DAtRNet: Disentangling Fashion Attribute Embedding for Substitute Item Retrieval

Gaurab Bhattacharya, Nikhil Kilari, Jayavardhana Gubbi, Bagya Lakshmi V.,
Arpan Pal, Balamuralidhar P.
TCS Research, India
{bhattacharya.gaurab, kilari.nikhil, j.gubbi, bagyalakshmi.v, arpan.pal, balamurali.p}@tcs.com

Abstract

Interactive substitute recommendation for fashion products improves the online retail customer experience. Traditional fashion search platforms incorporate product metadata between the query products and the products to be retrieved. In this paper, we propose DAtRNet, an attribute representation network to disentangle the features in the query product. It is used to recommend attribute-aware substitute items based on the conditional similarity of the retrieved products. The proposed architecture relies on attribute-level similarity providing a fine-grained recommendation. In addition, a concurrent axial attention mechanism is proposed that generates global information embedding and adaptively re-calibrates the soft attention masks. Overall, the end-to-end framework enables the system to disentangle the attribute features and independently deals with them to enhance its flexibility towards one or multiple attributes. The proposed method outperforms the state-of-the-art by a significant margin.

1. Introduction

Fashion item retrieval plays a pivotal role in selecting desired products from e-commerce websites. In different shopping scenarios, users desire to retrieve similar products based on visual appearance. In traditional substitute recommendation systems, similar products are shown to the customer using metadata of the product that are manually generated. In realistic scenarios, there is a need to manipulate the product attributes such as color, style or pattern based on customer suggestions. This is illustrated in Fig. 1, where a customer may look for a substitute shirt with a different color or a substitute shirt with the same color. Such requirements manifest in the profitability of the retail industry and the premise for building the next-generation retail recommendation engine for large-scale fashion item retrieval.

There are several challenges in building an effective recommendation system with attribute feedback; such as overlapping fine-grained attributes, variations in attribute style and visual appearance, small inter-class variation and class imbalance. To address them, several solutions have been proposed [4, 6, 12]. One of the limitations of these works is the entanglement of features where changes in one alter the other. However, for understanding the attributes individually in feature sub-spaces, we need to *disentangle* the feature sub-space into several disjoint attribute-aware sub-spaces. Such sub-spaces must preserve the semantic embedding of the attribute. This aspect of disentanglement of attribute features has received limited attention in substitute recommendation [1–3, 16].

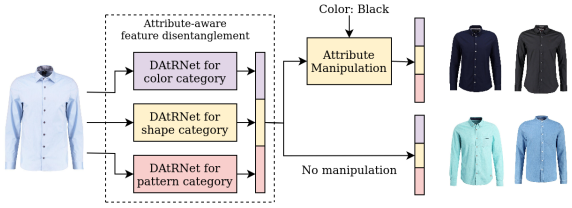

Figure 1. The attribute-aware substitute recommendation. Here, we perform substitute recommendation with (color: blue → black) & without attribute manipulation.

In this paper, we propose Disentangled Attribute Representation Network (DAtRNet) and Attentive Style Embedding module (ASE) to tackle the above mentioned limitations. Our contributions are given below:

1. The Attentive Style Embedding (ASE) module is proposed that tackles visually-similar attributes with low inter-class variance using the multi-scale feature extraction sub-network. The overlapping attributes and style and visual appearance variations are addressed by the concurrent axial attention sub-network.
2. A new architecture, DAtRNet, is proposed that uses ASE modules as fundamental blocks. It generates disentangled attribute-aware style embedding for super-category specific fine-grained attributes.

3. DAtRNet has been shown to successfully incorporate the attribute manipulation to provide an attribute-aware substitute fashion search and outperforms state-of-the-art methodologies by a significant margin on DeepFashion [12] and Shopping100k [3] datasets.

2. Related Work

Attribute-aware substitute recommendation: In recent years, several methods have attempted attribute-aware substitute recommendation [1–3, 16]. However, they consider complex multi-step process, such as localization unit [1–3] or memory unit [16] with simpler feature extraction network without understanding the fine-grained representation in the latent space. In this paper, an attentive style embedding (ASE) module is proposed to extract multi-scale fine-grained features where discriminatory regions across three axes are highlighted using a multi-axis attention mechanism.

Attention based methods: The impact of the channel and spatial regions for effective feature extraction has been investigated in [4, 10, 13, 15]. Recently, Wang *et al*. [14] proposed axial attention to extract features across width and height axes. However, these methods capture features across one axis [4, 8, 10], jointly encode both height and width axes [13, 15] or separately encode features using self-attention [14] without *attending* to the channel features. In our network, we have proposed concurrent axial attention, which separately treats all three axes using *Conv* layers to capture local features and uses global max and average pooling to calibrate channel information adaptively.

3. Proposed Methodology

For attribute-aware substitute recommendation, the core modules should localize and discriminate the fine-grained attributes by disentangling the attribute representation. As a first step, attentive style embedding (ASE) module is proposed to extract local attribute information. Multiple ASE modules are combined to form the proposed disentangled attribute representation network (DAtRNet).

3.1. Attentive Style Embedding Module

The proposed Attentive Style Embedding (ASE) module (Figure 2) extracts fine-grained attribute features across multiple scales and discriminatory regions are selected across height, width and channel dimension for holistic feature representation.

Multi-scale feature extraction: The multi-scale feature extraction sub-network enables DAtRNet to encapsulate fine-grained attribute information to alleviate the inter-class similarity problem. Motivated from [4], this sub-network generates a holistic representation of attributes for better discrimination. We extract the mid-level attribute representation by concatenating responses from two parallel paths involving *Conv* and pooling layers of different kernel sizes to encapsulate the variation of attribute instances. We then send it through a multi-scale feature extraction mechanism to address minute variations and visual similarity across different attributes.

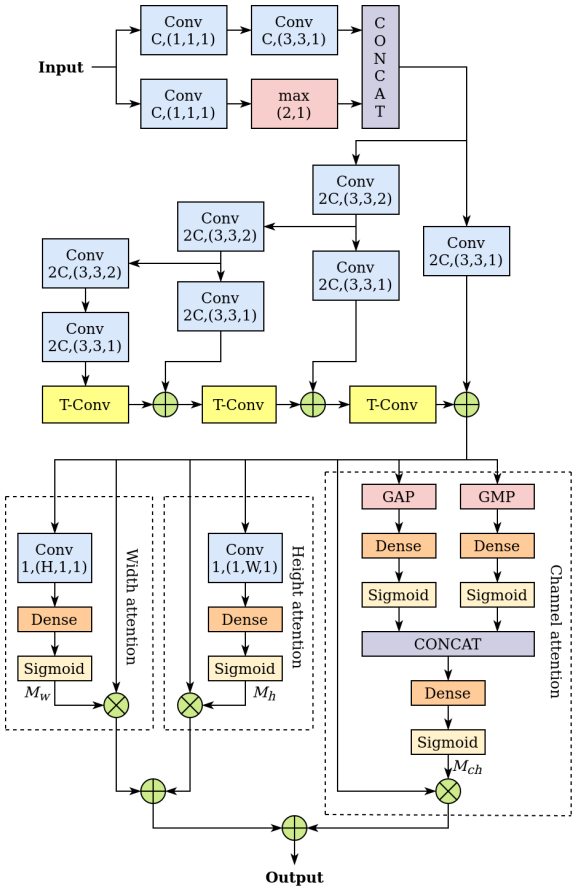

Figure 2. The proposed Attentive Style Embedding module. Here, input and output are *Conv* layers with C and $2C$ number of channels, respectively. The $Conv\ F,(Y,Z,S)$ represents a *convolution* layer with $kernel\ size$ of (Y,Z), $stride$ as S and $filter_{num}$ as F. $Max(A,B)$ represents *max pooling* layer with $pool\ size$ of (A,A) and $stride$ as B.

Concurrent axial attention: The proposed concurrent axial attention sub-network focuses on axial information by proposing a series of solutions. Firstly, this sub-network adaptively investigates height, width and channel axes to reduce the effect of redundant regions. Secondly, we incorporate the use of *Conv* layers to extract global height and width embedding vector. Thirdly, to calibrate channel weights, unlike others, we consider both the global average pooling and global max pooling operation [4, 10].

The *adaptive channel embedding* is created to separately investigate the spatial planes with relevant channels. We consider the input to be $X(H, W, D)$, where H, W and D represent the height, the width and the number of channels. The global information is embedded for every channel using global average pooling and global max pooling. The dense layers with sigmoid activation creates the attention mask of embedding vectors with complex variations, as given in Eq. 1a and Eq. 1b. To obtain the global representative for each channel, we use another *dense* layer followed by a sigmoid activation for the generation of aggregated channel mask M_{ch}, as in Eq. 1c. The discriminatory regions of the input feature are highlighted by multiplying this mask.

$$x_1 = \sigma(Dense(Dense(GMP(X), D/4), D)) \quad (1a)$$
$$x_2 = \sigma(Dense(Dense(GAP(X), D/4), D)) \quad (1b)$$
$$M_{ch} = \sigma(Dense(Concat(x_1, x_2), D)) \quad (1c)$$

In the case of *adaptive height and width embedding*, the spatial dimension covers the local information across the height and width axis. The state-of-the-art methods either encode these features jointly [13, 15] or use axial attention [14], which uses position-sensitive self-attention in both the axes. Contrary to this, we hypothesize that the position-sensitive global descriptors can be obtained using convolution operation for height and width squeezing. Considering input to be X, Eq. 2a, Eq. 2b specify how global height(M_h) and width(M_w) embeddings are generated. The outputs of these attention modules are generated by multiplying the masks with the input. The sigmoid activation has been chosen for mask creation following [10, 15].

$$M_h = \sigma(Dense(Dense(Conv(X), H/4), H)) \quad (2a)$$
$$M_w = \sigma(Dense(Dense(Conv(X), W/4), W)) \quad (2b)$$

3.2. The DAtRNet architecture

The description of the proposed overall DAtRNet architecture is given in Table 1. In DAtRNet, we choose an embedding dimension to generate attribute-specific feature embedding in style space. For our application, we have considered this value to be 128. In the dense layer, after global average pooling, we have used a dropout of 0.25 to prevent the network from overfitting.

3.3. Attribute-aware Substitute Recommendation

In DAtRNet, the attribute features are embedded within its super-category specific style features that can be independently extracted by creating triplets. The triplet attributes should be chosen to ascertain that the anchor and positive images have the same attribute features from the same category, although they might differ for other categories. It should be noted that our intention is not to find image similarity, instead to find attribute similarity. Following

Table 1. The DAtRNet architecture.

Input size	Output size	Layer name
(256,256,3)	(256,256,32)	Conv (32, (3,3), 2)
(256,256,32)	(128,128,32)	Max Pool (2,2)
(128,128,32)	(128,128,64)	ASE Module-1
(128,128,64)	(64,64,64)	Max Pool (2,2)
(64,64,64)	(64,64,128)	ASE Module-2
(64,64,128)	(32,32,128)	Max Pool (2,2)
(32,32,128)	(32,32,256)	ASE Module-3
(32,32,256)	(256)	Global Avg. Pool.
(256)	(256)	Dense(256)
(256)	(128)	Dense(128)

Figure 1, we obtain the 128-dimensional disentangled attribute embedding for six categories from the proposed network and concatenate them to get the attribute-aware representation of the product image. For the training of our network, we have used triplet loss with hard mining. For training, we have used the image dimension of 256, kernel size of 3×3 and the number of filters in first *Conv* layer as 32. These values have been chosen from extensive experiments of hyper-parameter tuning, given in Section 4 of the supplementary material.

The query images and the attribute manipulation instructions are considered to obtain the closest k neighbors from the retrieval gallery. It is achieved by calculating the L_2 distance of the attribute-aware embedding of query images from the feature representation of images from the retrieval gallery. To incorporate the attribute manipulation instructions in feature embedding, we have replaced the 128-dimensional vector of the undesired attributes with a generic representation of the desired attribute. The generic representation for each attribute is obtained by passing all images with desired attributes to the corresponding model and finding the average response. The averaging operation normalizes the response and prevents the performance from worsening due to the class imbalance problem. A detailed explanation of attribute-aware substitute fashion search can be found in Section 3 of the supplementary section. We also performed experiments using a product- or category-specific generic attributes and documented the results in Section 7 of the supplementary material.

4. Results and analysis

We evaluate the performance of DAtRNet on three use-cases: 1) substitute fashion search with query images, which retrieves products with the same set of attributes; 2) substitute fashion search with one attribute manipulation from the query product; and 3) substitute product search with multiple attribute manipulations.

To evaluate the performance, we conduct a series of experiments on DeepFashion [12] and Shopping100k [3] datasets. The DeepFashion dataset provides product images with people consisting of 26 categories, including six attribute super-categories, namely texture, sleeve, length,

Table 2. Comparison of category-specific Top-30 retrieval accuracy of the proposed DAtRNet for Shopping100k and DeepFashion datasets with the state-of-the-art.

	Shopping100k						DeepFashion		
	Color	**Collar**	**Fastening**	**Neckline**	**Pattern**	**Sleeve**	**Category**	**Shape**	**Texture**
Attribute-based [11]	0.175	0.195	0.181	0.137	0.299	0.101	0.118	0.138	0.115
AMNet [16]	0.433	0.477	0.248	0.350	0.388	0.360	0.218	0.249	0.273
FashionSearchNet w/o Loc [1]	0.583	0.599	0.336	0.494	0.552	0.524	0.202	0.409	0.330
FashionSearchNet [1]	0.649	0.642	0.423	0.532	0.575	0.640	0.380	0.409	0.338
DAtRNet (Ours)	**0.738**	**0.626**	**0.503**	**0.641**	**0.674**	**0.761**	**0.667**	**0.652**	**0.636**

Table 3. Comparison of Top-30 retrieval accuracy w.r.t the state-of-the-art methods using Shopping100k and DeepFashion datasets.

Methods	**Shopping100k**		**DeepFashion**	
	Query	Query+att.	Query	Query+att.
Attribute-based: AlexNet [11]	0.593	0.216	0.464	0.124
Attribute-based: ResNet-50 [7]	0.601	0.245	0.503	0.232
Attribute-based: ViT [5]	0.582	0.186	0.401	0.151
AMNet: AlexNet backbone [16]	0.637	0.429	0.483	0.229
AMNet: ResNet backbone [16]	0.658	0.434	0.496	0.309
AMNet: ViT backbone [16]	0.593	0.385	0.438	0.318
FashionSearchNet w/o Loc [3]	0.611	0.512	0.448	0.305
FashionSearchNet [3]	0.651	0.572	0.469	0.381
ADDE-M [9]	0.682	0.598	0.538	0.315
DAtRNet (Ours)	**0.752**	**0.677**	**0.731**	**0.662**

neckline, category, and shape. For the Shopping100k dataset, we have used six attribute categories, namely color, collar, fastening, neckline, pattern, and sleeve spanning 87 fine-grained attributes. For training the Siamese networks, we generate 90,000 triplets for each super-category for both datasets. In Table 2, the attribute-specific top-30 retrieval accuracy for various architectures is compared with DAtRNet. From these results, we observe that the proposed method results in better recognition performance by outperforming super-category specific attributes for all cases, except for collar attribute in Shopping100k (0.626 compared to 0.642 given by FashionSearchNet). The variation of the category-specific results for k is given in Sections 6 and 8 of the supplementary material. In Table 3, we compare the overall performance of the proposed DAtRNet with existing methodologies [1–3, 16] for both the search strategies. Our model consistently outperforms the existing architectures with a significant difference for both search strategies on both datasets. We validate these results by visual examples in Figure 3 considering query image and one attribute manipulation instruction. The t-SNE visualization of the attribute embeddings is illustrated in Section 10 of the supplementary material. Also, we demonstrate the variation in performance *w.r.t.* the number of manipulated attributes in Section 9 of the supplementary material.

Ablation experiments are conducted to understand the impact of the sub-networks in DAtRNet. Table 4 provides the results of these experiments performed using 20,000 triplets of two categories of the DeepFashion dataset. For a fair comparison, we re-trained DAtRNet with same setup,

Table 4. Comparison of top-30 retrieval accuracy for ablation study of DAtRNet using DeepFashion dataset. Here, MSF: Multi-scale feature extraction sub-network.

Ablation Experiments	**Only query**	**Query + Attribute**
Only three MSF modules	0.768	0.525
MSF + channel attention	0.836	0.789
MSF + spatial attention	0.849	0.776
MSF + height	0.838	0.768
MSF + width	0.841	0.751
One ASE module	0.624	0.457
ResNet replacing MSF	0.861	0.803
DAtRNet (Ours)	**0.894**	**0.846**

Figure 3. Visual examples of fashion search by query image and desired attribute instruction. The first and second columns contain the query image and the desired attribute information respectively. The next four columns show the attribute-aware substitute recommendation.

resulting in better performance than Table 3. From Table 4, we observe that DAtRNet outperforms other network configurations. More details including class activation maps are given in Section 5 of the supplementary material.

5. Conclusions

A novel architecture for style embedding, DAtRNet, is proposed that consists of two sub-networks: multi-scale fine grained feature extraction and concurrent axial attention for obtaining discriminatory cues. The DAtRNet successfully consolidates key features for improved attribute representation and can be trained end-to-end without augmenting with other complex sub-networks. Extensive experimental results on two large publicly available datasets show that the proposed network outperforms the state-of-the-art methods for all the search strategies.

References

[1] Kenan E Ak, Ashraf A Kassim, Joo Hwee Lim, and Jo Yew Tham. Fashionsearchnet: Fashion search with attribute manipulation. In *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*, pages 45–53, 2018. 1, 2, 4

[2] Kenan E Ak, Ashraf A Kassim, Joo Hwee Lim, and Jo Yew Tham. Learning attribute representations with localization for flexible fashion search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7708–7717, 2018. 1, 2, 4

[3] Kenan E Ak, Joo Hwee Lim, Jo Yew Tham, and Ashraf A Kassim. Efficient multi-attribute similarity learning towards attribute-based fashion search. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1671–1679, 2018. 1, 2, 3, 4

[4] Gaurab Bhattacharya, Nikhil Kilari, Jayavardhana Gubbi, V Bagya Lakshmi, and P Balamuralidhar. F-attnet: Towards multi-scale feature fusion for fashion attribute prediction. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021. 1, 2

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[6] Sheng Guo, Weilin Huang, Xiao Zhang, Prasanna Srikhanta, Yin Cui, Yuan Li, Hartwig Adam, Matthew R Scott, and Serge Belongie. The imaterialist fashion attribute dataset. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 3113–3116, 2019. 1

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 4

[8] Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. *arXiv preprint arXiv:1912.12180*, 2019. 2

[9] Yuxin Hou, Eleonora Vig, Michael Donoser, and Loris Bazzani. Learning attribute-driven disentangled representations for interactive fashion retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12147–12157, 2021. 4

[10] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018. 2, 3

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems 25 (NIPS)*, pages 1097–1105, 2012. 4

[12] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1096–1104, 2016. 1, 2, 3

[13] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Bam: Bottleneck attention module. *arXiv preprint arXiv:1807.06514*, 2018. 2, 3

[14] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 108–126, 2020. 2, 3

[15] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018. 2, 3

[16] Bo Zhao, Jiashi Feng, Xiao Wu, and Shuicheng Yan. Memory-augmented attribute manipulation networks for interactive fashion search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1520–1528, 2017. 1, 2, 4