Strategies for mitigating the global energy and carbon impact of artificial intelligence

Daniel Xue, University of Virginia, USA (dlx3ud@virginia.edu)

Abstract

With the advent of innovations in artificial intelligence such as ChatGPT, such technologies have the ability to revolutionize communications, research, and human expression. Given that the energy efficiency of the information and communication technologies (ICTs) is estimated to contribute 2-4% of the world's global greenhouse gas emissions, however, the energy and carbon demand of such machine learning models poses a threat to achieving net zero emissions by 2050, as demand for artificial intelligence applications only grows. While work is already underway for reducing the electricity consumed and equivalent carbon dioxide emitted for model training and inference, further progress will require action from policymakers, researchers, and industry experts. Three initial recommendations for addressing this issue include 1) investment in and collaboration on green AI research efforts, 2) development of better reporting and tracking practices of energy and carbon impact of machine learning models, and 3) creation of industry standards for both reducing and offsetting greenhouse gas emission in pursuit of the 2050 net zero emissions goal.

"ChatGPT is a large language model developed by OpenAI, designed to simulate human-like conversation. It uses deep learning algorithms to understand natural language and generate responses that are contextually relevant and coherent. ChatGPT has been trained on a massive corpus of text data, including books, articles, and online content, which allows it to generate a wide range of responses on a variety of topics. The goal of ChatGPT is to provide users with an engaging and informative conversational experience that feels as close to talking to a human as possible" (ChatGPT, 2023).

Launched in November 2022, ChatGPT has already reshaped the way that people across the globe communicate, taking everyday writing tasks and churning out responses within seconds. While it should be noted that ChatGPT's "outputs may be inaccurate, untruthful, and otherwise misleading at times" per its FAQs (Staudacher, 2023), its sheer speed and flexibility allows the model to be broadly applicable for basic writing tasks, from cover letters, to public service announcements, to even this policy brief itself.

Because of this, usage of the ChatGPT web app has skyrocketed, hitting 616 million visits to the chat.openai.com website in January and over 1 billion visits this past month (Carr, 2023). As of writing, estimates of the model's energy consumption for January 2023 are on the order of 4,200 MWh, assuming that its 13 million users made 15 requests per day and the model was run on Nvidia A100 GPUs (Ludvigsen, 2023b). This makes the model's electricity usage roughly equivalent to 30,000 Danish citizens, not including the additional 1,287 MWh required for developing its underlying model, OpenAI's GPT-3 (Ludvigsen, 2023a). These energy figures are on the magnitude as other large language models (LLMs), such as BigScience Workshop's BLOOM which consumed 433 MWh (Luccioni et al., 2022) or Google's GLaM which consumed 456 MWh (Patterson et al., 2022) for training. BLOOM was also found to consume a similar amount of energy as ChatGPT, at roughly 0.00396 kWh per query (Ludvigsen, 2023a).

If ChatGPT drew its electricity from the average American power grid, using the EPA's estimate of 0.855 lbs. (0.388 kgs.) of equivalent CO_2 (CO_2e) emitted per kWh of electricity in 2021 (EIA, 2022) means that the model's training and January deployment energy usage would have released about 2,100 metric tons of CO_2e into the atmosphere, which is little less than that released from the entire lifespan of 37 cars, given an average of 126,000 lbs. (approx. 57,000 kgs) (Strubell et al., 2019). This estimate does not include the additional emissions from powering the servers, network equipment, and client-side devices for using ChatGPT's web application.

Given that the carbon footprint of the Information Communication Technology (ICT) sector is already estimated to be 2.1%-3.9% of global greenhouse gas (GHG) emissions (Freitag et al., 2021), the unprecedented growth and energy consumption of training and deploying LLMs such as that used for ChatGPT poses a threat to the IPCC's goal of reducing net greenhouse gas emissions to zero by 2050 (IPCC, 2018). Even with current efforts to reduce and offset GHG emissions, broad agreement exists that the energy demand of ICTs will continue to grow due to rebound effects and significant investment in emerging technologies, such as the developments in natural language processing (NLP) described above (Freitag et al., 2021). Thus, generating policy for the tracking and development of greener AI applications would not only be prudent for addressing the ongoing climate crisis, but would moreover provide opportunities for further pursuing the sustainable development goals of renewable energy, resilient infrastructure, and global partnerships.

Energy Consumption

Energy consumed for the training and inference of machine learning models can be grouped into two categories: (1) dynamic, which includes the energy drawn from all CPU, GPU, and main memory sockets, and (2) idle, which consists of cooling and network infrastructure.

For training, dynamic power consumption is calculated by multiplying the number of training hours by the sum of the product of thermal design power times the number present in the system of each component (Strubell et. al., 2019). Thermal design power can accurately represent GPU power during training, since GPUs ideally have100% utilisation (Luccioni et al., 2022). Idle power consumption can then be factored in via separate measurement and calculation (like in Luccioni et al., 2022) or power usage effectiveness (PUE), an industry metric that gives the ratio between the total data center energy consumed and the energy consumed by its computing equipment (Patterson et. al., 2022). While the industry average for PUE was 1.58 in 2020, groups such as (Patterson et. al., 2022) can take advantage of the lower PUE, approx. 1.1, of cloud providers to reduce the energy consumption of their models.

The same methodology can be mostly followed for deployment, however, it should be noted that the real-time nature of queries means that optimization techniques such as batching and padding cannot be used to reduce energy consumption (Luccioni et. al., 2022). More research into tracking real-time model inference power usage is required to better estimate energy consumed during deployment.

Greenhouse Gas Emissions

Greenhouse gas emissions from training and deploying machine learning models can be grouped into two categories: (1) operational, which covers the GHG emitted from electricity consumed dynamically and idly, and (2) lifecycle, which includes the GHG emitted from the materials and processes required for the manufacture, transport, and disposal of all components employed (Patterson et al., 2022).

For operational GHG emissions, the methodology described in the energy consumption section can be reused for calculating the CO_2e emitted during both training and inference, with an additional step for multiplying the total energy used by the carbon intensity of the associated power grid. While the average carbon intensity of the national grid is sufficient for estimating the GHG emitted in some cases, such as in (Strubell et al., 2019), each data center's carbon intensity will vary based on its location and energy available. Estimating operational emissions can quickly become difficult for models deployed through web applications like ChatGPT, which can be hosted in multiple locations, each with its own energy mix that can even vary throughout the day (Pointon, 2023). While such metrics are not publicly available for all such centers, this fact has been successfully leveraged to decrease the total emissions of model training and inference, as shown in (Patterson et al., 2022) and (Luccioni et. al., 2022).

Likewise, calculating the lifetime GHG emissions of model training and deployment is not a simple task. The embodied emissions for a process is equal to the total GHG emitted during the manufacture over the time of use for a given component used (Luccioni et al., 2022). For an LLM such as BLOOM, embodied emissions were found to be 0.056 kg and 0.003 kg of CO_{2e} emitted for each hour of server and GPU time, respectively, resulting in an additional 11.2 metric tons of equivalent CO_2 (CO_2e) for training the model (Luccioni et. al., 2022). While this only represented 22.2% of the emitted CO_2e from the model's life cycle, this number did not account for the other network infrastructure or cooling equipment necessary for model training (Luccioni et. al. 2022). Characterising embodied emissions for model deployment would also have to include estimates for client-side devices, which would include phones, laptops, personal computers, TVs, and other smart devices. Further research is necessary to fully characterise the embodied emissions of artificial intelligence applications and other ICTs.

Policy recommendations

Given the above information, policymakers, researchers, and industry experts should collaborate on the following actions:

Research

1. Invest in development of more energy-efficient algorithms, hardware, data center equipment, and other practices
2. Pool resources for energy-efficient equipment and share best practices for data center management
3. Identify and promote locations across the globe that have the potential for more renewable energy and lower carbon intensity data centers

Transparency

1. Establish and maintain metrics and standards for evaluating the energy-efficiency of training and deploying machine learning models, such as that described in (Henderson et al., 2020)
2. Incentivize the reporting of embodied emissions and carbon intensity for hardware and data centers used for training and deploying machine learning models
3. Inform the public regularly about the energy and carbon impact of artificial intelligence, and what measures have been taken to mitigate said impact

Accountability

1. Create a plan for reducing machine learning models GHG emissions for pursuing net zero emissions by 2050
2. Measure and characterize the indirect impact of artificial intelligence on the GHG emissions of other sectors, including the potential rebound effects described in (Freitag et al., 2021)
3. Incentivize self-regulating companies and other organizations to pursue standards for permanence, verifiability, and additionality (Freitag et al., 2021) in their own efforts to offset their GHG emissions

While it can be tempting to adopt a pessimistic view of artificial intelligence and ICTs due to their rapid growth, it should always be noted that substantial work is already underway to reduce the energy consumed and GHG emitted by machine learning models. Patterson et al. report that their LLM GLaM only emitted 40 metric tons of CO_2e in the model's training, which is almost 14x smaller than the 552 tCO_2e emitted for GPT-3 (2022). As of writing, it is unfortunately unknown what the lifecycle emissions of the model are, as the embodied emissions for the TPUs used for training are not available. While it is important to recognize that improving LLM energy efficiency does not necessarily decrease their carbon footprint due to rebound effects (Freitag et. al., 2021), developing improvements in both metrics will be required for decreasing the overall impact of AI. With the proper actions taken for improving research, transparency, and accountability for machine learning models, however, such information could not only help mitigate climate change but also usher in a world with cleaner energy, greater innovation, and stronger partnerships.

Acknowledgments

I would like to acknowledge and thank Professor Rider Foley and Bill Kelly for their contributions in drafting and revising the above brief through the Policy Internship Program at the University of Virginia.

References

Boudreau, C. (2023, February 14). I asked ChatGPT about its carbon footprint and it didn't have a real answer. Business Insider.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. arXiv.

Carr, D. F. (2023, March 7). ChatGPT Topped 1 Billion Visits in February. Similarweb Blog.

ChatGPT. (2023, March 8). Re:What is ChatGPT?

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. de O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., … Zaremba, W. (2021). Evaluating large language models trained on code. arXiv.

Du, N., Huang, Y., Dai, A. M., Tong, S., Lepikhin, D., Xu, Y., Krikun, M., Zhou, Y., Yu, A. W., Firat, O., Zoph, B., Fedus, L., Bosma, M., Zhou, Z., Wang, T., Wang, Y. E., Webster, K., Pellat, M., Robinson, K., … Cui, C. (2022). Glam: Efficient scaling of language models with mixture-of-experts. arXiv.

EIA. (2022, November 25). Frequently asked questions (FAQs) . United States Energy Information Administration.

Freitag, C., Berners-Lee, M., Widdicks, K., Knowles, B., Blair, G. S., & Friday, A. (2021). The real climate and transformative impact of ICT: A critique of estimates, trends, and regulations. Patterns, 2(9), 100340.

Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., & Pineau, J. (2020). Towards the systematic reporting of the energy and carbon footprints of machine learning. The Journal of Machine Learning Research, 21(1), 248:10039-248:10081.

IPCC. (2018). Special Report—Global Warming of 1.5 ºC. Intergovernmental Panel on Climate Change.

Luccioni, A. S., Viguier, S., & Ligozat, A.-L. (2022). Estimating the carbon footprint of BLOOM, a 176B parameter language model.

Ludvigsen, K. G. A. (2023a, March 5). Chatgpt's electricity consumption. Medium.

Ludvigsen, K. G. A. (2023b, March 5). ChatGPT's electricity consumption, pt. II. Medium.

Patterson, D., Gonzalez, J., Holzle, U., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D. R., Texier, M., & Dean, J. (2022). The carbon footprint of machine learning training will plateau, then shrink. Computer, 55(7), 18–28.

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). Carbon emissions and large neural network training. arXiv.

Pointon, C. (2023, March 3). The carbon footprint of ChatGPT. Medium.

Roehrl, R. A. (2021). Impacts of new Internet applications and artificial intelligence on global energy demand – an issue of concern? (Emerging Science, Frontier Technologies, and the SDGs - Perspectives from the UN System and Science and Technology Communities, pp. 165–171). United Nations Interagency Task Team on Science, Technology and Innovation for the Sustainable Development Goals.

Staudacher , N. (2023, February). ChatGPT General FAQ. OpenAI.

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP.

UN. (2022a). Goal 7 | Ensure access to affordable, reliable, sustainable and modern energy for all. Sustainable Development; United Nations Department of Economic and Social Affairs.

UN. (2022b). Goal 9 | Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation. Sustainable Development; United Nations Department of Economic and Social Affairs.

UN. (2022c). Goal 17 | Strengthen the means of implementation and revitalize the Global Partnership for Sustainable Development. Sustainable Development; United Nations Department of Economic and Social Affairs.