Neural network optimization strategies and the topography of the loss landscape

Alexandre Morozov[1]

[1]Rutgers University, Piscataway, United States

Artificial neural networks (ANNs) are general-purpose machine learning models inspired by networks of neurons in the brain. ANNs are trained by optimizing high-dimensional sets of fitting parameters on non-convex loss landscapes. Low-loss regions of the landscapes correspond to the parameter sets that perform well on the training data. A key issue in machine learning is generalizability - the performance of trained ANNs on previously unseen test data. Here, we investigate neural network training by stochastic gradient descent (SGD) - a non-convex global optimization algorithm which relies only on the gradient of the objective function. We contrast SGD solutions with those obtained via a non-stochastic quasi-Newton method, which utilizes curvature information to determine step direction and Golden Section Search to choose step size. We use several computational tools to investigate ANN parameters obtained by these two optimization methods, including kernel Principal Component Analysis and FourierPathFinder - a general-purpose algorithm for finding low-height paths between pairs of points on loss or energy landscapes. We find that the choice of the optimizer profoundly affects the nature of the resulting solutions. SGD solutions tend to be separated by lower barriers than quasi-Newton solutions, even if both sets of solutions are regularized by early stopping to ensure adequate performance on test data. When allowed to fit extensively on the training data, quasi-Newton solutions occupy deeper minima on the loss landscapes that are not reached by SGD. However, these solutions are less generalizable to the test data. Overall, SGD explores smooth basins of attraction, while quasi-Newton optimization is capable of finding deeper, more isolated minima. Our findings help understand the fundamental role of landscape exploration strategies in creating robust, transferrable ANN models.