QoS AWARE AUTOMATED BATCH INFERENCE

Dilushan G.*, Riza M. S. I. and Jananie J.
Faculty of Engineering, University of Jaffna, Ariviyal Nagar, Sri Lanka
**2019e032@eng.jfn.ac.lk*

The complexity of the predominant deployment of machine learning inference and batch processing needs attention for careful resource utilization without degrading the performance. End user side, when the cost is considered and goes for the model inference on CPUs, the systematic setting of batching will be a challenge. Determining batch size requires analysis of different settings depending on various resource configurations, which becomes a challenge. Then the decision space becomes large and domain expertise may be necessary to satisfy the end users' cost and performance requirements. The performance of the inference mainly relies on the computation time of the given neural network, task processing density (p), and the compute power of the nodes (c). The experiment of comparing the estimated end-to-end inference time and real measured time of AlexNet and ResNet50 for a given configuration and different batch sizes results in different coefficients for the suggested batch sizes by the proposed, designed and implemented profiler. The inference time estimation formulas handled by the different researchers and our experiment results motivated us to propose a representation to place the right batch size to the right node is flops(NN)*p/c. Now the remaining challenge is how to manipulate the batching through these representations efficiently without degrading the performance. As another contribution, we design the optimizer which minimizes the end-to-end inference time considering the batch size. This research mainly focuses on the end users of cloud platforms where users can get pay-per-use configurations for their inferencing workloads. Our proposed representation and optimization technique could be extended in cloud platforms. While other batching techniques of different researchers mainly consider various scheduling mechanisms, our framework motivates the future researchers to deeply analyze the computation and the configurations. Hence the proposed system provides a configuration oriented batching mechanism and give mitigation techniques to improve the inferencing performance.

Keywords: *Batch scheduling, CPU clusters, Deep learning inference, Quality of service (QoS), Resource utilization.*