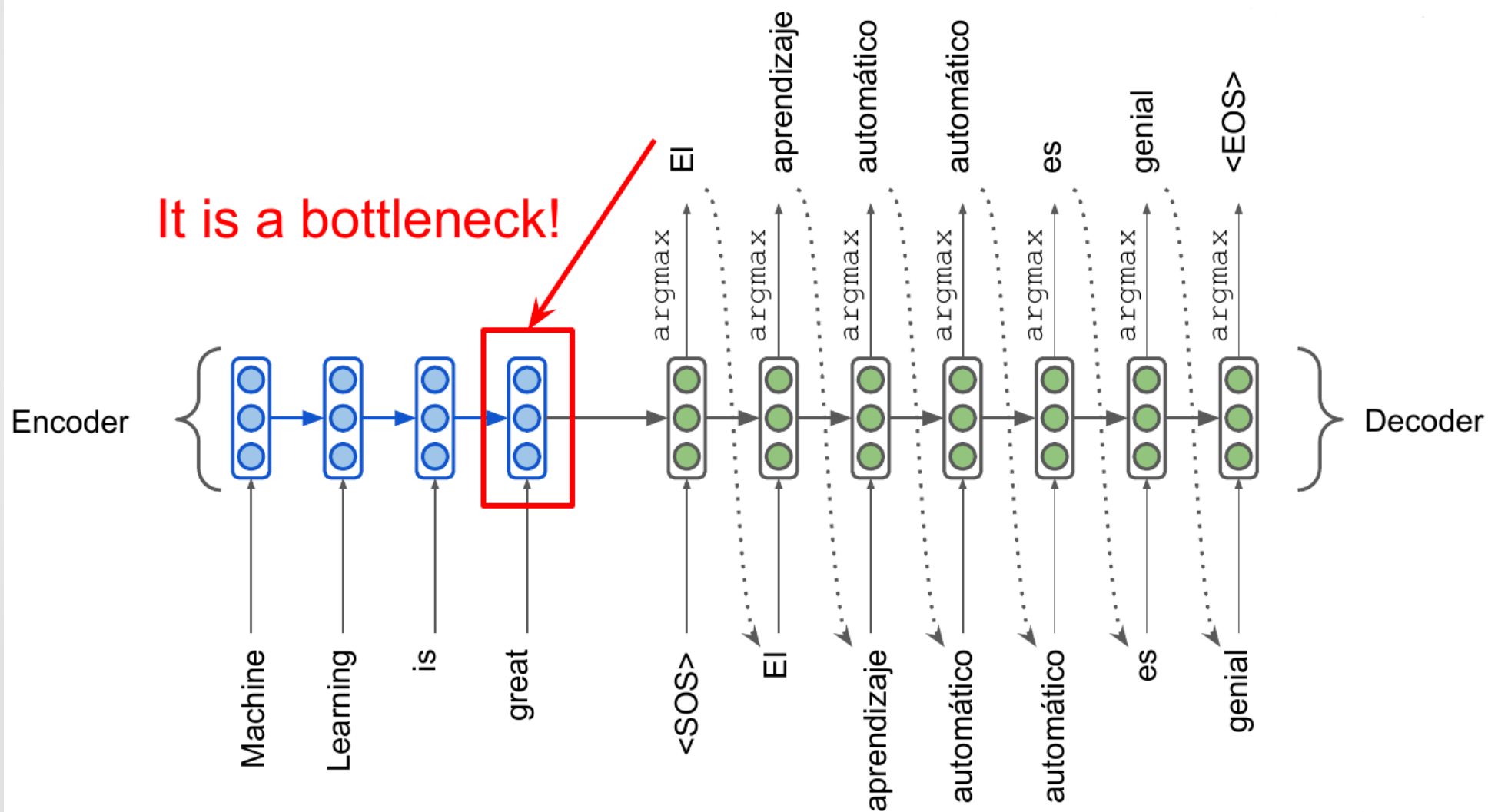




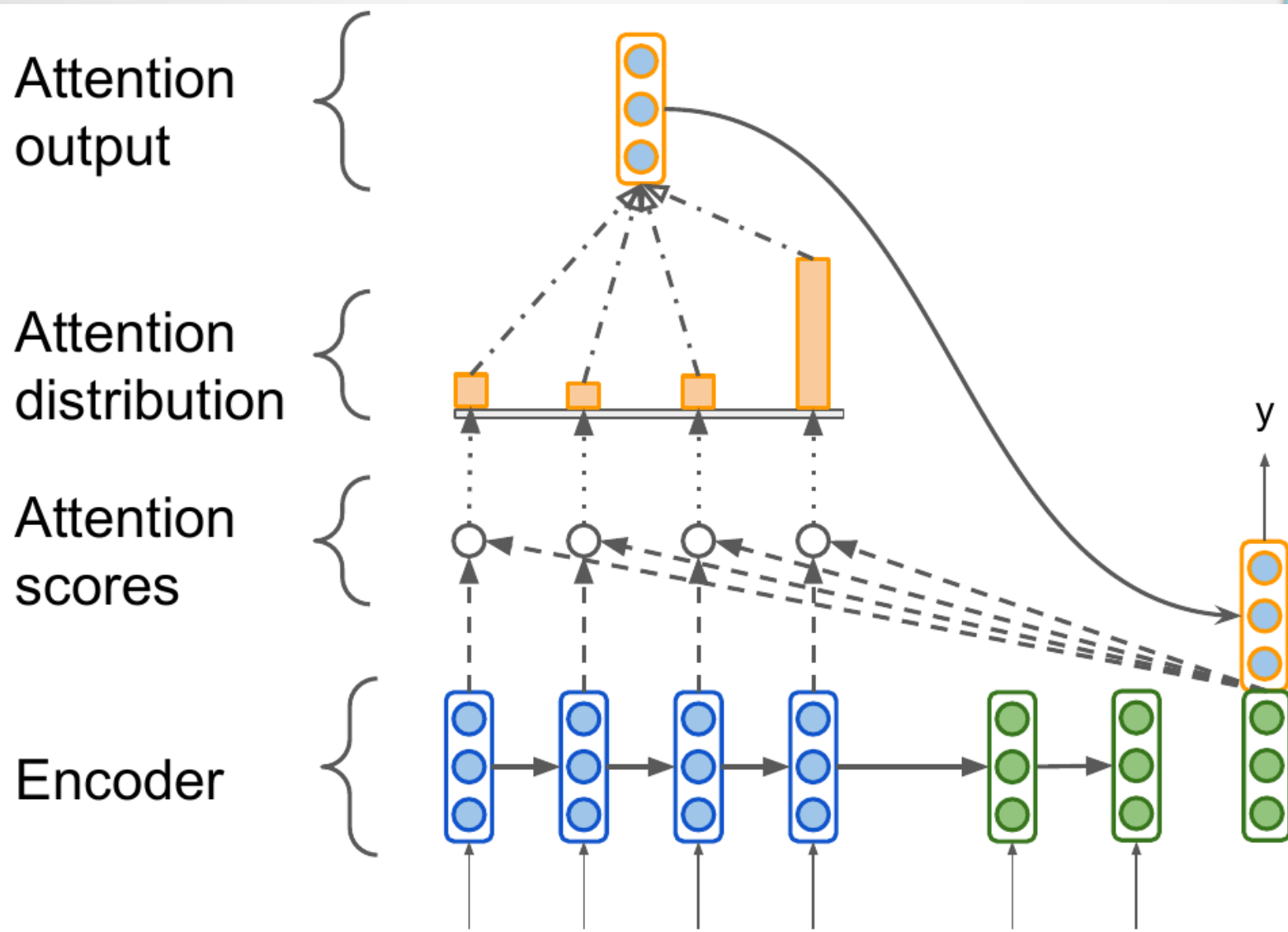
Модель Transformer и механизм внимания Self-Attention

Евгений Борисов

SEQ2SEQ NMT



SEQ2SEQ NMT with ATTENTION

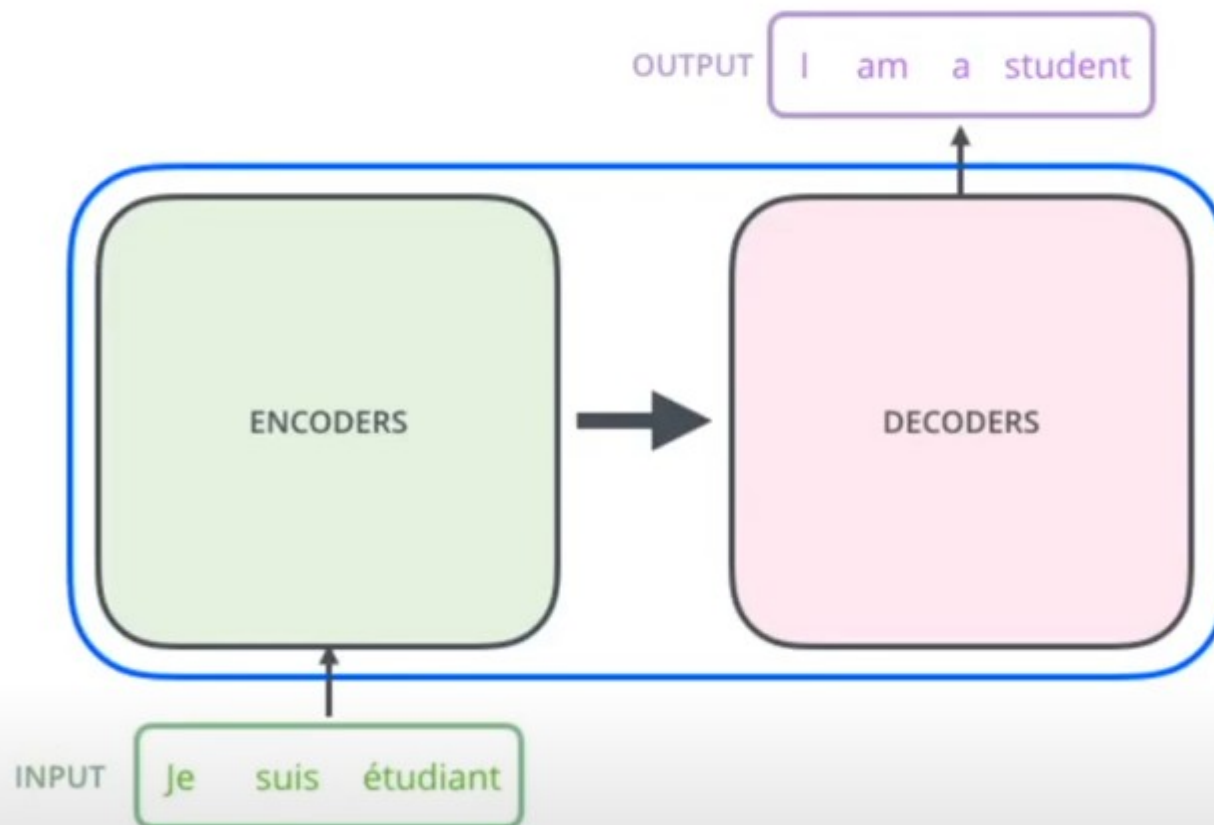


Transformer

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

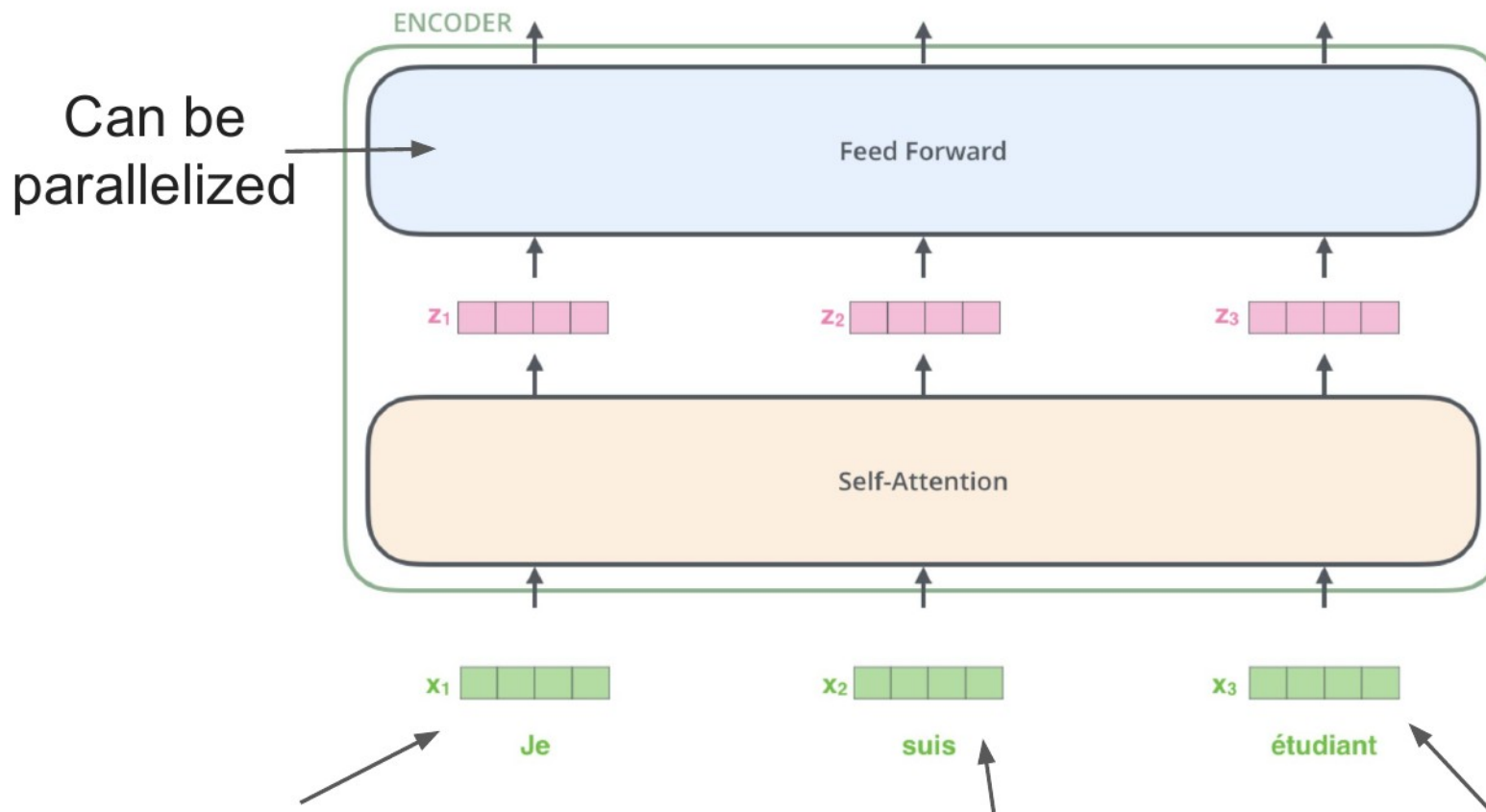
Attention Is All You Need, **2017**

The Transformer



Transformer

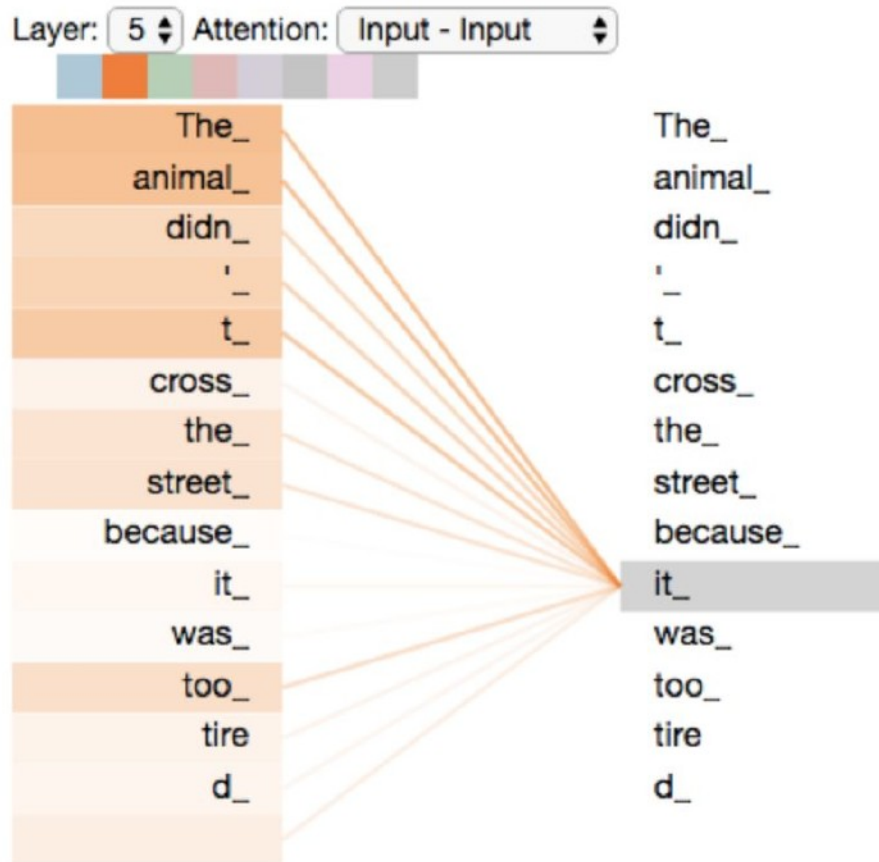
The Encoder Side



the word in each position flows through its own path in the encoder

Transformer

Self-Attention at a High Level



Transformer

Self-Attention

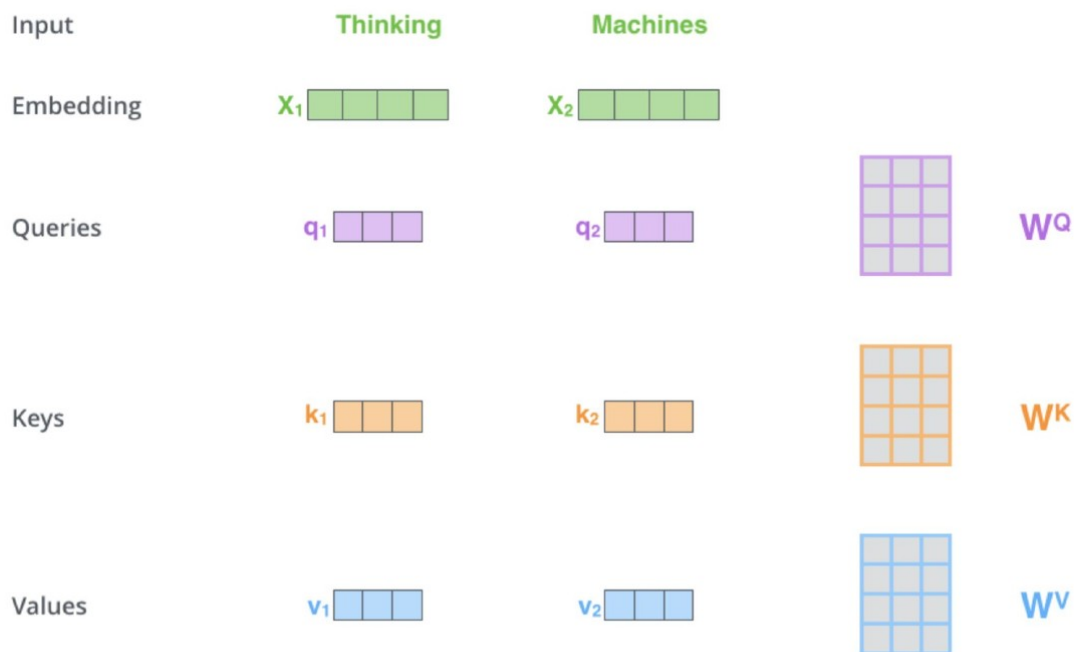
Transformer Self-Attention

query - откуда смотрим (из какого слова)

key - куда смотрим (на какое слово)

value - смысл (условно) слова

Self-Attention: detailed explanation

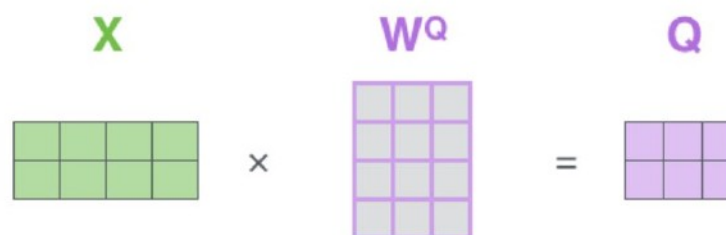


Transformer Self-Attention

Self-Attention: Matrix Calculation

Pack embeddings into matrix **X**

Multiply **X** by weight matrices we've trained (**W_k**, **W_q**, **W_v**)

$$\mathbf{X} \times \mathbf{W}^Q = \mathbf{Q}$$


$$\mathbf{X} \times \mathbf{W}^K = \mathbf{K}$$

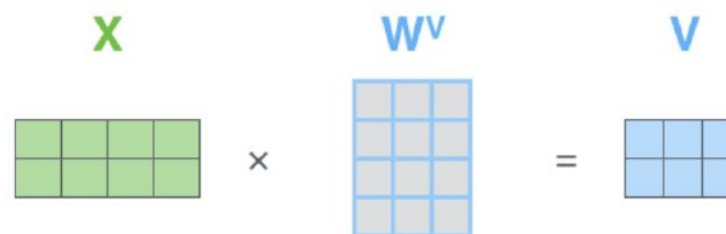
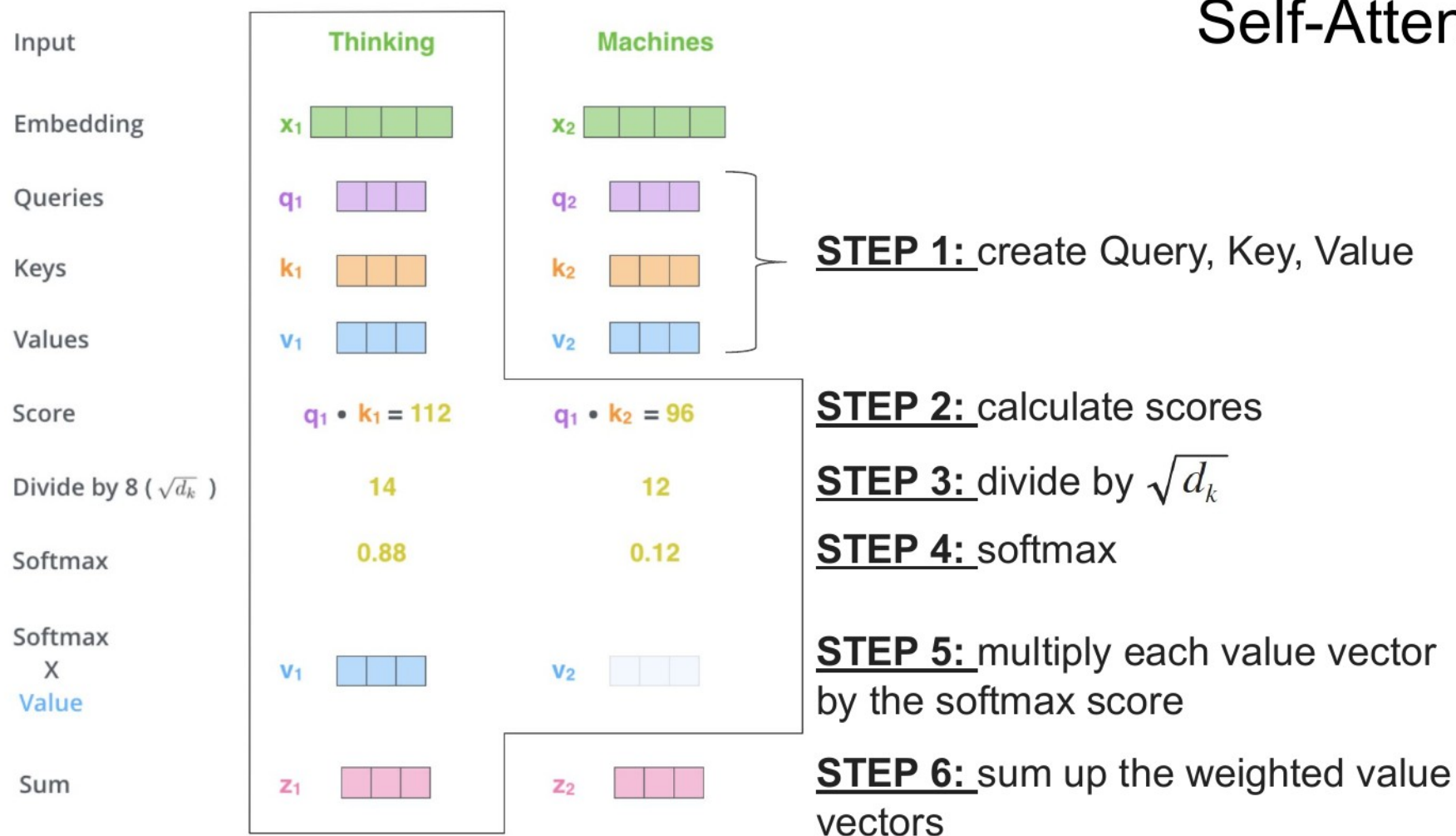

$$\mathbf{X} \times \mathbf{W}^V = \mathbf{V}$$


Image source: <https://jalammar.github.io/illustrated-transformer/>

Transformer Self-Attention

Self-Attention



Transformer Self-Attention

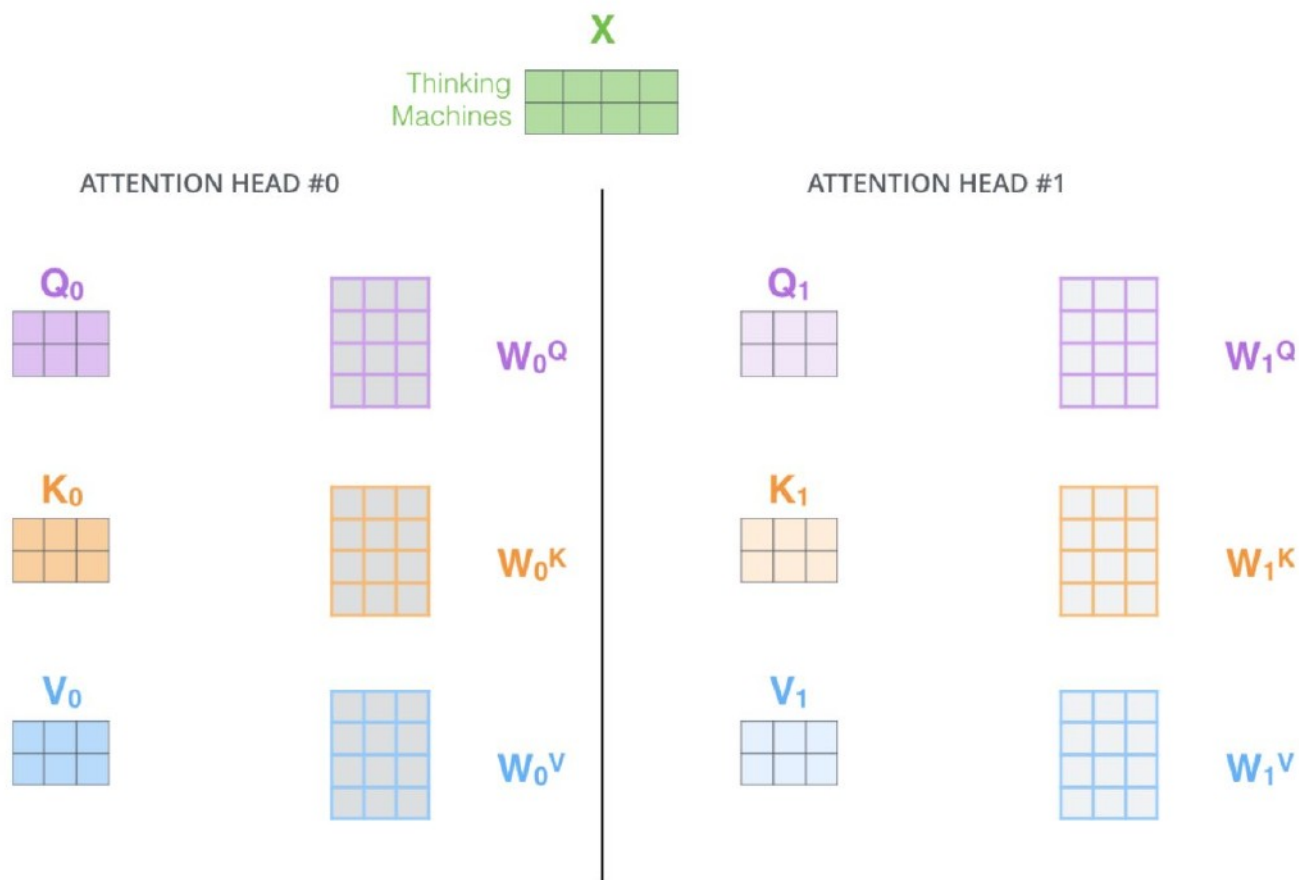
Self-Attention: Matrix Calculation

$$\text{softmax} \left(\frac{\overset{\text{Q}}{\begin{array}{|c|c|c|} \hline & & \\ \hline & & \\ \hline \end{array}} \times \overset{\text{K}^T}{\begin{array}{|c|c|} \hline & \\ \hline & \\ \hline & \\ \hline \end{array}} \right) \overset{\text{V}}{\begin{array}{|c|c|c|} \hline & & \\ \hline & & \\ \hline \end{array}} = \overset{\text{Z}}{\begin{array}{|c|c|c|} \hline & & \\ \hline & & \\ \hline \end{array}}$$

Transformer Self-Attention

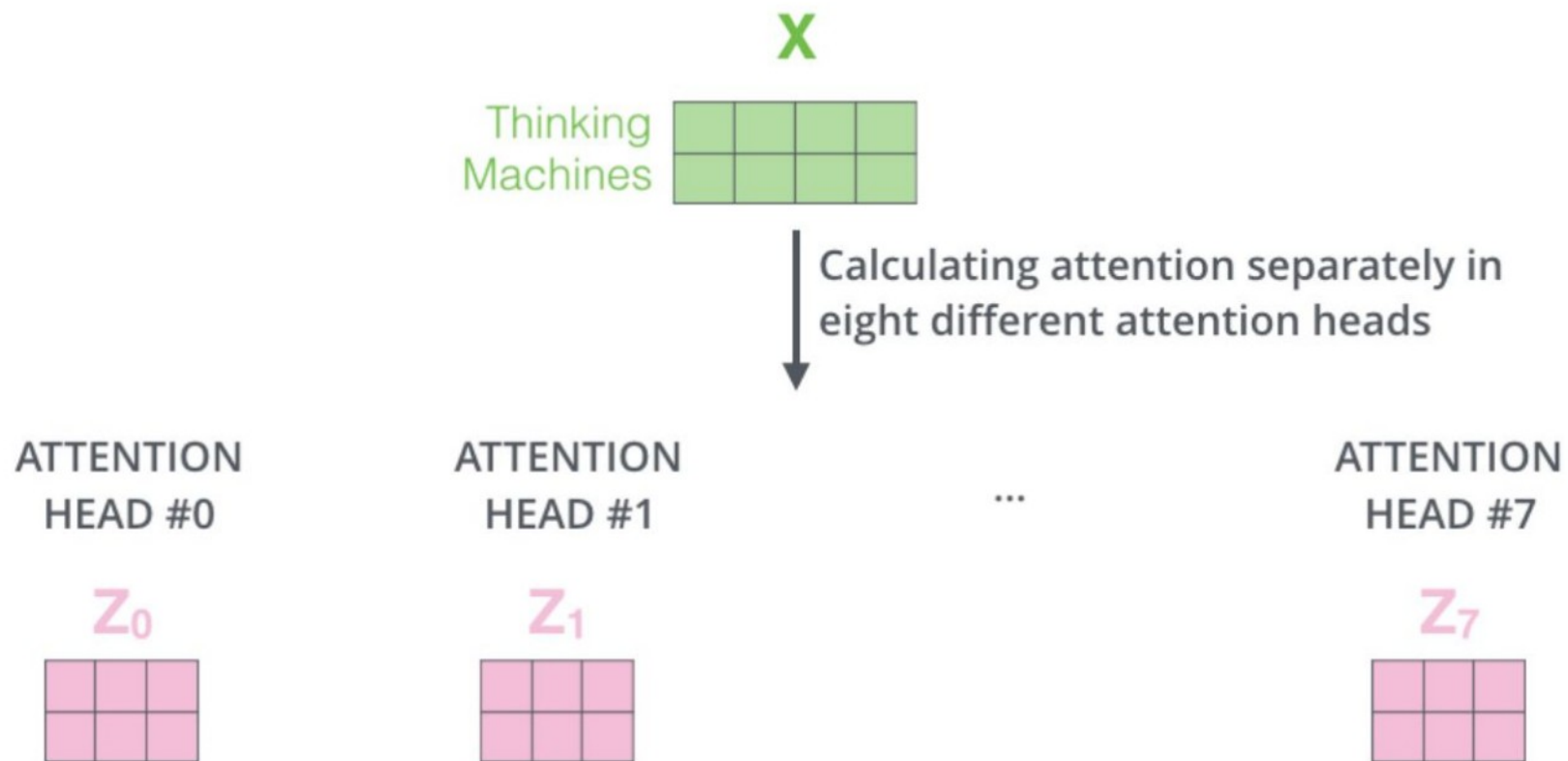
каждая голова MHSA ищет свои связи между словами

Multi-Head Attention



Transformer Self-Attention

Multi-Head Attention



Transformer Self-Attention

можно сделать так, чтобы размерность входа и выхода MHSE была одинаковая

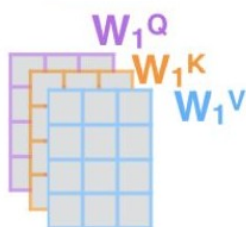
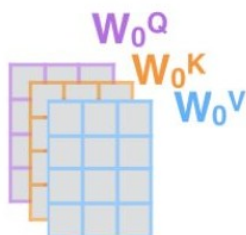
1) This is our input sentence*

Thinking
Machines

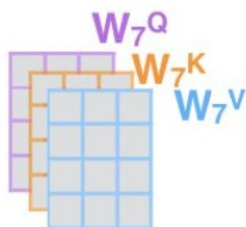
2) We embed each word*



3) Split into 8 heads.
We multiply X with weight matrices



...



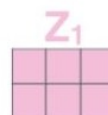
4) Calculate attention using the resulting $Q/K/V$ matrices



...



5) Concatenate the resulting Z matrices, then multiply with weight matrix W^O to produce the output of the layer



...

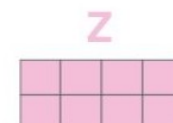
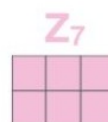
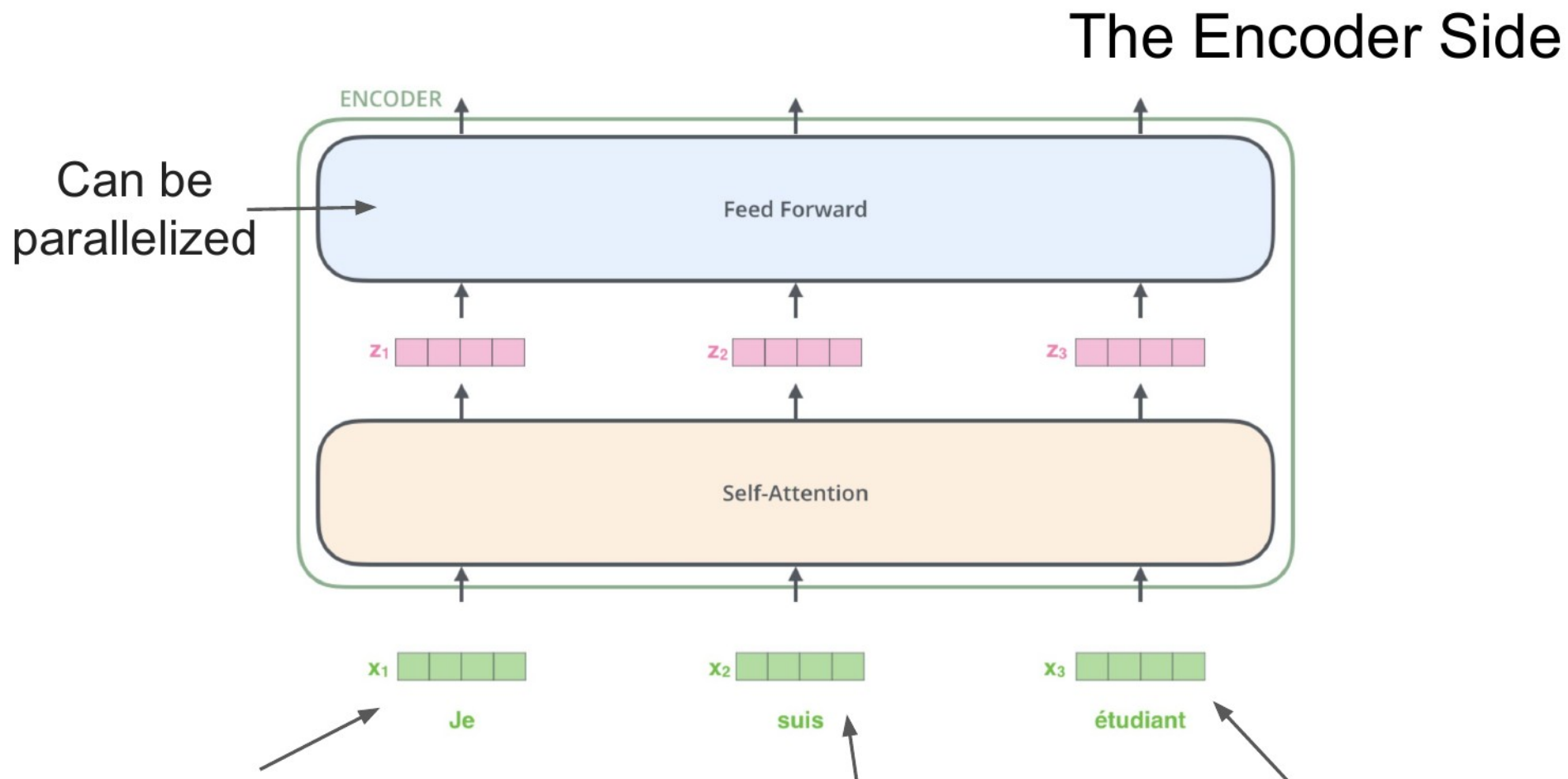


Image source: <https://jalammar.github.io/illustrated-transformer/>

Transformer Positional Encoding

Positional Encoding

Transformer Positional Encoding



the word in each position flows through its own path in the encoder

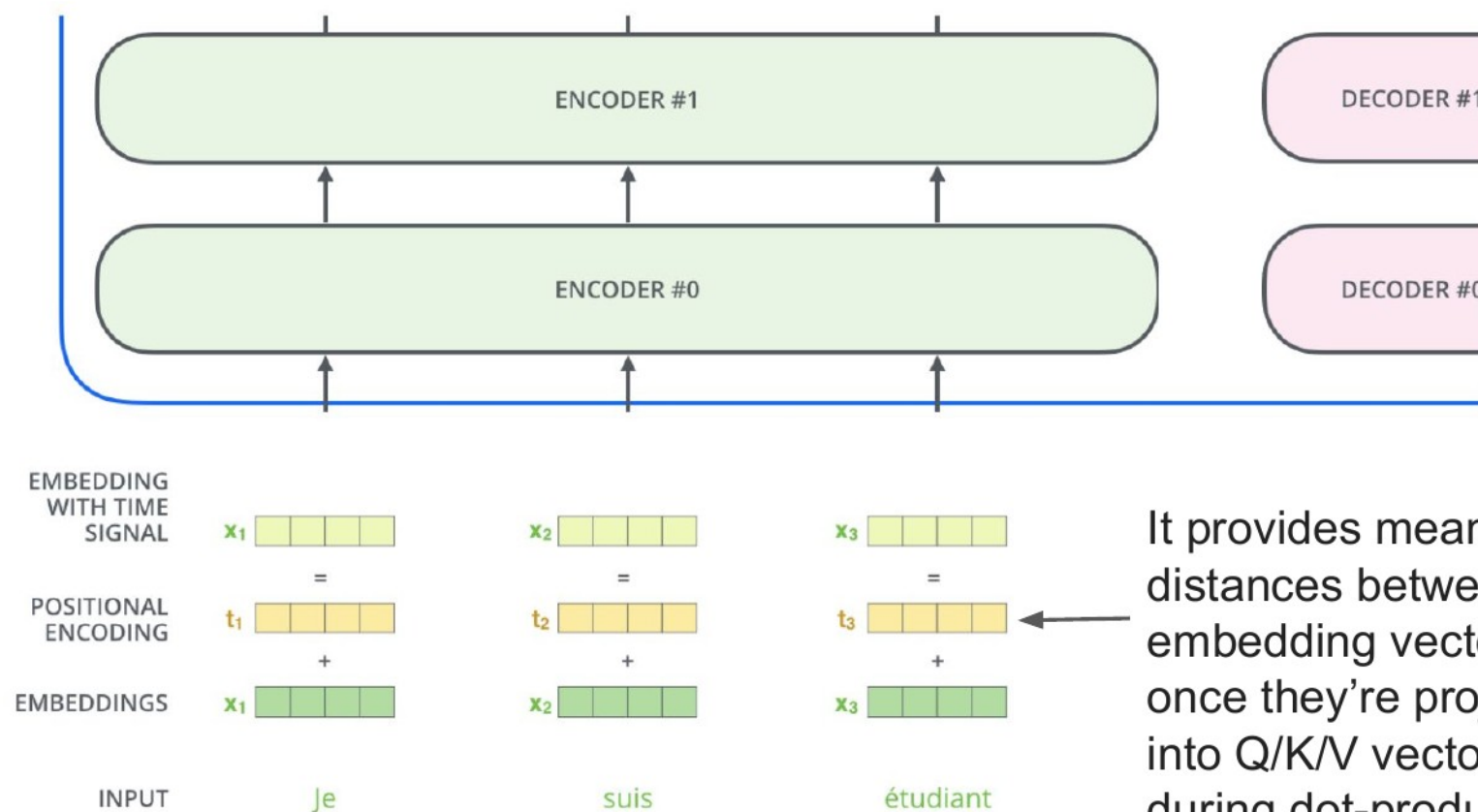
Transformer Positional Encoding

Необходимо обозначить позицию слова
выполняя условия

- уникальность для каждого слова
- не зависит от длины предложения
- детерминирован (не стохастический)

Transformer Positional Encoding

Positional Encoding



Transformer Positional Encoding

Positional Encoding: why sin and cos?

$$\vec{p}_t^{(i)} = f(t)^{(i)} = \begin{cases} \sin(\omega_k t), & \text{if } i = 2k \\ \cos(\omega_k t), & \text{if } i = 2k + 1 \end{cases}$$

$$\omega_k = \frac{1}{10000^{2k/d}}$$

t stays for position in the original sequence
k is the index of the element in the positional vector

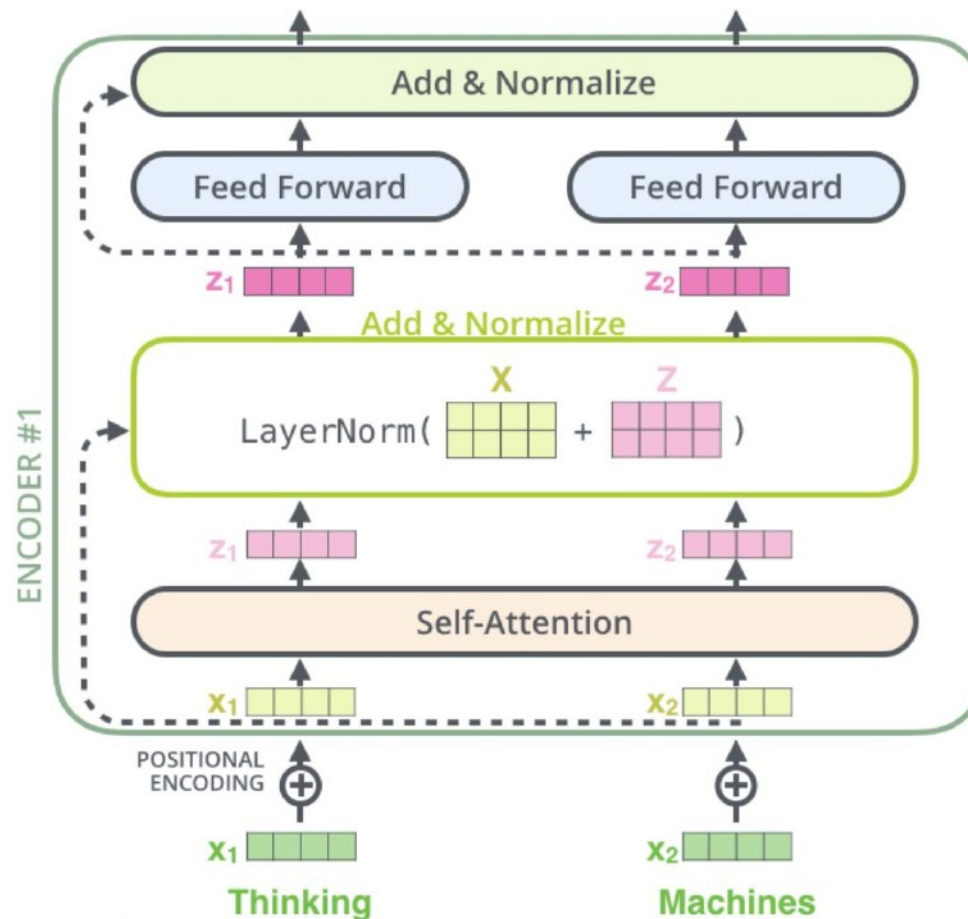
$$\vec{p}_t = \begin{bmatrix} \sin(\omega_1.t) \\ \cos(\omega_1.t) \\ \\ \sin(\omega_2.t) \\ \cos(\omega_2.t) \\ \\ \vdots \\ \\ \sin(\omega_{d/2}.t) \\ \cos(\omega_{d/2}.t) \end{bmatrix}_{d \times 1}$$

Transformer

Layer Normalization

Like BatchNorm

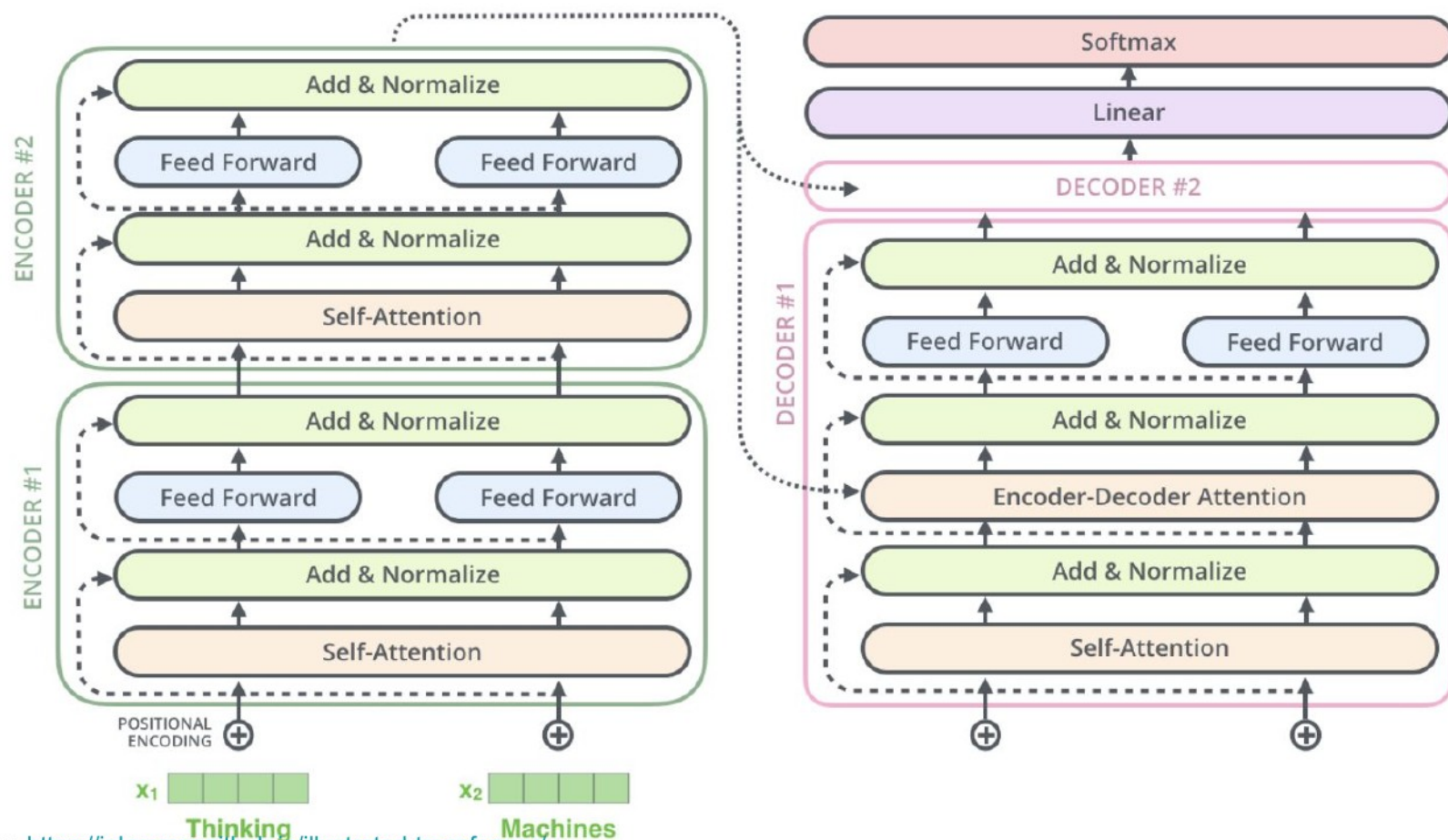
but normalize along
all features
representing latent
vector



More info:
[Layer Normalization](https://jalammar.github.io/illustrated-transformer/)

Transformer

можно сделать так, чтобы размерность входа и выхода MHSE была одинаковая и состыковать несколько Encoder



Transformer: литература

git clone https://github.com/mechanoid5/ml_nlp.git

Евгений Борисов Неросетевой транслятор текстов. Использование рекуррентных нейронных сетей для создания систем машинного перевода и чатботов.

<http://mechanoid.su/ml-chatbot.html>

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin
Attention Is All You Need

<https://arxiv.org/abs/1706.03762>

Радослав Нейчев Прикладное машинное обучение. 4. Self-Attention. Transformer overview.

https://www.youtube.com/watch?v=UETKUIIYE6g&list=PL4_hYwCyhAvY7k32D65q3xJV08X8dc3Ye&index=5