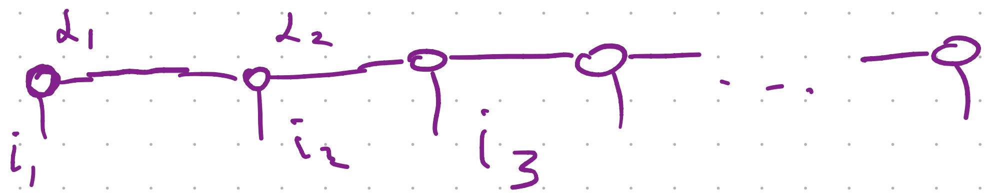
1. Сеть LeNet

2. TT разложение, ядра

TT

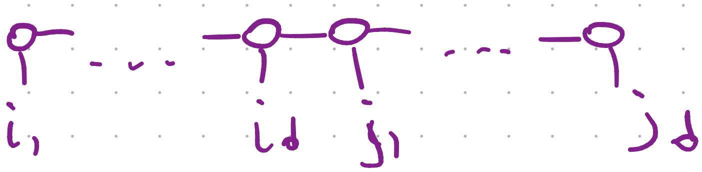
$$X_{i_1\dots i_d} = G_1(i_1)\dots G_d(i_d) =$$

$$= \sum_{\alpha_1,\dots,\alpha_{d-1}=1}^{r_1,\dots,r_{d-1}} g^{(1)}_{i_1\alpha_1}\, g^{(2)}_{\alpha_1 i_2 \alpha_2}\, g^{(3)}_{\alpha_2 i_3 \alpha_3}\dots g^{(d)}_{\alpha_{d-1} i_d}$$



$$y_i = \sum a_{ij} x_j$$

$$y_{i_1\dots i_d} = \sum_{j_1,\dots,j_d} \underline{a_{i_1\dots i_d\, j_1\dots j_d}}\ \underline{x_{j_1\dots j_d}}$$

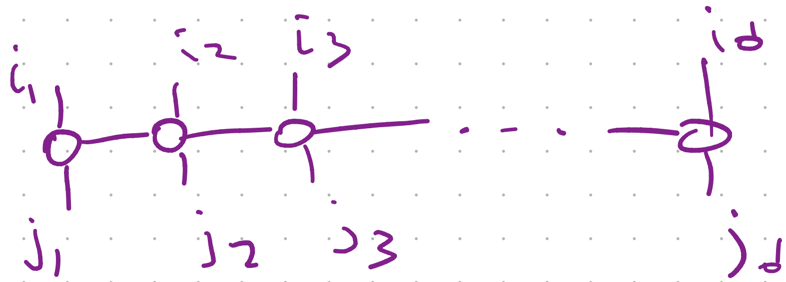


$$a_{\overline{i_1\dots i_d},\ \overline{j_1\dots j_d}}$$

$$A(i_1, \dots, i_7) =$$

$$= \sum_{\alpha_1, \dots \alpha_6 = 1}^{r} g_{i_1 \alpha_1} \cdot g_{\alpha_1 i_2 \alpha_2} \cdot \cdots \cdot g_{\alpha_5 i_6 \alpha_6} \cdot g_{\alpha_6 i_7}$$

$$\tilde{a}_{\overline{i_1 j_1}\,\overline{i_2 j_2}\ldots\overline{i_d j_d}}$$



$$X_{ij} = \sum_{\alpha_1,\ldots,\alpha_{d-1}}^{r_1,\ldots,r_d} a_{i_1,j_1,\alpha_1} \cdot a_{i_2,j_2,\alpha_1,\alpha_2} \cdot \ldots \cdot a_{i_d,j_d,\alpha_{d-1}}$$

$$\delta_{ij} = \delta_{i_1 j_1}\,\delta_{i_2 j_2}\ldots\delta_{i_d j_d}$$

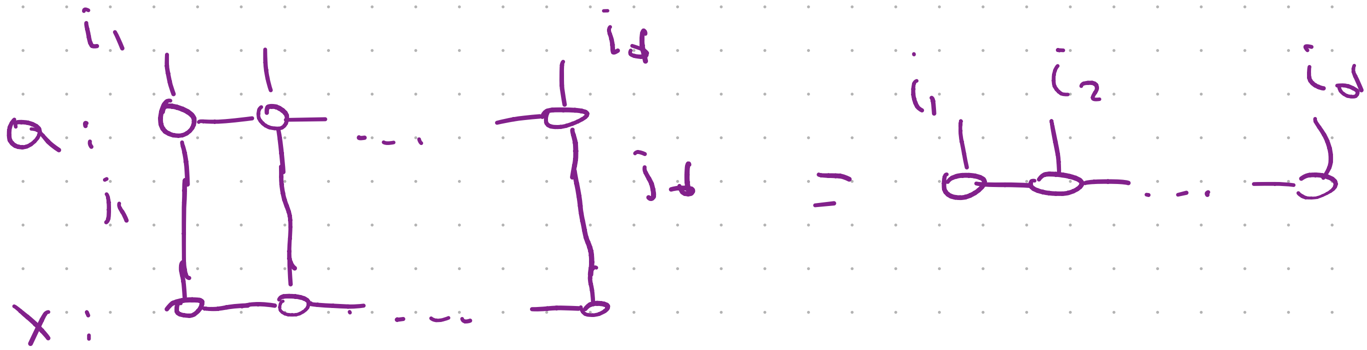
$$\equiv$$

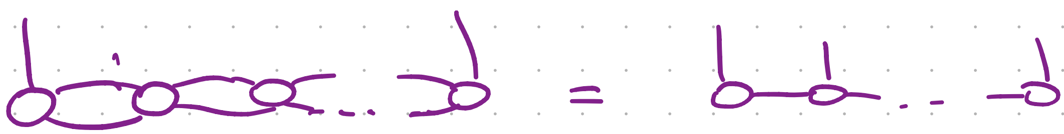
$$\delta_{\overline{i_1\ldots i_d},\,\overline{j_1\ldots j_d}}$$

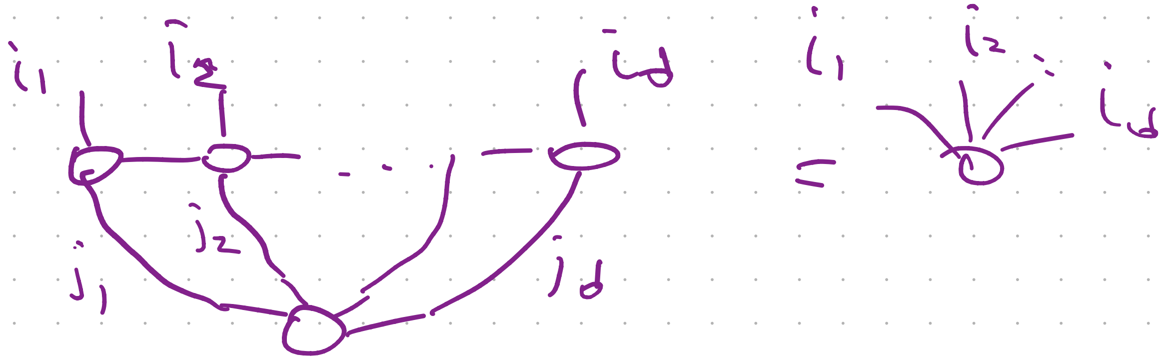
$$A_{(k)} \in \mathbb{R}^{(n_1\ldots n_k)\times(n_{k+1}\ldots n_d)}$$

$$r_k = \min\left(\prod_{j=1}^{k} n_j,\ \prod_{j=k+1}^{d} n_j\right)$$

$$y_{i_1 \ldots i_d} = \sum_{j_1, \ldots, j_d} \underline{a_{i_1 \ldots i_d \, j_1 \ldots j_d}} \; \underline{x_{j_1 \ldots j_d}}$$







$2^d \rightarrow 2 \times \cdots \times 2$

$2^d \quad y = Wx$

$$a = \begin{array}{cc} & 8 \\ 8 & \begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{pmatrix} \end{array}$$

демонстрация, что TT ранги больше 1 =

a=a.reshape([2]*6, order='f')

$a_{i_1 i_2 i_3 j_1 j_2 j_3}$

tt.tensor(a, 1e-8)

(2, 4, 8, 4, 2)

$a_{i_1 j_1 i_2 j_2 i_3 j_3}$

a = np.traspose(a, [0, 3, 1, 4, 2, 5])

a = a.reshape([4, 4, 4], order='f')

= но если исп. формат TT-матрицы, то ранги = 1.

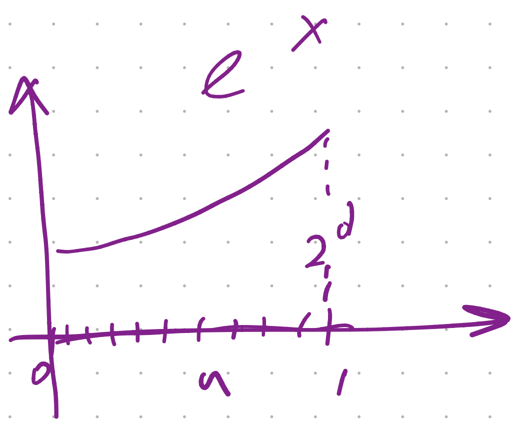
здесь делали TT-SVD и видим, что ранги полные

$$i+j+k = \frac{(1+\varepsilon i)(1+\varepsilon j)(1+\varepsilon k)-1}{\varepsilon} + O(\varepsilon)$$

$$i_1+i_2+\dots+i_d$$

$$\overset{-1/\varepsilon}{\overset{\|}{u_i}}\ \overset{1}{\overset{|}{v_j}}\ \overset{1}{\overset{|}{w_k}}$$

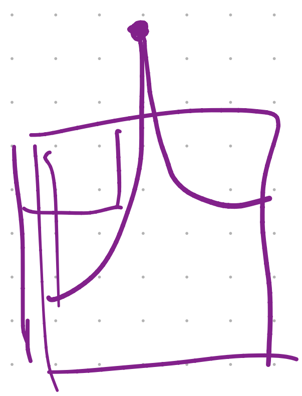
$$\sum u_{id}\, v_{jd}\, w_{kd}$$



$$e^{hi}, \quad i = 0,\dots,\underset{\overset{\|}{n}}{2^d-1}$$

$$i = i_1 2^{d-1}+\dots+i_d$$

$$Q_{ijk}$$



$$e^{(i_1 2^{d-1}+\dots+i_d)h} =$$

$$= e^{i_1 2^{d-1}h}\cdot e^{i_2 2^{d-2}h}\cdots e^{i_d h}$$

* есть 2^d знач.

$1 \times 2 \times 1$ экспонента

$O(\log_2 n)$ сжимаем в TT ранга 1 => храним только TT-ядра.

BLESS of DIMENSIONALITY

CP - единственность. суть вещей из-за единственност

CP-ALS по 3х мерному $i+j+k$.

---

1. CPALS
2. TT → TT един
   TT → TT-слой.

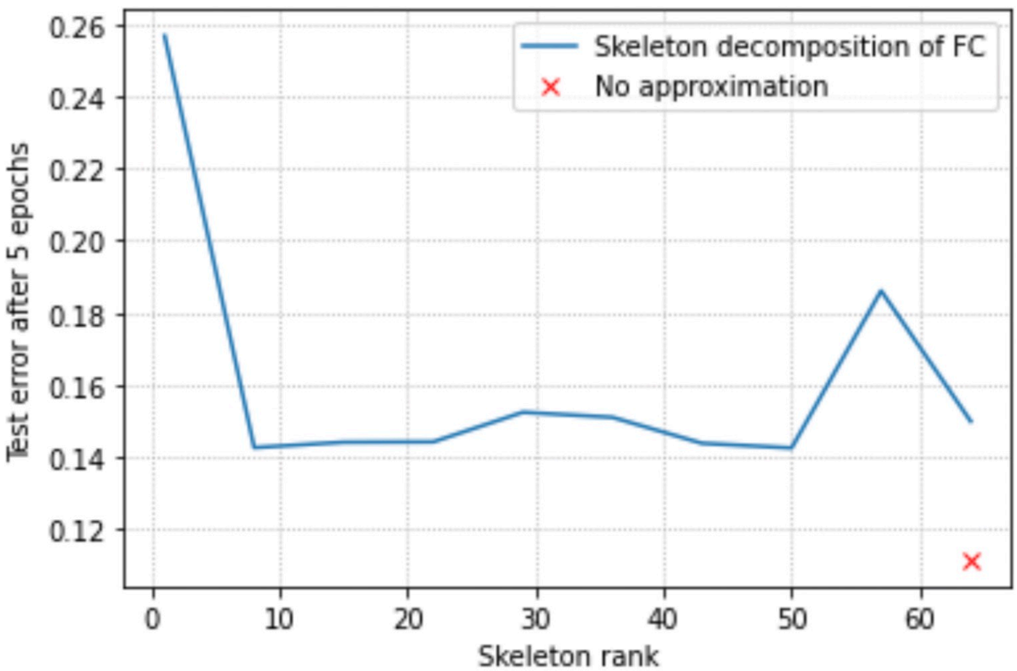
Memory, Quality, Time.

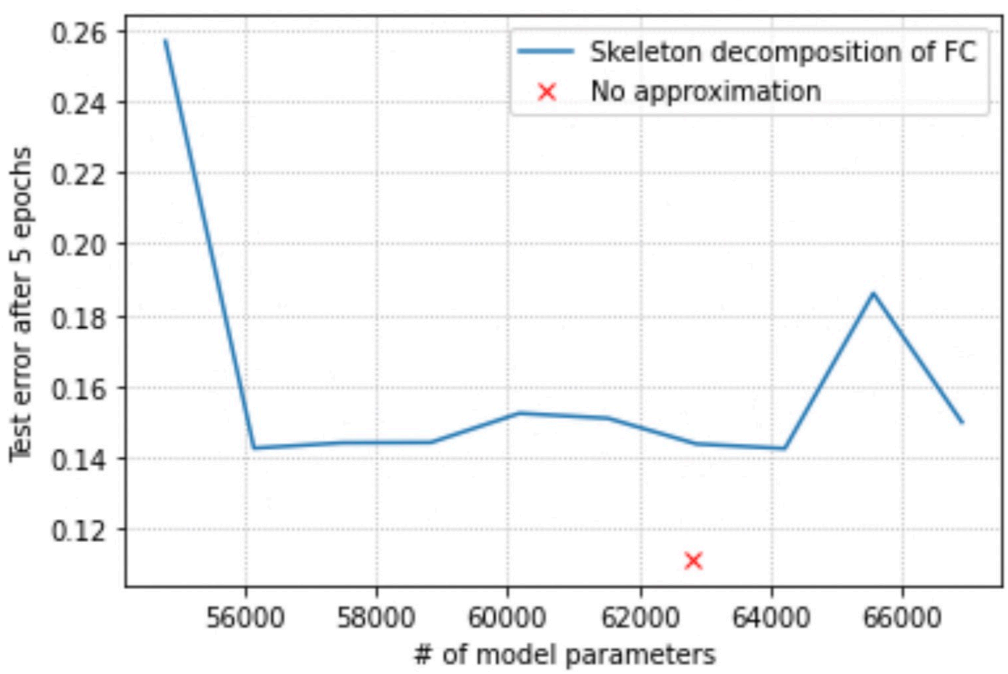
1. обученная сеть AS USUAL
2. обученная с TT-слоем вместо FC-слоя.

① ТТ-матрицы и ТТ-разложение матриц

② Skeleton и ТТ слой вместо FC

③ CP для тензора $i_1+i_2+\dots+i_d$





① TT разложение $I_n = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}$

A_1 $\quad$ A_2 $\quad n = 2^d$

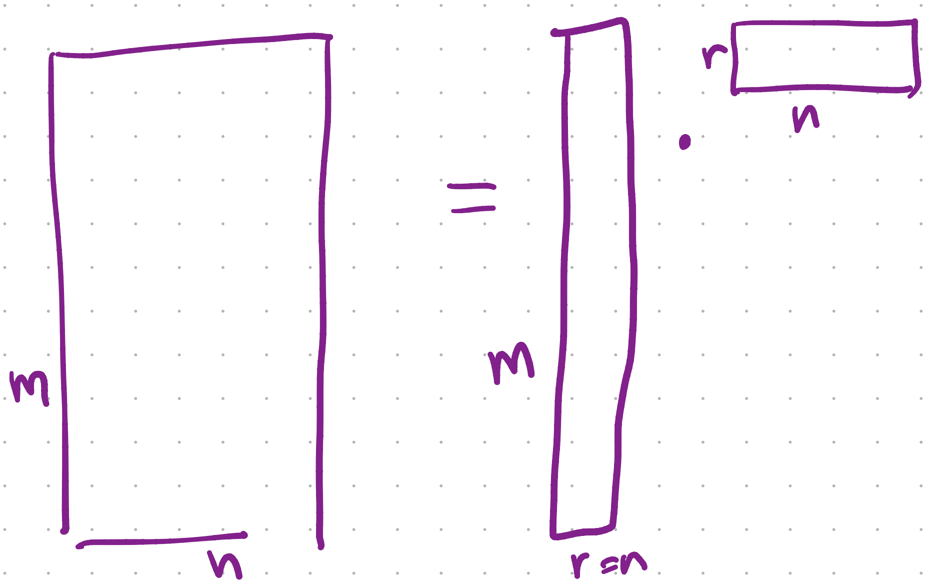
$$2^d \qquad \underbrace{2 \times 2 \times \ldots \times 2}_{d \text{ раз}}$$

② Сжимать FC нейросети с Skeleton и TT

③ CP

$$A_{m \times n} = \underset{m \times r}{U} \; \underset{r \times n}{V}$$

SKELETON



$m \cdot n + nn$

$r = n$

$$I = \delta_{ij} = \begin{cases} 1, & i=j \\ 0, & i \neq j \end{cases}$$

ТТ разложение будет иметь полный ранг

$$A_1 = 2\times 2\times 2\times 2\times 2\times 2$$

$$A_{i_1 \dots i_6}$$

ТТ ранги . ТТ развертки

$$A_{(1)} = \dots 2\times 2^5$$

$$A_{(2)} = 2^2\times 2^4$$

$$\vdots$$

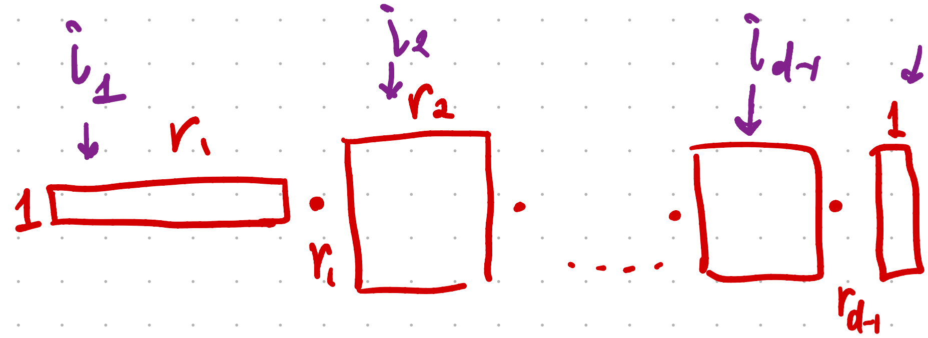
$$A_{(5)}$$

ранги матриц $A_{(i)}$ $i=1,5$ – ТТ-ранги

$$r \in R^{d-1}$$

скаляр

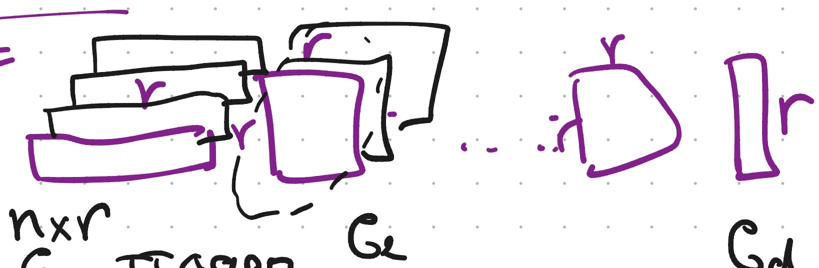
$$A_{i_1,\dots i_d} = 1$$



$$r_1 = \dots = r_{d-1} = r$$

n^d

$$\underbrace{A_{i_1 \dots i_d}}_{\substack{n\times n\times n \dots n \\ d \text{ раз}}} =$$



$i_1 = 1,\dots n$

$i_2 = 1,\dots n$

общий размер ядер $\quad nr \quad n \cdot r^2(d-2) \quad nr$

было:

$$I_{ij} \longrightarrow I_{i_1 i_2 i_3 i_4 i_5 i_6}$$

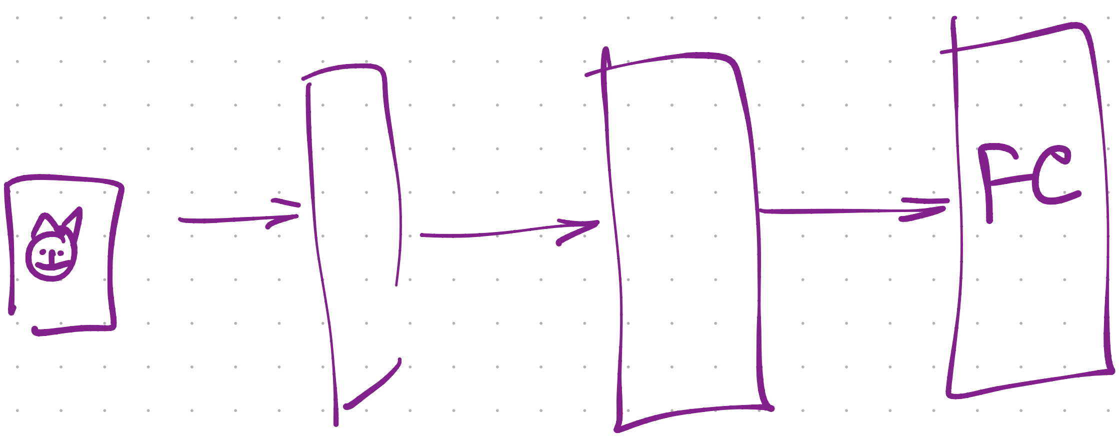
$$\downarrow$$

$$I_{\underbrace{i_1 i_4}\underbrace{i_2 i_5}\underbrace{i_3 i_6}} =$$

$$\delta_{ij} =$$

$$\delta_{\underbrace{i_1 i_4} i_2 i_5 i_3 i_6} = \delta_{i_1 i_4} \cdot \delta_{i_2 i_5} \cdot \delta_{i_3 i_6}$$

Neural Network



$$y_{out} = \underset{out \times in}{W} \cdot x_{in} + b$$

$W_{64\times128} = \underset{U}{\overset{}{64\,\boxed{?}\,r}} \cdot r\,\overset{128}{\boxed{?}}\,V$ $\quad r = 1, 64$

обучаемые пар-ры

$$\frac{(64+128)r}{64\times128}$$

Наивный ТТ слой нейросети (вместо FC)

$W_{64\times128}$ всего элементов 2^{13} order = 'F' order = 'C'

① $\underbrace{\overset{0}{2}\times\overset{1}{2}\times\ldots\ldots\times\overset{12}{2}}_{13}$

ТТ – МАТРИЦА

② permute

$\underbrace{0\ 6}\ \underbrace{1\ 7}\ 2 \ldots\ldots 12$

$\overset{0}{2}\times\overset{1}{2}\times\overset{2}{2}\times\overset{3}{2}\times\overset{4}{2}\times\overset{5}{2}\times\overset{6}{2}\times\overset{7}{2}\times\overset{8}{2}\times\overset{9}{2}\times\overset{10}{2}\times\overset{11}{2}\times\overset{12}{2}$

$\underbrace{0\ 6}\ \underbrace{1\ 7}\ 3\ 8 \ldots\ldots\ldots\ 5\ 11\ 12$

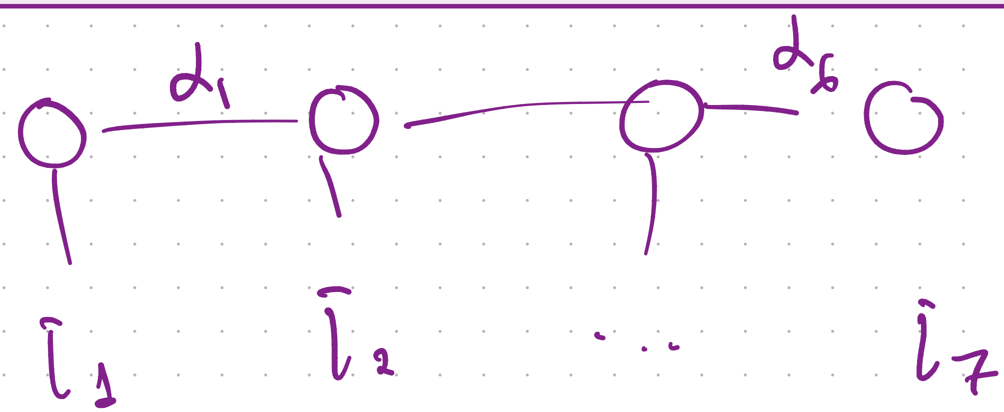
③ RESHAPE

$$\boxed{T = 4\times4\times4\times4\times4\times4\times2.}$$

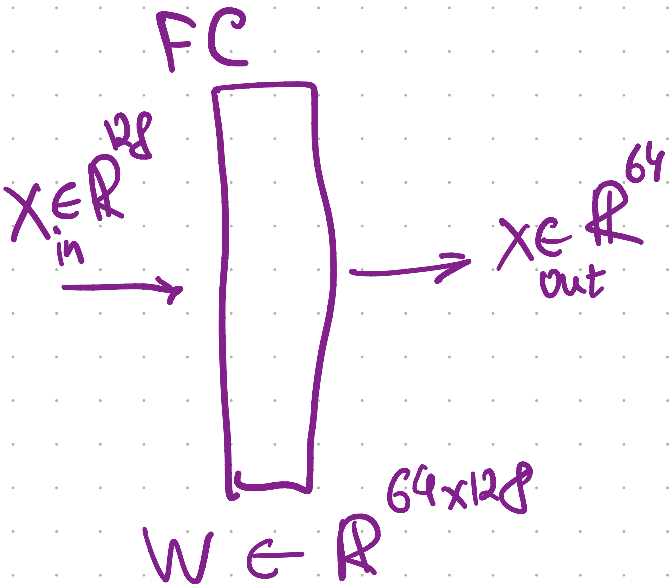
СЖАТИЯ НЕТ!

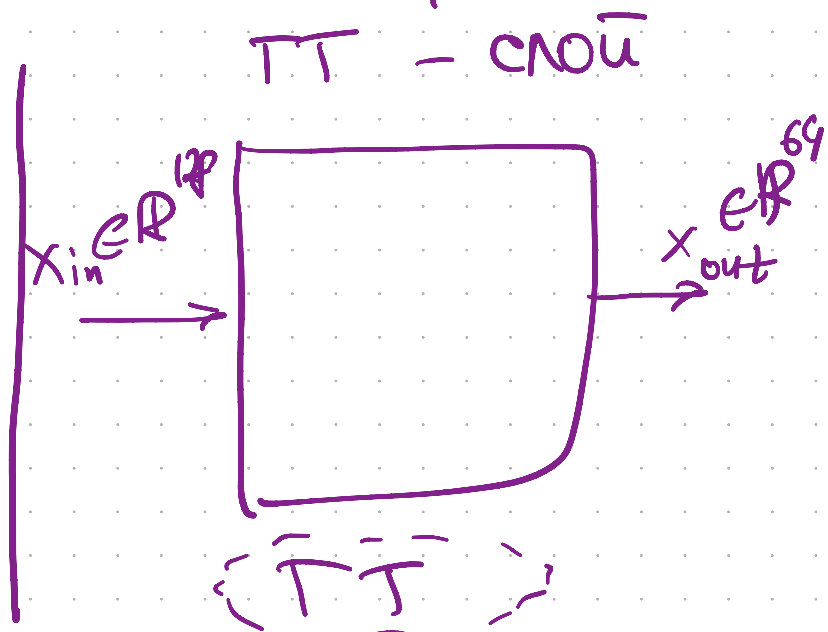
④ ищем T в TT формате (собирать из ядер)

$$\boxed{T_{i_1,\dots,i_d} = \sum_{\alpha_1\dots\alpha_6=1}^{r} g_{i_1\alpha_1}\, g_{\alpha_1 i_2 \alpha_2} \cdots g_{\alpha_6 i_7}}$$



есть сжатие, если r – достат. мало

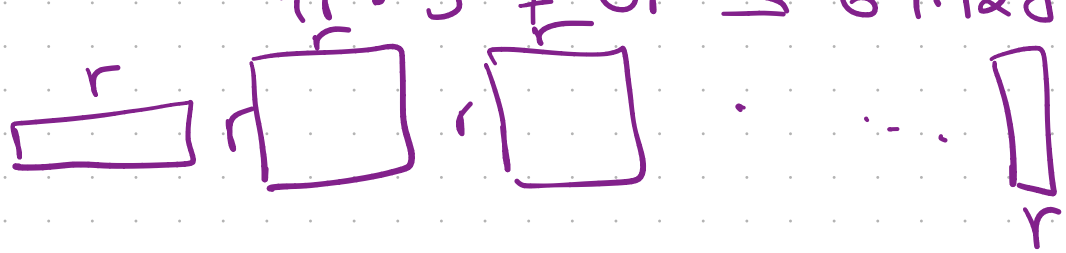




$G_1, \dots\dots G_7$

$4\times r,\ 4r^2,\ 4r^2,\ 4r^2,\ 4r^2,\ 4r^2,\ r\times2$

$$4r^2\cdot 5 + 6r \le 64\cdot128$$



$$i + j + k + \ell = \frac{(1+\varepsilon i)(1+\varepsilon j)(1+\varepsilon k)(1+\varepsilon \ell)^d - 1^d}{\varepsilon} + O(\varepsilon)$$

ep $\uparrow$ rank

d

$r = 2$