Master's Thesis

Neural architecture search via sparse optimization

Guidance

Professor Nobuo YAMASHITA
Program-Specific Associate Professor Hiroyuki SATO

Ren Ito

Department of Applied Mathematics and Physics

Graduate School of Informatics

Kyoto University

February 2022

Abstract

In recent years, deep learning has been actively studied and has been successfully applied to various tasks such as image class classification and object detection. The architecture of neural networks for better performance is becoming more complex and large. Under such circumstances, Neural Architecture Search (NAS), which automatically constructs a structure of neural networks, has attracted much attention. The NAS can be formulated as a bilevel optimization problem consisting of a higher-level problem that optimizes the architecture of the neural network and a lower-level problem that optimizes the weights of the neural network. This bilevel optimization problem is non-convex and large-scale, and it takes a huge amount of time to optimize it. Differentiable architecture search (DARTS) is a gradient-based optimization method that significantly reduces the computation time. The DARTS represents a supergraph encompassing all the networks in the search space by a set of continuous architectural parameters. The computational cost is reduced by simultaneously learning the supergraph weights and structural parameters using the gradient-based optimization method, without learning the weights of individual networks separately. Unfortunately, sparsity of a solution is not taken into account when training on the supergraph, although the individual neural network is sparse. As a result, there is a difference between the optimal solution obtained by DARTS and the sparse network used in practical applications, and the performance may deteriorate.

In this paper, we propose a NAS method based on the gradient descent method that takes sparsity into account. The proposed model is an optimization problem based on the DARTS model with simplex constraints on the architectural parameters and an $\ell 1$-regularization term for the architecture parameters added to the objective function. With these additions, the difference between the optimal solution and the actual network can be reduced. In this paper, we conduct numerical experiments using the proposed method for the image classification problem using the CIFAR-10 dataset. The numerical results show that the proposed method yields sparse optimal solutions, and in addition, it tends to find architectures with better performance than DARTS for certain problems.