MViTv2: Improved Multiscale Vision Transformers for Classification and Detection

陈加壹

摘要

本文对多尺度视觉 transformers(MViTv2) 进行了复现，MViTv2 是 MViT 的改进版本，主要有两个方面的改进，第一个是加入了分解的相对位置 embedding，第二个是在池化的 self-attention 计算中加入了查询张量 q 的残差连接来补偿池化产生的信息损失。在 ImageNet-1k 数据集上进行了训练，对其分类的准确率为 82.16%，原论文的准确率为 88.8%，与原论文相比低了 6.6 个百分点。

关键词：MViTv2；ImageNet-1k

1 引言

基于 self-attention 的体系结构，特别是 Transformers，已经成为自然语言处理 (NLP) 的首选模型。主要的方法是在大型文本语料库上进行预训练，然后在较小的特定任务数据集上进行微调。由于 Transformers 的计算效率和可伸缩性，可以训练规模空前的模型。随着模型和数据集的增长，仍然没有表现饱和的迹象。然而，在计算机视觉中，卷积体系结构仍然占据主导地位。受 NLP 成功的启发，多个工作尝试将类似 CNN 的架构与自我关注相结合，有些工作完全取代了卷积。后一种模型虽然理论上是有效的，但由于使用了专门的注意模式，还没有在现代硬件加速器上有效地扩展。因此，在大规模图像识别中，经典的 Resnetlike 架构仍然是最先进的。尽管如此，从 2020 年年底开始，Transformer 还是在 CV 领域中展现了革命性的性能提升。目前 ViT 在诸多领域展现出可以与 CNN 相媲美的性能，由于 self-attention 的计算复杂度是分辨率的平方，所以即使 ViT 在图像分类中表现优异，但是将其应用在高分率目标检测和时空理解任务中依然具有挑战性，目前主流的两种优化方法是：(1) 在窗口内计算局部注意力；(2) 在视频任务中使用池化注意力聚合局部特征。后者启发了多尺度视觉 transformers(MViT) 的提出，这是一种简单易扩展的框架：它在整个网络中没有使用固定的分辨率，而是使用了从高分辨率到低分辨率的多个阶段的特征层次。在本篇文章中，作者提出了两个简单的技术改进来进一步提高其性能：(1) 使用可分解的相对位置编码来注入位置信息；(2) 使用残差池化连接来弥补 self-attention 计算中步长的损失。

2 相关工作

2.1 CNNs

卷积神经网络是一种前馈型神经网络, 受生物自然视觉认知机制启发而来的。现在，CNNs 已经成为众多科学领域的研究热点之一，特别是在模式分类领域，由于该网络避免了对图像的复杂前期预处理，可以直接输入原始图像，因而得到了更为广泛的应用。可应用于图像分类，目标识别，目标检测，语义分割等等。

2.2 Vision transformers

ViT 提出了视觉 transformer 的概念，将 transformer 的体系结构应用于图像 patch 上，并在图像分类上显示出极具竞争力的结果。Vit 模型会将图像分成固定大小的 patch，然后通过线性变换 embed 每个 patch，并且添加上位置 embedding，用来表征每个 patch 在图像中的绝对位置，为了进行分类，在序列中添加一个额外的可学习的 class token。然后将得到的向量序列提供给 transformer 编码器进行特征提取。用于分类的序列始终在序列的第一个位置，最后只需要将这个序列通过一个 MLP 层的分类头就可以实现对原图像的分类输出。Tranformer 编码器主要是通过 self-attention 机制来实现特征提取，可以理解为计算序列中的相关性，然后通过相关性的不同进行序列的权重相加来表征这幅图像的信息。编码器使用了多头注意力机制，把原始的 q、k、v 张量分割后分别进行 attention 再将结果进行拼接。这样可以让它学到多重含义的表达。

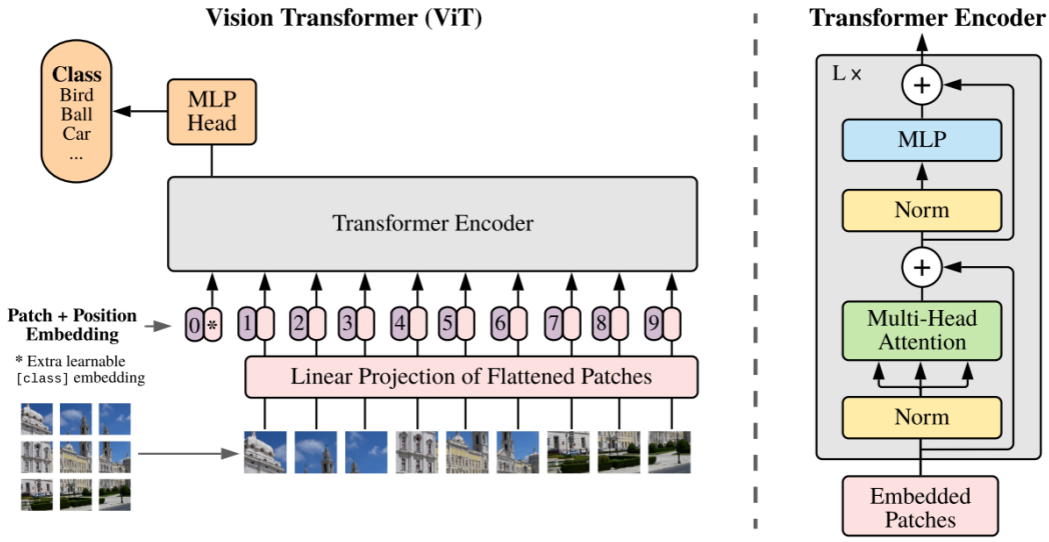

图 1: Vision Transformer

自 Vit 的工作以来，视觉 transformer 已经引起了人们的极大兴趣。此后，为了进一步改进 ViT，人们开展了不同的工作，包括高效的训练策略、多尺度 transformer 结构和更先进的自注意力机制。

2.3 Multiscale Vision transformers

Mvit 模型是作者提出的对 vit 模型的优化，原始的 vit 模型的 tranformer 输入和输出的序列尺寸是相同，Mvit 也就是多尺度视觉 tranformer，使用不同尺度的 transformer，逐级增加中间潜在序列的通道容量，同时减小其长度，从而减小空间分辨率，让通道得到更丰富的特征。前边层用高空间分辨率操作，来模拟简单的低级视觉信息，而更深的层以空间粗糙但复杂的高维特征操作。这样的好处是一方面可以在低分辨率的情况下降低计算需求，另一方面是低分辨率的情况下可以更好地理解场景的 context，从而指导高分辨率下的处理。

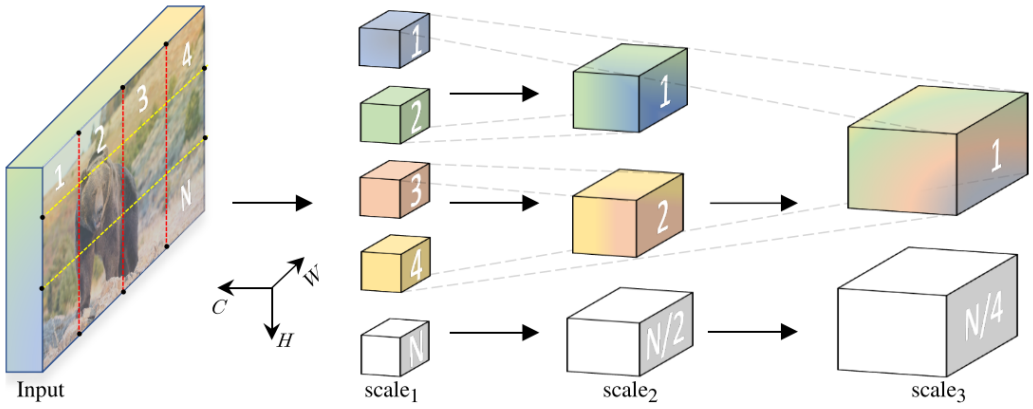

图 2: Multiscale Vision Transformer

实现分辨率跨级的基础就是在 attention 中加入对张量的池化操作，K 和 V 张量上的步幅比 Q 张量的步幅大，Attention 输出的尺寸和池化后的 Q 张量相同，所以 Q 张量只有在输出序列的分辨率跨级变化时才被下采样。这样的话，通过池化 q 张量，原始的输入就可以得到减小分辨率后的输出。

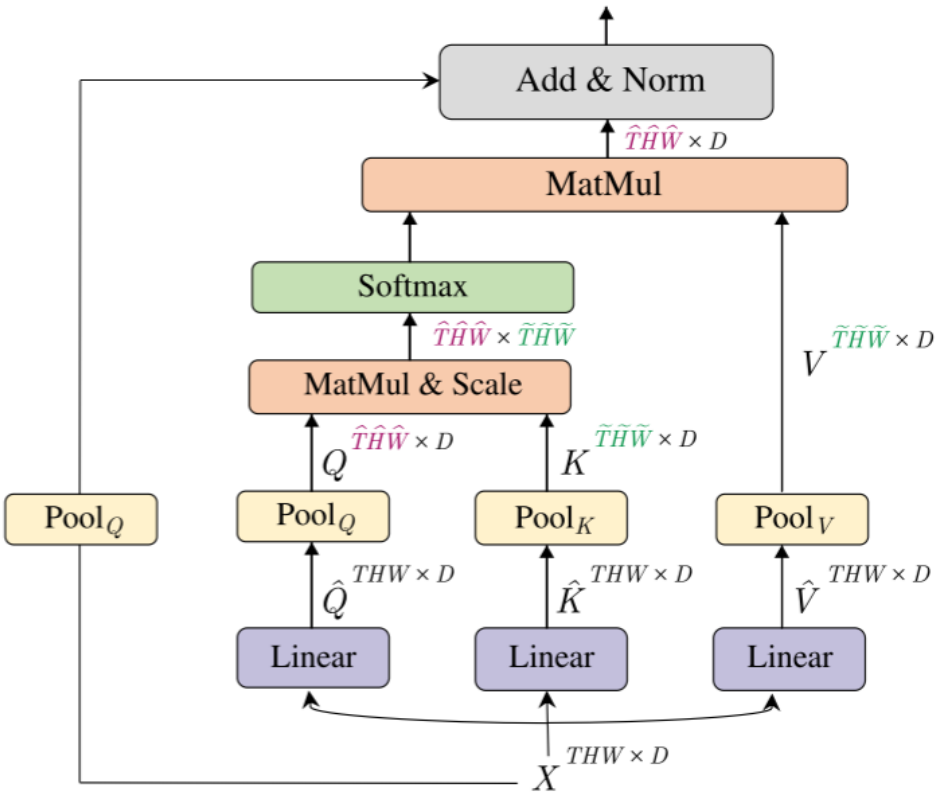

图 3: Pooling Attention

3 本文方法

3.1 本文方法概述

这篇文章主要对 MViT 中的池化自注意力计算模块做了两项改进：此部分对本文将要复现的工作进行概述，图的插入如图 5所示：

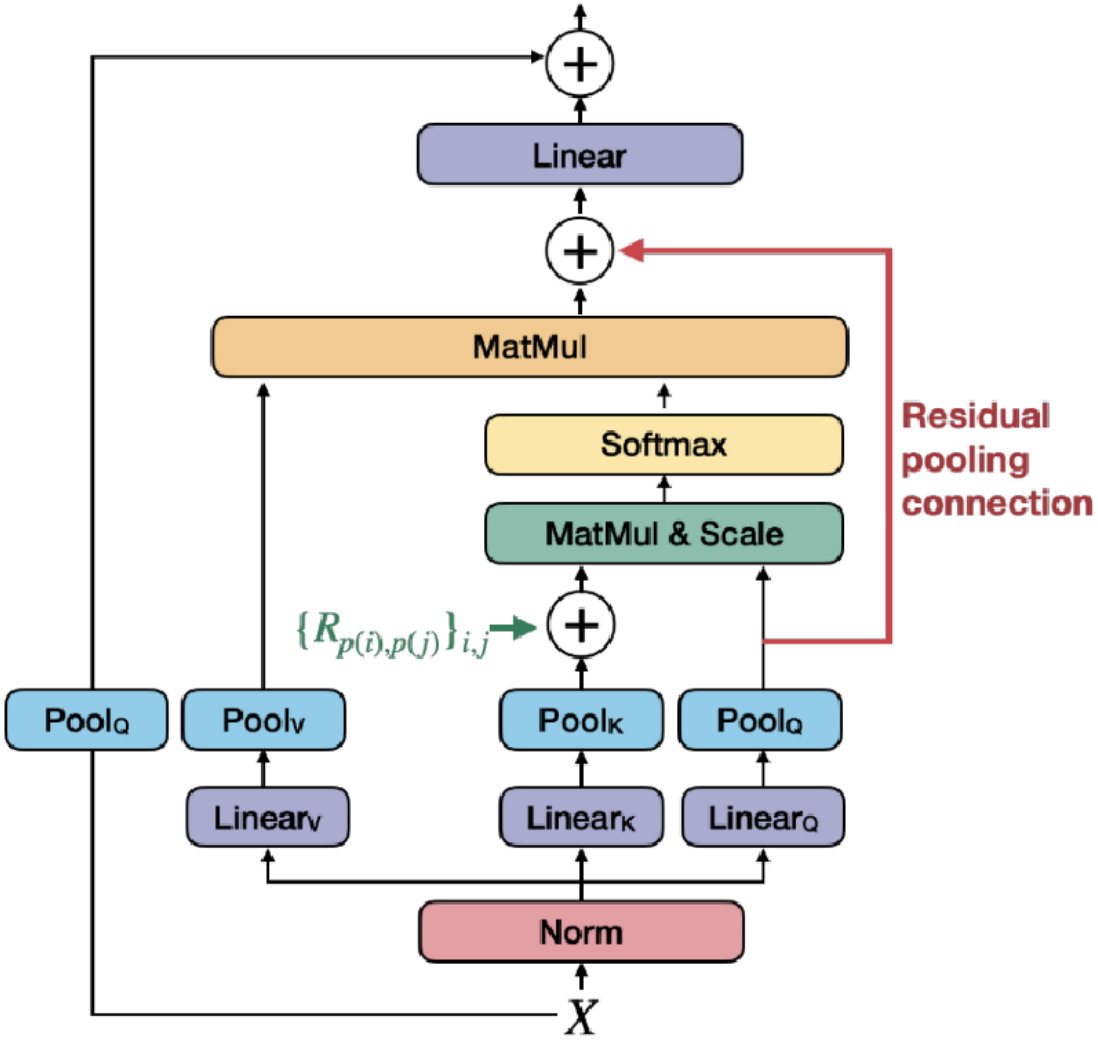

图 4: 将分解相对位置 embedding 和池连接模块结合在注意块中的改进池注意机制

3.2 分解的相对位置 embedding

虽然 MViT 在捕捉 token 之间的关系方面已经展示了优异的性能，但是这种注意力计算更关注内容而不是结构。时空结构建模完全依靠“绝对”位置 embedding 来提供位置信息，这忽略了视觉中平移不变性的基本原理。也就是说，即使 MViT 中两个 patch 的相对位置保持不变，它们之间相互作用的方式也会随着它们在图像中的绝对位置而改变。为了解决这个问题，文章引入了相对位置 embedding，它只依赖于 token 之间的相对位置距离。我们将两个输入元素 i 和 j 之间的相对位置编码为位置嵌入，然后将成对编码表示嵌入到自注意力模块中：

$$Attn(Q,K,V) = Softmax((QK^{\top} + E^{(rel)})/\sqrt{d})V, where \quad E_{ij}^{(rel)} = Q_i \cdot R_{p(i),p(j)} \tag{1}$$

然而，O(TW) 中可能的计算代价较高。为了降低复杂度，我们将元素 i 和元素 j 之间的距离计算沿着空间轴进行分解：

$$R_{p(i),p(j)} = R^{h}_{h(i),h(j)} + R^{w}_{w(i),w(j)} \tag{2}$$

其中 Rh、Rw 是沿高度、宽度的位置嵌入，以及 h(i)、w(i) 分别表示 token i 的垂直和水平位置。相比之下，使用分解嵌入将学习嵌入数减少到 O(t+w)，这对早期高分辨率特征映射有很大的效果。

3.3 残差池化连接

MViTv1 中引入的 MSPA 池化注意力可以大大减少 SA 的计算量，主要会在 Q-K-V 进行线性映射后在进行一步池化操作，MViTv1 在 K 和 V 张量上的步幅比 Q 张量的步幅大，Q 张量只有在输出序列的分辨率跨级变化时才会被下采样。这就需要在 pooling attention module 的计算中加入残差连接来增加信息流动。文章在注意力模块中引入一种新的残差池化连接，表示为以下公式：

$$Z := Attn(Q, K, V) + Q \tag{3}$$

4 复现细节

4.1 与已有开源代码对比

与开源代码相比引入了了卷积操作，在每个阶段对特征图进行一次卷积操作。保证每个阶段都能减小特征图的尺寸，增加特征图通道数，使得 token 能够在越来越大的空间维度上表示越来越复杂的视觉模式。同时增加局部注意力机制，在自注意力机制中，每个 token 都是跟所有 token 进行交互，利用卷积的思想让局部的特征图之间进行自注意力机制的运算，使得网络增加了卷积操作的平移不变性以及越相邻的像素之间相关性越高的先验知识。

4.2 实验环境

实验在 vscode 平台上进行，使用 python 语言进行代码的编写。模型训练在服务器上进行，利用 Anaconda 进行服务器上 python 环境的管理。使用 8 张 3090 的显卡进行模型的训练，数据集采用 ImageNet-1k，包含了 1000 个分类。

4.3 创新点

虽然 Transformer 具备动态关注、全局上下文和更好的泛化等优点，但是对细节和局部特征的提取能力不强。所以将 CNN 和 Transformer 结合是一个好方法。Transformer 的效果之所以比 CNN 好，是因为它对于距离较远的两个 patch 之间的交互能力要高于卷积，而卷积有越相近的像素点之间的关联性越高的先验，所以将两者结合起来，先使用卷积神经网络提取低层特征，再使用 transformer 模块来弥补卷积对于相隔较远的 patch 之间交互能力的不足，同时卷积还具有共享权值和空间下采样，能够让网络更好地适应较高分辨率的图像，这可能对于 Transformer 应用于高分辨率图像的目标检测任务来说是至关重要的。

5 实验结果分析

在服务器上使用 8 张 3090 的 gpu 进行训练，训练集包含了 1281167 张图片，测试集包含了 50000 张图片，训练了 300 个 epoch 大概使用了 3 天的时间。经过测试，模型在分类任务上的准确率为 82.16%，原作者的实验中，分类任务准确率达到了 88.8%，与原论文相比降低了 6.6 个百分点。

使用模型对图像进行分类，结果显示模型可以对图像进行正确的分类:

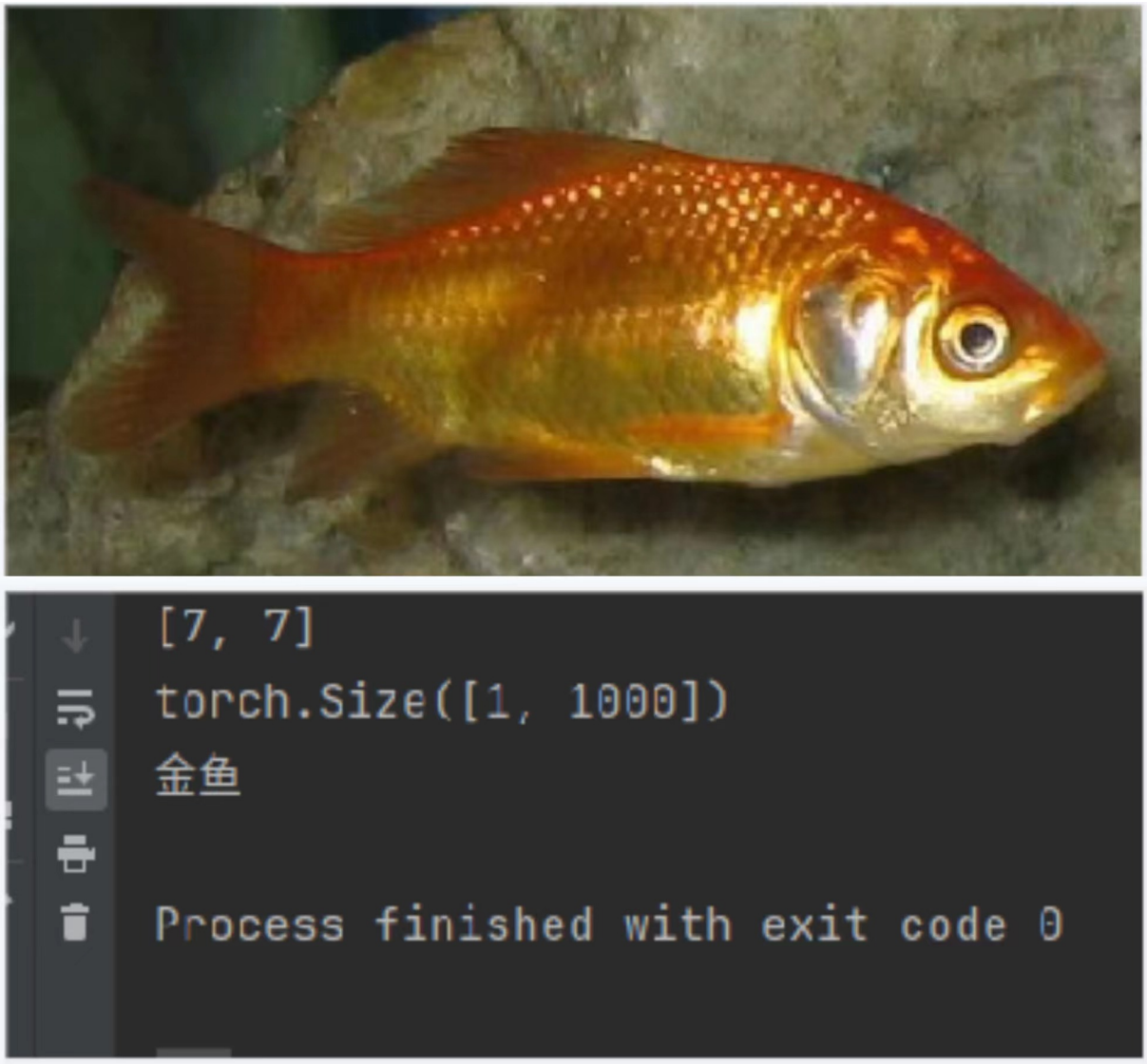

图 5: 使用模型对图像进行分类

6 总结与展望

本文对视觉 transformers 的模型结构和多尺度视觉 transformers(MViTv2) 的改进内容以及创新点进行了说明，对其复现后在 ImageNet-1k 数据集上进行了训练，最终分类的准确率为 82.16%，原论文的准确率为 88.8%，与原论文相比低了 6.6 个百分点。验证了 MViTv2 在图像分类上的任务具备一定的有效性，希望今后能对这个视觉识别任务进行进一步的研究。可以尝试在网络中加入语义关系的推理方法，在 few-shot 的情况下增加网络学习新概念的能力。