

《Language Modeling with Gated Convolutional Networks》阅读笔记

 bear8133
熊大

22 人赞了该文章

转载请注明出处：[西土城的搬砖日常](#)

原文链接：[Language Modeling with Gated Convolutional Networks](#)

问题介绍：目前语言模型主要基于RNN，这篇文章提出了一种新颖的语言模型，仿照LSTM中的门限机制，利用多层的CNN结构，每层CNN都加上一个输出门限。文中提出的GLU模型在两个常用数据集上的测试效果超过了目前循环模型，并且速度更快。

主要方法

统计学语言模型，通过条件概率的形式来估计此序列的分布：

$$P(W_0, \dots, W_N) = P(W_0) \prod_{i=1}^N P(W_i | W_0, \dots, W_{i-1})$$

，神经网络通过：

$H = [h_0, \dots, h_N]$ 为词序列 W_0, \dots, W_N 各个词的向量表示，其中 $h_i = f(h_{i-1}, w_{i-1})$ ，给词序列词之间的依赖关系建模。为了缓解梯度消失的问题，LSTM引入了门限机制：输入门，遗忘门和输出门。由于循环神经网络中每个时刻状态都不仅和输入相关，也和前一个时刻状态相关。所以在序列中无法并行化处理。这篇论文提出了一种新颖的模型：门限机制的卷积模型，文中具体提出了GTU和GLU两种模型，并在实验阶段作了比较。这两个模型整体类似，主要是激活函数不一样。图一是模型的结构图：

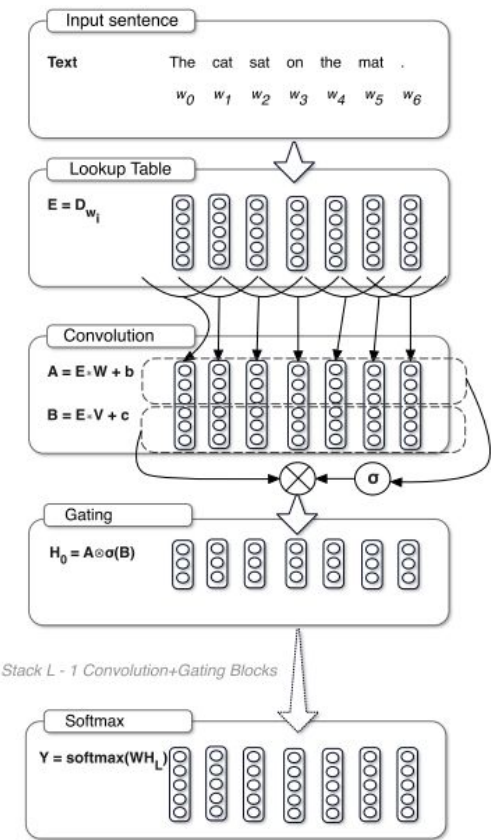


Figure 1. Architecture of the gated convolutional network for language modeling.

图一

GLU模型： $h_l(X) = (X * W + b) \otimes$

赞同 22

21 条评论

分享

收藏

...

《EFFICIENT SL

GTU模型: $h_l(X) = \tanh(X * W + b) \otimes \delta(X * V + c)$

从公式中可以看出两个模型不同之处在激活函数不同，GLU的几乎函数是线性的，GTU的激活函数是tanh，非线性的。作者稍后从梯度的角度分析了GLU比GTU更优。

公式中的X是上一层的输出向量（或者初始输入的词序列向量）， $X \in R^{N \times m}$ ， $W \in R^{k \times m \times n}$ ， $V \in R^{k \times m \times n}$ ，N为词序列的长度，m的词向量的维度，k为卷积核的大小，b，c为偏置。最后每层的输出每个词的向量表示 $H = h_L \circ \dots \circ h_0(E)$ ，(E为输入，L为模型的层数)。

从门限机制比较GLU和GTU两个模型：

LSTM通过引入门限机制减缓梯度消失问题，论文中通过在输出时加入输出门引入门限机制。GTU模型梯度：

$$\begin{aligned} \nabla[\tanh(X) \otimes \sigma(X)] &= \tanh'(X) \nabla X \otimes \sigma(X) \\ &\quad + \sigma'(X) \nabla X \otimes \tanh(X). \end{aligned}$$

梯度中相加的两个部分有 $\tanh'(X)$ 和 $\sigma'(X)$ 衰减项，而GLU模型的梯度：

$$\nabla[X \otimes \sigma(X)] = \nabla X \otimes \sigma(X) + X \otimes \sigma'(X) \nabla X$$

第一项没有衰减项，从这个角度分析作者认为GLU比GTU更优，并在实验中将两者进行比较。

实验

一 数据集：Google Billion Word和WikiText-103

二 训练：借鉴了Nesterov' s momentum中梯度下降的方法，借鉴了adaptive softmax中的softmax，同时借鉴了梯度截断的方法。

三 实验结果：

Model	Test PPL	Hardware
Sigmoid-RNN-2048 (Ji et al., 2015)	68.3	1 CPU
Interpolated KN 5-Gram (Chelba et al., 2013)	67.6	100 CPUs
Sparse Non-Negative Matrix LM (Shazeer et al., 2014)	52.9	-
RNN-1024 + MaxEnt 9 Gram Features (Chelba et al., 2013)	51.3	24 GPUs
LSTM-2048-512 (Jozefowicz et al., 2016)	43.7	32 GPUs
2-layer LSTM-8192-1024 (Jozefowicz et al., 2016)	30.6	32 GPUs
LSTM-2048 (Grave et al., 2016a)	43.9	1 GPU
2-layer LSTM-2048 (Grave et al., 2016a)	39.8	1 GPU
GCNN-13	38.1	1 GPU

Table 1. Results on the Google Billion Word test set.

上面的结果看到，论文中提出的模型GCNN-13超过了之前所有基于循环神经网络的模型，GCNN-13中13指使用了13层卷积。

Model	Test PPL
LSTM-1024 (Grave et al., 2016b)	48.7
GCNN-8	44.9

Table 2. Results on the WikiText-103 dataset.

	Throughput		Responsiveness
	(CPU)	(GPU)	(GPU)
LSTM-2048	169	45,622	2,282
GCNN-22	179	45,878	45,878

Table 3. Processing speed in tokens/s at test time for an LSTM with 2048 units and GCNN with 22 layers achieving 43.9 and 43.8 perplexity, respectively on Google Billion Word. The GCNN improves the responsiveness by 20 times while maintaining high throughput.

上面表中比较了LSTM-2048和GCNN的训练处理速度，Responsiveness表示处理单个句子的速度。

	Parameters	FLOPs/token
LSTM-2048	289M	19M
GCNN-22	185M	14M

Table 4. Number of parameters and FLOPs for the models of Figure 3. FLOPs exclude the operations required by the softmax layer which are identical.

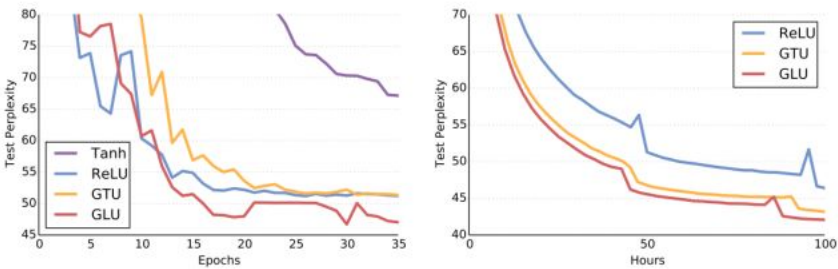
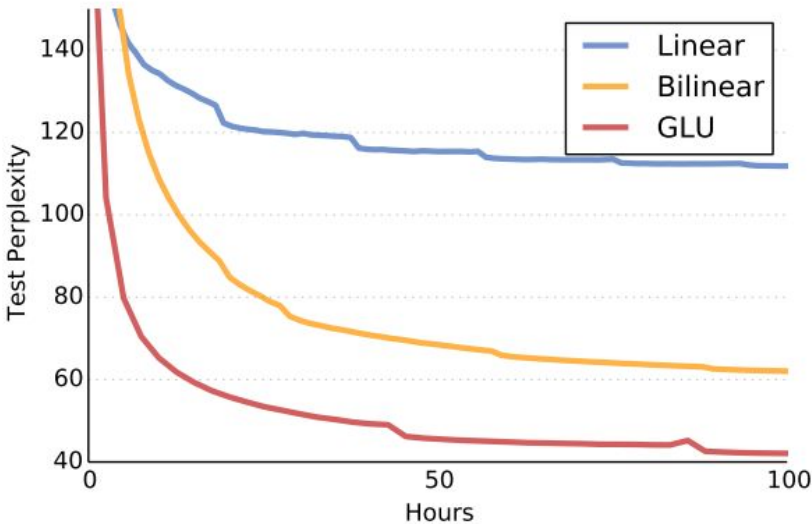


Figure 2. Learning curves on WikiText-103 (left) and Google Billion Word (right) for models with different activation mechanisms. Models with gated linear units (GLU) converge faster and to a lower perplexity.

Tanh是GTU去掉输出门部分后的模型，将其和GTU比较研究门限影响和贡献。从实验中对比可以看到GTU取得了最优的结果。



Linear代表将GLU中的输出门去掉后的模型，Bilinear表示将GLU中的输出门部分替换成另外一个线性部分的模型，将三者进行比较，说明门限的影响。

赞同 22 21 条评论 分享 收藏

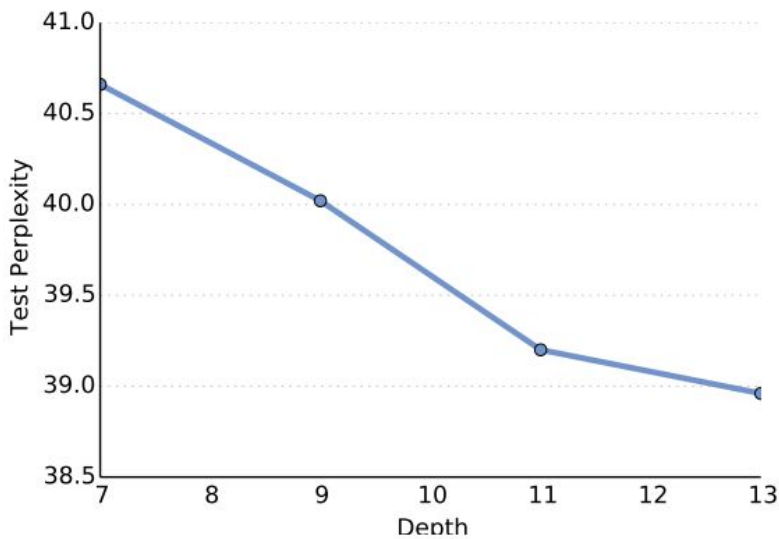


Figure 4. Impact of network depth on test perplexity for Google Billion Word. Deeper models perform better.

简评

这篇文章提出了基于卷积神经网络和门限机制的深度学习模型，将其运用到语言模型中，取得了比循环神经网络模型好的效果，同时由于卷积神经网络局部性的特点使得其可以在词序列中进行并行训练，提高了处理的速度，同时引入门限机制，减缓梯度消失，加快了模型的收敛速度。通过叠加多层来学习词序列的前后依赖关系，使得其在长文本WikiText-103语言模型的学习中也取得不错的效果。

编辑于 2017-01-09

自然语言处理 深度学习（Deep Learning） 卷积神经网络（CNN）

文章被以下专栏收录



西土城的搬砖日常
机器学习，深度学习等各种人工智能分享


进入专栏

推荐阅读

《Semi-supervised Multitask Learning for...

来源：ACL 2017 原文：Semi-supervised Multitask Learning for Sequence LabelingIntroduction 序列标注任务（Sequence Labeling）在自然语言处理中有广泛的应用，包括...

yifannnn



DeepLesion系列论文之 JMI2018-阅读笔记

livew... 发表于每周一篇机...

《Effective LSTMs for Target-Dependent...

转载请注明出处：西土城的搬砖 原文链接：Effective LSTM Target-Dependent Sentiment Classification 来源：COLIN 问题：target-dependent 情析一、target-dependent情 simpl... 发表于西土城