A Clothing Retrieval Model Based on DenseNet and Multi-Similarity Loss

Xu Feifei\textsuperscript{1,a}, Tian Yu\textsuperscript{2,b}, Zheng Shuting\textsuperscript{3}

\textsuperscript{1}School of Computer Science and Technology Shanghai University of Electric Power, Shanghai, China
\textsuperscript{2}School of Computer Science and Technology Shanghai University of Electric Power, Shanghai, China
\textsuperscript{3}School of Computer Science and Technology, Shanghai University of Electric Power, Shanghai, China
\textsuperscript{a}xufeifei@shiep.edu.cn
\textsuperscript{b}xufeifei1983@hotmail.com, tianyu@mail.shiep.edu.cn

\textbf{Abstract—}With the rapid development of deep learning in recent years, Deep Neural Network structure has been applied to all aspects of image algorithm, including image classification and recognition, target detection, image retrieval. Image retrieval is suffering from some challenges, such as crease, cover, background. Based on these challenges, using the local information of the image to replace the image is proposed which learns embedding vectors by local information to reduce noise interference. Inspired by the above discussion, a new deep learning model named IAPM (Integrated and Partial Model) is proposed in this paper. According to global information and local information, Joint training is carried out. For feature extraction, DenseNet is used to extract the feature vector. And the idea of transfer learning is applied to transform classification problems to retrieval problems. Meanwhile, the multi-task model is adopted to construct a framework, and Multi-Similarity Loss is used to pull the position of the embedded vector in feature space. Experimental results show its effectiveness.

\section{1. INTRODUCTION}

As a new research hotspot, clothing image retrieval has attracted more and more attention from researchers. With the rapid development of e-commerce, to erode the market share, customer experience has become a new field for competition. By improving customer experience, satisfying customers, E-commerce companies try to design a variety of new smart experiences to get their customer loyalty which is the main reason for their repeating purchasing. For example, people like to buy clothes in TaoBao. In the past, we could only input some keywords and select clothes in thousands of search results. With the new trends of clothing image retrieval technology, we can directly input the clothes picture we like, and the returned retrieval results will be more accurate. For customers, it will be more convenient and fast to find the clothes they want to buy. Clothing retrieval technology can also be applied to many different scenarios. Although the new technology has been attempted to use, it still needs to be improved since the accuracy is not high enough. The challenging problem has inspired lots of researchers to explore.

In the past, image retrieval was mainly based on text. It aims at the retrieval and matching of the image’s text information. Although the accuracy is high, it has an obvious defect, that it requires a lot of
manual annotation. Owing to a lot of manpower and material resources are needed, content-based image retrieval has been developed. Content-based image retrieval mainly deals with each pixel, and then uses the image’s pixel value to extract features, which can obtain the image semantics by its coding vector in another embedded space, and then look for similar images in this space.

In the image retrieval task, the focus of scholars is from the original traditional image algorithm SIFT to the increasingly popular convolution neural network algorithm\(^1\), which is a change from manual feature extraction to automatic feature extraction. Image retrieval can also be regarded as a process of metric learning. There are two main development directions for deep metric learning. One is to design new network structures, such as Siamese network, Triple network\(^{2,3}\). The other one is improvements for the loss function, such as contrast loss, center loss, multi-similarity loss\(^{4,5,6}\). These methods mainly push or pull feature vectors in different degrees and directions in the embedded space.

In this paper, a new network model named IAPM (Integrated and Partial Model) based on DenseNet\(^7\) is proposed, and the multi-similarity loss function proposed by Xun Wang\(^6\) is quoted. To show the effectiveness of our proposed model, in the experiments, the data stream is encoded into a specific feature map. From the experimental results, the two closest clothes can be found faster with our model which indicates our model's efficiency.

2. RELATED WORK

2.1. Convolutional Neural Network
Convolutional Neural Network (CNN) is a deep network that contains a series of convolution operations. It is one of the representative networks in the deep learning model. The common convolution neural networks generally include convolution layer, pooling layer, full connection layer, and activation layer. The main function of convolution operation is to extract feature information. Convolution neural network is widely used in image processing. However, with the expansion of the model scale, there will be a series of problems, such as gradient dispersion, model degradation. Recently, VGGNet, Inception (V1, V2, V3, V4), ResNet\(^{8-13}\), and other models are proposed based on the improvement of convolutional neural network.

2.2. DenseNet
As shown in figure 1,2,3, DenseNet\(^7\) is the best paper of CVPR2017, which is different from the Inception model\(^9\). The DenseNet model is inspired by the residual network model proposed by Kaiming He\(^{12}\). The lower layer features are not only affected by the upper layer features but also affected by the higher layer features. It can be regarded as a short-circuit connection. Compared with ResNet model, DenseNet model has more radical measures of dense connection. The characteristics of each layer are affected by all the above layers. This greatly improves the feature reuse, and alleviates the disappearance of gradient, and reduces the number of parameters.

![Figure 1. Inception structure](image-url)
2.3. **Multi-Similarity Loss**

In image classification or retrieval, there are always some samples to be easily misclassified, which is inevitable. Learning these difficult samples is a good method for model optimization. This can be achieved by designing a loss function. The loss function is a constraint condition of the model. In terms of retrieval, the loss functions based on sample pairs include the Comparative loss\[^4\] function for Siamese networks, Triplet loss\[^3\] function with three input models, Lifted Structure loss\[^4\]}, and Quadruplet loss\[^15\].

At CVPR 2019, the global top-level computer vision conference, \[^6\] proposed a framework called general sample pair weighting (GPW). It analyzes the loss functions based on sample pairs. And they also proposed a new loss function called Multi-Similarity Loss. The authors integrated most of the loss functions based on sample pairs and found the relationship between them. And then, they put forward three kinds of similarity: self-similarity (S), negative relative similarity (N), and positive relative similarity (P). Self-similarity is the calculation of self sample pairs, and the distance between positive sample and anchor is less than that between negative sample and anchor. Negative relative similarity refers to the influence of the distance between negative samples on similarity, and positive relative similarity refers to the influence of the distance between positive samples on similarity. The multi-
similarity loss function is a comprehensive consideration of these three directions. Table 1 compares the similarity of several loss functions (where MS represents multi-similarity loss function):

| similarity | Contrastive | triplet | Lifted structure | MS |
|------------|-------------|---------|------------------|----|
| S          | 1           | 0       | 0                | 1  |
| N          | 0           | 0       | 1                | 1  |
| P          | 0           | 1       | 0                | 1  |

The multi-similarity loss function integrates three similarity cases proposed by the authors (self-similarity, negative similarity, and positive similarity). Because it is difficult to integrate them in one step, it is divided into two steps: sampling and weighting of sample pairs. Its combined mathematical expression is shown in formula (1) (where $\alpha$, $\beta$, and $\gamma$ are hyperparameters):

$$L_{MS} = \frac{1}{m} \sum_{i=1}^{m} \left\{ \frac{1}{\alpha} \log \|1 + \sum_{k \in P_i} e^{\alpha (S_i - k)}\| + \frac{1}{\beta} \log \|1 + \sum_{k \in N_i} e^{\beta (S_i - k)}\| \right\}$$

3. A CLOTHING RETRIEVAL ALGORITHM MODEL

Generally, the retrieval of clothing will encounter a series of practical problems. For example, in a data set, clothes may wrinkle or hide. These problems will affect the accuracy of the experimental results. When the clothes are wrinkled, the eigenvalues that extracted from the training samples may be different from those of the flat clothes. This will be a big interference in the test. Of course, there will also be many problems in the reverse. In our cognition, the colors of the same clothes may not be the same, but the styles are the same. We generally think that they are the same kind of clothes. However, color discrimination is the main classification method for image recognition. But, through a large number of the same category images, the output results of the model may be corrected. However, in the actual situation, we may have a large data set, but there are not many images of the same type. So, we can not get the great output of the model. In this paper, we use data enhancement and model modification to enhance these features.

Figure 4 is the structure diagram of IAPM, in which there are two inputs and two outputs. They are image input, ROI input, DenseNet’s output, and joint feature output of ROI Pooling and DenseNet.
When the training set as an input, the data flow first passes through DenseNet, as shown in figure 5. DenseNet is divided into three parts and Output_1 is the output of the last transition layer. It is also an input for the dimension reduction layer together with the result of ROI pooling. Output_2 is the output of the last dense block and flows to the dimension reduction layer on the left. Output_3 is the output of the classification layer, and it eventually flows out of the model directly. The reason why Output_1 and Output_2 use the outputs of different layers is feature Pyramid [16] that different scales can detect different sizes of details. So, different feature details can be obtained by the output of multiple layers.

As shown in figure 6, the dimension reduction layer is composed of convolution, pooling, and activation. Due to the large output feature dimensions of DenseNet, the efficiency of the algorithm is affected and the memory is over-occupied. Therefore, reducing the dimension and downsampling operation is helpful to release the potential of the algorithm and increase the efficiency. The connection form of 1x1,3x3,1x1 convolution is also used to reduce the amount of calculation.

The results of the dimension reduction layer on the left side are equivalent to the global features of clothes. And the dimension reduction layer on the right side obtains the features of the interested region, that is, local features. The results of the last two dimension reduction layers are spliced and fully connected to get the final global and local features.
4. EXPERIMENTAL ANALYSIS

4.1. Data Set
DeepFashion\textsuperscript{[17]} is a large-scale data set for clothing and it is opened by the Chinese University of Hong Kong. It contains 800000 pictures from different backgrounds, different angles, different models, sellers' shows, buyers' shows, etc. Each image also has rich annotation information, including 50 categories, 1000 attributes, Bbox, and feature points. There are about 300000 different poses and pictures in different scenes. DeepFashion is composed of four parts: Category and Attribute Prediction Benchmark for classification and attribute prediction, In-shop Clothes Retrieval Benchmark, and Consumer-to-shop Clothes Retrieval Benchmark for image retrieval, and Fashion Landmark Detection Benchmark is used to detect key areas.

In the experiment, the In-Shop Clothes Retrieval data set is mainly used. It is a collection of sellers' shows taken from different angles, including 7982 pieces of clothing and a total of 52712 pictures. Through cleaning and screening, 3975 kinds of clothing were selected as a training set and verification set, and the remaining 3985 types were used as prediction set.

4.2. Experimental Process
For training, we adopt the method of pre-training. Firstly, DenseNet is used to train the training set, and the weight of ImageNet is applied. The cross-entropy loss shown as formula (2) is chosen. And the method of random descent gradient is used. The learning rate is set to 0.01 and the learning decline rate is $e^{-6}$. When reaching the level of goal expectation, we stop training. Afterward, we obtain the weight of the pre-training model and put it into the IAPM model to continue training.

Because the training set is too large, the method of batch processing is generally used to train the data set. However, the multi-similarity loss function used in this paper is different from the general batch processing method. It constructs a batch using structural random sampling. Specifically, n clothes of the same kind and m different kinds of clothes are randomly selected. The batch size is $m*n$.

The Input_ROI is to mainly select the region of interest. Inspired by Yolo\textsuperscript{[18–20]}, the K-means method is used to preprocess the training set to find multiple regions of interest. In this paper, four ROI regions are found in the training set. all ROIs are integrated and output backward through the RoiPooling layer\textsuperscript{[21]} proposed by Shaoqing Ren.

The test results show that the relative optimum of the model can not be found by using multi-similarity loss alone. Therefore, IAPM uses the joint training of multiple loss functions. Firstly, the cross-entropy loss function of image classification is given priority, and the multi-similarity loss function is supplemented. When the total loss drops to a certain degree, the weight ratio of the multi-similarity loss is increased to correctly pull the feature vector.

For testing, the purpose of the whole model is to adjust the weight of the model to adapt to the distribution of the data set. So this paper directly intercepts the feature map above the DenseNet classification layer and then finds the most similar clothes according to the cosine similarity.

It is a demand that different colors but the same style are in the same category. However, the data of our single category clothing is limited, and the whole model mainly depends on RGB to distinguish styles, so data enhancement is used to expand the color gamut with a ratio of 1:11. As shown in Figure 7:
4.3. Evaluation Index

This experiment belongs to the classification task. It uses the cross-entropy loss function and multi-similarity loss function as constraints.

According to the results of training and testing, the samples are generally divided into four situations:

TP(True Positive): The number of positive samples detected by the model is the number of positive examples.

FP(False Positive): Negative samples detected by the model is the number of positive cases.

TN(True Negative): The number of negative samples detected by the model is the number of negative cases.

FN(False Negative): Positive samples through the model check out the number of negative cases.

In this paper, three evaluation indexes are used (accuracy, recall, precision). As followed in formula (3) (4) (5):

\[
\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}
\]

\[
\text{recall} = \frac{TP}{TP + FN} \tag{4}
\]

\[
\text{precision} = \frac{TP}{TP + FP} \tag{5}
\]

4.4. Experimental Results and Analysis

In the following table 2, four algorithms are compared, mainly through three evaluation indicators (accuracy, recall, precision). IAPM⁺ refers to the data enhancement of the model data, that is, the color gamut expansion mentioned above. It can be seen from this table that with the data enhancement, the improvement in model accuracy is obvious, which increases about 19 percent. From table 2, the accuracy of our proposed model is significantly better than the other three models, which is 13% higher than that of the baseline (FashionNet).

| Model (%) | accuracy | recall | precision |
|-----------|----------|--------|-----------|
| IAPM⁺     | 65       | 64     | 68        |
| IAPM      | 46       | 46     | 49        |

TABLE II. THE PERFORMANCE OF MODELS
| Method       | 38 | 37 | 40 |
|--------------|----|----|----|
| SiameseNet   |    |    |    |
| DenseNet     | 34 | 34 | 35 |
| FashionNet   | 52 | 53 | -  |
| HDC          | -  | 62 | -  |
| WTB          | 35 | -  | -  |
| DARN         | 39 | -  | -  |

Figure 8 shows the accuracy of the top k (1-50) index tested by five training methods. The horizontal axis means the top n are sorted according to the similarity, and the vertical axis represents the accuracy. From the chart, the accuracy rate of IAPM+ at the top 5 is as high as 82%, and it is still rising. When it reaches the top 30, the accuracy rate has exceeded 90%. From figure 8, the accuracy of IAPM+ is always the highest.

5. CONCLUSION

The clothing retrieval model based on DenseNet and Multi-Similarity loss is proposed in this paper. The DeepFashion data set is used for experiments. It achieves 65% accuracy in top 1 and 82% accuracy in top 5 which shows obvious advantages compared with other algorithms in the experiment. In addition, due to the characteristics of clothing that different colors of clothing but the same style is considered as the same category, and there are few same kinds of clothes in the data set, it is more illustrative that our proposed model is extremely effective by carrying out specific data enhancement gamut expansion. Further improvement will be studied for better performance.

ACKNOWLEDGMENT

This work was supported by the Natural Science Foundation of Shanghai with Grant No. 19ZR1420800

REFERENCES

[1] Zheng L, Yang Y, Tian Q. SIFT meets CNN: A decade survey of instance retrieval[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(5): 1224-1244.
[2] Chopra S, Hadsell R, LeCun Y. Learning a similarity metric discriminatively, with application to face verification[C]/2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). IEEE, 2005, 1: 539-546.
[3] Hoffer E, Ailon N. Deep metric learning using triplet network[C]/International Workshop on Similarity-Based Pattern Recognition. Springer, Cham, 2015: 84-92.
[4] Hadsell R, Chopra S, LeCun Y. Dimensionality reduction by learning an invariant mapping[C]//2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). IEEE, 2006, 2: 1735-1742.

[5] Wen Y, Zhang K, Li Z, et al. A discriminative feature learning approach for deep face recognition[C]//European conference on computer vision. Springer, Cham, 2016: 499-515.

[6] Wang X, Han X, Huang W, et al. Multi-similarity loss with general pair weighting for deep metric learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 5022-5030.

[7] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[8] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.

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

[10] 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.

[11] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning[J]. arXiv preprint arXiv:1602.07261, 2016.

[12] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[13] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-4708.

[14] Oh Song H, Xiang Y, Jegelka S, et al. Deep metric learning via lifted structured feature embedding[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 4004-4012.

[15] Chen W, Chen X, Zhang J, et al. Beyond triplet loss: a deep quadruplet network for person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 403-412.

[16] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.

[17] Liu Z, Luo P, Qi S, et al. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 1096-1104.

[18] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[19] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.

[20] Redmon J, Farhadi A. YoloV3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.

[21] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]//Advances in neural information processing systems. 2015: 91-99.

[22] Oh Song H, Jegelka S, Rathod V, et al. Deep metric learning via facility location[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 5382-5390.

[23] Chen H, Gallagher A, Girod B. Describing clothing by semantic attributes[C]//European conference on computer vision. Springer, Berlin, Heidelberg, 2012: 609-623.
[24] Huang J, Feris R S, Chen Q, et al. Cross-domain image retrieval with a dual attribute-aware ranking network[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1062-1070.