论文译文 S^2-MLPv2- Improved Spatial-Shift MLP Architecture for Vision

论文译文 S^2-MLPv2: Improved Spatial-Shift MLP Architecture for Vision

Abstract

最近，基于MLP的视觉骨干出现了。与CNNs和视觉Transformer相比，基于MLP的视觉架构具有较小的感应偏差，在图像识别中获得了具有竞争力的性能。其中，空间移动 MLP（S2-MLP）采用了直接的空间移动操作，比 MLP-mixer、ResMLP 等先驱作品取得了更好的性能。最近，Vision Permutator (ViP) 和 Global Filter Network (GFNet) 使用金字塔结构的较小补丁，取得了比 S2-MLP 更好的性能。本文改进了 S2-MLP 视觉骨干。我们沿通道维度扩展特征图，并将扩展后的特征图分割成若干部分。我们对分割的部分进行不同的空间移动操作。同时，我们还利用了分割关注操作来融合这些分割部分。此外，与同行一样，我们也采用了较小尺度的斑块，并使用金字塔结构来提高图像识别精度。我们将改进的空间平移MLP视觉骨架称为S2 - MLPv2。使用55M的参数，我们的中等规模模型S2MLPv2 - Medium在ImageNet - 1K基准测试集上使用224 × 224图像，在没有自注意力和外部训练数据的情况下，取得了83.6 %的top - 1准确率。

1 Introduction

最近，人们对计算机视觉进行了广泛研究，以期在减少感应偏差的情况下实现高性能。出现了两类架构，包括视觉转换器（Dosovitskiy 等人，2021 年；Touvron 等人，2020 年）和基于 MLP 的骨干（Tolstikhin 等人，2021 年；Touvron 等人，2021a）。与具有精心设计的卷积内核的实际视觉主干 CNN （He et al.， 2016） 相比，视觉 Transformer 和基于 MLP 的主干在图像识别方面都取得了有竞争力的性能，而无需昂贵的手工设计。具体来说，视觉 Transformer 模型堆叠了一系列 Transformer 模块，实现了全局接收场。

基于MLP的方法，如MLP - Mixer (Tolstikhin et al , 2021)和ResMLP (图夫龙等, 2021a)，通过沿MLP实现的不同补丁的投影来实现补丁之间的通信。与 MLP-Mixer 和 ResMLP 不同，空间移动 MLP（S2-MLP）（Yu 等人，2021b）采用了一种非常简单的操作--空间移动，来实现补丁之间的通信，从而在 ImageNet1K 数据集上实现了更高的图像识别精度，而无需外部训练数据。与此同时，Vision Permutator（ViP）（Hou 等人，2021 年）沿高度和宽度维度对特征表示进行编码，同时利用两级金字塔结构对更精细的补丁尺寸进行利用，取得了比 S2-MLP 更好的性能。CCS-MLP（Yu 等人，2021a）设计了一种环状标记混合 MLP，以实现翻译不变性特性。全局滤波器网络（GFNet）（Rao 等人，2021b）利用二维傅里叶变换将空间斑块特征映射到频域，并在频域中进行跨斑块通信。正如 Rao 等人（2021b）所指出的，频域中的令牌混合操作等同于带有环状权重的深度卷积。为了达到较高的识别精度，GFNet 还利用了具有金字塔结构的较小尺寸补丁。最近，AS-MLP（Lian 等人，2021 年）轴向移动了特征图的通道，并设计了一个四层金字塔，取得了出色的性能。与此同时，Cycle-MLP（Chen 等人，2021a）为空间投影设计了多个伪核，也取得了出色的性能。值得注意的是，AS-MLP（Lian 等人，2021 年）和 Cycle-MLP（Chen 等人，2021a）都是基于精心设计的四级金字塔。

在这项工作中，我们重新思考了空间移动 MLP（S2-MLP）（Yu 等人，2021b）的设计，并提出了一种改进的空间移动 MLP（S2-MLPv2）。与原来的 S2-MLP 相比，主要从两个方面进行了改进：

- 如图 1 (b) 所示，我们沿通道维度扩展特征图，并将扩展后的特征图分割成多个部分。对于不同的部分，我们会进行不同的空间移动操作。我们利用分割关注操作（Zhang 等人，2020 年）来融合这些分割部分。

- 与现有的基于 MLP 的架构（如 ViP（Hou 等人，2021 年）、GFNet（Rao 等人，2021 年b）、ASMLP（Lian 等人，2021 年）和 Cycle-MLP （Chen 等人，2021 年a））一样，我们采用了较小尺度的斑块和分层金字塔结构。

我们将改进后的空间移动 MLP 架构称为 S2-MLPv2。图 1 显示了原始空间移动 MLP（S2-MLP）与改进后的 S2-MLPv2 之间的差异。我们在公共基准 ImageNet-1K 上进行的实验证明，所提出的 S2-MLPv2 具有最先进的图像识别准确性。具体来说，使用 5500 万个参数，我们的中型模型

S2-MLPv2-Medium 在使用 224 × 224 图像（无自我关注）和外部训练数据的情况下，达到了 83.6% 的最高准确率。

2 Related Work

……

3 PRELIMINARY

1. SPATIAL-SHIFT MLP (S2-MLP)

在这一部分中，我们简要回顾了S2 - MLP（ Yu et al , 2021b)体系结构。它由块嵌入层、S2 - MLP块堆栈和分类头组成。

Patch embedding layer.首先将一幅大小为W × H × 3的图像裁剪成w × h的小块。每个斑块大小为p × p × 3，p = W w = H h。然后通过全连接层将每个面片映射为d维向量。

Spatial-shift MLP block.如图2所示，它由用于混合通道的4个MLP层和用于混合面片的空间偏移层组成。下面我们只介绍Spatial - shift模块。给定一个输入张量X∈Rw × h × c，它首先沿通道维度将X等分为四部分{ Xi } 4i = 1，并沿四个方向平移：

值得注意的是，S2-MLP（ Yu et al , 2021b)堆叠了相同设置的N个Spatial - shift MLP块，并且没有像MLP -脊柱的对应结构如Vision转换开关(Hou et al , 2021)和Global Filter Network（ GFNet) (饶品贵等, 2021b)那样使用金字塔结构。

2. SPLIT ATTENTION

Vision Permutator（Hou 等人，2021 年）采用了 ResNeSt（Zhang 等人，2020 年）中提出的分割注意力来增强来自不同操作的多个特征图。具体来说，我们用 [X1, X2, - - - , XK] 表示大小相同的 n × c 的 K 个特征图，其中 n 是斑块数，c 是通道数。

其中，1∈Rn 是包含所有 1 的 n 维行向量。然后，a ∈ Rc 经过一堆 MLP，生成

其中，σ 是 GELU 实现的激活函数，W1∈ Rc×˜ c 和 W2∈ R˜ c×Kc 是 MLP 的权值，输出ˆ a∈ RKc。然后，ˆ A 被重塑为矩阵 ˆ A∈ RK×c，再由软最大函数沿第一维进一步处理，生成˜ A = softmax(ˆ A)∈ RK×c。然后生成被观测特征图 ˆ X，其中 ˆ X 的每一行 ˆ X[i, :] 都是通过以下方式计算的

其中，表示两个向量之间的元素相乘。

4 S2-MLPv2

在这一部分，我们介绍了我们提出的S2 - MLPv2架构。与S2 - MLP骨架类似，S2 - MLPv2骨架由块嵌入层、S2 - MLPv2块堆栈和分类头组成。由于我们在上一节中已经引入了patch嵌入层，下面我们只介绍本文提出的S2 - MLPv2块。

1. S2v2 Block

S2 - MLPv2块由S2 - MLPv2分量和通道混合MLP (CM-MLP)分量两部分组成。给定一个输入特征图X∈Rw × h × c，执行

通道混合MLP (CM-MLP)采用与MLP混频器(Tolstikhin et al , 2021)和Res MLP (图夫龙等, 2021a)相同的结构，这里略去其细节。下面我们只对本文提出的S2 - MLPv2组件进行详细介绍。

给定输入特征图X∈Rw × h × c，提出的S2 - MLPv2分量首先通过一个MLP将X的通道从c扩展到3c：

然后将沿通道维度扩展的特征图(X)等分为三部分：

它通过两个空间平移层SS1 (·)和SS2 (·)对X1和X2进行平移。SS1 (·)进行与方程1相同的空间平移操作。相反，SS2 (·)相对于SS1 (·)进行非对称的空间位移操作。具体来说，给定特征映射X2，SS2 (X2)进行：

值得注意的是，我们有意地将SS1 (·)和SS2 (·)设计成一个非对称结构，使它们互为补充。同时，我们不对X3进行移位，只保留X3。

然后，将{ Xk } 3k = 1重塑为矩阵{ Xk } 3k = 1，其中Xk∈Rwh × c，作为方程2、方程3和方程4输入到分裂注意力(split- attention，SA)模块中生

成

然后，被选中的特征图 A 被进一步输入到另一个 MLP 层，以生成输出结果

拟议的 S2-MLP 模块结构如图 3 所示，具体细节见算法 1。

2. PYRAMID STRUCTURE