The Illustrated Transformer

Jay Alammar
Visualizing machine learning one concept at a time.
@JayAlammar on Twitter. YouTube Channel

Artificial Intelligence

Creating the Future

Dong-A University

Division of Computer Engineering & Artificial Intelligence

References

Main

- https://jalammar.github.io/illustrated-transformer/

Illustrated Transformer

Transformer

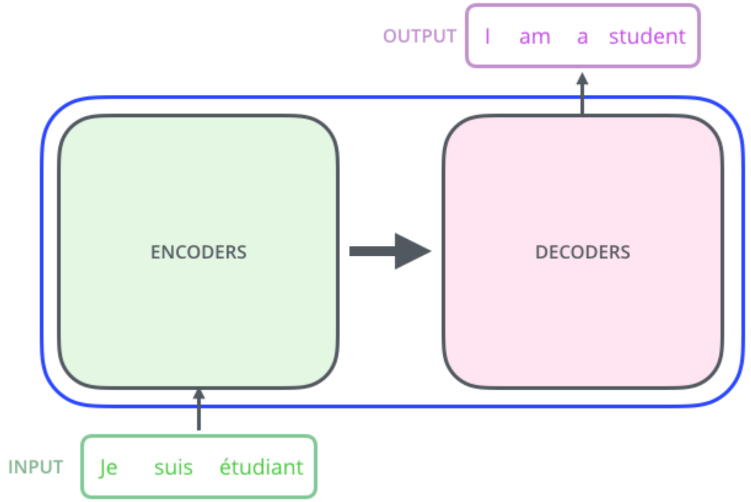
- A model that uses **attention** to boost the speed with which these models can be trained.
- The Transformers outperforms the Google Neural Machine Translation model in specific tasks.
- The biggest benefit, however, comes from how The Transformer lends itself to **parallelization**. It is in fact Google Cloud's recommendation to use The Transformer as a reference model to use their Cloud TPU offering.

- A TensorFlow implementation of it is available as a part of the Tensor2Tensor package.
- Harvard's NLP group created a guide annotating the paper with PyTorch implementation.

A High-Level Look

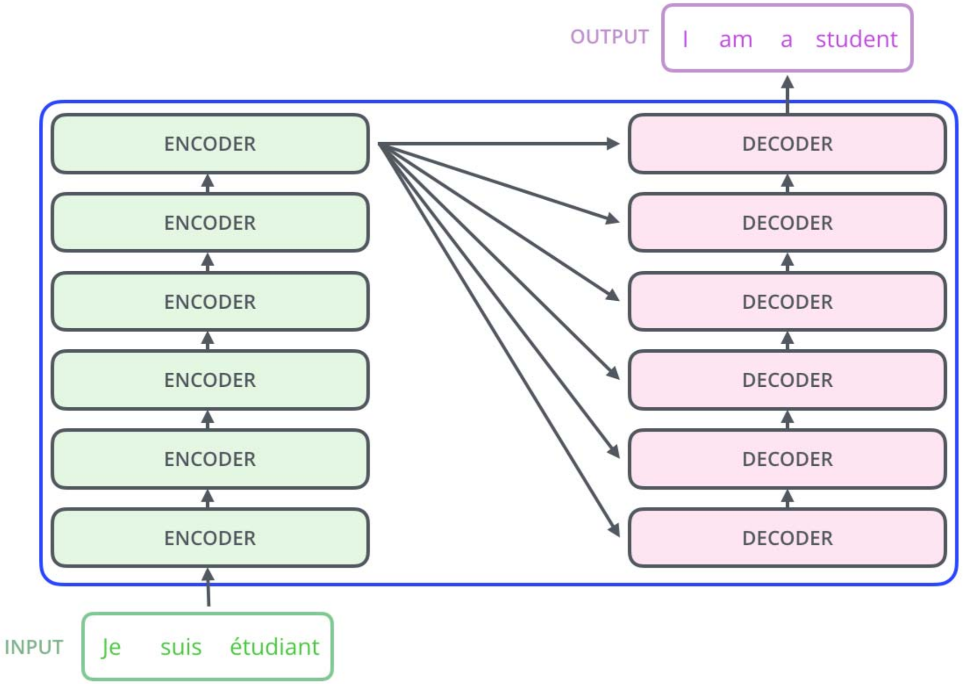
Illustrated Transformer

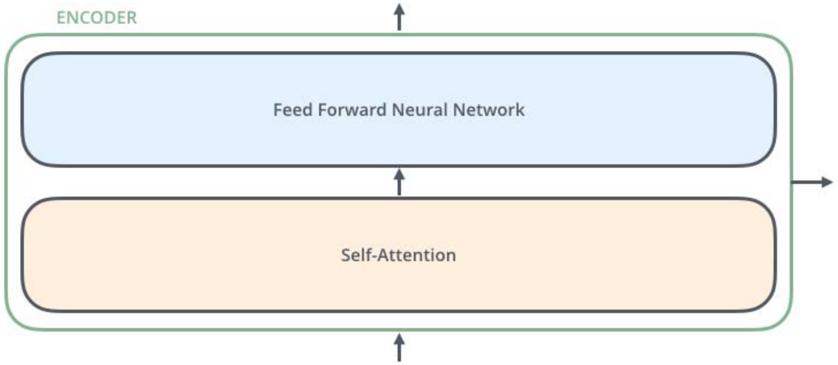
A High-Level Look

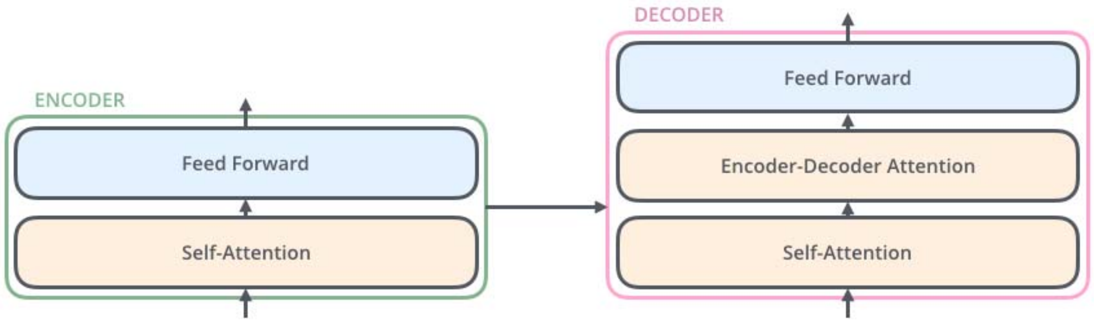


Illustrated Transformer

A High-Level Look

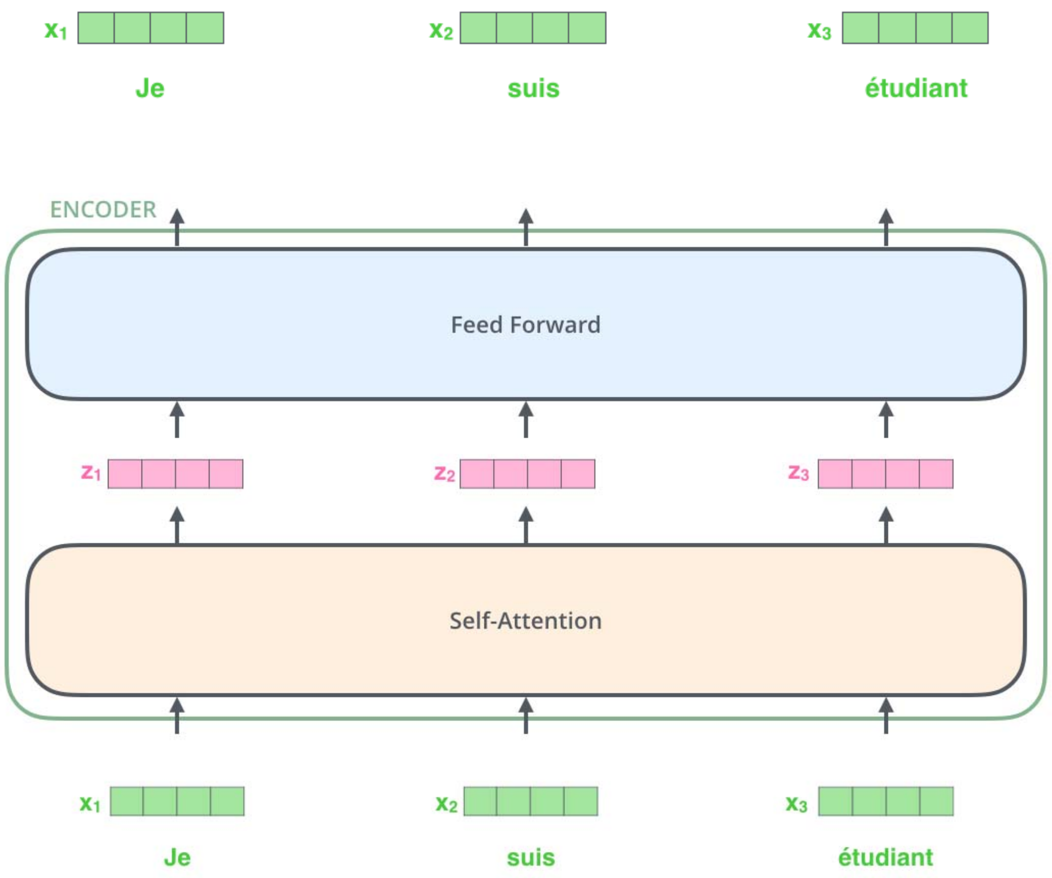




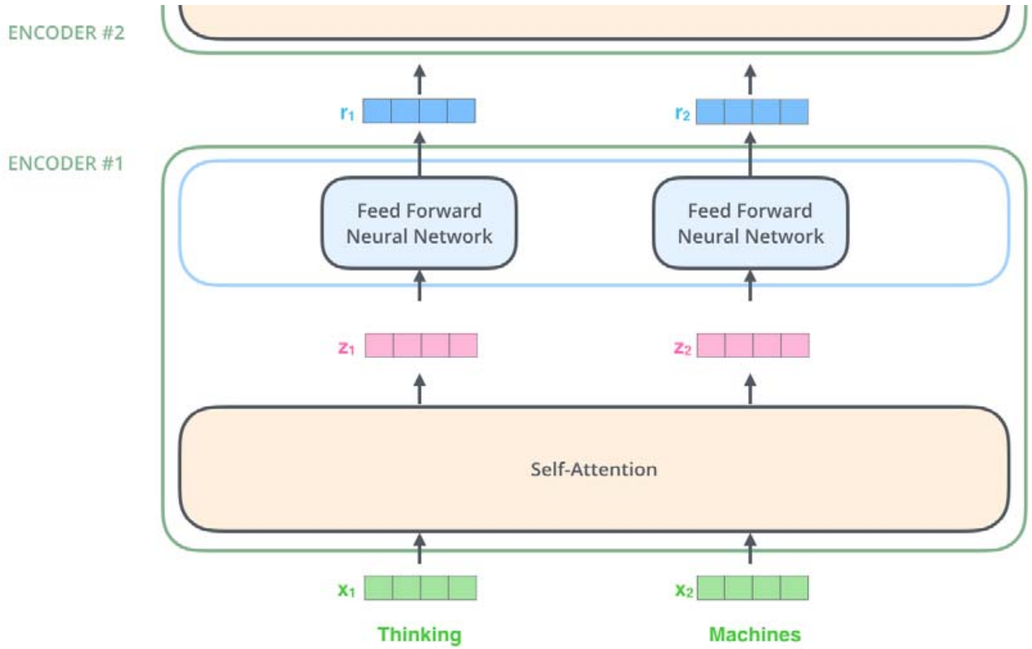


Illustrated Transformer

Bringing The Tensors Into The Picture

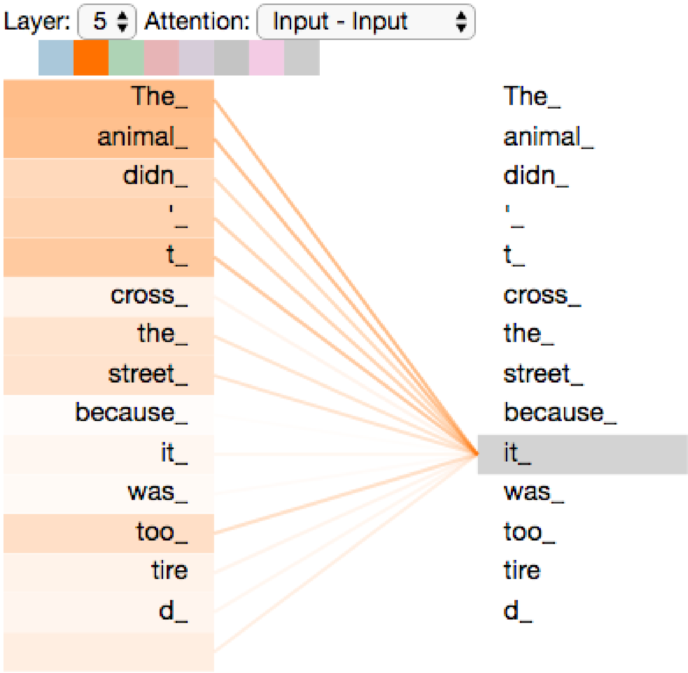


Bringing The Tensors Into The Picture

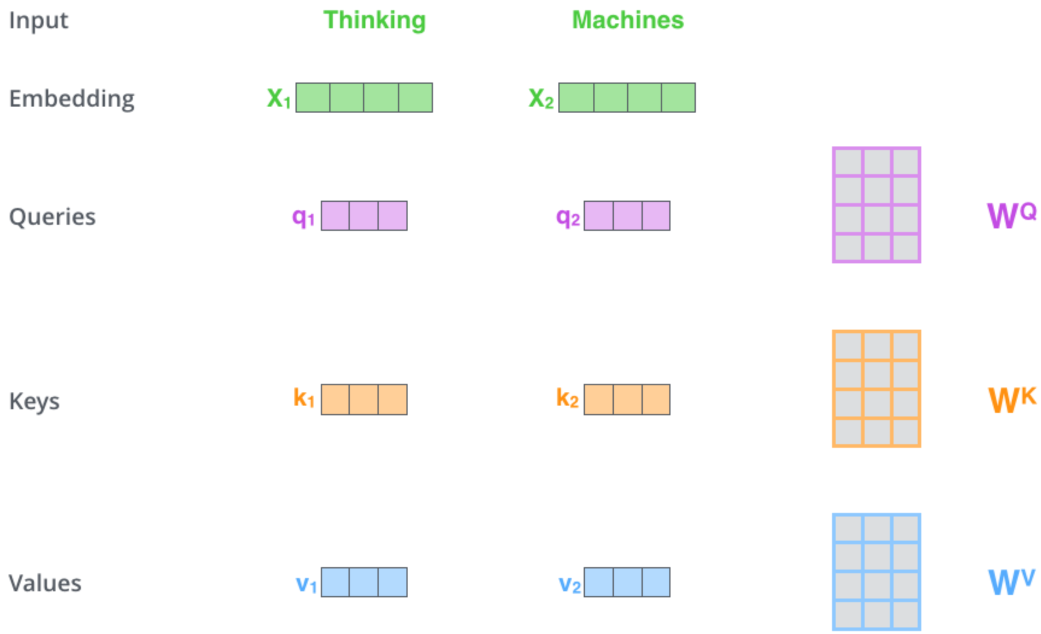


Illustrated Transformer

Self-Attention at a High Level

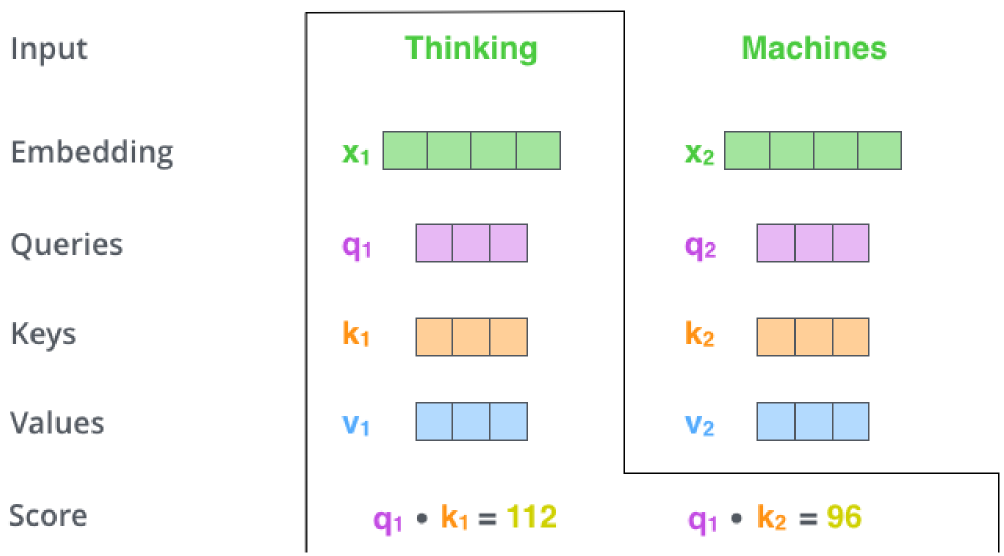


Self-Attention in Detail

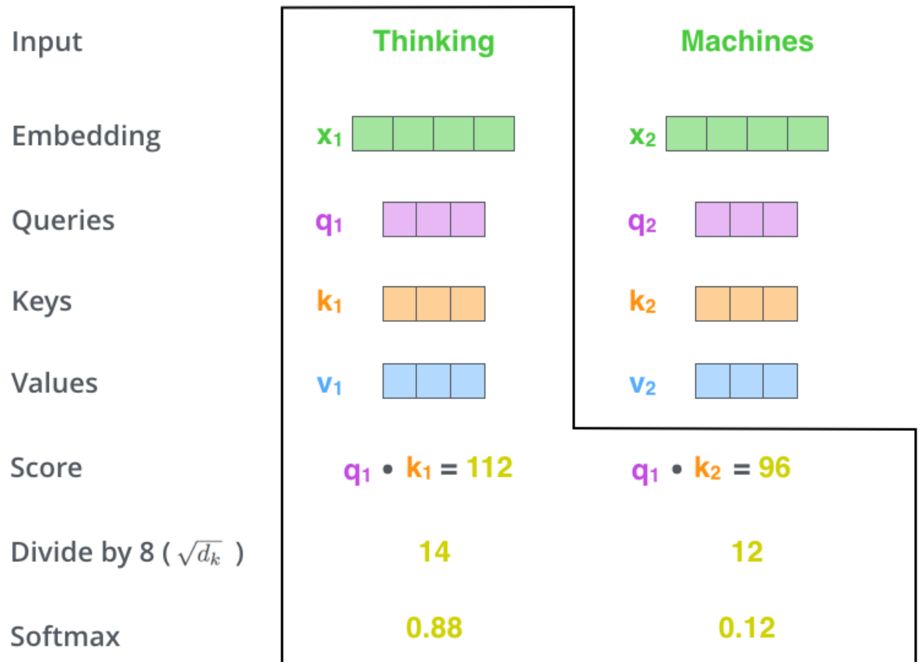


Illustrated Transformer

Self-Attention in Detail

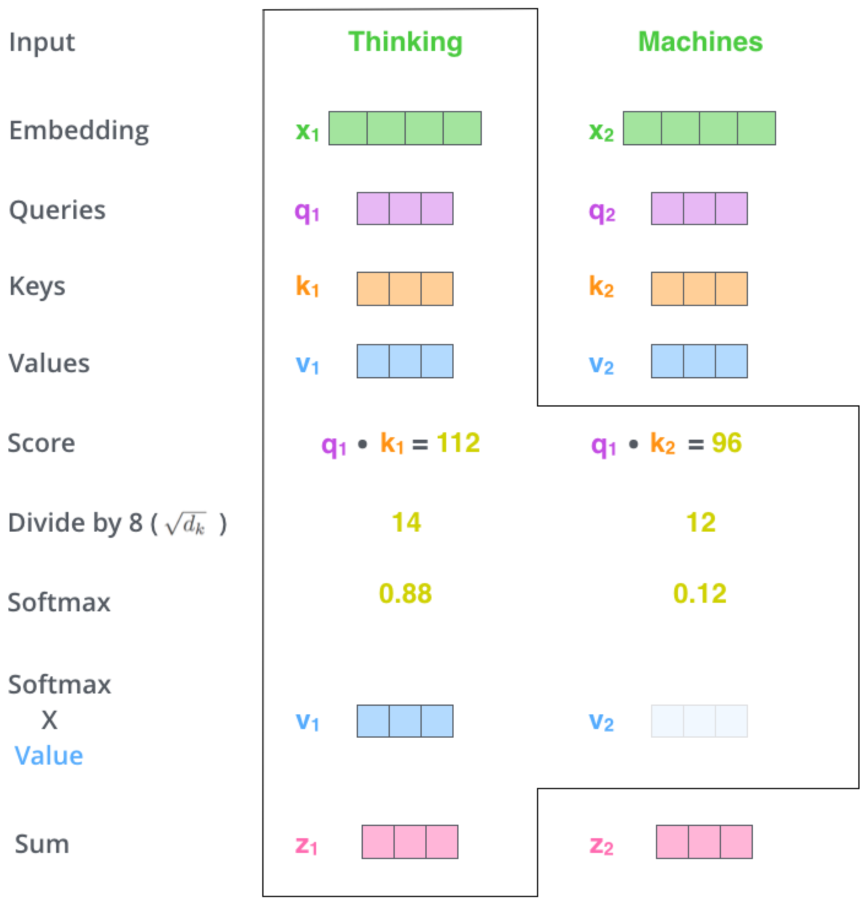


Self-Attention in Detail

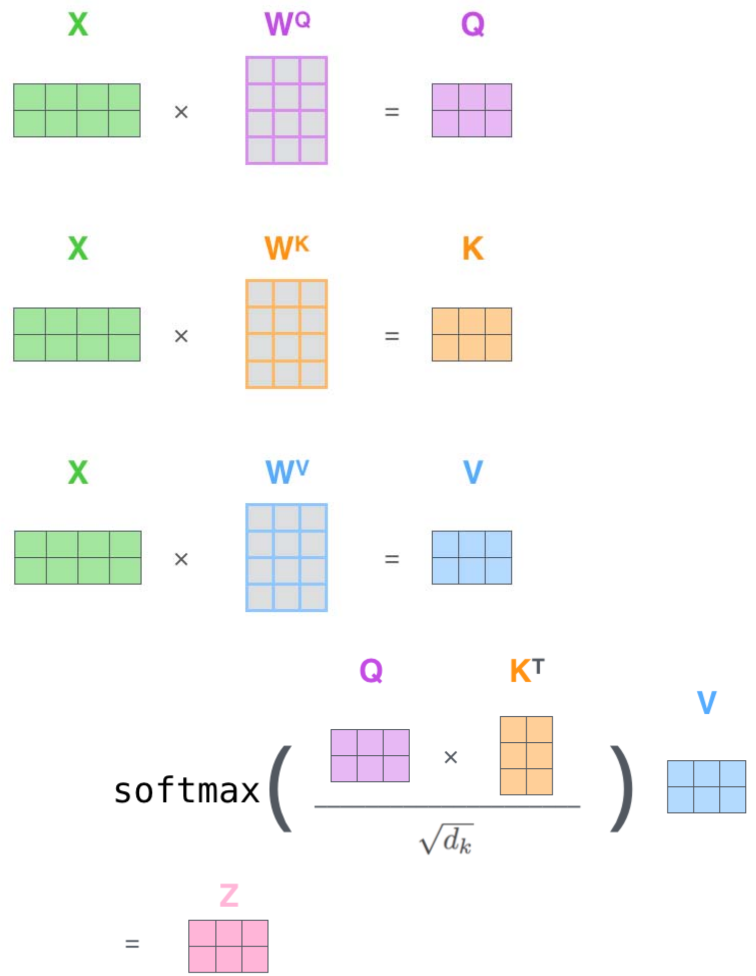


Illustrated Transformer

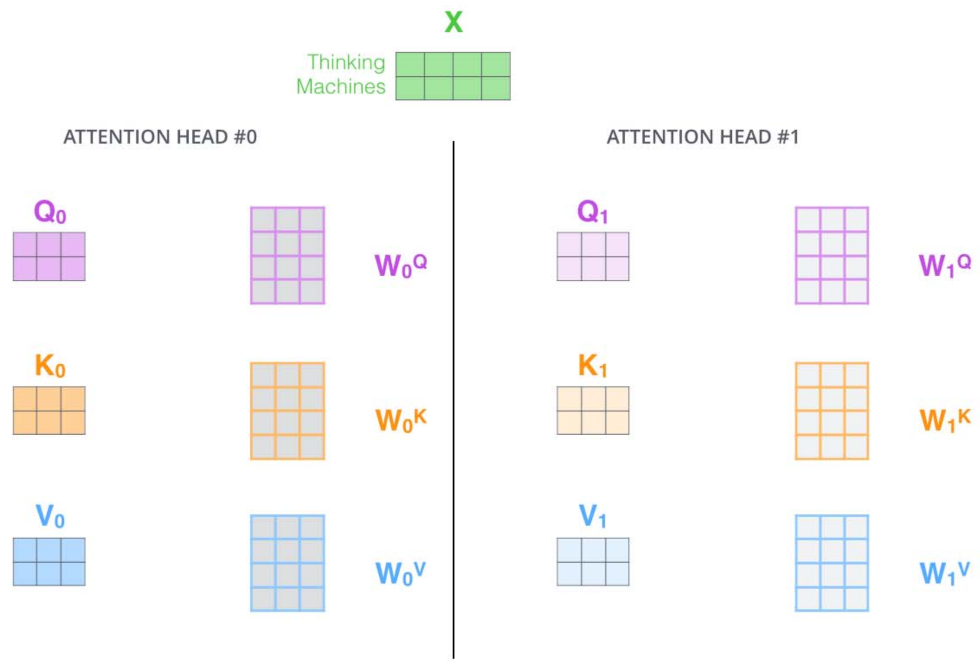
Self-Attention in Detail

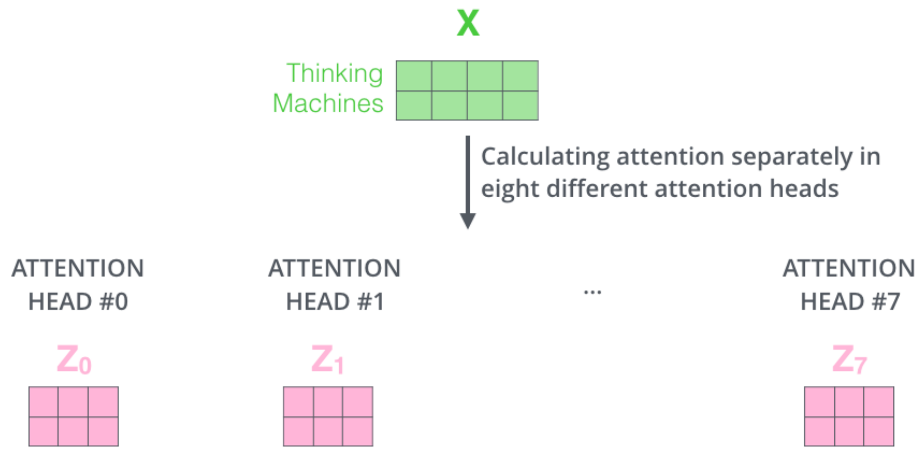


Matrix Calculation of Self-Attention

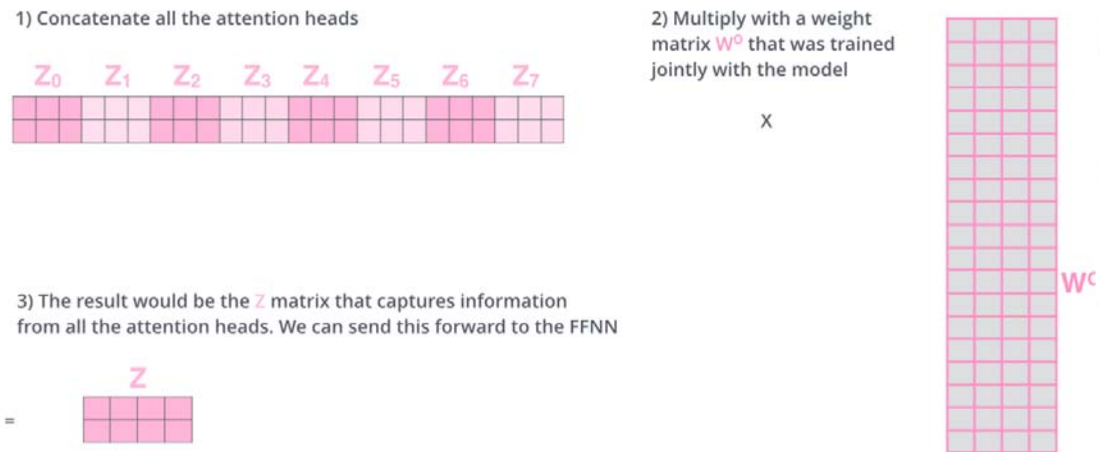


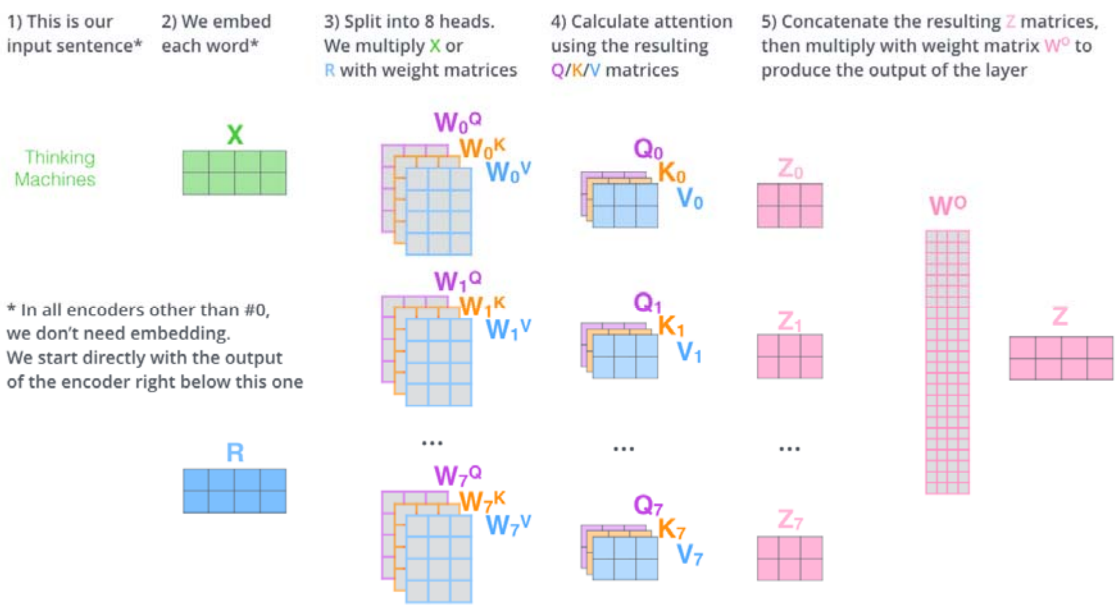
The Beast With Many Heads



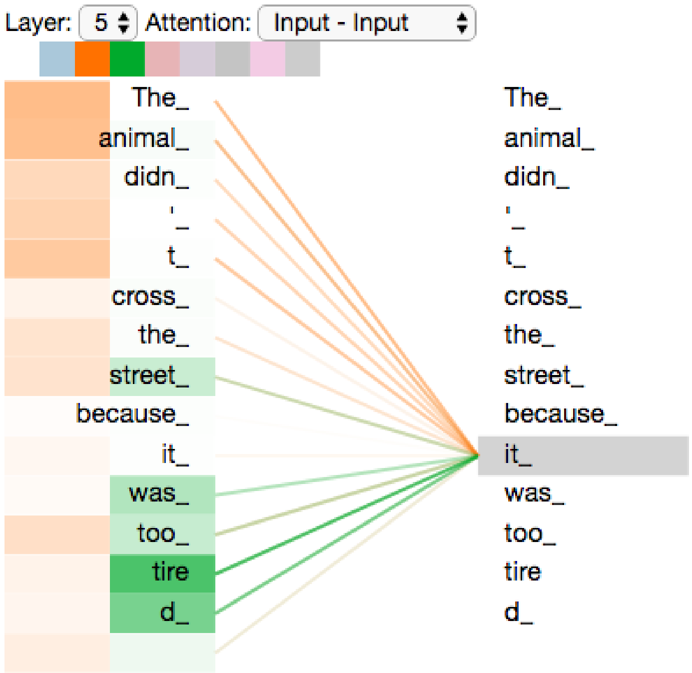


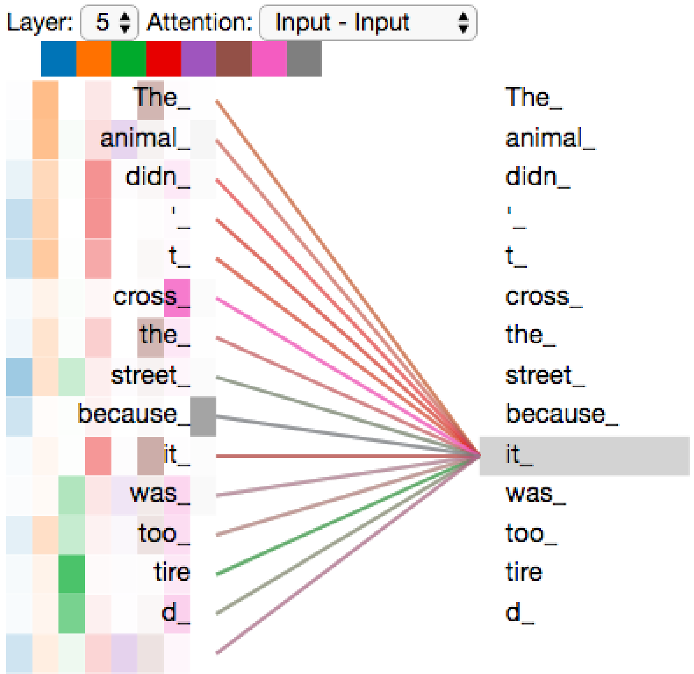
The Beast With Many Heads



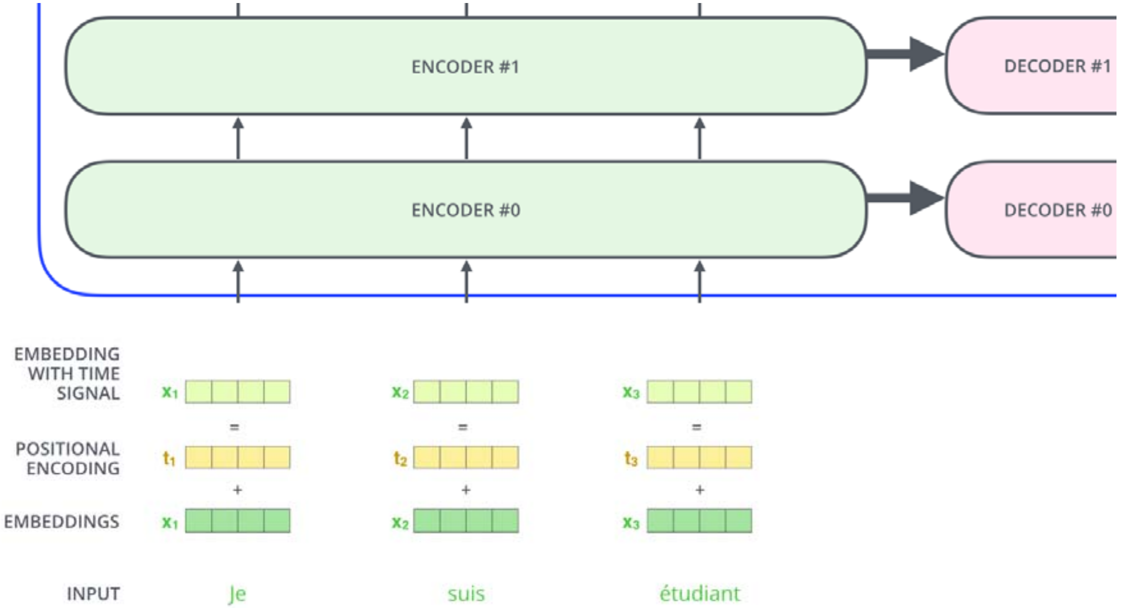


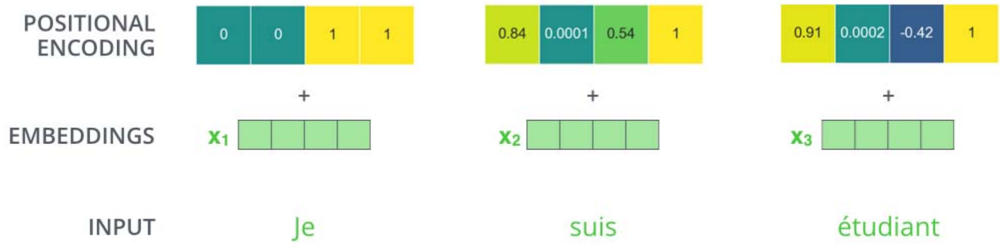
The Beast With Many Heads



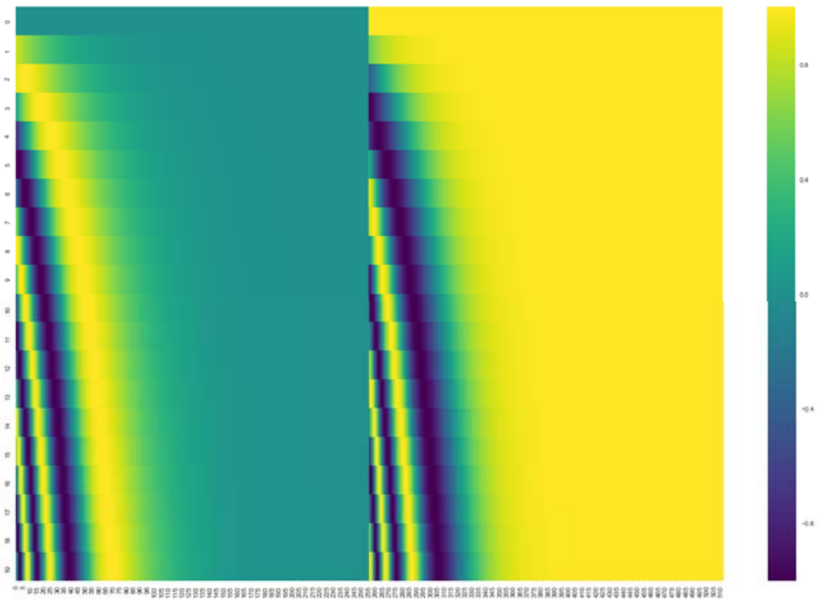


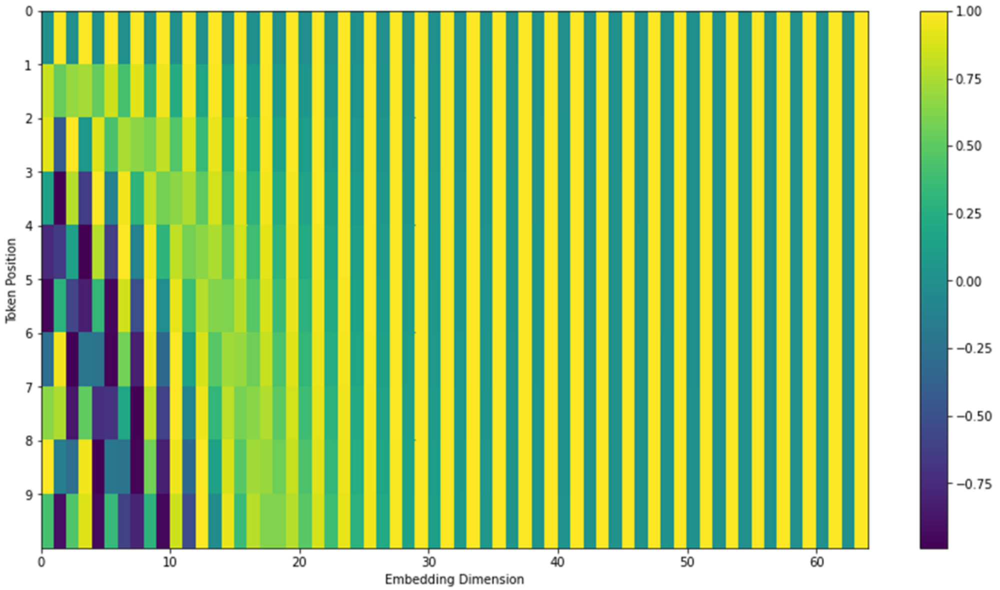
Representing The Order of The Sequence Using Positional Encoding





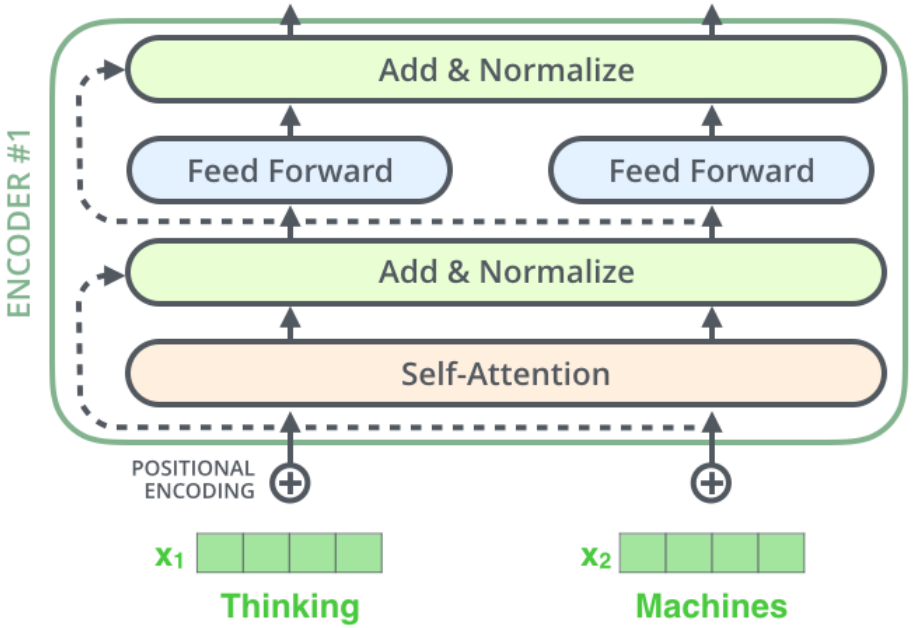
Representing The Order of The Sequence Using Positional Encoding

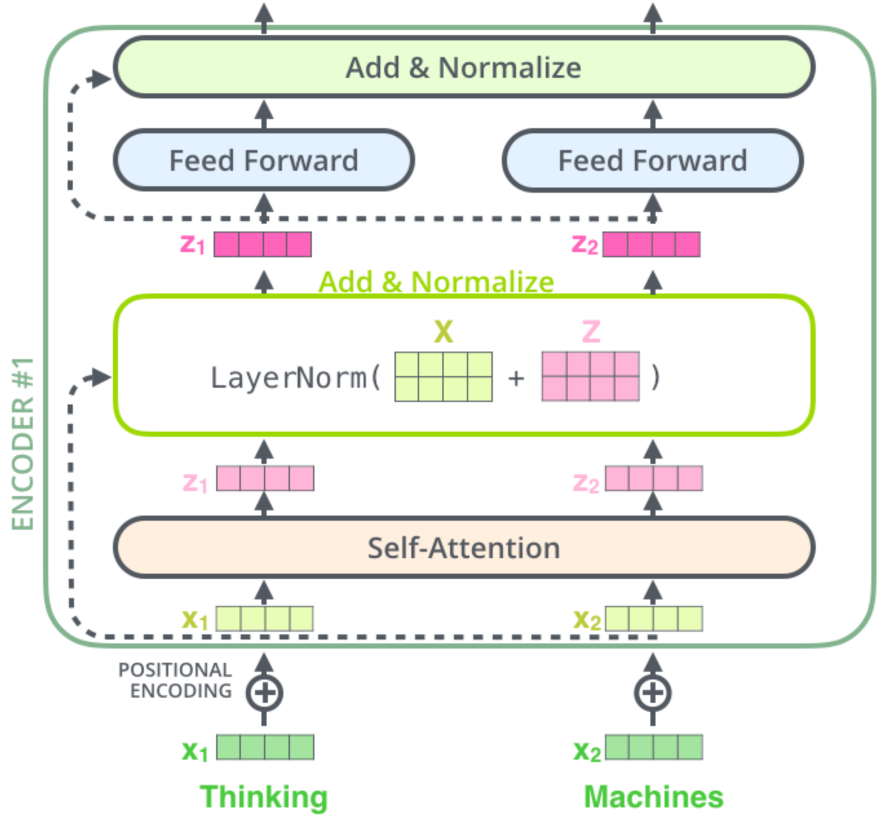




Illustrated Transformer

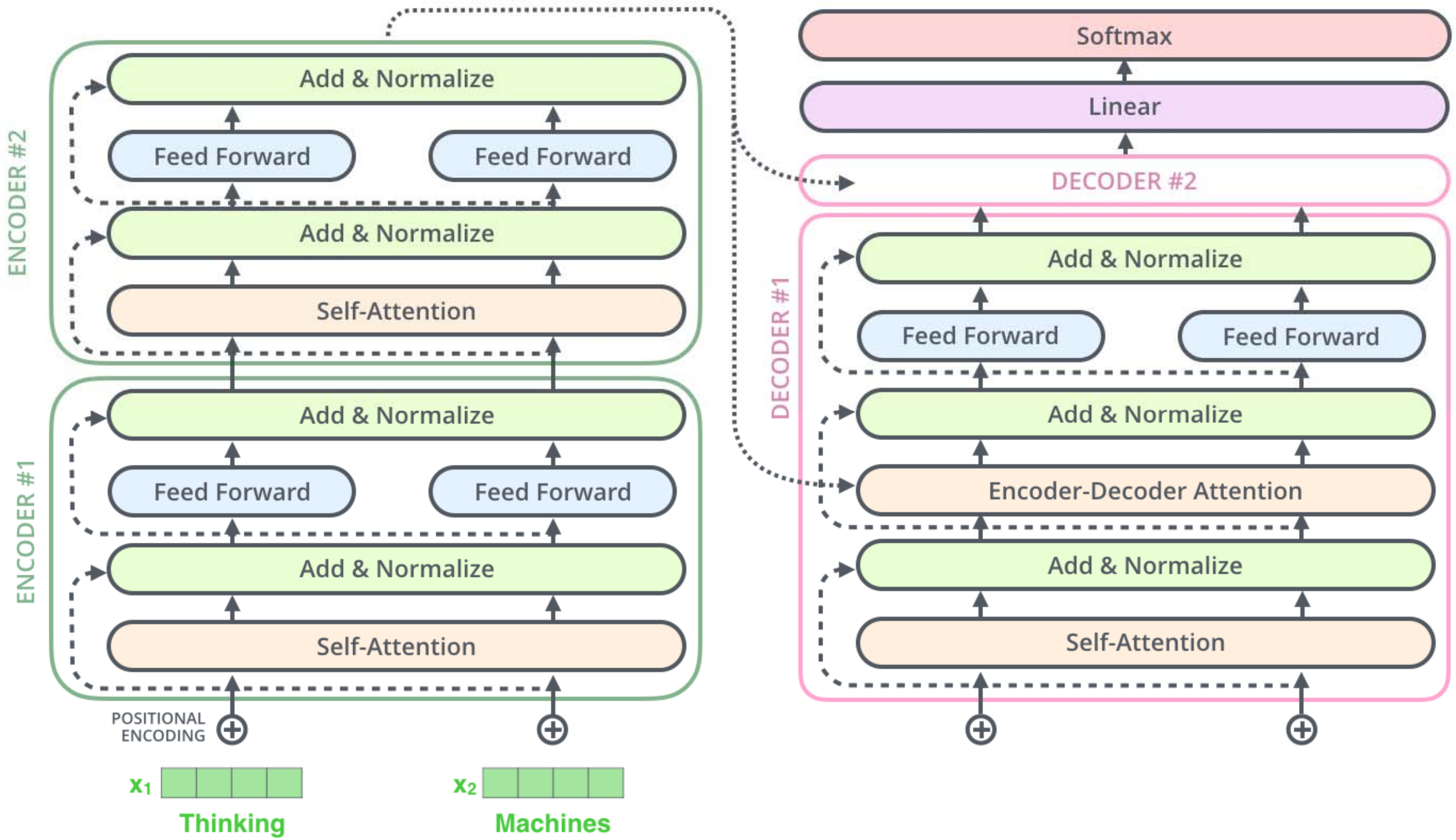
The Residuals





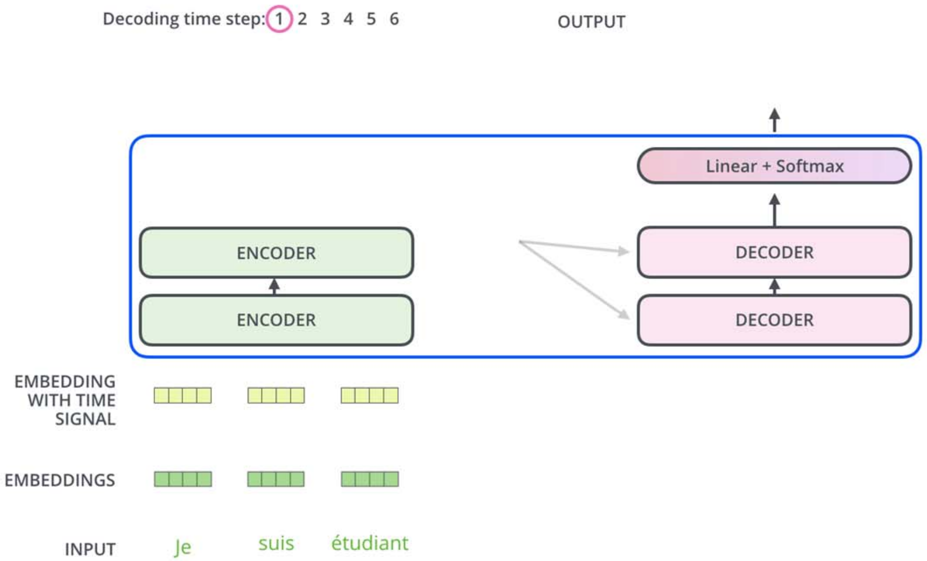
Illustrated Transformer

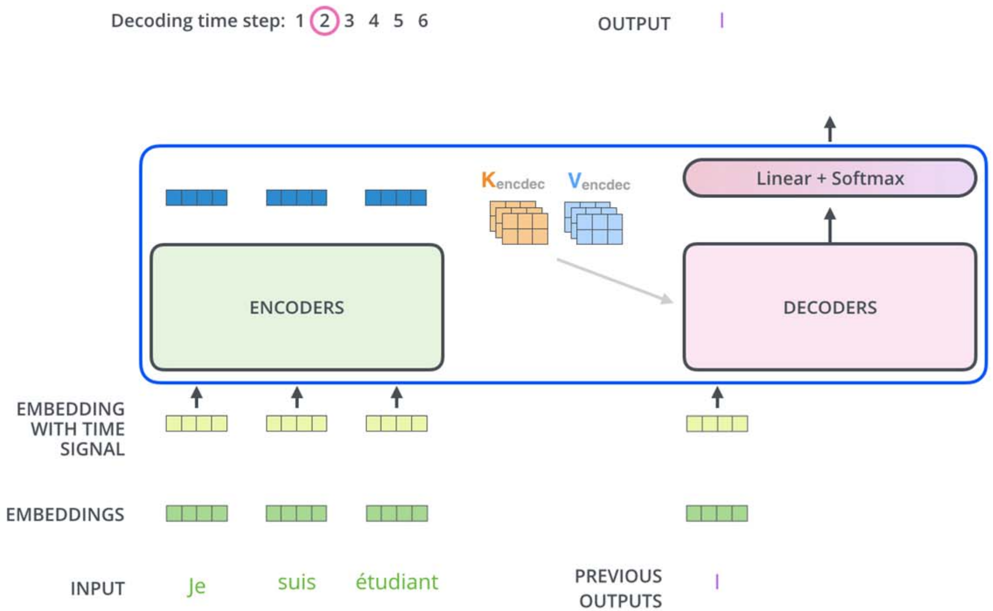
The Residuals



Illustrated Transformer

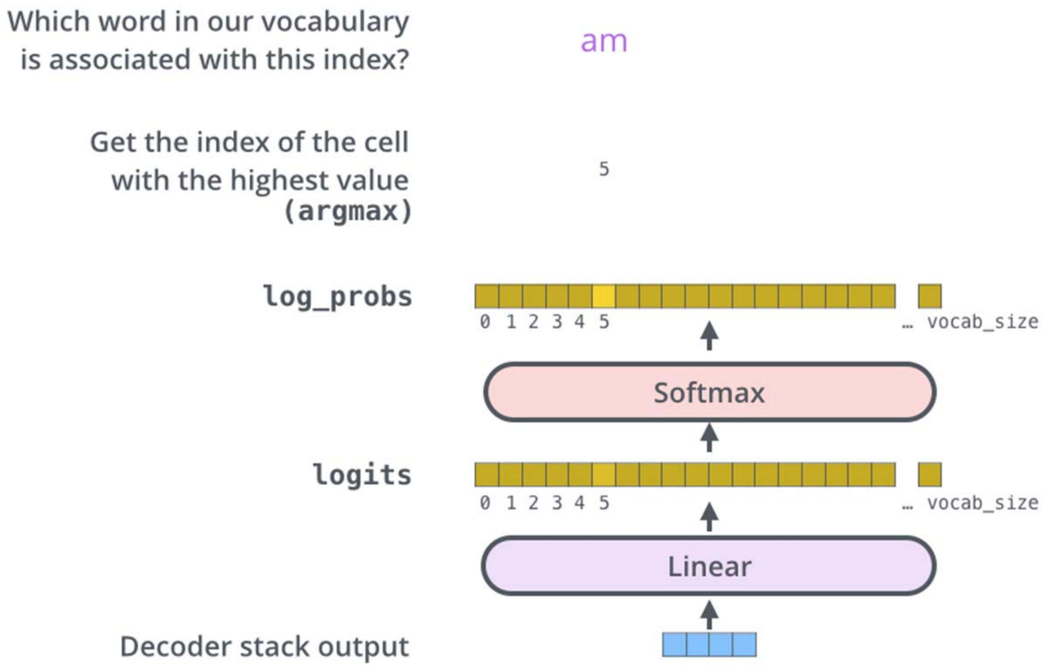
The Decoder Side





Illustrated Transformer

The Final Linear and Softmax Layer



Recap Of Training

Output Vocabulary

WORD	a	am	I	thanks	student	<eos>
INDEX	0	1	2	3	4	5

Output Vocabulary

WORD	a	am	I	thanks	student	<eos>
INDEX	0	1	2	3	4	5

One-hot encoding of the word "am"

0.0	1.0	0.0	0.0	0.0	0.0

Illustrated Transformer

The Loss Function

Untrained Model Output

a	am	I	thanks	student	<eos>
0.2	0.2	0.1	0.2	0.2	0.1

Correct and desired output

a	am	I	thanks	student	<eos>
0.0	0.0	0.0	1.0	0.0	0.0

Target Model Outputs

Output Vocabulary:	a	am	I	thanks	student	<eos>
position #1	0.0	0.0	1.0	0.0	0.0	0.0
position #2	0.0	1.0	0.0	0.0	0.0	0.0
position #3	1.0	0.0	0.0	0.0	0.0	0.0
position #4	0.0	0.0	0.0	0.0	1.0	0.0
position #5	0.0	0.0	0.0	0.0	0.0	1.0
	a	am	I	thanks	student	<eos>

Trained Model Outputs

Output Vocabulary:	a	am	I	thanks	student	<eos>
position #1	0.01	0.02	0.93	0.01	0.03	0.01
position #2	0.01	0.8	0.1	0.05	0.01	0.03
position #3	0.99	0.001	0.001	0.001	0.002	0.001
position #4	0.001	0.002	0.001	0.02	0.94	0.01
position #5	0.01	0.01	0.001	0.001	0.001	0.98
	a	am	I	thanks	student	<eos>