-Celeb-1M dataset [17])

Method	MAE	Parameters	Year
ORCNN [24]	3.27	479.7K	2016
RGAN* [6]	2.61	-	2017
VGG-16 CNN + LDAE* [2]	2.35	138M	2017
SSR-Net* [40]	3.16	40.9K	2018
DRFs[32]	2.17	138M	2018
M-V Loss* [25]	2.16	138M	2018
DLDL-V2† [9]	1.97	3.7M	2018
C3AE* [43]	2.75	39.7K	2019
DHAA [34]	**1.91**	100M	2019
AVDL*	1.94	11M	-

4.3 Evaluation Criteria

According to previous works [31, 33], we measure the performance of age estimation by the Mean Absolute Error (MAE) which is calculated using the average of the absolute errors between estimated age and chronological age.

4.4 Comparisons With State-of-the-arts

On Morph II. We first compare the proposed method with other state-of-the-art methods on MORPH II dataset in Setting I, as illustrated in Table 1. We achieve the second best performance, which is slightly lower than DHAA [34] by 0.03. It is worth to note that DHAA is large and complex, their parameters are around 10 times larger than ours, though without additional face dataset for pre-training. Moreover, using the same pre-training dataset, we surpass the M-V Loss by a significant margin of 0.22.

Table 2 shows the test result under Setting II. We achieve the best performance, which is slightly higher than BridgeNet [22] by 0.01. Nevertheless, we have fewer parameters than BridgeNet. That is to say, we achieve the performance nearly to theirs with a significantly lower model complexity at the same

time. As Table 3 shows, we achieve MAE of 2.53 under Setting III. Our method performs much better than the current state-of-the-art. All of the above comparisons consistently demonstrate the effectiveness of the proposed method.

Table 2. The comparisons between the proposed method and other state-of-the-art methods on MORPH II dataset (Setting II) and FG-NET dataset. Bold indicates the best (* indicates the model was pre-trained on the IMDB-WIKI dataset)

Method	MORPH II	FG-NET	Parameters	Year
OHRANK [4]	6.07	4.48	-	2011
CA-SVR [5]	5.88	4.67	-	2013
Human [18]	6.30	4.70	-	2015
DEX* [31]	2.68	3.09	138M	2018
DRFs [32]	2.91	3.85	138M	2018
M-V Loss* [25]	-	2.68	138M	2018
AGEn* [33]	2.52	2.96	138M	2018
C3AE* [43]	-	2.95	39.7K	2019
BridgeNet* [22]	2.38	2.56	138M	2019
DHAA* [34]	-	2.59	100M	2019
AVDL*	**2.37**	**2.32**	11M	-

On FG-NET. As shown in Table 2, we compare our model with state-of-the-art models on FG-Net. Our method achieves the lowest MAE of 2.32, which improves the state-of-the-art performance by a large margin of 0.24. Experimental results show that our method is effective even when there are only a few training images.

4.5 Ablation Study

In this subsection, we conduct ablation study on MORPH II dataset under Setting I to conduct ablation study.

The superiority of adaptive variance to fixed variance value. We train a set of baseline models, which all adopt ResNet-18 and K-L divergence loss but with different fixed variance values. Theoretically, the larger variance indicates the smoother distribution which refers to the stronger correlation in that age group. In comparison, the smaller variance represents the sharper distribution and the weaker correlation. If the variance is set too high, i.e., the label distribution is too smooth, the age estimation may not perform well. As Fig. 3 shows, the MAE increases along with the growth of variance when it is higher than 3, which indicates the worse performance. When the variance reduces to 0, it assumes there is no correlation between ages which is similar to the assumption when regarding age estimation as classification problem. However, considering

Table 3. The comparisons between the proposed method and other state-of-the-art methods on MORPH II under Setting III. Bold indicates the best (* indicates the model was pre-trained on the IMDB-WIKI dataset)

Method	Train	Test	MAE	Avg
KPLS [14]	S1	S2+S3	4.21	4.18
	S2	S1+S3	4.15	
BIF+KCCA [15]	S1	S2+S3	4.00	3.98
	S2	S1+S3	3.95	
CPLF [42]	S1	S2+S3	3.72	3.63
	S2	S1+S3	3.54	
DRFs [32]	S1	S2+S3	-	2.98
	S2	S1+S3	-	
DOEL [39]	S1	S2+S3	-	2.75
	S2	S1+S3	-	
AGEn* [33]	S1	S2+S3	2.82	2.70
	S2	S1+S3	2.58	
BridegNet* [22]	S1	S2+S3	2.74	2.63
	S2	S1+S3	2.51	
AVDL*	S1	S2+S3	**2.64**	**2.53**
	S2	S1+S3	**2.41**	

the gradual changing of face in aging, taking a proper use of age correlation can help age estimation. As illustrated in Fig. 3, when the fixed variance is less than 3, the MAE fluctuates. It validates that setting a fixed variance is suboptimal because the age correlation can not be the same for different people in different ages. The best performance of baseline is achieved with a variance of 3. However, it is still much worse than our proposed method, AVDL. In Fig. 3, we also show the performance achieved by training the ResNet-18 with cross-entropy loss, which is our baseline method by treating age estimation as classification task. In summary, Fig 3 demonstrates the superiority of the adaptive variance. Actually, for each dataset and experiment setting, our approach is compared to the baseline method with fixed variance. We observed from Fig 3 that the variation in fixed variance value within a certain range has little impact on performance, due to limited time, we only search the best variance for MORPH II(Setting I) and apply it to other experiments. In addition, the baseline with the fixed variance of MORPH II(Setting II), MORPH II(Setting III), and FGNET are 2.66, 2.79,2.64, respectively.

The influence of different sample number in validation set. As [28] shows, a balanced meta dataset could provide balanced class knowledge. For the same purpose, we choose an unbiased validation set as meta dataset. As for the composition of the clean validation set, we try different sizes of validation set. We respectively random select 1, 2 and 3 images from each class in the training set to form the validation set for experiment. From Table 4, we can find that

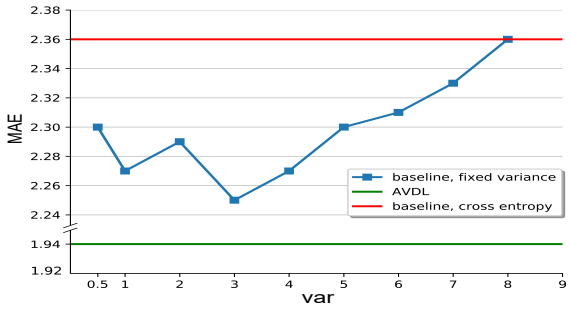

Fig. 3. The MAE results on MORPH II under Setting I. The blue line denotes the results of the baseline model trained with different fixed variance. The red line is the result of the baseline model trained with cross-entropy loss and the green one is the result of AVDL

Table 4. The performance comparison on selecting different number of facial images of each age to form validation set

Number of images	1	2	3
MAE of AVDL	1.98	1.96	1.94

the larger the validation set is, the better the model performs. However, since all validation set is used in each iteration, it needs more time and memory as the size of validation set increases. Considering the time and space cost, for each dataset setting, we randomly chose three image from each class to form the validation set.

4.6 Visualization and Discussion

Considering the affordance and credibility, here we display some visual results of AVDL in age estimation and variance adaptation.

We use the learned variance of samples to show the effectiveness of AVDL and to justify our motivation. Under the Setting I on MORPH II, each age, ranging from 16 to 60, possesses a group of face images belonging to different person identities. While there is no person whose images covering the full age range. We select images of several persons at different ages with their adapted variances in Fig. 4(a). As [11] mentioned, the age variances of younger or older people tend to be smaller than those of middle age. And the variances vary between people in the same age group. Besides, the variance in Fig. 4(b) shows the visualization of the adjusted variances in a mini-batch. The initial variance for each sample, as indicated in Section 4.2, is set to 1. The learned variances are shown in the horizontal band above in which each block represents a sample and the color of the block indicates the magnitude of variance. The blocks are arranged from left to right according to the ages of samples. The band below is the legend which indicates the relationship between the magnitude of variance

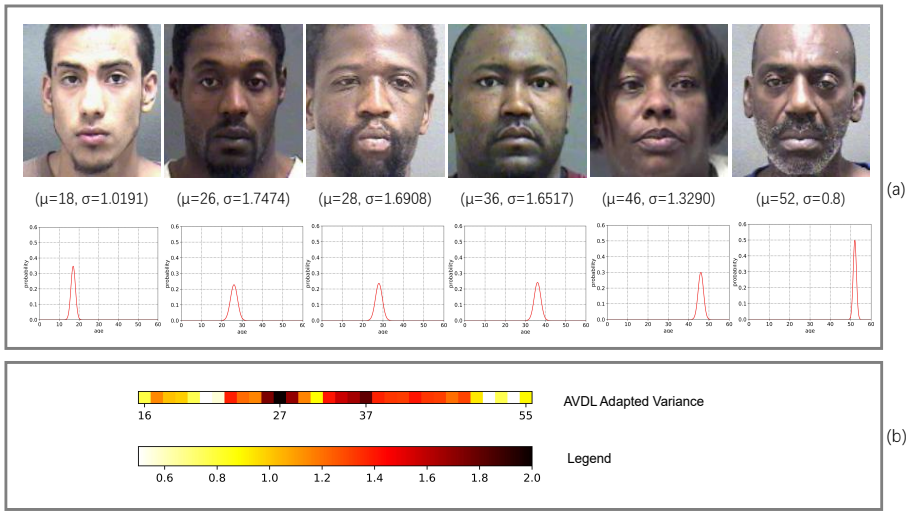

Fig. 4. Examples of age estimation results by AVDL. (a) shows some samples at different ages on Morph II with adapted variances. According to the Gaussian curves, it can be proved that for younger and older people, the variances tend to be smaller while for middle age, larger. (b) uses heat map to visualize the adaptively learned variances σ corresponding to different ages.

and the color of block. Same as Fig. 4(a) shows, the variance in young age and old age is smaller. Besides, the variances in the band fluctuate slightly which demonstrates the variance is different for people.

5 Conclusions

In this paper, we propose a novel method for age estimation, named adaptive variance based distribution learning(AVDL). AVDL introduces meta-learning to adaptively adjust the variance for each image in single iteration. It achieves better performances than others on multiple age estimation datasets. Our experiments also show that AVDL can guide variance to get close to real facial aging law. The idea that using meta-learning to guide key hyper-parameters is inspirational and we will explore more possibilities of it.

Acknowledgments

This work was supported by Key-Area Research and Development Program of Guangdong Province (No.2019B010153001), National Natural Science Foundation of China (No.61772527,61806200,61976210), China Postdoctoral science Foundation(No.2019M660859), Open Project of Key Laboratory of Ministry of Public Security for Road Traffic Safety (No.2020ZDSYSKFKT04).

References

1. Andrychowicz, M., Denil, M., Gomez, S., Hoffman, M.W., Pfau, D., Schaul, T., Shillingford, B., De Freitas, N.: Learning to learn by gradient descent by gradient descent. In: Advances in neural information processing systems. pp. 3981–3989 (2016)
2. Antipov, G., Baccouche, M., Berrani, S.A., Dugelay, J.L.: Effective training of convolutional neural networks for face-based gender and age prediction. Pattern Recognition **72**, 15–26 (2017)
3. Cao, D., Zhu, X., Huang, X., Guo, J., Lei, Z.: Domain balancing: Face recognition on long-tailed domains. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5671–5679 (2020)
4. Chang, K.Y., Chen, C.S., Hung, Y.P.: Ordinal hyperplanes ranker with cost sensitivities for age estimation. In: Computer vision and pattern recognition (cvpr), 2011 ieee conference on. pp. 585–592. IEEE (2011)
5. Chen, K., Gong, S., Xiang, T., Change Loy, C.: Cumulative attribute space for age and crowd density estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2467–2474 (2013)
6. Duan, M., Li, K., Li, K.: An ensemble cnn2elm for age estimation. IEEE Transactions on Information Forensics and Security **13**(3), 758–772 (2017)
7. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 1126–1135. JMLR. org (2017)
8. Gao, B.B., Xing, C., Xie, C.W., Wu, J., Geng, X.: Deep label distribution learning with label ambiguity. IEEE Transactions on Image ProceRing **26**(6), 2825–2838 (2017)
9. Gao, B.B., Zhou, H.Y., Wu, J., Geng, X.: Age estimation using expectation of label distribution learning. In: IJCAI. pp. 712–718 (2018)
10. Geng, X.: Label distribution learning. IEEE Transactions on Knowledge and Data Engineering **28**(7), 1734–1748 (2016)
11. Geng, X., Wang, Q., Xia, Y.: Facial age estimation by adaptive label distribution learning. In: 2014 22nd International Conference on Pattern Recognition. pp. 4465–4470. IEEE (2014)
12. Geng, X., Xia, Y.: Head pose estimation based on multivariate label distribution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1837–1842 (2014)
13. Geng, X., Yin, C., Zhou, Z.H.: Facial age estimation by learning from label distributions. IEEE transactions on pattern analysis and machine intelligence **35**(10), 2401–2412 (2013)
14. Guo, G., Mu, G.: Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. In: CVPR 2011. pp. 657–664. IEEE (2011)
15. Guo, G., Mu, G.: Joint estimation of age, gender and ethnicity: Cca vs. pls. In: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). pp. 1–6. IEEE (2013)
16. Guo, J., Zhu, X., Zhao, C., Cao, D., Lei, Z., Li, S.Z.: Learning meta face recognition in unseen domains. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6163–6172 (2020)
17. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: European conference on computer vision. pp. 87–102. Springer (2016)

18. Han, H., Otto, C., Liu, X., Jain, A.K.: Demographic estimation from face images: Human vs. machine performance. IEEE transactions on pattern analysis and machine intelligence **37**(6), 1148–1161 (2014)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
20. Hou, P., Geng, X., Huo, Z.W., Lv, J.Q.: Semi-supervised adaptive label distribution learning for facial age estimation. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
21. Li, P., Hu, Y., Wu, X., He, R., Sun, Z.: Deep label refinement for age estimation. Pattern Recognition **100**, 107178 (2020)
22. Li, W., Lu, J., Feng, J., Xu, C., Zhou, J., Tian, Q.: Bridgenet: A continuity-aware probabilistic network for age estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1145–1154 (2019)
23. Liu, Z., Zhu, X., Hu, G., Guo, H., Tang, M., Lei, Z., Robertson, N.M., Wang, J.: Semantic alignment: Finding semantically consistent ground-truth for facial landmark detection. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3462–3471 (2019)
24. Niu, Z., Zhou, M., Wang, L., Gao, X., Hua, G.: Ordinal regression with multiple output cnn for age estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4920–4928 (2016)
25. Pan, H., Han, H., Shan, S., Chen, X.: Mean-variance loss for deep age estimation from a face. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5285–5294 (2018)
26. Panis, G., Lanitis, A., Tsapatsoulis, N., Cootes, T.F.: Overview of research on facial ageing using the fg-net ageing database. Iet Biometrics **5**(2), 37–46 (2016)
27. Ren, M., Triantafillou, E., Ravi, S., Snell, J., Swersky, K., Tenenbaum, J.B., Larochelle, H., Zemel, R.S.: Meta-learning for semi-supervised few-shot classification. arXiv preprint arXiv:1803.00676 (2018)
28. Ren, M., Zeng, W., Yang, B., Urtasun, R.: Learning to reweight examples for robust deep learning. arXiv preprint arXiv:1803.09050 (2018)
29. Ricanek, K., Tesafaye, T.: Morph: A longitudinal image database of normal adult age-progression. In: 7th International Conference on Automatic Face and Gesture Recognition (FGR06). pp. 341–345. IEEE (2006)
30. Rothe, R., Timofte, R., Van Gool, L.: Dex: Deep expectation of apparent age from a single image. In: The IEEE International Conference on Computer Vision (ICCV) Workshops (December 2015)
31. Rothe, R., Timofte, R., Van Gool, L.: Deep expectation of real and apparent age from a single image without facial landmarks. International Journal of Computer Vision **126**(2-4), 144–157 (2018)
32. Shen, W., Guo, Y., Wang, Y., Zhao, K., Wang, B., Yuille, A.L.: Deep regression forests for age estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2304–2313 (2018)
33. Tan, Z., Wan, J., Lei, Z., Zhi, R., Guo, G., Li, S.Z.: Efficient group-n encoding and decoding for facial age estimation. IEEE transactions on pattern analysis and machine intelligence **40**(11), 2610–2623 (2017)
34. Tan, Z., Yang, Y., Wan, J., Guo, G., Li, S.Z.: Deeply-learned hybrid representations for facial age estimation. In: IJCAI. pp. 3548–3554 (2019)
35. Tan, Z., Yang, Y., Wan, J., Wan, H., Guo, G., Li, S.: Attention-based pedestrian attribute analysis. IEEE Transactions on Image Processing **PP**, 1–1 (07 2019). https://doi.org/10.1109/TIP.2019.2919199

36. Thrun, S., Pratt, L.: Learning to learn. Springer Science & Business Media (2012)
37. Vanschoren, J.: Meta-learning: A survey. arXiv preprint arXiv:1810.03548 (2018)
38. Wang, G., Han, H., Shan, S., Chen, X.: Cross-domain face presentation attack detection via multi-domain disentangled representation learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
39. Xie, J.C., Pun, C.M.: Deep and ordinal ensemble learning for human age estimation from facial images. IEEE Transactions on Information Forensics and Security **15**, 2361–2374 (2020)
40. Yang, T.Y., Huang, Y.H., Lin, Y.Y., Hsiu, P.C., Chuang, Y.Y.: Ssr-net: A compact soft stagewise regression network for age estimation. In: IJCAI. vol. 5, p. 7 (2018)
41. Yang, X., Gao, B.B., Xing, C., Huo, Z.W., Wei, X.S., Zhou, Y., Wu, J., Geng, X.: Deep label distribution learning for apparent age estimation. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 102–108 (2015)
42. Yi, D., Lei, Z., Li, S.Z.: Age estimation by multi-scale convolutional network. In: Asian conference on computer vision. pp. 144–158. Springer (2014)
43. Zhang, C., Liu, S., Xu, X., Zhu, C.: C3ae: Exploring the limits of compact model for age estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 12587–12596 (2019)
44. Zhao, X., Liang, X., Zhao, C., Tang, M., Wang, J.: Real-time multi-scale face detector on embedded devices. Sensors **19**(9), 2158 (2019)
45. Zhu, B., Chen, Y., Tang, M., Wang, J.: Progressive cognitive human parsing. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
46. Zhu, B., Chen, Y., Wang, J., Liu, S., Zhang, B., Tang, M.: Fast deep matting for portrait animation on mobile phone. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 297–305 (2017)