1 任务 1：PCA 数据降维

本文使用的 CIFAR-10 数据集含有 50k 张训练图像和 10k 张测试图像，每张图像的分辨率为 32*32，共有 RGB 三个通道，对图像作 flatten 操作并加上偏置项，则特征维度为：32*32*3 + 1 = 3073，维度过高，处理起来的计算复杂度大，且图像中存在大量信息冗余，可以使用 PCA 降维。

具体来说，先对数据矩阵进行奇异值分解，求得奇异值，再对奇异值平方得到特征值并按降序排列，随后累计求和与特征值的总和计算比值，反映信息的保留比例，这里保留比例分别为：40%，60%，80% 以及 100%，得到的特征维数分别为 4，10，35，3073（最后 1 维均为 SVM 的偏置项)。

2 任务 2：多分类 SVM 的相关讨论

SVM 是十分经典的传统机器学习分类器，相对 logistic regression，SVM 会关注与决策边界的鲁棒性，即会考虑最大化决策边界间隔，决策边界间隔约束的强弱取决于超参数 C 的大小，C 越大，对于决策边界间隔的约束也就越强，决策边界间隔约束的存在会让 SVM 拥有更低的 variance。同时，由于核函数的引入，可以让 SVM 获得非线性的决策边界，以解决不能线性可分的数据。

首先看看不同降维程度特征对性能的影响，先采用了 5-fold 交叉验证方式，得到的结果分别为:

4-d	0.245	0.240	0.247	0.243	0.249
10-d	0.313	0.329	0.319	0.319	0.322
35-d	0.374	0.382	0.382	0.376	0.377
3073-d	0.374	0.380	0.383	0.373	0.373

Table 1: 5-fold cross-validation

综合来讲，选择 PCA 降维到 35 维，可以达到计算复杂度和性能的 trade-off。

下面再看看不同降维程度在测试集上的表现，为了让横轴间差距更小，这里对特征维度取了对数进行画图，结果如 fig.1所示，可以看到前期随着保留的信息比率的升高，正确率上升，当保留信息比率超过 80% 后，剩下的 20% 信息由 90% 以上的参数贡献，反而让模型难以学习，性能变差。

以上结果均采用的是线性 SVM，下面讨论使用高斯核函数的 SVM 在不同 C 值下的测试集表现，这里采用了 5-fold 交叉验证的结果，即特征降维维度选择为 35 维，结果如 fig.2所示，可以看到，C 的升高可以让模型在测试集上性能能好，即 variance 更低。

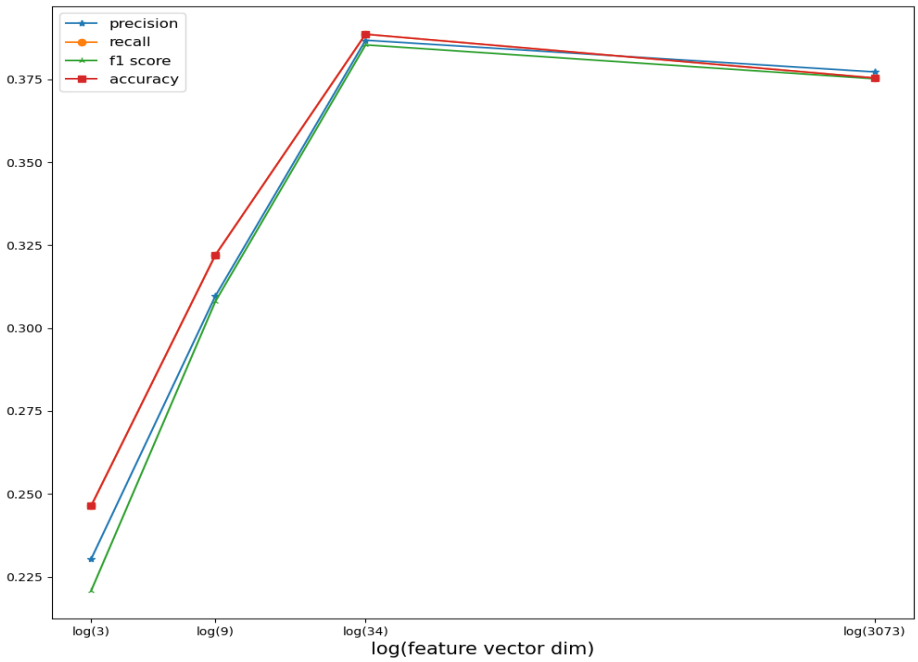

Figure 1: test scores

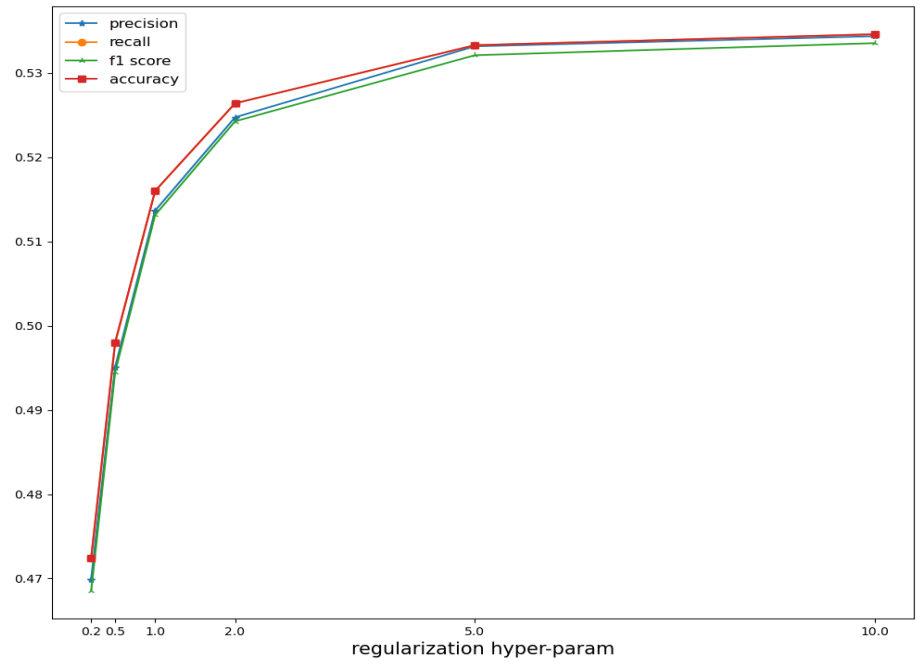

Figure 2: RBF-kernel SVM test scores

3 任务 3：神经网络的相关讨论

本部分采用了四种架构的卷积神经网络，测试其在 CIFAR-10 数据集上的表现，分别采用了纯粹的卷积神经网络、引入 Residual 模块、引入 Reidual 模块和 SE 模块以及引入 ResidualShuffle 模块，后面三者分别借鉴了 ResNet、SENet 以及 ShuffleNet。每个模型提供了两种设置，即按参数量/每个 stage 的通道数分为两种，参数量更少/每个 stage 的通道数量更少的模型叫做 small 模型，结构如下表所示：

	basic block	layers per stage	channels per block	parameters/M
conventional	convblock	[1,2,3,3]	[64,256,512,512] small: [64,128,256,256]	19.08(small:4.85)
residual	residual block			34.46(small:8.72)
se_res	se_residual block			34.68(small:8.78)
shuffle_res	shuffle_residual block			3.99(small:1.06)

Table 2: model architecture

每个模型的 basic block 如下图所示，每个模型的下参样均采用 stride=2 的卷积，并完成 stage 间通道数的转换，basic block 中是不改变通道数的：

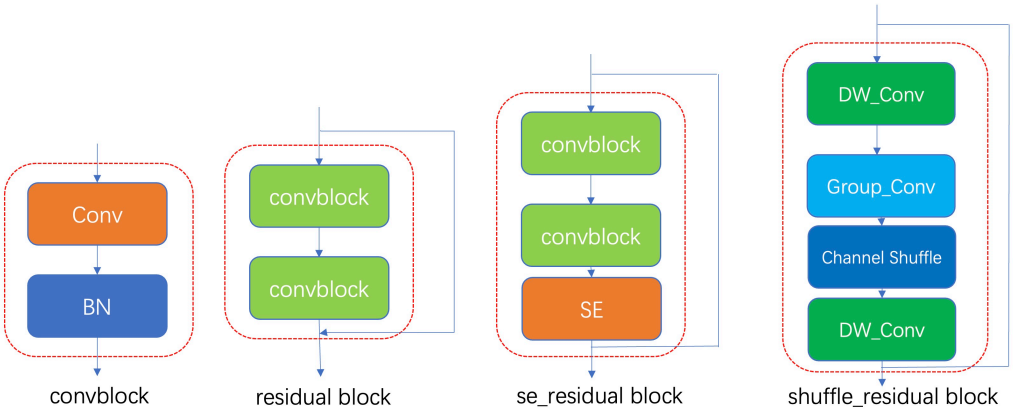

Figure 3: Basic Block

模型的最佳测试准确率和模型参数量的关系如下图所示，左图为正常模型，右图为 small 模型：

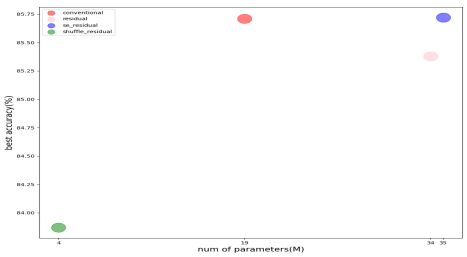

Figure 4: base model

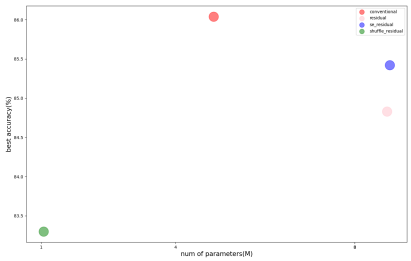

Figure 5: small model

具体准确率如下表所示：

	conventional	residual	se_residual	shuffle_residual
base	85.71	85.38	85.72	83.87
small	86.04	84.83	85.42	83.3

Table 3: test accuracy

模型在学习时的超参数保持一致，均为，优化器采用 Adam，学习率 1e-3，weight-decay 为 1e-5，epoch 数为 30。

4 任务 4：结果分析

下面均选择 conventional CNN 的学习曲线进行分析，首先分析 base 版本，其学习曲线如下图所示：

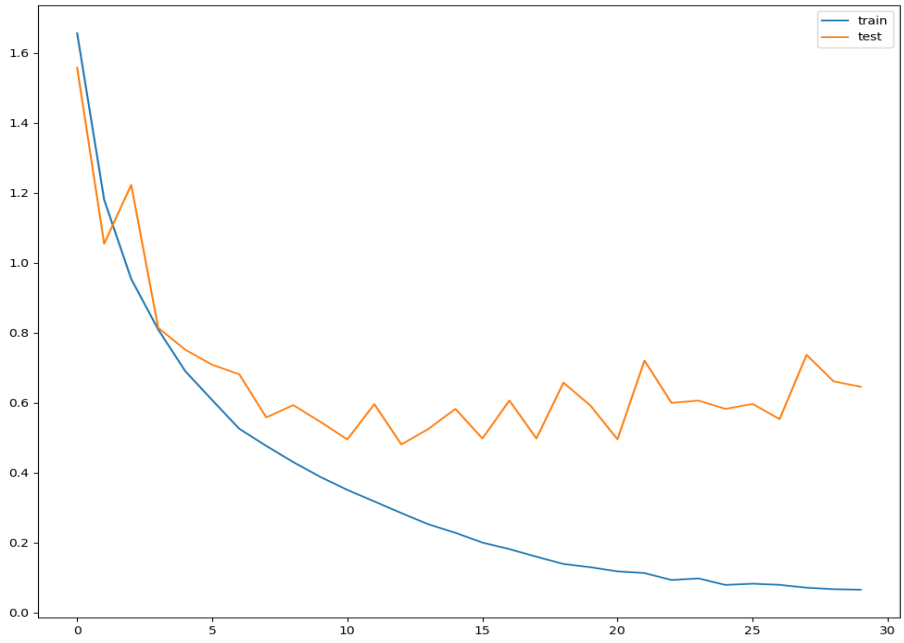

Figure 6: learn curve of conventional CNN (base version)

可以看到，随着训练的进行，train loss 是不断在下降直至收敛的，而 test loss 经历了很大程度的波动，甚至波动越来越大，这是比较轻微的 overfitting 的现象。

再看看 small 模型的学习曲线：

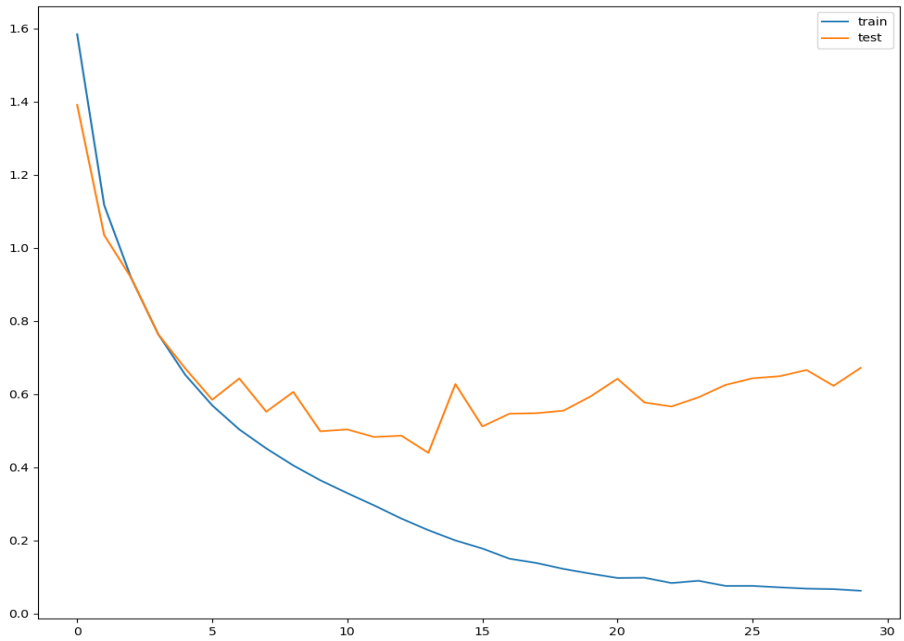

Figure 7: learn curve of conventional CNN (small version)

可以看到，在相同的超参设定下，small 版本的模型明显过拟合程度更低，说明模型的过拟合程度是和模型的大小相关的。

结合 task2 SVM 板块的分析，我们在 task2 中为了达到计算复杂度（时间成本）和 precision、recall、f1、accuracy 等指标（性能）的 trade-off，我们选择了降维信息保留率为 80%（即 35 维）的特征，同样，这里我们也可以根据 task3 中模型参数量和 test accuracy 性能的 trade-off，选择简单的 conventional CNN 或 shuffle-residual CNN 作为我们的实际使用的模型，尤其是 small 版本的 shuffle-residual CNN，由于分组卷积（Group Conv）和深度可分离卷积（DepthWise Conv），其参数量仅有 1M，而由于采用 residual connection 和 channel shuffle 的操作，其性能仍能达到 83.3%，是非常高性价比的选择。

对于 overfitting 问题，非常常用的方法是本文中用到的 BN，weight decay，以及本文未用到的 dropout，根据 task2 中 RBF-kernel 的 SVM 在不同 C 值下的性能可知，L2 regularization（和 weight decay 类似）对于 overfitting 是一个不错的解决方案，因为从 SVM 实验中可以看出合适的 C 对模型的 variance 有较大影响，而 variance 正好和模型的泛化性一定程度正相关。

因此，对于 weight decay 的调整，或是引入 dropout 操作隐式改变模型 variance，都是缓解 overfitting 的重要方法。