Copyright Notice

Sequence to sequence models

Basic models

deeplearning.ai

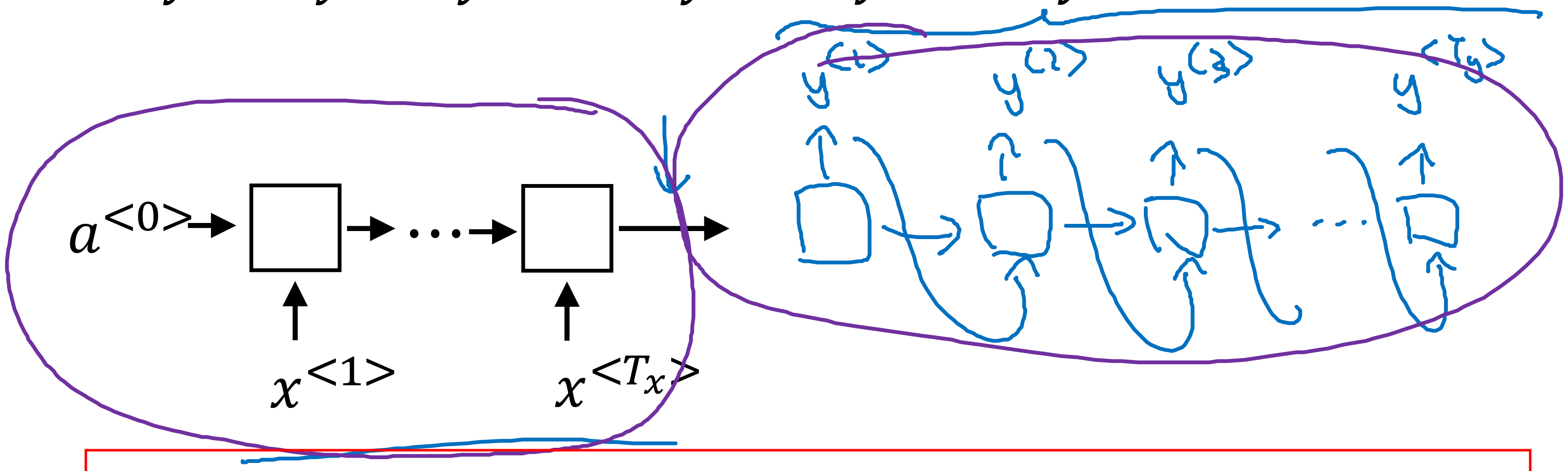
Sequence to sequence model

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

Jane visite l'Afrique en septembre

→ Jane is visiting Africa in September.

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

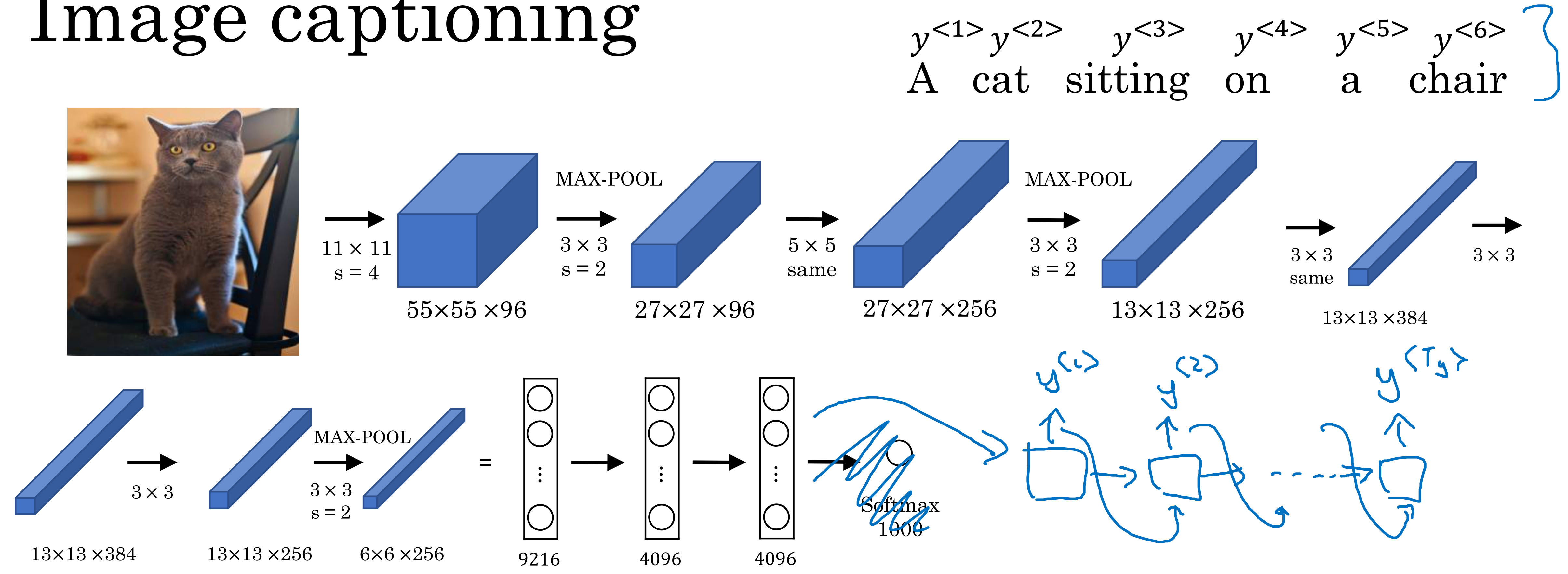


Encoder Network . and That side is the decoder network and the one connecting them is the

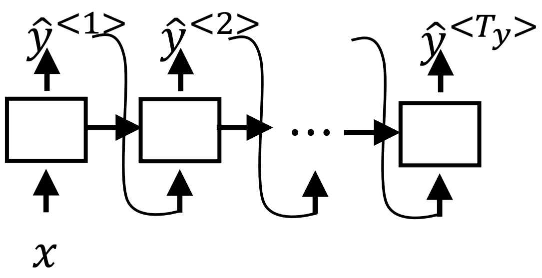
[Sutskever et al., 2014. Sequence to sequence learning with neural networks]

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

Image captioning



We train on the convnet output obtained by deleting that last output final layer.



[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

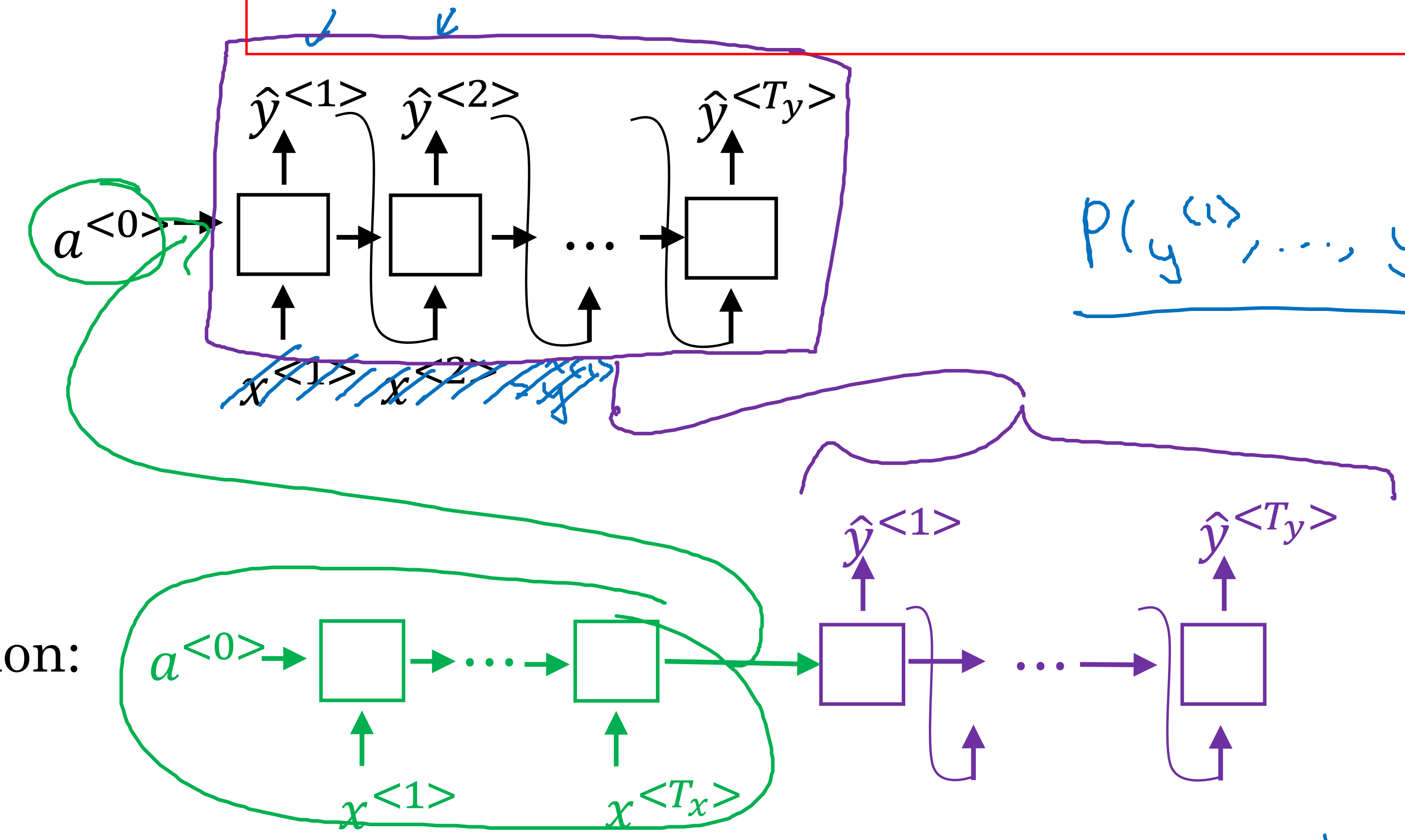
deeplearning.ai

Machine translation as building a conditional language model

Very similar to word prediction model but instead of feeding the constant a0 we feed in the output o

Language model:

Machine translation:



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

"Conditional language model"

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

Andrew Ng

Finding the most likely translation

Jane visite l'Afrique en septembre.

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

→ Jane is visiting Africa in September.

→ Jane is going to be visiting Africa in September.

→ In September, Jane will visit Africa.

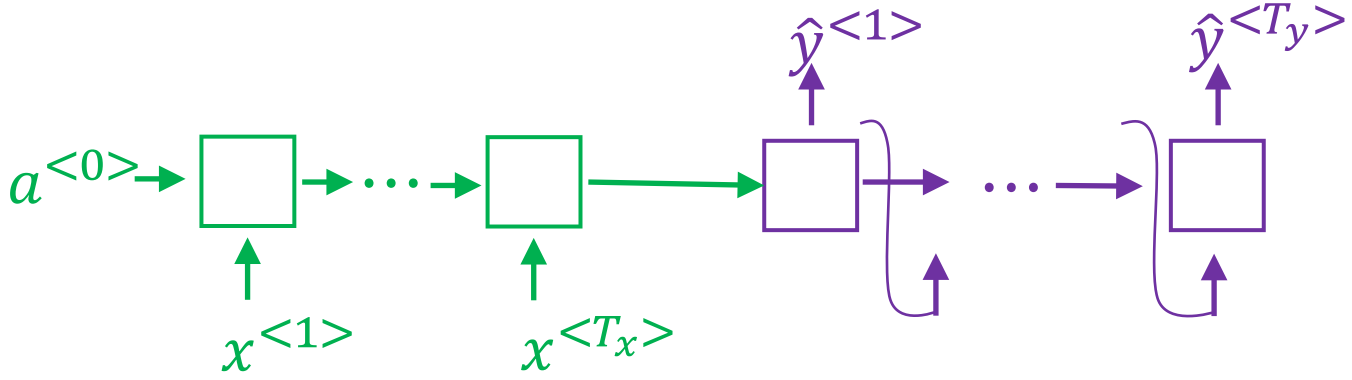
→ Her African friend welcomed Jane in September.

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

We can think of this machine t

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

$10,000^{10}$

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

→ Jane is visiting Africa in September.

→ Jane is going to be visiting Africa in September.

P(Jane is going | x) > P(Jane is visiting | x)

Sequence to sequence models

deeplearning.ai

Beam search

For machine translation and for audio to text generation.

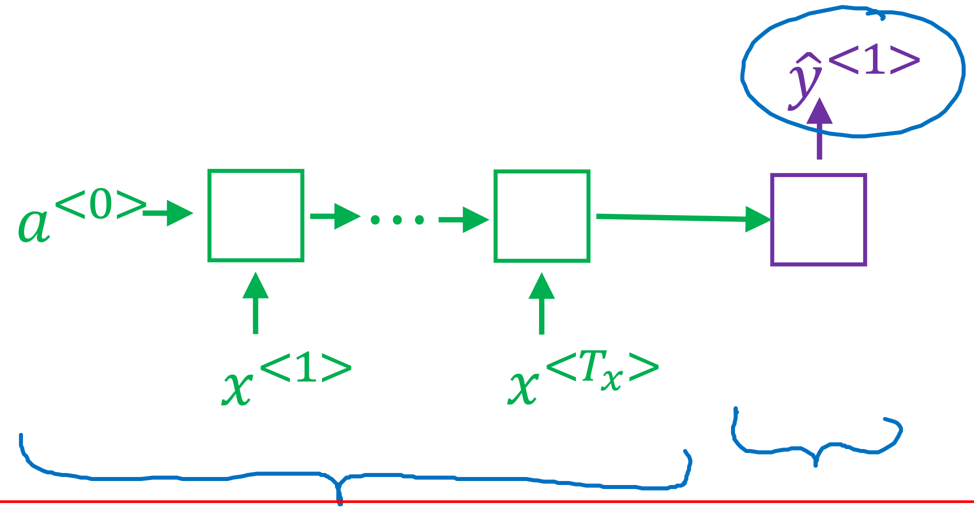
Beam search algorithm

$B = 3$ (beam width)

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

Step 1

$$10000 \left[\begin{array}{c} \text{a} \\ \vdots \\ \text{in} \\ \vdots \\ \text{jane} \\ \vdots \\ \text{september} \\ \vdots \\ \text{zulu} \end{array}\right]$$

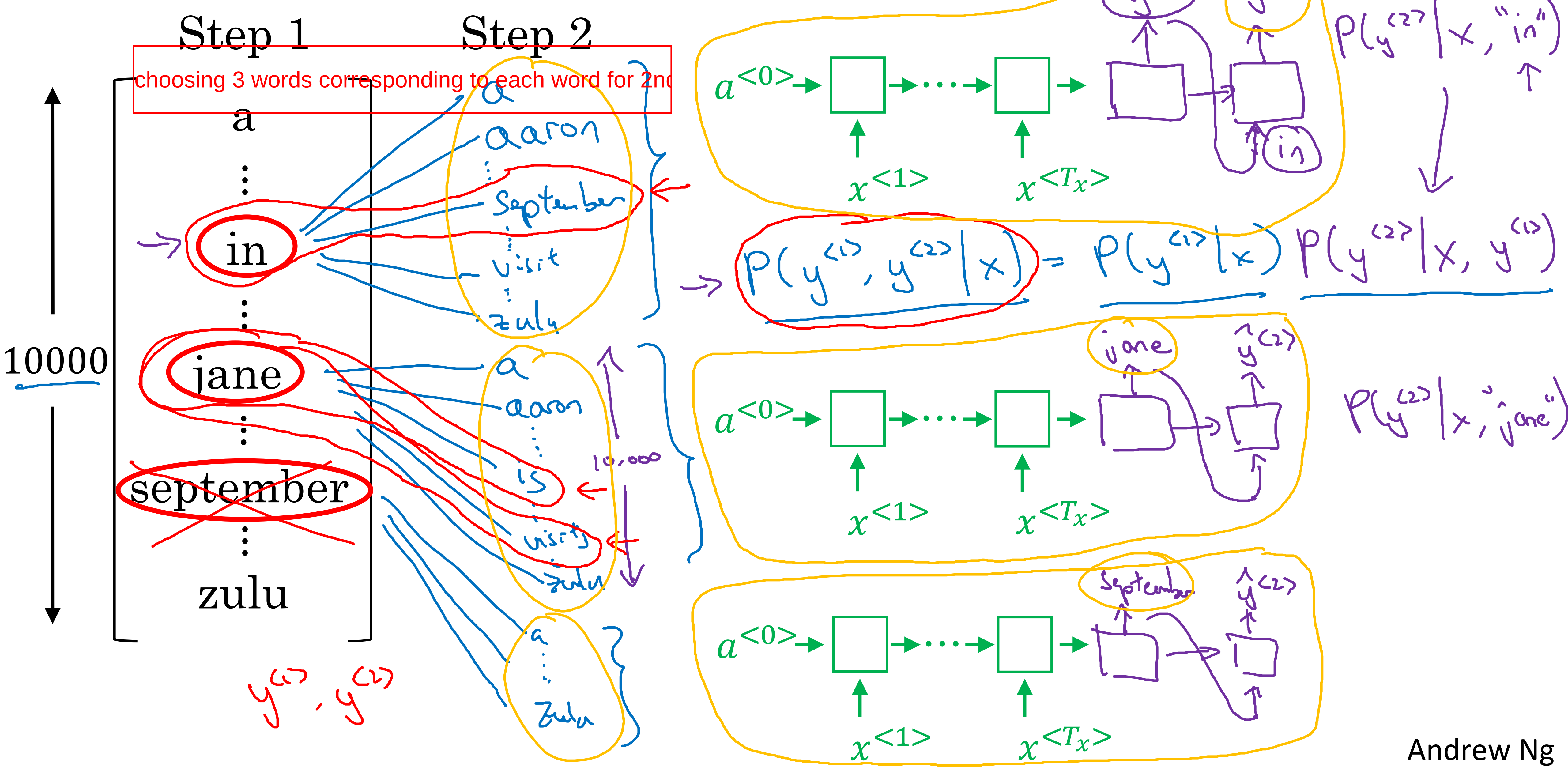


In greedy given the encoded representation we choose that what is the best word here. But it sometimes results

Andrew Ng

Beam search algorithm

$(B=3)$



Andrew Ng

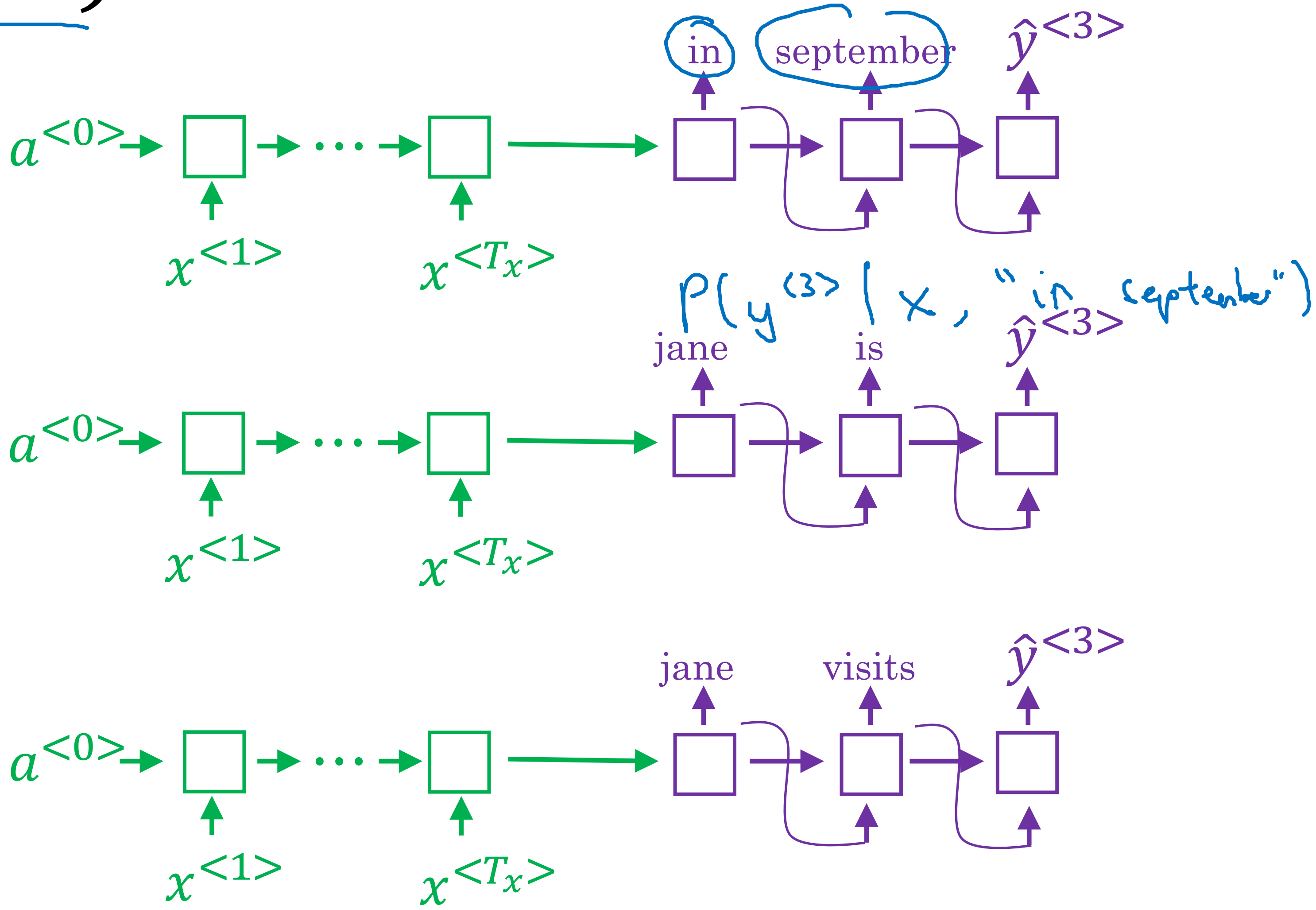
Beam search ($B = 3$)

$B = 1 \rightsquigarrow$ greedy search

in september — a, aaron, jane, zulu

jane is — a, visits, zulu

jane visits — a, africa, zulu



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

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

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

$\log P(y|x)$

$P(y|x)$

convert ot log so as to convert produ

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

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

$\alpha = 0.7$

$\alpha = 1$

$\alpha = 0$

Beam search discussion

Beam width B?

1 → 3 → 10, 100, 1000, → 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 $\arg\max_y P(y|x)$.

Andrew Ng

deeplearning.ai

Sequence to sequence models

Error analysis on beam search

Example

→ RNN

→ Beam Search

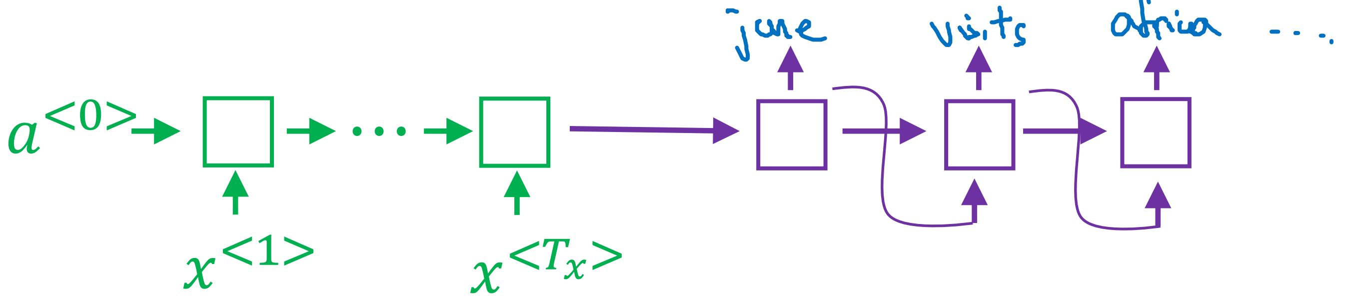
$\boxed{BT}$

Jane visite l'Afrique en septembre.

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

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

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



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.

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

There are many translation possible for a givven sequence of words, but bleu score tells us how good is a particular translation.

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: Modified precision:

THE the predict kar raha par ye dono dond

Bleu

bilingual evaluation understudy

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

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. ($\hat{y}$)

$p_1, p_2, = 1.0$

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

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

n-gram

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

unigram

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

Andrew Ng

Bleu details

p_1, p_2, p_3, p_4

p_n = Bleu score on n-grams only

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

BP = brevity penalty

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

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

Andrew Ng

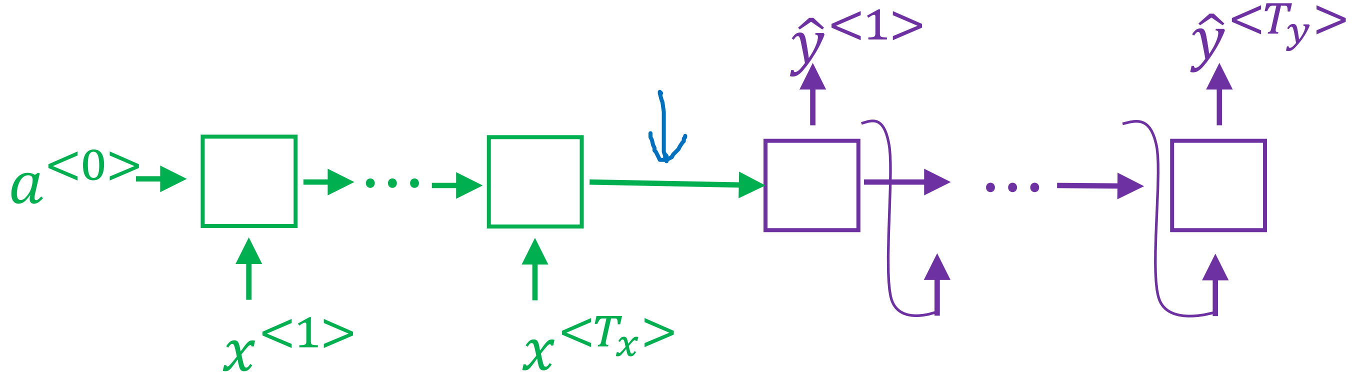
If human needs to translate that whole sentence into the another langurage is that they willn't read the whole paragranph remenber it and then do the tran

Sequence to sequence models

deeplearning.ai

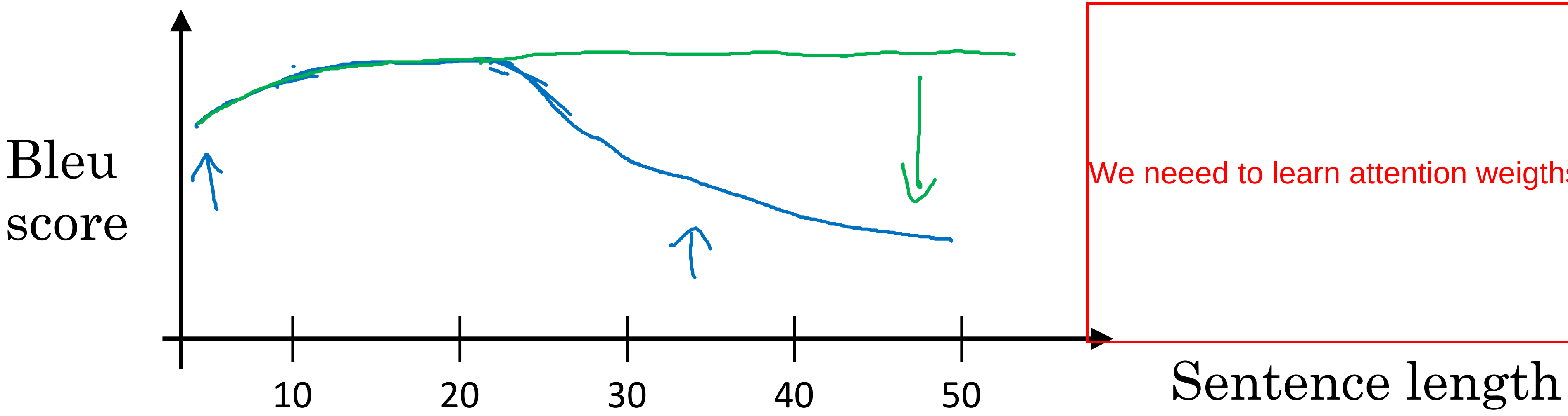
Attention model intuition

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.

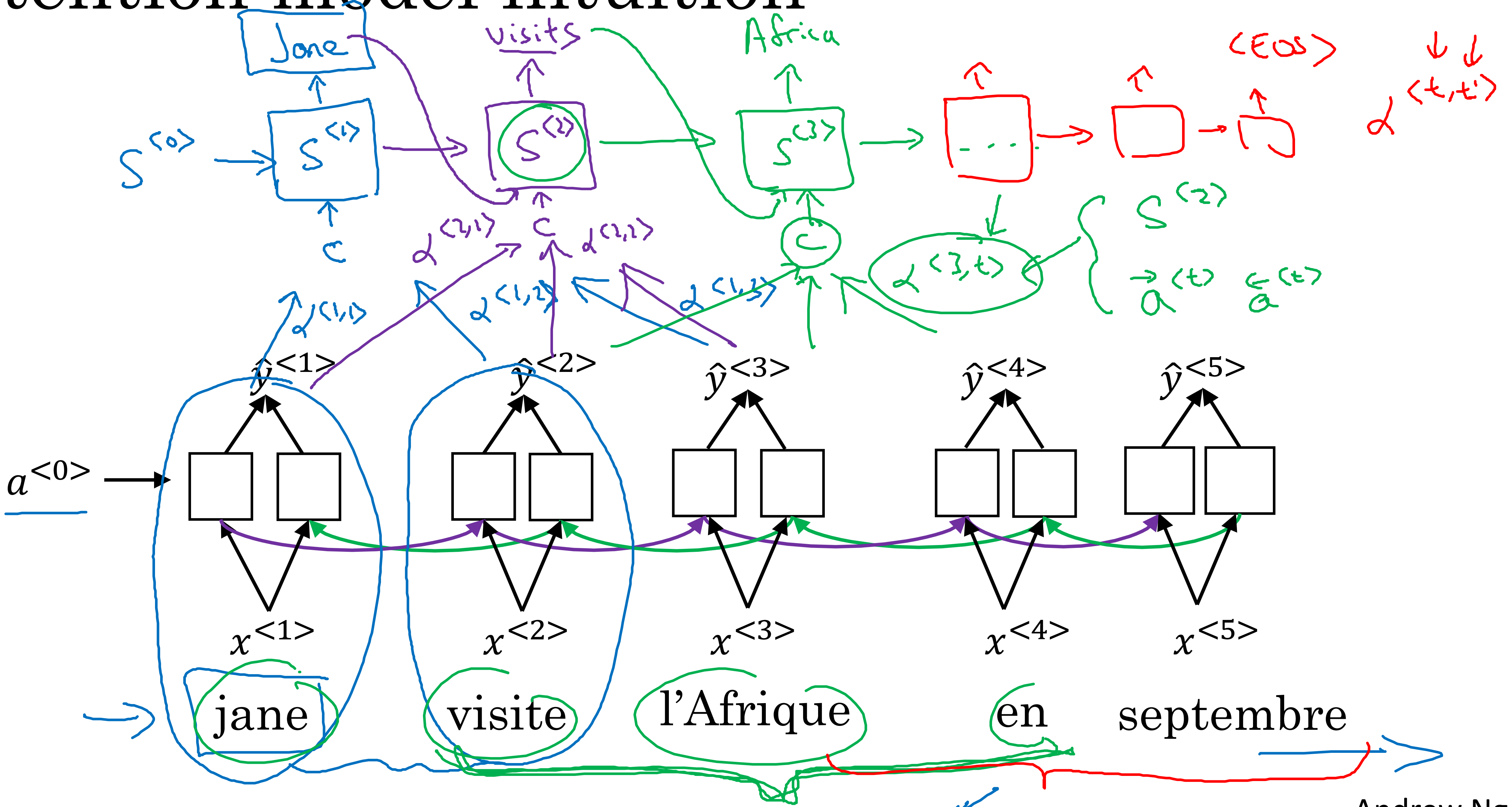
Jane went to Africa last September, and enjoyed the culture and met many wonderful people; she came back raving about how wonderful her trip was, and is tempting me to go too.



We neeed to learn attention weigths alphas denoted below in the slide

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

deeplearning.ai

Attention model

Attention model

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

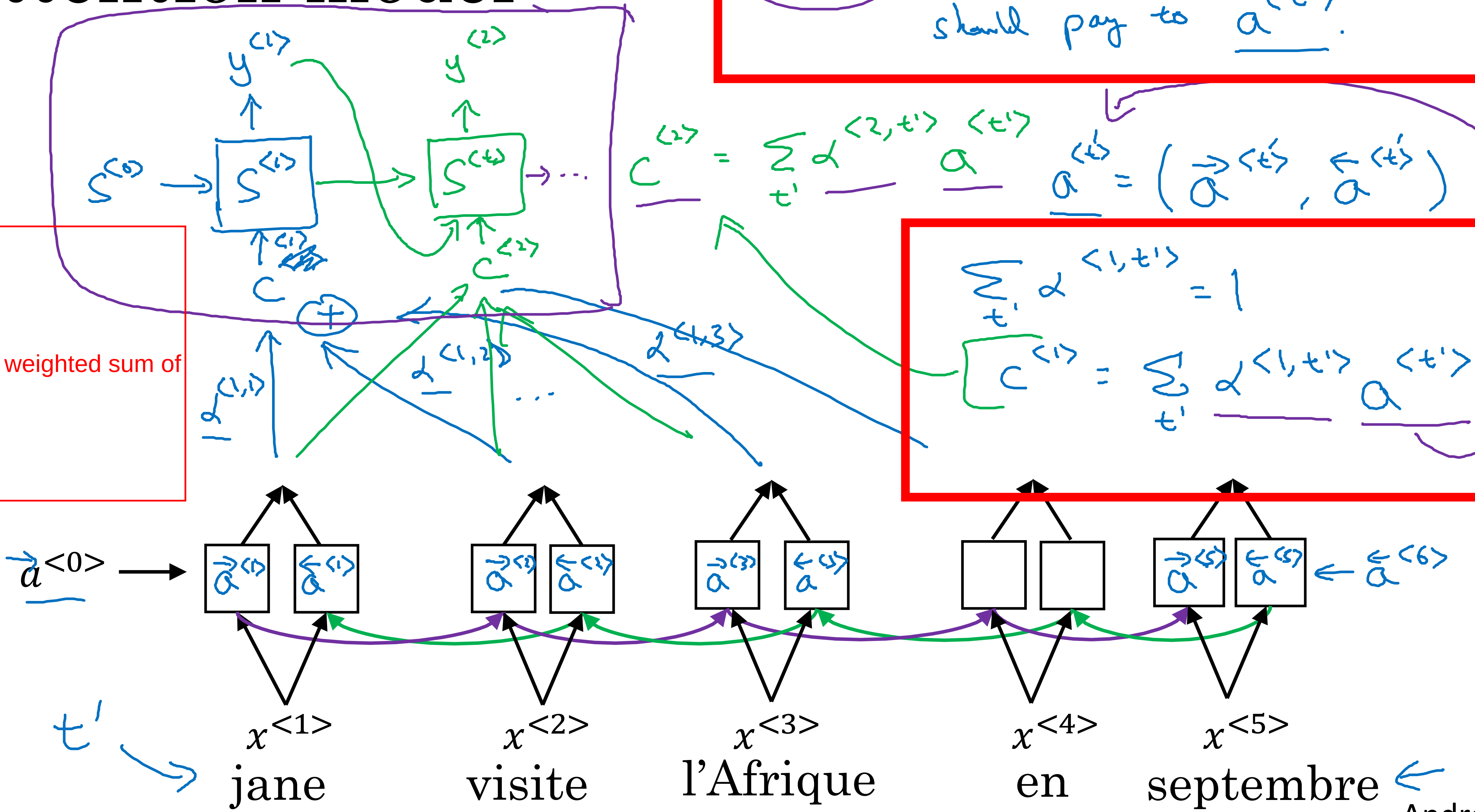
$c^{<2>} = \sum_{t'} \alpha^{<2,t'>} a^{<t'>}$

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

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

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

C is the weighted sum of



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

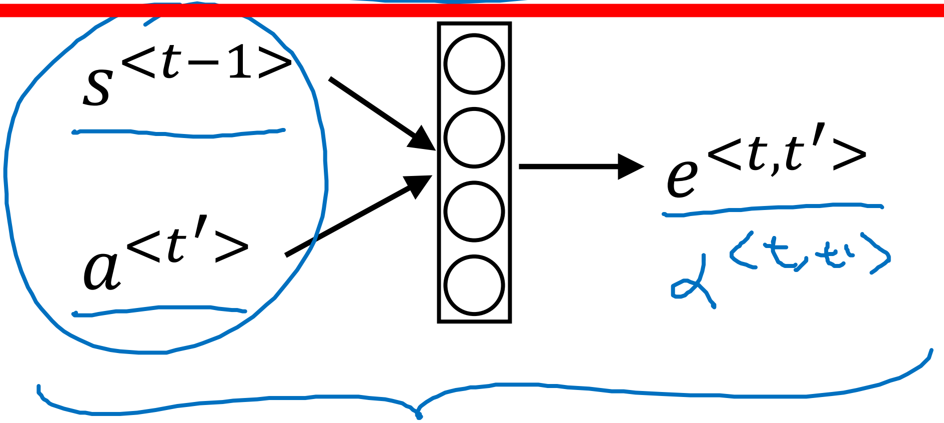
Andrew Ng

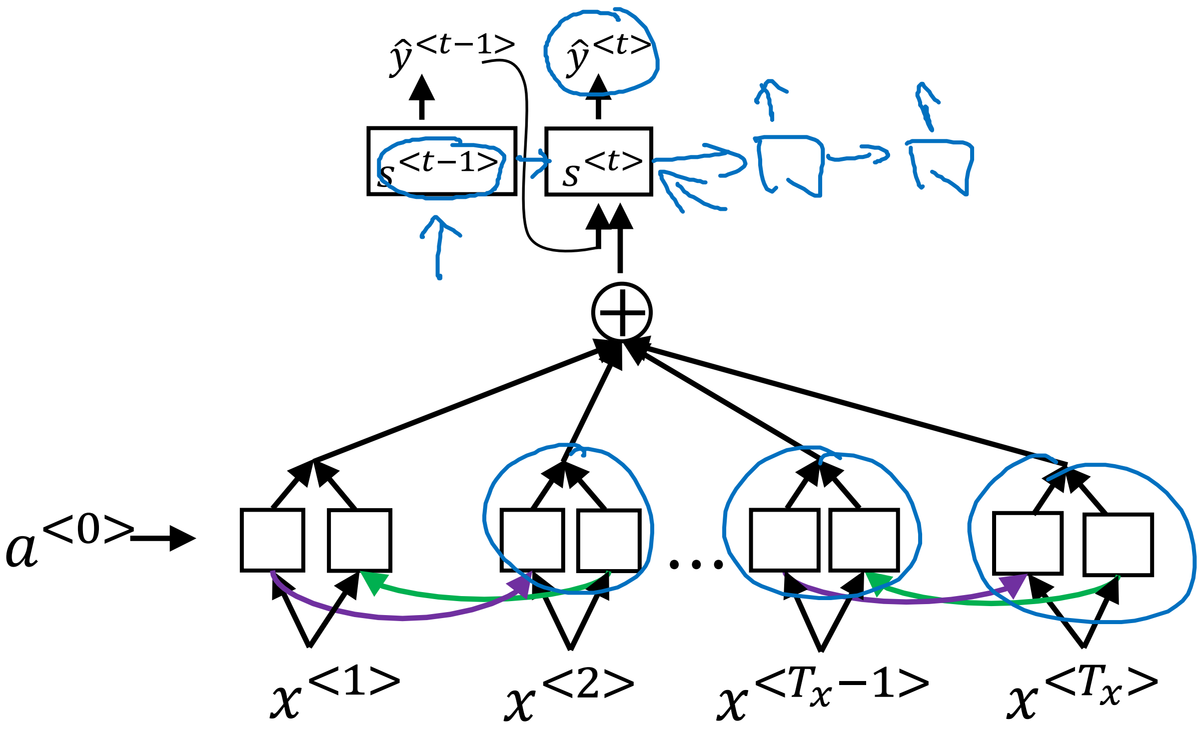
Computing attention $\alpha^{<t,t'>}$

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]

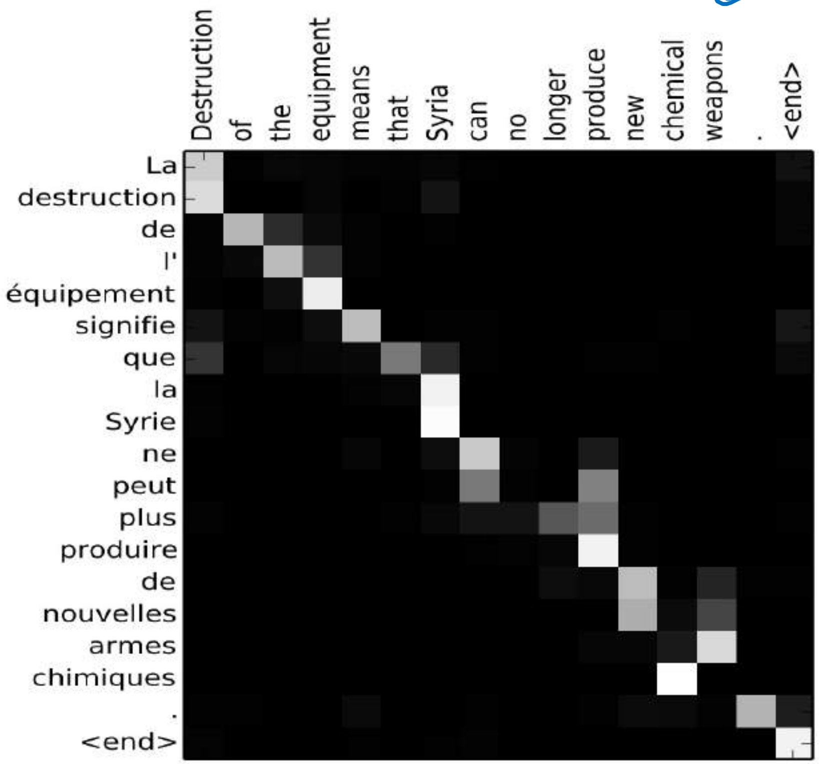
Andrew Ng

Attention examples

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

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

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



Andrew Ng

deeplearning.ai

Audio data

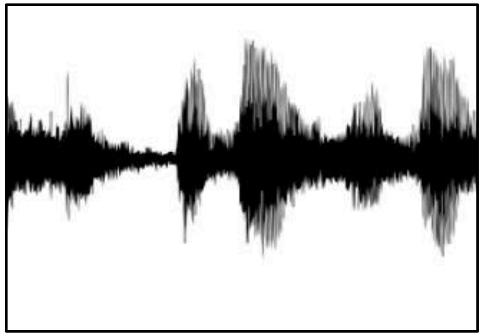
Speech recognition

Speech recognition problem

x ⟶ y

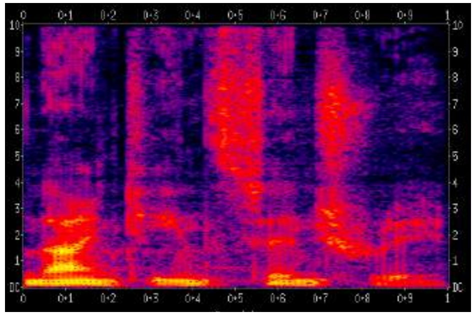
audio clip

transcript



⟶ "the quick brown fox"

⟶ phonemes: de kwik braun



Spectrogram where x is time and y is

300h

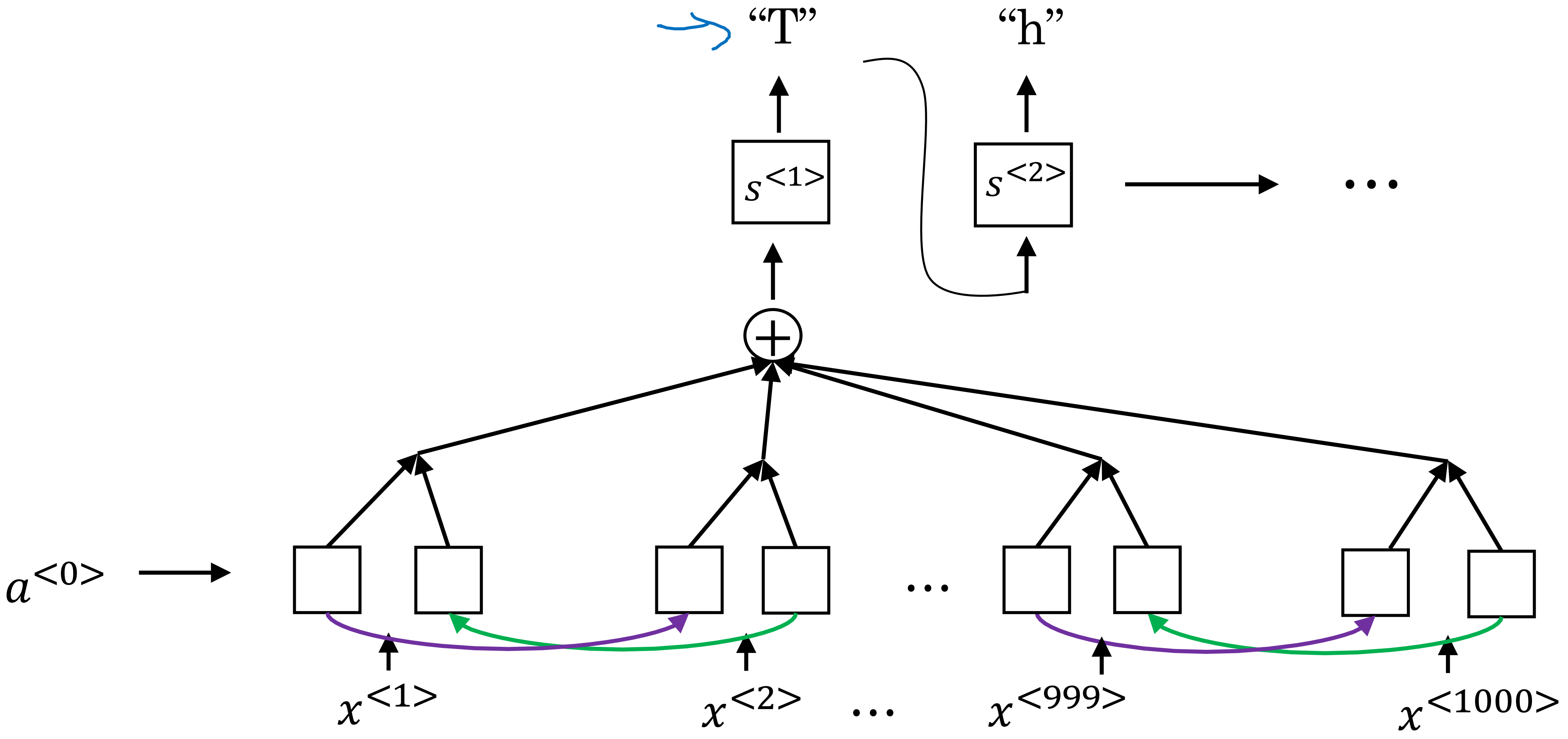
3000h

100,000h

Earlier speech recognistion used to be pe

Andrew Ng

Attention model for speech recognition



Andrew Ng

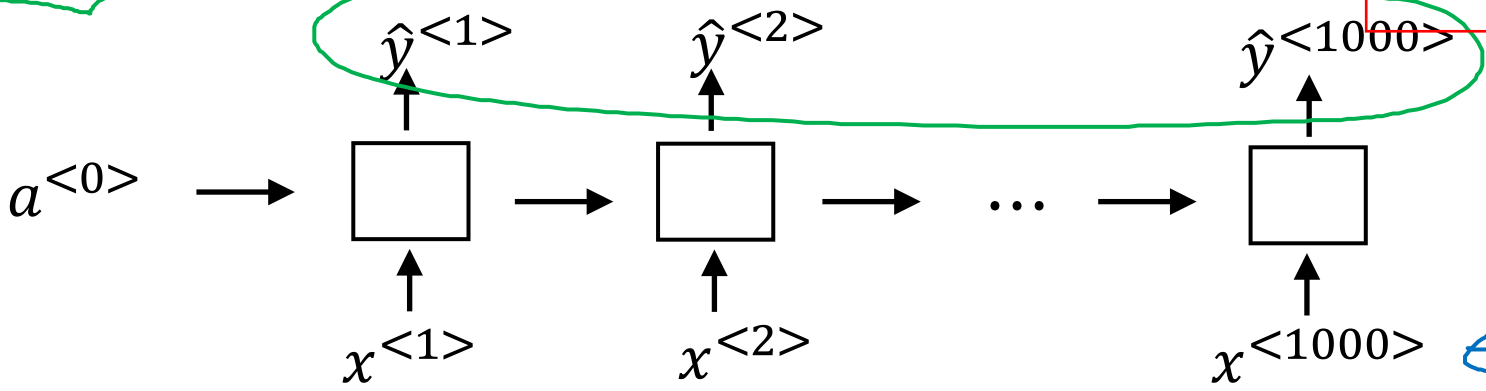
CTC cost for speech recognition

Isme kya hota hai ek large size ka bidirectio

(Connectionist temporal classification)

"the quick brown fox"

19 characters



ttt_h_eee___⎵___qqq__

"space"

"blank"

the⎵q

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?

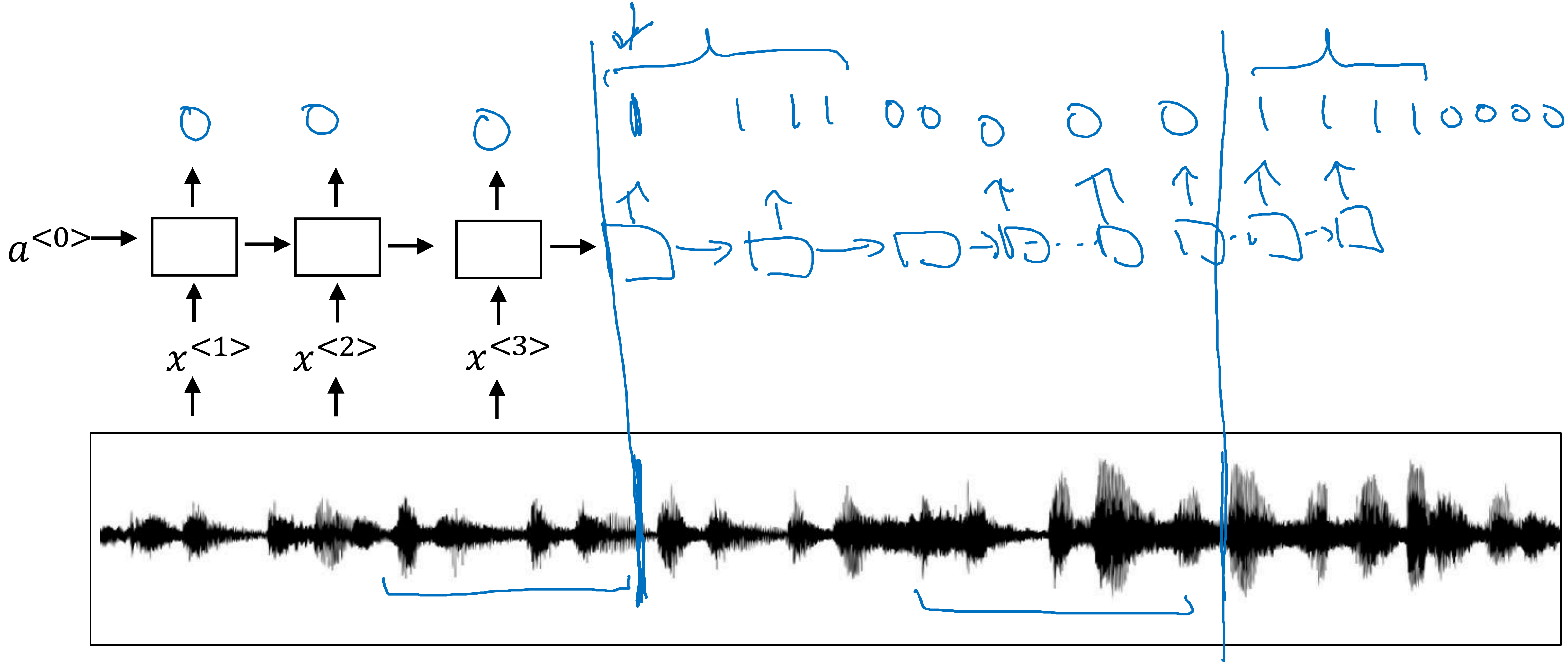
Amazon Echo
(Alexa)

Baidu DuerOS
(xiaodunihao)

Apple Siri
(Hey Siri)

Google Home
(Okay Google)

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

Deep learning is a super power

Please buy this from shutterstock and replace in final video.

Thank you.

-Andrew Ng