Curse of High Dimensionality Issue in Transformer for Long-context Modeling

South China University of Technology

Shuhai Zhang*, Zeng You*, Yaofo Chen, Zhiquan Wen, Qianyue Wang, Zhijie Qiu, Yuanqing Li, Mingkui Tan†

shuhaizhangshz@gmail.com mingkuitan@scut.edu.cn

Code Paper

Background

Long-context modeling faces computational inefficiencies due to **redundant** attention computations: while attention weights are often **sparse**, all tokens consume **equal** computational resources.

From Sequence Modeling to Supervised Reformulation

- Given a token sequence $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L\}$, the **traditional probabilistic sequence modeling** is:

$$\max_{\theta} \prod_{i=1}^{L} P_\theta(\mathbf{x}_i | \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{i-1}).$$

- Given a training corpus $\{(C(\mathbf{y}_i), \mathbf{y}_i)\}_{i=1}^{n}$, where $\mathbf{y}_i \in \mathcal{V}$ is a token from the vocabulary $\mathcal{V}$, $C(\mathbf{y}_i) = \{\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}\}$ is its **context**, we can model sequence modeling as a **supervised learning task**:

$$\max_{\theta} \prod_{\mathbf{y} \in \mathcal{V}, C(\mathbf{y})} P_\theta(\mathbf{y} | C(\mathbf{y})).$$

- Observation: By casting the problem as supervised learning, we can precisely determine which tokens in $C(\mathbf{y})$ are vital for predicting $\mathbf{y}$, for which we can assign more computations.

$$C(\mathbf{y}) = \{\mathbf{x}_1^{\mathrm{R}}, \mathbf{x}_2^{\mathrm{IR}}, \mathbf{x}_3^{\mathrm{R}}, \mathbf{x}_4^{\mathrm{IR}}, \mathbf{x}_5^{\mathrm{IR}}, \ldots, \mathbf{x}_{L-1}^{\mathrm{R}}, \mathbf{x}_L^{\mathrm{IR}}\},$$

where $\mathbf{x}^{\mathrm{R}}$ denotes relevant tokens that contribute to predicting the target token $\mathbf{x}$, and $\mathbf{x}^{\mathrm{IR}}$ denotes **irrelevant tokens** that introduce unnecessary noise or computation.

Redundancy in Self-attention Mechanism

Self-attention: $\mathbf{Att} = \mathrm{softmax}\left(\mathbf{Q}\mathbf{K}^\top/\sqrt{d}\right)\mathbf{V}$, where $\mathbf{Q}=\mathbf{X}\mathbf{W}^Q$, $\mathbf{K}=\mathbf{X}\mathbf{W}^K$, and $\mathbf{V}=\mathbf{X}\mathbf{W}^V$.

- **Theorem 1.** ***(Sparsity on attention weights)*** Consider $\rho \in (1/L, 1]$ as a sparse rate, we say that the weight α is ρ-sparse when there exists at least one probability greater than $1/(L\rho)$. Let $\xi = \mathbf{K}\mathbf{Q}_i \in \mathbb{R}^L$, then the probability of α being ρ-sparse $P_{sparse}(L, \rho)$ is given by

$$P_{sparse}(L, \rho) \geq \max_{x>0} 1 - [P_{head} P_{tail}]^L,$$

where $P_{head} = P\{\exp(\xi_j) \leq x\}$ and $P_{tail} = P\{(L\rho - 1)x \leq \sum_{k \neq j}^{L} \exp(\xi_k)\}$. For sufficiently large L, the exponential amplification of L suppresses $P_{head}P_{tail}$, leading to a large $P_{sparse}(L, \rho)$.

- Simplifying the optimization of self-attention as a ***linear coding problem***:

$$\min_{\alpha} \left\| \sum_{j=1}^{L} \alpha_j \mathbf{V}_j - \mathbf{y} \right\|_2^2, \text{ s.t. }, \sum_{j=1}^{L} \alpha_j = 1,\ \alpha_j > 0,$$

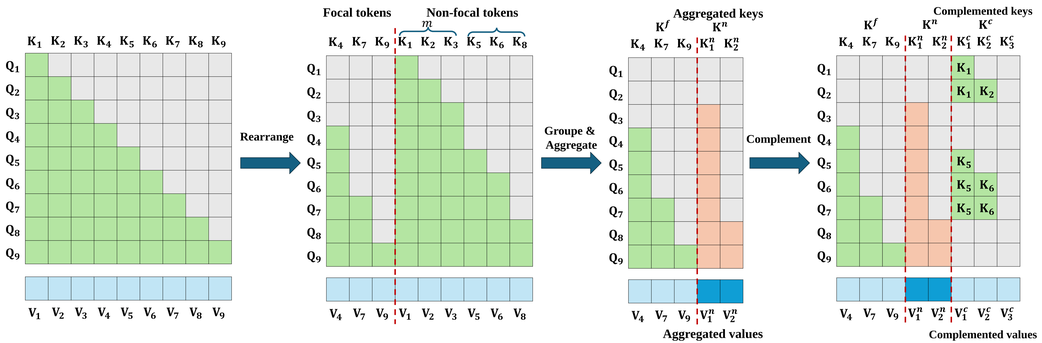
where α_j is the weight of the j-th token's $\mathbf{V}_j$ and $\mathbf{y}$ is the embedding of the target token.

- Since achieving stable and efficient learning is often hindered by noise sensitivity and overfitting to ***irrelevant tokens***, we can model the optimization as a ***group coding problem***:

$$\min_{\bar{\alpha}} \left\| \sum_{g=1}^{k} \frac{\bar{\alpha}_g}{|G_g|} \sum_{j \in G_g} \mathbf{V}_j - \mathbf{y} \right\|_2^2, \text{ s.t. }, \sum_{g=1}^{k} \alpha_g = 1,\ \alpha_g > 0,$$

where $\bigcup_{g=1}^{k} G_g = \{1, \ldots, L\}$ and $G_i \neq G_j$ for $\forall i \neq j$, the group weight $\bar{\alpha}_g = \frac{1}{|G_g|} \sum_{i \in G_g} \alpha_i$. This group mechanism **reduces both optimization dimensionality** and **attention redundancy**, cutting computational overhead while enhancing model robustness and efficiency.

Dynamic Group Attention (DGA) for Long-context Modeling



- Fast focal token identification for grouping:

$$s_i = \frac{\sum_{j=1}^{L} \widetilde{\mathbf{A}}_{j,i}}{\|\widetilde{\mathbf{A}}_{\cdot,i}\|_0}, \quad \widetilde{\mathbf{A}} = \mathrm{softmax}\left(\widetilde{\mathbf{Q}}\mathbf{K}^\top/\sqrt{d}\right).$$

Use s to divide tokens into **focal** and **non-focal tokens**.

$$\mathbf{K}^f = [\mathbf{K}'_1; \ldots; \mathbf{K}'_r],\ \mathbf{V}^f = [\mathbf{V}'_1; \ldots; \mathbf{V}'_r] \in \mathbb{R}^{r \times d}.$$

- Group and aggregate keys and values in i-th group:

$$\mathbf{K}_i^n = \sum_{j \in G_i} [P_{ij}\mathbf{K}'_j],\ \mathbf{V}_i^n = \sum_{j \in G_i} [P_{ij}\mathbf{V}'_j],$$

where $\mathbf{P}_{i,\cdot} = \mathrm{softmax}(\mathbf{Q}_{\max G_i}\mathbf{K}^\top_{[\min G_i : \max G_i]})$.

- Concatenate: $\mathbf{K}^{\mathrm{group}} = [\mathbf{K}^f; \mathbf{K}^n; \mathbf{K}^c]$, $\mathbf{V}^{\mathrm{group}} = [\mathbf{V}^f; \mathbf{V}^n; \mathbf{V}^c]$.
- Construct complementary KV pairs for i-th query $\mathbf{Q}_i$: $\mathbf{K}^c = [\mathbf{K}_j], \mathbf{V}^c = [\mathbf{V}_j], j \in G_z$, if $i \in [\min G_z, \max G_z]$.
- Attention for $\mathbf{Q}_i$: $\mathbf{Att}_i = \sum_j (\mathbf{A}_{i,j}^{\mathrm{group}} \odot \mathbf{M}_{i,j})\mathbf{V}_j^{\mathrm{group}}$, $\mathbf{A}_i^{\mathrm{group}} = \mathrm{softmax}\left(\mathbf{Q}_i\mathbf{K}^{\mathrm{group}\top}/\sqrt{d}\right)$, $\mathbf{M}$ is the causing mask.

Comparisons with SOTA Methods

- **Results on LongBench-E, where the 95% text length quantile is 31K. "ITL" denotes inter-token latency.**

Methods	Single-Doc. QA	Multi-Doc. QA	Summary	FS. Learning	Synthetic	Code	Avg. ↑	ITL / ms ↓
LLaMA2-7B (Vanilla)	6.43	2.37	13.65	56.65	3.04	48.0	**21.69**	69.70
MInference	5.86	2.65	14.33	55.99	2.63	48.41	21.64	94.34
StreamingLLM	4.99	4.13	11.51	45.43	2.16	30.38	16.43	78.28
LM-Infinite	3.54	2.61	3.31	48.97	1.33	35.26	15.84	102.22
DGA-LLM (Ours)	3.61	3.58	6.81	57.90	1.47	53.45	21.14	**28.80**

Contributions

- We reformulate the sequence modeling as a **supervised learning** task, enabling the separation of **relevant** and **irrelevant** tokens in long-context modeling. It provides a clear understanding of redundancy and inspires more efficient attention.
- We theoretically analyze attention sparsity: only a few tokens significantly contribute to the target representation. Framing attention optimization as a linear coding, we propose a **group coding** strategy, showing its **robustness** and **efficiency**.
- We propose Dynamic Group Attention (DGA) by **aggregating less important tokens into meaningful groups** to reduce redundancy in attention computation while maintaining accuracy for long-context modeling.

More Experiments

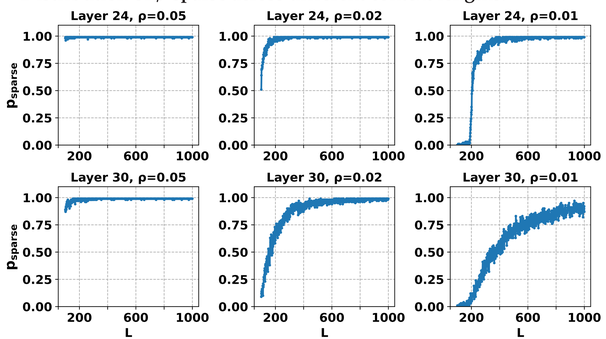
- **Results on EM score evaluated at lengths from 4K to 32K.**

Methods	4K	8K	16K	32K	Avg.
LLaMA2-7B (Vanilla)	**37.2**	**36.4**	**33.8**	**26.8**	**33.6**
StreamingLLM	30.2	25.8	22.2	20.8	24.8
LM-Infinite	29.4	28.6	23.8	22.4	26.1
MInference	29.2	24.8	23.6	17.0	23.7
DGA-LLM (Ours)	35.0	27.4	25.6	22.6	27.7

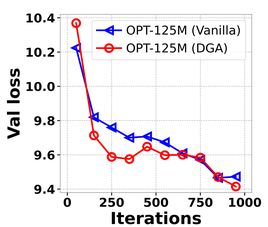
- **Results on LongBench-E over ITL, where "ITL" is inter-token latency, measuring the time delay between generating consecutive tokens.**

Methods	4K	8K	16K
LLaMA2-7B (Vanilla)	36.87	42.32	69.70
MInference	93.13	93.93	94.34
StreamingLLM	40.00	46.67	78.28
LM-Infinite	49.70	64.34	102.22
DGA-LLM (Ours)	**26.26**	**26.87**	**28.79**

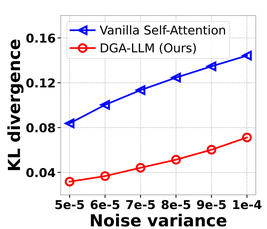
- **Distribution of ρ-sparse across different context lengths.**



- **Optimization efficiency and robustness.**



Optimization efficiency



Robustness