众包数据标注及其问题

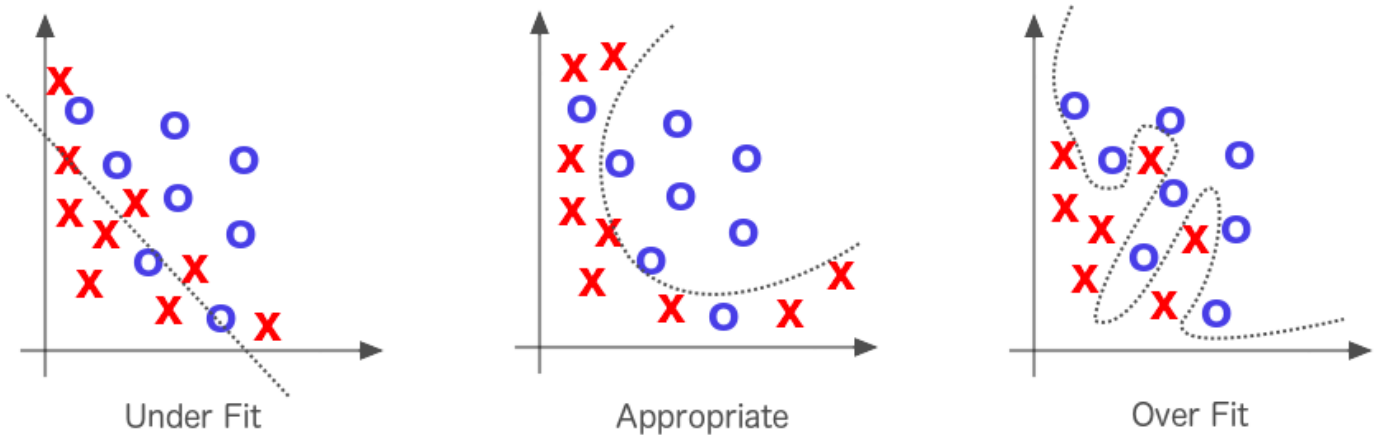
软件学院-2018

概要

- 数据标注与众包
- 众包数据标注存在的问题

数据规模与深度学习

- 数据规模过大-欠拟合
- 数据规模过小-过拟合



数据规模与深度学习

- Revisiting Unreasonable Effectiveness of Data in Deep Learning Era

Initialization	mIOU
ImageNet	73.6
300M	75.3
ImageNet+300M	**76.5**

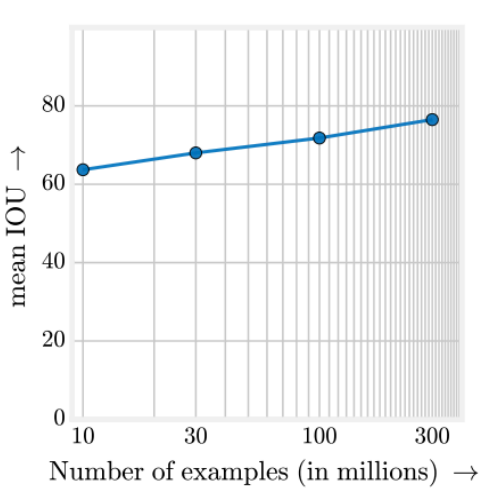
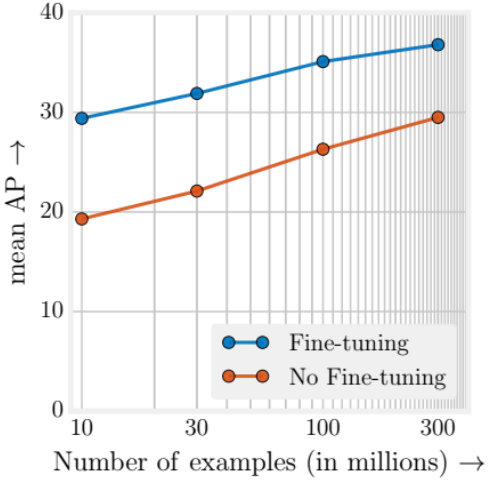


Figure 6. Semantic segmentation performance on Pascal VOC



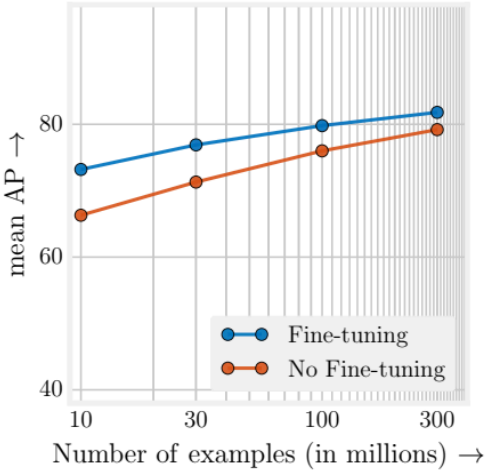
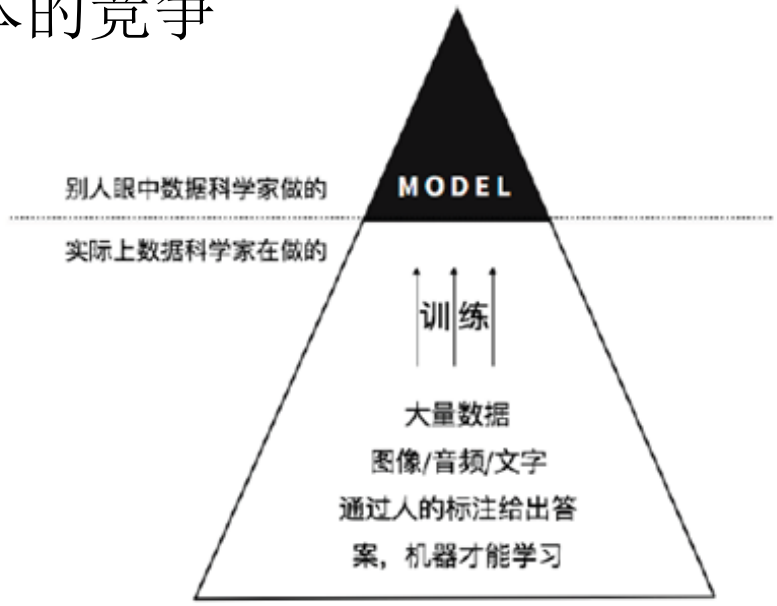


Figure 4. Object detection performance when initial checkpoints

数据规模与深度学习

- [大量数据+普通模型]＞[普通数据+高级模型]
- AI界最根本的竞争



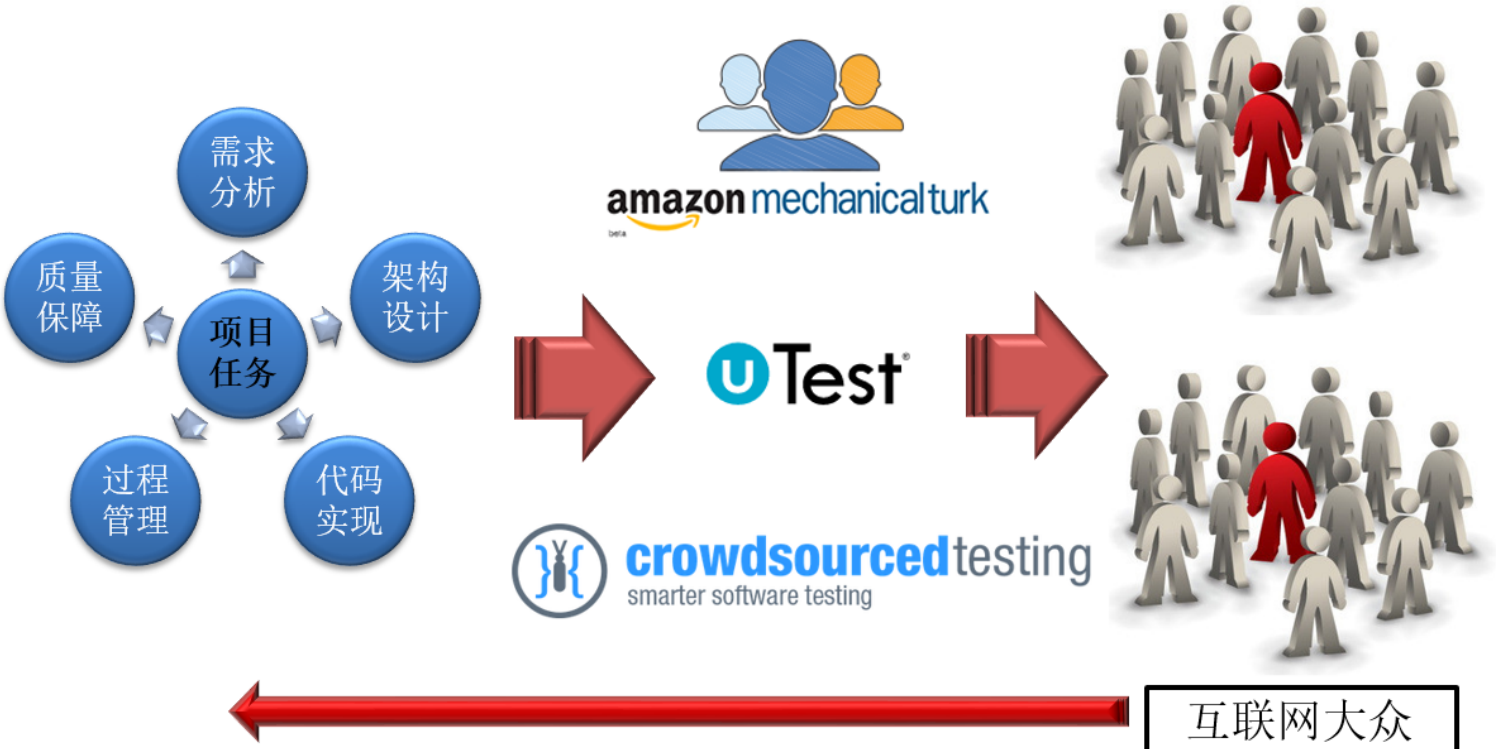
数据来源

- 人工标注 v.s. 智能化标注

数据标注

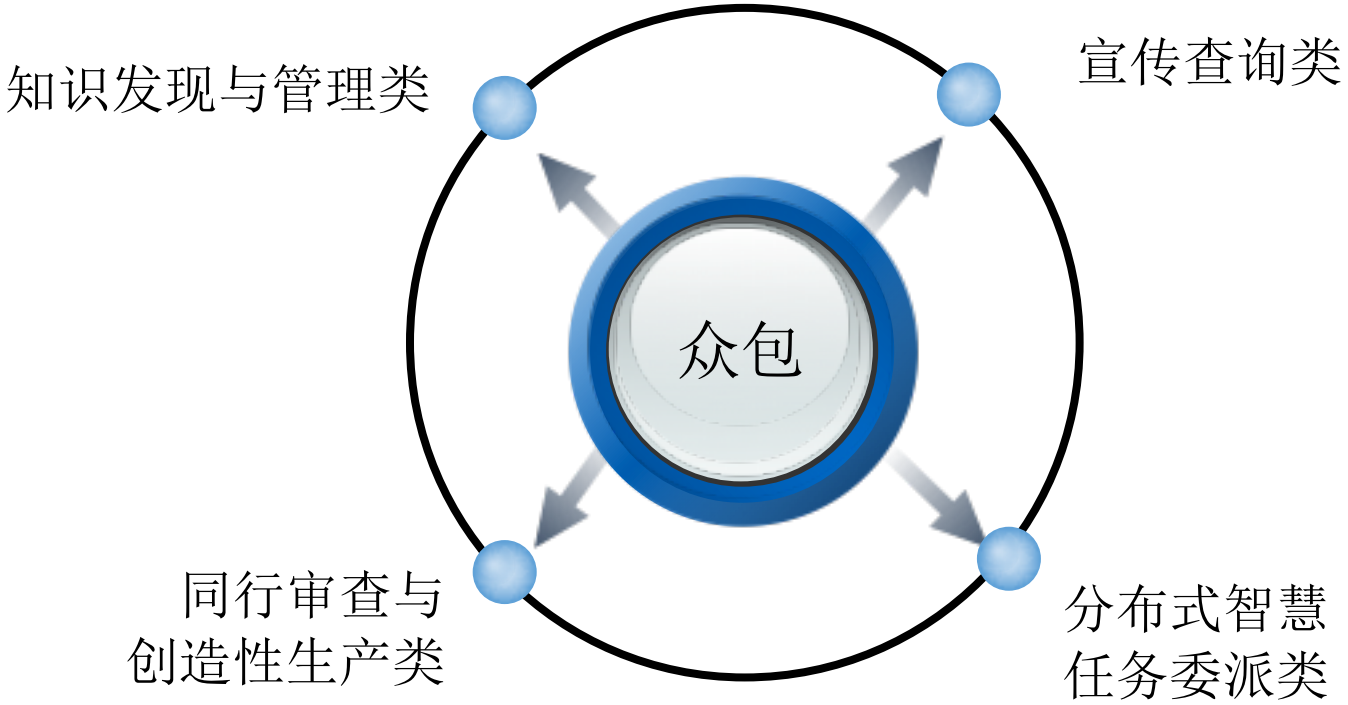
- 数据标注是指对数据进行结构化处理
  - 累计先验知识
  - 注入人工智能
- 数据标注类型
  - Classification
  - Detection
  - Segmentation
  - Caption
  - Attribute

众包



众包是指任务以自由的形式（通常基于互联网）外包给非特定大规模群体的组织模式

众包分类



众包流程

众包平台对比

众包平台	用户	功能服务	任务类型	界面
Mturk	任务请求人：企业或个人 工人：以个人为主	为请求人提供任务发布、接收结果等服务 为工人提供任务搜索、结果提交、收取报酬等服务	微观任务	图形化、API 接口
CrowdFlower	任务请求人：企业或个人 工人：其他众包平台上的工人	只为请求人提供服务：任务分解、任务设计、任务发布、结果质量检查等	微观任务	图形化
samasource	任务请求人：企业 工人：由 samasource 认证的贫困人员	为请求人提供任务分解、任务发布、结果整合等完整的服务 为工人进行基本的培训	微观任务	未公开
CloudCrowd	任务请求人：企业 工人：以个人为主	为请求人提供任务设计、结果整合等完整的业务服务 为工人提供结果提交、收取报酬等服务	微观任务	图形化
脑力库	任务请求人：以企业为主 工人：专业设计人员	为请求人提供任务发布、保密任务等服务，为长期发布任务的请求人提供专栏显示 为工人提供与请求人沟通、收取报酬等服务	宏观任务	图形化
猪八戒	任务请求人：企业或个人 工人（服务商）：企业或个人	为请求人提供选择服务商、发布任务需求等服务 服务商可以发布提供的服务内容，服务商可以交换意见	宏观任务 微观任务	图形化
三打哈	任务请求人：以企业为主 工人：以个人为主	为请求人提供推广任务方案、任务渠道等服务 为工人提供任务搜索、工人交流等服务	宏观任务 微观任务	图形化

众包数据标注存在的问题

- 任务准备
- 任务执行
- 任务答案整合
- 数据质量

众包数据标注存在的问题-任务准备

- 任务准备
  - 任务标价
  - 欺诈者处理
  - 花费、质 量、时 间 的 平 衡
  - 任务界面
- 任务选择
  - 任务搜索
  - 任务推荐

众包平台	关键词搜索	地域搜索	类别搜索	按时间排序	按价格排序
Mturk	√			√	√
CloudCrowd	√			√	√
脑力库	√			√	√
猪八戒	√	√	√		√
三打哈	√		√	√	√

众包数据标注存在的问题-任务执行

- 动态任务分配
- 动态标价

定价策略	定价依据	策略评价
固定标价	任务难度、任务颗粒、奖励机制	优点：实现简单 缺点：价格过低导致任务完成时间增加；价格过高导致任务花费代价大
在线标价	工人预期	优点：价格更加合理 缺点：要对任务价格进行多次调整，增加了任务选择时间

众包数据标注存在的问题-任务答案整合

- 多数投票原则
- 固定准确率
- 动态准确率

推断方法	基本思想	方法评价
单纯利用工人答案	多数投票原则 文献[28,33,35,84]	优点:方法简单,利于实现 缺点:结果准确率不高
结合工人答案和答题准确率	固定的工人答题准确率 文献[71,78]	优点:结果准确率较高 缺点:工人准确率的变化导致结果质量降低,需要预知工人答题准确率
	变化的工人答题准确率 文献[51,59,94]	优点:结果准确率很高,不需要预知工人答题准确率 缺点:基于 EM 的方法时间代价较高

众包数据标注存在的问题-数据质量

- 数据获取
- 数据清洗
- 数据标注
- 数据审查