Toward Energy-Efficient Transformers: Recent Trends in Optimization Techniques

Jaehyun Chung, Seungcheol Oh, and Joongheon Kim
Department of Electrical and Computer Engineering, Korea University, Seoul, Republic of Korea
E-mails: {rupang1234, seungoh, joongheon}@korea.ac.kr

***Abstract*—With the rapid advancement of artificial intelligence technologies, transformer-based models have garnered significant attention by achieving outstanding performance across a wide range of application domains, including natural language processing, computer vision, and speech recognition. However, the increasing scale of these models has led to substantial growth in computational demands and memory consumption, thereby raising critical concerns regarding energy efficiency and sustainability. As transformer models continue to evolve and integrate into diverse real-world applications, the need for efficient and environmentally conscious design has become increasingly vital. These issues are particularly pressing in environments with limited resources, such as edge devices and mobile platforms. In this paper, we conduct a comprehensive survey of recent optimization techniques designed to enhance the energy efficiency of transformer architectures. We focus on analyzing the structural characteristics, design principles, and real-world application cases of these methods. Furthermore, we explore the feasibility and limitations of lightweight transformer models, offering insights into future directions for developing efficient and scalable AI systems. Ultimately, this survey aims to provide insights into the development of sustainable AI technologies by identifying key strategies for reducing energy consumption without compromising model performance.**

***Index Terms*—Transformer Optimization, Energy Efficiency**

I. Introduction

In recent years, artificial intelligence has seen rapid progress across various domains, including reinforcement learning [1]. Among these developments, deep learning has achieved remarkable breakthroughs, and among its various architectures, the Transformer model has emerged as a core technology [2]. Due to its strong performance in capturing long-range dependencies through self-attention mechanisms, the transformer has rapidly replaced traditional models such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), especially in natural language processing [3]. Its applications have since expanded into computer vision, speech recognition, and multimodal learning, further solidifying its role in modern artificial intelligence (AI) systems [4]. One of the most powerful features of transformers is their ability to leverage large-scale pretraining to achieve generalization and transfer learning across diverse tasks [5].

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2024-RS-2024-00436887) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation); and also by IITP grant funded by MSIT (RS-2024-00439803, SW Star Lab). *(Corresponding author: Joongheon Kim)*

However, the impressive performance gains come at the cost of significant computational overhead and high energy consumption [6]. These drawbacks present major challenges for deploying transformer models in real-time systems, edge devices, and energy-constrained environments. For instance, large-scale models such as GPT-3 consume significant amounts of energy for a single inference, posing challenges for deployment in energy-constrained or real-time environments, such as mobile and edge devices [7]. As a result, improving the energy efficiency of transformer models has become a key research agenda in both academia and industry. Recent studies have proposed a range of strategies, including model compression, low-precision quantization, structured pruning, and dynamic inference, to reduce computational complexity while preserving model accuracy. These techniques are especially valuable in mobile and Internet of Things (IoT) environments, where resources are limited and energy efficiency is crucial. This paper provides a comprehensive overview of recent transformer optimization techniques aimed at enhancing energy efficiency. We analyze structural characteristics and representative applications of various methods and explore their practical implications for the development of lightweight, scalable AI systems.

This paper is structured as follows. Section II provides an overview of the transformer architecture. Section III surveys recent optimization techniques aimed at improving energy efficiency. Section IV concludes the paper and outlines future research directions.

II. Transformer Structure

The Transformer architecture, illustrated in Fig. 1, is based entirely on attention mechanisms and eliminates the need for recurrence or convolution [8]. This design enables highly parallel computation and improves training efficiency by removing sequential dependencies.

The architecture consists of an encoder-decoder structure, where both components are composed of multiple stacked layers [9]. In the encoder, each layer includes a multi-head self-attention mechanism followed by a position-wise feed-forward neural network. The decoder has a similar structure but introduces two key additions: the masked self-attention layer, which prevents information leakage from future positions during training, and the cross-attention layer, which enables the decoder to attend to the encoder's output representations. At the core of the Transformer is the self-attention mechanism, which allows each token in a sequence to weigh the relevance

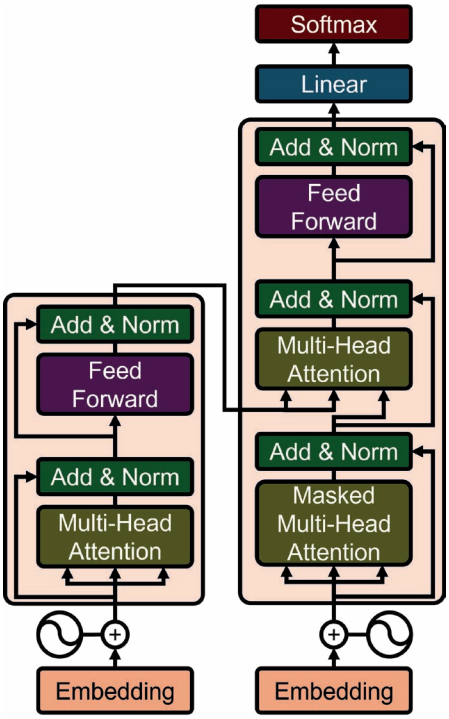

Fig. 1: Transformer structure.

of other tokens. Given input representations $X \in \mathbb{R}^{n \times d}$, the attention scores are computed as,

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{1}$$

In this formulation, Q, K, and V represent the query, key, and value matrices, respectively, obtained through learned linear projections of the input. The term d_k denotes the dimensionality of the key vectors, which is used to scale the dot product for numerical stability. To enhance representational capacity, the Transformer employs multi-head attention, where multiple attention heads operate in parallel to capture information from different subspaces of the input. The outputs of these heads are then concatenated and linearly transformed, as defined by,

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{2}$$

Here, h is the number of attention heads, each of which performs an independent self-attention operation, and W^O is the output projection matrix that integrates the results into a unified representation. Each layer in both the encoder and decoder includes a residual connection that directly adds the input of a sub-layer to its output, which helps preserve information and improves gradient flow during training. This is followed by a normalization layer that stabilizes the learning process. Additionally, since the attention mechanism does not inherently encode the order of input tokens, the model incorporates positional encodings to inject information about the relative or absolute position of each token in the sequence.

Overall, the Transformer is highly effective at modeling long-range dependencies and has demonstrated state-of-the-art performance across a wide range of sequence-based tasks, particularly in natural language processing and increasingly in vision and speech domains.

III. Energy-Efficient Optimization Techniques for Transformers

The high computational cost and memory demands of transformer models have posed considerable challenges for deployment in low-power or real-time environments, such as edge devices and battery-powered systems. To address these issues, a wide range of optimization strategies have been proposed to improve energy efficiency without compromising model performance.

One research direction focuses on the multi-head attention mechanism, a core component of the transformer architecture. Studies have shown that not all attention heads are equally important, and some contribute negligibly to overall performance [10]. By identifying and pruning less important heads, it is possible to significantly reduce computation while maintaining model accuracy. This approach not only improves efficiency but also enhances the interpretability of the model.

Hardware-aware optimization has also emerged as a prominent area of study. In this context, block-circulant matrix transformations have been applied to re-structure weight matrices for more efficient computation on hardware platforms such as field-programmable gate array (FPGA). By exploiting the circulant property, these transformations have achieved up to a $16\times$ reduction in parameter count, and experiments have reported approximately $8\times$ and $27\times$ energy efficiency improvements over graphics processing units (GPUs) and central processing units (CPUs), respectively [11].

Several studies have addressed optimization at the inference stage, particularly for time-series tasks. By combining structured pruning with quantization, where floating-point operations are replaced with low-precision integer arithmetic, researchers have reduced the size of the model and computational load. Quantization improves speed and reduces power usage, while pruning removes low-importance connections, simplifying the network architecture. These techniques have demonstrated up to 60% improvement in inference speed and approximately 30% reduction in energy consumption [12].

Other approaches focus on redesigning the self-attention mechanism itself. Spiking neural networks (SNNs), which use event-driven computation instead of continuous activations, have been applied to transformer models to drastically lower energy usage. By replacing multiplication-based operations with mask-based addition and introducing spike-event flows, some models have reported up to $87\times$ reductions in computational energy [13]. Similarly, sparse attention mechanisms reduce the quadratic complexity of traditional self-attention to linear or sublinear levels, improving both speed and energy efficiency for long input sequences [14].

Dynamic inference is another promising strategy. Instead of computing all layers or blocks uniformly, models can selectively activate parts of the network based on input complexity or importance. This adaptive execution reduces redundant

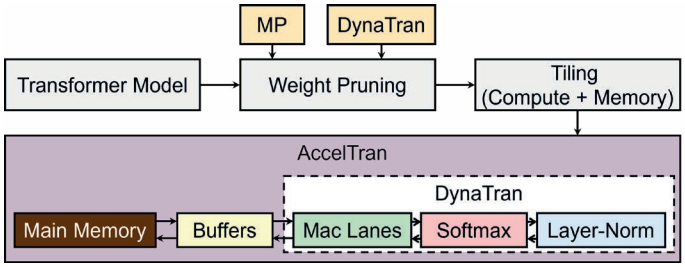

Fig. 2: Overview of the AccelTran pipeline for hardware acceleration of the transformer model.

computation and lowers power consumption without sacrificing accuracy. One study reports that sparsity-aware accelerators can be used to maximize data reuse and minimize memory bandwidth, further improving energy efficiency in edge computing environments [15]. Fig. 2 presents the overall workflow of AccelTran, a sparsity- and dynamic-inference-aware accelerator for transformer models. The diagram highlights how weight pruning, memory tiling, and selective execution are integrated to reduce computational cost and energy usage during inference.

Finally, some studies propose system-level co-optimization of algorithms and hardware. Techniques such as quantization-aware training (QAT), dynamic sparsity, and hardware-friendly pruning are designed with deployment constraints in mind, and have been integrated into custom application-specific integrated circuit (ASIC) or FPGA designs. These methods aim to minimize memory access, organize parallel execution, and optimize scheduling at the chip level [16].

Collectively, these energy-efficient optimization techniques demonstrate significant potential for enabling high-performance transformer models in resource-constrained environments such as mobile platforms, IoT devices, and embedded systems.

IV. Conclusion

This paper reviewed recent optimization techniques designed to improve the energy efficiency of transformer models from architectural, algorithmic, and hardware perspectives. As transformer-based models continue to grow in complexity and computational demand, energy-efficient solutions have become increasingly critical, especially for applications in mobile, edge, and IoT environments. We examined structural improvements that reduce the computational complexity of self-attention, including pruning and quantization techniques that decrease parameter counts without compromising accuracy. Event-driven spiking models and sparse attention mechanisms were also discussed as promising directions to minimize energy usage while maintaining performance. Furthermore, dynamic inference techniques allow selective execution paths based on input complexity, which can significantly lower power consumption. System-level co-optimization strategies that combine model design with hardware-aware deployment were also highlighted. These techniques collectively contribute to enabling practical, sustainable deployment of transformer models in energy-constrained scenarios. Moving forward, future research is expected to explore hybrid approaches that combine multiple optimization techniques and develop adaptive mechanisms that adjust computation dynamically according to energy budgets. Enhancing energy efficiency will remain a central challenge in the pursuit of scalable, high-performance AI systems.

V. Acknowledgement

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2024-RS-2024-00436887) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation); and also by the National Research Foundation of Korea (NRF) grant funded by MSIT (RS-2025-00561377). *(Corresponding author: Joongheon Kim)*

References

[1] J. Chung, C. Im, J. Choi, Y. Yoon, and S. Park, "DDPG-based deep reinforcement learning tactics for defending torpedo attacks," *IEEE Transactions on Intelligent Vehicles*, pp. 1–10, 2024 (Early Access).

[2] L. Papa, P. Russo, I. Amerini, and L. Zhou, "A survey on efficient vision transformers: Algorithms, techniques, and performance benchmarking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 7682–7700, December 2024.

[3] W. Zhou, S.-I. Kamata, H. Wang, and X. Xue, "Multiscanning-based RNN–transformer for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–19, May 2023.

[4] S. Islam, H. Elmekki, A. Elsebai, J. Bentahar, N. Drawel, G. Rjoub, and W. Pedrycz, "A comprehensive survey on applications of transformers for deep learning tasks," *Expert Systems with Applications*, vol. 241, pp. 122 666–122 213, May 2024.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, June 2019, pp. 4171–4186.

[6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, Vancouver, Canada, December 2020, pp. 1877–1901.

[7] I. Yoon, J. Mun, and K.-S. Min, "Comparative study on energy consumption of neural networks by scaling of weight-memory energy versus computing energy for implementing low-power edge intelligence," *Electronics*, vol. 14, no. 13, pp. 2718–2736, July 2025.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, Long Beach, CA, USA, December 2017, pp. 5998–6008.

[9] Y. Oh, G. E. Bae, K.-H. Kim, M.-K. Yeo, and J. C. Ye, "Multi-scale hybrid vision transformer for learning gastric histology: Ai-based decision support system for gastric cancer treatment," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 8, pp. 4143–4153, August 2023.

[10] P. Michel, O. Levy, and G. Neubig, "Are sixteen heads really better than one?" in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, Vancouver, BC, Canada, December 2019, pp. 14 014–14 024.

[11] B. Li, S. Pandey, H. Fang, Y. Lyv, J. Li, J. Chen, M. Xie, L. Wan, H. Liu, and C. Ding, "FTRANS: Energy-efficient acceleration of transformers using FPGA," in *Proc. ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED)*, Boston, MA, USA, August 2020, pp. 175–180.

[12] A. Kermani, E. Zeraatkar, and H. Irani, "Energy-efficient transformer inference: Optimization strategies for time series classification," *CoRR*, vol. abs/2502.16627, February 2025.

[13] C. Du, Q. Wen, Z. Wei, and H. Zhang, "Energy efficient spike transformer accelerator at the edge," *Intelligent Marine Technology and Systems*, vol. 2, no. 24, pp. 1–10, September 2024.

[14] M. Yao, J. Hu, Z. Zhou, L. Yuan, Y. Tian, B. Xu, and G. Li, "Spike-driven transformer," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, New Orleans, LA, USA, December 2023, pp. 64 043–64 058.

[15] S. Tuli and N. K. Jha, "AccelTran: A sparsity-aware accelerator for dynamic inference with transformers," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 42, no. 11, pp. 4038–4051, November 2023.

[16] S. Huang, E. Tang, S. Li, X. Ping, and R. Chen, "Hardware-friendly compression and hardware acceleration for transformer: A survey," *Electronic Research Archive*, vol. 30, no. 10, pp. 3755–3785, August 2022.