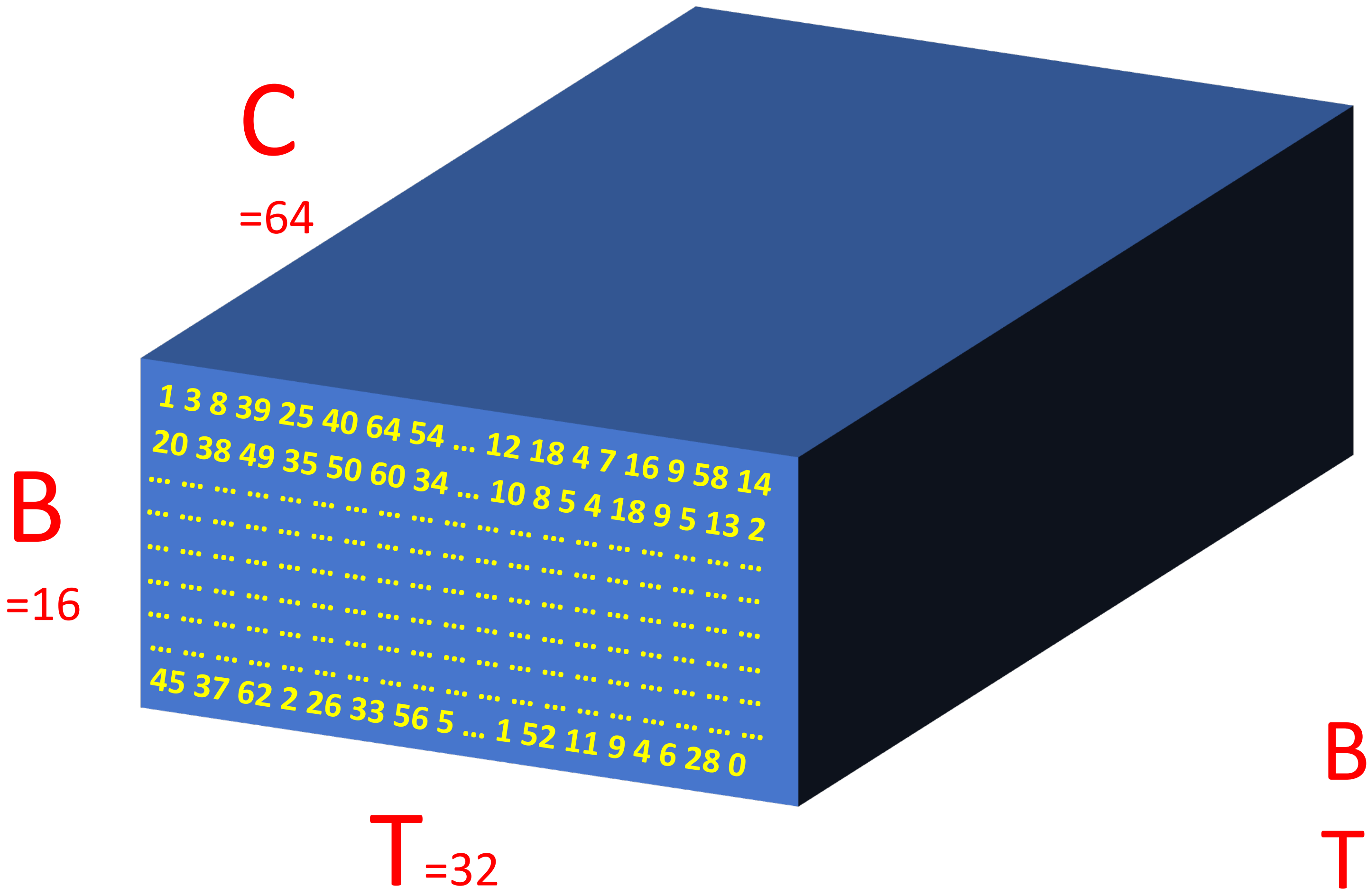
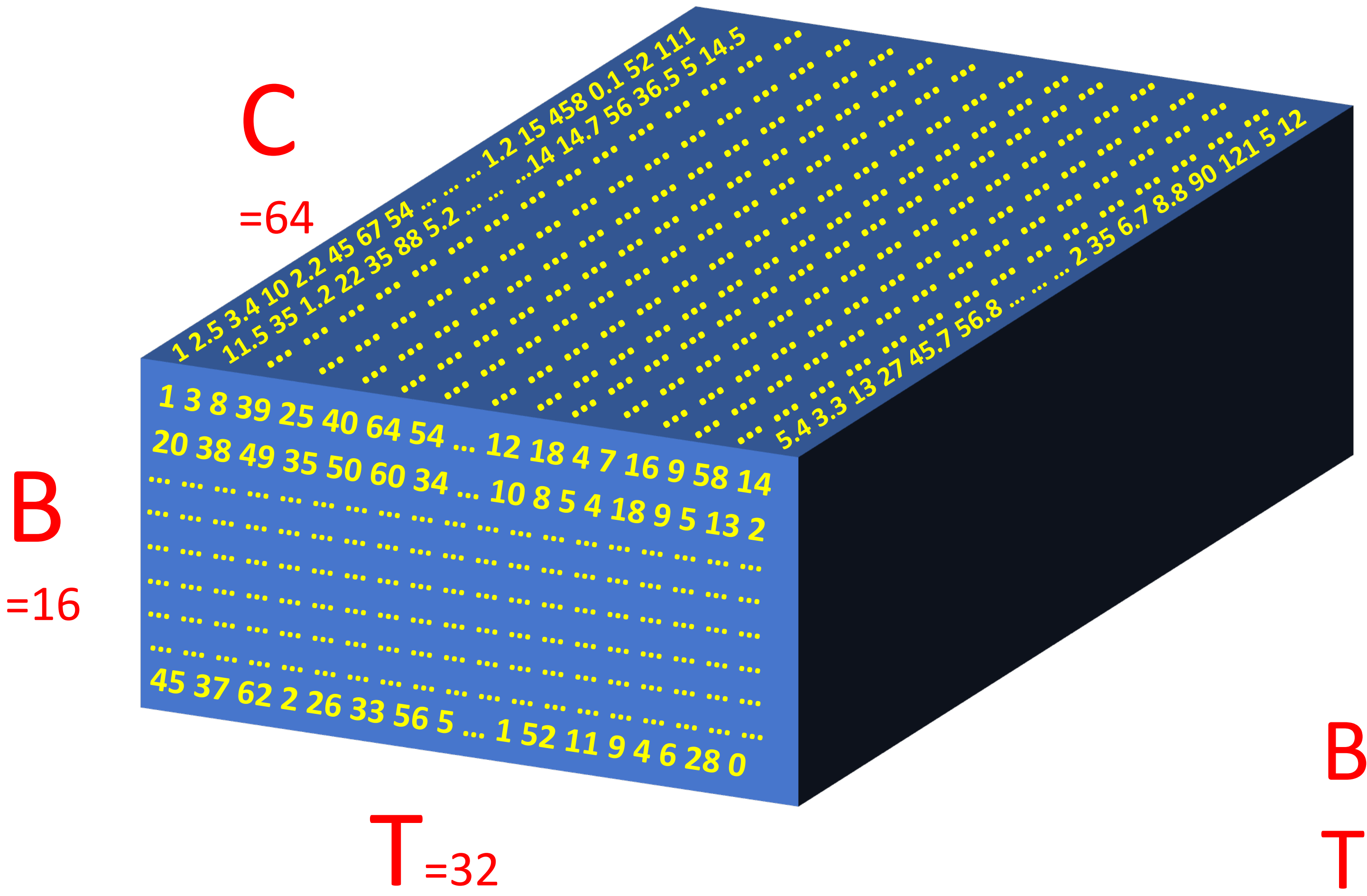
# Visualize a batch of data



B: batch_size

T: block_size

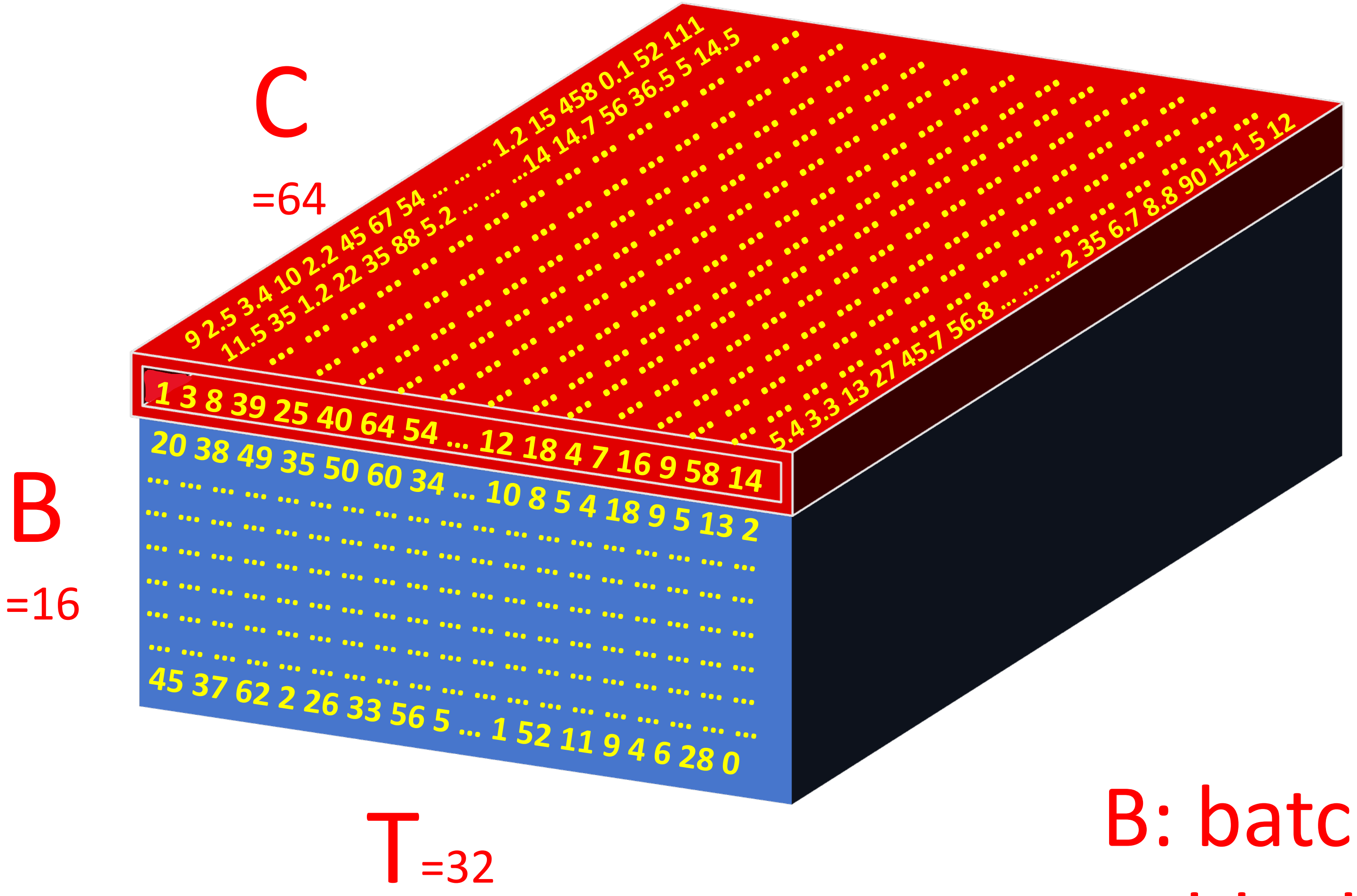
C: n_embd



B: batch_size

T: block_size

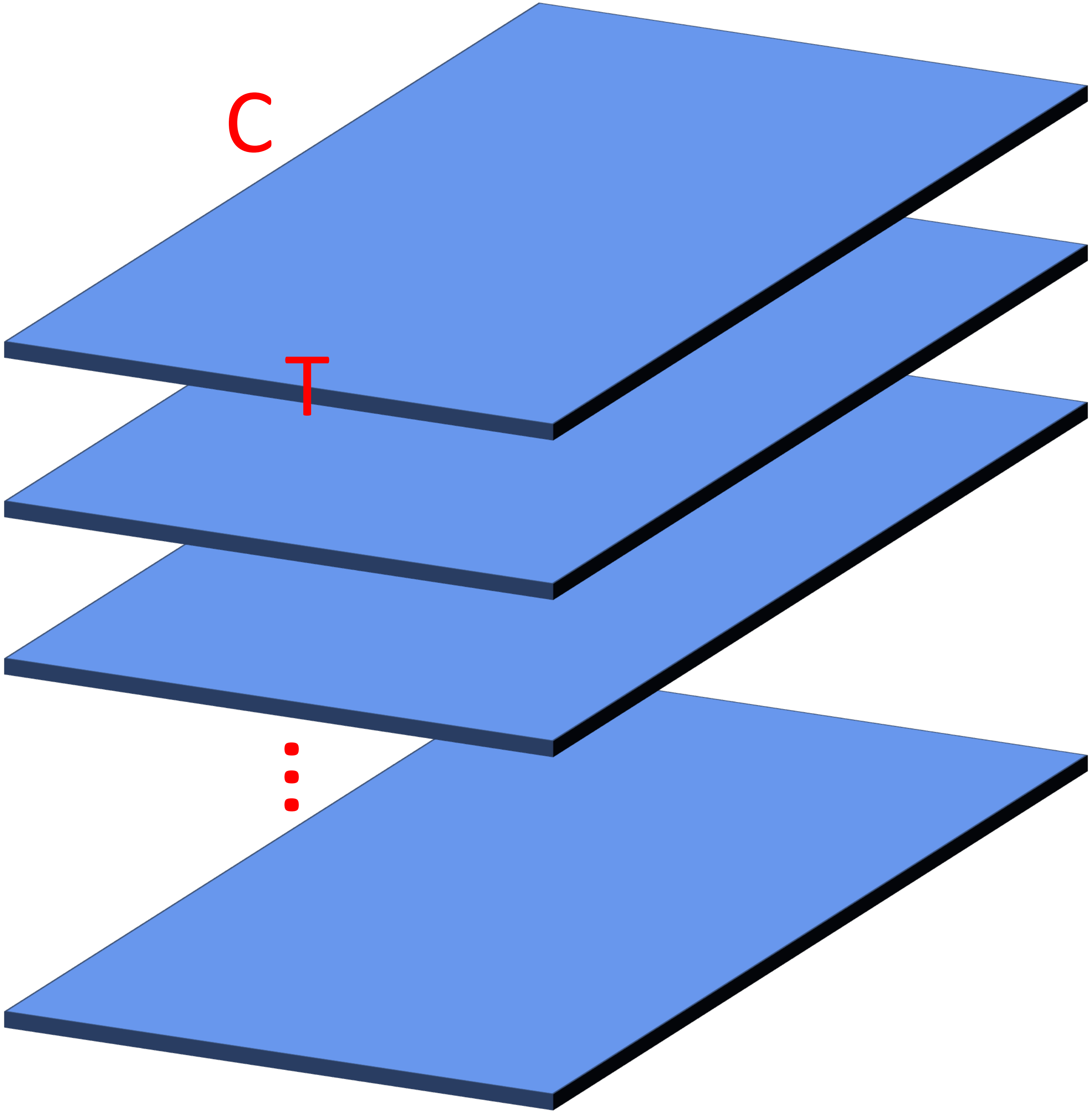
C: n_embd



B: batch_size

T: block_size

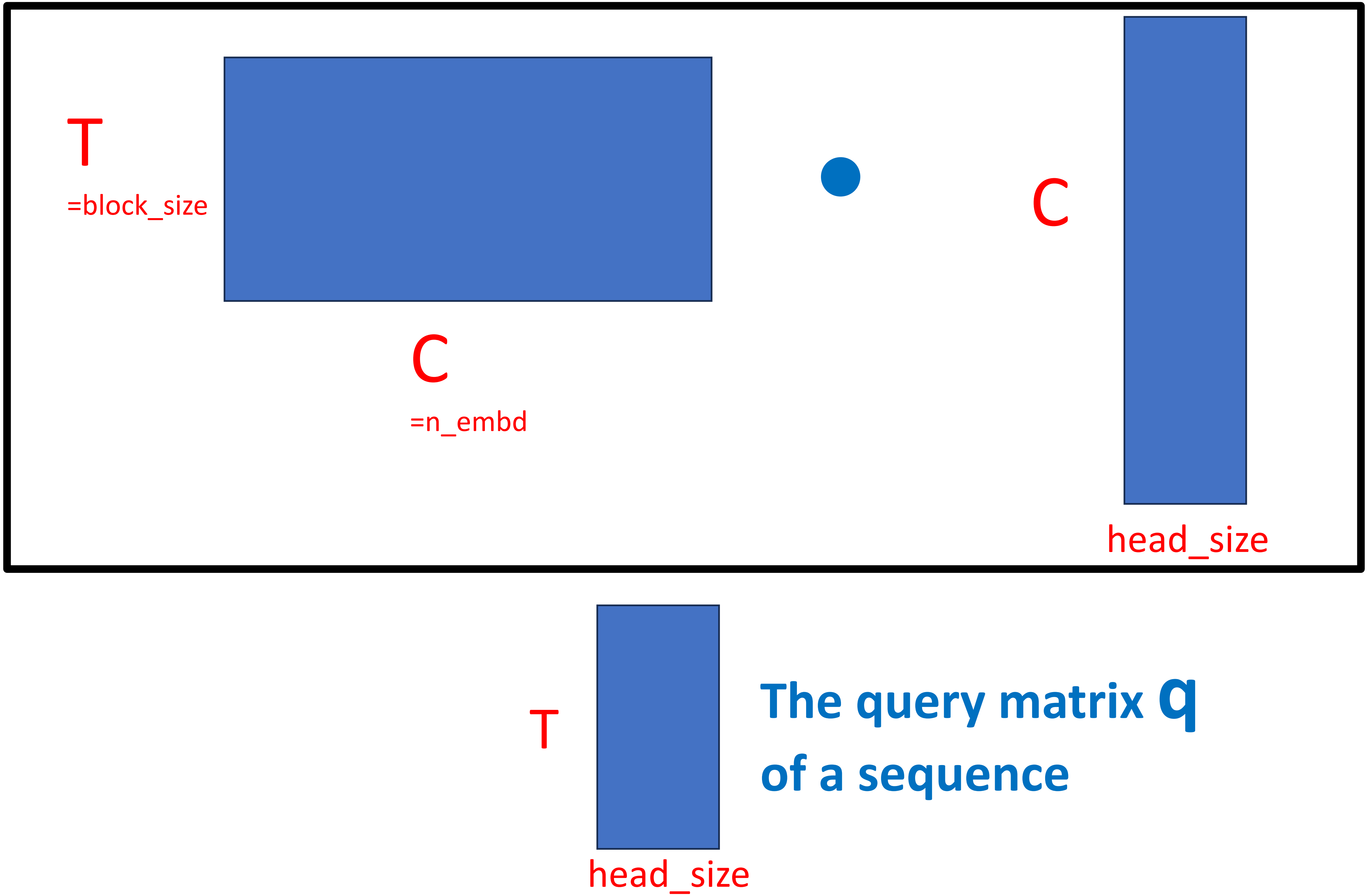
C: n_embd



16 slates stacked together

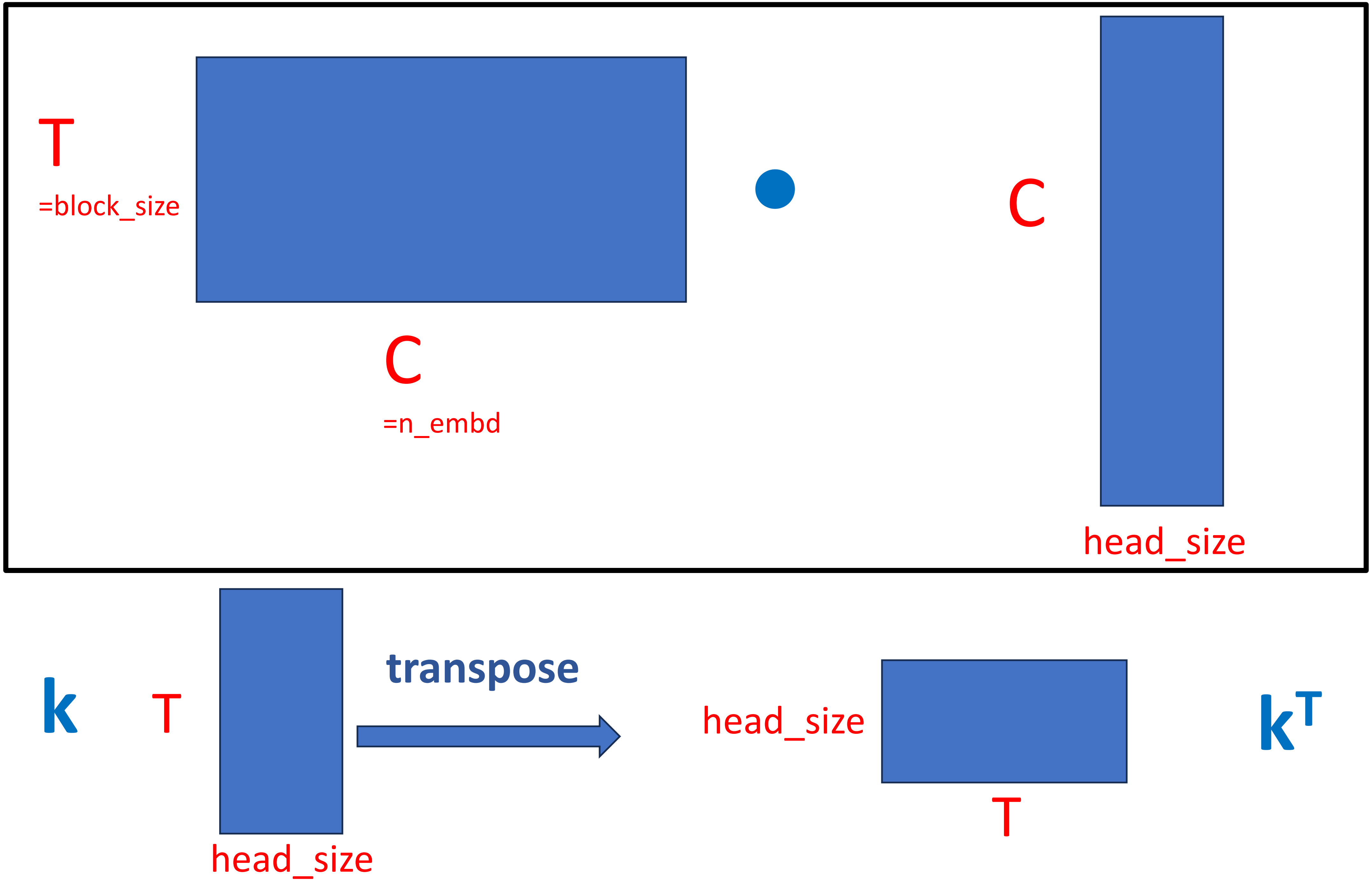
each slate is a sequence of chars in the batch and is a matrix of shape (T, C)

# Visualize a single head



**The query matrix q of a sequence**

$$\begin{array}{l|cccc|}
\mathbf{q}^{\mathrm{chr1}} & q_1^{chr1} & q_2^{chr1} & \cdots & q_{16}^{chr1} \\
\mathbf{q}^{\mathrm{chr2}} & q_1^{chr2} & q_2^{chr2} & \cdots & q_{16}^{chr2} \\
\mathbf{\dots} & \cdots & \cdots & \cdots & \cdots \\
\mathbf{q}^{\mathrm{chr32}} & q_1^{chr32} & q_2^{chr32} & \cdots & q_{16}^{chr32}
\end{array}$$



T

=block_size

C

=n_embd

●

C

head_size

k

T

head_size

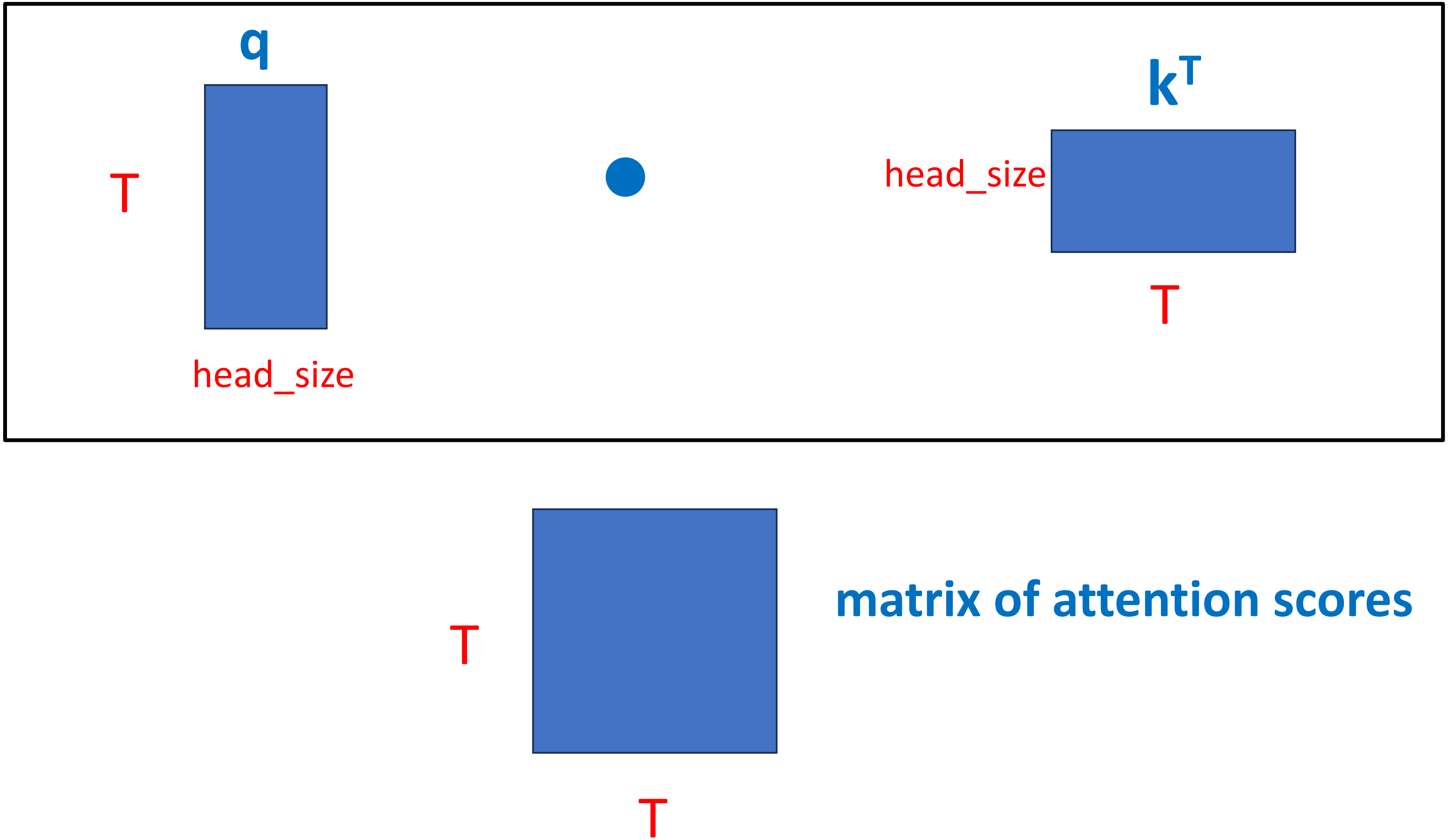
transpose

head_size

T

$k^T$

$$\begin{array}{c}
\begin{array}{cccc} \mathbf{k}^{\text{chr1}} & \mathbf{k}^{\text{chr2}} & \ldots & \mathbf{k}^{\text{chr32}} \end{array} \\
\begin{bmatrix}
k_1^{chr1} & k_1^{chr2} & \cdots & k_1^{chr32} \\
k_2^{chr1} & k_2^{chr2} & \cdots & k_2^{chr32} \\
\cdots & \cdots & \cdots & \cdots \\
k_{16}^{chr1} & k_{16}^{chr2} & \cdots & k_{16}^{chr32}
\end{bmatrix}
\end{array}$$



q

T

head_size

●

$k^T$

head_size

T

T

T

matrix of attention scores

$$\begin{array}{c|cccc|}
\mathbf{q}^{\text{chr1}} & q_{1}^{chr1} & q_{2}^{chr1} & \dots & q_{16}^{chr1} \\
\mathbf{q}^{\text{chr2}} & q_{1}^{chr2} & q_{2}^{chr2} & \dots & q_{16}^{chr2} \\
\dots & \dots & \dots & \dots & \dots \\
\mathbf{q}^{\text{chr32}} & q_{1}^{chr32} & q_{2}^{chr32} & \dots & q_{16}^{chr32}
\end{array}
\ast
\begin{array}{cccc}
\mathbf{k}^{\text{chr1}} & \mathbf{k}^{\text{chr2}} & \dots & \mathbf{k}^{\text{chr32}} \\
\hline
k_{1}^{chr1} & k_{1}^{chr2} & \dots & k_{1}^{chr32} \\
k_{2}^{chr1} & k_{2}^{chr2} & \dots & k_{2}^{chr32} \\
\dots & \dots & \dots & \dots \\
k_{16}^{chr1} & k_{16}^{chr2} & \dots & k_{16}^{chr32}
\end{array}$$

(T, T)

$$\begin{bmatrix} \mathbf{q}^{chr1} * \mathbf{k}^{chr1} & \mathbf{q}^{chr1} * \mathbf{k}^{chr2} & \dots & \mathbf{q}^{chr1} * \mathbf{k}^{chr32} \\ \mathbf{q}^{chr2} * \mathbf{k}^{chr1} & \mathbf{q}^{chr2} * \mathbf{k}^{chr2} & \dots & \mathbf{q}^{chr2} * \mathbf{k}^{chr32} \\ \dots & \dots & \dots & \dots \\ \mathbf{q}^{chr32} * \mathbf{k}^{chr1} & \mathbf{q}^{chr32} * \mathbf{k}^{chr2} & \dots & \mathbf{q}^{chr32} * \mathbf{k}^{chr32} \end{bmatrix}$$

(T, T)

$$\begin{bmatrix} \mathbf{q}^{chr1} * \mathbf{k}^{chr1} & \mathbf{q}^{chr1} * \mathbf{k}^{chr2} & \dots & \mathbf{q}^{chr1} * \mathbf{k}^{chr32} \\ \mathbf{q}^{chr2} * \mathbf{k}^{chr1} & \mathbf{q}^{chr2} * \mathbf{k}^{chr2} & \dots & \mathbf{q}^{chr2} * \mathbf{k}^{chr32} \\ \dots & \dots & \dots & \dots \\ \mathbf{q}^{chr32} * \mathbf{k}^{chr1} & \mathbf{q}^{chr32} * \mathbf{k}^{chr2} & \dots & \mathbf{q}^{chr32} * \mathbf{k}^{chr32} \end{bmatrix}$$

(T, T)

$$\begin{bmatrix} \mathbf{q}^{chr1} * \mathbf{k}^{chr1} & \textbf{-inf} & \ldots & \ldots & \textbf{-inf} \\ \mathbf{q}^{chr2} * \mathbf{k}^{chr1} & \mathbf{q}^{chr2} * \mathbf{k}^{chr2} & \textbf{-inf} & \ldots & \textbf{-inf} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \mathbf{q}^{chr32} * \mathbf{k}^{chr1} & \mathbf{q}^{chr32} * \mathbf{k}^{chr2} & \ldots & \ldots & \mathbf{q}^{chr32} * \mathbf{k}^{chr32} \end{bmatrix}$$

(T, T)

$$\begin{bmatrix} \mathbf{q}^{chr1} * \mathbf{k}^{chr1} & \textbf{-inf} & \dots & \dots & \textbf{-inf} \\ \mathbf{q}^{chr2} * \mathbf{k}^{chr1} & \mathbf{q}^{chr2} * \mathbf{k}^{chr2} & \textbf{-inf} & \dots & \textbf{-inf} \\ \dots & \dots & \dots & \dots & \dots \\ \mathbf{q}^{chr32} * \mathbf{k}^{chr1} & \mathbf{q}^{chr32} * \mathbf{k}^{chr2} & \dots & \dots & \mathbf{q}^{chr32} * \mathbf{k}^{chr32} \end{bmatrix}$$

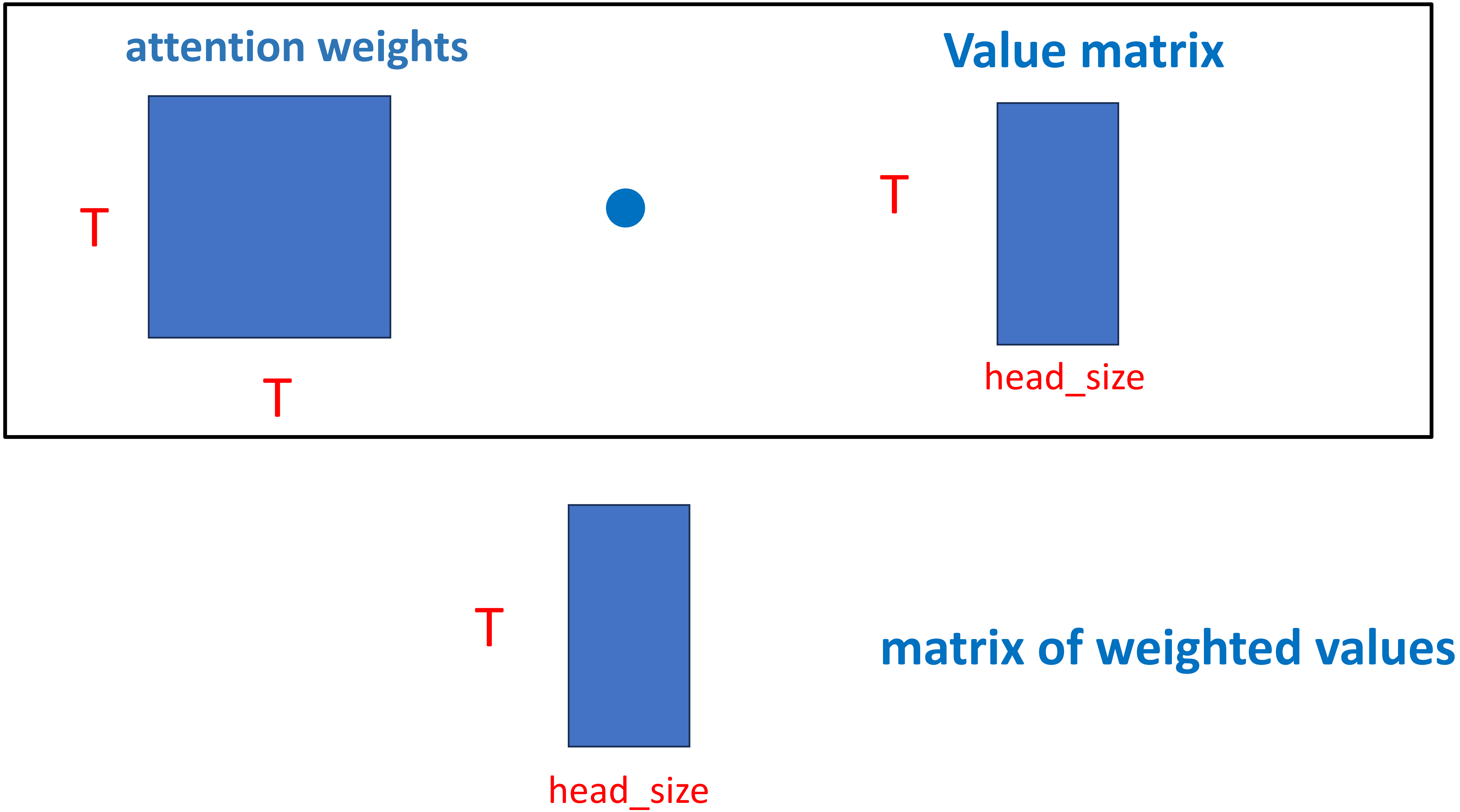
**Softmax**
**to each row**

(T, T)

$$\begin{array}{cl}\begin{bmatrix} w_{\mathbf{q}^{chr1} * \mathbf{k}^{chr1}} & 0 & \dots & \dots & 0 \\ w_{\mathbf{q}^{chr2} * \mathbf{k}^{chr1}} & w_{\mathbf{q}^{chr2} * \mathbf{k}^{chr2}} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & 0 \\ w_{\mathbf{q}^{chr32} * \mathbf{k}^{chr1}} & w_{\mathbf{q}^{chr32} * \mathbf{k}^{chr2}} & \dots & \dots & w_{\mathbf{q}^{chr32} * \mathbf{k}^{chr32}} \end{bmatrix} & \begin{array}{l} \text{sum up to 1} \\ \text{sum up to 1} \\ \text{sum up to 1} \\ \text{sum up to 1} \end{array}\end{array}$$

(T, T)

$$\begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0.22 & 0.78 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0.02 & 0.12 & 0.24 & \dots & 0.45 \end{bmatrix} \begin{array}{l} \text{sum up to 1} \\ \text{sum up to 1} \\ \text{sum up to 1} \\ \text{sum up to 1} \end{array}$$

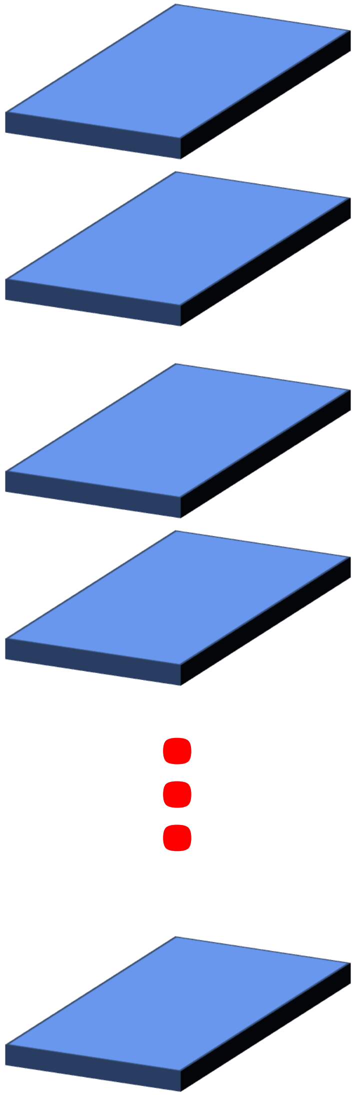


attention weights

Value matrix

matrix of weighted values

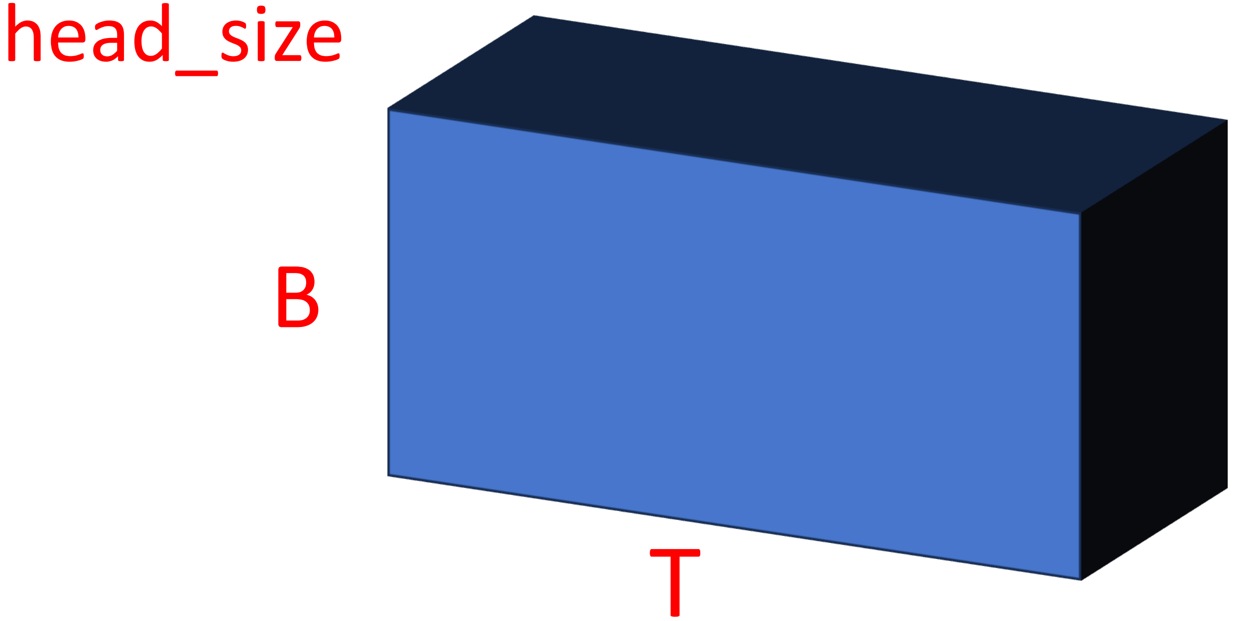
$$\overset{\text{(T, T)}}{\begin{bmatrix} w_{\mathbf{q}^{chr1} * \mathbf{k}^{chr1}} & 0 & \ldots & \ldots & 0 \\ w_{\mathbf{q}^{chr2} * \mathbf{k}^{chr1}} & w_{\mathbf{q}^{chr2} * \mathbf{k}^{chr2}} & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & 0 \\ w_{\mathbf{q}^{chr32} * \mathbf{k}^{chr1}} & w_{\mathbf{q}^{chr32} * \mathbf{k}^{chr2}} & \ldots & \ldots & w_{\mathbf{q}^{chr32} * \mathbf{k}^{chr32}} \end{bmatrix}} \bullet \overset{\text{(T, head\_size)}}{\begin{bmatrix} v_1^{chr1} & v_2^{chr1} & \ldots & v_{16}^{chr1} \\ v_1^{chr2} & v_2^{chr2} & \ldots & v_{16}^{chr2} \\ \ldots & \ldots & \ldots & \ldots \\ v_1^{chr32} & v_2^{chr32} & \ldots & v_{16}^{chr32} \end{bmatrix}}$$



16 slates stacked together

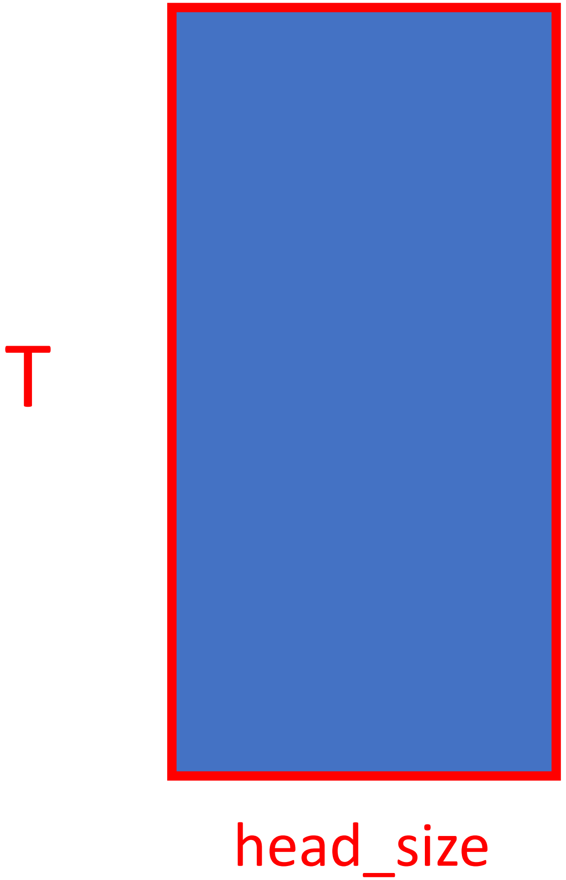
each slate is of shape (T, head_size)
and is a matrix of weighted values
of one sequence in the batch

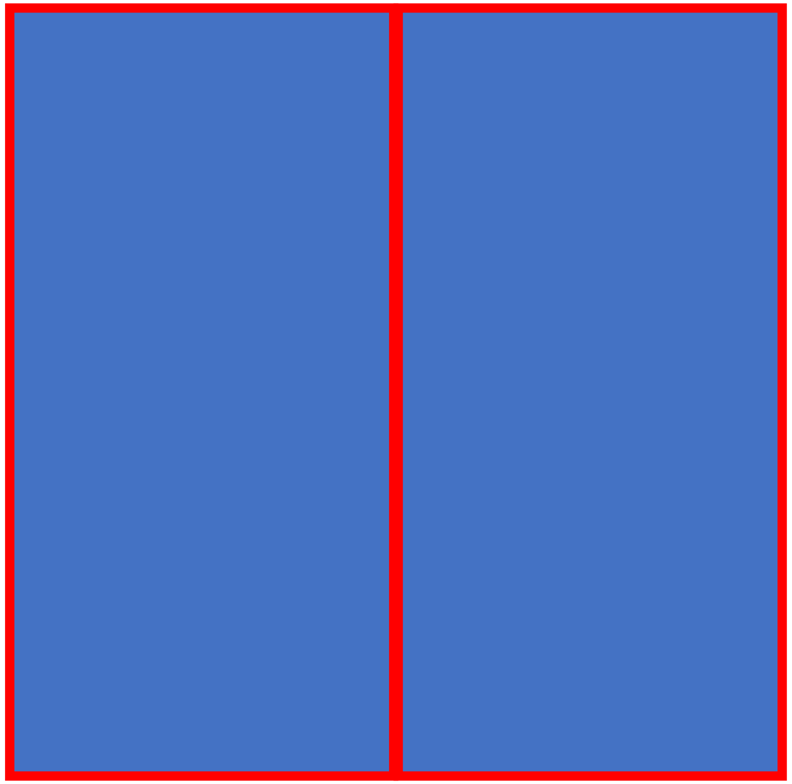
The cuboid formed by these slates
is of shape (B, T, head_size)

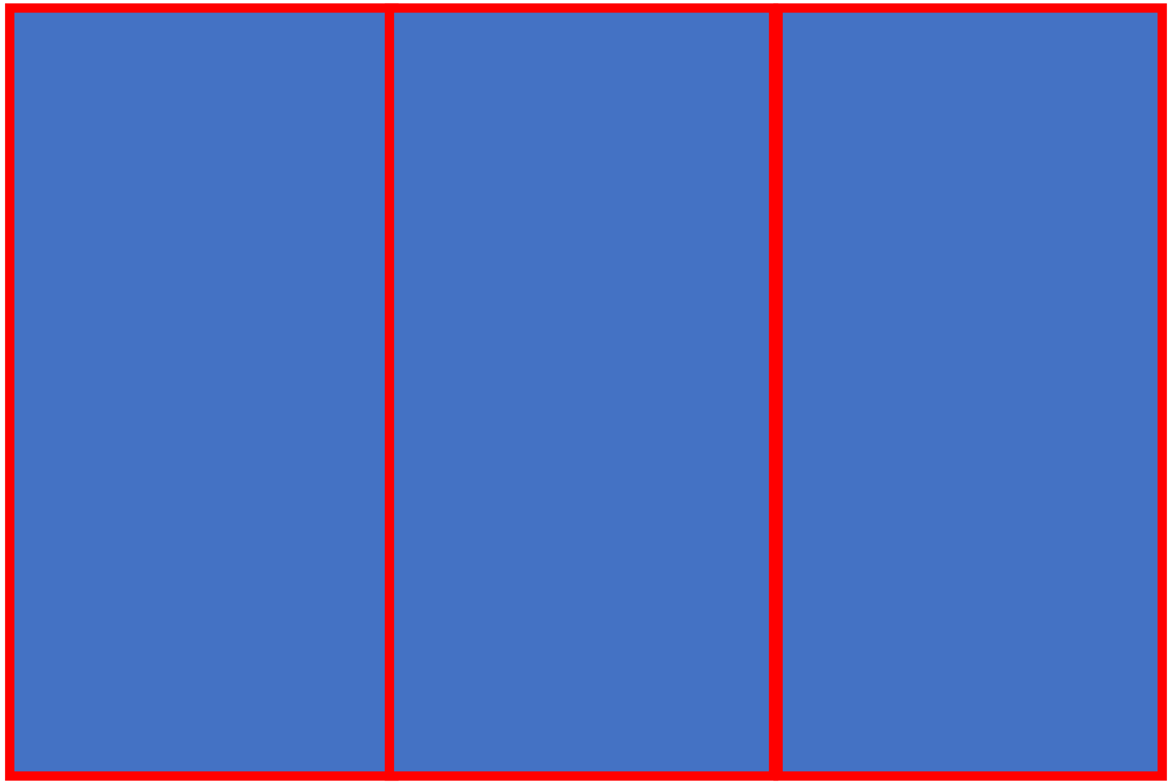


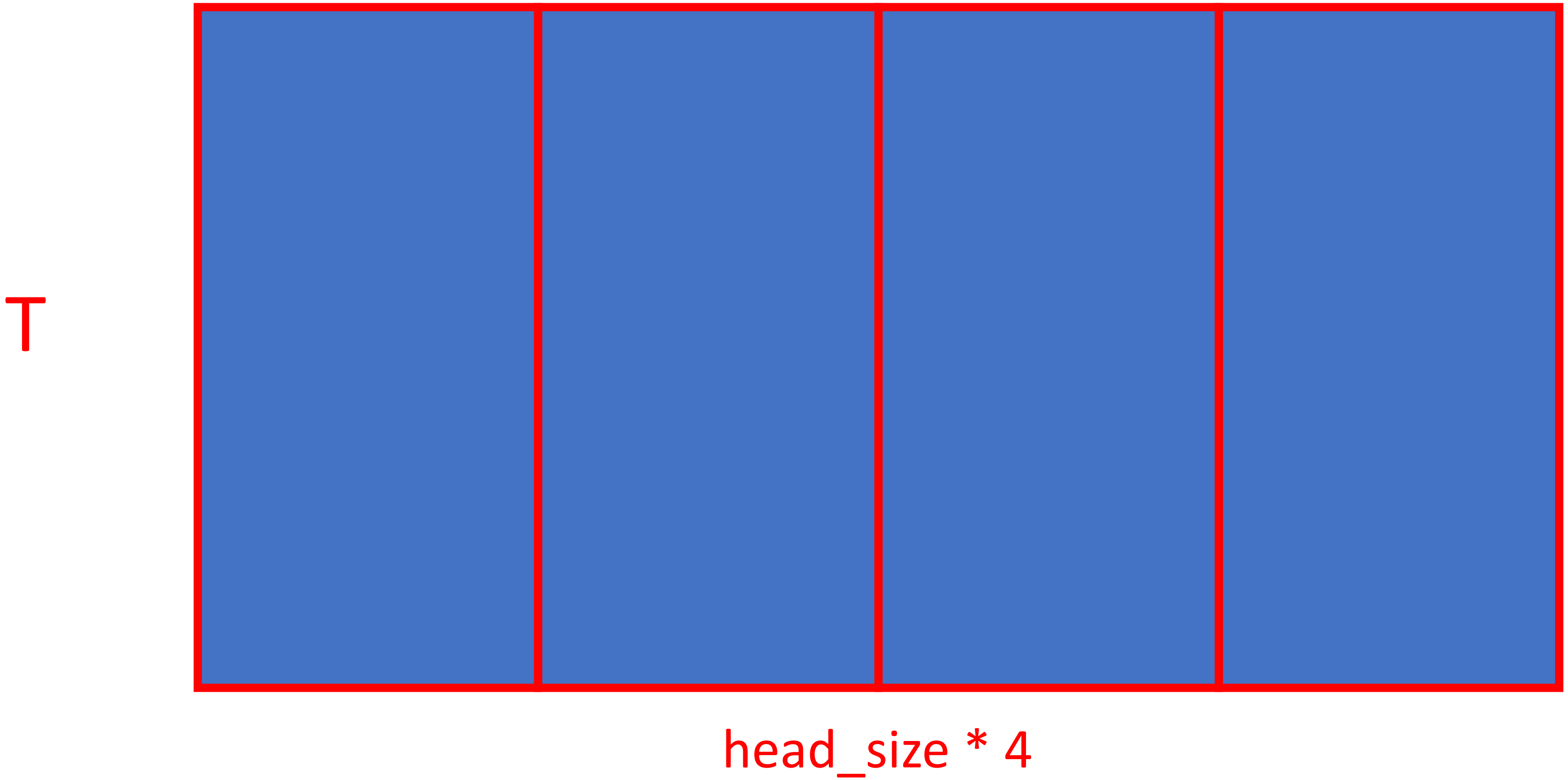
(B, T, head_size)

# Visualize a multi-head

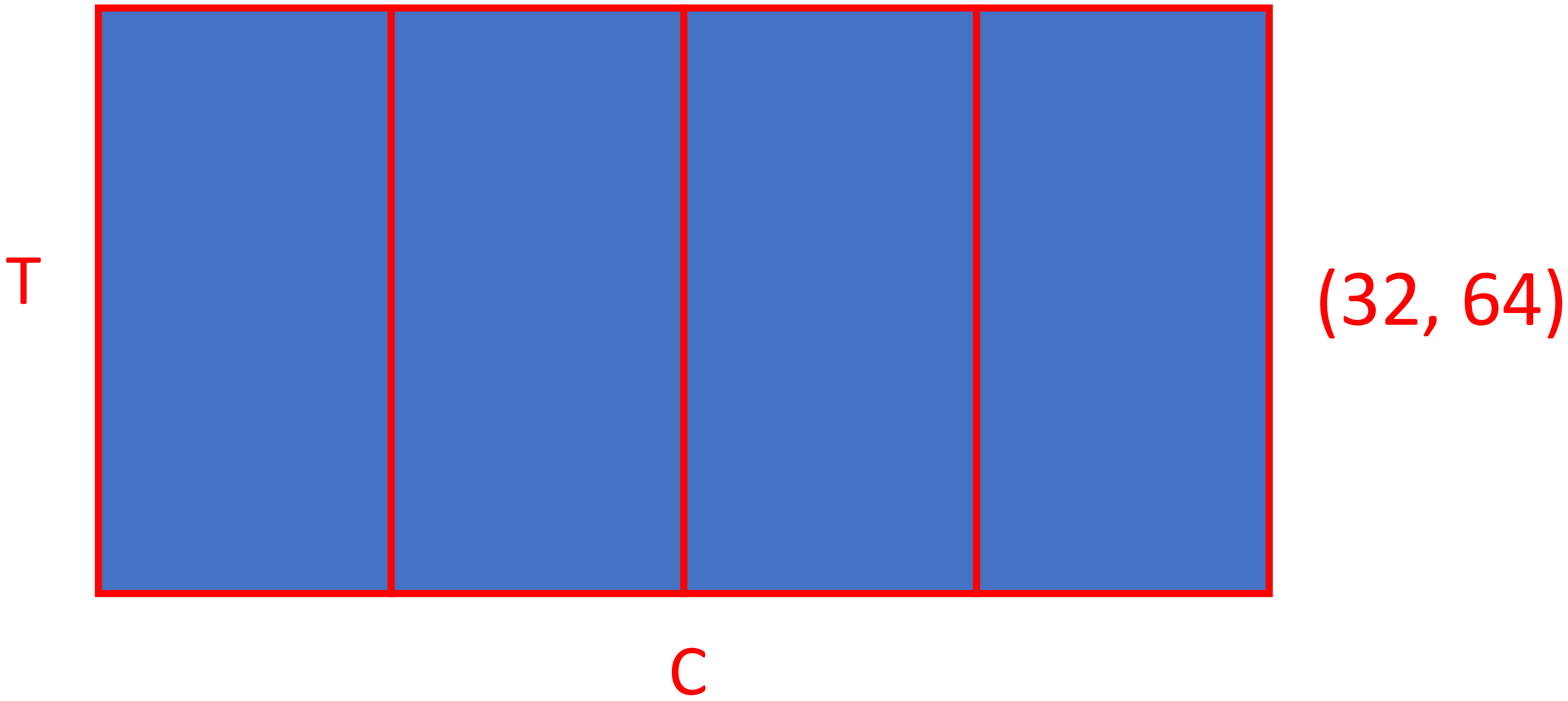

T
head_size

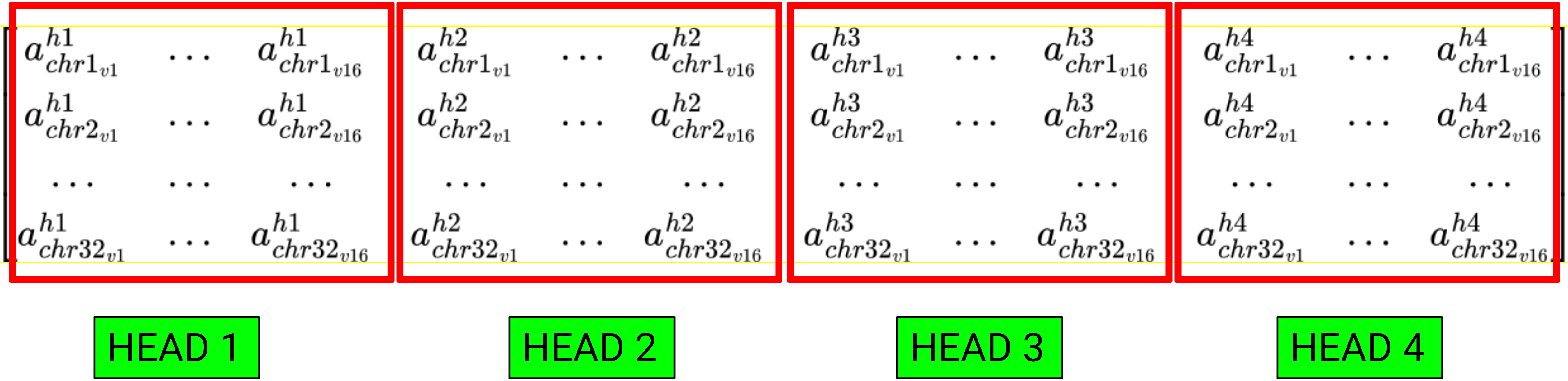




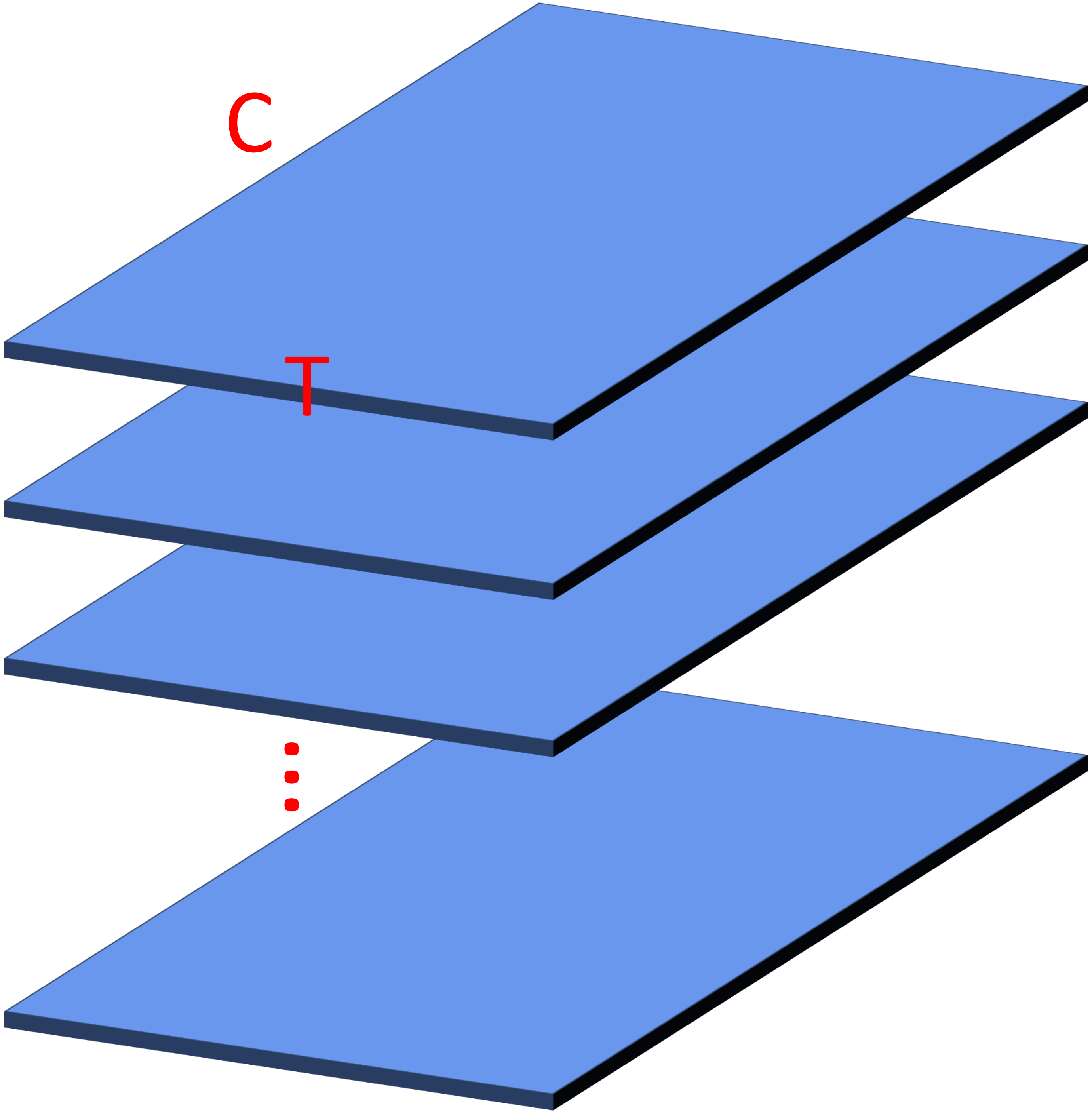


C = head_size * num_heads


T
(32, 64)
C

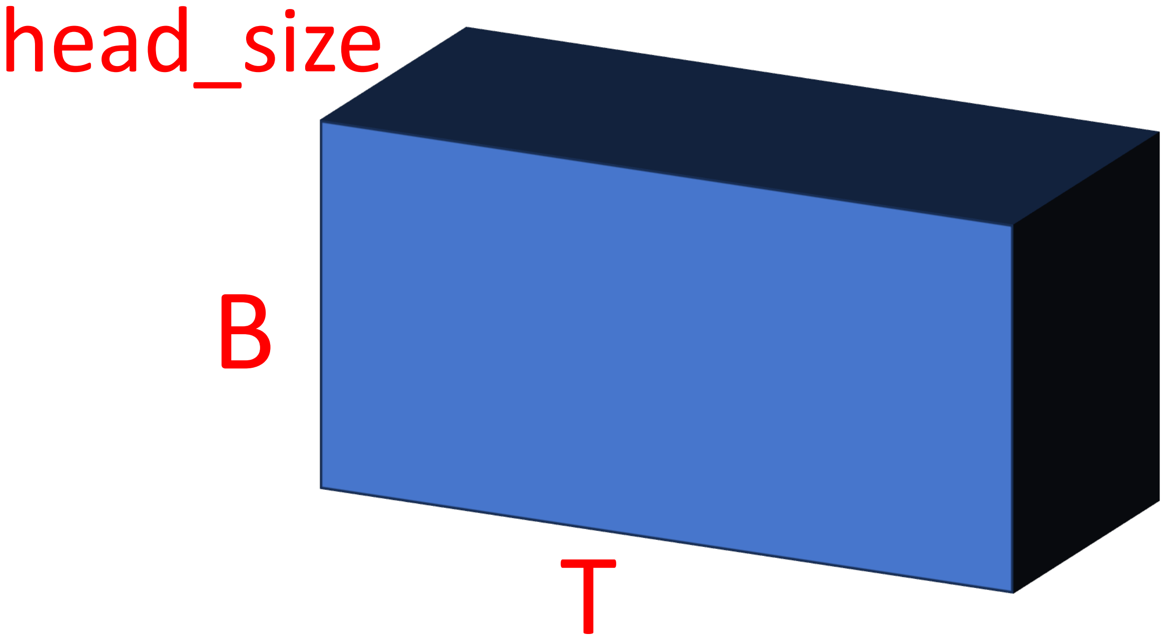


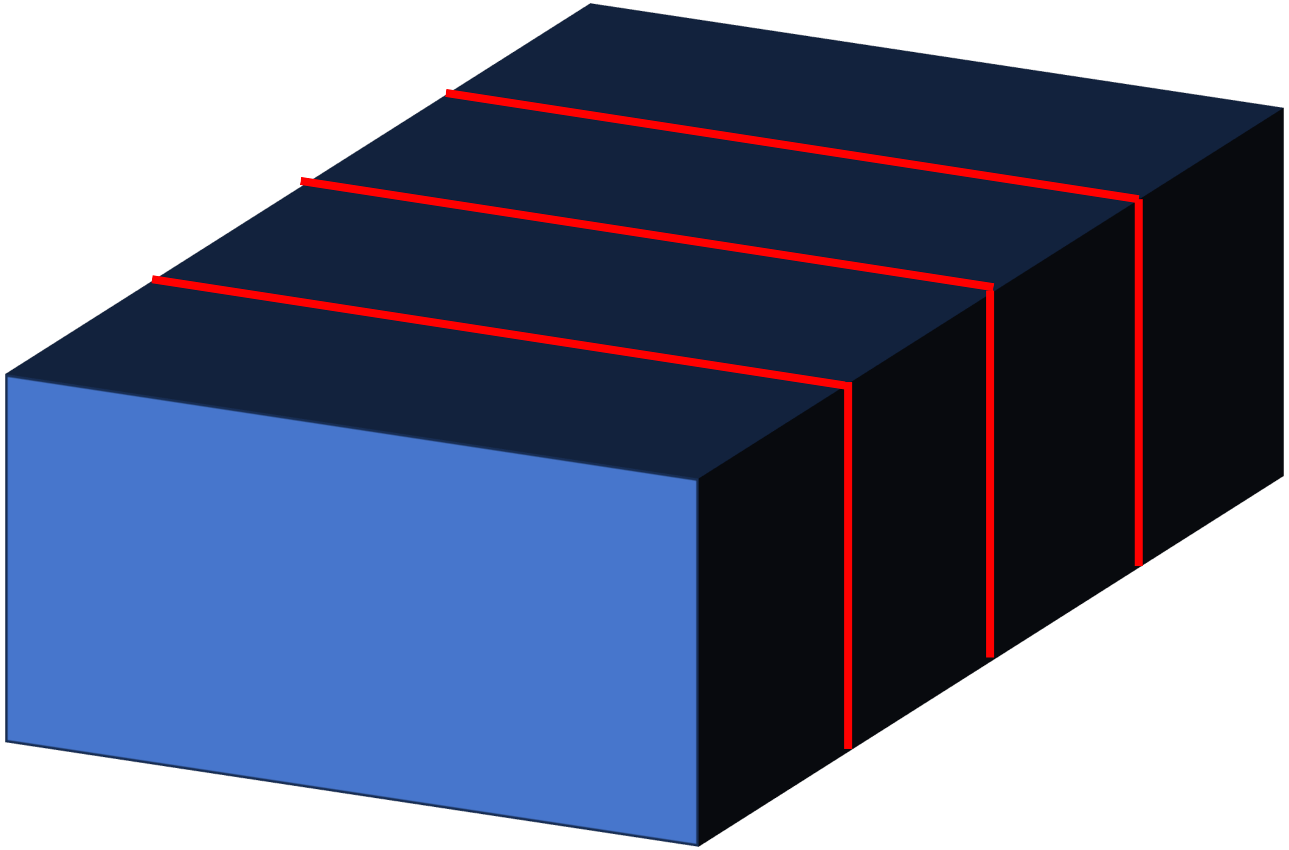
$$\begin{array}{c}
\left[\begin{array}{ccc|ccc|ccc|ccc}
a^{h1}_{chr1_{v1}} & \dots & a^{h1}_{chr1_{v16}} & a^{h2}_{chr1_{v1}} & \dots & a^{h2}_{chr1_{v16}} & a^{h3}_{chr1_{v1}} & \dots & a^{h3}_{chr1_{v16}} & a^{h4}_{chr1_{v1}} & \dots & a^{h4}_{chr1_{v16}} \\
a^{h1}_{chr2_{v1}} & \dots & a^{h1}_{chr2_{v16}} & a^{h2}_{chr2_{v1}} & \dots & a^{h2}_{chr2_{v16}} & a^{h3}_{chr2_{v1}} & \dots & a^{h3}_{chr2_{v16}} & a^{h4}_{chr2_{v1}} & \dots & a^{h4}_{chr2_{v16}} \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
a^{h1}_{chr32_{v1}} & \dots & a^{h1}_{chr32_{v16}} & a^{h2}_{chr32_{v1}} & \dots & a^{h2}_{chr32_{v16}} & a^{h3}_{chr32_{v1}} & \dots & a^{h3}_{chr32_{v16}} & a^{h4}_{chr32_{v1}} & \dots & a^{h4}_{chr32_{v16}}.
\end{array}\right] \\
\begin{array}{cccc}
\text{HEAD 1} & \text{HEAD 2} & \text{HEAD 3} & \text{HEAD 4}
\end{array}
\end{array}$$

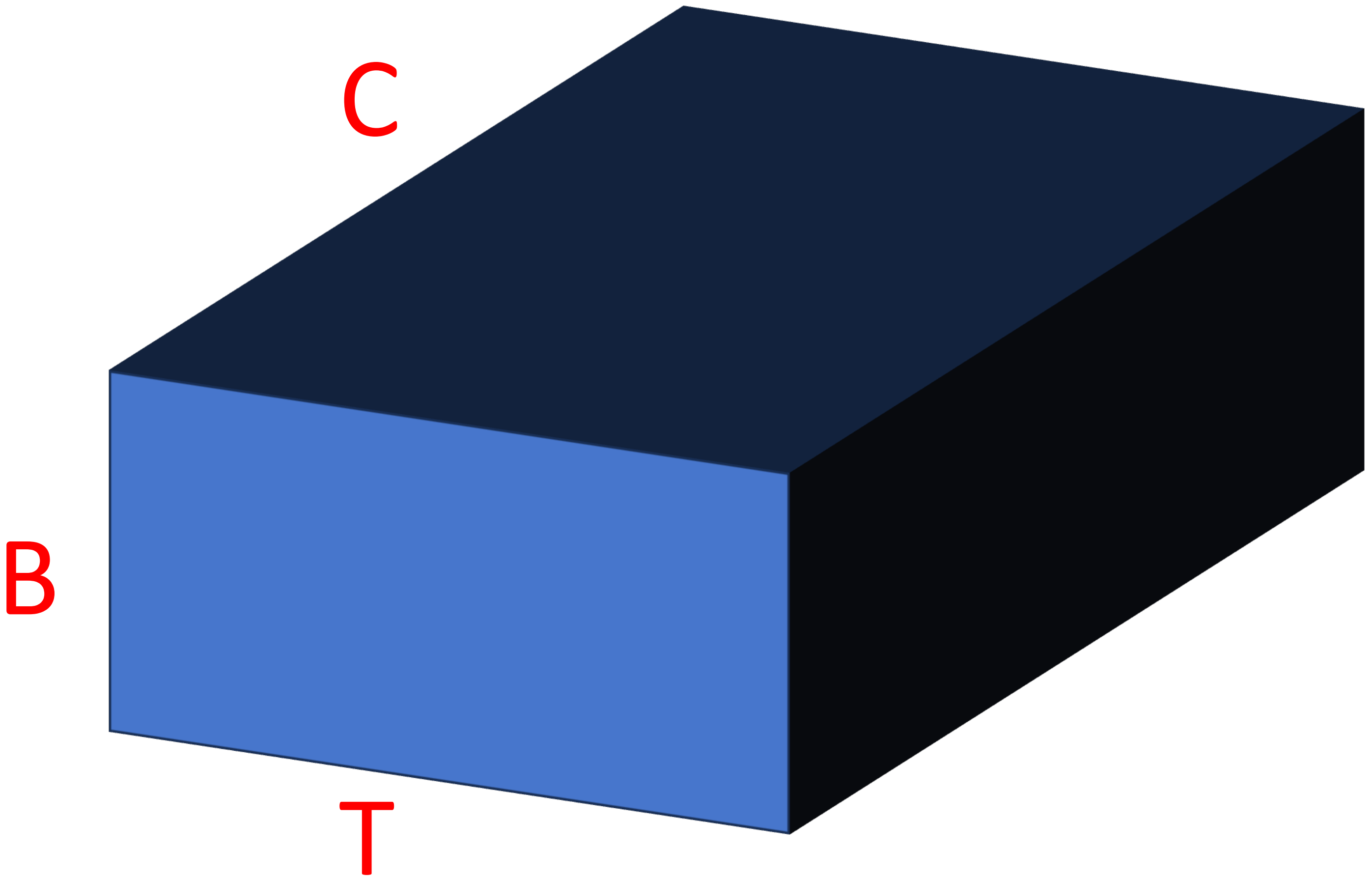


16 slates stacked together

each slate is an output from a multi-head layer and is a matrix of shape (T, C)


head_size
B
T




C
B
T