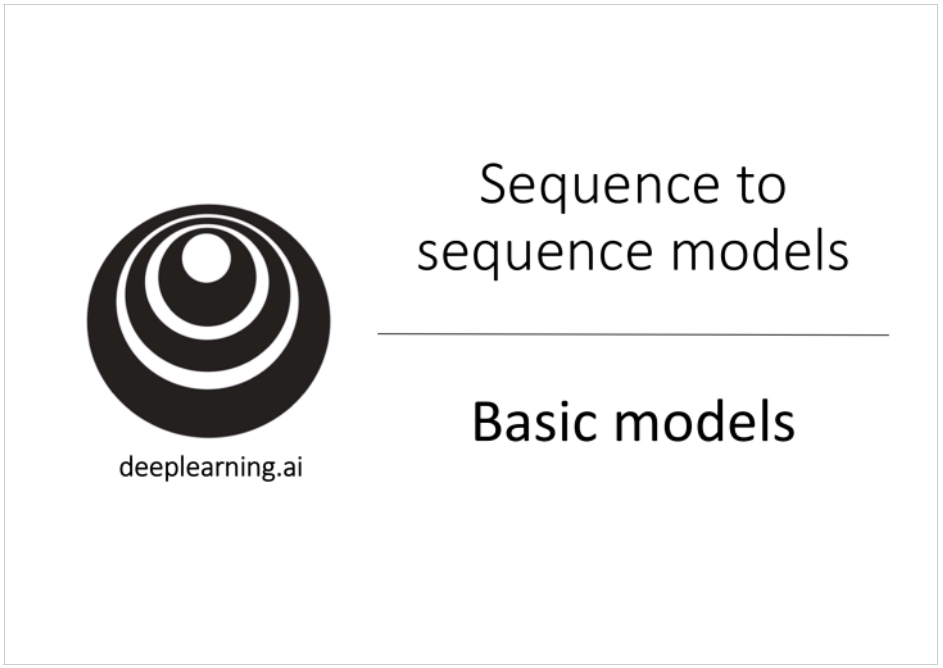
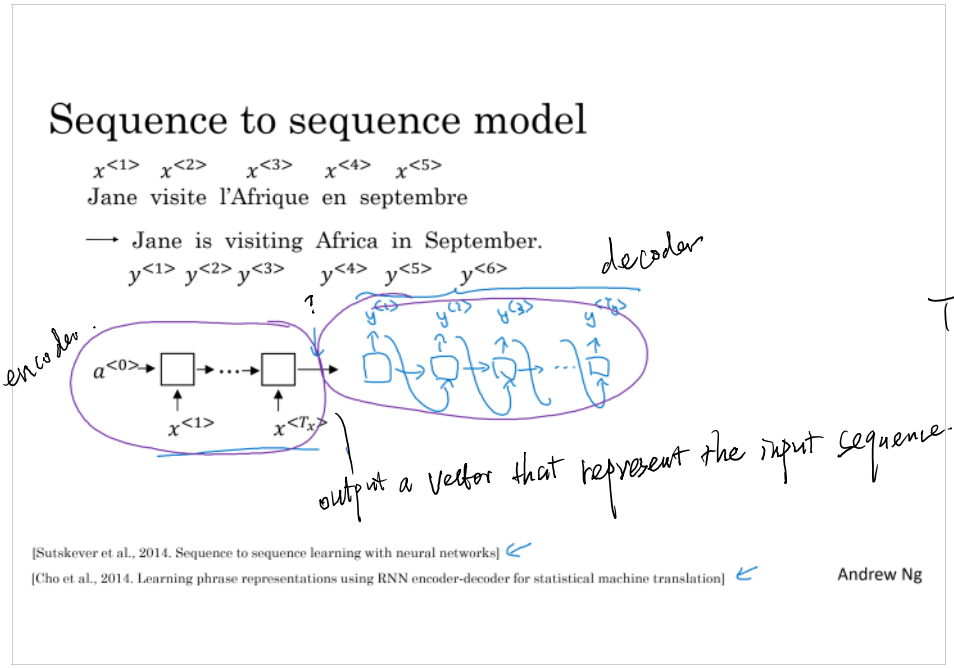
Various Sequence to sequence architectures

Saturday, June 9, 2018 10:19 AM

translate





encoder

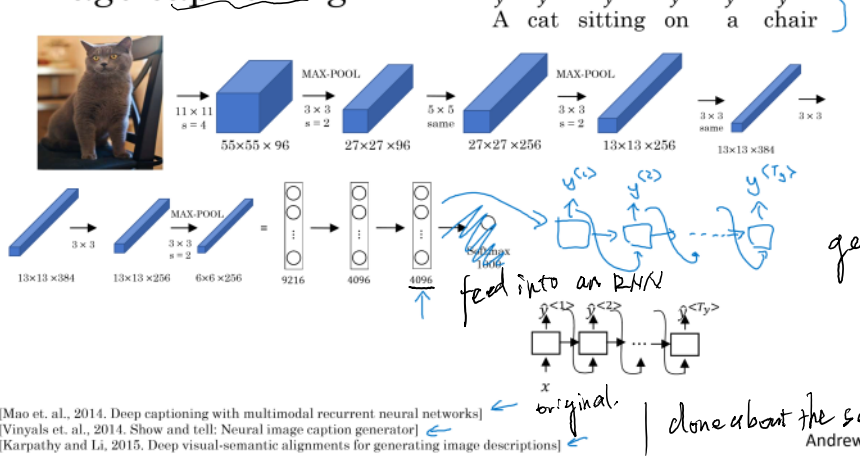
decoder

Translate French into English

output a vector that represent the input sequence.

Another task this method can do.

Image captioning



Alexnet architecture.
get rid of softmax

generate an output sequence.

feed into an RNN

[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]

original. | done about the same time

Andrew Ng

There are some difference. Don't want a randomly chosen. 为了得到Sample sentence 的话. (是这样).

Want Most likely

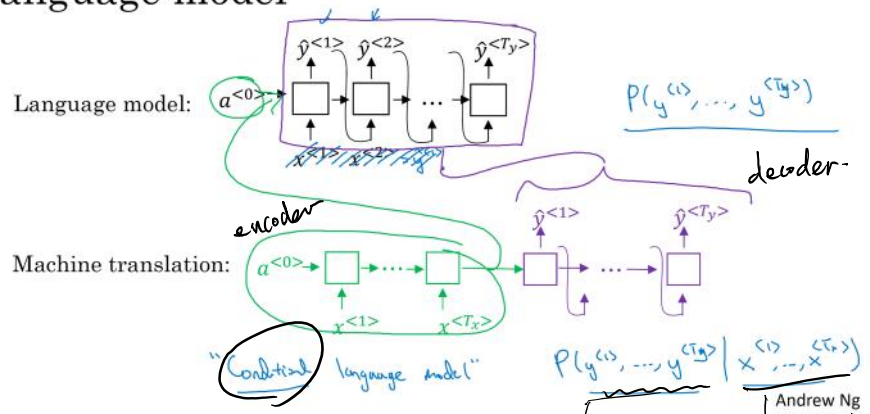
picking

deeplearning.ai

Sequence to sequence models

Picking the most likely sentence

Machine translation as building a conditional language model



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

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

the decoder network look identical to the language

Conditional — Difference start from the output of the encoder network.

output English sentence (translation). conditional on French sentence

Finding the most likely translation

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

Do not want random sampling.

- 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.

$\arg\max_{y^{<1>}, \dots, y^{<T_y>}} P(y^{<1>}, \dots, y^{<T_y>} \mid x)$ ⟹ find an English sentence that max the probability.

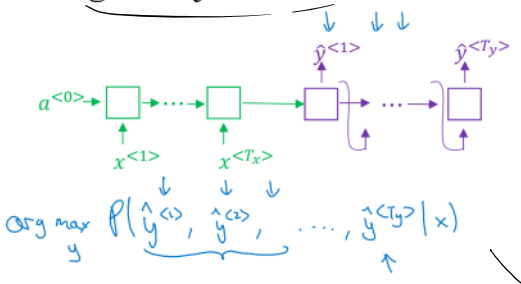
Andrew Ng

greedy search may end up with suboptimal 可以看 top

Why not a greedy search?

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

greedy search — search for the most likely word.



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

10,000 / 10 / $10{,}000^{10}$

$P(y \mid x)$

goal. pick the words to maximize the JOINT probability.

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

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

Jane is going to be visiting Africa in September.

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

Andrew Ng

locally optimal. not globally max.

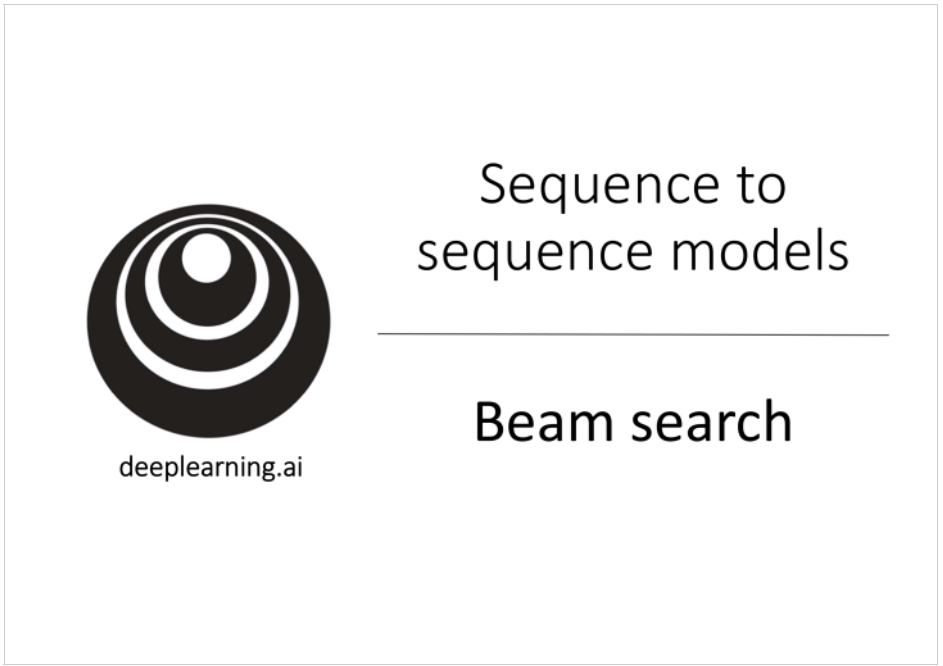
The total # of combinations is exponentially large ⇒ can't calculate all

1000 words in dict. 10 word long sentence

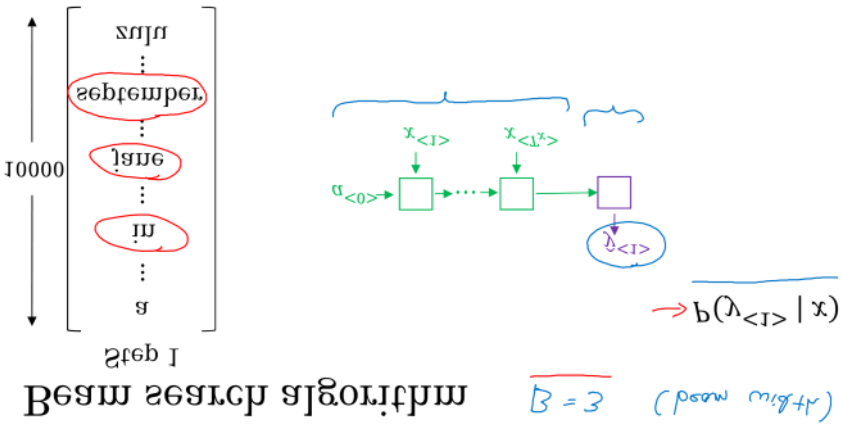
1000^{10} sentences

The common thing to do use approximate algorithm

beam search



Beam width: consider the 3 most likely possibilities



Keep a memory of the top 3.

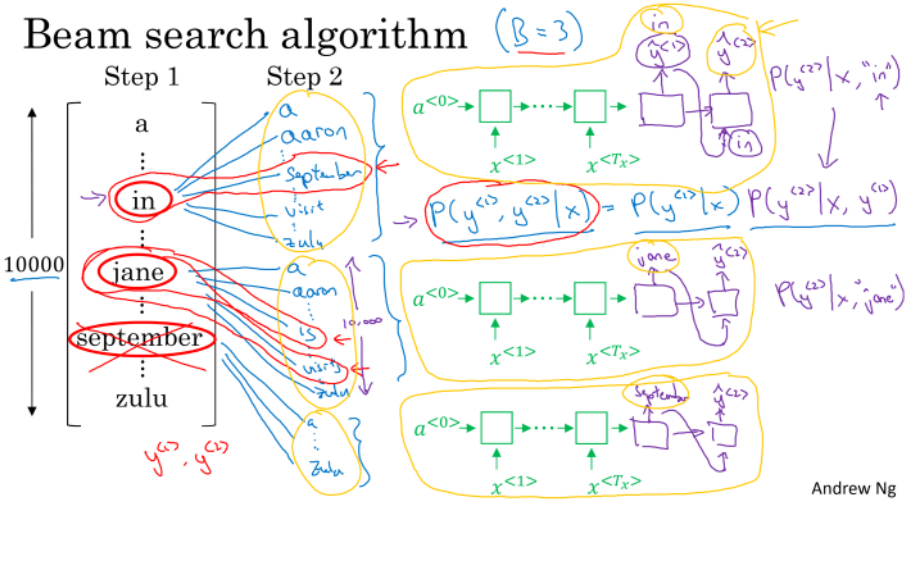
B = 3

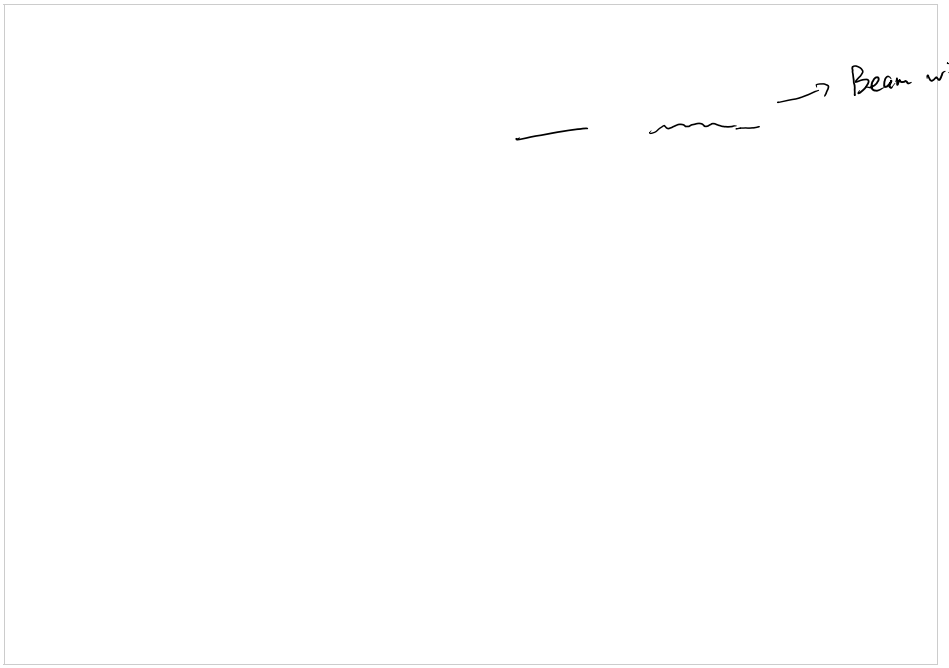
B × 10000 possibilities

⇒ pick top B.

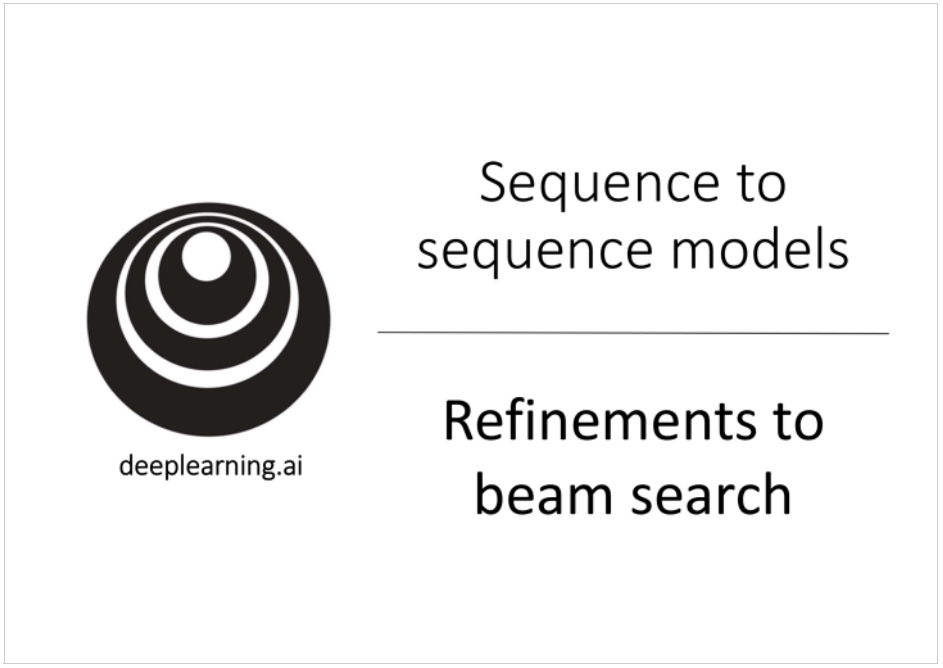
每步都 pick top 3. path 不会增多.

(一步特别坏怎么办?)




→ Beam width = 1. greedy search

refine beam


deeplearning.ai
Sequence to sequence models
Refinements to beam search

Length normalization

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

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

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

Andrew Ng

normalize by length

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

middle: 1: full normalization; 0: no normalization

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

If not normalize by length $\Rightarrow$ reward short translation $\Rightarrow$ not good. normalize.

larger B. larger computational power required.

$\rightarrow$ for research, sometimes used.

very large for production system.

Beam search: faster, not guaranteed to find exact maximum.

error analysis beam search

deeplearning.ai

Sequence to sequence models

Error analysis on beam search

Example

Jane visite l'Afrique en septembre.

→ RNN
→ Beam Search

B↑

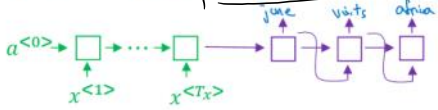
Decide which part is to blame
— Always tempting to ↑B. But it may not help.

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

Algorithm: Jane visited Africa last September. ($\hat{y}$) ⇐ Bad translation, changed meaning.

Use error analysis to decide which part to focus on fixing

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



understand this and know which part is problematic

Andrew Ng

Error analysis on beam search

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

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

Case 1: $P(y^*|x) > P(\hat{y}|x)$

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.			

Figures out what faction of errors are "due to" beam search vs. RNN model

Andrew Ng

blue score

Sequence to sequence models

Bleu score

deeplearning.ai

Bleu score (optional)

★ Evaluate how much your predicted output is ~~the same~~ similar to one or more reference.

Multiple equally good reference

The Bleu score gives us human evaluation.

Ask human to evaluate.

Evaluating machine translation

French: Le chat est sur le tapis.

Bleu: bilingual evaluation understudy.

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.

one way: look at each word, see if appears in the references.

=> precision.

Precision: 7/7

Modified precision: a unique word get credit at most twice. 2/7

A reader-friendly paper. Ng recommended

[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

4/6

共6个bigrams

有4个.

[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. ($\hat{y}$)

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

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

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

Andrew Ng

$p_1, p_2 = 1.0$

→ Sometimes possible even if the $\hat{y}$ not exact the same with any of the reference

total # of ngram

Bleu details

p_n = Bleu score on n-grams only

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

p_1, p_2, p_3, p_4

Brevity penalty.

BP = brevity penalty

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

short words → high score because higher chance all terms are in the reference

=> So use BP to adjust for that (penalize shorter translations)

find details here in the paper

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

Andrew Ng

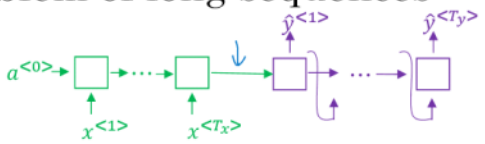
attention model

deeplearning.ai

Sequence to sequence models

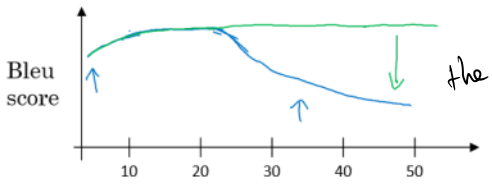
Attention model intuition

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.

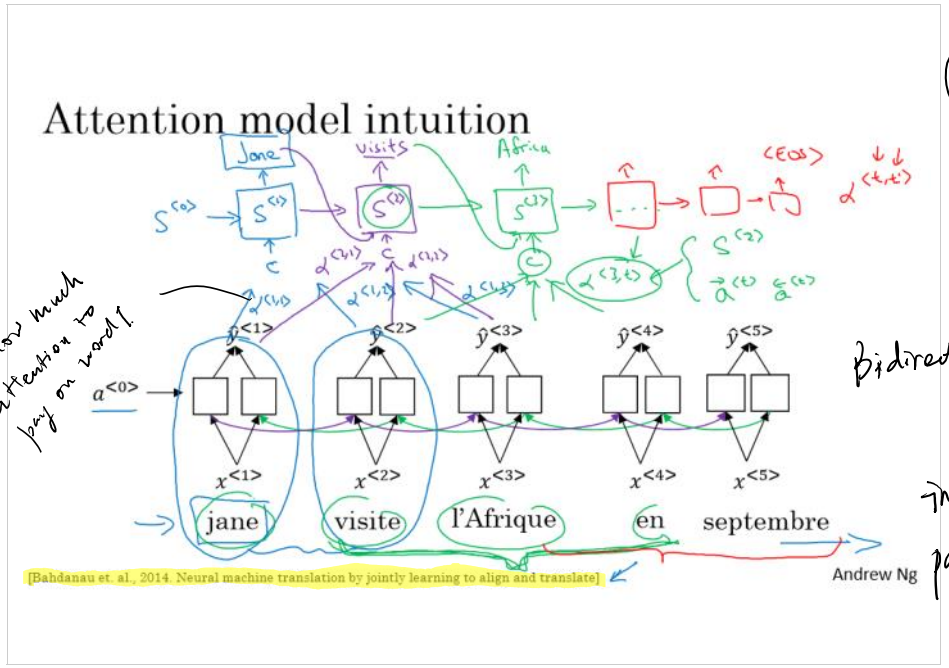


Andrew Ng

human translator will read the sentence part by part.

把句子拆成部分 translate

the ability of the neural network to memorize is limited.



? context. Mainly on attention. next few slides.

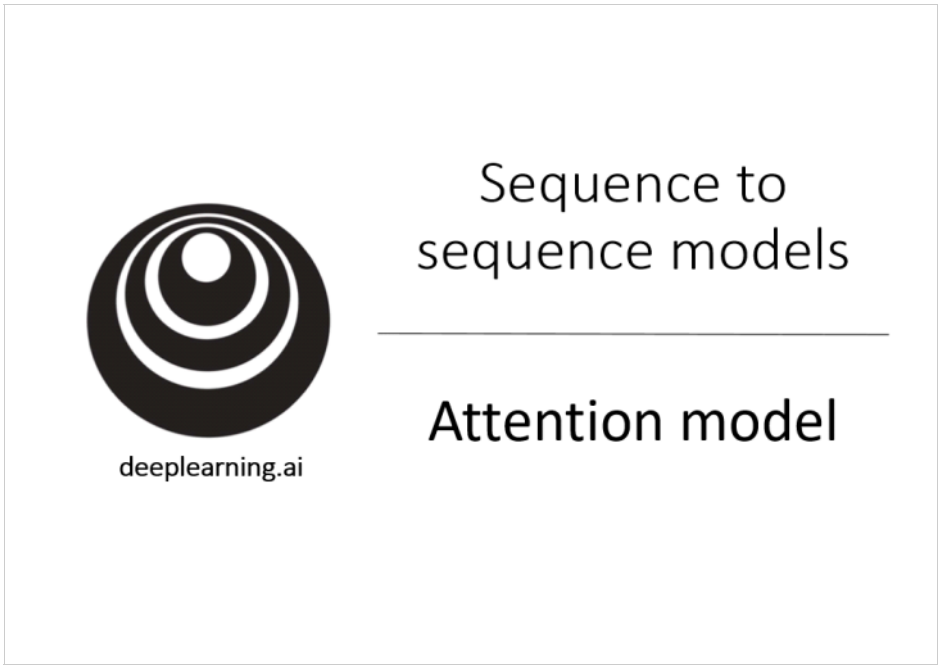
$\alpha^{<t,t'>}$ attention weight.

for each step of generation only focus on a few words

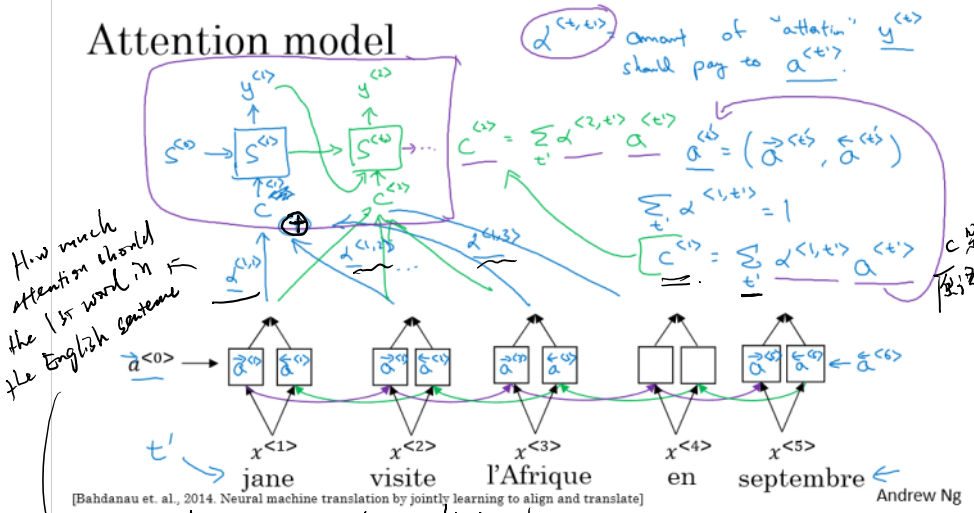
Bidirectional RNN

Instead of attention weights we pay attention to $a^{<1,1>}$.

attention model 2



Attention model

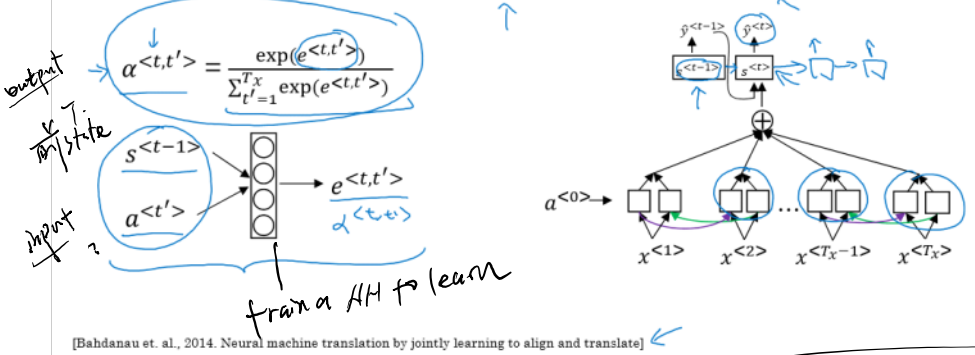


[Bahdanau et. al., 2014. Neural machine translation by jointly learning to align and translate]

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

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

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



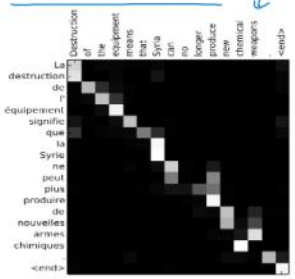
[Bahdanau et. al., 2014. Neural machine translation by jointly learning to align and translate]
[Xu et. al., 2015. Show, attend and tell: Neural image caption generation with visual attention]

Attention examples

July 20th 1969 ⟶ 1969 – 07 – 20

23 April, 1564 ⟶ 1564 – 04 – 23

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



Andrew Ng Deep Learning. Page 14

high — white in the figure

Andrew Ng