NEURAL INFORMATION PROCESSING SYSTEMS

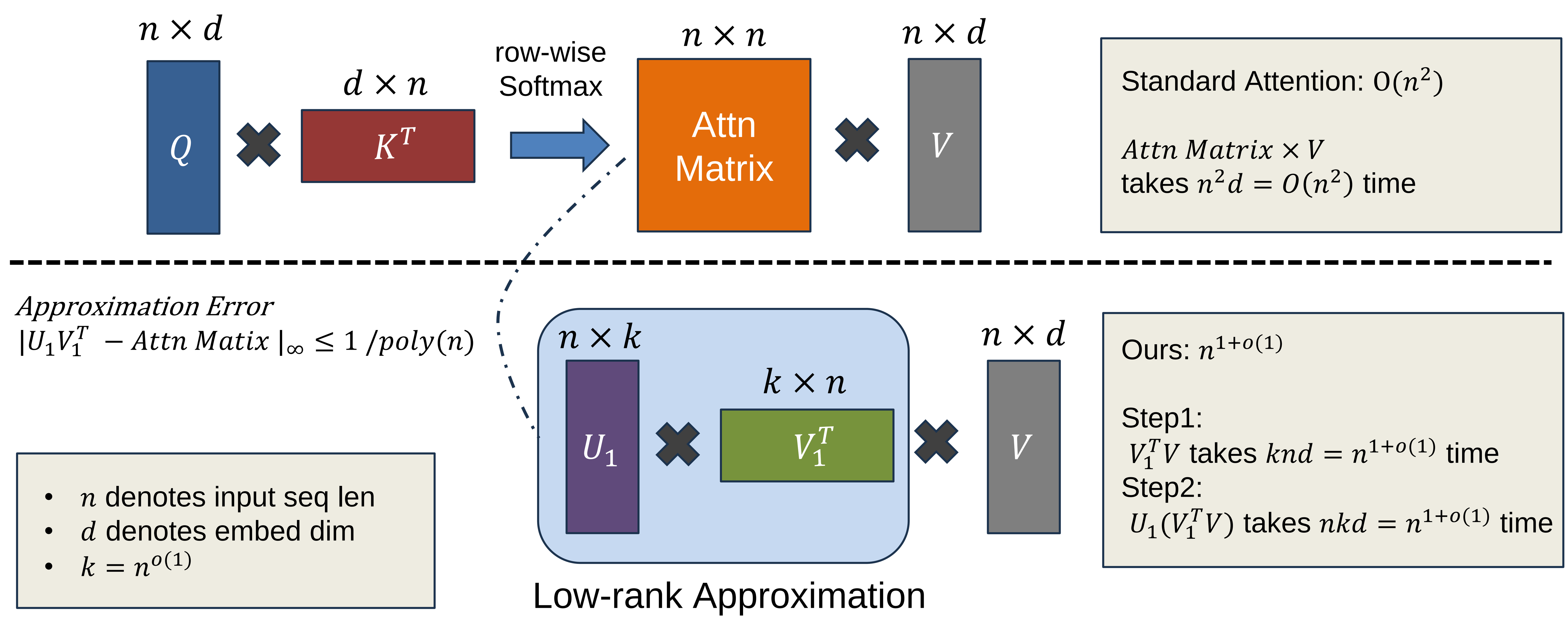
Multi-Layer Transformers Gradient Can be Approximated in Almost Linear Time

Yingyu Liang, Zhizhou Sha, Zhenmei Shi, Zhao Song, Yufa Zhou

香 港 大 學 THE UNIVERSITY OF HONG KONG · 清華大學 Tsinghua University · WISCONSIN · Berkeley · Penn

Transformer Training is too slow?

We have proved your Transformer Training can speed up from $O(n^2)$ to $n^{1+o(1)}$



Problem Setup

- Self-attention module $\mathrm{Attn}(X) = \mathrm{Softmax}(XW_QW_K^\top X^\top/d)\cdot XW_V$

  $\mathrm{Attn}(X) = f(X)\cdot XW_V$

  where (1) $A := \exp(XW_QW_K^\top X^\top/d) \in \mathbb{R}^{n\times n}$ (2) $D := \mathrm{diag}(A\mathbf{1}_n) \in \mathbb{R}^{n\times n}$

  (3) $f(X) := D^{-1}A \in \mathbb{R}^{n\times n}$
- Multilayer Transformers $\mathsf{F}_m(X) := g_m \circ \mathrm{Attn}_m \circ g_{m-1} \circ \mathrm{Attn}_{m-1} \circ \cdots \circ g_1 \circ \mathrm{Attn}_1 \circ g_0(X)$

  where (1) Attn_i denotes self-attention module

  (2) g_i denotes components other than

  (3) $\circ$ denotes function composition

Theoretical Results

Theorem 1 (Single-layer gradient approximation)
Our algorithm can approximate the gradient on X, $W_QW_K^T$, W_V in almost linear time $n^{1+o(1)}$, with approximation error bounded by $1/poly(n)$.

Theorem 2 (Multi-layer gradient approximation)
The number of layers m can be treated as an constant.
Our algorithm can approximate the gradient on X, $W_QW_K^T$, W_V in almost linear time $n^{1+o(1)}$, with approximation error bounded by $1/poly(n)$.

Extensions We have also proved that our almost linear time algorithm also can easily extend to supporting other components in Transformers, such as residual connection, multi-head attention, causal mask, etc.

Take-Home Message We leverage the low-rank nature of the attention matrix to accelerate the gradient computation of multi-layer Transformers from $O(n^2)$ to $n^{1+o(1)}$. Our findings will inspire the further study and usage of the low-rank patterns within the Transformer architecture.