# Interventions to Reduce AI Energy Requirements

Daniel Edelman*, Joseph McDonald†, David Bestor†, Michael Jones †, Baolin Li ¶
Devesh Tiwari ¶, Dan Zhao‡, Siddharth Samsi†, Vijay Gadepally†§
* MIT EECS, † MIT Lincoln Laboratory, ‡ MIT CSAIL, § MIT Connection Sciences ¶ Northeastern University

## I. Introduction

The ever-growing computational requirements of AI and its associated development and deployment costs are a widely understood trend [1]–[3]. This increasing computational demand naturally translates into increased energy usage and, in most cases, increased carbon emissions from datacenters where these models are developed and deployed. For example, particular Natural Language Processing (NLP) models can consume as much $CO_2$ emission as the lifetime emission of 5 cars [4]. In particular, the widespread adoption, proliferation, and development of large neural networks has made it increasingly important for AI practitioners to account for the environmental and climate impacts of AI development. Creating power and energy efficient methods to train neural networks could reduce the carbon footprint of these models and thus lessen the environmental impact of AI. While there are numerous examples of research into efficient machine learning models [5], [6], the focus of our paper is on easy-to-implement interventions that can be readily applied by ML practitioners without significant modifications to their code. Further, some of these interventions seem to provide energy efficiency gains almost "for free" in that they do not affect the accuracy or precision of the trained model and may incur minimal changes in computational performance. To illustrate the potential impact of these interventions, we highlight selected results via a popular neural network architecture on a common computer vision benchmark, to quantify improvements in energy efficiency and corresponding changes to model accuracy with relatively simple, straightforward tweaks. In particular, we explore simple modification to the training algorithm, such as altering the learning rate schedule, along with simple hardware-level interventions such as capping the power draw of GPUs or reducing the level of precision in numerical representations.

Our paper also describes a new effort on open-sourcing datasets that can be used by ML researchers interested in better understanding the relationship between machine learning applications, energy usage and carbon emissions. We hope this proposed *Green AI Challenge* will spur research in this fledgling but important field of study.

This material is based upon work supported by the Assistant Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001, United States Air Force Research Laboratory and the United States Air Force Artificial Intelligence Accelerator under Cooperative Agreement Number FA8750-19-2-1000. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Assistant Secretary of Defense for Research and Engineering, or the United States Air Force. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

## II. What are Interventions?

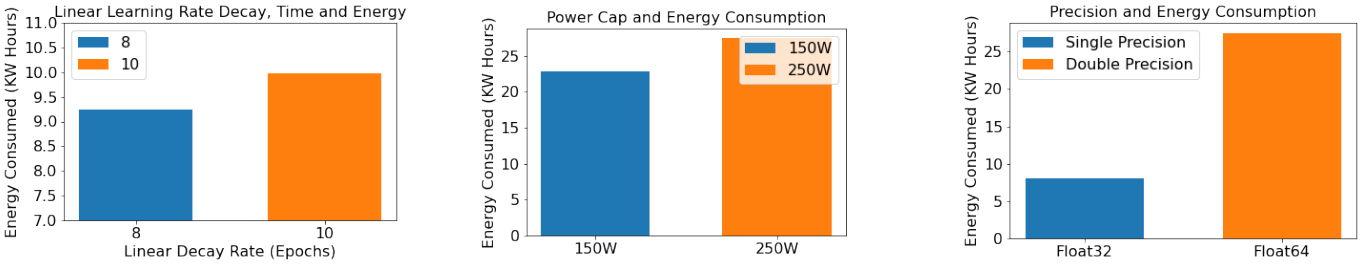
Approaches to reducing energy requirements and associated carbon usage of the training and inference of machine learning models can occur in many ways. For example, one could move the training of a large-scale workload from an inefficient datacenter to an efficient datacenter. One could better optimize how software leverages existing hardware or even develop more efficient algorithms. From our perspective, we see the variety of approaches being roughly broken in to three categories:

- Data Center Level Interventions: This class of interventions focus on techniques to make how one uses a data centers more efficient. For example, these interventions may move workloads towards times where data centers are more efficient [7] or migrate workloads to lower PUE data centers [8].
- Hardware Level Interventions: This class of interventions focus on making efficient hardware choices. For example, these interventions may focus on tuning and optimizing general purpose computing or suggest specialized computing platforms for particular computing kernels.
- Software and Algorithmic Interventions: This class of interventions focus on improving the efficiency of software systems and algorithm development. For example, debloating sofware systems that make them inherently more efficient or techniques that make tasks such as neural architecture search more efficient [9].

We admit that the above categories may be viewed as a simplification of the numerous avenues of related research but hope that they provide a birds-eye view of different approaches.

## III. Do these interventions actually work?

The high-level interventions described in the previous section can have a huge impact to energy consumption. For example, shifting workloads to more efficient times of day (e.g., day to night) can yield 10-20% energy reductions which are more pronounced during heat waves [7]. Similarly, if you are performing neural architecture search, it is possible to terminate hyper parameter combinations that are unlikley to yield meaningful results which can reduce energy usage by nearly 80% [10]. Similarly, hardware level interventions such as power capping can yield impressive savings with minimal impact to the application. Below, we highlight a few results based on hardware and algorithm interventions.




(a) Linear Decay Rate and Energy Consumption

(b) GPU Power Cap and Energy Consumption

(c) Precision and Energy Consumption

Fig. 1: A few simple interventions in action. These simple tweaks such as modifying the learning rate schedule, power capping, and reducing training precision can lead to nearly 70-80% reduction (cumulatively) in energy consumed for an image classification task

The reference implementation used for these results is the ResNet50 [11] model trained on the ImageNet [12] dataset to a top-1 accuracy of 70% on our institution's computing cluster. Each node on this cluster consists of two Nvidia V100 GPUs and our implementation is based on the MLPerf [13] Image Classfication benchmark. Several experiments were constructed that involved adjusting the batch-size chosen for training (default is 256) as well as various parameters such as the learning rate (default 0.1), momentum (default 0.9) and weight decay (default 1e-4). Selected results for three interventions – modifying the learning rate, implementing hardware power limits, and reducing computational precision –are shown in Figure 1.

In Figure 1a, we explored the effects of changing the learning rate decay scheduler. A simple linear decay rate was chosen initially (learning rate would decay after 10 epochs) and results are likely to translate to other learning rate schedulers. As seen in the figure, simply changing the decay occuring every 10 Epochs instead of every 8 epochs reduces the overall energy for the entire job by nearly 10%.

In Figure 1b, we explore a simple but powerful change to our GPUs power draw. Rather than the default 250W power cap, we limit our NVidia V100 GPU to draw up to 150W. This simple change reduces the overall energy consumption by over 15% with a negligible difference in run-time.

Finally, in Figure 1c, we explore changing the computational precision being used for training the image classification model. This simple change of training using signle precision (float32) instead of the default double precision (float64) results in an impressive reduction of energy needed for training by nearly 75% with no significant reduction in accuracy. Further experiments show promise of completing training faster and more efficiently beyond the 70% benchmark with no signifacnt reduction in accuracy.

## IV. Show me your data – the Green AI Challenge

3It is our belief that driving research into fundamental problems such as datacenter usage and optimization can be largely limited by the availability and accessibility of relevant data. For example, datacenter oriented data sets such as [14], [15] have led to new innovations and insight into the operation of modern datacenters [16], [17]. Similarly, in order to drive research into understanding and potentially mitigating the environmental impact of AI, our team is developing a Green AI Challenge. The chief goals of the Challenge are:

1) Development of power-efficient approaches to both AI training and inference with the goal of improving petaflops/watt performance.
2) Data-efficient computing to reduce training resources.
3) Informed machine learning to simplify data or model design by using or encoding prior knowledge.
4) Energy-efficient neural network design.
5) Adaptive, scalable, energy-efficient data center management.

To provide a budding research community with real problems, we will open-source detailed records from our institution's datacenter and call on others to do the same. Such a collaborative and sustained effort will not only spur research into this critical area, but also train a generation of machine learning developers where energy efficiency is a first-order priority.

## V. Conclusions

In this article, we describe early approaches to reducing computational demands, energy impact, and associated carbon emissions for machine learning – a growing source of data center computing usage. We motivate these approaches with a few selected results and describe our intention to release data to help drive further research into this field.

## Acknowledgements

Omitted for anonymity purposes. he authors acknowledge the MIT Lincoln Laboratory Supercomputing Center (LLSC) for providing HPC resources that have contributed to the research results reported in this paper. The authors wish to acknowledge the following individuals for their contributions and support: Bob Bond, Andrew Bowne, Garry Floyd, Jeff Gottschalk, Tim Kraska, CK Prothmann, Charles Leiserson, Dave Martinez, John Radovan, Steve Rejto, Daniela Rus, Marc Zissman, Matthew L Weiss, Michael Jones, Albert Reuther,

William Arcand, William Bergeron, Chansup Byun, Michael Houle, Matthew Hubbell, Hayden Jananthan, Jeremy Kepner, Kurt Keville, Anna Klein, Adam Michaleas, Peter Michaleas, Lauren Milechin, Julia Mullen, Charles Yee, Andrew Prout, and Antonio Rosa.

## REFERENCES


[1] N. C. Thompson, K. Greenewald, K. Lee, and G. F. Manso, "The computational limits of deep learning," 2020. [Online]. Available: https://arxiv.org/abs/2007.05558

[2] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green ai," *Communications of the ACM*, vol. 63, no. 12, pp. 54–63, 2020.

[3] P. Dhar, "The carbon impact of artificial intelligence." *Nat. Mach. Intell.*, vol. 2, no. 8, pp. 423–425, 2020.

[4] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 3645–3650. [Online]. Available: https://aclanthology.org/P19-1355

[5] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 781–10 790.

[6] H. Wang, Z. Zhang, and S. Han, "Spatten: Efficient sparse attention architecture with cascade token and head pruning," in *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2021, pp. 97–110.

[7] J. McDonald, B. Li, N. Frey, D. Tiwari, V. Gadepally, and S. Samsi, "Great power, great responsibility: Recommendations for reducing energy for training language models," *arXiv preprint arXiv:2205.09646*, 2022.

[8] D. Gmach, Y. Chen, A. Shah, J. Rolia, C. Bash, T. Christian, and R. Sharma, "Profiling sustainability of data centers," in *Proceedings of the 2010 IEEE International Symposium on Sustainable Systems and Technology*. IEEE, 2010, pp. 1–6.

[9] R. Ru, C. Lyle, L. Schut, M. Fil, M. van der Wilk, and Y. Gal, "Speedy performance estimation for neural architecture search," *Advances in Neural Information Processing Systems*, vol. 34, pp. 4079–4092, 2021.

[10] N. C. Frey, D. Zhao, S. Axelrod, M. Jones, D. Bestor, V. Gadepally, R. Gómez-Bombarelli, and S. Samsi, "Energy-aware neural architecture selection and hyperparameter optimization," in *2022 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. IEEE, 2022, pp. 732–741.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[13] P. Mattson, C. Cheng, G. Diamos, C. Coleman, P. Micikevicius, D. Patterson, H. Tang, G.-Y. Wei, P. Bailis, V. Bittorf *et al.*, "Mlperf training benchmark," *Proceedings of Machine Learning and Systems*, vol. 2, pp. 336–349, 2020.

[14] S. Samsi, M. L. Weiss, D. Bestor, B. Li, M. Jones, A. Reuther, D. Edelman, W. Arcand, C. Byun, J. Holodnack *et al.*, "The mit supercloud dataset," in *2021 IEEE High Performance Extreme Computing Conference (HPEC)*. IEEE, 2021, pp. 1–8.

[15] M. Jeon, S. Venkataraman, A. Phanishayee, J. Qian, W. Xiao, and F. Yang, "Analysis of {Large-Scale}{Multi-Tenant}{GPU} clusters for {DNN} training workloads," in *2019 USENIX Annual Technical Conference (USENIX ATC 19)*, 2019, pp. 947–960.

[16] D. Narayanan, K. Santhanam, F. Kazhamiaka, A. Phanishayee, and M. Zaharia, "{Heterogeneity-Aware} cluster scheduling policies for deep learning workloads," in *14th USENIX Symposium on Operating Systems Design and Implementation (OSDI 20)*, 2020, pp. 481–498.

[17] B. Li, R. Arora, S. Samsi, T. Patel, W. Arcand, D. Bestor, C. Byun, R. B. Roy, B. Bergeron, J. Holodnak *et al.*, "Ai-enabling workloads on large-scale gpu-accelerated system: Characterization, opportunities, and implications," in *2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2022, pp. 1224–1237.