

KernelView of Attention

- Suppose there exists ρ such that $\sin(q, k) = \langle \rho(q), \rho(k) \rangle$

• If $Q' = \varphi(Q)$ and $K' = \varphi(K)$, output is

• Why write this way?

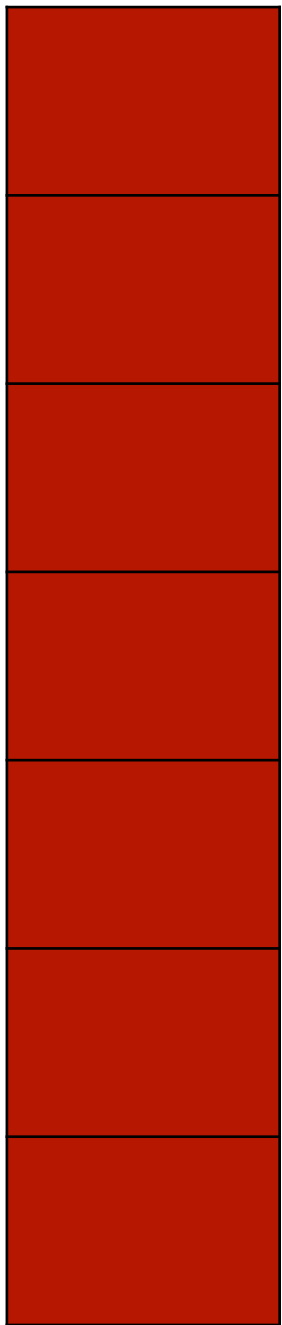
- **Linear time algorithm for computing $LT(A \cdot B^T) \cdot C$**

- What about ρ for softmax?

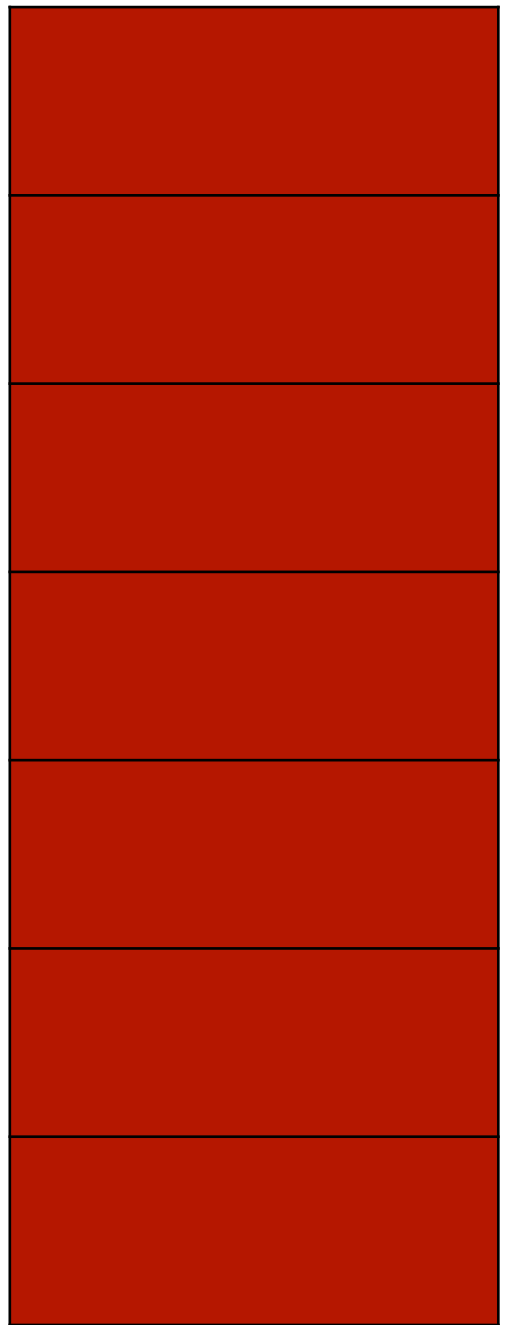
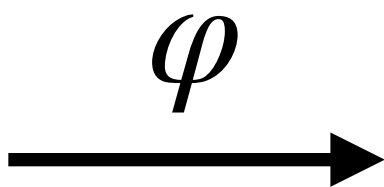
• **No finite dimensional feature maps**



$$D^{-1} \cdot {}^{\text{LT}}(Q' \cdot (K')^{\text{T}}) \cdot V$$



Q



Q'

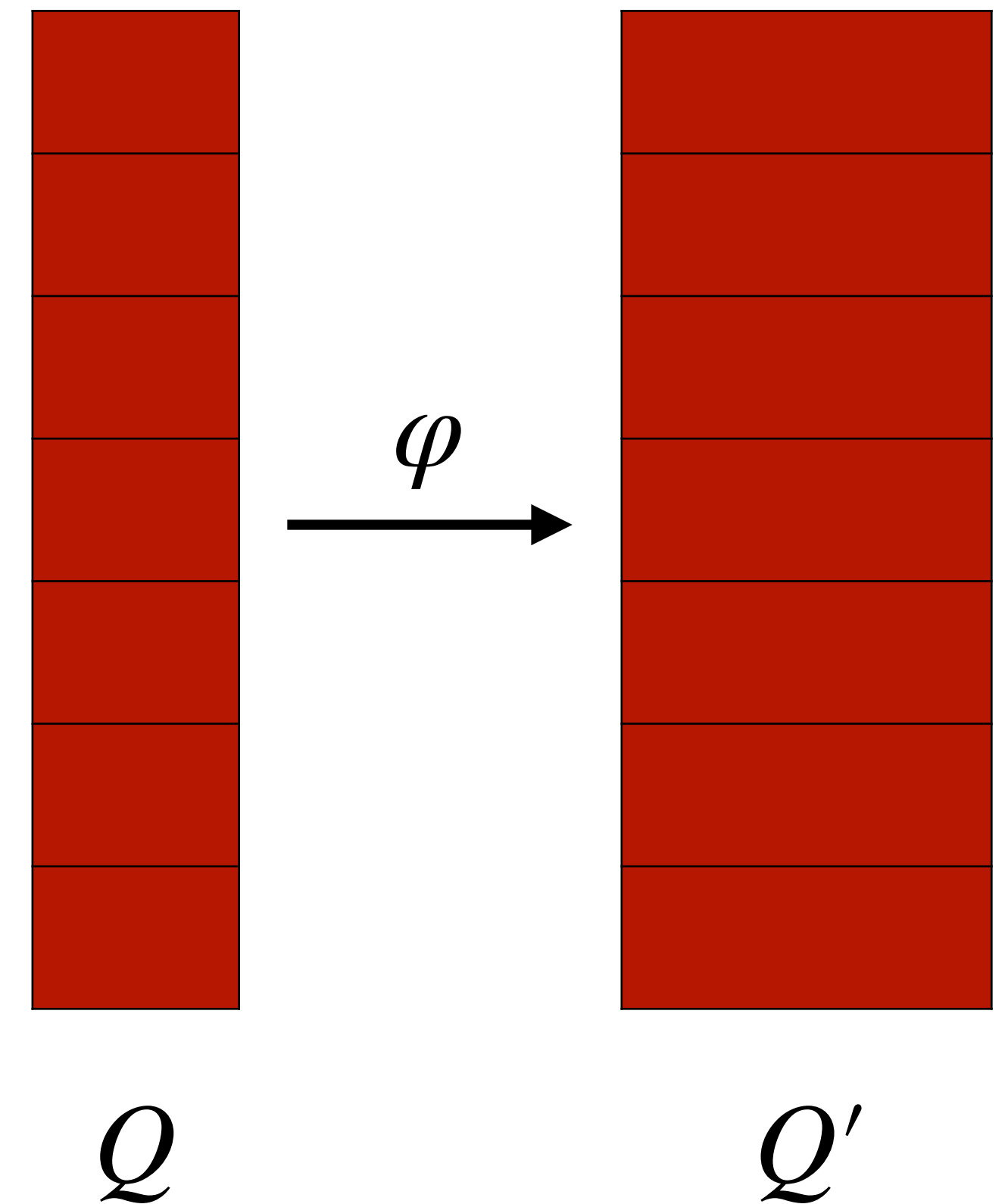
Kernel View of Attention

- Suppose there exists φ such that $\text{sim}(q, k) = \langle \varphi(q), \varphi(k) \rangle$

- If $Q' = \varphi(Q)$ and $K' = \varphi(K)$, output is

$$D^{-1} \cdot \text{LT}(Q' \cdot (K')^T) \cdot V$$

- Why write this way?
 - **Linear time algorithm** for computing $\text{LT}(A \cdot B^T) \cdot C$
- What about φ for softmax?
 - No finite dimensional feature maps



Previous Works