

Integrating Exact Attention Locally

- Algorithm: Given Q, K and V , let $Q' = \varphi(Q)$ and $K' = \varphi(K)$ and compute $\text{LT}(Q' \cdot (K')^T) \cdot V$

- Instead of using $\text{LT}(Q' \cdot (K')^T)$ in the diagonal blocks, what if we use $\text{LT}((Q \cdot K^T)^4)$?

This is exactly what we are trying to approximate

The computation that is needed is of the same order

• Improve the model quality as we!!!

36

Integrating Exact Attention Locally

- Algorithm: Given Q, K and V , let $Q' = \varphi(Q)$ and $K' = \varphi(K)$ and compute $\text{LT}(Q' \cdot (K')^\top) \cdot V$
- Instead of using $\text{LT}(Q' \cdot (K')^\top)$ in the diagonal blocks, what if we use $\text{LT}((Q \cdot K^\top)^4)$?
 - This is exactly what we are trying to approximate
 - The computation that is needed is of the same order
 - Improves the model quality as well!

Model Perplexities

Perplexities on Wiki-40B

