COMPSCI 690AB SYSTEMS FOR DEEP LEARNING

Reading Assignment 9

Proteus: A High-Throughput Inference-Serving System with Accuracy Scaling– Summary

The paper mainly addresses the challenge of efficiently serving the inference requests in resource constrained environment, while also trying to maintain high accuracy and meeting Service Level Objectives also known as SLOs. This problem seems to be crucial since the demand for AI powered applications are continuing to grow, which leads to increased pressure on the inference-serving systems to handle varying workloads without having to compromise on the accuracy. The major hypothesis of this work is that through dynamic scaling the accuracy of the model variants that are based on the workload demands, system throughput can be optimized while minimizing SLO violations and accuracy degradation.

The proposed solution in the paper, Proteus tries to leverage accuracy scaling as a means to handle varying query workloads. This is done by considering few solutions, including formulating the resource management problem as a mixed-integer linear programming optimization to adapt micro-scale variations, employing an adaptive batching algorithm to improve throughput and absorb micro-scale variations and decoupling resource allocation from the critical path of inference-serving to ensure responsiveness to workload changes. One key drawback of the proposal is that the reliance on a MLP solver for resource management might introduce overhead in solving complex optimization problem, mainly in the system scales. In order to address this limitation, one could explore alternative optimization techniques or optimizations specific to MILP solving to improve scalability and reduce overhead. Future work can also be done in exploration of fairness considerations as optimizing for accuracy scaling could inadvertently lead to unequal treatment of difference query types or users.

Srimathi Mahalingam