2

Transformer实验过程

数据集:WMT14 英德数据集

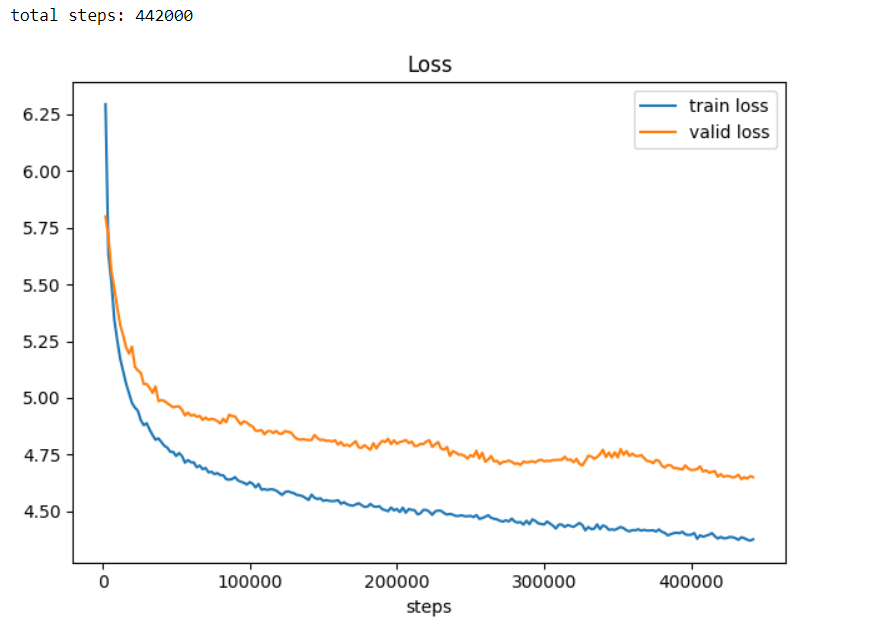
Transformer为Attention is all you need.中的base

RTX2080单卡

(batch\_size=8)

10 hours 15steps

30 hours 44.2steps loss 4.38 ppl 79.55



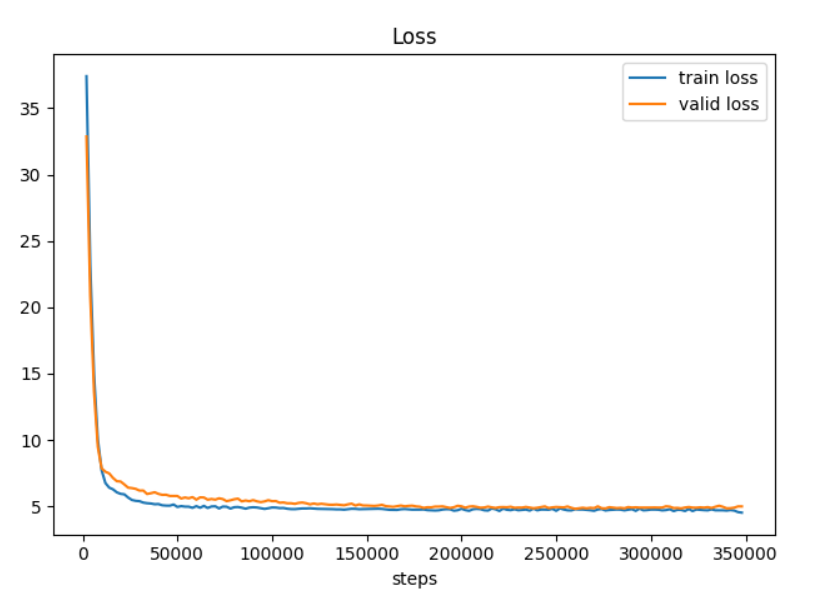
共44.2万步,最后一部分给丢了

两天时间下降到4~5之间,还有下降趋势。

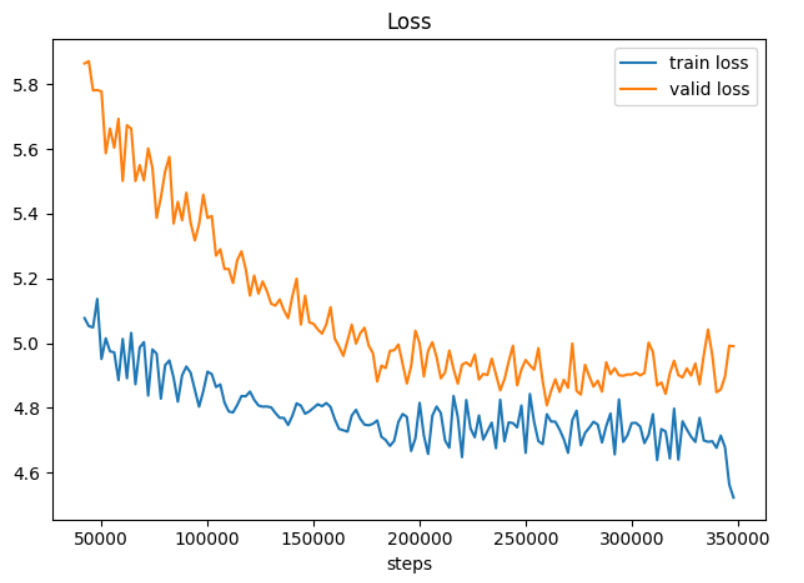
问题是为什么在第一个epoch中会出现过拟合的现象，而且valid是验证模式,train是训练模式。好怪。

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:

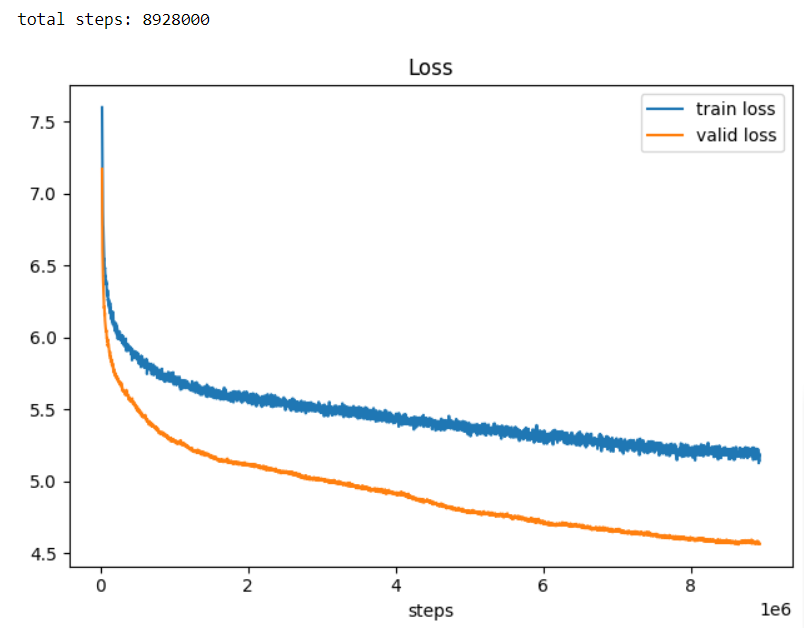
Valid\_loss = 4.62459 train\_loss=5.21 train\_ppl = 182.37

3月13日 21:04:

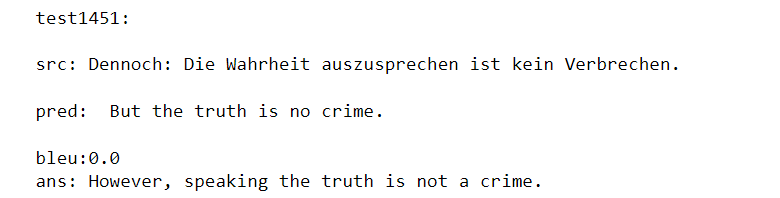
第十三天晚上

Log.txt显示上次更新是 a day ago,所以训练十个epoch花了12天

Valid\_loss = 4.56293 train\_loss=5.14 train\_ppl = 171.30



Train loss更高应该是因为dropout



BLEU4可能并非完美,这，我不想吐槽。“the truth is”刚好只占3个Token。

平均BLEU为3.53

比第8个epoch低了(epoch8为5)

会不会过拟合了，比较valid\_loss和valid\_BLEU不一样。

真不好判断是否过拟合了，毕竟批大小、学习率和激活函数都不同。

这BLEU分数也太低了，valid的BLEU有20多，valid和test之间的gap也太大了。

不对，valid的BLEU不是自回归，艹，根本不能判断。重练吧。

3月14日:

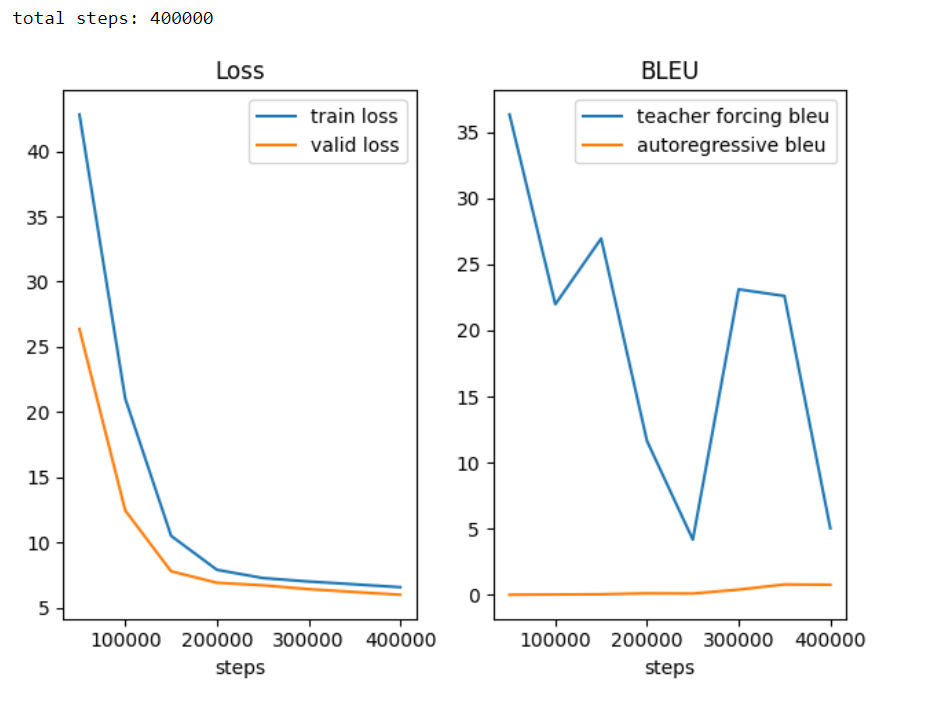
提前backward确实不占用那么多内存了,可以考虑增大一个batch里的tokens数和序列长度。

3月15日:

昨天运行了大概十几个小时之后超内存了。今天减少batch\_token到1400,截断长度为650。

进程号为17139

昨天的数据：



可见teacher forcing bleu 和 autoregressive有很大的gap。

autoregressive采取和预测时完全一致的设置，即自回归预测。

Emmm，忘了把checkpoint给删了，所以现在继承了昨天的训练，不过问题不大，与一次性训练的区别只有：

1. epoch1有些数据被多训练了一次（随着epoch的增多，这个影响会越来越小）
2. Batch\_token变小了,截断长度也变小了。（这个也没有影响）

3/15

第一晚：

| steps 500000|lr 0.00015073317636017444 | s/step 0.10 | loss 6.23 | ppl 510.21

==============================

A strategy of strategy to be- Obama-inction of Obama............ Obama Obama Obama Obama Obama Obama Obama Obama Obama Obama

A Republican strategy to counter the re-election of Obama

valid\_loss:5.62950

teacher forcing bleu:10.502298921346664

==============================

autoregressive bleu: 1.3754440773657577

best autoregressive bleu score: 1.3754440773657577

3/16

第二晚：

| steps 1350000|lr 0.00020937112928702688 | s/step 0.10 | loss 4.73 | ppl 113.67

==============================

A strategy of strategy to make Obama Obama-exction of Obama

A Republican strategy to counter the re-election of Obama

valid\_loss:4.21882

teacher forcing bleu:18.535153567790985

==============================

autoregressive bleu: 5.72969986582261

best autoregressive bleu score: 5.72969986582261

save to best\_bleu.pt

save to checkpoint.tar

Valid Loss下降得多快，一天下降1.4,之前下降只有0.1

