华中科技大学

计算机视觉实验三报告

姓　　名：

专　　业：

班　　级：

学　　号：

指导教师：

分数	
教师签名	

年月日

目录

实验一：基于剪枝算法的深度神经网络压缩

1.1 任务要求

对实验二构建的 CIFAR-10 数据集分类神经网络进行权重剪枝实现模型压缩。

例如：若最后一层卷积层的权重大小为 D×3×3×P，输出特征图大小为 M×N×P，在测试数据集上对 P 个输出特征图的神经元激活（test_dataset_size×M×N）求平均并进行排序。按激活水平由低到高，对前 K 个神经元权重进行剪枝，K=1 to P-1。剪枝后的卷积层权重大小为$D \times 3 \times 3 \times (P - K)$，测试此时神经网络分类准确率。

可将待剪枝的神经元权重、偏置设为 0，即相当于神经元剪枝而不用改变网络架构。

1.2 实验内容

本实验中，我对实验二中训练的 ResNet9 和 vgg13_bn 两个网络进行剪枝。

剪枝算法的相关代码实现在 tools/prune.py 文件中，网络模型的代码实现在 modules/文件夹下，实验二训练好的各模型权重文件在 work_dir/文件夹下，实验过程在 exp3_pruning.ipynb 文件中。

1.2.1 ResNet9

1、修改实验二中实现的模型代码（modules/mymodel.py），在前向传播时保存最后一层卷积层的输出特征图，并添加新的、可调用的方法返回该特征图。网络模型结构和最后一层卷积层（红框标出）如下图所示:

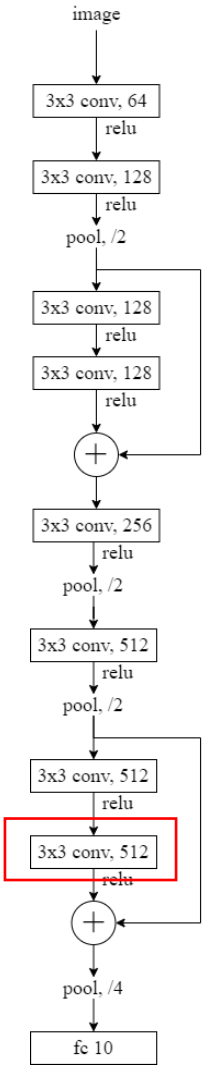

图 1-1：resnet9 网络模型结构图和其最后一层卷积层

由图可知：最后一层卷积层的输出特征图大小为 4×4×512，即有 512 各神经元。

2、加载 CIFAR-10 测试数据集和实验二训练好的模型权重；

3、在测试数据集上对最后一层卷积层的输出特征图求平均，并可视化。共 512 个平均特征图，每个特征图大小为 4×4，用灰度图表示，如下图为 16 行 32 列：

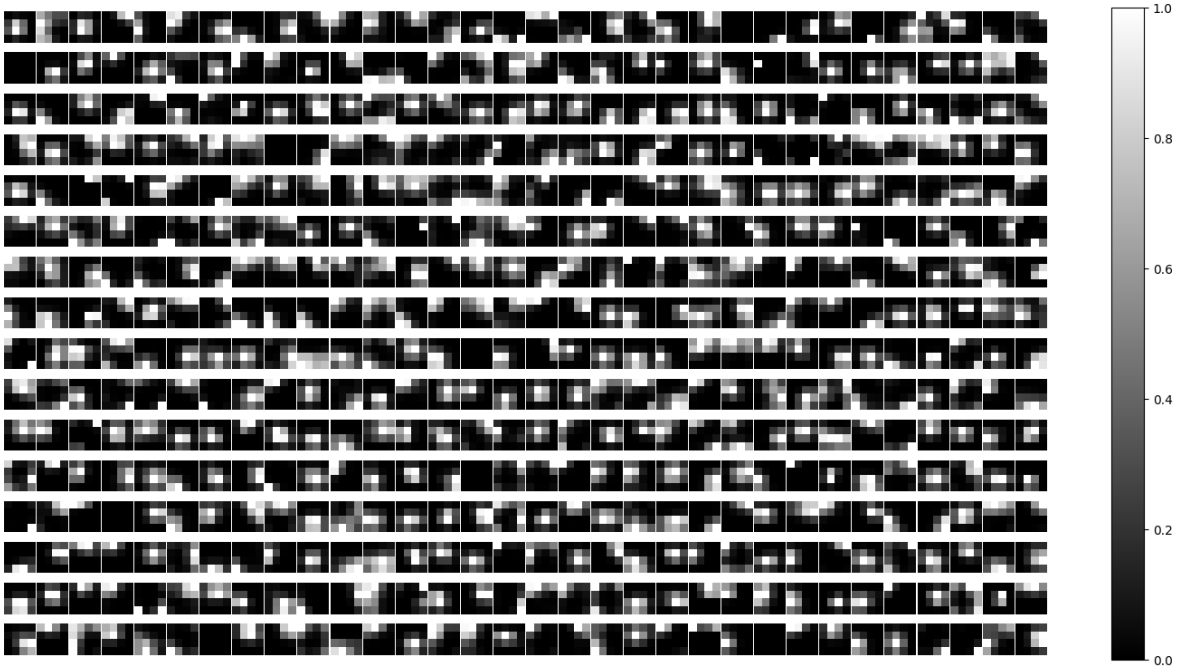

图 1-2：resnet9 最后一层卷积层的输出特征图

由上图可以看出：resnet9 最后一层卷积层的神经元激活水平较高，很少有完全没激活的神经元。

3、开始神经元剪枝。先根据 2 中的平均输出特征图，对 512 个神经元的激活水平由低到高进行排序。然后优先剪枝激活水平低的神经元，每次剪枝 32 个神经元（即将这 32 个神经元的权重和偏置置为 0），并测试每次剪枝后模型的准确率，直至 512 个神经元全部剪枝完。

剪枝过程的打印信息如下：

```
0 neurons: 90.56%
32 neurons: 90.50%
64 neurons: 90.53%
96 neurons: 90.28%
128 neurons: 90.13%
160 neurons: 90.04%
192 neurons: 90.13%
224 neurons: 89.98%
256 neurons: 89.66%
288 neurons: 89.50%
320 neurons: 89.40%
352 neurons: 89.14%
384 neurons: 88.90%
416 neurons: 88.29%
448 neurons: 87.63%
480 neurons: 87.41%
512 neurons: 86.52%
```

图 1-3：resnet9 剪枝过程打印信息

将上述信息画成折线图，横坐标为剪枝神经元个数，纵坐标为模型准确率，如下图所示：

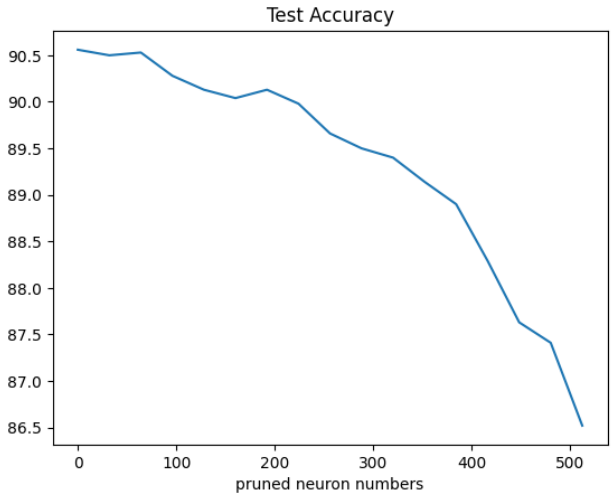

图 1-4：模型准确率与剪枝神经元个数折线图

分析：

（1）由图 1-2：resnet9 最后一层卷积层的输出特征图可知，resnet9 最后一层卷积层的神经元激活水平较高，很少有完全没激活的神经元。所以刚开始剪枝时，模型准确率就开始下降；

（2）由于剪枝顺序是按照神经元激活水平由低到高，即剪枝的神经元激活水平越来越高，对模型影响越来越大，所以模型准确率下降得越来越快；

（3）由于 resnet9 模型中的 shortcut connections，即使将最后一层卷积层的所有 512 个神经元全部剪枝掉，模型也只是退化成了 6 层的网络，只有最后的 building block 失去作用，所以在剪枝掉最后一层卷积层的所有神经元后，模型仍然有 85%以上的准确率。

1.2.1 VGG13_bn

1、修改实验二中实现的 vgg 模型代码（modules/vgg.py），在前向传播时保存最后一层卷积层的输出特征图，并添加新的、可调用的方法返回该特征图。

对于 vggnet，最后一层卷积层的输出特征图大小为 2×2×512，即有 512 各神经元。

2、加载 CIFAR-10 测试数据集和实验二训练好的模型权重；

3、在测试数据集上对最后一层卷积层的输出特征图求平均，并可视化。共 512 个平均特征图，每个特征图大小为 2×2，用灰度图表示，如下图为 16 行 32 列：

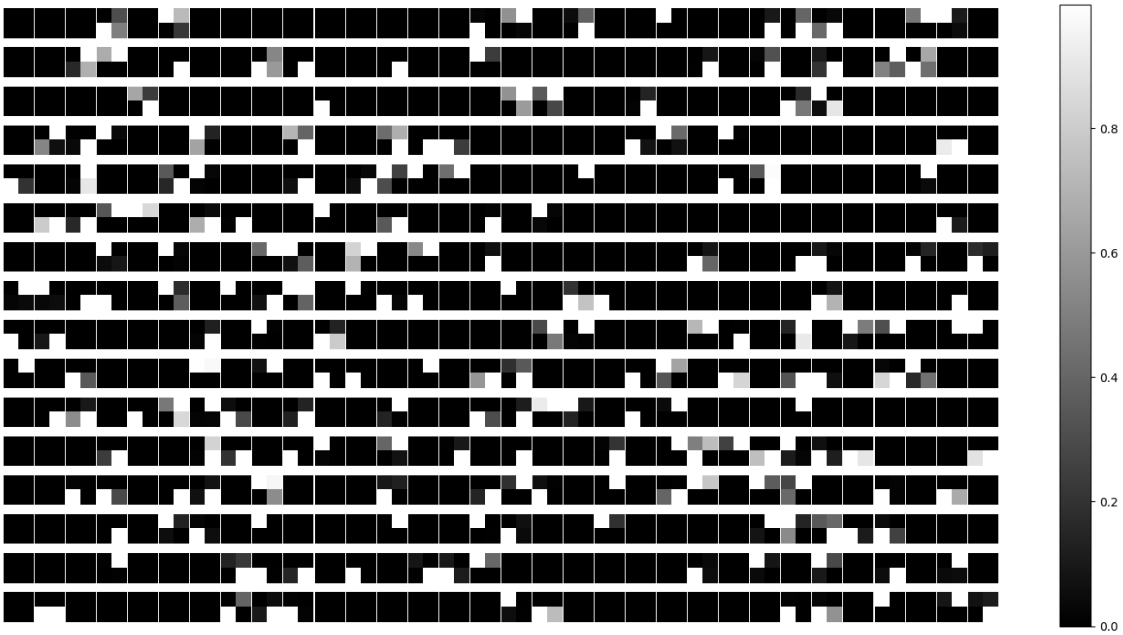

图 1-5：vgg13_bn 最后一层卷积层的平均输出特征图

由上图可以看出：vgg13_bn 最后一层卷积层的神经元激活水平较低，有大量完全没激活的神经元。

3、开始神经元剪枝。先根据 2 中的平均输出特征图，对 512 个神经元的激活水平由低到高进行排序。然后优先剪枝激活水平低的神经元，每次剪枝 32 个神经元（即将这 32 个神经元的权重和偏置置为 0），并测试每次剪枝后模型的准确率，直至 512 个神经元全部剪枝完。

剪枝过程的打印信息如下：

```
0 neurons: 92.18%
32 neurons: 92.18%
64 neurons: 92.25%
96 neurons: 92.24%
128 neurons: 92.09%
160 neurons: 91.92%
192 neurons: 91.94%
224 neurons: 91.28%
256 neurons: 91.32%
288 neurons: 91.20%
320 neurons: 91.08%
352 neurons: 90.54%
384 neurons: 90.23%
416 neurons: 82.94%
448 neurons: 83.38%
480 neurons: 39.02%
512 neurons: 10.00%
```

图 1-6：vgg13_bn 剪枝过程打印信息

将上述信息画成折线图，横坐标为剪枝神经元个数，纵坐标为模型准确率，如下图所示：

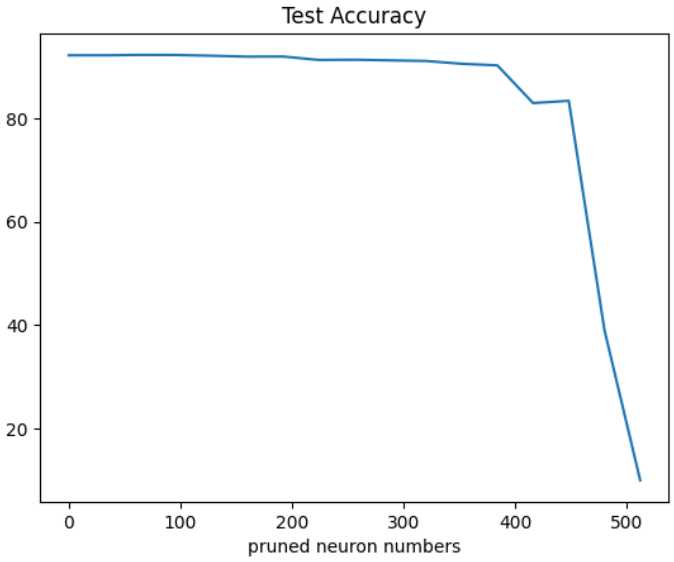

图 1-7：模型准确率与剪枝神经元个数折线图

分析：

（1）由图 1-5：vgg13_bn 最后一层卷积层的平均输出特征图可知，vgg13_bn 最后一层卷积层的神经元激活水平较低，有大量完全没激活的神经元。所以剪枝前 400 个神经元时，模型准确率没有什么变化，只轻微下降了一点；

（2）由于 vgg13_bn 模型中没有类似 resnet 的 shortcut connections，剪枝激活的神经元时，模型准确率急剧下降。并且剪枝完全部的 512 个神经元后，最后一层卷积层及其前面的所有卷积层全都失去作用，模型准确率只有 10%。