Kernel View of Attention

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

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

- Why write this way?

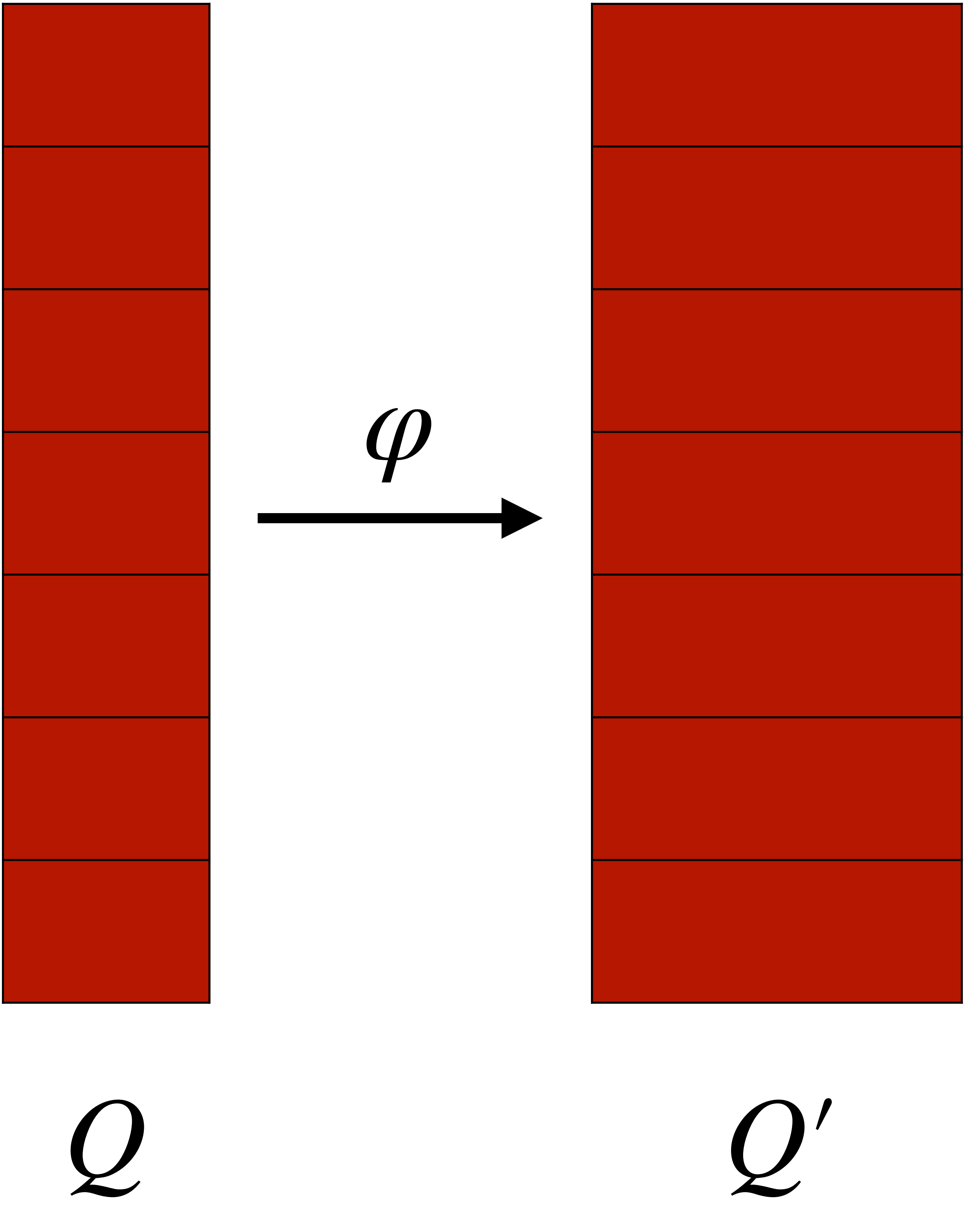
- **Linear time algorithm** for computing $\mathrm{LT}(A \cdot B^\top) \cdot C$

- What about φ for softmax?

- No finite dimensional feature maps

17

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

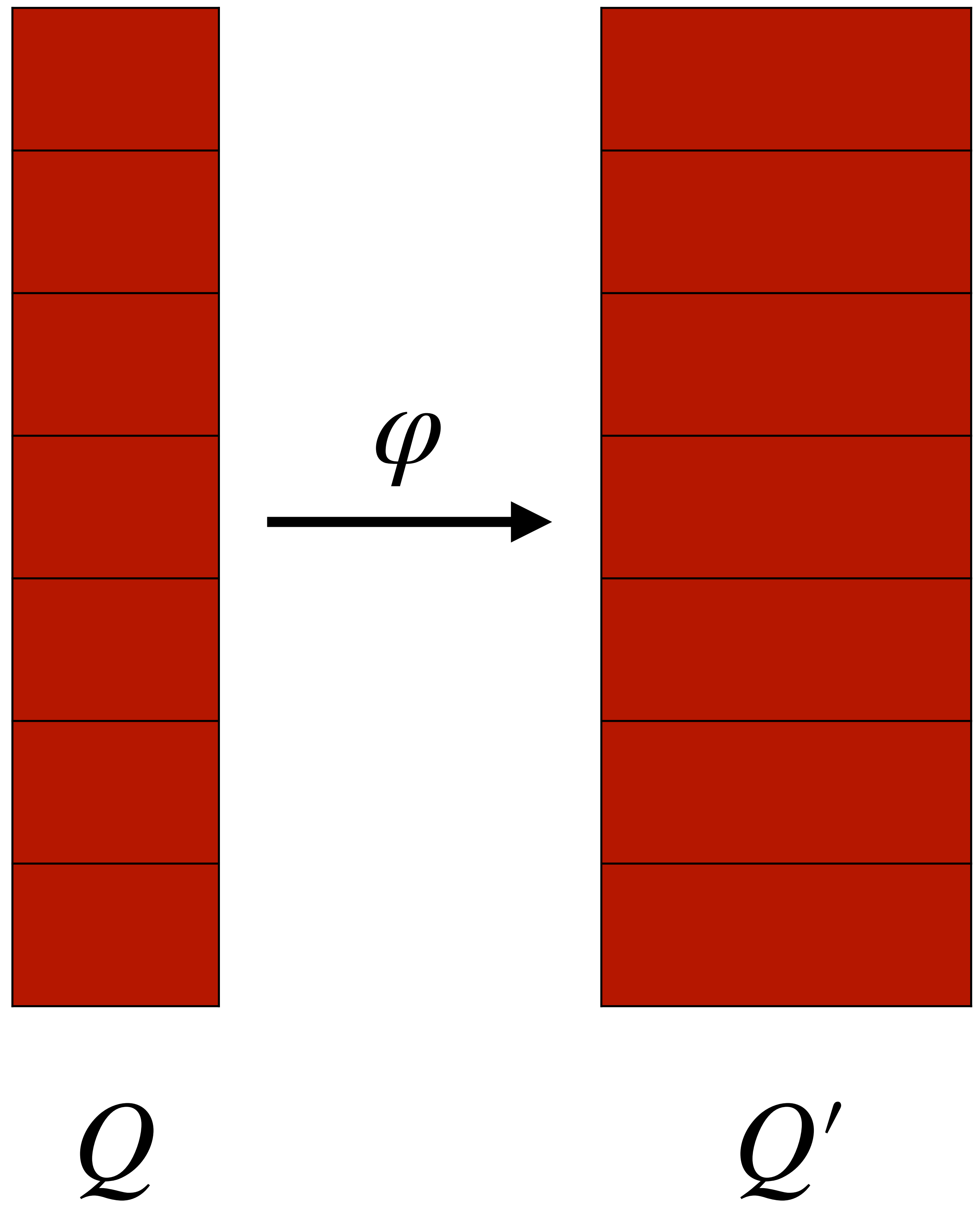

φ
Q
Q′

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

$$D^{-1} \cdot \text{LT}(Q' \cdot (K')^{\top}) \cdot V$$

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



Previous Works