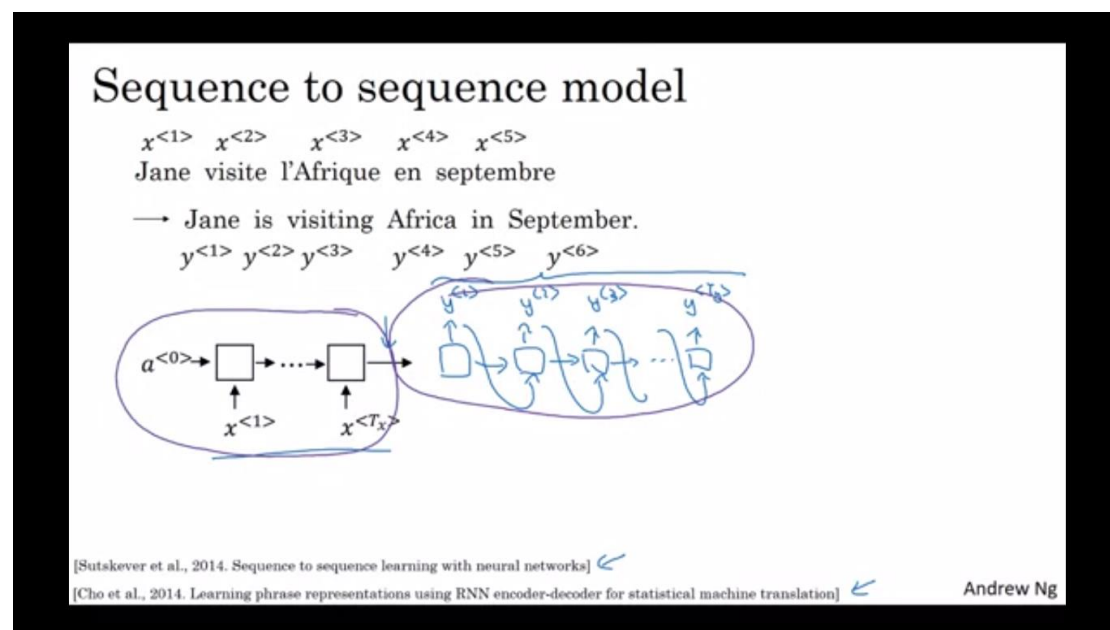


Ahmad Hussameldin Hamed Hassan

Shared Git-hub link: <https://github.com/ahmadhassan1993/sharing-github>

Sequence-to-Sequence

Sequence-to-Sequence is generating sequence from sequence. Like Machine translation, where we use Encoder then Decoder (Language Model). This is called Conditional Language Model (conditional probability).



Machine translation as building a conditional language model

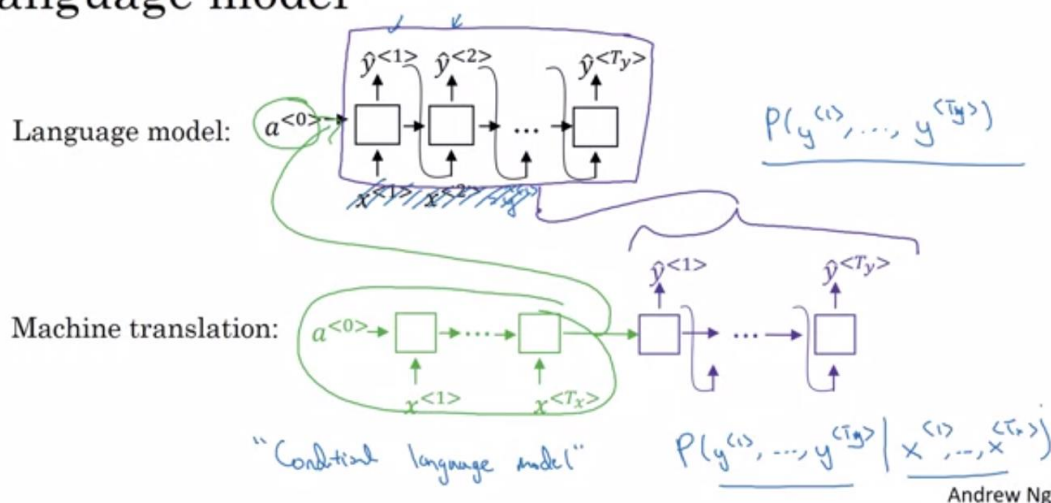
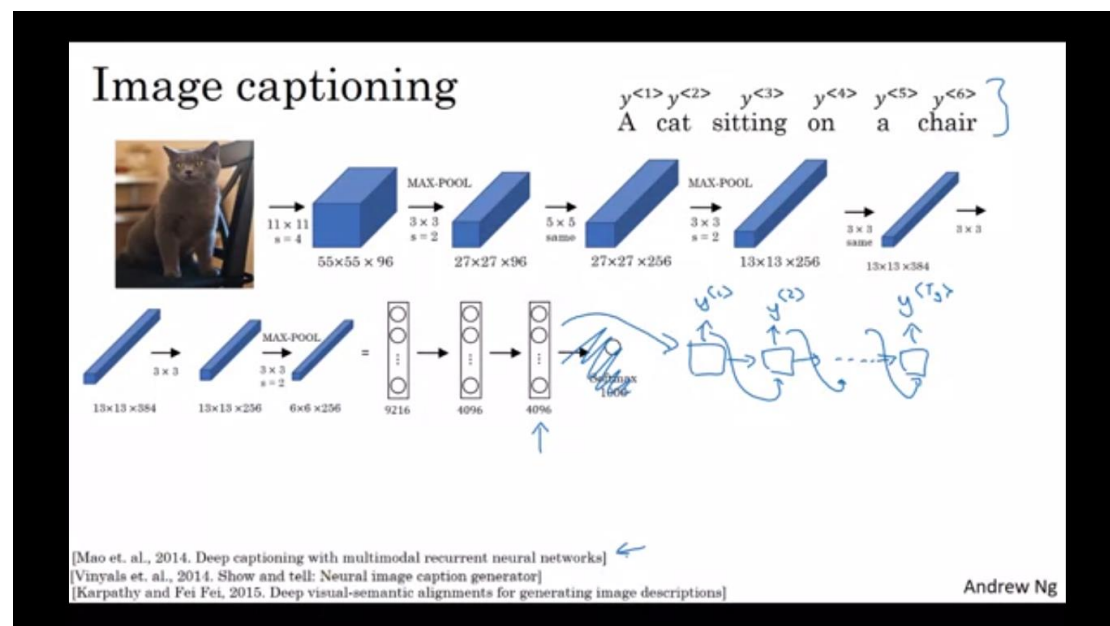


Image-to-Sequence is like making caption to an image, so we use image to generate sequence.



Beam Search is taking maximum of that conditional probability.

Finding the most likely translation

Jane visite l'Afrique en septembre.

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

English

French

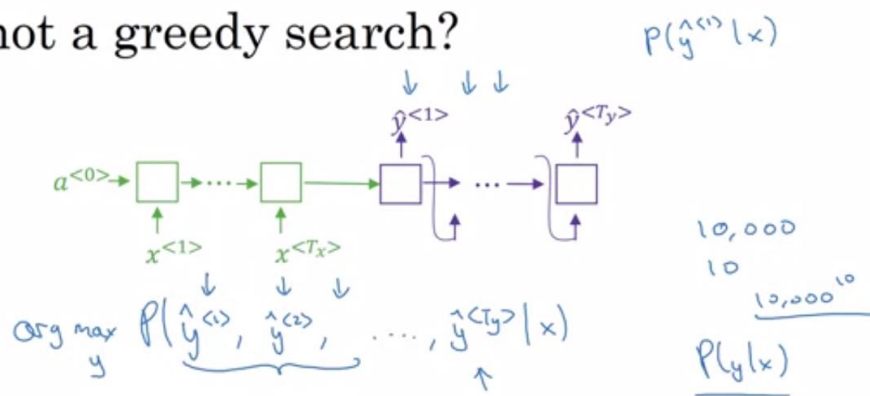
- Jane is visiting Africa in September.
- Jane is going to be visiting Africa in September.
- In September, Jane will visit Africa.
- Her African friend welcomed Jane in September.

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

Andrew Ng

Rather than Greedy Search algorithm, where we just generating sequence without investigating whether it is the most suitable translation or not.

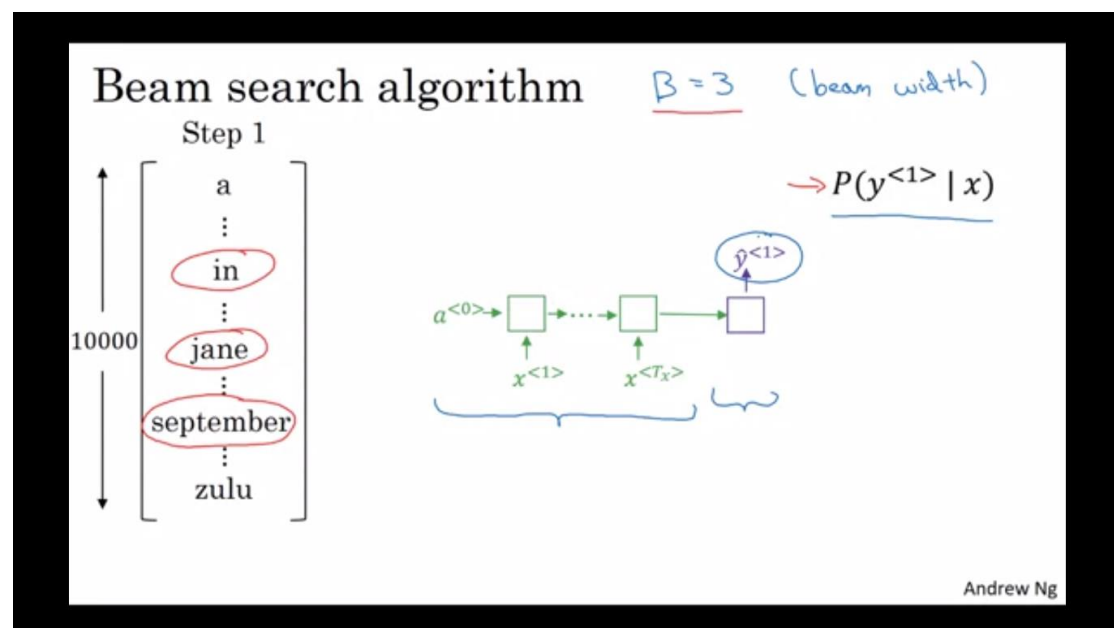
Why not a greedy search?



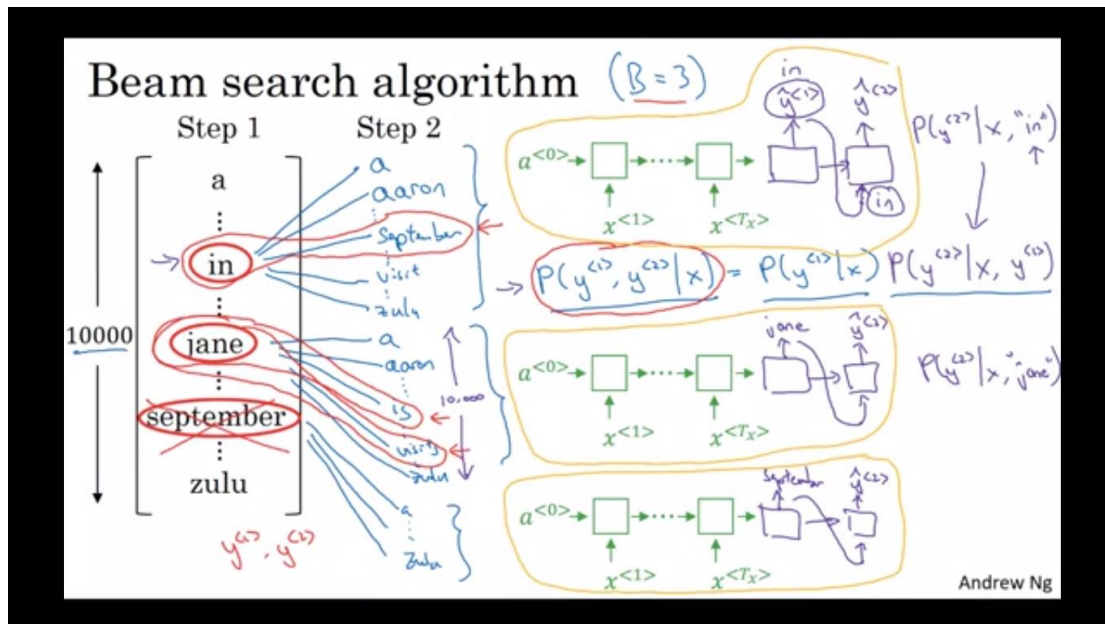
- Jane is visiting Africa in September.
 - Jane is going to be visiting Africa in September.
- $P(\text{Jane is going} | x) > P(\text{Jane is visiting} | x)$

Andrew Ng

In Beam Search, we search the vocabulary vector for the number of words defined by parameter B (Beam Width) to find (from them) the most suitable words.

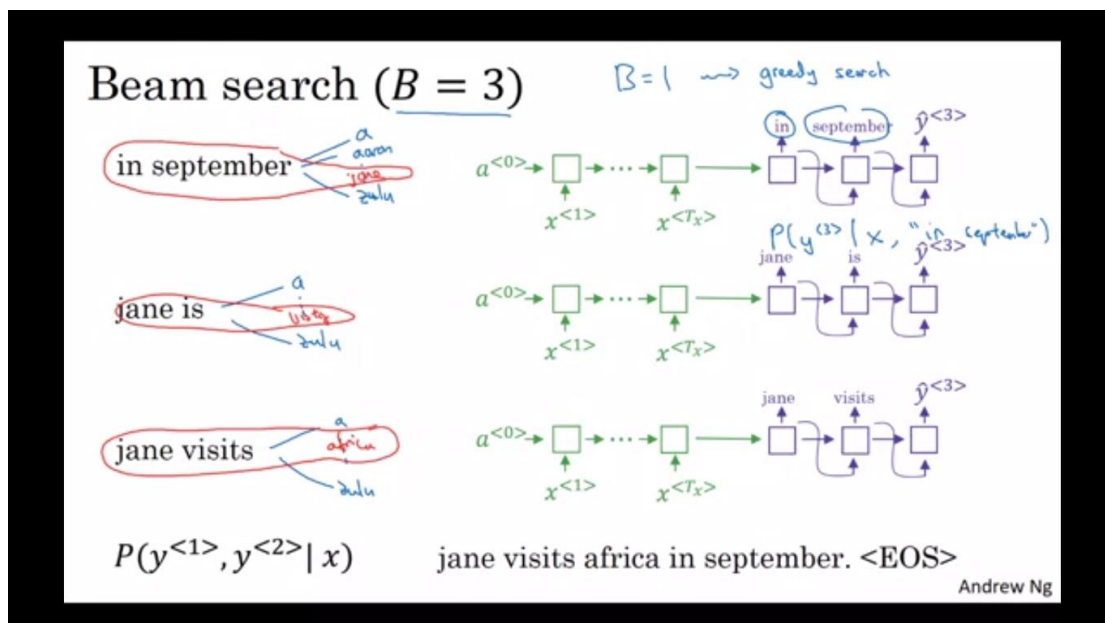


Andrew Ng



We need to use only B copies of the network to find the suitable words from $(B \times \text{Vocabulary vector length})$ words.

Greedy search is Beam Search but with $B=1$.



To enhance the Beam Search, we use length normalization and log function. The more words, the more negative results in log output (Normalized log probability objective). We can make full length normalization or partially normalization depending on the application.

Length normalization

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

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

log

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

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

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

$\alpha = 0.7$ $\frac{\alpha=1}{\alpha=0}$

Andrew Ng

Beam search discussion

Beam width B?

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

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

Detecting the error, whether it is in Beam Search algorithm or RNN Model, is based on comparing the conditional probabilities between human and machine translations. If human translation probability is higher than that of machine translation, the Beam Search is at fault.

Example

Jane visite l'Afrique en septembre.

→ RNN

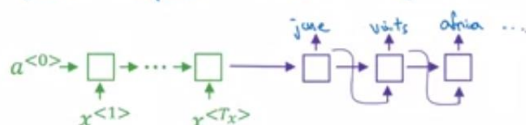
→ Beam Search

BT

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

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

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



Andrew Ng

Error analysis on beam search

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

$P(y^*|x)$

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

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

Case 1: $P(y^*|x) > P(\hat{y}|x)$ ←

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

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

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

Conclusion: RNN model is at fault.

Andrew Ng

Error analysis process

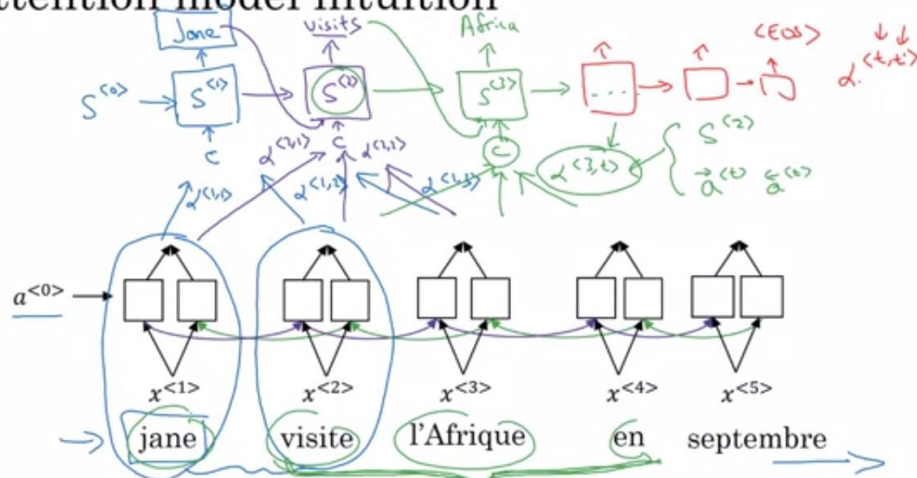
Human	Algorithm	$P(y^* x)$	$P(\hat{y} x)$	At fault?
Jane visits Africa in September.	Jane visited Africa last September.	2×10^{-10}	1×10^{-10}	(B) (R) Q R R ...
...	...	—	—	
...	...	—	—	

Figures out what fraction of errors are “due to” beam search vs. RNN model

Andrew Ng

Attention Models: How much attention should we give for input sequence words near specific word to generate a relative specific word in output sequence.

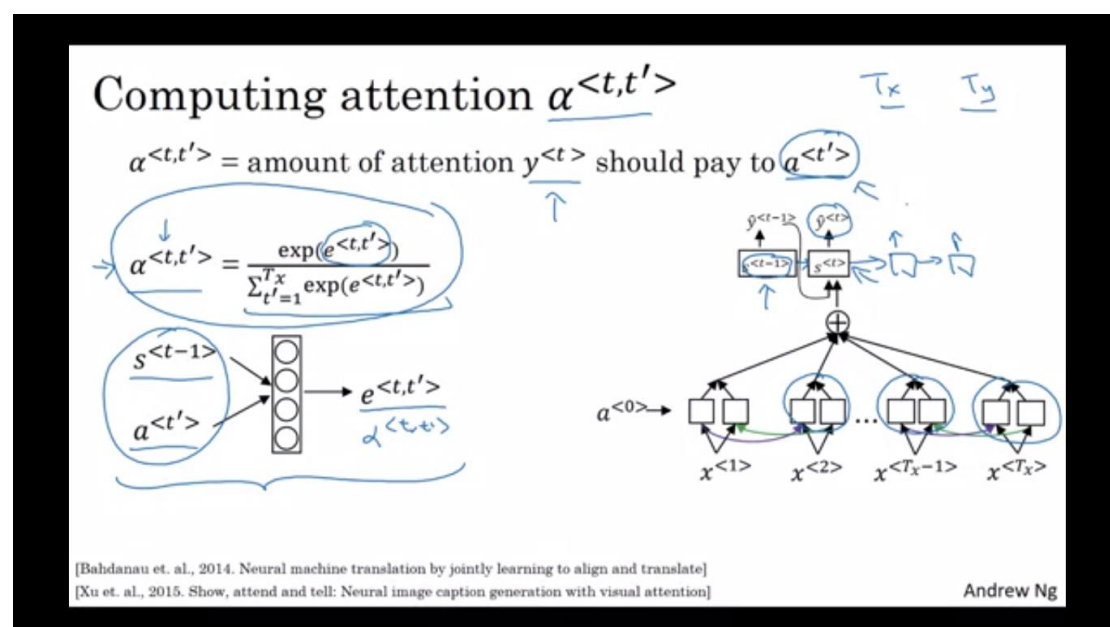
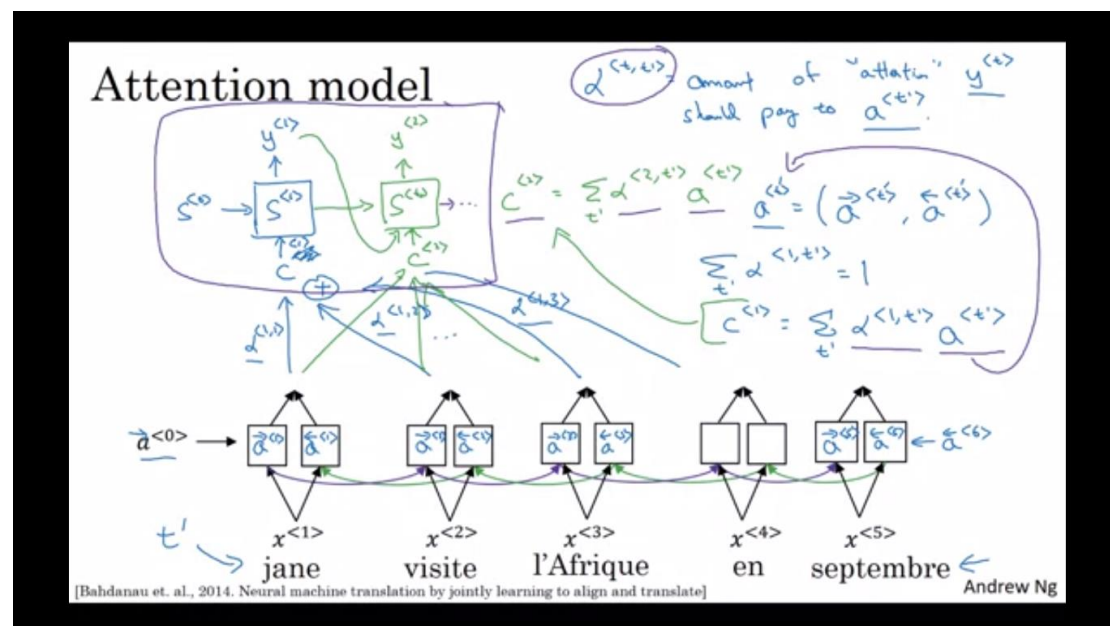
Attention model intuition



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

Andrew Ng

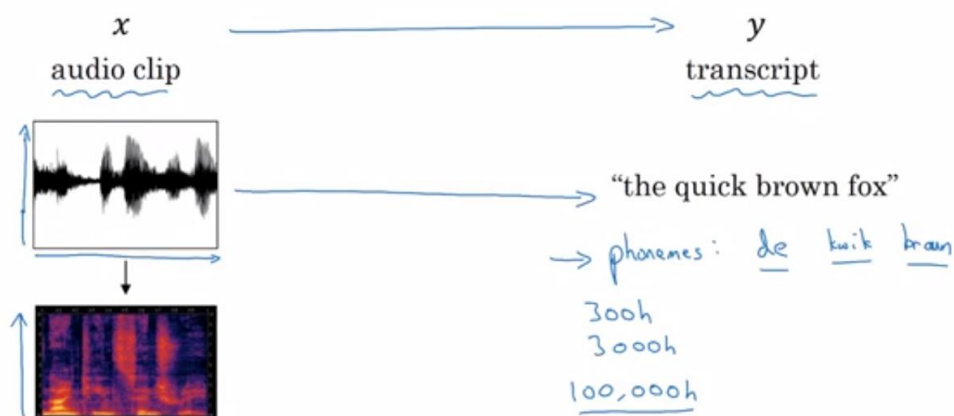
These attention weights are proportional to activation functions and state value of previous stage. Multiplying attention weights with activation functions generates context (c).



We use very small standard NN to compute the attention function by using gradient descent.

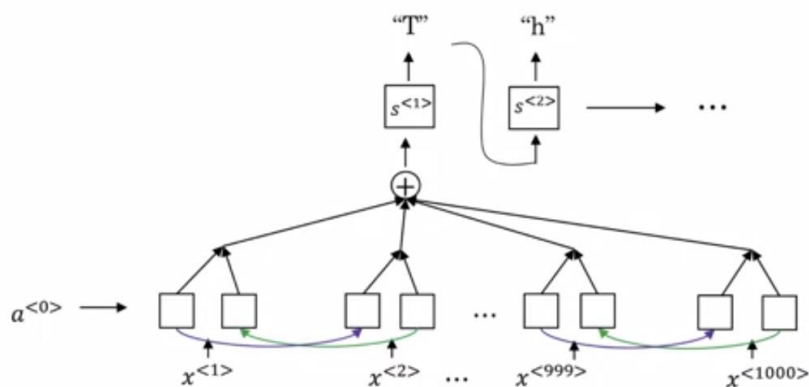
Speech Recognition:

Speech recognition problem



Andrew Ng

Attention model for speech recognition

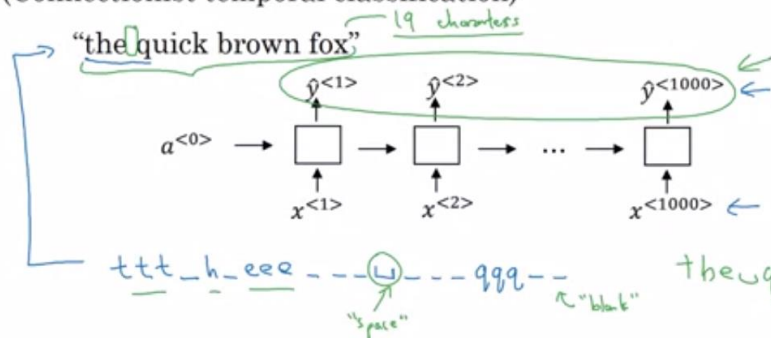


Andrew Ng

We can use attention model or CTC (Connectionist Temporal Classification) model. In CTC, we collapse repeated characters not separated by blank after ensuring that output sequence is same length as input speech sequence.

CTC cost for speech recognition

(Connectionist temporal classification)

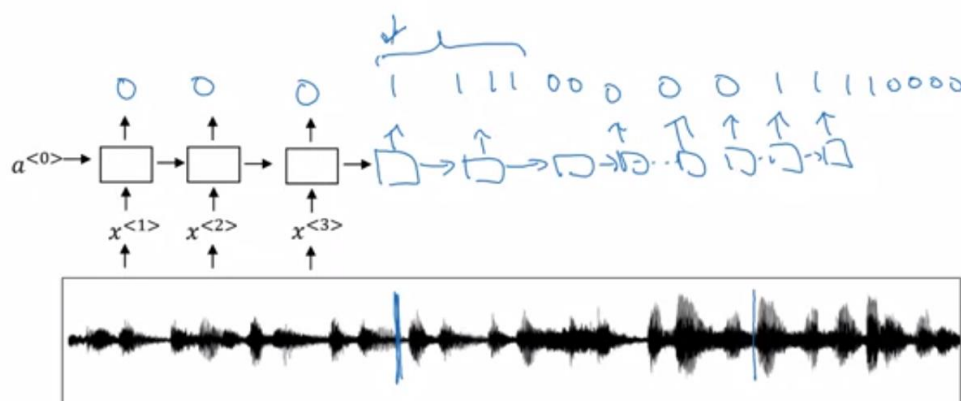


Basic rule: collapse repeated characters not separated by "blank"

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

Trigger word application: It is speech recognition application where we output 1 in output sequence if trigger word was detected.

Trigger word detection algorithm



Andrew Ng