Deans' Undergraduate Research Fund 2020

Project Report

Facial Age Estimation with Deep Neural Network

Yijian Liu

supervised by
Guo Li

Abstract

Age estimation aims to produce an accurate age label for a given facial image. It has been an important task in that it can be employed in many applications but the age progression varies from person to person, making it especially challenging. In this paper, we propose a combination of SE-ResNeXt-50 and label distribution learning (LDL), preserving the ordinal information between age labels and ensuring sufficient training for all age groups. As a result, we achieve decent performance on MORPH among some of the state-of-the-art methods, and best performance on MegaAge-Asian to our knowledge.

1 Introduction

Human faces have always been an important topic in computer vision. Faces are relatively easy to identify and contain many biological features that can reveal a person's identity, gender, ethnicity, and other background information. Specifically, facial age estimation refers to the task that labeling a given facial image with an exact age value or categorizing such an image into an age group, which can then be used in various applications such as video surveillance, customer profiling, human-computer interaction, Internet security, as well as facilitating age progression research and biometric identification.

Facial age estimation remains a challenging task by nature nowadays due to its granularity and high intragroup variance. Unlike face detection, age estimation is more fine-grained than searching for a common pattern of faces across all human beings. But it also needs to selectively neglect the features that are more personal and cannot maximize the difference between different age groups. On top of that, age progression is not linear and can vary significantly among the same age group. From the perspective of research, the difficulty of gathering a large amount of accurate age data leads to the rarity of a large, high-quality face-age dataset.

To fully exploit the characteristics of age estimation, Deep Label Distribution Learning (DLDL) is purposed in [1]. It uses a normal distribution centered at the ground-truth as the label. The goal of training is to learn to predict a probability distribution with a mean $\mu =$ age from a facial image. Kullback–Leibler divergence is employed as a loss function to reduce the distance between the ground-truth age distribution and the predicted distribution. An improved version, DLDL-v2, is purposed in [2], achieving state-of-the-art performance on the MORPH dataset. It adds an L1 loss to the original setting to minimize the possible inconsistency between the distribution and the final regressed value. We adopt the idea of DLDL-v2 based on its great performance in our experiments. In this paper, we refer to the setting of DLDL-v2 simply as label distribution learning or LDL.

With the development of convolutional neural network (CNN), it is witnessed that features extracted by CNN yields more satisfactory results when dealing with large datasets than conventional hand-designed features. And it is often witnessed that a new network architecture can boost performance and reduce the difficulty of training. Compared to the classic Architectures, such as VGG-16 (with batch normalization) and ResNet-50, SE-ResNeXt-50 (32×4d) is shown to be better at the task in the setting of LDL.

In this way, our best practice can be described as a combination of label distribution learning strategy and a SE-ResNeXt-50 backbone. We further show that the performance of the model can be improved through data augmentation and multi-task learning.

2 Related Work

Due to the ordinal nature of the age estimation problem, much effort has been devoted to how this ordinal relationship can be preserved across age classes and training data. OR-CNN [3] uses the whole training dataset to train a series of binary classifiers, all of which are then summed up to form the regressed value. Ranking-CNN [4] further investigates this idea and proposes to use separate CNNs to extract features of different ages. Compared to simple regression, DEX [5] calculates the expectation of the softmax probability. Importantly, a preprocessing pipeline is proposed in [5], including a 40% expansion of the margin area when cropping out the face from the original image, which is adopted in our preprocessing stage. DAG-CNN proposed in [6] is also an expectation-based method. It extracts features from multiple stages in the network and performs score-level fusion to get the final estimation. This allows some discriminative features in shallow layers to be considered, which inspires us to construct the multi-task model based on the existing model.

Codes accompanying this paper are available at `https://github.com/yjl450/age-estimation-ldl-pytorch`.

Recently, probability distribution learning has been popular. In [7], an encoding method call Label Distribution Age Encoding (LDAE), which is also used in DLDL[1] and DLDL-v2 [2]. It is proposed in [7] that multi-task learning of gender has a positive effect on the task of age estimation. This idea is later adopted as a multi-task model for gender, ethnicity, and age prediction. In [1], Kullback-Leibler divergence loss is employed to evaluate and minimize the difference between the ground-truth label distribution and the predicted label distribution. However, there is the problem that the goal of optimization is different from the goal of output, which is addresses in DLDL-v2 [2] by adding a $L1$ loss to minimize the error between the regressed expectation and the ground-truth age. We will follow the setting of DLDL-v2 in our experiments.

Furthermore, [8] proposed mean-variance loss and [9] proposed Label Refinery Network (LRN), both of which aim to adaptively learn to shape the label distribution to better fit the varying rate of age progression. Compared to these methods, our method is more of a stationary method that variance σ, which controls the shape of the distribution, is a hyperparameter.

3 Our Approach

3.1 Problem Formulation

Age estimation is the task that, for each given input of facial image, produces an estimated age value or predicts the corresponding age group. Here, we focus on estimating the age value. We define $X = \{x_n | n = 1, 2, 3, \cdots, N\}$ as the set of facial images for training, $Y = \{y_n \in K\}$ as the corresponding set of ground-truth ages, where N is the total number of samples and K is the range of all possible ages, which is also the set of all age classes. The goal of age estimation is to find a function F to minimize the error between the predicted age, denoted by $\hat{y_n} = F(x_n)$, and the ground-truth age, that it, to minimize $|\hat{y_n} - y_n|$.

3.2 Label Distribution Learning

Label distribution learning takes age estimation as a problem of learning a mapping from an image to a probability distribution of age. We follow the setting of DLDL-v2 in [2]. We will omit the indices of samples for simplicity. We first encode the age label of a training sample x into a probability distribution, denoted by $\mathbf{d}$. This is done by sampling each age class from the normal distribution centered at the ground-truth age. The activation of age class i in $\mathbf{d}$ is defined as

$$d_i = f(i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(i-\mu)^2}{2\sigma^2}} \text{ for } i \in K \tag{1}$$

where μ is the mean of the distribution and the ground-truth age, σ is the variance, which defines the shape of the normal distribution. Naturally, if an age class is too far away from the ground-truth, it will get a very small value.

In the forward propagation, without the need of modifying the backbone model, the output of the last linear layer, denoted by $\mathbf{z}$, is normalized to a probability distribution by Softmax function, from which we have

$$\hat{d_j} = \text{Softmax}(\mathbf{z})_j = \frac{e^{z_j}}{\sum_{i=1}^{K} e^{z_i}} \text{ for } j \in K \tag{2}$$

where z_j is the element corresponding to the age class of i in $\mathbf{z}$. And the regressed estimation of age $\hat{y}$ is given by by taking the expectation of the distribution.

$$\hat{y} = \begin{bmatrix} k_1 & k_2 & k_3 & \cdots & k_{|K|} \end{bmatrix}^T \cdot \hat{\mathbf{d}} \tag{3}$$

To optimize our model, Kullback-Leibler divergence loss is employed to minimize the distance between the predicted distribution and the ground-truth distribution. $L1$ loss (distance) is employed to minimize the regressed expectation and the ground-truth age value. The final loss

function is the sum of the two loss functions

$$L = L_{KLDiv} + L_{L1} = \sum_{i \in K} d_i \ln \frac{d_i}{\hat{d}_i} + \lambda |\hat{y} - y| \tag{4}$$

which will then be minimized.

3.3 Backbone Models

In our research, we mainly look into the three network architectures: VGG-16 with Batch Normalization, ResNet-50, and ResNeXt-50 (32 × 4d). Eventually, we choose SE-ResNeXt-50 as part of our best practice after extensive experiments, which are discussed in section 4.4. Here, we explain in detail the design of the three networks we will compare.

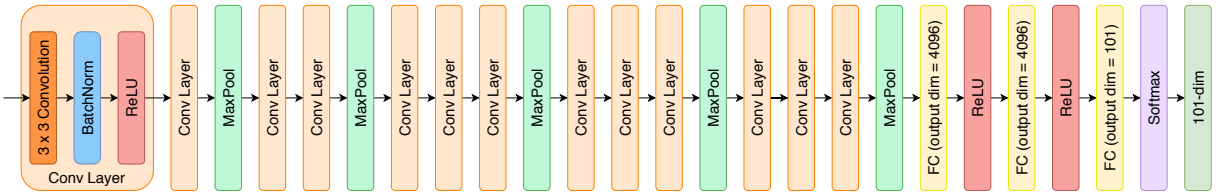


Figure 1: Architecture of VGG-16 with Batch Normalization

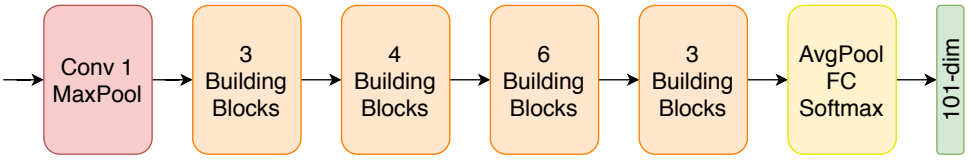
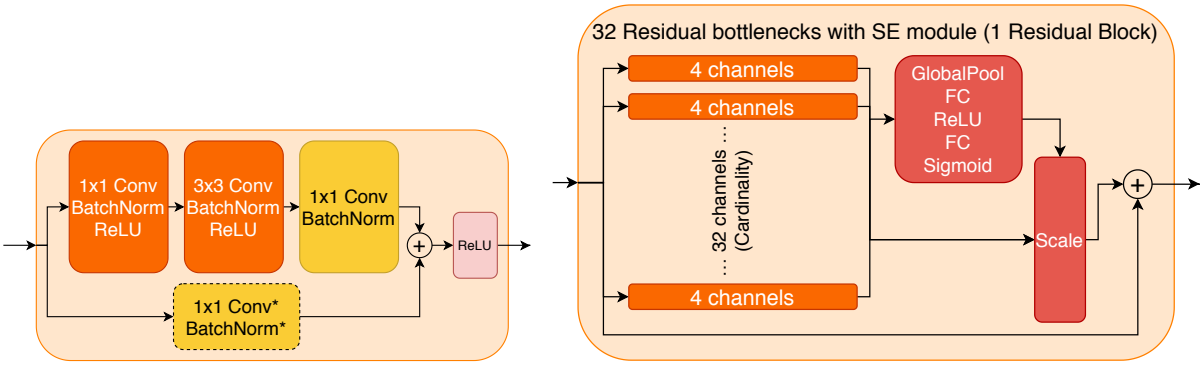


Figure 2: Basic Structure of ResNet-50 and SE-ResNeXt-50 (32 × 4d)



(a) ResNet-50(*: Only apply to the first block in each layer)

(b) SE-ResNext-50

Figure 3: Detail of One building block in SE-ResNeXt-50 (32 × 4d)

VGG-16 with Batch Normalization: VGG-16 [10] is a classic network architecture for image recognition and classification. The architecture of VGG-16 is straightforward and easy to understand. It is also widely used in many papers in age estimation, such as [5] and [7]. In our experiment, a batch normalization [11] layer is added in between each convolution layer and ReLU layer, which is expected to accelerate the training and improve the performance.

ResNet-50: ResNet-50 [12] is proposed to address the degradation problem, that deeper network leads to a higher training error and decreased accuracy. By adding an identity shortcut between each residual block, ResNet enables the training of a deeper neural network. ResNet-50 has a basic structure shown in Figure 2. It has 4 main layers built upon the residual block visualized in Figure 3 (a). Note that the downsampling module, including a 1×1 convolution and a ReLU activation, is only present in the first block of each layer. Compared to VGG-16, ResNet-50 has fewer parameters, trains faster, and achieves a higher accuracy on ImageNet image classification.

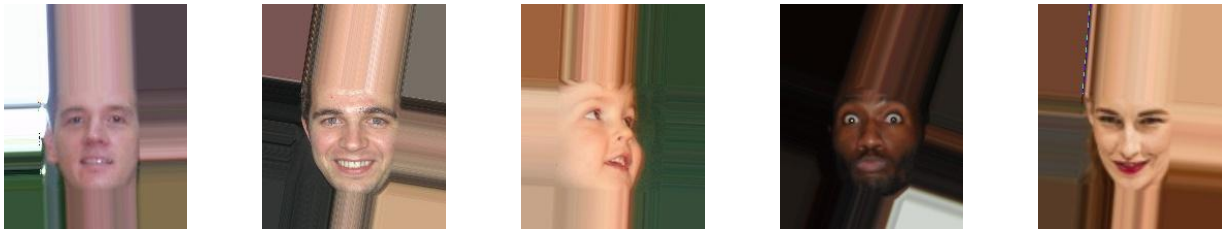
SE-ResNeXt-50 (32 × 4d): SE-ResNeXt-50 [13] has the same basic architecture from a high-level perspective (Figure 2). In detail, its structure, as is shown in Figure 3 (b), is a combination of the Squeeze-and-Excitation (SE) [13] block and the ResNeXt [14] architecture. ResNeXt, in each block, splits the trivial residual bottleneck into a number (defined as *cardinality*) of groups of convolution operations, known as grouped convolution [15]. It is proved that increasing cardinality, other than increasing the width or the depth of the network, can also improve performance. SE-ResNeXt-50 introduces the SE block to the ResNeXt. SE module *squeeze* the output of a residual block to form a channel descriptor, then perform *excitation* to adaptively learn a $x \in (-1, 1)$ to recalibrate the learned features to be more discriminative. As a result, a residual block is shown in Figure 3 (b). Here, we use SE-ResNeXt-50 with a template of 32 groups in each residual block (cardinality = 32), and 4 channels in each convolution (bottleneck width = 4), which is denoted by (32 × 4d).

4 Experiments and Analysis

4.1 Datasets

(a) 68, Male, Black (b) 58, Male, White (c) 56, Male, Black (d) 39, Female, White (e) 17, Male, Hispanic

Figure 4: Sample Images from MORPH



(a) 19 (b) 35 (c) 5 (d) 36 (e) 22

Figure 5: Sample Images from MegaAge

MORPH: MORPH [16] is one of the biggest longitudinal face databases available to the public and the most used benchmark for age estimation models. It contains 55,000 images taken

(a) 1 (b) 21 (c) 35 (d) 58 (e) 35*

Figure 6: Sample Images from MegaAge-Asian (*: An obviously wrong data sample of non-Asian)

in a constrained manner. Age ranges from 16 to 77. Metadata, including accurate age, gender and ethnicity are given, making this dataset ideal for multitasking learning. However, 77% of the images are of African and 19% are of European, suggesting that this dataset is highly unbalanced. This dataset uses mean absolute error (MAE) as its evaluation metric. There are multiple evaluation protocols among publications. Here, we randomly split 80% of the images into the training set and rest into the validation set. The MAE on the validation set is reported and compared.

MegaAge: MegaAge [17] is a relatively new dataset featuring 41,941 images crawled from the Internet. The images are taken in unconstrained conditions and vary significantly. Very few of the images are of low quality or are not human faces at all. The evaluation metric is cumulative accuracy (CA). This dataset provides separate lists for training and testing. We report CA(3), CA(5) and CA(7) on the testing set.

MegaAge-Asian: Released along with MegaAge, MegaAge-Asian is a large dataset allegedly containing images of Asians only. There are 40,000 images in total. In practice, this dataset suffers from noises as well, such as images of Europeans or images without an identifiable human face. However, this dataset if still of great value due to the lack of facial age data of Asian people. This dataset uses CA as the evaluation metric and provides training/testing separation. We report CA(3), CA(5) and CA(7) on the given testing set.

4.2 Evaluation Metrics

Mean Absolute Error (MAE) is widely used as an evaluation metric of the performance of age estimation models. It is defined as the average error between the predicted age and the ground-truth, which can be calculated as:

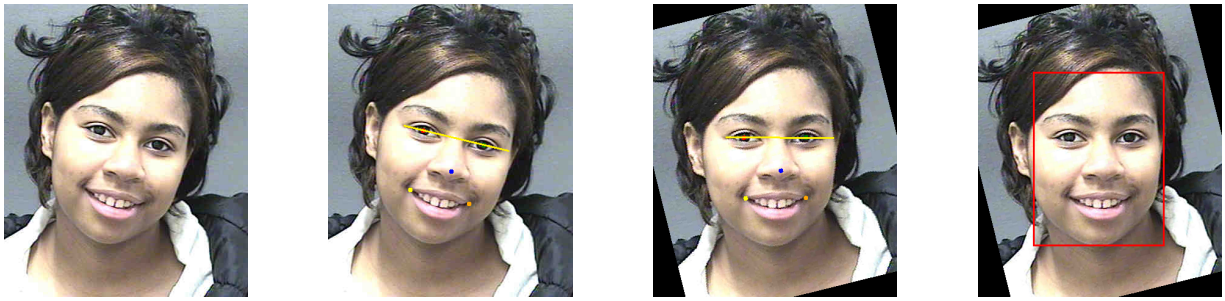
$$MAE = \frac{\sum_{i=1}^{N} |\hat{y}_i - y_i|}{N} \tag{5}$$

where N is the total number of samples, $\hat{y}_i$ is the predicted age of the i-th sample, and y_i is the ground-truth age of the i-th sample. The unit of MAE is year. A smaller MAE is better.

Cumulative Accuracy (CA) is used to evaluate the performance on MegaAge and MegaAge-Asian, which can be calculated as:

$$CA(n) = \frac{K_n}{K} \times 100\% \tag{6}$$

where n is the maximum accepted error, K_n is the number of samples whose absolute error (absolute difference between the predicted age and the ground-truth age) is smaller than n, and K is the total number of samples. The unit of CA is %. A bigger CA(n) is better when n is the same.



(a) Original (b) MTCNN Pass 1 Facial Landmarks (c) Rotation (d) MTCNN Pass 2 Bounding Box*

Figure 7: Preprocessing (*: The face area is not cropped out at this point. Only the coordinates are saved.)

4.3 Preprocessing

Face detection and alignment are carried out before each image is sent into the model for training and validation. This is done using an MTCNN [18] detector, which is a popular detector for face detection and extracting facial landmarks. Each image will go through 2 passes. In the first pass, facial landmarks are extracted. We use the coordinates of the two eyes to calculate an angle of rotation. After rotation, the two eyes should form a horizontal line, which indicates the face is aligned. In the second pass, MTCNN is used again to get the bounding box of the aligned face. The coordinates of the bounding box are saved so that we can adjust the margin area when cropping out the face from the image.

In our experiment, we use a PyTorch implementation of MTCNN. Other implementations should suffice as well.

4.4 Model Comparison

In this section, we compare three network architectures: VGG-16 with Batch Normalization, ResNet-50, and the relatively new SE-ResNeXt-50 (32 $\times$ 4d). Note that all the models are pretrained models on the ImageNet dataset, that is, they are not randomly initialized. We also benchmark each model in two ways of formulating the task: classification, and label distribution learning (LDL).

For all the experiments in this paper, unless otherwise specified, we report the best reasonable results out of several runs, that is, the training processes producing these results did not fail, and these results do not deviate far from results obtained under the same settings.

4.4.1 Age Estimation as Classification

To formulate age estimation as a conventional classification problem, each age is considered an individual class. Therefore, we encode age labels into one-hot vectors. The output of the network is a 101-class vector, referring to a range from 0 to 100 years old. The output is normalized with softmax and the class with the highest probability (argmax) is considered the predicted age value. Cross entropy is used as the loss function in the training process.

We can draw a conclusion from Table 1 that VGG-16 with Batch Normalization performs the best on all three datasets when age estimation is cast as a classification problem. SE-ResNeXt-50 follows VGG-16 BN and the original ResNet-50 performs generally the worst.

Architecture	MORPH	MegaAge			MegaAge-Asian		
	MAE	CA(3)	CA(5)	CA(7)	CA(3)	CA(5)	CA(7)
VGG-16 BN	**2.5203**	**30.03**	**48.30**	**64.04**	**52.52**	**74.33**	**86.29**
ResNet-50	2.9587	29.47	47.05	61.87	46.81	69.13	81.89
SE-ResNeXt-50	2.8500	29.23	47.44	63.69	50.06	72.64	84.09

Table 1: Comparison of models on age estimation as a classification problem

Architecture	MORPH	MegaAge			MegaAge-Asian		
	MAE	CA(3)	CA(5)	CA(7)	CA(3)	CA(5)	CA(7)
VGG-16 BN	2.5253	33.02	51.42	66.33	61.03	80.07	90.32
ResNet-50	2.6490	34.46	52.78	67.43	60.34	78.96	89.29
SE-ResNeXt-50	**2.4447**	**37.14**	**55.21**	**71.19**	**64.03**	**81.68**	**90.87**

Table 2: Comparison of models on age estimation as a label distribution learning problem

4.4.2 Age Estimation as Label Distribution Learning

To train an age estimation model with the label distribution strategy, we follow the setting described in section 3.2. Each model still outputs a 101-class vector as does in the classification setting. The sum (weight $\lambda = 1$) of Kullback-Leibler divergence loss and $L1$ loss is adopted as the loss function. When encoding the labels, we use a normal distribution with mean $\mu =$ ground-truth age and standard deviation $\sigma = 2$ in our experiments.

As is shown in Table 2, with label distribution learning, SE-ResNeXt-50 outperforms both VGG-16 BN and ResNet-50, producing the best results on all three datasets. In addition, models generally perform much better with label distribution learning, which indicates LDL does provide a universal mechanism that would benefit many neural network architectures rather than a specific one.

We will use SE-ResNeXt-50 with label distribution learning as the primary setting in the following experiments.

4.5 Data Augmentation

(a) Original (b) HF (c) HSV (d) Gamma (e) BC

Figure 8: Sample Images after Data Augmentation

Data augmentation is utilized to generate more training data from a relatively small dataset and help to improve the generalization capability of a model by simulating possible alternative forms of existing samples. Some commonly used data augmentation methods may not be suitable for the task of age estimation, such as severe color shift, channel shuffle, distortion, blur, random

crop, etc. Because we want to ensure that the faces, which we have aligned and cropped out, retain the shape and color of human faces after the transformations so that the neural network will not be wasting time learning some extreme cases that are impossible to happen in real life. Thus, the goal is to mimic the possible effects that can be generated in the process of obtaining a facial image, such as wrong camera configurations, different lighting conditions, photo editing, or even the compression induced by converting the image from one format to another. In experiments, we adopt the following four methods of data augmentation, which are visualized in Figure 8:

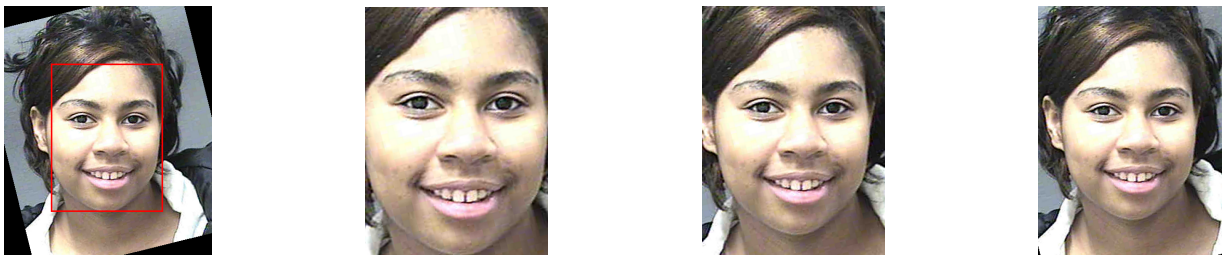
- Random horizontal Flip (HF)
- Random Hue Saturation Value Shift (HSV)
- Random Gamma Shift (Gamma)
- Random Brightness & Contrast Shift (BC)

Each transformation is initialized with a probability of 0.3 in order to lay more emphasis on the original data. It is possible for more than one transformation to take effect and produce a more complicated augmented sample.

Data Augmentation	MORPH	MegaAge			MegaAge-Asian		
	MAE	CA(3)	CA(5)	CA(7)	CA(3)	CA(5)	CA(7)
With Augmentation	**2.3269**	36.89	**55.64**	**71.36**	63.49	**81.89**	**92.13**
Without Augmentation	2.4447	**37.14**	55.21	71.19	**64.03**	81.68	90.87

Table 3: Comparison of LDL with or without Data Augmentation

We can see from Table 3 that data augmentation works well on MORPH dataset. On MegaAge and MegaAge-Asian, however, CA(3) is not improved, which may suggest that some features that are crucial to highly accurate predication are not well preserved after these transformations. Important features, such as wrinkles, could easily be covered by overexposure and other effects. Still, since one image can contribute to a whole age group in label distribution learning, the increase of samples can lead to an improvement in the rough estimation, which is reflected in the increased CA(5) and CA(7). It is not safe to say that adding data augmentation is a universally good strategy that will suit all datasets. Both options will be tested in the next experiment.



(a) Before Cropping (b) No Expansion (c) 20 % Expansion (d) 40 % Expansion

Figure 9: Sample Images with Different Margin Expansion

Inspired by [5], we further consider the effect of slightly expanding the face area when cropping out the face. While such expansion is not part of the data augmentation pipeline, it is performed at this stage and seems to work well with data augmentation as they both alter the input data. In [5], each cropped face is expanded a little to include a 40% margin, which is supported by

performance increase. In Figure 9, it can be seen that the bounding box generated by MTCNN strictly contains the face area only. By expanding the cropped face by 40 %, hair, ears, neck, clothing, and some background are included. We doubt that this is the optimal setting as the clothing and the background seem to take much area. Thus, we add another configuration of expanding the cropped face by 20 %, which contains some hair, ear, and neck information but eliminates the appearance of the unrelated background. To do such expansion, the height and width of the bounding box are multiplied by the factor (20 % or 40 %). If the expanded bounding box exceeds the size of the image, we simply take the border of the image to form a new bounding box.

Aug.	Expansion	MORPH	MegaAge			MegaAge-Asian		
		MAE	CA(3)	CA(5)	CA(7)	CA(3)	CA(5)	CA(7)
With	0	2.3269	36.89	55.64	71.36	63.49	81.89	92.13
	20%	**2.3041**	37.04	56.08	70.89	63.57	81.86	91.11
	40%	2.3337	**37.37**	**56.41**	**71.58**	**64.92**	**83.45**	**92.36**
Without	0	2.4447	37.14	55.21	71.19	64.03	81.68	90.87
	20%	2.4925	36.45	54.93	69.24	62.98	81.96	91.39
	40%	2.4523	36.63	54.60	70.35	63.52	82.14	91.03

Table 4: Comparison of LDL with Different Margin Expansion

As a result, performance is further improved compared to applying data augmentation only. In Table 4, Expansion = 0 means they are baseline models from the previous experiment. On MORPH dataset, an expansion of 20% along with data augmentation produces the best result. On MegaAge and MegaAge-Asian, the best expansion factor is 40%. Considering the bigger variance of the images in MegaAge and MegaAge-Asian, it is reasonable for them to require additional area to include information that is not always at the expected position. Overall, data augmentation with margin expansion produces decent results. In the following experiment, we will follow the best practice so far and use data augmentation with 20 % expansion on MORPH, and 40 % expansion on MegaAge and MegaAge-Asian.

4.6 Multi-task Learning

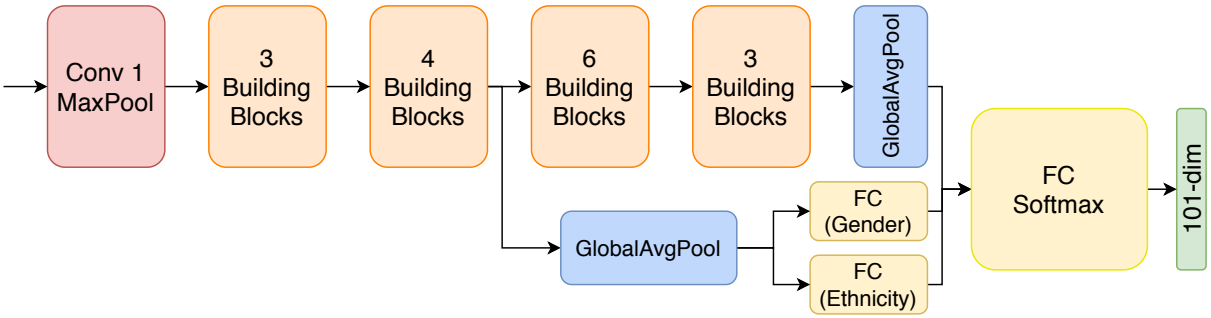


Figure 10: Modified Multi-task network based on SE-ResNext-50 (32 × 4d)

In [7], researchers observed that multi-task learning of both gender recognition and age estimation has a positive regularization effect, and age estimation is more complex than gender recognition. Based on the observation, we introduce a small modification to the original SE-ResNeXt-50 (32 × 4d) architecture. As is shown in Figure 10, from the original SE-ResNeXt-50 (32 × 4d), we take out the output of the second residual block and pass it to a global average

Multi-task learning	MORPH
	MAE
Gender & Ethnicity	**2.2310**
Solely Age Estimation	2.3041

Table 5: Comparison of LDL with or without Gender and Ethnicity Recognition

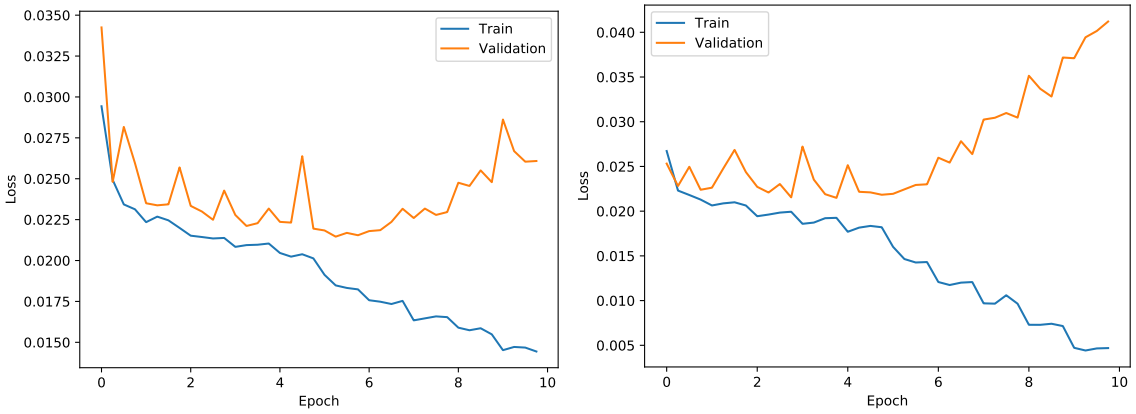
pooling layer, forming a 512-dimension feature vector. This feature vector is then passed to two fully connected layers, one for gender recognition (output dimension = 2) and one for ethnicity detection (output dimension = 5). The output of these two FC layers and the output of the last average pooling layer are then concatenated to form a 2055-dimension feature vector, which will be sent into the last FC layer to form a predicted age distribution. By extracting features from an early stage in the network, we do not introduce extra time and computation that are required if we choose to train separate networks.

Even with only two simple fully connected layers, gender recognition and ethnicity detection reach 91% accuracy after the first epoch and eventually achieve 98-99 % accuracy. More importantly, as is shown in Table 5, this model achieves the lowest MAE among all the experiments we conduct, reducing MAE by 0.073 year. One predicament of employ this model, however, is that only MORPH provides both gender and ethnicity labels.

5 Discussion

5.1 The Advantages of Label Distribution Learning

In section 4.4, when formulating age estimation as a classification problem, despite the generally inferior performance, it is also important to note that ResNet-50 and SE-ResNext-50 converge rather quickly. ResNet-50 stops getting better results after 9 epochs on MORPH, 5 epochs on MegaAge, and 7 epochs on MegaAge-Asian. SE-ResNeXt-50 stops getting better results after 7 epochs on MORPH, 3 epochs on MegaAge, and 9 epochs on MegaAge-Asian. Compared to label distribution learning, classification models converge faster but cannot improve over more epochs.

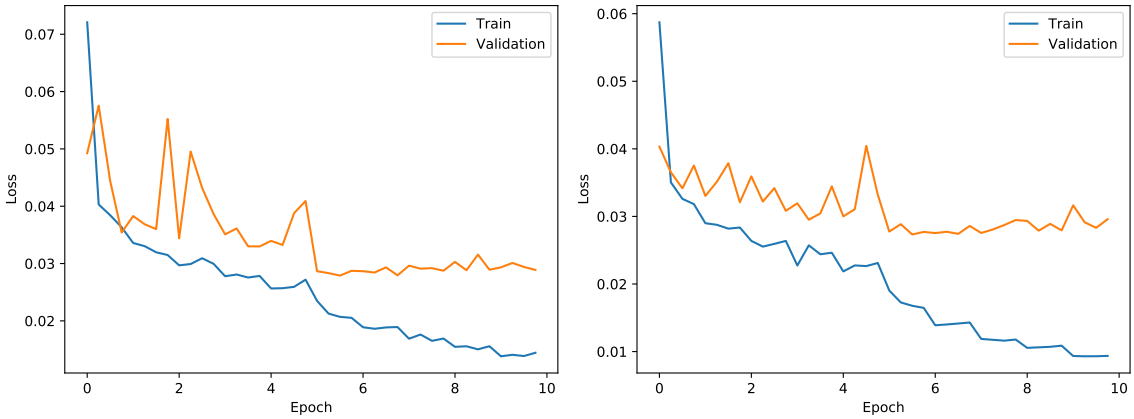


(a) VGG-16 BN on MORPH

(b) SE-ResNeXt-50 on MORPH

Figure 11: Training and validation loss of VGG-16 BN and SE-ResNeXt-50 on MORPH (Classification)

To better understand why models stop improving at a suboptimal state, we visualize the train-



(a) VGG-16 BN on MORPH

(b) SE-ResNeXt-50 on MORPH

Figure 12: Training and validation loss of VGG-16 BN and SE-ResNeXt-50 on MORPH (LDL)

ing and validation loss of another batch of baseline experiments with the only two variables being the model (VGG-16 BN or SE-ResNext-50, considering the similarity between ResNet-50 and SE-ResNeXt-50) and the problem formulation (classification or label distribution learning), without data augmentation, margin expansion, multi-task learning. We record four validations losses during one training epoch and visualize only the first ten epochs to achieve a higher resolution. These model reaches similar *level* of performance to those previous experiments after only 10 epochs but may improve slightly with more epochs. Note that the lower value of loss does not necessarily lead to a lower MAE, but the behavior of the loss reflects the behavior of the models.

In Figure11, we can observe that the validation loss of both classification models starts to grow shortly after the start of the training session while their training loss is still decreasing, which suggests that these models may be subject to severe overfitting. Specifically, the validation loss of SE-ResNeXt-50 grows at a higher rate than that of VGG-16 BN. This corresponds to the phenomenon that SE-ResNeXt-50 stops getting better results faster than VGG-16 BN and VGG-16 is able to obtain good results occasionally after 20 epochs. SE-ResNext-50 effectively deteriorates faster than VGG-16 BN with more epochs.

A possible explanation for this is that each class is completely unrelated to other classes in this problem setting, which ignores the ordinal nature of age. And one-hot encoding makes each image only contributes to its labeled class of age. In this way, each class has an average of a few more than 400 sample images. Each class is not sufficiently trained, leading to high predication error and overfitting.

In the setting of label distribution learning, as is shown in Figure 12, the two models are much more stable when trained for more epochs, allowing the model to be trained for longer, which is why models with LDL can achieve better results but they typically need more epochs. By encoding the age label into a probability distribution, one sample can activate both the targeted age class and ages around it. This fits our intuition that aging is continuous progress and that people within a certain age range may look quite similar in terms of details that indicate age progression. Each class is effectively trained on many more samples, which prevents overfitting. Through shared training data, adjacent age classes are connected and rectified by each other, that is, one class with a high activation value will activate slightly its adjacent classes. This helps the model to preserve the ordinal information between each age class and better learn to predict a probability distribution.

Method	MAE
OR-CNN [3]	3.27*
SSR-NET [19]	3.16
Ranking-CNN [4]	2.96*
DAG-CNN [6]	2.81*
DEX [5]	2.68
VGG-16 CNN + LDAE [7]	2.35
Our Method	2.3041
Our Method (Multi-task)	2.2310
DRFs [20]	2.17*
Mean-variance Loss [8]	2.16*
DLDL-v2 [2]	1.969
LRN [9]	**1.905**

Table 6: Comparison of different methods on MORPH (*: a slightly different validation protocol is adopted)

Method	CA(3)	CA(5)	CA(7)
SSR-NET [19]	54.9	74.1	N/A
MSFCL-KL [21]	62.89	82.46	N/A
JCSPL [17]	64.23	82.15	90.8
LRN [9]	64.52	82.03	91.7
3-stage Network [22]	64.8	83.2	91.4
Our Work	**64.92**	**83.45**	**92.36**

Table 7: Comparison of different methods on MegaAge-Asian

5.2 Comparison with Other State-of-the-art Methods

In Table 6, we compare our results on MORPH with some of the state-of-the-art methods. MORPH has been used in almost all studies of age estimation algorithm, which means it has been explored thoroughly. However, our method avoids the exhausting pretraining process on a large dataset required by models such as DLDL-v2 [2], which is pretrained on Microsoft's large-scale face dataset MS-Celeb-1M [23]. Such pretraining is also not very practical at this point since Microsoft has ceased the distribution of this dataset. Instead, we opt to adopt a much more competent network architecture to extract sufficient discriminative features from the training data. Methods such as LRN [9] and Mean-variance Loss [8], both of which aim to adaptively shape the predicted age distribution, should be able to be conveniently integrated into our current model and yield satisfactory results.

Furthermore, we compare the result we achieve on MegaAge-Asian to some existing methods in Table 7. CA(7) is not reported in SSR-NET [19] and MSFCL-KL [21]. we can see that, to the best of our knowledge,our model produces the best result. Though the performance difference between each method is relatively small. If data cleansing is conducted on this dataset, removing images of non-human objects and non-Asian, better performance is expected on all these methods. Specifically to our method, because MegaAge-Asian does not provide gender or ethnicity labels, we cannot directly employ the multi-task model on it to improve performance. However, semi-supervised learning sheds light on how to train a model on unlabelled data with a small portion of external, labeled data.

Results on MegaAge are not listed here because it is relatively less used among researches compared to the other two datasets.

6 Conclusion

In this paper, we proposed a combination of a strong CNN network, SE-ResNeXt-50 (32 × 4d), and a learning strategy, label distribution learning. Label distribution learning enhances the learning process by preserving the ordinal information between ages and making one sample activate multiple age classes, which prevents overfitting and helps to learn accurate age distribution. Through experiments, we show that these two together form a best practice. We further show that the performance of the purposed method can be improved through data augmentation, margin expansion, and multi-tasking learning. As a result, our method performs well on MORPH, MegaAge, and MegaAge-Asian dataset, producing one of the best results on MegaAge-Asian.

While our method produces decent results, it is not the best on MORPH dataset, suggesting possible ways of further improvement. It is applicable to introduce the adaptive mechanism of reshaping the age distribution to our current model so that it can adapt to the various aging rate at different stages of age progression. Our multi-task learning model also has the potential to be generalized by generating gender and/or ethnicity label using semi-supervised learning techniques.

References

[1] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, and X. Geng, "Deep label distribution learning with label ambiguity," *IEEE Transactions on Image Processing*, vol. 26, no. 6, p. 2825–2838, Jun 2017. [Online]. Available: http://dx.doi.org/10.1109/TIP.2017.2689998

[2] B.-B. Gao, H.-Y. Zhou, J. Wu, and X. Geng, "Age estimation using expectation of label distribution learning," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI 2018)*, 2018, pp. xx–xx.

[3] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Ordinal regression with multiple output cnn for age estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[4] S. Chen, C. Zhang, M. Dong, J. Le, and M. Rao, "Using ranking-cnn for age estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[5] R. Rothe, R. Timofte, and L. V. Gool, "Deep expectation of real and apparent age from a single image without facial landmarks," *International Journal of Computer Vision*, vol. 126, no. 2-4, pp. 144–157, 2018.

[6] S. Taheri and Önsen Toygar, "On the use of dag-cnn architecture for age estimation with multi-stage features fusion," *Neurocomputing*, vol. 329, pp. 300 – 310, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231218313110

[7] 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*, vol. 72, pp. 15 – 26, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320317302534

[8] H. Pan, H. Han, S. Shan, and X. Chen, "Mean-variance loss for deep age estimation from a face," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5285–5294.

[9] P. Li, Y. Hu, X. Wu, R. He, and Z. Sun, "Deep label refinement for age estimation," *Pattern Recognition*, vol. 100, p. 107178, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320319304789

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[11] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[13] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," 2018.

[14] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," *arXiv preprint arXiv:1611.05431*, 2016.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[16] A. W. Rawls and K. Ricanek, "Morph: Development and optimization of a longitudinal age progression database," in *Biometric ID Management and Multimodal Communication*, J. Fierrez, J. Ortega-Garcia, A. Esposito, A. Drygajlo, and M. Faundez-Zanuy, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 17–24.

[17] Y. Zhang, L. Liu, C. Li, and C. C. Loy, "Quantifying facial age by posterior of age comparisons," in *British Machine Vision Conference (BMVC)*, 2017.

[18] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, p. 1499–1503, Oct 2016. [Online]. Available: http://dx.doi.org/10.1109/LSP.2016.2603342

[19] T.-Y. Yang, Y.-H. Huang, Y.-Y. Lin, P.-C. Hsiu, and Y.-Y. Chuang, "Ssr-net: A compact soft stagewise regression network for age estimation," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. International Joint Conferences on Artificial Intelligence Organization, 7 2018, pp. 1078–1084. [Online]. Available: https://doi.org/10.24963/ijcai.2018/150

[20] W. Shen, Y. Guo, Y. Wang, K. Zhao, B. Wang, and A. L. Yuille, "Deep regression forests for age estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[21] M. Xia, X. Zhang, W. Liu, L. Weng, and Y. Xu, "Multi-stage feature constraints learning for age estimation," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 2417–2428, 2020.

[22] Y. Tingting, W. Junqian, W. Lintai, and X. Yong, "Three-stage network for age estimation," *CAAI Transactions on Intelligence Technology*, vol. 4, no. 2, pp. 122–126, 2019.

[23] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *European conference on computer vision*. Springer, 2016, pp. 87–102.