

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

摘 要.....	I
ABSTRACT.....	II
插图清单.....	VI
附表清单.....	VII
1 引言.....	1
1.1 研究背景与研究意义.....	1
1.2 国内外研究现状.....	2
1.3 本文主要工作和贡献.....	3
1.4 论文组织结构.....	4
2 相关理论与工作.....	5
2.1 神经网络架构.....	5
2.1.1 卷积神经网络架构.....	5
2.1.2 Transformer 架构.....	7
2.2 模型优化.....	11
2.2.1 神经网络剪枝.....	11
2.2.2 神经网络量化.....	13
2.2.3 结构重参数化.....	14
2.2.4 神经网络架构搜索.....	16
2.2.5 神经网络自增长.....	17
2.3 相关工作述评.....	19
3 数据集自适应的模型自增长方法.....	20
3.1 引言.....	20
3.2 模型自增长训练框架.....	21
3.3 卷积神经网络模型的自增长方案.....	23
3.3.1 卷积神经网络的自增长流程.....	23
3.3.2 模型自增长的决策机制.....	27
3.3.3 网络模块的选择和增长.....	32
3.3.4 初始化策略与优化器调整.....	35
3.4 Transformer 模型的自增长方案.....	38
3.4.1 Transformer 模型的自增长流程.....	38

3.4.2 Transformer 模型的自增长策略寻优	40
3.5 实验验证与案例分析	42
3.5.1 数据集介绍与实验配置	42
3.5.2 卷积神经网络模型的数据集适应性评估	44
3.5.3 Transformer 模型的数据集适应性评估	50
3.5.4 与模型剪枝的对比	51
3.5.5 与其他自增长方法的对比	54
3.6 本章小结	55
4 基于误差弥补和权重重排的自增长模型量化方法	56
4.1 自增长模型的量化挑战	56
4.2 自增长模型的量化方案	57
4.2.1 误差弥补	58
4.2.2 权重重新排布	58
4.3 自增长模型的量化加速实践	60
4.4 本章小结	62
5 结论与展望	63
参考文献	65
致 谢	74
作者简历及在学研究成果	75
独创性说明	76
关于论文使用授权的说明	77
学位论文数据集	78

插图清单

图 2-1 卷积神经网络的主要模块构成	5
图 2-2 Transformer 模型架构.....	8
图 2-3 Transformer 架构的关键模块组成.....	9
图 2-4 结构化、非结构化和半结构化剪枝	11
图 2-5 量化的映射过程	13
图 2-6 ACNet、DBB 和 RepVGG 的可重参数化特征提取模块.....	16
图 2-7 神经架构搜索的基本流程	16
图 2-8 基于随机深度的模型自增长方法	18
图 3-1 智慧生命体自然增长的智能发展模式	20
图 3-2 卷积神经网络的自增长流程	25
图 3-3 模型规模、自增长训练轮次与模型性能的关系	30
图 3-4 神经网络中原生存在的缩放的重要性	32
图 3-5 宽度增长中待插入分支的搜索空间	34
图 3-6 模型自增长中的新参数初始化和优化器调整需求	36
图 3-7 Transformer 的自增长流程.....	39
图 4-1 自增长模型的低过参数化现象	56
图 4-2 可重参数化模型的尖锐权重分布现象	57
图 4-3 基于误差弥补和权重重新排布的自增长模型量化方法	59

附表清单

表 2-1 经典卷积神经网络的创新结构优化.....	6
表 3-1 两次下采样时的自增长卷积神经网络模块组成.....	24
表 3-2 单次深度增长的规模对模型自增长性能的影响.....	27
表 3-3 单次宽度增长的规模对模型自增长性能的影响.....	28
表 3-4 增长后模型的训练轮次数对模型自增长性能的影响.....	29
表 3-5 终止条件对模型自增长的影响.....	31
表 3-6 深度增长策略对模型自增长性能的影响.....	33
表 3-7 宽度增长中待插入分支选取方式对模型自增长性能的影响.....	34
表 3-8 模型深度增长和宽度增长的协调匹配对性能的影响.....	35
表 3-9 新增参数初始化方法对自增长性能的影响.....	36
表 3-10 优化器调整策略对模型自增长性能的影响.....	37
表 3-11 Transformer 深度增长策略对模型性能的影响.....	39
表 3-12 Transformer 宽度增长方式对模型性能的影响.....	40
表 3-13 Transformer 模型单次增长后训练轮次数对性能的影响.....	40
表 3-14 Transformer 模型深度增长与宽度增长的配比对性能的影响.....	41
表 3-15 Transformer 模型新增参数初始化方法对自增长性能的影响.....	41
表 3-16 优化器调整对 Transformer 模型自增长性能的影响.....	42
表 3-17 实验涉及数据集介绍.....	43
表 3-18 模型自增长配置方案.....	44
表 3-19 CIFAR-10 卷积神经网络自增长效果.....	45
表 3-20 CIFAR-100 卷积神经网络自增长效果.....	45
表 3-21 SVHN 卷积神经网络自增长效果.....	46
表 3-22 MNIST 卷积神经网络自增长效果.....	47
表 3-23 ImageWoof 卷积神经网络自增长效果.....	47
表 3-24 ImagenetTE 卷积神经网络自增长效果.....	48
表 3-25 Tiny-Imagenet 卷积神经网络自增长效果.....	49
表 3-26 Mini-Imagenet 卷积神经网络自增长效果.....	49
表 3-27 计算机视觉任务上的 Transformer 模型自增长效果.....	50
表 3-28 自然语言处理任务上的 Transformer 模型自增长效果.....	51
表 3-29 ResNet 剪枝结果与模型自增长的对比.....	52
表 3-30 Vision Transformer 剪枝结果与模型自增长的对比.....	53

表 3-31 不同自增长方法的效果对比	54
表 4-1 自增长卷积神经网络的量化加速实践	60
表 4-2 自增长 Vision Transformer 的量化加速实践	61