Recurrent Neural Networks III

Suleyman Demirel University

CSS634: Deep Learning

PhD Abay Nussipbekov

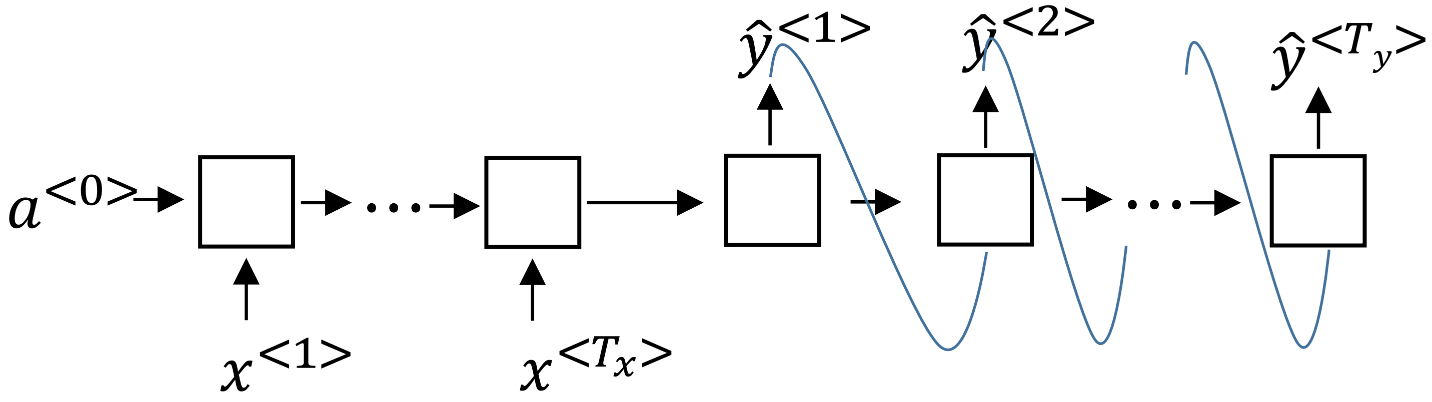
Basic Sequence to Sequence Model (Translation)

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

Jane visite l'Afrique en septembre

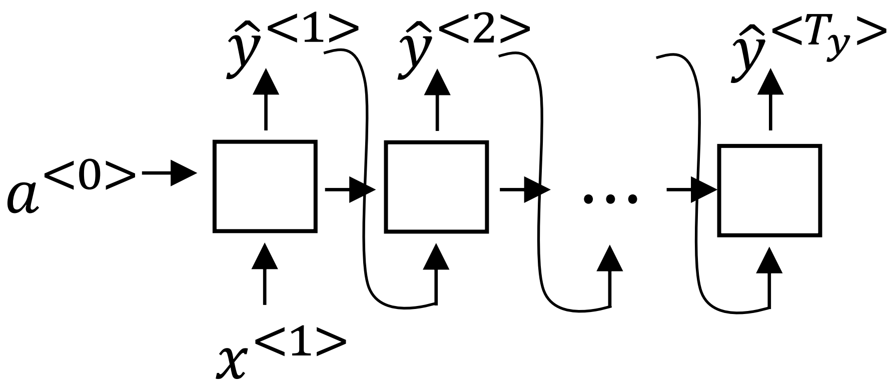
$\longrightarrow$ Jane is visiting Africa in September.

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



Machine Translation as Building a Conditional Language Model

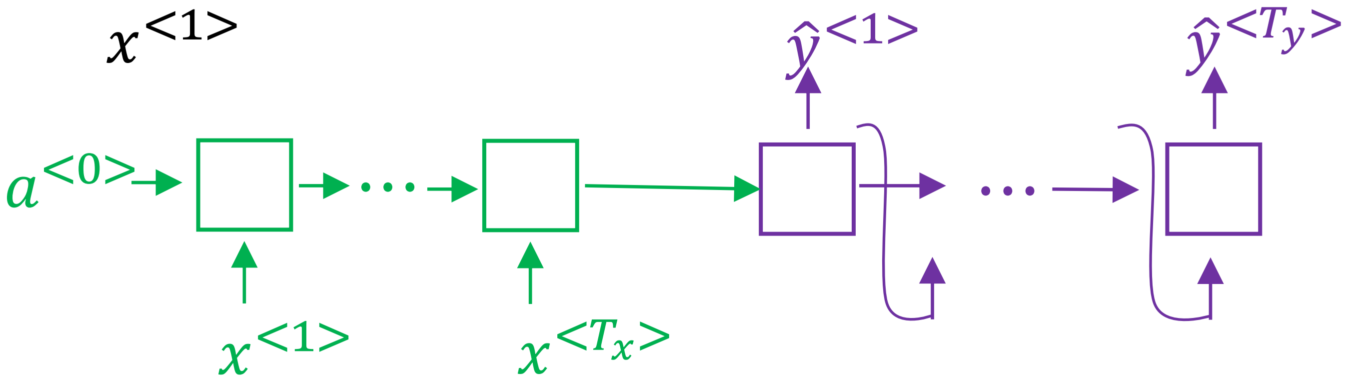
Language model:



Language mode is essentially calculating the probability of an output sequence

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

Machine translation:



Same as in language model but we give an encoded input as a first hidden state rather that all zeros.

"Conditional language model": $P\left(\hat{y}^{<1>}, \dots, \hat{y}^{<T_y>} \mid x^{<1>}, \dots, x^{<T_x>}\right)$

Finding the Most Likely Translation

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

→ Jane is visiting Africa in September.

→ Jane is going to be visiting Africa in September.

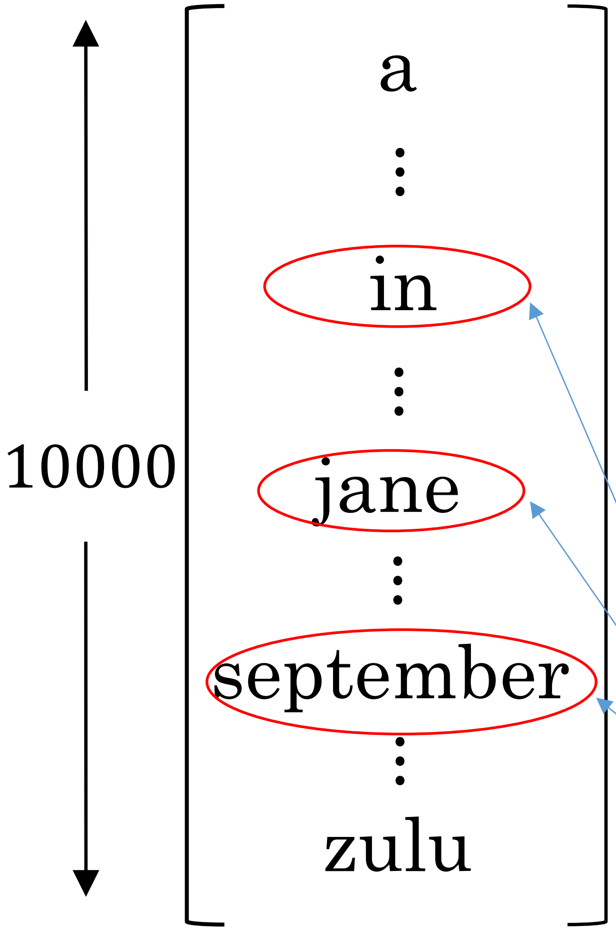
→ In September, Jane will visit Africa.

→ Her African friend welcomed Jane in September.

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

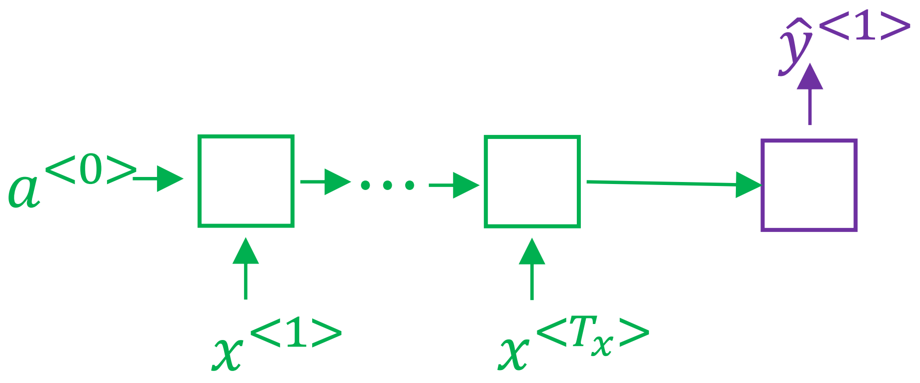
Beam Search Algorithm

Step 1



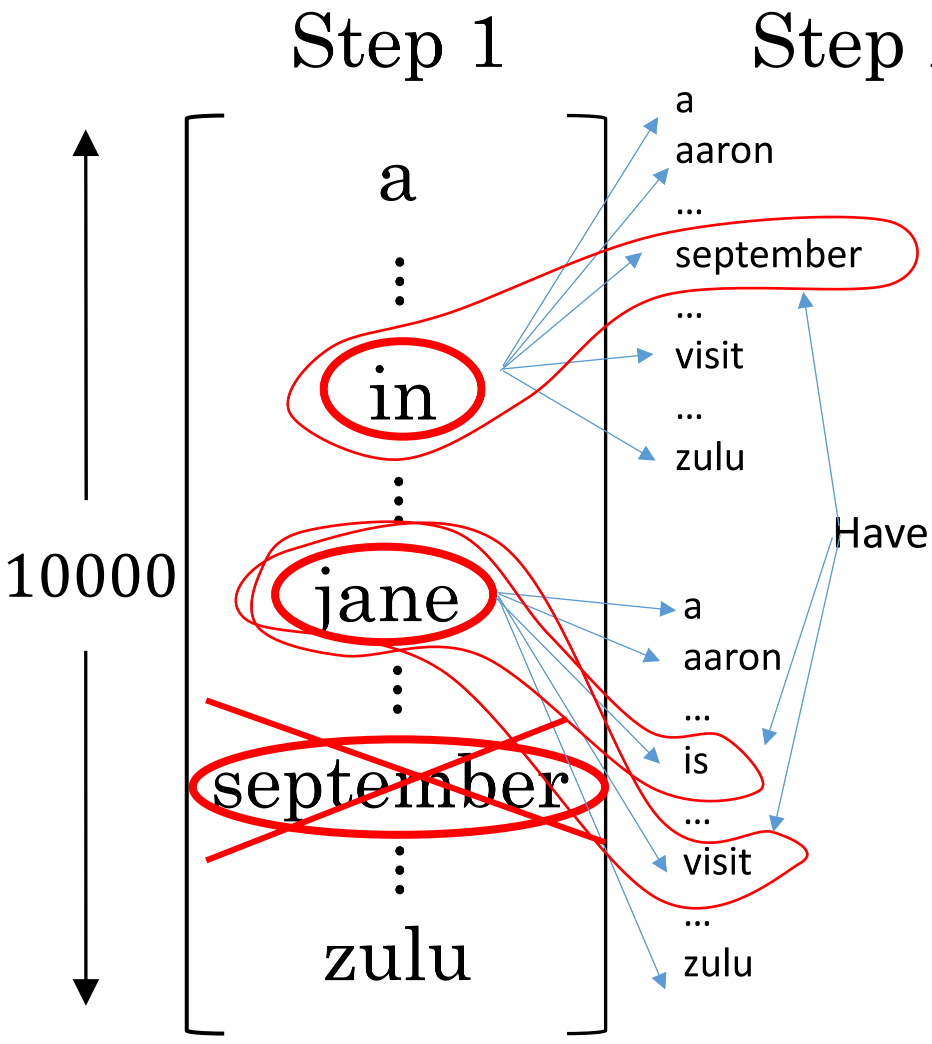
B = 3 (beam width)

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

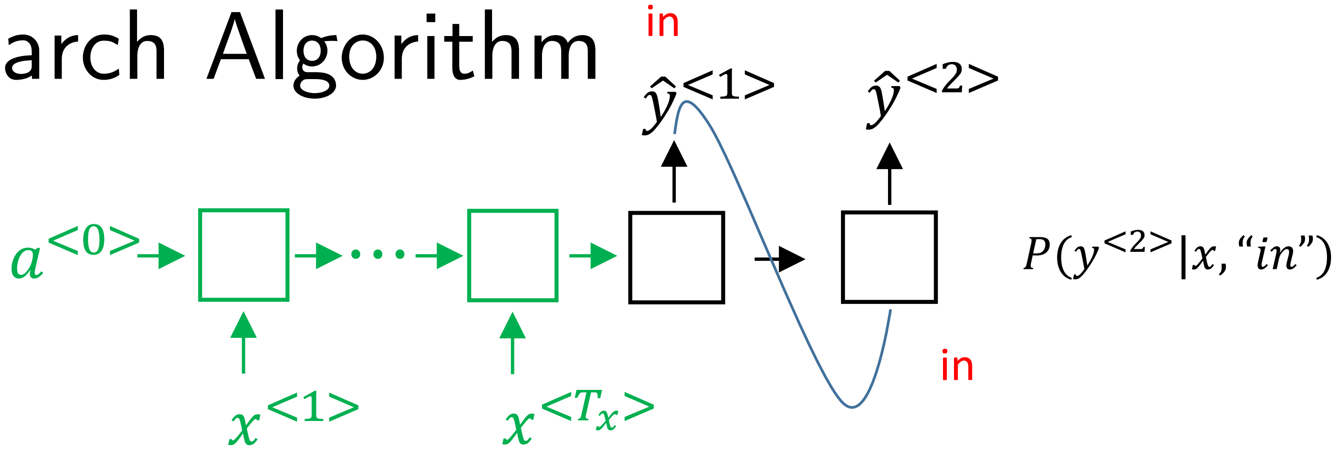


We start not with the one with highest probability but with top B (3) items

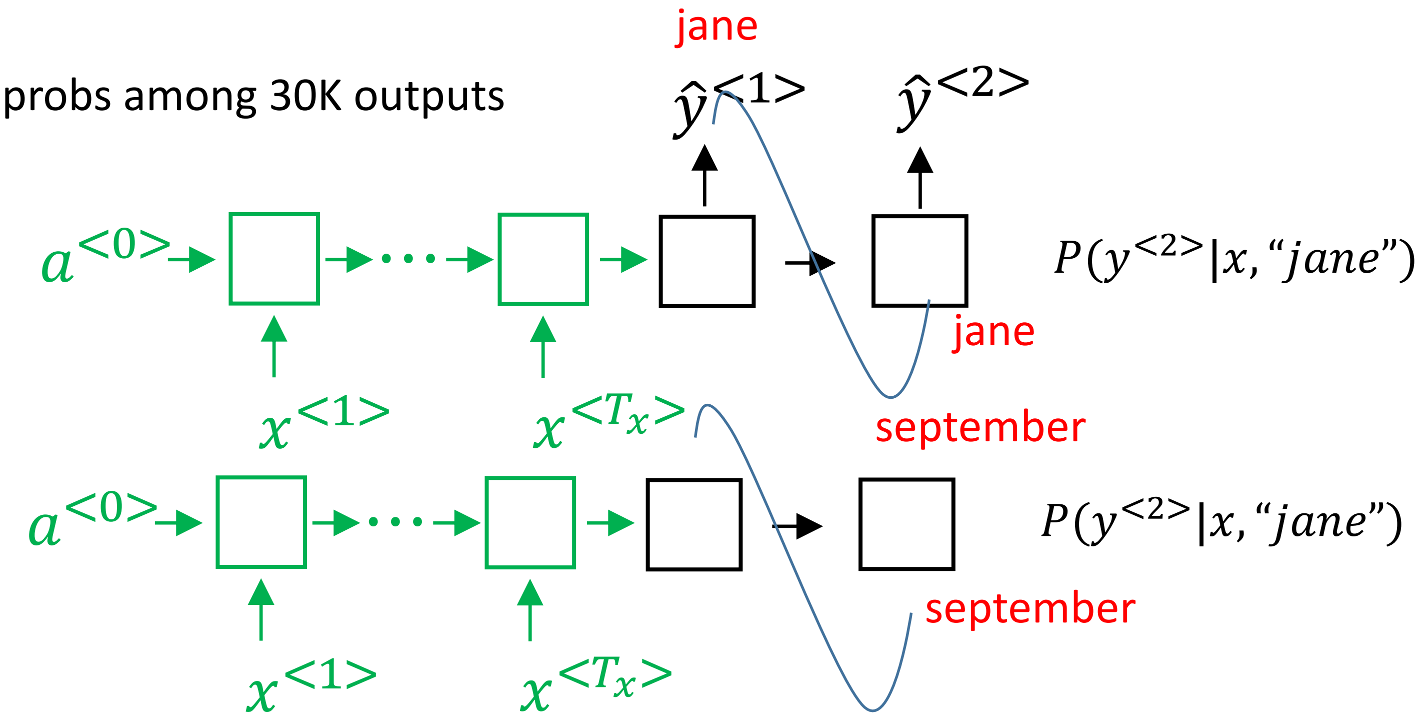
Beam Search Algorithm



Have highest probs among 30K outputs



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



Beam Search ($B = 3$)

in september

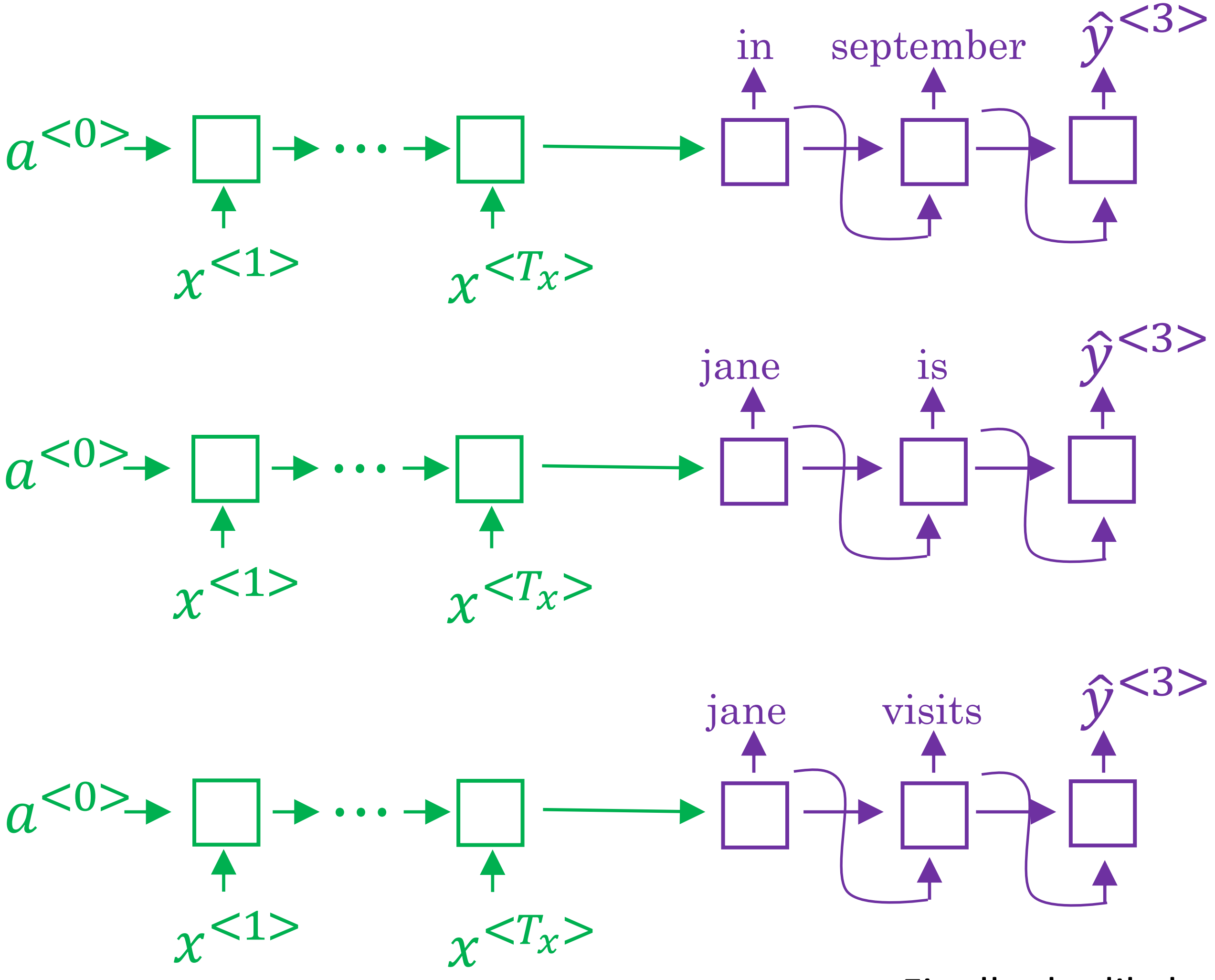
a
...
jane
...
zulu

jane is

a
...
visiting
...
zulu

jane visits

a
...
africa
...
zulu



Finally the likely output

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

jane visits africa in september. <EOS>

Beam Search Discussion

Beam width B?

Large B: better result, slower
Small B: worse result, faster

Up to 10 is fine for production

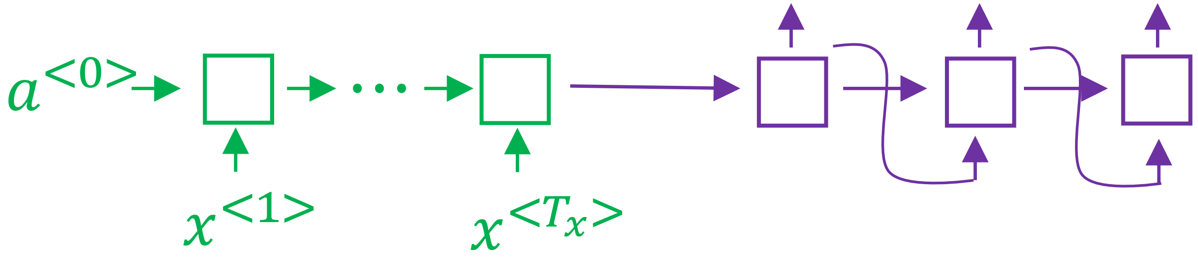
Unlike exact search algorithms like BFS (Breadth First Search) or DFS (Depth First Search), Beam Search runs faster but is not guaranteed to find exact maximum for $\arg\max_{y} P(y|x)$.

Error Analysis on Beam Search

Jane visite l'Afrique en septembre.

Human: Jane visits Africa in September.

Algorithm: Jane visited Africa last September.



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.

Evaluating Machine Translation

French: Le chat est sur le tapis.

2 appearances

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.

Precision: $\frac{7}{7}$ Modified precision: $\frac{2}{7}$

Count “the”

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	$\textbf{Count}_{\textbf{clip}}$
the cat	2	1
cat the	1	0
cat on	1	1
on the	1	1
the mat	1	1

$$\frac{4}{6}$$

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.

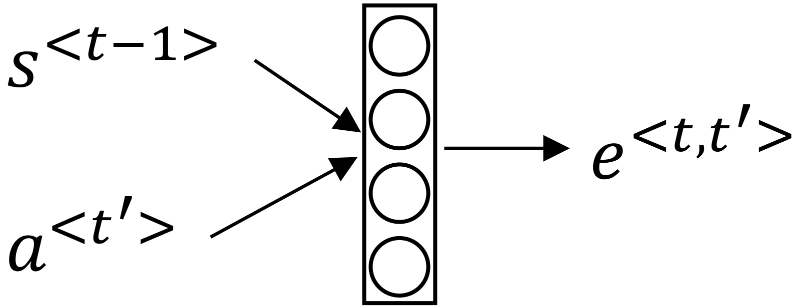
$$p_1 = \frac{\sum_{unigram \in \hat{y}} count_{clip}\,(unigram)}{\sum_{unigram \in \hat{y}} count\,(unigram)}$$

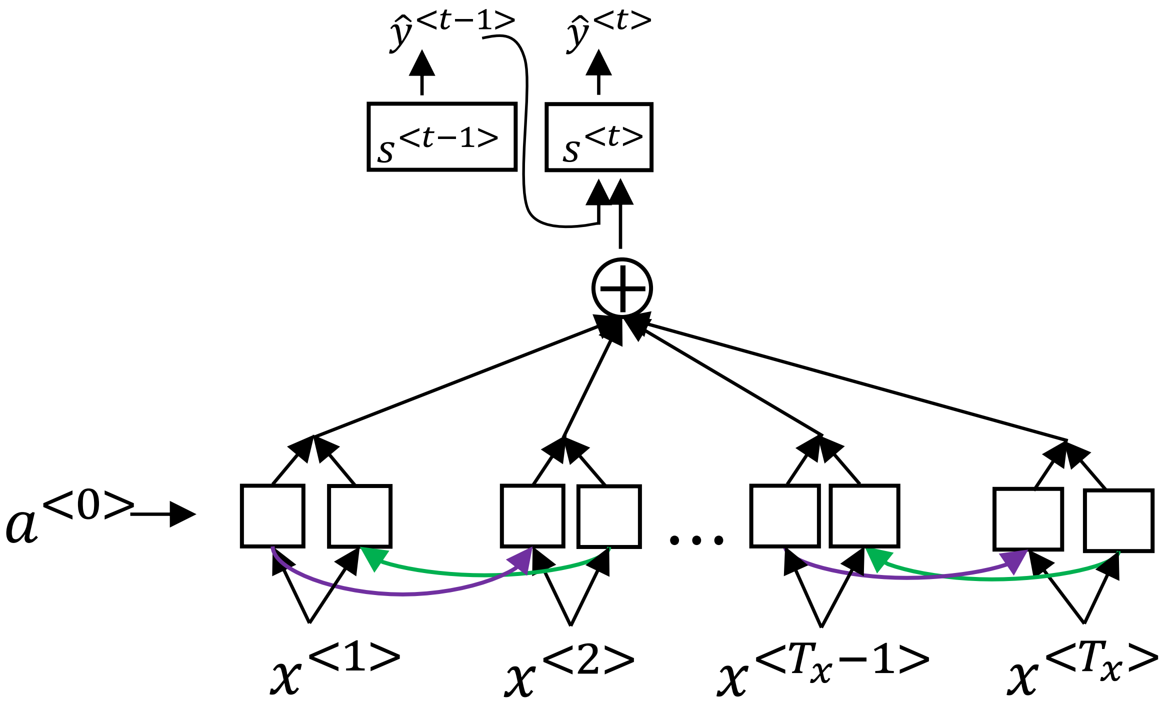
$$p_n = \frac{\sum_{ngram \in \hat{y}} count_{clip}\,(ngram)}{\sum_{ngram \in \hat{y}} count\,(ngram)}$$

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

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

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

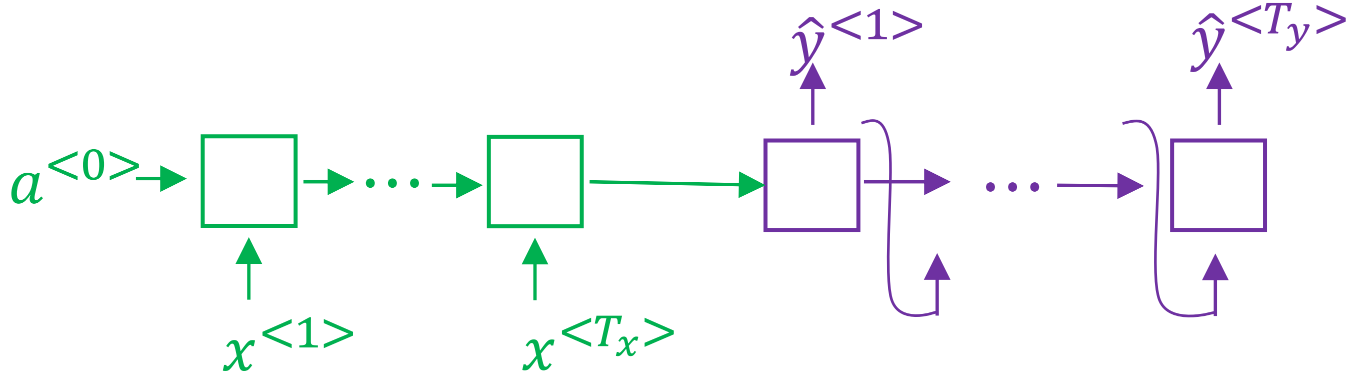




[Bahdanau et. al., 2014. Neural machine translation by jointly learning to align and translate]

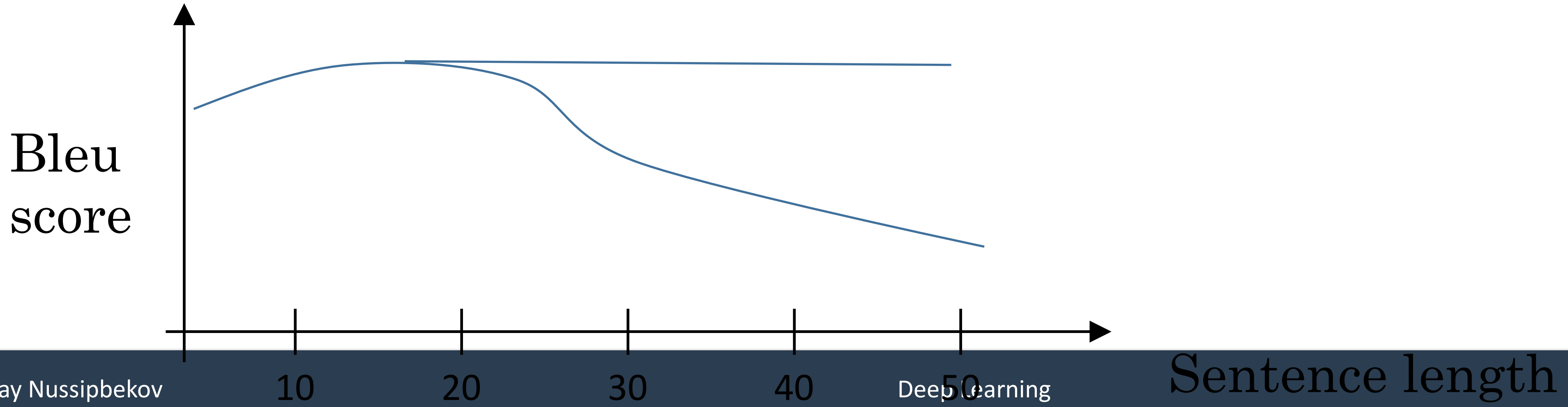
[Xu et. al., 2015. Show, attend and tell: Neural image caption generation with visual attention]

The Problem of Long Sequences

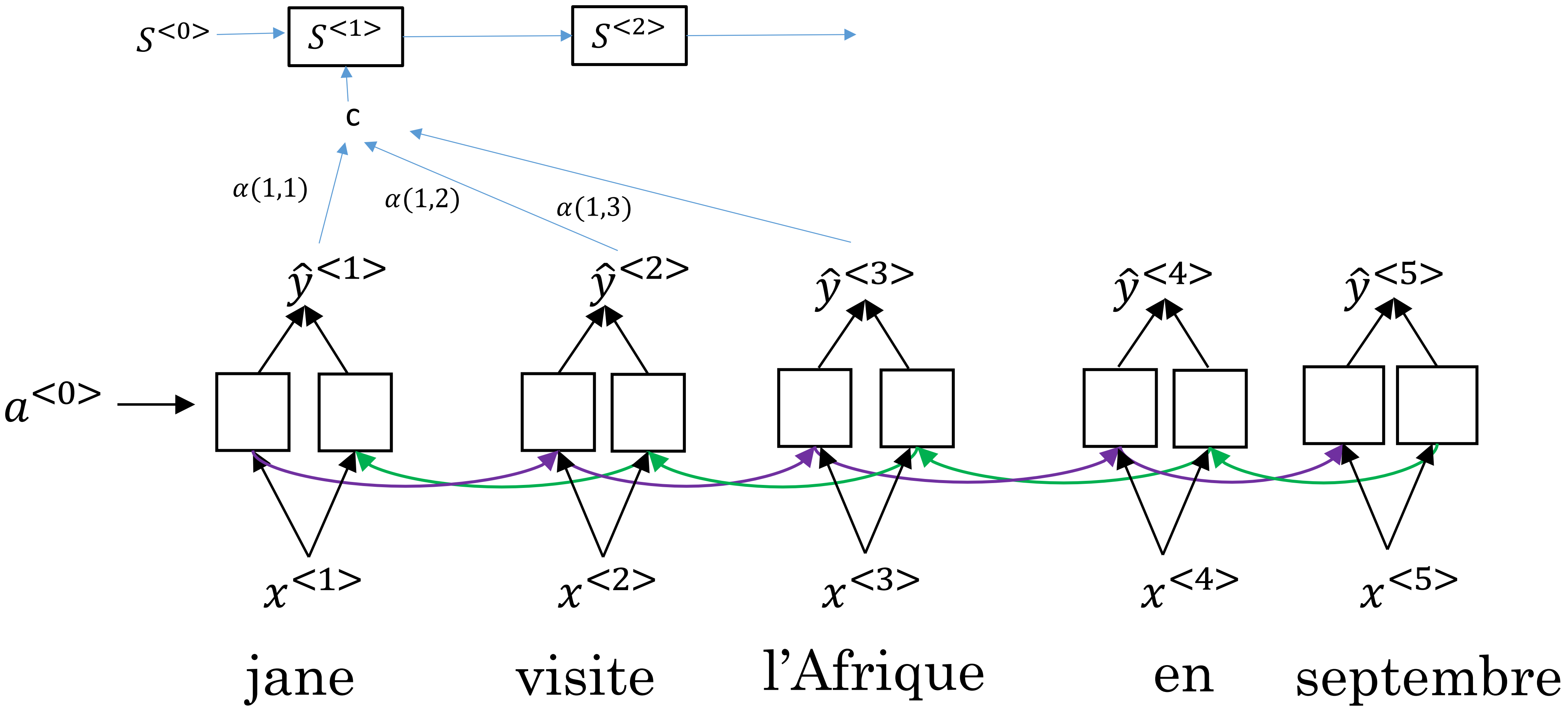


Jane s'est rendue en Afrique en septembre dernier, a apprécié la culture et a rencontré beaucoup de gens merveilleux; elle est revenue en parlant comment son voyage était merveilleux, et elle me tente d'y aller aussi.

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.

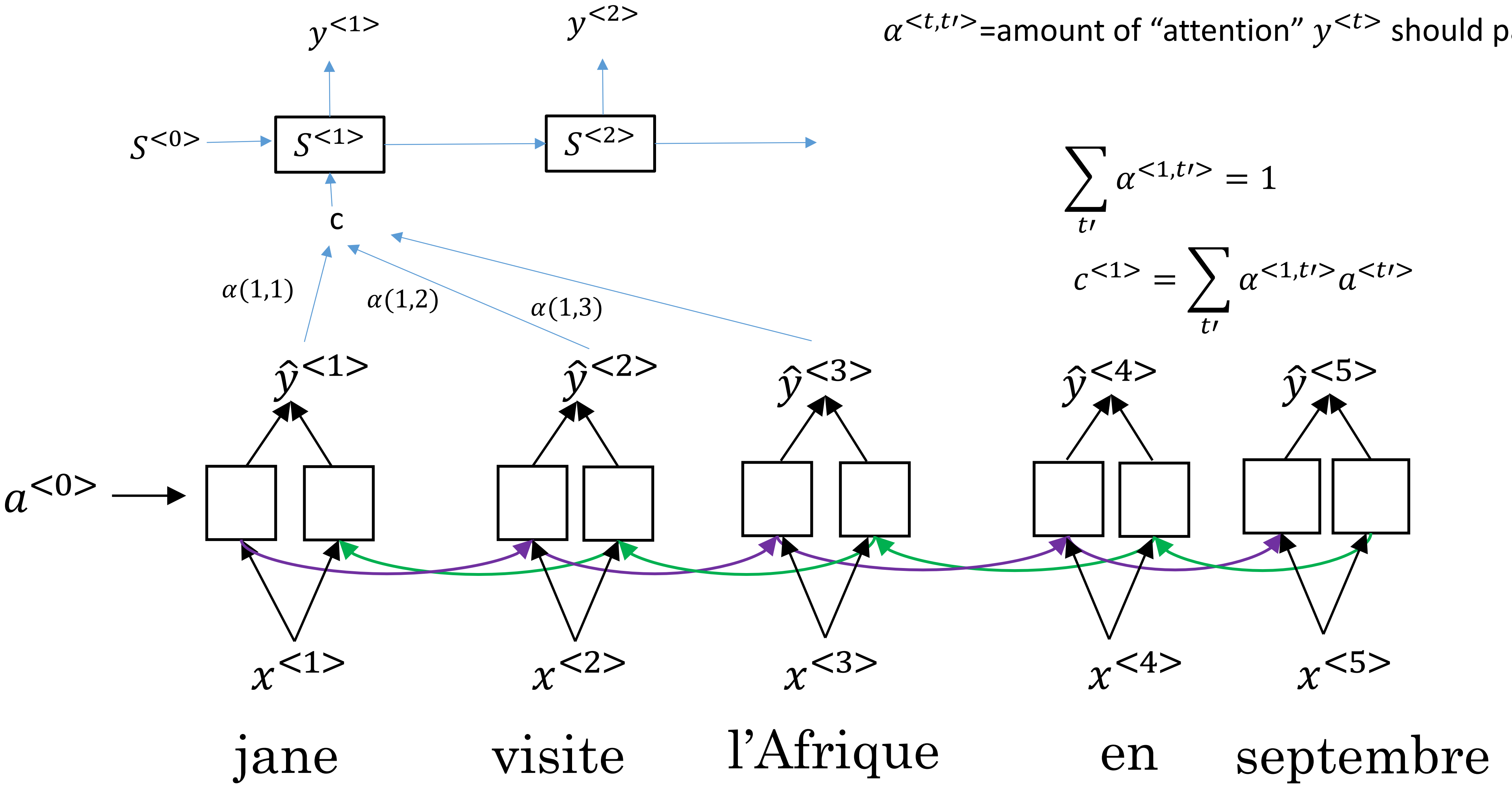


Attention Model Intuition



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

Attention Model



$\alpha^{<t,t\prime>}$=amount of "attention" $y^{<t>}$ should pay to $a^{<t\prime>}$

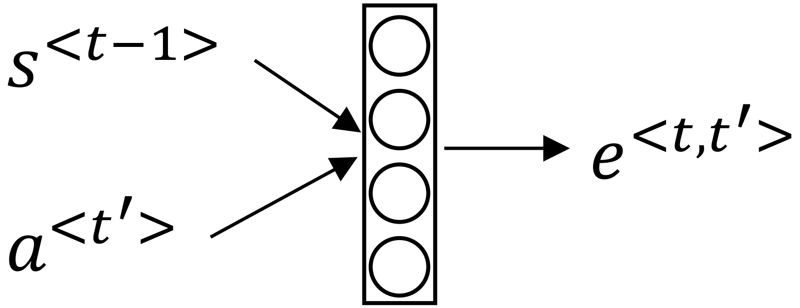
$$\sum_{t\prime} \alpha^{<1,t\prime>} = 1$$

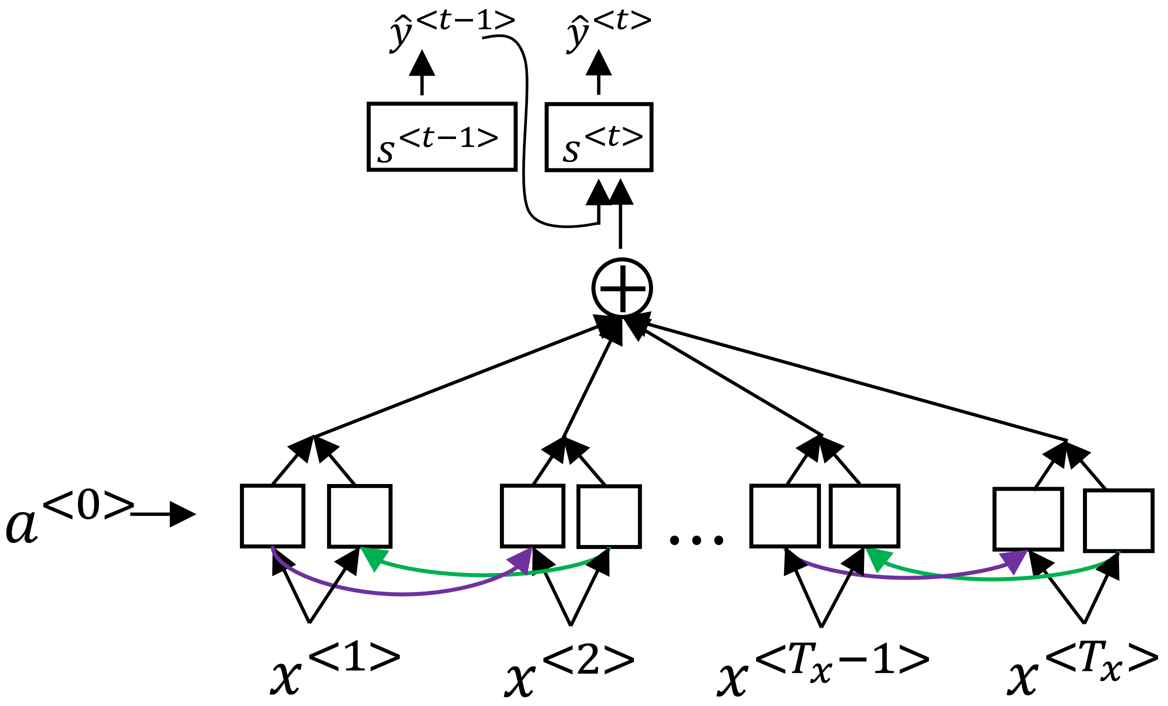
$$c^{<1>} = \sum_{t\prime} \alpha^{<1,t\prime>} a^{<t\prime>}$$

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

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

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





Resources Used

- Deeplearning.ai by Andrew Ng