

蔡泽凡 | Zefan Cai

zefncai@gmail.com · Google Scholar

学术简历 · 中文版

教育经历

威斯康星大学麦迪逊分校 · 计算机学院 | 2024 - 至今

计算机科学博士在读

导师：计算机、数据与信息科学学院 Junjie Hu 教授

北京大学 · 软件与微电子学院 | 2022 - 2024

计算机科学与技术硕士；GPA：3.71/4.0 (前 5%)

导师：北京大学计算语言学研究所 (ICL) 常宝宝教授

北京交通大学 · 计算机与信息技术学院 | 2018 - 2022

计算机科学学士 (工学)；GPA：3.61/4.0 (排名 2/31)

论文、投稿与预印本

此前研究主要包括 KV 缓存压缩 [1][2][12]、效率 [1][2][8][12]、指令微调 [3][4][5]、推理 [14][17]、评估 [10]、偏好学习 [21]、智能体 [15][18]、视觉语言理解 [4][8][16][18]、视觉语言生成 [13]，以及传统自然语言处理 [6][7][9][20]。

1. KVCache-Factory：面向自回归模型的统一 KV 缓存压缩方法

Zefan Cai

开源项目 | [代码](#)

2. PyramidKV：基于金字塔式信息汇聚的动态 KV 缓存压缩

Zefan Cai, Yichi Zhang, ..., Baobao Chang, Junjie Hu, Wen Xiao

预印本 | [PDF](#) · [代码](#)

3. 提升零样本事件检测中的事件定义遵循能力

Zefan Cai*, Po-Nien Kung*, ..., Wei Wang, Nanyun Peng

ACL 2024 主会长文 | [PDF](#) · [代码](#)

4. MMICL：通过多模态上下文学习增强视觉语言模型

Haozhe Zhao*, Zefan Cai*, ..., Zixuan Liu, Sheng Wang, Wenjuan Han, Baobao Chang

ICLR 2024, 海报 | [PDF](#) · [代码](#) · [数据](#) · [模型](#)

5. 大语言模型的组合式任务表示

Nan Shao*, Zefan Cai*, Chonghua Liao, Yanan Zheng, Zhilin Yang

ICLR 2023, 海报 | [PDF](#) · [代码](#)

6. DialogVCS：对话系统升级中的鲁棒自然语言理解

Zefan Cai*, Xin Zheng*, ..., Baobao Chang, Yunbo Cao

NAACL 2024 主会长文 | [PDF](#)

7. 通过教师语言选择与跨语言蒸馏缓解多语言预训练语言模型的语言间性能差异

Haozhe Zhao*, Zefan Cai*, Shuzheng Si, ..., Baobao Chang

NAACL 2024 主会长文 | [PDF](#) · [代码](#)

8. VeCAF：由视觉语言模型增强、感知训练目标的协同主动微调

Rongyu Zhang*, Zefan Cai*, ..., Kurt Keutzer, Baobao Chang, ..., Shanghang Zhang

ACMMM 2024 | [PDF](#)

论文、投稿与预印本 (续)

9. SANTA : 针对远程监督命名实体识别中不准确与不完整标注噪声的独立策略

Shuzheng Si*, Zefan Cai*, ..., Jiaying Lin, Baobao Chang

ACL 2023 Findings 长文 | [PDF](#) · [代码](#)

10. 大语言模型并非公平的评估者

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, ..., Zhifang Sui

ACL 2024 主会长文 | [PDF](#) · [代码](#)

11. PCA-Bench : 在感知、认知与行动链中评估多模态大语言模型

Liang Chen, ..., Zefan Cai, ..., Baobao Chang

ACL 2024 Findings 长文 | [PDF](#) · [代码](#)

12. 并非所有注意力头都同样重要 : 融合检索与推理的注意力头级 KV 缓存压缩方法

Yu Fu, Zefan Cai, ..., Wen Xiao

预印本 | [PDF](#)

13. 视觉语言智能的火花 : 用于高效精细图像生成的二维自回归 Transformer

Liang Chen, Sinan Tan, Zefan Cai, ..., Baobao Chang

预印本 | [PDF](#)

14. 通过思维链实现人在环中参与

Zefan Cai, Baobao Chang, Wenjuan Han

预印本 | [PDF](#)

15. ML-Bench : 大语言模型利用开源库完成机器学习任务

Xiangru Tang*, Yuliang Liu*, Zefan Cai*, ..., Baobao Chang, ..., Arman Cohan, Mark Gerstein

预印本 | [PDF](#) · [代码](#) · [项目主页](#)

16. DiffCap : 探索图像描述生成中的连续扩散

Yufeng He*, Zefan Cai*, Xu Gan, Baobao Chang

预印本 | [PDF](#) · [代码](#)

17. 大语言模型评审器助力发现数学错误 : 利用自然语言反馈构建更好的数学验证器

Bofei Gao, Zefan Cai, ... Baobao Chang

预印本 | [PDF](#)

18. COMMA : 具备通信能力的多模态多智能体基准

Timothy Ossowski, ..., Zefan Cai, ..., Junjie Hu

预印本 | [PDF](#)

19. Omni-Math : 面向大语言模型的通用奥林匹克级数学基准

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, ...

预印本 | [PDF](#)

20. CENSOR : 基于不确定性感知教师学习与学生间协同学习的远程监督命名实体识别

Helan Hu, ..., Zefan Cai, Baobao Chang

ACL 2024 Findings 长文 | [PDF](#)

21. 迈向大语言模型偏好学习的统一视角 : 综述

Bofei Gao, Feifan Song, Yibo Miao, Zefan Cai, ... Baobao Chang

预印本 | [PDF](#) · [代码](#) · [项目主页](#)

研究项目

威斯康星大学麦迪逊分校 | 导师：Junjie Hu | 2024 年 9 月 - 至今

KVCache-Factory：面向自回归模型的统一 KV 缓存压缩方法 [1]

- 创建面向基于 Transformer 的基础模型的高效 KV 缓存优化库。
- 统一实现不同阶段、不同类别的 KV 缓存压缩方法。
- 涵盖预填充 (Prefilling)、KV 缓存管理与解码 (Decoding) 阶段的加速方法。
- 在 KV 缓存管理方面，涵盖合并 (Merge)、量化 (Quantization) 和淘汰 (Eviction) 等压缩方法。

PyramidKV：基于金字塔式信息汇聚的动态 KV 缓存压缩 [2]

- 观察到大语言模型的信息聚合呈现金字塔式信息汇聚：低层注意力广泛分散，高层最终聚焦于关键词元。
- 开发 PyramidKV，在不同层之间动态调整 KV 缓存大小。
- 实验表明，PyramidKV 表现优异，尤其在资源需求较高的情况下，并能有效保留长上下文能力。
- 实现表明，PyramidKV 可与预填充阶段的方法 (如 MInference)、其他 KV 缓存管理方法 (如 KV 缓存量化) 及不同框架 (如 Huggingface 和 vLLM) 集成。

加利福尼亚大学洛杉矶分校 | 导师：Nanyun Peng | 2023 年 6 月 - 9 月

提升零样本事件检测中的事件定义遵循能力 [3]

- 探索基于大语言模型的事件检测模型能否根据给定的事件定义泛化到未见过的事件。
- 通过在不同规模下训练大语言模型，评估事件抽取的缩放规律；规模维度包括事件类型数、样本数及事件定义数。
- 使用大语言模型生成的事件类型、事件定义与样本微调大语言模型，并在未见过的事件类型上评估。

荣誉与奖励

北京大学三好学生 (前 10%) | 2023 年 9 月

北京交通大学三好学生 (前 10%) | 2020 年 9 月

学术服务

审稿人：EMNLP、ICML、ICLR、NeurIPS、ARR