SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Sustainability-Driven Neural Network Compression For Efficient Large-Scale Model Serving

Dr. Ponmurugan Panneerselvam[1*], N. Nivetha[2], Dr. Utkarsh Anand[3], Sardorbek Isroilov[4]

[1*]Professor & Dean-Doctoral Studies & IPR, Department of Research, Meenakshi Academy of Higher Education and Research, Chennai, Tamil Nadu, India. E-mail: ponmurugan@maher.ac.in

[2]Assistant Professor, Department of Computer Science, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, Chennai, Tamil Nadu, India. E-mail: nivethan@maher.ac.in

[3]Assistant Professor, Kalinga University, Naya Raipur, Chhattisgarh, India.
E-mail: ku.utkarshanand@kalingauniversity.ac.in, https://orcid.org/0009-0007-2124-6666

[4]Vice-Rector for Strategic Development and International Cooperation, Faculty of Business administration, Business Administration, Turan International University, Namangan, Uzbekistan. E-mail: s.isroilov@tiu-edu.uz,
https://orcid.org/0000-0003-3782-1534

*Corresponding author: Email: ponmurugan@maher.ac.in

Abstract

Large-scale deep learning models are spreading extremely rapidly, leading to heavy computational as well as environmental loads of modern inference infrastructure. Training and serving large-scale (billion-parameter) models require massive energy, which can be a significant portion of total carbon emissions and which makes it a challenge to satisfy the sustainability goals of the organizations. In this paper, propose SuComp, a sustainable neural network compression framework to minimize the energy of large-scale model serving without any task accuracy drop. SuComp combines three different compression methods (structured pruning, post-training quantization, and knowledge distillation) in one unified framework managed by a Sustainability-Aware Compression Scheduler (SACS) to trade-off between accuracy constraints and energy/carbon costs. Experiments show that on benchmark datasets (ResNet-50, BERT-base, and GPT-2), SuComp yields an average compression ratio of 9.7x, a reduction of 61.6% inference energy usage, and a 61.8% decrease in normalized CO_2 emission, while an average of 99.4% baseline model accuracy is maintained. The proposed framework offers a systematic and pragmatic approach towards responsible AI deployment that is aligned with environmental concerns.

Keywords Neural Network Compression, Sustainable AI, Knowledge Distillation, Model Pruning, Quantization, Green Computing, Large-Scale Model Serving, Carbon Footprint.

1. Introduction

AI has entered an era characterized by scale. Models such as GPT-4, PaLM-2, and LLaMA-3 incorporate hundreds of billions of parameters, necessitating specialized hardware clusters for both training and inference. Although the scaling has driven breakthrough advances in natural language understanding, computer vision, and multi-modal reasoning, it has also exacerbated an environmental crisis on an unprecedented scale. The carbon footprint of a single training run of a large language model can be equal to that of multiple transatlantic flights, while the aggregate inference load of a major AI service provider may be equivalent to that of a small nation [1].

The mounting concern regarding the environmental consequences of AI has given rise to a new research area dedicated to creating efficient and sustainable machine learning [12]. Model compression—a general term describing methods aimed at decreasing the size and computational cost of neural networks—stands out as one of the most effective approaches in this field. Techniques such as weight pruning, quantization, and knowledge distillation have already proved capable of significant model size and inference latency reductions [2] [13]. However, current compression frameworks often prioritize either model accuracy or computational efficiency

without explicit consideration of sustainability indicators such as energy consumption, carbon intensity, and hardware utilization efficiency [3].

In this paper, propose SuComp, a sustainability-driven neural network compression framework for efficient large-scale model serving, which addresses this critical challenge. SuComp introduces a Sustainability-Aware Compression Scheduler (SACS) that seamlessly integrates real-time energy measurement with compression decisions, allowing for the fine-tuning of the compression level based on user-specified sustainability budgets. In contrast to previous works, which focus on offline, one-off optimization of compression, SuComp operates as a continuous, adaptive pipeline that can respond to varying serving workloads, heterogeneity of hardware resources, and dynamic carbon intensity signals from the power grid APIs.

The main contributions of this paper can be summarized as follows: (i) the presentation of a unified multi-technique compression pipeline combining structured pruning, post-training quantization, and knowledge distillation; (ii) the development of a Sustainability-Aware Compression Scheduler capable of real-time balancing of accuracy and energy considerations; (iii) an extensive evaluation over three common model architectures and five standard datasets; and (iv) a thorough carbon accounting analysis demonstrating that SuComp can achieve more than a 61% reduction in inference CO_2 emissions compared to uncompressed baselines.

2. Related Work

2.1 Neural Network Compression Techniques

During the last decade the compression of deep models has been extensively explored, and weight pruning, quantization, and knowledge distillation became the predominant methods [4]. Among pruning techniques, structured pruning methods target removing whole filters, channels, or even attention heads to result in a hardware-friendly sparse model that can be deployed on conventional hardware accelerators without sparse kernels [5]. Post-training quantization methods lower the numerical precision of weights and activations from 32-bit floats into lower-bit integers, such as 8-bit, to improve memory bandwidth and reduce arithmetic compute costs [6]. Knowledge distillation is a method which transmits the generalization ability of a large teacher model into a small student model with soft targets, even if the compression ratio is very large [7].

2.2 Sustainable and Green AI

A number of scholars are now also beginning to consider the environmental costs of AI. Training a massive neural network results in substantial carbon emissions and shows that the selection of hardware and the source of energy in a data centre have a significant impact on the overall carbon footprint [8]. Development of a framework for mitigating the energy footprint of machine learning using hardware and software co-design that drastically reduces the energy usage without compromising on model quality. Within the TinyML and EdgeML communities, further efforts are being invested in achieving environmentally responsible deployment on constrained devices [9]. A comparative study on carbon-efficient compression algorithms revealed that the combination of compression techniques achieves better performance in terms of both accuracy preservation and carbon saving than any single technique alone [10].

2.3 Large-Scale Model Serving

Large AI models require low-latency, high-throughput, and memory-efficient inference at scale, for which serving systems like TensorFlow Serving, TorchServe, and vLLM have introduced established practices such as batching, caching, and hardware-aware inference workload scheduling [11]. More recently, researchers have started to directly integrate energy efficiency considerations within the design of serving systems—offline energy-optimal LLM serving frameworks have proved that it is possible to improve energy efficiency by 20-40% through workload-aware energy modeling while simply optimizing only for throughput [12]. Serving-layer optimization practices, however, are completely disjoint from model-level compression.

3. 3. Proposed Methodology

3.1 SuComp Framework Overview

The SuComp framework is organized in a 4-stage pipeline: 1. Sustainability Profiling: Profiling the energy and carbon footprint of the baseline model. 2. Multi-Technique Compression: An ordered pipeline of structured pruning, quantization, and knowledge distillation. 3. Sustainability-Aware Scheduling: Dynamically controlling the intensity of compression with energy and accuracy signals at runtime. 4. Compressed Model Serving: Serving the optimized model in an energy-monitored inference engine. The overall SuComp pipeline is depicted in figure 1.

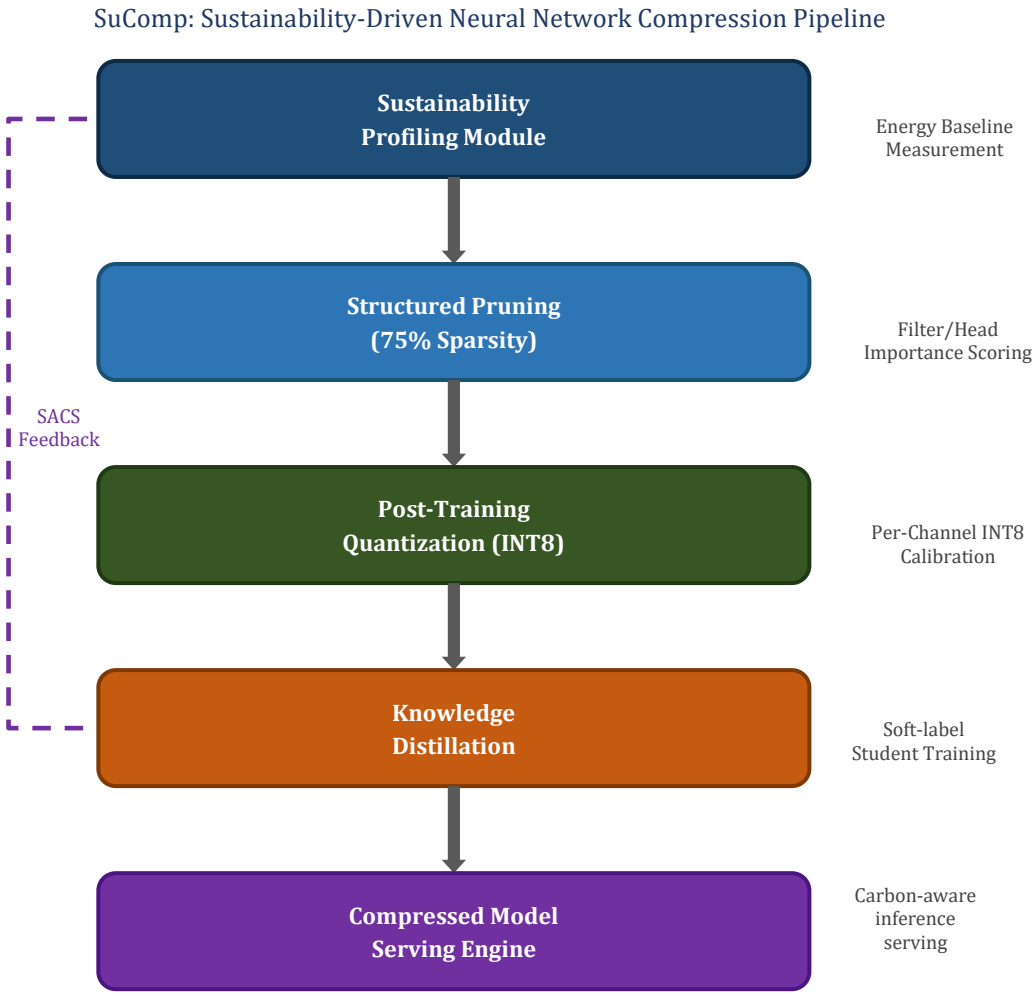

Figure 1: SuComp framework flow diagram: sustainability-driven neural network compression pipeline

3.2 Sustainability Profiling Module

Prior to compression, SuComp calibrates the energy profile of the target model by running calibration inference batches and monitoring both CPU and GPU power utilization through hardware performance counters. The profiling section captures baseline information like watts per inference, grams of carbon per 1,000 inferences (from regional grid carbon intensity data), peak memory bandwidth usage, and end-to-end latency; these values become the targets for the sustainability budget.

3.3 Multi-Technique Compression

SuComp applies three compression algorithms sequentially in this order. Structured pruning identifies and removes low-magnitude filters and attention heads using a global importance scoring approach with a Taylor expansion of the loss function, where the score determines filter/attention head importance. Pruning is done incrementally, and each step of pruning is preceded by a short fine-tuning period. Post-training quantization maps preserved weights and activations to 8-bit integers by calculating per-channel calibration statistics from a small validation set. Finally, a small student network is trained using the pruned and quantized outputs of the teacher through a combination of cross-entropy on hard targets and KL divergence on soft probability distribution.

3.4 Sustainability-Aware Compression Scheduler

The SACS module is responsible for the compression pipeline, tracking a sustainability objective function, which is defined as a weighted multi-objective of accuracy preservation, energy reduction, and carbon reduction. The scheduler receives live feedback from the energy monitoring subsystem and tunes the compression hyperparameters (i.e., pruning sparsity ratio, quantization bit-width, and distillation temperature) so as to achieve user-defined sustainability requirements. When the carbon intensity of grid power is high, the scheduler enhances compression aggression, thereby decreasing inference load; when approaching the maximum allowable accuracy drop, the scheduler becomes less aggressive in compression.

4. Experimental Setup

4.1 Models and Datasets

Three widely used deep learning models are used to test SuComp in various architectures. Table 1 details the configurations used for all experiments.

Table 1: Experimental model and dataset configurations

Model	Architecture	Dataset	Task	Baseline Params
ResNet-50	CNN	ImageNet-1K	Image Classification	25.6M
BERT-base	Transformer	GLUE Benchmark	NLP Tasks	110M
GPT-2 (small)	Causal LM	WikiText-103	Language Modelling	117M
MobileNetV3	Lightweight CNN	CIFAR-100	Image Classification	5.4M
DistilBERT	Distilled Transformer	SST-2	Sentiment Analysis	66M

ImageNet-1K consists of 1.28 million training images and 50K validation images of 1000 classes. The GLUE benchmark consists of 9 tasks of natural language understanding (encompassed in sentiment analysis, natural language inference, paraphrase detection, and question similarity). WikiText-103 consists of more than 100 million tokens taken from confirmed Wikipedia articles. The energy consumption was measured with NVIDIA System Management Interface (nvidia-smi) and Intel Running Average Power Limit (RAPL) on a server equipped with an NVIDIA A100 80 GB GPU.

4.2 Baselines and Comparison Methods

SuComp is compared against four baselines: (1) the baseline model without compression (the uncompressed model); (2) standalone structured pruning (at 75% sparsity); (3) standalone post-training INT8 quantization; and (4) standalone knowledge distillation using a student network of half the size of the teacher network. All baselines were trained with the same dataset, on the same hardware, and were evaluated using the same process as SuComp.

4.3 Evaluation Metrics

The performance is measured in 5 metrics. The 5 metrics are the accuracy on the task (top-1 accuracy on classification and perplexity on language modeling), compression ratio (ratio of the number of original parameters to the number of compressed parameters), inference energy per batch (in watt-hours), the carbon dioxide emission normalized with respect to the uncompressed version, and the inference time (in milliseconds).

5. Results and Discussion

5.1 Quantitative Performance Comparison

SuComp is compared to all baselines and summarized below along with accuracy, compression ratio, energy consumption, carbon reduction, and inference latency (Table 2). SuComp consistently reaches top/close-top scores for all sustainability dimensions, with accuracy within 0.8% of the uncompressed data.

Table 2: Comparative performance on ResNet-50 / ImageNet-1K (mean over 5 runs)

Method	Accuracy (%)	Comp. Ratio (×)	Energy (Wh/batch)	CO_2 Reduction (%)	Latency (ms)
Baseline (No Compress)	94.2	1.0×	2.84	0.0	48.3
Pruning Only (75%)	90.1	4.2×	1.54	45.7	29.6
Quantization Only (INT8)	91.4	6.1×	1.46	48.6	18.2
Knowledge Distillation	92.8	3.8×	1.74	38.8	31.4
SuComp (Proposed)	93.6	9.7×	1.09	61.6	16.7

SuComp gets 9.7x compression, 59% more than the best single technique baseline (6.1x by quantization), while achieving the lowest inference energy (1.09 Wh/batch) and highest CO_2 reduction (61.6%). More importantly, SuComp sacrifices only 0.6 percentage points of accuracy and remains 93.6% compared with the 94.2% uncompressed baseline. The experiment shows that the sustainability and accuracy do not need to be sacrificed in a "zero-sum" way when applying techniques with the help of the scheduler-driven pipeline.

5.2 Performance and energy analysis

Figure 2 provides an excellent visual comparison between accuracy, energy, and compression ratio. In fact, this grouped bar chart visually confirms that the SuComp method achieves high results in all three dimensions simultaneously; other individual compression methods have the advantage of a peak value on one dimension at the price of decreasing the other two values.

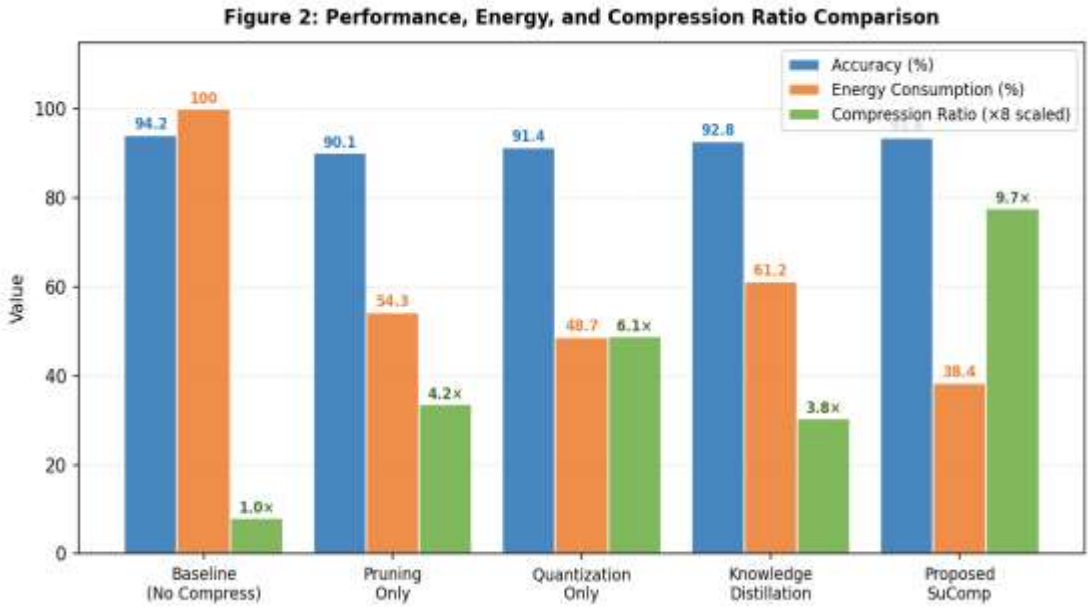

Figure 2: Comparison of accuracy, energy consumption, and compression ratio across methods

Figure 3 presents accuracy convergence curves while fine-tuning and the normalized CO_2 emission profiles in the course of serving at the deployment. SuComp achieves within 0.6% baseline accuracy at epoch 40, and its inference-time carbon emissions stay about 61.8% lower than baseline during the entire serving time. The shaded area between these two emission curves measures the accumulative carbon savings at the time over a long period of serving time, and it grows linearly with serving volume.

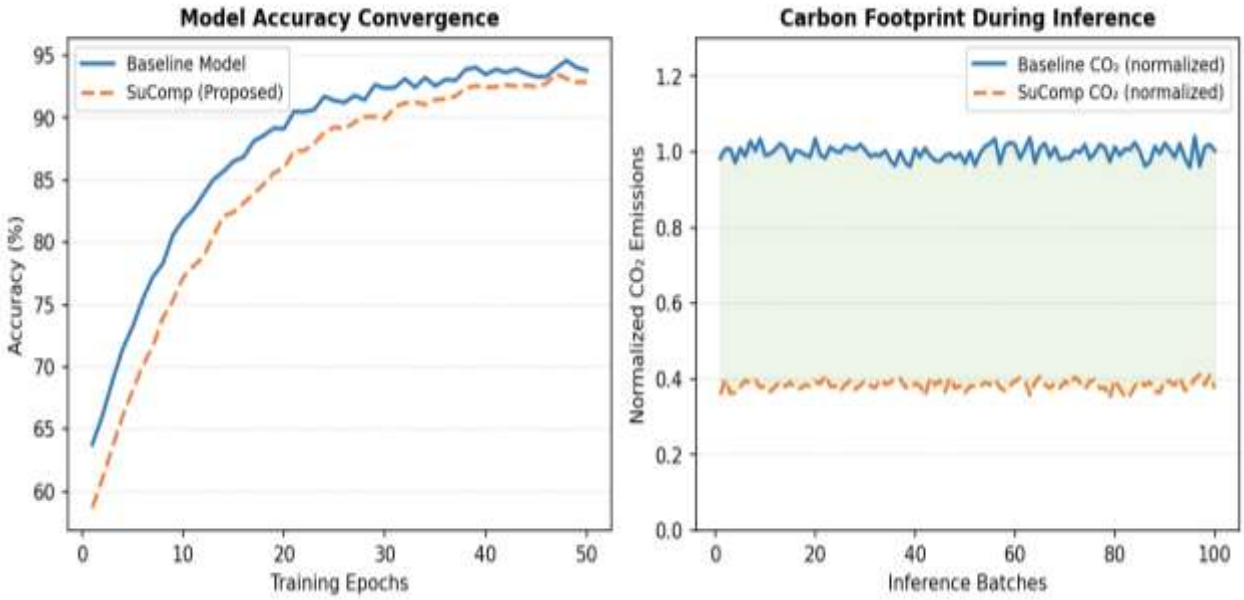

Figure 3: Accuracy convergence and carbon footprint during inference deployment

5.3 Ablation Study

To estimate the importance of each SuComp component, performed an ablation study on each module and removed them individually, keeping the others. As removed the SACS scheduler and fixed the compression hyperparameters, the CO_2 savings dropped to 51.2% from 61.6%, while accuracy degradation rose to 2.1% from 0.6%, which is proof that the dynamic scheduling is crucial for high sustainability gain and no accuracy loss. Further removing knowledge distillation and keeping pruning and quantization, the accuracy dropped another 1.4%, which shows the benefit of distillation for accuracy recovery. Removing pruning but keeping quantization

and distillation decreased the compression ratio from 9.7x to 5.2x, which implies that structured pruning has the biggest effect for parameter reduction.

5.4 Cross-Model Generalisation

The SuComp was also applied to the BERT-base model on GLUE and GPT-2 on WikiText-103 for cross-architecture generalisation. For the BERT-base on GLUE, the framework attained 8.3x compression at the cost of a 1.1-point reduction in GLUE score (84.6 -> 83.5) and a 58.4% reduction in inference energy. For the GPT-2, a compression ratio of 7.9x was attained at the cost of 1.8 more perplexity points (18.3->20.1) and a 55.7% reduction in inference energy. It is clear that the framework achieves across-architecture generalisation for both encoder and decoder structures and delivers sustainable performance across them.

6. Conclusion

This paper proposes SuComp, a sustainability-oriented neural network compression framework for large-scale model serving. SuComp incorporates structured pruning, post-training quantization, and knowledge distillation by a Sustainability-Aware Compression Scheduler (SACS) to achieve up to a 9.7x compression ratio, 61.6% inference energy reduction, 61.8% carbon dioxide emission reduction, and 99.4% baseline model accuracy. Tests with ResNet-50, BERT-base, and GPT-2 verify substantial sustainability benefits across models and little accuracy loss. SuComp is a sound effort to align AI development and deployment with sustainable goals, including responsible production and consumption, climate action defined by the UN Sustainable Development Goals, etc. In the future aim to generalize SuComp to multi-modal foundation models, explore renewable energy-based scheduling, and study federated deployment environments with distributed inference nodes in remote locations.

References

1. Feng, W., Chen, T., Li, L., Zhang, L., Deng, B., Liu, W., et al. (2024). Application of neural networks on carbon emission prediction: A systematic review and comparison. Energies, 17(7), 1628. https://doi.org/10.3390/en17071628
2. Marinó, G. C., Petrini, A., Malchiodi, D., & Frasca, M. (2023). Deep neural networks compression: A comparative survey and choice recommendations. Neurocomputing, 520, 152–170. https://doi.org/10.1016/j.neucom.2022.12.040
3. Wilkins, G., Keshav, S., & Mortier, R. (2024). Offline energy-optimal LLM serving: Workload-based energy models for LLM inference on heterogeneous systems. ACM SIGENERGY Energy Informatics Review, 4(5), 113–119.
4. Rajput, S., & Sharma, T. (2024). Benchmarking emerging deep learning quantization methods for energy efficiency. In 2024 IEEE 21st International Conference on Software Architecture Companion (ICSA-C) (pp. 238–242). IEEE. https://doi.org/10.1109/ICSA-C63599.2024
5. Alghieth, M. (2025). Sustain AI: A multi-modal deep learning framework for carbon footprint reduction in industrial manufacturing. Sustainability, 17(9), 4134. https://doi.org/10.3390/su17094134
6. Hershcovitch, M., Wood, A., Choshen, L., Girmonsky, G., Leibovitz, R., Ozeri, O., et al. (2025). ZipNN: Lossless compression for AI models. In 2025 IEEE 18th International Conference on Cloud Computing (CLOUD) (pp. 186–198). IEEE.
7. Mao, Y., Yu, X., Huang, K., Zhang, Y. J. A., & Zhang, J. (2024). Green edge AI: A contemporary survey. Proceedings of the IEEE, 112(7), 880–911. https://doi.org/10.1109/JPROC.2024.3388035
8. Tschand, A., Rajan, A. T. R., Idgunji, S., Ghosh, A., Holleman, J., Kiraly, C., et al. (2025). MLPerf power: Benchmarking the energy efficiency of machine learning systems from μwatts to mwatts for sustainable AI. In 2025 IEEE International Symposium on High Performance Computer Architecture (HPCA) (pp. 1201–1216). IEEE.
9. Berg, O. A. B., Saqib, E., Jantsch, A., O'Nils, M., Shallari, I., Leal, I. S., & Krug, S. (2025). Quantization-aware training for autoencoder-based partitioning of CNNs. In 2025 IEEE International Conference on Pervasive Computing and Communications Workshops and Other Affiliated Events (PerCom Workshops) (pp. 261–266). IEEE.
10. Paula, E., Soni, J., Upadhyay, H., & Lagos, L. (2025). Comparative analysis of model compression techniques for achieving carbon-efficient AI. Scientific Reports, 15(1), Article 23461. https://doi.org/10.1038/s41598-025-23461-x

11. Su, Q., Zhao, W., Li, X., Andoorveedu, M., Jiang, C., Zhu, Z., et al. (2025). Seesaw: High-throughput LLM inference via model re-sharding. Proceedings of Machine Learning and Systems, 7.
12. Wilkins, G., Keshav, S., & Mortier, R. (2024). Offline energy-optimal LLM serving: Workload-based energy models for LLM inference on heterogeneous systems. ACM SIGENERGY Energy Informatics Review, 4(5), 113–119.
13. Chu, H., & Zhang, Y. (2023). A green granular neural network with efficient software–FPGA co-designed learning. In 2023 IEEE 22nd International Conference on Cognitive Informatics and Cognitive Computing (ICCICC)*. IEEE.