唐子聪

电话: 187 727 22454 **邮箱:** tangzc@whu.edu.cn

课程学习

武汉大学 (WHU)，本科生 **武汉, 中国**
计算机科学与技术 *(US News 排名 16)* 2022 年 9 月–至今
均分&绩点：94.15/100.0，Rank **1**/240；**3.97**/4.00，Rank **2**/240
核心课程：高等数学 100，线性代数 100，离散数学 98，操作系统 98，软件设计与体系结构 98，编译原理 98，高级语言程序设计 97，数据库系统 96，计算机网络 96
研究兴趣：大语言模型（LLMs），多模态大语言模型（MLLMs），上下文压缩（KV Cache Compression），多模态表征学习（Multi-Modal Representation），投机解码（Multi Token Prediction）

论文发表

SpindleKV: A Novel KV Cache Reduction Method Balancing Both Shallow and Deep Layers **Preprint**
– **Zicong Tang**, Luohe Shi, Zuchao Li, Baoyuan Qi, Guoming Liu, Lefei Zhang
– **ACL2025** 在审，一作。ARR 评分 3，3，4，AC 评分 4。

研究经历

多模态大语言模型上下文压缩和计算效率提升 2025 年 2 月 - 至今
武汉大学，*Sigma Lab* 指导老师：李祖超
– 探索多模态大模型中视觉和文本上下文的特点和区别,从空间和时间两个维度对视觉上下文的特点进行研究。
– 开发一种跨帧合并的方法同时在空间和时间上对视觉上下文进行压缩，以此减少显存访问瓶颈，提升计算效率。

平衡深浅层冗余的大语言模型 KV Cache 优化方法 2024 年 9 月 - 2024 年 12 月
武汉大学，*Sigma Lab* 指导老师：李祖超
– 发现了存在于大语言模型 KV Cache 中的两种冗余：KV Cache 中不重要序列的冗余和 KV Cache 构造成分的冗余，分别表现为注意力分数的稀疏性和令牌之间较高的余弦相似性。
– 提出了令牌驱逐与合并向结合的方法：通过基于注意力分数的驱逐策略和基于余弦相似度的替换策略减少不同层中存在的不同冗余并且在同类方法中达到最优表现。
– 解决了被以往方法忽略的难题：我们的方法解决了以往的驱逐策略在兼容分组查询注意力（GQA）机制时面临的挑战并在 GQA 结构的模型中远超基线方法。

荣誉和奖励

– **国家奖学金**, 全国前 2% 2023 年 10 月
– 甲等奖学金, 两次, 全校前 5% 2023 年 10 月 & 2024 年 10 月
– 三好学生, 两次, 全校前 5% 2023 年 10 月 & 2024 年 10 月
– 二等奖, 队长, 计算机设计大赛中南地区赛 (4C) 2024 年 7 月

项目经历

计算机视觉 & 自然语言处理 (课程项目) 2024 年 7 月
– 用 LeNet 在 CIFAR-10 数据集上进行图像分类并且通过调整模型结构显著提升了分类效果。
– 使用 QLora 方法在 AdvertiseGen 数据集上微调 chatglm2-6b 模型并对主模型采用 4-bit 量化。

RISC-V 指令集 CPU 设计 (课程项目) 2024 年 5 月
– 使用 Verilog 语言设计了一个包含 IF/ID/EX/MEM/WB 五个阶段的五级流水线 CPU.
– 实现了 RISC-V 指令集中的算术逻辑运算、分支跳转和访存等指令并在 Nexys A7 fpga 上成功运行。

任职 & 活动

– 优秀青年志愿者，132h 志愿时长 2024 年 8 月
– 武汉大学跆拳道协会秘书部长 2023 年 10 月 - 至今
– 优秀助教，担任数据结构课程助教 2024 年 3 月 - 2024 年 7 月
– 暑期支教，贵州遵义凤岗二中支教物理和人工智能介绍 2024 年 7 月 - 2024 年 8 月