Kubeflow用户研究：Data Scientist是一群什么生物?

李一帆 Caicloud 行业AI部

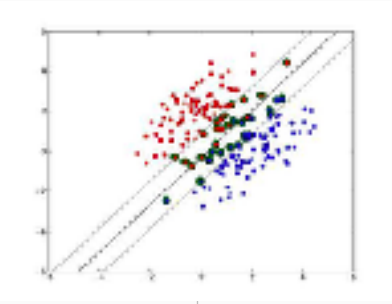
IT大咖说

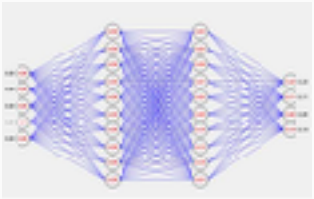
caicloud 才云

目录

caicloud 才云

开篇：工业界的AI





统计机器学习算法在互联网广告业务中被广泛adopt

机器学习算法被推广到了不同的场景

深度学习被用到图像，语音和文本数据之中，接连理论突破

2010左右

Tensorflow, Caffe等算法框架将深度学习的能力带出了学术圈

2014左右

开篇：工业界的AI

将深度学习算法用于商业问题的创业公司如雨后春笋

大型互联网公司也加入了AI军备竞赛，All in AI 成了日常战略

传统企业进入密集的数字化转型期，互联网+，AI+被带到了传统企业之中，诸如零售，能源，金融，制造等行业

百万AI人才荒

开篇：工业界的AI

工具的出现，附能了更多的人来加入AI事业

事业的发展，催生了新一代的AI工具

我们有了强大的算法基础工具：Tensorflow
但是发展的事业带来了万块GPU并行的AI业务，带来了百人到千人的AI算法团队

开篇：工业界的AI

DS触达算力的平台

DS合作的平台

GPU合作的平台

一个Solo DS

计算平台

数据

IDE

老师与FAQ

花书

证明自己的比赛

一个DS的工作流程

深度学习时代，90%的时间被花在等模型训练完

90%的时间被花在了研究数据，特征工程

一个Solo DS

“训练慢”：多GPU训练

“没数据”：批量的数据整理与标注

“又要配环境了”：社区极广泛和不同意的框架与库

一个DS团队

计算平台

老师与FAQ

数据

代码库

IDE

模型库?

发布流水线?

资源如何共享与协调?

模型场景化应用?

批训练?

在线大规模Serving?

批量推理?

一个DS团队

“你好了么？好了我要跑了”

“咦，这谁给我关了？”

“这个Cudnn怎么又坏了”

“这个cuda版本对么”

GPU资源管理难

GPU 1	GPU 2	GPU 3
DS 1正在占用到4.30	DS 1正在占用到4.30	DS 1正在占用到4.30

一个DS团队

“不对啊，这个模型文件和代码不匹配？”

“怎么回事儿，这个预测结果不对啊？”

AI Devops

“我上个版本的模型去哪儿了？”

不够

一个DS团队

分布式训练

不会分布式训练的DS

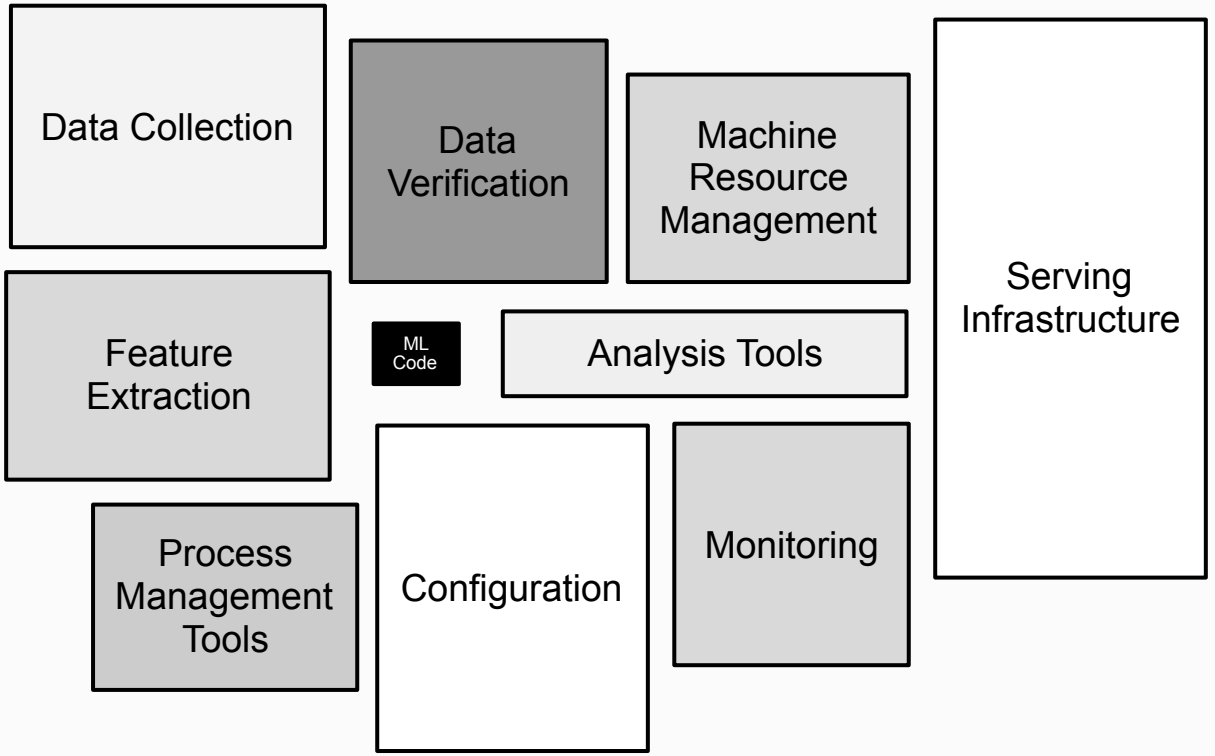
会分布式训练的DS

一个DS团队

一个DS团队最大的挑战在于，他们聚在一起是为了一个清晰的目标的，需要**有质量，高速度**的完成一个AI任务。不能有任何闪失！

我们面临各种开发上的挑战，都是AI快速发展给我们带来的技术债务！

Kubeflow 给我们的期待



caicloud 才云

才云行业AI部

通过AI和云技术，帮助行业客户完成AI的最后一公里落地

世界一流的算法团队

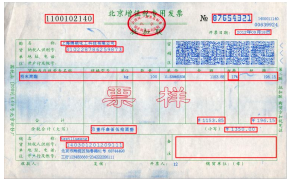
领先的深度学习训练平台

Clever
企业级分布式深度学习平台

领先的容器云平台

Compass
混合云部署，跨集群细粒度控制与联动

定制化OCR



物体检测与识别

人脸和人体系列检测与识别

某快消集团图像识别平台搭建

痛点问题一：手持海量图像无法全面检核

- 基于欣和庞大体量图像数据，单纯依靠人力无法做到每日、每店的陈列检查，只能降低检查频率和数量，达到初步的抽检目标

痛点问题二：终端货架陈列表现和数据无法高效全面获取

- 零售业日趋激烈竞争环境，需要全面掌控终端货架陈列表现和吸引力，并且需要实时动态报告

痛点问题三：线下活动数据采集数据质量、精细化、效率有待提升

- 线下活动（CE活动、移动烹饪教室等）数据采集人力成本、管理成本较高，数据质量难以满足精细化、高效管理

痛点问题四：会议管理无法实时智能管理

- 会前、会中、会后管控需要更加智能化，比如会议签到、现场管理、会后统计等方面

某电商图像+推荐识别平台搭建

In this e-commerce go global project, we are mainly providing two deep learning empowered backend modules in the project:

- Image Tagging module:
  - Trained on 3 on-premise K80 GPUs
  - Batch offline tagging every few days for SKU update, around 200k images for each update and 2 million SKU images from the starting point.
  - Online tagging for each active user, growing traffic with e-commerce site opening in new countries.
- User behavior based product recommendation module:
  - Deep learning recommendation engine based on data collected from click, like, purchase and other user behaviour events.

某银行智慧网点平台搭建

AI模型：视频人体追踪模型和生物识别模型

输入数据：监控摄像头数据

应用场景：在智慧银行网点的客户会在不同的智能售卖机之间移动，进行浏览和购买操作。通过网点中的摄像头对网点客户进行捕捉，并且描绘他们在网点中运动的轨迹，可以刻画出他们在不同售卖机的停留时间，进而推断他们对于不同理财产品的关注程度。

应用对于银行网点的价值：全方位的把握客户在网点中的动作有助于网点的精准营销推荐和网点管理，提升大堂经理和总行对于网点中客户行为的理解。


CAI
CLOUD
ACCELERATION
INTELLIGENCE