COMPSCI 690AB SYSTEMS FOR DEEP LEARNING

Reading Assignment 4

MLPerf Inference Benchmark– Summary

The paper mainly addresses why there is a need for frameworks to measure machine learning inference performance across different hardware architecture, use cases and software framework. Performance of ML inference systems is essential because it is used in a wide range of applications like mobile devices and data centres. And, there is a need for standard benchmarks to do fair comparison between different systems and in order to improve ML inference technology. This paper proposes such a standardized benchmarking framework which is predefined with metrics like latency, throughput, accuracy, scenarios and quality targets. MLPerf Inference majorly has four evaluation scenario, single-stream, multistream, server, and offline. Through this it is feasible to accurately measure and compare the performance of various ML inference systems across different applications. This complements MLPerf Training and it is developed by industry and academia. MLPerf Inference version 0.5 showed significant performance variations around different tasks hardware platforms. Furthermore, there are plans to extend its scope to also include more new models, scenarios and metrics to support the constant evolution of ML Technology. One limitation of the proposal in the given paper would be the how reliant the predefined benchmarks and scenarios are. They may not completely capture the diversity of real-world ML inference use cases. Therefore, in the future we could incorporate more dynamic and different scenarios in order to get better idea of the complexities of the real-world deployment environments.

Question: How much impact can this standardized benchmarking framework actually have in various industries?

Srimathi Mahalingam