Transformer 相比 CNN 的不足

虽然 Transformer 在全局建模和长程依赖方面具有显著优势，但与传统 CNN 相比，也存在一些劣势：

1. 缺乏局部归纳偏置

- **平移不变性弱**：CNN 的卷积操作天然对平移具有不变性，而标准自注意力对位移敏感，需要额外设计（如位置编码、数据增强）来补偿。
- **局部模式捕捉能力差**：小尺度的边缘、纹理等局部特征，CNN 通过共享卷积核更高效地提取；Transformer 则需通过自注意力层反复学习。

2. 数据需求高

- **预训练依赖大规模数据**：ViT 等模型通常需要在百万级以上的图像数据（如 ImageNet-21k、JFT-300M）上预训练，否则在中小规模数据集上难以收敛或性能不佳。
- **过拟合风险**：在小数据集上直接训练，自注意力参数多、缺少强归纳偏置，容易出现过拟合。

3. 计算与内存开销大

- **二次方复杂度**：标准自注意力对 (N) 个 token 的计算和内存开销都是 (O(N^2))，当图像分成很多 patch（token）时，计算量和显存需求暴涨。
- **高分辨率图像受限**：对于高分辨率输入，token 数激增，导致 Transformer 难以直接应用于超高分辨率任务。

4. 推理速度与部署成本

- **推理延迟高**：大模型的多头注意力层和全连接层，尤其在没有硬件加速（如 TPU）时，推理速度往往慢于同等规模的 CNN。
- **硬件友好性差**：CNN 的卷积操作高度优化、易于并行；而自注意力涉及全局矩阵乘法，对 GPU/ASIC 调优要求更高。

5. 训练不稳定 & 收敛慢

- **优化难度大**：自注意力层缺乏 CNN 那样的归纳偏置，需要更精细的学习率调度、正则化策略和长时间预热。
- **梯度分布不均**：在深层 Transformer 中，梯度可能集中在某些头或层，导致部分注意力头“死亡”或冗余。

6. 对细粒度局部信息不够敏感

- **Patch 粒度限制**：将图像切成 (P\times P) 的 patch 后，patch 内的微小细节在嵌入阶段就被“压平”处理，不如 CNN 的多层小卷积逐层提炼细节。
- **额外混合结构需求**：很多工作不得不引入 Hybrid ViT、局部卷积模块或可变形卷积来补强细节建模。

7. 模型体积大 & 部署复杂

- **参数量多**：同等性能下，Transformer 模型通常参数更多，难以在边缘设备或移动端部署。
- **裁剪与量化难度**：注意力机制带来的全局依赖，使得剪枝、蒸馏和量化更加复杂。