Copyright Notice

deeplearning.ai

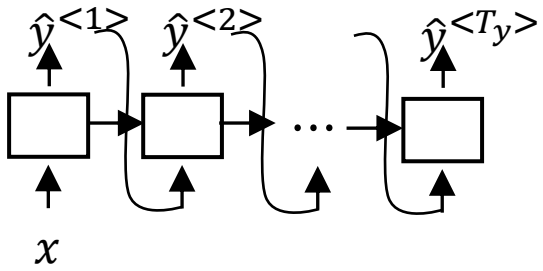
Sequence to sequence models

Transformers Intuition

Transformers Motivation

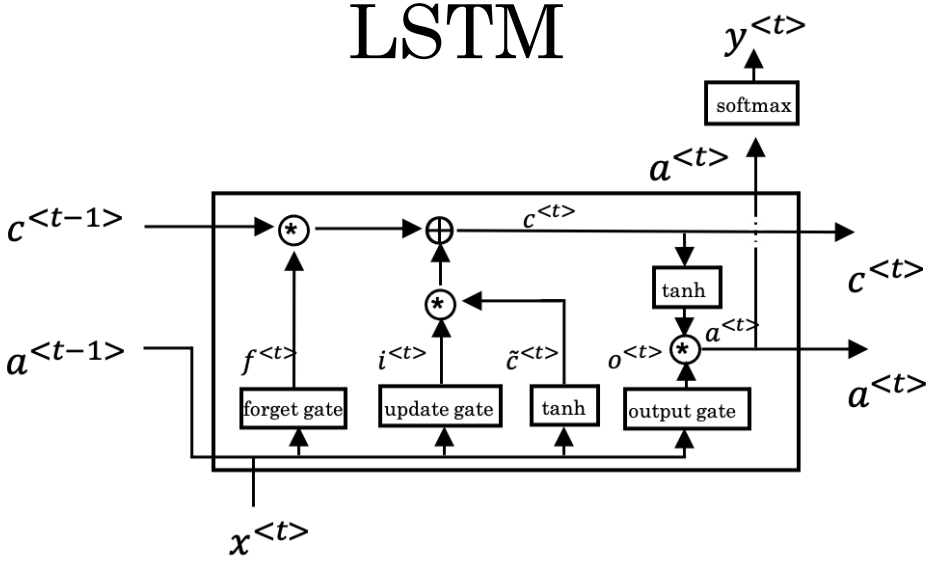
Increased complexity, sequential

RNN



GRU

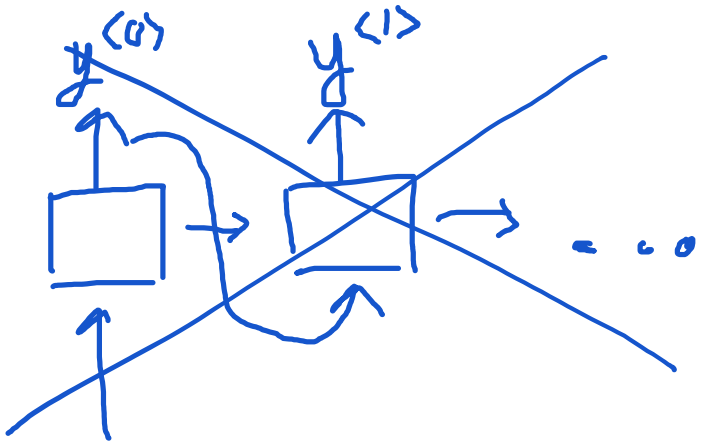
LSTM

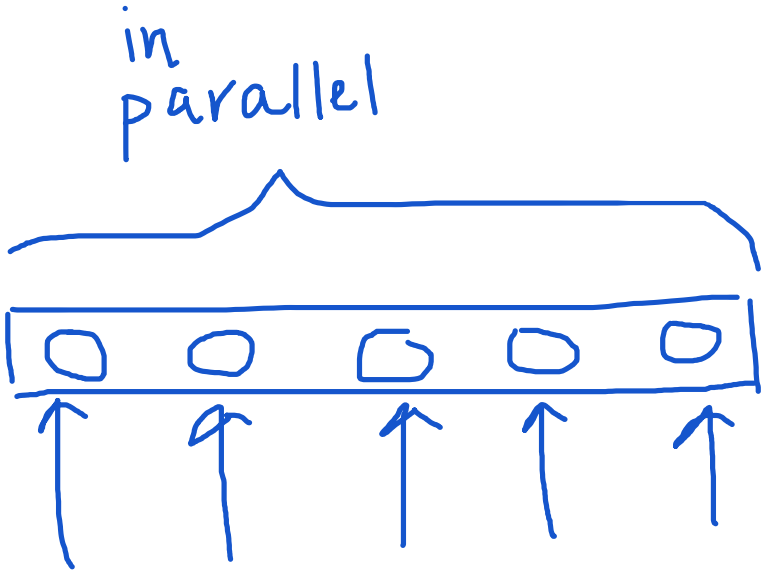


Andrew Ng

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

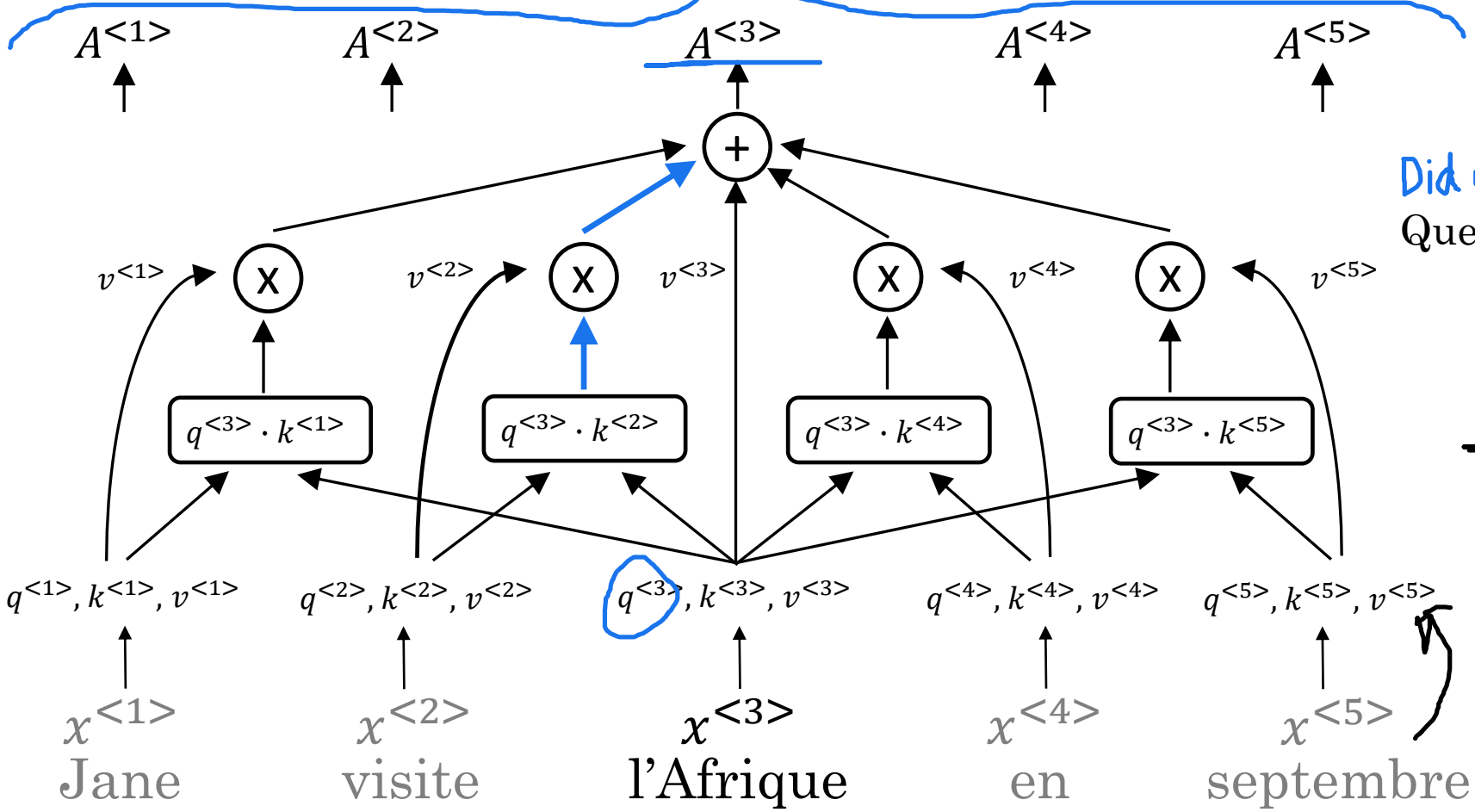
[Vaswani et al. 2017, Attention Is All You Need]

Self-Attention

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

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

Values of q,k,v received from :

q<5> = W_q * x<5>
k<5> = W_k * x<5>
b<5> = W_v * x<5>

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

Andrew Ng

deeplearning.ai

Sequence to sequence models

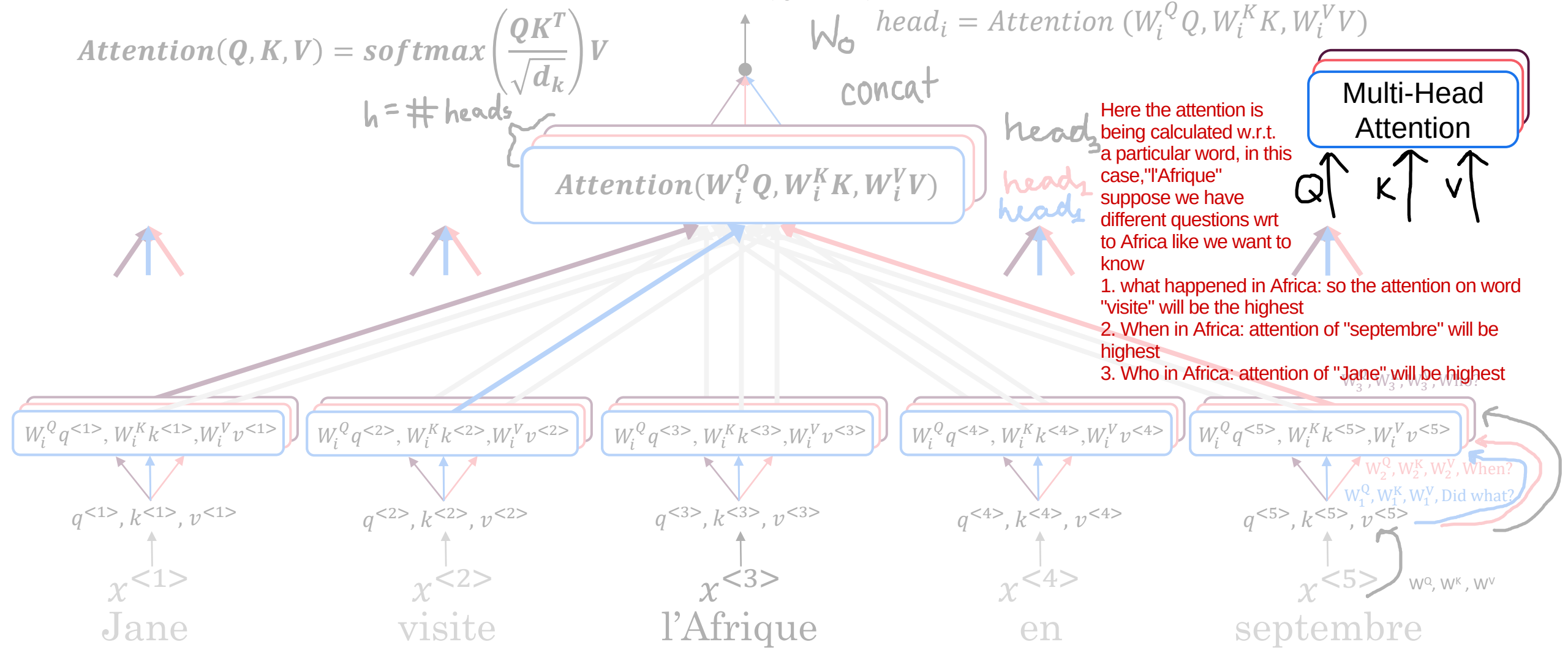
Multi-Head Attention

Multi-Head Attention

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

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

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



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

Similarly, there can be "h" no of heads and each word will have "h" attention heads

Andrew Ng

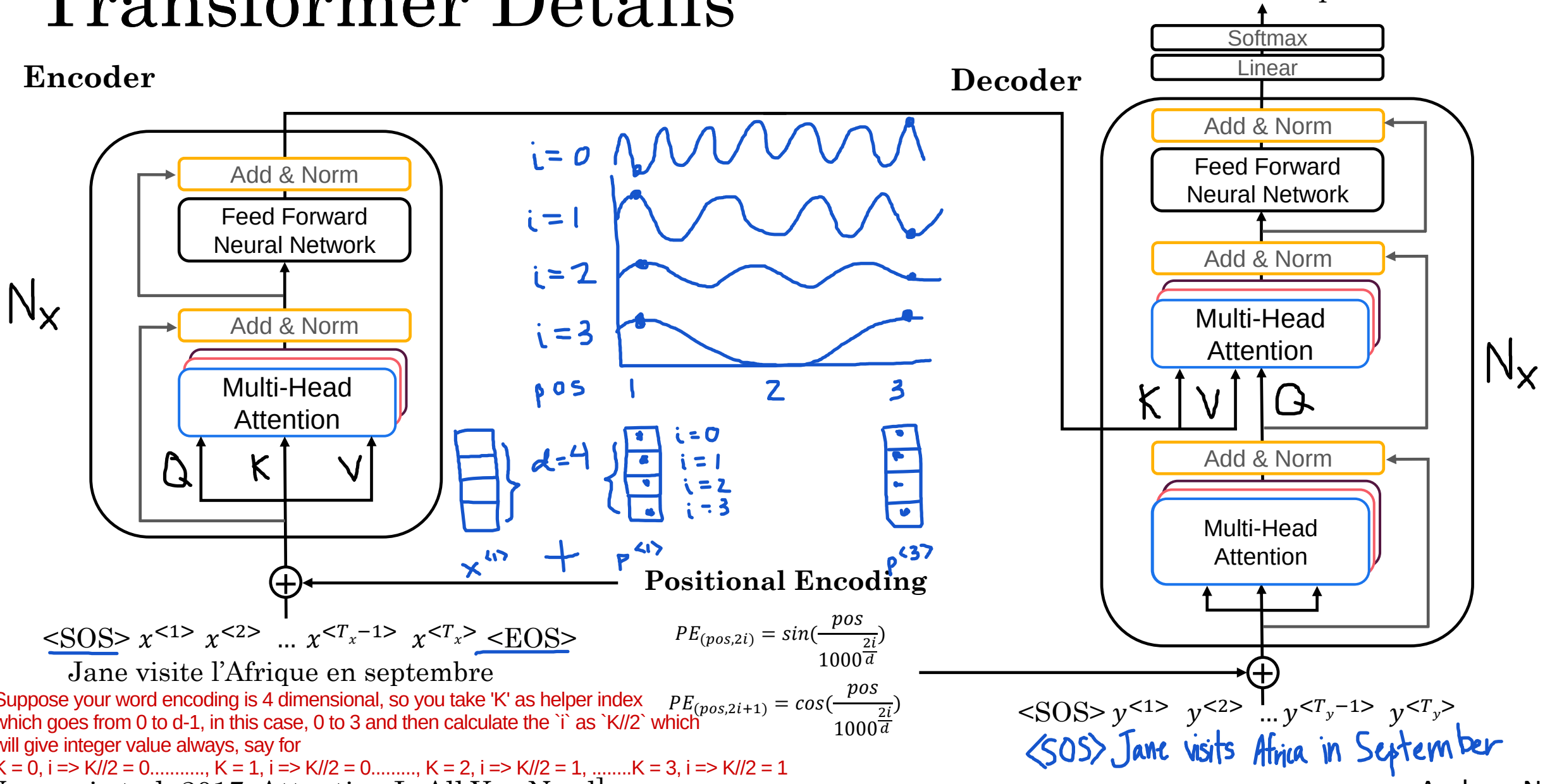
deeplearning.ai

Sequence to sequence models

Transformers

Transformer Details

<SOS> Jane visits Africa in September <EOS>



Suppose your word encoding is 4 dimensional, so you take 'K' as helper index which goes from 0 to d-1, in this case, 0 to 3 and then calculate the `i` as `K//2` which will give integer value always, say for
K = 0, i => K//2 = 0....., K = 1, i => K//2 = 0....., K = 2, i => K//2 = 1, ...K = 3, i => K//2 = 1

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

Andrew Ng