深度学习-第十一章-序列模型

黄海广 副教授

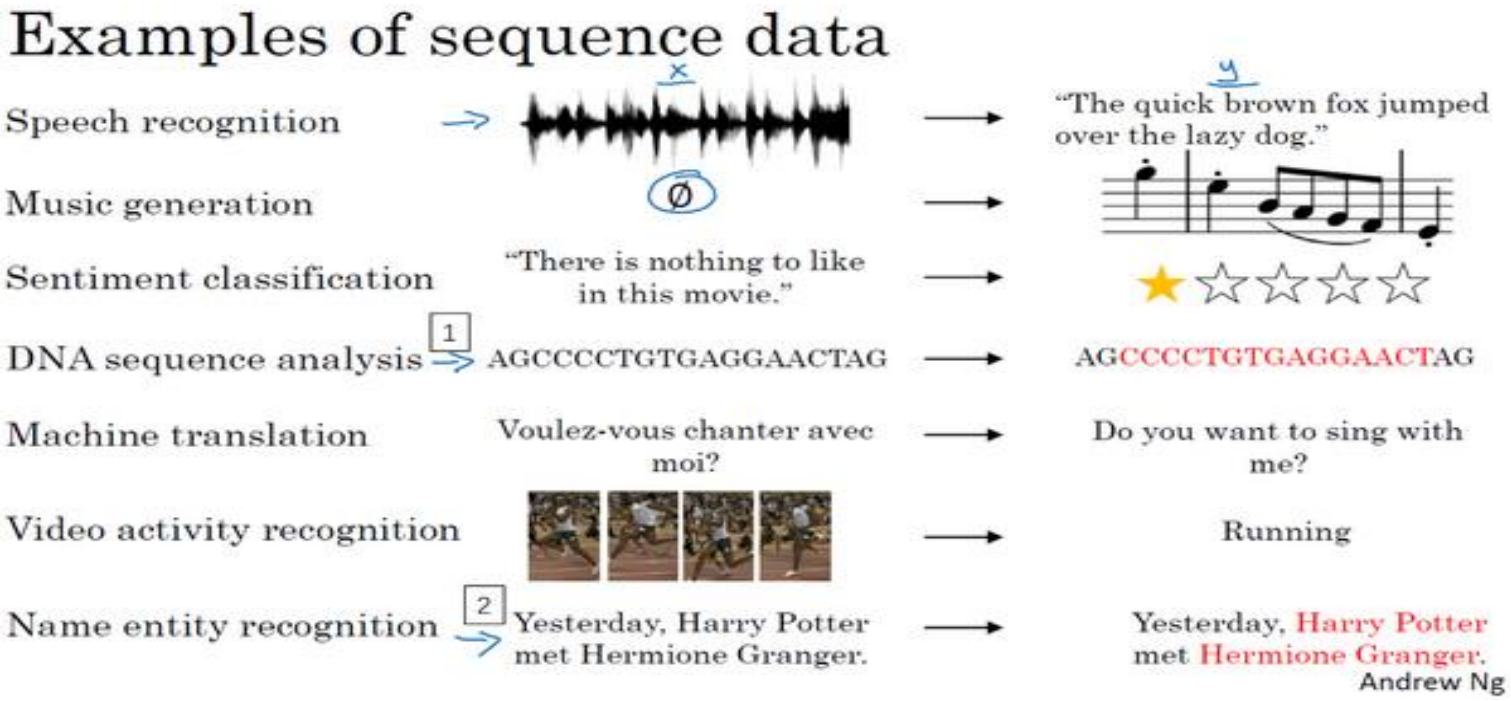
2021年05月

本章目录

1.序列模型概述

1.序列模型概述

循环神经网络（**RNN**）之类的模型在语音识别、自然语言处理和其他领域中引起变革。



数学符号

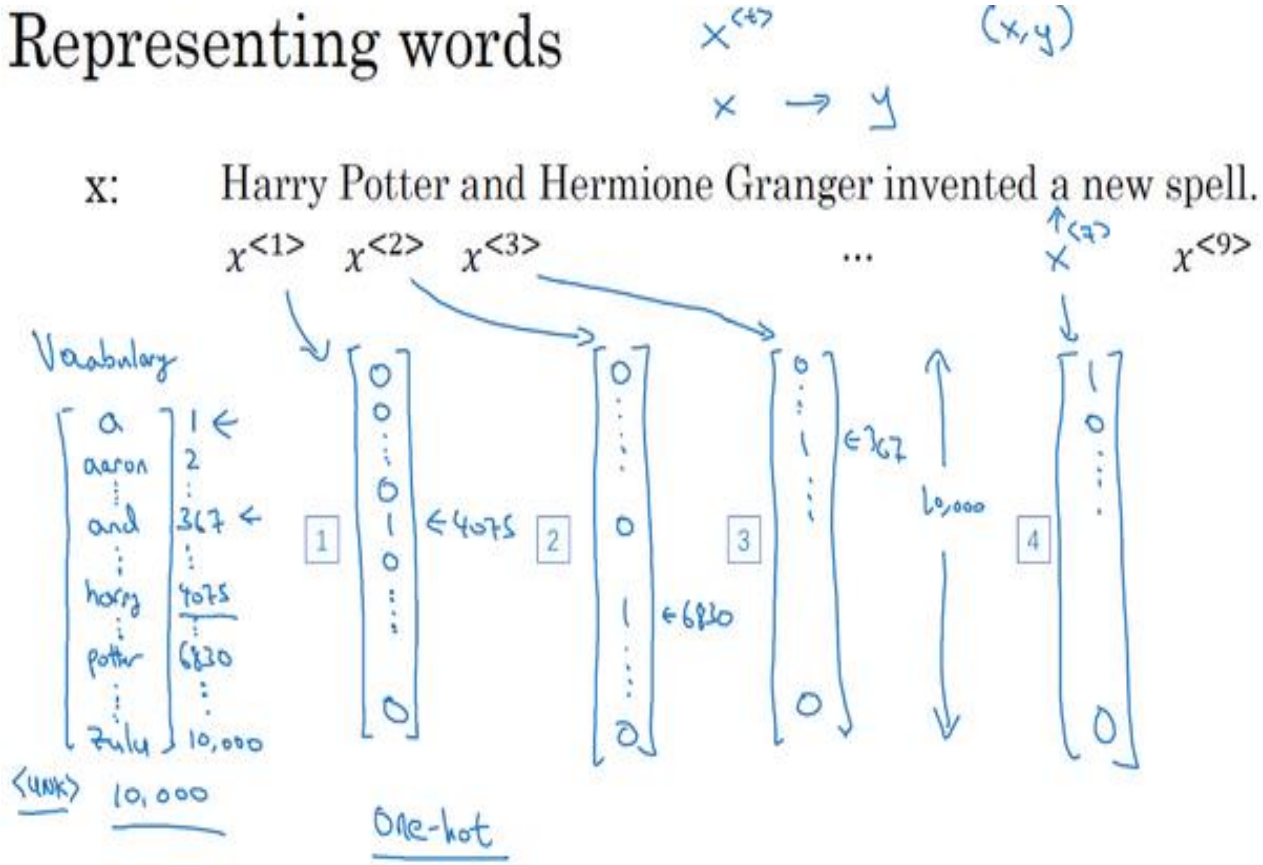
在这里$x^{<1>}$表示**Harry**这个单词，它就是一个第4075行是1，其余值都是0的向量（上图编号1所示），因为那是**Harry**在这个词典里的位置。

$x^{<2>}$是第6830行是1，其余位置都是0的向量（上图编号2所示）。

and在词典里排第367，所以$x^{<3>}$就是第367行是1，其余值都是0的向量（上图编号3所示）。

因为**a**是字典第一个单词，$x^{<7>}$对应**a**，那么这个向量的第一个位置为1，其余位置都是0的向量（上图编号4所示）。

Unknow Word的伪单词，用<**UNK**>作为标记。

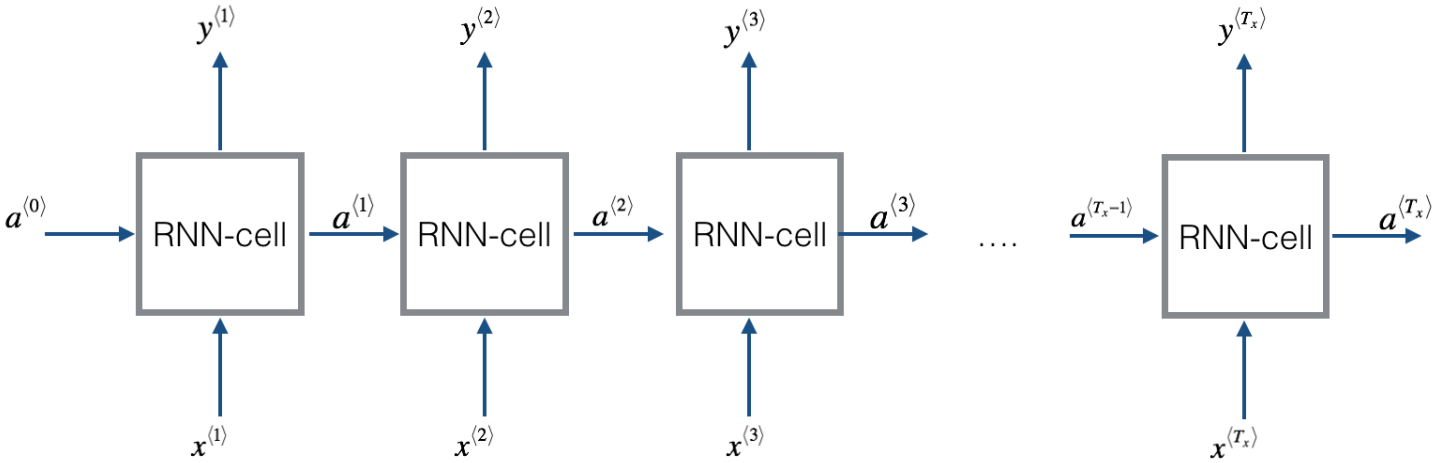


如果你的词典大小是10,000的话，那么这里的每个向量都是10,000维的。

2.循环神经网络(RNN)

2.循环神经网络(RNN)

RNN的前向传播



$$a^{<0>} = 0$$

$$a^{<1>} = g_1(W_{aa}a^{<0>} + W_{ax}x^{<1>} + b_a)$$

$$\hat{y}^{<1>} = g_2(W_{ya}a^{<1>} + b_y)$$

$$a^{<t>} = g_1(W_{aa}a^{<t-1>} + W_{ax}x^{<t>} + b_a)$$

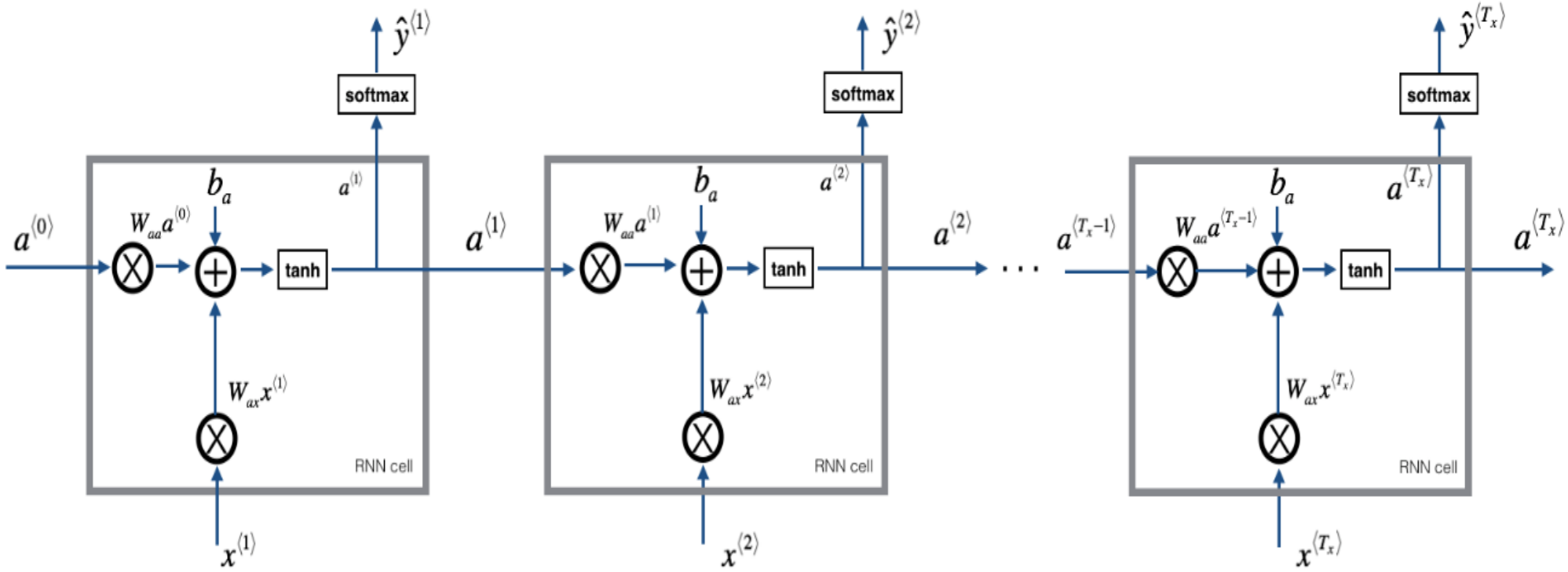
$$\hat{y}^{<t>} = g_2(W_{ya}a^{<t>} + b_y)$$

2.循环神经网络(RNN)

RNN的前向传播

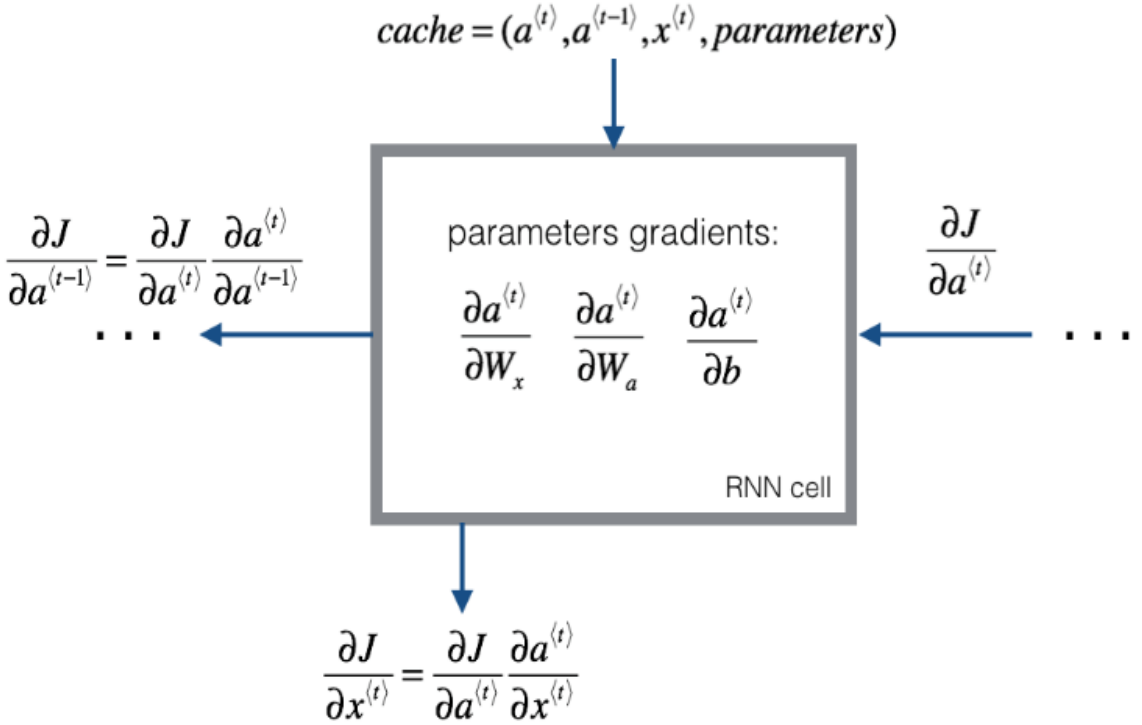
$$a^{<t>} = g_1(W_{aa}a^{<t-1>} + W_{ax}x^{<t>} + b_a)$$

$$\hat{y}^{<t>} = g_2(W_{ya}a^{<t>} + b_y)$$



2.循环神经网络(RNN)

RNN的反向传播



$$a^{\langle t \rangle} = \tanh(W_{ax} x^{\langle t \rangle} + W_{aa} a^{\langle t-1 \rangle} + b)$$

$$\frac{\partial \tanh(x)}{\partial x} = 1 - \tanh(x)^2$$

$$\frac{\partial a^{\langle t \rangle}}{\partial W_{ax}} = (1 - \tanh(W_{ax} x^{\langle t \rangle} + W_{aa} a^{\langle t-1 \rangle} + b)^2) x^{\langle t \rangle T}$$

$$\frac{\partial a^{\langle t \rangle}}{\partial W_{aa}} = (1 - \tanh(W_{ax} x^{\langle t \rangle} + W_{aa} a^{\langle t-1 \rangle} + b)^2) a^{\langle t-1 \rangle T}$$

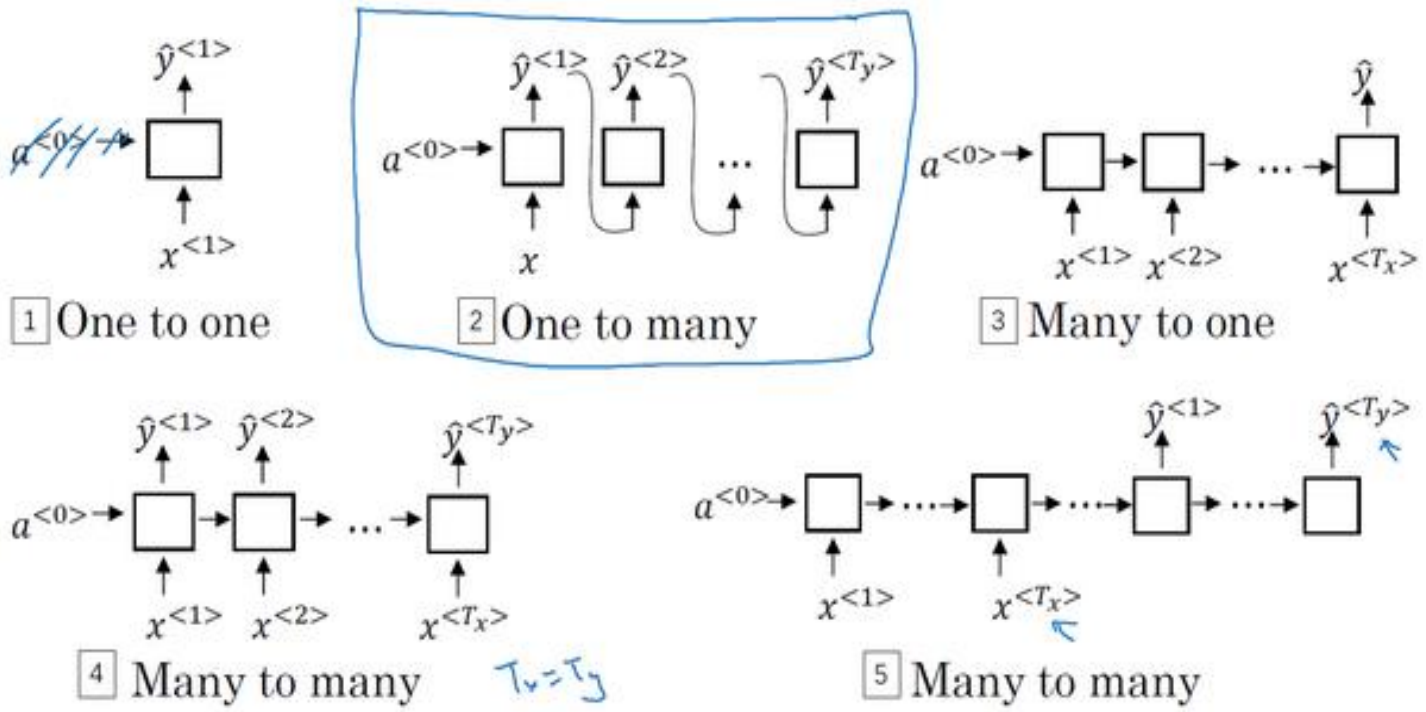
$$\frac{\partial a^{\langle t \rangle}}{\partial b} = \sum_{batch} (1 - \tanh(W_{ax} x^{\langle t \rangle} + W_{aa} a^{\langle t-1 \rangle} + b)^2)$$

$$\frac{\partial a^{\langle t \rangle}}{\partial x^{\langle t \rangle}} = W_{ax}{}^T . (1 - \tanh(W_{ax} x^{\langle t \rangle} + W_{aa} a^{\langle t-1 \rangle} + b)^2)$$

$$\frac{\partial a^{\langle t \rangle}}{\partial a^{\langle t-1 \rangle}} = W_{aa}{}^T . (1 - \tanh(W_{ax} x^{\langle t-1 \rangle} + W_{aa} a^{\langle t-1 \rangle} + b)^2)$$

2.循环神经网络(RNN)

RNN的类型



2.循环神经网络(RNN)

语言模型和序列生成

The apple and pear（pair） salad was delicious.

第一句话的概率是:

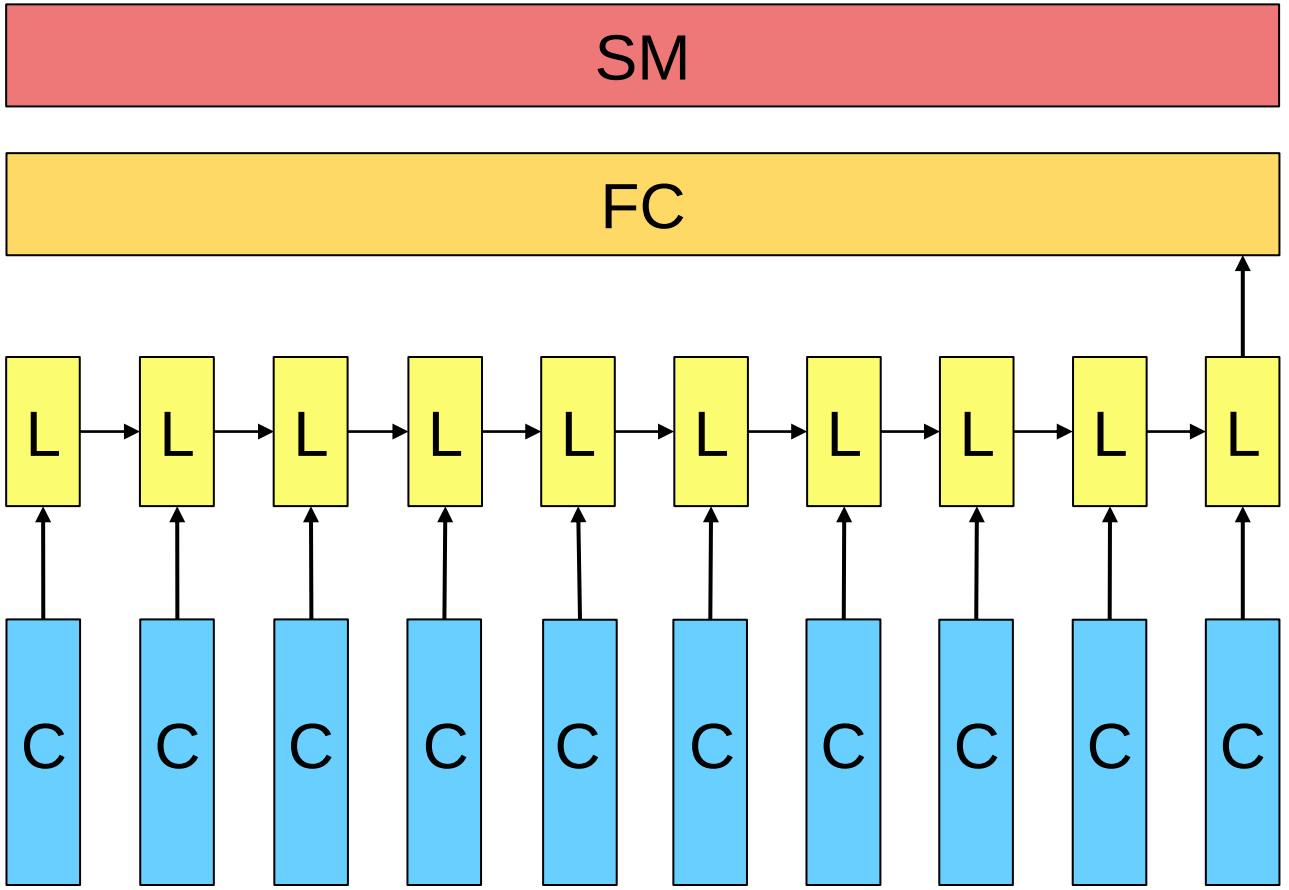
$P(\text{The apple and pair salad}) = 3.2 \times 10^{-13}$,

而第二句话的概率是:

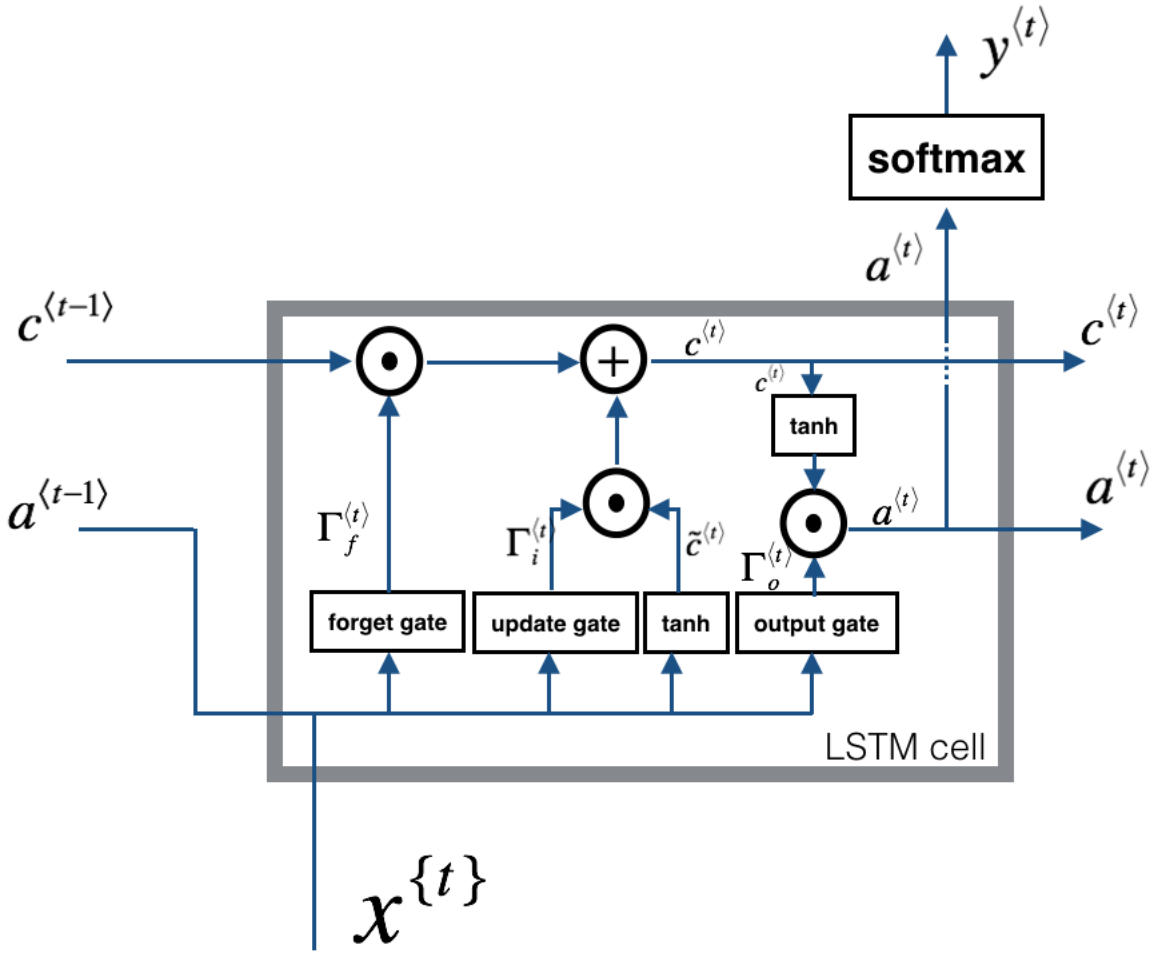
$P(\text{The apple and pear salad}) = 5.7 \times 10^{-10}$,

3.长短期记忆(LSTM)

3.长短期记忆(LSTM)



3.长短期记忆(LSTM)



$$\Gamma_f^{\langle t \rangle} = \sigma(W_f[a^{\langle t-1 \rangle}, x^{\langle t \rangle}] + b_f)$$

$$\Gamma_u^{\langle t \rangle} = \sigma(W_u[a^{\langle t-1 \rangle}, x^{\langle t \rangle}] + b_u)$$

$$\tilde{c}^{\{t\}} = \tanh(W_C[a^{\langle t-1 \rangle}, x^{\langle t \rangle}] + b_C)$$

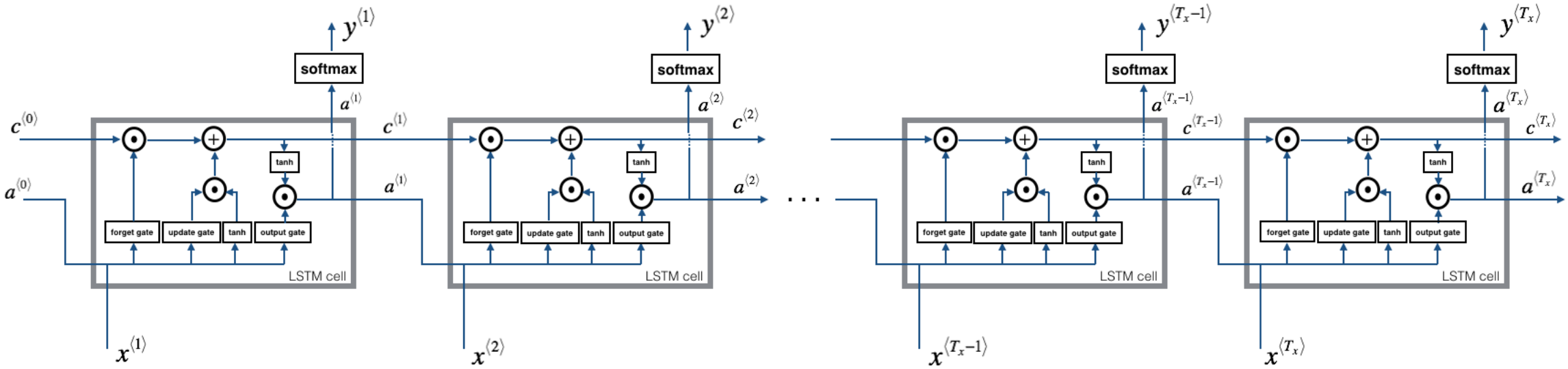
$$c^{\langle t \rangle} = \Gamma_f^{\langle t \rangle} \circ c^{\langle t-1 \rangle} + \Gamma_u^{\langle t \rangle} \circ \tilde{c}^{\langle t \rangle}$$

$$\Gamma_o^{\langle t \rangle} = \sigma(W_o[a^{\langle t-1 \rangle}, x^{\langle t \rangle}] + b_o)$$

$$a^{\langle t \rangle} = \Gamma_o^{\langle t \rangle} \circ \tanh(c^{\langle t \rangle})$$

3.长短期记忆(LSTM)

LSTM的前向传播



3.长短期记忆(LSTM)

LSTM的反向传播

$$d\Gamma_o^{\langle t\rangle} = da_{next} * \tanh(c_{next}) * \Gamma_o^{\langle t\rangle} * (1 - \Gamma_o^{\langle t\rangle})$$

$$d\tilde{c}^{\langle t\rangle} = dc_{next} * \Gamma_i^{\langle t\rangle} + \Gamma_o^{\langle t\rangle}(1 - \tanh(c_{next})^2) * i_t * da_{next} * \tilde{c}^{\langle t\rangle} * (1 - \tanh(\tilde{c})^2)$$

$$d\Gamma_u^{\langle t\rangle} = dc_{next} * \tilde{c}^{\langle t\rangle} + \Gamma_o^{\langle t\rangle}(1 - \tanh(c_{next})^2) * \tilde{c}^{\langle t\rangle} * da_{next} * \Gamma_u^{\langle t\rangle} * (1 - \Gamma_u^{\langle t\rangle})$$

$$d\Gamma_f^{\langle t\rangle} = dc_{next} * \tilde{c}_{prev} + \Gamma_o^{\langle t\rangle}(1 - \tanh(c_{next})^2) * c_{prev} * da_{next} * \Gamma_f^{\langle t\rangle} * (1 - \Gamma_f^{\langle t\rangle})$$

3.长短期记忆(LSTM)

LSTM的反向传播

parameter derivatives

$$dW_f = d\Gamma_f^{\langle t \rangle} * \begin{pmatrix} a_{prev} \\ x_t \end{pmatrix}^T$$

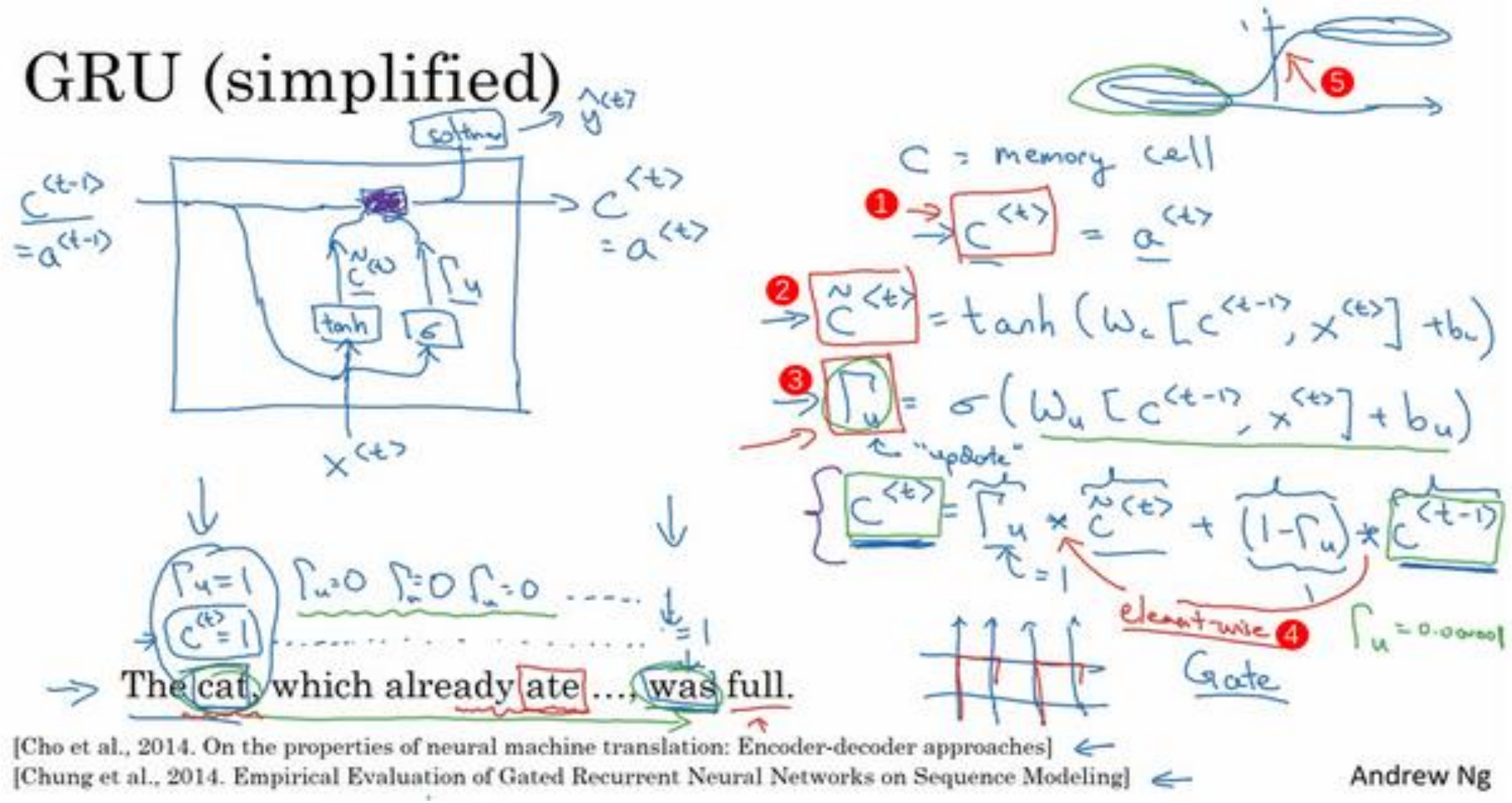
$$dW_c = d\tilde{c}^{\langle t \rangle} * \begin{pmatrix} a_{prev} \\ x_t \end{pmatrix}^T$$

$$dW_o = d\Gamma_o^{\langle t \rangle} * \begin{pmatrix} a_{prev} \\ x_t \end{pmatrix}^T$$

To calculate db_f, db_u, db_c, db_o you just need to sum across the horizontal (axis= 1) axis on $d\Gamma_f^{\langle t \rangle}, d\Gamma_u^{\langle t \rangle}, d\tilde{c}^{\langle t \rangle}, d\Gamma_o^{\langle t \rangle}$ respectively. Note that you should have the `keep_dims = True` option.

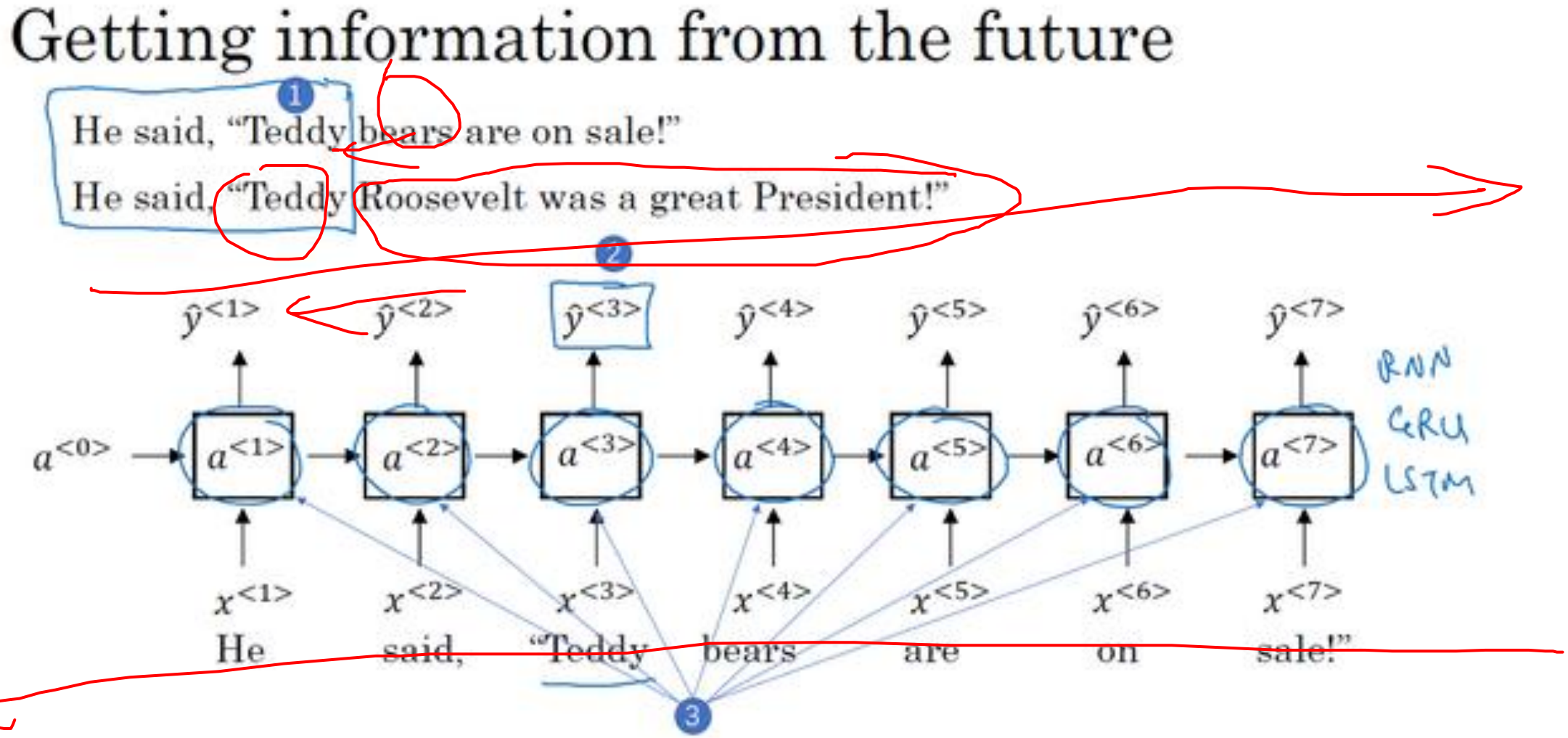
Finally, you will compute the derivative with respect to the previous hidden state.

3.长短期记忆(LSTM)


GRU (simplified)
softmax
tanh
c = memory cell
"update"
element-wise
Gate
The cat, which already ate ..., was full.
[Cho et al., 2014. On the properties of neural machine translation: Encoder-decoder approaches]
[Chung et al., 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling]
Andrew Ng

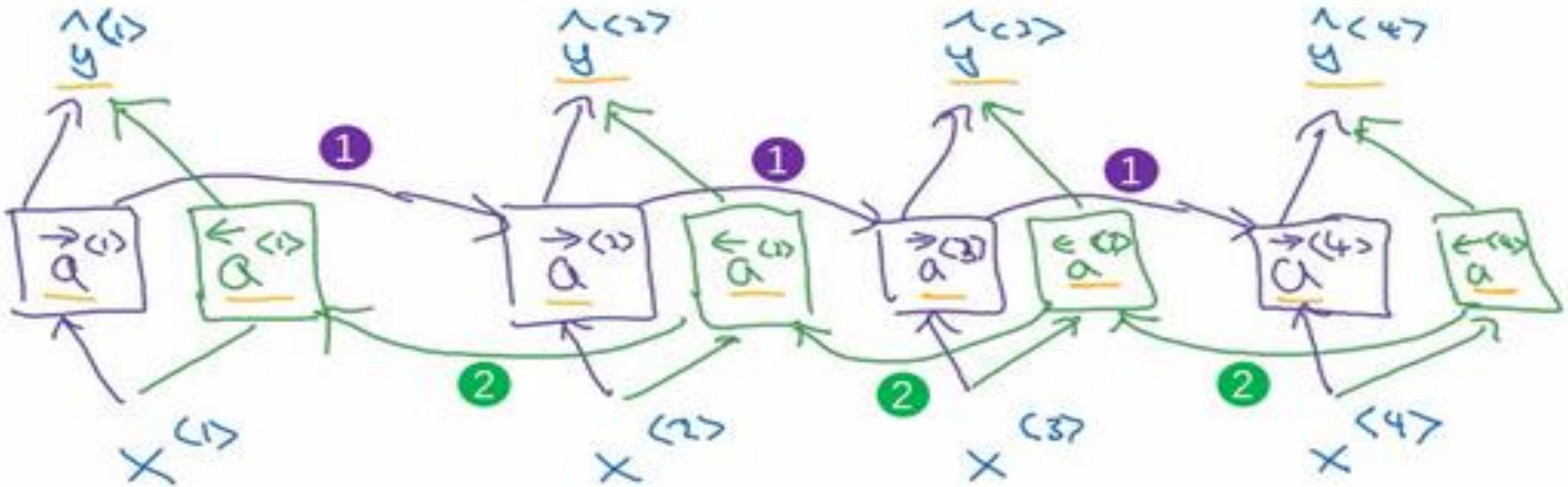
2.循环神经网络(RNN)

4.双向循环神经网络


Getting information from the future
1
He said, "Teddy bears are on sale!"
He said, "Teddy Roosevelt was a great President!"
2
$\hat{y}^{<1>}$ $\hat{y}^{<2>}$ $\hat{y}^{<3>}$ $\hat{y}^{<4>}$ $\hat{y}^{<5>}$ $\hat{y}^{<6>}$ $\hat{y}^{<7>}$
$a^{<0>}$ $a^{<1>}$ $a^{<2>}$ $a^{<3>}$ $a^{<4>}$ $a^{<5>}$ $a^{<6>}$ $a^{<7>}$
RNN
GRU
LSTM
$x^{<1>}$ $x^{<2>}$ $x^{<3>}$ $x^{<4>}$ $x^{<5>}$ $x^{<6>}$ $x^{<7>}$
He said, "Teddy bears are on sale!"
3

4.双向循环神经网络

Bidirectional RNN (BRNN)



Acyclic graph

5.深层循环神经网络

01 序列模型概述

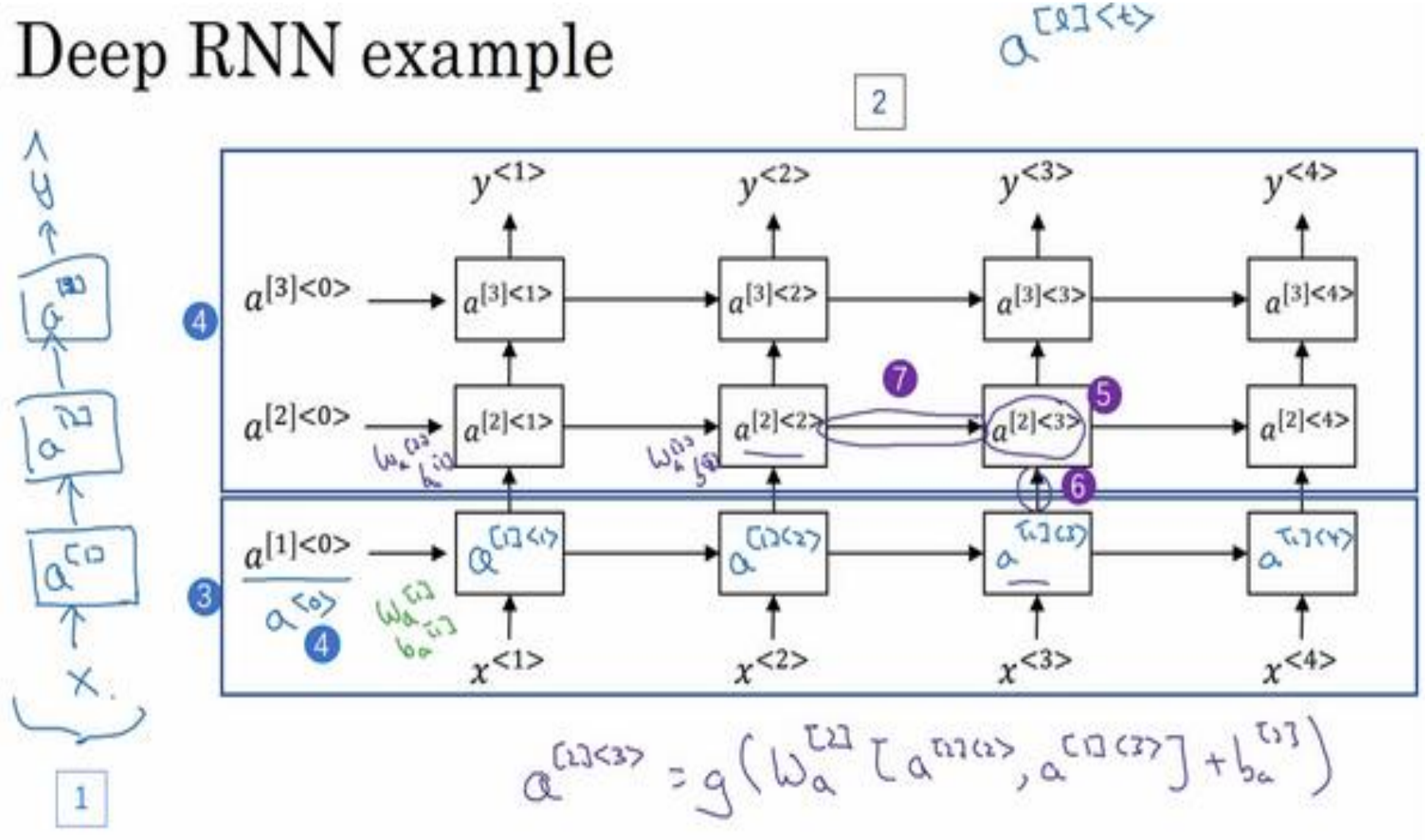
02 循环神经网络(RNN)

03 长短期记忆(LSTM)

04 双向循环神经网络

05 深层循环神经网络

5.深层循环神经网络


Deep RNN example
$y^{<1>}$
$y^{<2>}$
$y^{<3>}$
$y^{<4>}$
$a^{[3]<0>}$
$a^{[3]<1>}$
$a^{[3]<2>}$
$a^{[3]<3>}$
$a^{[3]<4>}$
$a^{[2]<0>}$
$a^{[2]<1>}$
$a^{[2]<2>}$
$a^{[2]<3>}$
$a^{[2]<4>}$
$a^{[1]<0>}$
$a^{[1]<1>}$
$a^{[1]<2>}$
$a^{[1]<3>}$
$a^{[1]<4>}$
$x^{<1>}$
$x^{<2>}$
$x^{<3>}$
$x^{<4>}$
1
2
3
4
5
6
7
$a^{[2]<t>}$
$a^{[2]<3>} = g(W_a^{[2]}[a^{[2]<2>}, a^{[1]<3>}] + b_a^{[2]})$

参考文献

1. IAN GOODFELLOW等，《深度学习》，人民邮电出版社，2017
2. Andrew Ng，http://www.deeplearning.ai


谢 谢!