张立宇

(+86)177-2798-3753 | chrisly.zhang@outlook.com | github.com/chrislyz

教育经历

新南威尔士大学 | 信息技术工程 | *硕士研究生* 2019/06 - 2020/09

Distinction 优异成绩毕业并曾担任计算机图形学助教, 主要课程包括: 计算机图形学、计算机视觉、信息检索与网络搜索、数据仓库与数据挖掘、统计机器学习高等课题、大数据管理

新南威尔士大学 | 计算机科学与人工智能 | *理学学士* 2015/06 - 2019/04

主要课程包括: 编程语言与编译器、人工智能、神经网络与深度学习、网页数据压缩与搜索、数据库系统实现等

工作经历

智寻 | *机器学习工程师* 2023/06 - 2024/04

- 利用开源框架（Llamaindex 等），成功搭建了一个高效且可扩展的检索增强生成（Modular RAG）自主知识库
- 通过整合 OpenAI 的 API，借鉴 AutoGPT 实现了 Prompting 工程的优化，有效提高了语言模型的响应质量和效率。通过对 Prompt 的设计和调整，使得模型能够更精准地理解和执行复杂指令，显著提升了用户体验和满意度
- 采用 LoRA 以及 DPO 强化学习对指令微调模型进行 SFT 和 RLHF，以适应多场景任务的需求
- 研究大模型分布式训练以及推理优化的实现技术，分析 TP/SP/PP 的使用场景，CUDA 融合算子和 PTQ 量化等

华为技术有限公司 | *算法/研发工程师* 2020/11 - 2022/04

- **主导电子纸智能交互系统核心模块研发**，攻克鱼眼摄像头动态捕捉点阵坐标的实时识别难题。基于 OpenCV 框架设计多级图像处理算法，创新融合分水岭算法与形态学腐蚀技术，实现复杂光照下的亚像素级点阵定位；开发低延时处理管线，通过完成图像转正-分割-解码全流程验证，将单帧处理时间从 120ms 压缩至 28ms，成功支持项目立项。后续成功验证基于阿诺德置乱和离散余弦变换将数字点阵作为隐水印写入电子纸中
- **研发基于 Mesa 3D 的新一代异构 GPU 图形驱动架构**，支撑华为自研 PC 端 GPU 实现 OpenGL 4.X 规范兼容。针对多线程渲染性能瓶颈，基于开源框架实现"Vulkan 底层渲染管线 +OpenGL 高层 API 映射" 双引擎架构，基于 SPIR-V 中间语言构建着色器转译层，通过 Command Buffer 并行化改造将几何绘制与光栅化任务解耦；设计硬件指令级批处理机制，结合 Vulkan 显式多队列特性提升 GPU 利用率，最终驱动通过 Piglit - OpenGL Driver Testing Framework 137 项严格测试，单指令流渲染效率较传统方案提升 3 倍，支撑海思 Kirin GPU 芯片流片前关键验证
- **优化华为实时光追渲染器 CUDA 计算效能，攻克追光渲染器内存带宽瓶颈的问题**。基于 NSight 工具链开展全栈性能剖析: 通过 Nsight Compute 分析 PTX 中间代码，定位因临时变量频繁创建引发的寄存器溢出问题，重构着色内核改用左值引用策略，降低 25% 寄存器压力；根据场景复杂度动态修改启动内核时的 BlockDim 优化可达成的 Occupancy；重新修改内存排布形式为 Tiled SOA，使 Memory Coalescing 效率从 50% 提升至 89%。最终单帧渲染时延从 400ms 降至 268ms（降幅 33%）

项目经历

业余写作 | Freelance

- Happyville个人英语写作技术分享博客
- 10 周快速了解计算机图形学

经典机器学习及神经网络实现

- 使用 Numpy 实现 KNN, K-means 等聚类算法以及感知机, 朴素贝叶斯等简单网络模型
- 使用 TensorFlow 实现 CNN 识别手写数字以及借助 OpenAI Gym 库实现 Deep Q-Learning 等强化学习算法
- 通过 NLTK 库实现基于 TF-IDF 特征和 XGDBoost 分类算法的传统文本分类器系统

DRIVE **数字视网膜图像血管提取** |*Kaggle 比赛项目* 2019/10 - 2019/11

- 分析来自 DRIVE 数据集的视网膜图像数据，进行血管分割
- 使用包括 CNN、FCN、V-Net 和 U-Net 在内的知名深度神经网络模型进行训练，并比较不同模型之间的准确性
- 使用传统 CV 算法如 Hough 圆等提取瞳孔信息并与深度学习方法进行对比

Burrows-Wheeler-Transform 2018/10 - 2018/11

- 通过 3-way Quicksort 算法加速 BWT 编码速度, 将编码 50MB 文件的时间缩短至 10 秒以内
- 实现了 BWT 反向搜索，可以高效搜索 BWT 编码的文件而无需解码，同时提供解码器，以无损方式解码 BWT 编码的文件
- 该实现仅使用不超过 30 兆字节 (MB) 的运行时内存来编码不超过 200 兆字节的文件, 实现内部通过磁盘 Offload 写入和读取临时文件来编码大文件

个人总结

作为一名专注于自然语言处理与大型语言模型研发的机器学习工程师，我致力于将技术创新与工程实践相结合。在 BERT、GPT 等架构应用方面拥有扎实经验，参与设计与优化不同规模的对话系统与文本生成项目，注重语义理解鲁棒性与知识融合的落地效果。熟悉分布式训练框架的部署调优，曾在模型推理优化方向积累实践经验。工作方式上注重系统性沉淀，建立了模型迭代的技术文档库，定期整理实验分析与行业动态，并通过建立规范的代码审查流程促进团队协作效率。生活中坚持跑步健身，将运动培养的专注力融入技术攻坚。

- 2021 年获得华为明日之星奖项
- 2019 年获得新南威尔士大学计算机图形学课程高荣誉奖 (Top-5)

技术能力

- **语言**: 编程不受特定语言限制。常用 Python, C/C++, CUDA；熟悉 JavaScript；了解 Lua, Java 等
- **工作流**: Linux, Shell, (Neo)Vim, Git, GitHub
- **其他**: 熟练深度学习框架 PyTorch；熟悉 Huggingface；熟悉 GPGPU 硬件