NEURAL INFORMATION PROCESSING SYSTEMS

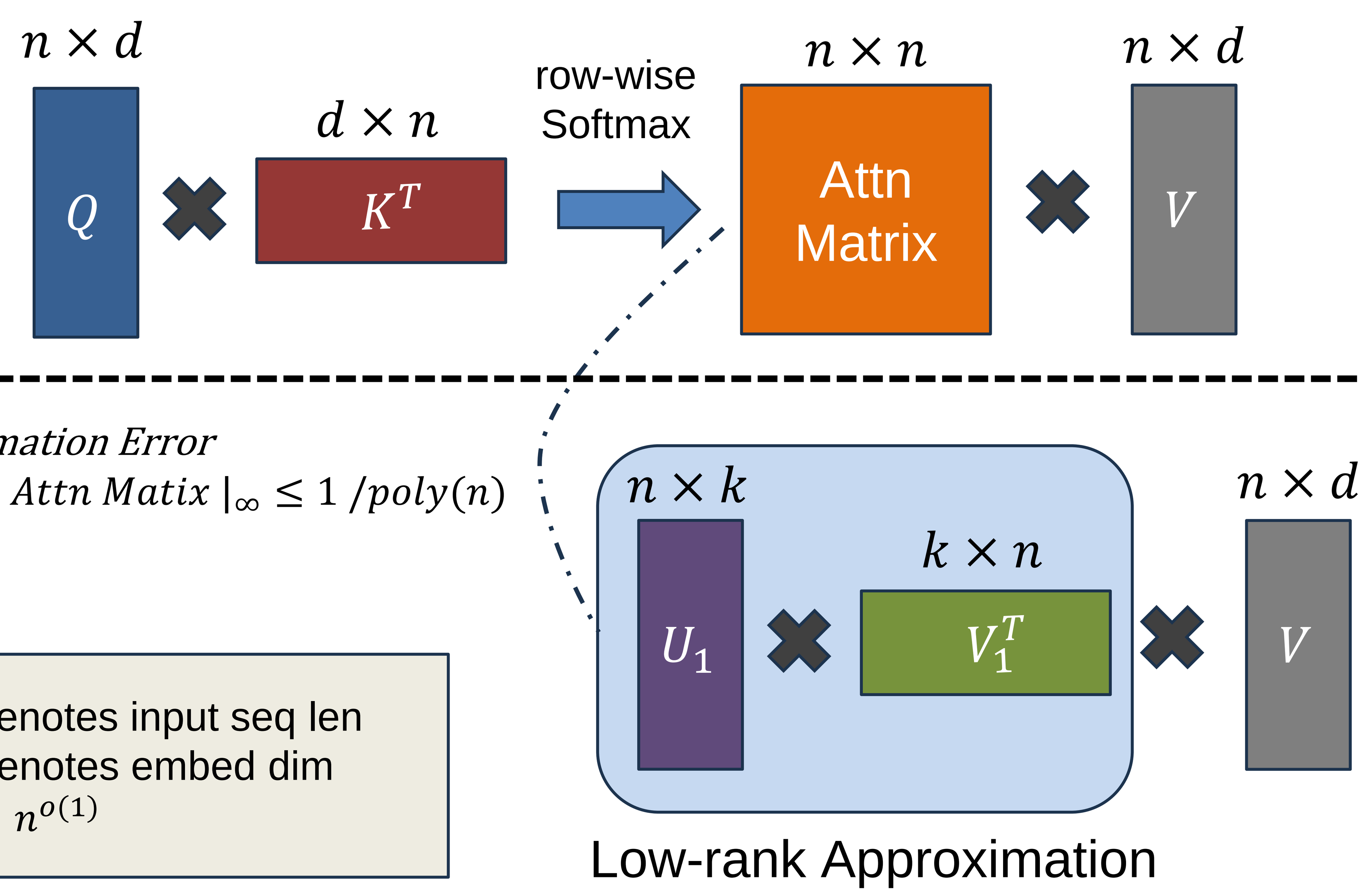
# Multi-Layer Transformers Gradient Can be Approximated in Almost Linear Time


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

THE UNIVERSITY OF HONG KONG · Tsinghua University · WISCONSIN · Berkeley · Penn


*Transformer Training is too slow?*

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




Standard Attention: $O(n^2)$

$Attn\ Matrix \times V$ takes $n^2 d = O(n^2)$ time

*Approximation Error*
$|U_1 V_1^T - Attn\ Matix|_\infty \le 1/poly(n)$

- $n$ denotes input seq len
- $d$ denotes embed dim
- $k = n^{o(1)}$

Ours: $n^{1+o(1)}$

Step1: $V_1^T V$ takes $knd = n^{1+o(1)}$ time

Step2: $U_1(V_1^T V)$ takes $nkd = n^{1+o(1)}$ time

## Problem Setup

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

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

  where (1) $A := \exp(XW_Q W_K^\top X^\top / d) \in \mathbb{R}^{n \times n}$ (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 $\mathsf{F}_m(X) := g_m \circ \mathsf{Attn}_m \circ g_{m-1} \circ \mathsf{Attn}_{m-1} \circ \cdots \circ g_1 \circ \mathsf{Attn}_1 \circ g_0(X)$

  where (1) $\mathsf{Attn}_i$ 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 $X$, $W_Q W_K^T$, $W_V$ 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 a constant.
Our algorithm can approximate the gradient on $X$, $W_Q W_K^T$, $W_V$ 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.