Towards Sustainable AI with IPA: Experiments in Energy Consumption of Machine Learning Models

Regan Willis, Chase Bryson, Osasuyi Agho

Abstract

Energy usage is becoming an important topic in many computer science research areas, but especially in machine learning. However, other factors such as accuracy, latency, and cost are also critical and commonly take priority. After all, if a machine learning pipeline is not reasonably accurate enough, fast enough, and cheap enough, the energy consumption will not matter because the pipeline will not be used. We analyze energy consumption as it compares to both accuracy and latency to expand this trade-off space. We use the tradeoff space outlined by the Inference Pipeline Adaptation system (IPA) (Ghafouri et al., 2024). We have found that energy consumption is not proportional to latency, and that tuning the batch size can help find the model configuration that takes slightly longer but is more energy efficient. Energy consumption can be included as a tunable metric in IPA to increase energy awareness and reduce energy consumption for machine learning pipelines that use IPA.

1. Introduction

With the recent Intergovernmental Panel on Climate Change's (IPCC) reports showing that 1.1 degrees C of global warming has already caused changes to Earth's climate that affect every region of the world (IPCC, 2023), sustainability is more important than ever. At the same time, machine learning models are being used more and more widely, with no sign of slowing down. Machine learning models are known for their high energy usage. This has led to a growing sub-field of research known as Sustainable AI. Sustainable AI is not just about using machine learning to create sustainability solutions, it is also about reducing the energy consumption of AI models that serve a variety of applications that may or may not be energy related (Wynsberghe, 2021). We apply this concept of "sustainability *of* AI" to the Inference Pipeline Adaptation system (IPA) (Ghafouri et al., 2024). IPA is a framework that can be applied to any multi-model inference pipeline. It balances accuracy, latency, and cost trade-offs to create the optimal user experience (Ghafouri et al., 2024). However, it does not currently consider energy consumption. We conducted experiments to determine if energy consumption expands IPA's trade-off space. The solver should choose an optimal model for accuracy, latency, and energy consumption, with these factors weighted according to developer preferences. Less energy-conscious developers will not be harmed by allowing a more energy efficient model to be used when it is at no cost to their Service Level Agreements (SLAs). More energy-conscious developers should be empowered to prioritize reducing energy consumption, choosing slightly slower or less-accurate models if necessary.

Our experiments are conducted using YOLOv7, a widely-used object detection architecture (Wang et al., 2022). We measured energy consumption as it compared to accuracy and latency over different sets of model weights (trained and provided by the YOLO developers to target different speeds and accuracies) (Wang et al., 2022) and batch sizes. Our first experiment found that latency and energy consumption were highly correlated. We created a second experiment to investigate the effect of batch size more deeply and explore the concept of energy proportionality. We found that energy consumption is a separate metric to both accuracy and latency, providing different information that expands the trade-off space.

2. Related Work

This section provides information on the works related to our experiments.

IPA: Inference Pipeline Adaptation to Achieve High Accuracy and Cost-Efficiency We based our experiment off of the IPA framework (Ghafouri et al., 2024) and expanded the trade-off space by adding energy consumption to IPA's three-dimensional trade-off space consisting of accuracy, latency, and cost.

RAPL in Action: Experiences in Using RAPL for Power Measurements Power meters are a reliable way of measuring energy consumption, but they are complex to deploy, especially with cloud computers, which are commonly used for compute-intensive processes such as machine learning, where a developer may not have physical access to the machine. Software methods of measuring energy consumption are much simpler and are accurate enough for general purposes (Khan et al., 2018). RAPL is the tool used to provide the energy metrics used in this paper.

YOLOv7 YOLOv7 is a recent improvement on the You Only Look Once (YOLO) object detection models. It is considered state-of-the-art and is widely used. We use this object detector for our experiments in this paper.

3. Data

This section provides information on the data that was used for inferencing during the experiments.

Common Objects in Context For our experiments we used the Common Objects in Context (COCO) dataset (Lin et al., 2014). The COCO 2017 dataset contains 123,287 images with 886,284 instances of 80 different object classes. It is widely used for benchmarking (Lin et al., 2017). We ran the models on this data while measuring energy consumption.

For Experiment 1 we attempted to load 500 images of the validation set. 496 images were loaded successfully and were used in the experiment.

For Experiment 2 we used the full validation set from COCO 2017, almost 5000 images. We expanded the test dataset so that the models would run for longer, providing a more accurate energy consumption metric.

4. Methods

This section provides a description of the various methods that were used throughout the experiments.

Energy Consumption Measurement We measured energy consumption using perf, a tool for Linux operating systems that can statistically profile the system, including energy consumption estimates (Linux Foundation, 2024). As stated by Intel, energy consumption metrics are calculated using the Running Average Power Limit (RAPL) provided by the CPU (Intel, 2022). RAPL is a measurement of the transistors being switched on the processor, which uses energy. It has been shown by Khan et al. (2018) that "RAPL readings are highly correlated with plug power" and "have negligible performance overhead", meaning they are accurate enough to be used reliably for energy measurement on their own. We have found that perf is the most accurate energy consumption measurement tool we can use while still being reasonably easy to use.

Accuracy Measurement The YOLOv7 code repository which we used to run the models returns both precision and recall metrics for each trial. We used these metrics to calculate the F_1 score with the below equation:

Equation 1. Calculation of F_1 score (Taha et al., 2015).

$$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = 2\frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} = \frac{2\mathrm{tp}}{2\mathrm{tp} + \mathrm{fp} + \mathrm{fn}}.$$

This F_1 score was used for our comparison.

Latency Measurement Speed information was returned by the YOLOv7 repository in the form of milliseconds per inference. This metric was used for our experiment as-is.

Hardware We use hardware in the Chameleon UC datacenter, specifically a compute_cascadelake_r_ib node recommended by IPA. The Chameleon Hardware Discovery page (Chameleon) shows these machines to have x86_64 processors with 2 CPUs, 96 threads, and 192GiB of RAM (Chameleon, 89e48f7e-d04f-4455-b093-2a4318fb7026 (P3-CPU-038)). The processors are Intel(R) Xeon(R) Gold 6240R CPU @2.40GHz.

Software We use a combination of bash scripting and Python scripting to run the experiments and plot the results. The Python plotting scripts read from a log file that is output from the bash script. Plotting is done using Matplotlib (Matplotlib development team, 2012 - 2024).

5. Experiments

This section provides a description of the experiments that were conducted to determine if energy consumption expands the IPA trade-off space. Two experiments were conducted.

5.1 Experiment One: Energy Consumption vs. Accuracy vs. Latency

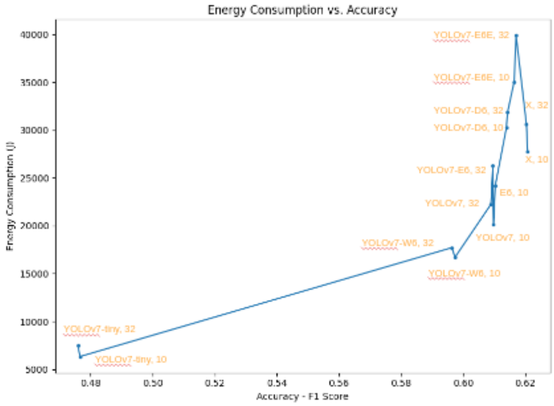
For our first experiment, we ran multiple trials of YOLOv7 with seven different weights and two different batch sizes. The goal of this experiment was to determine if there was a correlation between energy consumption and accuracy or energy consumption and latency. The batch sizes were 10 and 32. The batch size 32 was chosen because it was used throughout the YOLOv7 documentation, and the batch size of 10 was arbitrarily chosen as a comparison to 32. The model weights used in the experiment are shown in Table 1 below.

Table 1. The YOLOv7 models used in the experiments, including their average precision (AP) and average latency for a batch size of 32 (Wang et al., 2022).

Model	AP	Batch 32 average time (ms)
YOLOv7	51.4%	2.8
YOLOv7-X	53.1%	4.3
YOLOv7-W6	54.9%	7.6
YOLOv7-E6	56.0%	12.3
YOLOv7-D6	56.6%	15.0
YOLOv7-E6E	56.8%	18.7
YOLOv7-tiny	--	--

Accuracy Our results show that as the accuracy of models increases, there is an upward trend in energy consumption. However, the two metrics are not exactly correlated. As we can see in Figure 1, the YOLOv7-X weights have the highest accuracy, but have much lower energy consumption than the YOLOv7-E6E weights, which are slightly less accurate.

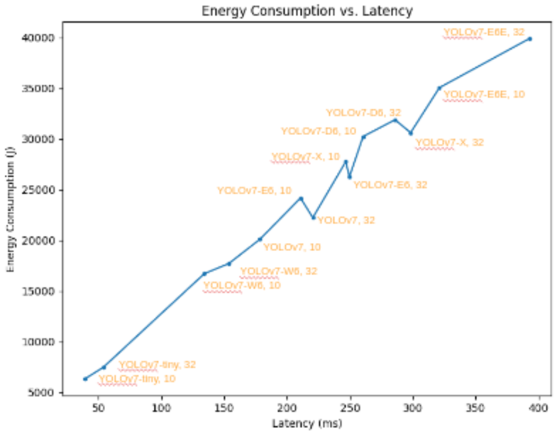
Figure 1. Plot of the relationship between energy consumption and accuracy resulting from experiment 1.



YOLOv7-E6 is also more accurate than the YOLOv7 weights, but their energy consumption is quite similar. We can also see that a smaller batch size generally uses less energy. From these results we conclude that accuracy and energy consumption are not redundant metrics.

Latency The comparison between energy consumption and latency was less conclusive. The energy consumption of model weights and batch sizes appear highly correlated with latency, as shown in Figure 2. Bigger models and bigger batch sizes generally consume more energy and inference at lower speeds. These results appear to show that energy consumption does not expand the trade-off space, as it will correlate with latency very closely.

Figure 2. Plot of the relationship between energy consumption and latency resulting from experiment 1.



5.2 Experiment Two: Energy Consumption vs. Latency

Our results from the first experiment necessitated a deeper analysis of the relationship between energy consumption and latency. The purpose of this experiment is to uncover more about how these two factors interact and determine if energy consumption expands a trade-off space that includes latency.

Design For this analysis, we focused on batch size and explored the concept of energy proportionality. **Energy proportionality** compares the amount of power being consumed by a computer with its utilization. Power and utilization are not proportional, meaning higher CPU utilization is more energy efficient (Green Software Foundation, 2024). In machine learning, and with IPA, the batch size for inferencing is a tunable metric. **Batch size** refers to the amount of data inputs that will be inferenced together. Increasing batch size will increase the utilization of the processor, decreasing the energy consumption per inference. However, increasing the batch size will increase latency, as every instance in the batch will need to wait for every other instance to be inferenced before it can be returned to the user.

We designed an experiment that focuses on this trade-off between batch size and energy usage. We used one constant set of model weights and 11 different batch sizes ranging from 5 to 1000. We also chose a large set of model weights (YOLOv7-E6E) to force the trials to run longer which increases the accuracy of the energy consumption metric.

Results Figure 3 shows that as batch size increased, latency (per input) increased, as expected. Energy consumption also generally increased, with exceptions at batch size 100, 150, and 300, which decreased slightly. The concept of energy proportionality is demonstrated, with slight changes in smaller batch sizes causing large changes in energy consumption, while larger batch sizes caused a smaller change in energy consumption.

Figure 3. Plot of the relationship between energy consumption and latency (per input) resulting from experiment 2.

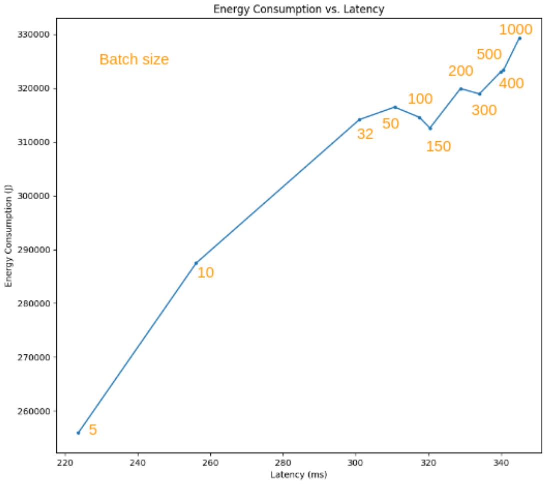


Figure 4. Plot of the relationship between energy consumption (per input) and latency (per input) resulting from experiment 2.

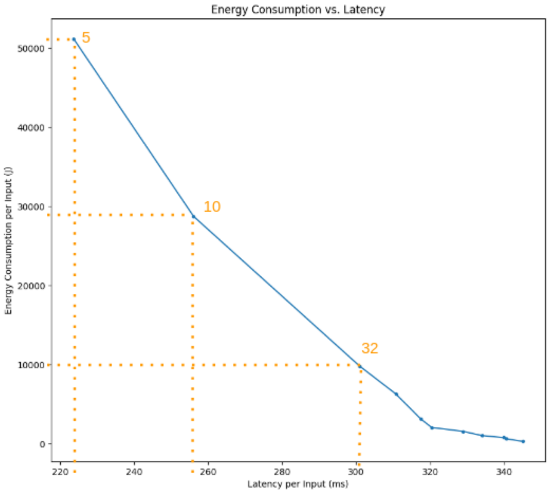


Figure 4 shows a comparison of latency per input with the energy consumption per input, showing that as batch size increases, the energy used per input decreases significantly, but latency per image increases, As batch size increases, the effect of the change on both energy consumption and latency is significantly decreased. These results suggest a trade-off between latency and energy consumption that is determined by utilization and would be useful to consider in IPA when selecting an optimal configuration.

6. Conclusion

These experiment results show that energy consumption can expand the trade-off space because it provides additional information from both latency and accuracy. As shown, the optimal configuration for any machine learning system (set of model weights, batch size) may often be different when energy efficiency is included as an objective.

These results from the experiment show that if energy consumption was not a factor, a batch size of five would have optimal latency. However, just by considering energy consumption we can see that increasing to a batch size of 10 or 32 increases latency by only 40 or 80 milliseconds, with a large energy usage decrease of 20,000 or 40,000 joules.

Future Work These experiments show that there is a complex relationship between energy consumption, latency, and accuracy, but it does not add the energy consumption metric into IPA. The IPA code could be expanded to include an energy consumption metric with a weight that can be set by the developer. Pre-existing energy consumption metrics could be used via a configuration file created by the developer. Additionally, energy consumption could be measured directly from within IPA, saving the developer from needing to do that work manually.

We could also perform these experiments on a wider range of model weights and batch sizes, and also perform them on different types of machine learning models such as Large Language Models (LLMs).

7. Supplementary Materials

The code to recreate the experiments is hosted at https://github.com/csce585-mlsystems/Sustainable-IPA/tree/main/scripts, including instructions for how to run the scripts to recreate the experiment. A video is published at the link https://youtu.be/vWdzPodgtU0?si=gun1-qlPhcyhxaxD where the authors present the work. The presented slide deck is available at the link https://github.com/csce585-mlsystems/Sustainable-IPA/blob/main/documentation/slides/Towards%20Sustainable%20AI%20with%20IPA_%20Experiments%20in%20Energy%20Consumption%20of%20Machine%20Learning%20Models.pdf.

References

Chameleon. Hardware Discovery. https://www.chameleoncloud.org/hardware/

Chameleon. Node 89e48f7e-d04f-4455-b093-2a4318fb7026 (P3-CPU-038). https://chameleoncloud.org/hardware/node/sites/uc/clusters/chameleon/nodes/89e48f7e-d04f-4455-b093-2a4318fb7026/

Ghafouri, S et al. IPA: Inference Pipeline Adaptation to Achieve High Accuracy and Cost-Efficiency. *Journal of Systems Research (JSys)*. 2024. https://arxiv.org/pdf/2308.12871

Green Software Foundation. Energy Efficiency. 2024. https://learn.greensoftware.foundation/energy-efficiency/#:~:text=Energy%20proportionality%2C%20first%20proposed%20in,usually%20given%20as%20a%20percentage.

Intel. Running Average Power Limit Energy Reporting. 2022. https://www.intel.com/content/www/us/en/developer/articles/technical/software-security-guidance/advisory-guidance/running-average-power-limit-energy-reporting.html

IPCC, 2023: Sections. In: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate

Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland, pp. 35-115, doi: 10.59327/IPCC/AR6-978929169164 https://www.ipcc.ch/report/ar6/syr/downloads/report/IPCC_AR6_SYR_LongerReport.pdf

Khan et al. RAPL in Action: Experiences in Using RAPL for Power Measurements. *ACM Transactions on Modeling and Performance Evaluation of Computing Systems (TOMPECS)*, Volume 3, Issue 2, Article No. 9, pp. 1-26. 2018. https://dl.acm.org/doi/10.1145/3177754

Lin et al. Microsoft COCO: Common Objects in Context. 2014. https://arxiv.org/abs/1405.0312

Lin et al. COCO: Common Objects in Context. 2017 https://cocodataset.org/#home

Linux perf wiki Contributors. Perf: Linux profiling with performance counters. 2024. https://perfwiki.github.io/main/

Matplotlib development team. Matplotlib. 2024. https://matplotlib.org/

Taha, A. A. and Hanbury, A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging*. 2015. https://pmc.ncbi.nlm.nih.gov/articles/PMC4533825/

Wang, C. Y., Bochkovskiy, A., and Liao, H. Y. M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. 2022. https://arxiv.org/abs/2207.02696

Wynsberghe, A. Sustainable AI: AI *for* sustainability and the sustainability *of* AI. *AI and Ethics*, Volume 1, pages 213-218, 2021 https://link.springer.com/article/.10.1007/s43681-021-00043-6