Sustainable Machine Learning: A Comparative Study of Representative Subset Selection Methods

Sebastian Mair
sebastian.mair@liu.se smair.github.io

October 6, 2025

Background

Deep learning models have achieved remarkable performance across a wide range of learning tasks. However, the training of neural networks can be computationally expensive, often requiring large datasets and extensive GPU resources. This leads to high energy consumption and significant carbon emissions, raising concerns about the sustainability of machine learning practices.

A promising approach to mitigate these issues is representative subset selection, also known as coreset selection. The goal is to identify smaller subsets of the training data that preserve model performance while reducing computational cost. Several methods exist for deep learning, including CRAIG (Mirzasoleiman et al., 2020), GRAD-MATCH (Killamsetty et al., 2021a), GLISTER (Killamsetty et al., 2021b), RETRIEVE (Killamsetty et al., 2021c), GraNd (Paul et al., 2021), and CREST (Yang et al., 2023).

While prior work has demonstrated these methods' ability to reduce training time or data requirements, little attention has been paid to their impact on energy consumption and carbon emissions. A systematic evaluation of these methods from a sustainability perspective is lacking.

Objectives

1. Implement a selection of representative subset selection methods.
2. Evaluate these methods on a set of benchmark datasets for classification and regression tasks.
3. Compare the training on the selected subsets and the full data in terms of:
   (a) Predictive performance (accuracy, RMSE, etc.)
   (b) Training time and computational cost
   (c) Energy consumption and estimated carbon emissions
4. Analyze trade-offs between model performance and sustainability metrics, and identify which methods offer the best compromise.

Research questions

- Which tool is best suited to measure and track the carbon emissions of machine learning models? There are several tools available, e.g., carbontracker[1] (Anthony et al., 2020), experiment-impact-tracker[2] (Henderson et al., 2020), and codecarbon[3] (Courty et al., 2024).
- How well do representative subset selection methods preserve predictive performance on deep learning models compared to training on the full dataset?
- Which subset selection methods provide the best balance between model accuracy, training efficiency, and environmental impact?
- How sensitive are the methods to hyperparameters (e.g., number of pretraining epochs, number of epochs between subset selections) and what are the implications for the energy-consumption?

Eligibility requirements

- Solid knowledge of machine learning and deep learning (i.e., very good grades in relevant courses are required)
- Strong background in Python and deep learning frameworks (i.e., PyTorch)
- Interest in sustainable AI and measuring energy consumption

Please attach your CV and transcripts when applying.

References

Anthony, L. F. W., Kanding, B., and Selvan, R. (2020). Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. ICML Workshop on Challenges in Deploying and monitoring Machine Learning Systems. arXiv:2007.03051.

Courty, B., Schmidt, V., Luccioni, S., Goyal-Kamal, MarionCoutarel, Feld, B., Lecourt, J., LiamConnell, Saboni, A., Inimaz, supatomic, Léval, M., Blanche, L., Cruveiller, A., ouminasara, Zhao, F., Joshi, A., Bogroff, A., de Lavoreille, H., Laskaris, N., Abati, E., Blank, D., Wang, Z., Catovic, A., Alencon, M., Stęchły, M., Bauer, C., de Araújo, L. O. N., JPW, and MinervaBooks (2024). mlco2/codecarbon: v2.4.1.

[1] `https://github.com/lfwa/carbontracker/`
[2] `https://github.com/Breakend/experiment-impact-tracker`
[3] `https://github.com/mlco2/codecarbon`

Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., and Pineau, J. (2020). Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, 21(248):1–43.

Killamsetty, K., Durga, S., Ramakrishnan, G., De, A., and Iyer, R. (2021a). Grad-match: Gradient matching based data subset selection for efficient deep model training. In *International Conference on Machine Learning*, pages 5464–5474. PMLR.

Killamsetty, K., Sivasubramanian, D., Ramakrishnan, G., and Iyer, R. (2021b). Glister: Generalization based data subset selection for efficient and robust learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8110–8118.

Killamsetty, K., Zhao, X., Chen, F., and Iyer, R. (2021c). Retrieve: Coreset selection for efficient and robust semi-supervised learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 14488–14501.

Mirzasoleiman, B., Bilmes, J., and Leskovec, J. (2020). Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pages 6950–6960. PMLR.

Paul, M., Ganguli, S., and Dziugaite, G. K. (2021). Deep learning on a data diet: Finding important examples early in training. In *Advances in Neural Information Processing Systems*, volume 34, pages 20596–20607.

Yang, Y., Kang, H., and Mirzasoleiman, B. (2023). Towards sustainable learning: Coresets for data-efficient deep learning. In *International Conference on Machine Learning*, pages 39314–39330. PMLR.