Performance Optimization Lab

Aim

Profile and optimize ML pipelines.

Theory

Bottlenecks arise from compute, IO, or memory. Profiling reveals hotspots.

Algorithm

1. Measure baseline.

2. Profile data/loading.

3. Apply batching, workers, fp16, caching.

4. Compare results.

Pseudocode

profile()

if bottleneck == io: increase_workers()

Results

Higher throughput and lower latency.

Conclusion

Measure first, optimize second.