DeepEn2023: Energy Datasets for Edge Artificial Intelligence

Xiaolong Tu
Department of Computer Science
Georgia State University
Atlanta, GA 30302
xtu1@student.gsu.edu

Anik Mallik
Department of Electrical and Computer Engineering
The University of North Carolina at Charlotte
Charlotte, NC 28223
amallik@uncc.edu

Haoxin Wang
Department of Computer Science
Georgia State University
Atlanta, GA 30302
haoxinwang@gsu.edu

Jiang Xie
Department of Electrical and Computer Engineering
The University of North Carolina at Charlotte
Charlotte, NC 28223
linda.xie@uncc.edu

Abstract

Climate change poses one of the most significant challenges to humanity. As a result of these climatic changes, the frequency of weather, climate, and water-related disasters has multiplied fivefold over the past 50 years, resulting in over 2 million deaths and losses exceeding $3.64 trillion USD. Leveraging AI-powered technologies for sustainable development and combating climate change is a promising avenue. Numerous significant publications are dedicated to using AI to improve renewable energy forecasting, enhance waste management, and monitor environmental changes in real time. However, very few research studies focus on making AI itself environmentally sustainable. This oversight regarding the sustainability of AI within the field might be attributed to a mindset gap and the absence of comprehensive energy datasets. In addition, with the ubiquity of edge AI systems and applications, especially on-device learning, there is a pressing need to measure, analyze, and optimize their environmental sustainability, such as energy efficiency. To this end, in this paper, we propose large-scale energy datasets for edge AI, named DeepEn2023, covering a wide range of kernels, state-of-the-art deep neural network models, and popular edge AI applications. We anticipate that DeepEn2023 will improve transparency in sustainability in on-device deep learning across a range of edge AI systems and applications. For more information, including access to the dataset and code, please visit https://amai-gsu.github.io/DeepEn2023.

1 Introduction

Environmentally-sustainable AI refers to the design and use of artificial intelligence (AI) and machine learning (ML) technologies to tackle environmental issues and advance sustainability [1], which is a two-sided research area: AI for sustainability and sustainability of AI [2, 3]. While there is growing interest in using AI to achieve the Sustainable Development Goals (SDGs) [4] related to climate change, research addressing the environmental impact of AI itself remains limited [5] [6] [7] [8]. For instance, a sophisticated AI-empowered Internet of Things (IoT) system can be deployed to monitor and predict the total carbon emissions of a building or factory, aligning with the objective of AI for sustainability. However, this raises new questions: *How much carbon does this AI system emit? How sustainable is the AI system itself?*

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2023.

On-device learning on edge devices, such as smartphones, IoT devices, and connected vehicles, is increasingly prevalent for model personalization and enhanced data privacy, yet its impact in terms of carbon emission is often overlooked [1, 9, 10]. This oversight might be attributed to the typically modest power consumption and carbon footprint of individual edge devices. However, when considering the immense proliferation of these AI-empowered devices worldwide, their cumulative carbon footprint would be substantial and cannot be overlooked. For instance, consider a scenario where an individual uses AI-powered applications on their smartphone for one hour every day. The average power consumption of a smartphone is 3W [11]. With 6.4 billion smartphone connections reported in 2022 [12], the cumulative energy consumption of these smartphones amounts to $19,200$ MWh per day. Based on the U.S. electricity generation carbon intensity of 371.2kg of carbon per MWh [13], the estimated daily carbon emissions from these smartphones would be 7127.04 metric tons. For comparison, this is equivalent to the annual carbon footprint of $1,848$ gasoline-powered passenger vehicles. [14]. Therefore, to understand and evaluate the sustainability of AI systems, especially edge AI systems, we have developed three large-scale energy consumption datasets: *kernel-level, model-level, and application-level*. We hope our energy datasets, named *DeepEn2023*, will encourage both the research community and end-users to prioritize sustainability in on-device learning and edge AI, a principle that drives our research.

2 Energy Measurement Platform

We developed an energy measurement platform employing the Monsoon Power Monitor to capture power consumption data during model execution. This power data, combined with inference latency, is used to generate energy datasets. The Monsoon Power Monitor is selected for its millisecond-level data granularity. Since most DNN model latencies, typically between 10 to 200 ms on mobile CPUs, can be significantly decreased to 1 to 50 ms on mobile GPUs. Compared to built-in smartphone sensors, the Monsoon provides more accurate and detailed power consumption data, especially for models running on edge devices. Fig.1 illustrates the power measurement platform we have implemented. We connected battery-removed smartphones to the power monitor using power cables. Then use Monsoon power monitor to power on the devices and measure power consumption during model execution with a granularity of up to 0.2 ms. We generated thousands of TensorFlow Lite models across various levels and executed them on different hardware platforms to create a comprehensive dataset.

For our study, we selected eight modern edge devices featuring eight different mobile SoCs, including at least one high-end and one mid-range SoC from leading chipset vendors such as Qualcomm, HiSilicon, and MediaTek. These SoCs have been chosen for their status as representative and advanced mobile AI silicon widely used in the last two years.

3 DeepEn2023: Energy Consumption Dataset

In this section, we provide details of our datasets and how it contributes to understanding the energy consumption and carbon emissions of edge AI systems. We have generated comprehensive datasets for typical kernels, models and applications across various configurations. We also discuss how each dataset can facilitate research efforts aimed at accessing the adverse impact of AI carbon emissions on global climate change.

3.1 Kernel-level Energy Consumption Dataset

Kernels constitute the fundamental units of execution in deep learning frameworks, with their types and configuration parameters significantly influencing the energy consumption during DNN model executions. In Table 1 we list nine typical kernels that are present in almost all CNN models, with the energy consumption and the carbon emission range for different configurations. The primary configurations include input height and width (HW)[1], input channel number (C_{in}), output channel number (C_{out}), kernel size (KS), and stride (S). Here are the key observations : 1) Energy consumption varies significantly for same kernel with different configurations on CPU and GPU. 2) Different configuration parameters have varying impacts for the kernels energy consumption 3) `conv+bn+relu` kernels typically consume more energy than other kernel types. 4) Across almost all

[1] In CNN models, input height usually is equal to input width.

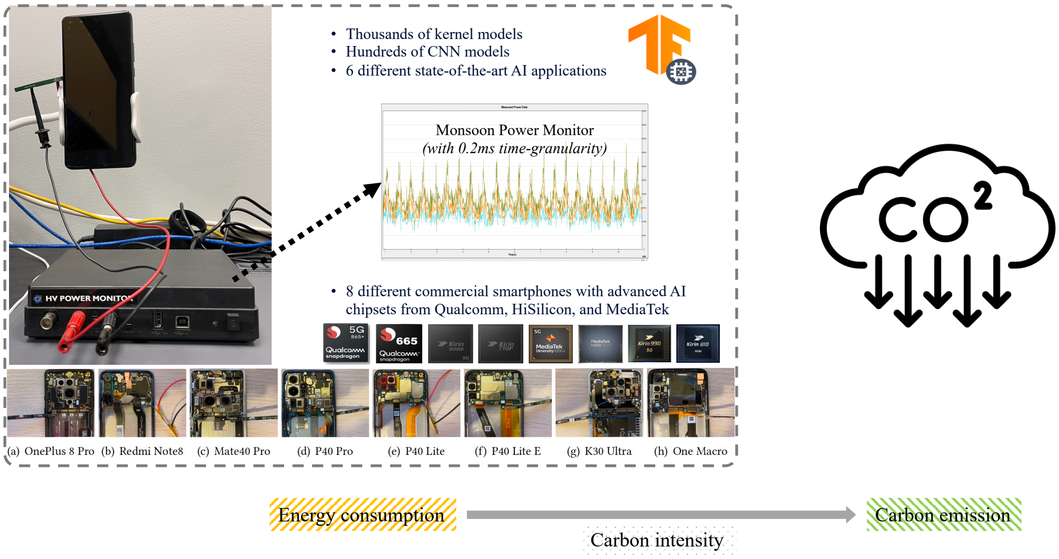


Figure 1: Power Measurement Platform utilizing the Monsoon Power Monitor to capture energy consumption data. Then, carbon intensity are used to convert this energy data into carbon emission estimates.

Table 1: Measured kernels per device in our kernel-level dataset.

Kernels	Energy Consumption (mJ)		Carbon Emission (gCO2eq/kWh)[2]		# Measured kernels		Configurations
	CPU min - max	GPU min - max	CPU min - max	GPU min - max	CPU	GPU	
`conv+bn+relu`	0.002 - 1200.083	0.002 - 120.152	1.762×10^{-10} - 1.057×10^{-4}	1.762×10^{-10} - 1.058×10^{-5}	1032	1032	$(HW, C_{in}, C_{out}, KS, S)$
`dwconv+bn+relu`	0.022 - 222.609	0.016 - 0.658	1.938×10^{-9} - 1.961×10^{-5}	1.409×10^{-9} - 5.797×10^{-8}	349	349	(HW, C_{in}, KS, S)
`bn+relu`	0.002 - 161.334	0.001 - 14.594	1.762×10^{-10} - 1.421×10^{-5}	8.811×10^{-11} - 1.285×10^{-6}	100	100	(HW, C_{in})
`relu`	0.001 - 141.029	0.003 - 6.86	8.811×10^{-11} - 1.242×10^{-5}	2.643×10^{-10} - 6.044×10^{-7}	46	46	(HW, C_{in})
`avgpool`	0.066 - 7.711	0.034 - 1.142	5.815×10^{-9} - 6.794×10^{-7}	2.995×10^{-9} - 1.006×10^{-7}	28	28	(HW, C_{in}, KS, S)
`maxpool`	0.054 - 7.779	0.032 - 1.214	4.758×10^{-9} - 6.854×10^{-7}	2.819×10^{-9} - 1.069×10^{-7}	28	28	(HW, C_{in}, KS, S)
`fc`	0.038 - 94.639	-	3.348×10^{-9} - 8.338×10^{-7}	-	24	-	(C_{in}, C_{out})
`concat`	0.001 - 42.826	0.066 - 3.428	8.811×10^{-11} - 3.773×10^{-6}	5.815×10^{-9} - 3.020×10^{-7}	142	142	$(HW, C_{in1}, C_{in2}, C_{in3}, C_{in4})$
`others`	0.001 - 132.861	0.003 - 10.163	8.811×10^{-11} - 1.170×10^{-5}	2.643×10^{-10} - 8.954×10^{-7}	98	72	(HW, C_{in})

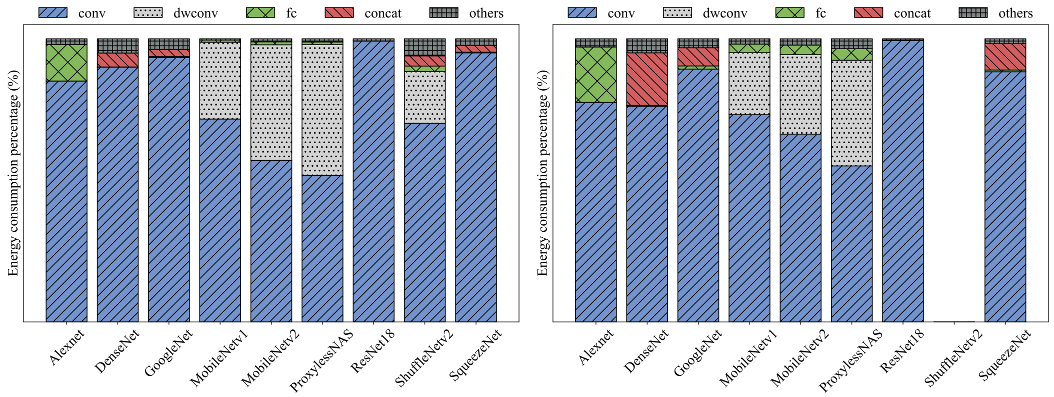
the kernels, GPU exhibit better energy efficiency under same configurations. Studying the impact of kernel configurations on energy consumption lays the foundation for a comprehensive understanding of energy usage during DNN model executions on edge devices. This emphasizes the importance of adaptive configuration selecting, in enhancing the energy efficiency of DNN models and how it can benefit researchers working toward carbon netural goal.

To build the dataset, we initially generate a large number of kernels with a variety of types (16 types for CPU and 10 types for GPU) featuring a range of configurations in the `tflite` format (e.g., 1032 `conv+bn+relu` and 349 `dwconv+bn+relu` kernels). These kernel configurations are randomly sampled. The number of sampled configurations for each kernel type hinges on two main factors: its configuration dimension and its impact on the overall energy consumption during DNN executions. This dataset provides researchers with detailed insights into how energy is consumed within models and which configurations or parameters affect kernel energy efficiency. Researchers can use this dataset to adapt configurations with best erergy efficiency on edge devices, consequently reducing carbon emissions.

3.2 Model-level Energy Consumption Dataset

We also introduce our model-level energy dataset, which collects nine state-of-the-art DNN models. These models represent a mix of both manually-designed and NAS-derived models, each with distinct kernel types and configurations. For each model, we generate 50 variants for conducting power and energy measurements by re-sampling the C_{out} and KS for each layer. Specifically, we randomly sample the new output channel number from a range of 20% to 180% of the original C_{out},

[2] The unit of measurement typically used for quantifying and comparing carbon emissions is CO2 equivalents.



(a) Mobile CPU

(b) Mobile GPU

Figure 2: DNN model energy consumption percentage breakdown. The top four most energy-consuming kernel types are `conv+bn+relu` (`conv`), `dwconv+bn+relu` (`dwconv`), `fc`, and `concat`.

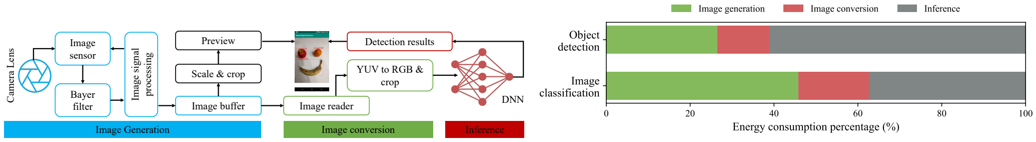
while the KS is sampled from the set of values: $\{1, 3, 5, 7, 9\}$. Generally, running these models on mobile GPUs results in an energy consumption reduction of approximately 49% to 79%, compared to the execution on mobile CPUs. Fig. 2 presents the energy consumption breakdown of individual models by kernel types. The four kernel types that consume the most energy are `conv+bn+relu`, `dwconv+bn+relu`, `fc`, and `concat`. They account for 79.27%, 14.79%, 2.03%, and 1.5% of the total model energy consumption on the mobile CPU, respectively. On the mobile GPU, these kernels represent 78.17%, 10.91%, 4.01%, and 4.28% of the total model energy consumption. Furthermore, in most models, `conv+bn+relu` and `dwconv+bn+relu` account for the main energy percentages. On average, `conv+bn+relu` and `dwconv+bn+relu` take 93.97% and 87.74% of the total model energy consumption on the mobile CPU and GPU, respectively. With this model-level energy consumption dataset, researchers can visually see the energy consumption of different models on various platforms, helping them choose he most energy-efficient models according to their needs.

3.3 Application-level Energy Consumption Dataset

The kernel- and model-level datasets can be beneficial for researchers and developers in understanding, modelling, and optimizing power and energy efficiency of DNN executions. However, the energy efficiency of applications on edge devices has a more direct impact on carbon emissions. To adress this, we create an application-level dataset, which uncovers the end-to-end energy consumption of six popular edge AI applications, covering three main categories: vision-based (object detection, image classification, super resolution, and image segmentation), NLP-based (natural language question answering), and voice-based applications (speech recognition). As shown in Table 2, we measure the power and energy consumption of each application with multiple reference DNN models that operate under four distinct computational settings, including CPU with a single thread, CPU with four threads, GPU delegate, and the NNAPI delegate. The dataset can serve as a resource for exploring the energy consumption distribution throughout the end-to-end processing pipeline of an edge AI application. For example, we can use the dataset to examine the energy consumed in generating image frames, converting these frames from YUV to RGB, and conducting DNN inference within an object detection application. It demonstrates that our application-level dataset can provide interpretable observations for comprehending who is the primary energy consumer in the end-to-end edge AI application. Fig. 3 depicts the energy consumption breakdown based on the processing phases in the object detection. It demonstrates that our application-level dataset can provide interpretable observations for comprehending who is the primary energy consumer in the end-to-end edge AI application.

3.4 Beneficial For Global Climate Change

These three datasets can contribute to addressing global climate change from different perspectives. For example, the kernel-level dataset can assist researchers in identifying the most energy-efficient



(a) End-to-end processing pipeline for object detection and image classification

(b) Energy consumption percentage breakdown

Figure 3: End-to-end energy consumption breakdown for object detection and image classification based on our application-level dataset.

Table 2: Measured edge AI applications per device in our application-level dataset.

Category	Application	Reference DNN models	Delegate				Model size
			CPU1	CPU4	GPU	NNAPI	(MB)
Vision-based	Image detection	MobileNetv2, FP32, 300 × 300 pixels	✓	✓		✓	24.2
		MobileNetv2, INT8, 300 × 300 pixels	✓	✓		✓	6.9
		MobileNetv2, FP32, 640 × 640 pixels	✓	✓		✓	12.3
		MobileNetv2, INT8, 640 × 640 pixels	✓	✓		✓	4.5
	Image classification	EfficientNet, FP32, 224 × 224 pixels	✓	✓	✓	✓	18.6
		EfficientNet, INT8, 224 × 224 pixels	✓	✓		✓	5.4
		MobileNetv1, FP32, 224 × 224 pixels	✓	✓	✓	✓	4.3
		MobileNetv1, INT8, 224 × 224 pixels	✓	✓		✓	16.9
	Super resolution	ESRGAN , FP32, 50 × 50 pixels	✓		✓		5
	Image segmentation	DeepLabv3 , FP32, 257 × 257 pixels		✓			2.8
NLP-based	Natural language question answering	MobileBERT , FP32	✓	✓		✓	100.7
Voice-based	Speech recognition	Conv-Actions-Frozen , FP32	✓	✓		✓	3.8

kernel configurations and parameters, finding the balance between computing performance and carbon emissions. We have used our dataset to train a random forest model to predict the energy consumption and carbon emissions of unseen models, and the accuracy is quite promising [15]. The model-level dataset aids researchers in discovering the most energy-efficient models based on various deployment requirements. For instance, Model A deployed on a CPU may exhibit better energy efficiency than Model B with the same accuracy for image classification. The application-level dataset provides researchers with insights into the end-to-end energy consumption of an application on edge devices, enabling them to implement more comprehensive measures to reduce energy consumption.

4 Conclusion

In this paper, we present our energy consumption datasets, DeepEn2023, from kernel-level, model-level, and application-level to facilitate research and development aimed at improving the energy efficiency and reducing the carbon emissions of AI applications on diverse edge devices. These datasets are valuable resources and tools for researchers and community to design energy-efficiency AI systems with fewer greenhouse gas emissions, thus contributing to the global climate change mitigation. We hope DeepEn2023 can help shift the mindset of both end-users and the research community towards sustainable edge AI, a principle that drives our research.

Acknowledgments and Disclosure of Funding

This work was supported by the US National Science Foundation (NSF) under Grant No. 1910667, 1910891, and 2025284.

References

[1] Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, Jinshi Huang, Charles Bai, et al. Sustainable AI: Environmental implications, challenges and opportunities. *Proceedings of Machine Learning and Systems*, 4:795–813, 2022.

[2] Aimee Van Wynsberghe. Sustainable AI: AI for sustainability and the sustainability of AI. *AI and Ethics*, 1(3):213–218, 2021.

[3] Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Daniela Langhans, Max Tegmark, and Francesco Fuso Nerini. The role of artificial intelligence in achieving the sustainable development goals. *Nature communications*, 11(1):1–10, 2020.

[4] Sustainbale Development Goal. `https://sdgs.un.org/goals/goal13`. Accessed on September 2023.

[5] Jie You, Jae-Won Chung, and Mosharaf Chowdhury. Zeus: Understanding and optimizing {GPU} energy consumption of {DNN} training. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*, pages 119–139, 2023.

[6] Manni Wang, Shaohua Ding, Ting Cao, Yunxin Liu, and Fengyuan Xu. Asymo: scalable and efficient deep-learning inference on asymmetric mobile cpus. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, pages 215–228, 2021.

[7] Simin Chen, Mirazul Haque, Cong Liu, and Wei Yang. Deepperform: An efficient approach for performance testing of resource-constrained neural networks. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*, pages 1–13, 2022.

[8] Dongqi Cai, Qipeng Wang, Yuanqiang Liu, Yunxin Liu, Shangguang Wang, and Mengwei Xu. Towards ubiquitous learning: A first measurement of on-device training performance. In *Proceedings of the 5th International Workshop on Embedded and Mobile Deep Learning*, pages 31–36, 2021.

[9] Carole-Jean Wu, David Brooks, Kevin Chen, Douglas Chen, Sy Choudhury, Marat Dukhan, Kim Hazelwood, Eldad Isaac, Yangqing Jia, Bill Jia, et al. Machine learning at facebook: Understanding inference at the edge. In *Proceedings of 2019 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pages 331–344, 2019.

[10] Stefano Savazzi, Sanaz Kianoush, Vittorio Rampa, and Mehdi Bennis. A framework for energy and carbon footprint analysis of distributed and federated edge learning. In *Proceedings of 2021 IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pages 1564–1569, 2021.

[11] Haoxin Wang, BaekGyu Kim, Jiang Xie, and Zhu Han. Energy drain of the object detection processing pipeline for mobile devices: Analysis and implications. *IEEE Transactions on Green Communications and Networking*, 5(1):41–60, 2020.

[12] The Mobile Economy 2023. `https://www.gsma.com/mobileeconomy/wp-content/uploads/2023/03/270223-The-Mobile-Economy-2023.pdf`. Accessed on September 2023.

[13] How much carbon dioxide is produced per kilowatthour of U.S. electricity generation? `https://www.eia.gov/tools/faqs/faq.php?id=74&t=11`. Accessed on September 2023.

[14] Greenhouse Gas Equivalencies Calculator. `https://www.epa.gov/energy/greenhouse-gas-equivalencies-calculator#results`. Accessed on September 2023.

[15] Xiaolong Tu, Anik Mallik, Dawei Chen, Kyungtae Han, Onur Altintas, Haoxin Wang, and Jiang Xie. Unveiling energy efficiency in deep learning: Measurement, prediction, and scoring across edge devices. In *Proc. The Eighth ACM/IEEE Symposium on Edge Computing (SEC)*, pages 1–14, 2023.