ISSN: 2664-9640

Face attribute recognition based on cascaded network

Linlin Luo*

Southwest Minzu University, Chengdu 610000, China.

Abstract

With the development of deep learning, important progress has been made in the field of face recognition. As an important feature of the body, the face contains rich facial information, but the current face attributes recognition is mostly for single attribute. For this reason, we propose a cascaded network model for identifying multiple attributes of human faces. In the experiment, data augmentation technology is used to improve the generalization ability of the model. And using the transfer learning technology, the Xception pre-training weight is introduced in the experiment for feature extraction, and a new classification layer is established on the output to complete the attribute recognition. Experiments show that this method has achieved decent accuracy on the three attributes: Male, Smiling, and Young.

Keywords

CelebA dataset, data enhancement, transfer learning.

1. Introduction

Face analysis has been a research hotspot in the field of computer vision. The research goal is to extract as much information as possible from a face, such as identity, gender, Age and expression. Many semantics of face attributes are relatively intuitive, and we can better describe the object through feature extraction, so as to better understand the target. In recent years, face detection technology based on deep learning has been greatly developed. Facetess [1] trained a series of CNN for face attribute recognition to detect partially occluded faces. Li et al. [2] developed a cascade structure based on CNN with powerful detection capabilities. Qin et al. [3] proposed joint training of cascaded neural networks to achieve end-to-end optimization.

For face gender recognition, Jiang et al. [4] uses a deep network model that combines high-level facial feature learning and low-level feature learning for gender recognition, which has good learning and generalization capabilities; Shi et al. [5] proposed L-MFCNN model for face recognition, which is based on a convolutional neural network combines multi-layer feature fusion and an adjustable supervision function mechanism. Dong et al. [6] adopts a combination of deep learning and random forest for gender recognition, and has high accuracy in recognition of facial images with complex lighting and posture changess.

For face age estimation, Rothe et al. [7] used the probability value of the softmax classifier to perform a weighted average for age estimation. Chen et al. [8] considered the order relationship between different ages and proposed a ranking-CNN for facial age estimation. Han et al. [9] used the improved AlexNet to construct a multi-task learning method for joint estimation of multiple attributes, including shared feature learning and attribute group-specific feature learning.

For facial expression recognition, Zhang et al. [10] proposed a cross-domain facial expression recognition method based on sparse subspace transfer learning. Su Zhi et al. [11] proposed a model based on multi-scale bilinear pooling neural network, which solves the difficulty of facial expression recognition due to subtle inter-class differences and significant intra-class changes in facial expressions, which leads to the recognition rate Low problem. Yin et al. [12] proposed a lightweight facial expression recognition method based on convolutional attention.

ISSN: 2664-9640

This paper is mainly based on cascaded neural networks, and studies the issues related to face attributes, including gender, expression and age. The main contributions of this paper are: (1) Data enhancement operations are performed on the training data to make the data as rich and diverse as possible, thereby improving the generalization ability of the model. (2) Use transfer learning methods to strengthen the ability to extract features.

The structure of the paper is as follows. The second section introduces the data set used in the experiment. The third section is an introduction to the experimental model. The fourth section is the analysis of the experimental results. The fifth section is a summary.

2. Dataset

This article uses the CelebA dataset [13], which is an open source large-scale face detection benchmark dataset of the Chinese University of Hong Kong. It contains 202,599 face images of 10,177 celebrities. The images in this dataset cover large pose changes and background clutter. Each image has 40 attribute annotations, such as Blad, Chubby, Male, Smiling, Young and other features. It contains 118,165 face pictures for women and 138704 face pictures for men. Some pictures are randomly selected from the CelebA dataset, as shown in Figure 1.

Figure1. CelebA dataset

Each picture has its corresponding attribute. By combining the image with the attribute file in the CelebA data set, the attribute corresponding to each image can be obtained. In the experiment, we only recognize the three attributes: Male, Smiling, and Young in the CelebA data set. The properties file is shown in Figure 2. We randomly select an image and visualize its properties, as shown in Figure 3.

image_id	partition	Male	Smiling	Young
000001.jpg	0	0	1	1
000002.jpg	0	0	1	1
000003.jpg	0	1	0	1
000004.jpg	0	0	0	1
000005.jpg	0	0	0	1

Figure2. The attributes corresponding to the face image

ISSN: 2664-9640

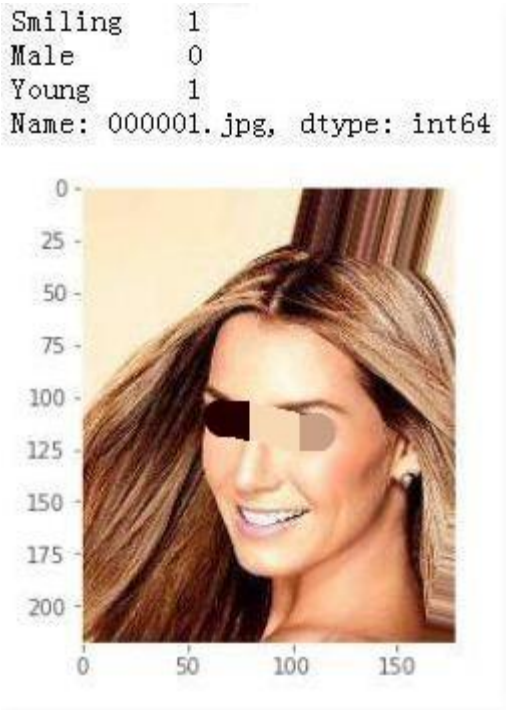

Figure3. The visualization of face attributes

3. Model introduction

In the experiment, we first preprocess the data, then put the preprocessed data into the pre-training model for feature extraction, and then fine-tune the pre-training model to finally recognize the face attributes. The model structure in the experiment is shown in Figure 4.

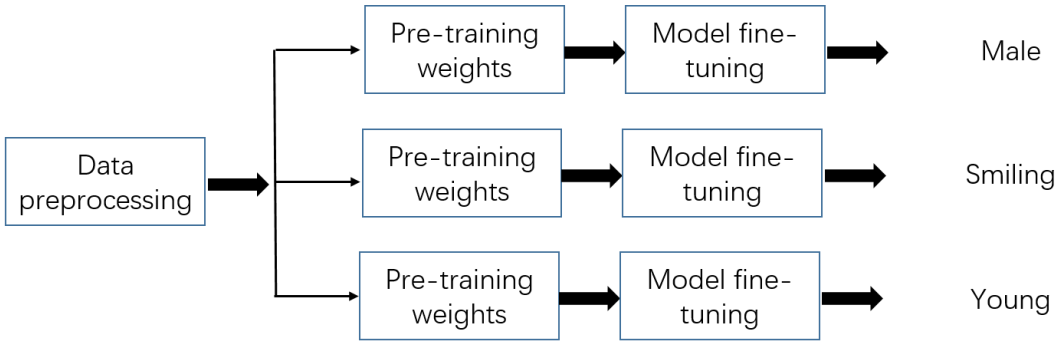

Figure 4. Model structure

3.1. Data enhancement

Data enhancement can solve the problem of few samples or unbalanced sample categories in the dataset . When preprocessing the dataset, the Image Data Generator class in keras was used in the experiment to enhance the data. Using this method can make the model generate pictures while training, so as to improve efficiency. Secondly, the generated image can increase the generalization ability and recognition accuracy of the model. Data enhancement methods, we use translation, random rotation, scale transformation and other methods, the generated image is shown in Figure 5.

ISSN: 2664-9640

Figure 5. Data enhancement

3.2. Transfer learning

Migration learning is a new machine learning method developed for image recognition due to the small amount of dataset that cannot make the training task of the deep model from scratch. Migration learning puts the trained model into a new classification task to perform image recognition again, and recognizes the original features learned before as features of the new data again. The workflow is to first instantiate the basic model and load the pre-trained weights into it, then freeze all the layers of the basic layer in the model, and then create a new fully connected layer on top of the output of one or more layers of the basic model or the classifier layer, and finally train a new model on the new dataset.

3.3. Xception model

Convolutional neural network (CNN) [14] consists of a series of convolutional layers, activation layers, pooling layers and fully connected layers. The convolution process extracts different features of the input signal. Each convolution kernel extracts a single feature on the entire feature map. Multi-core convolution makes the features fully extracted. The extracted features are then passed as input to the next layer. These features progress from low-level to high-level. Therefore, the deep network structure makes the learned features more global. In order to solve the problem of too many parameters, the Inception module in the Inception v3[15] network uses a multi-branch structure to increase the network width, adds 1×1 convolution to the branches to reduce the number of channels involved in the operation, and uses global average pooling Instead of the fully connected layer, as shown in Figure 6.

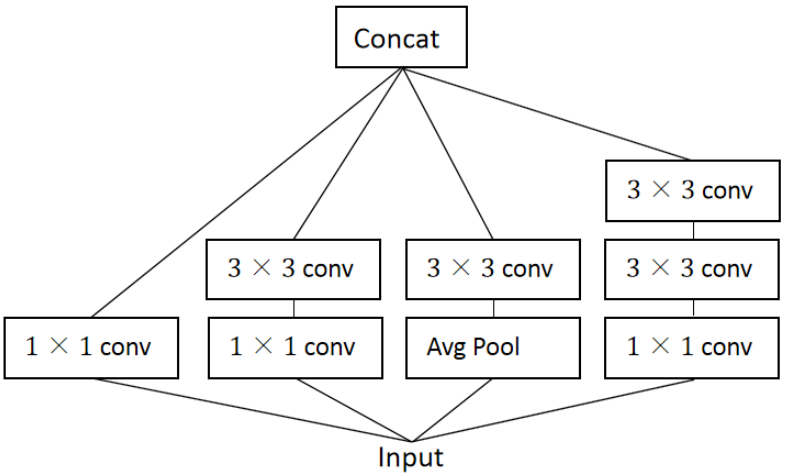

Figure 6. A cannonical Inception module(Inception V3)

ISSN: 2664-9640

The Xception model [16] is an improved model based on the InceptionV3 model, which replaces the Inception module with a deep separable convolution, called Xception. The composition is mainly composed of residual network and depth separable convolution. Xception contains 36 convolutional layers, divided into 14 blocks, and the middle 12 blocks all contain linear residual connections. At the same time, the model refers to the characteristics of depth separable convolution [17], and performs spatial layer-by-layer convolution on each channel of the input data independently, and then performs point-by-point convolution on the result. The Xception model structure shown in Figure 7 is between ordinary convolution and the above separable convolution. The first step is to separate channels through 1×1 convolution, and the second step is to independently plot the spatial correlation of each output channel. Use 3×3 to process separately, and finally merge.

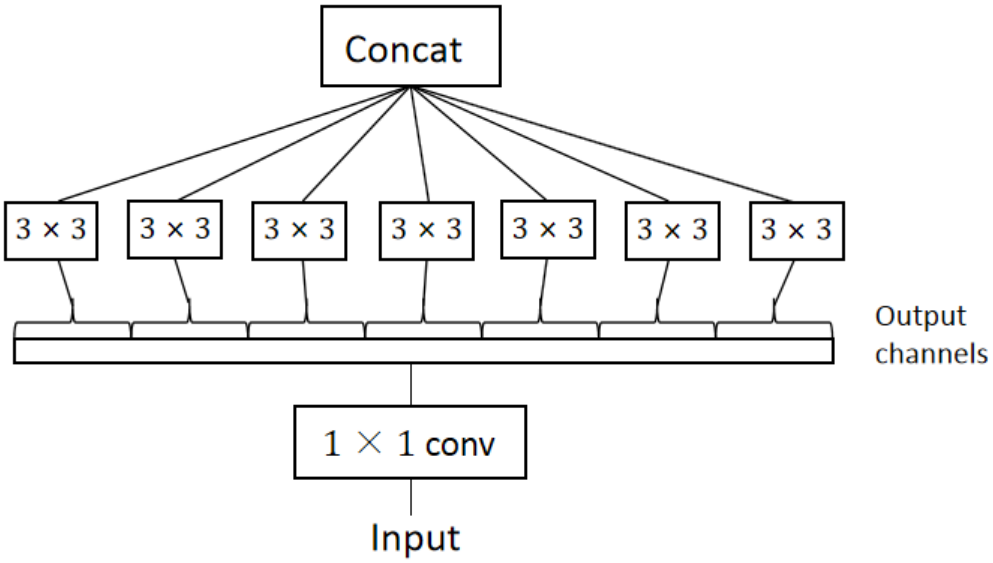

Figure 7. Xception basic module

4. Experimental results and analysis

4.1. Experimental environment

In the experiment, the model training was carried out on a PC equipped with a 3.6 GHz Intel Core processor. The programming language used in the experiment is Python, and the deep learning framework used is Tensorflow.

4.2. Experimental results

In the experiment, we first divide the data set into training set, validation set and test set. The training set is used to fit the model to complete the training process of the model. The purpose of the validation set is to find the best model. The test set is to test the performance of the obtained model. In the experiment, we use Adam as the optimizer to continuously reduce the loss value, where the learning rate is set to 0.01 and the decay rate parameter is set to 5e-4. In the experiment, the training batch size is 32. After 50 training sessions, the test accuracy rates of 94.5%, 93%, and 82.4% were obtained on the three attributes of Male, Smiling, and Young faces in the CelebA dataset. As shown in Table 1-3, it can be seen that our model in this paper has significantly improved the testing accuracy.

Table 1. Accuracy comparison of age recognition on CelebA dataset

Algorithm	Test accuracy
Cost-sensitive[18]	75%

ISSN: 2664-9640

PANDA-w[19]	77%
DeepID2[20]	76%
Triplet-knn[21]	75%
MT-RBM PCA[22]	81%
Ours	82.4%

Table 2. Accuracy comparison of gender recognition on CelebA dataset

Algorithm	Test accuracy
Cost-sensitive[18]	93%
PANDA-w[19]	93%
DeepID2[20]	94%
Triplet-knn[21]	91%
MT-RBM PCA[22]	90%
Ours	94.5%

Table 3. Accuracy comparison of smiling recognition on CelebA dataset

Algorithm	Test accuracy
FaceTracer[23]	89%
PANDA-w[19]	89%
PANDA-1[24]	92%
Ours	93%

5. Conclusion

This paper uses the Xception pre-training model based on migration learning to extract features from the data, and add a classification layer to the output layer for face attribute recognition. In the data preprocessing stage, data enhancement technology is used to strengthen the generalization ability of the model. On the CelebA data set, the three attributes of Male, Smiling, and Young faces are recognized by this model, and the test accuracy rates are respectively 94.5%, 93%, and 82.4%. However, due to the choice of cascade, the recognition of face attributes will cause a large amount of parameters, so the next research is to use a lightweight model to recognize face attributes.

References

[1] Yang S, Luo P, Loy C C, et al. From Facial Parts Responses to Face Detection: A Deep Learning Approach. 2015 IEEE International Conference on Computer Vision (ICCV). IEEE, 2015:3676-3684.
[2] Li H, Lin Z, Shen X, et al. A convolutional neural network cascade for face detection. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2015:5325-5334.
[3] Qin H, Yan J, Li X, et al. Joint Training of Cascaded CNN for Face Detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016:3456-3465.
[4] Lin Feng, Zhang Lei, and Liang Mingliang. Design of HEVC encoding and transmission system based on NVIDIA Jetson TX1. Journal of Shenyang Aerospace University 035.005(2018):51-56.

[5] Duanzhen Qin. Gstreamer-based streaming media player design under Linux. Science Popular (Science Education) 03(2014):149-149.
[6] NVIDIA Corporation. NVIDIA Jetson. The AI Platform for Autonomous Machines. https:// developer.nvidia.com/embedded/develop/hardware.Html, 2018-05-10.
[7] Rothe R, Timofte R, Van Gool L. Deep Expectation of Real and Apparent Age from a Single Image Without Facial Landmarks. International Journal of Computer Vision, 2016:1-10
[8] Chen S, Zhang C, Dong M, et al. Using Ranking CNN for Age Estimation. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017:5183-5192.
[9] Han H, K. Jain A, Shan S, et al. Heterogeneous Face Attribute Estimation: A Deep Multi-Task Learning Approach. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(11): 2597-2609.
[10] Zhang Wenjing, et al. Cross-domain facial expression recognition based on sparse subspace transfer learning. Data collection and processing 36.01(2021):113-121. doi:10.16337/j.1004-9037. 2021.01.011.
[11]Su Zhiming, Wang Lie, and Lan Zhengjie. "Fine-grained facial expression recognition based on multi-scale hierarchical bilinear pooling network." Computer Engineering. ():. doi:10.19678/j.issn.1000-3428.0060133.
[12] Yin Pengbo, Pan Weimin, and Zhang Haijun. "Lightweight facial expression recognition method based on convolutional attention." Progress in Laser and Optoelectronics .():. doi:10.3788 /lop58.1210023.
[13] Liu Z, Ping L, Wang X, et al. Deep Learning Face Attributes in the Wild. IEEE International Conference on Computer Vision. IEEE, 2016.
[14] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Proceedings of the 25th International Conference on Neural Information Processing System, 2012: 1097-1105.
[15] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. arXiv preprint arXiv:1512.00567, 2015.
[16] Chollet F. Xception: Deep Learning with Depthwise Separable Convolutions. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017.
[17] Howard, A. G., et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. (2017).
[18] He H, Garcia E A. Learning from imbalanced data. IEEE Transactions on Knowledge & Data Engineering, 2008 (9): 1263-1284.
[19] Zhang N, Paluri M, Ranzato M A, et al. Panda: Pose aligned networks for deep attribute modeling. Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 1637-1644.
[20] Sun Y, Chen Y, Wang X, et al. Deep learning face representation by joint identification- verification. Advances in neural information processing systems. 2014: 1988-1996.
[21] Zhang N, Paluri M, Ranzato M A, et al. Panda: Pose aligned networks for deep attribute modeling. Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 1637-1644.
[22] Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015:815-823.
[23] Kumar, N., PN Belhumeur, and SK Nayar. FaceTracer: A Search Engine for Large Collections of Images with Faces. Computer Vision-ECCV 2008, 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part IV Springer-Verlag, 2008.
[24] Zhang, N., et al. PANDA: Pose Aligned Networks for Deep Attribute Modeling. (2013).