International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 OPEN ACCESS
Available Online at : www.ijsrcseit.com
doi : https://doi.org/10.32628/CSEIT2410758

Survey on Green AI with the Help of Data-Centric Methods

Prisha Thakur, Mrs. Deepali Gohil, Dr. Dipalee Rane
Department of Computer Engineering, D. Y. Patil College of Engineering Akurdi, Pune, Maharashtra, India

ARTICLEINFO

Article History:

Accepted : 20 Nov 2024
Published: 12 Dec 2024

Publication Issue
Volume 10, Issue 8
November-December-2024

Page Number
61-68

ABSTRACT

The domain of Artificial Intelligence (AI) continues to extend and the environmental impact of training large-scale deep learning models have increased. The term "Red AI" highlights the high computational demands and energy consumption associated with these models. In response, "Green AI" advocates for more sustainable approaches, aiming to balance AI progress with environmental responsibility. This paper explores the change from a model centric to a data centric AI paradigm, emphasizing the critical role of data quality and efficient dataset usage. Specifically, we examine dataset distillation techniques, which downsize training datasets while preserving important information, enabling more efficient and sustainable learning. By leveraging data-centric methods, deep learning models can achieve similar or better performance with fewer computational resources, contributing to the broader goals of Green AI. This shift is essential for advancing AI in a way that minimizes its carbon footprint while maintaining its potential for innovation.

Keywords—Artificial Intelligence (AI), Model-Centric AI (MCAI), Data-Centric Artificial Intelligence (DCAI), Green AI, Deep Learning (DL), Dataset Distillation (DL)

Introduction

Artificial Intelligence (AI) refers to intelligent machines that are designed to think, learn, and perform tasks that usually need human intelligence. These tasks include solving problems, understanding language, pattern recognition , and making decisions according to different situations. "Red AI" is coined to signify artificial intelligence (AI) models that go through training with the help of resource-intensive techniques on huge datasets. This method is harmful as it has generous energy consumption and emissions of carbon. The Green AI models are made in such a way that they have resembling efficiency and minimal environmental impact. With the help of methods such as less computationally intensive training methodologies, optimal usage of smaller datasets, or sustainable energy resources. While Red AI focuses more on precision and performance, Green AI gives more stress on efficiency and sustainability [3]. The increased use of computational powers over the years has made impressive developments in the field of artificial intelligence (AI). The primary way to

do so is to focus on it expansion in size of deep learning models to an immense scale and then is to train them on enormous data quantities. Although impact in the AI field has moved from mainly focus on achieving quantifiable results to precision, which comes at the cost of energy efficiency. Neglecting the power and environmental expense of the enormous amounts of data erasing and deploying these models is the cost of in the obsession with quantifiable outcomes. Since the call for even more advanced ML models will increase, it is increasingly imperative then to continue ignoring this method and look for ways to combat this technology by being more self-sustained. Not able to achieve this can lead to severe environmental problems, and as a result would constrain the long-term success and impact of ML [1]. Every 3-4 months the amount of computing being used to train AI models have been doubling [2].

As a result different methods are used to reduce the computational costs of AI algorithms at multiple steps in an AI system life cycle including training, inference phase, architecture design and data use[1]. Data is the most powerful and effective technology that can immensely cut down computational expenses. Selecting a subset of data, cleaning the chosen subset to remove redundancy in observed samples without overfitting help drive down data size and computation cost as envisioned for more affordable AI technology. Thus, major benefits can be bought by giving attention to efficiently using data and this should not be ignored [1]. A pivotal shift in this movement is the change from a model centric to data centric methods in AI research and development. Traditional AI efforts have primarily focused on refining model architectures and scaling their size, often at the expense of computational efficiency. Instead, the data-centric methods emphasizes the value of high-quality, well-curated datasets, enabling enhanced model performance without requiring larger models or continuous training cycles. Techniques such as dataset distillation—condensing large datasets into smaller, more informative subsets—are central to this shift, allowing for more efficient learning while reducing resource usage. This paper explores the evolving landscape of AI, with a focus on the transition toward data-centric methodologies and the role of dataset optimization in advancement of Green AI. We study how deep learning models can benefit from these techniques, reducing their environmental impact and contributing to more sustainable AI development practices.

Model Centric ML

Fig.1. Work flow in model-centric models [5] Firstly we get knowledge from the data and then we make use of that knowledge within a model this technique helps a lot in the advancement of AI systems based on machine learning (ML). Model-centric models usually use this approach. In model-centric AI (MCAI),the data is kept constant after its collection and preparation and then we focus on improving the model. We make an assumption that if we have more data than the model will perform better, this leads to the only focus being collection of as much data as possible. This data is studied in dept and then given as reference for shallow and deep learning [7].The Fig 1. Shows the work flow of Model-centric models. The model-centric approach focuses on finding the best model with a fixed set of data, and improving model accuracy is achieved through hyperparameter tuning. However, data can contain features that do not improve precision, resulting in overfitting, and erroneous data that is hard to detect in larger datasets and cannot be fixed by hyperparameter tuning. To compensate for these weaknesses, the most common approach is to collect more data. Model-centric AI mainly focuses on optimizing model architecture and hyperparameters, with data created almost only one time and is maintained the same throughout the AI system's lifecycle. However, this approach has been under considerable strain because of its vulnerability to adversarial samples, a narrow scope of business applicability, and low generalization capacity [8].

A. Limitations

So, with a model-centric AI there is an assumption that ML methods cater to building better models by concentrating much on optimizing the algorithm and underlying hyper-parameters. This means that you generate the data one time and save any generated data even as development of AI system progresses. Model-centric AI has been the hot trend for several years — but it also had its limitations. It really shines in organizations and industries that have consumer platforms with millions of users perhaps more than happy to rely on common solutions. Although this setup would be adequate for most consumers in an ideal world, the outliers are rendered virtually useless. For example, in industries when we talk to like advertising companies using lots of data (often with standard formats), Google/Baidu/Amazon/Facebook could have enough scale already just on the model side. However, as is the case in most industries such as manufacturing and agriculture for example or even healthcare services require tailored solutions significantly outperform one-size-fits-all approaches so these organizations need to strategize on how their algorithms can learn from data that is more extensive capturing all important cases but still consistently labelled.[7].

Data Centric ML

Data centric AI ensures the ability to replay repeat evaluations as new data arrives. Such balanced solutions of model-centric and data- centric blends can offer the best performance in our AI system. Model-centric is about building the best possible algorithms. Data centric metrics concerns with how well data trains good models (relevant and clean etc..) The data-centric methods are devoted to the naming of completeness, relevance, consistency and heterogeneity issues. It is a repetitive process that needs to be homogenized with the model-centric workflow to build a code and data efficient AI system. These models can be for data classification, clustering to recommendation systems and natural language processing. AI algorithms like deep learning, reinforcement learning and decision trees can process large number of data set and it takes the decisions or predictions by using that processed dataset. AI usage in data-centric models provides the ability to discover hidden patterns, relationships and understand this with humans so organization takes better decisions based on data. By integrating AI into data-centric models, various fields like finance, retail healthcare stand a chance to adapt revolutionary changes [5].

Data-centric AI is an engineering method that aims to boost the performance of a machine learning system by methodically maximizing the quality of each individual piece of data in training the model. Unavailability of a large enough data of all types and in huge amount. Compared to manufacturing industries, which may be relatively data-poor by comparison of Goole or Baidu. A model that is trained to spot problems by methods suitable for hundreds of millions of data points simply would break down if it has no enough proper parade examples. The most effective solutions should also be more personalized, especially for a manufacturing business which in turn involves many different products. If you had to use an AI system on a one-size-fits-all basis it probably would not be efficient in detecting product quality issues across all products as every product would require its own personalized trained ML system. [6]. Through comprehensive yet different techniques and categories we can enhance the working of the AI system by systematically enhancing the quality of the data used to train the model. AI that is data-centric uses advanced data engineering methods to do this. Data-centric AI can improve the performance of AI models and services through augmentation, extrapolation, and interpolation. AI services can be made more accurate and dependable with the help of data-centric AI, which involves expanding the accessible data and enabling more optimal use of it. This innovative methodology creates data-centric AI by utilizing training data from various sources such as synthetic data, public datasets, and private datasets.

Data-centric AI also help to improve the accuracy and reliability of AI services by increasing both data that can be used by them as well how effectively they are able to harness it. This methodology is the culmination of training data from various sources ranging over synthetic data, public datasets to private datasets which when combined, produces this first-of-its-kind Data-Centric AI. Adopting this strategy can help reduce the effort and time involved in generating training data, thus contributing immensely to boosting your data quality as well. This is able to improve the efficiency of utilizing AI services that perform computations using training data as well. In addition, data centric AI can handle more types of datasets as the data is tailored. This means that no matter how large the dataset, data-centric AI can analyse and work with it to get satisfactory predictions [5].

Right now, what the attention that DCAI is getting to shows how vital of a role data plays when it comes to constructing intelligent systems. Data quality and reliability in model development because of the realization that AI gating factors become more about "garbage-in (data), garbage-out" DCAI emphasizes consideration for increased quality, recognizing data used to train an AI system is a key determinant factor when factoring accuracy or performance. The underlying principle of DCAI is the assumption that producing good intelligence demands both, high quality consistent data indexed by unique identifiers and trust-worthiness maintained through detailed record keeping. Through this emphasis on data quality, DCAI is enabling new innovation in data management and planning technology that will ensure AI systems make right decisions with confidence [1]. To socially enable the research team will need to spend time and effort on quality of the data so that AI systems can learn from small quantities which are typical for most industries. For example, this means that we want to make the data collection in a way where it shows neatly what kind of content should be learned by AI. AI experts, felt that data engineering should be done by the ones using AI. This will serve to make AI more convenient and therefore used by any industry making use of it. What it demonstrates is the importance of selecting high quality training data for algorithms. We also see a paradigm shift from model centric to data centric approach.

Data Centric Frameworks

A. Deep Learning

Deep learning plays a major role in performing many tasks, but it requires a lot of effort. In addition to hug e financial costs, it also produces more carbon, caus ing more climate change. Developers face many decisions that affect the result of their deep learning solutions, including options in the framework, optimization, architecture, batch size, and job schedule that affect the performance and energy footprint of the model [13]. Deep learning has worked well on problems, such as image classification/ face detection / text categorization/video recognition. Instead, deep learning models learn on the training data for an extended period of time and use hidden features which represent dense vectors to make better predictions for these tasks This necessitates a great deal of accounting and resources. In recent years deep learning has come into its own, using computational resources to replace the explicit knowledge in cross-reference books that told us how best to handle very large amounts of data with cumbrous techniques like linear approximations. The GPT-3 language model, with 175 billion parameters relies on the learning of thousands of Graphics processing units (GPUs) and has learned up to ˜45 terabytes worth of text [10]. However, every day an unmanageable amount of information is produced which puts a very serious threat to the education process and there are increasing problems with storing all paper volume of information as well incompatibility between numbers. Dataset distillation (DD) is born precisely to cope with this huge data volume issue. Recently, one of the

approaches to address big data is Dataset-Distillation (DD) [9]. Deep Learning Research has been growing at an exponential rate, especially those using deep learning in Computer Vision. One main cause of the suboptimal performance by a deep learning models is likely to be because you have used terrible training and testing data. Some of the popular databases for image classification are: MNIST CIFA ImageNet Tiny Net This data is a huge dataset to use for deep learning model deployment. But deep learning architectures and model hyperparameters make it fairly expensive to compute. Thus, a technique called data distillation is proposed to solve this problem in deep learning models [10].

B. Dataset Distillation

Deep learning has experienced unprecedented growth over the past decade, becoming the go-to choice for many applications. This progress is mostly driven by collaboration, and the rapidly increasing use of computer hardware allows algorithms to process large amounts of data. However, managing the ever-increasing amount of data with limited power consumption is becoming increasingly difficult. Different methods are introduced to enhance the efficiency of processing data. Dataset distillation is a technique used in deep learning to break large datasets into smaller, more efficient pieces while preserving the important data needed to inform the model. The idea is to create a compact, optimized dataset that allows the model to learn from the entire original dataset. It is a data compression technology that can combine some high-level data points to record real data. A base model trained on small mixed data can achieve similar performance to a model trained on real data [9]. The Fig 2. Shows the illustration of dataset distillation. This process is especially useful when dealing with resource constraints like limited memory or computational power, as it can reduce the amount of data required to train a model while maintaining performance. It can also speed up training by minimizing the time needed to process large datasets. Dataset distillation typically involves using optimization techniques to extract key patterns or representative samples from the original dataset. These distilled datasets can then be used for tasks such as model training, fine-tuning, or transfer learning.

Fig.2.Dataset distillation example. Training examples on large datasets and small synthetic datasets show si milar performance on test datasets.[9]

Dataset distillation is a data reduction technique that solves this problem by merging small datasets from large datasets and has gained a lot of attention. Current dataset distillation techniques can be divided based on if they are consistent with the performance of required dataset or not. It can be divided into meta-learning and dataset matching [9]. Dataset distillation is an exciting idea and has the potential to solve this challenge, which has become a popular research topic in recent years. Data Distillation eclipsed the coreset selection method for dataset reduction. We could do this by using a small coreset of prototypes chosen from the original training set and then only train on that to save processing time This would in theory limit how much performance drops. However, the coreset is not elements-modifiable and they are real-data-constrained due to which it restricts its expression drastically especially under bounded budgets. The removal of limitation of uneditable elements and precise modification of a small number of examples can be done by Dataset distillation. This is done to conserve more information rather than how things work in coreset selection.

After summarizing the data from an originally large dataset via a small synthetic one, models trained on such distilled datasets in general are able to achieve better performance compared to uneditable coresets. It is challenging to distil these into a handful of points for high dimensional data, primarily since the nature of true high dimension in deep learning characteristic states that information from this additional dimension is largely ambiguous with respect to any specific concept. Depending on what has been used as the goals to mimic target data, these dataset distillation

techniques can be carried out jointly in both meta-learning and data matching framework; each of such methods for every context could further fall into an easy classification. In the domain of metrics-based meta-learning, we consider this distilled data to be a hyperparameter and performing nested loop optimization over it with regard to the model learnt on distilled-data. Data Coordination can be explained as the step updates distilled data by simulating how the target data would affect model training from either parameters or feature space [9]. Data Distillation can be further divided into meta learning, and data match-based frameworks given in Fig. 3. A meta learning perspective Trained distilled data as a hyperparameter with the objective of Data Distillation to learn how to process input data for enhancing general performance across models [1]. Dataset distillation helps in reducing the computational costs of learning model and obtain some partial solutions for security concerns, such as data privacy issues. Nonetheless, a majority of dataset distillation techniques implemented previously have focused on image classification with large scale image datasets. And various data types such as video, text and so on including 5 others that need to be processed [10]. Fig.3. DD frameworks based on [1]

Conclusion

The increased developments in Artificial Intelligence (AI), particularly in deep learning, have led to significant breakthroughs across various domains. However, this progress comes at a substantial computational and environmental cost. We have come across the point that the nature of training large AI models consumes a lot of energy, this gives rise to the concept of "Red AI," which emphasizes high computational power without considering its environmental impact. In contrast, "Green AI" has emerged as a paradigm that wants to achieve AI advancements in a more resource-efficient and environmentally sustainable manner. The main change in this movement is the transition from a model oriented to a data-oriented methods in AI research and development. Traditional AI has focused heavily on improving model architectures and increasing their size, often at the expense of computational efficiency. However, the data-centric approach stresses on the importance of high-quality, well-curated datasets to optimize model performance without necessitating larger models or excessive training cycles. Techniques such as dataset distillation, which condenses large datasets into smaller, more informative subsets, play a crucial role in this shift by enabling efficient learning with fewer resources. This paper explores the evolving landscape of AI, focusing on the transition toward data-centric methodologies and the importance of dataset optimization in achieving the goals of Green AI. We examine how deep learning models can benefit from these methods while reducing their environmental footprint, ultimately leading to more sustainable AI development practices

References

[1]. S. Salehi and A. Schmeink, "Data-Centric Green Artificial Intelligence: A Survey," in IEEE Transactions on Artificial Intelligence, vol. 5, no. 5, pp. 1973-1989, May 2024, doi: 10.1109/TAI.2023.3315272.

[2]. R. Verdecchia, L. Cruz, J. Sallou, M. Lin, J. Wickenden and E. Hotellier, "Data-Centric Green AI An Exploratory Empirical Study," 2022 International Conference on ICT for Sustainability (ICT4S), Plovdiv, Bulgaria, 2022, pp. 35-45, doi: 10.1109/ICT4S55073.2022.00015

[3]. E. Barbierato and A. Gatti, "Toward Green AI: A Methodological Survey of the Scientific Literature," in IEEE Access, vol. 12, pp. 23989-24013, 2024, doi: 10.1109/ACCESS.2024.3360705.

[4]. P. Heck, "What About the Data? A Mapping Study on Data Engineering for AI Systems," 2024 IEEE/ACM 3rd International Conference

on AI Engineering – Software Engineering for AI (CAIN), Lisbon, Portugal, 2024, pp. 43-52.

[5]. S. Kumar, S. Datta, V. Singh, S. K. Singh and R. Sharma, "Opportunities and Challenges in Data-Centric AI," in IEEE Access, vol. 12, pp. 33173-33189, 2024, doi: 10.1109/ACCESS.2024.3369417.

[6]. O. H. Hamid, "From Model-Centric to Data-Centric AI: A Paradigm Shift or Rather a Complementary Approach?," 2022 8th International Conference on Information Technology Trends (ITT), Dubai, United Arab Emirates, 2022, pp. 196-199, doi: 10.1109/ITT56123.2022.9863935.

[7]. P. -P. Luley, J. M. Deriu, P. Yan, G. A. Schatte and T. Stadelmann, "From Concept to Implementation: The Data-Centric Development Process for AI in Industry," 2023 10th IEEE Swiss Conference on Data Science (SDS), Zurich, Switzerland, 2023, pp. 73-76, doi: 10.1109/SDS57534.2023.00017.

[8]. N. P. Sable, S. Singh, M. K. Sharma, N. Patil, A. Mishra and L. B, "Machines Thinking Fast: AI-Powered Automating of Information Capture and Retrieval," 2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), Bangalore, India, 2023, pp. 1-6, doi: 10.1109/SMARTGENCON60755.2023.10441985

[9]. M. Nieberl, A. Zeiser and H. Timinger, "A Review of Data-Centric Artificial Intelligence (DCAI) and its Impact on manufacturing Industry: Challenges, Limitations, and Future Directions," 2024 IEEE Conference on Artificial Intelligence (CAI), Singapore, Singapore, 2024, pp. 44-51, doi: 10.1109/CAI59869.2024.00018.

[10]. S. Lei and D. Tao, "A Comprehensive Survey of Dataset Distillation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 1, pp. 17-32, Jan. 2024, doi: 10.1109/TPAMI.2023.3322540.

[11]. T. -T. -H. Le, H. T. Larasati, A. T. Prihatno and H. Kim, "A Review of Dataset Distillation for Deep Learning," 2022 International Conference on Platform Technology and Service (PlatCon), Jeju, Korea, Republic of, 2022, pp. 34-37, doi: 10.1109/PlatCon55845.2022.9932086.

[12]. T. Yarally, L. Cruz, D. Feitosa, J. Sallou and A. van Deursen, "Uncovering Energy-Efficient Practices in Deep Learning Training: Preliminary Steps Towards Green AI," 2023 IEEE/ACM 2nd International Conference on AI Engineering – Software Engineering for AI (CAIN), Melbourne, Australia, 2023, pp. 25-36, doi: 10.1109/CAIN58948.2023.00012.

[13]. N. Alizadeh and F. Castor, "Green AI: A Preliminary Empirical Study on Energy Consumption in DL Models Across Different Runtime Infrastructures," 2024 IEEE/ACM 3rd International Conference on AI Engineering – Software Engineering for AI (CAIN), Lisbon, Portugal, 2024, pp. 134-139.

[14]. S. Sikand, V. S. Sharma, V. Kaulgud and S. Podder, "Green AI Quotient: Assessing Greenness of AI-based software and the way forward," 2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE), Luxembourg, Luxembourg, 2023, pp. 1828-1833, doi: 10.1109/ASE56229.2023.00115

[15]. Gadekar, D. P., Sable, N. P., & Raut, A. H. (2019). Exploring Data Security Scheme into Cloud Using Encryption Algorithms. International Journal of Recent Technology and Engineering (IJRTE), Published By: Blue Eyes Intelligence Engineering & Sciences Publication, ISSN, 2277-3878.

[16]. A. Kannagi, T. Agrawal, M. K. Singar, K. R. Singh, K. K. Senthilkumar and N. P. Sable, "From Theory to Practice: Analyzing the Feasibility of Augmenting Human Intelligence with AI Technologies," 2024 15th International Conference on Computing Communication and

Networking Technologies (ICCCNT), Kamand, India, 2024, pp. 1-8, doi: 10.1109/ICCCNT61001.2024.10724873.