International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 12 | Dec - 2021 SJIF Rating: 6.714 ISSN: 2582-3930

Energy Efficient Large Language Models: Advancements and Challenges

Vishakha Agrawal vishakha.research.id@gmail.com

***Abstract*—The rapid growth of Large Language Models (LLMs) has revolutionized the field of natural language processing (NLP), achieving unparalleled levels of linguistic nuance and precision. The debut of ChatGPT, a cutting-edge LLM, has catalyzed widespread adoption across diverse applications. Built on the GPT-3.5 architecture, ChatGPT's 175 billion parameters were honed through reinforcement learning from human feedback, yielding remarkable performance. Nevertheless, the escalating scale and intricacy of LLMs, epitomized by ChatGPT's substantial parameter count, have sparked pressing concerns regarding their ecological footprint and energy expenditure. As demand for LLMs continues to surge, developing sustainable, energy-efficient solutions is crucial to mitigating their environmental impact. This paper undertakes a comprehensive examination of energy consumption monitoring, challenges, and areas for improvement necessary for the development of energy efficient Large Language Models.**

***Index Terms*—BERT, Large Language Model, Energy-aware training, GPU, Footprint, GPT-3, Profiling, CO_2 emissions**

I. Introduction

Large Language Models (LLMs) have revolutionized artificial intelligence by enabling groundbreaking applications in natural language processing, such as chatbots, translation systems, content generation, and semantic search. However, the extraordinary capabilities of these models come with a significant trade-off: their energy consumption. Training and deploying LLMs require immense computational resources, which contribute to high electricity use and a substantial carbon footprint. As the global demand for AI-powered solutions grows, the environmental and economic costs associated with running these models are becoming increasingly unsustainable.

Training a state-of-the-art LLM, such as GPT or similar models, often involves billions of parameters and requires petabytes of data processed over weeks or months on power-hungry GPUs or TPUs. For instance, training a model like GPT-3 reportedly consumes as much energy as powering several households for a year, generating tens or even hundreds of metric tons of CO_2 emissions. Beyond training, deploying these models at scale—serving billions of queries daily—adds a continuous energy burden.

The carbon emissions from AI training and inference contribute to climate change, especially when energy is sourced from non-renewable resources. High energy consumption translates to significant operational costs, limiting accessibility for smaller organizations and researchers. As models grow larger and more sophisticated, their energy demands scale disproportionately, challenging the feasibility of developing and deploying next-generation models.

II. Related Work

Recent research has highlighted the significant energy consumption and environmental impact of training large AI models, particularly in the field of Natural Language Processing (NLP) and Large Language Models (LLMs). This section discusses key studies that have quantified these costs and proposed methods for addressing them.

Strubell et al[8] conducted a comprehensive analysis of the computational and environmental costs associated with training deep learning models for NLP tasks. Their study examined four popular off-the-shelf NLP models: Tensor2Tensor (T2T), ELMo, BERT, and GPT-2. The authors estimated the energy consumption and carbon footprint of training these models by measuring power draw during training and extrapolating to full training time. They found that training a single BERT base model on GPUs consumed approximately 1507 kWh and emitted 719 lbs of CO_2 equivalent.

To estimate the CO_2 emissions of GPT-3, we can extrapolate from the findings by [8] regarding BERT, considering key differences in the scale and computational requirements of the models. Strubell's study provided detailed energy consumption and carbon footprint data for BERT base, which consists of 110 million parameters, while GPT-3, with its 175 billion parameters, represents a much larger computational effort.

A. *Key Factors in Estimating GPT-3's* CO_2 *Emissions*

Following are the key factors:

- Model Size: GPT-3 is roughly 1,590 times larger than BERT base in terms of parameters.This scale dramatically increases the computational complexity, as the number of computations required for training scales non-linearly with the number of parameters.
- Training Iterations: BERT was trained on a corpus of 16 GB of text data, while GPT-3 was trained on a significantly larger dataset of 570 GB, approximately 36 times more data. This alone implies that GPT-3's training required far more compute cycles.
- Training Infrastructure: Strubell's analysis assumed the use of GPUs, while OpenAI used supercomputing clusters with thousands of NVIDIA V100 GPUs or similar hardware to train GPT-3.The efficiency of modern hardware could slightly mitigate energy consumption per computation, but the sheer scale of GPT-3's requirements would still dominate the total energy usage.
- Training Time: Training GPT-3 reportedly took weeks or months, with continuous utilization of tens of thousands of GPUs. This represents orders of magnitude more energy use than the relatively smaller-scale BERT training.

We can estimate power draw of GPUs based on these factors. The GPUs used for GPT-3 training, such as NVIDIA V100, have a TDP (thermal design power) of around 300 watts per card. A supercomputing cluster with thousands of such GPUs operating for months would lead to massive energy consumption.

Hence, approximate calculation of GPT-3's CO_2 emissions would look something like this:

1) If BERT base training consumed 1,507 kWh, we can extrapolate based on model size and data scale. GPT-3 is 1,590 times larger, but due to efficiencies in scaling, let's conservatively assume energy consumption grows at a non-linear factor of 1,000.Additionally, GPT-3 uses 36 times more training data than BERT base, further increasing energy demands.Thus, we estimate the energy required for GPT-3 training as: 1,507kWh×1,000×36=54,252,000kWh(conservatively)

2) The carbon footprint depends on the energy source mix for the data center where GPT-3 was trained.Assuming the average carbon intensity of electricity in the U.S. is around 0.41 kg CO_2 per kWh: 54,252,000 kWh x 0.41 kg CO_2 per kWh = 22,243,320 , kg CO_2 (22,243 metric tons)

3) Contextualizing Emissions: 22,243 metric tons of CO_2 is equivalent to driving approximately 4,800 passenger vehicles for a year. The annual energy consumption of around 2,000 U.S. homes.

Even though, the specific figures provided above is only a rough estimate, as the energy consumption and carbon emissions of LLMs can vary widely based on factors such as model architecture, training techniques, hardware efficiency, and energy sources used, It gives us key insight that The jump in energy use and emissions from BERT to GPT-3 underscores the exponential costs of scaling LLMs, emphasizing the urgency of pursuing energy-efficient model architectures.

Given that GPT-4 and other advanced models are even larger than GPT-3, their environmental footprint is likely to be even greater unless significant efficiency gains or renewable energy solutions are implemented.

This analysis highlights the pressing need for energy-efficient training methods, model optimization, and renewable energy integration to make the future of AI development more sustainable. Complementing the work on AI energy consumption, Almeida et al [1] focused on energy monitoring in Ultra Scale Systems (USS).

Their paper discusses the challenges of energy monitoring in large-scale parallel and distributed computing systems, which are often used for training and deploying LLMs. The authors highlight the importance of developing standardized APIs and tools for energy monitoring, which could be adapted for more precise energy monitoring in LLM training and inference.

Patterson et al. [6] conducted a comprehensive study on the emissions and energy consumption of large neural network training, focusing on natural language processing models. Their work, "Emissions and Large Neural Network Training," provides valuable insights into the environmental impact of training large language models and proposes strategies for reducing their carbon footprint.The authors analyzed several recent large models, including T5, Meena, GShard, Switch Transformer, and GPT-3, calculating their energy use and CO_2 equivalent emissions (CO_2e). They identified three key areas for improving energy efficiency: the use of sparsely activated neural networks, strategic geographic placement of ML workloads, and leveraging efficient datacenter infrastructure and ML-oriented accelerators.Patterson et al. demonstrated that combining these strategies could lead to a 100-1000 X reduction in carbon footprint. Their work emphasizes the importance of considering energy consumption and CO_2e as crucial metrics in model evaluation, aligning with our research on energy-efficient large language models. The paper also highlights the need for transparency in reporting energy usage and emissions in ML research, which is a valuable consideration for future work in this field.

III. Main Challenges of Energy Efficient LLMs

Here are the main challenges of energy-efficient Large Language Models:

1) Lack of research : Despite the growing concern about the energy consumption of LLMs, there is a lack of research focused on monitoring and improving the energy efficiency of these models. The development of energy-efficient LLMs requires a comprehensive understanding of the factors that contribute to their energy consumption, including:
   a) Model architecture: The design of the LLM architecture, including the number of layers, parameters, and activations, significantly impacts energy consumption.
   b) Computational intensity: The computational intensity of LLMs, including the number of floating-point operations (FLOPs) and memory accesses, contributes to energy consumption.
   c) Hardware and software platforms: The choice of hardware and software platforms, including GPUs, TPUs, and CPUs, as well as frameworks and libraries, affects energy consumption.
   d) Data characteristics: The characteristics of the input data, including size, complexity, and distribution, influence energy consumption.

2) Lack of transparency from AI companies:This is a significant deterrent to research on energy efficiency in Large Language Models (LLMs). Here are some points highlighting this issue:
   a) Many AI companies develop proprietary LLM architectures and models. This makes it challenging for researchers to study and optimize energy efficiency, as the underlying architectures and models are not publicly disclosed.
   b) Many AI companies use closed-source software stack, which can make it difficult for researchers to optimize energy efficiency. Open-source software,

on the other hand, allow researchers to modify and optimize the code for energy efficiency.

c) AI companies often patent their innovations, including LLM architectures and optimization techniques. This can limit the ability of researchers to study and build upon existing work, hindering progress in energy efficiency research.

3) Limited access to computational resources: AI companies often have access to large-scale computational resources, including custom-built hardware accelerators like TPUs and GPUs. However, these resources are typically not available to researchers, making it difficult to replicate and study the energy efficiency of LLMs.

IV. Potential solutions

1) Shared responsibility:BIG AI companies, such as Google, Facebook, Microsoft, and Amazon, have a significant share of the responsibility for the environmental impact of LLMs. Their collective contribution can help mitigate this issue.Collective contribution from these companies can facilitate standardization and collaboration, enabling the development of more efficient and sustainable LLMs.
2) Energy monitoring and profiling: Creating efficient tools like machine learning emissions calculator [4] and methodologies for monitoring [3] and profiling the energy consumption of LLMs, including hardware-based and software-based approaches.
3) Energy-aware training methods: Developing training methods that incorporate energy efficiency as a primary objective[7], including energy-constrained optimization and green AI.
4) Energy-efficient model design: Investigating techniques for designing energy-efficient LLM architectures, including model pruning, quantization, and knowledge distillation.
5) Nudging Energy Efficiency in Large Language Models using Model Cards: Model cards are documentation that provides detailed information about a Large Language Model’s performance, energy consumption, and environmental impact. By presenting this information in a clear and concise manner or in other words eco-labeling [2], developers and users can be ”nudged” [5] towards making more energy-efficient choices when selecting or designing LLMs.

V. Conclusion

Research in energy-efficient LLMs is essential not only for reducing their environmental and economic costs but also for ensuring equitable access to AI. Without significant strides in this area, the benefits of LLMs may remain restricted to large corporations with the resources to afford their high costs. Moreover, improving energy efficiency aligns with broader global sustainability goals, such as the United Nations’ Sustainable Development Goals (SDGs), particularly those addressing climate action and affordable, clean energy.

By focusing on monitoring and improving the energy efficiency of LLMs, the AI community can ensure that innovation remains sustainable, accessible, and environmentally responsible. As AI continues to shape the future of technology, addressing its energy demands will be critical to maximizing its benefits while minimizing its ecological and societal costs.

References

[1] Francisco Almeida, Marcos Assuncao, Jorge Barbosa, Vicente Blanco, Ivona Brandic, Georges Da Costa, Manuel F. Dolz, Anne Elster, Mateusz Jarus, Helen Karatza, Laurent Lefe`vre, Ilias Mavridis, Ariel Oleksiak, Anne-Ce´cile Orgerie, and Jean-Marc Pierson. Energy monitoring as an essential building block towards sustainable ultrascale systems. *Sustainable Computing: Informatics and Systems*, 17, 11 2017.

[2] Abhijit Banerjee and Barry D. Solomon. Eco-labeling for energy efficiency and sustainability: a meta-evaluation of us programs. *Energy Policy*, 31:109–123, 2003.

[3] Peter Henderson, Jie Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. Towards the systematic reporting of the energy and carbon footprints of machine learning. *ArXiv*, abs/2002.05651, 2020.

[4] Alexandre Lacoste, Alexandra Sasha Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *ArXiv*, abs/1910.09700, 2019.

[5] Richard G. Newell and Juha Siikama¨ki. Nudging energy efficiency behavior: The role of information labels. *Journal of the Association of Environmental and Resource Economists*, 1:555 – 598, 2013.

[6] David A. Patterson, Joseph Gonzalez, Quoc V. Le, Chen Liang, Llu´ıs-Miquel Mungu´ıa, Daniel Rothchild, David R. So, Maud Texier, and Jeff Dean. Carbon emissions and large neural network training. *ArXiv*, abs/2104.10350, 2021.

[7] Roy Schwartz, Jesse Dodge, Noah Smith, and Oren Etzioni. Green ai. *Communications of the ACM*, 63:54–63, 11 2020.

[8] Emma Strubell, Ananya Ganesh, and Andrew Mccallum. Energy and policy considerations for modern deep learning research. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:13693–13696, 04 2020.