Transformer Network

Learning Objectives:

- Create positional encodings to capture sequential relationships in data
- Calculate scaled dot-product self-attention with word embeddings
- Implement masked multi-head attention
- Build and train a Transformer model
- Fine-tune a pre-trained transformer model for Named Entity Recognition
- Fine-tune a pre-trained transformer model for Question Answering
- Implement a QA model in TensorFlow and PyTorch
- Fine-tune a pre-trained transformer model to a custom dataset
- Perform extractive Question Answering

Transformer Network

1 Transformers Intuition
2 Self-Attention
3 Multi-Head Attention
4 Transformers

FPT UNIVERSITY

FPT UNIVERSITY

Transformer Network

Transformers Intuition

Transformers Motivation

- The development of Transformers was motivated by the limitations of traditional sequence-to-sequence models in capturing long-term dependencies in input sequences due to the vanishing gradient problem.

- Transformers use self-attention mechanisms to attend to different parts of the input sequence at different time steps, enabling them to capture long-term dependencies more effectively and leading to improved performance in natural language processing tasks.

Transformers Motivation

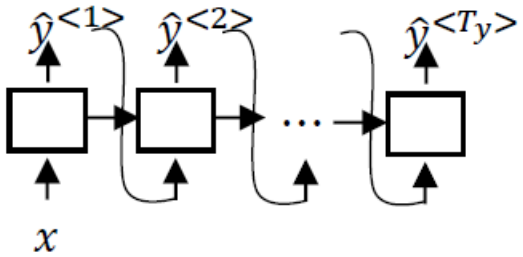
- The transformer architecture is very popular in the field of natural language processing (NLP).
- It is a complex architecture, but it allows for parallel processing of entire sequences, rather than ingesting input one word or token at a time. This is in contrast to RNNs, GRUs, and LSTMs, which are all sequential models that process input one unit at a time. As the complexity of sequence tasks increases, so does the complexity of the model.
- The major innovation of the transformer architecture is the combination of attention-based representations and a CNN-style of processing.

Transformers Motivation

FPT Education
FPT UNIVERSITY

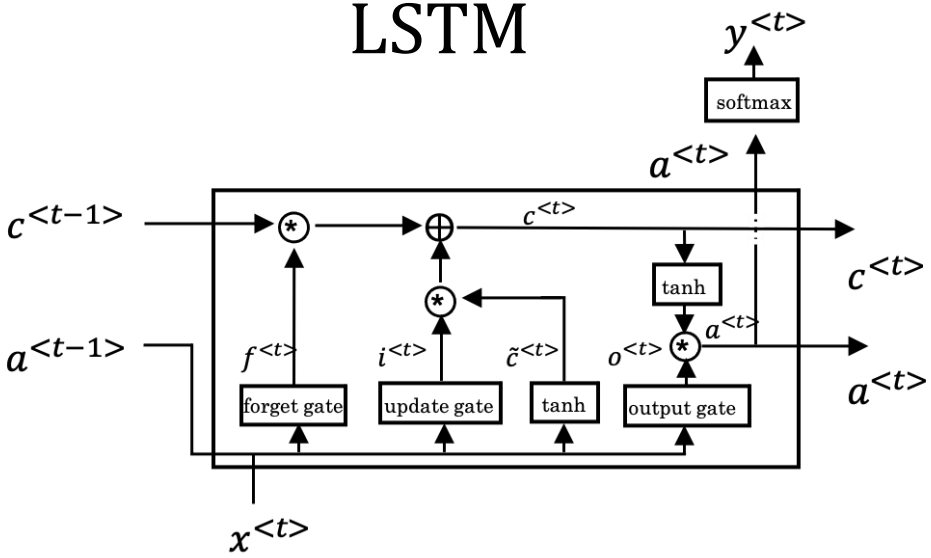
Increased complexity, sequential

RNN



GRU

LSTM



Transformers Intuition

- The attention network is a way of computing rich and useful representations of words in parallel using self-attention and multi-headed attention.
- The transformer architecture overcomes the limitations of traditional sequence-to-sequence models in capturing long-term dependencies.
- The next section will cover the self-attention and multi-headed attention processes in detail and put all the pieces together to give a complete understanding of how the transformer works.

Transformers Intuition

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

FPT Education

FPT UNIVERSITY

Sequence to sequence models

Self-Attention

Self-Attention Intuition

- The attention values are then computed using the dot product of the query and key, and scaled by the square root of the dimension of the key.
- The resulting attention weights are used to compute a weighted sum of the values, which gives the attention-based representation for each word.
- The self-attention mechanism allows for the parallel computation of these representations, making it more efficient than traditional sequence-to-sequence models.

Self-Attention Intuition

- $A(q, K, V)$= attention-based vector representation of a word

RNN Attention

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

Transformers Attention

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

Self-Attention

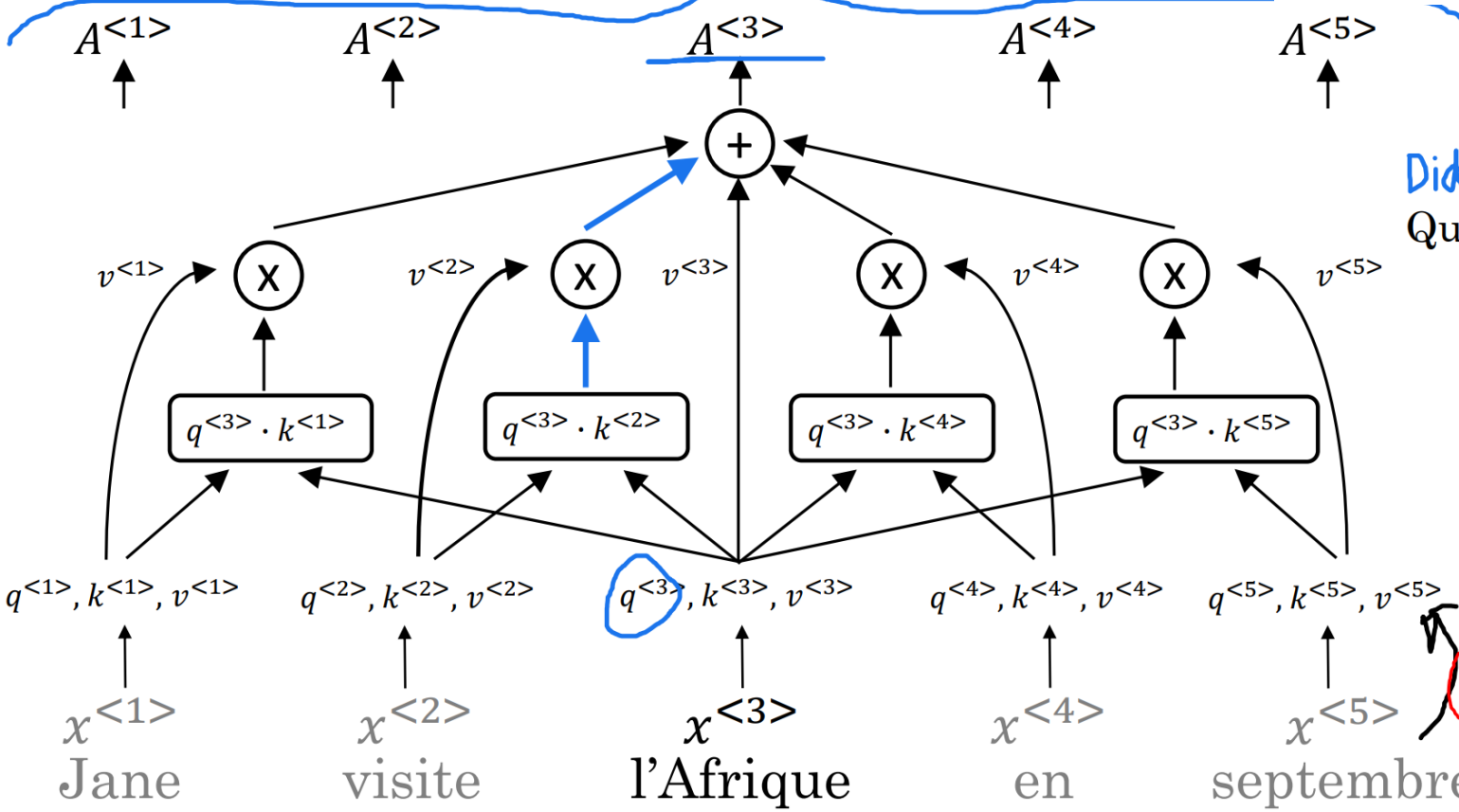
FPT Education
FPT UNIVERSITY

- The value of each word in a sentence can be plugged into the attention-based representation for the sentence, allowing for a richer and more nuanced representation of words than fixed word embeddings.
- This self-attention mechanism can be used to create attention-based representations for all the words in a sentence, which can be combined in a compressed or vectorized representation using the scaled dot-product attention mechanism.
- The multi-headed attention mechanism adds a loop over the self-attention mechanism, creating multiple attention-based representations for each word and allowing for even more complex and nuanced representations of input sequences.

Self-Attention

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

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



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
Jan
visit
What's happening there?

W^Q, W^K, W^V

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

Andrew Ng

FPT Education

FPT UNIVERSITY

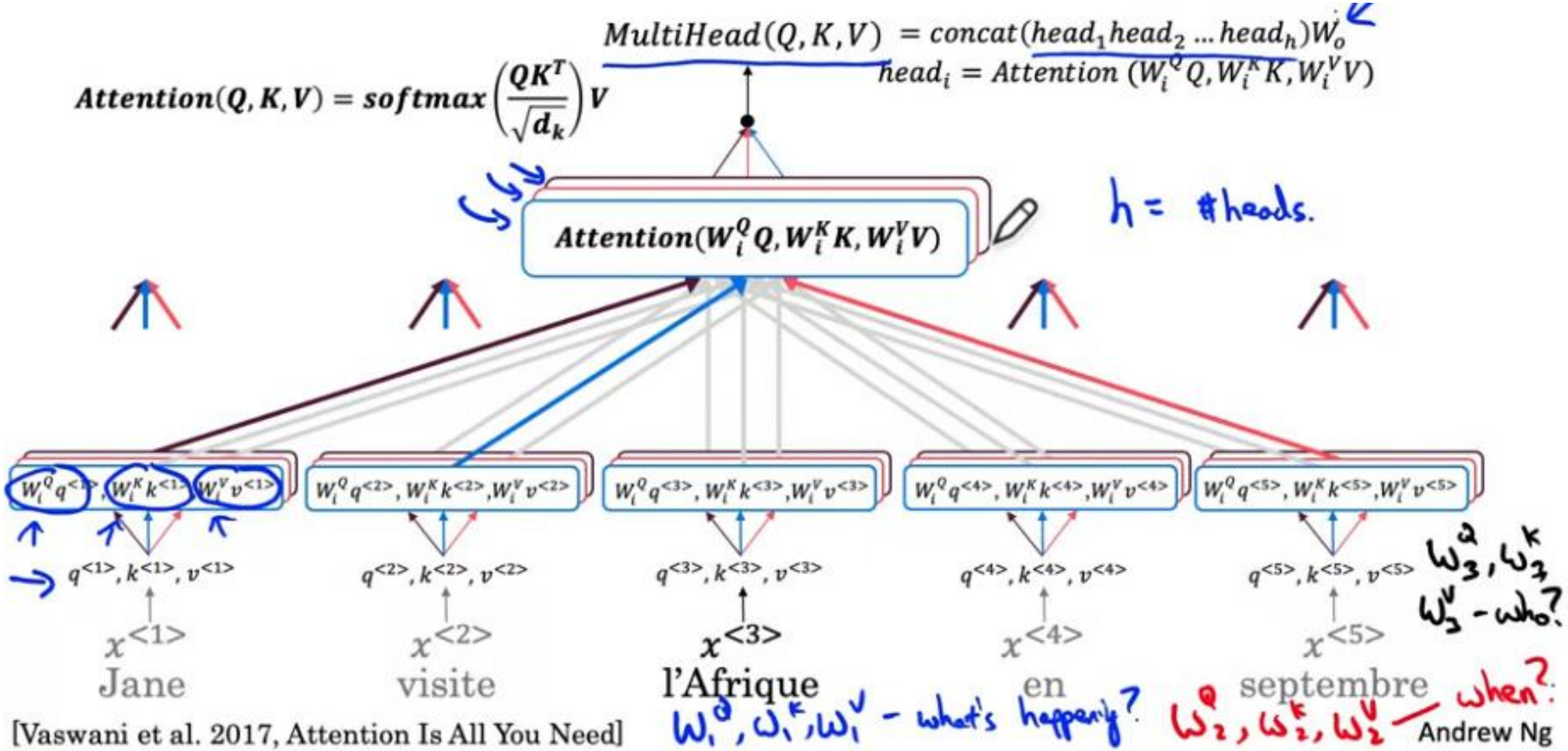
Sequence to sequence models

Multi-Head Attention

Multi-Head Attention

- In multi-head attention mechanism, the input is multiplied with weight matrices to obtain query, key, and value vectors.
- Self attention is computed multiple times, each time using a different set of weight matrices referred to as "heads". Each head represents a different feature and asks a different question about the input sequence.
- The output of each head is concatenated and multiplied by a final weight matrix to produce the final output. This computation can be done in parallel, making it computationally efficient and suitable for use in large-scale deep learning models such as the transformer network.

Multi-Head Attention



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

Andrew Ng

FPT Education

FPT UNIVERSITY

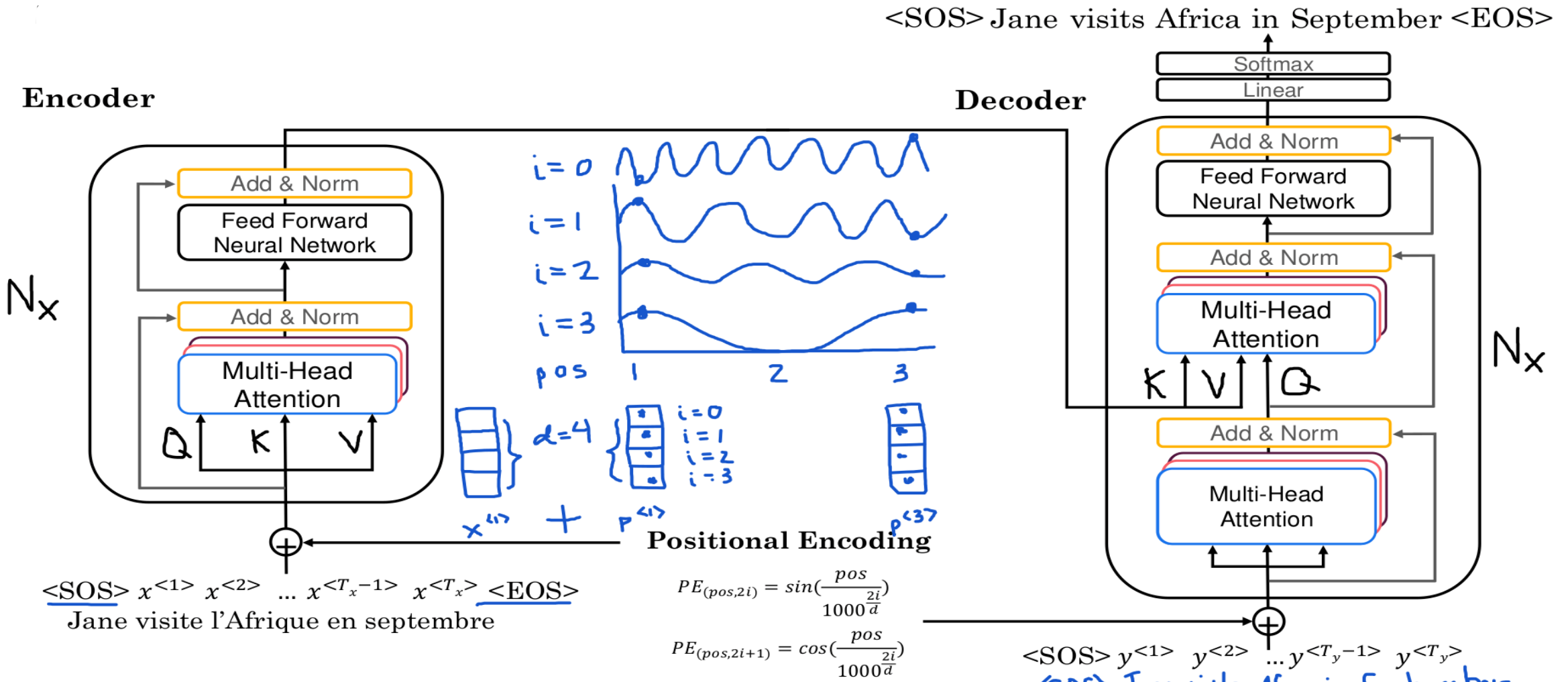
Sequence to sequence models

Transformers

Transformer Details

- The Transformer network is an architecture that uses self-attention and multi-headed attention mechanisms to perform sequence-to-sequence tasks, such as machine translation.
- The network has an encoder and decoder block, each with multiple layers.
- The encoder produces a contextualized embedding for each word using multi-headed attention and feed-forward neural networks.
- The decoder generates the translated output using a similar multi-headed attention mechanism, generating the output one word at a time.
- Additional features such as positional encoding, residual connections, and masking during training are used to improve performance. Variations of the model, such as BERT and BERT distill, have also been proposed.

Transformer Details



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

Andrew Ng

Summarization

- Transformers, introduced in "Attention is All You Need," use self-attention for sequential data.
- Self-attention allows dynamic weighing of words based on relationships. Multi-head attention enhances this by focusing on different parts simultaneously.
- Transformers consist of encoder and decoder layers with self-attention and feed-forward networks. They excel in NLP, surpassing RNNs and CNNs due to parallelization and long-range dependency handling.