SisterNetwork: Enhancing Robustness of Multi-label Classification with Semantically Segmented Images

Holim Lim[1*], Jeeseung Han[1*] and Sang-goo Lee[1]

[1] IntelliSys Corp., Seoul National University, Seoul, South Korea
{ihl7029, jshan, sglee}@europa.snu.ac.kr
[*] **These authors contributed equally to this work.**

Abstract. The rapid growth of the online fashion market has raised the demand for fashion technologies, such as clothing attribute tagging. However, handling fashion image data is challenging since fashion images likely contain irrelevant backgrounds and involve various deformations. In this paper, we introduce SisterNetwork, a deep learning model to tackle the multi-label classification task for fashion attribute tagging. The proposed model consists of two different CNNs to leverage both the original image and the semantic segmentation information. We evaluate our model on the DCSA dataset which contains tagged fashion images, and we achieved the state-of-the-art performance on the multi-label classification task.

Keywords: Multi-label classification, Semantic segmentation, Fashion.

1 Introduction

The online fashion market is growing rapidly, and the demand for related technology is increasing. When dealing with fashion, it is important to exploit image data since visual information is the most principal feature in the fashion domain. Among those technologies related for online fashion retail, extracting human-understandable fashion attributes from images is one of the most crucial tasks since it serves as a basic module for most retail applications such as fashion search [1] or fashion recommendation [2].

However, dealing with unconstrained fashion images is a highly challenging task in computer vision for two major reasons. First, "user photos" such as outfit-of-the-day photos and selfies are likely to contain irrelevant information such as background and other objects. Second, fashion images frequently suffer various deformations such as deviation of focus, diverse contrast depending on the position of the light source, pose of the human model and etc. Therefore, it is difficult to extract robust image features that can handle a wide range of the fashion domain.

In this paper, we introduce SisterNetwork, a deep learning model that leverages the augmented information to tackle a robust multilabel classification task for tagging fashion images. The model consists of two different Convolution Neural Network (CNN) that accept original images and semantically segmented images, respectively, followed by “fusing” layer and fully-connected layer for classifying. A semantically segmented

image is a preprocessed image from which irrelevant background pixels are removed. To “fuse” two different outcomes of CNNs, we first apply max-pooling respectively to obtain vectors of the same dimension and then fuse them through the fusing layer. In the fusing layer, we adopt two methods: elementwise addition and Hadamard product.

For evaluation, we performed a multi-label classification task for fashion attribute tagging on the DCSA dataset [9], which contains 1,856 “user photo” images with 26 labels. Through the experiment, we found that exploiting semantic segmentation information yield, considerable improvement, compared to the baseline model that only considers original images. As a result, the proposed model achieved the state-of-the-art performance.

2 Related works

2.1 Learning from Noisy Fashion Data

In the clothing detection task, focusing on a region of interest is necessary in order to minimize noises. For this purpose, there are object detection methods [3, 4, 5] that detects the bounding box localization of objects of interest and semantic segmentation methods [6, 7, 8] to segment the specific object for guiding fashion-related models.

In addition to, pose-based approaches have been suggested. Chen et al [9] used the pose estimation method to estimate the skeleton on the upper body and remove the background except the upper body area using the grab-cut based on the ratio set by the heuristic method. Liu et al [10] proposed an approach called Landmark detection. They used the method to detect the position of a certain point which can be a feature of a fashion item and to use the feature by pooling the surrounding information.

We propose that the semantic segmentation method approach to solving the multi-label classification problem would improve the supervised CNN network.

2.2 Dataset

As a fashion domain dataset, we used the DCSA dataset, which was introduced by Chen et al [9]. DCSA dataset contains 1856 photos with 26 ground truth clothing attributes such as "long sleeves", "has a collar", and "striped pattern". The labels were collected using Amazon Mechanical Turk. We find that the tagset designed in DCSA is intuitive and clearly identifiable. For these reasons, we choose DCSA as the target task for our studies.

3 Approach

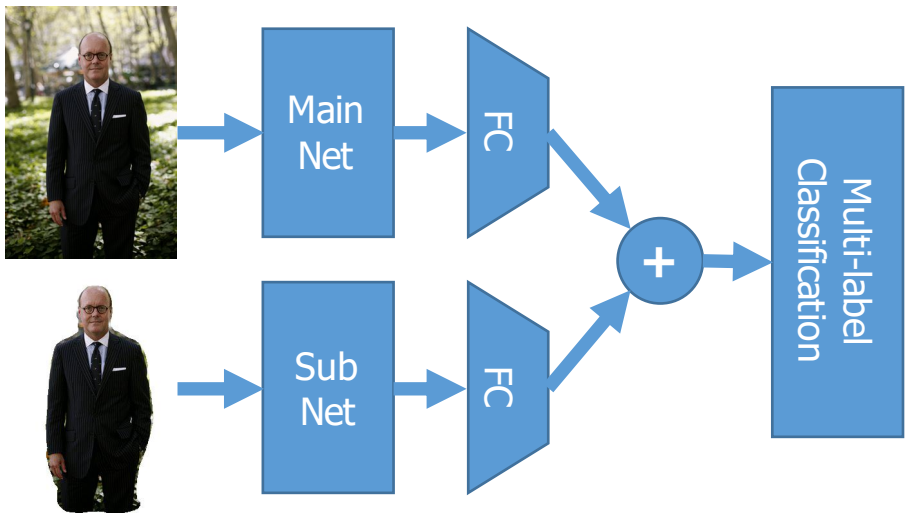

Fig. 1. Overview of Sister-A structure

We propose a novel deep model structure, SisterNetwork which simultaneously exploits original images and semantically segmented images. We name this structure "Sister" as original images go through the bigger backbone network and preprocessed images go through the other smaller network. Then two outputs of sister networks are combined and fed into the last multi-label classification layers.

3.1 Semantic Segmentation

In this paper, our goal is to classify annotated labels on unconstrained images. Hence, we argue that using images preprocessed by semantic segmentation models could be useful. As [9] shows, color labels and bigger pattern like 'solid' or 'graphics' achieve higher accuracy. But labels like 'neckline' which can be inferred from fine-grained features exhibit lower accuracy. So to exploit these local features, we need to use both original images and semantic segmentation images at the same time.

3.2 Sister Structure

The most intuitive approach to encoding two similar images is to use a symmetric (or twin) network in which two identically shaped image-encoding networks encode respective images. However, in our experiments, we find that the network for encoding processed images requires less expressive power, hence employing a completely same network architecture performs worse than a more specialized network. In this paper, we show that our asymmetric architecture is more effective than the counterpart.

4 Experiments and Details

We trained and evaluated our models on DCSA [9] using 16-fold cross-validation. DCSA contains 23 binary labels and 3 multiclass labels. For the binarized labels, we compute scores indicating the likelihood of the visual presence for those labels. And the other multi-class labels, probabilities for each class are computed. We evaluate our models using the unweighted mean of all label accuracies.

4.1 Base Settings and Segmentation models

We use Resnet[11] pretrained on ImageNet. Input images are feed into Resnet to predict 26 labels. The last average pooling layer is replaced with 7x7 max pooling layer. Also, the last fully connected layer is changed to 2 fully connected layers. The first fully connected layer is followed by ReLU and the number of filters is set to 1024. Also, the model is optimized with Adam at learning rate 10^{-5}. We denote this experimental setting as *the base setting*.

Before the main experiments, we select the best segmentation under the base setting. We choose the model that achieves the highest mean accuracy using images processed by each approach. Specifically, we use Resnet-50 at batch size 64 to select the segmentation model.

Table 1. MACC using Segmented Features on Resnet50

	V3-Black	V3-White	FCN-White
Title Set	CFPD	CFPD	VOC
MACC	90.43	90.45	**91.06**

We consider FCN-8s [6] and DeepLab V3 [7] as our candidates. Table 1 shows the results of each model and training data for semantic segmentation. DeepLab V3 models are trained on CFPD [13] and remove backgrounds in the DCSA [9] images. V3-Black and V3-White make backgrounds black or white each. FCN model is trained on VOC data, and only leaves pixels classified as a person. As FCN achieves the best results, we use it for other experiments.

4.2 Baselines and Compared Methods

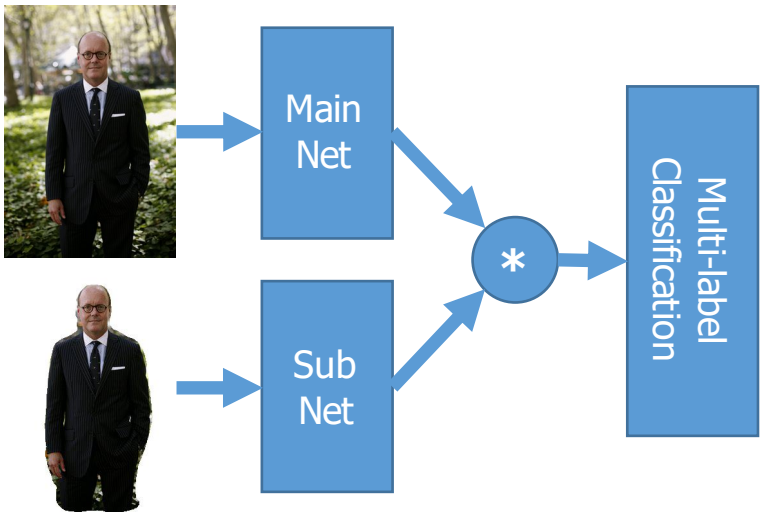

Fig. 2. Overview of Sister-B structure

Baseline. The baseline for each experimental setting is the network that predicts based on the unprocessed images. ResNet variants are used as the encoder network. We trained baseline models under the base setting. Batch size was 64 in the Resnet-50 baseline and 32 for other cases.

Sister-A. These approaches are shown in Fig. 1. Sister-As utilize both processed images and original images. Unprocessed images are fed into the backbone and processed images are fed into the small one which is Resnet-50. And each output goes through one fully-connected layer followed by ReLU. The number of filters is set to 1024. Then they are fused with elementwise addition. The last layer computes scores from this summed vector. In the Resnet-152 case, batch size was 16. And in other cases, batch sizes were 32.

Sister-B. These approaches are shown in Fig. 2. These models are similar to Sister-As. But they do not have a fully connected layer following the backbone neither the small one. Also, the output features are not summed but element-wise multiplied (Hadamard product). Batch sizes were the same as in Sister-A.

Table 2. Comparison between Naïve and Twin Models

Backbone	Baseline	Sister-A	Sister-B
Resnet-50	91.40	**91.63**	91.62
Resnet-101	91.66	**91.73**	90.50
Resnet-152	91.65	**91.91**	90.97

Table 2 shows mean accuracies of the approaches. We see that the baselines using Resnet-101 and Resnet-152 show similar results although Resnet-152 has a larger number of parameters. On the other hand, our Sister-A models gain higher accuracies as backbone networks get larger.

We also reevaluate Resnet-101 Sister-A model with a small modification that the small network is set to be Resnet-101. In this case, we can get a mean accuracy of 91.59%. This shows that the asymmetricity of our model is effective.

5 Discussion

5.1 Employing Gated Methods

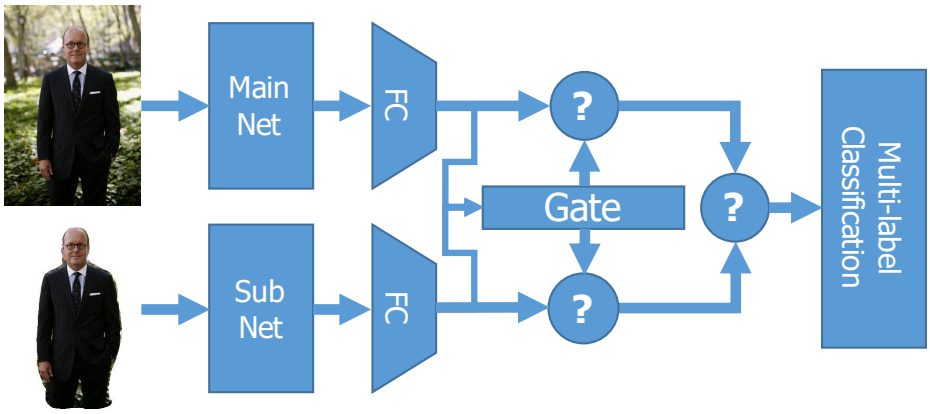

Fig. 3. Structure Employing Gated Methods

As sister models present higher accuracies than baselines, we conduct additive experiments using other gate methods. These experiments use Resnet-50 as backbone networks. Detailed structures are shown in Fig. 3. These models are optimized with Adam at learning rate 10^{-5} and batch sizes are 32. We note that the gating approaches exhibit lower accuracies compared to Resnet-50 baseline.

Table 3. Gated Twin Models with Resnet-50

Model	Single-value Gate	Element-wise Gate (Add)	Element-wise Gate (Concat)
MACC	90.12	90.32	90.43

5.2 Micro and Macro Features

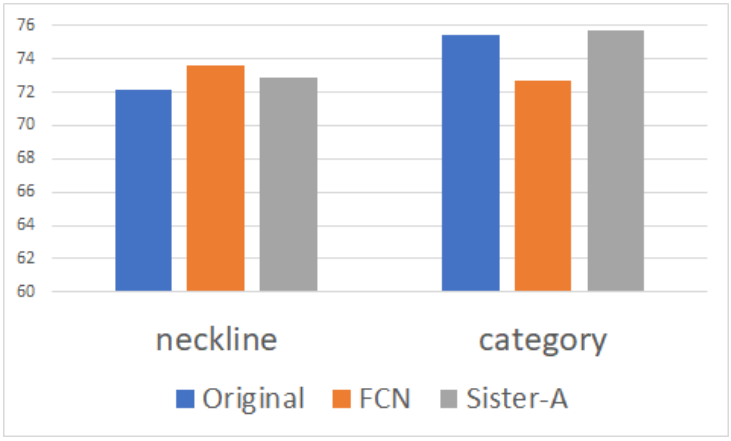

Fig. 4. 'neckline' and 'category' accuracies using Resnet-50

Fig. 4 shows 'neckline' and 'category' label accuracies using Resnet-50 baseline, FCN-White, and Resnet-50 Sister-A. The baseline model achieves higher accuracy in the category. And FCN-White achieves higher accuracy in the neckline. We argue that using images processed by semantic segmentation leads the model to learn detail features and using whole images leads the model to learn overall features. Combining these two, our model can gain more accuracies from both micro and macro features.

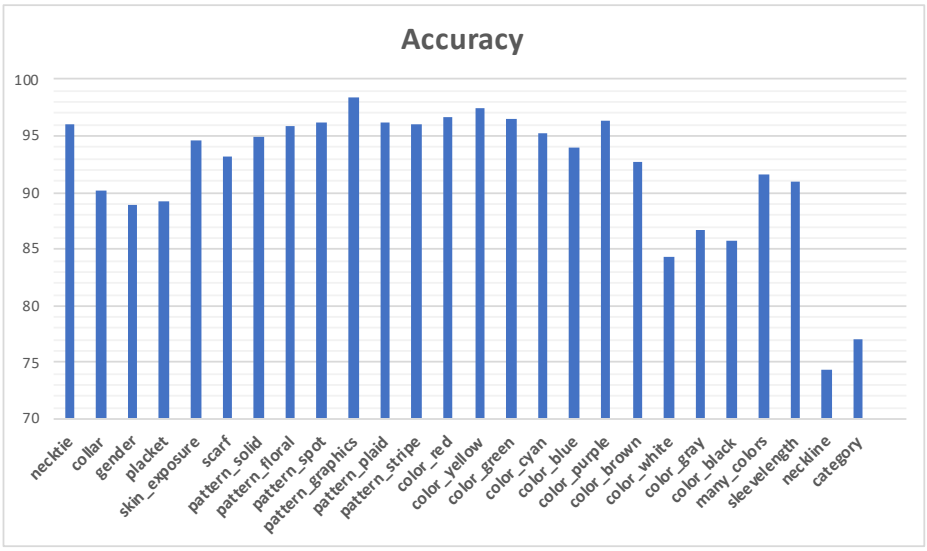

Fig. 5. Sister-A model label accuracy

6 Conclusions

In this paper, we proposed and addressed the multi-label, multiclass classification in the fashion domain. We proposed novel deep learning architecture, namely SisterNetwork which leverages both whole images and preprocessed images. Exploiting semantic segmentation models, we conducted diverse experiments to achieve higher accuracies. A key feature of our model is the use of semantic segmentation information and an asymmetric network architecture.

Albeit our model is quite simple and elegant, there might be a room for developing more elaborate approaches. Although gated methods that we used were not significant yet, we believe further work would be beneficial for the multi-label classification task.

Acknowledgments. This work was supported by the Technology development Program (S2646078) funded by the Ministry of SMEs and Startups (MSS, Korea).

References

1. Wei, Di, et al. "Style Finder: fine-grained clothing style recognition and retrieval." Computer Vision and Pattern Recognition Workshops. 2013
2. Zhou, Wei, et al. "Fashion recommendations using text mining and multiple content attributes." (2017).
3. Girshick, Ross. "Fast R-CNN." Proceedings of the IEEE international conference on computer vision. 2015.
4. Liu, Wei, et al. "SSD: Single shot multibox detector." European conference on computer vision. Springer, Cham, 2016.
5. Redmon, Joseph, and Ali Farhadi. "YOLO9000: better, faster, stronger." arXiv preprint (2017).
6. Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
7. Chen, Liang-Chieh, et al. "Encoder-decoder with atrous separable convolution for semantic image segmentation." arXiv preprint arXiv:1802.02611 (2018).
8. He, Kaiming, et al. "Mask R-CNN." Computer Vision (ICCV), 2017 IEEE International Conference on. IEEE, 2017.
9. Chen, Huizhong, Andrew Gallagher, and Bernd Girod. "Describing clothing by semantic attributes." European conference on computer vision. Springer, Berlin, Heidelberg, 2012.
10. Liu, Ziwei, et al. "DeepFashion: Powering robust clothes recognition and retrieval with rich annotations." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
11. He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
12. Liu, Si, et al. "Fashion parsing with weak color-category labels." IEEE Transactions on Multimedia 16.1 (2014): 253-265.