Supplementary Information

1 Dataset Modification and Model Structure

1.1 Dataset Modification

The dataset we choose, DeepFashion (category and attribute prediction benchmark), has some labeling problems. Bounding-box-wise, we removed bounding boxes with odd aspect ratios (height/weight lower than 0.2 or higher than 5) or extremely small area (less than 0.21% of the whole image). Attribute-wise, we manually removed 45 unclear attributes (such as "girl" and "please") and merged semantically similar attributes, (for example, "abstract geo" vs. "abstract geo print" vs "geo" vs "geo pattern" vs "geo print"). The cleaned dataset ended up with 544 diverse clothing attributes and 50 clothing categories. There are still wrongly labeled (false positive) categories, attributes and bounding boxes in this dataset, e.g., recognizing a skirt as a dress, but we are not able to deal with them due to expensive labor cost.

1.2 Model Architecture

We extend the Faster R-CNN object detection framework [1] with ResNet 101 and ROI-align (implemented by Google Research [2]) with two modifications: a pruning mechanism and additional clothing attributes branches parallel to category branch. Figure 1 shows the overall model architecture.

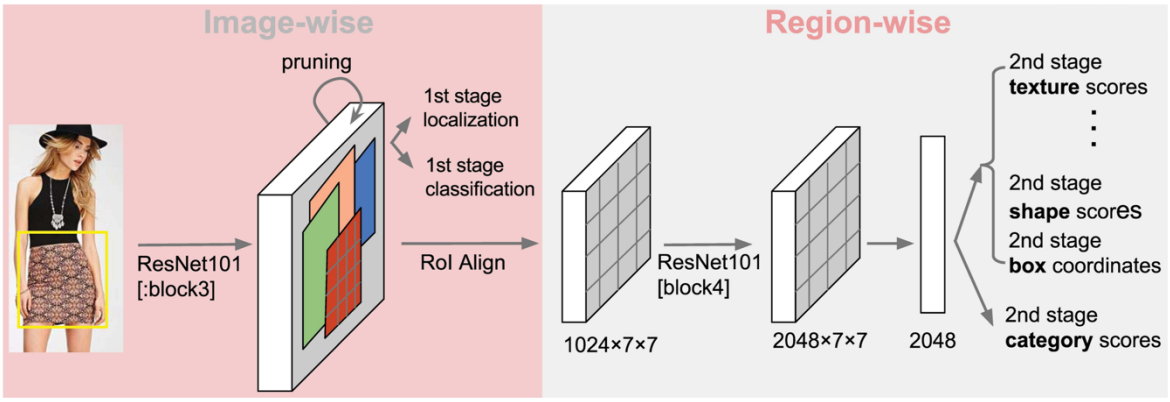

Figure 1: Model Architecture

Additional pruning. Faster R-CNN classifies objectiveness of each densely distributed proposals at the first Region Proposal Network (RPN) stage. Each proposal is labeled as positive/negative or ignored based on its IoU value with the groundtruth boxes. Since DeepFashion dataset gives only one box for a single image, other clothing items in this image (especially upper body vs. lower body) will be classified as background. This would confuse the detection model and decrease the performance. In order to solve such problem, we propose an additional pruning process at the first stage. Specifically, we introduce groundtruth people boxes for each image, and prune away any proposals that classified as non-objects but have an IoU of a certain value or higher with groundtruth people boxes (Figure 2). We use SSD-Mobilenet [2] to extract groundtruth people

boxes. It is worth mentioning that a small fraction (9.9%) of images display clothes other than models, out of which a large portion display single clothing item.

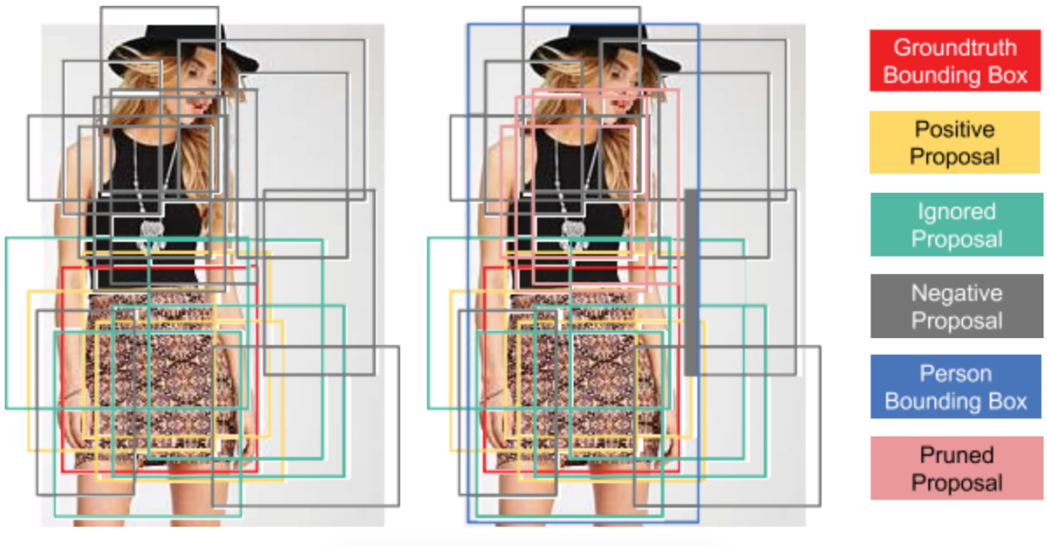

Figure 2: Pruning Mechanism During Training

Attribute branches. We approach learning both attributes and category as a multi-task learning problem. The attributes branches reuse features extracted by RPN. We propose three different structures as indicated in Figure 3. In detail, (a) uses 3 convolution layers ($1 \times 1 \times 512$ with padding, $7 \times 7 \times 512$ without padding, and $1 \times 1 \times 2048$ without padding) followed by 5 fully connected layers for each attribute type scores; (b) shares the same flattened proposal features as category classifier, followed by a fully connected layers (1024) and 5 fully connected layers for each attribute type scores; (c) shares the same flattened proposal features as category classifier, followed by 5 fully connected layers for each attribute type scores.

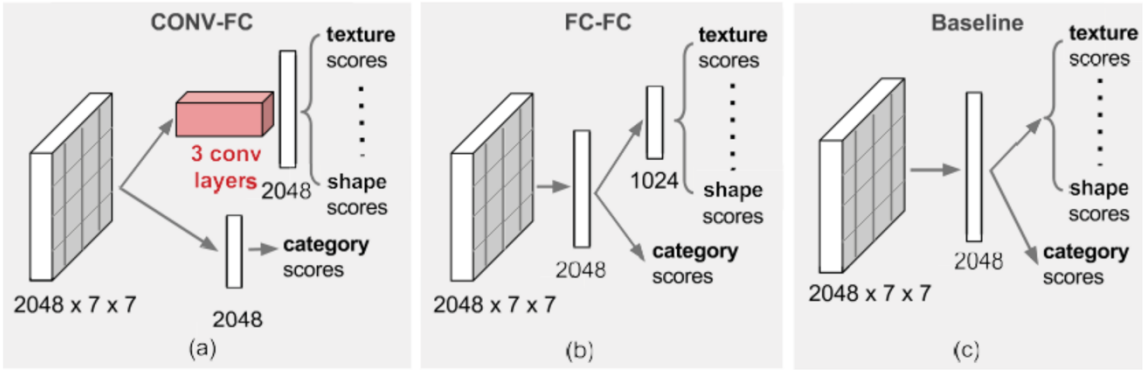

Figure 3: Attribute Branches

2 Experiment and Analysis

2.1 Test and Results

We test our trained model on three datasets: (a) 2,094 selected images from DeepFashion (category and attribute prediction benchmark) test partition; (b) 92 images from ready-to-wear runway photos; (c) 92 images from fashion technical sketches. By default, 300 detections are generated for each image without score threshold. For evaluation of category prediction, we consider two metrics: **(i) average precision (AP) per class and weighted mAP.** We use all the predictions with scores higher than 0.5 per image to compute true-positive and false-positive labels per class with a matching IoU threshold of 0.5 with groundtruth boxes. Then these labels are used to calculate AP per class and weighted mAP (concatenating all labels regardless of classes); **(ii) CorLoc per class and weighted mean CorLoc.** For each image, we pick the top 5 detections per image and check if the grountruth is correctly detected per class with a matching IoU threshold of 0.5 with groundtruth boxes. CorLoc is computed as the ratio of number of detected groundtruth instances over number of total groundtruth instances per class, and weighted mean CorLoc is such ratio regardless of classes. For evaluation of attribute prediction (only on DeepFashion dataset), we label attribute detections as positive and negative with a score threshold of 0.5. For each image, we merge all detections that have an IoA higher than 0.7 over the detection with the highest category score, using logical "and" operation. Then we calculate true-positive and false-positive labels for each attribute and generate **precision and recall per attribute and precision and recall per attribute type.** Table 1 presents the corresponding test results. We used two IoU threshold (0.3 and 0.7) for additional pruning and three structures as attribute branch. Comparisons are made with models restored from checkpoints under the same epoch. Note that conv-fc attribute branch doesn't give any positive attribute prediction.

Dataset	weighted mAP			weighted mean CorLoc		
	no pruning	pruning 0.3	pruning 0.7	no pruning	pruning 0.3	pruning 0.7
DeepFashion	0.1425	0.1603	0.1715	0.6418	0.6336	0.6772
Runway	0.0996	0.0865	0.0980	0.4130	0.4130	0.4891
Sketches	0.0639	0.0422	0.0748	0.3804	0.3261	0.3913

Dataset	Attribute Precision			Attribute Recall		
	baseline	fc-fc	conv-fc	baseline	fc-fc	conv-fc
DeepFashion	0.0448	0.0899	-	0.2932	0.2127	-

Table 1: Test Results

2.2 Analysis

From the experimental results, we can see the performance of our model can be improved in four areas:

Data imbalance. The training data we used consists of 46 categories and some categories have only tens or hundreds of images while other categories can have over 70, 000 images. This imbalance makes it really hard to train the model so that it can detect those minor categories.

Wrongly labeled data. There are still a lot of wrongly labeled categories and attributes in the dataset even after our data cleaning.

Unbounded objects. We tackled this issue using proposal pruning, however, there are still such cases considering the limitation of pruning criterion and they also occur in the test data, which adversely affects the evaluation.

Too many negative attribute labels. In the training data, each attribute has way more negative instances than positive ones. We didn't deal with issue, and as a result, the attribute classifier is not well established.

3 Future Work

Optimization methods. To improve optimization performance, we want to compare different optimization strategies. We'd like to explore using Adam optimizer with manual learning rate decay compare different optimization strategies and using batch size of 1 instead of mini-batch for gradient descent.

Dealing with data imbalance: As mentioned above, there's significant data imbalance among categories and between positive and negative labels for attributes. Stratified sampling [3] and weighted cross-entropy loss [4] might be of help.

Finer feature recognition. We use feature maps with low resolutions but large receptive field, thus detailed attributes on clothes (such as side-zippers, or small embroideries at collars) may not be easily recognized. We may consider using Feature Pyramid Networks (FPN) or multi-scale DenseNet to improve it.

Domain adaptation. We will explore how to improve detection performance of a model trained from one domain on another domains (haute couture runway photos, artistic fashion drawings, etc.).

References

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in Advances in neural information processing systems, 2015, pp.91–99.

[2] J.Huang,V.Rathod,C.Sun,M.Zhu,A.Korattikara,A.Fathi,I.Fischer,Z.Wojna,Y.Song,S.Guadarrama et al., "Speed/accuracy trade-offs for modern convolutional object detectors," arXiv preprint arXiv:1611.10012, 2016.

[3] K.Matzen,K.Bala,andN.Snavely,"Streetstyle:Exploring world-wide clothing styles from millions of photos," arXiv preprint arXiv:1706.01869, 2017.

[4] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, “Deepfashion: Powering robust clothes recognition and retrieval with rich annotations,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1096–1104.