深度学习第二次小作业：MNIST

519030910400 朱皓怡

1. 环境

在此次任务中，我将使用 Paddle 框架解决 MNIST 手写数字识别任务。为了更明显地展示我的方法效果，所有实验中我都**不使用**任何的数据增强。为了公平的对比，所有实验的都使用 Adam 优化器训练 40 个 epoch，学习率设为 0.001 并且使用 $T_{max} = 100$ 的 CosineAnnealing 衰减策略，batchsize 都为 200。本次任务中所有实验都使用基于 ResNet-50 的 backbone。

2. 方法

为了解决样本不均衡的问题，我一共尝试和实现了三种方法。下面一一进行介绍。

2.1. 基于 batch 内分布的权重

一个最自然的想法就是在损失函数上给少数样本更大的权重。而这里我基于 batch 的采样来实现这一点。具体来说，对每一个 batch 内的样本，给每个类别加上与 batch 内该类别样本数成反比的权重，这样一来，就可以使得网络更关注少数样本：

$$weight_i = C \cdot \frac{1/N_i}{\sum_j 1/N_j}$$

其中 $weight_i$ 是给第 i 个类别的 loss 加的权重，C 代表总的类别个数，而 N_i 表示 batch 内第 i 个样本的个数，C 和分母是为了权重之和保持不变的归一化系数。

2.2. 标签平滑

标签平滑（label smoothing）[1, 2] 是一种用来缓解过拟合的方法，可以起到正则化和平衡样本权重的效果。一个直觉的想法是，当样本不均衡的时候，网络会更容易过拟合，所以通过标签平滑提高网络的泛化性能或许能够有所帮助。具体来说，标准的交叉熵函数的标签使用的是 one-hot 编码：

$$y_i = \begin{cases} 1, & i = \text{ target} \\ 0, & i \neq \text{ target} \end{cases}$$

而标签平滑通过向上述 0-1 分布中加入均匀分布的噪声来使得避免模型对正确标签过于自信（即过拟合），它相当于使用 soft one-hot 编码的标签：

$$\hat{y}_i = \begin{cases} 1-\epsilon, & i = \text{ target} \\ \epsilon/K, & i \neq \text{ target} \end{cases}$$

其中 K 是类别总数，ϵ 是一个小的超参数。

然而，我在实际实验中发现（Sec. 3），直接使用标签平滑反而会使得结果下降。我想这是因为标签平滑同时作用于所有样本，反而拉大了不平衡样本之间性能的差距。于是我对其进行改进，只对少数样本（即 0-4）做标签平滑。实验结果（Sec. 3）证明如此就可以有效提升模型性能。

2.3. 自监督预训练

[3] 证明了利用自监督预训练可以在类别不均衡的情况（长尾分布）下大大提升模型的性能。受此启发，于是有了这一部分。具体来说，针对 MNIST 数据集的特点，我采用了 [4] 中的自监督任务，即将每张图片随机旋转 90 度的整数倍，并让网络预测是哪一种旋转。需要注意的是，数字 6 和数字 9 具有对称性，所以在实际实施的时候，我给数字 6 和 9 的旋转标签不是 one-hot，而是给两个合理值各赋为 0.5。

3. 实验

针对上述方法，我进行了一些消融实验，所有实验都在同样条件下进行（Sec.1）。结果如表 1所示。

从表中可以看出，我提出的几个方法确实都

数据集	方法	测试集上准确率 (%)
原始 MNIST	baseline	99.37
划分 MNIST	baseline	98.95
	+batch 内权重	99.02(+0.07)
	+ 给 0-9 加标签平滑	98.90(-0.05)
	+ 给 0-4 加标签平滑	99.07(+0.12)
	+ 自监督预训练	99.16(+0.21)
	+batch 内权重 + 给 0-4 加标签平滑	99.07(+0.12)
	+batch 内权重 + 自监督预训练	99.29(+0.34)
	+ 给 0-4 加标签平滑 + 自监督预训练	99.20(+0.25)
	+batch 内权重 + 给 0-4 加标签平滑 + 自监督预训练	**99.33(+0.38)**

表 1: 所有方法的消融实验结果。标签平滑中 $\epsilon = 0.1$。

是有效的，其中自监督预训练的效果最明显。实际实验中我发现，batch 内加权重是一个能稳定提升性能的方法，而标签平滑虽然只给 0-4 加可以有一定提升，但是实际过程中我发现它很不稳定，有的时候仍然会效果不佳（表格中取了多次实验中最好的结果），我猜测可能它对超参数比较敏感或者需要更小的学习率和更多的 epoches 来收敛到稳定。

4. References

[1] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in International conference on machine learning. PMLR, 2015, pp. 448–456.

[2] Rafael Müller, Simon Kornblith, and Geoffrey Hinton, "When does label smoothing help?," arXiv preprint arXiv:1906.02629, 2019.

[3] Yuzhe Yang and Zhi Xu, "Rethinking the value of labels for improving class-imbalanced learning," arXiv preprint arXiv:2006.07529, 2020.

[4] Spyros Gidaris, Praveer Singh, and Nikos Komodakis, "Unsupervised representation learning by predicting image rotations," arXiv preprint arXiv:1803.07728, 2018.