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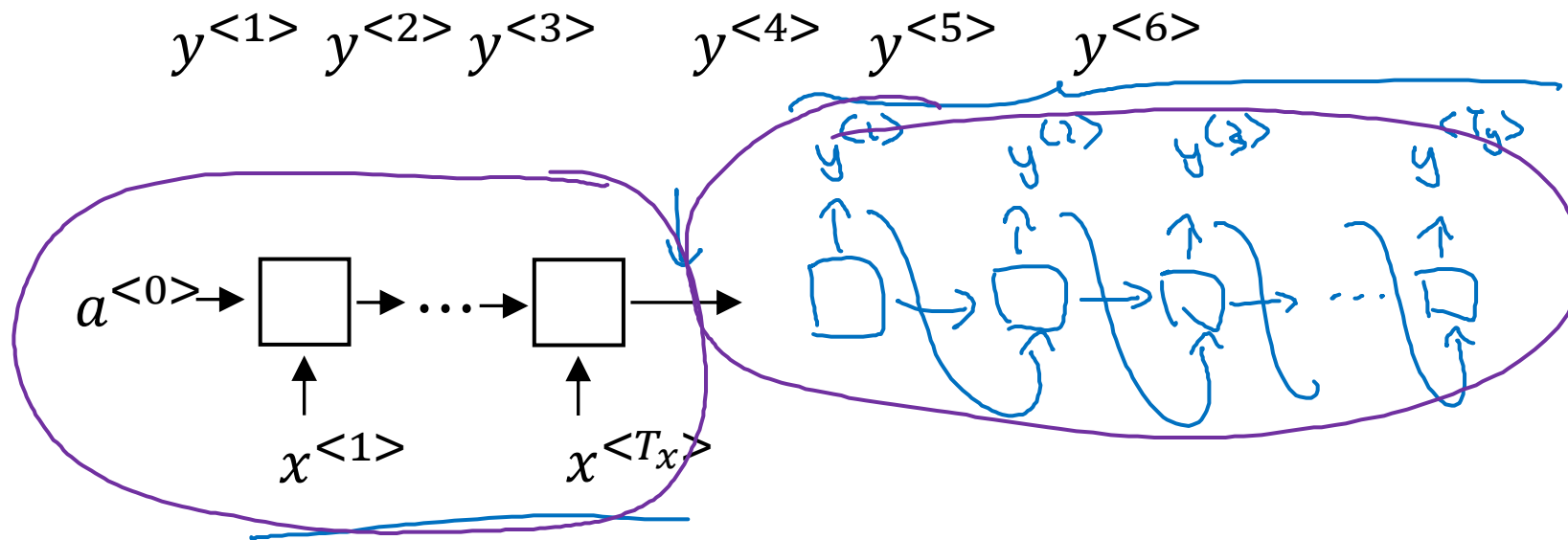
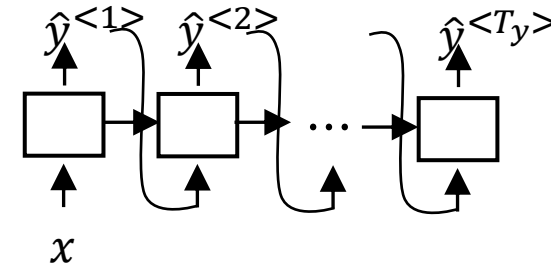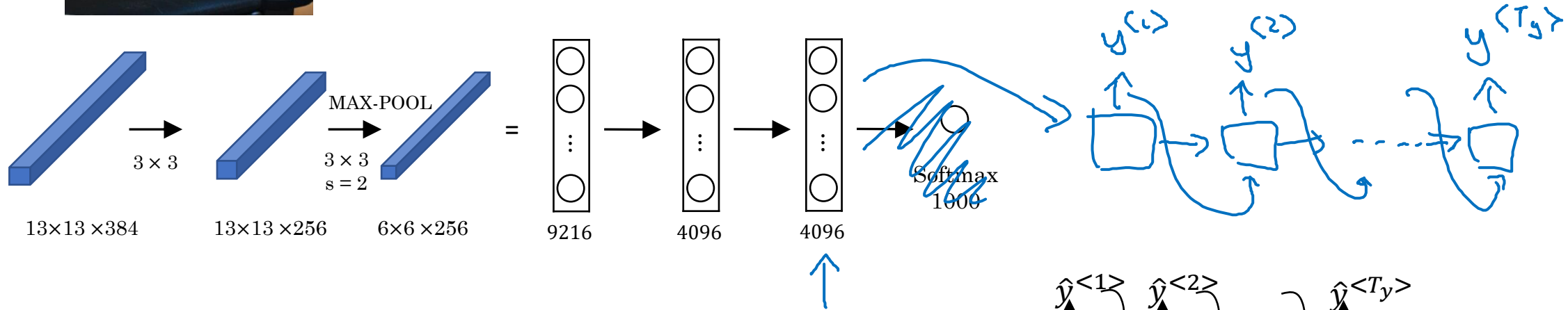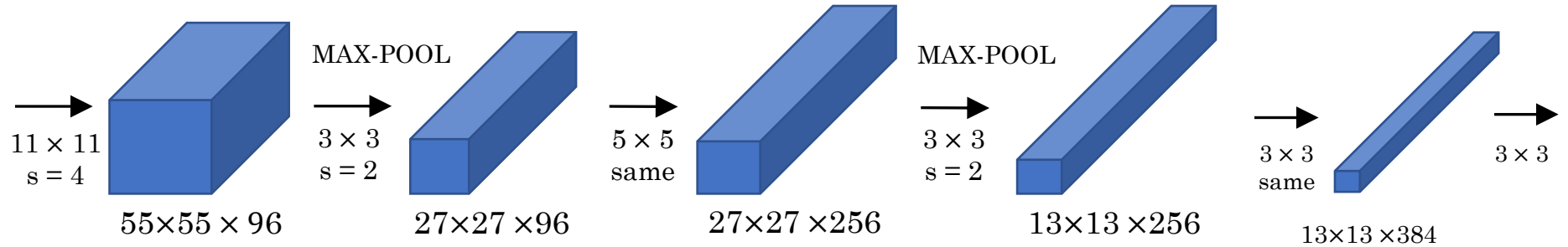$y^{<1>}$ $y^{<2>}$ $y^{<3>}$ $y^{<4>}$ $y^{<5>}$ $y^{<6>}$

[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]

Andrew Ng
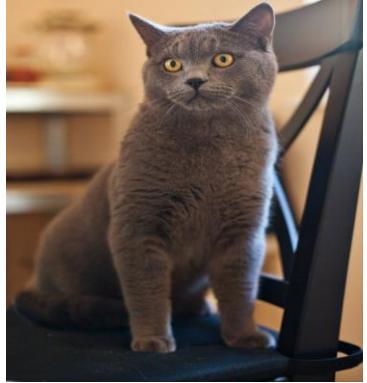
Image captioning

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

A cat sitting on a chair

11×11
$s = 4$

$55 \times 55 \times 96$

MAX-POOL

3×3
$s = 2$

$27 \times 27 \times 96$

5×5
same

$27 \times 27 \times 256$

MAX-POOL

3×3
$s = 2$

$13 \times 13 \times 256$

3×3
same

$13 \times 13 \times 384$

3×3

$13 \times 13 \times 384$

3×3

$13 \times 13 \times 256$

MAX-POOL

3×3
$s = 2$

$6 \times 6 \times 256$

9216

4096

4096

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]
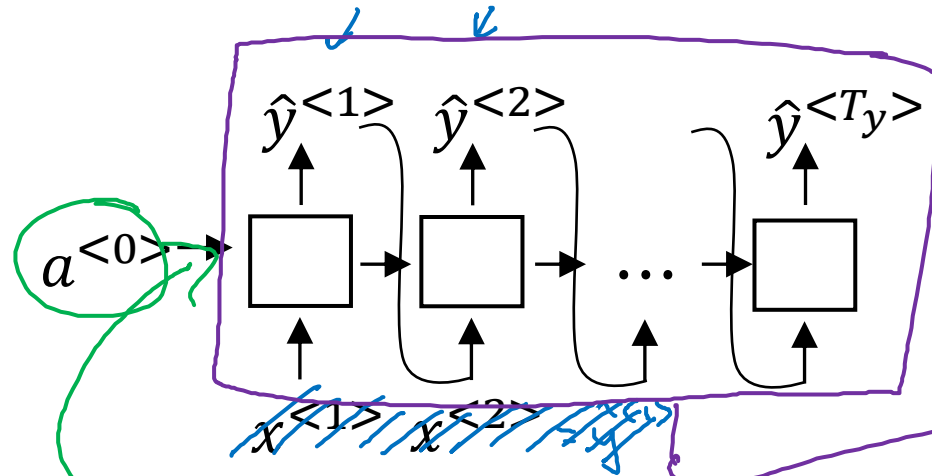
Andrew Ng

Sequence to sequence models

Picking the most likely sentence
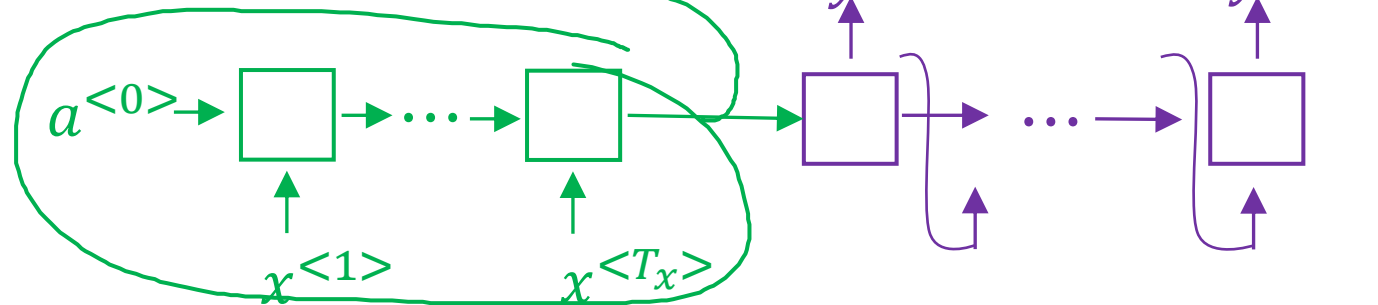
deeplearning.ai

Machine translation as building a conditional language model

Language model:

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

Machine translation:

"Conditional language model"

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

Andrew Ng

Finding the most likely translation
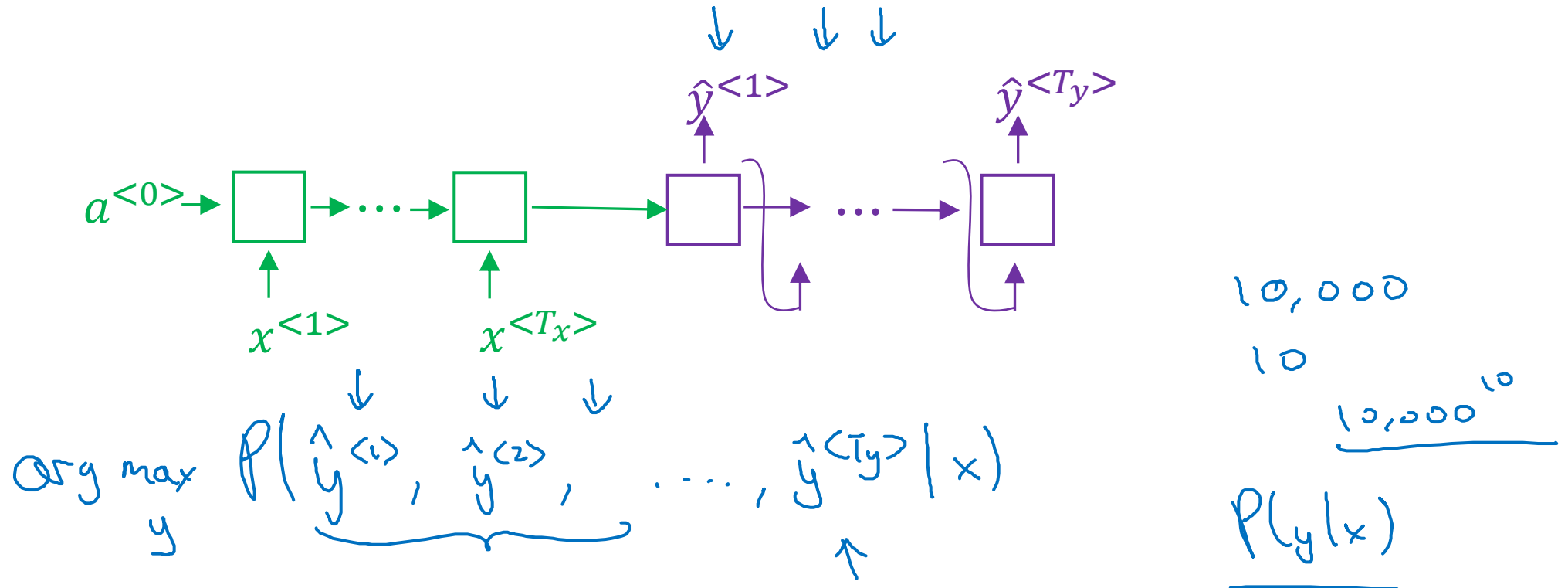
English   French

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.

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

Why not a greedy search?

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

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

$a^{<0>}$

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

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

$10,000$

10

$\frac{10,000^{10}}{\;}$

$P(y|x)$

→ Jane is visiting Africa in September.

→ Jane is going to be visiting Africa in September.

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

Sequence to sequence models

Beam search

deeplearning.ai

Beam search algorithm

$B = 3$ (beam width)

Step 1

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

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

$a^{<0>} \rightarrow \square \rightarrow \cdots \rightarrow \square \rightarrow \square \rightarrow \hat{y}^{<1>}$

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

Andrew Ng

Beam search algorithm

$(B = 3)$

Step 1

Step 2

a
.
.
.
in
.
.
.
jane
.
.
.
september
.
.
.
zulu

10000

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

a
aaron
.
September
visit
.
zulu

a
aaron
.
is
.
visit
.
zulu

10,000

a
.
.
zulu

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

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

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

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

in

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

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

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

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

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

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

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

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

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

Andrew Ng
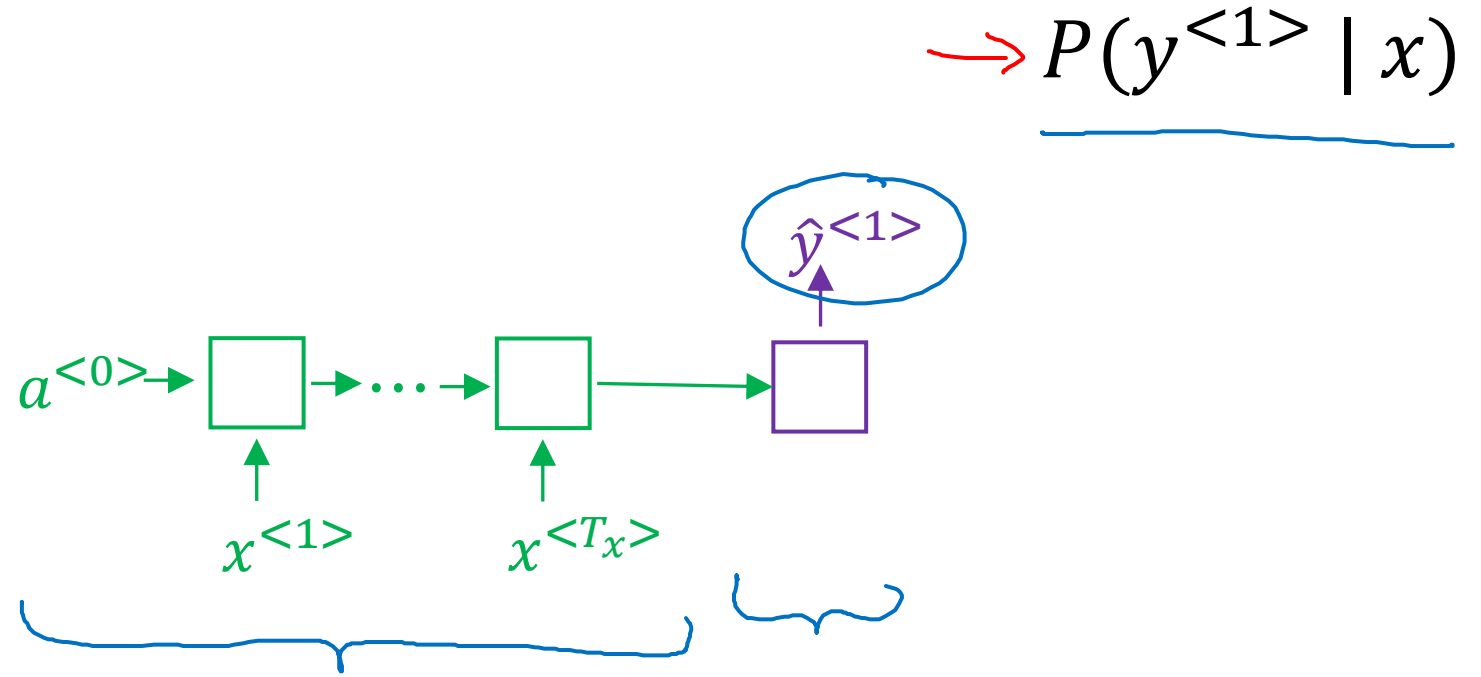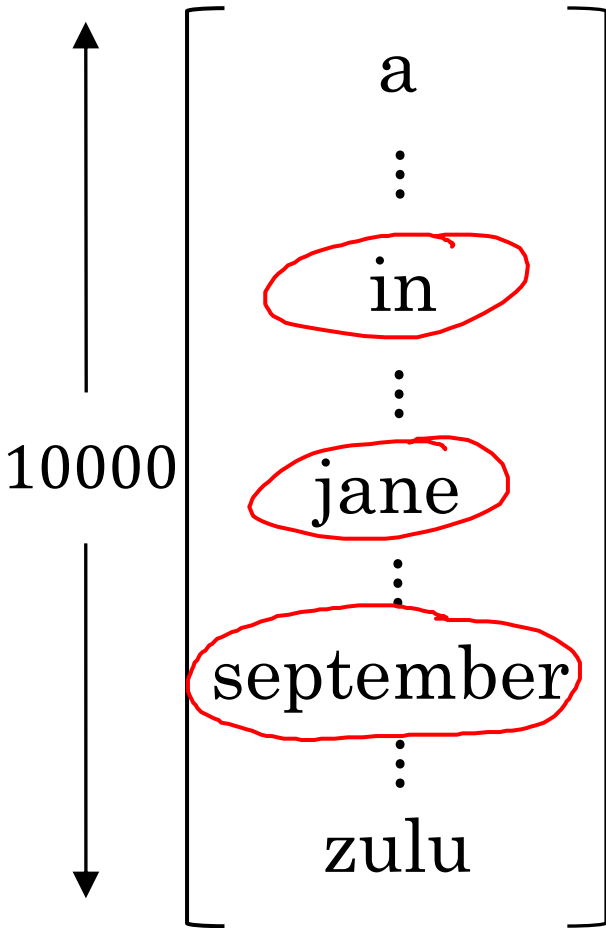
Beam search ($B = 3$)

$B = 1 \rightsquigarrow$ greedy search

in september
$\begin{array}{l} a \\ aaron \\ jane \\ zulu \end{array}$

jane is
$\begin{array}{l} a \\ visits \\ zulu \end{array}$

jane visits
$\begin{array}{l} a \\ africa \\ zulu \end{array}$

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

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

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

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

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

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

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

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

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

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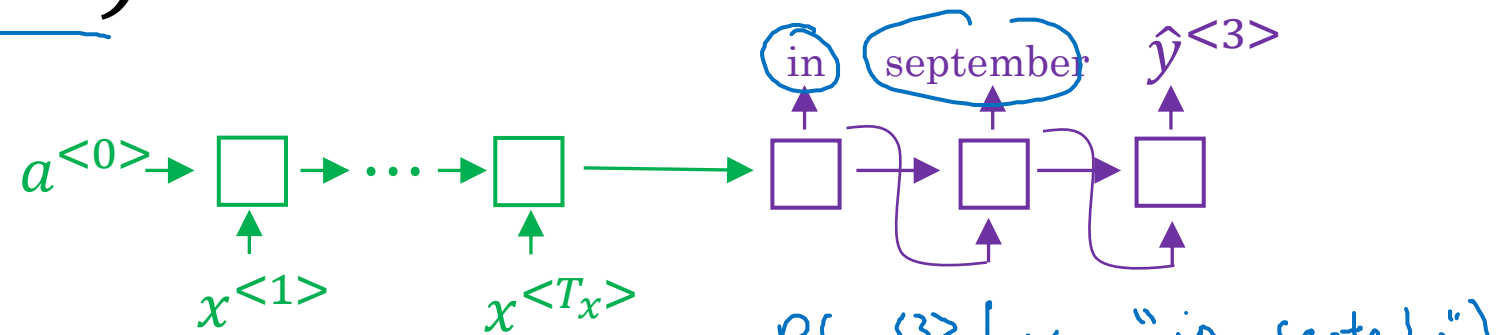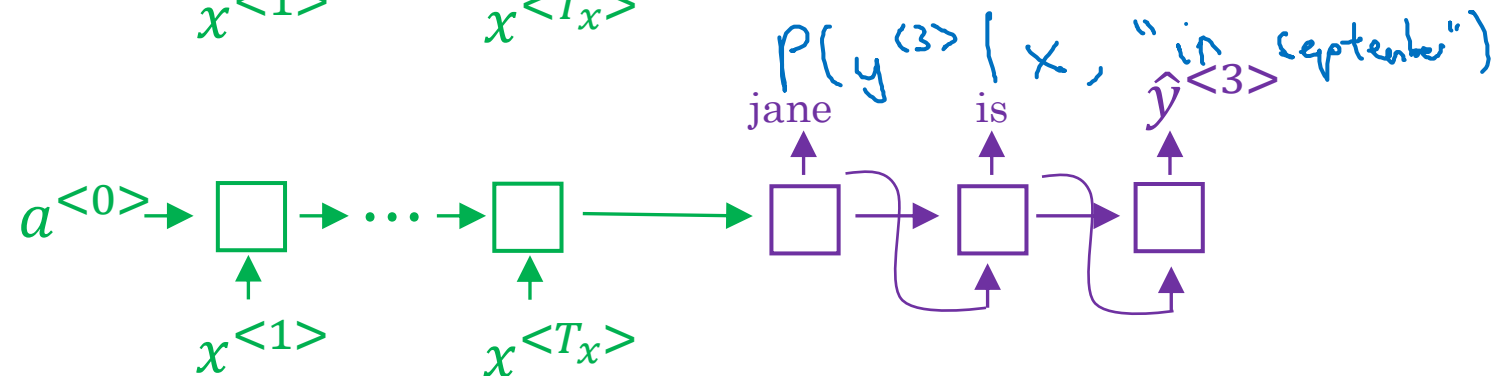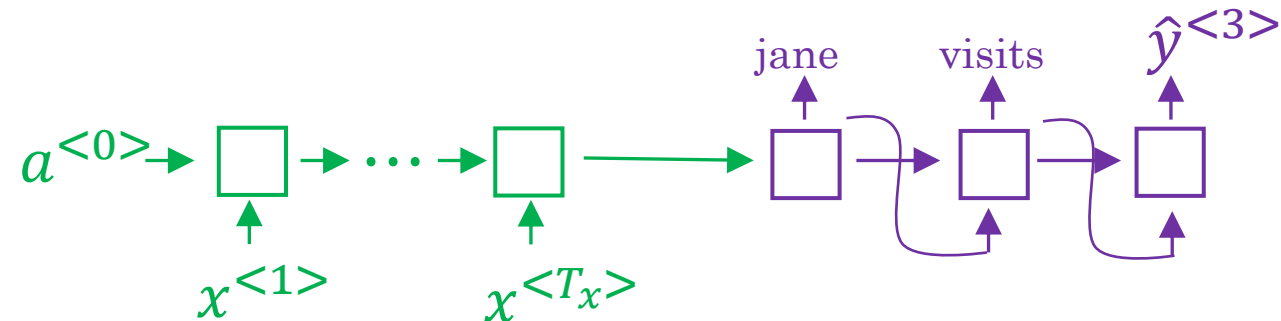
$P(y^{<1>}, y^{<2>} \mid x)$

jane visits africa in september. <EOS>

Andrew Ng

Sequence to sequence models

Refinements to beam search

deeplearning.ai

Length normalization

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

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

log

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

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

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

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

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

$$\alpha = 0.7 \qquad \alpha = 1$$
$$\alpha = 0$$

Andrew Ng
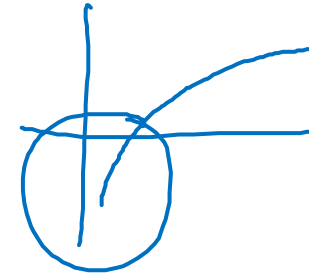
Beam search discussion

Beam width B?

large B: better result, slower

small B: worse result, faster

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

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

deeplearning.ai

Sequence to sequence models

Error analysis on beam search

Example

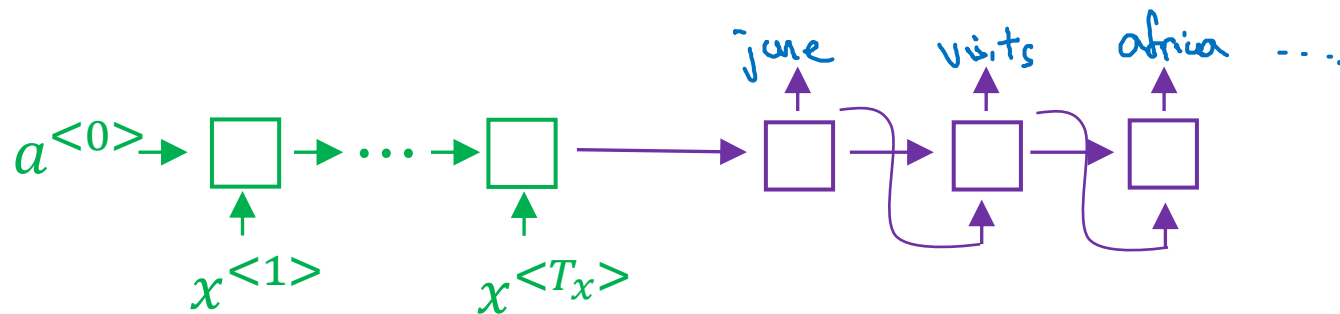Jane visite l'Afrique en septembre.

→ 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: $P(y^* | x) > P(\hat{y} | x)$ ←

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

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 $\boxed{P(y^* | x)} < \underline{P(\hat{y} | x)}$.

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

Sequence to sequence models

Bleu score (optional)

deeplearning.ai

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. $(\hat{y})$

$p_1, p_2 = 1.0$

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

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

n-gram

[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

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

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

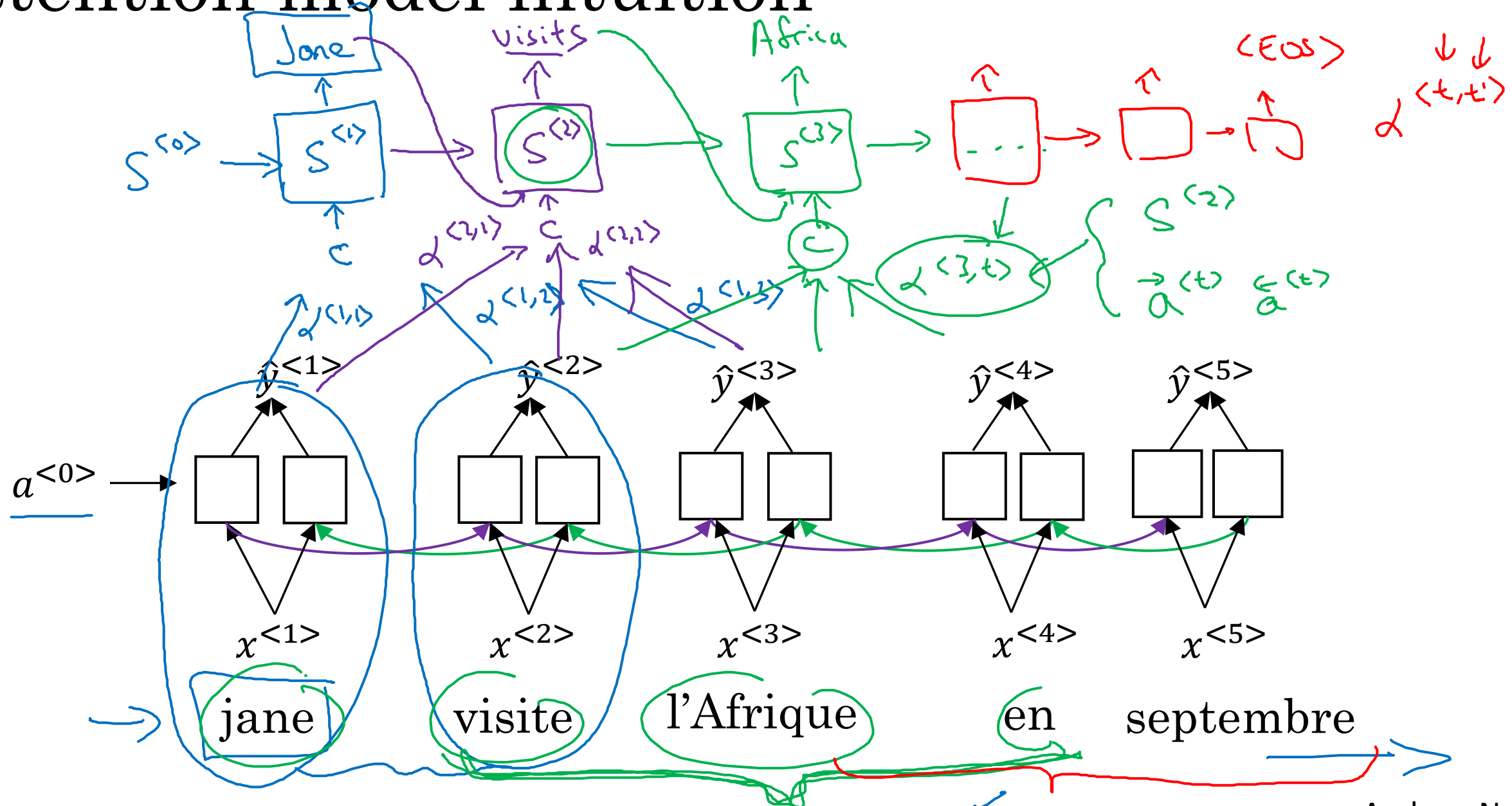$$\hat{y}^{<1>} \qquad \hat{y}^{<T_y>}$$
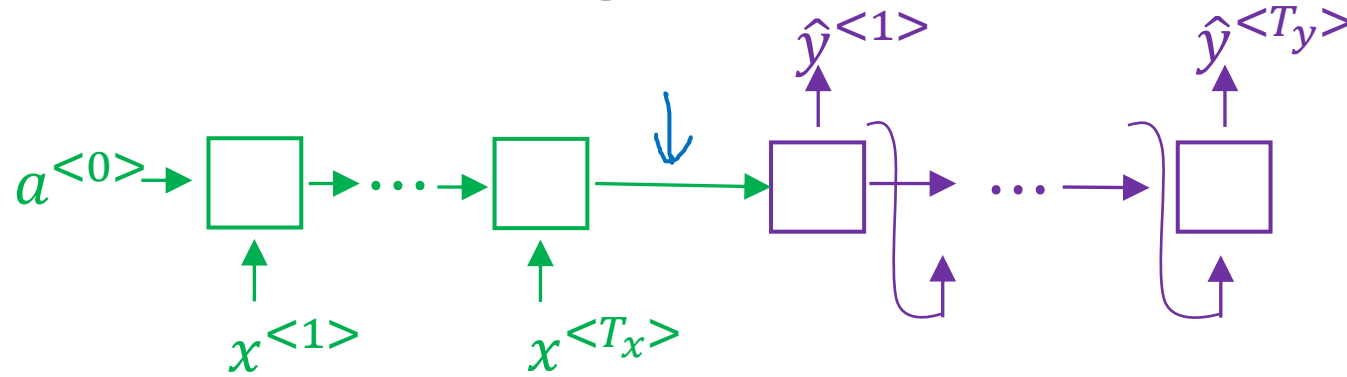
Jane s'est rendue en Afrique en septembre dernier, a apprécié la culture et a rencontré beaucoup de gens merveilleux; elle est revenue en parlant comment son voyage était merveilleux, et elle me tente d'y aller aussi.

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.
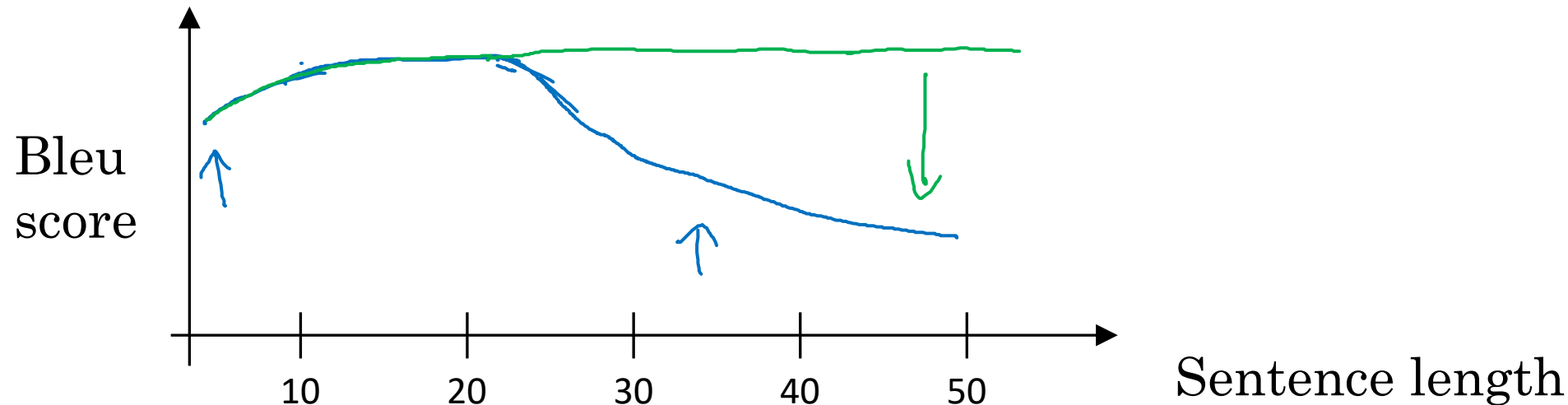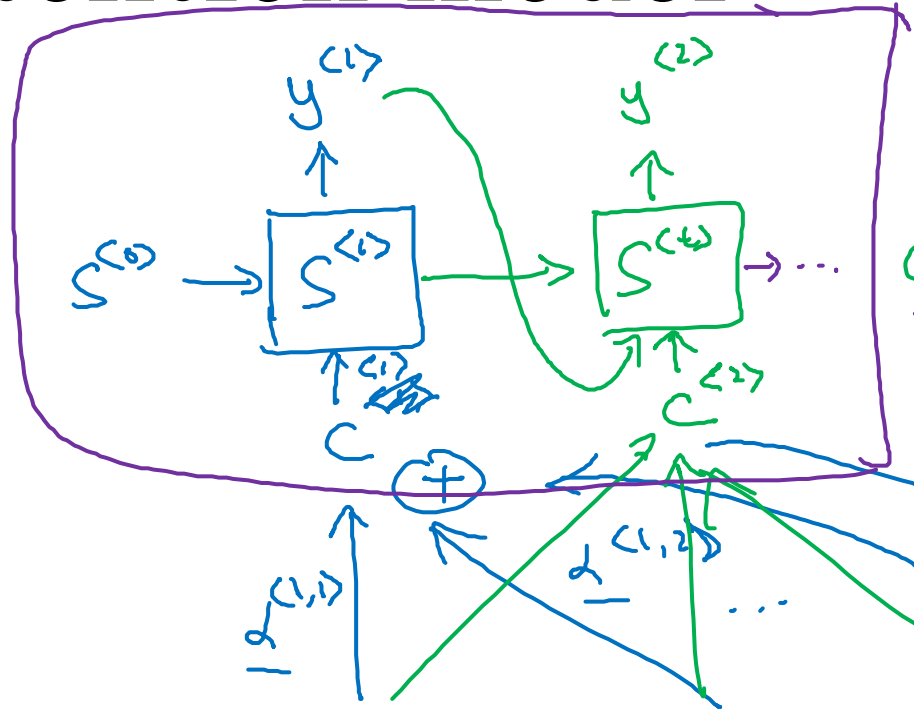
Bleu score

Sentence length

10   20   30   40   50

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 $a^{<t'>}$.

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

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

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

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

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

$x^{<1>}$ jane

$x^{<2>}$ visite

$x^{<3>}$ l'Afrique

$x^{<4>}$ en

$x^{<5>}$ septembre

t'

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

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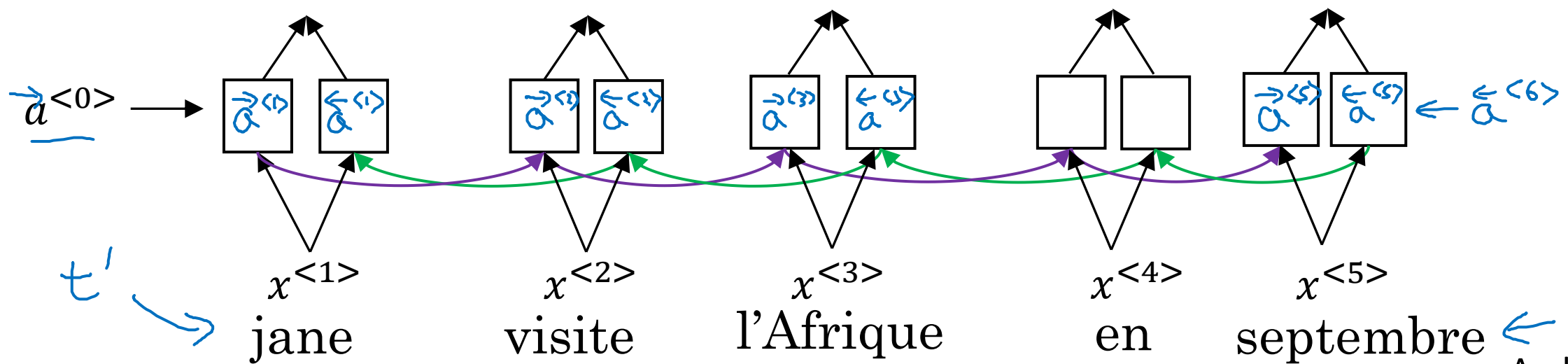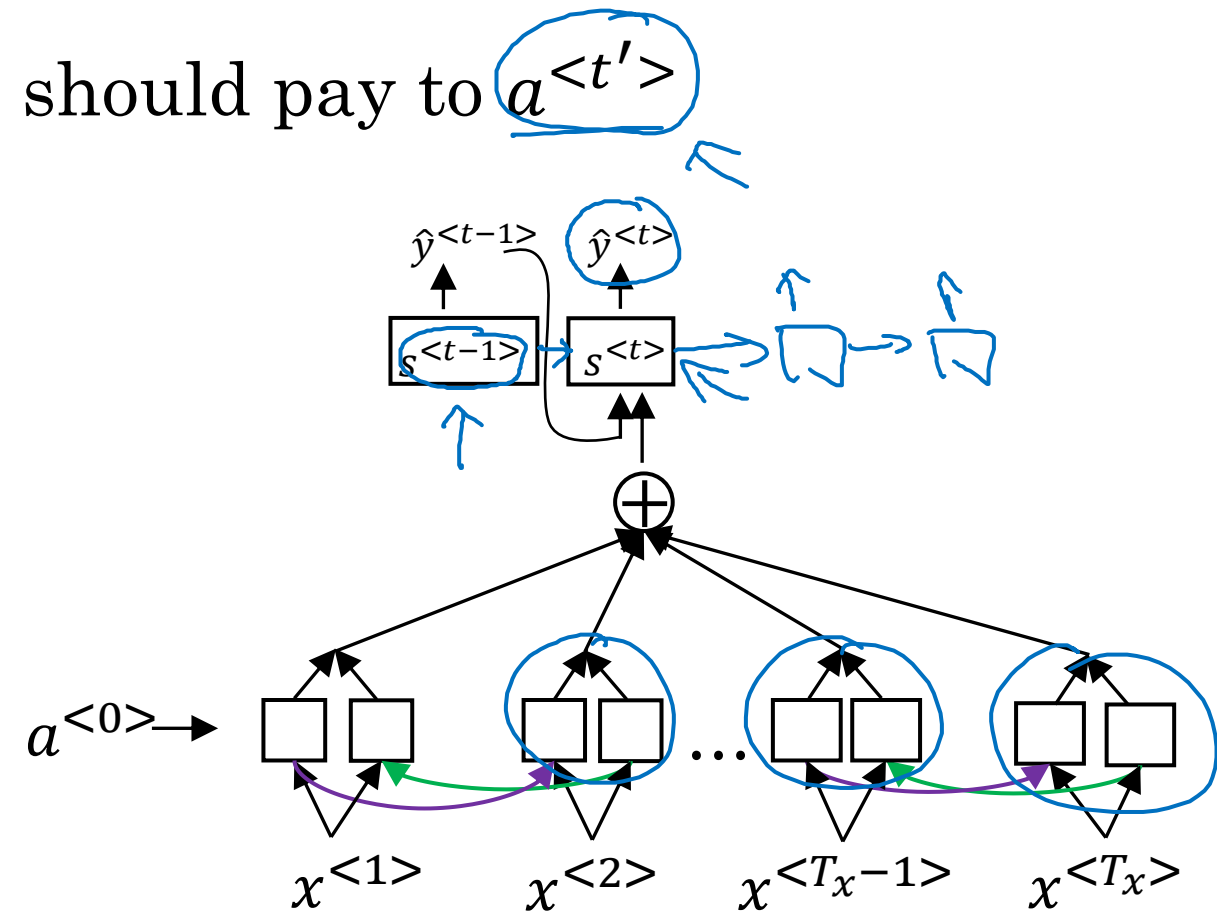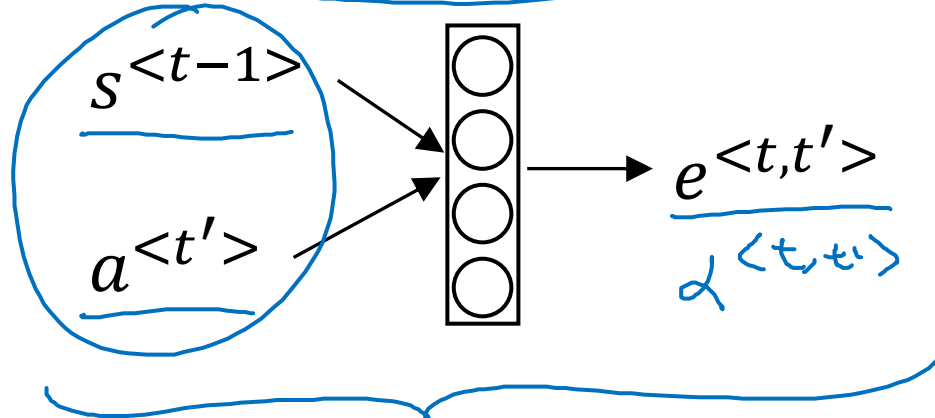
Computing attention $\alpha^{<t,t'>}$

T_x T_y

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

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

$s^{<t-1>}$

$a^{<t'>}$

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

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

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

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

$\oplus$

$a^{<0>}$

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

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