Convolutional Neural Networks for Clothing Image Style Recognition
Chun-yan DONG*, You-qun SHI and Ran TAO
Donghua University, Shanghai, 201620, China
*Corresponding author

Keywords: Convolutional neural networks, VGG-Net, Spatial pyramid pooling, Clothing style recognition.

Abstract. Automatic recognition of the clothing image's style is important for quite a few applications, including apparel automatic labeling, recommendation for clothing, and clothing retrieval, etc. Convolutional neural networks cope with the image recognition well. However, the networks require a fixed-size input via cropping or scaling the image arbitrarily, which may reduce the recognition accuracy for the images. This paper equipped the fine-tuned VGG-Net with spatial pyramid pooling to eliminate the restriction of a fixed-size input image. The study showed that the combined network had a higher cross-validation accuracy of the style recognition in clothing images compared with the Google-Net and the fine-tuned VGG-Net. The network for the style recognition of clothing images flexibly addresses the issue of the dataset with different sizes and scales. This study also improves the accuracy of the style recognition in clothing images. Moreover, the network is beneficial to the classification or recognition of other datasets.

Introduction
The style in a clothing image plays an important role in the recommendation system for clothing based on expert knowledge. The system recommends clothes referring to rules defined by the expert for users' style and the style of clothing. Users' style type is diagnosed according to their face types and personality characteristics [1]. The clothing style based on the experience of the expert in fashion image design is complex taking the factors of colors, patterns, design and fabric into account. The labels of styles from original clothing images are inconsistent with the expert, which influences what the system recommends. With the advent of massive garments, labeling garment styles manually takes a lot of time and effort under the complex rules. In addition, manually labeling clothing styles is often subjective.

The style recognition of apparel images can be achieved by using image recognition or image classification techniques. In recent years, CNNs (Convolutional Neural Networks) have been widely used in the field of image processing [2], and have achieved favorable results. CNNs can automatically extract the features of the detection area, which can be combined with the recognition of clothing images. Alex-Net and Fashion-Net (whose network architecture is similar to VGG-Net) were used in the classification of clothing categories [3], [4]. This turns out to be true that convolutional neural network is an efficient tool for the image recognition.

However, these networks require a fixed-size input image. Cropping or scaling the input image to a fixed size loses part of the original image information, thus affecting the accuracy of recognition when the images of dataset are arbitrary size and proportion. In the SPP-Net [5], a pooling strategy of "spatial pyramid pooling" was introduced to eliminate the limitation above. The main contribution of this work is combining VGG-19 network with the SPP (Spatial Pyramid Pooling) to recognize the clothing image style. This study handles the limit of fixed-size images flexibly and improves the accuracy of the style recognition in clothing images. This study also applies to the classification of other datasets.

Related Work
Traditional way to process the clothing image basically uses artificially designed image feature extraction algorithms to extract features. Bossard et al. [6] combined HOG, LBP and other features to
use SVM, random forest, and transfer forests to classify apparel types, achieving average accuracy rates of 35.03%, 38.29%, and 41.36%, respectively. Common methods for image recognition include: Bayesian classification, template matching, etc. [7]. In the Bayesian classification method, Bayes theorem assumes that the effect of an attribute value on a given class is independent of the values of other attributes, and this assumption is difficult to establish in the recognition of clothing image styles, for example, keywords in style quantification the color attributes and the pattern attributes are not completely independent, and the colors include the mainstream colors of the clothes and patterns. The template matching method has its own limitations. Provided that the matching target in the original image rotates or alerts in size, the algorithm is invalid [7], which undoubtedly imposes strict requirements on the input image.

The Alex-Net [8] won the first place in the classification task of ILSRVC-2012 and deep learning has received much attention afterwards. Network models such as ZF-Net [9], Google-Net [10], ResNet [11], VGG-Net [12] make convolutional neural networks a breakthrough in image classification. As in [3], Alex-Net was used in the clothing category classification. After that, Simonyan and Andrew [12] proved that the depth of the convolutional network is conducive to classification accuracy, and the proposed VGG-Net [12] deepens the depth of the network based on the Le-Net [13] architecture. As in [4], the Fashion-Net (whose network architecture is similar to VGG-Net) was used in the classification of clothing categories.

The above studies provide guidance for our study. However, these studies mainly used convolutional neural networks in the classification of clothing categories. What’s more, deep convolutional neural networks have the requirement of a fixed-size input image. This work is combining VGG-19 network with the SPP (Spatial Pyramid Pooling) to recognize the clothing image style.

The Model

Convolutional Neural Network

CNNs have been able to achieve such great success mainly as it can represent different levels of complex natural images. CNN consists of an input layer, an output layer, and multiple hidden layers. The hidden layer of CNN is usually composed of a convolutional layer, a pooling layer and a fully connected layer.

The generation of convolutional layers uses non-linear activation functions, typically sigmoid function, tanh function, and so on. In order to cope with the exploding gradient or vanishing gradient problem and accelerate the convergence speed, the unsaturated non-linear function is now commonly used as the convolutional layer activation function such as the ReLU function. The ReLU function is calculated as follows:

\[ f(x) = \max(0, x). \]  

(1)

Let \( h_i \) be the characteristic map of the layer \( i \) of the convolutional neural network. Assuming \( h_i \) is a pooling layer, the generation process of the pooling layer can be described as:

\[ h_i = \text{subsampling}(h_{i-1}). \]  

(2)

The function \( \text{subsampling}(x) \) refers to the rules adopted for the pooling layer, such as max-pooling and average pooling.

After the convolutional layer and the pooling layer are alternately transmitted, the convolutional neural network classifies the extracted features through a fully connected network. The ReLU function is generally used as activation function for each neuron of the full connected layer. The output value of the last fully connected layer is passed to an output layer, which can be classified using softmax regression. It is called softmax layer as well. In order to avoid training over fitting, the regularization method—dropout is often used in the fully connected layer.
VGG-19 Network

The VGG-Net [12] team secured the first and the second places in the localisation and classification tracks respectively in the ILSVRC-2014 contest. The VGG-19 network model has 19 layers, including 16 convolution layers and 3 fully connected layers. Each convolutional layer uses the same convolution kernel of size $3 \times 3$ and joins five pooling layers (using the max-pooling method) to perform a 5-phase convolutional feature extraction. The architecture of the entire network is extremely symmetrical. It uses ReLU activation function instead of sigmoid or tanh to shorten the training time and introduces dropout to prevent over-fitting.

SPP-Net

The VGG-19 network requires a fixed input image size (e.g., $224 \times 224$ RGB image). In the SPP-Net [5], a pooling strategy of "spatial pyramid pooling" was introduced to eliminate the limitation. Spatial Pyramid pooling is extremely robust to deformation of objects. Actually, what requires the fixed images size is not the convolutional layer but the fully connected layer. The spatial pyramid pooling layer is placed behind the last convolutional layer and produces a fixed size output used as the input by the fully connected layer avoiding cropping or scaling the input image at the beginning. The output of the spatial pyramid pooling layer is $kM$-dimensional, where $k$ represents the number of convolution kernels in the last convolution layer and $M$ represents the number of blocks. This fixed $kM$-dimension vector is the input of the fully connected layer.

The Clothing Image Style Recognition Network

In order to make the VGG-19 network more suitable for the style recognition of clothing images, the following improvements have been made on the basis of the original network architecture.

It is calculated that the number of neurons in the first fully connected layer is 4096, the same as the second fully connected layer, and the parameters are as high as 102764544. What’s more, the second fully connected layer contains 16777216 parameters. The total number of parameters of two fully connected layers accounts for more than 85% of all parameters, which seriously affects the overall speed of the network. In a convolutional neural network, the convolutional layer mainly extracts image features, while the role of the fully connected layer is mainly to integrate local information with class differentiation in the convolution layer, which has not great impact on the network. So the two fully connected layer are removed, leaving the last fully connected layer merely. Additionally, the overall parameters of the network can be greatly reduced by reducing the number of fully connected layers, so that more calculation time and memory space can be saved [14].

Modify the neurons’ number of the last fully connected layer. The original network targets 1000 categories of ImageNet dataset and we have six kinds of clothing styles used in the recommendation system for clothing based on expert knowledge. Therefore, the number of neurons in the last fully connected layer is modified from 1000 to 6.

In order to use the original input images instead of cropping or scaling the input image to the fixed size, we add the spatial pyramid pooling layer between the last convolutional layer and the fully connected layer of the fine-tuned VGG-19 network.

Figure 1 lists the structure of the network model we used for recognizing the style of clothing images.

![Figure 1. The structure of the clothing style recognition network model.](image-url)
The input is the original size of image instead of the fixed-size image. Next, the images are calculated by the means of convolution and max-pooling to perform a five-phase convolution feature extraction, the output of which is different sizes of feature images. Then, the images output fixed size ones used as the input by the fully connected layer through the SPP layer. Afterwards, the output value of the last fully connected layer is passed to an output layer, which can be classified using softmax regression. Each convolutional layer uses the same convolution kernel of size 3×3 (e.g., expressed in Conv3) and joins five pooling layers using the max-pooling method. The number after the Conv3 refers to the quantity of the kernels.

**Experimental Results**

**DataSet**

At present, there are few datasets for clothing style published on the Internet and the styles given are not what we want. Nevertheless, we use part of the data in the database of recommendation system for clothing based on expert knowledge. The data source comes from the information of women’s clothing on the e-commerce platform, mainly including the name, details and pictures of the clothing. The apparel style tagged in the e-commerce platform may lack specialization so the data requires being pre-processed and re-evaluated in accordance with the standard definition for style given by the expert. The styles used in recommendation system for clothing include retro, simple, cute, elegant, romantic and ladylike. Each image is a garment without a model and the background of it is white, which reduces the interference with the extraction of the main features of the clothing. Figure 2 shows example images of our dataset.

![Figure 2. Example images in the dataset.](image)

Among them, the criteria of each style definition comprehensively take the four aspects of the color, the pattern, design and fabric into account. Firstly, we extract the main colors of the clothing to determine the range of styles that it belongs to. When experiencing the color repetition, we must consider the pattern, the design and fabric to define the final style furthermore. Finally, a total of 35,446 clothing images are selected as the dataset. We did random split on the original dataset as train set and test set with ratio 7:3. Table 1 lists the quantity of clothing images for each style.

<table>
<thead>
<tr>
<th>The type of style</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>elegant</td>
<td>5472</td>
</tr>
<tr>
<td>cute</td>
<td>4564</td>
</tr>
<tr>
<td>retro</td>
<td>5636</td>
</tr>
<tr>
<td>romantic</td>
<td>5488</td>
</tr>
<tr>
<td>ladylike</td>
<td>7644</td>
</tr>
<tr>
<td>simple</td>
<td>6622</td>
</tr>
</tbody>
</table>

**Settings**

The purpose of the experiment is to verify the efficiency of the clothing image style recognition model. At the same time, we use our dataset on the VGG-19 network and Google-Net as a contrast.
When it comes to evaluation indicators, we use training accuracy, cross-validation accuracy, training loss, and verification loss to evaluate the efficiency of recognition.

In terms of input data, the input of VGG-Net and Google-Net is a fixed-size 224×224 RGB image. The clothing images require being pre-processed. The shorter side of an image is normalized to 224 pixels, and the longer side of an image is scaled. What's more, the image area of 224×224 in the middle of the image is intercepted and the image is finally saved as 224×224. Nevertheless, our model does not need to do such processing and input the original image directly.

Results

The experiment is based on the Keras framework. Figure 3 shows the training accuracy, training loss, accuracy of the cross validation set and verification loss for each iteration of our network.

Figure 3. Evaluation indicators of our network.

Figure 3 shows that the training loss of our network gradually decreases with the increase of the epoch. With the increasing epoch, the cross-validation accuracy keeps upward tendency slowly and remains the same about 0.7959 at the end. All of the four indicators change sharply before the 20th epoch. Notice that the verification loss reaches the bottom near the 43rd epoch when the network performs the best, while the training accuracy is 0.8728 and cross-validation accuracy up to 0.7678.

Similarly, we recorded the results of VGG-19 network and Google-Net at the best level. Table 2 lists the comparison of three networks.

Table 2. Comparison of three networks.

<table>
<thead>
<tr>
<th>Network Model</th>
<th>Training Accuracy</th>
<th>Cross-Validation Accuracy</th>
<th>Training Loss</th>
<th>Verification Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG-Net</td>
<td>0.7402</td>
<td>0.6890</td>
<td>0.7125</td>
<td>0.8567</td>
</tr>
<tr>
<td>Our Network</td>
<td>0.8728</td>
<td>0.7678</td>
<td>0.4137</td>
<td>0.7968</td>
</tr>
<tr>
<td>Google-Net</td>
<td>0.7493</td>
<td>0.7070</td>
<td>0.7507</td>
<td>0.8203</td>
</tr>
</tbody>
</table>

Three networks' cross-validation accuracy is 0.6890, 0.7678, and 0.7070, respectively. Our network performs the best with the least loss of training and verification in contrast to VGG-Net and Google-Net. VGG-Net has less loss of training and verification than Google-net. Although Google-Net has a higher cross-validation accuracy than VGG-Net, VGG-Net is fit for our task of the style recognition in clothing images better considering the complex of network architecture whose loss of training and verification is less. Our network is 0.0788 higher as to the accuracy of cross-validation than VGG-19 network owing to the added SPP layer.

Conclusions

In this paper, we apply the fine-tuned VGG-19 network (with SPP layer) to the style recognition of clothing images. In experiments, we verified our network by comparison to VGG-Net and Google-Net. This study flexibly copes with the issue of any size and proportion of input images. This study
improves the recognition accuracy of the clothing image style. Our approach of automatic recognition of the clothing image's style is useful for many applications, including apparel automatic labeling, recommendation for clothing, and clothing retrieval, etc. Moreover, our network also has a strong generalization ability which could be used for the classification or recognition of other datasets. For future research, we will consider image texture recognition algorithms to improve our network possibly.

Acknowledgment

This research is supported by the Dongguan professional town innovation service platform construction project, Dongguan Humen garment collaborative innovation center, and the Guangdong provincial cooperation innovation and platform environment construction special funds (Grant No. 2014B090908004).

References


