Report: MNIST Digit Classification with MLP

2018011365 张鹤潇

实验内容

在本次实验中，我们需要手动实现 MLP 的各个组件，包括线性层、激活函数、损失函数，并在 MNIST 数据集上测试模型的性能。

实验设置

我共实现了两类 MLP 网络，分别有1个和2个隐藏层，详细结构如下：

- MLP with One Hidden Layer (MLP1)

Layer	Type	In Dim	Out Dim
0	Linear	784	256
1	Activation	256	256
2	Linear	256	10
3	Activation	10	10

- MLP with Two Hidden Layer (MLP2)

Layer	Type	In Dim	Out Dim
0	Linear	784	512
1	Activation	512	512
2	Linear	512	256
3	Activation	256	256
4	Linear	256	10
5	Activation	10	10

为了节约训练时间，我在实验中引入了 Early Stopping，即在测试集准确率**连续 10 个 epoch 不上升**时停止训练。

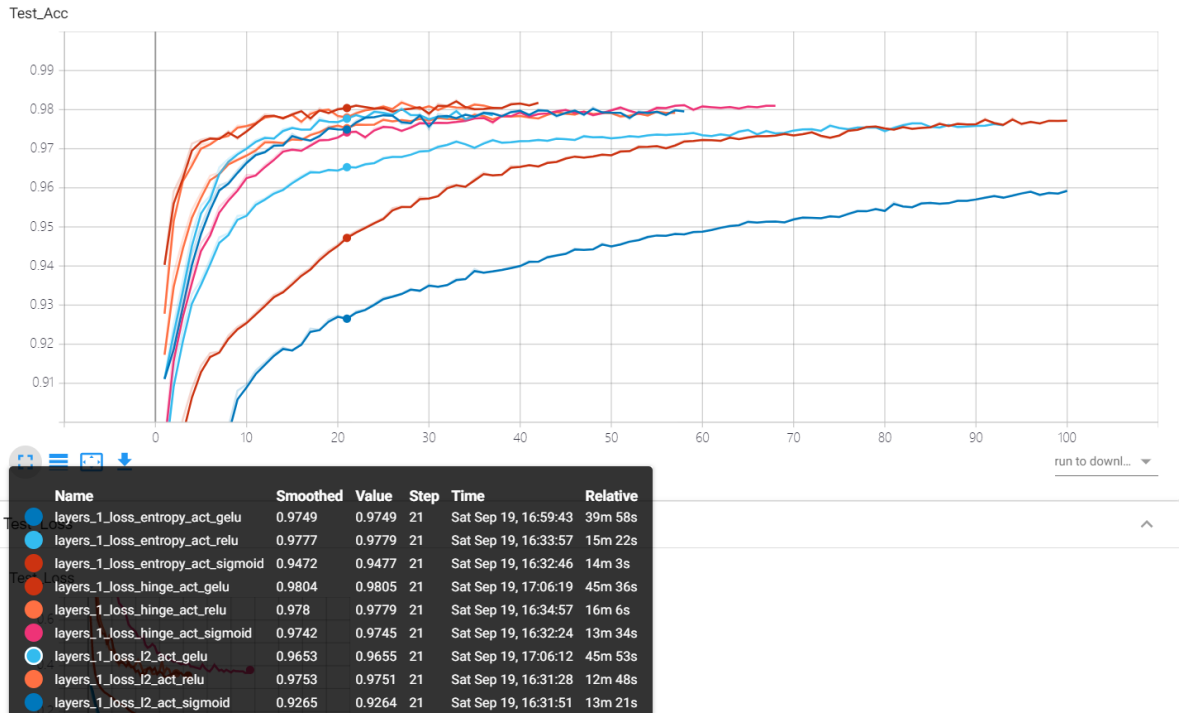
其它超参数设置如下:

```
INIT_STD = 0.01
HINGE_LOSS_THRESHOLD = 5
config = {
    'learning_rate': 0.01,
    'weight_decay': 0.0,
    'momentum': 0.9,
    'batch_size': 100,
    'max_epoch': 100
}
```

实验结果

Layer	Loss	Activation	Test Acc	Test Loss	Train Acc	Train Loss
1	L2	Sigmoid	95.94	0.0155	96.22	0.0156
		Relu	**97.97**	0.013	98.88	0.0115
		Gelu	97.66	0.014	98.43	0.0128
	Cross Entropy	Sigmoid	97.78	0.076	98.79	0.0472
		Relu	98.05	0.0642	99.16	0.0316
		Gelu	**98.06**	0.0667	99.75	0.0145
	Hinge	Sigmoid	98.12	0.3836	99.55	0.0864
		Relu	98.22	0.3587	99.7	0.0591
		Gelu	**98.24**	0.3516	99.86	0.0316
2	L2	Sigmoid	94.69	0.0156	94.67	0.016
		Relu	**98.56**	0.0091	99.61	0.0065
		Gelu	98.26	0.0101	99.03	0.0087
	Cross Entropy	Sigmoid	97.63	0.0781	98.91	0.0403
		Relu	**98.25**	0.0652	99.78	0.0095
		Gelu	98.12	0.0683	99.21	0.0256
	Hinge	Sigmoid	97.91	0.4238	99.37	0.1058
		Relu	98.32	0.3599	99.52	0.0712
		Gelu	**98.31**	0.3664	99.83	0.0268

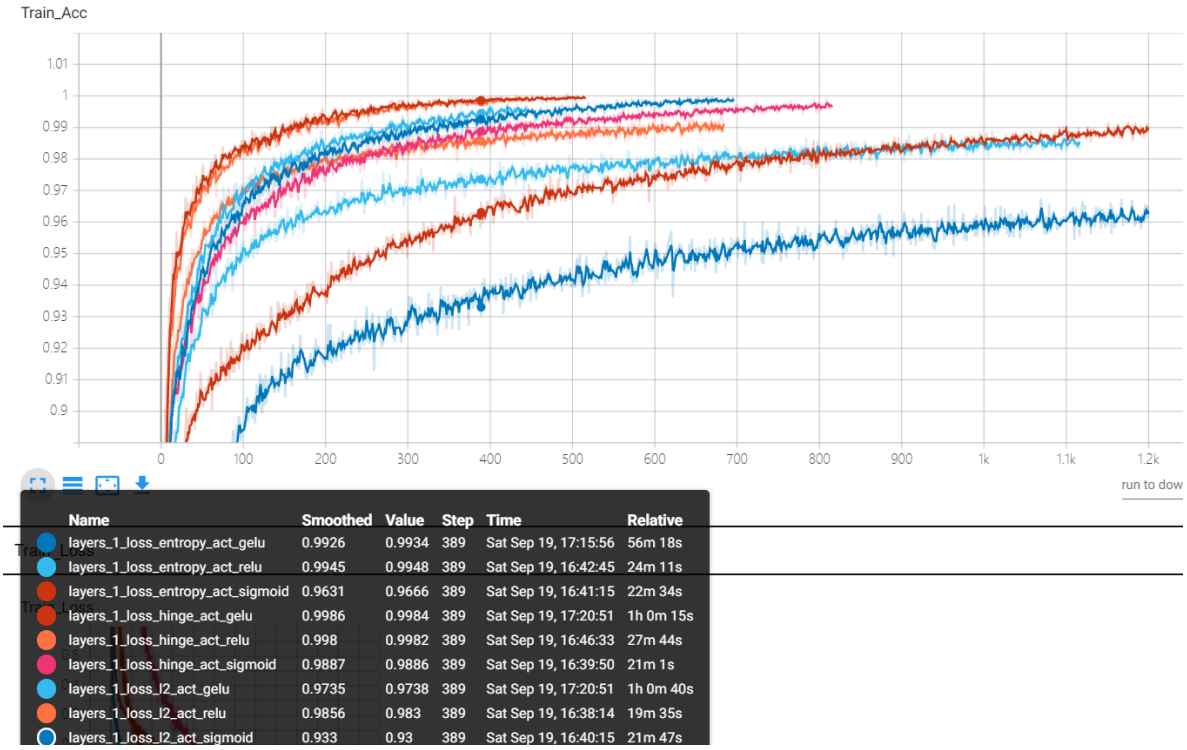
MLP1的Test Acc曲线



MLP1的Test Loss曲线

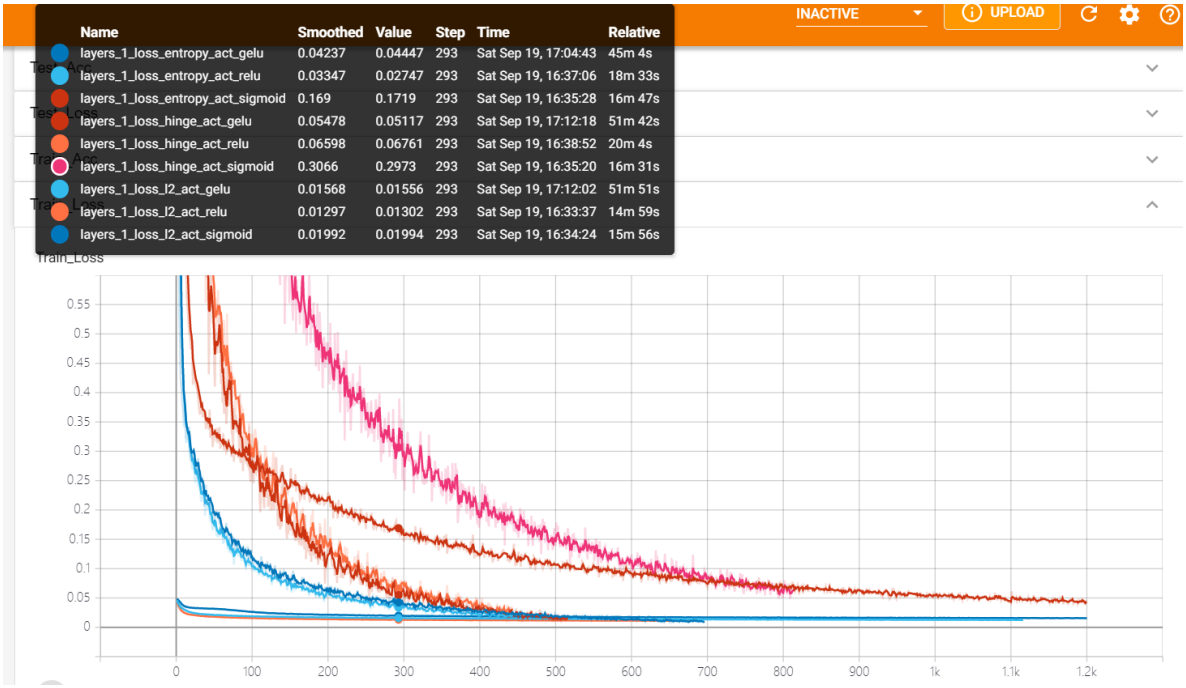
Name	Smoothed	Value	Step	Time	Relative
layers_1_loss_entropy_act_gelu	0.07184	0.07161	32	Sat Sep 19, 17:14:52	55m 7s
layers_1_loss_entropy_act_relu	0.06806	0.06838	32	Sat Sep 19, 16:41:53	23m 18s
layers_1_loss_entropy_act_sigmoid	0.1397	0.1389	32	Sat Sep 19, 16:40:31	21m 48s
layers_1_loss_hinge_act_gelu	0.3516	0.3473	32	Sat Sep 19, 17:20:21	59m 38s
layers_1_loss_hinge_act_relu	0.3686	0.3561	32	Sat Sep 19, 16:45:37	26m 46s
layers_1_loss_hinge_act_sigmoid	0.4414	0.4416	32	Sat Sep 19, 16:39:26	20m 36s
layers_1_loss_l2_act_gelu	0.01522	0.0152	32	Sat Sep 19, 17:20:10	59m 51s
layers_1_loss_l2_act_relu	0.01338	0.01338	32	Sat Sep 19, 16:37:34	18m 54s
layers_1_loss_l2_act_sigmoid	0.01828	0.01824	32	Sat Sep 19, 16:39:34	21m 4s

MLP1的Train Acc曲线

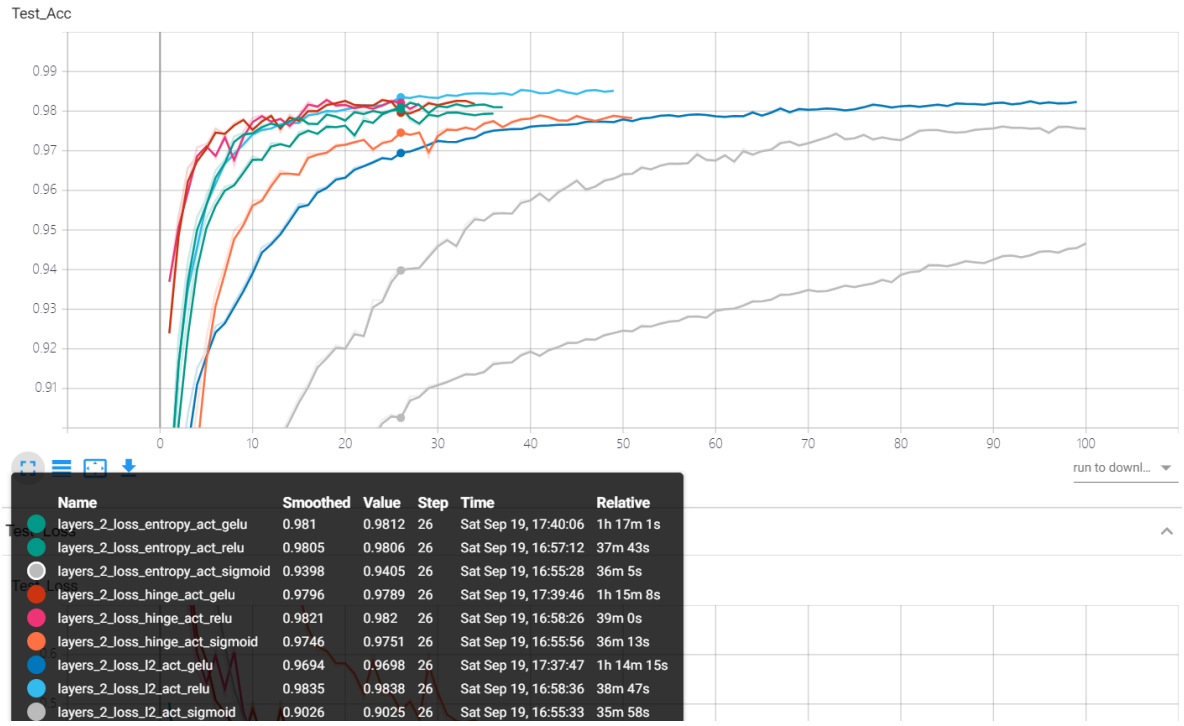


Name	Smoothed	Value	Step	Time	Relative
layers_1_loss_entropy_act_gelu	0.9926	0.9934	389	Sat Sep 19, 17:15:56	56m 18s
layers_1_loss_entropy_act_relu	0.9945	0.9948	389	Sat Sep 19, 16:42:45	24m 11s
layers_1_loss_entropy_act_sigmoid	0.9631	0.9666	389	Sat Sep 19, 16:41:15	22m 34s
layers_1_loss_hinge_act_gelu	0.9986	0.9984	389	Sat Sep 19, 17:20:51	1h 0m 15s
layers_1_loss_hinge_act_relu	0.998	0.9982	389	Sat Sep 19, 16:46:33	27m 44s
layers_1_loss_hinge_act_sigmoid	0.9887	0.9886	389	Sat Sep 19, 16:39:50	21m 1s
layers_1_loss_l2_act_gelu	0.9735	0.9738	389	Sat Sep 19, 17:20:51	1h 0m 40s
layers_1_loss_l2_act_relu	0.9856	0.983	389	Sat Sep 19, 16:38:14	19m 35s
layers_1_loss_l2_act_sigmoid	0.933	0.93	389	Sat Sep 19, 16:40:15	21m 47s

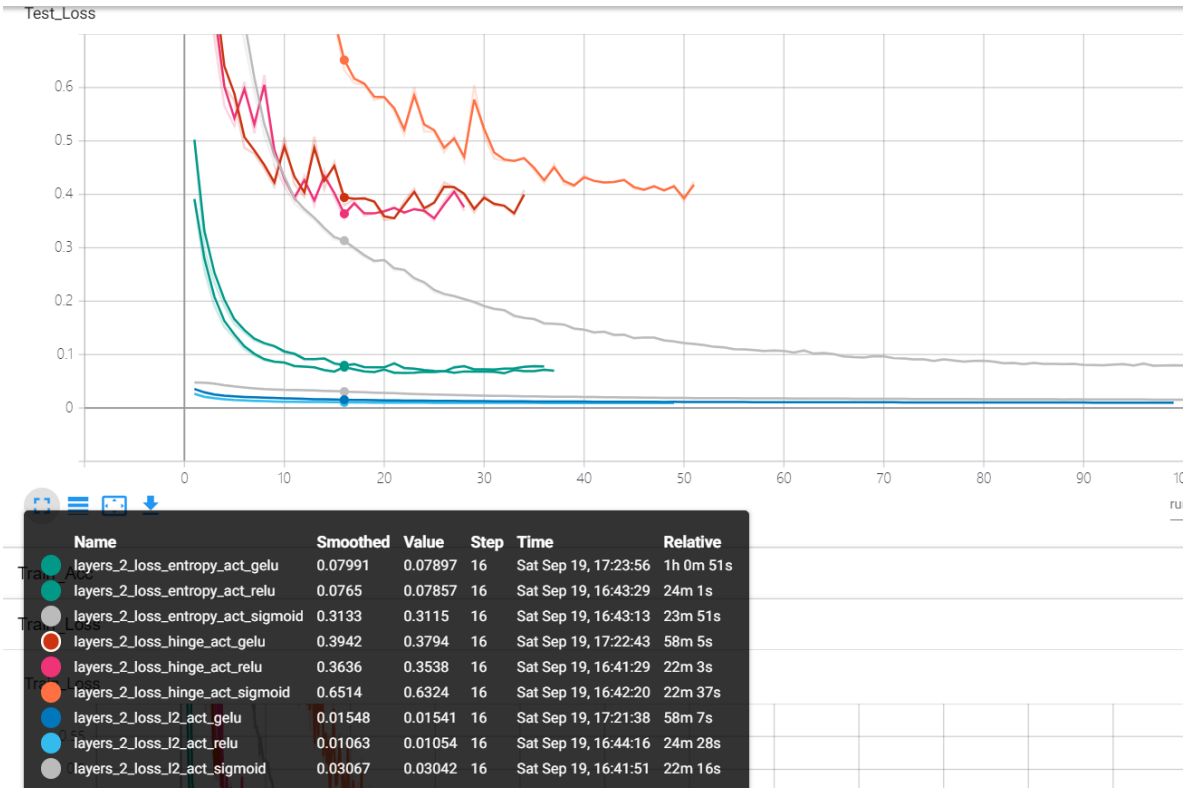
MLP1的Train Loss曲线



MLP2的Test Acc曲线

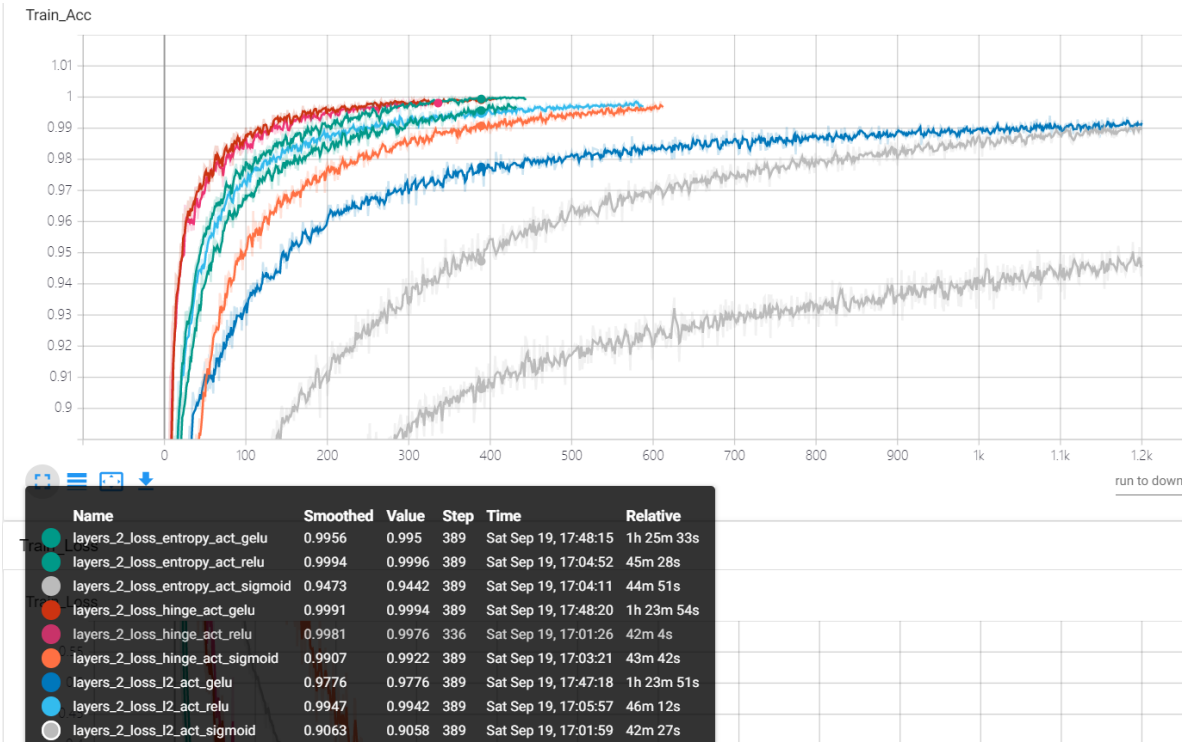


MLP2的Test Loss曲线



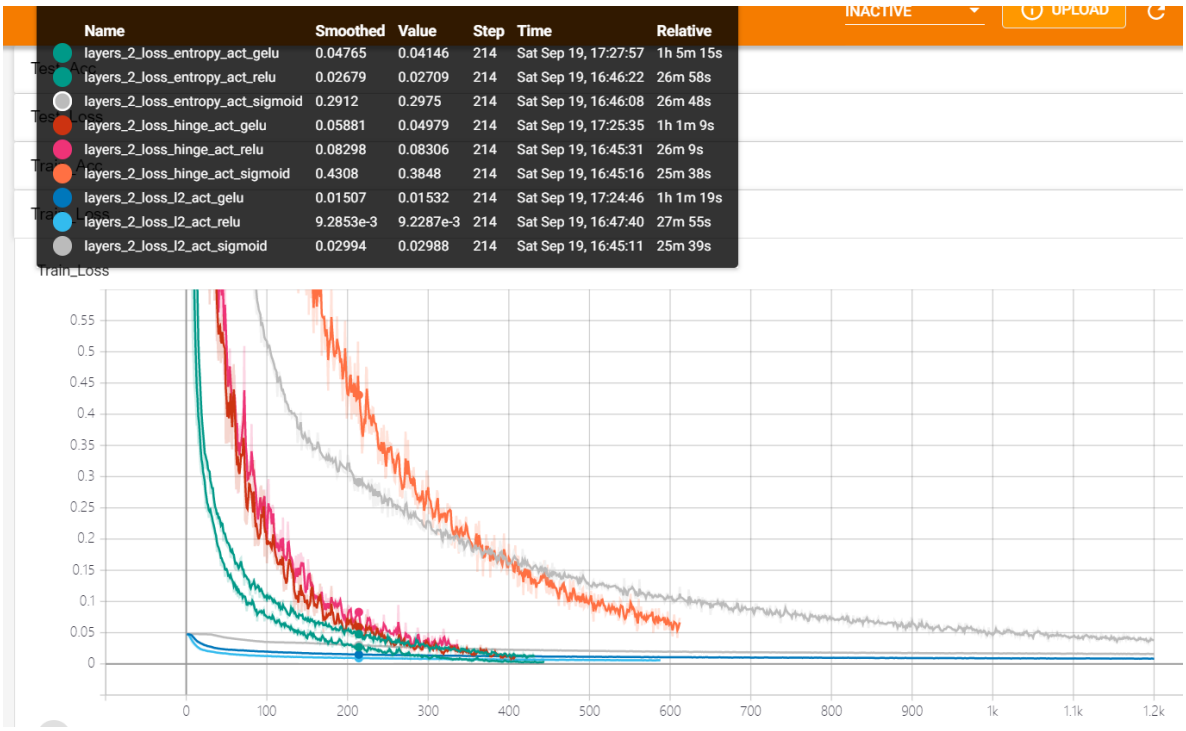
Name	Smoothed	Value	Step	Time	Relative
layers_2_loss_entropy_act_gelu	0.07991	0.07897	16	Sat Sep 19, 17:23:56	1h 0m 51s
layers_2_loss_entropy_act_relu	0.0765	0.07857	16	Sat Sep 19, 16:43:29	24m 1s
layers_2_loss_entropy_act_sigmoid	0.3133	0.3115	16	Sat Sep 19, 16:43:13	23m 51s
layers_2_loss_hinge_act_gelu	0.3942	0.3794	16	Sat Sep 19, 17:22:43	58m 5s
layers_2_loss_hinge_act_relu	0.3636	0.3538	16	Sat Sep 19, 16:41:29	22m 3s
layers_2_loss_hinge_act_sigmoid	0.6514	0.6324	16	Sat Sep 19, 16:42:20	22m 37s
layers_2_loss_l2_act_gelu	0.01548	0.01541	16	Sat Sep 19, 17:21:38	58m 7s
layers_2_loss_l2_act_relu	0.01063	0.01054	16	Sat Sep 19, 16:44:16	24m 28s
layers_2_loss_l2_act_sigmoid	0.03067	0.03042	16	Sat Sep 19, 16:41:51	22m 16s

MLP2的Train Acc曲线



Name	Smoothed	Value	Step	Time	Relative
layers_2_loss_entropy_act_gelu	0.9956	0.995	389	Sat Sep 19, 17:48:15	1h 25m 33s
layers_2_loss_entropy_act_relu	0.9994	0.9996	389	Sat Sep 19, 17:04:52	45m 28s
layers_2_loss_entropy_act_sigmoid	0.9473	0.9442	389	Sat Sep 19, 17:04:11	44m 51s
layers_2_loss_hinge_act_gelu	0.9991	0.9994	389	Sat Sep 19, 17:48:20	1h 23m 54s
layers_2_loss_hinge_act_relu	0.9981	0.9976	336	Sat Sep 19, 17:01:26	42m 4s
layers_2_loss_hinge_act_sigmoid	0.9907	0.9922	389	Sat Sep 19, 17:03:21	43m 42s
layers_2_loss_l2_act_gelu	0.9776	0.9776	389	Sat Sep 19, 17:47:18	1h 23m 51s
layers_2_loss_l2_act_relu	0.9947	0.9942	389	Sat Sep 19, 17:05:57	46m 12s
layers_2_loss_l2_act_sigmoid	0.9063	0.9058	389	Sat Sep 19, 17:01:59	42m 27s

MLP2的Train Loss曲线



结论

网络层数的影响

除激活函数为 Sigmoid 的情况外，MLP2 的准确率略优于 MLP1，但是训练时间也明显变长。

激活函数的影响

使用 Gelu 和 Relu 的模型准确率优于使用 Sigmoid 的模型。Relu 的收敛速度快于 Sigmoid，与 Gelu 相当；但 Gelu 计算复杂，训练时间明显变长。

损失函数的影响

交叉熵和 Hinge Loss 的性能相当，略优于 L2 Loss.

L2 Loss 在与 Sigmoid 搭配时性能明显偏差。

总结反思

Sigmoid 与梯度消失

从实验结果可知，以 Sigmoid 为激活函数的单隐层模型比双隐层模型的性能更好，这种反常现象是由**梯度消失**导致的。

$$\sigma(x) = \frac{1}{1+e^{-x}} \in (0,1)$$

$$\sigma'(x) = \sigma(x)(1-\sigma(x)) \leq \frac{1}{4}$$

分析可知，Sigmoid 函数的导数$\sigma'(x) \in (0, \frac{1}{4}]$，$\sigma'(x)^n \leq 4^{-n}$。随着网络层数的增加，在 Sigmoid 函数间回传的梯度会迅速减小，导致模型得不到有效训练。

Sigmoid 与 L2 Loss

对于 L2 Loss，

$$C = \frac{1}{2}||y - y_t||_2^2$$
$$y = \sigma(z),\ z = x^T W + b$$
$$\frac{\partial C}{\partial W} = (y - y^T)\sigma'(z)x$$
$$\frac{\partial C}{\partial b} = (y - y^T)\sigma'(z)$$

权重 W 和偏置 b 的梯度与激活函数的导数成正比。激活函数的梯度越大，w和b的大小调整得越快，训练收敛得就越快。而对于 Sigmoid 函数来说，其导数在输入为0时取最大，输入较大导数反而变小。

当激活值很大，网络需要快速调整时，Sigmoid 的梯度较小，权重更新的步幅较小，这会导致网络收敛变慢；而当激活值很小时，sigmoid的梯度会比较大，权重更新的步幅也会较大，这会导致网络的震荡，使致模型难以收敛。

这就是 Sigmoid 与 L2 Loss 搭配时模型性能不佳的原因。

输出层是否需要激活函数

我个人认为输出层加或不加激活函数都是可以的。但要注意，如果在使用 Hinge Loss 的模型输出层加激活函数，需要调整 Hinge Loss 的 Δ 参数，否则模型可能不收敛。