DeepMark++: CenterNet-based Clothing Detection

Alexey Sidnev[1,2], Alexander Krapivin[1], Alexey Trushkov[1], Ekaterina Krasikova[1], Maxim Kazakov[1,3]

[1]Huawei Research Center, Nizhny Novgorod, Russia
[2]Lobachevsky State University of Nizhny Novgorod, Russia
[3]National Research University Higher School of Economics, Nizhny Novgorod, Russia

{sidnev.alexey, krapivin.alexander, trushkov.alexey, krasikova.ekaterina, kazakov.maxim}@huawei.com

Abstract

The single-stage approach for fast clothing detection as a modification of a multi-target network, CenterNet [11], is proposed in this paper. We introduce several powerful post-processing techniques that may be applied to increase the quality of keypoint localization tasks. The semantic keypoint grouping approach and post-processing techniques make it possible to achieve a state-of-the-art accuracy of 0.737 mAP for the bounding box detection task and 0.591 mAP for the landmark detection task on the DeepFashion2 validation dataset [2]. We have also achieved the second place in the DeepFashion2 Challenge 2020 with 0.582 mAP on the test dataset. The proposed approach can also be used on low-power devices with relatively high accuracy without requiring any post-processing techniques.

1. Introduction

In recent studies, keypoints, which are also referred to as landmarks, have proved to be one of the most distinctive and robust representations of visual analysis. The class of keypoint-based methods in computer vision includes the detection and further processing of keypoints. These methods can be utilized in tasks such as object detection, pose estimation, facial landmark recognition, and more.

The performance of models that operate with keypoints highly depends on the number of unique landmarks defined in the task, which may be considerably large for some modern datasets. Deepfashion2, one of the newest fashion datasets, provides annotations for 13 classes, each being characterized by a certain set of keypoints, with 294 unique ones in total.

In this paper, we propose a study on the clothing landmark detection task on the the DeepFashion2 dataset, as well as an approach to deal with it efficiently.

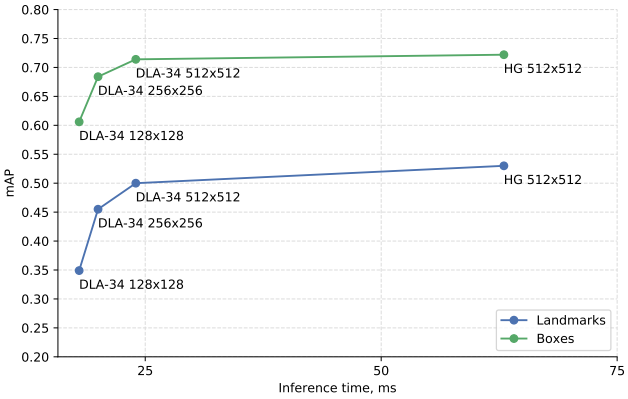

Figure 1. Speed-accuracy trade-off for object detection and landmark estimation on the DeepFashion2 validation dataset [2]. NMS post-processing is applied to every model.

2. Related work

In general, there are numerous applications for the keypoints estimation task. For instance, keypoints can be used directly to identify human pose [7] or locate facial landmarks [10], and as the main part of the object detection pipeline [4]. Keypoint-based object detection methods are gaining popularity in recent papers, especially because they are simpler, faster, and more accurate than the corresponding bounding box-based detectors.

Previous approaches, such as [5], required anchor boxes to be manually designed to train detectors. A series of anchor-free object detectors were then developed, with the aim of predicting the bounding box's keypoints, instead of trying to fit an object to an anchor. Without relying on manually designed anchors to match objects, CornerNet's [4] performance on MS COCO datasets improved significantly. Subsequently, several other variants of keypoint detection-based one-stage detectors came into existence, one of which was CenterNet [11].

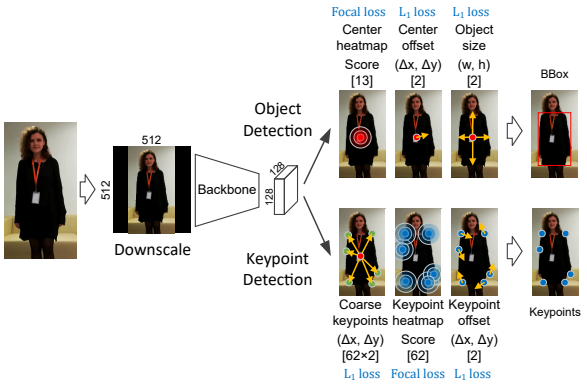

Figure 2. Scheme of the proposed approach.

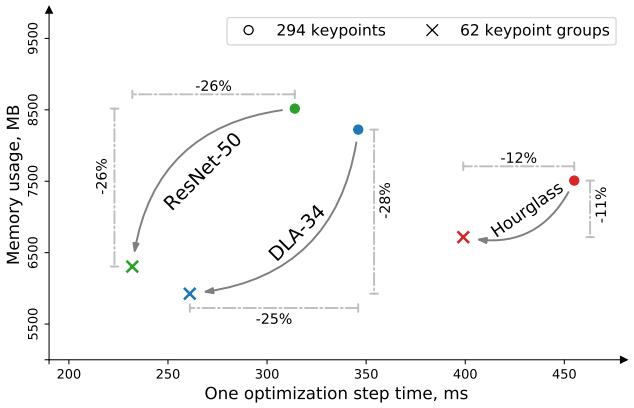

Figure 3. GPU memory consumption and training iteration time on RTX 2080ti. The input resolution is 256×256, the batch size is 32 for both DLA-34 and ResNet-50, and 8 for Hourglass. Time in ms was measured for 1 optimization step: batch loading to GPU, forward pass and backward pass. GPU memory was measured by using the nvidia-smi tool.

This paper focuses on clothing landmark prediction and clothing detection tasks using the DeepFashion2 datasets [2]. The baseline approach for landmark estimation was built on Mask R-CNN [3] due to its two-stage nature, which is substantially heavy and difficult to use on low-power devices. We aim to propose a lightweight architecture that is not as heavy. These requirements are perfectly met by CenterNet, which operates directly with keypoints.

3. Proposed approach

Our approach is based on the CenterNet [11] architecture (see Figure 2). It solves two tasks simultaneously: object detection and keypoint location estimation.

The DeepFashion2 dataset contains 13 classes; therefore, 13 channels are used to predict the probabilities of object centers for all classes (Center heatmap in Figure 2). An object center is defined as the center of a bounding box. Two additional channels in the output feature map $\triangle$x and $\triangle$y are used to refine the center coordinates, and both width and height are predicted directly.

The fashion landmark estimation task involves estimating 2D keypoint locations for each item of clothing in one image. The coarse locations of the keypoints are regressed as relative displacements from the box center (Coarse keypoints in Figure 2). To refine a keypoint location, a heatmap with probabilities is used for each keypoint type. Local maximum with high confidence in the heatmap is used as a refined keypoint position. Similar to the detection case, two additional channels $\triangle$x and $\triangle$y are used to obtain more precise landmark coordinates. During model inference, each coarse keypoint location is replaced with the closest refined keypoint position.

3.1. Semantic keypoint grouping

One of the first steps involved in solving keypoint detection tasks is defining the model output. The number of keypoints for every category varies from 8 for a skirt to 39 for a long sleeve outerwear in the DeepFashion2 dataset. The total number of unique keypoints is 294. The simple approach is to concatenate keypoints from every category and deal with each keypoint separately. Directly predicting 294 keypoints leads to a huge number of output channels: $901 = 13 + 2 + 2 + 294 \cdot 2 + 294 + 2$ ($294 \cdot 2$ from coarse keypoints, 294 from keypoint heatmap).

It is evident that certain clothing landmarks are a subset of others. For example, shorts do not require unique keypoints because they can be represented by a subset of trousers keypoints. The semantic grouping rule is defined as follows: identical semantic meaning keypoints (collar center, top sleeve edge, etc.) with different categories can be merged into one group. This approach enables the formation of 62 groups and reduces the number of output channels from 901 to 205.

The semantic grouping approach reduces training and memory consumption times by up to 26% and 28%, respectively, without accuracy drops (see Figure 3). The latter reduction enables the use of larger batches during model training.

3.2. Post-processing techniques

We have developed 4 post-processing techniques that increase the model's accuracy without compromising performance.

3.2.1 Center rescoring

The first technique is a recalculation of the detection confidence score using keypoint scores from the keypoint heatmap. Let $Score_{bbox}$ be the original detection confi-

dence score from the center heatmap and $Score_{kps}$ be the average score of refined keypoints for the predicted category from the keypoint heatmap. The final detection confidence scores are calculated through the following formula:

$$Score = \alpha \cdot Score_{bbox} + (1 - \alpha) \cdot Score_{kps}, \quad (1)$$

where $\alpha \in [0, 1]$.

3.2.2 Heatmap rescoring with Gaussian kernel

The second technique is a general approach that can be applied to any keypoint-based architecture. Let H be a heatmap with the center or keypoint scores. Taking the training procedure into account, you can expect the 8-connected neighbors of each item to be related to the same object. This fact can be used to improve the estimation of each heatmap value. Therefore, we applied the following formula:

$$\hat{H} = H \circledast G(\sigma), \quad (2)$$

where $\circledast$ is the convolution operation, $G(\sigma)$ is the 3×3 Gaussian kernel with the standard deviation σ. Experimental results show that in our model, the proposed technique improves the localization of peaks and their values that correspond to object centers or keypoints and their scores. A similar operation has been considered in [9] as a part of the proposed method.

3.2.3 Keypoint location refinement

The third technique is a recalculation of the refined keypoint locations using coarse positions. Let $(x, y)_{refined}$ be the refined keypoint location from the heatmap and $(x, y)_{coarse}$ be coarse positions predicted as offsets from object centers. The final keypoint locations were calculated through the following expression:

$$(x, y) = \gamma \cdot (x, y)_{refined} + (1 - \gamma) \cdot (x, y)_{coarse}, \quad (3)$$

where $\gamma \in [0, 1]$.

3.2.4 Keypoint heatmap rescoring

The fourth technique is a keypoint heatmap rescoring with the $mask$ from a coarse keypoint location. Let $mask$ be a heatmap with zero values by default. We set 1 into the $mask$ in the coarse keypoint position and fill neighbor values with 2D Gaussian function with standard deviation $sigma = \min(width, height)/9$, where $width$ and $height$ are the object size. The keypoint heatmap is rescored through the following expression:

$$\hat{H}_{kps} = H_{kps} \cdot mask. \quad (4)$$

Post-processing	mAP_{pt}	mAP_{box}	Infer. time, ms
Base	0.529	0.720	62
NMS	0.530	0.722	62
Technique (1)	0.538	0.717[1]	64
Technique (2)	0.533	0.720	73
Technique (3)	0.536	0.720	62
Technique (4)	0.534	0.720	93

Table 1. Different post-processing techniques applied independently to Hourglass 512×512. The technique numbers correspond to the numbers in section 3.2.

3.3. Multi-inference strategies

We consider 2 extra inference strategies: fusing model outputs from original and flipped images with equal weights; fusing model results with the original image downscaled/upscaled through certain multipliers. The proposed techniques increase accuracy but require several model inferences, significantly affecting the entire processing time.

3.4. Keypoint Refinement Network

At the final stage, detection results are refined with the PoseFix [6] model-agnostic pose refinement method. The method learns the typical error distributions of any other pose estimation method and corrects mistakes at the testing stage.

We trained a set of 13 PoseFix models by using the number of classes in the DeepFashion2 dataset. The inference results of our method on the training set are used to train each of the 13 models. Subsequently, we applied trained PoseFix models to the result.

4. Results

All experiments were performed on the publicly available DeepFashion2 Challenge dataset [2], which contains 191,961 images in the training set and 32,153 images in the validation set.

We used the CenterNet MS COCO model for object detection as the initial checkpoint and performed experiments with Hourglass backbone and Adam optimizer to achieve the state-of-the-art results (Table 2) for object detection and keypoint estimation tasks on the DeepFashion2 validation dataset. The Hourglass 512×512 model was trained for 100 epochs with a batch size = 46 images. Learning rate schedule: 1e-3 - 65 epochs, 4e-4 - 20 epochs, 4e-5 - 15

[1]Certain techniques can increase mAP_{pt} and reduce mAP_{box} simultaneously. Note that bounding box detection and keypoint estimation results for the same object may have different IoU and OKS with the ground truth, for example, when bounding box was detected correctly, but keypoints were not. In this case, technique (1) involves lowering a score for false positive keypoints, which is advisable. The corresponding true positive bounding box also suffers from this lowered score.

Approach	mAP_{pt}	mAP_{box}
Mask R-CNN [2]	0.529	0.638
DeepMark [8]	0.532	0.723
DAFE [1]	0.549	—
Hourglass 512×512	0.583	0.735
Hourglass 768×768	**0.591**	**0.737**

Table 2. Accuracy comparison of the proposed and alternative approaches with the DeepFashion2 validation dataset.

Metric	mAP_{pt}		mAP_{box}	
Resolution	512×512	768×768	512×512	768×768
Base	0.529	0.520	0.720	0.695
+ Post-processing	0.545	0.540	0.713	0.698
+ NMS	0.548	0.549	0.718	0.712
+ Flip	0.561	0.563	0.731	0.727
+ Multiscale	0.568	0.578	**0.735**	**0.737**
+ PoseFix	**0.583**	**0.591**	**0.735**	**0.737**

Table 3. Clothing detection and landmark estimation. Hourglass with 512×512 and 768×768 resolution were used in the experiments. The next technique is added to each of the previous ones. The post-processing refers to applying all techniques from section 3.2.

Approach	Parameter value
Technique (1)	$\alpha = 0.8$
HG 512×512 technique (2)	Center heatmap $\sigma = 0.40$ Keypoint heatmap $\sigma = 0.85$
HG 768×768 technique (2)	Center heatmap $\sigma = 0.45$ Keypoint heatmap $\sigma = 0.90$
Technique (3)	$\gamma = 0.75$
Mutiscale	Multipliers: 0.85, 0.95, 1.1

Table 4. Parameters of post-processing and multi-inference techniques.

epochs. Hourglass 768×768 model was fine-tuned from Hourglass 512×512 for 25 epochs with a batch size = 22 images: 2e-5 - 20 epochs, 1e-5 - 5 epochs.

We considered 5 fast post-processing techniques: bounding box non-maximum suppression and 4 techniques from section 3.2. The individual (Table 1) and combined (Table 3) effectiveness of each technique has been shown. During all experiments, our target was to increase the keypoint estimation accuracy instead of the object detection accuracy. Due to this reason, object detection increased only by 0.015 mAP_{box}, but all the techniques added more than 0.07 mAP to mAP_{pt}.

The parameters of the post-processing techniques α, σ and γ were determined through grid searching with step 0.05 on a small validation subset (1285 images).

5. Conclusion

This new approach is proposed as an adaptation of CenterNet [11] for clothing landmark estimation tasks. The state-of-the-art accuracy was achieved on the DeepFashion2 dataset by applying several post-processing techniques: clothing detection hit 0.735 mAP and clothing landmark estimation – 0.591 mAP. The proposed approach can also be used without post-processing techniques. It takes 24 ms per image on RTX 2080ti for DLA-34 512×512, and yields considerably high accuracy (0.5 mAP_{pt} and 0.714 mAP_{box}) for clothing detection tasks (see Figure 1).

References

[1] Ming Chen, Yingjie Qin, Lizhe Qi, and Yunquan Sun. Improving fashion landmark detection by dual attention feature enhancement. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.

[2] Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo. DeepFashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5337–5345, 2019.

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017.

[4] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 734–750, 2018.

[5] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[6] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Posefix: Model-agnostic general human pose refinement network, 2018.

[7] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation, 2016.

[8] Alexey Sidnev, Alexey Trushkov, Maxim Kazakov, Ivan Korolev, and Vladislav Sorokin. Deepmark: One-shot clothing detection. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.

[9] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. *arXiv preprint arXiv:1910.06278*, 2019.

[10] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *European conference on computer vision*, pages 94–108. Springer, 2014.

[11] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.