**当前存在问题：**

虽然transformer最近在各种视觉任务上都表现出了出色的表现。有时甚至是全局的，接受域赋予transformer模型比CNN模型具有更高的表示能力。然而，简单地扩大接受域也会引起几个问题。

一方面，使用密集的注意力，如在ViT中，会导致过多的内存和计算成本，而特征可能会受到超出感兴趣区域的无关部分的影响。

另一方面，在PVT或Swin transformer中采用的稀疏注意是数据不可知的，可能会限制建模长期关系的能力。

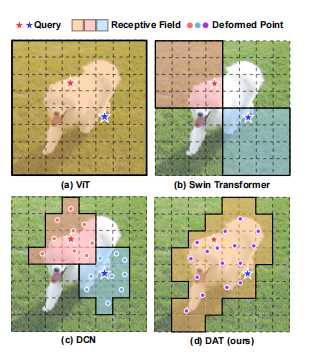
**解决问题的方式：**

提出了一种新的可变形的自注意模块DA，其中自注意中的键对和值对的位置以数据依赖的方式进行选择。这种灵活的方案使自我注意模块能够专注于相关区域，并捕获更多的信息特征.

**运用场景：**

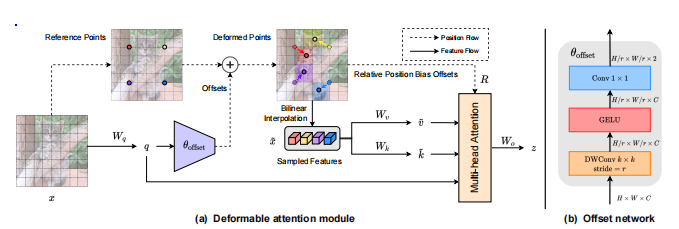
用于图像分类和各种密集预测任务。

**模型思路：**



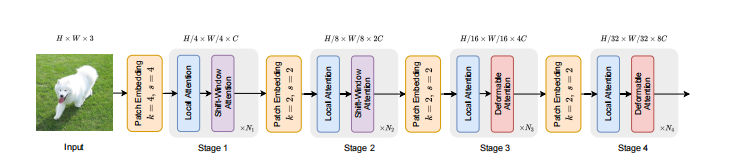
不同于DCN学习不同的抵消不同像素的整个特性地图，我们建议学习几组查询不可知偏移键和值到重要区域，基于观察全球关注通常导致几乎相同的注意模式不同的查询。这种设计既具有线性空间的复杂性，又为transformer的主干结构引入了一种可变形的注意模式。具体来说，对于每个注意模块，参考点首先作为统一的网格生成，它们在输入数据之间是相同的。然后，偏移量网络以查询特征作为输入，为所有参考点生成相应的偏移量。通过这种方式，候选键/值向重要区域转移，从而增强原始的自我注意模块具有更高的灵活性和效率，以捕获更多的信息特征。

**模块流程：**



a)表示了变形注意的信息流。在左边，一组参考点均匀地放置在特征图上，偏移量查询通过偏移网络学习其偏移量。然后根据变形点从采样特征中投影变形键和值，如右图所示。相对位置偏差也计算了变形点，增强了多头注意力，输出转换后的特征。我们只展示了4个参考点来进行清晰的表示，在实际实现中还有更多的点。(b)揭示了偏移生成网络的详细结构，用特征地图的大小进行标记

**模型流程：**



可变形注意transformer具有相似的金字塔结构，广泛适用于需要多尺度特征图的各种视觉任务。，首先将H×W×3形状的输入图像与4步进行4×4非重叠卷积，然后通过归一化层得到H4×W4×C补丁嵌入。为了建立一个层次化的特征金字塔，主干包括4个阶段，步幅逐渐增加。在两个连续的阶段之间，有一个不重叠的2×2卷积与步幅2，以向下采样特征图，使空间尺寸减半，并使特征尺寸翻倍。