状态

- 有pytorch开源实现
- 文章链接

简介

DCNv1虽然取得了一定效果，但是其sample的位置相对于roi还是不够dense而过于spread，而且实验基本用的是poscal VOC数据集数据量不够大，因此提出一些改进方法，并在coco数据集上进行了测试。

1. 将更多的卷积层换为DCN
2. 加入modulation机制使得不仅可以选择采样的位置，还可以控制采样的幅度(amplitude)。
3. 还采取了知识蒸馏的方法来训练模型，利用RCNN模型作为teacher进行了蒸馏。因为RCNN的输入是经过crop的image content，所以其feature不会被RoI之外的区域所影响。采用了mimicking loss的方法进行蒸馏。

可视化

- Effective receptive fields， 计算输出对输入的导数，可以反映出输出相对于输入扰动的变化
- Effective sampling/ bin locations， 之前只是可视化采样点的位置，但是没有揭示每个采样位置的影响大小。这个新的指标会the grad of network node with respect to the sampling / bin locations
- Error-bounded saliency regions 不同部分对最后的结果影响贡献是不同的。目的是找到对应响应区域的最小输入区域。

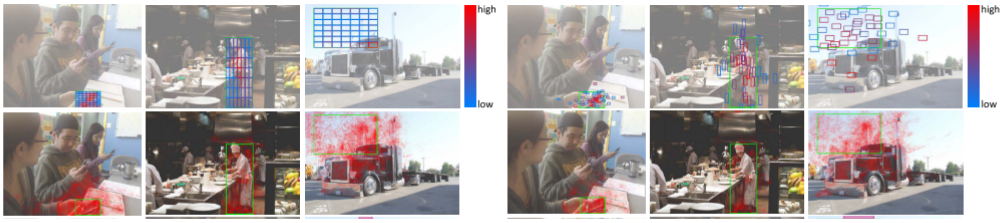
结果

(a) regular conv

(b) deformable conv@conv5 stage (DCNv1)

(c) modulated deformable conv@conv3∼5 stages (DCNv2)

Figure 1. Spatial support of nodes in the last layer of the conv5 stage in a regular ConvNet, DCNv1 and DCNv2. The regular ConvNet baseline is Faster R-CNN + ResNet-50. In each sub-figure, the effective sampling locations, effective receptive field, and error-bounded saliency regions are shown from the top to the bottom rows. Effective sampling locations are omitted in (c) as they are similar to those in (b), providing limited additional information. The visualized nodes (green points) are on a small object (left), a large object (middle), and the background (right).



(a) aligned RoIpooling, with deformable conv@conv5 stage

(b) deformable RoIpooling, with deformable conv@conv5 stage (DCNv1)

(c) modulated deformable RoIpooling, with modulated deformable conv@conv3∼5 stages

(d) with R-CNN feature mimicking on setting (c) (DCNv2)

(e) with R-CNN feature mimicking in regular ConvNet

Figure 2. Spatial support of the 2*fc* node in the per-RoI detection head, directly followed by the classification and the bounding box regression branches. Visualization is conducted on a regular ConvNet, DCNv1 and DCNv2. The regular ConvNet baseline is Faster R-CNN + ResNet-50. In each subfigure, the effective bin locations, effective receptive fields, and error-bounded saliency regions are shown from the top to the bottom rows, except for (c)∼(e) where the effective bin locations are omitted as they provide little additional understanding over those in (a)∼(b). The input RoIs (green boxes) are on a small object (left), a large object (middle), and the background (right).

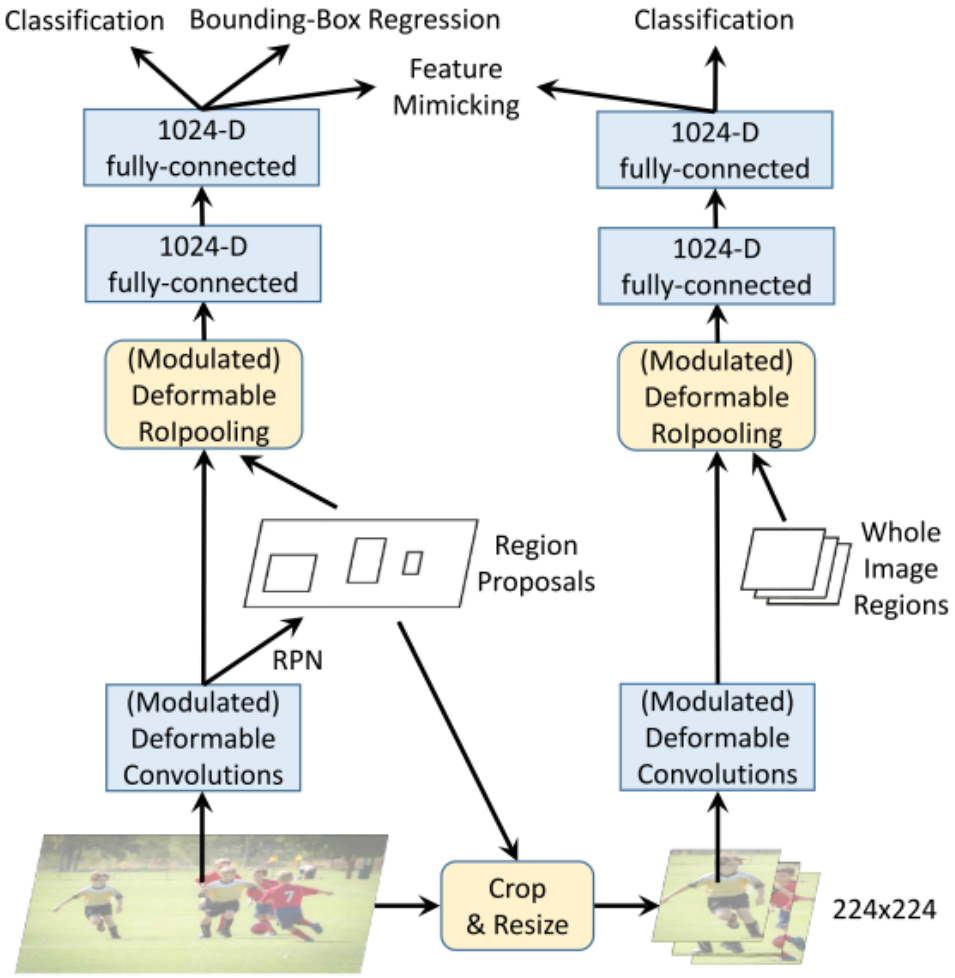


Figure 3. Network training with R-CNN feature mimicking.

实验

method	setting (shorter side 1000)	Faster R-CNN						Mask R-CNN			
		AP^{bbox}	AP^{bbox}_S	AP^{bbox}_M	AP^{bbox}_L	param	FLOP	AP^{bbox}	AP^{mask}	param	FLOP
baseline	regular (RoIpooling)	32.1	14.9	37.5	44.4	51.3M	326.7G	-	-	-	-
	regular (aligned RoIpooling)	34.7	19.3	39.5	45.3	51.3M	326.7G	36.6	32.2	39.5M	447.5G
	dconv@c5 + dpool (DCNv1)	38.0	20.7	41.8	52.2	52.7M	328.2G	40.4	35.3	40.9M	449.0G
enriched deformation	dconv@c5	37.4	20.0	40.9	51.0	51.5M	327.1G	40.2	35.1	39.8M	447.8G
	dconv@c4∼c5	40.0	21.4	43.8	55.3	51.7M	328.6G	41.8	36.8	40.0M	449.4G
	dconv@c3∼c5	40.4	21.6	44.2	56.2	51.8M	330.6G	42.2	37.0	40.1M	451.4G
	dconv@c3∼c5 + dpool	41.0	22.0	45.1	56.6	53.0M	331.8G	42.4	37.0	41.3M	452.5G
	mdconv@c3∼c5 + mdpool	41.7	22.2	45.8	58.7	65.5M	346.2G	43.1	37.3	53.8M	461.1G

Table 1. Ablation study on enriched deformation modeling. The input images are of shorter side 1,000 pixels (default in paper). In the setting column, "(m)dconv" and "(m)dpool" stand for (modulated) deformable convolution and (modulated) deformable RoIpooling, respectively. Also, "dconv@c3∼c5" stands for applying deformable conv layers at stages conv3∼conv5, for example. Results are reported on the COCO 2017 validation set.

method	setting (shorter side 800)	Faster R-CNN						Mask R-CNN			
		AP^{bbox}	AP^{bbox}_S	AP^{bbox}_M	AP^{bbox}_L	param	FLOP	AP^{bbox}	AP^{mask}	param	FLOP
baseline	regular (RoIpooling)	32.8	13.6	37.2	48.7	51.3M	196.8G	-	-	-	-
	regular (aligned RoIpooling)	35.6	18.2	40.3	48.7	51.3M	196.8G	37.8	33.4	39.5M	303.5G
	dconv@c5 + dpool (DCNv1)	38.2	19.1	42.2	54.0	52.7M	198.9G	40.3	35.0	40.9M	304.9G
enriched deformation	dconv@c5	37.6	19.3	41.4	52.6	51.5M	197.7G	39.9	34.9	39.8M	303.7G
	dconv@c4∼c5	39.2	19.9	43.4	55.5	51.7M	198.7G	41.2	36.1	40.0M	304.7G
	dconv@c3∼c5	39.5	21.0	43.5	55.6	51.8M	200.0G	41.5	36.4	40.1M	306.0G
	dconv@c3∼c5 + dpool	40.0	21.1	44.6	56.3	53.0M	201.2G	41.8	36.4	41.3M	307.2G
	mdconv@c3∼c5 + mdpool	40.8	21.3	45.0	58.5	65.5M	214.7G	42.7	37.0	53.8M	320.3G

Table 2. Ablation study on enriched deformation modeling. The input images are of shorter side 800 pixels. Results are reported on the COCO 2017 validation set.

setting	regions to mimic	Faster R-CNN	Mask R-CNN	
		AP^{bbox}	AP^{bbox}	AP^{mask}
mdconv3∼5 + mdpool	None	41.7	43.1	37.3
	FG & BG	42.1	43.4	37.6
	BG Only	41.7	43.3	37.5
	FG Only	43.1	44.3	38.3
regular	None	34.7	36.6	32.2
	FG Only	35.0	36.8	32.3

Table 3. Ablation study on R-CNN feature mimicking. Results are reported on the COCO 2017 validation set.

backbone	method	Faster R-CNN	Mask R-CNN	
		AP^{bbox}	AP^{bbox}	AP^{mask}
ResNet-50	regular	35.1	37.0	32.4
	DCNv1	38.4	40.7	35.5
	DCNv2	43.3	44.5	38.4
ResNet-101	regular	39.2	40.9	35.3
	DCNv1	41.4	42.9	37.1
	DCNv2	44.8	45.8	39.7
ResNext-101	regular	40.1	41.7	36.2
	DCNv1	41.7	43.4	37.7
	DCNv2	45.3	46.7	40.5

Table 4. Results of DCNv2, DCNv1 and regular ConvNets on various backbones on the COCO 2017 test-dev set.