Recurrent Neural Networks

Why sequence models?

Examples of sequence data
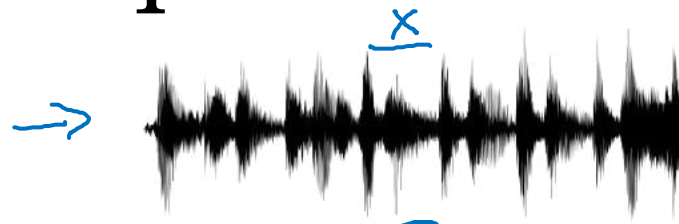
Speech recognition	x	y "The quick brown fox jumped over the lazy dog."
Music generation	$\emptyset$	
Sentiment classification	"There is nothing to like in this movie."	★☆☆☆☆
DNA sequence analysis	AGCCCCTGTGAGGAACTAG	AGCCCCTGTGAGGAACTAG
Machine translation	Voulez-vous chanter avec moi?	Do you want to 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.

$\rightarrow \quad x^{<1>} \qquad x^{<2>} \qquad x^{<3>} \qquad - - - \cdots \qquad x^{<t>} \qquad - - - \cdots \qquad x^{<9>}$

$T_x = 9$

$\rightarrow \quad y: \qquad\qquad 1 \qquad\quad 1 \qquad\quad 0 \qquad\quad 1 \qquad\qquad 1 \qquad\qquad 0 \quad\; 0 \quad 0 \qquad 0$

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

$T_y = 9$

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

$y^{(i)<t>} \qquad \uparrow$

$T_x^{(i)} = 9 \qquad\qquad 15$

$T_y^{(i)}$

Andrew Ng

Representing words

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

$x \longrightarrow y$

x: Harry Potter and Hermione Granger invented a new spell.
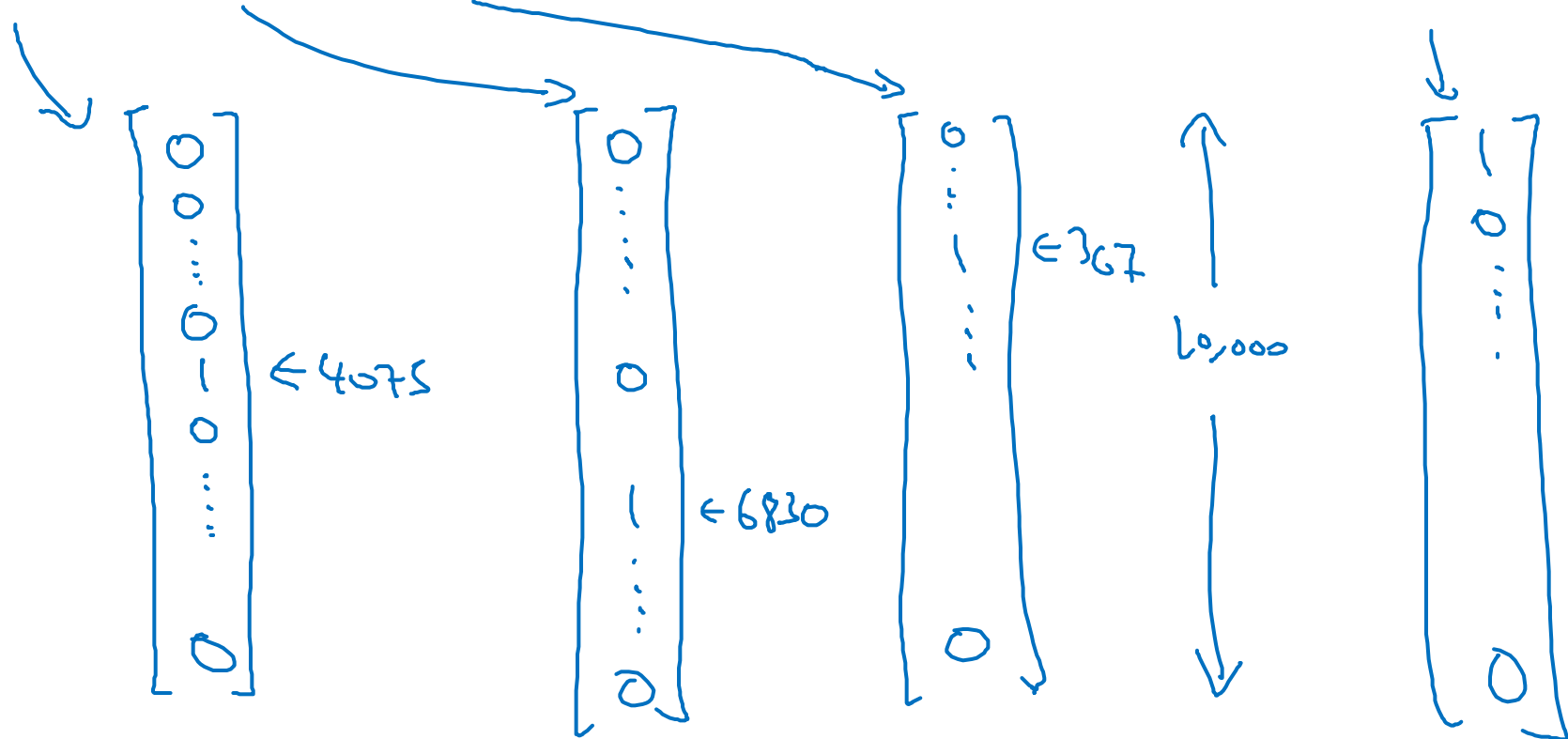
$x^{<1>}$ $x^{<2>}$ $x^{<3>}$ ... $x^{<9>}$

Vocabulary

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

<UNK>   10,000

One-hot

Andrew Ng

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
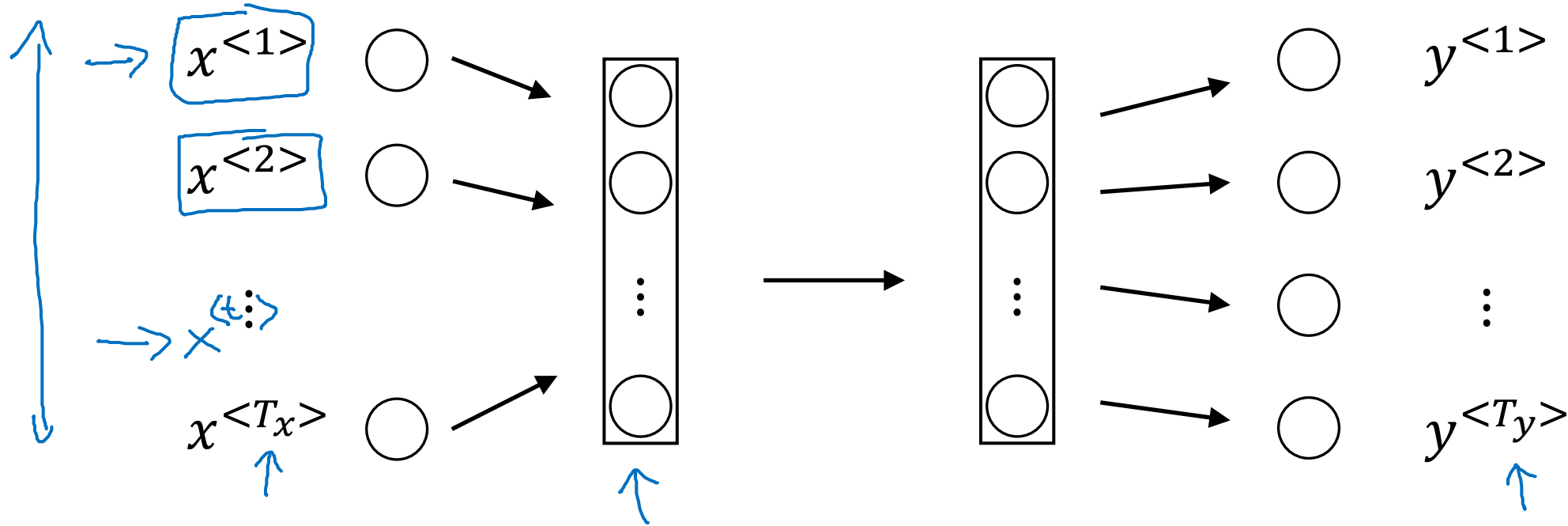Harry = 4075
Potter = 6830
Hermione = 4200
Gran… = 4000

deeplearning.ai
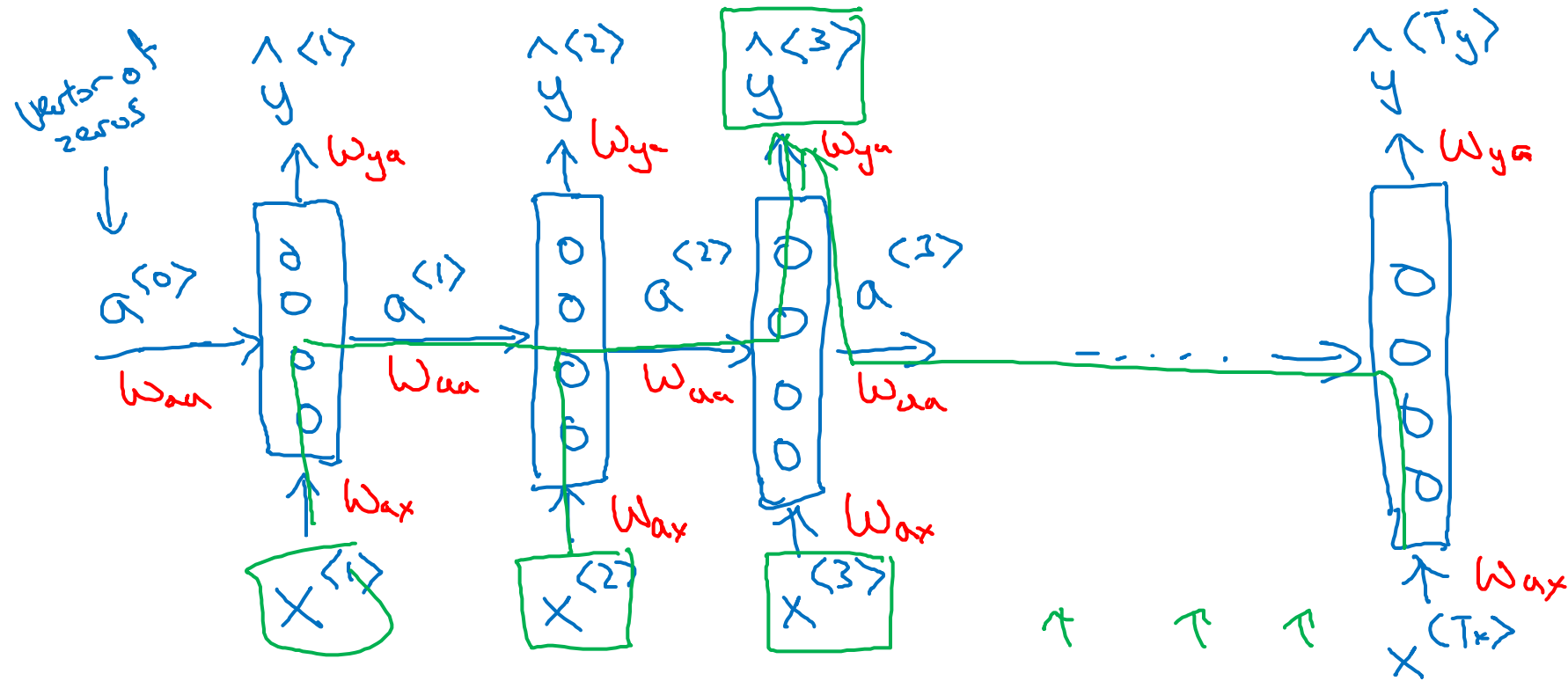
Recurrent Neural Networks

Recurrent Neural Network Model

Why not a standard network?
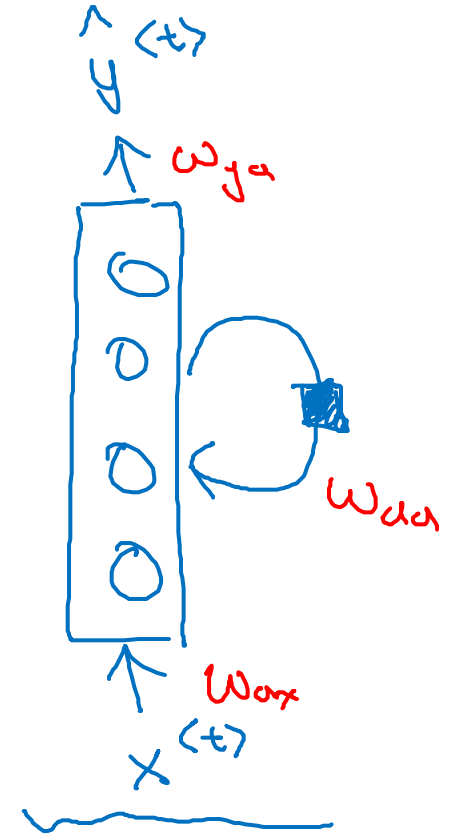

Problems:
- Inputs, outputs can be different lengths in different examples.
- Doesn't share features learned across different positions of text.
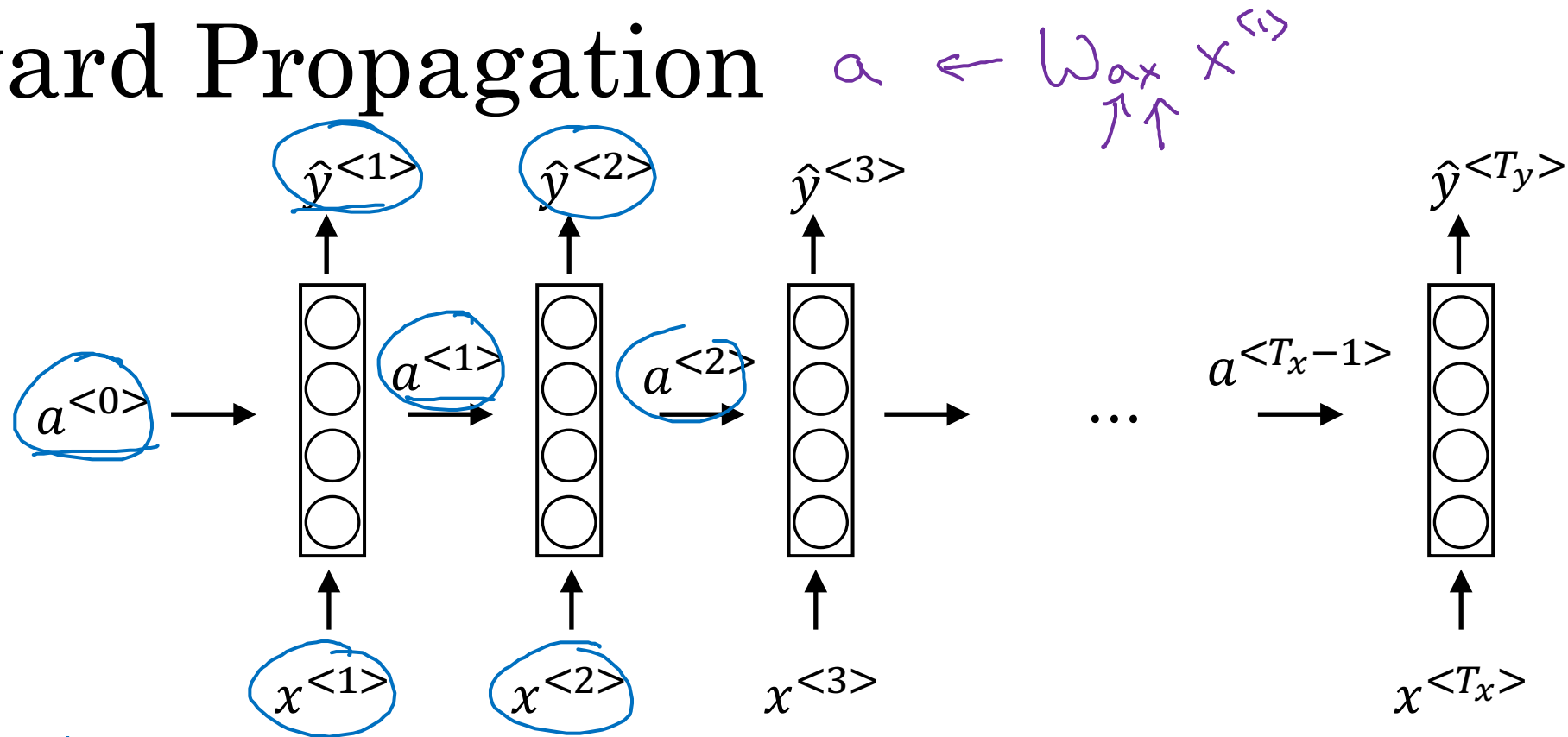
Recurrent Neural Networks

$T_x = T_y$

Bidirectional RNN (BRNN)

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

He said, "Teddy bears are on sale!"

Andrew Ng

Forward Propagation

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

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

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

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

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

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

Andrew Ng

Simplified RNN notation

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

$(100,100)$ 100 $(100,10,000)$ $10,000$

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

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

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

$$\left[W_{aa} ; W_{ax} \right] = W_a$$

100 100 $10\,000$ $(100, 10100)$

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

100 10000 10100

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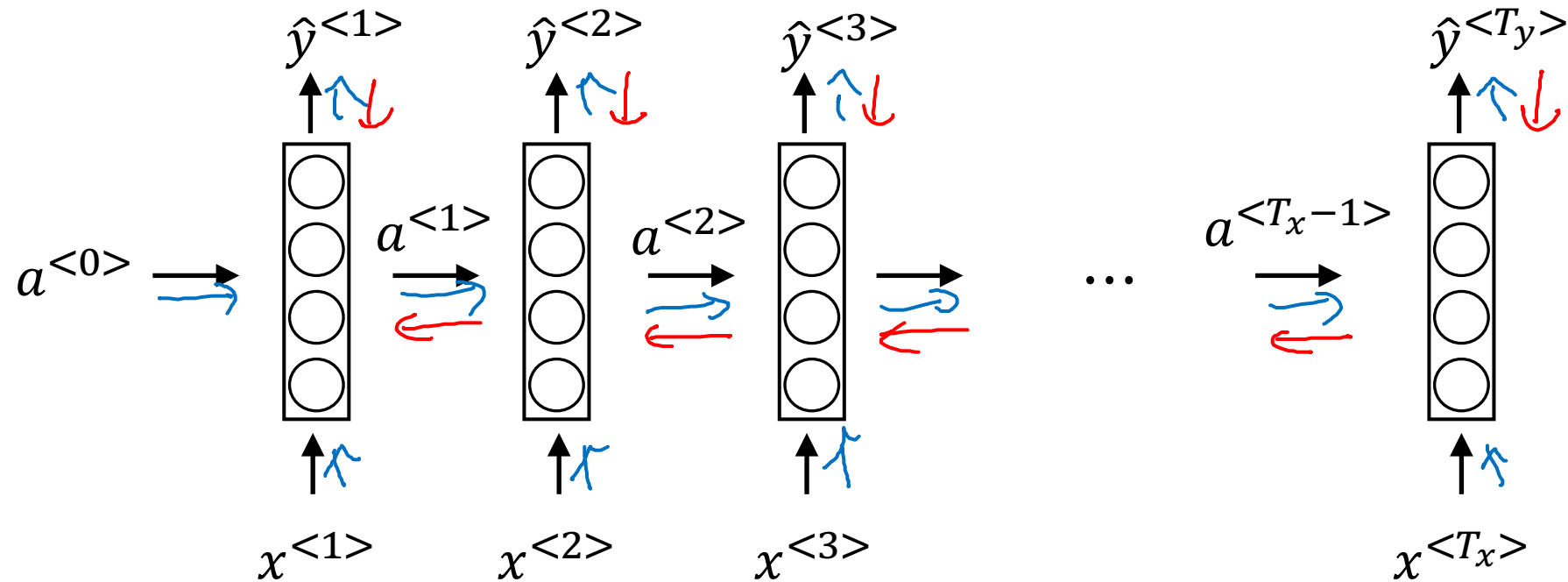
Recurrent Neural Networks

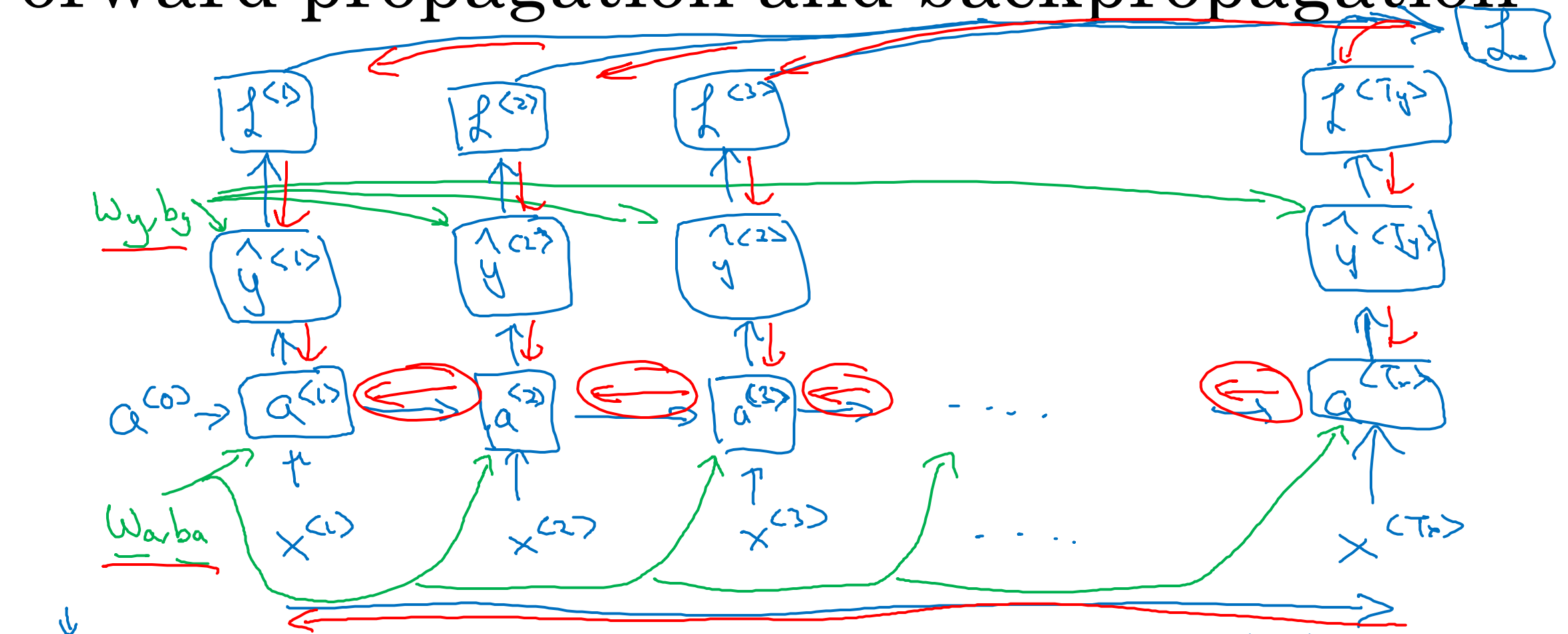deeplearning.ai

Backpropagation through time

Forward propagation and backpropagation

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

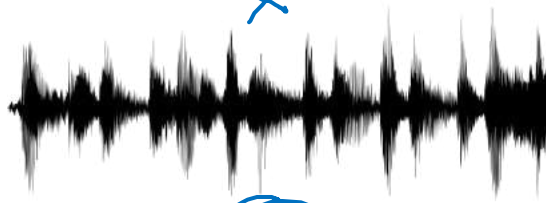# Recurrent Neural Networks

deeplearning.ai

Different types of RNNs

Examples of sequence data

T_x T_y

Speech recognition — x (audio waveform) → y "The quick brown fox jumped over the lazy dog."

Music generation — ∅ → (musical notation)

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

DNA sequence analysis — AGCCCCTGTGAGGAACTAG → AG CCCCTGTGAGGAACT AG

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

Video activity recognition — (video frames) → Running

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

Andrew Ng

Examples of RNN architectures

$T_x = T_y$

Sentiment classification
$x = text$
$y = 0/1 \qquad 1 \cdots 5$

Many-to-many

Many-to-one

one-to-one

Examples of RNN architectures

Music generation

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

One-to-many

$x = \phi$

Machine translation

encoder

decoder

Many - to - many

Andrew Ng

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
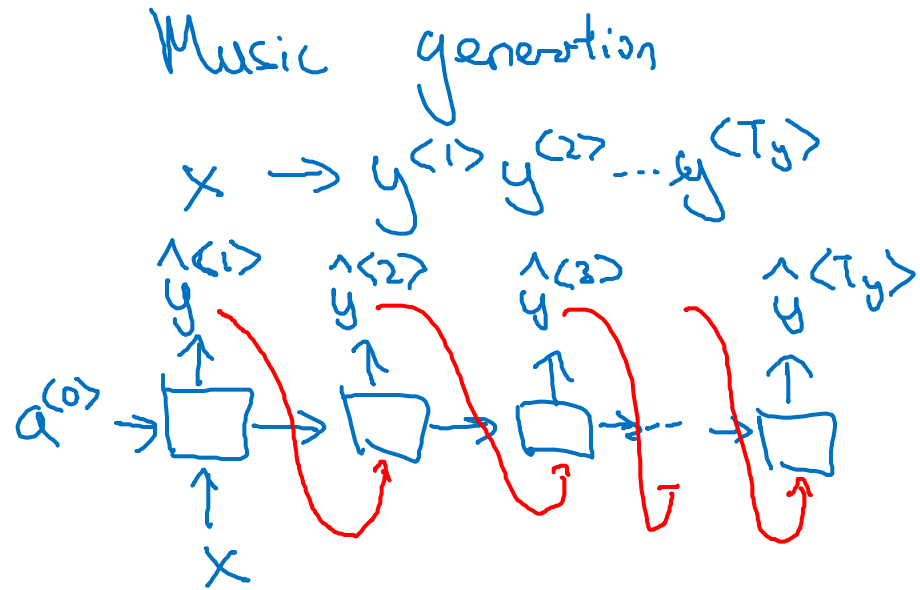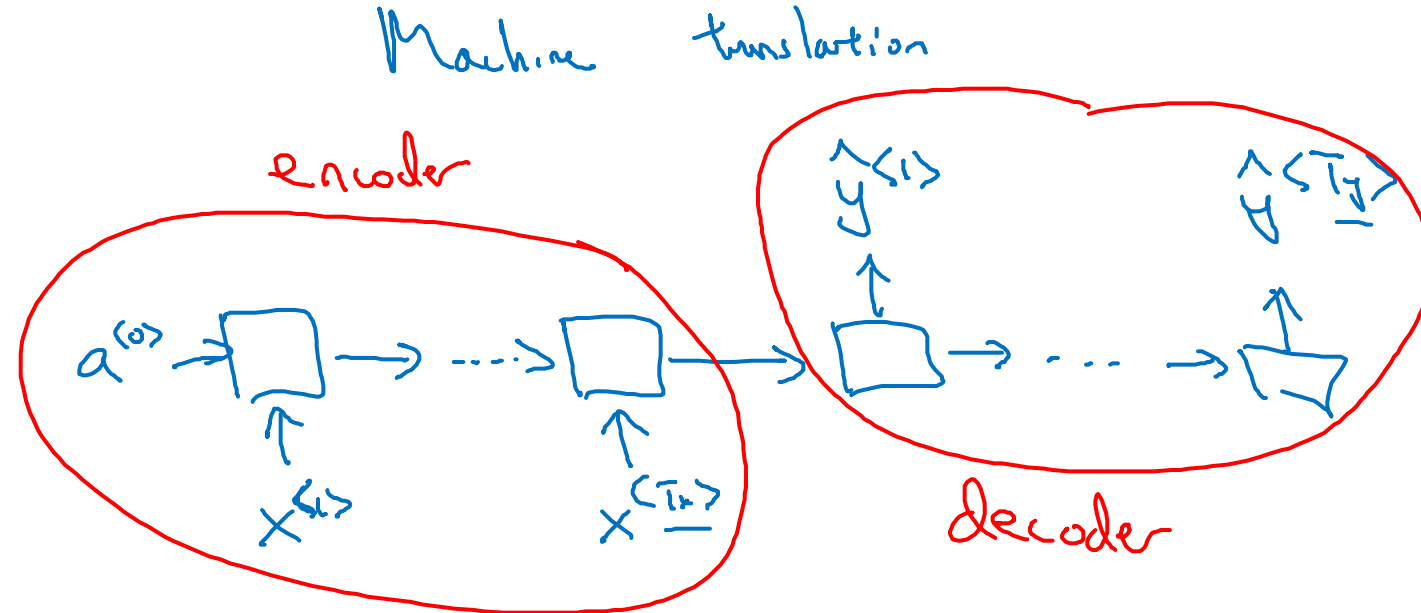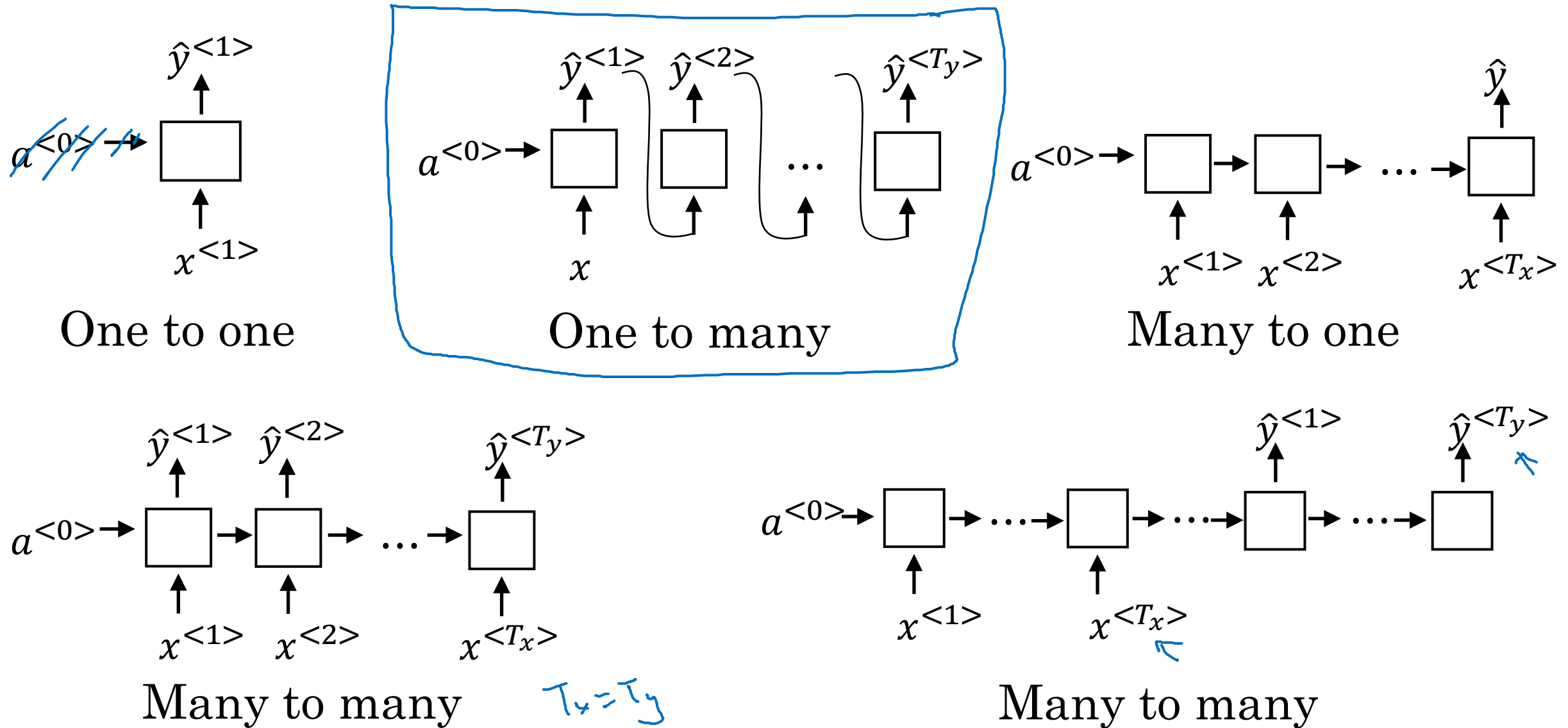
Recurrent Neural Networks

Language model and sequence generation

deeplearning.ai

What is language modelling?

Speech recognition

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

$\rightarrow$ The apple and <u>pear</u> salad.

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

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

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

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

Andrew Ng

Language modelling with an RNN

Training set: large <u>corpus</u> of english text.

Tokenize

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

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

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

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

$<UNK>$

10,000

Andrew Ng

RNN model

$P(a) \; P(aaron) \cdots P(cats) \cdots P(zulu)$
$P(\text{<UNK>})$
$P(\text{<EOS>})$

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

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

$P(\text{<EOS>} \mid \dots)$

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

$a^{<0>} = \vec{0}$ $a^{<1>}$ $a^{<2>}$ $a^{<3>}$ $a^{<9>}$

$x^{<1>} = \vec{0}$ $x^{<2>} = y^{<1>}$ $x^{<3>} = y^{<2>}$ $x^{<9>} = y^{<8>}$
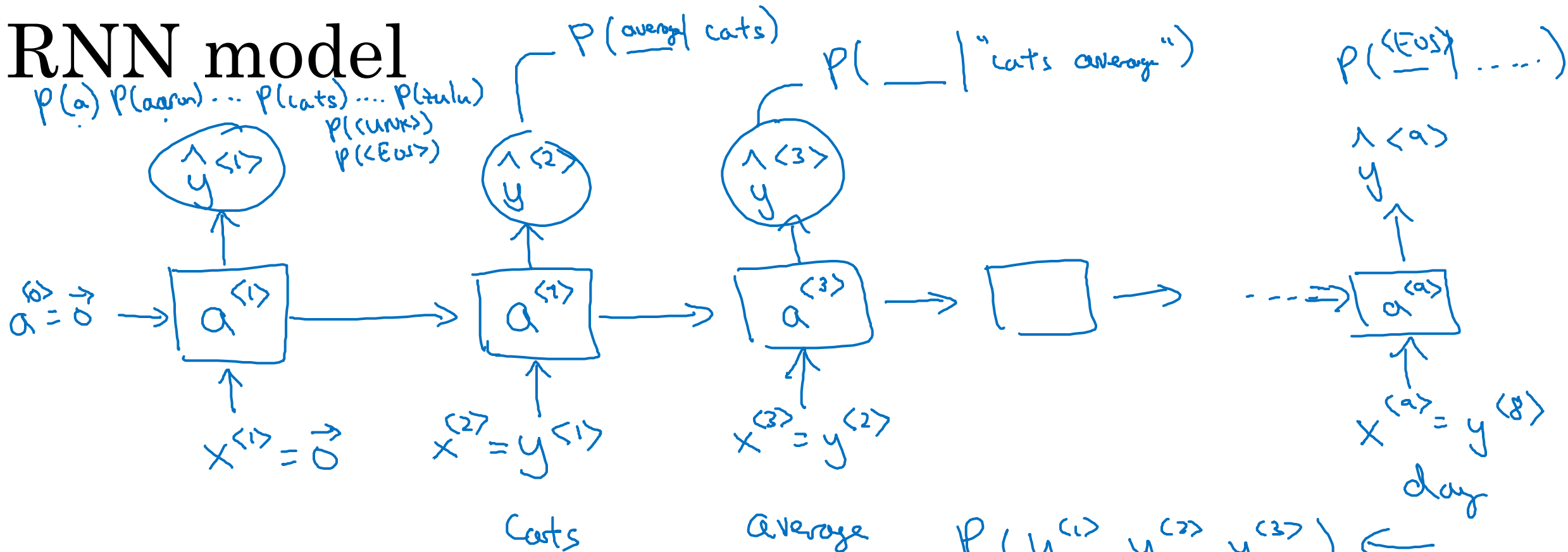
Cats   average   day

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

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

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

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

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

Andrew Ng

deeplearning.ai

Recurrent Neural Networks

Sampling novel sequences

Sampling a sequence from a trained RNN

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

Training:

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

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

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

Sampling:

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

$a^{<0>} \rightarrow \boxed{a^{<1>}} \rightarrow \boxed{a^{<2>}} \rightarrow \boxed{a^{<3>}} \rightarrow \dots \rightarrow \boxed{}$
$= 0$

$x^{<1>} = 0$ The $x^{<2>} = \hat{y}^{<1>}$ $y^{<T_x - 1>}$

$<EOS>$

$<UNK>$

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

Andrew Ng
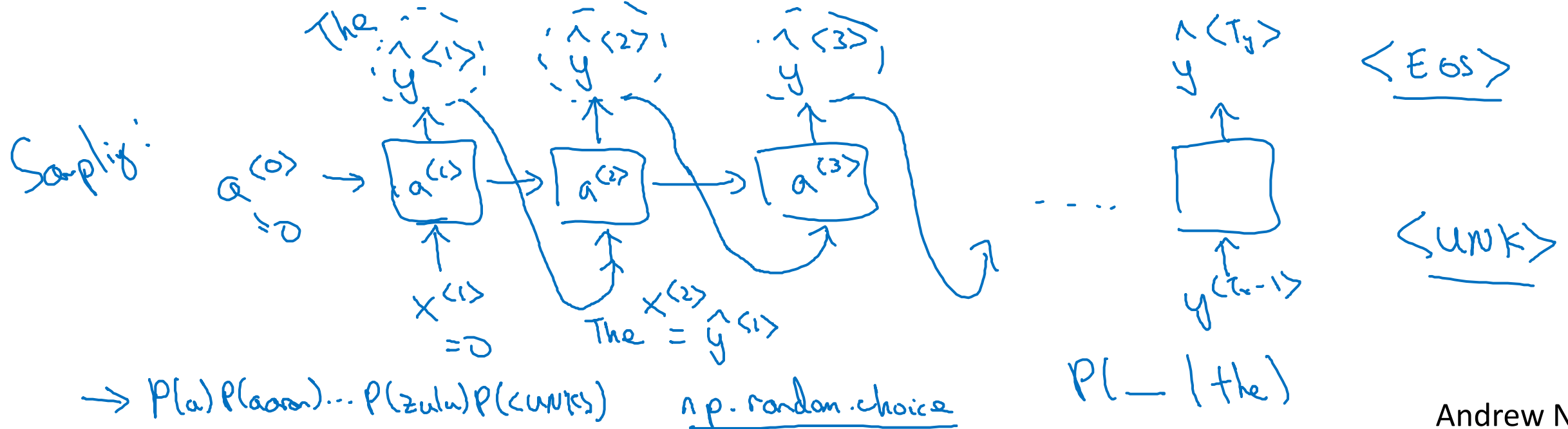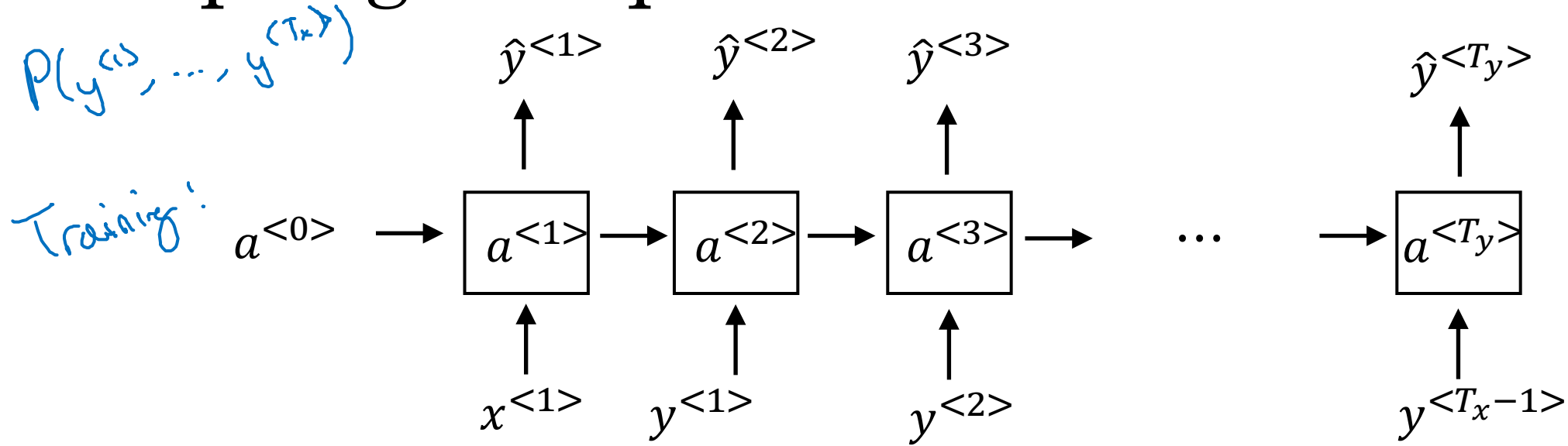
Character-level language model

→ Vocabulary = [a, aaron, ..., zulu, <UNK>] ←

→ Vocabulary = [a, b, c, ..., z, ⎵, ., , , ;, 0, ..., 9, A, ..., Z]

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

Cat   average
↑ ↑ ↑ ↑ ...

May

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

$a^{<0>}$ → $a^{<1>}$ → $a^{<2>}$ → $a^{<3>}$ → ... → $a^{<T_y>}$

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

Andrew Ng
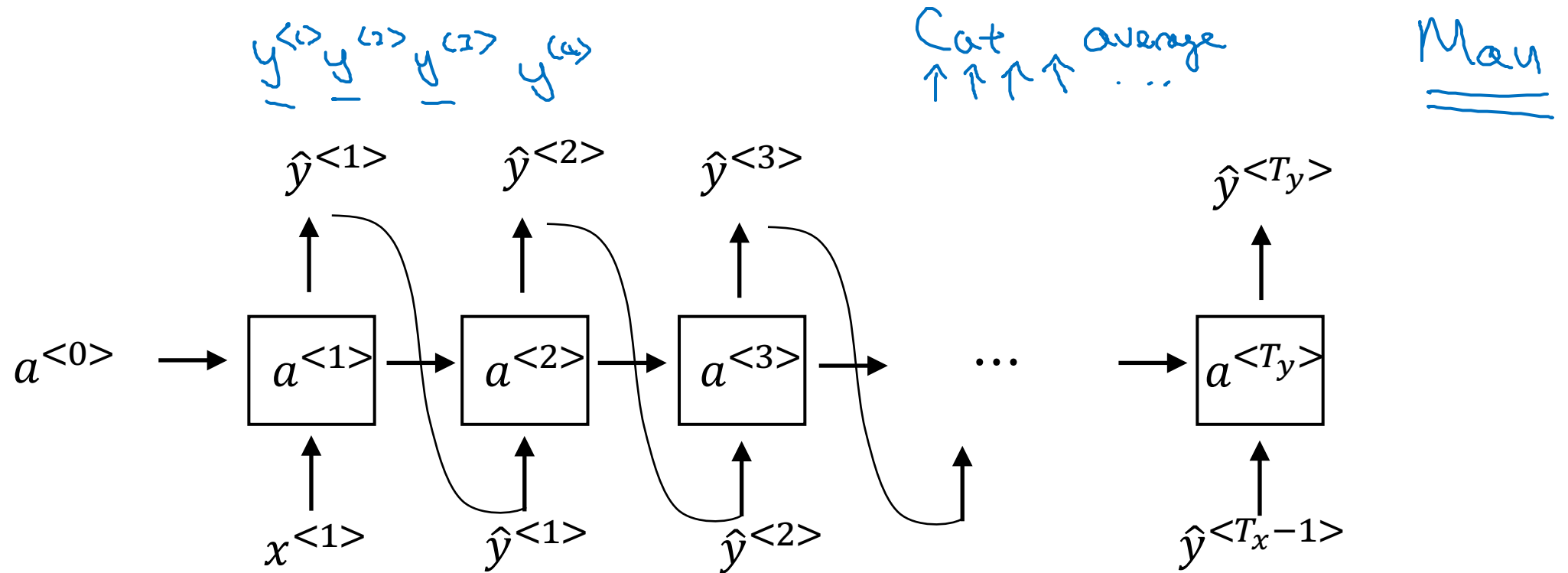
Sequence generation

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.
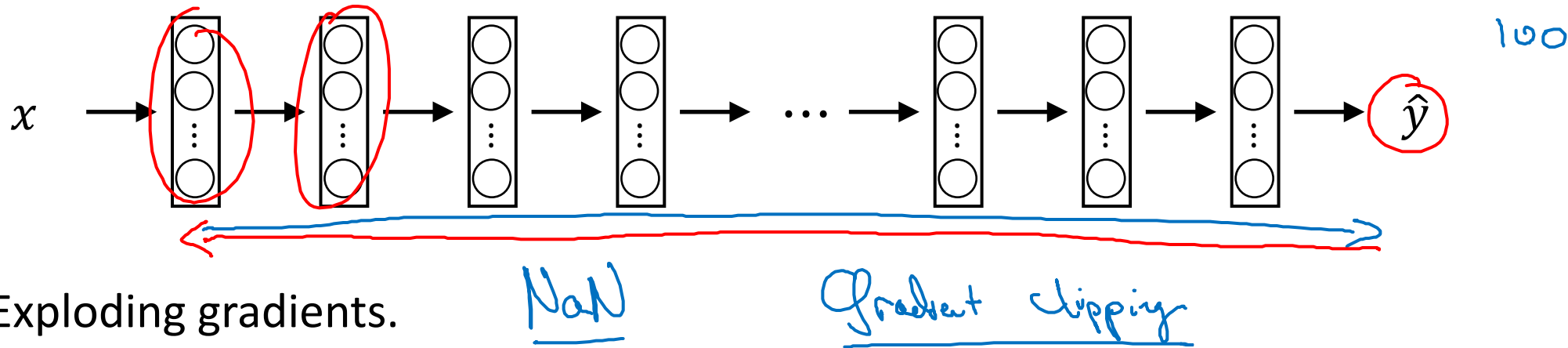
For whose are ruse of mine eyes heaves.

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 $\cdots$ $\hat{y}$

100

Exploding gradients.   NaN   Gradient clipping
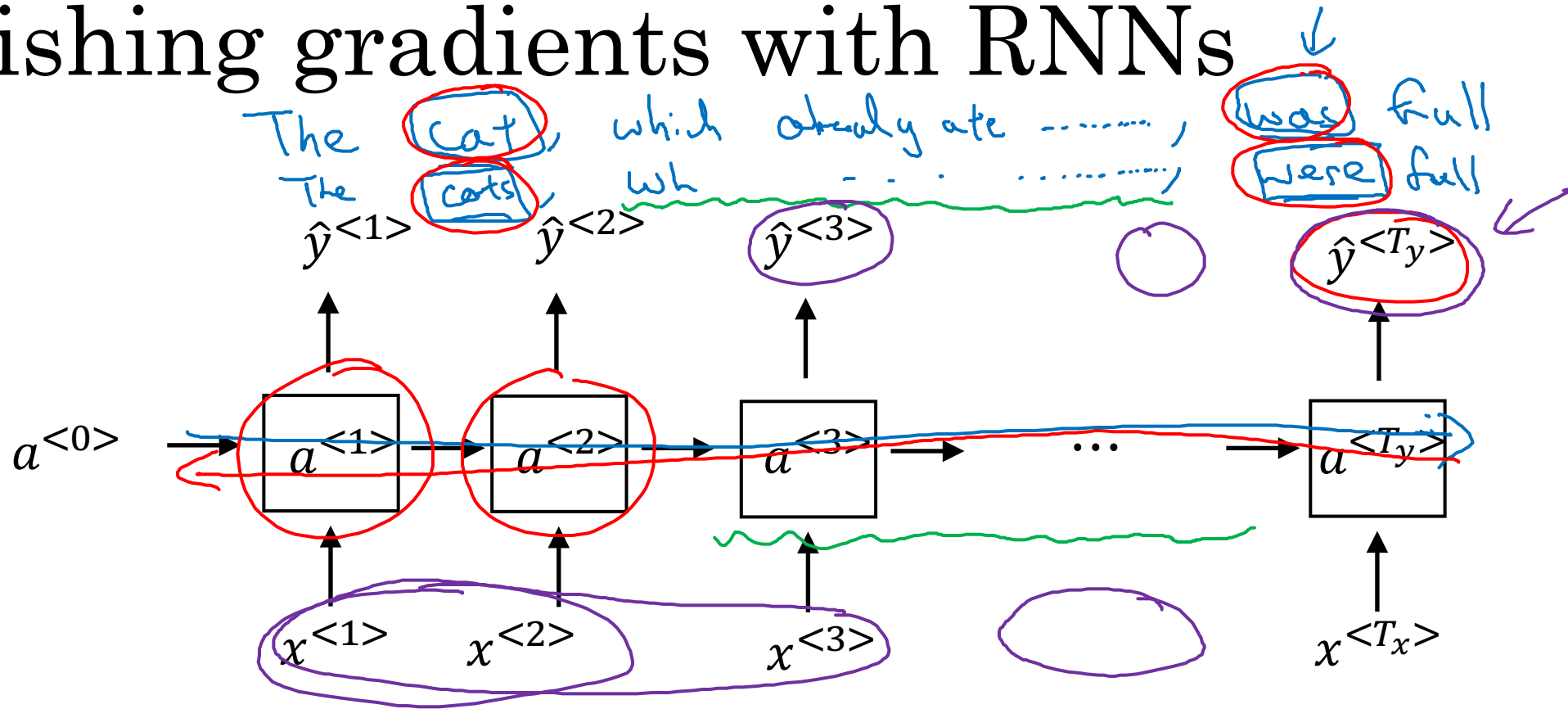
Recurrent Neural Networks

deeplearning.ai

Gated Recurrent Unit (GRU)

RNN unit

$y^{<t>}$

Softmax

$a^{<t-1>}$

$a^{<t>}$

tanh

$x^{<t>}$
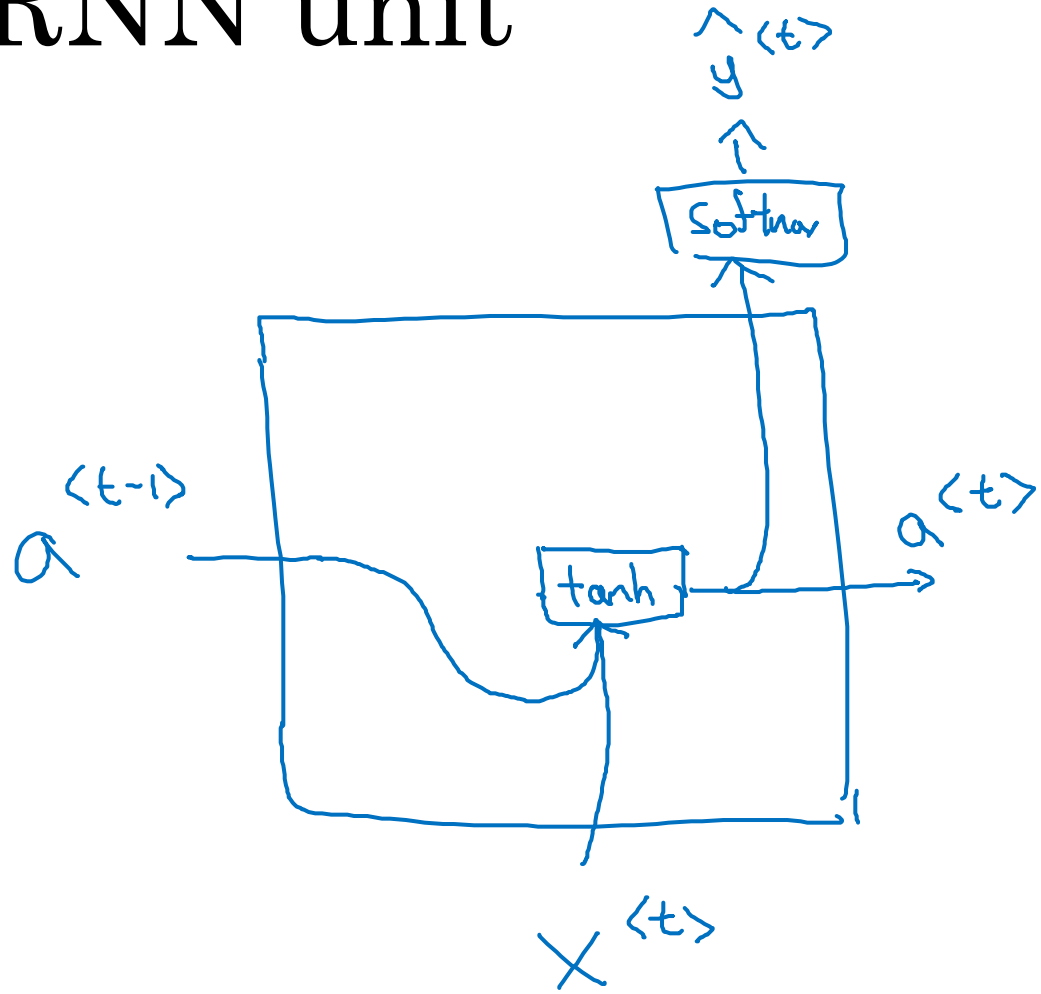
tanh

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

GRU (simplified)

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

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

$\tilde{C}^{\langle t \rangle}$

Γ_u

tanh σ

$x^{\langle t \rangle}$

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

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

$c^{\langle t \rangle} = 1$

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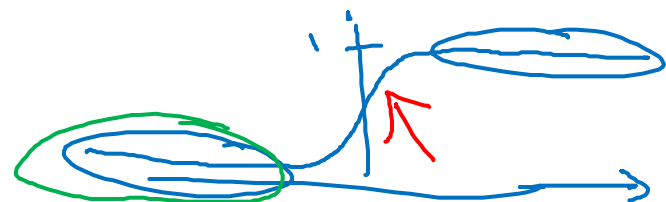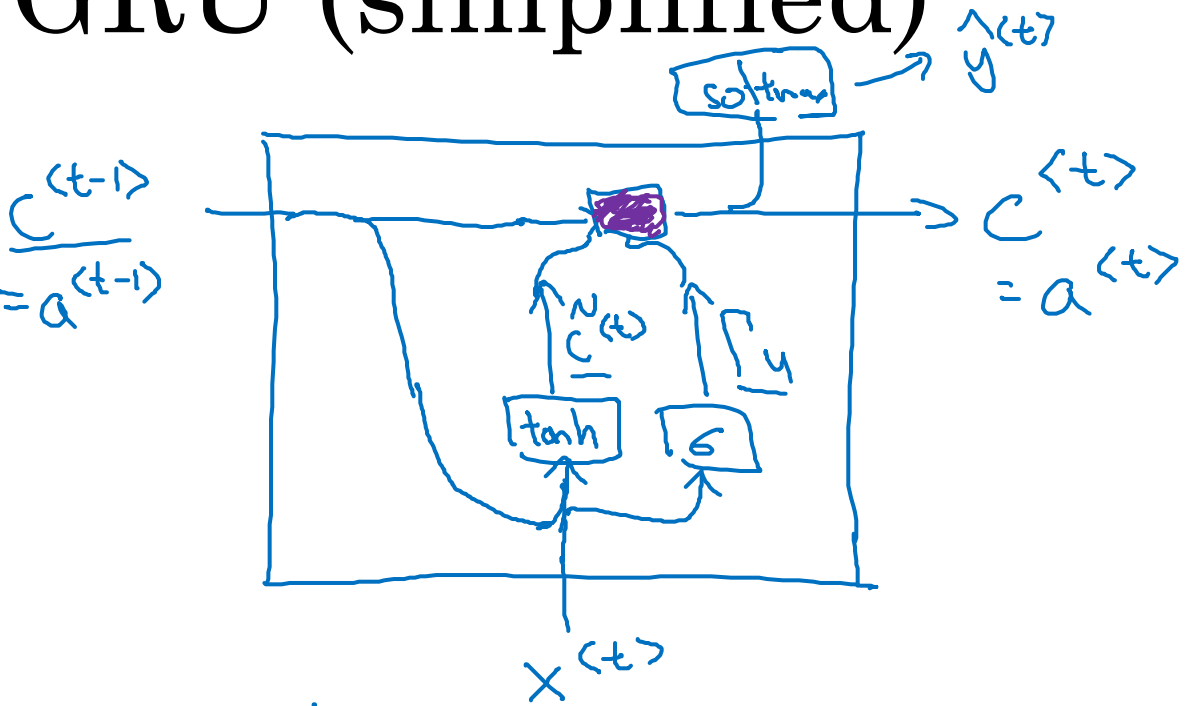
C = memory cell

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

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

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

"update"

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

$\Gamma_u = 1$

element-wise

$\Gamma_u = 0.0000001$
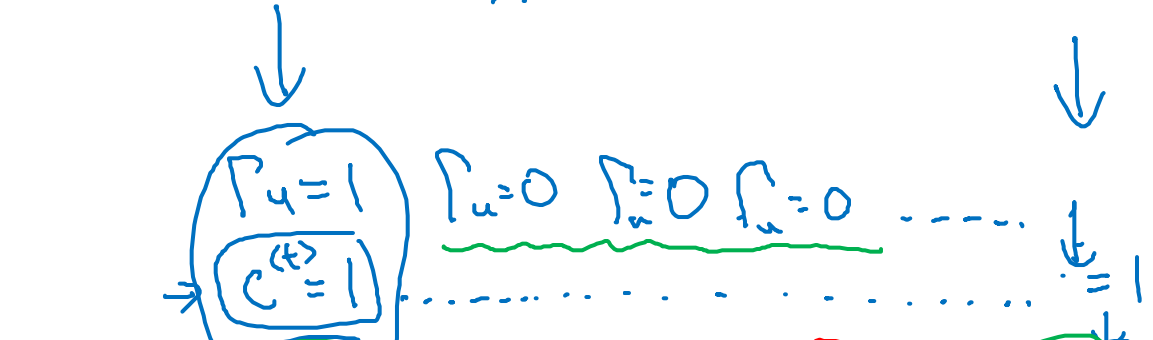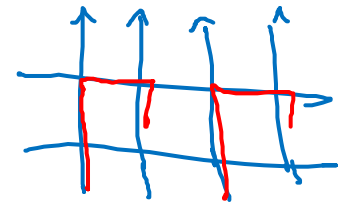
Gate

[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

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

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

$*$

The cat, which ate already, was full.

Andrew Ng

Recurrent Neural Networks

deeplearning.ai

LSTM (long short term memory) unit

GRU and LSTM

GRU

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

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

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

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

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

Γ_f

LSTM

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

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

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

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

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

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

*tanh

[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

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

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

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

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

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

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

Andrew Ng

LSTM in pictures

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

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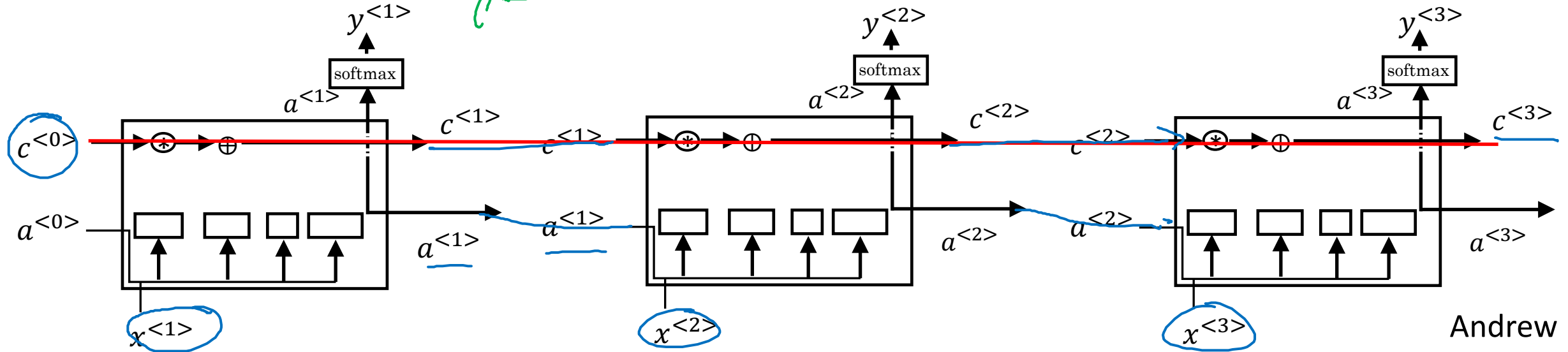
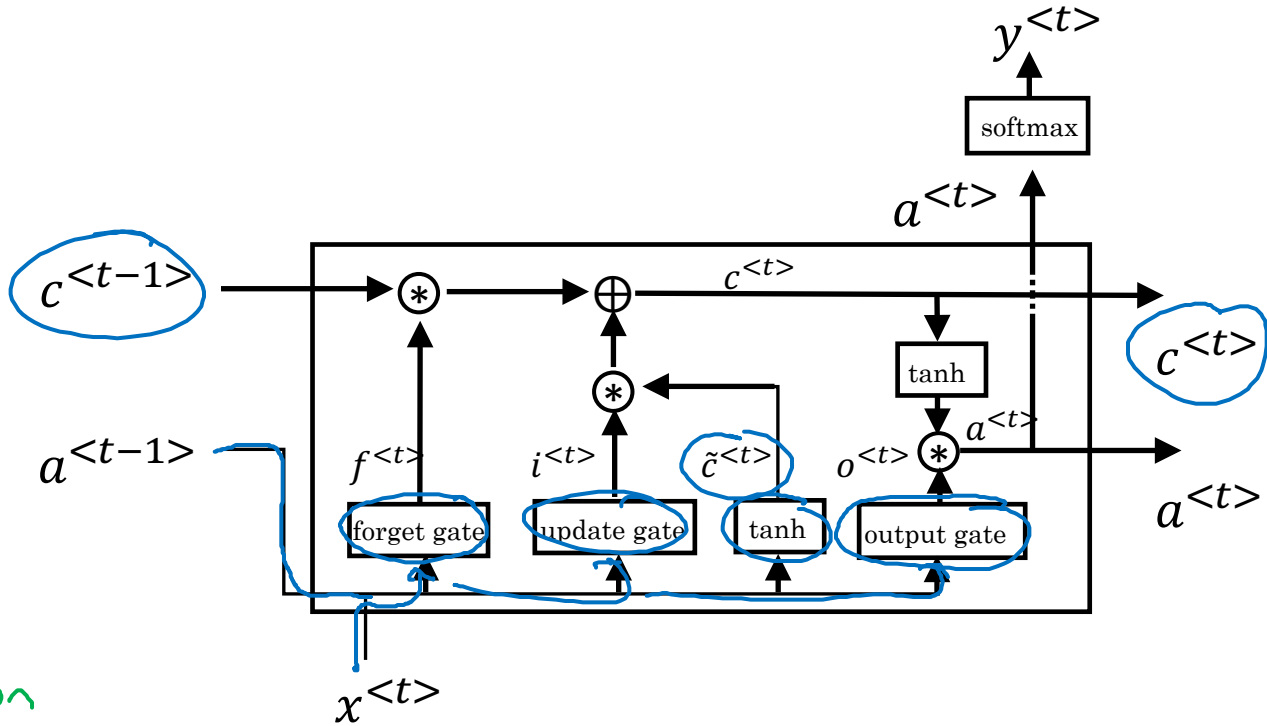$$\Gamma_f = \sigma(W_f[a^{<t-1>}, x^{<t>}] + b_f)$$

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

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

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

peephole connection

$c^{<t-1>}$

Andrew Ng

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)

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

GRU
LSTM

Acyclic graph

He said, "Teddy Roosevelt ..."

BRNN w/ LSTM

Andrew Ng
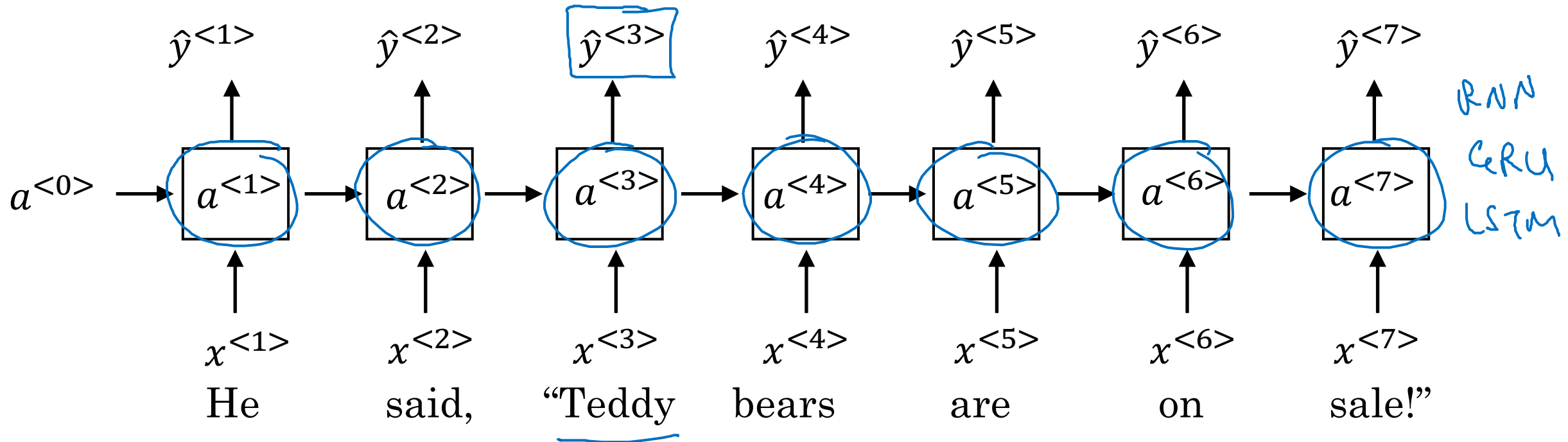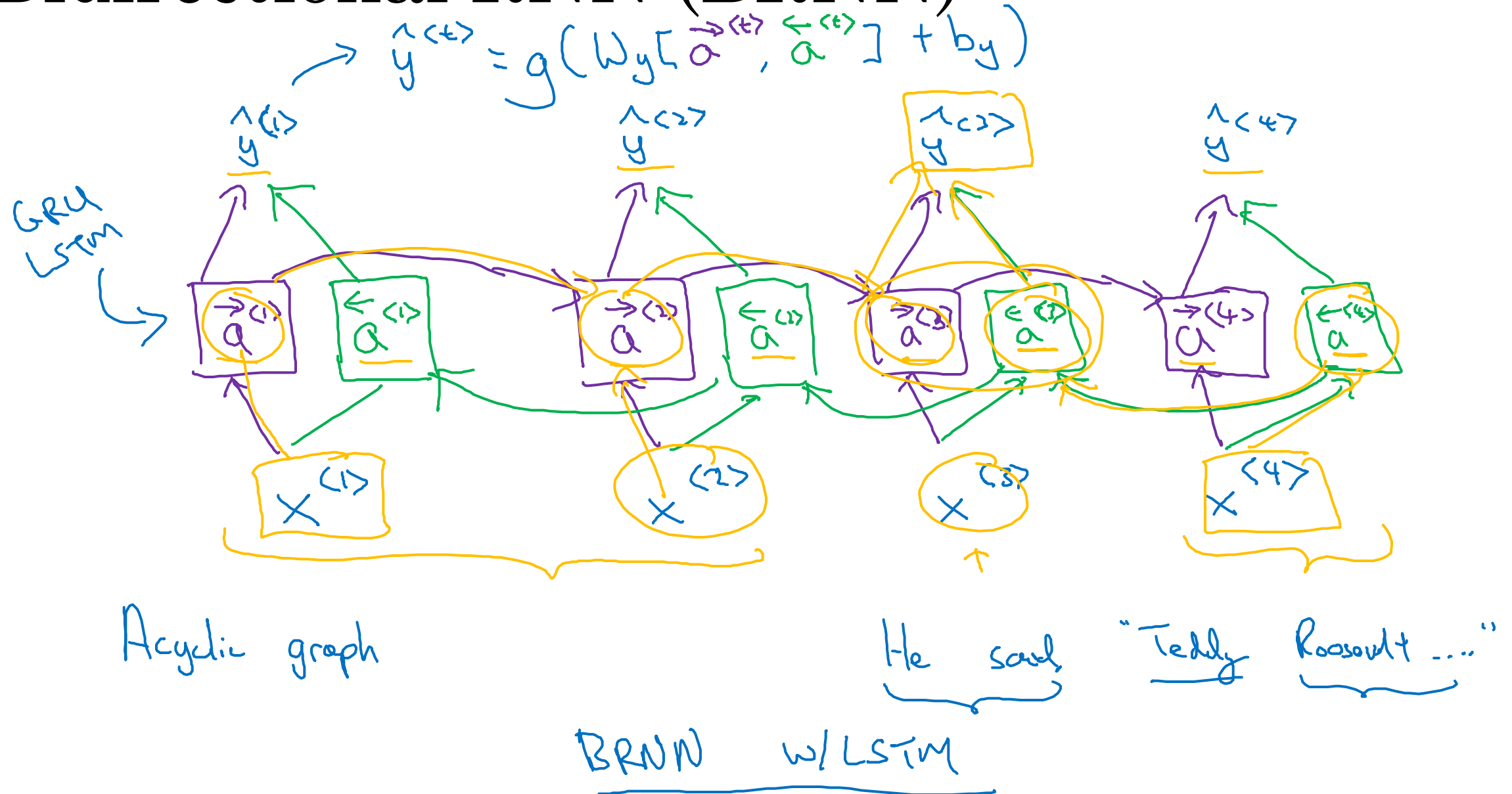
Recurrent Neural Networks

Deep RNNs

deeplearning.ai

Deep RNN example

$y^{<1>}$

$y^{<2>}$

$a^{[\ell]<t>}$

$y^{<3>}$

$y^{<4>}$

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

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

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

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

$a^{<0>}$

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

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

$\boxed{a^{[3]}}$

$\boxed{a^{[2]}}$

$\boxed{a^{[1]}}$

x
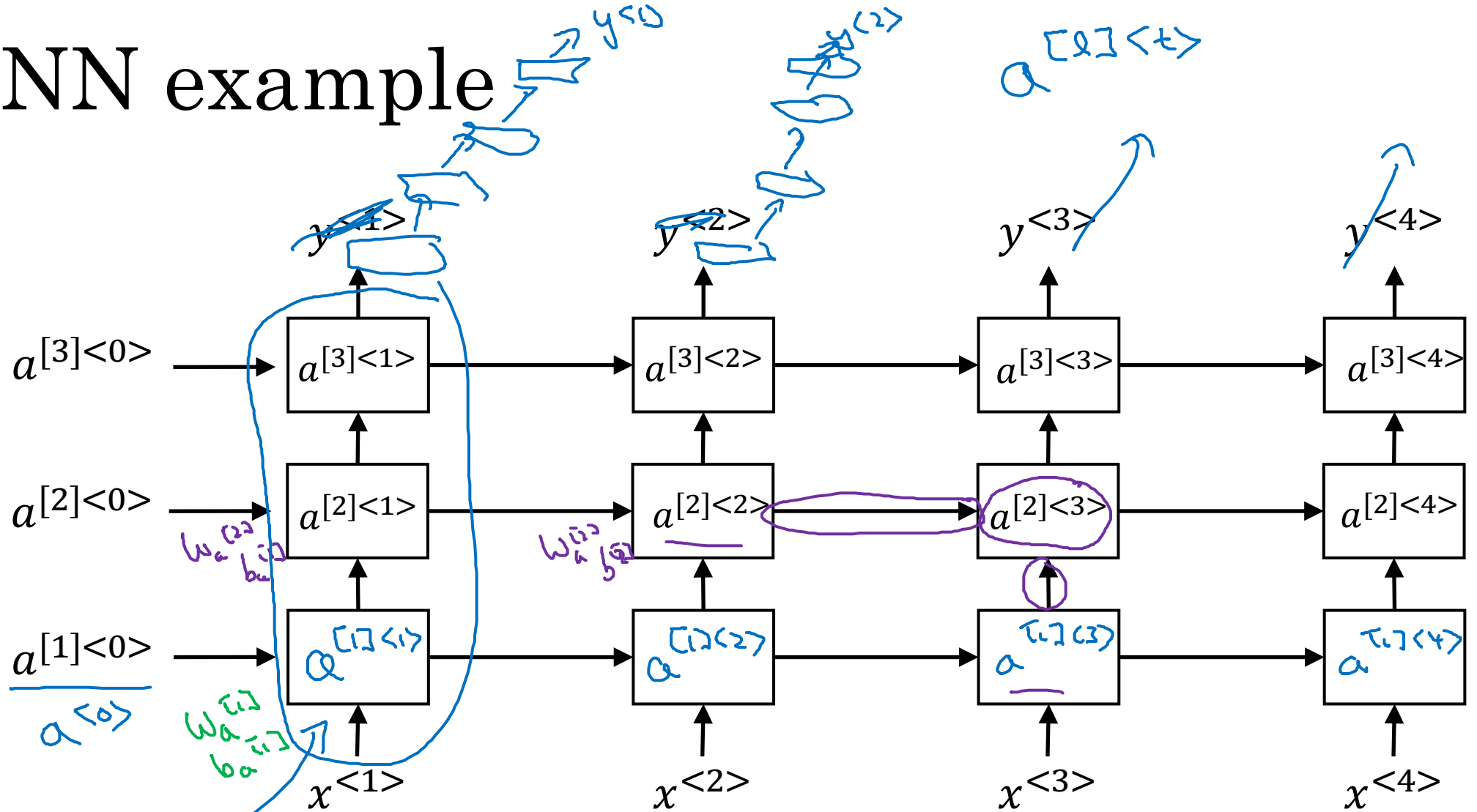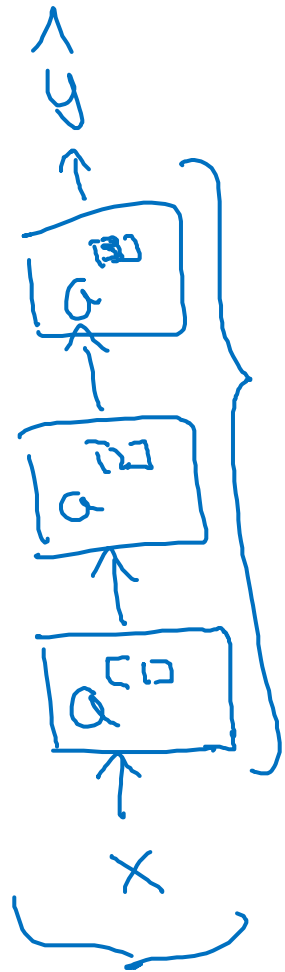
RNN
GRU
LSTM
BRNN

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

Andrew Ng

NLP and Word Embeddings

Word representation

deeplearning.ai

Word representation

V = [a, aaron, ..., zulu, <UNK>]

$|V| = 10,000$

1-hot representation

Man (5391)	Woman (9853)	King (4914)	Queen (7157)	Apple (456)	Orange (6257)
$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$

O_{5391} O_{9853}

I want a glass of orange _juice_.

I want a glass of apple _____.

Andrew Ng
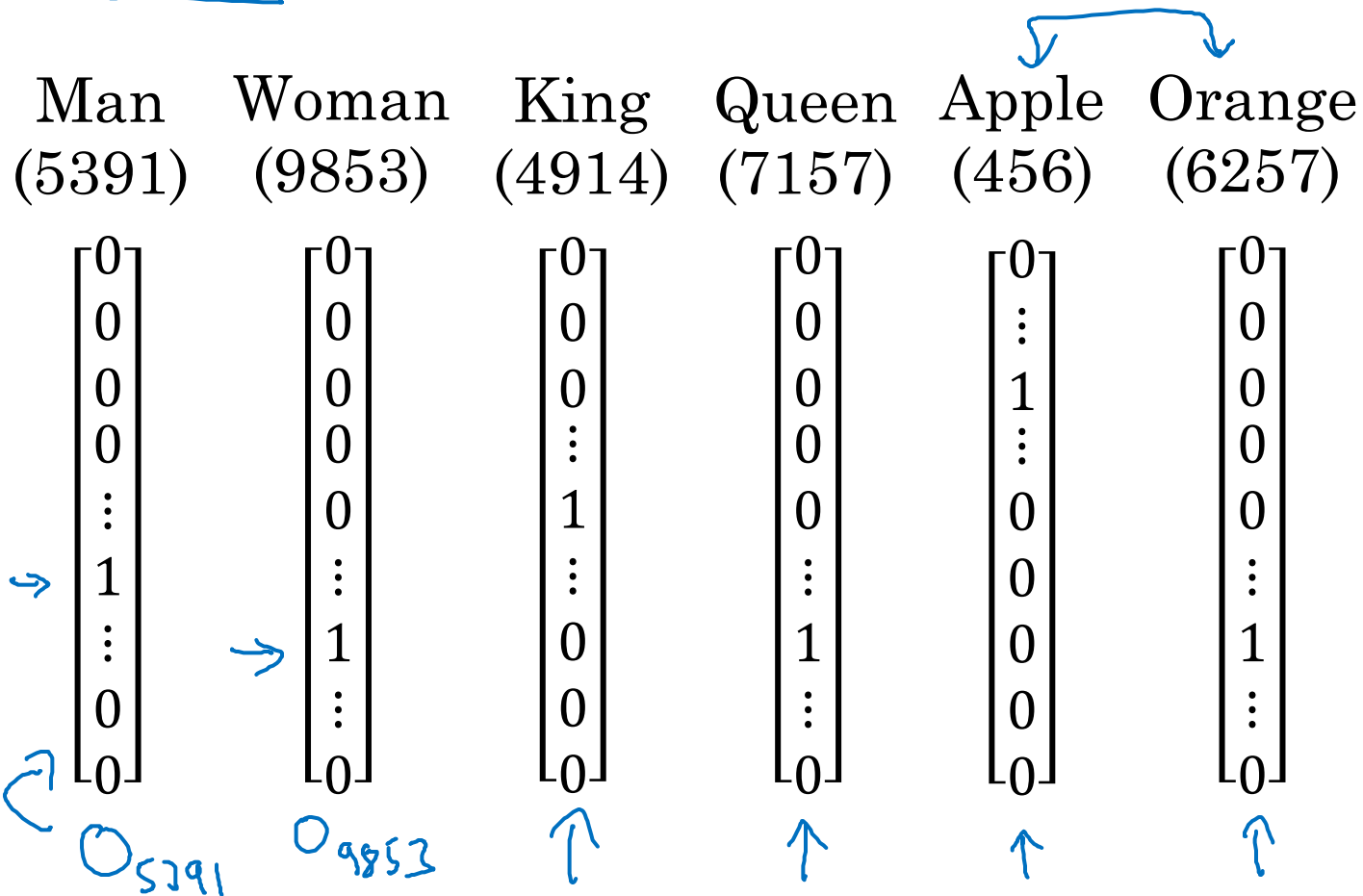
Featurized representation: word embedding

	Man (5391)	Woman (9853)	King (4914)	Queen (7157)	Apple (456)	Orange (6257)
Gender	−1	1	−0.95	0.97	0.00	0.01
Royal	0.01	0.02	0.93	0.95	−0.01	0.00
Age	0.03	0.02	0.7	0.69	0.03	−0.02
Food	0.04	0.01	0.02	0.01	0.95	0.97

300

size
cost
alive   verb

e_{5391} e_{9853}

I want a glass of orange juice.

I want a glass of apple juice.

Andrew Ng

Visualizing word embeddings

$\rightarrow 300D$

$\downarrow$

$2D$

t-SNE

apple   orange

$300D$

[van der Maaten and Hinton., 2008. Visualizing data using t-SNE]

Andrew Ng
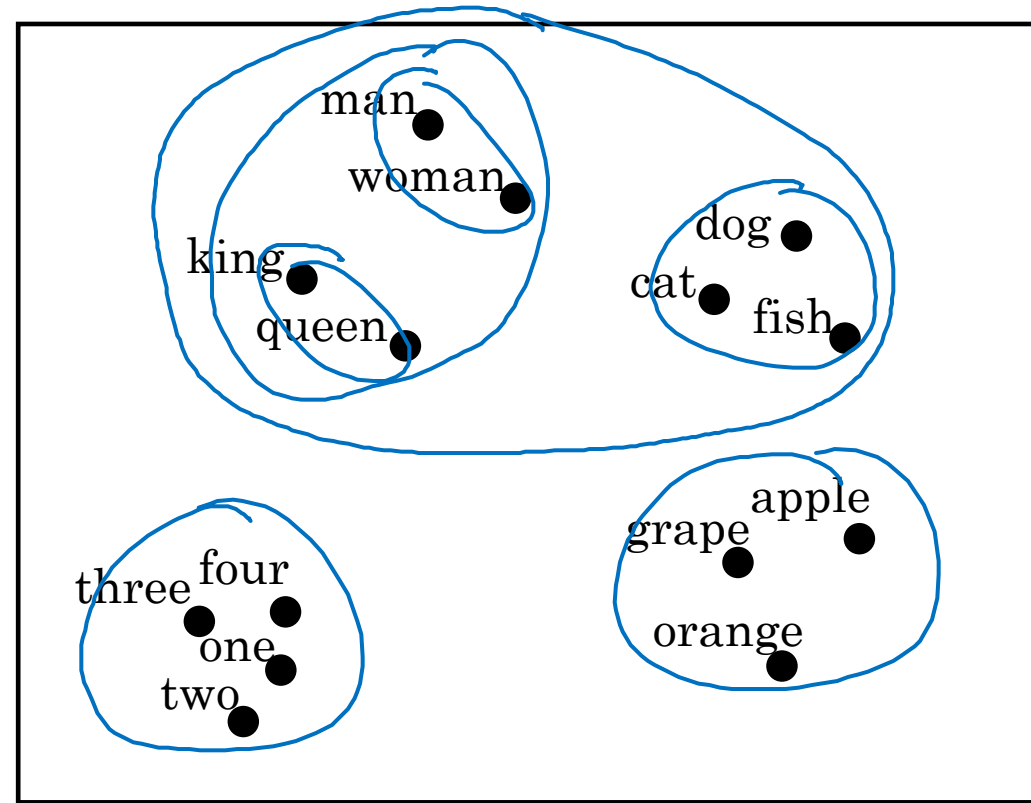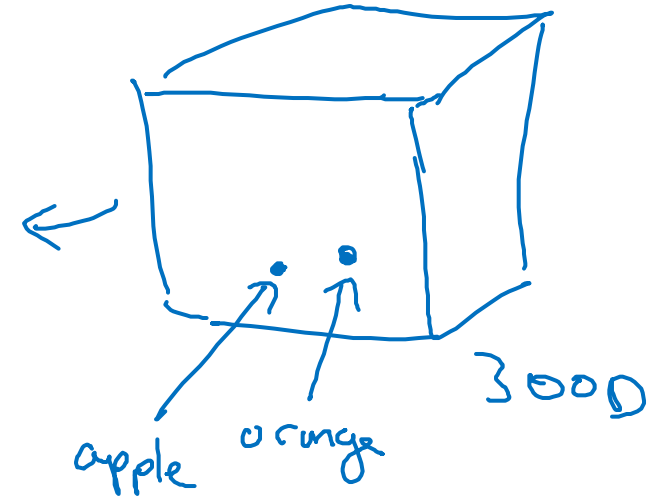
NLP and Word Embeddings

deeplearning.ai

Using word embeddings

Named entity recognition example

1   1   0   0   0   0

$x^{(1)}$ $x^{(2)}$ $x^{(3)}$

Sally   Johnson   is   an   orange   farmer

Robert   Lin   is   an   apple   farmer

a   durian   cultivator.

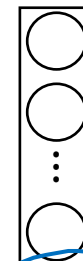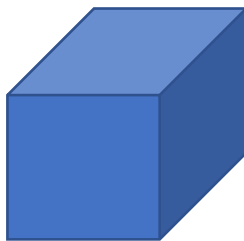→ 1 B words – 100B words
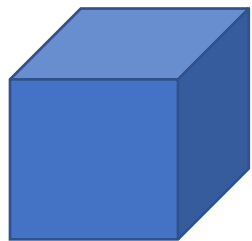
→ 100k words

BRNN

Andrew Ng
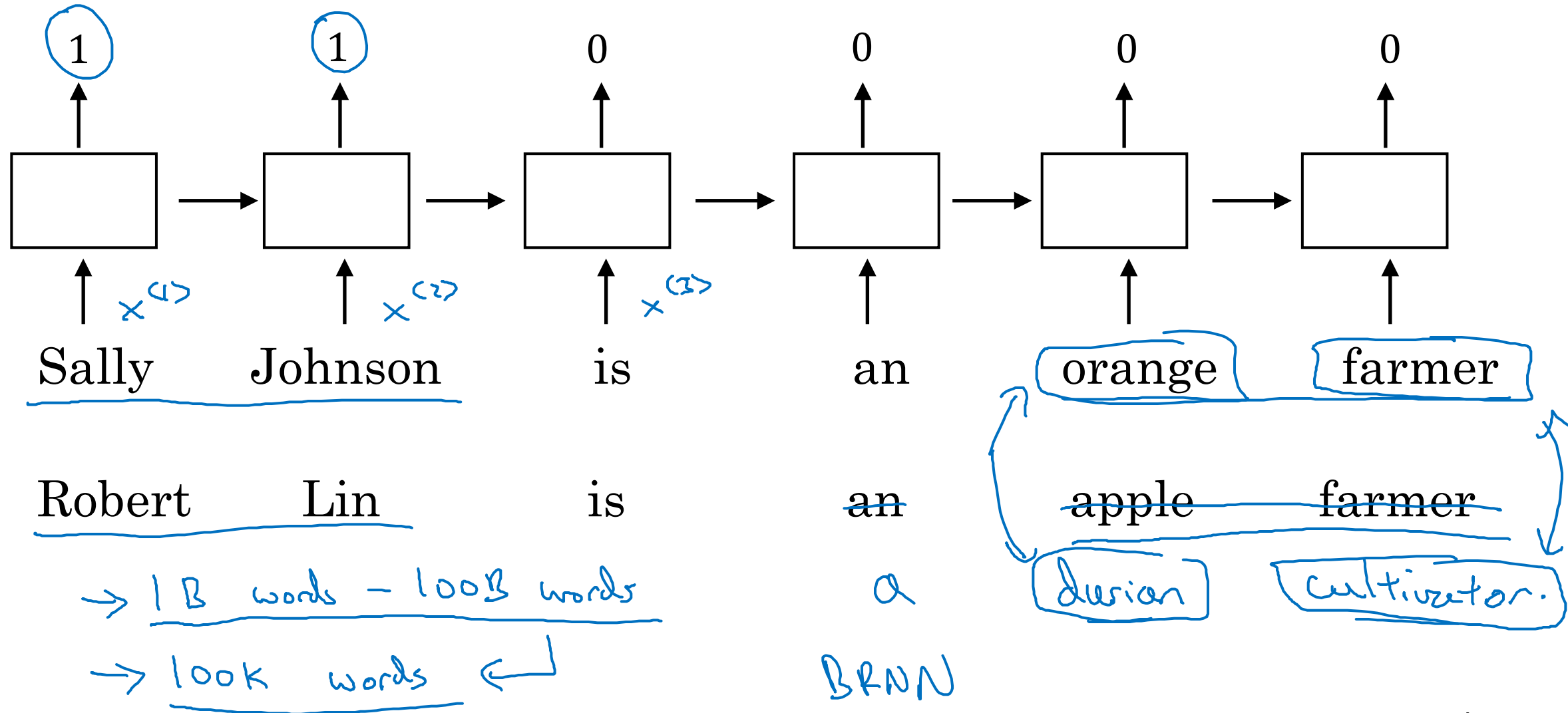
Transfer learning and word embeddings

1. Learn word embeddings from large text corpus. (1-100B words)

   (Or download pre-trained embedding online.)

2. Transfer embedding to new task with smaller training set. (say, 100k words) $\rightarrow 10,000$ $\rightarrow 300$

3. Optional: Continue to finetune the word embeddings with new data.

Andrew Ng

Relation to face encoding (embedding) 128D

f(x^{(i)})

f(x^{(j)}) 12D

$|V| = 10,000$

$e_1, \ldots, e_{10,000}$

$x^{(i)}$

$x^{(j)}$

$\hat{y}$

[Taigman et. al., 2014. DeepFace: Closing the gap to human level performance]   Andrew Ng
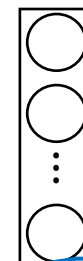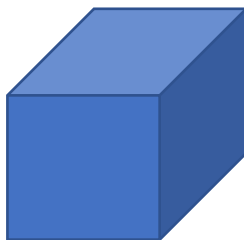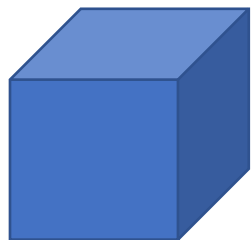
NLP and Word Embeddings

Properties of word embeddings

deeplearning.ai

Analogies

	Man (5391)	Woman (9853)	King (4914)	Queen (7157)	Apple (456)	Orange (6257)
Gender	-1	1	-0.95	0.97	0.00	0.01
Royal	0.01	0.02	0.93	0.95	-0.01	0.00
Age	0.03	0.02	0.70	0.69	0.03	-0.02
Food	0.09	0.01	0.02	0.01	0.95	0.97

e_{5391}
e_{man} e_{woman}

$Man \rightarrow Woman$ as $King \rightarrow ?$ $Queen$

$e_{man} - e_{woman} \approx e_{king} - e_{?}$
$\quad \uparrow Queen$

$e_{man} - e_{woman} \approx \begin{bmatrix} -2 \\ 0 \\ 0 \\ 0 \end{bmatrix}$

$e_{king} - e_{queen} \approx \begin{bmatrix} -2 \\ 0 \\ 0 \\ 0 \end{bmatrix}$

[Mikolov et. al., 2013, Linguistic regularities in continuous space word representations]

Andrew Ng

Analogies using word vectors

man ·   dog ·

king ·   woman ·   cat ·   ·

queen ·   fish

three ·   four ·   apple

· · one ·   grape · ·

two   orange ·
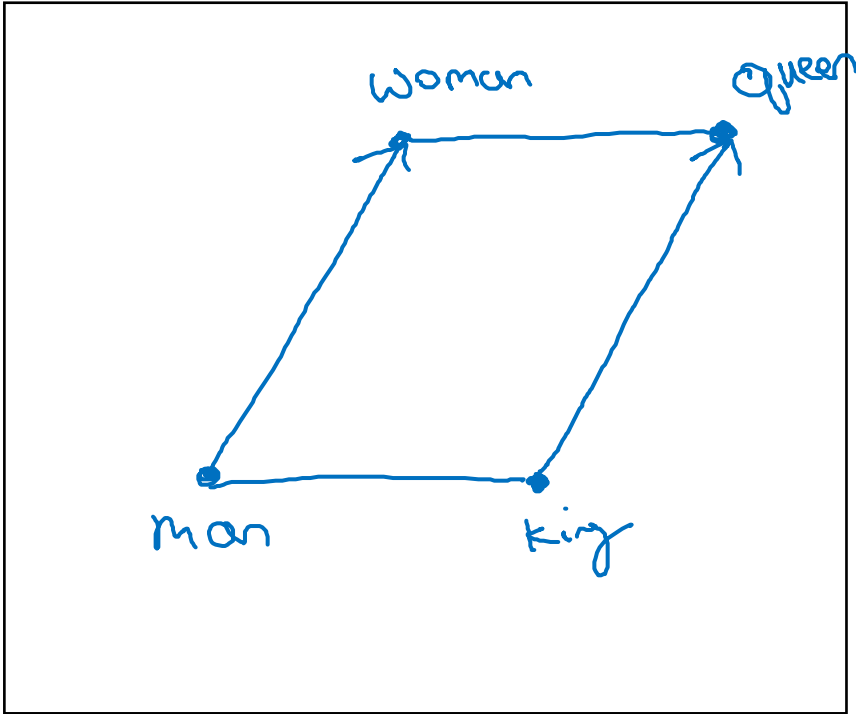
$3000 \rightarrow 2D$
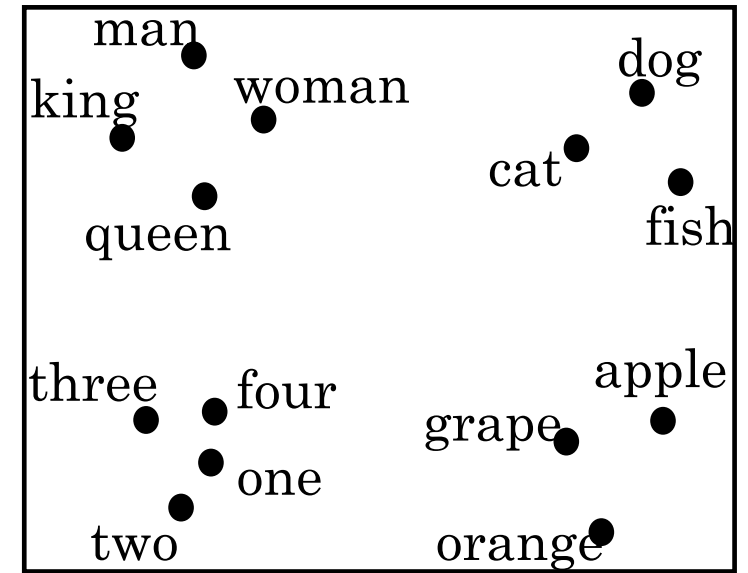
t-SNE

$$e_{man} - e_{woman} \approx e_{king} - e_{?} \quad e_w$$

$300D$

Find word w: $\arg\max\limits_{w} \; Sim(e_w, \; e_{king} - e_{man} + e_{woman})$

$30 - 75\%$

Andrew Ng

Cosine similarity

$$sim(e_w, e_{king} - e_{man} + e_{woman})$$

$$Sim(u,v) = \frac{u^T v}{\|u\|_2 \|v\|_2}$$
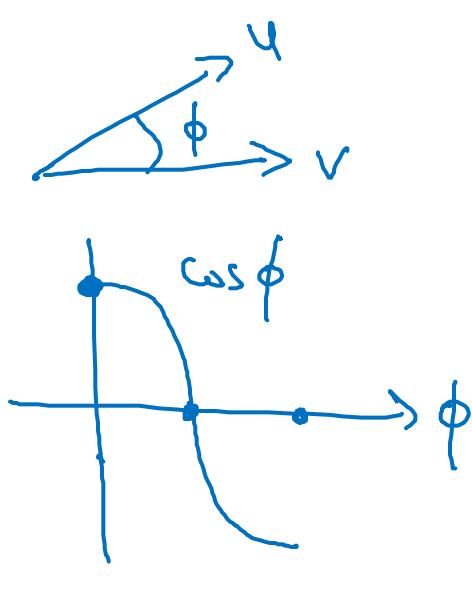
$\|u - v\|^2$

$\cos\phi$

Man:Woman as Boy:Girl

Ottawa:Canada as Nairobi:Kenya

Big:Bigger as Tall:Taller

Yen:Japan as Ruble:Russia

Andrew Ng

NLP and Word Embeddings

Embedding matrix

deeplearning.ai

Embedding matrix

6257

a aaron $\cdots$ orange $\cdots$ zulu   \<UNK\>

300

E

10,000

O_{6257}

$\leftarrow$ 6257

10,000

$E \cdot O_{6257} \quad = \quad = e_{6257}$

(300, 10K)   (10K, 1)

(300, 1)

$\rightarrow E \cdot O_j = e_j$

= embedding for word j
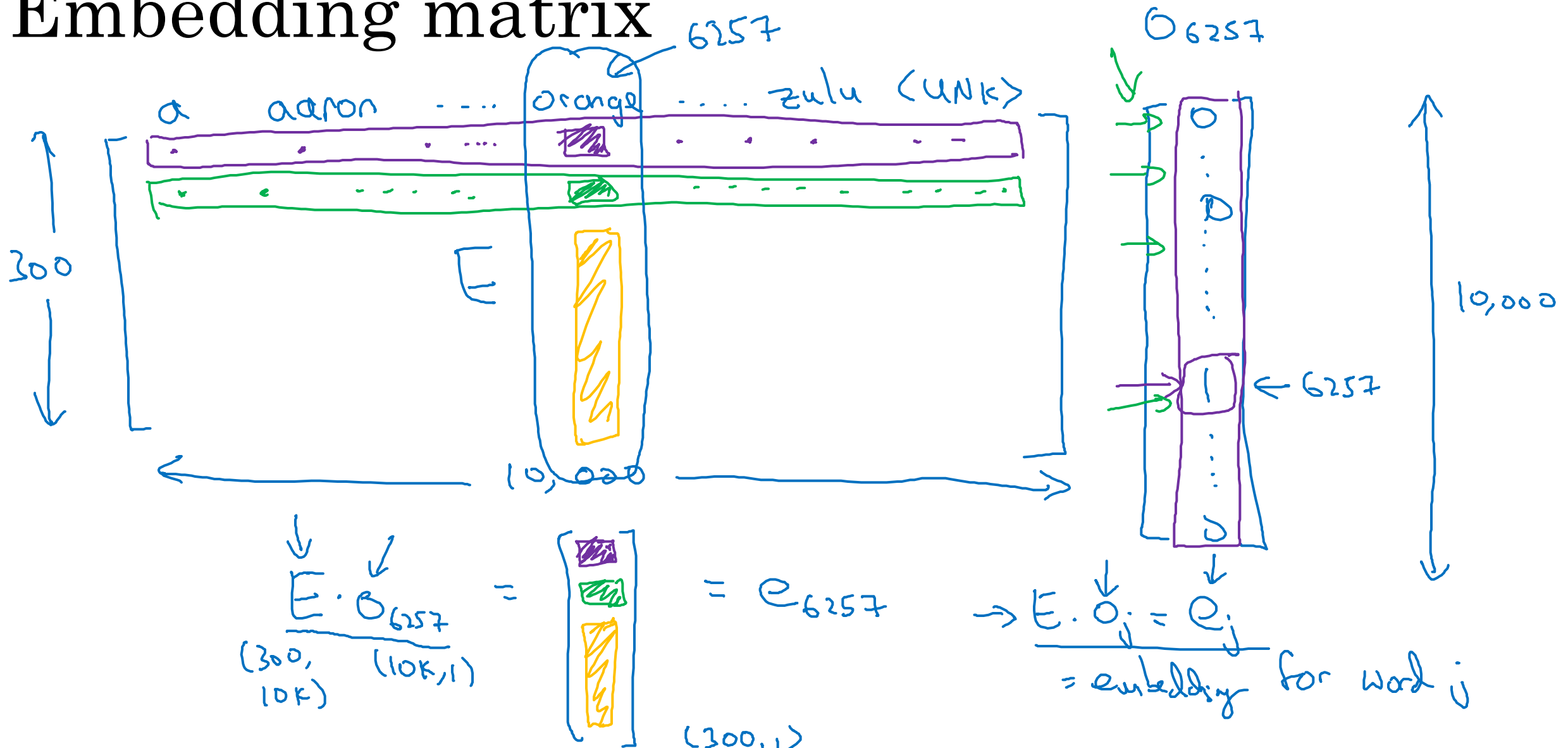
In practice, use specialized function to look up an embedding.

$\rightarrow$ Embedding

Andrew Ng

NLP and Word Embeddings

Learning word embeddings

deeplearning.ai

Neural language model

I   want   a   glass   of   orange   juice .

4343   9665   1   3852   6163   6257

apple juice.

$e_{4343} = E \, o_{4343}$

I	o_{4343}	E	e_{4343}
want	o_{9665}	E	e_{9665}
a	o_1	E	e_1
glass	o_{3852}	E	e_{3852}
of	o_{6163}	E	e_{6163}
orange	o_{6257}	E	e_{6257}

Softmax

10,000

$W^{[1]}, b^{[1]}$

$W^{[2]}, b^{[2]}$

1800 1200

[Bengio et. al., 2003, A neural probabilistic language model]

Andrew Ng
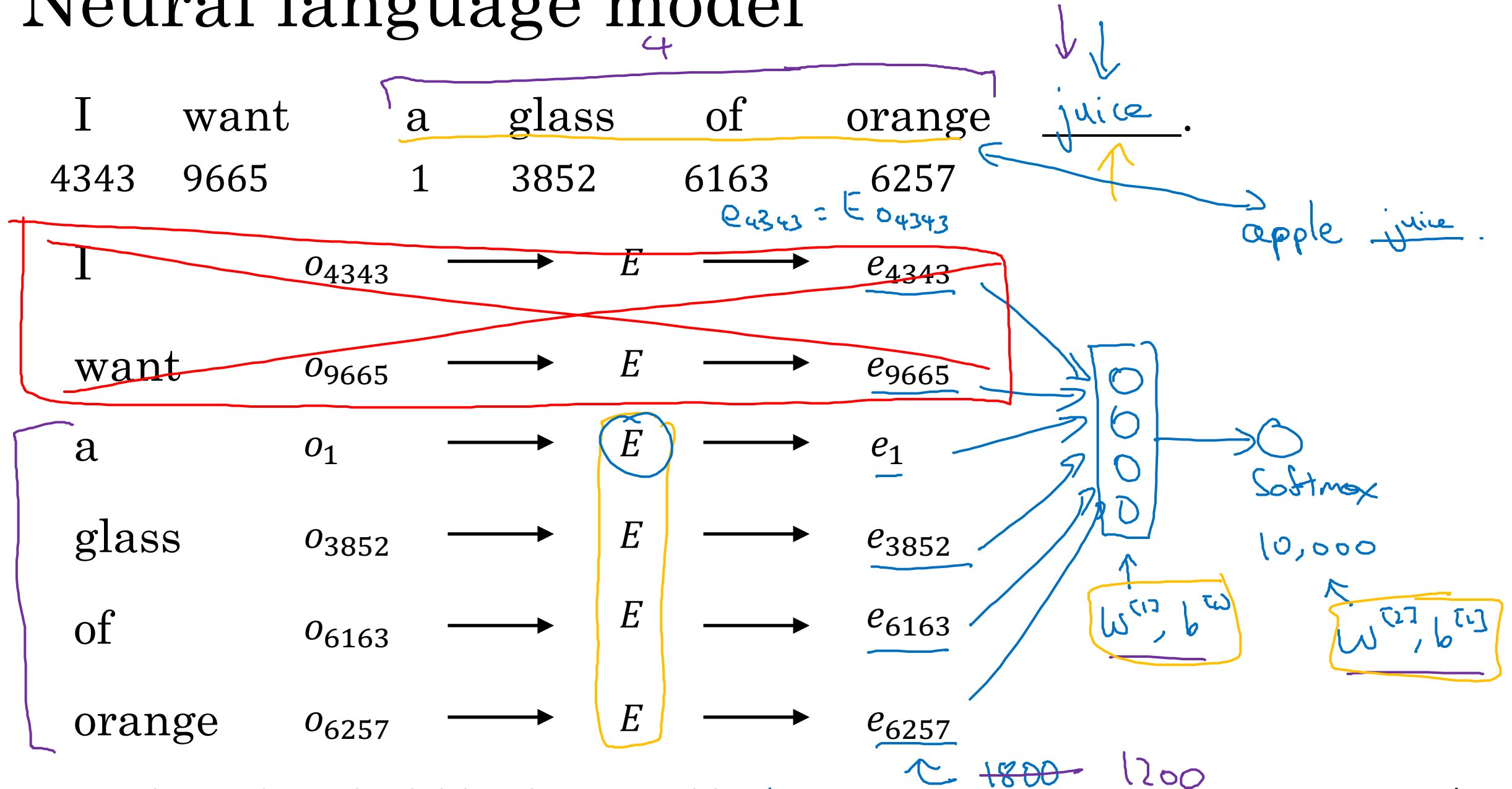
Other context/target pairs

I want a [glass] of [orange] juice to go along with my cereal.

context

target

Context: Last 4 words.

- 4 words on left & right
- Last 1 word
- Nearby 1 word

a glass of orange ___ ? to go alg with

orange ___ ?

glass ___ ?

skip gram

NLP and Word Embeddings

Word2Vec

deeplearning.ai

Skip-grams

I want a glass of orange juice to go along with my cereal.

Context	Target
orange	juice
orange	glass
orange	my

Model

Vocab size = 10,000k

$x \longrightarrow \underline{\underline{y}}$

Context c ("orange") $\longrightarrow$ Target t ("juice")

6257

4834

$O_c \rightarrow \boxed{E} \rightarrow e_c \rightarrow \text{Softmax} \rightarrow \hat{y}$

$e_c = E O_c$
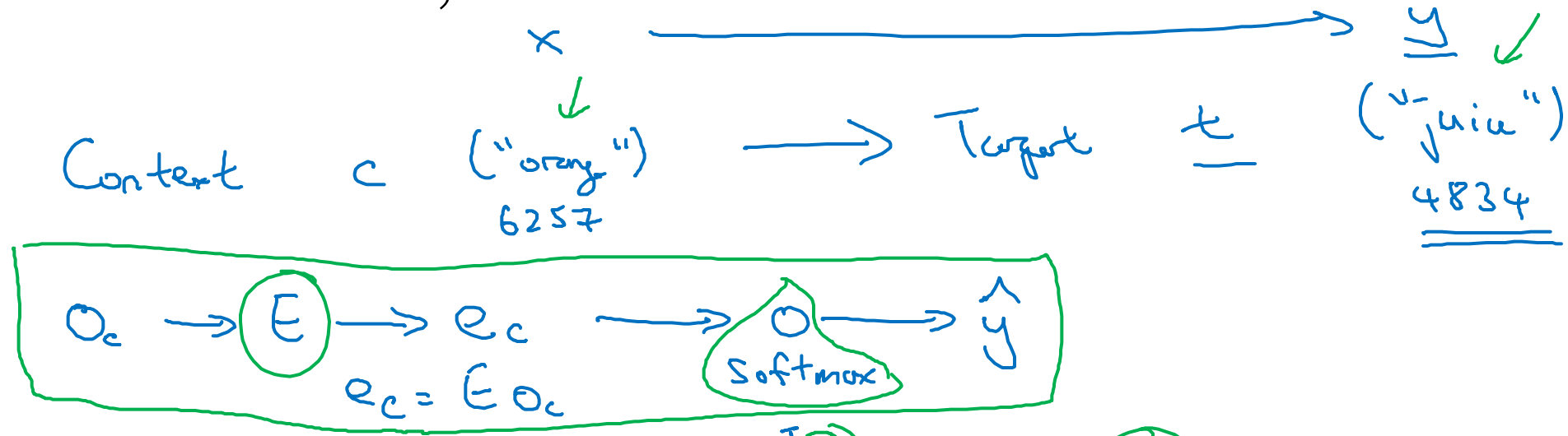
Softmax: $p(t|c) = \dfrac{e^{\theta_t^T e_c}}{\sum\limits_{j=1}^{10,000} e^{\theta_j^T e_c}}$

θ_t = parameter associated with output t

$\mathcal{L}(\hat{y}, y) = -\sum\limits_{i=1}^{10,000} y_i \log \hat{y}_i$

$y = \begin{bmatrix} 0 \\ \vdots \\ i \\ \vdots \\ 0 \end{bmatrix} \leftarrow 4834$

Andrew Ng

Problems with softmax classification

$$p(t|c) = \frac{e^{\theta_t^T e_c}}{\sum_{j=1}^{10,000} e^{\theta_j^T e_c}}$$

Hierachial   softmax.

$\log |V|$

5K   5K

2500   2500
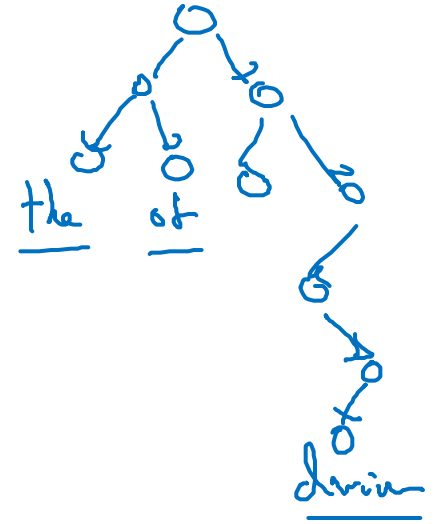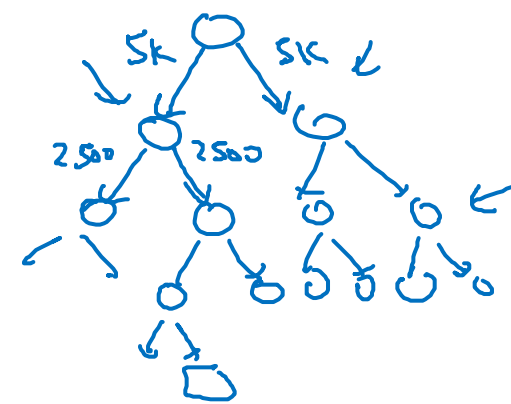
the   of

durian

How to sample the context c?

→ the, of, a, and, to, ...

→ orange, apple, durian

durian

t

$c \rightarrow t$

$P(c)$

NLP and Word Embeddings

Negative sampling

deeplearning.ai

Defining a new learning problem

I want a glass of orange juice to go along with my cereal.

Context	word	target?
orange	juice	1
orange	king	
orange	book	
orange	the	
orange	of	

$x \longrightarrow y$

$k = 5-20$ smaller. datasets
$k = 2-5$ larger dataset

[Mikolov et. al., 2013. Distributed representation of words and phrases and their compositionality]

Andrew Ng
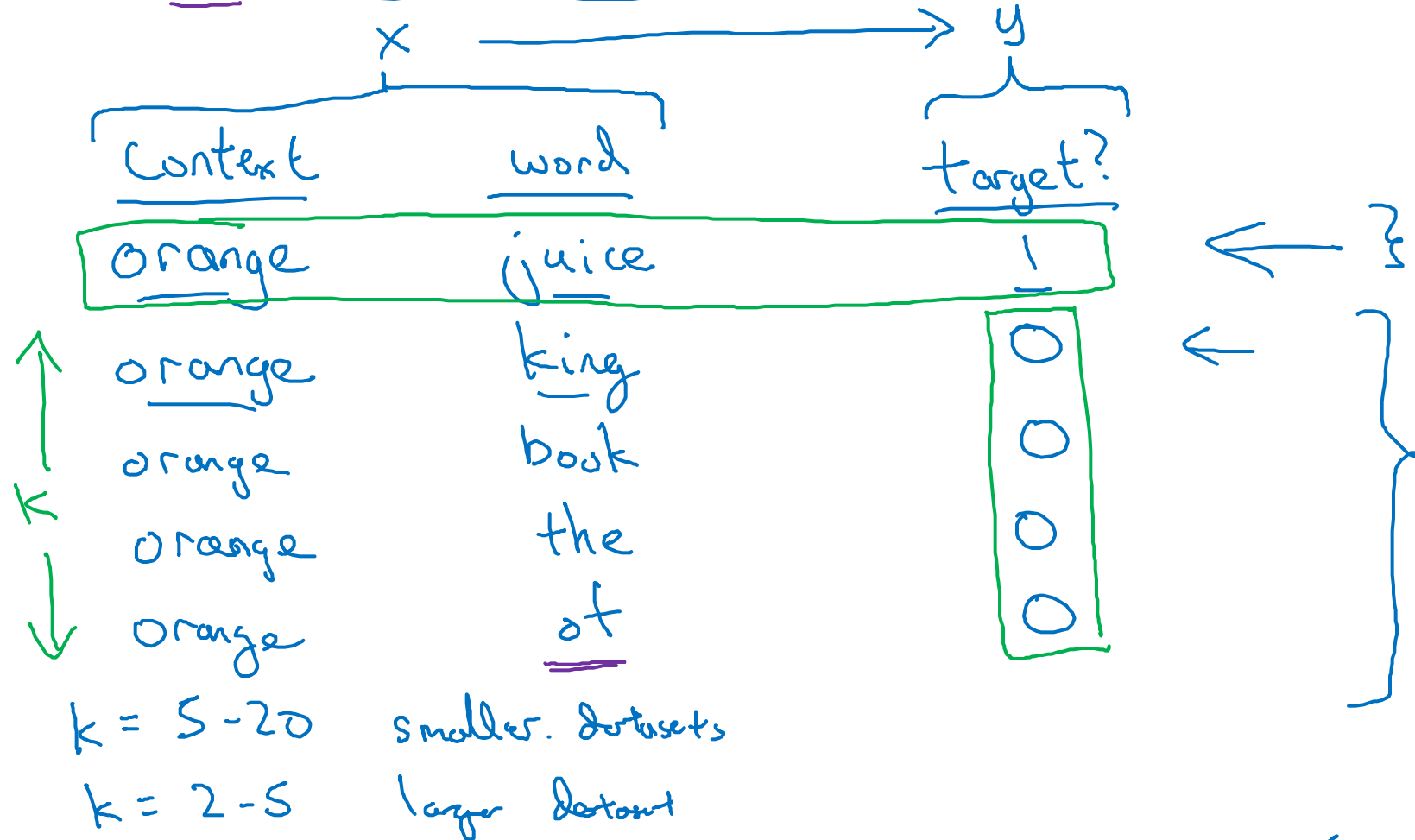
Model

Softmax: $$p(t|c) = \frac{e^{\theta_t^T e_c}}{\sum_{j=1}^{10,000} e^{\theta_j^T e_c}}$$

10,000-way softmax

x y

context	word	target?
orange	juice	1
orange	king	0
orange	book	0
orange	the	0
orange	of	0

c t y

$$P(y=1 \mid c, t) = \sigma(\theta_t^T e_c)$$

Orange
6257

$o_{6257} \rightarrow E \rightarrow e_{6257}$ juice?
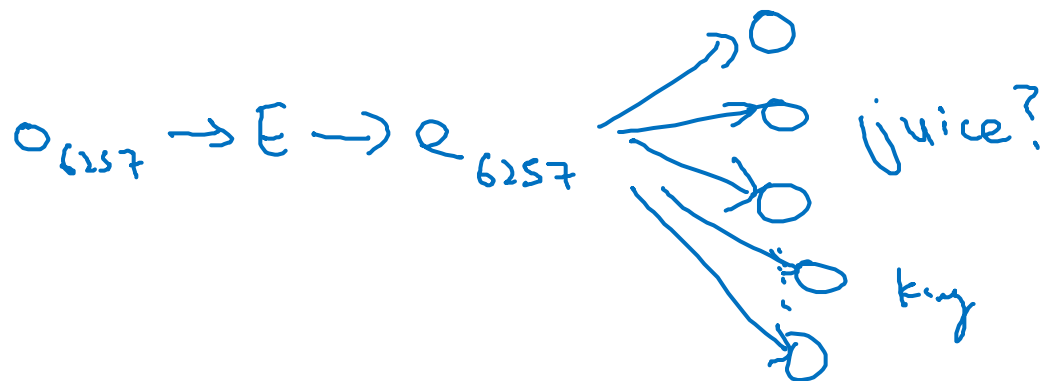
king

10,000

10,000 binary classification problem

$k+1$

Andrew Ng

Selecting negative examples

$\overbrace{}^{x}$ $\overbrace{}^{y}$

context	word	target?
orange	juice	1
orange	king	0
orange	book	0
orange	the	0
orange	of	0

$\uparrow$
t

the , of, and, ...

$$P(w_i) = \frac{f(w_i)^{3/4}}{\sum\limits_{j=1}^{10,000} f(w_j)^{3/4}}$$

$$\frac{1}{|V|}$$
$\uparrow$

NLP and Word Embeddings

GloVe word vectors

deeplearning.ai

GloVe (global vectors for word representation)

I want a glass of orange juice to go along with my cereal.

c, t

X_{ij} = # times i appears in context of j.

$X_{ij} = X_{ji}$ ←

[Pennington et. al., 2014. GloVe: Global vectors for word representation]   Andrew Ng

Model

$$\text{minimize} \quad \sum_{i=1}^{10,000} \sum_{j=1}^{10,000} f(X_{ij}) \left(\Theta_i^T e_j + b_i + b_j' - \log X_{ij} \right)^2 \quad \overset{>0?}{\longleftarrow}$$

$$\overset{t \quad c}{\text{``} \Theta_t^T e_c \text{''}}$$

weighting term

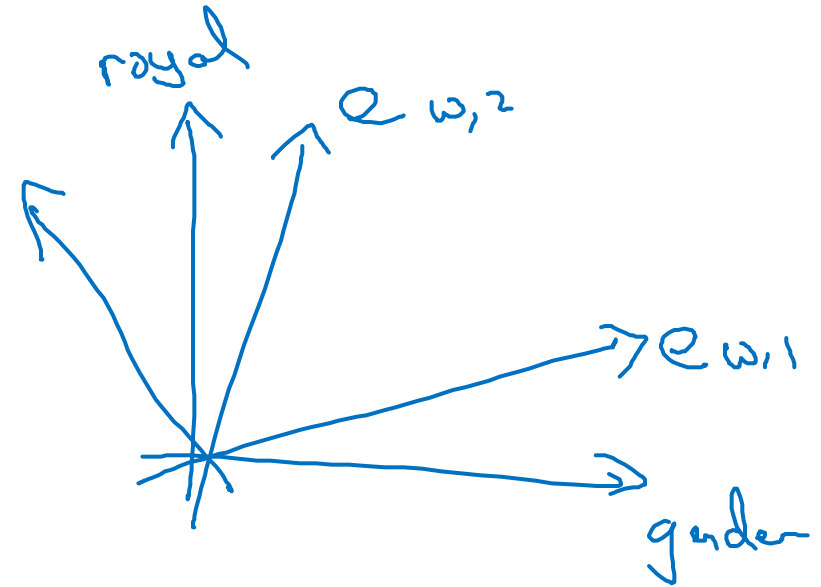$f(X_{ij}) = 0$ at $X_{ij} = 0$.   "$0 \log 0$" $= 0$

$\to$ this, is, of, a, ...

$\to$ durian

Θ_i, e_j are symmetric

$$e_w^{(final)} = \frac{e_w + \Theta_w}{2}$$

A note on the featurization view of word embeddings

	Man (5391)	Woman (9853)	King (4914)	Queen (7157)
Gender	-1	1	-0.95	0.97
Royal	0.01	0.02	0.93	0.95
Age	0.03	0.02	0.70	0.69
Food	0.09	0.01	0.02	0.01

$$\text{minimize} \sum_{i=1}^{10,000} \sum_{j=1}^{10,000} f(X_{ij}) \left(\overbrace{\theta_i^T e_j} + b_i - b_j' - \log X_{ij} \right)^2$$

$$(A\Theta_i)^T (A^{-T} e_j) = \theta_i^T A^T A^{-T} e_j$$

Andrew Ng

NLP and Word Embeddings

deeplearning.ai

Sentiment classification

Sentiment classification problem

x $\longrightarrow$ y

The dessert is excellent. ★★★★☆

Service was quite slow. ★★☆☆☆

Good for a quick meal, but nothing special. ★★★☆☆

Completely lacking in good taste, good service, and good ambience. ★☆☆☆☆
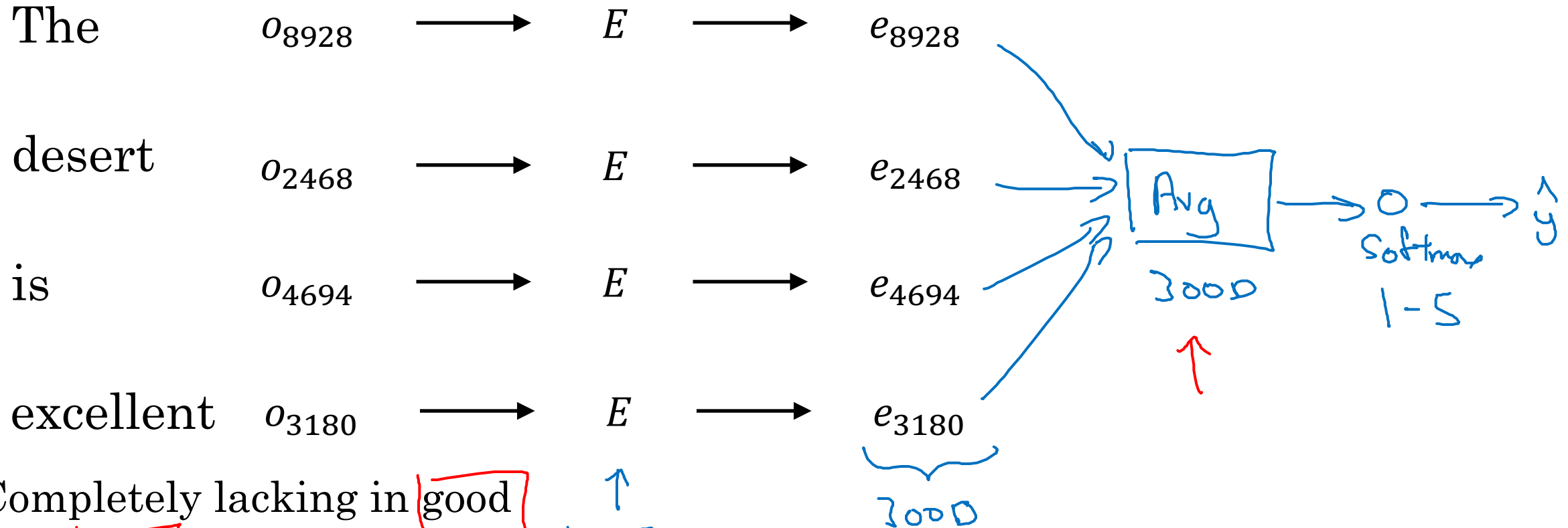
10,000 → 100,000 words

Andrew Ng

Simple sentiment classification model

The   dessert   is   excellent   ★★★★☆

8928   2468   4694   3180

The o_{8928} $\longrightarrow$ E $\longrightarrow$ e_{8928}

desert o_{2468} $\longrightarrow$ E $\longrightarrow$ e_{2468}

is o_{4694} $\longrightarrow$ E $\longrightarrow$ e_{4694}

excellent o_{3180} $\longrightarrow$ E $\longrightarrow$ e_{3180}

Avg → O → $\hat{y}$

Softmax

300D   1-5

300D

"Completely lacking in good taste, good service, and good ambience."

100 B words

Andrew Ng

RNN for sentiment classification

$\hat{y}$

softmax

$a^{<0>} \rightarrow \boxed{a^{<1>}} \rightarrow \boxed{a^{<2>}} \rightarrow \boxed{a^{<3>}} \rightarrow \boxed{a^{<4>}} \rightarrow \cdots \rightarrow \boxed{a^{<10>}}$

e_{1852} e_{4966} e_{4427} e_{3882} e_{330}

E E E E E

Completely   ~~lacking~~   ~~in~~   good   ....   ambience

absent

of

many-to-one

"not good"
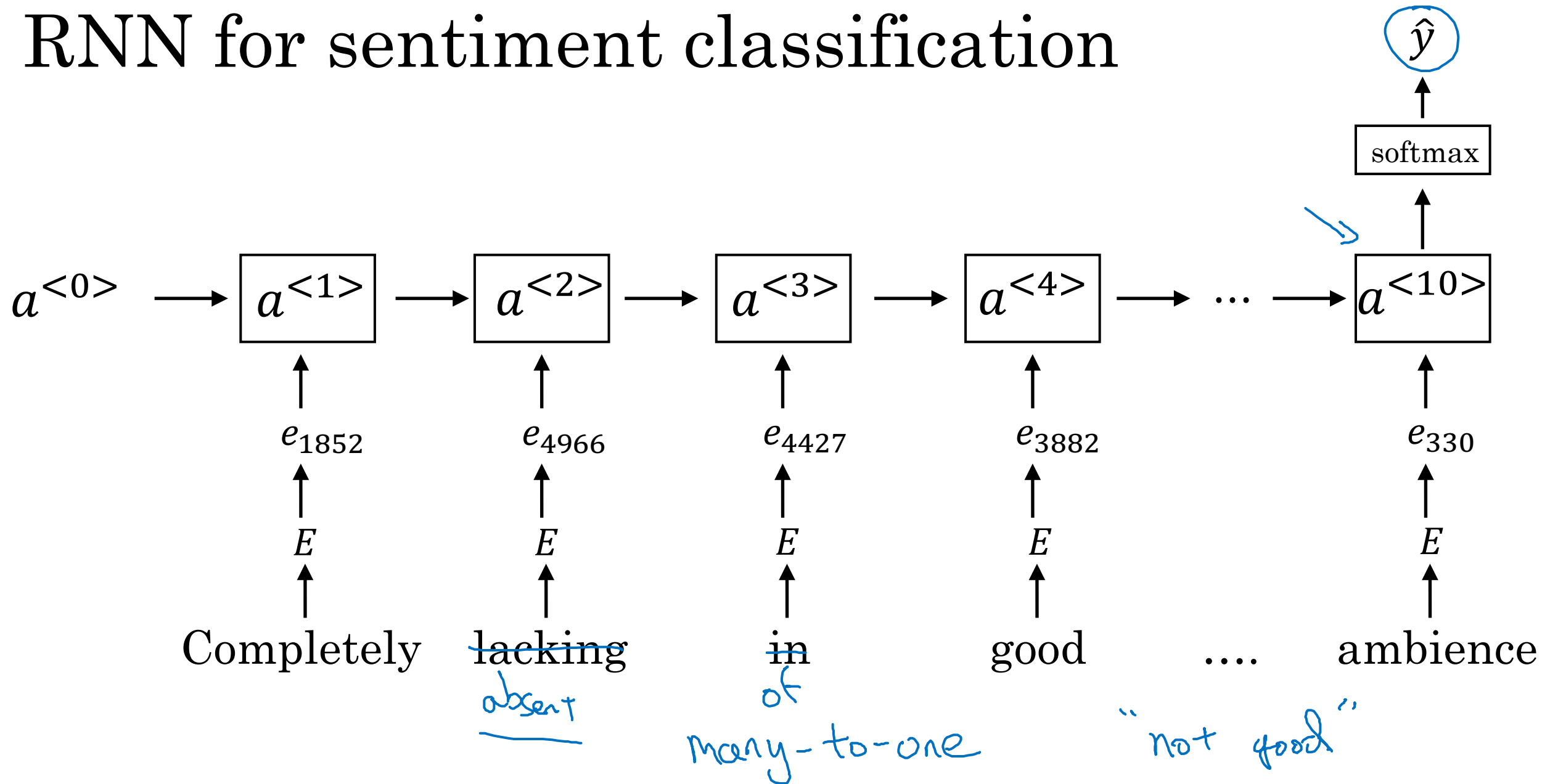
NLP and Word
Embeddings

Debiasing word
embeddings

deeplearning.ai

The problem of bias in word embeddings
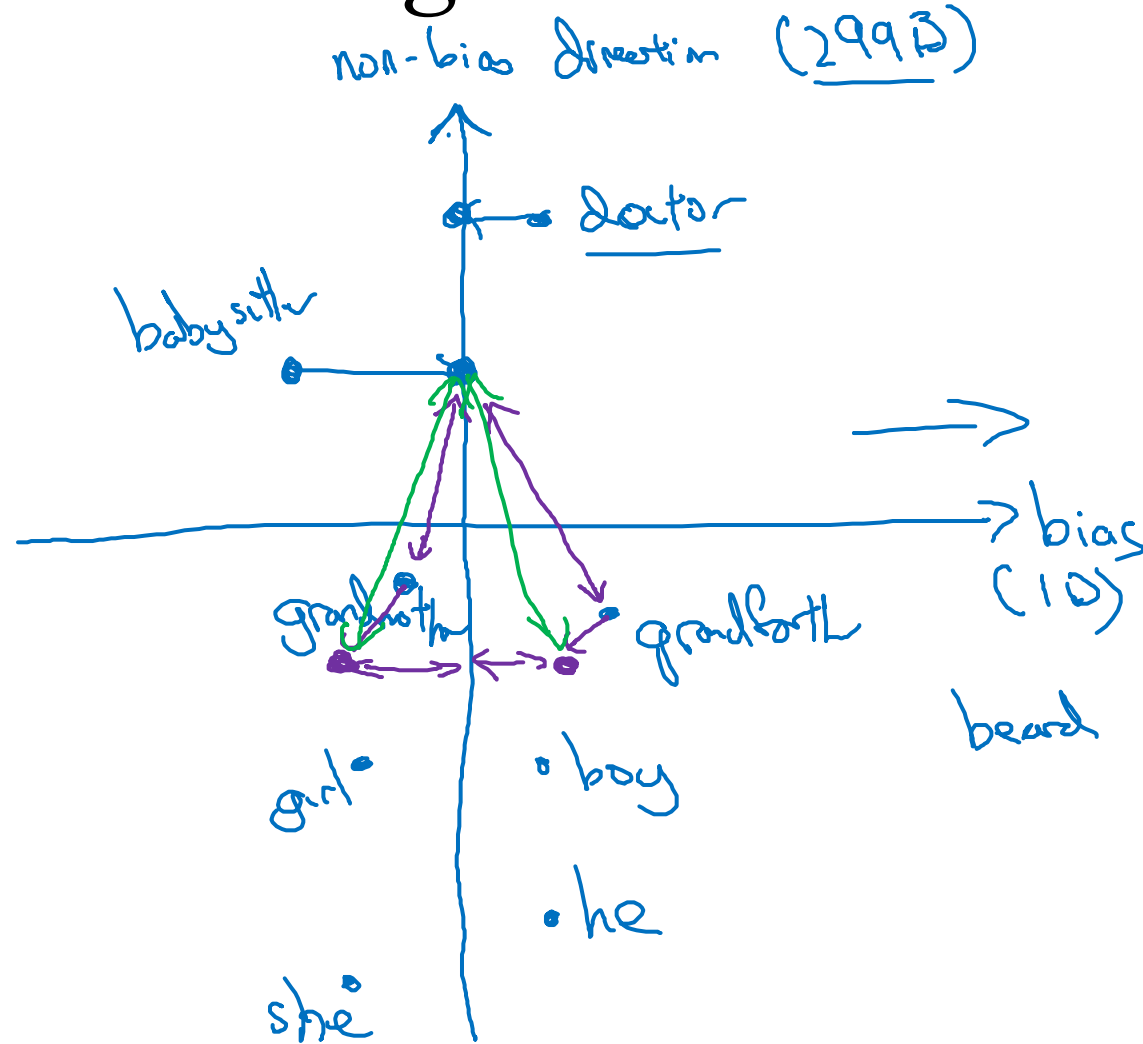
Man:Woman as King:Queen

Man:Computer_Programmer as Woman:Homemaker ✗

Father:Doctor as Mother:Nurse ✗

Word embeddings can reflect gender, ethnicity, age, sexual orientation, and other biases of the text used to train the model.

[Bolukbasi et. al., 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings]

Andrew Ng

Addressing bias in word embeddings

non-bias Direction (299D)

doctor

babysitter

grandmother   grandfather

girl   boy

he

she

bias (1D)

beard

1. Identify bias direction.

$$e_{he} - e_{she}$$
$$e_{male} - e_{female}$$
$$\vdots$$
$$\rightarrow average$$

2. Neutralize: For every word that is not definitional, project to get rid of bias.

3. Equalize pairs.

$$\rightarrow \text{grandmother} - \text{grandfather}$$
$$\text{girl} \qquad \text{boy}$$

[Bolukbasi et. al., 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings]
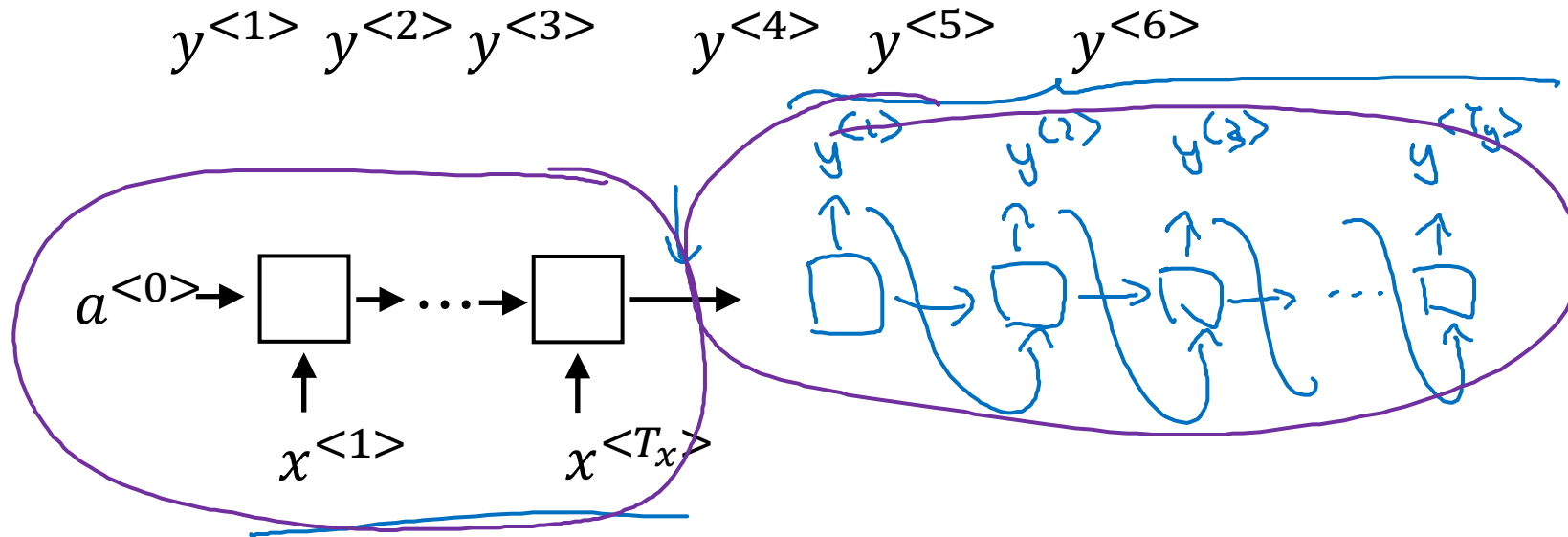
Andrew Ng

Sequence to
sequence models

Basic models

deeplearning.ai

Sequence to sequence model

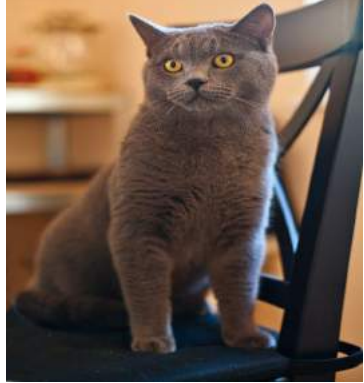$x^{<1>} \quad x^{<2>} \qquad x^{<3>} \qquad x^{<4>} \quad x^{<5>}$

Jane visite l'Afrique en septembre

$\longrightarrow$ Jane is visiting Africa in September.

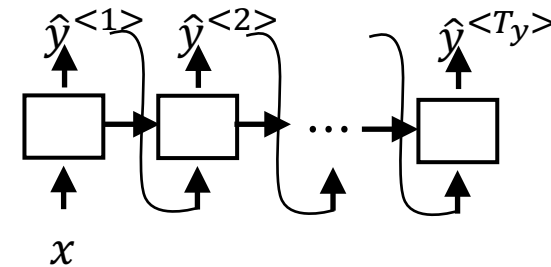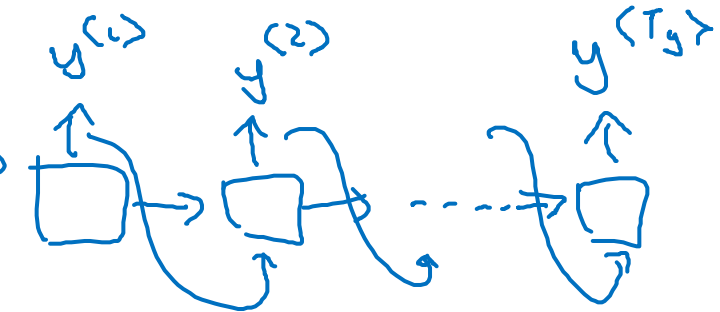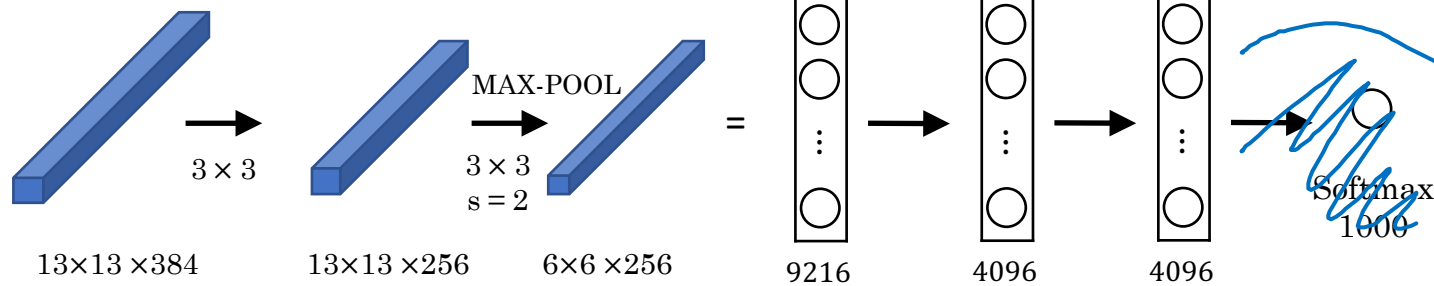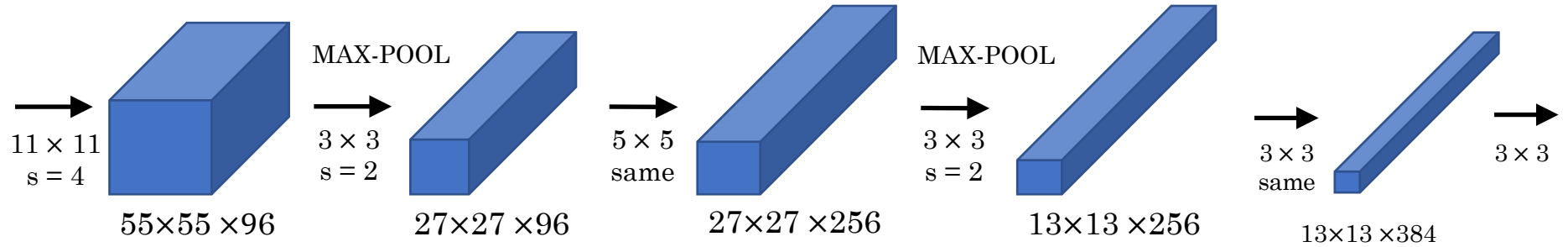$y^{<1>} \; y^{<2>} \, y^{<3>} \qquad y^{<4>} \quad y^{<5>} \qquad y^{<6>}$

[Sutskever et al., 2014. Sequence to sequence learning with neural networks]

[Cho et al., 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation]

Andrew Ng

Image captioning

$y^{<1>}\ y^{<2>}\quad y^{<3>}\quad y^{<4>}\quad y^{<5>}\quad y^{<6>}$

A   cat   sitting   on   a   chair

MAX-POOL

11×11
$s = 4$

3×3
$s = 2$

5×5
same

3×3
$s = 2$

3×3
same

3×3

$55 \times 55 \times 96$ $27 \times 27 \times 96$ $27 \times 27 \times 256$ $13 \times 13 \times 256$ $13 \times 13 \times 384$

3×3

MAX-POOL
3×3
$s = 2$

$=$

Softmax
1000

$y^{<1>}$ $y^{<2>}$ $y^{<T_y>}$

$13 \times 13 \times 384$ $13 \times 13 \times 256$ $6 \times 6 \times 256$ 9216 4096 4096

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

$\cdots$

x

[Mao et. al., 2014. Deep captioning with multimodal recurrent neural networks]
[Vinyals et. al., 2014. Show and tell: Neural image caption generator]
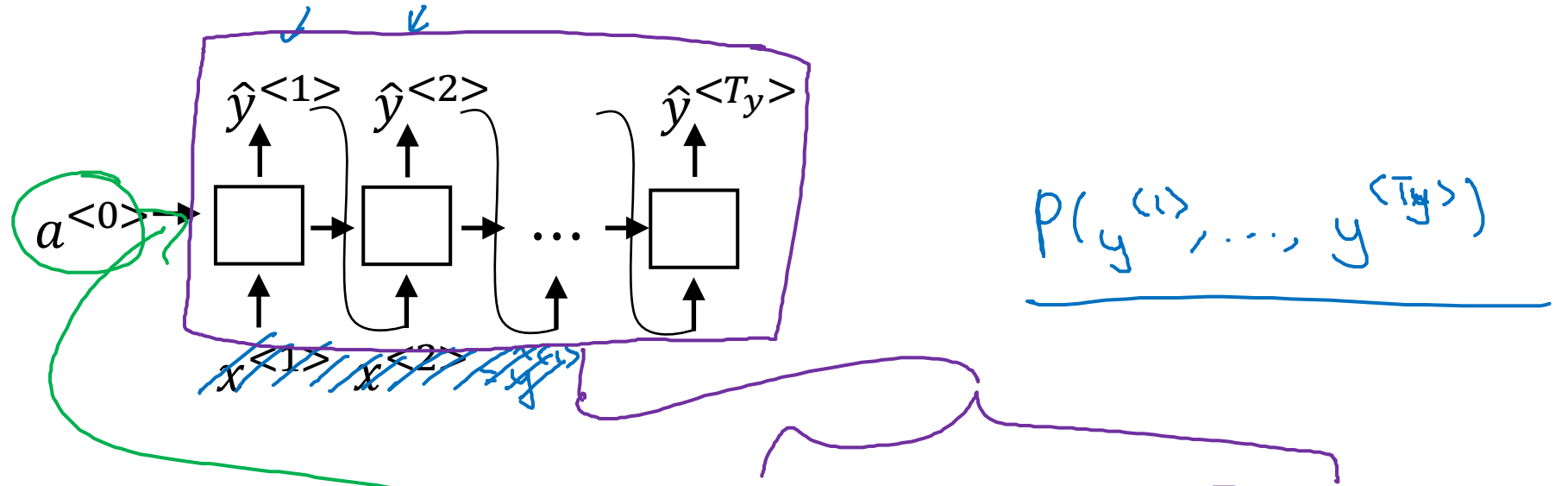[Karpathy and Li, 2015. Deep visual-semantic alignments for generating image descriptions]
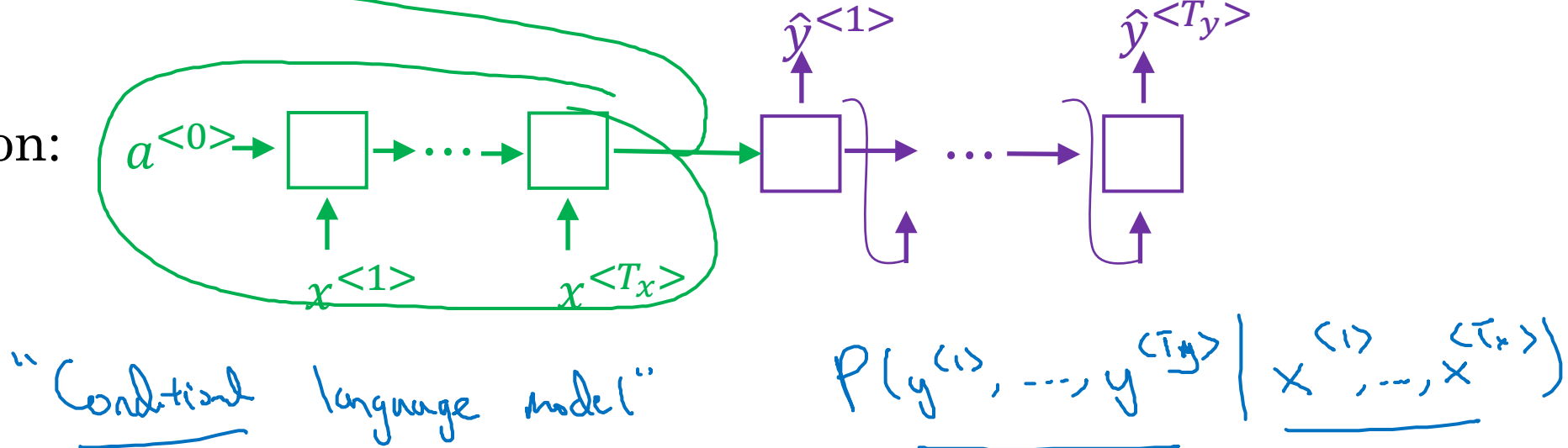
Andrew Ng

deeplearning.ai

Sequence to sequence models

Picking the most likely sentence

Machine translation as building a conditional language model

Language model:

Machine translation:

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

"Conditional language model"

$$P(y^{<1>}, \ldots, y^{<T_y>} \mid x^{<1>}, \ldots, x^{<T_x>})$$
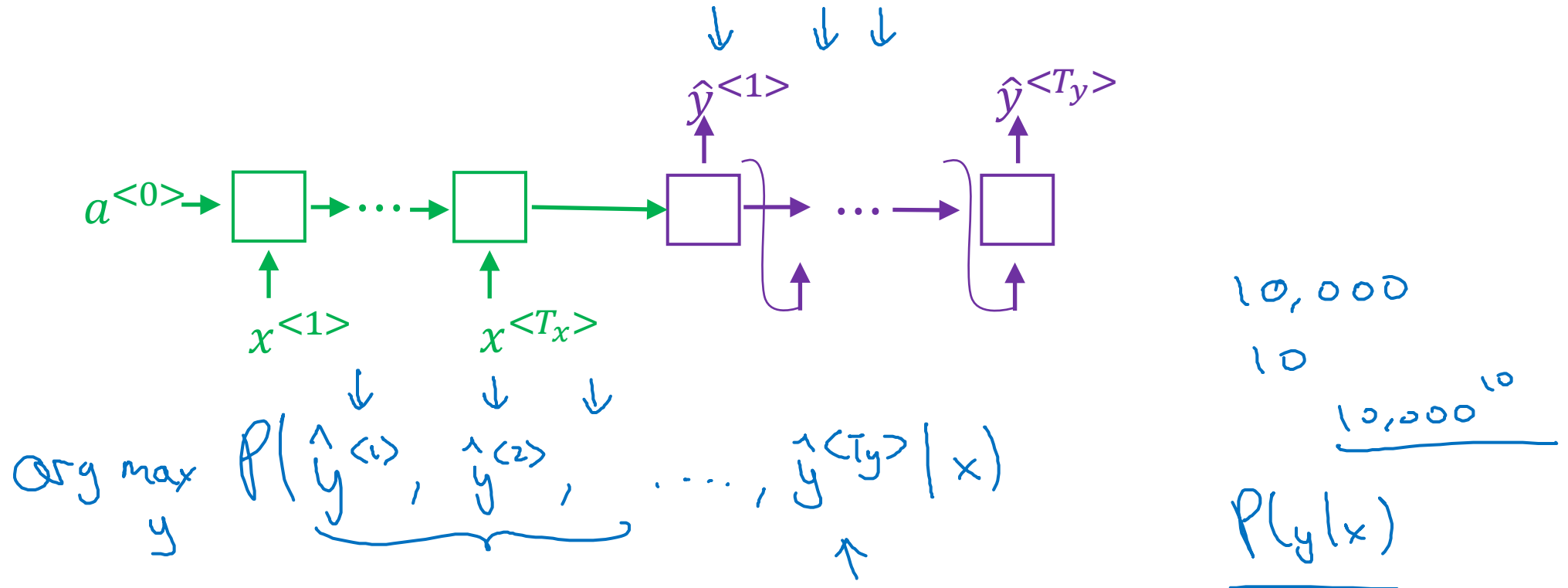
Finding the most likely translation

Jane visite l'Afrique en septembre.

English French

$$P(y^{<1>}, \dots, y^{<T_y>} | x)$$

→ Jane is visiting Africa in September.

→ Jane is going to be visiting Africa in September.

→ In September, Jane will visit Africa.

→ Her African friend welcomed Jane in September.

$$\arg\max_{y^{<1>}, \dots, y^{<T_y>}} P(y^{<1>}, \dots, y^{<T_y>} | x)$$

Why not a greedy search?

$P(\hat{y}^{<1>}|x)$

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

$a^{<0>}$

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

$\arg\max_y P(\hat{y}^{<1>}, \hat{y}^{<2>}, \ldots, \hat{y}^{<T_y>}|x)$

$10,000$

10

$\dfrac{10,000^{10}}{}$

$P(y|x)$

→ Jane is visiting Africa in September.

→ Jane is going to be visiting Africa in September.

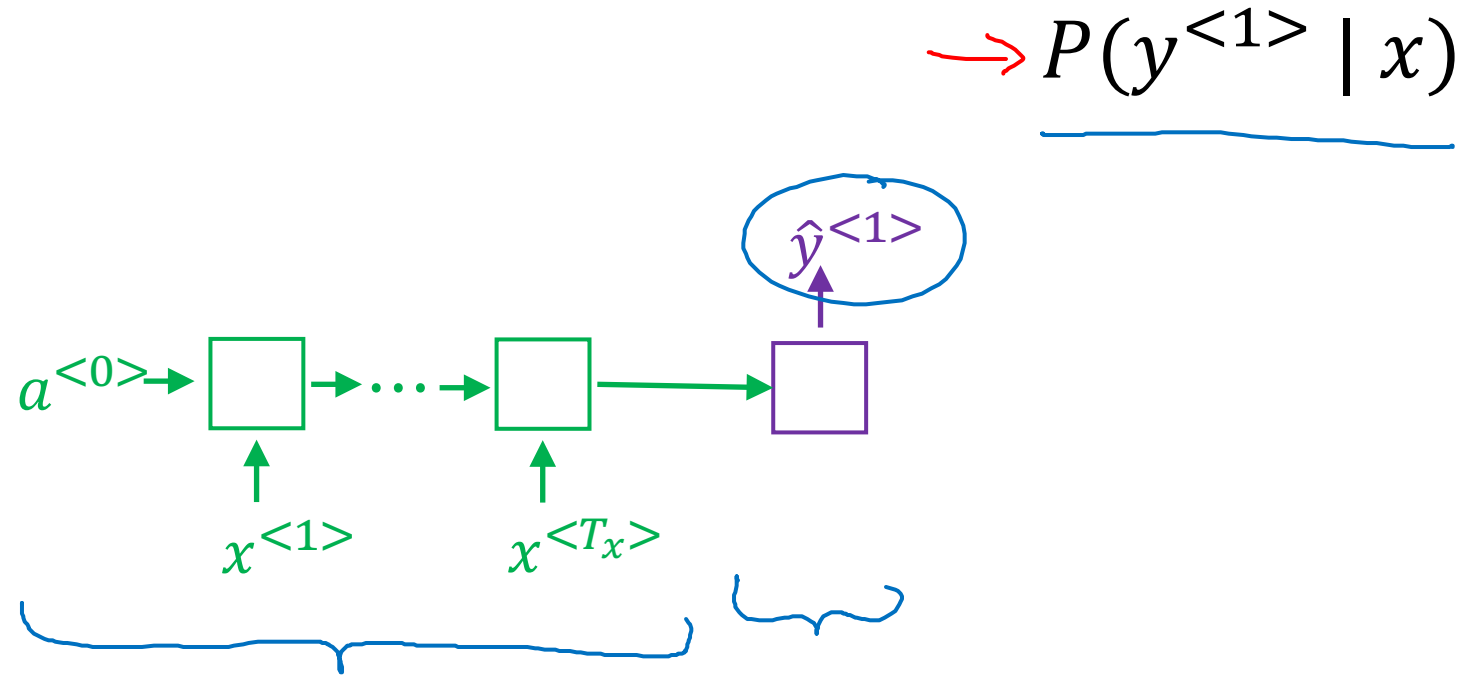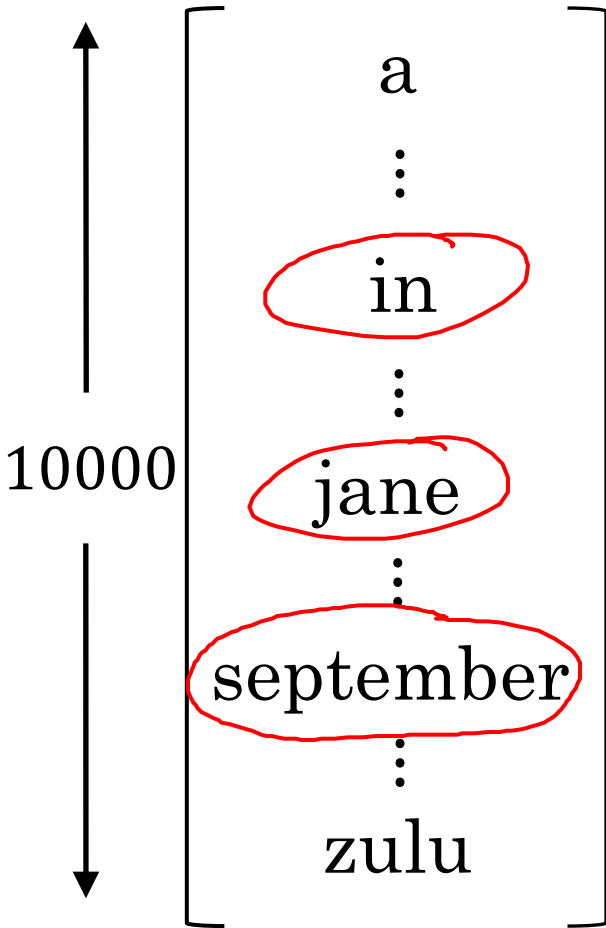$P(\text{Jane is going}|x) > P(\text{Jane is visit}|x)$

Andrew Ng

Sequence to sequence models

Beam search

deeplearning.ai

Beam search algorithm

$B = 3$ (beam width)

Step 1

$$\to P(y^{<1>} \mid x)$$

$$\begin{bmatrix} a \\ \vdots \\ in \\ \vdots \\ jane \\ \vdots \\ september \\ \vdots \\ zulu \end{bmatrix}$$ 10000

$\hat{y}^{<1>}$

$a^{<0>} \to \square \to \cdots \to \square \to \square$

$x^{<1>}$　　　$x^{<T_x>}$

Beam search algorithm

$(B = 3)$

Step 1   Step 2

$\begin{bmatrix} a \\ \vdots \\ in \\ \vdots \\ jane \\ \vdots \\ september \\ \vdots \\ zulu \end{bmatrix}$ 10000

a
aaron
September
visit
zulu

a
aaron
is
visit
zulu
10,000

a
zulu

$y^{<1>}, y^{<2>}$

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

in
$\hat{y}^{<1>}$ $\hat{y}^{<2>}$

$a^{<0>} \rightarrow \square \rightarrow \cdots \rightarrow \square \rightarrow \square \rightarrow \square$
$x^{<1>}$ $x^{<T_x>}$

$P(y^{<2>} | x, "in")$

in

jane $\hat{y}^{<2>}$

$a^{<0>} \rightarrow \square \rightarrow \cdots \rightarrow \square \rightarrow \square \rightarrow \square$
$x^{<1>}$ $x^{<T_x>}$

$P(y^{<2>} | x, "jane")$

September $\hat{y}^{<2>}$

$a^{<0>} \rightarrow \square \rightarrow \cdots \rightarrow \square \rightarrow \square \rightarrow \square$
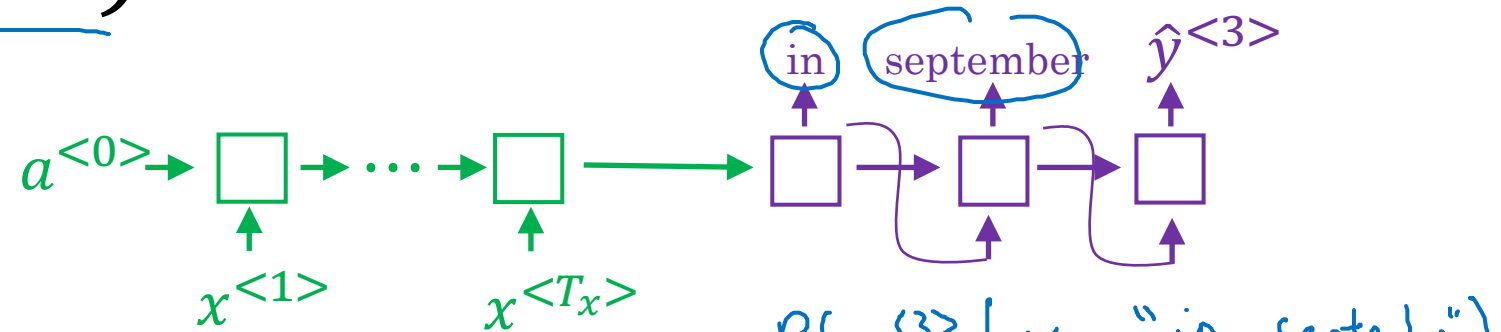$x^{<1>}$ $x^{<T_x>}$

Andrew Ng

Beam search ($B = 3$)

$B = 1 \rightsquigarrow$ greedy search

in september
- a
- aaron
- jane
- zulu

jane is
- a
- visits
- zulu
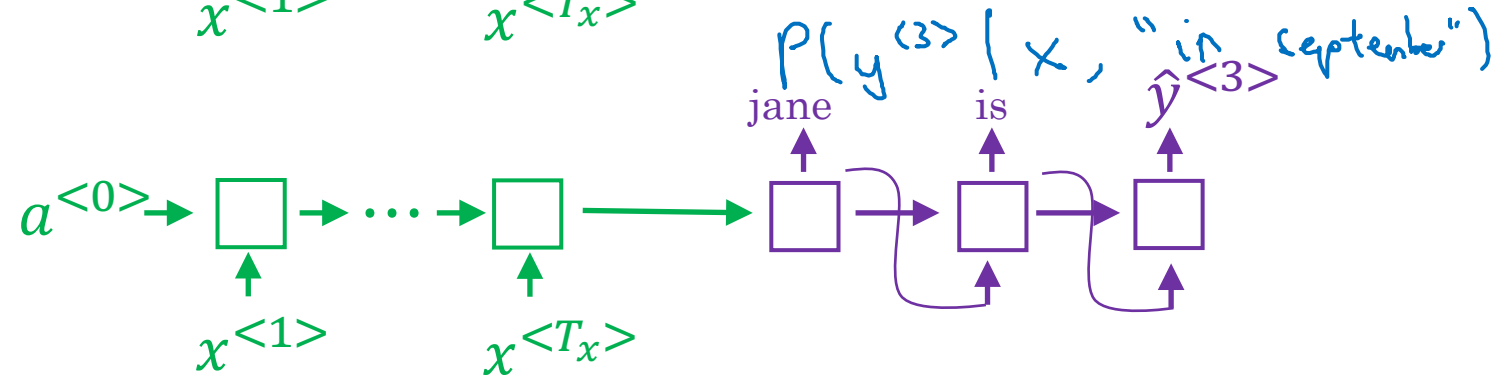
jane visits
- a
- africa
- zulu
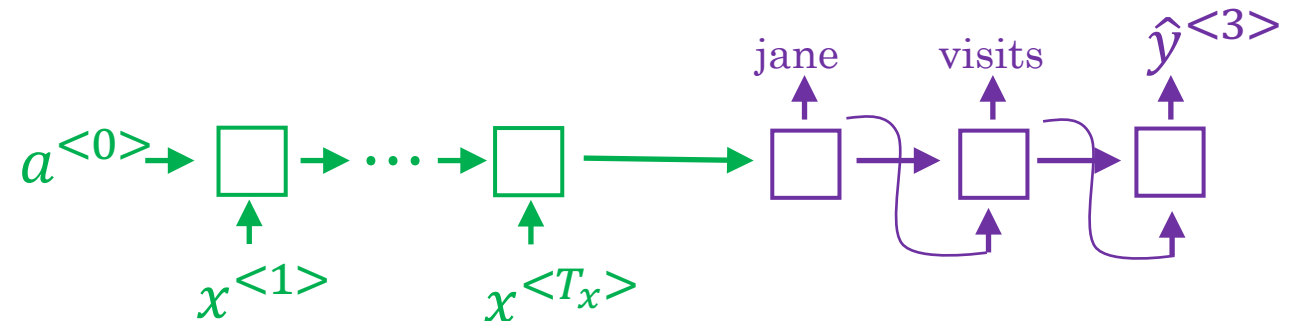
$P(\hat{y}^{<3>} | x, \text{"in september"})$

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

jane visits africa in september. <EOS>

Andrew Ng

deeplearning.ai

Sequence to sequence models

Refinements to beam search

Length normalization

$$P(y^{<1>}...y^{<T_y>}|x) = P(y^{<1>}|x)P(y^{<2>}|x, y^{<1>})...$$
$$P(y^{<T_y>}|x, y^{<1>}...,y^{<T_y-1>})$$

$$\arg\max_y \prod_{t=1}^{T_y} P(y^{<t>}|x, y^{<1>},...,y^{<t-1>})$$

$\log$

$\log P(y|x) \leftarrow$

$P(y|x) \leftarrow$

$$\arg\max_y \sum_{t=1}^{T_y} \log P(y^{<t>}|x, y^{<1>},...,y^{<t-1>}) \leftarrow$$

$$T_y = 1, 2, 3, ..., 30.$$

$$\frac{1}{T_y^\alpha} \sum_{t=1}^{T_y} \log P(y^{<t>}|x, y^{<1>},...,y^{<t-1>})$$
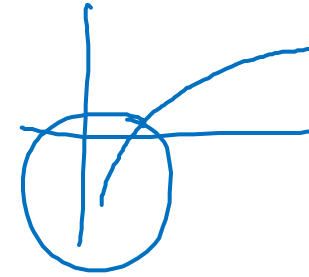
$\alpha = 0.7$

$\alpha = 1$

$\alpha = 0$

Beam search discussion

Beam width B?

$1 \to 3 \to 10, \quad 100, \quad 1000 \to 3000$

large B: better result, slower

small B: worse result, faster

Unlike exact search algorithms like BFS (Breadth First Search) or DFS (Depth First Search), Beam Search runs faster but is not guaranteed to find exact maximum for $\arg\max\limits_{y} P(y|x)$.

Sequence to sequence models

Error analysis on beam search

deeplearning.ai

Example

Jane visite l'Afrique en septembre.
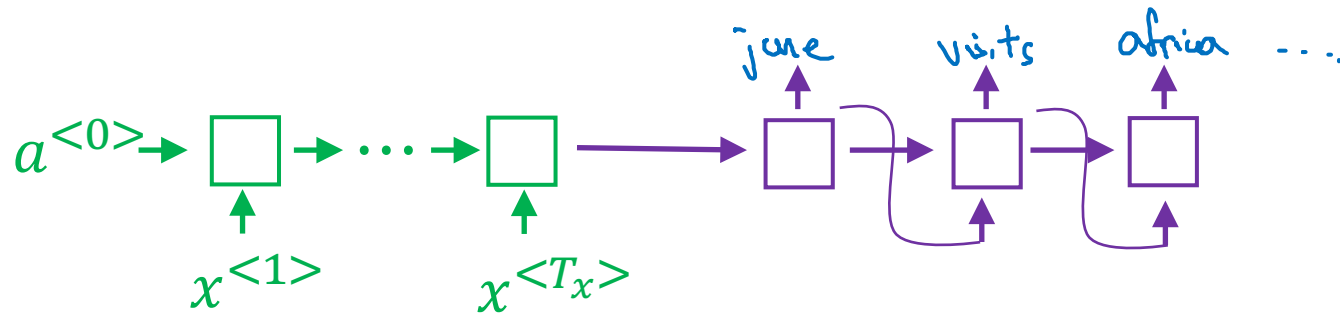
→ RNN

→ Beam Search   BT

Human: Jane visits Africa in September. (y^*)

Algorithm: Jane visited Africa last September. $(\hat{y})$ ←

RNN computes $P(y^*|x) \gtrless P(\hat{y}|x)$

Error analysis on beam search

$P(y^* | x)$

$P(\hat{y} | x)$

Human: Jane visits Africa in September. (y^*)

Algorithm: Jane visited Africa last September. $(\hat{y})$

Case 1: $P(y^* | x) > P(\hat{y} | x) \leftarrow$

$\underset{y}{\arg\max} \; P(y | x)$

Beam search chose $\hat{y}$. But y^* attains higher $\boxed{P(y | x)}$.

Conclusion: Beam search is at fault.

Case 2: $P(y^* | x) \lesssim P(\hat{y} | x) \leftarrow$

y^* is a better translation than $\hat{y}$. But RNN predicted $\boxed{P(y^* | x)} < P(\hat{y} | x)$.

Conclusion: RNN model is at fault.

Error analysis process

Human	Algorithm	$P(y^*\|x)$	$P(\hat{y}\|x)$	At fault?
Jane visits Africa in September.	Jane visited Africa last September.	2×10^{-10}	1×10^{-10}	B
				R
				B
				R
				R
				...

Figures out what faction of errors are "due to" beam search vs. RNN model

Andrew Ng

Sequence to sequence models

deeplearning.ai

Bleu score (optional)

Evaluating machine translation

French: Le chat est sur le tapis.

Reference 1: The cat is on the mat. ←

Reference 2: There is a cat on the mat. ←

MT output: the the the the the the the.

Precision: Modified precision:

Bleu
bilingual evaluation understudy

[Papineni et. al., 2002. Bleu: A method for automatic evaluation of machine translation]

Andrew Ng

Bleu score on bigrams

Example: Reference 1: The cat is on the mat. ←

Reference 2: There is a cat on the mat. ←

MT output: The cat the cat on the mat. ←

	Count	Count$_{clip}$
the cat	2 ←	1 ←
cat the	1 ←	0
cat on	1 ←	1 ←
on the	1 ←	1 ←
the mat	1 ←	1 ←

$$\frac{4}{6}$$

[Papineni et. al., 2002. Bleu: A method for automatic evaluation of machine translation]   Andrew Ng

Bleu score on unigrams

Example: Reference 1: The cat is on the mat.

Reference 2: There is a cat on the mat.

→ MT output: The cat the cat on the mat. $(\hat{y})$

$p_1, p_2 = \underline{1.0}$

$$p_1 = \frac{\displaystyle\sum_{unigram \in \hat{y}} count_{clip}(unigram)}{\displaystyle\sum_{unigram \in \hat{y}} count(unigram)}$$

$$p_n = \frac{\displaystyle\sum_{ngram \in \hat{y}} count_{clip}(ngram)}{\displaystyle\sum_{ngram \in \hat{y}} count(ngram)}$$

n-gram

$count_{clip}(n\text{-}gram)$

n-grams $\in \hat{y}$

$count(n\text{-}gram)$

n-grams $\in \hat{y}$

unigram $\in \hat{y}$

count(unigram)

unigram

[Papineni et. al., 2002. Bleu: A method for automatic evaluation of machine translation]   Andrew Ng

Bleu details

p_n = Bleu score on n-grams only

P_1, P_2, P_3, P_4

Combined Bleu score: $BP \exp\left(\frac{1}{4}\sum_{n=1}^{4} P_n\right)$

BP = brevity penalty

$$BP = \begin{cases} 1 & \text{if MT\_output\_length} > \text{reference\_output\_length} \\ \exp(1 - \text{MT\_output\_length}/\text{reference\_output\_length}) & \text{otherwise} \end{cases}$$

[Papineni et. al., 2002. Bleu: A method for automatic evaluation of machine translation]   Andrew Ng

deeplearning.ai

Sequence to
sequence models

Attention model
intuition

The problem of long sequences

$a^{<0>} \rightarrow \square \rightarrow \cdots \rightarrow \square \rightarrow \square \rightarrow \cdots \rightarrow \square$

$x^{<1>}$ $x^{<T_x>}$ $\hat{y}^{<1>}$ $\hat{y}^{<T_y>}$
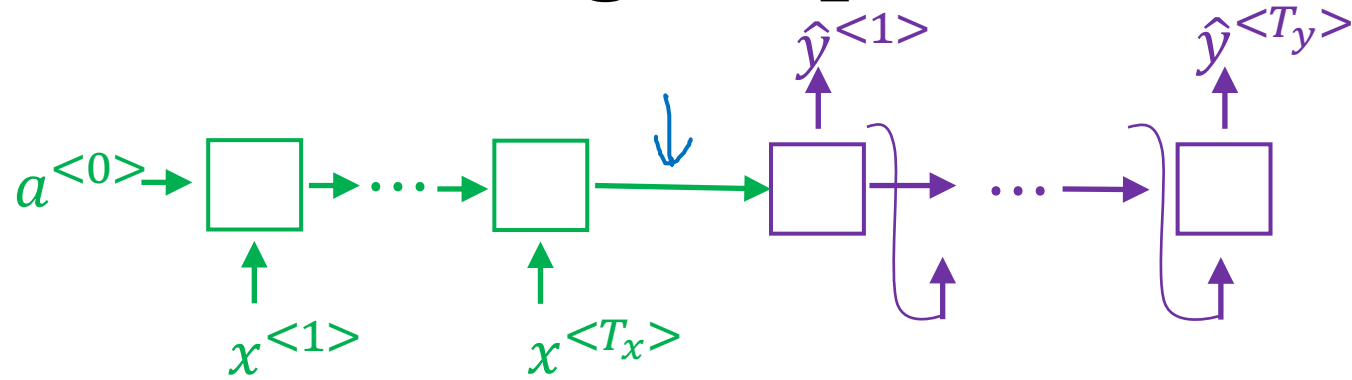
Jane s'est rendue en Afrique en septembre dernier, a apprécié la culture et a rencontré beaucoup de gens merveilleux; elle est revenue en parlant comment son voyage était merveilleux, et elle me tente d'y aller aussi.

Jane went to Africa last September, and enjoyed the culture and met many wonderful people; she came back raving about how wonderful her trip was, and is tempting me to go too.
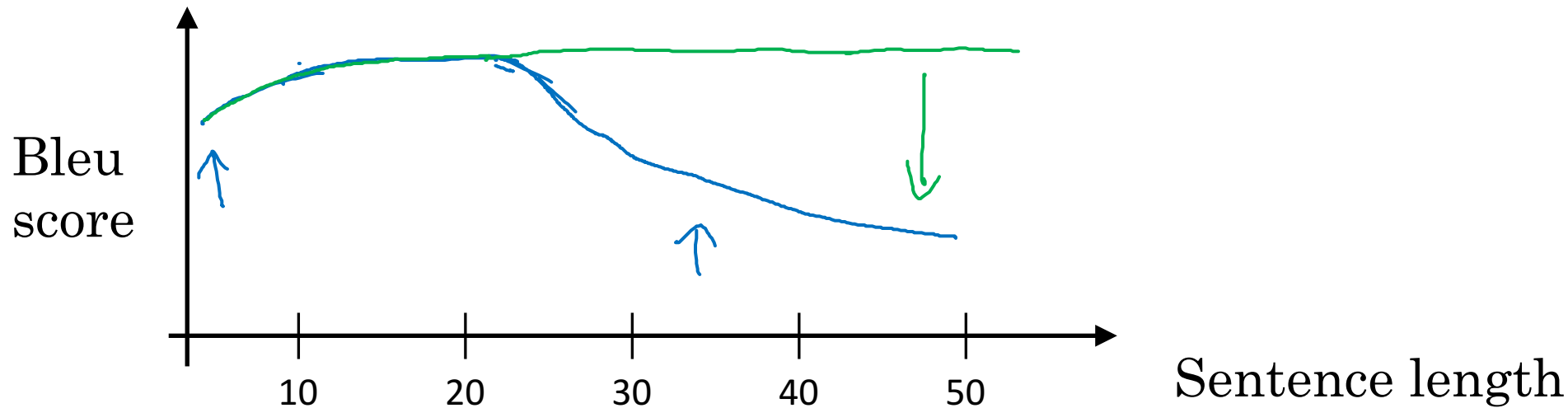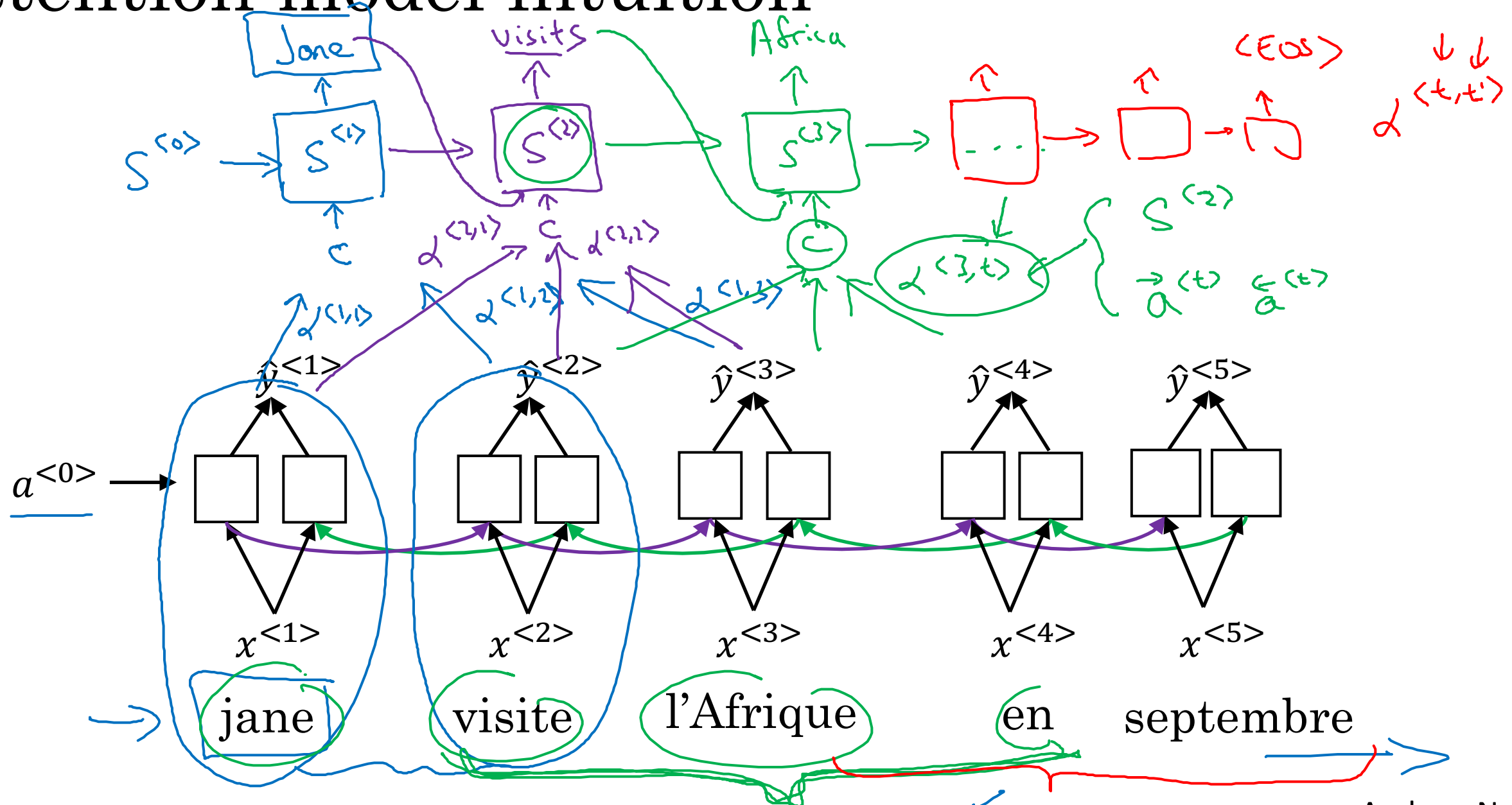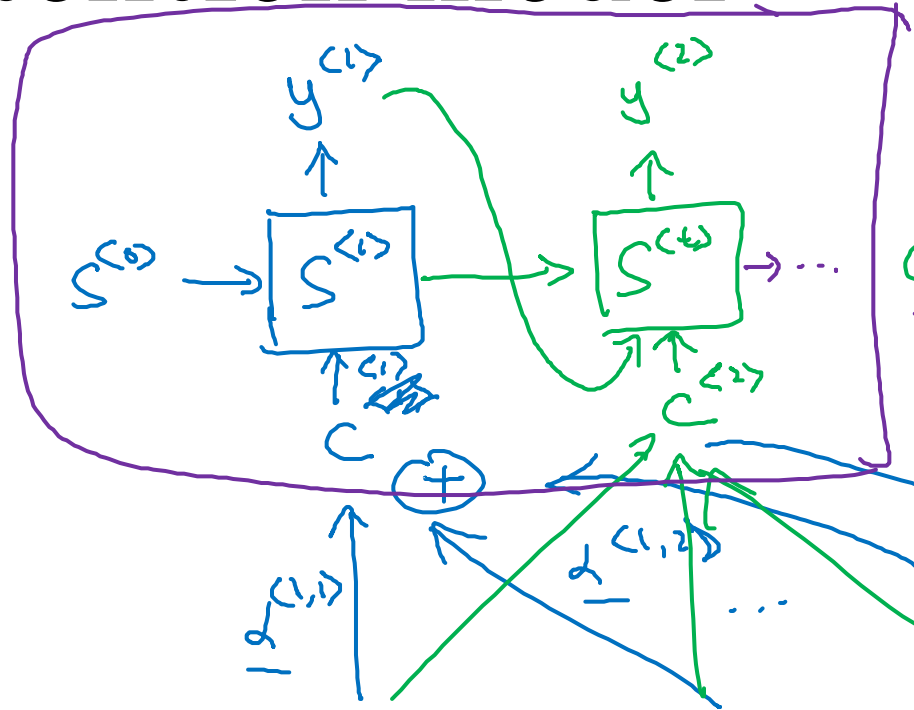
Bleu score

10   20   30   40   50

Sentence length

Andrew Ng

Attention model intuition

[Bahdanau et. al., 2014. Neural machine translation by jointly learning to align and translate]

Andrew Ng

Sequence to
sequence models

Attention model

deeplearning.ai

Attention model

$\alpha^{<t, t'>}$ — amount of "attention" $y^{<t>}$ should pay to $a^{<t'>}$.

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

$S^{<0>} \to \boxed{S^{<1>}} \to \boxed{S^{<2>}} \dashrightarrow \dots$

$c^{<1>}$

$c \quad \oplus$

$c^{<2>} = \sum_{t'} \alpha^{<2, t'>} \, a^{<t'>}$

$a^{<t>} = (\overrightarrow{a}^{<t>}, \overleftarrow{a}^{<t'>})$

$\sum_{t'} \alpha^{<1, t'>} = 1$

$c^{<1>} = \sum_{t'} \alpha^{<1, t'>} \, a^{<t'>}$

$\alpha^{<1,1>}$ $\alpha^{<1,2>}$ $\dots$ $\alpha^{<1,3>}$

$\overrightarrow{a}^{<0>} \to \boxed{\overrightarrow{a}^{<1>}}\;\boxed{\overleftarrow{a}^{<1>}} \qquad \boxed{\overrightarrow{a}^{<2>}}\;\boxed{\overleftarrow{a}^{<2>}} \qquad \boxed{\overrightarrow{a}^{<3>}}\;\boxed{\overleftarrow{a}^{<3>}} \qquad \boxed{}\;\boxed{} \qquad \boxed{\overrightarrow{a}^{<5>}}\;\boxed{\overleftarrow{a}^{<5>}} \leftarrow \overleftarrow{a}^{<6>}$

$x^{<1>}$ $x^{<2>}$ $x^{<3>}$ $x^{<4>}$ $x^{<5>}$

t'

jane   visite   l'Afrique   en   septembre

[Bahdanau et. al., 2014. Neural machine translation by jointly learning to align and translate]

Andrew Ng
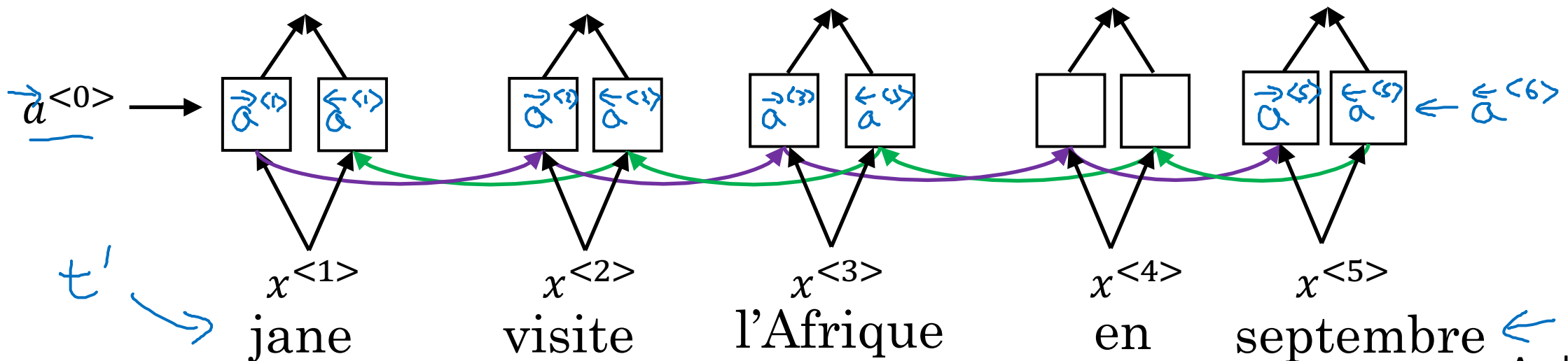
Computing attention $\alpha^{<t,t'>}$

T_x T_y

$\alpha^{<t,t'>}$ = amount of attention $y^{<t>}$ should pay to $a^{<t'>}$

$$\alpha^{<t,t'>} = \frac{\exp(e^{<t,t'>})}{\sum_{t'=1}^{T_x} \exp(e^{<t,t'>})}$$
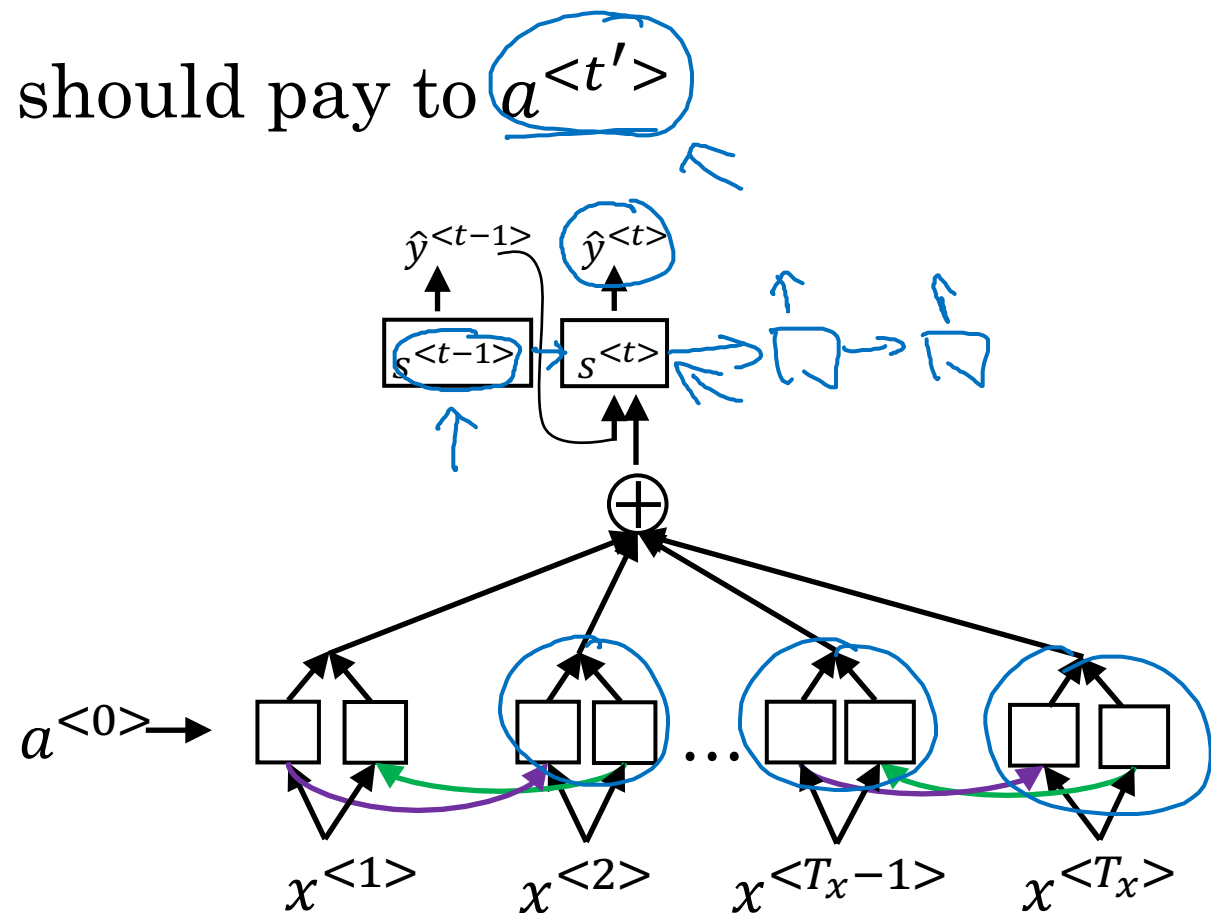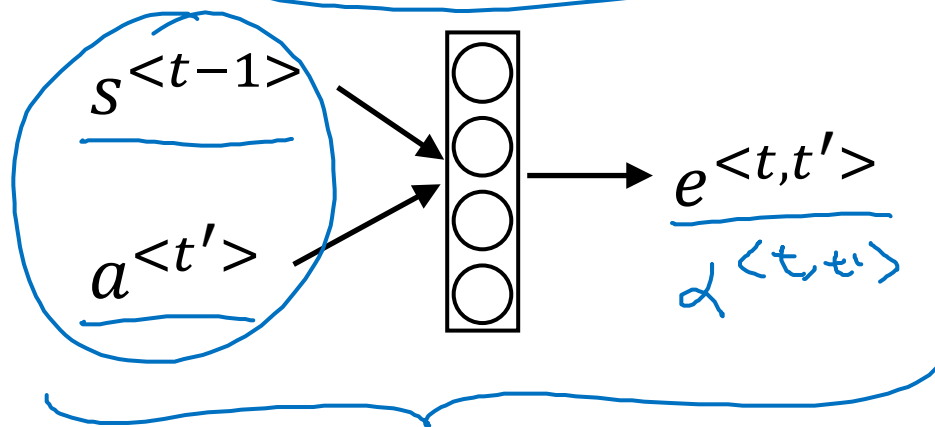
$s^{<t-1>}$

$a^{<t'>}$

$e^{<t,t'>}$

$\alpha^{<t,t'>}$

$\hat{y}^{<t-1>}$ $\hat{y}^{<t>}$

$s^{<t-1>}$ $s^{<t>}$

$\bigoplus$

$a^{<0>}$

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

[Bahdanau et. al., 2014. Neural machine translation by jointly learning to align and translate]

[Xu et. al., 2015. Show, attend and tell: Neural image caption generation with visual attention]

Andrew Ng

Attention examples

July 20th 1969 $\longrightarrow$ $1969 - 07 - 20$

23 April, 1564 $\longrightarrow$ $1564 - 04 - 23$

Visualization of $\alpha^{<t,t'>}$:
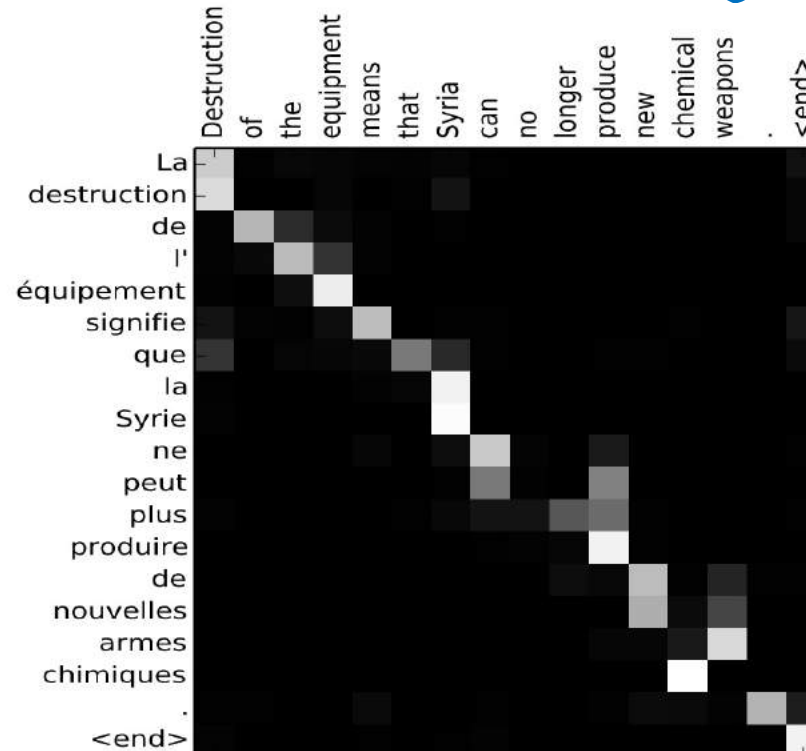
deeplearning.ai

Audio data

Speech recognition

Speech recognition problem

x y

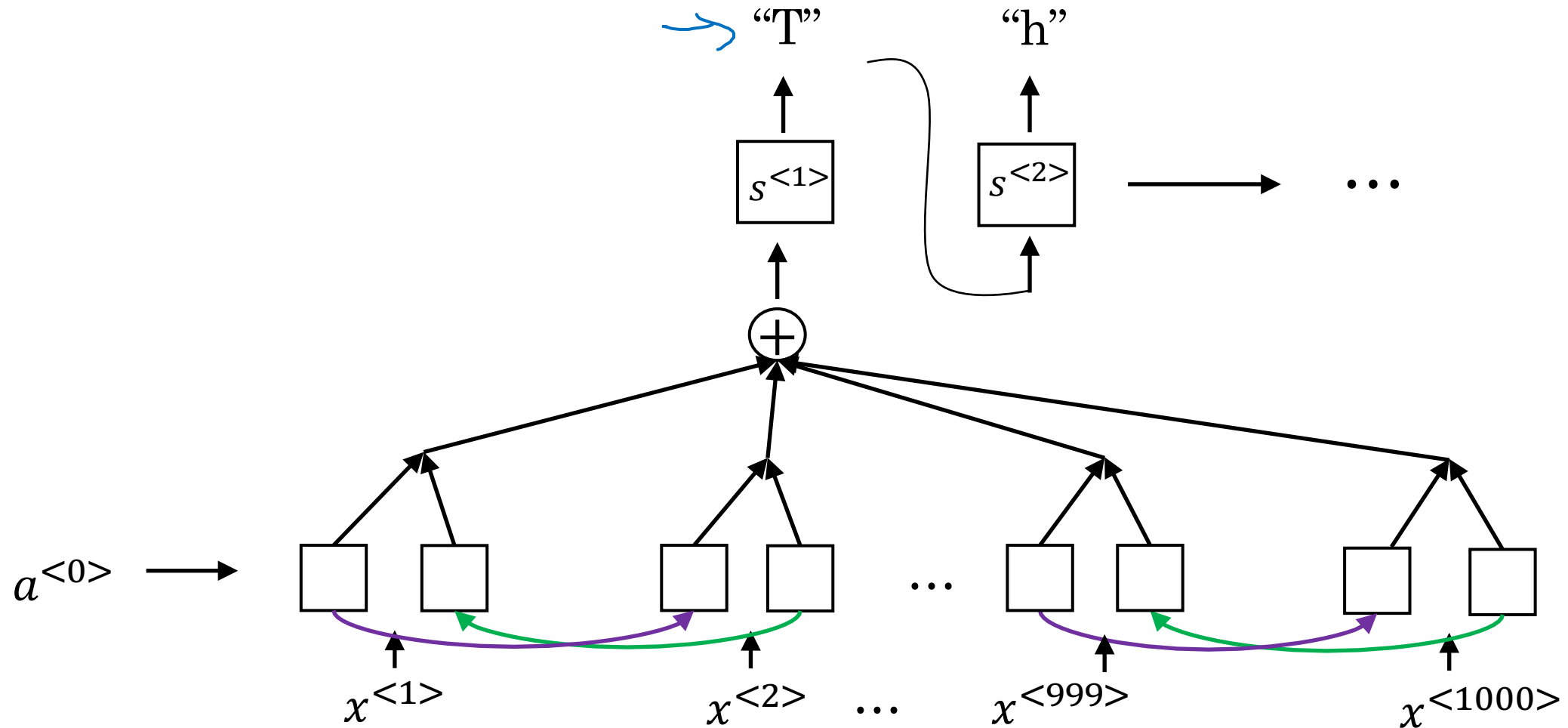audio clip   transcript
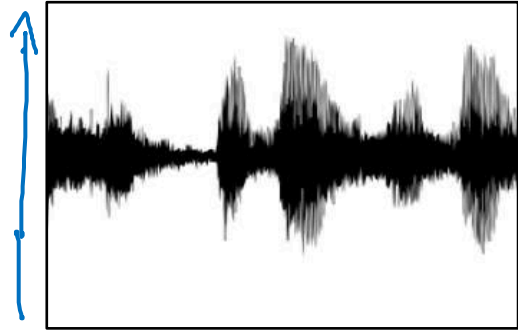
"the quick brown fox"

phonemes: de   kwik   braun

300h

3000h

100,000h

Andrew Ng
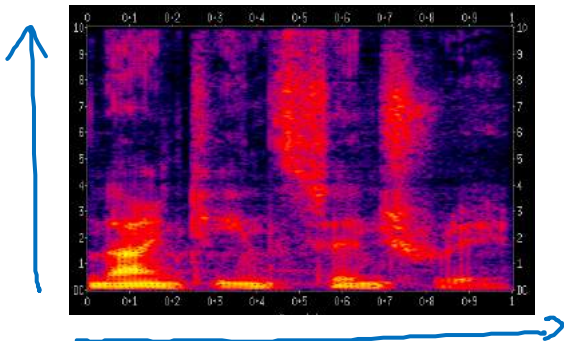
Attention model for speech recognition

CTC cost for speech recognition

(Connectionist temporal classification)

19 characters

"the quick brown fox"

$a^{<0>} \rightarrow \boxed{} \rightarrow \boxed{} \rightarrow \cdots \rightarrow \boxed{}$

$\hat{y}^{<1>} \quad \hat{y}^{<2>} \quad \hat{y}^{<1000>}$

$x^{<1>} \quad x^{<2>} \quad x^{<1000>}$

ttt_h_eee ----ᴗ---- qqq--   theᴗq

"space"   ⤺ "blank"
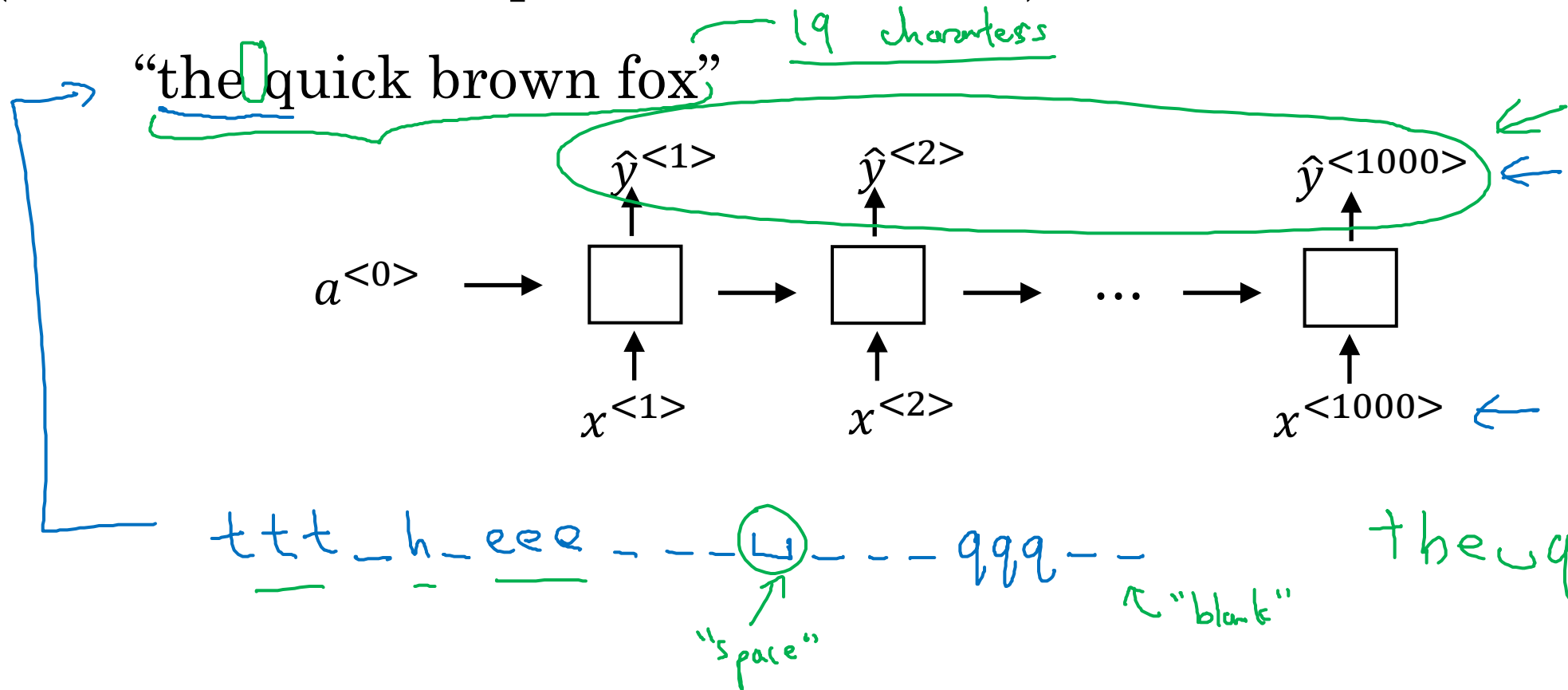
Basic rule: collapse repeated characters not separated by "blank"

[Graves et al., 2006. Connectionist Temporal Classification: Labeling unsegmented sequence data with recurrent neural networks]   Andrew Ng

Audio data

Trigger word detection

deeplearning.ai

What is trigger word detection?
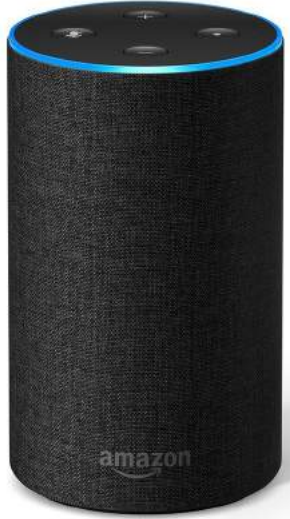
Amazon Echo
(Alexa)

Baidu DuerOS
(xiaodunihao)

Apple Siri
(Hey Siri)

Google Home
(Okay Google)

Andrew Ng

Trigger word detection algorithm
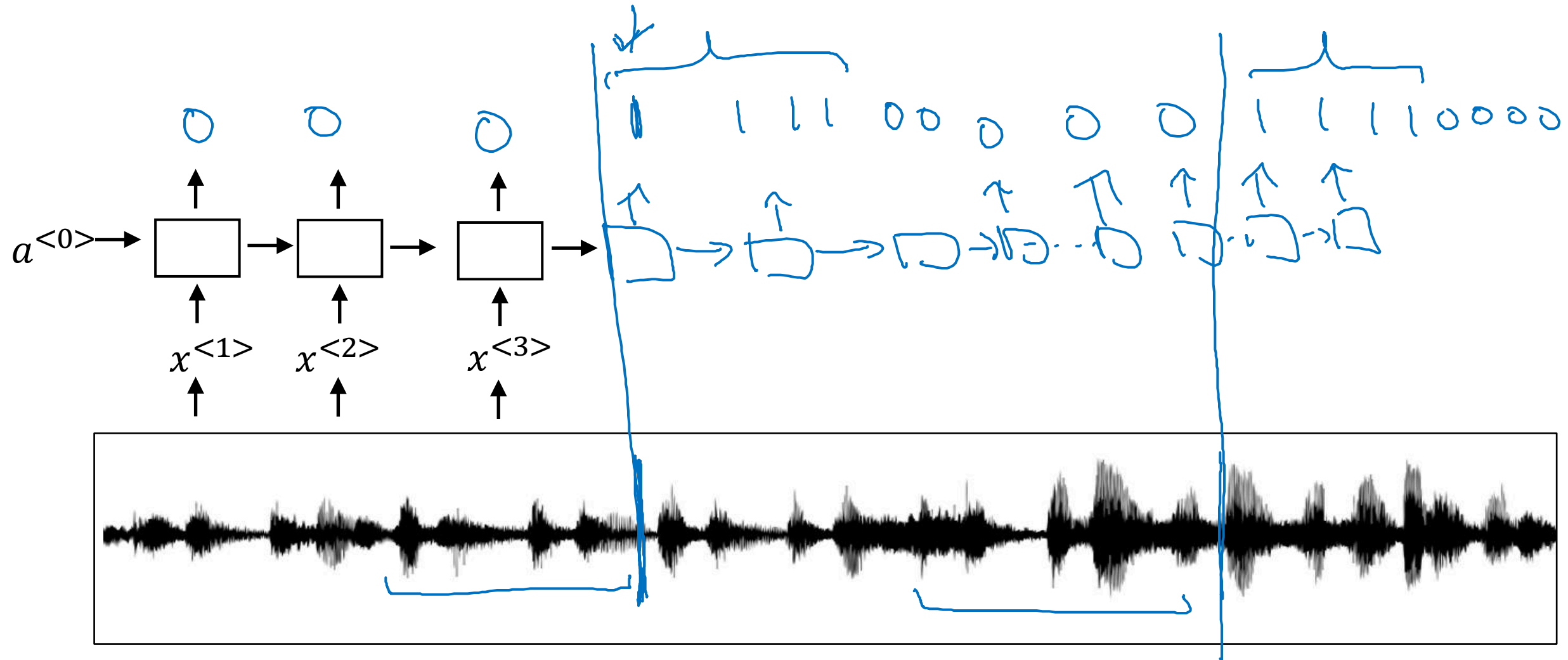
Conclusion

Summary and thank you

deeplearning.ai

Specialization outline

1. Neural Networks and Deep Learning

2. Improving Deep Neural Networks: Hyperparameter tuning, Regularization and Optimization

3. Structuring Machine Learning Projects

4. Convolutional Neural Networks

5. Sequence Models

Deep learning is a super power

Please buy this from shutterstock and replace in final video.

Andrew Ng

Thank you.

- Andrew Ng

Sequence to sequence models

Transformers Intuition

deeplearning.ai

Transformers Motivation

Increased complexity, sequential

RNN

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

x

GRU

LSTM

$y^{<t>}$

softmax

$a^{<t>}$

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

tanh

$a^{<t-1>}$ $a^{<t>}$

$f^{<t>}$ $i^{<t>}$ $\tilde{c}^{<t>}$ $o^{<t>}$ $a^{<t>}$

| forget gate | update gate | tanh | output gate |

$x^{<t>}$

Andrew Ng
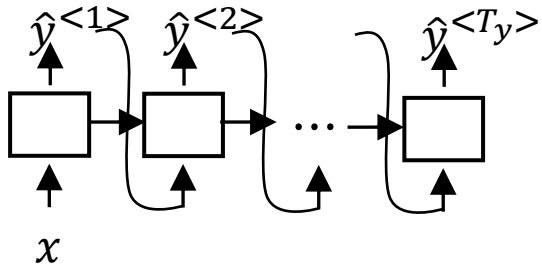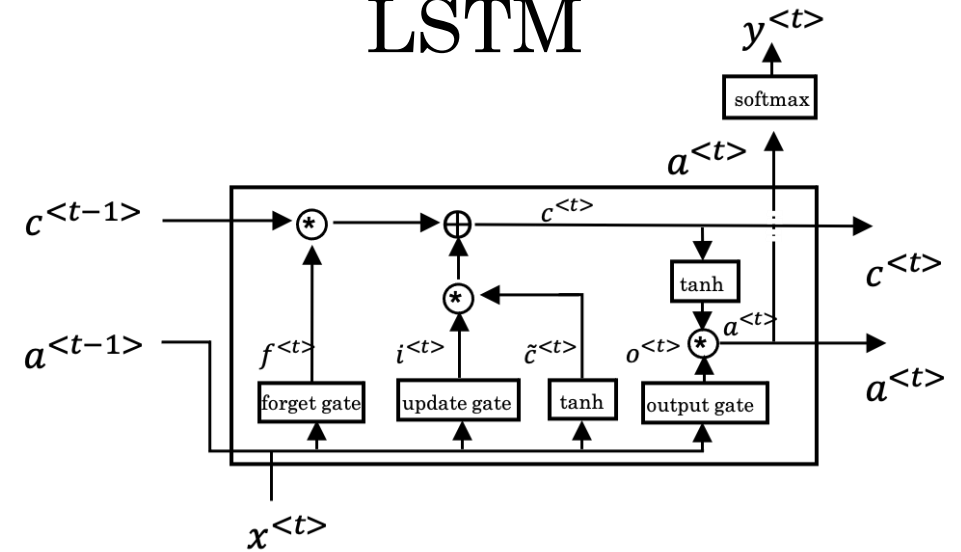
Transformers Intuition

- Attention + CNN
  - Self-Attention
  - Multi-Head Attention

Andrew Ng
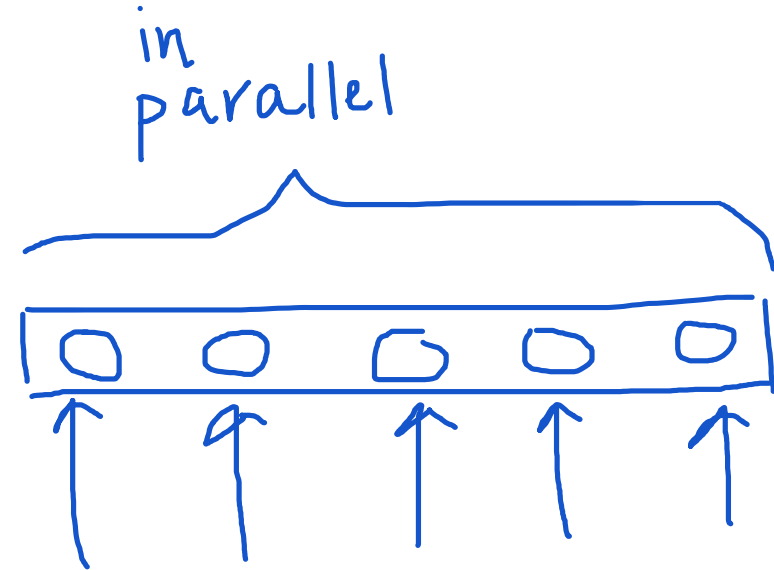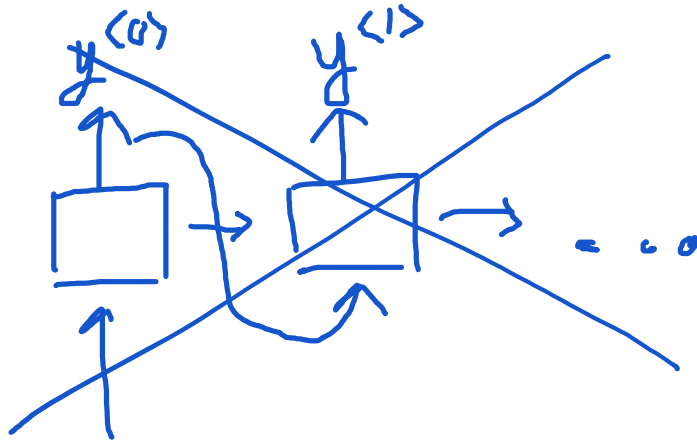
Sequence to sequence models

Self-Attention

deeplearning.ai

Self-Attention Intuition

$A(q, K, V)$ = attention-based vector representation of a word

↳ calculate for each word

RNN Attention

$$\alpha^{<t,t'>} = \frac{\exp(e^{<t,t'>})}{\sum_{t'=1}^{T_x} \exp(e^{<t,t'>})}$$

Transformers Attention

$$A(q, K, V) = \sum_i \frac{\exp(e^{<q \cdot k^{<i>}>})}{\sum_j \exp(e^{<q \cdot k^{<j>}>})} v^{<i>}$$

$x^{<1>}$ $x^{<2>}$ $x^{<3>}$ $x^{<4>}$ $x^{<5>}$

Jane   visite   l'Afrique   en   septembre

Andrew Ng

Self-Attention

$$A(q, K, V) = \sum_i \frac{\exp(e^{<q \cdot k^{<i>}>})}{\sum_j \exp(e^{<q \cdot k^{<j>}>})} v^{<i>}$$

softmax

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$A^{<1>}$ $A^{<2>}$ $A^{<3>}$ $A^{<4>}$ $A^{<5>}$

(+)

$v^{<1>}$ (X) $v^{<2>}$ (X) $v^{<3>}$ (X) $v^{<4>}$ (X) $v^{<5>}$

$q^{<3>} \cdot k^{<1>}$ $q^{<3>} \cdot k^{<2>}$ $q^{<3>} \cdot k^{<4>}$ $q^{<3>} \cdot k^{<5>}$

$q^{<1>}, k^{<1>}, v^{<1>}$ $q^{<2>}, k^{<2>}, v^{<2>}$ $q^{<3>}, k^{<3>}, v^{<3>}$ $q^{<4>}, k^{<4>}, v^{<4>}$ $q^{<5>}, k^{<5>}, v^{<5>}$

$x^{<1>}$ $x^{<2>}$ $x^{<3>}$ $x^{<4>}$ $x^{<5>}$
Jane   visite   l'Afrique   en   septembre
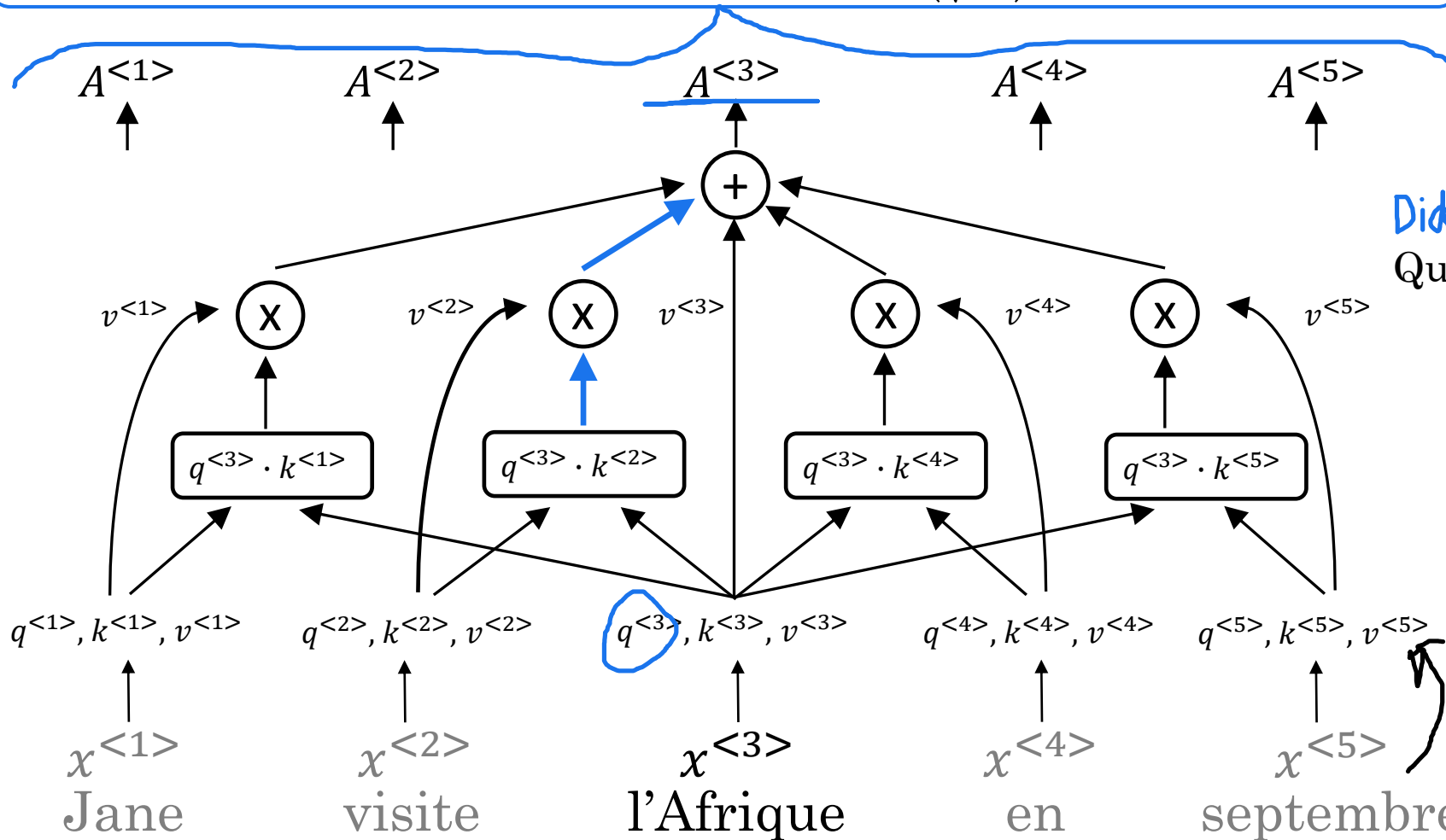
Did what?

Query (Q) Key (K) Value (V)

$q^{<1>}$ $k^{<1>}$
$q^{<2>}$ $k^{<2>}$ person $v^{<1>}$ Jane
$q^{<3>}$ $k^{<3>}$ action $v^{<2>}$ visit
$q^{<4>}$ $k^{<4>}$ $v^{<3>}$
$q^{<5>}$ $k^{<5>}$ $v^{<4>}$
                                $v^{<5>}$

What's happening there?

W^Q, W^K, W^V

[Vaswani et al. 2017, Attention Is All You Need]

Andrew Ng

Sequence to sequence models

Multi-Head Attention

Multi-Head Attention

$$MultiHead(Q, K, V) = concat(head_1 head_2 \ldots head_n)W_0$$

$$head_i = Attention(W_i^Q Q, W_i^K K, W_i^V V)$$

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

W_o

concat

$h = \# \text{ heads}$

$head_3$

$head_2$

$head_1$

$Attention(W_i^Q Q, W_i^K K, W_i^V V)$

Multi-Head Attention

Q K V

$W_3^Q, W_3^K, W_3^V, \text{Who?}$

$W_i^Q q^{<1>}, W_i^K k^{<1>}, W_i^V v^{<1>}$ $W_i^Q q^{<2>}, W_i^K k^{<2>}, W_i^V v^{<2>}$ $W_i^Q q^{<3>}, W_i^K k^{<3>}, W_i^V v^{<3>}$ $W_i^Q q^{<4>}, W_i^K k^{<4>}, W_i^V v^{<4>}$ $W_i^Q q^{<5>}, W_i^K k^{<5>}, W_i^V v^{<5>}$

$W_2^Q, W_2^K, W_2^V, \text{When?}$

$W_1^Q, W_1^K, W_1^V, \text{Did what?}$

$q^{<1>}, k^{<1>}, v^{<1>}$ $q^{<2>}, k^{<2>}, v^{<2>}$ $q^{<3>}, k^{<3>}, v^{<3>}$ $q^{<4>}, k^{<4>}, v^{<4>}$ $q^{<5>}, k^{<5>}, v^{<5>}$
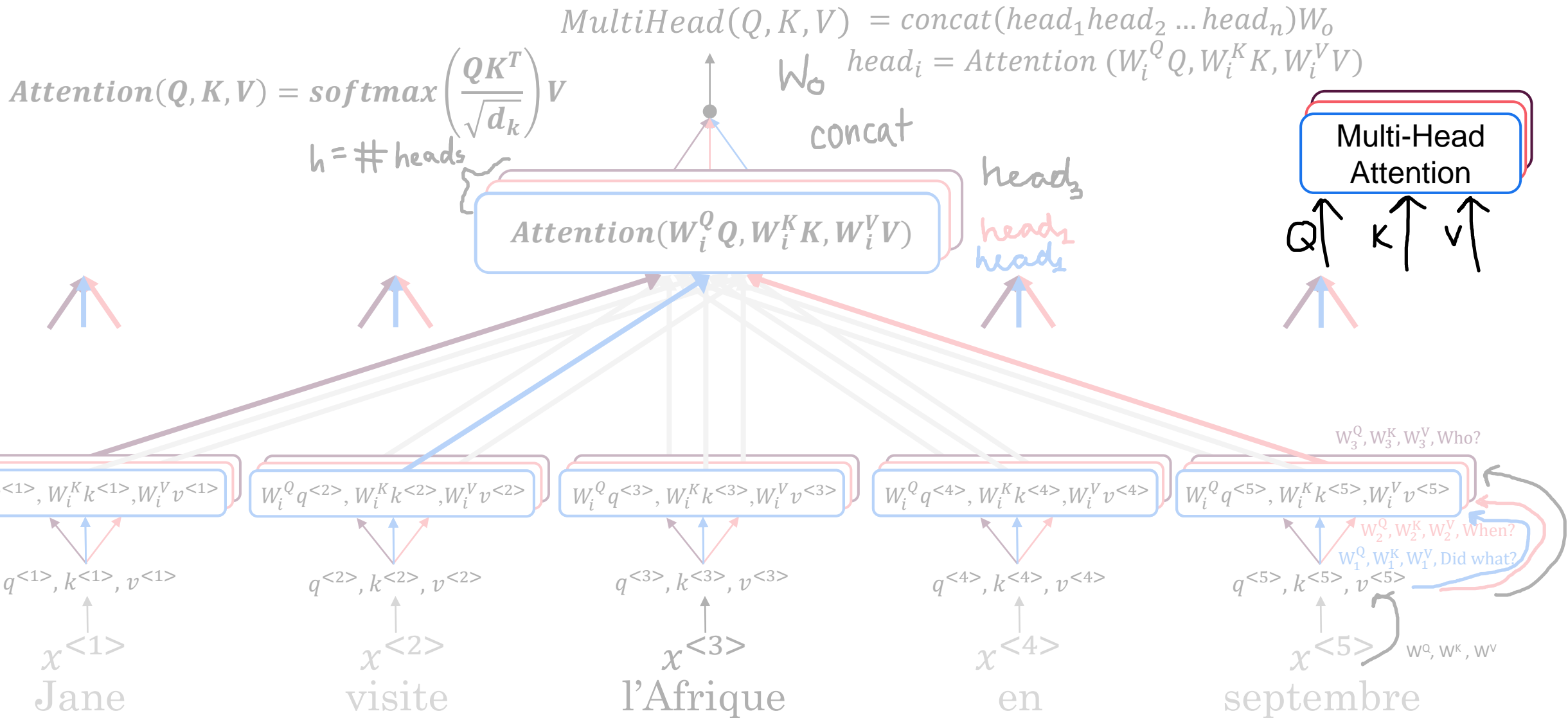
$x^{<1>}$ $x^{<2>}$ $x^{<3>}$ $x^{<4>}$ $x^{<5>}$ W^Q, W^K, W^V

Jane   visite   l'Afrique   en   septembre

[Vaswani et al. 2017, Attention Is All You Need]

Andrew Ng

Sequence to sequence models

Transformers

deeplearning.ai

Transformer Details

<SOS> Jane visits Africa in September <EOS>

Encoder

Decoder

Softmax

Linear

Add & Norm

Feed Forward
Neural Network

Add & Norm

Multi-Head
Attention

N_x

Add & Norm

Feed Forward
Neural Network

Add & Norm

Multi-Head
Attention

K V Q

N_x

Add & Norm

Multi-Head
Attention

Q K V

$i=0$
$i=1$
$i=2$
$i=3$

pos 1   2   3

$d=4$ $\{$ $i=0$ $i=1$ $i=2$ $i=3$

$x^{<1>}$ + $p^{<1>}$

$p^{<3>}$

$\oplus$

Positional Encoding
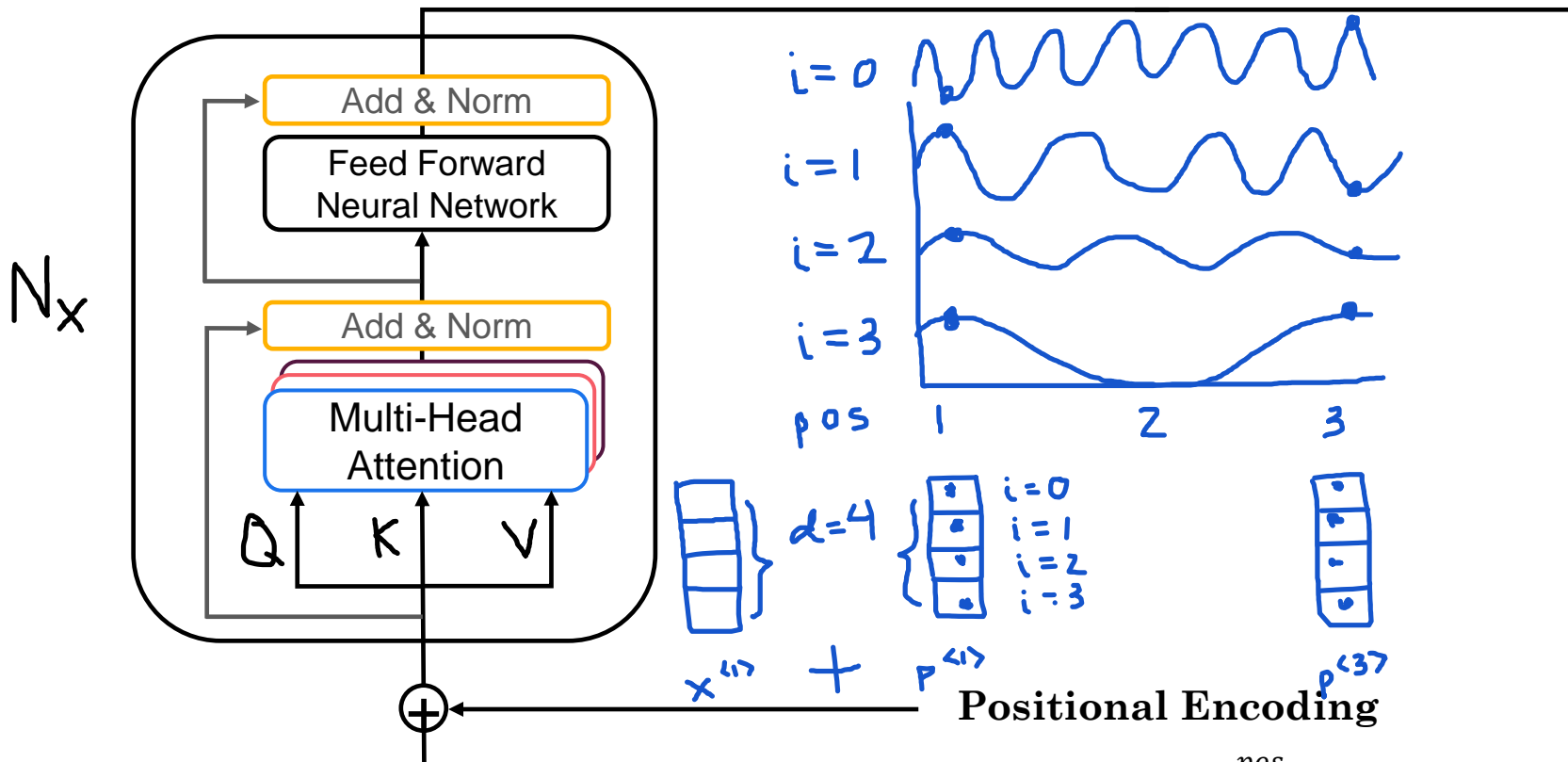
<SOS> $x^{<1>}$ $x^{<2>}$ ... $x^{<T_x-1>}$ $x^{<T_x>}$ <EOS>

Jane visite l'Afrique en septembre

$$PE_{(pos,2i)} = sin\left(\frac{pos}{1000^{\frac{2i}{d}}}\right)$$

$$PE_{(pos,2i+1)} = cos\left(\frac{pos}{1000^{\frac{2i}{d}}}\right)$$

$\oplus$

<SOS> $y^{<1>}$ $y^{<2>}$ ... $y^{<T_y-1>}$ $y^{<T_y>}$

<SOS> Jane visits Africa in September

[Vaswani et al. 2017, Attention Is All You Need]

Andrew Ng