学界 | LSTM 网络的分解技巧

2017-11-17 机器海岸线

选自 arXiv

作者：Oleksii Kuchaiev,Boris Ginsburg 等

机器海岸线编译

参与：方建勇

FACTORIZATION TRICKS FOR LSTM NETWORKS

Oleksii Kuchaiev
NVIDIA
okuchaiev@nvidia.com

Boris Ginsburg
NVIDIA
bginsburg@nvidia.com

论文链接：https://arxiv.org/pdf/1703.10722

摘要：我们提出了两个简单的方法来减少参数数量，加速大型长时间短期记忆（LSTM）网络的训练：第一个是将 LSTM 矩阵的“由设计矩阵分解”成两个较小矩阵的乘积， 第二个是将 LSTM 矩阵，其输入和状态划分为独立的组。 这两种方法都使我们能够更快地训练大型 LSTM 网络，以应对最先进的困惑。 在十亿字基准测试中，我们将单一模型的复杂度降低到 23.36。

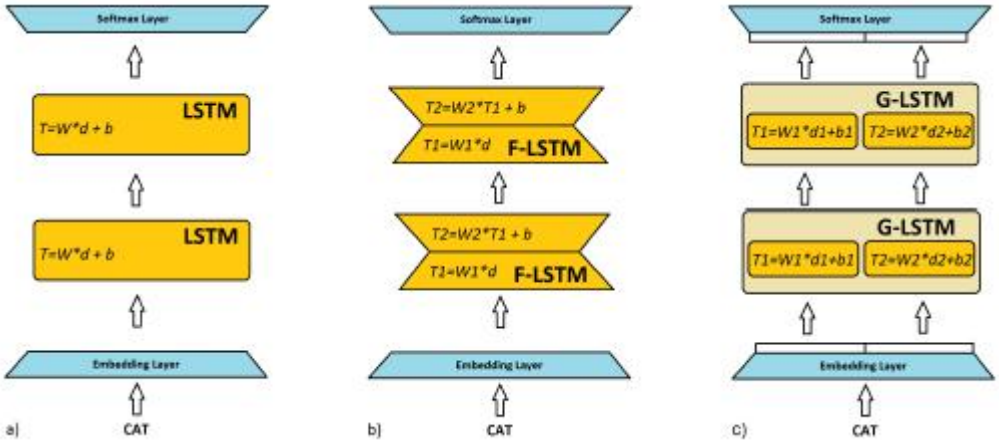

图 1：语言模型使用：（a）2 个常规 LSTM 层，（b）2 个 F-LSTM 层，以及（c）每层 2 个 G-LSTM 层。 单元格内的方程式显示在每个时间步骤中由这些单元格计算的仿射变换的类型。

Model	Perplexity	Step	Num of RNN parameters	Words/sec
BIGLSTM baseline	31.001	584.6K	151,060,480	20.3K
BIG F-LSTM F512	**28.11**	1.217M	52,494,336	42.9K
BIG G-LSTM G-4	**28.17**	1.128M	50,397,184	41.1K
BIG G-LSTM G-16	34.789	850.4K	25,231,360	41.7K

表 1：使用一台 DGX-1 培训 1 周后的十亿字基准评估结果。

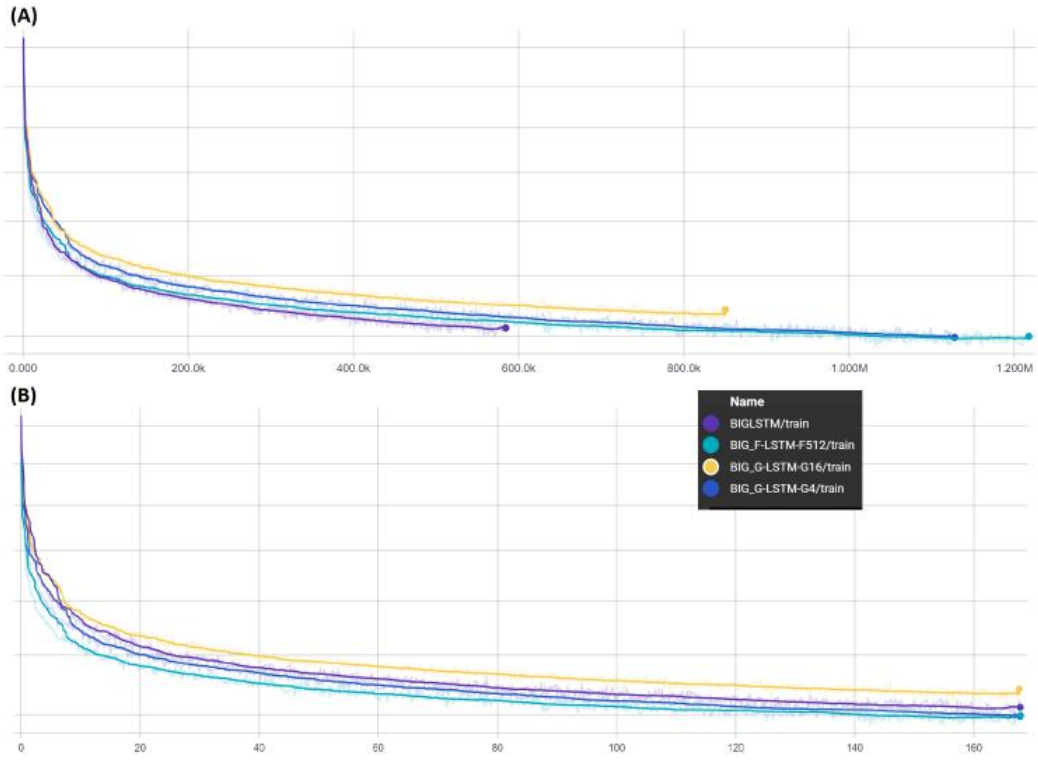

图 2：Y 轴：（A）和（B） - 训练损失对数标度相同，X 轴：（A） - 步骤或小批量计数，（B） - 小时。 BIGLSTM 基线，BIG G-LSTM-G4，BIG G-LSTM-G16 和 BIG F-LSTM-F512 均训练整整一周。 清楚地看到，在相同的步数下，具有更多参数的模型获胜。 另一方面，分解模型可以在给定的时间内做更多的迭代，因此在相同的时间内得到更好的结果。 （（A）和（B）的 X 轴的全部范围是 1 周）。

✂----------