DIVERSITY IN FASHION RECOMMENDATION USING SEMANTIC PARSING

Sagar Verma[1] *Sukhad Anand*[1] *Chetan Arora*[1] *Atul Rai*[2]

[1]IIIT Delhi [2]Staqu Technologies

ABSTRACT

Developing recommendation system for fashion images is challenging due to the inherent ambiguity associated with what criterion a user is looking at. Suggesting multiple images where each output image is similar to the query image on the basis of a different feature or part is one way to mitigate the problem. Existing works for fashion recommendation have used Siamese or Triplet network to learn features between a similar pair and a similar-dissimilar triplet respectively. However, these methods do not provide basic information such as, how two clothing images are similar, or which parts present in the two images make them similar. In this paper, we propose to recommend images by explicitly learning and exploiting part based similarity. We propose a novel approach of learning discriminative features from weakly-supervised data by using visual attention over the parts and a texture encoding network. We show that the learned features surpass the state-of-the-art in retrieval task on DeepFashion dataset. We then use the proposed model to recommend fashion images having an explicit variation with respect to similarity of any of the parts.

Index Terms— Fashion Recommendation, Semantic Parsing of Clothing Items

1. INTRODUCTION

The recent advancements in computer vision and machine learning techniques have made it possible to match images and objects captured under different pose and environmental conditions. One popular application of such a system is in online shopping for clothing items, where researchers have used the technique for product retrieval, fashion recommendation, style recognition, and clothing item parsing, etc. The image-based recommendation is a crucial part of any fashion e-commerce website which primarily undertakes two kinds of retrieval tasks; in-shop and consumer-to-shop retrieval. In-shop retrieval task requires searching for similar product images, all captured in a shop environment, whereas consumer-to-shop retrieval task requires matching shop images that are similar to user-shot query images. Understandably, the later problem is much harder due to uncontrolled lighting conditions, body pose and object occlusion in the user-shot images.

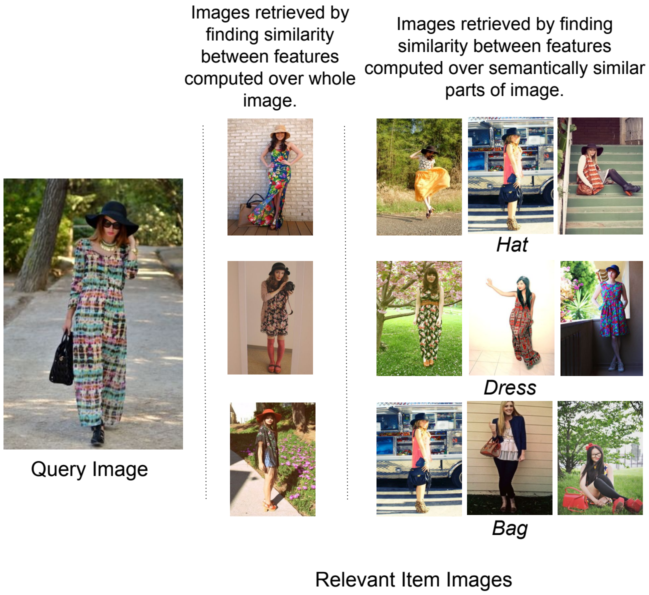

Images retrieved by finding similarity between features computed over whole image.
Images retrieved by finding similarity between features computed over semantically similar parts of image.
Query Image
Hat
Dress
Bag
Relevant Item Images

Fig. 1: Finding similarity based on the visual features computed over the whole image gives results that look similar because of some large clothing item present in the images as shown on the middle column. Proposed work recognizes the distinct and fine clothing items present in the query image and then finds similar images to those items category wise as show on right column.

Existing methods for fashion recommendation focus on finding similarity between two images using Siamese network [1, 2, 3]. In an extension of the work, it is also possible to learn similarity-dissimilarity between three images using a Triplet network [4, 5]. Both Siamese, and Triplet Networks typically focus on the whole image to learn visual features and usually capture only the dominant visual cue. This becomes problematic, especially in the case of consumer-to-shop fashion recommendation tasks, since the pose of the person in query image is uncontrolled. This makes it difficult to infer, which part or clothing item the user is interested in. The problem becomes even more challenging, if the desired item is visually smaller, making it harder for convolutional neural networks, a technique of choice in many such systems, to focus on such non-dominant cues.

To handle inherent ambiguity in user's intentions, in this paper, we propose a novel fashion recommendation system, which implicitly hedges its suggestions on the basis of similarity with various parts in the query image. We propose to use attention module in deep neural networks to focus on distinct parts of an image and retrieve similar products for each part based on the texture similarity. We treat the problem as multi-label classification and use a single convolutional stream to learn visual features and an attentional LSTM to focus on different locations. Our convolutional stream contains a texture encoding layer in the end to learn texture based similarity from each attention region.

The specific contributions of this work are as follows :

- Given the ambiguous nature of user intent, we propose a new recommendation system to produce diverse recommendations on the basis of similarity of different parts in the query image.
- To generate semantically meaningful parts in the fashion image, we propose to use attention based deep neural network which learns to attend to different parts using the weakly labeled data available in the benchmark dataset.
- Instead of features from standard pre-trained neural networks, we suggest using texture-based features which, as we show in our experiments, are better suited for finding clothing similarity.
- Apart from diversity in fashion recommendation, our experiments and evaluations on multiple datasets demonstrate the superiority of the proposed model in attribute classification and cross-scene image retrieval tasks.

2. RELATED WORK

In the past few years there has been a tremendous increase in the research on fashion and related domains [6, 7] such as cloth parsing [8, 9], and clothing attribute recognition [10, 11]. Detecting fashion style [2] and more subjective attribute such as "how fashionable a person looks" have also been looked at [12]. A task closely related to our goal is cross-domain image retrieval [13, 1], where most of the current techniques have focussed on the whole image for computing similarity score. Wang et. al. [14] used attention for the task, whereas Chen et. al. [15] have shown that the visual attention architecture of image captioning task cannot be used for multi-label classification as no sequence information is present. Wang et. al. [16] have proposed a similar architecture, as used in this paper, using spatial transformer network and hard attention.

In this paper, we propose to learn semantically important regions and then generate diverse recommendations on the basis of similarity of any of these. We note that most of the state-of-the-art learning systems for image matching, are geared towards global shape matching. The networks learn to be invariant against the fine variations and distortions. However, in case of fashion and clothing, the fine-grained differences in style and texture are the most important cues. Therefore we propose to match the recovered semantic regions on the basis of texture rather than standard features obtained from ResNet [17] or VGGNet [18]. Texture has been a well-studied problem in computer vision where both traditional hand tuned [19, 20] as well as deep learned features [21, 22] have been proposed. We use the texture encoding layer proposed by Zhang et. al. [23] in our architecture. The encoding layer generalizes robust residual encoders such as VLAD [24] and Fisher Vectors [25] and can be trained end to end, along with the rest of the model.

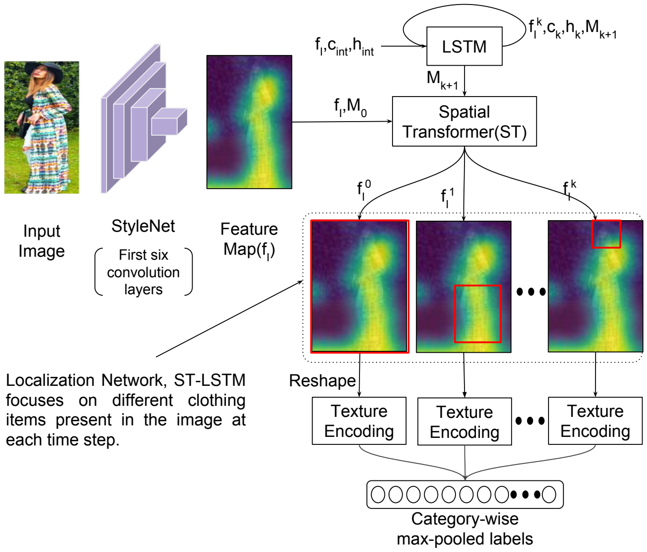

f_l^k, c_k, h_k, M_{k+1}
f_l, c_{int}, h_{int}
LSTM
M_{k+1}
f_l, M_0
Spatial Transformer(ST)
Input Image
StyleNet
First six convolution layers
Feature Map(f_l)
f_l^0
f_l^1
f_l^k
Localization Network, ST-LSTM focuses on different clothing items present in the image at each time step.
Reshape
Texture Encoding
Texture Encoding
Texture Encoding
Category-wise max-pooled labels

Fig. 2: Proposed architecture

3. ARCHITECTURE

Our proposed architecture consists, a six-layered convolutional neural network to extract global image features, a visual attention module to generate attention over an image, and a texture encoding layer to capture clothing texture from the generated attentions. Since we would like the proposed model to capture similarity of various semantic regions in the image, we first train the model to explicitly learn these important regions. We do so by using an attention module, which learns to focus on and predict different objects present in the image. Owing to the space restriction, we briefly describe below, important aspects of the proposed architecture. More details can be found at the project page: `www.iiitd.edu.in/~chetan/projects/fashion/`.

CNN for Global Image Features: We use the first three convolutional and max-pooling layers of the model used in [5] to extract global image features. The model has been trained specifically for clothing and fashion images and worked best in our experiments. We note that other similar architectures could have been used as well.

Visual Attention Module: For attention we use a recurrent memorized-attention module, which combines the recurrent computation process of an LSTM network, a spatial transformer[26] and a texture encoding layer [23] described in the next section. It iteratively searches different parts and predicts the scores of label distribution for them and also the corresponding feature vectors.

Texture Encoding Layer: The features extracted at each time step of the LSTM network are passed through a texture encoding layer which learns the texture of the attentional region. The encoding layer incorporates dictionary learning, feature pooling and classifier learning into it. The layer assumes k codewords and assigns descriptors to them using soft assignment so that the function is differentiable and can be trained along with the model using Stochastic gradient descent with backpropagation as proposed by Zhang et. al.[23].

Training: Given a dataset with N training sample images and C class labels, during training each image is passed through the complete network. Multiple labels for an input image is represented as one-hot-vector with 1 corresponding to all the labels present. A CNN generates global spatial features which are fed into the attention module which further generates localized texture features of the attentional region and corresponding labels. The LSTM network is unrolled for k time steps and hence produce k labels by category wise max pooling.

Loss: We use Euclidean distance as the objective function for classification loss as given in [16]. In order to force localization network to look at a different location at each time step, we use divergence loss given in [27]. Divergence loss is computed by finding the correlation between two subsequent attention maps. We also use localization loss given in [16] which removes redundant locations and forces localization network to look at small clothing parts. All three loss values are added together with multiplicative factors of 1 for classification loss and localization loss and 0.01 for divergence loss for all of our experiments.

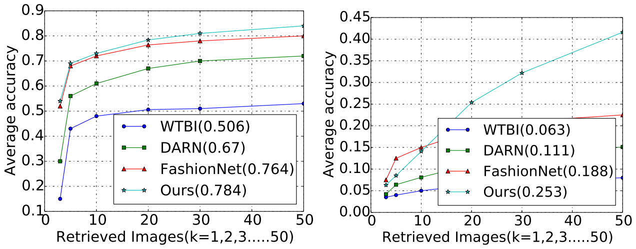
4. DATASETS

We have conducted our experiments on Fashion144K [5], Fashion550k [28], and DeepFashion dataset [30].

The cleaned version of Fashion144K dataset [5] contains $90,000$ images. It has multi-label annotations available. There is no similar pair annotation available in the dataset. Along with multiple labels, each image has a fashionability score. The total number of labels are 128 comprising of parts and color tags. Resolution of images is 384×256. We use this dataset for discriminative feature learning due to the availability of multiple labels. Fashion550k [28] has similar properties as that of its predecessor, Fashion144k [5]. Size is four times that of Fashion144k. The total number of labels are 66.

Dataset	Fashion144k [5]		Fashion550k [28]	
Model	AP_{all}	mAP	AP_{all}	mAP
StyleNet [5]	65.6	48.34	69.53	53.24
Baseline [28]	62.23	49.66	69.18	58.68
Viet et al. [29]	NA	NA	78.92	63.08
Inoue et al. [28]	NA	NA	79.87	**64.62**
Ours	**82.78**	**68.38**	**82.81**	57.93

Table 1: Multi-label classification on Fashion144k [5] and Fashion550k [28]



(a) In-Shop retrieval (b) Consumer-to-shop retrieval

Fig. 3: Retrieval results for In-shop and Consumer-to-shop retrieval tasks on DeepFashion dataset [30]. Proposed method works better in both the tasks.

DeepFashion dataset [30] contains $800,000$ images. Similar pair information along with bounding box is available for both consumer-to-shop and in-shop retrieval tasks. For consumer-to-shop retrieval task a total of $33,841$ clothing items are available in $239,557$ consumer/shop image pairs. For in-shop retrieval task $7,982$ clothing item are available in $52,712$ in-shop images.

5. EXPERIMENTS AND RESULTS

All the experiments have been conducted on a workstation with 1.2 GHz CPU, 32 GB RAM, NVIDIA P5000 GPU and running Ubuntu 14.04. We use pytorch for the network implementation. We train our model on Fashion144k dataset [5], with 59 item labels, we exclude the color labels for training.

Though not the focus of the work, to confirm the hypothesis that our network is able to detect different clothing regions, we first evaluate the proposed architecture on item recognition task on Fashion144k [5] and Fashion550k dataset [28], In these experiments, we use Adam optimizer with a batch size of 64, momentum of 0.9 and learning rate of 0.00001 and trained the models for 40 epochs. 32 codewords are used

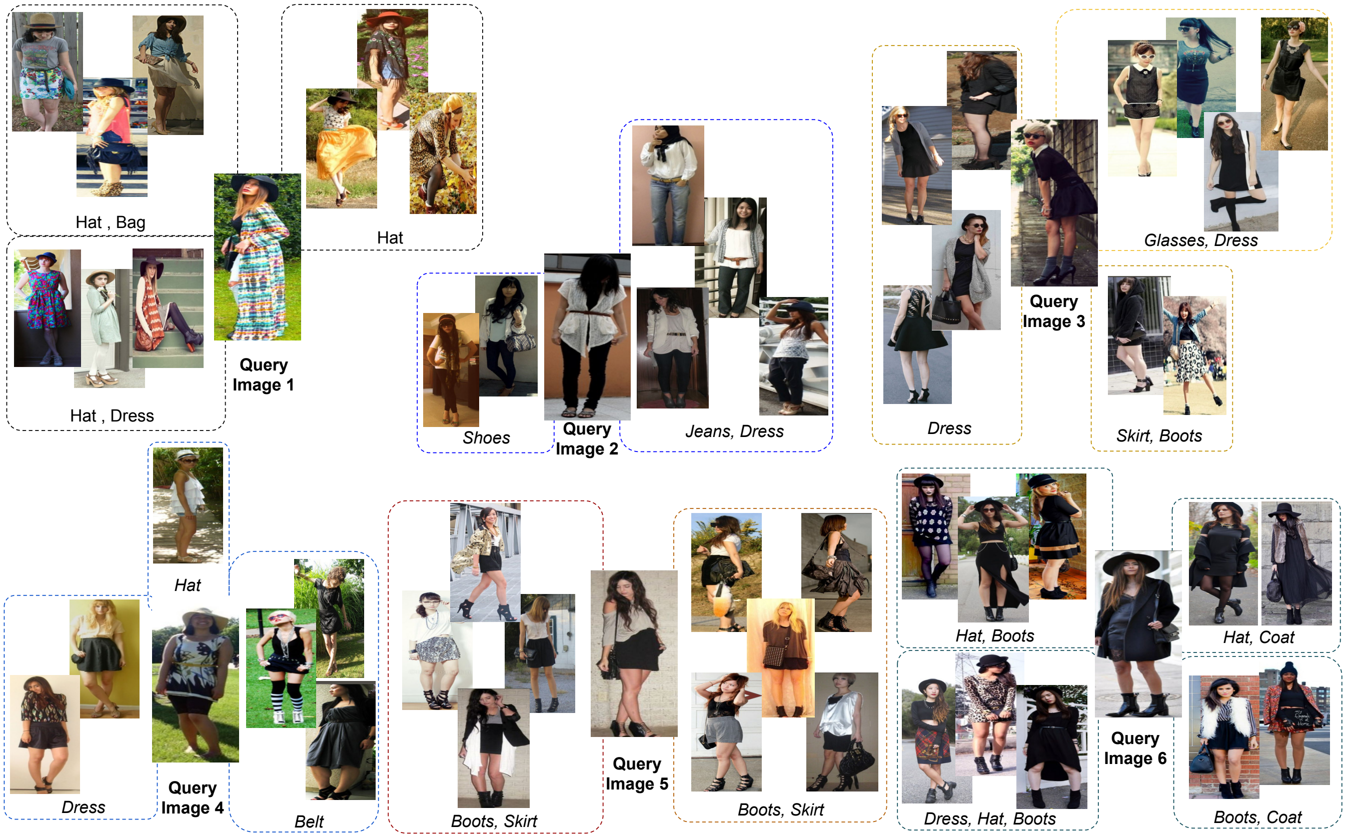


Fig. 4: Semantically similar results for some of the query images from Fashion144k dataset [5] using our method. Retrieved images for six query images are shown from left to right, top to bottom. For each query image, retrieved images having similar clothing items are shown together.

in texture encoding layer. For each image, we assign top 6 highest-ranked labels to the image and compare with the ground-truth labels. As evaluation metrics for the item recognition task, the class-agnostic average precision (AP_{all}), and the mean of each class-average precision (mAP) are used. Table 1 shows the comparison.

We also tested our model for item retrieval tasks, for in-Shop as well as consumer-to-shop retrieval tasks on DeepFashion dataset [30]. For this task, we extract texture features from the model trained on Fashion144k dataset [5] and use Euclidean distance to find similarity between two image features. Fig. 3 show the top-k retrieval accuracy of all the compared methods with k ranging from 1 to 50. We also list the top-20 retrieval accuracy after the name of each method. For in-shop retrieval, our model achieves best performance(0.784) which is better than state-of-the-art FashionNet [30] accuracy (0.764). For consumer-to-shop, the corresponding numbers are 0.253 for our model and 0.188 for FashionNet [30].

Finally, we test our model for similar item retrieval task. Fig. 4 shows the retrieval results for various query images. It is easy to see that recommended results have lots of variation in terms of similarity of various parts.

6. CONCLUSION

We have presented a novel attention based deep neural network which learns to attend to different semantic parts using the weakly labeled data. Unlike state of the art techniques using global similarity features, we suggest to extract per part fine-grained features using texture cues: we use DeepTEN in our experiments. We have conducted experiments on three widely accepted tasks in clothing recognition and retrieval. We show that the proposed model outperforms previous state-of-the-art models on deep fashion dataset for both consumer-to-shop and in-shop image retrieval tasks. Our experiment on recommendation task shows that the proposed model is able to successfully recommend a wide variety of objects matching with different parts of the query image.

7. REFERENCES

[1] Junshi Huang, Rogerio Feris, Qiang Chen, and Shuicheng Yan, "Cross-domain image retrieval with a

dual attribute-aware ranking network," in *ICCV*, 2015, pp. 1062–1070.

[2] Andreas Veit et al., "Learning visual clothing style with heterogeneous dyadic co-occurrences," in *ICCV*, 2015, pp. 4642–4650.

[3] Xi Wang, Zhenfeng Sun, Wenqiang Zhang, Yu Zhou, and Yu-Gang Jiang, "Matching user photos to online products with robust deep features," in *ICMR*, 2016, pp. 7–14.

[4] Jiang Wang et al., "Learning fine-grained image similarity with deep ranking," in *CVPR*, 2014.

[5] E. Simo-Serra and H. Ishikawa, "Fashion style in 128 floats: Joint ranking and classification using weak data for feature extraction," in *CVPR*, 2016, pp. 298–307.

[6] Edgar Simo-Serra, Sanja Fidler, Francesc Moreno-Noguer, and Raquel Urtasun, "A high performance crf model for clothes parsing," in *ACCV*, 2014, pp. 64–81.

[7] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Parsing clothing in fashion photographs," in *CVPR*, 2012, pp. 3570–3577.

[8] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Retrieving similar styles to parse clothing," in *TPAMI*, 2015, vol. 37, pp. 1028–1040.

[9] Wei Yang, Ping Luo, and Liang Lin, "Clothing co-parsing by joint image segmentation and labeling," in *CVPR*, 2014, pp. 3182–3189.

[10] Huizhong Chen, Andrew Gallagher, and Bernd Girod, "Describing clothing by semantic attributes," in *ECCV*, 2012, pp. 609–623.

[11] Kota Yamaguchi et al., "Mix and match: Joint model for clothing and attribute recognition," in *BMVC*, 2015, vol. 1, p. 4.

[12] Edgar Simo-Serra, Sanja Fidler, Francesc Moreno-Noguer, and Raquel Urtasun, "Neuroaesthetics in fashion: Modeling the perception of fashionability," in *CVPR*, 2015, vol. 2, p. 6.

[13] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan, "Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set," in *CVPR*, 2012, pp. 3330–3337.

[14] Zhonghao Wang, Yujun Gu, Ya Zhang, Jun Zhou, and Xiao Gu, "Clothing retrieval with visual attention model," in *VCIP*, 2017.

[15] Shang-Fu Chen et al., "Order-free RNN with visual attention for multi-label classification," in *AAAI*, 2018.

[16] Zhouxia Wang, Tianshui Chen, Guanbin Li, Ruijia Xu, and Liang Lin, "Multi-label image recognition by recurrently discovering attentional regions," in *ICCV*, 2017.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[18] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *CoRR*, 2014, vol. abs/1409.1556.

[19] K. S. Loke, "Automatic recognition of clothes pattern and motifs empowering online fashion shopping," in *ICCE-TW*, 2017, pp. 375–376.

[20] Si Liu et al., "Hi, magic closet, tell me what to wear!," in *Proceedings of the 20th ACM international conference on Multimedia*, 2012, pp. 619–628.

[21] Mircea Cimpoi, Subhransu Maji, and Andrea Vedaldi, "Deep filter banks for texture recognition and segmentation," in *CVPR*, 2015, pp. 3828–3836.

[22] Vincent Andrearczyk and Paul F Whelan, "Using filter banks in convolutional neural networks for texture classification," in *Pattern Recognition Letters*, 2016, vol. 84, pp. 63–69.

[23] Hang Zhang, Jia Xue, and Kristin Dana, "Deep TEN: Texture encoding network," in *CVPR*, 2017.

[24] H. Jgou, M. Douze, C. Schmid, and P. Prez, "Aggregating local descriptors into a compact image representation," in *CVPR*, 2010, pp. 3304–3311.

[25] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *CVPR*, 2007, pp. 1–8.

[26] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu, "Spatial transformer networks," in *NIPS*, 2015, pp. 2017–2025.

[27] B. Zhao, X. Wu, J. Feng, Q. Peng, and S. Yan, "Diversified visual attention networks for fine-grained object classification," in *IEEE Transactions on Multimedia*, 2017.

[28] Naoto Inoue, Edgar Simo-Serra, Toshihiko Yamasaki, and Hiroshi Ishikawa, "Multi-label fashion image classification with minimal human supervision," in *ICCVW*, 2017, pp. 2261–2267.

[29] Andreas Veit et al., "Learning from noisy large-scale datasets with minimal supervision," in *CVPR*, 2017.

[30] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *CVPR*, 2016, pp. 1096–1104.