MViTv2: Improved Multiscale Vision Transformers for Classification and Detection

Yanghao Li Chao-Yuan Wu Haoqi Fan

摘要

将多尺度特征层次结构的思想与 Transformer 模型相结合，提出了用于视频和图像识别的多尺度视觉 Transformer(MVIT)。多尺度 Transformer 有几个通道分辨率尺度级。从输入分辨率和较小的信道维数开始，分级扩展信道容量，同时降低空间分辨率。这创建了一个多尺度的特征金字塔，前几层以高空间分辨率操作，以模拟简单低维视觉信息，而高层以低空间分辨率操作，以模拟高维负责的视觉信息。

关键词：多尺度 transformer；高分辨率图像处理

1 引言

视觉加工层次结构的主要方面已经被建立起来：(1) 随着处理层次结构的上升，降低空间分辨率。(2) 增加通道数量，每一个通道都对应着更加高维的特征信息。同时计算机视觉开发了多尺度处理，有时被称为“金字塔”策略。有两个动机 (1) 在低分辨率下工作可以降低计算需求。(2) 在低分辨率下可以更好地理解“上下文”，也就是全局信息，从而指导高分辨率下的视觉信息处理。

Transformer 体系结构允许学习定义在集合上的任意函数，并且在语言理解和机器翻译等序列任务中获得了可扩展的成功。实质上，一个 transformer 使用拥有两个基本操作的块。首先，是一个用于建模元素间关系的注意操作。其次，是一个多层感知机 (MLP)，他对元素内的关系进行建模。将这些操作与归一化和残差链接融合起来，使 transformer 可以推广到各种各样的任务。最近，Transformer 被应用于图像分类等关键计算机视觉任务上，Vision Transformer 达到了卷积模型在各种数据和计算体系中的性能，而模型仅仅在第一层将输入图片达成”patch”，之后拼接上多个 transformer blocks 即可。Vision Transformer 旨在展示 Transformer 架构使用很少的偏置便可达到很好的效果。

我们的意图是将多尺度特征层次思想与 Transformer 模型联系起来,提出多尺度视觉 Transformer(MVIT)。与传统 transformer 不同，传统 transformer 在整个网络中保持恒定的通道容量和分辨率，多尺度 transformer 有多个通道分辨率“尺度”级。从图像分辨率和较小的通道信息，分级扩展通道容量，同时降低空间分辨率。这在 transformer 网络中创建了一个多尺度的特征激活金字塔，有效地将 transformer 的原理与多尺度特征层次联系起来。

2 相关工作

2.1 卷积神经网络

ConvNets 结合了下采样、移位不变性和共享权重，是用于图像[1][2][3][4] 和视频[5][6][7][8]的计算机视觉任务的标准骨干。

2.2 注意力机制

自我注意机制已被用于图像理解[9]、无监督物体识别[10]以及视觉和语言[11]。自注意操作和卷积网络的混合也被应用于图像理解[12]和视频识别[13]。

2.3 Vision Transformer

目前将 Transformer[14]应用于视觉任务的大部分热情始于 Vision Transformer(VIT)[15]和 Detection Transformer。我们直接在 VIT 的基础上建立了一个允许通道扩展和分辨率下采样的分阶段模型。DeiT[16]提出了一种数据有效的 VIT 训练方法，我们的训练方法建立在相同设置下。一个新兴的工作路线旨在将 transformer 应用于视觉任务，如对象检测[17]、语义分割[18]、姿态估计、图像检索、点云、视频实例分割、对象重新识别、视频检索、视频对话、视频对象检测和多模态任务。

3 本文方法

3.1 本文方法概述

多尺度 transformer 架构建立在阶段这个核心概念之上，每个级是由多个具有特定时空分辨率和信道维数的 transformer 块组成。多尺度 transformer 的主要思想是逐步扩大信道容量，同时池化网络从输入到输出的分辨率，图的插入如图 1所示：

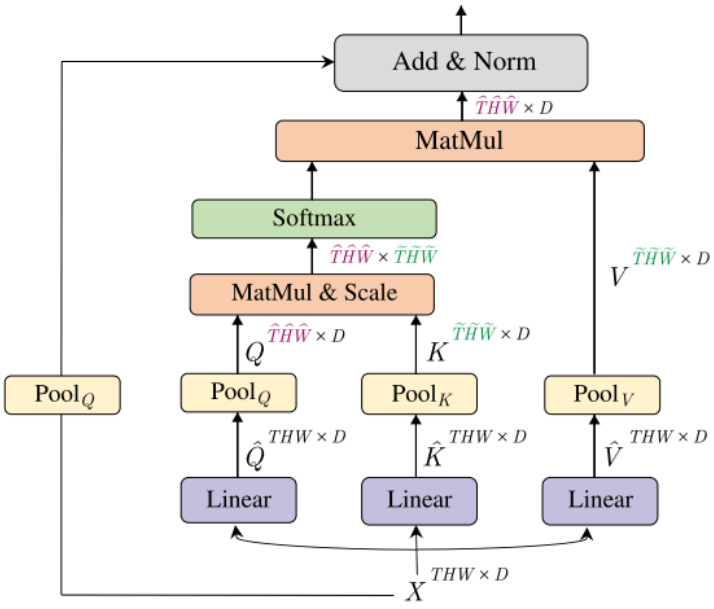

图 1: 方法示意图

3.2 多头集中注意机制

多头集中注意力机制 (Multi Head Pooling Attention)(MHPA), 是一个自注意力操作，它能够在一个 transformer 块中灵活地建立分辨率模型，允许多尺度 transformer 在逐渐变化的时空分辨率下工作，与最初的多头注意力 (MHA) 操作不同，其中通道维数和时空分辨率保持不变，MHPA 汇集潜在张量序列以减少参与输入的序列长度（分辨率）具体来说，考虑序列长度为 L，输入张量 X 维度是 D，$X \in \mathbb{R}^{L \times D}$。在 MHA 的基础上，MHPA 通过线性运算将输入 X 投影到查询张量 $\hat{Q} \in \mathbb{R}^{L \times D}$，键值张量 $\hat{K} \in \mathbb{R}^{L \times D}$，和值张量 $\hat{V} \in \mathbb{R}^{L \times D}$。$W_Q, W_K, W_V$ 的维度均为 $D \times D$。然后使用池化算子（P）分别处理查找，键和值张量：

$$Q = P_Q(XW_Q) \quad K = P_K(XW_K) \quad V = P_V(XW_V)$$

池化操作完成后，序列长度就会由原来的 L 减少到 $\tilde{L}$，Q $\in \mathbb{R}^{\tilde{L}\times D}$，K 键矩阵与 V 值矩阵也类似。之后继续进行注意力操作：

$$Z := Attn(Q,K,V) = Softmax(QK^T/\sqrt{D})V$$

计算输出序列 $Z \in \mathbb{R}^{\tilde{L}\times D}$。集中注意力可以通过池化张量 Q 来降低多尺度 transformer 不同阶段的分辨率，并通过池化键张量 K 和值张量 V 来显著降低计算和内存复杂度。

3.3 分解相对位置嵌入

虽然 MVIT 在建模 tokens 之间交互的能力方面显示出了希望，但它们关注的是内容，而不是结构。时空结构建模完全依靠“绝对”位置嵌入来提供位置信息。这忽略了视觉中平移不变性的基本原理 [47]。也就是说，即使它们的相对位置保持不变，MVIT 对两个贴片之间相互作用的建模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，我们将相对位置嵌入 [65]，在集中注意力机制中它只依赖于令牌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 $R_{p(i),p(j)} \in \mathbb{R}^d$，其中 p(i) 和 p(j) 代表元素 i 和元素 j 的空间位置，然后将成对的编码表示嵌入到自关注模块中：

$$Attn(Q,K,V) = Softmax((QK^T + E^{(rel)}/\sqrt{d})V$$

$$where E_{ij}^{(rel)} = Q_i \cdot R_{p(i),p(j)}$$

然而，可能的嵌入 $R_{p(i),p(j)}$ 的数量达到 O(TWH), 计算代价较高。为了降低计算复杂度，我们将元素 i 和元素 j 之间的距离沿着时空轴进行分解：

$$R_{p(i),p(j)} = R^h_{h(i),h(j)} + R^w_{w(i),w(j)} + R^t_{t(i),t(j)}$$

R^h, R^w, R^t 分别是沿高度、宽度和时间轴的位置嵌入，$h(i), w(i), t(i)$ 分别表示 token i 的垂直、水平和时间位置。值得注意的是 R_t 是可选的，并且只有在视频情况下支持时间维度时才需要。相比之下，我们的分解嵌入将学习嵌入数减少到 O(T+W+H)，这对前面几个阶段高分辨率特征映射有很大的效果。

3.4 残差池化连接

正如 [21] 所证明的，集中注意力对于减少注意力块中的计算复杂度和内存需求是非常有效的。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被下采样。这促使我们增加剩余的池连接与（池化的）Q 张量，以增加信息流，并促进训练池注意块在 MVIT。

我们在注意块中引入了一个新的剩余池连接，如图 2所示. 具体地说，我们将池化的查询张量添加到输出序列 Z, 所以注意力机制公式改为：

$$Z := Attn(Q,K,V) + Q$$

注意输出序列 Z 具有与池查询张量 Q 相同的长度

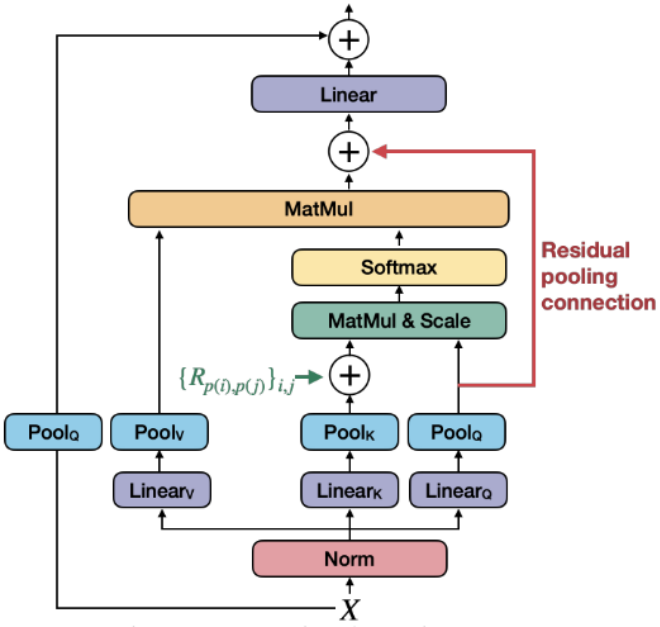

图 2: 方法示意图

4 复现细节

4.1 与已有开源代码对比

注意力机制中，查询矩阵和键矩阵的乘积 QK^T 是核心步骤，文章对 Query 矩阵使用池化操作从而实现多头集中注意力机制。为了降低计算量而忽略掉 Query 矩阵中的细节，是一种用效率换精度的一个做法。我提出的改进是将 PoolQ 操作换成 Linear 操作，在 Query 矩阵上模拟卷积操作，不仅实现 PoolQ 中的缩小分辨率的操作，实现多尺度的需求，同时也考虑到了 Query 矩阵元素的局部细节，不至于全部丢失。这种做法虽然降低了计算速度，但是精度会有一定的提升。如图 3所示。

MVIT 对两个贴片之间相互作用的建模方式也会随着它们在图像中的绝对位置而改变。因而文章将相对位置嵌入，在集中注意力机制考虑到了 token 之间的相对位置距离。元素 i 和元素 j 之间的距离沿着时空轴进行分解在视频作业任务上表现优秀，但是在单张图片识别中可以将相对位置距离设置的更加精确，思想同样是牺牲计算速度换取精度，在单张图片中我们采用欧式距离来替代时空轴分解后的位置，使得在处理图片任务时精度有一定的提升。如图 3所示。

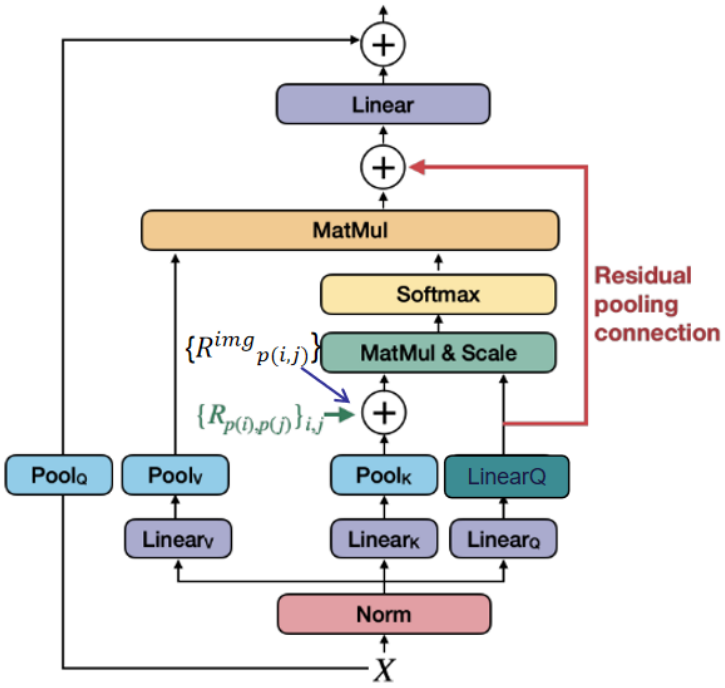

图 3: 改进方法示意图

4.2 实验环境搭建

需下载 1.7 版本的 Pytorch 和对应版本的 torchvision，安装 Facebookresearch 的 fvcore、FairScale，以及 iopath、simplejson、psutil，并下载 ImageNet1K 作为训练集。

4.3 创新点

根据不同下游任务搭建不同分解相对位置嵌入。

保留 Query 矩阵细节信息，增大 Attention 运算中的信息交互。

5 实验结果分析

Top-1 error：模型预测图片的类别，且模型只输出 1 个预测结果，这个结果错误的概率则成为 Top-1 error。Top-5 error：模型预测图片的类别，但模型会输出来 5 个预测结果，预测输出的这五个结果里没有正确结果概率则成为 Top-5 error。对模型进行训练 300 个 epoch，在第 120 个 epoch 时，损失等于 4.5，学习率调整到 0.00023，top1err 等于 29，top5err 等于 16。在第 180 个 epoch 时，损失等于 3.7，学习率等于 0.00014，top1err 等于 19，top5err 等于 10。如图 4所示。最后在测试中，top1err 等于 15，top5err 等于 4。如图 5所示。

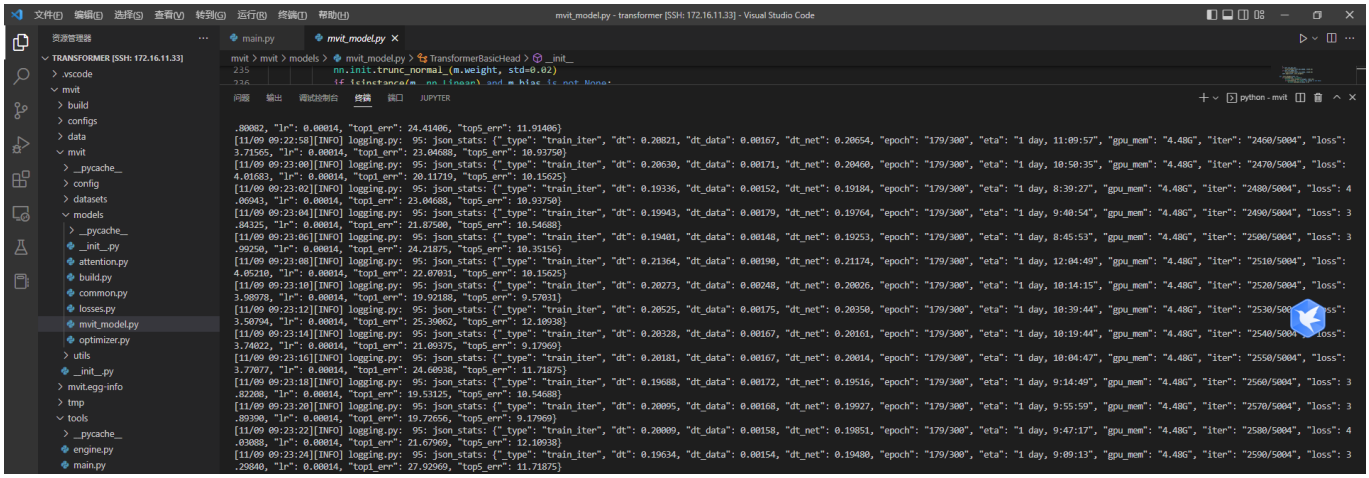

图 4: 训练示意图

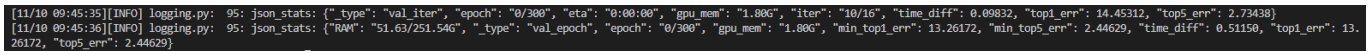

图 5: 测试结果示意图

6 总结与展望

图像具有尺度上的问题。因为比如说现在有一张街景的图片，里面有很多车和行人，里面的物体都大大小小，那这时候代表同样一个语义的词，比如说行人或者汽车就有非常不同的尺寸，另外一个挑战是图像的 resolution 太大了，如果要以像素点作为基本单位的话，序列的长度就变得高不可攀，所以说之前的工作要么就是用后续的特征图来当做 Transformer 的输入，要么就是把图片打成 patch 减少这个图片的 resolution，要么就是把图片画成一个一个的小窗口，然后在窗口里面去做自注意力，所有的这些方法都是为了减少序列长度。MVIT 作为多尺度 transformer，减少空间分辨率并增加通道信息，有效的启发了人们如何处理高分辨率图像，但是其计算规模还是随着图像的尺寸平方级增长，而 Swin transformer 提出了 hierarchical Transformer，它的特征是通过一种叫做移动窗口的方式学来的。移动窗口的好处：不仅带来了更大的效率，因为跟之前的工作一样，现在自注意力是在窗口内算的，所以这个序列的长度大大的降低了；同时通过 shifting 移动的这个操作，能够让相邻的两个窗口之间有了交互，所以上下层之间就可以有 cross-window connection，从而变相的达到了一种全局建模的能力。然后作者说这种层级式的结构不仅非常灵活，可以提供各个尺度的特征信息，同时因为自注意力是在小窗口之内算的，所以说它的计算复杂度是随着图像大小而线性增长，而不是平方级增长，从而让他们可以在特别大的分辨率上去预训练模型。因为 Swin Transformer 拥有了像卷积神经网络一样分层的结构，有了这种多尺度的特征，所以它很容易使用到下游任务里。

参考文献

[1] GAO S H, CHENG M M, ZHAO K, et al. Res2net: A new multi-scale backbone architecture[J]. IEEE transactions on pattern analysis and machine intelligence, 2019.

[2] TAN M, LE Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]// International conference on machine learning. 2019: 6105-6114.

[3] RADOSAVOVIC I, KOSARAJU R P, GIRSHICK R, et al. Designing network design spaces[C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 10428-

10436.

[4] ZHANG H, WU C, ZHANG Z, et al. Resnest: Split-attention networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 2736-2746.

[5] GIRDHAR R, CARREIRA J, DOERSCH C, et al. Video action transformer network[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 244-253.

[6] FEICHTENHOFER C. X3d: Expanding architectures for efficient video recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 203-213.

[7] ZHOU B, ANDONIAN A, OLIVA A, et al. Temporal relational reasoning in videos[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 803-818.

[8] JIANG B, WANG M, GAN W, et al. Stm: Spatiotemporal and motion encoding for action recognition [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 2000-2009.

[9] ZHAO H, JIA J, KOLTUN V. Exploring self-attention for image recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 10076-10085.

[10] LOCATELLO F, WEISSENBORN D, UNTERTHINER T, et al. Object-centric learning with slot attention[J]. Advances in Neural Information Processing Systems, 2020, 33: 11525-11538.

[11] LU J, BATRA D, PARIKH D, et al. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks[J]. Advances in neural information processing systems, 2019, 32.

[12] HU H, GU J, ZHANG Z, et al. Relation networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 3588-3597.

[13] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7794-7803.

[14] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[15] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[16] TOUVRON H, CORD M, DOUZE M, et al. Training data-efficient image transformers & distillation through attention[C]//International Conference on Machine Learning. 2021: 10347-10357.

[17] BEAL J, KIM E, TZENG E, et al. Toward transformer-based object detection[J]. arXiv preprint arXiv:2012.09958, 2020.

[18] ZHAO H, JIANG L, JIA J, et al. Point transformer[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 16259-16268.