密　　级：公开

加密论文编号：

论文题目：面向数据集自适应的模型自增长研究

学　　号：M202120812

作　　者：李冠辰

专 业 名 称：计算机技术

2024 年 5 月 29 日

面向数据集自适应的模型自增长研究

Research on Model Auto-Growing for Dataset Adaptation

研究生：李冠辰

指导教师：何杰

北京科技大学计算机与通信工程学院

北京 100083，中国

Master Degree Candidate：Guanchen Li

Supervisor：Jie He

School of Computer & Communication Engineering

University of Science and Technology Beijing

30 Xueyuan Road，Haidian District

Beijing 100083，P.R.CHINA

中图分类号：TP391　　学校代码：10008
U　D　C：　　密　　级：公开

北京科技大学硕士学位论文

论文题目： 面向数据集自适应的模型自增长研究

作者： 李冠辰

指　导　教　师： 何杰　**单位：** 北京科技大学　**职称：** 教授

指 导 小 组 成 员：　**单位：**　**职称：**

单位：　**职称：**

论文提交日期：2024 年 5 月 29 日

学位授予单位：北 京 科 技 大 学

摘 要

神经网络设计通常采取预设模型规模并预训练的构建范式，这种范式要求用户基于预训练的模型进行自定义的调整以满足各自的需求，暴露了预设模型规模与用户动态变化的数据集及推理效率需求之间的矛盾。

模型自增长通过在训练过程中逐步扩展模型的规模，有潜力找到能适应特定数据集的、充分平衡性能与效率的模型。本研究以为模型更重要的位置新增功能更丰富的特征提取结构为理念，深入探索了影响模型自增长的主要因素，包括深度与宽度增长策略、增长频率、增长终止条件以及新增模块初始化等，并据此建立了一套最优的模型自增长框架。其中，本框架通过结构重参数化技术，为卷积神经网络实现了模拟宽度增长，有效地避免了实际宽度增长将导致的效率损失。同时，本框架还广泛适用于基于 Transformer 架构的多种主流模型。此外，本框架还简化了模型自增长训练中的正则化设计、学习率配置和优化器选择等环节，有效提升了训练效率。本研究进一步注意到自增长模型在量化加速中面临的高量化损失的挑战，这是由低过参数化现象和结构重参数化引起的异常参数分布现象引起的。为此，本研究提出了一种创新的量化方法，通过误差弥补和权重的重新排布有效降低了量化过程中的性能损失，为自增长网络的高效部署提供了新方案。

实验数据表明：（1）本研究提出的模型自增长框架在涵盖计算机视觉和自然语言处理等十个数据集上，平均精度和推理速度分别比固定版本的最优模型提高了 7.7%和 205.5%。（2）与模型剪枝以及其他模型自增长方法相比，本研究所提框架展现出了超过 2.7%的精度优势和超过 45.9%的推理速度优势。（3）在相同的量化配置下，基于误差弥补和权重重新排布的量化方法平均比现有量化方法提升了 0.7%的精度。

关键词：数据集自适应；模型自增长；结构重参数化；模型压缩