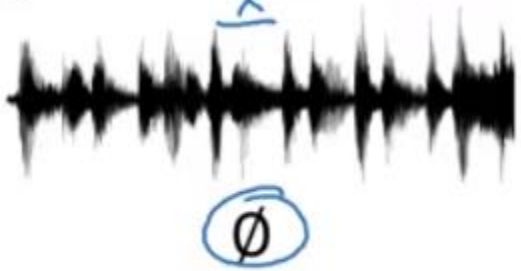
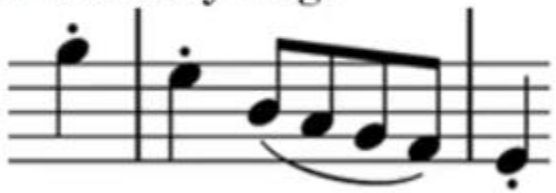
deeplearning.ai

# Recurrent Neural Networks

## Why sequence models?

# Examples of sequence data

| | | |
|---|---|---|
| Speech recognition |  | "The quick brown fox jumped over the lazy dog." |
| Music generation | |  |
| 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

deeplearning.ai

# Recurrent Neural Networks

## 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

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

$T_y = 9$

$X^{(i)<t>}$

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

$T_x^{(i)} = 9$

$T_y^{(i)}$

15

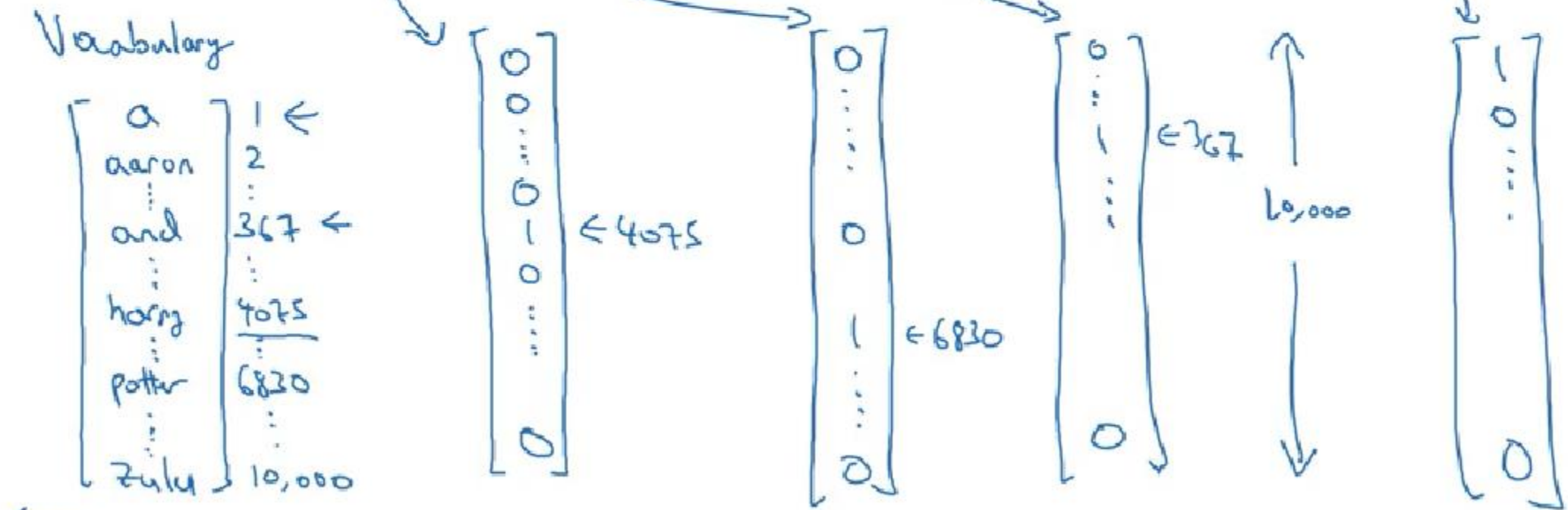
# Representing words

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

$x \rightarrow y$

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

$x^{<1>}$ $x^{<2>}$ $x^{<3>}$ ... $x^{<7>}$ $x^{<9>}$




Andrew Ng

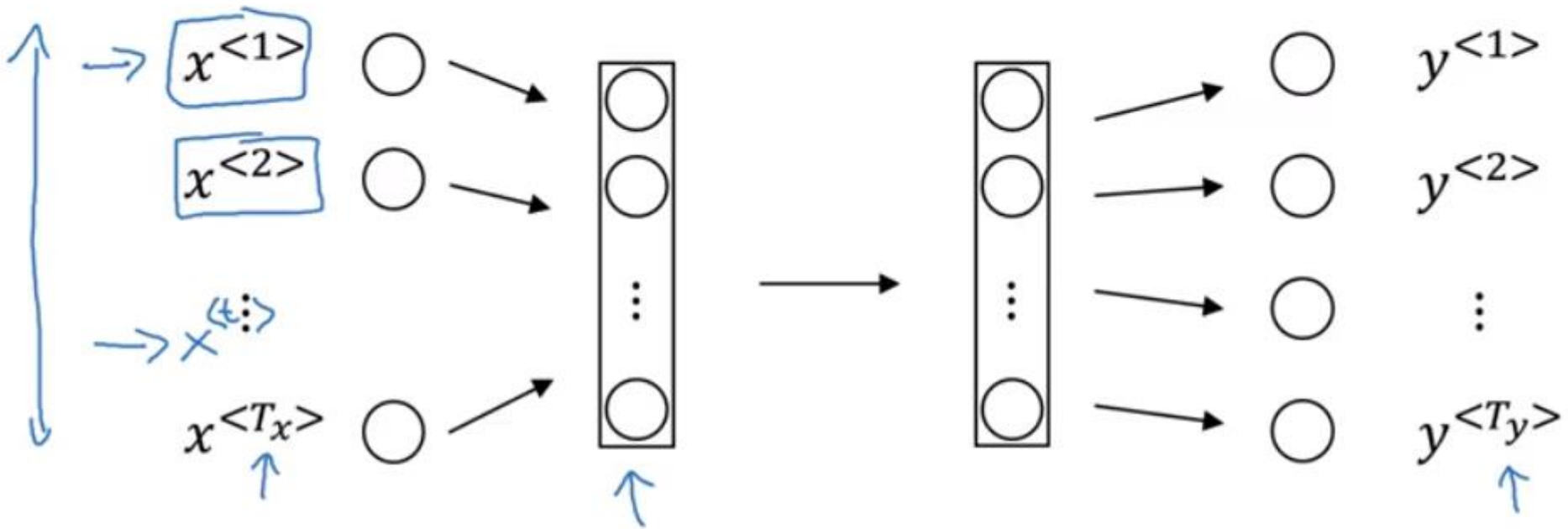
deeplearning.ai

# Recurrent Neural Networks

---

# Recurrent Neural Network Model

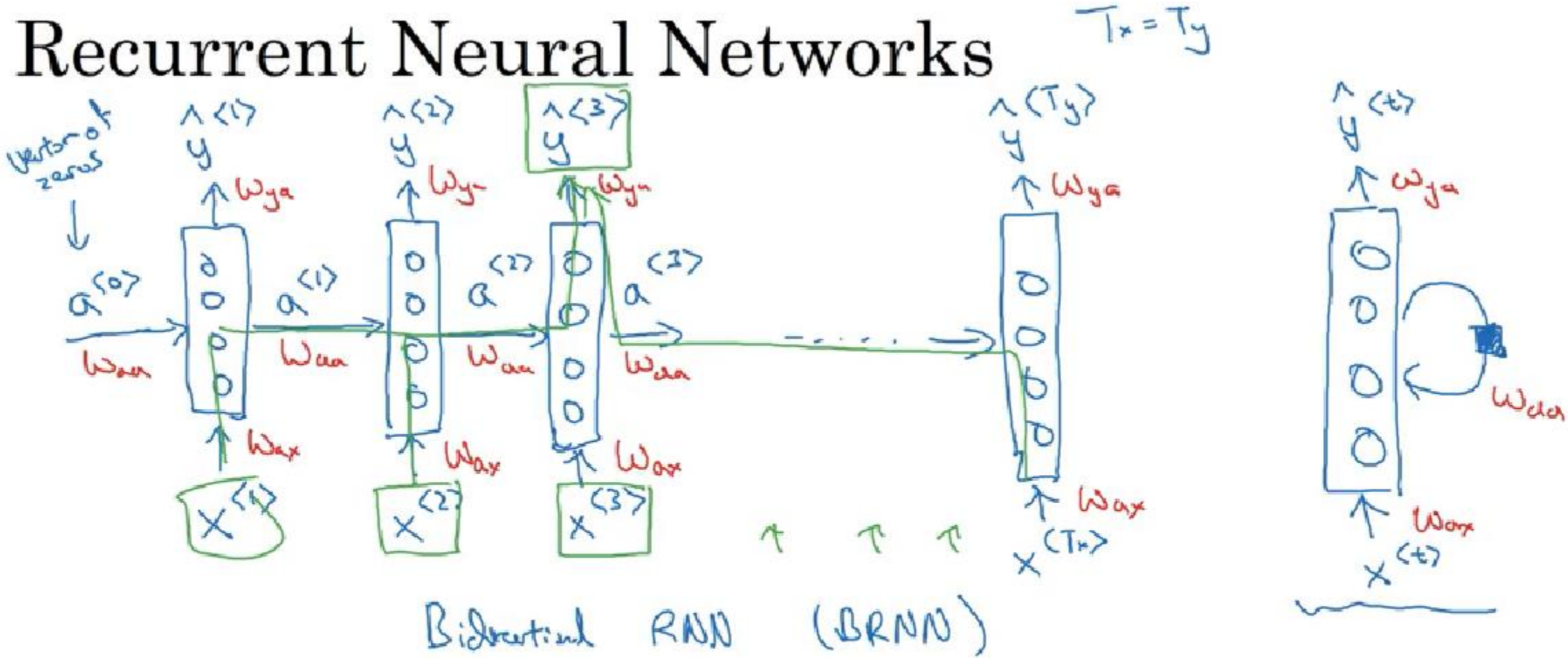
# Why not a standard network?




Problems:

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

# Recurrent Neural Networks



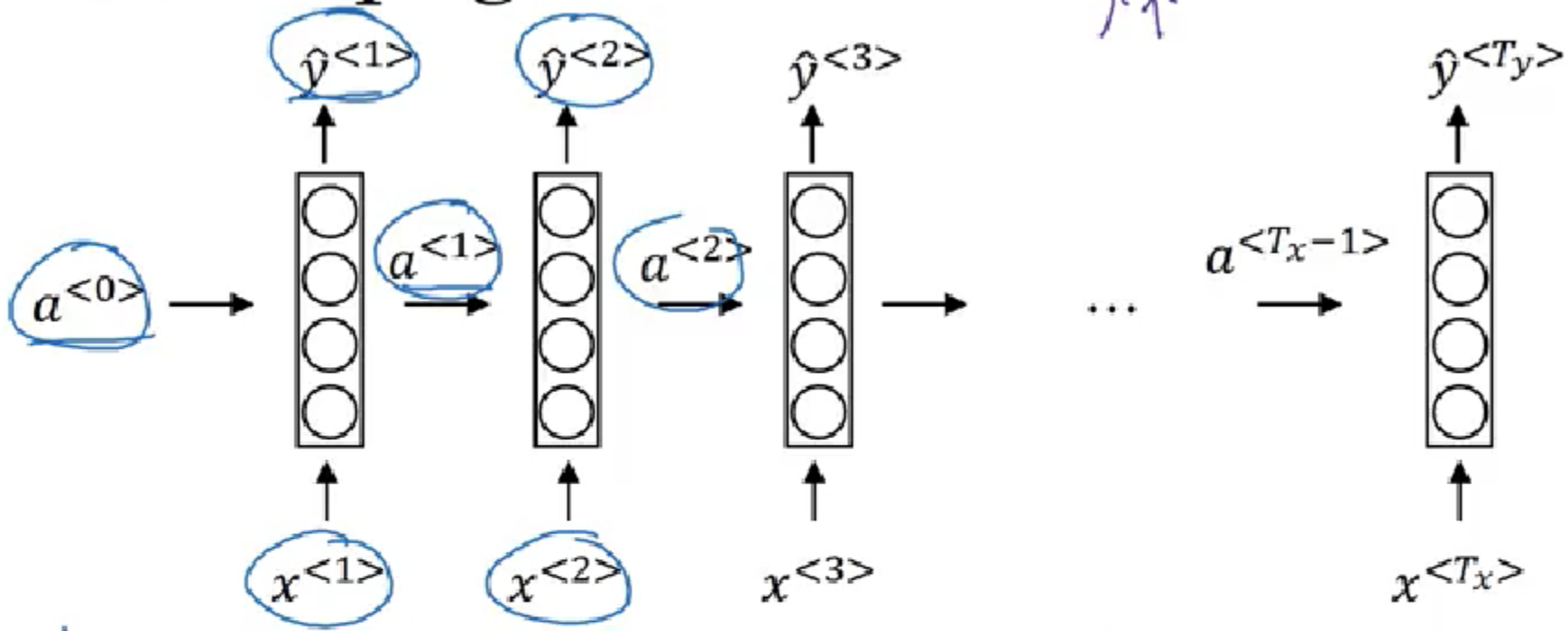

He said, "Teddy Roosevelt was a great President."

He said, "Teddy bears are on sale!"

Andrew Ng

# Forward Propagation

$a \leftarrow W_{ax} x^{\langle 1 \rangle}$




$a^{\langle 0 \rangle} = \vec{0}$.

$$a^{\langle 1 \rangle} = g_1(W_{aa} a^{\langle 0 \rangle} + W_{ax} x^{\langle 1 \rangle} + b_a) \leftarrow \tanh / \text{ReLU}$$

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

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

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

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)

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

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

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

$$\begin{bmatrix} 100 \uparrow \\ \downarrow \end{bmatrix} \begin{bmatrix} W_{aa} & W_{ax} \\ \leftarrow 100 \rightarrow & \leftarrow 10000 \rightarrow \end{bmatrix} = W_a \quad (100, 10100)$$

$$[a^{<t-1>}, x^{<t>}] = \begin{bmatrix} a^{<t-1>} \\ x^{<t>} \end{bmatrix} \begin{matrix} \updownarrow 100 \\ \updownarrow 10000 \end{matrix} \quad \updownarrow 10100$$

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

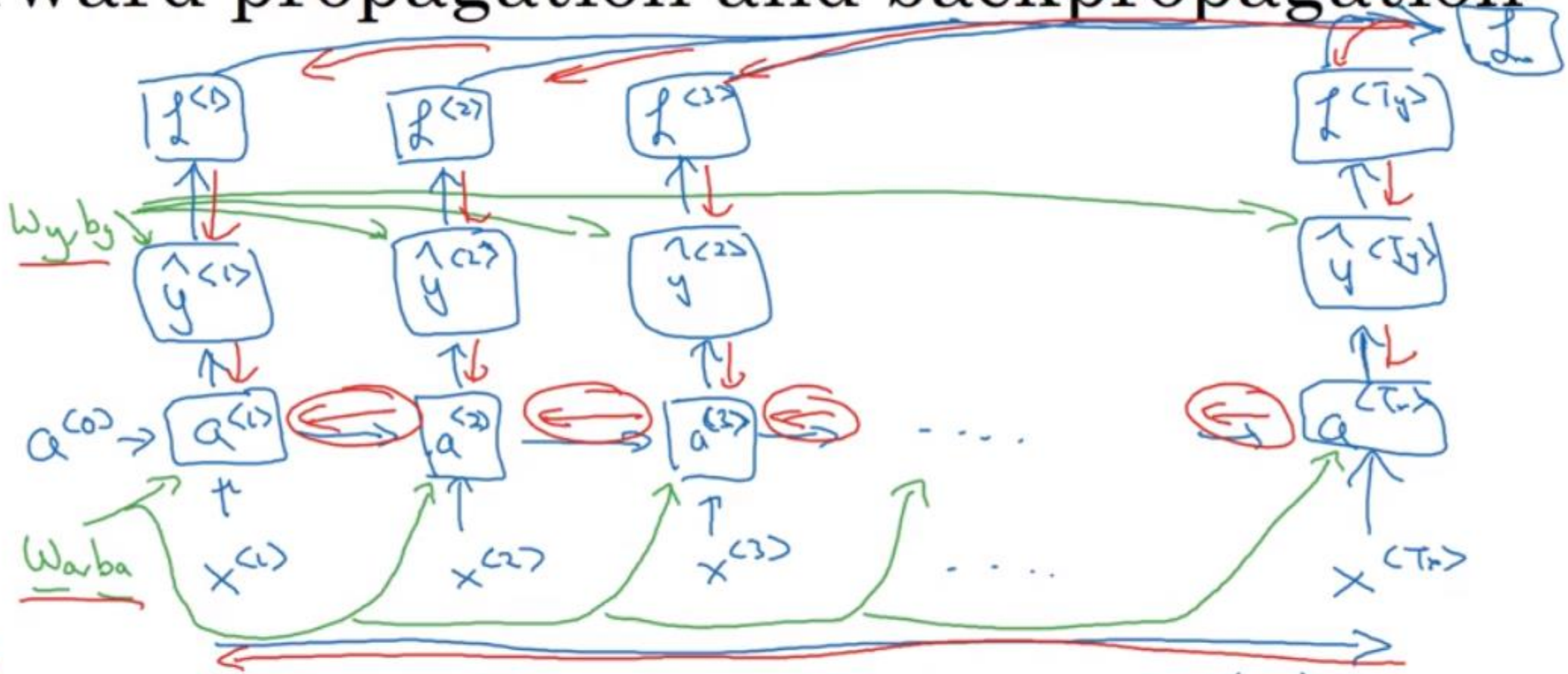
Andrew Ng

deeplearning.ai

# Recurrent Neural Networks

## Backpropagation through time

# Forward propagation and backpropagation




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

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

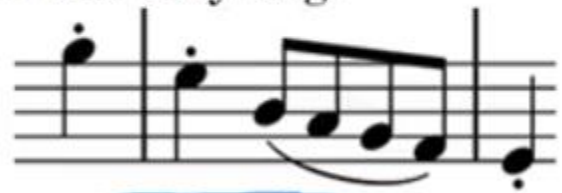
Backpropagation through time

deeplearning.ai

# Recurrent Neural Networks

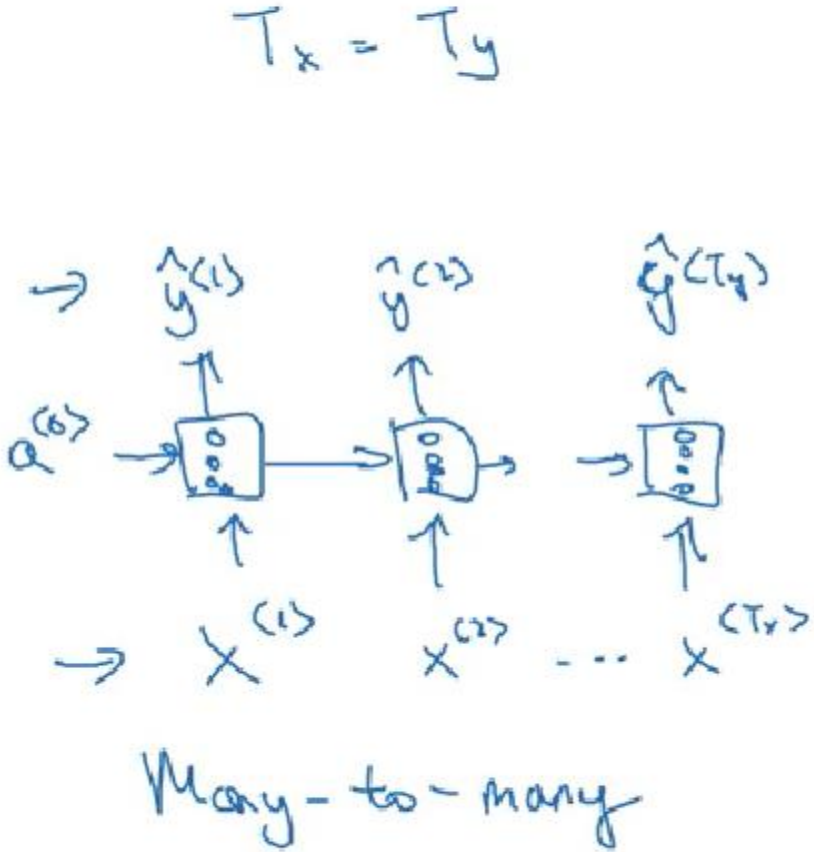
## Different types of RNNs

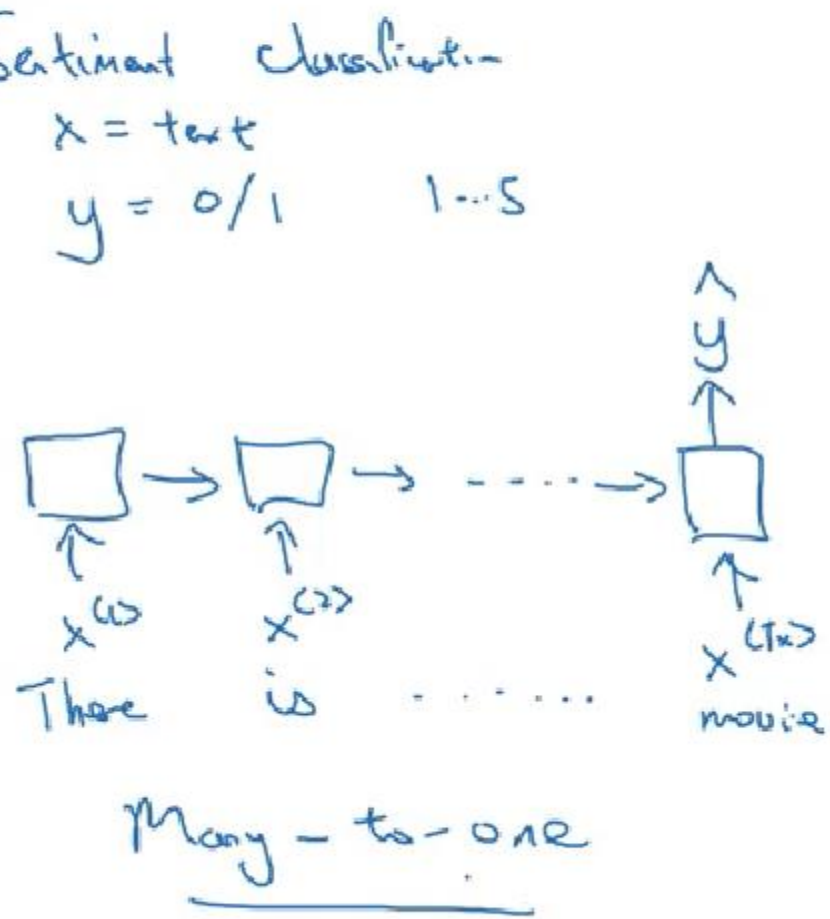
# Examples of sequence data

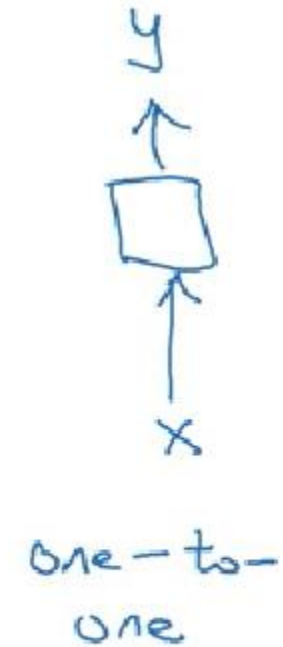
| | | |
|---|---|---|
| Speech recognition |  | "The quick brown fox jumped over the lazy dog." |
| Music generation | ∅ |  |
| Sentiment classification | "There is nothing to like in this movie." | ★☆☆☆☆ |
| DNA sequence analysis | AGCCCCTGTGAGGAACTAG | AGCCCCTGTGAGGAACTAG |
| Machine translation | 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

# Examples of RNN architectures

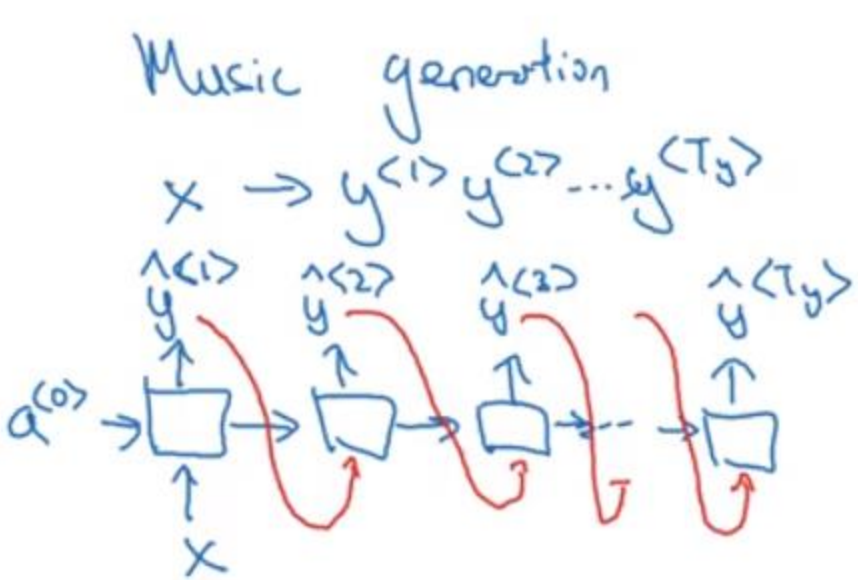


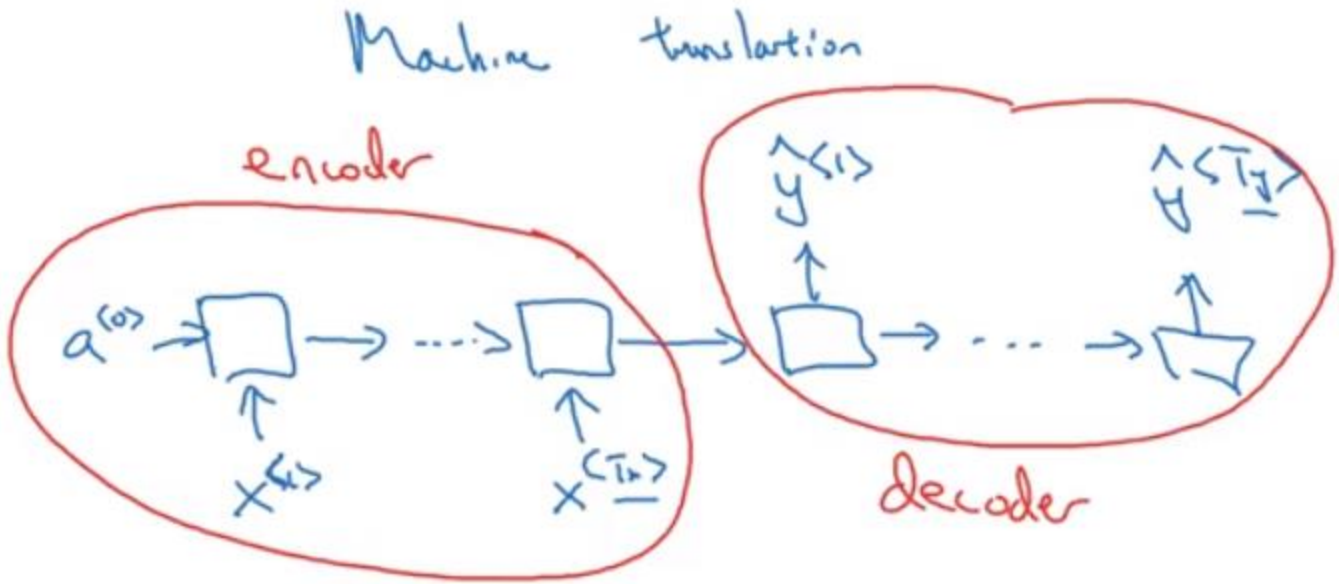

Andrew Ng

# Examples of RNN architectures



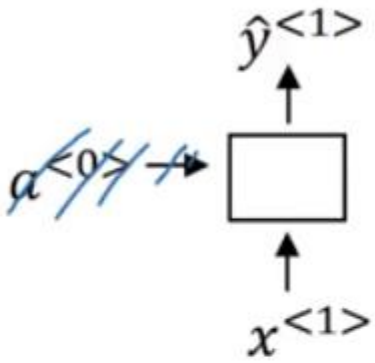

One-to-many

$x = \phi$

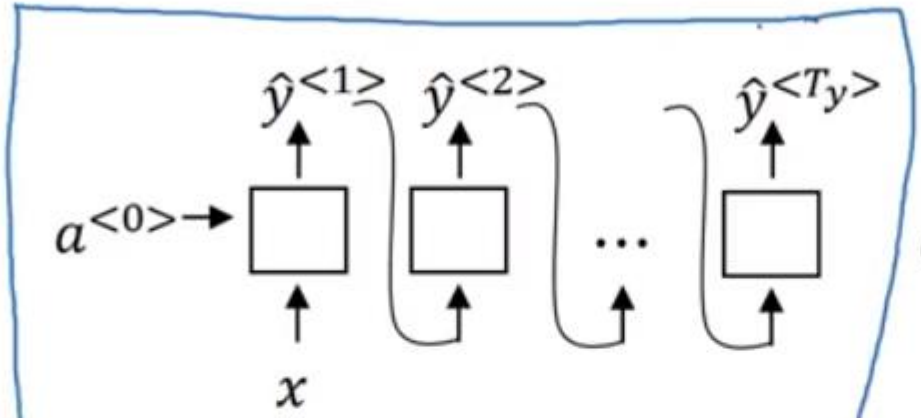



Many-to-many

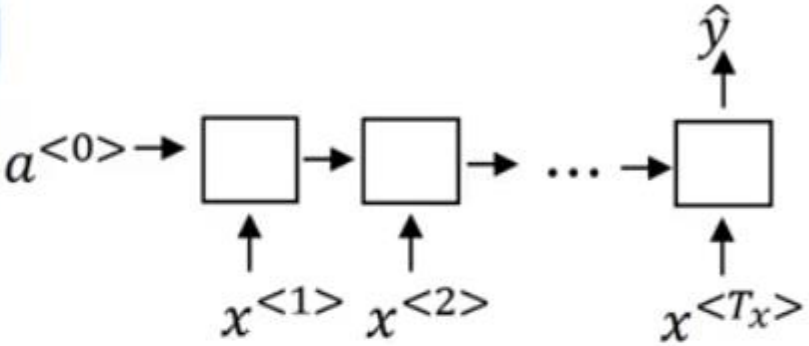
# Summary of RNN types



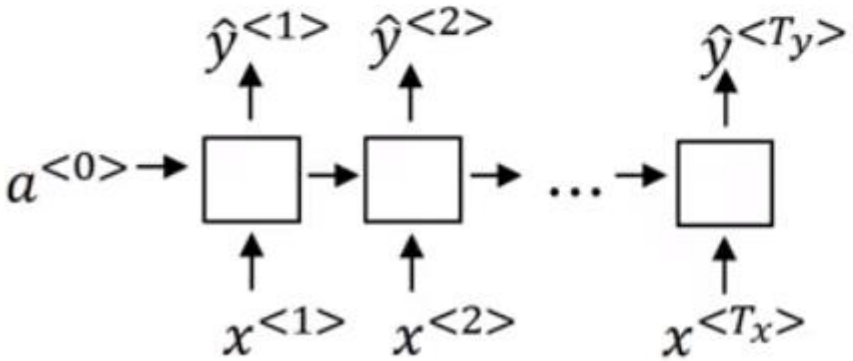

One to one




One to many

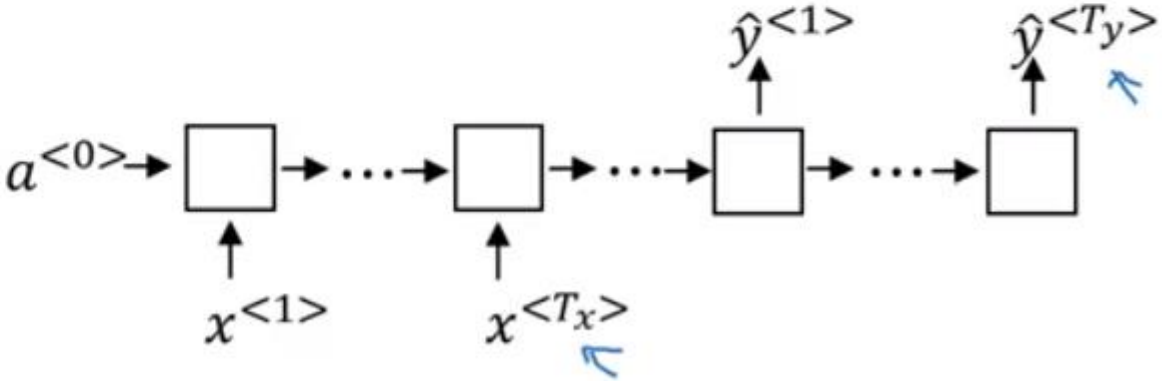



Many to one




Many to many

$T_x = T_y$




Many to many

deeplearning.ai

# Recurrent Neural Networks

---

# Language model and sequence generation

# What is language modelling?

Speech recognition

The apple and pair salad.

The apple and pear salad.

$P(\text{The apple and pair salad}) =$

$P(\text{The apple and pear salad}) =$

# Language modelling with an RNN

Training set: large corpus of english text.

Cats average 15 hours of sleep a day.

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

# RNN model

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

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

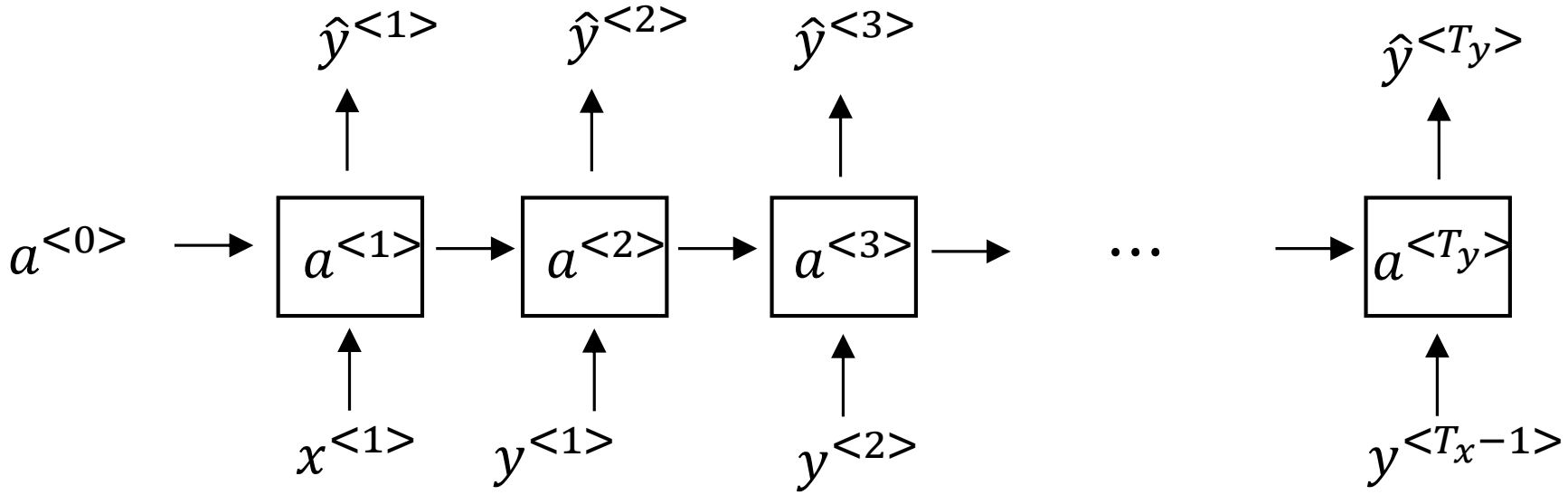
$$\mathcal{L} = \sum_t \mathcal{L}^{<t>}(\hat{y}^{<t>}, y^{<t>})$$

deeplearning.ai

# Recurrent Neural Networks

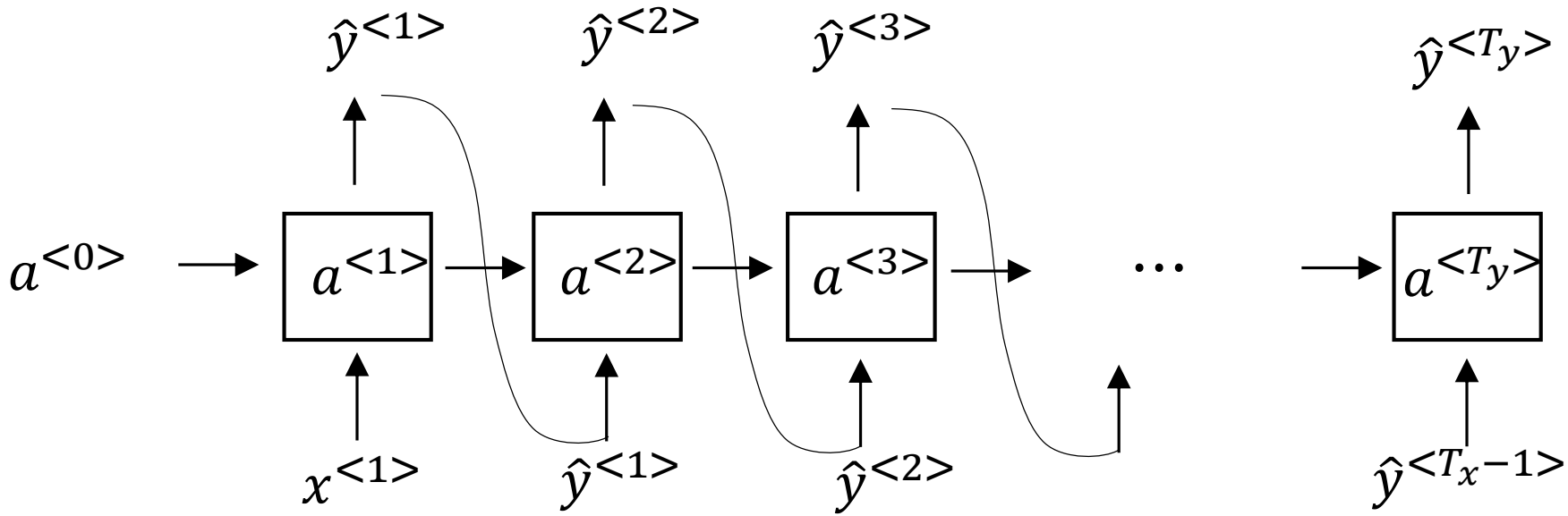
## Sampling novel sequences

# Sampling a sequence from a trained RNN



# Character-level language model

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



# 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.

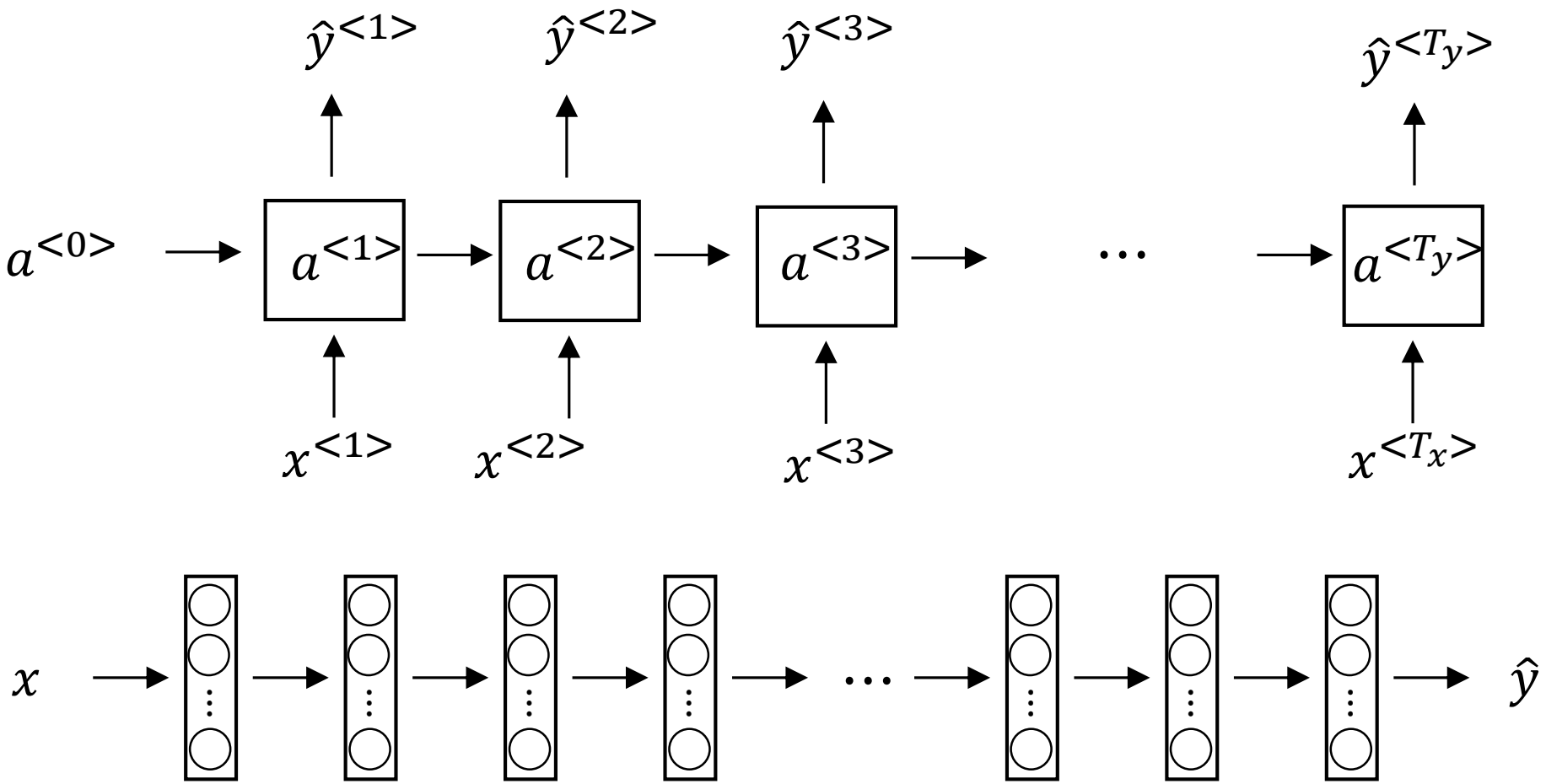
deeplearning.ai

# Recurrent Neural Networks

---

# Vanishing gradients with RNNs

# Vanishing gradients with RNNs




Exploding gradients.

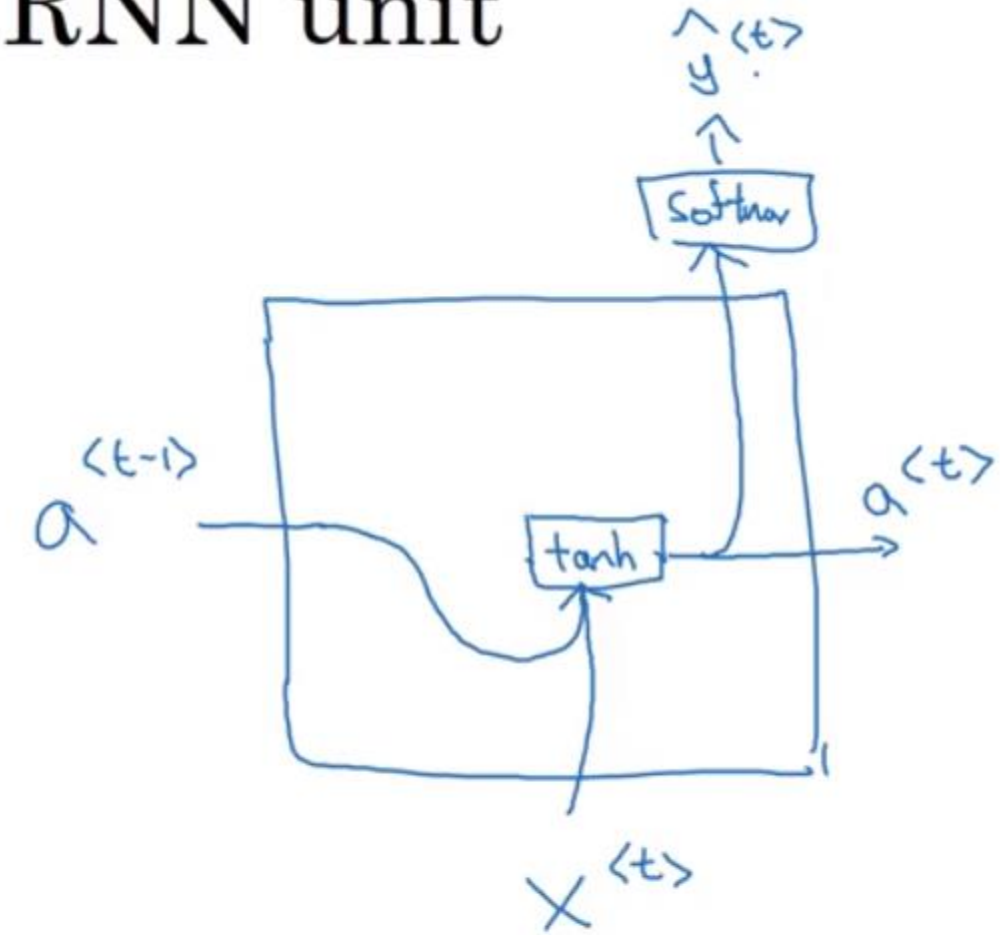
deeplearning.ai

# Recurrent Neural Networks

---

# Gated Recurrent Unit (GRU)

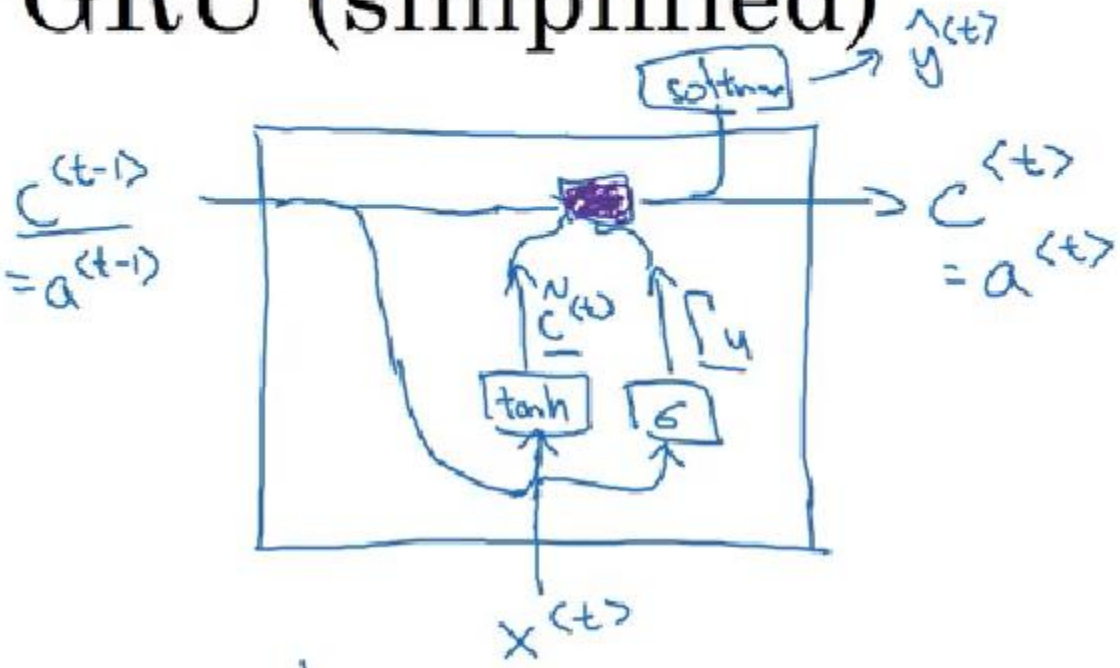
# RNN unit




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

tanh

# GRU (simplified)




$c$ = memory cell

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

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

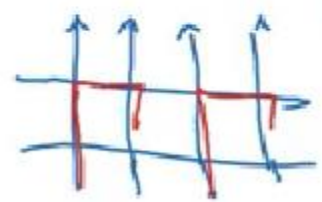
$\Gamma_u = \sigma(W_u [c^{\langle t-1 \rangle}, x^{\langle t \rangle}] + b_u)$

"update"

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

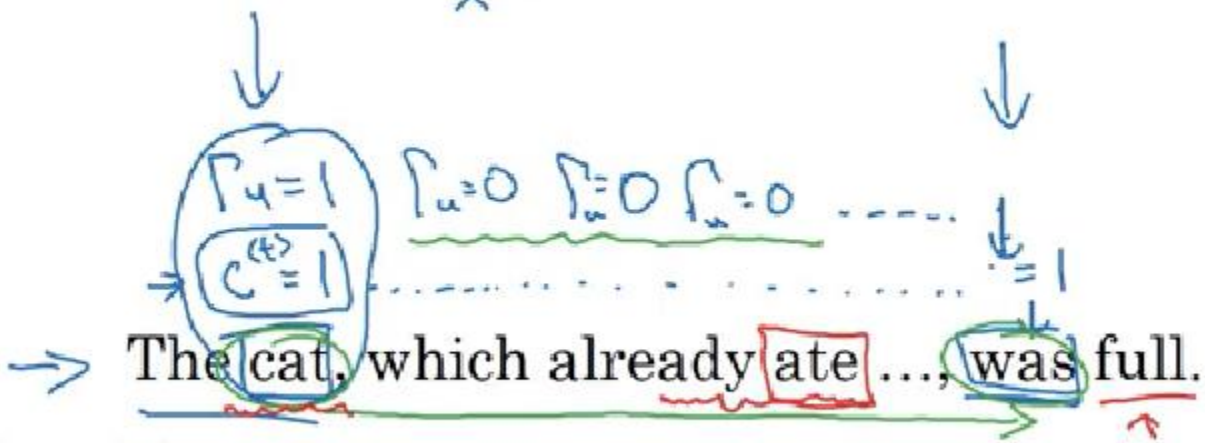
$\Gamma_u = 1$

element-wise



Gate

$\Gamma_u = 0.000001$




[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

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

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

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

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

LSTM

The cat, which ate already, was full.

deeplearning.ai

# Recurrent Neural Networks

---

# 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

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

# LSTM in pictures

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

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

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

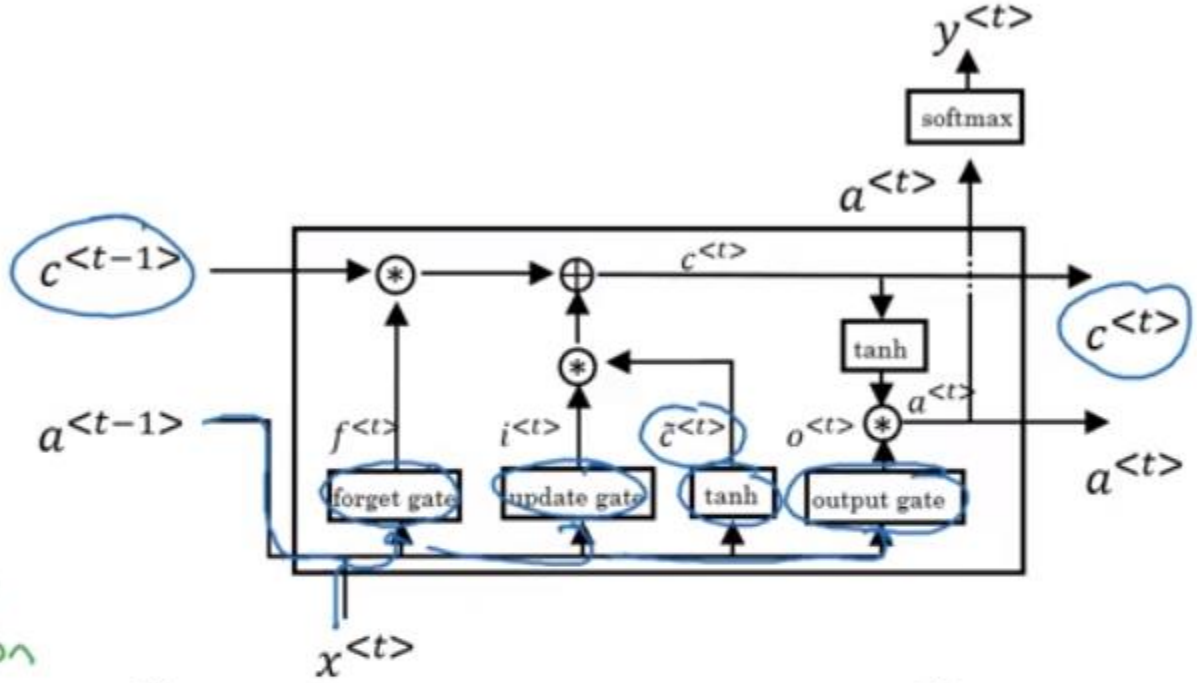
$$\Gamma_o = \sigma(W_o[a^{<t-1>}, x^{<t>}] + b_o)$$

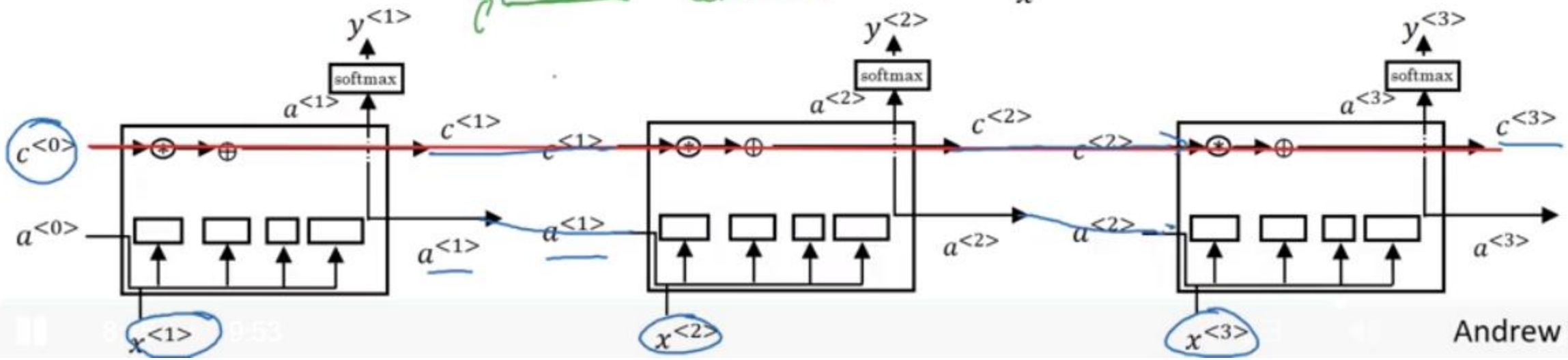
$$c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + \Gamma_f * c^{<t-1>}$$

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

$c^{<t-1>}$

peephole connection




Andrew Ng

deeplearning.ai

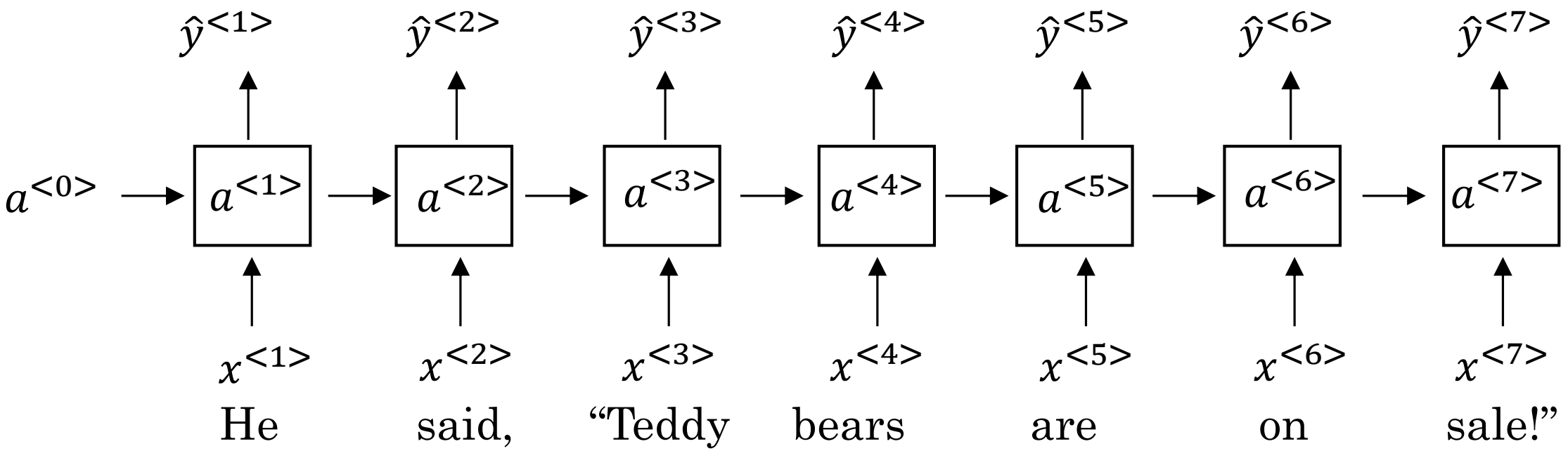
# Recurrent Neural Networks

## Bidirectional RNN

# Getting information from the future

He said, “Teddy bears are on sale!”

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



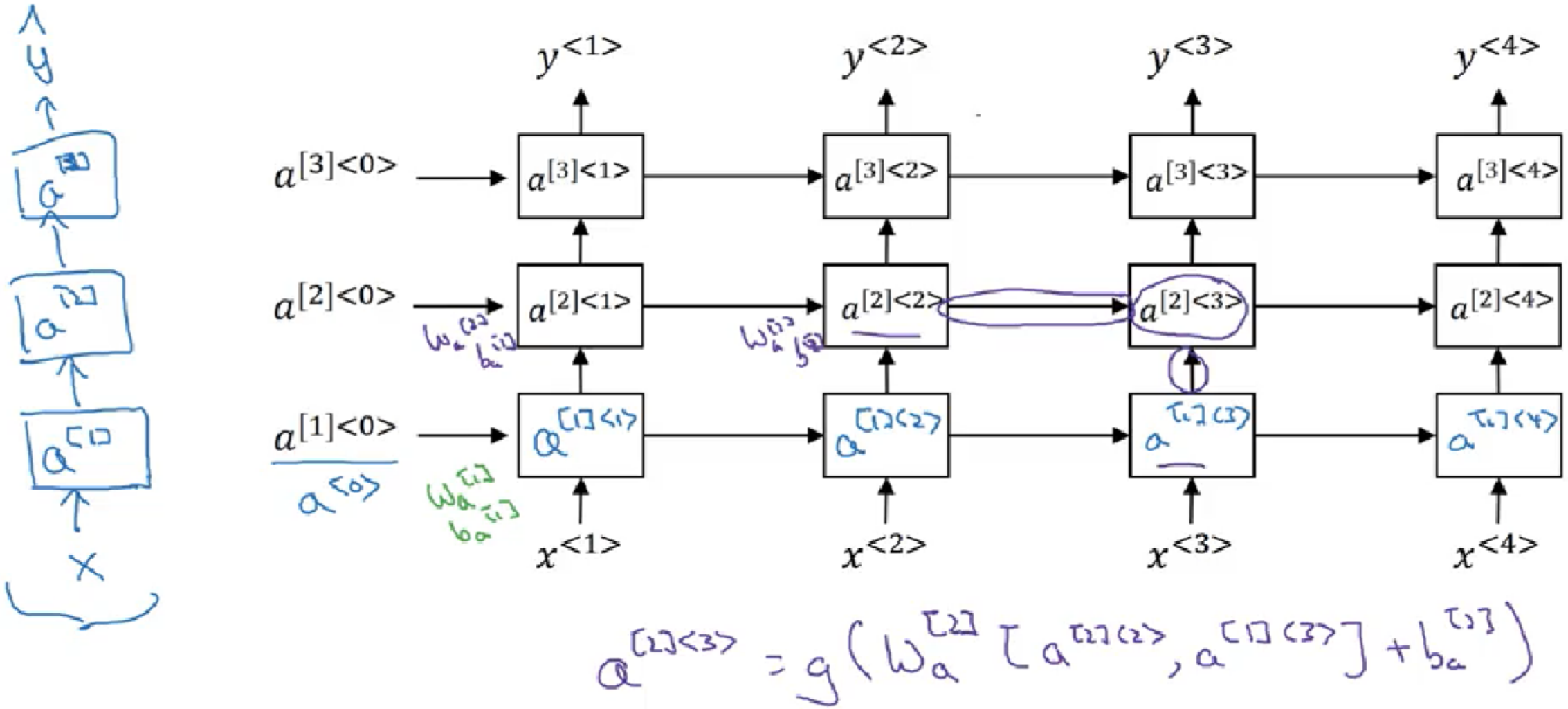
# Bidirectional RNN (BRNN)

deeplearning.ai

# Recurrent Neural Networks

## Deep RNNs

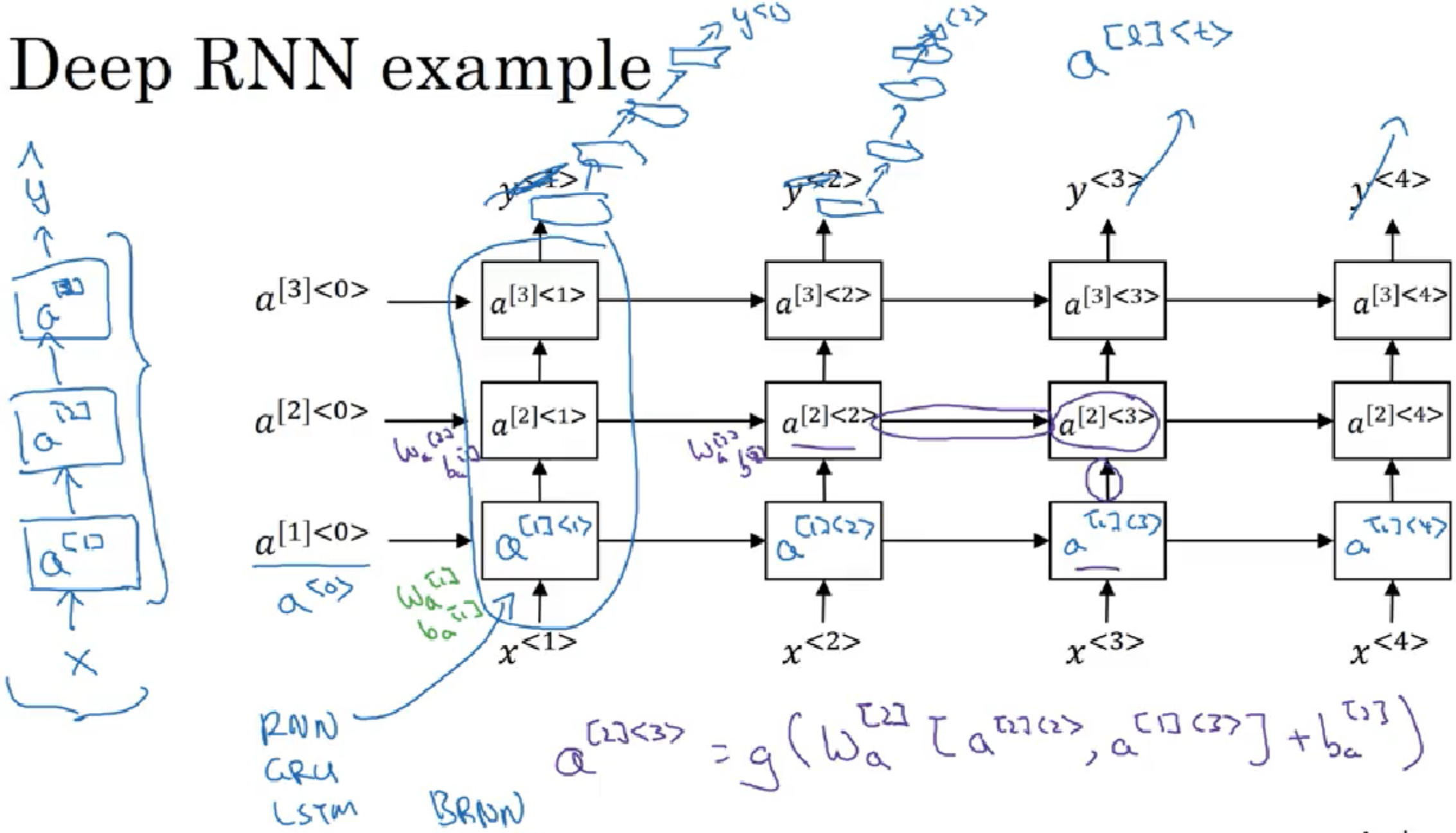
# Deep RNN example




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

Andrew Ng

# Deep RNN example




Andrew Ng