deeplearning.ai

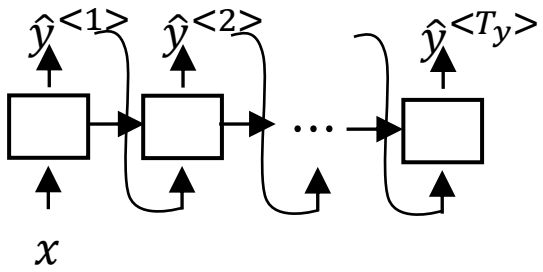
Sequence to sequence models

Transformers Intuition

Transformers Motivation

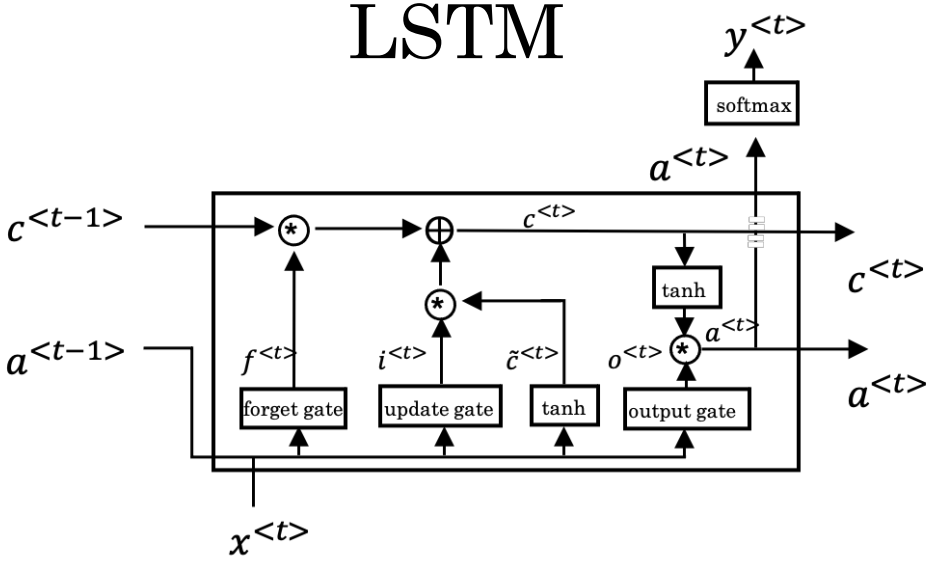
Increased complexity, sequential

RNN



GRU

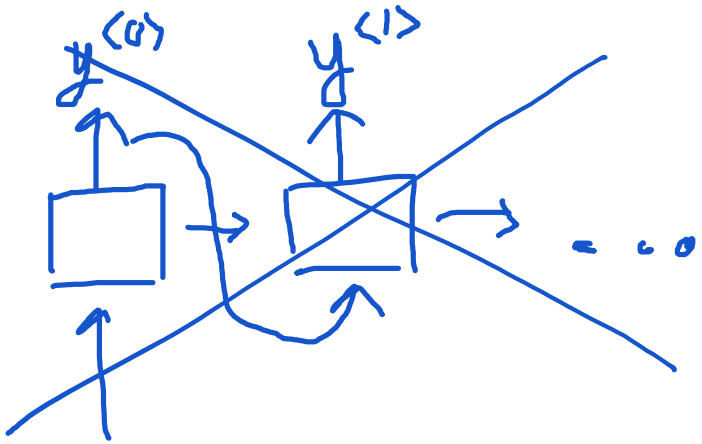
LSTM

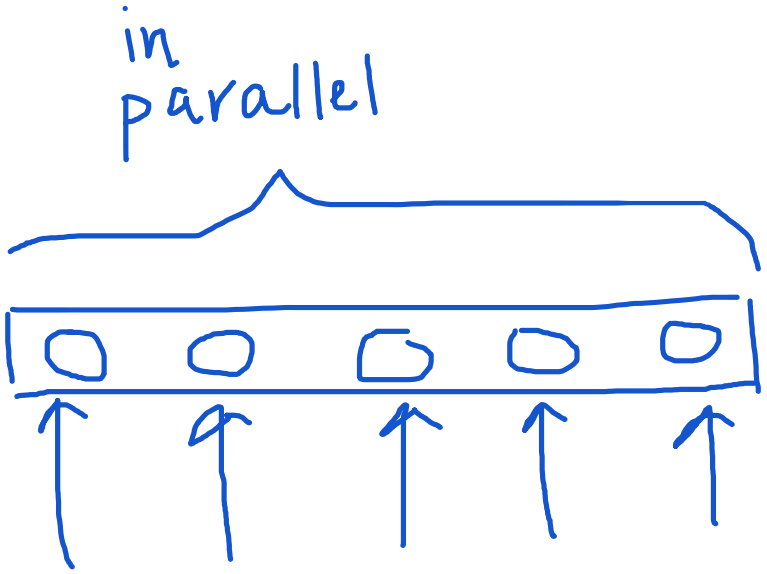


Andrew Ng

Transformers Intuition

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





[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

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

Transformers Attention

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

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

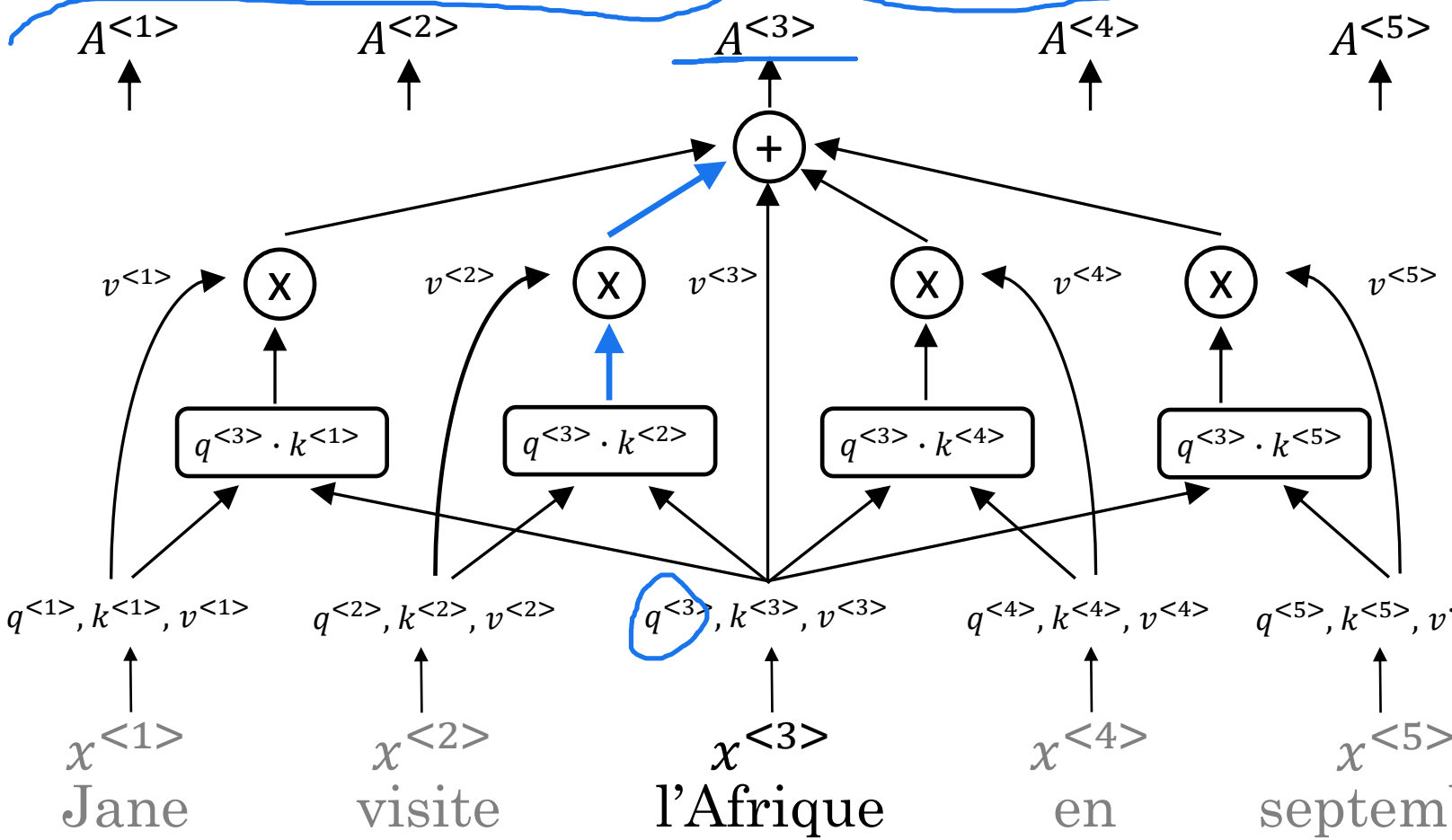
[Vaswani et al. 2017, Attention Is All You Need]

Self-Attention

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

Softmax

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



Did what?

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

What's happening there?

person action

Jane visit

W^Q, W^K, W^V

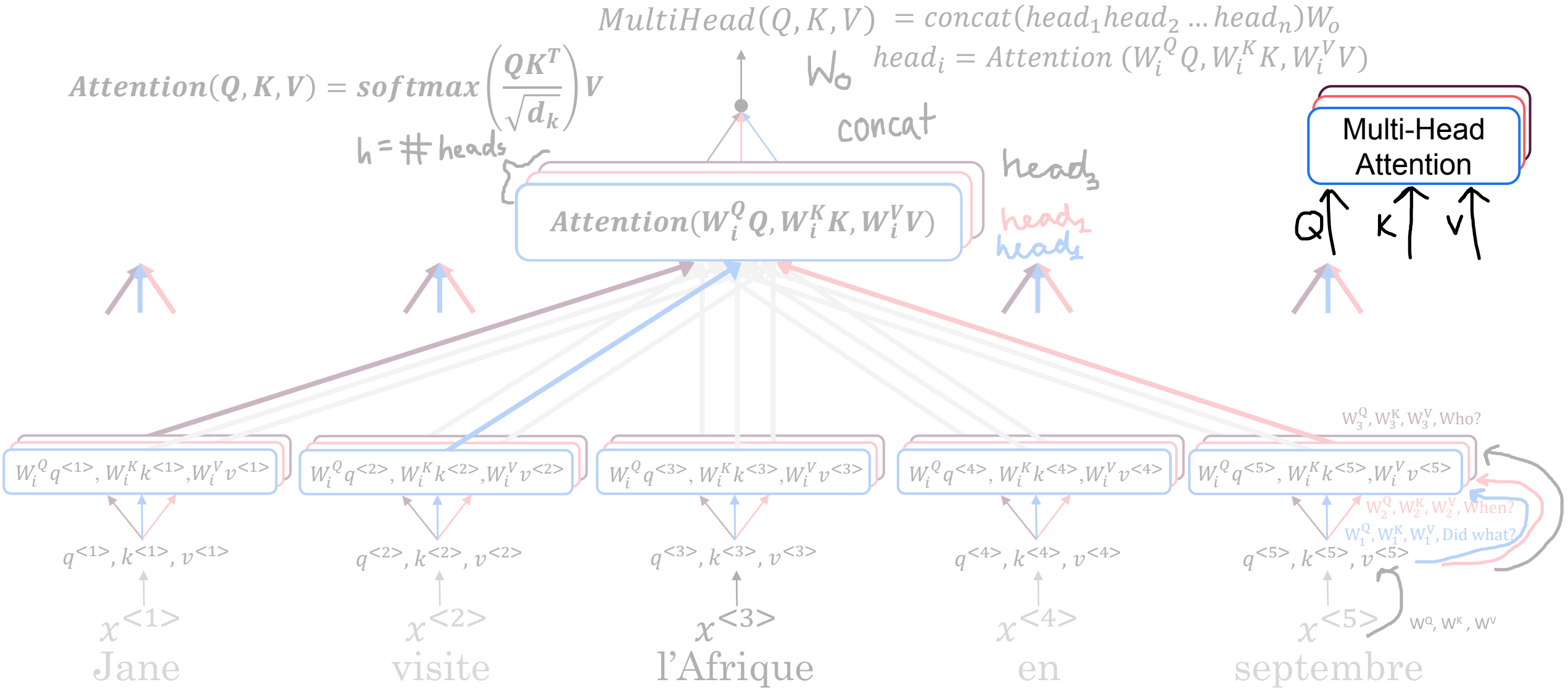
[Vaswani et al. 2017, Attention Is All You Need]

deeplearning.ai

Sequence to sequence models

Multi-Head Attention

Multi-Head Attention



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

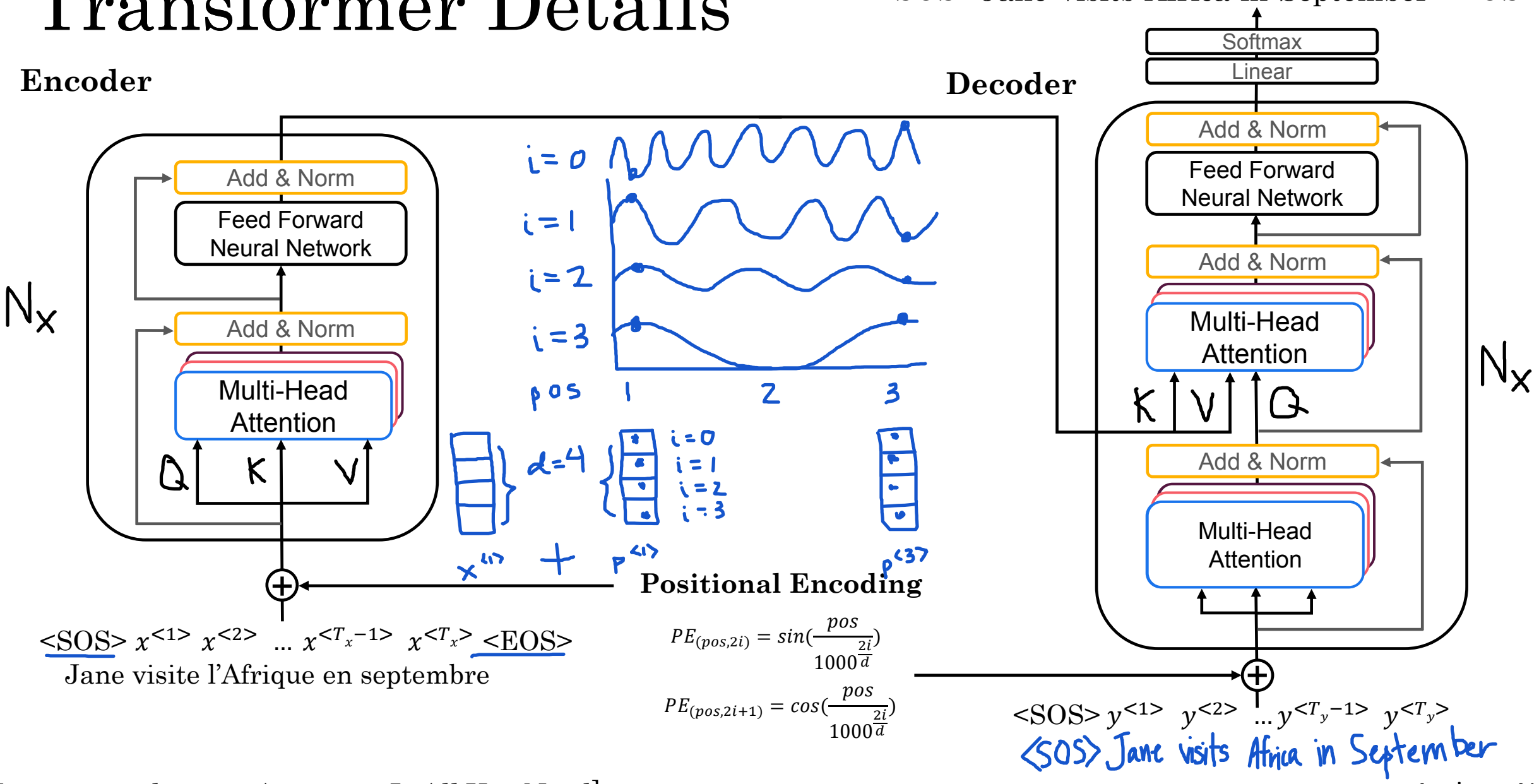
Andrew Ng

deeplearning.ai

Sequence to sequence models

Transformers

Transformer Details



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

Andrew Ng