实验二：基于 ViT 的 CIFAR-10 图像分类实验报告

概述

任务目标

- 学习如何使用深度学习框架来实现和训练一个 ViT 模型，以及 ViT 中的Attention 机制
- 进一步掌握使用深度学习框架完成任务的具体流程：如读取数据、构造网络、训练模型和测试模型等
- 实现 CIFAR-10 数据集的 10 类别图像分类任务，验证 Vision Transformer（ViT）模型在小尺度图像上的有效性.

数据集

- **CIFAR-10 数据集**：
  - 60,000 张 32×32 彩色图像（50k训练集 + 10k测试集）
  - 均匀分布在 10 个类别：飞机、汽车、鸟、猫、鹿、狗、青蛙、马、船、卡车
  - 数据增强策略：

```
transforms.RandomCrop(32, padding=4)
transforms.RandomHorizontalFlip()
```

解决方案

采用 Vision Transformer 架构，核心创新点：

- 将 32×32 图像分割为 4×4 的小块（共 64 patches）
- 通过线性映射获取 patch embeddings
- 引入可学习的 [CLS] token 用于分类
- 6 层 Transformer 编码器堆叠

网络结构设计

ViT 架构图

```
ViT(
  (to_patch_embedding): Sequential(
    Rearrange('b c (h p1) (w p2) -> b (h w) (p1 p2 c)', p1=4, p2=4),
    LayerNorm(48),
    Linear(48, 512),
    LayerNorm(512)
  )
  (transformer): Encoder(
    (layers): ModuleList(
      [ModuleList(Attention + FeedForward) ×6]
    )
```

```
  )
  (mlp_head): Linear(512, 10)
)
```

核心组件

1. **Patch Embedding**：

```
Rearrange('b c (h p1) (w p2) -> b (h w) (p1 p2 c)', p1=4, p2=4)
Linear(48, 512)  # 4×4×3=48 → 512
```

2. **位置编码**：

```
self.pos_embedding = nn.Parameter(torch.randn(1, 65, 512))  # 64+1
```

3. **Transformer Encoder**：

```
Encoder(
  dim=512, depth=6, heads=8,
  mlp_dim=512, dim_head=64
)
```

4. **分类头**：

```
self.mlp_head = nn.Linear(512, 10)
```

损失函数与优化器

损失函数

```
nn.CrossEntropyLoss()  # 交叉熵损失
```

优化策略

```
optim.Adam(lr=0.0003)  # 初始学习率 3e-4
scheduler = ReduceLROnPlateau(
  optimizer,
  mode='min',
  patience=3
)  # 动态学习率调整
```

创新点

1. **小尺度图像适配**：
   - 采用 4×4 的 patch 划分策略（原论文为 16×16）

- 调整位置编码维度适配 32×32 输入

2. **轻量化设计**：
   - 仅使用 6 层 Transformer（Base 版本为 12 层）
   - 隐藏维度 512（Base 版本为 768）
3. **训练优化**：
   - 引入随机裁剪+水平翻转增强
   - 使用动态学习率调整策略

实验分析

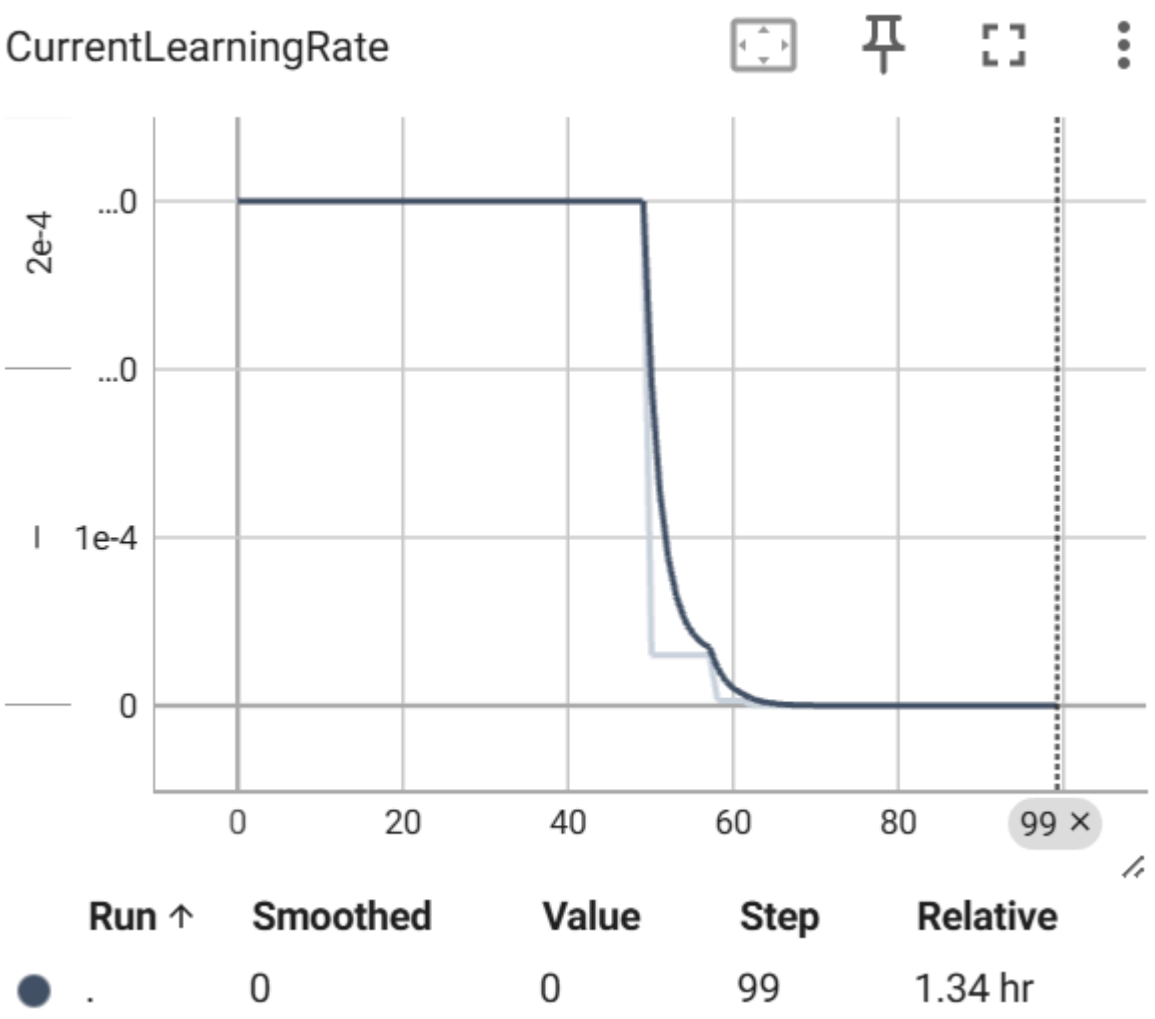
训练配置

参数	值
Batch Size	256
Epochs	100
初始 LR	3e-4
优化器	Adam
设备	GPU（CUDA）

实验结果

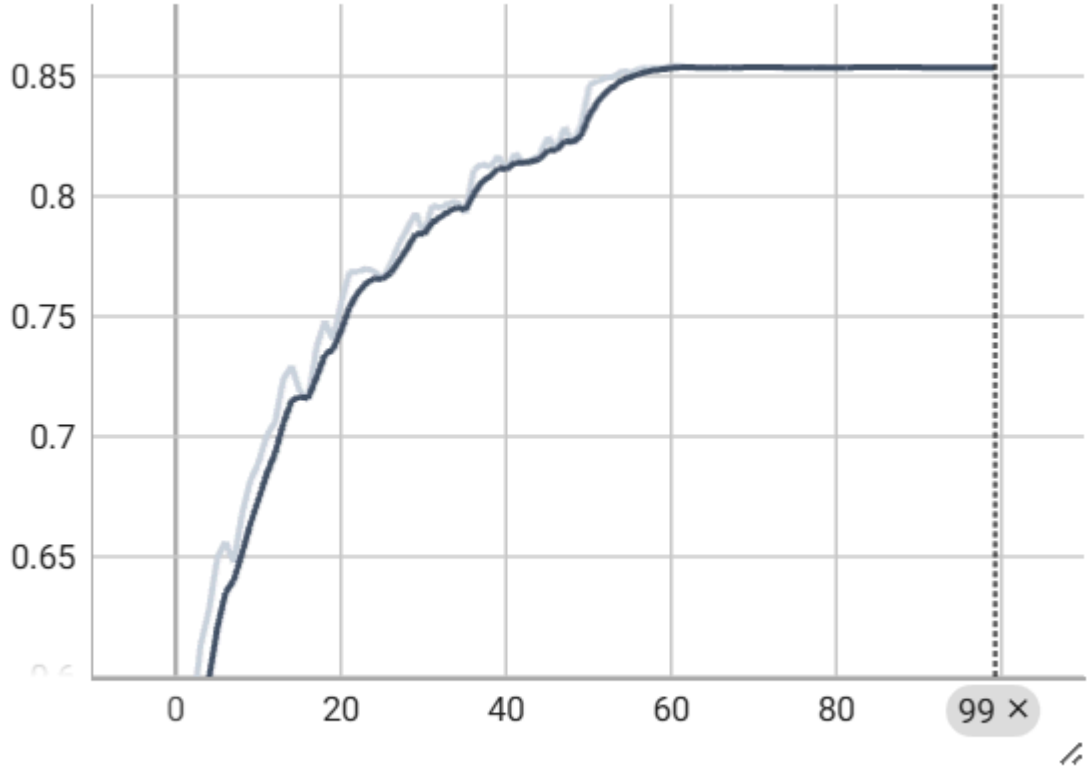
指标	训练集	测试集
最佳准确率	91.07%	85.36%
最终损失值	0.2491	0.4691

学习曲线：



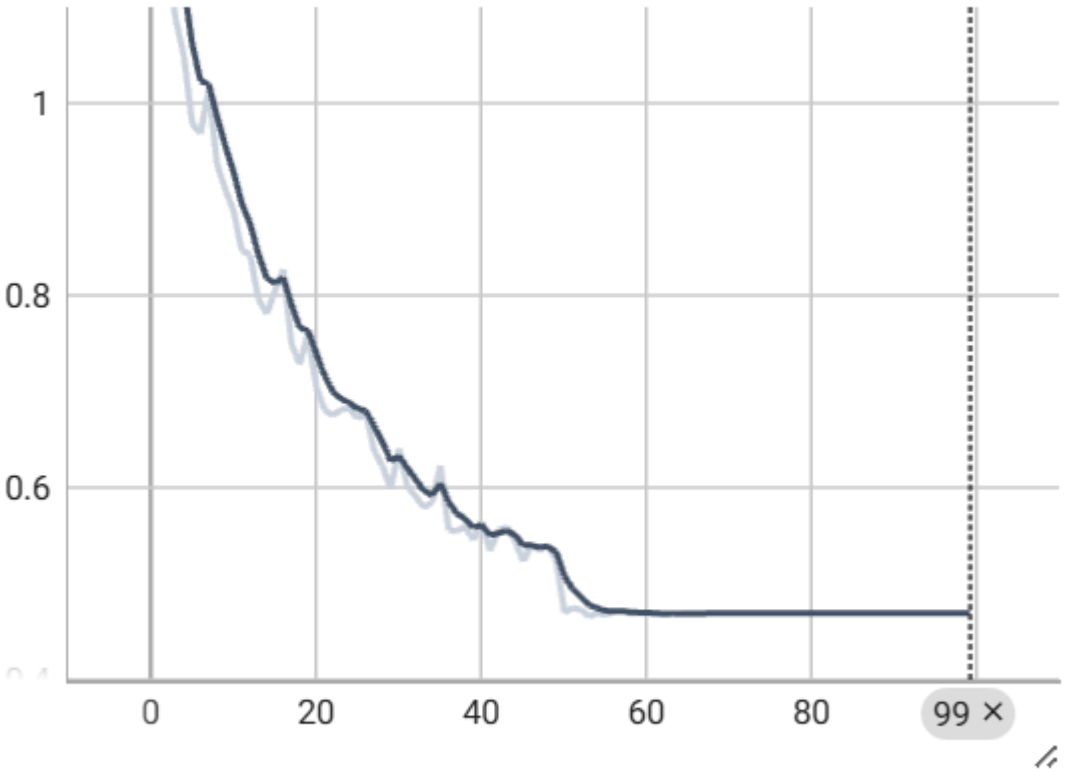
Run ↑	Smoothed	Value	Step	Relative
.	0	0	99	1.34 hr

TestAccuracy



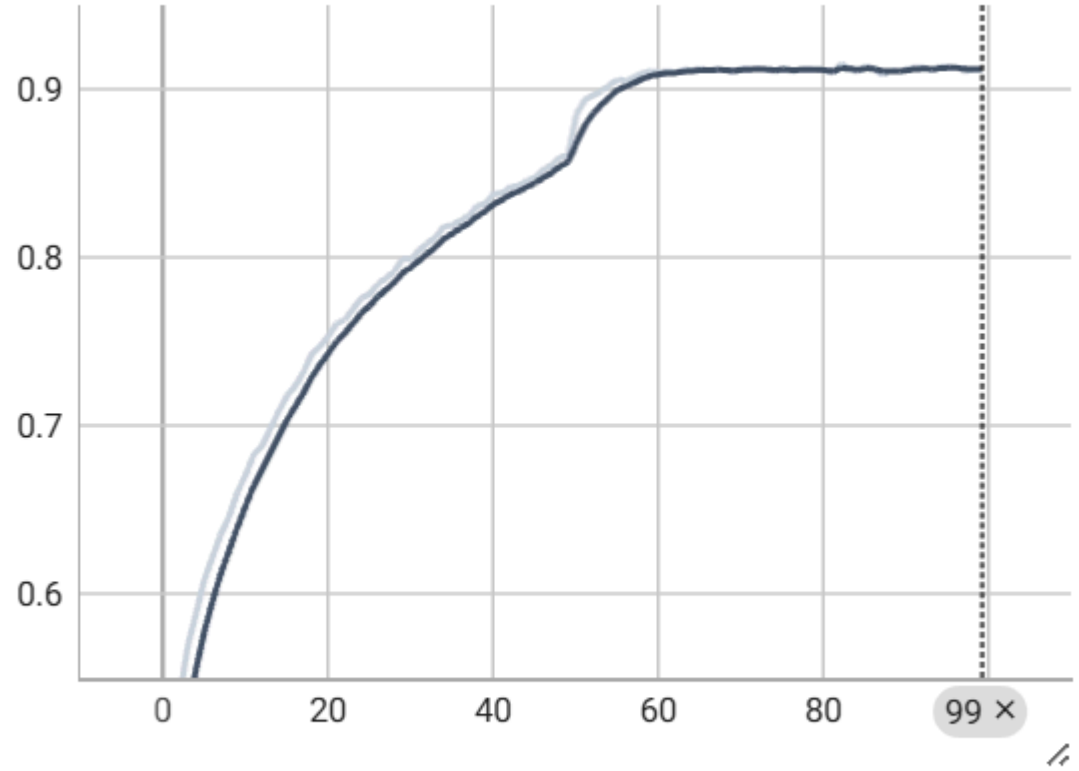
Run ↑	Smoothed	Value	Step	Relative
.	0.8536	0.8536	99	1.34 hr

TestLoss

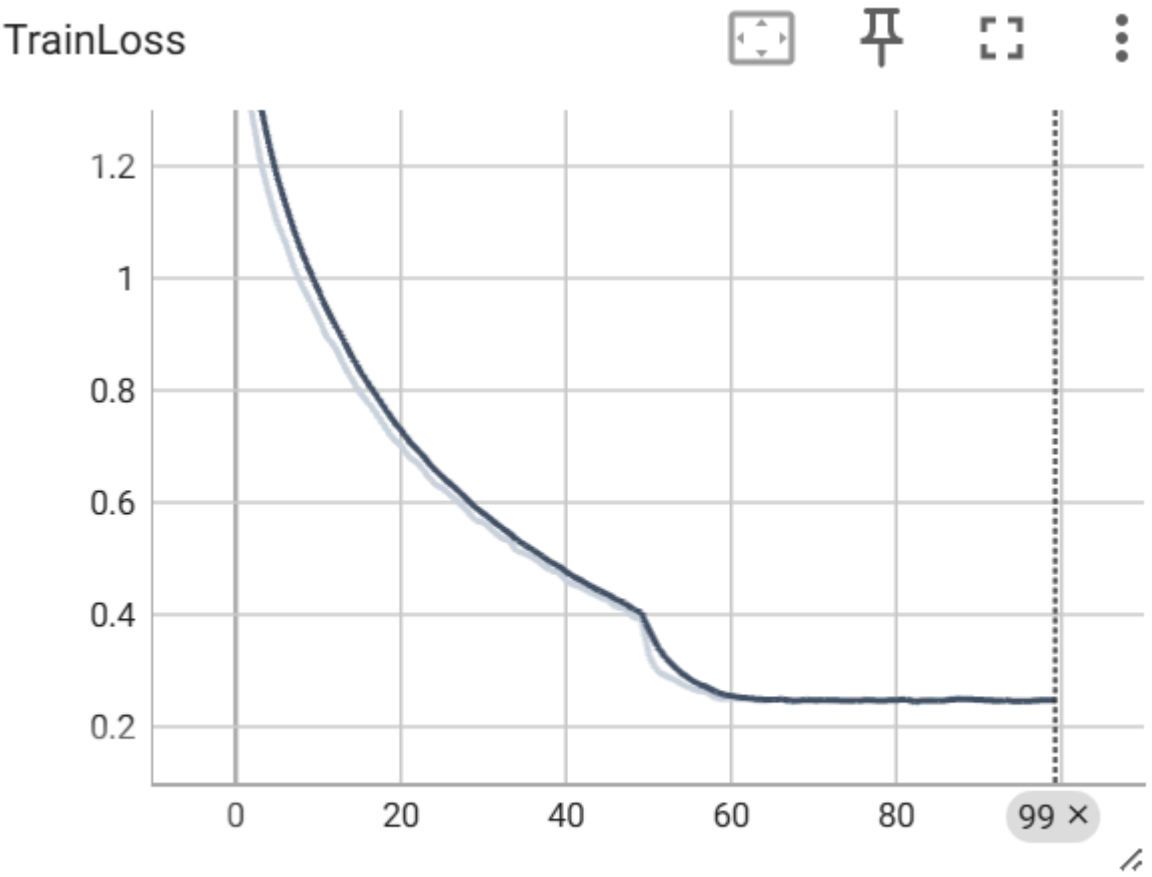


Run ↑	Smoothed	Value	Step	Relative
.	0.469	0.469	99	1.34 hr

TrainAccuracy



Run ↑	Smoothed	Value	Step	Relative
.	0.9118	0.9117	99	1.34 hr



Run ↑	Smoothed	Value	Step	Relative
.	0.2474	0.2481	99	1.34 hr

结果分析

1. **收敛性**：模型在 60 个 epoch 后进入稳定收敛状态
2. **泛化能力**：训练集与测试集存在约 6% 的准确率差距，表明可能存在一定的过拟合
3. **优化效果**：动态学习率调整使测试损失下降显著

总结

主要成果

- 成功将 ViT 应用于小尺度图像分类任务
- 在 CIFAR-10 上实现 85.36% 的测试准确率
- 验证了 Transformer 架构在 CV 任务中的有效性

改进方向

1. 引入更强的正则化策略（DropPath, Stochastic Depth）
2. 尝试混合架构（CNN + Transformer）
3. 使用更大的预训练模型进行迁移学习

附录：

- 完整代码见 [**CICV_zyh.py**]
- 权重文件见[**model_weights_pth**]
- 模型可视化结果通过 Netron 生成，部分结构图见[**1.png, 2.png, 3.png**]