Fast Attention Mechanisms: a Tale of Parallelism

Jingwen Liu, Hantao Yu, Clayton Sanford, Alexandr Andoni, Daniel Hsu
Columbia University, Google

Neural Information Processing Systems

Motivation

Standard attention mechanism

- Slow, takes quadratic time in sequence length
- Expressive
  - Can simulate MPC protocol (MapReduce)
  - Solves multi-step reasoning problems with optimal depth

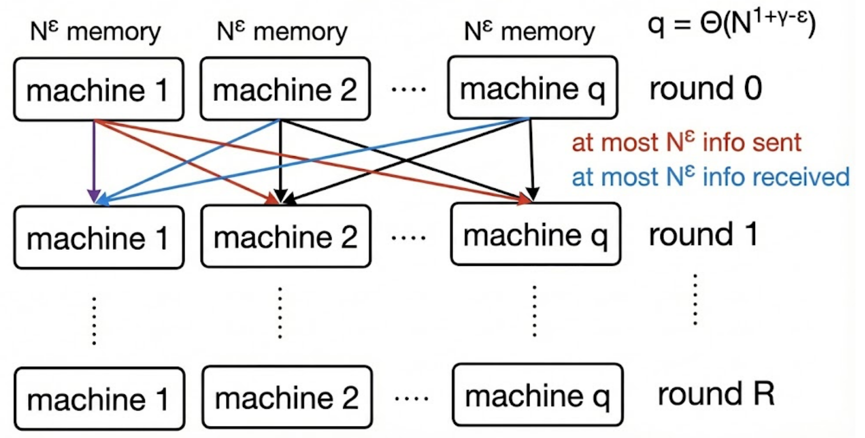
Sub-quadratic variants of attention

- Fast, near-linear time in sequence length
- Parameter-inefficient for algorithmic reasoning tasks
  - RNN, LSTM, Mamba
  - Performer, Poly-Sketchformer, Longformer, etc

Is there any efficient attention mechanism that maintains the key representational advantages of standard attention over non-parallel mechanisms?

Massively Parallel Computation (MPC)

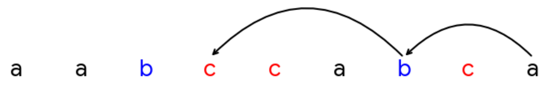
- A model for processing big datasets with parallel and distributed computation on clusters.
- An R-round (γ, ε)-MPC protocol on input N words specifies the computation represented by $q = N^{1+\gamma-\varepsilon}$ machines, each with local memory $s = N^{\varepsilon}$ words.



k-Hop Induction Heads

- Induction heads are the identified as a mechanism for model's capability for in-context learning
- Related problem: multi-step reasoning
  - John is in the playground. Helen is playing with John. Helen picked up a football. Where is the football?
- 1-hop induction heads: find the last occurrence, output the next token
- k-hop induction heads:

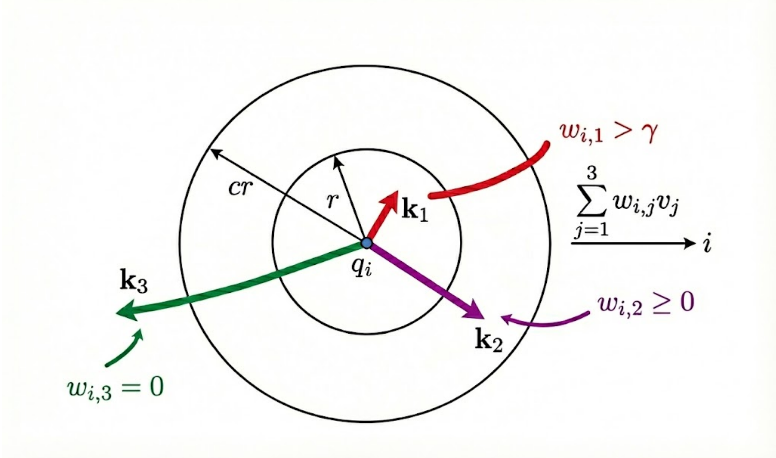
  repeat the procedure k times



Main Results

Approximate Nearest Neighbor Attention (ANNA)

- Standard attention can be seen as exact nearest neighbor search
- Approximate nearest neighbor search
  - Find neighbors within cr distance with the query (r is the NN distance)
  - Near-linear time $O(N^{1+1/c^2})$ when c is large
- Given the embedded $Q, K, V \in \mathbb{R}^{N\times m}$, for each query only compute attention with Approximate NNs
  - $ANNA_{Q,K,V}(X)_i = \sum_{j}^{N} w_{i,j} v_j$
  - $\sum_{j}^{N} w_{i,j} = 1$
  - $w_{i,j} > 0 \to \|k_j - q_i\| \le cr$
  - $\|k_j - q_i\| \le r \to w_{i,j} \ge \tau, \tau > 0$



ANNA-transformer is equivalent to MPC

- Theorem (ANNA-transformer simulates MPC): Any R-round (γ, ε)-MPC protocol can by simulated by an ANNA-transformer with depth $O(R)$ and width (number of heads $\times$ m) $O(N^{\varepsilon+\delta})$, for any fixed $\delta > 0$.
  - Sub-quadratic time simulation
  - Ties ANNA-transformer in the existing MPC hierarchy
  - $O(1)$-layer ANNA-transformer can solve 3-SUM with width $O(N^{1/2+\delta})$
- Theorem (MPC simulates ANNA-transformer): Any L-layer ANNA-transformer with width $O(N^{\varepsilon})$ can be simulated by a $O(L)$-round MPC protocol with local memory $s = O(N^{\varepsilon+\delta})$ and $q = O(N^{1+\delta+3/c^2})$ machines.
  - Sub-quadratic number of machines
  - Round complexity lower bound for MPC $\to$ depth lower bound

Comparison with other efficient mechanisms

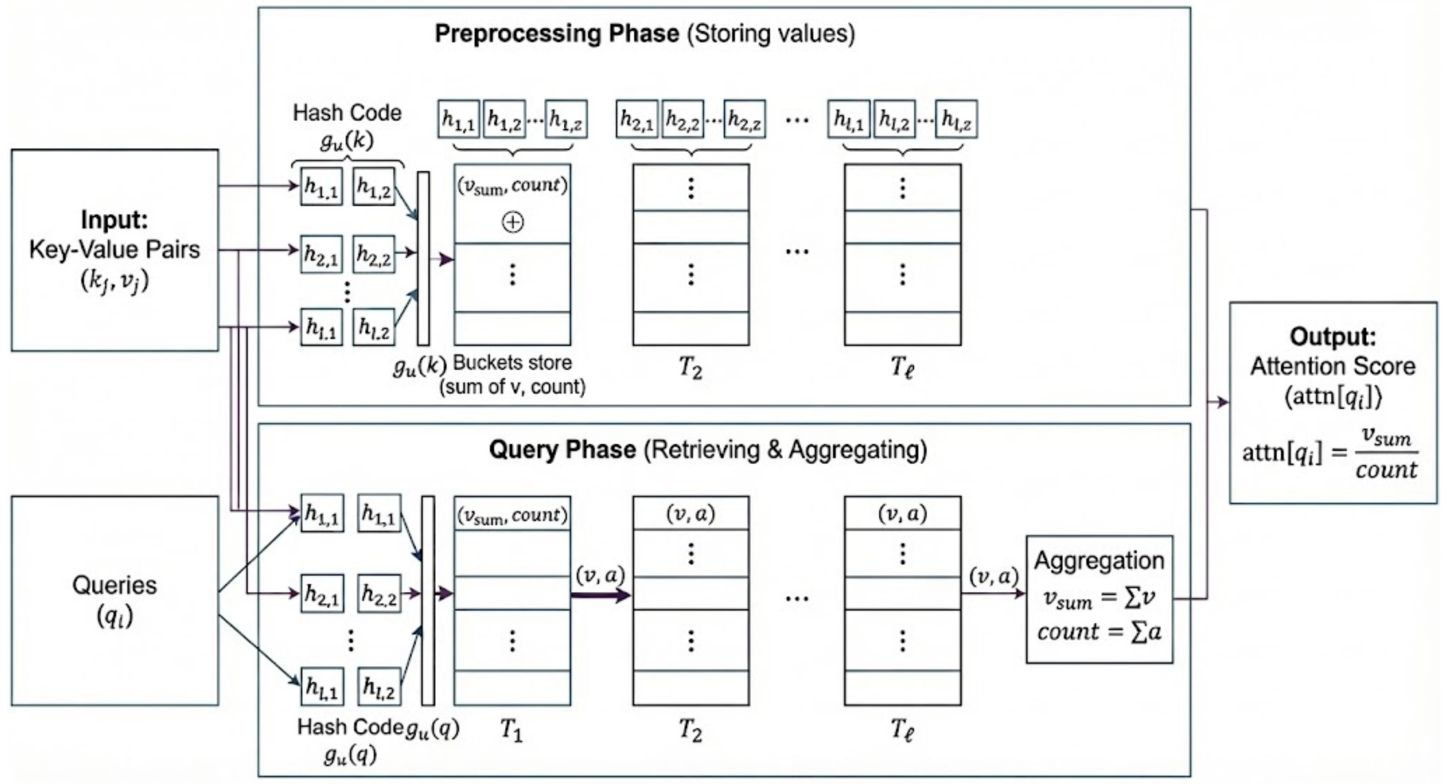
- Theorem (ANNA-simulates low-rank/kernel-based attention): Any low-rank attention-based transformer with L layers, rank $\times$ width $O(N^{\varepsilon})$ can be simulated by an ANNA-transformer with depth $O(L)$ and width $O(N^{\varepsilon+\delta})$.

Near-optimal multi-step reasoning

- Theorem: Depth $O(\log k)$ ANNA-transformers with width $O(N^{\varepsilon})$ can solve k-hop.
  - RNN, LSTM, State-Space model requires either depth k or linear width [1]
  - Low-rank/kernel-based and masking-based sub-quadratic attention require either depth k or near-quadratic computation [1]

Algorithm

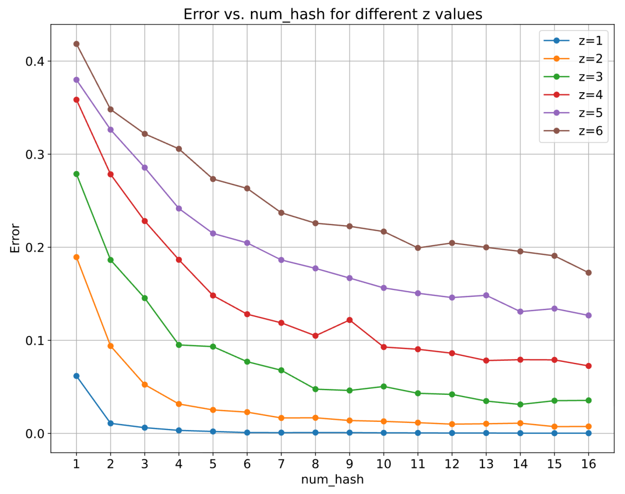
- Theorem: Fix $c > \sqrt{3}$. An LSH-based algorithm can compute ANNA with high probability, using time $\tilde{O}(mN^{1+3/c^2})$ and space $\tilde{O}(mN)$.
- Locality sensitive hashing (LSH): A family of hash functions that maps nearby points into the same hash buckets
  - $\|x - y\| \le r \to \Pr[h(x) = h(y)] \ge p_1$
  - $\|x - y\| > cr \to \Pr[h(x) = h(y)] \le p_2$



Experiments

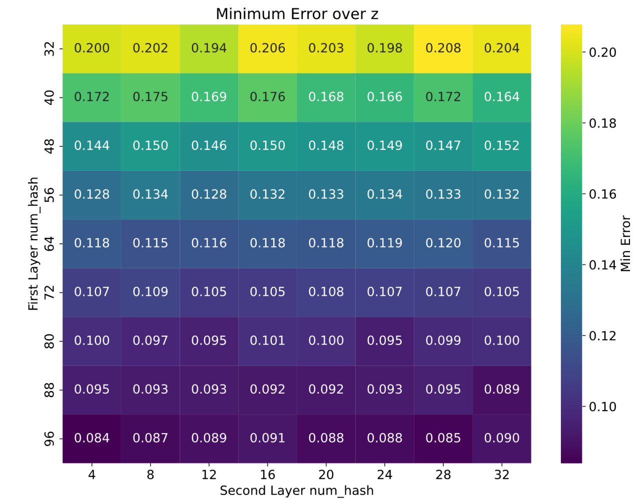
Match2

- Sequence length $N = 32$
- 1-layer ANNA-transformer



Induction heads (1-hop)

- Sequence length $N = 100$
- 2-layer ANNA-transformer



[1] Clayton Sanford, Daniel Hsu and Matus Telgarsky. Transformers, parallel computation, and logarithmic depth. ICML 2024.