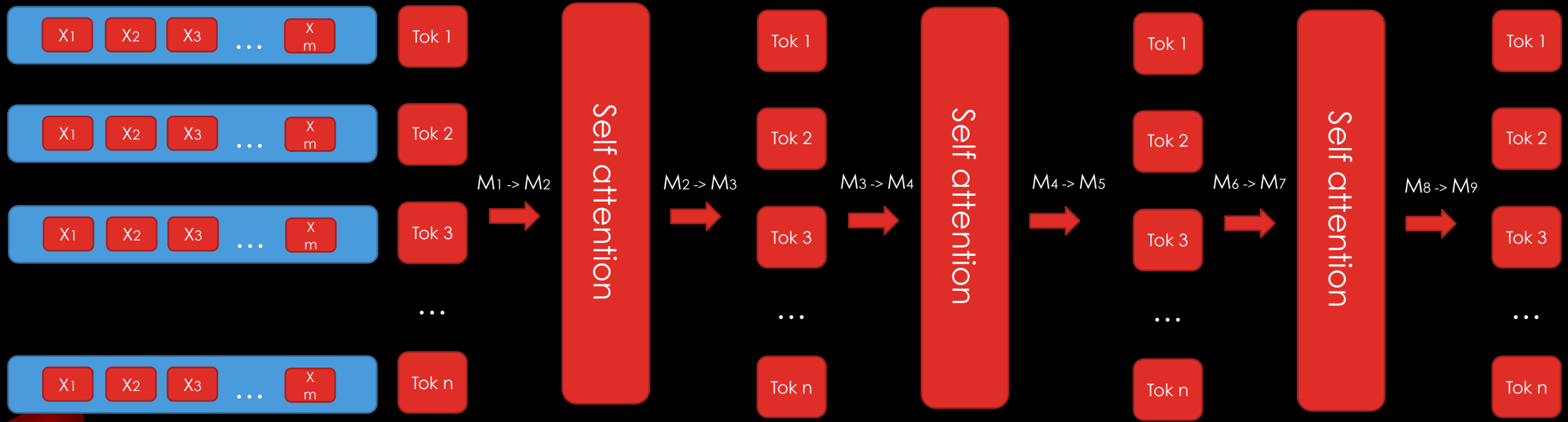
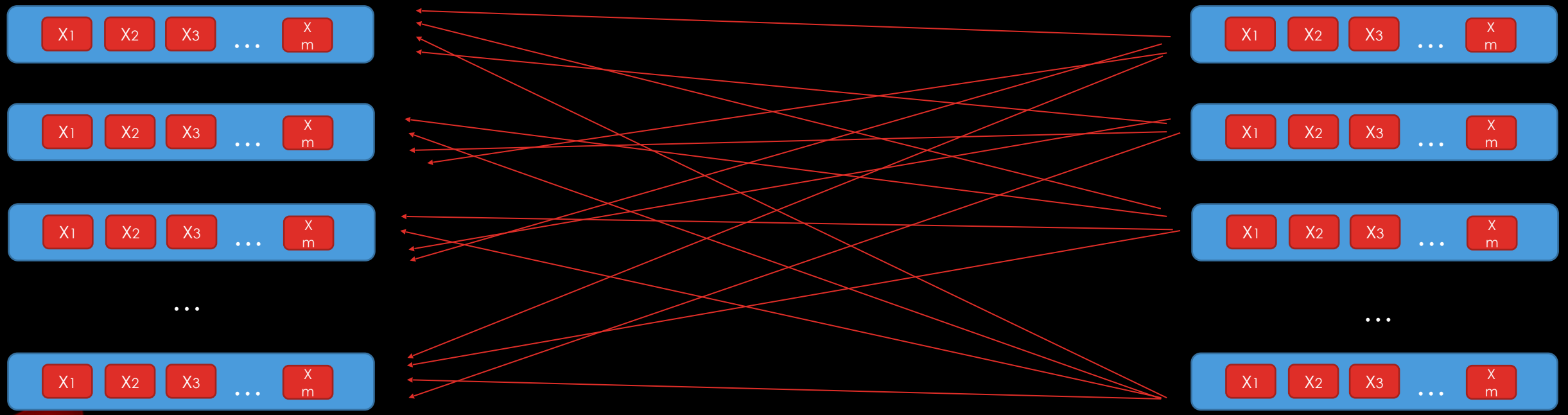
TRANSFORMERS ARE RNNS

(Не совсем)

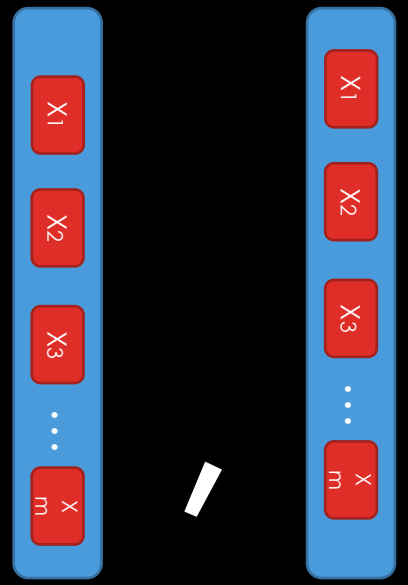
TRANSFORMER



SELF ATTENTION



SELF ATTENTION

Sim (

)

Sim ( , ) = 10

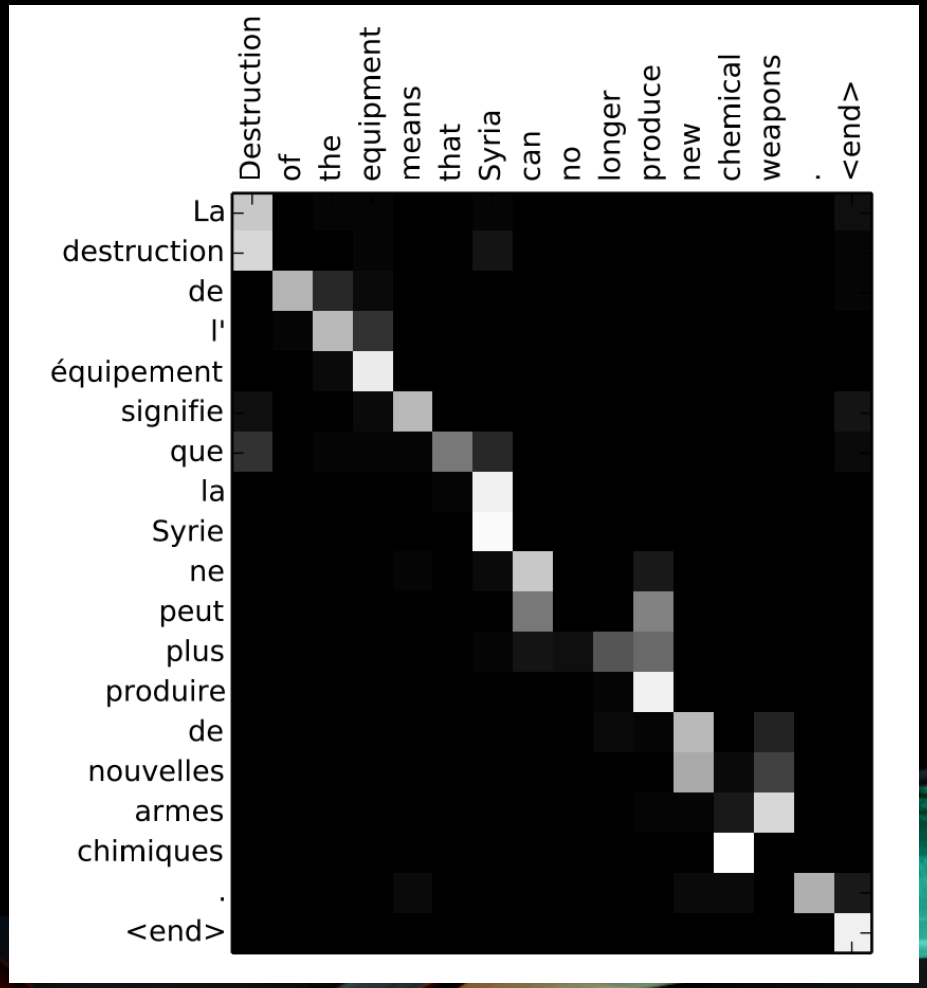
Sim ( , ) = -34

$$Sim(a,\ b) = a^T b$$

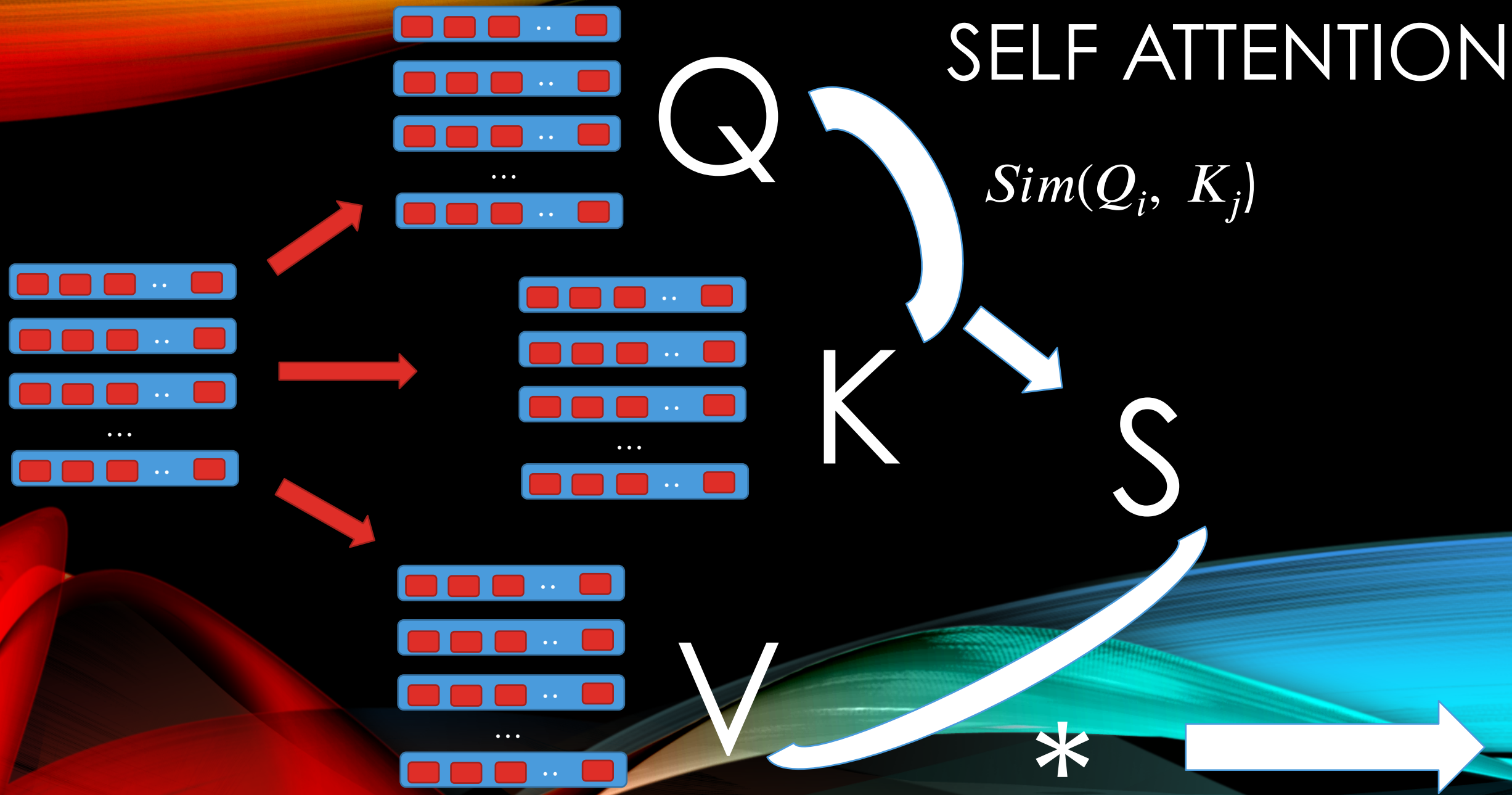
SELF ATTENTION

$$\begin{pmatrix} 10 & 9 & -20 & -34 \\ 10 & -34 & 29 & 0 \\ -29 & -56 & 99 & -3 \\ -22 & 13 & 13 & -33 \end{pmatrix} \xrightarrow{\text{softmax}} \begin{pmatrix} 0.6 & 0.4 & 0 & 0 \\ 0.2 & 0 & 0.8 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0.5 & 0.5 & 0 \end{pmatrix}$$

ATTENTION


Destruction
of
the
equipment
means
that
Syria
can
no
longer
produce
new
chemical
weapons
.
<end>
La
destruction
de
l'
équipement
signifie
que
la
Syrie
ne
peut
plus
produire
de
nouvelles
armes
chimiques
.
<end>

SELF ATTENTION



SELF ATTENTION

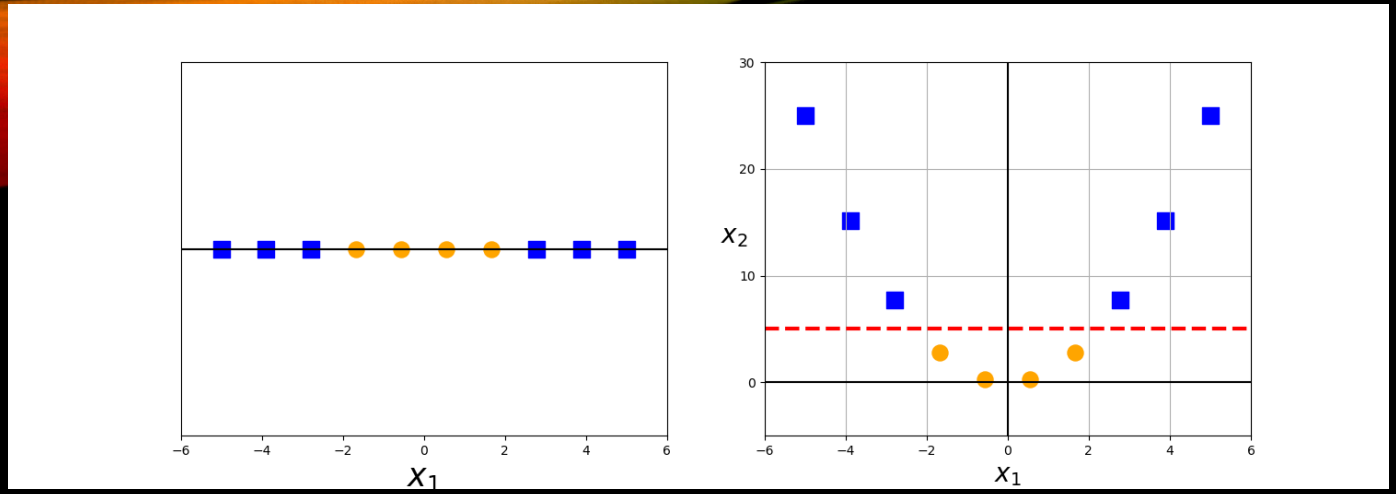
$$\text{Softmax}(Q^T K)V\,[O(n^2 d)]$$

SELF ATTENTION

$$\frac{Sim(Q,\ K)}{\sum_i Sim(Q,\ K)_i} V$$

$$Sim(a,\ b) = e^{a^T b}$$

KERNELS



$$\varphi(x) = (x,\ x^2)$$

$$K(x,\ y) = \varphi(x)^T\varphi(y) = (x,\ x^2)^T(y,\ y^2) = xy + x^2y^2$$

TRANSFORMERS ARE RNNS

$$\frac{Sim(Q,\ K)}{\sum_i Sim(Q,\ K)_i} V \qquad \frac{\varphi(Q)^T \varphi(K)}{\sum_i \varphi(Q)_i^T \varphi(K)_i} V$$

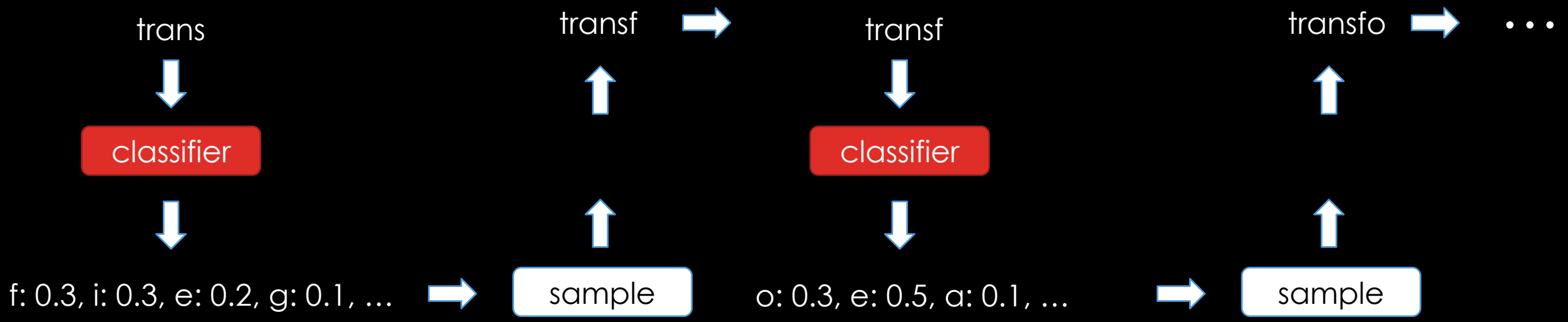
TRANSFORMERS ARE RNNS

$$X_k = \frac{\varphi(Q)^T \varphi(K)}{\sum_i \varphi(Q)_i^T \varphi(K)_i} V_k = \frac{\sum_j \varphi(Q_k)^T \varphi(K_j) V_j}{\sum_j \varphi(Q_k)^T \varphi(K_j)}$$

$$= \frac{\varphi(Q_k)^T \sum_j \varphi(K_j) V_j^T}{\varphi(Q_k)^T \sum_j \varphi(K_j)} [O(n^2 d) \rightarrow O(nd)]$$

$$\varphi(\mathrm{x}) = \mathrm{elu}(\mathrm{x}) + 1$$

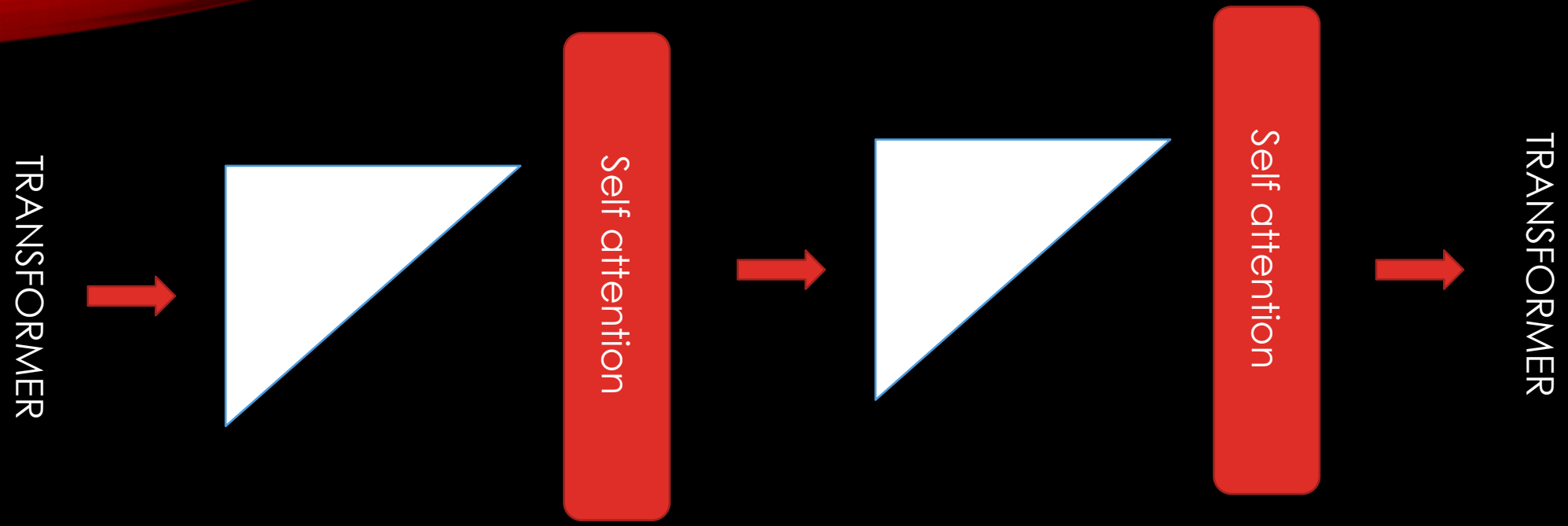
AUTOREGRESSIVE MODELS



$O(n^2d)$ - наивная имплементация kernel трансформером

$O(n^3d)$ - наивная имплементация softmax трансформером

TEACHER FORCING + CAUSAL MASKING

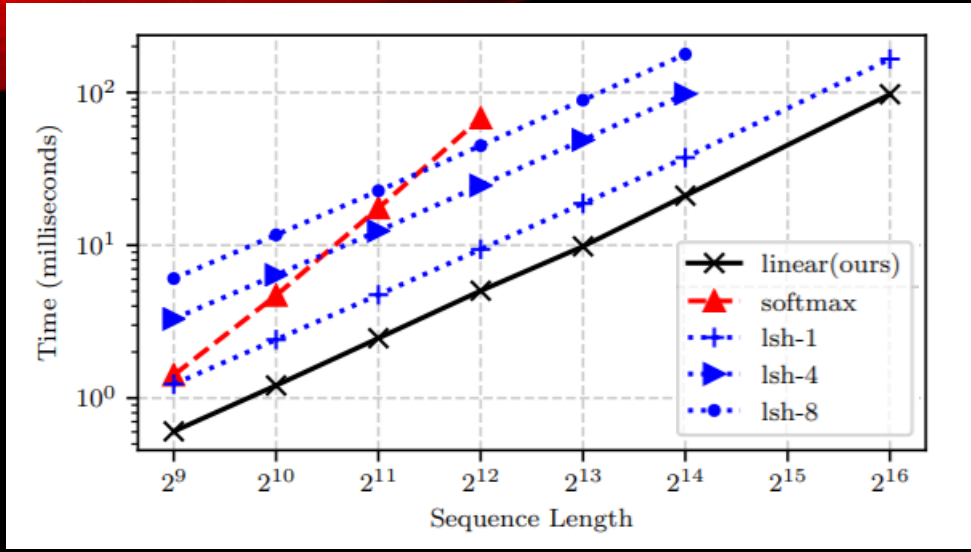


$O(nd)$ - наивная имплементация kernel трансформером

$O(n^2d)$ - наивная имплементация softmax трансформером

$O(nd)$ - RNN

TRANSFORMERS ARE RNNS



MNIST autoregressive

Method	Bits/dim	Images/sec	
Softmax	0.621	0.45	(1×)
LSH-1	0.745	0.68	(1.5×)
LSH-4	0.676	0.27	(0.6×)
Linear (ours)	0.644	**142.8**	**(317×)**

CIFAR-10 autoregressive

Method	Bits/dim	Images/sec	
Softmax	3.47	0.004	(1×)
LSH-1	3.39	0.015	(3.75×)
LSH-4	3.51	0.005	(1.25×)
Linear (ours)	3.40	**17.85**	**(4,462×)**

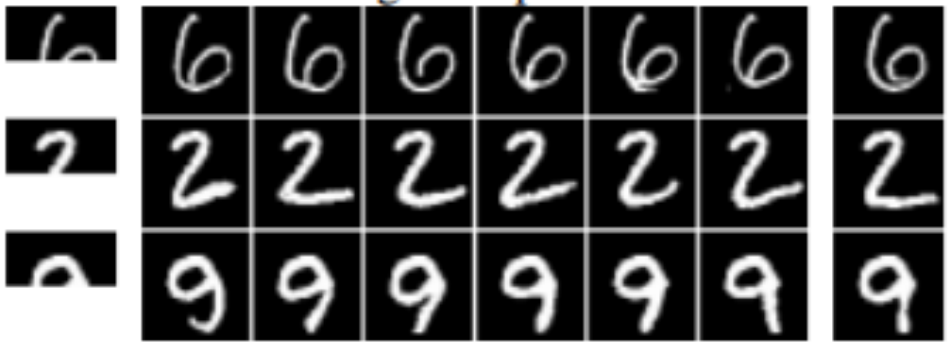
Speech recognition

Method	Validation PER	Time/epoch (s)
Bi-LSTM	10.94	1047
Softmax	5.12	2711
LSH-4	9.33	2250
Linear (ours)	8.08	**824**

TRANSFORMERS ARE RNNS

Unconditional samples

Image completion



(a) (b) (c)

Unconditional samples

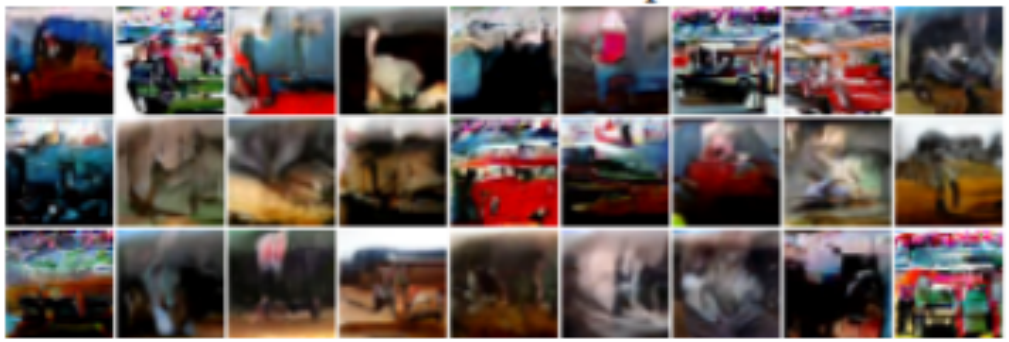


Image completion

(a) (b) (c)

TRANSFORMERS ARE RNNS

$$
\begin{aligned}
s_0 &= 0, \\
z_0 &= 0, \\
s_i &= s_{i-1} + \phi\left(x_i W_K\right)\left(x_i W_V\right)^T, \\
z_i &= z_{i-1} + \phi\left(x_i W_K\right), \\
y_i &= f_l\left(\frac{\phi\left(x_i W_Q\right)^T s_i}{\phi\left(x_i W_Q\right)^T z_i} + x_i\right).
\end{aligned}
$$

QUESTIONS

1) какова асимптотическая сложность обычного трансформера в зависимости от длины последовательности токенов n? какова асимптотическая сложность rnn в такой-же ситуации? какова асимптотическая сложность трансформера из статьи?

2) выписать формулу attention стандартного softmax трансформера

3) выписать формулу attention произвольного трансформера через функцию похожести sim.

4) выписать формулу attention произвольного трансформера, через функцию добавляющую новые признаки соответствующие ядру трансформера (ядро наоборот) phi.

SOURCES

1) https://arxiv.org/abs/2006.16236

2) https://arxiv.org/abs/1706.03762

3) https://www.youtube.com/watch?v=hAooAOFRsYc