International Journal of Information Technology Vol. 28 No. 1 2022

Towards Green AI: Interdisciplinary Advances

Wei Yang Bryan Lim[a], Huanhuan Chen[b] and Chunyan Miao[a,c]

[a]Alibaba-NTU Joint Research Institute, Singapore
[b]University of Science and Technology of China, China
[c]Nanyang Technological University, Singapore
bryan.limwy@ntu.edu.sg, hchen@ustc.edu.cn, ascymiao@ntu.edu.sg

Abstract

Recently, AI has achieved remarkable success in several applications. However, its conventional focus on leveraging big data on large models, trained in powerful cloud servers is not environmentally sustainable. In this paper, we will discuss a multi-level framework comprising four components to achieve Green AI through a comprehensive end-to-end framework built on interdisciplinary foundations.

Keywords: Green AI, Environmental Sustainability, Interdisciplinary Research.

I. Introduction

In recent years, the effects of climate change have disrupted natural habitats and adversely impacted human societies at an alarming rate [4]. At present, the approaches adopted to improve Artificial Intelligence (AI) applications in academia and industry have been centered on training large models on big data. While this has led to significant progress in AI research, the ever-growing carbon footprints of AI model training, model parameters exchange, and model inferences have placed a significant burden on the environment. For example, the cloud computation cost to train a large Transformer model with neural architecture search may reach up to US$3 million and emits five

times as much greenhouse gases as that of the lifetime usage of a vehicle [25]. This conventional school of thought has been classified as "Red AI" [23].

Given the rapid degradation of the environment, there exists an urgent need to introduce a new paradigm known as "Green AI" that aims to make all stages of AI model development environmentally sustainable, from data collection to model training and deployment. In this paper, we review the literature to address the question: *How can the sustainable development of AI be promoted?* We introduce an end-to-end Green AI framework built on interdisciplinary foundations that will aim to greenify AI through the "Reduce, Reuse, Recycle" philosophy applied to multi-level enablers.

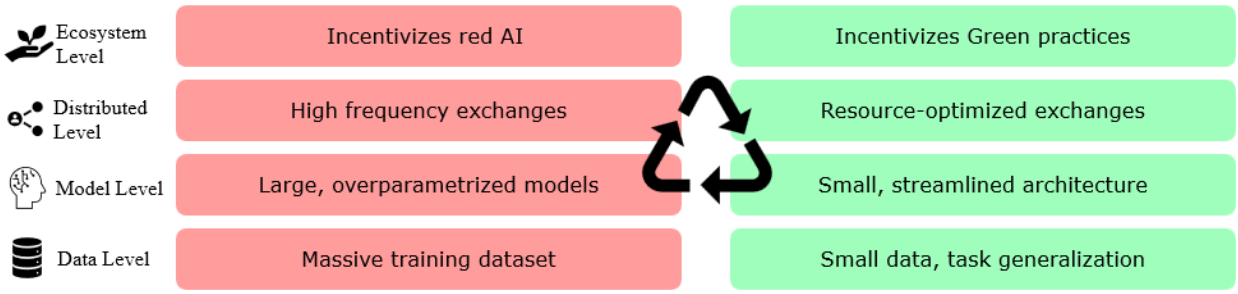

Fig 1 A multi-level Green AI framework.

II. Data Level

The training data required for the state-of-the-art AI models today can involve terabytes of text data or billions of images [5]. However, data is environmentally expensive to collect and preprocess. Moreover, the size of the training dataset is directly proportional to the environmental cost that model training imposes on the environment [23]. Despite this, the rapid growth of AI has been credited to the access to vast amounts of data and powerful computing resources. Fortunately, numerous studies have investigated how to train AI models with limited data, albeit for motivations unrelated to the environment, e.g., due to training data scarcity in the healthcare [24] and low-resource natural language processing (NLP) domains [7].

One emerging method is the increasing use of models that have been pre-trained on extensive data. These are widely known as foundation models such as BERT [9], CLIP [16], and GPT-3 [1]. Using small data or learnable prompts that are domain-specific to downstream tasks, the foundation models can be adapted through techniques such as zero-shot inference with hand-crafted prompts, prompt tuning, and fine-tuning [30]. This mitigates the need to implement training or data collection/pre-processing from scratch.

Learning paradigms to reduce the number of *annotated* training samples required can also reduce the environmental cost of sensing and pre-processing. This group of techniques, one of which is known as active learning [3, 15], seeks the input of an "oracle" (i.e., human-in-the-loop) to label influential data samples that are algorithmically singled out in iterations, rather than rely on a completely labeled dataset right from the beginning. This approach can significantly reduce the number of annotated training data required and therefore, the environmental cost of pre-processing.

To reduce the environmental impact of data collection and processing, the *reuse* of training samples is a viable option. This can be achieved through participating in data marketplaces, which are platforms for transactions to obtain related data to enrich the buyer's internal dataset. The efficient facilitation of data trading has been well-studied [22], with the literature addressing topics such as efficient pricing of data from the crowdsourcing and network economics perspective [2]. With the advent of blockchain, the necessity of a central platform has been challenged, leading to the proposal of decentralized marketplaces that leverage blockchains for peer-to-peer transactions [17]. However, with privacy concerns and the introduction of stringent privacy laws such as the General Data Protection Regulation (GDPR), the trading of data has come under scrutiny. To enable the reuse of data without direct sharing, techniques such as Federated and Split Learning [14, 12] have been proposed to exchange model parameters towards the development of a global

AI model. Through these means, the value of data is unlocked and the environmental cost incurred by an over-reliance on big data is reduced.

III. Model Level

State-of-the-art models today tend to be large to the extent of being over-parametrized [28]. For example, GPT-3 has about 175 billion parameters and requires 800GB of storage. Large models are costly to train, store, and infer. In response, early research [8, 21] has focused on designing more efficient model architectures, particularly for their implementation on mobile devices. Besides taking up less storage, streamlined models offer important benefits such as reduced environmental impact and quicker inference times, though they may result in some loss of accuracy. However, these trade-offs between accuracy and model size can be adjusted to optimize the overall performance of the model, depending on various user-centric factors.

Retraining deep learning models from scratch can be both time-consuming and resource-intensive. To address this issue, the reuse of model parameters has been explored in the field, including the use of parameter sharing [18]. This technique involves sharing the same weights/filters across multiple layers in the network, reducing the number of parameters that need to be learned and potentially leading to improved learning performance. Parameter sharing has been applied across various tasks from translation [20] and vision models [18] to reinforcement learning [26]. A related concept is multi-task learning [29] in which parameter sharing is utilized to reuse parameters for multiple separate tasks, in order to allow for task generalization.

For the recycle dimension, learning algorithms such as leveraging knowledge distillation [6] techniques to fine-tune pre-trained models instead of starting from scratch, can lead to improved task generalization and reduction of the environmental cost of model training. Moreover, the uti-

lization of foundation models for fine-tuning to accomplish downstream tasks can achieve the best of both worlds of low-resource training data requirements and lower computation costs.

IV. Distributed Computation Level

Inefficient model training in the cloud can have a significant impact on the environment due to high energy consumption. Processes such as data shuffling and server communication during training consume a large amount of energy, resulting in a significant carbon footprint. As such, energy-efficient algorithms to reduce the amount of data that has to be transferred will offset carbon emissions. Moreover, the optimized scheduling of workloads among data centers with varying carbon intensities can lead to reduced emissions as well as cost savings in the carbon market [11].

While cloud computing is still the dominant approach for model training, edge computing has emerged as a viable alternative due to growing privacy concerns [13]. Edge computing utilizes the resources of end devices and edge servers for edge caching, training, and inference, bringing the computation closer to where the data is generated. A key enabling technology of edge computing is Federated Learning (FL), in which data owners carry out model training locally before transmitting the model parameters or gradient updates, rather than the raw data, to a model owner for aggregation [14]. This enables privacy-preserving collaborative machine learning while leveraging the computation capabilities of workers. As with cloud computing, FL faces the issues of redundant parameter updates to the parameter server and surplus local computations, all of which result in carbon emissions that can be reduced. It follows that research on importance-aware updating (i.e., that selectively chooses key updates to be communicated with the server) [19], energy-aware client selection (i.e., of which protocol selects clients that emits least carbon emissions during training) [12], and the utilization of ambient energy or green sources of energy for distributed model training

can be most impactful to reduce the carbon footprint in edge AI.

V. Ecosystem Level

At present, red AI is prevalent given that the conventional AI ecosystem rewards "red" practices. Through developing ecosystem level enablers, we are able to empower green AI. Incorporating carbon footprint costs as a penalty in AI model training is a key enabler for promoting environmentally-friendly AI. Considering the tradeoff between performance and environmental cost will encourage researchers and practitioners to work towards a greener AI. This can be feasibly accomplished, given that tools to quantify the carbon emissions of AI now widely exist [10].

Applying economic principles to govern the AI domain is another promising direction for Green AI. To effectively implement the "reduce, reuse, recycle" philosophy using pre-trained models or public datasets, these models/datasets must be readily available for sharing. Yet, without incentives to share, it is unlikely that these models/datasets will be fully made public. In fact, premium/paid versions of large language models now exist. Just as cloud computing provides computations or functions as a service [27], economic tools can be utilized to model the interactions of buyers and sellers in the data market/model-as-a-service market. Similarly, reputation systems can be introduced to consolidate the opinions of participants in the ecosystem into quantifiable ratings, thereby improving the valuation of each participant's data/model resources.

VI. Lessons Learned and Conclusion

In this paper, we highlight four key levels in which Green AI can be realized. We provide a high level summary of enablers within these levels, and in the process, provide readers with insights on

the way forward towards realizing Green AI. The lessons learned from our survey can be summarized in the following:

1. *The lack of incentives to greenify AI is a pressing concern:* This paper primarily examines techniques aimed at addressing other challenges, such as utilizing small data learning paradigms to overcome data scarcity in AI. Although these techniques have the potential to decrease the carbon impact of AI, they are rarely positioned as such due to the lack of academic incentives or interests to make AI more environmentally friendly.

2. *Benchmarks and standards for Green AI have to be established:* At present, there is a widespread focus on creating leaderboards that showcase the highest model accuracies, as research groups compete to outperform current state-of-the-art results. Unfortunately, this is not the case for Green AI, where there are no established benchmarks. Although there are online resources that can calculate the carbon footprint of machine learning models, there has not been enough research conducted to determine how to achieve optimal performance while minimizing carbon impact. As a result, a set of standards to promote and guide green practices during AI development is absent.

3. *The trajectory towards large models can both benefit and hinder the effort to create environmentally sustainable AI:* The rise of large models have transformed AI. However, the trend toward using large models presents a dilemma. On one hand, even though large models have shown superior results, training these models demand considerable amounts of computing power and energy, which results in a larger carbon footprint and exacerbates climate change. On the other hand, large models (particularly foundation models) can be effectively utilized and applied to various downstream tasks, providing more energy-efficient solutions. Moving

forward, research on how ecosystem mechanisms such as green premium pricing and green AI regulations to encourage green practices will be most beneficial.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[2] Lingjiao Chen, Paraschos Koutris, and Arun Kumar. Towards model-based pricing for machine learning in a data marketplace. In *Proceedings of the 2019 International Conference on Management of Data*, pages 1535–1552, 2019.

[3] David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145, 1996.

[4] Anthony Costello, Mustafa Abbas, Adriana Allen, Sarah Ball, Sarah Bell, Richard Bellamy, Sharon Friel, Nora Groce, Anne Johnson, Maria Kett, et al. Managing the health effects of climate change: lancet and university college london institute for global health commission. *The lancet*, 373(9676):1693–1733, 2009.

[5] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems*, 34:3965–3977, 2021.

[6] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819, 2021.

[7] Michael A Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. A

survey on recent approaches for natural language processing in low-resource scenarios. *arXiv preprint arXiv:2010.12309*, 2020.

[8] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[9] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[10] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019.

[11] Trung Le and David Wright. Scheduling workloads in a network of datacentres to reduce electricity cost and carbon footprint. *Sustainable Computing: Informatics and Systems*, 5:31–40, 2015.

[12] Wei Yang Bryan Lim, Nguyen Cong Luong, Dinh Thai Hoang, Yutao Jiao, Ying-Chang Liang, Qiang Yang, Dusit Niyato, and Chunyan Miao. Federated learning in mobile edge networks: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 22(3):2031–2063, 2020.

[13] Yuyi Mao, Changsheng You, Jun Zhang, Kaibin Huang, and Khaled B Letaief. A survey on mobile edge computing: The communication perspective. *IEEE communications surveys & tutorials*, 19(4):2322–2358, 2017.

[14] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[15] Soham Pal, Yash Gupta, Aditya Shukla, Aditya Kanade, Shirish Shevade, and Vinod Ganapathy. Activethief: Model extraction using active learning and unannotated public data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 865–872, 2020.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[17] Gowri Sankar Ramachandran, Rahul Radhakrishnan, and Bhaskar Krishnamachari. Towards a decentralized data marketplace for smart cities. In *2018 IEEE International Smart Cities Conference (ISC2)*, pages 1–8. IEEE, 2018.

[18] Siamak Ravanbakhsh, Jeff Schneider, and Barnabas Poczos. Equivariance through parameter-sharing. In *International conference on machine learning*, pages 2892–2901. PMLR, 2017.

[19] Jinke Ren, Yinghui He, Dingzhu Wen, Guanding Yu, Kaibin Huang, and Dongning Guo. Scheduling for cellular federated edge learning with importance and channel awareness. *IEEE Transactions on Wireless Communications*, 19(11):7690–7703, 2020.

[20] Devendra Sachan and Graham Neubig. Parameter sharing methods for multilingual self-attentional translation models. In *Conference on Machine Translation*, 2018.

[21] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[22] Fabian Schomm, Florian Stahl, and Gottfried Vossen. Marketplaces for data: an initial survey. *ACM SIGMOD Record*, 42(1):15–26, 2013.

[23] Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. Green ai. *Communications of the ACM*, 63(12):54–63, 2020.

[24] Torgyn Shaikhina and Natalia A Khovanova. Handling limited datasets with neural networks in medical applications: A small-data approach. *Artificial intelligence in medicine*, 75:51–63, 2017.

[25] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*, 2019.

[26] Justin K Terry, Nathaniel Grammel, Ananth Hari, Luis Santos, and Benjamin Black. Revisiting parameter sharing in multi-agent deep reinforcement learning. *arXiv preprint arXiv:2005.13625*, 2020.

[27] Quyuan Wang, Songtao Guo, Jiadi Liu, Chengsheng Pan, and Li Yang. Profit maximization incentive mechanism for resource providers in mobile edge computing. *IEEE Transactions on Services Computing*, 15(1):138–149, 2019.

[28] Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3635–3673. PMLR, 2020.

[29] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609, 2021.

[30] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.