
FCS 上海交通大学王肇国等：大语言模型压缩技术组合性的综合研究与形式化框架

作者：writer 来源：科学网

本文原地址：<https://www.iikx.com/news/progress/41513.html>

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

FCS

上海交通大学王肇国等：大语言模型压缩技术组合性的综合研究与形式化框架。论文标题：A study and formal framework of the composability of LLM compression techniques

期刊：Frontiers of Digital Education

作者：Gansen HU, Zhaoguo WANG

发表时间：17 Jul 2025

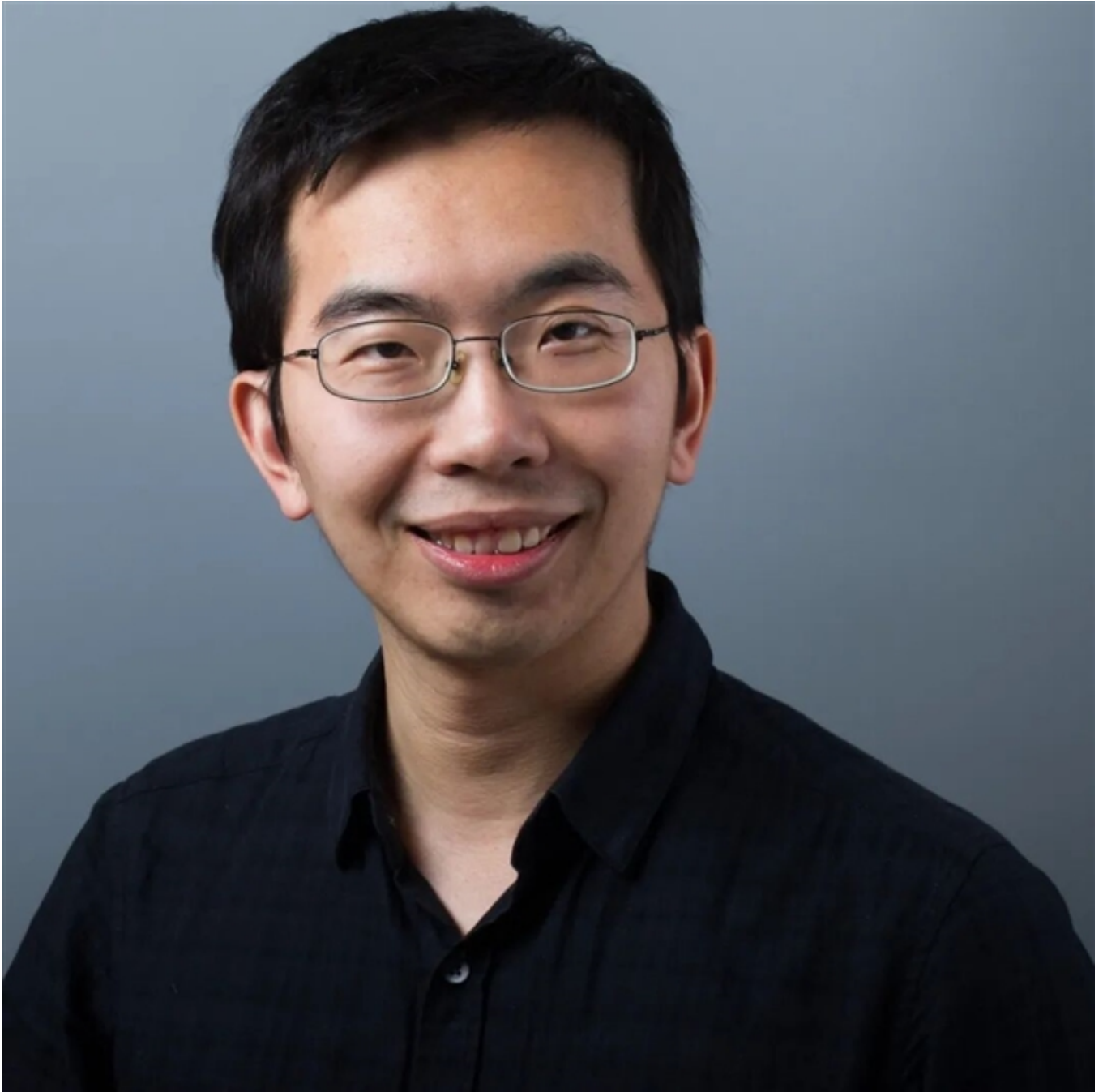
DOI：10.1007/s11704-025-50814-1

微信链接：[点击此处阅读微信文章](#)

FCS优秀青年计算机科学家论坛于2019年启动，以尊重科学贡献、传播更多优秀成果为宗旨。论坛作者审视自己的研究领域，介绍研究方向和研究进展。本论坛所有文章均为特邀稿件。

本期优秀青年计算机科学家论坛推出上海交通大学王肇国团队的成果——大语言模型压缩技术组合性的综合研究与形式化框架。

作者简介



王肇国，上海交通大学教授。主要从事分布式系统与数据库系统研究，成果发表在OSDI、SIGMOD、VLDB、NSDI、PPoPP、PODC等相关领域权威会议上。

A study and formal framework of the composability of LLM compression techniques

Gansen HU, Zhaoguo WANG✉

Institute of Parallel and Distributed Systems, Shanghai Jiao Tong University, Shanghai 200240, China

Received June 10, 2025; accepted July 15, 2025

E-mail: zhaoguowang@sjtu.edu.cn

© Higher Education Press 2026

引用格式：

Gansen HU, Zhaoguo WANG. A study and formal framework of the composability of LLM compression techniques. *Front. Comput. Sci.*, 2026, 20 (9) : 2009616

阅读原文：



问题概述

目前，像DeepSeek R1（6710亿参数）这样的大语言模型能够实现突破性的自然语言处理能力，但也需要海量的计算资源，这可能阻碍其在边缘设备上的部署。虽然单一的压缩方法（如剪枝、量化等）提供了部分解决方案，但其单独应用所能带来的提升有限。本文研究发现，对这些技术（尤其是上下文剪枝与量化）进行策略性的组合，可以在保持模型精度的同时，将模型大小大幅削减90%以上。尽管现有工作大多手动设计压缩算法组合方案，但手动设计组合方案效率低下且难以扩展，需一种自动化、系统化的框架。

技术步骤

通过综合分析研究32项关于组合多种大语言模型压缩技术的研究，本文发现了了观察：

- (1) 组合顺序至关重要：应用压缩技术的顺序会显著影响性能。这种影响在高稀疏度下可能更加显著。
- (2) 联合优化可能更优：统一压缩目标（例如，剪枝+量化）相较于顺序堆叠这些技术，可能减少累积误差。

为了自动化组合过程，本文构建了一个形式化框架的原型，能够在一定条件下自动搜索最优的压缩序列。其核心思想是将压缩操作转化为可兼容SMT求解器（如Z3的符号逻辑。本文的方法利用求解器来寻找最优顺序，从而减少了昂贵的试错和实验。

实验结果

该文的原型框架侧重于数学可保证的压缩。当前的工作处理小规模型矩阵，未来目标包括：

- 扩展到十亿参数级的 LLM。
- 整合低秩逼近/蒸馏算法。
- 延伸分析两种以上的组合，确定哪些技术组合能产生最佳性能。

- 研究其他特性（如鲁棒性、可解释性和泛化性等），以更全面地分析压缩技术间的相互作用。

期刊简介

Frontiers of Computer Science (FCS) 是由教育部主管、高等教育出版社和北京航空航天大学共同主办，南京大学支持，SpringerNature 公司海外发行的英文学术期刊。本刊于 2007 年创刊，月刊，全球发行。主要刊登计算机科学领域具有创新性的综述论文、研究论文等。本刊主编为周志华院士，执行主编为周昆教授。编委会及青年 AE 团队由国内外知名学者及优秀青年学者组成。本刊被 SCI、Ei、DBLP、INSPEC、SCOPUS 和中国科学引文数据库 (CSCD) 核心库等收录，为 CCF 推荐 B 类期刊；两次入选中国科技期刊国际影响力提升计划；入选第 4 届中国国际化精品科技期刊；两次入选中国科技期刊卓越行动计划（一期梯队、二期领军）。

来源：Frontiers of Digital Education

更多 科学进展 请访问 <https://www.iikx.com/news/progress/>

本文版权归作者所有，请勿用于商业用途，[爱科学iikx.com](https://www.iikx.com)转发