Dual attention mechanism networks for fabric image retrieval

Hao Wei*, Zhixiang Wang
School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214000, Jiangsu, China

ABSTRACT

In recent years, with the booming development of deep learning, content using images has become a mainstream research hotspot for retrieval. Fabric image retrieval, as a specific application scenario for retrieval, has potential value in many areas such as textile product design, e-commerce and inventory management. However, in fabric retrieval, fabric image has unique striped texture features, and it is still a challenging task to better extract its features for retrieval. To address these problems, this paper proposes a method for fabric retrieval based on a dual attention mechanism. The method inserts a dual attention mechanism module into the convolutional neural network, firstly strip attention module, which connects long and distant image information to capture narrow feature regions, and secondly the channel attention module, which adaptively recalibrates the channel feature responses and selects the prominent and effective feature maps. Experimental results show that the proposed method performs superiorly in fabric image retrieval, and DAMNet is proposed to solve the problem of feature extraction and retrieval of such fabric images.

Keywords: Dual attention mechanism, fabric image retrieval, convolutional neural networks

1. INTRODUCTION

Fabric image retrieval, a special application scenario in the field of image retrieval on fabrics, has potential value and significance for its wide range of applications in the apparel market trade.

Colour, shape and texture are the three most basic low-level features of an image. Many previous studies have extracted these features to retrieve images. This is also the case for fabric images as a specific application vehicle. This strategy of using colour histograms coupled with segmentation in the image is used to solve the retrieval of patterned fabrics[1]. Further research follows in which algorithms combining colour moments and feature descriptors called Gist are used for printed fabrics[2]. Both approaches combine fabric colour and shape features. Li et al.[3] describe a content-based retrieval system for lace fabrics. Nanik et al.[4] develop a fabric retrieval system that combines fractal-based texture features with HSV features. In these studies[5-7], they use Fourier transforms, local binary patterns and local affine invariants for retrieval, respectively, all using the two features of colour and texture.

Among these methods, Convolutional Neural Networks (CNN) have become the dominant approach to deep learning. Focus ranking is embedded in CNNs to learn image features[8]. A method called triplet CNN is applied to learn image representations[9]. Since each layer of CNN features has different meanings, and also these features from different layers can complement each other, a strategy based on hierarchical retrieval was born[10].

In a nutshell, there are two main problems with the current research on fabric retrieval:

(1) Many manual feature-based methods cannot adequately express the semantic features of the image and are poorly adapted.

(2) Existing network perceptual fields are square convolutional to extract fabric features, however, for fabric images with repetitive long-distance band structure features, as shown in Figure 1, the ability of square perceptual fields to extract fabric band information is limited and it is difficult to adequately characterise the visual properties of the fabric.

In order to solve this problem, dual attention mechanism (DAM) based fabric image retrieval algorithm is proposed, where DAM consists of a fusion of two attention mechanisms, the strip and the channel attention mechanism. After the bar attention module, the feature map is changed for narrower parts of itself in a more targeted manner, after which the channels of the feature map are recalibrated to optimise the features, allowing for efficient capture of the fabric's bar space and

* 6201613052@stu.jiangnan.edu.cn

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 1250639
 · 0277-786X · doi: 10.1117/12.2662003

continuous correction of self-features. Both mechanism modules are lightweight and can be inserted into any current backbone network as add-on modules. The proposed DAM method performs well on the fabric dataset with much higher retrieval accuracy and good results.

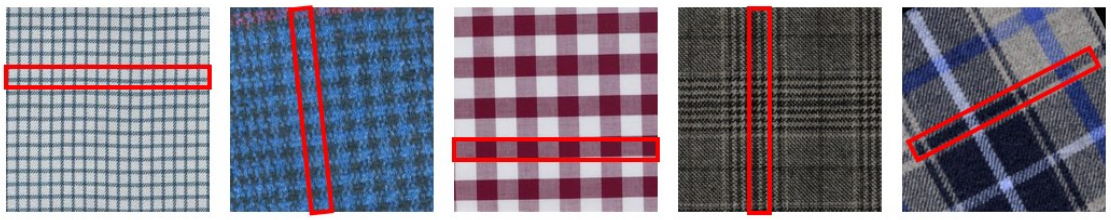

Figure 1. The red box shows the narrow striped textural features evident in the fabric image.

2. METHOD

The first is the bar attention mechanism module, which is proposed here to solve the above fabric image narrow distance problem by using a strip pool window to perform computational operations along the horizontal or vertical direction to connect the spatially long-distance information of the image content as a way to capture the bar texture.

A feature map $x \in \mathrm{R}^{H\times W}$ is assumed to be a two-dimensional tensor, where H and W are height and width respectively. In the strip merge layer, the required spatial extent is $H \times 1$ and $1 \times W$. Unlike the averaging pool, the proposed strip pool is computed by averaging all the feature values in a row or column of the feature map obtained by convolution. Thus, from the definition given it follows that the output of the horizontal strip pool can be written as:

$$y_i^h = \frac{1}{W}\sum_{0\le j<W} x_{i,j} \tag{1}$$

Similarly, the output of a vertically oriented strip pool can be written as:

$$y_j^v = \frac{1}{H}\sum_{0\le i<H} x_{i,j} \tag{2}$$

As can be seen from the two equations above, the strip pool is designed to be long and narrow, making it easy to establish long-term dependencies between discrete distributions of remote regions in the horizontal and vertical strip pool layers and to encode regional image information using the strip shape. At the same time, because the strip shape is narrower along one dimension, it not only connects remote image information, but also captures a certain degree of local image detail information. These properties unique to strip pooling make it different from the traditional spatial merging of square pools.

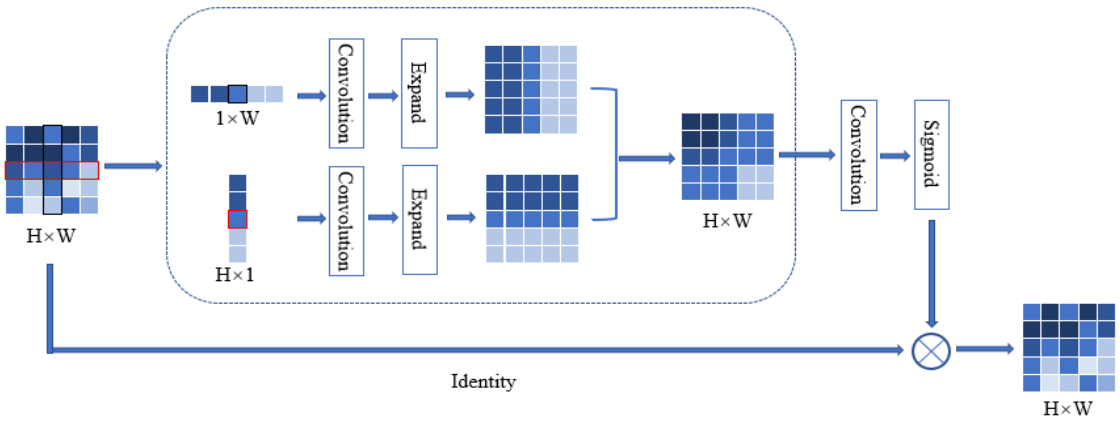

Figure 2. Diagram of the bar attention mechanism.

Next, we combine the two strip pools described above to propose a strip attention mechanism module. It combines horizontal and vertical strip pooling operations to collect remote image information from these two spatial dimensions,

helping the backbone to capture a more complete picture of the unique features of the fabric images. As shown in Figure 2, it depicts the strip-attention mechanism. Assume that the feature map $x \in R^{C\times H\times W}$ in the network is an input tensor, where C is the number of channels. First, we load x into two parallel scan paths, each containing a horizontal or vertical strip pooling layer, and then feed the output of the upper layer into a one-dimensional convolutional layer with a convolutional kernel of 3, which is used to adjust the current spatial position and its neighboring features.

Here we first give the two components, horizontal output $y^h \in R^{C\times H}$ and vertical output $y^v \in R^{C\times W}$, in order to obtain the final output $z \in R^{C\times H\times W}$. z is the result of y passing through the convolutional layer, so it contains more useful global information. The strip attention mechanism then first combines the two components y^h and y^v obtained above, as follows, to obtain $y \in R^{C\times H\times W}$:

$$y_{c,i,j} = y_{c,i}^{h} + y_{c,j}^{v} \tag{3}$$

After convolution, activation function and corresponding element multiplication operations, we can obtain z:

$$z = \mathrm{Scale}(x,\ \sigma(\mathrm{f}(y))) \tag{4}$$

The function Scale represents two matrices where each element is multiplied in turn, the function σ represents the sigmoid activation function so that the values are mapped between (0, 1) and the function f represents a 1 x 1 convolution operation. This solution is computationally more intensive, while the results perform similarly to the approach we have adopted. Considering efficiency, accuracy and the lightweight of the module, our proposed computational method is used to construct the strip attention mechanism module.

In contrast to the previous square perceptual field, the proposed bar attention mechanism module considers a long and narrow spatial range rather than the entire feature map, which fits well with the textural characteristics of fabric images and never avoids the creation of unnecessary connections between locations that are close to each other.

The feature map, after the bar attention mechanism module described above, has fused spatial and channel information from the local perceptual field to fit the extraction operator for the fabric image features. To continue to enhance the expressiveness of the network, as shown in Figure 3, the channel attention mechanism module is applied to the subsequent feature optimization.

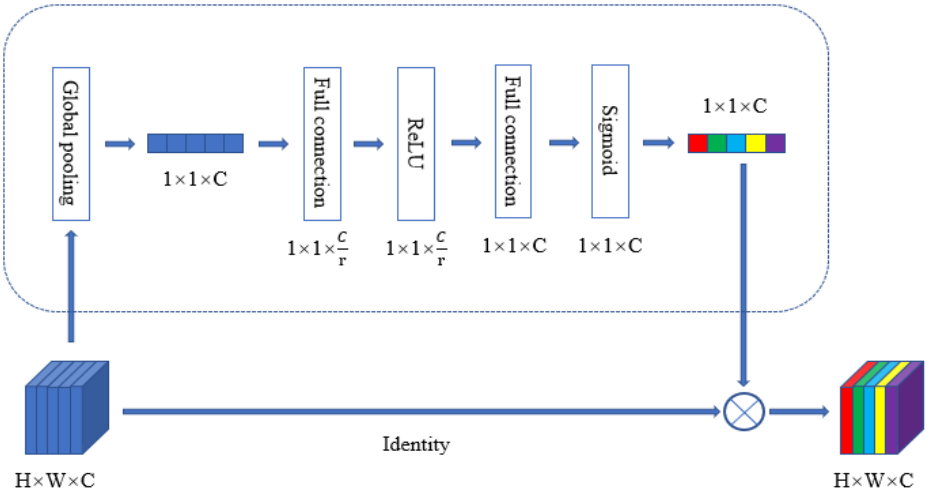

Figure 3. Diagram of the channel attention mechanism.

The image is fed into the network and passes through a convolutional layer to represent the local spatial connectivity patterns of the input channels. The resulting feature map enters the bar attention mechanism module to capture the narrow texture features of the fabric image. Recent research has shown that the performance of the network can then be improved by embedding the channel attention mechanism module, a self-learning mechanism, which allows the network channel features to be recalibrated to capture correlations in channel space without additional supervision.

Given a feature map z output by the upper layer, the signal of each channel in the feature map is first output, i.e. the feature map z is first subjected to a squeezing operation to aggregate the feature maps on the spatial dimension $H \times W$ to generate a channel descriptor which embeds the global distribution of the channel feature responses so that information from the global sensory field of the network can be used by its lower layers. Assuming that the generated channel descriptor is $u \in \mathrm{R}^C$, then the c-th element of z is calculated by the following equation:

$$u_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} z_c(i,j) \tag{5}$$

The output u can be interpreted as a collection of local descriptors whose statistical information represents the whole image. The use of this information is very common in feature engineering work.

Then comes the weighting operation, which learns weights for each channel through a self-selecting channel mechanism based on channel dependencies, reweighting the original feature map to fully capture the channel dependencies. Here the weighting is not simply a corresponding multiplication of z with the original feature map, but must be able to learn non-linear interactions between channels and a non-mutually exclusive relationship. Thus, after u has been obtained, u is then fed into a multilayer perceptron to learn the final weights. Assuming that the weights after non-linear learning are $s \in \mathrm{R}$, then it is calculated as:

$$s = \sigma(\mathrm{W}_1 \delta(\mathrm{W}_2(u))) \tag{6}$$

where W_1 and W_2 denote the weight matrix of the perceptron, $\mathrm{W}_1 \in \mathrm{R}^{\frac{C}{r} \times C}$ is the weight matrix of a descending layer, $\mathrm{W}_2 \in \mathrm{R}^{C \times \frac{C}{r}}$ is the weight matrix of an ascending layer, the value of C is the number of channels in the input feature map, $\frac{C}{r}$ denotes the number of nodes in the hidden layer of the perceptron, δ is the ReLU activation function. After the channel attention mechanism module, the final output of the feature map is obtained by rescaling and transforming the original feature map using the learned weights:

$$o_c = \mathrm{F}_{\mathrm{scale}}(z_c, s_c) \tag{7}$$

where o_c denotes the feature map of O on channel C, z_c denotes the feature map of the input feature map z on channel C, s_c denotes the weights on channel C learned by the multilayer perceptron, and the $\mathrm{F}_{\mathrm{scale}}$ function denotes the one-to-one multiplication of a feature mapping value with all its corresponding channel values. The channel attention mechanism module essentially introduces dynamic adjustment conditional on the input, which facilitates improved feature discriminability.

It captures the long and thin texture features of the fabric image, while the channel attention mechanism dynamically adjusts and optimizes each feature map. By inserting the proposed dual attention mechanism module into the network, the extraction of fabric image features is facilitated, laying the foundation for the improvement of image retrieval accuracy later on.

3. EXPERIMENTATION AND ANALYSIS

The experimental environment for this paper uses an Ubuntu system, Python as the programming language and the Pytorch framework as the neural network learning framework. The dataset used for the experiments is a fabric image dataset created by a school teacher for a project requirement. It contains a training set, a validation set and a test set with five categories. There are 10,322 images in the training set and 3,500 images in both the validation and test sets. Due to the limited number of images in the home-made dataset, using a large network model with a very deep layer count tends to cause overfitting. Therefore, we chose AlexNet, VGGNet-11, ResNet-18 and ResNet-34, which have fewer layers, as the backbone network, while adding the validation set during training and testing the validation set every epoch during the network training to retain the model parameters with the highest correctness. The above two measures will greatly reduce the model fitting and improve the generalization ability of the model. To verify the effectiveness of the proposed method, four sets of comparison experiments were done with the four networks mentioned above, without and with DAM. We first validated the effectiveness of the DAM module on AlexNet, ResNet-18 and ResNet-34. In this paper, we fuse the most effective base network VGGNet-11 and DAM module to propose a new network model DAMNet to solve the fabric image problem.

The above networks were trained and tested separately on a test set with the correct and recall rates shown in Table 1. The network models with DAM added were all more correct than those without. The images from the fabric image library were fed into these networks, and the depth-hashed image representations were saved. Here we compare four feature vector

lengths to retrieve comparisons, hashed to 16, 32, 64 and 128 bits respectively. The data set has five categories, and the retrieved images are set to return 10 relevant images, and the retrieval performance metric mAP is calculated as shown in Table 2, which shows that the retrieval accuracy of the network with the dual attention mechanism module is higher than that without it, thanks to the richer image features captured by the proposed method.

Table 1. Comparison of the Accuracy and Recall@K of different methods on the test set.

Methods	Accuracy	Recall@5	Recall@10	Recall@15
AlexNet	92.5%	74.6%	81.1%	88.3%
AlexNet-DAM	92.8%	74.8%	81.7%	88.7%
ResNet-18	92.9%	75.2%	82.8%	88.5%
ResNet-18-DAM	93.5%	75.7%	83.2%	89.1%
ResNet-34	92.5%	74.8%	82.3%	88.5%
ResNet-34-DAM	93.4%	75.9%	83.4%	89.6%
VGGNet-11	91.5%	73.8%	80.5%	87.4%
DAMNet	**93.6%**	**76.1%**	**84.3%**	**90.2%**

Table 2. Comparison of the mAP of the proposed method on the test set at different bits.

Methods	16bits	32bits	64bits	128bits
AlexNet	0.705	0.805	0.837	0.864
AlexNet-DAM	0.713	0.810	0.849	0.878
ResNet-18	0.709	0.800	0.841	0.874
ResNet-18-DAM	0.728	0.825	0.863	0.885
ResNet-34	0.711	0.804	0.850	0.882
ResNet-34-DAM	0.726	0.826	0.868	0.890
VGGNet-11	0.726	0.821	0.862	0.893
DAMNet	**0.737**	**0.836**	**0.873**	**0.906**

4. CONCLUSIONS

This paper proposes a dual attention mechanism-based method for fabric retrieval that captures narrow band structure by connecting remote dependencies in the image through a strip attention mechanism, after which the channel attention mechanism then self-corrects the feature map to optimize the features. The above proposed dual attention mechanism module is lightweight and can be easily inserted into any backbone network for use. Combining the two network models and the two dimensionality reduction algorithms, we did four sets of comparison experiments. The retrieval results show that the retrieval accuracy of the methods with the added dual attention mechanism are all higher than those without. This shows that it is effective and feasible.

REFERENCES

[1] Jing, J., Li, Q., Li, P., Zhang, H. and Zhang, L., "Patterned fabric image retrieval using color and space features," Journal of Fiber Bioengineering and Informatics 8(3), 603-614 (2015).

[2] Jing, J., Li, Q., Li, P. and Zhang, L., "A new method of printed fabric image retrieval based on color moments and gist feature description," Textile Research Journal 86(11), 1137-1150 (2016).
[3] Li, Y., Luo, H., Jiang, G. and Cong, H., "Content-based lace fabric image retrieval system using texture and shape features," The Journal of The Textile Institute 110(6), 911-915 (2019).
[4] Suciati, N., Herumurti, D. and Wijaya, A. Y., "Fractal-based texture and HSV color features for fabric image retrieval," International Conf. on Control System, Computing and Engineering (ICCSCE), 178-182 (2015).
[5] Zhang, N., Xiang, J., Wang, L., Gao, W. and Pan, R., "Image retrieval of wool fabric. Part I: Basedon low-level texture features," Textile Research Journal 89(19-20), 4195-4207 (2019).
[6] Zhang, N., Xiang, J., Wang, L., Xiong, N., Gao, W. and Pan, R., "Image retrieval of wool fabric. Part II: based on low-level color features," Textile Research Journal 90(7-8), 797-808 (2020).
[7] Li, Y., Zhang, J., Chen, M., Lei, H., Luo, G. and Huang, Y., "Shape based local affine invariant texture characteristics for fabric image retrieval," Multimedia Tools and Applications 78(11), 15433-15453 (2019).
[8] Deng, D., Wang, R., Wu, H., He, H., Li, Q. and Luo, X., "Learning deep similarity models with focus ranking for fabric image retrieval," Image and Vision Computing 70, 11-20 (2018).
[9] Cai, Z., Gao, W., Yu, Z., Huang, J. and Cai, Z., "Feature extraction with triplet convolutional neural network for content-based image retrieval," 2017 12th IEEE Conf. on Industrial Electronics and Applications (ICIEA), 337-342 (2017).
[10] Xiang, J., Zhang, N., Pan, R. and Gao, W., "Fabric image retrieval system using hierarchical search based on deep convolutional neural network," IEEE Access 7, 35405-35417 (2019).