姓名：魏嘉辰
通讯地址：北京市昌平区城北街道府学路 18 号中国石油大学（北京）汇才公寓 4 号楼 4317

志愿活动：参加 2019 西安灞河国际半程马拉松、大明宫净扫志愿活动、省图书馆整理志愿活动、地铁站台安全员志愿活动、参加会议室清扫志愿活动

个人简介备注：
本科：陕西科技大学，GPA3.4，大二转专业进入计算机类网络工程专业，分别于大二大三两个学年获得专业综测第一，大四综测第二的成绩，获得过多次奖学金，获得保研资格，并荣获校级优秀毕业生称号。

本科研究方向：主要涉及人工智能中的机器学习算法与深度学习神经网络，包括 CNN、RNN、LSTM、注意力机制、医学交叉、心电图分类等。

本科主修课程及分数：
Python 数据分析与应用、JAVA 网络应用开发、C++课程设计、操作系统、数据结构、计算机网络原理、计算机组成与体系结构、网络协议分析与应用、网络安全基础

Python 数据分析-97
计算机组成原理与系统结构-93
微机原理与接口技术-93
网络协议分析与应用-93
数据库原理与应用-91
C/C++语言程序设计-89
大数据技术原理与应用-88
Java 网络应用开发-87

研究生：中国石油大学（北京），GPA3.4，本科保研进入该校攻读硕士研究生，分别于研一研二两个学年获得研究生学业奖学金一等奖和二等奖，目前研三在读。

研究生研究方向：主要涉及人工智能中的机器学习算法与深度学习神经网络，前期主要面向目标检测模型，现在主要面向基于 Transformer 的生成式大语言模型/多模态大模型的优化方法，包括模型剪枝、量化、分布式并行等加速技术研究。

本科主修课程及分数：
人工神经网络、最优化方法、应用数理统计、随机过程、矩阵理论、现代通信系统概论

科技论文写作-94
现代通信系统概论-92
计算理论基础-84
随机过程-83
矩阵理论-80
应用统计方法-80
最优化方法-78

2018.08 – 2019.08 心理委员
1、心理健康教育与宣传：心负责向同学或成员普及心理健康知识，提高大家对心理健康重要性的认识，以及如何维护和促进个人心理健康。
2、情绪支持与倾听：为同学或成员提供情绪支持，倾听他们的烦恼和问题，给予同情和理解，帮助他们缓解压力和焦虑。
3、危机干预与报告：在发现同学或成员有心理危机或异常行为时，及时进行干预，并在必要时向专业人士或上级报告，以确保及时获得专业帮助。

2018.08 – 2019.08 宣传委员
1、策划与执行宣传活动：负责策划和组织各种宣传活动，包括设计宣传材料、组织演讲等，以提高组织或活动的知名度和影响力。
2、媒体关系管理：与各种媒体建立和维护良好的关系，确保组织的信息能够通过媒体渠道有效传播，同时处理媒体采访和报道请求。
3、内容创作与发布：负责创作和编辑宣传内容，包括新闻稿、社交媒体帖子、博客文章等，确保信息的准确性和吸引力，并及时发布到各种宣传渠道。

担任本科宿舍宿舍长
1、制定值日计划：跟踪执行过程，曾多次获得文明寝室称号；
2、关系维护：跟进舍友沟通与反馈，有效协调舍友关系，舍友争吵发生率几近于 0；
3、上传下达：及时接收和传达宿舍活动及学校各项通知，尤其是在疫情蔓延期间，及时地上传下达保证了舍友们平安快速返家。

2022.10 – 2022.11 主楼会议室净扫行动
1、收拾整理打扫会议室的桌面卫生及地面卫生，还有桌椅板凳摆放；
2、整理收纳会议室屏幕阵列的各种线材，并用扎带捆绑好。

2021.3 – 2021.5　西安凤城五路地铁站站台志愿者
1、引导乘客：在站台对乘客进行引导，帮助乘客了解排队候车的流程和纪律，以及车站的基本情况和岗位分布；
2、帮扶特殊群体：在服务过程中，帮扶老弱病残孕等特殊乘客乘车，确保他们能够安全、便捷地使用地铁服务；
3、维护秩序：负责监督站台内的秩序，规范乘客行为，如制止拥挤排队、防止乘客乱扔垃圾等，以维护地铁站内良好的公共秩序。

2019.11　2019 西安灞河国际半程马拉松志愿者
1、负责将每位运动员的随身物品进行编号并妥善保管；
2、对每位运动员的个人物品与运动员编号一一对应，保证在比赛结束后，大量的取包服务时能快速检索并归还。

2019.3　大明宫净扫行动志愿者
1、监督游客的文明出行，对游客的不文明行为礼貌劝阻；
2、对园区内地面即草丛上的垃圾进行清理、打扫。

中文题目：面向卷积神经网络协同推理的交错式算子划分

论文题目：Cooperative Inference with Interleaved Operator Partitioning for CNNs
录用期刊/会议：International Conference on Intelligent Computing (ICIC) 2024 (CCF C)
原文链接：http://poster-openaccess.com/files/icic2024/2251.pdf
中文摘要：https://www.cup.edu.cn/cupai/kxyj/kydt/fc3484670fa249998c0c39f90babb0be.htm
目前，智能物联网（AIoT）已广泛应用于工业生产、自动驾驶、智能家电等多个领域。随着深度学习技术的兴起，智能模型在执行推理过程中对设备的计算和内存需求正在急剧增加。一方面，物联网设备的内存容量十分有限；另一方面，许多实际应用场景具有严格的实时响应需求。例如阀门泄漏的检测，需要毫秒级的响应时间，否则将会导致严重的安全隐患。协同推理是解决这一问题的重要方法。现有的协同推理方法通常将算子的输出通道或特征图的高和宽作为划分维度。由于算子的激活值分布在多个设备上，需要在传递给后继算子之前进行拼接操作，这将会引入额外的通信开销，增加推理延迟。针对这一问题，本文提出了一种新颖的 AIoT 协同推理方案——交错式算子划分（IOP）以减少智能模型的推理延迟。
本文的主要内容如下：
（1）提出了 IOP，一种适用于 CNN 的协同推理加速方法，通过减少推理过程中所需的通信次数来降低推理延迟。
（2）基于 IOP 方案，对模型最小化推理延迟问题进行了建模。
（3）提出了一种启发式划分算法，该算法在所有包含两个算子的分段中应用 IOP，以最小化协同推理延迟。
（4）使用多个 CNN 模型评估了 IOP 策略，表现出了优越的性能。

[1] Liu, Z., Xu, C.*, Liu, Z., Huang, L., **Wei, J.**, & Li, C. (2024). Cooperative Inference with Interleaved Operator Partitioning for CNNs.（ICIC2024 已录用）
[2] Lu, Z., Xu, C.*, **Wei, J.** (2024). Data Sinks Deployment for Backbone-Assisted Real-Time PD-NOMA Networks Based on Reinforcement Learning.（UIC2024 已录用）

基于强化学习的骨干辅助实时 PD-NOMA 网络的数据基站部署
录用期刊：IEEE International Conference on Ubiquitous Intelligence and Computing (UIC) 2024 (CCF C)
我们研究了在给定数据接收器数量的情况下，为数据接收器找到最佳位置的问题，以最小化 BA-PDNOMAWN 中的上行链路传输延迟。我们制定了问题，然后将问题形式化为马尔可夫决策过程（mdp）。此外，我们提出了一种基于经典多智能体深度确定性策略梯度的强化学习算法，该算法精心设计其奖励函数来加速收敛。
本文的主要内容如下：
（1）我们对在 BA-PDNOMAWN 中找到给定接收器的最佳位置的问题进行建模，以最小化上行传输访问延迟。
（2）我们将问题形式化为 mdp，形成了低复杂度算法的理论基础。
（3）我们提出了一种基于多代理深度确定性策略梯度（MADDPG）的骨干辅助多数据基站放置（BAMDSPs）算法，其奖励函数是在我们之前的工作基础上精心设计的。

2023-2024 学年 中国石油大学（北京）研究生学业奖学金（二等奖）
2023 年 4 月 第十四届蓝桥杯全国软件和信息技术专业人才大赛省赛个人赛（三等奖）
2022-2023 学年 中国石油大学（北京）研究生学业奖学金（一等奖）
2022 年 6 月 陕西科技大学校级优秀毕业生

2020-2021 学年 陕西科技大学校级优秀学业奖学金（一等奖）4.28
2019-2020 学年 陕西科技大学校级优秀学业奖学金（二等奖）4.28
2018-2019 学年 陕西科技大学校级优秀学业奖学金（三等奖）6.12
2019 年 6 月 陕西科技大学第十三届高等数学竞赛校级三等奖

1、2023-2024 学年-研究生学业奖学金-校级二等奖-中国石油大学（北京）
2、2023 年 4 月-第十四届蓝桥杯全国软件和信息技术专业人才大赛省赛个人赛-省级三等奖-工业和信息化部人才交流中心
3、2022-2023 学年-研究生学业奖学金-校级一等奖-中国石油大学（北京）
4、2022 年 6 月-陕西科技大学校级优秀毕业生-校级-陕西科技大学
5、2020-2021 学年-优秀学业奖学金-校级一等奖-陕西科技大学
6、2019-2020 学年-优秀学业奖学金-校级二等奖-陕西科技大学
7、2018-2019 学年-优秀学业奖学金-校级三等奖-陕西科技大学
8、2019 年 6 月-陕西科技大学第十三届高等数学竞赛-校级三等奖-陕西科技大学

项目经历：

智能计算系统优化加速技术研究 2023.11 - 2024.07
项目简述：针对各种生成式大模型、多模态大模型，在搭载了不同型号 GPU 的异构计算系统上进行部署推理测试，以从各个角度如参数量、计算量、显存占用量、SM 占用率等，分析现今流行大模型的痛点，并进行相应的优化加速技术研究。

项目职责：
1、绘制 LLaMA、Stable-Diffusion 等大模型的算子流图及张量 shape 变换，同时统计不同参数量下模型推理时的计算量、参数量和显存占用量；
2、使用 TensorRT/TensorRT-LLM 构建 Stable-Diffusion 及 LLaMA 大模型的推理引擎并在服务器上部署；
3、结合 NVIDIA 官方文档，统计收集整理 A40/A100/4090/V100 等型号的硬件参数，如 SM 数量、SP 数量、CUDA/Tensor Core 数量、峰值算力、峰值带宽、每 SM 最大并行 block 数、每 SM 最大共享内存等指标，为后续分析模型推理时占用资源打下基础；
4、使用 DCGM 工具观测、记录大模型推理时在不同型号 GPU 上的 SM 占用率及带宽等指标，并使用 Nsight Compute 针对 Self-Attention 和 FFN 两部分的 kernel 进行分析，深挖其 SM 利用率、吞吐量等指标与模型的关联；
5、针对 30B/70B 参数规模的大模型，使用仅权重 INT8 量化，同时在推理时结合 DejaVu 动态剪枝、L2 Norm 剪枝减少了参数量和计算量，解决了其不能在单卡 A40 上部署推理的问题，同时保持精度，并降低了推理延迟；
6、完成相关表格及绘制分析用柱状图线形图等。

目标识别网络及前后处理在 MLU 上的一体化移植、适配和优化加速算法（2022）横向 2023.6 - 2023.10
项目链接：https://github.com/Jayce0625/Cambricon.git

项目简述：使用离线生成寒武纪目标检测模型，在不使用深度学习框架的前提下，使用 C++ 离线部署到搭载寒武纪 MLU 和英特尔 CPU 的异构计算平台上，其目的是解决 PyTorch 的

复杂张量操作和动态逻辑所带来的移植困难，极大地简化了工业应用的部署。

项目职责：
1、使用 cnrtTransDataOrder 对输入张量和输出张量进行 NCHW 和 NHWC 两种数据排布的互转，解决了异构平台上寒武纪 MLU 端与主机 CPU 端的数据排布不一致的问题；
2、使用 cnml 和 libtorch 两种方式实现了能正确处理数据排布并将结果回存 CPU 的 torch.gather 算子，以实现寒武纪算子库中不支持的算子；
3、使用自定义的 gather 算子，解决了由于不支持 gather 算子进而被迫分成三段的离线模型的顺序部署，并成功执行推理；
4、统计整个离线模型的推理时间及各部分关键模块、算子的执行时间，并筛选出执行时间最长的几个算子。

拥挤场景下的微小目标检测系统　　2023.3 - 2023.6
项目背景：在现如今的遥感图像领域，其图像往往分辨率非常大，这就导致需要进行检测的目标尺寸很小，尤其是对于“人”这一类别，在整幅图像中只占了几个像素点，进一步导致了现有的目标检测网络如——yolov5、efficientdet、faster-rcnn 等检测精度非常低，所以能够适配这种微小且拥挤的目标同时达到高精度水准的算法/网络就十分重要。

项目链接：https://github.com/Jayce0625/glsan_wjc.git

项目简述：实现一个用于无人机视觉微小目标检测的网络，解决无人机场景下，待检测目标过于拥挤与微小造成的精度下降的问题，通过自适应分片切分、超分放大、数据增强、子图原图结果融合等技术和手段，最终实现一个高精度的微小拥挤目标检测模型。

项目整体分为以下几部分：自适应分片算法编写、超分放大算法编写、数据集收集制作增强、骨干神经网络选取组合与搭建、训练与微调及测试这几部分组成。

项目成果：
在使用以 ResNet-50 作为骨干网络的 Faster-RCNN 上，经由我们改进的方法，使得在 VisDrone2019 数据集上，对十个类别的检测精度平均比 vanilla 版本的 Faster-RCNN-ResNet-50 高出约 10-20 个百分点。

项目职责：
1、使用 K-means 聚类算法对输入图像进行自适应密集区域裁剪，从而避免对无目标的背景进行检测；
2、通过 EDSR 超分网络对小尺寸的图像切片执行超分辨率从而将微小目标清晰放大，并构建增强数据集；
3、使用以 ResNet-50 为骨干的 Faster-RCNN 在增强数据集上进行训练、微调；
4、使用训练好的模型执行推理时同时结合自适应裁剪与超分放大，将目标检测的各类别精度提升约 10%；
5、编写接口文档及其他相关文档说明。

中法录井智能传感器系统　2022.8 - 2023.3　中法渤海地质服务有限公司
项目全称：基于通用嵌入式处理器平台的录井传感器智能模块

项目编号: CFB_TG_TI_2021_001

项目链接：https://github.com/Jayce0625/zhongfaV2.git

项目背景及简述：实现一个海上录井智能传感器系统的转发器部分。转发器接收传感器数据并进行处理，使用 ATmega16/128 单片机控制 USART 连接的 WIFI/4G 模块，转发至 Socks5 代理服务器，经用户认证后，最终转发至目标数据服务器。解决了海上复杂环境长距离通信传输的可靠性，以及使用 Socks5 代理认证增加安全性。

项目整体分为以下几部分：底板硬件设计、采集器软件设计、转发器软件设计和 Socks5 代理认证、数据库设计这几部分，本人负责转发器软件设计和 Socks5 代理认证这一部分。

项目成果：
最终实现了一个 STM32 采集器底板+ATmega16/128 转发器底板的录井智能传感器系统，解决了海上录井复杂环境长距离通信传输的可靠性以及使用 Socks5 代理认证增加安全性的问题。项目已经成功落地投入使用一年有余，期间未发生过问题。

项目职责：
1、编写数据处理算法，实现对 RS-485 串口转发的、由上游传感器采集的数据进行大小端转换；
2、编写中断函数及 C++代码，通过 I2C 与 SPI 协议实现 ATmega16 与 ATmega128 单片机之间的串行通信，同时利用 ATmega128 单片机通过 USART 串口对连接的 WIFI 和 4G 模块进行控制，以实现数据通信、转发；
3、部署 Socks5 代理服务器，并增加用户代理认证功能，使得重组转发后的数据先经由 Socks5 代理服务器认证后，再转发到目标数据服务器；
4、编写接口文档及其他相关文档说明。

基于 Spark 的农产品分析系统　2022.3 - 2022.5
项目背景：我国农业现代化建设稳步发展，农产品消费结构转型加快，预计农产品市场价格将保持小幅上涨态势。面对错综复杂的国际形势和国内经济下行压力，对农业产品市场的统计分析就显得尤为重要。

项目链接：https://github.com/Jayce0625/agricultural_analysis.git

项目简述：设计一个农产品分析系统，针对存放于 HDFS 分布式文件系统上的全国各省农产品市场数据进行统计分析并存入 MySQL 数据库，最终将数据分析结果进行可视化。（基于 Spark 大数据内存计算框架，使用 Scala 函数式编程语言开发实现）

项目整体分为以下几个部分：原始数据清洗分类、农产品市场分析模块、农产品分析模块、可视化展示模块这几部分组成。

项目成果：
最终成品对农产品及农产品市场行情、走向、地域等要素进行了详细的数据分析，并经由

Davinci 可视化工具绘制了一个精美的数据大屏，并以数据大屏的方式进行了展出、汇报。

项目职责：
1、负责对存放于 HDFS 分布式文件系统上的数据利用 Spark Core（Scala）进行数据清洗及离线批处理操作；
2、负责将处理好的数据通过 Scala 的 JDBC 库导入 MySQL 数据库；
3、使用 Davinci 可视化工具将分析结果进行可视化展示；
4、编写接口文档及其他相关文档说明。

个人优势：在国家级实验室之江实验室实习一年时间；有 AIGC 模型部署测试微调的经验；有生成式大语言模型/多模态大模型部署优化压缩的经验；熟悉英伟达 GPU 硬件模型和软件模型；有国产寒武纪 MLU 异构平台的模型移植经验；有 Linux/嵌入式等平台的 C/C++开发经验。

自我评价：我具有出色的学习能力，无论是本专业的新颖前沿知识，亦或是跨专业相关知识，我都能将知识迅速吸收转化为自己的实际能力。我同样也是一个工作态度严谨的人，对于每一项任务，我都能够认真对待，确保高质量完成。工作学习中最重要的就是团队协作能力，我始终能在团队中保持积极地沟通，努力为团队贡献自己的力量。我同样也有一些缺点，我的处事不够果断，做决定时总是瞻前顾后，为了提高效率，现在的我在首次遇到的某种情况时才会仔细分析研判，而对于熟悉的任务，则坚定信心，果断行动。

未来五年职业规划：我将首先通过校园招聘加入公司，利用入职初期的机会，通过参与培训和实践，打好专业基础，全面了解公司的业务流程和文化。在职业生涯的前两年，我将专注于提升个人技能，包括技术能力和沟通协调能力，同时积极完成岗位职责，争取在团队中发挥积极作用。随着经验的积累，我希望能够逐步承担更多的责任，为团队贡献更多的价值。在第三年及以后，我将寻求合适的机会，通过展示我的工作成果和潜力，争取获得更高层次的工作挑战。我的目标是逐步提升自己的职业地位，同时保持学习和进步的态度，为公司的长远发展做出贡献。在整个职业发展过程中，我都将保持谦虚学习的心态，不断适应变化，努力成为公司需要的复合型人才。

实习经历：

之江实验室
天基计算系统研究中心　AIGC 模型部署微调实习生　2023.06 – 2024.07
1、使用寒武纪 MLU 加速卡，利用 CNRT 及 CNML 对 AI 模型进行离线部署，实现无深度学习框架下执行推理；
2、绘制 LLaMA、Stable-Diffusion 等大模型的算子流图及张量的尺寸变换，同时统计不同参数量下模型推理时的计算量、参数量和显存占用量；
3、使用 TensorRT/TensorRT-LLM 构建 Stable-Diffusion 及 LLaMA 大模型的推理引擎并在服务器上部署；
4、结合 NVIDIA 官方文档，统计收集整理不同型号 GPU 的硬件参数，如 SM 数量、CUDA Core 数量、峰值算力/带宽、每 SM 最大并行 block 数等指标，为后续分析模型推理时占用资源打下基础；
5、使用 DCGM 工具观测、记录大模型推理时在不同型号 GPU 上的 SM 占用率及带宽等指

标，并使用 Nsight Compute 针对 Self-Attention 和 FFN 两部分的 kernel 进行分析，深挖其 SM 利用率、吞吐量等指标与模型的关联；
6、针对 30B/70B 参数规模的大模型，使用仅权重 INT8 量化，同时在推理时结合 DejaVu 动态剪枝、L2 Norm 剪枝减少了参数量和计算量，解决了其不能在单卡 A40 上部署推理的问题，同时保持精度，并降低了推理延迟；
7、使用 Visio 完成各类项目的图表绘制并撰写相关文档材料。

西安鲲豚网络科技有限公司
网络安全工程师培训中心　网络安全工程师实习生 2021.03 – 2021.06
1、与网络安全团队合作，使用 Kali Linux 对组织的 IT 基础设施进行全面渗透测试和漏洞评估；
2、在 Linux 环境中设计并实施 LAMP 和 LNMP 软件栈；
3、使用 iptables 防火墙工具实现并维护网络访问控制；
4、利用 Wireshark 的过滤器、解包器并结合 Burp Suite 工具，捕获并分析网络流量，从而识别并修复安全漏洞；
5、绘制并撰写相关图表与文档，制作汇报用 PPT。