2

Transformer实验过程

数据集:WMT14 英德数据集

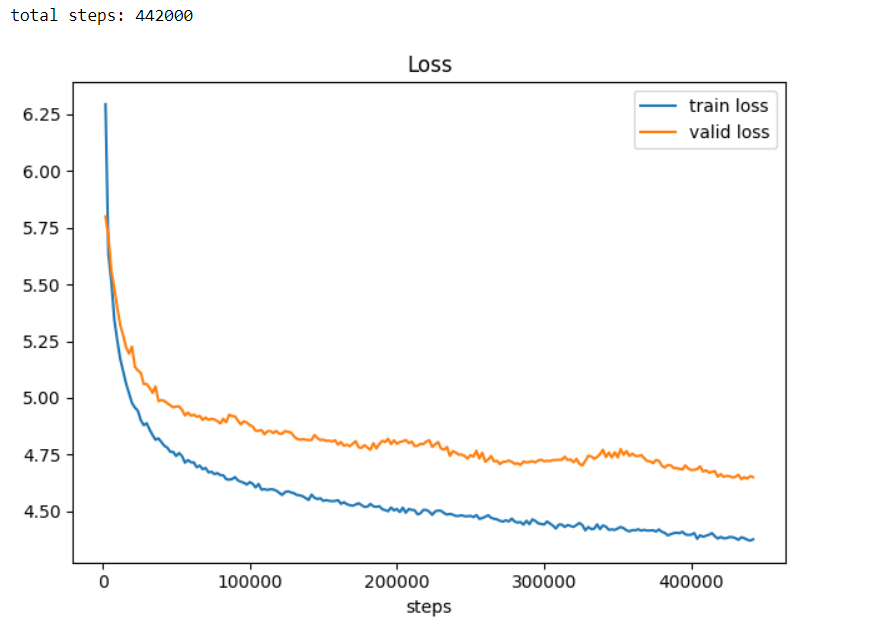
Transformer为Attention is all you need.中的base

RTX2080单卡

(batch\_size=8)

10 hours 15steps

30 hours 44.2steps loss 4.38 ppl 79.55



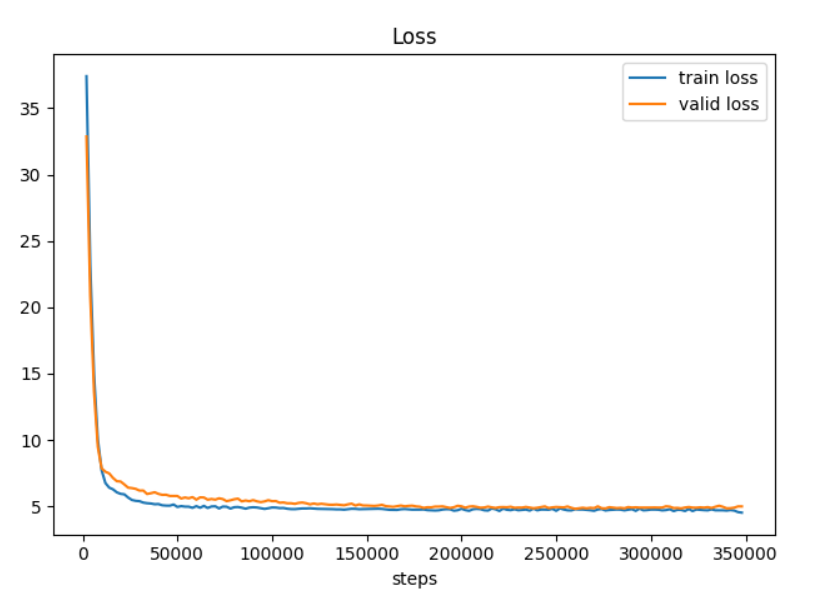
共44.2万步,最后一部分给丢了

两天时间下降到4~5之间,还有下降趋势。

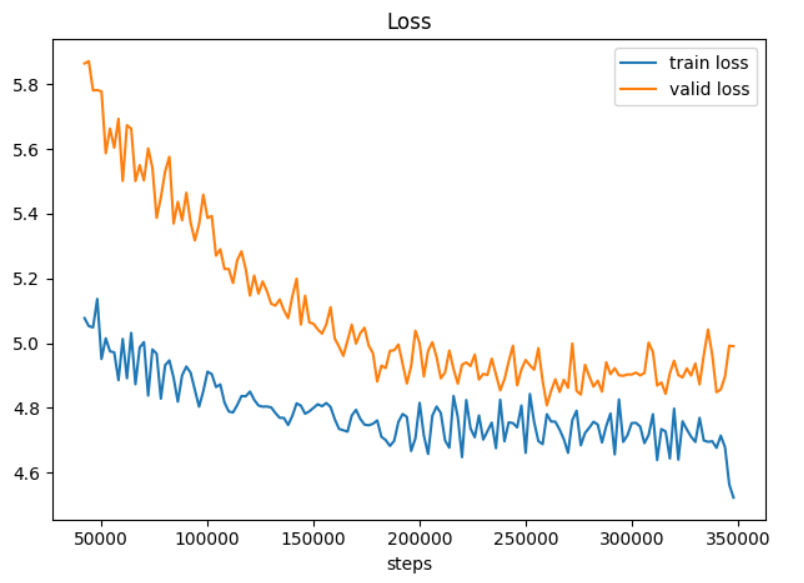
问题是为什么在第一个epoch中会出现过拟合的现象，而且valid是验证模式,train是训练模式。好怪。

第二版Transformer

改了很多:换了Tokenizer(之前Sentencepiece的写法有问题,只用2048个句子训练Tokenizer),重写了数据生成(按Token数量生成一批数据,截断了太长的数据),权重绑定,用TransformerLayer重写模型,加了LabelSmooth ,激活函数换成gelu,



只看40000步的loss



正式训练的记录:

前几天没记,因为log会记,现在记是为了看loss是否真的下降了。根据知乎的一篇文章要训练23轮之后才会拟合好，现在只一次训练了10轮。大概一天一轮左右吧。

每天晚上看一次，不要太频繁了，不然没有信心了。

第三天晚上：

Valid\_loss = 5.12346 train\_ppl = 260.88

第四天晚上18:10:

Valid\_loss = 5.06045 train\_ppl = 253.34

第五天晚上 21:37:

Valid\_loss = 4.95481 train\_ppl = 237.43

第六天晚上 20:08:

Valid\_loss = 4.88754 train\_loss=5.42 train\_ppl = 226.73

3月7日

第七天晚上 21:36:

Valid\_loss = 4.78126 train\_loss=5.36 train\_ppl = 212.25

3月8日

第八天晚上 22:11:

Valid\_loss = 4.72332 train\_loss=5.35 train\_ppl = 210.97

学习率不对劲。lr已经是1e-5了

应该到25000个Token后才加一步才对，和梯度累积一样。

3月10日

第十天早上 9:36:

Valid\_loss = 4.64546 train\_loss=5.26 train\_ppl = 192.73

第十天晚上 20:35:

Valid\_loss = 4.62459 train\_loss=5.21 train\_ppl = 182.37