张望

+86-1896-980-3679 wang.zhang@aol.com

教育背景

卡内基梅隆大学, 匹兹堡, 美国

机械工程 硕士, 2014年1月 - 2015年12月

GPA: 3.85 / 4.0

浙江大学, 杭州, 浙江

应用力学 学士, 2009年9月 - 2013年7月

GPA: 3.46 / 4.0

工作经历

高级软件工程师, 才云科技有限公司 — 2018年12月 - 2020年4月

- 设计、编写基于 Kubernetes 的深度学习模型 Inference 平台，实现模型自动转换、弹性伸缩、AB 测试、请求记录、反馈记录、针对 inference 设备的集成优化
- 设计、编程适用于 Kubernetes GPU 共享方案，实现 GPU 在容器间的共享和显存限制
- 针对 PS/Worker 和 AllReduce 两种分布式训练模式，根据带宽提供亲和性调度
- 设计面向 AllReduce 式分布式训练的可容错训练方案: github.com/caicloud/ftlib

高级系统工程师, 网易·人工智能事业部 — 2018年5月 - 2018年12月

- 利用 Conda 实现轻量化向 HPC 集群提交深度学习任务
- 针对 Slurm 集群改造 Caffe 以执行非均衡分布的训练任务
- 提供分布式训练加速和基于 TensorRT 的推理加速，取得美团视频多标签分类第一名

销售工程师, Wolfram Research — 2016年2月 - 2018年4月

- 在 Modelica 中添加神经网络推理模块
- 利用 Modelica 模型制作Plant用以强化学习

研究助理, Particle Flow and Tribology Lab, CMU — 2015年7月 - 2016年1月

- 用 CUDA-C 编写基于沉浸边界法的 3 维结构网格求解器较之单线程加速 80 倍（GTX970）
- 针对流固耦合问题，增加了粘性力的计算, 克服 GPU 单精度浮点数造成的截断误差
- 利用 OpenCL 和 MPI 重构求解器

技能

编程语言: Python, C++, Go, Shell

Github: github.com/zw0610