高级 AI CCF 选题

1. 基于文心 CV 大模型的智慧城市视觉多任务识别-图像分类

/3🌟/5w

背景：

近年来预训练大模型一次次刷新记录，展现出惊人的效果。但受算力和存储的限制，大模型无法直接部署在边缘设备上。针对大模型的开发和部署问题，VIMER-UFO 给出了 One for All 的解决方案，将不同参数量、不同任务功能和不同精度的模型训练过程变为训练一个超网络模型。

。VIMER-UFO All in One 研发模式可被广泛应用于各类多任务 AI 系统，以智慧城市场景为例，VIMER-UFO 可以用单模型实现人脸识别、人体和车辆 ReID 等多个任务的 SOTA 效果，同时多任务模型可获得显著优于单任务模型的效果，证明了多任务之间信息借鉴机制的有效性。

针对大模型的开发和部署问题，UFO 给出了 One for All 的解决方案，通过引入超网络的概念，超网络由众多稀疏的子网络构成，每个子网络是超网络中的一条路径，将不同参数量、不同任务功能和不同精度的模型训练过程变为训练一个超网络模型。训练完成的 One for All UFO 超网络大模型即可针对不同的任务和设备低成本生成相应的可即插即用的小模型，实现 One for All Tasks 和 One for All Chips 的能力

任务

选手需使用飞桨（PaddlePaddle）深度学习框架，基于人脸、人体、车辆、商品四大场景任务，训练视觉大一统模型。

参考网页：

https://blog.csdn.net/m0_63642362/article/details/126679950

数据集介绍

我们使用了脸、人体、车辆、商品的公开数据集具体如下:

训练集

任务	数据集	图片数	类别数
人脸	MS1M-V3	5,179,510	93,431
人体	Market1501-Train	12,936	751
人体	MSMT17-Train	30,248	1,041
车辆	Veri-776-Train	37,778	576
车辆	VehicleID-Train	113,346	13,164
车辆	VeriWild-Train	277,797	30,671
商品	SOP-Train	59,551	11,318

测试集

任务	数据集	图片数	类别数
人脸	LFW	12,000	-
人脸	CPLFW	12,000	-
人脸	CFP-FF	14,000	-
人脸	CFP-FP	14,000	-
人脸	CALFW	12,000	-
人脸	AGEDB-30	12,000	-
人体	Market1501-Test	19,281	750
人体	MSMT17-Test	93,820	3,060
车辆	Veri-776-Test	13,257	200
车辆	VehicleID-Test	19,777	2,400
车辆	VeriWild-Test	138,517	10,000
商品	SOP-Test	60,502	11,316

2. 系统访问风险识别 - 分类问题/2🌟/5w

赛题背景

在当前 IAM 的探索进程当中，最易落地的方法是基于规则的行为分析技术。它的可理解性很高，且很容易与身份鉴别技术进行联动，但它是基于经验的，缺少从数据层面来证明是否有人正在尝试窃取/验证非法获取的身份信息，或正在使用窃取的身份信息。

赛题任务

基于用户历史的系统访问日志及是否存在风险标记等数据，结合行业知识，构建必要的特征工程，建立机器学习、人工智能或数据挖掘模型，并用该模型预测将来的系统访问是否存在风险。

参考资料

https://www.heywhale.com/mw/project/6319f91e62548faa8911761e （看起来是比较完整的代码。感觉偏向于数据挖掘，主要做特征工程？）

3. 高端装备制造知识图谱自动化构建技术评测任务-知识图谱/关系抽取/★★★/5.5 万人民币

赛题背景

各种高端装备领域的故障案例文本是由业务专家或者专业维修人员撰写的描述相关设备异常以及故障排查步骤的记录，它的利用受到数据结构化程度的影响，因而识别数据中的部件单元、性能表征、故障状态、故障检测工具等核心实体及其之间的组成关系至关重要。

赛题任务

通过从大量故障案例文本抽取出部件单元、性能表征、故障状态、检测工具等实体及其关系，为后续高端装备制造业故障知识图谱构建和故障智能检修和实时诊断打下坚实基础。本任务需从故障案例文本自动抽取 4 种类型的关系和 4 种类型的实体。

参考资料

https://aistudio.baidu.com/aistudio/projectdetail/4485704 （某大佬在飞桨平台写的，感觉是比较完整的代码）;

4. 返乡发展人群预测-分类问题/★★★/5w

赛题背景

近年来，随着新一线城市的快速发展，带领着我国经济稳步发展，新一线城市对人才的吸引力也逐年递增，紧追一线城市。越来越多的年轻人不再局限于在一线城市谋求就业机会，而是选择回到家乡就业。

赛题任务

基于中国联通的大数据能力，通过使用对联通的信令数据、通话数据、互联网行为等数据进行建模，对个人是否会返乡工作进行判断。

参考资料

https://blog.csdn.net/JinbaoSite/article/details/126876073 提供了思路和代码；

https://aistudio.csdn.net/631441cb6a097251580d51d9.html 飞桨平台上某用户；

5. 小样本数据分类任务

赛题背景

常见的分类体系有国际专利分类(IPC)、联合专利分类(CPC)、欧洲专利分类(ECLA)等，但这些分类体系比较复杂，专业性强，对非 IP 人员而言使用有一定的困难。

赛题任务

为了解决以上困难，智慧芽构建了一种新的分类体系，并提供 958 条训练数据，及对应 36 个类别的分类标签，要求选手设计一套算法，完成测试专利数据的分类任务。

参考资料

https://zhuanlan.zhihu.com/p/563822559

https://blog.csdn.net/CallMeYunzi/article/details/127016647 有完整代码和得分，看了一下好像主要是处理文本内容。

6. 基于文心 NLP 大模型的阅读理解可解释评测-知识挖掘/阅读理解/★★★/5w

赛题背景

神经网络（NN）模型已经成功地应用于很多 NLP 任务并取得了不错成绩，但 NN 模型的黑盒性质降低了使用者对其结果的信任度，因此 NN 模型的可解释性、鲁棒性等问题受到广泛关注。为进一步推动模型可解释性研究的发展。

赛题任务

选手需使用飞桨（PaddlePaddle）深度学习框架，根据给定的一段文本 T 及与其相关的问题

Q，从文本 T 中抽取问题 Q 对应的的答案，同时给出模型预测答案所依赖的证据。

参考资料

https://beta-www.datafountain.cn/competitions/589 更详细的介绍；

https://www.pudn.com/news/632ca4ee272bb74d44e72287.html 记录（1）-飞桨的安装，题目理解；

https://blog.csdn.net/weixin_61083660/article/details/127003095 记录（2）一些代码

7. 基于昇思 MindSpore AI 框架的肾脏肿瘤分割-/文本分类 /★★★★/5w

赛题背景

近年来深度学习方法在图像分割领域引起广泛关注，特别是 UNet 等网络在肿瘤分割中取得显著的效果，但其在结构设计、训练模式等方面仍有进一步提升空间。

赛题任务

参赛团队利用赛方所提供数据，分析数据特点，在华为昇思 MindSpore AI 框架上设计、开发可用于肾脏和肾脏肿瘤语义分割的算法模型，在测试集中得到最好的肾脏和肾脏肿瘤语义分割结果的模型为获胜者。

参考资料

暂无

8. 基于 TPU 平台实现人群密度估计 - 算能/人群计数 /★★★★/5w

赛题背景

人群密度估计是计算机视觉中的一项重要任务，旨在同时识别各种情况下的任意大小的目标，包括稀疏和杂乱的场景。它主要应用于现实生活中的自动化公共监控，能够在公共安全管理、

公共空间设计、数据收集分析等方面发挥重要的作用。

赛题任务

参赛者选用预训练的模型部署在算能 TPU 芯片上，**无需自己训练模型**；在实现模型部署的过程中，参赛者需兼顾精度与推理速度。

参考资料

https://zhuanlan.zhihu.com/p/562910290 更详细的介绍

https://beta-www.datafountain.cn/competitions/583 官网的详细介绍

https://zhuanlan.zhihu.com/p/500703139 《第一届 TPU 编程竞赛》代码思路分享-Conv2D 和 Depthwise2D

https://blog.csdn.net/u011285477/article/details/106264498

https://blog.csdn.net/weixin_44523062/article/details/109628792 人群密度估计的一些笔记

其他相关代码和参考思路暂无

9. 基于 openLooKeng SQL 语句内存使用量预测系统-性能优化/★★★★★/5w

赛题背景

全内存查询引擎由于其高性能被广泛应用于各大厂商，但是它普遍面临的问题是，当 SQL 语句使用的内存超出了系统的内存大小时，就会导致 SQL 语句执行失败，甚至可能会导致系统崩溃。因此，在 SQL 运行前准确预估 SQL 内存的使用量对系统的稳定性尤为重要。

赛题任务

本赛题只涉及 SQL 语句内存使用量预测系统，重点在于根据 SQL 涉及的算子、表格数据量和参与计算的列等信息，建立一个内存使用量预测系统。

参考资料

https://my.oschina.net/frankwu/blog/4900574 只有一些 openLooKeng 的介绍资料

相关代码和参考思路暂无

10. 数据湖流批一体性能优化-性能优化/★★★★★/5w

赛题背景

湖仓存储框架的流批一体读写性能，关系到数据能否快速、准确的摄入到湖仓中，并做高效的

数据处理分析。而数据湖通常使用计算存储分离的设计，并且需要支持多种计算框架、支持对象存储等，给读写性能优化带来很大的挑战。

赛题任务

参赛选手可以选择一个数据湖存储框架，在保证数据正确性的前提下，通过调优参数、优化代码的方式来优化性能，提升写入速度。

参考资料

https://zhuanlan.zhihu.com/p/141182199 阿里的一个数据湖流批一体的综述

11. 大规模金融图数据中异常风险行为模式挖掘-图计算 /★★★★/5w

赛题背景

业界常用的频繁子图挖掘算法可以帮助发现高频出现的子图结构，如何使用频繁子图挖掘算法高效地进行异常风险行为模式挖掘显得尤为重要。

赛题任务

赛题使用简化的金融仿真数据，数据为带有时间戳和金额的账户间交易、转账等数据。基于此数据自动挖掘出不小于频繁度（f>=10000）的频繁子图模式集合。

参考资料

无

其他

其他的专题赛感觉这俩基本也是做 NLP 的，难度应该挺高的。其他两个，一个是 Linux 系统的，一个是 Web 攻击漏洞的，感觉没啥兴趣....orz

所属赛道 数字安全专题赛

出题单位 中国科学院信息工程研究所
360未来安全研究院工业互联网实验室

技术方向 文本分类 关系抽取

赛题奖金 8万人民币

难易程度 ★★★★

13

基于人工智能的漏洞数据分类

赛题背景

CVE平台的漏洞信息包含有CVE编号、漏洞评分、漏洞描述等内容，为更好地理解和持续研究，需将这些漏洞信息按照一定规则进行筛选分类。但人工筛选分类效率较低、耗时耗力，利用人工智能、通过自然语言处理则可能更好地解决这一问题。

赛题任务

本赛题分为A、B榜，平台将提供数据分类规则和具体内容，参赛选手需通过平台给出的已标注数据、未标注的漏洞数据，设计软件算法模型并不断完善模型对于分析数据结果的正确率。

赛题详情

所属赛道 数字安全专题赛

出题单位 北京科技大学网络空间安全与大数据智能应用实验室

技术方向 文本分类

赛题奖金 8万人民币

难易程度 ★★★

14

大数据平台安全事件检测与分类识别

赛题背景

日志审计分析是数据安全问题一种非常有效的风险应对手段，基于大数据平台日志、安全设备日志和平台网络流量等多源异构数据进行分析，可有效实现攻击行为的发现或预测，并进行溯源，保护企业或组织内部数据安全。

赛题任务

根据赛题提供的无标签大数据平台日志数据，参赛队利用机器学习、深度学习，UEBA等人工智能方法，构建系统用户使用大数据平台的行为基线和数据安全事件识别及分类模型。