Advances in Computer, Signals and Systems (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2024.080618
ISSN 2371-8838 Vol. 8 Num. 6

Sparse Attention Mechanisms in Large Language Models: Applications, Classification, Performance Analysis, and Optimization

Jingxuan Bai[1,*]

[1]*School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, 100083, China*
[*]*Corresponding author: baijingxuan1013@126.com*

Keywords: Sparse Attention Mechanism, Large Language Models, Performance Improvement Strategies, Transformer Model, Time Complexity

Abstract: This paper explores the application and performance analysis of sparse attention mechanisms in large language models (LLMs), highlighting their ability to reduce the computational complexity of the traditional Transformer architecture for long sequences, it also reviews various sparse attention strategies that enhance efficiency by minimizing token interactions while preserving model performance, addressing the limitations of conventional models. A novel classification framework categorizes these mechanisms into global, local, and hybrid strategies. Through performance analyses of key models such as Longformer, Reformer, and BIGBIRD, this paper demonstrates their advantages in tasks like document understanding, information extraction, and image generation. Additionally, this paper proposes strategies for performance enhancement, including multimodal potential, integration with knowledge distillation, and anchor-based methods, to further optimize the effectiveness of sparse attention mechanisms in large language models and identify their potential pathways for development. These contributions provide a comprehensive understanding for beginners studying sparse attention mechanisms and offer possible directions for future research to improve performance and efficiency in large-scale NLP tasks.

1. Introduction

Large language models have brought about a revolutionary shift in the field of natural language processing (NLP), excelling in core tasks such as language generation, text classification, and machine translation [1]. The Transformer's global attention mechanism allows interactions between all tokens, enabling comprehensive contextual modeling and establishing it as a key NLP innovation [2]. However, as task scales expand and long-sequence text problems emerge, traditional models such as the Transformer expose significant computational bottlenecks. Its global attention leads to an $O(L^2)$ complexity. It means that as sequence length increases, the computational and memory demands of the Transformer grow exponentially, severely limiting its application in long-text processing [2].

Sparse attention mechanisms address this issue by reducing token interactions, significantly lowering computational complexity while preserving model performance, making them ideal for long-sequence tasks [3]. Recent large language models such as Longformer, BIGBIRD, and Reformer, have demonstrated the effectiveness of sparse attention in handling long texts [4]. However, there remains space for optimization in the effectiveness of sparse attention. How to enhance model performance while further reducing computational costs remains a key focus of current research.

This paper introduces the key applications of sparse attention mechanisms in large language models across NLP and other domains. It then proposes a novel classification method and conducts a comprehensive performance analysis of classical models based on this classification. Finally, it explores potential optimization strategies, offering specific recommendations aimed at enhancing the efficiency of sparse attention mechanisms.

2. Applications of Sparse Attention Mechanisms in Large Language Models

Building on the introduction of sparse attention mechanisms as a solution to the limitations of traditional Transformers in large language models (LLMs), this section summarizes their practical applications in key tasks across both NLP and other domains, illustrating the importance of sparse attention mechanisms in improving LLM performance.

Document Understanding and Question-Answering Tasks: In the domain of tasks that necessitate the comprehension of extensive documents and the engagement in question-answering paradigms, Longformer shows significant advantages over traditional Transformer models. The attention mechanism in traditional Transformers cannot handle inputs longer than 512 tokens [5]. In contrast, Longformer, utilizing sparse attention, can process sequences up to 4096 tokens [3]. Additionally, in question-answering tasks like WikiHop and TriviaQA, Longformer has outperformed RoBERTa and set new state-of-the-art (SOTA) records [3].

Information Extraction and Text Classification Tasks: Within the information extraction tasks (e.g., summarization, named entity recognition) and the text classification tasks (e.g., sentiment analysis, topic classification), sparse attention mechanisms offer notable advantages. Reformer, leveraging Locality-Sensitive Hashing (LSH) attention, significantly may reduces the memory usage and computational cost for processing long sequences [6]. And the experiments on the enwik8 dataset demonstrate that Reformer achieves much faster processing speeds than traditional Transformers while maintaining comparable accuracy [6].

The Image Generation Tasks: Empirical studies conducted on datasets such as CIFAR-10 and ImageNet-64 have illuminated that the Sparse Transformer is capable of generating unconditional image samples of high fidelity, characterized by global coherence and diversity[7]. Compared to traditional Transformers, Sparse Transformer shows significant improvement in handling long-term dependencies, making it more suitable for large-scale image generation tasks with strong dependency requirements [7].

3. A New Classification Framework for Sparse Attention Mechanisms and Performance Analysis of Representative Models in This Classification

3.1. Classification

Sparse attention mechanisms are designed to enhance computational efficiency by mitigating the extent of token-to-token interactions. This can be achieved through two main approaches: The first methodology entails a global attentiveness to all tokens, selecting a subset for interaction, while the other limits the tokens attended to from the beginning, interacting only with the selected few.

Therefore, based on the number of tokens attended to, sparse attention mechanisms can be classified into three types: global, local, and a combination of both. The global strategy attends to all tokens but limits interaction using mechanisms like Locality-Sensitive Hashing (LSH), while the local strategy narrows focus to a smaller subset (e.g., through sliding windows or random attention) and interacts only within this scope. The combined global-local strategy utilizes both approaches simultaneously at the same time.

3.2. Performance Analysis of Classic Models

3.2.1. Global Approximation Model: Reformer

The Reformer model leverages Locality-Sensitive Hashing (LSH) attention to optimize long-sequence processing efficiency [6]. Compared to Transformer, Reformer reduces token interaction range while maintaining similar performance, lowering the computational complexity from $O(L^2)$ to $O(LlogL)$, significantly improving its ability to handle long sequences [6].

Random Projection: Each input token is projected into a lower-dimensional subspace, ensuring sparse distribution of the queries and the keys [6].

Bucket Assignment: Tokens are allocated to discrete buckets based on their hashed values only allows the tokens within the same or adjacent buckets interact [6].

After bucketing, the sequence is divided into chunks for parallel computation. Each token attends only to tokens within same bucket in own chunk or previous chunk, reducing interactions while maintaining some long-range dependencies [6].

$$LSH\ Attention(Q,K,V) = soft\max(\frac{QK^T}{\sqrt{d_k}} \cdot M(Q,K))V \tag{1}$$

Where $M(Q,K)$ is an encoding matrix that indicates whether q_i and k_j belong to the same bucket. If both of them are in the same bucket, then $M_{i,j} = 1$; otherwise $M_{i,j} = 0$.

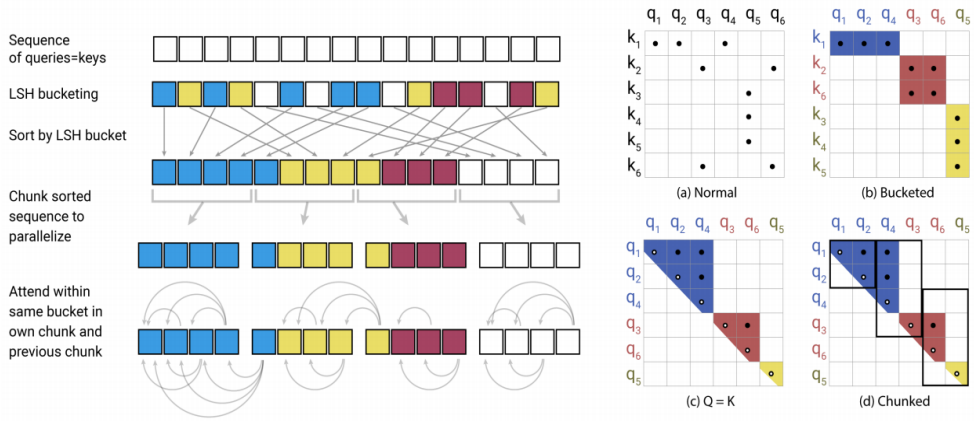

Figure 1: Simplified illustration of LSH Attention in Reformer.

According to figure 1, (a) Full attention where each query attends to all tokens, but at high computational cost [6]. (b) LSH attention groups tokens into hash buckets, restricting attention within each bucket, though uneven bucket sizes complicate processing [6]. (c) Shared transformations for queries and keys ensure similar tokens group together, simplifying computation [6]. (d) Splitting the sequence into chunks limits interactions to the same or previous chunk, reducing costs while preserving long-range dependencies [6].

3.2.2. Local Approximation Model: Longformer

The Longformer constitutes an attention model predicated on localities, which diminishes the expanse of token attention via its proprietary local attention mechanism, optimizing the efficiency of processing long sequences [3]. For the tasks which involving long sequences, compared to the Transformer model, Longformer achieves similar performance by attending to a smaller range of tokens, reducing the computational complexity from $O(L^2)$ to $O(L)$, significantly improving efficiency [3].

Sliding Window Attention: Longformer uses a sliding window attention mechanism, where each token attends to neighboring tokens within a limited range above and below in the sequence, capturing broader context as layers stack while maintaining efficiency [3].This mechanism effectively reduces the computational complexity to $O(wL)$, where w is the window size. Even as sequence length grows, Longformer retains the ability to capture long-range dependencies [3].

Dilated Sliding Window Attention: In NLP, words that occupy similar positions within a sentence typically convey analogous semantic information [8]. Longformer extends this mechanism by introducing dilated gaps, allowing tokens to attend to farther ones, by reducing the capture of closely related redundant information, while maintaining efficiency and expanding the contextual range [3].

$$Attention(Q,K,V) = soft\max(\frac{QK^T}{\sqrt{d_k}})V \tag{2}$$

Where Q,K,V are the projection matrices based on window size.

Longformer primarily uses windowed attention, but can incorporate global attention in specific tasks, allowing key tokens (such as question or [CLS] tokens) to attend to all others in the sequence, improving performance in classification or question-answering tasks [3].

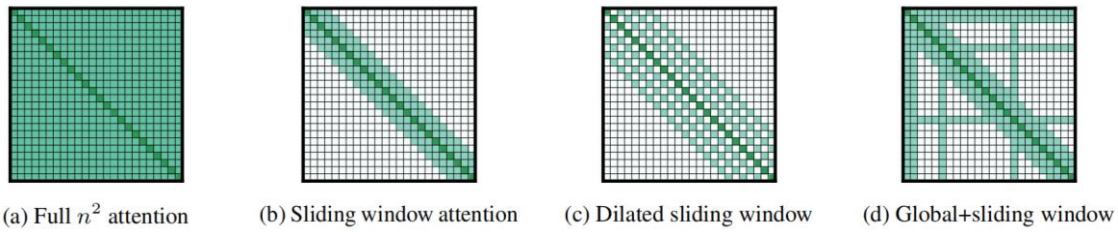

Figure 2: Comparison of self-attention patterns, showing diagonal lines for standard attention and varying shades indicating attention strength in Longformer.

According to figure 2, the full attention mechanism (a) allows all tokens to interact with each other but has high computational costs [3]. The sliding window attention (b) limits interactions to a fixed number of neighboring tokens, reducing complexity [3]. The dilated sliding window (c) expands the receptive field by introducing gaps between attended tokens, while still keeping computational costs low [3]. Finally, the global attention combined with sliding window attention (d) allows certain tokens to attend to the entire sequence, optimizing both local and global context processing [3].

3.2.3. Global and Local Hybrid Strategy Model: BIGBIRD

The BIGBIRD model is an instantiation of a sparse attention mechanism that amalgamates global and local attentiveness strategies, purposefully architected to facilitate the expeditious handling of tasks involving extended sequences [4]. Compared to traditional Transformer models,

BIGBIRD reduces the attention range, and lowering the computational complexity from $O(L^2)$ to linear O(L) while maintaining comparable accuracy [4]. It incorporates a combination of global tokens, local window attention, and random attention, thereby reducing the number of tokens attended to while capturing more comprehensive information.

BIGBIRD employs Sliding Window Attention and Global Attention, where in a subset of tokens engages exclusively with their proximate counterparts, while key tokens interact with the entire sequence. This configuration effectively reduces computational complexity compared to Transformer, while capturing contextual information more effectively. A distinguishing feature of BIGBIRD is Random Attention, allowing the model to randomly connect some distant tokens [4].

$$Attention(Q,K,V) = soft\max(\frac{QK^T}{\sqrt{d_k}} \cdot M(Q,K))V \tag{3}$$

Where $M(Q,K)$ represents the masking matrix; if q_i attends to k_j, then $M_{i,j} = 1$; otherwise $M_{i,j} = 0$.

Theoretically, under the condition of identical sliding window sizes and the same number of attended tokens, BIGBIRD exhibits superior robustness compared to Longformer. Therefore, it can be inferred that for more complex generative tasks—especially for those involving greater uncertainty, latent relationships, or numerous local dependencies—BIGBIRD's random attention mechanism can capture a wider variety of connections, potentially leading to improved performance. However, in the tasks where structural clarity and the global dependencies of certain tokens are critically important (such as question-answering tasks), Longformer, with its ability to define global token positions manually, may demonstrate superior performance.

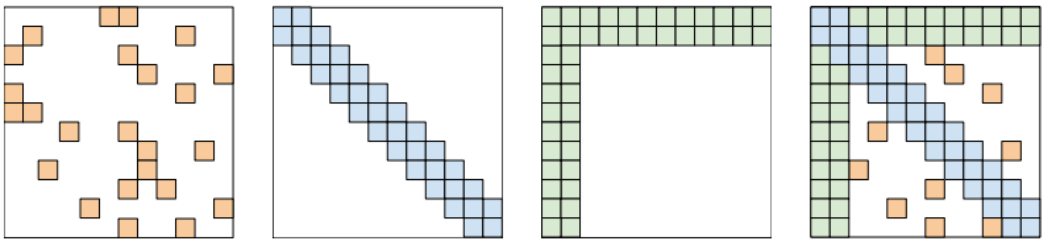

Figure 3: Overview of the attention mechanism components in BIGBIRD.

According to figure 3, (a) Random attention selects a subset of tokens to reduce complexity while maintaining some global context [4]. (b) Sliding window attention limits token interactions to nearby tokens, focusing on local context [4]. (c) Global attention allows certain tokens to attend to all tokens, providing broader sequence awareness [4]. (d) The BIGBIRD model combines these approaches, balancing efficiency and global-local interactions for long-sequence processing [4].

4. Performance Enhancement Strategies

4.1. Multimodal Potential

Recent scholarly inquiries have illuminated that models such as Sparse Transformer and Reformer exhibit superior performance over traditional Transformers within single-domain tasks like text extraction and image generation [7]. Notably, Transformers have been widely applied in multimodal tasks (such as visual-language integration, image captioning, and visual question answering), demonstrating robust multimodal learning capabilities [9]. Sparse attention mechanisms efficiently capture long-range dependencies, significantly reducing computational overhead in

multimodal tasks, thereby excelling when handling large-scale multimodal data, such as images and texts [7]. The strategy of sparse connectivity not only maintains model efficiency but also enhances the model's adaptability to complex multimodal inputs [10]. Therefore, Transformer-based architectures, along with the strong performance in single-task scenarios, provide great potential for sparse attention models in multimodal tasks, warranting further exploration.

4.2. Integration with Knowledge Distillation

Current optimization methodologies predominantly rely on dynamic adjustment of the model's sparse connections, an approach that entails considerable computational expense due to the necessity of processing for each input [11]. Knowledge distillation enables the student model to inherit the teacher model's effective sparse connection patterns [12]. By introducing knowledge distillation, it is expected that this overhead can be effectively reduced. In this framework, the global attention model can be trained as a teacher model to acquire weight information during the training process. Subsequently, this weight experience is transferred to the student model, allowing it to utilize learned representations without the need for dynamic adjustments for each input and directly determine the optimal sequence of sparse connection configurations, or enabling it to apply directly in similar tasks. This method significantly reduces the computational overhead associated with recalibrating sparse connections and is particularly suited for efficiently processing domain-specific tasks such as news summarization, legal document extraction, and medical literature analysis, improving their processing efficiency and accuracy.

4.3. Replacing LSH with Anchor-based Methods

By combining the anchor-based methods with the core concepts of Locality-Sensitive Hashing (LSH), the anchor points can be viewed as representative points in hash buckets, achieving local clustering effects similar to those of LSH. Unlike LSH, which requires adjusting the hash functions to optimize similarity searches, anchor methods conduct distribution learning through subsets of data, allowing for greater flexibility in adapting to varying data distributions [13]. This renders anchor-based methods more malleable and adaptable when navigating complex, nonlinear high-dimensional data spaces. Moreover, the efficacy of anchor methods on large-scale datasets.

5. Conclusion

This paper has explored the transformative role of sparse attention mechanisms in large language models (LLMs), demonstrating their effectiveness in overcoming the computational limitations of traditional Transformer architectures. As the demand for processing longer sequences continues to grow, sparse attention mechanisms provide a viable solution by significantly reducing the $O(L^2)$ complexity associated with global attention, thereby enhancing efficiency and scalability.

This paper's review of models such as Longformer, Reformer, and BIGBIRD highlights their high performance across various tasks, including document understanding, information extraction, and image generation. By categorizing these models into global, local, and hybrid strategies, readers can gain a clearer understanding of the objectives and implementation methods of sparse attention mechanisms. Furthermore, this paper has theoretically proposed several optimization strategies, such as integrating with knowledge distillation and utilizing anchor-based methods, aimed at further enhancing the performance of sparse attention in large language models and outlining potential future development directions for these models.

Looking ahead, the ongoing evolution of sparse attention mechanisms presents exciting opportunities for further research. Future work can delve deeper into integrating these mechanisms

with emerging technologies in machine learning, thereby enhancing their applicability in complex tasks. Ultimately, the insights from this study offer a solid foundation for understanding sparse attention mechanisms and contribute to ongoing efforts in improving large-scale NLP models.

References

[1] Patil R, Gudivada V. A Review of Current Trends, Techniques, and Challenges in Large Language Models (LLMs). Appl Sci. 2024; 14(5):2074.

[2] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention Is All You Need. Adv Neural Inf Process Syst. 2017; 30: 5998–6008.

[3] Beltagy I, Peters ME, Cohan A. Longformer: The Long-Document Transformer. arXiv preprint [cs.CL]. 2020.

[4] Zaheer M, Guruganesh G, Dubey KA, Ainslie J, Alberti C, Ontanon S, Pham P, Ravula A, Wang Q, Yang L, Ahmed A. Big Bird: Transformers for Longer Sequences. Adv Neural Inf Process Syst. 2020; 33: 17283-17297.

[5] Hao C, Zhang P, Xie M, Zhao D. Recurrent Transformers for Long Document Understanding. In: CCF International Conference on Natural Language Processing and Chinese Computing. Cham: Springer Nature Switzerland; 2023. p. 57-68.

[6] Kitaev N, Kaiser Ł, Levskaya A. Reformer: The Efficient Transformer. arXiv preprint [cs.LG]. 2020.

[7] Child R, Gray S, Radford A, Sutskever I. Generating Long Sequences with Sparse Transformers. arXiv preprint [cs.LG]. 2019.

[8] Griffiths TL, Steyvers M. (2004) Distributional semantics and the problem of semantic similarity. Proc Natl Acad Sci U S A, 101: 8171-8176.

[9] Tan, H., & Bansal, M. (2019). LXMERT: Learning Cross-Modality Encoder Representations from Transformers. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1-14.

[10] Tay Y, Dehghani M, Abnar S, Shen Y, Bahri D, Pham P, Rao J, Yang L, Ruder S, Metzler D. Long range arena: A benchmark for efficient transformers. arXiv preprint arXiv:2011.04006. 2020 Nov 8.

[11] Mostafa H, Wang X. Parameter efficient training of deep convolutional neural networks by dynamic sparse reparameterization. InInternational Conference on Machine Learning 2019 May 24 (pp. 4646-4655). PMLR.

[12] Hinton G. Distilling the Knowledge in a Neural Network. arXiv preprint arXiv:1503.02531. 2015.

[13] De Kergorlay HL, Higham DJ. Consistency of anchor-based spectral clustering. Information and Inference: A Journal of the IMA. 2022 Sep; 11(3):801-822.