

Department: Head

Evaluating Sustainability and Social Costs of Adversarial Training in Machine Learning

Syed M. Hasan, Abdur R. Shahid, and Ahmed

Imteaj

Southern Illinois University Carbondale

Abstract—The growing prevalence of adversarial attacks on machine learning models in consumer electronics necessitates enhancing adversarial robustness. Although adversarial training improves the robustness of a model against adversarial attacks, its sustainability remains a critical concern due to carbon emissions and the environmental impact of the extensive computational demands. To address this, we use the Robust Carbon Trade-Off Index metric, which establishes a relationship between robustness and carbon emissions, and introduce the Cost Per Unit of Robustness Change metric to quantify the economic impact of increasing robustness in terms of carbon emission costs measured by an economic metric quantifying the costs associated with carbon emissions. By examining the theoretical foundations, practical quantification techniques, and interdisciplinary research areas, we shed light on the multifaceted aspects of building sustainable and scalable models with robust adversarial defenses.

■ **ADVERSARIAL MACHINE LEARNING** explores the dynamics of understanding and defending against adversarial attacks on machine learning (ML) models, particularly in the context of consumer electronics. Conventionally, ML models are trained and tested under the assumption that the data encountered during deployment will mirror the distribution seen during training. However, adversaries often manipulate input

data in subtle ways to deceive or degrade the performance of ML models, posing significant risks of cyber attacks in consumer devices. These attacks on ML models are of two main types: white-box attacks where a attacker has access to training data of ML model and black-box attacks in which a attacker does not have access to the internal details of the target model. Evasion attacks, the most prevalent form of black-box attacks, are designed to mislead AI systems into making incorrect decisions which emphasize the need for robust defenses in intelligent systems of consumer electronics. Some of the popular evasion at-

Digital Object Identifier 10.1109/MCE.YYYY.Doi Number

Date of publication DD MM YYYY; date of current version DD MM YYYY

tacks include the Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), etc. Adversarial robustness is pivotal in ensuring the reliability and efficacy of ML models by protecting against adversarial attacks to manipulate model predictions [1]. One of the ways to achieve adversarial robustness is adversarial training [2]. Adversarial training involves continually updating the model with new adversarial examples, which helps the system adapt and improve its defenses over time. Based on these, we address emission quantification of adversarial training along with monetary cost to provide a better view of environmental consequence of this process.

MOTIVATION

The extensive computational demands of ML models involve significant use of CPUs, GPUs, and RAM, which in turn leads to higher carbon emissions [3], [4], [5]. To address these environmental concerns, initiatives like sustainable AI [6] and green AI [7] have been developed with the aim to reduce the environmental impact by minimizing the computational demands of AI technologies [8]. Ongoing research focuses on exploring the carbon emissions of these systems by considering various factors such as the efficiency of computational resources, data centers, energy sources, model complexity, and operational efficacy [9], [10], [11]. For instance, Li et al. [12] proposed a framework named DeSVig that is designed to enhance the robustness and security of industrial AI systems against adversarial attacks; However, it does not include the environmental consequences of adversarial training. In an attempt to generalize carbon quantification for adversarial training, we discover that the adversarial robust ML models has the broader impacts, which include environmental, societal, and economic effects.

The social cost of carbon (SCC) can be used to assess the economic value of environmental effects. The SCC is a metric that quantifies the long-term economic damage of carbon emissions as applied to consumer electronics. We calculate the SCC using the Dynamic Integrated Climate Economy model, which converts potential future damages into present-day dollar values. These calculations hinge critically on the discount rate, which is the societal time preference. A higher discount rate prioritizes near-term benefits, lowers the SCC, and potentially undervalues future climate burdens [13].

Our previous work [14] provides empirical evi-

dence of a direct relationship between the robustness of an adversarial ML model and its associated carbon emissions in introducing Robust Carbon Trade-Off Index (RCTI). This metric establishes a crucial environmental dimension in the design of secure and robust ML systems with another important findings for quantifying emission concerns.

CONTRIBUTION

We introduce the **Cost Per Unit of Robustness Change (CRC)**, which assigns a monetary value to the carbon-robustness trade-off. This metric is crucial as it allows stakeholders to balance the costs of robustness improvements against other priorities, such as budget constraints or sustainability goals. In essence, while the RCTI measures the trade-off between robustness and carbon emissions, the CRC extends this by incorporating economic considerations, clarifying the financial implications of this trade-off. The main contributions of this article are as follows:

- **Extension of RCTI:** We expand RCTI to cover a wider range of conditions, offering a more detailed analysis of the environmental costs associated with enhancing robustness in ML models. This provides a finer assessment of how effectively adversarial training converts emissions into robustness gains.
- **CRC Metric:** We propose CRC metric, adding an economic dimension to the carbon-robustness trade-off. It assigns a monetary value to robustness improvements related to carbon emissions, helping evaluate the financial implications of adversarial training. We also define robustness elasticity to explore varying CRCs.
- **Integrated Sustainability Framework:** By combining CRC with the extended RCTI, we create a framework that evaluates both environmental and economic factors, enabling a more comprehensive analysis of adversarial training's sustainability and cost-effectiveness.

METHODOLOGY

Our methodology to establish a framework for quantifying the environmental and economical affects of carbon emissions in adversarial ML involves enhancing the RCTI to cover a broader range of conditions and introducing a new metric, the CRC. These metrics involve measuring carbon emissions and the associated costs among a baseline model and adversarial trained model, parameterized by ϵ , to achieve

Table 1. Relationships among ΔC , $\Delta \mathcal{R}$, RCTI, Elasticity of Robustness, and CRC

ΔC	$\Delta \mathcal{R}$	RCTI	Elasticity of Robustness	CRC	Interpretations
+	+	> 1	Eco-Costly Robust	High	Significant environmental impact for each unit of robustness gained yielding very high CRC
+	+	$= 1$	Eco-Neutral Robust	Moderate	Shows a balanced trade-off between carbon emissions and robustness gain leading to moderate CRC
+	+	< 1	Eco-Efficient Robust	Low	Suggests improved robustness where the environmental cost per unit of robustness is lower, resulting in a lower CRC
+	-	< 0	Unrobust	High	A decrease in robustness despite increased emissions; leading to high CRC
-	+	< 0	Eco-Ideal Robust	Ideal	A decrease in emissions with an increase in robustness, which is an ideal scenario with ideal CRC
-	-	$= 1$	Balanced Unrobust	Low	A decrease of emission with decreased accuracy which is not desirable.
-	-	> 1	Unrobust	High	The robustness decreases more rapidly than the emissions causing high CRC
0	+	$= 0$	Eco-Ideal Robust	Zero	No change in emissions with an increase in robustness, showing no environmental cost for robustness gain consequently zero CRC
0	-	$= 0$	Unrobust	Zero	No change in emissions but decrease in robustness yielding high CRC
+	0	∞	Baseline Robust	High	Increase in emission without any improvement in robustness compared to the baseline model; high value for CRC as it is ∞ with higher emissions cost
-	0	∞	Eco-Ideal Baseline Robust	Ideal	Decrease in emissions without any improvement in robustness compared to the baseline model; ideal CRC as it is ∞ with lower emissions cost
0	0	∞	Eco-Neutral Baseline Robust	Zero	No changes in either parameter; no CRC

+: Increase, -: Decrease, 0: No change

robustness its in training. to highlight the trade-off between robustness against adversarial attacks and environmental impacts. In our previous work [14], we defined RCTI metrics that quantifies the carbon efficiency or inefficiency incurred for each unit increase in robustness. This metric involves the relative change in carbon emissions in the robust model compared to the baseline model, defined as ΔC . In parallel, it involves the change in model performance, or robustness, due to these enhancements compared to the baseline model's performance, which can be measured by the relative change in robustness $\Delta \mathcal{R}$.

On the other hand, the Social Cost of Carbon (SCC) is expressed in terms of the cost per ton of CO_2 emitted expressed in the CO_2eq . This measures a standardized way of measuring the impact of different greenhouse gases based on their global warming potential. The CO_2eq is calculated by the amount of CO_2 released in terms of grams in the environment times the quantity of electricity used during some computational procedure that measured CO_2eq emitted per kilowatt-hour of electricity. This metric allows us to aggregate and compare emissions from various sources effectively. Our new metric, CRC, combines the ratio of carbon emission change per unit to per unit change of performance (accuracy) due to adversarial training multiplied by the value of SCC. This metric provides a comprehensive measure of the impacts associated with robust ML models. To apply this metric, we first calculate the carbon emissions after implementing adversarial training to enhance robustness (C) and the

emissions under baseline training (C_{base}) and take the difference between these emissions. In similar fashion, we can measure performance (accuracy) achieved by adversarial training (P) and the baseline performance (P_{base}) and calculate the difference between them. The CRC is then calculated by the ratio of these two differences – the change in carbon emissions to the change in performance times the value of the SCC.

$$\text{CRC} = \frac{C - C_{base}}{P - P_{base}} \times \text{SCC} \quad (1)$$

The CRC provides an initial projection of the robustness enhancements in terms of their environmental impact. Furthermore, the CRC not only quantifies the financial costs associated with the projected increase in emissions, but also provides a clear perspective on the economic implications of each proposed robustness enhancement. The implementation and review phase of the ML project involves continuous monitoring of the impacts of robustness and emissions following the deployment of the model. This monitoring is crucial as it allows for real-time tracking of the actual environmental and economic impacts. If the impacts significantly deviate from the initial estimates, adjustments to the robustness strategies may be necessary. In examining the correlation between the Elasticity of Robustness and the CRC within our framework presented in Table 1, we observe a subtle relationship that aligns the environmental impact with economic costs. For instance, when the Elasticity of Robustness is classified as “Eco-Costly Robust,” the corresponding CRC is typically very high. This classification indicates that significant environmental impacts result

from each unit of robustness gained, leading to steep economic costs associated with achieving higher robustness levels. Similarly, scenarios labeled as “Eco-Efficient Robust” show low CRC values, reflecting less severe environmental costs compared to the Eco-Costly category. On the other hand, when robustness enhancements yield a more balanced trade-off between carbon emissions and robustness improvements, classified under “Eco-Neutral Robust,” the CRC tends to be moderate. This reflects a more sustainable approach where the costs and environmental impacts are kept in equilibrium. At the best scenario, the “Eco-Ideal Robust”, where emissions decrease or remain constant while robustness increases, typically corresponds to an ideal or zero CRC. This scenario represents the most desirable outcome, minimizing environmental impact while maximizing the robustness of the model without incurring significant environmental costs. Next, we explore the proposed mechanism in various edge cases and discuss how these scenarios can be managed based on the trade-offs reflected in our proposed metrics.

The first case occurs when ΔC is +, ΔR is +, and RCTI is slightly less than 1 (e.g., 0.99). In such a case, we level the model as “Eco-Efficient Robust”. This scenario is likely to have a low CRC because it minimizes the environmental cost per unit of gained robustness. In such a case, adversarial training of the ML model will increase efficiency with no environmental cost. In a different scenario, when ΔC is +, ΔR is + and RCTI is slightly greater than 1 (e.g., 1.01) and the CRC is high, we will label the model as “Eco-Costly Robust”. A high CRC signifies that the model is generating more carbon emissions with adversarial robustness and is thus not ideal. Another important case is when $\Delta C \approx 0^+$ (slightly greater than 0), ΔR is +, and RCTI is low (close to 0 but positive). We will label these models as “Eco-Ideal Robust” or “Unrobust” based on the increase in robustness. The increased robustness can make the model ideal, while the decreased robustness can make it unrobust.

Analysis of all the possible edge cases reveals that the precision with which we calibrate the values of ΔC , ΔR and RCTI plays a crucial role in determining the classification, and consequently, the trade-offs involved in enhancing model robustness. This sensitivity of precision highlights the importance of the fine-tuning model parameters with a degree of accuracy. Even small improvements in precision of measuring carbon emissions, robustness, or RCTI, can have a significant impact of the overall assessment

of model’s sustainability. Furthermore, the trade-off between robustness and environmental cost is highly context-dependent. In critical applications where reliability and security of the model are paramount, a slightly higher cost might be acceptable. On the other hand, if sustainability is a priority, the goal would be to minimize the RCTI value. This would involve optimizing the model to achieve robustness with minimal environmental impact. Also, the precision and calibration of these metrics highly depended on the specific goals and constraints of the application. For instance, in resource-constrained environments, where both computational resources and environmental impact are limited, the focus should be on achieving the best possible robustness with the lowest RCTI.

The takeaway here is that the ability to precisely measure and calibrate these values determines the effectiveness of the trade-offs between robustness and cost associated with the carbon emissions. Accurate calibration ensures that the balance between enhancing model robustness and minimizing environmental impact is optimized for a more sustainable and efficient outcomes.

EXPERIMENTAL SETUP

Our experimental setup utilizes the MNIST dataset, a collection of 28x28 pixel grayscale images of handwritten digits (0-9). We deploy a deep neural network with Tensorflow and use epsilon (ϵ) to represent the perturbation level in adversarial training. The ϵ values used are 0, 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9. In adversarial training, ϵ defines the maximum allowable perturbation added to input data to create adversarial examples. $\epsilon = 0$ represents no perturbation, serving as the baseline. Smaller increments at the beginning (e.g., 0.01, 0.1) help capture subtle changes in model behavior and robustness with minimal perturbations. Larger increments towards the higher end (e.g., 0.8, 0.9) allow for exploration of the model’s limits and the point at which perturbations significantly degrade performance or exceed acceptable thresholds. In the attack phase, we generate adversarial samples with $\epsilon = 0.49$ to test the robustness of the models trained during the training phase. This helps evaluate the effectiveness of adversarial training in enhancing model robustness against strong adversarial attacks. As ϵ increases, the generation and training on adversarial samples require more computational power, leading to higher energy consumption and increased carbon emissions, which

Model	ϵ	Energy	Emission	Accuracy	ΔC	$\Delta \mathcal{R}$	RCTI	CRC
Training Phase								
Baseline	0	0.0055	0.0016	0.9916	0	0	0	0
Adversarial _{0.01}	0.01	0.0178	0.0051	0.9905	2.1875	-0.0011	-1971.9	-0.6705
Adversarial _{0.1}	0.1	0.0179	0.0051	0.9928	2.1875	0.0012	1807.6	0.6146
Adversarial _{0.2}	0.2	0.0179	0.0051	0.9918	2.1875	0.0002	10845.6	3.6875
Adversarial _{0.3}	0.3	0.0182	0.0052	0.9910	2.2500	-0.0006	-3718.5	-1.2643
Adversarial _{0.4}	0.4	0.0179	0.0051	0.9879	2.1875	-0.0037	-586.25	-0.1993
Adversarial _{0.5}	0.5	0.0178	0.0051	0.9885	2.1875	-0.0031	-699.7	-0.2379
Adversarial _{0.6}	0.6	0.0181	0.0052	0.9895	2.2500	-0.0021	-1062.4	-0.3612
Adversarial _{0.7}	0.7	0.0180	0.0051	0.9897	2.1875	-0.0019	-1141.6	-0.3882
Adversarial _{0.8}	0.8	0.0179	0.0051	0.9877	2.1875	-0.0039	-556.2	-0.1891
Adversarial _{0.9}	0.9	0.0177	0.0051	0.9890	2.1875	-0.0026	-834.3	-0.2837
Attack Phase								
Baseline	0.49	5.35E-05	7.43E-06	0.2791	0	0	0	0
Adversarial _{0.01}	0.49	8.35E-05	1.16E-05	0.3126	0.5590	0.1201	4.6600	1.58E-03
Adversarial _{0.1}	0.49	4.49E-05	6.23E-06	0.3751	-0.1620	0.3440	-0.4700	-1.60E-04
Adversarial _{0.2}	0.49	4.82E-05	6.69E-06	0.4185	-0.0991	0.4995	-0.2000	-6.80E-05
Adversarial _{0.3}	0.49	3.17E-05	4.40E-06	0.4747	-0.4070	0.7009	-0.5810	-1.98E-04
Adversarial _{0.4}	0.49	5.37E-05	7.46E-06	0.639	0.0038	1.2896	0.0030	1.00E-06
Adversarial _{0.5}	0.49	5.39E-05	7.48E-06	0.9026	0.0067	2.2340	0.0030	1.01E-06
Adversarial _{0.6}	0.49	5.36E-05	7.43E-06	0.8756	0.00051	2.1373	0.0002	8.04E-08
Adversarial _{0.7}	0.49	5.38E-05	7.46E-06	0.8544	0.0040	2.0613	0.0019	6.48E-07
Adversarial _{0.8}	0.49	5.58E-05	7.74E-06	0.7166	0.0414	1.5676	0.0264	8.99E-06
Adversarial _{0.9}	0.49	3.16E-05	4.38E-06	0.7294	-0.4110	1.6134	-0.2550	-8.66E-05

ϵ = Perturbation, Energy = kWh, Emission = gCO₂eq, where SCC is 340 US Dollar (USD) considering discount rate is 1.5% [15]

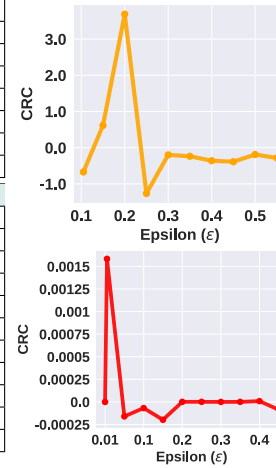


Figure 1. The table shows the results of an analysis of SCC for Adversarial Robustness of ML on the MNIST dataset (left) and Adversarial Training of the MNIST dataset based on ϵ and CRC on the training (top right) and attack (bottom right) phases.

we monitor using Codecarbon [16]. Codecarbon, a popular library that integrates with Python code to measure the carbon emissions, energy usage, location, and other metadata. For adversarial attack simulation, we use the FSGM attack method to test the robustness of the models. For the simulation, we use Google's Colab, which provides a Linux OS with two CPUs of Intel (R) Xeon (R) 2.20GHz, 12.67 GB of RAM, and no GPU. From the result, we generate RCTI, and to evaluate the monetary impact on monetary value, we calculate CRC.

RESULT ANALYSIS

Our analysis of the MNIST dataset has been summarized in the table described in figure 1, where we take two phases of the DNN models: the training phase and the attack phase. Both phases utilize two types of models: baseline, without adversarial training, and robust, which undergoes training with adversarial samples. When we train our DNN model with an adversarial sample, we observe notable changes in the emission. However, when adversarial attacks occur, a robust model can cause almost the same amount of emission in inference. But, for adversarial robustness, there is additional computation overhead, which may be subject to robustness parameter (ϵ). To validate our hypothesis, we train our adversarial model using targeted evasion attacks, in our case, FSGM. Adversarial training equips the model to sustain evasion attacks. For example, a higher number of ϵ can make the

attack so visible that common people can identify and report it, whereas a lower number of this parameter can make the attack so imperceptible that it is hard to detect with the naked eye. Robustness training of DNN can minimize the number of tragedies, thereby preventing such attacks in mission-critical areas like connected autonomous vehicles. However, greater robustness comes with more computation, which may affect emissions. Thus, it is necessary to analyze each of the perturbation emissions and perform a holistic analysis. So, we focus on the training and attack phases of the ML models. The figure 1 shows a correlation between ϵ and CRC of the MNIST dataset with selected the perturbation parameter. In the CRC value a lower value means lower emissions cost compared to a higher one, thus being more desirable. For example, in attack phase, Adversarial_{0.9} has lower CRC value than Adversarial_{0.8} which is why Adversarial_{0.9} is more environmental cost friendly adversarial training option.

CONCLUSION AND FUTURE WORKS

As the demand for intelligent systems in consumer electronics continues to grow, it is imperative to prioritize the development of adversarial robust ML models. However, we must balance these endeavors with considerations for economic, social, and environmental sustainability. While adversarial training techniques enhance model robustness against adversarial attacks, the extensive computational demands associated with this process lead to significant carbon emissions, con-

tributing to the complex and pressing challenge of climate change. Our work highlights the importance of considering the triple interplay of robustness, carbon emissions, and monetary costs in the design of robust ML, promoting a balanced perspective that accounts for both security and sustainability concerns in consumer electronics. The introduction of CRC and the extension of RCTI applicability provide a comprehensive framework for evaluating sustainability, which allows stakeholders to balance the benefits of increased robustness against the associated carbon emissions and financial costs. This holistic approach ensures that the advancement of adversarial robustness does not come at the expense of environmental and economic well-being, promoting a more balanced and responsible development of machine learning technologies. For future work, we will explore the feasibility of applying these metrics to quantify carbon emissions and their associated monetary costs in adversarial federated learning and other distributed learning architectures. Additionally, we plan to investigate how these metrics can be integrated into model compression techniques to better understand and optimize the trade-offs between robustness and carbon emissions.

ACKNOWLEDGMENTS

This research was supported by NSF grant IIS-2348145. The opinions, findings, and conclusions expressed here do not necessarily reflect NSF views.

REFERENCES

1. F. et al., "Adversarial attacks on medical machine learning," *Science*, vol. 363, no. 6433, pp. 1287–1289, 2019.
2. S. et al., "Opportunities and challenges in deep learning adversarial robustness: A survey," 2020.
3. D. P. et al., "Carbon emissions and large neural network training," 2021.
4. E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in nlp," 2019.
5. Z. E. et al., "Art and the science of generative ai," *Science*, vol. 380, no. 6650, pp. 1110–1111, 2023.
6. W. et al., "Sustainable ai: Environmental implications, challenges and opportunities," in *Proceedings of Machine Learning and Systems*, D. Marculescu, Y. Chi, and C. Wu, Eds., vol. 4, 2022, pp. 795–813.
7. S. et al., "Green ai," *Communications of the ACM*, vol. 63, no. 12, pp. 54–63, 2020.
8. R. Verdecchia, J. Sallou, and L. Cruz, "A systematic review of green ai," 2023.
9. H. et al., "Towards the systematic reporting of the energy and carbon footprints of machine learning," *Journal of Machine Learning Research*, vol. 21, no. 248, pp. 1–43, 2020.
10. J. e. a. Dodge, "Measuring the carbon intensity of ai in cloud instances," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 1877–1894.
11. K. et al., "Aligning artificial intelligence with climate change mitigation," *Nature Climate Change*, vol. 12, no. 6, pp. 518–527, 2022.
12. L. et al., "Desvig: Decentralized swift vigilance against adversarial attacks in industrial artificial intelligence systems," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 5, pp. 3267–3277, 2020.
13. H. et al., "The environmental price of intelligence: Evaluating the social cost of carbon in machine learning," in *2024 IEEE Conference on Technologies for Sustainability (SusTech)*, 2024, pp. 397–403.
14. S. M. Hasan, A. R. Shahid, and A. Imteaj, "Towards sustainable secureml: Quantifying carbon footprint of adversarial machine learning," in *2024 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2024, pp. 1359–1364.
15. U. E. P. Agency, "Inventory of u.s. greenhouse gas emissions and sinks: 2023 report," U.S. Environmental Protection Agency, Tech. Rep., 2023.
16. B. C. et al., "mlco2/codecarbon: v2.4.1," May 2024.

Syed M. Hasan (syedmhamudul.hasan@siu.edu) is pursuing Ph.D in Computer Science and a member of SHIELD Lab at Southern Illinois University Carbondale. His research interests are adversarial AI, sustainability, AI fairness, and Cyber-Physical Systems.

Abdur R. Shahid (shahid@cs.siu.edu) is an assistant professor and the director of the SHIELD Lab in the School of Computing at Southern Illinois University Carbondale. He holds M.S. and Ph.D. degrees in Computer Science from Florida International University. His research focuses on secure, trustworthy, and privacy-enhanced AI solutions for Cyber-Physical Systems, emphasizing sustainability and trust.

Ahmed Imteaj (imteaj@cs.siu.edu) is an assistant professor and director of SPEED Lab in the School of Computing at Southern Illinois University, Carbondale. He received his Ph.D. in Computer Science from Florida International University. His current research focuses on theoretical and practical aspects of Federated Learning, Large-Language Models, Vision-Language Models, Cybersecurity, and AI.