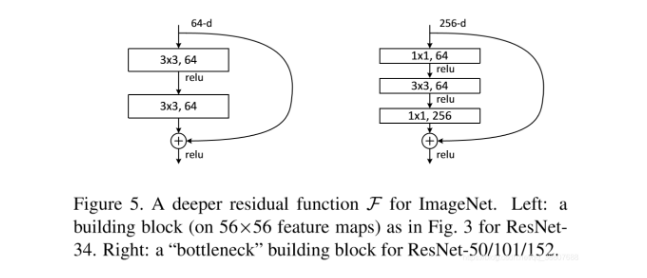
**实验一 图像分类**

1. **实验要求**

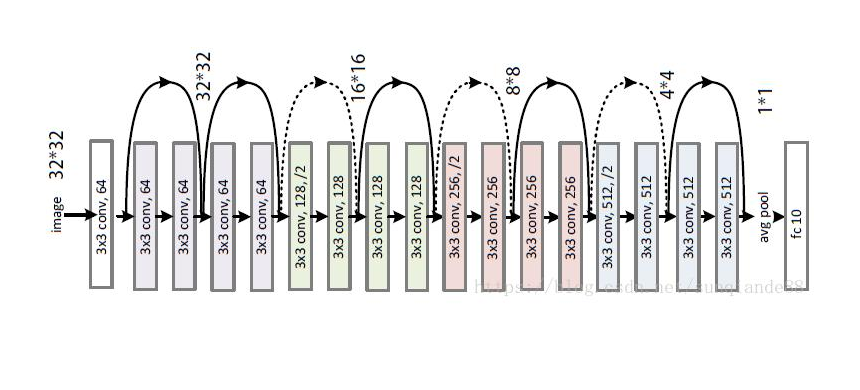
使用深度学习方法在 CIFAR10 训练集上进行训练，然后将训练后的模型在 CIFAR10 测试集上进行测试。使用的深度模型包括：

1. 一个基于 CNN 的模型，例如 ResNet，MobileNet 等；
2. 一个基于 Vision Transformer 的模型，例如 ViT，DeiT，TNT 等；如果计算资源不足可在原始模型基础上减少网络层数或者通道数；
3. 一个基于 CNN 和 Transformer 的混合模型，例如将 CNN 融合进 ViT 模型，反之亦然。
4. **实验原理**
5. ResNet原理简介

ResNet是继VGG之后出现的最具影响力的卷积神经网络，其主要特征是首次引入了残差结构（捷径连接）：



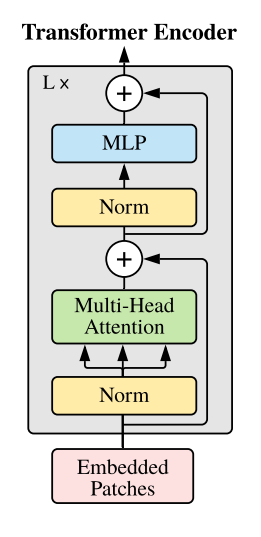
这种残差结构使得卷积层的输出能够直接受到输入的作用，从而实现更快地学习。其基本结构方式与之前已经出现过的深度卷积神经网络无异，都是卷积模块堆叠作为特征提取器，之后接一层或三层全连接层作为分类器。以ResNet18为例，其结构如下：



可以看到它由一个卷积层、8个残差卷积模块（每个模块包括两层卷积层和一个捷径连接）和一个全连接层组成，共有18层，因此称为ResNet18。

1. ViT原理简介

Transformer模型是继CNN、RNN后出现的又一类全新的神经网络模型，其特点是不需要任何卷积运算和串行计算，仅仅使用attention机制来得到输入的特征，具有非常好的长期依赖提取能力和并行能力，速度快、准确率高，使得它在自然语言处理领域尤其是序列到序列的问题上取得了非常大的成功。而Vision Transformer（ViT）是Google在2019年提出的将transformer模型基本不做修改地直接应用于处理图像分类问题的一种新的解决方案，其模型结构如下图所示：



不难发现，由于图像分类并不是一个seq-seq的任务，所以ViT并不具备原版Transformer的encoder-decoder结构，而是只保留了后者的编码器作为全部的模型。

Transformer原理的特殊性导致其只能以序列作为输入，因此ViT的输入也必须是序列，于是与CNN将整张图片直接作为模型输入不同，ViT将一张完整的图片划分成了一个个相同大小的小片，并将这些小片构成的序列作为模型的输入。在处理NLP任务时，由于Transformer在并行化过程中丢失了各输入间的时序关系，所以需要在词嵌入的基础上再加上sin函数调制的位置嵌入才能作为最终Transformer的输入，ViT中延续了这种做法，首先使用全连接层生成了上述各小片的嵌入向量，然后又与可学习的位置向量叠加，才得到了每一个小片最终的嵌入向量，各小片的最终嵌入向量组成序列，作为ViT的输入。ViT的其他运算操作与Transformer基本一致，最后使用一个MLP作为分类器。

1. AlexNet+ViT

通过上述分析，我们不难发现，CNN可以直接对矩阵表示的图片进行处理，而ViT只能对经过嵌入处理的图片（图片的嵌入向量）进行处理，标准ViT中使用全连接层对图片进行嵌入，那么由于常见CNN的输入也是一个向量，能不能直接将CNN的输出当做图片的嵌入向量作为ViT的输入呢。

于是本次实验中，我们欲将ResNet18与ViT结合使用，也即将图片先输入ResNet18，得到一个一维的向量，再将该向量视为ResNet18提取到的图片的特征向量，并与可学习的位置向量混合后得到图片最终的嵌入向量送入ViT，这个过程中相当于原图片被分割为了若干1x1的小片。实际实现时，由于计算设备所限，为加快训练速度，我们最终将ResNet18换成了AlexNet，相比于ResNet18，前者没有捷径连接，层数更少，只有五层卷积，其他方面在实验中均无区别，下表给出了实验中所用AlexNet的结构：

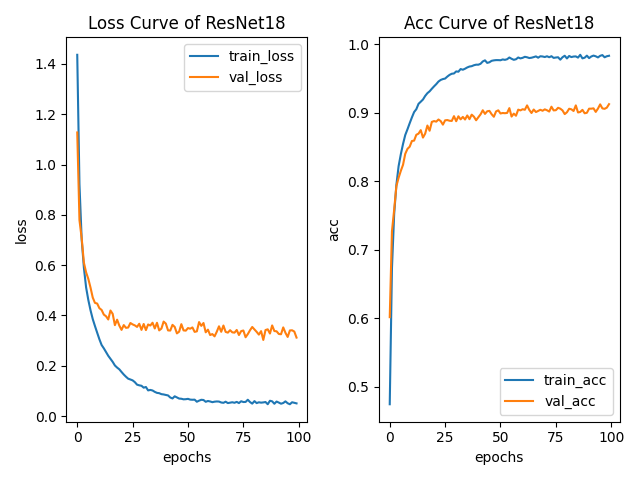
|  |  |
| --- | --- |
|  | Inputs |
| Layer1 | 5x5Conv,64,2+ReLU+2x2MaxPool,2 |
| Layer2 | 3x3Conv,192+ReLU+2x2MaxPool,2 |
| Layer3 | 3x3Conv,384+ReLU |
| Layer4 | 3x3Conv,256+ReLU |
| Layer5 | 3x3Conv,256+ReLU+2x2MaxPool,2 |
| Layer6 | FC |

其中“5x5Conv,64,2”表示卷积核大小为5x5，通道数为64，步长为2，无特别说明，默认卷积中步长为1、padding为1；“2x2MaxPool,2”表示最大池化，池化窗口的大小为2x2，步长为2.

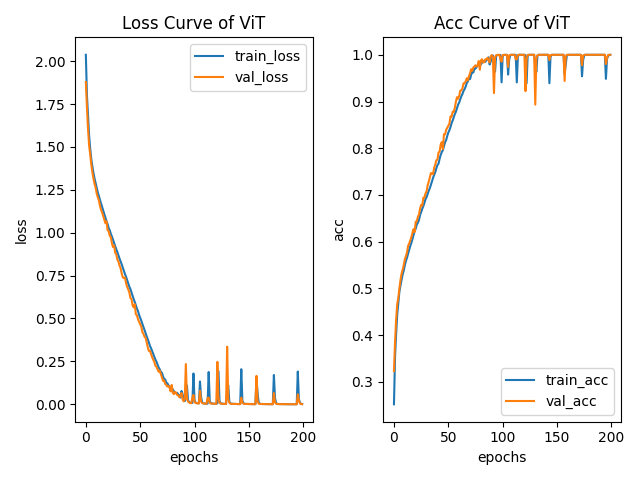
1. **实验结果**

本次实验共采用三种模型来实现图像分类任务，分别是ResNet18、ViT和二者的混合模型，其中混合模型首先使用ResNet18对输入图片做特征提取，然后再将其输出结果直接作为ViT的输入。实验中所用数据集为CIFAR10，该数据集包含 10 类，共计 60000 张大小为 32x32 的标注图像。其中，训练集有 50000 张图像，测试集有 10000 张图像。

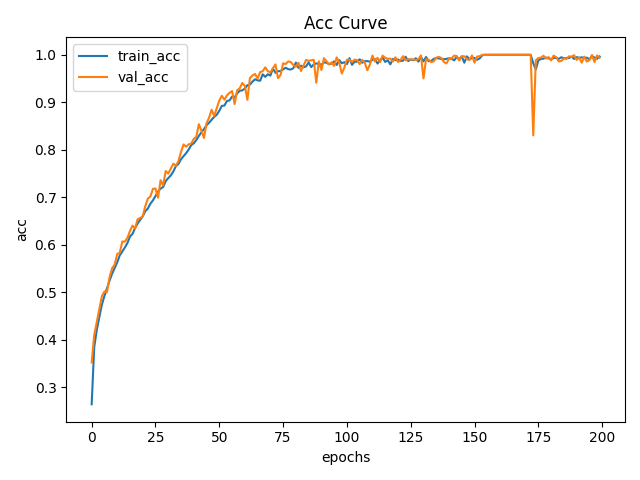
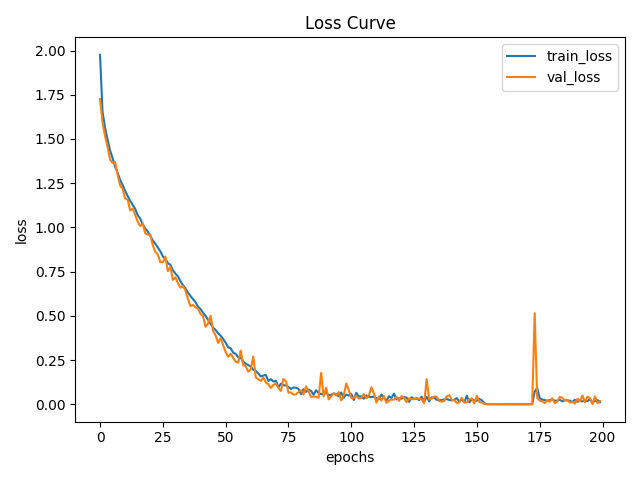
各模型均采用SGD方法训练，损失函数均为交叉熵，图1、图2分别给出了ResNet18在该数据集上的损失曲线和准确率变化曲线，图3、图4分别给出了ViT在该数据集上的损失曲线和准确率变化曲线，图5、图6分别给出了混合模型在该数据集上的损失曲线和准确率变化曲线。表1中前三列给出了各模型训练的相关参数，最后一列展示了它们在数据集上的top1准确率。



**图1 ResNet18训练情况**



**图2 ViT训练情况**



**图3 混合模型训练情况**

**表1 各模型训练参数及准确率情况**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 模型 | 学习率 | 批大小 | 训练轮数 | Top-1 |
| ResNet18 | 0.01 | 128 | 50 | 91.26 |
| ViT | 3e-5 | 128 | 200 | 100.00 |
| ResNet18+ViT | 3e-5 | 128 | 200 | 100.00 |

从训练结果不难看出，对于CIFAR10这种小数据集，ResNet18已经可以实现不错的效果，相比之下，ViT在准确率方面虽然有优势，但相比ResNet而言，其训练时间明显更长，需要更多的迭代次数才能够收敛。因此一般数据集上ResNet完全可以满足需求。此外，由于ViT在CIFAR10数据集上取得了相当不错的训练效果，所以与CNN混合并没有体现出准确率方面的优势，并且模型更大、需要花更多时间训练。总体来看，CIFAR10数据集并不适用于比较区分这三种模型的性能。

还需要注意的是，本实验中在将CIFAR10数据集中的图片输入模型前先做了标准化处理，对于ResNet，图片三个通道的均值设置为0.4914, 0.4822, 0.4465，方差设置为0.2023, 0.1994, 0.2010，而对于ViT和混合模型，三个通道的均值方差全部设置为0.5，如果不做上述标准化处理，三个模型则都会出现过拟合现象。