deeplearning.ai

Sequence to sequence models

Beam search

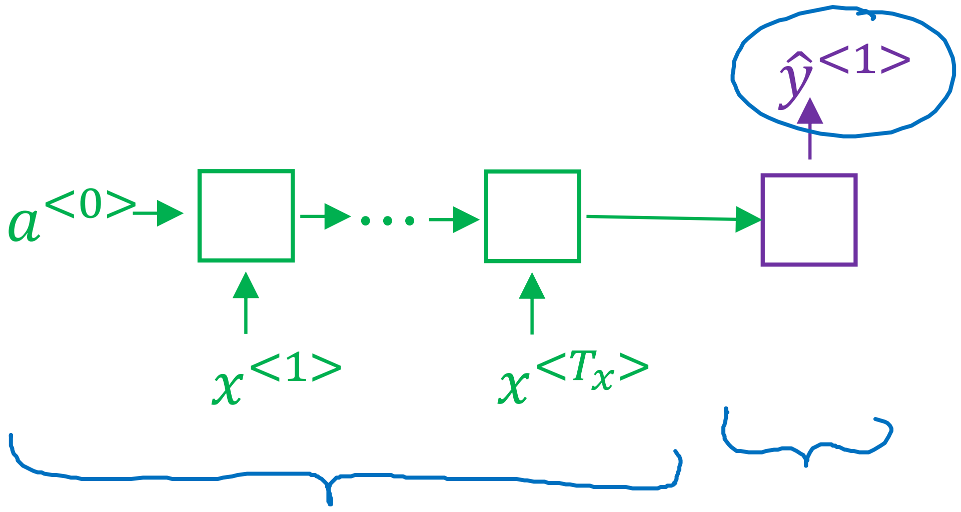
Beam search algorithm

$B = 3$ (beam width)

Step 1

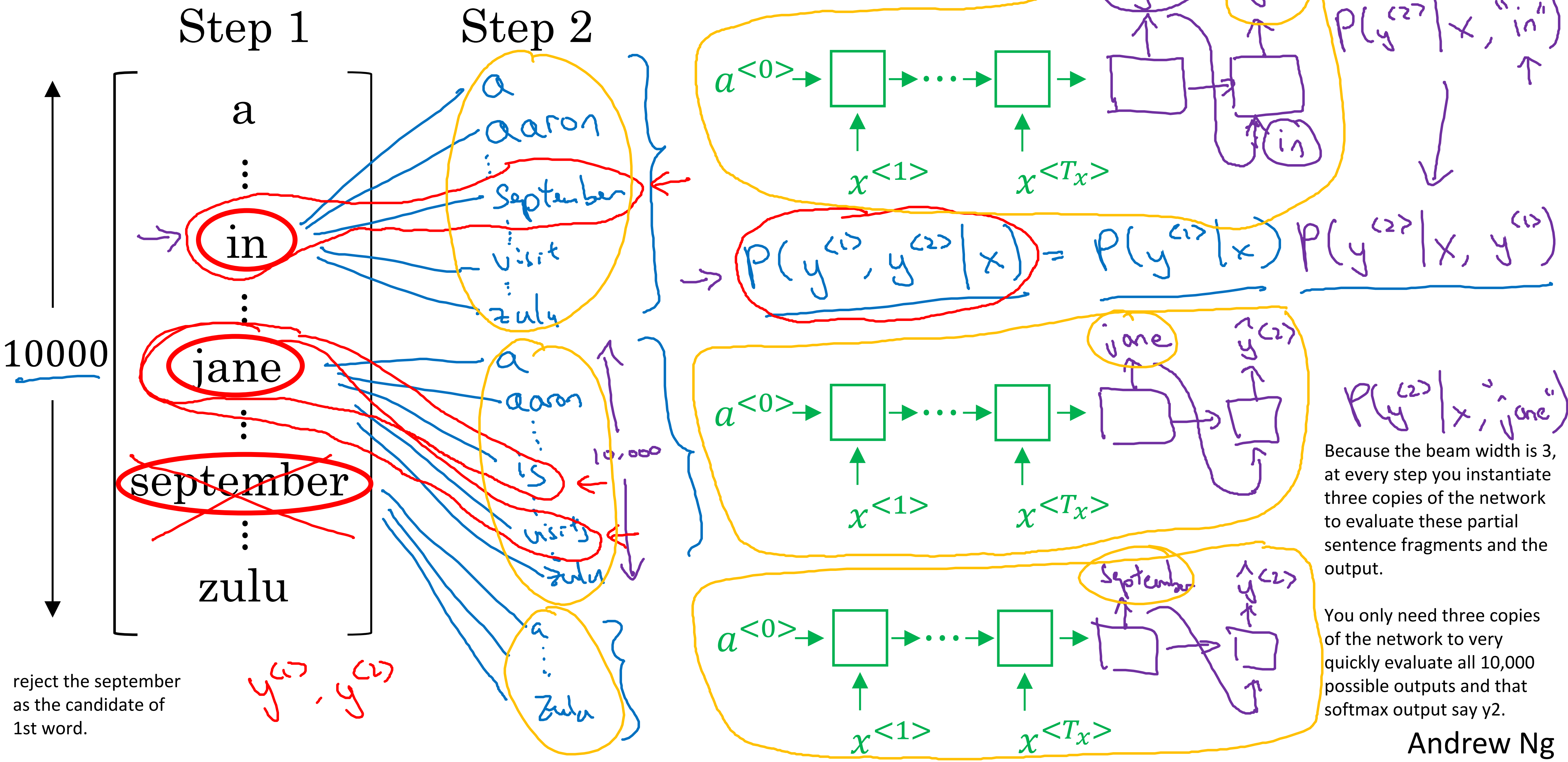
$$10000 \begin{bmatrix} \text{a} \\ \vdots \\ \text{in} \\ \vdots \\ \text{jane} \\ \vdots \\ \text{september} \\ \vdots \\ \text{zulu} \end{bmatrix}$$

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



Beam search algorithm

(B = 3)



Step 1

Step 2

10000

a

in

jane

september

zulu

$a^{<0>}$

$x^{<1>}$

$x^{<T_x>}$

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

reject the september as the candidate of 1st word.

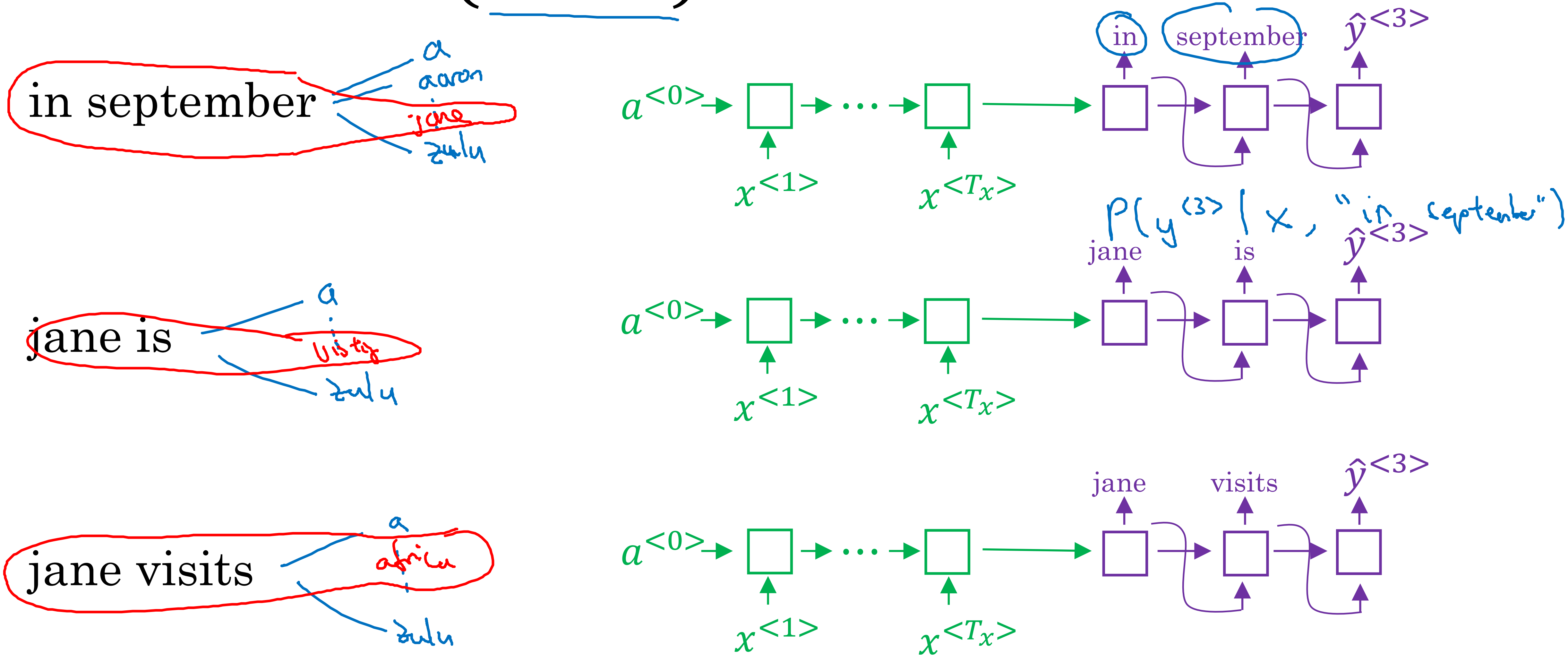
Because the beam width is 3, at every step you instantiate three copies of the network to evaluate these partial sentence fragments and the output.

You only need three copies of the network to very quickly evaluate all 10,000 possible outputs and that softmax output say y2.

Andrew Ng

Beam search ($B = 3$)

$B = 1$ ⇝ greedy search



$P(y^{<3>} \mid x, \text{"in september"})$

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

jane visits africa in september. <EOS>

Andrew Ng