

李天天



个人信息

- 性别：男
 - 手机：18092430560
 - 专业：机器人技术工程
- 年龄：28
 - 邮箱：lixiamomo@outlook.com
 - 岗位：通软/底软件开发工程师

工作及教育经历

- 华为2022.8~至今
 - 西安电子科技大学2019.9~2022.7
 - 西安工业大学2014.9~2018.7
- 机器人技术工程-研究生
 - 机械电子工程-本科

个人介绍

- 性格开朗，拥有强烈的技术追求，具备较强的学习能力和求知欲。精通端侧AI软件相关技术栈并多次获得部门奖项，具有一定的 AI Infra 算法及 AI 算法的项目积累。
- AI软件开发**：精通 C/C++ 语言的架构设计与开发，拥有丰富的多线程编程和性能优化经验。入职两年来获得**华山论剑个人金奖**、**上海海思研发好作品奖**、**芯火奖**等多个荣誉。负责AIPQ-Service模块的架构设计与开发，以及AI软件栈的重构方案设计与开发；熟悉驱动开发，具备多颗海思画质芯片的驱动维护与开发经验。
- AI Infra**：掌握AI Infra相关技术栈，**独立开发**  **Needle神经网络推理库**。了解CUDA和OpenCL算子开发，熟悉反向传播算法及常用高性能算子优化，如Conv(im2col高并行度)、Matmul (2D tiling 达到 cuBLAS 70性能)。
- AI算法**：熟悉常用神经网络算法的原理，能够独立复现常见、经典AI算法，曾复现过  **Vision Transformer**、FCOS等经典算法。能够熟练运用**迁移学习**、**模型蒸馏**等模型训练与优化技术。

获奖经历

- 华山论剑个人金奖（2024）(所有获奖者中唯一14级，其余获奖者均为16+) [\[获奖照片\]](#)[\[华为内网链接\]](#)
- 上海海思 研发好作品奖 [\[获奖照片\]](#)
- 芯星奖 (2024) [\[华为内网链接\]](#)
- 互联网+省级金奖 [\[获奖照片\]](#)
- 互联网+省级银奖 [\[获奖照片\]](#)
- 研究生二等奖学金（2020）
- 研究生二等奖学金（2019）

项目经历

AI软件方向

AIPQ-Service - [华为-海思] - 软件架构设计、开发 - 2024.2-2024.8

项目背景：通过用户态进程运行分类网络，动态计算图像优化参数，以提升码流播放场景下的图像质量。

项目成果：

- 架构设计获得 **华山论剑个人金奖**；
- 在 45 天内完成方案架构设计，开发代码 9000+ 行，模块上线一年内仅收到一例提单；
- 利用 AIPQ 特性，成功使 6xx 系中端芯片与 8xx 系芯片在画质效果上具有相似竞争力。

技术方案简述：通过用户态守护进程实时多线程运行多组 AIPQ 模型，动态计算图像优化参数。

- **技术难点：**
 - 实现 OpenCL 算子编写，满足模型前后处理需求。
 - 对性能要求极高，根据场景不同，需要并行运行多组 AIPQ 算法，部分算法需要多 worker 协作，以高并发掩盖数据处理过程，同时确保各个场景下模块的资源管理和线程安全。
 - 考虑到架构演进，需要处理多芯片、多形态算法的差异，对架构设计提出更高要求。
- **算法与性能优化：**
 - **性能优化：**设计 DMA 零拷贝方案、模型硬化方案等，有效降低内存拷贝次数，**优化 DDR 带宽 50%**；识别并推动 Cast 取帧方案落地，**优化 CPU 占用 12%**。
 - **算法优化：**通过迁移学习和模型蒸馏，将模型准确率从 **83% 提升至 89%**。

AI 软件栈重构 - [华为-海思] - 开发 - 2023.10-2024.2

项目背景：旧架构严重腐化，测试发现 20 余个致命问题，功能基本处于不可用状态，同时旧架构实现与 Ascend 生态耦合。本次重构旨在重新设计和实现软件栈，解决现有架构问题的同时，以适配 AI-HAL 模块，支持多家AI生态的导入。

项目成果：

- 成功完成 AI 软件栈重构，核心代码上库 6000+ 行，测试用例代码 2 万+ 行。
- 上线后首轮测试问题单数量仅 2 例，系统稳定性显著提高。

技术方案简述：AI软件栈为客户提供JAVA与NDK C++两种接口，通过安卓 Binder 机制 System 和 Vendor 侧进程通信，完成神经网络推理的同时，支持后续架构演进。

- **技术难点：**
 - 安卓架构的通信机制复杂，需要深入了解安卓Binder通信机制。
 - 业务场景复杂，需要支持多Client, 多Model调用，在待机唤醒、异步推理等多场景下需要确保资源管理和线程安全问题。

AI Infra

Needle Project Link - 2024.6-2025.1

项目背景：Needle 是一个简易版 PyTorch 的实现，其项目框架源自 [CMU 10-414/714 Deep Learning System](#) 课程。通过 Needle，用户能够在 Python 侧使用简洁的接口定义模型，轻松完成神经网络的训练和推理，降低了深度学习模型开发的复杂性。

- **自动微分 (Automatic Differentiation)**：通过抽象类 `NArray`、`Tensor` 和 `Device`，实现了自动梯度推导，支持两种模式：`eager mode` 和 `lazy model`，满足多种使用场景。
- **算子支持 (Operator Support)**：支持了常用的大多数算子，如 `Stack`、`Permute` 等。其中，`Conv (Im2Col)` 和 `Matmul (2D Tiling)` 算子的性能达到了 **NVIDIA cuBLAS 的 68%**，极大提升了模型的运行速度。
- **后端支持 (Backend)**：实现了对 `numpy`、`CUDA` 和 `CPU` 三种后端的支持，确保了模型能够在不同硬件环境下高效运行。
- **优化器 (Optimizers)**：实现了 `Adam` 和 `SGD` 两种常用的优化器，满足不同模型的需求。
- **神经网络模型 (Models)**：编写了 `ResNet9` 模型，并在 CIFAR 数据集上成功验证，展示了 Needle 可靠性。

AI 算法

端侧分类模型优化 - [华为-海思] - 开发 - 2024.10

项目背景：在 6XX 低端芯片有限的 GPU 能力下，单次推理速度要求达到 10ms，而训练数据集仅有数万张的情况下，需最大化提升模型推理的准确性。

项目结果：原 XX 平台算法部交付的模型在测试集上的准确率为 83%，通过该方案，成功将 **模型准确率提升至 89%**。

- **技术难点**：
 - **数据集受限**：由于使用场景限制，分类标准与现有公开数据集不兼容，自建数据集仅有数万张图片，训练模型容易出现过拟合。
 - **性能受限**：在低端 6XX 芯片 GPU 能力不足的情况下，单帧推理时间需要达到 10ms，计算量限制了模型的大小。
- **技术方案**：
 - 使用 ResNet50 进行迁移学习，在自建数据集上达到 99% 的精度，建立教师模型 (teacher model)。
 - 采用模型蒸馏技术，使用软标签损失 (soft target loss) 通过教师模型蒸馏得到学生模型 (student model)。

Vision Transformer [Project Link](#) 实现 - 图像分类模型开发

- **项目概要**：实现一个基于 Vision Transformer 的图像分类模型，以提升图像识别的精度和效率。
- **技术概要**：
 - 从零开始实现 Vision Transformer，不依赖于 PyTorch 现有模型，深入阅读相关论文，手动实现每个细节，包括位置嵌入 (Position Embedding)、图像补丁嵌入 (Patch Embedding) 等。这一过程不仅加深了对 Transformer 基本原理的理解，也锻炼了实现能力。
- **实现版本**：

1. **纯手写版本**：由于需要从头训练模型，时间成本较高，在小型数据集上运行了几个 epoch，确认模型的收敛。
 2. **迁移学习版本**：基于 PyTorch 的预训练模型，重训练分类头（head），在小型数据集上达到了 99% 的分类准确性，显著提高了模型的效率和准确性。
- **结果**：成功复现 Vision Transformer，加深了对模型机制的理解，而迁移学习版本则有效提升了模型的实际应用性能。
-

个人账号

- blog 地址 [主要分享架构设计、测试驱动开发等内容](#)
- Wechat && 手机号 18092430560

其他信息

- 英语雅思6.5，能够无压力听懂英文课程、讲座等
- 性格开朗，易于相处