AI for sustainability should be sustainable in its own: results from project FAIR-Spoke 9

Luigi Pontieri[1,*], Pietro Sabatino[1] and Francesco Scala[1]

[1] Institute for High Performance Computing and Networking (ICAR-CNR), via P. Bucci 8/9C, 87036 Rende (CS), Italy

Abstract

Current Machine Learning technology can help address sustainability challenges in various sectors and better assess and mitigate natural risks. However, this technology has a substantial ecological footprint, and its accelerating adoption across industries risks turning it into a threat to sustainability in itself. This paper summarizes the results of research activities of project FAIR-Spoke 9 (*Green-aware AI*) that concern developing foundational methods for improving the eco-efficiency and sustainability of data-driven AI pipelines. This goal has been pursued by two research directions: reducing the amount of training data in a task-aware way; and leveraging background information. The first research direction has led to methods for selecting training data in a task-oriented adaptive way. The second line has yielded training methods exploiting users' feedback and prior models or devoted to training multi-task modular models. A high-level technical description of two representative methods of both types is offered in the paper, along with a discussion of research implications, limitations, and future research directions.

Keywords

Sustainability, Green AI, Data reduction, Informed Machine Learning, Compositional Machine Learning

1. Introduction

Current Machine Learning technology offers transformative potential for addressing sustainability challenges across disparate sectors, for better assessing and mitigating natural risks. However, this technology has a substantial ecological footprint, and its accelerating adoption across industries risks turning it into a threat to sustainability in itself. As observed in [1], most research efforts in Green ML were devoted to reducing the model size and improving hardware efficiency [2]. And yet, the sustainability of complex machine learning and deep learning tasks can be improved as well by reducing the amount of training data and using background information.

A variety of Dataset Pruning (DP) methods (e.g., based on gradient matching or bilevel-optimization) have been proposed in the literature to extract a condensed version (a.k.a. coreset) of a given dataset that retains enough information to efficiently perform a target ML task. [3, 4, 5, 6, 7] However, these methods are computationally expensive and fail to neatly overcome cheaper solutions (e.g., pure random sampling) when training a Deep Neural Network (DNN)[8].

In general, background information can reduce the risk of overfitting arising from using a reduced amount of training data, besides possibly helping prune suboptimal regions of the search space. In the literature, several attempts to infuse knowledge encoded via logical constraints in the training of ML models were made, most of which rely on extending the loss function with ad-hoc semantic-regularization terms, [9]. While these approaches may improve model quality in extreme cases where very few examples are available, there is no clear evidence that they can ensure a satisfactory trade-off between the level of compliance to constraints and training efficiency (the auxiliary loss can lead to slower convergence).

This paper summarizes the results of research activities of project FAIR-Spoke 9 ("Green-aware AI") for improving the efficiency of ML processes by both reducing the amount of training data and leveraging background information. Different data-reduction solutions have been developed for three relevant general kinds of use cases, paying suitable attention to the peculiarities of the respective data type and/or ML task: (1) Adaptive pruning of tensorized data within supervised deep learning [10, 11, 12];

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy.

[*] Corresponding author.

luigi.pontieri@icar.cnr.it (L. Pontieri); pietro.sabatino@icar.cnr.it (P. Sabatino); francesco.scala@icar.cnr.it (F. Scala)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2025-12-07

(2) Graph compression for community-focused network analysis on large graph data [13, 14, 15]; (3) Learning lightweight conditional generative models for temporal data analysis (the unstructured and high-dimensional nature of which makes classic data pruning methods unsuitable) [16, 17]. As to the exploitation of background knowledge, complementary research lines have been pursued: (1) Exploiting knowledge coming from users (via Active Learning) and/or existing models (via model adaptation) [18, 19]; (2) Using modular model architectures enabling inter-module knowledge sharing in a compositional fashion.

Decomposing a complex ML task into easier interrelated sub-tasks by training a modular model allows us to leverage the knowledge or inductive bias shared between sub-models in balancing the lack of training data. The sub-models can be complementary ML models, as in the semi-supervised deep learning approach proposed in [13] a combination of sub-symbolic and symbolic modules, in the spirit of neural-symbolic AI, as in the approach to the analysis of low-level temporal log data proposed in [20], or an ensemble of specialized predictors as in the sparse Mixture-of-Experts (MoE) proposed in [16, 21, 17]. In the latter approach, model modularity is exploited synergistically with a conditional computation strategy, to substantially reduce the computational costs and energy footprint of learning/inference processes.

A high-level technical description of two representative methods of both types is offered in the paper, in Sections 2 and 3 along with related experimental findings. Research implications and limitations, open issues and directions for future research are discussed in the concluding section.

2. Adaptive data reduction

In [12], a novel algorithm, dubbed *Play it Straight* , for efficiently learning a DNN model was proposed which exploits cheap data sampling and (Active-Learning-like) data-selection mechanisms to incrementally refine the model using a growing small subset of (informative) examples. This algorithm was shown to than existing DP-based solutions in the training of (ResNet) DNN classifiers from benchmark image datasets.

The algorithm synergistically combines an *RS* -based DNN warm-up step with an iterative AL-like scheme to efficiently refine the DNN with selections of informative data instances.

In general, AL methods rely on repeatedly choosing a subset of unlabeled data to retrain a model, till a desired accuracy level is reached or a predefined (labeling) budget is consumed. A similar iterative model refinement scheme is here extended to address the problem of efficiently training a DNN model against a large collection of labeled data, starting from a preliminary version of the model, obtained with the help of algorithm *RS* [22].

In more detail, the two-phase training approach of *Play it Straight* begins with a “boot" phase where a given (randomly initialized) DNN model is partially trained over the entire dataset by running the *RS* procedure using a low value for its reduction factor (so that a small number of model optimization steps are performed); as confirmed by our experimental analysis, this boot phase is expected to efficiently produce an informed initial setting of the DNN that allows for reliable enough importance scores to the data instances. The second, “fine-tune", phase then consists in repeatedly selecting small-sized instance subsets, based on their associated importance scores, and exploiting them to incrementally fine-tune the DNN model. This fine-tune loop ends when either the target accuracy, computed on the test set, or the maximum energy budget stated by the user has been overcome.

In this AL-like training process, a key design choice concerns the data selection strategy to be used at each fine-tune round: since the selection procedure must be applied to a potentially large amount of data, it must introduce a small computation overhead. Thus, we propose to perform each data selection step by using a combination of uncertainty-based scores (namely, least-confidence, entropy, or margin-based) with error-based scores, both of which are quite cheap to compute.

3. Learning modular models

A *Mixture of Experts* (*MoE*) is a neural net that implements both the gate and the local predictors ("experts") through the composition of smaller interconnected sub-nets. In particular, in classical (dense) MoEs [23], once provided with an input data vector x, the gate is a feed-forward sub-net that computes a vector of (softmax-ed) weights, one for each expert, estimating how competent they are in predicting x. The overall prediction $M(x)$ results from linearly combining all the experts' predictions for x according to competency weights they are assigned by the gate. In Sparse MoEs [24], this formulation is adapted to activate only a given number k of experts, associated with the top-k competency weights. In such an ensemble, the different expert sub-nets are pushed to specialize over different data sub-populations and eventually used selectively by the trainable routing logic of the "gate" sub-net.

Model The neural network model proposed in [16], named hereinafter *Green-MoE* for the sake of presentation, has a general purpose and can be used to approximate a given target variable. It can be regarded as a tuple $\mathscr{N} = \langle \mathscr{N}_g, \mathscr{N}_1, \ldots, \mathscr{N}_m \rangle$ where $\mathscr{N}_E$ is the sub-net consisting of all its experts $\mathscr{N}_1, \ldots, \mathscr{N}_m$, and $\mathscr{N}_g$ is the gate sub-net. The model as a whole encodes a function, for instance in the case of a regression problem $f : \mathbb{R}^d \to \mathbb{R}$ (this is the case considered in this exposition for the sake of simplicity). The function is defined as follows: $f(x) \triangleq \mathscr{N}_k(x)$ such that $k = \arg\max_{k \in \{1,\ldots,m\}} \mathscr{N}_g(x)[k]$ and $\mathscr{N}_g(x)[k]$ is the k-th component of the probability vector returned by the gate sub-net $\mathscr{N}_g$ when applied to x.

Thus, for any novel input instance x, a decision mechanism is applied to the output of the gate, which transforms it into an "argmax"-like weight vector where all the entries are zeroed but the one corresponding to the expert which received the highest competency score (which is turned into 1); this makes the gate implement a hard-selection mechanism. For the sake of interpretability, the following design choices are often taken: *(i)* each expert $\mathscr{N}_i$, for $i \in [1..m]$ is implemented as one-layer feed-forward nets with linear activation functions, followed by a standard sigmoid transformation: *(ii)* the gate $\mathscr{N}_g$ is implemented as a one-layer feed-forward network with linear activation functions followed by a softmax normalization layer.

Training algorithm The proposed training algorithm takes two auxiliary arguments as input: the desired number m of expert sub-nets and the maximal number $kTop \in \mathbb{N} \cup \{\text{ALL}\}$ of input features per gate/expert sub-net (where ALL means no actual upper bound is fixed).

The algorithm performs four main steps: **(1)** A *Green-MoE* instance M is created, according to the chosen number of experts m, and initialised randomly. In a variant of the training algorithm, leveraging the modular nature of a *Green-MoE*, pre-trained expert sub-models are loaded. These models can be partially or totally frozen during training of the model's gate and remaining experts to retain their internal background knowledge. **(2)** M is trained end-to-end using a batch-based SGD-like optimization procedure (using different learning rates for the gate and the experts and a variant of the loss function proposed in [23], favoring expert specialization and competency weights' skewness). Within each epoch, $\mathscr{N}_g$ employs a differentiable Gumbel-softmax layer to approximate a sparse expert selection policy in an end-to-end fashion. As a temperature parameter decreases, the Gumbel-softmax distribution approaches a hard argmax-like selection, identifying a single expert for the input data instance. Additionally, a Gumbel-softmax mechanism (with the same temperature annealing schedule as above) is used to select a specific subset of k active features (features with non-zero associated coefficients) for each *LR* sub-network (either the gate or an expert). To implement the adaptive feature selection policy, integrated with the training algorithm itself, the differentiable relaxation of the function *top-k* proposed in [25] is employed. This allows the algorithm to perform adaptive feature selection throughout the training process. **(3)** M is optimized again with the same procedure but keeping the expert sub-nets frozen, to fine-tune the gate one only. **(4)** Feature-wise parameter pruning is performed on both the gate and experts to make all these LR-like sub-nets base their predictions on $kTop$ data features at most.

The loss function utilized in the training algorithm combines an accuracy term like the one proposed in [23] (favoring expert specialization) with a regularization term summing up the absolute values of all the model parameters. The influence of this regularization term can be controlled via weighting factor λ_R.

The last step of the algorithm consists of applying an ad hoc, magnitude-based, structured parameter pruning procedure to both the gate $\mathcal{N}_g$ and all experts $\mathcal{N}_1, \ldots, \mathcal{N}_m$. In this procedure, each parameter block gathers the weights of the connections reachable from a distinct input neuron, and all the parameters that do not belong to any of the $kTop$ blocks are eventually zeroed. This corresponds to making all the sub-nets $\mathcal{N}_g$, $\mathcal{N}_1, \ldots, \mathcal{N}_m$ to only rely on $kTop$ input features.

4. Discussion and future work

Discussion Based on our analysis, *Play it Straight* emerges as an efficient method for training DNN models on large datasets. It delivers significant computational savings compared to standard training and AL approaches, without compromising model accuracy. In addition, *Play it Straight* consistently outperforms other data pruning techniques in terms of energy consumption when considering different levels for the target accuracy to achieve. The computational efficiency of *Play it Straight* makes it particularly well-suited for resource-constrained devices and aligns with the goals of Green AI, an increasingly important field in light of the climate crisis. Furthermore, its capacity to handle large datasets expands the potential applications of deep learning models, contributing to more efficient and sustainable systems.

Our approach to the discovery of a *model* has been applied to the prediction of business process outcomes in [16] and as the core of the clinical decision support framework proposed in [21]. Notably, the latter framework uses interpretable, reliable, and transparent machine learning models, i.e. logistic regressors. Te MoE-based architecture was empirically to well balance accuracy, interpretability, and efficiency, so helping better align AI solutions with clinical practice needs. Indeed, unlike typical AI solutions based on black-box models, this framework offers case-level predictions accompanied by uncertainty scores and feature-based explanations. These explanations are derived directly from interpretable logistic regression components, similar to familiar risk models used in healthcare. The framework was also shown in [21] to support the integration of prior knowledge, enabling the use of existing models or domain expertise. This facilitates trust, adaptability, and human oversight in Clinical Decision Support Systems. The interpretability and flexibility of this framework promote more transparent, safe, and accountable decision-making. This potentially increases the acceptance of ML-driven tools in real-world clinical settings.

MoE-based models are also at the core of the federated learning framework proposed in [17], which faces the challenging problem of collaboratively training a neural classifier in a federation of parties that only store different subsets of features of the data instances. The framework is designed to limit the risk of information leakage and computation/communication costs in both model training (through data sampling) and application (via conditional computation). Experiments on real data have shown the proposed approach to ensure a better balance between efficiency and model accuracy, compared to other VFL-based solutions.

Limitations, Open issues, and Future work While *Play it Straight* demonstrates promising results, it is important to acknowledge some limitations of it. First of all, the current implementation of *Play it Straight* requires manual setting of different hyperparameters, improper choices for these may undermine the energy-saving ability of *Play it Straight* . Additionally, the choice of the AL-like instance ranking function is critical, as it needs to strike a balance between energy efficiency and effectiveness in data selection. If the selected function is too computationally intensive, it may vanish part of the energy savings achieved through data reduction. To address these limitations, our future work will focus on several areas. First, we plan to investigate adaptive methods for tuning the above hyperparameters, to alleviate the burden of manual tuning and potentially improve *Play it Straight* 's performance across different scenarios. Second, we will explore other cheap data selection strategies, combined with model-training acceleration techniques, to further improve *Play it Straight* 's energy efficiency and effectiveness. Moreover, we aim to extend this approach to the training of large language models (LLMs), in the fine-tuning and alignment stages. This includes enhancing state-of-the-art techniques such as DPO [26] and SimPO

[27], for example. Given that LLMs are highly energy-intensive, our objective is to improve training efficiency—particularly where it matters most—by leveraging intelligent data selection strategies to reduce computational costs.
Both implementations in [17, 21] are limited to structured, static, tabular inputs and lack native support for multimodal or sequential data. To address this limitation, one could extend the frameworks to handle such data types by adopting more expressive neural network architectures for the experts' sub-models and developing ad hoc explanation mechanisms to generate accurate, cross-modal justifications. Furthermore, we intend to integrate domain-specific knowledge directly into our solutions through user-defined models or constraints, ensuring predictions are more aligned with established domain-specific reasoning processes.

5. Acknowledgments

This work was partially supported by research project FAIR (PE00000013), funded by the EU under the program NextGeneration EU.

6. Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT, Grammarly in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] S. Srivastava, Towards Green AI: Cost-Efficient Deep Learning using Domain Knowledge, Ph.D. thesis, The Ohio State University, 2022. Ph.D. Thesis.

[2] L. Deng, et al., Model compression and hardware acceleration for neural networks: A comprehensive survey, Proceedings of the IEEE 108 (2020) 485–532.

[3] O. Sener, S. Savarese, Active learning for convolutional neural networks: A core-set approach, arXiv preprint arXiv:1708.00489 (2017).

[4] M. Paul, S. Ganguli, G. K. Dziugaite, Deep learning on a data diet: Finding important examples early in training, in: A. Beygelzimer, Y. Dauphin, P. Liang, J. W. Vaughan (Eds.), Advances in Neural Information Processing Systems, 2021.

[5] B. Mirzasoleiman, J. Bilmes, J. Leskovec, Coresets for data-efficient training of machine learning models, in: H. D. III, A. Singh (Eds.), Proceedings of the 37th International Conference on Machine Learning, volume 119 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 6950–6960.

[6] Z. Yang, H. Yang, S. Majumder, J. Cardoso, G. Gallego, Data pruning can do more: A comprehensive data pruning approach for object re-identification, Transactions on Machine Learning Research (2024).

[7] R. Iyer, N. Khargoankar, J. Bilmes, H. Asanani, Submodular combinatorial information measures with applications in machine learning, in: V. Feldman, K. Ligett, S. Sabato (Eds.), Proceedings of the 32nd International Conference on Algorithmic Learning Theory, volume 132 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 722–754.

[8] P. Okanovic, R. Waleffe, V. Mageirakos, K. Nikolakakis, A. Karbasi, D. Kalogerias, N. M. Gürel, T. Rekatsinas, Repeated random sampling for minimizing the time-to-accuracy of learning, in: The Twelfth International Conference on Learning Representations, 2024.

[9] E. Giunchiglia, M. C. Stoian, T. Lukasiewicz, Deep learning with logical constraints, in: L. D. Raedt (Ed.), Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22, International Joint Conferences on Artificial Intelligence Organization, 2022, pp. 5478–5485. doi:10.24963/ijcai.2022/767, survey Track.

[10] F. Scala, S. Flesca, L. Pontieri, Data filtering for a sustainable model training, in: Proceedings of the 32nd Symposium of Advanced Database Systems, Villasimius, Italy, June 23rd to 26th, 2024, volume 3741 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 205–216.

[11] F. Scala, L. Pontieri, S. Flesca, DPET: A data and parameter efficient training framework for green AI (SHORT PAPER), in: R. Cantini, D. M. Longo, D. Thakur (Eds.), Proceedings of the 1st AIxIA Workshop on Green-Aware Artificial Intelligence co-located with the 23rd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2024), Bolzano, Italy, November 27-28, 2024, volume 3934 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 24–28.

[12] F. Scala, S. Flesca, L. Pontieri, Play it straight: An intelligent data pruning technique for green-ai, in: D. Pedreschi, A. Monreale, R. Guidotti, R. Pellungrini, F. Naretto (Eds.), Discovery Science, Springer Nature Switzerland, Cham, 2025, pp. 69–85.

[13] A. Socievole, C. Pizzuti, Community detection in complex networks exploiting spectral graph sparsification for efficient disaster response, in: 16th Intl Conf on Advances in Social Networks Analysis and Mining (ASONAM 2024), 2025.

[14] A. Socievole, C. Pizzuti, Effective resistance and kernel-based graph sparsification for community detection in complex networks, Soft Computing (2024).

[15] A. Socievole, C. Pizzuti, Efficient community detection in disaster networks using spectral sparsification, Pervasive and Mobile Computing - To Appear (2024).

[16] F. Folino, L. Pontieri, P. Sabatino, Sparse mixtures of shallow linear experts for interpretable and fast outcome prediction, in: Intl. Conf. on Process Mining Workshops, Revised Selected Papers, 2024, pp. 141—-152.

[17] F. Folino, G. Folino, F. S. Pisani, L. Pontieri, P. Sabatino, Efficiently approaching vertical federated learning by combining data reduction and conditional computation techniques, Journal of Big Data 11 (2024) 77. doi:10.1186/s40537-024-00933-6.

[18] F. Folino, G. Folino, M. Guarascio, L. Pontieri, Data- & compute-efficient deviance mining via active learning and fast ensembles, Journal of Intelligent Information Systems 62 (2024) 995–1019. doi:10.1007/s10844-024-00841-4.

[19] F. Folino, G. Folino, M. Guarascio, L. Pontieri, P. Zicari, Towards data- and compute-efficient fake-news detection: An approach combining active learning and pre-trained language models, SN Computer Science 5 (2024) 470. doi:10.1007/s42979-024-02809-1.

[20] B. Fazzinga, S. Flesca, F. Furfaro, L. Pontieri, F. Scala, Combining abstract argumentation and machine learning for efficiently analyzing low-level process event streams, 2025. arXiv:2505.05880.

[21] A. Cuzzocrea, F. Folino, M. Samami, L. Pontieri, P. Sabatino, Efficiently approaching vertical federated learning by combining data reduction and conditional computation techniques, Neural Computing and Applications - To Appear (2025).

[22] P. Okanovic, R. Waleffe, V. Mageirakos, K. Nikolakakis, A. Karbasi, D. Kalogerias, N. M. Gürel, T. Rekatsinas, Repeated random sampling for minimizing the time-to-accuracy of learning, in: The Twelfth International Conference on Learning Representations, 2024.

[23] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, G. E. Hinton, Adaptive mixtures of local experts, Neural Computation 3 (1991) 79–87.

[24] N. Shazeer, et al., Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, in: Proc. Intl. Conf. on Learning Representations (ICLR), 2017, pp. 1–17.

[25] S. M. Xie, S. Ermon, Reparameterizable subset sampling via continuous relaxations, in: Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI'19), 2019, p. 3919–3925.

[26] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, C. Finn, Direct preference optimization: Your language model is secretly a reward model, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems, volume 36, Curran Associates, Inc., 2023, pp. 53728–53741.

[27] Y. Meng, M. Xia, D. Chen, Simpo: Simple preference optimization with a reference-free reward, in: Advances in Neural Information Processing Systems (NeurIPS), 2024.