IPA-Ext

Sabah S. Anis
Computer Science
ML Engineer

Misagh Soltani
Computer Science
ML Research Scientist,
ML Engineer

Xeerak Muhammad
Computer Science
ML Engineer, Scribe, Team
Lead

Problem Statement

- The InfAdapter system lacks reporting on energy consumption across different experiments
- Lacks energy efficiency benchmarks comparing against other inference-serving systems
- This limits the evaluation of InfAdapter's overall effectiveness and sustainability in real-world deployments
- Addressing this gap is crucial for meeting industry standard sustainable AI/ML deployment
- Experimenting with integration of RL (Q-Learning) into the existing framework

Feature	MS [38]	INFaaS [30]	Cocktail [20]	VPA [9]	InfAdapter
Cost Optimization	✗	✓	✓*	✓	✓
Accuracy Maximization	✓	✗	✓	✗	✓
Predictive Decision-Making	✓	✗	✓	✓	✓
Container as a Service (CaaS)	✗	✗	✗	✓	✓
Latency SLO-aware	✓	✓	✓	✗	✓

Feature Comparison Table

Salmani, M., Ghafouri, S., Sanaee, A., Razavi, K., Mühlhäuser, M., Doyle, J., ... & Sharifi, M. (2023, May). Reconciling high accuracy, cost-efficiency, and low latency of inference serving systems. In Proceedings of the 3rd Workshop on Machine Learning and Systems (pp. 78-86).

Technical Challenges

- Power usage varies during different stages in a pipeline thus fine-grained analysis is required to give an accurate energy profile of the pipeline
- Profiling tools may have limitations in some cases
- Integrating real-time energy consumption adaptations to an inference pipeline can be challenging to optimize
- Integrating RL into the framework poses challenges regarding the formulation of the problem in an effective way.
- Early iterations of the Q-Learning optimizer can face technical failures and required repeated refinement.

Related Works

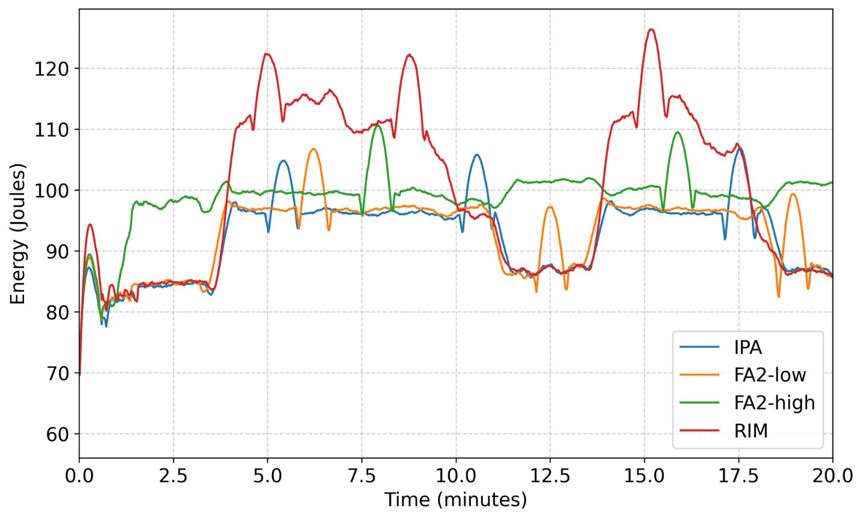
- Existing tools like InferLine, Loki, AutoInfer, and Swayam overlook direct energy and carbon impact.
- Clover [1], introduces carbon-aware inference by mixing high- and low-quality models and optimizing GPU resource partitioning, reducing carbon emissions while meeting SLAs.
- SPROUT [2] tailors sustainability to generative LLM inference, using "generation directives" to curb carbon footprint without degrading generation quality.
- DeepLine [3] leverages deep reinforcement learning and hierarchical action filtering to efficiently generate high-performance machine learning pipelines.
- The approach introduced in 'Reinforcement Learning for Multi-Objective AutoML' [4] focuses on optimizing AutoML pipelines by addressing trade-offs between competing objectives such as accuracy and computational efficiency.
- Our approach directly integrates energy profiling into IPA for a flexible optimization framework that dynamically balanced accuracy, cost, and energy efficiency.
- Our Reinforcement Learning extension adds Q-learning into IPA which enables dynamic adaptation to changing computational and accuracy requirements by optimizing model variants and configurations.

Our Approach and Results

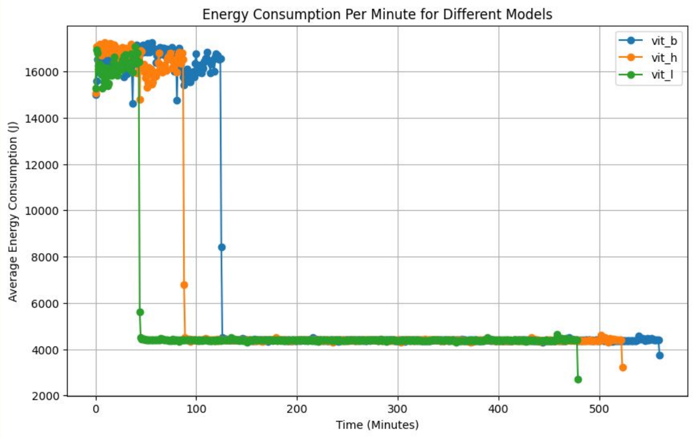
1. Monitored energy consumption using perf library

2. Monitored energy consumption for three-weights of a zero-shot segmentation method (SAM)

3. Attempted to optimize for energy in addition to latency, accuracy, and cost



Energy measurement of fluctuating workload



Energy measurement of three one-shot segmentation methods

Integrating Q-Learning

Q-Learning operates through a state-action-reward framework:

- **State Space:** Represents pipeline configurations including model variants, replica counts, and batch sizes.
- **Action Space:** Defines potential adjustments like switching models, modifying replicas, or changing batch sizes.
- **Reward Function:** Designed to optimize accuracy, minimize resource usage, adhere to SLAs, and balance throughput and latency.

Algorithm 1 Q-Learning for Pipeline Optimization

Initialize Q(s, a) $\leftarrow 0 \quad \forall s \in \mathcal{S}, a \in \mathcal{A}$
Set parameters:
  $\alpha_q \leftarrow 0.1$ ▷ Learning Rate
  $\gamma_q \leftarrow 0.9$ ▷ Discount Factor
  $\epsilon \leftarrow 1.0$ ▷ Initial Exploration Rate
  $\epsilon_{\min} \leftarrow 0.01$ ▷ Minimum Exploration Rate
  $\kappa \leftarrow 0.001$ ▷ Exploration Decay Rate
  $N \leftarrow$ Number of Episodes
  $T \leftarrow$ Maximum Steps per Episode
for episode $\leftarrow 1$ **to** N **do**
  Initialize state s randomly
  for step $\leftarrow 1$ **to** T **do**
    if Random number $> \epsilon$ **then**
      Choose action $a \leftarrow \arg\max_{a'} Q(s, a')$
    else
      Choose random action $a \in \mathcal{A}$
    end if
    Execute action a, observe next state s' and reward r
    if Constraints violated in s' **then**
      $r \leftarrow -1000$
    else
      $r \leftarrow \alpha \cdot \text{Accuracy}(s') - \beta \cdot \text{Resource}(s') - \gamma \cdot \text{Batch}(s')$
    end if
    Update Q-value:
  $Q(s, a) \leftarrow Q(s, a) + \alpha_q \cdot [r + \gamma_q \cdot \max_{a'} Q(s', a') - Q(s, a)]$
    $s \leftarrow s'$
    Decay exploration rate:
  $\epsilon \leftarrow \max(\epsilon_{\min}, \epsilon \cdot e^{-\kappa \cdot \text{episode}})$
    if Termination condition met **then**
      break
    end if
  end for
end for
Extract Optimal Policy π^*:
for each state $s \in \mathcal{S}$ **do**
  $\pi^*(s) \leftarrow \arg\max_a Q(s, a)$
end for
Return π^*

Broader Impacts

- Having scripts to monitor and analyze energy consumption can be generalized to other inference pipelines
- Can help developers track and optimize their carbon footprint

References

[1] Li, B., Samsi, S., Gadepally, V., & Tiwari, D. (2023, November). Clover: Toward sustainable ai with carbon-aware machine learning inference service. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (pp. 1-15).

[2] Li, B., Jiang, Y., Gadepally, V., & Tiwari, D. (2024). Toward sustainable genai using generation directives for carbon-friendly large language model inference. arXiv preprint arXiv:2403.12900.

[3] Yuval Heffetz, Roman Vainshtein, Gilad Katz, and Lior Rokach. 2020. DeepLine: AutoML Tool for Pipelines Generation using Deep Reinforcement Learning and Hierarchical Actions Filtering. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '20). Association for Computing Machinery, New York, NY, USA, 2103–2113. https://doi.org/10.1145/3394486.3403261

[4] Armin Dadras Eslamlou, Shiping Huang, Reinforcement learning for multi-objective AutoML in vision-based structural health monitoring, Automation in Construction, Volume 166, 2024, 105593, ISSN 0926-5805, https://doi.org/10.1016/j.autcon.2024.105593