Zeus Inference on Mobile: Enhancing Energy Efficiency

Ante Tonkovic-Capin, Ashu Kumar, Drishan Poovaya, Pratik Sanghavi

8th March 2024

1 Introduction

Our project, Zeus Inference on Mobile [14] focuses on the problem of energy efficiency during AI inference on mobile devices. Accordingly, our primary goal is to explore the feasibility of measuring and tuning GPU usage and energy consumption during inference to enable, and be able to target, the most efficient energy usage on mobile devices. As AI powered breakthroughs and innovations continue to permeate every aspect of modern life, the demand for access is only growing. Currently, there's an estimated 17 billion mobile devices in operation worldwide[7], an average of about 2 devices for every person on the planet.

Enabling efficient energy consumption across all these devices is critical for many reasons. First, allowing efficient usage and power consumption is crucial to the performance and accuracy of the models themselves[3]. Models often require input data to be pre-processed on-device with tasks like feature extraction, down-sampling, pooling, multi-modal transformations and many more tasks to ensure the models they feed provide reliable, consistent, timely and accurate responses[15]. Second, and perhaps more critically, enabling efficient power consumption during AI inference is essential to the long-term sustainability of the technological breakthroughs being powered by advances in all fields of Artificial Intelligence[3]. Ensuring that these breakthroughs are accompanied by efficient energy-usage and a commensurate carbon footprint should be seen as a responsibility shared by all researchers and advocates of the technology.

Given how crucial the issue is, there have been numerous studies into the the energy usage and efficiency of models during training, yet attention on inference is lacking[8]. This may sound counter-intuitive, as most of the work and focus has been on efficient GPU usage and power-consumption during model training rather than inference[6]. However, several sources, including the likes of NVIDIA and Amazon, estimate that inference can exceed the cost of training in popular LLMs, with inference accounting for up to 90% of the costs for deployed models[4, 1]. As we celebrate the exponential growth and power of AI models, it is crucial that their energy usage, their carbon footprint, does not grow exponentially with them.

With this objective, we find an opportunity in energy consumption specifically during the inference phase on mobile devices. An energy efficient approach that is both real-time adaptive and performance, will unlock avenues for more powerful model deployments on the edge.

By adapting Zeus[14] for mobile energy tuning during inference, we solve the twin problems of energy efficiency and battery longevity paving way for energy conscious systems. The key innovation lies in leveraging the existing capabilities of mobile GPUs, particularly those of Qualcomm devices, to dynamically adjust processing frequencies. By integrating a modified version of Zeus with Qualcomm's APIs, we aim to create a system that can intelligently adjust power usage based on the device's current state and the demands of the AI model being run.

The reason for choosing Zeus for this endeavour is multi-fold. The real-time adaptive nature saves us from profiling runs[12] or static predictive model training[13] and can be generalized to run on several models under different hardware load conditions. With programmer guided prioritization of performance vs energy efficiency, we can allow application to dictate the inference latency - whats not to like?

We anticipate that our approach will enable a more optimal utilization of the device's resources, balancing power consumption with computational performance. The initial phase of our project will focus on leveraging Qualcomm's APIs to dynamically adjust GPU power consumption during inference tasks. Despite the potential increase in inference time, we expect the benefits in terms of energy savings to significantly outweigh the slight decrease in performance. This hypothesis will be tested through experiments using Qualcomm's Hardware Development Kits (HDK) and a selection of pre-trained neural network models, simulating real-world inference scenarios. Should it prove infeasible to adjust the power level of mobile GPUs, we plan to pivot and test our hypothesis by tweaking the power level of CPU-based devices.

2 Related Work

The literature review will categorize existing research into three main buckets: model-level optimizations, energy-efficient AI model training and model selection strategies for inference on general-purpose computing devices as well as popular accelerators. This will allow us to explore the range of methodologies employed and the resulting outcomes achieved. By examining the state of energy efficiency in AI applications, we can lay the groundwork for finding opportunities for mobile efficient inference jobs.

The earliest methods[2] that sought to achieve this relied on static and dynamic optimization techniques. Static techniques such as quantization, pruning, knowledge distillation and collaborative inference reduce time, memory and energy requirements but they are input-independent causing such deployments to miss the less salient inputs due to their simplified representations. Dynamic optimizations solve this but place the burden of determining the correct technique on the engineer in addition to placing additional demands on memory. Since edge devices are resource constrained, striking a balance between computational efficiency and model accuracy is crucial for successful deployment in edge computing scenarios.

Modern body of work takes a different approach. A recurring theme of papers such as [5, 9, 14] is that although a tradeoff between energy consumption and performance exists, such a tradeoff is not linear. This non-linearity can be leveraged to obtain energy efficient, model training and serving.

Other popular techniques focus on selecting the right model for the right inference target at run time, as in Mobile Deep Inference(MODI)[11] or by performing the inference itself on the cloud, as in CNNSelect[10] . By varying model size and the batching strategy, we're changing the power draw by the core act of inference. However, this does place a requirement for a remotely located repository of models.

While the first group addresses model training which is run only periodically, the second group of papers addresses inference but needs to sync with a remotely located server to get the appropriate model and/or inference results which isn't suitable for the more latency/privacy sensitive workloads.

There is a third group which includes polythrottle[13] which addresses both these pain points. It relies on a static optimization approach as opposed to a more adaptive strategy that responds to real-time conditions. This raises the question: why not explore a hybrid strategy that integrates the merits of both approaches?

3 Timeline and Evaluation Plan

For evaluating our project we plan to do the following:

- Utilize Qualcomm's HDK and a suite of pre-trained models to conduct thorough testing.
- Quantify the trade-offs between energy consumption and inference performance, aiming to establish benchmarks for energy-efficient AI on mobile devices.

Table 1: Timeline for the Project

Action	Time
Feasibility of adjusting GPU power on device in situ	1-2 weeks
Develop run experiment/process to test GPU power usage during inference given available APIs	2-4 weeks
Modify Zeus for Mobile to explore possibility of dynamic adjustment during inference	1-2 weeks
Findings and analysis report results	1-2 weeks
Time permitting look at Poly Throttle integration	TBD

4 Conclusion

By extending Zeus to support energy-efficient neural network inference on mobile devices, this project addresses a critical need in the era of mobile AI. Through innovative adaptations and leveraging existing technologies, we aim to pave the way for more sustainable and efficient AI applications, making a significant contribution to both the field of artificial intelligence and environmental conservation.

References

[1] J. Barr. *Amazon ec2 update–inf1 instances with aws inferentia chips for high performance costeffective inferencing*.

[2] F. Daghero, D.J. Pagliari, and M. Poncino. "Energy-efficient deep learning inference on Edge Devices". In: *Advances in Computers* (2021), pp. 247–301. DOI: 10.1016/bs.adcom.2020.07.002.

[3] Radosvet Desislavov, Fernando Martínez-Plumed, and José Hernández-Orallo. "Trends in AI inference energy consumption: Beyond the performance-vs-parameter laws of deep learning". In: *Sustainable Computing: Informatics and Systems* 38 (2023), p. 100857. ISSN: 2210-5379. DOI: https://doi.org/10.1016/j.suscom.2023.100857. URL: https://www.sciencedirect.com/science/article/pii/S2210537923000124.

[4] McDonald J. et al. "Great power, great responsibility: Recommendations for reducing energy for training language models". In: (2022).

[5] A. Krzywaniak, P. Czarnul, and J. Proficz. "Dynamic GPU power capping with online performance tracing for Energy Efficient GPU computing using Depo Tool". In: *Future Generation Computer Systems* 145 (2023), pp. 396–414. DOI: 10.1016/j.future.2023.03.041.

[6] Anthony L.F.W., Kanding B., and Selvan R. "Carbontracker: Tracking and predicting the carbon footprint of training deep learning models". In: (2020).

[7] Federica Laricchia. *Number of Mobile Devices Worldwide 2020-2025*. https://www.statista.com/statistics/245501/multiple-mobile-device-ownership-worldwide/. Mar. 2023.

[8] D. Li et al. "Evaluating the Energy Efficiency of Deep Convolutional Neural Networks on CPUs and GPUs". In: *2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom), BDCloud-SocialCom-SustainCom*. 2016, pp. 477–484. DOI: 10.1109/BDCloud-SocialCom-SustainCom.2016.76.

[9] K. Ma et al. "GreenGPU: A Holistic Approach to Energy Efficiency in GPU-CPU Heterogeneous Architectures". In: *2012 41st International Conference on Parallel Processing*. Pittsburgh, PA, USA, 2012, pp. 48–57. DOI: 10.1109/ICPP.2012.31.

[10] Samuel S. Ogden and Tian Guo. "Characterizing the Deep Neural Networks Inference Performance of Mobile Applications". In: *CoRR* (2019). eprint: 1909.04783. URL: http://arxiv.org/abs/1909.04783.

[11] Samuel S. Ogden and Tian Guo. "MODI: Mobile Deep Inference Made Efficient by Edge Computing". In: (2018). URL: https://www.usenix.org/conference/hotedge18/presentation/ogden.

[12] Yasuhiro Watashiba, Yuki Matsui, and Susumu Date. "Evaluation of Resource Management System for InfaaS-adaptive Disaster Management Application Platform". In: *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*. 2020, pp. 987–992. DOI: 10.23919/MIPRO48935.2020.9245080.

[13] Minghao Yan, Hongyi Wang, and Shivaram Venkataraman. "PolyThrottle: Energy-efficient Neural Network Inference on Edge Devices". In: *ArXiv* abs/2310.19991 (2023). URL: https://api.semanticscholar.org/CorpusID:264824281.

[14] J. You, J. W. Chung, and M. Chowdhury. "Zeus: Understanding and Optimizing GPU Energy Consumption of DNN Training". In: *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*. 2023, pp. 119–139.

[15] C. V. G. Zelaya. "Towards explaining the effects of data preprocessing on machine learning". In: *2019 IEEE 35th international conference on data engineering (ICDE)*. IEEE. Apr. 2019, pp. 2086–2090.