# 各式各樣的 Attention

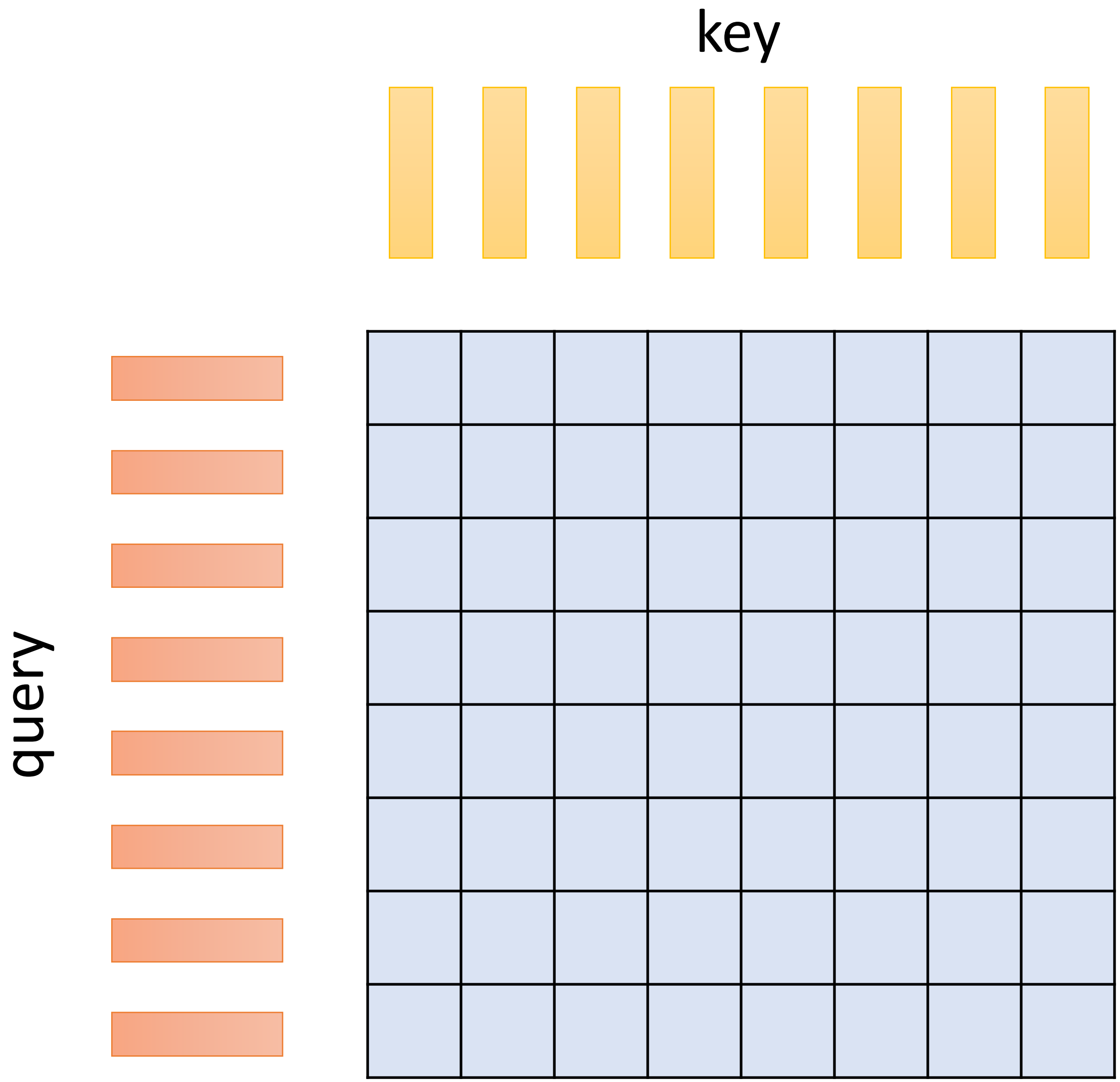
Hung-yi Lee 李宏毅

# Prerequisite

https://youtu.be/hYdO9CscNes

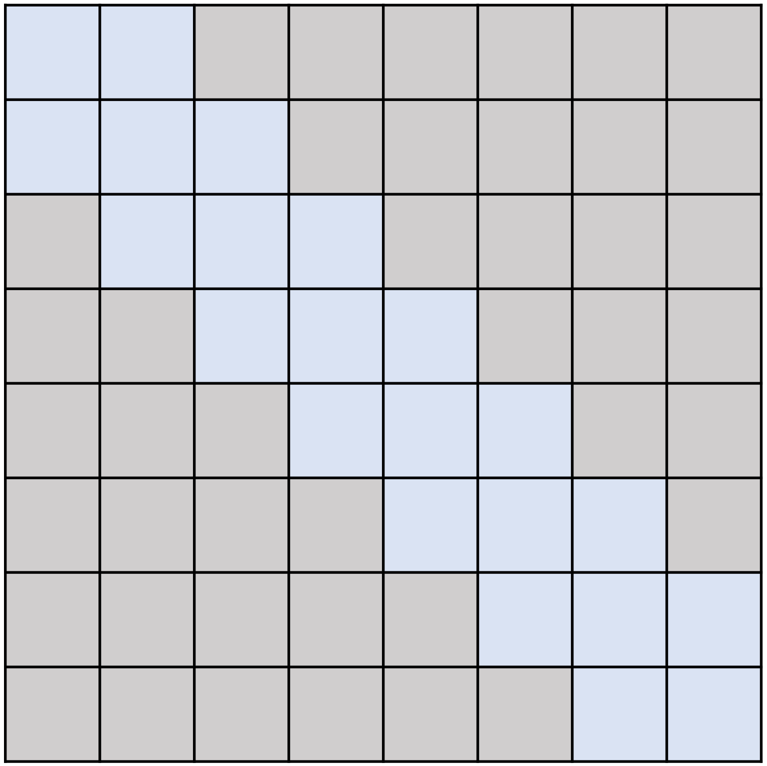
https://youtu.be/gmsMY5kc-zw

# Review

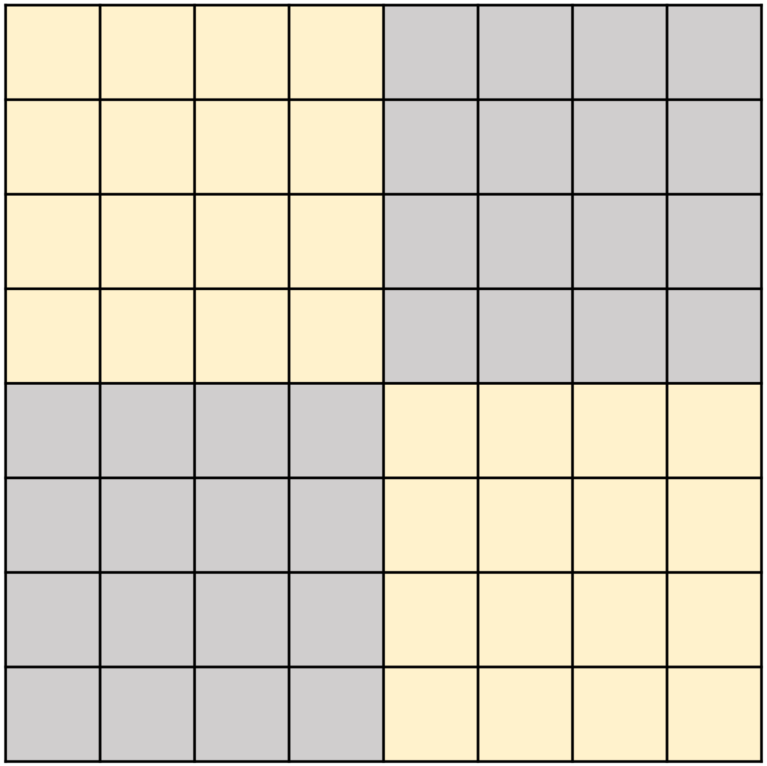



*Attention Matrix*

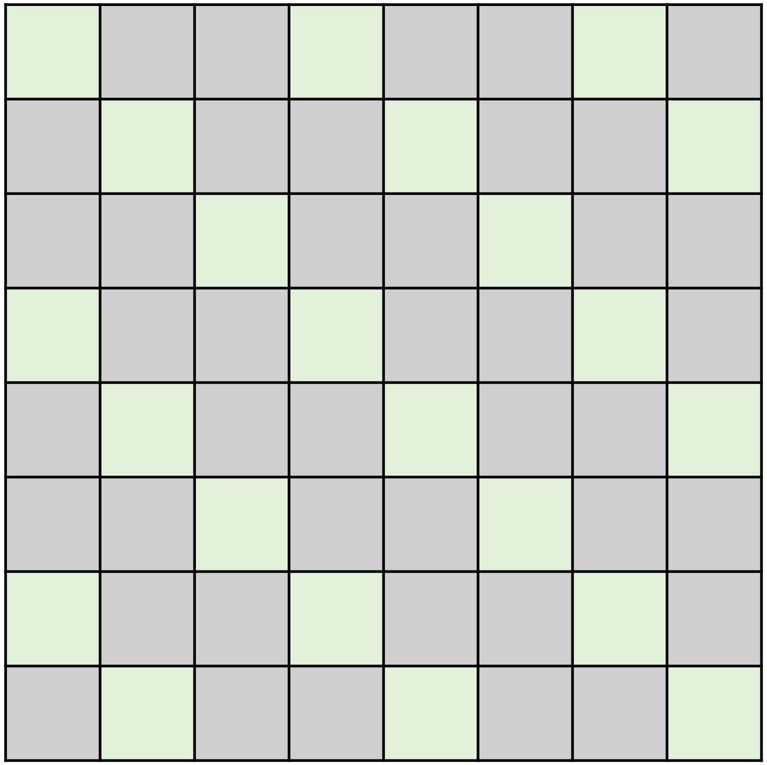
# Fixed Pattern

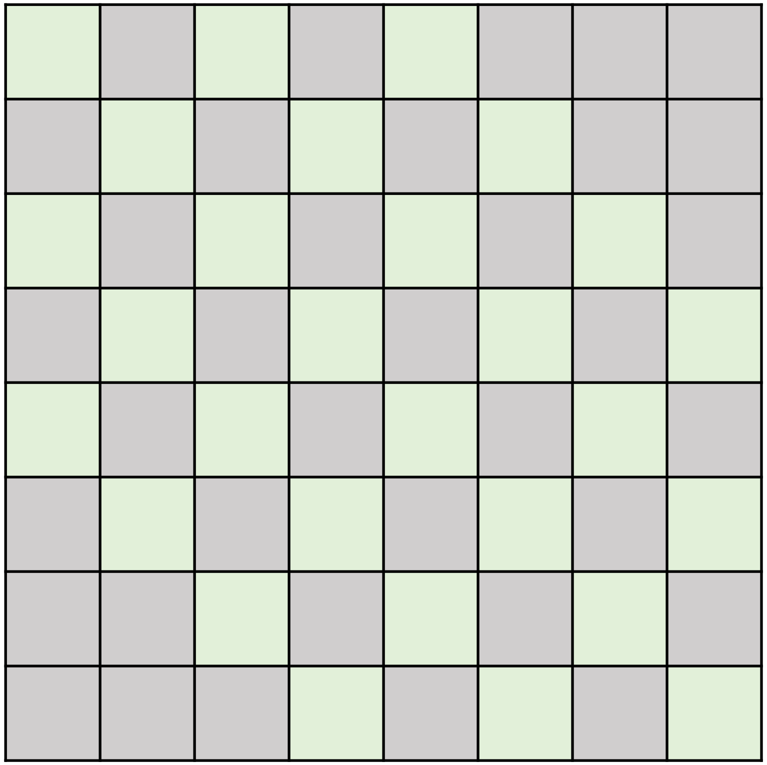


c.f. CNN

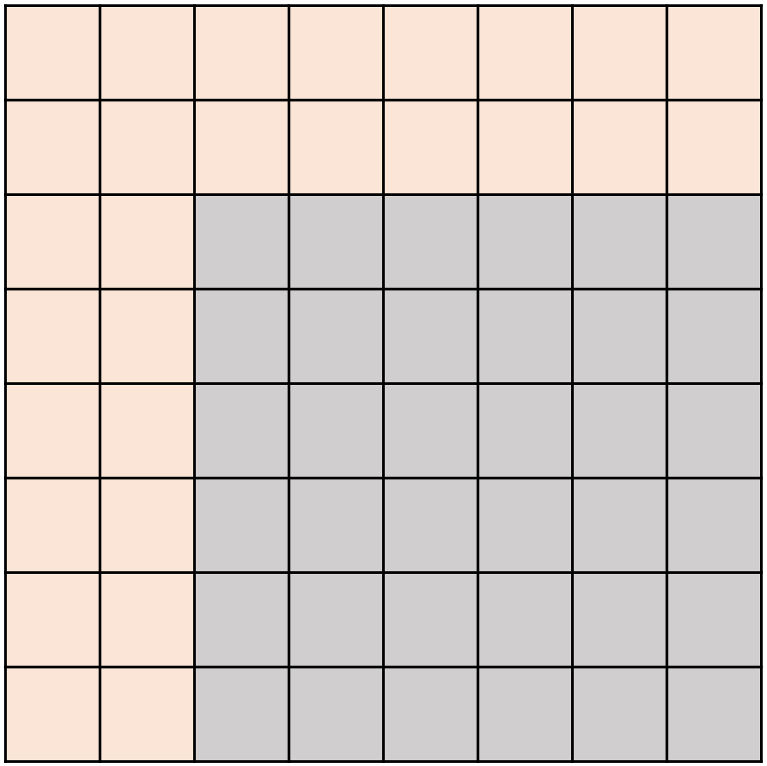


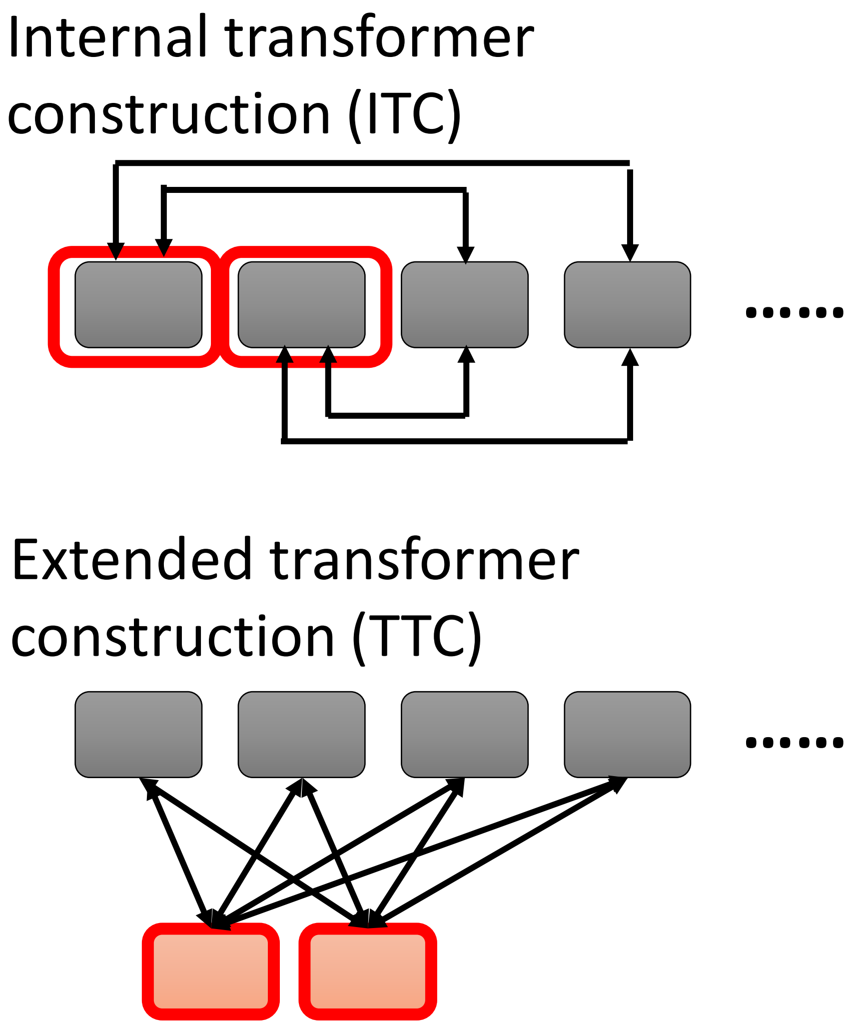
# Fixed Pattern



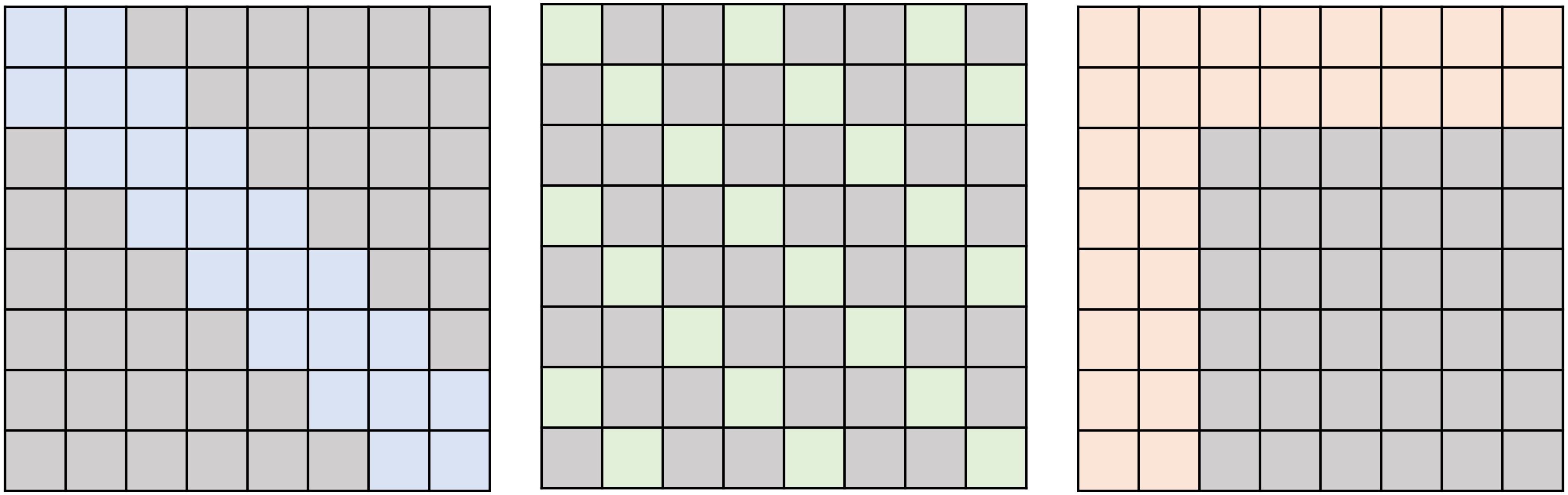


# Fixed Pattern





# Fixed Pattern

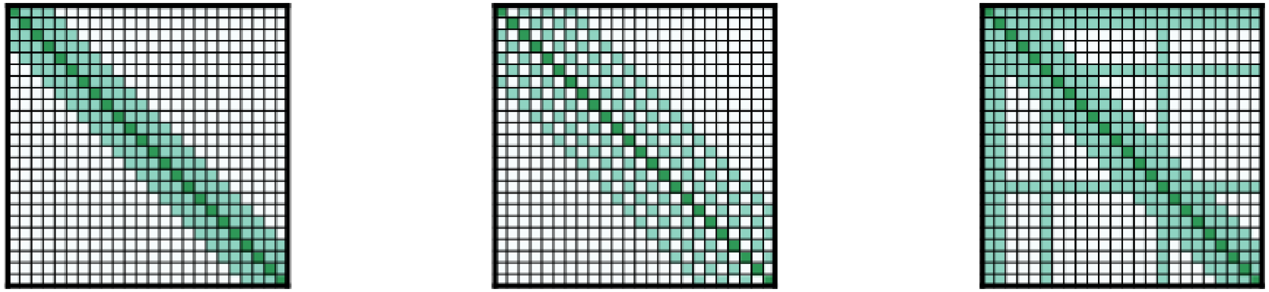


Different heads use different patterns.

# Fixed Pattern

- Longformer

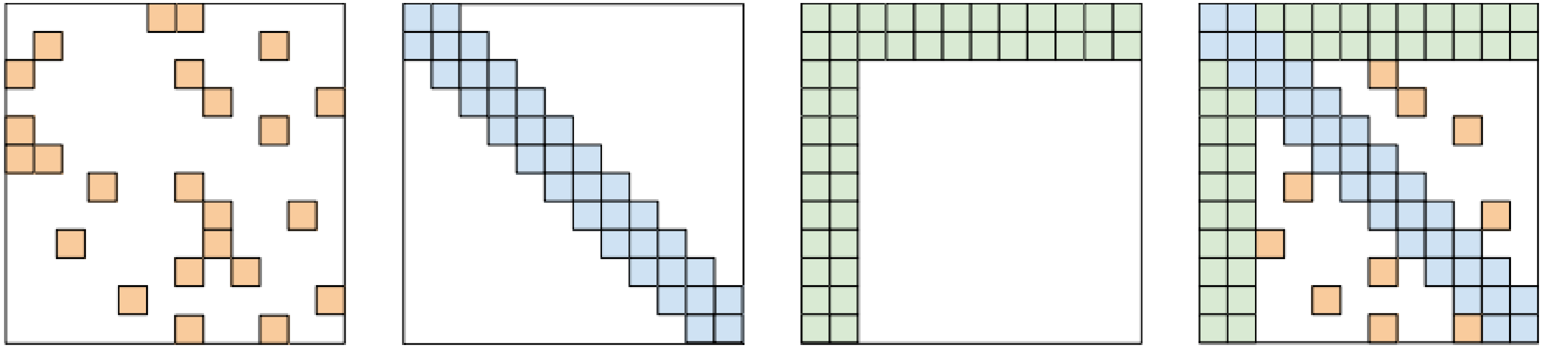
https://arxiv.org/abs/2004.05150



(b) Sliding window attention (c) Dilated sliding window (d) Global+sliding window

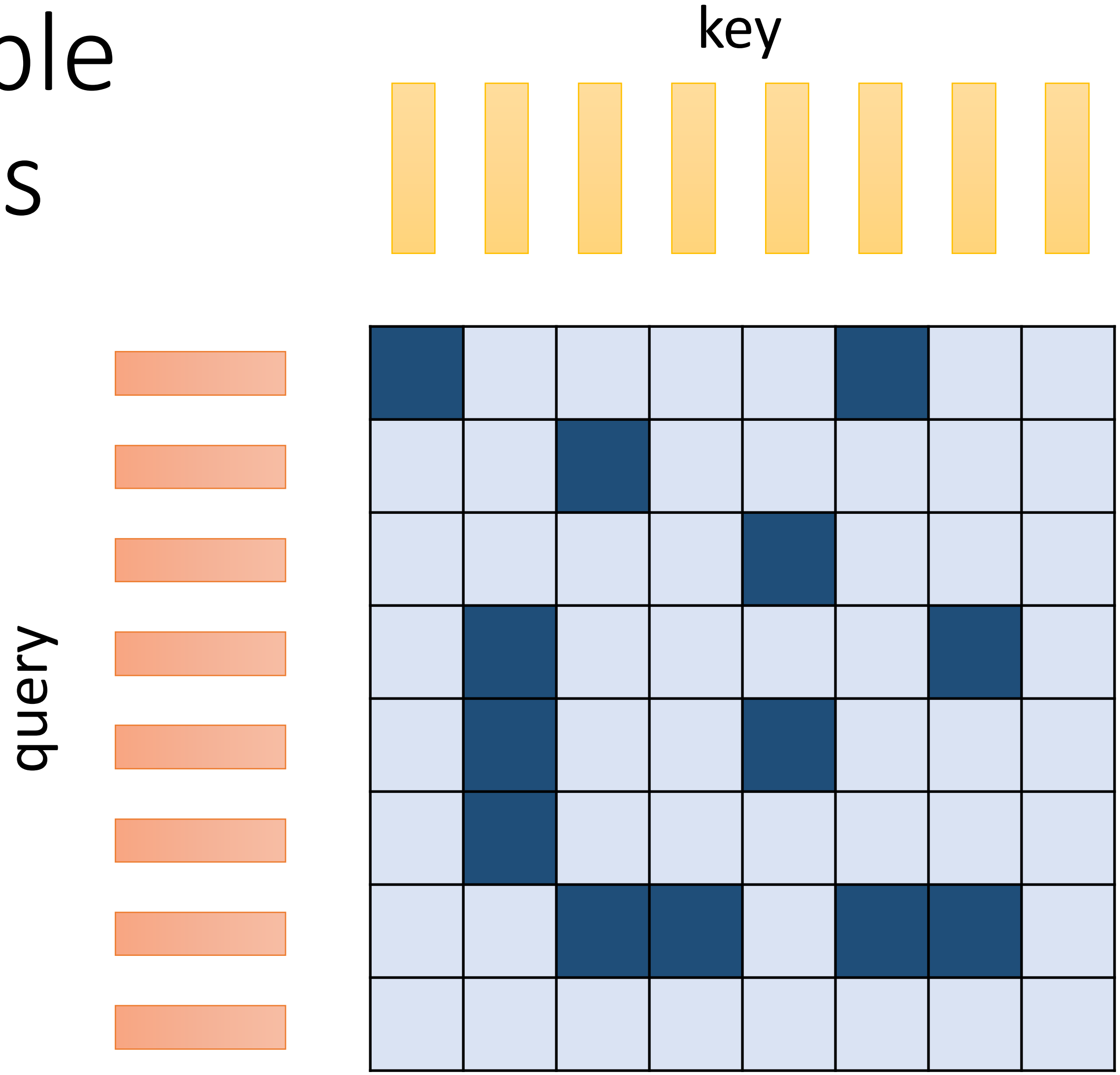
- Big Bird

https://arxiv.org/abs/2007.14062



(a) Random attention (b) Window attention (c) Global Attention (d) BIGBIRD

# Learnable Patterns



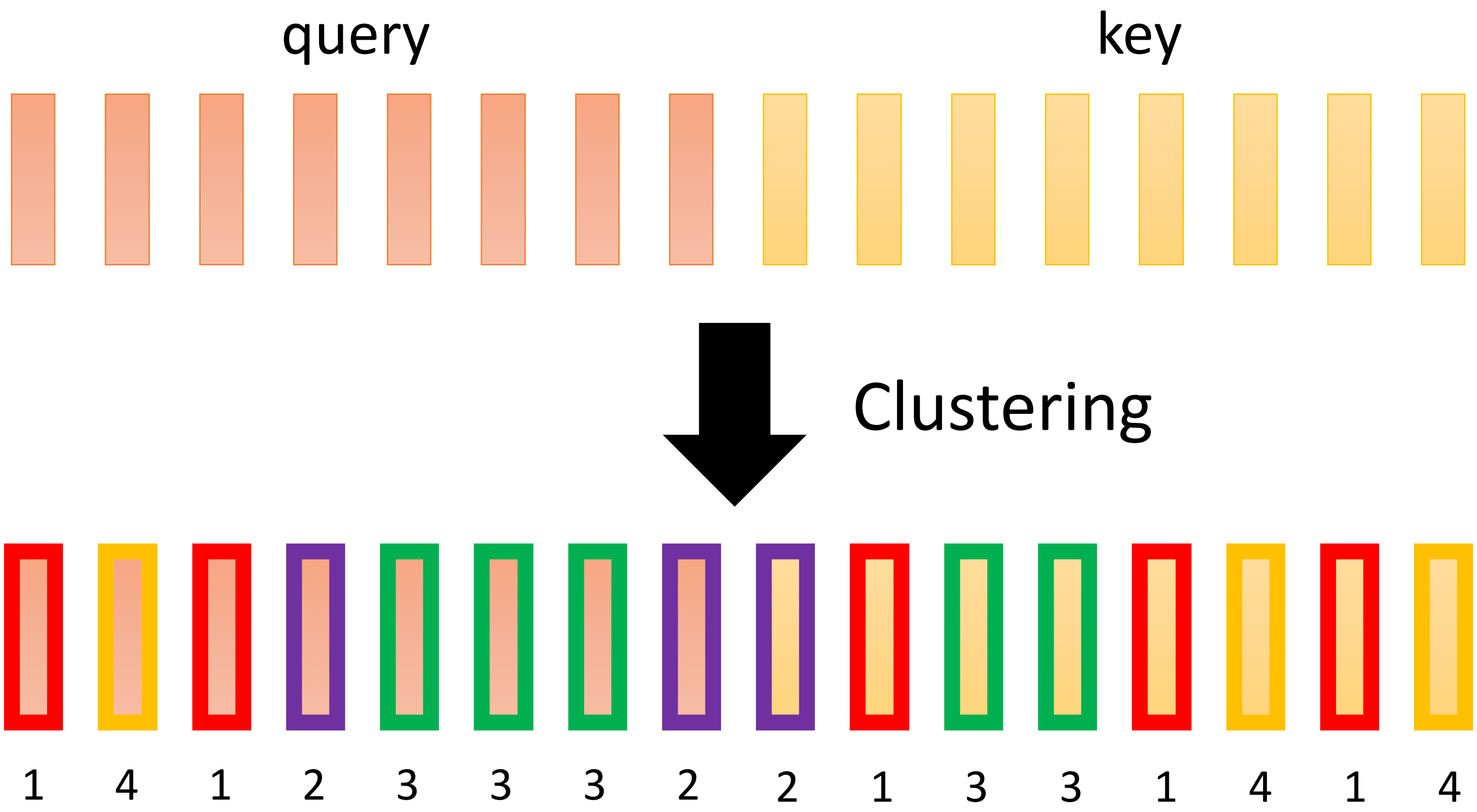
# Learnable Patterns

## Reformer

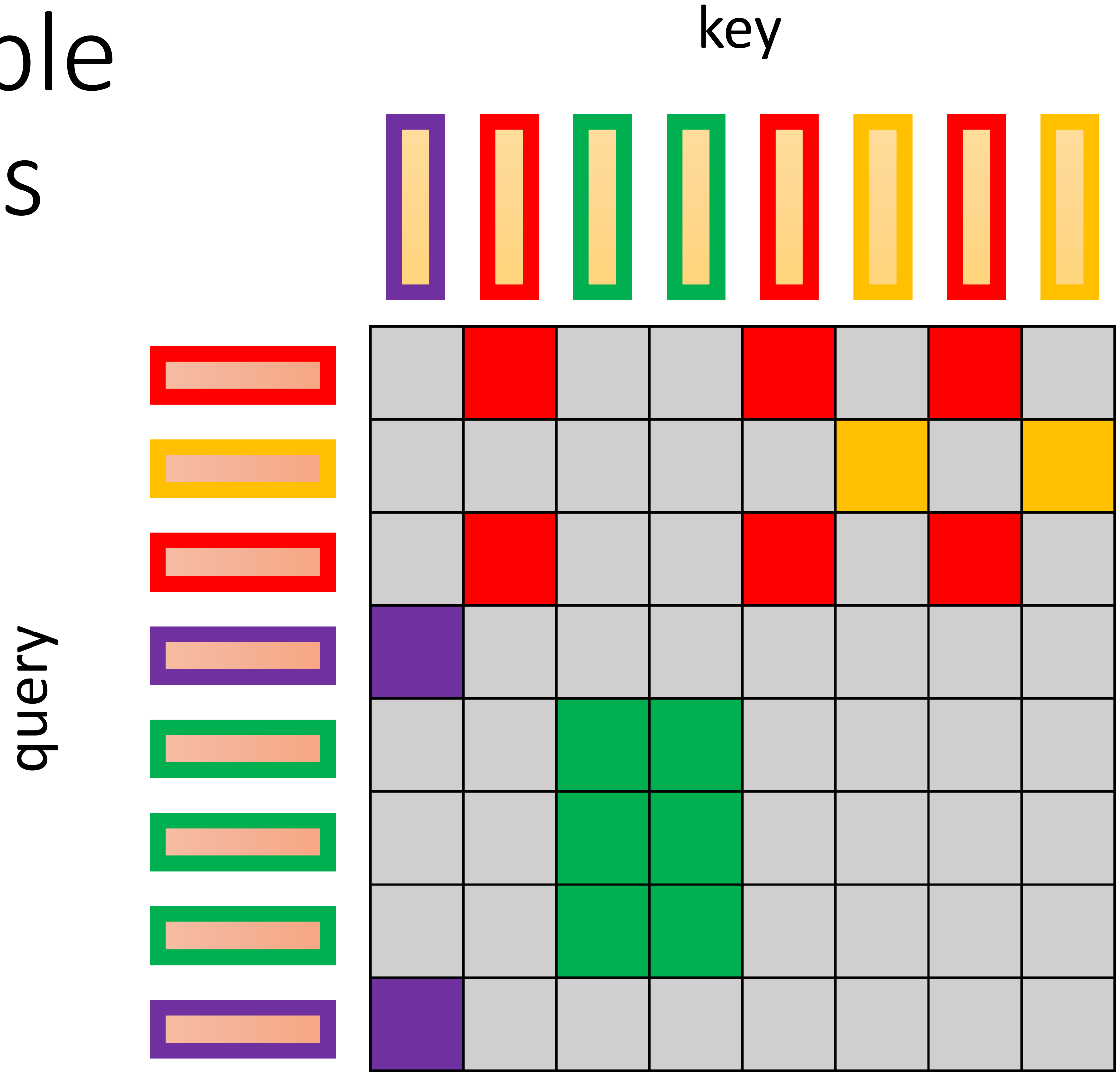
https://openreview.net/forum?id=rkgNKkHtvB

## Routing Transformer

https://arxiv.org/abs/2003.05997



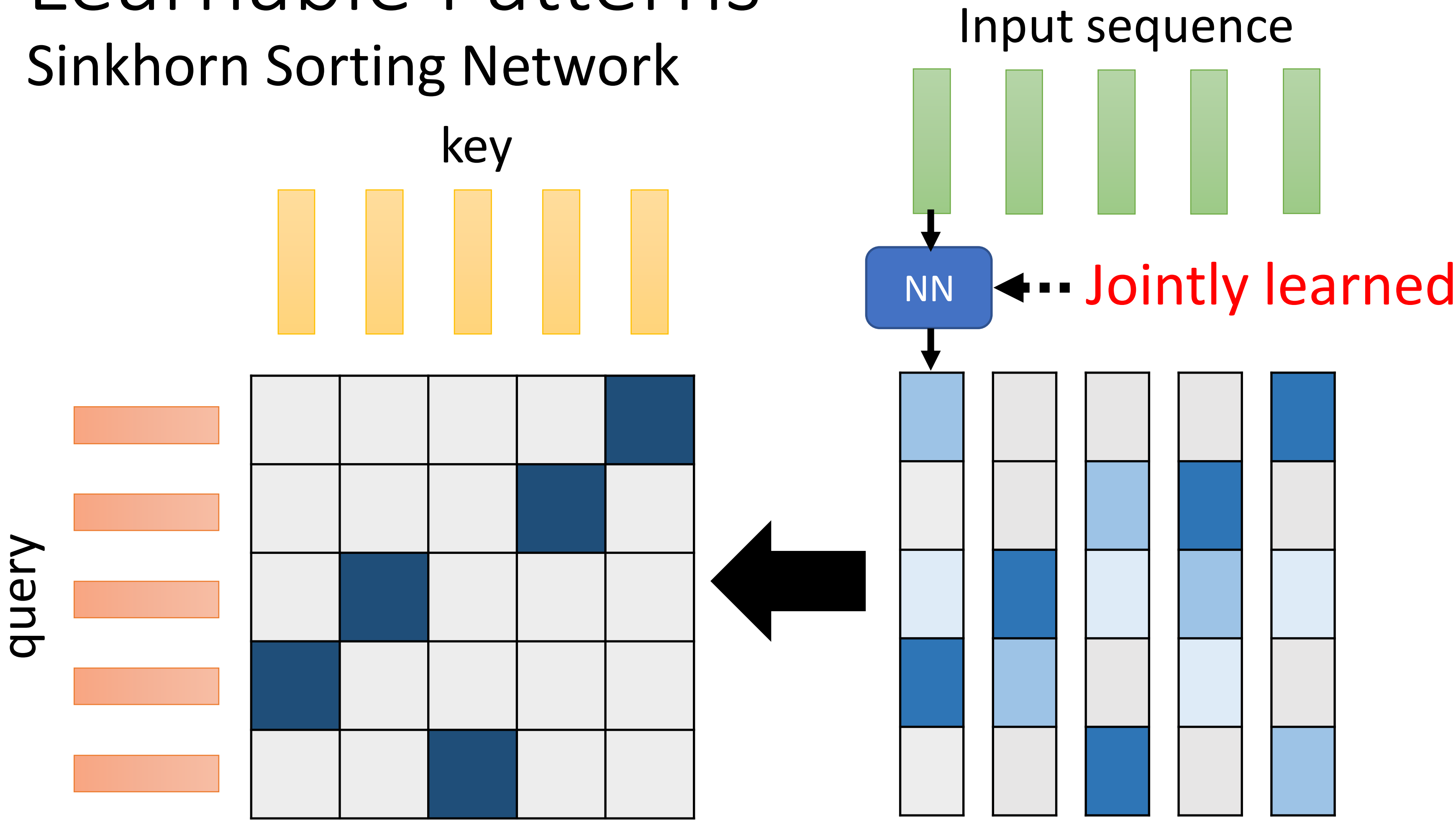
# Learnable Patterns



https://arxiv.org/abs/2002.11296

# Learnable Patterns

## Sinkhorn Sorting Network

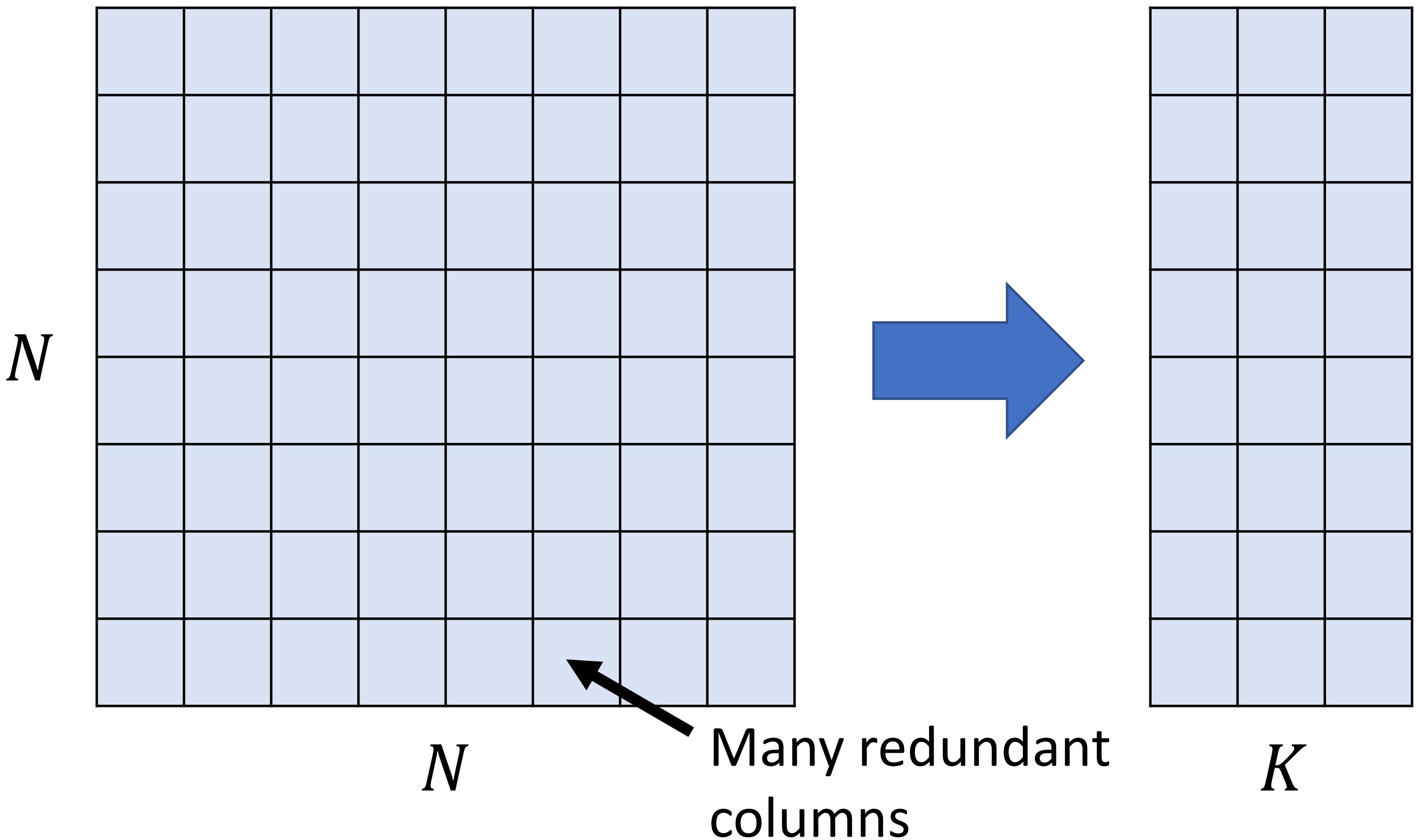


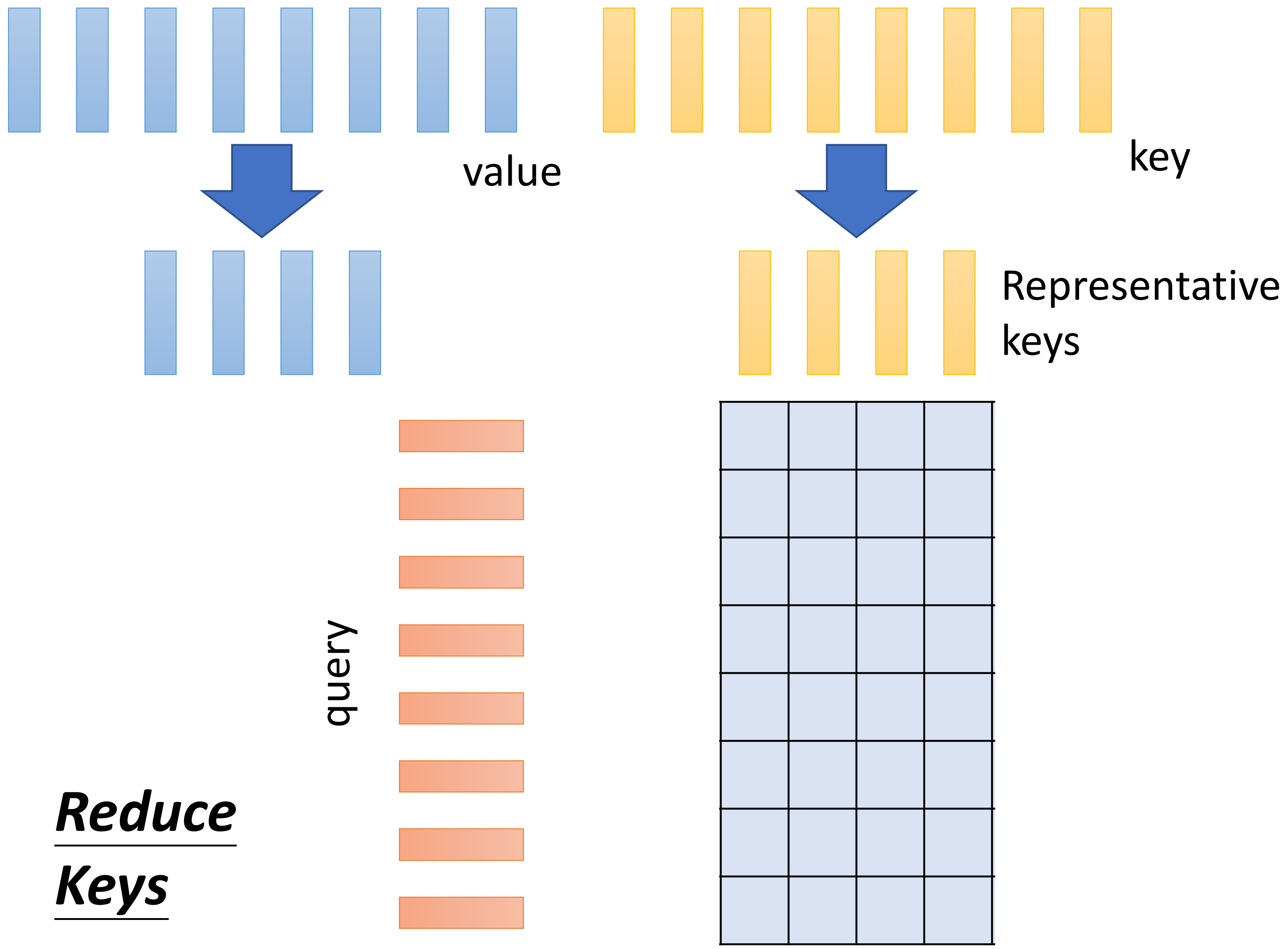

(simplified version)

# Do we need full attention matrix?

Linformer

https://arxiv.org/abs/2006.04768

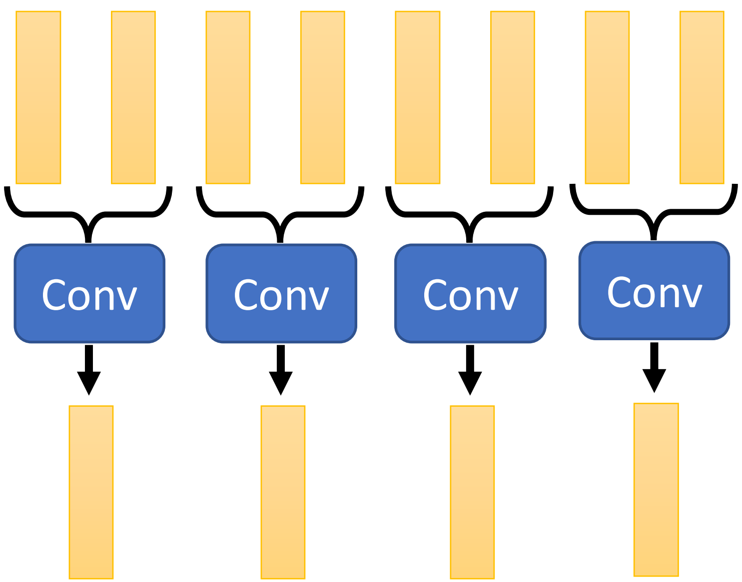



value
key
Representative
keys
query
Reduce
Keys

# Reduce Number of Keys

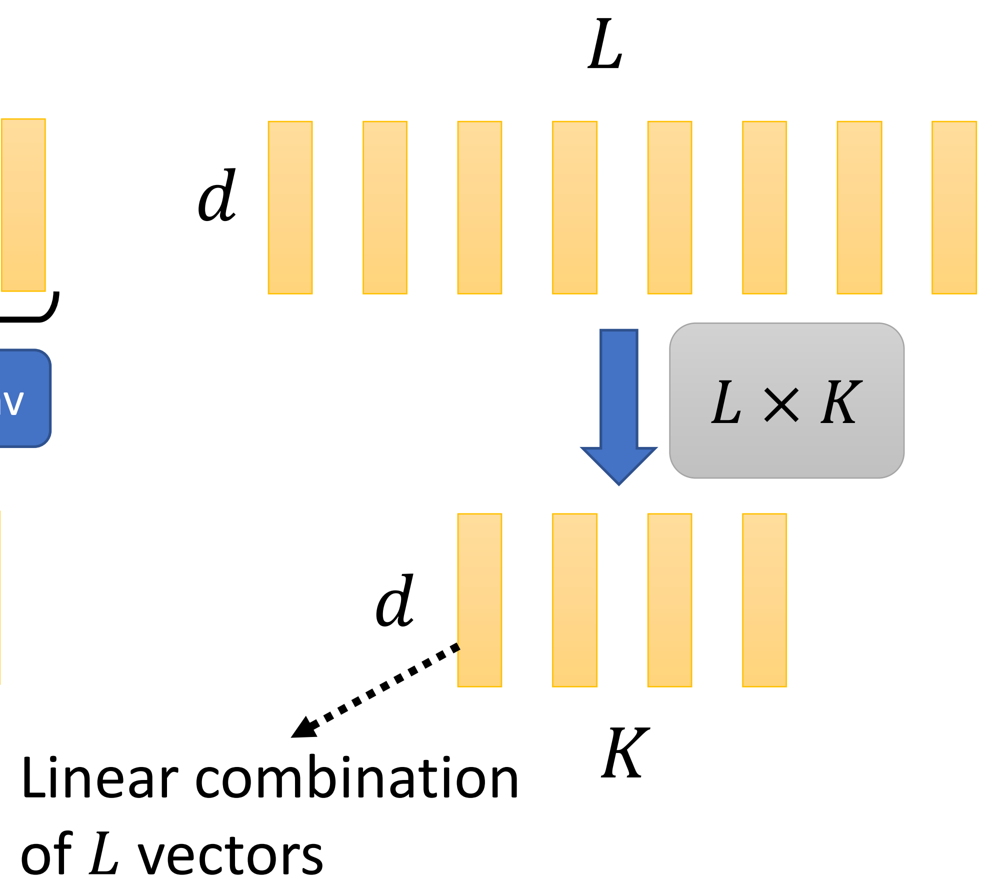
## Compressed Attention

https://arxiv.org/abs/1801.10198



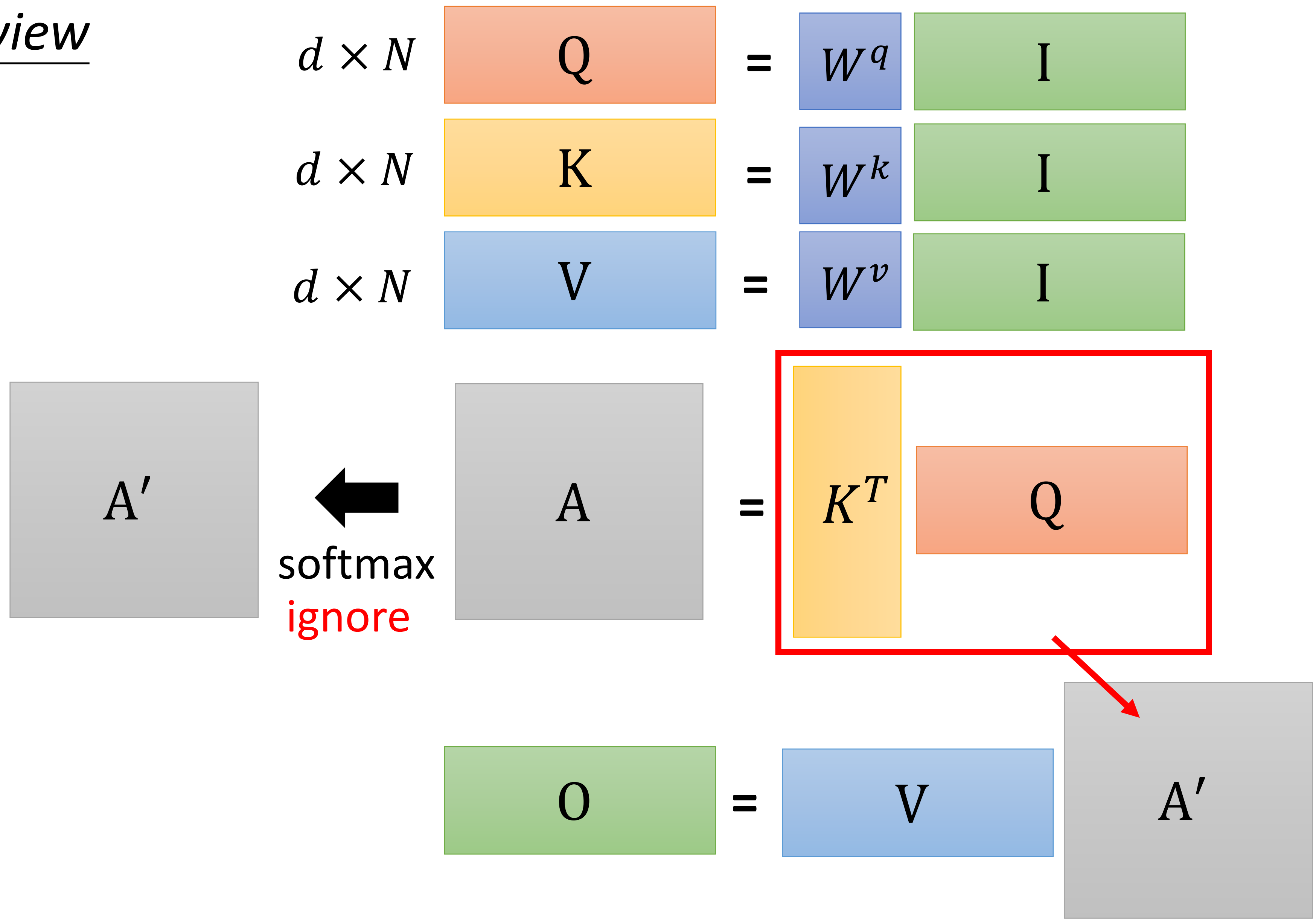

## Linformer

https://arxiv.org/abs/2006.04768



# *Attention Mechanism is three-matrix Multiplication*

*Review*



# Attention Mechanism is three-matrix Multiplication

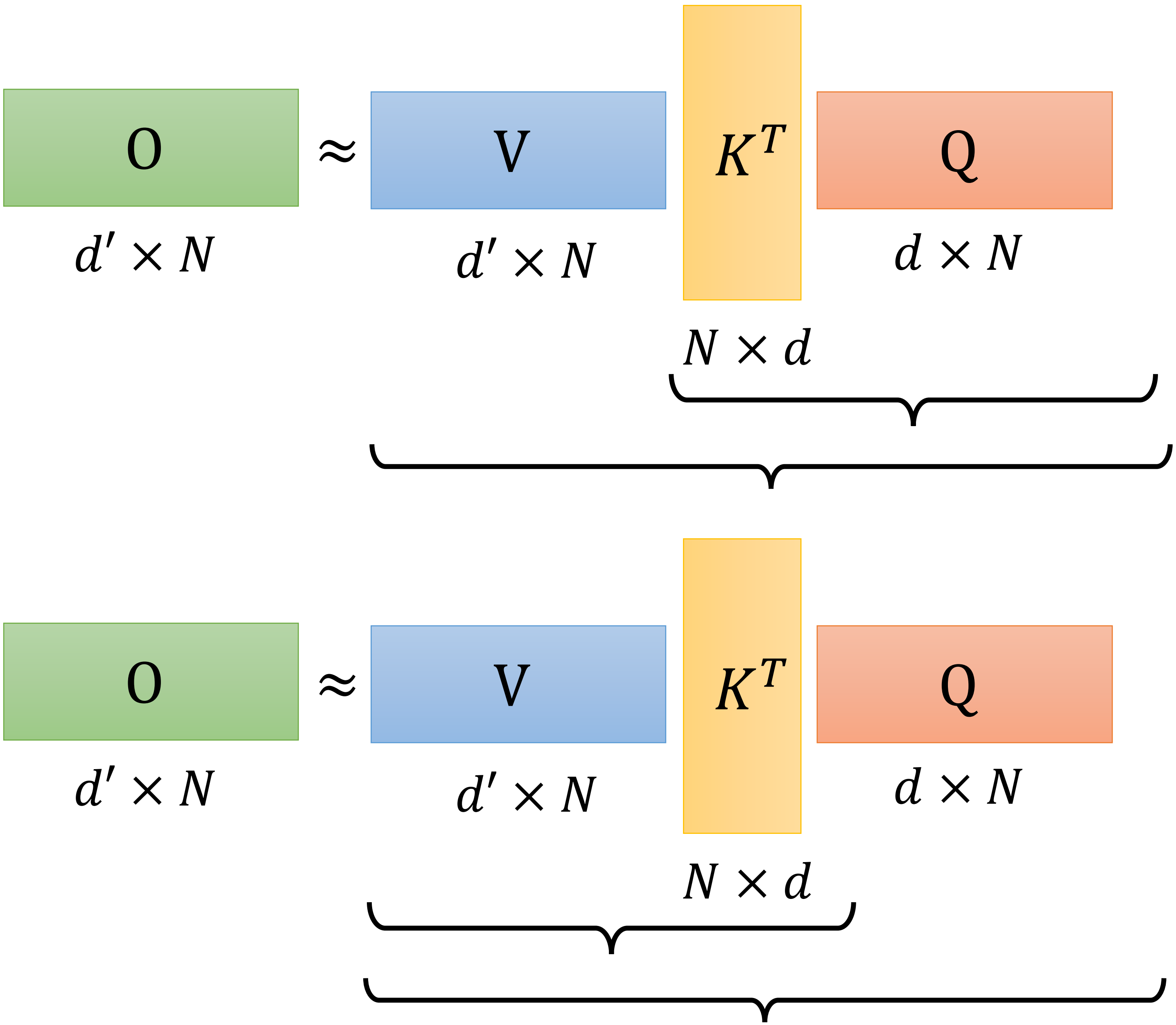
*Review*

$$\underset{d \times N}{Q} = W^q \, I$$

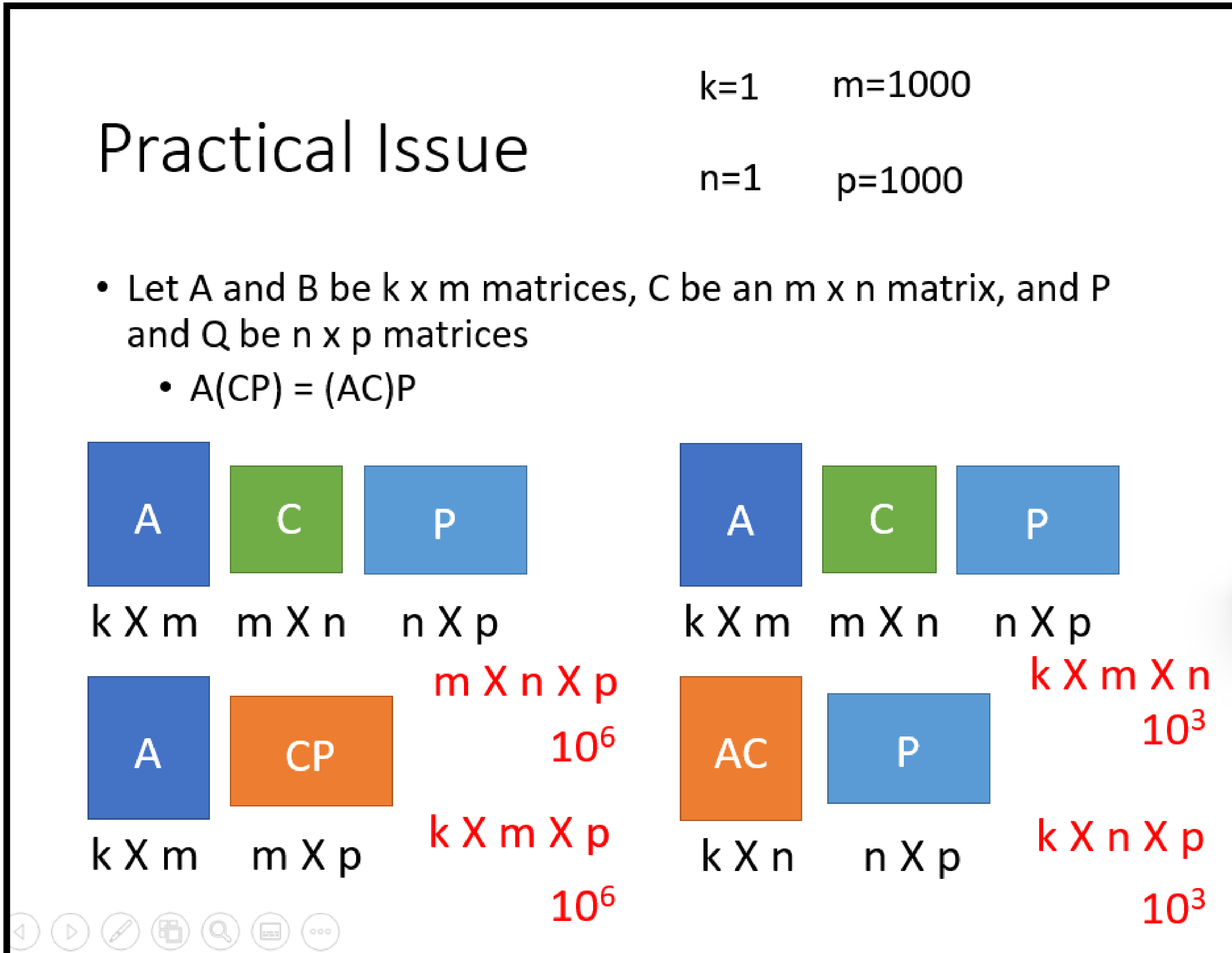
$$\underset{d \times N}{K} = W^k \, I$$

$$\underset{d' \times N}{V} = W^v \, I$$

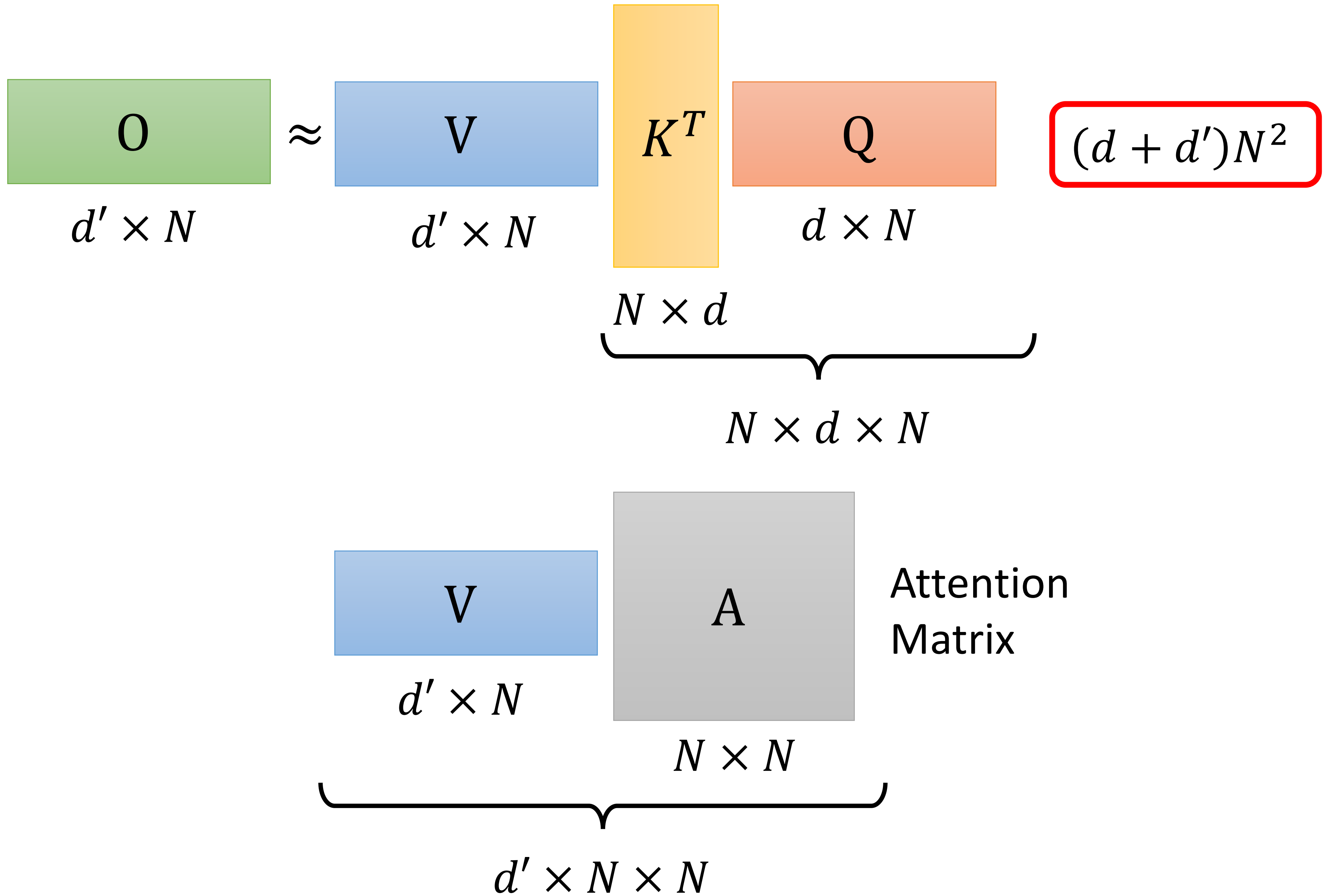
$$\underset{d' \times N}{O} \approx \underbrace{\underset{d' \times N}{V} \; \underbrace{\underset{N \times d}{K^T} \; \underset{d \times N}{Q}}}$$


O
≈
V
$K^T$
Q
$d' \times N$
$d' \times N$
$d \times N$
$N \times d$
O
≈
V
$K^T$
Q
$d' \times N$
$d' \times N$
$d \times N$
$N \times d$

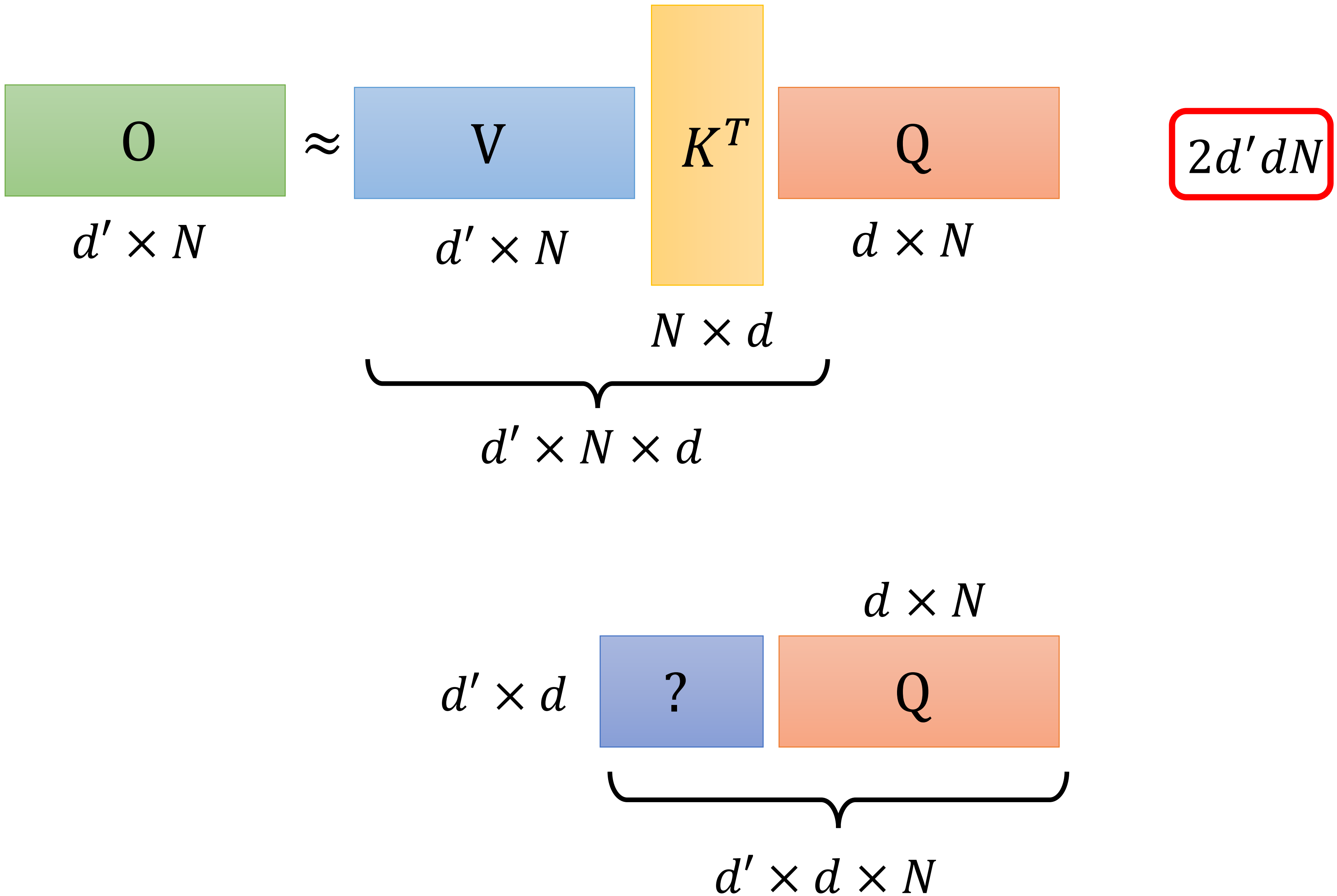
# *Review Linear Algebra*

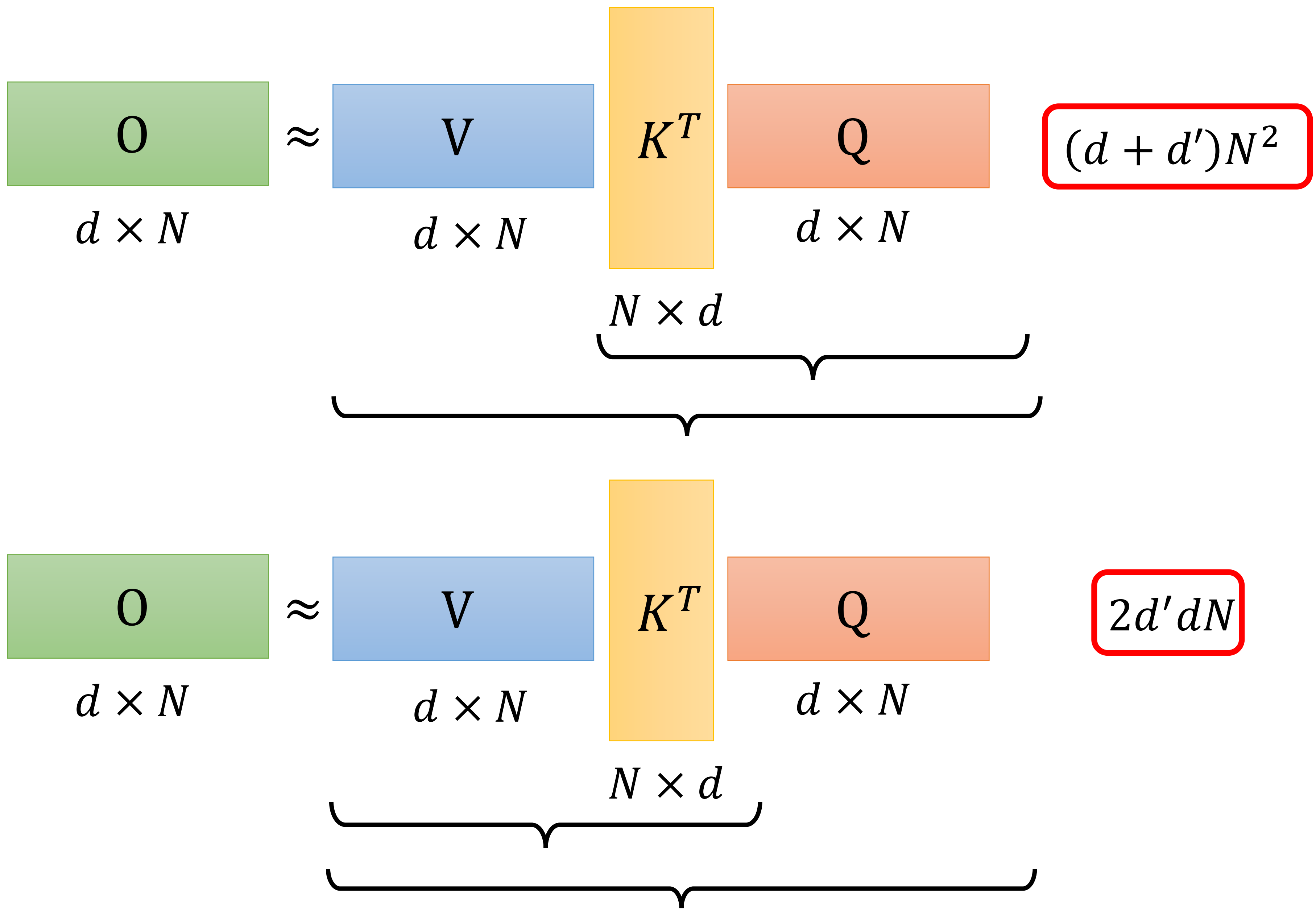



https://youtu.be/yO8lDzf4jMs



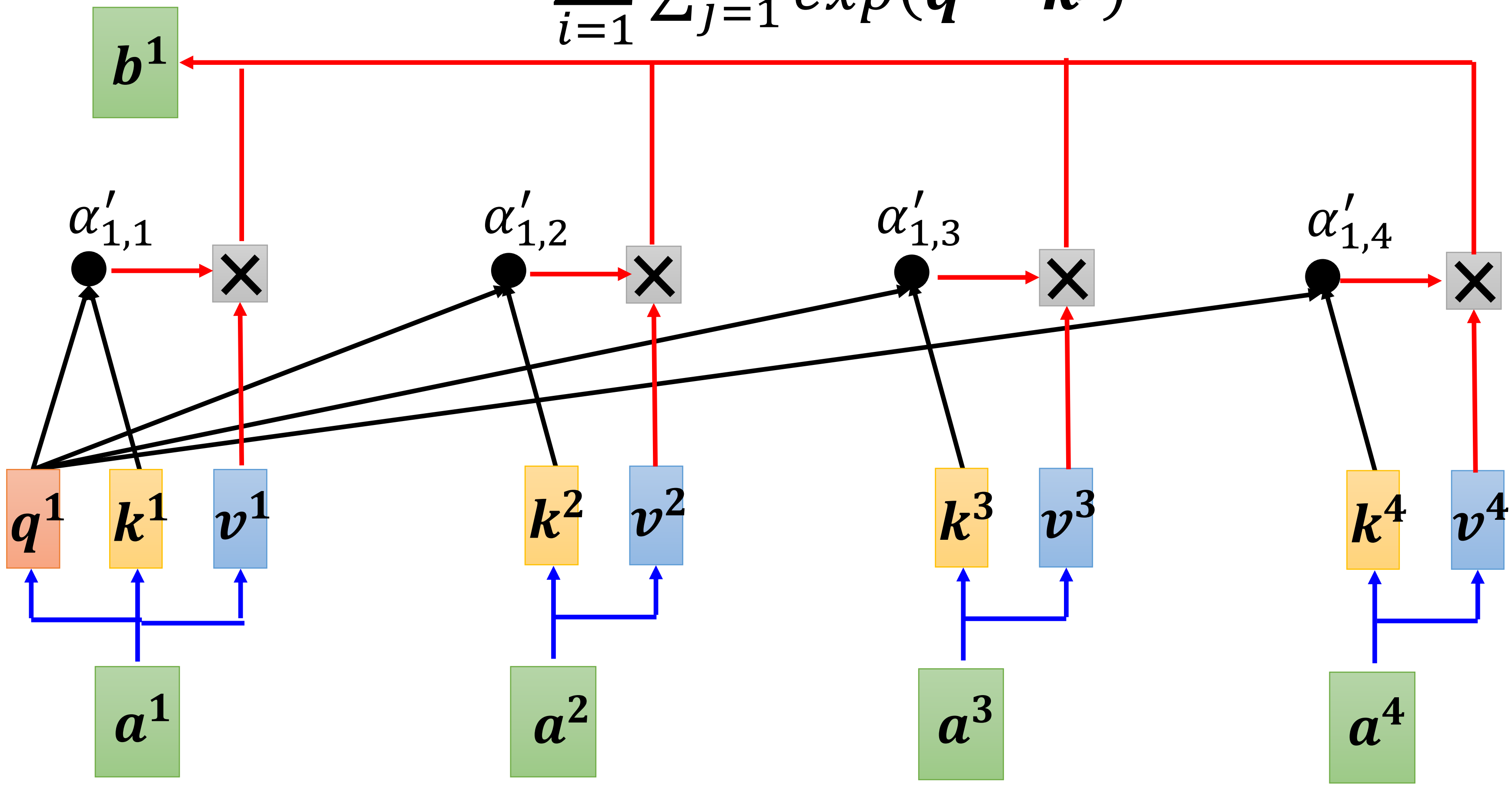
O
≈
V
$K^T$
Q
$(d + d')N^2$
$d' \times N$
$d' \times N$
$d \times N$
$N \times d$
$N \times d \times N$
V
A
Attention
Matrix
$d' \times N$
$N \times N$
$d' \times N \times N$


O
≈
V
$K^T$
Q
$2d'dN$
$d' \times N$
$d' \times N$
$d \times N$
$N \times d$
$d' \times N \times d$
$d \times N$
$d' \times d$
?
Q
$d' \times d \times N$


O
≈
V
$K^T$
Q
$(d + d')N^2$
$d \times N$
$d \times N$
$d \times N$
$N \times d$
O
≈
V
$K^T$
Q
$2d'dN$
$d \times N$
$d \times N$
$d \times N$
$N \times d$

$$\boldsymbol{b}^{\mathbf{1}} = \sum_{i=1}^{N} \alpha'_{1,i}\boldsymbol{v}^{\boldsymbol{i}} = \sum_{i=1}^{N} \frac{exp(\alpha_{1,i})}{\sum_{j=1}^{\mathrm{N}} exp(\alpha_{1,j})}\boldsymbol{v}^{\boldsymbol{i}}$$

$$= \sum_{i=1}^{N} \frac{exp(\boldsymbol{q}^{\mathbf{1}} \cdot \boldsymbol{k}^{\boldsymbol{i}})}{\sum_{j=1}^{\mathrm{N}} exp(\boldsymbol{q}^{\mathbf{1}} \cdot \boldsymbol{k}^{\boldsymbol{j}})}\boldsymbol{v}^{\boldsymbol{i}}$$

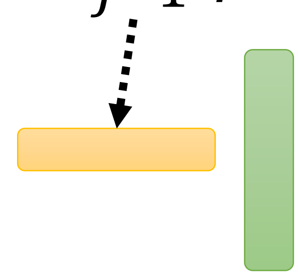


$$\boldsymbol{b}^1 = \sum_{i=1}^{N} \alpha'_{1,i} \boldsymbol{v}^i = \sum_{i=1}^{N} \frac{exp(\alpha_{1,i})}{\sum_{j=1}^{N} exp(\alpha_{1,j})} \boldsymbol{v}^i$$

$$\boldsymbol{q} \rightarrow \boxed{\phi} \rightarrow \phi(\boldsymbol{q}) \qquad = \sum_{i=1}^{N} \frac{exp(\boldsymbol{q}^1 \cdot \boldsymbol{k}^i)}{\sum_{j=1}^{N} exp(\boldsymbol{q}^1 \cdot \boldsymbol{k}^j)} \boldsymbol{v}^i$$

$$\begin{aligned} &exp(\boldsymbol{q} \cdot \boldsymbol{k}) \\ &\approx \phi(\boldsymbol{q}) \cdot \phi(\boldsymbol{k}) \end{aligned} \qquad = \sum_{i=1}^{N} \frac{\phi(\boldsymbol{q}^1) \cdot \phi(\boldsymbol{k}^i)}{\sum_{j=1}^{N} \phi(\boldsymbol{q}^1) \cdot \phi(\boldsymbol{k}^j)} \boldsymbol{v}^i$$

$$= \frac{\sum_{i=1}^{N} [\phi(\boldsymbol{q}^1) \cdot \phi(\boldsymbol{k}^i)] \boldsymbol{v}^i}{\sum_{j=1}^{N} \phi(\boldsymbol{q}^1) \cdot \phi(\boldsymbol{k}^j)} = \frac{\sum_{i=1}^{N} [\phi(\boldsymbol{q}^1) \cdot \phi(\boldsymbol{k}^i)] \boldsymbol{v}^i}{\phi(\boldsymbol{q}^1) \cdot \sum_{j=1}^{N} \phi(\boldsymbol{k}^j)}$$



$\phi(\boldsymbol{q}^1)$

$$\boldsymbol{b}^{\mathbf{1}} = \sum_{i=1}^{N} \alpha'_{1,i}\boldsymbol{v}^{\boldsymbol{i}} = \frac{\boxed{\sum_{i=1}^{N}\left[\phi\left(\boldsymbol{q}^{\mathbf{1}}\right)\cdot\phi\left(\boldsymbol{k}^{\boldsymbol{i}}\right)\right]\boldsymbol{v}^{\boldsymbol{i}}}}{\phi\left(\boldsymbol{q}^{\mathbf{1}}\right)\cdot\sum_{j=1}^{N}\phi\left(\boldsymbol{k}^{\boldsymbol{j}}\right)}$$

$$\boxed{\sum_{i=1}^{N}\left[\phi\left(\boldsymbol{q}^{\mathbf{1}}\right)\cdot\phi\left(\boldsymbol{k}^{\boldsymbol{i}}\right)\right]\boldsymbol{v}^{\boldsymbol{i}}} \qquad \phi\left(\boldsymbol{q}^{\mathbf{1}}\right) = \begin{bmatrix} q_1^1 \\ q_2^1 \\ \vdots \end{bmatrix} \qquad \phi\left(\boldsymbol{k}^{\mathbf{1}}\right) = \begin{bmatrix} k_1^1 \\ k_2^1 \\ \vdots \end{bmatrix}$$

$$= \left[\phi\left(\boldsymbol{q}^{\mathbf{1}}\right)\cdot\phi\left(\boldsymbol{k}^{\mathbf{1}}\right)\right]\boldsymbol{v}^{\mathbf{1}} + \left[\phi\left(\boldsymbol{q}^{\mathbf{1}}\right)\cdot\phi\left(\boldsymbol{k}^{\mathbf{2}}\right)\right]\boldsymbol{v}^{\mathbf{2}} + \cdots$$

$$= (q_1^1 k_1^1 + q_2^1 k_2^1 + \cdots)\boldsymbol{v}^{\mathbf{1}} + (q_1^1 k_1^2 + q_2^1 k_2^2 + \cdots)\boldsymbol{v}^{\mathbf{2}} + \cdots$$

$$= q_1^1 k_1^1 \boldsymbol{v}^{\mathbf{1}} + q_2^1 k_2^1 \boldsymbol{v}^{\mathbf{1}} + \cdots + q_1^1 k_1^2 \boldsymbol{v}^{\mathbf{2}} + q_2^1 k_2^2 \boldsymbol{v}^{\mathbf{2}} + \cdots + \cdots$$

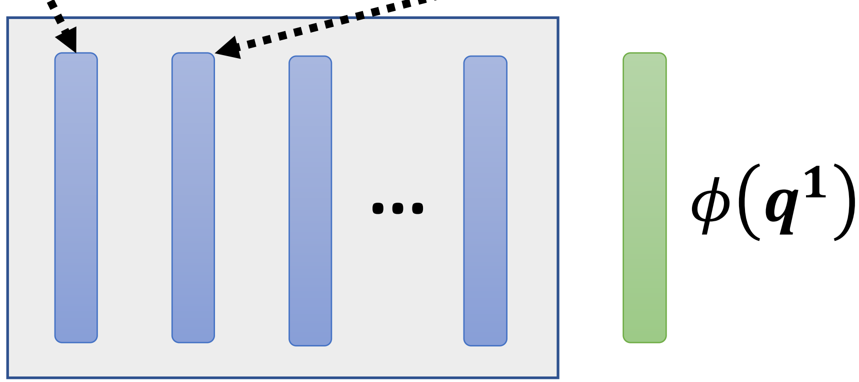
$$= q_1^1\left(k_1^1 \boldsymbol{v}^{\mathbf{1}} + k_1^2 \boldsymbol{v}^{\mathbf{2}} + \cdots\right) + q_2^1\left(k_2^1 \boldsymbol{v}^{\mathbf{1}} + k_2^2 \boldsymbol{v}^{\mathbf{2}} + \cdots\right)$$

$$\boldsymbol{b}^{1} = \sum_{i=1}^{N} \alpha'_{1,i}\boldsymbol{v}^{i} = \frac{\boxed{\sum_{i=1}^{N}\left[\phi(\boldsymbol{q}^{1})\cdot\phi(\boldsymbol{k}^{i})\right]\boldsymbol{v}^{i}}}{\phi(\boldsymbol{q}^{1})\cdot\sum_{j=1}^{N}\phi(\boldsymbol{k}^{j})}$$

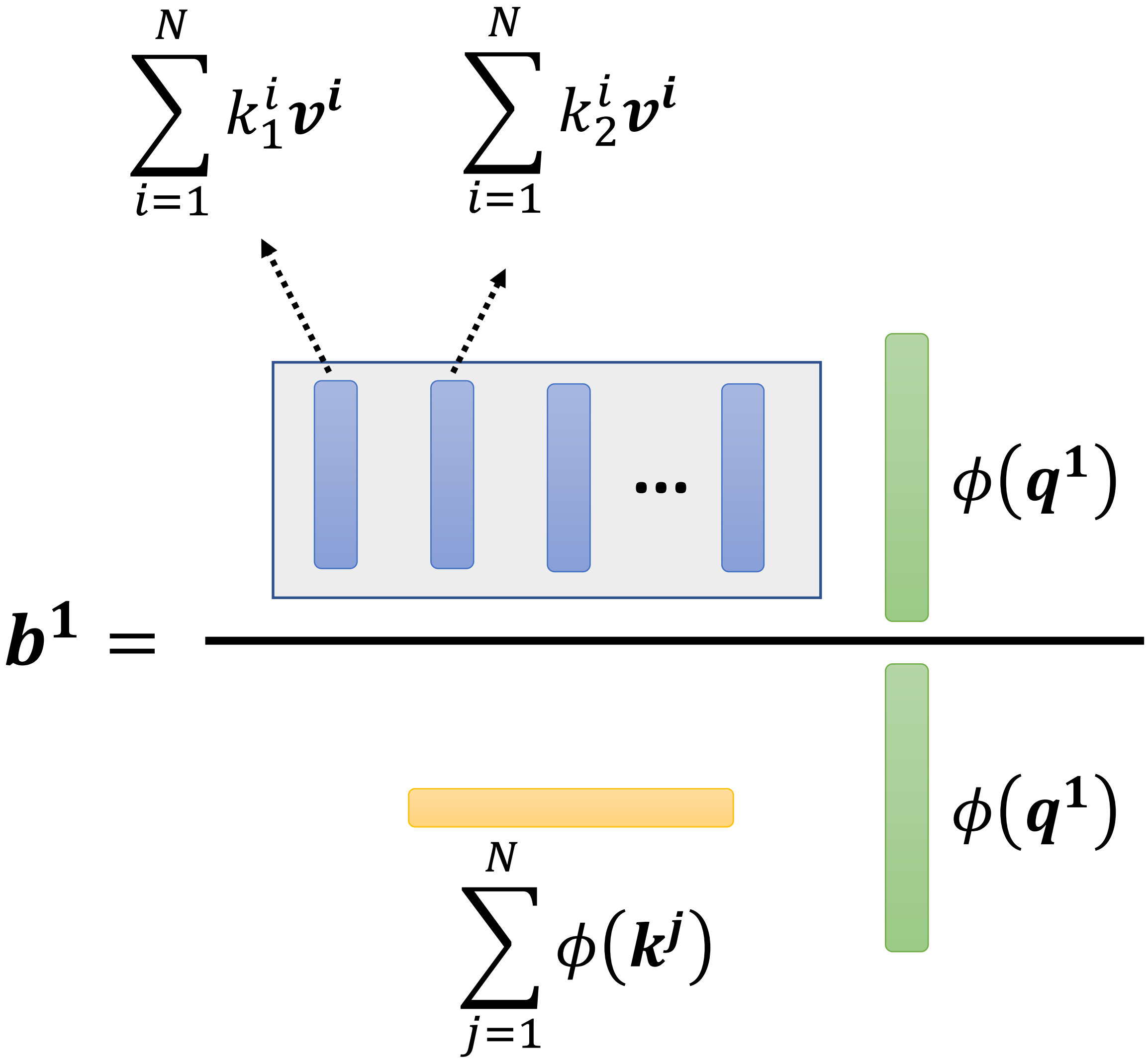
$$\boxed{\sum_{i=1}^{N}\left[\phi(\boldsymbol{q}^{1})\cdot\phi(\boldsymbol{k}^{i})\right]\boldsymbol{v}^{i}} \qquad \phi(\boldsymbol{q}^{1}) = \begin{bmatrix} q_1^1 \\ q_2^1 \\ \vdots \end{bmatrix} \qquad \phi(\boldsymbol{k}^{1}) = \begin{bmatrix} k_1^1 \\ k_2^1 \\ \vdots \end{bmatrix}$$

$$= q_1^1\left(k_1^1\boldsymbol{v}^{1} + k_1^2\boldsymbol{v}^{2} + \cdots\right) + q_2^1\left(k_2^1\boldsymbol{v}^{1} + k_2^2\boldsymbol{v}^{2} + \cdots\right)$$

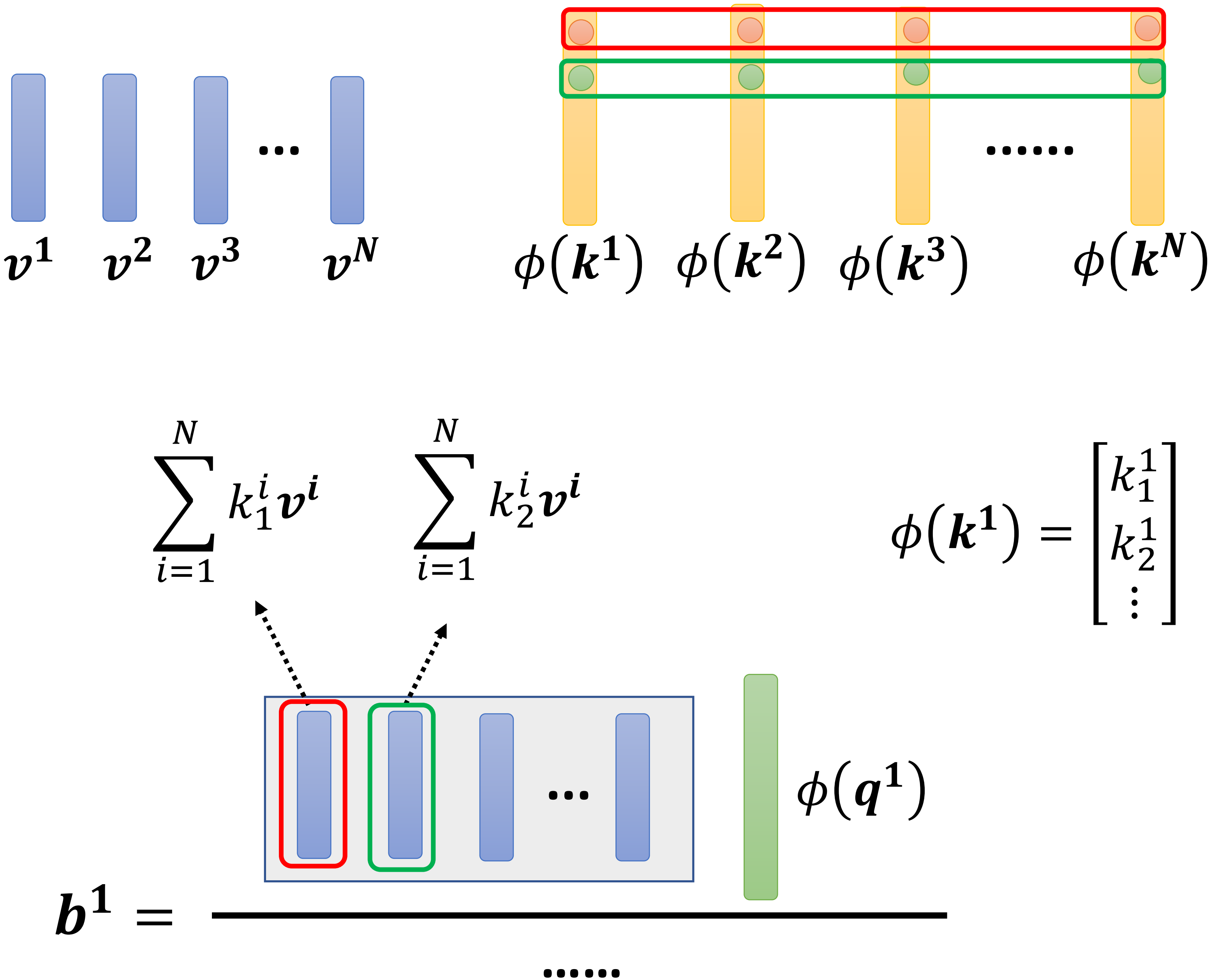
$$\sum_{i=1}^{N} k_1^i\boldsymbol{v}^{i}$$



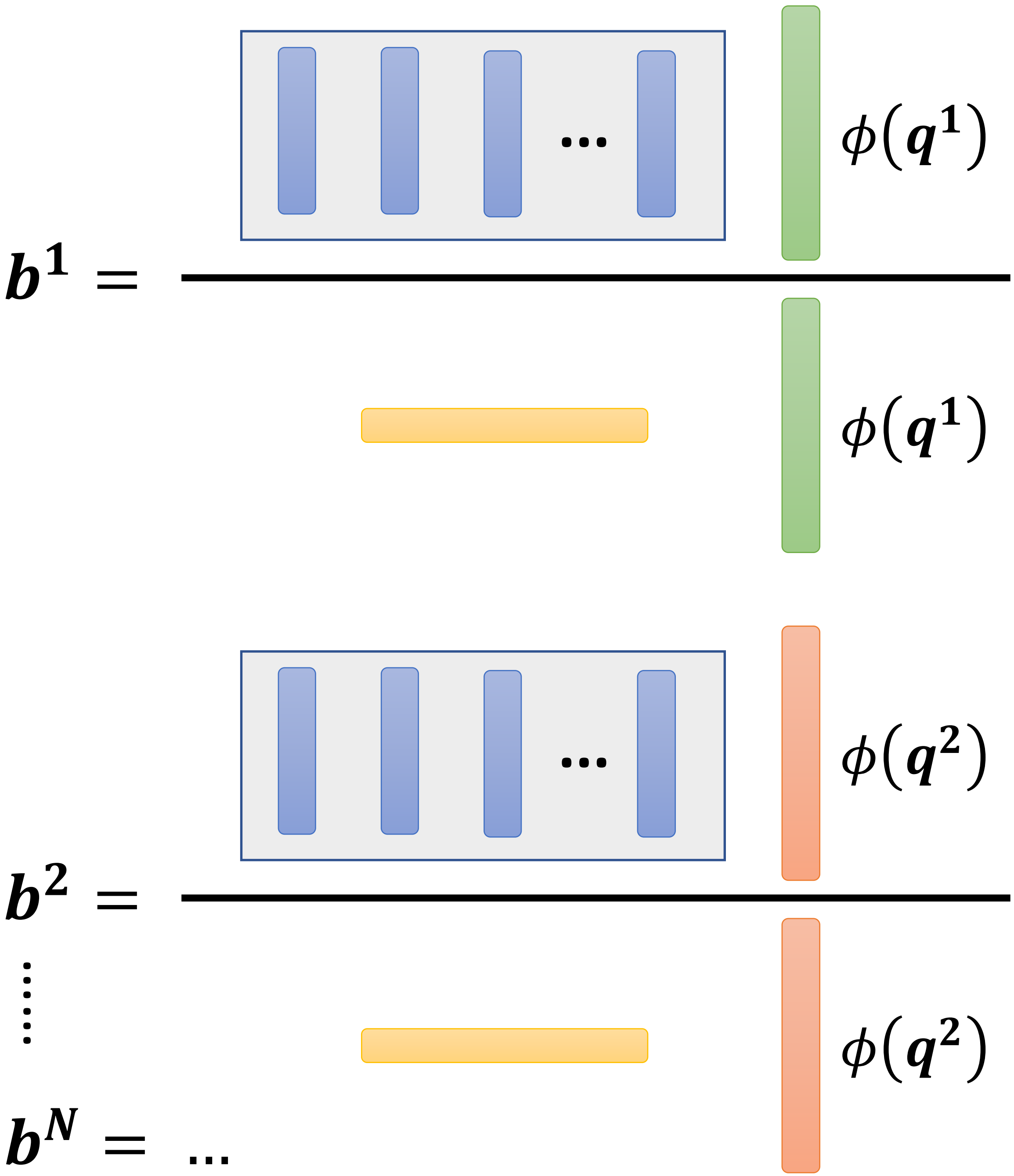

$$\sum_{i=1}^{N} k_2^i\boldsymbol{v}^{i}$$



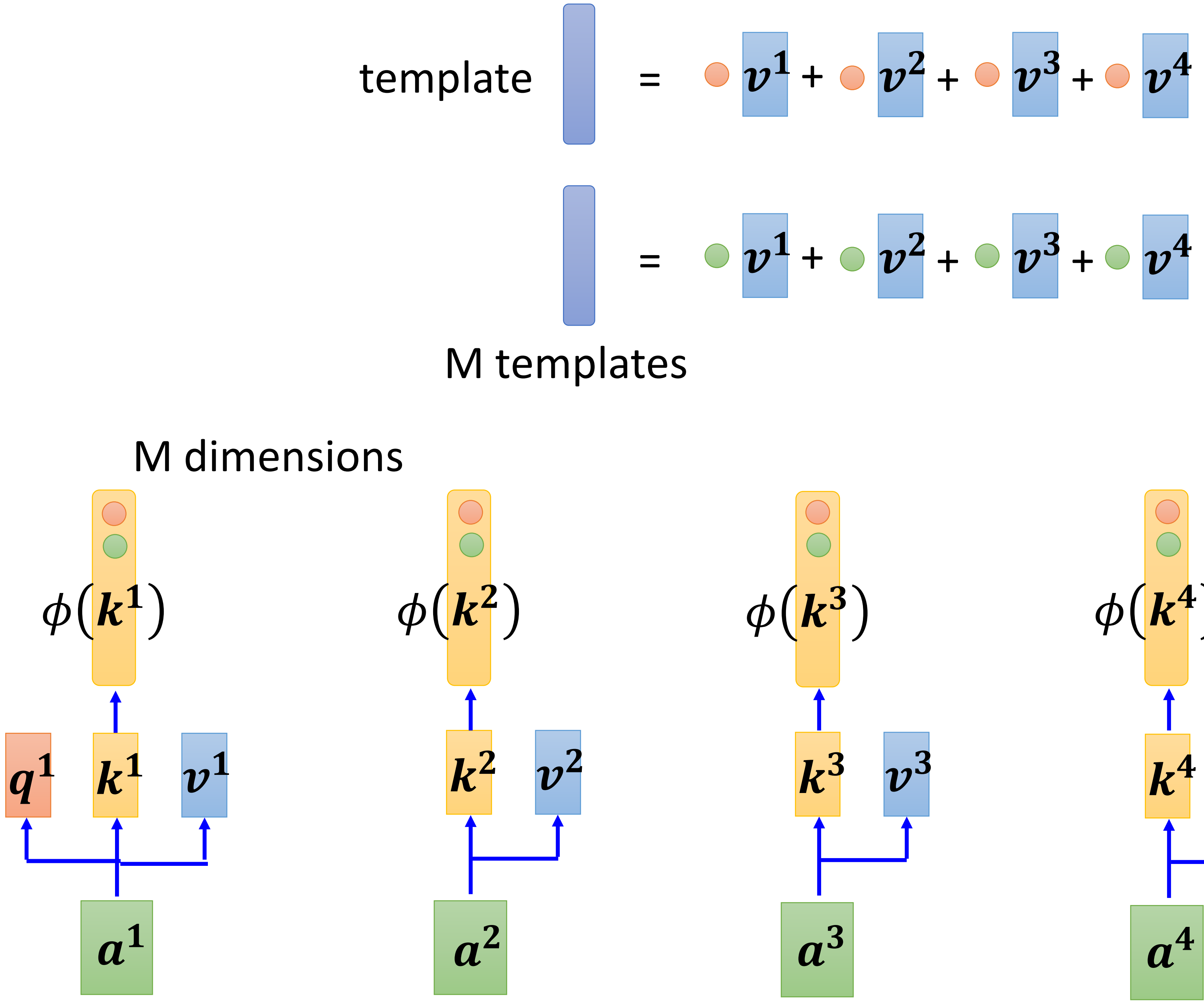
$\sum_{i=1}^{N} k_1^i \boldsymbol{v}^i$
$\sum_{i=1}^{N} k_2^i \boldsymbol{v}^i$
$\boldsymbol{b}^1 =$
$\phi(\boldsymbol{q}^1)$
$\phi(\boldsymbol{q}^1)$
$\sum_{j=1}^{N} \phi(\boldsymbol{k}^j)$



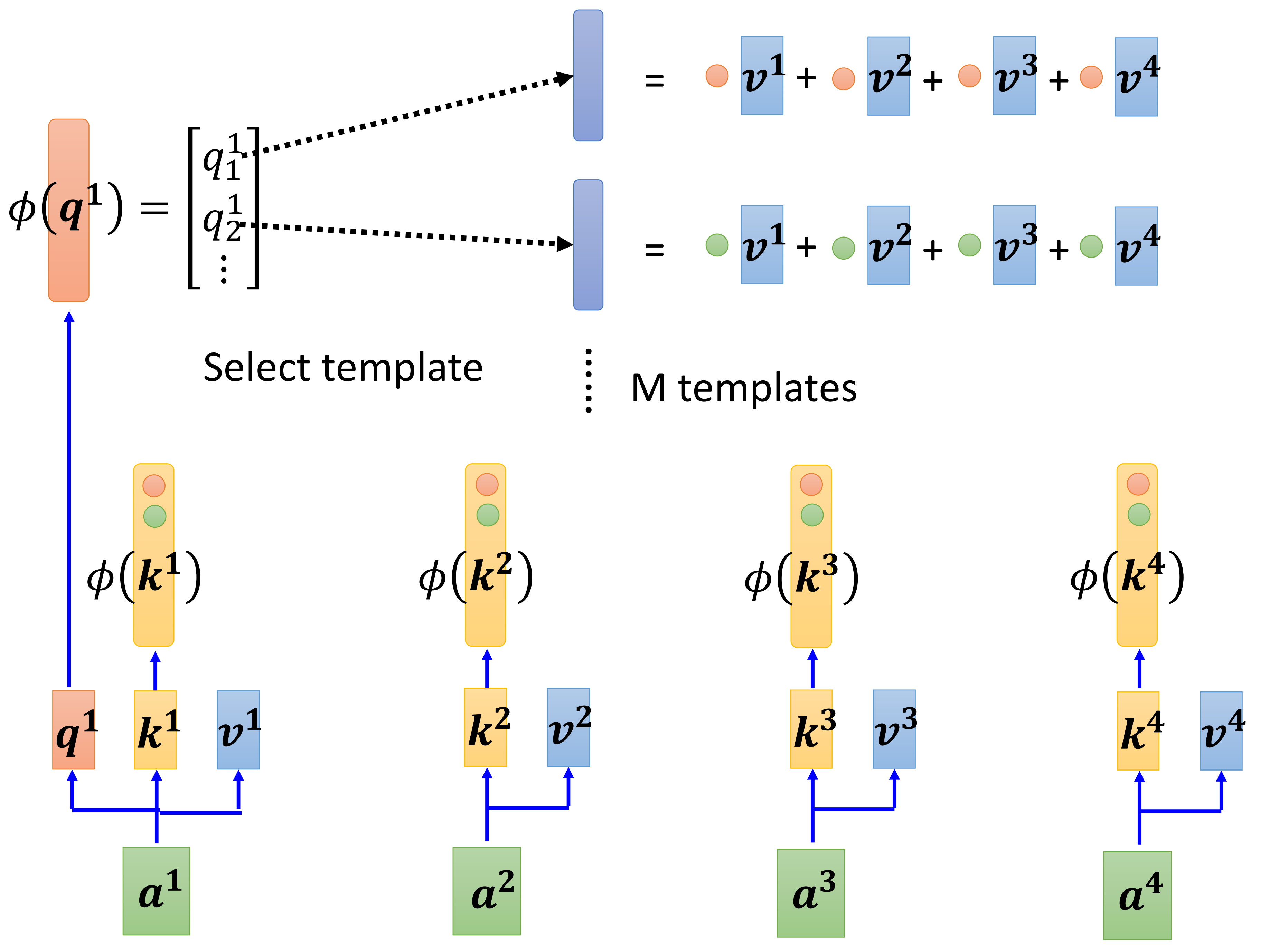
$\boldsymbol{v}^1$ $\boldsymbol{v}^2$ $\boldsymbol{v}^3$ … $\boldsymbol{v}^N$
$\phi(\boldsymbol{k}^1)$ $\phi(\boldsymbol{k}^2)$ $\phi(\boldsymbol{k}^3)$ …… $\phi(\boldsymbol{k}^N)$
$\sum_{i=1}^{N} k_1^i \boldsymbol{v}^i$
$\sum_{i=1}^{N} k_2^i \boldsymbol{v}^i$
$\phi(\boldsymbol{k}^1) = \begin{bmatrix} k_1^1 \\ k_2^1 \\ \vdots \end{bmatrix}$
$\phi(\boldsymbol{q}^1)$
$\boldsymbol{b}^1 =$
……



$\boldsymbol{b}^1 =$
$\phi(\boldsymbol{q}^1)$
$\phi(\boldsymbol{q}^1)$
$\boldsymbol{b}^2 =$
$\phi(\boldsymbol{q}^2)$
$\phi(\boldsymbol{q}^2)$
$\boldsymbol{b}^N = \ldots$

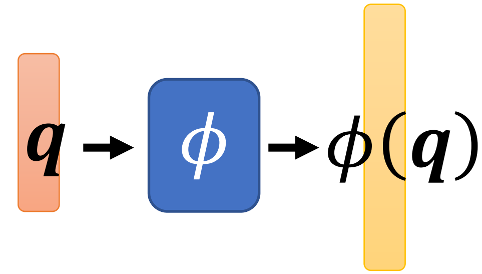


template
=
$v^1$ + $v^2$ + $v^3$ + $v^4$
=
$v^1$ + $v^2$ + $v^3$ + $v^4$
M templates
M dimensions
$\phi(k^1)$
$\phi(k^2)$
$\phi(k^3)$
$\phi(k^4)$
$q^1$
$k^1$
$v^1$
$k^2$
$v^2$
$k^3$
$v^3$
$k^4$
$v^4$
$a^1$
$a^2$
$a^3$
$a^4$



$\phi(q^1) = \begin{bmatrix} q_1^1 \\ q_2^1 \\ \vdots \end{bmatrix}$
$= \; v^1 + \; v^2 + \; v^3 + \; v^4$
$= \; v^1 + \; v^2 + \; v^3 + \; v^4$
Select template
M templates
$\phi(k^1)$
$\phi(k^2)$
$\phi(k^3)$
$\phi(k^4)$
$q^1$ $k^1$ $v^1$
$k^2$ $v^2$
$k^3$ $v^3$
$k^4$ $v^4$
$a^1$
$a^2$
$a^3$
$a^4$

# Realization




$$exp(\boldsymbol{q} \cdot \boldsymbol{k}) \approx \phi(\boldsymbol{q}) \cdot \phi(\boldsymbol{k})$$

- Efficient attention
  https://arxiv.org/pdf/1812.01243.pdf
- Linear Transformer
  https://linear-transformers.com/
- Random Feature Attention
  https://arxiv.org/pdf/2103.02143.pdf
- Performer
  https://arxiv.org/pdf/2009.14794.pdf