moco

30% faster ML inference = higher throughput, lower breach risk, and fewer GPUs.

Problem:
Large networks must inspect trillions of packets per year (e.g., Walmart: ~6T/year). Each packet must be analyzed in real time, but the inference latency of ML detection models determines how much traffic can be processed.

- Low throughput → packets dropped → threats missed.
- High latency → delayed decisions → attacks spread before containment.
- Current fix: throw expensive and power-hungry GPUs at the problem.

Solution
moco analyzes your dataset and derives simple rules consistent with your model's predictions.

How it works:
- As input, moco takes data and the model's classifications.
- moco clusters the data and identifies groups of data points that have the same prediction. Then, rules are fit to predict membership within these clusters.
- At runtime, the new system first checks if a data point is within any of the clusters.
    If it is, the model outputs the relevant prediction.
    If not, the model runs as usual.

Safe Mode: If impacting p99 latency is not an option, we can run rule + model in parallel; use whichever returns first.

Infra Saver: If impacting p99 latency is an option, use the rule to skip unnecessary model calls, reducing GPU and energy cost.

Benefits:

Security Impact: Higher throughput + Lower Dwell Time → Fewer Missed Threats and Less
Spread of Threats. Preventing a single breach averts an average $4.9M loss

Financial Impact: Reduced GPU and energy spend.
30% faster inference reduces GPU fleet size by 25–40%, saving hundreds of thousands to millions annually in Fortune 500 SOCs

Interested? Next steps
I want to validate the concept on models serving in production that are currently rate-limited. Schedule a consultation with me where we'll put together a plan to apply the technology to your model.

Contact:
Sam Randall
https://www.linkedin.com/in/sam-randall-9a3068110/
https://compressmodels.github.io/
quickmlmodels@gmail.com