EVOLVING DEEP NEURAL NETWORKS ARCHITECTURES

Deep neural networks (DNNs) have been regarded as fundamental tools for many disciplines. Meanwhile, they are known for their large-scale parameters, high redundancy in weights, and extensive computing resource consumptions, which pose a tremendous challenge to the deployment in real-time applications or on resource-constrained devices. To cope with this issue, *compressing DNNs* for accelerating its inference has drawn extensive interest recently. The basic idea is to prune parameters with little performance degradation. However, the over-parameterized nature and the conflict between parameters reduction and performance maintenance make it prohibitive to manually search the pruning parameter space. In this talk, we will formally establish filter pruning as a multiobjective optimization problem, and propose a knee-guided evolutionary algorithm (KGEA) that can automatically search for the solution with quality tradeoff between the scale of parameters and performance, in which both conflicting objectives can be optimized simultaneously. In particular, by incorporating a minimum Manhattan distance approach, the search effort in the proposed KGEA is explicitly guided toward the knee area, which greatly facilitates the manual search for a good tradeoff solution. Moreover, the parameter importance is directly estimated on the criterion of performance loss, which can robustly identify the redundancy. In addition to the knee solution, a performance-improved model can also be found in a fine-tuning-free fashion. The experiments on compressing fully convolutional LeNet, VGG-19 and Inception networks validate the superiority of the proposed algorithm over the state-of-the-art competing methods.