专栏 | MSRA视觉计算组提出第二代可变形卷积网络，增强形变，更好效果

作者：朱锡洲、胡瀚、Stephen Lin（林思德）、代季峰

微软亚洲研究院

2017 年提出的可变形卷积网络自提出以来取得了学界和业界的广泛关注，在目前计算机视觉里重要的 COCO 识别竞赛中，该方法被广泛地采用。近日，来自微软亚洲研究院视觉计算组的研究员提出了第二代可变形卷积网络（Deformable ConvNets v2，简称 DCNv2），新一代的可变形卷积网络通过在网络中应用更多可变形卷积层和引入幅度调制机制，进一步大大增强了网络的形变建模能力。

为了有效地利用这一更强的形变建模能力，研究员们提出了一种利用更精细的驱动力量来引导网络学习的方法，具体来说，考虑到 R-CNN 框架在进行候选框特征提取时能排除无关背景的干扰，在网络训练过程中通过额外引入要求网络特征模仿 R-CNN 特征的损失函数，使得所学习到的形变更专注在前景物体上。通过引入以上更强的建模能力和更优的训练策略，新一代可变形卷积网络在多个主流的识别任务上取得了相比于第一代可变形卷积网络好得多的性能。

以 ResNet-50 基本网络为例，DCNv2 在物体检测的最主要数据集 COCO 上相比于 DCNv1 能带来近 5 个点（mAP）的提升，在 ImageNet 分类上能带来 1.7 个点（top-1 准确率）的提升。在更好的基本网络 ResNeXt-101 上，DCNv2 相比于 DCNv1 在 COCO 物体检测和 ImageNet 分类上依旧能分别带来 3.6 个点以及 1.0 个点的提升。在其它多种识别任务上，DCNv2 也取得了广泛的显著效果。

尺度、姿态、视角的变化和局部形变所导致的几何多样性一直是困扰物体识别和检测的一大难题，为了解决这一难题，MSRA 视觉计算组曾在 17 年提出第一代可变形卷积网络（Deformable ConvNets v1，下称 DCNv1），其包括两个基本模块，可变形卷积层（Deformable Convolution）和可变形兴趣区域池化层（Deformable RoI Pooling）。通过引入这两个模块，卷积神经网络获得了自动适应物体形态变化的特征表达能力，从而大大提升物体检测和分割的精度。

为了理解 DCN，在 DCN 原始的文章里作者们通过在 Pascal VOC 数据集上可视化学习到的卷积采样点和池化位置的分布，发现它们会主要聚集到前景物体区域。然而，研究者们再次仔细地检查这些分布后发现采样点或池化区域往往并不是完全聚集到前景物体区域的，它们常常出现在无关的背景区域，这一现象在更具挑战性的 COCO 更为普遍，甚至常常无法观察到显著的聚集效应。这些现象暗示第一代可变形卷积网络依旧有提升的空间，也激发了研究员们去进一步深入地研究这一问题。

在新一代可变形卷积网络研究过程中，研究员们采用了更好更丰富的工具来深入研究可变形卷积层以及可变形池化层的形变建模能力，具体来说，包括有效感受野（Effective Receptive Field）、有效采样点（Effective Sampling/Bin Locations）和有界误差下的显著性区域等。这些工具能有效地分析网络的空间支持区域（spatial support），利用它们对第一代可变形卷积网络进行全面诊断，DCNv1 存在的问题被进一步验证，也进一步坚定了研究员们去尝试提出更好的可变形卷积网络，即第二代可变形卷积网络（称为 DCNv2）。

这一新的可变形卷积网络主要做了两个方面的改进，包括对网络本身的改进，使其具备更强的形变建模能力，以及一个更好的训练策略来释放这一更强形变建模能力的潜力。

对网络本身的改进使其具备更强的形变建模能力主要包括两点，一是在网络中增加可变形卷积层的使用，和 DCNv1 中仅将其应用到 conv5 的 3 层 3x3 卷积相比，DCNv2 将可变形卷积层应用到 conv3，conv4 和 conv5 的所有 3x3 卷积层。通过引入更多可变形卷积层，DCNv2 能控制更广泛层级特征的采样点，从而使网络整体上具备更精细地学习空间支持区域的能力，这一更强的能力也被前述各种可视化工具所验证（详见图 1 和图 2）。二是在可变形卷积层和可变形兴趣池化层中引入了幅度调制机制，其让

网络学习到每一个采样点偏移量的同时也学习到这一采样点的幅度，从而使得网络在具备学习空间形变能力的同时也具备了区分采样点重要性的能力。

此外，为了更好地挖掘这一更强形变建模能力的潜力，DCNv2 进一步提出了一种更有效的训练机制。受到神经网络中知识蒸馏（Knowledge Distillation）等相关工作的启发，提出了利用 R-CNN 网络作为教师网络（Teacher Network）来更好地指导 DCNv2 中形变参数学习的方法。由于 R-CNN 在裁剪过的图像区域上进行特征提取，其特征不会受到感兴趣区域外的无关信息的影响，通过让 DCNv2 的物体特征模仿 R-CNN 网络提取的对应物体特征，DCNv2 学习到的空间支持区域能更好地聚焦到物体前景区域，从而获得更好的识别和检测效果。

注意到上述改进并没有影响第一代可变形卷积网络的一些优点，新一代可变形卷积网络依旧保持保持了轻量级以及可以很容易地融入到现有网络架构中地特点。研究员们将 DCNv2 广泛地应用到各种识别任务中，特别是重要的物体检测和实例分割问题上，在重要基线（baseline）系统 Faster R-CNN 和 Mask R-CNN 和最主要公开数据集 COCO 上，DCNv2 相比于 DCNv1 取得了显著的精度提升。在其他多种识别任务（包括 ImageNet 图像分类）上 DCNv2 也取得了广泛的性能提升。

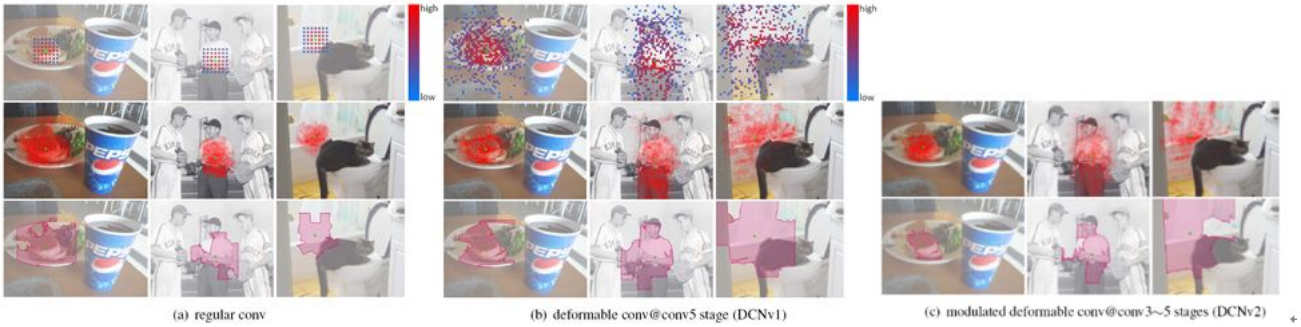


图 1. 常规卷积网络（Regular ConvNets）、DCN v1 和 DCN v2 中基础网络 Stage5 最后一层卷积输出图像特征的空间支持（Spatial Support）可视化效果。常规卷积网络的基准方法是以 ResNet-50 作为基本网络的 Faster R-CNN。在每个子图中，自上而下的每一行分别为有效采样点（Effective Sampling Locations）、有效感受野（Effective Receptive Field）和有界误差下的显著性区域（Error-bounded Saliency Regions）的可视化效果。由于图（c）中的有效采样点和图（b）中的类似，因此图（c）中的有效采样点可视化结果被省略。被可视化的节点（绿点所示）在每张子图中自左到右分别位于小物体（左），大物体（中）和背景（右）上。

图 2. 在检测网络 2fc Node 上的空间支持（Spatial Support）可视化结果。与图 1 相似，可视化的网络包括常规卷积网络（Regular ConvNets）、DCN v1 和 DCN v2，常规卷积网络的基准方法是以 ResNet-50 作为基本网络的 Faster R-CNN。在每个子图中，自上而下的每一行分别为有效采样点（Effective Bin

Locations）、有效感受野（Effective Receptive Field）和有界误差下的显著性区域（Error-bounded Saliency Regions）的可视化效果。（c）--（e）中的有效采样点可视化被省略。输入的感兴趣区域（绿框所示）在每张子图中自左到右分别位于小物体（左），大物体（中）和背景（右）上。

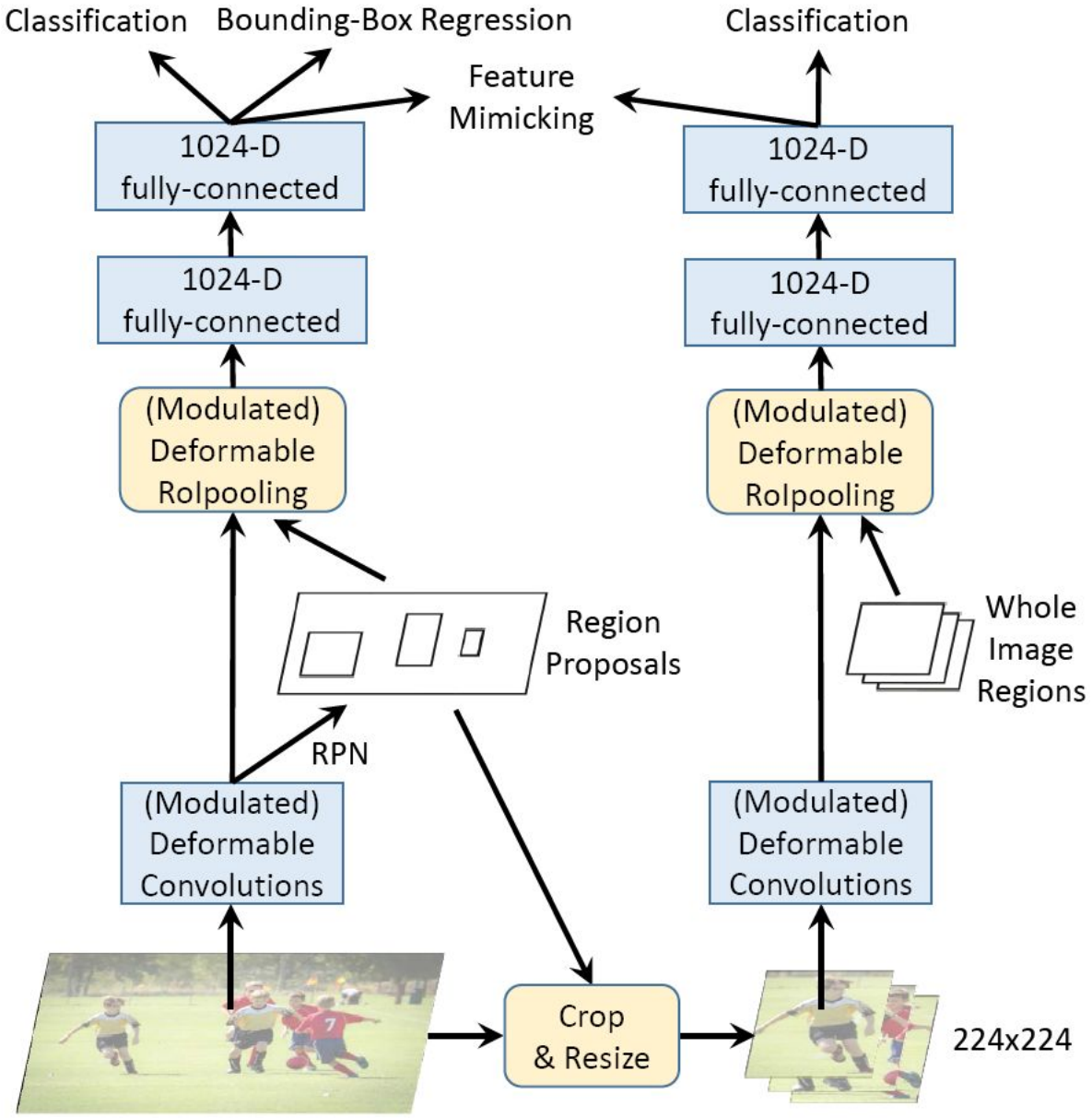


图 3. 引入 R-CNN 特征模仿机制后的网络训练示意图。

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}	para
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
	dconv@c5 + dpool (DCNv1)	38.0	20.7	41.8	52.2	52.7M	328.2G	40.4	35.3	40.9M
enriched deformation	dconv@c5	37.4	20.0	40.9	51.0	51.5M	327.1G	40.2	35.1	39.8M
	dconv@c4~c5	40.0	21.4	43.8	55.3	51.7M	328.6G	41.8	36.8	40.0M
	dconv@c3~c5	40.4	21.6	44.2	56.2	51.8M	330.6G	42.2	37.0	40.1M
	dconv@c3~c5 + dpool	41.0	22.0	45.1	56.6	53.0M	331.8G	42.4	37.0	41.3M
	mdconv@c3~c5 + mdpool	41.7	22.2	45.8	58.7	65.5M	346.2G	43.1	37.3	53.8M

表 1. 采用不同配置增强可变形建模能力的对比实验，输入图片的短边均为 1000 像素（论文中的默认值）。配置（setting）一栏中的「(m)dconv」和「(m)dpool」分别代表（引入幅度调制机制的）可变形卷积层和（引入幅度调制机制的）可变形兴趣区域池化层；「dconv@c3~c5」则代表将可变形卷积层应用在基础网络的 stage3~stage5 上，其他配置同理。实验结果均在 COCO 2017 验证集（Validation Set）上得到。

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}	para
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
	dconv@c5 + dpool (DCNv1)	38.2	19.1	42.2	54.0	52.7M	198.9G	40.3	35.0	40.9M
enriched deformation	dconv@c5	37.6	19.3	41.4	52.6	51.5M	197.7G	39.9	34.9	39.8M
	dconv@c4∼c5	39.2	19.9	43.4	55.5	51.7M	198.7G	41.2	36.1	40.0M
	dconv@c3∼c5	39.5	21.0	43.5	55.6	51.8M	200.0G	41.5	36.4	40.1M
	dconv@c3∼c5 + dpool	40.0	21.1	44.6	56.3	53.0M	201.2G	41.8	36.4	41.3M
	mdconv@c3∼c5 + mdpool	40.8	21.3	45.0	58.5	65.5M	214.7G	42.7	37.0	53.8M

表 2. 采用不同配置增强可变形建模能力的对比试验，输入图片的短边均为 800 像素。实验结果均在 COCO 2017 验证集（Validation Set）上得到。

setting	regions to mimic	Faster R-CNN	Mask R-CNN	
		AP^{bbox}	AP^{bbox}	AP
mdconv3∼5 + mdpool	None	41.7	43.1	37
	FG & BG	42.1	43.4	37
	BG Only	41.7	43.3	37
	FG Only	43.1	44.3	38
regular	None	34.7	36.6	32
	FG Only	35.0	36.8	32

表 3. 基于不同 R-CNN 特征模仿策略的对比实验。模仿区域（regions to mimic）一栏中的「FG」和「BG」分别代表要模仿的是前景（foreground）区域还是背景（background）区域。实验结果均在 COCO 2017 验证集（Validation Set）上得到。

backbone	method	Faster R-CNN	Mask R-CNN	
		AP^{bbox}	AP^{bbox}	AP^{mas}
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

表 4. 使用不同基础网络的常规卷积网络（Regular ConvNets）、DCN v1 和 DCN v2 的对比实验结果。实验结果均在 COCO 2017 测试集（Test-Dev Set）上得到。

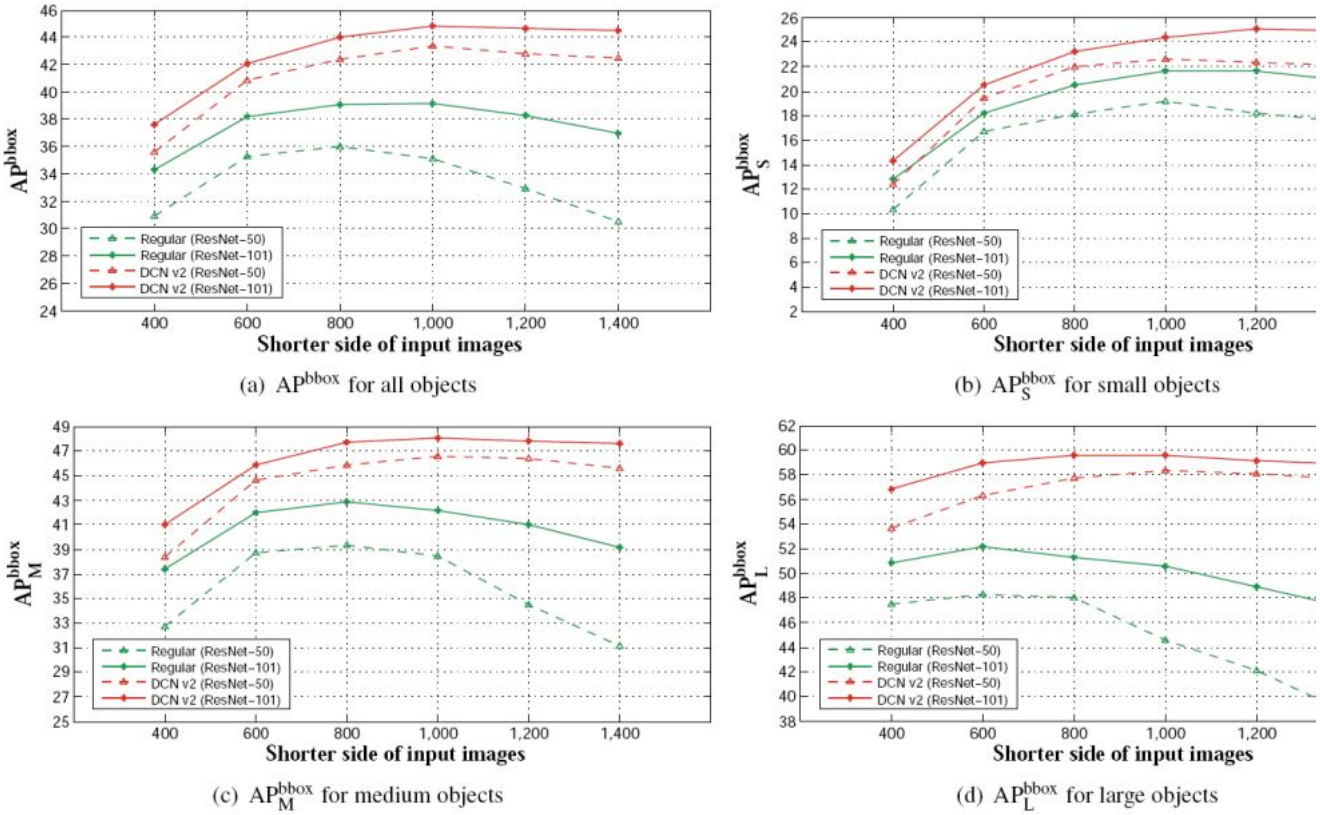
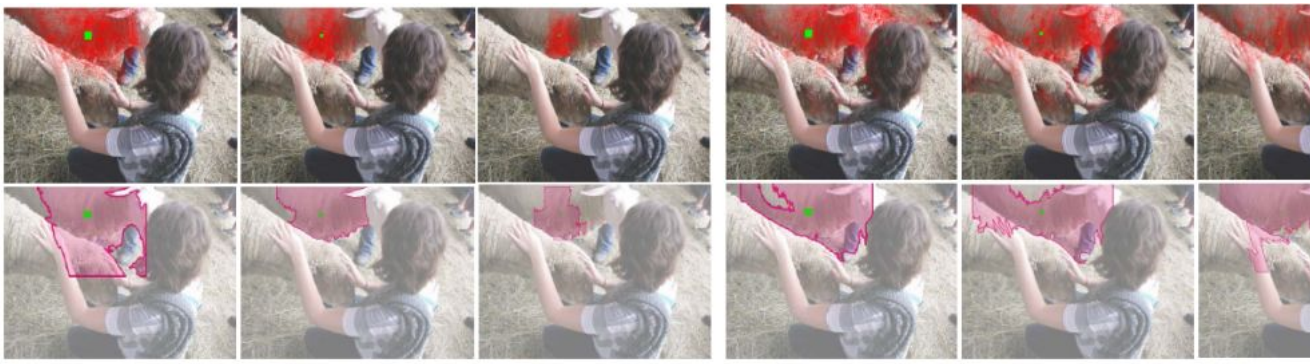


图 4. 常规卷积网络（基于 ResNet-50/ResNet-101 的 Faster R-CNN）和 DCN v2 在 COCO 2017 测试集（Test-Dev Set）上的 APbbox 得分随图像分辨率的变化曲线。



(a) regular conv (b) modulated deformable conv@conv3~5 stages (DC

图 5. 常规卷积网络（Regular ConvNets）和 DCN v2 中基础网络 Stage5 最后一层卷积输出特征图对应的空间支持（Spatial Support）可视化图。在每个子图中，自左向右输入图像的短边长度分别为 400、800 和 1400 像素，自上而下分别为有效感受野（Effective Receptive Field）和有界误差下的显著性区域（Error-bounded Saliency Regions）。

method	setting	Faster R-CNN + ResNet-101				
		AP^{bbox}	AP^{bbox}_{50}	AP^{bbox}_{75}	AP^{bbox}_{S}	AP^{bbox}_{M}
regular	single-scale, shorter side 800	39.1	60.6	42.2	20.5	42.9
	single-scale, shorter side 1000 (best)	39.2	60.6	42.4	21.6	42.2
	multi-scale test	41.2	62.4	45.2	24.6	44.3
DCNv2	single-scale, shorter side 800	44.0	65.9	48.1	23.2	47.7
	single-scale, shorter side 1000 (best)	44.8	66.3	48.8	24.4	48.1
	multi-scale test	46.0	67.9	50.8	27.8	49.1

表 5. 不同分辨率输入图片下的对比实验。实验结果均在 COCO 2017 测试集（Test-Dev Set）上得到。

architecture	method	top-1 acc (%)	top-5 acc (%)	param	FLOP
ResNet-50	regular	76.5	93.1	26.6M	4.1G
	DCNv1	76.6	93.2	26.8M	4.1G
	DCNv2	78.2	94.0	27.4M	4.3G
ResNet-101	regular	78.4	94.2	45.5M	7.8G
	DCNv1	78.4	94.2	45.8M	7.8G
	DCNv2	79.2	94.6	47.4M	8.2G
ResNeXt-101	regular	78.8	94.4	45.1M	8.0G
	DCNv1	78.9	94.4	45.6M	8.0G
	DCNv2	79.8	94.8	49.0M	8.7G

表 6. 常规卷积网络（Regular ConvNets）、DCN v1 和 DCN v2 在 ImageNet 图像分类任务上的准确率。

method	offset&modulation pretraining	VOC det		VOC seg	ImageNet VID det	COCO det
		AP^{bbox}_{50}	AP^{bbox}_{70}	mIoU	AP^{bbox}	AP^{bbox}
regular	none	81.9	68.2	72.0	74.9	39.2
DCNv2	none	83.7	72.4	76.1	79.2	44.8
DCNv2	ImageNet	84.9	73.5	78.3	80.7	44.9

表 7. 在不同的数据集和任务上微调（finetune）在 ImageNet 上预训练的 DCN v2 模型，使用的基础网络为 ResNet-101。

SYNCED

论文链接：https://arxiv.org/abs/1811.11168?from=timeline&isappinstalled=0