Sequence to sequence models

deeplearning.ai

Basic models

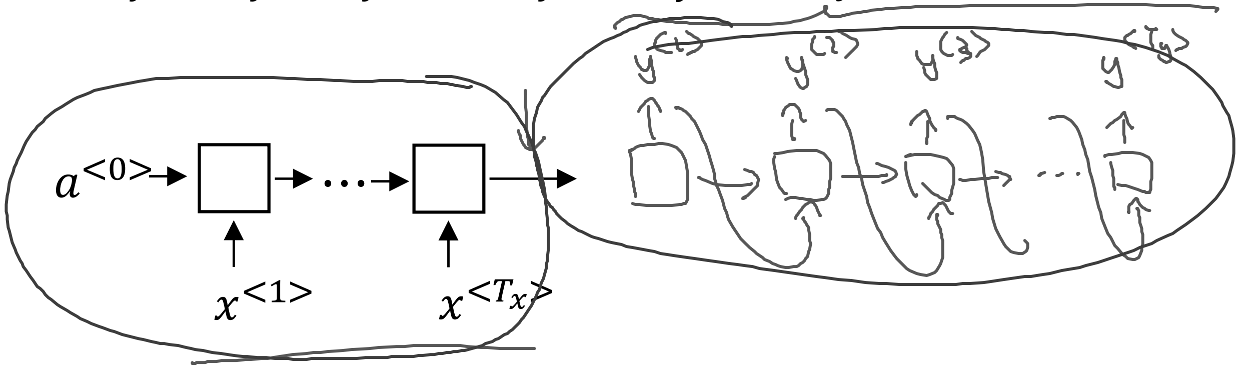
Sequence to sequence model

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

Jane visite l'Afrique en septembre

$\longrightarrow$ Jane is visiting Africa in September.

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



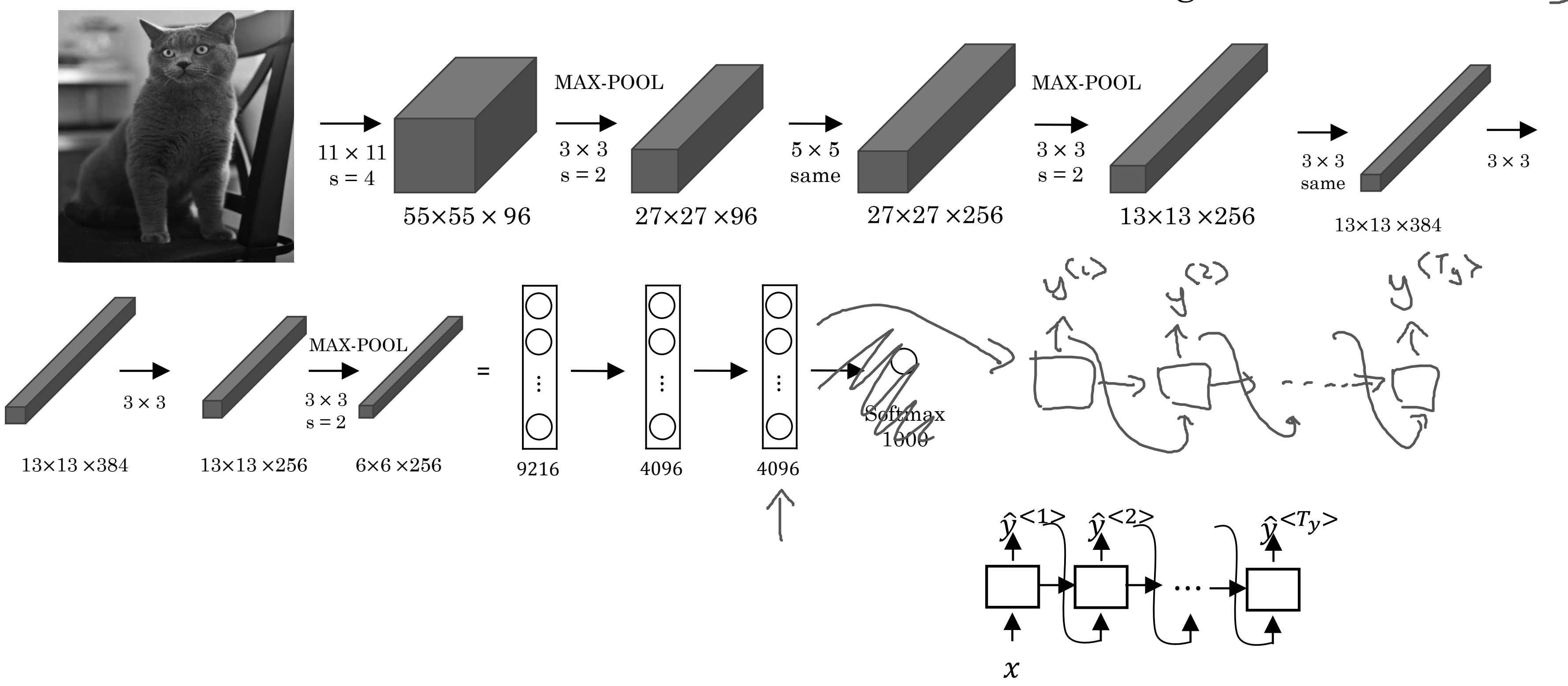
[Sutskever et al., 2014. Sequence to sequence learning with neural networks]

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

Andrew Ng

Image captioning

$y^{<1>}$ $y^{<2>}$ $y^{<3>}$ $y^{<4>}$ $y^{<5>}$ $y^{<6>}$
A cat sitting on a chair



[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]

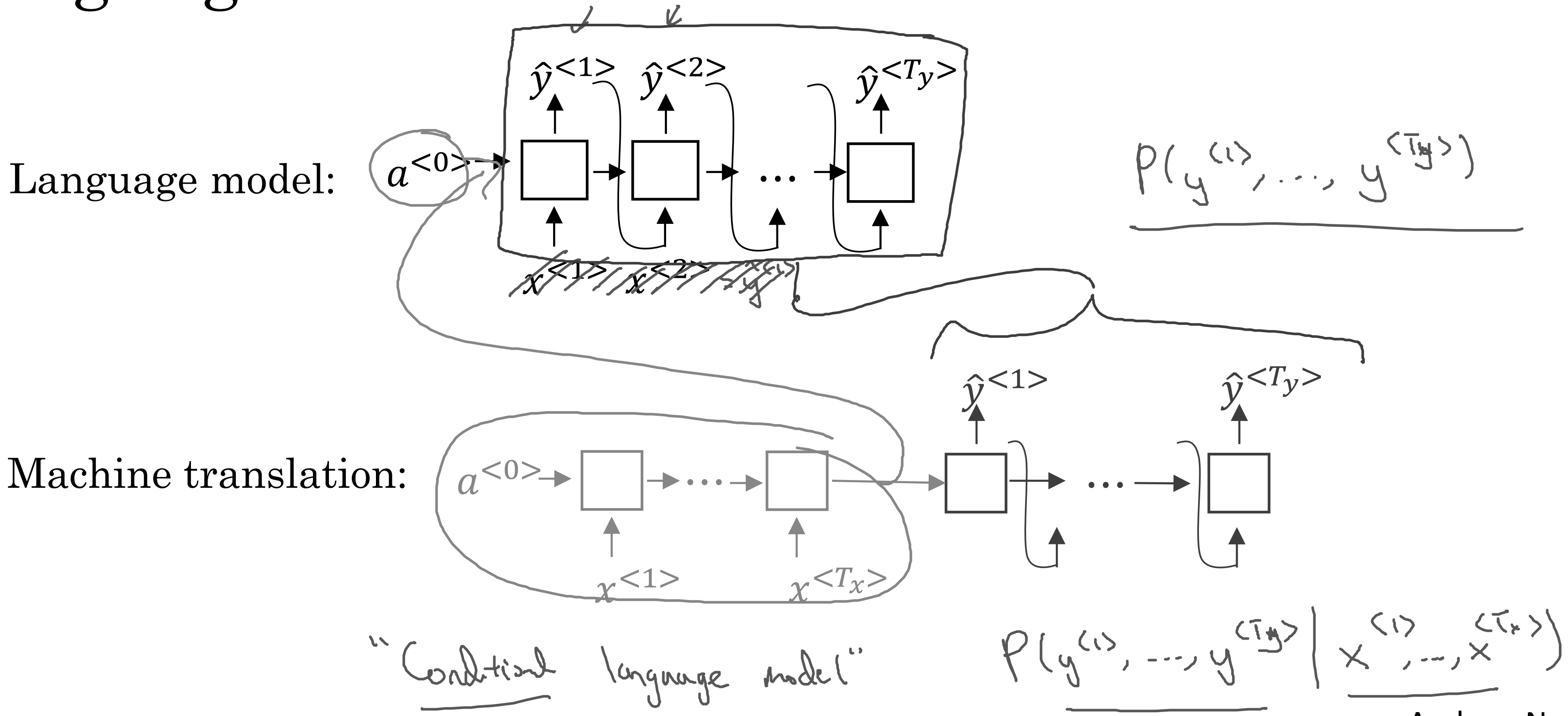
Andrew Ng

deeplearning.ai

Sequence to sequence models

Picking the most likely sentence

Machine translation as building a conditional language model



Andrew Ng

Finding the most likely translation

Jane visite l'Afrique en septembre.

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

→ Jane is visiting Africa in September.

→ Jane is going to be visiting Africa in September.

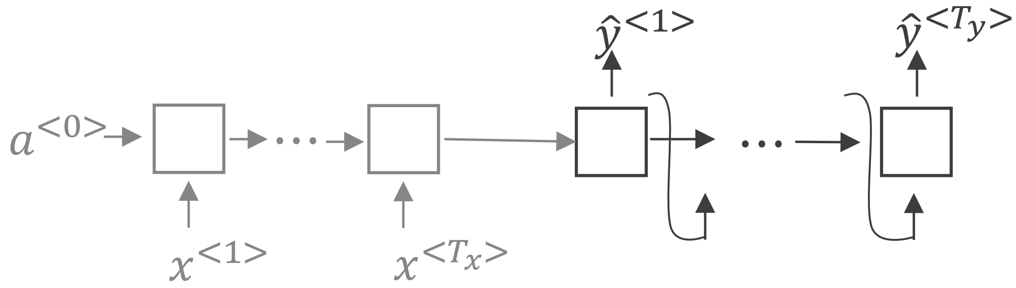
→ In September, Jane will visit Africa.

→ Her African friend welcomed Jane in September.

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

Why not a greedy search?

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



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

10,000

10

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

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

→ Jane is visiting Africa in September.

→ Jane is going to be visiting Africa in September.

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

Andrew Ng

deeplearning.ai

Sequence to sequence models

Beam search

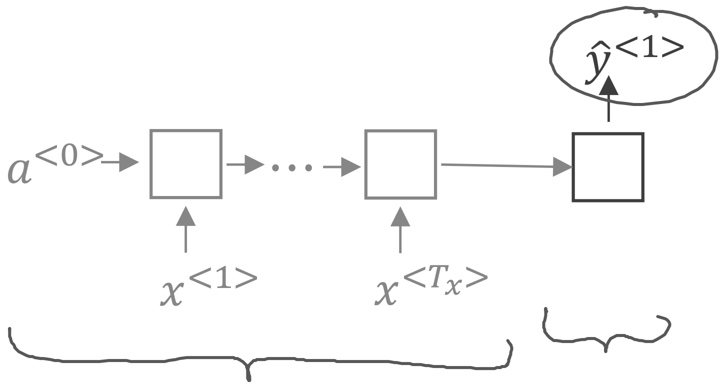
Beam search algorithm

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

Step 1

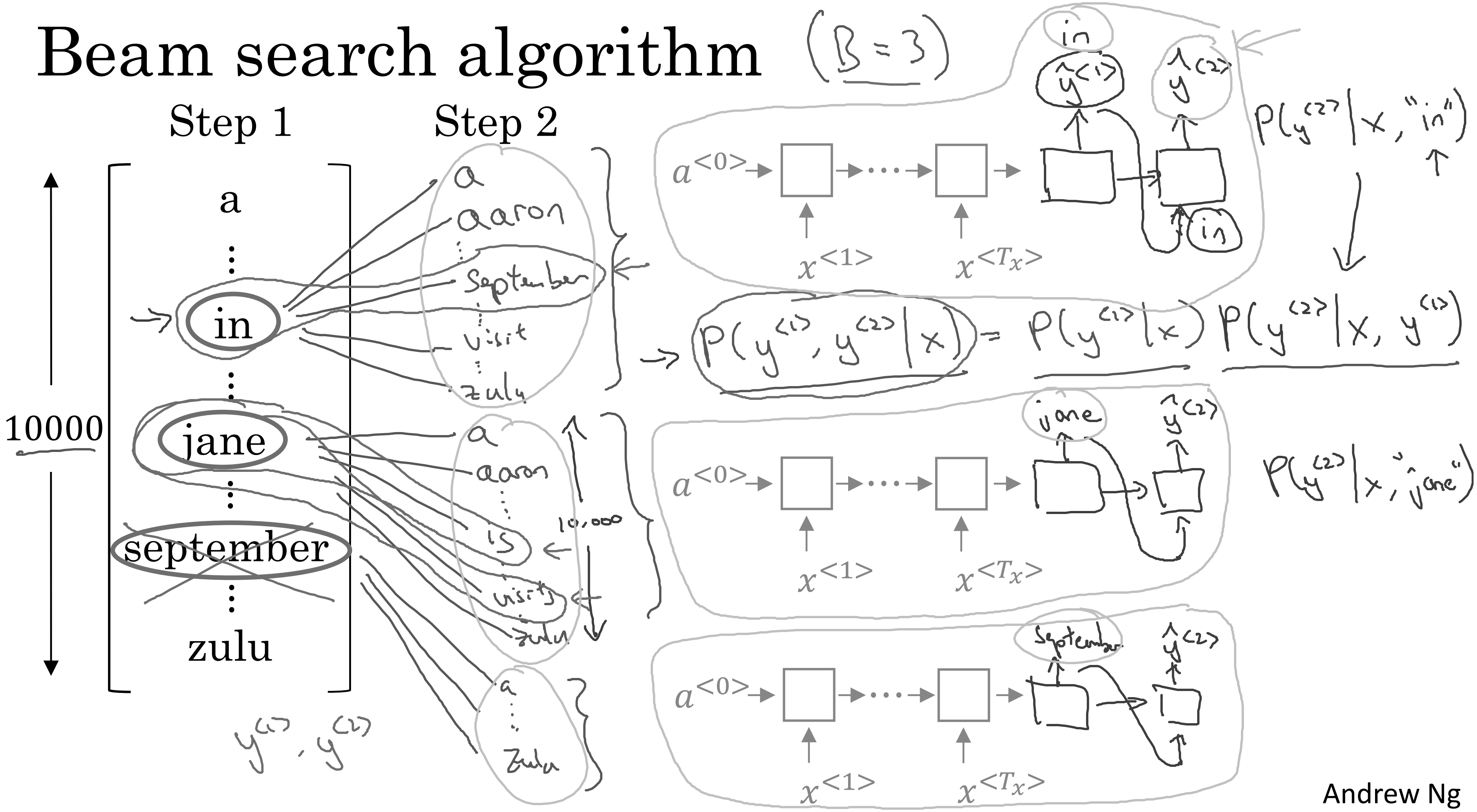
$$10000 \left[\begin{array}{c} \text{a} \\ \vdots \\ \text{in} \\ \vdots \\ \text{jane} \\ \vdots \\ \text{september} \\ \vdots \\ \text{zulu} \end{array}\right]$$

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



Beam search algorithm

(B = 3)



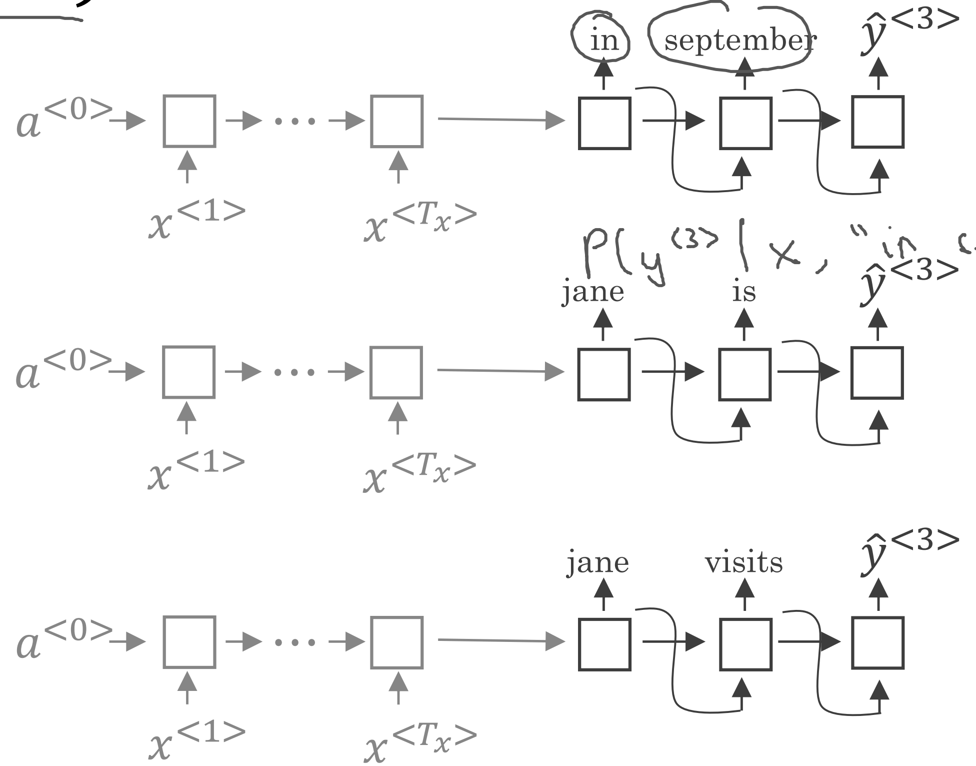
Beam search ($B = 3$)

$B = 1 \rightsquigarrow$ greedy search

in september
a
aaron
jane
zulu

jane is
a
visits
zulu

jane visits
a
africa
zulu



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

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

jane visits africa in september. <EOS>

Andrew Ng

deeplearning.ai

Sequence to sequence models

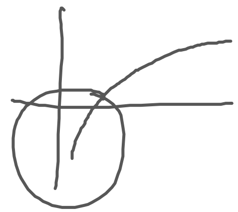
Refinements to beam search

Length normalization

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

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

log



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

$P(y|x) \leftarrow$

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

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

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

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

Andrew Ng

Beam search discussion

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

Beam width B?

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

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

deeplearning.ai

Sequence to sequence models

Error analysis on beam search

Example

→ RNN

→ Beam Search

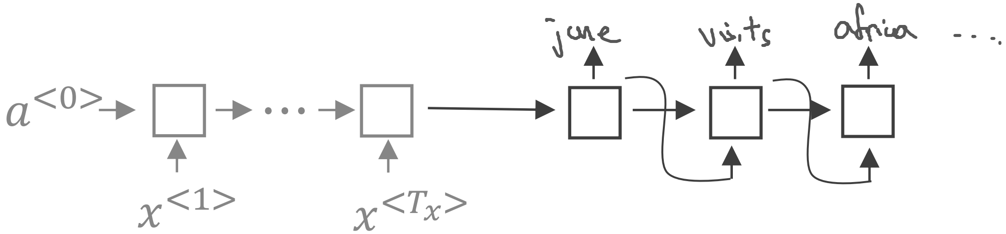
$\boxed{B\uparrow}$

Jane visite l'Afrique en septembre.

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

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

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



Andrew Ng

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

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

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

Conclusion: Beam search is at fault.

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

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

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	$\frac{4}{6}$
cat on	1	1	
on the	1	1	
the mat	1	1	

[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. $(\hat{y})$

$P_1, P_2, = \underline{1.0}$

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

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

$\sum_{unigram \in \hat{y}} count_{clip}(unigram)$

$\sum_{unigram \in \hat{y}} Count\,(unigram)$

$\underline{\text{Unigram}}$

$\underline{\text{n-gram}}$

$\sum_{n\text{-grams} \in \hat{y}} Count_{clip}\,(n\text{-gram})$

$\sum_{n\text{-grams} \in \hat{y}} Count\,(n\text{-gram})$

[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

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

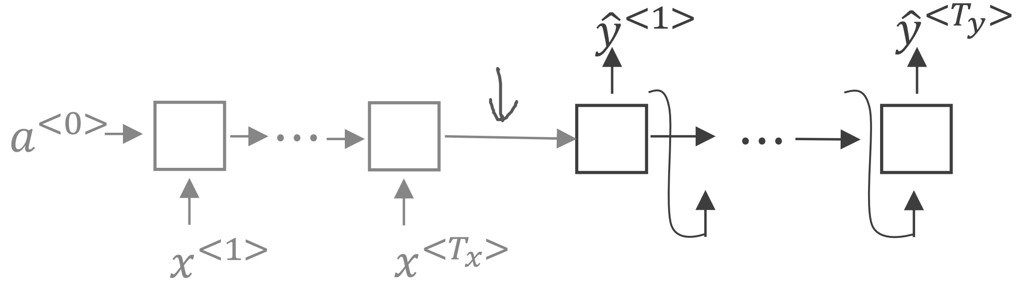
[Papineni et. al., 2002. Bleu: A method for automatic evaluation of machine translation]

deeplearning.ai

Sequence to sequence models

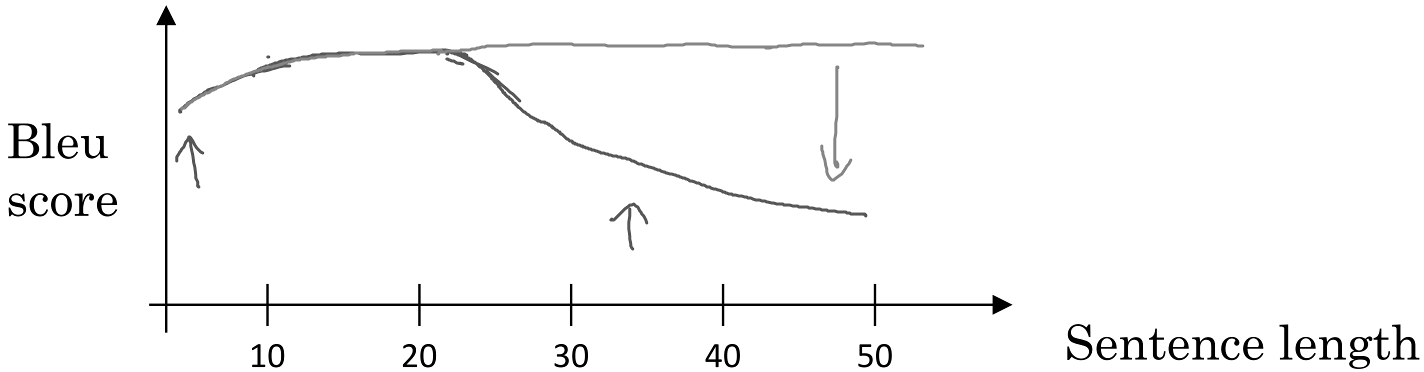
Attention model intuition

The problem of long sequences



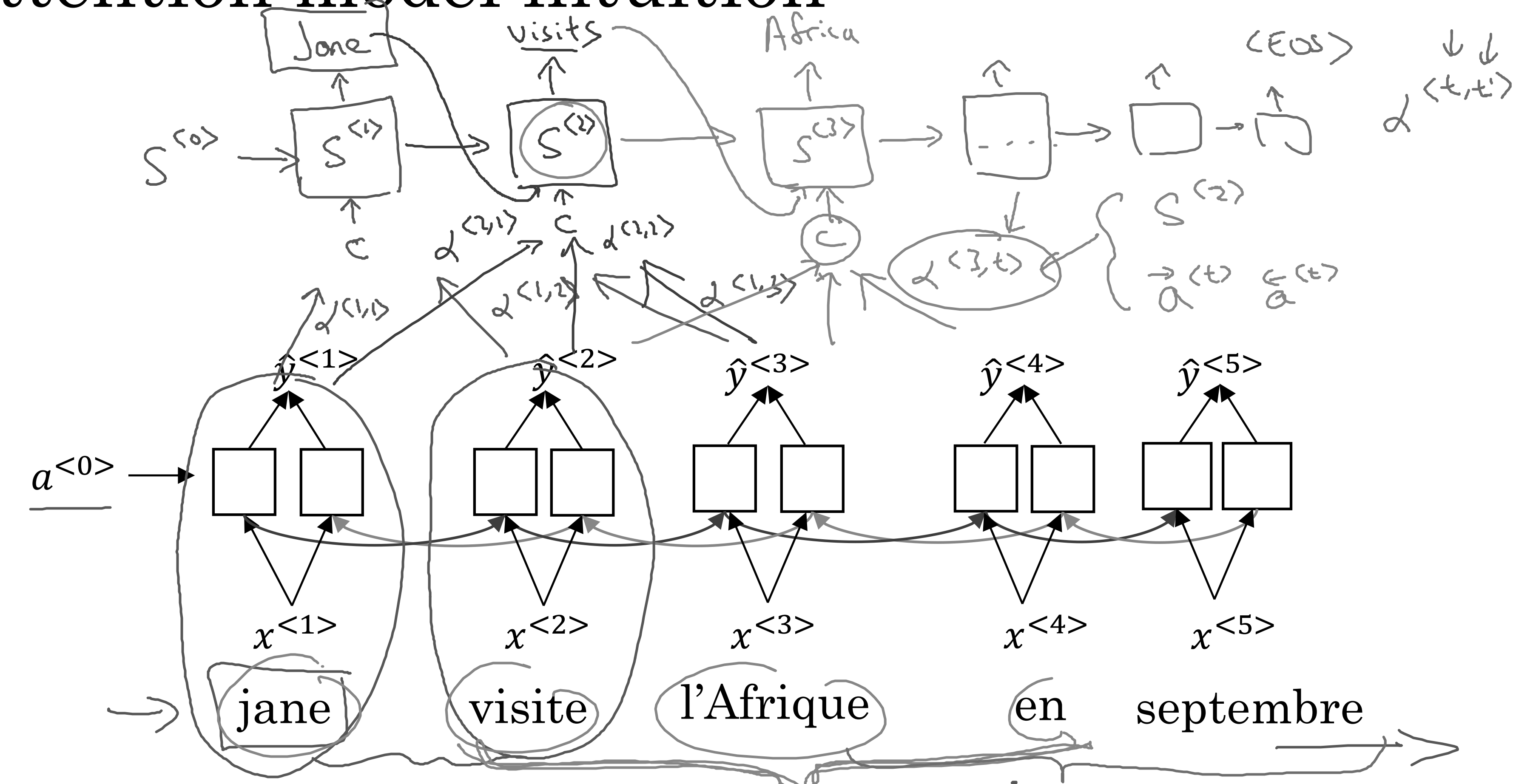
Jane s'est rendue en Afrique en septembre dernier, a apprécié la culture et a rencontré beaucoup de gens merveilleux; elle est revenue en parlant comment son voyage était merveilleux, et elle me tente d'y aller aussi.

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]

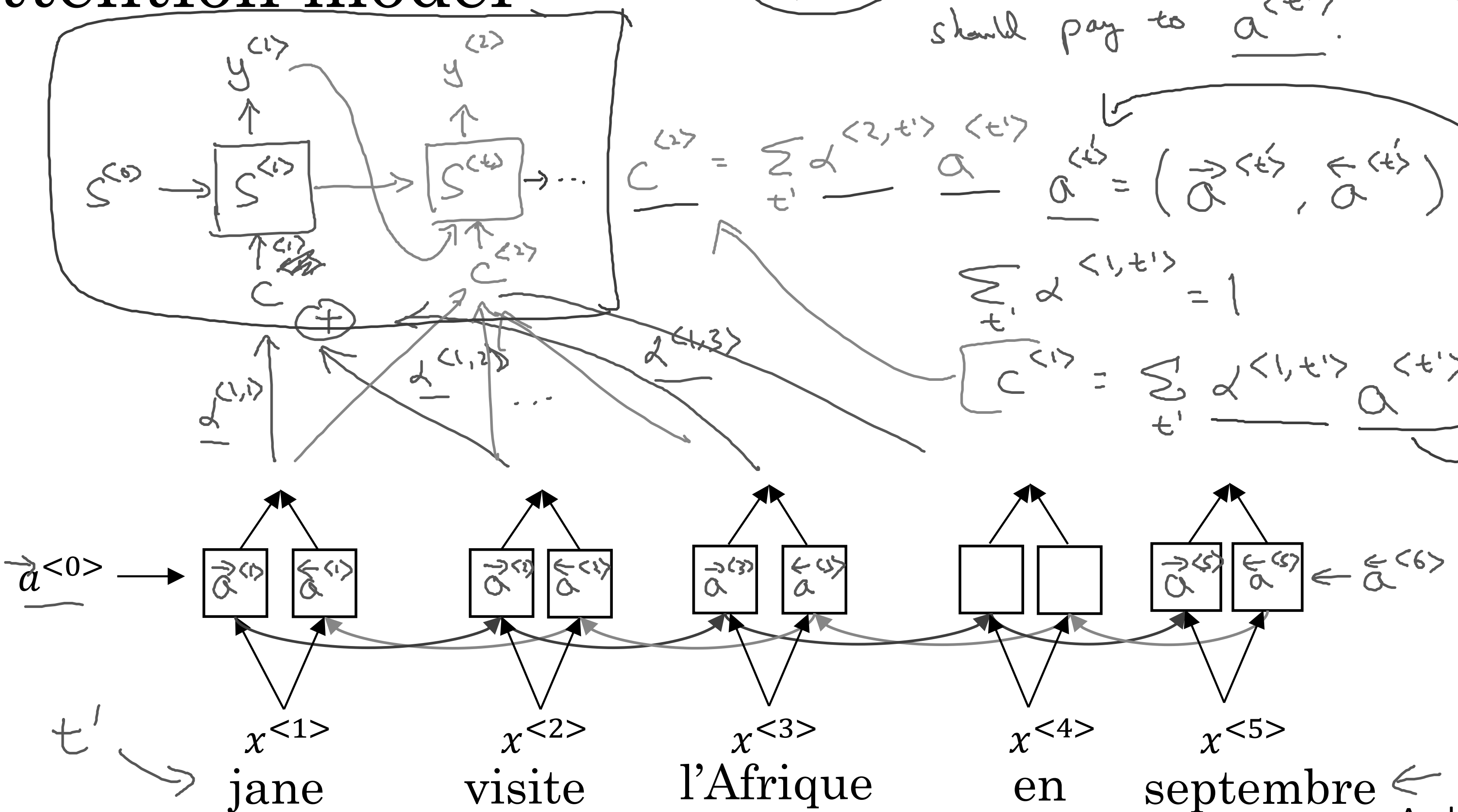
deeplearning.ai

Sequence to sequence models

Attention model

Attention model

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



$c^{\langle 2\rangle} = \sum_{t'} \alpha^{\langle 2,t'\rangle} a^{\langle t'\rangle}$

$a^{\langle t'\rangle} = (\vec{a}^{\langle t'\rangle}, \overleftarrow{a}^{\langle t'\rangle})$

$\sum_{t'} \alpha^{\langle 1,t'\rangle} = 1$

$\boxed{c^{\langle 1\rangle}} = \sum_{t'} \alpha^{\langle 1,t'\rangle} a^{\langle t'\rangle}$

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

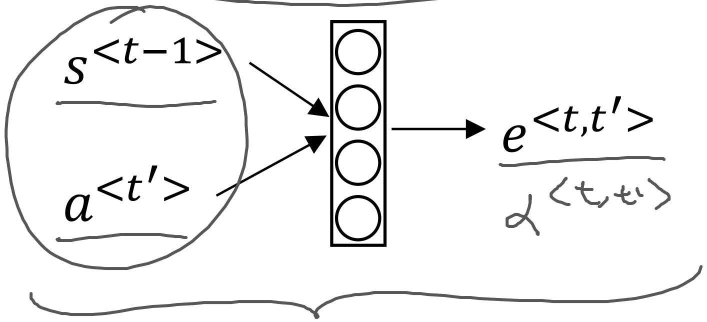
Andrew Ng

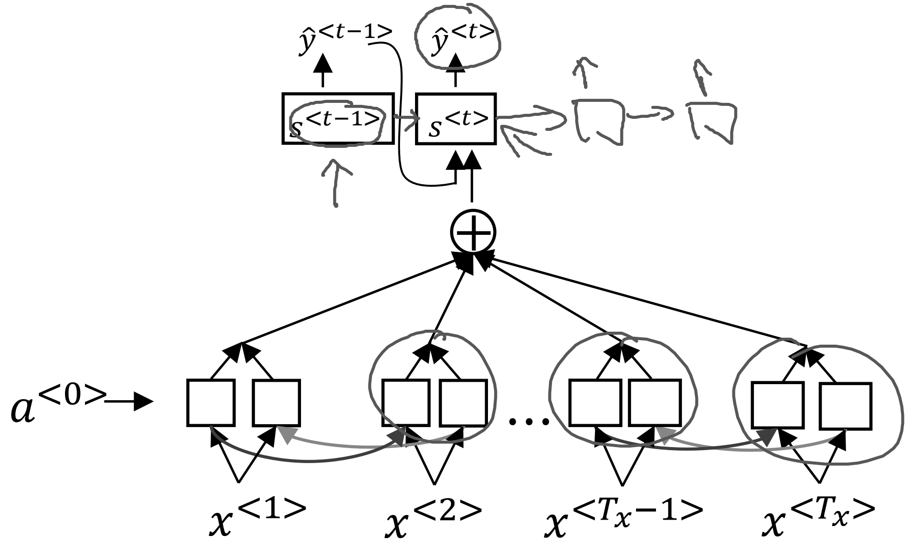
Computing attention $\alpha^{<t,t'>}$

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]

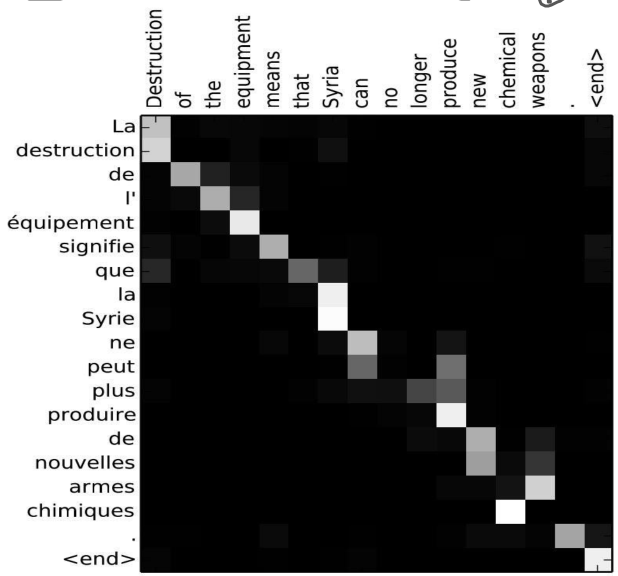
Andrew Ng

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

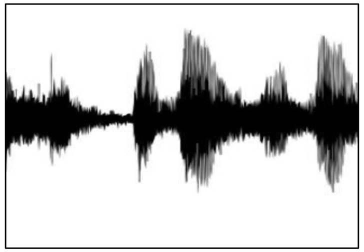
Speech recognition

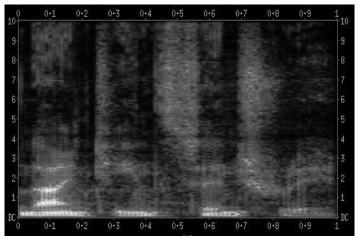
Speech recognition problem

$x \longrightarrow y$

audio clip

transcript





"the quick brown fox"

phonemes: de kwik braun

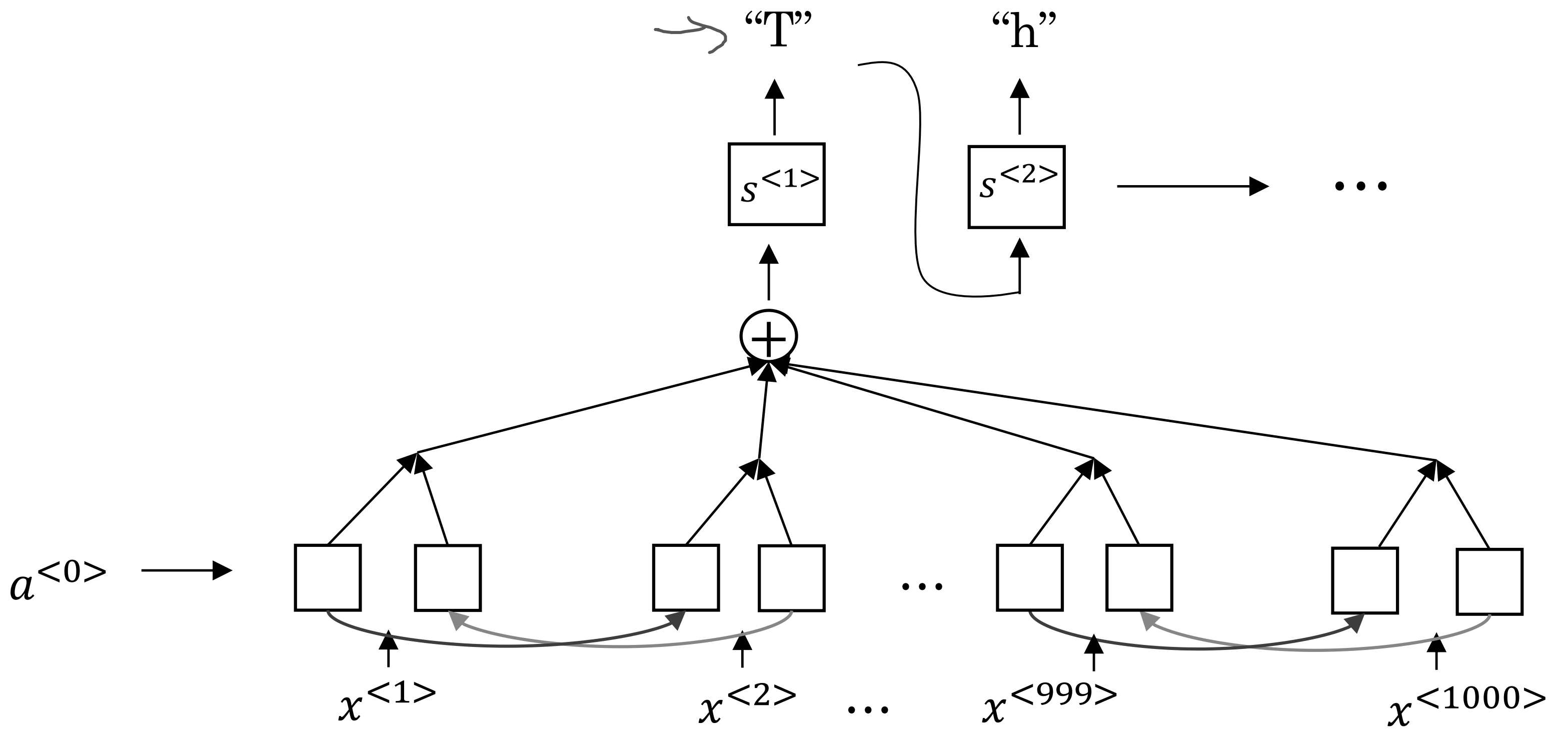
300h

3000h

100,000h

Andrew Ng

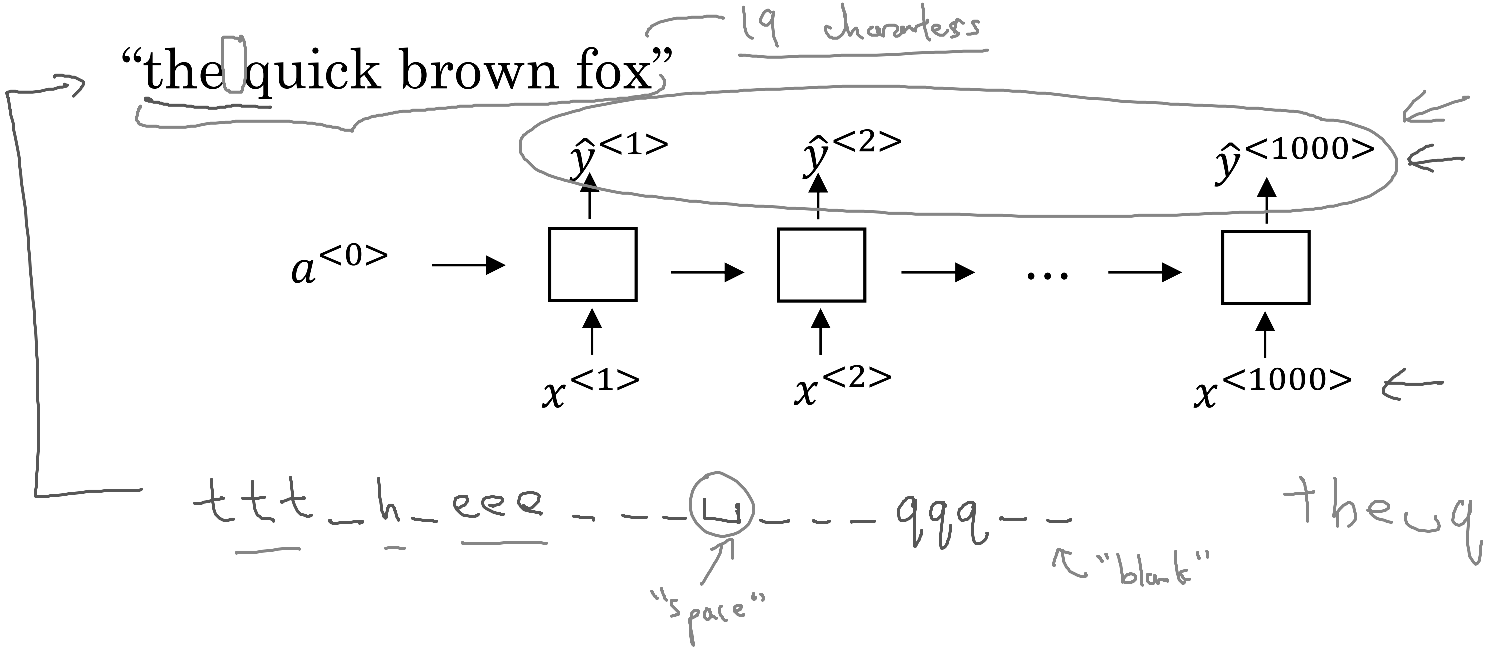
Attention model for speech recognition



Andrew Ng

CTC cost for speech recognition

(Connectionist temporal classification)



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)

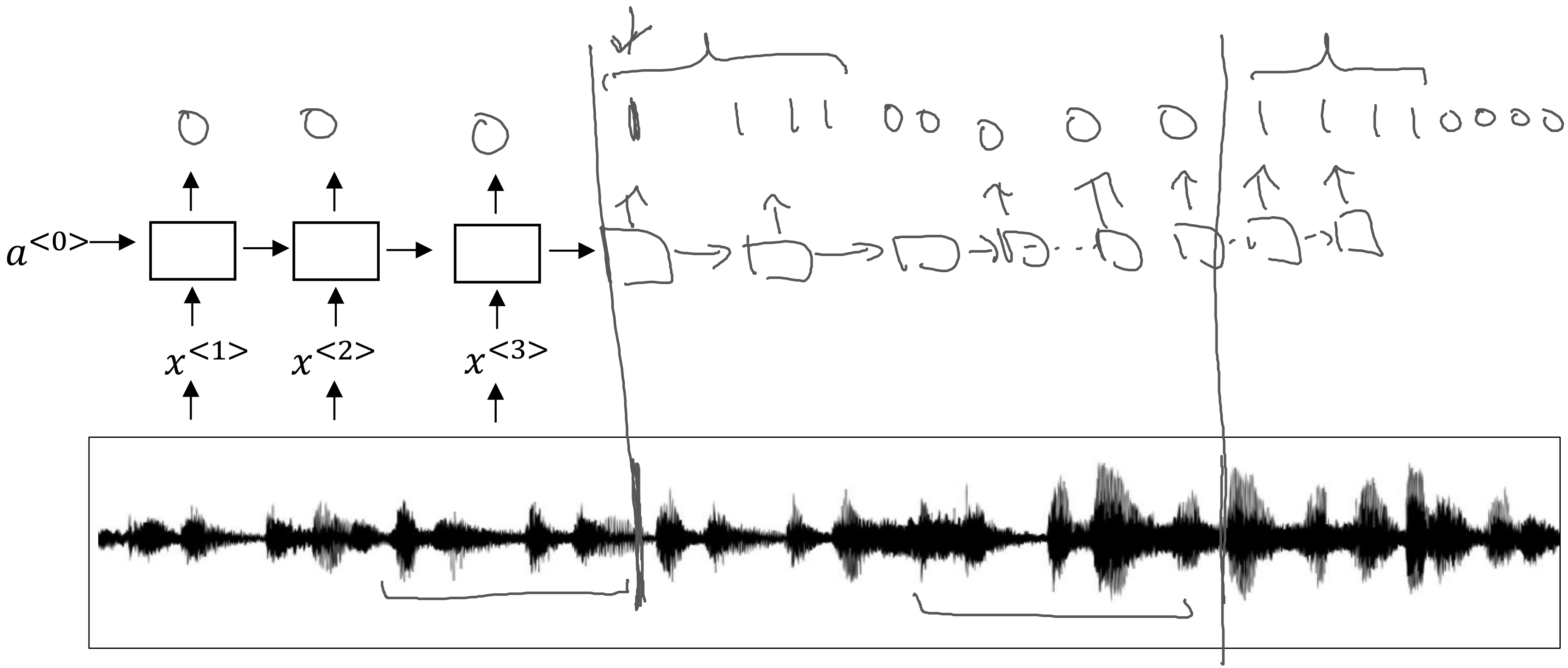
Baidu DuerOS
(xiaodunihao)

Apple Siri
(Hey Siri)

Google Home
(Okay Google)

Andrew Ng

Trigger word detection algorithm



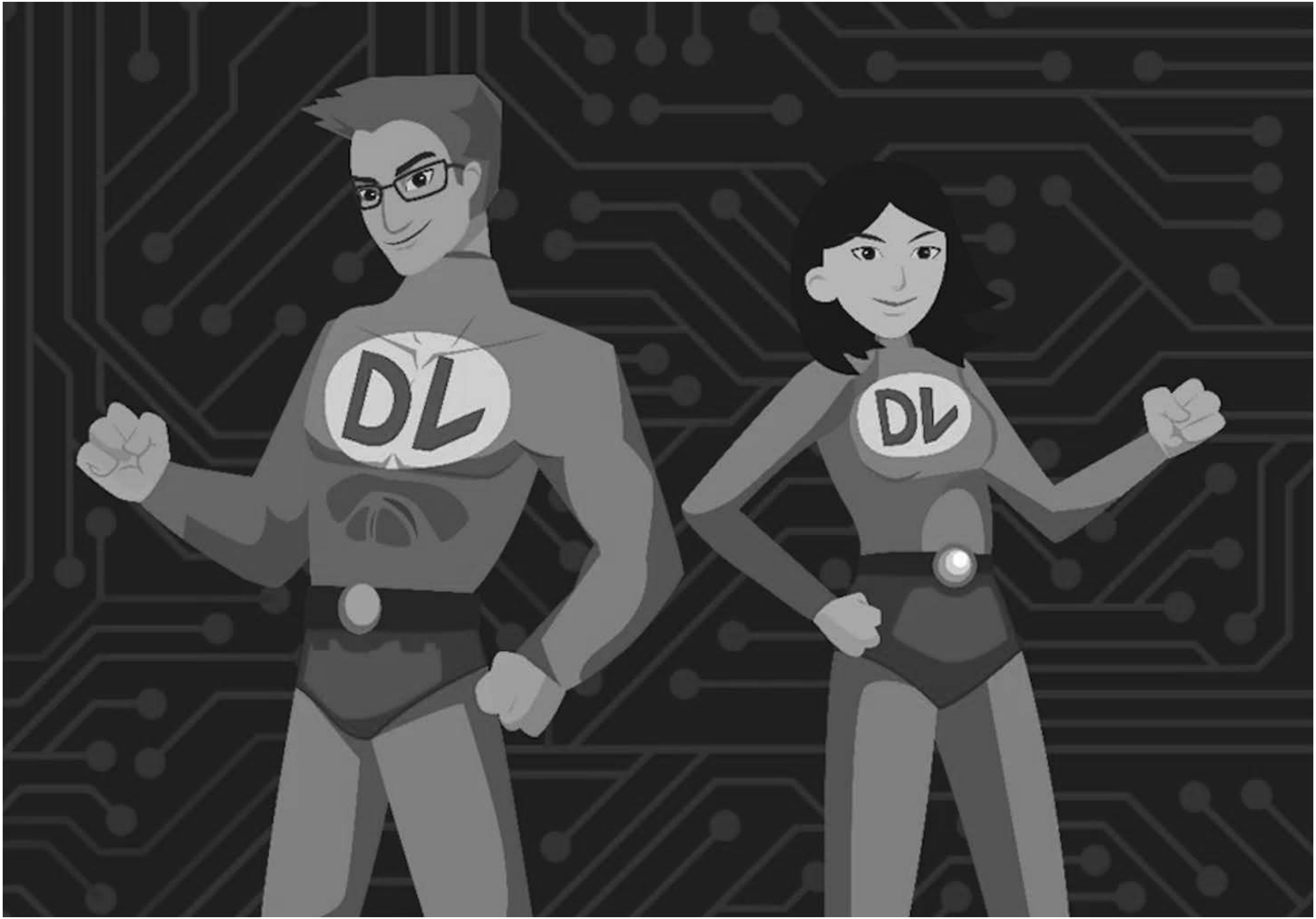
deeplearning.ai

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


DL
DL

Thank you.

-Andrew Ng