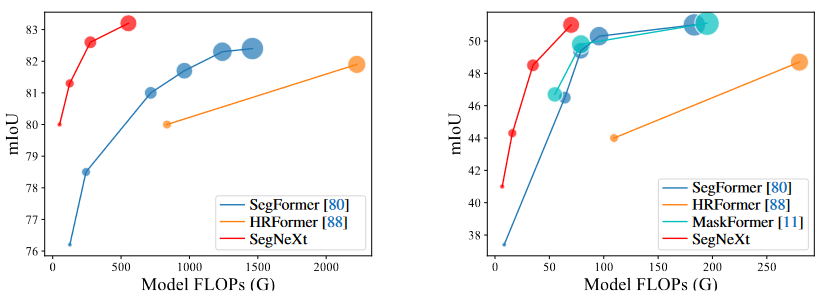
SegNeXt: Rethinking Convolutional Attention Design for Semantic Segmentation

SegNeXt: 重新思考用于语义分割的卷积注意力设计·

我们提出了SegNeXt，一个用于语义分割的简单卷积网络架构。由于自我注意在编码空间信息方面的效率，最近基于变换器的模型在语义分割领域占主导地位。在本文中，我们表明卷积注意是一种比变压器中的自我注意机制更有效率和效果的编码上下文信息的方式。通过重新审视成功的分割模型所拥有的特征，我们发现了导致分割模型性能提高的几个关键成分。这促使我们设计了一个新颖的卷积注意力网络，它使用廉价的卷积操作。在没有任何附加条件的情况下，我们的SegNeXt大大改善了以前最先进的方法在流行基准上的性能，包括ADE20K、Cityscapes、COCO-Stuff、Pascal VOC、Pascal Context和iSAID。值得注意的是，SegNeXt的性能优于EfficientNet-L2 w/NAS-FPN，在Pascal VOC 2012测试排行榜上只用了它的1/10的参数就达到了90.6%的mIoU。与最先进的方法相比，SegNeXt在ADE20K数据集上以相同或更少的计算量平均实现了约2.0%的mIoU改进。代码是可用的。

代码链接：<https://github.com/Visual-Attention-Network/SegNeXt>

介绍  
作为计算机视觉中最基本的研究课题之一，语义分割旨在为每个像素分配一个语义类别，在过去十年中引起了极大的关注。从早期基于CNN的模型，如FCN[53]和DeepLab系列[4, 6, 8]，到最近基于变换器的方法，如SETR[96]和SegFormer[80]，语义分割模型在网络架构方面经历了重大的革命。基于上述观察，我们认为一个成功的语义分割模型应该具有以下特征：（i）强大的骨干网络作为编码器。与之前基于CNN的模型相比，基于变换器的模型的性能提升主要来自于更强的骨干网络。(ii) 多尺度信息交互。与图像分类任务大多识别单一物体不同，语义分割是一个密集的预测任务，因此需要处理单一图像中不同大小的物体。(iii) 空间注意力。空间注意力允许模型通过对语义区域内的区域进行优先排序来进行分割。(iv) 低计算复杂性。在处理遥感和城市场景的高分辨率图像时，这一点尤其关键。考虑到上述分析，在本文中，我们重新思考了卷积注意的设计，并为语义分割提出了一个高效而有效的编码器-解码器架构。与之前的基于变压器的模型不同，我们的方法是将解码器中的卷积作为特征细化器，倒置了变压器-卷积编码器-解码器架构。具体来说，对于我们编码器中的每个块，我们翻新了传统卷积块的设计，并利用多尺度卷积特征，通过一个简单的元素相乘来唤起空间注意力，遵循[24]。我们发现这样一种简单的建立空间注意力的方法比标准卷积和空间信息编码中的自我注意力都要有效。对于解码器，我们从不同的阶段收集多层次的特征，并使用Hamburger[21]来进一步提取全局背景。在这种情况下，我们的方法可以获得从局部到全局的多尺度背景，实现空间和通道维度的适应性，并从低级到高级的信息聚合。我们的网络被称为SegNeXt，除了解码器部分，大部分由卷积运算组成，其中包含一个基于分解的汉堡模块[21]（Ham）用于全局信息提取。这使得我们的SegNeXt比以前严重依赖变换器的分割方法要高效得多。如图1所示，SegNeXt明显优于最近基于变换器的方法。



特别是我们的SegNeXt-S在处理Cityscapes数据集的高分辨率城市场景时，只用了大约1/6（124.6G vs. 717.1G）的计算成本和1/2的参数（13.9M vs. 27.6M）就超过了SegFormer-B2（81.3% vs. 81.0%）。我们的贡献可以归纳为以下几点。  
- 我们确定了一个好的语义分割模型应该拥有的特征，并提出了一个新颖的定制网络架构，称为SegNeXt，它通过多尺度卷积特征唤起空间注意。

- 我们表明，一个具有简单而廉价的卷积的编码器仍然可以比视觉变换器表现得更好，特别是在处理物体细节时，而它所需要的计算成本要低得多。

- 我们的方法在各种分割基准上，包括ADE20K、Cityscapes、COCO-Stuff、Pascal VOC、Pascal Context和iSAID，将最先进的语义分割方法的性能提高了很多。

2 相关工作

2.1 语义分割

语义分割是一项基本的计算机视觉任务。自FCN[53]提出以来，卷积神经网络（CNN）[1, 64, 86, 94, 19, 87, 71, 20, 45]已经取得了巨大的成功，成为语义分割的流行架构。最近，基于变换器的方法[96, 80, 88, 65, 63, 44, 11, 10]已经显示出巨大的潜力，并超过了基于CNN的方法。在深度学习时代，分割模型的结构可以大致分为两部分：编码器和解码器。对于编码器，研究人员通常采用流行的分类网络（如ResNet[27]、ResNeXt[81]和DenseNet[32]），而不是定制的架构。然而，语义分割是一种密集预测任务，它与图像分类不同。分类的改进可能不会出现在具有挑战性的分割任务中[28]。因此，出现了一些定制的编码器，包括Res2Net[20]、HRNet[71]、SETR[96]、SegFormer[80]、HRFormer[88]、MPViT[38]、DPT[63]，等等。对于解码器来说，它经常与编码器合作使用，以达到更好的效果。针对不同的目标，有不同类型的解码器，包括实现多尺度接受场[94, 7, 78]，收集多尺度语义[64, 80, 8]，扩大接受场[5, 4, 62]，加强边缘特征[95, 2, 16, 42, 90]，以及捕捉全局背景[19, 34, 89, 40, 23, 26, 91]。在本文中，我们总结了那些为语义分割设计的成功模型的特点，并提出了一个基于CNN的模型，命名为SegNeXt。与我们的论文最相关的工作是[62]，它将k×k卷积分解为一对k×1和1×k卷积。虽然这项工作表明大卷积核在语义分割中很重要，但它忽略了多尺度接受场的重要性，也没有考虑如何利用大卷积核提取的这些多尺度特征，以注意力的形式进行分割。

2.2 多尺度网络

设计多尺度网络是计算机视觉中的一个流行方向。对于分割模型，多尺度块出现在编码器[71, 20, 67]和解码器[94, 86, 6]两部分。GoogleNet[67]是与我们的方法最相关的多尺度架构之一，它使用多分支结构来实现多尺度特征提取。另一个与我们方法相关的工作是HRNet[71]。在深层阶段，HRNet也保留了高分辨率的特征，与低分辨率的特征进行聚合，从而实现多尺度特征提取。与之前的方法不同，SegNeXt除了在编码器中捕获多尺度特征外，还引入了一个有效的注意力机制，并采用了更便宜和更大的内核卷积。这些都使我们的模型能够达到比上述的分割方法更高的性能。

2.3 注意机制

注意机制是一种自适应的选择过程，其目的是使网络集中于重要部分。一般来说，它在语义分割中可以分为两类[25]，包括通道注意和空间注意。不同类型的注意起着不同的作用。例如，空间注意主要关心重要的空间区域[17, 14, 57, 51, 22]。不同的是，使用通道注意的目的是使网络有选择地注意那些重要的物体，这在以前的工作中已经被证明是很重要的[30, 9, 72]。说到最近流行的视觉变换器[17, 51, 82, 74, 73, 50, 80, 33, 49, 88]，它们通常忽略了通道维度的适应性。视觉注意力网络（VAN）[24]是与SegNeXt最相关的工作，它也提出利用大核注意力（LKA）机制来建立通道和空间注意力。虽然VAN在图像分类中取得了很好的表现，但它在网络设计过程中忽略了多尺度特征聚合的作用，而这对于类似分割的任务是至关重要的。

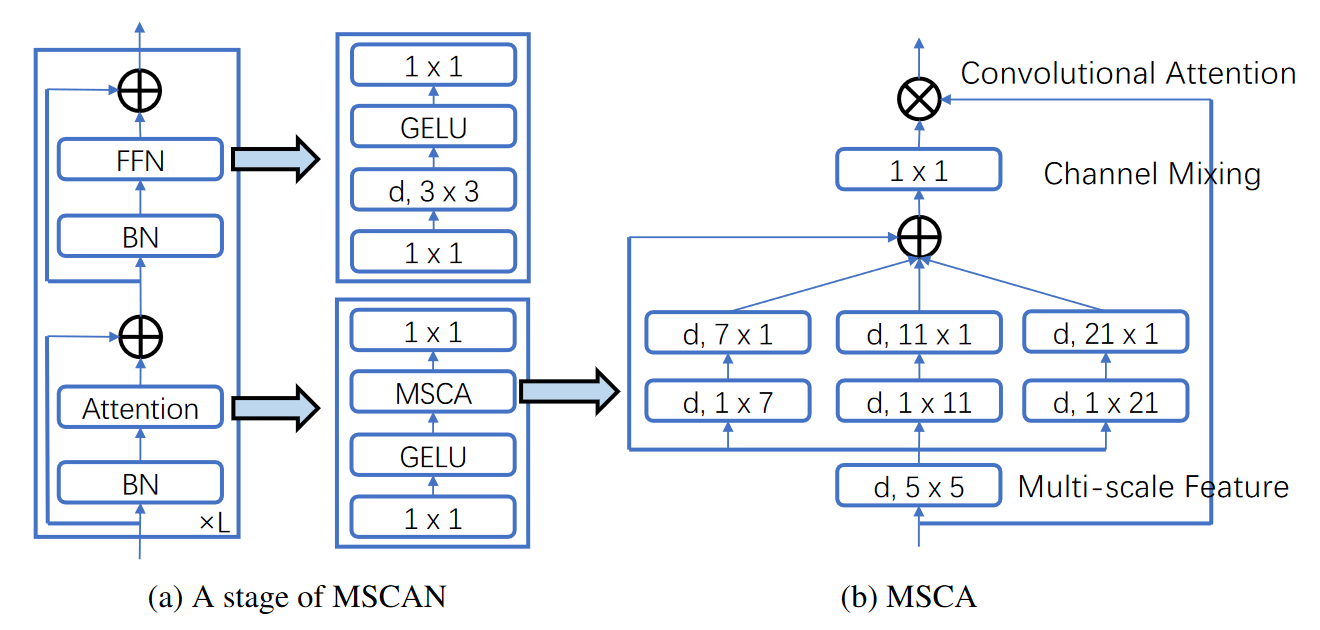
3 方法

在本节中，我们将详细描述所提出的SegNeXt的结构。基本上，我们采用的是一个编码器-解码器的架构，这与以前的大多数工作一样，简单易行。

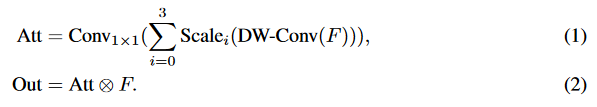
基础介绍：分解卷积

3.1 卷积编码器

我们的编码器采用的是金字塔结构，这与之前的大多数工作[80, 5, 19]相同。对于我们编码器中的构件，我们采用了与ViT[17, 80]类似的结构，但不同的是，我们没有使用自我注意机制，而是设计了一个新颖的多尺度卷积注意（MSCA）模块。如图2(a)所示，



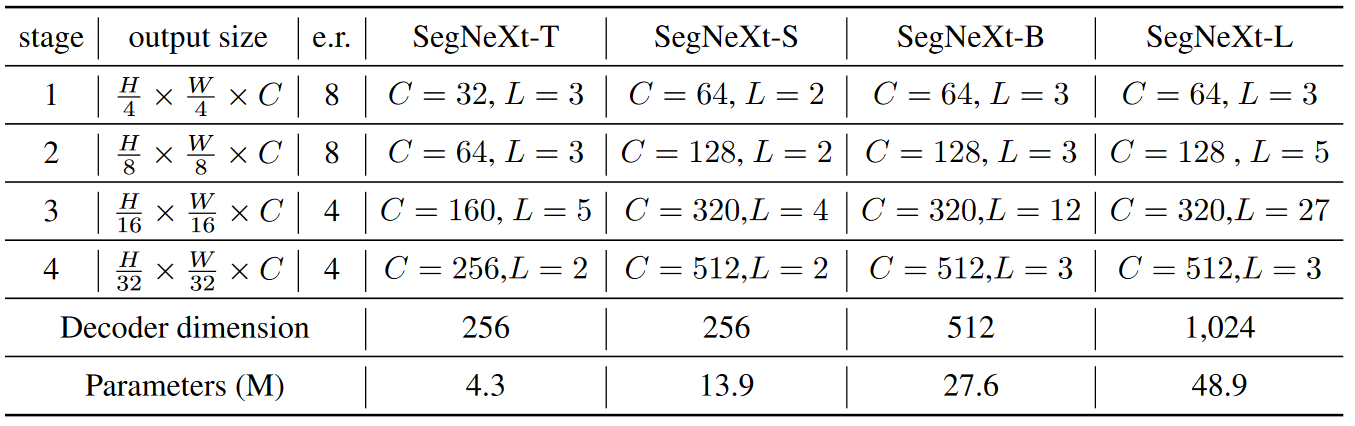
MSCA包含三个部分：一个深度卷积来聚集局部信息，多分支深度卷积来捕捉多尺度背景，以及一个1×1卷积来模拟不同通道之间的关系。1×1卷积的输出被直接用作注意力权重，以重新权衡MSCA的输入。在数学上，我们的MSCA可以写成：



总结：这里的Att就是权重系数

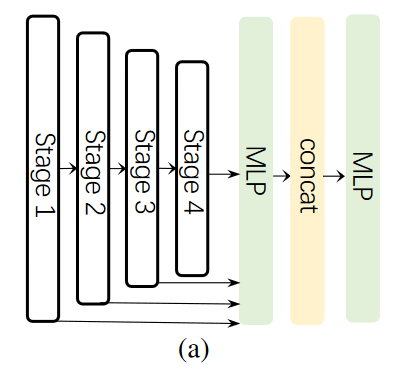
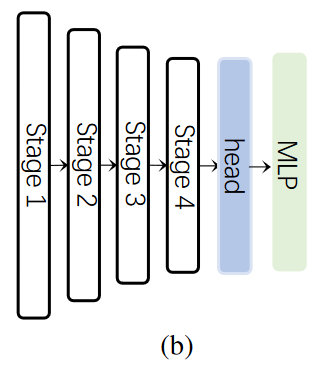
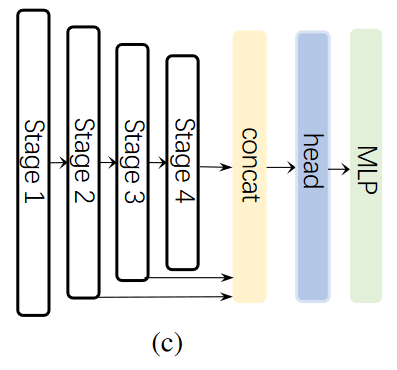
其中F代表输入特征。Att和Out分别为注意图和输出。⊗是逐元的矩阵乘法运算。DW-Conv表示深度卷积，Scalei，i∈{0，1，2，3}，表示图2（b）中的第i个分支。Scale0是身份连接。按照[62]，在每个分支中，我们使用两个深度明智的带状卷积来近似于大核的标准深度明智的卷积。这里，每个分支的核大小分别被设定为7、11和21。我们选择深度的条状卷积的原因有两个方面。一方面，带状卷积是轻量级的。为了模仿核大小为7×7的标准二维卷积，我们只需要一对7×1和1×7的卷积。另一方面，在分割场景中存在一些条状物体，如人和电线杆。因此，条状卷积可以作为网格卷积的补充，有助于提取条状特征[62, 29]。将一连串的构件堆叠在一起，就得到了提议的卷积编码器，命名为MSCAN。对于MSCAN，我们采用一个普通的分层结构，它包含四个阶段，空间分辨率依次为H/4 ×W/4,H/8 ×W/8,H/16 ×W/16和H/32 ×W/32。这里，H和W分别是输入图像的高度和宽度。如上所述，每个stage都包含一个下采样块和一个堆叠的构建块。下采样块有一个跨度为2、核大小为3×3的卷积，然后是一个批量归一化层[35]。请注意，在MSCAN的每个构建块中，我们使用批量归一化而不是层归一化，因为我们发现批量归一化对分割性能的增益更大。  
我们设计了四个不同规模的编码器模型，分别命名为MSCAN-T、MSCAN-S、MSCAN-B和MSCAN-L。相应的整体分割模型分别被称为SegNeXt-T, SegNeXt-S, SegNeXt-B, SegNeXt-L。详细的网络设置显示在表2.

表2: 提出的SegNeXt的不同规模的详细设置。在这个表中，'e.r.'代表前馈网络的扩展率。'C'和'L'分别是通道和构件的数量。'解码器尺寸'表示解码器中的MLP尺寸。'参数'是在ADE20K数据集上计算的[98]。由于不同的数据集中类别的数量不同，参数的数量可能略有变化。



3.2 解码器

在分割模型[80, 96, 5]中，编码器大多是在ImageNet数据集上预训练的。为了捕捉高层次的语义，通常需要一个解码器，它被应用在编码器上。在这项工作中，我们研究了三种简单的解码器结构，如图3所示。

第一种，在SegFormer[80]中采用，是一种纯粹的基于MLP的结构。第二种是主要采用基于CNN的模型。在这种结构中，编码器的输出被直接用作重型解码器头的输入，如ASPP[5]、PSP[94]和DANet[19]。最后一种是我们的SegNeXt中采用的结构。我们将后三个阶段的特征聚集起来，并使用一个轻量级的Hamburger[21]来进一步建立全局背景模型。结合我们强大的卷积编码器，我们发现使用一个轻量级的解码器可以提高性能-计算效率。

值得注意的是，与SegFormer不同的是，SegFormer的解码器聚集了第一阶段到第四阶段的特征，而我们的解码器只接收最后三个阶段的特征。这是因为我们的SegNeXt是基于卷积的。第一阶段的特征包含了太多的低层次信息，会影响性能。此外，对第一阶段的操作带来了沉重的计算开销。在实验部分，我们将展示我们的卷积SegNeXt比最近最先进的基于变换器的SegFormer[80]和HRFormer[88]表现得更好。

5 结论和讨论

在本文中，我们分析了以前成功的分割模型，并找到了它们所拥有的良好特性。基于这些发现，我们提出了一个量身定做的卷积注意模块MSCA和一个CNN式网络SegNeXt。实验结果表明，SegNeXt在相当大的程度上超过了目前最先进的基于变压器的方法。最近，基于变换器的模型在各种分割排行榜上占主导地位。相反，本文表明，当使用适当的设计时，基于CNN的方法仍然可以比基于变压器的方法表现得更好。我们希望本文能够鼓励研究人员进一步研究CNN的潜力。我们的模型也有其局限性，例如，将此方法扩展到具有100M以上参数的大规模模型以及在其他视觉或NLP任务上的表现。这些问题将在我们未来的工作中得到解决。