A Landmark-free Attention Network for Fashion Clothes Classification and Attribute Prediction

Shajini Majuran and Amirthalingam Ramanan
Department of Computer Science, Faculty of Science, University of Jaffna
shayu.kiri@gmail.com, a.ramanan@univ.jfn.ac.lk

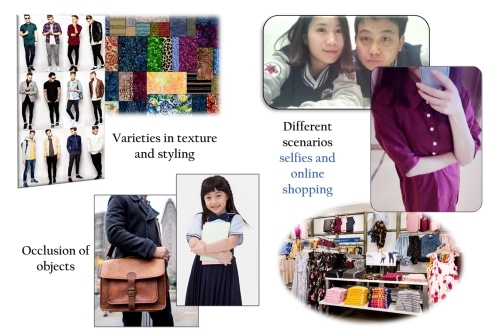
1. Introduction

In recent years, with the rapid growth of e-commerce and fashion-based applications, the visual fashion clothing analysis has gained more attention. An evolving technique for clothing-related tasks is the deep learning method, which uses deep convolutional neural networks (CNN) to simultaneously learn feature illustration and classify clothing images. Basically, fashion analysis work on there core problems: landmark localization, category classification and attribute prediction. However, the existing landmark-driven models take up a lot of manpower for landmark annotation and also suffers from inter- and intra-individual variability. Besides, CNN has a semantic gap between the lower and upper level features from different levels of feature representation of a network. From these inspirations, we construct an attentive network without extracting knowledge from clothing landmarks (e.g. collar, waist, sleeve, and hemline points).

Professional etiquette

Cultural etiquette

Applications



Challenges

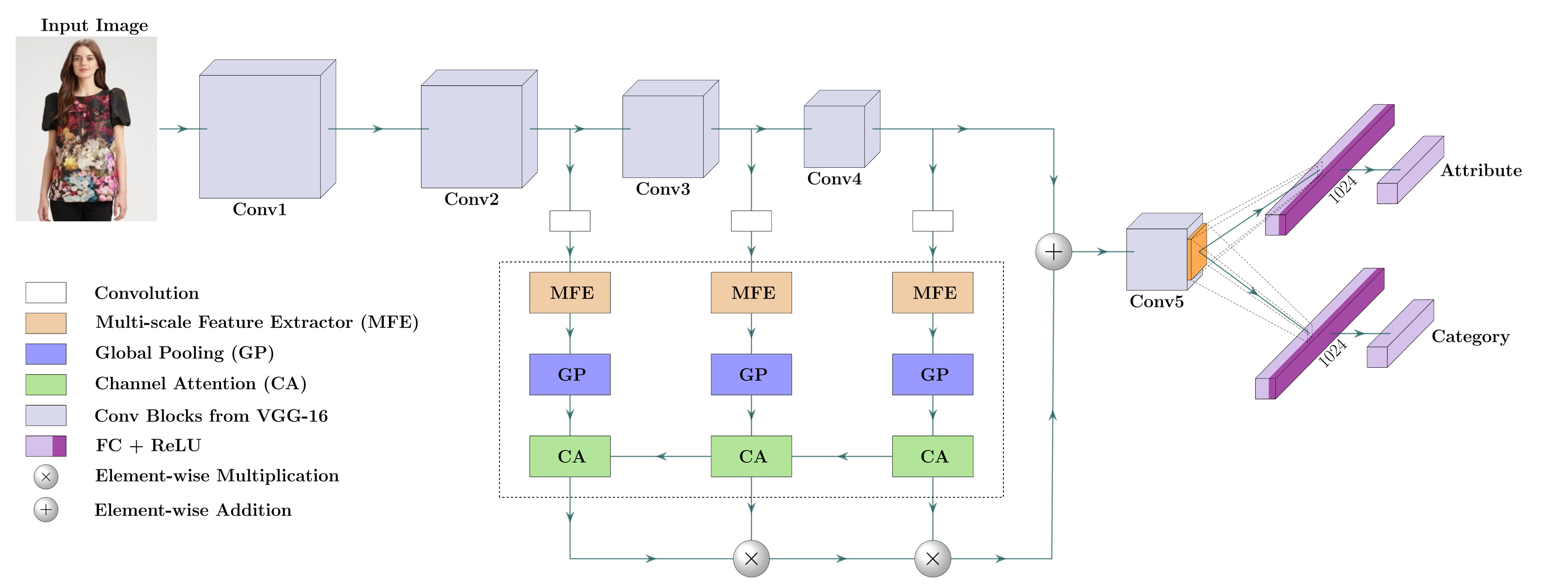
2. Problem Specification

The landmark annotation is very expensive and time consuming. It also suffers from inter- and intra-individual variability on clothes.

3. Contribution

1. Formalised fashion analysis into a multitask deep neural network for fashion clothing analysis.
2. Proposed a landmark-free multi-staged attention network to experiment the multitask learning network for fashion clothing classification and its corresponding attribute prediction.

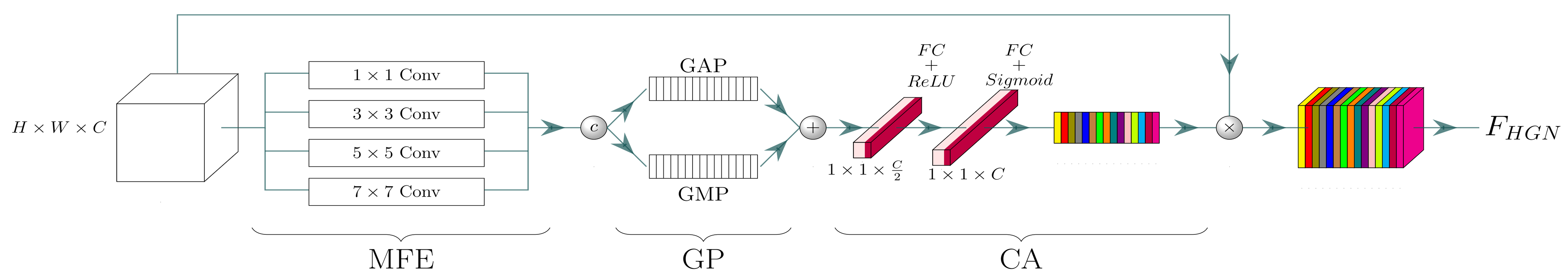
4. Landmark-free Attention Network



- Conv1: $224 \times 224 \times 64$
- Conv2: $112 \times 112 \times 128$
- Conv3: $56 \times 56 \times 256$
- Conv4: $28 \times 28 \times 512$
- Conv5: $14 \times 14 \times 512$
- Learning rate: 0.0001
- Batch size: 16
- Optimiser: Adam

The multiscale features which are extracted from multilevel of deep neural network architecture help the network to locate more characteristic features and also the high-level features are able to guide in rectifying low-level features. Generally, the top-level features of CNN contain rich semantic information while the initial layers possess rich spatial information which refers to the simple understanding of clothing images through networks. Thus, the multilevel convolutional features are combined to form a new attention, which gives our network a flexible way to focus on the most important features of clothes for category and attribute prediction. Further, the model increases the adaptive attention by incorporating global attention which exploit the need for contextual and semantic relationship among features.

5. Hierarchical Global Attention (HGA) Block



HGA block is an attention of the proposed network consisting of three major components: (i) Multiscale feature extractor (MFE), (ii) Global pooling (GP), and (iii) Channel attention (CA). This creates a combination of multilevel features with the added benefit of effectively representing features.

6. Dataset

DeepFashion-C Dataset [2]

- 289,222 well-posed shop images including consumer photos
- (training/val/testing)=(209222/40000/40000) images
- 50 categories of cloth items (top, blouse, shorts, skirt, etc.)
- 5 attribute types (shape, style, texture, fabric, and part)
- 1000 descriptive attributes (abstract, dot, animal print, etc.)
- 8 landmarks (left/right of sleeve, collar, waist, and hem)

7. Experimental Setup

The proposed architecture is built based on *VGG-16* pretrained network. Further, we use standard cross-entropy loss ($Loss_c$) and weighted cross-entropy loss ($Loss_a$) to train category classification and attribute estimation, respectively. To facilitate multitask learning throughout model, we utilise a weighted loss combination to calculate the total loss which is used to optimise the model weights. The *combined loss* is as follows,

$$L = w_c \times Loss_c + w_a \times Loss_a \tag{1}$$

where w_c and w_a are the weights for the losses computed from category and attribute branches, respectively. The weighted cross-entropy loss to predict attributes incorporates the weighting factors by the ratio of the numbers of positive and negative samples in the training set as illustrated in [2].

8. Testing Results

Table I: Quantitative results for category classification and attribute prediction on DeepFashion-C dataset using top-k accuracies and top-k recall, respectively. The notations, LM stands for landmark-free models and AM stands for the model which has attention mechanism

Methods	LM	AM	Category		Attribute	
			top-3	top-5	top-3	top-5
Huang *et al.* [1]	✓	×	59.48	79.58	42.35	51.95
Liu *et al.* [2]	×	✓	82.58	90.17	45.52	54.61
Corbiere *et al.* [3]	✓	×	86.30	92.80	23.10	30.40
Wang *et al.* [4]	×	✓	90.99	95.78	51.53	60.95
Ye *et al.* [5]	✓	✓	90.06	95.04	52.82	62.49
Lee *et al.* [6]	✓	✓	91.37	95.26	47.70	57.28
Cho *et al.* [7]	✓	×	91.24	95.68	-	-
Zhang *et al.* [8]	×	✓	91.99	96.44	50.58	60.43
Shajini *et al.* [9]	×	✓	91.06	96.35	51.22	61.63
Ours	✓	✓	**92.78**	**96.85**	**55.26**	**64.02**

Table II: Quantitative results for top five categories and all attribute types using top-k accuracies and top-k recall, respectively

Top 5 categories			Attribute types		
Categories	top-3	top-5	Attributes	top-3	top-5
Dress	97.91	98.98	Texture	60.18	70.76
Tee	97.54	98.61	Fabric	43.49	54.91
Blouse	97.68	98.75	Shape	62.35	71.21
Shorts	96.98	98.05	Part	50.15	61.76
Skirt	96.93	98.01	Style	68.13	73.27

9. Discussion and Conclusion

The multilevel attention network recalibrates the feature extraction by focusing on spatial and contextual information from different levels of network through more attention. The experimental results demonstrated that our proposed architecture improves the feature representation of clothes images and achieves comparable performance to the landmark-driven methods.

10. References

[1] Huang, J., Feris, R.S., Chen, Q., Yan, S., 2015. Cross-domain image retrieval with a dual attribute-aware ranking network, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1062–1070.
[2] Liu, J., Lu, H., 2018. Deep fashion analysis with feature map upsampling and landmark-driven attention, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 30–36.
[3] Corbiere, C., Ben-Younes, H., Ramé, A., Ollion, C., 2017. Leveraging weakly annotated data for fashion image retrieval and label prediction, in: Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 2268–2274.
[4] Wang,W., Xu, Y., Shen, J., Zhu, S.C., 2018. Attentive fashion grammar network for fashion landmark detection and clothing category classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4271–4280.
[5] Ye, Y., Li, Y., Wu, B., Zhang, W., Duan, L., Mei, T., 2019. Hard-aware fashion attribute classification. arXiv preprint arXiv:1907.10839.
[6] Lee, S., Eun, H., Oh, S., Kim, W., Jung, C., Kim, C., 2019a. Landmark-free clothes recognition with a two-branch feature selective network. IET Electronics Letters 55(13), 745–747.
[7] Cho, H., Ahn, C., MinYoo, K., Seol, J., Lee, S.G., 2019. Leveraging class hierarchy in fashion classification, in: Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 3197–3200.
[8] Zhang, Y., Zhang, P., Yuan, C.,Wang, Z., 2020. Texture and shape biased two-stream networks for clothing classification and attribute recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13538–13547.
[9] Shajini, M., Ramanan, A., 2020. An Improved Landmark-driven and Spatial-channel Attentive Convolutional Neural Network for Fashion Clothes Classification, The Visual Computer, Springer, 37, pp. 1517–1526.