Transformer 与全局感受野

Transformer 架构天生具有 **全局感受野（Global Receptive Field）**，这是它相比 CNN 最大的结构性优势之一，尤其在图像理解等计算机视觉任务中意义重大。

一、什么是全局感受野?

在视觉领域，全局感受野意味着：

> **一个位置上的特征可以直接获取整个图像（或输入）中的所有信息，进行全局建模。**

传统 CNN 要通过层层堆叠、扩大感受野来"间接"获得全局信息。而 Transformer 中，只需一个注意力层，每个 token 就能直接与 **所有其他 token 交互**。

二、Transformer 中如何实现全局感受野?

核心机制是 **自注意力（Self-Attention）**：

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$

- (Q, K, V)：分别是 query、key、value（从输入中线性映射得到）。
- 计算的是每个位置与所有其它位置之间的相关性。
- 输出即为每个位置融合全图信息后的表示。

因此，无论是图像的中心、边缘，还是小目标区域，都能在一个注意力层内获取整个图像的信息。

三、相比 CNN 的优势

特性	Transformer（自注意力）	CNN
感受野	全局（一个层即可）	局部（层层堆叠扩展）
建模长程依赖	强（直接对所有位置建模）	弱（需要很多层传播）
空间不变性	无（对 token 顺序敏感）	有（对位移鲁棒）
对非结构化输入适应性	强（可用于文本、图、序列等）	弱（主要用于栅格图像）

四、在图像中的表现：Vision Transformer (ViT)

ViT 将图像划分为固定大小的 patch,如 (16×16)，展平并嵌入为 token，然后送入标准 Transformer：

- 每个 token 能"看见"全图，学习全局上下文依赖。
- ViT 在大数据集（如 ImageNet-21k）上表现优于传统 CNN。

补充关于patch选择方面的问题

ViT 模型中的 Patch 分解

在 Vision Transformer (ViT) 模型中，图像会被分解为一系列不重叠的 **patch**，每个 patch 被当作一个 **token** 来处理。下面详细解释这一过程以及 patch 大小选择的意义。

1. 分解过程

假设输入图像的尺寸为 $(H \times W)$（例如 224×224 像素），ViT 会按照设定的 patch 大小（例如 $(P \times P)$，常见的是 16×16 或 32×32）将图像划分成若干个小块。

- 沿水平方向分为$(\frac{W}{P})$个$patch$，垂直方向分为$(\frac{H}{P})$ 个 patch。
- 总的 patch 数量为：$N = \frac{H}{P} \times \frac{W}{P}$

例如，对于 224×224 的图像和 16×16 的 patch，大约会得到 196个 patch。

2. Patch 展平与嵌入

每个$P \times P$的 patch 会被展平成一个一维向量（即将 patch 中所有像素按顺序排列），然后通过一个线性层（或全连接层）映射到一个固定维度的特征向量（也称为 patch embedding）。
这一步相当于将每个 patch 转换为 Transformer 可处理的 token。

Patch 展平后向量维度

对于常见的 RGB 图像（通道数 C=3）和 $(P \times P)$ patch 大小，patch 展平后得到的原始向量维度是：
$P^2 \times C$ 例如，ViT 中常用的 (P=16) 时，展平后就是：$16 \times 16 \times 3 = 768$维。

这个 768 维的向量再通过一个线性层（patch embedding）投影到模型的**隐藏维度**（也叫 embedding size）——通常记作 (D)。常见配置有：

- **ViT-Base/16**：(D = 768)
- **ViT-Large/16**：(D = 1024)
- **ViT-Small/16**：(D = 384)

也就是说，最终每个 patch 会被表示成一个 (D) 维的向量，(D) 的选取决定了 Transformer 模型的宽度（参数量与表达能力）。

3. Patch 大小的影响

- **较小的 patch（如 16×16 或更小）**
  - **优点**：能够捕获更多的局部细节，对细粒度任务（如精细目标检测、语义分割）更有帮助。
  - **缺点**：生成的 patch 数量较多，序列长度增加，计算量和内存占用随之上升，训练和推理时间更长。
- **较大的 patch（如 32×32 或更大）**
  - **优点**：生成的 token 数量减少，可以降低计算复杂度，加快模型训练和推理速度。
  - **缺点**：可能会丢失部分局部细节，导致对微小特征的捕捉不足，影响任务性能（尤其是在需要高分辨率细节的任务中）。

4. 平衡与设计考量

选择合适的 patch 大小需要在捕捉细节和计算成本之间做出平衡：

- 对于大规模图像分类任务，如果图像本身存在丰富的全局信息，适当增大 patch 大小可以降低计算量。
- 对于细粒度识别或者需要高分辨率信息的任务，较小的 patch 更能保留细节信息，但可能需要更多的计算资源。
- 近年来一些变体引入了局部卷积或混合结构（如 Hybrid ViT），在初期通过 CNN 提取局部特征，再结合 Transformer 的全局建模能力，以弥补纯 ViT 在局部细节上的不足。

ViT 模型将图像分解为 patch 是其将图像转化为序列数据、利用 Transformer 架构建模全局上下文信息的重要步骤，而 patch 大小的选择则直接影响模型的细粒度表达能力和计算效率。

五、全局感受野带来的挑战

1. 计算量大

- 注意力矩阵大小是 $N \times N$，其中 N 是 token 数（图像分成多少块）。
- 图像越大，计算和显存越吃紧。

解决方案：

- 降采样 token 数（如 Patch Embedding + Pooling）
- 局部注意力（如 Swin Transformer）
- 稀疏注意力（如 Performer、Linformer）

2. 缺乏局部归纳偏置

- CNN 在视觉上有天然优势：平移不变性、局部连接、权重共享。
- ViT 在小数据集上学习效率较低，收敛慢、精度低。

解决方案：

- 加入 CNN 模块（如 ConvStem）
- 提前引入局部性（如 Hybrid ViT、Conformer）
- 数据增强 + 预训练（如 MAE）

六、设计进化：平衡局部与全局

为了解决 Transformer 全局感受野虽大但计算/训练成本高的问题，很多结构引入了 **局部感受野 + 全局建模** 的折中方案：

模型	特点
Swin Transformer	局部窗口注意力 + 滑动窗口移位实现跨区域建模
PVT	使用逐步下采样控制 token 数量，减小计算
ViT-Hybrid	前几层用 CNN 提取局部特征，后接 Transformer
ConvNeXt	保留 CNN 结构，引入部分 Transformer 设计理念

七、小结

Transformer 的 **全局感受野** 赋予它以下关键优势：

- 更强的上下文理解能力
- 可建模长程依赖
- 更适合多模态（图+文）场景

但也带来：

- 高计算成本
- 对局部细节建模不足

因此，现代视觉模型的主流趋势是：**融合 Transformer 的全局建模能力 与 CNN 的局部感知能力，走向局部–全局协同建模。**