deeplearning.ai

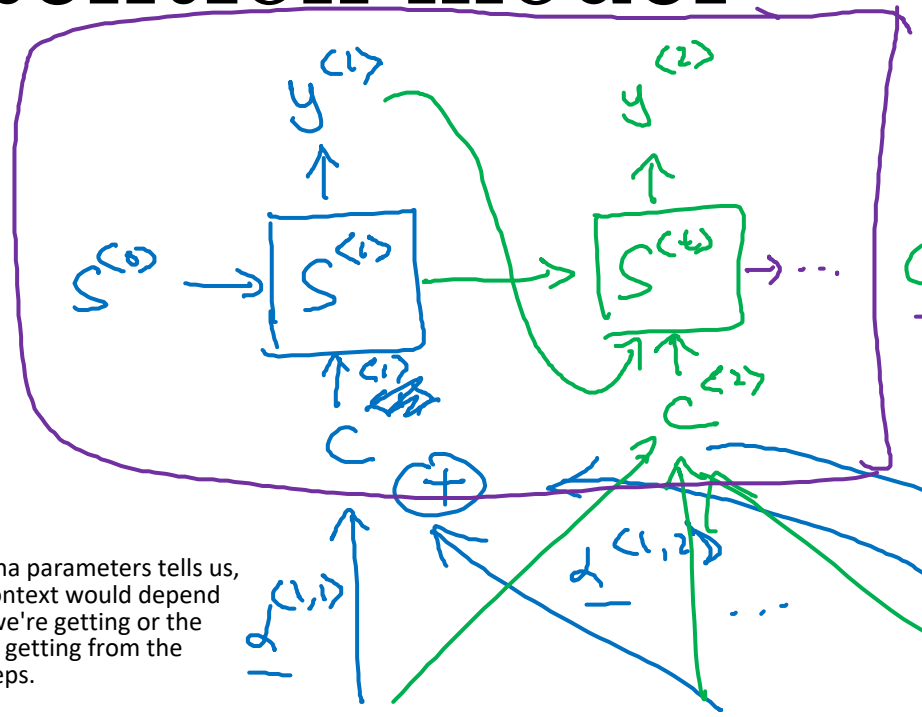
Sequence to sequence models

Attention model

Let's now formalize that intuition into the exact detail of how you would implement an attention model.

Attention model

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

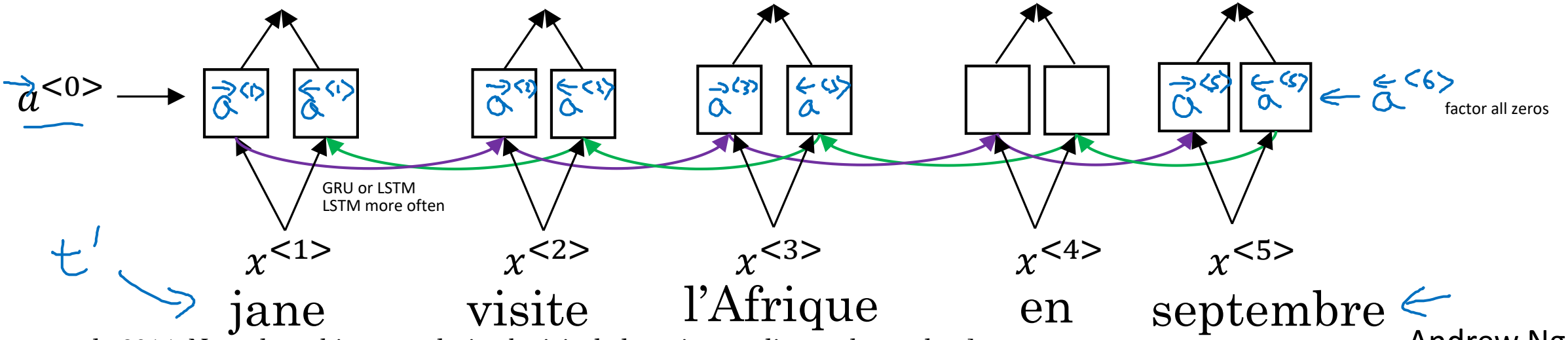


$$c^{<1>} = \sum_{t'} \alpha^{<1,t'>} a^{<t'>}$$

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

$$\sum_{t'} \alpha^{<1,t'>} = 1$$ zero positive sum to 1

And so these alpha parameters tells us, how much the context would depend on the features we're getting or the activations we're getting from the different time steps.



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

Andrew Ng

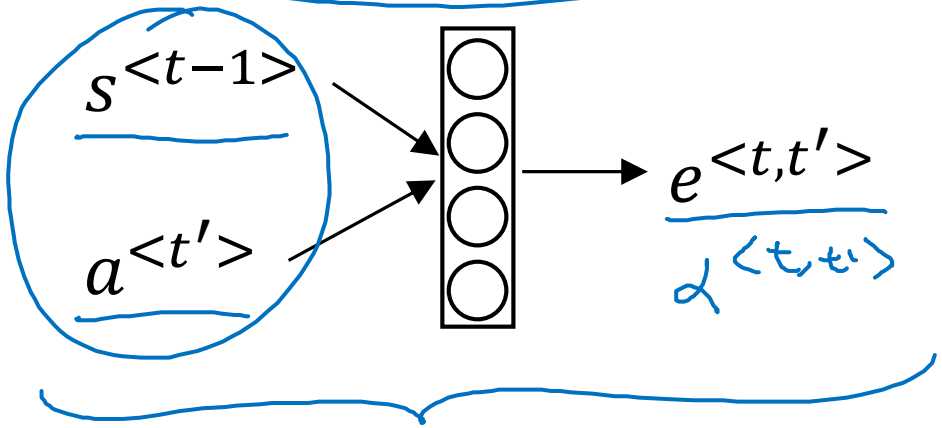
Computing attention $\alpha^{<t,t'>}$

total # of attention parameters are going to be Tx times Ty

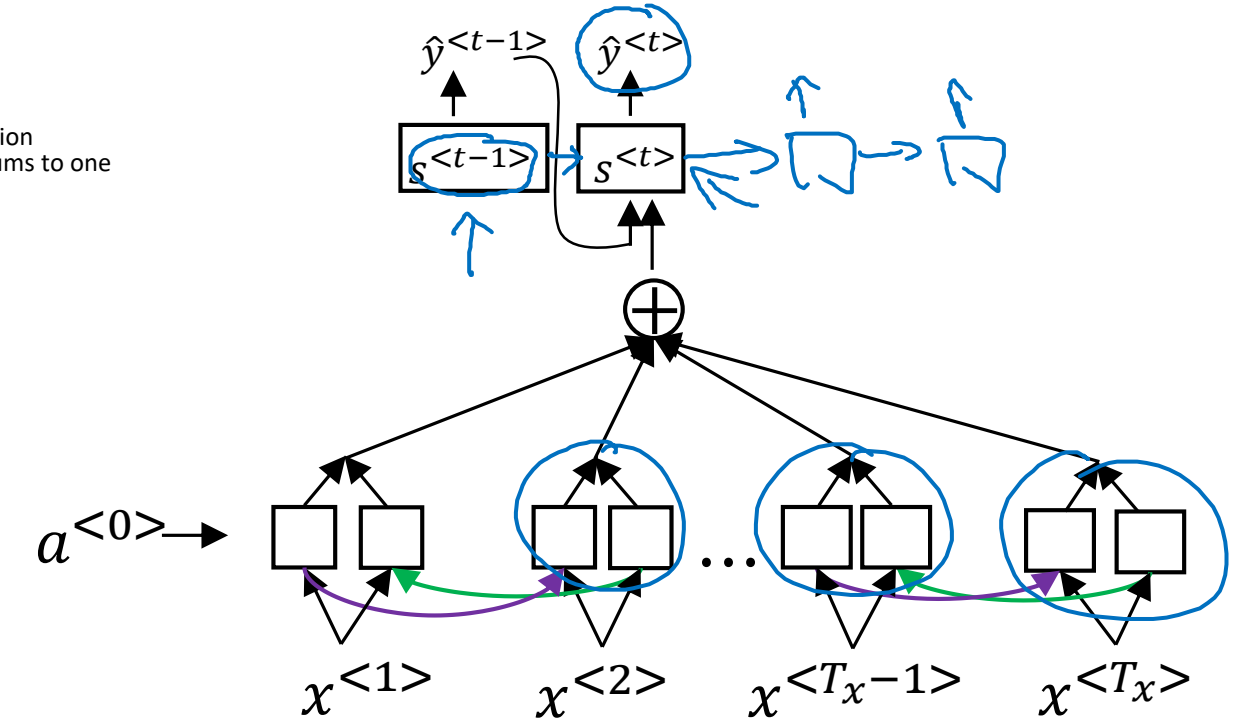
$\alpha^{<t,t'>}$ = amount of attention $y^{<t>}$ should pay to $a^{<t'>}$

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

using this softmax priorization just ensure this property sums to one



Although in machine translations where neither input nor output sentences is usually that long, maybe quadratic is actually acceptable.
Although there is some research work on trying to reduce costs as well.



One downside of this algorithm is that it does take quadratic time or quadratic cost to run this algorithm.

[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]

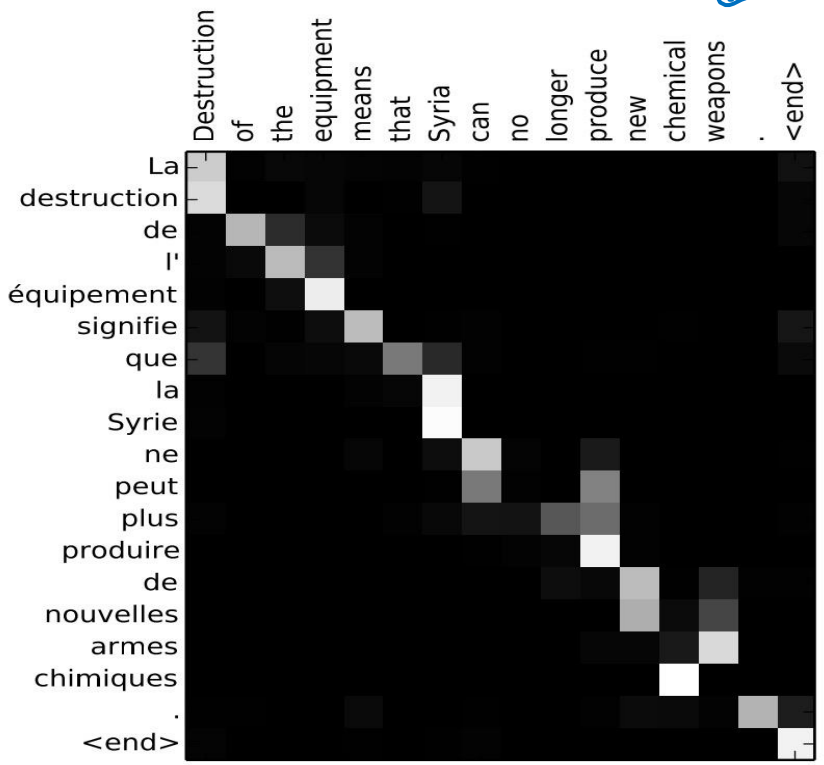
Andrew Ng

Attention examples

July 20th 1969 $\longrightarrow$ 1969 – 07 – 20

23 April, 1564 $\longrightarrow$ 1564 – 04 – 23

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



Andrew Ng