Copyright Notice

deeplearning.ai

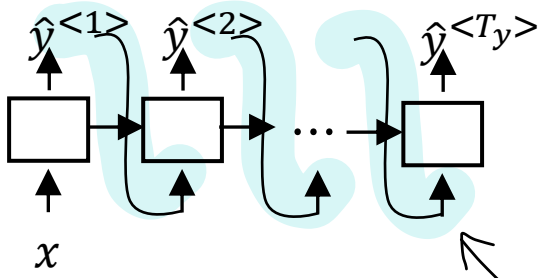
Sequence to sequence models

Transformers Intuition

Transformers Motivation

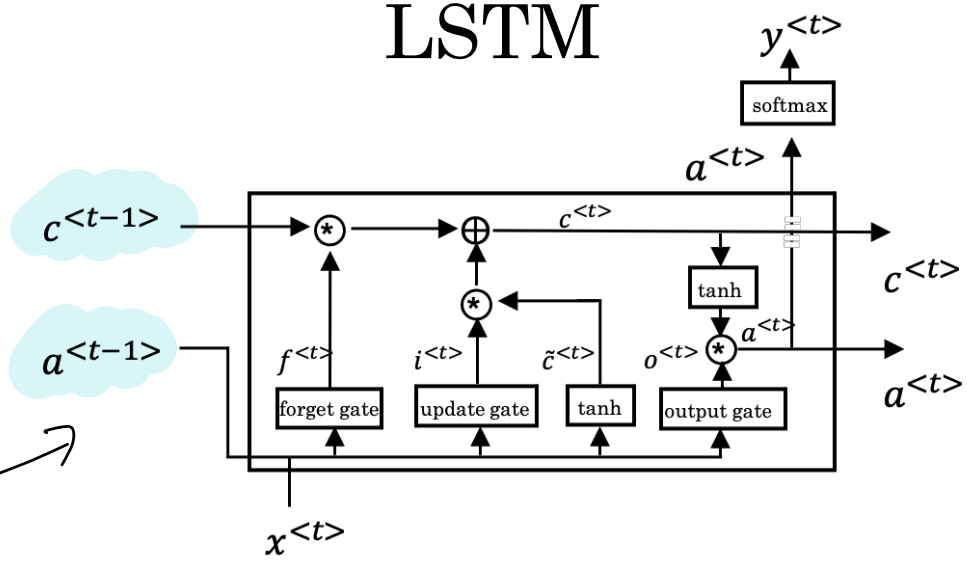
Increased complexity, sequential

RNN



GRU

LSTM

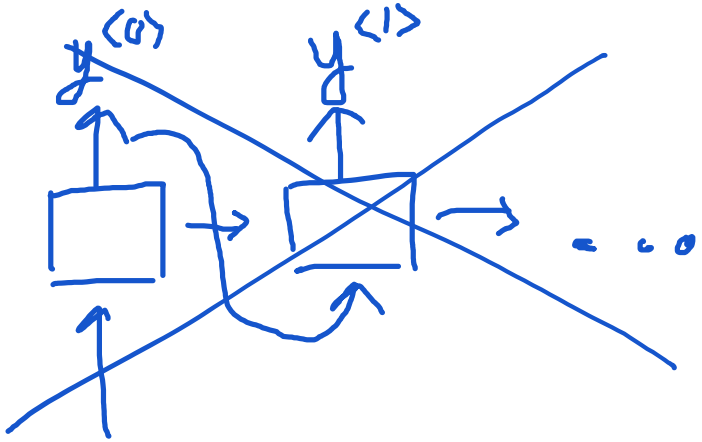


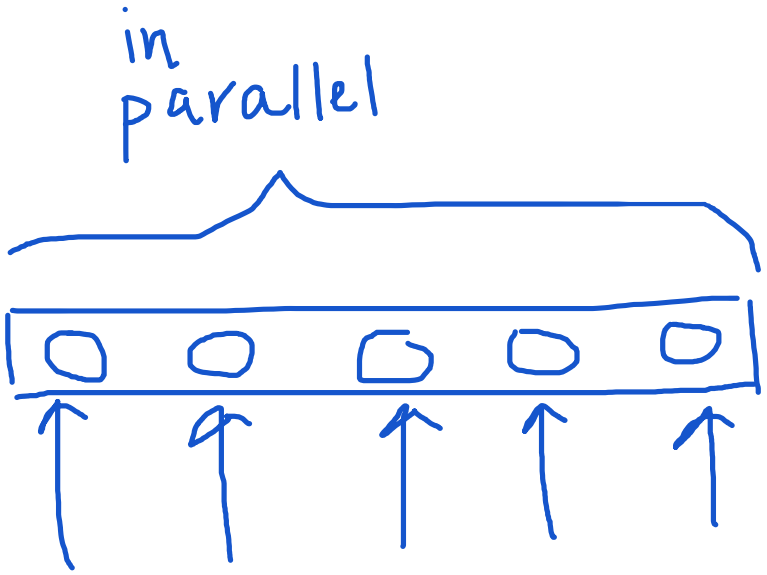
are sequential — need to wait for result from previous (& next) timestep

Andrew Ng

Transformers Intuition

- Attention + CNN
  - Self-Attention
  - Multi-Head Attention for loop over 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 in parellel

↑ is found using context (surrounding words)

RNN Attention

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

Transformers Attention

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

$x^{<1>}$ 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

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

Softmax

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

for scaling

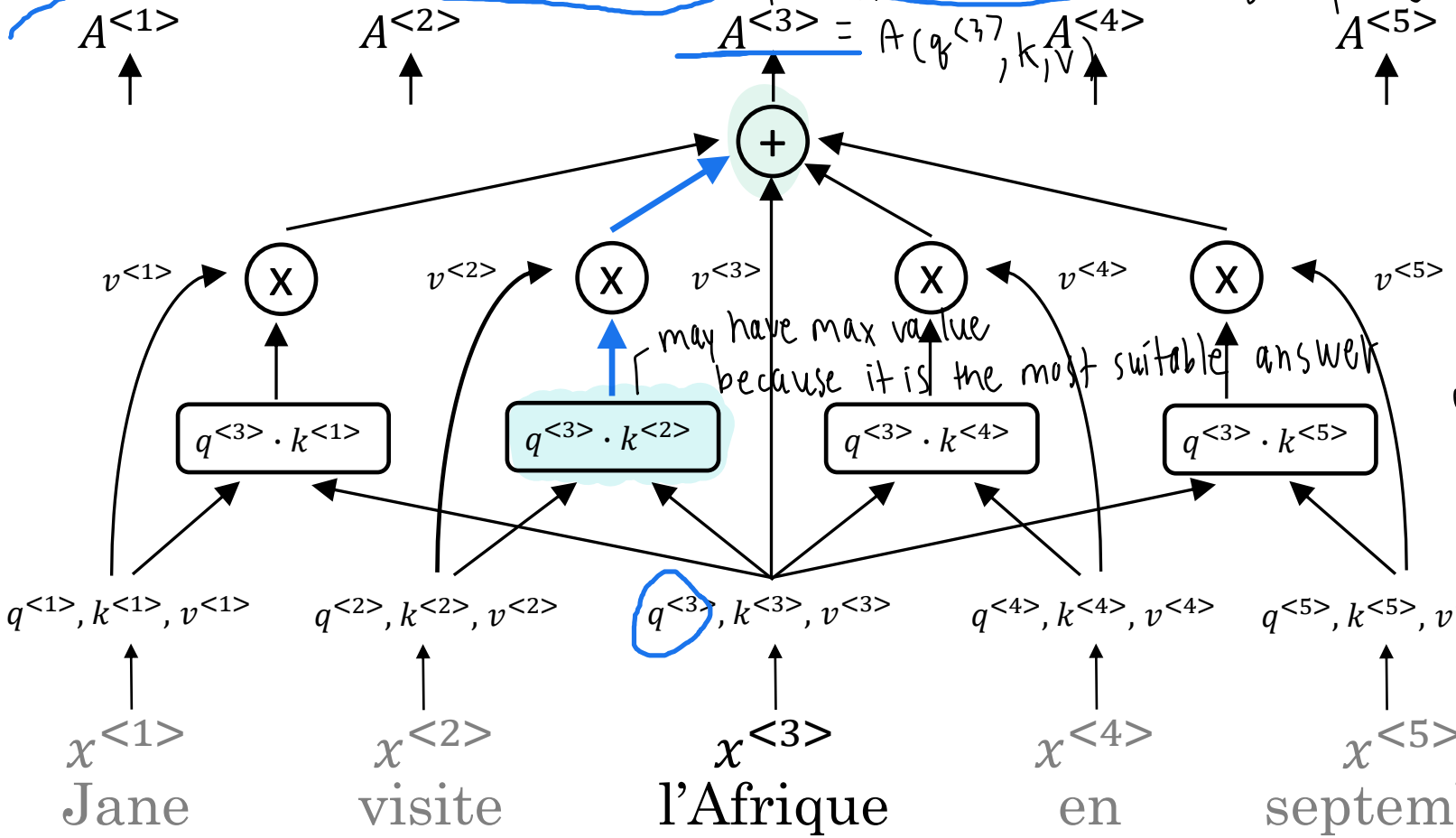
$A^{\langle 3 \rangle} = A(q^{\langle 3 \rangle}, K, V)$ can represent $x^{\langle 3 \rangle}$ as a place

$q^{(i)} = W^Q \cdot x^{(i)}$

$k^{(i)} = W^K \cdot x^{(i)}$

$v^{(i)} = W^V \cdot x^{(i)}$

are learnable



Did what?

Query (Q)	Key (K)	Value (V)
question		
$q^{\langle 1 \rangle}$	$k^{\langle 1 \rangle}$	$v^{\langle 1 \rangle}$
$q^{\langle 2 \rangle}$	$k^{\langle 2 \rangle}$	$v^{\langle 2 \rangle}$
$q^{\langle 3 \rangle}$ What's happening there?	$k^{\langle 3 \rangle}$	$v^{\langle 3 \rangle}$
$q^{\langle 4 \rangle}$	$k^{\langle 4 \rangle}$	$v^{\langle 4 \rangle}$
$q^{\langle 5 \rangle}$	$k^{\langle 5 \rangle}$	$v^{\langle 5 \rangle}$

person / action — Jane / visit

W^Q, W^K, W^V

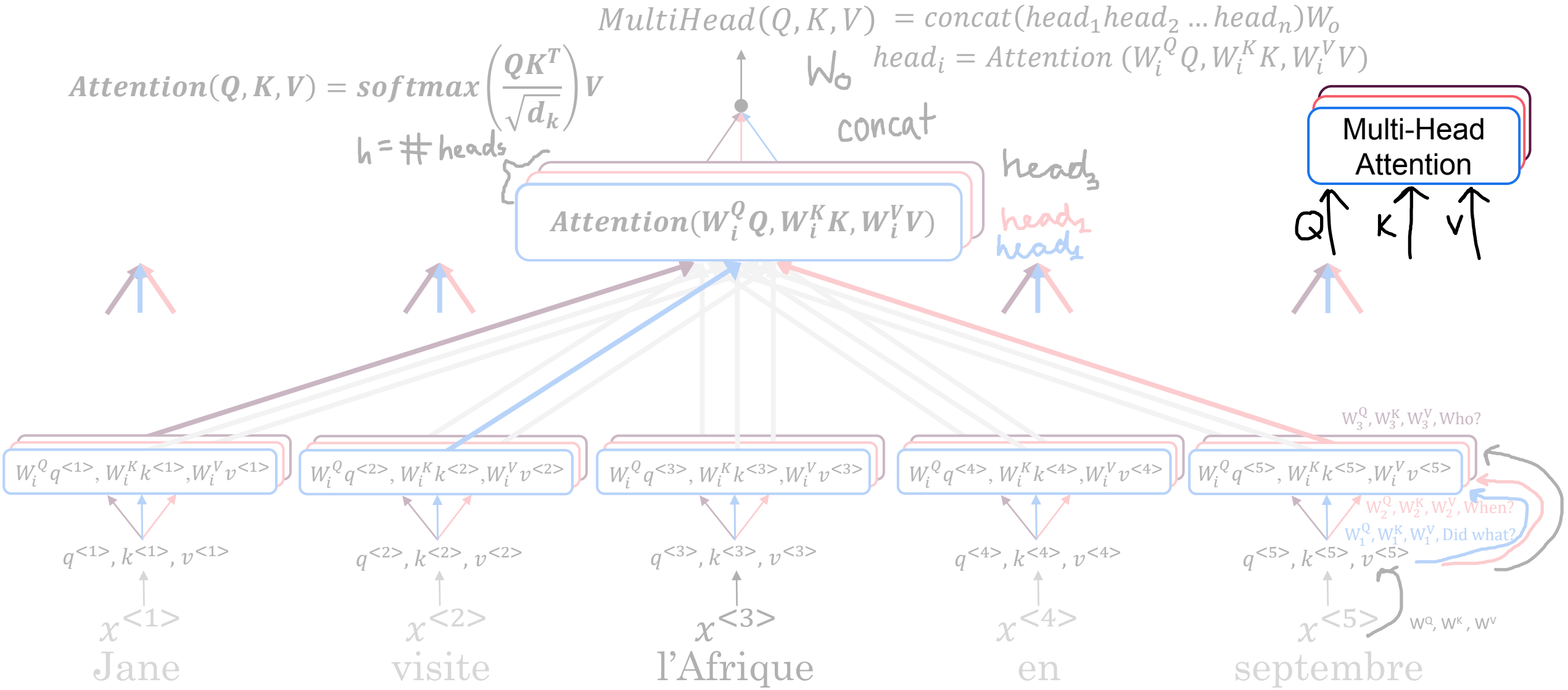
[Vaswani et al. 2017, Attention Is All You Need]

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

masking = hiding part of input

Transformer Details

is predicted from <SOS> (previous word) & K, V from encoder

<SOS> Jane visits Africa in September <EOS>

DONE

output shape = d x max_sequence_length

Encoder

repeated N_x

Add & Norm

Feed Forward Neural Network

Add & Norm

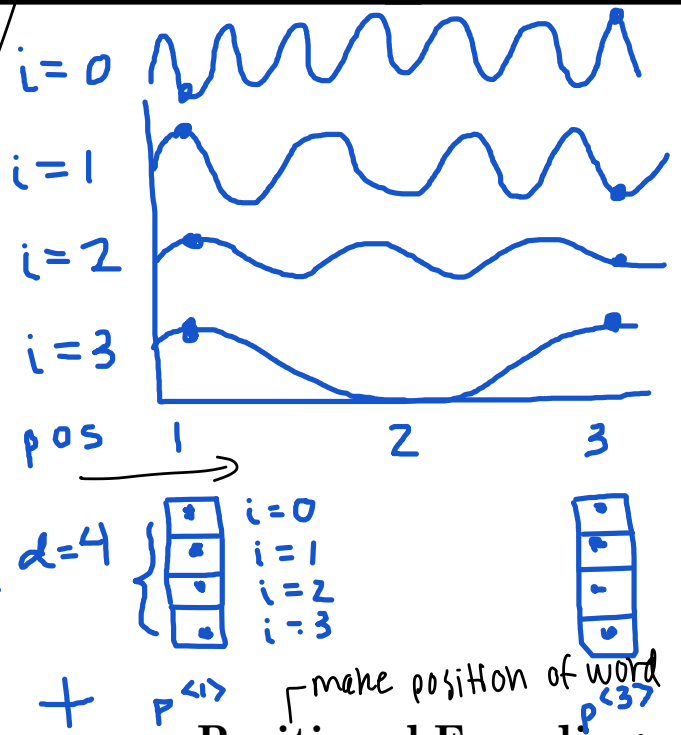
Multi-Head Attention

Q K V

residual connections to pass along positional encoding

<SOS> $x^{<1>}$ $x^{<2>}$ ... $x^{<T_x-1>}$ $x^{<T_x>}$ <EOS>

Jane visite l'Afrique en septembre



make position of word affect output, too.

Positional Encoding

$$PE_{(pos,2i)} = sin(\frac{pos}{1000^{\frac{2i}{d}}})$$

$$PE_{(pos,2i+1)} = cos(\frac{pos}{1000^{\frac{2i}{d}}})$$

i starts at zero

pos = position of word

d = dimension of word embedding

Decoder

Softmax

Linear

Add & Norm

Feed Forward Neural Network

Add & Norm

Multi-Head Attention

K V Q

Add & Norm

Multi-Head Attention

repeated N_x

Q,K,V is computed using $y^{<t>}$

<SOS> $y^{<1>}$ $y^{<2>}$ ... $y^{<T_y-1>}$ $y^{<T_y>}$

<SOS> Jane visits Africa in September

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

Andrew Ng