

- 论文：了解bert和transformer结构，语言模型
- 实验：
  - 参考代码：[LSTM and QRNN Language Model Toolkit](#)，其中包含多个任务的LSTM模型，训练源码
  - 数据集：
    - Hutter Prize dataset (enwik8):
      - A byte-level dataset
      - 100Mb of a Wikipedia XML dump
      - 205 unique tokens.

这里不看

- word level language models over the Penn Treebank (PTB), [WikiText-2](#) (WT2), and [WikiText-103](#) (WT103) datasets
- character level language models over the Penn Treebank (PTBC) and Hutter Prize dataset (enwik8)

- 任务：
  - word level language model:  
通过Word2Vec构建词典，将word矢量化，每个word表示为一个vector  
**对于大型数据库，要构建词典，得到的tokens太多**  
**含有单词语义，建模和训练简单，使用LSTM可以完成**
  - **character level language model:**  
用数字表示每一个字符，其实相当于用0,1表示每个字符，简单粗暴。

```
ww.auburn.edu/academic/liberal_arts/foreign/russian/icons/ Russian
Icons from 12th to 18th century]
119 119 46 97 117 98 117 114 110 46 101 100 117 47 97 99 97 100 101
109 105 99 47 108 105 98 101 114 97 108 95 97 114 116 115 47 102
111 114 101 105 103 110 47 114 117 115 115 105 97 110 47 105 99 111
110 115 47 32 82 117 115 115 105 97 110 32 73 99 111 110 115 32 102
114 111 109 32 49 50 116 104 32 116 111 32 49 56 116 104 32 99 101
110 116 117 114 121 93
```

**忽略了语法和单词语义，神经网络建模困难**

**之前LSTM一般不能完成，[该论文](#)通过调参实现了并且做了进一步优化的QRNN**

参考：<https://blog.csdn.net/CoderPai/article/details/80216301>

论文：An Analysis of Neural Language Modeling at Multiple Scales [2018]

- 模型：

Character PTB		enwik8	WikiText-103
RNN Cell	LSTM	LSTM	QRNN
Layers	3	3	4
RNN hidden size	1000	1840	2500
AR/TAR	0/0	0/0	0/0
Dropout (e/h/i/o)	0/0.25/0.1/0.1	0/0.01/0.01/0.4	0/0.1/0.1/0.1
Weight drop	0.5	0.2	0
Weight decay	1.2e-6	1.2e-6	0
BPTT length	150	200	140
Batch size	128	128	60
Input embedding size	128	400	400
Learning rate	0.002	0.001	0.001
Epochs	500	50	14
LR reduction ( $\frac{lr}{10}$ )	[300, 400]	[25, 35]	[12]
Total parameters	13.8M	47M	151M
Training time (hours)	8	47	12

Table 6. Hyper-parameters for word- and character-level language modeling experiments. Training time is for all noted epochs on an NVIDIA Volta. Dropout refers to embedding, (RNN) hidden, input, and output.

- 只有训练代码，没有公布模型，所以需要自己从头训练，目前在做enwik8，**训练进度33epochs/50epochs，大概进行了33hours了，今天晚上应该能训练完成。**
- 作者说调参非常重要，随机试验了200多个模型得到的参数，所以我直接按照作者参数来。
- 基于这个模型的column combining代码正在写
- 和原来模型的对比：

现在	原来
<pre>[WeightDrop(   (module): LSTM(400, 1840) ), WeightDrop(   (module): LSTM(1840, 1840) ), WeightDrop(   (module): LSTM(1840, 400) )]</pre>	<pre>[ (   (module): LSTM(28, 64) )]</pre>
47MB	0.024MB

- 工作计划：

	实验	目的
1	继续enwik8数据实验	还原了论文结果
2	修改该模型为column combine训练（正在修改中）	尽量减少精度损失
3	允许conflict实验（刘诗玮想法）	尽量在上面的基础上提高效果

- 考虑到enwik8数据集训练时间过长（47hours）问题，会考虑换成Character PTB
- 在3的实验过程中，column combine的规则需要结合硬件制定一下
- 1,2,3实验进行的同时了解一下量化方法，3个实验结束后尝试量化实验