IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 27, 2019, accepted March 8, 2019, date of publication March 12, 2019, date of current version April 1, 2019.

*Digital Object Identifier 10.1109/ACCESS.2019.2904536*

# A Multi-Task Learning Model for Better Representation of Clothing Images

**CAIRONG YAN[1], LINGJIE ZHOU[1], AND YONGQUAN WAN[2]**
[1]School of Computer Science and Technology, Donghua University, Shanghai, China
[2]College of Information Technology, Shanghai Jian Qiao University, Shanghai, China

Corresponding author: Cairong Yan (cryan@dhu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61402100.

**ABSTRACT** Clothing images vary in style and everyone has a different understanding of style. Even with the current popular deep learning methods, it is difficult to accurately classify style labels. A style representation learning model based on the deep neural networks called StyleNet is proposed in this paper. We adopt a multi-task learning framework to build the model and make full use of various types of label information to represent the clothing images in a finer-grained manner. Due to the semantic abstraction of image labels in the current fashion field, using a simple migration learning method cannot fully meet the requirements of clothing image classification. An objective function optimization method is put forward by combining the distance confusion loss and the traditional cross entropy loss to improve the accuracy of StyleNet further. The experimental results show that by applying the multi-task representation learning framework, StyleNet can achieve a better classification accuracy, the optimized loss function can also bring performance improvement for deep learning models, and the classification effect of StyleNet becomes better as the size of the data set increases. In order to verify the robustness and effectiveness of the deep learning method in StyleNet, we also apply a Faster R-CNN module to pre-process the clothing images and use the result as the input of StyleNet. The classifier can only get a limited performance improvement, which is negligible compared with the methods proposed in this paper of increasing the depth of the neural network and optimizing the loss function.

**INDEX TERMS** Clothing image, multi-task learning, representation learning, cross entropy loss, distance confusion loss, Faster R-CNN.

## I. INTRODUCTION

With the rise of social platforms such as Facebook, Twitter, and Weibo, image sharing by street shooting and taking selfies has become the most popular social activity. Whether popular clothing elements can be retrieved from social networking platforms has become one of the standards of measurement of popularity. At the same time, the massive clothing image information resources accumulated by major social platforms also provide data support for artificial intelligence research in the field of fashion apparel. They allow us to examine fashion from an unprecedented perspective, and then to explore more intelligence in the field of fashion apparel.

Traditional clothing image classification technology mainly focuses on extracting global features such as color, shape, texture through ingeniously designed feature extraction algorithms and uses these features or their combination as input for classification. In recent years, more research has been done by extracting local features of clothing images. The commonly used local feature extraction algorithms include Harris Corner Detection [1], SIFT [2], and HOG features [3]. However, these algorithms, while effective, face two problems. Firstly, the ingenious feature extraction algorithms are time-consuming and resource intensive. They rely heavily on the experience of experts to give heuristic guide expertise. Secondly, feature extraction of clothing images is affected by many factors such as viewpoint changing, light changing, complex background, and multiple target objects, so it is impossible to hard code an algorithm to deal with such situations and achieve good feature extraction and sometimes even rely on luck. There are currently two acceptable strategies to deal with these situations. The first is to perform clothing image segmentation and extract image feature based on the segmented clothing area. The second is

The associate editor coordinating the review of this manuscript and approving it for publication was Larbi Boubchir.

to extract image feature by locating the position of the person in the clothing image and selecting key points.

With the breakthrough progress of deep learning in the field of image recognition, the deep neural network has become a research hotspot. Most of the current deep learning methods use neural networks to build multi-layered structure models and train them to learn computer-readable representations through large-scale data and a large number of iterative calculations. The classic deep learning models for image recognition include AlexNet [4], GoogLeNet [5], and ResNet [6]. These models have achieved remarkable grades in ILSVRC competition. The training of these models is based on massive data and they are mainly suitable for coarse-grained classification, e.g., this picture is about cats and that picture is about dogs. However, in the fashion field, the purpose of style classification is to determine whether an image shows formal, sport or other style. Furthermore, the sample information is incomplete. The classification task at this time belongs to a fine-grained classification. Although the classic deep learning models have strong generalization ability, when they encounter these problems, they have difficulty obtaining better performance. This is the issue of concern in this paper.

The clothing image classification process is generally divided into three steps, image pre-processing, feature extraction, and final classification. The accuracy of the entire classification task depends on the effect of each partial processing, so every part is needed to be independently studied.

The purpose of this paper is to focus on the characteristics of clothing images, design an efficient representation learning model, propose an optimization method of the loss function, and improve the classification accuracy of clothing images. The main contributions are as follows.

1) Aiming at the problem that there are many details hidden in the clothing image and it is difficult to accurately classify labels, a style representation learning model based on the deep neural networks called StyleNet is proposed. The model makes full use of different label information of the data set to classify the clothing image in a fine-grained manner. The experiment shows that multi-task learning framework can help improve the classification accuracy.
2) Aiming at the problem of semantic abstraction of labels in clothing image classification, simple transfer learning cannot improve the accuracy of clothing image classification. A loss function optimization method combining distance confusion loss and traditional cross entropy loss is proposed. The experiment shows that the new loss function can bring performance improvement.
3) We explore the auto focus and adjustment capabilities of deep neural networks. By using two kinds of test data set, the original clothing image set and the clothing image set pre-processed by Faster R-CNN module [7], we find that the classification accuracy of the latter does not get much improvement. In the representation learning model, if we properly deepen the depth of the neural network without pre-processing the original images, we can also obtain good performance.

## II. RELATED WORK

### A. REPRESENTATION LEARNING FOR FASHION IMAGES

With the continuous development of social network and mobile computing, there are more and more images in the fashion field, which provides the basis for researchers to analyze data in the fashion field and promote the development of the fashion industry. Because images can provide more hidden and anomalous information, how to maximize the extraction of image information and make it easier to obtain useful information when constructing classifiers or other predictors is the problem to be studied by representation learning [8]. Since the understanding of computers to all things in the real world is based on zero and one, different representations can more or less hide or explain the changing factors behind the data. To find a better representation for different tasks, many teams have conducted research in this area. Saha *et al.* [9] proposed two simple variants of multimodal representation learning model which can learn disentangled common representations for the fashion domain where each dimension would correspond to a specific attribute. It led to better performance for cross-modal image retrieval, visual search, image tagging, and query expansion. Taheri *et al.* [10] proposed a method for learning graph representations by using the recurrent neural network-based auto encoders for graph classification and comparison tasks. Ma *et al.* [11] proposed a novel fashion-oriented multimodal deep learning-based model called bimodal correlative deep autoencoder (BCDA) to capture the intrinsic relevance of clothing collocation and achieved good results in brand trend analysis, clothing matching recommendation and other applications. Hsiao *et al.* [12] proposed an unsupervised representation method based on probabilistic polylingual topic models to discover a set of latent style factors. It can be seen from the related references that representation learning is widely researched in the fashion field [13]–[15]. This paper will focus on the fashion field and try to find a more efficient representation learning model for improving the classification accuracy of clothing images.

### B. MULTI-TASK LEARNING

Representation learning is a key issue in the field of artificial intelligence. Over the past few years, there has been an increasing interest in nonlinear representation learning from multiple tasks by using multiple layers of deep networks [16]. Multi-task learning is a learning paradigm that uses other related tasks to improve the generalization performance of learning tasks. It is widely applied in transfer learning [17], especially in the field of NLP [18]. Tang *et al.* [19] proposed a collaborative joint training approach based on multi-task recurrent neural network models and demonstrated that the multi-task approach could improve performance of both automatic speech and speaker recognition tasks compared to single-task systems. Collobert and Weston [20] proposed a convolutional neural network based on multi-task

framework for natural language processing including part-of-speech tagging, sentence syntactic component partitioning, named entity recognition, semantic role tagging, and also obtained state-of-the-art performance. Multi-task learning is also widely used in the field of image processing. Yang *et al.* [21] proposed a brand-new detection model based on multi-task rotational region convolutional neural network to solve the problems above, and it got a competitive performance in a detailed evaluation based on SRSS for rotation detection. Rezaei *et al.* [22] presented an end-to-end multi-task learning architecture and it was tailored to picture tumors in magnetic resonance imaging (MRI) and computed tomography (CT) images. It is widely accepted that multi-task learning is more beneficial than independent task learning, so we will apply multi-task framework to clothing style representation modeling, design a multi-task network structure, and add a corresponding constraint for each subtask to learn a better representation.

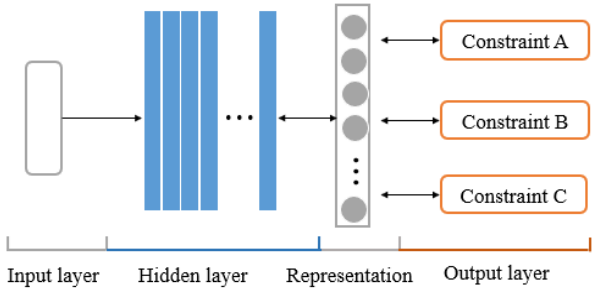


**FIGURE 1.** The framework for deep representation learning. It is made up of four parts. The last part is used to train the representation learning model.

## III. FORMULATION

How to make computers better understand the style of clothing is defined as a problem of representation learning in the fashion field. According to the image features obtained by representation learning model, we can classify clothing into different categories or styles. Fig. 1 shows the framework of deep representation learning. It is made up of four parts, input layer, hidden layer, representation layer, and output layer.

In the input layer, images are the input to the model. The traditional representation learning models use numeric or text as input. With the development of visual technology, images became the hot research objects in the fashion field. Given a data set $D$, each sample in it consists of an image and its corresponding classification labels. For clothing images, they can be classified according to style task, seasonal task, and feature task. Different tasks correspond to different classification labels. In the hidden layer, features are extracted from the images by a deep learning model such as a custom CNN model. VGGNet is a classic CNN model, which consists of a 5-layer convolutional layer, a 3-layer fully-connected layer, and a softmax output layer. The layers are separated by max-pooling, and the activation units of all hidden layers use ReLU function. These extracted feature representations are then modified by supervised constrained tasks. It shows the error backpropagation training process of the neural network. The constraint task here is usually a supervised classification task, e.g., if there are three classification tasks, there will be three constraints, as shown in Fig. 1. After the model is trained, for a clothing image, a representation can be generated that can be used to determine which classification the image belongs to. In this paper, we use a classification task to verify the representation learning model. The formalization process of this multi-task framework is as follows.

For a sample $s$ ( $s \in D$), s = ($v_i$, $l_a$, $l_b$, $l_c$), $v_i$ is the image, $l_a$, $l_b$, and $l_c$ are three classification label set. The framework also supports more classification tasks. $w_i = < w_{i1}, w_{i2}, \ldots, w_{in} > (w_{ij} \in [0, 1], 1 \leq j \leq n)$ is the output feature vector of the hidden layer. The value of $w_i$ will be adjusted according to the constraints in the representation layer. $l_a$, $l_b$, and $l_c$ are mapped to three constraints.

## IV. METHODOLOGIES

### A. STYLENET

The labeling of the clothing images is usually done by vendors or sellers. Generally, there are two different ways to label them. One is to label images based on the overall characteristics of them, e.g., style and season. Such tag values are usually mutually exclusive, and the corresponding classification task belongs to the category of single-label classification. The other is to mark the local features of the image, such as tops, skirts, and bags. Such tags are primarily used to describe whether an image contains such features. This kind of classification task is usually in the category of multi-label classification. For the latter, the training process of the model is relatively complicated.

In this paper, we will consider both cases and improve the accuracy of single-label classification by training with multi-label classification to improve the results of style classification and make full use of the label information to make the model better understand the embedded representation of style.

According to the above analysis, a style representation learning model based on the deep neural network called StyleNet is proposed in this paper. Fig. 2 shows the structure of it, which consists of four parts.

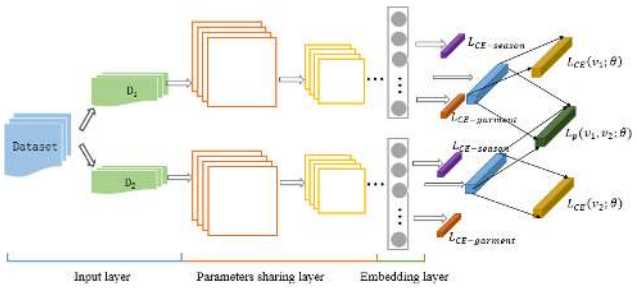


**FIGURE 2.** The structure of model StyleNet based on multi-task representation learning framework. The dataset is divided into two subsets to train the model. This approach does not increase the complexity of learning because both models share the same parameters.

1) Input layer. The data set is divided into two groups for training in a random manner. They will be used as the input of two CNN models.
2) Parameters sharing layer. Two data sets are used to train two CNN models separately. For each layer of convolution in CNN, the most efficient features of the data will be extracted, e.g., the most basic feature in the image, such as edges or corners in different directions can be extracted. The outputs of two CNN models will be combined to form higher-order features. The two CNN models share the same parameters so that the training efficiency of backpropagation of the whole model can be improved.
3) Embedding layer. The learning process benefits from the backpropagation algorithm, which can be adjusted according to the loss function of the next layer to obtain the feature representation of image.
4) Constraint layer. A new objective function $L_{all}$ is trained together by adding three supervised predictors, e.g., season, costume, and style, behind the embedded layer. The objective function $L_{all}$ is expressed as

$$L_{all} = \lambda_1 L_{season} + \lambda_2 L_{p-style} + \lambda_3 L_{garment} \quad (1)$$

where $\lambda_1$, $\lambda_2$, $\lambda_3$ are weights, $L_{season}$ and $L_{\text{garment}}$ represent the corresponding cross entropy losses, and $L_{p-style}$ is the confusion loss with distance constraint. The confusion loss will be explained in detail in the next section.

The multi-task representation learning model can make full use of multiple kinds of the label information. It is an effective method to solve the problem of fine-grained classification in the fashion field. In addition to the fashion field, Model StyleNet can also be applied in other fields. However, in the fashion field, the model faces two important challenges. One is the lack of high-quality public data sets, so it is difficult to obtain a better model based on existing data sets. The other is that there are even semantic intersections and conflicts between some tag values. We will solve these problems from the perspective of objective function optimization.

### B. LOSS FUNCTION OPTIMIZATION

The following will be divided into three parts to introduce the optimization method of the objective function. First, we will introduce two kinds of commonly used loss functions and then propose an improved method.

#### 1) CROSS ANDENTROPY LOSS

The square root error is widely used as the loss function in the early neural networks. With the rise of deep learning, cross entropy loss function becomes a recognized effective loss function [23]. It is expressed as

$$L_{CE} = -\frac{1}{N}\sum_{n=1}^{N} y_n^T \log P_n \quad (2)$$

where $y_n$ is the category label, $\log(\cdot)$ is the bitwise logarithmic function, and $P_n \in R^k$ is the output of softmax function. By using softmax function, the neural network can provide a classification probability distribution.

Since our goal is to match the distribution of the results produced by the trained model to the real distribution, we guide the model by minimizing the cross entropy to learn a probability distribution close to the target distribution. The cross entropy is also called negative log likelihood, and in information theory, it is closely related to Kullback-Leibler (KL) divergence, which is a measure for evaluating the distance between two distributions. It is expressed as

$$D_{KL}(p||q) = \sum_{x\in X} p(x) \log \frac{p(x)}{q(x)} \quad (3)$$

Assume $p$ is a true distribution, then the information entropy $H(p)$ of $p$ is a certain value. At this time, the cross entropy is equivalent to the relative entropy, i.e., KL divergence $D_{KL}$. During the optimization process, when the cross entropy becomes small, KL divergence will also become small.

Therefore, in Eq. (2), $y_n$ can be considered as the target probability distribution and the goal of training with cross entropy loss function is to make the probability distribution of the network output as close as possible to the target distribution. Compared to the minimum mean square error function, cross entropy loss function has fewer flat regions, so the network is easier to train and more likely to jump out of local minimum points [24].

Since cross entropy loss function will make the network learn a target distribution close to the data set as much as possible in the training process, when the information in the data set is richer and the coverage is wider, the trained model will be more accurate and the confidence will be higher. This is one of the reasons why large data sets are needed to train deep learning models. Unfortunately, compared with other image classification fields, there are fewer public data sets in the fashion field and their quality is not high, so we need to find a way to make up for this shortcoming.

#### 2) DISTANCE CONFUSION CONSTRAINTS

Convolutional neural networks benefit from the convolution kernel structure and can learn more visual information about images. In the classification application of fashion images, there is such a phenomenon where images of the two samples are very similar but labeled by different tags. It is very difficult for some general classification models to solve this problem. Confusion training is a method that adds Euclidean distance constraint between samples during model training and force the neural network model to distinguish the features of two images with high confidence [25].

For a neural network with parameter $\theta$, after accepting an image $v$, the probability distribution $p_\theta(y|v)$ of the image under N categories can be generated. In this paper, a new confusion loss function $L_p$ is defined as

$$L_p(v_1, v_2; \theta) = \sum_i (p_\theta(y_i|v_1) - p_\theta(y_i|v_2))^2 \quad (4)$$

where $v_1$ and $v_2$ are single samples randomly selected in the training set, and they belong to different categories.

By applying a penalty of Euclidean distance to different categories of samples in the process of learning, this method can help neural network narrow the distance of different categories to avoid confusing a single cross entropy loss. The model can focus on the confusing visual part of the sample during training to minimize the training error. Especially in the fine-grained image classification task, it can highlight this advantage. Moreover, over-fitting can also be prevented to some extent. Fig. 3 shows the calculation process of applying distance confusion constraints.

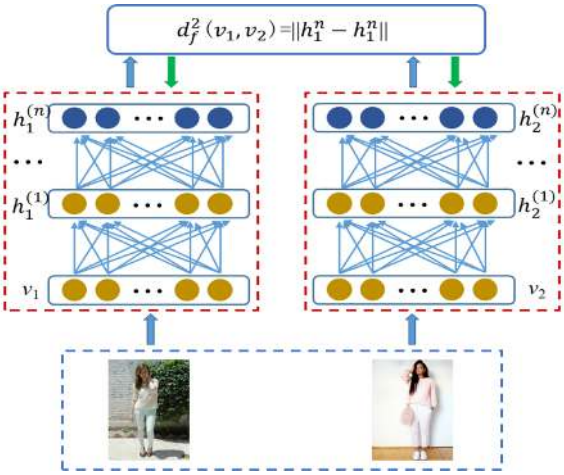


**FIGURE 3.** The calculation process of distance confusion constraint. Two images are randomly selected from the training set as input and they belong to different categories. Samples from different categories are imposed a penalty of Euclidean distance so that they can be pulled a little closer to confuse the single cross entropy loss.

### 3) OPTIMIZED LOSS FUNCTION

For the image classification task of a specific domain, the distribution of pixels in the image dataset will have distinct domain characteristics.

Fig. 4 shows an example, where we do the mean processing for the left three samples and get the rightmost mean image. The shape of the portrait of the rightmost image can be clearly seen. Based on such inspiration, when a deep network is used to realize the classification task of the clothing images, the loss function can be adaptively adjusted according to the characteristics of the data set itself.

For such a data set, it is very challenging to let the model learn the abstract label of clothing style or season. People who do not know much about fashion cannot accurately label the style of the clothes. In some cases, it is a difficult question to distinguish whether clothing belongs to spring or autumn.

For these specific problems of clothing image classification, simply adjusting the network structure of the model does not have much performance improvement. In addition, due to the characteristics of cross entropy loss function, in the process of continuous iteration, the network may learn some special weights for some images that are visually similar but belong to different classes. In this case, the problem of overfitting is more likely to occur. According to the above analysis, we will combine distance confusion loss and the traditional cross entropy loss to form a new loss function $L_c$ as follows:

**FIGURE 4.** The three clothing images on the left are samples and the rightmost image is the mean processing result of the data set. The left three images above show the landscape and the left three images below are from the fashion field. The right image below shows a clear portrait outline while the right image above cannot be seen.

$$L_c \Leftarrow L_{CE}(v_2; \theta) + \gamma \cdot I \cdot L_p(v_1, v_2; \theta) \tag{5}$$

where $L_{CE}(\cdot)$ is the traditional cross entropy loss function, $L_P(\cdot)$ is used to calculate Euclidean distance between a given set of features, $v_1$ and $v_2$ are randomly sampled from the data set for calculating the corresponding Euclidean distance, respectively, $\theta$ is the internal parameter of the neural network, $\gamma$ is the penalty parameter brought by the added Euclidean distance, and $I$ is the indication function.

If the samples from $v_1$ and $v_2$ belong to the same category, then the value of $I$ is set to 0, otherwise $I$ is set to 1. During the model training process, the Euclidean distance is calculated by randomly selecting two subsets from the dataset, and then the distance metric is added to the final objective function. Finally, the model is trained together to achieve the purpose of confusion. The training process with the optimized loss function is shown in Algorithm 1.

During the training process, the sample is divided into two parts. Each part is trained individually. First, the same method is used to calculate the cross entropy loss of the two subsets, and then the distance loss is calculated for the sample pair, at last the two losses are used together as the objective function of the training.

During the testing process, only one branch of the network is active and it can output the prediction result for the input image, so adding confusion training does not bring additional computational overhead for model testing or classification task.

## V. EXPERIMENTS AND EVALUATION

### A. EXPERIMENT SETTING

#### 1) DATA SETS

To evaluate the related machine learning methods in the fashion field, three data sets, Fashionista [26], Fashion 144k [27], and SFS [28] are widely used. The detailed information

**Algorithm 1** Model Training Algorithm With the Optimized Loss Function

Input: training set $D$, test set $\hat{D}$, iteration number max_iter
Output: parameters $\theta$ and hyper-parameter $\hat{\theta}$
01: Initialize $\theta$ and $\hat{\theta}$ randomly;
02: for j $\in \lceil 0, max\_iter \rceil$
03: $D_1 \Leftarrow$ shuffle data set $D$;
04: $D_2 \Leftarrow$ shuffle data set $D$;
05: for i $\in$ [0, batch size]
06: $L_{batch} = 0$;
07: for $(v_1, v_2)_i \in (D_1, D_2)$
08: $I \Leftarrow 1$ if the label of $v_1 \neq$ the label of $v_2$;
09: $L_c \Leftarrow L_{CE}(v_1; \theta) + L_{CE}(v_2; \theta) + \lambda \cdot I \cdot L_p(v_1, v_2; \theta)$;
10: $L_{batch} = L_{batch} + L_c$;
11: end for
12: minimize $L_{batch}$;
13: update $\theta$;
14: end for
15: update $\hat{\theta}$;
16: end for

is shown in Table 1. Although these three data sets are published by different authors, they come from the same source Chictopia, which is the largest style community in the world where bloggers share style posts and online clothing. According to the literature, there are very few data sets in the fashion field and the sources of these data sets are very single. This problem severely hinders the application of artificial intelligence in the fashion field. High-quality image datasets such as ImageNet created by professional teams have promoted the development of the image field [29]. We hope that in the near future, there will also be teams that can create professional data sets in the field of fashion.

**TABLE 1. Detailed statistics of data sets in the fashion field. they come from the same source chictopia. Dataset SFS is the latest one and its label information is also the most abundant.**

| Data set | Source | Size(k) | Information | Quality |
|---|---|---|---|---|
| Fashionista | chictopia.com | 150 | postures, items, etc. | medium |
| Fashion 144k | chictopia.com | 30 | items, user comments | medium |
| SFS | chictopia.com | 290 | style, items, etc. | high |

In the subsequent experiments data set SFS is chosen to verify our model because it is the latest one among these three data sets and the label information of SFS is also the most abundant. In order to further improve the data quality, we select about 80,000 images with perfect label information and high image quality from it. Although the multimodal representation method of combining multiple kinds of features such as text and image is more effective [30], the intuitiveness of image prompts us to study it. We divide the training set and test set into 8:2 in a random manner.

#### 2) MODEL COMPARISON

With the development of deep learning, the improvement in model structure has developed rapidly. AlexNet uses ReLU function as the activation unit, adds dropout technique to selectively ignore individual neurons to avoid overfitting, and chooses the largest pool to avoid the average effect of average pooling. VGGNet deepens on the network structure by using multiple $3 \times 3$ convolution kernels to mimic a larger receptive field. GoogLeNet uses a network structure called inception which splices convolution kernels of different sizes and merges features of different scales. By learning the residual function, ResNet allows the original input information to be directly transmitted to the subsequent layers so as to ensure the completeness of the information and make the depth of the network a breakthrough. The achievements of these classic network structures in the ILSVRC competition are sufficient to demonstrate their success in image classification.

The main problem discussed in this paper is how to make the computer better understand the clothing image and abstract the conceptual labels such as styles under the problem of fine-grained classification. In the experiment, we will verify the improvement by adding the confusion loss on the classic CNN models, verify the improvement by applying the multi-task representation learning framework, analyze the influence of different data set size, and analyze the influence of Faster R-CNN. All of these experiments will focus on the feasibility and effectiveness of multitasking frameworks and loss function optimization.

#### 3) EXPERIMENTAL PLATFORM AND ENVIRONMENT

All experiments are conducted on a server with Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz, Nvidia®Tesla K80 GPU, 128GB RAM, and 1TB hard drive. The operating system is Ubuntu 16.04. The experiment runs under Caffe platform and we implement all models in Python.

### B. RESULTS AND ANALYSIS

The following four experiments will be carried out to verify our model and method proposed in the paper.

#### 1) VERIFY THE IMPROVEMENT BY ADDING THE CONFUSION LOSS ON THE CLASSIC CNN MODELS

The four CNN models mentioned above are implemented and trained with the optimized loss function. We set the value of the hyper-parameter $\gamma$ to 0.01. Two kinds of loss function are adopted, the cross entropy loss function with and without distance confusion constraints. The classification accuracy results are shown in Table 2 and Table 3. Table 2 is the experimental result of classifying the seasonal labels and Table 3 is the result of classifying the style labels. In addition, Table 3 counts the accuracy of Top-5. Here, the multi-task learning framework is not applied in the models.

**TABLE 2. Classification results of seasonal labels on four classic CNN models. Two kinds of loss functions are applied in these models. The representation results produced by different loss functions have different effects on the classification.**

| Deep neural network | Softmax classification | Confusion added | Percentage increased |
|---|---|---|---|
| AlexNet | 0.4028 | 0.4532 | 12.5 |
| VGGNet | 0.4169 | 0.4687 | 12.4 |
| GoogleNet | 0.4292 | 0.4713 | 9.81 |
| ResNet-50 | 0.4348 | 0.4734 | 8.87 |

**TABLE 3. Classification results of style labels on four classic CNN models. Two kinds of loss functions are applied in these models. The representation results produced by different loss functions have different effects on the classification.**

| Deep neural network | Softmax classification | Confusion added | Percentage increased | Top-5 accuracy |
|---|---|---|---|---|
| AlexNet | 0.2029 | 0.2571 | 26.7 | 0.7503 |
| VGGNet | 0.2121 | 0.2625 | 23.8 | 0.7531 |
| GoogleNet | 0.2146 | 0.2753 | 28.2 | 0.7685 |
| ResNet-50 | 0.2152 | 0.2789 | 29.6 | 0.7662 |

According to the experimental results in Table 2 and Table 3, we can get the following conclusions.

1) In large classification tasks, the deeper network models are proven to produce better results. For the four models AlexNet, VGGNet, GoogleNet, and ResNet-50, the deeper their network structure is, the better the classification effect is. Although the classification effect in Table 3 is not ideal, it also follows this rule.
2) We adopt softmax as the classifier in the experiment. Due to the lack of label semantics, the accuracy of models trained with the cross entropy loss function is relatively low. When the loss function is optimized by adding confusion constraint to train the models, the experimental results are greatly improved. The improvement effect is especially noticeable in models with a shallower network structure, e.g., AlexNet gains 12.5% accuracy improvement for seasonal label classification. Therefore, for the fine-grained classification task similar to clothing image classification, it is not enough to simply use the cross entropy loss to train models. It is a solution to design a new objective function to complicate the training objectives of the network.
3) Understanding the style of clothing is a more abstract issue. According to the values in Table 3, we can find that it is hard to obtain a good classification effect by using the deep learning model, regardless of whether the network structure of the model is sufficiently deep. However, after adding the confusion constraint to the loss function, the overall performance is significantly improved. Moreover, for models with deeper network structures, the effect is even better. The accuracy of Top-5 rate is even more than 75% after the models are trained with the optimized loss function.

### 2) VERIFY THE IMPROVEMENT BY APPLYING THE MULTI-TASK REPRESENTATION LEARNING FRAMEWORK

We implement VGGNet using the optimized loss function proposed in this paper. For convenience, we call it Vgg-Lp. We also implement model StyleNet, which adopts multi-task learning framework and uses Vgg-Lp in the hidden layer to extract features. In fact, StyleNet applies a multi-task framework and Vgg-Lp just applies a single-task framework.

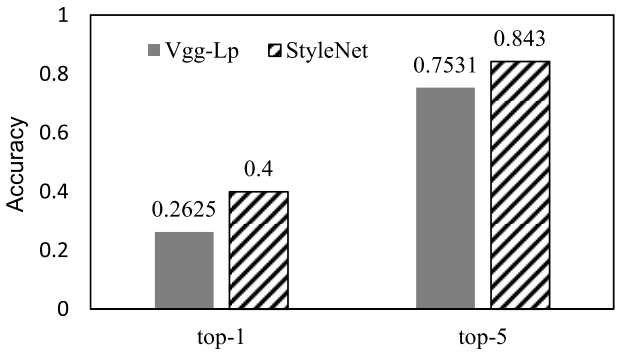


**FIGURE 5. Comparison of StyleNet and Vgg-Lp. StyleNet adopts a multi-task representation learning framework and Lgg-Lp adopts a single-task representation learning framework.**

Fig. 5 shows the accuracy comparison between Vgg-Lp and StyleNet. The values of parameters $\lambda_1$, $\lambda_2$, and $\lambda_3$ in StyleNet are set to 0.2, 0.3, and 0.5, respectively. We use 0.1 as the step size and take the best results after multiple tests. From Fig. 5, we can see that model StyleNet works much better than Vgg-Lp. The classification accuracy of StyleNet is increased by 52% for Top-1 and 11% for Top-5.

### 3) ANALYZE THE INFLUENCE OF DIFFERENT DATA SET SIZE

In order to analyze the impact of different data set size on model StyleNet, we randomly divide the training data set into 5 equal parts and then take 1/5, 2/5, 3/5, 4/5 data samples to train the model. Because of the particularity of the style labels, we choose to classify the season labels. Fig. 6 shows the experimental result.

### 4) ANALYZE THE INFLUENCE OF FASTER R-CNN

Faster R-CNN is an image target detection model based on deep learning, and it is also a well-recognized model. Generally, the effect of the image has a large impact on the classification model. This section will examine whether Faster R-CNN module contributes to the pre-processing of data sets. In the experiment, two data sets were used, one is the original data set and the other is processed by Faster R-CNN. Table 4 and Table 5 show their classification results of seasonal labels and style labels, respectively.

As depicted in Table 4 and Table 5, after adding Faster R-CNN module for data set pre-processing, the classification

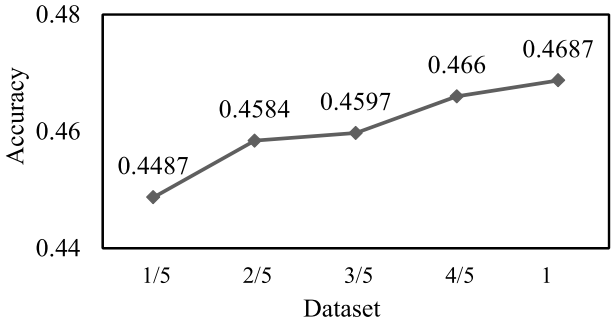


**FIGURE 6.** Comparison of accuracy for data sets of different sizes. The x axis represents the size of the training data set. The total training data set is randomly divided into 5 equal parts.

**TABLE 4.** Classification results comparison of seasonal labels by adding Faster R-CNN module. All models are trained by using the optimized loss function. The input images in column three are pre-processed by Faster R-CNN.

| Deep neural network | Original images | Faster R-CNN added | Percentage increased |
|---|---|---|---|
| AlexNet | 0.4532 | 0.4673 | 3.1 |
| VGGNet | 0.4687 | 0.4803 | 2.5 |
| GoogleNet | 0.4713 | 0.4824 | 2.4 |
| ResNet-50 | 0.4734 | 0.4852 | 2.5 |

**TABLE 5.** Classification results comparison of style labels by adding Faster R-CNN module. All models are trained by using the optimized loss function. The input images in column three are pre-processed by Faster R-CNN.

| Deep neural network | Original images | Faster R-CNN added | Percentage increased |
|---|---|---|---|
| AlexNet | 0.2571 | 0.2587 | 0.6 |
| VGGNet | 0.2625 | 0.2691 | 2.5 |
| GoogleNet | 0.2753 | 0.2803 | 1.8 |
| ResNet-50 | 0.2789 | 0.2814 | 0.8 |

accuracy is improved. For seasonal labels, the accuracy rate is increased by 2.6% on average, and for style labels, the accuracy rate is increased by 1.4% on average. Comparing to the average improvement of 10.9% and 27.1% by using the optimized loss function, the effect of adding Faster R-CNN processing can be ignored. Further, if we deepen the network depth of the model, we can achieve better performance than Faster R-CNN.

This experiment illustrates a phenomenon from the side, i.e., as the iteration deepens, the deep learning model can gradually focus and adjust the image automatically when extracting features from images. Therefore, when using a deep learning model for feature representation and classification, it is not necessary to pre-process the image because the image features it learns are robust and effective.

## VI. CONCLUSIONS

Nowadays, the effectiveness and performance of deep learning is widely recognized and how to make advantage of deep learning methods in different fields is a research hotspot. The understanding of clothing images and classification is a good area for verifying machine intelligence. However, the efficiency of deep learning methods applied in the fashion field is restricted by the lack of public data sets of high quality. Although we can occasionally find very some data sets on the Internet, their size is small and the label information is incomplete. These problems hinder further research. In this paper, a multi-task learning model StyleNet based on the deep neural network is proposed to represent clothing images. We use several different types of labels to train the model and adopt an optimized loss function which is the combination of distance confusion loss and traditional cross entropy loss. By adopting a multi-task framework and optimizing the loss function, model StyleNet can minimize the problem of insufficient semantic information of clothing images so as to improve the classification accuracy.

During the process of data set selection, we found that the quality of the clothing image data set is still unsatisfactory, which affects the performance of model StyleNet. In addition, the attribute values of style labels in some data sets have a certain degree of semantic overlap, so our future work will focus on improving the quality of data sets by integrating more information and further verifying the validity of our model.

## REFERENCES

[1] C. G. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. 4th Vis. Conf.*, Alvey, U.K., 1988, pp. 147–152.
[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[3] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. CVPR*, San Diego, CA, USA, Jun. 2005, pp. 886–893.
[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 2, pp. 84–90, Jun. 2012.
[5] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. CVPR*, Boston, MA, USA, Jun. 2015, pp. 1–9.
[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, Las Vegas, NV, USA, 2016, pp. 770–778.
[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[8] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.
[9] A. Saha, M. Nawhal, M. M. Khapra, and V. C. Raykar, "Learning disentangled multimodal representations for the fashion domain," in *Proc. WACV*, Lake Tahoe, NV, USA, Mar. 2018, pp. 557–566.
[10] A. Taheri, K. Gimpel, and T. Berger-Wolf, "Learning graph representations with recurrent neural network autoencoders," in *Proc. KDD Deep Learn. Day*, London, U.K., 2018, pp. 1–8.
[11] Y. Ma, J. Jia, S. Zhou, J. Fu, Y. Liu, and Z. Tong, "Towards better understanding the clothing fashion styles: A multimodal deep learning approach," in *Proc. AAAI*, San Francisco, CA, USA, 2004, pp. 91–110.
[12] W.-L. Hsiao and K. Grauman, "Learning the latent 'look': Unsupervised discovery of a style-coherent embedding from fashion images," in *Proc. ICCV*, Venice, Italy, Oct. 2017, pp. 4213–4222.
[13] J. Jia *et al.*, "Learning to appreciate the aesthetic effects of clothing," in *Proc. AAAI*, Phoenix, AZ, USA, 2016, pp. 1216–1222.
[14] P. Date, A. Ganesan, and T. Oates. (2017). "Fashioning with networks: Neural style transfer to design clothes." [Online]. Available: https://arxiv.org/abs/1707.09899
[15] K. Vaccaro, S. Shivakumar, Z. Ding, K. Karahalios, and R. Kumar, "The elements of fashion style," in *Proc. UIST*, Tokyo, Japan, 2016, pp. 777–785.

[16] A. Maurer, M. Pontil, and B. Romera-Paredes, "The benefit of multitask representation learning," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2853–2884, 2013.
[17] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *Proc. ICANN*, Rhodes, Greece, 2018, pp. 270–279.
[18] E. Cambria and B. White, "Jumping NLP curves: A review of natural language processing research," *IEEE Comput. Intell. Mag.*, vol. 9, no. 2, pp. 48–57, May 2014.
[19] Z. Tang *et al.*, "Collaborative joint training with multitask recurrent model for speech and speaker recognition," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 3, pp. 493–504, Mar. 2017.
[20] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proc. ICML*, Helsinki, Finland, 2008, pp. 160–167.
[21] X. Yang, H. Sun, X. Sun, M. Yan, Z. Guo, and K. Fu, "Position detection and direction prediction for arbitrary-oriented ships via multitask rotation region convolutional neural network," *IEEE Access*, vol. 6, pp. 50839–50849, 2018.
[22] M. Rezaei, H. Yang, and C. Meinel, "Instance tumor segmentation using multitask convolutional neural network," in *Proc. IJCNN*, Rio de Janeiro, Brazil, Jul. 2018, pp. 1–8.
[23] X. Chen, S. Kar, and D. A. Ralescu, "Cross-entropy measure of uncertain variables," *Inf. Sci.*, vol. 201, pp. 53–60, Oct. 2012.
[24] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. AISTATS*, Sardinia, Italy, 2010, pp. 249–256.
[25] A. Dubey, O. Gupta, P. Guo, R. Raskar, R. Farrell, and N. Naik, "Pairwise confusion for fine-grained visual classification," in *Proc. ECCV*, Munich, Germany, 2018, pp. 70–86.
[26] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Parsing clothing in fashion photographs," in *Proc. CVPR*, Providence, RI, USA, Jun. 2012, pp. 3570–3577.
[27] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun, "Neuroaesthetics in fashion: Modeling the perception of fashionability," in *Proc. CVPR*, Boston, MA, USA, Jun. 2015, pp. 869–877.
[28] X. Gu, Y. Wong, P. Peng, L. Shou, G. Chen, and M. S. Kankanhalli, "Understanding fashion trends from street photos via neighbor-constrained embedding learning," in *Proc. 25th ACM Int. Conf. Multimedia*, Mountain View, CA, USA, 2017, pp. 190–198.
[29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR*, Miami, FL, USA, Jun. 2009, pp. 248–255.
[30] G. Chen and S. N. Srihari. (2015). "Generalized K-fan multimodal deep model with shared Representations." [Online]. Available: https://arxiv.org/abs/1503.07906

**CAIRONG YAN** received the Ph.D. degree in computer science from Xi'an Jiaotong University, China, in 2006. She is currently an Associate Professor with the School of Computer Science and Technology, Donghua University of China. Her research interests include cloud computing, big data, and machine learning.

**LINGJIE ZHOU** is currently pursuing the master's degree with the School of Computer Science and Technology, Donghua University, China. His research interests include machine learning and image processing.

**YONGQUAN WAN** is currently a Lecturer with the College of Information Technology, Shanghai Jian Qiao University, China. His research interests include machine learning and recommender systems.