deeplearning.ai

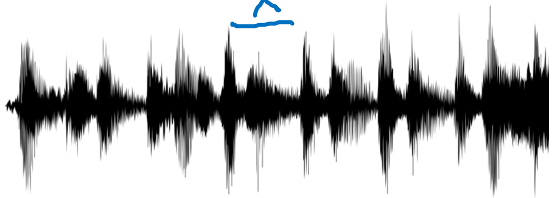
Recurrent Neural Networks

Why sequence models?

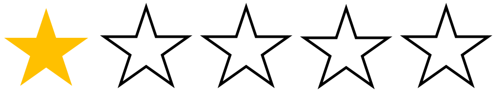
Examples of sequence data

x

y

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

Recurrent Neural Networks

Notation

deeplearning.ai

Motivating example

NLP

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

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

$T_x = 9$

y: 1 1 0 1 1 0 0 0 0

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

$T_y = 9$

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

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

Andrew Ng

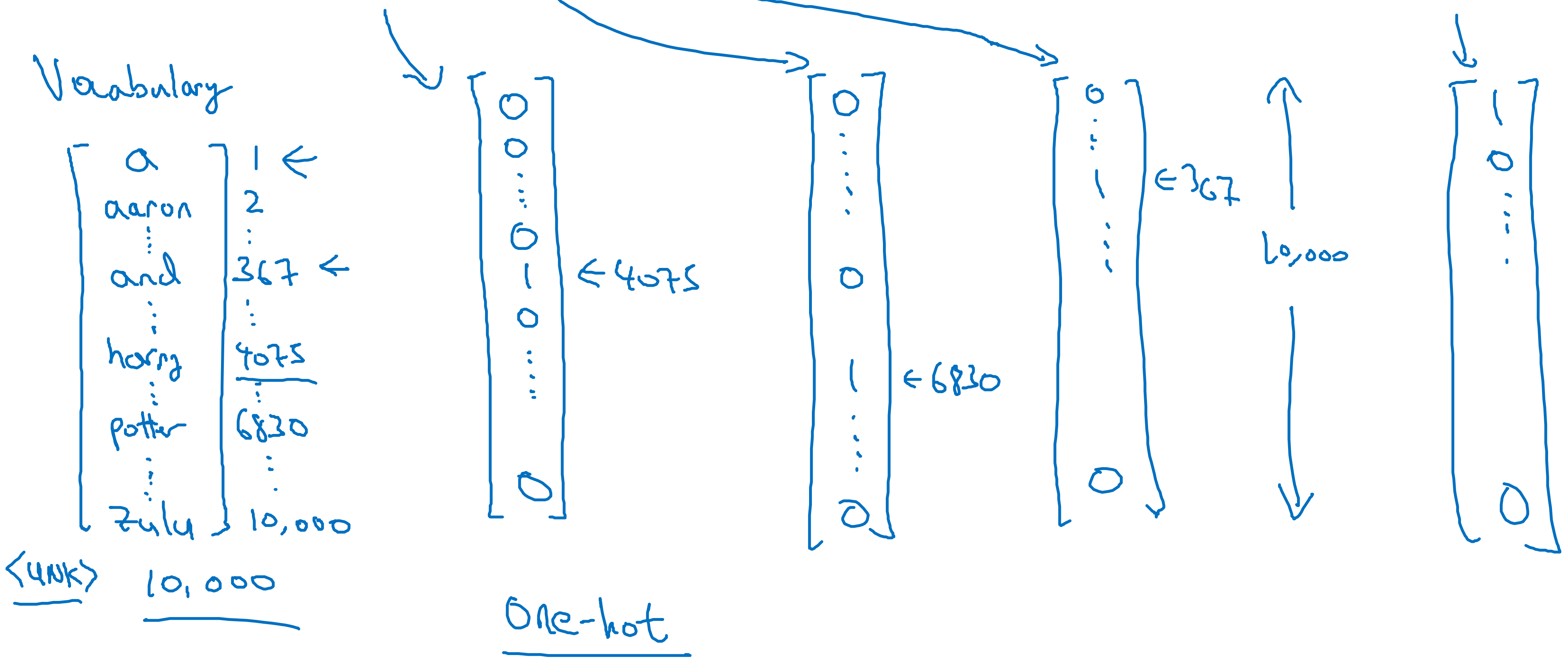
Representing words

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

$x \rightarrow y$

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

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



Andrew Ng

Representing words

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

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

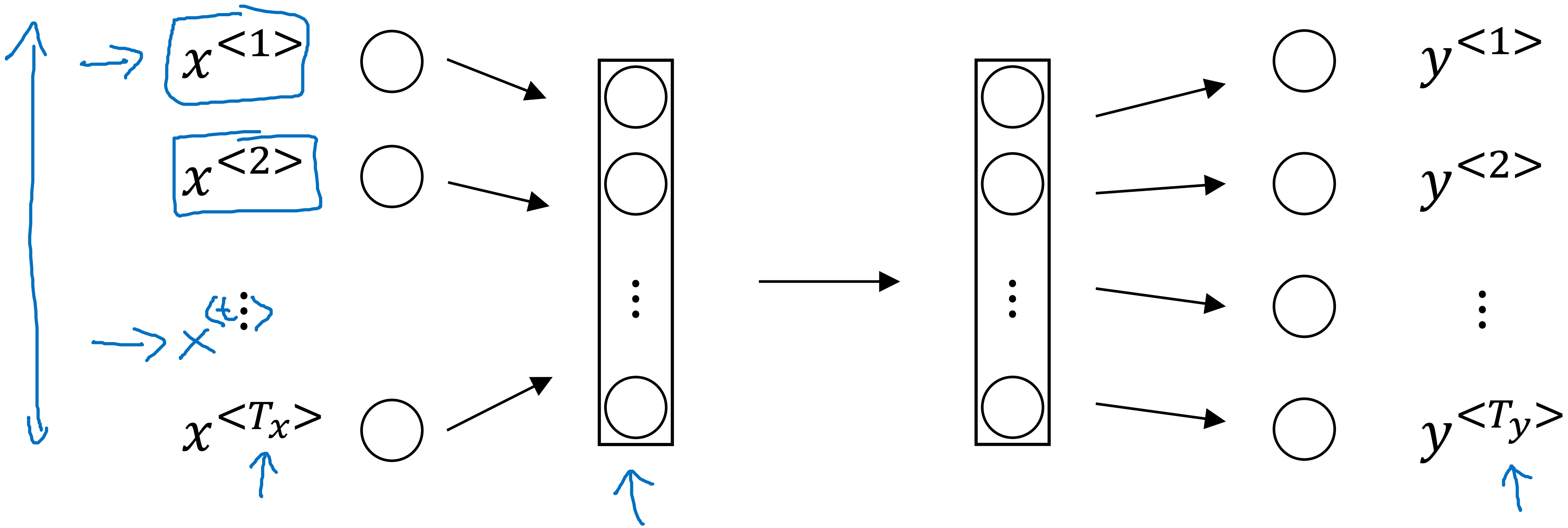
And = 367
Invented = 4700
A = 1
New = 5976
Spell = 8376
Harry = 4075
Potter = 6830
Hermione = 4200
Gran… = 4000

deeplearning.ai

Recurrent Neural Networks

Recurrent Neural Network Model

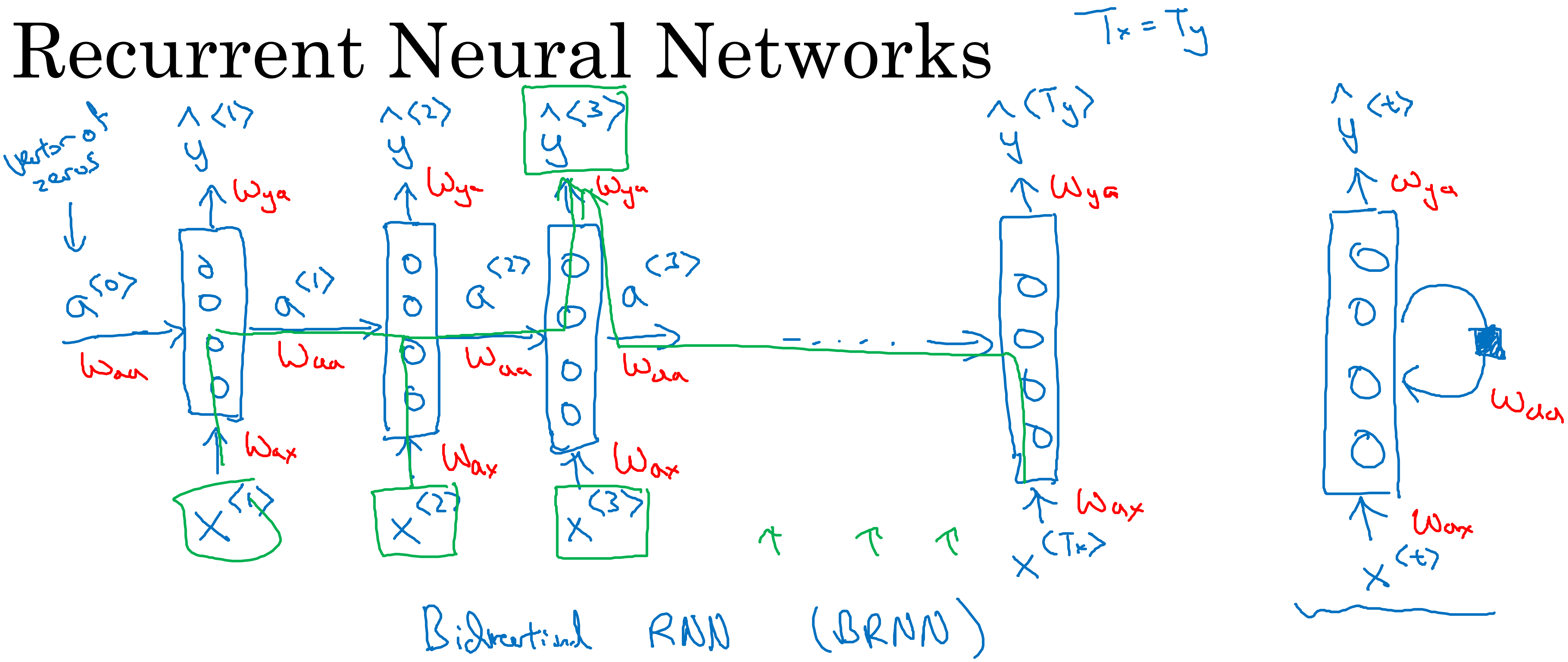
Why not a standard network?



Problems:

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

Recurrent Neural Networks



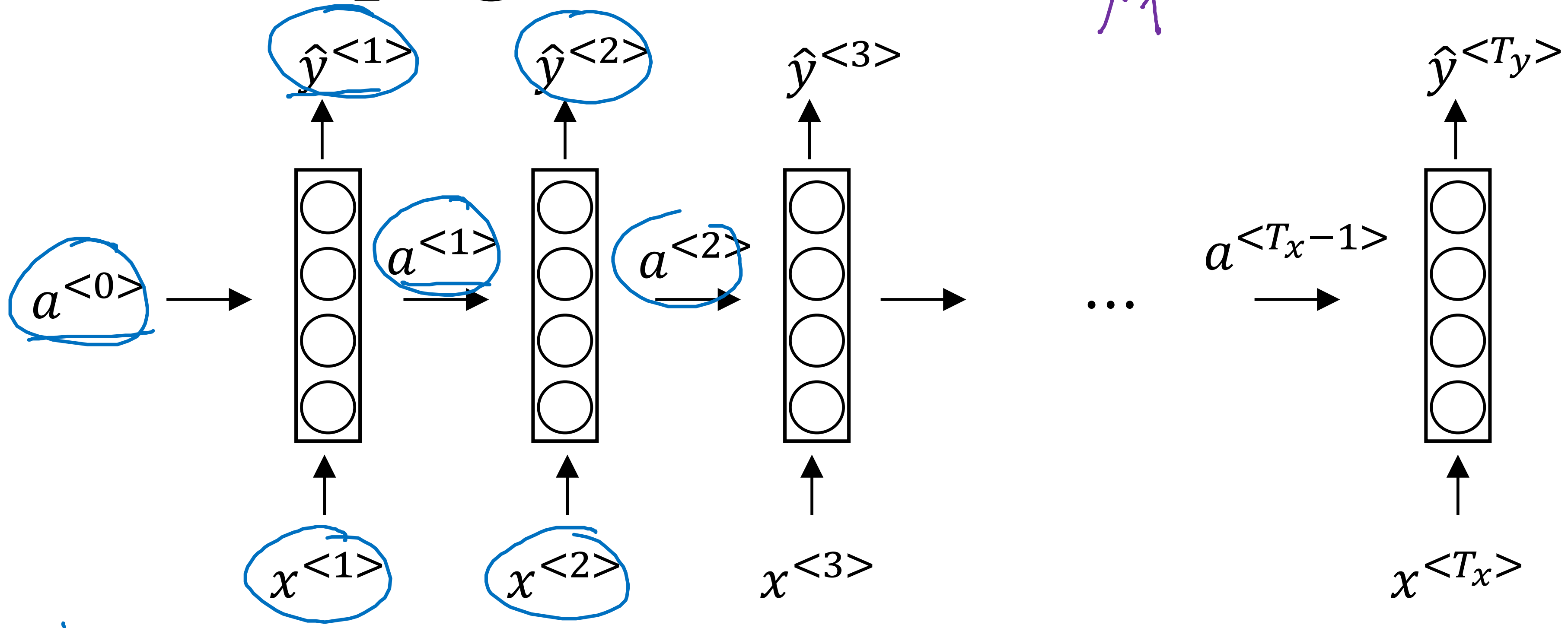
He said, "Teddy Roosevelt was a great President."

He said, "Teddy bears are on sale!"

Andrew Ng

Forward Propagation

$a \leftarrow W_{ax} x^{<1>}$



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

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

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

$$\boxed{\begin{aligned} a^{<t>} &= g(W_{aa} a^{<t-1>} + W_{ax} x^{<t>} + b_a) \\ \hat{y}^{<t>} &= g(W_{ya} a^{<t>} + b_y) \end{aligned}}$$

Simplified RNN notation

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

W_{aa}: (100,100), $a^{<t-1>}$: 100, W_{ax}: (100,10,000), $x^{<t>}$: 10,000

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

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

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

$$\underbrace{\left[\, \underbrace{W_{aa}}_{100} \;\vdots\; \underbrace{W_{ax}}_{10\,000} \,\right]}_{100 \text{ rows}} = W_a$$

$(100, 10100)$

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

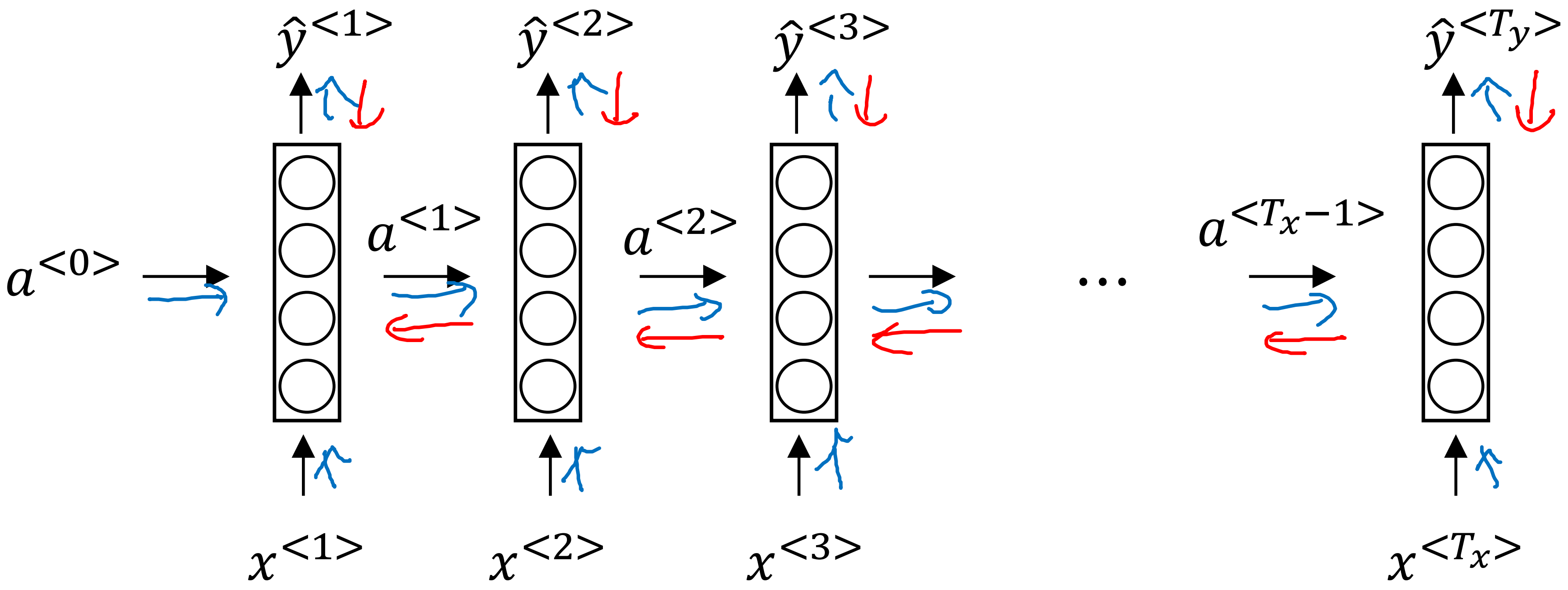
$$[W_{aa} \vdots W_{ax}] \begin{bmatrix} a^{(t-1)} \\ x^{(t)} \end{bmatrix} = W_{aa}a^{(t-1)} + W_{ax}x^{(t)}$$

deeplearning.ai

Recurrent Neural Networks

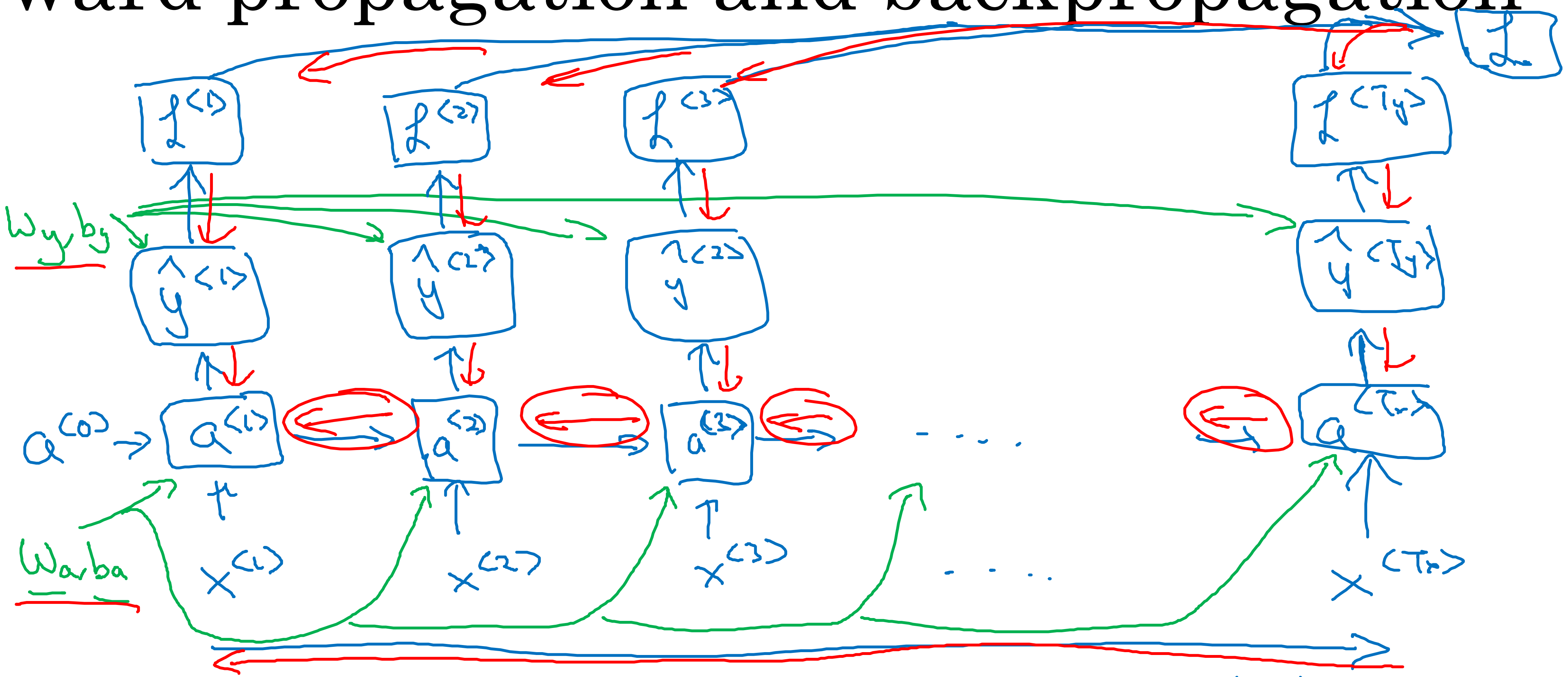
Backpropagation through time

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

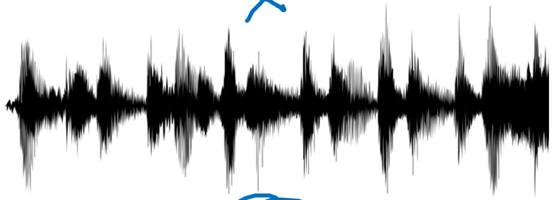
deeplearning.ai

Recurrent Neural Networks

Different types of RNNs

Examples of sequence data

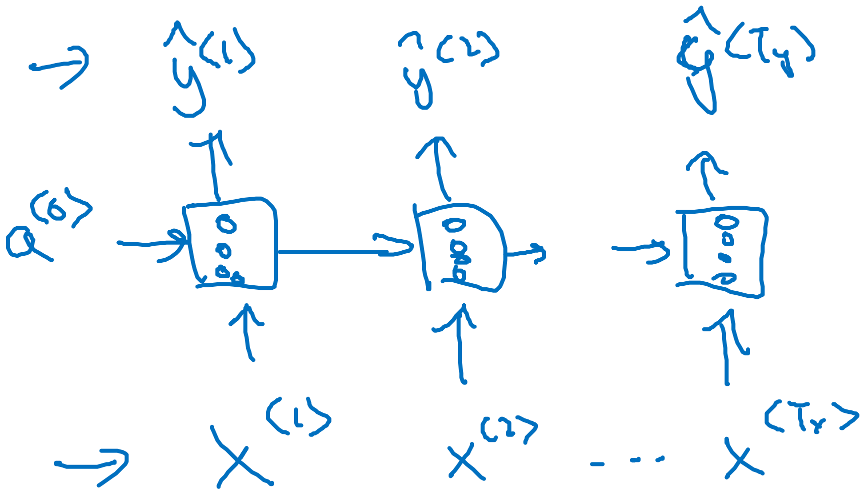
T_x T_y

	x		y
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

$T_x = T_y$

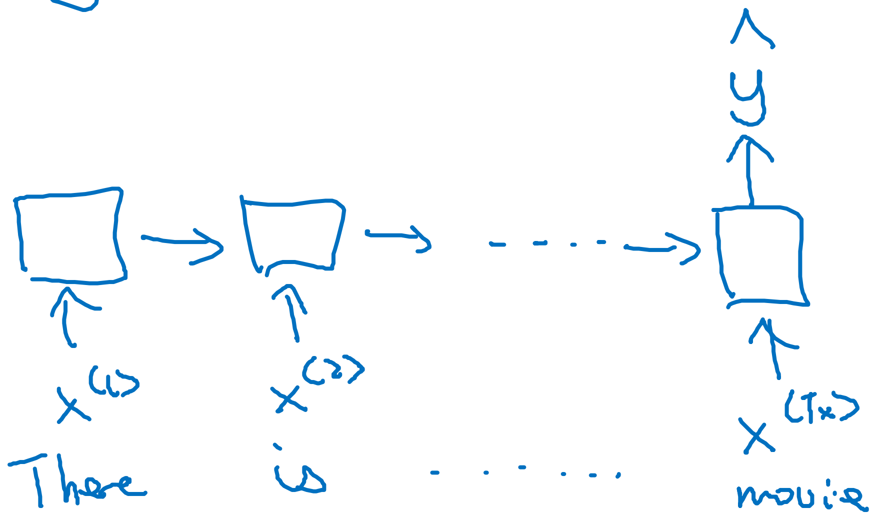


Many-to-many

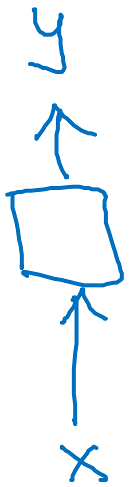
Sentiment classification

x = text

$y = 0/1 \qquad 1 \ldots 5$



Many-to-one

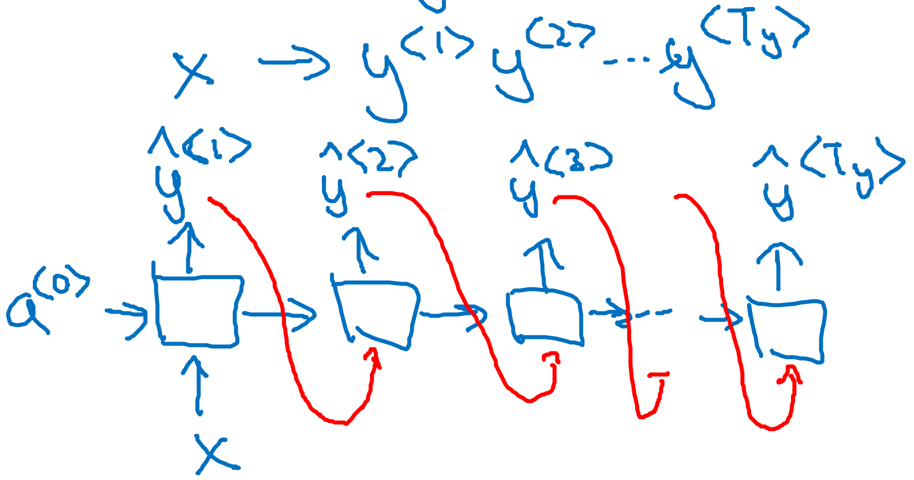


one-to-one

Andrew Ng

Examples of RNN architectures

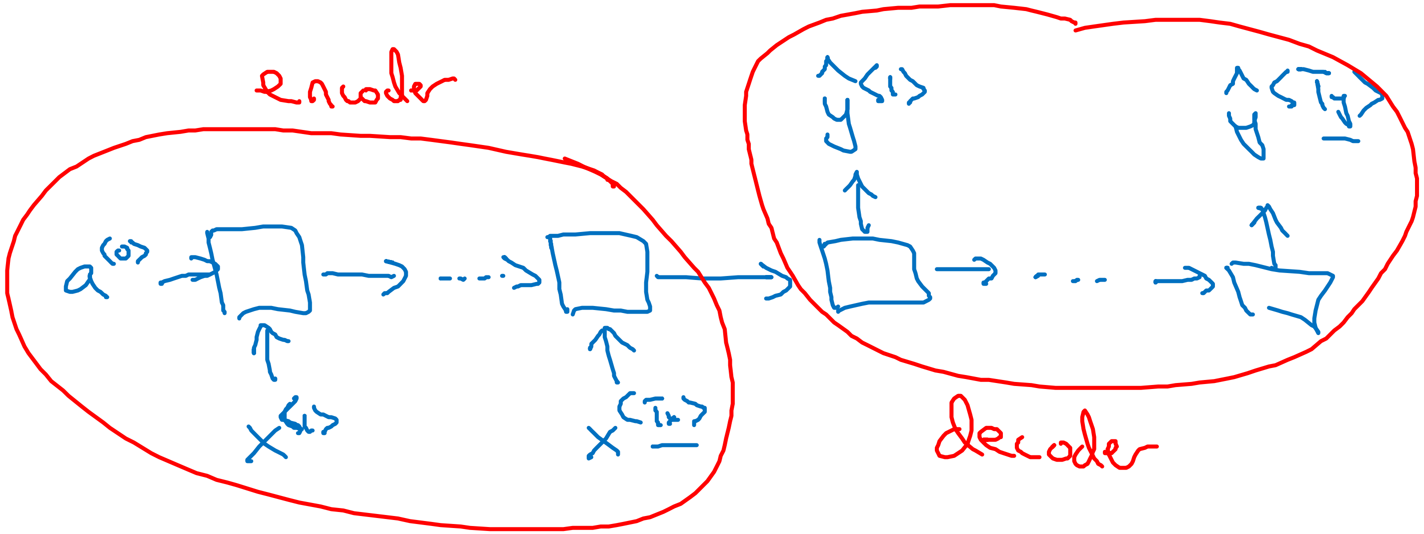
Music generation



One-to-many

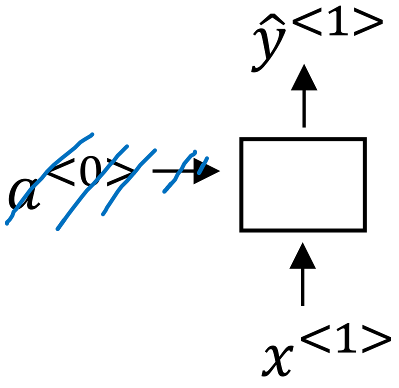
$x = \emptyset$

Machine translation

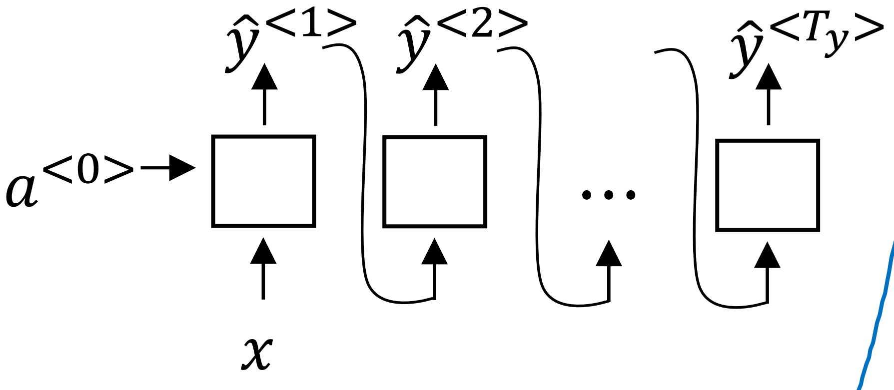


Many-to-many

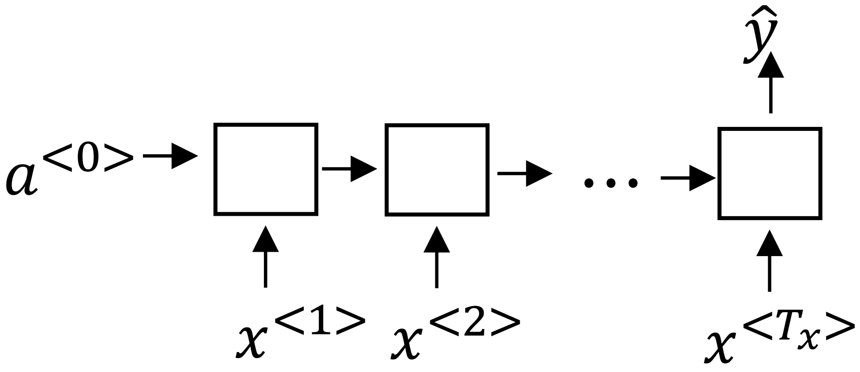
Summary of RNN types



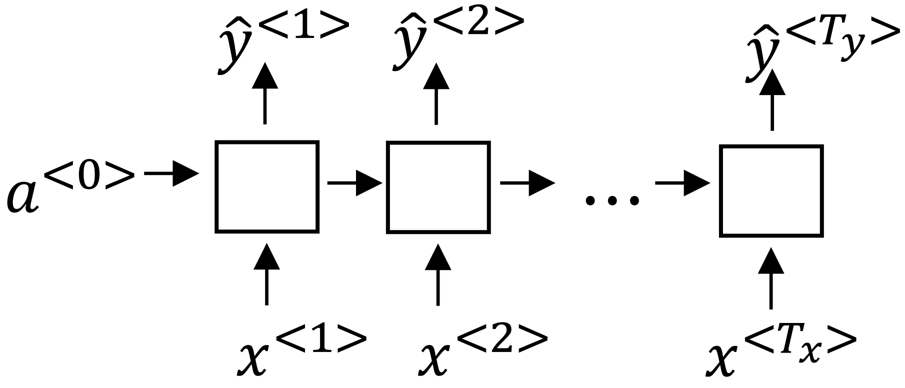
One to one



One to many

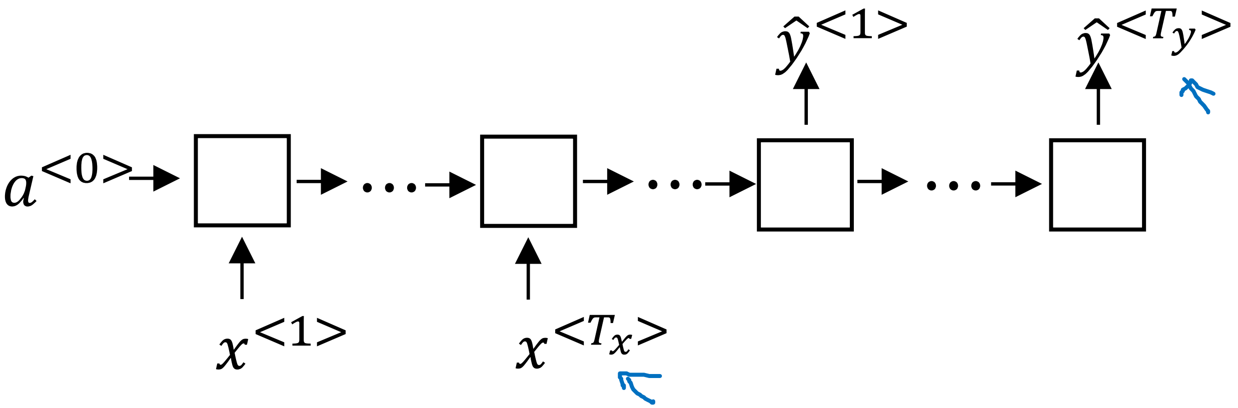


Many to one



Many to many

$T_x = T_y$



Many to many

Andrew Ng

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.

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

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

$P(\text{sentence}) = ?$ $\qquad P(y^{<1>}, y^{<2>}, \ldots, y^{<T_y>})$

Andrew Ng

Language modelling with an RNN

Training set: large corpus of english text.

Tokenize

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

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

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

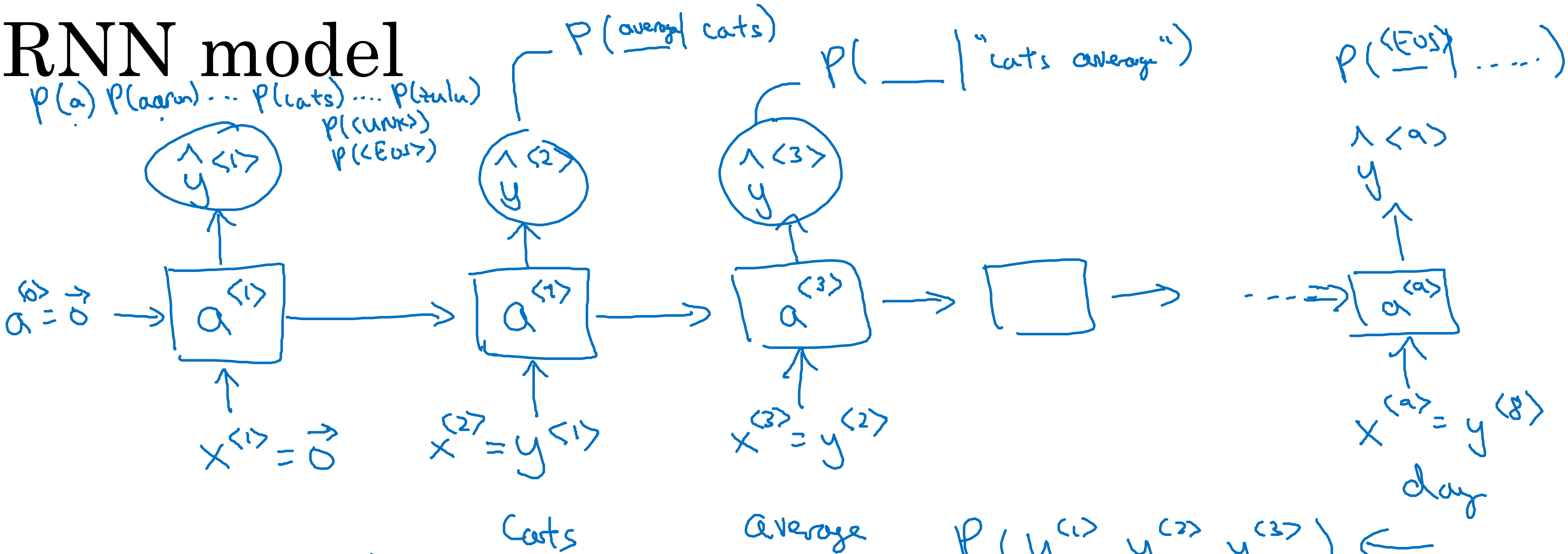
The Egyptian ~~Mau~~ is a bread of cat. <EOS>

<UNK>

10,000

Andrew Ng

RNN model



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

$\boxed{\text{Cats average 15}}$ hours of sleep a day. <EOS>

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

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

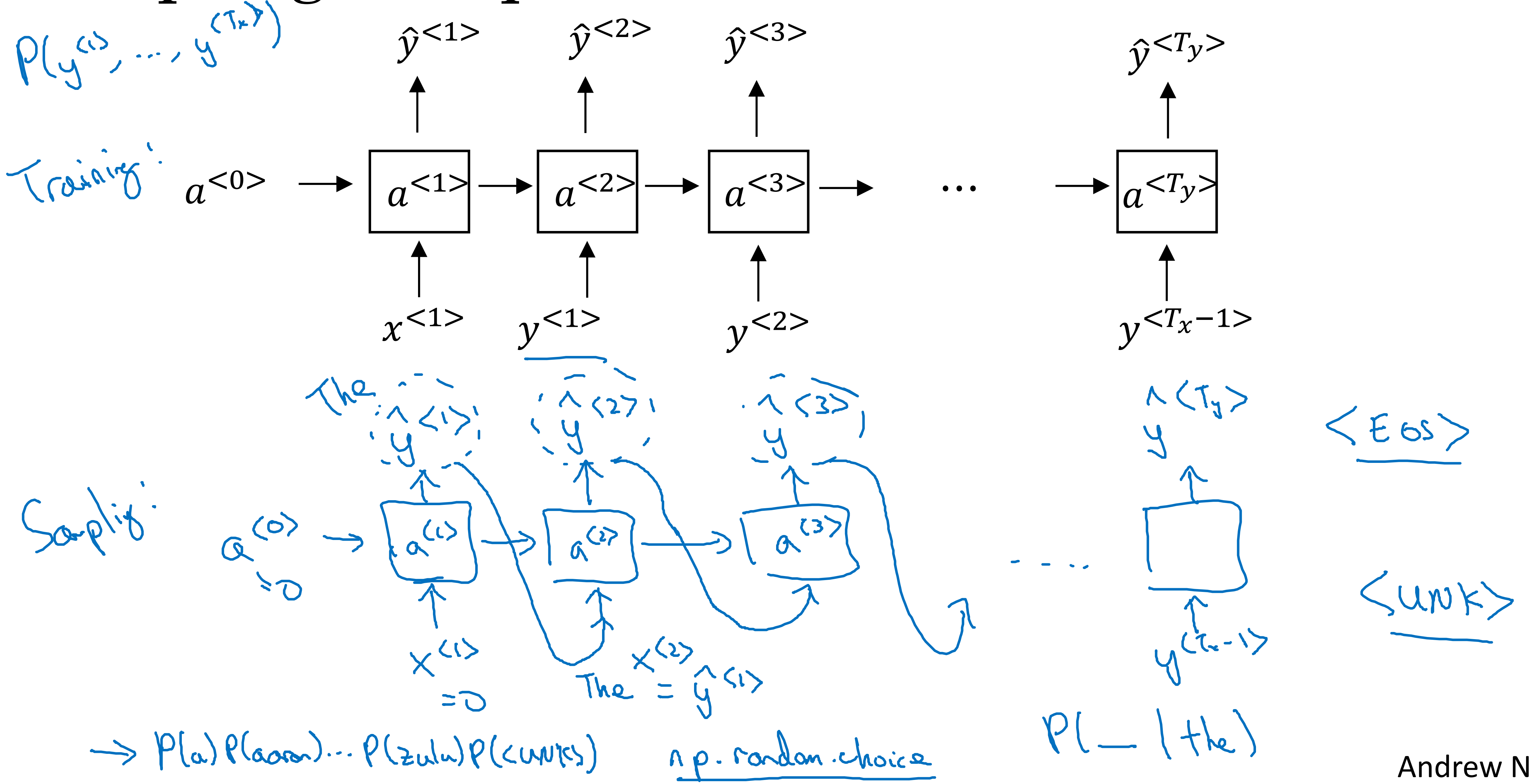
Andrew Ng

deeplearning.ai

Recurrent Neural Networks

Sampling novel sequences

Sampling a sequence from a trained RNN



Character-level language model

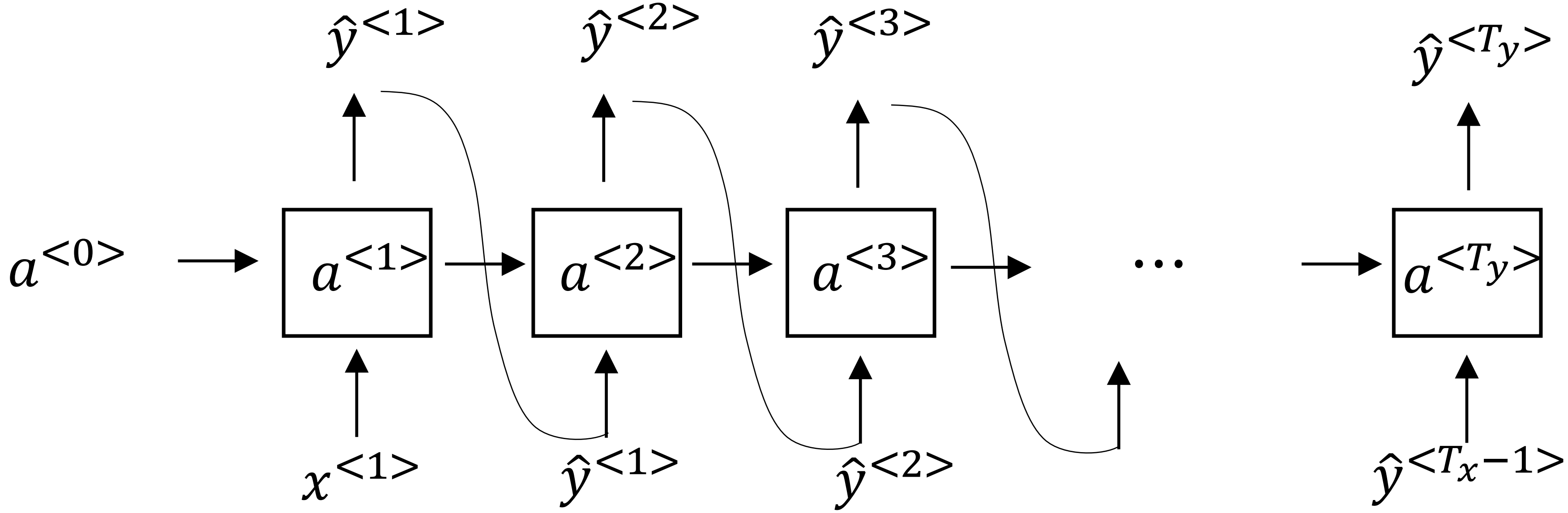
→ Vocabulary = [a, aaron, …, zulu, <UNK>] ←

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

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

Cat average …

Mau



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.

deeplearning.ai

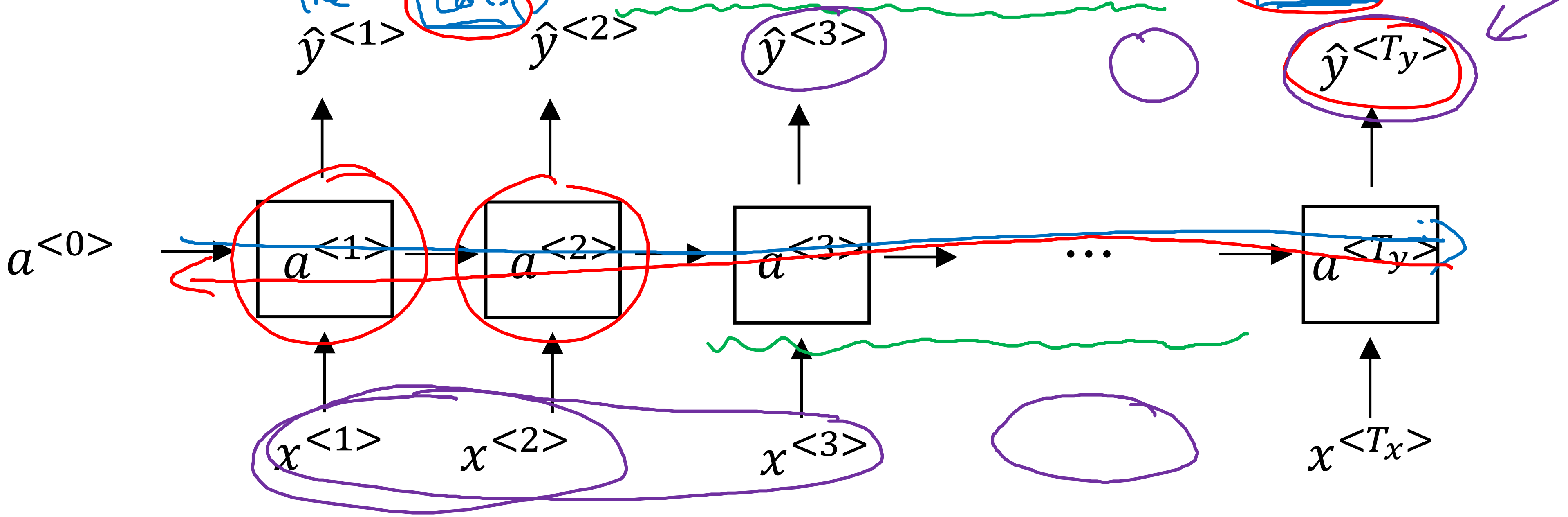
Recurrent Neural Networks

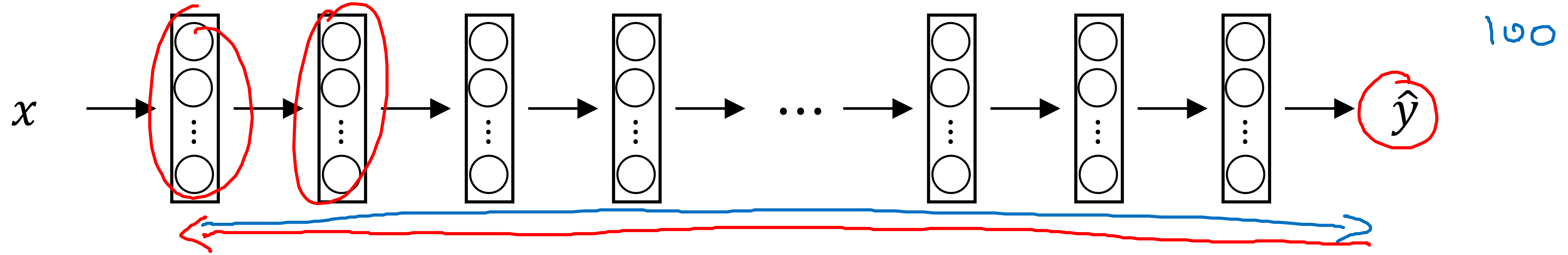
Vanishing gradients with RNNs

Vanishing gradients with RNNs

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

The cats, wh ... ., were full





Exploding gradients. NaN Gradient clipping

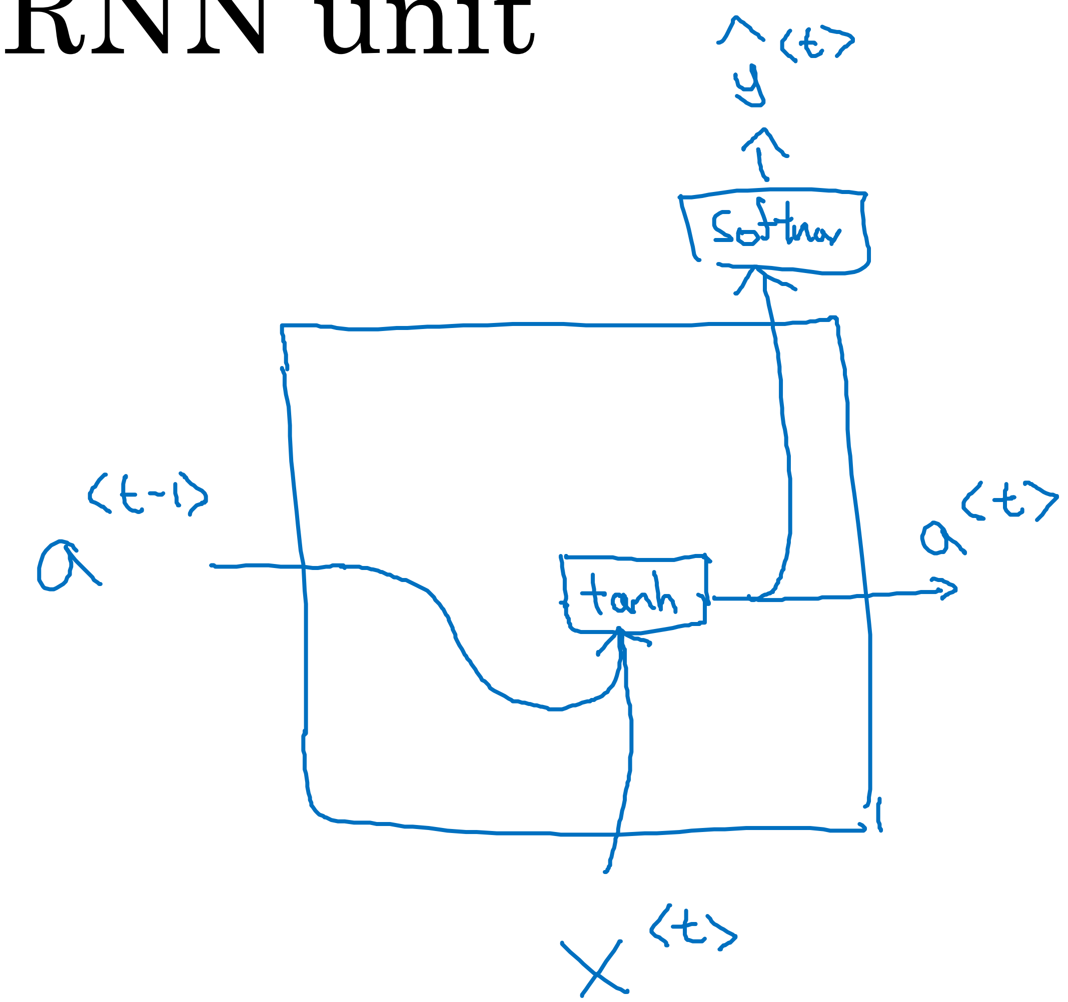
Andrew Ng

deeplearning.ai

Recurrent Neural Networks

Gated Recurrent Unit (GRU)

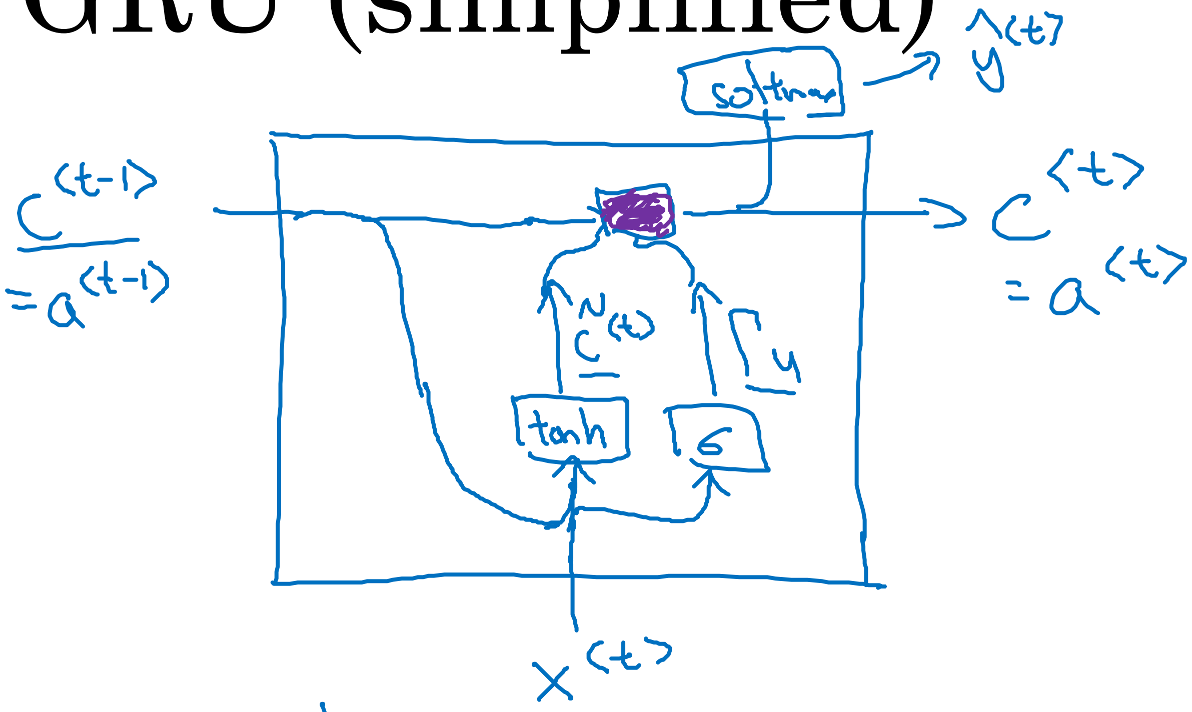
RNN unit

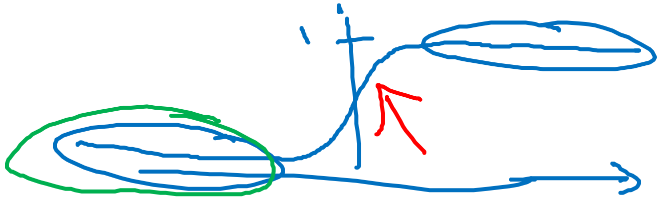


tanh

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

GRU (simplified)





c = memory cell

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

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

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

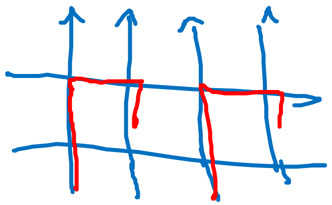
"update"

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

$\Gamma_u = 1$

element-wise

$\Gamma_u = 0.000001$



Gate

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

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

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

[Cho et al., 2014. On the properties of neural machine translation: Encoder-decoder approaches]
[Chung et al., 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling]

Full GRU

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

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

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

LSTM

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

The cat, which ate already, was full.

Andrew Ng

deeplearning.ai

Recurrent Neural Networks

LSTM (long short term memory) unit

GRU and LSTM

GRU

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

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

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

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

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

LSTM

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

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

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

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

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

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

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

LSTM 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]

LSTM in pictures

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

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

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

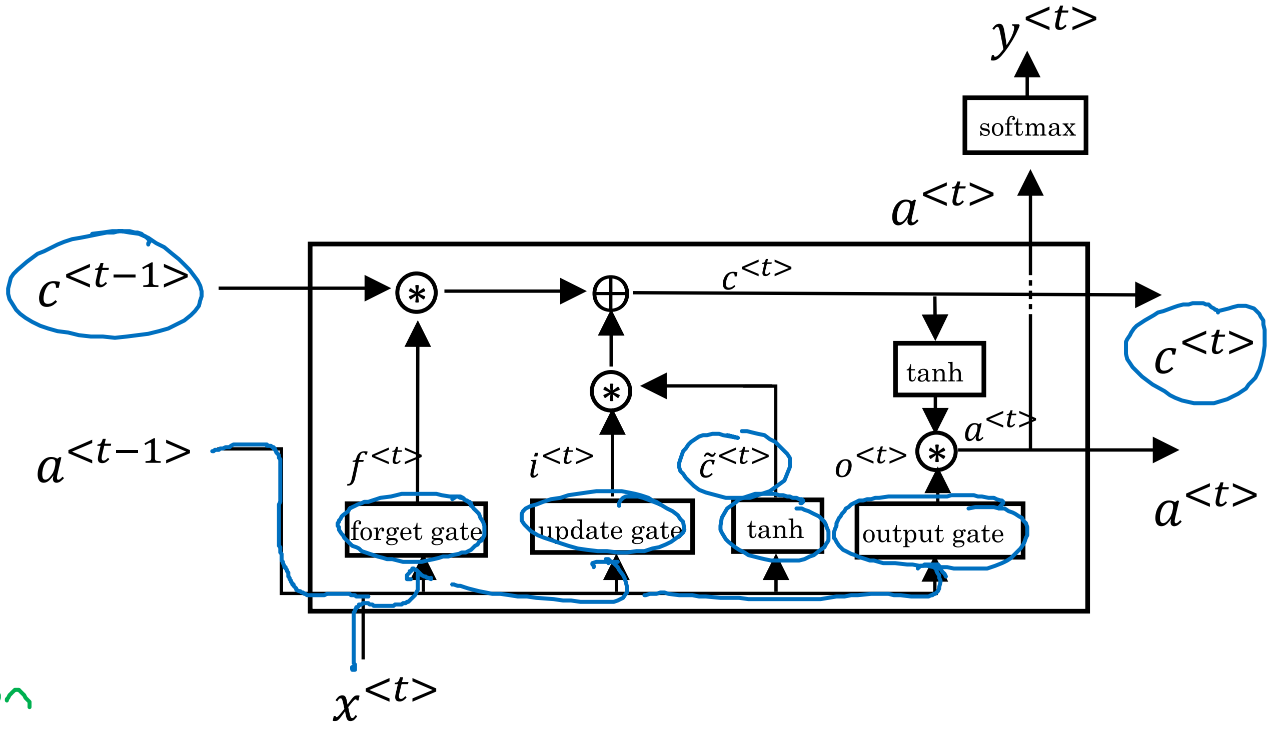
$$\Gamma_o = \sigma(W_o[\, a^{<t-1>}, x^{<t>}] + b_o)$$

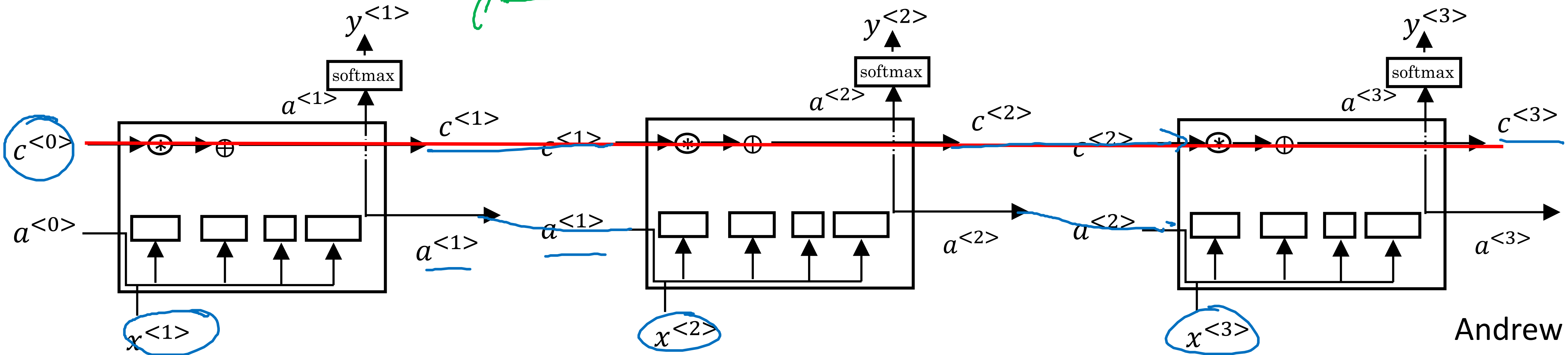
$$c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + \Gamma_f * c^{<t-1>}$$

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

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

peephole connection





deeplearning.ai

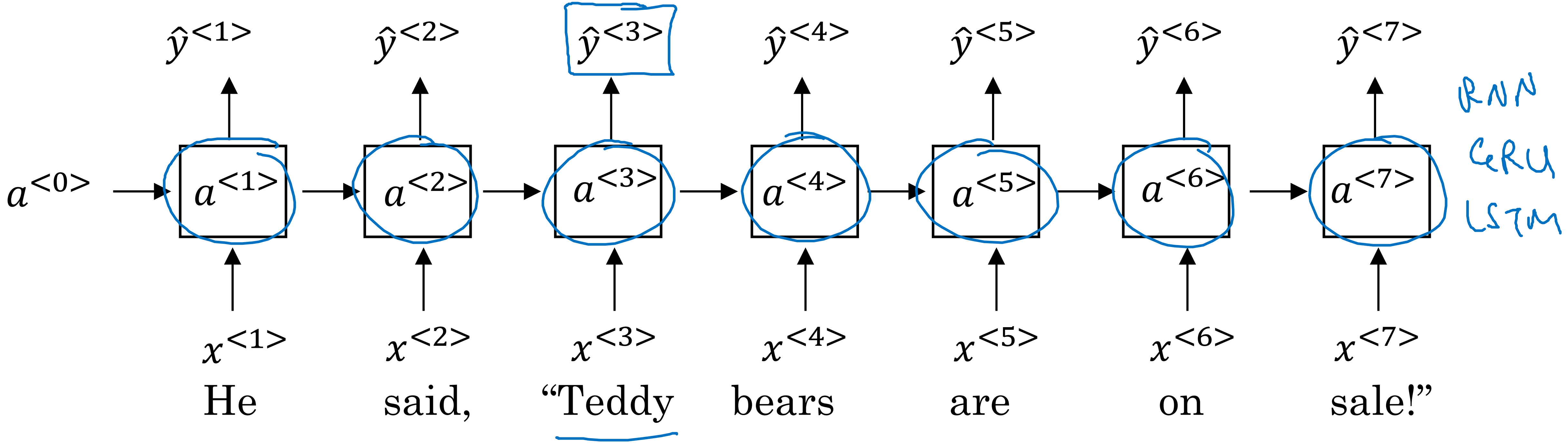
Recurrent Neural Networks

Bidirectional RNN

Getting information from the future

He said, "Teddy bears are on sale!"

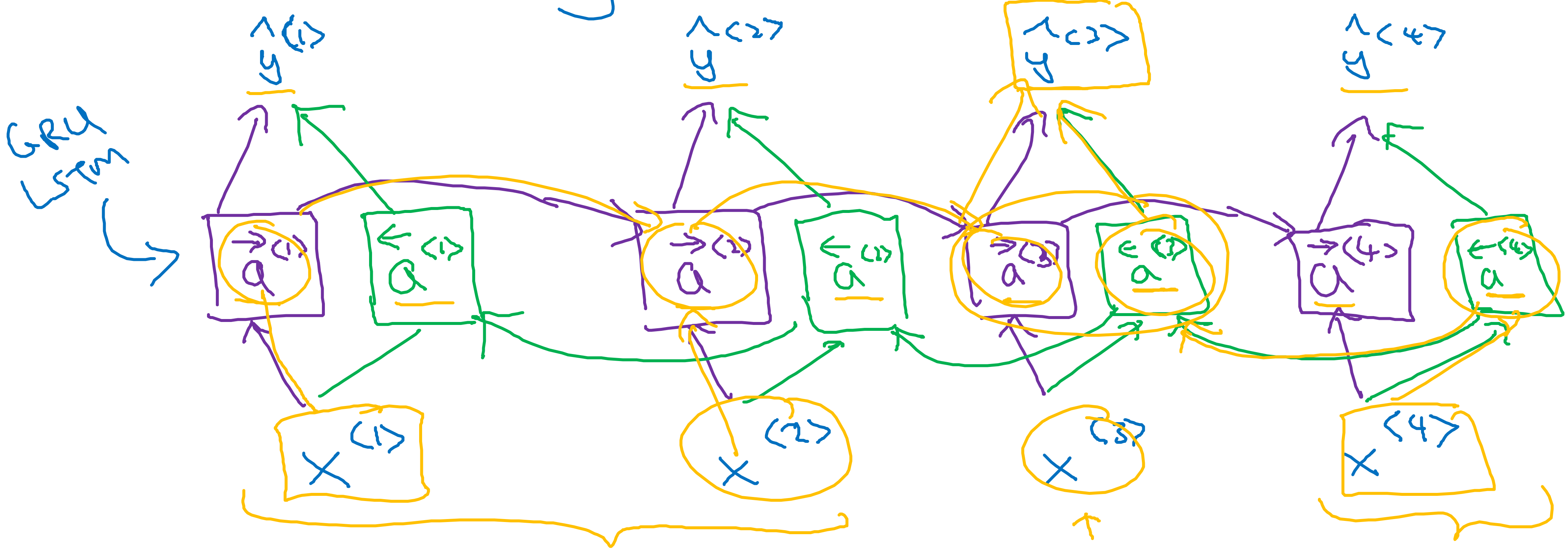
He said, "Teddy Roosevelt was a great President!"



Andrew Ng

Bidirectional RNN (BRNN)

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



Acyclic graph

He said, "Teddy Roosevelt ...."

BRNN w/ LSTM

Andrew Ng

Sequence to sequence models

Basic models

deeplearning.ai

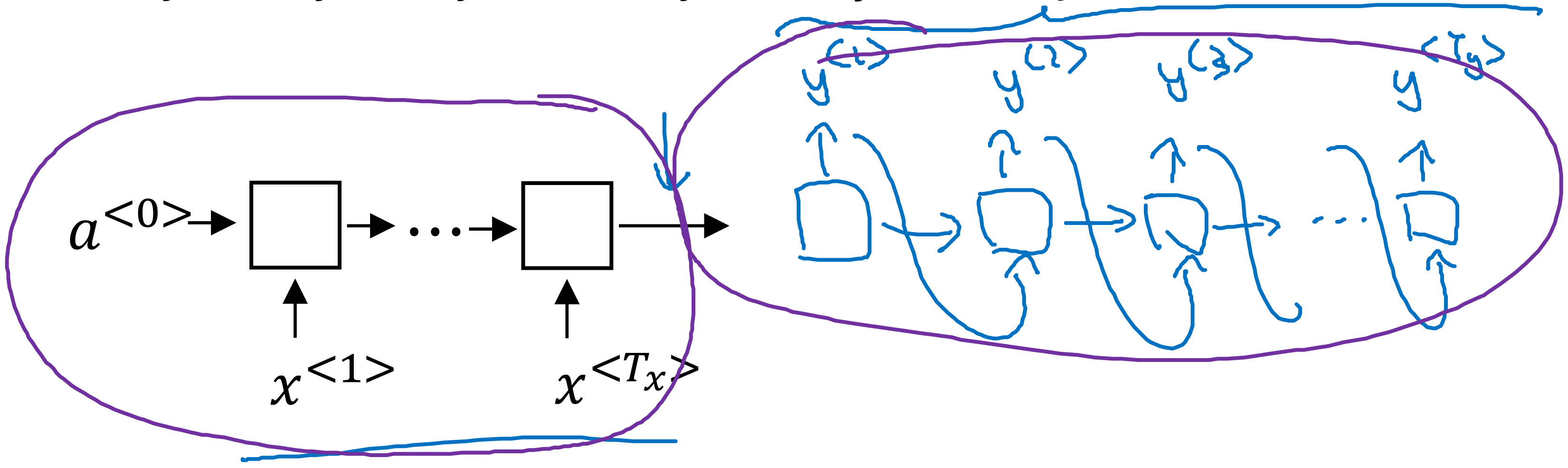
Sequence to sequence model

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

Jane visite l'Afrique en septembre

$\longrightarrow$ Jane is visiting Africa in September.

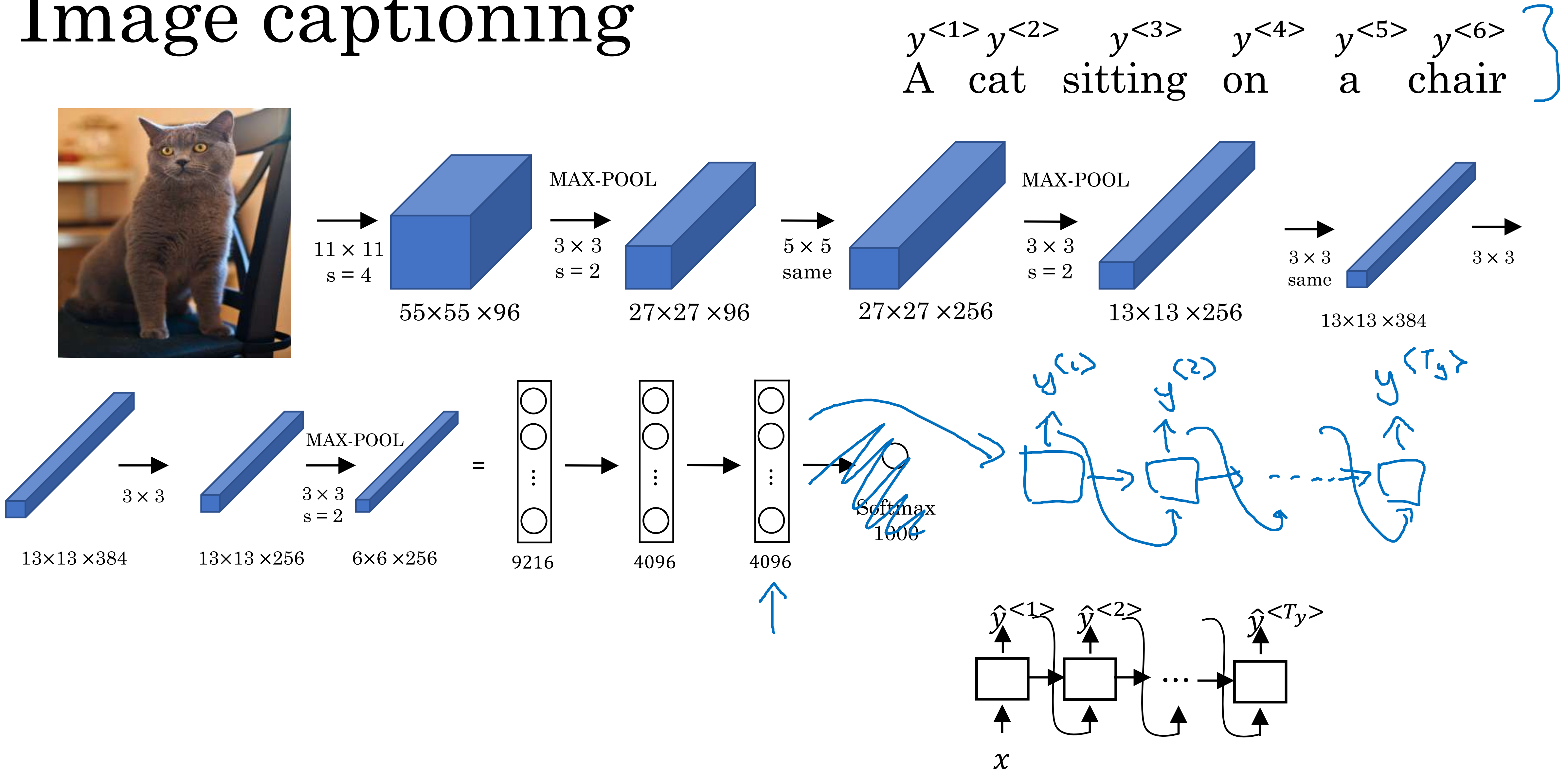
$y^{<1>}$ $y^{<2>}$ $y^{<3>}$ $y^{<4>}$ $y^{<5>}$ $y^{<6>}$



[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]

Image captioning



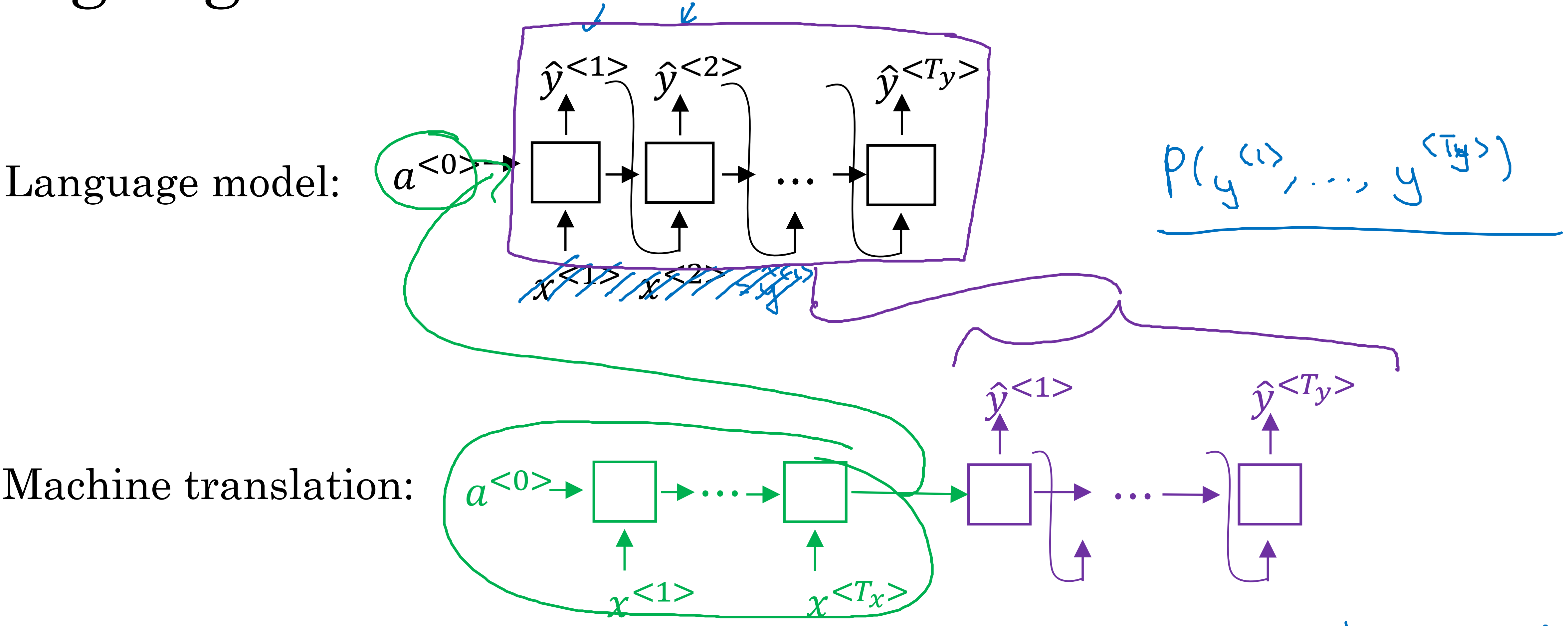
[Mao et. al., 2014. Deep captioning with multimodal recurrent neural networks]
[Vinyals et. al., 2014. Show and tell: Neural image caption generator]
[Karpathy and Li, 2015. Deep visual-semantic alignments for generating image descriptions]

deeplearning.ai

Sequence to sequence models

Picking the most likely sentence

Machine translation as building a conditional language model



"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

$$\underbrace{P(\overbrace{y^{<1>}, \dots, y^{<T_y>}}^{\text{English}} \mid \overset{\text{French}}{x})}$$

→ Jane is visiting Africa in September.

→ Jane is going to be visiting Africa in September.

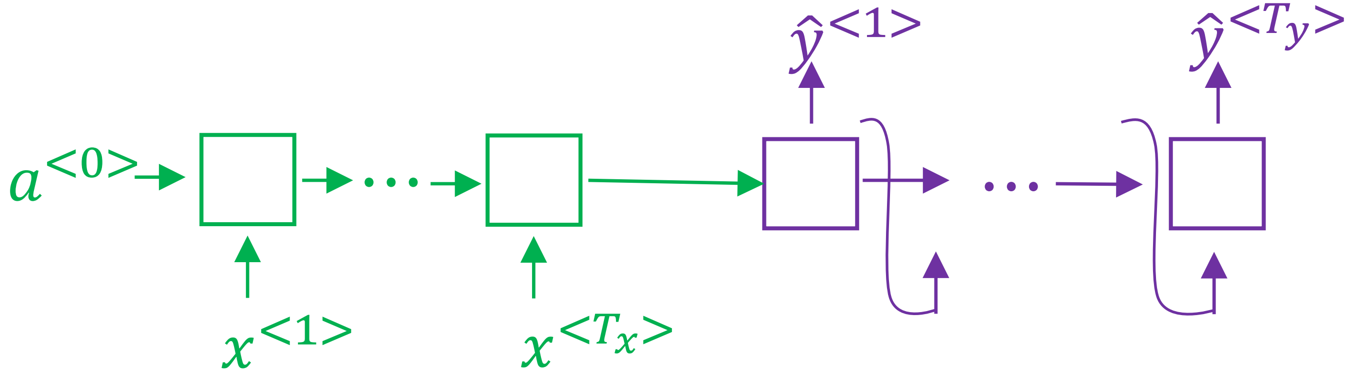
→ In September, Jane will visit Africa.

→ Her African friend welcomed Jane in September.

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

Why not a greedy search?

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



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

10,000

10

$10{,}000^{10}$

$P(y|x)$

→ Jane is visiting Africa in September.

→ Jane is going to be visiting Africa in September.

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

deeplearning.ai

Sequence to sequence models

Beam search

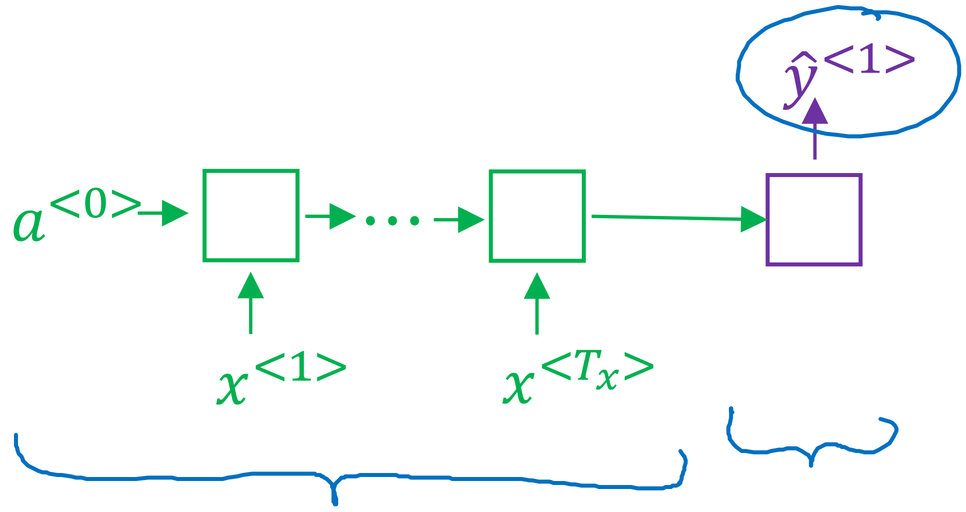
Beam search algorithm

$\underline{B = 3}$ (beam width)

Step 1

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

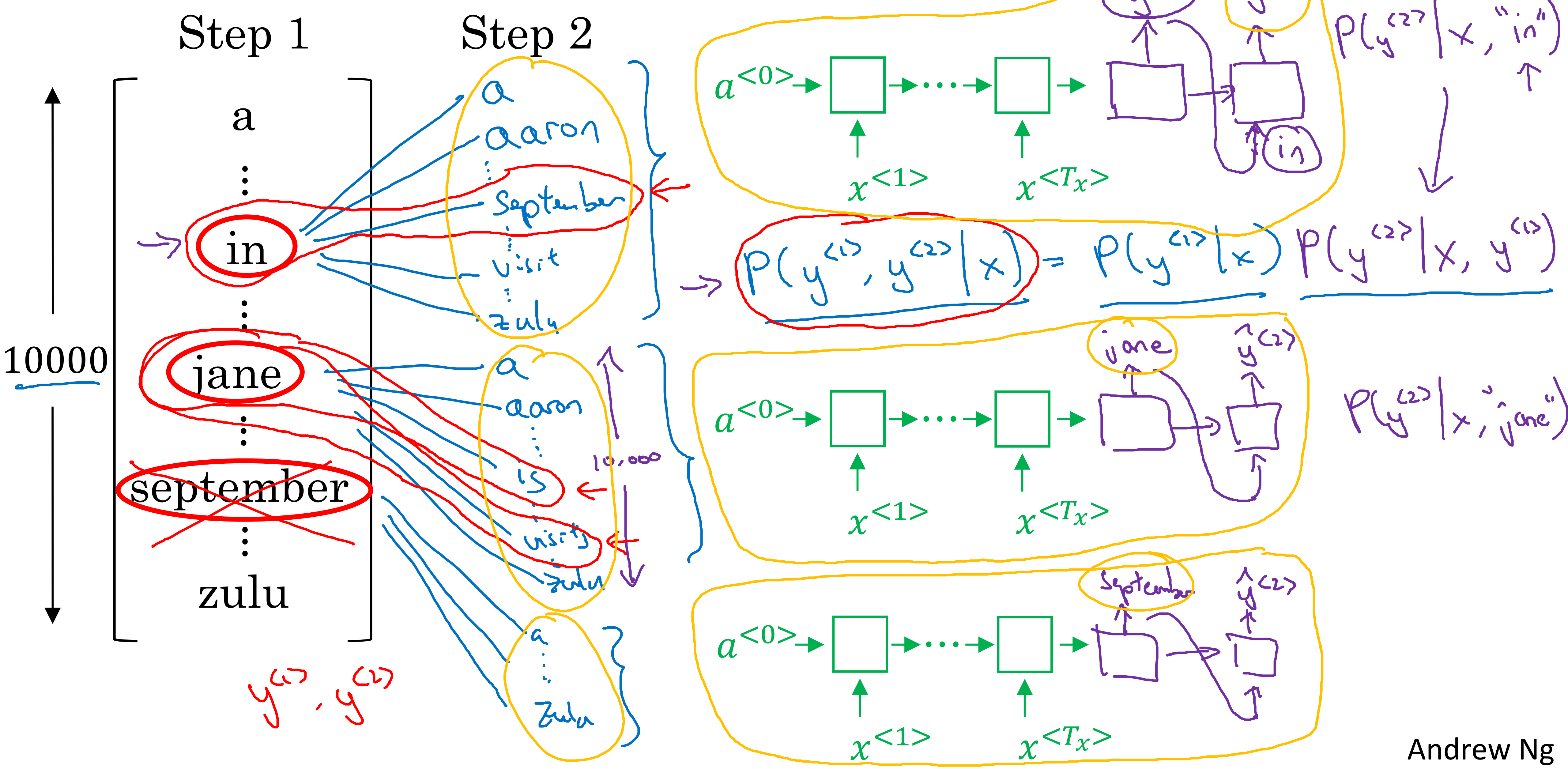
$\rightarrow \underline{P(y^{<1>} \mid x)}$



Andrew Ng

Beam search algorithm

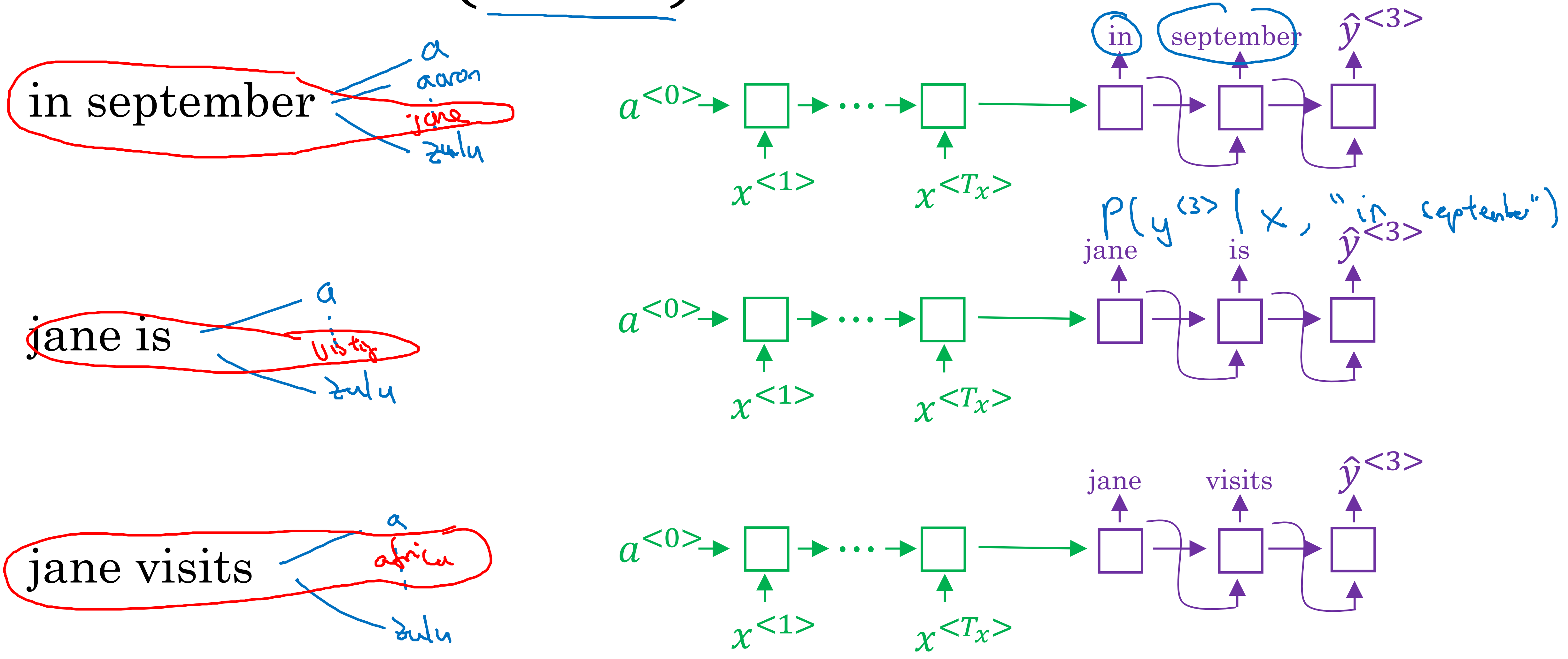
$(B=3)$



Andrew Ng

Beam search ($B = 3$)

$B = 1 \rightsquigarrow$ greedy search



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

jane visits africa in september. <EOS>

deeplearning.ai

Sequence to sequence models

Refinements to beam search

Length normalization

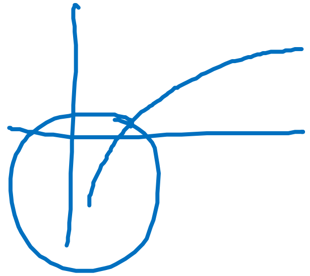
$P(y^{<1>} \ldots y^{<T_y>}|x) = P(y^{<1>}|x)\, P(y^{<2>}|x, y^{<1>}) \ldots$

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

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

log

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



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

$P(y|x) \leftarrow$

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

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

$\alpha = 0.7$

$\alpha = 1$

$\alpha = 0$

Andrew Ng

Beam search discussion

large B: better result, slower
small B: worse result, faster

Beam width B?

$1 \rightarrow 3 \rightarrow 10$, 100, $1000, \rightarrow 3000$

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 $\arg\max_{y} P(y|x)$.

deeplearning.ai

Sequence to sequence models

Error analysis on beam search

Example

→ RNN

→ Beam Search

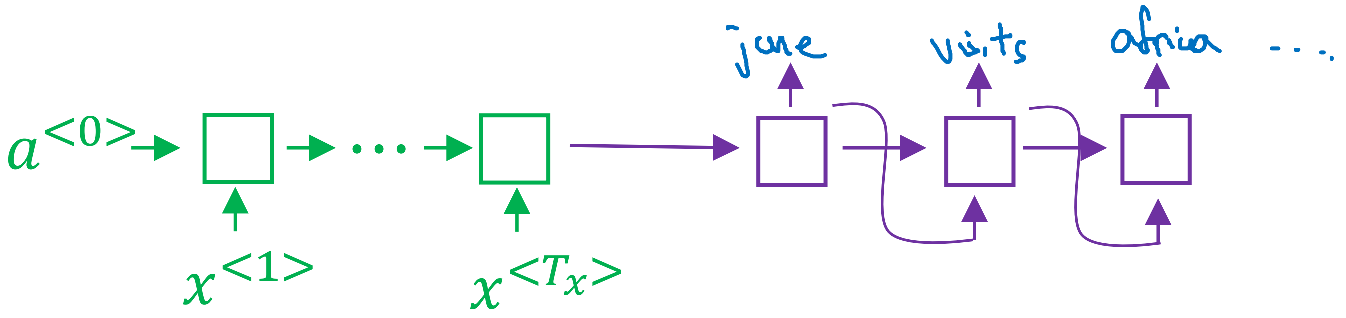
$\boxed{BT}$

Jane visite l'Afrique en septembre.

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

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

RNN computes $P(y^*|x) \underset{\leq}{>} P(\hat{y}|x)$



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. $(\hat{y})$

$\arg\max_y P(y | x)$

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

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

Conclusion: Beam search is at fault.

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

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.

Andrew Ng

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

deeplearning.ai

Sequence to sequence models

Bleu score (optional)

Evaluating machine translation

Bleu

bilingual evaluation understudy

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:

[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]

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. ($\hat{y}$)

$p_1, p_2, = 1.0$

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

unigram

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

n-gram

$$P_n = \frac{\sum_{n\text{-grams} \in \hat{y}} Count_{clip}(n\text{-gram})}{\sum_{n\text{-grams} \in \hat{y}} Count(n\text{-gram})}$$

[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

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

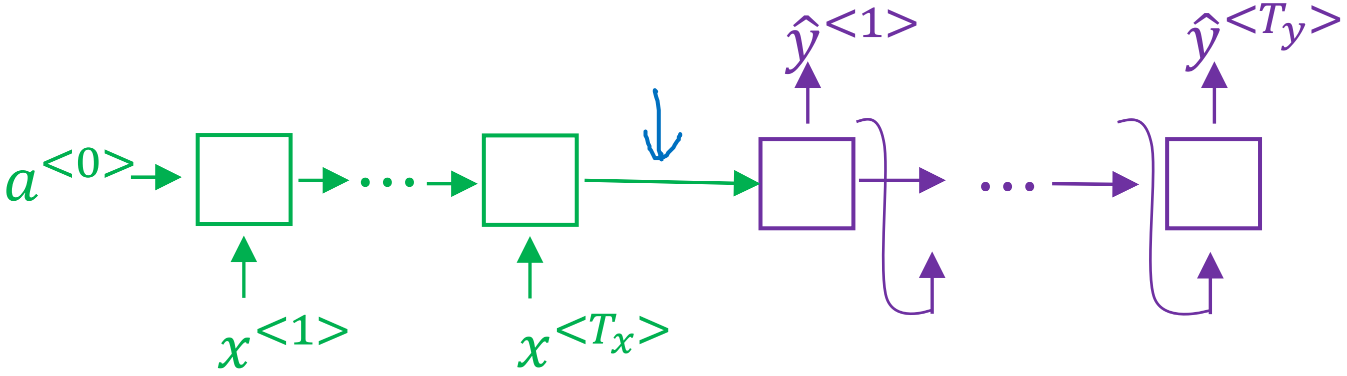
[Papineni et. al., 2002. Bleu: A method for automatic evaluation of machine translation]

deeplearning.ai

Sequence to sequence models

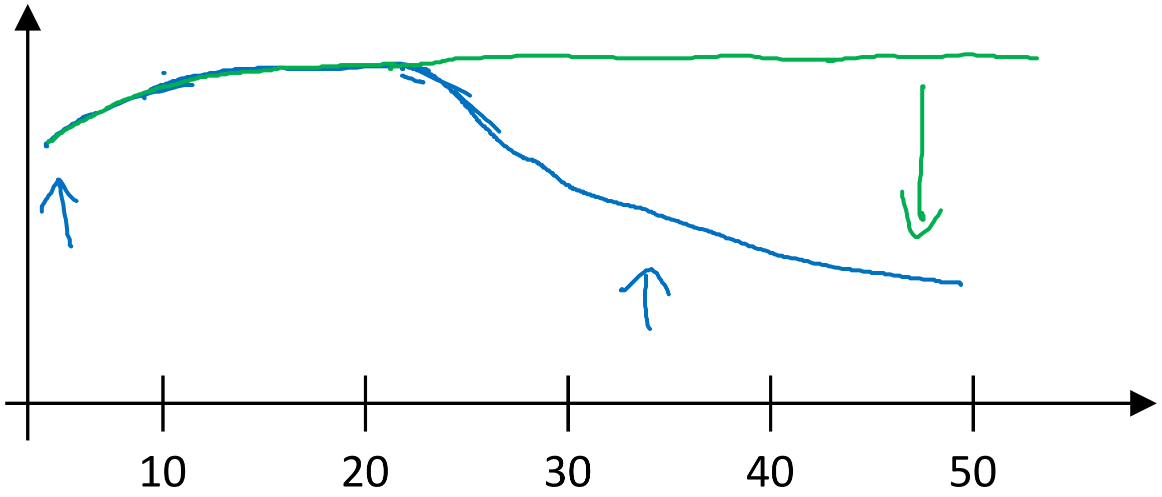
Attention model intuition

The problem of long sequences



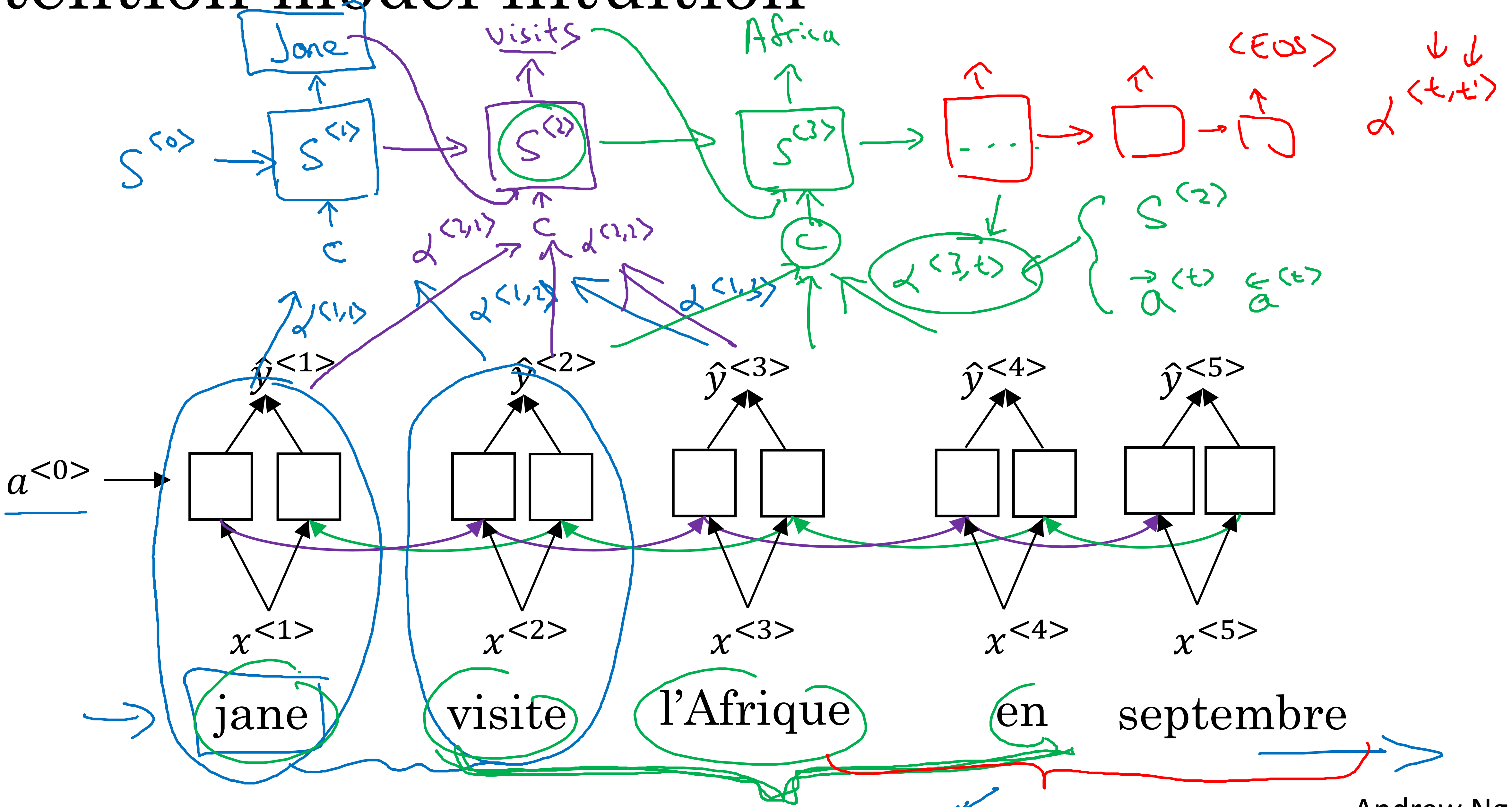
Jane s'est rendue en Afrique en septembre dernier, a apprécié la culture et a rencontré beaucoup de gens merveilleux; elle est revenue en parlant comment son voyage était merveilleux, et elle me tente d'y aller aussi.

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.



Andrew Ng

Attention model intuition



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

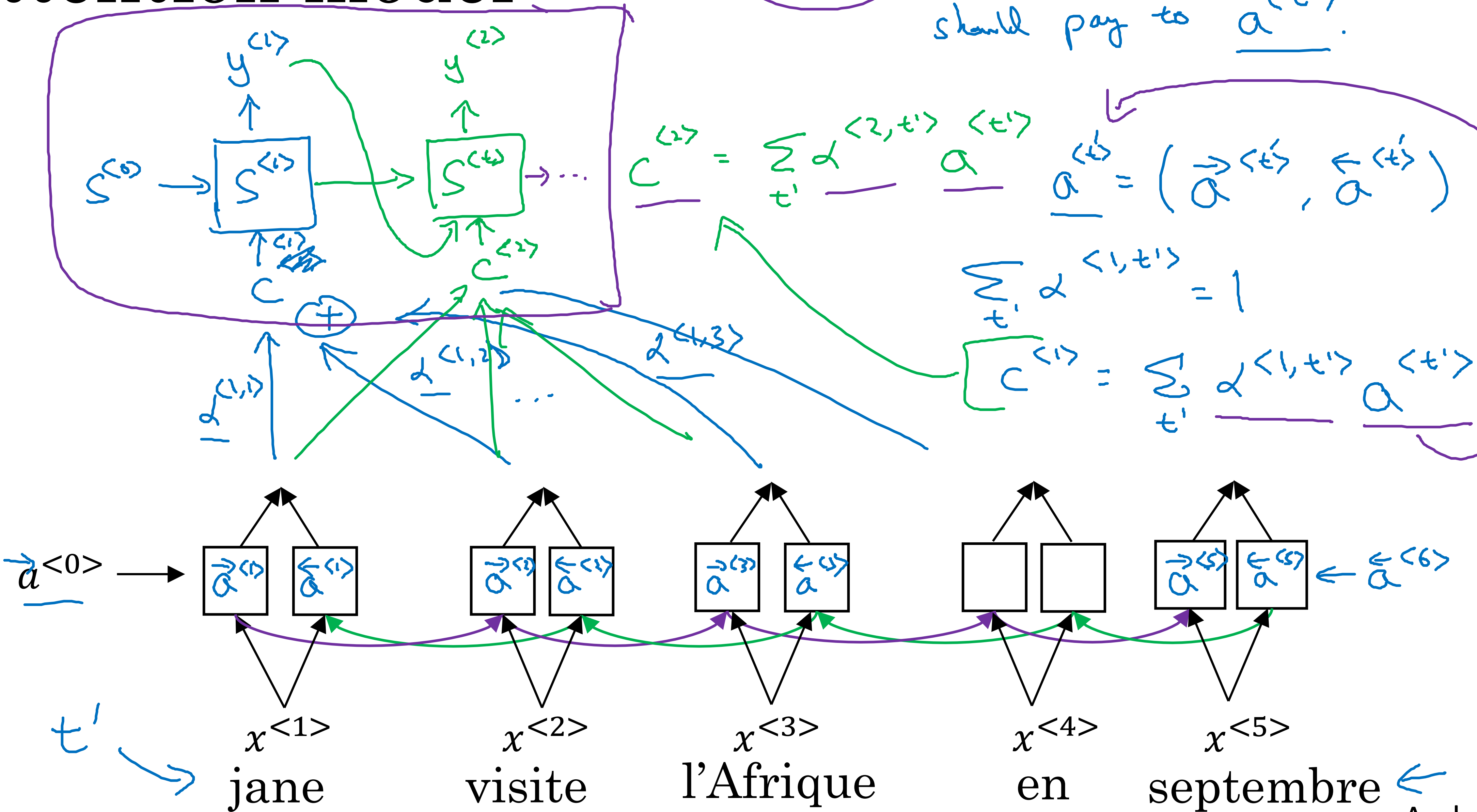
Sequence to sequence models

Attention model

deeplearning.ai

Attention model

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



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

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

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

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

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

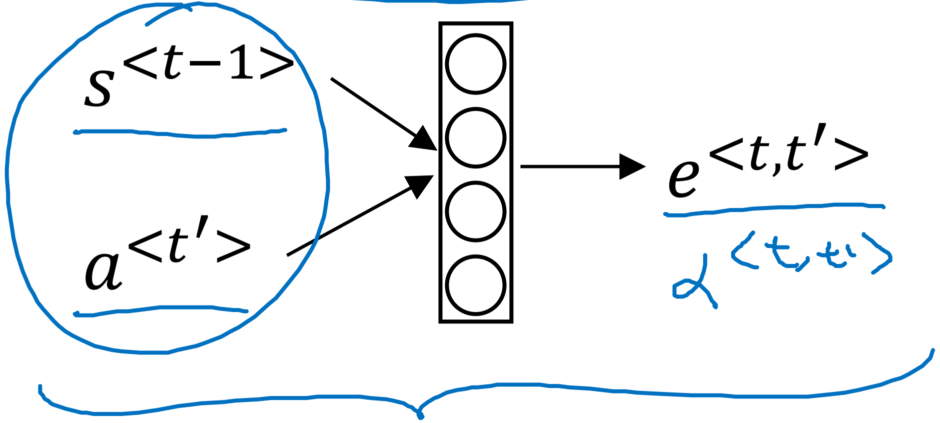
Andrew Ng

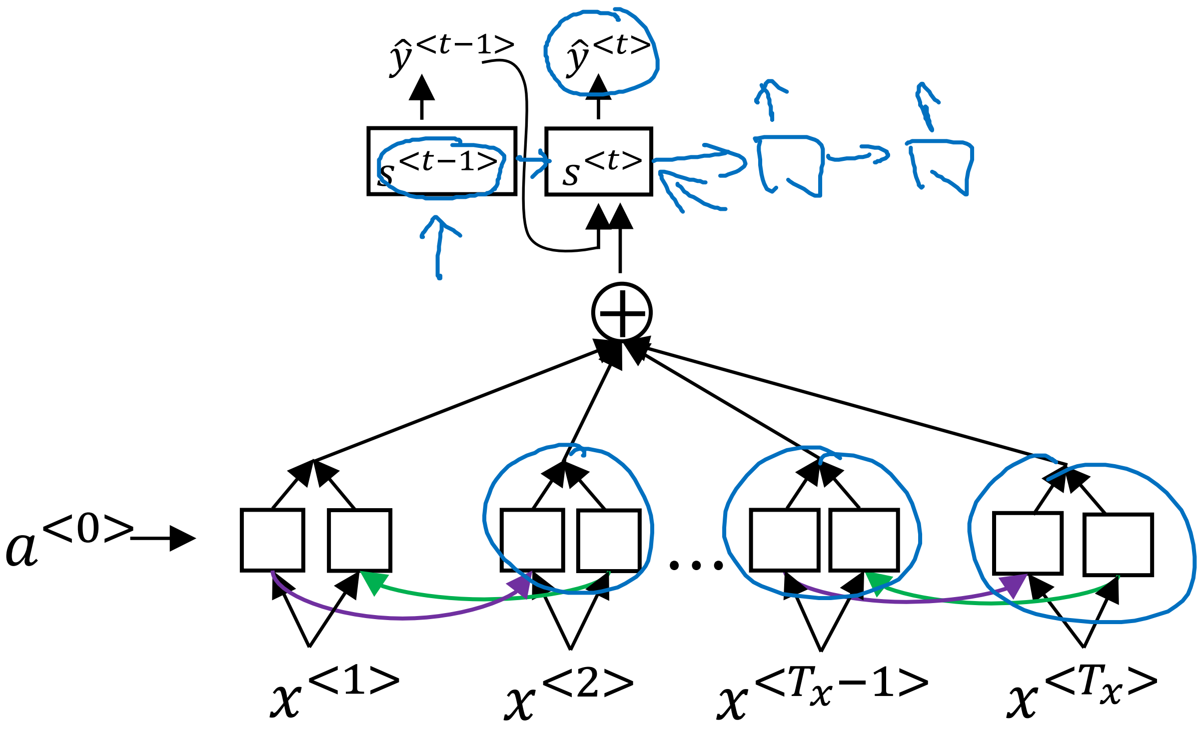
Computing attention $\alpha^{<t,t'>}$

T_x T_y

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

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





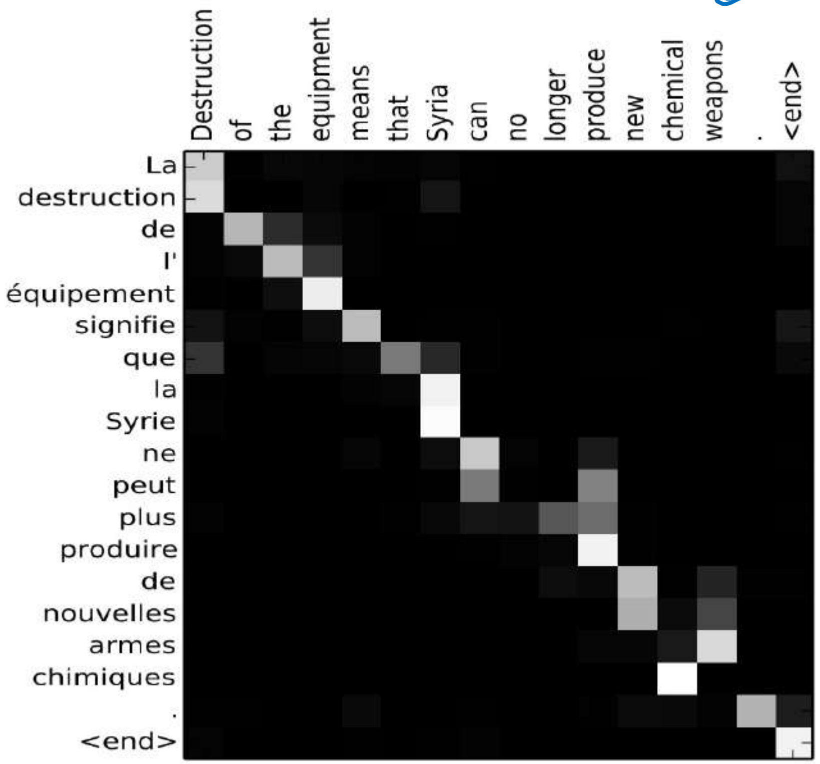
[Bahdanau et. al., 2014. Neural machine translation by jointly learning to align and translate]
[Xu et. al., 2015. Show, attend and tell: Neural image caption generation with visual attention]

Attention examples

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

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

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



deeplearning.ai

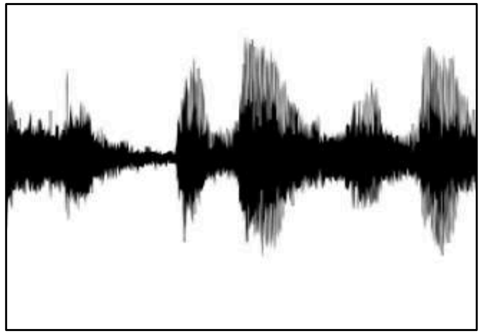
Audio data

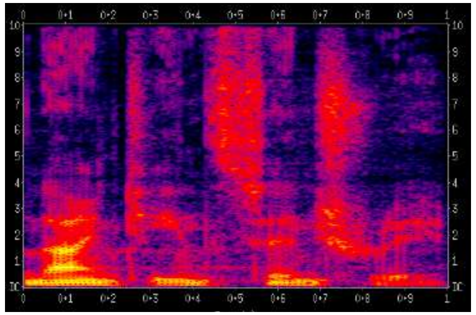
Speech recognition

Speech recognition problem

$x \longrightarrow y$

audio clip — transcript





"the quick brown fox"

phonemes: de kwik braun

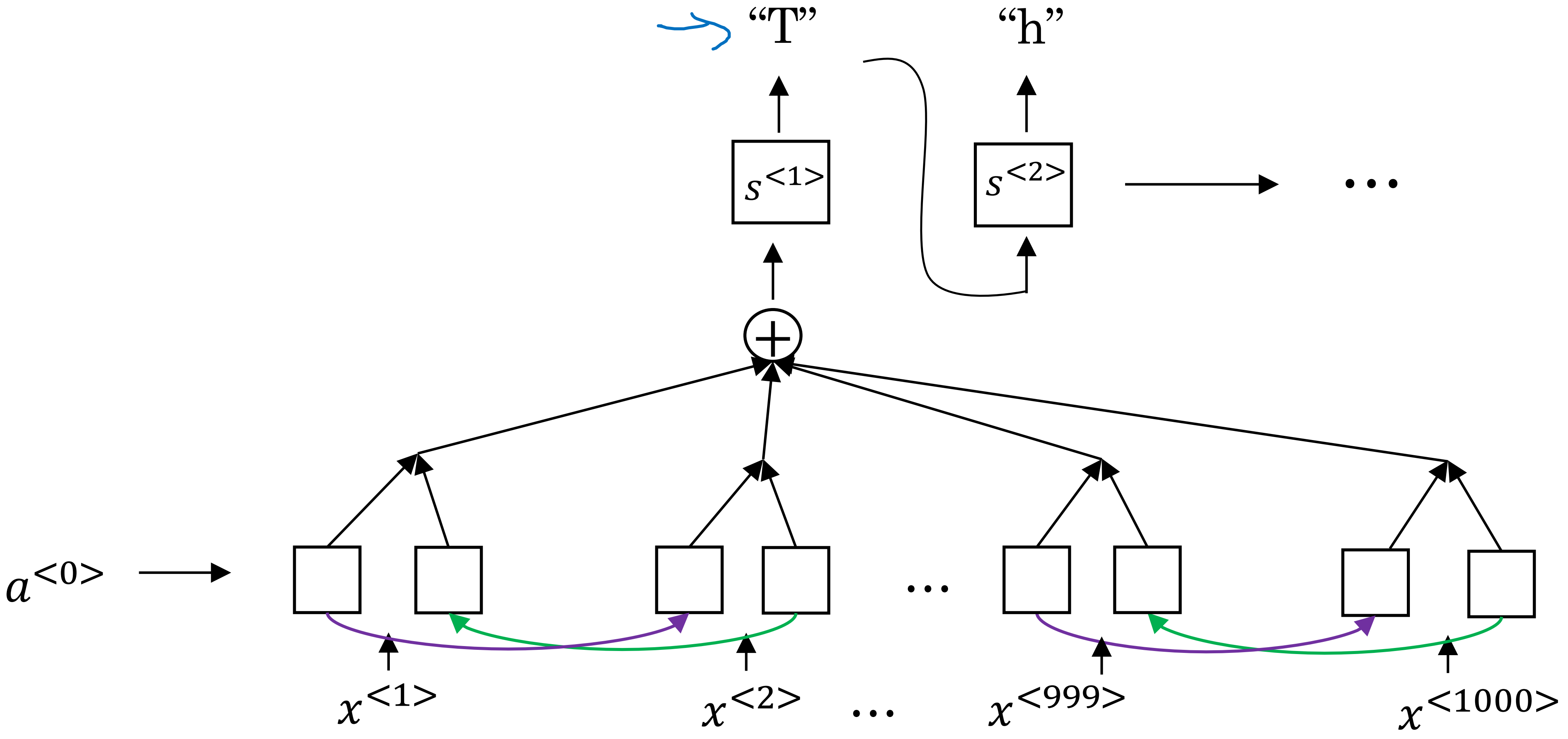
300h

3000h

100,000h

Andrew Ng

Attention model for speech recognition

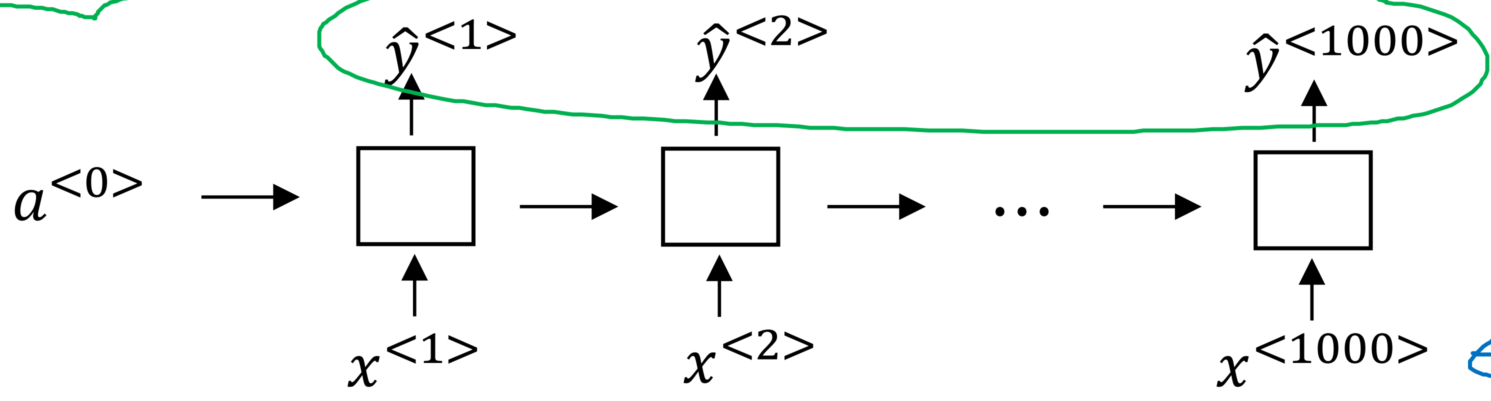


Andrew Ng

CTC cost for speech recognition

(Connectionist temporal classification)

"the quick brown fox" 19 characters



ttt _ h _ eee _ _ _ ␣ _ _ _ qqq _ _   the␣q

"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]

deeplearning.ai

Audio data

Trigger word detection

What is trigger word detection?

Amazon Echo
(Alexa)

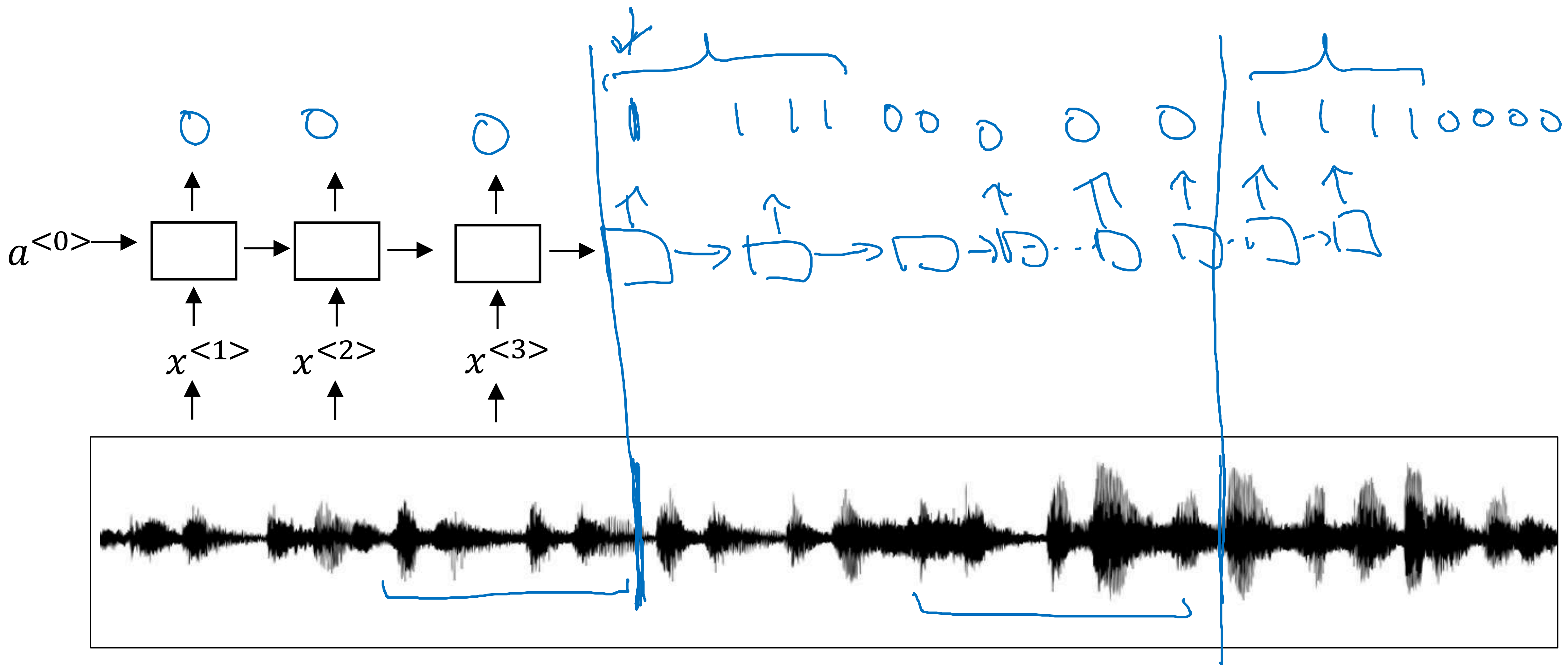
Baidu DuerOS
(xiaodunihao)

Apple Siri
(Hey Siri)

Google Home
(Okay Google)

Andrew Ng

Trigger word detection algorithm



Andrew Ng