Survey on the Challenges and Solutions of Energy Efficient Computing

CIS 454 Fall 2024

Hardik Polamarasetti
Department of Computer Science
Cleveland State University
Cleveland, Ohio USA
h.polamarasetti@vikes.csuohio.edu

ABSTRACT

With the exponential growth of data and computing power demands, energy efficiency has become a critical concern in computing. Traditional computing models, suchas centralized cloud based models, struggle to meet the growing demands and strain resource constrained devices. This paper explores new energy-efficient strategies with a primary focus on edge computing, a technique where computation is performed closer to data sources to reduce latency, bandwidth, and energy consumption. It also covers a promising breakthrough in edge computing regarding a new model for streaming time series classification called Attentive Power Iteration (StrAPI). Through an analysis of current trends and innovations in edge computing, and by exploring advancements in AI algorithms, hardware design, and data compression techniques, the top approaches to reducing energy consumption are explored. Additionally, this paper discusses the challenges of implementing these solutions and emphasizes the importance of continued innovation in achieving a sustainable technological future.

KEYWORDS

Energy efficiency, edge computing, machine learning, IoT, sustainability, optimization

1 Challenges in Energy Efficient Computing

The rapid expansion of demanding, data driven technologies, such as artificial intelligence (AI), the Internet of Things (IoT), and cloud computing, have led to exponential increases in energy consumption. In 2020 alone, data centers accounted for approximately 1% of all global electricity usage, a number that continues to grow as industries demand more and more computational power and memory resources. Large scale AI systems further the issue, and some complex models consume the amount of energy that is equivalent to that used by dozens of households over the course of an entire year. These trends present environmental challenges, such as higher carbon emissions, and also financial concerns for organizations and individuals that are trying to minimize operational costs and maximize the efficiency of their system.

Energy efficient computing faces several critical obstacles. First, the increasingly monumental amount of energy required for modern computing infrastructures, particularly in data centers, contributes significantly to global carbon emissions. This is especially true for data centers, which are much more demanding of computational and memory resources. Businesses are constantly under increasing pressure to adopt greener practices but often find the costs of implementing such technologies a barrier to making these necessary changes. Additionally, resource constrained devices, such as IoT sensors and wearables, are devices that are limited in on board computational power and memory and are often required to send data to centralized servers where the computations can take place, and the results are sent back to be applied. These devices struggle to balance performance with energy efficiency. These challenges highlight the urgent need for innovative approaches to sustainable computing.

2 Edge Computing

Edge computing has newly emerged as a major solution to the energy challenges faced by traditional cloud computing. it works by processing data closer to its source such as locally on devices or nearby servers instead of transmitting data to far removed cloud servers. By doing this, edge computing reduces the energy required for data transmission and minimizes network congestion.

One real world use case for this technology is in healthcare, where edge devices can process patient data from wearable sensors in real time, and provide immediate alerts for critical conditions without the need to transmit the data to distant servers. Another example is in the automotive industry, where autonomous

vehicles can use edge computing to quickly analyze sensor data to make split second decisions on the road while also conserving energy by avoiding cloud based processing.

This approach is extremely beneficial because it minimizes latency but also optimizes bandwidth utilization. This makes it particularly valuable for applications requiring real time processing in industries like healthcare, automotive and manufacturing, where precision and speed are crucial.

2.1 StrAPI: Edge Computing Model

One of the most promising advancements in edge computing is the Attentive Power Iteration (StrAPI) model, introduced by researchers Hao Huang, Tapan Shah, Scott Evans, and Shinjae Yoo. This advancement was proposed in the research article, "Energy Efficient Streaming Time Series Classification with Attentive Power Iteration" for the 38th AAAI Conference on Artificial Intelligence. StrAPI was specifically designed to address the challenges of streaming time series classification, which is a task that is one of the most demanding components of edge computing from a resource standpoint.

Traditional methods of time series classification process entire datasets in bulk and require high computational power that may not be available on resource constrained devices. StrAPI uses an innovative and novel approach that incrementally updates its calculations as new data arrives.

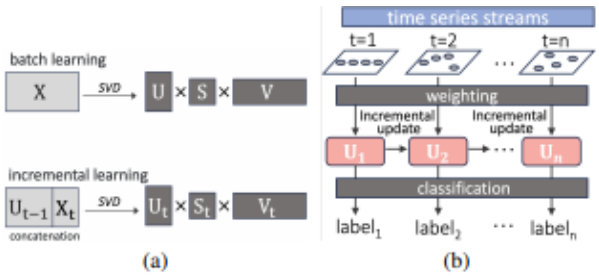

Figure 1: Comparison of data reduction with traditional batch learning and incremental learning.

This allows for quicker real time processing which conserves both energy and computational overhead. The model is particularly suited for scenarios where only partial access to a time series is available, such as for resource constrained devices. StrAPI's advantages are can be seen in experimental results where this new method demonstrated up to 70% energy savings and three times faster processing speeds compared to traditional edge computing methods. These findings highlight StrAPI's potential to revolutionize energy-efficient computing in various applications.

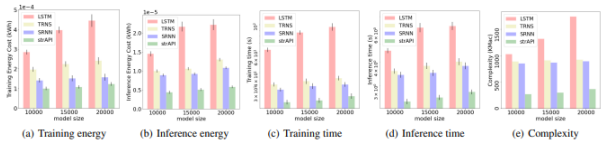

Figure 2: Comparison of running costs and complexity with different model sizes.

This approach aligns closely with the goals of energy efficient computing. By focusing on incremental updates rather than processing entire datasets, StrAPI reduces redundant calculations, therefore saving energy. Additionally, its ability to perform well on resource constrained devices means it is not dependent on the power demanding centralized infrastructures. This model not only lowers the environmental footprint of computing tasks but also makes advanced analytics accessible to industries and regions with limited energy resources. This makes it potentially pivotal in global energy conservation efforts.

3 Other Advancements

Edge computing shows promise for energy efficiency, however, other developments in AI algorithms, hardware optimization, and data compression also contribute to energy efficiency. Some of these developments are related to edge computing and can work in conjunction with edge computing to further the progress in energy efficient computing.

3.1 Algorithmic Optimization

Algorithmic innovations such as pruning, quantization, and spiking neural networks have revolutionized the way AI models process information. Pruning removes redundant parameters from neural networks, which reduces the computational load without significantly compromising accuracy. Quantization is a technique that simplifies computations by using integer arithmetic instead of floating points. This approach speeds up processing and also significantly reduces energy consumption as well. Spiking neural networks work by processing information through discrete spikes, requiring far less energy than traditional continuous data streams. These innovations make it possible to deploy advanced AI models on devices with limited power, such as IoT sensors, without sacrificing performance.

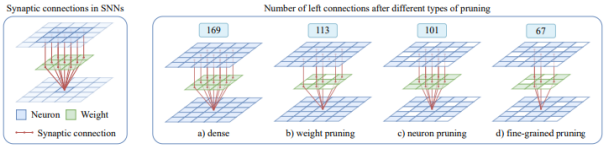

Figure 3: Diagram of synaptic connections in SNNs and number of connections left after pruning.

3.2 Hardware Innovations

Energy-efficient hardware designs are another critical area of innovation. Field Programmable Gate Arrays (FPGAs) and Graphics Processing Units (GPUs) are increasingly optimized for specific tasks, reducing energy waste. For example, Tesla's Full Self-Driving (FSD) chip uses this technology to handle complex autonomous driving computations while minimizing power usage. This chip works by using neural network accelerators to process large volumes of data from sensors around the car in real time which makes autonomous driving both energy efficient and safer. Additionally, specialized GPUs for AI tasks, such as NVIDIA's Tensor Cores, have shown significant improvements in both speed and energy consumption for deep learning models. These advancements show how hardware improvements for computational applications can result in greater energy efficiency while meeting the growing demands of modern technologies.

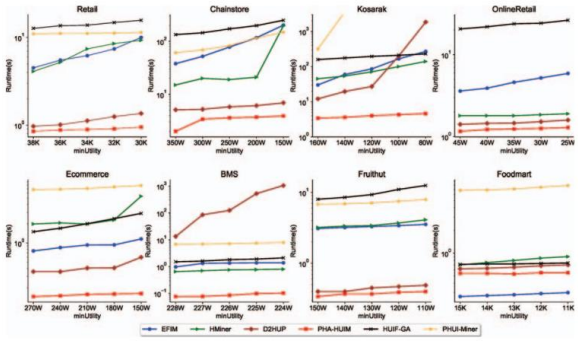

Figure 4: Comparison of GPU based algorithm runtime (PHA-HUIM) with other techniques on various datasets.

3.3 Data Handling

Efficient data management is essential for reducing energy consumption, especially in data intensive applications. Techniques such as DeepMapping use neural networks to create compact data while retaining essential features. This allows systems to store and retrieve information more quickly, reducing the energy required for storage and computation. For example, DeepMapping has been shown to optimize data lookup times and reduce the memory footprint of large datasets by as much as 70.8% in some cases, making it a great solution for edge devices that are resource constrained. By compressing data without losing critical information, these methods conserve energy and also allow for faster processing, which increases the overall efficiency of computing systems.

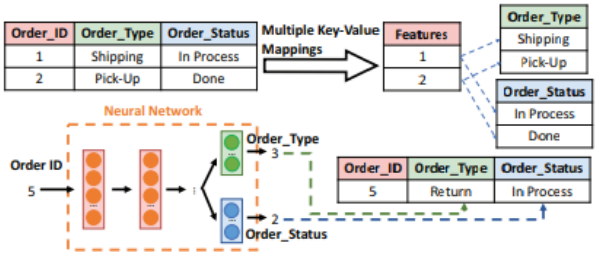

Figure 5: Diagram of DeepMapping use of Neural Networks to memorize key-value mapping in tabular data.

4 Conclusion

Energy-efficient computing requires a comprehensive approach that integrates advancements across various fields. Edge computing advancements such as the StrAPI model show a benefit to localized processing that reduces energy demands while maintaining high performance. By processing real time data efficiently, StrAPI demonstrates how developments and new strategies can significantly lower computational overhead.

To supplement the improvements made by edge computing, algorithmic optimization such as pruning, quantization, and spiking neural networks improve processing efficiency without sacrificing accuracy. Specialized hardware improvements such as GPUs and FPGAs further reduce energy consumption by optimizing computations to specific tasks. These hardware developments can be seen in applications such as autonomous driving. Additionally, data handling techniques such as DeepMapping optimize storage and retrieval, enabling faster and more energy efficient operations.

These innovations offer a comprehensive solution for energy efficiency in computing. However, widespread adoption will require overcoming challenges such as cost barriers, integration with legacy systems, and scalability. More efforts between the technology industry and research communities are crucial to continue to progress. Continuing to conduct research in these fields is important to meeting the growing demands of modern computational requirements while significantly reducing its environmental impact.

REFERENCES

[1] P. Huang et al., "Towards efficient verification of quantized neural networks," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 19, pp. 21152–21160, Mar. 2024. doi:10.1609/aaai.v38i19.30108

[2] Y. Wang et al., "Towards ultra-high performance and energy efficiency of deep learning systems: An algorithm-hardware co-optimization framework," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, Apr. 2018. doi:10.1609/aaai.v32i1.11653

[3] H. Huang, T. Shah, S. Evans, and S. Yoo, "Energy efficient streaming time series classification with Attentive Power Iteration," Proceedings of the AAAI Conference

on Artificial Intelligence, vol. 38, no. 11, pp. 12574–12582, Mar. 2024. doi:10.1609/aaai.v38i11.29151
[4] L. Zhou, K. S. Candan, and J. Zou, “Deepmapping: Learned data mapping for lossless compression and efficient lookup,” 2024 IEEE 40th International Conference on Data Engineering (ICDE), pp. 1–14, May 2024. doi:10.1109/icde60146.2024.00008
[5] K. Zhang et al., “Duet: Efficient and scalable hybrid neural relation understanding,” 2024 IEEE 40th International Conference on Data Engineering (ICDE), pp. 56–69, May 2024. doi:10.1109/icde60146.2024.00012
[6] W. Fang et al., “GPU-based efficient parallel heuristic algorithm for high-utility itemset mining in large transaction datasets (extended abstract),” 2024 IEEE 40th International Conference on Data Engineering (ICDE), pp. 5733–5734, May 2024. doi:10.1109/icde60146.2024.00498
[7] A. Tyspin, L. Ugadiaro, K. Khrabrov, and A. Telepov, “Gradual Optimization Learning for Conformational Energy Minimization,” ICLR, vol. 2024, Nov. 2023. doi:https://arxiv.org/abs/2311.06295
[8] H. Huang, L. He, F. Liu, R. Zhao, and L. Shi, “Neural Dynamics pruning for energy-efficient spiking neural networks,” 2024 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6, Jul. 2024. doi:10.1109/icme57554.2024.10688279