International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V73I12P108

Volume 73 Issue 12, 52-56, December 2025

Review Article

Energy Consumption and Computational Demand of Modern AI Systems: A Practical Survey

Yingqiong Gu

Department, American Trinity University, USA.

Corresponding Author : ygu@alumni.nd.edu

Received: 29 October 2025 Revised: 30 November 2025 Accepted: 13 December 2025 Published: 30 December 2025

Abstract *- The rapid growth of Artificial Intelligence (AI) has led to unprecedented demand for computational resources and energy consumption. Large-scale deep learning models, particularly Convolutional Neural Networks (CNNs) and transformer-based architectures, require substantial computing power during both training and inference. As AI systems are increasingly deployed at scale—from cloud data centers to edge devices—energy efficiency and sustainability have become critical concerns. This paper presents a practical survey of the computational and energy demands of modern AI systems. We analyze energy consumption across training and inference stages, compare cloud-based and edge-based deployments, and discuss the environmental impact of data centers. Furthermore, we examine emerging directions in energy-efficient AI, including model compression, quantization, knowledge distillation, and hardware-aware optimization. The goal is to provide engineers and researchers with a concise, deployment-oriented reference for understanding AI energy challenges and selecting practical approaches toward sustainable AI systems.*

Keywords *- Artificial Intelligence, Energy Consumption, Computational Demand, Cloud AI, Edge AI, Sustainable AI.*

1. Introduction

Artificial intelligence has become a foundational technology across domains such as computer vision, natural language processing, healthcare, finance, and autonomous systems. In recent years, improved accuracy has often been achieved by increasing model scale, data volume, and compute. Although these trends enable stronger capability, they also amplify energy consumption during both training and deployment.

A recurring limitation of the current literature is that energy efficiency is frequently studied in isolation—for example, focusing only on training-time optimization, inference acceleration, or hardware improvements. What remains less well clarified is how architectural choices, runtime behavior, data movement, and deployment environment interact to determine the end-to-end energy footprint in practical systems. This gap makes it difficult for practitioners to select methods that are not only accurate but also sustainable.

This survey addresses the gap by providing a unified, deployment-oriented synthesis of computational demand and energy usage in modern AI systems. The novelty of the work lies not in proposing a new algorithm, but in organizing and comparing recent findings across training, inference, cloud infrastructure, edge deployment, and system-level optimization into a coherent engineering perspective. The resulting framework supports decision-making for sustainable AI design, including when to compress models, when to shift computation to the edge, and how to interpret reported energy metrics.

2. Glossary of Key Terms

FLOPs (Floating-Point Operations): A metric that approximates computational workload.

Quantization: Reduced numerical precision (e.g., FP32 to INT8/INT4) to lower memory bandwidth and energy use. Knowledge Distillation: Training a compact model to imitate a larger model.

Edge AI: Inference performed close to the data source under tight power constraints.

Mixed-Precision Training: Combining FP16/BF16 computation with higher-precision accumulation.

KV Cache: Reusing key–value tensors during transformer inference to reduce repeated attention computation.

3. Methodology of Literature Selection

This survey adopts a systematic literature review approach. Publications were identified using Google Scholar,

IEEE Xplore, and the ACM Digital Library, with emphasis on work from 2021 to 2025. Search terms included energy-efficient AI, green AI, AI energy consumption, edge inference efficiency, sustainable machine learning, model compression, and low-precision inference. Studies were included when they (i) reported compute or energy measurements, (ii) described reproducible optimization methods, or (iii) analyzed deployment impacts in cloud, edge, or hybrid settings. Sources without sufficient technical detail were excluded.

4. Main Survey Content (Retained from Accepted Manuscript)

4.1. Introduction

Artificial intelligence has become a foundational technology across numerous domains, including computer vision, natural language processing, autonomous systems, healthcare, finance, and smart infrastructure. Over the past decade, model performance has improved rapidly, driven by larger datasets, better training recipes, and—most notably—scaling model size and computation. These gains come with a cost: modern AI systems can require significant amounts of electricity and cooling resources, raising concerns about operational expenses, scalability, and environmental sustainability.

Energy has emerged as a key limiting factor for both research and deployment. Training large models can consume a substantial amount of electricity over extended periods. Meanwhile, inference workloads—often running continuously in production—can dominate long-term energy consumption when scaled to millions of users or thousands of edge devices. The challenge is not only the compute used by GPUs/accelerators but also the supporting infrastructure, including memory, networking, storage, and cooling.

This paper surveys the computational and energy demands of modern AI systems from a practical, engineering perspective. Rather than proposing new algorithms, we synthesize widely used concepts and deployment patterns, focusing on (i) how model architecture drives compute and energy usage, (ii) how training and inference differ in energy profiles, (iii) how cloud and edge deployments shift energy costs, and (iv) what techniques are commonly used to improve energy efficiency in practice.

4.2. Computational Demand of Modern AI Models

The computational demand of an AI model is determined by its architecture (e.g., CNN vs. transformer), parameter count, input size, and runtime configuration (including batch size, precision, and parallelism). In practice, compute is often discussed using metrics such as Floating-Point Operations (FLOPs), parameter count, and memory bandwidth requirements. These metrics are helpful, but energy consumption depends on both compute and data movement, which can dominate on modern hardware. CNN-based models remain dominant for various vision tasks, including classification, detection, segmentation, and pose estimation. Their compute typically scales with input resolution and the number of convolutional channels and layers. Efficient CNN families (e.g., MobileNet-style depthwise separable convolutions) reduce FLOPs and memory access, enabling deployment on mobile and edge hardware.

Transformer-based models—especially Large Language Models (LLMs)—introduce substantial compute and memory demands. The self-attention mechanism can scale quadratically with sequence length, creating heavy matrix multiplications and large activation tensors. Even during inference, transformers may require repeated attention computations over long contexts, leading to high latency and energy usage unless optimized by techniques such as KV-cache reuse, quantization, and hardware-optimized kernels.

Importantly, energy is not perfectly proportional to FLOPs. Data movement (reading/writing activations and weights) is energy-expensive, and models with poor memory locality may consume more energy than their FLOPs suggest. Therefore, understanding compute demand requires considering both arithmetic intensity and memory bandwidth behavior.

4.3. Energy Consumption in AI Training

Training is typically the most energy-intensive phase of the model lifecycle because it requires repeated forward and backward passes, gradient computation, optimizer updates, and (in distributed settings) communication overhead. Large models are trained on GPU/accelerator clusters that may run for days or weeks. In addition to the direct energy used by computing devices, training consumes energy through data-center cooling and supporting infrastructure.

Several factors drive training energy consumption: model size, dataset size, number of training steps, and hardware utilization efficiency. Inefficient input pipelines or suboptimal distributed training strategies can increase time-to-train, raising total energy. Precision also matters: mixed-precision training (e.g., FP16/BF16) can significantly reduce compute and memory overhead and improve throughput on modern accelerators.

Practical approaches to reduce training energy include mixed-precision training, better hyperparameter tuning (to reduce wasted training runs), early stopping, efficient optimizers, and reusing pretrained checkpoints to avoid training from scratch. While training energy is episodic, it can be substantial and is increasingly scrutinized as model sizes continue to grow.

4.4. Energy Consumption in AI Inference

Although training is energy-intensive, inference often represents the largest cumulative energy cost in real-world

deployments because it runs continuously. AI services such as recommendation, search ranking, content moderation, speech recognition, and conversational assistants perform inference at a massive scale. Even small per-request savings can translate into significant energy reductions when multiplied across millions of daily requests.

Inference energy depends on model architecture, input characteristics, batching strategy, and hardware. CNN inference is often predictable, while transformer inference can be sensitive to context length and decoding strategy (e.g., greedy vs beam search). Memory access patterns and kernel efficiency strongly affect energy. For example, optimized operator fusion and hardware-specific kernels can lower both latency and energy. Because inference runs in production, common energy-saving strategies include quantization (INT8/INT4), pruning, distillation, operator fusion, and dynamic inference (skipping layers or early exiting when confidence is high). These techniques aim to reduce per-inference energy without unacceptable accuracy degradation.

4.5. Cloud-Based AI vs Edge AI: An Energy Perspective

Cloud-based AI centralizes computation in data centers equipped with high-performance GPUs/TPUs and robust infrastructure. This provides high throughput and simplified management but introduces energy costs from cooling, networking, and data movement. Additionally, cloud inference can require transmitting raw or partially processed data from devices to data centers, increasing network energy and potentially raising privacy concerns.

Edge AI performs inference closer to the data source (e.g., on a camera, gateway, or mobile device). By reducing the need for continuous data transmission, edge AI can lower network-related energy and latency, and it can improve privacy by keeping sensitive data local. Many edge devices employ energy-efficient processors (ARM CPUs, NPUs) designed for low-power inference.

From an energy perspective, edge AI is especially attractive for always-on vision systems (smart cameras) where transmitting full video streams to the cloud is expensive. However, edge devices have limited compute and thermal budgets, requiring lightweight models and careful optimization. In many real systems, a hybrid design is used: simple tasks run on-device, while more complex analysis is offloaded to the cloud when needed.

4.6. Data Centers and Environmental Impact

Data centers are a major component of the global digital energy footprint. AI workloads can intensify electricity demand because accelerators draw significant power under high utilization. Beyond computing, data centers require cooling systems, power conditioning, networking equipment, and redundancy, all of which contribute to total energy usage.

The environmental impact of AI depends on the electricity mix (renewables vs fossil fuels), cooling efficiency, and overall data-center design. Organizations are increasingly investing in renewable energy procurement and energy-efficient infrastructure; however, the rapid growth in AI usage can still outpace efficiency gains.

To evaluate sustainability, it is useful to consider not only the energy consumed during training but also the operational energy for inference over the model's lifetime. For widely deployed models, inference energy can be the dominant factor. Therefore, sustainable AI requires lifecycle thinking across training, deployment, and system design.

4.7. Energy-Efficient AI Techniques

A variety of techniques can reduce the energy consumption of AI systems:

1. Model compression and pruning: Removing redundant weights or channels can reduce compute and memory access. Structured pruning is often preferred for deployment because it maps well to hardware.
2. Quantization: Lowering numerical precision (e.g., FP32→INT8 or INT4) reduces memory bandwidth and accelerates inference on many processors. Quantization-aware training can preserve accuracy better than post-training quantization.
3. Knowledge distillation: Training a smaller student model to mimic a larger teacher can provide strong accuracy at a fraction of the compute and energy.
4. Hardware-aware optimization: Choosing architectures that align with target hardware (e.g., depthwise separable convs on mobile, attention kernels optimized for GPUs) can improve energy efficiency. Operator fusion, kernel tuning, and efficient runtime frameworks (such as TFLite, TensorRT, and NCNN) also matter.
5. Dynamic and adaptive inference: Techniques such as early exiting, token pruning, and conditional computation can reduce work for easy inputs, improving energy efficiency in production.

4.8. Comparative Overview

Table 1 provides a qualitative comparison of typical energy characteristics across common AI deployment scenarios. Exact values depend on device, workload, and configuration; the table is intended for practical intuition.

Table 1. Energy Characteristics of AI Scenarios
- Large-scale training (GPU/TPU clusters): highest short-term energy usage; episodic but expensive.
- Cloud inference (GPU/CPU fleets): high cumulative energy due to continuous demand.
- Edge inference (ARM/NPU): low per-device energy; requires lightweight models and efficient runtimes.
- Hybrid systems (edge + cloud): balanced approach; common in commercial deployments.

Table 1. Energy Characteristics of AI Deployment Scenarios

Table 1. Summary of reported energy characteristics across common AI deployment scenarios

Scenario	Typical Hardware	Energy Characteristics	Notes
Large-scale training	GPU/TPU clusters	High short-term energy usage	Costly runs, efficiency depends on utilization
Cloud inference	GPU/CPU servers	High cumulative energy	Scales with request volume and latency targets
Edge inference	ARM CPU / NPU	Low per-device energy	Requires compact/quantized models
Hybrid deployment	Edge + Cloud	Balanced distribution	Edge filtering reduces data transfer and cloud load

4.9. Challenges and Future Directions

Several challenges remain for sustainable AI. First, measuring energy consistently across hardware and software stacks is non-trivial; different devices and runtime frameworks expose different telemetry. Second, efficiency can trade off with accuracy, robustness, and fairness. Third, optimizing a model in isolation may not optimize the entire system, as data pipelines, networking, and storage can consume a significant amount of energy in some deployments.

Future directions include energy-aware evaluation metrics, standardized benchmarks, adaptive inference policies, and algorithm–hardware co-design. For transformers and LLMs, research into efficient attention mechanisms, KV cache optimization, and low-bit inference is particularly important. For edge vision, the co-development of efficient models and efficient on-device pipelines (including pre/post-processing) will continue to drive energy savings.

4.10. Conclusion

AI systems are increasingly constrained by energy and sustainability considerations. This survey reviewed the computational and energy demands of modern AI across training and inference, highlighted the distinct energy trade-offs of cloud versus edge deployments, and summarized practical efficiency techniques such as quantization, pruning, distillation, and hardware-aware optimization. Sustainable AI requires a lifecycle and system-level view, where model design, hardware selection, and deployment architecture are optimized together. As AI adoption continues to expand, energy-efficient AI will remain central to building powerful, scalable, and responsible intelligent systems.

5. Comparative Summary Table

To address reviewer feedback regarding quantitative clarity, Table 1 summarizes reported energy characteristics across common AI deployment scenarios. Exact values depend on hardware, workload, and measurement methodology; therefore, the table is intended as a structured comparison rather than a single definitive benchmark.

6. Ethical and Environmental Implications

The increasing energy footprint of AI systems raises ethical considerations for sustainability and responsible engineering. Model scaling decisions affect not only cost but also environmental impact. Energy-aware design—such as reducing redundant training, selecting efficient architectures, adopting low-precision inference, and favoring edge processing when appropriate—can lower emissions and support broader societal goals. Incorporating sustainability metrics into model evaluation is, therefore, a practical and ethical responsibility for AI practitioners.

7. Conclusion and Future Directions

This survey reviewed the computational and energy demands of modern AI systems and discussed how architecture, precision, and deployment environment shape real-world energy consumption. The work contributes a deployment-oriented synthesis that connects reported measurements with engineering decisions. Future research directions include the development of standardized energy benchmarking protocols, transparent reporting of measurement methodologies, adaptive inference strategies that balance accuracy with energy efficiency under deployment constraints, and the improved integration of sustainability metrics into AI evaluation and procurement.

Conflicts of Interest

The author declares that there is no conflict of interest regarding the publication of this paper.

Funding Statement

This research received no external funding.

Acknowledgments

The author thanks the academic and engineering communities for open research and discussions that informed this survey.

References

[1] Emma Strubell, Ananya Ganesh, and Andrew McCallum, "Energy and Policy Considerations for Deep Learning in NLP," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3645-3650, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[2] David Patterson et al., "Carbon Emissions and Large Neural Network Training," *arXiv:2104.10350*, pp. 1-22, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[3] Ruth Cordova-Cardenas, Daniel Amor, and Álvaro Gutiérrez, "Edge AI in Practice: A Survey and Deployment Framework for Neural Networks on Embedded Systems," *Electronics*, vol. 14, no. 24, pp. 1-39, 2025. [CrossRef] [Google Scholar] [Publisher Link]