RACE Attention: A Strictly Linear-Time Attention for Long-Sequence Training

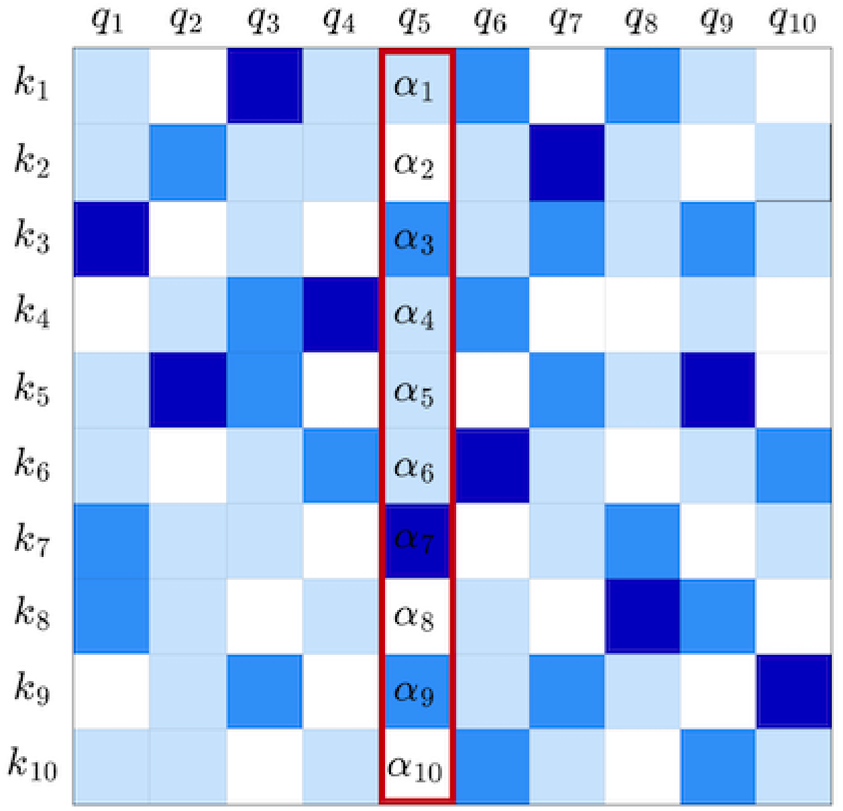
Sahil Joshi, *Agniva Chowdhury*, *Amar Kanakamedala*, *Ekam Singh*, *Evan Tu*, *Anshumali Shrivastava*

Department of Computer Science, Rice University

RICE UNIVERSITY

Discussing the Problem

- Softmax Attention incurs a quadratic complexity in sequence length during training, which becomes prohibitive to run at long contexts.
- Attention can be written as $O_i = \frac{\sum_{j=1}^{N} \mathrm{sim}(Q_i, K_j) V_j}{\sum_{j=1}^{N} \mathrm{sim}(Q_i, K_j)}$
- Calculating the similarities for each pair of (Q_i, K_j) is expensive for long context, even with fully optimized GPU implementation – FlashAttention.
- In Softmax Attention, $\mathrm{sim}(Q_i, K_j) = \exp\left(\frac{Q_i^T K_j}{\sqrt{d}}\right)$

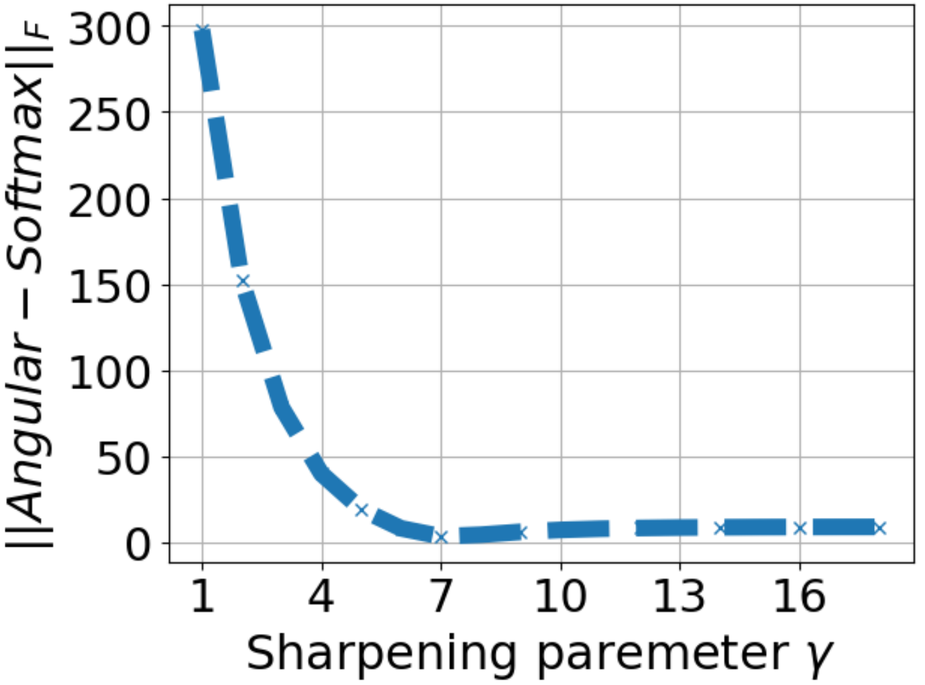


$$o_5 = \sum_{i=1}^{10} \alpha_i v_i \quad \text{where } \alpha_i = \mathrm{softmax}(q_5 k_i)$$

Softmax Attention

Our Finding

- $\mathrm{sim}(Q_i\ K_j) = 1\ - \frac{1}{\pi}\cos^{-1}\left(\frac{Q_i^T K_j}{||Q_i||_2 \left|\left|K_j\right|\right|_2}\right)$. Note that it depends only on the angle between the vectors Q_i and K_j and is invariant to their norms. So, we assume $||Q_i||_2 = \left|\left|K_j\right|\right|_2 = 1$.
- Although this angular kernel appears to be less discriminative than the exponential, raising it to a constant power γ, yields a sharpened angular similarity $\mathrm{sim}(\mathrm{Q_i}, \mathrm{K_j}) = \left(1\ - \frac{1}{\pi}\cos^{-1}\left(Q_i^T K_j\right)\right)^{\gamma}$.
- The figure shows that as γ increases the sharpened angular similarity closely mimics the softmax similarity!



Introduction to RACE Sketch

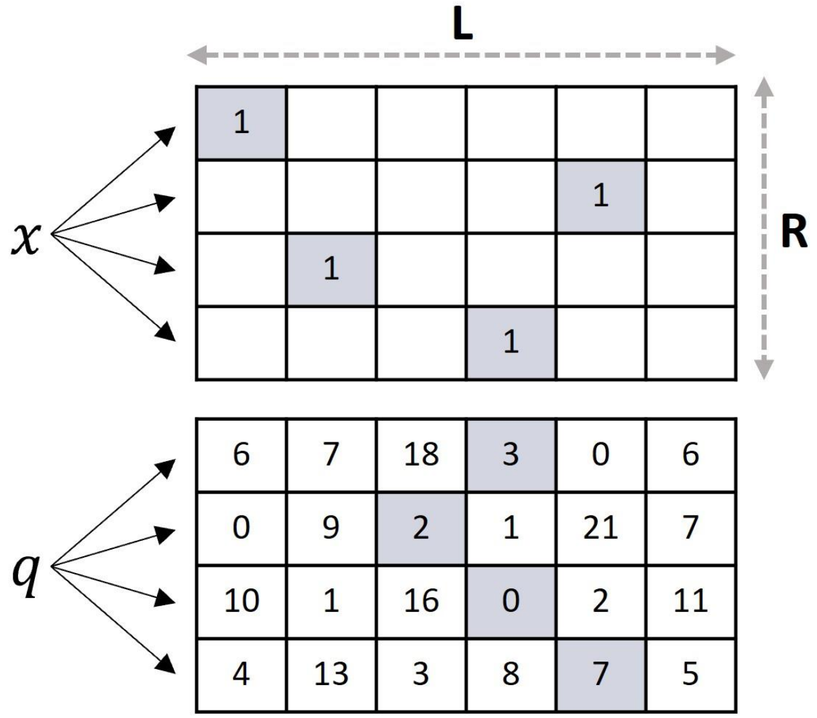
- Let $h(x) = sign(r^T x)$ with $r \sim N(0, I)$. For unit vectors,

$$\Pr[h(q) = h(x)] = 1 - \frac{1}{\pi}\cos^{-1}(q^T x)\,.$$

- Concatenating γ independent hashes, $g(x) = (h_1(x), \ldots, h_\gamma(x))$, yields $\Pr[g(q) = g(x)] = \mathrm{sim}(q, x)$.
- RACE Sketch: For each query q, and L tables, the estimator can be written as

$$s(q) = \frac{1}{L}\sum_{l=1}^{L}\sum_{j=1}^{N} 1\{g^l(x_j) = g^l(q)\}, \qquad E[s(q)] = \sum_{j=1}^{N} \mathrm{sim}(q, x_j)$$

- RACE estimates the total similarity of dataset around the query. In the context of attention computation, it can estimate the denominator directly, and the numerator with a value-weighted variant of it.
- However, traditional RACE estimator is discrete and hence non-differentiable, making it unsuitable for end-to-end training. Our algorithm on the next slide introduces a differentiable alternative.



$$O_i = \frac{\sum_{j=1}^{N} \mathrm{sim}(Q_i, K_j) V_j}{\sum_{j=1}^{N} \mathrm{sim}(Q_i, K_j)}$$

Our Algorithm

Soft bucketization: Queries and Keys are softly assigned to $L \times R$ hash buckets via random projections.

Bucket aggregation: Key assignment weights accumulate per-bucket value sums and bucket masses.

Global normalization: Queries softly mix bucket statistics across tables, followed by a single normalization.

Complexity: Time $\mathcal{O}(LNRd)$, Space $\mathcal{O}\left(L(NR + Rd)\right)$; with $R, L \ll d$, scaling is linear in N and d.

Causal variant: Bucket statistics are replaced by cumulative sums.

Input: $\mathrm{Q}, \mathrm{K}, V \in \mathbb{R}^{\mathrm{N}\times\mathrm{d}}$, # hash tables L; # hyperplanes P; temperature $\beta > 0$.

Output: $\hat{\mathrm{O}} \in \mathbb{R}^{\mathrm{N}\times\mathrm{d}}$.

for $\ell = 1, \ldots, L$ **do**

Draw $W^{(\ell)} \in \mathbb{R}^{\mathrm{P}\times\mathrm{d}}$ with rows $w_p^{(\ell)} \sim \mathcal{N}(0, I_d)$

Define the corner set $\mathcal{V} = \{\pm 1\}^P$ with $R = 2^P$ corners.

Build $\Phi_Q^{(\ell)}, \Phi_K^{(\ell)} \in \mathbb{R}^{\mathrm{N}\times\mathrm{R}}$ with rows

$$\left[\phi^{(\ell)}(x)\right]_r = \frac{\exp\{\beta(\tanh(W^{(\ell)}x))^\top v_r\}}{\sum_{r'} \exp\{\beta(\tanh(W^{(\ell)}x))^\top v_{r'}\}}$$

Compute per-table bucket statistics:

$$A^\ell = \left(\Phi_K^{(\ell)}\right)^\top 1_N, \qquad B^\ell = \left(\Phi_K^{(\ell)}\right)^\top V.$$

end for

Compute averaged statistics:

$$Num = \frac{1}{L}\sum_{\ell=1}^{L} \Phi_Q^{(\ell)} B^{(\ell)}, \qquad Den = \frac{1}{L}\sum_{\ell=1}^{L} \Phi_Q^{(\ell)} A^{(\ell)}$$

Return: $\hat{\mathrm{O}} \leftarrow \left(\mathrm{diag}(\mathrm{Den})\right)^{-1}\mathrm{Num}$

Theoretical Insights

- **Assumptions:**
  - Stable Normalization (A1): Normalization factor is bounded away from zero and scales with N.
  - Bounded Operator Norm (A2): The spectral norm of the angular kernel scales at most linearly with N.

- Let $Q, K, V \in \mathbb{R}^{N \times d}$ be the query, key, and value matrices. For parameters $L, P,$ and β, and under conditions (A1) and (A2), the estimator $\hat{O}$ produced by the Algorithm satisfies

$$\left|\left|\hat{O} - O\right|\right|_{\text{rms}} = \mathcal{O}\left(\frac{P}{\beta} + \sqrt{\frac{\log\left(\frac{N}{\delta}\right)}{L}}\right)\left|\left|V\right|\right|_F$$

- As $\beta \uparrow$ (i.e., bias $\downarrow$) and $L \uparrow$ (i.e., variance $\downarrow$), the error vanishes

Experiments 1/2

- We show that RACE Attention achieves competitive performance on diverse tasks such as LM, MLM, and text/image classification.

Food-101 (16K)

Method	Train ↓	Test ↓	Acc. ↑
RACE (P=2, L=2)	**891s**	**37s**	42.4%
RACE (P=3, L=3)	950s	40s	**43.5%**
RACE (P=4, L=4)	1042s	42s	40.3%
Linear	1166s	44s	41.4%
Linformer-128	1250s	49s	20.2%
Performer-256	2546s	105s	42.4%
FlashAttention2	2600s	95s	42.1%

ArXiv Classification (64K)

Method	Train ↓	Test ↓	Acc. ↑
RACE (P=2,L=2)	**561s**	22s	97.14%
RACE (P=3,L=3)	584s	22.5s	**97.92%**
RACE (P=4,L=4)	594s	22.9s	97.4%
Linear	591s	22.8s	96.35%
Linformer-128	616s	**15.2s**	97.4%
Performer-256	952s	35s	96.61%
FlashAttention2	1645s	47s	97.0%

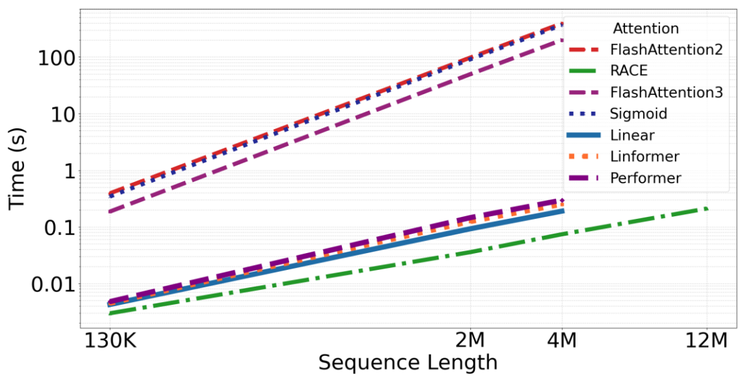
LM and MLM @ 1024

Method	WikiText PPL ↓	Tiny Stories PPL ↓
RACE (P=2, L=2)	23.9	4.2
RACE (P=2, L=3)	23.4	3.2
RACE (P=3, L=3)	21.9	3.2
RACE (P=3, L=4)	21.5	2.9
RACE (P=4, L=4)	<u>20.9</u>	<u>2.6</u>
FlashAttention2	<u>20.9</u>	2.7
Angular (γ=8)	**19.0**	**2.5**

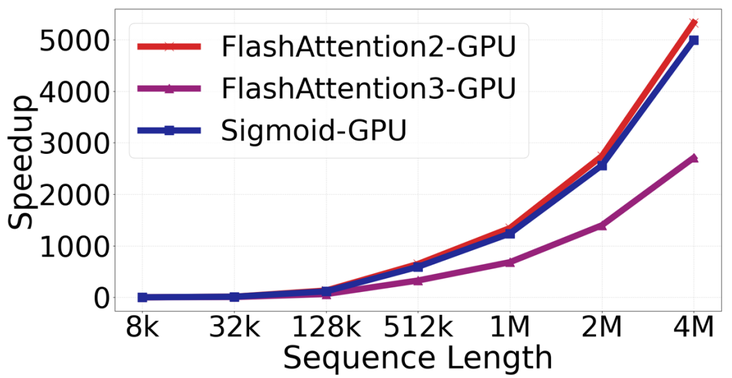
Experiments 2/2

- We analyze scaling behavior using a single forward–backward pass of one attention layer.
- For large sequence lengths (4M tokens), **GPU-RACE** substantially outperforms existing methods, achieving speedups of up to **5,500×** over FlashAttention2, **5,000×** over Sigmoid attention, and **2,600×** over FlashAttention3.
- **CPU-RACE** also delivers significant gains, running about **40× faster** than FlashAttention2 and Sigmoid attention, and **20× faster** than FlashAttention3.

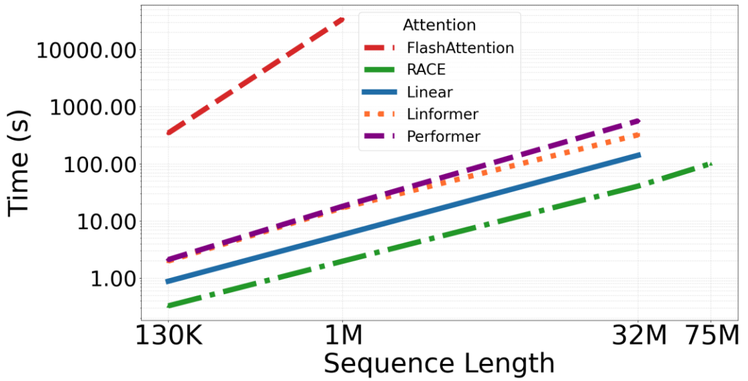
GPU Scaling



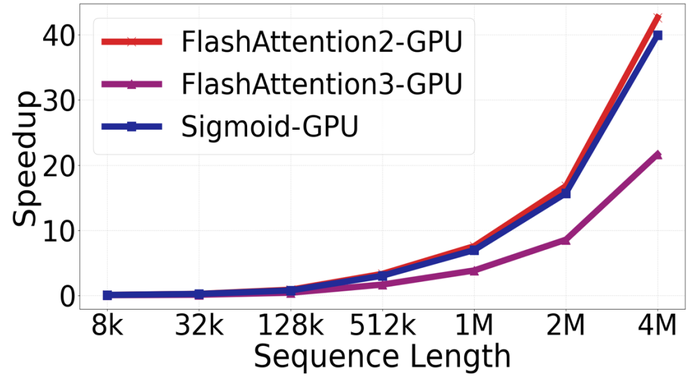
Speedup GPU-RACE



CPU Scaling



Speedup CPU-RACE



Conclusion

- We propose RACE Attention, a randomized, differentiable sketching approach that enables linear-time and memory-efficient attention, making long-sequence training practical on modern hardware.
- The method approximates a sharpened angular kernel and provides a scalable alternative to Softmax Attention, with clear benefits for efficient key–value caching during inference.

Sahil Joshi

RICE UNIVERSITY