SOCAMS 2019, CalTech, Saturday, April 27th, 2019

Quasi-Newton Optimization For Large-scale Machine Learning

Jacob Rafati
Electrical Engineering and Computer Science
University of California, Merced
http://rafati.net

Abstract:
Deep learning algorithms often require solving a highly non-linear and nonconvex unconstrained optimization problem. Methods for solving the optimization problems in large-scale machine learning, deep learning and deep reinforcement learning (RL) are restricted to the class of first-order algorithms, like stochastic gradient descent (SGD). The major drawback of the SGD methods is that they have the undesirable effect of not escaping saddle-points. Furthermore, these methods require exhaustive trial-and-error to fine-tune many learning parameters. Using second-order curvature information to find search directions can help with more robust convergence for non-convex optimization problems. However, computing Hessian matrices for large-scale problems is not computationally practical. Alternatively, quasi-Newton methods construct an approximate of the Hessian matrix to build a quadratic model of the objective function. Quasi-Newton methods, like SGD, require only first-order gradient information, but they can result in superlinear convergence, which makes them attractive alternatives to SGD. The limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) approach is one of the most popular quasi-Newton methods that construct positive definite Hessian approximations. In our research, we have proposed efficient optimization methods based on L-BFGS quasi-Newton methods using line search and trust-region strategies. Our method bridges the disparity between first order methods and second order methods by using gradient information to calculate low-rank updates to Hessian approximations. We have studied formal convergence analysis as well as empirical results on variety of the large-scale machine learning tasks, such as image classification tasks, and deep reinforcement learning on classic ATARI 2600 games. Our results show a robust convergence with preferred generalization characteristics as well as fast training time.