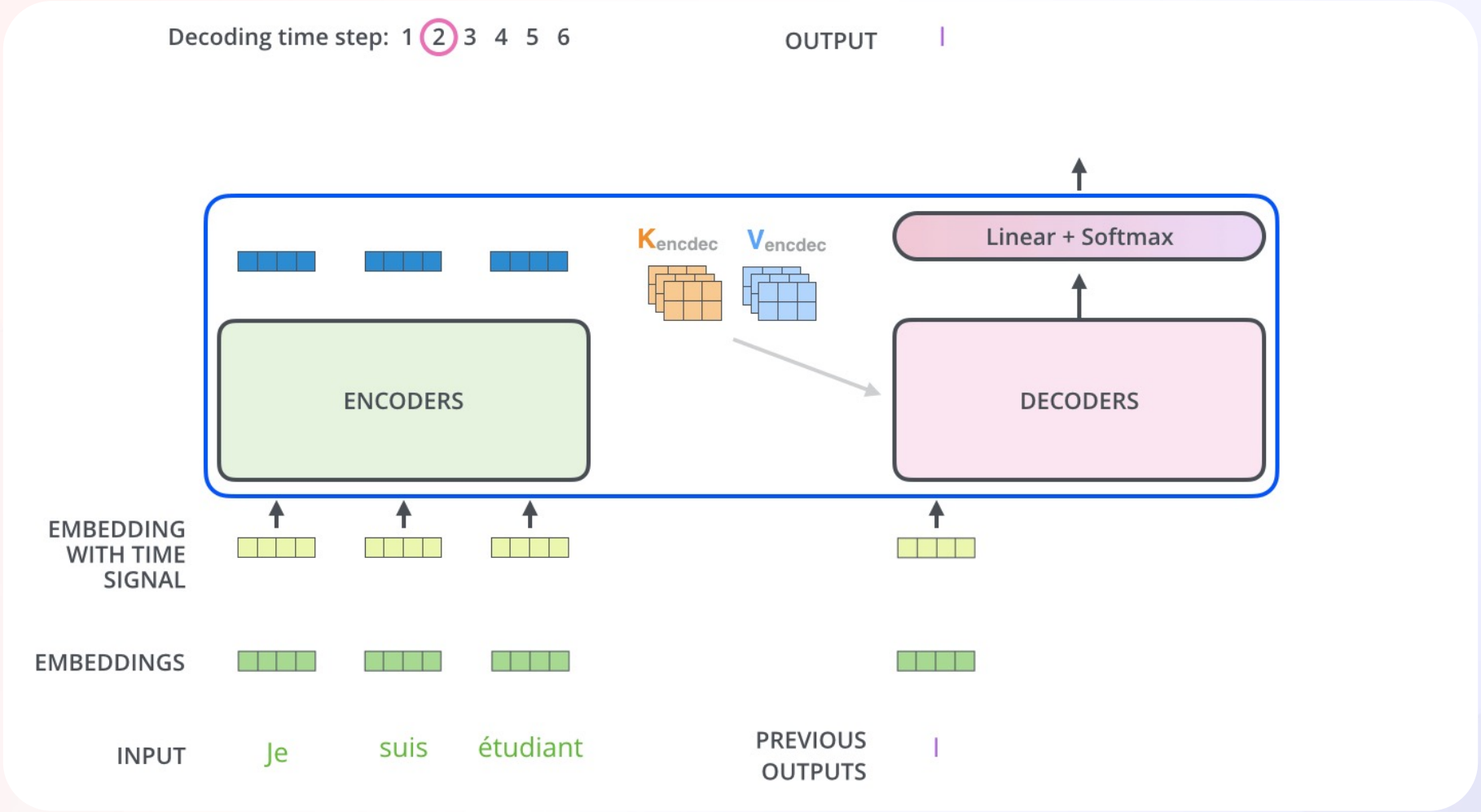
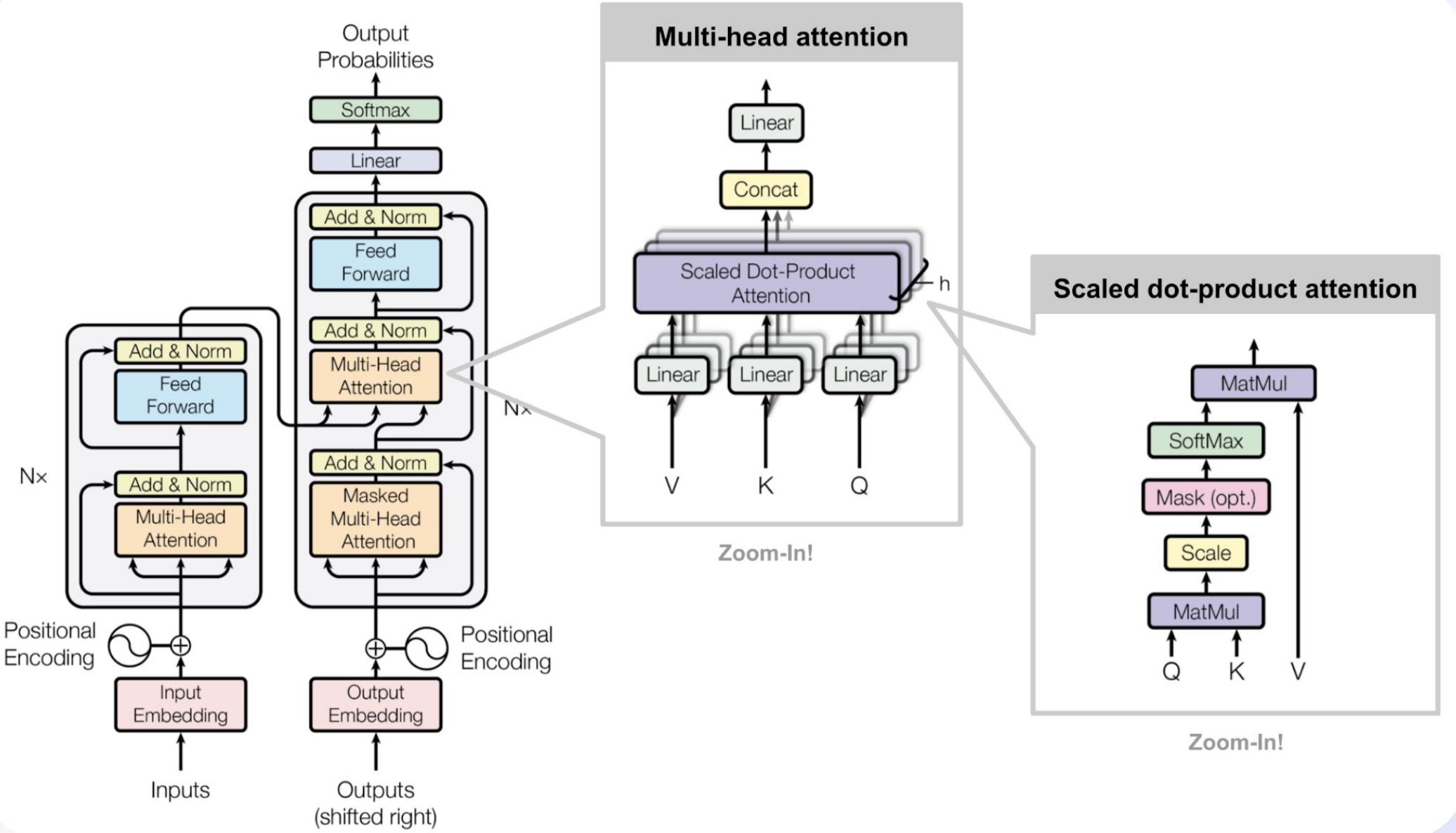
Проблемы трансформеров и их решения

Денис Козлов

Трансформер




Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)
Multi-head attention
Linear
Concat
Scaled Dot-Product Attention
h
Linear
Linear
Linear
V
K
Q
Zoom-In!
Scaled dot-product attention
MatMul
SoftMax
Mask (opt.)
Scale
MatMul
Q
K
V
Zoom-In!

Нотация

Входная последовательность длины L

Выходная последовательность длины $\dot{L}$

Эмбеддинги размера d

Входная последовательность $X \in \mathbb{R}^{L \times d}$

Матрицы весов $W^k, W^q \in \mathbb{R}^{d \times d_k}$; $W^v \in \mathbb{R}^{d \times d_v}$

Как правило $d_k = d_v = d$

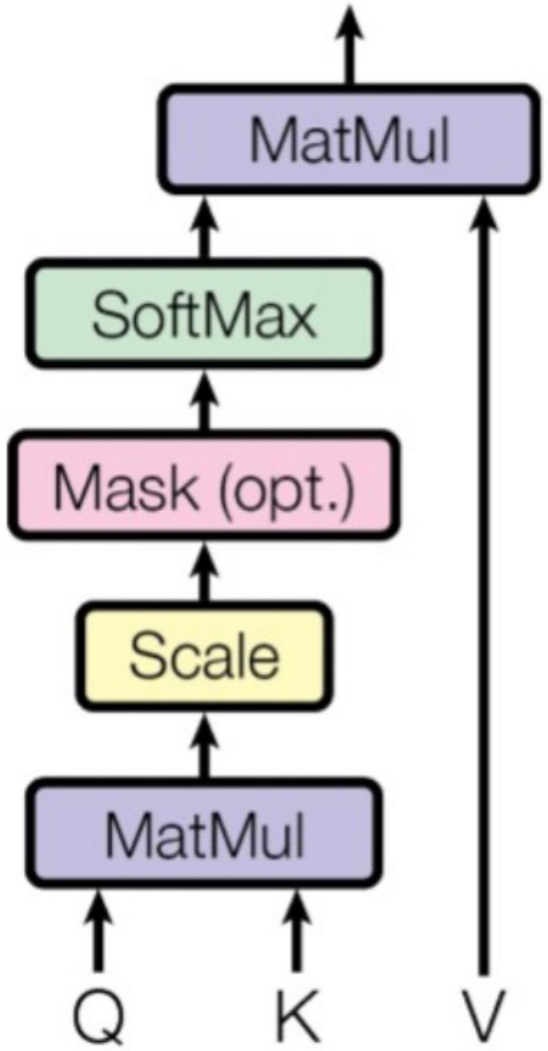
Attention

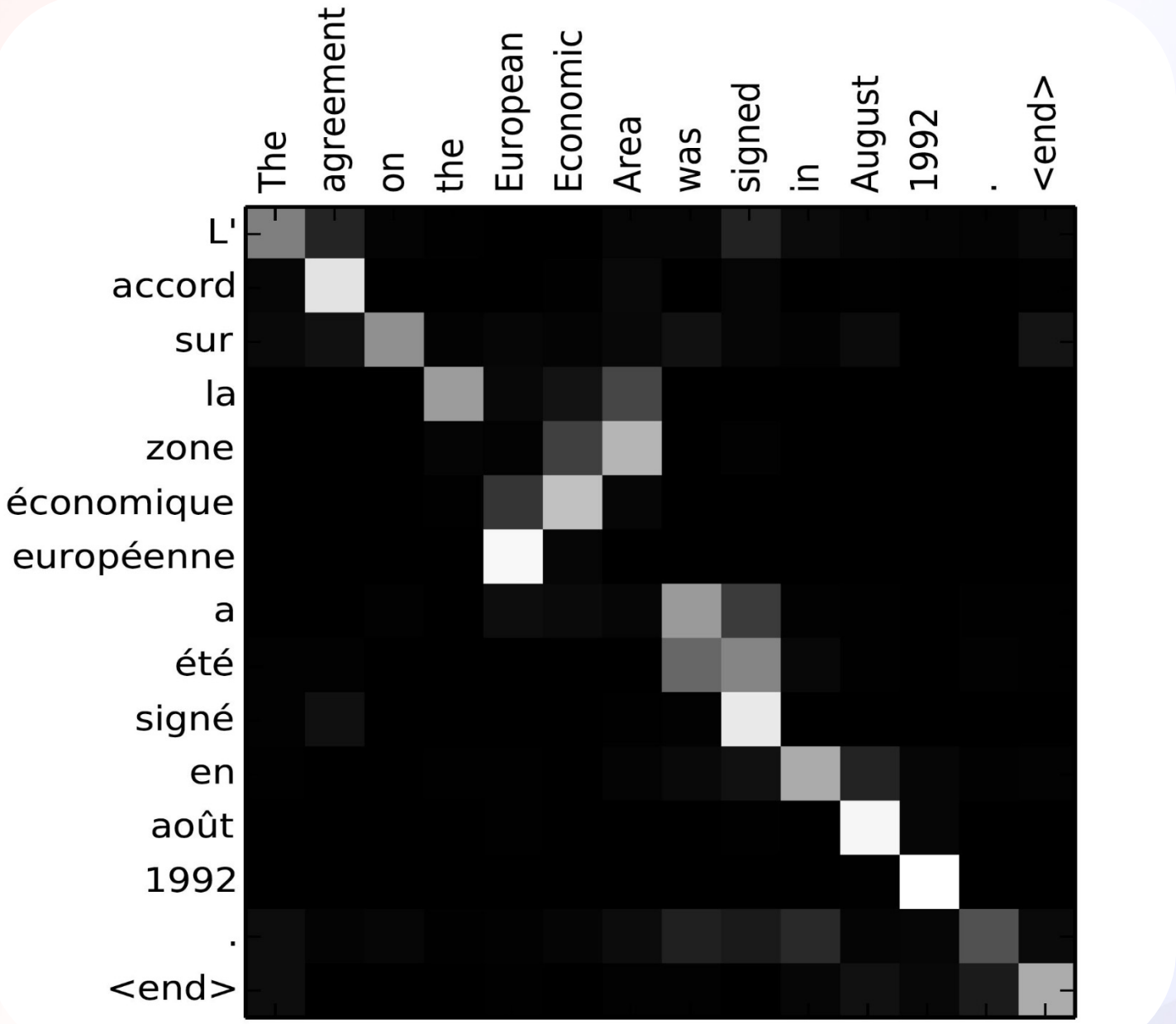
Query $Q = XW^q \in R^{L \times d_k}$

Key $K = XW^k \in R^{L \times d_k}$

Value $V = XW^v \in R^{L \times d_v}$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \in \mathbb{R}^{L \times d_v}$$




The
agreement
on
the
European
Economic
Area
was
signed
in
August
1992
.
<end>
L'
accord
sur
la
zone
économique
européenne
a
été
signé
en
août
1992
.
<end>

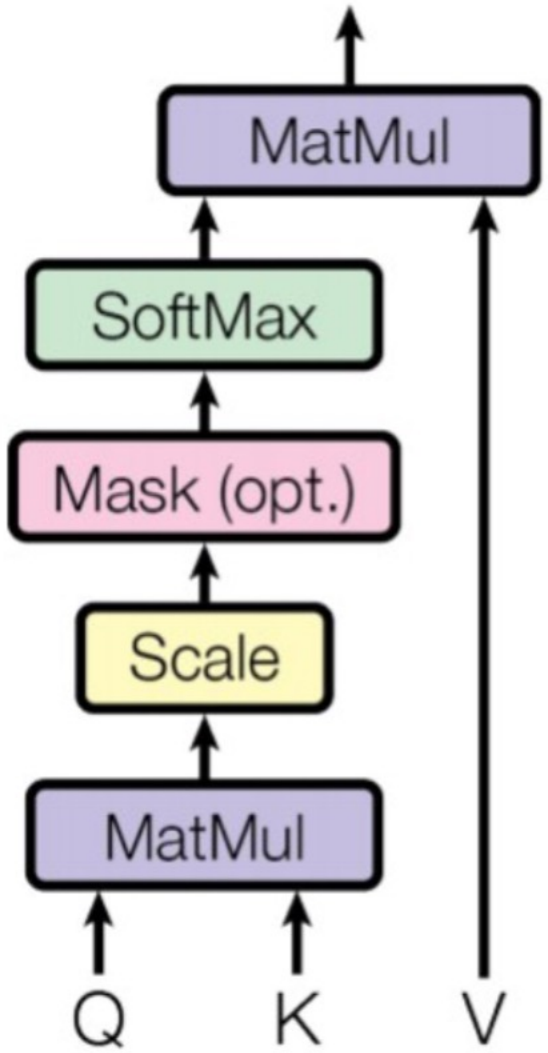
Attention

Query $Q = XW^q \in R^{L \times d_k}$

Key $K = XW^k \in R^{L \times d_k}$

Value $V = XW^v \in R^{L \times d_v}$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \in \mathbb{R}^{L \times d_v}$$



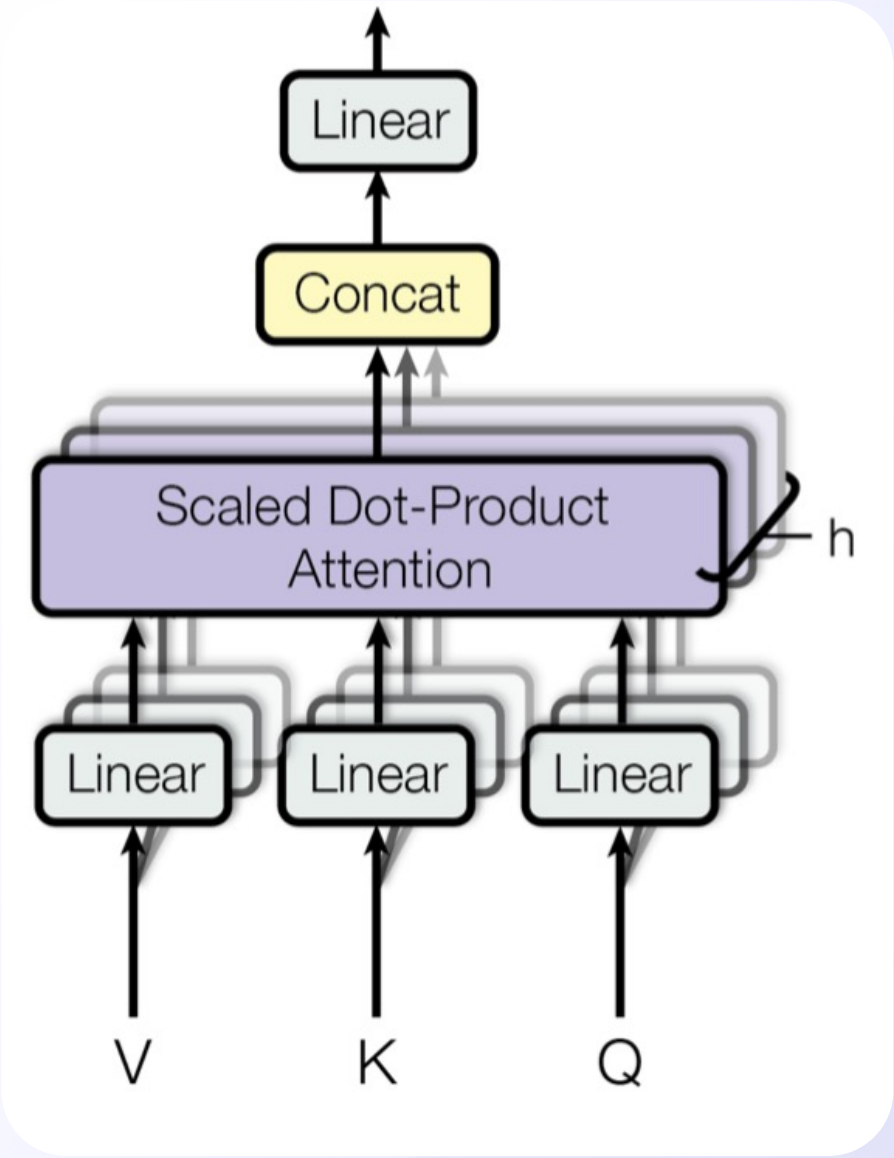
Multi-head attention

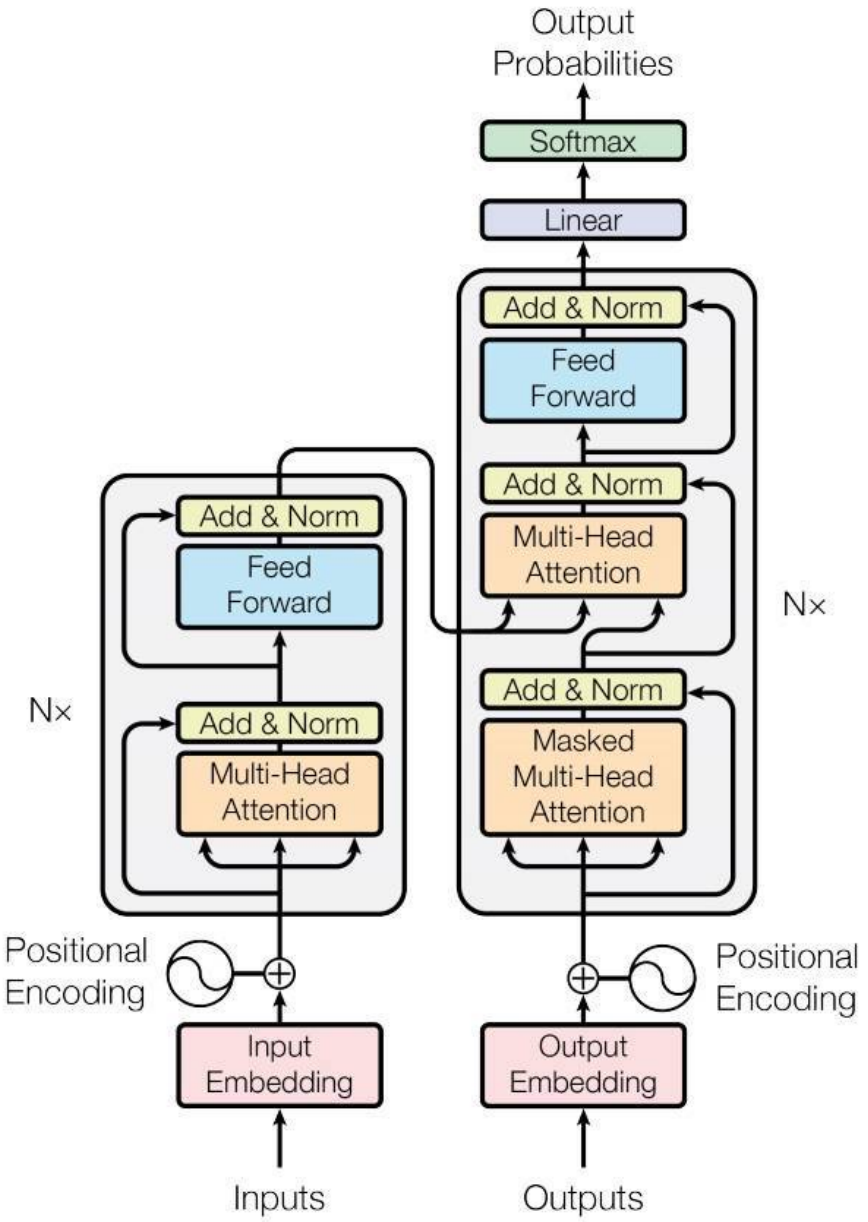
Количество голов h

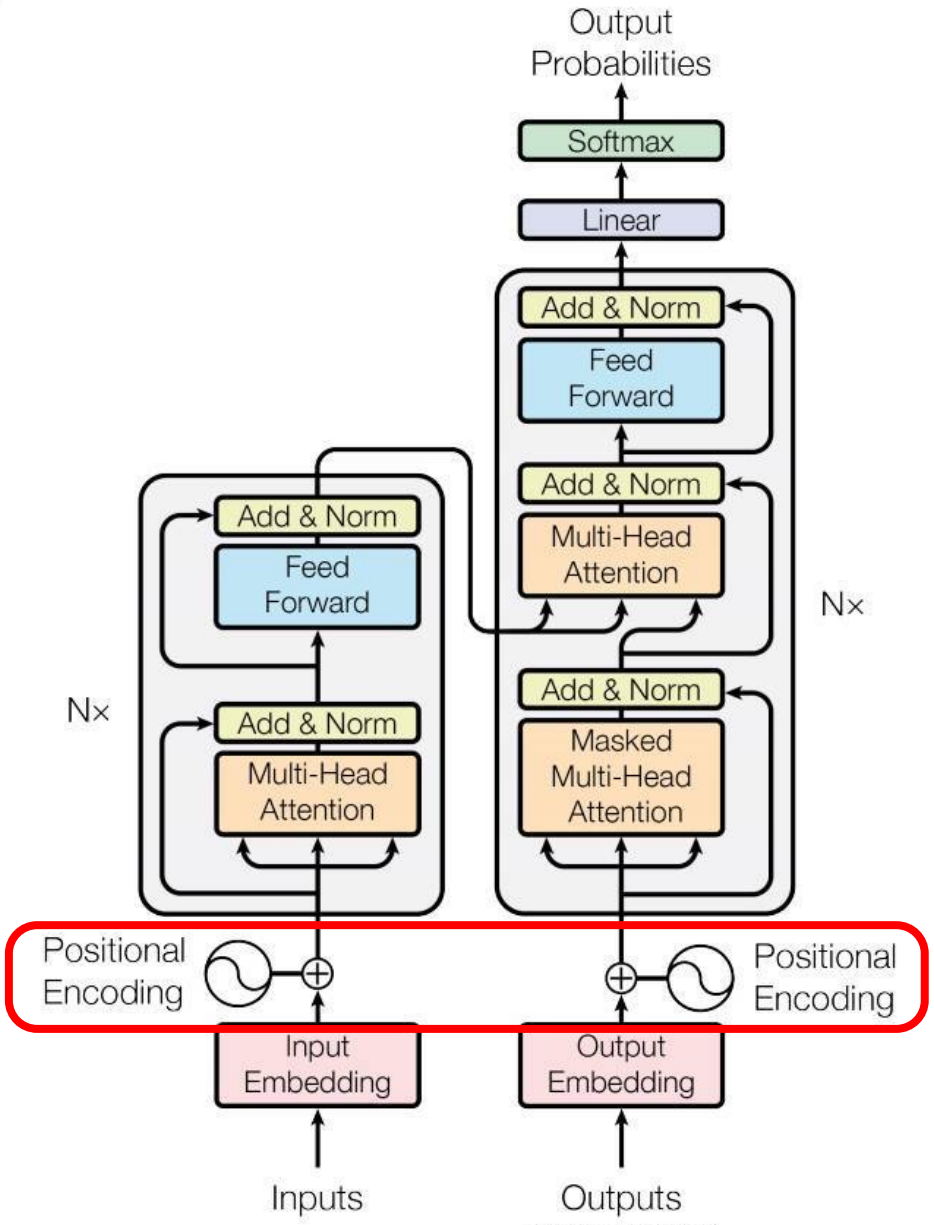
Матрица весов $W^o \in \mathbb{R}^{d_v \times d}$

$$\text{MultiHeadAttention}(X) = \\ = \text{concat}(head_1; \dots; head_h)W^o \\ \in \mathbb{R}^{L \times d}$$

$$head_i = \text{Attention}(\text{X}W_i^q, XW_i^k, XW_i^v)$$




Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs


Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
N×
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
N×
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs

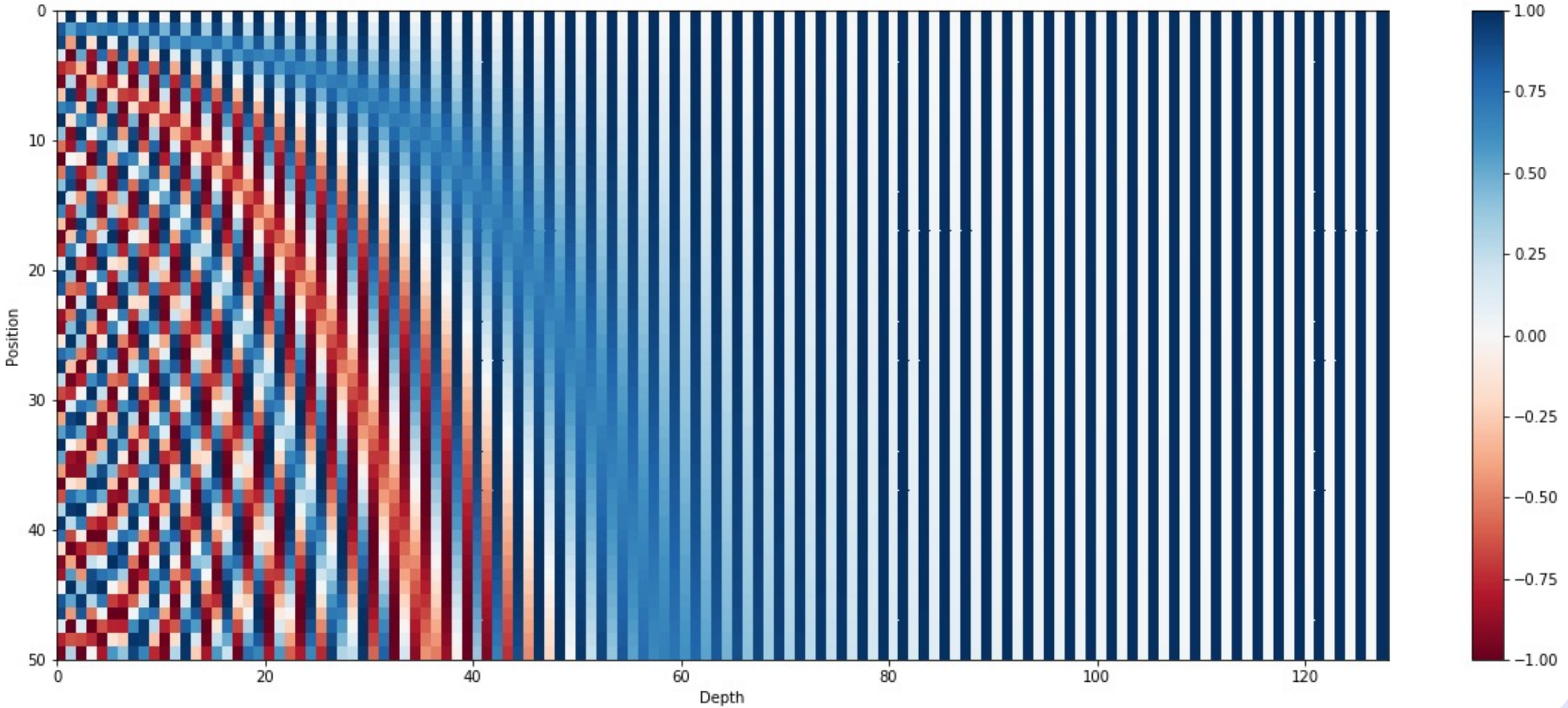
Positional encoding

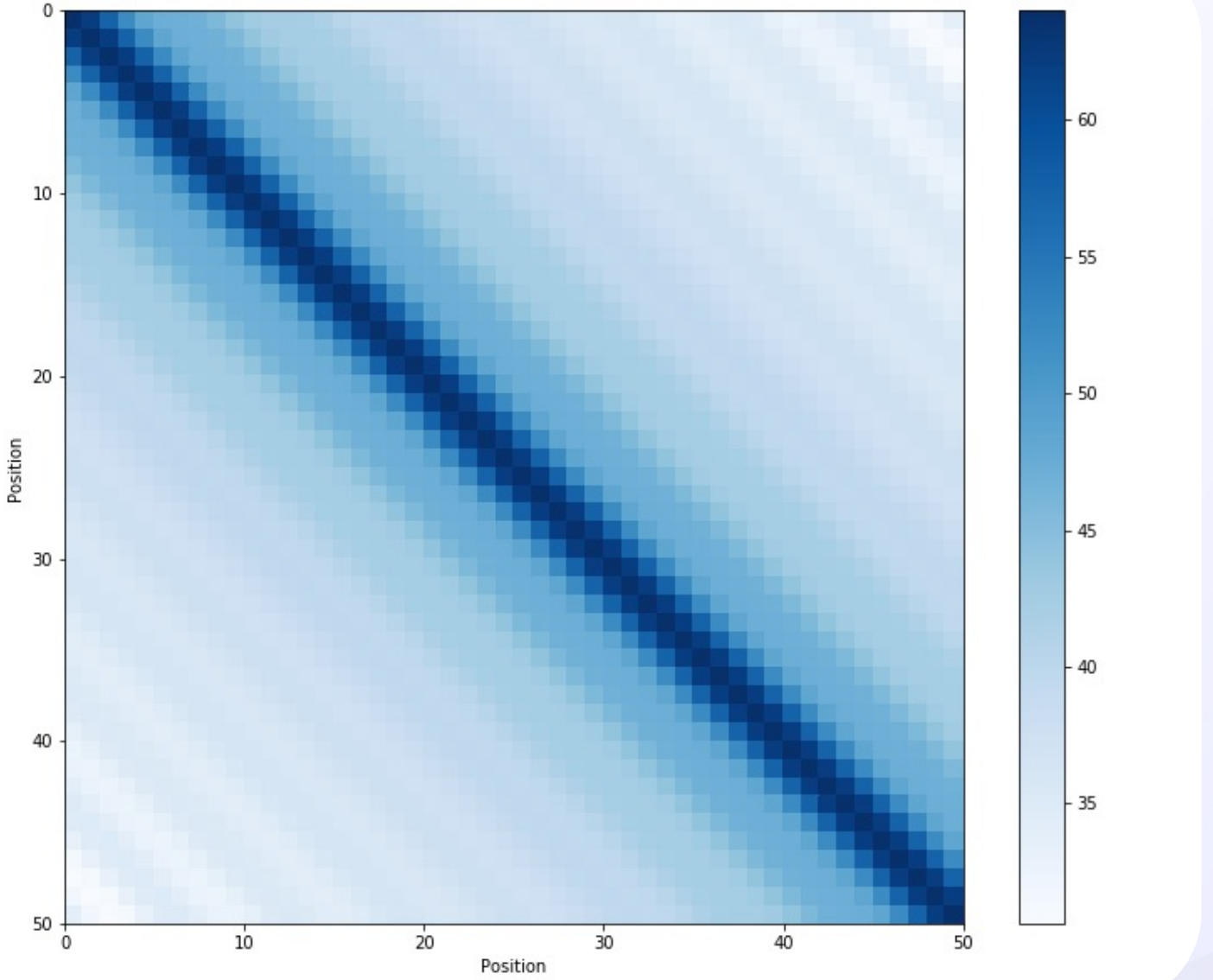
0000
0001
0010
...
1111

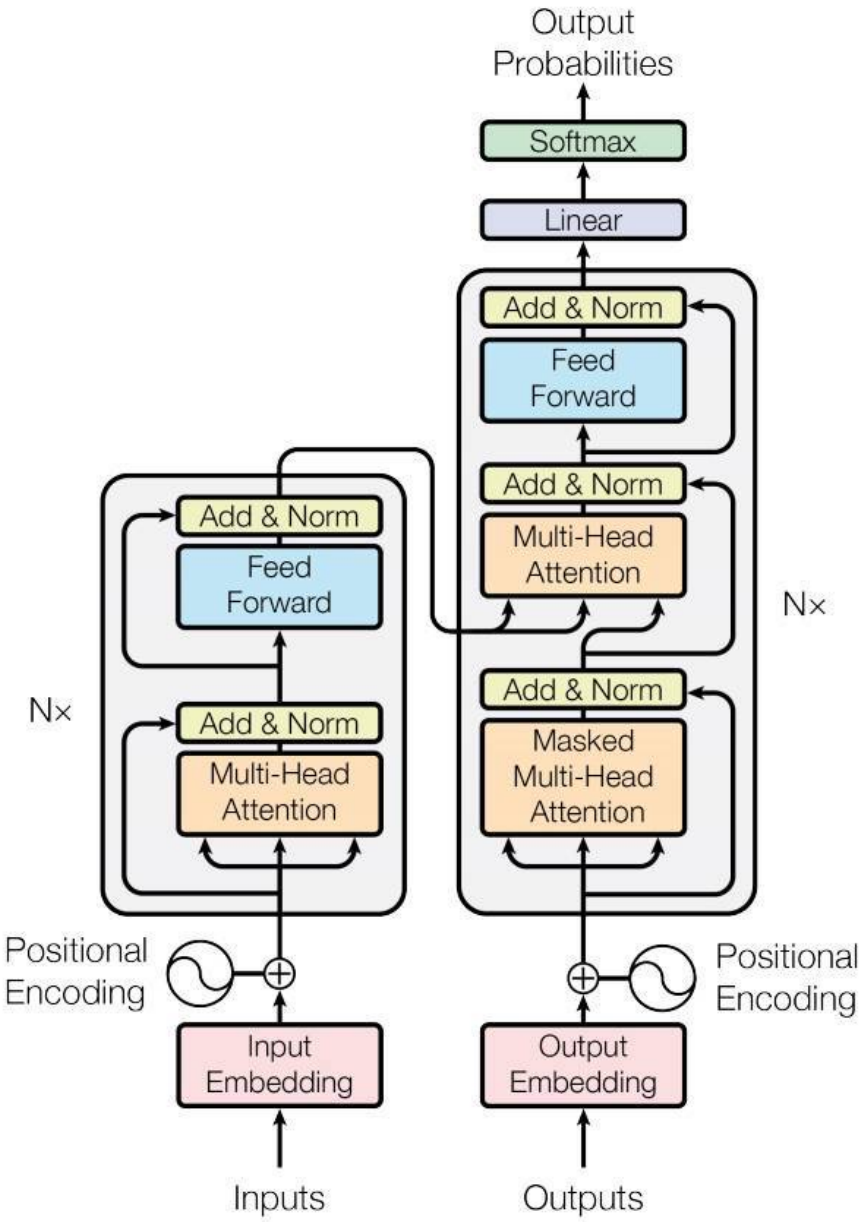
Но у нас же float!

$$\overrightarrow{p_t}^{(i)} = f(t)^{(i)} := \begin{cases} \sin(\omega_k . t), & \text{if } i = 2k \\ \cos(\omega_k . t), & \text{if } i = 2k+1 \end{cases}$$

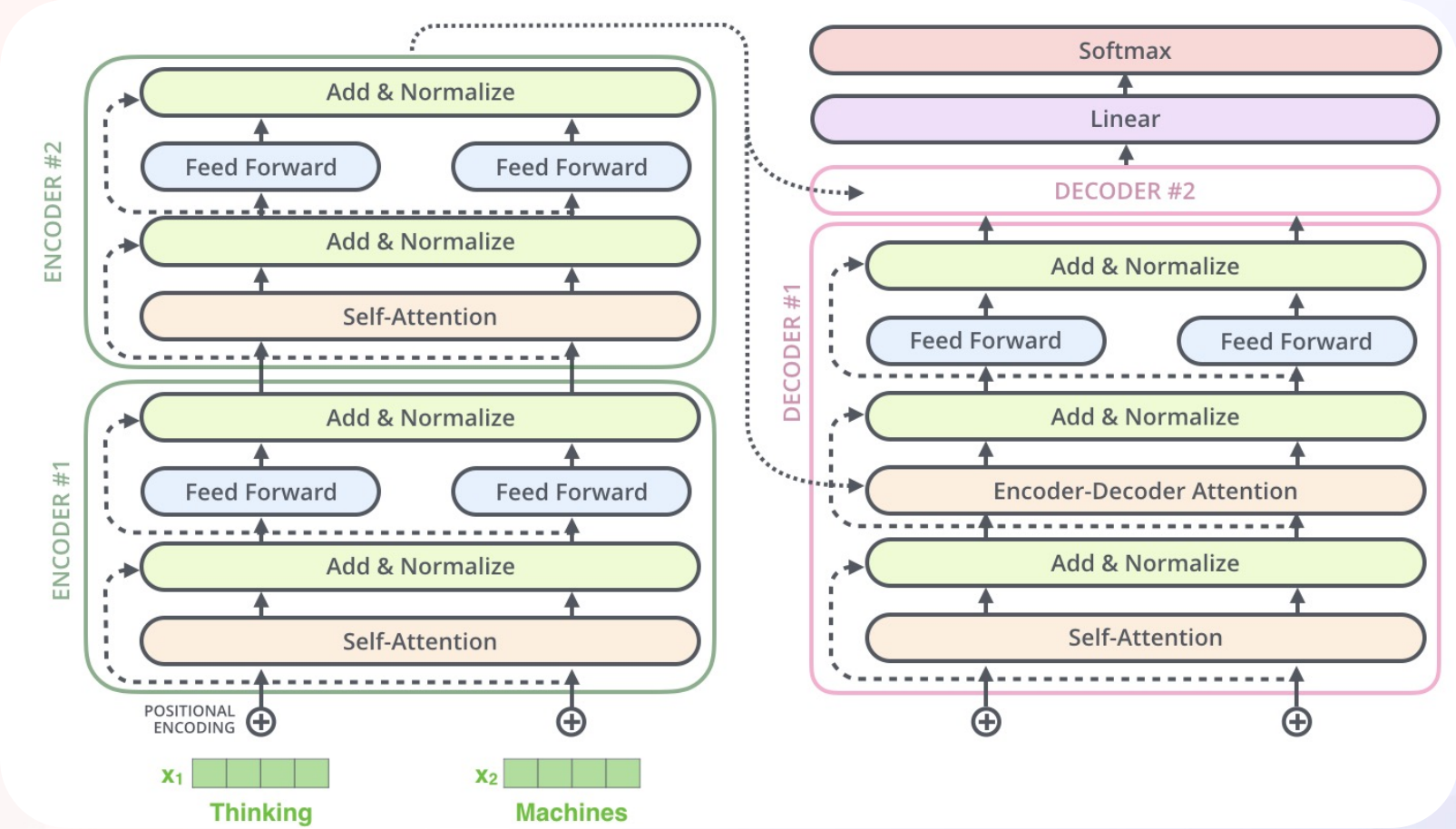
$$\omega_k = \frac{1}{10000^{2k/d}}$$


Position
Depth
0
10
20
30
40
50
0
20
40
60
80
100
120
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


0
10
20
30
40
50
Position
0
10
20
30
40
50
Position
60
55
50
45
40
35


Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs

Много уровней



Все ли хорошо?

Все ли хорошо?

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

Query $Q = XW^q \in R^{L \times d_k}$

Key $K = XW^k \in R^{L \times d_k}$

Value $V = XW^v \in R^{L \times d_v}$

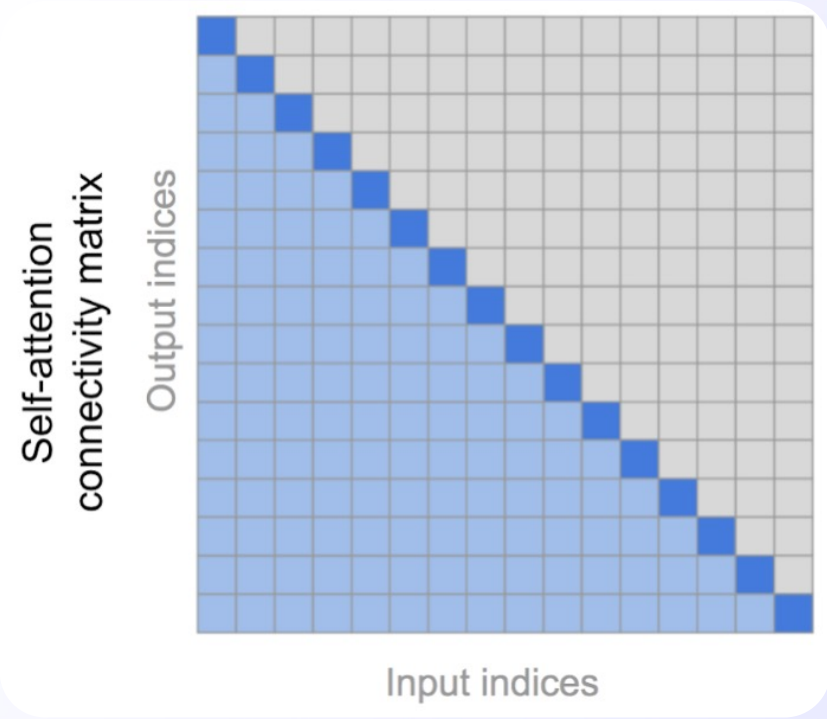
Требуется $O(L^2)$ памяти!

Для простоты примем $n = L$

Тяжелый attention

Decoder-only Transformer (GPT)

Для предсказания смотрим
на все предыдущие элементы
последовательности



Тяжелый attention

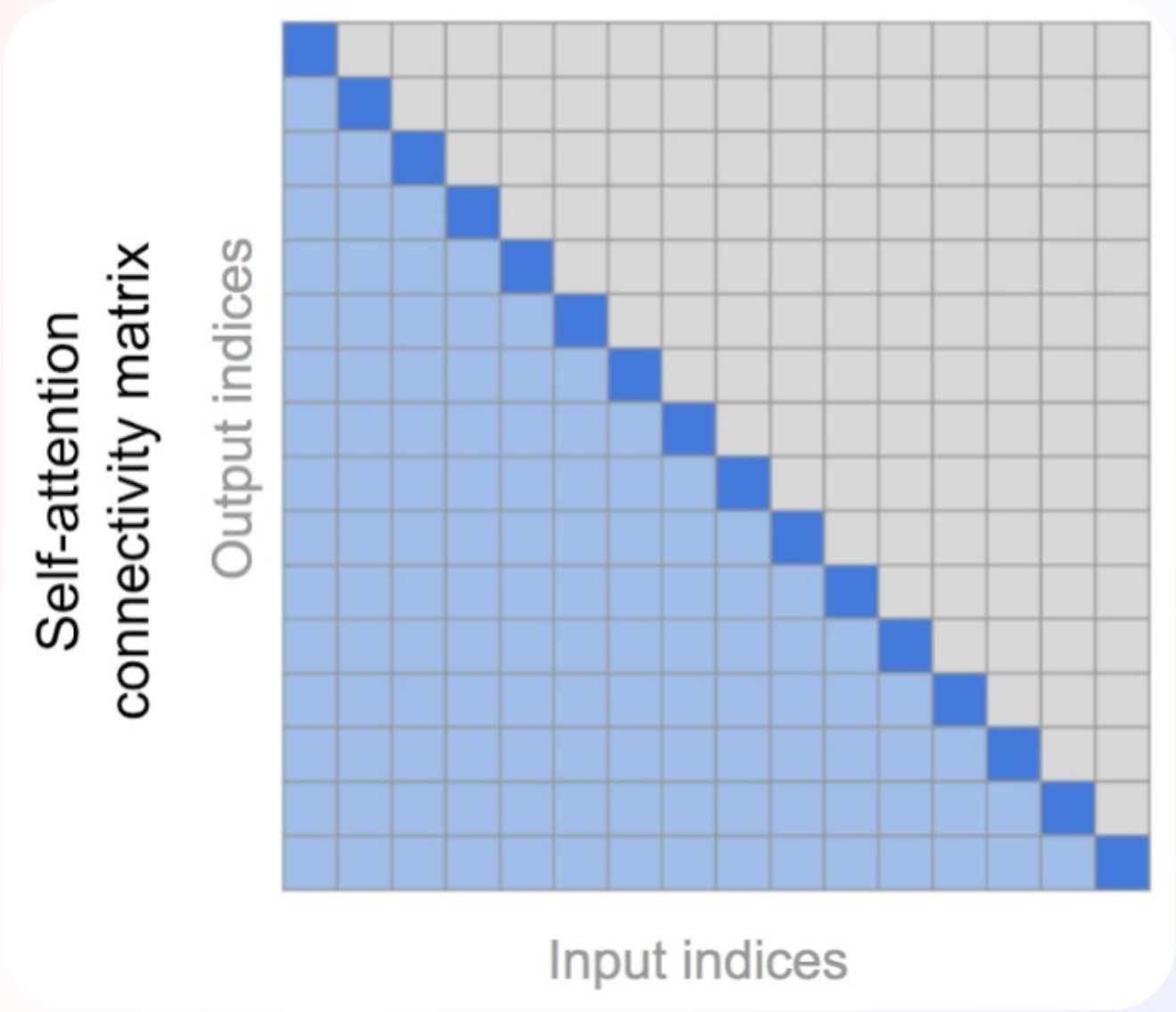
$$\text{Attend}(X) = \text{softmax}\left(\frac{(XW^q)(XW^k)^T}{\sqrt{d_k}} * \text{mask}\right)XW^v$$

$$\text{Attend}(\mathbf{X}, \mathcal{S}) = \left(a(\mathbf{x}_i, S_i)\right)_{i\in\{1,\ldots,L\}}$$

$$\mathcal{S} = \{S_1, \ldots, S_n\}$$

$$\text{where } a(\mathbf{x}_i, S_i) = \text{softmax}\left(\frac{(\mathbf{x}_i\mathbf{W}^q)(\mathbf{x}_j\mathbf{W}^k)^\top_{j\in S_i}}{\sqrt{d_k}}\right)(\mathbf{x}_j\mathbf{W}^v)_{j\in S_i}$$

Пока что $S_i = \{j | j \leq i\}$

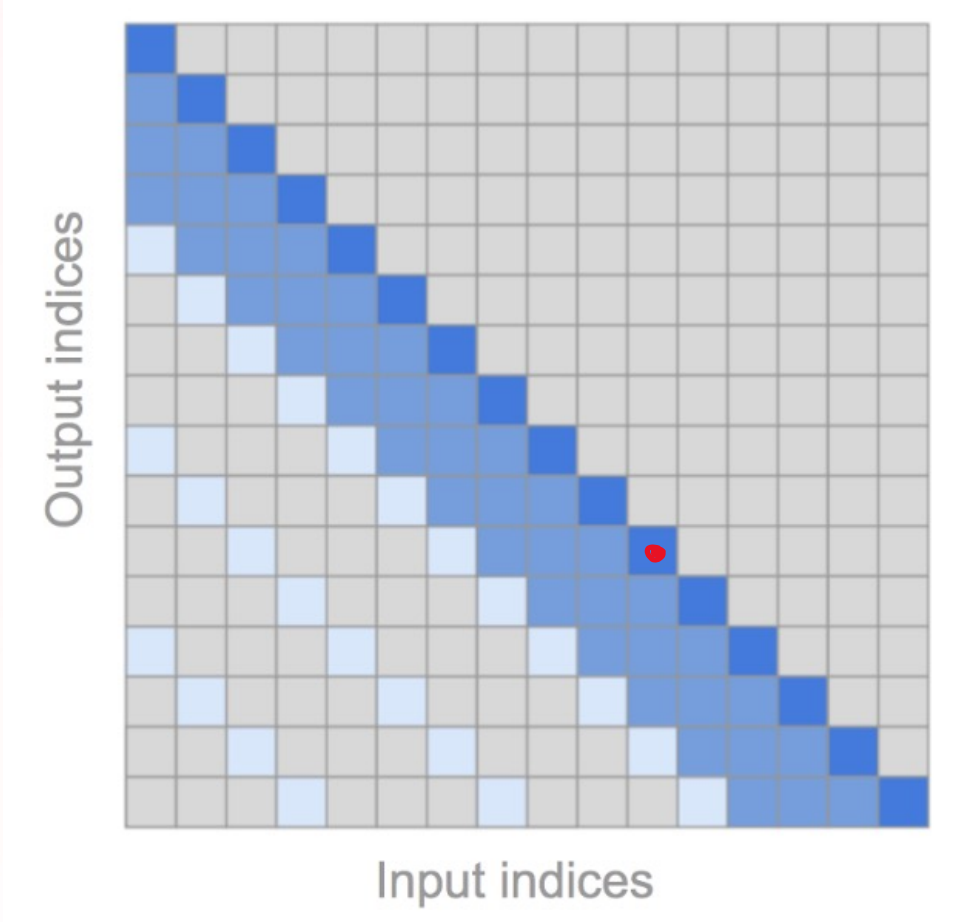

Self-attention
connectivity matrix
Output indices
Input indices

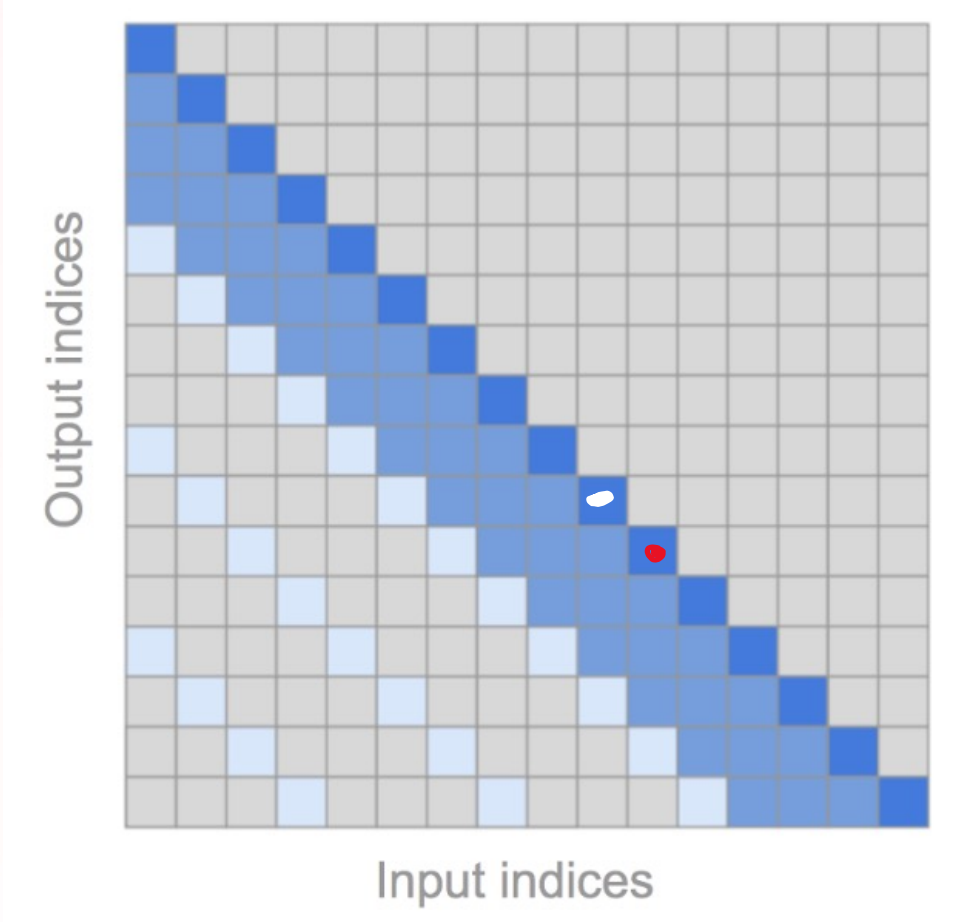
Почти Factorized self-attention

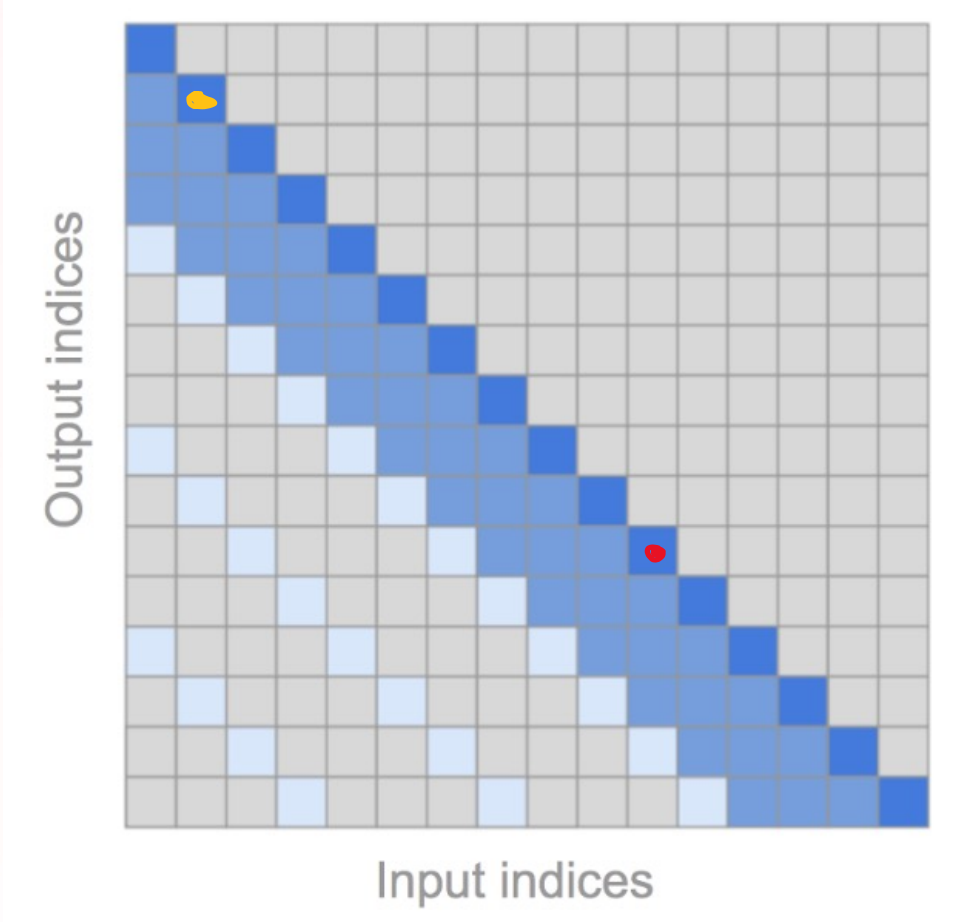
$S_i \subseteq \{j | j < i\}$, и верно:
для каждой пары $j \leq i$ существует «путь» от i до j

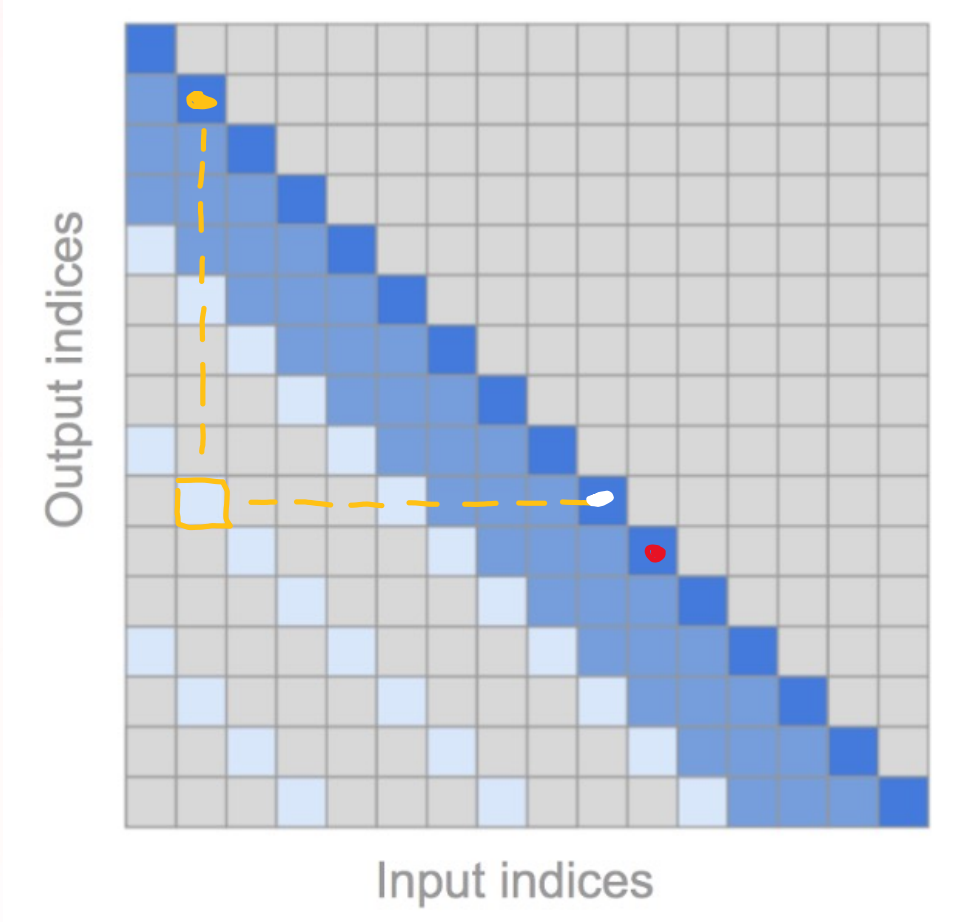
К примеру:

- $j \in S_i$ — *напрямую*
- $j \in S_{t_1}, t_1 \in S_{t_2}, \ldots, t_k \in S_i$ — *есть маршрут*


Output indices
Input indices


Output indices
Input indices


Output indices
Input indices


Output indices
Input indices

Почти Factorized self-attention

$S_i \subseteq \{j | j < i\}$, и верно:
для каждой пары $j \leq i$ существует «путь» от i до j

К примеру:

- $j \in S_i$ — *напрямую*
- $j \in S_{t_1}, t_1 \in S_{t_2}, \ldots, t_k \in S_i$ — *есть маршрут*

Factorized self-attention

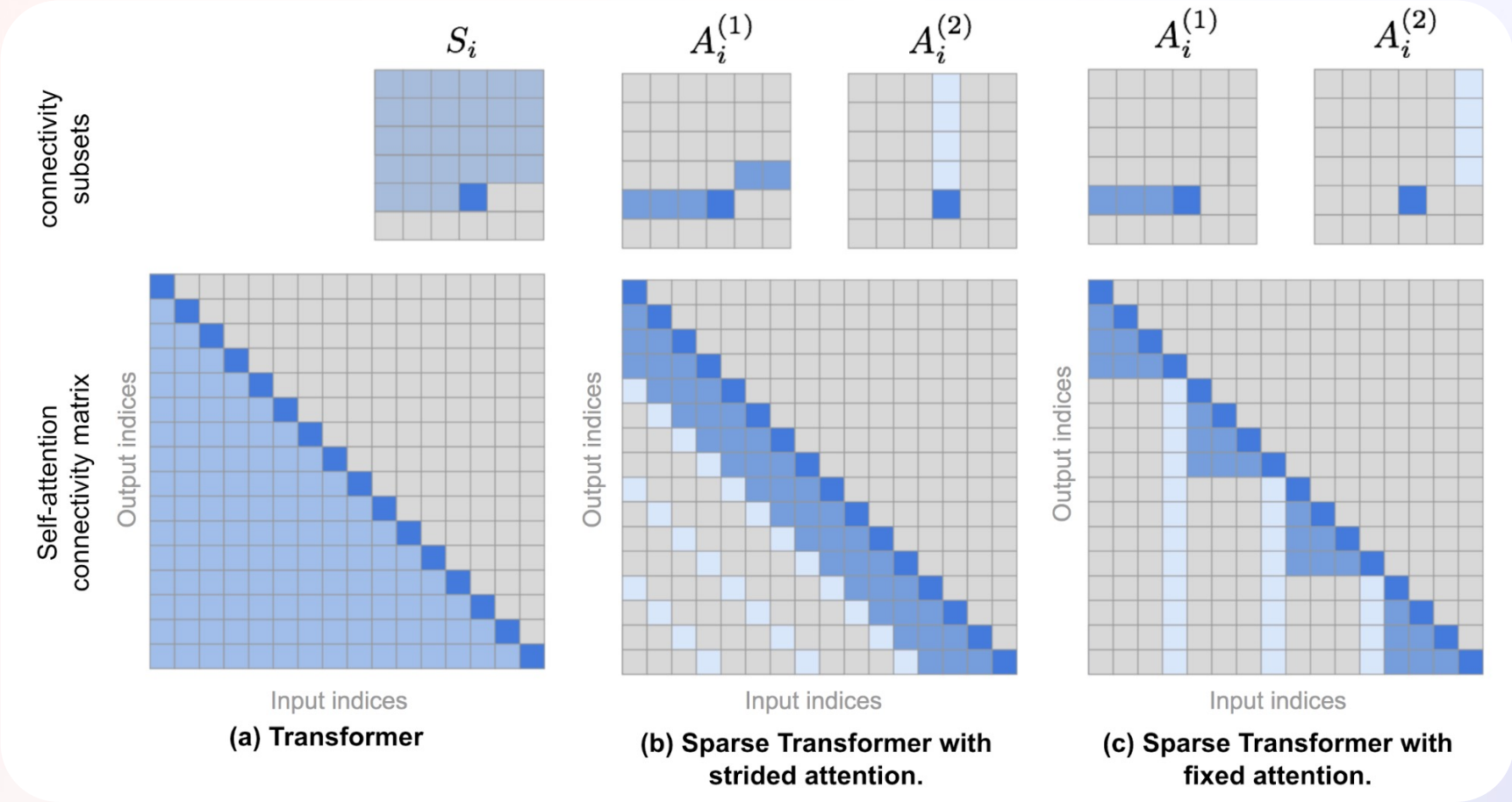
$S_i \subseteq \{j | j < i\}$ делится на p непересекающихся множеств $A_i^1, \ldots, A_i^p$ и верно, что:
для каждой пары $j \leq i$ существует «путь» от i до j

К примеру:

- $j \in S_i$ — *напрямую*
- $j \in A_{t_1}^1, t_1 \in A_{t_2}^2, \ldots, t_k \in A_i$
  есть маршрут $(j, t_1, t_2, \ldots, t_k, i)$ *не длиннее* $p + 1$

Посмотрим $p = 2$

Factorized self-attention

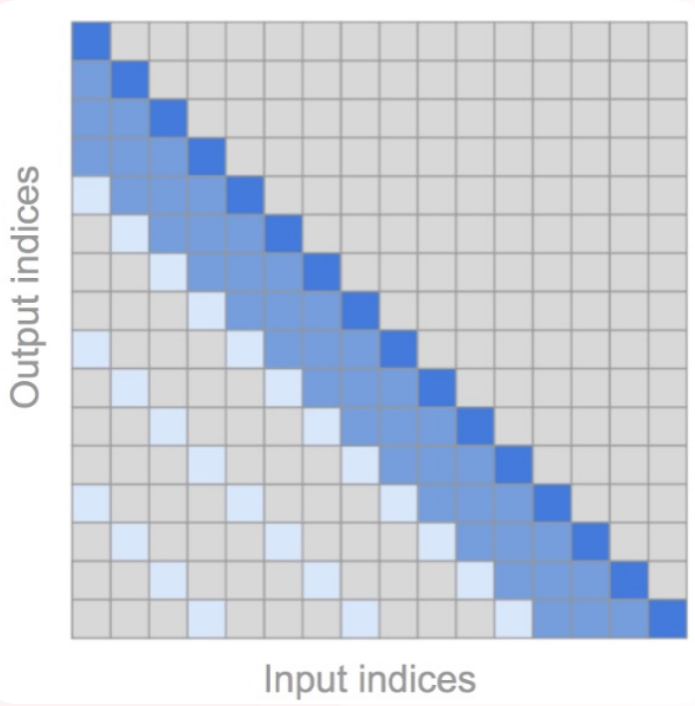


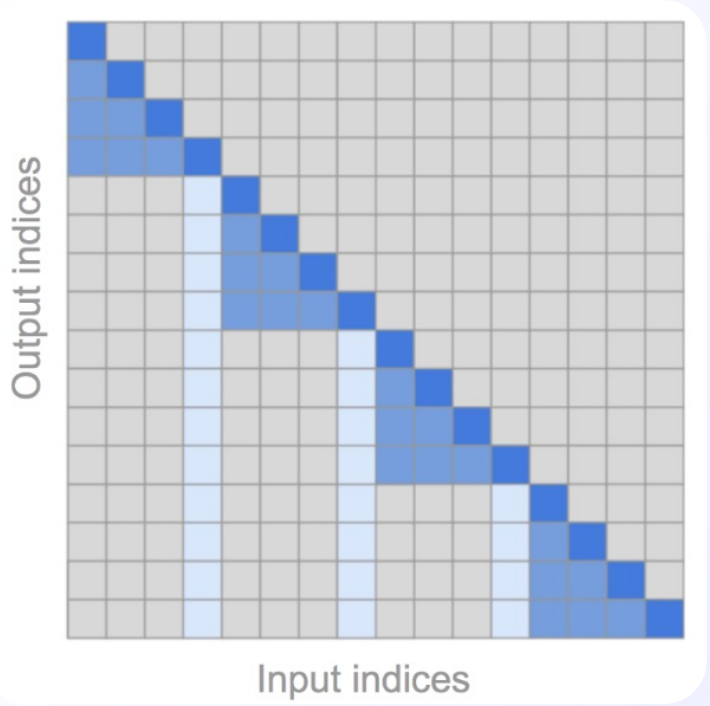
(a) Transformer

(b) Sparse Transformer with strided attention.

(c) Sparse Transformer with fixed attention.

Factorized self-attention





$$A_i^{(1)} = \{t, t+1, \ldots, i\}, \text{ where } t = \max(0, i - \ell)$$
$$A_i^{(2)} = \{j : (i - j) \mod \ell = 0\}$$

$$\ell \sim \sqrt{n}$$

$$A_i^{(1)} = \{j : \lfloor \frac{j}{\ell} \rfloor = \lfloor \frac{i}{\ell} \rfloor\}$$
$$A_i^{(2)} = \{j : j \mod \ell \in \{\ell - c, \ldots, \ell - 1\}\}$$

Factorized self-attention

- $\mathrm{Attention(X)} = \mathrm{Attend}\left(X, A^{n\%p}\right)W^o$
- $\mathrm{Attention(X)} = \mathrm{Attend}\left(X, \bigcup_{m=1}^{p} A^m\right)W^o$

Результаты