The Hidden Joules: Evaluating the Energy Consumption of Vision Backbones for Progress Towards More Efficient Model Inference

Zeyu Yang [1] **Wesley Armour** [1]

Abstract

Deep learning has achieved significant success but poses increasing concerns about energy consumption and sustainability. Despite these concerns, there is a lack of understanding of their energy efficiency during inference. In this study, we conduct a comprehensive analysis of the inference energy consumption of 1,200 ImageNet classification models—the largest evaluation of its kind to date. Our findings reveal a steep decline in accuracy gains relative to the increase in energy usage, highlighting sustainability concerns in the pursuit of marginal improvements. We identify key factors contributing to energy consumption and demonstrate methods to improve energy efficiency. To promote more sustainable AI practices, we introduce an energy efficiency scoring system and develop an interactive web application that allows users to compare models based on accuracy and energy consumption. By providing extensive empirical data and practical tools, we aim to facilitate informed decision-making and encourage collaborative efforts in the development of energy-efficient AI technologies.

1. Introduction

Over the past decade, AI has achieved remarkable capabilities in various fields. However, these accomplishments have come at the cost of significant computational demands. AI research has traditionally prioritized achieving the highest possible accuracy, often disregarding considerations of model size, complexity, and data requirements.

As the field matures and more AI products and services transition into commercial deployment, computational cost is becoming a major concern. Google reported that the energy consumption of machine learning workloads constituted 10–15% of its total energy usage from 2019 to 2021, with training accounting for 40% and inference for 60% (Patterson et al., 2022). Similarly, Meta observed a power capacity distribution of 10:20:70 among experimentation, training, and inference in their AI infrastructure (Wu et al., 2022).

Moreover, the electricity consumption of these tech giants has been rising steadily. Google's electricity usage increased by an average of 21% per year over the past decade, growing from 3.7 TWh in 2013 to 25.3 TWh in 2023 (Google, 2019; 2024). Meta's electricity consumption grew by an average of 32% per year over the past five years, from 4.9 TWh in 2018 to 15.0 TWh in 2023 (Meta, 2024). Data centers globally were estimated to consume about 1% of global electricity and contribute 0.3% of greenhouse gas emissions in 2018 (Masanet et al., 2020; Jones et al., 2018). Furthermore, the widespread adoption of autonomous vehicles could require as much electricity as all current data centers combined (Sudhakar et al., 2022).

The high energy consumption associated with AI leads to several negative consequences. Economically, it results in higher capital costs for purchasing computing hardware and increased operating expenses for electricity and cooling. Environmentally, it produces a large carbon footprint, exacerbating climate change. Additionally, the substantial computational demands impede further development and innovation in AI. The soaring costs of purchasing or renting computing power close the door to many researchers, leaving only large tech companies as major players in the field. This centralization contradicts the open-source ethos that has traditionally driven AI and software development. Moreover, high energy consumption hinders the deployment of AI in edge scenarios where battery life and thermal design power (TDP) are constrained (Yang et al., 2022; Pereira et al., 2023).

Despite these concerns, there is a lack of comprehensive understanding of the energy consumption during inference and the energy efficiency of different models. In this study, we address this gap by measuring the inference energy consumption of 1,200 ImageNet classification models—a scale that is orders of magnitude larger than any previous work.

[1]Department of Engineering Science, University of Oxford, Oxford, UK. Correspondence to: Zeyu Yang <zeyu.yang@eng.ox.ac.uk>, Wesley Armour <wes.armour@oerc.ox.ac.uk>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

Our research aims to answer the following questions: How much additional cost are we incurring for marginal increases in accuracy? What are the contributing factors to energy consumption in AI models? Do current acceleration techniques improve energy efficiency? How to trade-off between energy consumption and accuracy?

Our key contributions are as follows:

- Extensive Dataset: We provide a comprehensive dataset of inference energy consumption metrics for 1,200 ImageNet classification models, enabling better understanding and comparison of model efficiencies.
- Insights into Energy Consumption Factors: We identify factors contributing to energy consumption, correct existing misunderstandings, and assess the impact of throughput improvements on energy consumption.
- Two new ways to quickly estimate the energy consumption of a model: 34% error just from model param, gmacs and activation count, and less than 1% error from throughput and TDP of GPU.
- Energy Efficiency Scoring System: We introduce a scoring system to rank models based on their energy efficiency, providing a standardized metric for evaluation and informed decision-making.
- Interactive Web Application: We develop a web app that allows users to visualize and compare models based on energy efficiency and other metrics, promoting energy-conscious choices within the community.

Our study empirically confirms that current deep learning models are significantly less energy-proportional than idealized hardware expectations suggest, aligning with insights from (Barroso & Hölzle, 2007). The steep diminishing returns we identify reinforce the need for more energy-proportional model architectures and inference methods.

Our findings directly support the goals outlined by (Schwartz et al., 2020), advocating for Green AI principles that prioritize energy efficiency alongside accuracy. By demonstrating the marginal gains in accuracy versus exponential energy demands, our work underscores the urgency emphasized by the Sustainable AI movement (Van Wynsberghe, 2021) to embed sustainability criteria into AI model evaluation and selection.

2. Related Work

Several researchers have called on the AI community to raise awareness of the energy consumption of AI models and their subsequent environmental consequences. Schwartz et al. (Schwartz et al., 2020) proposed the concept of Green AI, which emphasizes computational efficiency alongside model quality, as opposed to Red AI that prioritizes higher accuracy regardless of computational cost - a norm in the field. They also advocated for reporting a model's FLOP count as a standard practice in publications. Van Wynsberghe (Van Wynsberghe, 2021) introduced the Sustainable AI movement to promote changes throughout the AI life cycle - including training, fine-tuning, implementation, and governance - towards ecological integrity and sustainable development.

Li et al.(Li et al., 2016) were among the first to investigate energy efficiency on GPUs, testing both the training and inference of AlexNet, OverFeat, VGG, and GoogleNet on NVIDIA K20m and TITAN X GPUs. They identified the energy consumption of different CNN layers and analyzed the impact of hardware settings such as batch size, hyper-threading, ECC, and DVFS on energy efficiency.

Canziani et al.(Canziani et al., 2016) conducted a comparative analysis of more than a dozen models, including variants of Inception, VGG, and ResNet, evaluating metrics such as accuracy, memory usage, inference time, and power consumption on an NVIDIA Jetson TX1. Their work aimed to guide efficient DNN design for practical applications by highlighting the trade-offs between accuracy and computational requirements. Yao et al. (Yao et al., 2021) tested three CNNs - VGG16, ResNet50, and Inception-V3 - on three GPUs: NVIDIA Tesla M40, P4, and V100. They highlighted the impact of different configurations and optimizations, including quantization and the use of TensorRT and Tensor Cores, on energy consumption, providing insights for more energy-efficient deployment of CNNs in high-performance computing environments. Overall, the number of models evaluated in these works is limited, and the GPUs used are outdated by today's standards.

Henderson et al. (Henderson et al., 2020) proposed a standardized framework for consistent reporting of energy and carbon emissions in ML research, aiming to raise awareness, enable cost-benefit analyses, and promote energy-efficient practices in model development and deployment. One of their main arguments was that a model's parameter count and FLOPs do not necessarily correlate with energy consumption. They tested over 20 models, including VGG, ResNet, MobileNet, and SqueezeNet. However, they did not specify which GPU they used. Most importantly, they ran all models with a batch size of one, which underutilizes any reasonably modern GPU and gives larger models an unfair advantage.

Desislavov et al. (Desislavov et al., 2023) conducted one of the most extensive analyses to date, examining 94 different ImageNet classification models. They showed that efficiency gains from hardware advances and algorithmic improvements mitigate energy growth despite increasing model complexity. However, they estimated the energy consumption of a model by dividing the model's FLOPs by the

GPU's FLOPs per second and multiplying by the GPU's TDP. This is a highly idealized and optimistic assumption, which, as we show in our results section, differs significantly from real-world scenarios.

Shifting focus away from computer vision, Samsi et al.(Samsi et al., 2023) benchmarked the inference energy and compute requirements of various configurations of the LLaMA model across GPU setups to highlight energy usage patterns and identify optimization opportunities for resource efficiency. Luccioni et al.(Luccioni et al., 2024) benchmarked 80 models across 10 specific tasks and 8 general-purpose models, providing a systematic comparison of energy consumption and carbon emissions in deployment. They emphasized the significantly higher costs of deploying general-purpose models compared to task-specific ones and urged careful consideration of these environmental impacts.

Beyond evaluating 50–100× more models, our work addresses key limitations of prior studies, such as using batch size = 1 and TDP-based energy estimates. Our rigorous methodology yields accurate, realistic conclusions. We clarify deployment details often omitted in past work, showing PyTorch can be up to 10× more energy-consuming than TensorRT. Measurements were conducted under an industry-standard inference scenario to ensure real-world relevance. Prior studies also used outdated models and GPUs. By evaluating models up to 2024 on Hopper and Blackwell Generation GPUs, our study offers the most comprehensive analysis to date. Our open-source framework on GitHub[1] supports easy evaluation of new models and GPUs. We invite the community to contribute by utilizing our framework to assess their models and GPUs and contribute to our GitHub repo and web app.

3. Methodology & Experimental Setup

Model Selection To comprehensively analyze energy efficiency across diverse model architectures, we included all available pre-trained models from the Hugging Face PyTorch Image Models (Timm) library (Wightman, 2019). Timm is widely used for its extensive collection of state-of-the-art vision models—encompassing convolutional neural networks, vision transformers, and hybrid architectures. We ultimately selected over 1,200 models in order to capture distinct energy–accuracy trade-offs across various architectures, depths, and publication times. This breadth ensures our results are not simply redundant and that they capture the true diversity of existing approaches, covering a wide range of model sizes, complexities, and design philosophies.

Hardware Configuration All main experiments used two NVIDIA GPUs: the A100 PCIe 40GB (NVIDIA, 2020) from the "Ampere" generation and the H100 PCIe 80GB (NVIDIA, 2023) from the "Hopper" generation. Both deliver state-of-the-art performance and efficiency in deep learning computations. Key hardware and software configurations can be found in the appendix.

Inference Deployment Methods We evaluated the models with two inference methods: standard PyTorch (Ansel et al., 2024) at FP32 precision and NVIDIA TensorRT at FP16 precision. Standard PyTorch serves as a non-optimized baseline, while TensorRT represents a production-level optimization for NVIDIA GPUs.

Accuracy Metrics To thoroughly evaluate the models' accuracy, robustness, and generalization, we used six validation/test datasets: the original ImageNet validation set (Russakovsky et al., 2015), along with five widely recognized datasets: ImageNet Real Labels (Beyer et al., 2020), ImageNet V2 Matched Frequency (Recht et al., 2019), ImageNet Sketch (Wang et al., 2019), ImageNet Adversarial (Hendrycks et al., 2021b), and ImageNet Rendition (Hendrycks et al., 2021a). These datasets assess standard classification accuracy as well as performance across diverse visual domains and distribution shifts.

Measurement Procedure We developed an automated script to evaluate all selected models. We iteratively increased the batch size, starting from 1 and doubling until reaching the GPU's memory limit, ensuring maximum hardware utilization and fair comparisons across different model sizes. For each batch size, we performed two runs: a warm-up to handle potential out-of-memory errors and prepare the GPU, followed by a measured run. Inference was continued until more than 13 repetitions and 10 seconds of runtime were reached, ensuring sufficient data for both large and small models. More details can be found in the Appendix.

Energy Measurement Following the guidelines in (Yang et al., 2024), we measured GPU energy consumption using onboard power sensors via nvidia-smi. We recorded power usage and other GPU metrics at 100 Hz, then logged the data for subsequent analysis.

Additional Metrics We collected key model statistics—such as parameter counts and FLOPs—using ptflops (Sovrasov, 2018-2024) and torchinfo (Yep, 2020). We also recorded GPU utilization, VRAM usage, and temperature to investigate how these metrics relate to model characteristics, performance, and energy consumption.

Result Integrity & Reproducibility We had exclusive machine access during experimentation, preventing interference from other processes. All experiments used the same A100 and H100 GPU cards for consistency. The servers were housed in a data center with controlled cooling, and GPU temperatures remained within operational ranges. We have made our source code available in a GitHub repository to facilitate replication and verification of our results.

[1] https://github.com/JimZeyuYang/DL-Inference-Energy-Efficiency

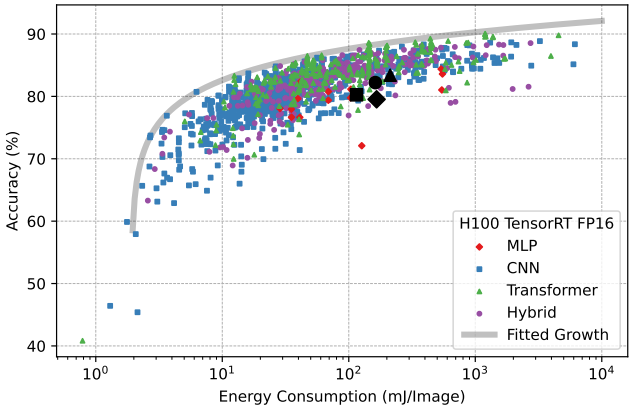

Figure 1. Energy consumption data plotted against the ImageNet accuracy for the H100 TensorRT FP16 inference setup. Models are categorized into general architectures, and the large black markers shows the center of the individual clusters.

Table 1. Pearson Correlation Coefficient (PCC) and Spearman's Rank Correlation Coefficient (ρ) of the energy consumption of models across different deployment setups: comparing same GPU with different software and different GPU with same software.

PCC / ρ	A100 - PT	H100 - TRT
A100 - TRT	0.8553 / 0.9345	0.9840 / 0.9943
H100 - PT	0.9939 / 0.9904	0.8299 / 0.9201

4. Results

In this section, we present our findings in four parts. First, we provide the overall energy consumption data of all tested models, analyzing observations and trends. Next, we investigate the factors that contribute to energy consumption, aiming to correct common misunderstandings. We then explore methods that improve energy efficiency and examine the relationship between other metrics. Finally, we discuss the trade-off between accuracy and energy consumption and introduce our interactive web app designed to facilitate this and promote efficient and responsible Machine Learning.

4.1. Energy Consumption Results and Trends

Overall Results Fig.1 plots the energy consumption of each model against its ImageNet classification accuracy for the most efficient H100 TensorRT FP16 setup. The model distribution remains largely consistent across different inference methods (see Table 1), and data for the other three setups appear in the Appendix.

These models are categorized by their architecture into Multi-Layer Perceptrons, Convolutional Neural Networks, Transformers, and hybrid CNN-Transformer models. Note that the energy consumption axis is on a logarithmic scale.

A prominent finding is the steep diminishing returns in accuracy as energy consumption grows. Energy usage spans four orders of magnitude (a factor of 10,000). In the first order of magnitude, a tenfold increase in energy consumption

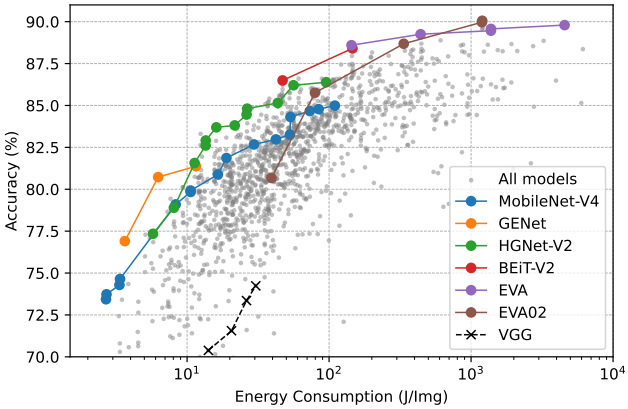

Figure 2. Six highlighted models that largely forms the efficient frontier. ML veteran VGG is also highlighted for reference.

approximately doubles accuracy (from 40% to 80%). The second decade increases accuracy by only 7% (to 87%), and a further tenfold increase gains just 3% (to 90%), with negligible improvements thereafter. This pattern exhibits logarithmic growth on a logarithmic scale, indicating a nested logarithmic relationship. We fit this growth function to the models on the efficient Pareto front of the H100 TensorRT data (gray line in Fig.1). Extrapolating suggests that achieving 100% accuracy would require about 207 MWh of electricity per image—enough to power San Francisco for approximately 15 minutes.

When comparing architectures, MLPs generally yield lower accuracy at higher energy consumption, CNNs occupy the lower accuracy and energy range, and Transformers reside in the higher accuracy and energy range. Hybrid CNN–Transformers lie between these extremes. The large black markers in Fig.1 show the centers of these clusters.

Efficient Frontier Models As shown in Fig.2, Models on the efficient frontier primarily come from just six families — MobileNet-V4 (Qin et al., 2025), GENet (Lin et al., 2020), HGNet-V2 (Contributors, 2023), BEiT-V2 (Peng et al., 2022), EVA (Fang et al., 2023), and EVA02 (Fang et al., 2024) — ranging from lower to higher accuracy and energy consumption. The first three are CNN-based, while the latter three are Transformers, indicating no single architecture uniformly maximizes energy efficiency.

MobileNet-V4, originally designed for resource-constrained edge devices, remains highly efficient on the H100 80G, one of the largest GPU. However, as the model scales up, other families surpass its performance. Both GENet and HGNet-V2 are optimized for low-latency GPU inference via grouped and depthwise-separable convolutions, tuned network depth and width, residual bottlenecks, memory-efficient access patterns, and parallel-friendly operations. Although energy consumption correlates strongly with throughput (shown in later sections), it is notable that these latency-oriented designs also excel in energy efficiency.

Moving to higher accuracy and energy consumption, BEiT-V2, EVA, and EVA02 were not explicitly designed for efficiency. Rather than reducing energy for the same accuracy, they offer higher accuracy at equivalent energy levels — totally another valid path toward more efficient progress.

Performance on Other Datasets As the original ImageNet dataset has largely reached performance saturation, Fig.3 plots energy consumption against the average accuracy across the other five datasets described in the Methodology. Accuracy naturally decreases here because some of these datasets are more challenging. Notably, lower-energy CNNs experience larger performance drops, whereas higher-energy Transformers maintain stronger accuracy, consistent with the expectation that larger Transformer models offer greater robustness against domain shifts. Although the diminishing-return effect appears less pronounced than on ImageNet, it still follows a beyond-exponential pattern.

One might wonder if the efficient frontier models for the original ImageNet lose their advantage on these additional datasets. The red line in Fig.3 marks the Pareto front for ImageNet. We find that most frontier models remain efficient, except in the 60%–70% accuracy range, where several ResNeXt (Xie et al., 2017) models surpass them. This result likely arises from ResNeXt's relatively high accuracy on ImageNet Sketch, attributable to its wider "cardinality" architecture that enhance the out-of-distribution performance.

Other GPUs To check whether the observed trends are generalizable to other GPUs, we tested the 128 most efficient models on H100 on 4 other GPUs: RTX 3090, 4090, 5090 and a mobile laptop GTX 1650 Ti. The results, shown in Fig. 4, strongly confirm the original trends observed on A100 and H100, further validating the broad applicability of our findings.

We found that the RTX 5090 offered limited efficiency gains. Across 128 models, the RTX 4090 achieved a 21.0% geometric mean energy reduction over the RTX 3090, while the RTX 5090 improved by only 5.7% over the RTX 4090. This slowdown in GPU generational energy efficiency gains highlights the importance of our research, as relying on newer GPUs being more efficient alone is no longer sufficient.

4.2. Understanding of Energy consumption

We investigate the relationship between energy consumption and various model metrics, including the number of parameters, FLOPs, activations, and input image size. Additionally, we aim to correct some misconceptions about the energy consumption of models.

Error of Naive Energy Estimation As mentioned in the related work section, (Desislavov et al., 2023) estimated

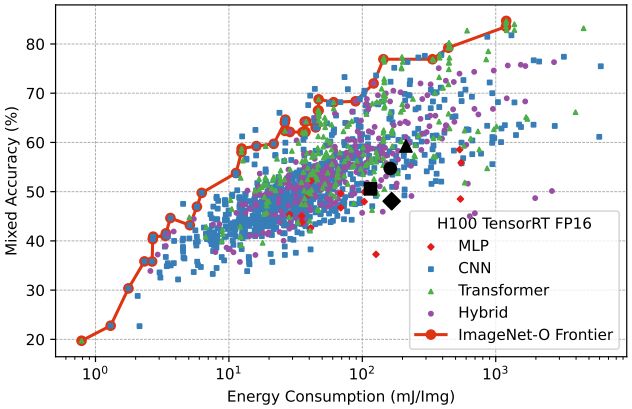

Figure 3. Energy consumption plotted against average accuracy of the five other datasets for robustness analysis. The red line highlights the pareto front models for the original ImageNet dataset.

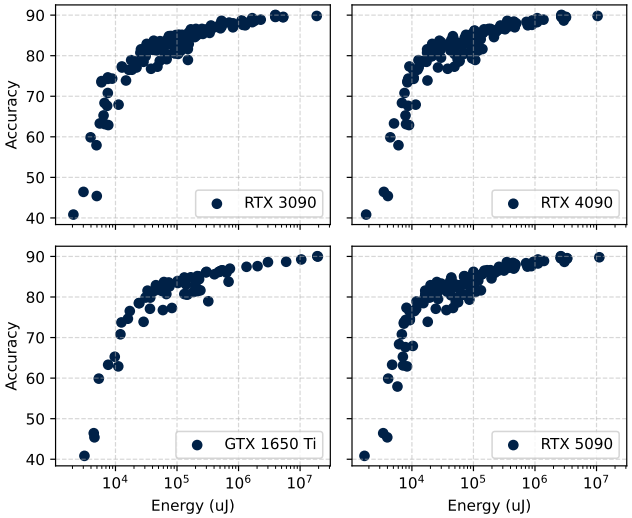

Figure 4. 128 most efficient models on H100 tested on 4 other GPUs: RTX 3090, 4090, 5090, and a Laptop GTX 1650 Ti.

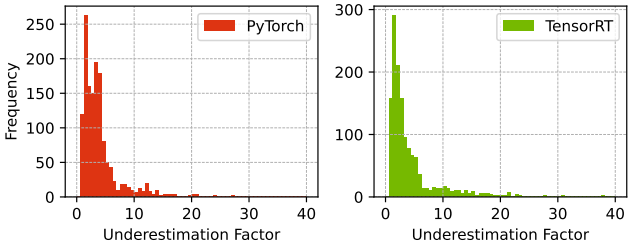

Figure 5. Error in naive estimation of energy consumption based on FLOPs compared with actual measurements on the A100 GPU. The geometric mean of the underestimation factor is 3.13 and 3.16.

the energy consumption of a model by dividing the model's FLOPs by the GPU's FLOP/second and multiplying by the GPU's TDP. We replicated this calculation and compared the estimated energy consumption with our actual measurements. Fig.5 shows the distribution of underestimation. On average, this method underestimates energy consumption by approximately three times and can underestimate by nearly 40 times in some cases.

Parameters, FLOPs, and Activations Intuitively, larger and more complex models demand more computation and thus consume more energy. However, (Henderson et al., 2020) concluded that "FPOs and Params have no strong correlation with Energy Consumption." In contrast, our results reveal a moderately strong linear correlation between parameter count and energy consumption, and a very strong correlation for both FLOPs and activations (Fig.6). Building on these correlations, we fitted a simple multiple linear regression (MLR) model to estimate energy consumption from Params, Gmacs, and activations, achieving a Mean Absolute Percentage Error of 34.73%. Although this empirical model is not highly accurate, it offers a quick way to gauge energy efficiency from basic model statistics - an improvement of nearly an order of magnitude over the naive FLOPs estimates mentioned above. Details of the fitted formula are provided in the Appendix.

Input Image Size For models that accept variable input sizes, increasing the input size yields only negligible improvements in accuracy but results in a substantial increase in energy consumption. Fig.7 illustrates the increase in accuracy and energy consumption as the input size increases for a subset of models that support variable input sizes.

4.3. Efforts to improve energy efficiency

GPU Utilization We examine how GPU utilization affects inference energy consumption. Larger batch sizes process multiple inputs in parallel to increase hardware utilization, hence reducing per-sample overhead, thereby increasing efficiency. Fig.8A shows the energy consumption of EfficientViT on the A100 GPU with progressively larger batch sizes up to the GPU memory limit. Energy consumption drops significantly as batch size grows—until memory bandwidth and processing units become fully utilized. Beyond this point, further increases yield minimal improvements. In some extreme cases, CUDA libraries may adopt less efficient strategies to fit larger batches into memory, ultimately increasing energy consumption.

Energy consumption, Throughput, and TDP Fig.8B shows the throughput and latency achieved at varying batch sizes. Latency increases linearly with batch size, while throughput increases initially and then plateaus. We observed an inverse relationship between throughput and energy consumption. Throughput is measured in images per second, and energy consumption is measured in joules per image. The product of these two gives the average power draw of the GPU during model execution:

$$\frac{Imgs}{s} \times \frac{Joules}{Img} = \frac{Joules}{s} = Watt = AvgPwrDraw \tag{1}$$

The maximum power a GPU can draw is its Thermal Design Power (TDP), and the GPU maintains power draw not to exceed the TDP under heavy load.

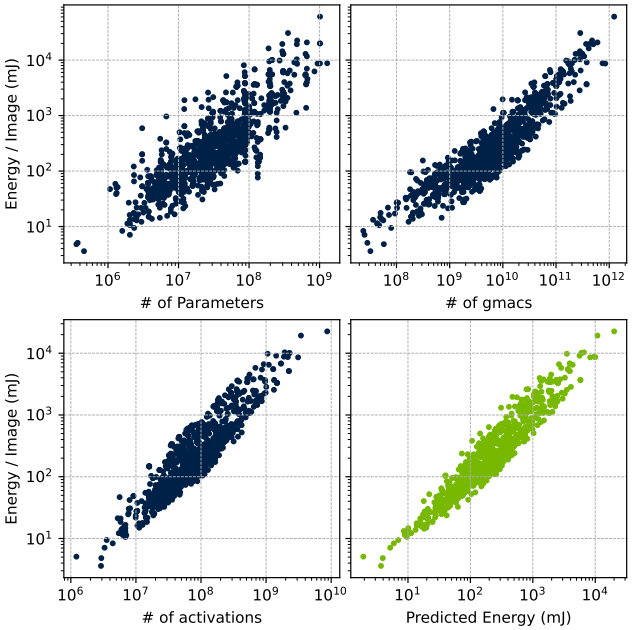

Figure 6. Relationship between energy consumption per image and number of parameters, FLOPs, and activations (A100 with PyTorch FP32). Pearson Correlation Coefficients are 0.6679, 0.8021, and 0.9023, respectively. The plot on the lower left shows the predicted energy using the emprical model vs. the actual Energy consumption per image.

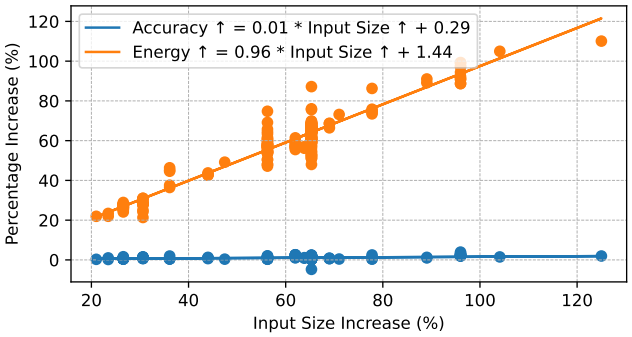

Figure 7. Increase in accuracy and energy consumption as the input image size increases (A100 with TensorRT). The increase in accuracy is minimal, whereas energy consumption is almost directly proportional with input size.

To confirm this inverse relationship, we plotted throughput against energy consumption in Fig.8C, along with the maximum possible product of the two (the TDP line). As batch size increases, energy consumption decreases, while throughput increases. Although the relationship is not strictly linear, when batch size is small, the GPU is underutilized, resulting in an average power draw below the TDP. As batch size increases, GPU utilization improves, and the throughput-energy consumption combination approaches the TDP limit. The conclusion, although counterintuitive, is that higher average power draw leads to greater energy

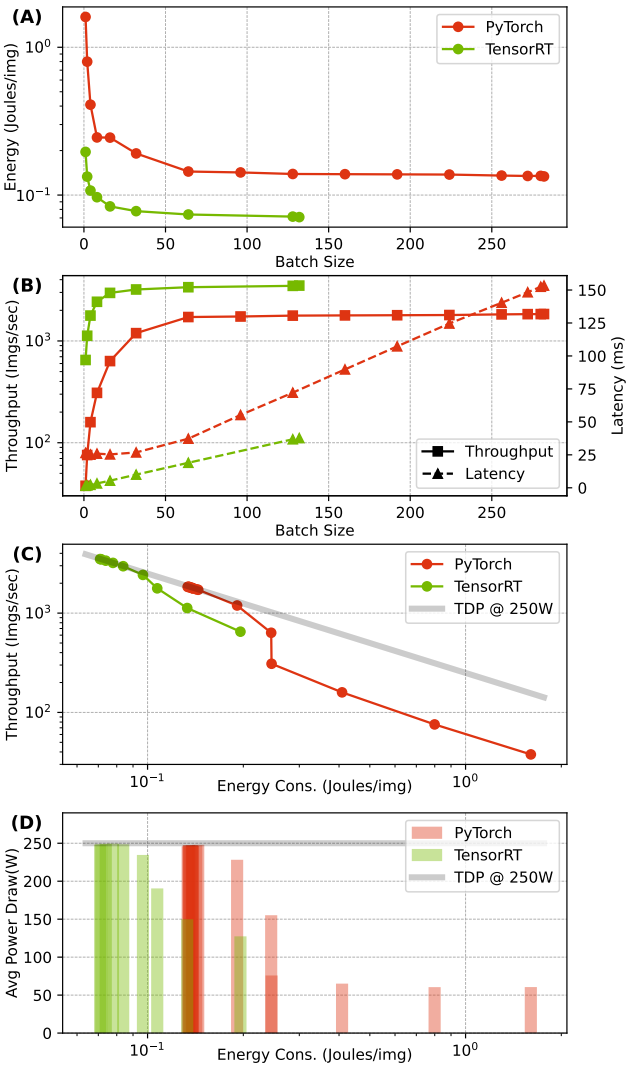

Figure 8. Analysis of batch size effects on EfficientViT on H100: (A) Energy consumption at different batch sizes, (B) Throughput and latency, (C) Throughput vs energy consumption with the TDP limit indicated, and (D) Average power draw relative to batch size.

efficiency due to better GPU utilization.

This observation underscores the importance of making sure the inference setup for a particular model fully utilize the hardware of a particular GPU, ensuring a fair comparison between different models. Simply performing inference with a batch size of one (as mentioned in related works) would unfairly favor larger and more complex models, as they naturally consume more GPU resources.

Infer Energy Consumption from Throughput Fig.9A plots energy consumption against throughput for all models at their most efficient batch size on the A100 GPU. Each model lies on the TDP line. One can accurately estimate a model's energy usage from its throughput alone with a Mean Absolute Percentage Error of 0.97% (e.g., when the GPU is unavailable, but throughput is reported in a paper).

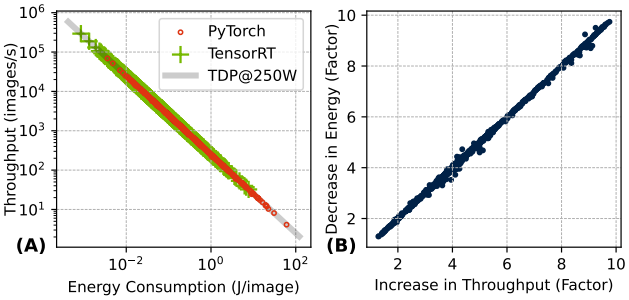

Figure 9. Relationship between energy consumption, throughput, and TDP: (A) Energy consumption versus throughput for all models, showing the TDP limit, and (B) Improvement in energy consumption versus throughput when using TensorRT instead of PyTorch on the A100 GPU.

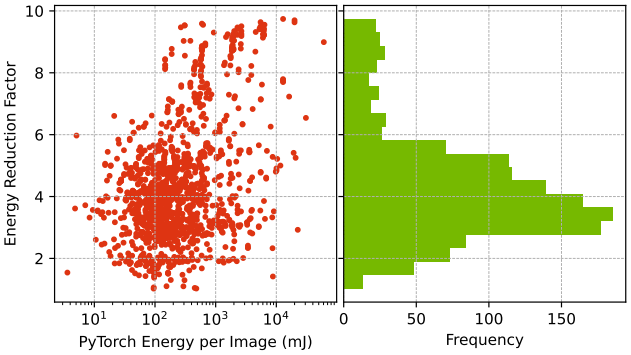

Figure 10. Reduction in energy consumption when using TensorRT instead of PyTorch on the A100. The geometric mean reduction factor is 3.89 with a geometric standard deviation of 1.53. On the H100, the geometric mean reduction factor is 4.02 with a geometric standard deviation of 1.55.

Fig.9B illustrates the proportional relationship between increased throughput and decreased energy consumption when switching from PyTorch FP32 to TensorRT FP16. Since many inference accelerators focus on throughput, and the TDP ceiling enforces a fixed power limit, any improvement in throughput directly translates into lower energy consumption.

Energy Savings from TensorRT FP16 Fig.10 shows the reduction in energy consumption for each model when switching from PyTorch FP32 to TensorRT FP16. On average, energy consumption decreases by about fourfold. This improvement stems from lower computational and memory overhead with FP16, as well as the layer/kernel fusion and kernel tuning provided by TensorRT. Notably, models that were highly energy-intensive under PyTorch see greater gains (up to 10×), likely due to less efficient PyTorch implementations. This also explains the relatively weak correlation between PyTorch and TensorRT energy consumption on the same GPU compared to using the same software on different GPUs.

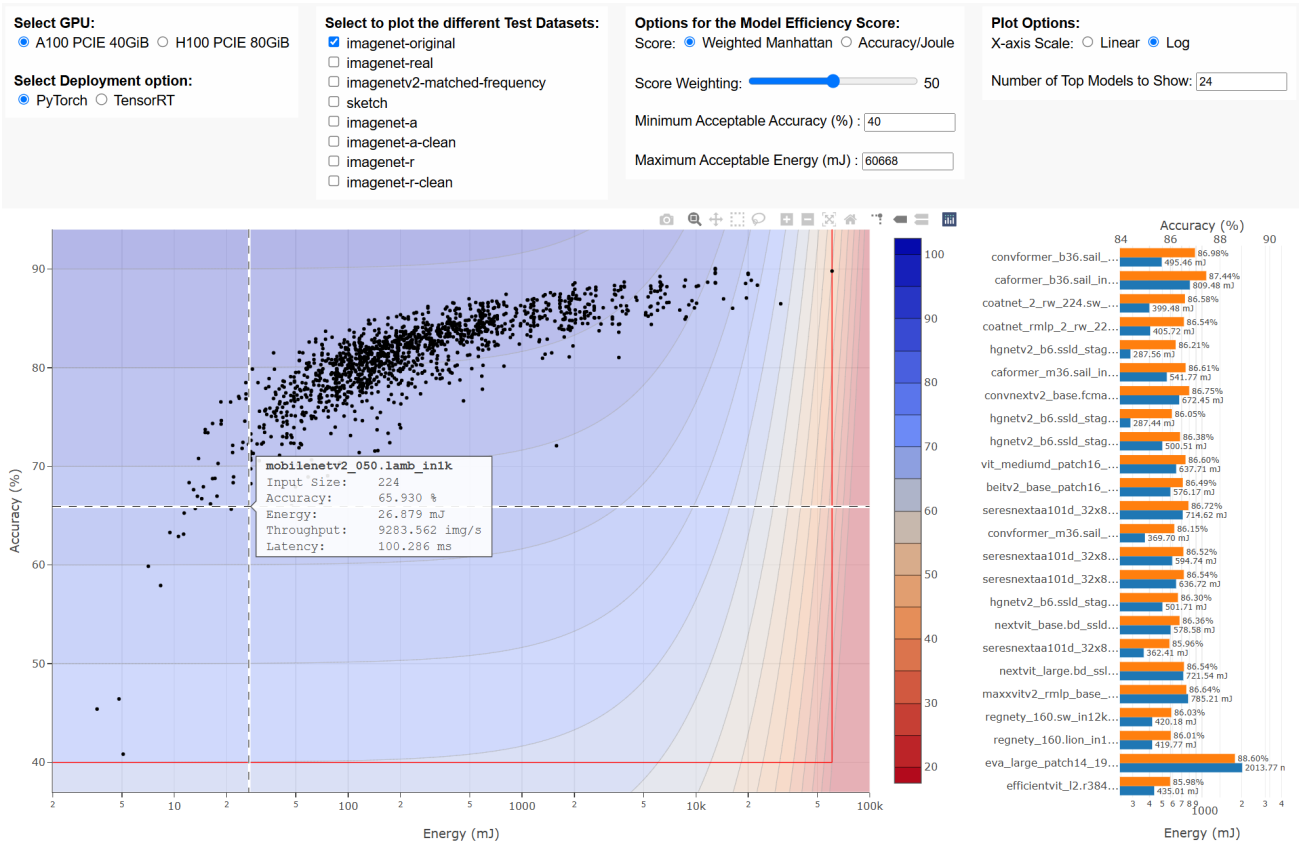

Figure 11. Screenshot of the interactive web application. The top menu lets users to select inference setups, test datasets, scoring metrics, and plotting options. The scatter plot shows energy consumption versus accuracy for all models based on these selections. Hovering over a point reveals the model's details, and clicking opens its Hugging Face page. The red lines indicate user-defined accuracy and energy thresholds, while the background displays a contour map of the selected efficiency score. The bar plot on the right highlights the top-performing models under the chosen metric. The web app is available at: https://jimzeyuyang.github.io/DL-Inference-Energy-Efficiency/.

4.4. Trade-off between Model Accuracy & Energy

We propose two methods to evaluate the trade-off between energy consumption and achieved accuracy. A standard "bang for the buck" calculation measures efficiency as the ratio of accuracy to energy consumption (in percentage per joule). However, this metric may unfairly favor less accurate models. For instance, a trivial model that always outputs "goldfish" could achieve 0.1% accuracy on ImageNet-1K while consuming negligible energy, yielding a misleadingly high efficiency score despite being practically useless.

To mitigate this issue, we recommend using the efficiency ratio alongside a minimum accuracy threshold. This approach ranks models based on energy efficiency only if they meet a baseline level of accuracy. Additionally, as with prioritizing accuracy, maximizing efficiency alone may not be ideal in certain use cases. In critical applications such as autonomous driving or medical diagnostics, accuracy is often paramount, regardless of power constraints.

Therefore, we propose a second metric based on a weighted Manhattan distance to the ideal point (100% accuracy and 0 energy consumption):

$$score = 100 - \left(W \left(\frac{E}{N} \right) + (1 - W)\,(100 - A) \right) \quad (2)$$

where E is the energy consumption, A is accuracy, N is the maximum energy consumption among all models (for normalization), and W is a weight between 0 and 100. A weight of 100 prioritizes energy consumption entirely, while 0 prioritizes accuracy. Users can adjust W to suit specific requirements: for example, a medical diagnostic model might use $W = 5$ for high accuracy, whereas categorizing personal photo albums might work well with $W = 90$.

To help researchers and practitioners explore these trade-offs, we developed an interactive web application that visualizes and compares the energy efficiency and accuracy of all models in our dataset under various configurations. Users can select inference setups (e.g., GPU type or software library), evaluate accuracy on specific or combined test sets, and apply different scoring metrics—all in real time. One of the primary aims of this tool is to promote

energy efficiency in machine learning by guiding the selection of feature backbones for broader tasks such as image segmentation and detection. Models that excel on ImageNet Adversarial or Rendition may suit security-critical applications, while those performing well on ImageNet Sketch could be ideal for abstract or low-resolution imagery. A screenshot of the webpage is shown in Fig. 11, illustrating how users can tailor parameters to their needs and instantly observe the impact on model rankings.

Through these complementary evaluation methods and our interactive tool, we hope to provide a balanced perspective on how to select architectures that meet both performance and sustainability goals.

5. Discussion

Our extensive benchmarking demonstrates that while state-of-the-art models achieve higher accuracy, they also incur substantial energy costs with diminishing returns. The logarithmic rise in energy consumption for minimal accuracy gains raises sustainability concerns, both environmentally and economically.

Practitioners must balance accuracy, throughput, and energy consumption—the "iron triangle." As shown in Section 4.3, throughput and energy consumption are bounded by the GPU's TDP, emphasizing the need for efficiency-oriented model design. To aid in navigating these trade-offs, the interactive tool introduced in Section 4.4 allows users to visualize and balance key performance metrics, promoting informed decision-making.

Our study significantly extends previous work by benchmarking 1,200 models on the latest A100 and H100 GPUs, offering a fair comparison and comprehensive insights. Although our experiments focus on these specific GPUs, the trends observed—especially the high correlation of power consumption across different GPUs—suggest broader applicability. The comprehensive coverage of vision backbones would also help downstream tasks like object detection and semantic segmentation.

The considerable energy costs tied to marginal accuracy gains highlight the need to adopt more sustainable AI practices. By prioritizing efficiency and providing practical tools to measure energy consumption, we encourage the development and deployment of models that balance performance with environmental impact, aligning with the growing emphasis on Sustainable AI (Van Wynsberghe, 2021).

Looking ahead, we plan to update our dataset and interactive web application as new models and GPUs emerge. We invite the community to use our open-source code to measure additional models and submit their findings. Such collaborative efforts will build a comprehensive resource to track energy efficiency trends over time, foster transparency, and accelerate sustainable AI innovation.

This comprehensive study establishes a strong foundation in the underexplored field of energy-efficient machine learning and sustainable AI by this comprehensive study on the current trends, factors that influence energy consumption, rectified existing misconceptions, best practices for efficient model deployment, and provided various tools to estimate energy consumption and tradeoff between accuracy. The data, results, and conclusions presented in our study are entirely novel, offering groundbreaking insights into the energy efficiency of deep learning models. We hope this work serves as a catalyst for the development of sustainable AI.

6. Conclusion

In summary, our work provides a comprehensive analysis of the energy consumption of 1200 vision models, illuminating the significant trade-offs between model accuracy and energy efficiency. By highlighting the diminishing returns in accuracy gains and introducing practical tools and metrics, we hope to shift the focus towards more sustainable AI practices. Our findings lay the groundwork for further exploration into optimizing deep learning models for energy efficiency, encouraging a paradigm shift in how we evaluate and prioritize model performance.

Acknowledgements

The authors would like to thank Karel Adámek for insightful and valuable discussions, which significantly influenced this work. WA would like to acknowledge support from the EPSRC Grant (EP/T022205/1). The authors would like to acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work (Richards, 2015).

Impact Statement

The increasing energy consumption of deep learning models, particularly in computer vision, presents critical economic and environmental challenges. This study provides the most comprehensive analysis to date of inference energy consumption across 1,200 ImageNet classification models, offering empirical insights and practical tools to promote more energy-efficient AI development.

Our findings reveal that improvements in model accuracy come with steeply diminishing returns in energy efficiency, raising concerns about the sustainability of current AI scaling trends. The introduction of an energy efficiency scoring system and an interactive web tool empowers researchers and practitioners to make informed decisions, balancing

model performance with environmental impact.

From an ethical standpoint, this work aligns with the growing emphasis on Sustainable AI, encouraging a shift towards more responsible AI development and deployment. By providing transparency on the energy trade-offs associated with different architectures and inference methods, this study fosters awareness and accountability within the AI community.

While our focus is on vision backbones and possible downstream tasks, the implications extend to broader AI applications, including natural language processing and edge computing. We anticipate that this research will contribute to industry-wide efforts in reducing AI's carbon footprint and driving innovations in energy-efficient model design. Ultimately, we hope this work serves as a catalyst for greener AI solutions that balance computational advancements with long-term sustainability.

References

Ansel, J., Yang, E., He, H., Gimelshein, N., Jain, A., Voznesensky, M., Bao, B., Bell, P., Berard, D., Burovski, E., Chauhan, G., Chourdia, A., Constable, W., Desmaison, A., DeVito, Z., Ellison, E., Feng, W., Gong, J., Gschwind, M., Hirsh, B., Huang, S., Kalambarkar, K., Kirsch, L., Lazos, M., Lezcano, M., Liang, Y., Liang, J., Lu, Y., Luk, C., Maher, B., Pan, Y., Puhrsch, C., Reso, M., Saroufim, M., Siraichi, M. Y., Suk, H., Suo, M., Tillet, P., Wang, E., Wang, X., Wen, W., Zhang, S., Zhao, X., Zhou, K., Zou, R., Mathews, A., Chanan, G., Wu, P., and Chintala, S. PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. ACM, April 2024. doi: 10.1145/3620665.3640366. URL https://pytorch.org/assets/pytorch2-2.pdf.

Barroso, L. A. and Hölzle, U. The case for energy-proportional computing. *Computer*, 40(12):33–37, 2007.

Beyer, L., Hénaff, O. J., Kolesnikov, A., Zhai, X., and Oord, A. v. d. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

Canziani, A., Paszke, A., and Culurciello, E. An analysis of deep neural network models for practical applications. *arXiv preprint arXiv:1605.07678*, 2016.

Contributors, P. Pp-hgnetv2 model documentation, October 2023. URL https://github.com/PaddlePaddle/PaddleClas/blob/develop/docs/zh_CN/models/ImageNet1k/PP-HGNetV2.md.

Desislavov, R., Martínez-Plumed, F., and Hernández-Orallo, J. Trends in ai inference energy consumption: Beyond the performance-vs-parameter laws of deep learning. *Sustainable Computing: Informatics and Systems*, 38:100857, 2023.

Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., and Cao, Y. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19358–19369, 2023.

Fang, Y., Sun, Q., Wang, X., Huang, T., Wang, X., and Cao, Y. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024.

Google. Google environment report 2019. https://www.gstatic.com/gumdrop/sustainability/google-2019-environmental-report.pdf, 2019.

Google. Google environment report 2024. https://www.gstatic.com/gumdrop/sustainability/google-2024-environmental-report.pdf, 2024.

Hashemi, S., Anthony, N., Tann, H., Bahar, R. I., and Reda, S. Understanding the impact of precision quantization on the accuracy and energy of neural networks. In *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2017*, pp. 1474–1479. IEEE, 2017.

Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., and Pineau, J. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, 21(248):1–43, 2020.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8340–8349, 2021a.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15262–15271, 2021b.

Jones, N. et al. How to stop data centres from gobbling up the world's electricity. *Nature*, 561(7722):163–166, 2018.

Li, D., Chen, X., Becchi, M., and Zong, Z. 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), pp. 477–484. IEEE, 2016.

Lin, M., Chen, H., Sun, X., Qian, Q., Li, H., and Jin, R. Neural architecture design for gpu-efficient networks. *arXiv preprint arXiv:2006.14090*, 2020.

Luccioni, S., Jernite, Y., and Strubell, E. Power hungry processing: Watts driving the cost of ai deployment? In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 85–99, 2024.

Masanet, E., Shehabi, A., Lei, N., Smith, S., and Koomey, J. Recalibrating global data center energy-use estimates. *Science*, 367(6481):984–986, 2020.

Meta. Meta 2024 sustainability report. https://sustainability.atmeta.com/wp-content/uploads/2024/08/Meta-2024-Sustainability-Report.pdf, 2024.

NVIDIA. Nvidia a100 tensor core gpu architecture. https://images.nvidia.com/aem-dam/en-zz/Solutions/data-center/nvidia-ampere-architecture-whitepaper.pdf, 2020. Accessed: 2024-11-07.

NVIDIA. Nvidia h100 tensor core gpu architecture. https://resources.nvidia.com/en-us-tensor-core/gtc22-whitepaper-hopper, 2023. Accessed: 2024-11-07.

Patterson, D., Gonzalez, J., Hölzle, U., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D. R., Texier, M., and Dean, J. The carbon footprint of machine learning training will plateau, then shrink. *Computer*, 55(7):18–28, 2022.

Peng, Z., Dong, L., Bao, H., Ye, Q., and Wei, F. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022.

Pereira, L. C., Guterres, B., Sbrissa, K., Mendes, A., Vermeulen, F., Lain, L., Smith, M., Martinez, J., Drews, P., Duarte, N., et al. The not-so-easy task of taking heavy-lift ml models to the edge: A performance-watt perspective. In *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, pp. 699–706, 2023.

Qin, D., Leichner, C., Delakis, M., Fornoni, M., Luo, S., Yang, F., Wang, W., Banbury, C., Ye, C., Akin, B., et al. Mobilenetv4: Universal models for the mobile ecosystem. In *European Conference on Computer Vision*, pp. 78–96. Springer, 2025.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Richards, A. University of oxford advanced research computing, 2015. URL https://doi.org/10.5281/zenodo.22558.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252, 2015.

Samsi, S., Zhao, D., McDonald, J., Li, B., Michaleas, A., Jones, M., Bergeron, W., Kepner, J., Tiwari, D., and Gadepally, V. From words to watts: Benchmarking the energy costs of large language model inference. In *2023 IEEE High Performance Extreme Computing Conference (HPEC)*, pp. 1–9. IEEE, 2023.

Schwartz, R., Dodge, J., Smith, N. A., and Etzioni, O. Green ai. *Communications of the ACM*, 63(12):54–63, 2020.

Sovrasov, V. ptflops: a flops counting tool for neural networks in pytorch framework. https://github.com/sovrasov/flops-counter.pytorch, 2018-2024. URL https://github.com/sovrasov/flops-counter.pytorch.

Sudhakar, S., Sze, V., and Karaman, S. Data centers on wheels: emissions from computing onboard autonomous vehicles. *IEEE Micro*, 43(1):29–39, 2022.

Tmamna, J., Ayed, E. B., Fourati, R., Gogate, M., Arslan, T., Hussain, A., and Ayed, M. B. Pruning deep neural networks for green energy-efficient models: A survey. *Cognitive Computation*, 16(6):2931–2952, 2024.

Van Wynsberghe, A. Sustainable ai: Ai for sustainability and the sustainability of ai. *AI and Ethics*, 1(3):213–218, 2021.

Wang, H., Ge, S., Lipton, Z., and Xing, E. P. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019.

Wightman, R. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

Wu, C.-J., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., Chang, G., Aga, F., Huang, J., Bai, C., et al. Sustainable ai: Environmental implications, challenges and opportunities. *Proceedings of Machine Learning and Systems*, 4:795–813, 2022.

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.

Yang, Z., Clark, A. B., Chappell, D., and Rojas, N. Instinctive real-time semg-based control of prosthetic hand with reduced data acquisition and embedded deep learning training. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 5666–5672. IEEE, 2022.

Yang, Z., Adamek, K., and Armour, W. Accurate and convenient energy measurements for gpus: A detailed study of nvidia gpu's built-in power sensor. In *SC24: International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–17. IEEE, 2024.

Yao, C., Liu, W., Tang, W., Guo, J., Hu, S., Lu, Y., and Jiang, W. Evaluating and analyzing the energy efficiency of cnn inference on high-performance gpu. *Concurrency and Computation: Practice and Experience*, 33(6):e6064, 2021.

Yep, T. torchinfo. https://github.com/TylerYep/torchinfo, 3 2020. URL https://github.com/TylerYep/torchinfo.

A. Appendix

A.1. Source Code

All the code used in this work are available at:

https://github.com/JimZeyuYang/DL-Inference-Energy-Efficiency

A.2. Hardware and Software Configurations

	A100 PCIe 40G	H100 PCIe 80G
TDP	250W	310W
CPU	EPYC 7452	Xeon Gold 6342
RAM	1TB	512GB
OEM	GIGABYTE	DELL
OS	CentOS 8.1.1911	CentOS 8.1.1911
GPU Driver	525.116.04	525.116.04
CUDA vers.	11.8	11.8
PyTorch vers.	2.4	2.4

Table 2. Key hardware and software configurations.

A.3. Automated Measurement Procedure

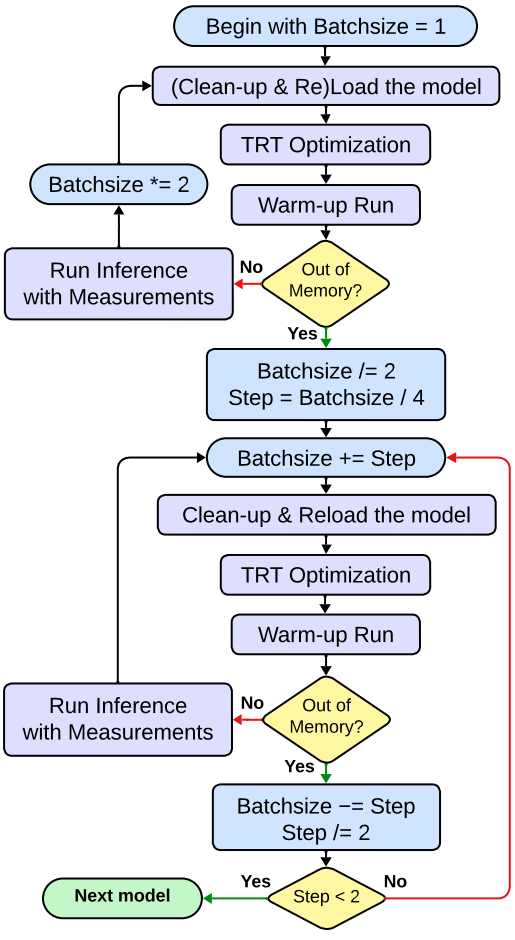

Figure 12. Automated model testing procedure.

A.4. Overall Result for the other 3 Inference Setups

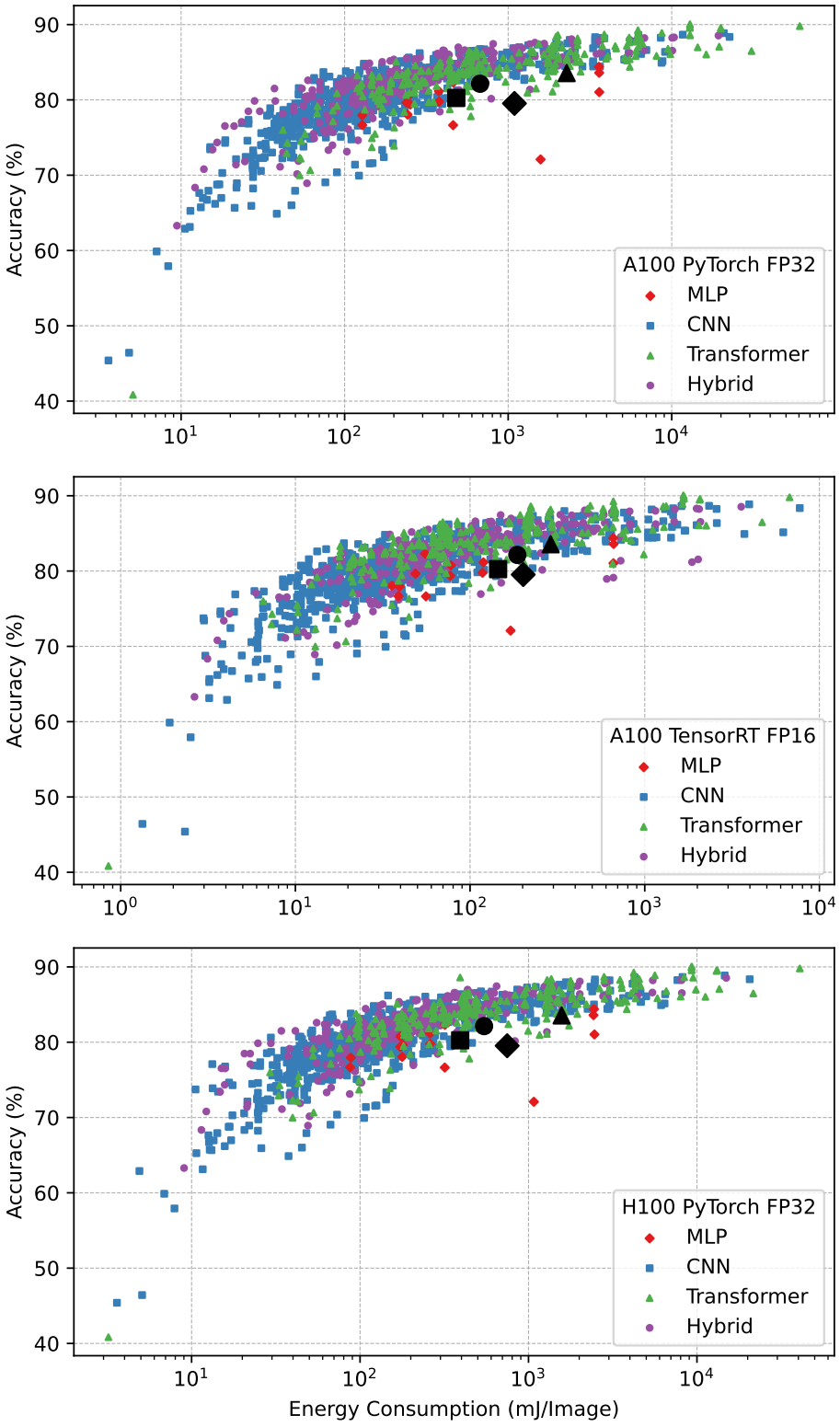

Figure 13. The overall scatter plot for the other 3 inference setups.

A.5. Fitted MLR Model for Energy Consumption Estimation

Since the all four of the Energy Consumption, Params, Gmacs and Activations exhibits uniformly in log-scale, the data follows a power-law or multiplicative distribution. Hence the Multiple Linear Regression model was done after taking the natural logarithm of both the depend and independent variables. The Log-Linear Regression Model has the form:

$$\log(\text{Energy Consumption}) = const + A\log(\text{params}) + B\log(\text{gmacs}) + C\log(\text{activations}) \tag{3}$$

$$\text{Energy Consumption} = e^{const} \times \text{params}^{A} \times \text{gmacs}^{B} \times \text{activations}^{C} \tag{4}$$

The parameters for the 4 inference setups are as follows:

	const	A	B	C
A100 PyTorch FP32	-14.1597	0.1426	0.1328	0.7687
A100 TensorRT FP16	-15.5186	0.1560	0.1622	0.7280
H100 PyTorch FP32	-14.0926	0.1348	0.0952	0.8103
H100 TensorRT FP16	-15.0298	0.1062	0.1141	0.7961

Table 3. Energy Estimation Model Parameters

A.6. Additional Justifications on Experiment Design Choices

Quantization and Pruning We acknowledge the ability of quantization and pruning to improve efficiency. The TensorRT inference setup used FP16. We excluded a more detailed investigation because: 1 - Quantization and pruning are not architecture changes, but rather optimization techniques applied to any existing architectures. Our study focuses on the inherent energy efficiency characteristics of different architectures themselves; 2 - Often needs post-quantization fine-tuning or QAT, and similar for pruning. This process is infeasible for a large-scale study like this; 3 - There are some existing works already on this topic (Tmamna et al., 2024; Hashemi et al., 2017).

nvidia-smi Query Frequency nvidia-smi updates the power draw reading at 10Hz to 50Hz. Querying nvidia-smi at 100 Hz was intentional, balancing the need to capture recent updates without unnecessary polling overhead. According to the Nyquist theorem, sampling above twice the signal frequency preserves all information. A higher sampling frequency would not do any harm.

A.7. Training & Year-on-year Improvement

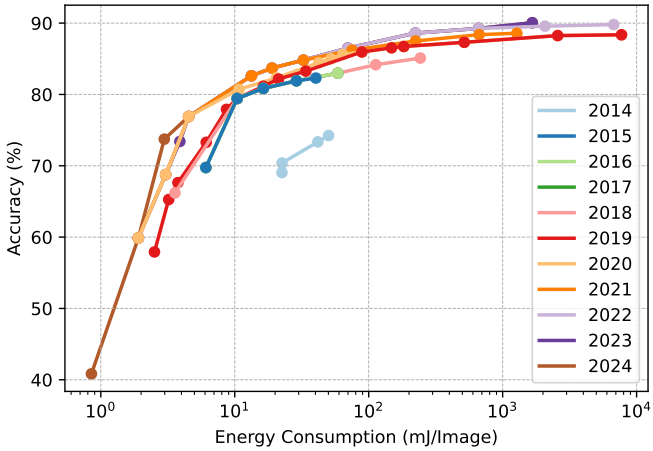

Figure 14. Yearly progress on model accuracy and energy consumption. (A100 TensorRT FP16 as an example for illustration).

Training has significantly evolved over the past decade, involving larger datasets (LAION), self-supervised methods, and various training recipes. We categorized the models by the year their corresponding papers were published. Fig.14 illustrates how each year's new models push the existing convex hull of models towards greater accuracy and efficiency. As the

field evolved and more models were introduced, the efficient frontier expanded in both the high accuracy, high energy consumption direction and the low accuracy, low energy consumption direction. Notably, we observe a consistent vertical increase in accuracy in the high consumption region and a somewhat inconsistent horizontal shift towards lower consumption in the lower accuracy region. However, improvements in the middle region—towards the top-left corner representing high accuracy and low consumption—appear to be more stagnant compared to others.

vit_base_patch16_clip_224 trained with laion2b are 0.2% more accurate than the openai version, but with significantly more training data. While selecting a pretrained model for inference, one would pick the higher accuracy version; the question of whether consuming significantly more energy during training for marginal accuracy gains (e.g., 0.2%) is justified is beyond the scope of this work. This topic alone deserves a future investigation by itself.

A.8. Robustness Analysis

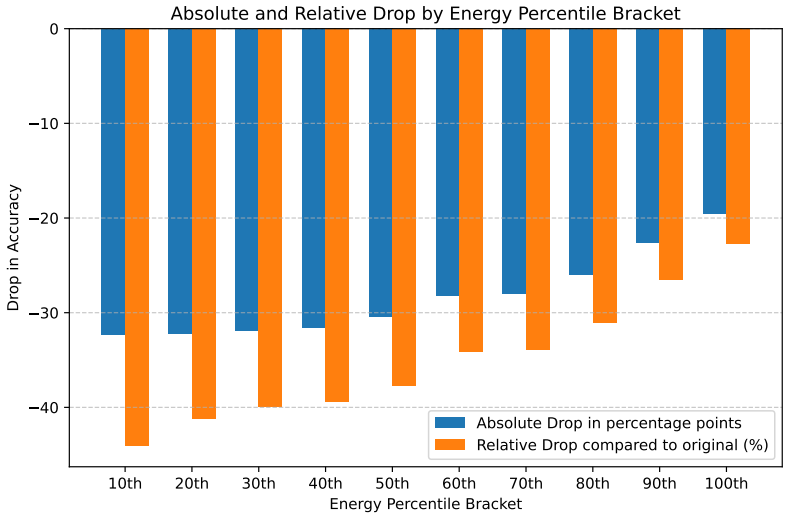

Figure 15. Drop in accuracy by energy consumption percentile.

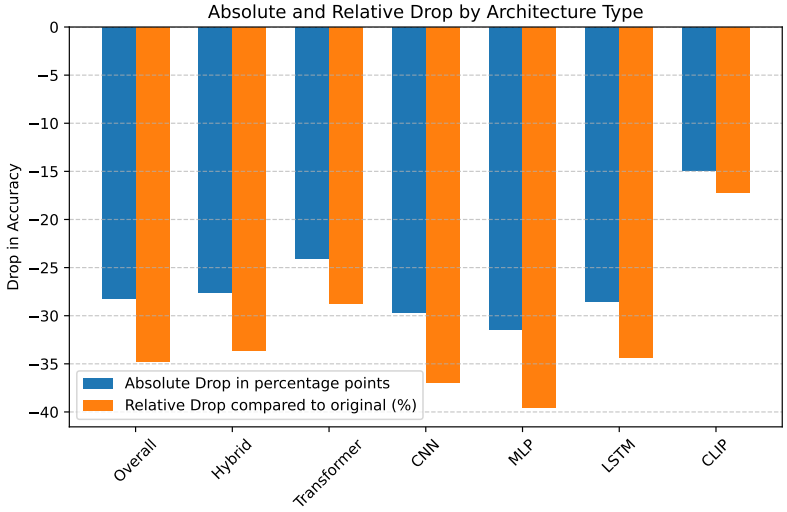

Figure 16. Drop in accuracy by model architecture type.

We performed some additional analysis on the robustness of the models. We measured and calculated the absolute and relative accuracy decrease from the original ImageNet dataset with the average of the other 5 robustness datasets. We grouped models into 10 deciles by energy consumption, as shown in Fig. 15, and observed that higher-energy-consuming models experience smaller accuracy drops, aligning with our findings presented in paper Fig. 3. Looking at different model architectures, as shown in Fig. 16, drop in accuracy for CNNs are 29.7%, hybrid models are 27.6%, and Transformers are 24.0%. We also observed that CLIP models only droped by 15%, demonstrating notably stronger out-of-distribution robustness.

A.9. Extension

Down-stream Tasks While our study directly evaluates image classification backbones, many popular detection and segmentation models (e.g., Faster R-CNN, Mask R-CNN, RetinaNet, DeepLab) commonly reuse these vision backbones as feature extractors. Given this standard practice, the energy efficiency characteristics we observed for various architectures—such as CNN, Transformer, and Hybrid models—serve as meaningful indicators for downstream task efficiency. For instance, backbones identified as energy-efficient at classification tasks could likely translate to efficient feature extraction in detection and segmentation pipelines, influencing overall inference efficiency. However, we do recognize the additional computational complexity and varying bottlenecks introduced by downstream modules (e.g., region proposal networks, mask heads, upsampling layers). Hence, explicitly measuring these downstream tasks remains important for precise validation. We plan to extend our methodology to directly measure and confirm how the observed backbone energy characteristics generalize to these more complex tasks in future work.

LLMs Our current evaluation framework was specifically tailored toward vision backbones. However, we fully recognize the importance and widespread use of large language models (LLMs) in various real-world applications, making their energy efficiency evaluation critically relevant to the sustainable AI community. Although our current study does not include LLMs explicitly, the core principles of our measurement methodology—such as GPU utilization optimization, real-time energy monitoring, and accuracy-performance trade-off metrics—are inherently transferable to large-scale LLM inference scenarios. To adapt our framework effectively to LLMs, we would primarily need to consider: 1 - Different inference characteristics: Token-based generation and longer context windows in LLM inference, compared to fixed-size image inference in vision models; 2 - Adaptation of evaluation metrics: Metrics such as perplexity, generation quality (e.g., BLEU, ROUGE), or task-specific evaluations (e.g., accuracy on reasoning benchmarks) instead of classification accuracy; 3 - Adjustments in batching strategies: Optimal GPU utilization patterns for LLM inference, including considerations for sequence length and context size variability. In future research, we plan to explicitly extend our methodology to evaluate and analyze energy-accuracy trade-offs for large-scale language models, providing analogous insights that could significantly benefit the LLM research and deployment community.