基于结构嵌套的视觉Transformer网络

Transformer in Transformer. NeurIPS 2021 (CCF-A)

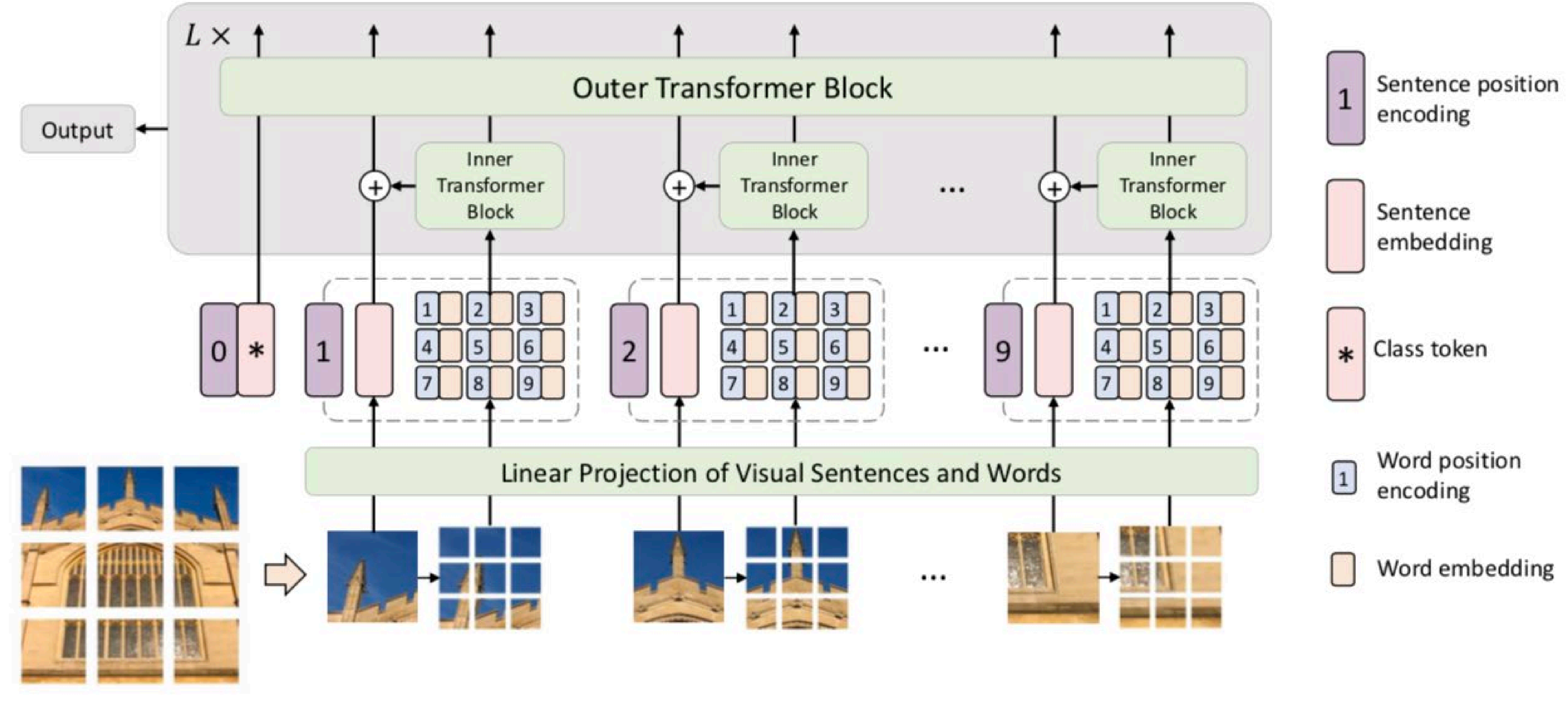
韩凯 吴恩华

摘要

Transformer 网络推动了诸多自然语言处理任务的进步，而近期 transformer 开始在计算机视觉领域崭露头角。本文提出一种基于结构嵌套的 Transformer 结构，图像块内部结构信息，在计算机视觉任务中表现好于传统Transformer。

方法

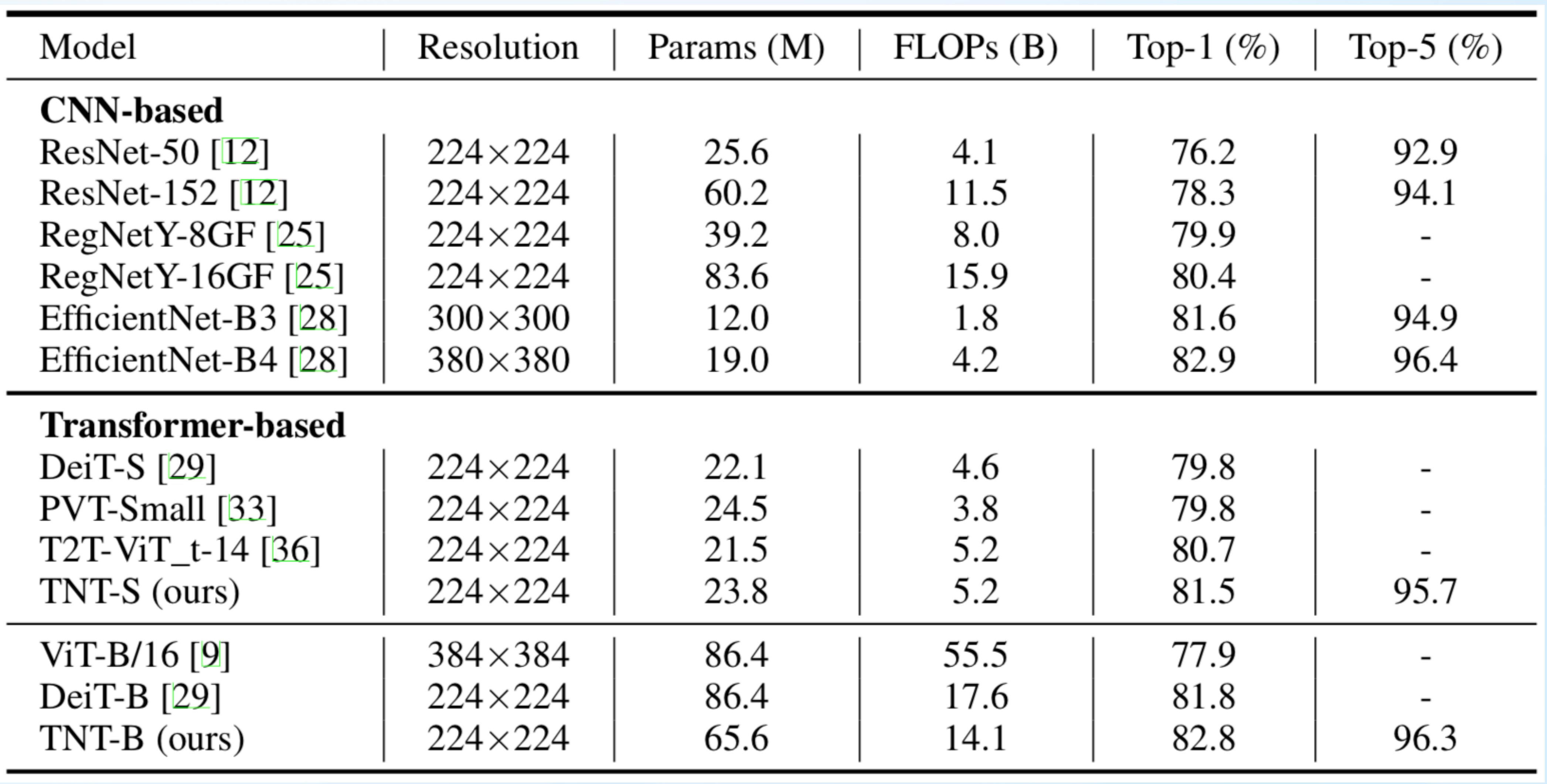
我们提出一种用于基于结构嵌套的 Transformer 结构，被称为 Transformer-iN-Transformer (TNT) 架构。如图1所示，TNT 将图像切块，构成图像块序列。不过，TNT 不把图像块拉直为向量，而是将 图像块看作像素（组）的序列。具体而言，新提出的 TNT模块使用一个外 Transformer模块来对图像块之间的关系进行建模，用一个内 Transformer模块来对像素之间的关系进行建模。



结构嵌套式视觉Transformer

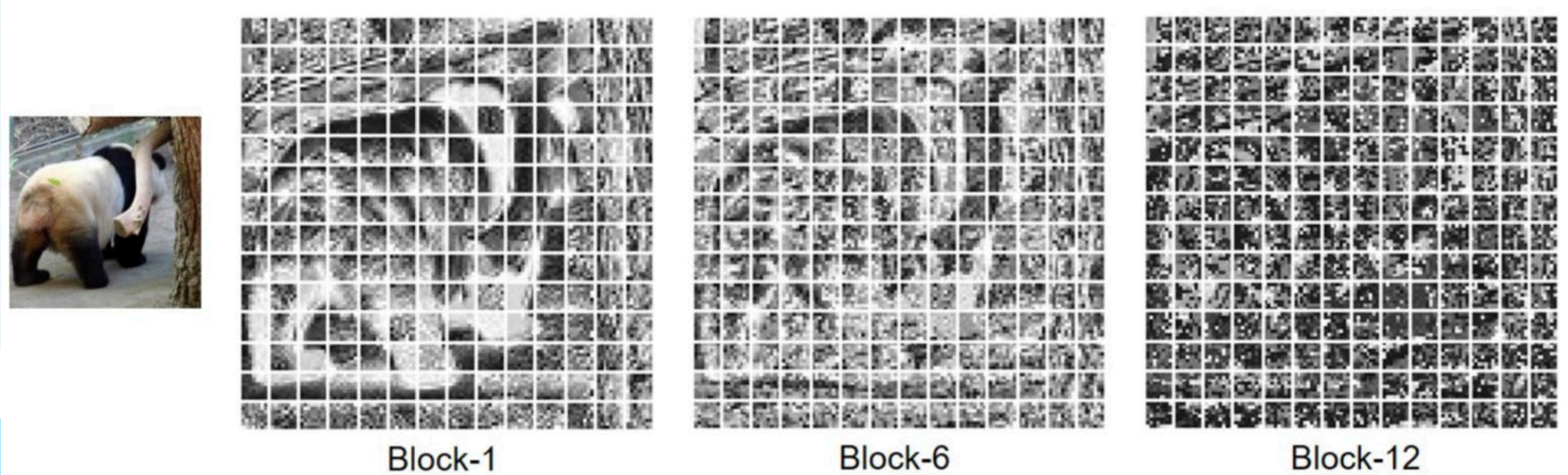
实验

ImageNet图像分类任务表现



Model	Resolution	Params (M)	FLOPs (B)	Top-1 (%)	Top-5 (%)
CNN-based					
ResNet-50 [12]	224×224	25.6	4.1	76.2	92.9
ResNet-152 [12]	224×224	60.2	11.5	78.3	94.1
RegNetY-8GF [25]	224×224	39.2	8.0	79.9	-
RegNetY-16GF [25]	224×224	83.6	15.9	80.4	-
EfficientNet-B3 [28]	300×300	12.0	1.8	81.6	94.9
EfficientNet-B4 [28]	380×380	19.0	4.2	82.9	96.4
Transformer-based					
DeiT-S [29]	224×224	22.1	4.6	79.8	-
PVT-Small [33]	224×224	24.5	3.8	79.8	-
T2T-ViT_t-14 [36]	224×224	21.5	5.2	80.7	-
TNT-S (ours)	224×224	23.8	5.2	81.5	95.7
ViT-B/16 [9]	384×384	86.4	55.5	77.9	-
DeiT-B [29]	224×224	86.4	17.6	81.8	-
TNT-B (ours)	224×224	65.6	14.1	82.8	96.3

特征图可视化



联系方式：韩凯 hankai@ios.ac.cn 13621227542