问题一

以CNN为基础的视觉表征模型有哪些特点，这些特点是Attention机制具备的吗?

CNN的特点

1. 高效的信息提取能力:
   局部连接和权值共享大幅减少参数量，降低计算复杂度。
   卷积操作高度并行化，便于利用GPU进行加速
2. 局部性:
   局部感受野
3. 平移等变性:
   无需显式学习同一物体在不同位置的特征表示。

这些基于尝试的先验知识使其特别适合图像任务。

Attention机制

1. 缺乏图像特有的结构性先验知识:
   需要从数据中学习对模型图像感知特征的能力，计算复杂度高
2. 全局感受野:
   能够捕捉到图像很远的信息，解决了CNN中信息需要经过多层传播的问题

问题二

使用Transformer取代CNN将面临哪些挑战?

1. 复杂度过高:
   调整超参优化难度更大
2. 归纳偏置缺乏:
   需要从数据中重头学习新的空间关系
3. 位置信息依赖:
   必须显式引入位置编码
4. 依赖大规模数据:
   模型在大数据集上才能发挥优势且需要进行预训练

问题三

ViT 模型取得巨大成功的关键点在哪里?

1. 大规模预训练:
   NLP+微调从Bert模型迁移到CV领域
   可应用于CV领域常见任务：如图像分类、检测与切割
2. 高效的序列建模:
   全局上下文理解
3. 可扩展性:
   随着数据规模的增大，性能有良好表现。
   实现多模态融合，统一NLP和CV