神经网络第十四章

2022年7月23日 星期六 00:40

Why Sequence Models?

Notation

x: Harry potter an ... a new spell

$x^{<1>}$ $x^{<2>}$ $x^{<3>}$ $x^{<4>}$ ... $x^{<t>}$ ... $x^{<9>}$

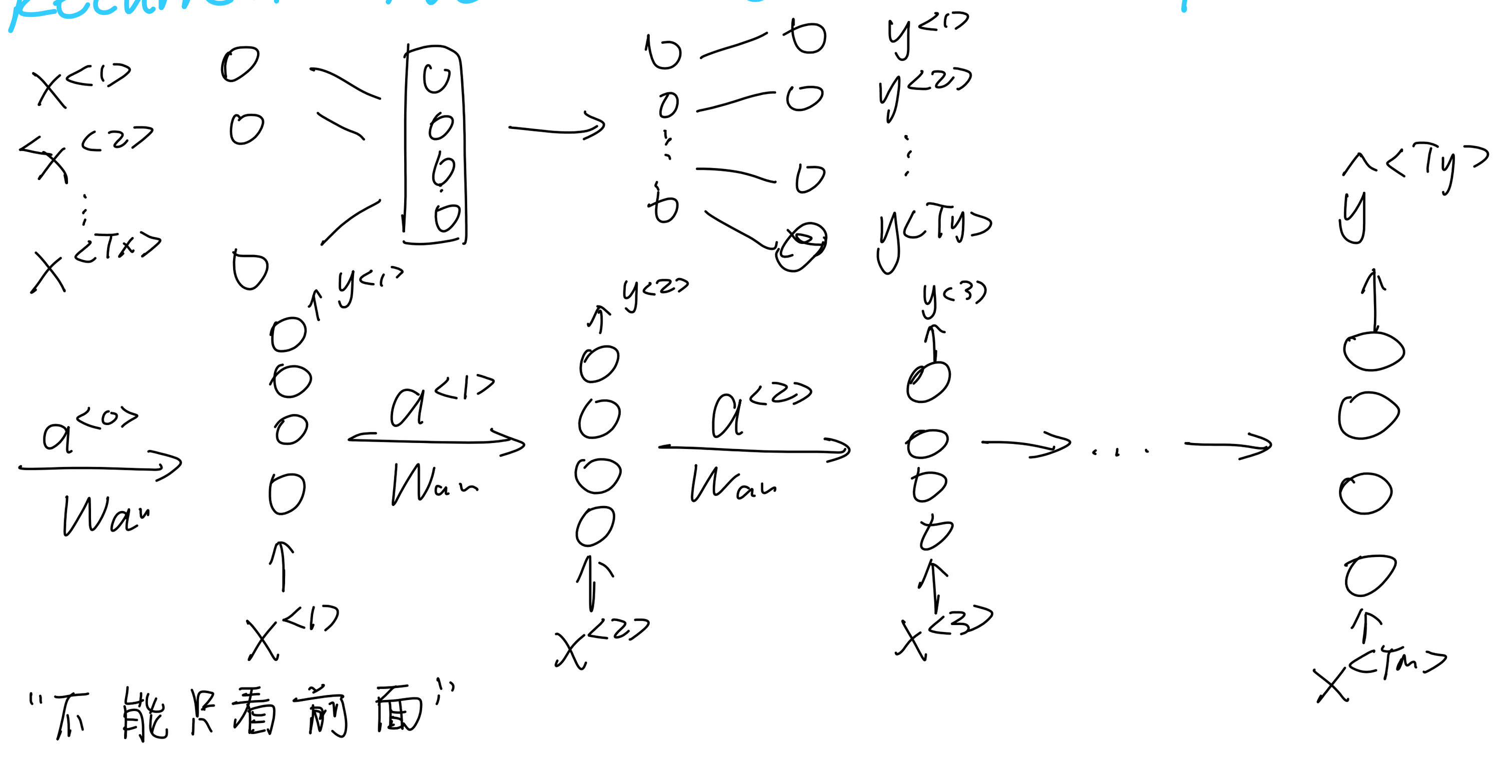
y: 1 1 0 1 1 0 0 0 0

$y^{<1>}$ $y^{<2>}$ $y^{<3>}$ $y^{<4>}$ $y^{<5>}$ $y^{<6>}$ $y^{<7>}$ $y^{<8>}$

Vocabulary

$$\begin{bmatrix} a \\ aron \\ \vdots \\ and \\ \vdots \\ harry \\ \vdots \\ potter \\ \vdots \\ Rulu \end{bmatrix} \begin{matrix} 1 \\ 2 \\ \vdots \\ 367 \\ \vdots \\ 4075 \\ \vdots \\ 6830 \\ \vdots \\ 10{,}000 \end{matrix}$$

Recurrent Neural Network Model



"不能只看前面"

"Teddy was a great President"

"Teddy bears ~~are~~ on sale."

$a^{<0>} = \vec{0}$ $\quad a^{<1>} = g(W_{aa} a^{<0>} + W_{ax} x^{<1>} + b_a) \leftarrow$ tanh/Relu

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

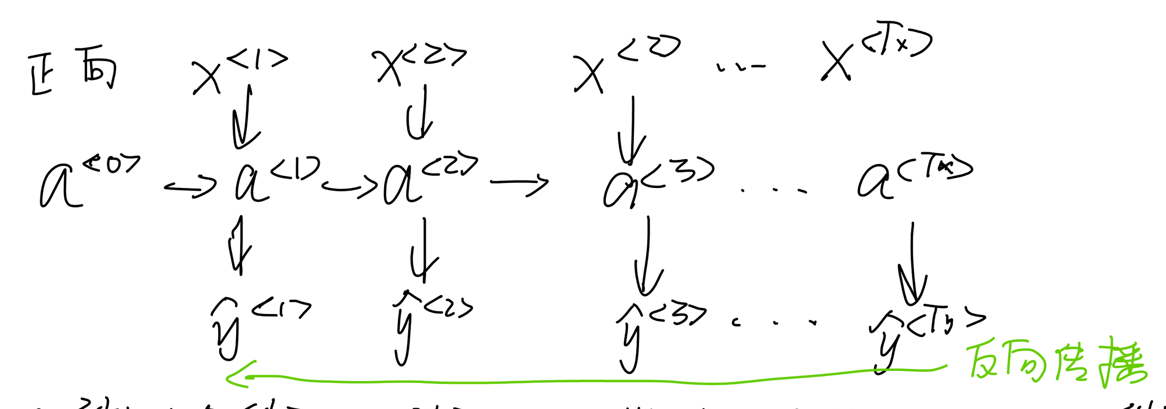
$a^{<t>} = g(W_{aa} a^{<t-1>} + W_{ax} x^{<t>} + b_a)$

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

$a^{<t>} = g(W_a[a^{<t-1>}, x^{<t>}] + b_a)$

↳ 指 $\begin{bmatrix} a^{<t-1>} \\ x^{<t>} \end{bmatrix}$ $\updownarrow 100$ $\updownarrow 10000$ $\updownarrow$ 10100 维向量

Backpropagation through time



$\mathcal{L}^{<t>}(\hat{y}^{<t>}, y^{<t>}) = -y^{<t>} \log \hat{y}^{<t>} - (1 - y^{<t>}) \log(1 - \hat{y}^{<t>})$

$\mathcal{L}(\hat{y}, y) = \sum_{t=1}^{T_y} \mathcal{L}^{<t>}(\hat{y}^{<t>}, y^{<t>})$

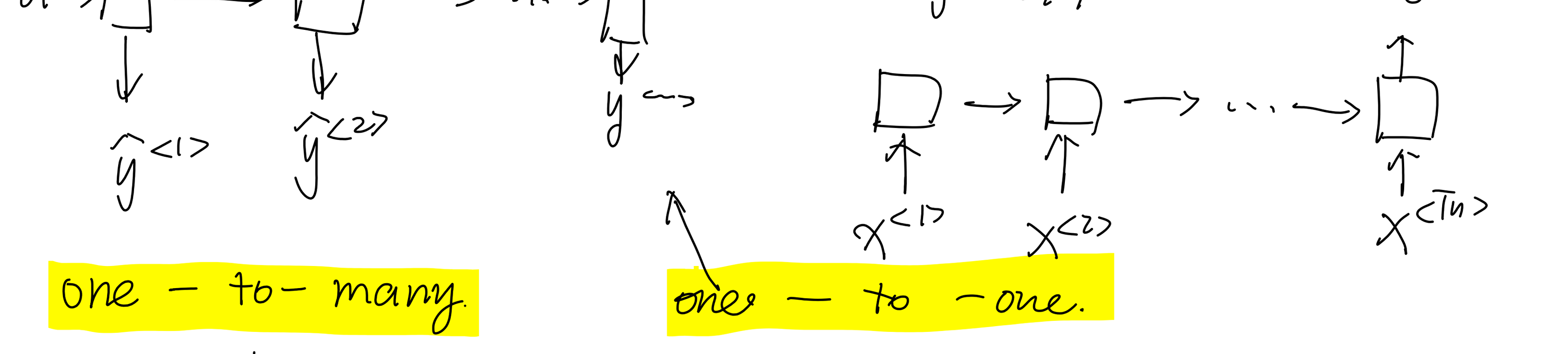
Different types of RNNs

$T_x = T_y$

setiment classifacation

$x =$ text

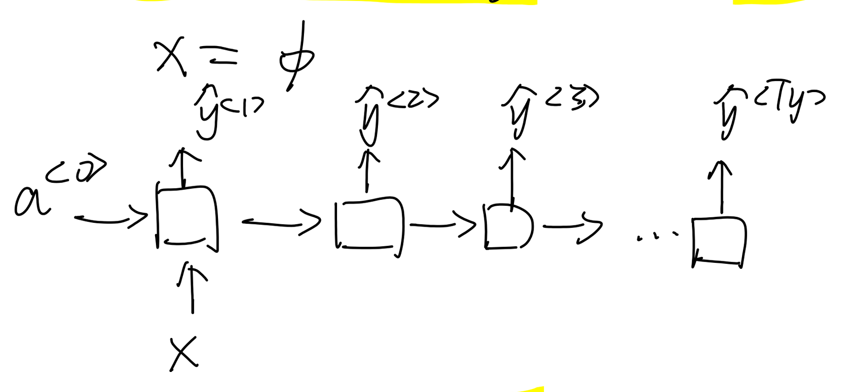
$y = 0/1 \quad 1..5$



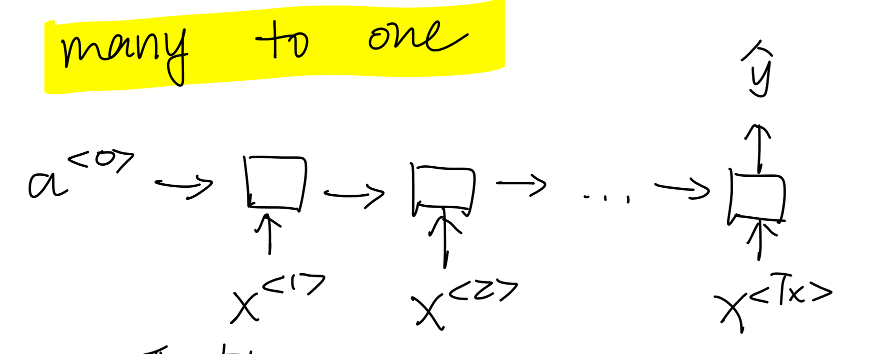
one-to-many

one-to-one.

$x = \phi$



many to one



(需要输入一段影评，生成...)

Language Model and Sequence Generation

Speech recognition

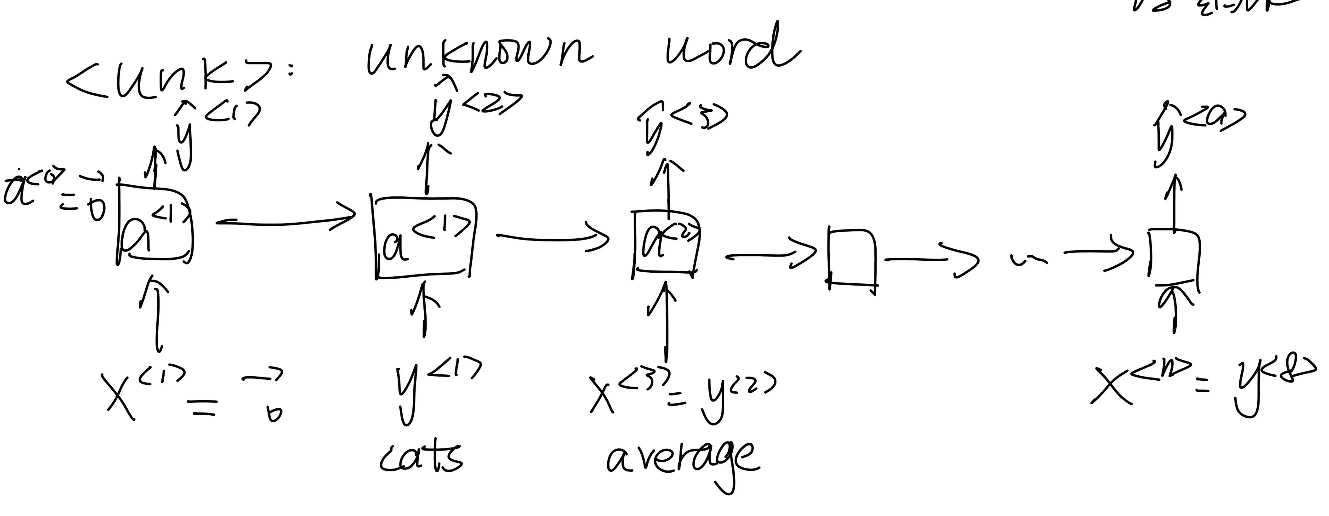
P(The apple and pair salad) $= 3.2 \times 10^{-13}$

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

P(sentence) = ?

$y^{<1>}, y^{<2>}, \ldots, y^{<T_y>}$ <EOS> token: 判断语句的结束

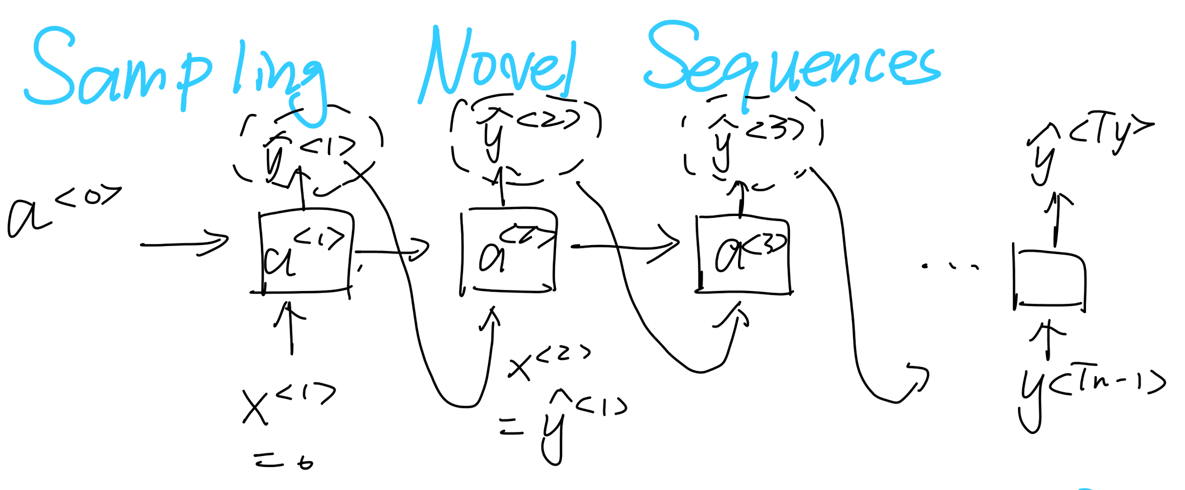
<unk>: unknown word



$P(y^{<1>}, y^{<2>}, y^{<3>}) =$

$P(y^{<1>}) P(y^{<2>} | y^{<1>}) P(y^{<3>} | y^{<1>}, y^{<2>})$

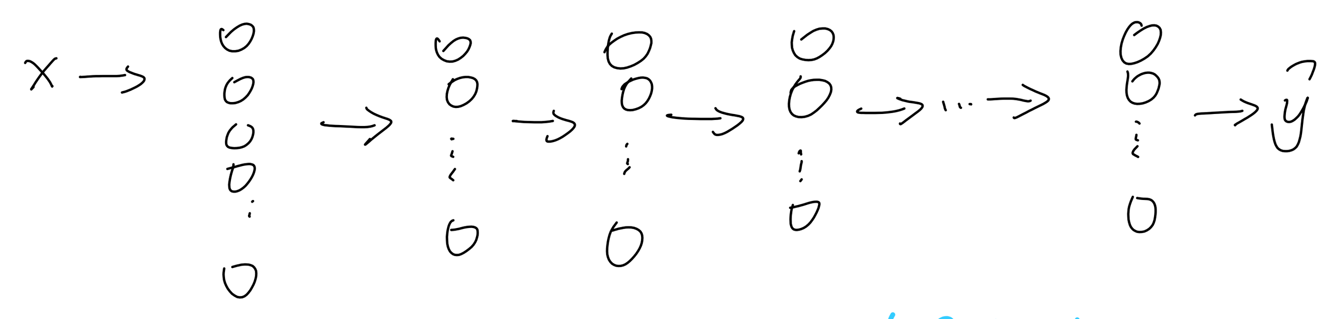
Sampling Novel Sequences



Vanishing Gradients with RNNs

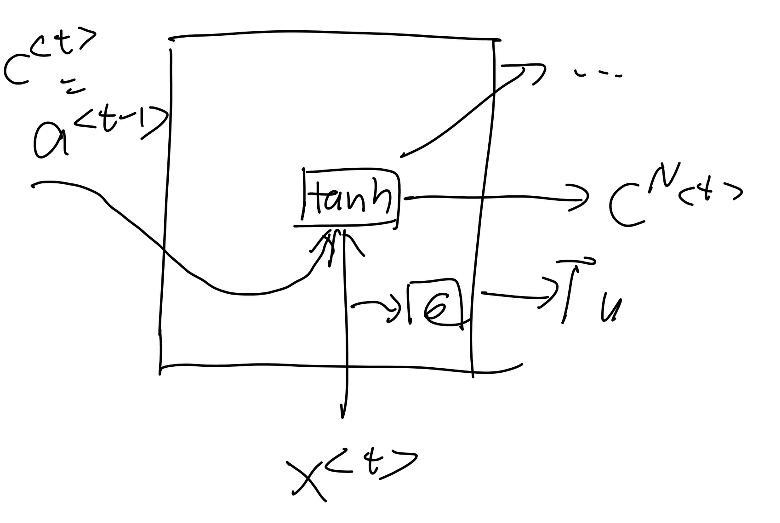
eg: The cat, which ......, was full

The cats, which ....., were full



Gated Recurrent Unit (GRU)

$a^{<t>} = g(W_a[a^{<t-1>}, x^{<t>}] + b_a)$



GRU (simplified)

c = memory cell

$c^{<t>} = a^{<t>}$

$\tilde{c}^{<t>} = \tanh(W_c[c^{<t-1>}, x^{<t>}] + b_c)$

$\Gamma_u = \sigma(W_u[c^{<t-1>}, x^{<t>}] + b_u)$

LSTM

$\tilde{c}^{<t>} = \tanh(W_c[a^{<t-1>}, x^{<t>}] + b_c)$

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

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

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

$c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + \Gamma_f * c^{<t-1>}$

$a^{<t>} = \Gamma_o * c^{<t>}$