NEIGHBORHOOD LOSS FOR AGE ESTIMATION FROM FACE IMAGE USING CONVOLUTIONAL NEURAL NETWORKS

Hyok Kwak, Chol Nam Om, Il Han and Jang Su Kim
Institute of Information Technology, High-Tech Research and Development Centre, Kim Il Sung University, Democratic People's Republic of Korea

Abstract

Convolutional Neural Network (CNN) is widely used in estimating age from face image. In many CNN applications such as image classification, face recognition and other computer vision scopes, the cross-entropy loss is used as a supervision signal to train CNN model. However, the cross-entropy loss only enhances the separability of classes and does not consider their correlation in age estimation task. In this paper we propose a novel loss function called neighborhood loss which regards the correlation between classes in age estimation by modifying standard cross entropy loss. To evaluate the effectiveness of the proposed neighborhood loss, we present CNN architecture based on the residual units. Through some experiments, we show that neighborhood loss provides superior performance compared to prior works in age estimation.

Keywords:

Age Estimation, Neighborhood Loss, Convolutional Neural Network

1. INTRODUCTION

Age estimation or classification technology from face image is widely used in commercial and service-based applications such as people analytic and people flow management, intelligent signage, customer engagement, photo indexing and sorting. Approaches for age estimation or classification can be divided into two categories: the ones not using convolutional neural networks (CNN) called “shallow” and ones using CNN called “deep”.

Shallow methods [1]-[3], not using CNN, start by extracting shallow level features from the face image by using handcrafted feature descriptors such as BIF, LBP and Gabor filter. After feature extraction, the age estimation can be considered as regression or multi-class classification problem such as LDA, SVM. The aging process of human face varies greatly for different individuals and the mechanism of aging of human is still an unresolved problem. Therefore, it is difficult to investigate feature descriptors which represent aging of the human face. Moreover, traditional machine learning algorithms do not fully exploit the big data. Prof. Andrew NG indicated that the performance of older learning algorithms plateaued even though you give it more data.

Most recently, the trend of age estimation is the usage of CNN. The well-designed CNN trained on a large scale dataset can get much representable features compared to hand-crafted features as shown various applications such as image classification and face recognition [4], [5].

In this paper, we propose a significant method for apparent age estimation of face image by using CNN.

Our main contributions are as follows: First, we propose the neighborhood loss regarding the relationship between age classes to learn the CNN classification. Second, we present overall CNN architecture for age estimation and an effective age prediction approach through the classification. Finally, we evaluate the performance of our method compared to prior works through experiments on FG-NET, MORPH and CACD datasets.

The remainder of this paper is organized as follows. In section 2 we review several related works such as DEX and Deep Rank etc. In section 3 and 4 we introduce the neighborhood loss, model architecture and age prediction method. In section 5, we show that through experiments the proposed method improves the performance compared to prior works. We conclude this paper in section 6.

2. RELATED WORKS

Reviewing widely research results for age estimation is out of the scope of this paper. In this paper we refer several meaningful research results for age estimation using CNN.

Recently the application of CNN has been achieved state-of-art performance in computer vision such as image classification, object detection, face recognition and age estimation etc.

ImageNet Large Scale Visual Recognition Challenge (ILSVRC) is an epitome showing the trend of CNN. Trimps Soushen team, the Third Research Institute of the Ministry of Public Security in China, took first place on the ILSVRC 2016 classification task has achieved results of 2.99% error, which is better 16.7% results than 3.6% in 2015 [6]. In ILSVRC 2017, the WMW team (Oxford University and Momenta joint team) took first place as 2.25% which improves by 24.7% than 2016 [7].

In the face recognition tests hosted by NIST (National Institution of Standard and Technology), deep learning-based face recognition algorithms using CNN show excellent results. The research of face recognition is related to other research aspects such as gender and age estimation, emotion recognition. For example, the approaches and training database used in face recognition can be easily adopted also in age estimation. The online face recognition service such as Microsoft Face API [8] and Face++ [9] and offline SDK (Software Development Kit) such as NeoFacePro (NEC Corporation) [10], Luxand FaceSDK(Luxand) [6] include face recognition and age estimation as well.

Gil Levi et al. [11] designed CNN architecture based LeNet-5 for age classification and has achieved better accuracy than the 3D face shape estimation and the method using SVM in Audience Benchmark Test Set. This work shows the result which can be achieved superior accuracy than the traditional ‘shallow’ methods even if use simple CNN architecture.

Huei-Fang Yang et al. [12] proposed DeepRank+, age estimation model using CNN. DeepRank+ combined previous ‘shallow’ method with deep learning approach. DeepRank+

model is composed of ScatNet with 3-layers, PCA dimension reduction and 3-fc layers. ScatNet extracts face representations robust to translations and small deformations; PCA is to reduce the high dimension of the concatenated scattering coefficients (SCs) produced by each node in ScatNet; the fully-connected network learns to predict the age rank. ScatNet extracts scattering representation using Morlet wavelet transform and combined it with pooling layer. Therefore, ScatNet has no learning parameters and has both advantages of scattering representation and pooling. DeepRank+ has achieved excellent performance in MORPH, Lifespan and FACES dataset.

It is known that DEX [13] has excellent performance for age classification. The CNN architecture consists of VGGNet and apply two stage fine- tuning method. Firstly, it fine-tunes a VGG-16 model pretrained for ImageNet classification in IMDB-Wiki dataset. Secondly, further fine-tune the resulting networks on 20 different splits of the ChaLearn LAP dataset. The final prediction result of CNN is the average of ensemble of 20 networks. DEX took first place in ChaLearn LAP challenge 2015.

3. NEIGHBORHOOD LOSS FOR AGE CLASSIFICATION

The cross entropy which is widely used in multi-class classification is defined as follows.

$$L=\sum_{n=1}^{N}L_n \tag{1}$$

where L_n is the loss value for the n^{th} sample x_n and N is the batch size. The L_n is formulated as follows.

$$L_n=-\sum_{k=1}^{K}d_k\log y_k \tag{2}$$

where $\mathbf{d}=[d_1,\cdots d_K]^T$ and $\mathbf{y}=[y_1,\cdots y_K]^T$ is the target output and the final output of neural network for x_n, respectively and K is class number.

As the activation function of the final layer is softmax function, y_k is as follows;

$$y_k=\frac{\exp(x_k)}{\sum_{j=1}^{K}\exp(x_j)} \tag{3}$$

where $x_{(\bullet)}$ denotes the output of last fully connected layer.

The target value $\mathbf{d}=[d_1,\cdots d_k]^T$ is ground truth which represents as one-hot encoding format in image classification task. For example, if x_n belongs to 2nd class label, the 2nd element value of target value **d** is one and all the other values are zero. e.g., $\mathbf{d}=[0100...0]^T$.

In image classification, the classes are considered to be mutually exclusive and uncorrelated each other, and the cross-entropy loss works well.

However, due to the aging characteristics of human face images, exist the correlation between classes in age estimation task. The human development can be divided into infant, childhood, adolescence, middle age and old age according to the similarity of aging characteristics such as wrinkle, eyelid sag. In particular, the aging characteristics of faces belonging to the age classes around specified age class are very similar. Therefore, it is not appropriate to treat age estimation as "hard" multi-class classification which represents the target value as one-hot encoding format.

From the above analysis, we consider the age estimation problem as “soft” multi-class classification. That is, we assume that the target output value follows a Gaussian distribution instead of one-hot encoding, as shown in Fig.1. The d_k is formulated as follows.

$$d_k=\exp\left(-\frac{(k-k_n)^2}{\sigma^2}\right) \tag{4}$$

where k_n is the class label corresponding to x_n and σ is hyper-parameter called “neighborhood margin”.

Therefore, Eq.(2) can be written as follows.

$$L_n=-\sum_{k=1}^{K}\exp\left(-\frac{(k-k_n)^2}{\sigma^2}\right)\log y_k \tag{5}$$

We define Eq.(5), the modified cross entropy loss as neighborhood loss. Our neighborhood loss is a modified version of the cross-entropy loss which regards the correlation between age classes.

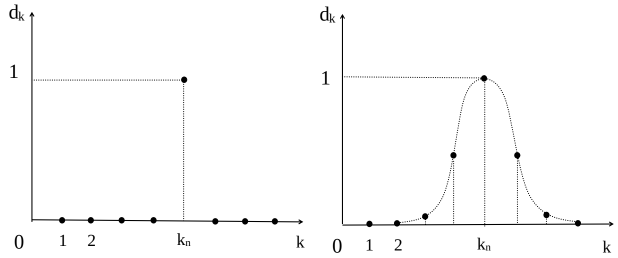

Fig.1. One-hot encoding representation (left) and Gaussian representation (right) of the target value d_k

4. MODEL ARCHITECTURE

Similar to most previous works, our model predicts the apparent age of human from a single face image. Our model architecture consists of face detection, face alignment and preprocessing, CNN architecture, 100-FC layer, softmax layer (output layer) and age prediction.

4.1 FACE ALIGNMENT AND PREPROCESSING

We detected the face bounding box and landmarks by using RetinaFace [14] and applied similarity transform to normalize the face images. The faces are cropped and resized to 224 × 224, and each pixel in RGB images is normalized by subtracting 127.5 then dividing by 128.

4.2 NETWORK SETTING

Application examples of CNN architecture which used to estimate age are LeNet, VGGNet etc. The high-capacity deep convolutional networks may provide high level features which represents aging property of face image. We use residual units [4] in our CNN architecture because they show better performance compared to VGGNet used in image classification.

The Table.1 shows our CNN architectures with 18 and 34 depths. The Conv1_x, Conv2_x, Conv3_x, Conv4_x and Conv5_x denotes convolution units that may contain multiple convolution layers and residual units are shown in double-column brackets. E.g., [3×3, 64] ×3 denotes 3 cascaded convolution layers with 64 filters of size 3×3, and S2 denotes stride 2. The Conv2.x includes max pooling layer of size 3×3 and S2.

Table.1. Our CNN architectures with 18 and 34 depths

Layer	Output size	18-layer CNN	34-layer CNN
Conv1_x	112 × 112	$[3\times3,64]\times1$	$[3\times3,64]\times1$
Conv2_x	56 × 56	$[3\times3]$ max pooling, $S2$	
		$\begin{bmatrix}3\times3,64\\3\times3,64,\ S2\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,64\\3\times3,64,\ S2\end{bmatrix}\times3$
Conv3_x	28 × 28	$\begin{bmatrix}3\times3,128\\3\times3,128,\ S2\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,128\\3\times3,128,\ S2\end{bmatrix}\times4$
Conv4_x	14 × 14	$\begin{bmatrix}3\times3,256\\3\times3,256,\ S2\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,256\\3\times3,256,\ S2\end{bmatrix}\times6$
Conv5_x	7 × 7	$\begin{bmatrix}3\times3,512\\3\times3,512,\ S2\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,512\\3\times3,512,\ S2\end{bmatrix}\times3$
	1 × 1	100-d FC, softmax layer	100-d FC, softmax layer

The input image size is same as 224 × 224 as ImageNet classification task. However, to preserve higher aging feature map resolution, we use conv3×3 and stride = 1 in the first convolutional layer instead of using conv7×7 and stride = 2. The convolution layers follow by fully-connected layer with 100- way age classification. The final layer is the softmax layer.

4.3 OUTPUT LAYER AND AGE PREDICTION

In fact, age estimation can be considered as a regression and not a classification problem, as age is continuous rather than a set of discrete classes.

Unfortunately training a CNN directly for regression is relatively unstable as outliers cause a large error term. This results in very large gradients which makes it difficult for the network to converge and leads to unstable predictions.

So, we handle the age prediction problem as a classification problem, i.e., the deep features extracted from CNN is passed to FC layer and softmax layer to compute class posterior probabilities.

The CNN model for age classification is trained on training dataset with known age (label) by using neighborhood loss. At test time we compute the predicted age value as the weighted sum using output probabilities.

$$Age = \sum_{i=1}^{n} y_i p_i \tag{6}$$

where p_i denotes a probability which the output (classification result) of network belongs to class y_i, where n is the number of classes.

5. EXPERIMENTS

5.1 TRAINING DATA

Our training data includes two existing datasets - IMDB [15] and WIKI [16] released by Rasmus et al. [13]. The IMDB and Wiki includes 461,871 images and 62,359 images respectively, total 524,230 images. However, IMDB-WIKI dataset includes lots of noise samples with false label. We merge IMDB and WIKI and refine by using Microsoft Face API [8] and the manual filtering to decrease the noise of IMDM-WIKI and get a training data of high quality.

To equalize the age distribution of training dataset and prevent the neural network from overfitting, if the number of images in each class is less than 5000 then we perform data augmentation, otherwise randomly ignore partially among images ranging in age from 20 to 60 years.

5.2 TEST DATA

We use three datasets, FG-NET [17], MORPH [18] and CACD [19], as test data.

- **FG-NET:** The Face and Gesture Recognition Research Network (FG-NET) aging dataset contains of 1002 images taken under uncontrolled environment. Its age ranges from 0 to 69.
- **MORPH;** The Craniofacial Longitudinal Morphological Face Database (MORPH) is the large scale public available face dataset which contains more than fifty thousand images. In our experiments, we extract a subset 6000 images which age ranges from 16 to 77. To compare to previous work, we use almost same number of images as it.
- **CACD:** The Cross-Age Celebrity Dataset (CACD) contains 163, 446 images of 2000 celebrities collected from web. In CACD, age is estimated subtracting from taken year of birth. We use 500 images in experiment.

5.3 IMPLEMENTATION AND HYPER-PARAMETER SETTING

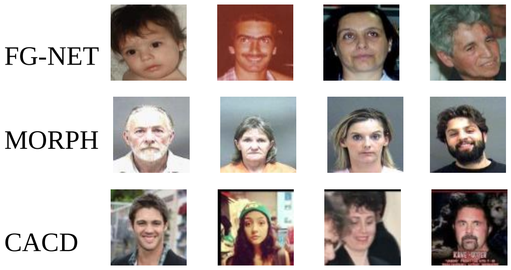

Fig.2. Face image examples from FG-NET, MORPH and CACD

Our implementation is written using python and TensorFlow [5], deep learning framework released by Google. We trained our model on four NVIDIA Titan XP GPUs.

The training variables is initialized by using Xavier-He method [20]. The learning rate is started from 0.1 and decreased by factor 10 at the 50k, 100k and 150k. We use the Adam optimizer with momentum 0.9 and set weight decay at 5e-4. We apply dropout with keep probability of 0.8 in last fully connected layer and adopt batch normalization in each convolutional layer. We empirically set σ of Eq.(5) as 5 in our experiments.

5.4 AGE ESTIMATION IN TEST DATA

Our implementation is written using Python and TensorFlow, deep learning framework released by Google. We trained our model on four NVIDIA Titan XP GPUs. In this section we present the performance of evaluation results in three test datasets as mentioned above.

We use the Mean Absolute Error (MAE) as evaluation metric. It is the average of the absolute error between the predicted age and the ground truth age.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right| \tag{7}$$

where y_i is the true age, $\hat{y}_i$ is the estimated age, and N is number of test images. The lower MAE means, the better performance of an age estimator attains.

MAE is a commonly used measure in the literature and a standard evaluation metric for age estimation.

5.4.1 Effectiveness of Neighborhood Loss:

To validate the effectiveness of our neighborhood loss, we evaluate compare the MAE values for our neighborhood loss and standard cross entropy loss. In this experiment, we train the CNN model with 34 layers described in 4.2 section from scratch. As shown in Table.2, we find that our neighborhood loss improves high accuracy compared to cross entropy loss in age estimation.

5.4.2 Comparison to Others:

We compare our method to other methods such as DEX [13], DIF [21], CA-SVR [22], SVR [23], OHRank [24] and DLA [25]. In this experiment, we train two CNN models which have 18 layers and 34 layers.

Table.2. MAE evaluation for neighborhood loss and cross entropy loss

Loss	MORPH	FG-NET	CACD
Cross entropy loss	4.57	3.94	6.52
Neighborhood loss	2.53	2.78	5.23

As can be seen from Table.3, our method shows superior performance compare to other methods.

Table.3. MAE evaluation results in three test datasets

Method	MORPH	FG-NET	CACD
DEX [7]	2.68	3.09	6.52
DIF [9]	3.80	4.80	8.13
CA-SVR [10]	5.88	4.67	7.98
SVR [11]	5.77	5.66	8.39
OHRank [12]	5.69	4.85	7.54
DLA [13]	4.77	4.26	7.87
Our method (18 layers)	2.71	3.25	6.82
Our method (34 layers)	2.53	2.78	5.23

6. CONCLUSION

In this paper we proposed the neighborhood loss for age estimation. The neighborhood loss is a modified version of the cross-entropy loss, which regards the relationship between age classes. We trained the CNN architecture based the residual units using the neighborhood loss from scratch, without fine-tuning. Our experiments showed that our method compared favorably to prior works.

REFERENCES

[1] M. Riesenhuber and T. Poggio, “Hierarchical Models of Object Recognition in Cortex”, *Nature neuroscience*, Vol. 2, No. 11, pp. 1019-1025, 1999.

[2] K.H. Liu, T.J. Liu, H.H. Liu and S.C. Pei, “Facial Makeup Detection via Selected Gradient Orientation of Entropy Information”, *Proceedings of IEEE International Conference on Image Processing*, pp. 4067-4071, 2015.

[3] T. Ahonen, A. Hadid and M. Pietikainen, “Face Description with Local Binary Patterns: Application to Face Recognition”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 28, No. 12, pp. 2037-2041, 2006.

[4] K.M. He, X.Y. Zhang, S.Q. Ren and J. Sun, “Identity Mappings in Deep Residual Networks”, *Proceedings of European Conference on Computer Vision*, pp. 630-645, 2016.

[5] F. Schroff, D. Kalenichenko and J. Philbin, “FaceNet: A Unified Embedding for Face Recognition and Clustering”, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815-823, 2015.

[6] W. Liu, “ImageNet Large Scale Visual Recognition Challenge (ILSVRC) Overview”, Available at https://www.image-net.org/challenges/LSVRC/2016/index.php, Accessed at 2016.

[7] W. Liu, “ImageNet Large Scale Visual Recognition Challenge (ILSVRC) Overview”, Available at https://image-net.org/challenges/LSVRC/2017/ , Accessed at 2017.

[8] Microsoft Face API, Available at: http://microsoft.com/cognitive-services/en-us/faceapi, Accessed at 2021.

[9] Face++, Available at: http://www.faceplusplus.com/demo-detect/, Accessed at 2021.

[10] Software Development Kit, Available at: http://uxand.com, Accessed at 2021.

[11] G. Levi and T. Hassner, “Age and Gender Classification using Convolutional Neural Networks”, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-13, 2015.

[12] H.F Yang, B.Y. Lin, K.Y. Chang and C.S. Chen, “Automatic Age Estimation from Face Images via Deep Ranking”, *Proceedings of British Machine Vision Conference*, pp. 1-8, 2015.

[13] R. Rothe, R. Timofte and L.V. Gool, “DEX: Deep Expectation of Apparent Age from a Single Image”, *Proceedings of International Conference on Computer Vision Workshop*, pp. 252-257, 2015.

[14] J.K. Deng, J. Guo, Y.X. Zhou, J.K. Yu, I. Kotsia and S. Zafeiriou, "Retinaface: Single-Stage Dense Face Localisation in the Wild", Proceedings of International Conference on Computer Vision and Pattern Recognition, pp. 1-7, 2019.

[15] IMDB Face Dataset, Available at: https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/static/imdb_crop.tar, Accessed at 2022.

[16] Wiki face dataset, https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/static/wiki_crop.tar, Accessed at 2021.

[17] G. Panis, A. Lanitis, N. Tsapatsoulis and T.F. Cootes, "Overview of Research on Facial Ageing using the FG-NET Ageing Database", *IET Biometrics*, Vol. 5, No.2, pp. 37-46, 2016.

[18] K. Ricanek and T. Tesafaye, "Morph: A Longitudinal Image Database of Normal Adult Age-Progression", *Proceedings of International Conference on Automatic Face and Gesture Recognition*, pp. 1-8, 2006.

[19] B.C. Chen, C.S. Chen and W.H. Hsu, "Face Recognition and Retrieval using Cross-Age Reference Coding with Crossage Celebrity Dataset", *IEEE Transactions on Multimedia*, Vol. 17, No. 6, pp. 804-815, 2015.

[20] K.M. He, X.Y. Zhang, S.Q. Ren and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification", *Proceedings of IEEE International Conference on Computer Vision*, pp. 1026-1034, 2015.

[21] H. Han, C. Otto, X.M. Liu and A.K. Jain, "Demographic Estimation from Face Images: Human vs Machine Performance", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 37, No. 6, pp. 1148-1161, 2015.

[22] K. Chen, S.G. Gong, T. Xiang and C.C. Loy, "Cumulative Attribute Space for Age and Crowd Density Estimation", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2467-2474, 2013.

[23] G.D. Guo, Y. Fu, C.R. Dyer and T.S. Huang, "Image-Based Human Age Estimation by Manifold Learning and Locally Adjusted Robust Regression", *IEEE Transactions on Image Processing*, Vol. 17, No. 7, pp. 1178-1188, 2008.

[24] K.Y. Chang, C.S. Chen and Y.P. Hung, "Ordinal Hyperplanes Ranker with Cost Sensitivities for Age Estimation", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 585-592, 2011.

[25] X.L. Wang, R. Guo and C. Kambhamettu, "Deeply-Learned Feature for Age Estimation", *Proceedings of IEEE Winter Conference on Applications of Computer Vision*, pp. 534-541, 2015.