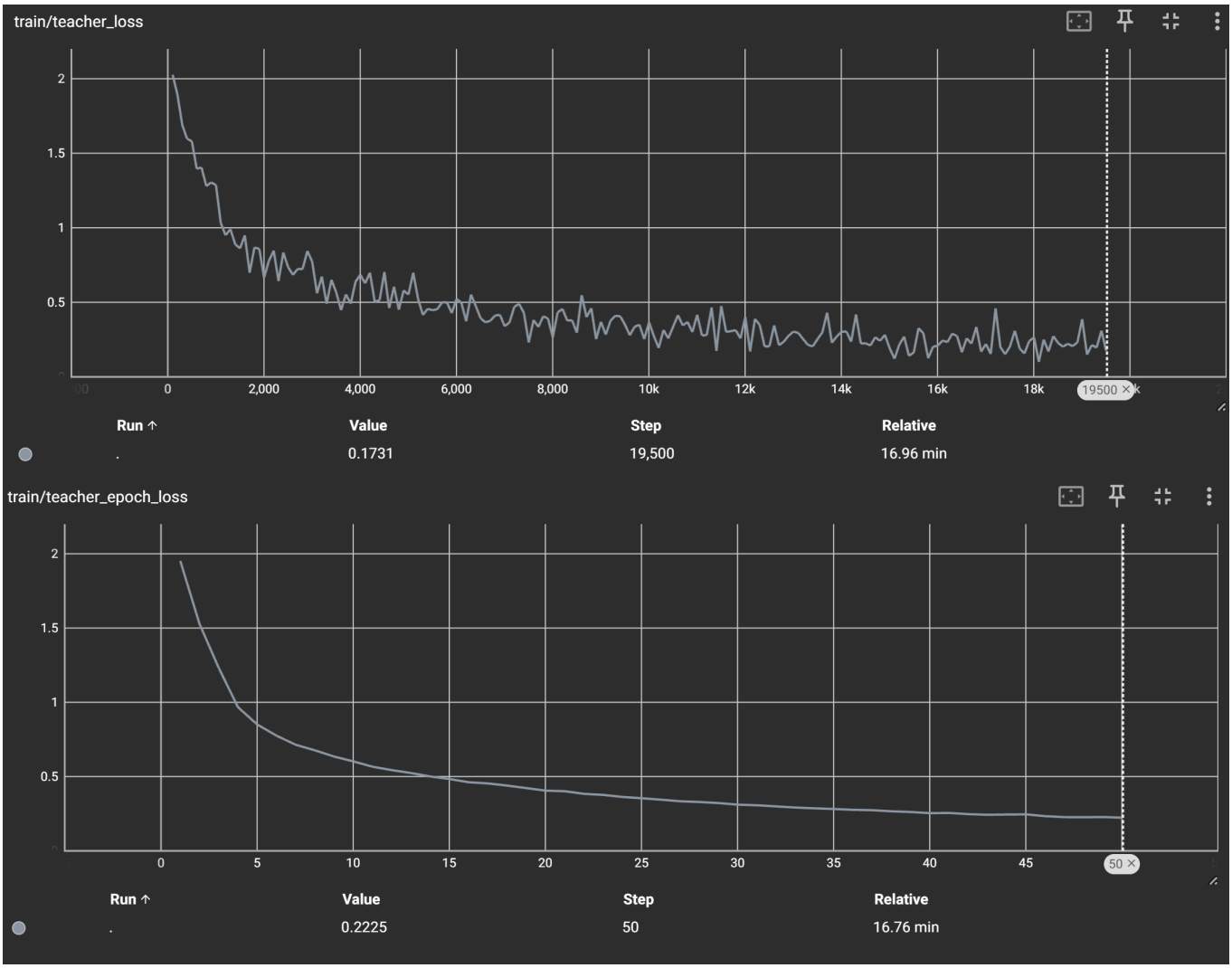
试验

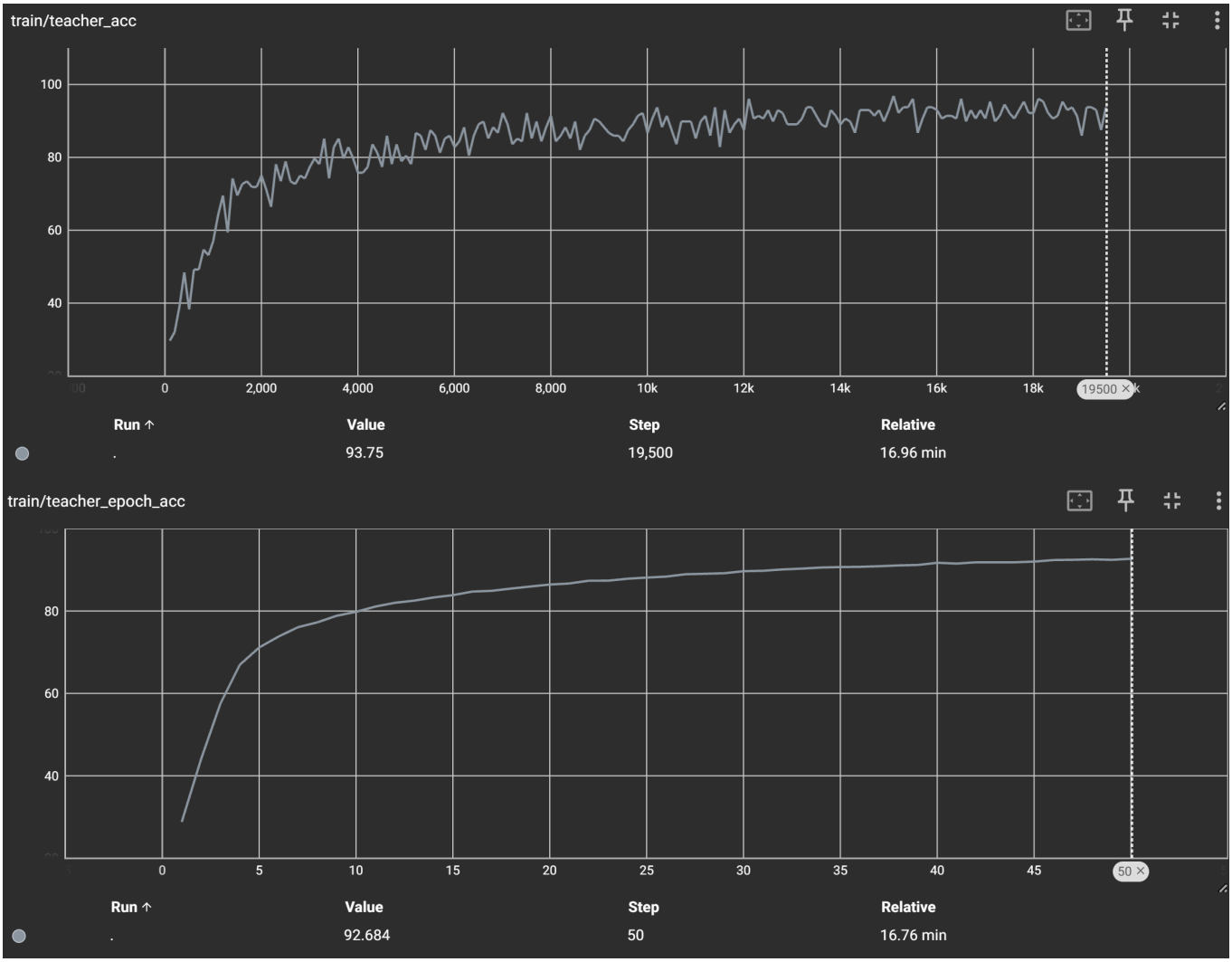
验证在引导的时候，多加的一个矩阵是否会影响到模型的性能，当验证集的准确率相差超过0.8%时，认为有差距。

两个试验，需要保证的前提是，

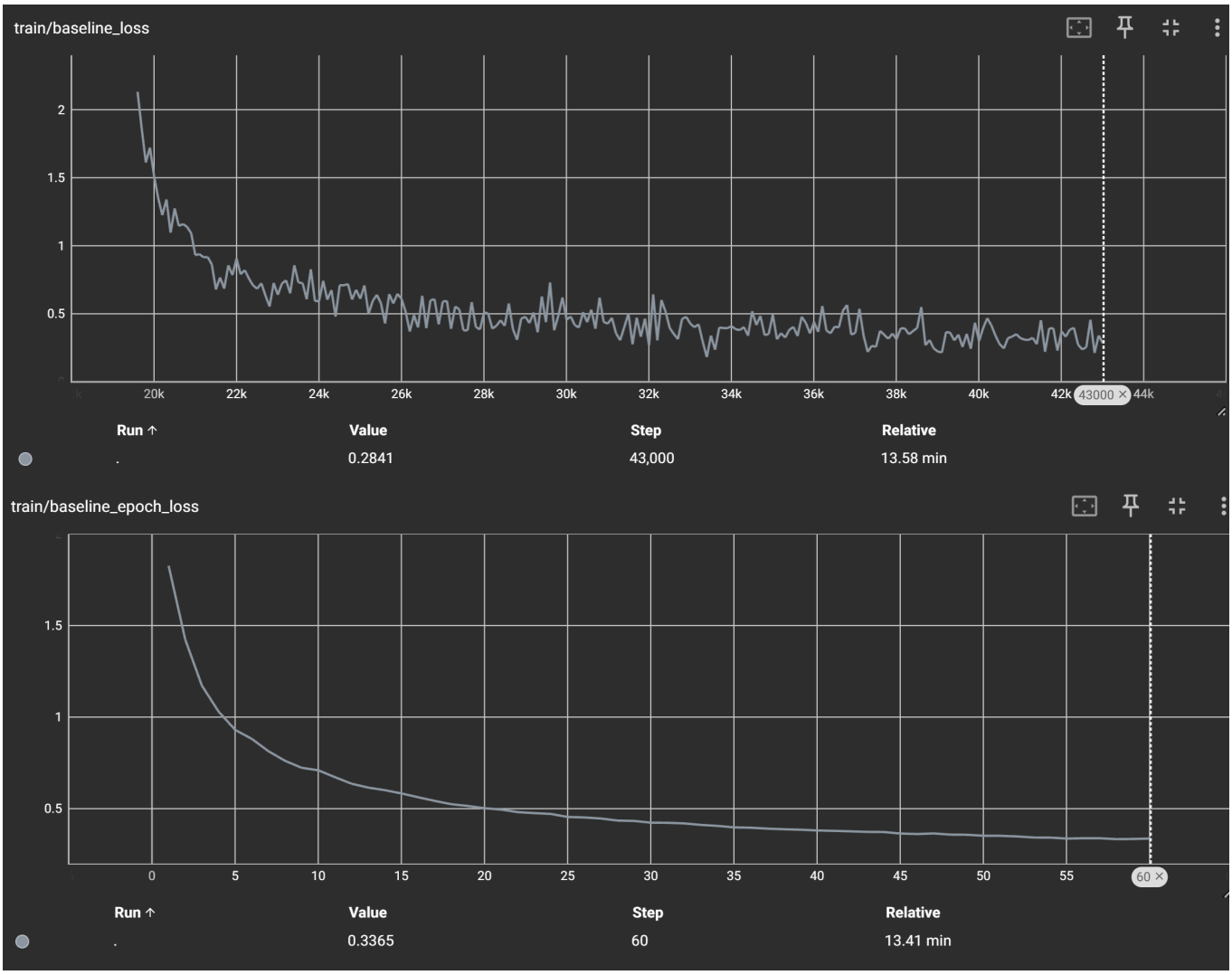
1. 使用同样的教师模型，使用同样的层引导
2. 必要时改变学生模型，在相同的位置进行特征引导
3. 两种基线模型的表现要一致

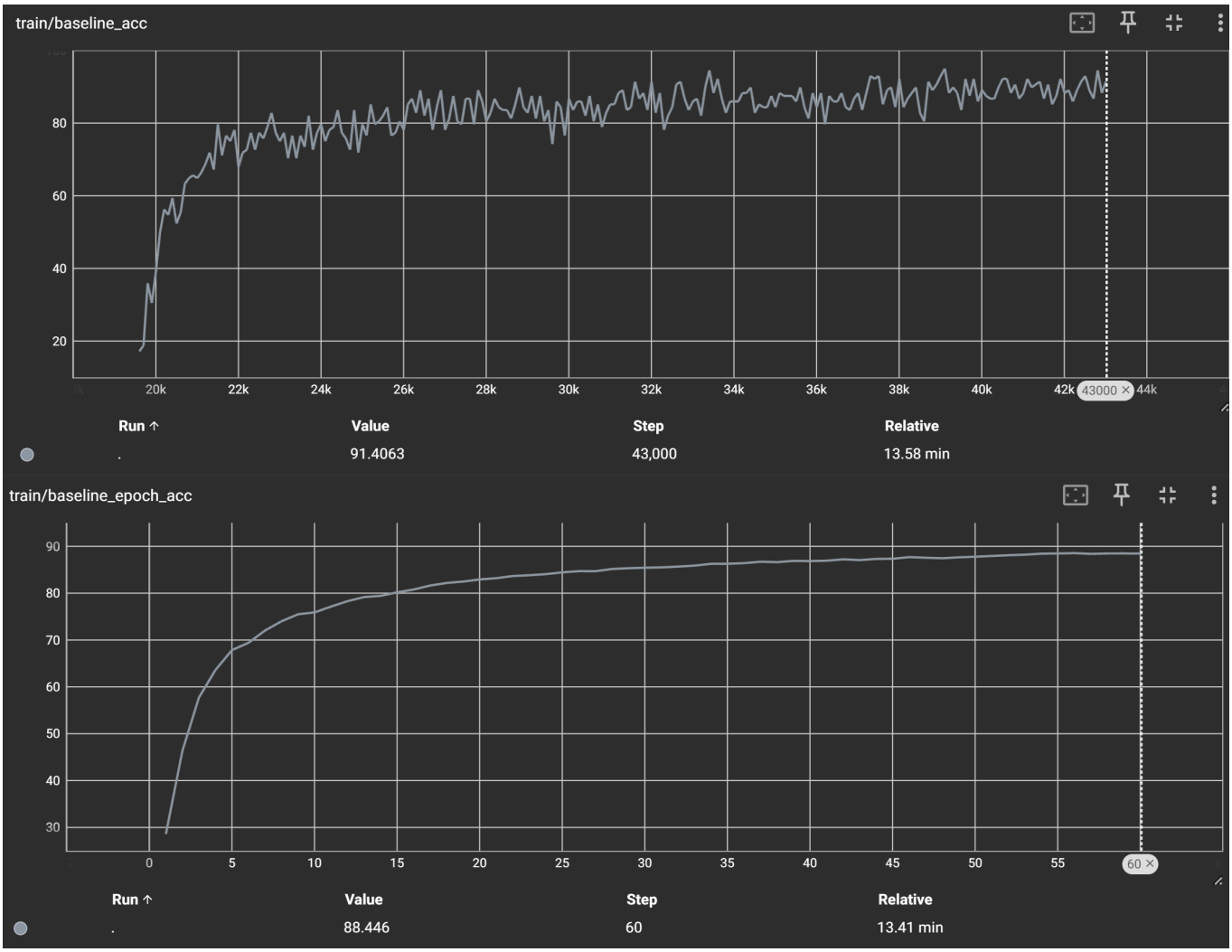
教师网络训练结果




train/teacher_acc
100
80
60
40
0
2,000
4,000
6,000
8,000
10k
12k
14k
16k
18k
19500
Run
Value
Step
Relative
.
93.75
19,500
16.96 min
train/teacher_epoch_acc
80
60
40
0
5
10
15
20
25
30
35
40
45
50
Run
Value
Step
Relative
.
92.684
50
16.76 min

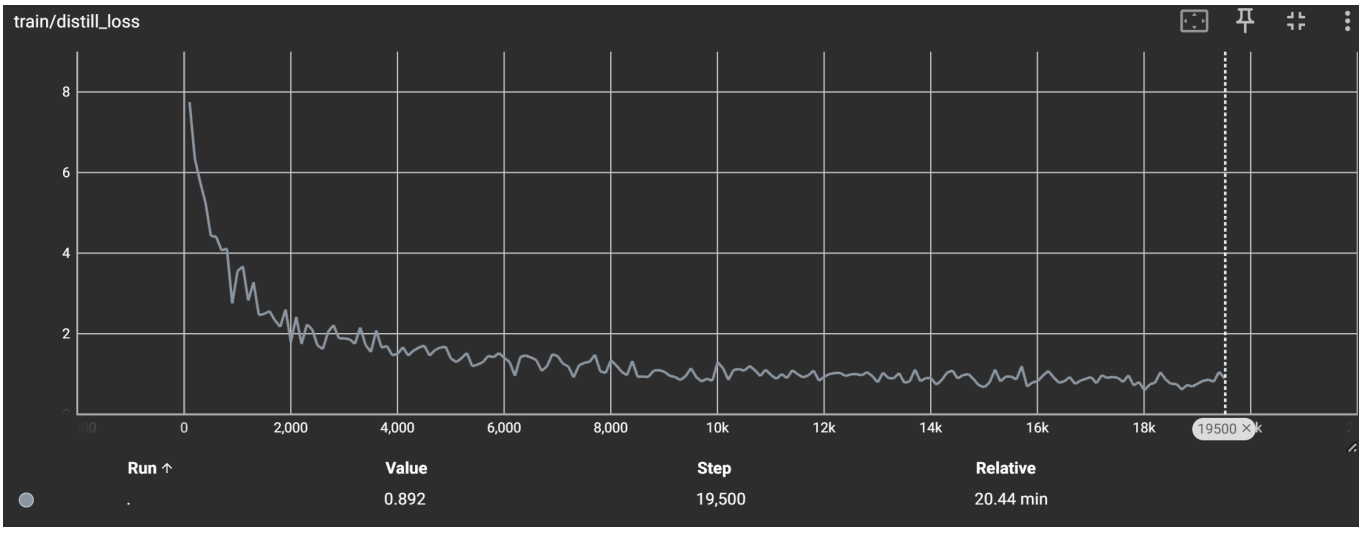
基线网络训练结果

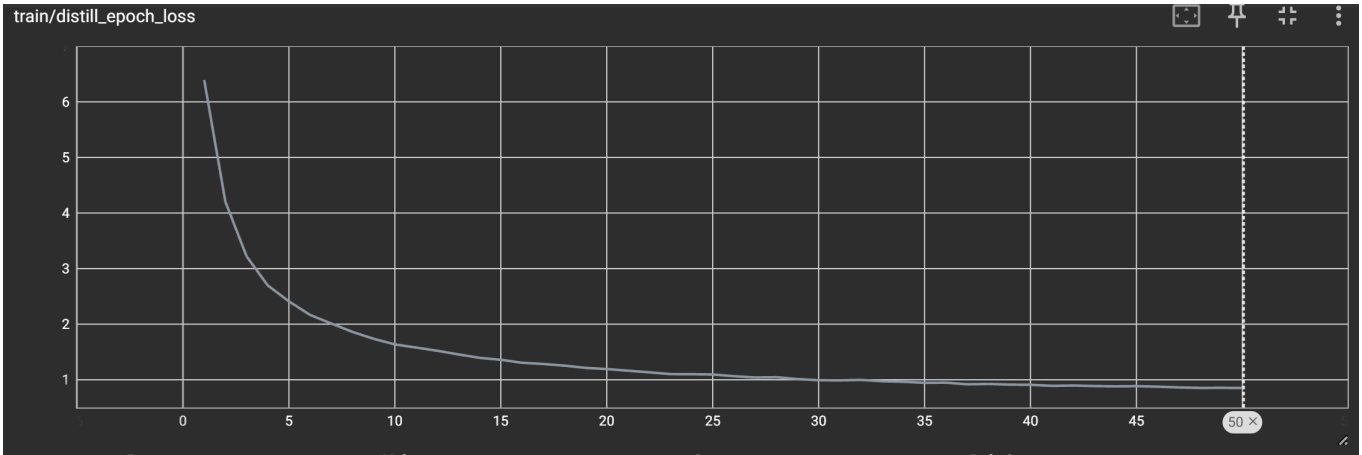

train/baseline_loss
2
1.5
1
0.5
20k
22k
24k
26k
28k
30k
32k
34k
36k
38k
40k
42k
43000
44k
Run
Value
Step
Relative
.
0.2841
43,000
13.58 min
train/baseline_epoch_loss
1.5
1
0.5
0
5
10
15
20
25
30
35
40
45
50
55
60
Run
Value
Step
Relative
.
0.3365
60
13.41 min



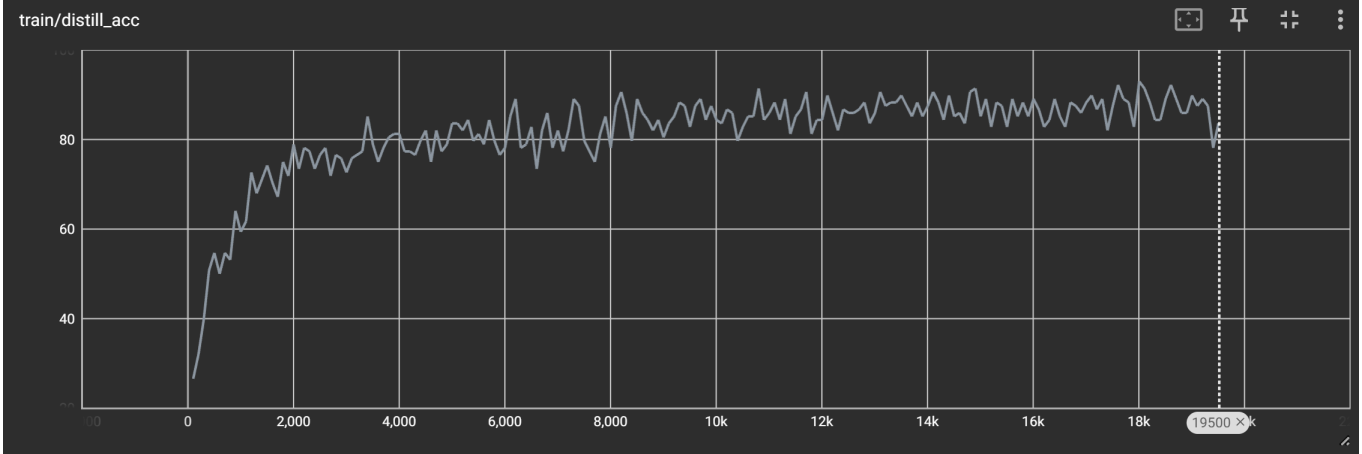
学生网络蒸馏效果

只进行蒸馏，不进行特征引导

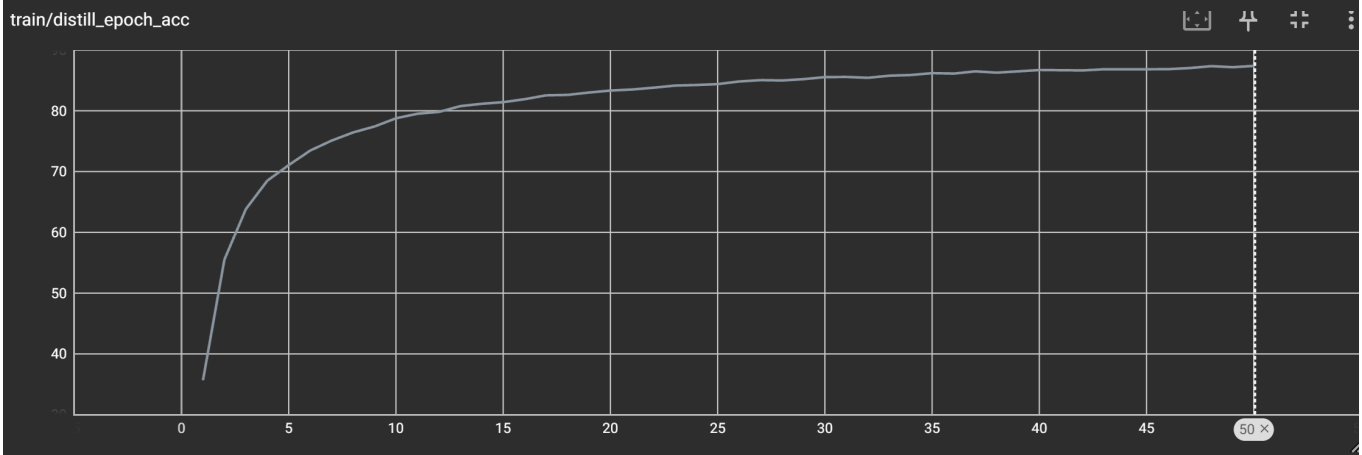




Run	Value	Step	Relative
.	0.8514	50	20.2 min

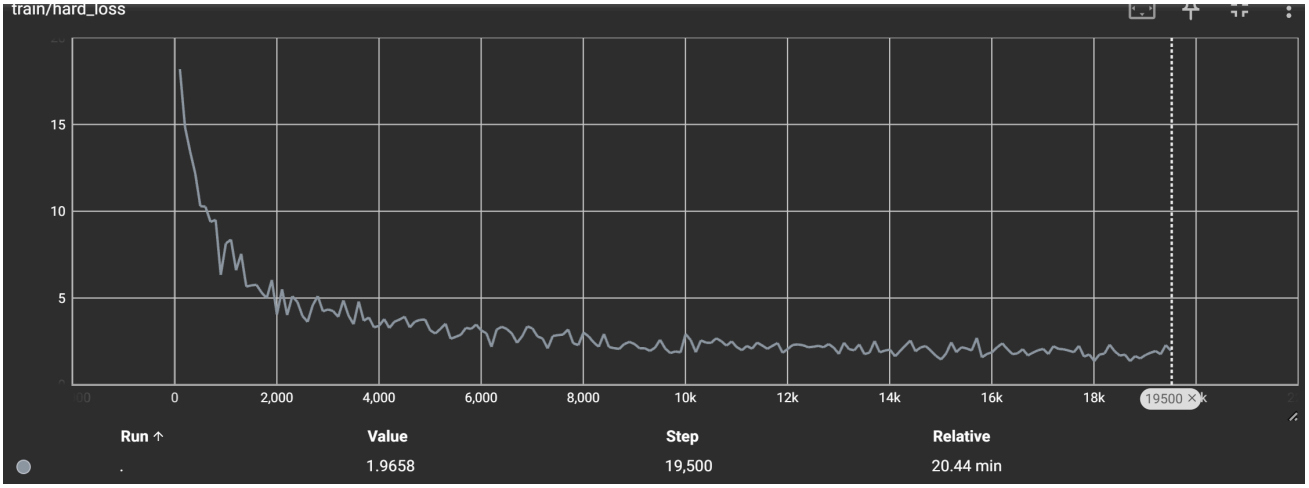


Run	Value	Step	Relative
.	83.5938	19,500	20.44 min



Run	Value	Step	Relative
.	87.386	50	20.2 min

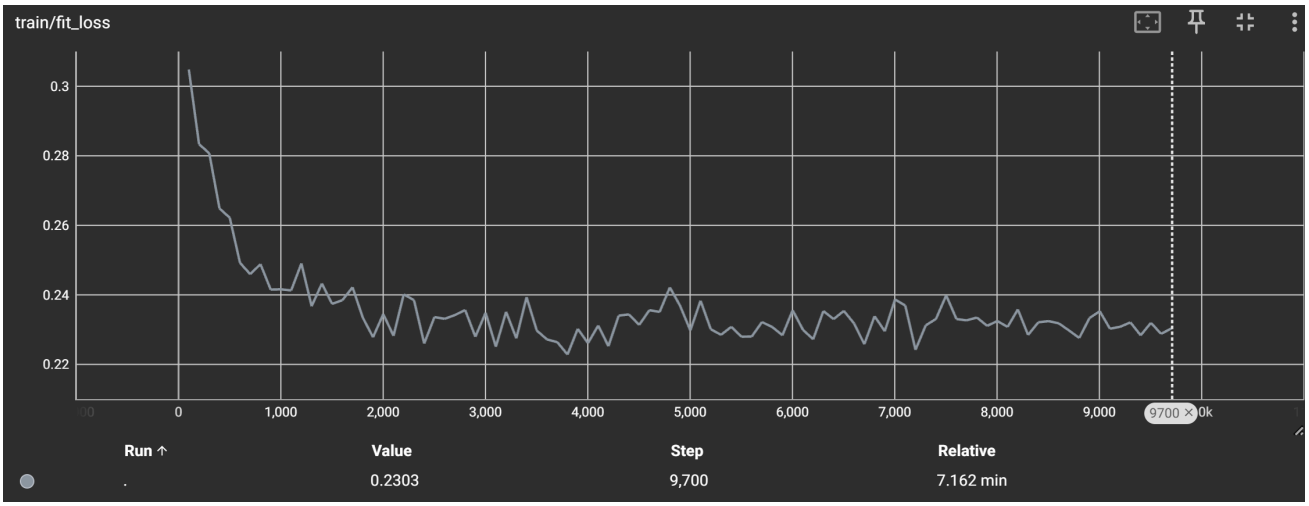
- 蒸馏分类损失

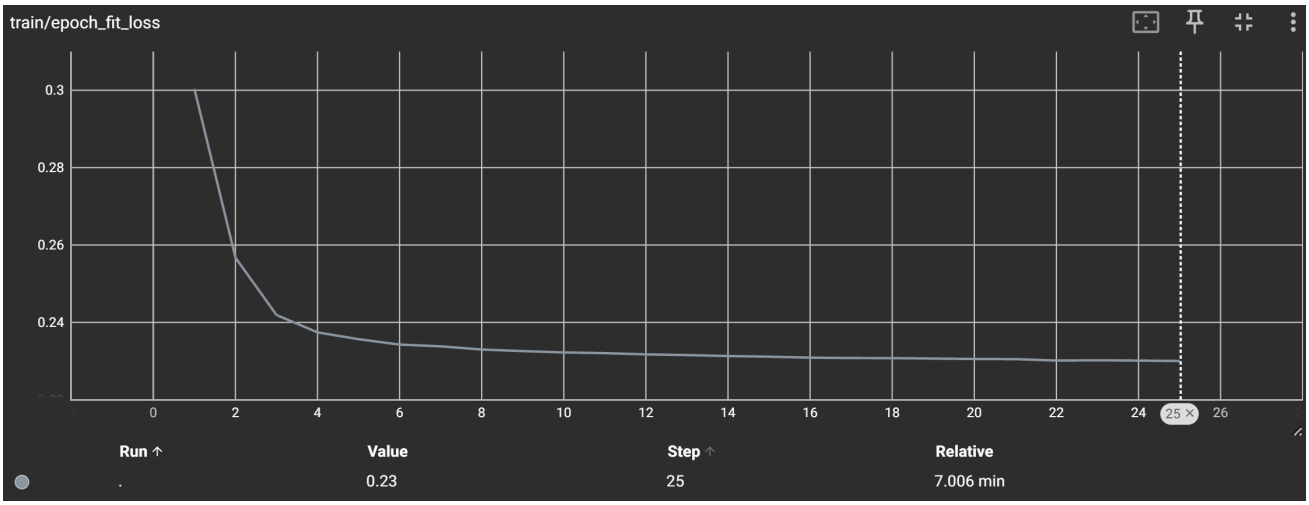


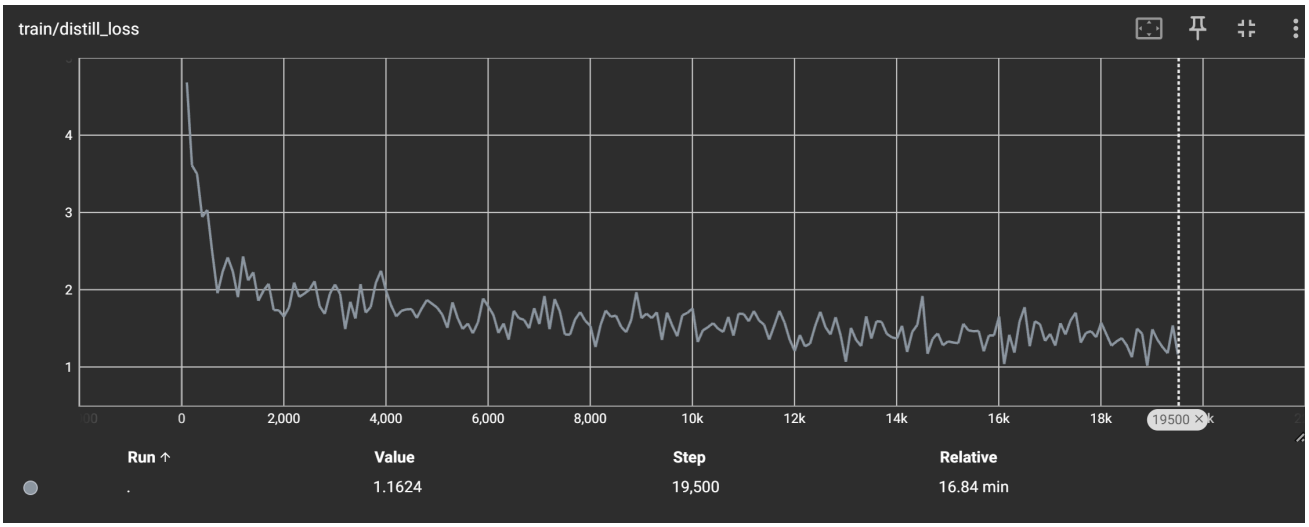
试验一

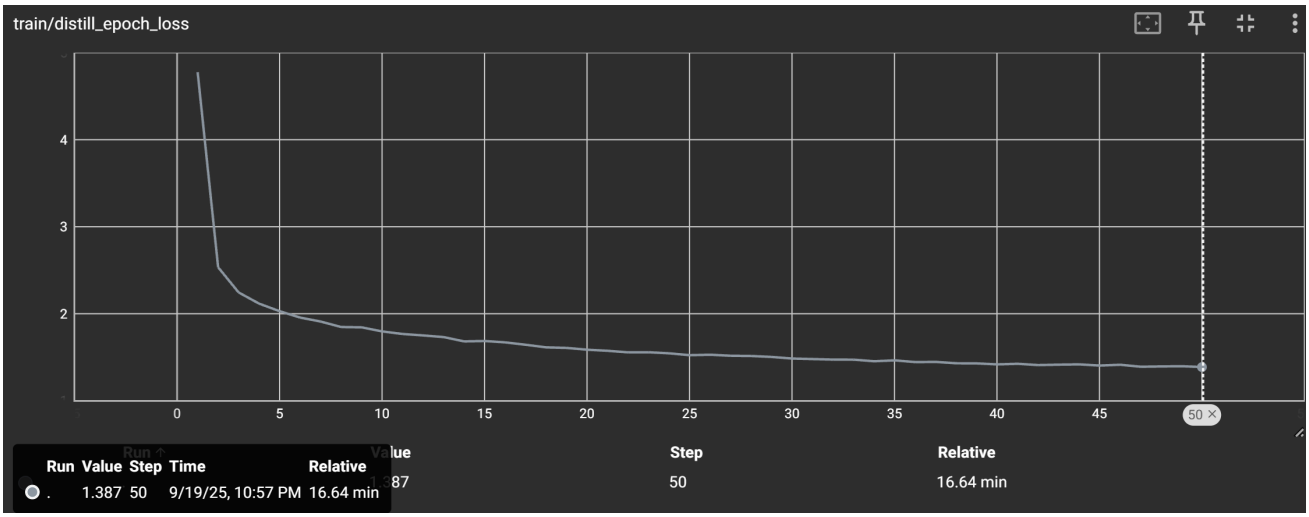
前后分布训练，

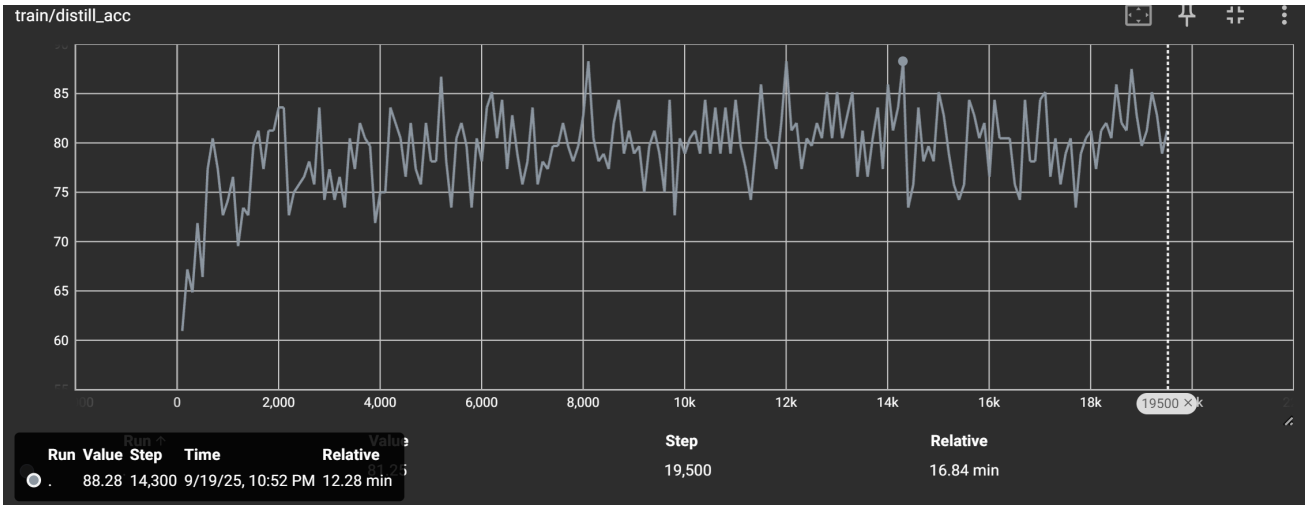
- stage one： 引导层前面的网络进行fitnet训练
- stage two : 固定引导层以及前面的网络，进行后续网络的单独训练

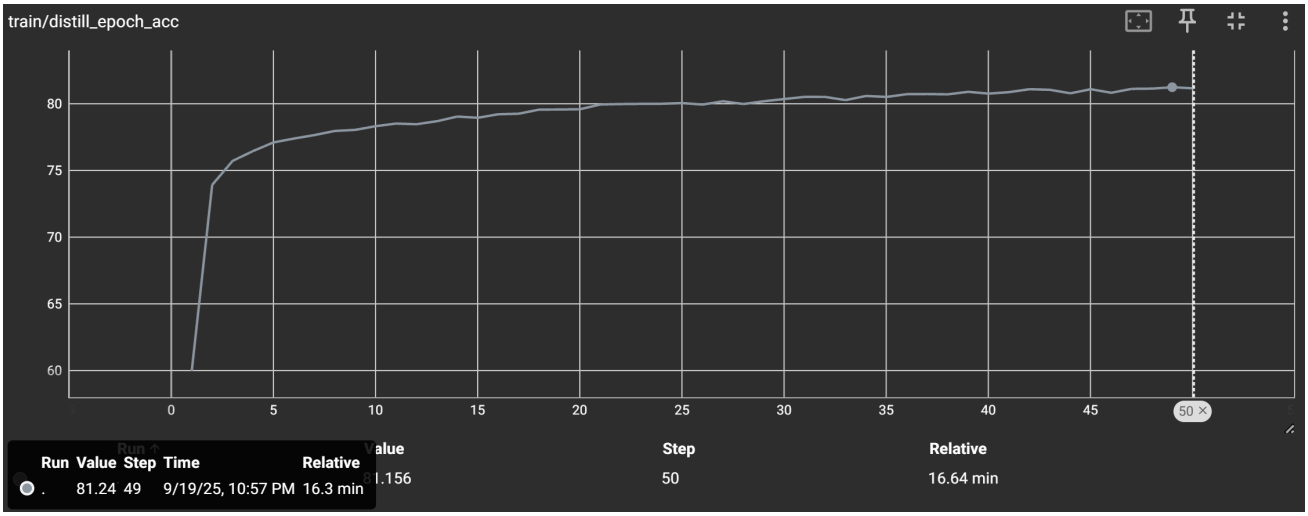




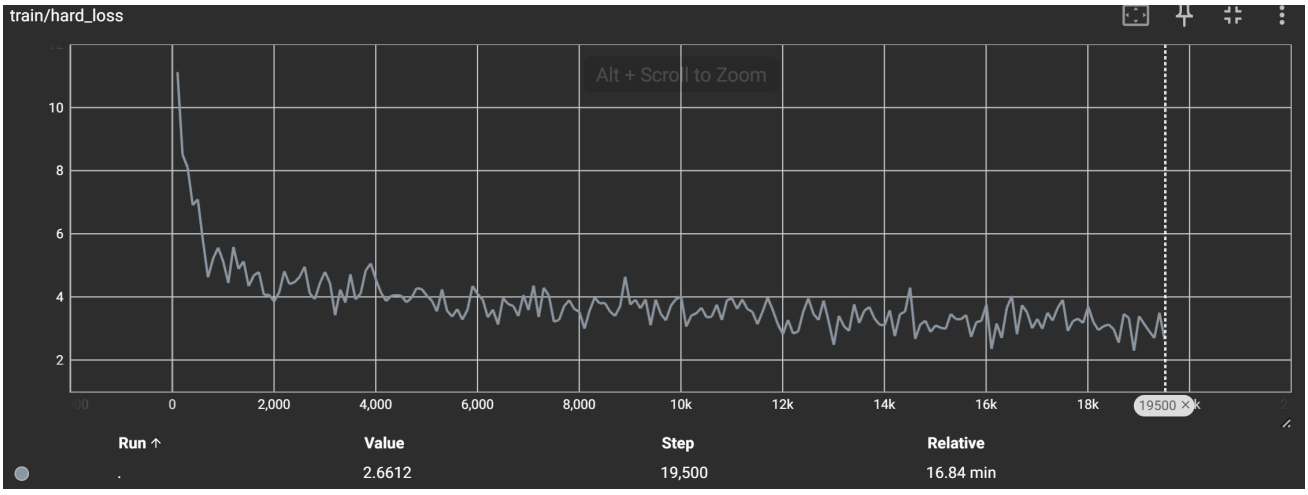

train/distill_loss
4
3
2
1
0
2,000
4,000
6,000
8,000
10k
12k
14k
16k
18k
19500
Run
Value
Step
Relative
.
1.1624
19,500
16.84 min

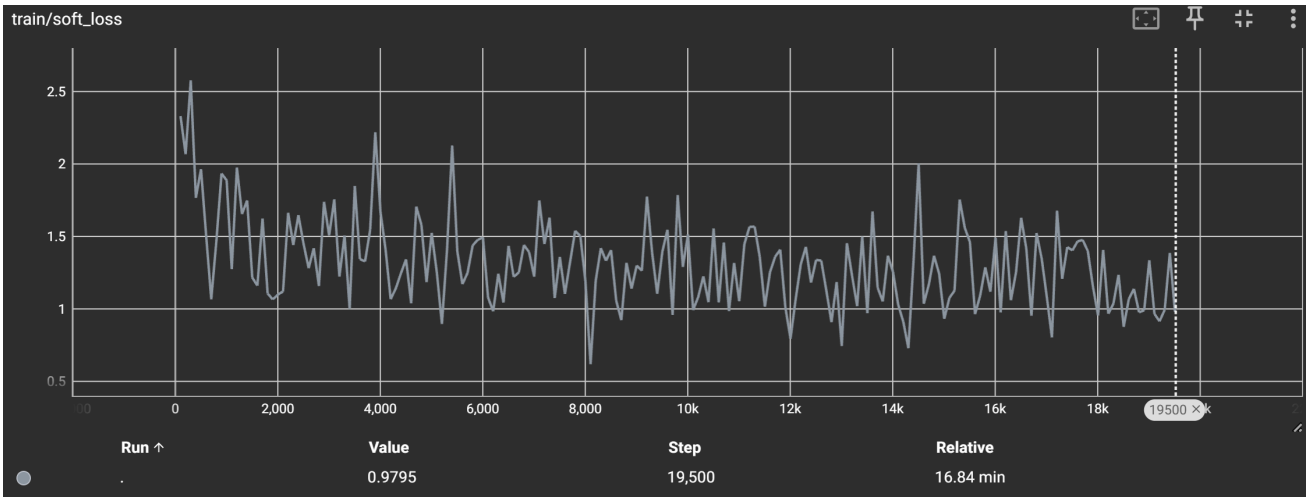

train/distill_epoch_loss
4
3
2
0
5
10
15
20
25
30
35
40
45
50
Run
Value
Step
Relative
. 1.387 50 9/19/25, 10:57 PM 16.64 min
87
50
16.64 min


train/distill_acc
85
80
75
70
65
60
0
2,000
4,000
6,000
8,000
10k
12k
14k
16k
18k
19500
Run
Value
Step
Time
Relative
. 88.28 14,300 9/19/25, 10:52 PM 12.28 min
Step
19,500
Relative
16.84 min


train/distill_epoch_acc
80
75
70
65
60
0
5
10
15
20
25
30
35
40
45
50
Run
Value
Step
Time
Relative
. 81.24 49 9/19/25, 10:57 PM 16.3 min
Step
50
Relative
16.64 min

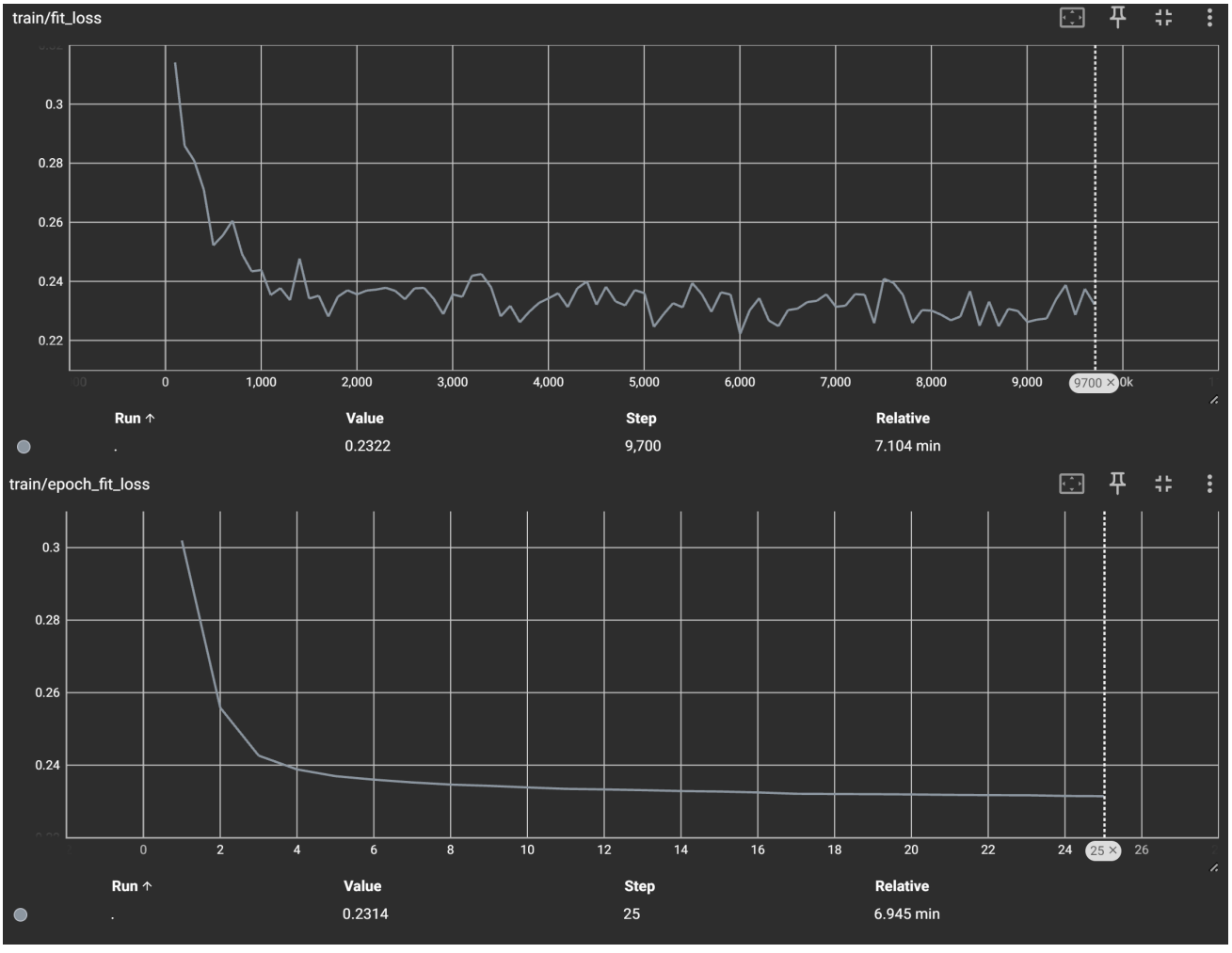
- 蒸馏各个损失记录

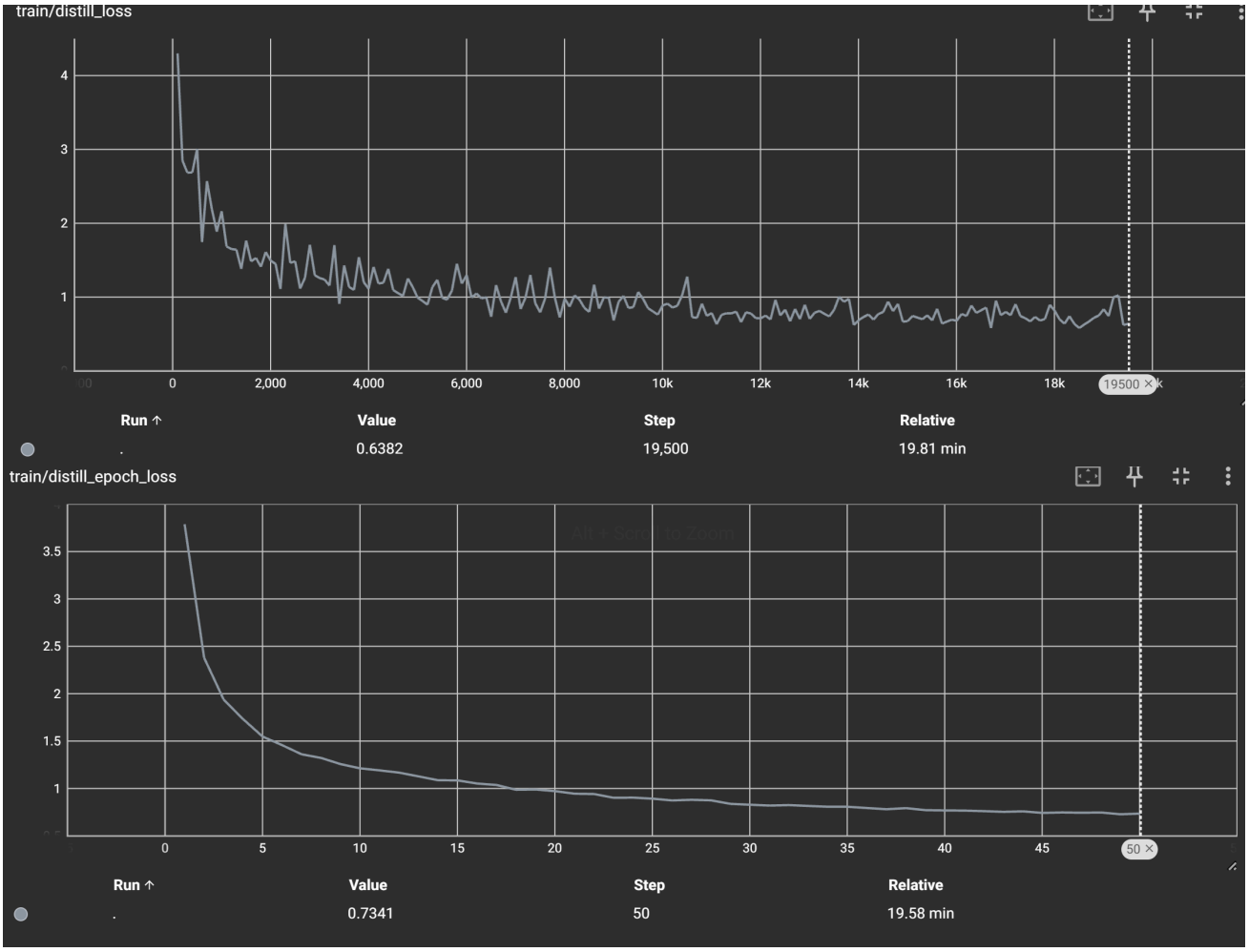

train/hard_loss
Alt + Scroll to Zoom
10
8
6
4
2
0
2,000
4,000
6,000
8,000
10k
12k
14k
16k
18k
19500
Run
Value
Step
Relative
. 2.6612 19,500 16.84 min

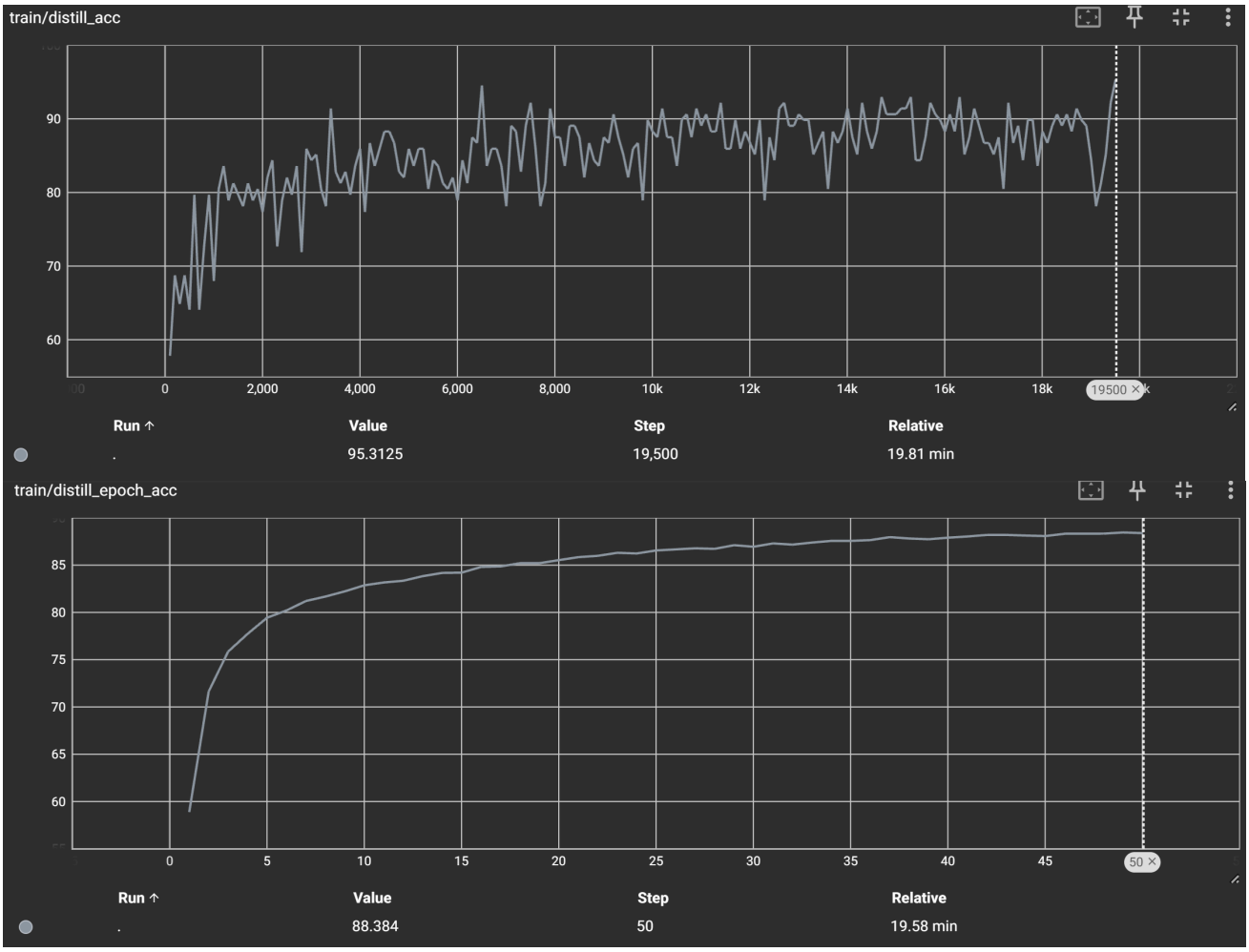


试验二

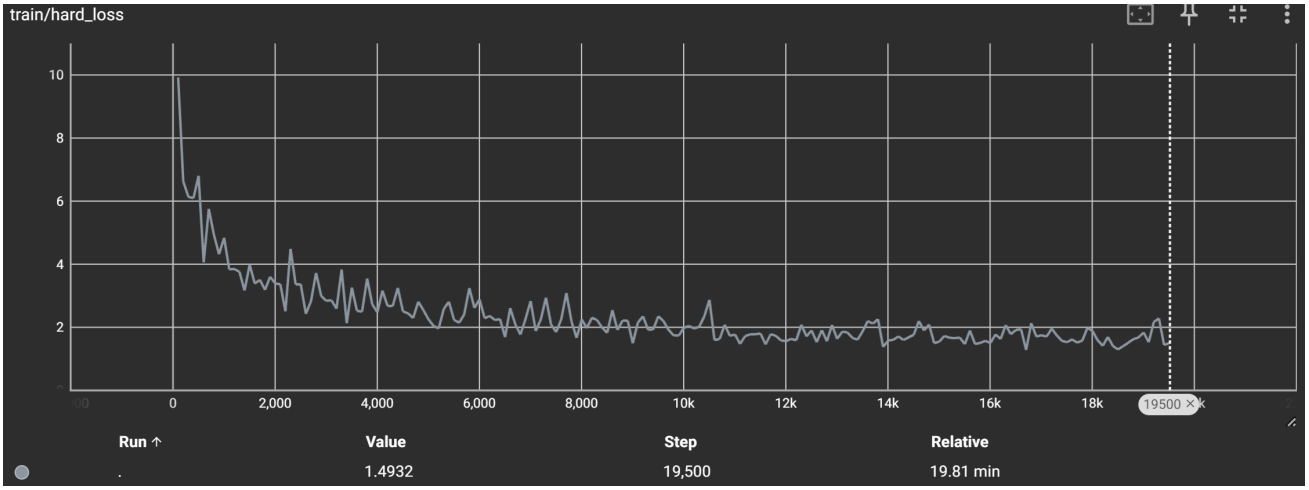
- stage one : 在引导层以及之前做训练
- stage two : 将stage one当作预训练来看，在预训练的基础上再次进行训练，观察否能够得到更好的效果

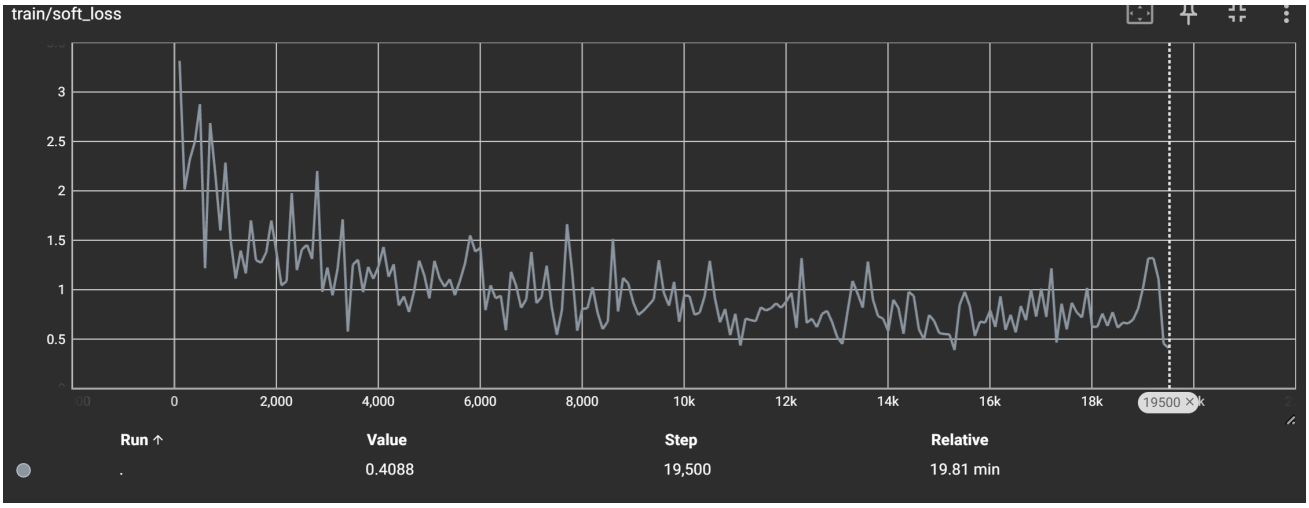



train/distill_loss
4
3
2
1
0
2,000
4,000
6,000
8,000
10k
12k
14k
16k
18k
19500
Run ↑
Value
Step
Relative
.
0.6382
19,500
19.81 min
train/distill_epoch_loss
3.5
3
2.5
2
1.5
1
0
5
10
15
20
25
30
35
40
45
50
Run ↑
Value
Step
Relative
.
0.7341
50
19.58 min

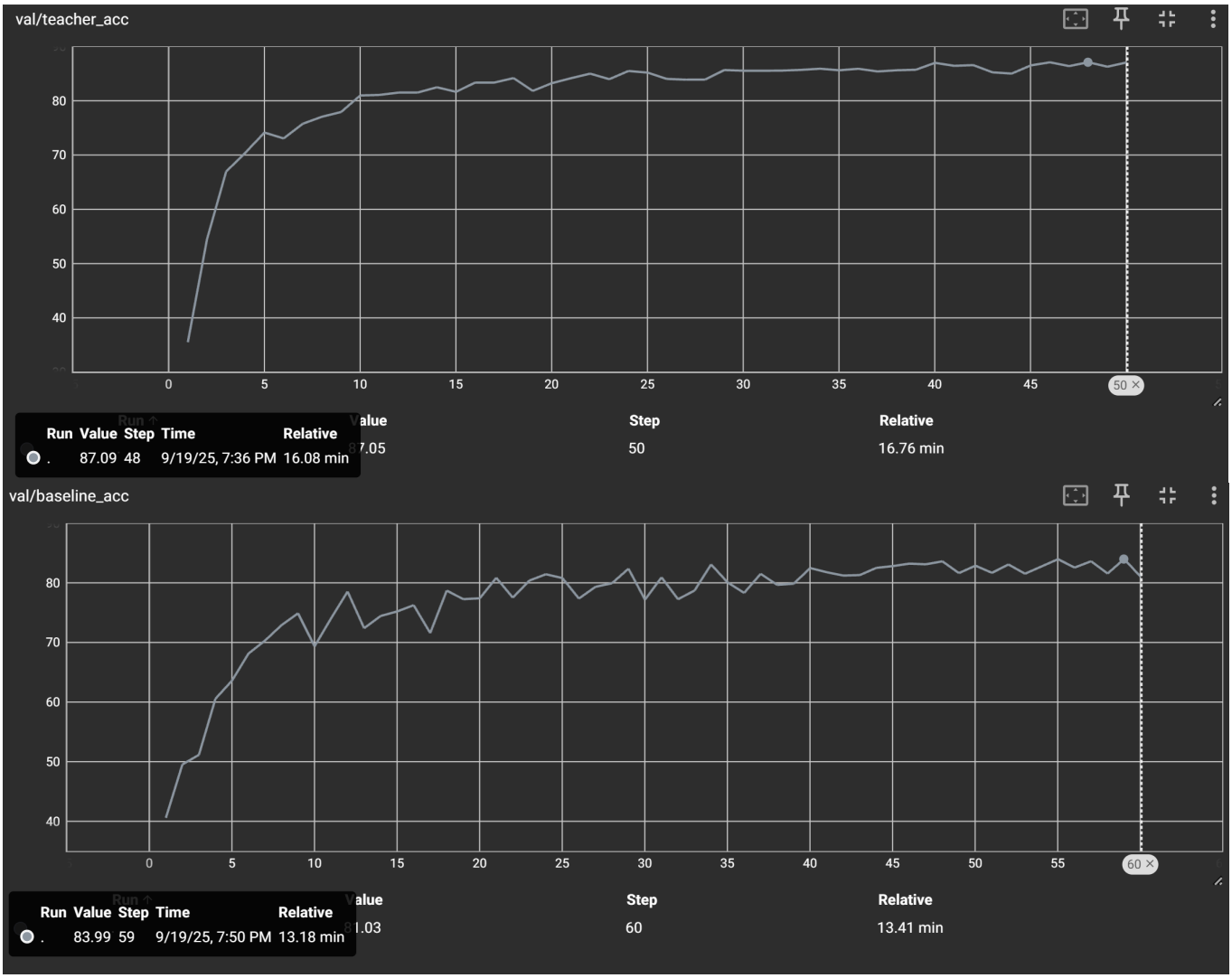

train/distill_acc
90
80
70
60
0
2,000
4,000
6,000
8,000
10k
12k
14k
16k
18k
19500
Run
Value
Step
Relative
.
95.3125
19,500
19.81 min
train/distill_epoch_acc
85
80
75
70
65
60
0
5
10
15
20
25
30
35
40
45
50
Run
Value
Step
Relative
.
88.384
50
19.58 min

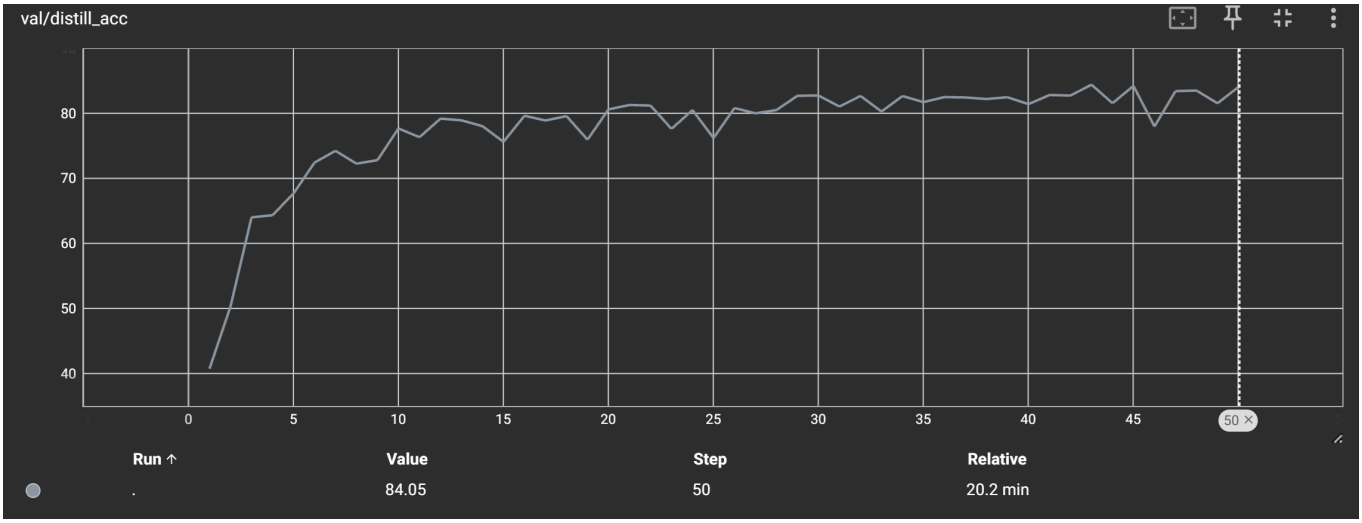
- 蒸馏的各个损失记录


train/hard_loss
10
8
6
4
2
0
2,000
4,000
6,000
8,000
10k
12k
14k
16k
18k
19500
Run
Value
Step
Relative
.
1.4932
19,500
19.81 min

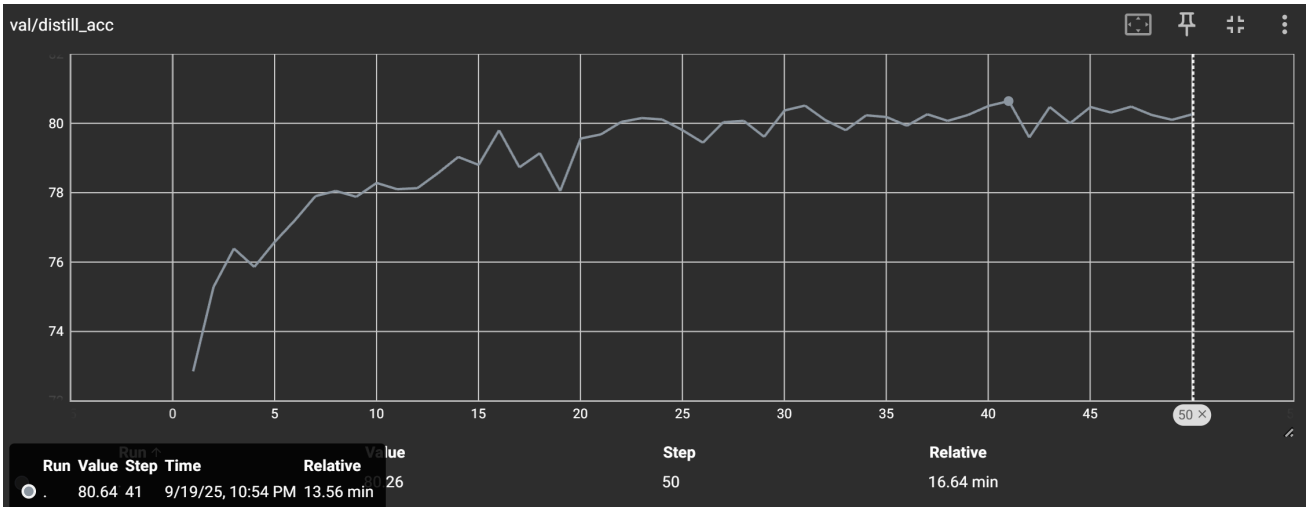

train/soft_loss
3
2.5
2
1.5
1
0.5
0
2,000
4,000
6,000
8,000
10k
12k
14k
16k
18k
19500
Run
Value
Step
Relative
.
0.4088
19,500
19.81 min

测试结果

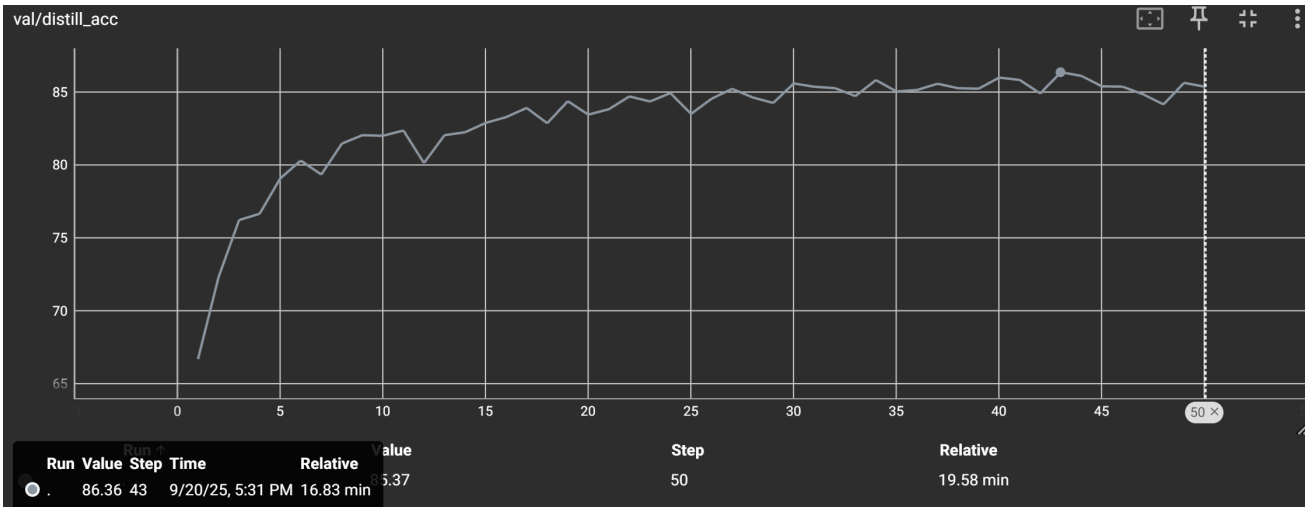

val/teacher_acc
80
70
60
50
40
0
5
10
15
20
25
30
35
40
45
50
Run
Value
Step
Time
Relative
. 87.09 48 9/19/25, 7:36 PM 16.08 min
Step
50
Relative
16.76 min
val/baseline_acc
80
70
60
50
40
0
5
10
15
20
25
30
35
40
45
50
55
60
Run
Value
Step
Time
Relative
. 83.99 59 9/19/25, 7:50 PM 13.18 min
Step
60
Relative
13.41 min



- stage one



- stage two



试验总结

模型	教师	基线	学生（蒸馏）	学生（实验一）	学生（实验二）
损失	***0.222***	0.339	0.85	1.38	**0.73**

模型	教师	基线	学生（蒸馏）	学生（实验一）	学生（实验二）
训练准确率	***92.7%***	88.42%	87.38%	81.24%	**88.4%**
验证准确率	***87.09%***	83.99%	84.05%	80.64%	**86.36%**
fit损失	-	-	-	0.23	0.23
软损失	-	-	1.05	0.97	**0.40**
硬损失	-	-	1.96	2.66	**1.49**
参数量	1.35M	306k~0.3M	306K~0.3M	306k~0.3M	306k~0.3M

教师＆基线比较

- 损失：两者使用了不同的学习率，学习率比较大的基线模型在收敛速度上较快；但是教师网络最终的损失比较小，这是在意料之内的，因为教师模型的表示能力比基线模型更强；两者在收敛之后稳定性都很好。
- 准确率：教师模型，基线模型在训练集上的准确率分别达到了**92.6% 和 88.5%**，可以看出，两者的差距比较明显，提升了绝对4%，两者都没有出现过拟合现象
- **测试**：两者在测试集上最佳准确率分别是**87%和84%，教师模型的表示能力高于基线模型**。在正常训练时，教师模型在epoch25时大概达到收敛，基线模型在15时达到收敛。

总结：教师模型能力强于基线模型，在更少的训练epoch之中达到了更好的效果，准确率有不小的差距。

试验一＆二比较

- fit损失：两次试验在stage one没有区别，损失都在0.23左右，表现不错，在2000×100步，也就是epoch6的时候都趋向于收敛。
- distill损失：损失分别为**1.3 和 0.7**，差距较大。在epoch程度损失函数较为平稳，但是从微观来看，实验一在不同的步骤损失浮动比较大。实验二在试验后期趋向于平稳。实验一收敛比较快。
  两者在hard和soft损失上表现差距很大，实验二明显优于实验一
- 准确率：**81% 和 88%**，差距很大
- **测试集**：**80.6% 和 86.3%**，差距很大

总结：实验二的效果显著优于实验一，我归结为以下要点

1. 冻结了前面的层之后，模型只需要调整后面的部分参数，所以试验一收敛比较快。这可以认为在已有输出的后面接上一个**分类头**，只需要调整分类头的部分参数。模型比较小，收敛比较快。但是，前面的层都接也就意味着，前面输出的特征图就是**一定程度上的原始图片**，这个微笑的模型要处理这个图片，是很艰难的。

2. 试验一在stage one之后，不一定能够保证引导层前面的参数学习教师模型的效果很优秀，**少参数经过学习之后不一定能够正确/完整替代多参数的表达效果**。考虑到两次试验都收敛到同一个损失，我们有理由相信这两次的训练都到达了最佳的效果。但是这个最佳效果在后续的实验一蒸馏之中并没有带来很好的效果。
3. **冻结参数影响了模型的调整能力**。冻结参数扼杀了根据正确标签纠正的机会。相反，解冻参数能够让模型在当前的模型之中，将引导层的输出调整向更适合小参数模型的方向。

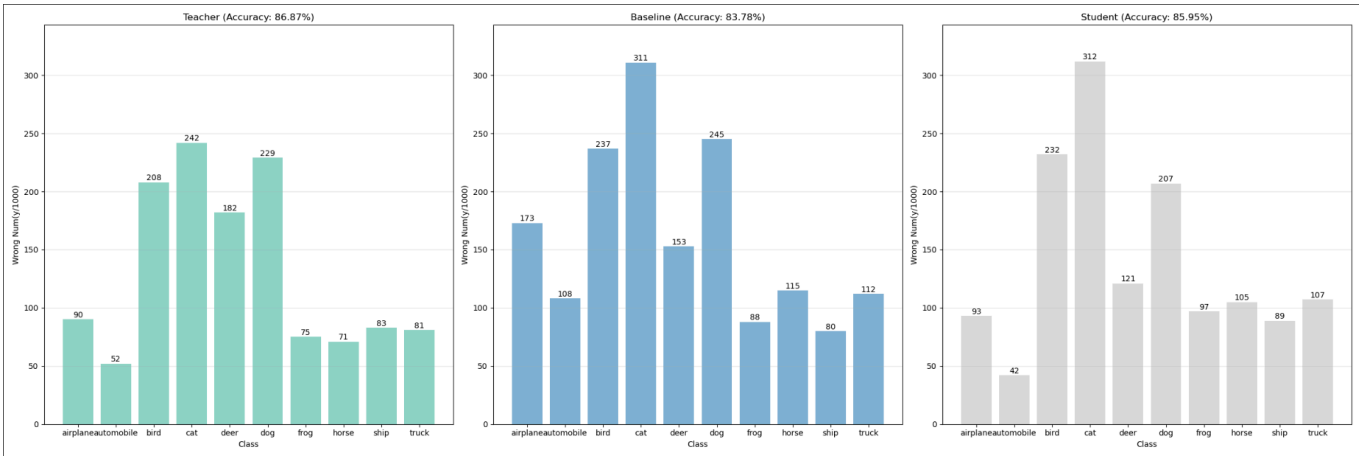
基线＆学生＆实验二比较

- 损失：学生和基线没有比较的必要，两者性质不同/实验二损失低于学生模型
- 准确率：两者都达到了88.4%的准确率，这一点惊人的相似
- 测试：**84% 和 86.3%**，fitnet模型有所提升。基于相同准确率考量，第二个模型的泛化效果更好

总结：

1. 基线和实验二在训练集上的正确率相等，但是实验二的泛化能力更好，一个观点就是**教师引导学生走向了正确方向**，一个观点是**蒸馏本身加强了模型的泛化能力**。但是学生和实验二表明，两者都发挥了作用。两者叠加效果更加。
2. **教师模型带领学生模型走向正确的方向**。虽然我们提到了，实验一之中少参数不能很好学习多参数的结论，但是它引导模型从局部最小值中走出，走向接近于教师的正确方向，因此，实验二训练之后相较于基线模型能够收敛向一个更好的结果。

分类结果分析



- dog 和 cat 两者错误最多
- student相较于teacher在一些方面有所提升。