2d80f8d57c35d10e1Xd52t-6FlVYwou_WfqZWOGrlvbTNhFm2A~~

张志远

男 | 38岁 13917215657 zhangzhiyuan303@163.com

17年工作经验 | 深度学习 | 期望薪资：20-40K | 期望城市：上海

个人优势

国产NPU， 大模型适配，FA等融合算子开发，熟练使用msdebug msprof op等工具，进行算子优化.成功把视频生成模型推理场景下的PFA算子性能提升15%。熟悉SD,图片理解等多模态算法原理；有910系列和A800多卡训练、推理部署，优化相关工程经验。

1.精通 Diffuser、Transformer、OpenCV 等库；熟练进行实际项目的 fine-tuning/lora训练。

2掌握Attention、VIT、CLIP 、DDPM等多模态和生成模型原理；有图生图，图片理解落地项目。

3.熟练使用docker,shell,Linux系统。熟练运用Profiling工具，算子调优，模型蒸馏等， 提升显卡性能和存储上限。

4.扎实的python,c++编程功底；线代、概率、统计、数值分析等数学基础；有面相对象思想，熟悉常用设计模式，熟练运用多线程，独力完成Ascend C融合算子开发和优化，良好的开发感觉。

5.了解开源VLLM推理框架，Qwen系列VL、Llama等多模态和大语言模型。

6主导过3款软件产品从0到1的设计，研发(产生两个爆款) .严谨的逻辑思维和良好的抽象思维;善于提炼用户需求痛点;挖掘产品价值与定义软件.熟练使用XMind、Axure、Visio等工具.

工作经历

软通动力技术服务有限公司 ai工程师 2024.08-至今

模型适配，优化。gpu到npu迁移。要解决各种问题: 数据类型，算法，精度，性能，显存上限、算子优化等，让npu跑大模型更流畅，高效！使用profiling，batchmark，Ascend C，量化等工具。

大语言、SD类、多模态模型的lora训练和Fine-tuning以及蒸馏.日常使用Linux操作系统,xshell,docker 等；服务器多为8卡npu，也有A800等。

具体：vllm+qwen2 vl视频理解、CogVideo、LTXVideo等npu适配优化实际项目。

太仓布百试软件 深度学习 2023.04-2024.07

追踪最新 AI 算法，满足客户 AI 方面需求落地。主要集中在工业缺陷检测，图片识别、分类（Fine-tuning），图像生成模型训练以及基于 comfyui 搭建个性化客户应用（lora 风格模型训练，contrlnet 特定修复模型训练），详见项目经历。

港中旅华茂物流 产品经理 2022.11-2023.03

在空运项目组，主要负责国际物流空运相关 分子公司需求调研，采集，分析;参加供应商演示会，了解各家供应商产品优势，为下一代软件开阔思路。主要从以下几个方面进行比较:

1.技术框架

2.系统柔性(支持分子公司业务特点灵活配置)

3.报表框架(BI)(是否支持终端用户自定义维度,自定义数据集相关解决方案)

4.业务沉淀(供应商当前系统对行业积累程度)

5.现有通用产品组件(托书识别，规则引擎，配置引擎,智能调度等)

6.供应商团队规模，以及大项目管理经验和能力

7.乙方成长和获利等维度进行量化分析,打分评比

参与编写完成软件功能清单，需求场景说明书;关键模块风险评估，技术验证。需求文档技术方案

上海布百试创意 产品 2016.01-2022.11

阶段一 2016/1-2018/12 软件产品经理

面料试衣渲染引擎服务器插件研发，运营。渲染引擎主要功能是把客户提供的面料图片，渲染为成衣图片返回，更加直观的方便用户挑选面料。

插件通过 json 等接口形式，把渲染号的图片返回给 app ，网页等应用场景应用(2B端市场)。

这3年是职业生涯相对比较顺利的时刻,赶上了互联网+行业风口，插件能帮创业公司把故事讲完整。产品得到了创业型网络公司，以及纺织行业内一些大的公司，大的网络平台的认购。包括全球纺织网，广州志达家居，以及很多互联网+行业创业公司。

产品定位:在别人创业路上卖水。

具体工作:

1.客户需求沟通，调研，出解决方案。

2.渲染引擎核心模块的研发。

3.建模软件需求分析，开发任务分配。

4.用户接口，以及 web 前端开发任务分配。

阶段二： 2019年1月-2022年11月 --软件产品经理 &项目经理

面料试衣 APP 的开发,运营，这款软件旨在通过后台数据库搭建的模特图形实现不同面料的成衣效果，以此来服务于面料工厂和商户，期待面料商铺可以直接展示面料的成衣效果来提升销售量。整个开发过程中反复研究了各种解决方案来降低模型制作成本，图像渲染时间，仿真效果，服务器大访问量的并发处理，用户界面操作友好性等系列问题。后期有1年左右时间研究深度学习算法，并尝试解决用户AI需求，并完成简单项目交付。

推行之初，得到了很多用户，互联网平台的认购，近2年随着市场供求关系的变化，仅仅通过面料成衣仿真已不能再促进成单率，市场需求的是更加完整的全套的解决方案，以及更加专业的业内经验，认清前景之后，准备放弃。

产品定位:为小B端用户(偏向个人)提供服务，目标是做成一个卖款式的 mini 商城。

具体工作内容:

1.生产环节:拆分建模流程，组建模型批量制作流水线(包括建模软件研发，以及后期建模工作岗位流水线组建)。

2.销售和售后：组建小的外呼和客服中心。购买探迹会员，通过招募有经验的 On Call ，组织 On call 话术(价值型话术)，以及培养其他同事进行电话陌拜推荐产品。各种激励提成政策的制定等。

3.终端展示产品：需求分析，开发任务分配。包括 IOS+Android 版本 原生 APP ； 面料博物馆(Unity 开发的 IPad 端版本); 微信小程序端的在线面料试衣间；Web 版本的在线面料试衣间以及客户化定制的一些 web 程序的任务分解，以及分配开发。

4.日常客户反馈问题的收集，分析，产品更新迭代。以及应付各种突发事件。保证项目正常运转。

科源软件 C/C++ 2013.01-2015.12

科源色织王-面料设计软件2.0及后续版本的的研发(PC软件，代码行数10万+)

该软件具备2个主要功能，面料设计和投产工艺。旨在降低面料工艺员的重复手工劳动，通过面料设计后的仿真实现打小织样的功能，并直接生产工艺投产单。该软件在南通，绍兴纺织行业应用广泛，但受制于纺织行业前景及重复需求性低的限制，持续扩展性不高。

上海卫宁软件 C# 2009.01-2012.12

这是职业生涯里，第一次做的超过3年的一份工作，从最初的具体业务需求开发，慢慢负责某一单独产品。体会到了，要专注做谋一行，就会有超预期的收获(比如：参见公司组织的开发主管级别以上的培训，并考取信息系统项目管理师高级证书)。

1 C#版本结构化电子病历（EMR ）研发，熟悉结构化文档编辑器，数据存储模型，文档打印机制及医院病历书写业务流程及国家卫生部相关要求(2010年10月至今)。EMR 可以理解为 rtf 版结构化文档编辑器

2电子病历工具及其 HIS 相关模块需求开发. 电子病历工具:支持实施现场画病历的一套快速开发工具.特定控件开发及修改.

这段工作对于 windows 消息处理机制，医院业务流程，控件开发，及电子病历（结构化及非结构化电子病历）有了较深入的理解。

项目经历

大模型适配，优化加速　开发　2024.08-至今

1.使能模型库快速构建应用

SD,LLM,MOE,多模态模型推理，训练环境搭建；profiling调优,量化等适配工具熟练在工作中使用。

2.独立完成Npu侧模型适配，丰富模型库

vllm+qwen2 vl视频理解、CogVideo视频生成、LTXVideo等

3.相关算子优化

3.1搬运类算子 repeat_interleave 在大shape,-1纬度下性能瓶颈优化Python：

3.2cann-ops高性能算子下发c++

3.3fa等融合算子泛化性能优化c++：基于昇腾硬件属性（L1 L0A/B/C UB空间大小，CV分离，L2Cache等），结合模型实际Shape,合理设置tiling；通过合理分核，设置double buffer等方法提高数据并行度；设置各种数据对齐，提升数据搬运效率，避免mte bound；GM数据合理切分，提升L2 Cache命中率，提升GM和UB/L1数据交互速度。

AI实际落地应用　开发　2023.03-2024.08

1工业检测项目：

需求：汽车零部件行业，特殊产品，要监测操作工关键操作步骤是否合格。

实现方法：我们选用 yolo 模型，然后设计好特征；搜集合格+不合格数据图片（训练数据和验证数据）标注数据，训练，验证，web 服务部署。

实际运行：通过摄像头获取当前图片，通过训练好的模型进行判断是否合格，不合格的时候，会暂停（通过 plc 通信，给操作工提醒，整改合格后才能继续）.

二：图片生成(图生图 aigc)

1需求：重现类似 Deep face 功能（图像可控确定性方面）

2实现：通过 huggingface 了解扩散模型原理和实际 demo 项目，找到灵感:特定数据集的局部重绘（inpaint)

3通过训练特定 controlnet .实际运行在 comfyui 框架下：可以实现类似人体局部换皮肤/手指等的功能。扩展下去，可以实现换脸等功能。

AI 算法工程师描述：

1.持续关注最新算法

2.了解从事行业的特定需求

3.匹配算法解决实际需求

除此之外，我的个人优势还在于：卓越的实际问题抽象化能力，编程技术扎实，软件产品落地经验丰富。

深度学习　开发　2022.12-2023.03

内容:

1.跟着教程. 练习 coding 各种深度算法，了解原理；重现并局部改动经典算法（引进，消化，吸收，再创新模式）

2.日常工作基于 huggingface get 项目,微调流程。跟踪业界主流，然后应用于实际项目。

业绩:

1.在计算机视觉 CV 方向，有了实操的经验图像分类，检测，分割等；比如使用卷积网络 LeNet、 AlexNet、GoogLeNet 对手写字数据库，进行训练，输出验证模型。

2.AIGC 方向:基于 SD的 OOTDiffusion 换装项目及其背后相关的算法原：LDM,diffusions ,其他基于 sd 的微调项目.阅读论文查找 sd 等文生图片图生图的核心原理算法--ddpm。

3.可以熟练应用 python，pytorch ，配置运行环境，超参数设置, 微调训练，并落地实际项目。

机器学习 开发 2021.03-2022.12

内容:

描述：一.机器学习使用场景高度不确定性:知识(经验)已经不能完全给出答案的时候.回归：找线把点串起来分类：找线把点分开(离着线最近)二.机器学习基本流程使用学习算法--通过训练数据（含标记类别=答案）--得到模型 ； 对新样本类别标记未知.

1.准备大量真实数据集(涉及数据降维，清洗等)

2.选择算法

3.参数设置

4.验证性能-调整参数

5.模型上线服务:真实数据输入-得出结果；

6.通过后期人工打赏和纠错，实现模型内部系数调整，从而实现算法越用越准确.三.开发工具和常用算法熟悉 Python 环境下的数据挖掘工作；了解各类算法数学原理(感知机，决策树，神经网络，支持向量机-SVM ，集成学习，回归，聚类等) ；熟悉监督学习，无监督学习，加强学习；四.机器学习知识体系渐进：

S

1.主成分分析(PCA)学习方法实例(空间三维数据降维):

1.对数据进行坐标平移（数据中心放到原点）

2.通过2次左乘初等矩阵 RSD ，实现坐标轴的旋转及拉伸X轴

3.通过协方差矩阵找到新坐标轴的特征向量(表示方向)和特征值(放缩大小) 。

4. 可实现三维降二维：找到二维特定平面，投影上去 ； 二维降直线，也是同理.小结: PCA 降维，有点类似图片处理：就是通过左乘矩阵初等变换(放缩，旋转等)对点阵/数据实现特征突出变换(数据预处理)，然后结合其他统计方法(协方差等)找到边界，实现对新数据的预测.

S

2.支持向量机 SVM 模型(环形数据分类):个人感觉比较好的二分类算法.支持线性分类，以及非线性分类(比如环形数据)。在非线性分类方面: 可以通过数据升维到特征空间后(二维变三维，同样也是类似图像处理坐标变换),再求分类边界线;奇妙的地方在于：不用求解升维后的复杂方程表达式，而是通过同维度的核函数实现.

S

3.朴素贝叶斯方法(垃圾邮件分类)：通过数据集学习条件概率分布(假设条件独立)；单个分类的先验概率分布P(w）(注意学习数据集和真实数据集的比例要尽量接近)。通过学习到的模型计算后验概率，并将最大的类作为目标分类。 在垃圾分类邮件实例中：输入文件d，某一分类下(1垃圾邮件 2 非垃圾邮件) 多个关键字出现次数的概率(出现次数/总的关键字出现次数)乘积作为条件概率最终取值，哪个数大，就算哪一类。此方法为概率模型下的生成模型的机器学习算法。

S

4.EM 算法(这是一个解决问题的通用策略，非独力算法，是一个局部极值算法)用于含有隐形变量的概率模型参数的极大值估计，或极大后验概率估计。Em 算法步骤 1求期待E (数学期待) 2 求极大值M,3 更新系数 4 (迭代)重复23,4直到收敛 EM 是数学期待和极大值的英文首字母缩小。可以用在以下算法：高斯混合模型，K均值聚类 。

S

5.决策树(多类型分类，类似自动找出多个 if ...then...)

1.要选择合适的特征(字段列)

2.可以通过信息熵，或者基尼指数 作为算法策略，选定根节点字段，叶节点字段。3通过剪枝限定树的层级(够指定百分比的才能够单独成子节点等条件)，提高效率，防止过拟合等问题。

S

6.感知机模型 (解决是非问题)是二分类线性模型 f(x)=sign(wx+b)，输入为实例的特征向量，输出为实例的类别(-1,1).通过误分类的损失函数，利用梯度下降法对损失函数最小化，求的感知机模型(正负分离超平面--二维坐标下的直线或者三维坐标下的的平面)。特点事简单而且易于实现，是很多其他模型的基础，所以被戏称为入门垫脚石(实际中用的很少)。S

7.K近邻模型给定一个训练数据集，其中类别已定。分类时，对新的实例，根据k个最近邻的训练实例的类别，通过多数表决(投票)等方式进行预测。里边会用到2点间距离，一般使用欧式距离(平面2点间的连线距离)。K紧邻模型对是很简单的模型。

业绩:

熟练掌握机器学习算法 数学基本原理；掌握 PyTorch 深度学习框架建模;掌握 Python 语言；能读懂英文算法论文，并用代码复现。

实战:

教育经历

河北科技大学 本科 服装设计与工程 2005-2009

资格证书

国家高级项目管理师资格证书