

Lower Energy Large Language Models (LLMs)

Hsiao-Ying Lin¹, Huawei France

Jeffrey Voas², IEEE Fellow

This message offers ideas about how to reduce the energy consumption associated with large language models.

We continually hear about the high energy costs required for artificial intelligence (AI) and machine learning (ML). So, we decided to see what has already been written about the energy costs specific to large language models (LLMs). Here is a little bit of what we found and a better explanation of our curiosity.

Since the introduction of ChatGPT, the business influence of LLMs has increased greatly. Various LLM-based applications are emerging and bring a huge potential for increasing work productivity, such as customized online chatting services¹ and topic-focused document analysis.² Meanwhile, a variety of new LLMs, such as HuggingChat,³

LLaMA,⁴ and stableLM,⁵ are continually being introduced. While these LLMs unlock a fruitful set of automatically generative capabilities, they also raise public concerns about the energy consumption of the LLM training and inference processes. For example, training the largest GPT-3 model, which has 175 billion parameters, consumes approximately 1,287 MWh,⁶ while the average annual electricity consumption for a

U.S. residential utility customer was only approximately 10 MWh in 2021.⁷ As new LLM models are emerging, such as GPT-4 (170 trillion parameters), the energy consumption momentum will continue.

As global heads of state have announced ambitious targets to manage the climate crisis, green and sustainable technologies are developing in all sectors, including AI. As current LLMs consume considerable energy, innovative concepts and technologies for lower energy consumption LLMs are a new research topic.

DISCLAIMER

The authors are completely responsible for the content in this message. The opinions expressed here are their own.




IN THIS ISSUE

In the second part of this special issue on “Telemedicine,” I’m adding two additional articles. I thank these authors for their patience in waiting for their accepted articles to be published. *Computer* has seen an uptick in the number of submissions, and we have a backlog of accepted articles.

In the article “VD-HEN: Capturing Semantic Dependencies for Source Code Vulnerability Detection With a Hierarchical Embedding Network,”^{A1} the authors propose a semantic dependency capture method for source code vulnerability detection. Their method extracts syntactic information and structural information through a statement embedding network and program embedding network. A recursive neural network is used to learn the statement representation from an AST subtree that corresponds to each statement, and the tree-based convolutional neural network is used to extract program structure information.

In the article “Enabling Cost-Benefit Analysis of Data Sync Protocols,”^{A2} the authors discuss data

synchronization in networked applications. The article states that data synchronization improves performance and describes a new middleware called *GenSync* that abstracts the subtleties of data synchronization protocols that allow users to choose protocols that are based on comparative evaluations under operational conditions. The article includes a case study where *GenSync* is integrated into a large wireless emulator.

—Jeffrey Voas, Editor in Chief 

APPENDIX: RELATED ARTICLES

- A1. J. Hao, S. Luo, L. Pan, and C. Chen, “VD-HEN: Capturing semantic dependencies for source code vulnerability detection with a hierarchical embedding network,” *Computer*, vol. 56, no. 10, pp. 49–61, Oct. 2023, doi: 10.1109/MC.2022.3228924.
- A2. N. Boškov, A. Trachtenberg, and D. Starobinski, “Enabling cost-benefit analysis of data sync protocols,” *Computer*, vol. 56, no. 10, pp. 62–71, Oct. 2023, doi: 10.1109/MC.2023.3251195.

Digital Object Identifier 10.1109/MC.2023.3295521
Date of current version: 20 September 2023

So, which lower energy technologies for LLMs are accessible now? Available approaches can be classified into two categories.

reasonable overhead. Hence, various customized LLMs are generated by slightly fine-tuning existing available LLMs. For example, one open source

running on a GPU-Nvidia A100. The base model (based on LLaMA) has 6.7 billion parameters, of which 4.2 million are trainable during fine-tuning. Thus, higher energy efficiency can be achieved by employing existing models by amortizing the training cost.

The second category focuses on how to develop from-scratch LLMs with better energy efficiency. Three examples are introduced here.

First, energy efficiency can be improved by increasing the reuse rate, that is, partly amortizing the training energy cost.

First, energy efficiency can be improved by increasing the reuse rate, that is, partly amortizing the training energy cost. After LLMs are trained at a considerable energy expense, a promising approach is to reuse the resulting LLMs as much as possible with


tool called *xTuring*⁸ enables developers to fine-tune available typical LLMs for customized purposes with indicated datasets. So, in a little experiment that we ran, it took approximately 100 computing units and 7 h on Google Colab to fine-tune a language model

1. MIT researchers presented a strategy called *Learned Linear Growth Operator (LiGO)*, which converts a smaller model to a larger model.⁹ This strategy can

reduce the computational cost of training vision and language models by approximately 50%.

2. A smaller model size is considered in the beginning while maintaining comparable performance, such as LLaMA (65 billion parameters) and Claude (52 billion parameters).
3. Energy efficiency can be obtained via leveraging sustainable high-performance computing technology, including customized hardware innovations¹⁰ and green-oriented software orchestrators.¹¹

Note: *Computer's* "Artificial Intelligence/ Machine Learning" column opens fresh perspectives, novel concepts, and sustainable solutions related to AI and ML, so please consider submitting column articles to it at hsiaoying.lin@gmail.com.

task for sustainability. This is something that we all need to keep an eye on as tools such as ChatGPT become the new norm. 

ACKNOWLEDGMENT

The corresponding author is Hsiao-Ying Lin.

REFERENCES

1. "Character.ai." Accessed: May 7, 2023. [Online]. Available: <https://beta.character.ai/>

2. "ChatPDF." Accessed: May 7, 2023. [Online]. Available: <https://www.chatpdf.com/>
3. "HuggingChat." Accessed: May 7, 2023. [Online]. Available: <https://huggingface.co/chat/>
4. H. Touvron et al., "LLaMA: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.
5. "StableLM: Stability AI language models." GitHub. Accessed: May 7, 2023. [Online]. Available: <https://github.com/Stability-AI/StableLM>
6. D. Patterson et al., "The carbon footprint of machine learning training will plateau, then shrink," *Computer*, vol. 55, no. 7, pp. 18–28, Jul. 2022, doi: 10.1109/MC.2022.3148714.
7. "How much electricity does an American home use?" U.S. Energy Inf. Admin., Washington, DC, USA, Oct. 2022. Accessed: May 7, 2023. [Online]. Available: <https://www.eia.gov/tools/faqs/faq.php?id=97&t=3>
8. "xTuring." Accessed: May 7, 2023. [Online]. Available: <https://xturing.stochastic.ai/>
9. P. Wang et al., "Learning to grow pretrained models for efficient transformer training," 2023, *arXiv:2303.00980*.
10. W. Wan et al., "A compute-in-memory chip based on resistive random-access memory," *Nature*, vol. 608, no. 7923, pp. 504–512, Aug. 2022, doi: 10.1038/s41586-022-04992-8.
11. J. McDonald, B. Li, N. Frey, D. Tiwari, V. Gadepally, and S. Samsi, "Great power, great responsibility: Recommendations for reducing energy for training language models," 2022, *arXiv:2205.09646*.

So, as LLMs open new chapters for human-machine interactions, how to develop and maintain greener LLMs becomes an essential



IEEE Software offers pioneering ideas, expert analyses, and thoughtful insights for software professionals who need to keep up with rapid technology change. It's the authority on translating software theory into practice.

www.computer.org/software

HSIAO-YING LIN is a principal researcher at Huawei France, 92100 Boulogne-Billancourt, France. Contact her at hsiaoying.lin@gmail.com.

JEFFREY VOAS, Gaithersburg, MD 20899 USA, is the editor in chief of *Computer*. He is a Fellow of IEEE. Contact him at j.voas@ieee.org