ISSN 2616-5775 Vol. 7, Issue 5: 130-135, DOI: 10.25236/AJCIS.2024.070517

Age-InceptionNet: A Deeper and More Robust Facial Age Prediction Model

Yaohui Wang

Petroleum School, China University of Petroleum-Beijing at Karamay, Karamay, 834000, China

Abstract: *With the rapid advancement of computer vision and artificial intelligence, facial age estimation has become an important area of research and application in various fields such as preventing adolescent gaming addiction, identity verification, and target advertising. Existing methods primarily suffer from sensitivity to dataset biases, limited applicability, and shallow neural network depths. Addressing these issues, this paper propose an innovative approach based on convolutional neural networks, namely Age-InceptionNet. Drawing upon the excellent structure of GoogLeNet, this method optimizes data preprocessing, feature extraction, fusion, and age prediction regression, thereby enhancing the accuracy and robustness of the model. Experimental results demonstrate that Age-InceptionNet achieves favorable outcomes with a Mean Absolute Error of 3.33 on the Morph-II dataset. This paper provides a new solution for age estimation, improving model performance and contributing to further advancements in this field.*

Keywords: *Age Estimation, Convolutional Neural Network (CNN), Feature Extraction, Feature Fusion*

1. Introduction

In recent years, driven by rapid advancements in computer vision and deep learning, facial recognition and age estimation have become prominent topics in both research and application domains. Age estimation finds extensive use across various contexts, including facial recognition systems, security surveillance, adolescent addiction prevention, and age-specific targeted advertising.

With the rise of deep learning technologies, researchers are exploring novel approaches like multi-task learning and feature enhancement networks for age estimation. These methods not only enhance accuracy but also reveal the significant potential of deep learning in this domain. Despite progress, challenges such as data bias and model robustness persist. Hence, this study aims to propose innovative solutions to improve facial age estimation's accuracy and reliability.

Facial age estimation comprises data preprocessing, age feature extraction, and age prediction modules. Data preprocessing is crucial for effectiveness, alongside age feature extraction. Recent research has introduced innovative age grouping schemes and explored multi-task learning models, such as EGroupNet [1] and methods embedding age difference information based on Kullback-Leibler [2] divergence, addressing factors like gender and ethnicity.

However, challenges remain. Some models are sensitive to dataset biases or exhibit instability with diverse facial attributes. To address these, this study draws from GoogLeNet [3], known for its robust feature extraction and age prediction. Leveraging the Inception module, it captures features at multiple scales, aiding in comprehensive feature capture across age groups. GoogLeNet's structure endows it with strong generalization capability, exhibiting robust performance across different facial data types. Its deep learning features, optimized objective functions, and utilization of unlabeled image data enable state-of-the-art performance in facial age estimation tasks.

1) We propose an end-to-end age estimation method based on the Inception structure, named Age-InceptionNet, which extensively employs the Inception architecture. Age-InceptionNet comprises a convolutional neural network with multiple layers of depth and complex structures, effectively increasing the network’s width and depth and enhancing its feature extraction capability.

2) We initially pre-train the model on ImageNet [4] and subsequently conduct numerous experiments on the Morph-II [5] dataset. The accuracy of age prediction exhibits a certain improvement as a result.

ISSN 2616-5775 Vol. 7, Issue 5: 130-135, DOI: 10.25236/AJCIS.2024.070517

2. Related work

Facial age estimation refers to the process in which machines infer the approximate age or age range of individuals based on facial images. It is an interdisciplinary research task spanning multiple fields. Facial age estimation systems typically consist of four main components: face detection and localization, age feature extraction, age estimation, and system performance evaluation. These systems find applications in various domains such as video surveillance, consumer behavior analysis, targeted advertising, security systems, and adolescent gaming addiction prevention.

With the continuous expansion of deep learning technologies, significant achievements have been made in the research of facial age estimation. Levi et al. developed a shallow CNN architecture, utilizing "three convolutional layers and two fully connected layers [4]" to learn feature representations, effectively addressing the overfitting issue caused by limited availability of large facial aging databases. Gurpinar er al. employed "Kernel extreme learning [5]" machines for classification, utilizing deep learning to estimate the apparent age of facial images. Duan et al. proposed a CNN2ELM integrated structure combining convolutional neural networks and ensemble learning machines: RaceNet+AgeNet+GenderNet+ElM [6] classifiers+ELM regressors, to address the impact of gender, ethnicity, and other intrinsic and extrinsic factors on age prediction, making the age prediction more robust. Liao et al. introduced a "divide-and-conquer" learning model: "AgeNet [7]" and "divide-and-rule" architecture in age estimation, addressing the "ordinal regression" problem related to age estimation. Duan et al. proposed the feature enhancement network EgroupNet [1], utilizing a multi-task learning model for age grouping and prediction, resolving the impact of various factors such as (ethnicity, smile, gender) on age estimation. Liu et al. proposed a method called Similarity-Aware Deep Adversarial Learning(SADAL [8]), enhancing the discriminative ability of learned feature representations for facial age, addressing the overfitting issue caused by existing methods enhancing indiscriminateness within each category [9].

3. Preliminary Information

The convolutional neural network (CNN) is a deep feedforward neural network characterized by features such as local connectivity and weight sharing [10]. Its structure typically consists of three main parts: the input layer, hidden layers, and output layer, making it a framework for deep supervised learning [11]. The input layer receives two-dimensional image information, while the hidden layers comprise three types of networks: convolutional layers, pooling layers, and fully connected layers. In the convolutional layer, each neuron is connected to a local receptive field from the previous layer, extracting features from these local receptive fields through filtering and non-linear transformations. In the pooling layer, feature dimensionality reduction can be applied to the features extracted by the convolutional layer, aiming to mitigate negative effects resulting from excessively large model parameters. In the fully connected layer, neurons in the layer are connected to all neurons in the previous layer, a component present in any model with at least one fully connected layer. During computer vision tasks, neural networks initially extract local features from input images, encode these features into one-dimensional vectors, evaluate them, and finally fuse these local features through a series of operations.

4. Architecture of the Age-InceptionNet

In this section, I will provide a detailed description of the structural design of Age-InceptionNet, as illustrated in the figure. Age-InceptionNet comprises two stages: feature fusion and age regression. The feature fusion stage consists of convolutional layers, pooling layers, and fusing layers. In the Age-InceptionNet network, a 3X224X224 image is extracted into a 1X1024 feature vector, which is then passed through a multi-layer perceptron in the feature fusion stage to perform age regression, thereby accomplishing age estimation. Model evaluation is conducted using the Mean Absolute Error function. The following provides a comprehensive overview of the Age-InceptionNet network.

4.1 The design of Age-InceptionNet

The Age-InceptionNet, as depicted in Figure 1, is composed of multiple Inception modules [3] along with a small number of pooling layers and a multi-Layer perceptron stacked together. The Inception portion is divided into five stages, comprising a total of nine Inception blocks. The first stage consists of a 7X7 convolutional layer followed by a 3X3 max-pooling layer. The second stage consists of 1X1 and

ISSN 2616-5775 Vol. 7, Issue 5: 130-135, DOI: 10.25236/AJCIS.2024.070517

3X3 convolutional layers along with a 3X3 max-pooling layer. The third stage consists of two Inception blocks and a 3X3 max-pooling layer. The fourth stage comprises five Inception blocks, a 3X3 max-pooling layer, and an auxiliary classifier branch. The fifth stage consists of two Inception blocks, a global average pooling layer, and an auxiliary classifier. As illustrated in Figure 2, the process of facial age estimation in the Age-InceptionNet unfolds as follows: the Inception network initially processes the input 3X224X224 feature image through stages 1 and 2, resulting in a 192X28X28 feature map. Stage 3 further processes this feature map into a 480X14X14size. Stage 4 transforms the 480X14X14 feature map into 832X7X7. Finally, stage 5 processes the 832X7X7 feature map into a 1024X1X1 output, representing a 1024-dimensional feature vector, which is then fed into a multi-layer perceptron for feature regression. Inception, through the combination of filters of various sizes, effectively identifies image details across different ranges, resulting in highly efficient recognition performance across the entire network.

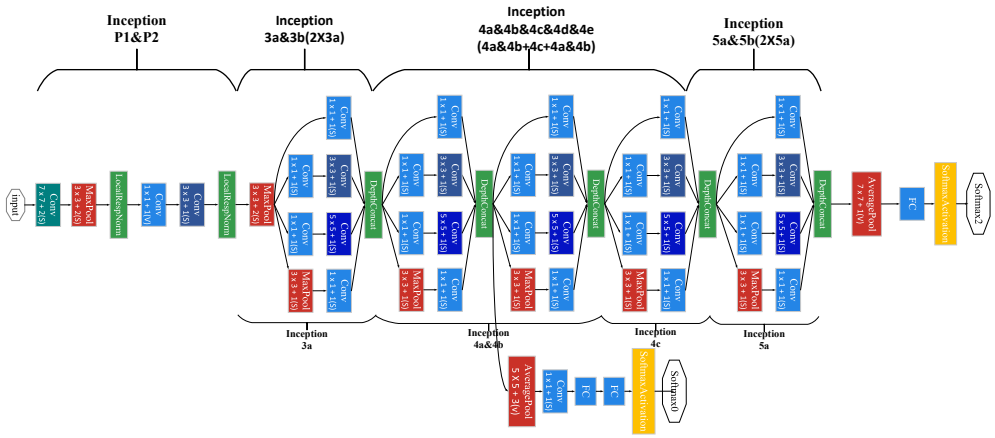

Figure 1: The structure of the Age InceptionNet network consists of 5 segments, a total of 9 Inception modules, and a total of 22 convolutional neural layers and a MLP

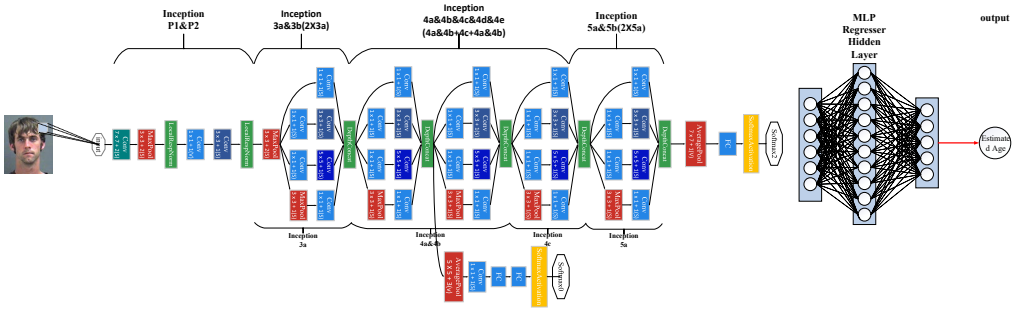

Figure 2: Age-InceptionNet Processing: After processing, the image is transformed into a 224X224 structure, which is then subjected to an Inception network (9 layers of Inception modules and 22 layers of neural networks) for age prediction. Finally, multiple layers of perceptions are used for age regression, and the average absolute error is used to determine the quality of model fitting.

1) Inception Structure

In the Inception [3] network, the fundamental convolutional block is referred to as the Inception block. As illustrated in Figure 2, the Inception block consists of four parallel paths: a 1X1 convolutional layer, a concatenation of a 1X1 convolutional layer followed by a 3X3 convolutional layer with padding of 1, a concatenation of a 1X1 convolutional layer followed by a 5X5 convolutional layer with padding of 2, and a 3X3 max-pooling layer with padding of 1. The first three paths extract information from different spatial sizes. The middle two paths perform 1X1 convolutions on the input to reduce the number of channels, thus reducing model complexity. The fourth path utilizes a 3X3 max-pooling layer, followed by a 1X1 convolutional layer to alter the number of channels. Finally, the outputs of each path are concatenated along the channel dimension to form the output of the Inception module.

2) Convolution Calculation

For a given image $X \in \mathbb{R}^{M \times N}$ and a filter $W \in \mathbb{R}^{U \times V}$, typically with $U << M$ and $V << N$, their convolution is defined as:

$$y_{i,j} = \sum_{u=1}^{U} \sum_{v=1}^{V} \omega_{uv} x_{i-u+1,j-v+1} \quad (1)$$

ISSN 2616-5775 Vol. 7, Issue 5: 130-135, DOI: 10.25236/AJCIS.2024.070517

where ω represents the learned weights. The two-dimensional convolution of input information X and filter W is defined as:

$$Y = W * X \tag{2}$$

when extracting local information from two-dimensional images, neurons are typically organized into a three-dimensional structure of neural layers, with a height of H, width of W, and depth of D, consisting of $D \times W \times H$ feature maps. Here, a feature map represents a feature extracted after convolution from an input image. Let x be the input feature map group, X^p be the feature map matrix, y be the output feature map group, Y^p be the feature map matrix, and $W^{p,d}$ be the two-dimensional convolution kernel. Then:

$$Z^p = W^p \otimes X + b^p = \sum_{d=1}^{D} W^{p,d} \otimes X^d + b^p \tag{3}$$

$$Y^p = f(Z^p) \tag{4}$$

where W^p is the three-dimensional convolution kernel, $f(\cdot)$ is the non-linear activation function, b^p is the scalar bias, and Z^p is the net input of the convolution layer.

3) Pooling Layer Calculation

Pooling computation involves two types: max pooling and average pooling. For a pooling layer input feature map group $X \in \mathbb{R}^{D \times W \times H}$ each feature map $X^d \in \mathbb{R}^{W \times H}$ is partitioned into random regions $R_{w,h}^d$. Then Max Pooling (taking the maximum activation value of all neurons in the region as the region representation):

$$y_{w,h}^d = \max x_i, i \in R_{w,h}^d \tag{5}$$

Average Pooling (taking the average value of all neurons in the region):

$$y_{w,h}^d = \frac{1}{\left|R_{w,h}^d\right|} \sum_{i \in R_{w,h}^d} x_i \tag{6}$$

Where x_i represents the activation value of each neuron in the region $R_{w,h}^d$.

4) Auxiliary Classifier

To address issues such as gradient vanishing or overfitting caused by excessively large neural network parameters, the Inception network introduces auxiliary classifiers to propagate gradients during forward propagation. The auxiliary classifier consists of a 5X5 average pooling layer with a stride of 3, a 1X1 convolutional layer with a stride of 1, two fully connected (FC) layers, and a softmax function. The auxiliary classifier utilizes the output of an intermediate layer for classification, thereby facilitating model fusion.

5) Softmax Regression

Also known as multinomial or multi-class logistic regression. For multi-class problems where the class labels y can take C different values $y \in \{1,2,\cdots,C\}$, given a sample x, the conditional probability of belonging to class c according to Softmax regression is:

$$p(y = c|x) = softmax(\omega_c^T x) \tag{7}$$

$$= \frac{e^{\omega_c^T x}}{\sum_{c'}^{C} e^{\omega_{c'}^T x}} \tag{8}$$

where ω_c is the weight vector for class c, and the vector representation of softmax is:

$$softmax(\omega_c^T x) = \frac{e^{\omega_c^T x}}{\sum_{c'}^{C} e^{\omega_{c'}^T x}} \tag{9}$$

In vector form, it can be written as:

$$\hat{y} = softmax(W^T x) \tag{10}$$

$$= \frac{e^{W^T x}}{1_C^T e^{W^T x}} \tag{11}$$

where $W = [\omega_1, \cdots, \omega_c]$ is a matrix composed of weight vectors for C classes, 1_C is a C dimensional vector of ones, $\hat{y} \in R^C$ is a vector composed of predicted conditional probabilities for all

ISSN 2616-5775 Vol. 7, Issue 5: 130-135, DOI: 10.25236/AJCIS.2024.070517

classes, and the c-th element is the predicted conditional probability for class c.

4.2 Evaluation Metrics

To assess the classification performance, we employ "train_cs" and "Void_cs" to denote the categorical scores on the training and validation sets, respectively. We use the Mean Absolute Error (MAE) loss to measure the average absolute error between the model-predicted ages and the actual ages.

5. Results and discussion

5.1 Dataset

We envision testing our model on Morph-II datasets, but recent research suggests that the FGNet database's small size is not conducive to algorithmic improvements. Therefore, we opt to utilize the Morph-II dataset.

The Morph-II dataset [11] comprises 55,134 facial images from 13,617 subjects aged 16 to 77 years. Each subject has at least six samples. Considering that facial samples are influenced by factors such as expressions, skin color, lighting conditions, resolution, and background, most facial samples exhibit variations. Each facial image is labeled with the corresponding subject's actual age.

5.2 Experiment Result

Our Age-InceptionNet was first pretrained on the ImageNet dataset and then trained on the Morph-II dataset using an Nvidia GeForce 3060 graphics card, achieving a mean absolute error of 3.33.

In this section, we compare our Age-InceptionNet network with existing methods for age prediction and find that Age-InceptionNet demonstrates excellent performance on the Morph-II dataset, as shown in the table 1 below. Different age estimation methods yield different MAE results, with our proposed Age-InceptionNet achieving the best result. A comparison with three methods listed in the table 1 is provided:

- CSOHR, although effectively utilizing the spatial temporal information of images and hierarchical residual learning, lacks the strong feature extraction and fusion capability of the Inception structure.

- SqueezeNet, while adopting Fire modules to reduce the number of input channels of feature maps through the Squeeze layer and 3x3 convolutional kernels, may suffer from performance loss due to reduced parameterization compared to larger, more complex models.

- CNN+ELM, despite reducing the risk of overfitting with a large amount of training data, suffers from weaker feature extraction capabilities with ELM and may not fully exploit the complex features in images. Additionally, its relatively shallow neural network layers may degrade model performance.

Compared to the aforementioned methods, our Age-InceptionNet surpasses them in both the depth of neural network layers and the composition of the neural network structure. Age-InceptionNet consists of 9 Inception modules, totaling 22 neural network layers, extensively employing 1x1 convolutional kernels to reduce channel numbers and model complexity. It comprehensively utilizes 3x3 and 5x5 convolutional kernels for feature extraction and fusion, resulting in higher model performance and parameter efficiency than the previous three networks.

Table 1: Comparison of mean absolute error of our method with various deep neural network architectures.

Method	MAE
CSOHR	3.82
CNN+ELM	4.03
SqueezeNet	3.77
Age-InceptionNet	3.33

Age-InceptionNet enhances model expressiveness by combining the Inception network with the MLP. The MLP layer performs complex combinations and transformations on advanced features, improving prediction capabilities. The Inception network cleverly merges 1x1 convolutional kernels with 3x3 and 5x5 convolutional kernels, enhancing both parameter calculation efficiency and feature extraction and fusion capabilities. Multiple invocations of the Inception network combined with the MLP enhance the

ISSN 2616-5775 Vol. 7, Issue 5: 130-135, DOI: 10.25236/AJCIS.2024.070517

fusion generalization capability of facial features, leading to superior performance and computational resource utilization compared to the previous three models.

6. Conclusions

We propose a neural network structure named Age-InceptionNet for age estimation. Age-InceptionNet optimizes feature extraction, fusion, and age prediction by combining the Inception structure with a multilayer perceptron (MLP). Our experiments on the Morph-II dataset yielded excellent results with a mean absolute error of 3.33, demonstrating the effectiveness of Age-InceptionNet. The experimental results validate that sufficient feature extraction and fusion can effectively enhance the performance and accuracy of predictive models. This work provides new insights and methods for addressing challenges such as data bias and model robustness in the field of facial age prediction. In future work, we will further explore the performance of Age-InceptionNet on the LAP-2016 dataset and continue to optimize it to achieve the best model parameters and performance.

References

[1] Shou Y, Meng T, Ai W, et al. A comprehensive survey on multi-modal conversational emotion recognition with deep learning[J]. arXiv preprint arXiv:2312.05735, 2023.
[2] Shou Y, Cao X, Meng D. Masked Contrastive Graph Representation Learning for Age Estimation[J]. arXiv preprint arXiv:2306.17798, 2023.
[3] Meng T, Shou Y, Ai W, et al. A multi-message passing framework based on heterogeneous graphs in conversational emotion recognition[J]. Neurocomputing, 2024, 569: 127109.
[4] Meng T, Shou Y, Ai W, et al. Deep imbalanced learning for multimodal emotion recognition in conversations [J]. arXiv preprint arXiv:2312.06337, 2023.
[5] Shou Y, Cao X, Meng D, et al. A Low-rank Matching Attention based Cross-modal Feature Fusion Method for Conversational Emotion Recognition [J]. arXiv preprint arXiv:2306.17799, 2023.
[6] Shou Y, Ai W, Meng T. Graph information bottleneck for remote sensing segmentation[J]. arXiv preprint arXiv:2312.02545, 2023.
[7] Edmonds, Emily, C. Cognitive Mechanisms of False Facial Recognition in Older Adults.[J]. Psychology & Aging, 2012.DOI:10.1037/a0024582.
[8] Shou Y, Ai W, Meng T, et al. Czl-ciae: Clip-driven zero-shot learning for correcting inverse age estimation [J]. arXiv preprint arXiv:2312.01758, 2023.
[9] Ai W, Shou Y, Meng T, et al. Der-gcn: Dialogue and event relation-aware graph convolutional neural network for multimodal dialogue emotion recognition[J]. arXiv preprint arXiv:2312.10579, 2023.
[10] Shou Y, Meng T, Ai W, et al. Adversarial representation with intra-modal and inter-modal graph contrastive learning for multimodal emotion recognition[J]. arXiv preprint arXiv:2312.16778, 2023.
[11] Shou Y, Meng T, Ai W, et al. Conversational emotion recognition studies based on graph convolutional neural networks and a dependent syntactic analysis[J]. Neurocomputing, 2022, 501: 629-639.