Artificial Neural Network HW3

刘泓尊　2018011446　计 84

Department of Computer Science, Tsinghua University

2020 年 10 月 24 日

目录

1 Network Structure and Hyperparameters

参数权重初始化均为默认，实现了 1 层和 2 层的模型。

超参：$embed\_units = 300, hidden\_units = 300, batch\_size = 32, lr = 1e^{-3}, weight\_decay = 0.0001$

Decoding 方式：默认为 random, temperature=1.0.

额外的**改动**：

```
add '--cell' option in ArgumentParser, options are ['rnn', 'lstm', 'gru']
weight_decay = 0.0001
add dropout=0.2 after embedding layer, add dropout=0.5 bettween rnn layers(
    only when layers > 1).
```

您可以在/codes 目录下运行如下命令复现我的结果，更多配置请见/codes/readme.md.

```
python main.py --num_epochs 100 --batch_size 32 --layers 1 --units 300 --
    decode_strategy random --cell gru
```

在/codes 目录下使用 tensorboard 可视化训练结果：

```
tensorboard --logdir=train --port=6006
```

2 Experiment Result

实验环境: NVIDIA TITAN Xp, CUDA Version: 10.1

2.1 Overview

下面是不同模型和参数配置的结果。

使用 random-decoding, temperature = 0.1 的不同模型对比：

Model/Best Metrics	Test PPL	Forward BLEU	Backword BLEU	Harmonic BLEU	Time(s)/epoch
RNN-1L	17.62	0.309	0.319	0.314	**16.249**
LSTM-1L	16.79	0.292	0.307	0.299	25.705
GRU-1L	16.22	0.309	0.313	0.311	25.313
LSTM-2L	16.29	**0.319**	0.323	0.321	42.127
GRU-2L	**15.92**	0.317	**0.332**	**0.324**	42.034

3 Analysis

3.1 1 层 RNN 的 Loss 曲线，三种 Cell 的性能对比

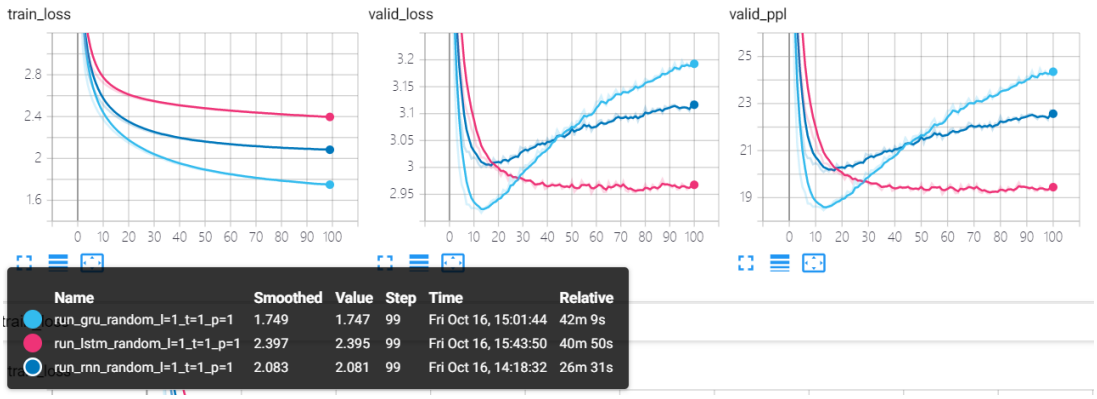

图 1: 1 层 RNN, GRU, LSTM 的 Loss 曲线对比

收敛速度上：GRU > RNN > LSTM. 一般来说 RNN 存在梯度消失问题，所以后期收敛速度不及 GRU。LSTM 的参数量比较大，所以收敛速度最慢。

性能上，从图1和表2.1来看，GRU > LSTM > RNN. 从不同 RNNCell 的构造来看，RNN 没有长程序列建模能力，只有短期记忆，并且存在梯度消失，所以生成的句子质量可能不如 LSTM 和 GRU. 但是本实验数据集上的句子都比较短，RNN 的劣势并没有被充分体现。

LSTM 和 GRU 在性能上并无明显差异，但是 GRU 存在较严重的过拟合问题，LSTM 的过拟合程度较低。在大数据集上，参数更多的 LSTM 表现一般更好。

3.2 2 层 RNN 的 Loss 曲线，1 层与 2 层 RNN 的性能对比

从图中可以看到，2 层的 RNN 比 1 层 RNN 收敛速度慢，这符合预期。而且 2 层 RNN 的过拟合程度比较低。

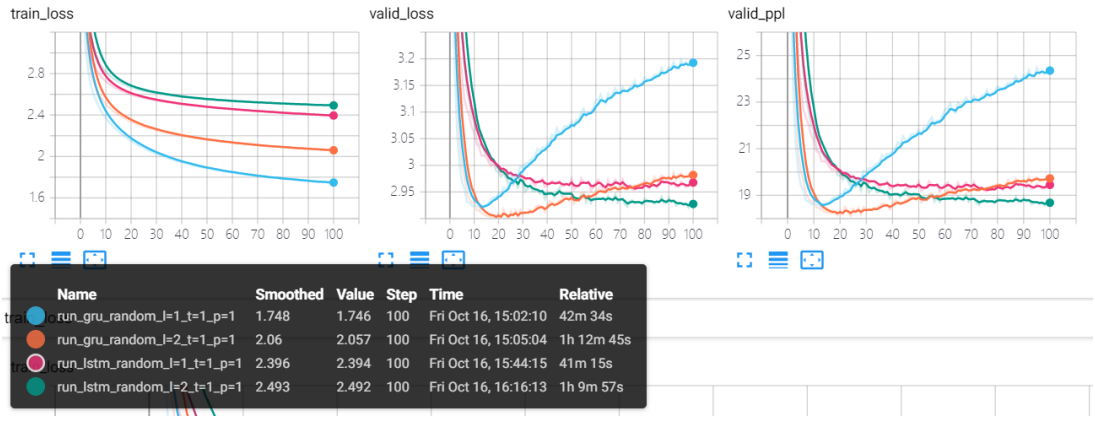

图 2: 1 层与 2 层 GRU, LSTM 的对比

从性能上看，2 层模型在各个指标上优于 1 层模型。增加层数可以提升模型的建模能力，模型自然性能更高，当然训练时间会增加 1 倍左右。因为本次实验是序列生成任务，所以采用双向 RNN 可能不太方便，只适合用 Stacked RNN。

3.3 2 层 GRU 不同 Decoding Strategy 的性能对比

使用 GRU-2L，不同 decoding strategy 的对比：

Model/Best Metrics	Test PPL	Forward BLEU	Backword BLEU	Harmonic BLEU
random(t=1.0)	15.92	0.317	0.332	0.324
random(t=0.8)	15.90	0.458	**0.351**	0.397
top-p(p=1.0, t=0.8)	15.94	0.445	0.334	0.382
top-p(p=0.8, t=0.8)	**15.85**	**0.571**	0.331	**0.419**

从表中可以看到，top-p decoding 优于 random decoding, 并且 temperatrue 适当降低会改善句子质量。

temperatrue 控制了概率分布的平滑程度，t 越小，分布越尖锐，增加了高概率单词的可能性并降低了低概率单词的可能性。可以推测 t 在较大时，sampling 过程引入的扰动会降低句子的质量。

对于 top-p decoding, p = 1.0 的时候退化为 random decoding. p 降低时，可选单词集合缩小。top-p decoding 比 top-k decoding 更加灵活，可以根据概率分布调整可选单词集合。

上述两种 decoding 策略都比 beam search 更加灵活，因为加入了 sample 过程来增加随机性。

3.4 2 层 GRU 不同 Decoding Strategy 生成的句子对比

random(t=1.0)

```
In an vans laying down to the bridge at the airport .
Two people walking down the street near a rolling .
An airplane in the sky with a wheels leading door .
A man sit on a runway on a pier next to a tree .
A plane is parked on a runway , and the jet are leaving .
A lady with a helmet is sitting on a bench next to a glass cloth commode .
A man standing in front of a red fire hydrant .
A bathroom that has a sink , toilet , and window .
```

```
An older woman sitting on a bench holding a toy .
A small plane is shown with lots of the wing removed .
```

random(t=0.8) 或 top-p(p=1.0, t=0.8)

```
Two buses in the middle of a city street .
A person is standing in a very cluttered kitchen with a cluttered kitchen
   table .
Black and white photograph of a man sitting on a bench .
An old , rusty , blue and blue bus driving down a street .
Two people are sitting on a bench looking for the river .
Three white sheep eating from a trough in a open field .
A fire hydrant is painted a blue and green leaves .
Some people are sitting in front of a business bus and a motorcycle .
A city street with cars and a bus in the background .
A traffic light with a red painted red sign sign and a building .
```

top-p(p=0.8, t=0.8)

```
The man is walking on the beach by the water .
A group of giraffes that are standing in the dirt .
A herd of sheep grazing in a field with green grass .
The front end of a plane that is being loaded .
The two cats are standing in front of the door .
The man is sitting on a bench near a large building .
A group of birds flying in the sky near a lake .
A black cat laying down in a bathroom with two monitors .
A group of people sitting on top of a wooden bench .
The back of a plane with it ' s wing wheels down .
```

使用 random decoding 时，可以看到生成的句子单词较为多样化，尤其是第一个单词，生成的句子长度也较为多样。而 top-p decoding 句子模式较为单一，句子长度也较为整齐，这是因为一定程度上缩小了采样空间。

temperature 变小时，分布更尖锐，提高了高频词的比重，所以句子的用词更为单一，高频词很多。

从句子质量上看，random(t=1.0) 生成句子语法错误较多，比如 “In an vans laying down...” 和 “Two people walking down the street near a rolling .”，大多数句子不太符合常识，只是在某一个窗口内是通顺的。当 temperature 降低时，语法和常识错误会减少。

对于 top-p decoding，常识错误较少，语法错误也几乎没有，可以看到它缩小了可选单词范围，能给句子带来更少的扰动，同时单词多样性会降低。综合来看，top-p decoding(t=0.8, p=0.8) 的句子质量是最优的。

上面的分析和 BLEU 指标是一致的，在 t 降低时，句子质量会提升。且 top-p decoding 优于 random decoding.

3.5 Pretrained WordVec v.s. Learnable WordVec

下图是 LSTM, GRU 分别使用 Pretrained WordVec 或 Learned WordVec 得到训练曲线和自动指标。

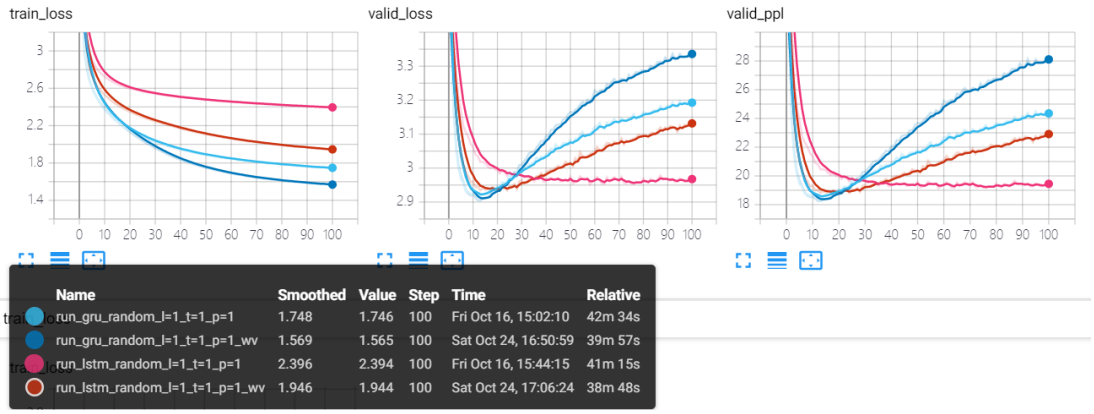

图 3: wv 后缀的为使用 Learnable WordVec，其他为使用 Pretrained WordVec

从图中可以看到，使用 Learnable WordVec 更容易出现过拟合，但是最佳性能却更好。学习的词向量能和后面的 RNN 层实现更好的配合。

Model/Best Metrics	Test PPL	Forward BLEU	Backword BLEU	Harmonic BLEU
GRU-1L(Pretrained)	16.22	0.309	0.313	0.311
LSTM-1L(Pretrained)	16.79	0.292	0.307	0.299
GRU-1L(Learned)	**16.05**	**0.320**	**0.330**	**0.325**
LSTM-1L(Learned)	16.46	0.301	0.323	0.311

4 Summary

本次实验我在 top-p decoding 上花了一些时间，感觉优雅地实现它需要深入地思考。此外我开始使用 tensorboard 来对训练指标进行实时观测，可以及时发现错误，而且很方便展示结果。

感谢老师和助教的悉心指导！