总结

validation loss on TinyStory Dataset:

Baseline: 1.34

w/o layer_norm: 1.54

w/o RoPE: 1.38

Use post-norm: 1.36

Use SiLU: 1.36

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
    ],
}
```

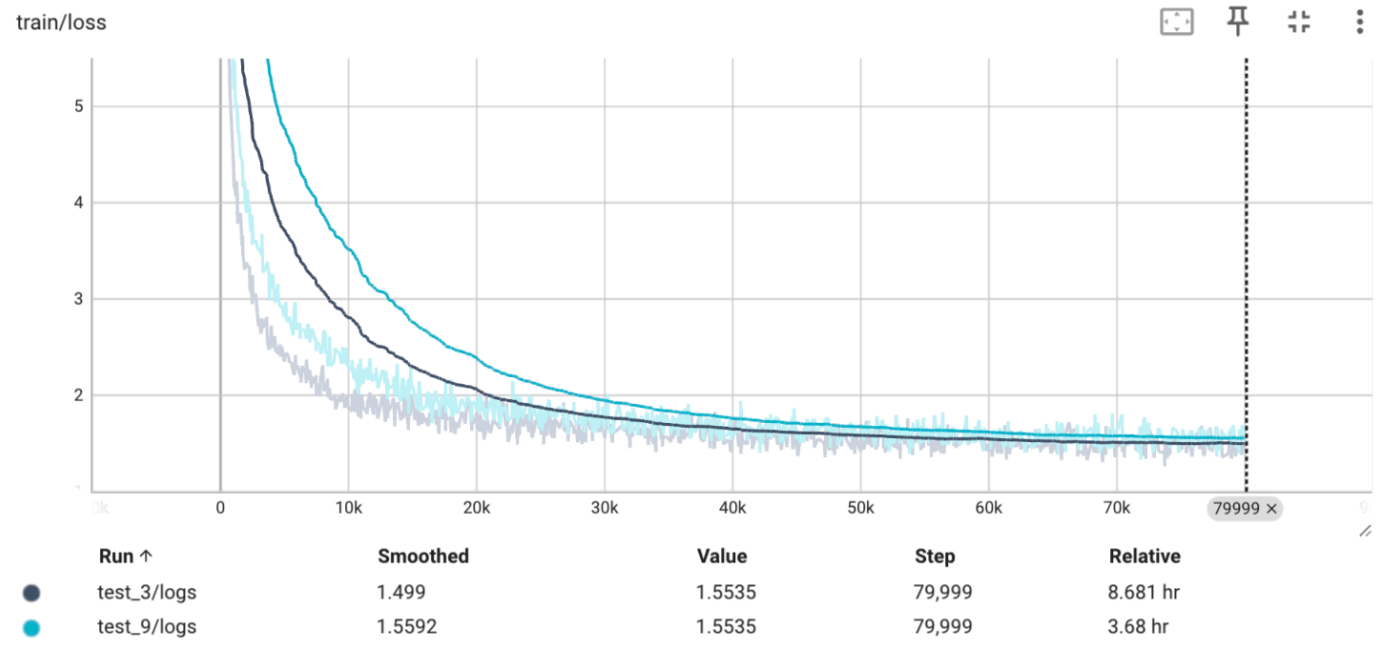
RoPE, Pre-Norm, SwiGLU, layer-norm, AdamW

在测试不同参数时，保持总训练token量为：327,680,000，同时，在测试项目影响模型总参数量时，对模型结构进行细微调整，保证总参数量一致。

layer norm层

去除layer norm层

以1e-3学习率（有layer norm层优选学习率）训练时，去除layer norm层的模型训练失败，中途损失爆炸为NAN，换成1e-4可以完成训练



Run	Smoothed	Value	Step	Relative
test_3/logs	1.499	1.5535	79,999	8.681 hr
test_9/logs	1.5592	1.5535	79,999	3.68 hr

test_3:

lr: 1e-4

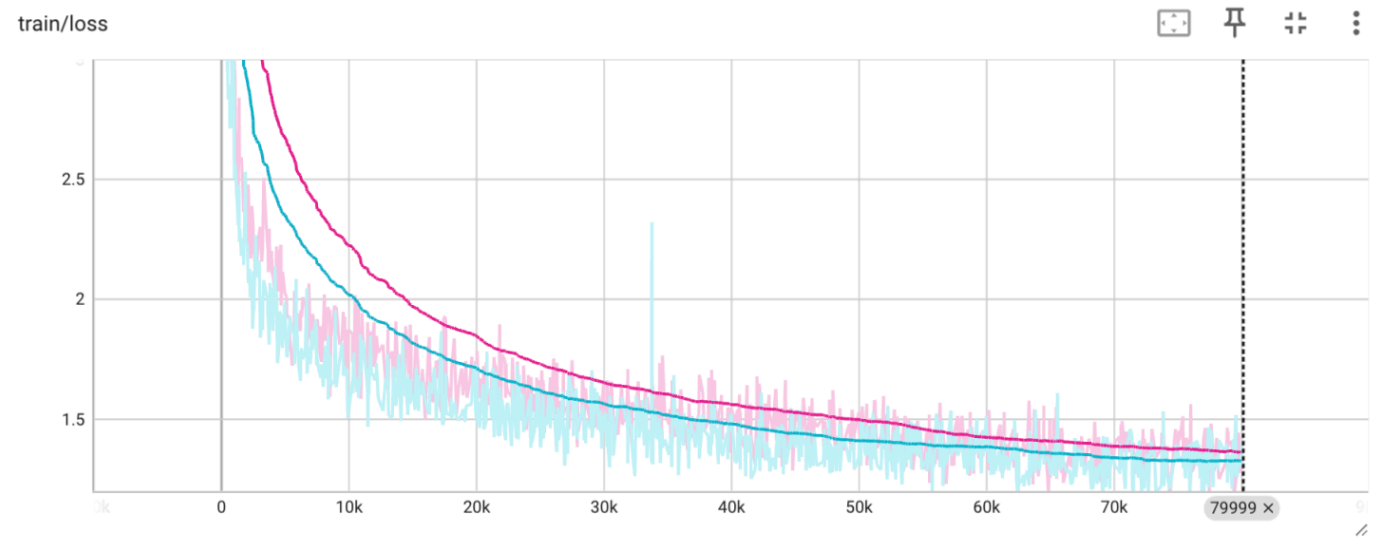
Validation loss: 1.48

test_9（无layer norm）：

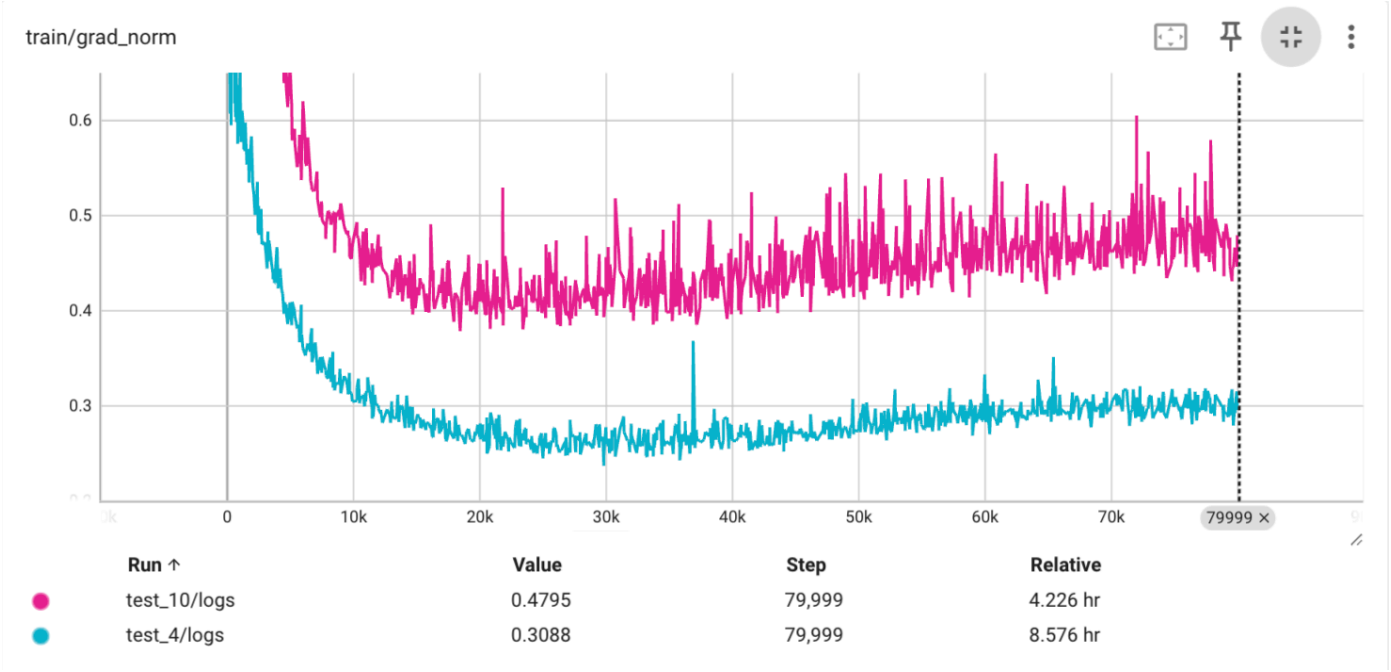
lr: 1e-4

validation loss: 1.54

将pre norm换成post norm



Run	Smoothed	Value	Step	Relative
test_10/logs	1.3621	1.3493	79,999	4.226 hr
test_4/logs	1.3277	1.3895	79,999	8.576 hr



test_4:

lr: 1e-3

validation loss: 1.34

Test_10（post_norm）：

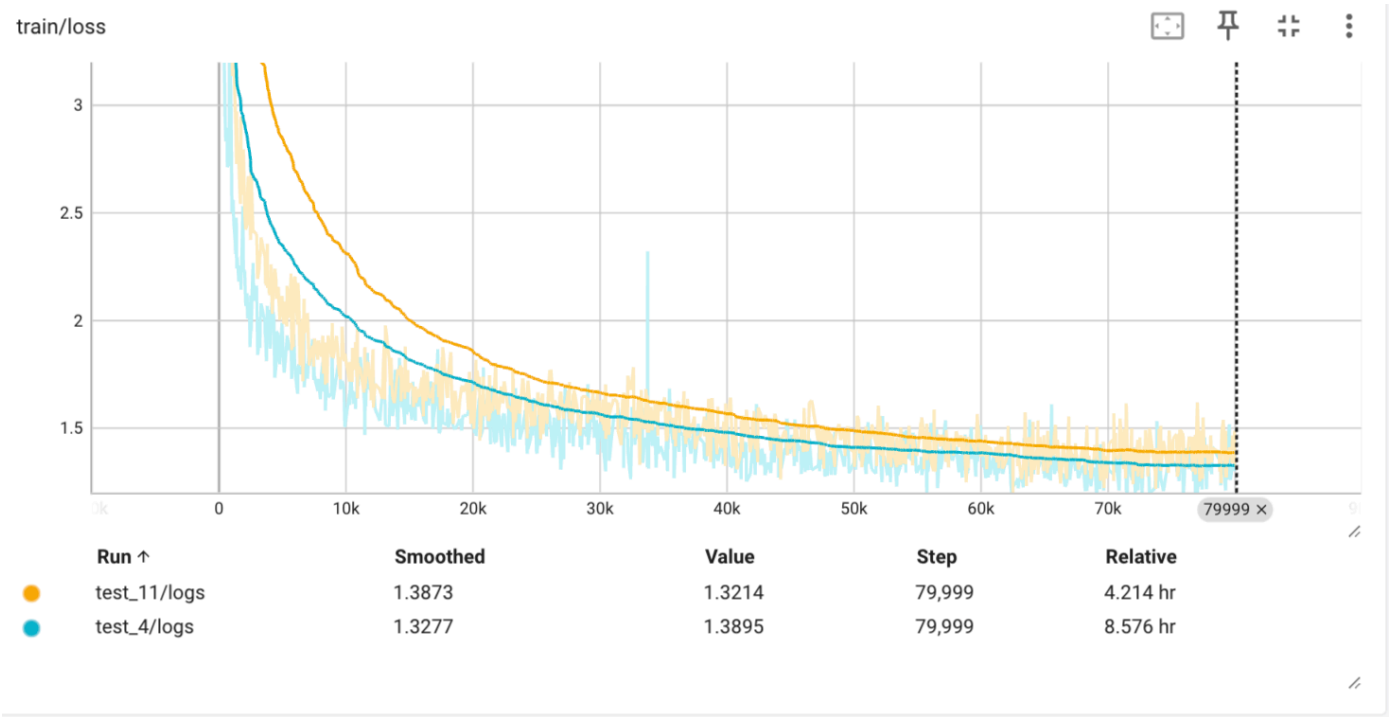
lr: 1e-3

validation loss: 1.36

结论：在 Transformer 或深层网络中，使用 pre-norm 更有助于梯度稳定、训练过程平稳、模型更容易收敛。post-norm 则更容易带来训练困难，尤其在 lr 相同的设置下表现略差。是体现在本小模型小样本中的结论，目前模型大部分都使用pre-norm，感觉是训练更稳定

RoPE

删除旋转位置编码



test_4:

lr: 1e-3

validation loss: 1.34

Test_11（删除位置编码）：

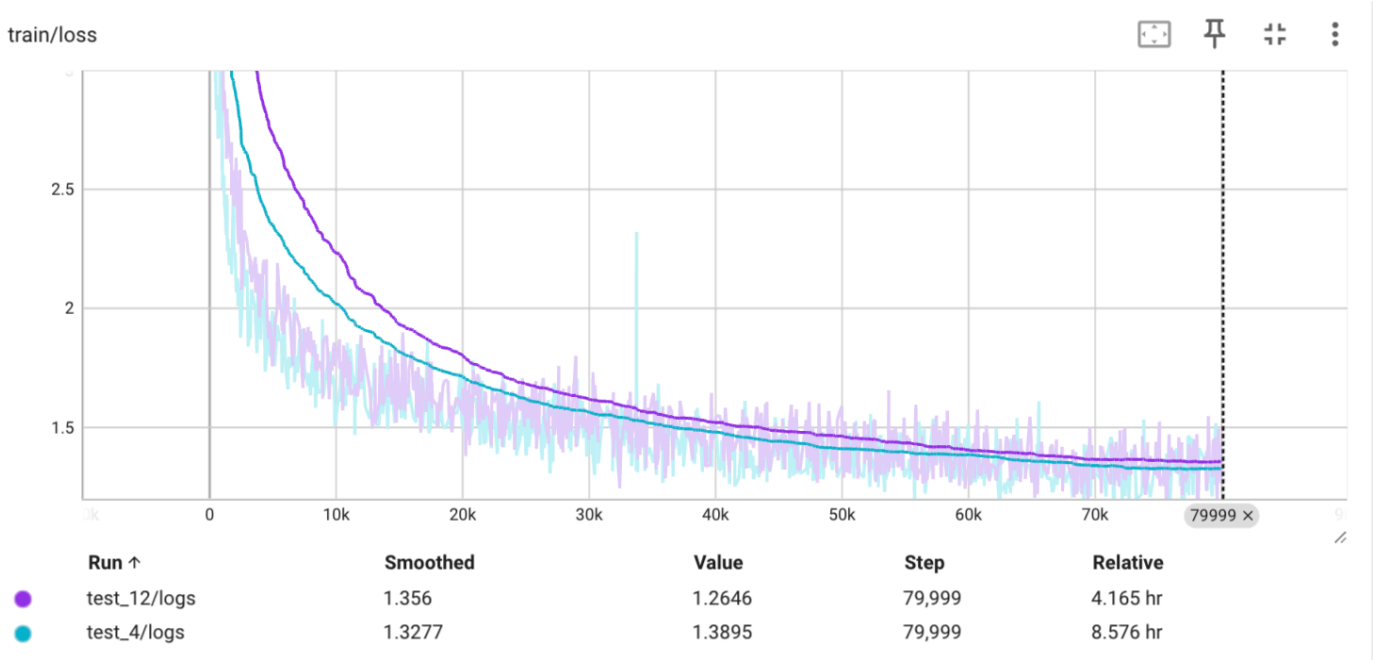
lr: 1e-3

validation loss: 1.38

结论：虽然理论上来说不需要显示位置编码通过transformer本身就可以学习到位置信息，但是加入RoPE模型效果优于不加位置编码

SwiGLU

将SwiGLU换成SiLU，去除门控机制



test_4:

lr: 1e-3

validation loss: 1.34

test_12:

lr: 1e-3

validation loss: 1.36

结论：前馈神经网络的选择上，使用SwiGLU的门控机制训练得到的模型效果优于使用不使用门控机制的SiLU