An Improved Selective Facial Extraction Model for Age Estimation

Chengwen Song
China Jiliang University
wadessong@gmail.com

Lingmin He
China Jiliang University
helm@cjlu.edu.cn

Wei Qi Yan
Auckland University of Technology
wyan@aut.ac.nz

Parma Nand
Auckland University of Technology
parma.nand@aut.ac.nz

Abstract

In this paper, we propose an improved end-to-end learning algorithm to address the aggregation of multiclass classification and regression for age estimation by using deep Convolutional Neural Networks (CNNs). Inspired by Soft Stagewise Regression Network (SSR-Net), we take residual units embedded with channel and spatial feature response correlation values and dynamically adopted the kernel corresponding to these feature maps into our model. In addition, we used weight normalization at each layer by using input samples at the beginning of training and this weight normalization was beneficial both in terms of accuracy as well as training time. We validate the proposed model based on benchmark datasets and compare the MAE with seven other mainstream networks. The results reveal that our model achieves an improved performance. Our contribution is an updated algorithm for age estimation by adopting the latest attention and normalization mechanisms for balancing the efficiency and accuracy of the proposed model.

1. Introduction

Estimating biological age and gender of a person based on a single facial image has always been a classic and challenging research field in computer vision and deep learning for many years. Besides, there are a great deal of applications benefitted from its potential usage, such as photo editor, surveillance system, market analysis, commercial television advertisement, and so on. However, it is still hard to precisely annotate the age of a person only through facial information since there are massive variations of facial information among a crowd of people with the same age. It is a fact that a small group of younger people indeed look older than those elder due to aging process. The purpose of age estimation is to minimize the errors between real age and its estimated one.

Based on previous work, the approaches of age estimation can be mainly divided into two parts: classification [2][9][13] and regression [1][8][15]. Great importance has been attached to the mind of multiclass classification for many years. The theory for regarding age prediction as a multiclass classification task is that age can be roughly divided into several discrete parts where each one can be viewed as an independent class, and then a proper classifier trained with data can be used to infer individual's age. For instance, Ranking-CNN [2] has been designed based on the age-related ordinal information by aggregating a series of binary outputs from CNN-based classifiers for age prediction.

On the other hand, there are massive researches devoting to casting the age prediction as a regression problem based on CNNs by deep learning because aging is a time series process [8]. The techniques of regression are introduced the construction of a continuous relationship between facial information and ages. For instance, DEX [9] designed an algorithm to estimate age based on regression of age value in a classification-based neural network. Niu et al. proposed a CNN learning algorithm with multiple outputs to address the ordinal regression problem for age predication. More recently, SSR-Net [15] proposed by Yang et al. was adopted to address the issue of age estimation based on multistage classification and regression operation in the end with soft stage strategy.

Our contribution to age estimation is that the construction of residual block normalized by weights of layers embedded with adaptively receptive field size and attention mechanism along channel and spatial dimension has been adopted in our scheme for enhancement of facial feature representation without imposing too much parameters at the same time. Meanwhile, the size of our model is relatively compact (about 2 MB), which is suitable to be deployed on those distributed or mobile devices with limited memory and computational resources.

In this paper, our literature review will be introduced in Section 2, our method will be depicted in Section 3, our experimental results will be detailed in Section 4, Section 5 will give the concluding remarks.

2. Related Work

2.1. Multi-class Classification for Age Estimation

It is more understandable and simpler for us to split a group of people into several classes according to facial information, which facilitates classification approaches for age estimation. DEX is one of the most significant projects among all classification-based methodologies. It firstly generated a vector of probabilities at the top layer of a CNN-based model, and then aggregated each probability in

978-1-7281-4187-9/19/$31.00 ©2019 IEEE

the vector according to a regression formulation of age value.

2.2. Selective Kernel Network

Based on previous work on CNNs, fixed receptive fields of neurons fail to have powerful ability in feature representation, Li et al. [6] proposed a novel method based on CNNs that allows each neuron to dynamically select its kernel size based on various scales of input data.

There are three stages existing in the block: splitting, reusing, and selecting. The splitting operation is responsible for creating multiple branches with different kernel sizes. Those feature maps are then integrated to achieve a global feature representation for weight distribution in the stage reusing. The final stage selecting reassigns those feature maps of various sizes according to the proportion of weight in the former stage. Our experiments for object classification and super resolution have proven the success of this method.

2.3. Weight Normalization

Weight normalization [10] is a method developed by OpenAI for overcoming defects of batch normalization [7], the operation of normalization is conducted based on the weights of layers rather than minibatch. As well known, a standard artificial neural network takes nonlinear operation on a weighed sum of input features:

$$\mathrm{y} = \emptyset(\boldsymbol{w}\cdot\boldsymbol{x} + b) \tag{1}$$

where w is a k-dimensional weight vector, b is a scalar bias, $\boldsymbol{x}$ is a k-dimensional vector of input features, $\emptyset(\cdot)$ indicates an elementwise nonlinearity, and y denotes the scalar output of the neuron.

Weight normalization redefines the weight vector $\boldsymbol{w}$ of any layer in terms of a parameter vector $\boldsymbol{v}$ and a scalar parameter g by using

$$\boldsymbol{w} = \frac{g}{\|\boldsymbol{v}\,\|}\boldsymbol{v} \tag{2}$$

where $\boldsymbol{v}$ is a $\boldsymbol{k}$-dimensional weight vector, g is a scalar, and $\|v\|$ indicates the Euclidean norm of $\boldsymbol{v}$. It splits the norm and optimizes both $\boldsymbol{v}$ and g by using gradient descent of a loss function L with new parameters $\boldsymbol{v}$ and g.

$$\nabla_g L = \frac{\nabla_w L\cdot\boldsymbol{v}}{\|\boldsymbol{v}\|}, \nabla_v L = \frac{g}{\|\boldsymbol{v}\|}\nabla_w L - \frac{g\nabla_g L}{\|\boldsymbol{v}\|^2}\boldsymbol{v} \tag{3}$$

where $\nabla_w L$ is the gradient of weights $\boldsymbol{w}$. This equation shows that the weight gradient of each layer has been scaled, and the gradient has been projected away from the current weight vector. Both operations benefit optimization of deep learning models.

Fig.1. Our algorithm for solving multistage regression with the multiple output CNN.

Input: training data $\boldsymbol{D} = \{\boldsymbol{x_i}, \boldsymbol{y_i}\}_{\boldsymbol{i=1}}^{\boldsymbol{N}}$ and testing images $\boldsymbol{D'} = \{\boldsymbol{x'_j}\}_{\boldsymbol{j=1}}^{\boldsymbol{M}}$

- Loop for $i = 1, 2, \cdots, \boldsymbol{N}$:
  - Data augmentation for $\boldsymbol{x_i}$ with rotation, scale, flip, clip, shift and shuffle.
- The learning of multistage regression:
  - The proposed multiple output CNN is trained with augmented training set according to the proposed learning algorithm (refers to Sec. 3.3, 3.4).
- For each testing image $\boldsymbol{x'_j}$ in $\boldsymbol{D'}$:
  - Forward $\boldsymbol{x'_j}$ to the trained model, and get $\boldsymbol{S}$ sets of values $\{\vec{p}^s\ ,\vec{\beta}^s, \theta_s\}$,(s = 1, 2, ⋯ , $\boldsymbol{S}$);
  - Estimate the age $\tilde{y}$ with previous outputs $\{\vec{p}^s\ ,\vec{\beta}^s, \theta_s\}_{s=1}^{S}$ according to Eq.7.

Output: the predicated values for testing images $\{\tilde{y}(\boldsymbol{x'_j})\}_{\mathrm{j=1}}^{M}$

Different from previous work, in this paper, we propose a novel method for human age estimation. To the best of our knowledge, this is the first time that the neural work has been applied to age estimation. Our contribution is to improve the existing methods and greatly promote the accuracy of age estimation.

3. Our Method

3.1. Problem Formulation

The age prediction based on a single face image is recently investigated underneath the deep convolution neural networks. Normally, there is a collection of training samples of human face images labeled with real ages. $\{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$, where x_i indicates the i-th facial image and y_i is the age label of x_i. The objective of the task is to learn a mapping equation $\mathcal{F}$ that obtains $\tilde{y} = \mathcal{F}(x)$ as the estimated age for an input facial image,

$$\tilde{y} = \vec{p}\cdot\vec{l} = \sum_{i=0}^{R} p_i \cdot l_i, \tag{4}$$

where $\boldsymbol{R}$ is the range of age, and p_i indicates the probability of age l_i. Besides, in order to guarantee the accuracy of our algorithm, Mean Absolute Error (MAE) is taken as the evaluation metric of our model for minimizing the error between the estimated age and the ground truth label,

$$\mathrm{J}(X) = \frac{1}{N}\sum_{n=1}^{N}|\widetilde{y_n} - y_n|, \tag{5}$$

where $\widetilde{y_n} = \mathcal{F}(x_n)$ indicates the predicted age for the input image x_n.

3.2. Residual Attention Block

Inspired by the mind of CBAM [14] and selective kernel mechanism, the block is proposed as a kind of attention units for the promotion of facial feature representation via the transformation, $\mathcal{F}\colon \mathrm{X} \rightarrow \tilde{X}$, where $\mathrm{X} \in R^{H \times W \times C}, \tilde{X} \in R^{H' \times W' \times C'}$. We can simply consider function $\mathcal{F}$ (•) as a standard convolutional operation along the channel and spatial dimensions. For channel attention, multiple sizes of kernel and pooling operations are adopted to feature maps for distinguishable feature vectors, and then the fusion result is obtained via channel-wise multiplication. The procedure of spatial dimension is as same as channel dimension. Figure 2 (b) illustrates more detail description about the block.

3.3. Multistage Regression

A large amount of previous work has casted the age estimation to a classification, and then aggregated the results for age regression. For instance, DEX divided the age into 101 classes, and trained a CNNs-based network for age classification. The procedure of their method can be visualized as

$$\tilde{y} = \vec{p} \cdot \vec{l} = \textstyle\sum_{i=0}^{100} p_i \cdot l_i, \qquad (6)$$

where $\vec{p} = (p_0, p_1, \cdots, p_{100})$ is obtained from the top layer of the model, which indicates the distribution of age probability based on input facial data, and $\vec{l}$ represents indexes of each probability. It is accurate and simple to separate the age duration based on the interval of one year. However, the number of parameters of the network is large, and plenty of computing resources are required, which is complicated and time-consuming for training the network.

To overcome that drawback as well as maintain the accuracy of age prediction, Soft Stagewise Regression Network (SSR) has been proposed to turn one regression into multistage regression by reducing the size of deep neural network to a relative compact and efficient one. The age is predicted through multistage regression

$$\tilde{y} = \vec{p} \cdot \vec{l} = \sum_{s=1}^{S} \sum_{i=0}^{s_i - 1} p_i^s \cdot l_i^s = \sum_{s=1}^{S} \sum_{i=0}^{s_i - 1} p_i^s \cdot i \left(\frac{100}{\sum_{p=0}^{s-1} s_p} \right) (7)$$

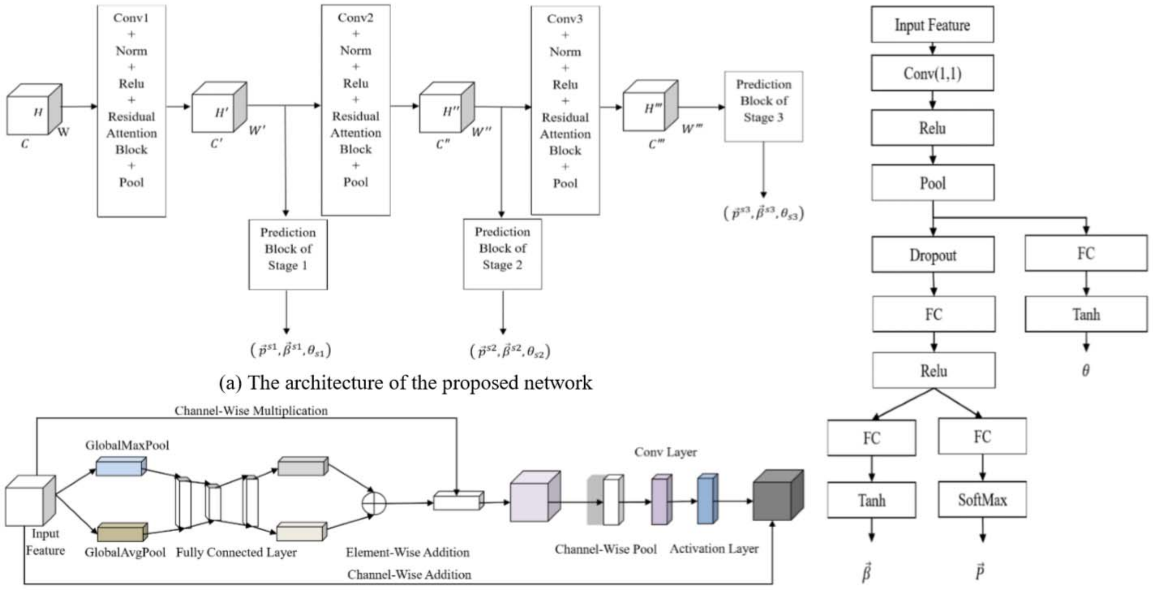
For each stage s, p_i^s , l_i^s indicates that the probability distribution of age within each stage and indexes of the classes in each stage, respectively.

Observations based on previous work reveal that dividing age duration into completely independent classes fails to cope with issues of aging process. The techniques of scaling stage width and indicator selection have been adopted to address the problem. The operation of stage width adjustment is explained as

$$\overline{s_\iota} = s_i(1 + \theta_i) \,, \qquad (8)$$

where $\overline{s_\iota}$ and s_i indicate the i-th stage width after and before the adjustment, respectively; θ_i is a factor that determines the degree of stage width change, which is a part of the network. Thus, the stage width is,

$$w_s = \frac{100}{\sum_{p=0}^{s-1} \overline{s_p}} . \qquad (9)$$

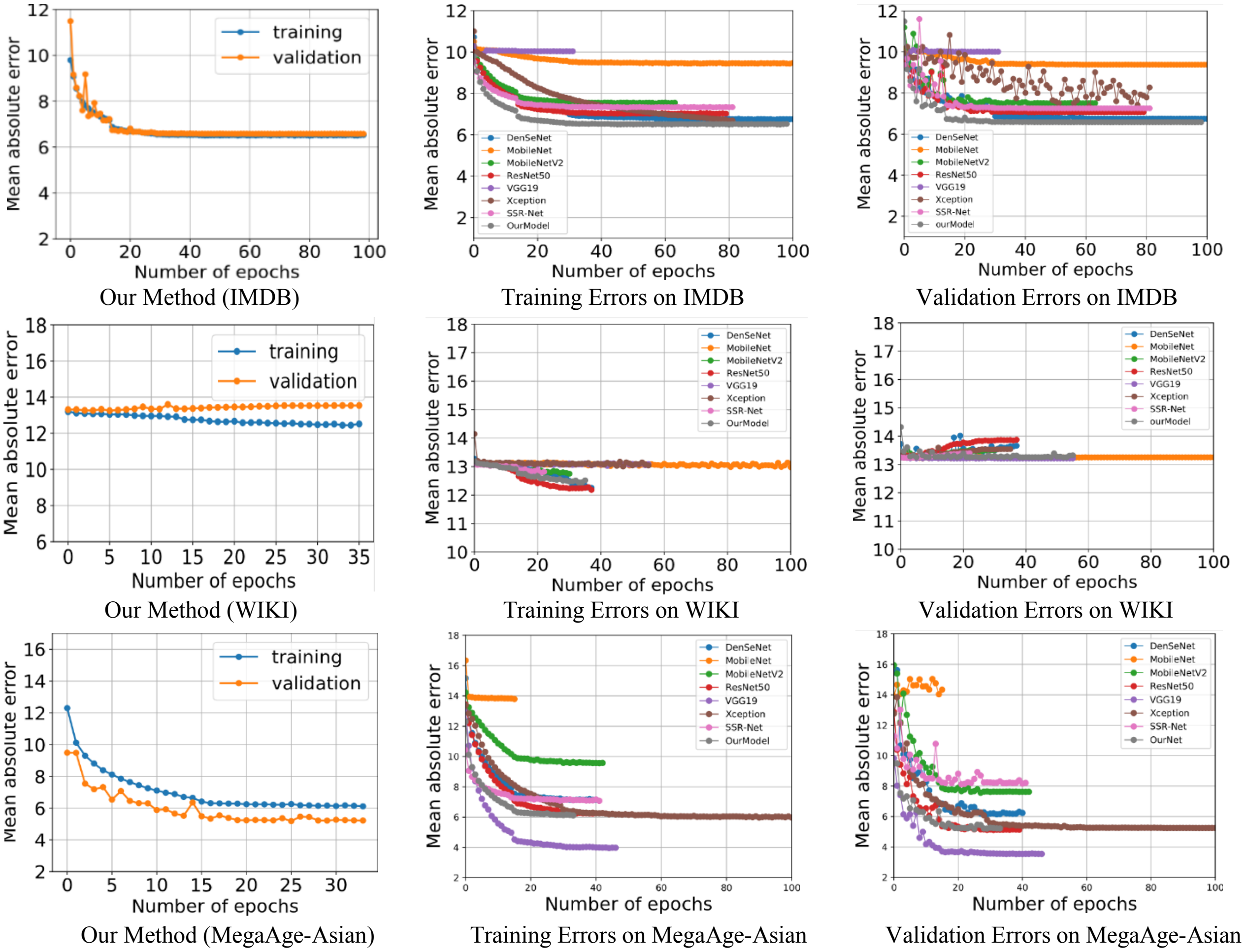


(a) The architecture of the proposed network

(b) The residual attention block

(c) The prediction block

Fig.2. The Architecture of our proposed scheme



Our Method (IMDB)

Training Errors on IMDB

Validation Errors on IMDB

Our Method (WIKI)

Training Errors on WIKI

Validation Errors on WIKI

Our Method (MegaAge-Asian)

Training Errors on MegaAge-Asian

Validation Errors on MegaAge-Asian

Fig. 3. Our method training results and comparisons of training and validation with DenseNet, MobileNet, MobileNetV2, ResNet50, VGG19, Xception and SSR-Net on IMDB, WIKI and MegaAge-Asian.

For shifting the indice of age classes within each stage, the offset vector $\vec{\beta}^s$ has the same size as l_i^s. The value of $\vec{\beta}^s$ is also one part of outputs in our model

$$\vec{\beta}^s = \left(\beta_0^s, \beta_1^s, \cdots, \beta_{s_i-1}^s\right). \tag{10}$$

Then, the index now becomes

$$\vec{\imath} = i + \beta_i^s. \tag{11}$$

Both operations depend on input facial information, allowing our model to learn a capability in dynamically predicting the apparent age.

3.4. Network Structure

In order to find a compromise solution in terms of efficiency and accuracy for age estimation, we decide to take residual structure with selective kernel and attention mechanism into consideration, and then propose our improved model in a way of deep learning. Figure 2(a) reveals the overall architecture of our network, 2-stream structure of SSR-Net is not used in our nets as we conclude that the construction of residual attention block with adjustable kernel has already achieved promising performance in extracting facial features for age prediction via experimental results. Besides, it will doubly increase parameters of the model and may affect the scope of usage.

There are three branches in our scheme. Each branch is composed of convolution, weight normalization, activation, three basic residual blocks, and pooling layer. Meanwhile, several residual attention blocks are embedded in every stage. Like the construction showed in Figure 2(b), the structure of our residual block is composed of convolution, weight normalization, channel and spatial attention, and fusion layer. Various filter kernels with different sizes and downsampling methods are applied to residual unit. The inside operation of element-wise multiplication on two feature maps produces in channel attention is applied to combine the feature maps extracted by different kernel sizes. After that, the integrated features are fed into spatial attention for constructing the relationship of aging process.

Hence, our algorithm can dynamically capture facial features for.

The features produced from three branches are then fed into prediction block for age estimation. This prediction block is responsible for outputting two vectors and a scalar of s-th stage for regression of age value, which are the vector of probability $\vec{p}^s$, the vector of index shifting $\vec{\beta}^s$, and the stage width factor θ_s, respectively. Fig 2 (c) illustrates the structure of prediction block. Before obtaining these values, each feature should be operated by transition layer, normalization layer, and downsampling layer.

The feature maps are then fed into a dense layer followed by normalization and nonlinear activation functions for these three values. The activation method $Tanh(\bullet)$ is used for θ_s. According to eq. (8), the interval of θ_s is $[-1,1]$. The procedure for vectors $\vec{p}^s$ and $\vec{\beta}^s$ is similar with θ_s but more complex. Meanwhile, softmax activation function is used for $\vec{p}^s$ while the $Tanh(\bullet)$ activation function is adopted for $\vec{\beta}^s$ because the index can shift to left or right.

4. Experiments

4.1. Preprocessing and Experimental Settings

In this paper, we use two public benchmark datasets for the evaluation of our model and comparison experiments with previous methods, including the IMDB-WIKI and MegaAge-Asian datasets.

We conduct data augmentations based on our datasets for reducing overfitting and improving generalization capability of our model, encapsulating shift, flipping, scaling, rotating, and adjusting on image contrast and luminance. Then, a square area with the size of 64×64 is cropped from face images to allow our network to contribute to the age prediction with less deviations. Meanwhile, our model is implemented by using Keras, which is a mainstream frame for deep learning. We redesigned the layers for multistage regression with residual attention block embedded with selective kernel strategy.

For our hardware usage, we used an Intel i7 CPU and an NVIDIA TITAN X. Followed the settings on SSR-Net, our model has three stages with another three patches for each one. The optimization function of our model is the modified Adam optimizer with weight normalization for 200 epochs. The learning rate begins at 0.001 and decays for a factor 0.1 every 20 epochs. Besides, the training data is divided by a factor 128 at the beginning of each epoch. From the view of saving time and training efficiency, the parameter of early stopping for model checkpoint is set as 15 epochs.

4.2. Experiments on IMDB-WIKI

There are several benchmark datasets consisting of face images with age labelled. We trained our method on IMDB-WIKI dataset, which is the biggest public dataset, containing 523,051 face images from the internet. There are about 90% images from IMDB while the rest are from Wikipedia.

The database is affected by a range of noises, such as blurred images and images without face or with multiple faces even if it is the largest one. Thus, we did not use it for performance evaluation, instead of pretraining. The dataset was used for validating the efficiency and effectiveness of our algorithm via the evaluation metric of Mean Absolute Error. Besides, at the beginning of the experiment, the datasets were divided into two components according to the settings of SSR-Net, one part with 80% of the whole images for training purpose, the other with 20% for validation purpose.

In order to demonstrate the superiority of our algorithm in terms of accuracy and time consumption, seven state-of-the-art CNNs were considered as comparative experiments, consisting of five compact networks, MobileNet [4], MobileNetV2[11], Xception [3], DenseNet [5] and SSR-Net, and two bulky models, ResNet50, VGG19 [12]. Figure 3 shows the process of convergence generated by each network during training and validation. In the first two rows, the blue curves of first column indicate the training errors of MAE and the orange ones were generated from the validation samples based on IMDB-WIKI dataset. Both are close to each other. Our model which was trained based on the training set can be applied to validation successfully. By observing curves from the last two columns, it is clear to see that the curves of our model have reached the best convergence point in minimal time when using IMDB dataset. However, all models failed to have a good performance based on WIKI.

4.3. Experiments on MegaAge-Asian

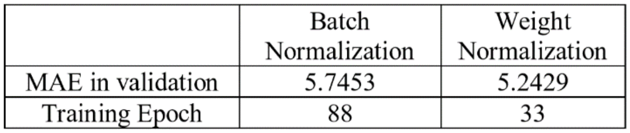
In IMDB-WIKI datasets, our training samples are almost captured from western countries. We assume that face data from different regions may be a key factor to affect the performance of our algorithm. For promoting the generalization capability of our model under multiple conditions, we took MegaAge-Asian dataset [15]. This dataset contains more than 40K face images of Asians. We took pretraining strategy for reducing the training time. The model trained with IMDB-WIKI dataset was used for starting the training and validation. The third row in Fig. 3 shows the results during training and validation. The blue and orange curves in the left have a much lower convergence than those generated from IMDB-WIKI dataset. It can be explained by the distinguishable and obvious division of age within this dataset, which is also proven by another fact that the speed of model convergence in MegaAge-Asian is faster than that in IMDB-WIKI. In other words, the model is precise to learn and distinguish facial feature representation for different age groups after trained with the dataset.

Among all models, it is obvious that our proposed model achieves the smallest MAE in both training and validating stages with minimal time-consuming. However, unlike the performance owned by using the two bulky models in IMDB-WIKI datasets, VGG19 has achieved the best performance while ResNet50 has a tiny gap. After obtaining the same results by multiple experiments, we test a hypothesis that it is designed for the integrity of data and architecture with deep layers that may affect the performance of model.

4.4. Experiments on Normalization

For validating the weight normalization to the performance of our deep learning model, we conducted comparative experiments on MegaAge-Asian dataset by adopting two different normalization ways, batch normalization and weight normalization. From the result of Table 1, we conclude that weight normalization and initialization speed up the convergence process and improve the performance of our model.

Table 1: Comparison between our models trained with batch and weight normalization.



	Batch Normalization	Weight Normalization
MAE in validation	5.7453	5.2429
Training Epoch	88	33

5. Conclusions

In this paper, we proposed an improved model embedded with residual attention block and selective kernel construction for deeply facial feature extraction and age value prediction to cope with age estimation based on single face image. The mechanism of weight normalization has been applied to this field for the first time, and we find that it achieves an improved performance both in terms of accuracy as well as computational efficiency. The attention mechanism intends to operate on the facial information by adopting different kernel sizes along dimensions of channel and spatial, and this improved its capability in extracting facial features from aging process when it is compared with the state-of-the-art models. The evaluation of our experiments reveal that our proposed algorithm has achieved the best results on both IMDB-WIKI and MegaAge-Asian datasets compared with several other mainstream models. In addition, our model is suited to be deployed on mobile or embedded devices due to its compact size and superior performance. In future, we plan to explore this mechanism for other applications related to facial information representation.

References

[1] Agustsson, E., Timofte, R., & Van Gool, L. (2017). Anchored regression networks applied to age estimation and super resolution. In IEEE International Conference on Computer Vision (pp. 1643-1652).

[2] Chen, S., Zhang, C., Dong, M., Le, J., & Rao, M. (2017). Using Ranking-CNN for age estimation. In IEEE Conference on Computer Vision and Pattern Recognition (pp. 5183-5192).

[3] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In IEEE conference on computer vision and pattern recognition (pp. 1251-1258).

[4] G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

[5] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In IEEE Conference on Computer Vision and Pattern Recognition (pp. 4700-4708).

[6] Li, X., Wang, W., Hu, X., & Yang, J. (2019). Selective Kernel Networks. In IEEE Conference on Computer Vision and Pattern Recognition (pp. 510-519).

[7] Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.

[8] Niu, Z., Zhou, M., Wang, L., Gao, X., & Hua, G. (2016). Ordinal regression with multiple output CNN for age estimation. In IEEE Conference on Computer Vision and Pattern Recognition (pp. 4920-4928).

[9] Rothe, R., Timofte, R., & Van Gool, L. (2015). DEX: Deep expectation of apparent age from a single image. In IEEE International Conference on Computer Vision Workshops (pp. 10-15).

[10] Salimans, T., & Kingma, D. P. (2016). Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In Advances in Neural Information Processing Systems (pp. 901-909).

[11] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). MobileNetv2: Inverted residuals and linear bottlenecks. In IEEE Conference on Computer Vision and Pattern Recognition (pp. 4510-4520).

[12] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[13] Tan, Z., Wan, J., Lei, Z., Zhi, R., Guo, G., & Li, S. Z. (2017). Efficient group-*n* encoding and decoding for facial age estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(11), 2610-2623.

[14] Woo, S., Park, J., Lee, J. Y., & So Kweon, I. (2018). CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 3-19).

[15] Yang, T. Y., Huang, Y. H., Lin, Y. Y., Hsiu, P. C., & Chuang, Y. Y. (2018). SSR-Net: A Compact Soft Stagewise Regression Network for Age Estimation. In IJCAI (Vol. 5, No. 6, p. 7).