

期末作业：自监督训练、Transformer 与三维重建

何益涵 20307110032

吕文韬 23210180109

摘要： 本项目的仓库地址为[该超链接](https://github.com/HeDesertFox/Neural-Networks-and-Deep-Learning-Homework-Group-Tasks.git)¹. 详见其中的文件夹 `final`. 其中包含自监督框架任务 `task1_self_supervised_learning`、Transformer 任务 `task2_Transformer_vs_CNN` 和三维重建任务 `task3_object_reconstruction_view_synthesis`.

本项目提供了训练好的模型，详见百度网盘²，若链接失效，也可以通过 Github Issue 联系作者.

¹<https://github.com/HeDesertFox/Neural-Networks-and-Deep-Learning-Homework-Group-Tasks.git>

²提取码

第一章 对比监督学习和自监督学习在图像分类任务上的性能表现

第 1 节 任务描述

1. 实现任一自监督学习算法并使用该算法在自选的数据集上训练 ResNet-18，随后在 CIFAR-100 数据集中使用 Linear Classification Protocol 对其性能进行评测；
2. 将上述结果与在 ImageNet 数据集上采用监督学习训练得到的表征在相同的协议下进行对比，并比较二者相对于在 CIFAR-100 数据集上从零开始以监督学习方式训练所带来的提升；
3. 尝试不同的超参数组合，探索自监督预训练数据集规模对性能的影响。

第 2 节 项目架构

第二章 在 CIFAR-100 数据集上比较基于 Transformer 和 CNN 的图像分类模型

第 1 节 任务描述

1. 分别基于 CNN 和 Transformer 架构实现具有相近参数量的图像分类网络;
2. 在 CIFAR-100 数据集上采用相同的训练策略对二者进行训练, 其中数据增强策略中应包含 CutMix;
3. 尝试不同的超参数组合, 尽可能提升各架构在 CIFAR-100 上的性能以进行合理的比较。

第 2 节 项目架构

此任务的所有文件在 `task2_Transformer_vs_CNN` 文件夹中, 其中包含以下文件:

1. `data_preparation.py` 文件: 包含数据下载函数与预处理函数, 其中包含 Cutmix 的实现.
2. `model.py` 文件: 包含 Resnet 和 ViT 的模型构造函数.
3. `training.py` 文件: 包含训练和超参数调优函数.
4. `test_notebook.ipynb` 文件: 包含 Transformer 架构和 CNN 架构的对比任务的全部流程, 如数据加载、调参和训练可视化.

第 3 节 实验设置

3.1. 数据增广

此项目的数据增广在常规图像变化的基础上增加了 Cutmix 增强方法。

CutMix 是一种数据增强技术, 旨在提升网络的泛化性能。CutMix 的主要思想是在训练图像之间剪切并粘贴图像块, 同时按比例混合相应的标签。此方法有助于模型从更多样化的训练样本中学习, 从而提高其鲁棒性和性能。

CutMix 的原理包括两个主要步骤：从一张图像中剪切出一个块并将其粘贴到另一张图像上，然后按比例调整两张图像的标签。具体来说，对于两张图像 x_A 和 x_B 及其对应的标签 y_A 和 y_B ，新的图像 \tilde{x} 和标签 \tilde{y} 的生成方式如下：

$$\tilde{x} = M \odot x_A + (1 - M) \odot x_B \quad (2.1)$$

$$\tilde{y} = \lambda y_A + (1 - \lambda) y_B \quad (2.2)$$

其中， M 是一个二进制掩码，表示图像块的应用位置， \odot 表示元素级乘法， $\lambda \in [0, 1]$ 是表示原始图像保留比例的随机变量，通常从 Beta 分布中采样。

CutMix 已被证明能通过以下几方面增强深度学习模型的性能：

- 提高泛化性能：通过使模型暴露于更广泛的图像变体，CutMix 有助于防止过拟合，从而提高在未见数据上的泛化性能。
- 正则化：该技术类似于 dropout，通过引入随机性防止模型过度依赖特定特征，起到正则化的作用。
- 标签平滑：标签混合有助于模型学习更柔和的标签，这对于分类任务中类间界限不明确的情况尤为有益。

实验证明，与传统的数据增强技术相比，使用 CutMix 训练的模型在准确率和对抗攻击的鲁棒性方面表现更佳。

CutMix 的实现包括以下步骤：

1. 随机选择两张图像：从训练集中选择两张图像 x_A 和 x_B 及其对应的标签 y_A 和 y_B 。
2. 生成二进制掩码：通过在第一张图像中选择一个随机矩形区域，并将该区域对应的第二张图像的部分设为零，创建一个二进制掩码 M 。
3. 应用 CutMix：使用二进制掩码组合两张图像，创建新的训练样本。根据掩码区域的面积按比例调整标签。
4. 更新训练数据：使用新的图像和标签作为训练过程的一部分。

在本项目的实现中，函数 `rand_bbox` 用于生成随机边界框，而 `cutmix` 函数则将 CutMix 增强应用于一批图像。在训练函数每次载入小批量数据时，我们调用 `cutmix` 函数。

3.2. 模型选择

本项目采用的 CNN 架构是 resnet 架构，具体采用了 resnet18 架构，也可以改为更大的 resnet 架构。

本项目采用的 Transformer 架构是经典的 ViT(Vision Transformer) 架构：

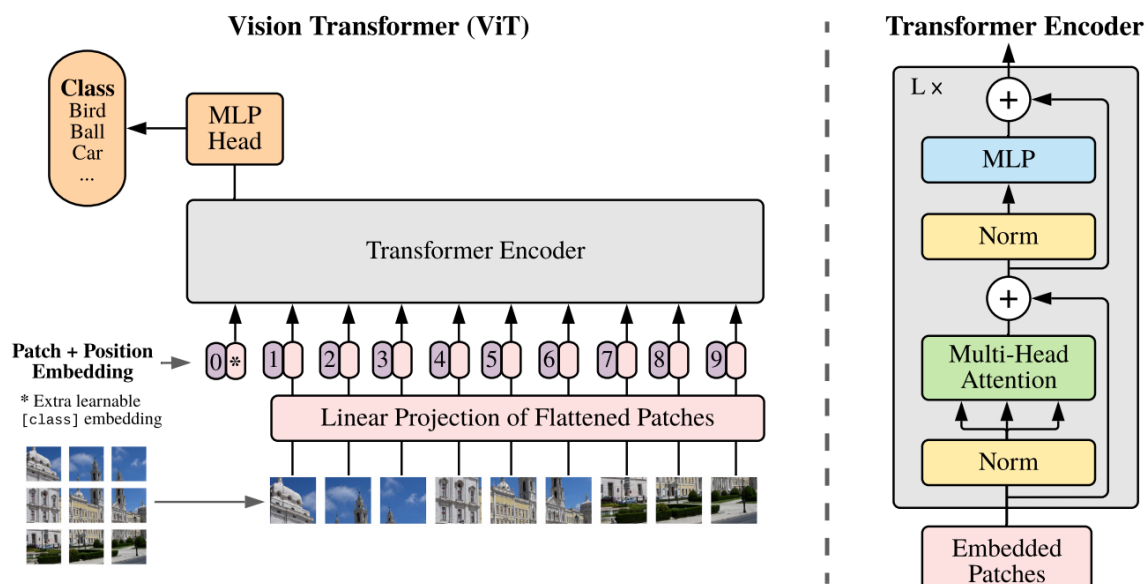


图 2.1: ViT 架构

ViT 是一种将 Transformer 应用于图像分类任务的新型架构。Transformer 最初在自然语言处理（NLP）任务中取得了巨大成功，ViT 则将其应用于计算机视觉领域，通过将图像划分为一系列图像块（patch），并将这些图像块视为序列数据进行处理。ViT 在许多图像分类任务上表现出色，特别是在大型数据集上。

ViT 的核心思想是将图像分割成固定大小的图像块，然后将这些图像块展平并嵌入到更高维度的向量空间中。每个图像块通过线性变换被映射为一个特征向量，这些特征向量连同位置编码一起作为 Transformer 编码器的输入。

具体步骤如下：

1. 图像分割：将输入图像 $x \in \mathbb{R}^{H \times W \times C}$ 划分为 N 个不重叠的图像块，每个图像块的大小为 $P \times P$ ，因此 $N = \frac{HW}{P^2}$ 。
2. 图像块展平和嵌入：将每个图像块展平并通过线性变换映射到 d 维特征向量，得到 $z_0 \in \mathbb{R}^{N \times d}$ 。
3. 位置编码：为每个图像块添加位置编码，以保留空间位置信息，位置编码与特征向量相加形成最终输入 $z_0 + E_{pos}$ 。

4. Transformer 编码器：将上述输入序列传递给标准的 Transformer 编码器，编码器由多层自注意力机制和前馈神经网络组成。
5. 分类：使用分类标记（class token）作为输入序列的一部分，通过 Transformer 编码器的输出进行图像分类。

ViT 在大型数据集上的表现优异，并且在参数数量相近的情况下，能够比传统卷积神经网络取得更好的性能。其主要优势包括：

- 灵活性：ViT 能够处理不同分辨率和大小的图像，只需调整图像块的大小即可。
- 全局信息：自注意力机制能够捕捉图像中的全局信息，而不是局限于局部感受野。
- 可扩展性：ViT 能够轻松扩展到更大的模型尺寸和更多的参数，从而进一步提高性能。

然而，ViT 的一个主要挑战是在小型数据集上容易过拟合，因此在使用 ViT 时，通常需要预训练模型或更强的数据增强技术（如 Cutmix）。

3.3. 模型参数量估计

3.4. 优化器设置

两种架构的优化器都选择 Adam 优化器。除学习率和权重衰退以外，其余参数都为默认参数。

第 4 节 调参结果

本实验调节两个参数以尽可能提高模型性能：学习率 `lr` 和权重衰退 `weight_decay`。经过多轮调参，最终确定 resnet18 的 `lr` 为 $1e-3$ ，`weight_decay` 为 0。ViT 的 `lr` 为 $4e-4$ ，`weight_decay` 为 $1e-5$ 。

第 5 节 实验结果

5.1. 数据分析

以下图片中，**橙线**都是 **resnet18 网络**的数据曲线，**蓝线**都是 **ViT 网络**的数据曲线。

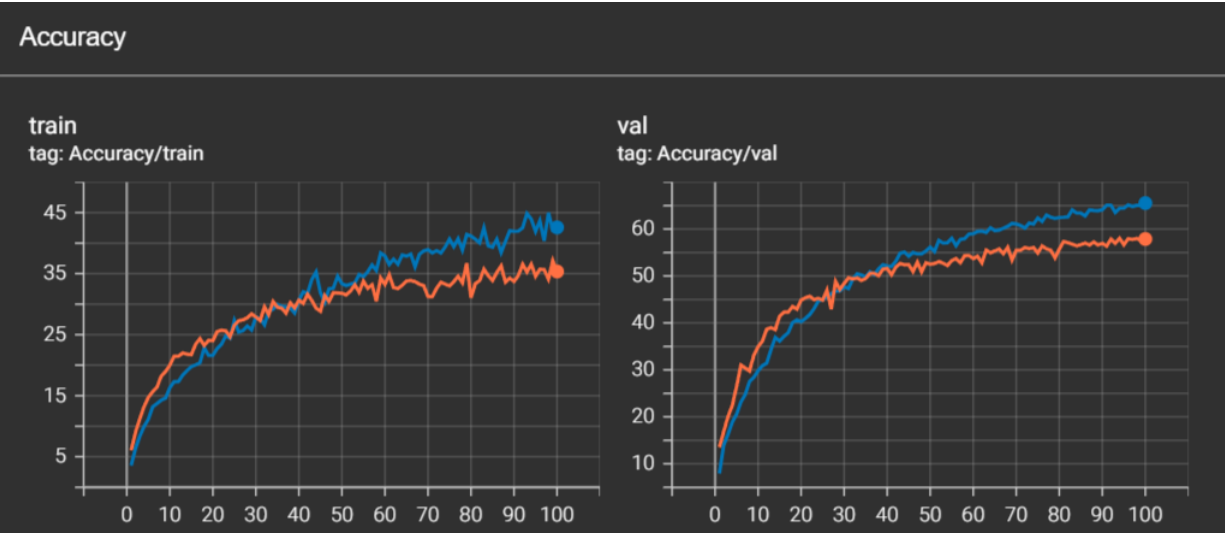


图 2.2: 准确率对比

可以看出两个模型的准确率在训练的过程中不断上升，并且都没有饱和，理论上如果训练更长时间，两者都能获得更好的效果。训练初始阶段 resnet18 性能更好，约 30 个 epoch 左右被 ViT 超越。resnet18 的最终验证集准确率为 57.88%，ViT 的最终验证集准确率为 65.53%。

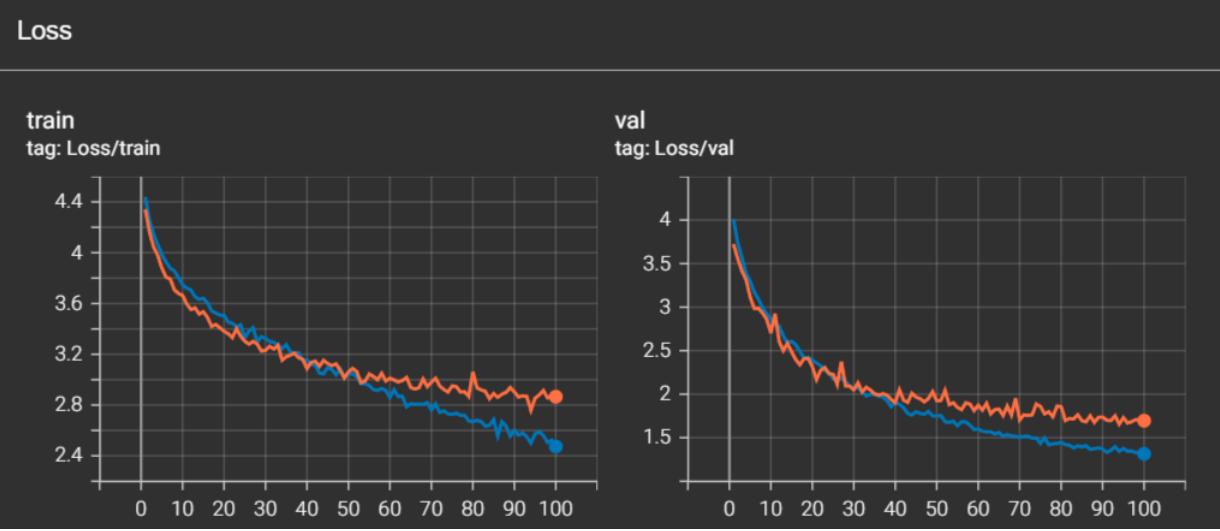


图 2.3: 损失函数对比

5.2. 结论

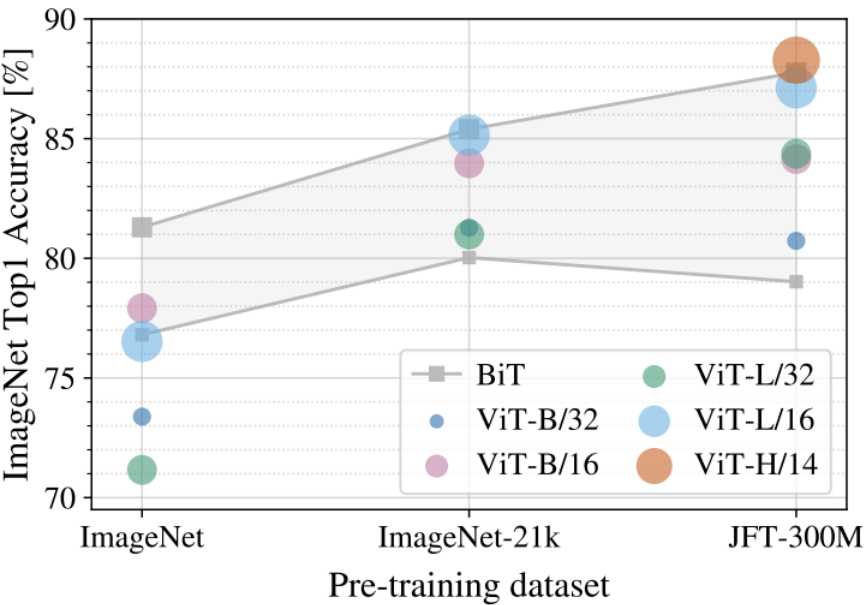


图 2.4: 数据集大小与模型性能的关系

第三章 基于 NeRF 的物体重建和新视图合成 3

第 1 节 任务描述

1. 选取身边的物体拍摄多角度图片/视频，并使用 COLMAP 估计相机参数，随后使用现成的框架进行训练；
2. 基于训练好的 NeRF 渲染环绕物体的视频，并在预留的测试图片上评价定量结果。