人工神经网络 · 第三次作业报告

吴明恒
2018011288

日期：2020 年 11 月 1 日

1 背景

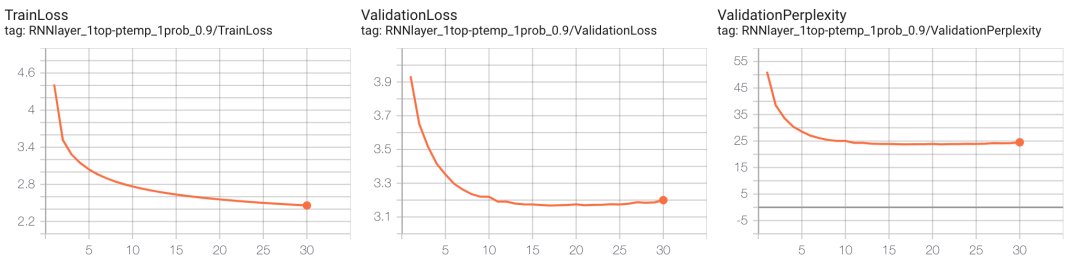
本次作业需要实现一个基于 RNN 的文本生成模型。

2 基础实验

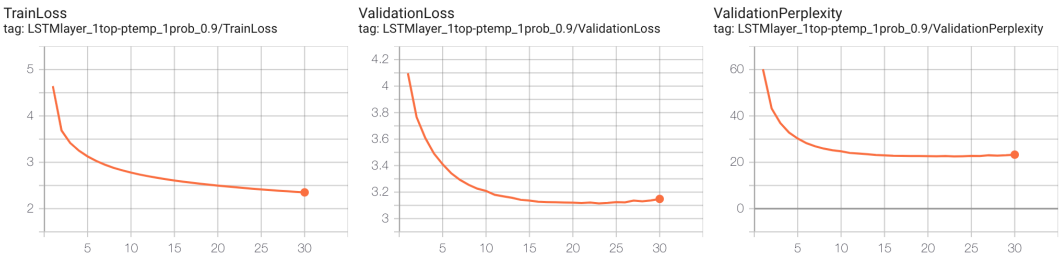
基本参数设置：`layer=1, decode_strategy=random, max_probability=0.9, num_epochs=30, temperature=1, learning_rate=0.001`

2.1 训练曲线

2.1.1 RNN cell



2.1.2 LSTM cell



2.1.3 GRU cell

2.2 测试结果

	Perplexity	Forward BLEU	Backward BLEU	Harmonic BLEU
RNN	20.56	0.342	0.318	0.330
LSTM	19.38	**0.381**	**0.336**	**0.357**
GRU	**18.77**	0.362	0.329	0.345

可以看出 LSTM 与 GRU 在相同的参数设置下，可以比单纯的 RNN cell 获得更好的效果，因为它们能更好地获取长程关系而不容易在序列过长时产生梯度消失。

GRU 相比 LSTM 结构更加简化了，但是效果却不输 LSTM，甚至在 Perplexity 评价中优于 LSTM。

另一方面，GRU 计算速度远优于 LSTM，因此在大多情况下 GRU 是一个更好的选择。

3 探究参数影响

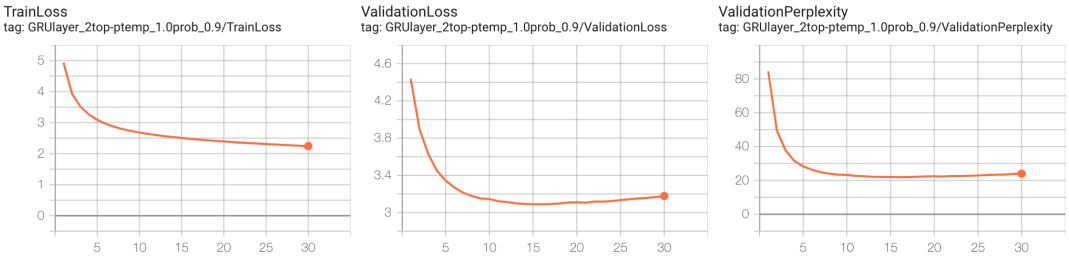
3.1 层数

将 GRU 网络的层数设为 2，与相同参数的单层网络进行对比，参数：`layer=2, decode_strategy=top-p, max_probability=0.9, num_epochs=30, temperature=1, learning_rate=0.001`

3.1.1 训练曲线

单层网络

双层网络



3.1.2 测试结果

	Perplexity	Forward BLEU	Backward BLEU	Harmonic BLEU
GRU(2-layer)	19.02	**0.366**	**0.333**	**0.349**
GRU(1-layer)	**18.77**	0.362	0.329	0.345

双层 GRU 在 BLEU 的表现稍好于单层 GRU，但在 Perplexity 上低于单层 GRU，总体来说表现相差不大。

3.2 解码策略

基础参数：`layer=1, num_epochs=30, learning_rate=0.001`

	Perplexity	Forward BLEU	Backward BLEU	Harmonic BLEU
GRU(top-p temp=1 prob=0.9)	18.77	0.362	0.329	0.345
GRU(ramdom temp=1)	18.76	0.304	0.317	0.310
GRU(top-p temp=0.8 prob=0.8)	18.72	**0.524**	**0.336**	**0.409**
GRU(random temp=0.8)	**18.71**	0.424	0.333	0.373
GRU(top-p temp=1 prob=0.8)	18.79	0.427	**0.336**	0.376

关于 random 与 top-p，相对 random，采用 top-p 在 perplexity 上表现基本一样，但在三项 BLEU 评价中 top-p 策略均会获得更佳的表现。top-p 可以在引入随机性的同时去除一些不太可能的选择，从而让句子更合理。

关于 temperature，取值 0.8 时相对取值 1 表现更优。引入 temperature 可以在 softmax 时微调概率分布，使概率大的词更容易被选择，从而增加句子生成的质量，但也要注意不应太小以保持选词的随机性。

对比的几个设置中，综合考虑最佳的参数为：`top-p, temperature=0.8, max_probability=0.8`

4 生成文本质量

4.1 top-p

1. A small plane flying through a yellow sky with the clouds sky .
2. A man that is laying on a laptop looking out in the kitchen .

3. A man riding a motorcycle on the street next to the road .
4. Two men are standing in a room in the kitchen .
5. A fire hydrant sitting on the side of a street .
6. A white plane is flying through a clear blue sky .
7. A jet is parked on the tarmac by a runway .
8. A little boy sitting on a bench next to a tree .
9. A giraffe standing on top of a field near some trees .
10. A cat sitting on the top of a wall near a door .

4.2 random

1. Two airplanes that are lined up in a city .
2. Line of motorcycles are parked in front of a building .
3. A cop with a surfboard on top of a blue tub .
4. A yellow fire hydrant on the side of a road .
5. A man on a bench next to water from the ocean .
6. A man wearing a red skirt and tie sitting on a motorcycle .
7. There is a small bathroom with a toilet with a sink .
8. A large white bowl filled with vegetables , muffin , and some it .
9. The giraffe is shown at the distance of trees .
10. A four tier of parked cars with a sign on a rainy day .

4.3 对句子的评价

在这十个句子中，观感上 random 生成的句子稍好一些，比如 random 的第 2 个句子，但这些句子都很容易出现语法错误，句意上也很少有通顺的，距人写的水平差距较大。

而评价指标为：

	Perplexity	Forward BLEU	Backward BLEU	Harmonic BLEU
GRU(top-p temp=0.8 prob=0.8)	18.72	**0.524**	**0.336**	**0.409**
GRU(random temp=0.8)	**18.71**	0.424	0.333	0.373

BLUE 评价分数反而 top-p 更高，评价生成的句子是否足够好确实很难有一个确定的标准，它必须兼顾原创性、规范性与通用性。

5 最终选择的模型

限于算力，我对使用 GRU cell 的模型尝试了以下的参数设置，结果如下：

Arguments	Perplexity	Forward BLEU	Backward BLEU	Harmonic BLEU
GRU layer_1 random temp_1 prob_0.9	18.76	0.304	0.317	0.310
GRU layer_1 random temp_0.8 prob_1	18.71	0.424	0.333	0.373
GRU layer_1 top-p temp_1 prob_0.9	18.77	0.362	0.329	0.345
GRU layer_1 top-p temp_1 prob_0.8	18.79	0.427	**0.336**	0.376
GRU layer_1 top-p temp_0.8 prob_0.9	18.81	0.480	0.335	0.395
GRU layer_1 top-p temp_0.8 prob_0.8	18.72	0.524	**0.336**	0.409
GRU layer_1 top-p temp_0.7 prob_0.8	**18.13**	0.561	0.324	**0.410**
GRU layer_2 top-p temp_1 prob_0.9	19.02	0.366	0.333	0.349
GRU layer_2 top-p temp_0.7 prob_0.8	19.23	**0.576**	0.319	**0.410**

综合考虑,我的最终模型参数设置为 `layer=1, decode_strategy=top-p, max_probability=0.8, num_epochs=30, temperature=0.7, learning_rate=0.001`

6 word embedding 的影响

将使用预训练词向量的 embedding 层改为直接生成词的 one-hot 向量，其他参数不变，实验结果如下：

	Perplexity	Forward BLEU	Backward BLEU	Harmonic BLEU
GRU one-hot	19.88	0.493	0.335	0.399
GRU embedding	**18.72**	**0.524**	**0.336**	**0.409**

可以看到采用词向量的表现比 one-hot 编码的句子表现要显著地好，尽管如此，one-hot 的表现依然比我预想的要好，词向量给句子生成带来了一些词义的信息，使训练更容易。

另一方面，one-hot 编码的长度为词表长度，即 3000 多，远比我们使用的词向量维度大（300），因此词向量也更易于计算。

7 隐藏层单元大小的影响

将使用 GRU cell 的模型中隐藏层大小改变，其他参数不变，实验结果如下：

	Perplexity	Forward BLEU	Backward BLEU	Harmonic BLEU
GRU hidden_32	21.71	0.499	0.320	0.390
GRU hidden_64	18.86	0.503	0.336	0.403
GRU hidden_128	17.49	**0.524**	**0.338**	**0.411**
GRU hidden_256	17.01	0.490	0.330	0.394
GRU hidden_512	**16.84**	0.515	0.331	0.403

从测试结果来看，hidden size 越大，Perplexity 越小，但 BLEU 没有这种单调性，当 hidden size 为 128 时表现最佳。

另一方面，当 hidden size 变大时，训练速度会显著下降。

8 思考题

1. Plot the loss value of one-layer RNN with 3 kinds of rnncells (i.e., RNNCell, GRUCell, LSTMCell) against every epoch during training (on both training parts and validation parts). Report the test results on 4 metrics (Perplexity, Forward BLEU, Backward BLEU, Harmonic BLEU). Compare and analyze the performance of 3 kinds of rnncells.

见基础实验部分。

2. (Optional, 0.5 point bonus. It is optional to report the result, but the code you submitted MUST implement the multilayer RNN.) Choose a kind of rnncell, plot the loss value of two-layer RNN against every epoch during training (on both training parts and validation parts). Report the test result on 4 metrics. Compare and analyze the performance of one- layer RNN and two-layer RNN.

见探究参数影响的层数部分。

3. Choose the best model and try different decoding strategies in inference (at least including, random sampling with temperature=1, random sampling with temperature=0.8, top-p sampling with p=0.8, top-p sampling with p=0.8 and temperature=0.8). Report the test results on 4 metrics. How the decoding strategies influence the generated results?

见探究参数影响的解码策略部分。

4. Read the sentences generated by the different decoding strategies above. Randomly choose 10 sentences for each strategies and list them in your report. Is there any grammar errors? Which strategies generate the best sentences? Discuss whether the 4 metrics (Perplexity, Forward BLEU, Backward BLEU, Harmonic BLEU) are consistent with your judgement?

见生成文本质量部分。

5. Describe your final network with the hyperparameters and decoding strategies. Report the result on 4 metrics, and submit output.txt with your report.

见最终选择的模型部分。

6. More explorations (e.g., discuss the effect of pretrained word vector, train a word-level language model) can be awarded with bonus scores. (At most 2 points. Total score for the homework is 15. It means that the maximum score for the homework is 17.)

见 word embedding 的影响、隐藏层单元大小的影响部分。

9 总结

本次实验实现了一个文本生成模型，产生英文短句，本实验对比了不同的参数设置与解码策略对实验效果的影响。