Reducing Carbon Emissions in Machine Learning with Representative Subsets

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

October 2, 2024

Background

Training machine learning models on large-scale data sets is time-consuming, costly, and not sustainable. A remedy is offered by so-called *representative subsets*, see, *e.g.*, Mair and Brefeld (2019); Paul et al. (2021). The idea is to reduce the amount of training data by focusing on informative data points while aiming on still learning approximately the same model as on the full data set. Here, we are trading off accuracy (being close to the model on full data) and efficiency (using less data means faster training). Usually, efficiency is only measured in terms of computation time. In this project, we aim at exploring the potential of representative subsets to improve the sustainability of machine learning by reducing carbon emissions (Luccioni and Hernandez-Garcia, 2023).

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).
- Do representative subsets also offer a practical trade off between accuracy and sustainability? In other words, can we sufficiently approximate a model on small but informative data subsets while significantly reducing carbon emissions?

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

Eligibility requirements

- Sound knowledge of machine learning (very good grades in relevant courses)
- Very good programming skills in Python
- Knowledge of frameworks such as PyTorch/JAX is beneficial

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.

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.

Luccioni, A. S. and Hernandez-Garcia, A. (2023). Counting carbon: A survey of factors influencing the emissions of machine learning. *arXiv preprint arXiv:2302.08476*.

Mair, S. and Brefeld, U. (2019). Coresets for archetypal analysis. In *Advances in Neural Information Processing Systems*, pages 7247–7255.

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*, pages 20596–20607.