JD AI Fashion Competition

line 1:Lijuan Ren [1st]
line 2: *Shandong University*
line 3: *School of Control Science and Engineering,*
line 4: *city of Jinan，China;*
line 5: 892276613@qq.com

line 1: Peng Song[2nd]
line 2: *Shandong University*
line 3: *School of Control Science and Engineering,*
line 4: *city of Jinan，China;*
line 5:631059153@qq.com

line 1: Peng yuan[3rd]
line 2: *Southeast university*
line 3:School of Information Science and Engineering
lie 4: *city of Nanjig，China;*
line 5: 736814652@qq.com

line 1: Ke tian[4th]
line 2: *Nagoya university*
line 3:School of Information Science and Engineering
line 4: *city of Nagoya，Japan;*
line 5: 285662034@qq.com

Abstract—An optimized Inception_ResnetV2 deep network model is proposed to solve the problem of difficult identification of apparel attributes. Firstly, create pictures of different clothing attributes. The ten thousand data set of images are established. Then the data is trained by the method of pre-training model. After the convolution layer, the global average pooling replaces the traditional full-link layer to connect the classifier, so as to reduce the over-fitting problem caused by the training model. What's more, the image data is enhanced. At the same time, Adam optimizer with high efficiency in local deep learning was adopted to accelerate the convergence speed of the model. It realizes the identification of different kinds of clothing attributes. In order to prevent the randomness of the experiment, the average value of the results of multiple experiments were adopted. The experimental results showed that compared with deep learning network models such as Inception_ResnetV2 and InceptionV4,Xception had a higher recognition rate. It has stronger ability to identify clothing attributes.

***Index Terms*—Costume attribute recognition; Neural network; Deep learning; Optimizer.**

I. INTRODUCTION

In recent years, the shopping online has been increasingly penetrating into people's life. Shopping websites such as jingdong, taobao and Vipshop have gradually become people's shopping platform.As clothing is the main sales product, it is of great importance for users to find their favorite clothes in the huge clothing products.The current classification of clothing attributes mostly adopts manual labeling of images, resulting in huge manpower.Moreover, they are highly subjective.The classification of clothing attributes is not very effective at present, and it is not easy for users to find the exact type of clothing.In order to better serve consumers, the targeted search of clothing can be conducted by searching keywords, such as an element of clothing, so as to meet the needs of consumers.However, in the recognition of clothing attributes, the diversity of clothing background causes certain interference to the recognition of clothing attributes.The visual difference of the same attribute is large.In order to solve this problem, many scholars have carried out a lot of research. Through the training of the visual features of clothing images, a method for automatic classification of clothing attributes is proposed. According to the classification of clothing, the literature [1] proposed the method of clothing region positioning and the association matrix to predict the clothing style.It works better than a style finder, but different categories don't. The literature [2] proposed to use deep convolutional neural network to extract convolution features and classify apparel attributes. The literature [3] adopted Girshick's RegionCNN (r-cnn) object detection model [4] to detect fashion items worn by individuals. The article [5] extracts the underlying features in an adaptive manner and combines the learning attribute classifier of complementary features.The prediction of the independent classifier is further improved by exploring the inter dependency among attributes through conditional random fields.

As for above factors, such as different dressing scenes, different regions, etc., which are not easy to be resolved, this paper uses the deep neural network of xception for type resolution, which achieves good results.

II. DATA TRAINING

A. Data download

Use the urllib function to perform the crawler to get more than 50,000 pictures for training.

B. image preprocessing and data enhancement

Because the pre-training model is used, the input image is redefined to be 299*299*3 to maximize the recognition ability of the graph.Make the data to be trained as close as possible to the original data.Where 299 represents the length and width, and 3 represents the RGB tee of the image.In addition, in order to enhance the generalization ability of the model, the enhancement of the discrimination ability of the specific information [8], first of all to random flip of the given data set, translation, rotation and other data to enhance operation, larger due to the existing data and to prevent explosive growth, the amount of data by online enhanced way, first about the input model of the small batch data to perform the corresponding operation.

C. Inception_Resnetv2 network architecture

Deep learning has made new progress in image classification [9], and the recognition rate has been increasing, which has a good recognition effect.The article USES

Inception_Resnet V2 to identify the attribute categories of various clothing.The Inception_ResNet structure is proposed by Szegedy and is a hybrid of Inception and Resnet network structure.

The network structure model of the Inception series [10] is a model proposed by Google, which does not restrict its use of specific convolution kernel, but USES all different sizes of convolution kernel at the same time, and then stitches the resulting matrix together.In order to reduce computation, 1*1 convolution kernel is added before 3*3 junction, which reduces the number of convolution kernel channels and reduces the dimension.In the stem structure of InceptionV4, multiple convolution plus 2 times for pooling were adopted for the pooling. The structure of convolution and pooling mentioned in the thesis of Inception-v3[7] was adopted for the pooling to prevent bottlenecks.The stem then used three different Inception modules, a total of 14.

Resnet is Kaiming proposed in ImageNet as deep as 152 layers of residual neural network structure [11], to the problem requires the deeper the more complex neural network, but the deeper the depth, can cause a fitting, gradient dispersion explosion problems such as [12] and gradient, residual structure of neural network is introduced into the shortcut [13], transfer the input layer and the convolution results together, reduced because of the depth of neural networks across deep brings the problems such as the gradient of dispersion.Inception-ResnetV2 is the addition of Resnet to inceptionV4.The concrete architecture of its model is shown in figure 2-1 below.It is divided into input blocks, such as Inception- resnet-a, Inception- resnet-b, Inception- resnet-c and reduce-b [14].The combined network structure is improved in both accuracy and computational efficiency.It is equal to the calculation of inceptionV4[15], but it has a higher accuracy rate.

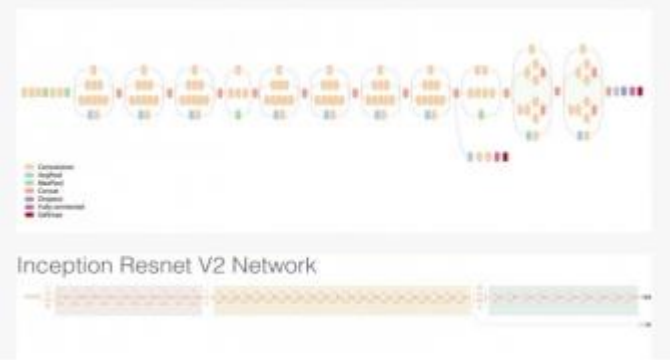


FIG. 2-1 network structuring of Inception_resnetV2

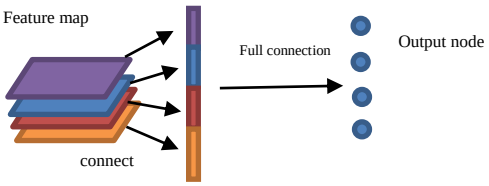
III. MODEL OPTIMIZATION

The traditional full connection layer [16] is shown in FIG. 2-2(a). Its function is to extend the feature map obtained by the last convolutional layer into a vector [17]. The calculation formula of the feature map is as follows:

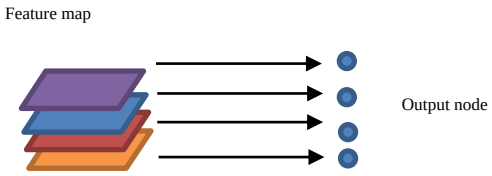
$$f_{i,j,k} = \max(w_k^T x_{i,j}, 0) \tag{1}$$

Where (i, j) represents the position index of image pixels, k represents the index of extracting feature graph. $x_{i,j}$ represents the image block in the convolution window, and w represents the weight value.

Of this vector multiplication, finally reduce the dimension of input to the sigmoid classifier in the corresponding probability value, but the connection layer number is too large, cause parameter redundancy, but the global average pooling as shown in figure 2-2 (b), the last layer of the characteristics of the figure for an average of the whole picture, to form a feature point, these feature points of the final characteristic vector, then as a result, directly into the sigmoid classifier.The advantage is that there are no parameters in the global average pooling and overfitting is avoided.



(a) Full Connection Layer



(b) Global Average Pooling

FIG. 2-2 full connection layer and global average pooling

Binary crossentropy loss function [18] was used in the selection of loss function, and the calculation formula is as follows:

$$y - y*t + \log(1+\exp(-y)) \tag{2}$$

Where, is the predicted value, is the target value, and the formula is transformed as follows to prevent overflow:

$$\max(y,0) - y*t + \log(1+\exp(-abs(y))) \tag{3}$$

In order to improve the learning rate of machine learning, to speed up the model convergence speed, choose Adam optimizer, he is an alternative to traditional stochastic gradient descent [19] first-order optimization algorithm, compared with the stochastic gradient algorithm, stochastic gradient descent to keep a single vector, update all the weights vector does not change during the process of training, and Adam optimizer is estimated by computing the gradient a moment and second moment estimation [20] and independent for different parameters design of adaptive vector.The formula is as follows:

$$m_t = \mu * m_{t-1} + (1-\mu) * g_t \tag{4}$$

$$n_t = v * n_{t-1} + (1-v) * g_t^2 \tag{5}$$

$$\hat{m}_t = \frac{m_t}{1-u^t} \tag{6}$$

$$\hat{n} = \frac{n_t}{1-v^t} \tag{7}$$

$$\Delta\theta_t = -\frac{\hat{m}}{\sqrt{\hat{n}_t + \varepsilon}} * \eta \tag{8}$$

Where m_t , n_t respectively for gradient first order moment estimation and second order moment estimation, can be regarded as the estimation of the expectation, $\hat{m}$, $\hat{n}$ is the correction of m_t , n_t . μ is the momentum factor and η is the learning rate. g_t is the gradient of hi .This approximates an unbiased estimate of expectations.It can be seen that direct torque estimation of gradient has no additional requirement on memory and can be dynamically adjusted according to gradient. $-\frac{\hat{m}}{\sqrt{\hat{n}_t + \varepsilon}}$ which forms a dynamic constraint on learning rate and has a clear range.

It can be seen from table 4-1 in the experimental results below that the Adam optimizer is more effective than the SGD optimizer.

Taking Mean Average Precision as the index of model evaluation [21], the expression is as follows:

$$\mathrm{m}AP = \frac{1}{|\mathrm{Q_R}|}\sum_{\mathrm{q}\in \mathrm{Q_R}} AP(\mathrm{q}) \tag{9}$$

Where P is the accuracy rate and R represents the recall rate [22].The value of P is the ratio of the predicted correct attribute value to the total predicted attribute value, and AP is the statistical average value of them.MAP is the AP synthesis calculated from all attribute dimensions, and their average value is calculated.

Experiment and result analysis

IV. Experiment and result analysis

A.Experimental platform

The experimental platform configuration is trained on the data by 3 blocks of nvidia's 1080ti GPU in parallel, and batchsize is set to 96.Ubuntu is 16.04 with 110G of ram, 1.4.1 for tesorflow-gpu and 3.5.2 for python.

B.Taining process

The training method adopts the method of transfer learning.First, all the parameters of the convolution layer are locked, followed by global average pooling, followed by rotation, the parameters dropout are set to 0.5.Finally connect the sigmoid classifier.The initial learning rate was 1e-4, and when the training was to 10 epochs, the fine-tuning method of unlocking part of convolution layer was adopted to continue the training to 20 epochs.And adjust the learning rate to 1e-5.In order to prevent randomness in the training process, the optimal parameter model was adopted after repeated training [23].Finally, the data set of the test set is inverted, translated and other operations are carried into the model to estimate the results, and the average is finally taken.

C. Experimental results

Since the sigmoid classifier was used to obtain the probability value, we used the probability value of any one of these 13 classes is greater than 0.5,we let his label be 1 and the final online result is 0.6076.In the end, we found that compared with other models, Inception_resnetV2 has the best effec.

V. Conclusion

Inception_resnetV2, Xception and InceptionV4 deep network models were used to train different kinds of clothing attributes. Finally, it was found that Inception-ResnetV2 performed better in the time of training and the accuracy of recognition.In the selection of the optimizer, Adam is faster than SGD, and the accuracy of test set is improved.Therefore, the deep network model based on Inception_ResnetV2 can well realize the recognition of clothing attributes.

References

[1]Sun G L, Wu X, Chen H , et al. Clothing style recognition using fashion attribute detection[C]//Proceedings of the 8th International Conference on Mobile Multimedia Communications. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2015: 145-148.

[2]厉智, 孙玉宝, 王枫, 等. 基于深度卷积神经网络的服装图像分类检索算法[J]. 计算机工程, 2016, 42(11): 309-315.

[3]Sun Y, Liu Q. Attribute Recognition from Clothing Using a Faster R-CNN based Multi-task Network[J]. International Journal of Wavelets, Multiresolution and Information Processing, 2018.

[4]Gao T, Packer B, Koller D. A segmentation-aware object detection model with occlusion handling[C]//Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011: 1361-1368.

[5]Chen H, Gallagher A, Girod B. Describing clothing by semantic attributes[C]//European conference on computer vision. Springer, Berlin, Heidelberg, 2012: 609-623.

[6]Erhan D, Bengio Y, Courville A, et al. Why does unsupervised pre-training help deep learning?[J]. Journal of Machine Learning Research, 2010, 11(Feb): 625-660.

[7] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2818-2826.

[8]刘茜, 卢心红, 李象霖. 基于多尺度 Retinex 的自适应图像增强方法[J]. 计算机应用, 2008, 29(08): 2077-2079.

[9]Perronnin F, Sánchez J, Mensink T. Improving the fisher kernel for large-scale image classification[C]//European conference on computer vision. Springer, Berlin, Heidelberg, 2010: 143-156.

[10]Mitchell A J, Shiri-Feshki M. Rate of progression of mild cognitive impairment to dementia–meta-analysis of 41 robust inception cohort studies[J]. Acta Psychiatrica Scandinavica, 2009, 119(4): 252-265.

[11]Targ S, Almeida D, Lyman K. Resnet in Resnet: generalizing residual architectures[J]. arXiv preprint arXiv:1603.08029, 2016.

[12]Mitra P P, Halperin B I. Effects of finite gradient-pulse widths in pulsed-field-gradient diffusion measurements[J]. Journal of Magnetic Resonance, Series A, 1995, 113(1): 94-101. Sun G L, Wu X, Chen H H, et al.
[13]He K, Zhang X, Ren S, et al. Identity mappings in deep residual networks[C]//European Conference on Computer Vision. Springer, Cham, 2016: 630-645.
[14]Hasani B, Mahoor M H. Spatio-temporal facial expression recognition using convolutional neural networks and conditional random fields[C]//Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on. IEEE, 2017: 790-795.
[15] Sevilla A, Bessonne L, Glotin H. Audio bird classification with inception-v4 extended with time and time-frequency attention mechanisms[J]. Working Notes of CLEF, 2017, 2017.
[16]Blelloch, Guy E., and Charles R. Rosenberg. "Network Learning on the Connection Machine." IJCAI. Vol. 87. 1987.
[17]Vedaldi A, Zisserman A. Efficient additive kernels via explicit feature maps[J]. IEEE transactions on pattern analysis and machine intelligence, 2012, 34(3): 480-492.
[18]De Boer P T, Kroese D P, Mannor S, et al. A tutorial on the cross-entropy method[J]. Annals of operations research, 2005, 134(1): 19-67.
[19] Bottou L. Large-scale machine learning with stochastic gradient descent[M]//Proceedings of COMPSTAT'2010. Physica-Verlag HD, 2010: 177-186.
[20]Blundell R, Bond S. Initial conditions and moment restrictions in dynamic panel data models[J]. Journal of econometrics, 1998, 87(1): 115-143.
[21]Yilmaz E, Aslam J A. Estimating average precision with incomplete and imperfect judgments[C]//Proceedings of the 15th ACM international conference on Information and knowledge management. ACM, 2006: 102.
[22]Litherland J C, Evans A J, Wilson A R M. The effect of hormone replacement therapy on recall rate in the National Health Service Breast Screening Programme[J]. Clinical radiology, 1997,
[23]Chilimbi T M, Suzue Y, Apacible J, et al. Project Adam: Building an Efficient and Scalable Deep Learning Training System[C]//OSDI. 2014, 14: 571-582.