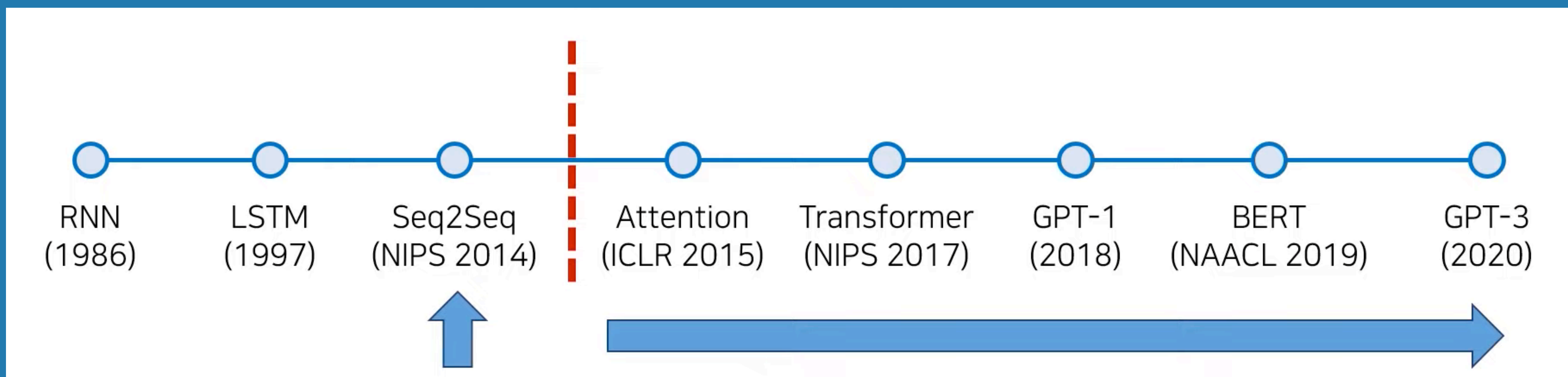
Attention is All You Need (2017)

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin

Hannah Do | Feb 20th, 2022

Development of Sequence Transduction Models

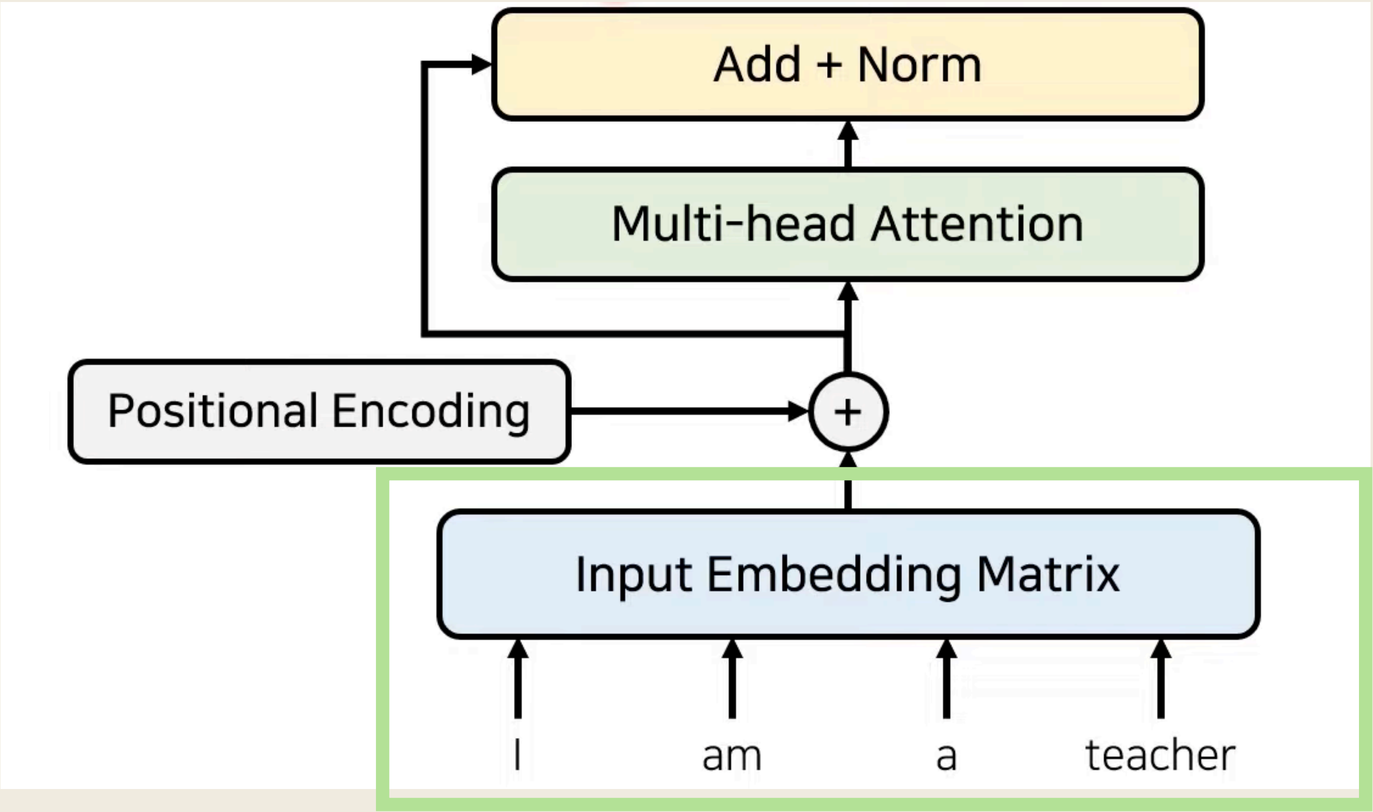


Fixed size of context vectors

Extracting information from the entire input sequence

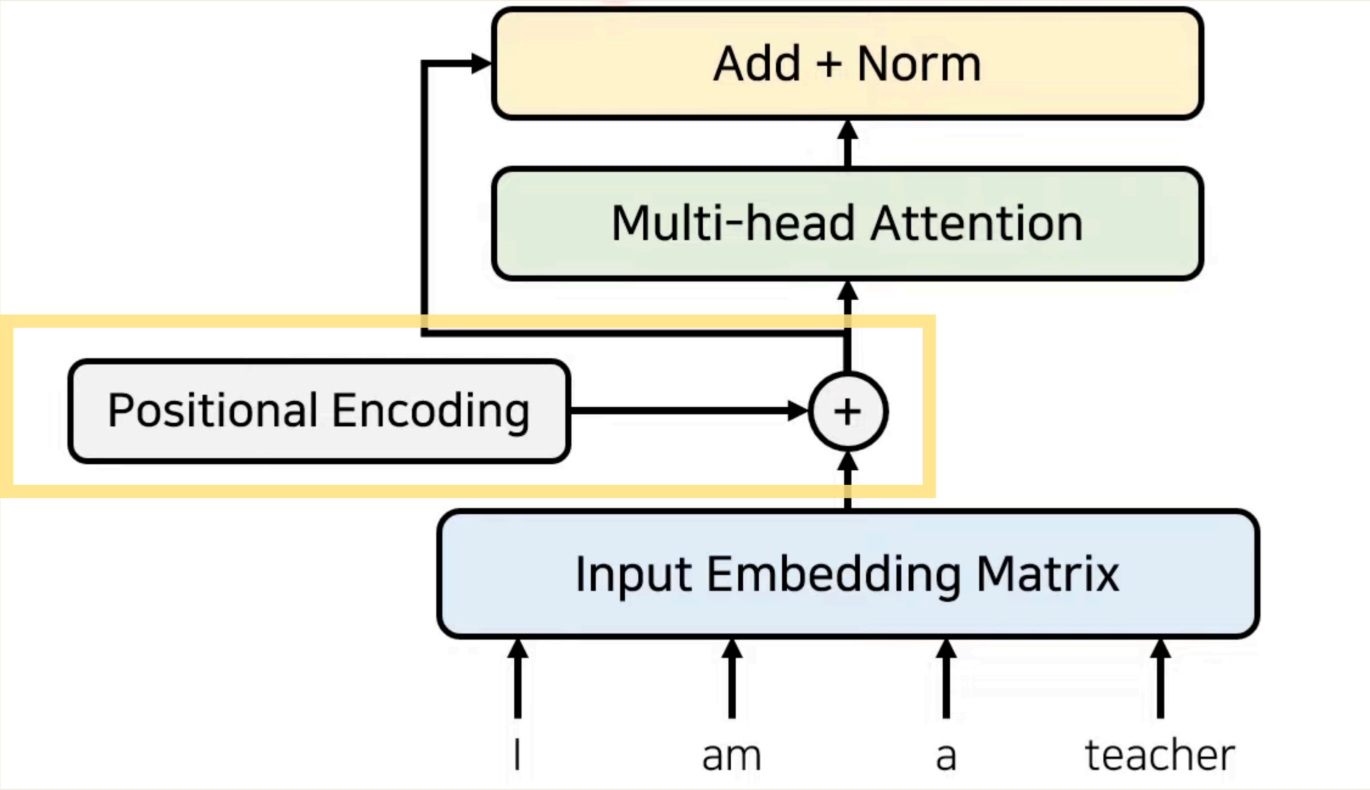
Transformer, the first sequence transduction model based entirely on **attention**, replaced the recurrent layers most commonly used in encoder-decoder architectures with **multi-headed self-attention**.

Part 1 : **Encoder**



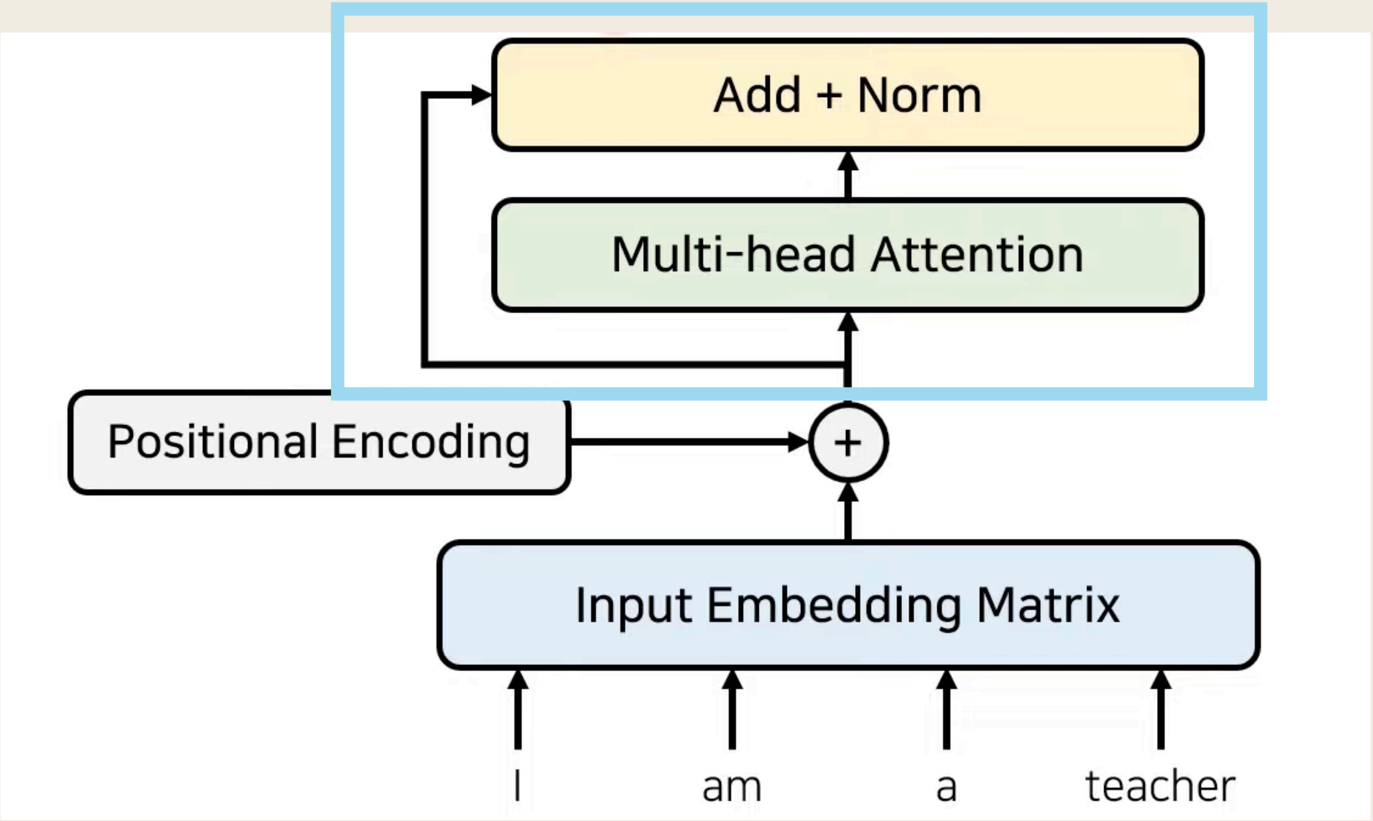
Sentence is converted into an **input embedding matrix**. The dimension of such matrix is # of tokens in a sentence x # of embedding columns.

Part 1 : **Encoder**



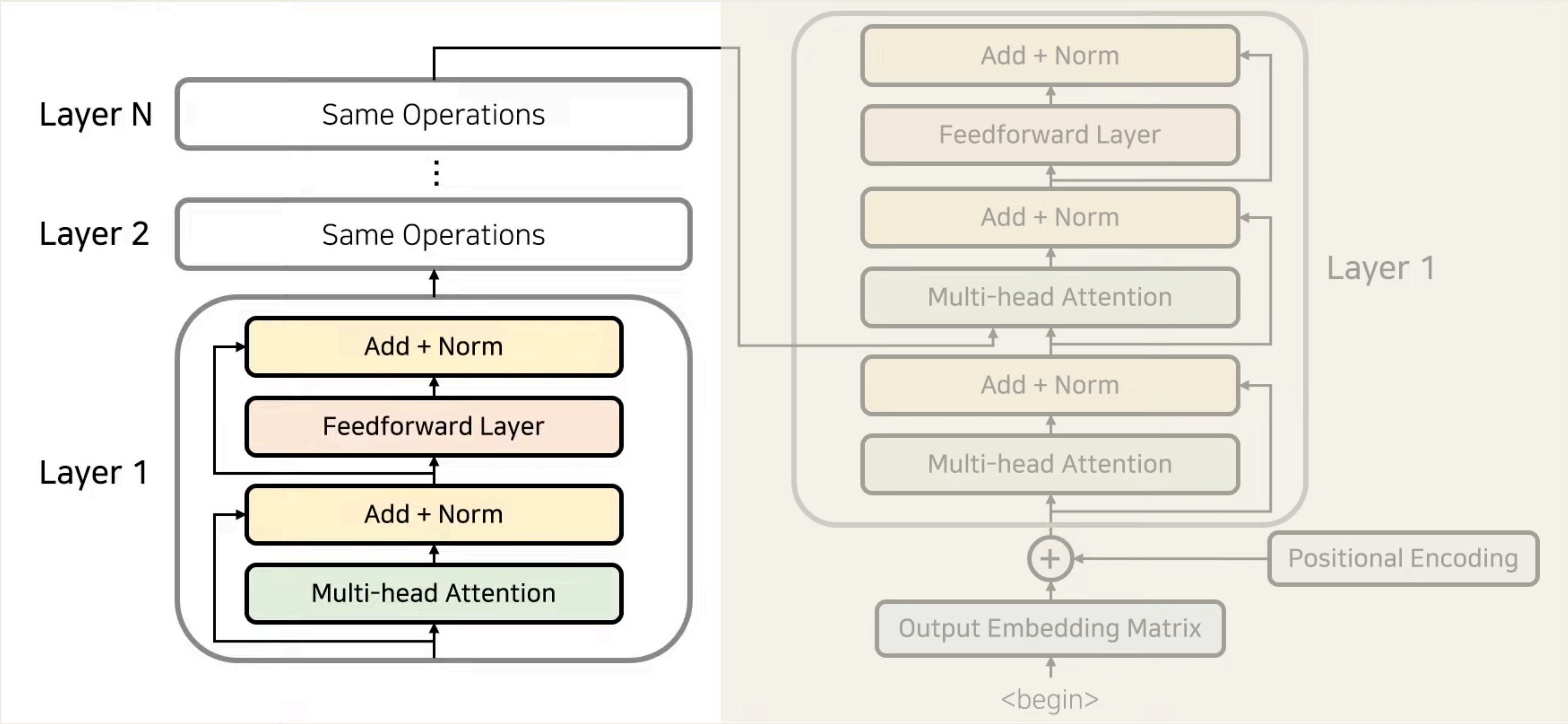
Positional Encoding allows a transformer to record the position of each token in a sentence.

Part 1 : **Encoder**



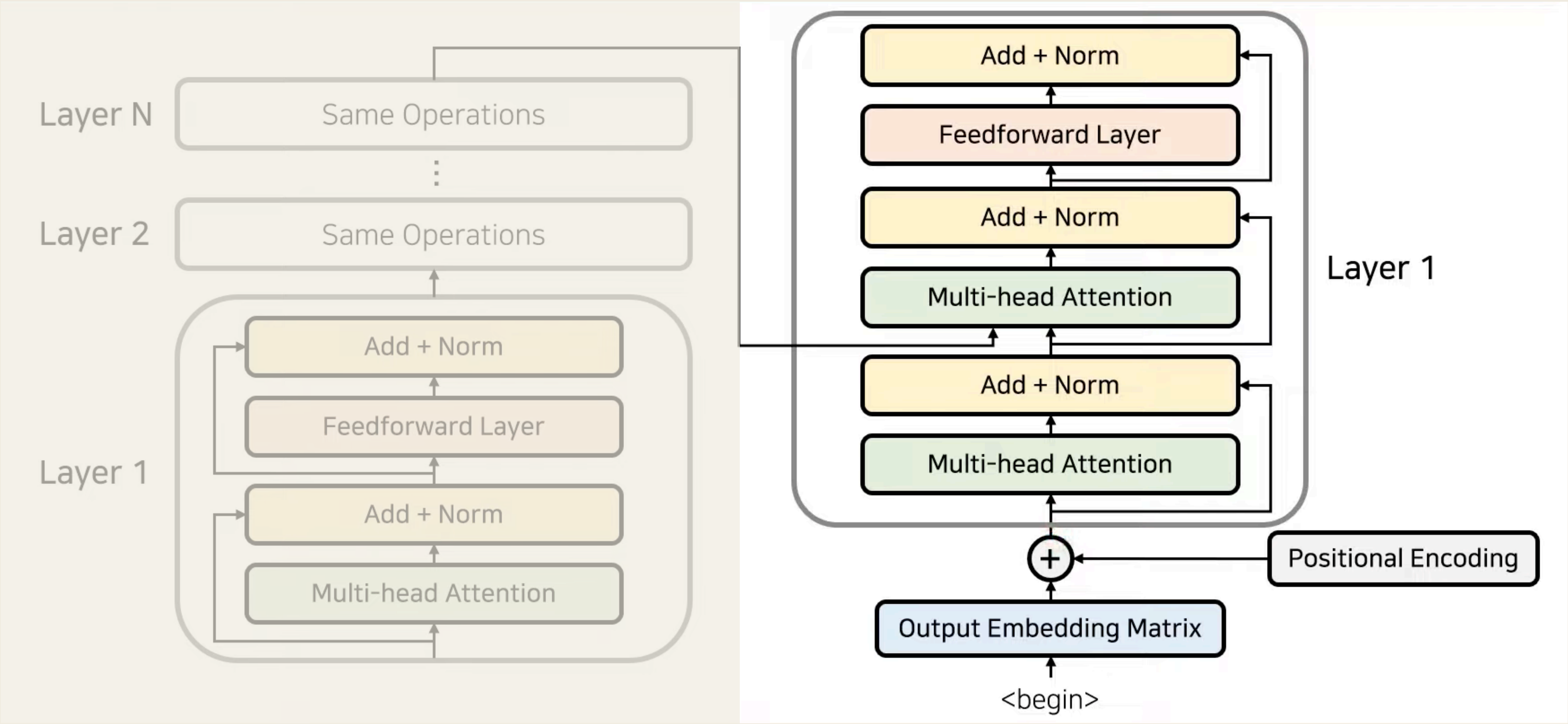
Residual Learning skips a layer to keep the previous information, allowing easier access to global optimization.

Part 1 : **Encoder**



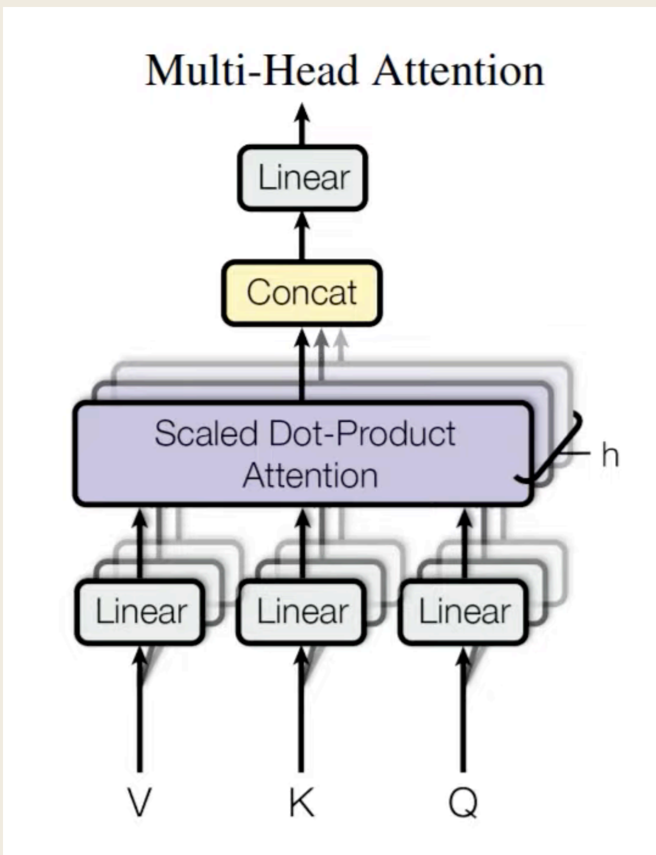
The **Encoder Layers** are composed of the multi-head attention, residual connection and normalization, and the output of the last encoder layer is passed on to the decoder.

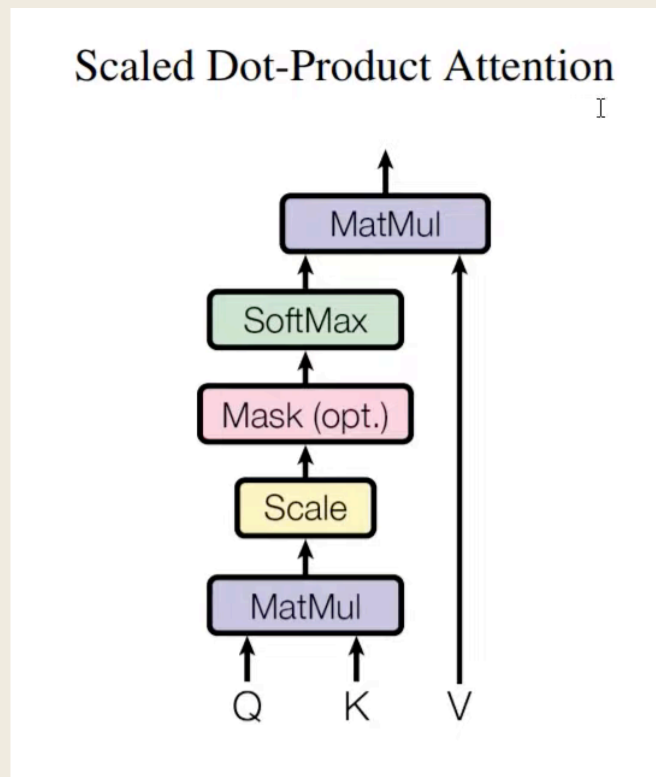
Part 2 : **Decoder**

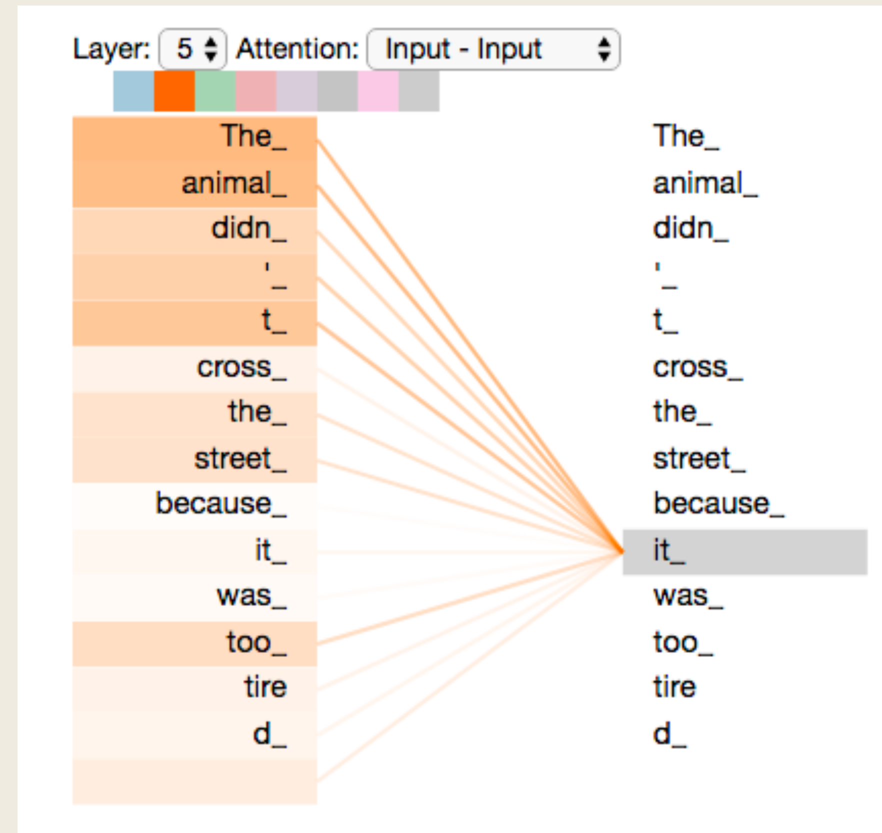


Decoder layer contains two multi-head attention. The purpose of the first attention layer serves similar purpose to that of an encoder, however **the second multi-head attention gets information from the encoder layer to determine the correlation between the current output token and the previously computed encoder outputs.**

Part 3 : **Multi-Head Attention**

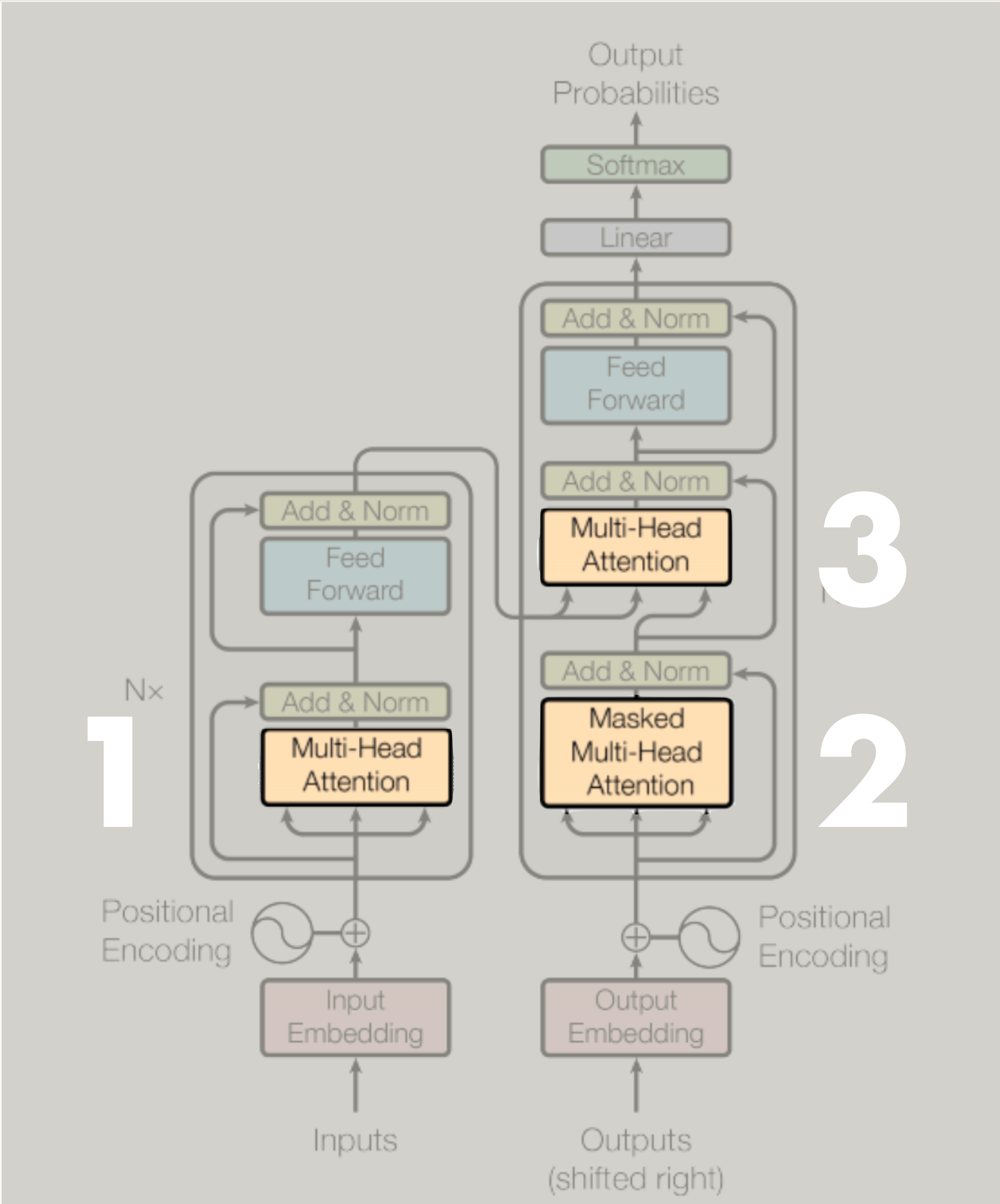






$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Multi-Head Attention uses different key, value, and queries for linear transformation, computes attention scores, and concatenates to create an output.



Part 3 :
Multi-Head Attention

1. **Multi-Head Attention (Encoder)**

   Finding relationship between a word and its surrounding context

2. **Masked Multi-Head Attention**

   Masking 'future' words in a sentence, leaving context of current and previous words

3. **Multi-Head Attention (Encoder-Decoder)**

   Computes scaled-dot product using context information from both the encoder and decoder

Summary

Significantly faster than architectures based on recurrent or convolutional layers.

On both WMT 2014 English-to-German and WMT 2014 English-to-French translation tasks, the transformer achieved a new state of the art in 2017.

Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(log_k(n))$
Self-Attention (restricted)	$O(r \cdot n \cdot d)$	$O(1)$	$O(n/r)$

References

1. **Attention Is All You Need** Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin (2017)
2. **[딥러닝 기계 번역] Transformer: Attention is All You Need (꼼꼼한 딥러닝 논문 리뷰와 코드 실습) -** https://www.youtube.com/watch?v=AA621UofTUA
3. **The Illustrated Transformer** Jay Alammar - https://jalammar.github.io/illustrated-transformer/