International Journal of Science and Engineering Applications
Volume 13-Issue 04, 34 - 39, 2024, ISSN:- 2319 - 7560
DOI:10.7753/IJSEA1304.1008

Analyzing the Impact of Large Language Models on Battery Consumption in Mobile Devices: An Empirical Study

Anton Novikau
Head of mobile Development, Talaera

Abstract:The rise of Large Language Models (LLMs) has led to a significant transformation across various applications, including natural language processing, machine learning, and artificial intelligence. Along with an increasing number of LLMs being launched in the cloud, a significant portion of them are also designed to be embedded in mobile devices. Therefore, their ability to influence battery consumption across mobile devices is going to be crucial. This work aims to empirically investigate the battery consumption of LLMs running on mobile devices across various configurations and runtime scenarios.
We methodically measure the energy usage of several popular LLMs, considering factors such as model size and the complexity of the tasks performed. Our methodology involves a series of controlled experiments with mobile devices running these models under standardized conditions to provide a comparative analysis of their energy efficiency.
Preliminary results indicate significant variances in battery consumption based on the model's operational parameters and the nature of the tasks executed. The study provides insights into the trade-offs between computational demands of LLMs and battery life, offering guidance for developers and researchers in optimizing LLM implementations for mobile environments. Furthermore, we discuss the implications of our findings for the future design and deployment of energy-efficient LLM applications on mobile devices. This research contributes to the emerging discourse on sustainable AI by highlighting the energy considerations of deploying advanced machine learning models in mobile computing contexts.

Keywords: LLM, Performance, iOS, Energy usage, AI

1. INTRODUCTION

1.1 Background information on Large Language Models

Large Language Models (LLMs) represent a significant advancement in the field of artificial intelligence, particularly within the realm of natural language processing (NLP). These models have evolved from simpler machine learning frameworks to complex systems capable of understanding, generating, and translating human language with remarkable accuracy. The development of LLMs has been propelled by increasing computational power and the availability of vast datasets, which are essential for training these models on a wide range of linguistic tasks.

Historically, language models were limited in scope and understanding, often restricted to specific domains or simple text generation. However, the introduction of models like GPT (Generative Pre-trained Transformer) and BERT (Bidirectional Encoder Representations from Transformers) marked a paradigm shift, demonstrating that models could learn from extensive corpuses of text and apply this knowledge across various language tasks. This ability to generalize from training data allows LLMs to perform a range of functions, from answering questions and summarizing texts to creating content and translating languages.
The architecture behind these models, primarily based on the transformer mechanism, enables them to process and interpret sequences of words, capturing nuances in language that were previously difficult for machines to grasp. The size of these models has grown exponentially, with newer versions containing billions, or even trillions, of parameters. This scale contributes to their depth of understanding but also presents challenges in terms of computational requirements and energy consumption.
LLMs have also sparked significant ethical and societal discussions, as their capabilities extend into areas like content creation, conversation, and even impersonation. The implications for privacy, misinformation, and bias are areas of active research and debate. Despite these concerns, the potential of

LLMs to revolutionize industries, education, and communication is undeniable.

The ongoing development and refinement of LLMs continue to push the boundaries of what artificial intelligence can achieve. As researchers address the limitations and ethical considerations of these models, LLMs are set to become even more integrated into our digital lives, transforming our interaction with technology and information.

1.2 Importance of LLMs in current technological landscapes

Large Language Models (LLMs) have become cornerstone technologies in the current technological landscape, profoundly impacting various sectors including healthcare, finance, education, and customer service. Their ability to understand and generate human-like text has revolutionized the way we interact with digital systems, making technology more accessible and user-friendly. LLMs drive the efficiency of search engines, enhance the relevance of content recommendations, and improve the accuracy of language translation services, breaking down language barriers and facilitating global communication.

In the realm of customer service, LLMs have enabled the development of advanced chatbots and virtual assistants, providing users with instant, round-the-clock support and significantly reducing operational costs for businesses. In education, they offer personalized learning experiences, adaptively supporting students based on their individual needs and progress. LLMs also play a crucial role in content creation, aiding writers, marketers, and creators in generating ideas, drafting articles, and producing diverse forms of content at scale.

Moreover, the analytical capabilities of LLMs are leveraged in data analysis and decision-making processes, helping businesses and researchers sift through vast amounts of information to identify trends, make predictions, and inform strategies. As the technology continues to evolve, the importance of LLMs in automating complex tasks, enhancing human productivity, and driving innovation is increasingly becoming a pivotal element of modern digital infrastructure.

1.3 Rationale for exploring the deployment of LLMs on mobile devices

Exploring the deployment of Large Language Models (LLMs) on mobile devices is driven by the goal of making advanced natural language processing capabilities widely accessible in everyday contexts. By integrating LLMs into mobile technology, users can benefit from real-time language translation, context-aware assistance, and sophisticated conversational interfaces, enhancing communication and productivity on the go. Additionally, deploying LLMs on mobile devices addresses the growing demand for instant information retrieval and decision-making support directly from one's pocket, catering to the fast-paced lifestyle of modern society.

Furthermore, mobile integration allows for personalized user experiences, as LLMs can leverage device-specific data to offer tailored advice, recommendations, and interactions. Lastly, expanding LLM capabilities to mobile platforms democratizes access to cutting-edge AI technologies, bridging the digital divide and ensuring a broader range of individuals can benefit from the advancements in artificial intelligence, irrespective of their access to traditional computing resources.

1.4 Brief analysis of computational power required by LLMs

Analyzing the computational power required by Large Language Models (LLMs) in the context of mobile devices reveals significant challenges due to the inherent limitations of mobile hardware. Mobile devices, such as smartphones and tablets, are equipped with processors and memory capacities that are significantly less powerful than those required to run LLMs effectively, which typically necessitate high-performance GPUs or TPUs found in desktop environments or data centers. The disparity in computational resources results in substantial constraints when attempting to deploy full-scale LLMs directly on mobile devices.

The energy efficiency of mobile devices is another critical factor; running LLMs demands substantial energy, which can rapidly deplete battery life, a significant concern for mobile users. Furthermore, the thermal management in mobile devices is not designed for the continuous, high-intensity computation required by LLMs, leading to potential overheating and hardware degradation. Additionally, the storage requirements for state-of-the-art LLMs, which may involve hundreds of gigabytes for the model parameters alone, far exceed the typical storage capacity available on mobile devices.

To mitigate these challenges, researchers and developers have explored various strategies, such as model compression techniques, including quantization, pruning, and knowledge distillation, to reduce the size and computational demands of LLMs without significantly compromising performance. Offloading computation to cloud services is another approach, though it introduces latency and requires a stable internet connection, which may not always be feasible in mobile contexts.

Moreover, the development of specialized hardware accelerators for mobile devices, similar to those used in desktop environments, presents a potential avenue for bridging the computational gap. However, this approach involves trade-offs between cost, energy consumption, and device size. In conclusion, while

the deployment of LLMs on mobile devices poses substantial technical challenges, ongoing research in model optimization and hardware innovation holds the promise of making this a more viable reality in the future.

1.5 Comparison with mobile device capabilities

The deployment of Large Language Models (LLMs) on mobile platforms is impeded by a fundamental mismatch: the substantial computational and storage demands of LLMs versus the constrained capabilities of mobile hardware. Mobile Central Processing Units (CPUs), despite continuous improvements, often falter under the complex computational burdens required for processing LLM inferences. Furthermore, the inherent storage limitations of mobile devices pose significant challenges for accommodating the extensive data footprint typical of LLM architectures.

In response to these challenges, the research community is actively pursuing avenues to reconcile the demands of LLMs with the operational constraints of mobile technology. Key strategies include the optimization of LLM frameworks for mobile contexts through methods such as model pruning and quantization, which aim to streamline the models while preserving their functional integrity. Additionally, the development of mobile-friendly inference engines and the adaptation of algorithms to be more hardware-conscious are under exploration to improve LLM feasibility on mobile platforms. Moreover, the investigation into alternative neural network architectures and advanced model compression techniques is ongoing, aiming to diminish the computational and storage burdens imposed by LLMs.

However, the effective mobilization of LLMs necessitates a nuanced equilibrium between maintaining model efficacy and optimizing for resource frugality. This balance involves evaluating the compromises between executing processes locally on the device versus leveraging cloud-based computational resources. For scenarios relying on cloud interaction, the creation of streamlined data communication protocols and the reduction of network latency are paramount. In essence, the advancement of LLM applications within mobile ecosystems will likely depend on a synergistic approach, fostering collaborative efforts between LLM researchers and mobile hardware developers to surmount the existing barriers and facilitate broader adoption of LLM technologies in mobile settings.

1.6 Impact of hardware limitations on LLM performance and functionality

The impact of hardware limitations on the performance and functionality of Large Language Models (LLMs) in mobile contexts is substantial and multifaceted. Primarily, the constrained computational resources of mobile devices, such as limited processing power and memory, directly affect the speed and efficiency of LLM inference, leading to slower response times and degraded user experiences. The absence of specialized hardware, like GPUs or TPUs in standard mobile devices, further exacerbates this issue, as these components are crucial for the parallel processing capabilities essential for efficient LLM operations.

Furthermore, the restricted memory and storage capacity inherent to mobile platforms pose significant challenges for hosting the extensive parameters of state-of-the-art LLMs, necessitating the use of compressed or simplified models which may not deliver the same level of accuracy or functionality as their full-sized counterparts. This reduction in model complexity can result in compromised linguistic understanding and generation capabilities, affecting tasks such as language translation, content creation, and contextual conversation.

Additionally, the energy consumption associated with running complex computational tasks like those required by LLMs can lead to rapid battery depletion, which is a critical concern for mobile device users. This energy inefficiency not only limits the practicality of deploying LLMs on mobile devices but also poses sustainability concerns. Moreover, the thermal output from continuous, intensive computation can cause devices to overheat, leading to potential hardware damage and further limiting the practical use of LLMs in mobile applications.

The limited bandwidth and higher latency of mobile networks compared to wired connections can also impact the functionality of LLMs, particularly those reliant on cloud-based computation and data access. In essence, the performance and functionality of LLMs on mobile devices are significantly hampered by hardware limitations, necessitating innovative solutions to bridge the gap between the advanced capabilities of LLMs and the practical realities of mobile computing.

1.7 Literature Review

As of today, the topic of Large Language Models (LLMs) is quite popular in the computer science community, and there is a plethora of research available in this field.

One of the major papers in this field is "Improving Language Understanding by Generative Pre-Training"[3] authored by OpenAI's employees. It demonstrates the use of a decoder-style LLM for generative modeling.

Another significant piece of research in the LLM space is "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"[4] by

Google. This article introduces a language representation model called BERT, which is used for models in language inference and question/answer tasks. Prior to this, Google published another research paper titled "Attention Is All You Need,"[5] proposing the Transformer architecture, which has greatly influenced the AI field.

With the release of ChatGPT, OpenAI also published a paper explaining its architecture and proposing that merely making language models larger does not improve them; instead, aligning them with human feedback fine-tunes the aspect of intent in the models[6].

Performance and resource consumption in LLM-related tasks are subjects studied in multiple research papers. A study titled "Full Parameter Fine-tuning for Large Language Models with Limited Resources"[7] focuses on reducing the memory consumption of full parameter fine-tuning.

There is also a conference paper exploring the resource consumption of locally deployed LLMs - "So You Want Your Private LLM at Home?: A Survey and Benchmark of Methods for Efficient GPTs"[8]. This research aligns well with the scope of the current paper and provides insights into GPU memory usage for privately managed LLMs.

2. METHODOLOGY

In this study, we aim to evaluate the impact of running the Phi-2 Large Language Model [9] on an iPhone 12 Pro, studying battery drain under varying computational demands. To facilitate a comprehensive analysis, we will employ Xcode's built-in diagnostic tools, which provide real-time insights into the device's performance metrics, including CPU utilization and battery consumption levels. Additionally, battery usage will be tracked with built-in iOS API.

2.1 MLC LLM

Machine Learning Compilation for Large Language Models (MLC LLM) is an advanced, universal system architecture engineered to facilitate native, on-device deployment of large language models. Its primary goal is to enable developers to build, optimize, and deploy AI models directly on users' devices using advanced ML compilation techniques. MLC LLM leverages the device's GPU to offload intensive computational tasks, significantly boosting processing speed while preserving energy efficiency. [1].

The MLC-LLM project is divided into three main components: model definition, model compilation, and runtime environments.[2]

Defining the model: This process is conducted using Python, enabling the use of various pre-built architectures including Llama (such as Llama2, Vicuna, OpenLlama, Wizard), GPT-NeoX (like RedPajama, Dolly), RNNs (for instance, RWKV), and GPT-J (such as MOSS)[1]. This allows developers to specify models purely through Python, avoiding direct interaction with code generation and runtime processes.

Model compilation: This process is also handled in Python. Models are compiled using the TVM Unity compiler, which allows for the configuration, quantization, and export of the Python-based model into a model library along with quantized weights. Developers can create quantization and optimization algorithms in Python to tailor LLMs for specific use cases.

Platform-native runtimes: MLCChat variants are available for different platforms, including C++ for the command line, JavaScript for the web, Swift for iOS, and Java for Android. These are configured with a JSON file configuration, simplifying the integration of MLC-compiled LLMs into various projects.

MLC LLM supports a wide array of platforms, including Linux, Windows, MacOS, iOS, and Android. Specifically, on iOS and Android, MLC LLM utilizes GPU hardware acceleration—Metal on iOS's A-series GPUs and OpenCL on Android's Mali & Adreno GPUs—enabling developers to leverage the full power of these devices for complex computational tasks.

2.2 Methodology Overview

Experimental Setup: We will configure an iPhone 12 with the latest iOS version to ensure compatibility and optimal performance with the Phi-2 LLM. The device will be set to airplane mode to minimize background processes and ensure that the observed effects are solely due to the LLM's operation. Additionally, all non-essential apps and services will be disabled to prevent interference with the tests.

Implementation of Phi-2 LLM: The Phi-2 LLM will be integrated into a custom iOS application using the MLC LLM framework. This setup will allow us to directly control the LLM's input and output, as well as to monitor its performance on the mobile hardware.

Test Execution: We will conduct two distinct tests to assess the LLM's impact on the device's load and battery life:

a. Short Query Test: In this initial test, the LLM will process a brief input, consisting of a small number of tokens (e.g., a single sentence or question). This input is designed in a way that the model generates relatively short output, approximately up to 30 tokens. This scenario is designed to represent a typical low-demand use case, such as generating a short text response or answering a simple query.

b. Large Query Test: The second test involves a significantly larger input, containing a high number of tokens (e.g., an extensive paragraph or a complex series of questions). The average size of input tokens is 50 tokens, while average size of output tokens is 300. This will simulate a high-demand scenario,

testing the LLM's behavior under substantial computational stress.

At the end of both tests, we will find a battery level delta - a difference between battery level before test started and finished. Both tests will be executed in multiple iterations and average battery level delta will be used further.

Data Analysis: The collected data will be analyzed to compare the effects of the two test scenarios on the device's performance and battery life. We will assess the correlation between input complexity and resource utilization, aiming to identify patterns and potential optimizations for running large language models on mobile devices efficiently.

By systematically evaluating the performance of the Phi-2 LLM on an iPhone 12, this methodology aims to provide insights into the practical implications of deploying advanced AI models on consumer-grade mobile hardware, particularly in terms of energy efficiency and computational demands.

3. EXPERIMENT RESULT & INTERPRETATION

Table 1. Results of running experiment

Test Type	Number of prompts	Mean battery level delta (10 iterations)
Short prompts	20	1%
Large prompts	20	16%

The energy consumption, even for small prompts, is significant and thus presents a considerable limiting factor for deploying local models onto mobile devices.

Even interacting with LLM using a small number of short prompts puts a significant pressure on battery level.

Both small and large prompts demonstrated a high impact on device battery life, suggesting that the computational requirements of LLMs, even at reduced scales, exceed the typical energy efficiency boundaries of current mobile hardware. This raises important questions regarding the practicality of integrating sophisticated AI functionalities directly into mobile devices without compromising user experience due to rapid battery depletion.

Furthermore, the study explores alternative strategies to mitigate these issues, such as optimizing model architecture for energy efficiency, offloading computation to cloud services while maintaining privacy and data security, and developing new hardware specifically designed to support AI applications. The paper also examines the potential of emerging technologies, such as edge computing, to bridge the gap between AI performance demands and mobile device capabilities.

4. CONCLUSION

In conclusion, this study has highlighted significant challenges associated with deploying Large Language Models (LLMs) on mobile devices, primarily due to their substantial energy consumption and computational requirements. Our findings indicate that even small-scale LLMs can drastically impact battery life, posing a severe limitation for their practical application in mobile contexts. Despite the transformative potential of LLMs across various sectors, their integration into mobile technology remains constrained by current hardware capabilities and energy efficiency considerations.

Our research underscores the necessity for developers and researchers to explore innovative solutions that balance the computational demands of LLMs with the operational limitations of mobile devices. Strategies such as model optimization, leveraging cloud services, and the development of specialized hardware emerge as essential components in addressing these challenges. Furthermore, the evolving landscape of edge computing presents a promising avenue for mitigating the energy and computational bottlenecks of mobile LLM deployment.

Ultimately, while the deployment of LLMs on mobile devices offers exciting prospects for enhancing user experiences and accessibility, it is imperative to navigate the trade-offs between advanced AI functionalities and the practical realities of mobile computing. Our study contributes to this ongoing discourse, providing a foundation for future research aimed at harmonizing the expansive capabilities of LLMs with the ubiquitous nature of mobile technology.

5. REFERENCES

[1] Welcome to MLC LLM — mlc-llm 0.1.0 documentation. (n.d.). https://llm.mlc.ai/docs/

[2] Project Overview — mlc-llm 0.1.0 documentation. (n.d.). https://llm.mlc.ai/docs/get_started/project_overview.html

[3] Radford, A., OpenAI, Narasimhan, K., Salimans, T., & Sutskever, I. (n.d.). Improving Language Understanding by Generative Pre-Training. OpenAI. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

[4] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018, October 11). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv.org. https://arxiv.org/abs/1810.04805

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[6] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askill, A., Welinder, P., Christiano, P., Leike, J., . . . OpenAI. (2022). Training language models to follow instructions with human feedback. In OpenAI [Journal-article]. https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf

[7] Lv, K., Yang, Y., Liu, T., Gao, Q., Guo, Q., & Qiu, X. (2023, June 16). Full Parameter Fine-tuning for Large Language Models with Limited Resources. arXiv.org. https://arxiv.org/abs/2306.09782

[8] Tuggener, L., Sager, P., Taoudi-Benchekroun, Y., Grewe, B. F., & Stadelmann, T. (2024). So you want your private LLM at home? : a survey and benchmark of methods for efficient GPTs. ZHAW Digitalcollection. https://doi.org/10.21256/zhaw-30279

[9] Hughes, A. (2023, December 16). Phi-2: The surprising power of small language models - Microsoft Research. Microsoft Research. https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/