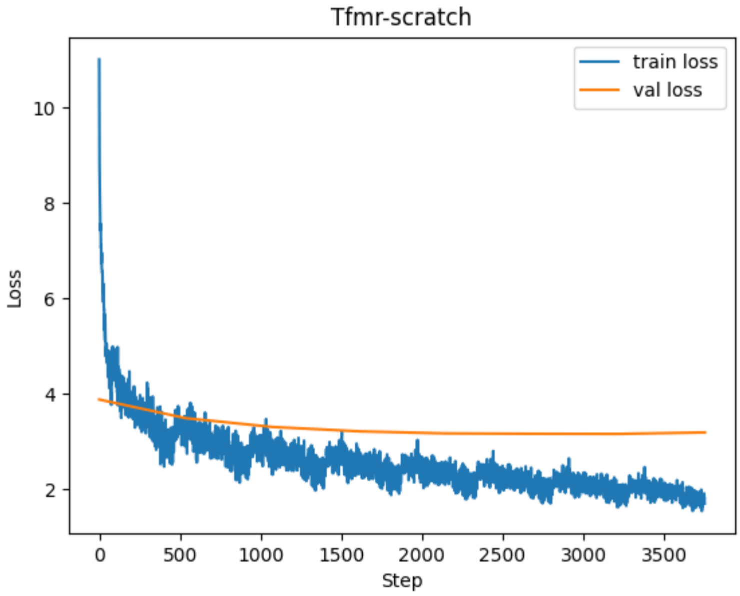
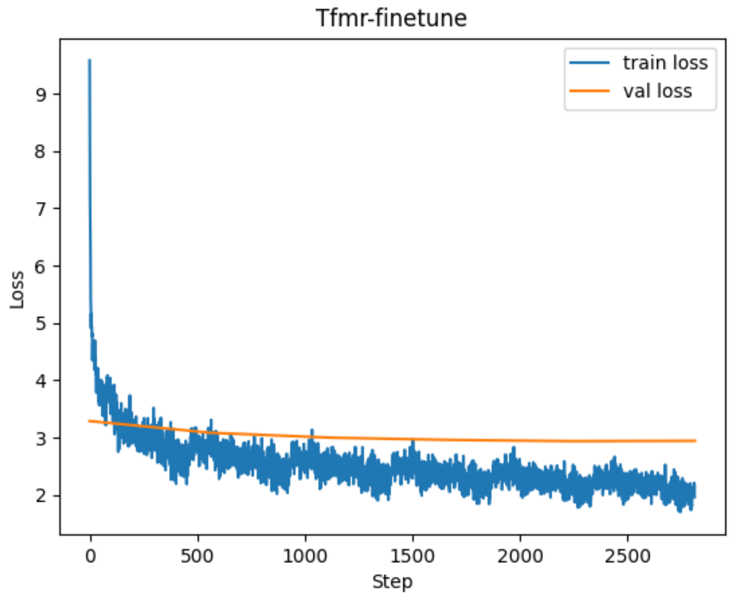
Text Generation with the Transformer Decoder

张益铭 车13 2021010552

Train two models

在默认的设置下，两个模型的loss values如下所示：





Metrics	Tfmr-scratch	Tfmr-finetune
Perplexity	18.91	15.48
Forward BLEU	0.576	0.569
Backward BLEU	0.429	0.433
Harmonic BLEU	0.492	0.492

- Tfmr-scratch和Tfmr-finetune的loss图像表明训练效率和收敛性存在差异。finetune的loss曲线收敛速度更快，数值更小，表明微调通过利用预先存在的知识能够帮助模型更有效地学习。
- Perplexity方面，与scratch（18.91）相比，finetune（15.48）有显著的改善。表明微调模型更善于生成类似于目标数据的序列，显示了从预训练模型开始的好处。
- BLEU方面，模型之间的差异很小，Tfmr-scratch的正向BLEU略有优势，Tfmr-finetune的反向BLEU 略有优势。

Generation results

- **Tfmr-scratch**

Metrics	random, $\tau = 1$	random, $\tau = 0.7$	top-p=0.9, $\tau = 1$	top-p=0.9, $\tau = 0.7$
Forward BLEU	0.576	0.806	0.696	0.873
Backward BLEU	0.429	0.382	0.421	0.306
Harmonic BLEU	0.492	0.518	0.525	0.453

- **Tfmr-finetune**

Metrics	random, $\tau = 1$	random, $\tau = 0.7$	top-p=0.9, $\tau = 1$	top-p=0.9, $\tau = 0.7$
Forward BLEU	0.569	0.810	0.690	0.887
Backward BLEU	0.433	0.378	0.417	0.312
Harmonic BLEU	0.492	0.516	0.520	0.461

- 在解码策略方面，random能够带来更高的反向BLEU值，表明具有更高的多样性；而top-p能够带来更高的正向BLEU值，表明更加准确。这是由于在随机策略下，所有token都有可能被选中，因此具有更高的多样性；而在top-p策略下，只有概率较大的token被选中，从而更加准确。
- 在温度方面，$\tau = 0.7$的正向BLEU更高，更低的温度能够在softmax分布中更加集中地选择概率较大的词，提高准确性；$\tau = 1$的反向BLEU更高，更高的温度能够使得softmax分布更加均匀，提高多样性。
- 对比scratch和finetune，finetune对BLEU的提升并不大。

10 random sentences

Tfmr-scratch

- random, $\tau = 1$

```
A red bird is looking down at something on a concrete road .
Two giraffesulsive and three zebra standing in jungle grass .
A tire tries to be driven off to the beach .
A photo of some traffic cones and riders on a city road .
A white passenger bus is parked in a parking lot .
A green fire hydrant sits by a sidewalk with a street sign on it .
A red fire hydrant sits on a sidewalk next to a street .
Two ladies in gas station .
A park bench is set for passengers flying underneath a tree .
Two city parking meters in front of a traffic light .
```

6 grammar errors

- random, $\tau = 0.7$

```
A red bus parked in front of a tall building .
A man is sitting on a bench in the grass .
A black bird is perched on the edge of a wooden bench .
A man sitting on a bench next to a tree .
A white , blue and yellow airplane flying over a body of water .
A giraffe standing next to a tree filled with rocks .
A red fire hydrant sits on the side of the street .
Two ladies in a bathing suit riding a horse is sitting on a bench .
A park bench is set up against a blue sky .
A truck is driving in the rain with pedestrians walking by .
```

4 grammar errors

- top-p=0.9, $\tau = 1$

```
A red bus passes towards a tree on the curb .
Two giraffes are walking through a wooded area .
A side of a train and station with a couple of people .
A photo of some traffic cones and people on a street .
A white passenger bus is parked in a parking lot .
A green fire hydrant sits in a snow covered park .
A red fire hydrant sits on a sidewalk next to a street .
Two ladies in gas station .
A park bench is set for passengers to eat outside .
Two city buses are traveling down a busy city street .
```

4 grammar errors

- top-p=0.9, $\tau = 0.7$

```
A red bus parked in front of a tall building .
A man is sitting on a bench in the grass .
A black bird is perched on the edge of a wooden bench .
A man sitting on a bench next to a tree .
A white , blue and yellow airplane flying over a field .
A giraffe standing next to a tree filled with rocks .
A red fire hydrant on a sidewalk next to a street .
A bench is next to a wall with some leaves .
A red fire hydrant sitting in the middle of a park .
A city street filled with lots of traffic on a street corner .
```

5 grammar errors

Tfmr-finetune

- random, $\tau = 1$

```
A red bird resting on the grass next to a metal object .
Two giraffes are walking around a bush while grazing .
A side view of a streetlight with a pedestrian walking and a lot .
A photo of some traffic cones and riders are jumping off .
A white , blue and yellow bus parked on the side of a street .
A green fire hydrant sits by a river with rocks and hills behind it .
A street scene with a bicycle parked on the grass .
Two ladies stand next to each other on the sidewalk .
a park bench is decorated with trees and trees outside .
Two city parking meters in front of a building on a wide city street .
```

3 grammar errors

- random, $\tau = 0.7$

```
A red and yellow bus is pulled up to a streetlight .
A man is sitting on a bench in the grassy area .
A black and white photo of a bus parked next to a bus stop .
A man sitting on a bench next to a tree .
A white , blue and yellow bus parked on the side of a street .
A green fire hydrant sits by a river with a street sign on it .
A red fire hydrant sits on the side of the street .
Two women stand next to each other on a bench .
A bus driving down a street with a few vehicles .
A truck is loaded onto the side of the road .
```

2 grammar errors

- top-p=0.9, $\tau = 1$

```
A red and yellow bus is pulled up to a metal pole .
Two giraffes are walking around a bush while grazing .
A side view of a streetlight with a pedestrian walking and a lot .
A photo of some traffic lights and people are jumping off .
A white , blue and yellow bus parked on the side of a street .
A green fire hydrant sits by a river with rocks and hills behind it .
A street scene with a bicycle parked on the side of the street .
Two ladies stand next to each other on the sidewalk .
A park bench is decorated for the camera , outside .
Two city buses are traveling down the busy streets on a wide city street .
```

4 grammar errors

- top-p=0.9, $\tau = 0.7$

```
A red and yellow bus is parked on the street .
A man is sitting on a bench in the grass .
A black and white photo of a bus parked next to a bus stop .
A man sitting on a bench next to a tree .
A white , blue and yellow bus parked on the side of a street .
A green fire hydrant sits in a grassy area near a tree .
A red fire hydrant sits on the side of the street .
A bench is next to a tree in the grass .
A bus driving down a street with a few vehicles .
A city street filled with lots of traffic on a sunny day .
```

3 grammar errors

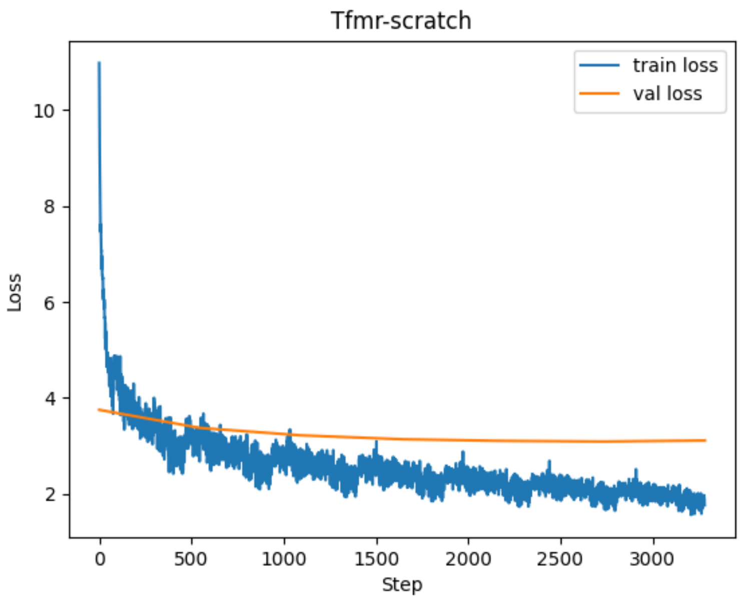
典型的错误有：

- 主谓不一致：主语和谓语不匹配。
- 缺少动词：句子缺少核心动词。
- 标点错误：如逗号漏用或误用。
- 句子拼接：用逗号连接独立分句。
- 句子残缺：缺少主语或谓语。
- 表达不清：结构不自然或含义模糊。
- 代词错误：代词使用不当或指代不清。
- 修饰语位置错误：修饰语位置不当，造成歧义。
- 时态不一致：混用不同时态。

总体上看，$\tau = 1$时语法错误更多，而Tfmr-finetune能够减少语法错误

Normalization

在random, $\tau = 1$下，Post-Norm训练结果如下



Metrics	Pre-Norm	Post-Norm
Perplexity	18.91	17.88

Metrics	Pre-Norm	Post-Norm
Forward BLEU	0.576	0.579
Backward BLEU	0.429	0.434
Harmonic BLEU	0.492	0.496

后向归一化的PPL更小，BLEU值更高，训练的结果更好一些。可能的原因有：

- 在后向归一化中，归一化的均值和方差是基于当前小批量数据动态计算的。随着训练的进行，模型参数逐渐更新，因此后向的统计信息更贴合模型当前的参数状态，从而使得归一化更具代表性，提升了稳定性。
- 后向归一化可以有效地缩放输出和梯度，避免梯度的爆炸或消失，使得梯度更稳定、模型更容易训练。

Final network

我选用的模型为在随机解码策略、温度为0.7下的finetune模型。metrics如下：

Perplexity	Forward BLEU	Backward BLEU	Harmonic BLEU
15.48	0.810	0.378	0.516

Questions

1. Compare Transformer and RNN from at least two perspectives such as time/space complexity, performance, positional encoding, etc.
    - performance
        - Transformer在机器翻译、文本生成和自然语言理解任务中表现非常优异，在长序列任务中能够较好地捕捉全局信息。自注意力机制能够灵活地学习不同位置间的依赖关系，不受固定顺序的限制。
        - 传统的RNN在捕捉长距离依赖上效果较差，容易遗忘前面的信息。尽管LSTM和GRU等变种改善了这一问题，但在长序列建模上依然不如Transformer。
    - positional encoding
        - Transformer没有RNN的递归结构，它无法直接获取输入序列的顺序信息，因此需要显式的位置编码来提供位置信息；而RNN则天然地具有顺序性，通过逐步处理序列的每个元素来隐式地编码位置信息。
2. Regarding the inference time complexity, answer the following question.
    1. During inference, we usually set `use_cache` in `model_tfmr.py` to `True` . What is the argument used for? What will happen if we set it to False?
        - `use_cache`主要是为了提升推理效率。`use_cache=True`时，模型会缓存前面生成的隐藏状态，在生成新词时会用到之前所有位置的隐藏状态。如果启用缓存，模型只需计算当前时间步的前向传播，将之前的隐藏状态缓存起来并复用，从而减少计算量。
        - 如果将`use_cache`设置为`False`，则在每次生成新词时，模型都将重新计算所有时间步的隐藏状态。会导致推理速度显著降低，尤其是在长文本生成中。同时还会导致资源占用增加，增加计算和内存的开销。

2. Denote the whole sequence as $L = (l_0 =< |\text{endoftext}| >, l_1, l_2, \dots, l_T)$, please give the inference **time complexity** when decoding the token l_t ,i.e., the t-th loop in the `inference` function `of model_tfmr.py` when decoding the first example, and the whole time complexity for decoding the whole sequence L. We denote the hidden state dimension as d (so that the dimension of the intermediate state of the feed forward layer is $4d$), the number of heads in multi-head attention as n, the number of hidden Transformer blocks as B, the vocab size as V.

   - decode the token l_t:
     1. The time complexity of calculating dot products between the query vector q_t and all t key vectors is $O(t \cdot d)$ per head. Since there are nnn heads, the multi-head attention has complexity $O(n \cdot t \cdot d)$.
     2. Each token passes through a two-layer feed-forward network with intermediate dimensionality $4d$, resulting in a time complexity of $O(4d^2)$ per token.
     3. Summing these, the time complexity per block is $O(n \cdot t \cdot d + 4d^2)$. With B Transformer blocks, the complexity for decoding l_t is: $O(B \cdot (n \cdot t \cdot d + 4d^2))$
     4. To generate the final output token l_t, we project the hidden state to the vocabulary space, which involves a matrix multiplication of complexity $O(d \cdot V)$.
     5. The time complexity for decoding a single token l_t is: $O(B \cdot (n \cdot t \cdot d + 4d^2) + d \cdot V)$
   - decoding the whole sequence L:
     1. To decode the entire sequence of length T, we sum the time complexity for each l_t from $t = 1$ to $t = T$: $\sum_{t=1}^{T} O(B \cdot (n \cdot t \cdot d + 4d^2) + d \cdot V)$
     2. Combining all components, the overall time complexity for decoding the sequence L is: $O(B \cdot n \cdot d \cdot T^2 + B \cdot 4d^2 \cdot T + d \cdot V \cdot T)$

3. Based on your analysis of the question No 2., in which case the self-attention module dominate the time complexity? And in which case the feed-forward layer is dominant?

   自注意力模块的时间复杂度为$O(B \cdot n \cdot d \cdot T^2)$，前馈层的时间复杂度为$O(B \cdot 4d^2 \cdot T)$。因此当序列长度$T$较大时，自注意力模块的时间复杂度会占主导地位；而当序列长度较小时，前馈层的时间复杂度会占主导地位。

3. Discuss the influence of pre-training regarding the generation results, convergence speed, etc. Considering the experimental setup (the training task, data, pre-trained checkpoints, etc.), does the influence of pre-training meet your expectation?
   - 生成结果的影响
     - 能够提升生成结果的质量。通过在大规模语料上进行预训练，模型能够学习到丰富的语言特征、语法结构和上下文信息
     - 能够捕捉多样化的生成模式。在文本生成、翻译等任务中，预训练的模型能生成更多样化且相关性更强的输出
   - 收敛速度
     - 往往能更快收敛。因为模型已经学习了许多基本的语言模式，微调时只需少量的训练步骤即可适应特定任务
     - 预训练有较好的初始参数设置，使得模型在微调过程中能更好地适应特定任务

- 在本次实验中，相比于从头开始训练，预训练的模型并没有显著地提高BLEU分数，文本生成的准确率也比较接近。这可能是由于使用的数据集规模不够，预训练模型的知识无法完全迁移。