学界 | 神经序列模型的动态评估

2017-11-17 机器海岸线

选自 arXiv

作者：**Ben Krause, Emmanuel Kahembwe, Iain Murray, & Steve Renals** 等

机器海岸线编译

参与：方建勇

DYNAMIC EVALUATION OF NEURAL SEQUENCE MODELS

Ben Krause, Emmanuel Kahembwe, Iain Murray, & Steve Renals
School of Informatics, University of Edinburgh
Edinburgh, Scotland, UK
ben.krause,e.kahembwe,i.murray,s.renals@ed.ac.uk

论文链接：https://arxiv.org/pdf/1709.07432

摘要：我们提出使用动态评估来改进神经序列模型的方法。 模型通过基于梯度下降的机制来适应近代历史，使它们将更高的概率分配给重新出现的连续模式。 动态评估在我们的比较中胜过现有的适应方法。动态评估改善了 Penn Treebank 和 WikiText-2 数据集中最先进的字级困惑，分别达到了 51.1 和 44.3，以及 text8 和 Hutter Prize 上最先进的字符级交叉熵数据集分别为 1.19 位/字符和 1.08 位/字符。

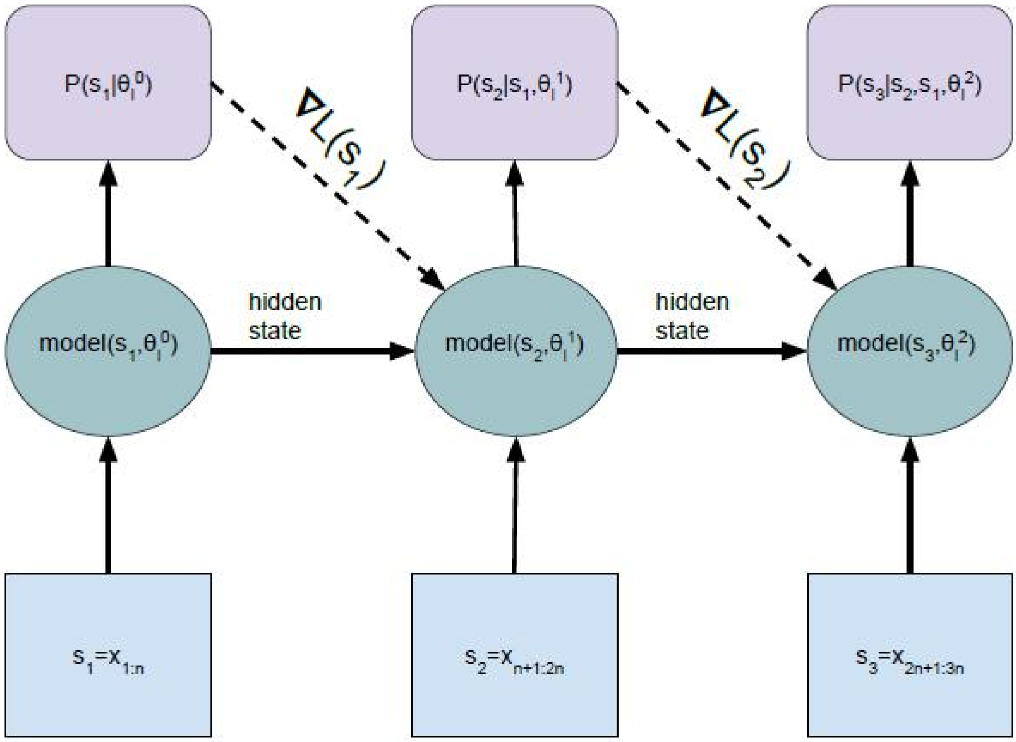

Figure 1: Illustration of dynamic evaluation. The model evaluates the probability of sequence segments s_i. The gradient $\nabla\mathcal{L}(s_i)$ with respect to the log probability of s_i is used to update the model parameters θ_l^{i-1} to θ_l^i before the model progresses to the next sequence segment. Dashed edges are what distinguish dynamic evaluation from static (normal) evaluation.

model	parameters	valid	test
RNN+LDA+kN-5+cache (Mikolov & Zweig, 2012)			92.0
CharCNN (Kim et al., 2016)	19M		78.9
LSTM (Zaremba et al., 2014)	66M	82.2	78.4
Variational LSTM (Gal & Ghahramani, 2016)	66M		73.4
Pointer sentinel-LSTM (Merity et al., 2017b)	21M	72.4	70.9
Variational LSTM + augmented loss (Inan et al., 2017)	51M	71.1	68.5
Variational RHN (Zilly et al., 2017)	23M	67.9	65.4
NAS cell (Zoph & Le, 2017)	54M		62.4
Variational LSTM + gradual learning (Aharoni et al., 2017)	105M		61.7
LSTM + BB tuning (Melis et al., 2017)	24M	60.9	58.3
LSTM (Grave et al., 2017)		86.9	82.3
LSTM + neural cache (Grave et al., 2017)		74.6	72.1
LSTM (ours)	**20M**	**88.0**	**85.6**
LSTM + traditional dynamic eval (sgd, bptt=1)	**20M**	**78.6**	**76.2**
LSTM + dynamic eval (sgd, bptt=5)	**20M**	**78.0**	**75.6**
LSTM + dynamic eval (sgd, bptt=5, global prior)	**20M**	**77.4**	**74.8**
LSTM + dynamic eval (RMS, bptt=5, global prior)	**20M**	**74.3**	**72.2**
LSTM + dynamic eval (RMS, bptt=5, RMS global prior)	**20M**	**73.5**	**71.7**
AWD-LSTM (Merity et al., 2017a)	24M	60.0	57.3
AWD-LSTM +neural cache (Merity et al., 2017a)	24M	53.9	52.8
AWD-LSTM (ours)	**24M**	**59.8**	**57.7**
AWD-LSTM + dynamic eval	**24M**	**51.6**	**51.1**

表 1：Penn Treebank 复杂度。 bptt 是指序列段长度。

model	parameters	valid	test
Byte mLSTM (Krause et al., 2016)	46M	92.8	88.8
Variational LSTM (Inan et al., 2017)	28M	91.5	87.0
Pointer sentinel-LSTM (Merity et al., 2017b)		84.8	80.8
LSTM + BB tuning (Melis et al., 2017)	24M	69.1	65.9
LSTM (Grave et al., 2017)		104.2	99.3
LSTM + neural cache (Grave et al., 2017)		72.1	68.9
LSTM (ours)	**50M**	**109.1**	**103.4**
LSTM + dynamic eval	**50M**	**63.7**	**59.8**
AWD-LSTM (Merity et al., 2017a)	33M	68.6	65.8
AWD-LSTM + neural cache (Merity et al., 2017a)	33M	53.8	52.0
AWD-LSTM (ours)	**33M**	**68.9**	**66.1**
AWD-LSTM + dynamic eval	**33M**	**46.4**	**44.3**

表 2：WikiText-2 的复杂度。

model	parameters	test
Stacked LSTM (Graves, 2013)	21M	1.67
Stacked LSTM + traditional dynamic eval (Graves, 2013)	21M	1.33
Multiplicative integration LSTM (Wu et al., 2016)	17M	1.44
HyperLSTM (Ha et al., 2017)	27M	1.34
Hierarchical multiscale LSTM (Chung et al., 2017)		1.32
Bytenet decoder (Kalchbrenner et al., 2016)		1.31
LSTM + BB tuning (Melis et al., 2017)	46M	1.30
Recurrent highway networks (Zilly et al., 2017)	46M	1.27
Fast-slow LSTM (Mujika et al., 2017)	47M	1.25
mLSTM (Krause et al., 2016)	46M	1.24
mLSTM + sparse dynamic eval ($d = 250k$)	**46M**	**1.13**
mLSTM + dynamic eval	**46M**	**1.08**

表 3：Hutter Prize 测试集错误位/字符。。

model	parameters	test
Multiplicative RNN (Mikolov et al., 2012)	5M	1.54
Multiplicative integration LSTM (Wu et al., 2016)	4M	1.44
LSTM (Cooijmans et al., 2017)		1.43
Batch normalised LSTM (Cooijmans et al., 2017)		1.36
Hierarchical multiscale LSTM (Chung et al., 2017)		1.29
Recurrent highway networks (Zilly et al., 2017)	45M	1.27
mLSTM (Krause et al., 2016)	45M	1.27
mLSTM + dynamic eval	**45M**	**1.19**

表 4：text8 测试设置错误位/字符。

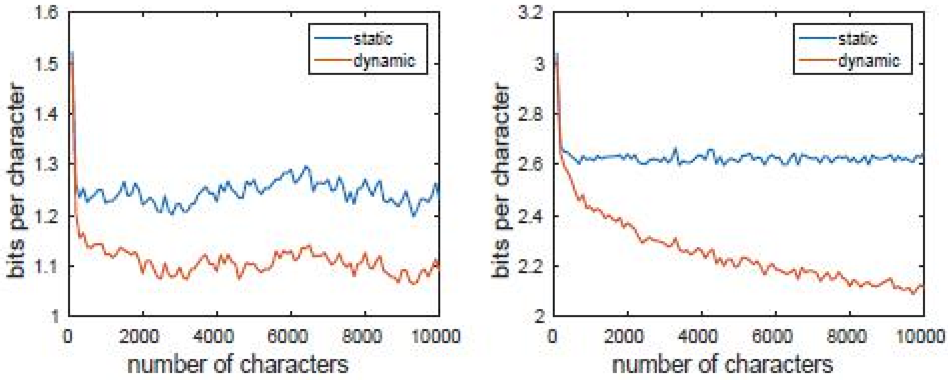

图 1：动态评估和静态评估的比特/字符的平均损失与处理字符的数量相关；来自 Hutter Prize 测试集（左）和欧洲议会数据集（西班牙语（右））的序列，平均每个测试 500 次以上。 每个数据点处的损失在长度为 100 的序列片段上平均，而不是累积的。 请注意两个图中不同的 y 轴比例。。

本文为机器海岸线编译，转载请联系 fangjianyong@zuaa.zju.edu.cn 获得授权。

✄---