Copyright Notice

Recurrent Neural Networks

deeplearning.ai

Why sequence models?

Examples of sequence data
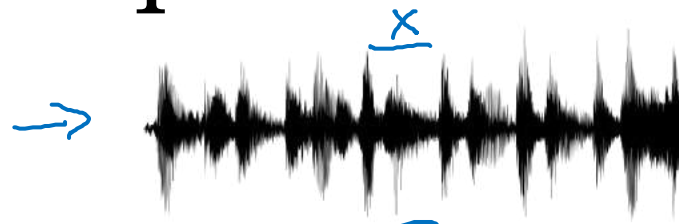
Speech recognition
x → y
"The quick brown fox jumped over the lazy dog."

Music generation
$\emptyset$ →

Sentiment classification
"There is nothing to like in this movie." → ★☆☆☆☆

DNA sequence analysis
AGCCCCTGTGAGGAACTAG → AGCCCCTGTGAGGAACTAG

Machine translation
Voulez-vous chanter avec moi? → Do you want to sing with me?

Video activity recognition
→ Running

Name entity recognition
Yesterday, Harry Potter met Hermione Granger. → Yesterday, Harry Potter met Hermione Granger.

Andrew Ng

Recurrent Neural Networks

Notation

deeplearning.ai

Motivating example

NLP

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

$\rightarrow$ $x^{\langle 1 \rangle}$ $x^{\langle 2 \rangle}$ $x^{\langle 3 \rangle}$ $- - - \cdots$ $x^{\langle t \rangle}$ $- - \cdots$ $x^{\langle 9 \rangle}$

$T_x = 9$

$\rightarrow$ y: 1   1   0   1   1   0   0   0   0

$y^{\langle 1 \rangle}$ $y^{\langle 2 \rangle}$ $y^{\langle 3 \rangle}$ $- - \cdots$ $y^{\langle 9 \rangle}$

$T_y = 9$

$x^{(i)\langle t \rangle}$ $T_x^{(i)} = 9$ 15

$y^{(i)\langle t \rangle}$ $T_y^{(i)}$
$\uparrow$
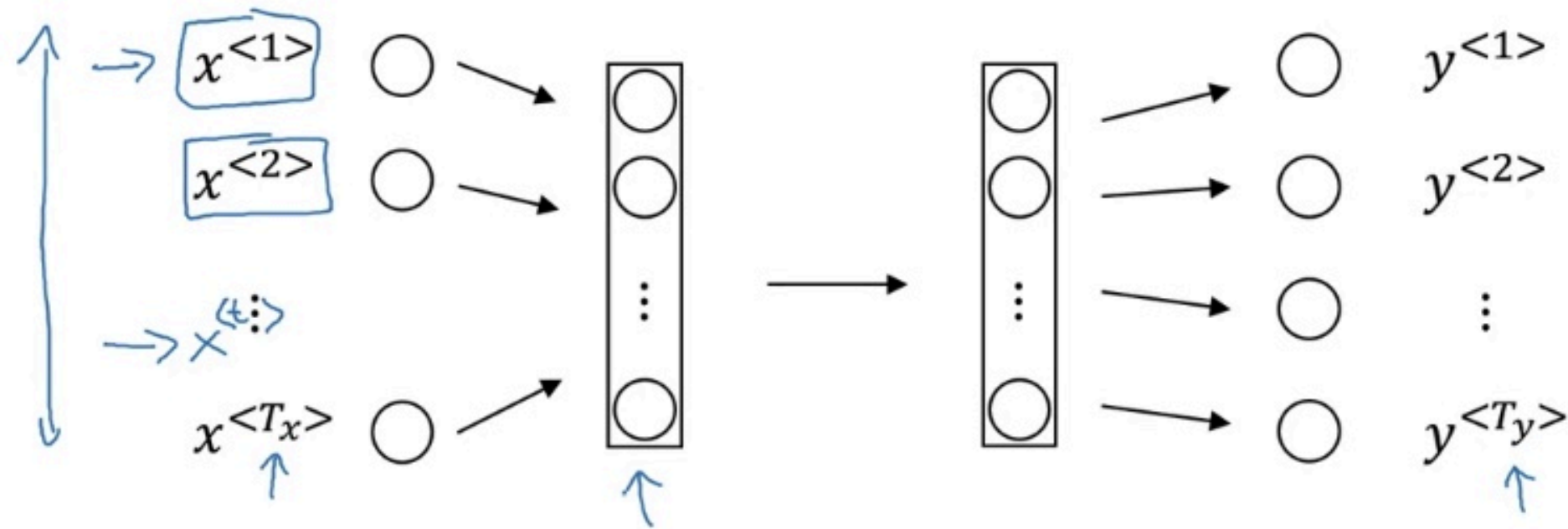
Andrew Ng

Representing words

$x^{<t>}$ (x, y)

$x \longrightarrow y$

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

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

Vocabulary

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

<UNK>   10,000

$$
\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \leftarrow 4075
$$

$$
\begin{bmatrix} 0 \\ \vdots \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \leftarrow 6830
$$

$$
\begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \leftarrow 367
$$

10,000

$$
\begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}
$$

One-hot

Andrew Ng

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.

Representing words

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

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

And = 367
Invented = 4700
A = 1
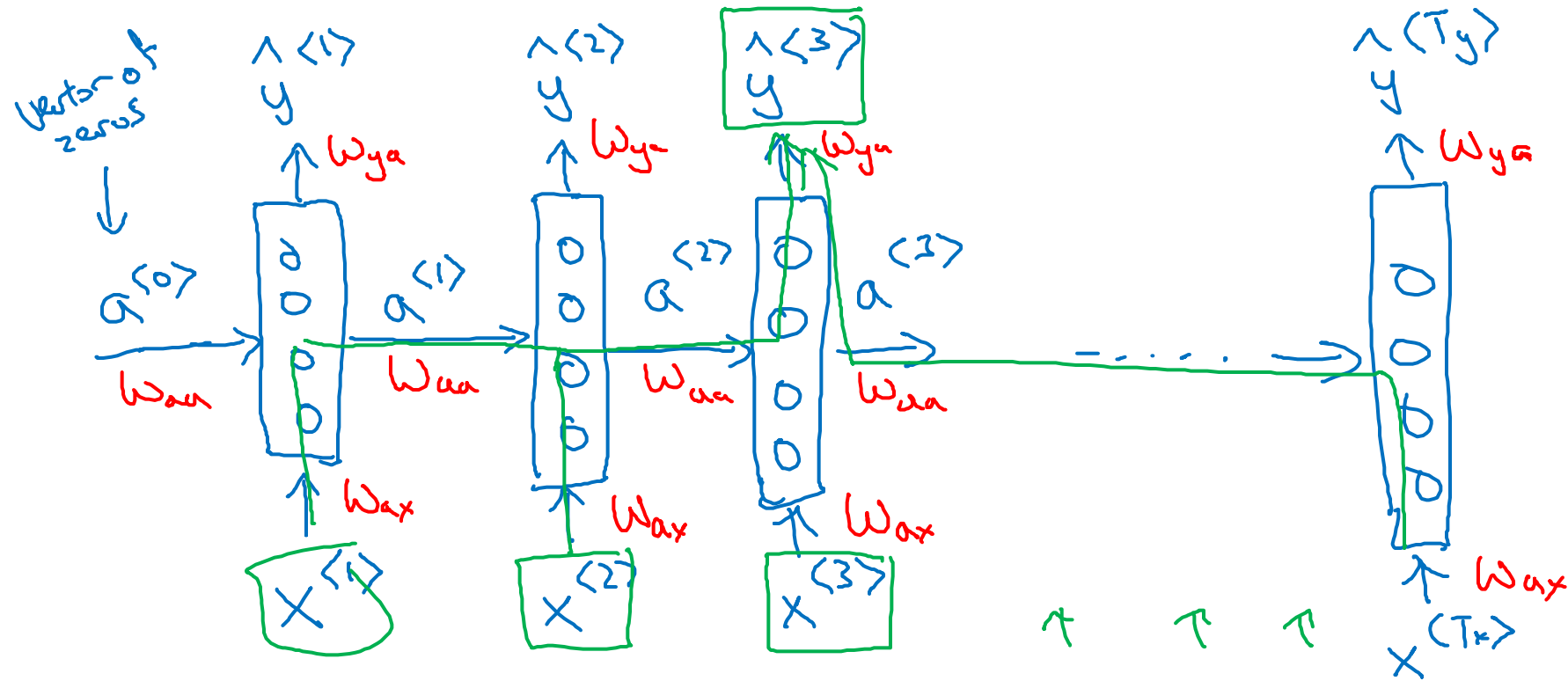New = 5976
Spell = 8376
Harry = 4075
Potter = 6830
Hermione = 4200
Gran... = 4000

Why not a standard network?
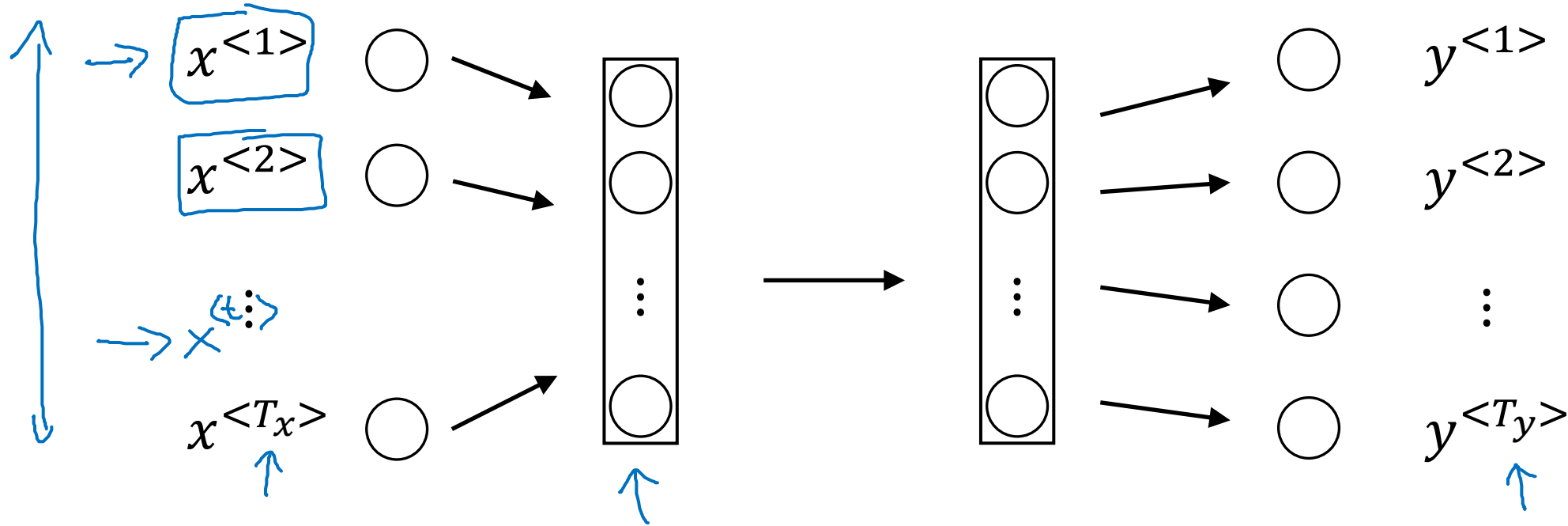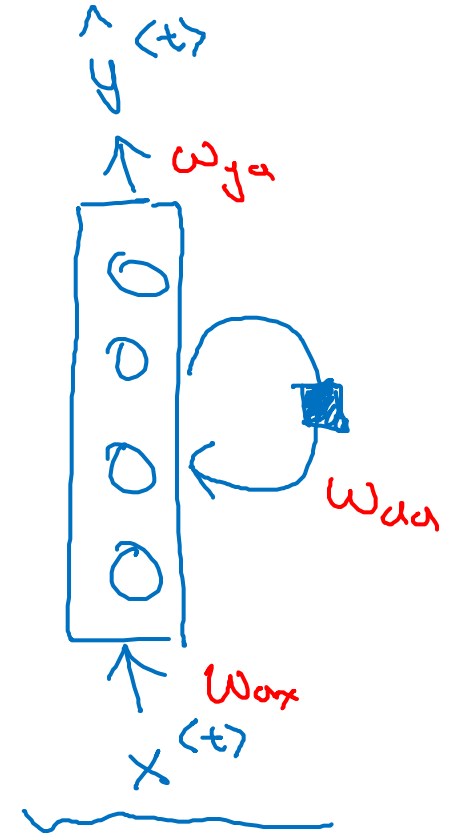

Problems:

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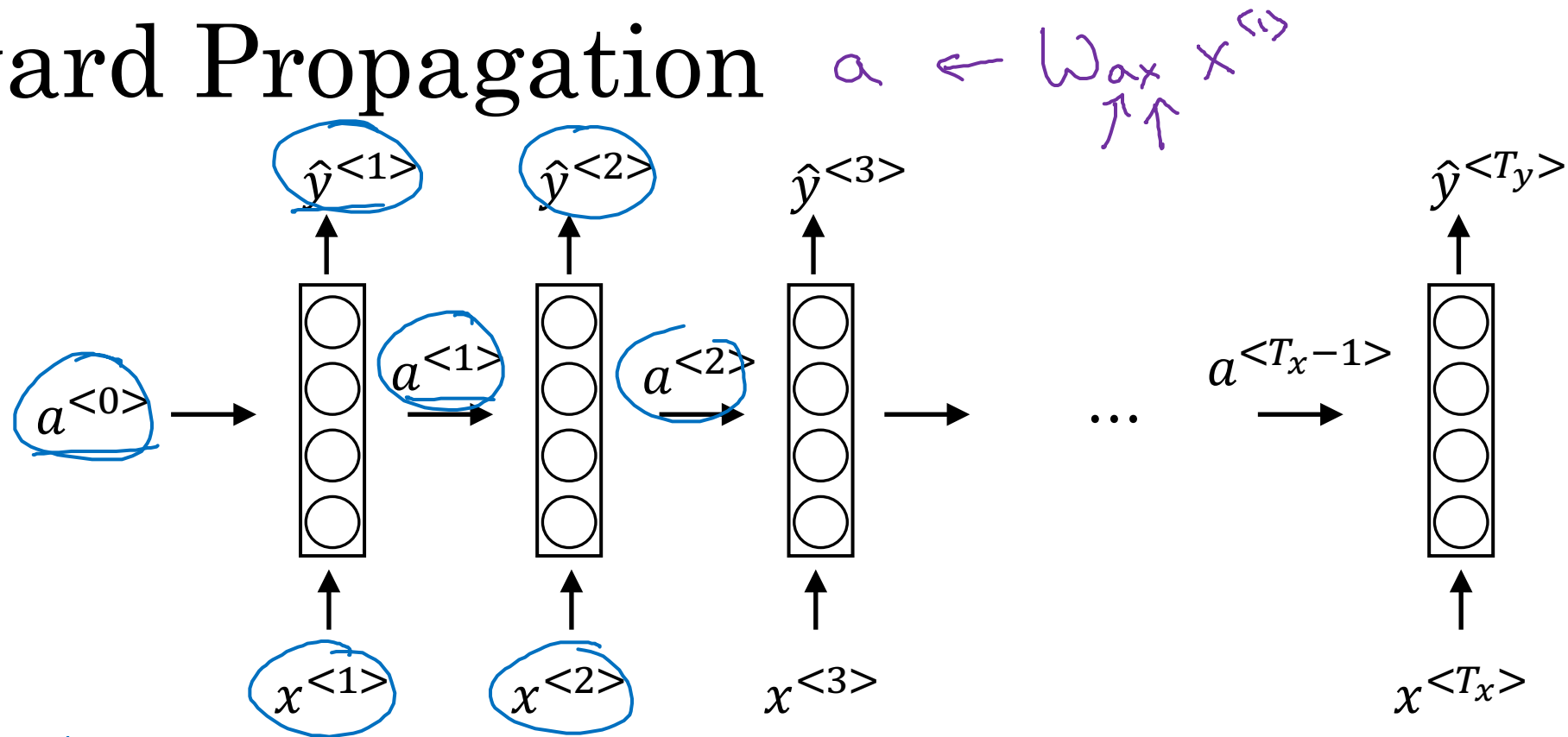
Recurrent Neural Networks

$T_x = T_y$

Bidirectional RNN (BRNN)

He said, "Teddy Roosevelt was a great President."

He said, "Teddy bears are on sale!"

Andrew Ng

Forward Propagation

$$a \leftarrow W_{ax} x^{(1)}$$

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

$$a^{<1>} = g_1(W_{aa} a^{<0>} + W_{ax} x^{<1>} + b_a) \leftarrow \text{tanh} / \text{ReLu}$$

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

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

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

Andrew Ng

Simplified RNN notation

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

100

(100,100)   10,000

(100,10,000)

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

$$y^{<t>} = g(W_y a^{<t>} + b_y)$$

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

$$100 \begin{bmatrix} W_{aa} & ; & W_{ax} \end{bmatrix} = W_a$$

100   10 000

(100, 10100)

$$[a^{<t-1>}, x^{<t>}] = \begin{bmatrix} a^{<t-1>} \\ x^{<t>} \end{bmatrix}$$

100

10000   10100

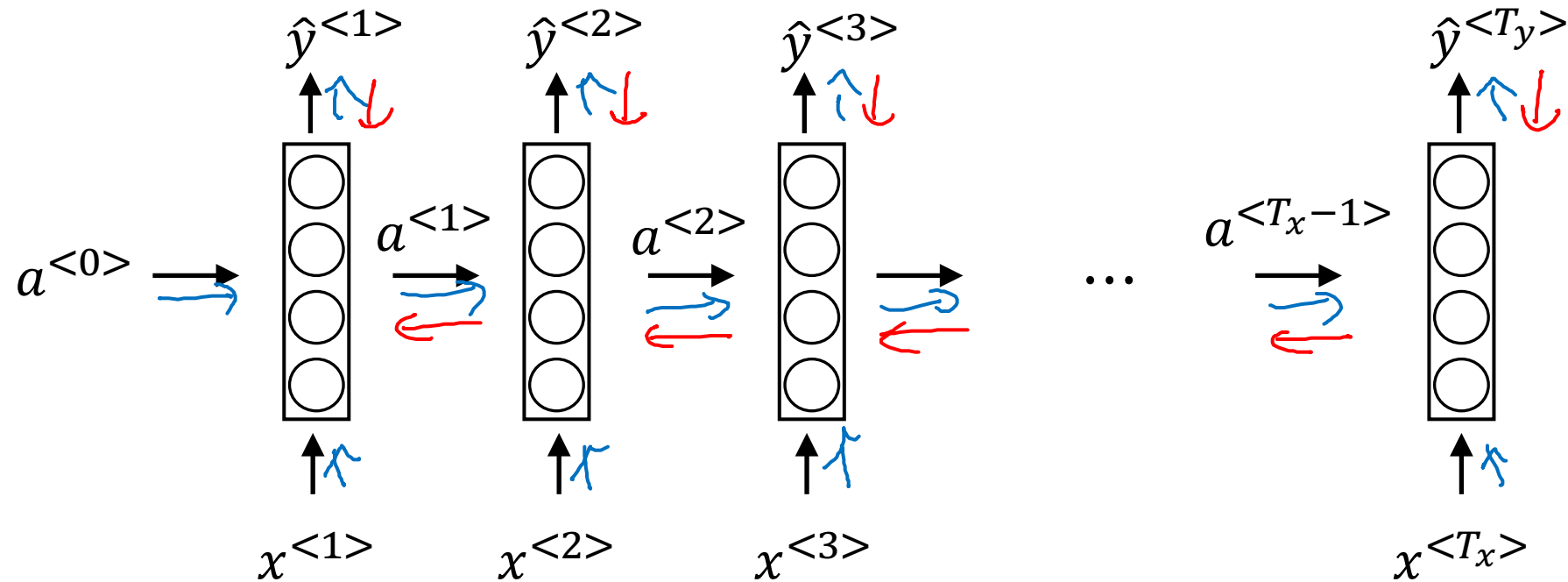$$[W_{aa} ; W_{ax}] \begin{bmatrix} a^{<t-1>} \\ x^{<t>} \end{bmatrix} = W_{aa} a^{<t-1>} + W_{ax} x^{<t>}$$
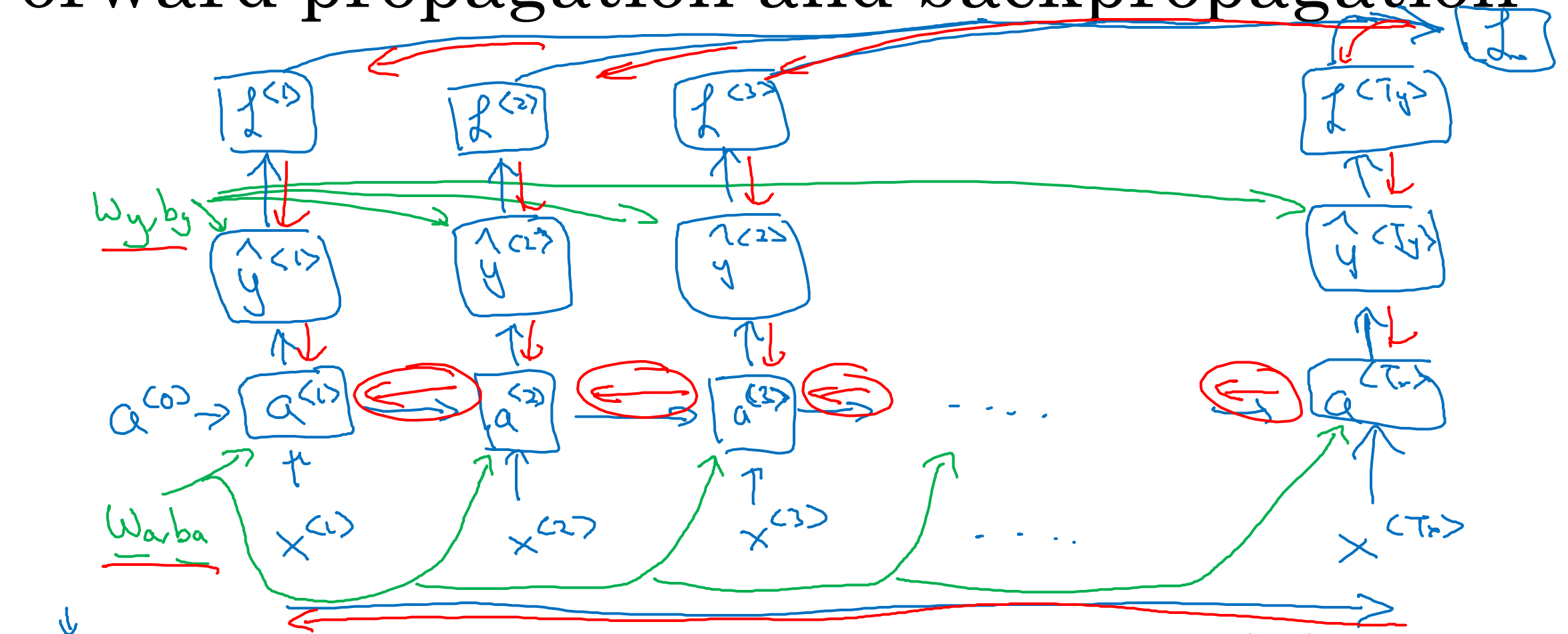
Andrew Ng

Recurrent Neural Networks

deeplearning.ai

Backpropagation through time

Forward propagation and backpropagation

Andrew Ng

Forward propagation and backpropagation

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

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

Backpropagation through time
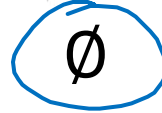
Andrew Ng

Recurrent Neural Networks

deeplearning.ai

Different types of RNNs

Examples of sequence data

T_x T_y

x y

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.

Andrew Ng

Examples of RNN architectures
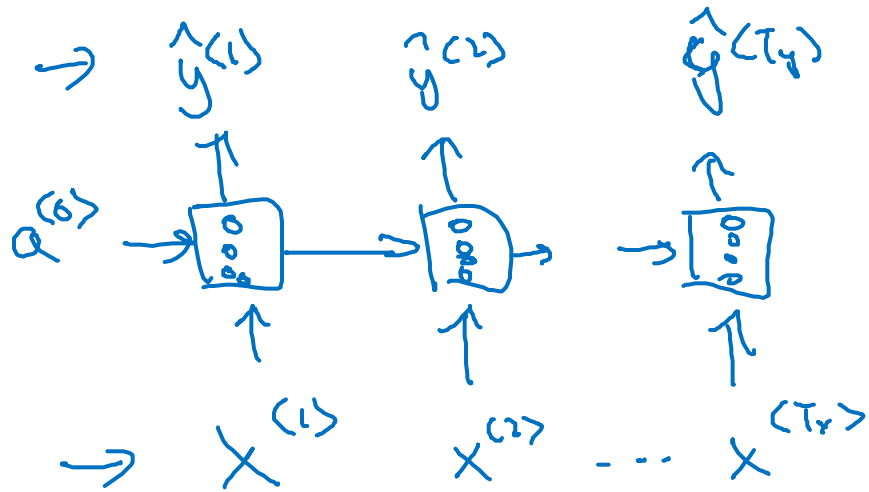
$T_x = T_y$

Sentiment classification
$x = text$
$y = 0/1$ $1 \cdots 5$
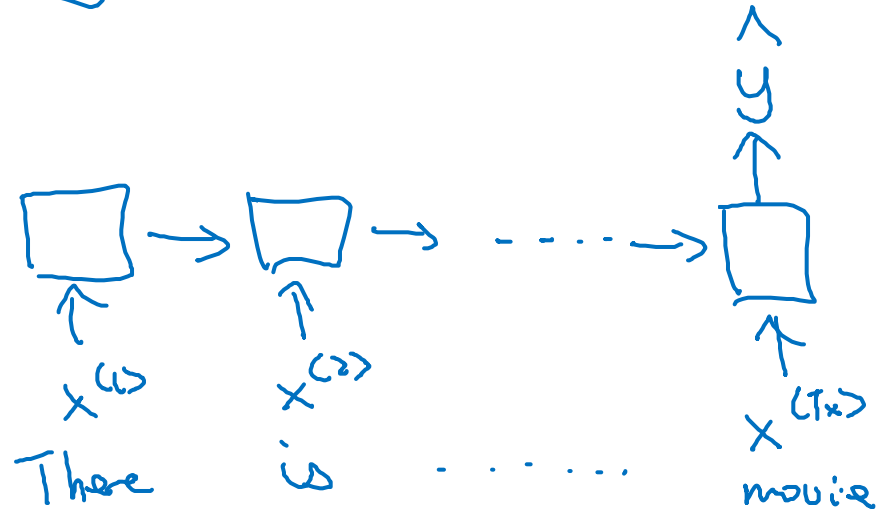
Many-to-many
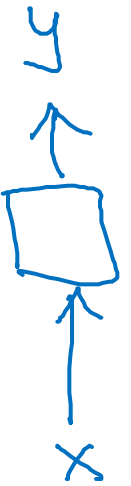
Many-to-one

one-to-one

Examples of RNN architectures

Music generation

$x \rightarrow y^{<1>} y^{<2>} \dots y^{<T_y>}$

$\hat{y}^{<1>} \quad \hat{y}^{<2>} \quad \hat{y}^{<3>} \quad \hat{y}^{<T_y>}$

$a^{<0>}$

x

One-to-many

$x = \phi$

Machine translation

encoder

decoder

$a^{<0>} \rightarrow \dots \rightarrow$

$x^{<1>} \quad x^{<T_n>}$

$\hat{y}^{<1>} \quad \hat{y}^{<T_y>}$
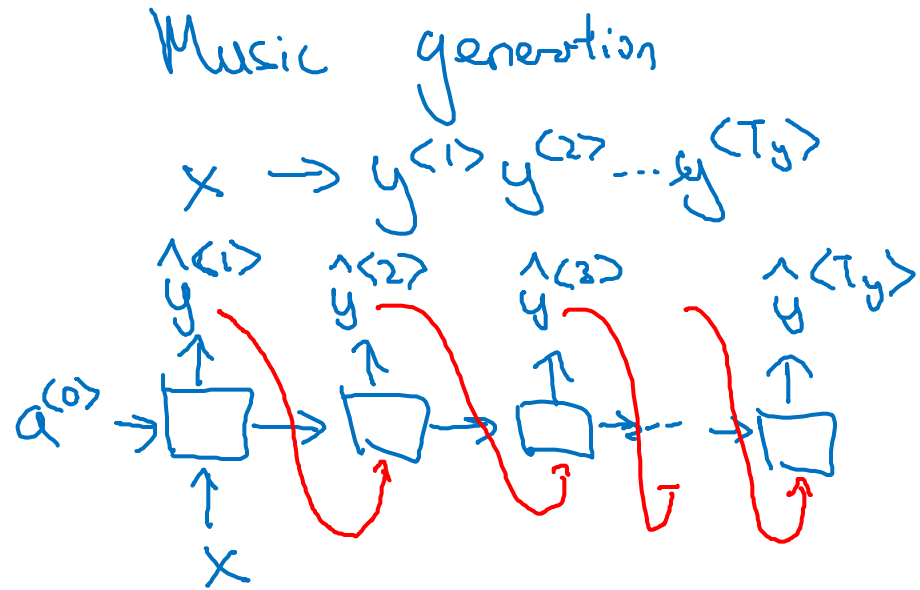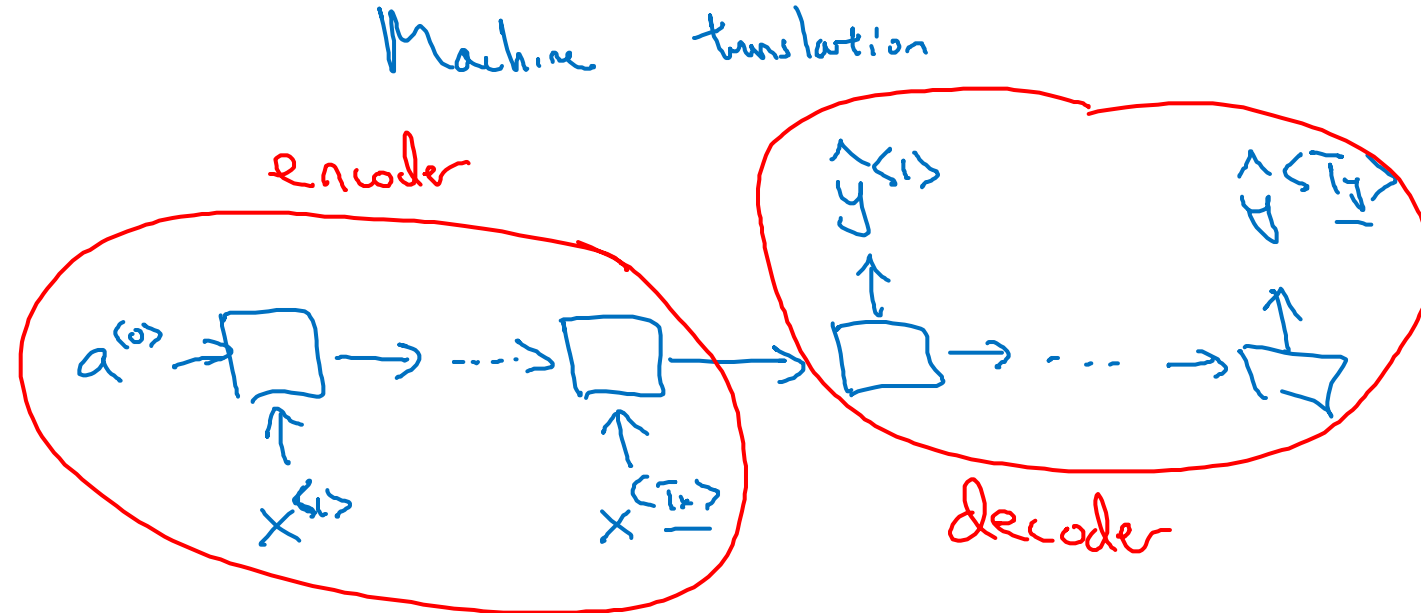
Many-to-many

Summary of RNN types

One to one

One to many

Many to one

Many to many

$T_x = T_y$

Many to many

Andrew Ng
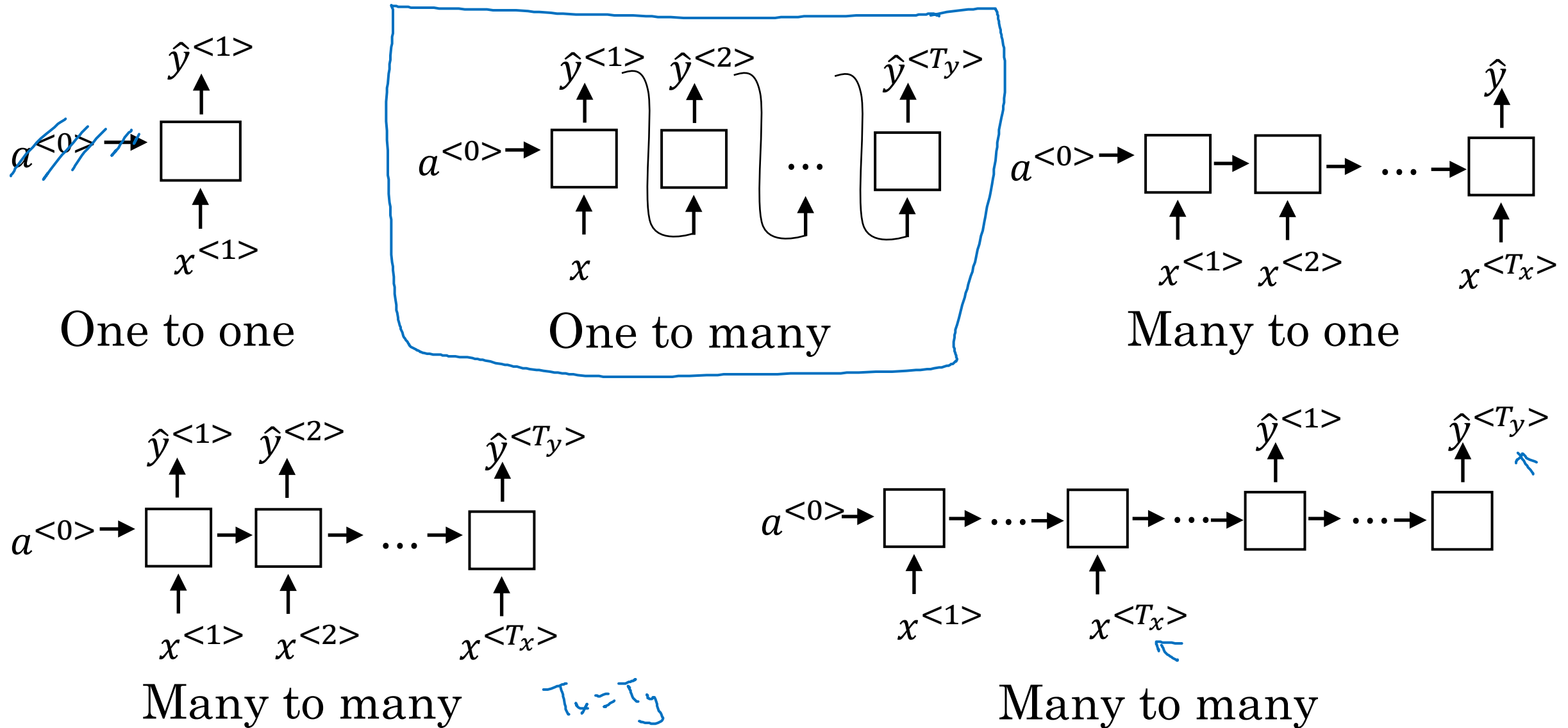
Recurrent Neural Networks

deeplearning.ai

Language model and sequence generation

What is language modelling?

Speech recognition

The apple and <u>pair</u> salad.

→ The apple and <u>pear</u> salad.

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

P(The apple and pear salad) = 5.7×10^{-10}

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

$$P\left(y^{(1)}, y^{(2)}, \ldots, y^{(T_y)}\right)$$

Language modelling with an RNN

Training set: large corpus of english text.

Tokenize

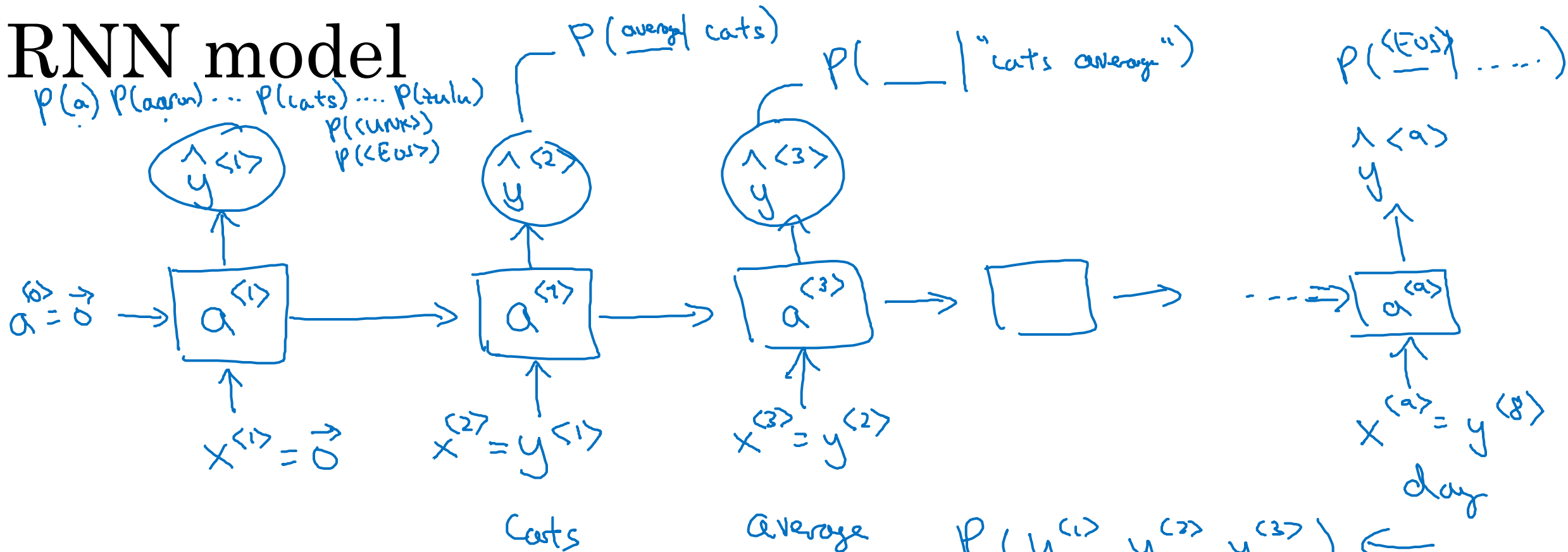Cats average 15 hours of sleep a day. $\swarrow$ <EOS>

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

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

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

<UNK>

10,000

RNN model

$P(a)\ P(aaron)\ \ldots\ P(cats)\ \ldots\ P(zulu)$
$P(<UNK>)$
$P(<EOS>)$

$P(average \mid cats)$

$P(\_\_ \mid \text{"cats average"})$

$P(<EOS> \mid \ldots)$

$\hat{y}^{<1>}$ $\hat{y}^{<2>}$ $\hat{y}^{<3>}$ $\hat{y}^{<a>}$

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

$a^{<1>}$ $a^{<2>}$ $a^{<3>}$ $a^{<a>}$

$x^{<1>} = \vec{0}$ $x^{<2>} = y^{<1>}$ $x^{<3>} = y^{<2>}$ $x^{<a>} = y^{<8>}$

Cats   average   day

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

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

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

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

Andrew Ng

Recurrent Neural Networks

deeplearning.ai

Sampling novel sequences

Sampling a sequence from a trained RNN

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

Training:

$\hat{y}^{<1>}$ $\hat{y}^{<2>}$ $\hat{y}^{<3>}$ $\hat{y}^{<T_y>}$

$a^{<0>} \rightarrow \boxed{a^{<1>}} \rightarrow \boxed{a^{<2>}} \rightarrow \boxed{a^{<3>}} \rightarrow \ldots \rightarrow \boxed{a^{<T_y>}}$

$x^{<1>}$ $y^{<1>}$ $y^{<2>}$ $y^{<T_x-1>}$

Sampling:

The $\hat{y}^{<1>}$ $\hat{y}^{<2>}$ $\hat{y}^{<3>}$ $\hat{y}^{<T_y>}$ $<EOS>$

$a^{<0>} \rightarrow \boxed{a^{<1>}} \rightarrow \boxed{a^{<2>}} \rightarrow \boxed{a^{<3>}} \rightarrow \ldots \rightarrow \boxed{}$ $<UNK>$

$=0$

$x^{<1>} \quad x^{<2>} = \hat{y}^{<1>}$ $y^{<T_x-1>}$

$=0 \quad$ The

$\rightarrow P(a) P(aaron) \ldots P(zulu) P(<UNK>)$ np.random.choice $P(\_ \mid the)$

Andrew Ng
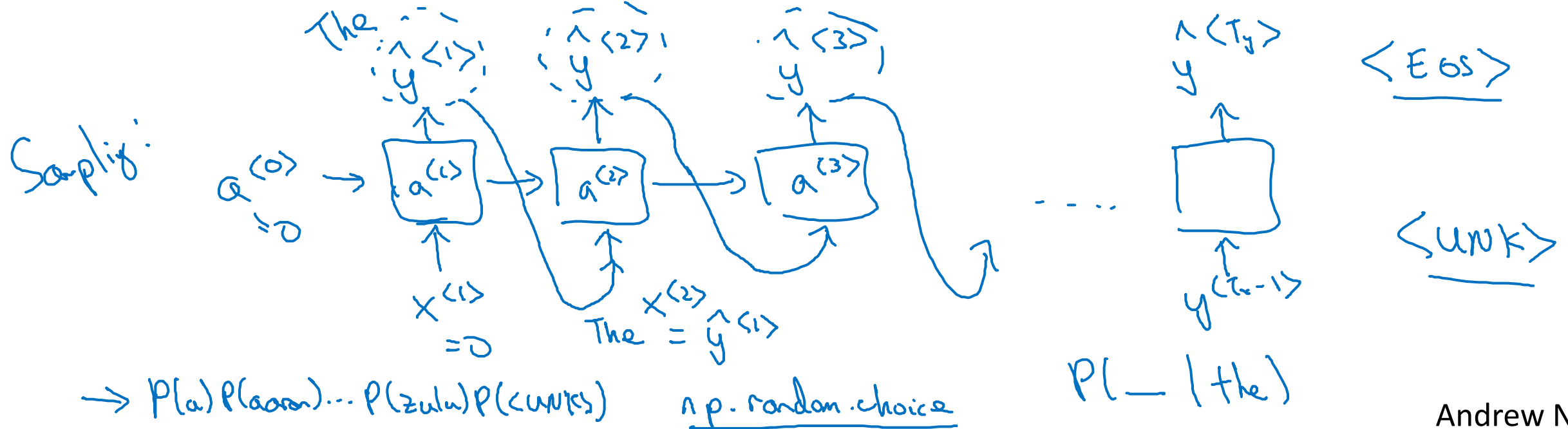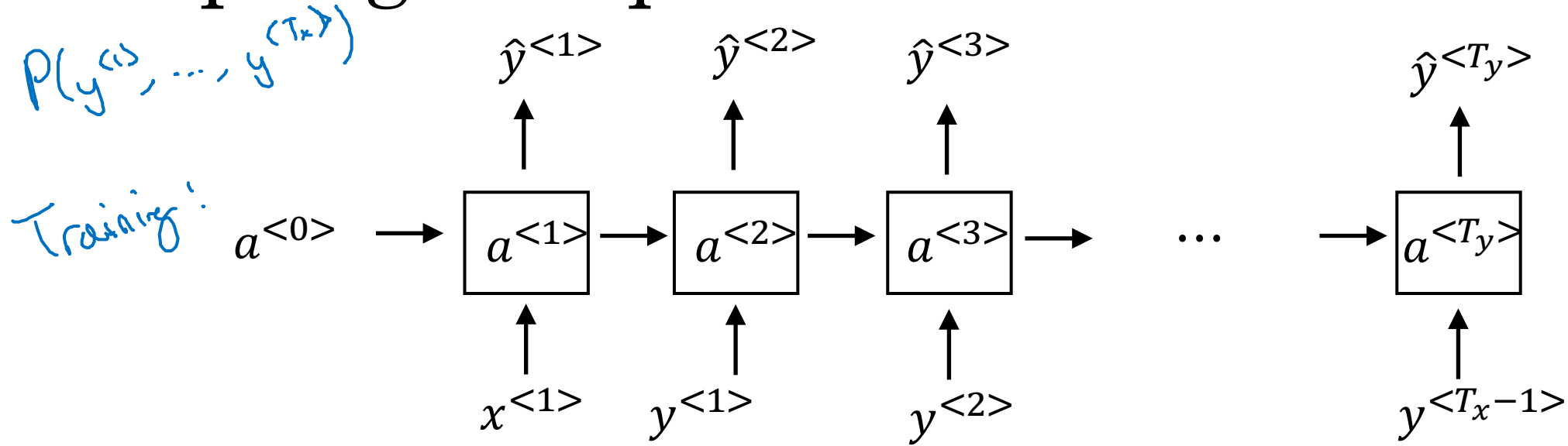
Character-level language model

$\rightarrow$ Vocabulary = [a, aaron, ..., zulu, <UNK>] $\leftarrow$

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

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

Cat average ...
↑ ↑ ↑ ↑

May

Andrew Ng
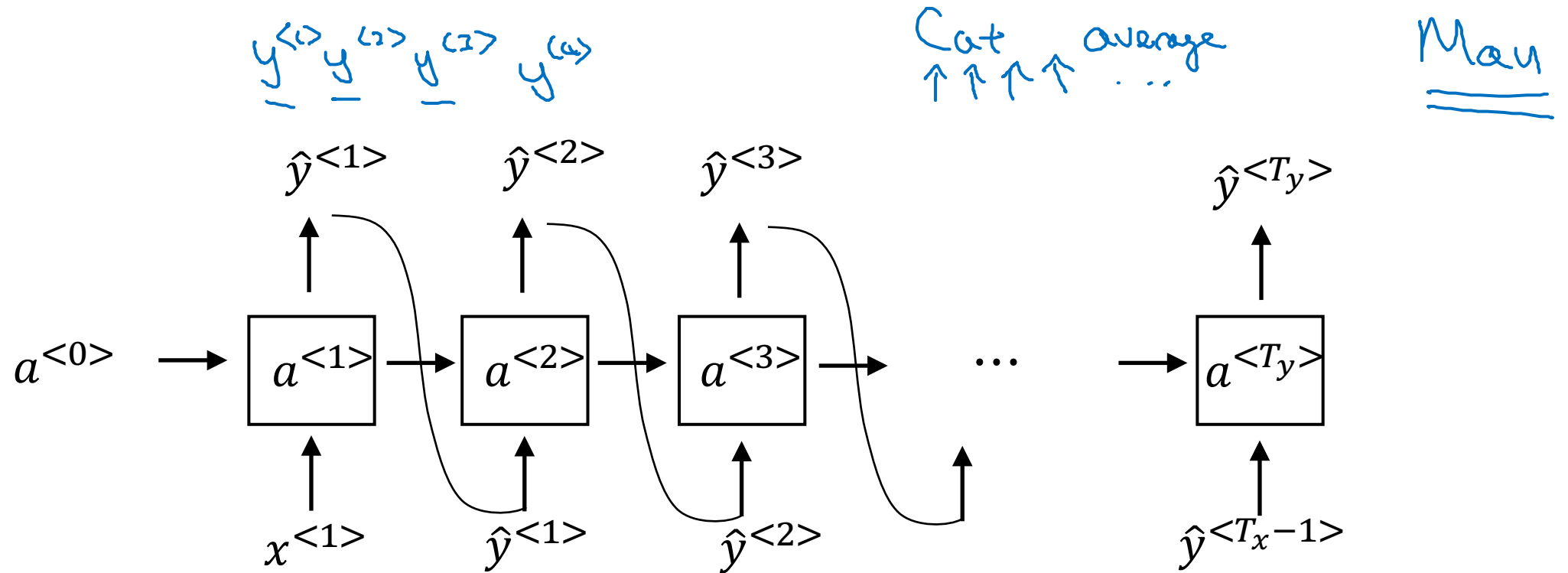
Sequence generation

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.

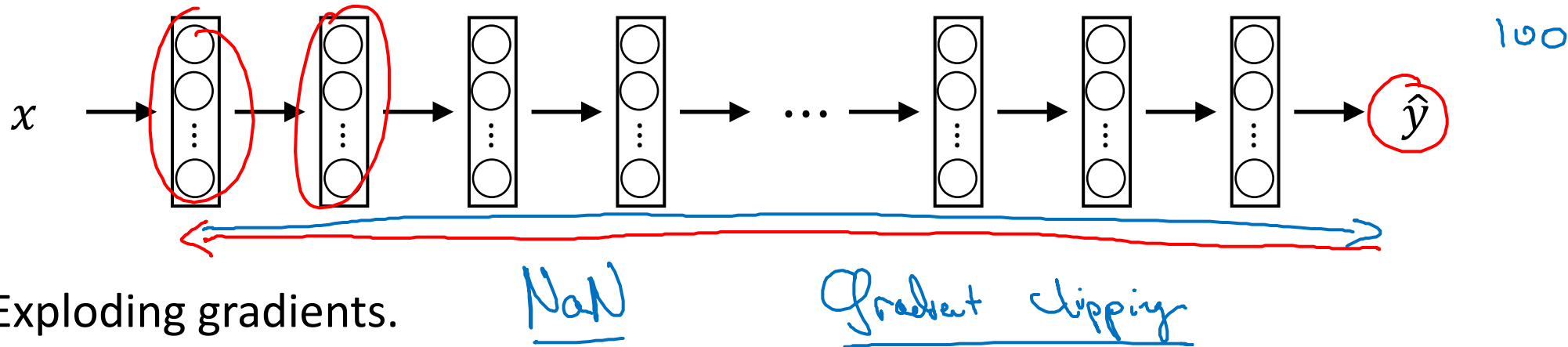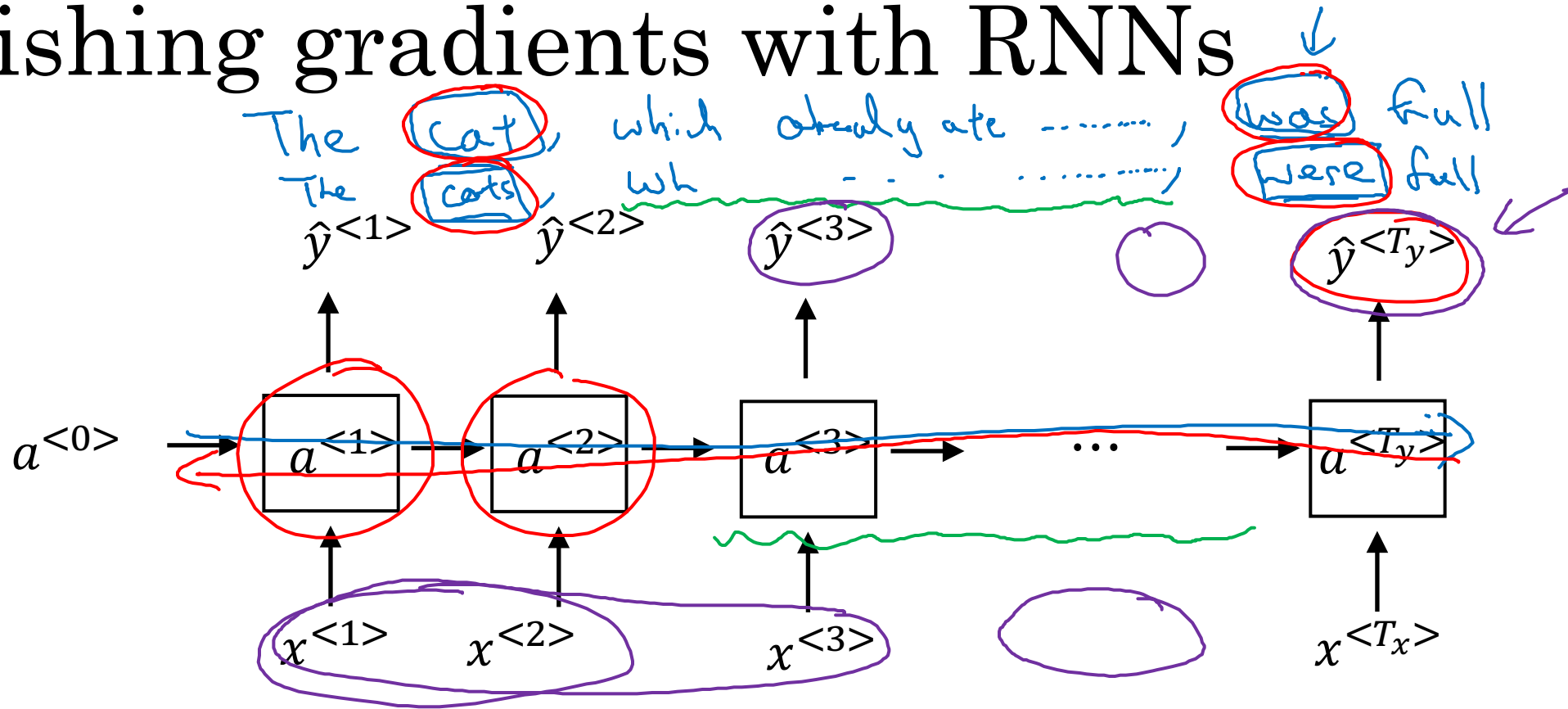For whose are ruse of mine eyes heaves.

Recurrent Neural Networks

Vanishing gradients with RNNs

deeplearning.ai
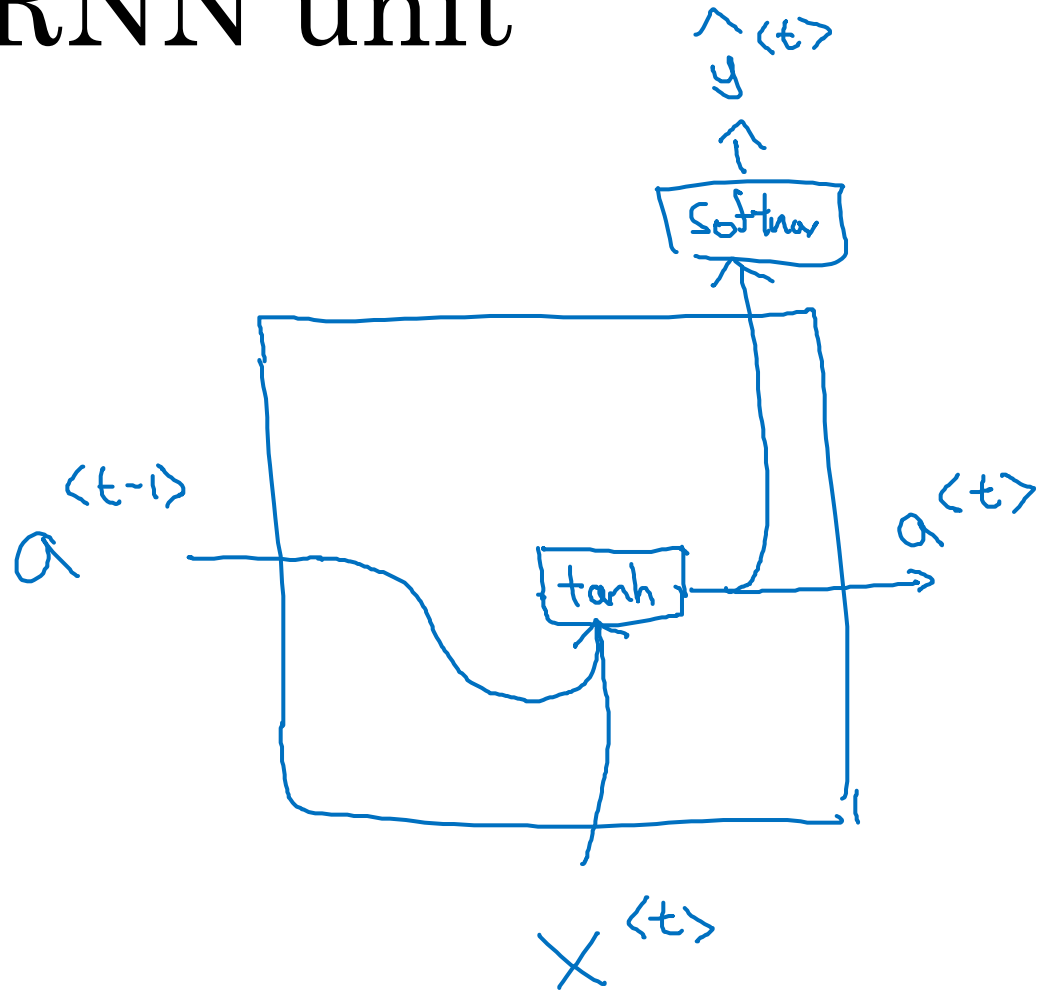
Vanishing gradients with RNNs

The (cat), which already ate ...., (was) full

The (cats), Wh .... (were) full

$\hat{y}^{<1>}$ $\hat{y}^{<2>}$ $\hat{y}^{<3>}$ $\hat{y}^{<T_y>}$

$a^{<0>}$ $a^{<1>}$ $a^{<2>}$ $a^{<3>}$ $\cdots$ $a^{<T_y>}$

$x^{<1>}$ $x^{<2>}$ $x^{<3>}$ $x^{<T_x>}$

x $\hat{y}$

100

Exploding gradients.   NaN   Gradient clipping

deeplearning.ai

Recurrent Neural Networks

Gated Recurrent Unit (GRU)

RNN unit

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

GRU (simplified)

softmax $\rightarrow y^{\langle t \rangle}$

$C^{\langle t-1 \rangle}$
$= a^{\langle t-1 \rangle}$

$\tilde{C}^{\langle t \rangle}$ Γ_u

tanh σ

$x^{\langle t \rangle}$

$C^{\langle t \rangle}$
$= a^{\langle t \rangle}$

$\Gamma_u = 1$ $\Gamma_u = 0$ $\Gamma_u = 0$ $\Gamma_u = 0$ .... $\downarrow = 1$
$c^{\langle t \rangle} = 1$

$\rightarrow$ The cat, which already ate ..., was full.
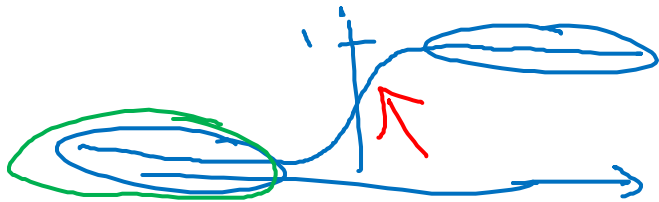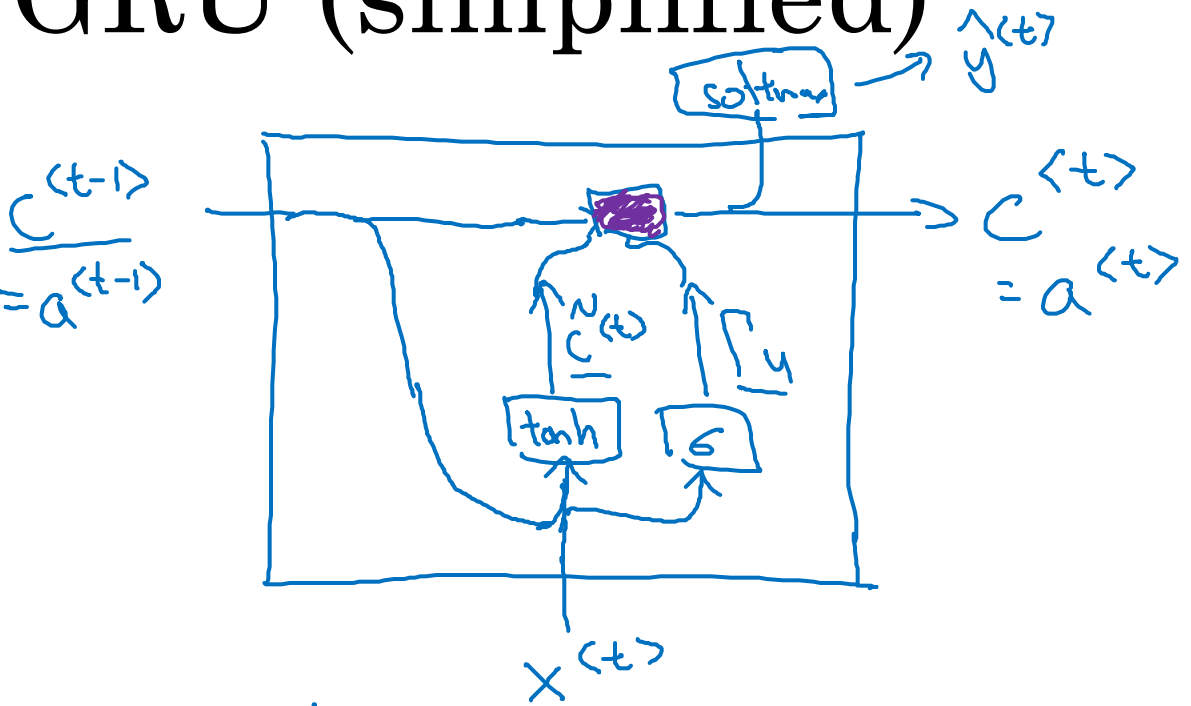
C = memory cell

$C^{\langle t \rangle} = a^{\langle t \rangle}$

$\tilde{C}^{\langle t \rangle} = \tanh\left(W_c\left[c^{\langle t-1 \rangle}, x^{\langle t \rangle}\right] + b_c\right)$

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

"update"

$C^{\langle t \rangle} = \Gamma_u * \tilde{C}^{\langle t \rangle} + \left(1 - \Gamma_u\right) * C^{\langle t-1 \rangle}$

$\Gamma_u = 1$

element-wise $\Gamma_u = 0.00000l$
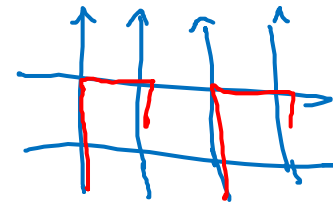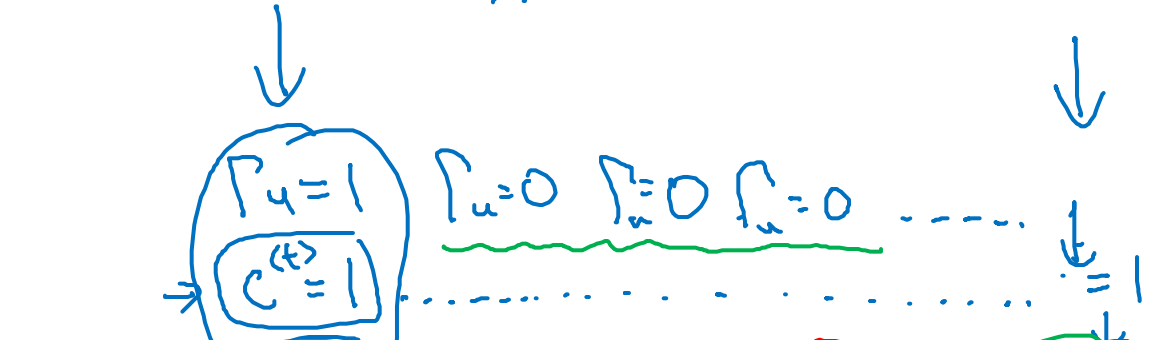
Gate

[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

Full GRU

$\tilde{h}$

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

u

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

r

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

LSTM

h

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

The cat, which ate already, was full.

Recurrent Neural Networks

deeplearning.ai

LSTM (long short term memory) unit

GRU and LSTM

GRU

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

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

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

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

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

LSTM

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

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

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

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

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

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

Γ_f

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

Andrew Ng

LSTM units

GRU

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

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

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

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

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

LSTM

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

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

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

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

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

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

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

LSTM in pictures

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

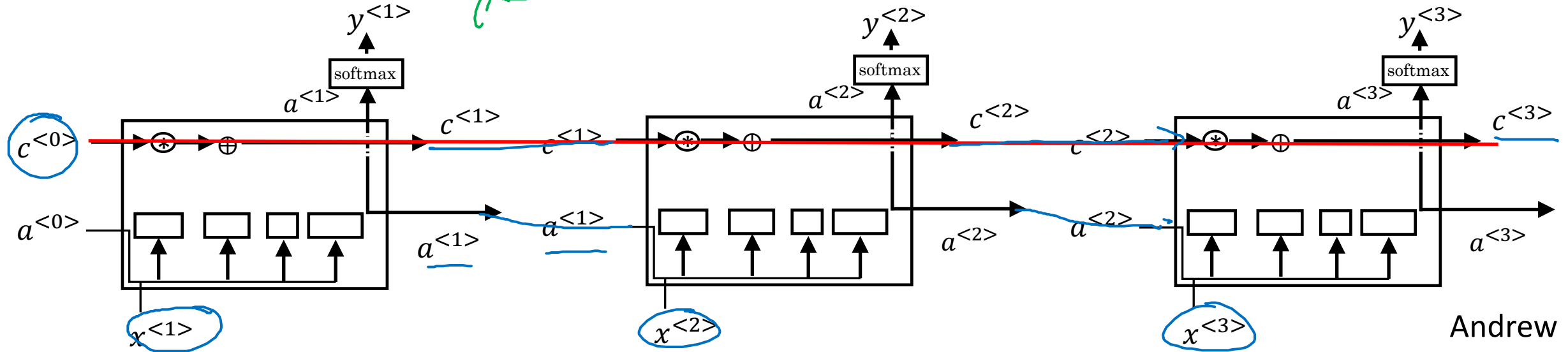$$\Gamma_u = \sigma(W_u[a^{<t-1>}, x^{<t>}] + b_u)$$
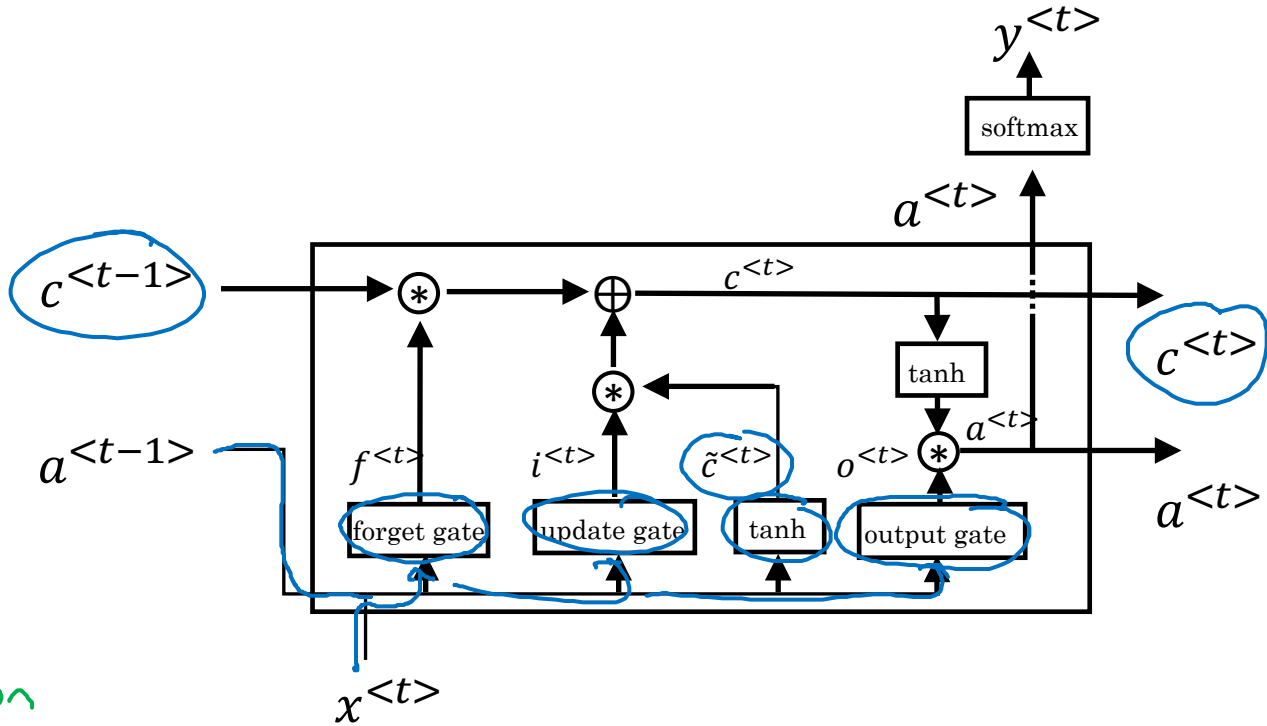
$$\Gamma_f = \sigma(W_f[a^{<t-1>}, x^{<t>}] + b_f)$$

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

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

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

peephole connection
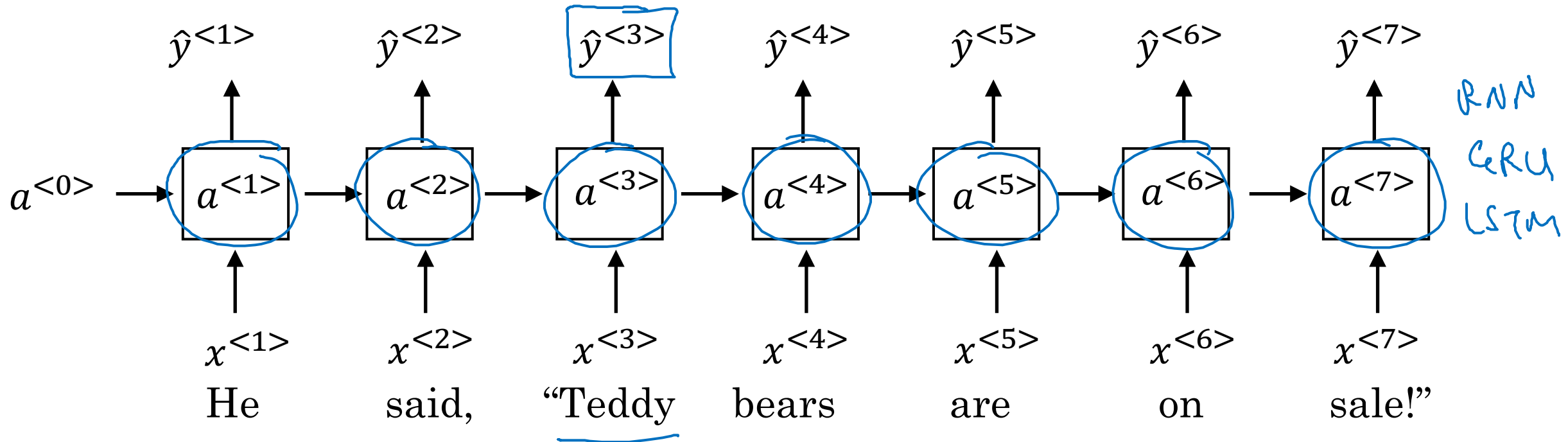
$c^{<t-1>}$

Andrew Ng

Recurrent Neural Networks

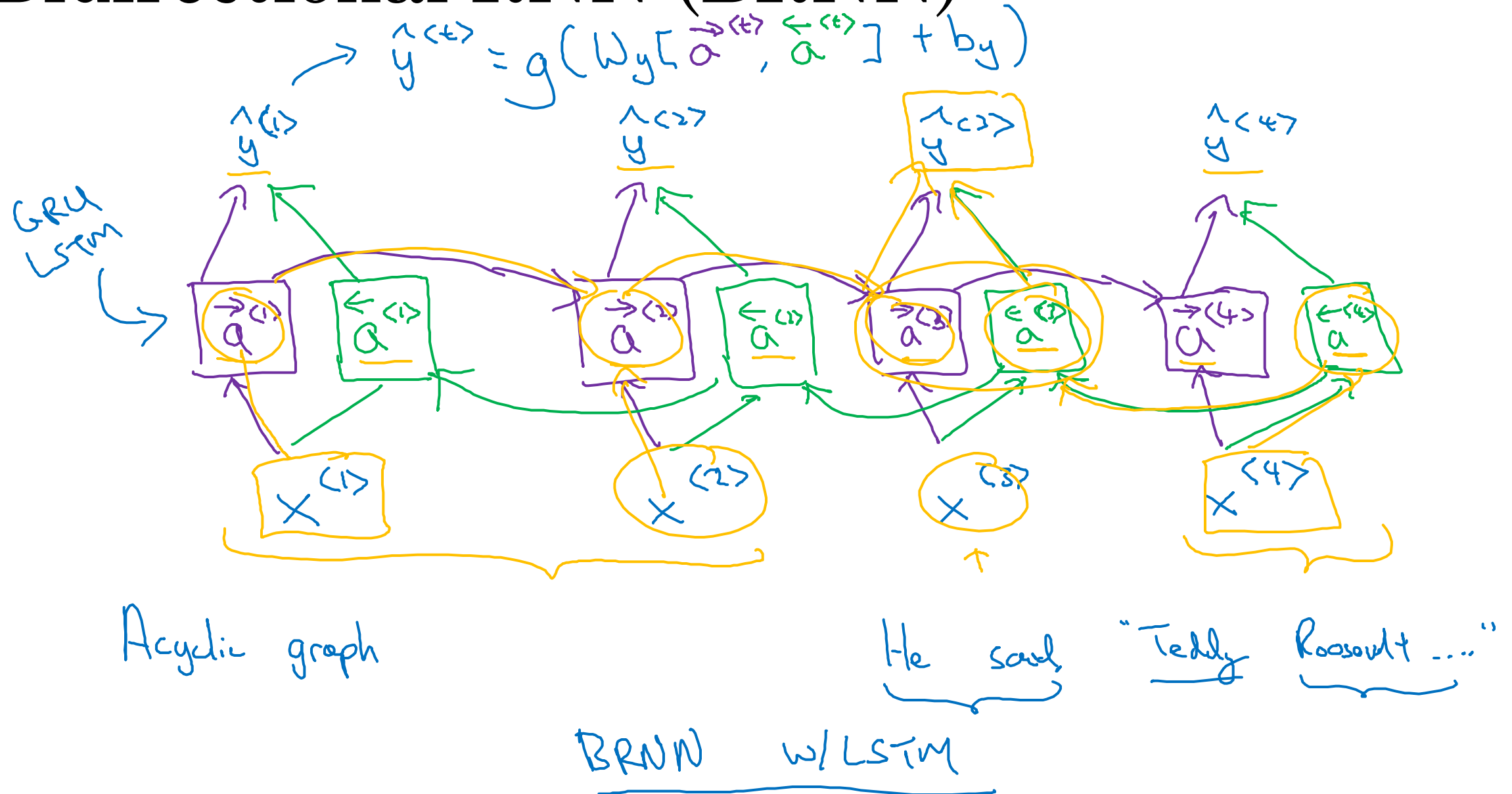## Bidirectional RNN

deeplearning.ai

Getting information from the future

He said, "Teddy bears are on sale!"

He said, "Teddy Roosevelt was a great President!"
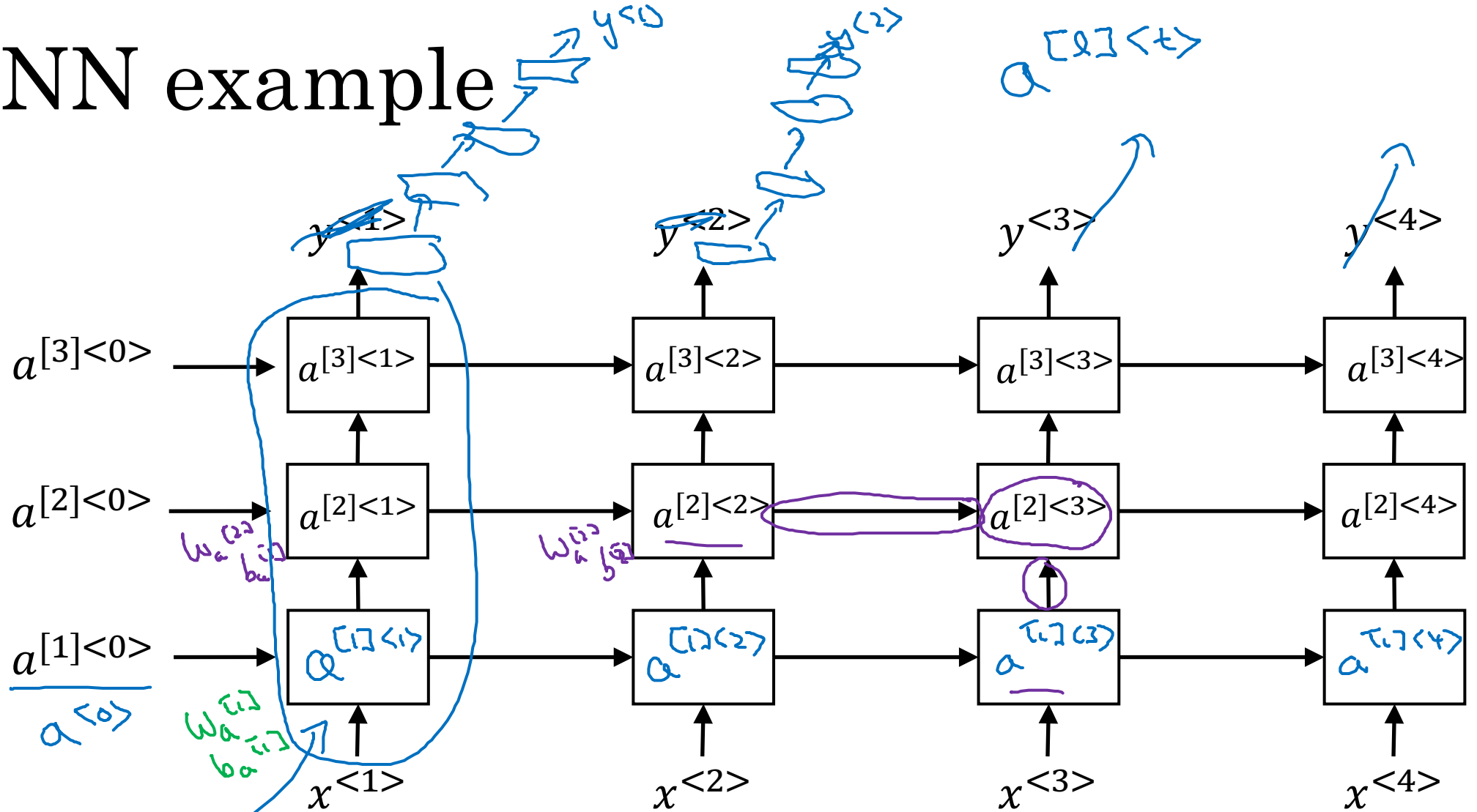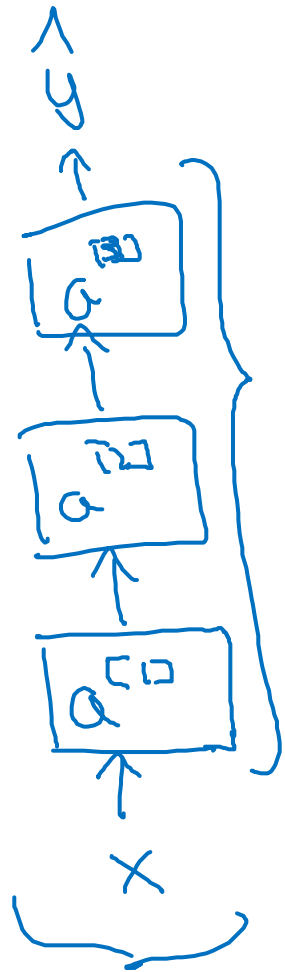
Bidirectional RNN (BRNN)

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

GRU
LSTM

Acyclic graph

He said, "Teddy Roosevelt ...."

BRNN w/ LSTM

Recurrent Neural Networks

Deep RNNs

deeplearning.ai

Deep RNN example

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

RNN
GRU
LSTM

BRNN

$a^{[3]<0>} \rightarrow a^{[3]<1>} \rightarrow a^{[3]<2>} \rightarrow a^{[3]<3>} \rightarrow a^{[3]<4>}$

$a^{[2]<0>} \rightarrow a^{[2]<1>} \rightarrow a^{[2]<2>} \rightarrow a^{[2]<3>} \rightarrow a^{[2]<4>}$

$a^{[1]<0>} \rightarrow a^{[1]<1>} \rightarrow a^{[1]<2>} \rightarrow a^{[1]<3>} \rightarrow a^{[1]<4>}$

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

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

$W_a^{[2]}, b_a^{[2]}$

$W_a^{[1]}, b_a^{[1]}$

$a^{[3]<t>}$

$a^{<0>}$

Andrew Ng