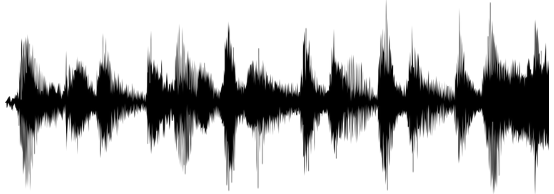
Chapter 08: Recurent Neural Network

Recurrent Neural Networks

deeplearning.ai

- Coursera course: *Sequence models*
- **Week 1: Recurrent Neural Networks**
- Outline:
  - Why sequence models?
  - Notation
  - Recurrent Neural Network Model
  - Backpropagation through time
  - Different types of RNNs
  - Language model and sequence generation
  - Sampling novel sequences
  - Vanishing gradients with RNNs
  - GRU, LSTM, BiRNN, Deep RNN

Why sequence models?

Speech recognition		$\longrightarrow$	"The quick brown fox jumped over the lazy dog."
Music generation	$\emptyset$	$\longrightarrow$	
Sentiment classification	"There is nothing to like in this movie."	$\longrightarrow$	★☆☆☆☆
DNA sequence analysis	AGCCCCTGTGAGGAACTAG	$\longrightarrow$	AGCCCCTGTGAGGAACTAG
Machine translation	Voulez-vous chanter avec moi?	$\longrightarrow$	Do you want to sing with me?
Video activity recognition		$\longrightarrow$	Running
Name entity recognition	Yesterday, Harry Potter met Hermione Granger.	$\longrightarrow$	Yesterday, Harry Potter met Hermione Granger.

Notation:

Motivating example

NLP

x: Harry Potter and Hermione Granger invented a new spell.

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

$T_x = 9$

y: 1 1 0 1 1 0 0 0 0

$y^{<1>}$ $y^{<2>}$ $y^{<3>}$ . . . . $y^{<9>}$

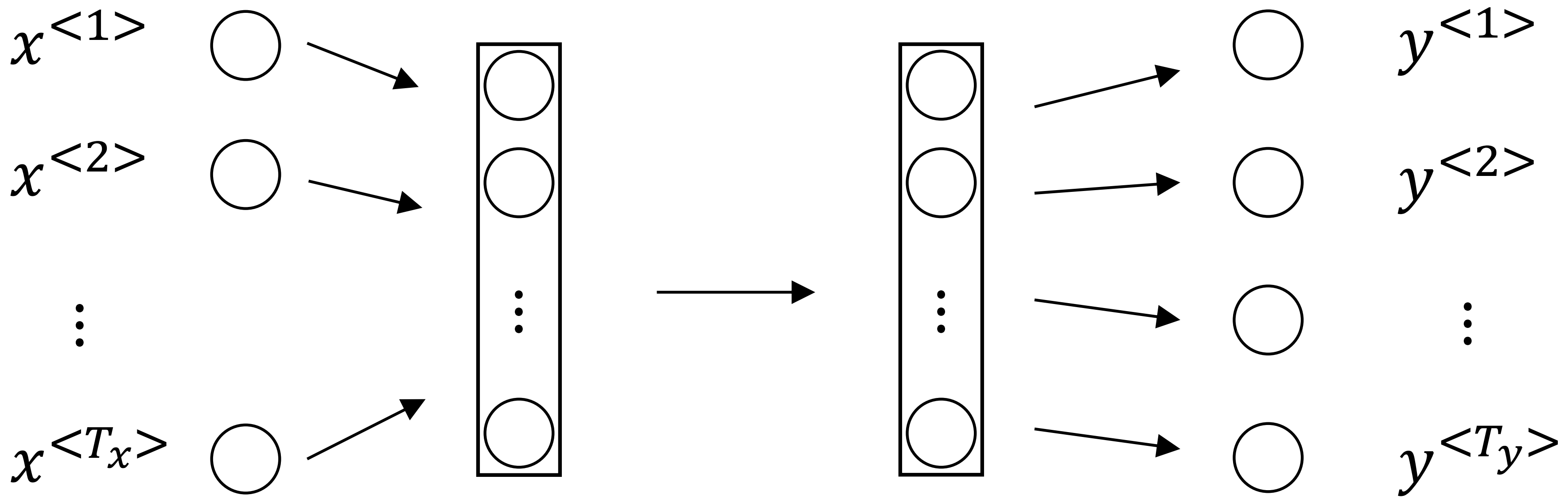
$T_y = 9$

$X^{(i)<t>}$ $T_x^{(i)} = 9$ 15

$y^{(i)<t>}$ $T_y^{(i)}$

Recurrent Neural Network Model

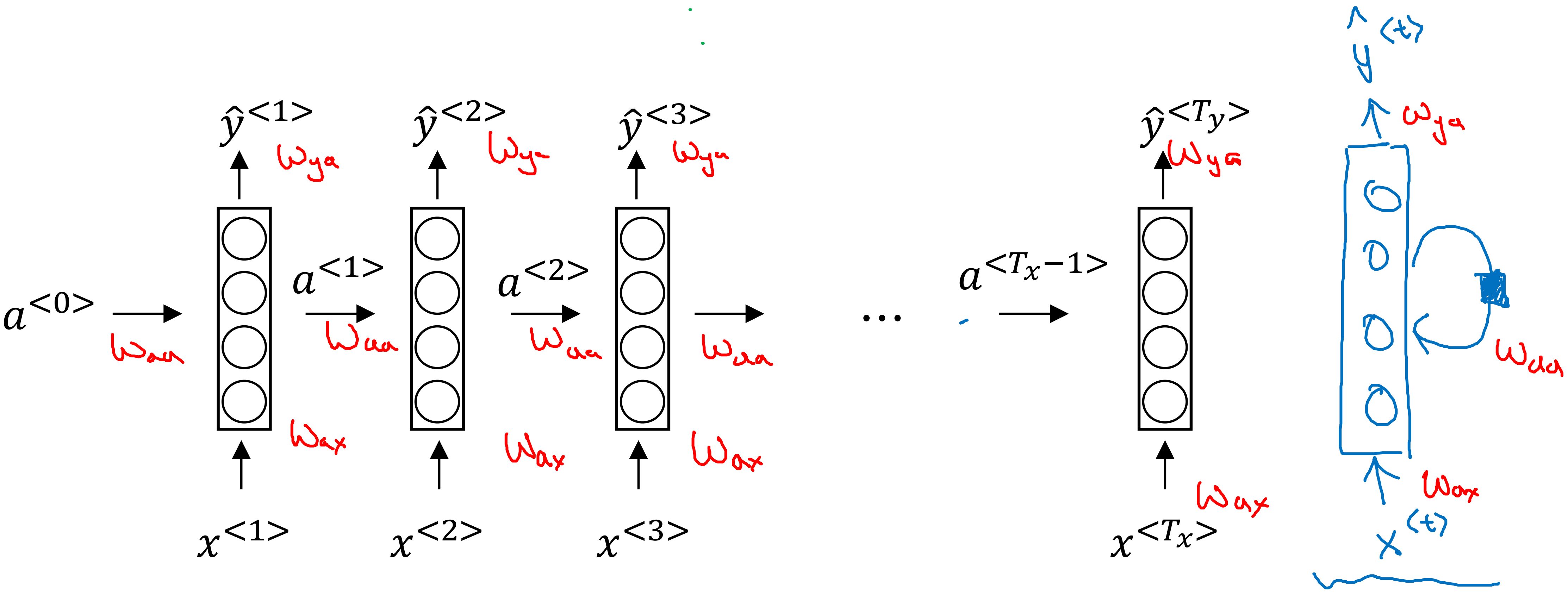
Why not a standard network?



Problems:

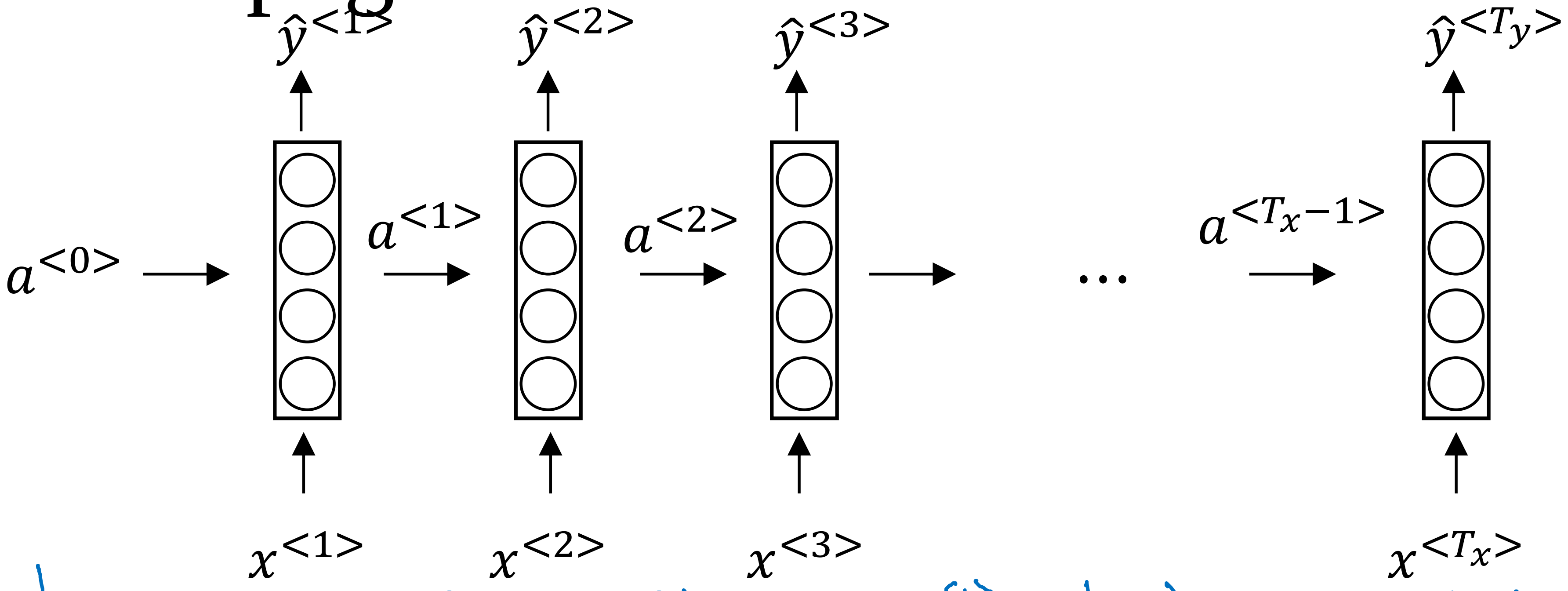
- Inputs, outputs can be different lengths in different examples.
- Doesn’t share features learned across different positions of text.

Recurrent Neural Network Model



Recurrent Neural Network Model

Forward Propagation



$a^{<0>} = \vec{0}.$

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

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

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

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

Recurrent Neural Network Model

Simplified RNN notation

$$a^{<t>} = g(\underbrace{\boxed{W_{aa}}}_{(100,100)} \underbrace{a^{<t-1>}}_{100} + \underbrace{\boxed{W_{ax}}}_{(100,10{,}000)} \underbrace{x^{<t>}}_{10{,}000} + b_a)$$

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

$$\hat{y}^{\langle t \rangle} = g(W_y a^{\langle t \rangle} + b_y)$$

$$a^{\langle t \rangle} = g(W_a [a^{\langle t-1 \rangle}, x^{\langle t \rangle}] + b_a)$$

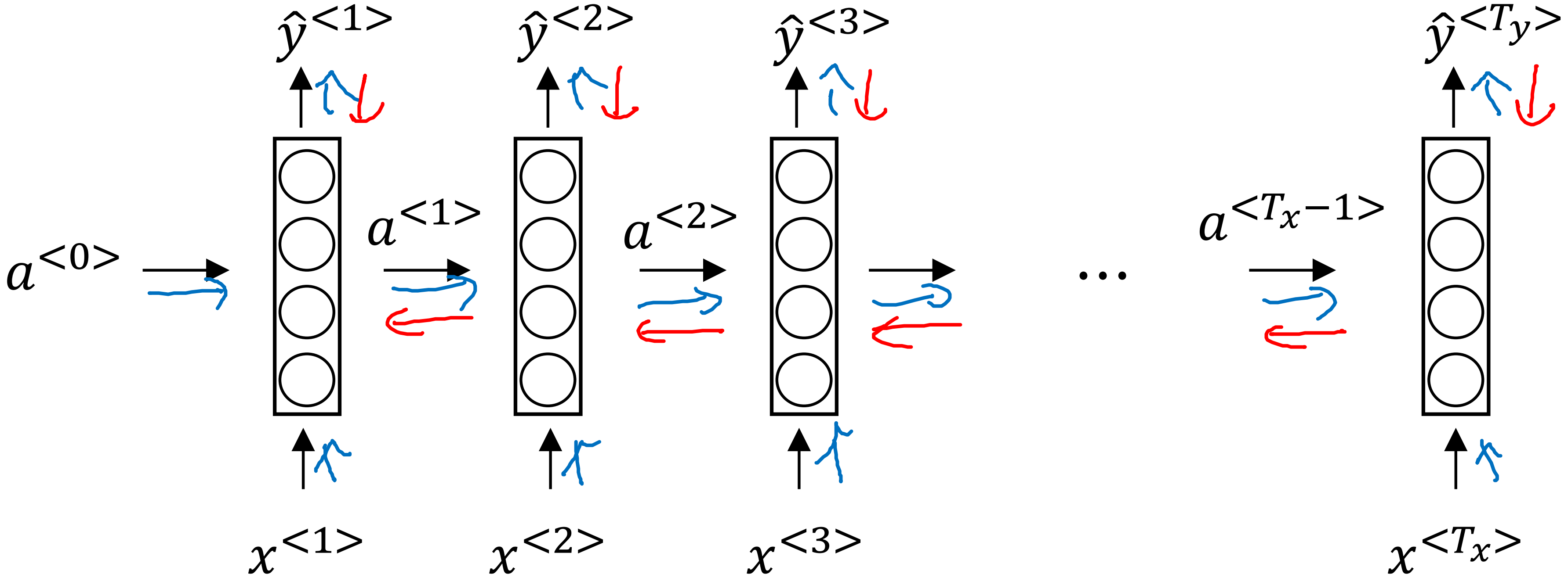
$$\begin{array}{c} \\ 100 \end{array} \underbrace{\begin{bmatrix} W_{aa} & \vdots & W_{ax} \end{bmatrix}}_{\substack{100 \qquad\qquad 10\,000}} = W_a \quad (100, 10100)$$

$$[a^{\langle t-1 \rangle}, x^{\langle t \rangle}] = \begin{bmatrix} a^{\langle t-1 \rangle} \\ \hline x^{\langle t \rangle} \end{bmatrix} \begin{array}{l} 100 \\ 10000 \end{array} \quad 10100$$

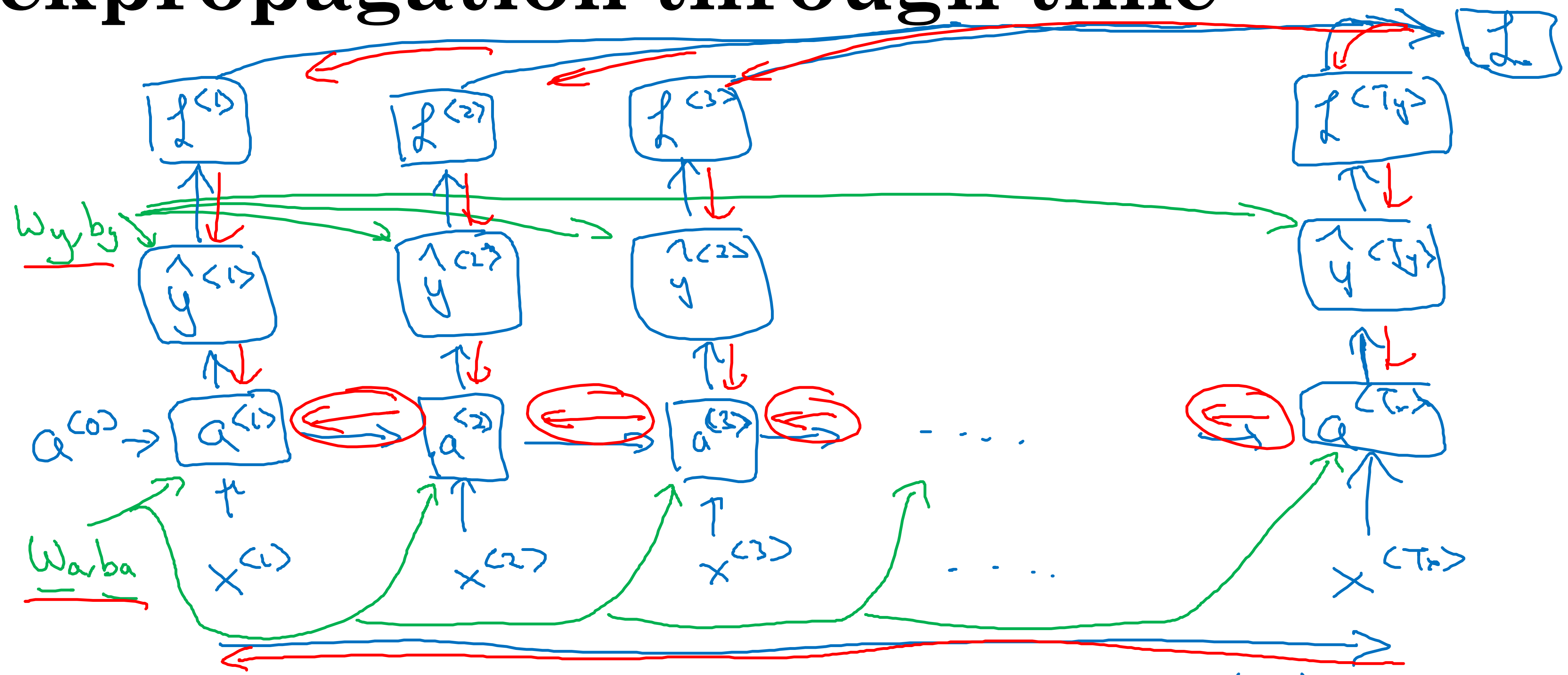
$$[W_{aa} \vdots W_{ax}] \begin{bmatrix} a^{\langle t-1 \rangle} \\ x^{\langle t \rangle} \end{bmatrix} = W_{aa} a^{\langle t-1 \rangle} + W_{ax} x^{\langle t \rangle}$$

Backpropagation through time

Forward propagation and backpropagation



Backpropagation through time

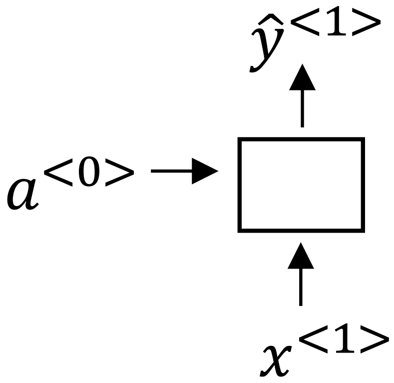


$$\mathcal{L}^{<t>}(\hat{y}^{<t>}, y^{<t>}) = - y^{<t>} \log \hat{y}^{<t>} - (1-y^{<t>}) \log(1-\hat{y}^{<t>})$$

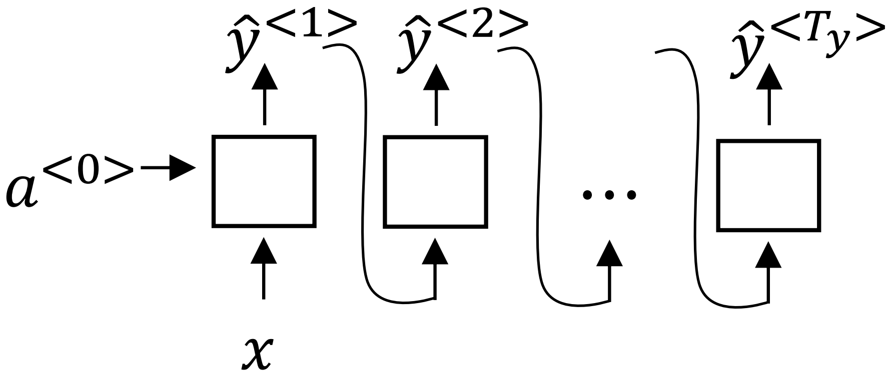
$$\mathcal{L}(\hat{y}, y) = \sum_{t=1}^{T_y} \mathcal{L}^{<t>}(\hat{y}^{<t>}, y^{<t>})$$

Backpropagation through time

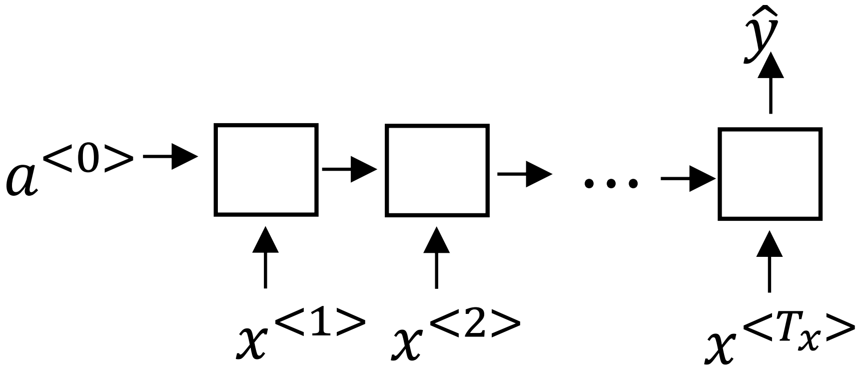
Different types of RNNs



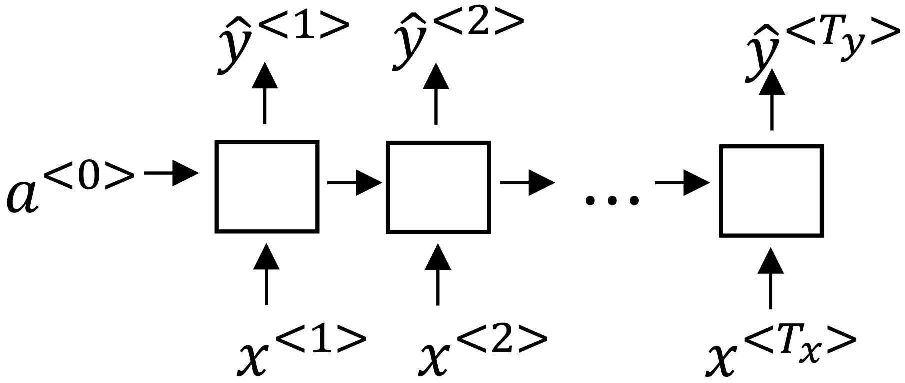
One to one



One to many

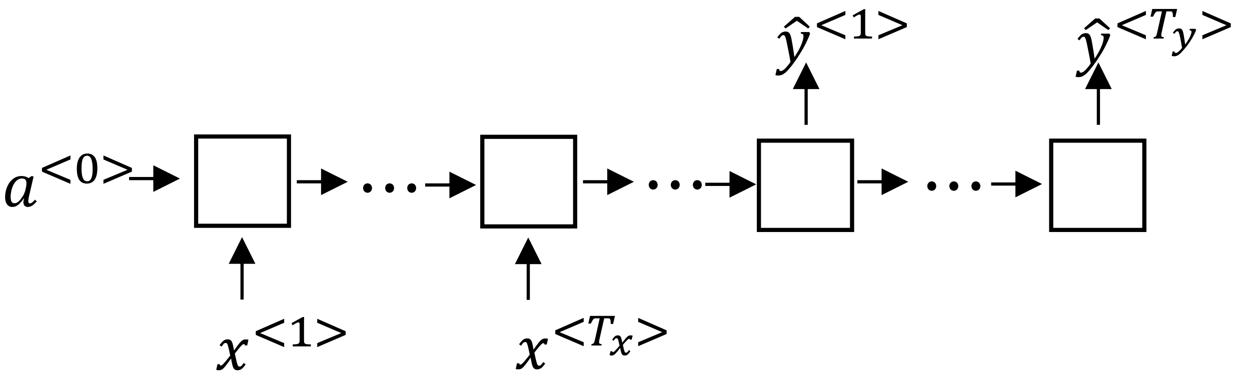


Many to one



Many to many

$T_x = T_y$



Many to many

Language model and sequence generation

What is language modelling?

Speech recognition

The apple and pair salad.

→ The apple and pear salad.

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

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

$P(\text{sentence}) = ?$

$P(y^{<1>}, y^{<2>}, \ldots, y^{<T_y>})$

Language model and sequence generation

Language modelling with an RNN

Training set: large corpus of english text.

Tokenize

Cats average 15 hours of sleep a day. <EOS>

$y^{<1>}$ $y^{<2>}$ $y^{<3>}$ . . . $y^{<8>}$ $y^{<9>}$

$x^{<t>} = y^{<t-1>}$

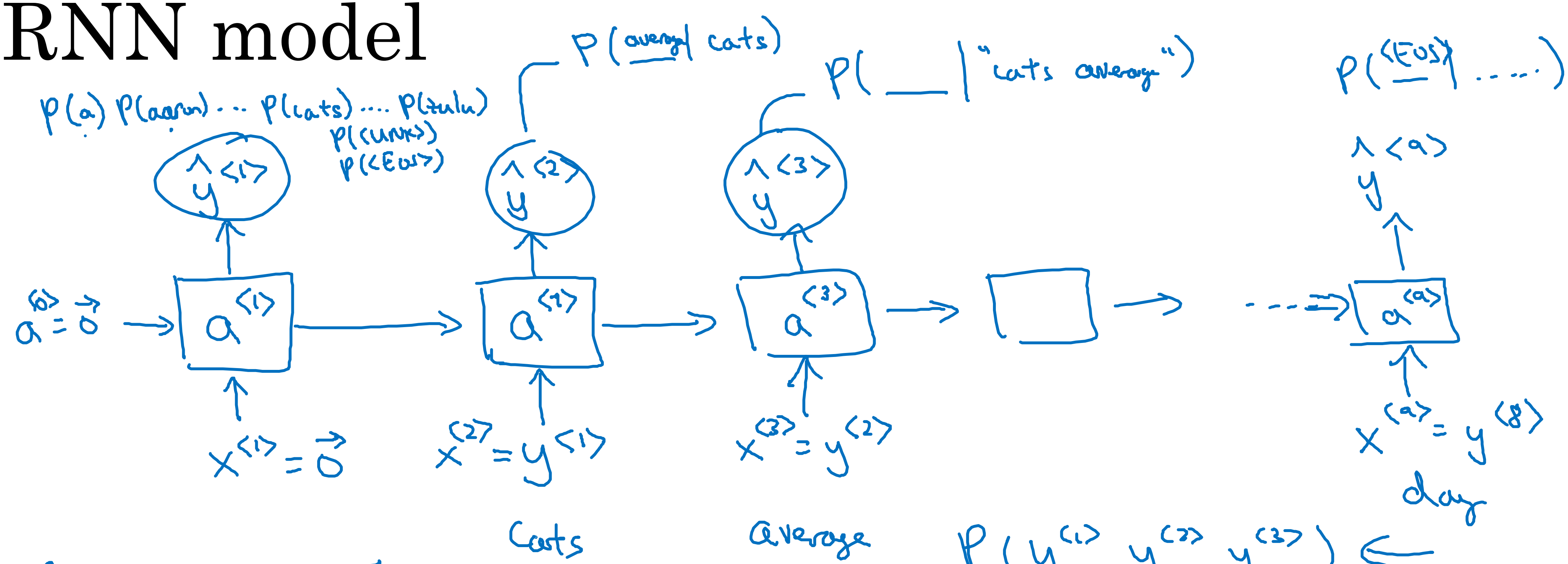
The Egyptian ~~Mau~~ is a bread of cat. <EOS>

<UNK>

10,000

Language model and sequence generation

RNN model



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

Cats average 15 hours of sleep a day. <EOS>

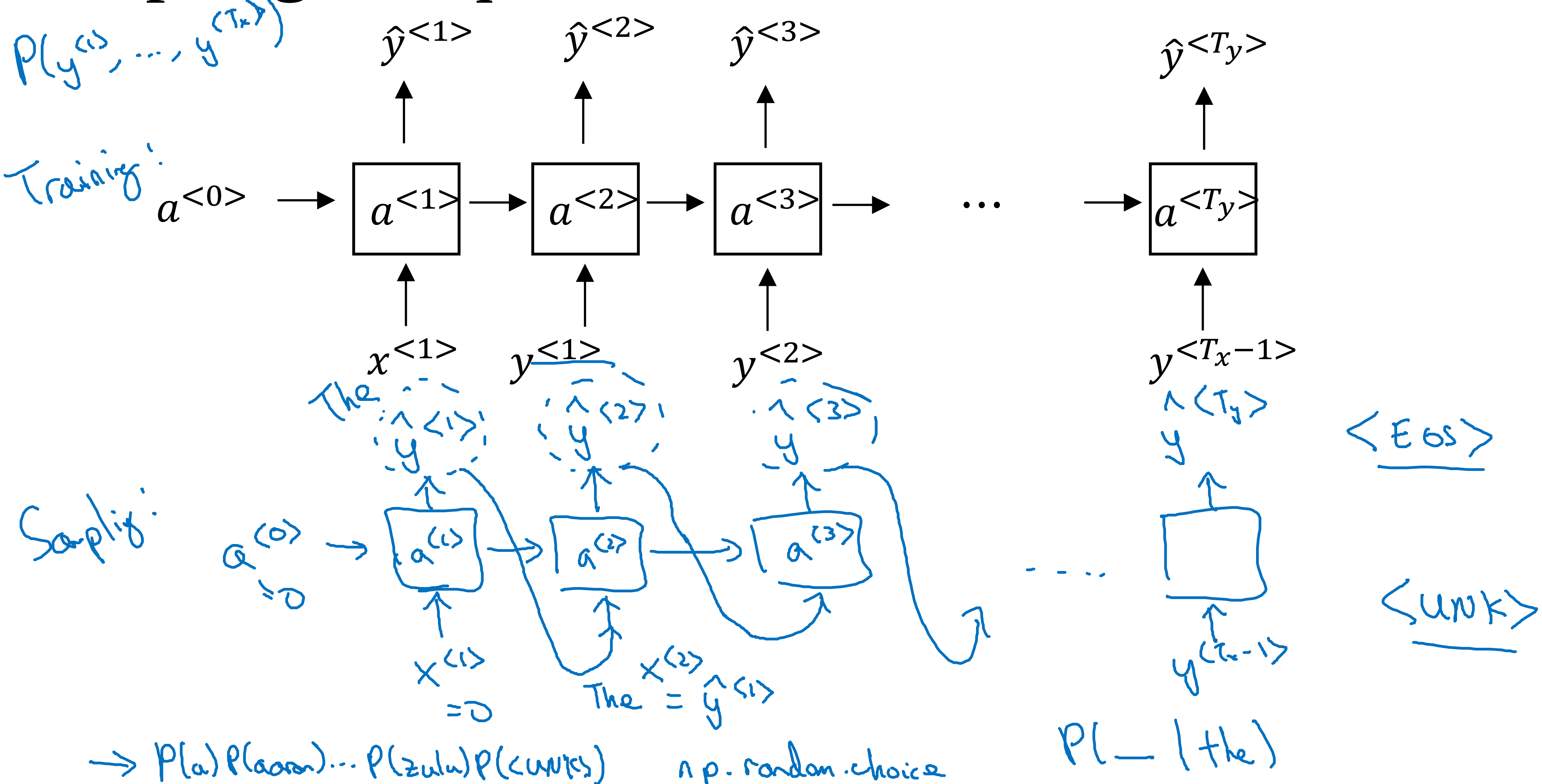
$= P(y^{<1>})\, P(y^{<2>} \mid y^{<1>})\, P(y^{<3>} \mid y^{<1>}, y^{<2>})$

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

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

Sampling novel sequences

Sampling a sequence from a trained RNN



Sampling novel sequences

Character-level language model

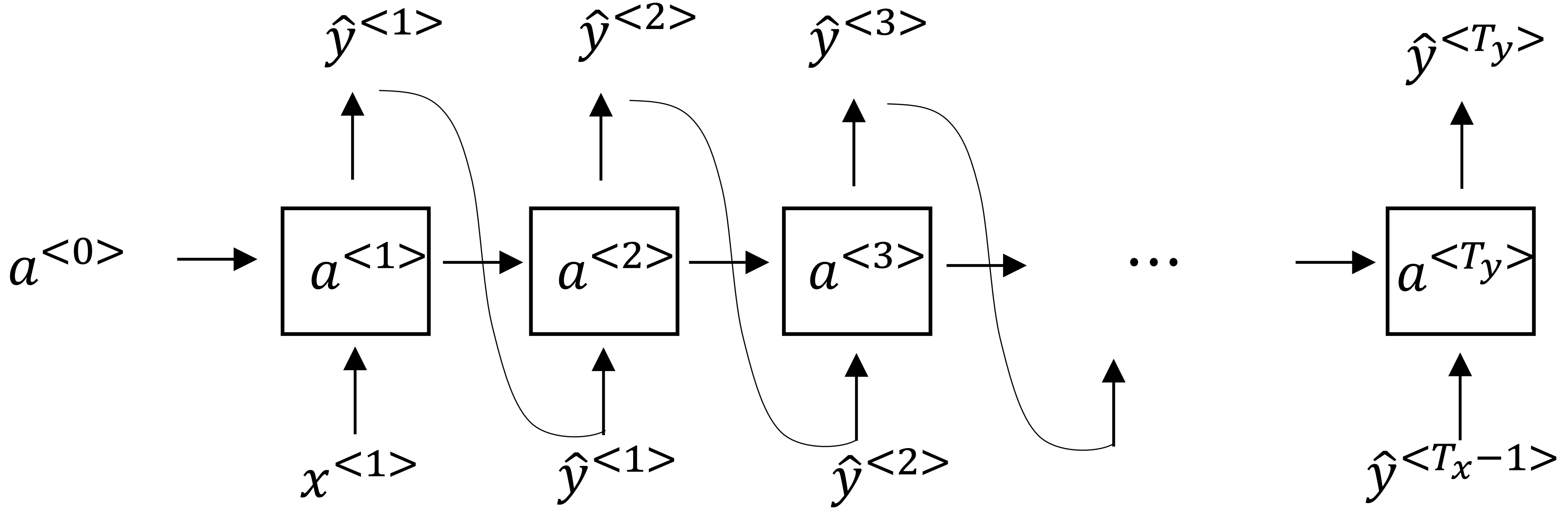
Vocabulary = [a, aaron, …, zulu, <UNK>]

Vocabulary = [a, b, c, …, z, ␣, ., ,, ;, 0, …, 9, A, …, Z]

$y^{<1>}$ $y^{<2>}$ $y^{<3>}$ $y^{<4>}$

Cat average …

Mau



Sampling novel sequences

Sequence generation

News

President enrique peña nieto, announced sench's sulk former coming football langston paring.

"I was not at all surprised," said hich langston.

"Concussion epidemic", to be examined.

The gray football the told some and this has on the uefa icon, should money as.

Shakespeare

The mortal moon hath her eclipse in love.

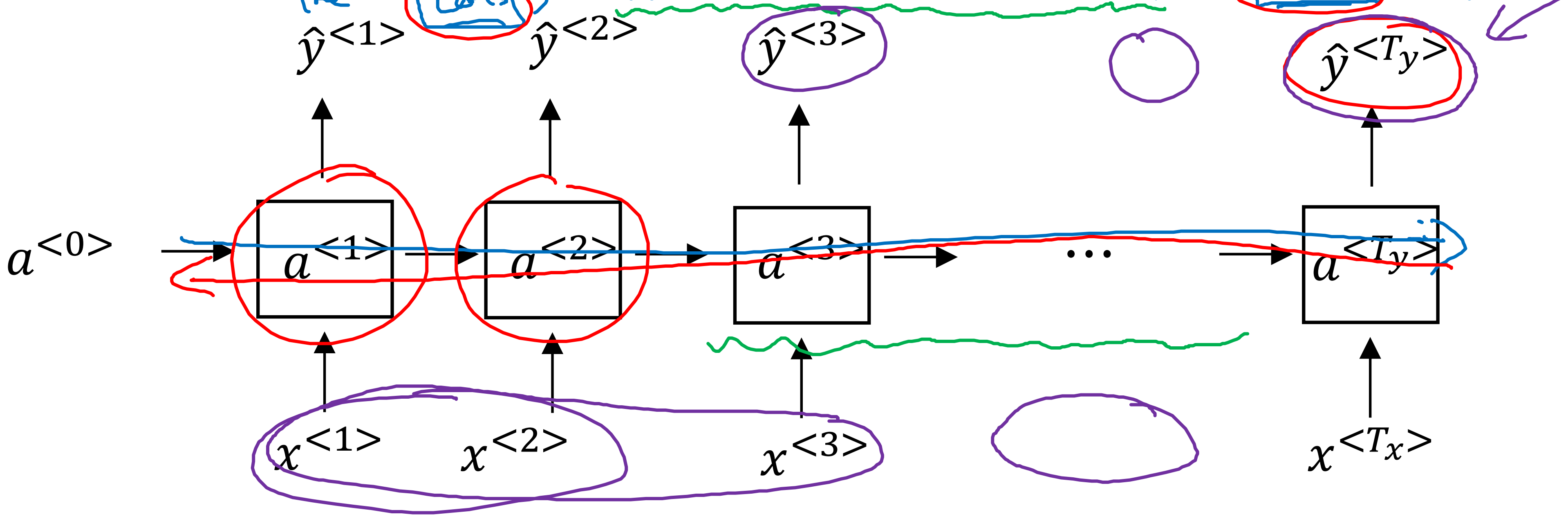
And subject of this thou art another this fold.

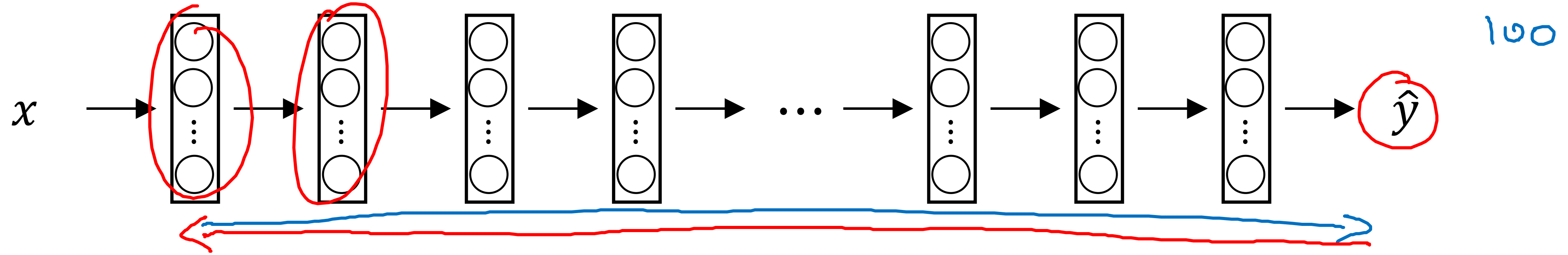
When besser be my love to me see sabl's.

For whose are ruse of mine eyes heaves.

Vanishing gradients with RNNs

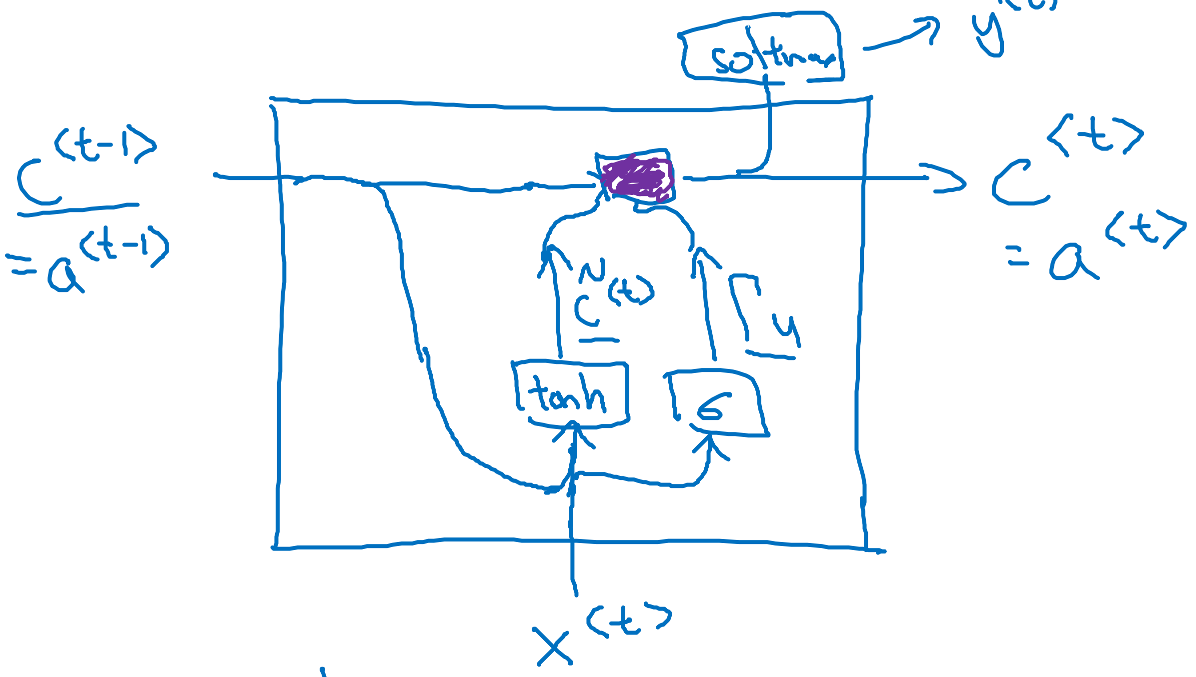
The cat, which already ate ...., was full
The cats, wh . . . ., were full





Exploding gradients. NaN Gradient clipping

Gated Recurrent Unit (GRU) (simplified)



c = memory cell

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

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

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

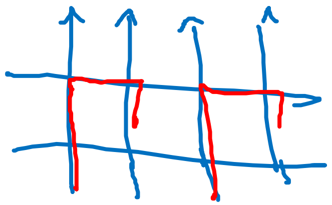
"update"

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

$\Gamma_u = 1$

element-wise

$\Gamma_u = 0.000001$



Gate

$\Gamma_u = 1$ $\Gamma_u = 0$ $\Gamma_u = 0$ $\Gamma_u = 0$ ......

$c^{<t>} = 1$ .... $c^{<t>} = 1$

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]

Full GRU

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

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

r $$\Gamma_r = \sigma(W_r[\, c^{<t-1>}, x^{<t>}] + b_r)$$

LSTM

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

The cat, which ate already, was full.

GRU and LSTM

GRU

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

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

$$\Gamma_r = \sigma(W_r[\, c^{<t-1>}, x^{<t>}] + b_r)$$

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

Γ_f

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

LSTM

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

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

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

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

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

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

[Hochreiter & Schmidhuber 1997. Long short-term memory]

LSTM in pictures

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

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

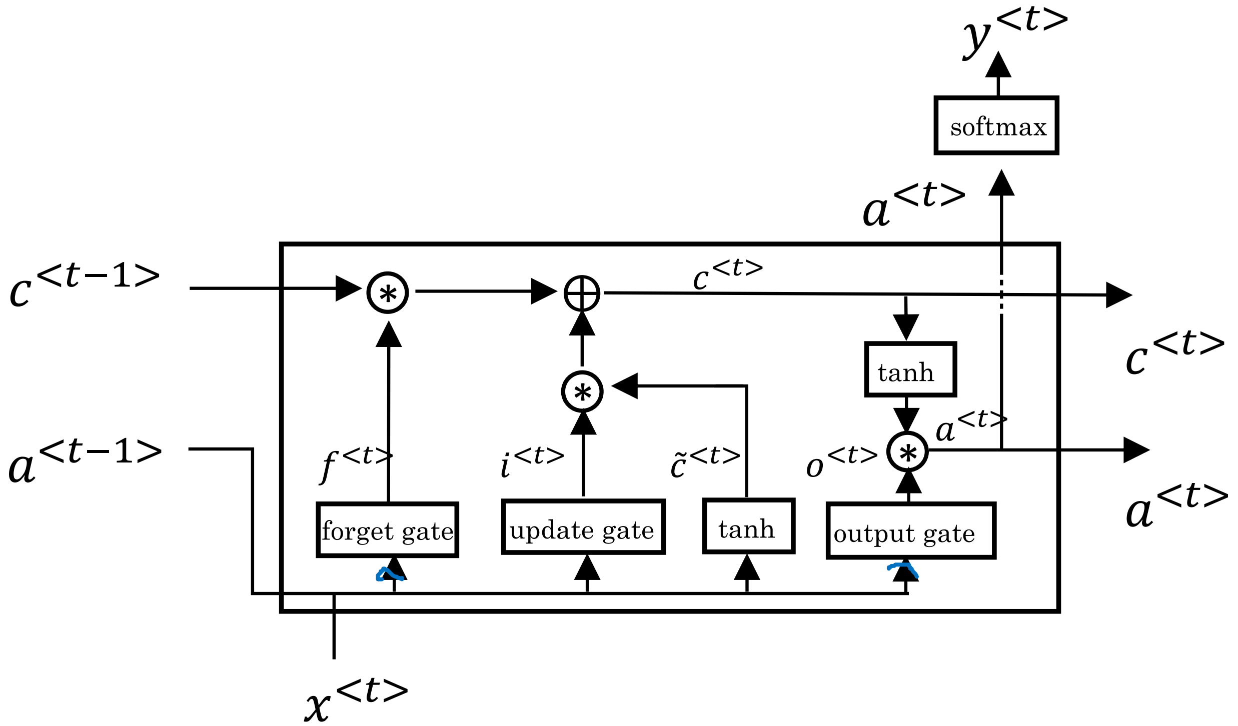
$$\Gamma_f = \sigma(W_f[\,a^{<t-1>}, x^{<t>}] + b_f)$$

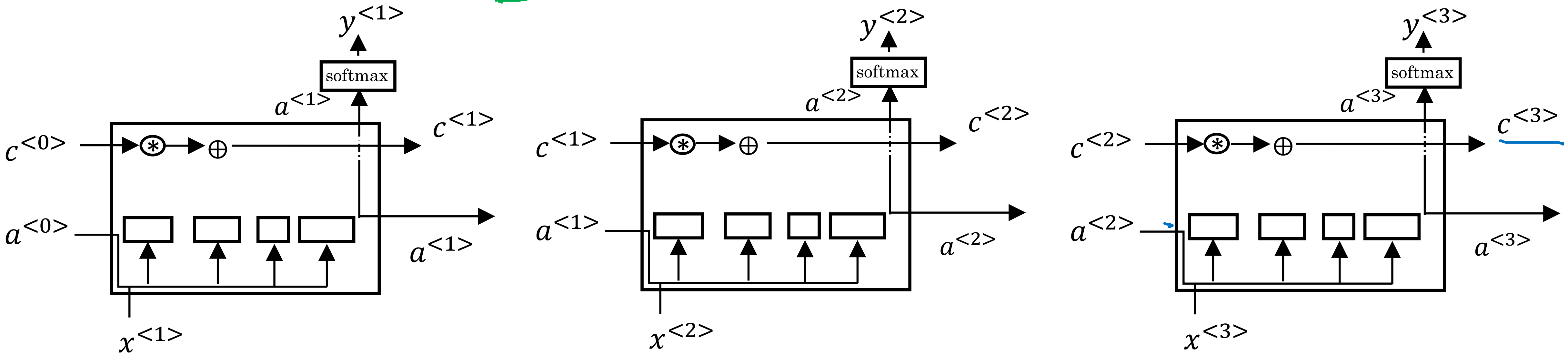
$$\Gamma_o = \sigma(W_o[\,a^{<t-1>}, x^{<t>}] + b_o)$$

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

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

$\boxed{c^{<t-1>}}$



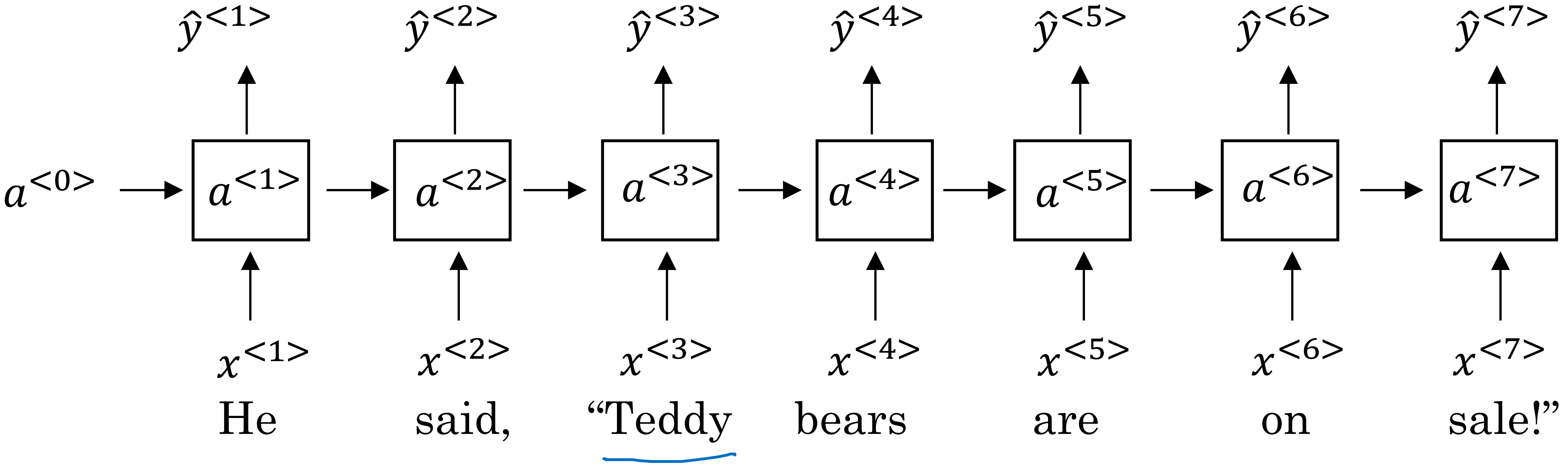


Bidirectional RNN

Getting information from the future

He said, “Teddy bears are on sale!”

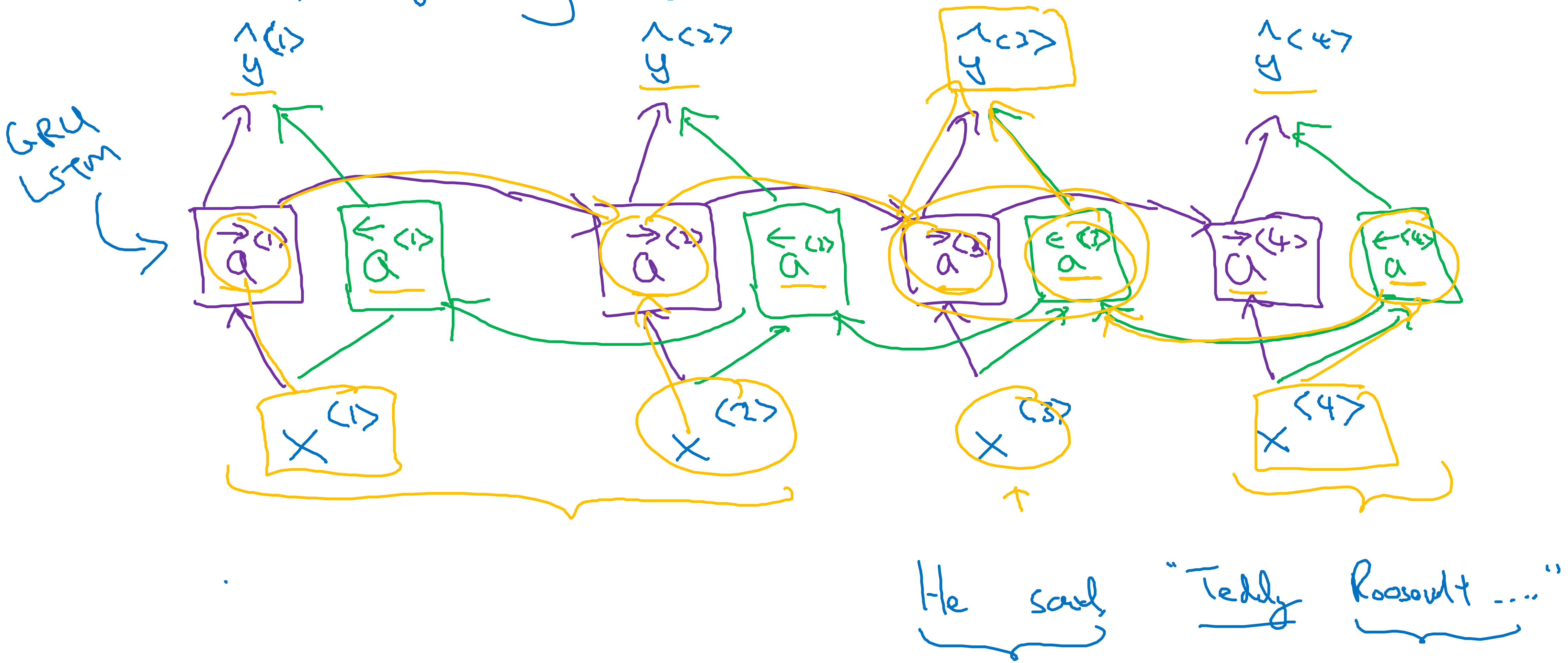
He said, “Teddy Roosevelt was a great President!”



RNN
GRU
LSTM

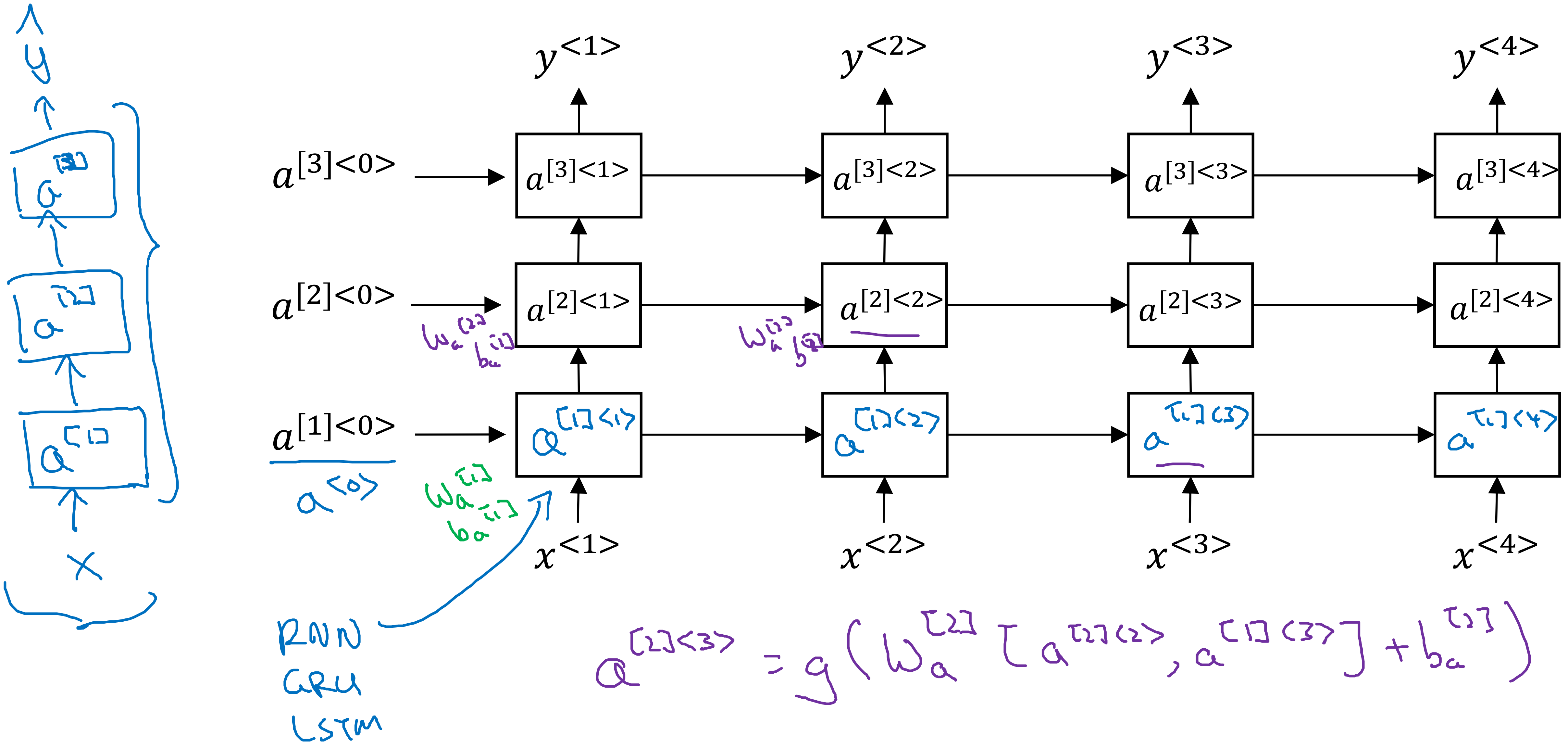
Bidirectional RNN (BRNN)

$$\hat{y}^{<t>} = g(W_y[\overrightarrow{a}^{<t>}, \overleftarrow{a}^{<t>}] + b_y)$$



BRNN w/ LSTM

Deep RNN example



$$a^{[2]<3>} = g\left(W_a^{[2]}\left[a^{[2]<2>}, a^{[1]<3>}\right] + b_a^{[2]}\right)$$