MORE CONVNETS IN THE 2020S: SCALING UP KERNELS BEYOND 51 × 51 USING SPARSITY

2020年更多的卷积网络：利用稀疏性将内核扩大到51×51以上

摘要

自从视觉变压器（ViTs）出现后，变压器迅速在计算机视觉领域大放异彩。卷积神经网络（CNN）的主导作用似乎受到了越来越有效的基于变换器的模型的挑战。最近，几个先进的卷积模型在局部窗口注意机制的激励下，以大内核的方式进行了反击，显示出了吸引人的性能和效率。虽然其中一个，即RepLKNet，令人印象深刻地设法将内核大小扩展到31×31，并提高了性能，但与Swin Transformer等高级ViTs的扩展趋势相比，随着内核大小的继续增长，性能开始饱和。在本文中，我们探索了训练大于31×31的极端卷积的可能性，并测试了是否可以通过战略性地扩大卷积来消除性能差距。这项研究最后从稀疏性的角度提出了一个应用极端大核的配方，它可以顺利地将核扩大到61×61，并且性能更好。在这个配方的基础上，我们提出了稀疏大核网络（SLaK），一个配备了稀疏因子化51×51核的纯CNN架构，在ImageNet分类以及广泛的下游任务上，包括ADE20K的语义分割、PASCAL VOC 2007的物体检测和MS COCO的物体检测/分割，可以与最先进的分层变压器和现代ConvNet架构相提并论，甚至更好。

代码链接：<https://github.com/VITA-Group/SLaK>

1介绍

自发明以来（Fukushima & Miyake, 1982; LeCun et al., 1989; 1998），卷积神经网络（CNN）（Krizhevsky et al., 2012a; Simonyan & Zisserman, 2015.; He et al., 2016; Huang et al., 2017; Howard et al., 2017; Xie et al., 2017; Tan & Le, 2019）已经迅速发展成为过去几十年机器学习中最不可或缺的架构之一。然而，在过去的几年里，CNN的主导地位受到了Transformer（Vaswani等人，2017）的极大挑战。源于自然语言处理，视觉变换器（ViTs）（Dosovitskiy等人，2021；d'Ascoli等人，2021；Touvron等人，2021b；Wang等人，2021b；Liu等人，2021e；Vaswani等人。2021）已经在各种计算机视觉任务中展示了强大的成果，包括图像分类（Dosovitskiy等人，2021；Yuan等人，2021b）、物体检测（Dai等人，2021；Liu等人，2021e）和分割（Xie等人，2021；Wang等人，2021a；c；Cheng等人，2021）。同时，关于理解ViTs的工作已经开花结果。ViTs成功背后的合理原因是较少的归纳偏差（Dosovitskiy等人，2021），长距离依赖（Vaswani等人，2017），先进的架构（Yu等人，2021），以及更像人类的表示（Tuli等人，2021）等。最近，有一种上升的趋势，即把ViTs的最高性能归结为捕捉大的感受野的能力。与在小的滑动窗口（如3×3和5×5）中用共享权重进行卷积的CNN相比，在ViTs中用较大的窗口尺寸进行全局注意或局部注意（Liu等人，2021e）直接使每个层能够捕获大的感受野。受到这一趋势的启发，最近一些关于CNN的工作（Liu et al., 2022b; Ding et al., 2022）通过设计先进的纯CNN架构并将大内核插入其中进行反击。例如，RepLKNet（Ding等人，2022）成功地将内核大小扩展到31×31，同时取得了与Swin Transformer（Liu等人，2021e）相当的结果。然而，大内核是出了名的难以训练。即使有小内核的并行分支的帮助，RepLKNet的性能也会随着内核大小的不断增加而开始饱和，与Swin Transformer等先进的ViTs的扩展趋势相比。因此，我们是否可以通过进一步扩大核大小到31×31以上来超越基于Transformer的模型，仍然是个谜。在本文中，我们试图通过利用人类视觉系统中常见的稀疏性来回答这个研究问题。稀疏性被认为是初级视觉皮层（V1）中最重要的原则之一（Tong，2003），在那里传入的刺激被假设为稀疏的编码和选择（Desimone & Duncan，1995；Olshausen & Field，1997；Vinje & Gallant，2000）。我们广泛地研究了大核的可训练性，并揭示了三个主要的观察结果：（i）现有的方法要么是天真地应用更大的核（Liu等人，2022b），要么是协助结构重参数化（Ding等人。2022）未能将核的大小扩大到31×31以上；（ii）（与分解卷积类似）用两个矩形的平行核（M×N和N×M，其中N <M）代替一个大的M×M核，可以将核的大小平滑地扩大到61×61，并提高了性能；（iii）用稀疏组构造，同时扩大宽度，可以明显提高性能。基于这些观察，我们提出了SLaK--稀疏大核网络--一种新的纯CNN架构，配备了前所未有的51×51的核大小。通过对各种任务的评估，包括ImageNet分类（Deng等人，2009）、ADE20K的语义分割（Zhou等人，2019）、PASCAL VOC 2007的物体检测（Everingham等人，2007）和COCO的物体检测/分割（Lin等人。2014），SLaK的表现比CNN先驱RepLKNet和ConvNeXt（Liu等人，2022b）以及基于SOTA的模型，如Swin（Liu等人，2021e）和Cswin（Dong等人，2022）在ImageNet上的变形器更好或更接近。我们对有效感受区（ERF）的分析表明，当插入最近提出的ConvNeXt时，我们的方法能够比现有的更大的内核范式覆盖更大的ERF区域，同时自然地引入类似人类的外围感应偏差（Lettvin等人，1976；Min等人，2022）。

2相关工作

**注意力中的大内核。**  
自我注意最初是为自然语言处理引入的（Vaswani等人，2017），并由Dosovitskiy等人（2021）在计算机视觉中进行了扩展，自我注意可以被看作是一个全局深度的内核，使每个层都有一个全局的感受野。Swin Transformer（Liu等人，2021e）是ViTs的一个变体，它采用了局部注意的移窗方式。与全局注意相比，局部注意（Ramachandran等，2019；Vaswani等，2021；Chu等，2021；Liu等，2021d；Dong等，2022）可以极大地提高记忆和计算效率，具有吸引人的性能。由于注意力窗口的大小至少为7，因此可以被看作是另一类大内核。最近的一项工作（Guo et al., 2022b）提出了一个新的大核注意模块，它使用堆叠深度、小卷积、扩张卷积以及点卷积来捕捉局部和全局结构。  
**卷积中的大内核。**  
卷积中的大核可以追溯到2010年代（Krizhevsky等人，2012b；Szegedy等人，2015；2017），如果不是更早的话，大核尺寸如7×7和11×11被应用。全局卷积网络（GCNs）（Peng等人，2017）通过采用1×M+M×1和M×1+1×M卷积的组合，将核大小扩大到15。然而，所提出的方法导致了ImageNet上的性能下降。Inceptions家族（Szegedy等人，2016；2017）允许利用不同的卷积核大小来学习不同尺度的空间模式。随着VGG（Simonyan & Zisserman, 2014）的普及，在过去的十年中，使用小核的堆叠（1×1或3×3）来获得大的感受野是很常见的（He等人，2016；Howard等人，2017；Xie等人，2017；黄等人，2017）。直到最近，一些工作开始恢复大核在CNN中的使用。Li等人（2021）提出了7×7大核的内卷化，在空间范围内使用不同的权重，同时在通道间共享权重。然而，当进一步扩大核的大小时，性能的提高趋于平稳。Han等人（2021b）发现，如果我们在Swin Transformer中用后者替代前者，动态深度卷积（7×7）的表现与局部注意机制相当。Liu等人（2022b）模仿Swin Transformer（Liu等人，2021e）的设计元素，设计了采用7×7核的ConvNeXt，超越了前者的性能。RepLKNet（Ding等人，2022）首次通过构建与之平行的小核（如3×3或5×5）将核的大小扩展到31×31，并取得了与Swin Transformer相当的性能。连续卷积核的一系列工作（Romero等人，2021；2022）可用于任意分辨率、长度和维度的数据。最近，Chen等人（2022）揭示了大核对三维网络也是可行的和有益的。之前的工作已经探索了平行（Peng等人，2017；Guo等人，2022a）或堆叠（Szegedy等人，2017）两个互补的M×1和1×M内核的想法。然而，他们将短边限制为1，并且没有将核的大小扩展到51×51以上。与这些先有技术不同的是，我们将一个大核分解成两个互补的非方形核（M×N和N×M），改善了大卷积核的训练稳定性和内存扩展性。  
**动态稀疏性。**  
作为一个长期的研究课题，最近关于稀疏性的尝试（Mocanu等人，2018；Liu等人，2021b；Evci等人，2020；Mostafa & Wang，2019；Dettmers & Zettlemoyer，2019；Chen等人，2021）仅使用一小部分参数和FLOPs（如图2所示）从头训练内在的稀疏神经网络。

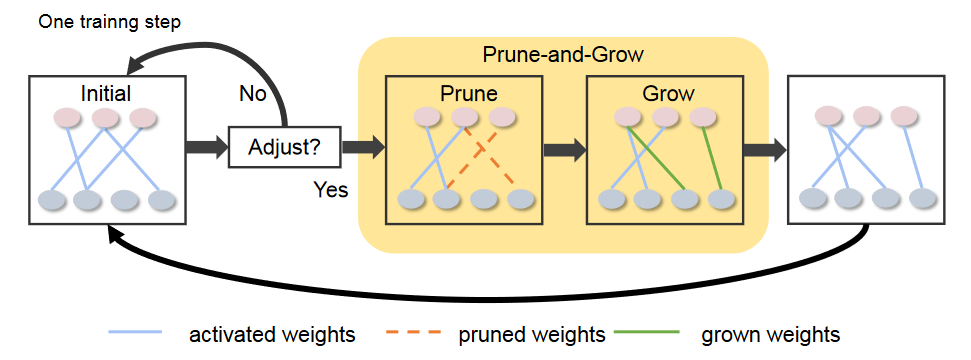


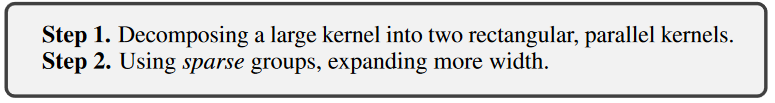
图2：动态稀疏性。动态稀疏性使我们能够从头开始构建和训练最初的稀疏神经网络（稀疏内核）。在训练过程中，它通过修剪最不重要的权重和增加新的权重来动态地调整稀疏权重。这样的动态程序逐渐将稀疏内核优化到一个良好的模式，从而鼓励对局部特征进行更细致的捕捉。

动态稀疏性能够从头开始训练稀疏模型，因此训练和推理的FLOPs和内存需求只是密集模型的一小部分。与训练后修剪不同（Han等人，2015；Frankle & Carbin，2019），用动态稀疏性建立的模型可以从头开始训练，以匹配它们的密集模型，而不涉及任何预训练或密集训练。动态稀疏性源于稀疏进化训练（SET）（Mocanu等人，2018；Liu等人，2021b），它随机初始化层间的稀疏连接，在训练过程中通过参数修剪和增长方案动态调整稀疏连接。参数修剪和增长方案使模型的稀疏结构逐渐演变，比天真地训练静态稀疏网络取得更好的性能（Liu等人，2021c）。在本文中，我们的目标不是要找到能够与相应的密集网络相匹配的稀疏网络。在ResNeXt（Xie等人，2017；Liu等人，2022b）原则的激励下--"使用更多的组，扩大宽度"，我们反而试图利用动态稀疏性来扩展具有极端内核的神经架构。  
  
3现有的方法未能超越31×31的内核

我们首先研究了大于31×31的极端内核大小的性能，并在此分享我们的三个主要观察结果。我们以最近在ImageNet-1K上开发的CNN架构ConvNeXt（Liu等人，2022b）为基准来进行研究。我们在本文中采用了MegEngine（Meg，2020）开发的高效大核实现。我们遵循最近的工作（Liu等人，2022b；Bao等人，2021；Liu等人，2021e；Ding等人。2022；Touvron等人，2021b）使用Mixup（Zhang等人，2017）、Cutmix（Yun等人，2019）、RandAugment（Cubuk等人，2020）和Random Erasing（Zhong等人，2020）作为数据增强。随机深度（Huang等人，2016）和标签平滑（Szegedy等人，2016）被用作正则化，其超参数与ConvNeXt中使用的相同。我们用AdamW（Loshchilov & Hutter，2019）训练模型。需要注意的是，在本节中，所有模型的训练长度都减少到120个历时，只是为了勾勒出大核尺寸的扩展趋势。在第5节的后面，我们将采用完整的训练配方，对我们的模型进行300个历时的训练，以便与最先进的模型进行公平的比较。更多细节请参考附录A。最近，RepLKNet（Ding等人，2022）通过结构上的重新参数化成功地将卷积扩展到31×31（Ding等人，2019；2021）。我们进一步将核的大小增加到51×51和61×61，看看更大的核是否能带来更多的收益。按照RepLKNet的设计，我们将每个阶段的核大小设定为[51, 49, 47, 13]和[61, 59, 57, 13]，并在表1中报告测试精度。正如预期的那样，天真地将核大小从7×7扩大到31×31会降低性能，而RepLKNet可以克服这个问题，将准确率提高0.5%。不幸的是，当我们进一步增加核的大小到51×51时，这种积极的趋势并没有继续。一个合理的解释是，尽管使用极大的核可以扩大感受野，但它可能无法保持理想的定位特性。由于标准ResNet（He等人，2016）和ConvNeXt中的干细胞导致输入图像的4倍下采样，51×51的极端内核已经大致等于典型的224×224 ImageNet的全局卷积。因此，这一观察是有意义的，因为在类似的ViTs机制中，精心设计的局部注意力（Liu等人，2021e；d；Chu等人，2021）通常优于全球注意力（Dosovitskiy等人，2021）。受此启发，我们认为有机会通过引入位置性，同时保留捕捉全局关系的能力来解决这个问题。

4一个超过31×31的超大核的配方

在这一节中，我们介绍了一个简单的两步法，用于超过31×31的极大型核。



**将一个大核分解成两个矩形的、平行的核，可以将核的大小平滑地扩展到61×61。**

虽然使用中等大小的卷积（如31×31）似乎可以直接避免这个问题，但我们想研究是否可以通过使用（全局）极端卷积进一步推动CNN的性能。我们这里的秘诀是用两个平行和矩形卷积的组合来近似大的M×M核，其核大小分别为M×N和N×M（其中N<M）， 如图1所示。

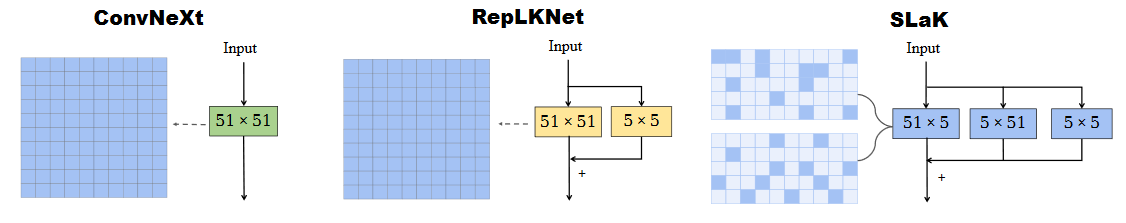


图1：ConvNeXt、RepLKNet和SLaK的大纵深核（例如51×51）范式。深蓝色的方块指的是卷积核中的密集权重。浅蓝色方块指的是卷积核中的稀疏权重。

按照Ding等人（2022）的做法，我们保留了一个与大核平行的5×5层，并在批量规范层后将其输出相加。这种分解在捕捉长距离的依赖性和提取局部细节特征（其边缘较短）之间取得了平衡。此外，现有的大核训练技术（Liu等人，2022b；Ding等人，2022）随着核大小的增加，存在着二次计算和内存开销。与此形成鲜明对比的是，我们的方法的开销只是随着核的大小而线性增加（图4）。

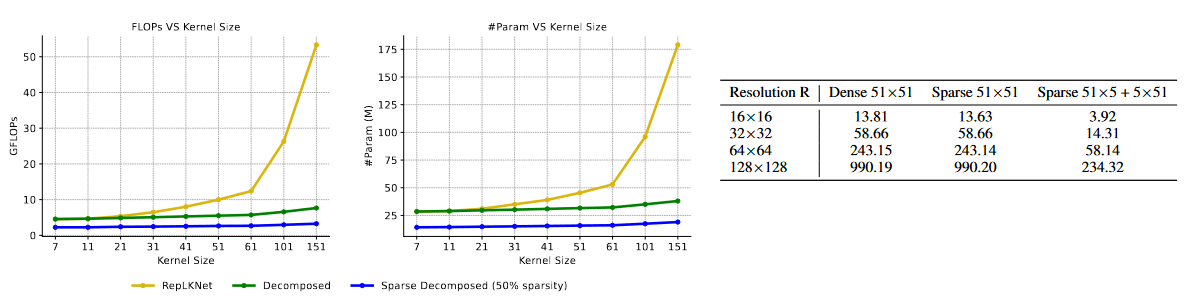
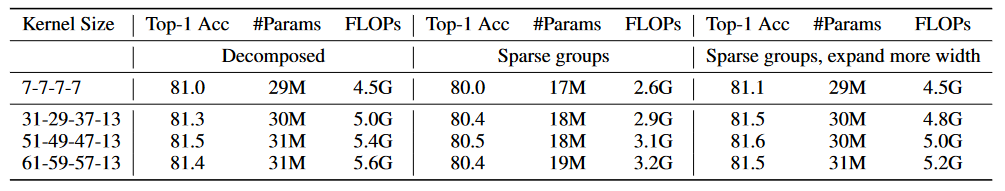


图4：左：扩展效率。随着ConvNeXt-T中内核大小的增加，GFLOPs的数量（左）和参数（右）。右图。实际推理时间延迟（ms）。不同内核大小的深度大小卷积的实际推理时间延迟。

表2中的 "分解 "组报告了N=5（关于N的影响见附录D）的内核分解的性能。



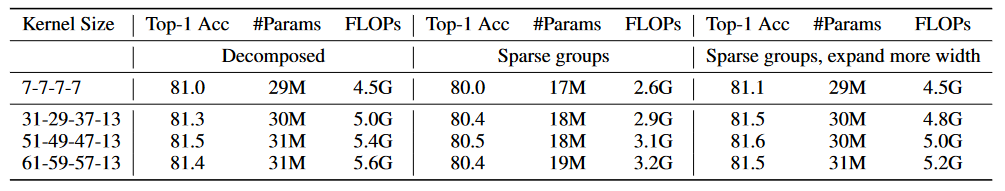
由于分解减少了可学习的参数和FLOPs，观察到我们的网络在最初与原始的RepLKNet相比，在中等核大小即31×31的情况下，精确度略有下降，这并不奇怪。然而，随着卷积大小的不断增加，我们的方法可以将核的大小扩展到61×61，并提高了性能。

**"使用稀疏的组，扩展更多的宽度 "大大提升了模型的容量。**

最近提出的ConvNeXt（Liu等人，2022b）重新审视了ResNeXt（Xie等人，2017）中引入的原理，将卷积滤波器分成小而多的组。ConvNeXt没有使用标准的组卷积，而是简单地采用了增加宽度的深度卷积，以达到 "使用更多组，扩大宽度 "的目的。在本文中，我们试图从稀疏性的角度来扩展这一原则--"使用稀疏组，扩展更多宽度"。具体来说，我们首先用稀疏卷积代替密集卷积，由于SNIP（Lee等人，2019）1在大规模模型上的强大性能，稀疏核是根据层间稀疏比随机构建的（Liu等人，2022a）。构建完成后，我们用动态稀疏性训练稀疏模型（Mocanu等人，2018；Liu等人，2021b），在训练过程中，通过修剪幅度最小的权重和随机增长相同数量的权重来动态调整稀疏权重。这样做可以实现稀疏权重的动态适应，从而获得更好的局部特征。由于内核在整个训练过程中是稀疏的，相应的参数数和训练/推理FLOPs只与密集模型成正比。关于动态稀疏性的配置，见附录G。为了评估，我们用40%的稀疏度对分解的内核进行稀疏化，并在 "稀疏组 "一栏中报告其性能。我们可以在表2的中间一栏中观察到，动态稀疏性明显减少了2.0GFLOPs以上，尽管造成了暂时的性能下降。我们接下来表明，上述动态稀疏性的高效率可以有效地转移到模型的可扩展性。动态稀疏性使我们能够在计算上友好地扩大模型的规模。例如，使用相同的稀疏度（40%），我们可以将模型宽度扩大1.3倍，同时保持参数数和FLOPs与密集模型大致相同。这给我们带来了显著的性能提升，在极端的51×51内核下，性能从81.3%提升到81.6%。令人印象深刻的是，在配备61×61内核的情况下，我们的方法超过了以前的技术水平（Liu等人，2022b；Ding等人，2022），同时节省了55%的FLOPs。

**使用我们的方法，大内核的通用性比小内核好。**

为了证明大内核的好处，我们还在表2中报告了7×7小内核的每个步骤的影响。



我们可以清楚地看到，性能始终随着内核大小的增加而增加，最高可达51×51。与51×51的内核相比，将我们提出的配方的每一部分应用于7×7的内核，要么没有收益，要么收益很小。这个分解实验证明了我们的主张：大内核是力量的根源，而我们提出的配方有助于从大内核中释放出这种力量。（这个方法得在大内核里面用才厉害）

4.1构建稀疏大核网络（slak）。

到目前为止，我们已经发现了我们的配方，它可以成功地将内核大小扩展到51×51而不影响性能。在这个配方的基础上，我们接下来构建了我们自己的稀疏大核网络（SLAK），一个采用极端的51×51核的纯CNN架构。SLaK是基于ConvNeXt的架构建立的。阶段计算率和stem cell的设计继承自ConvNeXt。SLaK-T的每个阶段的块数为[3, 3, 9, 3]，SLaK-S/B的块数为[3, 3, 27, 3]。 stem cell 只是一个具有4×4核和4个步长的卷积层。我们首先直接将ConvNeXt的核大小增加到每个阶段的[51, 49, 47, 13]，用M×5和5×M核的组合取代每个M×M核，如图1所示。我们发现在每个分解的核之后直接添加一个BatchNorm层，在将输出相加之前是至关重要的。遵循 "使用稀疏组，扩大宽度 "的准则，我们将整个网络进一步稀疏化，并将阶段的宽度扩大了1.3倍，最终得到SLaK。即使通过调整模型宽度和稀疏度之间的权衡，SLaK的性能仍有很大的提升空间（如附录C所示），我们在所有的实验中都保持一组超参数（1.3倍的宽度和40%的稀疏度），因此SLaK可以简单地 "开箱即用"，根本无需特别调整。  
  
7 结论和对更广泛影响的讨论

最近关于现代ConvNets的工作通过设计先进的架构和插入大内核来捍卫卷积在计算机视觉中的重要作用。然而，最大的内核尺寸被限制在31×31，并且随着内核尺寸的不断增加，性能开始饱和。在本文中，我们研究了用超过31×31的超大内核训练ConvNets的问题，并因此提供了一个受稀疏性启发的应用超大内核的秘诀。基于这个配方，我们建立了一个纯粹的ConvNet模型，该模型可以平滑地将内核大小扩展到51×51以上，同时取得比Swin Transformer和ConvNeXt更好的性能。我们强有力的结果表明，作为深度学习的 "老朋友"，稀疏性可以成为促进神经网络扩展的一个有前途的工具。这项工作是科学性质的，我们不期望有任何负面的社会影响。