基本设置

数据集：TinyStory **训练集: 2.12M, 验证集22k**

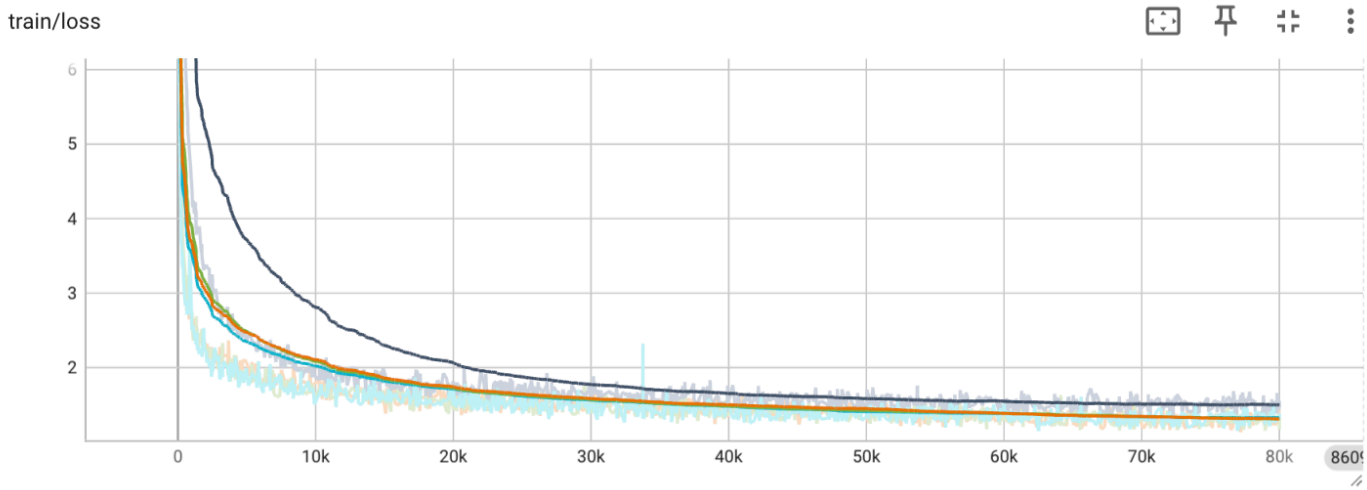
以下测试中若非专门测试项目，均使用了以下设置，模型参数量22M：

```
{
    "vocab_size": 10000,
    "context_length": 256,
    "d_model": 512,
    "num_layers": 4,
    "num_heads": 16,
    "d_ff": 1344,
    "rope_theta": 10000,
    "max_grad_norm": 5,
    "norm_type": 'pre',
    "betas": [
        0.9,
        0.999
    ],
}
```

RoPE, Pre-Norm, SwiGLU, layer-norm, AdamW

在测试不同参数时，保持总训练token量为：327,680,000

学习率测试



Run ↑	Smoothed	Value	Step	Relative
test_3/logs	1.499	1.5535	79,999	8.681 hr
test_4/logs	1.3277	1.3895	79,999	8.576 hr
test_6/logs	1.3114	1.3866	79,999	8.687 hr
test_7/logs	1.313	1.3227	79,999	8.523 hr

test_3:

lr: 1e-4

Validation loss: 1.48

test_4:

lr: 1e-3

validation loss: 1.34

test_6:

lr: 2e-3

validation loss: 1.30

test_7:

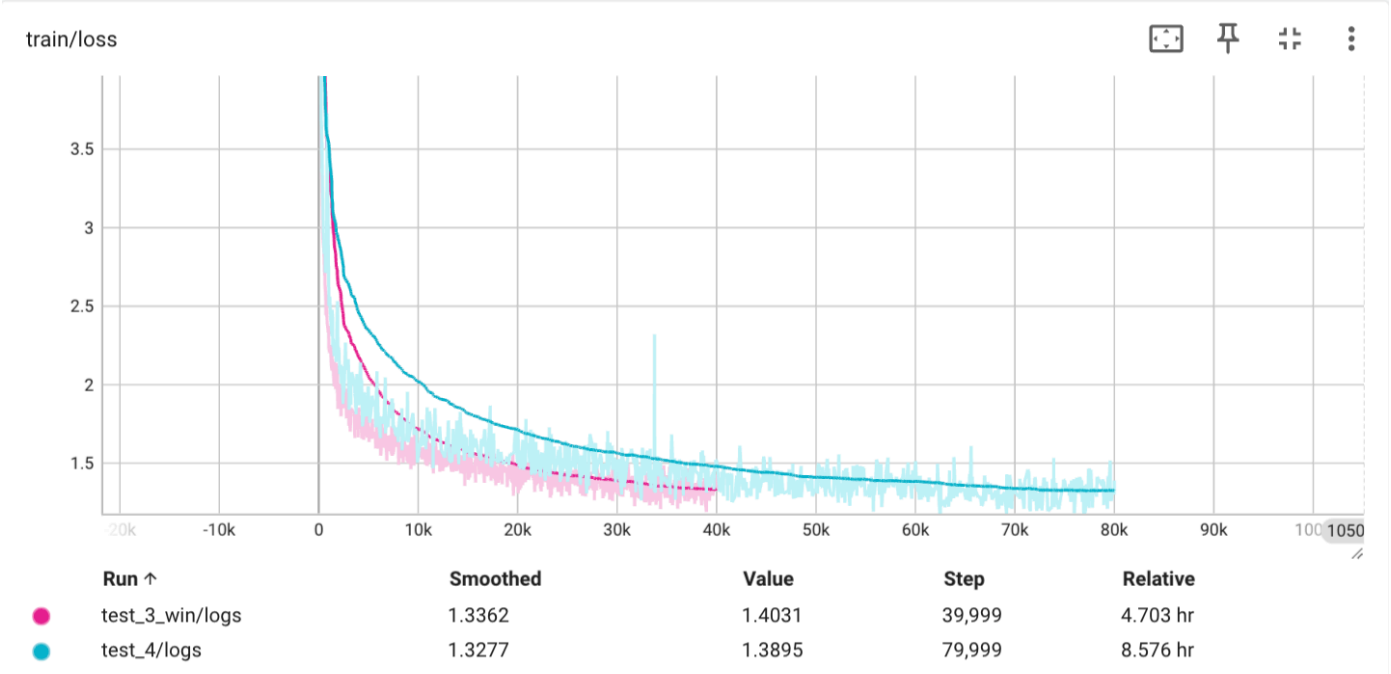
lr: 3e-3

validation loss: 1.31

再增加学习率得到不稳定的训练，损失函数发散

结论：随着学习率逐渐逼近发散临界值，收敛得更快更好，但是突破临界值则无法正常训练

Batch_size测试



test_4:

lr: 1e-3

validation loss: 1.34

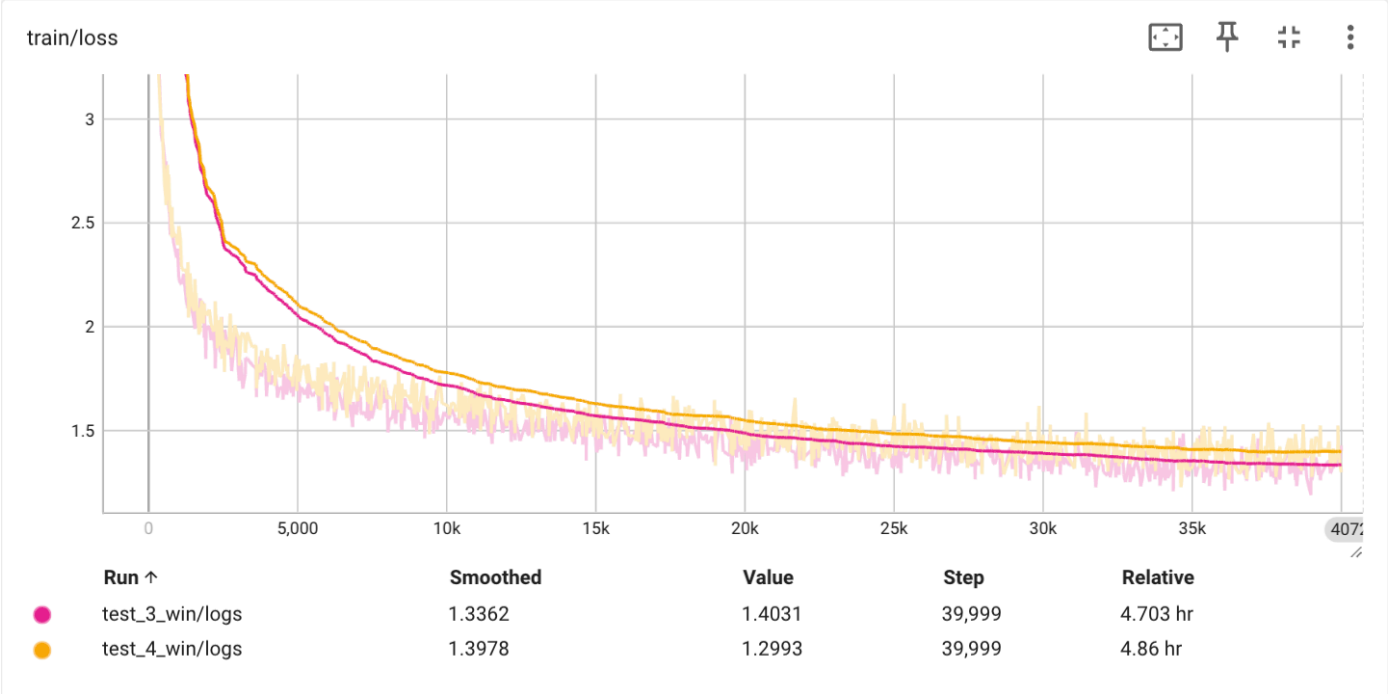
test_3_win:

lr: 1e-3

validation loss: 1.35

结论：在该小模型小数据测试上，batch_size对结果影响不大，收敛到的loss大致一致，不过更大的batch_size可以更好利用显卡资源，训练可能更快

AdamW betas测试



Run ↑	Smoothed	Value	Step	Relative
test_3_win/logs	1.3362	1.4031	39,999	4.703 hr
test_4_win/logs	1.3978	1.2993	39,999	4.86 hr

test_3_win:

lr: 1e-3

betas: (0.9, 0.999)

validation loss: 1.35

test_4_win:

lr: 1e-3

betas: (0.9, 0.95)

validation loss: 1.39

结论：更高的betas在本实验中收敛更好，测试损失更小。分析原因为低betas优化器对梯度方差的估计过于敏感，在本实验学习率较大（1e-3）的情况下，训练不稳定