MLProxy: SLA-Aware Reverse Proxy for Machine Learning Inference Serving on Serverless Computing Platforms

Authors : Nima Mahmoudi, Hamzeh Khazaei

Abstract : Serving machine learning inference workloads on the cloud is still a challenging task at the production level. The optimal configuration of the inference workload to meet SLA requirements while optimizing the infrastructure costs is highly complicated due to the complex interaction between batch configuration, resource configurations, and variable arrival process. Serverless computing has emerged in recent years to automate most infrastructure management tasks. Workload batching has revealed the potential to improve the response time and cost-effectiveness of machine learning serving workloads. However, it has not yet been supported out of the box by serverless computing platforms. Our experiments have shown that for various machine learning workloads, batching can hugely improve the system's efficiency by reducing the processing overhead per request. In this work, we present MLProxy, an adaptive reverse proxy to support efficient machine learning serving workloads on serverless computing systems. MLProxy supports adaptive batching to ensure SLA compliance while optimizing serverless costs. We performed rigorous experiments on Knative to demonstrate the effectiveness of MLProxy. We showed that MLProxy could reduce the cost of serverless deployment by up to 92% while reducing SLA violations by up to 99% that can be generalized across state-of-the-art model serving frameworks.

Keywords : serverless computing, machine learning, inference serving, Knative, google cloud run, optimization
Conference Title : ICAC 2022 : International Conference on Autonomous Communications
Conference Location : Paris, France
Conference Dates : August 30-31, 2022

Open Science Index, Computer and Information Engineering Vol:16, No:08, 2022 publications.waset.org/abstracts/148895.pdf