Towards Green, Accurate, and Efficient AI models through Multi-Objective Optimization

Udit Gupta, Daniel Jiang, Max Balandat, Carole-Jean Wu
Meta AI

1 Introduction

Machine learning is one of the fastest growing services in modern hyperscale data centers. In Google’s data centers, for instance, AI accounts for up to 15% of all infrastructure cycles (Patterson et al., 2021). At Meta, AI training and inference capacity grew by 2.9× and 2.5×, respectively over 1.5 years; during a similar time period, data capacity for Meta’s production recommendation models doubled (Wu et al., 2022). While the exponential scaling has enabled unprecedented capabilities across computer vision, natural language processing, generative AI, protein modeling, personalized recommendation, it comes at the expense of significant energy and environmental footprints.

Recent work demonstrates the high carbon footprint incurred by AI at scale. Patterson et al. (2021) illustrate that training a single natural language processing model, such as Meena, consumes the equivalent carbon footprint of thousands of miles traveled by an average passenger vehicle. However, such estimates account for only a portion of AI’s holistic environmental impact as they quantify emissions owing to *operational* energy consumption during *training* alone.

We must consider the real environmental impact of the end-to-end AI ecosystem going forward including, model life cycles (i.e., training, inference) and hardware life cycles (i.e., hardware manufacturing, operational hardware use). As an example, when considering only dynamic power consumption, training a 176 billion parameter BLOOM language model consumes 24.7 metric tonnes of CO_2; on the other hand, considering the holistic training process, including overheads of manufacturing training platforms, the training footprint increases to 50.5 metric tonnes of CO_2, doubling the environmental impact. And this does not include any of the impact incurred from serving – i.e., actually using – the model (which is often even larger). At Meta, data collection and storage, training and experimentation, and inference deployment have been found to account for roughly 30%, 30%, and 40% of AI infrastructure footprint (Wu et al., 2022).

The goal of this proposed project is to investigate the design of environmentally sustainable, accurate, and efficient AI models. We frame this as a *multi-objective optimization* problem with understudied and non-intuitive tradeoffs. Central to this work is holistically accounting for the environmental impact across both model and hardware life cycles.

2 Research proposal

In this section, we provide details on the problem formulation that we intend to study. At a high level, we are interested in maximizing *performance*-related objectives, namely accuracy and efficiency, while keeping the sustainability costs low. The levers (or decision variables) at our disposal are model and training hyperparameters (e.g., architectures, learning rates, batch sizes) and compute parameters (e.g., which datacenter to use, hardware platform). In general, both types of levers can influence both performance and sustainability objectives: for example, a model architecture with more parameters will have an adverse effect on sustainability costs. In the subsequent subsections, we detail the optimization methods and models used in our proposed study.

2.1 Multi-Objective Optimization

It is clear that there are difficult trade-offs to consider between performance and sustainability objectives. At the same time, evaluating the objectives is time-consuming and expensive, as it requires training a model to completion. We therefore propose to study this problem using the framework of *multi-objective Bayesian optimization* (MOBO) (Belakaria et al., 2019; Daulton et al., 2020; Konakovic Lukovic et al., 2020; Paria et al., 2020; Daulton et al., 2021), a set of techniques for sample-efficient multi-objective optimization. Given decision variables $\boldsymbol{x} \in \mathcal{X}$ and a

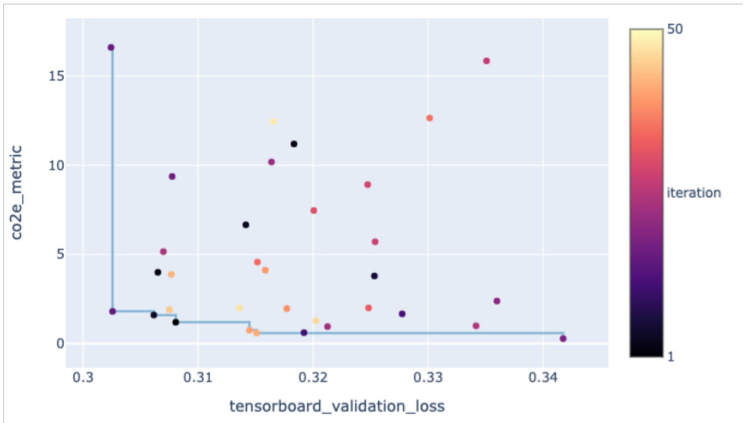

Figure 1: An example Pareto frontier after 50 trials of multi-objective Bayesian optimization on a toy recommendation model, where we tuned the optimizer learning rate, the embedding dimension, and the number of training epochs. Explicitly accounting for CO2 emissions as an optimization objective, we find configurations with low validation loss and significantly reduced CO2 impact.

vector-valued objective function $\boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}^M$, the goal of MOBO is to identify a Pareto frontier $\mathcal{P}^* = \{\boldsymbol{f}(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X}, \nexists\, \boldsymbol{x'} \in \mathcal{X} \text{ s.t. } \boldsymbol{f}(\boldsymbol{x}) \succ \boldsymbol{f}(\boldsymbol{x}')\}$, where $\boldsymbol{f}(\boldsymbol{x}) \succ \boldsymbol{f}(\boldsymbol{x}')$ means $\boldsymbol{f}(\boldsymbol{x}) \geq \boldsymbol{f}(\boldsymbol{x}')$ with at least one dimension's inequality being strict. In other words, the Pareto frontier represents the set of tradeoffs where one cannot improve one objective without hurting another. We plan to use, and build upon, the *noisy expected hypervolume improvement* (qNEHVI) method (Daulton et al., 2021), which are available as open source software (Balandat et al., 2020).

2.2 Optimization models

Post-deployment model. We first describe the *post-deployment* version of the problem, where we focus on the sustainability cost of the model while it is deployed. This includes costs due to inference and retraining, but excludes *experimentation* costs accrued during hyperparameter tuning. Let $\boldsymbol{x} \in \mathbb{R}^d$ represent our decision levers (hyperparameters and compute parameters) and let $\boldsymbol{\xi} \in \mathbb{R}^l$ be a vector a carbon intensity values, one for each datacenter location. We denote our performance objectives by $\boldsymbol{g}(\boldsymbol{x})$ and our sustainability cost objective by $c_{\text{dep}}(\boldsymbol{x})$, which considers the model's expected deployment lifetime. The *post-deployment* (multi-objective) optimization problem is

$$\max_{\boldsymbol{x} \in \mathcal{X}} \big[\boldsymbol{g}(\boldsymbol{x}), -c_{\text{dep}}(\boldsymbol{x})\big], \tag{1}$$

where the square brackets indicate concatenation of the objectives. Given that (1) takes the form of a standard MOBO problem, we can apply qNEHVI. Figure 1 shows an illustrative example of qNEHVI applied to a toy model with two optimization objectives: validation loss and CO_2.

Joint experimentation and deployment model. Experimentation costs can be substantial, even when compared against training and inference (Wu et al., 2022, Figure 3). A richer and more realistic optimization model would therefore *jointly examine* experimentation and deployment costs, with the central question during the experimentation phase is: does the potential improved sustainability cost of the deployed model outweigh the cost of additional model optimization? Letting c_{exp} be the sustainability cost of a single experimentation trial and T be the total number of experimentation trials, an extension of (1) is

$$\max_{\boldsymbol{x}_0,\ldots,\boldsymbol{x}_T} \Big[\boldsymbol{g}(\boldsymbol{x}_T), -\textstyle\sum_{t=0}^{T-1} c_{\text{exp}}(\boldsymbol{x}_t) - c_{\text{dep}}(\boldsymbol{x}_T)\Big], \tag{2}$$

where $\boldsymbol{x}_t \in \mathcal{X}$ represents the parameter tested at step t of experimentation. An effective strategy for (2) will require planning multiple steps ahead; we will explore using lookahead Bayesian optimization strategies, such as Jiang et al. (2020).

2.3 Carbon accounting

A core aspect of this research project is to holistically quantify the carbon footprint of AI models by also considering impacts across hardware life cycles. This includes emissions from physically manufacturing hardware and its operational use data centers. Gupta et al. (2021) illustrate that given the significant efficiency optimizations and presence of renewable energy powering data centers, embodied emissions dominate the carbon footprint of modern data centers. Below we describe how we account for both operational and embodied emissions:

Operational carbon. To compute the operational emissions we the energy consumed by the AI training or inference, and carbon intensity of energy powering the data center. The energy consumed by AI training and inference heavily depends on the workload and the hardware used. For instance, for AI training, important algorithmic attributes include the model architecture (e.g., compute, memory, and storage requirements) and time to convergence (e.g., iterations per epoch, number of epochs, learning rate, batching dimension). These attributes impact not only operational emissions but also model accuracy and efficiency. The choice in underlying hardware also impacts the accuracy achieved in a given time budget, training and inference efficiency, and energy consumed.

Going beyond energy consumed we will also consider system exogenous inputs such as the carbon intensity of the data centers' power consumption. Given the growing number of data centers procuring renewable energy we intend to consider two cases. First, we will compute carbon emissions assuming the carbon intensity of the data center tracks the local power grid's intensity under a variety of temporal and geographic scenarios to emulate a diverse set of environments. Second, we will compute carbon emissions assuming carbon intensity of the data center follow a renewable energy source such as solar or wind. We expect the scenarios to impact tradeoffs in sustainability, influencing optimal designs that balance accuracy, efficiency, and sustainability.

Embodied carbon. Going beyond operational carbon, we will quantify embodied carbon footprint. To estimate the embodied carbon footprint, which comes from high environmental impacts during hardware manufacturing, of running AI workloads on specific hardware platforms we will build on recent research proposals Gupta et al. (2021; 2022). These works model the footprint from hardware manufacturing using a combination of life cycle analyses and analytical hardware, architectural carbon models (e.g., ACT Gupta et al. (2022)). We attribute these emissions to a given workload on the hardware by amortizing the run-time of the application by the lifetime of the hardware (e.g., 3-4 years for a typical server): runtime/lifetime. A key aspect of this work will be to develop embodied carbon estimates for high performance AI hardware including CPUs and GPUs based on publicly available life cycle analyses and hardware models.

2.4 Models and datasets

We plan to begin this study by analyzing the accuracy, efficiency, and sustainability tradeoffs of open-source models and datasets. To understand differences across domains we intend to study computer vision (ResNet50 trained on ImageNet50), natural language processing (Transformer trained on WikiText-103), and recommendation (DLRM trained on Criteo Kaggle and Terabyte).

3 Expected Outcome

Overall, we want to answer whether taking sustainability into account yields distinct optimal model designs compared to current practices of accuracy, time-to-accuracy, efficiency driven design. In situations where sustainability driven optimization yields distinct model designs it will be crucial to understand how the designs vary and when they vary in order to guide higher-level decisions on setting priorities around accuracy-driven, efficiency-driven, and sustainability-driven efforts. This will impact model architectures optimized for accuracy vs. time-to-accuracy, vs. efficiency vs. sustainability. Additional questions we hope to tackle and expected outcomes include:

- How should hyperscale data centers prioritize model optimization? Data centers may have variable, diurnal loads and variable renewable energy availability. This varying availability of resources may influence how we pick models to minimize experimentation time and carbon footprint. For instance, in hyperparameter tuning systems we may evaluates models in parallel or serially to balance accuracy, performance/efficiency, and sustainability.
- We want to extend existing frameworks to enable ML researchers and engineers to co-optimize models for accuracy, efficiency, and sustainability. We envision this will be done on top of Ax, an open-source, industry-grade adaptive experimentation framework.

REFERENCES

Maximilian Balandat, Brian Karrer, Daniel Jiang, Samuel Daulton, Ben Letham, Andrew G Wilson, and Eytan Bakshy. Botorch: A framework for efficient monte-carlo bayesian optimization. *Advances in neural information processing systems*, 33:21524–21538, 2020.

Syrine Belakaria, Aryan Deshwal, and Janardhan Rao Doppa. Max-value entropy search for multi-objective Bayesian optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

Samuel Daulton, Maximilian Balandat, and Eytan Bakshy. Differentiable expected hypervolume improvement for parallel multi-objective bayesian optimization. *Advances in Neural Information Processing Systems*, 33:9851–9864, 2020.

Samuel Daulton, Maximilian Balandat, and Eytan Bakshy. Parallel bayesian optimization of multiple noisy objectives with expected hypervolume improvement. *Advances in Neural Information Processing Systems*, 34:2187–2200, 2021.

Udit Gupta, Young Geun Kim, Sylvia Lee, Jordan Tse, Hsien-Hsin S Lee, Gu-Yeon Wei, David Brooks, and Carole-Jean Wu. Chasing carbon: The elusive environmental footprint of computing. In *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pp. 854–867. IEEE, 2021.

Udit Gupta, Mariam Elgamal, Gage Hills, Gu-Yeon Wei, Hsien-Hsin S Lee, David Brooks, and Carole-Jean Wu. Act: Designing sustainable computer systems with an architectural carbon modeling tool. In *Proceedings of the 49th Annual International Symposium on Computer Architecture*, pp. 784–799, 2022.

Shali Jiang, Daniel Jiang, Maximilian Balandat, Brian Karrer, Jacob Gardner, and Roman Garnett. Efficient nonmyopic bayesian optimization via one-shot multi-step trees. *Advances in Neural Information Processing Systems*, 33:18039–18049, 2020.

Mina Konakovic Lukovic, Yunsheng Tian, and Wojciech Matusik. Diversity-guided multi-objective Bayesian optimization with batch evaluations. *Advances in Neural Information Processing Systems*, 33:17708–17720, 2020.

Biswajit Paria, Kirthevasan Kandasamy, and Barnabás Póczos. A flexible framework for multi-objective Bayesian optimization using random scalarizations. In *Uncertainty in Artificial Intelligence*, pp. 766–776. PMLR, 2020.

David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*, 2021.

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.