Copyright Notice

These slides are distributed under the Creative Commons License.

DeepLearning.AI makes these slides available for educational purposes. You may not use or distribute these slides for commercial purposes. You may make copies of these slides and use or distribute them for educational purposes as long as you cite DeepLearning.AI as the source of the slides.

For the rest of the details of the license, see https://creativecommons.org/licenses/by-sa/2.0/legalcode
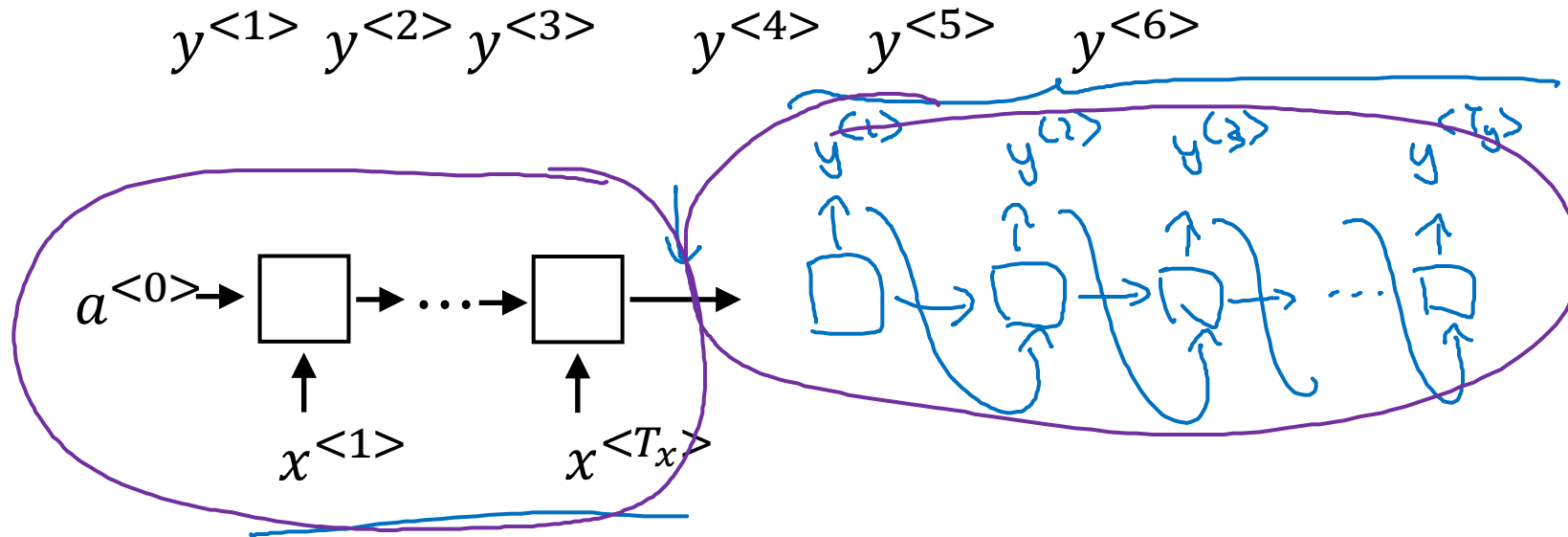
Sequence to sequence models

Basic models

deeplearning.ai

Sequence to sequence model

$x^{<1>} \quad x^{<2>} \quad x^{<3>} \quad x^{<4>} \quad x^{<5>}$
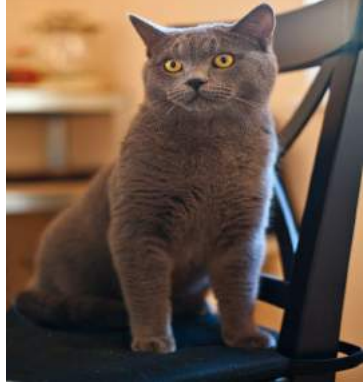
Jane visite l'Afrique en septembre

$\longrightarrow$ Jane is visiting Africa in September.

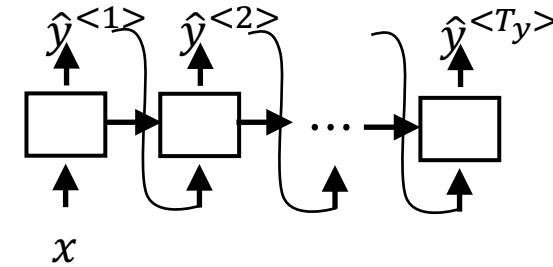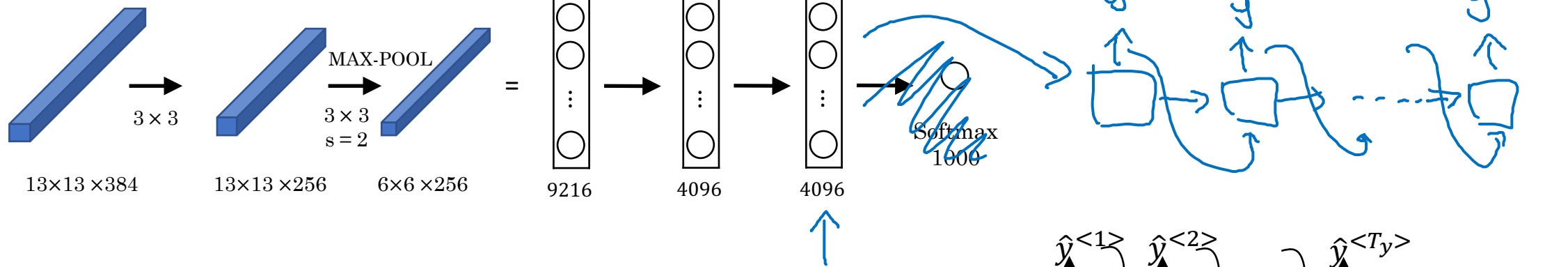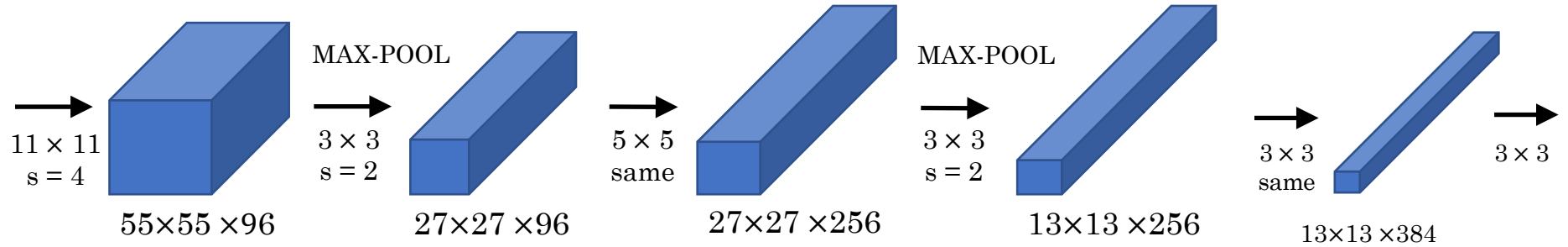$y^{<1>} \quad y^{<2>} \quad y^{<3>} \qquad y^{<4>} \quad y^{<5>} \qquad y^{<6>}$

[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

55×55 ×96 27×27 ×96 27×27 ×256 13×13 ×256 13×13 ×384

11×11 s = 4 3×3 s = 2   MAX-POOL 5×5 same 3×3 s = 2   MAX-POOL 3×3 same 3×3

13×13 ×384 13×13 ×256 6×6 ×256 9216 4096 4096

3×3 3×3 s = 2   MAX-POOL   Softmax 1000

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

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

x

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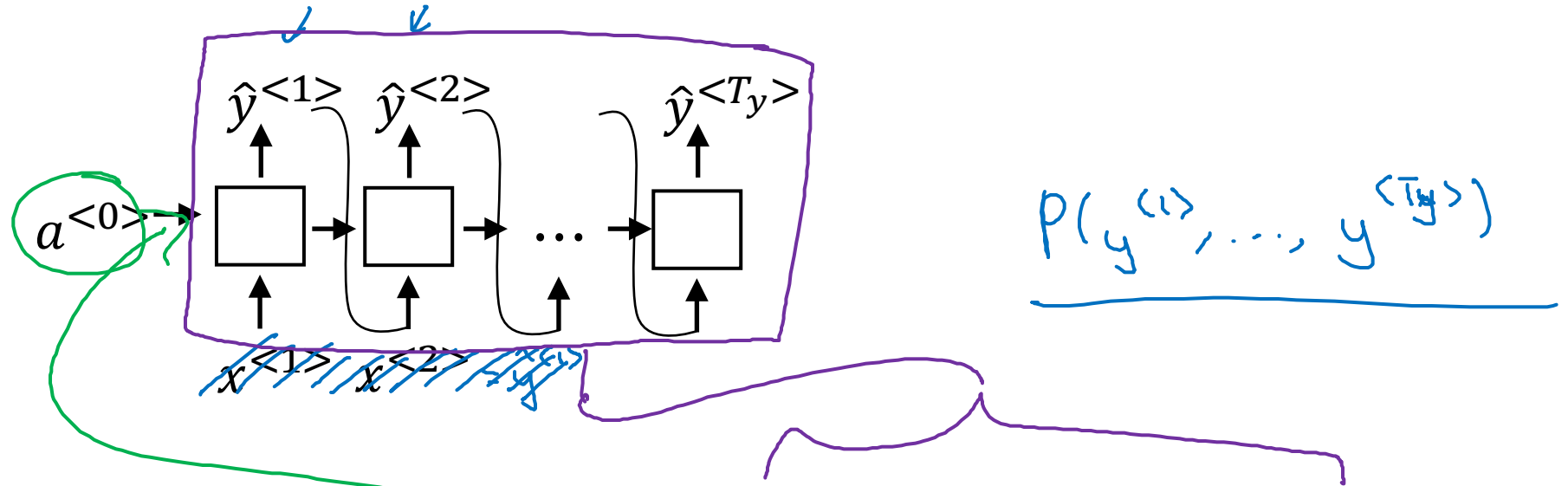[Karpathy and Li, 2015. Deep visual-semantic alignments for generating image descriptions]

Andrew Ng

Machine translation as building a conditional language model

Language model:

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

$$a^{<0>}$$

$$x^{<1>}, x^{<2>}, x^{<T_y>}$$

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

Machine translation:

$$a^{<0>} \quad x^{<1>} \quad x^{<T_x>}$$

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

output english sentence condition on the input french sentence

"Conditial language model"

$$P(y^{(1)}, \ldots, y^{(T_y)} \mid x^{(1)}, \ldots, x^{(T_x)})$$

Andrew Ng
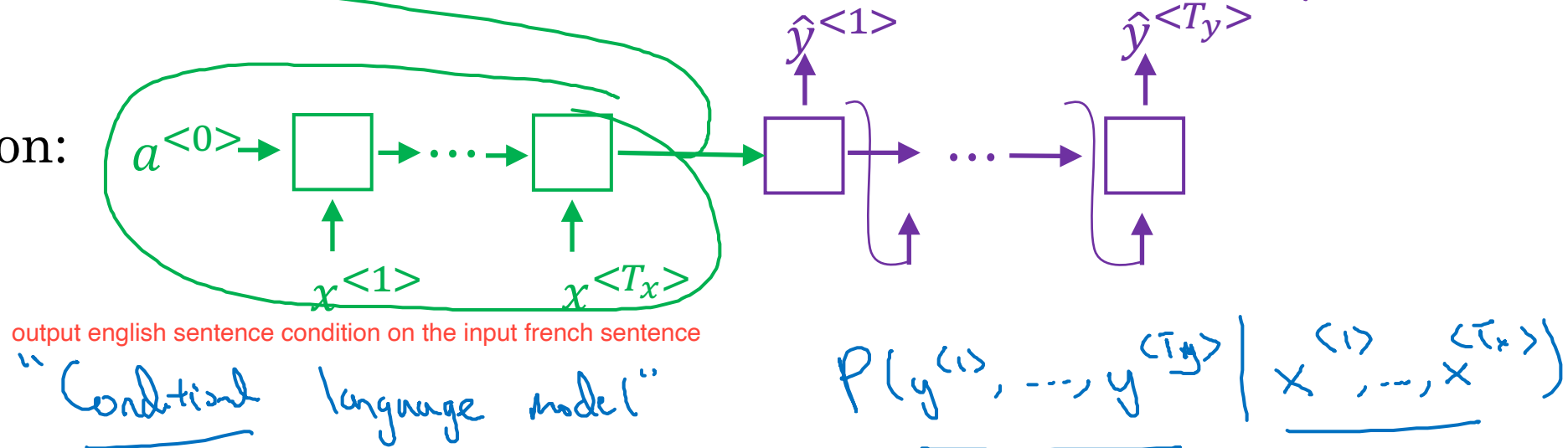
Finding the most likely translation

English French

Jane visite l'Afrique en septembre.
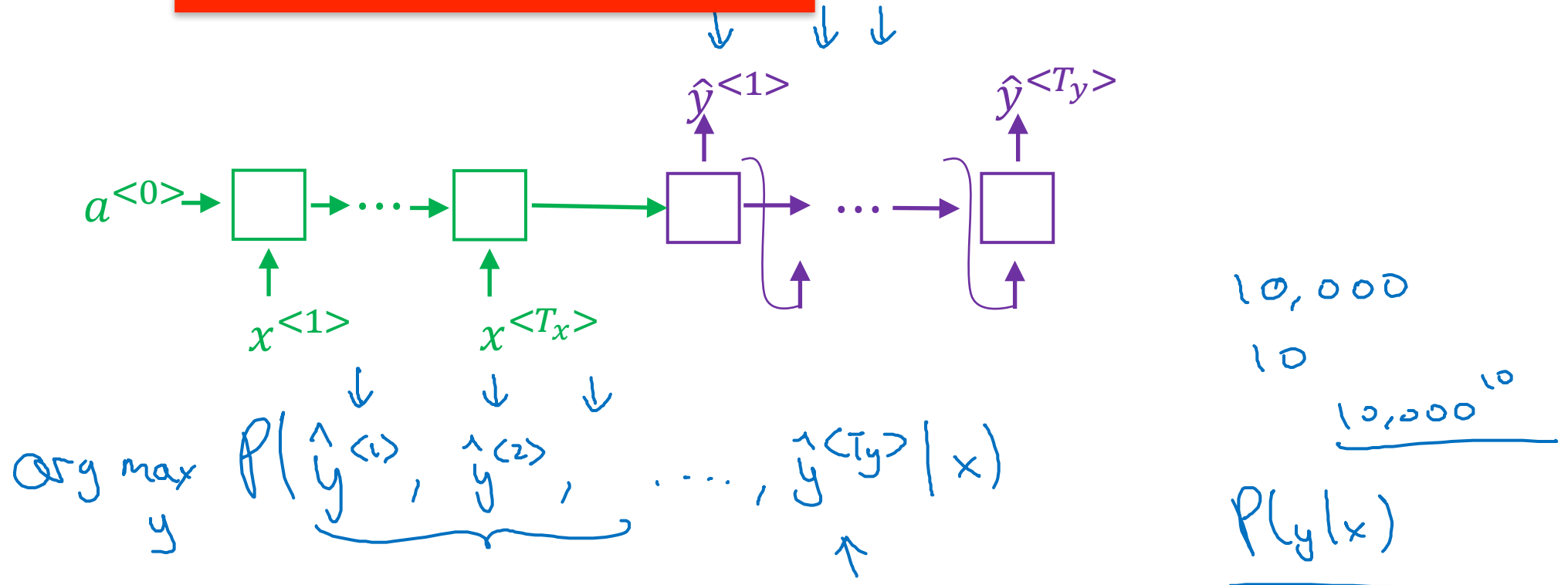
$$P(y^{<1>}, \ldots, y^{<T_y>} \mid x)$$

→ 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.

maximize the conditional prob

$$\underset{y^{<1>}, \ldots, y^{<T_y>}}{\arg \max} P(y^{<1>}, \ldots, y^{<T_y>} \mid x)$$

Andrew Ng

Why not a greedy search?

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

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

$a^{<0>}$

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

$10,000$

10

$10,000^{10}$

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

$P(y|x)$

→ Jane is visiting Africa in September.   picking 1 words at a time, that maximize the prob

→ Jane is going to be visiting Africa in September.

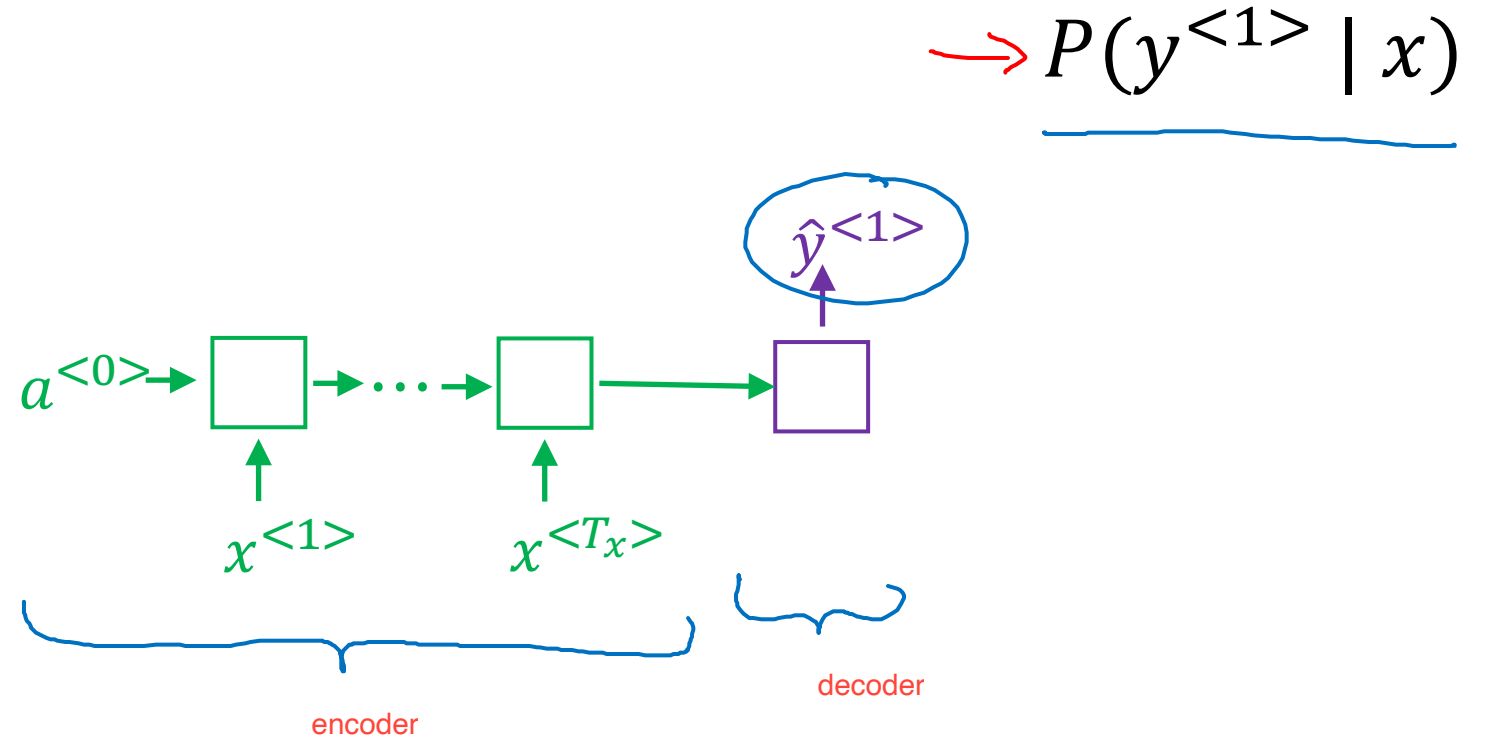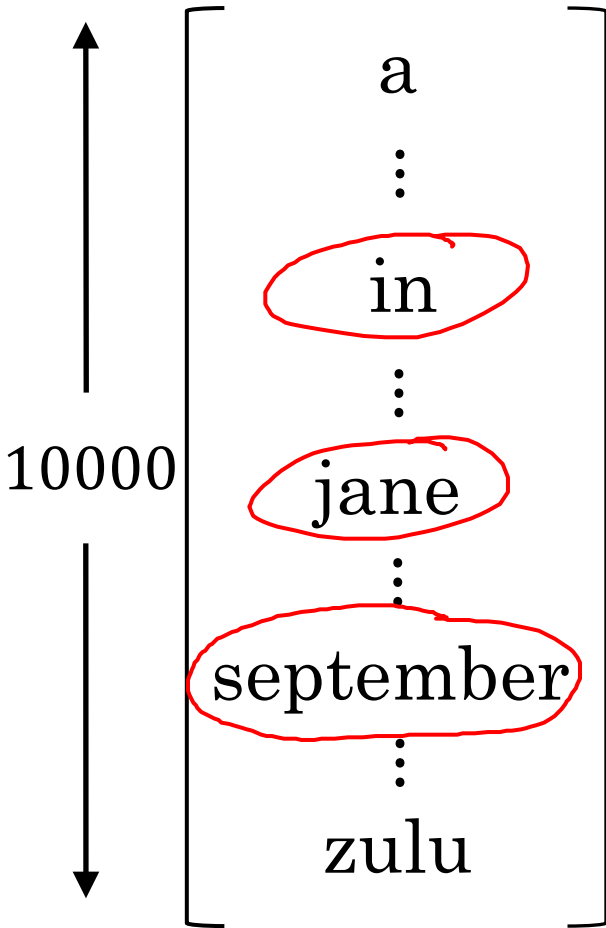$P(\text{Jane is going}|x) > P(\text{Jane is visit}|x)$

Andrew Ng

Sequence to sequence models

Beam search

deeplearning.ai

Beam search algorithm

$B = 3$ (beam width)

Step 1

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

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

10000

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

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

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

encoder

decoder

Andrew Ng

Beam search algorithm

$(B = 3)$

Step 1   Step 2

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

10000

$y^{(1)}, y^{(2)}$

a
aaron
$\vdots$
September
visit
$\vdots$
zulu

a
aaron
$\vdots$
is
$\vdots$
visits
$\vdots$
zulu

10,000

a
$\vdots$
zulu

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

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

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

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

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

September $\hat{y}^{<2>}$
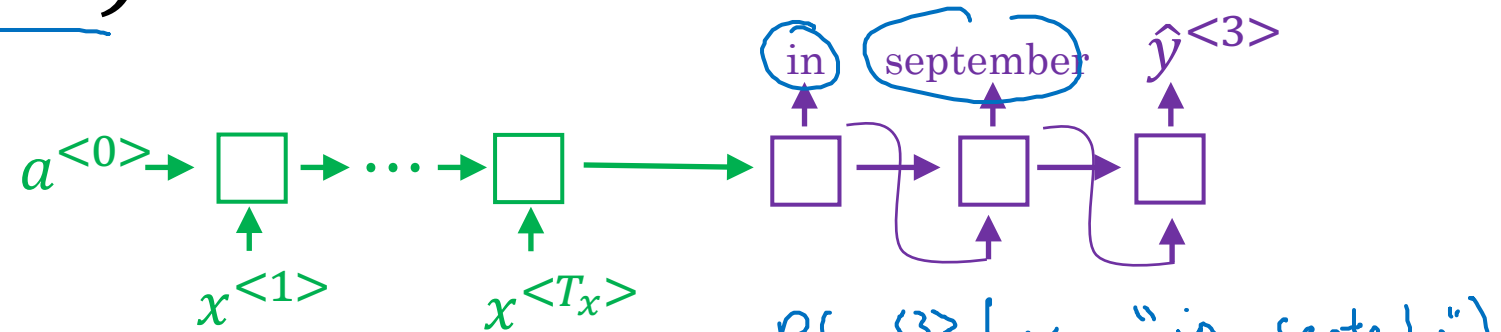
Andrew Ng
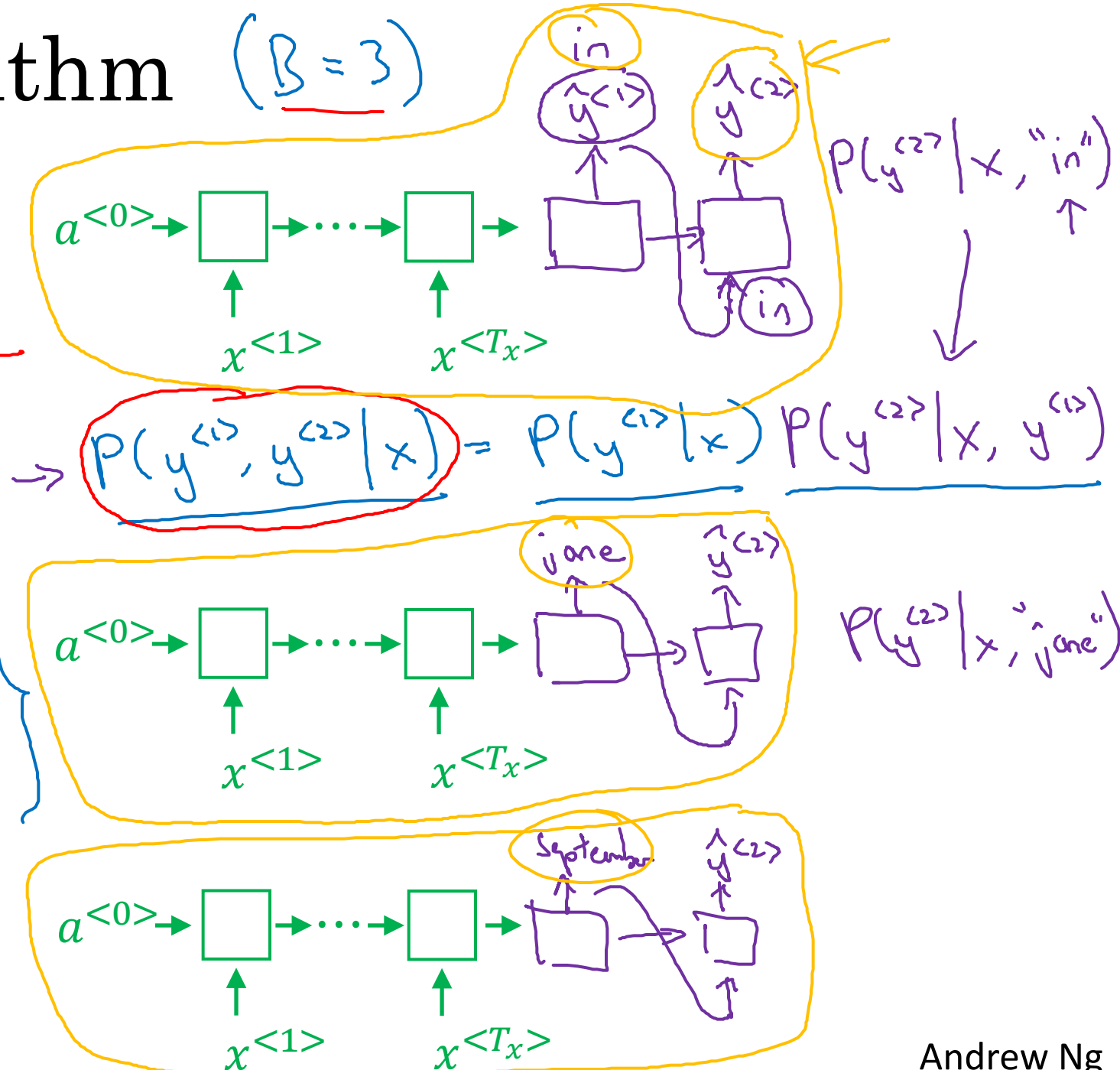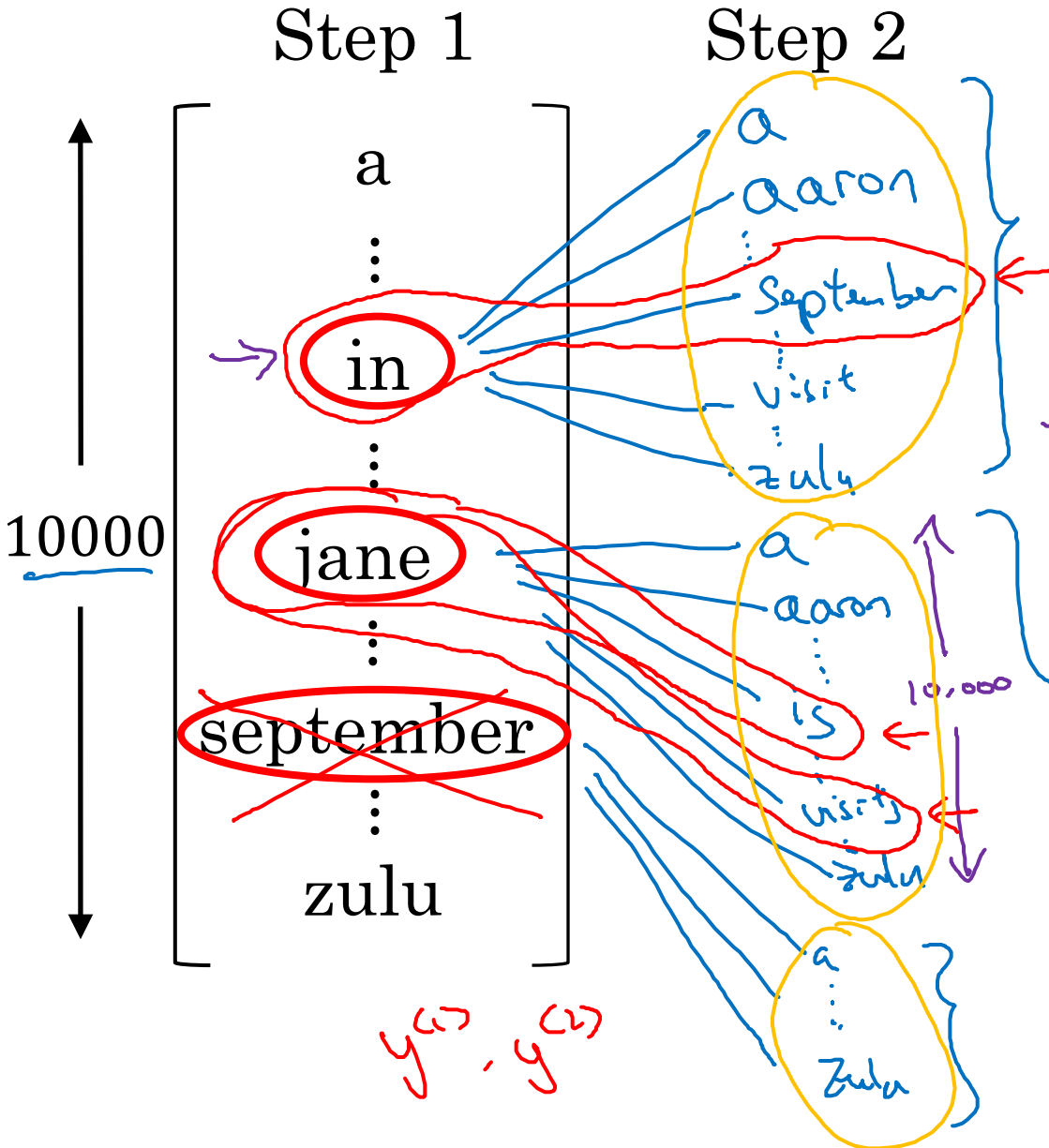
Beam search ($B = 3$)

$B = 1 \rightsquigarrow$ greedy search

in september
- a
- aaron
- jane
- zulu

jane is
- a
- visits
- zulu
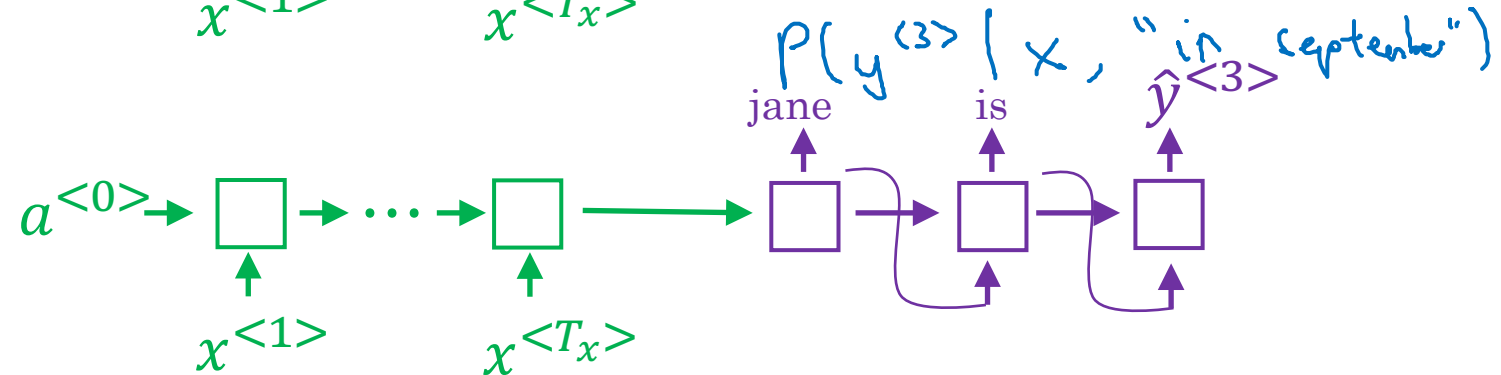
jane visits
- a
- africa
- zulu

in   september $\hat{y}^{<3>}$

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

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

$P(y^{<3>} | x, \text{"in september"})$

jane   is $\hat{y}^{<3>}$

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

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

jane   visits $\hat{y}^{<3>}$

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

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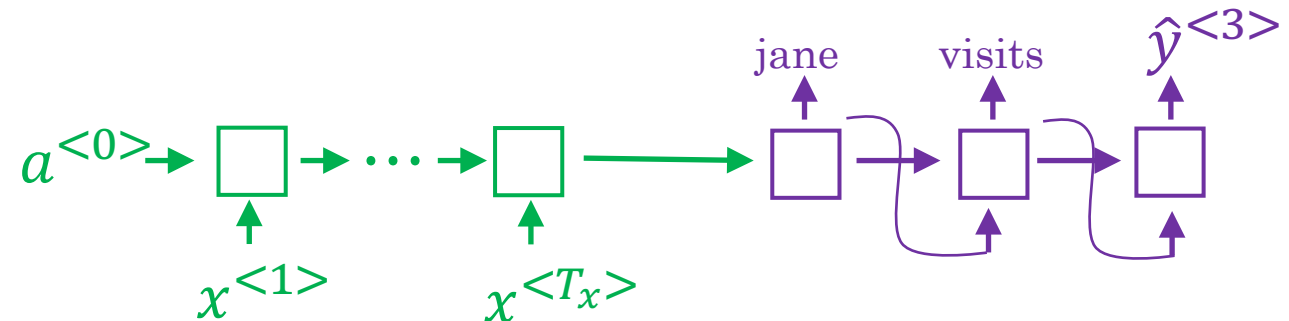
$P(y^{<1>}, y^{<2>} | x)$

jane visits africa in september. <EOS>

Andrew Ng

deeplearning.ai

Sequence to sequence models

Refinements to beam search
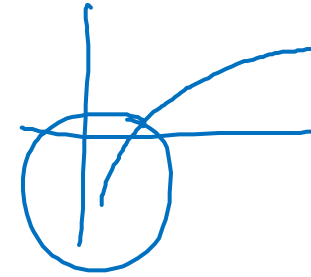
Length normalization

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

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

$\log$

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

$P(y|x) \leftarrow$

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

numerical stable -> avoid numerical underflow

normalize by length
so that it wont penalize by sentence length

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

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

$\alpha = 0.7$

$\alpha = 1$

$\alpha = 0$

Andrew Ng

Beam search discussion

Beam width B?

large B: better result, slower

Small B: worse result, faster

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

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

appromx search algo
and it is faster

deeplearning.ai

Sequence to
sequence models

Error analysis on
beam search

Example
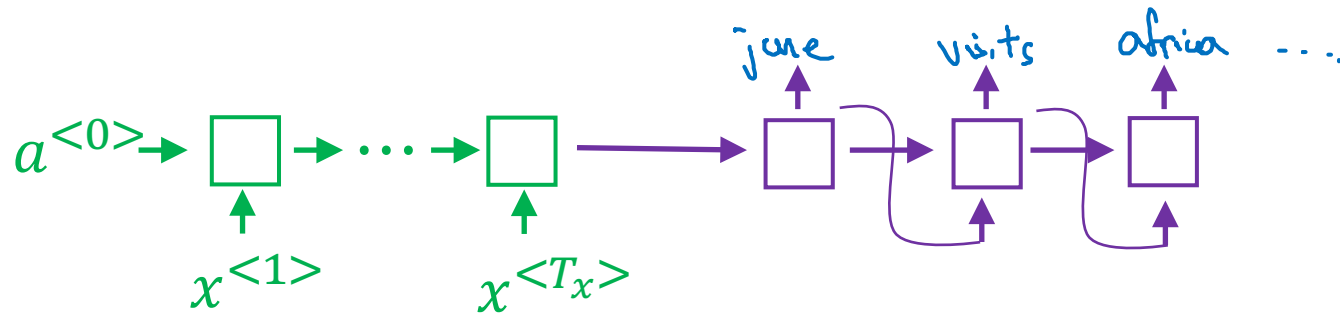
Jane visite l'Afrique en septembre.

$\rightarrow$ RNN

$\rightarrow$ Beam Search

BT

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

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

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

$a^{<0>}$ $x^{<1>}$ $x^{<T_x>}$ jane visits africa ...

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: $P(y^* | x) > P(\hat{y} | x)$ ←

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

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

Conclusion: Beam search is at fault.

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

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
				...

increase beam width
rnn:
change model architecture
tune hyperparameter
retrain

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

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: 7/7   Modified precision: 2/7

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.

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

$p_1, p_2 = 1.0$

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

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

n-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 = brevy 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]
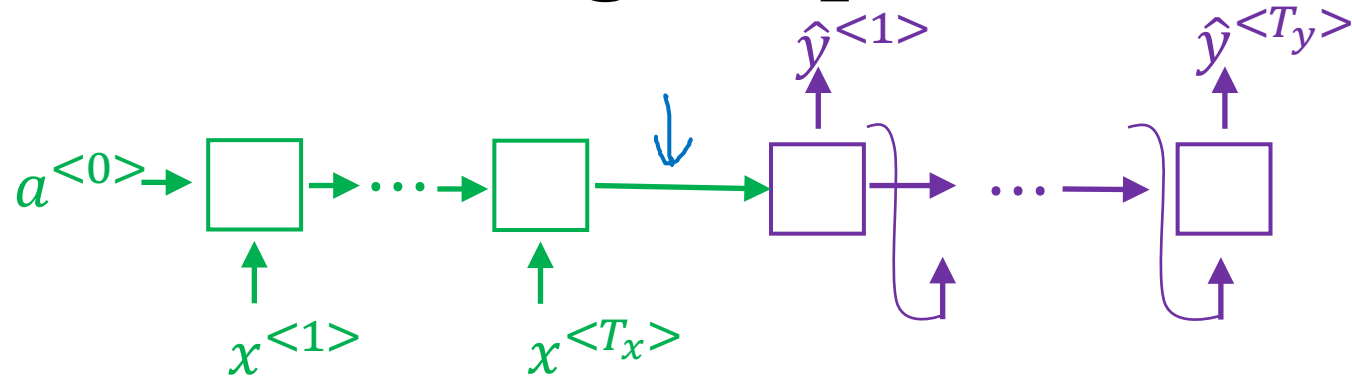
Andrew Ng
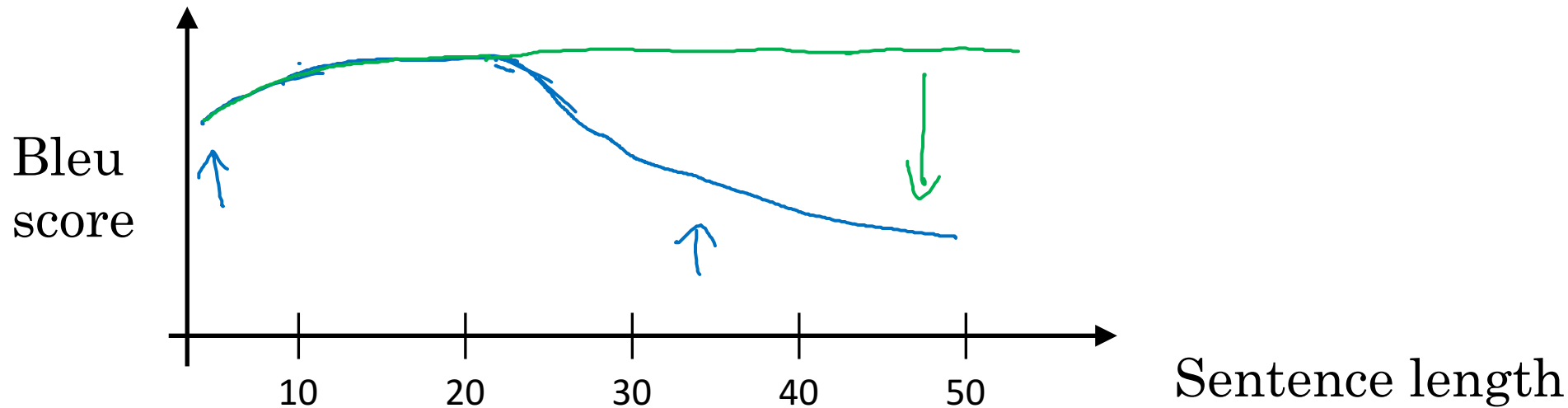
deeplearning.ai

Sequence to sequence models

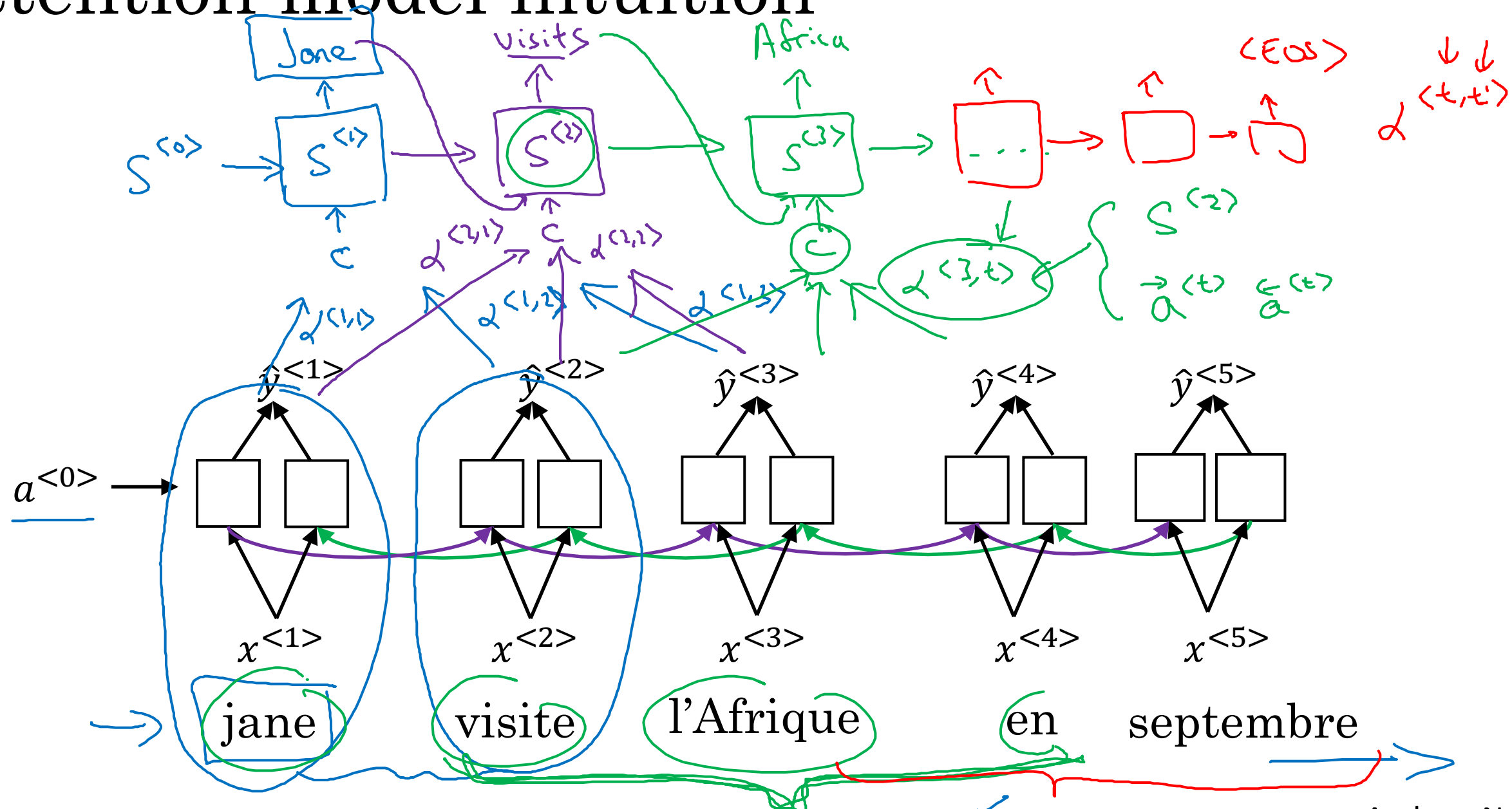Attention model intuition

The problem of long sequences

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.
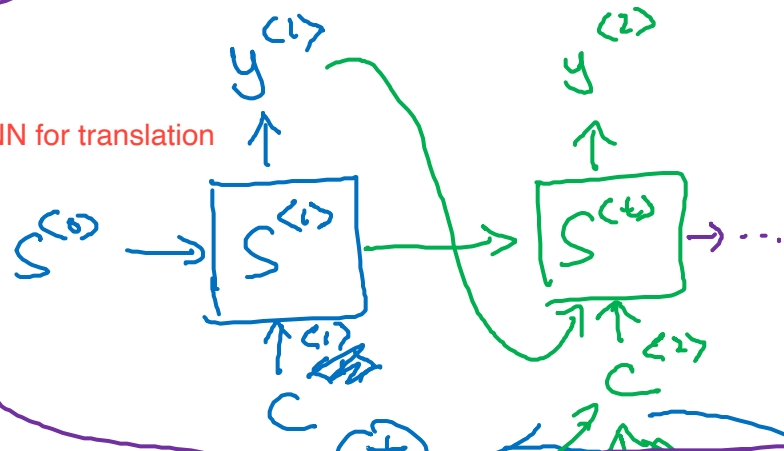

Bleu score

Sentence length

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 $a^{<t'>}$.

single direction RNN for translation

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

$S^{<0>} \rightarrow S^{<1>} \rightarrow S^{<t>} \rightarrow \cdots$

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

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

$C^{<1>}$

$C^{<2>}$

attention weight: how much attention needed

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

$\alpha^{<1,1>}$ $\cdots$

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

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

bi directional RNN

$\overrightarrow{a}^{<0>} \rightarrow$ $\overrightarrow{a}^{<1>}$ $\overleftarrow{a}^{<1>}$ $\overrightarrow{a}^{<2>}$ $\overleftarrow{a}^{<2>}$ $\overrightarrow{a}^{<3>}$ $\overleftarrow{a}^{<3>}$ □ □ $\overrightarrow{a}^{<5>}$ $\overleftarrow{a}^{<5>}$ $\leftarrow \overleftarrow{a}^{<6>}$

t'

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

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

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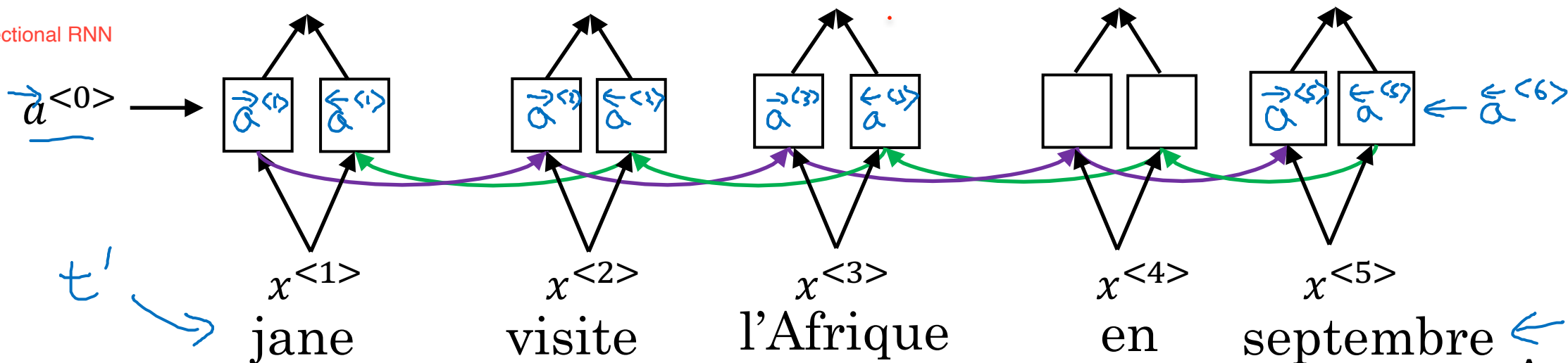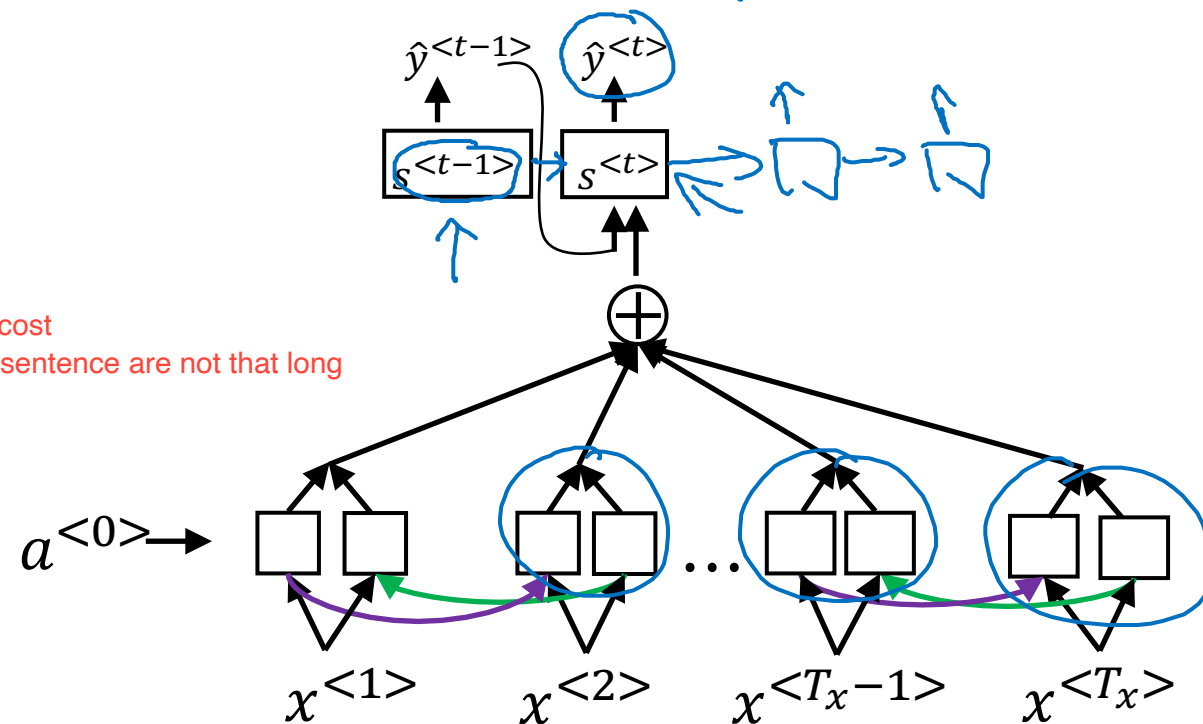
$\alpha^{<t,t'>}$ = amount of attention $y^{<t>}$ should pay to $a^{<t'>}$

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

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

$s^{<t-1>}$ $s^{<t>}$
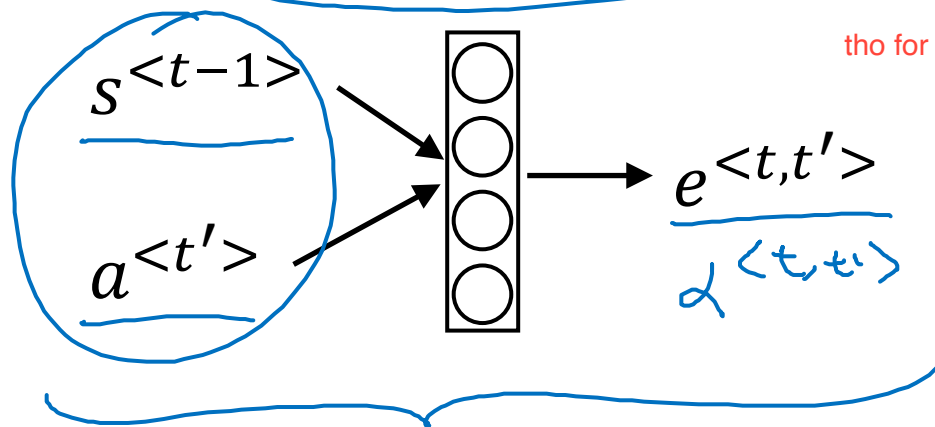
quatratic cost
tho for machine translation, the sentence are not that long

$s^{<t-1>}$

$a^{<t'>}$

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

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

$a^{<0>}$

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

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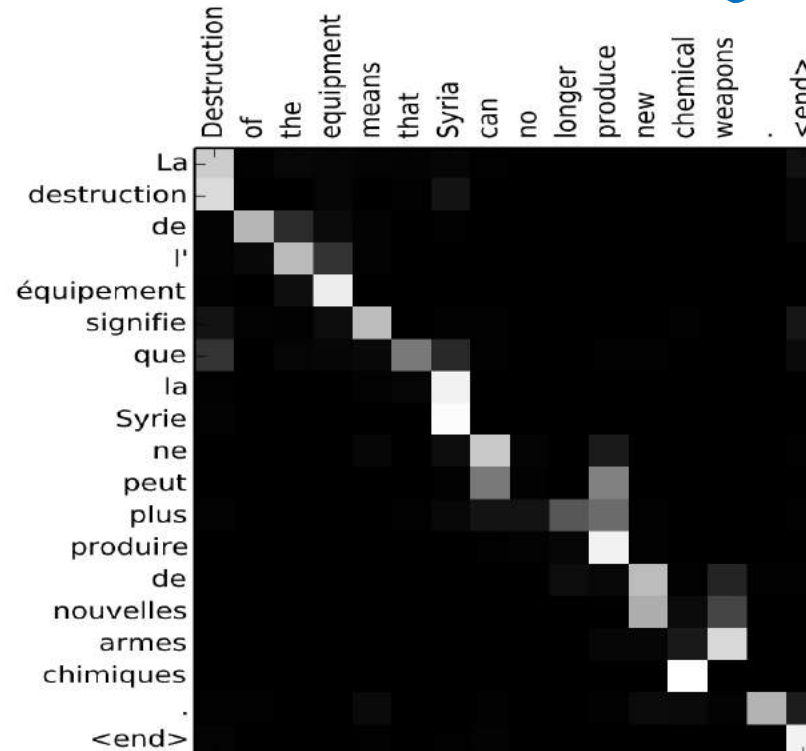[Xu et. al., 2015. Show, attend and tell: Neural image caption generation with visual attention]

Andrew Ng

Attention examples

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

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

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

Audio data

Speech recognition
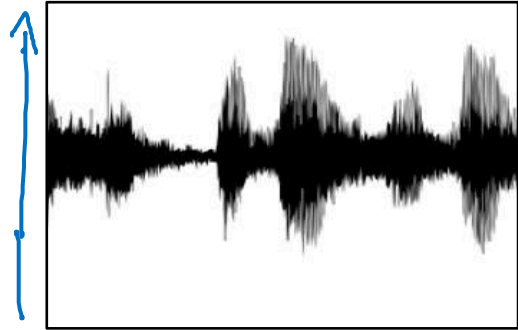
deeplearning.ai

Speech recognition problem
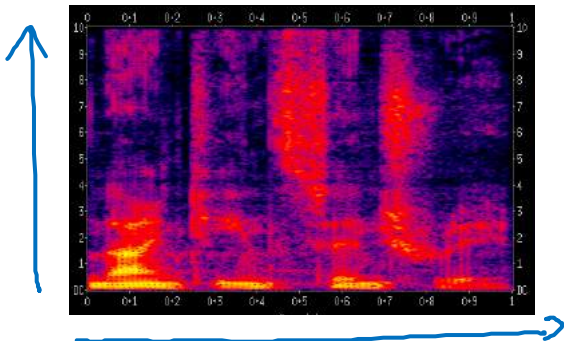
x

audio clip

y

transcript

"the quick brown fox"

phonemes: de kwik braun
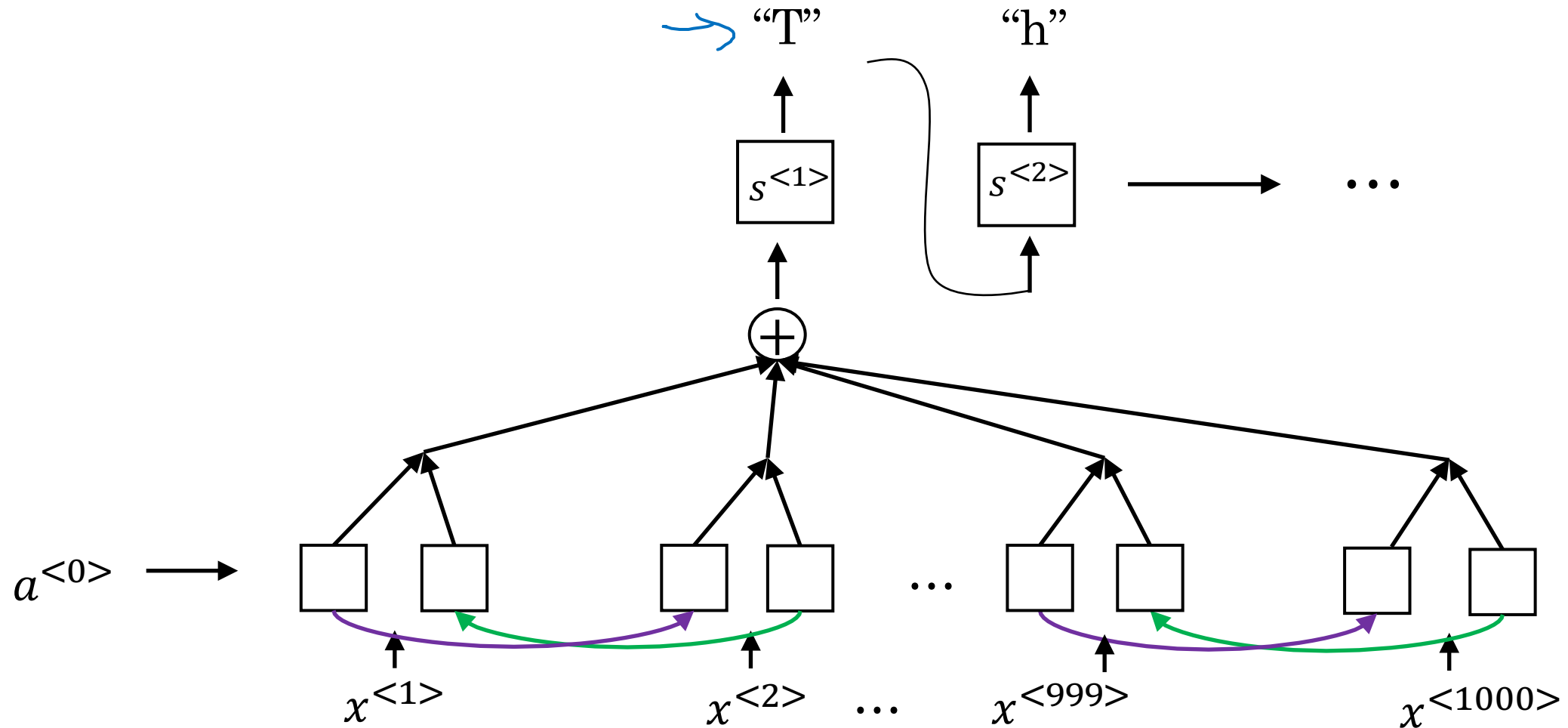
300h
3000h
100,000h

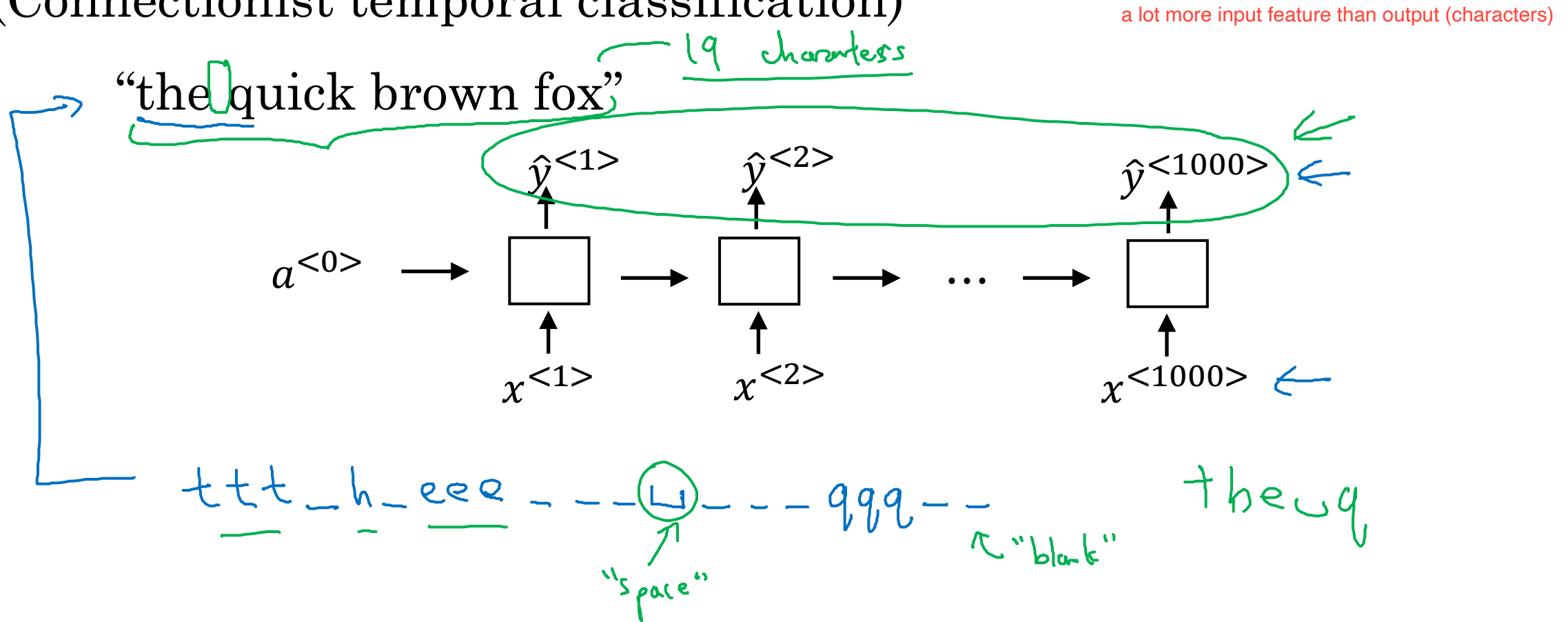# Attention model for speech recognition

CTC cost for speech recognition

(Connectionist temporal classification)

a lot more input feature than output (characters)

19 characters

"the quick brown fox"

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

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

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

ttt_h_eee - - - - ⊔ - - - - qqq - -   theˌq

"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
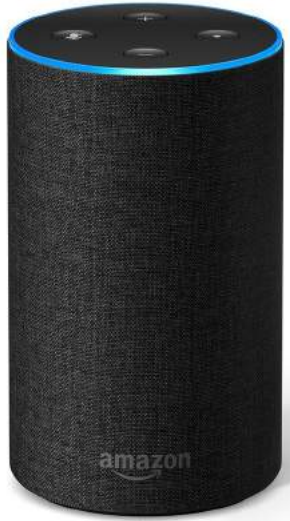
deeplearning.ai

Audio data
___

Trigger word
detection

What is trigger word detection?
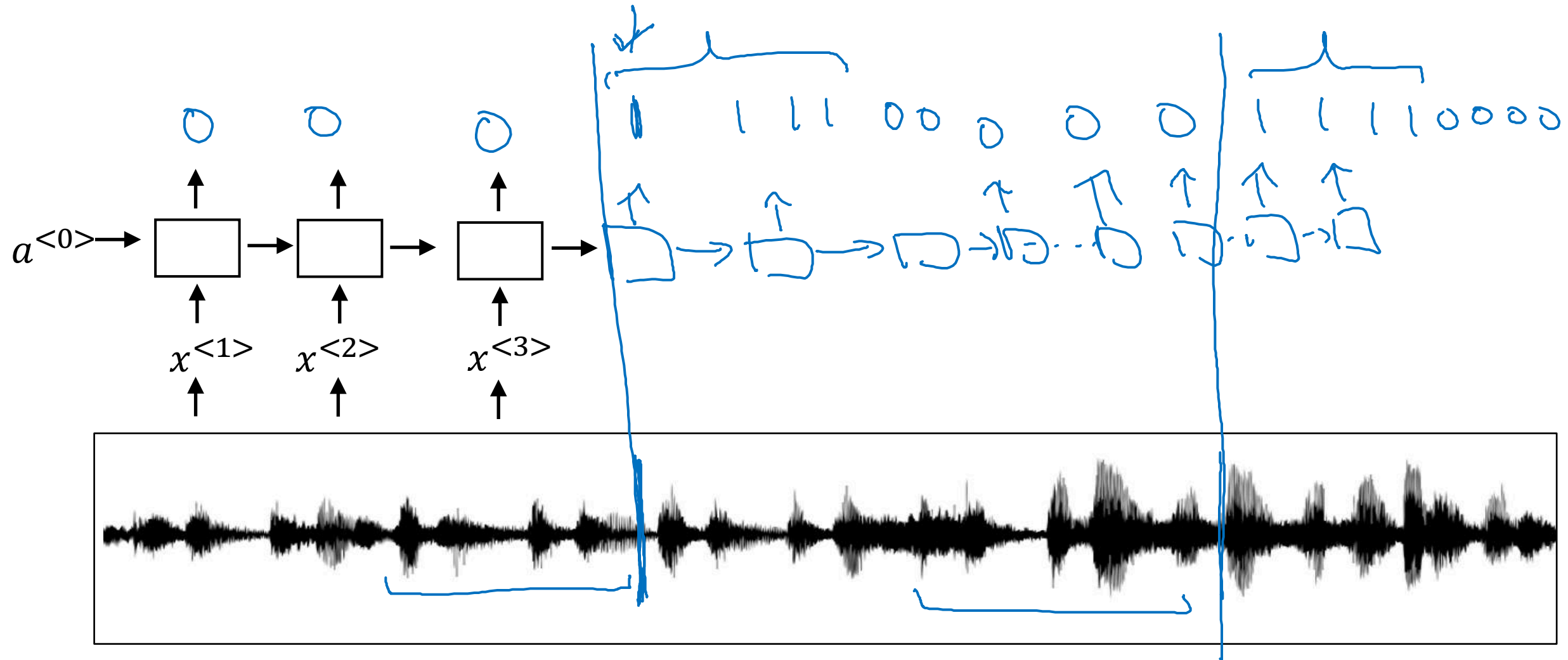

Amazon Echo
(Alexa)

Baidu DuerOS
(xiaodunihao)

Apple Siri
(Hey Siri)

Google Home
(Okay Google)

Andrew Ng

Trigger word detection algorithm

deeplearning.ai

Conclusion

Summary and thank you

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

Andrew Ng

Deep learning is a super power

Please buy this from shutterstock and replace in final video.

www.shutterstock.com · 331201091

Andrew Ng

Thank you.

- Andrew Ng