法律声明

- 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

- 课程详情请咨询
  - 微信公众号：小象
  - 新浪微博：ChinaHadoop

第二节课：RNN语言模型

- RNN语言模型理论基础
- RNN语言模型实践
- 从RNN到LSTM
  - LSTM与长期记忆
  - 两句话解释BPTT

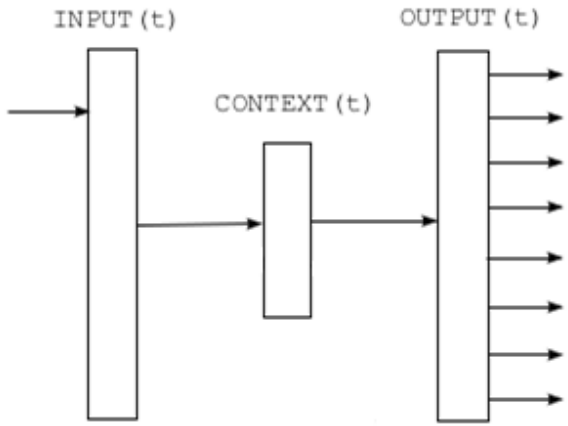
参考文献

- RNN语言模型理论基础
  - Recurrent neural network based language model. (Interspeech 2010, Tomas Mikolov et al.)
  - Unreasonable Effectiveness of Recurrent Neural Network
- RNN语言模型实践
  - https://github.com/sherjilozair/char-rnn-tensorflow, 教科书级别的代码
- 从RNN到LSTM
  - Understanding LSTM Networks –colah's blog
  - Deep Learning, Ian Goodfellow, Yoshua Bengio, Aaron Courville

RNN语言模型

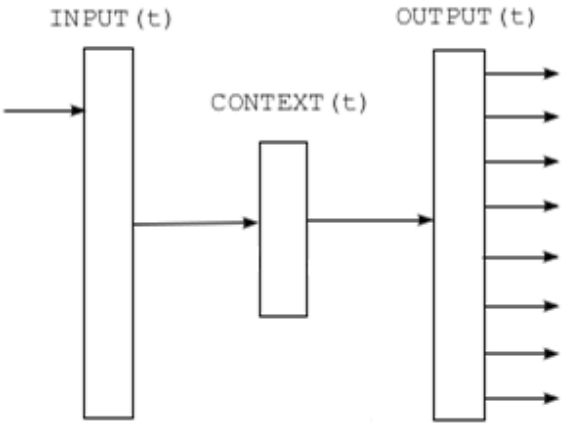
RNNLM

CBOW/skip-gram的局限



- CBOW/skip-gram 使用固定长度的语境.
  - 语境的长度需要人工调节
  - 只能利用固定长度的语境中的信息，不能利用更长范围内的词语之间的关联.

我们引入记忆



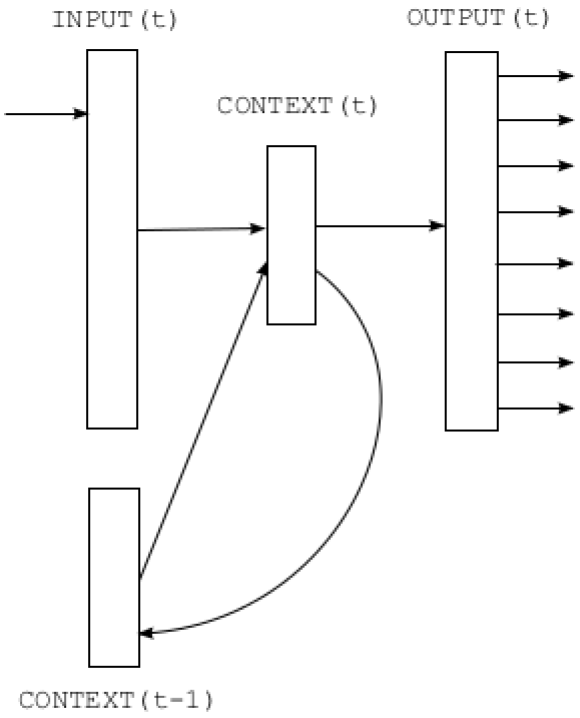


Figure 1: *Simple recurrent neural network.*

RNN语言模型细节

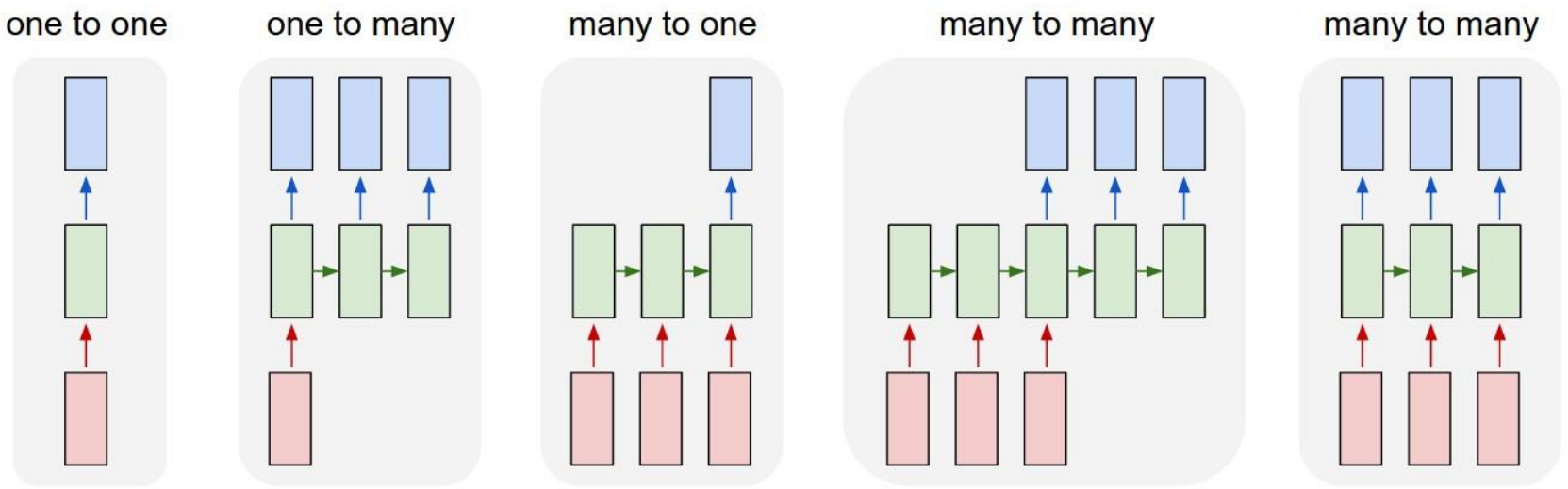
一个RNN cell
是一个经典的神经网络

RNN语言模型细节

RNN语言模型细节

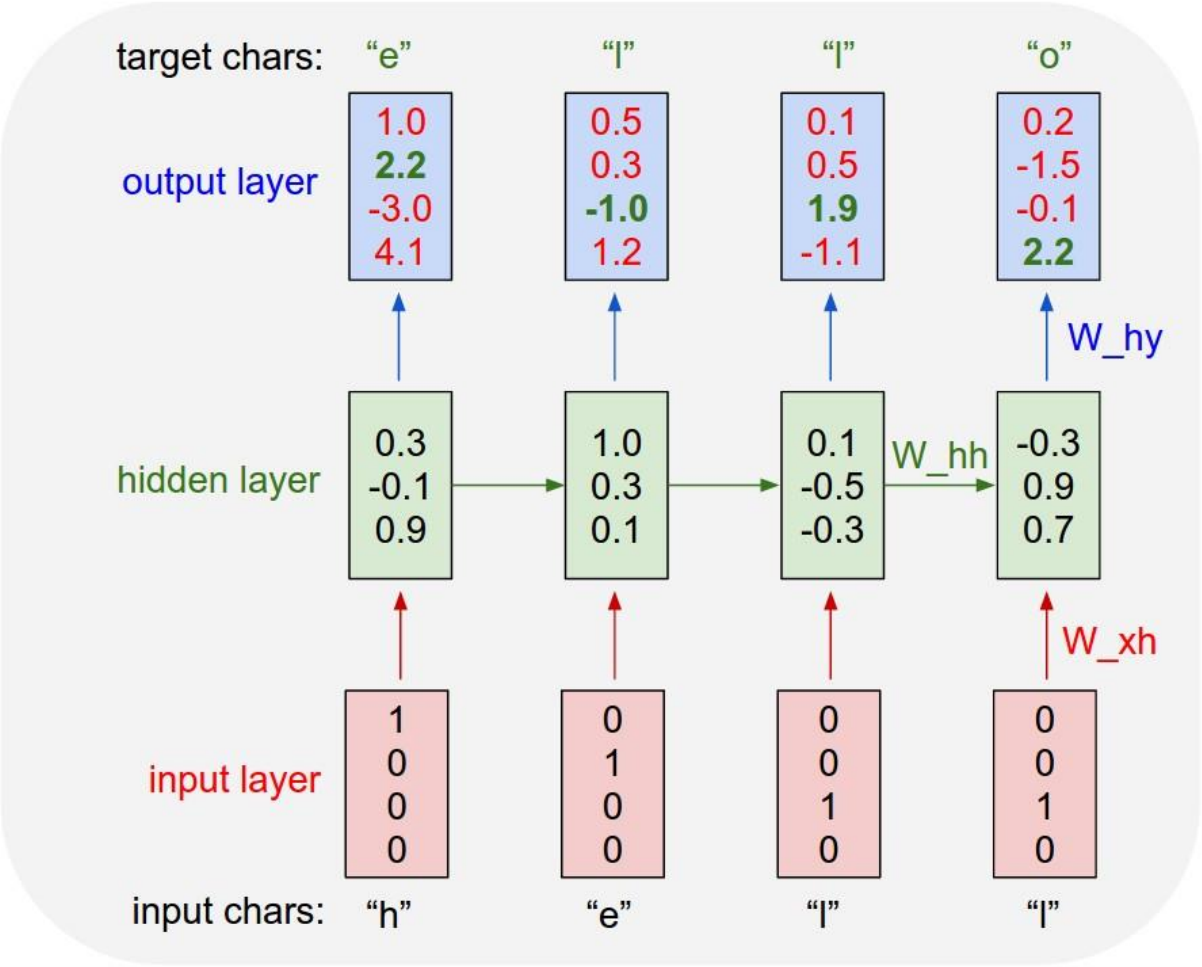
- $P(W_{t+1}|W_t, W_{t-1}, \dots W_0)$
  - 给定当前（W_t）以及更早（$W_{t-1}, \dots W_0$）的词语，预测句子中的下一个词语
  - 在当前时间，有**两个**信息来源
    1. 当前词语的embedding
    2. 一个总结了之前所有单词的状态向量 S_{t-1}

RNN语言模型细节



来自大牛Karpathy的blog

RNN语言模型细节


target chars:
"e"
"l"
"l"
"o"
output layer
1.0
2.2
-3.0
4.1
0.5
0.3
-1.0
1.2
0.1
0.5
1.9
-1.1
0.2
-1.5
-0.1
2.2
W_hy
hidden layer
0.3
-0.1
0.9
1.0
0.3
0.1
0.1
-0.5
-0.3
W_hh
-0.3
0.9
0.7
W_xh
input layer
1
0
0
0
0
1
0
0
0
0
1
0
0
0
1
0
input chars:
"h"
"e"
"l"
"l"

RNN语言模型

代码实践

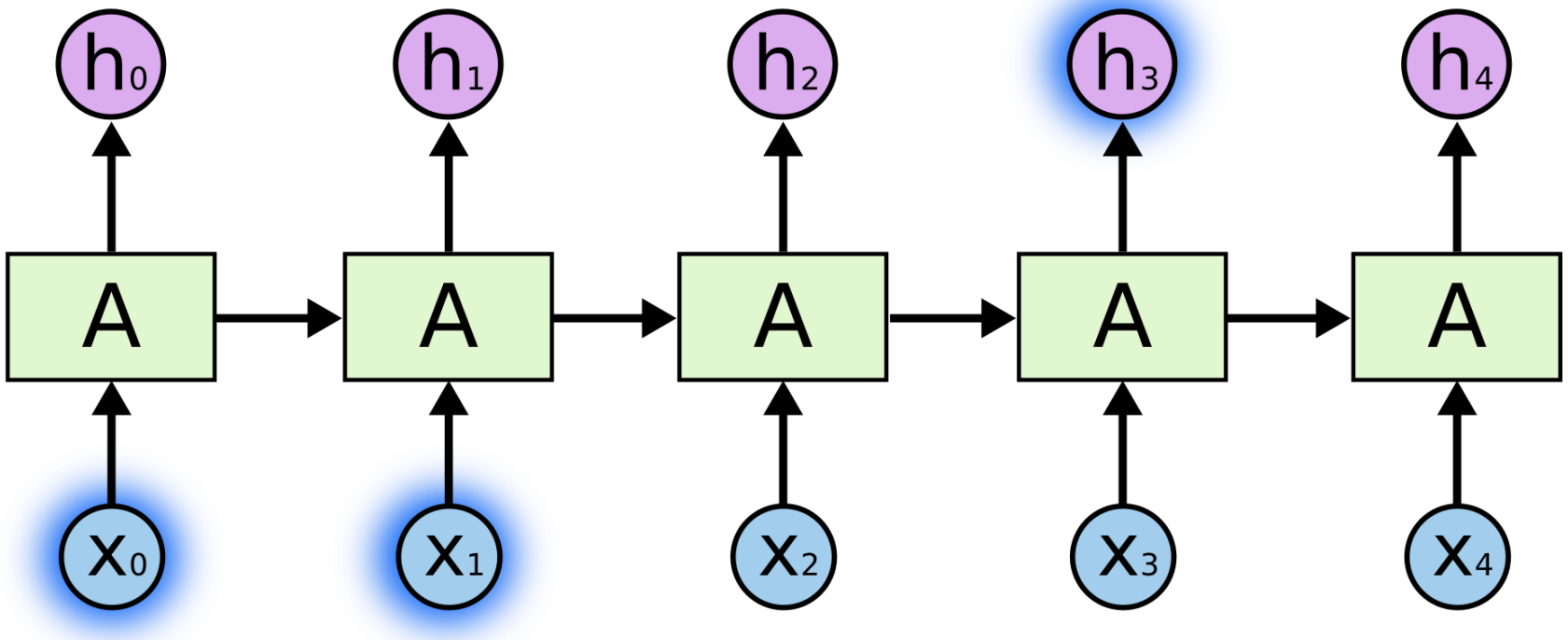
RNN语言模型实践

- RNN 序列模型 第二版
  - 使用 static_rnn
  - 实践 cross-validation 交叉验证

RNN语言模型

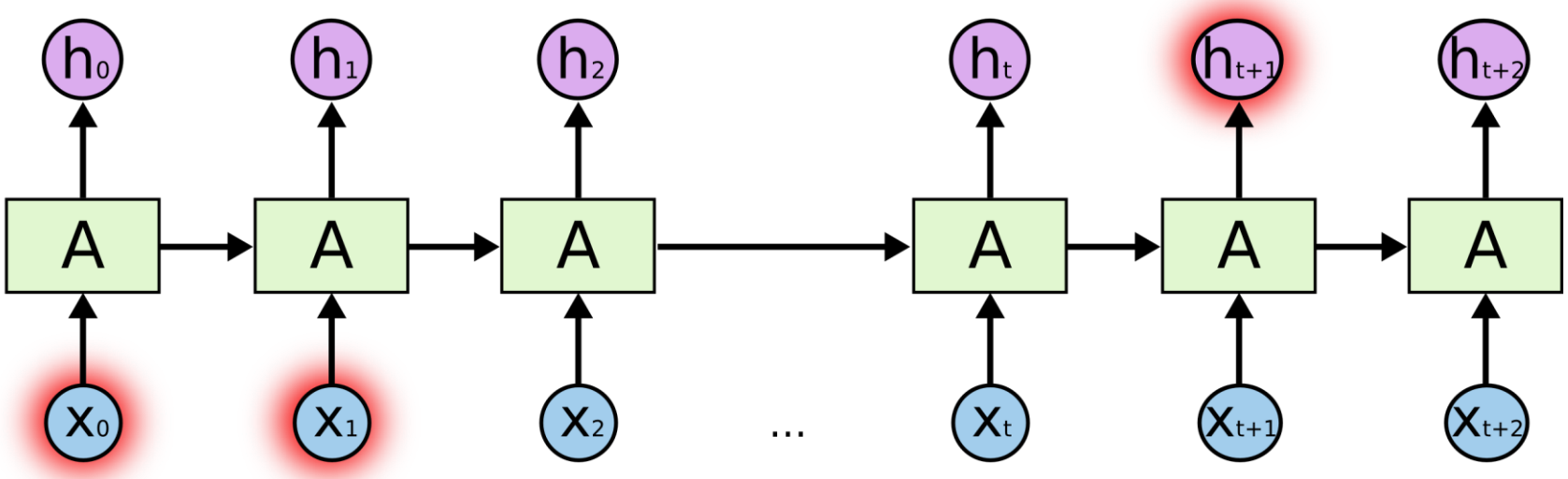
LSTM

相互关联的一些词距离很近时：



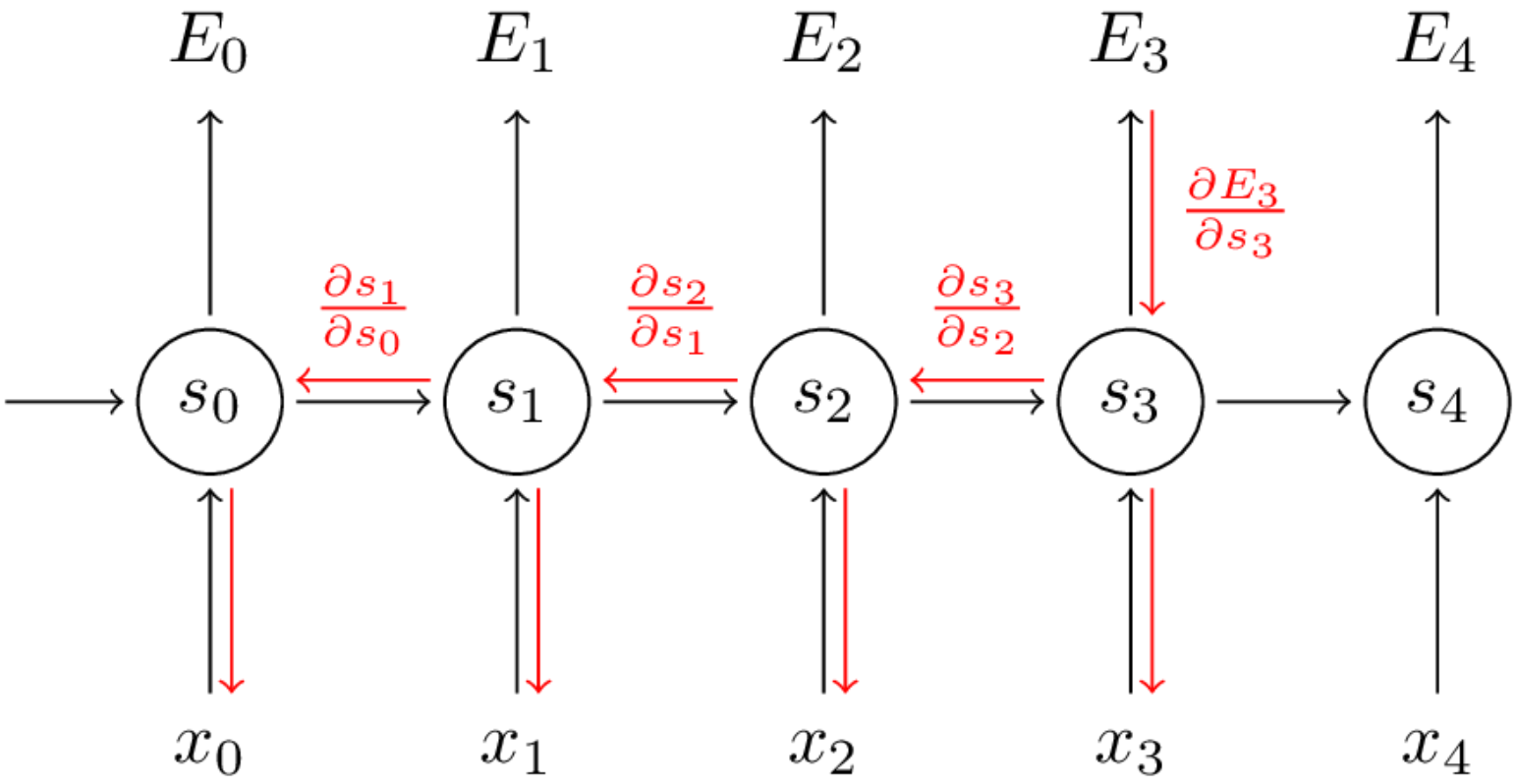
"the **clouds** are in the ***sky***,"

相互关联的一些词距离很远时：



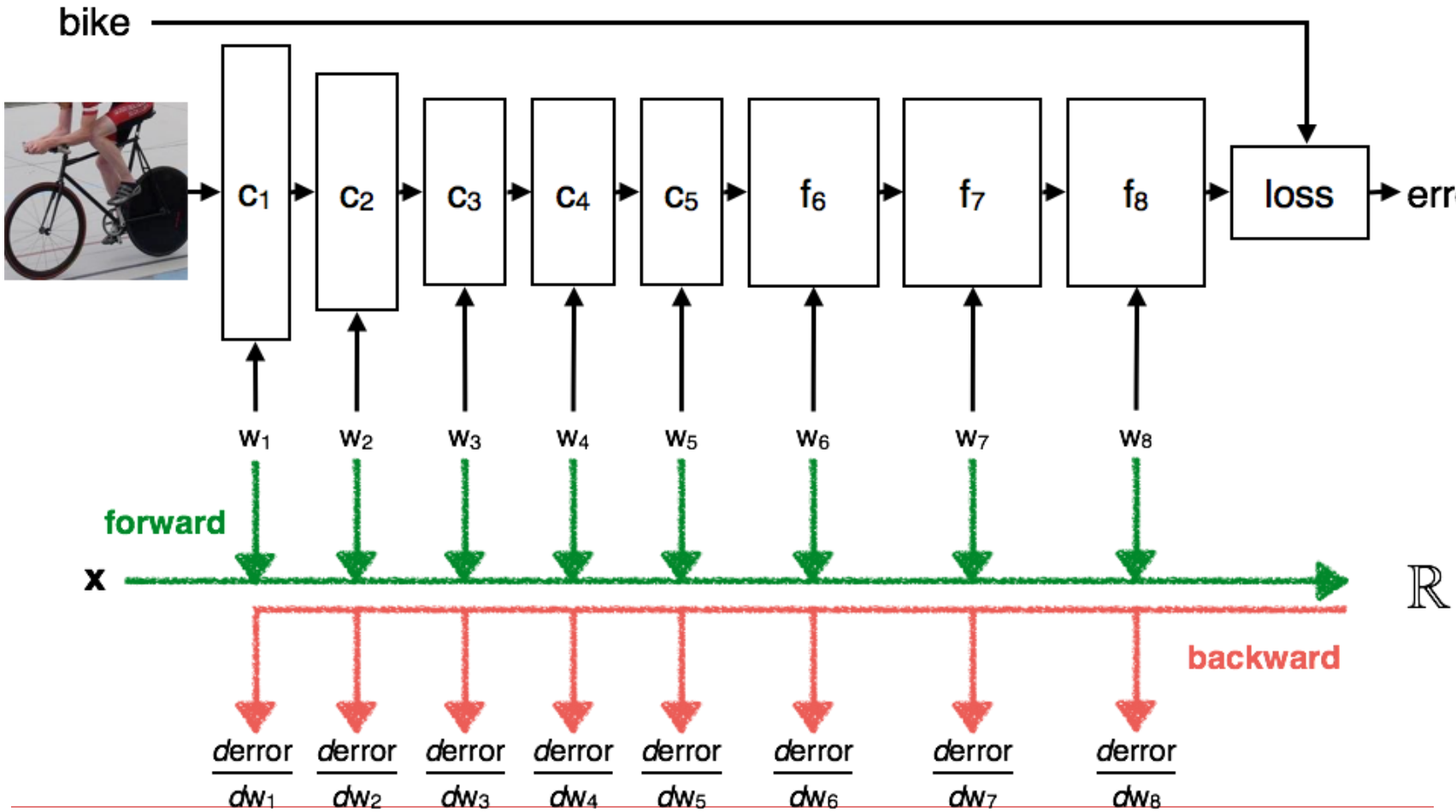
“I grew up in **France**…I ate a log of cheese…. I speak fluent ***French***.”

BPTT Gradient: 爆炸或者消失

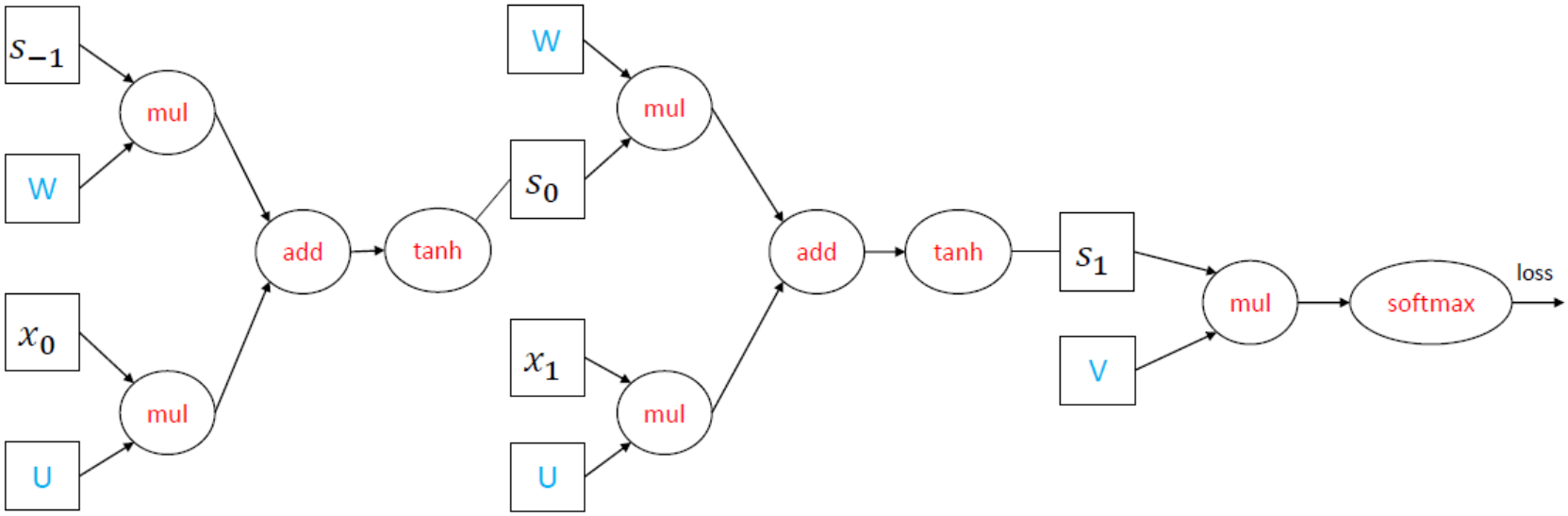


BPTT Gradient: 爆炸或者消失

问题来了，CNN的depth和RNN的depth有什么区别吗？



BPTT Gradient: 爆炸或者消失



$$S_0W + \ldots \rightarrow S_1$$
$$S_1W + \ldots \rightarrow S_2$$
$$S_2W + \ldots \rightarrow S_3$$
$$\ldots$$
$$S_{t-1}W + \ldots \rightarrow S_t$$

在每一个时间点，同一个参数W和状态向量相乘

在长度为T+1的序列里面，W和状态向量相乘T次

BPTT Gradient: 爆炸或者消失

and lacking inputs $\boldsymbol{x}$. As described in Sec. 8.2.5, this recurrence relation essentially describes the power method. It may be simplified to

$$\boldsymbol{h}^{(t)} = \left(\boldsymbol{W}^t\right)^\top \boldsymbol{h}^{(0)}, \tag{10.37}$$

and if $\boldsymbol{W}$ admits an eigendecomposition of the form

$$\boldsymbol{W} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top, \tag{10.38}$$

with orthogonal $\boldsymbol{Q}$, the recurrence may be simplified further to

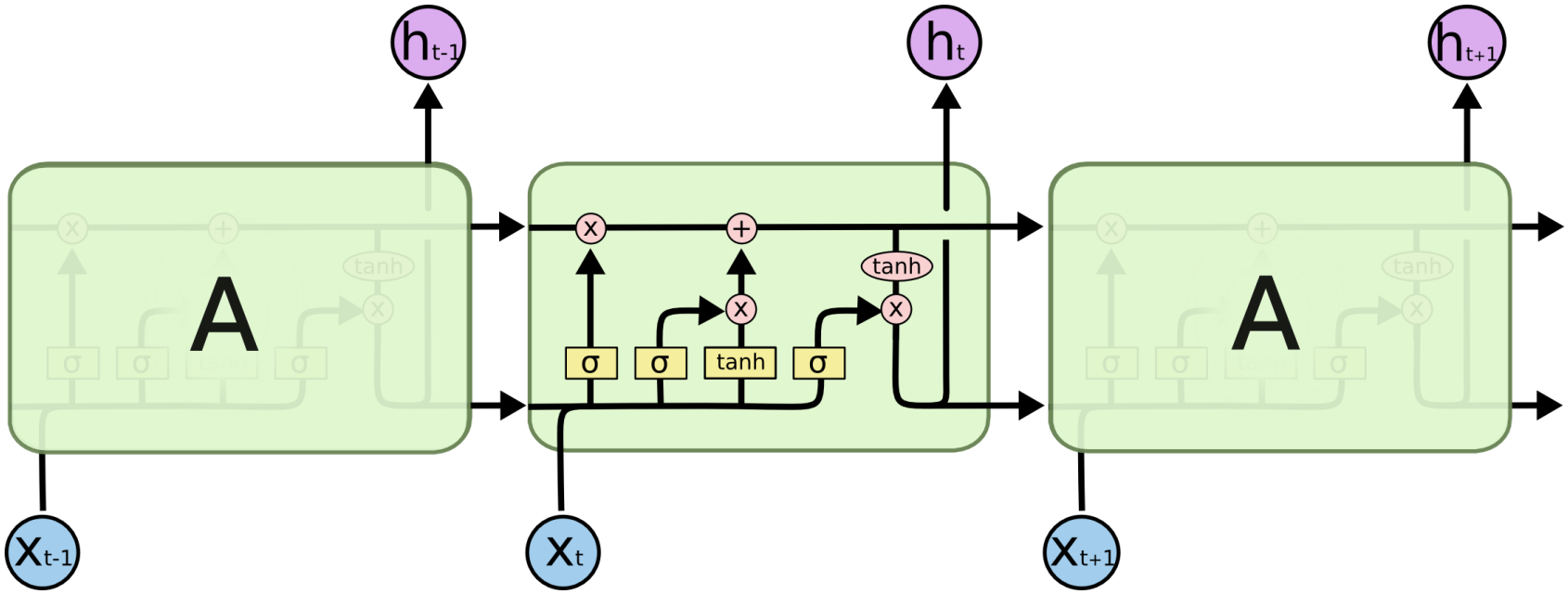
$$\boldsymbol{h}^{(t)} = \boldsymbol{Q}^\top \boldsymbol{\Lambda}^t \boldsymbol{Q}\boldsymbol{h}^{(0)}. \tag{10.39}$$

来自 Deep Learning, Ian Goodfellow

BPTT Gradient: 爆炸或者消失

- 在较长的时间跨度下，不稳定的gradient是学习长距离关联的主要障碍
- LSTMs 被设计用来描述长距离关联（long-term dependency）.
  - **L**ong
  - **S**hort **T**erm
  - **M**emory networks

LSTM



LSTM是一种特殊的RNN CELL

LSTM: Output Gate

中国 男足 踢得 真烂　　气得我 想 大吃 一顿

假定我们用256维的向量表示RNN的状态
(256在RNN应用中算不上特别高维)

LSTM : Output Gate

中国 男足 踢得 真烂　　气得我 想 大吃 一顿

我们需要使用所有的256个维度去表示
“中国男足踢得真烂”这个简单的意思吗？

LSTM : Output Gate

- 中国男足踢得真烂。气得我想大吃一顿。最近有没有什么新开的好吃的餐馆？哎呀我的手机好慢，你帮我去xx app 上查一下？
  1. 256个维度都用来既记录第一句，又记录第二句，还记录第三句
  2. 256个维度有100个记录第一句，78个记录第二句，78个记录第三句（better ☺）

LSTM : Output Gate

我们要求只有一部分维度用来记忆/总结目前为止的输入

中国 男足 踢得 真烂

气得我 想 大吃 一顿

LSTM : Output Gate

中国 男足 踢得 真烂

气得我 想 大吃 一顿

LSTM : Output Gate

中国 男足 踢得 真烂

气得我 想 大吃 一顿

LSTM : Input Forget Gate

中国 男足 踢得 真烂　　气得我 想 大吃 一顿

让我们切换到下一个词语（气得我...）的视角

LSTM : Input Forget Gate

中国 男足 踢得 真烂　　气得我 想 大吃 一顿

让我们切换到下一个词语（气得我...）的视角

LSTM : Input Forget Gate

中国 男足 踢得 真烂　　气得我 想 大吃 一顿

- 256个维度有100个记录“中国 男足 踢得 **真烂**”这句话
- 进一步，我们假设其中25个维度侧重表示“**真烂**”这个词

LSTM : Input Forget Gate

中国 男足 踢得 真烂　　气得我 想 大吃 一顿

- 进一步，我们假设其中25个维度侧重表示“真烂”这个词
- 考虑到当前的词语是“气得我”，和语境里面的“真烂”这个词关系最密切

LSTM : Input Forget Gate

中国 男足 踢得 真烂　　气得我 想 大吃 一顿

- 考虑到当前的词语是“气得我”，和语境里面的“真烂”这个词关系最密切
- 我们把历史状态里面和当前最密切的维度提取出来

LSTM : Input Forget Gate

根据当下的单词，我们检查哪些维度包含的信息和当下单词密切相关？

LSTM : Input Forget Gate

但是，我看到的是O_t处理过的二手信息——一手信息啥样的？

LSTM : Input Forget Gate

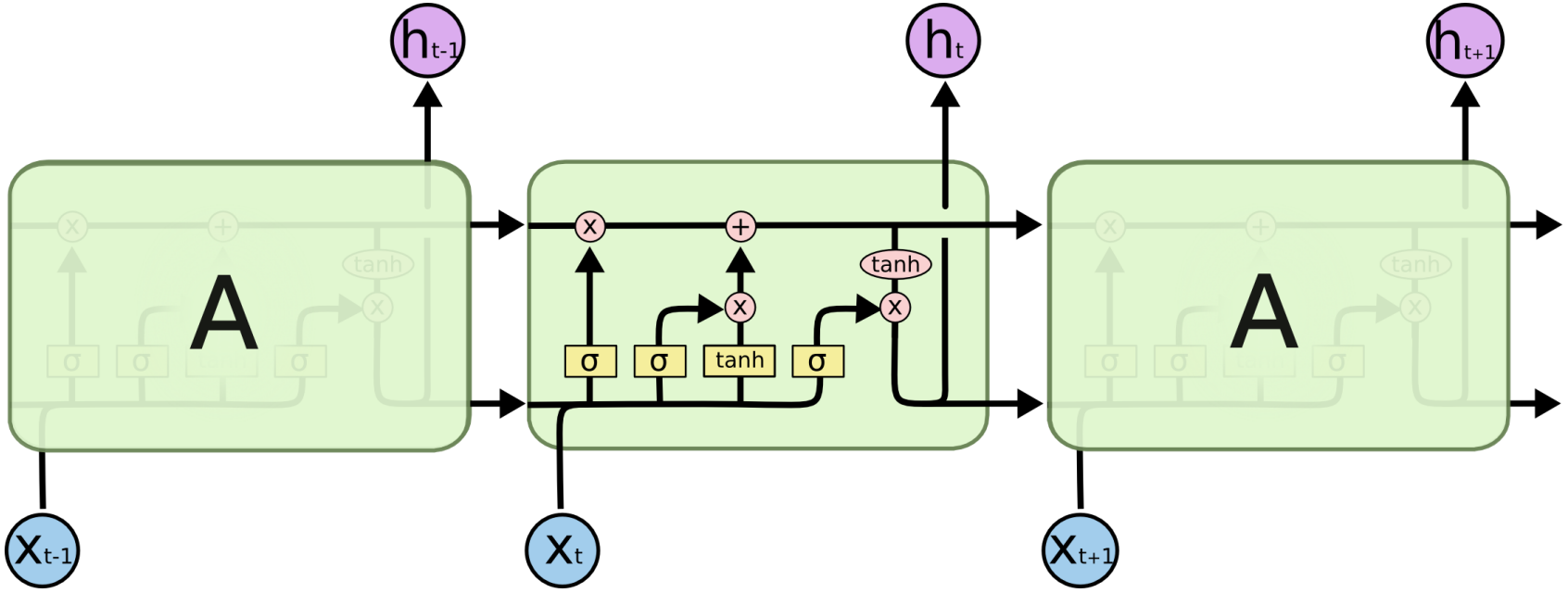
根据当前词语，
挑选的一手信息

LSTM : Input Forget Gate

两类信息如何结合起来？

LSTM : Input Forget Gate

LSTM 小结



LSTM 有两套state: C & h

两句话解释BPTT

- Back Propagation Through Time
  - 无论多么复杂的深度学习模型都是DAG
  - 对于任何一个node X, 想要计算$\frac{\partial L}{\partial X}$, 只需要计算所有children的$\frac{\partial L}{\partial \mathrm{y}}, \mathrm{y} \in Children(X)$: $\frac{\partial L}{\partial X} = \sum_y W_y \frac{\partial L}{\partial \mathrm{y}}$

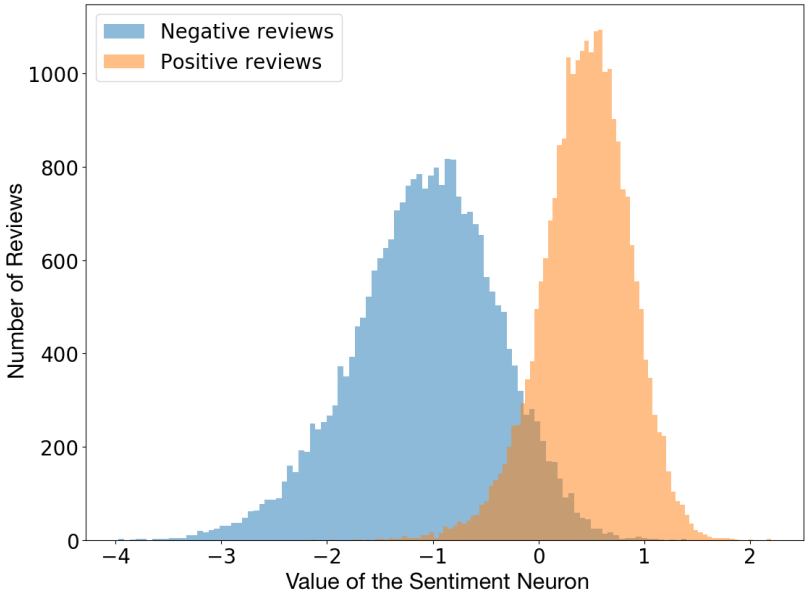
上述理解是否合理？

- 上述理解：
  - 256个维度有100个记录第一句，78个记录第二句，78个记录第三句
  - 256个维度有100个记录“中国 男足 踢得 真烂”这句话；进一步，我们假设其中25个维度侧重表示“真烂”这个词
  - mLSTM有4,096 个维度
  - 发现4096个维度里面有一个neuron node，可以相当准确地预测评论的褒贬

一个支持上述理解的工作

- OpenAI: **Unsupervised Sentiment Neuron**
  - 在82million Amazon 评论数据集上训练RNNLM
  - 使用一个LSTM的变种（multiplicative LSTM)

1. mLSTM有4,096 个维度
2. 这4096个维度里面有一个neuron node，可以相当准确地预测评论的褒贬



LSTM代码演示

疑问

- 问题答疑： http://www.xxwenda.com/
  - 可邀请老师或者其他人回答问题

联系我们

小象学院：互联网新技术在线教育领航者

- 微信公众号：大数据分析挖掘
- 新浪微博：ChinaHadoop