Copyright Notice

deeplearning.ai

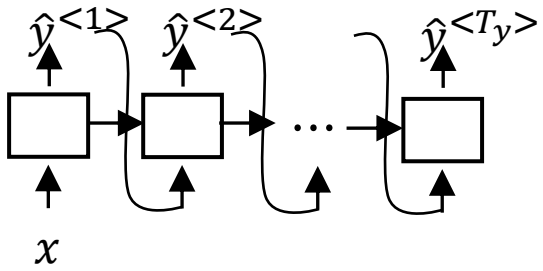
Sequence to sequence models

Transformers Intuition

Transformers Motivation

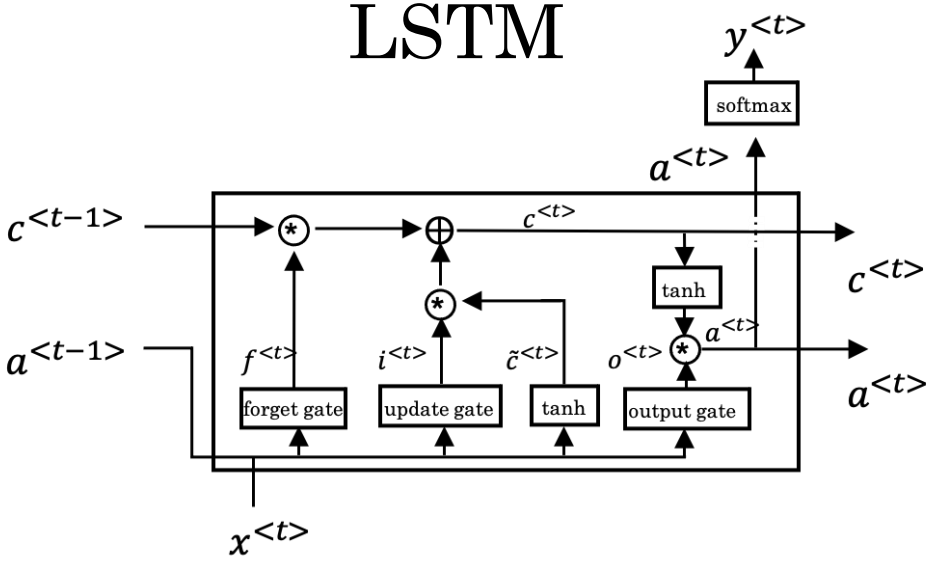
Increased complexity, sequential

RNN



GRU

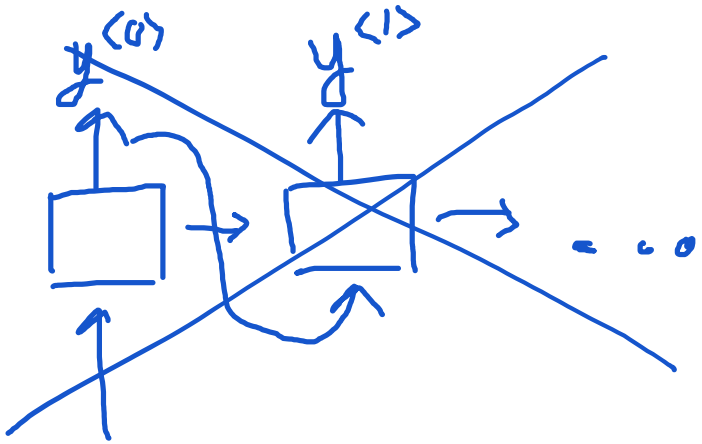
LSTM

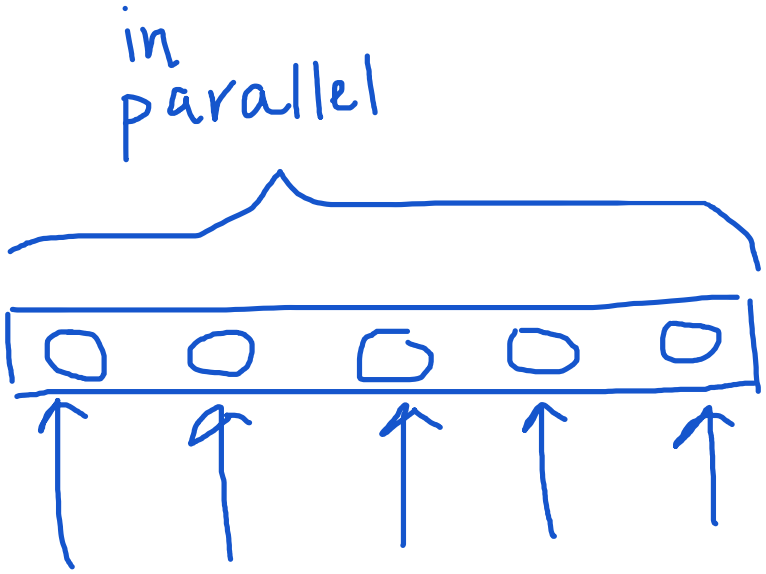


Andrew Ng

Transformers Intuition

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





[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

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

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

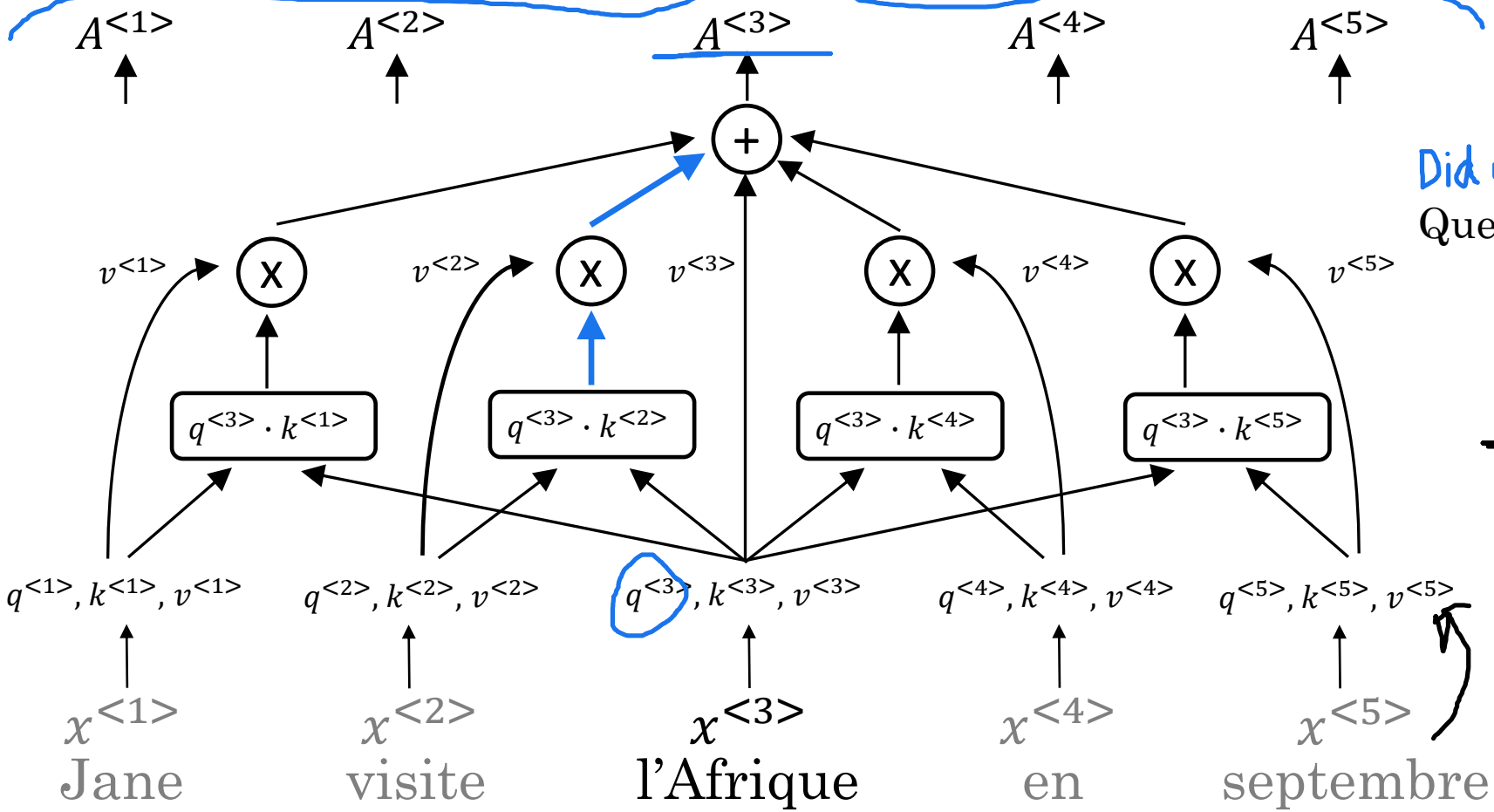
[Vaswani et al. 2017, Attention Is All You Need]

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

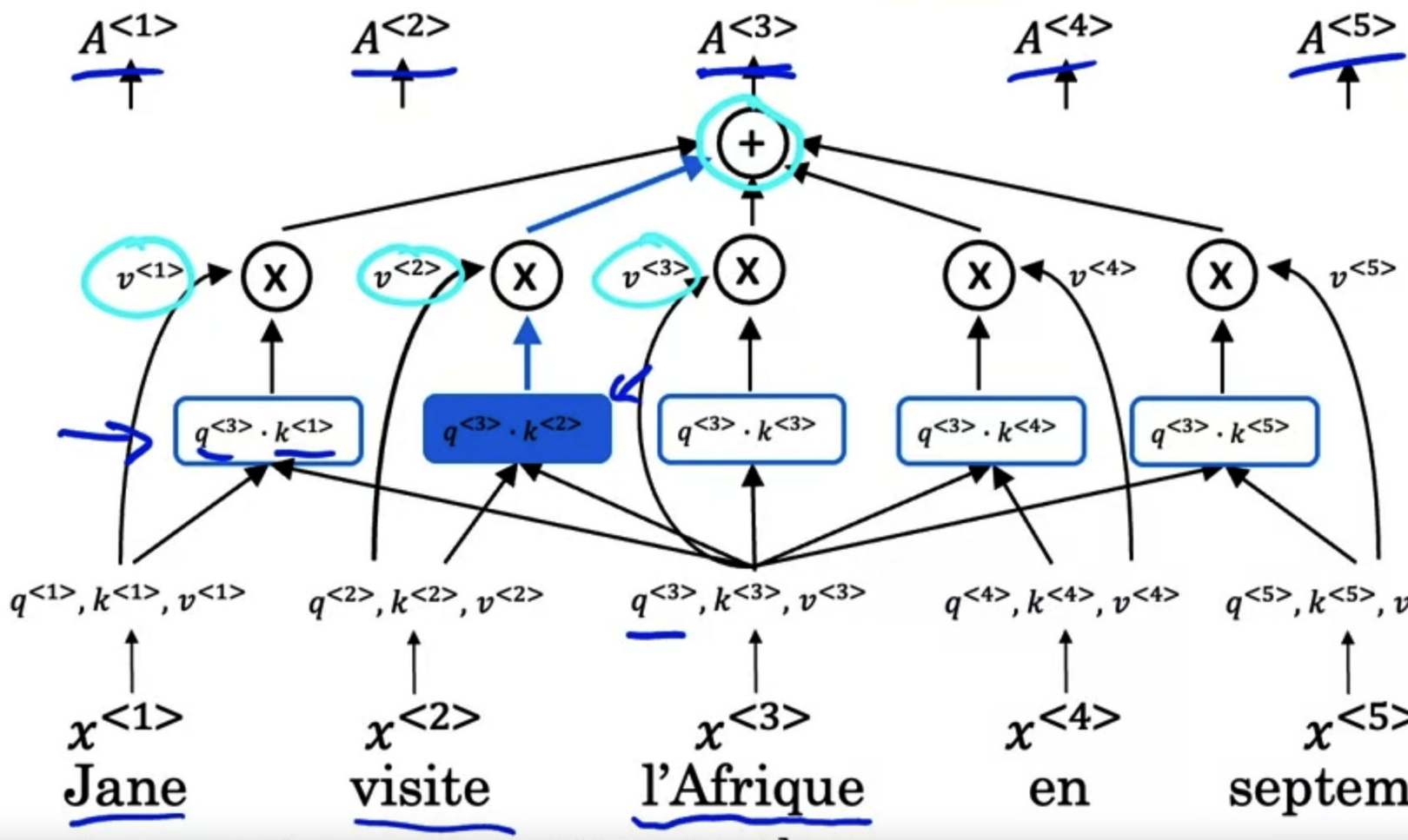
W^Q, W^K, W^V

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

Self-Attention

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

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



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
what's happening there

$$q^{<3>} = W^Q x^{<3>}$$
$$k^{<3>} = W^K x^{<3>}$$
$$v^{<3>} = W^V x^{<3>}$$

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

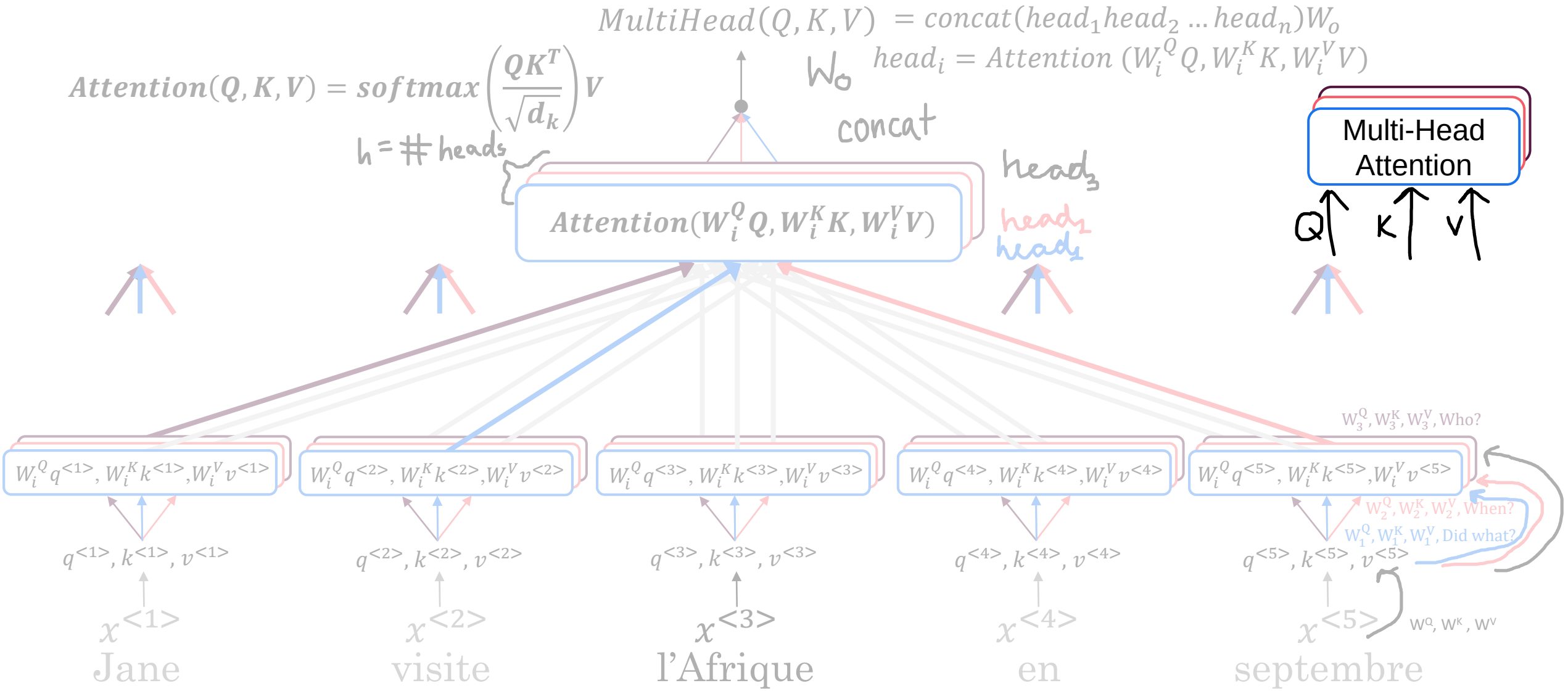
Andrew Ng

deeplearning.ai

Sequence to sequence models

Multi-Head Attention

Multi-Head Attention



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

Andrew Ng

Multi-Head Attention

"head"

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

$$MultiHead(Q,K,V) = concat(head_1 head_2 \dots head_h)W_o$$

$$head_i = Attention\,(W_i^Q Q, W_i^K K, W_i^V V)$$

h = #heads.

Multi-Head Attention

$Attention(W_i^Q Q, W_i^K K, W_i^V V)$

$W_i^Q q^{<1>}, W_i^K k^{<1>}, W_i^V v^{<1>}$ | $W_i^Q q^{<2>}, W_i^K k^{<2>}, W_i^V v^{<2>}$ | $W_i^Q q^{<3>}, W_i^K k^{<3>}, W_i^V v^{<3>}$ | $W_i^Q q^{<4>}, W_i^K k^{<4>}, W_i^V v^{<4>}$ | $W_i^Q q^{<5>}, W_i^K k^{<5>}, W_i^V v^{<5>}$

$q^{<1>}, k^{<1>}, v^{<1>}$ | $q^{<2>}, k^{<2>}, v^{<2>}$ | $q^{<3>}, k^{<3>}, v^{<3>}$ | $q^{<4>}, k^{<4>}, v^{<4>}$ | $q^{<5>}, k^{<5>}, v^{<5>}$

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

W_3^Q, W_3^K

W_3^V - who?

W_1^Q, W_1^K, W_1^V - what's happening?

W_2^Q, W_2^K, W_2^V — when?

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

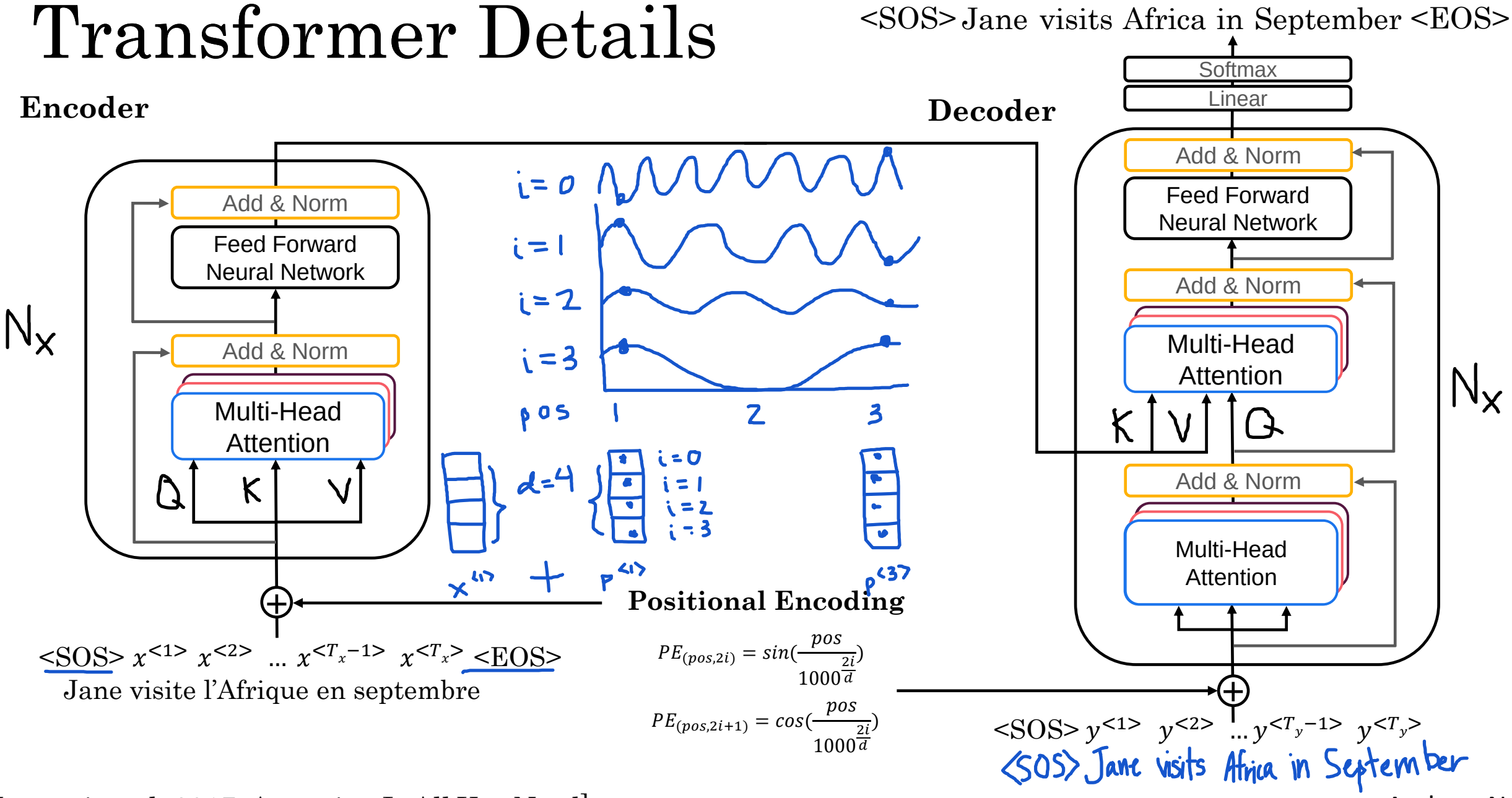
Andrew Ng

deeplearning.ai

Sequence to sequence models

Transformers

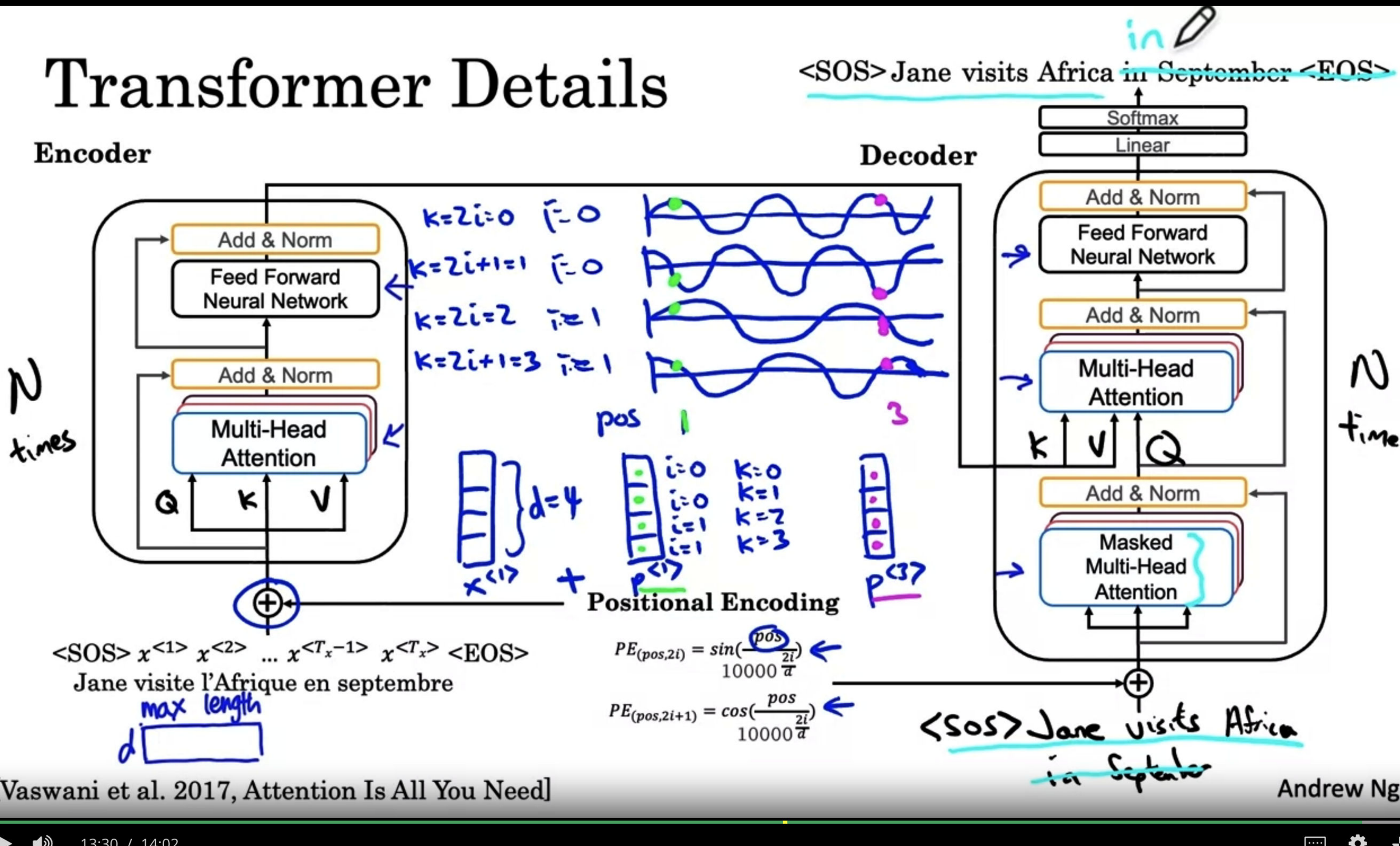
Transformer Details



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

Andrew Ng

Transformer Details



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

Andrew Ng