Towards Energy-Aware AI Deployment: Investigating the Interplay of Model Quantization and Hardware Platforms

Haoji Bian
Northwestern University
haojibian2027@u.northwestern.edu

Zinan Wang
Northwestern University
zinanwang2025@u.northwestern.edu

Renyuan Lu
Northwestern University
renyuanlu2026@u.northwestern.edu

Abstract

The increasing adoption of deep learning models, particularly large language models (LLMs), across a wide spectrum of hardware platforms, from resource-constrained embedded devices to powerful cloud-based accelerators, necessitates a critical focus on energy efficiency during inference. Optimizing energy consumption is crucial for sustainability, reducing operational costs, and enabling broader deployment on edge devices with limited power budgets. This research project aims to investigate the key factors influencing AI model inference energy efficiency, focusing on the interplay between model-level optimization techniques, specifically quantization, and the characteristics of the underlying hardware platforms. By synthesizing empirical findings from two distinct studies – one examining quantization's impact on an embedded system and another evaluating LLM efficiency across various GPU architectures – we will provide insights into energy-performance trade-offs and inform strategies for energy-aware AI deployment across a diverse hardware landscape.

ACM Reference Format:
Haoji Bian, Zinan Wang, and Renyuan Lu. 2025. Towards Energy-Aware AI Deployment: Investigating the Interplay of Model Quantization and Hardware Platforms. In . ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Deep learning models have become ubiquitous, powering applications from mobile assistants to complex data analytics. The inference phase of these models, despite being computationally less expensive than training, accounts for a significant portion of their total energy footprint due to its high frequency of execution in real-world scenarios. This energy consumption presents challenges for environmental sustainability and limits the deployment of sophisticated AI models on power-sensitive edge devices.

Achieving energy-efficient AI inference requires understanding and optimizing various influencing factors. These factors span both the software (model design, optimization techniques) and hardware (architecture, capabilities) domains. Model quantization, which reduces the precision of model parameters and computations, is a promising software technique for reducing computational cost and memory footprint, thereby potentially lowering energy consumption. Simultaneously, the choice of hardware platform, ranging from low-power microcontrollers to high-performance GPUs, fundamentally dictates the computational resources and energy profile available for inference.

This project addresses the critical need to understand how these factors interact across different computational environments. Specifically, we propose to investigate the interplay between model quantization and hardware platforms in determining AI model inference energy efficiency. We will draw upon the results of two prior empirical studies conducted by the team. The first study explored the direct impact of varying quantization levels on the energy consumption and performance of a deep learning model running on an embedded system (Raspberry Pi). The second study evaluated the energy efficiency of larger language models across a range of high-performance GPU accelerators, analyzing the influence of different hardware architectures and model characteristics like knowledge distillation.

By synthesizing the findings from these two investigations, our project aims to provide a more holistic perspective on AI energy efficiency. We seek to illustrate how model optimization techniques like quantization manifest in terms of energy savings on resource-constrained hardware, and how the underlying hardware platform significantly shapes the energy landscape for more complex models like LLMs. This understanding is vital for developing strategies to deploy AI models efficiently and sustainably across the ever-expanding spectrum of computing devices.

2 Proposed Work / Methodology

This research project will be conducted by synthesizing and analyzing the empirical data and findings from two distinct, previously completed investigations. No new experimental data collection will be performed. The core methodology involves a structured analysis and integration of the existing results to address the project's research questions regarding the interplay of model quantization and hardware platforms on AI energy efficiency.

The two investigations providing the data are:

- **Investigation 1: Quantization on an Embedded Platform**. This study focused on quantifying the energy consumption and performance of a deep learning model (specifically, word embedding algorithms and TinyBERT as a representative model) subjected to different levels of bit quantization (FP32, FP16, FP8, FP4). The experiments were conducted on a Raspberry Pi, representing a typical resource-constrained embedded platform. Energy consumption was measured directly using external hardware (Joulescope JS220), alongside monitoring of CPU and memory usage. Performance was evaluated based on task accuracy. The findings

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

from this investigation provide concrete evidence of the energy-saving potential and performance trade-offs of quantization in an embedded context.
- **Investigation 2: LLM Energy Efficiency on GPU Platforms**. This study focused on evaluating the energy efficiency of larger language models (7B parameter scale LLMs) across a variety of high-performance GPU accelerators, including different NVIDIA architectures (e.g., Ada Lovelace, Ampere, Volta) and types (consumer and professional cards). The study introduced and utilized metrics such as Energy-to-Output Ratio (EOR) and Time-Weighted Energy-to-Output Ratio (TWEOR) to quantify the relationship between model performance (evaluated on standard benchmarks) and energy consumption (monitored using tools like NVIDIA-SMI). This investigation provides insights into the significant impact of hardware architecture, as well as model characteristics like knowledge distillation, on the energy efficiency of LLMs running on more powerful platforms.

The project's methodology will involve:

- Data Synthesis and Consolidation: Review and consolidate the experimental data and results obtained from Investigation 1 (quantization, embedded energy, performance, resource usage) and Investigation 2 (LLM energy efficiency, hardware comparisons, EOR/TWEOR metrics, model variations on GPUs).
- **Factor-Based Analysis:** Structure the analysis around the key factors identified in the project theme:
  - Analyze the impact of **Model Quantization** on energy efficiency, primarily drawing from the results of Investigation 1, discussing the observed trade-offs on an embedded platform.
  - Analyze the impact of **Hardware Platform** characteristics on energy efficiency, primarily drawing from the results of Investigation 2, discussing how different GPU architectures and capabilities influence LLM energy consumption and performance per watt.
  - Discuss the influence of other **Model Characteristics** (beyond basic quantization), such as knowledge distillation and architectural differences, on energy efficiency, drawing insights from Investigation 2.
- **Interplay Exploration:** Explore the interactions between these factors. Discuss how the effectiveness of quantization might be influenced by hardware capabilities (e.g., native support for lower precision). Analyze how the choice of hardware platform dictates which model optimizations (quantization, distillation, etc.) are most impactful for energy savings. Contrast the energy efficiency landscape and optimization priorities between the resource-constrained embedded environment and the high-performance GPU environment based on the synthesized findings.
- **Discussion and Insights:** Provide a comprehensive discussion of the findings, highlighting the nuances of AI energy efficiency across the hardware spectrum. Discuss the implications for developers and engineers in selecting appropriate models, optimization techniques, and hardware for energy-aware AI deployment in various scenarios.
- **Report Writing:** Compile the analysis and discussion into a research report following the specified length requirements (3-3.5 pages minimum, excluding references) and formatting guidelines (Sigconf).

2.1 Expected Outcomes

The expected outcomes of this research project include:

- A consolidated analysis illustrating the distinct yet complementary insights into AI model energy efficiency provided by the two prior investigations.
- A clear presentation of the empirical evidence demonstrating the energy-saving potential and performance trade-offs of model quantization on an embedded platform.
- A detailed analysis showcasing the impact of different hardware platforms and LLM characteristics on energy efficiency in high-performance computing environments, utilizing metrics like EOR.
- A comprehensive discussion on the interplay between model-level optimizations (quantization) and hardware platforms, highlighting the varying importance and effects of different factors across the hardware spectrum (embedded vs. GPU).
- Practical insights and considerations for deploying energy-efficient AI models tailored to the capabilities and constraints of different hardware platforms.
- A final research report summarizing these findings and contributing to the understanding of energy-aware AI deployment strategies.