

Received September 17, 2020, accepted October 28, 2020, date of publication November 4, 2020, date of current version November 17, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3035781

A Novel Clothing Attribute Representation Network-Based Self-Attention Mechanism

YUTONG CHUN^{1,2}, CHUANSHENG WANG¹, AND MINGKE HE^{1,2}

¹School of Management and Engineering, Capital University of Economics and Business, Beijing 100070, China

²School of Logistics Management, Beijing Wuzi University, Beijing 101149, China

Corresponding authors: Chuansheng Wang (wangcs2020paper@163.com) and Mingke He (hemingke@vip.sina.com)

This work was supported by the Key Project of National Social Science Foundation of China under Grant 20AJY016.

ABSTRACT As highly increasing of on-line fashions retail industry, automatic recognition and representation of clothing items have huge potentials. With the help of deep learning methods, many clothing attribute representation models have been proposed. However, these models are mainly suitable for coarse-grained classification which are not suitable for clothing attribute representation. To address such a problem, in this article, we propose a novel network structure named SAC, which is a combination of CNNs and Self-attention mechanism and can represent clothing attributes more fine-grained. Besides, we use Grad-CAM to visualize which part of the clothing attributes is more concerned by customers. Finally, a new labeled clothing dataset is introduced in this article, which is expected to be helpful to the researchers who are working in fashion domains for image representation.

INDEX TERMS Clothing attribute, representation, self-attention mechanism, convolutional neural networks (CNNs), new dataset.

I. INTRODUCTION

With the rise of business-to-consumer (B2C) e-commerce website such as Amazon, Ebay and Tmall, online shopping for clothing has become a general situation. Online images are commonly used as visual cue to attract consumers' attention to enhance their perception of product understanding, which has direct impact on their purchasing decisions [1]–[3]. How to efficiently represent clothing elements from online images has become one of the research traction in the past couple of years. One major reason is that it can help to examine fashion from more diverse perspective and explore more intelligence in the field of fashion apparel, such as retrieve similar or identical fashion items from e-commerce website. However, Clothing images vary in style and different customers have different understanding of style. Different brands of clothing may deliver different type of clothing images on B2C e-commerce website. Some images will have mannequin and some may just contain clothing. Several examples of images from the fashion dataset are shown in Fig. 1. Even with the current popular deep learning methods, it is difficult to accurately classify style labels.

Traditional clothing image classification technologies mainly focus on extracting global features such as color, shape, texture through ingeniously designed feature extraction algorithms and use these features or their combination as

The associate editor coordinating the review of this manuscript and approving it for publication was Taous Meriem Laleg-Kirati¹.

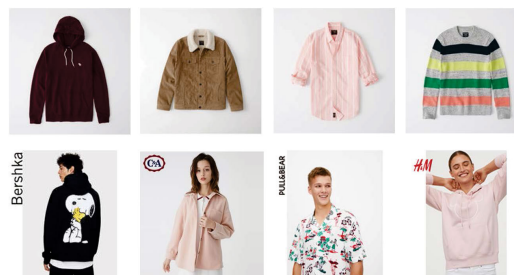


FIGURE 1. Examples from the fashion dataset.

input for classification [4]. The commonly features include SIFT [5], HOG [6] and LAB space [7]. However, these technologies have three problems. Firstly, they heavily rely on the experience of experts to give heuristic labels. Secondly, multiple feature extraction algorithms are time-consuming and resource intensive. Thirdly, feature extraction of clothing images can be affected by many noise factors such as light changing, complex background and promotion information, so it is hard to deal with such complex situation by hard coding an algorithm.

With the development of deep learning in the field of computer vision, the Convolutional Neural Networks (CNNs) have become a research hotspot. Most of CNNs are structured as multi-layered models and learned through large-scale data with a large number of iterative calculations to improve accuracy. The classic deep learning models for image recognition include VGG16 [8], ResNet50 [9], Inception [10] and

DenseNet121 [11], all of which have achieved remarkable grades in ILSVRC competition. The classic CNNs models have strong generalization ability. However, these models are mainly suitable for coarse-grained classification such as shirt or coat. In the fashion field, the purpose of clothing representation is to determine pattern, color even shape of collar, which is the classification task that belongs to a fine-grained classification. How to achieve better performance based on classic CNNs is the issue of concern in this article.

Motivated by the above needs and practice in clothing attribute representation, we explore to describe fashion style by developing a novel network structure named SAC (Self-attention and CNNs), which is a combination of CNNs and Self-attention mechanism and can classify attributes more accurate. To evaluate the ability of our network, we have created a dataset containing 15,025 fashion outfits from online shopping website and annotated all outfits with attributes. The attributes were selected from key attributes frequently retrieved in fashion learning field. Our major contributions are as follows:

1) A novel clothing style extraction network. We introduce a novel clothing style extraction network named SAC which improve the predicting accuracy by self-attention mechanism. The model is more suitable for describing fashion attributes.

2) Big-scale fashion dataset. We collected a big-scale annotated dataset of fashion, which contains tens of thousands of images and hundreds of attributes. We believe many applications can benefit from this dataset.

3) Viewing clothing attributes. We show which parts of clothing have more contribution on classification. It can verify effectiveness of the model and also be helpful to clothing designer modifying their content and making their clothing more attractive.

The rest of this article is organized as follows. Section 2 presents a more detailed literature review and demonstrates clearly the literature positioning of this article. Section 3 shows the proposed model. Section 4 validates our approach on the collected clothing dataset by providing both qualitative and quantitative evaluations. Section 5 concludes the paper and proposes future research.

II. RELATED WORK

A. ATTRIBUTE REPRESENTATION LEARNING

With the continuous development of social network and mobile computing, there are more and more images which can provide more hidden and anomalous information in the fashion field. How to maximize the extracted information from the images is a problem to be studied by representation learning. Related to the fashion domain, as a part of computer vision studies, attribute representation learning has been used for image retrieval [7], [12]–[16], clothing match rules [17] and fashion image ranking [18]. Chen *et al.* proposed a fully automated system, which extracted low-level features in a pose-adaptive manner and combined complementary features for learning attribute classifiers [7]. Fu *et al.* used semantic-preserving visual phrases

to address the problem of large scale cross-scenario clothing retrieval [12]. Di *et al.* constructed an attribute vocabulary using human annotations obtained on a fine-grained clothing dataset, and then trained a binary linear SVM for each attribute [13]. All works mentioned above mostly relied on hand-crafted features, such as SIFT, HOG and color histogram etc. The performance of these methods were limited by the expressive power of these features. Some researches turn to powerful CNNs feature driven by specific tasks. Xia *et al.* combined CNNs with multi-task learning to extract features related to customer's favorite clothing attributes [16]. Liu *et al.* proposed a new deep model namely FashionNet which learned clothing features by jointly predicting clothing attributes and landmarks. The network structure of FashionNet is similar to VGG-16, except the last convolutional layer which is replaced by three branches of layers to predict attributes. Besides, they introduced a large-scale clothes dataset named DeepFashion to enlarge existing clothes datasets [14]. He & Chen designed a fast fashion guided clothing image retrieval framework by efficiently converting float formatted features into binary codes to achieve much faster image retrieval without much accuracy reduction [15]. Liu *et al.* investigated clothing match rules based on semantic attributes according to the generative adversarial network (GAN) model, which can generate clothing-match pairs automatically [17]. Wang *et al.* proposed to learn a ranking SPN (sum product networks) to rank pairs of fashion images by modeling the semantic attributes and data-driven attributes of fashion images [18]. Yan *et al.* adopted a multi-task learning framework to build a model named StyleNet, which could make full use of various types of label information to represent the clothing images in finer-grained manner [4]. This article will focus on the fashion field and try to find a more efficient attribute representation learning model for improving the classification accuracy of clothing images.

B. SELF-ATTENTION MECHANISM

Self-Attention, is a mechanism of attention associated with different locations of a single sequence, with the goal of calculating the representation of the sequence [19]. It has been successfully used in many tasks including reading comprehension, semantic role labeling, image classification and scene segmentation [20]–[25]. Vaswani *et al.* proposed a new simple network architecture, the self-attentional Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely [19]. Cheng *et al.* integrated the long short-term memory architecture with self-attention mechanism to render sequence-level networks better at handling structured input [20]. Tan *et al.* presented a simple and effective architecture for semantic role labeling based on self-attention to handle structural information and long range dependencies [21]. The attention mechanism has enjoyed great popularity in the machine translation as well as NLP communities. However, in computer vision, CNNs are still the norm and self-attention just began to slowly creep into

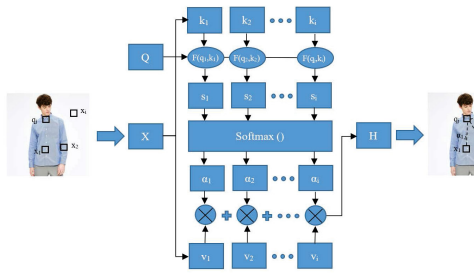


FIGURE 2. Attention mechanism.

the main body of research. Bello *et al.* considered the use of self-attention for discriminative visual tasks and introduced a novel two-dimensional relative self-attention mechanism for image classification [22]. Wang *et al.* proposed an non-local neural networks, which employ self-attention in convolutional architectures to improve video classification accuracy. The networks could be plugged into many computer vision architectures and were initialized in such a way that they did not break pretraining [23]. Woo *et al.* proposed a Convolutional Block Attention Module (CBAM), which sequentially inferred attention maps along two separate dimensions, channel and spatial [25]. Fu *et al.* proposed a Dual Attention Network (DANet) to adaptively integrate local features with their global dependencies based on the self-attention mechanism for scene segmentation [24]. Inspired by the recent successes of the attention mechanism in other domains [23], [26], we integrate CNNs with self-attention mechanism to propose a novel model which could describe clothing appearance with semantic attributes.

III. METHOD

In this section, we describe our model for representing attributes of new clothing. We first propose the classifier which can describe clothing attributes trained by on-sale clothing images. And then, we visualize parts of images that help interpret what features contribute to clothing’s style.

A. CLASSIFIER MODEL

In computer vision field, CNNs are normal and effective method. However, convolution only looks at local spatial information in a certain radius. To improve the classification accuracy, we use self-attention mechanism computing the response at a position as a weighted sum of the features at all positions in the input feature maps.

Attention in the deep learning can be seen as a vector of importance weights, which can enhance convolutional layers to selectively focus on segments of the image. With the help of self-attention, we can know ‘where’ is an informative part. Applying pooling operations along with the channel axis is shown to be effective in highlighting informative regions [25]. As shown in Fig.2, an attention mechanism can be described as mapping an image input sequence $X = [x_1, \dots, x_M]$ of length M to a target output sequence $H = [h_1, \dots, h_M]$ of length M .

Each vector $x_i, i \in [1, M]$ represents a feature map of image. The previous output is compressed into a query (Q of

dimension m) and the next output is produced by mapping this query and the set of keys and values. As shown in Equation (1) and (2), the *att* represents attention mechanism. The output $h_j, j \in [1, M]$ is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key:

$$h_j = att(X, q_j) = \sum_{i=1}^n \alpha_{i,j} x_i \tag{1}$$

$$\alpha_{i,j} = align(x_i, q_j) = softmax(s(x_i, q_j)) = \frac{exp((s(x_i, q_j)))}{\sum_{j=1}^N (exp(s(x_j, q_j)))} \tag{2}$$

The alignment model assigns a score $\alpha_{i,j}$ to the pair of input at position i and output at position j , where $j \in M$ and $i \in M$, based on how well they match. The set of $\alpha_{i,j}$ are weights defining how much of each source hidden state should be considered for each output. As shown in Equation (2), *softmax* is a normalization function. $s(x_i, q_j)$ correspond to alignment score functions. There are two most commonly used alignment score functions, namely additive attention [27], and scaled dot-product attention [19]. We choose the scaled dot-product attention to represent $s(x_i, q_j)$ as it is much faster and more space-efficient [19].

We consider input X as a set of key-value pairs, (K, V) , which means $X = (K, V) = [x_1, \dots, x_M] = [(k_1, v_1), \dots, (k_M, v_M)]$. The K is used to calculate $\alpha_{i,j}$ and the V normally represents the hidden state on conducting the K . So the output vector h_j is:

$$h_j = att((K, V), q_j) = \sum_{i=1}^N \alpha_{i,j} v_i = \sum_{i=1}^N softmax(s(k_i, q_j)) v_i \tag{3}$$

As one of the attention mechanism, self-attention has similar mechanism which relate different positions of a single sequence in order to compute a representation of the same sequence. In another word, the input X equals to output H in self-attention. Given an input tensor of shape (C, H, W, N_{in}) , where C, H, W and N_{in} represent to the channel, height, width and number of input filters of an activation map, we flatten it to a matrix.

We consider the input consists of queries (Q), keys (K) of dimension d_k , and values of dimension d_v . Each x_i of input X is transformed by an observation made by the collection, which means Query = Key = Value. The output is a weighted sum of the values, where the weight assigned to each value is determined by the dot-product of the query with all the keys:

$$H = softmax\left(\frac{QK^T}{\sqrt{n}}\right)V = softmax\left(\frac{(XW_q)(XW_k^T)}{\sqrt{d_k^h}}\right)(XW_v) \tag{4}$$

where $W_q, W_k \in R^{N_{in} \times d_k}$ and $W_v \in R^{N_{in} \times d_v}$ are parameter matrices to be learned, which could map the input X to queries $Q = XW_q$, keys $K = XW_k$ and values $V = XW_v$. With the

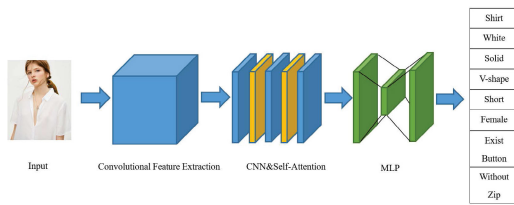


FIGURE 3. The structure of SAC network based on self-attention mechanism.

help of self-attention mechanism, each pixel i will calculate α with each other pixel including itself in input feature map.

Applying multiple self-attention blocks is shown to perform better improvement than single block. Messages can be delivered back and forth between distant positions in spacetime [23]. To enhance the attention of specified clothing attribute, we combine self-attention mechanism and CNNs to propose a new network named SAC (Self-attention and CNNs). We use a 2048-dimensional output of penultimate layer of the ResNet50 network [9] pretrained on the ImageNet dataset [28] to get a high-level image representation for input. To obtain a compact representation from ImageNet feature, we train a five-layer neural network with self-attention mechanism which allows to learn to focus on the most relevant elements in a image(similar to [23]), as shown in Fig.3. In order to improve the attention effect, we apply two self-attention blocks which have an interval of CNN block. Because small spatial size may insufficient to provide precise spatial information, the first CNN will enlarge the receptive field to 4096-dimensional output. Due to attention the network is able to learn which part from the image contributes the most to a clothing attribute. The clothing attributes are achieved after a multi-layer perceptron (MLP) with ReLU activation function as hidden layer.

B. VISUALIZATION

While the SAC can help judge attributes of a clothing, it cannot give any clues which part of a clothing attracts more attention. We propose to visualize how each attributes are judged based on attention. Being able to interpret which part of clothing is relevant to an attribute also can help clothing designer modify their content to better attract customer’s attention.

To that end, we use Grad-CAM [29] to generate heatmaps which can indicate the regions that contribute to clothing attribute in each image. Let $A \in R^{W \times H \times K}$ to be the output of the last convolutional layer in SAC. Firstly, we compute the gradient of a clothing attribute l , with respect to the feature maps activations A^k , *i.e.* $\frac{\partial l}{\partial A^k}$. These gradients are global-average-pooled over the width and height dimintions to obtain feature map importance coefficients α_k :

$$\alpha_k = \frac{1}{W \times H} \sum_i^W \sum_j^H \frac{\partial l}{\partial A_{i,j}^k} \quad (5)$$

We perform a weighted combination of forward activation maps followed by ReLU to obtain the final heatmap.

$$H = ReLU(\sum_k \alpha_k A^k) \quad (6)$$

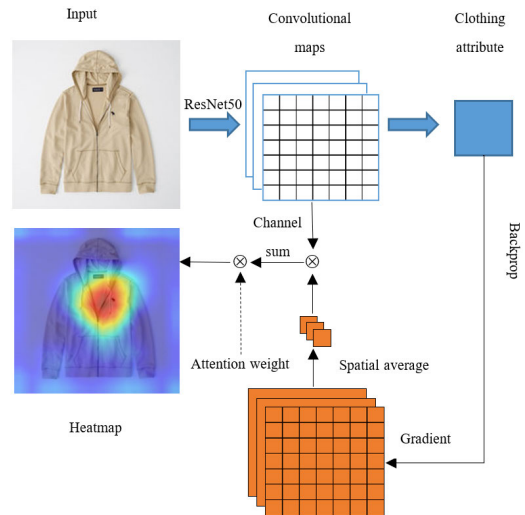


FIGURE 4. Illustration of Grad-CAM algorithm.

We then normalize heatmap values to range [0,1]. For ResNet50 network, this results in 7×7 heatmap of relative importance per image. We use attention weight $\alpha_{i,j}$ to scale the heatmap.

The results of Grad-CAM algorithm applied to a clothing can be seen in Fig.4 [26]. The SAC outputs the probability of a clothing image belonging to the attribute class as score and a set of values for visualization purposes.

IV. EXPERIMENTS

In this section, we describe more details about our new dataset, providing an experimental analysis for evaluating our proposed approach and showcasing a couple of applications.

A. DATASET

In order to train and evaluate our proposed model, we collected a real fashion outfit dataset from the Tmall, which is the largest online shopping website in China. Many famous fast fashion brands like ZARA, H&M and UNIQLO have online store on the Tmall. We crawled fashion outfits from 17 famous fast fashion brands on the Tmall. In this work, due to the lower clothing is occasionally occluded or otherwise not visible in some images, we only consider the upper clothing. After manual data cleaning, we obtained 15,025 images in 5 categories. Examples of fashion dataset are shown in Fig.5.

B. ANNOTATION

Based on the practice of searching fashion outfit from Tmall, we find most of fast fashion brands classifying the upper clothing into 5 categories, namely shirt, sweater, T-shirt, hoodies and outwear. In order to be closer to actual situation, we regard coat and jacket as outwear and top tank as shirt. By referring multiple related works [7], [16], [17], [30], we produce a list of common attributes which are representative of modern fashion trends and cover a large diversity of fashion style. However, the original images collected from online shopping website did not contain attributes information. We invited ten undergraduates to manually annotate

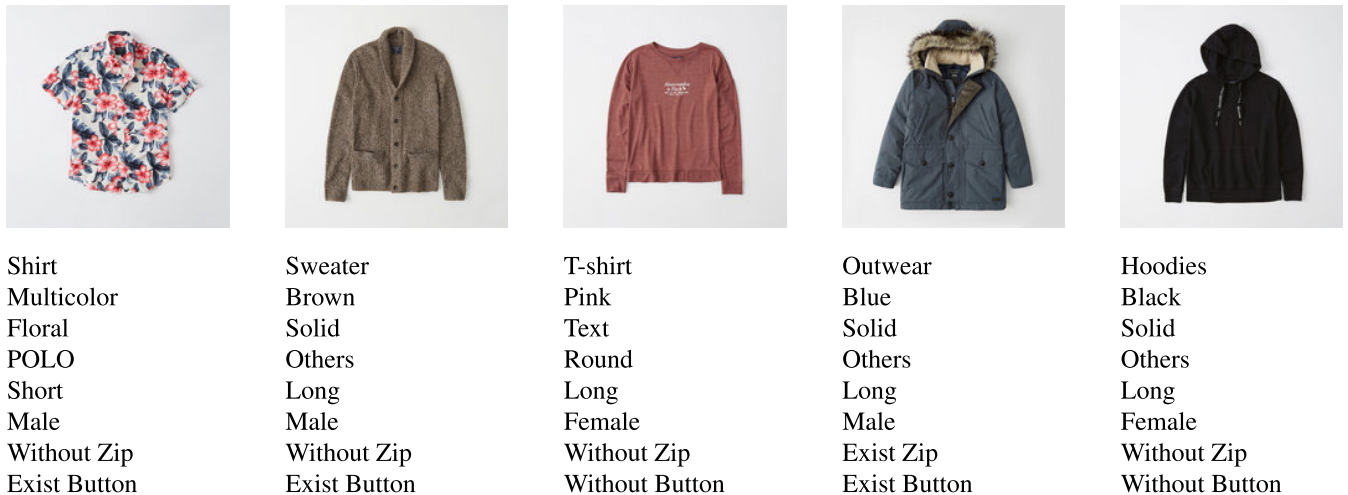


FIGURE 5. Attribute examples in the dataset.

TABLE 1. Statistics of the clothing attribute dataset (There are 8 attributes in total, including 5 multi-class attributes and 3 binary-class attributes).

Attributes	Attribute values
Category	Shirt, Sweater, T-shirt, Outwear, Hoodies White, Yellow, Gray, Green, Blue,
Color	Brown, Red, Cyan, Black, Purple, Multi-color
Pattern	Solid, Floral, Graphics, Spotted, Plaid, Striped, Text, Other patterns
Shape of collar	V-shape, Round, Turndown collar/POLO, Others
Shape of sleeve	Sleeveless, Short, Long
Gender	Male, Female
Zip	Exist, Without
Button	Exist, Without

the attributes of items in order to describe clothing in a variety of views. As shown in Table 1, there are 8 types of clothing attributes in total, including 5 multi-class attributes such as “clothing category” and 3 binary-class attributes like “Button”. The total number of attribute values is 37. The attribute examples of the dataset are shown in Fig.5.

Table 2 shows the cooccurrence matrix of the style attributes. We can find that some pairs of attributes have higher cooccurrence rates. For example, *t-shirt* is more likely to co-occur with *many color* than with other colors. Similarly, *shirt* tend to be *without zip* and *exist button*. It is also more likely to be *long* sleeves and *polo* collar, versus *sleeveless* and *other shape* of collar. Such representations are reasonable and meet common sense.

In addition, we note that there are some potential relationship that may not be easily discovered in daily life. For instance, a clothing with *solid* pattern has more probability to have *long* sleeve and *without zip*, versus other shape of sleeve and *exist zip*. Moreover, a clothing with *round* collar tend to *without zip* and *without button*.

C. PARAMETER SETTING

We use the ResNet50 architecture as feature extraction network because of its widespread use and its ability to easily scale across several computational budgets. In the training

stage, input samples were resized to $180 \times 180 \times 3$. Specifically, we set the number of epochs to 50 with batch size equal to 32. We use the cross entropy loss to train the network for attribute prediction. The network is trained using Adam [31] for stochastic optimization with an initial learning rate of 0.001 and a weight decay of $5e - 4$.

D. COMPARISON AND RESULTS

1) BASELINE

To demonstrate the effectiveness and efficiency of the SAC network, we compare with some recent approaches used in clothing attribute prediction tasks and competing works on deep learning. (1)baseline_AlexNet [30]. The network contains two-stages, one of which is used to learn the pattern of discriminative clothing representation from video and the other is used to predict clothing attributes. We just choose the second predicting clothing attributes stage as comparison baseline whose main construction is the standard AlexNet [32]. (2)baseline_VGG16 [14]. The network proposed in [14] is named FashionNet, which consists three stages, i.e., predicting clothing landmarks, extracting local features via estimated clothing landmarks and fusing local and global features for category and clothing attribute prediction. It needs extra clothing landmarks and category annotations which are missed in our proposed dataset. So we ignore these parts and choose the core prediction part which is the same convolutional parts with VGG16 network [8]. (3)baseline_CNN. To verify the effectiveness of self-attention blocks, we build a same deep network as SAC which has five CNN blocks but no self-attention blocks. (4)baseline_Self-Attention. As proposed in paper [23], self-attention blocks which were named non-local blocks have ability to replace CNN blocks and more blocks in general lead to better results. To verify the combination of CNNs and self-attention mechanism has better performance, we build a same deep network as SAC which just has self-attention blocks.(5)baseline_ResNet101 [9]. ResNets are deep residual learning framework which have been widespread used in

TABLE 2. Attribute-attribute cooccurrence statistics.

shirt	0	0	0	0	0	869	565	121	135	87	400	74	76	36	290	125	206	1758	291	244	24	125	26	104	231	581	1765	201	67	401	2149	1306	1472	2744	34	416	2362
-------	---	---	---	---	---	-----	-----	-----	-----	----	-----	----	----	----	-----	-----	-----	------	-----	-----	----	-----	----	-----	-----	-----	------	-----	----	-----	------	------	------	------	----	-----	------

TABLE 3. Confusion Matrix.

		Prediction Class	
		True	False
Actual Class	True	TP	FN
	False	FP	TN

TABLE 4. Accuracy of clothing attributes prediction. The Bl stands for Baseline.

Attribute	Bl_AlexNet	Bl_VGG16	Bl_Conv	Bl_Self-Attention	Bl_ResNet101	SAC
Gender	0.5076	0.4923	0.4923	0.8789	0.8849	0.8996
Zip	0.8989	0.9010	0.9157	0.9050	0.8896	0.9197
Pattern	0.6354	0.6361	0.4816	0.7979	0.8000	0.8220
Color	0.6795	0.5678	0.5799	0.6976	0.6957	0.6943
Button	0.6568	0.3431	0.3431	0.8214	0.8280	0.8334
Category	0.6408	0.4963	0.6688	0.6695	0.6474	0.6802
Sleeve length	0.9210	0.9143	0.9438	0.9237	0.9043	0.9424
Collar	0.6528	0.5765	0.6816	0.6528	0.6648	0.6896
Average	0.6991	0.6159	0.6383	0.7933	0.7893	0.8102

deep learning recent years. ResNets can gain accuracy with considerably increased depth, which means ResNet101 has better feature extraction performance than ResNet50. However, a deeper network will has more computational cost. To verify the SAC can increase accuracy with less cost, we build a same deep network as baseline_CNN using ResNet101 as feature extraction network.

We train all the baseline approaches using the same amount of data and protocol as SAC did. Furthermore, we also add pre-trained models before all the baselines to certificate the effectiveness of comparison.

2) EVALUATION CRITERION

Because the whole clothing dataset is almost balanced, we use accuracy(ACC) to measure the performance of the SAC network. The criterion is defined as follows [33]:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

All the abbreviations are defined as the confusion matrix shown in table 3.

3) RESULTS

The classification accuracy of each attribute of different methods is shown in Table 4. The dataset is randomly divided into three parts, training set, validating set and testing set at the ratio of 8:1:1.

Seeing from Table 4, the proposed SAC network obtains highest average accuracy, 81.02%. It much outperforms the Baseline_AlexNet, Baseline_VGG16, and Baseline_Conv, which demonstrates the effectiveness of self-attention mechanism. Comparing with Baseline_Self-Attention, the prediction accuracy of SAC is still 1.69% higher than just use

self-attention mechanism, which indicates the combination of CNNs and self-attention mechanism has better performance than single attention mechanism. Besides, although ResNet101 is much deeper than ResNet50, the average accuracy of SAC is 2.09% higher than Baseline_ResNet101, which validates the combination of CNNs and self-attention mechanism can improve performance with less cost.

We also find some interesting phenomenon. Firstly, we observe that a classifier with self-attention mechanism perform better on Gender attribute than the other networks in same depth. This may be because the self-attention mechanism take more attention on significant information about gender. As shown in Fig.1, clothing images from B2C website would be a clothing or clothing with mannequin or brand. Such improvement will be very useful to automatic attribute representation. Secondly, for attribute of Gender, Color, Pattern and Button, the accuracy of SAC, Baseline_Self-Attention and Baseline_ResNet101 are much higher than other methods. Such attributes are important features that need to be attentioned when customers want to buy a clothing. Although Baseline_ResNet101 has similar improvement as Baseline_Self-Attention and SAC, the depth of Baseline_ResNet101 is much deeper than the others, which indicates the self-attention mechanism is useful.

Visualization result are displayed in Fig.6. The red regions correspond to high score for class, while the blue regions correspond to evidence for the class. As shown in Fig.6, we can directly know the regions of an image that provide support for a particular prediction. For instance, regarding attribute of zip and button, the middle part of clothing corresponds to the highest score for classification, which are consistent with people's attention. Besides, we observe that the background color and colorful brand of picture will cause confuse on classifier. On the second row of color attribute, the classifier has great attention on brand and background, which should be the reason why all classifiers do not have good performance on the attribute.

5. APPLICATIONS

(1) AUTOMATIC ATTRIBUTE REPRESENTATION

Our proposed method can be used in automated image annotation. Given a submitted clothing image, the classifier can retrieve attributes automatically and generate a suitable

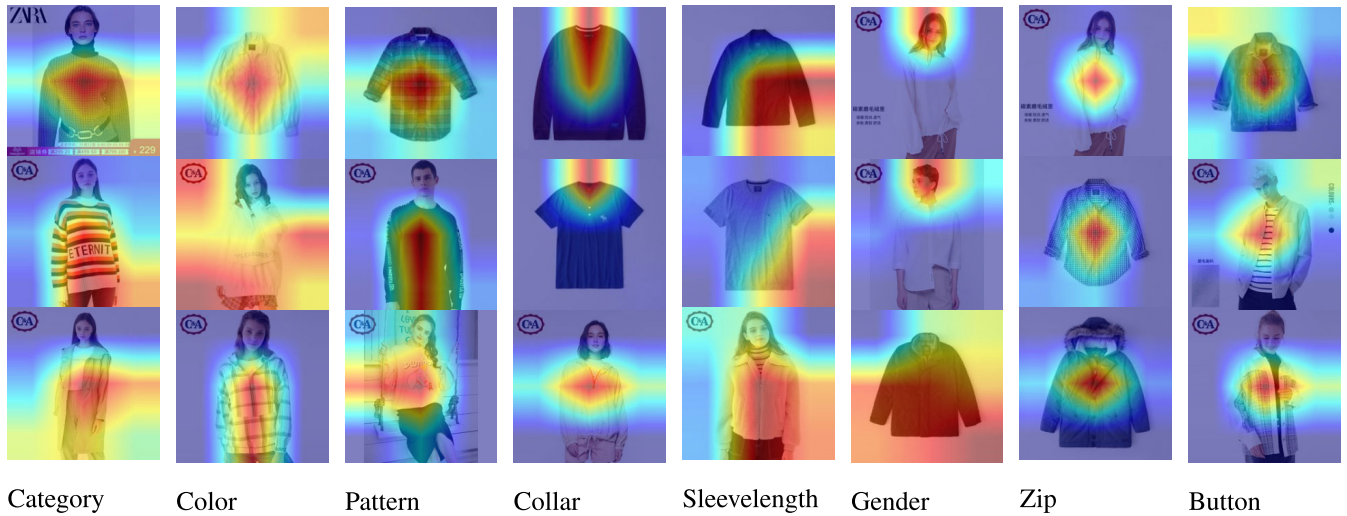


FIGURE 6. Visualization of the region of attention focused by the SAC network.

TABLE 5. Clothing attribute prediction results. The attribute with red indicates that it is judged wrong.

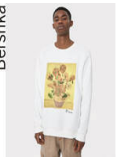

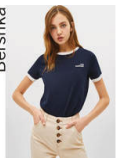

	Baseline_AlexNet	Baseline_VGG16	SAC		Baseline_AlexNet	Baseline_VGG16	SAC	
Bershka 	Sweater	Sweater	Hoodies		Sweater	Sweater	Hoodies	
	Many_colors	Many_colors	Many_colors		Many_colors	Many_colors	Many_colors	Many_colors
	Female	Female	Male		Female	Female	Female	Female
	Round	Round	Round		Other Shapes	Round	Round	Round
	Floral	Other	Graphics		Floral	Other	Text	Text
	Long sleeves	Long sleeves	Long sleeves		Long sleeves	Long sleeves	Long sleeves	Long sleeves
	Without Zip	Without Zip	Without Zip		Without Zip	Without Zip	Without Zip	Without Zip
	Without Button	Without Button	Without Button		Without Button	Without Button	Without Button	Without Button
Bershka 	T-shirt	T-shirt	T-shirt		Shirt	Outwear	Shirt	
	Many_colors	Black	Black		White	White	Gray	Gray
	Female	Female	Female		Female	Female	Male	Male
	Round	Round	Round		Round	Round	Round	Round
	Floral	Other	Solid		Floral	Other	Solid	Solid
	Short sleeves	Short sleeves	Short sleeves		Long sleeves	Long sleeves	Long sleeves	Long sleeves
	Without Zip	Without Zip	Without Zip		Without Zip	Without Zip	Without Zip	Without Zip
	Without Button	Without Button	Without Button		Without Button	Without Button	Without Button	Exist Button

image description. This can be useful for on-line fashions retail to assist sellers in creating listing descriptions.

The qualitative clothing attribute prediction results are shown in Table 5. Comparing with the two existed baseline methods [14], [30], Table 5 shows that our approach is better. The category, color, pattern and collar attributes can be judged by our method more precisely than the baseline methods. These results have also been verified in accuracy results. The wrong inferences of “round” on the male shirt are judged by all methods. In fact, all methods judge correctly. The reason of this error is because the groundtruth is wrongly labeled with ‘round’ value.

2) ARTIFICIAL DESIGN

Another application of our method is artificial fashion design. Fashion designers can label on-sale fashion images based on the number of sales. With a collection of hot-sale clothing photos, our method can be used to analyze and visualize what combination of attributes on a clothing can attract more attention by customers. For example, in our analysis, a hot-sale clothing with solid pattern is more likely to be with

zip than without zip. Such analysis should be helpful to find potential association rules on clothing design.

V. CONCLUSION

In this article, we propose a novel clothing attribute representation network, SAC, which combines self-attention mechanism with CNNs. We also present a new clothing dataset for fashion style prediction of 8 attributes. We conduct an experiment on the SAC network with several baselines from existing works, which proves the effectiveness of our network. Besides, based on Grad-CAM, we visualize which parts contribute to the prediction result, and which parts are useful to fashion designers. However, a limitation of our work is that it cannot generate clothing design draft automatically for fashion designers. Future research can include using generative adversarial networks (GAN) to generate new clothing design draft based on proposed approach.

REFERENCES

[1] J.-H. Ahn, Y.-S. Bae, J. Ju, and W. Oh, “Attention adjustment, renewal, and equilibrium seeking in online search: An eye-tracking approach,” *J. Manage. Inf. Syst.*, vol. 35, no. 4, pp. 1218–1250, Oct. 2018.

- [2] A. J. King, A. J. Lazard, and S. R. White, "The influence of visual complexity on initial user impressions: Testing the persuasive model of Web design," *Behav. Inf. Technol.*, vol. 39, no. 5, pp. 497–510, 2019.
- [3] H. Xia, X. Pan, Y. Zhou, and Z. Zhang, "Creating the best first impression: Designing online product photos to increase sales," *Decis. Support Syst.*, vol. 131, Apr. 2020, Art. no. 113235.
- [4] C. Yan, L. Zhou, and Y. Wan, "A multi-task learning model for better representation of clothing images," *IEEE Access*, vol. 7, pp. 34499–34507, 2019.
- [5] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
- [6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.
- [7] H. Chen, A. C. Gallagher, and B. Girod, "Describing clothing by semantic attributes," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 609–623.
- [8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1–14.
- [9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [10] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
- [11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.
- [12] J. Fu, J. Wang, Z. Li, M. Xu, and H. Lu, "Efficient clothing retrieval with semantic-preserving visual phrases," in *Proc. Asian Conf. Comput. Vis.*, 2012, pp. 420–431.
- [13] W. Di, C. Wah, A. Bhardwaj, R. Piramuthu, and N. Sundaresan, "Style finder: Fine-grained clothing style detection and retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2013, pp. 8–13.
- [14] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "DeepFashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1096–1104.
- [15] Y. He and L. Chen, "Fast fashion guided clothing image retrieval: Delving deeper into what feature makes fashion," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 134–149.
- [16] Y. Xia, B. Chen, W. Lu, F. Coenen, and B. Zhang, "Attributes-oriented clothing description and retrieval with multi-task convolutional neural network," in *Proc. 13th Int. Conf. Natural Comput., Fuzzy Syst. Knowl. Discovery (ICNC-FSKD)*, Jul. 2017, pp. 804–808.
- [17] L. Liu, H. Zhang, Y. Ji, and Q. M. Jonathan Wu, "Toward AI fashion design: An attribute-GAN model for clothing match," *Neurocomputing*, vol. 341, pp. 156–167, May 2019.
- [18] J. Wang, A. A. Nabi, G. Wang, C. Wan, and T. Ng, "Towards predicting the likeability of fashion images," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–10.
- [19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. U. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.
- [20] J. Cheng, L. Dong, and M. Lapata, "Long short-term memory-networks for machine reading," 2016, *arXiv:1601.06733*. [Online]. Available: <http://arxiv.org/abs/1601.06733>
- [21] Z. Tan, M. Wang, J. Xie, Y. Chen, and X. Shi, "Deep semantic role labeling with self-attention," in *Proc. Assoc. Advancements Artif. Intell.*, 2019, pp. 1–8.
- [22] I. Bello, B. Zoph, Q. Le, A. Vaswani, and J. Shlens, "Attention augmented convolutional networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3286–3295.
- [23] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.
- [24] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.
- [25] S. Woo, J. Park, J. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
- [26] A. Bielski and T. P. Trzcinski, "Understanding multimodal popularity prediction of social media videos with self-attention," *IEEE Access*, vol. 6, pp. 74277–74287, 2018.
- [27] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *Comput. Lang.*, vol. 1409, 2014.
- [28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
- [29] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017.
- [30] S. Zhang, S. Liu, X. Cao, Z. Song, and J. Zhou, "Watch fashion shows to tell clothing attributes," *Neurocomputing*, vol. 282, pp. 98–110, Mar. 2018.
- [31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, vol. 1412, 2014.
- [32] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1–9.
- [33] Y. Jo, J. Wi, M. Kim, and J. Y. Lee, "Flexible fashion product retrieval using multimodality-based deep learning," *Appl. Sci.*, vol. 10, no. 5, p. 1569, Feb. 2020.



YUTONG CHUN is currently pursuing the Ph.D. degree with the School of Management and Engineering, Capital University of Economics and Business, Beijing, China. His research interests include machine learning and big data analytics.



CHUANSHENG WANG is currently the Vice President and a Professor with the Capital University of Economics and Business, Beijing, China. His research interests include information systems and management science.



MINGKE HE is currently the Vice President and a Professor with Beijing Wuzi University, Beijing, China. His research interests include supply chain management and management science.