Considering Energy Consumption in IPA Towards Sustainable AI: Experiment Design

Regan Willis, Chase Bryson, Osasuyi Agho

Motivation

- **Recall**: IPA adjusts ML system pipelines to consider accuracy, latency, and cost at different request volumes.
- IPA does **not** consider energy consumption as a metric to tune for.
- Lower energy consumption is better for the environment.
- IPA will be more widely adopted if it provides an energy consumption measurement.

Experiment Question

- **Does energy consumption expand the adaptation search space for IPA?**
  - Are there adaptations for some pipelines that may have the similar accuracy, latency, and cost, but differing energy consumption?
    - More accurate models are larger, requiring more energy
    - Larger models take longer to run, raising latency
- **Future Goal**: A user of IPA can set an energy consumption SLA.
- Regardless of the experiment result, adding energy consumption as a metric in IPA can inform user choices.
  - Energy consumption can be difficult to measure and is often not prioritized by developers.

Experiment Design

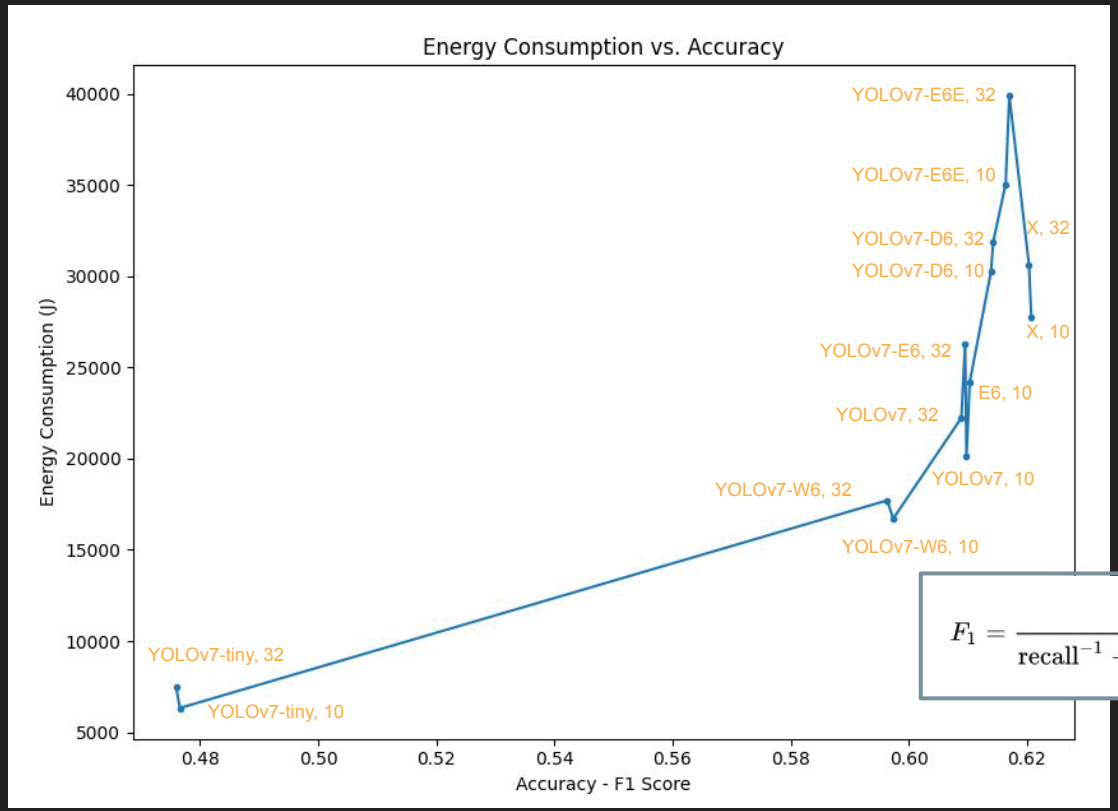
- Run different weights and batch sizes of **YOLOv7**, measure energy consumption, and collect information on accuracy and latency.
  - Independent variables:
    - YOLOv7 weights
    - Batch size
  - Dependent variables:
    - Accuracy
    - Latency
    - **Energy Consumption**
  - Control Variables
    - Dataset -> 496 images from COCO validation set 2017
    - Network pretraining dataset
    - Constant hardware (CPU)
- **Three averaged trials** to increase the signal-to-noise ratio of the energy consumption measurement
- Measure energy:
  - Bare Metal Experiment Pattern: uses perf

Search Space

- Weights: 7 YOLOv7 weights
- Batch sizes: 10, 32

Model	Test Size	AP^{test}	AP_{50}^{test}	AP_{75}^{test}	batch 1 fps	batch 32 average time
YOLOv7	640	**51.4%**	**69.7%**	**55.9%**	161 *fps*	2.8 *ms*
YOLOv7-X	640	**53.1%**	**71.2%**	**57.8%**	114 *fps*	4.3 *ms*
YOLOv7-W6	1280	**54.9%**	**72.6%**	**60.1%**	84 *fps*	7.6 *ms*
YOLOv7-E6	1280	**56.0%**	**73.5%**	**61.2%**	56 *fps*	12.3 *ms*
YOLOv7-D6	1280	**56.6%**	**74.0%**	**61.8%**	44 *fps*	15.0 *ms*
YOLOv7-E6E	1280	**56.8%**	**74.4%**	**62.1%**	36 *fps*	18.7 *ms*

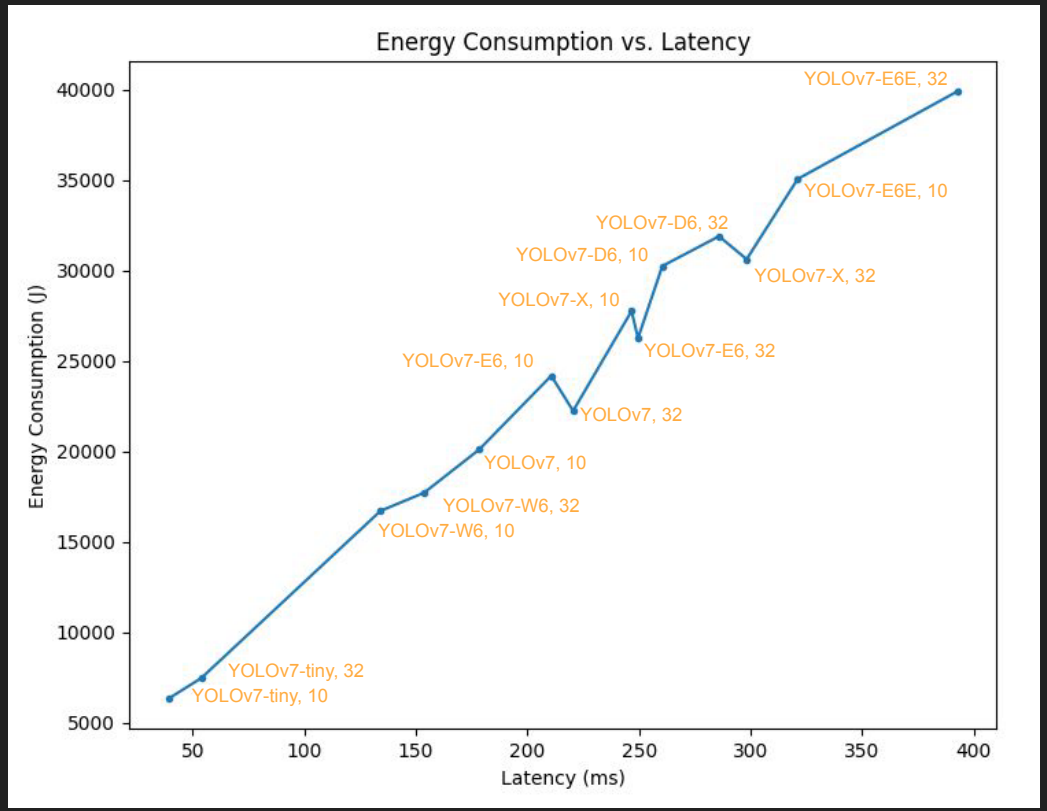
Results: Energy Consumption vs. Accuracy



- As accuracy of models increases, energy consumption generally increases.
- Different model weights affect energy consumption much more than different batch sizes.

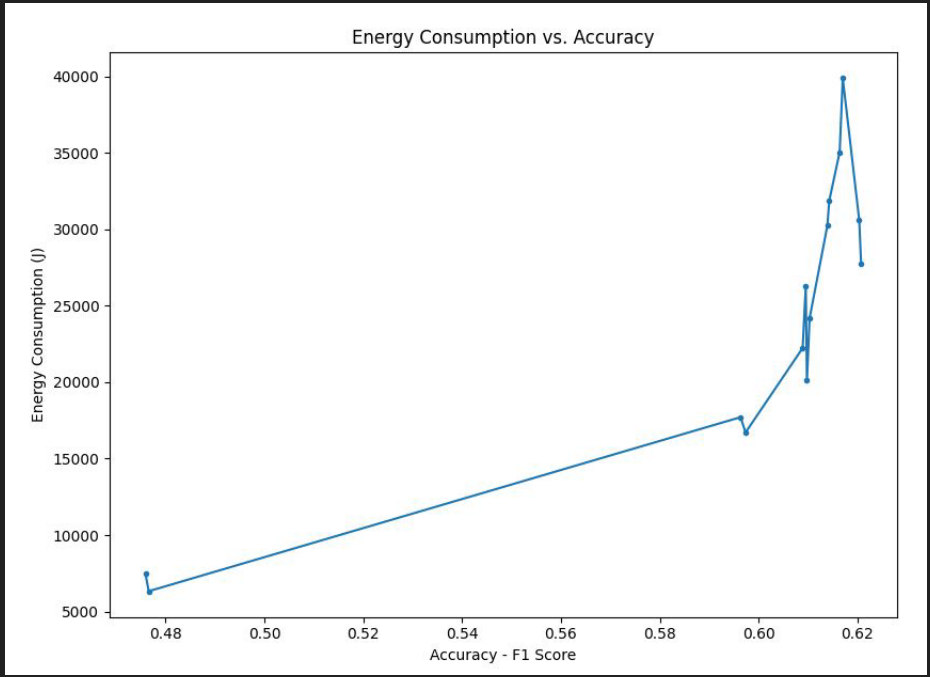
$$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = 2\frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} = \frac{2\mathrm{tp}}{2\mathrm{tp} + \mathrm{fp} + \mathrm{fn}}.$$

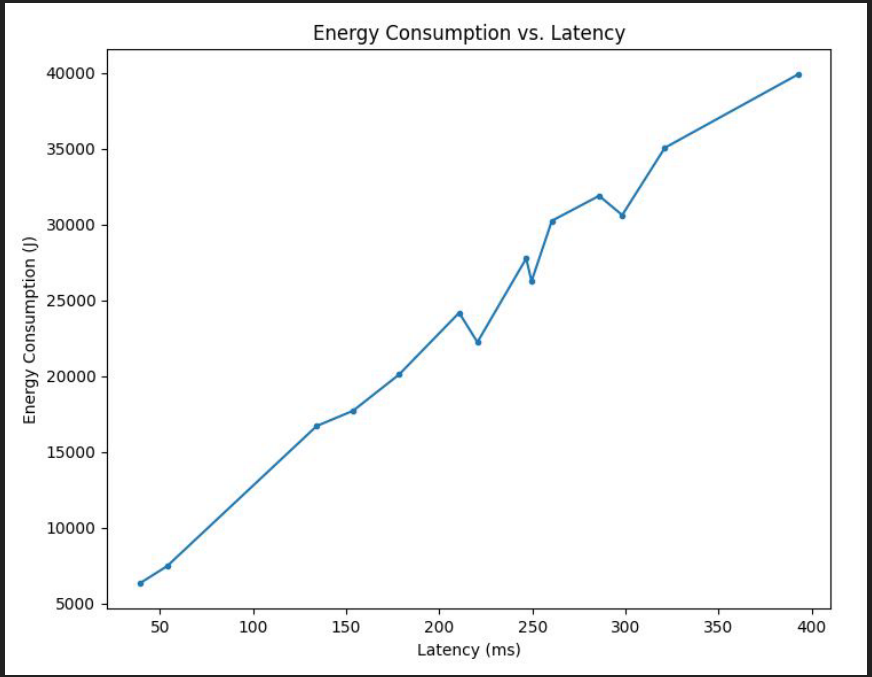
Results: Energy Consumption vs. Latency



- Bigger, more accurate models generally have a greater latency.
- Again, batch size has less relevance on latency than model weights.

Results


Energy Consumption vs. Accuracy
Energy Consumption (J)
Accuracy - F1 Score
40000
35000
30000
25000
20000
15000
10000
5000
0.48
0.50
0.52
0.54
0.56
0.58
0.60
0.62


Energy Consumption vs. Latency
Energy Consumption (J)
Latency (ms)
40000
35000
30000
25000
20000
15000
10000
5000
50
100
150
200
250
300
350
400

Analysis

- Energy consumption is heavily correlated with both accuracy and latency.
- Energy consumption is correlated more with latency than accuracy.
- IPA will be benefited by the addition of an energy consumption metric.

Future Work Directions

- Repeat experiment on different hardware
- **Add energy consumption metric into IPA code**
- Sensitivity analysis
- Experimenting with other pipeline types
- GPU processing
- Real-time energy consumption tracking
- Expanding energy consumption measurement methods
- Providing carbon footprint information