moco

30% fewer FLOPS in inference -> low energy models, low latency models at a faster development speed.

Problem:
Classification models—like wake word detection, behavior recognition, or fraud detection—must run efficiently at the edge. Optimizing these models for low energy and low latency is complex, often requiring **weeks or months of engineering time**.

Solution:
moco automates this process. It analyzes your model's input data and derives simple, accurate rules consistent with your model's predictions that can be used in lieu of inferencing the model for some data. This enables **in-place optimization**, producing models that are both **low-energy and low-latency**.

Key Benefits:

- **30% fewer FLOPS in inference** → lower energy usage
- **Faster development** → hours instead of weeks
- **Flexible optimization strategies:** skip unnecessary model calls or run rules + model in parallel

Energy Saver: Use rules to skip model calls when predictions are certain.
Latency Saver: Run rule + model in parallel and use whichever returns first.

Interested? Next steps
We're looking to **validate moco on production models** that are currently rate-limited. Schedule a consultation to see how your models can benefit.

Contact:
Sam Randall, LinkedIn, Email, Website