# 花分类识别

# ViT模型

Vision Transformer (ViT) 是一种基于纯Transformer架构的图像分类模型，它将输入图像分割为固定大小的图像块并线性嵌入为序列，通过添加位置编码保留空间信息后送入多层Transformer编码器进行特征提取，最终利用额外的可学习分类令牌实现图像分类，完全摒弃了传统卷积操作，在大规模数据集上表现出色但在小规模数据上可能需要更强的正则化或预训练。ViT模型如图 1所示。

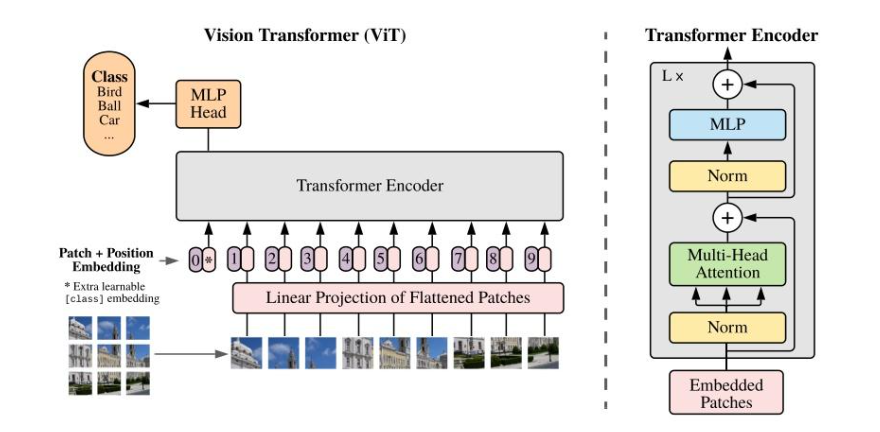


图 1 ViT模型结构示意图

其中，Transformer块的核心由多头自注意力机制和前馈神经网络组成，先通过层归一化处理输入数据，然后利用多头自注意力计算全局依赖关系并保留残差连接，接着再次层归一化后送入前馈网络进行非线性变换并叠加残差连接，最终输出经过双重特征交互和归一化的增强特征表示。注意力计算如下：

其中，、、是将输入通过一个线性投影映射得到，表示注意力计算的查询、键、值。表示激活函数，表示和的维度。

# 数据集与评价指标

花分类数据集（Flower Photos）是经典的细粒度图像分类基准，包含不同品种花卉的标注图像，典型规模为数千至数万张（如102类×每类40-250张），涵盖菊花、玫瑰、郁金香等常见品种，图像具有姿态多变、背景复杂、类间差异细微等特点，常用于测试模型在细粒度视觉识别中的特征提取和判别能力，是计算机视觉领域验证图像分类、迁移学习等任务的常用数据集之一。

评价指标为Top-1，Top-1准确率是分类任务中最常用的评估指标，表示模型预测概率最高的类别（即Top-1预测结果）与实际标签一致的样本比例，直接反映模型的单标签分类能力，例如在ImageNet分类任务中若模型对某张图片的最高预测概率对应"狗"且真实标签也是"狗"，则计入正确分类计数，最终Top-1准确率等于正确预测数除以总样本数，该指标计算简单但要求严格，必须最高置信度预测完全匹配真实值才视为正确。

# 计算量（FLOPs）与参数量（Params）

总计算量与参数量如下：

flops: 16.86 G, params: 86.24 M

每一层计算量与参数量如下：

{'patch\_embed': (115605504.0, 590592.0, {'proj': (115605504.0, 590592.0, {}),

'norm': (0.0, 0, {})}),

'pos\_drop': (0.0, 0.0, {}),

'blocks': (16746651648.0, 85054464.0, {

'0': (1395554304.0, 7087872.0, {'norm1': (605184.0, 1536.0, {}), 'attn': (464781312.0, 2362368.0, {'qkv': (348585984.0, 1771776.0, {}), 'attn\_drop': (0.0, 0.0, {}), 'proj': (116195328.0, 590592.0, {}), 'proj\_drop': (0.0, 0.0, {})}), 'drop\_path': (0.0, 0, {}), 'norm2': (605184.0, 1536.0, {}), 'mlp': (929562624.0, 4722432.0, {'fc1': (464781312.0, 2362368.0, {}), 'act': (0.0, 0, {}), 'fc2': (464781312.0, 2360064.0, {}), 'drop': (0.0, 0.0, {})})}),

'1': (1395554304.0, 7087872.0, {'norm1': (605184.0, 1536.0, {}), 'attn': (464781312.0, 2362368.0, {'qkv': (348585984.0, 1771776.0, {}), 'attn\_drop': (0.0, 0.0, {}), 'proj': (116195328.0, 590592.0, {}), 'proj\_drop': (0.0, 0.0, {})}), 'drop\_path': (0.0, 0, {}), 'norm2': (605184.0, 1536.0, {}), 'mlp': (929562624.0, 4722432.0, {'fc1': (464781312.0, 2362368.0, {}), 'act': (0.0, 0, {}), 'fc2': (464781312.0, 2360064.0, {}), 'drop': (0.0, 0.0, {})})}),

'2': (1395554304.0, 7087872.0, {'norm1': (605184.0, 1536.0, {}), 'attn': (464781312.0, 2362368.0, {'qkv': (348585984.0, 1771776.0, {}), 'attn\_drop': (0.0, 0.0, {}), 'proj': (116195328.0, 590592.0, {}), 'proj\_drop': (0.0, 0.0, {})}), 'drop\_path': (0.0, 0, {}), 'norm2': (605184.0, 1536.0, {}), 'mlp': (929562624.0, 4722432.0, {'fc1': (464781312.0, 2362368.0, {}), 'act': (0.0, 0, {}), 'fc2': (464781312.0, 2360064.0, {}), 'drop': (0.0, 0.0, {})})}),

'3': (1395554304.0, 7087872.0, {'norm1': (605184.0, 1536.0, {}), 'attn': (464781312.0, 2362368.0, {'qkv': (348585984.0, 1771776.0, {}), 'attn\_drop': (0.0, 0.0, {}), 'proj': (116195328.0, 590592.0, {}), 'proj\_drop': (0.0, 0.0, {})}), 'drop\_path': (0.0, 0, {}), 'norm2': (605184.0, 1536.0, {}), 'mlp': (929562624.0, 4722432.0, {'fc1': (464781312.0, 2362368.0, {}), 'act': (0.0, 0, {}), 'fc2': (464781312.0, 2360064.0, {}), 'drop': (0.0, 0.0, {})})}),

'4': (1395554304.0, 7087872.0, {'norm1': (605184.0, 1536.0, {}), 'attn': (464781312.0, 2362368.0, {'qkv': (348585984.0, 1771776.0, {}), 'attn\_drop': (0.0, 0.0, {}), 'proj': (116195328.0, 590592.0, {}), 'proj\_drop': (0.0, 0.0, {})}), 'drop\_path': (0.0, 0, {}), 'norm2': (605184.0, 1536.0, {}), 'mlp': (929562624.0, 4722432.0, {'fc1': (464781312.0, 2362368.0, {}), 'act': (0.0, 0, {}), 'fc2': (464781312.0, 2360064.0, {}), 'drop': (0.0, 0.0, {})})}),

'5': (1395554304.0, 7087872.0, {'norm1': (605184.0, 1536.0, {}), 'attn': (464781312.0, 2362368.0, {'qkv': (348585984.0, 1771776.0, {}), 'attn\_drop': (0.0, 0.0, {}), 'proj': (116195328.0, 590592.0, {}), 'proj\_drop': (0.0, 0.0, {})}), 'drop\_path': (0.0, 0, {}), 'norm2': (605184.0, 1536.0, {}), 'mlp': (929562624.0, 4722432.0, {'fc1': (464781312.0, 2362368.0, {}), 'act': (0.0, 0, {}), 'fc2': (464781312.0, 2360064.0, {}), 'drop': (0.0, 0.0, {})})}),

'6': (1395554304.0, 7087872.0, {'norm1': (605184.0, 1536.0, {}), 'attn': (464781312.0, 2362368.0, {'qkv': (348585984.0, 1771776.0, {}), 'attn\_drop': (0.0, 0.0, {}), 'proj': (116195328.0, 590592.0, {}), 'proj\_drop': (0.0, 0.0, {})}), 'drop\_path': (0.0, 0, {}), 'norm2': (605184.0, 1536.0, {}), 'mlp': (929562624.0, 4722432.0, {'fc1': (464781312.0, 2362368.0, {}), 'act': (0.0, 0, {}), 'fc2': (464781312.0, 2360064.0, {}), 'drop': (0.0, 0.0, {})})}),

'7': (1395554304.0, 7087872.0, {'norm1': (605184.0, 1536.0, {}), 'attn': (464781312.0, 2362368.0, {'qkv': (348585984.0, 1771776.0, {}), 'attn\_drop': (0.0, 0.0, {}), 'proj': (116195328.0, 590592.0, {}), 'proj\_drop': (0.0, 0.0, {})}), 'drop\_path': (0.0, 0, {}), 'norm2': (605184.0, 1536.0, {}), 'mlp': (929562624.0, 4722432.0, {'fc1': (464781312.0, 2362368.0, {}), 'act': (0.0, 0, {}), 'fc2': (464781312.0, 2360064.0, {}), 'drop': (0.0, 0.0, {})})}),

'8': (1395554304.0, 7087872.0, {'norm1': (605184.0, 1536.0, {}), 'attn': (464781312.0, 2362368.0, {'qkv': (348585984.0, 1771776.0, {}), 'attn\_drop': (0.0, 0.0, {}), 'proj': (116195328.0, 590592.0, {}), 'proj\_drop': (0.0, 0.0, {})}), 'drop\_path': (0.0, 0, {}), 'norm2': (605184.0, 1536.0, {}), 'mlp': (929562624.0, 4722432.0, {'fc1': (464781312.0, 2362368.0, {}), 'act': (0.0, 0, {}), 'fc2': (464781312.0, 2360064.0, {}), 'drop': (0.0, 0.0, {})})}),

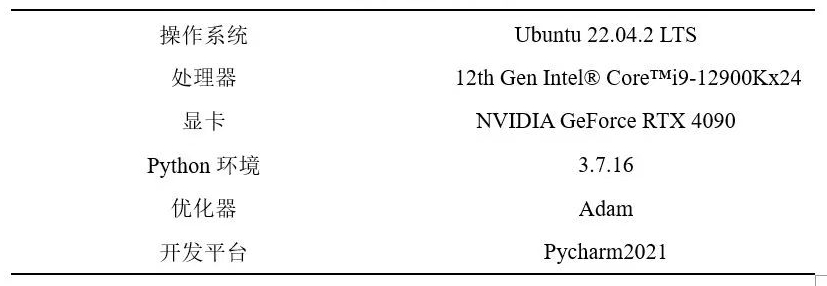
'9': (1395554304.0, 7087872.0, {'norm1': (605184.0, 1536.0, {}), 'attn': (464781312.0, 2362368.0, {'qkv': (348585984.0, 1771776.0, {}), 'attn\_drop': (0.0, 0.0, {}), 'proj': (116195328.0, 590592.0, {}), 'proj\_drop': (0.0, 0.0, {})}), 'drop\_path': (0.0, 0, {}), 'norm2': (605184.0, 1536.0, {}), 'mlp': (929562624.0, 4722432.0, {'fc1': (464781312.0, 2362368.0, {}), 'act': (0.0, 0, {}), 'fc2': (464781312.0, 2360064.0, {}), 'drop': (0.0, 0.0, {})})}),

'10': (1395554304.0, 7087872.0, {'norm1': (605184.0, 1536.0, {}), 'attn': (464781312.0, 2362368.0, {'qkv': (348585984.0, 1771776.0, {}), 'attn\_drop': (0.0, 0.0, {}), 'proj': (116195328.0, 590592.0, {}), 'proj\_drop': (0.0, 0.0, {})}), 'drop\_path': (0.0, 0, {}), 'norm2': (605184.0, 1536.0, {}), 'mlp': (929562624.0, 4722432.0, {'fc1': (464781312.0, 2362368.0, {}), 'act': (0.0, 0, {}), 'fc2': (464781312.0, 2360064.0, {}), 'drop': (0.0, 0.0, {})})}),

'11': (1395554304.0, 7087872.0, {'norm1': (605184.0, 1536.0, {}), 'attn': (464781312.0, 2362368.0, {'qkv': (348585984.0, 1771776.0, {}), 'attn\_drop': (0.0, 0.0, {}), 'proj': (116195328.0, 590592.0, {}), 'proj\_drop': (0.0, 0.0, {})}), 'drop\_path': (0.0, 0, {}), 'norm2': (605184.0, 1536.0, {}), 'mlp': (929562624.0, 4722432.0, {'fc1': (464781312.0, 2362368.0, {}), 'act': (0.0, 0, {}), 'fc2': (464781312.0, 2360064.0, {}), 'drop': (0.0, 0.0, {})})})}),

'norm': (605184.0, 1536.0, {}), 'pre\_logits': (589824.0, 590592.0, {'fc': (589824.0, 590592.0, {}), 'act': (0.0, 0, {})}), 'head': (3840.0, 3845.0, {})}

# 实验环境：



# 参数设置：

使用ViT\_L

参数设置如下：

# vit-L  
model = VisionTransformer(img\_size=224,  
 in\_c=in\_c,  
 patch\_size=16,  
 embed\_dim=1024,  
 depth=24,  
 num\_heads=16,  
 representation\_size=None,  
 num\_classes=num\_classes).to(device)

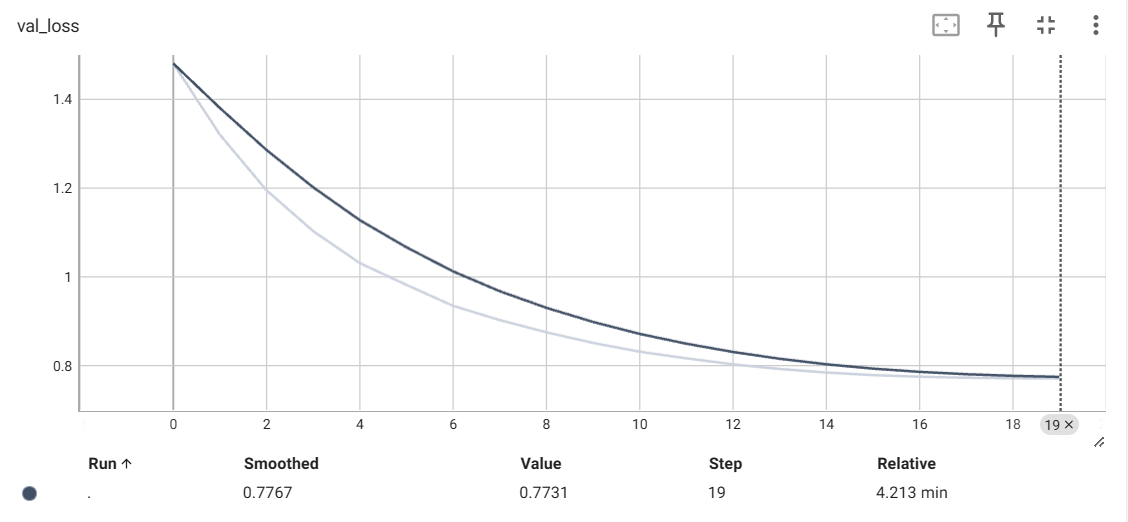
# 核心代码如下：

def forward\_features(self, x):  
 # [B, C, H, W] -> [B, num\_patches, embed\_dim]  
 x = self.patch\_embed(x) # [B, 196, 768]  
 # [1, 1, 768] -> [B, 1, 768]  
 cls\_token = self.cls\_token.expand(x.shape[0], -1, -1)  
 if self.dist\_token is None:  
 x = torch.cat((cls\_token, x), dim=1) # [B, 197, 768]  
 else:  
 x = torch.cat((cls\_token, self.dist\_token.expand(x.shape[0], -1, -1), x), dim=1)  
  
 x = self.pos\_drop(x + self.pos\_embed)  
 x = self.blocks(x)  
 x = self.norm(x)   
  
def forward(self, x):  
 x = self.forward\_features(x)  
 if self.head\_dist is not None:  
 x, x\_dist = self.head(x[0]), self.head\_dist(x[1])  
 if self.training and not torch.jit.is\_scripting():  
 # during inference, return the average of both classifier predictions  
 return x, x\_dist  
 else:  
 return (x + x\_dist) / 2  
 else:  
 x = self.head(x)  
 return x

# 实验结果

以下实验结果使用tensorboard绘制。

Loss曲线：



Top-1 Acc曲线：

