

Research on AI Computational Energy Consumption Optimization and Green AI

Jiawei Xu*

College of Arts and Science, Boston University, Boston, 02215, The United States

Abstract. With the rapid development of Artificial Intelligence (AI), many deep learning models have been extensively applied in various fields, and correspondingly, the environmental issues arising from the development of AI have become one of the pressing arguments at present. Due to the significant impact on the environment caused by the high energy consumption and high emissions of AI, it is crucial to figure out how to reduce the energy costs and environmental costs of AI. This study mainly aims to promote the development of green AI by recognizing, analyzing, and improving AI consumption. Firstly, it examines the sources of energy consumption and carbon footprint of AI, which stem from model training, data storage, and data transmission. Subsequently, it optimizes the energy consumption of AI computing, covering both software and hardware aspects. Moreover, it also elaborates on the applications of green AI in specific fields. Finally, it discusses the existing challenges and future trends of green AI.

1 Introduction

As artificial intelligence technology (AI) develops rapidly, especially its high-efficiency performance demonstrated in various fields, it has gradually become accessible to the general public and emerged as the mainstream technology in contemporary society. Along with the rapidly increased demand, the energy required for constructing AI models and training AI has also been gradually rising. Therefore, widespread attention has been paid to the energy consumption and environmental issues caused by AI.

AI models pose multiple challenges during the training and usage processes. The economic costs for the development and training of AI models are substantial. For instance, the costs of developing GPT3 and GPT4 soared from 4 million US dollars to 63 million US dollars. In terms of environmental costs, the training and calculation of AI models, as well as the operation and maintenance of data centers, need to be supported by a massive amount of electrical energy. Moreover, approximately 0.72 tons of carbon dioxide will be emitted for every 1,000 kWh of electricity consumed [1], which has already significantly affected the global environment. Under the background of the global advocacy for carbon neutrality and sustainable development, AI needs to be improved to be more environmentally friendly without consuming much energy. Therefore, promoting the development of green AI has become an important issue in contemporary artificial intelligence. The development of AI

* Corresponding author: xujiawei@bu.edu

should not only concentrate on the computational efficiency of models but also pay more attention to environmental protection, eliminate resource waste, and step towards sustainable development.

This paper mainly focuses on the energy consumption and carbon emission issues caused by AI models, as well as the development of green AI. The second part concentrates on the energy consumption problems of AI computing and the analysis of the carbon footprint. It studies the environmental issues caused by AI model training, data storage, and data transmission in data centers, respectively. The third part mainly analyzes the optimization paths for AI computing energy consumption, conducting research ranging from the optimization of model structures and parameters at the software level to the development of energy-saving chips and energy-efficient operations at the hardware level. The fourth part is the practical applications of green AI in different fields, highlighting its potential. The fifth part discusses the challenges and overall trends that will be encountered by green AI in the future. The sixth part summarizes the entire research and presents outlooks for the future.

2 Energy consumption issues of AI computing

Boosted by the rapid development of artificial intelligence (AI) technology, serious energy consumption problems have been caused by the high demand for data computing. Model training, data storage, and data transmission of AI account for the majority of the energy consumption in AI computing.

2.1 Sources of energy consumption in AI computing

In recent years, large-scale AI language models (such as ChatGPT/Deepseek) have been extensively applied in various fields, thus causing significant energy consumption issues in the domain of AI. The development of technical models requires several to hundreds of thousands of training accelerators for the training of AI models [2], consuming more than one million kilowatt-hours of electricity. Besides, training large-scale natural language processors not only demands a substantial investment of energy but also incurs huge environmental costs [2]. Therefore, how to reduce energy consumption and carbon emissions during the training process of AI models has become an important issue.

2.1.1 Energy consumption produced by data storage in data centers

The data sources for AI computing all come from data centers, which are core facilities that require continuous energy supply. Artificial intelligence models rely mostly on data, which stimulates the consumption of a vast amount of energy for data sets and large data centers [3]. Data centers are usually designed and equipped to meet the needs of high-density computing devices, which require dedicated uninterruptible power supplies and cooling and heat dissipation systems. Moreover, data centers also need to be equipped with lighting devices and environmental control equipment to maintain the temperature and humidity, and control the indoor air quality, aiming to achieve the efficient operation of data [4]. Therefore, how to optimize the electricity consumption during the data storage process has a profound impact.

2.1.2 Analysis of energy consumption in the data transmission process

The transmission of AI data also results in a large amount of energy consumption. The operation of network devices causes energy consumption of equipment such as routers,

switches, servers, and base stations. The energy consumed by data transmission is also affected by factors such as the scale of data and the transmission distance. With the gradual popularization of AI, the scale of data transmission will increase, so energy consumption will also gradually rise. The prediction in relevant literature reveals that if the technology and behavior remain unchanged, the power demand of data centers will increase from 286 TWh in 2016 to 321 TWh in 2030 [5]. Therefore, the primary goal is to study how to optimize energy consumption of data transmission and improve the energy transmission efficiency.

2.2 Analysis of the carbon footprint of AI computing

2.2.1 Analysis of the carbon emissions of AI models

The carbon emissions of AI models mainly come from the training of AI models and the application process. The carbon emissions required for training an AI model is equivalent to that of five cars [6]. For example, training a general language translation model (LM) generates 45.2 metric tons of carbon emissions, while the carbon emissions generated by a deep learning recommendation model (DLRM) are four times that of the LM [2]. Moreover, the amount of carbon emissions is directly proportional to the size of the training computational model, revealing that the cost of adjusting large models will surge [6]. Besides, AI models applied to actual production and those that need to fit the real-world environment require continuous updating and training, and the renewal process also exacerbates carbon emissions. As the scale of AI applications expands, the total scale of carbon emissions is still increasing.

2.2.2 Current situation of energy consumption and carbon emissions in data centers

Constructing and operating data centers is accompanied by huge energy consumption and carbon emissions. With the increasing demand for data calculation and processing in various fields, a large number of servers in data centers need to operate around the clock, so the carbon emissions of information and communication technology (ICT) are growing at a rate of 6% per year [7]. The main equipment causing carbon emissions in data centers is the servers and the cooling system, each accounting for 40% of the total energy consumption [7]. Therefore, to maintain the sustainable development of data centers, it is extremely crucial to reduce the operating costs and huge energy consumption of servers and optimize the carbon emissions of the cooling and heat dissipation systems. Relevant literature estimates that data centers consume more than 2.4% of the world's total electricity, and at the same time, data centers also emit 78.7 million metric tons of carbon dioxide, equivalent to 2% of global emissions [8].

3 Technical paths for optimizing the energy consumption of AI computing

3.1 Low-energy-consuming deep learning models (at the software level)

3.1.1 Simplification of model data and parameter optimization (removing unnecessary model structures and data)

Pruning of AI models and optimization of network structures can remove the redundant parts in the models without affecting their usability, thus reducing the demand for computing resources. For example, Autopruner is an end-to-end trainable filter pruning method that reduces the model size and computational cost by deleting unnecessary neurons [9]. Meanwhile, pruning AI data can also effectively help AI computing avoid unnecessary data in advance to achieve more efficient AI computing. If reasonably designed, data scaling, sampling, and rational management and deployment can effectively improve the efficiency of devices, reduce the time for training models, and provide higher-quality AI models [2].

3.1.2 Knowledge distillation

Knowledge distillation can omit repeated learning processes, and compressing the model can reduce the cost and energy consumption of model training. By transferring information from a large model or a group of models to a small model during its training, the accuracy of the small model will not be significantly reduced [10].

3.2 Green computing hardware (at the hardware level)

3.2.1 Developing energy-saving chips and optimizing the energy consumption of AI computing

The application of artificial intelligence is accompanied by a large amount of data processing. Therefore, high-efficiency and low-energy-consuming artificial intelligence chips are particularly crucial for green and energy-saving AI. High-efficiency energy-saving chips and dedicated AI accelerators, such as GPUs (Graphics Processing Units) and TPUs (Tensor Processing Units), can reduce the energy consumption of AI computing by utilizing their unique parallel operation structures. For example, a TPU can perform 4 trillion operations per second. When considering aspects such as control and data orientation and comparing it with a CPU, it can effectively reduce the training time when training large and complex neural networks, thus reducing computational energy consumption [11].

3.2.2 Optimization of data storage technology and effective heat dissipation during the computing process (reducing energy consumption)

In addition to AI computing, it is also necessary to optimize data storage in data centers to reduce the total Power Usage Effectiveness (PUE) and achieve green AI. Data storage facilities (disk arrangements) and built-in cooling systems (fans, coolers) are the main sources of energy consumption in data centers. When storing data, it is possible to distinguish between data with high-frequency usage and those with low-frequency usage. By placing low-frequency data in the low-energy-consuming storage layer, the storage energy consumption of these data can be reduced. Meanwhile, the stored data contains a certain amount of duplicate data, which will impose an unnecessary burden on the storage space. Therefore, deleting duplicate data is also one of the methods to reduce energy consumption. The cooling system accounts for a large proportion of the energy consumption in data centers, approaching 40% [7]. Both external and internal factors can be used to optimize the energy

consumption of the data center cooling system. External factors include the climate of the data center's location, which affects the overall power consumption of the cooling system. Under higher temperatures and relative humidity, the power consumption of the cooling system in a data center can increase by more than double compared to that in a data center with relatively lower temperatures [12]. Therefore, the external climate plays a crucial role in energy consumption. Besides, the actual heat dissipation rate of the traditional fan system inside the data center has no significant advantage compared to more efficient cooling media. On the contrary, the thermal conductivity of the cooling liquid is much higher than that of air, and its heat dissipation power is more than five times that of the traditional heat dissipation system [13].

3.2.3 Application of sustainable energy sources (such as solar energy and wind power) to reduce unnecessary energy consumption

From the perspective of data centers, the energy sources are from fuel emissions [2], which greatly increases the carbon emission level. By introducing sustainable energy sources, such as renewable energy sources like solar energy and wind power, into data centers, the dependence on traditional fuels can be significantly reduced. This not only saves fuel costs but also reduces carbon emissions. For example, Parasol is a solar-powered data center that balances the consumption of non-renewable energy by entirely relying on renewable energy sources [14].

4 Application scenarios of green AI

With the widespread popularization of AI, the environmental problems caused by AI have become a hot topic of discussion at present. Green AI has played a full role in solving the problems of high energy consumption and carbon emissions.

4.1 AI optimizes traffic carbon emissions

AI can predict and analyze future traffic trends, control the switching of traffic lights, and avoid traffic congestion based on past traffic flow data, which significantly shortens driving duration. In this way, the carbon emissions caused by fuel consumption can be reduced. For example, the combination of an artificial neural network (ANN) and statistics has enabled hourly-level prediction of traffic flow and the artificial bee colony (ABC) algorithm has been applied to coordinate the signal timing of road networks [15]. Thus, AI reduces the potential carbon emissions caused by traffic.

4.2 AI Applied in the green manufacturing industry

4.2.1 AI formulates solutions by monitoring carbon emissions in real time and analyzing the data

AI can monitor in real time the carbon emission data generated by various parts of a factory, analyze the carbon emissions produced in various links such as transportation, production, and storage, look for the sources of high carbon emissions for analysis, aiming to identify the parts where emissions can be reduced. The predictive analysis of data by AI can simulate various scenarios to optimize resource allocation and operational issues, thus minimizing environmental pollution [16].

4.2.2 Predicting future environmental changes through AI

AI can predict possible future extreme weather and urban heat island effects by analyzing current carbon emission data and the global images provided by meteorological satellites. Governments and factories can take corresponding measures in advance based on the predictions made by AI to avoid potential problems. For example, AI can be used to predict short-term or long-term carbon dioxide emissions to formulate artificial restriction measures in advance to limit the total amount of carbon emissions [17]. Moreover, traditional prediction methods require manually developing an observation parameter for each assimilated observation type, so a large amount of manpower needs to be invested in the development process [18]. In contrast, AI prediction can not only save labor costs but also learn the hidden relationships that physical models cannot capture [18].

5 Challenges and future trends of green AI

5.1 Challenges of green AI

5.1.1 Will energy consumption optimization and algorithm optimization impact the performance accuracy of AI computing? And how can we ensure the efficient operation of AI

Green AI has emerged as the current mainstream. Although AI models can curtail energy consumption via methods like model pruning, parameter optimization, and knowledge distillation, this process unavoidably impairs the models' performance and accuracy. Hence, it is of utmost importance to strike a balance between energy conservation and accuracy in order to reach the desired objectives.

5.1.2 Will the open sharing of data in data centers pose privacy problems

Sharing data in data centers for training AI models can cut down on the energy consumption resulting from repeated computations. However, there exists a potential privacy risk of information leakage. Data centers might contain sensitive data related to individuals and enterprises, some of which may not be encrypted. As a consequence, information leakage could happen during the training of AI models.

5.2 Future trends of green AI

5.2.1 Wide-scale popularization of green AI

Environmental protection and sustainable development are fundamental elements of future global development. Green AI is not merely confined to energy conservation and carbon emission reduction. Instead, it is applied in various fields such as finance, agriculture, and transportation in pursuit of more comprehensive sustainable development.

5.2.2 The steady development of AI makes AI more reliable in energy optimization

Green AI currently offers highly efficient returns and demonstrates excellent performance in energy optimization and energy forecasting. With the continuous exploration of AI in environmental protection, data models are becoming more mature, so the continuous

accumulation of data and the reinforcement learning of AI can make the results of green AI more reliable. This will lead enterprises and governments to rely more on green AI, aiming to achieve environmental optimization and optimal resource allocation.

6 Conclusion

This paper reviews the environmental impacts caused by the rapid development of AI. Then it analyzes the energy consumption and carbon emissions resulting from current AI model training, the storage in AI data centers, and data transmission. To address these issues, corresponding AI computing optimization technologies are proposed. Software optimization for AI energy consumption problems uses methods such as model pruning optimization, parameter optimization, and knowledge distillation. Besides, the development of energy-saving chips, the optimization of data center operation and heat dissipation, as well as the application of sustainable energy sources help to address the problems at the hardware level. The development of green AI is not confined to a single field. On the contrary, green AI provides substantial assistance in various fields. Relevant literature reveals that green AI also plays a role in reducing emissions in transportation and the manufacturing industry. However, it still faces multiple challenges, for example, after software-level optimization, how to ensure the accuracy of the model to achieve the energy-saving effect, and the privacy issues after data sharing. Despite these challenges, the future development trend of green AI is still steadily on the rise, which has a profound impact on global sustainable development. Therefore, Green AI will gradually enter more fields and help the global industry achieve green development.

References

1. M. Uddin, Y. Darabidarabkhani, A. Shah, J. Memon, Evaluating power efficient algorithms for efficiency and carbon emissions in cloud data centers: A review. *Renew. Sustain. Energy Rev.* **51**, 1553 – 1563 (2015)
2. C. J. Wu, B. Acun, R. Raghavendra, K. Hazelwood, Beyond efficiency: Scaling AI sustainably. *IEEE Micro* (2024)
3. R. Nishant, M. Kennedy, J. Corbett, Artificial intelligence for sustainability: Challenges, opportunities, and a research agenda. *Int. J. Inf. Manag.* **53**, 102104 (2020)
4. G. Lykou, D. Mentzelioti, D. Gritzalis, A new methodology toward effectively assessing data center sustainability. *Comput. Secur.* **76**, 327 – 340 (2018)
5. M. Koot, F. Wijnhoven, Usage impact on data center electricity needs: A system dynamic forecasting model. *Appl. Energy.* **291**, 116798 (2021)
6. K. Hao, Training a single AI model can emit as much carbon as five cars in their lifetimes. *MIT Technol. Rev.* **75**, 103 (2019)
7. H. Rong, H. Zhang, S. Xiao, C. Li, C. Hu, Optimizing energy consumption for data centers. *Renew. Sustain. Energy Rev.* **58**, 674 – 691 (2016)
8. S. Mondal, F. B. Faruk, D. Rajbongshi, M. M. K. Efaz, M. M. Islam, GEECO: Green data centers for energy optimization and carbon footprint reduction. *Sustain.* **15(21)**, 15249 (2023)
9. J. H. Luo, J. Wu, Autopruner: An end-to-end trainable filter pruning method for efficient deep model inference. *Pattern Recognit.* **107**, 107461 (2020)
10. J. Gou, B. Yu, S. J. Maybank, D. Tao, Knowledge distillation: A survey. *Int. J. Comput. Vis.* **129(6)**, 1789 – 1819 (2021)

11. G. S. Nikolić, B. R. Dimitrijević, T. R. Nikolić, M. K. Stojcev, A survey of three types of processing units: CPU, GPU and TPU. In Proc. 57th Int. Sci. Conf. Inf., Commun. Energy Syst. Technol. (ICEST), 1 - 6 (2022)
12. N. Lei, E. Masanet, Statistical analysis for predicting location-specific data center PUE and its improvement potential. *Energy*. **201**, 117556 (2020)
13. R. Kong, H. Zhang, M. Tang, H. Zou, C. Tian, T. Ding, Enhancing data center cooling efficiency and ability: A comprehensive review of direct liquid cooling technologies. *Energy*. **308**, 132846 (2024)
14. S. Bharany, S. Sharma, O. I. Khalaf, G. M. Abdulsahib, A. S. Al Humaimeedy, T. H. Aldhyani, ... H. Alkahtani, A systematic survey on energy-efficient techniques in sustainable cloud computing. *Sustain*. **14(10)**, 6256 (2022)
15. A. Kadkhodayi, M. Jabeli, H. Aghdam, S. Mirbakhsh, Artificial intelligence-based real-time traffic management. *J. Electr. Electron. Eng.* **2(4)**, 368 - 373 (2023)
16. B. Ameh, Digital tools and AI: Using technology to monitor carbon emissions and waste at each stage of the supply chain, enabling real-time adjustments for sustainability improvements. *Int. J. Sci. Res. Arch.* **13(1)**, 2741 - 2754 (2024)
17. Y. Meng, H. Noman, Predicting CO₂ emission footprint using AI through machine learning. *Atmosphere*. **13(11)**, 1871 (2022)
18. S. Dewitte, J. P. Cornelis, R. Müller, A. Munteanu, Artificial intelligence revolutionises weather forecast, climate monitoring and decadal prediction. *Remote Sens.* **13(16)**, 3209 (2021)