Vision GNN: An Image is Worth Graph of Nodes

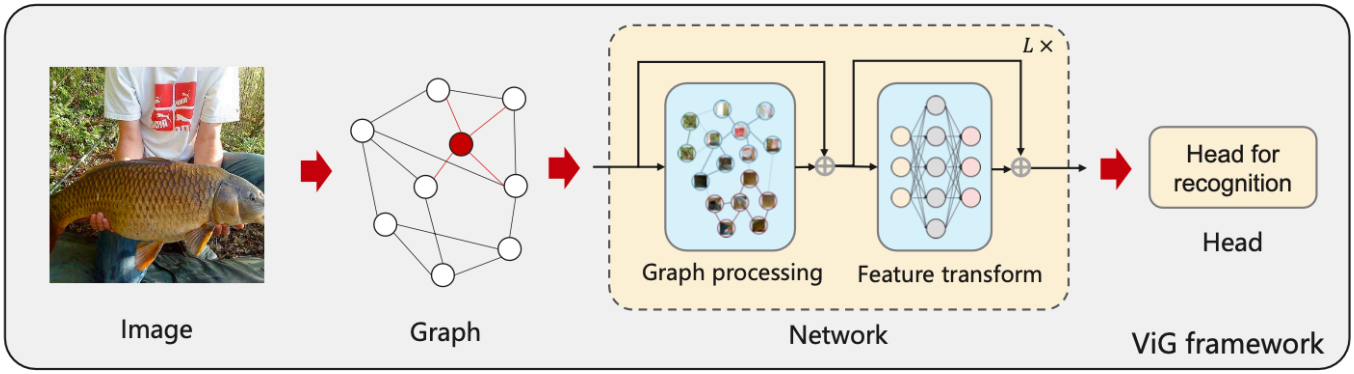
论文地址：https://arxiv.org/abs/2206.00272v2

代码地址：https://github.com/huawei-noah/CV-Backbones

论文出发点

在视觉任务中，CNNs，Transformer，MLP都取得了很好的效果。CNNs利用了平移不变性和局部性使用滑动窗口提取特征。Transformer和MLP将将图像视为一系列的patches。上述的考虑都是基于规则网格或序列表示，本文尝试一种更灵活的图结构。计算机视觉的一个基本任务是识别图像中的物体，由于对象通常不是规则的方形，而以往的网络如ResNet和ViT中常用的网格或序列结构往往会造成冗余，难以处理这些不规则对象。图是一种广义的数据结构，网格和序列可以看作是图的一种特殊情况。将图像视为图形对于视觉感知来说更加灵活和有效。本文基于图像的图表示提出了vision graph neural network，首次将图神经网络用于视觉任务，同时能取得很好的效果。

方法



1. 对图像的图表示

对于一张图片，首先将图片划分为N个patch，然后将进行特征变换得到每一个patch对应的特征。这些特性可以看作是一组无序的节点，表示为V=[v1,...,vN]。对于每一个节点vi找到穷最近的K个邻居N_vi，然后加入一条有向边eji从vj到vi。因此就得到了一个图结构G=（V，E），其中E表示所有的边集合。通过将图像视为图数据，因此可以利用GCN提取其表征。

2. 图层次的处理

图卷积层通过聚合相邻节点的特征来实现节点之间的信息交换。具体来说，聚合运算是通过聚合邻居节点的特征来计算节点的表示。更进一步，引入了多头注意力。将聚集特征分成h个头，然后分别用不同的权重更新这些头。多头更新操作使模型能够在多个表示子空间中更新信息，有利于特征的多样性。

3. ViG block

本文在图卷积前后应用线性层，将节点特征投影到同一个域，增加特征多样性。在图卷积后插入一个非线性激活函数以避免层坍塌。我们称升级后的模块为Grapher模块。为了进一步提高特征转换能力和缓解过平滑现象，在每个节点上使用前馈网络(FFN)。FFN模块是一个简单的多层感知器，有两个完全连接的层。由Grapher模块和FFN模块叠加而成的ViG块是构成网络的基本构建单元。因此构建面向视觉任务的ViG网络。与ResGCN相比，ViG随着层的深入能够保持特征多样性，学习出判别性的表征。

网络架构

1. 各向同性架构

各向同性架构意味着主体在整个网络中具有大小和形状相同的特征，如ViT和ResMLP。本文构建了三种不同模型尺寸的各向同性ViG架构，分别为ViG-ti、S和B，节点数设为N =196。为了逐渐扩大接收场，这三种模型中随着层深的增加，邻居节点数K从9线性增加到18。头的数量默认设置为h = 4。

2. 金字塔架构

金字塔架构考虑了图像的多尺度特性，即随着层越深提取空间尺寸越小的特征，如ResNet和PVT。经验证据表明，金字塔结构对视觉任务是有效的。因此，本文利用先进的设计和建立了四个版本的金字塔ViG模型。

3. 位置编码

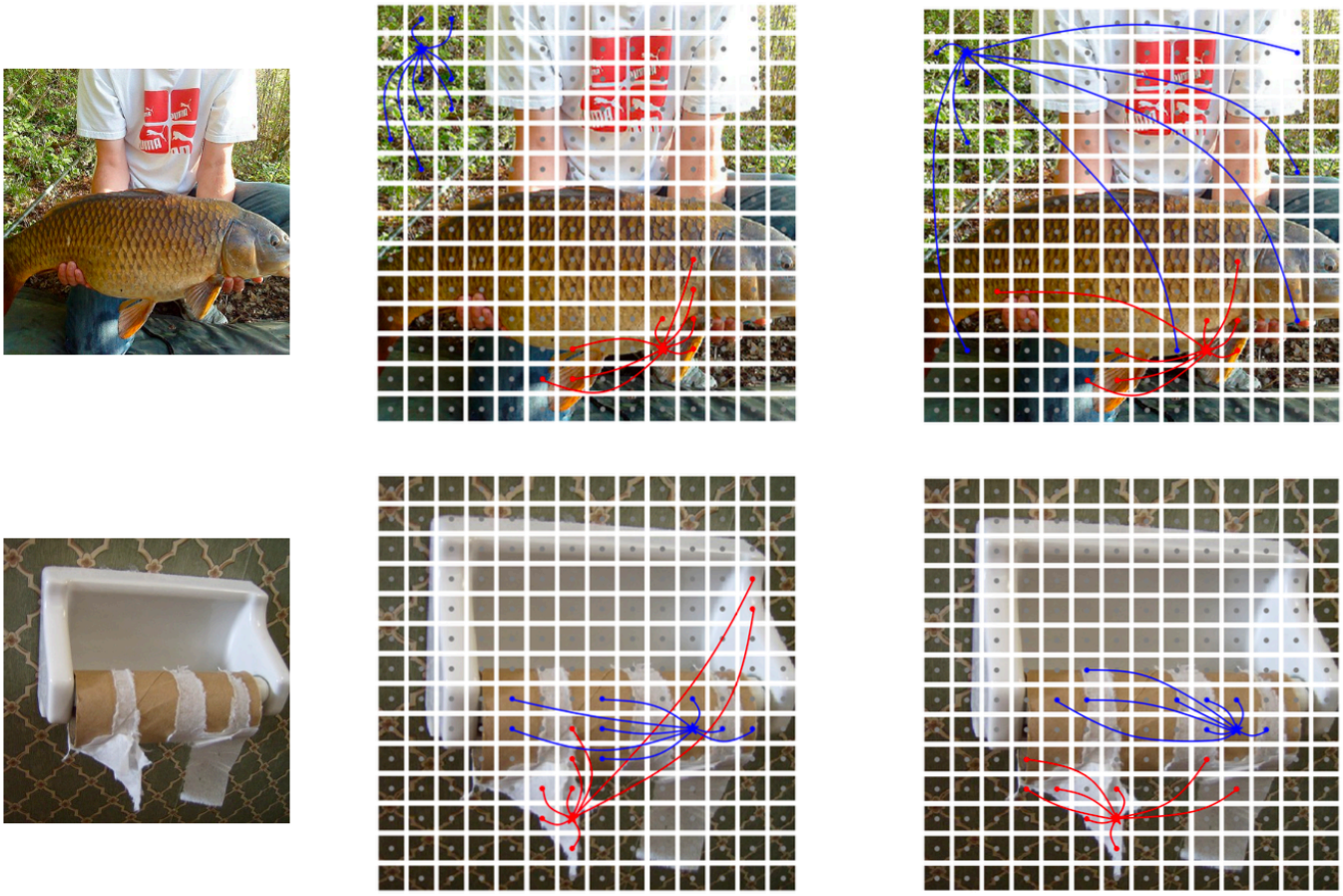
为了表示节点的位置信息，在每个节点特征中添加一个位置编码向量。对于金字塔ViG，进一步使用Swin Transformer等高级设计，例如相对位置编码。

实验结果

Model	Resolution	Params (M)	FLOPs (B)	Top-1	Top-5
♠ ResMLP-S12 conv3x3 [48]	224×224	16.7	3.2	77.0	-
♠ ConvMixer-768/32 [50]	224×224	21.1	20.9	80.2	-
♠ ConvMixer-1536/20 [50]	224×224	51.6	51.4	81.4	-
♦ ViT-B/16 [8]	384×384	86.4	55.5	77.9	-
♦ DeiT-Ti [49]	224×224	5.7	1.3	72.2	91.1
♦ DeiT-S [49]	224×224	22.1	4.6	79.8	95.0
♦ DeiT-B [49]	224×224	86.4	17.6	81.8	95.7
■ ResMLP-S24 [48]	224×224	30	6.0	79.4	94.5
■ ResMLP-B24 [48]	224×224	116	23.0	81.0	95.0
■ Mixer-B/16 [47]	224×224	59	11.7	76.4	-
★ ViG-Ti (ours)	224×224	7.1	1.3	**73.9**	**92.0**
★ ViG-S (ours)	224×224	22.7	4.5	**80.4**	**95.2**
★ ViG-B (ours)	224×224	86.8	17.7	**82.3**	**95.9**

Model	Resolution	Params (M)	FLOPs (B)	Top-1	Top-5
♠ ResNet-18 [16, 56]	224×224	12	1.8	70.6	89.7
♠ ResNet-50 [16, 56]	224×224	25.6	4.1	79.8	95.0
♠ ResNet-152 [16, 56]	224×224	60.2	11.5	81.8	95.9
♠ BoTNet-T3 [44]	224×224	33.5	7.3	81.7	-
♠ BoTNet-T3 [44]	224×224	54.7	10.9	82.8	-
♠ BoTNet-T3 [44]	256×256	75.1	19.3	83.5	-
♦ PVT-Tiny [54]	224×224	13.2	1.9	75.1	-
♦ PVT-Small [54]	224×224	24.5	3.8	79.8	-
♦ PVT-Medium [54]	224×224	44.2	6.7	81.2	-
♦ PVT-Large [54]	224×224	61.4	9.8	81.7	-
♦ CvT-13 [57]	224×224	20	4.5	81.6	-
♦ CvT-21 [57]	224×224	32	7.1	82.5	-
♦ CvT-21 [57]	384×384	32	24.9	83.3	-
♦ Swin-T [33]	224×224	29	4.5	81.3	95.5
♦ Swin-S [33]	224×224	50	8.7	83.0	96.2
♦ Swin-B [33]	224×224	88	15.4	83.5	96.5
■ CycleMLP-B2 [4]	224×224	27	3.9	81.6	-
■ CycleMLP-B3 [4]	224×224	38	6.9	82.4	-
■ CycleMLP-B4 [4]	224×224	52	10.1	83.0	-
■ Poolformer-S12 [64]	224×224	12	2.0	77.2	93.5
■ Poolformer-S36 [64]	224×224	31	5.2	81.4	95.5
■ Poolformer-M48 [64]	224×224	73	11.9	82.5	96.0
★ Pyramid ViG-Ti (ours)	224×224	10.7	1.7	**78.2**	**94.2**
★ Pyramid ViG-S (ours)	224×224	27.3	4.6	**82.1**	**96.0**
★ Pyramid ViG-M (ours)	224×224	51.7	8.9	**83.1**	**96.4**
★ Pyramid ViG-B (ours)	224×224	92.6	16.8	**83.7**	**96.5**

可视化结果



可以观察到，提出的模型可以选择与内容相关的节点作为一阶邻居。**在浅层中，倾向于根据颜色、纹理等低级和局部特征来选择邻居节点。在深层，中心节点的邻居语义性更强，属于同一类别。**因此VIG网络可以通过其内容和语义表示将节点逐渐连接起来，帮助更好地识别对象。