Carnegie Mellon University
Computer Science Department

Speaking Skills Talk

Parity Models: Erasure-Coded Resilience for Prediction Serving Systems

Jack Kosaian

Friday, December 6, 2019
12:00 pm
GHC 6501

Machine learning models are becoming the primary workhorses for many applications. Services deploy models through prediction serving systems that take in queries and return predictions by performing inference on models. Prediction serving systems are commonly run on many machines in cluster settings, and thus are prone to slowdowns and failures that inflate tail latency. Erasure coding is a popular technique for imparting resource-efficient resilience against data unavailability in storage and communication systems. However, existing approaches for imparting erasure-coded resilience to distributed computation apply only to a severely limited class of functions, precluding their use for many serving workloads, such as neural network inference.

We introduce parity models, a new approach for enabling erasure-coded resilience in prediction serving systems. A parity model is a neural network trained to transform erasure-coded queries into a form that enables a decoder to reconstruct slow or failed predictions. We implement parity models in ParM, a prediction serving system that makes use of erasure-coded resilience. ParM encodes multiple queries into a "parity query," performs inference over parity queries using parity models, and decodes approximations of unavailable predictions by using the output of a parity model. We showcase the applicability of parity models to image classification, speech recognition, and object localization tasks. Parity models enable accurate reconstructions of unavailable predictions and reduce tail latency in the presence of resource contention. These results highlight the potential of parity models to open new doors in imparting resource-efficient resilience to prediction serving systems.

Presented in Partial Fulfillment of the CSD Speaking Skills Requirement.