

# Copyright Notice

These slides are distributed under the Creative Commons License.

[DeepLearning.AI](#) makes these slides available for educational purposes. You may not use or distribute these slides for commercial purposes. You may make copies of these slides and use or distribute them for educational purposes as long as you cite [DeepLearning.AI](#) as the source of the slides.

For the rest of the details of the license, see <https://creativecommons.org/licenses/by-sa/2.0/legalcode>



deeplearning.ai

Sequence to  
sequence models

---

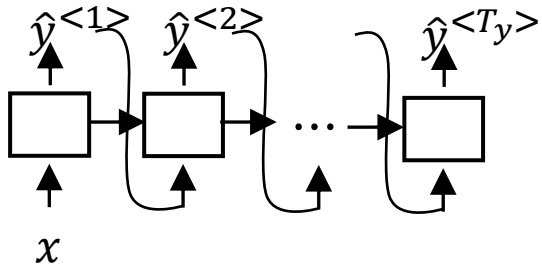
Transformers  
Intuition

# Transformers Motivation

Increased complexity,  
sequential

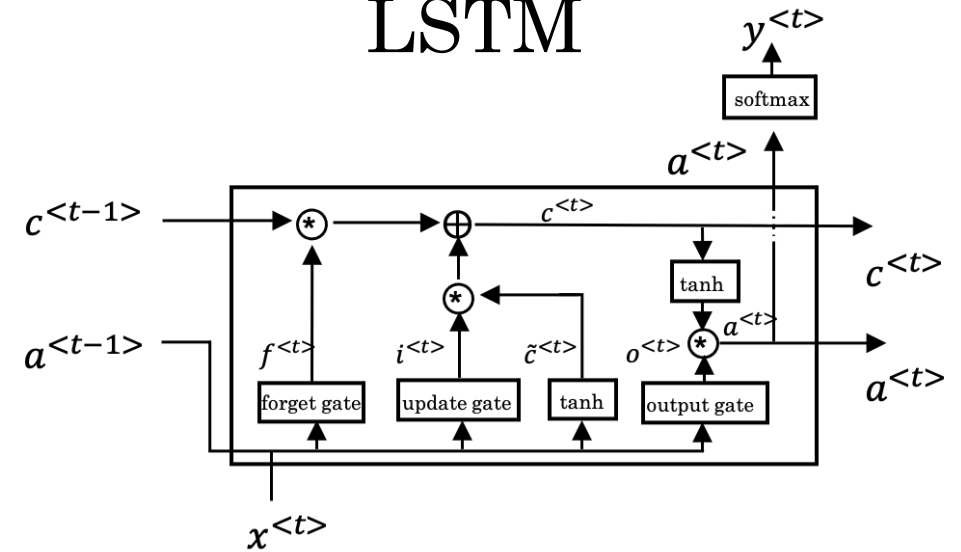


RNN



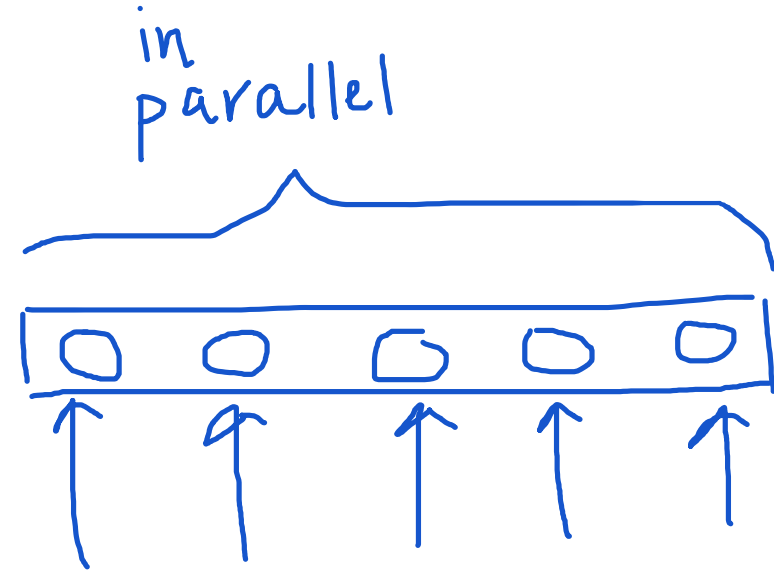
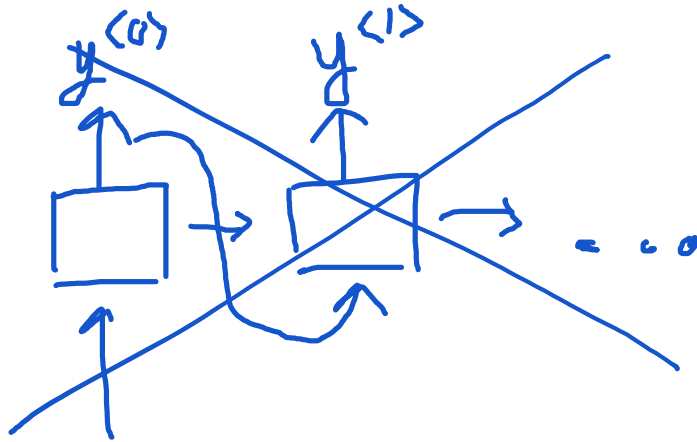
GRU

LSTM



# Transformers Intuition

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





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

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

$x^{<1>}$   
Jane

$x^{<2>}$   
visite

$x^{<3>}$   
l'Afrique

$x^{<4>}$   
en

$x^{<5>}$   
septembre

## Transformers Attention

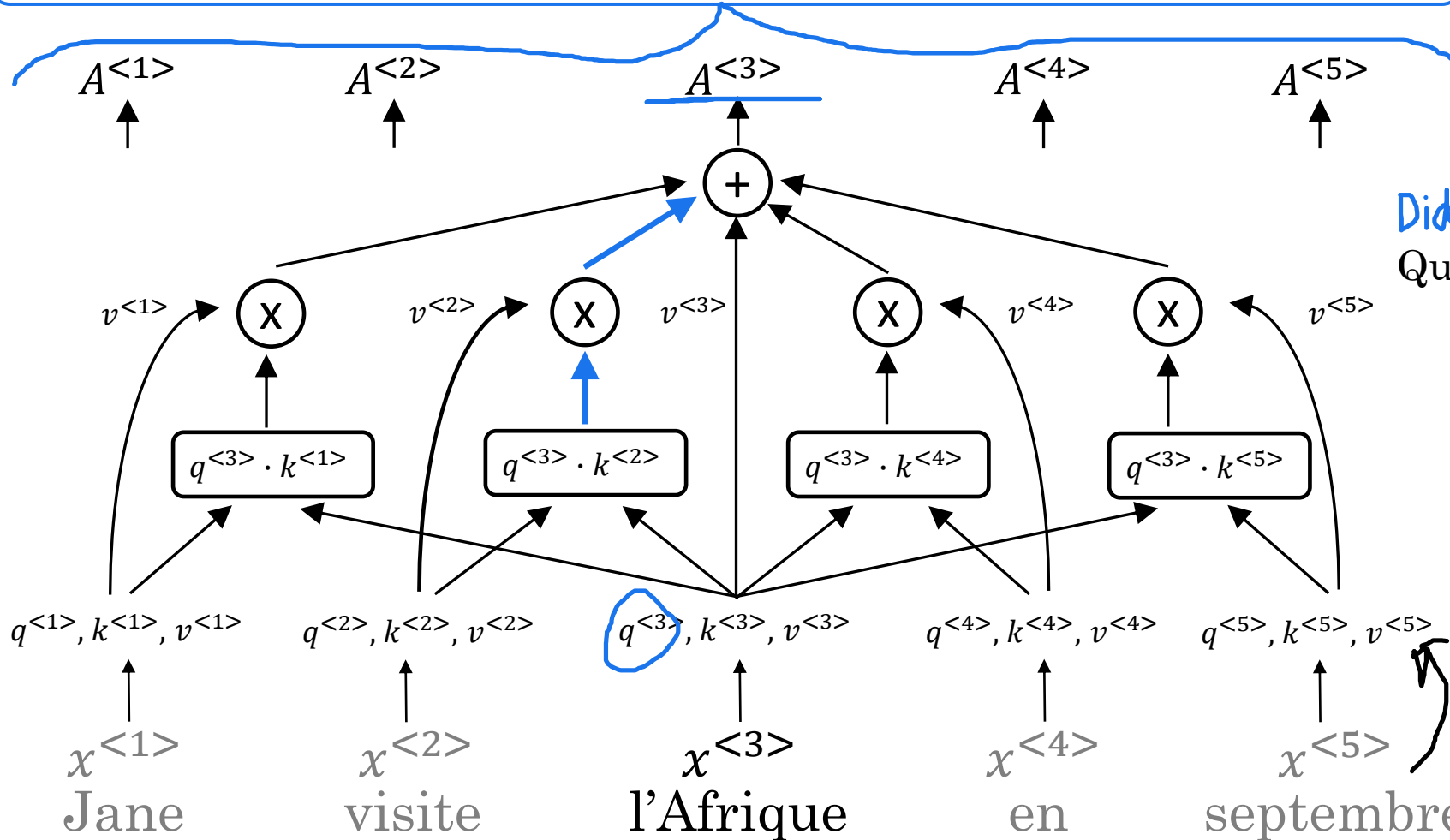
$$A(q, K, V) = \sum_i \frac{\exp(q \cdot k^{<i>})}{\sum_j \exp(q \cdot k^{<j>})} v^{<i>}$$

# 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

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



Did what?

Query (Q)

Key (K)

Value (V)

$q^{<1>}$

$k^{<1>}$

$v^{<1>}$

$q^{<2>}$

$k^{<2>}$

$v^{<2>}$

$q^{<3>}$

$k^{<3>}$

$v^{<3>}$

$q^{<4>}$

$k^{<4>}$

$v^{<4>}$

$q^{<5>}$

$k^{<5>}$

$v^{<5>}$

$q^{<1>}$

$k^{<1>}$

$v^{<1>}$

$q^{<2>}$

$k^{<2>}$

$v^{<2>}$

$q^{<3>}$

$k^{<3>}$

$v^{<3>}$

$q^{<4>}$

$k^{<4>}$

$v^{<4>}$

$q^{<5>}$

$k^{<5>}$

$v^{<5>}$

$q^{<1>}$

$k^{<1>}$

$v^{<1>}$

$q^{<2>}$

$k^{<2>}$

$v^{<2>}$

$q^{<3>}$

$k^{<3>}$

$v^{<3>}$

$q^{<4>}$

$k^{<4>}$

$v^{<4>}$

$q^{<5>}$

$k^{<5>}$

$v^{<5>}$

$q^{<1>}$

$k^{<1>}$

$v^{<1>}$

$q^{<2>}$

$k^{<2>}$

$v^{<2>}$

$q^{<3>}$

$k^{<3>}$

$v^{<3>}$

$q^{<4>}$

$k^{<4>}$

$v^{<4>}$

$q^{<5>}$

$k^{<5>}$

$v^{<5>}$

$q^{<1>}$

$k^{<1>}$

$v^{<1>}$

$q^{<2>}$

$k^{<2>}$

$v^{<2>}$

$q^{<3>}$

$k^{<3>}$

$v^{<3>}$

$q^{<4>}$

$k^{<4>}$

$v^{<4>}$

$q^{<5>}$

$k^{<5>}$

$v^{<5>}$

$q^{<1>}$

$k^{<1>}$

$v^{<1>}$

$q^{<2>}$

$k^{<2>}$

$v^{<2>}$

$q^{<3>}$

$k^{<3>}$

$v^{<3>}$

$q^{<4>}$

$k^{<4>}$

$v^{<4>}$

$q^{<5>}$

$k^{<5>}$

$v^{<5>}$

$q^{<1>}$

$k^{<1>}$

$v^{<1>}$

$q^{<2>}$

$k^{<2>}$

$v^{<2>}$

$q^{<3>}$

$k^{<3>}$

$v^{<3>}$

$q^{<4>}$

$k^{<4>}$

$v^{<4>}$

$q^{<5>}$

$k^{<5>}$

$v^{<5>}$

$q^{<1>}$

$k^{<1>}$

$v^{<1>}$

$q^{<2>}$

$k^{<2>}$

$v^{<2>}$

$q^{<3>}$

$k^{<3>}$

$v^{<3>}$

$q^{<4>}$

$k^{<4>}$

$v^{<4>}$

$q^{<5>}$

$k^{<5>}$

$v^{<5>}$

$q^{<1>}$

$k^{<1>}$

$v^{<1>}$

$q^{<2>}$

$k^{<2>}$

$v^{<2>}$

$q^{<3>}$

$k^{<3>}$

$v^{<3>}$

$q^{<4>}$

$k^{<4>}$

$v^{<4>}$

$q^{<5>}$

$k^{<5>}$

$v^{<5>}$

$q^{<1>}$

$k^{<1>}$

$v^{<1>}$

$q^{<2>}$

$k^{<2>}$

$v^{<2>}$

$q^{<3>}$

$k^{<3>}$

$v^{<3>}$

$q^{<4>}$

$k^{<4>}$

$v^{<4>}$

$q^{<5>}$

$k^{<5>}$

$v^{<5>}$

$q^{<1>}$

$k^{<1>}$

$v^{<1>}$

$q^{<2>}$

$k^{<2>}$

$v^{<2>}$

$q^{<3>}$

$k^{<3>}$

$v^{<3>}$

$q^{<4>}$

$k^{<4>}$

$v^{<4>}$

$q^{<5>}$

$k^{<5>}$

$v^{<5>}$

$q^{<1>}$

$k^{<1>}$

$v^{<1>}$

$q^{<2>}$

$k^{<2>}$

$v^{<2>}$

$q^{<3>}$

$k^{<3>}$

$v^{<3>}$

$q^{<4>}$

$k^{<4>}$

$v^{<4>}$

$q^{<5>}$

$k^{<5>}$

$v^{<5>}$

$q^{<1>}$

$k^{<1>}$

$v^{<1>}$

$q^{<2>}$

$k^{<2>}$

$v^{<2>}$

$q^{<3>}$

$k^{<3>}$

$v^{<3>}$

$q^{<4>}$

$k^{<4>}$

$v^{<4>}$

$q^{<5>}$

$k^{<5>}$

$v^{<5>}$

$q^{<1>}$

$k^{<1>}$

$v^{<1>}$

$q^{<2>}$

$k^{<2>}$

$v^{<2>}$

$q^{<3>}$

$k^{<3>}$

$v^{<3>}$

$q^{<4>}$

$k^{<4>}$

$v^{<4>}$

$q^{<5>}$

$k^{<5>}$

$v^{<5>}$

$q^{<1>}$

$k^{<1>}$

$v^{<1>}$

$q^{<2>}$

$k^{<2>}$

$v^{<2>}$

$q^{<3>}$

$k^{<3>}$

$v^{<3>}$

$q^{<4>}$

$k^{<4>}$

$v^{<4>}$

$q^{<5>}$

$k^{<5>}$

$v^{<5>}$

$q^{<1>}$

$k^{<1>}$



deeplearning.ai

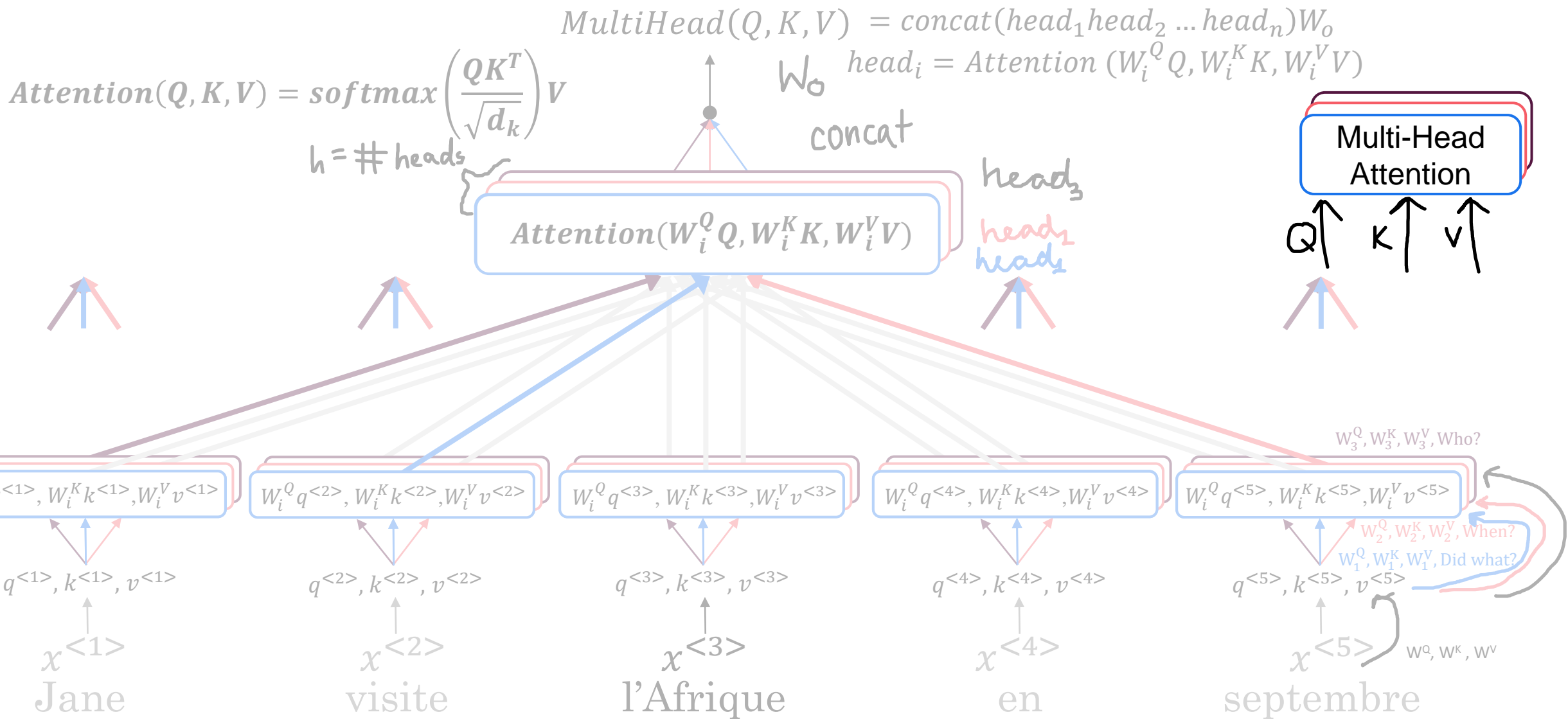
# Sequence to sequence models

---

## Multi-Head Attention



# Multi-Head Attention





deeplearning.ai

# Sequence to sequence models

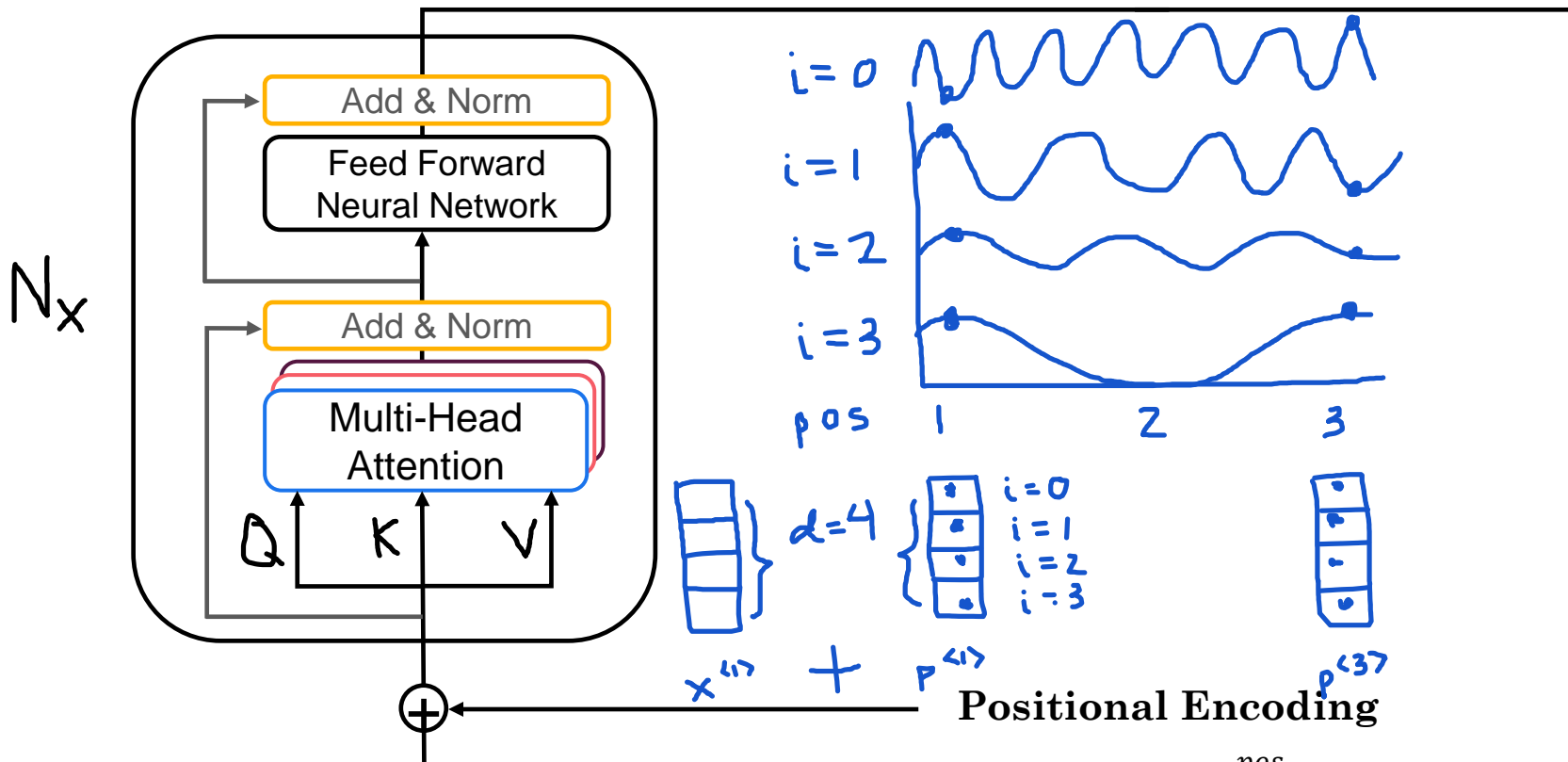
---

# Transformers

# Transformer Details

<SOS> Jane visits Africa in September <EOS>

Encoder

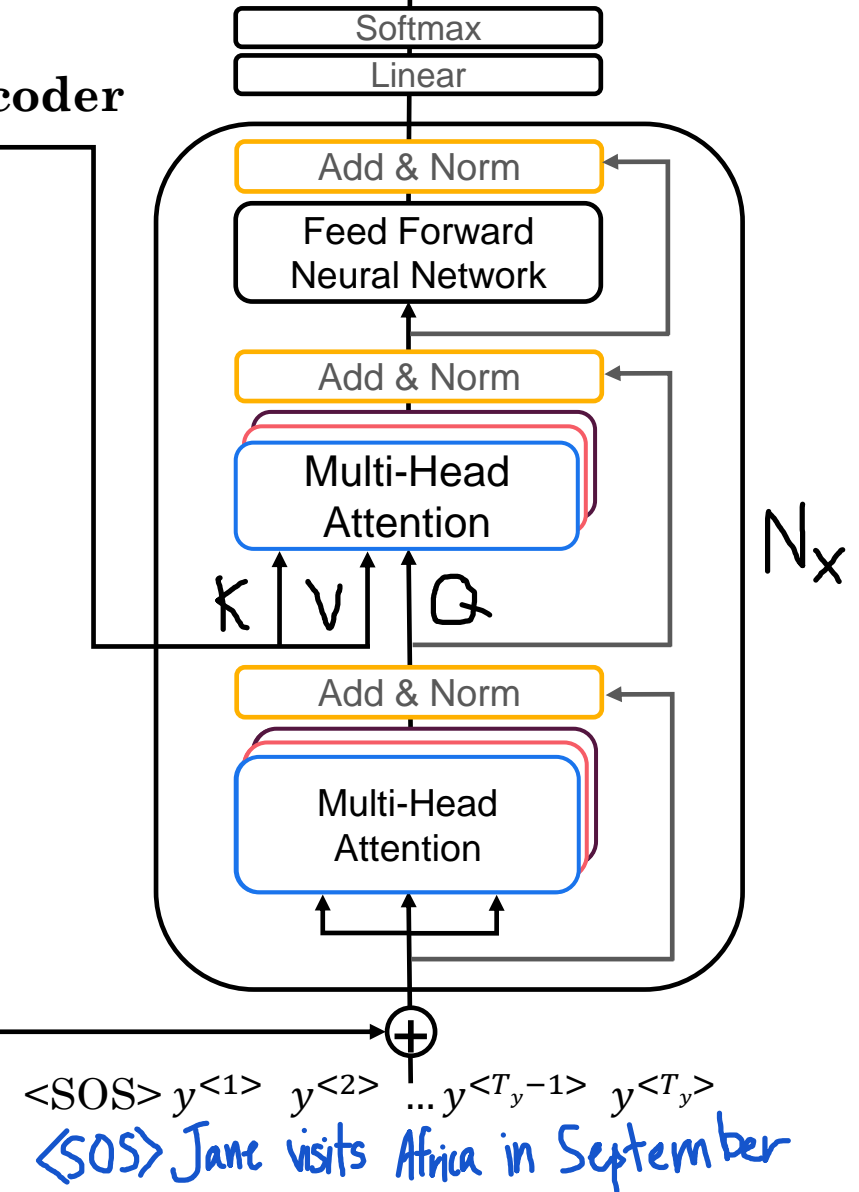


$\langle \text{SOS} \rangle x^{(1)} x^{(2)} \dots x^{(T_x-1)} x^{(T_x)} \langle \text{EOS} \rangle$   
Jane visite l'Afrique en septembre

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

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

Decoder



$\langle \text{SOS} \rangle y^{(1)} y^{(2)} \dots y^{(T_y-1)} y^{(T_y)}$   
<SOS> Jane visits Africa in September