CRUNCH Seminars at Brown, Division of Applied Mathematics

Friday – September 6, 2019

An empirical model of large batch training

Xiaowei Jin

In an increasing number of domains it has been demonstrated that deep learning models can be trained using relatively large batch sizes without sacrificing data efficiency. However the limits of this massive data parallelism seem to differ from domain to domain, ranging from batches of tens of thousands in ImageNet to batches of millions in RL agents that play the game Dota 2. To our knowledge there is limited conceptual understanding of why these limits to batch size differ or how we might choose the correct batch size in a new domain. In this paper, we demonstrate that a simple and easy-to-measure statistic called the gradient noise scale predicts the largest useful batch size across many domains and applications, including a number of supervised learning datasets (MNIST, SVHN, CIFAR-10, ImageNet, Billion Word), reinforcement learning domains (Atari and Dota), and even generative model training (autoencoders on SVHN). We find that the noise scale increases as the loss decreases over a training run and depends on the model size primarily through improved model performance. Our empirically-motivated theory also describes the tradeoff between compute-efficiency and time-efficiency, and provides a rough model of the benefits of adaptive batch-size training