赵英灼

(+86) 156-0337-9828 上海
28 岁 男 zhuozhuozhuo_97
推理引擎高级研发工程师
github.com/caijizhuo
zhao_yingzhuo@foxmail.com
https://caijizhuo.github.io

教育背景

2020.12 – 2019.09 **伦敦大学学院(University College London)** · 互联网工程 · 硕士
成绩优异，毕业时被授予 Distinction 一等荣誉学位 。

2019.09 – 2015.09 **东北大学(NorthEast University)** · 计算机科学与技术 · 本科学位
成绩优异，多门专业课高分，GPA：3.93（专业前 5%）。取得过河北省数学竞赛二等奖，蓝桥杯 编程竞赛三等奖，多次获得校综合奖学金。

专业技能

操作系统 Linux, Windows, Mac

编程语言 Python, C/C++, Shell

开发工具 Git, Vscode, Tmux, Vim, Gdb, Docker, CMake

英语水平 英语（CET-6）雅思（6.5）

工作经历

无问芯穹智能科技有限公司 高级研发工程师 2023 年 09 月 – 至今

C++, Python, Shell

推理计算组（上海）

- **从 0 到 1** 的构建自研大语言模型推理引擎：infini_llm（后更名为 alioth_llm)。其中，本人担任 runtime 阶段的 commiter，调研了竞品框架，主导运行时阶段的基础逻辑设计和编码，包含传入 CUDA 算子前的数据结构设计以及执行逻辑设计。
- 维护代码质量，为自研推理引擎仓库**搭建 CI/CD** 并添加测试用例，保障代码提交流程。
- 参与**从 0 到 1** 的构建推理引擎的 serving 模块，使得用户可以通过 openai api 格式的报文访问推理服务。
- 参与设计 prefix caching 模块编码，使得引擎面对相同前缀报文时可以复用已经缓存的 kv cache blocks。
- 自研引擎对接量化模型，支持了 AWQ-LLama 和 AWQ-DeepSeek-V1 的构图阶段，以及 FP8-Qwen1.5 到 FP8-Qwen2.5 全系列模型运行（W8A16，对接 marlin 算子)。其中 FP8 系列模型在**不减少精度的情况下节省 50% 的显存**，降低推理成本。
- 担任 Maas（Model As A Service）工作开发，参与发布镜像若干，包括：用户授权登陆功能，用户无授权码无法访问我们自研推理服务；对接 OpenTelemetry 功能，使得推理服务网络部分执行的详细时间可以上传远端系统查看；修复 bug 若干；
- 参与 ComfyUI 镜像的性能优化，优化自定义节点性能，通过属性劫持的方式添加模型加载缓存，使得该节点**热启动运行时间从 5 sec/次到 0.1 sec/次**。
- 阅读竞品代码，包含 **vllm** 和 **ComfyUI**。熟悉框架中的运行流程，可以根据已有代码做出自定义化的代码修改。

- **贡献项目 vllm**，发现该项目中存在流式 API 接口无法正常返回数据的 bug 并向社区提 pr 修复:https://github.com/vllm-project/vllm/pull/2756 。**该 pr 被 vllm 社区成员认可并合入主线分支**。

华为技术有限公司 软件开发工程师 2021 年 01 月 – 2023 年 06 月

C, C++, Python, Shell

2012 实验室 · 分布式与并行计算实验室（上海）

- 参与开源项目：MindSpore (AI 框架）的研发与维护。累计合入 PR 60+。主要工作内容为 ONNX/TF 模型在 Mindspore 上的推理**支持、精度、性能**调优。同时维护优化以及支持新算子。
- 负责语音模型在晟腾芯片的精度调测，并主动思考优化程序结构，减少推理时间。

数据通信产品线 · 子卡软件开发（南京）

- 编写每日构建大包的一键部署脚本，提升流水线每日用例部署效率 2h/天。
- 编写仿真建模脚本，快速生成建模 xml 文件，提高部门新增仿真产品形态时的适配效率。
- 仿真代码测试验证，为保证仿真设备 rpc 接口的正确建立连接，添加了守护进程。

其他编码经历

技术博客 在 github 上利用 hexo + markdown 的方式搭建维护个人博客。**博客链接** 2022 年 02 月 – 至今

YASTL 研读开源项目 Tiny STL 的源代码以及参考书籍《STL 源码剖析》，并自己对项目额外加了注释，且发现解决了 deque 内存泄露问题。**仓库地址** 2022 年 10 月 – 2023 年 01 月

室内智能系统 在硕士就读期间，和同学合作完成 IOT 设备的智能系统，个人负责云端 Node red 数据整合处理以 及温度传感器模块的设计。 2020 年 01 月 – 2020 年 02 月

关于我

- 热爱技术，自工作起时刻仍保持学习，阅读过书籍《程序员的自我修养》、《深度学习的数学》、《Linux 高性能服务器编程》、《深入理解计算机系统》(CSAPP)、《Unix 环境高级编程》(APUE)、《STL 源码剖析》。
- 喜欢分享，搭建个人博客并分享知识。
- 喜欢运动，保持健身运动习惯。