ResNet-18 与 Transformer

王逸群 19307110397

2022.5

GitHub repo 链接：https://github.com/quniLcs/cv-final
网盘链接：

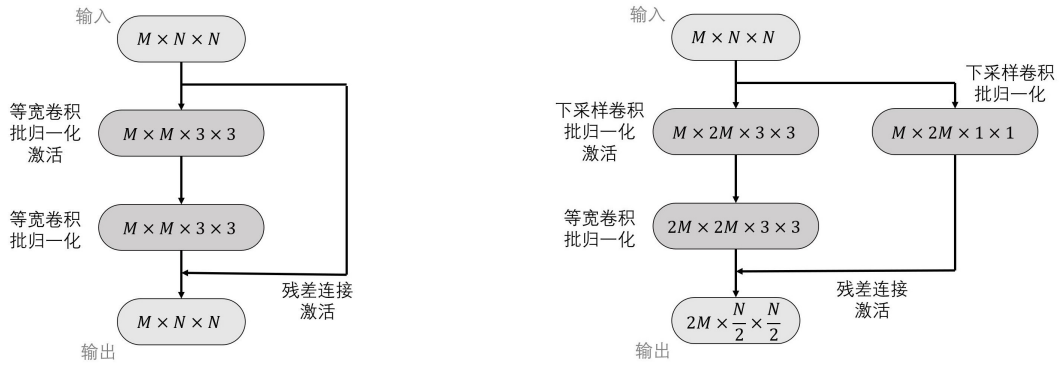
1 数据集

本项目使用 CIFAR-100 数据集，其中包含 60000 张 32×32 的彩色图片，其中训练集 50000 张，测试集 10000 张，被平均分为 100 类。

2 网络结构

2.1 ResNet

本项目使用的第一种网络结构是 ResNet-18，其中激活函数为 ReLU，最大的特征为残差连接。后者包括两种单元结构如图 1所示。



(a) 第一种单元结构　　(b) 第二种单元结构

图 1: 残差连接的两种单元结构

对于输入的图像，先进行步长为 1 的 $3 \times 64 \times 3 \times 3$ 卷积操作，并进行批归一化和激活，维度变为 $64 \times 32 \times 32$；接着通过两次第一种单元结构，维度不变；再通过第二

种单元结构，维度变为 $128 \times 16 \times 16$；再通过第一种单元结构，维度不变；再通过第二种单元结构，维度变为 $256 \times 8 \times 8$；再通过第一种单元结构，维度不变；再通过第二种单元结构，维度变为 $512 \times 4 \times 4$；再通过第一种单元结构，维度不变；最后通过全连接得到输出。由此，参数数量为 11220132.

2.2 ViT

本项目使用的第二种网络结构是 ViT(Vision Transformer)，其主体是一个 Transformer 编码器，其中激活函数为 GELU，如图 2所示。

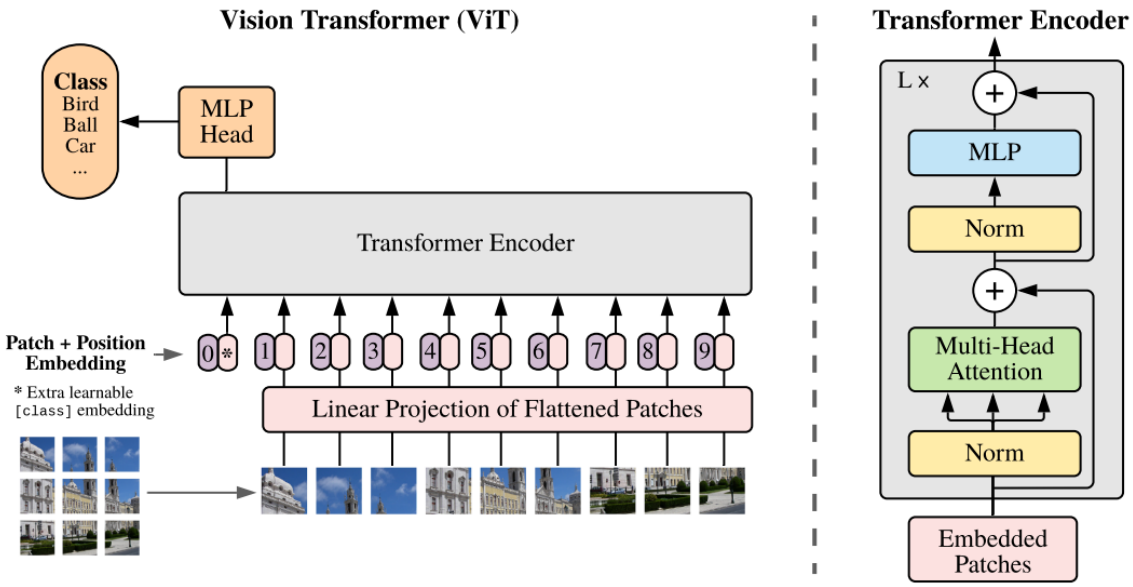


图 2: ViT 网络结构

对于输入的图像，根据给定的 patch size 划分成若干个 patch，并通过全连接层编码成给定维度的向量，即 patch 编码。这些 patch 编码和可学习的分类编码分别与各自的位置编码相加后进入 Transformer 编码器。在 Transformer 编码器中，每个 Transformer 层由层归一化、多头注意力机制、残差连接、层归一化、两层前馈神经网络、残差连接所组成。最后，分类编码对应的 Transformer 编码通过层归一化和全连接层得到输出。

参考 ViT 原论文、与 ViT 相关的 Jean-Baptiste Cordonnier 等人的工作、以及官方开源软件包中模型的默认参数，选择网络结构参数如表 1所示。值得注意的是，多头注意力机制中隐藏层的总数据维度应与模型隐藏层数据维度保持一致，因此注意力机制隐藏层的数据维度可以由模型隐藏层数据维度与注意力机制头数计算得到。由此，参数数量为 13056612。

参数名称	原论文	实际参数选择
图像边长	224	32
patch size	16	16
图像类别数	1000	100
模型隐藏层数据维度	768	512
Transformer 层数	12	6
注意力机制头数	12	8
前馈神经网络隐藏层数据维度	3072	1024
注意力机制隐藏层数据维度	64	64

表 1: ViT 网络结构参数

3 超参数设置

数据增强：裁剪、水平翻转；
学习率：由 0.02 开始每 20 个回合阶梯下降一个数量级；
优化器：带有 0.9 动量的随机梯度下降算法；
正则化参数：0.0005；
回合数：60；
批量大小：128；
每回合循环数：391；
总循环数：$60 \times 391 = 23460$；
损失函数：交叉熵损失函数；
评价指标：精确度。

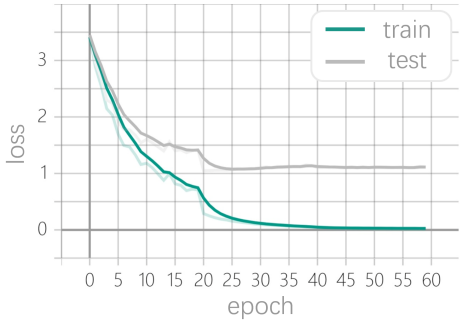
4 实验结果

实验结果如图 3和表 2所示。

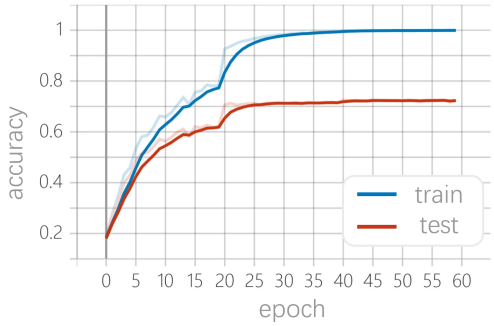
模型	训练集 top1	训练集 top5	测试集 top1	测试集 top5
ResNet	0.99904	1.00000	0.72480	0.91870
ViT	0.44518	0.75418	0.35280	0.64170

表 2: 实验结果

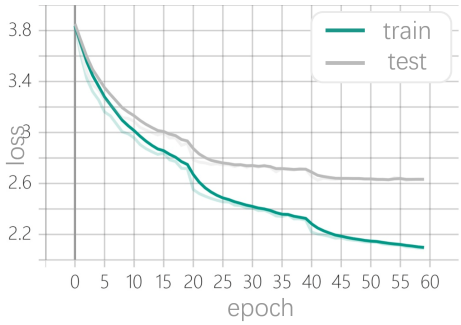
可以看到，相比 ResNet，Transformer 模型结果较差，原因是其缺乏归纳偏置，如局部性和平移等变形。尽管存在位置编码，它也是随机初始化、从零开始学习的。



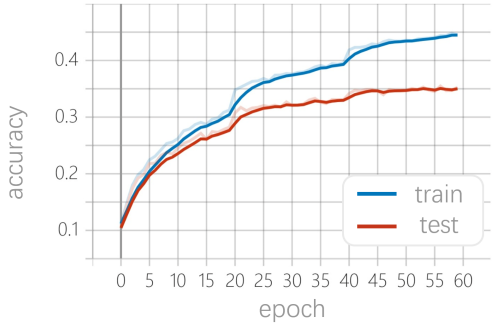
(a) ResNet loss



(b) ResNet accuracy



(c) Transformer loss



(d) Transformer accuracy

图 3: 实验结果