Research on Model Auto-Growing for Dataset Adaptation

ABSTRACT

Neural network design often adopts a paradigm of preset model sizes and pre-training, a paradigm that requires users to make customized adjustments based on pre-trained models to meet their respective needs, exposing the contradiction between preset model sizes and the dynamic needs of users for specific datasets and inference efficiency.

Model auto-growing has the potential to find a model that can be adapted to a specific dataset with an adequate balance of performance and efficiency by gradually scaling the model during training. Based on the concept of adding more functional feature extraction structures for more important positions of models, this study explored the main factors affecting model auto-growing, including the depth and width growing strategy, the growing frequency, the growing termination conditions, and the initialization of the new modules, etc., and established an optimal auto-growing framework accordingly. Notably, this framework achieves simulated width growing for convolutional neural networks through a structural re-parameterization technique, which effectively avoids the efficiency loss resulting from actual width growing. Meanwhile, this framework is widely applicable to many dominant models based on the Transformer architecture. In addition, the framework simplifies the design of regularization, configuration of learning rates, and selection of optimizers for model auto-growing, which effectively improves training efficiency. This study further notes the challenge of high quantization loss faced by auto-grown models in quantization acceleration, which is caused by the phenomenon of low over-parameterization and anomalous parameter distributions due to structural re-parameterization. To this end, this study proposes an innovative quantization method that effectively reduces the performance loss during the quantization process through error compensation and weight rearrangement, providing a new solution for the efficient deployment of auto-grown models.

Experiments reveal that the model auto-growing framework excels, outstripping the optimal fixed-size model by 7.7% in accuracy and 205.5% in inference speed across 10 datasets spanning computer vision and natural language

processing. Compared to model pruning and alternative model auto-growing methods, it offers an accuracy enhancement of over 2.7% and a speed enhancement exceeding 45.9%. Additionally, employing the error compensation and weight rearrangement-based quantization method enhances accuracy by an average of 0.7% over conventional quantization approaches, under the same quantization settings.

Key Words: Dataset adaptation; Model auto-growing; Structural re-parameterization; Model compression

目 录

插图清单

附表清单