# 通过自校准卷积改进卷积神经网络

Jiang-Jiang Liu<sup>1\*</sup>

Oibin  $Hou^{2*}$ Ming-Ming Cheng<sup>1</sup> <sup>1</sup>CS, Nankai University

<sup>2</sup>NUS

Jiashi Feng<sup>2</sup>

https://mmcheng.net/scconv/

## Abstract

目前关于卷积神经网络的大部分改进主要集中 在设计更复杂的网络结构来增强网络的表达能力。 在本文中,我们考虑在不调整模型结构的前提下改 进卷积神经网络的基本卷积特征转换过程。为此,我 们提出了一种全新的自校准卷积, 该卷积通过内部 通信显著扩展了每个卷积层的感受野,从而输出更 丰富的特征表示。特别是,与使用小卷积核 (比如, 3×3) 来融合空间和通道信息的标准卷积不同,我 们的自校准卷积通过全新的自校准操作,在每个空 间位置周围自适应地建立了远程空间和通道间依赖 性。因此, 它可以通过显式地合并更丰富的信息来 帮助卷积神经网络生成更具有区分度的特征。我们 的自校准卷积设计简单且通用,可以在不引入额外 参数和复杂性的前提下轻松地增强标准卷积层。大 量实验表明,将我们的自校准卷积应用于不同的主 干网络时,在不改变网络结构的前提下,可以在各 种视觉任务(包括图像识别,对象检测,实例分割和 关键点检测) 中显着提升基准模型表现。我们希望这 项工作可以为将来的研究提供一种设计全新的卷积 特征变换以改善卷积网络的有价值的方法。代码在 项目页面已开源。

# 1. 引言

使用大规模图像分类数据集 (如 ImageNet [30]) 训练的深度神经网络通常被用作主干网络来提取



Changhu Wang<sup>3</sup>

<sup>3</sup>BvteDance AI Lab

图 1. 使用不同的 Grad-CAM [31]习得的特征热点图的可视 化结果。所有的网络均在 ImageNet [30]训练得到。我们的结 果通过将提出的自校准卷积应用于 ResNet-50 获得。从热点 图可见,受限于他们卷积层的感受野,使用传统(分组)卷积 甚至 SE 模块 [16]的残差网络 [12, 40]都不能捕捉整个具有区 分度的区域。相反, 自校准卷积有助于我们的模型很好地捕 获整个具有区分度的区域。

具有较强代表能力的特征用于下游任务,如目标检 测 [23, 29, 2, 8], 分割 [45, 11]和人体关键点检测 [11, 39]。一个好的分类网络通常具有很强的特征转 换能力,因此可以提供较强的特征特征用于下游任 务 [20, 10, 27]。因此,非常需要增强卷积网络的特 征变换能力。

在文献中, 生成丰富特征表示的一种有效方法 是使用功能强大的人为设计的网络体系结构,如残 差网络 (ResNets) [12] 和它的各种变体网络 [40, 43, 34, 7]或基于 AutoML 技术 [47, 26]设计网络。最近,

<sup>\*</sup>作者贡献相等。

一些方法尝试通过将注意力机制 [38, 48, 16, 15]或 non-local 模块 [37, 3]引入成熟的网络中建模空间位 置间或通道间的依赖关系。上述方法背后的共同思 想集中在调整网络体系结构以产生丰富的特征表示, 这需要太多的人力。

在本文中,我们没有设计复杂的网络体系结构 来增强特征表示能力,而是引入了自核准卷积作为 一种有效的方法,通过增强每层的基本卷积变换来 帮助卷积网络学习有区分度的特征表示。类似于分 组卷积,它将特定层的卷积过滤器不均匀地分为多 个部分,每个部分中的过滤器以异构方式被利用。具 体地说,自校正卷积不是通过均匀地对原始空间中 的输入进行所有卷积操作,而是首先通过下采样将 输入转换为低维嵌入张量。通过一组卷积滤波器获 得的低维嵌入张量被用来校准另一个卷积滤波器组 的卷积变换。得益于这种异构卷积和滤波器间通信, 每个空间位置的感受野可以有效地扩大。

作为标准卷积的增强版本,我们的自校准卷积 具有两个优点。首先,它使每个空间位置都能自适 应地编码长程上下文信息,从而打破了在小区域内 (比如 3 × 3)进行卷积的传统。这使我们的自校准卷 积产生的特征表示更具区分度。在图 1中,我们可视 化了基于不用卷积 [12,40]的 ResNets 产生的特征热 点图。可以看出,使用自校准卷积的 ResNet 可以更 准确、整体地定位目标物体。其次,所提出的自校准 卷积是通用的,可以轻松应用于标准卷积层,而无 需引入任何参数和复杂度或更改超参数。

为了证明所提出的自校准卷积的有效性,我们 首先将其应用于大规模图像分类问题。我们以残差 网络 [12]及其变体 [40,16]为基准模型,在具有相当 的模型参数和计算能力的情况下,top-1 准确性得到 了很大的提高。除了图像分类,我们还进行了广泛 的实验,以证明所提出的自校准卷积在多种视觉应 用中的泛化能力,包括目标检测,实例分割和关键 点检测。实验表明,通过所提出的自校准卷积,以上 三个任务的基准结果都可以得到较大改善。

## 2. 相关工作

在本节中,我们简要回顾卷积网络的体系结构 设计和长程依赖关系构建方面的最新代表性工作。

结构设计:近年来,在全新结构设计领域已经取得 了显着的进步 [33, 35, 32, 44]。作为早期工作, VG-GNet [33]使用比 AlexNet [19]更小的卷积核 (3×3) 构建更深的网络,从而在使用更少参数的情况下获 得更好的性能。ResNets [12, 13]通过引入残差连接 并使用批正则化 [18]来改进顺序结构, 从而可以构建 非常深的网络。ResNeXt [40]和 Wide ResNet [43]通 过对 3 × 3 卷积层进行分组或增加其宽度来扩展 ResNet。GoogLeNet [35]和 Inceptions [36, 34]利用 精心设计的由一组具有多条并行路径的特殊卷积滤 波器 (3×3等) 组成的 Inception 模块来进行特征变 换。NASNet [48]通过探索预定义的搜索空间来学习 构建模型架构,从而实现可移植性。DenseNet [17]和 DLA [42]通过复杂的自上而下的跳跃连接来聚合特 征。Dual Path Networks (DPNs) [7]利用残差连接 和稠密连接来构建强大的特征表示。SENet [16]引入 squeeze-and-excitation 操作来显式地建模通道间的 依赖。

**长程依赖建模**: 建立长程依赖对大多数计算机视觉 任务很有帮助。SENet [16]是成功的例子之一, 它采 用 Squeeze-and-Excitation 模块在通道间建立相互 依赖关系。后续工作, 如 GENet [15]、CBAM [38]、 GCNet [3]、GALA [25]、AA [1]和 NLNet [37]通过 引入空间注意力机制或设计更高级的注意力模块进 一步扩展了这一思想。建模长程依赖关系的另一种 方法是利用具有大卷积核窗口的空间池化或卷积运 算。一些经典的例子, 如 PSPNet [45]采用多个不同 大小的空间池化运算来捕捉多尺度上下文。还有许 多工作 [28, 14, 41, 5, 22]利用大卷积核或空洞卷积 进行长程上下文聚合。我们的工作也不同于 Octave convolution [6], 后者旨在减少空间冗余和降低计算 成本。

与上述所有侧重于调整网络体系结构或添加其 他人为设计的模块以改善卷积网络的方法不同,我 们的方法考虑更有效地利用卷积层中的卷积滤波器, 并设计功能强大的特征转换以生成更具表达力的特征表示。

## 3. 方法

传统的 2D 卷积层  $\mathcal{F}$  由一组卷积滤波器集合 K = [k<sub>1</sub>, k<sub>2</sub>,..., k<sub> $\hat{C}$ </sub>] 构成,其中 k<sub>i</sub> 代表第 i 个尺寸 为 C 的卷积滤波器,它将输入 X = [x<sub>1</sub>, x<sub>2</sub>,..., x<sub>C</sub>]  $\in$  $\mathbb{R}^{C \times H \times W}$  变换为输出 Y = [y<sub>1</sub>, y<sub>2</sub>,..., y<sub> $\hat{C}$ </sub>]  $\in$  $\mathbb{R}^{\hat{C} \times \hat{H} \times \hat{W}}$ 。注意,为了符号上的便捷,我们省略了卷 积滤波器的空间大小和偏差项。给定以上标记,输出 的特征图的通道 i 可表示为

$$\mathbf{y}_i = \mathbf{k}_i * \mathbf{X} = \sum_{j=1}^C \mathbf{k}_i^j * \mathbf{x}_j, \tag{1}$$

其中 '\*' 代表卷积操作并且  $k_i = [k_i^1, k_i^2, ..., k_i^C]$ 。从 上面可以看出,每个输出特征图都是通过所有通道 求和得到的,并且都是通过多次重复方程 1统一生 成的。通过这种方式,卷积滤波器可以学习到相似 的模式。而且,卷积特征变换中每个空间位置的感 受野主要由预定义的卷积核大小控制,并且由这种 卷积层堆叠成的网络也缺少大的感受野来捕获组足 够多的高级语义信息 [46,45]。以上两个缺点都可能 导致特征图的区分度较低。为了缓解上述问题,我 们提出了自校准卷积,下面进行详细介绍。

## 3.1. 自校准卷积

在分组卷积中,特征转换过程在多个并行分支 中均匀且独立地执行,然后所有分支的输出连接起 来作为最终输出。与分组卷积相似,我们提出的自校 准卷积也将可学习的卷积滤波器分为多个部分,不 同的是,每个部分的滤波器并不相同而各自有特殊 功能。

#### 3.1.1 概述

我们提出的结构的工作流程如图 2所示。在我 们的方法中,我们考虑一个简单的情况,其中输入的 通道数 *C* 和输出的通道数  $\hat{C}$  相同, 即  $\hat{C} = C$ 。因 此,在下文中,为了符号上的方便我们使用 *C* 代替  $\hat{C}$ 。给定一组形状为 (*C*,*C*,*k<sub>h</sub>*,*k<sub>w</sub>*) 的卷积滤波器组 K,其中  $k_h$ 和  $k_w$ 分别是空间的高度和宽度,我们首 先将其均分为四个部分,每个部分负责不同的功能。 在不失一般性的前提下,假设 C 能被 2 整除。分组 后,我们有四个由卷积滤波器组组成的部分,符号为 {K<sub>i</sub>} $_{i=1}^4$ ,每组滤波器的形状分别为( $\frac{C}{2}, \frac{C}{2}, k_h, k_w$ )。

给定四部分滤波器后,我们将输入 X 均匀地分 为两个部分 {X<sub>1</sub>,X<sub>2</sub>},两部分分别被送入用于收集 不同类型的上下文信息的特殊路径中。在第一条路 径中,我们使用 {K<sub>1</sub>,K<sub>2</sub>,K<sub>3</sub>} 对 X<sub>1</sub> 进行自校准操 作,得到输出 Y<sub>1</sub>。在第二条路径中,我们执行一个 简单的卷积操作: Y<sub>2</sub> =  $\mathcal{F}_1(X_2) = X_2 * K_1$ ,其目的 是保留原始的上下文信息。两个中间结果 {Y<sub>1</sub>,Y<sub>2</sub>} 连接在一起作为输出 Y。在下面的内容中,我们将 详细介绍如何在第一条通路中执行自校准操作。

## 3.1.2 自校准

为了高效地收集每个空间位置的上下文信息, 我们提出在两个不同的尺度空间中进行卷积特征转 换:一个是原始比例空间,其中特征图与输入的分 辨率相同,另一个是下采样后较小的隐空间。经过 变换后进入较小的隐空间的嵌入向量有大的感受野, 因此可以作为参考用来指引原始特征空间中的特征 变换过程。

**自校准**: 给定输入 X<sub>1</sub>, 我们采用尺寸为 *r*×*r*、步长 为 *r* 的平均池化, 如下所示:

$$T_1 = \operatorname{AvgPool}_r(X_1).$$
(2)

基于 K<sub>2</sub> 对 T<sub>1</sub> 进行如下特征变换:

$$\mathbf{X}_1' = \mathrm{Up}(\mathcal{F}_2(\mathbf{T}_1)) = \mathrm{Up}(\mathbf{T}_1 * \mathbf{K}_2), \qquad (3)$$

其中 Up(·) 是双线性插值运算符, 它将中间参考值 从小比例的隐空间映射到原始特征空间。至此, 校 准操作可以表述为:

$$Y_1' = \mathcal{F}_3(X_1) \cdot \sigma(X_1 + X_1'), \qquad (4)$$

其中  $\mathcal{F}_3(X_1) = X_1 * K_3$ ,  $\sigma$  是 sigmoid 函数, '·'代 表逐元素相乘。如式 4所示,我们发现使用  $X'_1$  作为 残差来生成校准的权重是有用的。校准后的最终输



图 2. 自校准卷积的示意图。可以看出,在自校准卷积中,原始滤波器分为四个部分,每个部分负责不同的功能。这使得自校准 卷积与传统卷积或称分组卷积以同质的方式执行卷积不同。自校准卷积的更多细节见小节 3.1。



图 3. 由 ResNet-50 的不同设置生成的中间特征图的视觉效 果对比。特征图选自最后一个构建模块的 3×3 卷积层。最上 面的行,我们使用传统的卷积;最下面的行,我们使用自校 准卷积 (SC-Conv)。很显然,使用自校准卷积的 ResNet-50 能捕捉更丰富的上下文信息。

出可以写成如下形式:

$$Y_1 = \mathcal{F}_4(Y_1') = Y_1' * K_4.$$
 (5)

优势:我们提出的自校准操作有三方面优势。首先, 与传统卷积相比,通过引入公式4,每个空间位置都 不仅可以自适应地将其周围的信息环境视为来自潜 在空间的嵌入(在原始尺度空间的响应中用作标量), 还可以对通道间相关性进行建模。因此,使用自校准 的卷积层的感受野可以有效地扩大。如图3所示,使 用自校准的卷积层能编码更大且有更有区分度的区 域。第二,自校准操作不收集全局上下文信息,而仅 考虑每个空间位置周围的上下文信息,在一定程度 上避免了来自不相关信区域的无关信息。从图6的 右两栏可以看出,可视化最终的评分层时,具有自 校准的卷积可以准确地定位目标物体。第三,自校 准操作可编码多尺度信息,这是与目标检测相关的 任务非常需要的。我们将在小节4给出更多分析。

#### 3.2. 实例化

为了验证所提出的自校准卷积的性能,我们使 用残差网络的几种变体 [12,40,16]作为示例。同时 考虑了 50 层和 101 层的 bottleneck 结构。为了简 便起见,我们仅使用自校准卷积替换每个构建模块 中的 3 × 3 卷积运算,并保持所有相关超参数不变。 默认情况下,自校准卷积中的下采样率设置为 4。

**与分组卷积的关系**:分组卷积采用了拆分-变换-合并 的策略,其中在多个并行分支中均匀地 [40]或以分 层方式 [9]进行各自的卷积变换。与分组卷积不同, 我们的自校准卷积能够以异构的方式利用卷积滤波 器的不同部分。因此,转换过程中每个空间位置可 以通过自校准操作来融合两个不同尺寸比例空间中 的信息,当应用于卷积层时,这会大大增加感受野, 并因此产生更具有区分度的特征表示。

与基于注意力的模块的关系:我们的工作也与现有 的基于附加注意力模块 (如 SE block [16]、GE [15]或 CBAM [38])的方法不同。这些方法需要额外的可学 习参数,而我们的自校准卷积改变了卷积层内部使 用卷积滤波器的方式,因此不需要额外的可学习的 参数。此外,尽管 GE block [15]像我们一样在较低维 度的空间中对空间信息编码,但它并没有显式地保 留原始比例空间中的空间信息。在下面的实验部分 中,我们将展示在没有任何其他可学习参数的情况 下,我们的自校准卷积可以在图像分类的基准模型 和其他基于注意力机制的模型上取得显著提升。此 外,我们的自校准模块是对注意力模块的补充,因 此可以从附加的注意力模块中受益。

## 4. 实验

## 4.1. 实现细节

我们使用 PyTorch 框架<sup>1</sup>实现我们的方法。为 了公平的比较,除非特别声明,否则我们均采用 官方分类框架来进行所以分类实验。我们在 ImageNet [30]上进行测试。输入图像的尺寸为 224×224, 按照 [40]的方法在调整大小后的图像中随机裁剪获 得。我们使用 SGD 优化所有模型。权重衰减和动 量分别设置为 0.0001 和 0.9。使用四个 Tesle V100 GPU,最小批处理数量设置为 256(每个 GPU 64 个) 默认情况下,我们所有模型训练 100 轮,初始学习 率为 0.1,每 30 轮学习率除以 10。测试阶段,我们 按照 [40]的做法将图像短边缩放为 256 后在中央裁 剪出 224×224 的图像来测试准确率。除网络结构本 身外,所有消融比较中的模型都采用相同的运行环 境和超参数。表 1中的所有模型都在相同的策略下 训练并在相同的设定下测试。

## 4.2. 在 ImageNet 上的结果

我们进行消融实验以验证我们提出的体系结构 中每个组件的重要性,并在 ImageNet-1K 分类数据 集 [30]上与现有的基于注意力的方法进行比较。

## 4.2.1 消融分析

**泛化能力**:为了验证所提出结构的泛化能力,我们 采用三种广泛使用的分类结构作为基准模型,包 括 ResNet [12]、ResNeXt [40]和 SE-ResNet [16]。 与之相对应的使用自校准卷积的模型分别命名为 SCNet、SCNeXt 和 SE-SCNet。按照默认版本的 ResNeXt [40] (32 × 4d),我们设置 SCNeXt 的 bottleneck 宽度为 4。我们还调整我们结构中每组卷积 的基数来保证 SCNeXt 的参数量与 ResNeXt 接近。 对于 SE-SCNet,我们按照 [16]的方式将 SE 模块应 用于 SCNet。 表 1. 自校准卷积应用于不同分类框架时在 ImageNet-1K 上的比较。我们展示单边裁剪准确率 (%)。

在表 1中,我们展示了每个模型 50 层和 101 层版本获得的结果。与原始的 ResNet-50 架构 相比,SCNet-50 的准确率提高了 1.4%(77.8% vs. 76.4%)。此外,SCNet-50 获得的提升 (1.4%)也高 于 ResNeXt-50(1.0%)和 SE-ResNet-50 (0.8%)。这 表明自校准卷积比增加基数或引入 SE 模块 [16]要 好得多。当网络加深时,也可以观察到类似的现象。

研究所提出结构的泛化能力的另一种方式是观 察它在其他视觉任务基干中的表现,如目标检测和 实例分割。我们将在下一小节中给出更多实验比较。

自校准卷积 v.s. 原始卷积:为了进一步研究所提出的自校准卷积相比于原始卷积的有效性,我们按照 [21]的做法为两个网络在其中一个中间层 (即res3)后添加监督 (辅助损失)。侧面输出的结果可以反映深度变化时网络的性能以及不同层的特征表示的强度。图 5 描绘了在 res3 处的侧面监督得到的top-1 准确率结果。显然,SCNet-50 的侧边结果比ResNet-50 的侧边结果好得多。这一现象间接表明了与原始卷积相比,使用自校准卷积的网络能生成更丰富更有区分度的特征表示。为了进一步验证,我

Params MAdds FLOPs | Top-1 Top-5 Network 50-layer ResNet [12] 25.6M4.1G8.2G76.493.0SCNet 25.6M4.0G7.9G77.8 93.9ResNeXt [40] 25.0M4.3G 77.48.5G93.4ResNeXt 2x40d 25.4M4.2G 8.3G76.893.3 SCNeXt 25.0M4.3G8.5G78.394.0 SE-ResNet[16] 28.1M4.1G8.2G 77.293.4SE-SCNet 28.1M4.0G7.9G78.293.9 101-laver ResNet [12] 44.5M7.8G15.7G78.093.9SCNet 44.6M7.2G78.914.4G94.3ResNeXt [40] 44.2M 8.0G16.0G78.594.2SCNeXt 44.2M8.0G15.9G79.294.478.4SE-ResNet[16] 49.3M7.9G15.7G94.2SE-SCNet 78.9 49.3M7.2G14.4G94.3

 $<sup>^{1} \</sup>rm https://pytorch.org$ 



ResNet (res3) SCNet (res3) ResNet (res3) SCNet (res3) 图 4. 不同网络的 res3 侧边输出特征图的可视化 (ResNet v.s. SCNet)。我们将两个网络都设为 50 层。



图 5. ResNet-50 和 SCNet-50 的辅助损失曲线。我们在 res3 后添加辅助损失。可以看到, SCNet(红线) 比 ResNet(青色线)效果好得多。这表明深度较低时自校准卷积对网络效果提升更强。

们在图 4展示了一些评分层输的侧边输出的可视化 结果。显然, SCNet 可以在网络深度较低时更精确、 整体地定位目标物体。在小节 4.3中, 我们将给出两 种卷积在更多不同的视觉任务上的结果。

**注意力机制比较**:为了探究自校准卷积对于分类网络有帮助的原因,我们使用 Grad-CAM [31]作为注意力提取工具来可视化 ResNet-50、ResNeXt-50 和 SE-ResNet-50 产生的注意力图片,如图 6所示。可以清楚地看到,SCNet-50 产生的注意力图像可以更准确地定位目标物体而不会扩展到太多背景区域。当目标对象较小时,与其他三个网络产生的语义区域相比,我们的网络的注意力也更好地定位于语义区



图 6. Grad-CAM[31]产生的注意力图的可视化结果。显然, 无论目标物体多大或形状多不规则,我们的 SCNet 能比其 他网络更准确地定位前景物体。这极大地依赖于我们的自校 准操作,它有利于自适应地捕捉丰富的语义信息。我们将所 有网络设定为 50 层。

域。这表明我们的自校准卷积有助于发现更完整的 目标对象,即使它们的尺寸很小。

设计选择:如小节 3.1所示,我们引入下采样来实现 自校准,并证明了这对改进卷积神经网络有帮助。在 此,我们研究自校准卷积中的下采样率对分类性能 的影响。在表 2中,我们展示了在自校准卷积中使用 不同下采样率时的性能。可以看到,当不采用下采 样操作时 (r = 1),结果已经比原始的 ResNet-50 更 好 (77.38% v.s. 76.40%)。随着下采样率增加,可以 得到更好的性能。特别地,当下采样率为 4 时,我 们得到了 77.81% 的 top-1 准确率。注意,由于最后 残差模块的分辨率已经非常小了 (比如 7 × 7),我们 没有使用更大的下采样率。此外,我们发现通过添 加如图 2所示的 identity 连接 (在  $F_2$  之后) 以低分 辨率的特征图作为残差,也有助于提高性能。不加 额外的 identity 连接将导致性能降至 77.48%。

**平均池化** vs. 最大池化:除了上述设计选择之外,我 们还研究了不同池化类型对性能的影响。在我们的 实验中,我们尝试用最大池化替换自校准卷积中的 所有平均池化,并观察性能差异。如表 2所示,在

| Model                                | DS Rate $(r)$ | Identity | Pooling | Top-1 Accuracy | Network                      | Params            | MAdds | Top-1 | Top-5 |
|--------------------------------------|---------------|----------|---------|----------------|------------------------------|-------------------|-------|-------|-------|
| ResNet                               | -             | -        | _       | 76.40%         | ResNet [12]                  | $25.6 \mathrm{M}$ | 4.1G  | 76.4  | 93.0  |
| ResNeXt                              | -             | -        | -       | 77.40%         | ResNeXt $[40]$               | 25.0M             | 4.3G  | 77.4  | 93.4  |
| SE-BesNet                            | _             | _        | AVG     | 77.20%         | SE-ResNet [16]               | 28.1M             | 4.1G  | 77.2  | 93.4  |
| SCNet                                | 1             | 1        | ma      | 77 2907        | ResNet + CBAM [38]           | 28.1M             | 4.1G  | 77.3  | 93.6  |
| SONEL                                | 1             | v        | -       |                | GCNet [3]                    | $28.1 \mathrm{M}$ | 4.1G  | 77.7  | 93.7  |
| SCNet                                | 2             | 1        | AVG     | 77.48%         | ResNet + GALA [25]           | $29.4 \mathrm{M}$ | 4.1G  | 77.3  | 93.6  |
| SCNet                                | 4             | ×        | AVG     | 77.48%         | ResNet + AA [1]              | 28.1M             | 4.1G  | 77.7  | 93.6  |
| SCNet                                | 4             | 1        | MAX     | 77.53%         | ResNet + GE $[15]^{\dagger}$ | 31.2M             | 4.1G  | 78.0  | 93.6  |
| SCNet                                | 4             | 1        | AVG     | 77.81%         | SCNet                        | 25.6M             | 4.0G  | 77.8  | 93.9  |
| SCNeXt                               | 4             | 1        | AVG     | 78.30%         | $\mathrm{SCNet}^\dagger$     | $25.6 \mathrm{M}$ | 4.0G  | 78.2  | 94.0  |
| 素 2 SCNet 不同设计选择的消融实验 'Identity' 代之与 |               |          |         | SE-SCNet       | 28.1M                        | 4.0G              | 78.2  | 93.9  |       |

表 2. SCNet 不同设计选择的消融实验。'Identity'代之与 表 2中名称相同的对应组成成分。'DS Rate'是公式 2中的下 采样率。我们还展示了不同池化操作的结果:平均池化 (AVG) 和最大池化 (MAX)。

所有其他配置不变的情况下,使用最大池化会导致 top-1 准确率下降大约 0.3%(77.81 vs. 77.53)。我们 认为,这可能是由于以下原因造成:与最大池化不 同,平均池化在整个池化窗口内的位置之间建立了 连接,从而可以更好地捕获上下文信息。

**讨论**:根据上述消融实验,引入自校准卷积对于 ResNet和 ResNeXt 这样的分类网络很有帮助。但 是,应注意,探索最优架构设置超出了本文的范围。 本文仅提供有关如何改善原始卷积的初步研究。我 们建议读者进一步研究更有效的结构。在下一小节 中,我们将展示我们的方法应用于主流的视觉任务 时,作为预训练的主干网络表现如何。

## 4.2.2 基于注意力机制的方法比较

在此,我们将 SCNet 应用于 ResNet-50 结构 上,与现有的基于注意力机制的方法进行对比,包 括 CBAM [38]、SENet [16]、GALA [25]、AA [1]和 GE [15] 比较结果见表 3。显而易见,大多数基于注 意力或 non-local 的方法都需要额外的可学习的参数 来构建其相应的模块,然后将其应用于模型构建。我 们的方法完全不同,不依赖任何额外的可学习参数, 而只是异构地利用卷积滤波器。结果显然比其他方 法好。还应该提到的是,我们的自校准卷积也与上

表 3. 在 ImageNet-1K 数据集上与前述基于注意力机制的 方法的比较。所有方法都以 ResNet-50 为基准模型。我们展 示了单边裁剪准确率(%)并比较了复杂度。'†' 表示模型训 练了 300 轮。

31.1M

4.0G

78.3

94.0

述基于注意力机制的方法兼容。例如,当按照 [15]的 方法将 GE 模块添加到 SCNet 每个构建模块时,我 们可以将准确率进一步提高 0.5%。这也表明我们的 方法不同于其他附加模块。

#### 4.3. 应用

GE-SCNet

在本小节中,我们通过将 SCNet 作为基准网络 应用于主流的视觉任务,包括目标检测,实例分割 和人体关键点检测,来研究该方法的泛化能力。

#### 4.3.1 目标检测

网络设置:在目标检测任务中,我们采用广泛使用的 Faster R-CNN 结构 [29]和 feature pyramid networks (FPNs) [23]作为主干网络。我们采用广泛使 用的 mmdetection<sup>2</sup>来进行所有实验。按照以前的工 作 [23, 11],我们使用 80k COCO 训练图像和验证 集 (trainval35k) [24]中的 35k 图像的并集来训练模 型,并在其余 5k 验证集 (minival)上进行测试。

我们严格按照 Faster R-CNN [29]及其 FPN 版本 [23]设置超参数。重新调整了所有图像尺寸,以

 $<sup>^{2} \</sup>rm https://github.com/open-mmlab/mmdetection$ 

使其短边都为 800 像素。我们使用 8 块 Tesla V100 GPU 来训练每个模型,并且 mini-batch 设置为 16, 即每个 GPU 上 2 张图像。初始学习率设置为 0.02, 并使用 2× training schedule 来训练每个模型。权重 衰减和动量分别设置为 0.0001 和 0.9。我们使用标准 COCO 指标测试结果,包括 AP(不同 IoU 阈值的平 均准确率)、AP<sub>0.5</sub>、AP<sub>0.75</sub> 和 AP<sub>S</sub>、AP<sub>M</sub>、AP<sub>L</sub>(不 同尺度的 AP)。采用 50 层和 101 层的主干。

**检测结果**: 在表 4页部,我们展示了使用不同分类 主干网络进行目标检测的实验结果。以 Faster R-CNN [29]为例,采用 ResNet-50-FPN 主干可得到 37.6 的 AP 分数,而用 SCNet-50 代替 ResNet-50 则可显著提高 3.2(40.8 v.s. 37.6)。更有趣的是,以 SCNet-50 作为主干的 Faster R-CNN 的性能甚至 比以 ResNeXt-50 作为主干的 Faster R-CNN 更好 (40.8 v.s. 38.2)。这表明提出的利用卷积滤波器的 方法比直接对卷积滤波器分类有效得多。这可能是 因为自校准卷积包含自适应的响应校准操作,这有 助于更准确地定位目标对象的确切位置,如图 6所 示。此外,从表 4中可以看出,使用更深的主干会 得到与上述类似的现象 (ResNet-101-FPN: 39.9 → SCNet-101-FPN: 42.0)。

## 4.3.2 实例分割

对于实例分割,我们使用与 Mask R-CNN [11]相同的超参数和数据集进行公平比较。的实验结果都 是基于 mmdetection 框架 [4]获得的。

表 4底部为 SCNet 版本的 Mask R-CNN 和 ResNet 版本的 Mask R-CNN 对比。由于我们已 经详细介绍了目标检测结果,所以在此我们仅使用 mask APs 展示结果。可以看出,ResNet-50-FPN 版本和 ResNeXt-50-FPN 版本的 Mask R-CNN 的 Mask APs 值分别为 35.0 和 35.5。然而,测试 SCNet 时,Mask AP 值分别提升 2.2 和 2.0。使用更深层的 主干网络时,也可以观察到类似结果。这表明我们 的自校准卷积也有助于实例分割。

| Backbone                           | AP   | $AP_{0.5}$ | $\operatorname{AP}_{0.75}$ | $AP_S$ | $\mathrm{AP}_M$ | $\mathrm{AP}_L$ |  |
|------------------------------------|------|------------|----------------------------|--------|-----------------|-----------------|--|
| Object Detection (Faster R-CNN)    |      |            |                            |        |                 |                 |  |
| ResNet-50-FPN                      | 37.6 | 59.4       | 40.4                       | 21.9   | 41.2            | 48.4            |  |
| SCNet-50-FPN                       | 40.8 | 62.7       | 44.5                       | 24.4   | 44.8            | 53.1            |  |
| ResNeXt-50-FPN                     | 38.2 | 60.1       | 41.4                       | 22.2   | 41.7            | 49.2            |  |
| SCNeXt-50-FPN                      | 40.4 | 62.8       | 43.7                       | 23.4   | 43.5            | 52.8            |  |
| ResNet-101-FPN                     | 39.9 | 61.2       | 43.5                       | 23.5   | 43.9            | 51.7            |  |
| SCNet-101-FPN                      | 42.0 | 63.7       | 45.5                       | 24.4   | 46.3            | 54.6            |  |
| ResNeXt-101-FPN                    | 40.5 | 62.1       | 44.2                       | 23.2   | 44.4            | 52.9            |  |
| SCNeXt-101-FPN                     | 42.0 | 64.1       | 45.7                       | 25.5   | 46.1            | 54.2            |  |
| Instance Segmentation (Mask R-CNN) |      |            |                            |        |                 |                 |  |
| ResNet-50-FPN                      | 35.0 | 56.5       | 37.4                       | 18.3   | 38.2            | 48.3            |  |
| SCNet-50-FPN                       | 37.2 | 59.9       | 39.5                       | 17.8   | 40.3            | 54.2            |  |
| ResNeXt-50-FPN                     | 35.5 | 57.6       | 37.6                       | 18.6   | 38.7            | 48.7            |  |
| SCNeXt-50-FPN                      | 37.5 | 60.3       | 40.0                       | 18.2   | 40.5            | 55.0            |  |
| ResNet-101-FPN                     | 36.7 | 58.6       | 39.3                       | 19.3   | 40.3            | 50.9            |  |
| SCNet-101-FPN                      | 38.4 | 61.0       | 41.0                       | 18.2   | 41.6            | 56.6            |  |
| ResNeXt-101-FPN                    | 37.3 | 59.5       | 39.8                       | 19.9   | 40.6            | 51.2            |  |
| SCNeXt-101-FPN                     | 38.2 | 61.2       | 40.8                       | 18.8   | 41.4            | 56.1            |  |

表 4. 在 COCO minival 数据集上与最先进方法的比较。所 有结果都使用相同超参数并使用单模型获得。对于目标检测, AP 指 box IoU,对实例分割 AP 指 mask IoU。

#### 4.3.3 关键点检测

最后,我们将 SCNet 应用于人体关键点检测并 在 COCO 关键点检测数据集 [24]上测试结果。我们 使用最先进的方法 [39]作为基准。我们仅用 SCNet 替换 [39]中的主干网络 ResNet,所有其他训练和测 试设置<sup>3</sup>保持不变。我们使用基于 OKS 的标准 mAP 在 COCO val2017 数据集上评估结果,其中 OKS(物 体关键点相似度) 定义了不同人体姿势之间的相似 度。如 [39]的做法,在测试阶段采用了 Faster R-CNN 目标检测器 [29],该检测器在 COCO val2017 数据 集上对'人'类别的 AP 检测值为 56.4。

表 5为对比结果。可以看到, 简单地用 SCNet-50 替换 ResNet-50 可以将尺寸为 256 × 192 的输入的 AP 值提高 1.5%, 对尺寸为 384 × 288 的输入的 AP

<sup>&</sup>lt;sup>3</sup>https://github.com/Microsoft/human-pose-estimation.pytorch

| Backbone                    | Scale            | AP   | $\mathrm{AP}_{.5}$ | $\mathrm{AP}_{.75}$ | $AP_m$ | $\mathrm{AP}_l$ |
|-----------------------------|------------------|------|--------------------|---------------------|--------|-----------------|
| ResNet-50                   | $256\times192$   | 70.6 | 88.9               | 78.2                | 67.2   | 77.4            |
| SCNet-50                    | $256\times192$   | 72.1 | 89.4               | 79.8                | 69.0   | 78.7            |
| ResNet-50                   | $384 \times 288$ | 71.9 | 89.2               | 78.6                | 67.7   | 79.6            |
| SCNet-50                    | $384 \times 288$ | 74.4 | 89.7               | 81.4                | 70.7   | 81.7            |
| $\operatorname{ResNet-101}$ | $256\times192$   | 71.6 | 88.9               | 79.3                | 68.5   | 78.2            |
| SCNet-101                   | $256\times192$   | 72.6 | 89.4               | 80.4                | 69.4   | 79.4            |
| ResNet-101                  | $384 \times 288$ | 73.9 | 89.6               | 80.5                | 70.3   | 81.1            |
| SCNet-101                   | $384 \times 288$ | 74.8 | 89.6               | 81.8                | 71.2   | 81.9            |

表 5. 关键点检测实验 [24]。我们以最先进的方法 [39]为基准 方法,使用基于 OKS 的 mAP 作为标准在 COCO val2017 数 据集上测试结果。按照 [39]使用两种不同的输入尺寸 (256 × 192 和 384 × 288)。

值可提高 2.5%。这些结果表明,在卷积层中引入自 校准操作对人体关键点检测有益。如表 5所示,当使 用更深的网络作为主干网络时,我们的 AP 值性能 增益也超过 1%。

## 5. 总结与展望

本文提出了一种新的自校准卷积,它能够异构 地利用卷积层中的卷积滤波器。为了使卷积滤波器 有多种模式,我们引入了自适应响应校准操作。我 们的自校准卷积可以很容易地嵌入到现有的分类网 络中。在大规模图像分类数据集上的实验表明,在 构建模块中构建多尺度特征表示可以大大提高预测 准确率。为了探究我们方法的泛化能力,我们将其 应用于多种主流的视觉任务,并发现在基线模型上 有实质性的提升。我们希望异构地开发卷积滤波器 的想法可以为视觉研究社区提供有关网络结构设计 的不同观点。

**致 谢** 这 项 研 究 得 到 了 新 一 代 人 工 智 能 重 大 项 目 (2018AAA01004)、 国 家 自 然 科 学 基 金 (61620106008)、国家青年人才支持计划和天津自然 科学基金会 (18ZXZNGX00110)的部分支持。部分 工作为刘姜江在 ByteDance AI Lab 实习期间完成。

# 参考文献

- Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. arXiv preprint arXiv:1904.09925, 2019. 2, 7
- [2] Ali Borji, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li. Salient object detection: A survey. Computational Visual Media, 5(2):117–150, 2019.
- [3] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeezeexcitation networks and beyond. arXiv preprint arXiv:1904.11492, 2019. 2, 7
- [4] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. mmdetection. https://github.com/openmmlab/mmdetection, 2018. 8
- [5] Liang-Chieh Chen, Maxwell Collins, Yukun Zhu, George Papandreou, Barret Zoph, Florian Schroff, Hartwig Adam, and Jon Shlens. Searching for efficient multi-scale architectures for dense image prediction. In NeurIPS, pages 8699–8710, 2018. 2
- [6] Yunpeng Chen, Haoqi Fan, Bing Xu, Zhicheng Yan, Yannis Kalantidis, Marcus Rohrbach, Shuicheng Yan, and Jiashi Feng. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In ICCV, pages 3435–3444, 2019.
- Yunpeng Chen, Jianan Li, Huaxin Xiao, Xiaojie Jin, Shuicheng Yan, and Jiashi Feng. Dual path networks. In Advances in Neural Information Processing Systems, pages 4467–4475, 2017. 1, 2
- [8] Deng-Ping Fan, Ming-Ming Cheng, Jiang-Jiang Liu, Shang-Hua Gao, Qibin Hou, and Ali Borji. Salient objects in clutter: Bringing salient object detection to the foreground. In ECCV, pages 186–202, 2018. 1
- [9] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. IEEE TPAMI, pages 1–1, 2020. 4
- [10] Shiming Ge, Xin Jin, Qiting Ye, Zhao Luo, and Qiang Li. Image editing by object-aware optimal boundary searching and mixed-domain composition. Computational Visual Media, 4(1):71–82, 2018. 1

- [11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In ICCV, pages 2980–2988.
   IEEE, 2017. 1, 7, 8
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016. 1, 2, 4, 5, 7
- [13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In ECCV, pages 630–645. Springer, 2016. 2
- [14] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip Torr. Deeply supervised salient object detection with short connections. IEEE TPAMI, 41(4):815–828, 2019. 2
- [15] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Andrea Vedaldi. Gather-excite: Exploiting feature context in convolutional neural networks. In NeurIPS, pages 9401–9411, 2018. 2, 4, 7
- [16] Jie Hu, Li Shen, and Gang Sun. Squeeze-andexcitation networks. In CVPR, pages 7132–7141, 2018.
  1, 2, 4, 5, 7
- [17] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In CVPR, pages 4700–4708, 2017.
- [18] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In ICML, 2015. 2
- [19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In NeurIPS, pages 1097–1105, 2012.
   2
- [20] Thuc Trinh Le, Andrés Almansa, Yann Gousseau, and Simon Masnou. Object removal from complex videos using a few annotations. Computational Visual Media, 5(3):267–291, 2019. 1
- [21] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In Artificial intelligence and statistics, pages 562–570, 2015. 5
- [22] Yi Li, Zhanghui Kuang, Yimin Chen, and Wayne Zhang. Data-driven neuron allocation for scale aggregation networks. In CVPR, pages 11526–11534, 2019.
- [23] Tsung-Yi Lin, Piotr Dollár, Ross B Girshick, Kaiming He, Bharath Hariharan, and Serge J Belongie. Fea-

ture pyramid networks for object detection. In CVPR, 2017.  $1,\,7$ 

- [24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In ECCV, 2014. 7, 8, 9
- [25] Drew Linsley, Dan Shiebler, Sven Eberhardt, and Thomas Serre. Learning what and where to attend. In ICLR, 2019. 2, 7
- [26] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. arXiv preprint arXiv:1806.09055, 2018.
- [27] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. A simple pooling-based design for real-time salient object detection. In CVPR, pages 3917–3926, 2019. 1
- [28] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters–improve semantic segmentation by global convolutional network. In CVPR, pages 4353–4361, 2017. 2
- [29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE TPAMI, 39(6):1137–1149, 2016. 1, 7, 8
- [30] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. IJCV, 115(3):211–252, 2015. 1, 5
- [31] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In ICCV, pages 618–626, 2017. 1, 6
- [32] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In ICLR, 2014. 2
- [33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015. 2
- [34] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inceptionresnet and the impact of residual connections on learning. In AAAI, volume 4, page 12, 2017. 1, 2

- [35] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In CVPR, pages 1–9, 2015. 2
- [36] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In CVPR, pages 2818–2826, 2016. 2
- [37] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In CVPR, pages 7794–7803, 2018. 2
- [38] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In ECCV, pages 3–19, 2018. 2, 4, 7
- [39] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In ECCV, 2018. 1, 8, 9
- [40] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In CVPR, pages 5987–5995, 2017. 1, 2, 4, 5, 7
- [41] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In ICLR, 2016. 2
- [42] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In CVPR, pages 2403–2412, 2018. 2
- [43] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. arXiv preprint arXiv:1605.07146, 2016.
   1, 2
- [44] Xingcheng Zhang, Zhizhong Li, Chen Change Loy, and Dahua Lin. Polynet: A pursuit of structural diversity in very deep networks. In CVPR, pages 718–726, 2017.
   2
- [45] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In CVPR, 2017. 1, 2, 3
- [46] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In CVPR, pages 2921– 2929, 2016. 3
- [47] Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. arXiv preprint arXiv:1611.01578, 2016. 1

[48] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In CVPR, pages 8697– 8710, 2018. 2