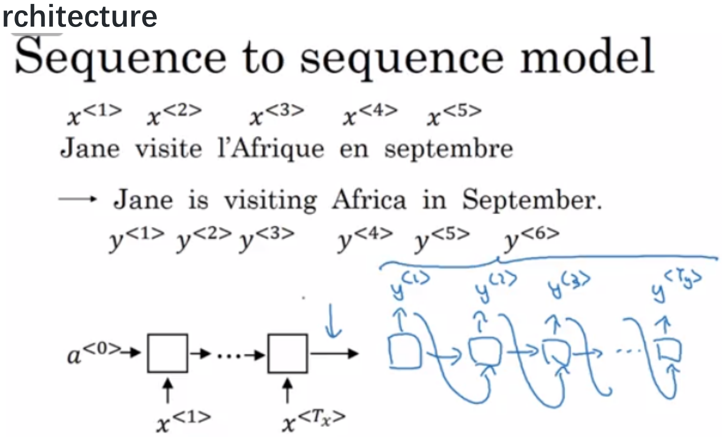
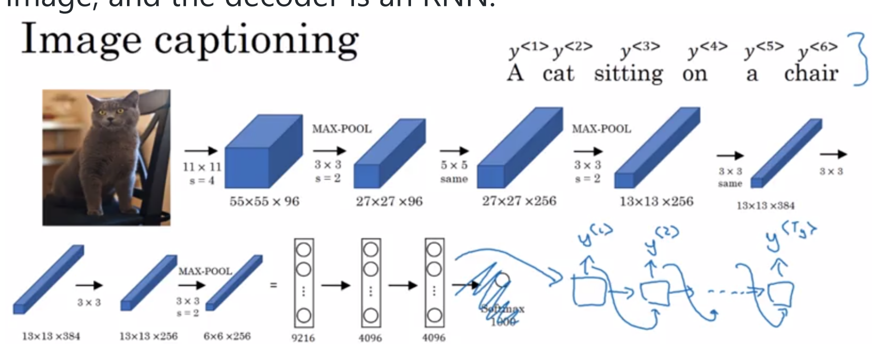
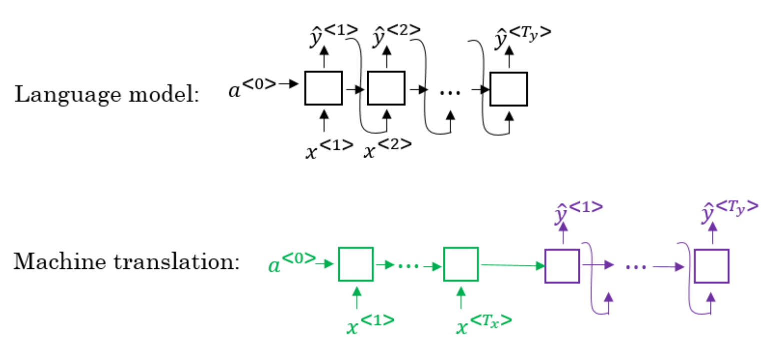
Created:	19/03/26 13:03	**Updated:**	19/03/28 13:23

Attention mechanism

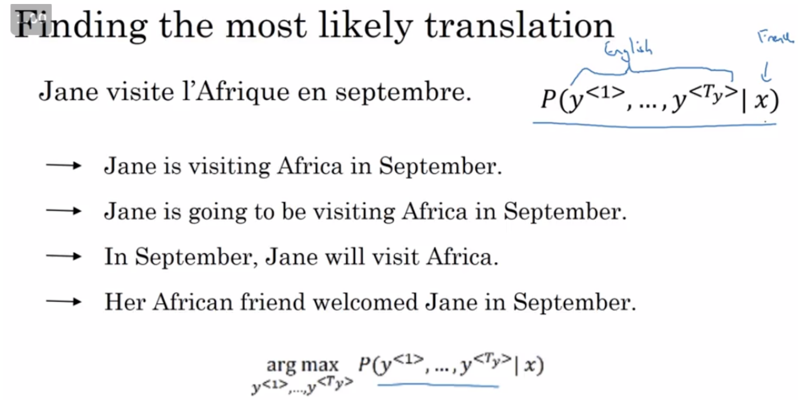
Various sequence to sequence architectures

- basic model
  - [encoder]→[decoder]
  - for machine translation task:
    - NN architecture
    - 

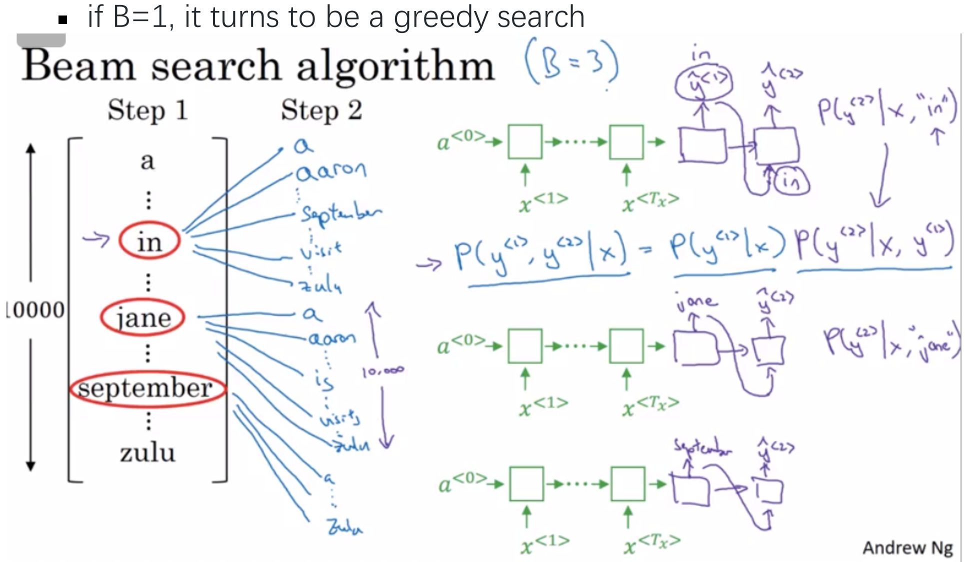
  - for image captioning
    - uses a pretrained CNN (like AlexNet) as an encoder for the image, and the decoder is an RNN.
    - 

- Picking the most likely model
  - comparison between language model and machine translation
    - NN
    - 

    - Problems formulations are different:
      - In language model: $P(y^{<1>}, \ldots, y^{<Ty>})$
      - In machine translation: $P(y^{<1>}, \ldots, y^{<Ty>} \mid x^{<1>}, \ldots, x^{<Tx>})$

- problem:
    - **we want the best translation instead of random answer**
        - 

- aim:
    - **best translation**
- **why not a greedy search?**
    - doesn't really work
    - the words in sentence have relation in long distance instead of just adjacent one.
    - distribution space is too large for greedy search.

- Beam Search
    - a heuristic search algorithm
        - by iterating and selecting the best results in each "B" outcomes.
        - optimize:
            - $$\arg\max_{y^{<1>},\dots,y^{<T_y>}} P(y^{<1>}, \dots, y^{<T_y>} \mid x)$$
        - caution. It's a bit computational expensive as **B** growing bigger.
    - parameter `B` : the beam width.
        - means that the top "B" number of possible outcomes.
        - if B=1, it turns to be a greedy search
    - 

- refinements to beam search
    - optimization:

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

may be too small

log

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

MLE

- using Maximum Likelihood Estimate
- another problem: The optimization function prefers small sequences rather than long ones.
    - Because multiplying more fractions gives a smaller value, so fewer fractions - bigger result.
- So: dividing by the number of elements in the sequence.

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

- how to choose B?

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

- what's more, error analysis could help.

- error analysis
    - help to find out which limit model performance, RNN or Beam Width
    - EX.

Example

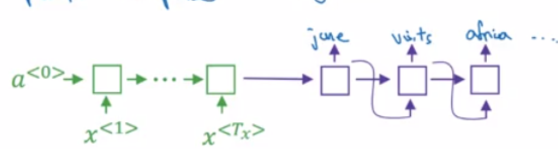
→ RNN
→ Beam Search

Jane visite l'Afrique en septembre.

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

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

RNN computes $P(y|x)$



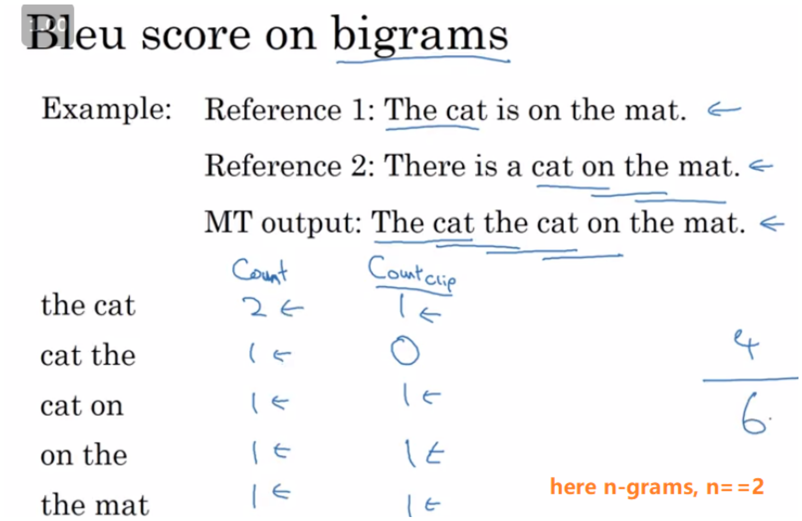
- calculate $P(y^* \mid X)$ and $P(\hat{y} \mid X)$
    - if $(P(y^* \mid X) > P(\hat{y} \mid X))$:
        - Beam search is at fault.
    - else: # $(P(y^* \mid X) <= P(\hat{y} \mid X))$

    - RNN model is at fault.
  - so, make a table, then get counts and decide what to work on next.
    -

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
				⋮

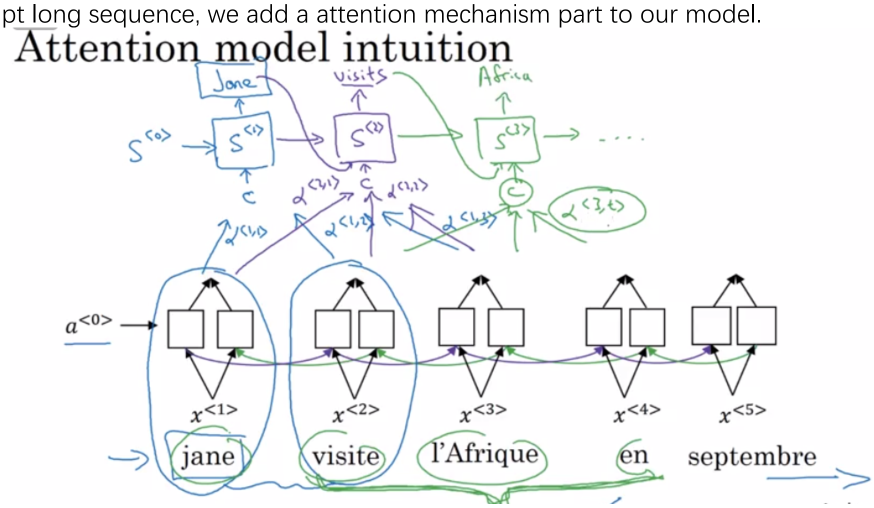
- **BLEU score**.
  - BLEU stands for *bilingual evaluation understudy*.
  - task:
    - given a sentence in a language there are many possible good translation. How to evaluate our results?
  - intuition:
    - choose the result which pretty closes to any of the references provided by humans
  - algorithm:
    - quite like a modified kind of "Template matching"
    - compute pn:
      - $$p_n = \frac{\sum_{ngram \in \hat{y}} count_{clip}(ngram)}{\sum_{ngram \in \hat{y}} count(ngram)}$$
    - combined Bleu Score:
      - BLEU score = BP * exp(1/n * sum(Pn))
      - BP here means:
        - BP = brevity penalty

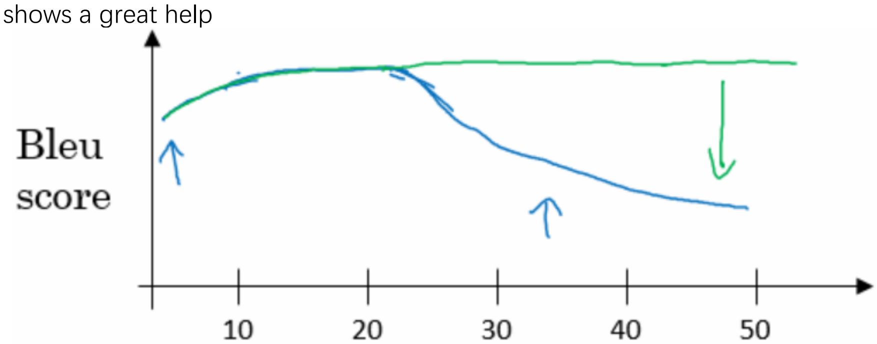
          $$BP = \begin{cases} 1 & \text{if MT\_output\_length} > \text{reference\_output\_length} \\ \exp(1 - \text{MT\_output\_length}/\text{reference\_output\_length}) & \text{otherwise} \end{cases}$$
        - It turns out that if a machine outputs a small number of words it will get a better score so we need to handle that.
    - Ex.

- 

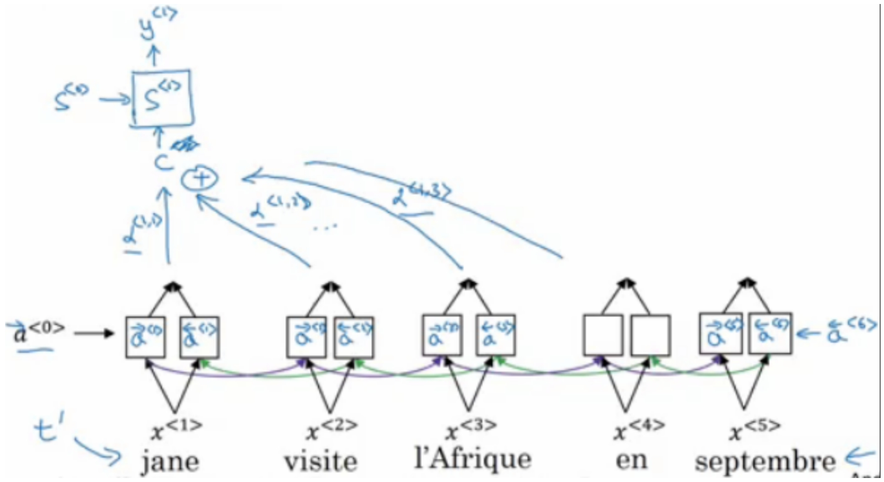
  - application:
    1. machine translation
    2. image captioning

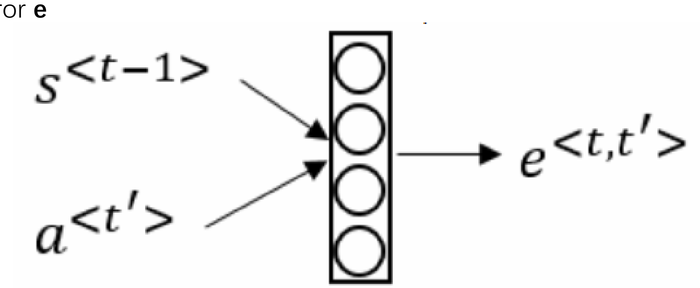
Attention Mechanism

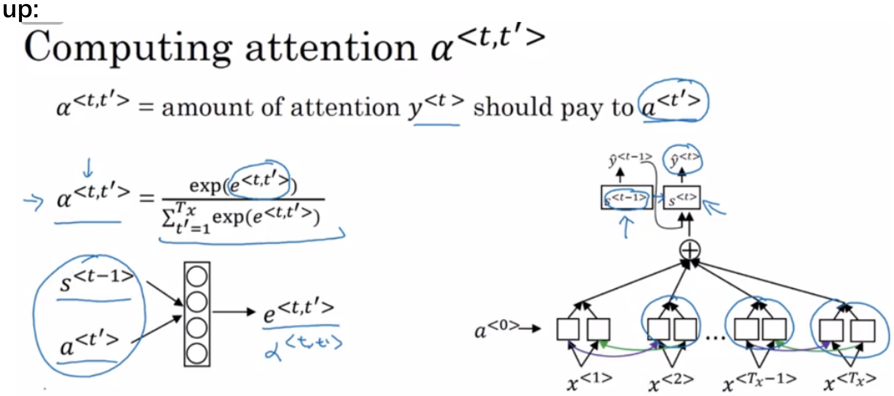
- intuition:
  - for a long sequence, instead of understanding it as a whole,we may pay more attention on certain part of it.
  - in fact, as sequence growing longer, the Bleu scores would go down conspicuously.
  - to adapt long sequence, we add a attention mechanism part to our model.
    - 

  - which shows a great help
    - 

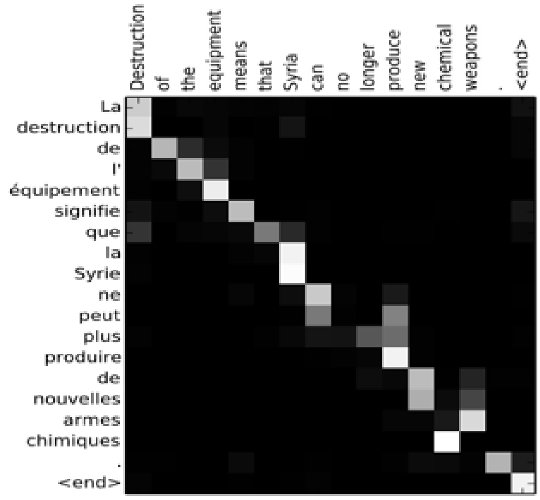
    - Blue is the normal model, while green is the model with attention mechanism.
- details
  - structure:

- - 

  - add on a RNN/LSTM/GRU
- formula
  - for each **a**:
    - $$\alpha^{<t,t'>} = \frac{\exp(e^{<t,t'>})}{\sum_{t'=1}^{T_x} \exp(e^{<t,t'>})}$$
    - and by summing up, it would be
      - $$\sum_{t'} \alpha^{<1,t'>} = 1$$
  - for context **c:**
    - $$c^{<1>} = \sum_{t'} \alpha^{<1,t'>} a^{<t'>}$$
  - for error **e**
    - 

- **wrap up:**
  - 

- visualizing weights


Destruction
of
the
equipment
means
that
Syria
can
no
longer
produce
new
chemical
weapons
.
<end>
La
destruction
de
l'
équipement
signifie
que
la
Syrie
ne
peut
plus
produire
de
nouvelles
armes
chimiques
.
<end>