1 目标任务

1. 利用神经网络完成 CIFAR-10 数据集的图像分类任务。
2. 将训练集划分，所有类别小于等于5的数据仅保留 10%，剩余部分不变。重新训练，比较结果，并尝试改进。

2 CIFAR-10 数据集

数据集包含了 10 类不同物体的图片（32×32， RGB）和对应的类别标签。训练集中每类数据各 5,000 对，共 50,000 对数据；测试集中每类数据各 1,000 对，共计 10,000 对数据。

为了对输入图片进行分类，模型的输入是 32×32 大小的 RGB 图片，输出则为类别预测。

3 CIFAR-10 数据集图像分类

3.1 网络结构

3.1.1 ResNet50

首先，尝试了 ResNet50 [1] 作为分类网络，效果优异，但是训练较慢，为更好地进行对比，故使用更简单的网络。

3.1.2 Simple Classifier

我设计了更简单的基于卷积的分类网络，其网络结构如下。它也能达到较好的效果。

```
nn.Conv2d(3, 64, 5)
nn.BatchNorm2d(64)
nn.ReLU()
nn.MaxPool2d(2, 2)
nn.Conv2d(64, 256, 5)
nn.BatchNorm2d(256)
nn.ReLU()
nn.MaxPool2d(2, 2)
nn.Linear(6400, 256)
nn.ReLU()
nn.Linear(256, 96)
nn.ReLU()
nn.Linear(96, 10)
```

3.2 对抗过拟合

训练过程中出现明显过拟合（训练损失下降，测试损失不变），故尝试多种方法解决。其中包括数据增强（随机翻转，随机裁剪）和 weight decay。前者从数据层面，后者从参数层面，共同引导模型关注关键特征，提升泛化能力。

4 非平衡数据集（CIFAR-10 IMBAlanced）图像分类

数据划分后，不同类别间出现不平衡。且训练集和测试集分布偏移。直接训练会导致不同类的关注度差异较大，直接表现为放大的劣势类错误量带来的高测试错误率。

4.1 再平衡数据集（CIFAR-10 REBAlanced）

原计划使用带权交叉熵损失，对不同类的惩罚加权，从而平衡各类对模型的影响。但 `Jittor` 的带权交叉熵行为不稳定（实测全一权重和不输入权重的训练结果不同），此外也避免劣势类异常点导致梯度爆炸，就改为对劣势类过采样，使得训练集中各类别的数据量相同的方法。由此得到的数据集称为再平衡数据集。

全部实验结果详见 `./checkpoint` 目录。

4.2 重述：对抗过拟合

REBA 数据集中存在大量重复样本，因此极其容易过拟合。此时一方面采用数据增强增加样本多样性，另一方面采用 weight decay 限制参数的增长，在均等关注各类的同时保障模型的泛化能力。

5 实验结果

实验结果如表 1, 2, 3 所示，此外，在表 3 的设置下最佳的实验结果为 F.3+C 和 weight decay 为 $5e-3$ 的组合，训练损失 0.5177，测试损失 0.8331，测试准确率 **72.73**%，比 IMBA 数据集的直接训练结果提升10.64%。

Model	ResNet50	Simple
Loss	**0.5786**	0.6540
Acc.	**86.00**%	78.20%

Table 1: Result of different models on CIFAR-10. Test losses and accuracies are reported.

Variation	Original	IMBA	REBA
Loss	0.6540	1.4708	1.7856
Acc.	78.20%	62.09%	62.68%

Table 2: Result on variations of CIFAR-10 with Simple Classifier.

Train Loss *Test Loss* *Test Acc.*	Augmentation				
Weight Decay	None	C	F.3+C	F.5+C	F.9+C
0	0.1198 2.7436 57.57%	0.5312 1.3350 60.60%	0.7723 1.1128 63.38%	0.8797 1.1291 61.61%	0.8810 1.3040 55.70%
$1e-3$	0.0263 1.7856 62.68%	0.1329 1.3967 66.66%	0.3496 1.0275 69.99%	0.6150 0.9844 66.67%	0.4338 1.1705 64.12%
$1e-2$	0.8659 1.1954 64.73%	0.4323 0.9248 69.04%	0.6833 **0.8616** **70.40**%	0.8101 0.9221 67.84%	0.8659 1.1295 59.05%

Table 3: Result of anti-overfitting methods with Simple Classifier on CIFAR-10 REBA. F.X stands for random flip (vertical and horizontal) with total probability of $0.X$. C stands for random crop.

6 附录

6.1 训练细节

默认情况下：学习率 $1e-2$，batch size 1024，共训练 $5,000,000$ 迭代，从第 $50,000$ 次迭代起，每 $200,000$ 迭代降低学习率为 0.5 倍。weight decay $1e-3$。数据增强不启用；若启用，方式为随机翻转和随机裁剪。

6.2 数据增强的使用场景

数据增强能增加样本多样性从而提高表现，但它是基于数据的，因此受到数据集本身影响。在 IMBA 数据集上，类别不均导致增强的表现集中体现在优势类，损失下降的同时准确性也下降了。

数据增强对模型的表达能力有要求。在 CIFAR-10 数据集上，数据增强使 Simple Extractor 的性能不升反减，却能使 ResNet50 的性能提升（Loss: 0.3838, Acc.: 87.35%）。

6.3 IMBA 数据集的实现

直接使用 mask 会导致 batch size，引发训练不稳定，同时也不能保证均匀、准确地删去九成劣势数据。同时为了便于实现 REBA 数据集，我采用先删减后读入的方式实现 IMBA 数据集。

全部实验结果详见 `./checkpoint` 目录。