论文译文 S^2-MLP- Spatial-Shift MLP Architecture for Vision

论文译文 S^2-MLP: Spatial-Shift MLP Architecture for Vision

Abstract

最近，视觉转换器（Visual Transformer，ViT）及其后续作品放弃了卷积，利用自注意操作，达到了与 CNN 相当甚至更高的精度。最近，MLP-Mixer 放弃了卷积和自注意操作，提出了一种只包含 MLP 层的架构。为了实现跨块通信，在信道混合MLP的基础上，增加了令牌混合MLP。在超大规模数据集上进行训练时，取得了很好的效果。但是在中等规模的数据集(如ImageNet1K和ImageNet21K)上训练时，它的性能不如CNN和ViT。MLP - Mixer性能的下降促使我们重新思考令牌混合MLP。我们发现令牌混合MLP是深度卷积的变体，具有全局接收场和特定的空间配置。但全局接收场和空间特定的属性使得令牌混合 MLP 容易出现过度拟合。在本文中，我们提出了一种新颖的纯 MLP 架构，空间移位 MLP (S2-MLP)。与 MLP-Mixer 不同，我们的 S2-MLP 仅包含通道混合 MLP。我们利用空间移位操作来进行补丁之间的通信。它具有局部接收场并且与空间无关。它是无参数且计算高效的。在 ImageNet-1K 数据集上训练时，所提出的 S2-MLP 比 MLP-Mixer 获得了更高的识别精度。同时，S2-MLP 在 ImageNet-1K 数据集上实现了与 ViT 一样出色的性能，并且架构相当简单，并且 FLOP 和参数更少。

1 Introduction

在过去的几年里，卷积神经网络（CNN）[17, 11]在计算机视觉领域取得了巨大的成功。最近，受到 Transformer [32] 在自然语言处理方面取得的胜利的启发，提出了视觉 Transformer (ViT) [7]。它用 Transformer 中使用的自注意力操作取代了 CNN 中的卷积操作，以对图像不同空间位置的局部块之间的视觉关系进行建模。ViT和后续工作[30,36,33,21,10,35,31]取得了与CNN模型相当甚至更好的性能。与CNN要求对卷积核进行精心设计相

比，ViT只是简单地堆叠一系列具有相同设置的标准Transformer块，减少了手工操作并减少了归纳偏差。

最近，MLP-Mixer [28] 提出了一种完全基于多层感知器（MLP）的更简单的替代方案，以进一步减少归纳偏差。 MLP-Mixer 中的基本块由两个组件组成：通道混合 MLP 和令牌混合 MLP。通道混合 MLP 沿通道维度投影特征图，以实现不同通道之间的通信。同时，token-mixing MLP沿着空间维度投影特征图并实现空间位置之间的通信。当在 JFT-300M [25] 等大规模数据集上进行训练时，MLP-Mixer 获得了有希望的识别精度。但在中等规模数据集 ImageNet1K 和 ImageNet-21K 上，MLP-Mixer 和 ViT 之间仍然存在精度差距[6]。具体来说，Mixer-Base-16 [28] 在 ImageNet-1K 上仅实现了 76.44% 的 top-1 准确率，而 ViT-Base-16 [7] 实现了 79.67% 的 top-1 准确率。

MLP - Mixer在ImageNet - 1K和ImageNet - 21K上的表现不尽人意，促使我们重新思考MLP - Mixer中的混合标记MLP。给定矩阵形式的N个patch特征，X = [x1 , · · · , xN]，令牌混合MLP执行XW，其中W∈RN × M是可学习的权重矩阵。可以很直观地观察到，令牌混合MLP的输出XW的每一列都是块特征(输入X中的列)的加权求和。求和中的权重类似于Transformer中的注意力。但Transformer中的注意力是数据依赖的，而令牌混合MLP中求和的权重与输入无关。在一定程度上，加权求和更类似于深度卷积[3、15、16]。但是深度卷积只具有局部接收域。相比之下，令牌混合MLP具有全局接收场。此外，深度卷积核在不同的空间位置之间共享，而令牌混合MLP中求和的权重在不同的空间位置上是不同的。与深度卷积相比，令牌混合MLP不受局部接收场的限制，也不受空间无关性的限制，具有更高的灵活性和更强的拟合能力。但是，在避免断链的同时，也存在过度拟合的风险。为了避免令牌混合 MLP 的过拟合，需要提供大量的训练样本。这就解释了为什么在超大规模数据集 JFT-300M 上进行预训练后，MLP-Mixer 与 ViT 之间的识别准确率差距缩短了。

为了缓解 MLP-Mixer 在只有中等规模训练数据时的过拟合问题，我们提出了空间移动 MLP（S2-MLP）架构，这是一种概念上简单的架构，只包含信道混合 MLP。为了在空间位置之间进行通信，我们采用了空间移动操作，这种操作不需要参数，计算效率高。空间移位操作与空间无关，并保持本地接收场。图 1 展示了拟议 S2-MLP 的结构。它将一幅图像裁剪成w × h个互不重叠的图像块。对于每个块，它通过一个全连接层获得块嵌入向

量。块嵌入向量进一步经过N个S2 - MLP块。每个S2 - MLP块包含4个全连接层。每个S2 - MLP块中的全连接层与MLP - Mixer中使用的通道混合MLP具有相同的功能。但是我们的S2 - MLP不需要令牌混合MLP。相反，不同空间位置之间的通信是通过所提出的空间移位模块实现的。它是无参数的，简单地将通道从一个斑块转移到其相邻的斑块。尽管空间移位模块只支持相邻块之间的通信，但通过堆叠一系列S2 - MLP块，可以实现远距离通信。

所提出的S2 - MLP在结构上是简单的。在ImageNet1K数据集上，与MLP - Mixer相比，在参数规模和FLOPs相当的情况下，取得了相当高的识别准确率。同时，它在ImageNet1K数据集上以相当简单的结构、较少的参数和FLOPs达到了与ViT相当的识别精度。

2 Related Work

.....

3 Method

在这一部分，我们介绍了空间移位MLP (S2-MLP)。

1. Preliminary

层归一化（LN）[1]广泛用于使用 Transformer 和 BERT 架构的模型。给定 c 维向量 x = [x1, · · · , xc]，层归一化计算平均值 μ = 1 c Σci=1 xi 和标准差 σ = √ 1 c Σci=1(xi − μ)2。它通过 xi = γ xi−μ σ + β 对 x 中的每个条目进行归一化，其中 β 和 γ 是可学习参数。

高斯误差线性单元（GELU）[12]是 Transformer 和 BERT 模型中广泛使用的激活函数。其定义为 GELU(x) = xΦ(x)，其中 Φ(x) 是标准高斯累积分布函数，定义为 Φ(x) = 1 2 [1 + erf(x/√2)]。

MLP-Mixer [28]堆叠了N个相同大小和结构的基本块。每个基本块由两种类型的 MLP 层组成：通道混合 MLP 和令牌混合 MLP。让我们用 pi ∈ Rc 表示一个 patch 特征，并用 P = [p1, · · · , pn] ∈ Rc×n 表示具有 n 个 patch 特征的图像。 ChannelMixing MLP 沿通道维度投影 P：

其中W1∈R（C × C），W2∈Rc ×（C）.同时，令牌混合MLP投影了沿空间维度的通道混合斑块特征(P)：

式中：W3∈Rn × ~ N，W4∈R ~ N × N。

2. Spatial-Shift MLP Architecture

如图1所示，我们的空间移位MLP主干由一个路径全连接层，N个S2MLP块和一个用于分类的全连接层组成。由于我们引入的全连接层用于分类是众所周知的，我们只引入了patchwise全连接层和提出的空间移位块。空间移位操作的提出与Shift [34]，4-连通Shift [2]和TSM [19]密切相关。我们的空间移位操作可以看作是4 - Connected Shift的一个特殊版本，不需要原点元素信息。与fc - shift - fc结构中的4-连通shift残差块[2]不同，我们的S2 - MLP块，如图1所示，在fc - shift - fc结构之后，仅对混合通道取另外两层全连接层。此外，4 -连接移位残差网络在早期层使用卷积，而我们的网络采用了pure - MLP结构。

Patch - Wise全连接层。我们用I∈RW × H × 3表示一幅图像.将其均匀分割为w × h个斑块，P = { Pi } wh i = 1，其中Pi∈Rp × p × 3，w = W p，h = H p。对于每个面片Pi，我们将其展开为一个向量pi∈R3p2，并通过一个全连接层将其投影为一个嵌入向量ei，然后进行层归一化：

其中W0∈Rc × 3p2和b0∈Rc是全连接层的参数，LN（·)表示我们将要介绍的层规范化.

S2-MLP block.我们的架构堆叠了相同大小和结构的N个S2 - MLP。每个Spatial - shift模块包含4个全连接层、2个层归一化层、2个GELU层、2个跳跃连接以及提出的Spatial - shift模块。值得注意的是，在我们的S2 - MLP中使用的所有全连接层只起到混合通道的作用。在MLPMixer中，我们没有使用令牌混合MLP。由于全连接层是众所周知的，并且我们已经在上面介绍了层归一化和GELU，因此这里只关注提出的空间移位模块。我们用T∈ Rw×h×c 表示空间平移模块输入的特征图，其中 w 表示宽度，h 表示高度，c 是通道数。空间平移操作可分解为两个步骤：1) 将信道分成若干组，2) 将每组信道向不同方向移动。

Group.我们沿通道维度均匀分割 T，得到 g 个更薄的张量 {T τ }g τ=1 其中 T τ∈ Rw×h×c/g。值得注意的是，组数 g 取决于第二步中移动方向的设

计。例如，默认情况下，我们只沿四个方向移动，因此在此配置中，g 设置为 4。

Spatial-shift operation.我们将不同的组向不同的方向移动。对于第一组通道 T 1，我们将其沿宽维度移动+1。同时，我们将第二组通道 T 1 沿宽维度移动-1。同样，对于 T 3，我们将其沿高度方向移动+1，而对于 T 4，我们将其沿高度方向移动-1。我们在公式 (3) 中阐明了空间移动操作的公式，并在算法 1 中演示了伪代码。

经过空间平移后，每个斑块从其相邻的斑块中吸收视觉内容。空间移位操作是无参数的，使得不同空间位置之间的通信成为可能。上述空间移位方式是最简单、最直接的移位方法之一。我们还评估了其他的空间转移方式。令人惊讶的是，与其他方法相比，上述简单的方法取得了优异的性能。采用空间移位操作，不再需要令牌混合器作为MLP - Mixer。我们只需要通道混合器就可以将分块特征沿通道维度投影。值得注意的是，单个块中的空间移位操作只能获得相邻块的视觉内容，而不能获得图像中所有块的视觉内容。但是我们堆叠N个S2 - MLP块，全局的视觉内容会逐渐扩散到每个块。

3. Relations with depthwise convolution

Depthwise convolution.给定一个定义为张量T∈Rw × h × c的特征图，深度卷积[3、15、16]在张量T的每个二维切片[: , : , i]∈Rw × h上可分离地使用一个二维卷积核Ki，其中i∈[1、c]。深度卷积具有廉价的计算成本，因此被广泛应用于高效的神经网络中进行快速推理。

Relations.事实上，空间移位操作相当于一个具有固定和特定组的核权重的深度卷积。令K = { K1，· · ·，Kc }表示一组深度卷积核。如果我们设定

基于K组核的深度卷积等价于我们的空间移位操作。

也就是说，我们的空间平移操作是上面定义的固定权重的深度卷积的变体。同时，空间移位操作在每组通道内共享核权重。如引言部分所述，MLP - Mixer中的令牌混合MLP是深度卷积的全局接收和空间特异性变体。与我们的空间移位操作和香草深度卷积不同，令牌混合中求和的权重是针对特定空间位置的共享交叉通道。相比之下，香草深度卷积针对不同的通

道学习不同的卷积核，我们的空间移位操作共享组内的权重，对不同的组采用不同的权重。

我们在表1中总结了它们的联系和区别。观察空间移位操作和深度卷积之间的联系，我们鼓励研究人员尝试使用不同设置的深度卷积来构建新的基于MLP的架构。

4. Complexity Analysis

片段全连接层（PFL）将从原始图像 P∈Rp×p×3 中裁剪出的每个片段投影到 c 维特征向量中。PFL 的权值满足 W0∈ Rc×3p2 和 b0∈ Rc。因此，PFL 的参数数为

补丁总数为 M = w × h = W p ×H p，其中 W 是输入图像的宽度，H 是输入图像的高度。在这种情况下，PFL 的浮点运算（FLOPs）为

值得注意的是，根据以前的研究成果 [30, 10]，我们在计算 FLOP 时只考虑浮点数之间的乘法运算。

S2-MLP 模块拟议的 S2-MLP 视觉架构由 N 个 S2-MLP 模块组成。所有区块的输入和输出大小相同。我们用张量 T (i) in 表示第 i 个 S2-MLP 块的输入，用 T (i) out 表示其输出。那么，这些张量满足

所有 S2-MLP 模块都采用相同的操作和配置。这导致所有区块的计算成本和参数数量相同。要获得拟议 S2-MLP 架构的总参数数和 FLOPs 数，我们只需计算每个基本模块的参数数和 FLOPs 数。