段清楠 | 个人简历

Email: duanqn96@gmail.com
GitHub: https://github.com/duanqn

教育经历

清华大学 计算机科学与技术系 本科　　2014.8-2018.7

- GPA: 91 / 100

清华大学 经济学 辅修　　2015.9-2018.7

滑铁卢大学 计算机科学 硕士　　2018.7-2020.8

- 导师：Bernard Wong, Srinivasan Keshav 方向：共识协议/区块链
- 毕业论文：In Search of a Scalable BFT Protocol for Global Deployment

工作经历

Microsoft Beijing

云计算与人工智能　　2020.9 至今

任职于 OCR 产品组，负责开发 Azure Document Intelligence 系列产品。

2020.9-2020.12　负责辅助开发训练/测试数据集的管理系统及对应的 python 命令行工具

2020.12-2021.5　负责维护和改进整个项目的构建系统，包括不同平台的编译、打包、生成 docker 镜像等

2021.5-2022.1　分析并优化系统性能，减少最坏情况下的运行时间。

- 对于多边形 NMS，重写 IOU 的计算逻辑进行加速，并使用 RTree 减少冗余计算
- 对堆上频繁分配的小对象启用内存池
- 限制单张图中最多识别的文本行数

2022.1-2023.1　条形码/二维码识别

2022.9-2023.2　探索 GPU 推理，校验模型在 CPU/GPU（TensorRT）的推理结果，使用半精度浮点数压缩显存占用。

2023.2-2023.7　对表格检测模型进行性能优化

- HRNet 中双线性插值由指定尺寸改为指定倍数（长、宽各扩大为 2 倍）
- 对 HRNet 部分进行量化

2023.8-2023.10　探索在 A10 显卡上的 GPU 推理

2023.10-至今　对基于 DETR 的下一代检测器模型进行产品化

- 打通包括 ONNX 模型导出、项目构建、识别准确率/性能测试的全部工程系统
- 实现包括 NMS 在内的运行时逻辑
- 探索 INT4/INT8 量化
- 为 CPU 推理实现 Multi-Scale Deformable Attention，达到 20x 加速
- 为 CPU 推理实现 Flash Attention

工作技能

熟练掌握 C++，有使用 AVX intrinsic 编写代码的经验，能读懂简单的 x86/x64 汇编代码。了解计算机底层细节，能使用 profiler 分析程序的性能瓶颈。

对 Python 等脚本语言的掌握可满足日常需要，了解如何编写 Python 包。