Baidu 百度

构建计算机视觉数据集的若干思考

孙逸鹏
百度视觉技术

RACV-2020@厦门 2020.8.28

构建计算机视觉数据集

①为什么要构建CV数据集

②什么样的CV数据集值得构建

③如何高效构建CV数据集

④业界期待怎样的数据集

① 为什么需要构建CV数据集

- **提出并定义新的研究问题和场景**
  - 建立研究问题范式，抽象实际应用痛点：开放问题 → 封闭问题（可量化、可复现）

研究模式：基于数据驱动的深度学习

✓ 引擎
深度学习等算法

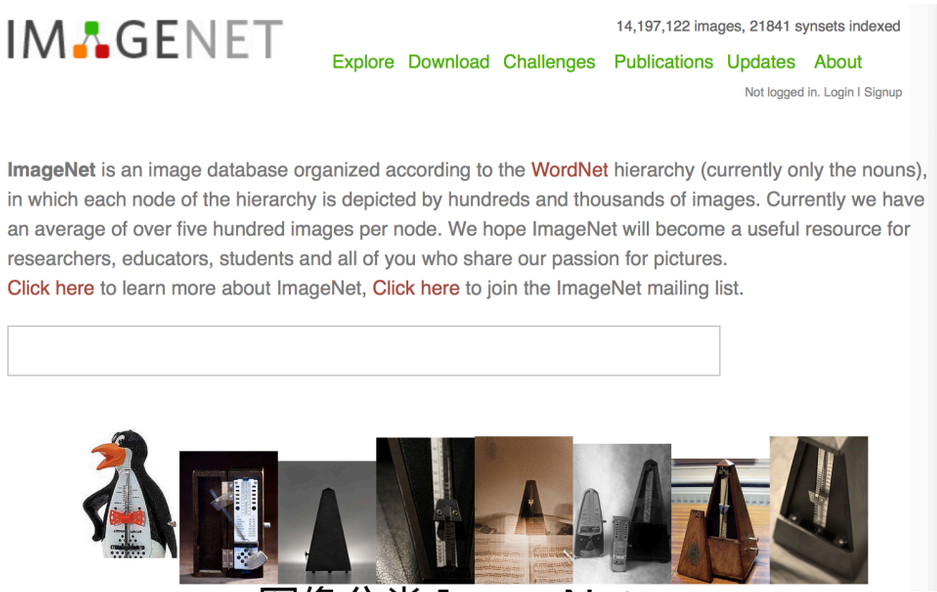
✓ 燃料
大量训练数据

CV研究员/开发者：“巧妇难为无米之炊”

✓ 需要丰富的“食材” → 创造新的“菜肴”

② 什么样的CV数据集值得构建

- **数据集是推动CV研究与应用发展的关键**



图像分类 ImageNet

人脸识别与生成

文字定位与识别

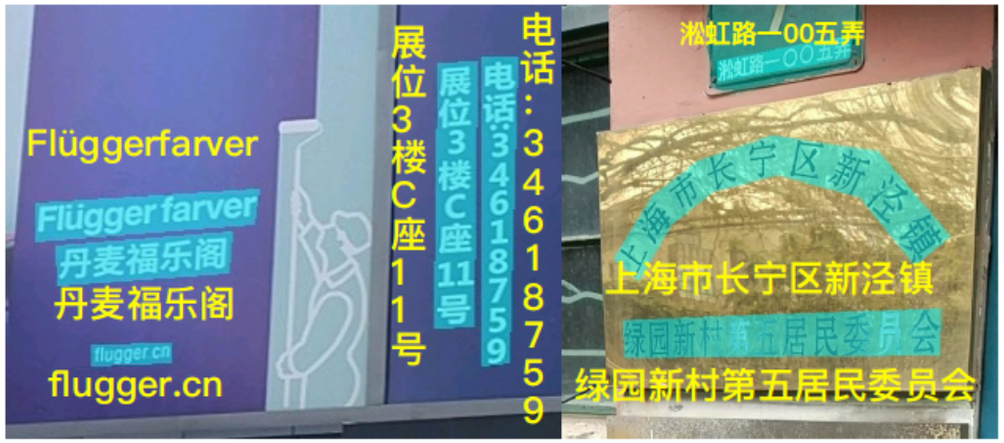
- **一个成功的CV数据集的主要特点**
  - **学术影响力**：开辟新的研究领域，并成为重要的benchmark评测标准
  - **实用性价值**：助力跨越可用门槛，可投入规模化生产帮助客户创造收益

③ 如何高效构建计算机视觉数据集

- **高效构建数据集**

  1. 基于应用痛点和研究动机，明确问题描述与关键任务

  - 输入$\mathcal{X}$、输出$\mathcal{Y}$、任务映射 $\mathcal{F}: \mathcal{X} \rightarrow \mathcal{Y}$



ICDAR 2019-LSVT 大规模街景文字识别

最大中文OCR集合：其它集合的14倍以上

（5万张位置/文字精标注 + 40万关键词弱标注）

ICDAR 2019-ArT 任意形状场景文字识别

最大任意形状OCR集合：任意形状文字可变点数标注

③ 如何高效构建计算机视觉数据集

2. 定义数据源、标注形式、评价准则

百度地图街景

少量精确位置标注 + 大量文本弱标注

采集、收集
爬取、生成

→

利用任务场景信息、用户反馈
自动/半自动生成、脱敏、人工标注

→

客观评价指标
体现主观效果感知

③ 如何高效构建计算机视觉数据集

3. **集合设计有效性**：Train, Val, Test比例合理，确保分布一致性

4. **数据标注专业化**：高效专业的众包模式标注平台、标注系统

百度众包标注平台

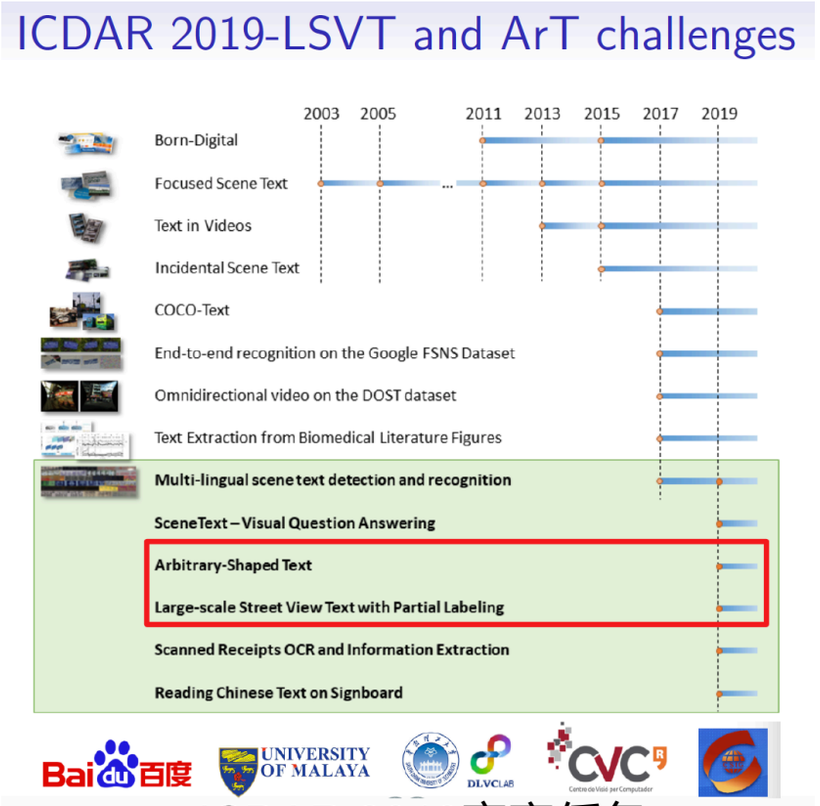
规范高效的标注培训及标注工具平台

③ 如何高效构建计算机视觉数据集

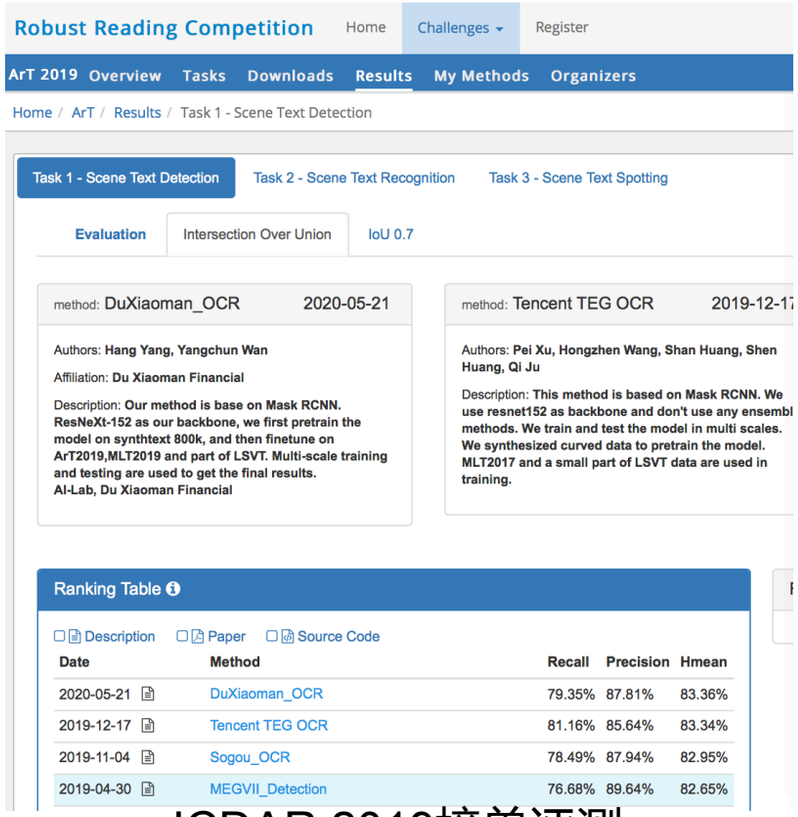
5. 可靠的算法基线：基于领域认可的SOTA方法给出可靠的baseline指标效果

PaddleOCR开源库：云+端训练部署一体 https://github.com/PaddlePaddle/PaddleOCR

6. 自动的评测系统：国际竞赛workshop，搭建维护benchmark榜单自动评测系统



ICDAR 2019竞赛任务



ICDAR 2019榜单评测

④ CV数据集新需求与期待

- **需要更贴近现实任务现状的数据集场景**
  - **无人/少人手工标注参与，抽象体现真实应用现状与视觉AI领域难题**

- **现实问题与场景现状**
  - 规模与标注质量
    - 海量无标注、弱标注、有噪声图像/视频数据
    - 有限的精确标注、有监督训练数据
  - 异构数据
    - 多种模态、不同标注体系、不同数据domain

移动互联网等海量视觉数据来源

总结与展望

- **工业界**
  - 丰富的数据来源
  - 真实的需求应用
  - 更丰富的计算资源
  - 实际可用的效果性能为准
  - 营收/增速是关键业绩目标

  …

- **学术界**
  - 前沿理论方法的探索与突破
  - 专注持久的研究投入
  - 注重研究方法凝练与思想拔高
  - 创新点带来的提升增量为主
  - 高水平论文/专利等为产出成果

  …

在碰撞与协作中求同存异、相互促进激发和谐共建，创造视觉AI领域新高度！

Q&A

谢谢!

sunyipeng@baidu.com