Integrating Exact Attention Locally

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

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

- 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!

36

Integrating Exact Attention Locally

- Algorithm: Given Q, K and V, let $Q' = \varphi(Q)$ and $K' = \varphi(K)$ and compute $\mathrm{LT}(Q' \cdot (K')^{\mathsf{T}}) \cdot V$
- Instead of using $\mathrm{LT}(Q' \cdot (K')^{\mathsf{T}})$ in the diagonal blocks, what if we use $\mathrm{LT}((Q \cdot K^{\mathsf{T}})^4)$?
  - 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