A review on green deployment for Edge AI - Abstract

Santiago del Rey[1,*], Silverio Martínez-Fernández[1] and Xavier Franch[1]

[1]Universitat Politècnica de Catalunya, Barcelona, Spain

Abstract

The convergence of edge computing and Artificial Intelligence, namely Edge AI, offers many opportunities to the industry for building competitive and innovative business models. However, this new paradigm has its own challenges in terms of latency, privacy, and energy. The latter is relevant considering that current AI requires expensive computation that is hard to achieve in existing edge devices. This work reviews 20 studies published between December 2018 and March 2023 on the subject of energy efficiency for the deployment of Edge AI. Most of the publications are devoted to improving the efficient deployment of Edge AI, while only a few focus on measuring the carbon footprint and energetic impact. Our work can help researchers quickly understand the state-of-the-art and learn which topics need more research.

Keywords

edge computing, energy-efficiency, deployment, edge AI, green AI

1. Technical description

With the appearance of the Internet of Things (IoT), new paradigms like edge computing have appeared to overcome the limitations of cloud computing. Given the high volume of data generated by edge devices, we are observing an increasing demand for systems that integrate edge computing and Artificial Intelligence (AI), which gives birth to the concepts of edge intelligence and intelligent edge, that make up what is known as Edge AI. [1, 2]. However, deploying machine learning (ML) models in edge devices is limited by the available resources of the devices and the communication network. Hence, Edge AI introduces new challenges in latency, cybersecurity, and especially, energy efficiency.

This work reviews 20 recent studies focusing on energy efficiency when deploying ML models for Edge AI. We extract and review 17 of them from a recent literature review [3]. We select the papers whose topic is deployment, according to the classification of the paper's authors. We perform forward snowballing as defined in [4] to add the three remaining studies. We group the studies by their main contributions. Figure 1 shows the themes identified and the number of papers for each theme. The complete list of papers reviewed is available at GitHub.[1]

In: B. Combemale, G. Mussbacher, S. Betz, A. Friday, I. Hadar, J. Sallou, I. Groher, H. Muccini, O. Le Meur, C. Herglotz, E. Eriksson, B. Penzenstadler, AK. Peters, C. C. Venters. Joint Proceedings of ICT4S 2023 Doctoral Symposium, Demonstrations & Posters Track and Workshops. Co-located with ICT4S 2023. Rennes, France, June 05-09, 2023.

*Corresponding author.

santiago.del.rey@upc.edu (S. del Rey); silverio.martinez@upc.edu (S. Martínez-Fernández); xavier.franch@upc.edu (X. Franch)

0000-0003-4979-414X (S. del Rey); 0000-0001-9928-133X (S. Martínez-Fernández); 0000-0001-9733-8830 (X. Franch)

[1]https://github.com/santidrj/green-edge-ai-deployment-papers

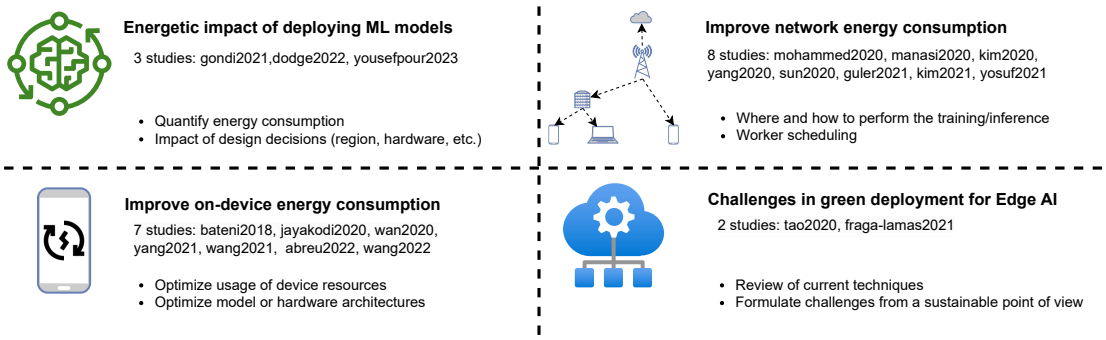

Figure 1: The 20 papers classified into four main themes based on their main contributions.

1.1. Energetic impact of deploying ML models (3 studies)

We find that little work has been done on analyzing the energetic impact of deploying ML models. Dodge et al. [5] show that the two most impactful factors on carbon footprint are geographical location and time of day, in this order. Hence, they propose two scheduling methods to optimize cloud workloads based on the time of day. Gondi and Pratap [6] evaluate the energy-accuracy trade-off of Automatic Speech Recognition (ASR) transformer models on an edge device. Their results show an exponential growth in CPU energy consumption as the word error rate (WER) improves linearly. Yousefpour et al. [7] quantify the carbon footprint of Federated Learning (FL). They find that asynchronous FL is faster than synchronous FL, but has higher carbon emissions. Moreover, they find that the overall benefits of higher concurrency (i.e., number of devices), considering resource consumption, do not scale linearly.

1.2. Improve network energy consumption (8 studies)

A significant portion of the studies reviewed proposes new frameworks to reduce energy by optimizing where and when is the training/inference performed [8]. Yosuf et al. [9] study how to place DNN inference models in a Cloud Fog Network architecture for energy efficiency. Their results show that significant savings can be achieved by the full utilization of edge devices. They also found that fog servers are bypassed in favor of cloud data centers. They argue this is caused due to the processing inefficiency and high Power Usage Effectiveness (PUE) of the fog servers. Kim and Wu [10] propose AutoScale, a tool that can select the optimal execution scaling decision based on the DNN characteristics, QoS and accuracy targets, underlying system profiles, and stochastic runtime variance. They improve inference energy efficiency by 9.8× and 1.6× compared to the baseline settings of mobile CPU and cloud offloading.

1.3. Improve on-device energy consumption (7 studies)

Many of the studies reviewed focus on optimizing the energy consumption in the edge devices [11]. Wang et al. [12] implement an online optimization framework connecting the asynchronous execution of federated training with application co-running to minimize energy consumption on mobile devices. By designating the training process to run in the background

while an application is running, they can save over 60% of energy with three times faster convergence speed compared to previous schemes. Abreu et al. [13] present a framework to facilitate the exploration of dedicated decision trees (DTs) and random forests (RFs) accelerators. The proposed framework translates tree-based structures to hardware description languages. Their approach achieves 10× power reduction compared to prior works.

1.4. Current challenges (2 studies)

Only two papers study the challenges of deploying ML models on the edge. Tao et al. [14] review the challenges of training DNNs with FPGA. They find these challenges mainly lie in the complexity of resource management and the requirements of both software and hardware design knowledge. Moreover, they propose an evaluation workflow and performance metric to consider on-chip resource usage, training efficiency, energy efficiency, and model accuracy. Fraga-Lamas et al. [15] provide a more general view and review the essential concepts related to the development of Edge AI Green IoT systems and their carbon footprint, and make a list of twelve open challenges.

2. Relevance and Novelty

With increased bandwidth and lower latency, edge computing promises to decentralize cloud applications. Meanwhile, the current AI methods assume computations are conducted in a powerful computational infrastructure, such as data centers with substantial computing and data storage capabilities. One of the main challenges of bringing edge computing and AI together remains in the energy constraints of edge devices.

This poster provides a brief overview of the state-of-the-art in green deployment for Edge AI. We find that some papers focus on very specific application areas, such as ASR [6], FL [7], or DTs [13], while some works are more general-purpose [5, 10]. In addition, excluding the two studies reporting current challenges, we find that 14 out of 18 papers report empirical results and the hardware used. We find that four papers report only using mobile phones or SoCs (e.g., Raspberry Pi, Nvidia Jetson), and two use a combination of both. While mobile phones vary greatly, we find that are the most commonly used devices for experimentation. Overall, we find that while most of the research is focused on improving energy efficiency by optimizing the edge devices' workload and communication, little work has been done on understanding the factors impacting energy consumption and carbon footprint (e.g., time of day, underlying hardware). This calls for putting more effort into understanding what elements contribute to increasing energy consumption and how. This can help to tackle the problem more accurately.

Acknowledgments

This work is part of the GAISSA project (TED2021-130923B-I00), which is funded by MCIN/AEI/10.13039/501100011033 and by the European Union "NextGenerationEU"/PRTR.

References

[1] X. Wang, Y. Han, V. C. M. Leung, D. Niyato, X. Yan, X. Chen, Edge AI: Convergence of Edge Computing and Artificial Intelligence, 2020. doi:10.1007/978-981-15-6186-3.

[2] S. Deng, H. Zhao, W. Fang, J. Yin, S. Dustdar, A. Y. Zomaya, Edge Intelligence: The Confluence of Edge Computing and Artificial Intelligence, IEEE IoT Journal 7 (2020) 7457–7469. doi:10.1109/JIOT.2020.2984887.

[3] R. Verdecchia, J. Sallou, L. Cruz, A systematic review of green ai, 2023. arXiv:2301.11047.

[4] C. Wohlin, Guidelines for snowballing in systematic literature studies and a replication in software engineering, in: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering, EASE '14, 2014. doi:10.1145/2601248.2601268.

[5] J. Dodge, T. Prewitt, R. Tachet des Combes, E. Odmark, R. Schwartz, E. Strubell, A. S. Luccioni, N. A. Smith, N. DeCario, W. Buchanan, Measuring the Carbon Intensity of AI in Cloud Instances, in: FAccT, 2022, pp. 1877–1894. doi:10.1145/3531146.3533234.

[6] S. Gondi, V. Pratap, Performance and efficiency evaluation of ASR inference on the edge, Sustainability 13 (2021). doi:10.3390/su132212392.

[7] A. Yousefpour, S. Guo, A. Shenoy, S. Ghosh, P. Stock, K. Maeng, S.-W. Krüger, M. Rabbat, C.-J. Wu, I. Mironov, Green Federated Learning, 2023. doi:10.48550/arXiv.2303.14604.

[8] T. Mohammed, A. Albeshri, I. Katib, R. Mehmood, UbiPriSEQ—Deep Reinforcement Learning to Manage Privacy, Security, Energy, and QoS in 5G IoT HetNets, Appl. Sci. 10 (2020) 7120. doi:10.3390/app10207120.

[9] B. A. Yosuf, S. H. Mohamed, M. M. Alenazi, T. E. H. El-Gorashi, J. M. H. Elmirghani, Energy-Efficient AI over a Virtualized Cloud Fog Network, in: e-Energy, 2021, pp. 328–334. doi:10.1145/3447555.3465378.

[10] Y. G. Kim, C.-J. Wu, AutoScale: Energy Efficiency Optimization for Stochastic Edge Inference Using Reinforcement Learning, in: MICRO-53, 2020, pp. 1082–1096. doi:10.1109/MICRO50266.2020.00090.

[11] S. Bateni, H. Zhou, Y. Zhu, C. Liu, PredJoule: A Timing-Predictable Energy Optimization Framework for Deep Neural Networks, in: RTSS, 2018, pp. 107–118. doi:10.1109/RTSS.2018.00020.

[12] C. Wang, B. Hu, H. Wu, Energy Minimization for Federated Asynchronous Learning on Battery-Powered Mobile Devices via Application Co-running, in: ICDCS, 2022, pp. 939–949. doi:10.1109/ICDCS54860.2022.00095.

[13] B. Abreu, M. Grellert, S. Bampi, A framework for designing power-efficient inference accelerators in tree-based learning applications, Engineering Applications of Artificial Intelligence 109 (2022) 104638. doi:10.1016/j.engappai.2021.104638.

[14] Y. Tao, R. Ma, M.-L. Shyu, S.-C. Chen, Challenges in Energy-Efficient Deep Neural Network Training With FPGA, in: CVPR Workshops, 2020, pp. 400–401.

[15] P. Fraga-Lamas, S. I. Lopes, T. M. Fernández-Caramés, Green IoT and Edge AI as Key Technological Enablers for a Sustainable Digital Transition towards a Smart Circular Economy: An Industry 5.0 Use Case, Sensors 21 (2021) 5745. doi:10.3390/s21175745.