Copyright Notice

deeplearning.ai

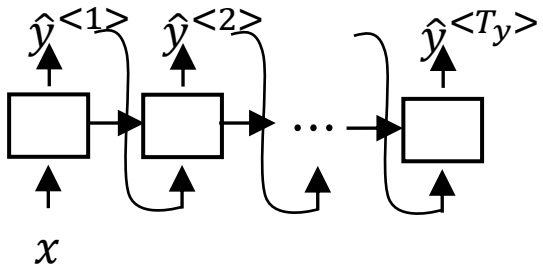
Sequence to sequence models

Transformers Intuition

Transformers Motivation

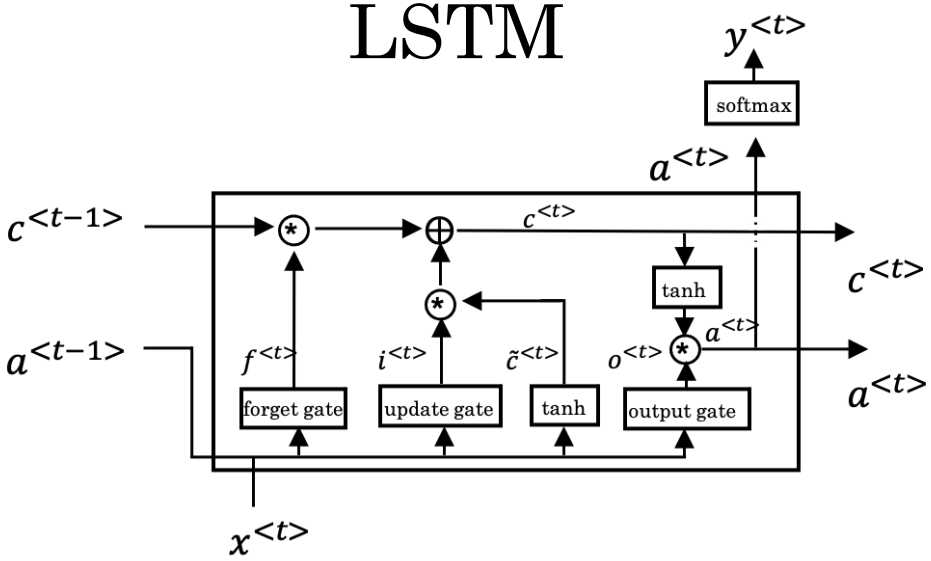
Increased complexity, sequential

RNN



GRU

LSTM

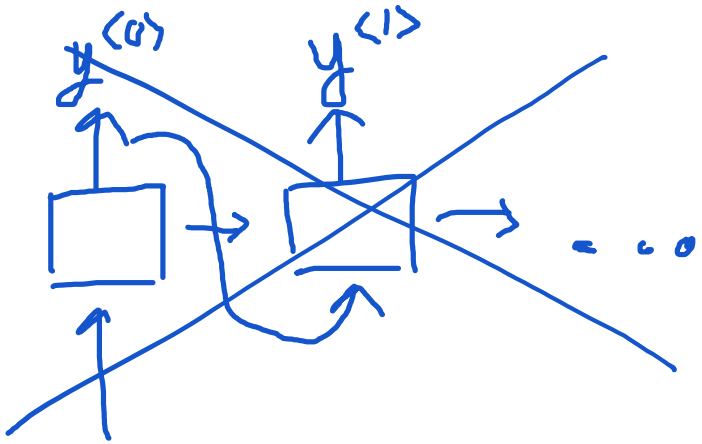


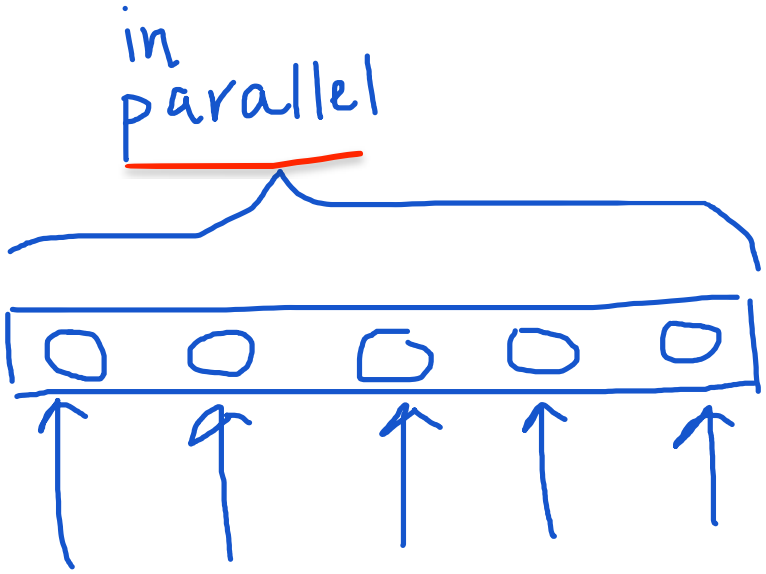
Andrew Ng

Transformers Intuition

- Attention + CNN
  - Self-Attention
  - Multi-Head Attention

for loop of self attention





[Vaswani et al. 2017, Attention Is All You Need]

deeplearning.ai

Sequence to sequence models

Self-Attention

Self-Attention Intuition

$A(q, K, V)$= attention-based vector representation of a word

calculate for each word

RNN Attention

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

softmax

Transformers Attention

$$A(q, K, V) = \sum_i \frac{\exp(q \cdot k^{<i>})}{\sum_j \exp(q \cdot k^{<j>})} v^{<i>}$$

"word embedding" that consider neighbours words

$x^{<1>}$	$x^{<2>}$	$x^{<3>}$	$x^{<4>}$	$x^{<5>}$
Jane	visite	l'Afrique	en	septembre

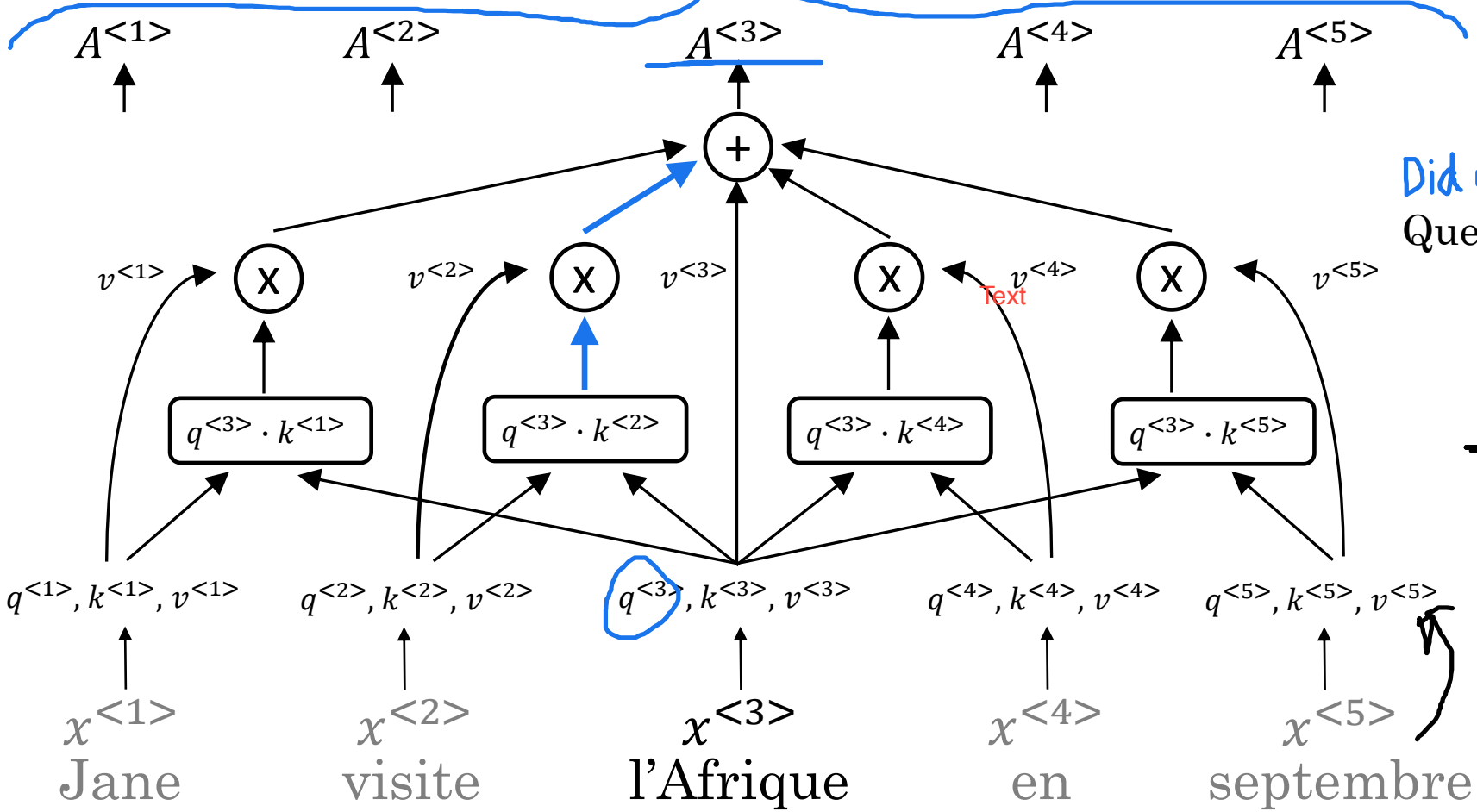
[Vaswani et al. 2017, Attention Is All You Need]

Self-Attention

$$A(q,K,V) = \sum_i \frac{\exp(e^{<q \cdot k^{<i>}>})}{\sum_j \exp(e^{<q \cdot k^{<j>}>})} v^{<i>}$$

Softmax

$$\textbf{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \textbf{softmax}\left(\frac{\boldsymbol{QK^T}}{\sqrt{\boldsymbol{d_k}}}\right)\boldsymbol{V}$$



Did what?

Query (Q)	Key (K)	Value (V)
$q^{<1>}$	$k^{<1>}$	$v^{<1>}$
$q^{<2>}$	$k^{<2>}$	$v^{<2>}$
$q^{<3>}$	$k^{<3>}$	$v^{<3>}$
$q^{<4>}$	$k^{<4>}$	$v^{<4>}$
$q^{<5>}$	$k^{<5>}$	$v^{<5>}$

person, action — Jane, visit

What's happening there?

W^Q, W^K, W^V

[Vaswani et al. 2017, Attention Is All You Need]

deeplearning.ai

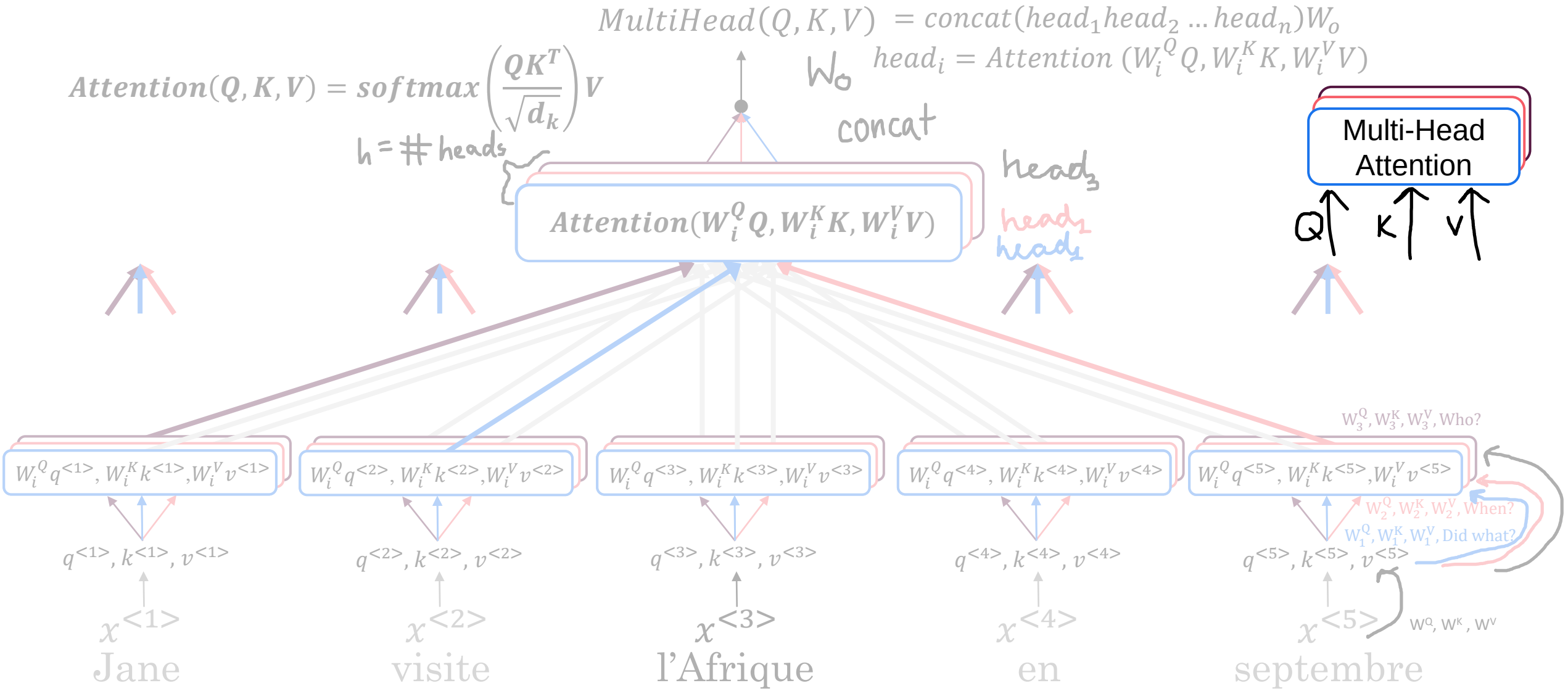
Sequence to sequence models

Multi-Head Attention

Note that in the previous video you learned that you calculate q, k and v (circled red) by multiplying x with matrices W. In case of the multi-head attention, you don't need to do this, as you already have the matrices W i

in each head, and you would effectively do the calculation twice if you did the multiplication here also.
In the simplest case of multi-headed self-attention you would actually use
q=k=v=x. The reason we anyway show q, k and v here as different values is that in one part of the transformer
(where you calculate the attention between the input and output) they are not all the same,
as they carry different information as you will learn in the next video.

Multi-Head Attention



[Vaswani et al. 2017, Attention Is All You Need]

Andrew Ng

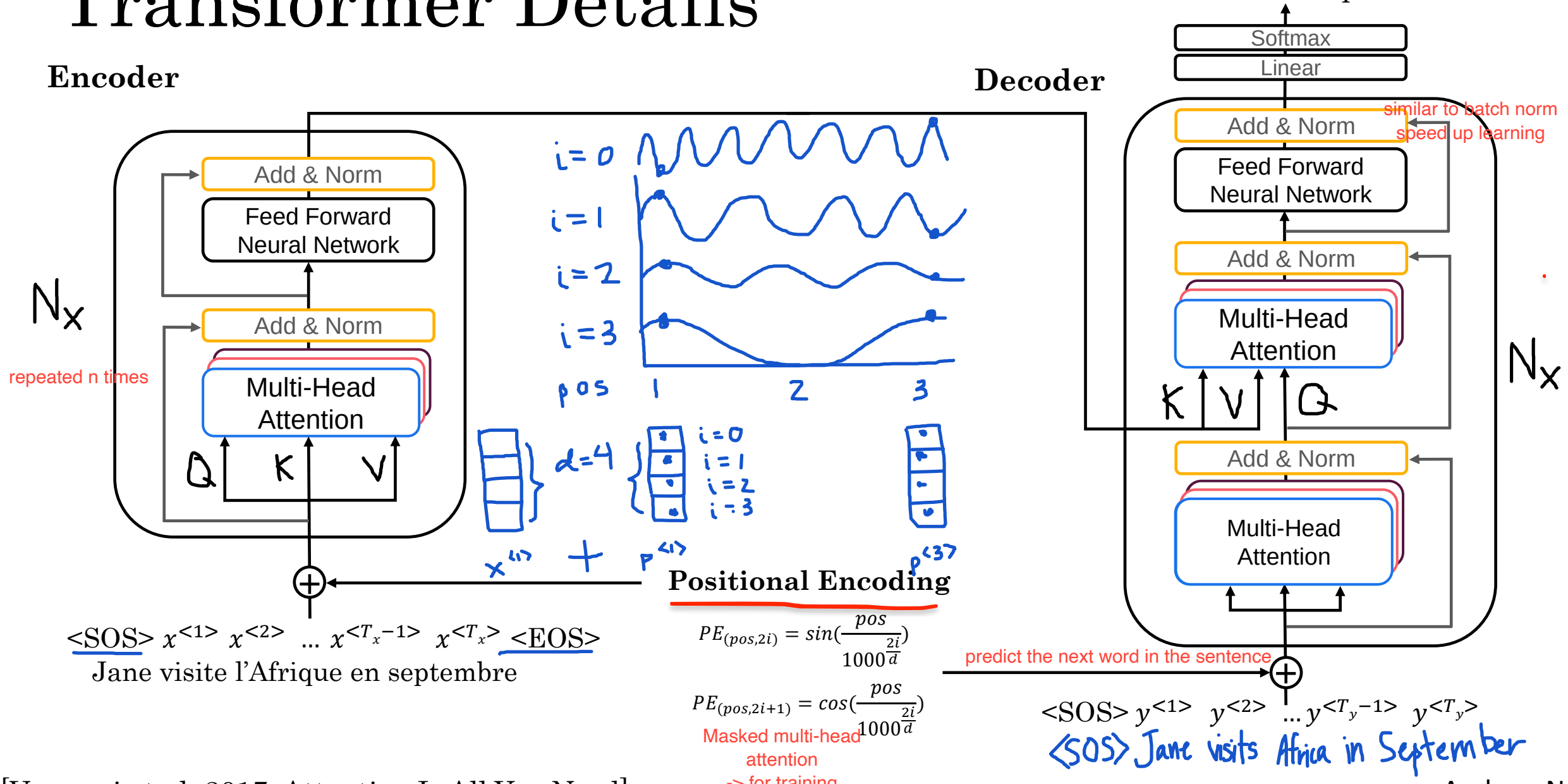
deeplearning.ai

Sequence to sequence models

Transformers

Transformer Details

<SOS> Jane visits Africa in September <EOS>



[Vaswani et al. 2017, Attention Is All You Need]

Andrew Ng