

Attention Outputs

• The computation can be represented as

• Q . • K^T is $n \times n$

$$\bullet \text{ diagonal of } \mathcal{D} = \text{LT}(\exp(\mathcal{Q} \cdot K^{\top})) \cdot \mathbf{1}_n$$

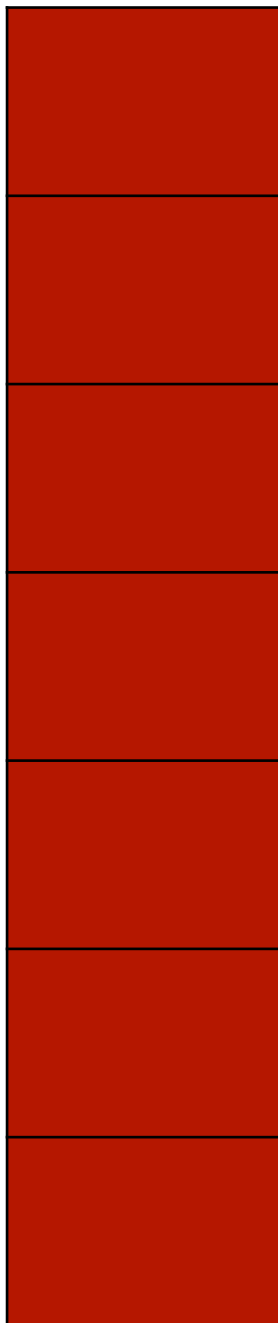
- Naively computing takes $\mathcal{O}(n^2)$ time

- Prohibitive when n is large



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

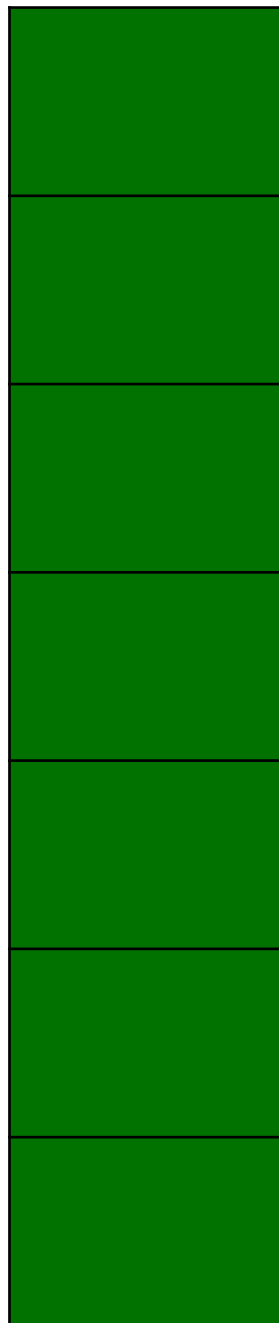
I
work
in
the
city
of
New



Q



K



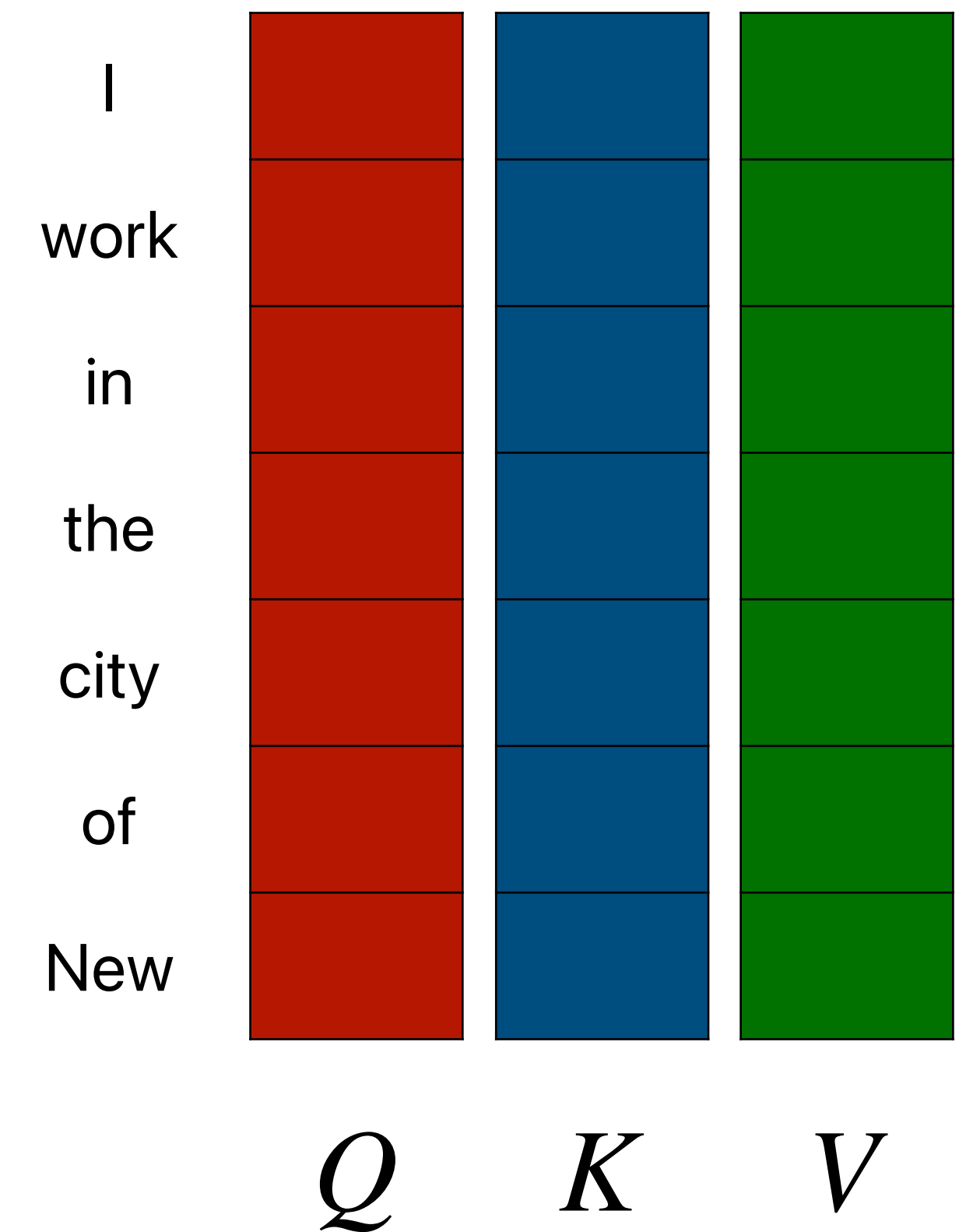
V

Attention Outputs

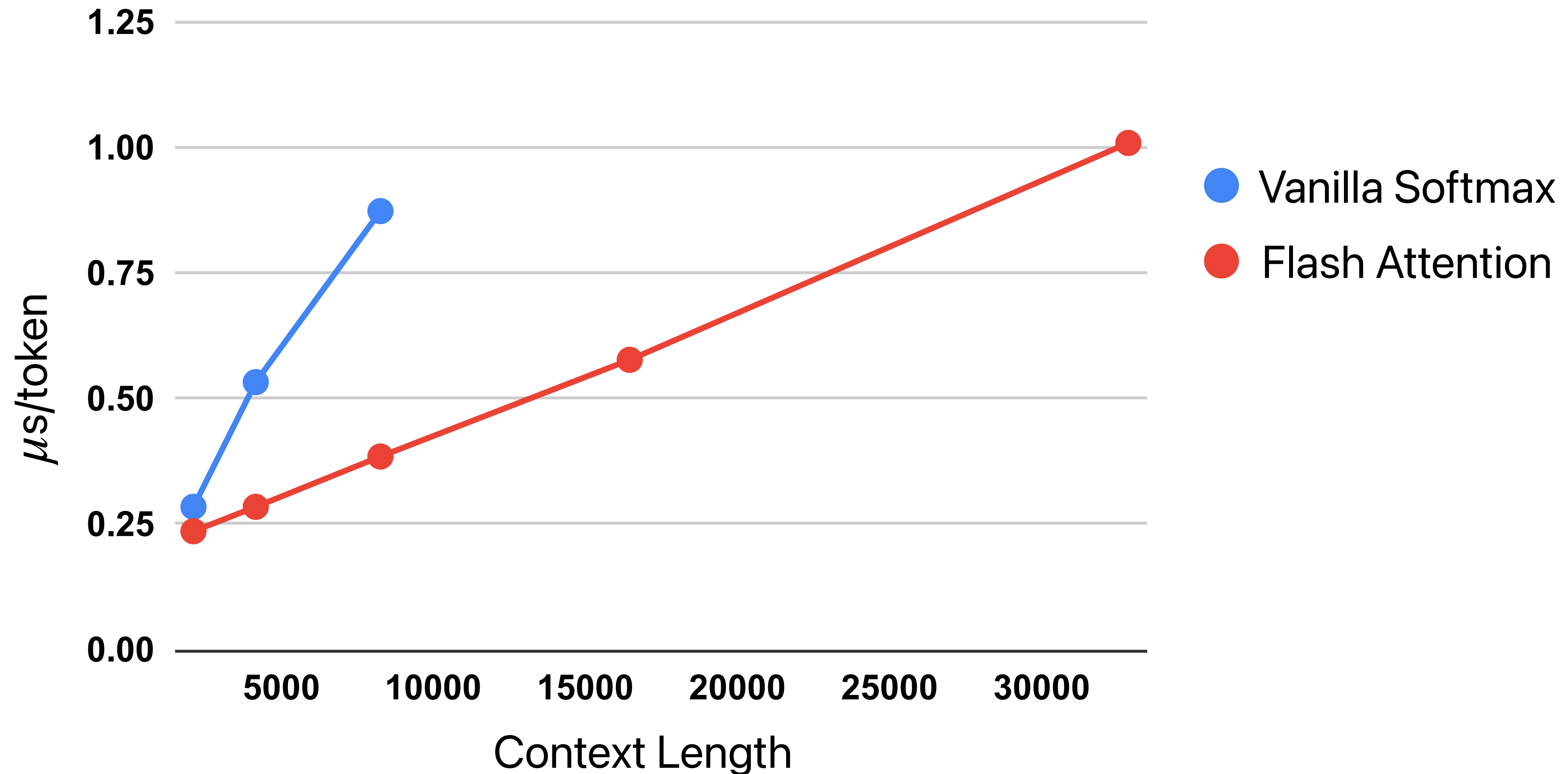
- The computation can be represented as

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

- $Q \cdot K^T$ is $n \times n$
- diagonal of $D = \text{LT}(\exp(Q \cdot K^T)) \cdot 1_n$
- Naively computing takes $O(n^2)$ time
 - Prohibitive when n is large



Train Step Latency Per Token



Each token in the training examples looks back at the whole context