Streaming Attention Approximation via Discrepancy Theory

Michael Kapralov

EPFL

Collaborators

Ekaterina Kochetkova
EPFL

Kshiteej Sheth
EPFL

Insu Han
KAIST

Amir Zandieh
Google AI

NeurIPS'25 (spotlight), Available at: arXiv:2502.07861

Transformers & Attention

Each token = query, key, value embeddings $(q, k, v) \in \mathbb{R}^d$

Attention between (q, k, v) and $(q_1, k_1, v_1), \ldots, (q_n, k_n, v_n)$:

$$\mathsf{Attn}(q; K, V) \coloneqq \mathbb{E}_{i \sim \mathsf{Softmax}} v_i$$

where $\Pr[\mathsf{Softmax} = i] \propto \exp\left(\frac{\langle k_i, q \rangle}{\sqrt{d}}\right)$

Problem: memory scales with context length

Because of attention, Transformers require memory

$$\Omega(n \cdot d)$$

n – number of processed tokens, d – dimension of embeddings

Problem: memory scales with context length

Because of attention, Transformers require memory

$$\Omega(n \cdot d)$$

n – number of processed tokens, d – dimension of embeddings

Problem

Reduce the memory without sacrificing attention

Problem: memory scales with context length

Because of attention, Transformers require memory

$$\Omega(n \cdot d)$$

n – number of processed tokens, d – dimension of embeddings

Problem

Reduce the memory without sacrificing attention

Approaches:

- Quantization: MiKV [Yang et al., 2024], WKVQuant [Yue et al., 2024]
- **Token pruning**: SnapKV [Li et al., 2024], H20 [Zhang et al., 2024]

LongBench [Bai et al., 2023] tests understanding of a long context ($7K$ words)

Example question: What is the primary conclusion Alice draws?

1. Adaptive heating always fails in real homes
2. User behavior is irrelevant to energy savings
3. ...

SnapKV [Li et al., 2024], H20 [Zhang et al., 2024]: take context tokens with largest attention to question

Token pruning: limitations of prior work

Goal

Design a **question independent** token eviction policy with provable guarantees and strong empirical performance

Token pruning: limitations of prior work

Goal

Design a **question independent** token eviction policy with provable guarantees and strong empirical performance

- Prior algorithms are **question-dependent**
- Prior algorithms are purely heuristic

- Theory

  1. Approximation of the denominator
  2. Approximation of the numerator

- Experiments

Reminder: formula for attention

$$\mathsf{Attn}(q; K, V) := \frac{\exp\left(\frac{\langle k_1, q\rangle}{\sqrt{d}}\right) v_1 + \ldots + \exp\left(\frac{\langle k_n, q\rangle}{\sqrt{d}}\right) v_n}{\exp\left(\frac{\langle k_1, q\rangle}{\sqrt{d}}\right) + \ldots + \exp\left(\frac{\langle k_n, q\rangle}{\sqrt{d}}\right)}$$

Problem setting

At a given moment, the key-value cache

$$K = (k_1, \ldots, k_n),\ V = (v_1, \ldots, v_n)$$

compress it by a factor of 2

$$K_1 \subset K,\ V_1 \subset V;\ K_2 \subset K$$

so that for any fixed q

$$\mathsf{Attn}(q; K, V) \approx \frac{\textsc{Est}_1(q; K_1, V_1)}{\textsc{Est}_2(q; K_2)}$$

Naive approach: uniform sampling

$$\textsc{Est}_1(q; K_1, V_1) := \sum_{(k,v)\in(K_1,V_1)} 2\exp\left(\frac{\langle k, q\rangle}{\sqrt{d}}\right) v,\ K_1, V_1 \sim \mathsf{Unif}(K, V)$$

$$\textsc{Est}_2(q; K_2) := \sum_{k\in K_2} 2\exp\left(\frac{\langle k, q\rangle}{\sqrt{d}}\right),\ K_2 \sim \mathsf{Unif}(K)$$

Naive approach: uniform sampling

$$\textsc{Est}_1(q; K_1, V_1) := \sum_{(k,v)\in(K_1,V_1)} 2\exp\left(\frac{\langle k, q\rangle}{\sqrt{d}}\right) v,\ K_1, V_1 \sim \mathsf{Unif}(K, V)$$

$$\textsc{Est}_2(q; K_2) := \sum_{k\in K_2} 2\exp\left(\frac{\langle k, q\rangle}{\sqrt{d}}\right),\ K_2 \sim \mathsf{Unif}(K)$$

Does not use the geometric properties of the vectors.

Naive approach: uniform sampling

$$\textsc{Est}_1(q; K_1, V_1) := \sum_{(k,v)\in(K_1,V_1)} 2\exp\left(\frac{\langle k, q\rangle}{\sqrt{d}}\right) v,\ K_1, V_1 \sim \mathsf{Unif}(K, V)$$

$$\textsc{Est}_2(q; K_2) := \sum_{k\in K_2} 2\exp\left(\frac{\langle k, q\rangle}{\sqrt{d}}\right),\ K_2 \sim \mathsf{Unif}(K)$$

Does not use the geometric properties of the vectors. Can do better!

- Theory
    1. Approximation of the denominator
    2. Approximation of the numerator
- Experiments

- Theory
  1. Approximation of the denominator
  2. Approximation of the numerator
- Experiments

Denominator approximation

$$\textsc{Est}_2(q; K_2) := \sum_{k \in K_2} 2 \exp\left(\frac{\langle k, q \rangle}{\sqrt{d}}\right)$$

Denominator approximation

$$\textsc{Est}_2(q; K_2) := \sum_{k \in K_2} 2 \exp\left(\frac{\langle k, q \rangle}{\sqrt{d}}\right)$$

Observation

Finding a good K_2

$$\textsc{Est}_2(q; K_2) \approx \sum_{i=1}^{n} \exp\left(\frac{\langle k_i, q \rangle}{\sqrt{d}}\right)$$

is equivalent to finding $\sigma : [n] \to \{+, -\}$

$$\sum_{i=1}^{n} \sigma(i) \cdot \exp\left(\frac{\langle k_i, q \rangle}{\sqrt{d}}\right) \approx 0$$

Phillips-Tai'20, Charikar-K.-Waingarten'24

Vector Balancing Problem

Our goal

Given vectors $\{k_i\}_{i\in[n]}$, find $\sigma : [n] \to \{+,-\}$

$$\sum_{i=1}^{n} \sigma(i) \cdot \exp\left(\frac{\langle k_i, q\rangle}{\sqrt{d}}\right) \approx 0 \text{ for any fixed } q$$

Vector Balancing Problem

Our goal

Given vectors $\{k_i\}_{i\in[n]}$, find $\sigma : [n] \to \{+,-\}$

$$\sum_{i=1}^{n} \sigma(i) \cdot \exp\left(\frac{\langle k_i, q\rangle}{\sqrt{d}}\right) \approx 0 \text{ for any fixed } q$$

Vector Balancing Problem:

Given vectors $\{u_i\}_{i\in[n]}$, find $\sigma : [n] \to \{+,-\}$:

$$\left\langle \sum_{i=1}^{n} \sigma(i) u_i, z \right\rangle \approx 0 \text{ for any fixed } z$$

Vector Balancing Problem

Fact

There exists a mapping ϕ such that for every $k, q \in \mathbb{R}^d$:

$$\langle \phi(k), \phi(q) \rangle = \exp\left(\frac{\langle k, q \rangle}{\sqrt{d}}\right)$$

Algorithm for the Vector Balancing Problem

Need

$$\left| \left\langle \sum_{i=1}^{n} \sigma(i) u_i, z \right\rangle \right| \leq \text{something small}$$

Algorithm for the Vector Balancing Problem

Need

$$\left| \left\langle \sum_{i=1}^{n} \sigma(i) u_i, z \right\rangle \right| \leq \text{something small}$$

What is "something small"?

Algorithm for the Vector Balancing Problem

Need

$$\left| \left\langle \sum_{i=1}^{n} \sigma(i) u_i, z \right\rangle \right| \leq \text{something small}$$

What is "something small"? What if z is known?

If z is known

Keep all prefixes balanced? That is

$$\left|\sum_{i=1}^{j} \sigma(i)\langle u_i, z\rangle\right| \leq \text{something small}$$

for all $j \leq n$?

If z is known

Keep all prefixes balanced? That is

$$\left|\sum_{i=1}^{j} \sigma(i)\langle u_i, z\rangle\right| \leq \text{something small}$$

for all $j \leq n$?

	u_1	u_2	u_3	u_4	u_5
$\langle u_i, z\rangle$	0.1	0.8	0.5	0.7	0.3
signs $\sigma(i)$					
$\sum_i \sigma(i)\langle u_i, z\rangle$					

If z is known

Need

$$\left|\sum_{i=1}^{j} \sigma(i)\langle u_i, z\rangle\right| \leq \text{something small}$$

for all $j \leq n$

	u_1	u_2	u_3	u_4	u_5
$\langle u_i, z\rangle$	0.1	0.8	0.5	0.7	0.3
signs $\sigma(i)$	$+$				
$\sum_i \sigma(i)\langle u_i, z\rangle$	0.1				

If z is known

Need

$$\left|\sum_{i=1}^{j} \sigma(i)\langle u_i, z\rangle\right| \leq \text{something small}$$

for all $j \leq n$

	u_1	u_2	u_3	u_4	u_5
$\langle u_i, z\rangle$	0.1	0.8	0.5	0.7	0.3
signs $\sigma(i)$	$+$	$-$			
$\sum_i \sigma(i)\langle u_i, z\rangle$	0.1	-0.7			

What if z is known?

Need

$$\left| \sum_{i=1}^{j} \sigma(i) \langle u_i, z \rangle \right| \leq \text{something small}$$

for all $j \leq n$

	u_1	u_2	u_3	u_4	u_5
$\langle u_i, z \rangle$	0.1	0.8	0.5	0.7	0.3
signs $\sigma(i)$	$+$	$-$	$+$		
$\sum_i \sigma(i) \langle u_i, z \rangle$	0.1	-0.7	-0.2		

What if z is known?

Need

$$\left|\sum_{i=1}^{j} \sigma(i)\langle u_i, z\rangle\right| \leq \text{something small}$$

for all $j \leq n$

	u_1	u_2	u_3	u_4	u_5
$\langle u_i, z\rangle$	0.1	0.8	0.5	0.7	0.3
signs $\sigma(i)$	$+$	$-$	$+$	$+$	
$\sum_i \sigma(i)\langle u_i, z\rangle$	0.1	-0.7	-0.2	0.5	

What if z is known?

Need

$$\left|\sum_{i=1}^{n} \sigma(i)\langle u_i, z\rangle\right| \leq \text{something small}$$

	u_1	u_2	u_3	u_4	u_5
$\langle u_i, z\rangle$	0.1	0.8	0.5	0.7	0.3
signs $\sigma(i)$	$+$	$-$	$+$	$+$	$-$
$\sum_i \sigma(i)\langle u_i, z\rangle$	0.1	-0.7	-0.2	0.5	0.2

What if z is known?

Need

$$\left|\sum_{i=1}^{n} \sigma(i)\langle u_i, z\rangle\right| \leq \text{something small}$$

	u_1	u_2	u_3	u_4	u_5
$\langle u_i, z\rangle$	0.1	0.8	0.5	0.7	0.3
signs $\sigma(i)$	$+$	$-$	$+$	$+$	$-$
$\sum_i \sigma(i)\langle u_i, z\rangle$	0.1	-0.7	-0.2	0.5	0.2

Proposition

If $|\langle u_i, z\rangle| \leq 1$ for all i then $\left|\sum_{i=1}^{j} \sigma(i)\langle u_i, z\rangle\right| \leq 1$ for all $j \leq n$

Banaszczyk's Theorem

Corollary of Banaszczyk's Theorem

For any $u_1, \ldots, u_n$, $\|u_i\|_2 \leq 1$, there exists a distribution $P : \{+,-\}^n \rightarrow [0,1]$ such that for any z, $\|z\|_2 = 1$:

$$\Pr_{\sigma \sim P}\left[\left|\left\langle \sum_{i=1}^{n} \sigma(i) u_i, z \right\rangle\right| \leq O(\log(n))\right] \geq 1 - \frac{1}{n^{100}}$$

Banaszczyk's Theorem

Corollary of Banaszczyk's Theorem

For any $u_1, \ldots, u_n$, $\|u_i\|_2 \leq 1$, there exists a distribution $P : \{+,-\}^n \rightarrow [0,1]$ such that for any z, $\|z\|_2 = 1$:

$$\Pr_{\sigma \sim P}\left[\left|\left\langle \sum_{i=1}^{n} \sigma(i) u_i, z \right\rangle\right| \leq O(\log(n))\right] \geq 1 - \frac{1}{n^{100}}$$

There is a simple algorithm for sampling from P!

Self-Balancing Walk: algorithm for VBP

input: vectors $u_1, \ldots, u_n$, parameter (normalizer) α

for j from 1 to n

$$w_j = \sum_{i<j} \sigma(i) u_i$$

$$p_j = \frac{1}{2} - \alpha \cdot \langle w_j, u_j \rangle$$

$$\sigma(j) = \begin{cases} +, & \text{with probability } p_j \\ -, & \text{with probability } 1 - p_j. \end{cases}$$

output: σ

Alweiss-Liu-Sawhney'21

Self-Balancing Walk: applied to denominator

input: vectors $k_1, \dots, k_n$, parameter (normalizer) α
for j from 1 to n **do**
$$p_j = \frac{1}{2} - \alpha \cdot \sum_{i<j} \sigma(i) \exp\left(\frac{\langle k_i, k_j \rangle}{\sqrt{d}}\right)$$
$$\sigma(j) = \begin{cases} +, & \text{with probability } p_j \\ -, & \text{with probability } 1 - p_j. \end{cases}$$
end for
output: σ

- Theory
    1. Approximation of the denominator
    2. Approximation of the numerator
- Experiments

- Theory
    1. Approximation of the denominator
    2. Approximation of the numerator
- Experiments

$$\mathsf{Attn}(q; K, V) = \frac{\exp\left(\frac{\langle k_1, q \rangle}{\sqrt{d}}\right) v_1 + \ldots + \exp\left(\frac{\langle k_n, q \rangle}{\sqrt{d}}\right) v_n}{\exp\left(\frac{\langle k_1, q \rangle}{\sqrt{d}}\right) + \ldots + \exp\left(\frac{\langle k_n, q \rangle}{\sqrt{d}}\right)}$$

Numerator approximation

Need

Find $\sigma : [n] \to \{+,-\}$

$$\sum_{i=1}^{n} \sigma(i) \cdot \exp\left(\frac{\langle k_i, q\rangle}{\sqrt{d}}\right) v_i \approx 0 \text{ for any fixed } q$$

Numerator approximation

Need

Find $\sigma : [n] \to \{+,-\}$

$$\sum_{i=1}^{n} \sigma(i) \cdot \exp\left(\frac{\langle k_i, q\rangle}{\sqrt{d}}\right) v_i \approx 0 \text{ for any fixed } q$$

Is this a vector balancing problem instance?

Numerator approximation

Need

Find $\sigma : [n] \to \{+, -\}$

$$\sum_{i=1}^{n} \sigma(i) \cdot \exp\left(\frac{\langle k_i, q\rangle}{\sqrt{d}}\right) v_i \approx 0 \text{ for any fixed } q$$

Is this a vector balancing problem instance? Yes!

Embedding construction

Fact

There exists a mapping ϕ such that for every $k, q \in \mathbb{R}^d$:

$$\langle \phi(k), \phi(q) \rangle = \exp\left(\frac{\langle k, q \rangle}{\sqrt{d}}\right)$$

Embedding construction

Fact

There exists a mapping ϕ such that for every $k, q \in \mathbb{R}^d$:

$$\langle \phi(k), \phi(q) \rangle = \exp\left(\frac{\langle k, q \rangle}{\sqrt{d}}\right)$$

Define ψ:

$$\psi(k, v) := \phi(k) \otimes v$$

where $\otimes$ is the tensor product

Self-Balancing Walk: applied to numerator

input: pairs of vectors $(k_1, v_1), \ldots, (k_n, v_n)$, parameter (normalizer) α

for j from 1 to n **do**

$$p_j = \frac{1}{2} - \alpha \cdot \sum_{i<j} \sigma(i) \exp\left(\frac{\langle k_i, k_j \rangle}{\sqrt{d}}\right) \langle v_i, v_j \rangle$$

$$\sigma(j) = \begin{cases} +, & \text{with probability } p_j \\ -, & \text{with probability } 1 - p_j. \end{cases}$$

end for

output: σ

Formal Problem Statement

Formal problem statement

Minimize K_1, V_1, K_2 so that for any q:

$$\left\| \mathsf{Attn}(q; K, V) - \frac{\textsc{Est}_1(q; K_1, V_1)}{\textsc{Est}_2(q; K_2)} \right\|_2 \le \varepsilon \cdot \|\mathsf{softmax}\left(K^T \cdot q\right)\|_2 \cdot \|V\|_F.$$

where

$$[\mathsf{softmax}\left(K^T \cdot q\right)]_j := \frac{\exp\left(\frac{\langle k_j, q\rangle}{\sqrt{d}}\right)}{\sum_{i=1}^{n} \exp\left(\frac{\langle k_i, q\rangle}{\sqrt{d}}\right)}$$

Theoretical Guarantees

Uniform sampling:

$$\approx \frac{d}{\varepsilon^2}\mathsf{poly}(\log(n))$$

Theoretical Guarantees

Uniform sampling:

$$\approx \frac{d}{\varepsilon^2}\mathsf{poly}(\log(n))$$

BalanceKV:

$$\approx \frac{d^{1.5}}{\varepsilon}\mathsf{poly}(\log(n))$$

Is this optimal?

Streaming Attention Approximation Space Complexity

Lower bound

Any ε-approximation algorithm has space complexity

$$\approx \Omega\left(\min\left\{\frac{1}{\varepsilon^2}, d\right\}\right)$$

Reduction from INDEX

Streaming Attention Approximation Space Complexity

Lower bound

Any ε-approximation algorithm has space complexity

$$\approx \Omega\left(\min\left\{\frac{1}{\varepsilon^2}, d\right\}\right)$$

Reduction from INDEX

Attention between (q, k, v) and $(q_1, k_1, v_1), \ldots, (q_n, k_n, v_n)$:

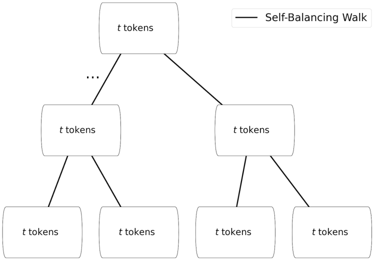
$$\text{Attn}(q; K, V) := \mathbb{E}_{i \sim \text{Softmax}} v_i$$

where $\Pr[\text{Softmax} = i] \propto \exp\left(\frac{\langle k_i, q \rangle}{\sqrt{d}}\right)$

Streaming?

Streaming?

Streamable by merge and reduce



- Theory
  1. Approximation of the denominator
  2. Approximation of the numerator
- Experiments

- Theory

  1. Approximation of the denominator
  2. Approximation of the numerator

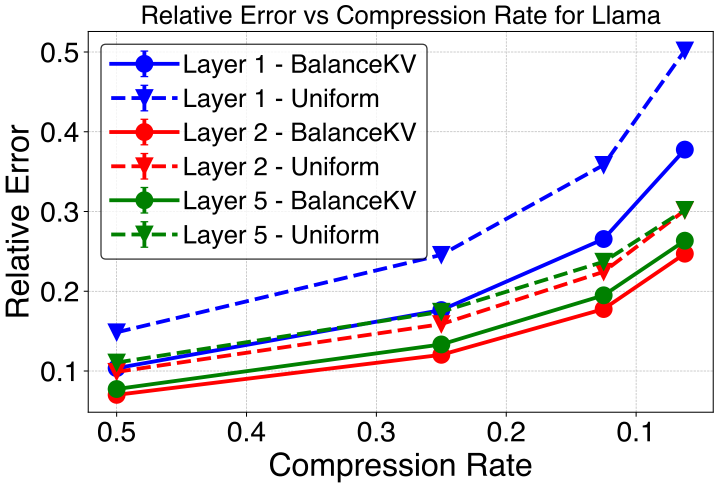
- Experiments

Models

`Llama-3.1-8B-Instruct`

`Qwen-2.5-14B/32B-Instruct`

Single layer attention approximation



Benchmarks

LongBench [Bai et al., 2023]

- Tests understanding of a long context ($7K$ words)

Example question: What is the primary conclusion Alice draws?

1. Adaptive heating always fails in real homes
2. User behavior is irrelevant to energy savings
3. ...

Needle-In-A-Haystack [Kamradt, 2023]

- Tests the ability to preserve surprising information in a context of $4K - 100K$ tokens
- **Example context:** an essay with inserted "The 5 best things to do in San Francisco are: ..."
- **Example question:** "What are the 5 best things to do in San Franscisco?"

End-to-end Evalution on LongBench

Context+question compression by a factor of 4

Method	Qwen2.5-32B	Qwen2.5-14B	Llama-3.1-8B
Exact (Baseline)	51.77	54.14	50.17
StreamingLLM	35.52	38.71	39.79
PyramidKV	47.13	49.80	45.64
SnapKV	48.77	50.22	46.12
Uniform	48.76	49.88	46.38
BalanceKV	**48.84**	**50.62**	**46.77**

Metrics: F1, Rouge-L, Accuracy, Edit Distance

Needle-In-A-Haystack

Context+question compression by a factor of 4

Heuristic: pick out the most "surprising" tokens and compress the rest

Method	Average Accuracy
SnapKV	0.83
PyramidKV	0.90
StreamingLLM	0.31
Uniform Sampling	0.90
BalanceKV	**0.99**

Future directions

- Tight bounds? Perhaps via data-dependent LSH?

Future directions

- Tight bounds? Perhaps via data-dependent LSH?
- Learned algorithms for discrepancy minimization?

Future directions

- Tight bounds? Perhaps via data-dependent LSH?
- Learned algorithms for discrepancy minimization?
- Applying discrepancy to feedforward layers?

Future directions

- Tight bounds? Perhaps via data-dependent LSH?
- Learned algorithms for discrepancy minimization?
- Applying discrepancy to feedforward layers?

Theory	Practice
SGD for convex functions	SGD for neural networks
Discrepancy-based methods	???

Future directions

- Tight bounds? Perhaps via data-dependent LSH?
- Learned algorithms for discrepancy minimization?
- Applying discrepancy to feedforward layers?

Theory	Practice
SGD for convex functions	SGD for neural networks
Discrepancy-based methods	???

Questions?

References I

Alweiss, R., Liu, Y. P., and Sawhney, M. (2021).
Discrepancy minimization via a self-balancing walk.
Proceedings of the 53rd ACM Symposium on the Theory of Computing (STOC '2021).

Bai, Y., Lv, X., Zhang, J., Lyu, H., Tang, J., Huang, Z., Du, Z., Liu, X., Zeng, A., Hou, L., et al. (2023).
Longbench: A bilingual, multitask benchmark for long context understanding.
arXiv preprint arXiv:2308.14508.

Kamradt, G. (2023).
Needle in a haystack-pressure testing llms.
Github Repository, page 28.

References II

Li, Y., Huang, Y., Yang, B., Venkitesh, B., Locatelli, A., Ye, H., Cai, T., Lewis, P., and Chen, D. (2024).
Snapkv: Llm knows what you are looking for before generation.
arXiv preprint arXiv:2404.14469.

Yang, J. Y., Kim, B., Bae, J., Kwon, B., Park, G., Yang, E., Kwon, S. J., and Lee, D. (2024).
No token left behind: Reliable kv cache compression via importance-aware mixed precision quantization.
arXiv preprint arXiv:2402.18096.

Yue, Y., Yuan, Z., Duanmu, H., Zhou, S., Wu, J., and Nie, L. (2024).
Wkvquant: Quantizing weight and key/value cache for large language models gains more.
arXiv preprint arXiv:2402.12065.

References III

Zhang, Z., Sheng, Y., Zhou, T., Chen, T., Zheng, L., Cai, R., Song, Z., Tian, Y., Ré, C., Barrett, C., et al. (2024).
H2o: Heavy-hitter oracle for efficient generative inference of large language models.
Advances in Neural Information Processing Systems, 36.