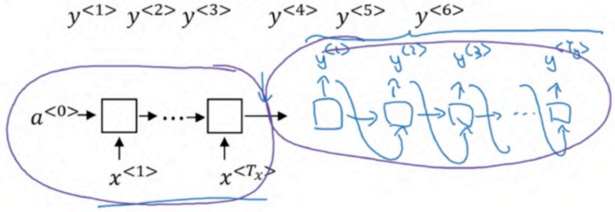
Sequence to sequence model

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

Jane visite l'Afrique en septembre

$\longrightarrow$ Jane is visiting Africa in September.

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



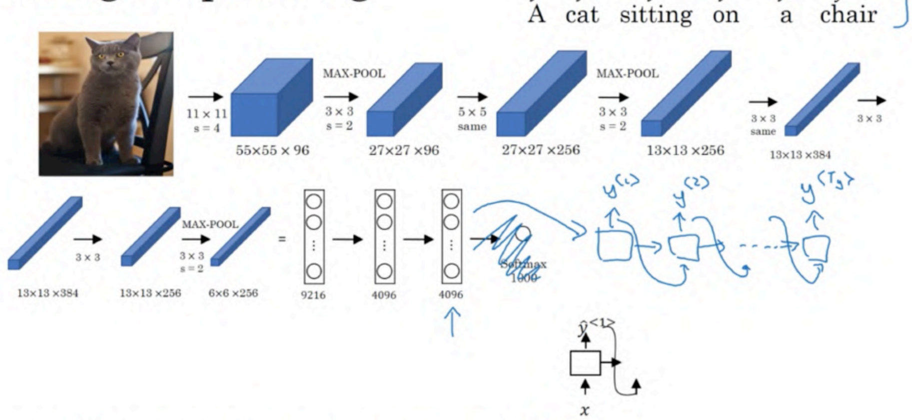
[Sutskever et al., 2014. Sequence to sequence learning with neural networks]
[Cho et al., 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation]

Andrew Ng

Image captioning

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

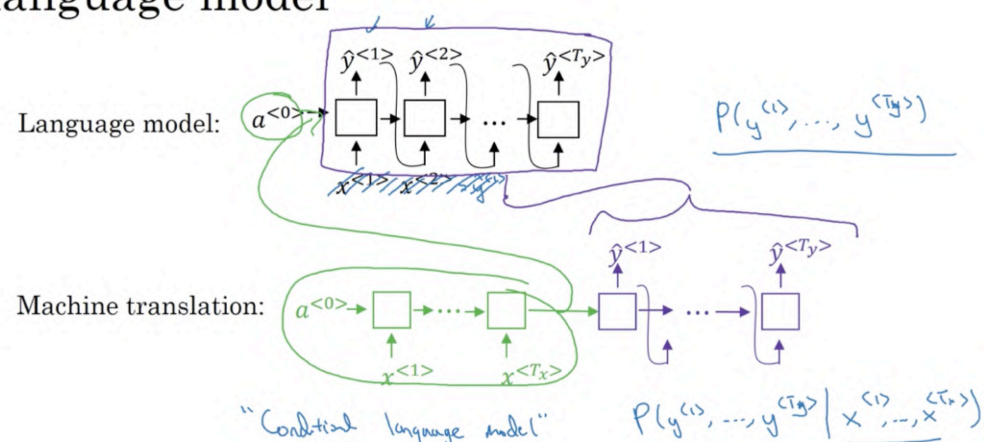
A cat sitting on a chair



[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

Machine translation as building a conditional language model



"Conditional language model"

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

Andrew Ng

Finding the most likely translation

Jane visite l'Afrique en septembre. $P(\overbrace{y^{<1>}, \dots, y^{<T_y>}}^{\text{English}} \mid \overset{\text{French}}{x})$

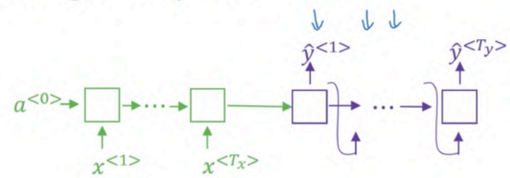
- → Jane is visiting Africa in September.
- → Jane is going to be visiting Africa in September.
- → In September, Jane will visit Africa.
- → Her African friend welcomed Jane in September.

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

Why not a greedy search?

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

Considers only 1 word at a time



10,000

10

$10{,}000^{10}$

$P(y \mid x)$

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

- → Jane is visiting Africa in September.
- → Jane is going to be visiting Africa in September.

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

Beam search algorithm

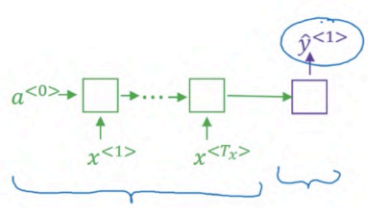
$\underline{B = 3}$ (beam width)

Step 1

10000

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

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



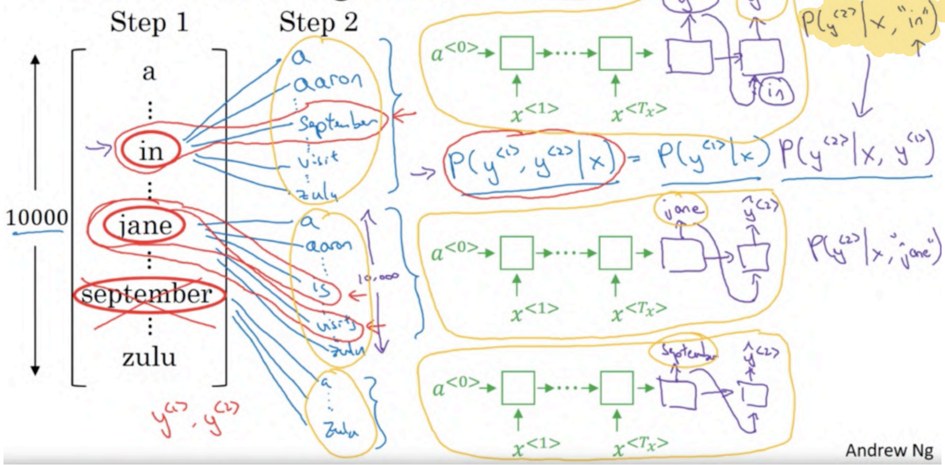
instantiates B copies of the network

Beam width 1

Considers B words!

Beam search algorithm

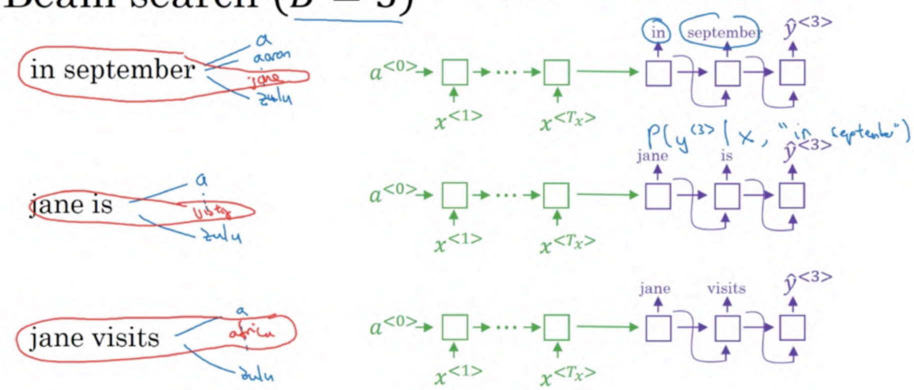
$(B=3)$



Andrew Ng

Beam search $(B = 3)$

$B=1 \leadsto$ greedy search



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

jane visits africa in september. <EOS>

Andrew Ng

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

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

$P(y|x) \leftarrow$

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

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

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

$\alpha = 0.7$

$\alpha = 1$

$\alpha = 0$

Andrew Ng

Beam search discussion

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

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

Andrew Ng

Length normalization

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

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

numbers less than 1

log

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

more numerically stable with logs

$\log P(y|x)$ ←

$P(y|x)$ ←

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

average of log probabilities

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

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

Andrew Ng

Beam search discussion

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

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

Andrew Ng

Example

→ RNN

→ Beam Search

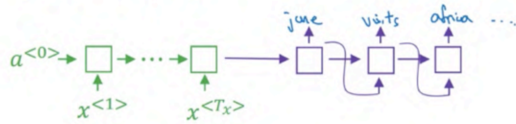
$\boxed{BT}$

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. $(\hat{y})$

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: $P(y^*|x) \leq P(\hat{y}|x)$ ←

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

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.

2 appearances

Bleu

bilingual evaluation understudy

Precision: $\frac{7}{7}$

Modified precision: $\frac{2}{7}$ ← $\frac{\text{Count}_{\text{clip}}(\text{"the"})}{\text{Count}(\text{"the"})}$

[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	$\text{Count}_{\text{clip}}$
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. $(\hat{y})$

$P_1, P_2, = 1.0$

$$P_1 = \frac{\sum_{\text{unigrams} \in \hat{y}} \text{Count}_{\text{clip}}(\text{unigram})}{\sum_{\text{unigram} \in \hat{y}} \text{Count}(\text{unigram})}$$

unigram

n-gram

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

[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 $\quad p_1, p_2, p_3, p_4$

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

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

Bleu details

p_n = Bleu score on n-grams only $\quad p_1, p_2, p_3, p_4$

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

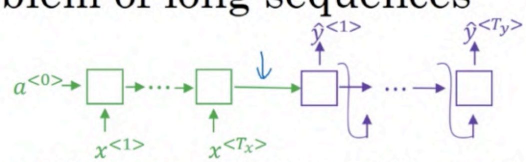
BP = brevity penalty

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

$\exp(1 - \text{reference\_output\_length/MT\_output\_length})$

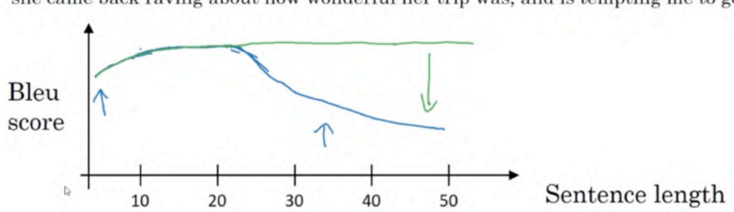
[Papineni et. al., 2002. Bleu: A method for automatic evaluation of machine translation] Andrew Ng

The problem of long sequences



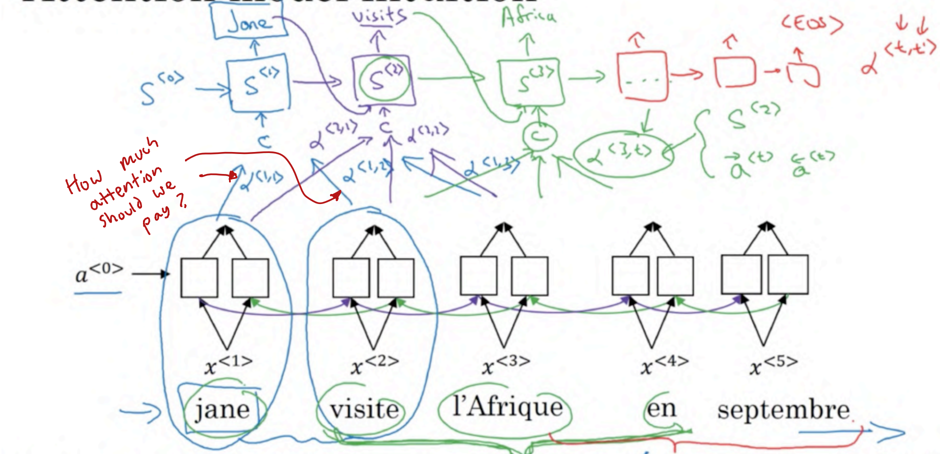
Jane s'est rendue en Afrique en septembre dernier, a apprécié la culture et a rencontré beaucoup de gens merveilleux; elle est revenue en parlant comment son voyage était merveilleux, et elle me tente d'y aller aussi.

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

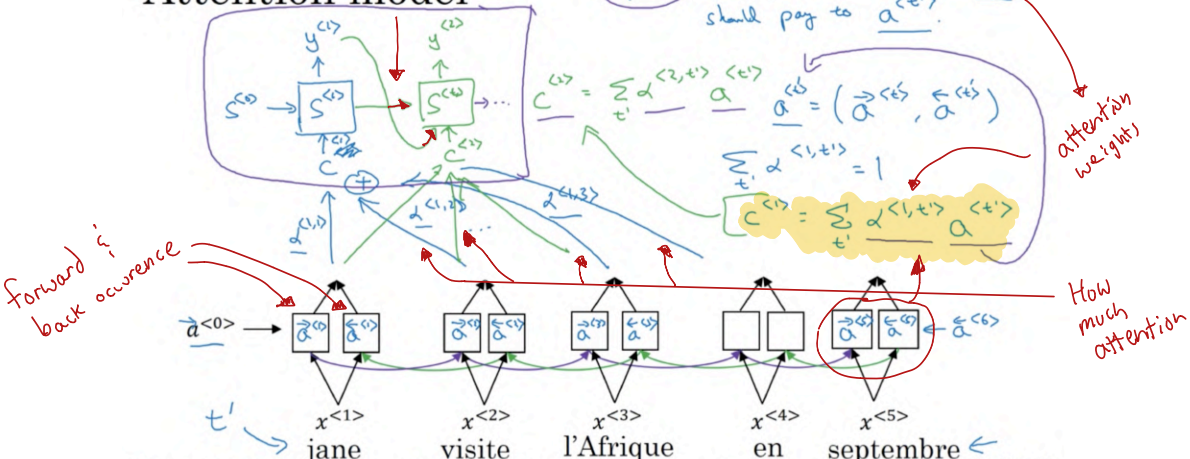
Attention model intuition



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

Andrew Ng

Attention model



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

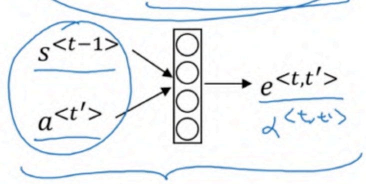
Andrew Ng

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

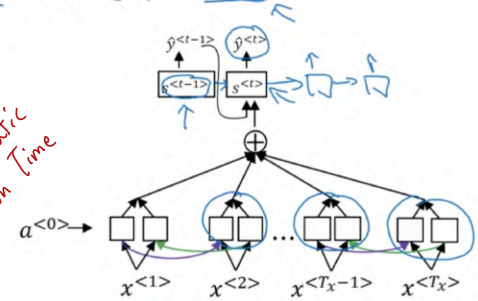
T_x T_y

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

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



Quadratic Run Time



[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]

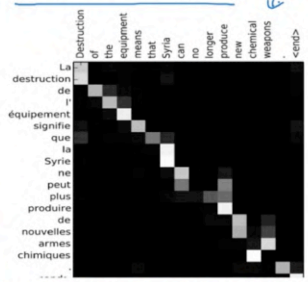
Andrew Ng

Attention examples

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

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

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

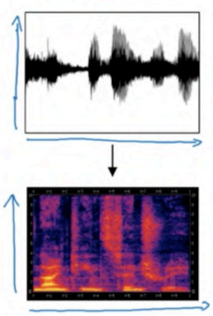


[Bahdanau et. al., 2014. Neural machine translation by jointly learning to align and translate]

Andrew Ng

Speech recognition problem

x audio clip $\longrightarrow$ y transcript



"the quick brown fox"

→ phonemes: de kwik brau

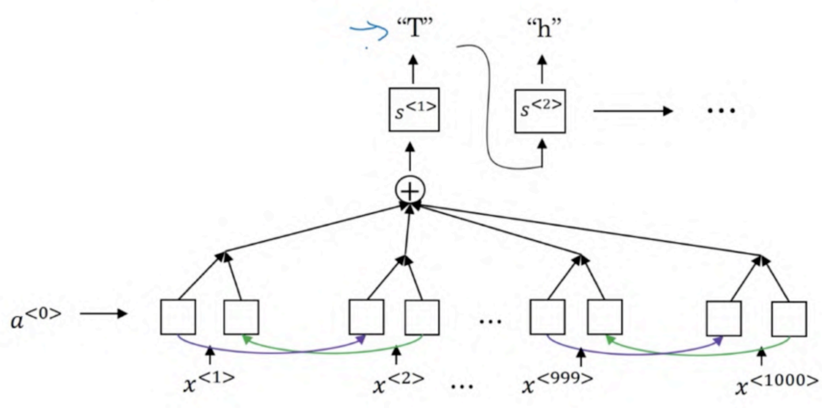
300h

3000h

100,000h

Andrew Ng

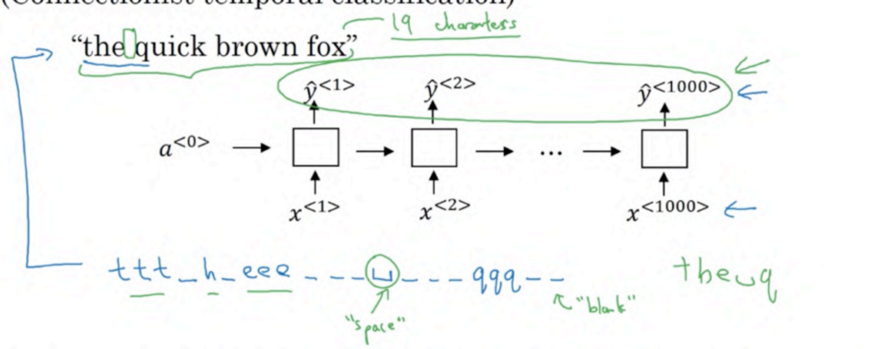
Attention model for speech recognition



Andrew Ng

CTC cost for speech recognition

(Connectionist temporal classification)



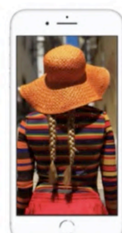
Basic rule: collapse repeated characters not separated by "blank"

[Graves et al., 2006. Connectionist Temporal Classification: Labeling unsegmented sequence data with recurrent neural networks] Andrew Ng

What is trigger word detection?

Amazon Echo (Alexa)

Baidu DuerOS (xiaodunihao)

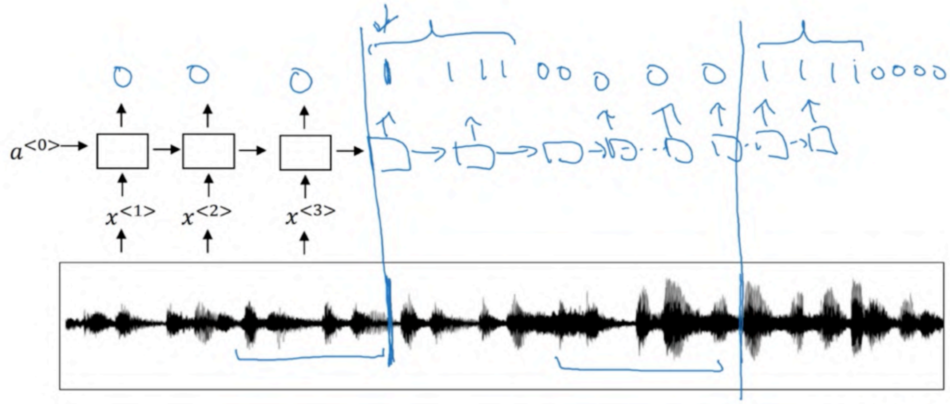


Apple Siri (Hey Siri)

Google Home (Okay Google)

Andrew Ng

Trigger word detection algorithm



Andrew Ng