ANN Homework 1

田思博 2020011840

实验概括

本次实验内容为实现多层感知机，并识别MNIST手写数字数据集。

多层感知机的实现主要分为三部分：线性层、激活函数层和损失函数；对于上述三部分，都需要实现前向计算和梯度反向传播。

在多层感知机的训练过程中，梯度反向传播至关重要。在本实验中，梯度被组织成形状为(batch_size, neuron_num)的numpy数组，其中第(i,j)个元素代表第i个样本损失对第j个神经元的偏导数。使用这种梯度表示形式可以简便地计算线性层中的梯度反向传播，仅通过一次矩阵乘法运算即可由输出神经元的梯度矩阵得到输入神经元的梯度矩阵。

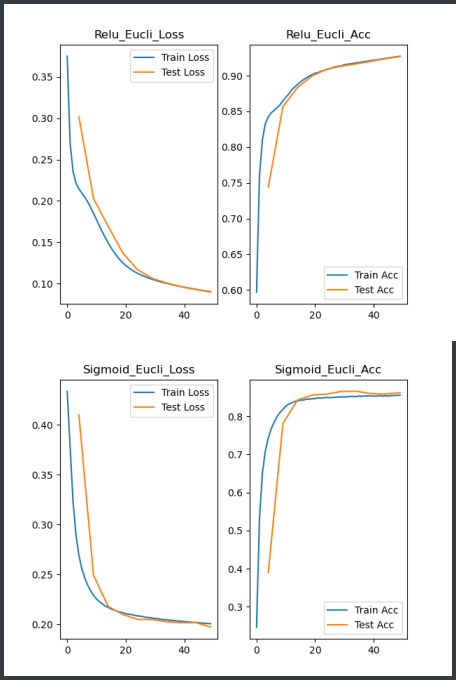
实验结果及分析

1. 在本报告中，为减少损失曲线、准确率曲线的噪声，以epoch为间隔绘制损失曲线、准确率曲线。用（input，output）表示线性层的结构，分别代表输入、输出神经元的数量。
2. 单隐藏层实验：
    1. 比较Relu/Gelu/Sigmoid三种激活函数的性能

        控制变量为：线性层结构为（784，512），（521，10）；损失函数采用EuclideanLoss，learning rate=0.01，weight_decay=0, momentum=0，batch_size=100，max_epoch=50；

        1. 实验结果

隐藏层数	激活函数	损失函数	Train Loss	Test Loss	Train Acc	Test Acc
1	Relu	Euclidean	0.087	0.087	0.933	0.934
1	Gelu	Euclidean	0.082	0.083	0.938	0.936
1	Sigmoid	Euclidean	0.201	0.197	0.855	0.861



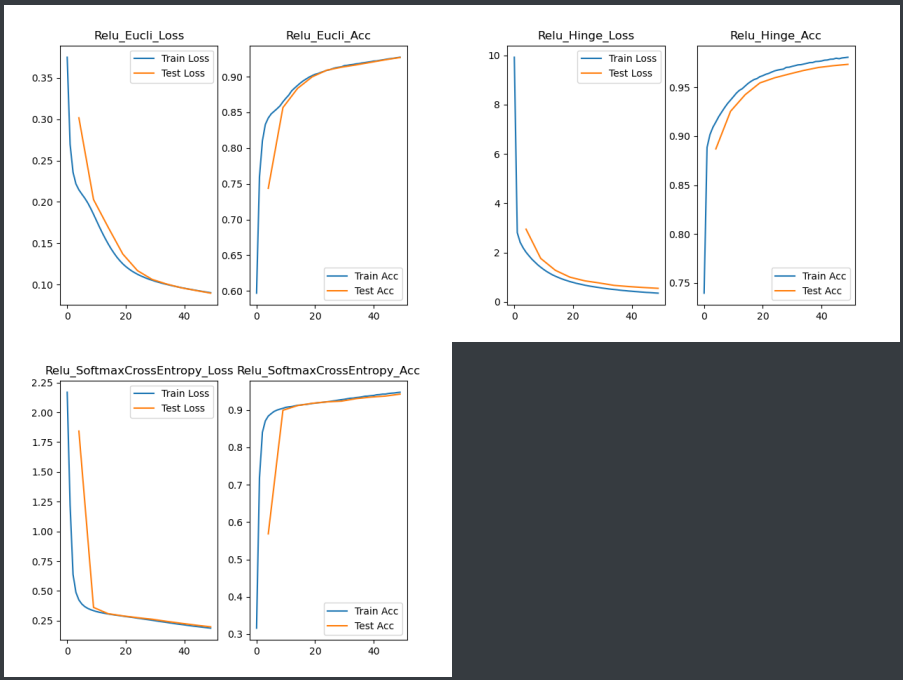
2. 实验分析
    1. 单隐藏层在训练过程中损失曲线平稳降低，未出现波动；
    2. Sigmoid激活函数在此任务中准确率明显低于另外两种激活函数；

2. 比较EuclideanLoss/SoftmaxCrossEntropyLoss/HingeLoss三种损失函数的性能

    控制变量为：线性层结构为（784，512），（512，10）激活函数采用Relu，learning rate=0.01，weight_decay=0, momentum=0，batch_size=100，max_epoch=50；

    1. 实验结果

隐藏层数	激活函数	损失函数	Train Loss	Test Loss	Train Acc	Test Acc
1	Relu	Euclidean	0.087	0.087	0.933	0.934
1	Relu	Hinge	0.354	0.550	0.980	0.973
1	Relu	SoftmaxCrossEntropy	0.187	0.198	0.947	0.942



2. 实验分析

    1. HingeLoss的表现明显优于另外两种损失函数，测试集准确率能达到97.3%；
    2. 基于不同的损失函数所训练的模型，在训练集和测试集上的准确率相仿，表明模型均不存在明显过拟合；

3. 双隐藏层实验

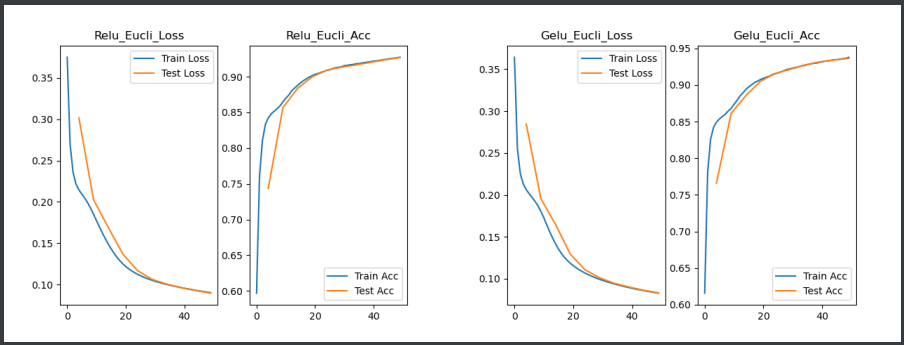
    1. 比较单层、双层网络性能

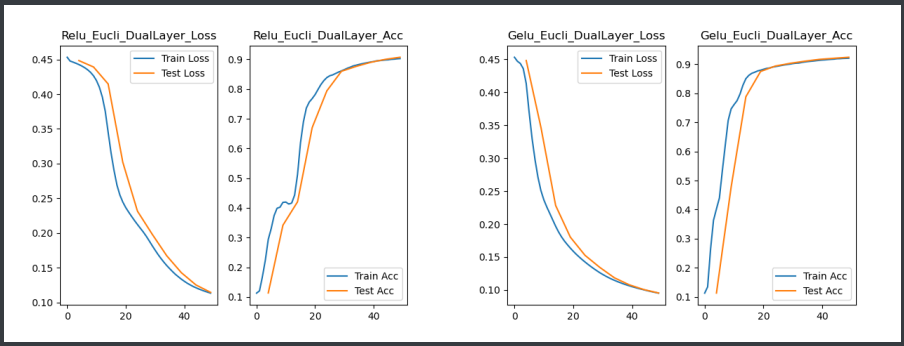
        1. 双层激活函数均为Relu
        2. 双层激活函数均为Gelu

        控制变量为：线性层结构为（784，512），（512，256），（256，10）损失函数采用EuclideanLoss，learning rate=0.01，weight_decay=0, momentum=0，batch_size=100，max_epoch=50；

    2. 实验结果

隐藏层数	激活函数	损失函数	Train Loss	Test Loss	Train Acc	Test Acc	Converge Time（CPU:M1 pro）
1	Relu	Euclidean	0.087	0.087	0.933	0.934	4m30s
1	Gelu	Euclidean	0.082	0.083	0.938	0.936	6m31s
2	Relu	Euclidean	0.113	0.114	0.902	0.907	5m53s
2	Gelu	Euclidean	0.095	0.095	0.920	0.924	8m3s





3. 实验分析

    1. 双层网络在MNIST分类问题上并无显著准确率提升；
    2. 双层网络较单层网络训练时间更长，且更容易出现准确率的抖动，如双层网络搭配Relu激活函数以及Euclidean损失函数的情况，考虑到数据点的间隔为一个epoch，这种抖动是显著的；

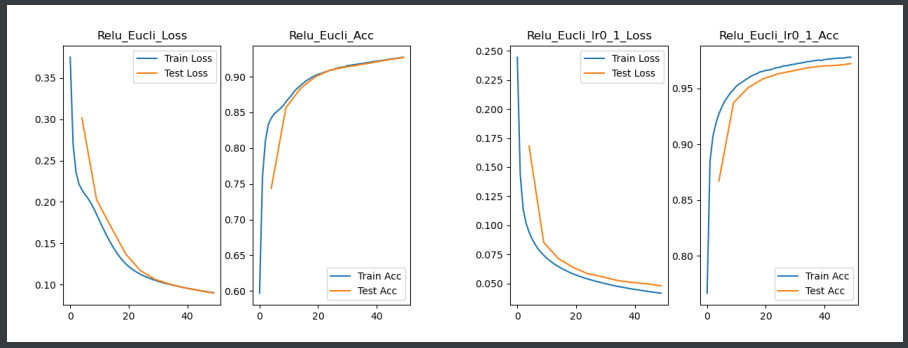
4. 超参数选择

在多层感知机的训练中，超参数的选择对于训练结果有显著影响。在本节中，简要讨论本实验的调参方法。作为对比，设置Baseline为单隐藏层+Relu+EuclideanLoss+{lr=0.01, wd=0, mm=0, batch_size=100, max_epoch=50}，其在测试集上的准确率为93.4%。

1. Learning rate

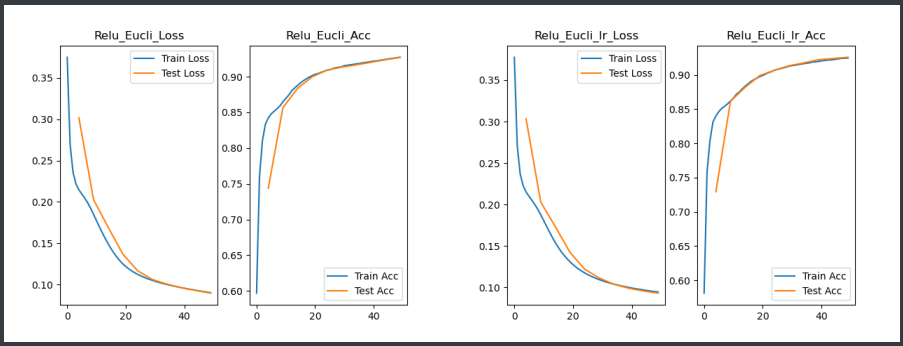
    1. 根据线性层系数与梯度的数量级差异修改学习率

        Linear Layer的权重矩阵取值范围 $10^{-2} ~ 10^{-4}$,梯度矩阵取值范围为 $10^{-5} ~ 10^{-7}$，相差3个数量级，原先选取的0.01（baseline）作为学习率偏小，调整至0.1可以加大步长；调整后改善显著，测试集准确率从93.4%提升至97.3%。



2. 根据训练进度动态调整学习率

   随着训练轮数的增加，当前系数与最优系数的距离减小，因此应当适度减小每次更新的步长来避免越过最优解；指数衰减，是指以固定频率减小learning rate。本实验中，采取了每5个epoch减小将learning rate乘0.95的方式，动态调整learning rate。最终在测试集上的准确率为92.6%，未有明显改进。



（左图为Baseline，右图为动态调整学习率的结果）

2. Weight decay

   权重衰减（weight decay）体现了正则化的思想，weight decay本质上是L2正则化项的系数。由于MNIST数据集较小（训练集仅有60000张图片），因此易出现过拟合现象；理论上引入正则化可以减少过拟合，提高模型在测试集上的准确率。

   但在本实验中，测试了$10^{-5} ~ 10^{-3}$之间的weight decay，并未获得明显的准确率提升；表现最好的10^{-5}weight decay的准确率也仅为92.9%，低于baseline。