Sequence to Sequence Architectures

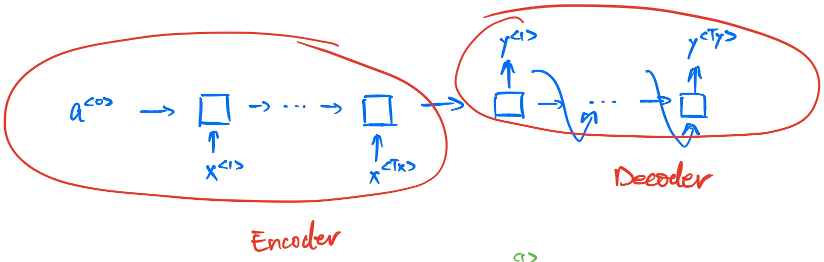
Translation

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

Jane visite l'Afrique en septembre

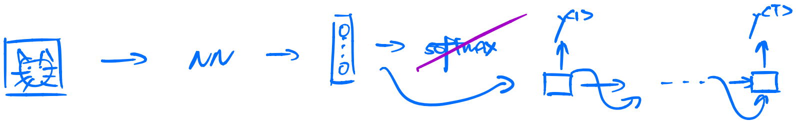
→ Jane is visiting Africa in September

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

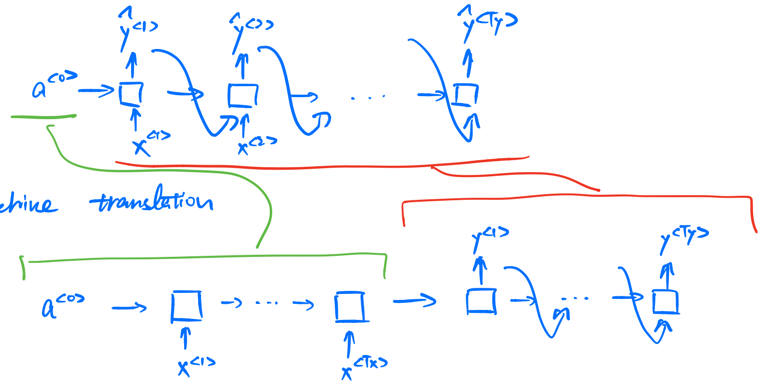


$y^{<1>}$ . . .

Image to Caption e.g. A cat is sitting on a chair



language model



"Conditioned language model"

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

⇒ Pick the most likely sentence

Why not greedy search ?

e.g. Jane is visiting Africa ✓

Jane is going to visit Africa ✗

$P(\text{Jane is going} \mid x)$ higher ⇒ less good sentence

Beam Search

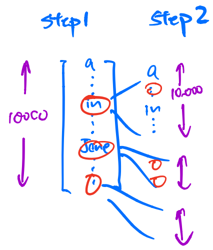
B = 3 (beam width)

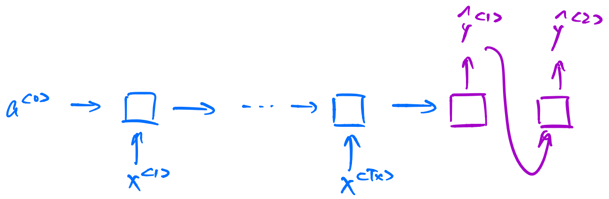
⇒ $P(y^{<1>} \mid x)$

B ↑ slower

B ↓ faster. worse result

3 copies





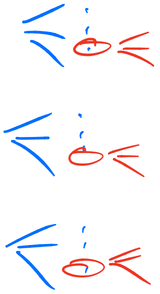
$$P(y^{<1>}, y^{<2>} \mid x) = P(y^{<1>} \mid x)\, P(y^{<2>} \mid x, y^{<1>})$$

step 3

in september

jane is

jane visits



Improvements to Beam Search

- length normalization

$P(y^{<1>} \mid x)\, P(y^{<2>} \mid x, y^{<1>}) \cdots$

- can be small
- prefer short sentences

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

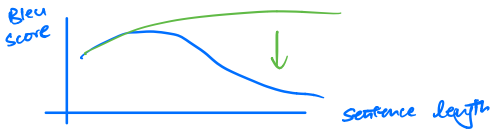
$$\hookrightarrow \arg\max_y \frac{1}{T_y^{\alpha}} \prod_{t=1}^{T_y} \log P(y^{<t>} \mid x, y^{<1>}, \ldots, y^{<t-1>})$$

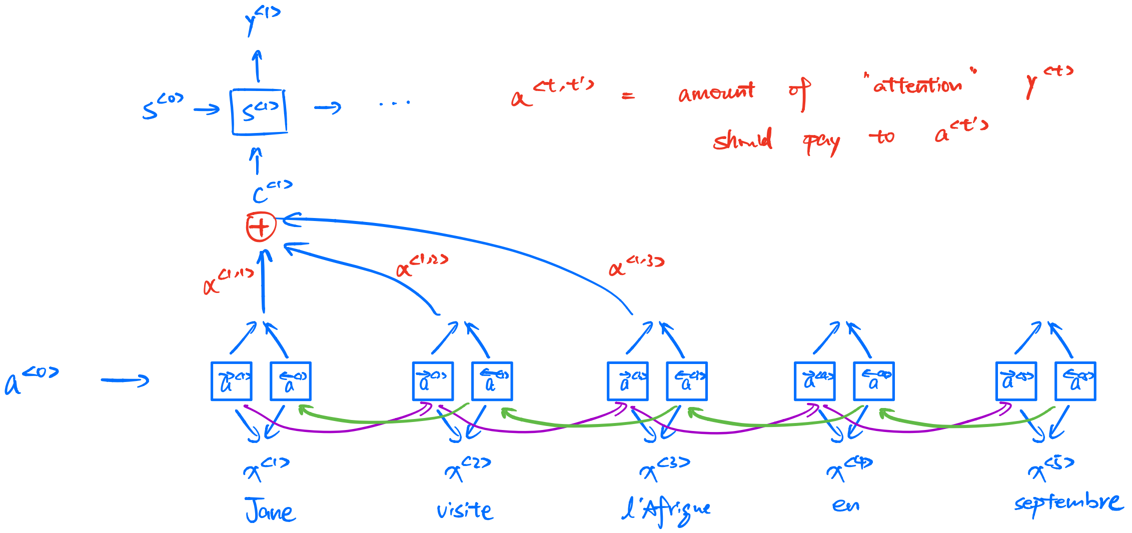
normalize (e.g. $\alpha = 0.7$)

Attention Model

The problem w/ long sequence

⇒ hard to memorize





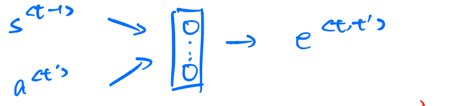
$\alpha^{<t,t'>}$ = amount of "attention" $y^{<t>}$ should pay to $a^{<t'>}$

$\sum_{t'} \alpha^{<1,t'>} = 1$ $\qquad a^{<t'>} = (\overrightarrow{a}^{<t'>}, \overleftarrow{a}^{<t'>})$

$c^{<1>} = \sum_{t'} \alpha^{<1,t'>} a^{<t'>}$

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

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



train a small NN