Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet

摘要

Transformer 在语言建模中很受欢迎，最近也被探索用于解决视觉任务，例如，用于图像分类的 Vision Transformer(ViT)。ViT 模型将每个图像分割成具有固定长度的 token 序列，然后应用多个 Transformer 层来模拟它们的全局关系以进行分类。然而，当在 ImageNet 这样的中型数据集上从头开始训练时，ViT 取得的性能不如 CNN。

为了克服这些局限，我们提出了一个新的 Token-To-Token Vision Transformer(T2T-ViT)，它包括：

1.一个分层的 Token-to-Token(T2T)变换，通过递归地将相邻的 Token 融合成一个 Token(Token-to-Token)，逐步将图像结构化为 Token，这样，由周围 Token 表示的局部结构可以被建模，使得 Token 长度可以被减少；

2.经过实证研究，在 CNN 架构设计的启发下，为 Vision Transformer 提供了一个具有"deep-narrow"结构的高效 backbone；

值得注意的是，T2T-ViT 将 vanilla ViT 的参数量与 MAC(Multi-Adds)减少了一半，而在 ImageNet 上从头开始训练时取得了超过 3.0%的提升。通过直接在 ImageNet 上训练，它的性能也超过了 ResNet，并达到了与 MobileNet 相当的性能。

关键词：机器学习；图像分类

1 引言

Vision Transformer(ViT)是第一个可以直接应用于图像分类的全 Transformer 模型。具体地说，ViT 将每个图像分割成固定长度的 14×14 或 16×16 块(也称为 tokens)；然后 ViT 应用 Transformer 层对这些 tokens 之间的全局关系进行建模以进行分类。

尽管 ViT 证明了全 Transformer 架构在视觉任务中很有前途，但在中型数据集(例如 ImageNet)上从头开始训练时，其性能仍逊于类似大小的 CNN 对等架构(例如 ResNets)。

论文假设，这种性能差距源于 ViT 的两个主要局限性：

(1)通过硬分裂对输入图像进行简单的 tokens 化，使得 ViT 无法对图像的边缘和线条等局部结构进行建模，因此它需要比 CNN 多得多的训练样本(如 JFT-300M 用于预训练)才能获得类似的性能；

(2)ViT 的注意力骨干没有很好地像用于视觉任务的 CNN 那样的设计，如 ViT 具有冗余性和特征丰富度有限的缺点，导致模型训练困难。

为了验证论文的假设，论文进行了一项初步研究，通过中可视化来调查 ViTL/16 和 ResNet5 的获知特征的差异。论文观察 ResNet 的功能，捕捉所需的局部结构(边、线、纹理等)。从底层(Cv1)逐渐向中间层(Cv25)递增。

然而，ViT 的特点却截然不同：结构信息建模较差，而全局关系(如整条狗)被所有的注意块捕获。这些观察结果表明，当直接将图像分割成固定长度的 tokens 时，原始 ViT 忽略了局部结构。此外，论文发现 ViT 中的许多通道都是零值(在图 2 中以红色突出显示)，这意味着 ViT 的主干不如 ResNet 高效，并且在训练样本不足的情况下提供有限的特征丰富度。

文章针对 ViT 中 tokenization 设计的不足进行了进一步的改进，让每个 token 能够捕捉到更加精细的 local structure，还探索了 CNN 中经典结构设计向 Vision Transformer 的迁移，基于一些传统的设计理念重新设计了 Vision Transformer 的 backbone 结构。

2 相关工作

(1) Transformers in Vision。

transfomer 是完全依赖自注意机制来绘制输入和输出之间的全局依赖关系的模型，目前它们主导了自然语言建模。变压器层通常由多头自注意层(MSA)和 MLP 块组成。在每一层和自注意层和 MLP 块中的剩余连接之前应用层 norm (LN)。最近的工作探索了将变压器应用于各种视觉任务:图像分类，目标检测，分割，图像增强，图像生成，视频处理，3D 点云处理。其中，Vision Transformer (ViT)证明了纯 Transformer

架构也可以在图像分类上获得最先进的性能。

然而，ViT 严重依赖 ImageNet-21k 和 JFT-300M(未公开)等大规模数据集进行模型预训练，需要巨大的计算资源。相比之下，我们提出的 T2T-ViT 更有效，可以在 ImageNet 上训练，而无需使用那些大规模的数据集。最近的并行工作 DeiT 应用 Knowledge Distillation 对原有的 ViT 进行改进，在类令牌的同时加入 KD 令牌，这与我们的工作是正交的，因为我们的 T2T-ViT 侧重于架构设计，并且我们的 T2T-ViT 在没有 CNN 作为教师模型的情况下可以获得比 DeiT 更高的性能。

(2) Self-attention in CNNs。

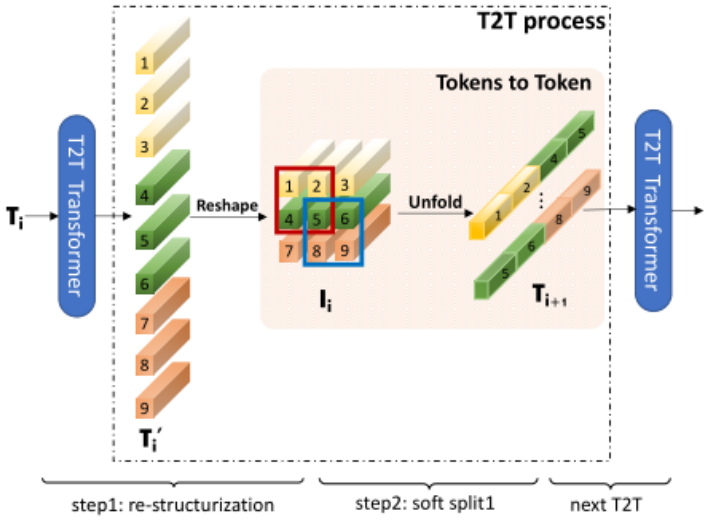
自注意机制已广泛应用于 CNN 在视觉任务中的应用。其中，SE 块[20]将注意力应用于信道维度，非本地网络设计用于通过全局注意力捕获长期依赖关系。与大多数研究图像全局注意的工作相比，一些工作也在局部补丁中研究自注意，以减少内存和计算成本。最近，SAN 研究了图像识别的成对和 patchwise 自注意，其中 patchwise 自注意是卷积的泛化。在这项工作中，我们也在实验中用多个卷积层替换了 T2T 模块，发现卷积层的性能并不比我们设计的 T2T 模块好。

3 本文方法

T2T-ViT 由两个主要部分组成：

(1) 一个层次化的“Tokens-to-Token 模块”(T2T 模块)，用于对图像的局部结构信息进行建模，并逐步减少 tokens 的长度。

Tokens-to-Token(T2T)模块旨在克服 ViT 中简单 tokens 化的限制。它将图像逐步结构化为表征，并对局部结构信息进行建模，这样可以迭代地减少表征的长度。每个 T2T 流程有两个步骤：重组和 Soft Split(SS)。



(2) 一个有效的“T2T-ViT 骨干”，用于从 T2T 模块中提取对 tokens 的全局关注关系。

论文探索了不同的 VIT 体系结构设计，并借鉴了 CNN 的一些设计，以提高骨干网的效率，增强学习特征的丰富性。由于每个 transformer 层都有跳跃连接，一个简单的想法是采用如 DenseNet 的密集连接来增加连通性和特征丰富性，或者采用 Wide-ResNets 或 ResNeXt 结构来改变 VIT 主干中的通道尺寸和头数。

论文探讨了从 CNN 到 VIT 的五种架构设计：

(i)密集连接如 DenseNet;

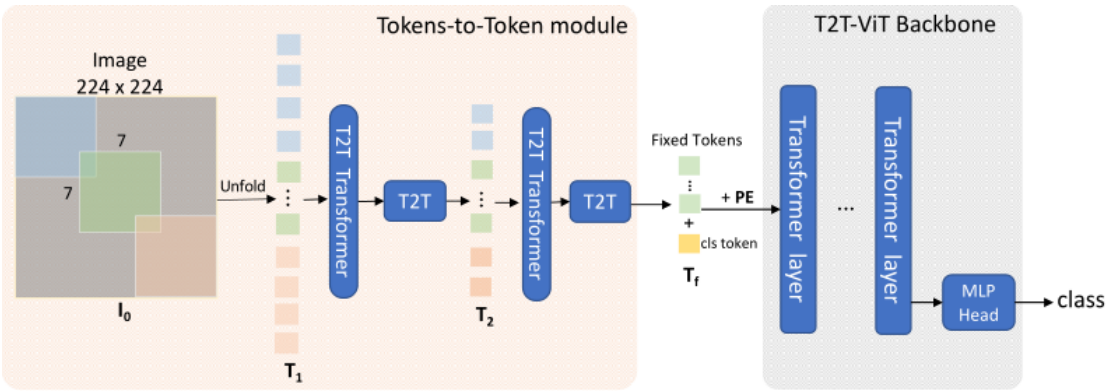
(ii)深-窄与浅-宽结构如宽 ResNet;

(iii)通道注意如挤压-激励(SE)网络;

(iv)多头注意层中更多的分头如 ResNext;

(v)Ghost 操作如 Ghost Net。

在研究了几种基于 CNN 的体系结构设计后，对主干采用深窄结构，以减少冗余度，提高特征丰富性。



4 实验结果分析

(1) T2T-ViT 与 ViT 在 ImageNet 上从头训练的比较。我们首先比较了 T2T-ViT 和 ViT 在 ImageNet 上的性能。结果如表所示。我们的 T2T-ViT 在参数和 mac 数量上比 ViT 小得多，但性能更高。例如，拥有 48.6M 和 10.1G mac 的小型 ViT 模型 ViT-s /16 在 ImageNet 上从头训练时具有 78.1%的 top-1 精度，而我们的 T2T-ViT-14 只有 44.2%的参数和 51.5%的 mac 则获得了超过 3.0%的改进(81.5%)。如果我们比较 T2T-ViT-24 和 ViT-L/16，前者减少了约 500%的参数和 mac，但在 ImageNet 上的改善超过 1.0%。将 T2T-ViT-14 与 DeiT-small 和 DeiT-small-distilled 进行比较，我们的 T2T-ViT 可以在没有大型 CNN 模

型作为教师的情况下获得更高的精度，以增强 ViT。
我们还采用更高的图像分辨率 384×384，并通过我们的 T2T-ViT↑384 获得 83.3%的精度

Models	Top1-Acc (%)	Params (M)	MACs (G)
ViT-S/16 [12]	78.1	48.6	10.1
DeiT-small [36]	79.9	22.1	4.6
DeiT-small-Distilled [36]	81.2	22.1	4.7
T2T-ViT-14	**81.5**	21.5	4.8
T2T-ViT-14↑384	**83.3**	21.5	17.1
ViT-B/16 [12]	79.8	86.4	17.6
ViT-L/16 [12]	81.1	304.3	63.6
T2T-ViT-24	**82.3**	**64.1**	**13.8**

(2) T2T-ViT 与 ImageNet 上 ResNets 的比较。为了进行公平的比较，我们建立了三个具有相似模型大小和 mac 的 T2T-ViT 模型，分别为 ResNet50、ResNet101 和 ResNet152。实验结果如表所示。本文提出的 T2T-ViT 比具有相似模型大小和 mac 的 ResNets 获得 1.4%-2.7%的性能提升。例如，与 ResNet50 的 255m 参数和 4.3G mac 相比，我们的 T2T-ViT-14 的 21.5M 参数和 4.8G mac 在 ImageNet 上获得了 81.5%的准确率。

Models	Top1-Acc (%)	Params (M)	MACs (G)
ResNet50 [15]	76.2	25.5	4.3
ResNet50*	79.1	25.5	4.3
T2T-ViT-14	**81.5**	21.5	4.8
T2T-ViT$_t$-14	**81.7**	21.5	6.1
ResNet101 [15]	77.4	44.6	7.9
ResNet101*	79.9	44.6	7.9
T2T-ViT-19	**81.9**	39.2	8.5
T2T-ViT$_t$-19	**82.2**	39.2	9.8
ResNet152 [15]	78.3	60.2	11.6
ResNet152*	80.8	60.2	11.6
T2T-ViT-24	**82.3**	64.1	13.8
T2T-ViT$_t$-24	**82.6**	64.1	15.0

(3) T2T-ViT 与 MobileNets 的比较。T2T-ViT-7 和 T2T-ViT-12 的

模型尺寸与 MobileNetV1 和 MobileNetV2 相似，但性能与 MobileNets 相当或更高。例如，我们的 T2T-ViT-12 参数为 6.9M，top1 精度为 76.5%，比 MobileNetsV2$_{1.4x}$ 高 0.9%。但我们也注意到，由于 transfomer 中的密集操作，T2T-ViT 的 mac 仍然比 mobilenet 大。但是，目前的 T2T-ViT-7 和 T2T-ViT-12 中并没有高效卷积等特殊的操作和技巧，我们只是通过降低隐维数、MLP 比和层深来减小模型尺寸，说明 T2T-ViT 作为 lite 模型也是很有前景的。总体而言，实验结果表明，我们的 T2T-ViT 在 ResNets 中等规模时能取得较好的性能，在 mobilenet 模型较小时能取得较好的结果。

Models	Top1-Acc (%)	Params (M)	MACs (G)
MobileNetV1 1.0x*	70.8	4.2	0.6
T2T-ViT-7	**71.7**	4.3	1.1
T2T-ViT-7-Distilled	**73.1**	4.3	1.1
MobileNetV2 1.0x*	72.8	3.5	0.3
MobileNetV2 1.4x*	75.6	6.9	0.6
MobileNetV3 (Searched)	75.2	5.4	0.2
T2T-ViT-12	**76.5**	6.9	1.8
T2T-ViT-12-Distilled	**77.4**	6.9	1.9

5 总结与展望

在这项工作中，我们提出了一种新的 T2T-ViT 模型，可以在 ImageNet 上从头开始训练，并实现与 cnn 相当甚至更好的性能。T2T-ViT 对图像的结构信息进行了有效的建模，增强了特征的丰富度，克服了 ViT 的局限性。它引入了新的令牌到令牌(T2T)过程，逐步将图像标记为令牌和结构化地聚合令牌。我们还从 cnn 中探索了提高 T2T - ViT 性能的各种架构设计选择，并通过实证发现深窄架构比浅宽结构性能更好。在 ImageNet 上从头开始训练时，我们的 T2TViT 的性能优于 ResNets，性能与模型大小相似的 mobilenet 相当。这为进一步开发基于 transformer 的视觉任务模型铺平了道路。