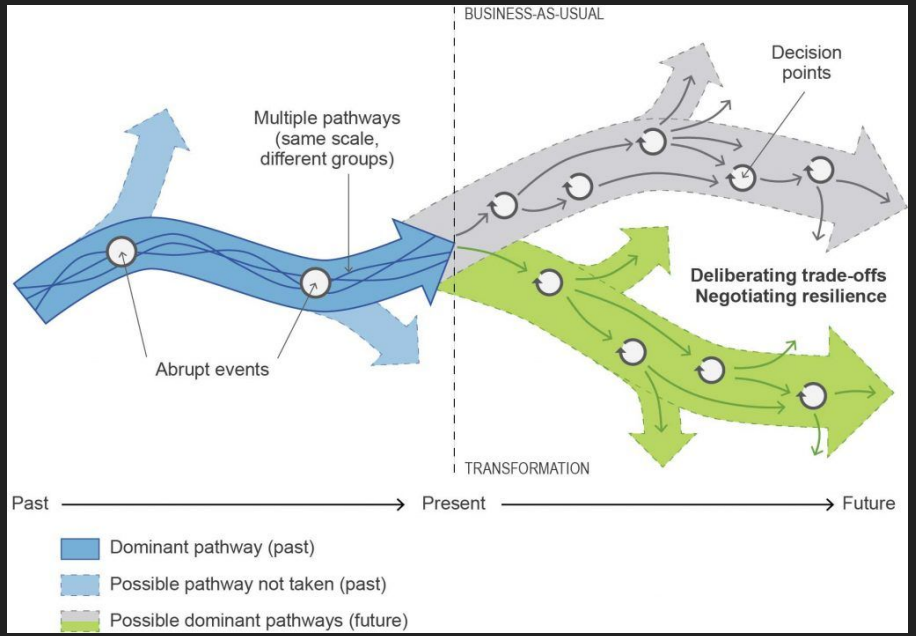
Towards Sustainable AI with IPA: Experiments in Energy Consumption of Machine Learning Models

Regan Willis, Chase Bryson, Osasuyi Agho

Sustainability is More Important Than Ever

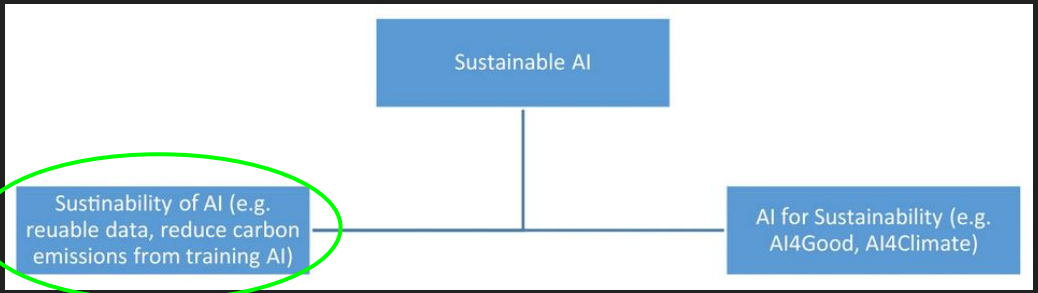
- According to the IPCC…
  - We are likely to reach **1.5 degrees Celsius of global warming** above pre-industrial levels between 2030 and 2052
  - Many areas are experiencing above average warming
  - We can reduce climate risk now by reaching net zero CO2 emissions



IPCC, 2018: Summary for Policymakers. In: *Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty* [Masson-Delmotte, V., P. Zhai, H.-O. Pörtner, D. Roberts, J. Skea, P.R. Shukla, A. Pirani, W. Moufouma-Okia, C. Péan, R. Pidcock, S. Connors, J.B.R. Matthews, Y. Chen, X. Zhou, M.I. Gomis, E. Lonnoy, T. Maycock, M. Tignor, and T. Waterfield (eds.)]. Cambridge University Press, Cambridge, UK and New York, NY, USA, pp. 3-24, doi:10.1017/9781009157940.001.

Sustainability *of* AI

- **Recall**: IPA adjusts ML system pipelines to consider accuracy, latency, and cost at different request volumes.
- Wynsberghe defines Sustainability *of* AI as an area “focused on sustainable data sources, power supplies, and infrastructures as a way of **measuring** and **reducing** the carbon footprint” for machine learning algorithms.
- IPA can join the cause by:
  - **Measuring** the energy consumption of machine learning pipelines that use IPA
  - Selecting models that **reduce** energy consumptions while also meeting other objectives

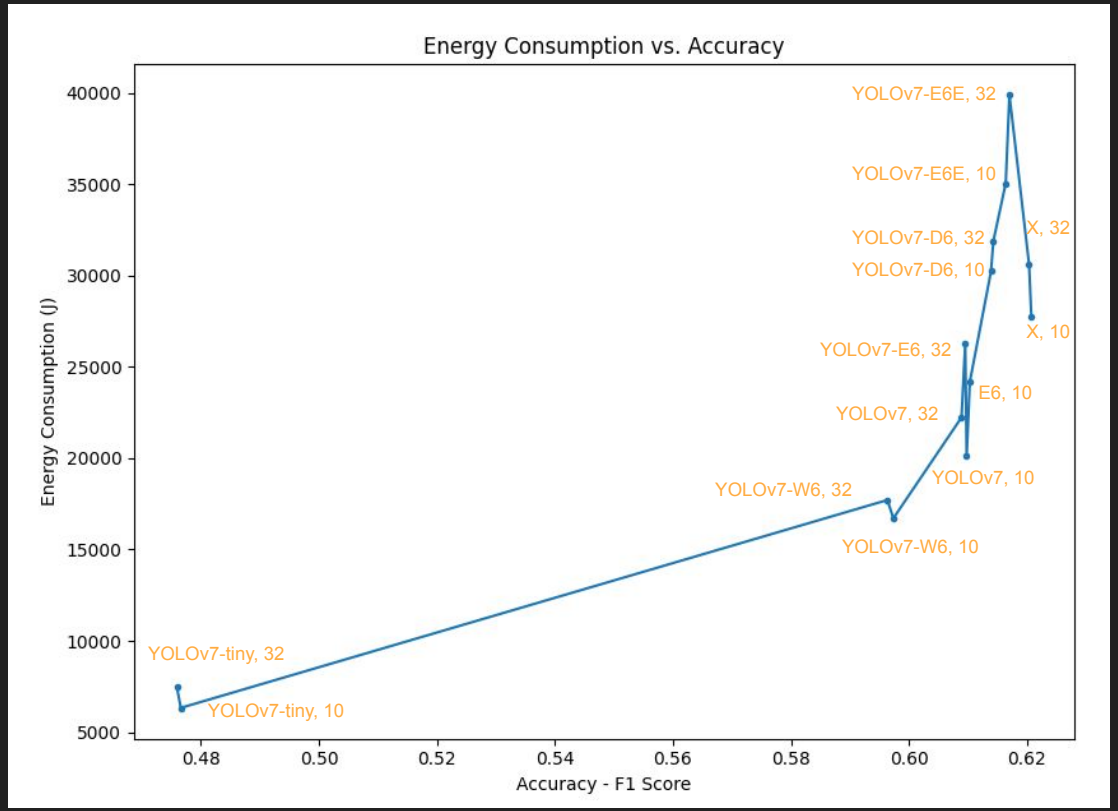


Wynsberghe, A. Sustainable AI: AI for sustainability and the sustainability of AI. AI and Ethics, Volume 1, pages 213-218, 2021 https://link.springer.com/article/.10.1007/s43681-021-00043-6

Does Considering Energy Consumption Change the Optimal Pipeline Configuration?

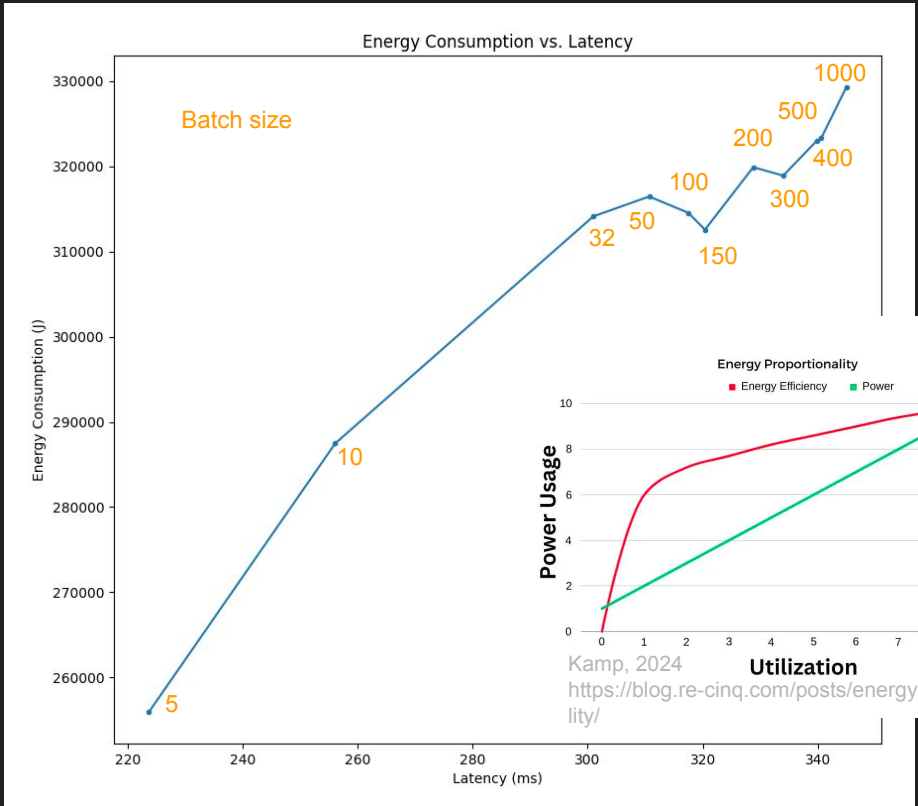
- **Does energy consumption expand the configuration search space for IPA?**
- Are there adaptations for some pipelines that have similar accuracy and latency but differing energy consumption?
- How correlated is energy consumption with the **accuracy** of models?
- How correlated is energy consumption with the **latency** of models?

Energy Consumption vs. Accuracy



- As accuracy of models increases, energy consumption *generally* increases.
- **Some more accurate models have lower energy consumption than less accurate models.**
- Energy consumption and accuracy are **not** perfectly correlated.

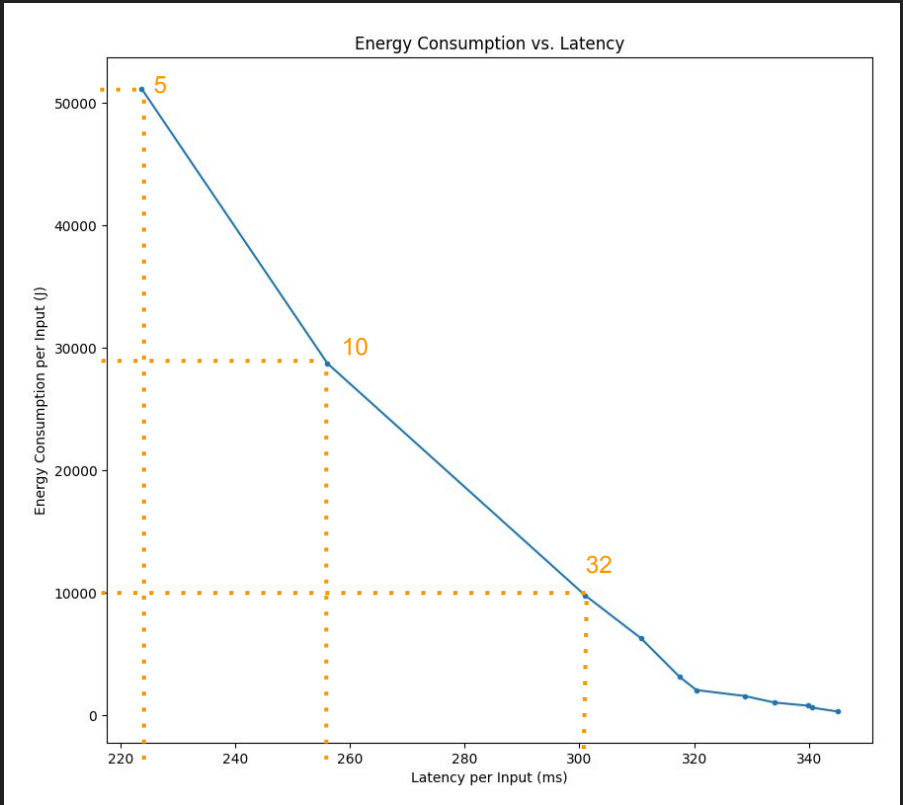
Results: Energy Consumption vs. Latency



Kamp, 2024
https://blog.re-cinq.com/posts/energy-proportionality/

- As latency of models increases, energy consumption increases.
- **Energy proportionality**: higher CPU utilization is more energy efficient.
- **At higher batch sizes, CPU utilization is increased, which decreases the impact of batch size on energy consumption.**

Results: Energy Consumption vs. Latency



- When plotting energy consumption per input we can directly see how increasing the batch size (ie. utilization) decreases the energy needed **for the same amount of work**.
- **At smaller batch sizes, the batch size has a greater impact on energy consumption.**

Future Work Directions

- **Reduce** energy consumption: **add energy consumption metric** into IPA
  - A user of IPA can set an energy consumption SLA
- **Measure** energy consumption: **measure energy consumption of pipelines** directly from within IPA
  - Energy consumption can be difficult to measure and is often not prioritized by developers.

- Other future work:
  - Repeat experiment on different hardware
  - Sensitivity analysis
  - Experimenting with other pipeline types
  - GPU processing
  - Real-time energy consumption tracking
  - Expanding energy consumption measurement methods
  - Providing carbon footprint information