NLP+VS丨深度学习数据集标注工具、图像语料数据库、实验室搜索ing...

原创

悟乙己 于 2017-02-07 12:12:01 发布

分类专栏： 图像丨相关技术跟踪与商业变现 文章标签： 数据标注 图像标注 数据集

本文链接：https://blog.csdn.net/sinat_26917383/article/details/54908389

~~因为不太会使用opencv、matlab工具，所以在找一些比较简单的工具。

.

一、NLP标注工具

来源：《构想：中文文本标注工具（附开源文本标注工具列表）》

Chinese-Annotator

来源：https://github.com/crownpku/Chinese-Annotator

能不能构建一个中文文本的标注工具，可以达到以下两个特点：

标注过程背后含有智能算法，将人工重复劳动降到最低；

标注界面显而易见地友好，让标注操作尽可能简便和符合直觉。

答案是可以的。事实上很多标注工具已经做到了这一点，最先进的如Explosion.ai的Prodigy；然而开发了著名的NLP开源包Spacy的explosion.ai选择了将Prodigy闭源，而Spacy支持中文也仍然遥遥无期。我们希望构建一个开源的中文文本标注工具，而本文很多的技术灵感正是来自Prodigy文档。
流程：

用户标一个label

主动学习的后台算法分为online和offline部分。online部分即时更新模型，可使用诸如SVM、bag of words等尽可能快的传统方法；offline部分当标注数据积累到一定数量时更新模型，可使用准确度较高的深度学习模型。

模型更新后，对尽可能多的example做预测，将确信度排序，取确信度最低的一个example作为待标注例子。重复1的过程。

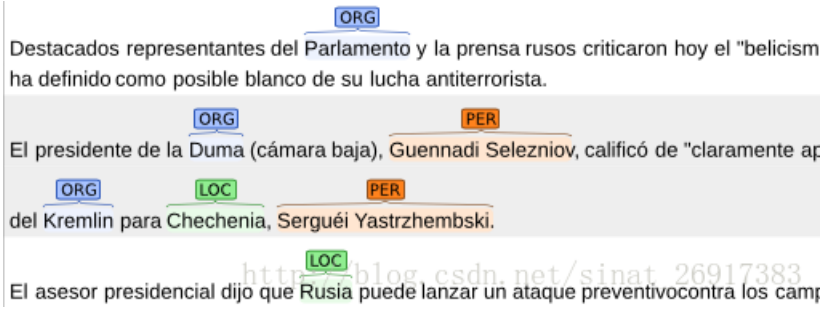
可以想象如果模型训练得好的话，这个过程将直接忽略掉确信度最大的那些例子，而把所有重点放在分类边界上的那些确信度小的例子。这样可以尽算法所能减少用户端的人工工作量。

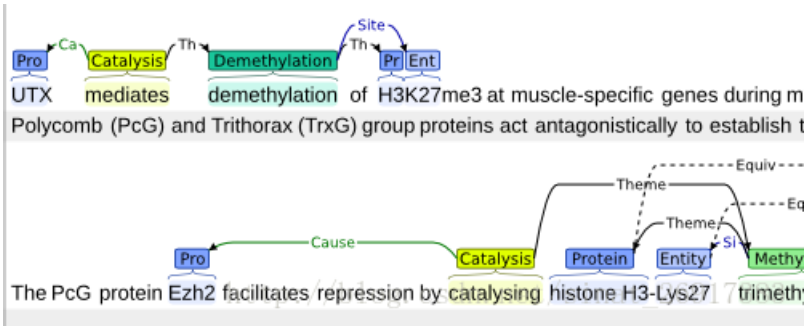
1 BRAT

BRAT是一个基于web的文本标注工具，主要用于对文本的结构化标注，用BRAT生成的标注结果能够把无结构化的原始文本结构化，供计算机处理。利用该工具可以方便的获得各项NLP任务需要的标注语料。以下是利用该工具进行命名实体识别任务的标注例子。

WeTest舆情团队在使用：http://wetest.qq.com/bee/

使用案例：http://blog.csdn.net/owengbs/article/details/49780225





2 Prodigy

Prodigy给了一个非常好的demo，每一次的标注只需要用户解决一个case的问题。以文本分类为例，对于算法给出的分类结果，只需要点击“正确”提供正样本，“错误”提供负样本，“略过”将不相关的信息滤除，“Redo”让用户撤回操作，四个功能键以最简模式让用户进行标注操作。

真正应用中，应该还要加入一个用户自己加入标注的交互方式，比如用户可以高亮一个词然后选择是“公司”，或者链接两个实体选择他们的关系等等。

3 IEPY

整个工程比较完整，有用户管理系统。前端略重，对用户不是非常友好

代码 https://github.com/machinalis/iepy

说明 http://iepy.readthedocs.io/en/latest/index.html

4、DeepDive (Mindtagger)

介绍 http://deepdive.stanford.edu/labeling

前端比较简单，用户界面友好。

前端代码 https://github.com/HazyResearch/mindbender

将DeepDive的corenlp部分转为支持中文的代码尝试：

https://github.com/SongRb/DeepDiveChineseApps

https://github.com/qiangsiwei/DeepDive_Chinese

https://github.com/mcavdar/deepdive/commit/6882178cbd38a5bbbf4eee8b76b1e215537425b2

5 SUTDAnnotator

用的不是网页前端而是pythonGUI，但比较轻量。

代码 https://github.com/jiesutd/SUTDAnnotator

Paper https://github.com/jiesutd/SUTDAnnotator/blob/master/lrec2018.pdf

支持中文

6 SnorkelPage: https://hazyresearch.github.io/snorkel/

Github: https://github.com/HazyResearch/snorkel

Demo Paper: https://hazyresearch.github.io/snorkel/pdfs/snorkel_demo.pdf

7 Slate

Code: https://bitbucket.org/dainkaplan/slate/

Paper: http://www.jlcl.org/2011_Heft2/11.pdf

8 Prodigy

和著名的spacy是一家做的

Website: https://prodi.gy/docs/

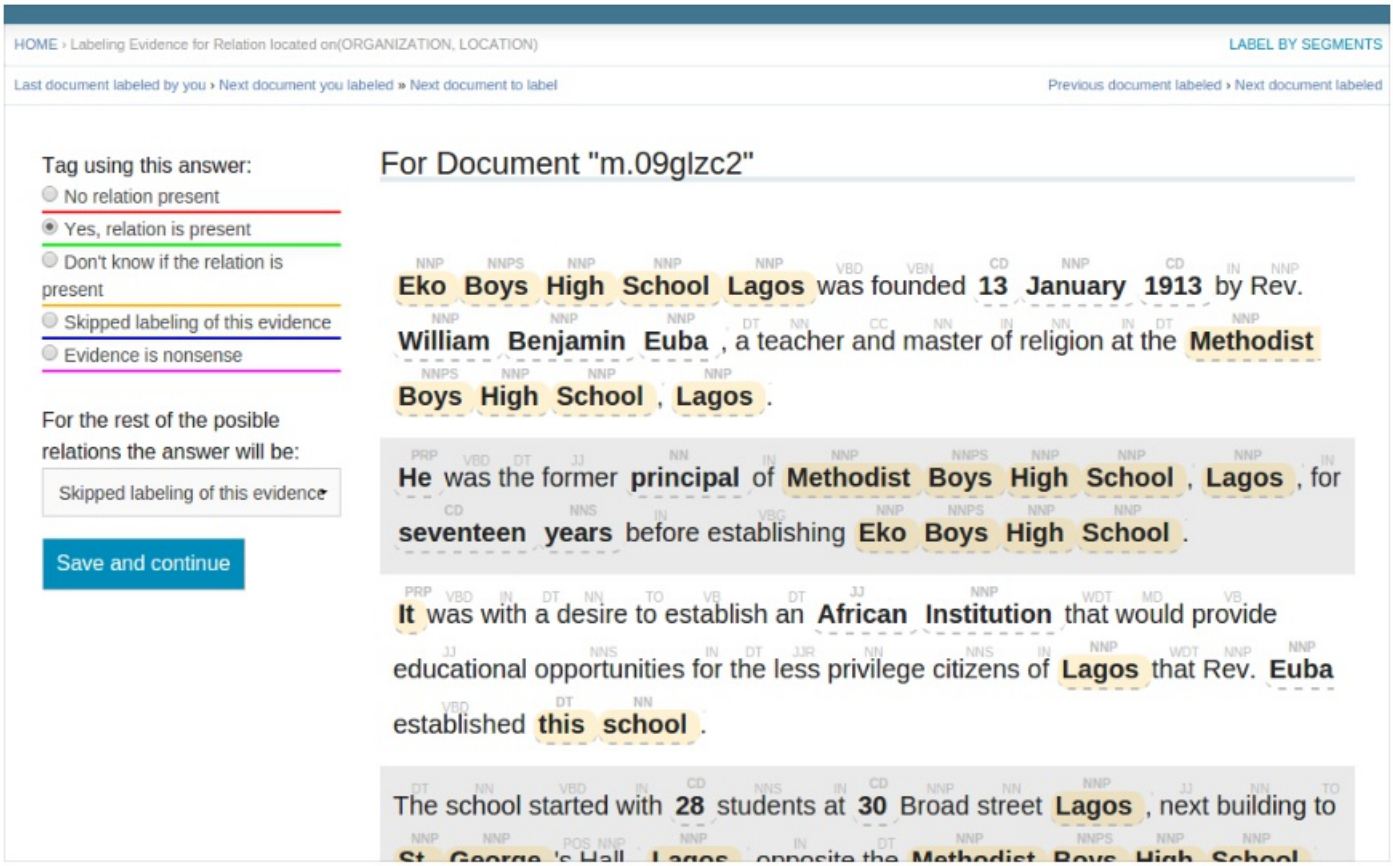
Blog: https://explosion.ai/blog/prodigy-annotation-tool-active-learning

开源：https://github.com/crownpku/Chinese-Annotator

9 几个开源文本标注工具的简单调研

https://github.com/deepwel/Chinese-Annotator/blob/master/docs/annotator_examples.md

IEPY



整个工程比较完整，有用户管理系统。前端略重，对用户不是非常友好

代码 https://github.com/machinalis/iepy

说明 http://iepy.readthedocs.io/en/latest/index.html

.

.

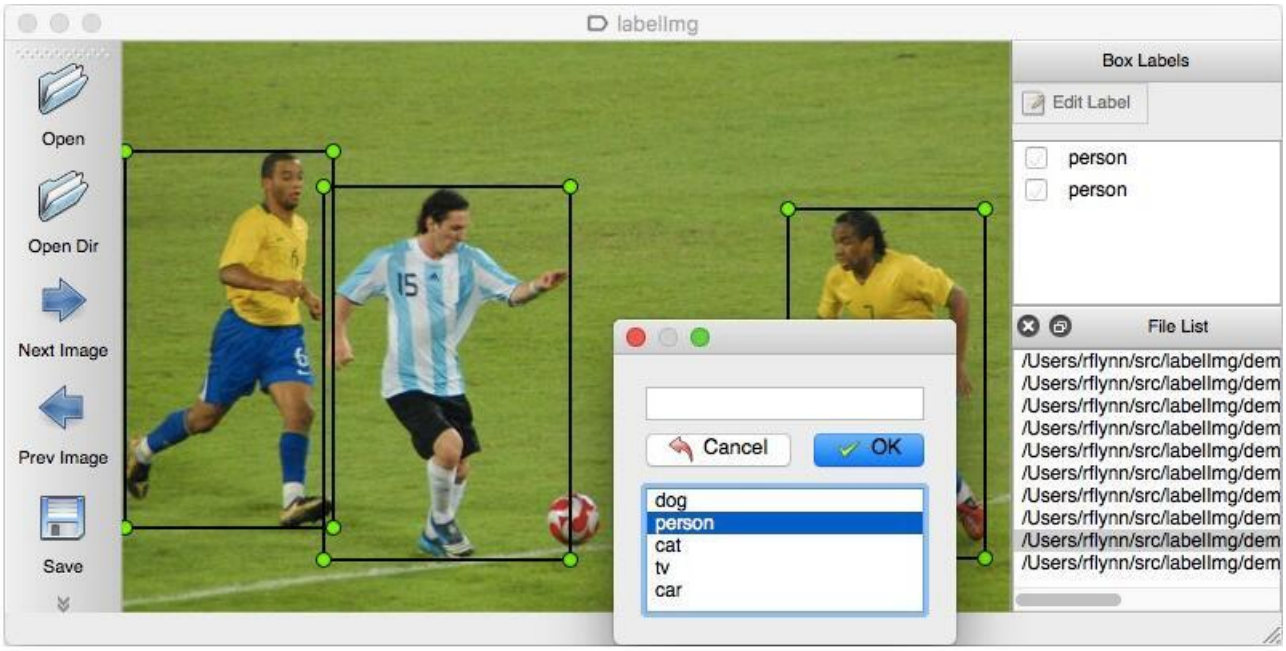
二、VS标注工具——LabelImg

1、PyQt

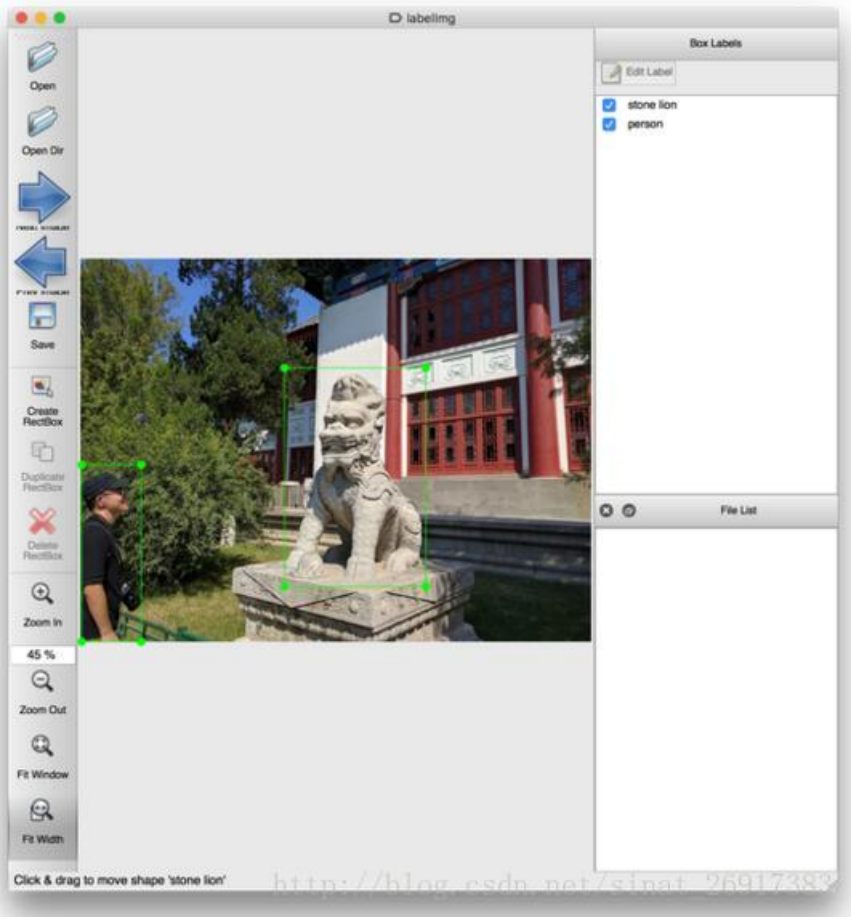
用 PyQt 写的, 很轻量, Linux/macOS/Windows 全平台均可运行.

工具github网址：https://github.com/tzutalin/labelImg

知乎介绍网址：有图像标注工具推荐或者分享吗？



Watch a demo video by author tzutalin



2、Vatic

参考：人工智能AI工具-视频标注工具vatic的搭建和使用

视频标注工具vatic，Vatic源自MIT的一个研究项目(Video Annotation Tool from Irvine, California)。输入一段视频，支持自动抽取成粒度合适的标注任务并在流程上支持接入亚马逊的众包平台Mechanical Turk。

网址：http://web.mit.edu/vondrick/vatic/

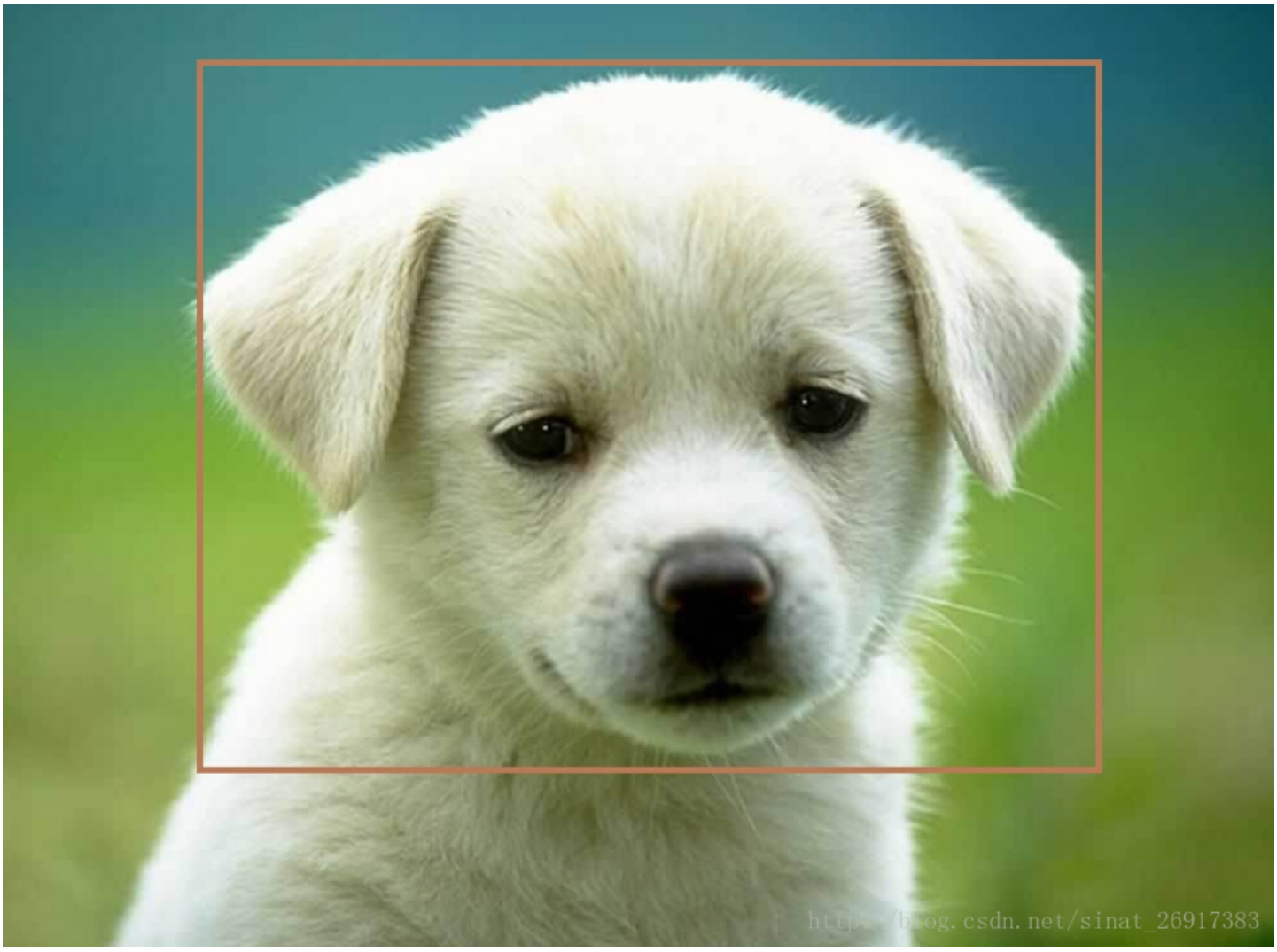
Vatic源自MIT的一个研究项目(Video Annotation Tool from Irvine, California)。输入一段视频，支持自动抽取成粒度合适的标注任务并在流程上支持接入亚马逊的众包平台Mechanical Turk。除此之外，其还有很多实用的特性：

1.简洁使用的GUI界面，支持多种快捷键操作

2.基于opencv的tracking，这样就可以抽样的标注，减少工作量

具体使用时，可以设定要标注的物体label，比如：水果，人，车，等等。然后指派任务给到众包平台（也可是自己的数据工程师）。现阶段支持的标注样式是框（box）。一个示例，下图标注了NBA直播比赛中的运动员

3、BBox-Label-Tool



4、图像标注VS2013项目

有人自己写了一个版本：

打框的代码(c++)我封装成了dll，下载地址：图像标注VS2013项目 （我的环境是win7vs2013旗舰版，win8 win10好像不能运行）

别人封装的opencv动态库，现在修改为opencv2.4.10，64位，vs2013，按网上教程配置

好opencv，资源地址：
图像标注EXE-2016-10-18
上面的代码好像忘写操作说明了，这里写一下：
（1）图片显示出来后，输入法切换到英文；
（2）在目标的左上角按下鼠标左键，拉一个包围框到目标右下角，然后键盘输入标签(一个字符)
（3）继续（2）操作，直到框完该张图片上的目标；
（4）按n进入下一张，esc退出。
注意：标签只能输入一个字符，你可以在生成的txt文件中替换成你实际的标签。
.
5、Yolo_mark

YOLO V2 准备数据的图形界面目标边界框标注工具 AlexeyAB/Yolo_mark

.
6、视频标注工具

CDVA（compact descriptor for video analysis），主要是基于CDVS中的紧凑视觉描述子来做视频分析，之前是紧凑视觉描述子主要应用在图像检索领域。需要制作新的数据集，对视频帧进行标注，所以根据网上一个博主的标注工具进行了一定的修改，实现的功能是在每一帧中将需要标注的区域用鼠标选取4个点，顺序是顺时针。因为四边形的范围更广，之前的一些人直接标注了矩形，但是在一些仿射变换中，往往矩形的定位效果不好，矩形定位应该比较适合于人脸定位和行人定位之中。
http://www.cnblogs.com/louyihang-loves-baiyan/p/4457462.html

7、微软发布的可视化图像/视频标记工具 VoTT

该工具支持以下功能：

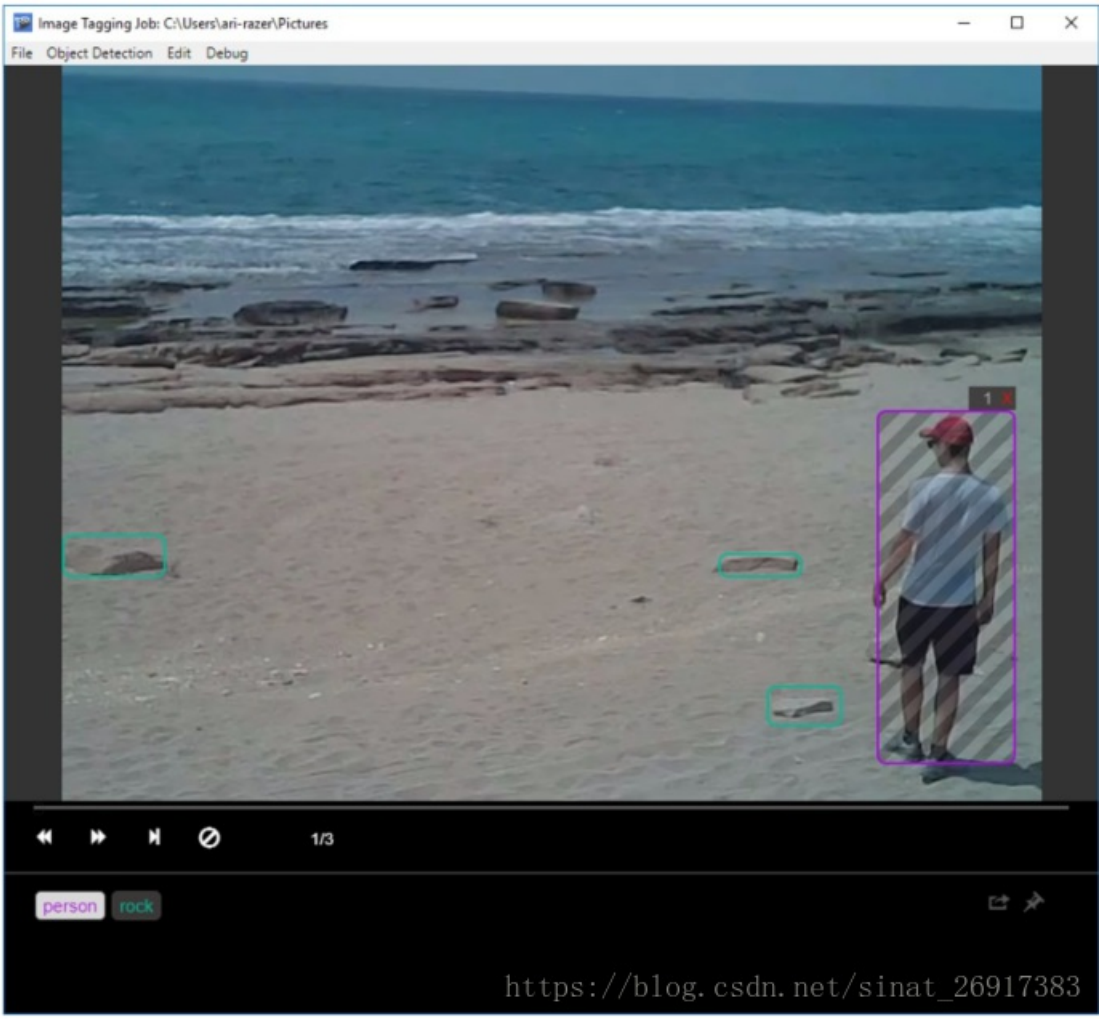
能够标记和注释图像目录或独立视频。

使用 Camshift 跟踪算法辅助计算机标记和跟踪视频中的物体。

将标签和资源导出到 Custom Vision Service CNTK，Tensorflow（PascalVOC）或YOLO 格式，用于训练对象检测模型。

在新视频中使用主动学习与训练对象检测模型（本地或远程）结合生成更强大的模型。

Github 链接：



8、精灵标注助手-AI数据集标注工具

阿里云市场上公开售卖的，厉害了！

.

.

三、Amazon’s Mechanical Turk 离线工作框架

一个开源的Amazon’s Mechanical Turk 离线工作框架，基于Django搭建的
github网址：https://github.com/hltcoe/turkle

.

.

四、用已训练来进行图像标注

《使用深度学习和Fisher向量进行图片标注》（paper）

主讲人Lior Wolf，特拉维夫大学的教员在一次伦敦深度学习会议上的一次公开演讲：
为了实现图像标注和搜索，他们最开始用CNNs将图片转换成向量，用Word2Vec将词语转换成向量。大部分研究工作都集中于如何将词语向量结合到语句向量之中，由此产生了基于Fisher向量的模型。一旦他们得到了语句向量，他们使用典型相关分析（CCA）将图片表示和语句表示投射到同一空间里，使图像和句子可以匹配，找到最近邻的部分。

参考自博客：2015伦敦深度学习峰会笔记：来自DeepMind、Clarifai等大神的分享

.

五、snape

人工数据集生成工具，来看一段有趣的独白：

> Snape is primarily used for creating complex datasets that challenge students and teach defense against the dark arts of machine learning.

专门是针对机器学习领域自动生成数据集。
安装：
Via Github

```
git clone https://github.com/mbernico/snape.git
cd snape
python setup.py install
```

来自：https://github.com/mbernico/snape

.
.

延伸一 国内一些众包的数据标注服务商

1、敲宝网——众包

里面确实有一些图像分类、图像标注的任务。但是也不是很多。

2、小鱼儿网

我的技能时间交易平台小鱼儿网成立最晚，但却走了最具互联网思维的盈利之路，增值服务盈利，平台在整个过程交易中不收取费用，提供大数据分析，筛选服务者等增值服务，主动权完全交给用户，互联网时代，流量为王，用户为王，小鱼儿网的盈利模式无疑向这个宗旨贴近的，长期来看，这种盈利模式或许最聪明。

挺大的，但是没有看到有图像的任务。

3、威客-创意,一品威客网

中国最专业威客网站一品威客网借鉴了猪八戒盈利模式的短板，对用户划分普通用户和vip用户，对普通用户实行免费，对VIP用户收取会员费，在互联网时代，有效的笼络住了大批用户的心，不失为一种好的盈利模式。
国内最大的众包了吧，但是图像标识项目很少，商家也几乎没有看到...

4、数据堂

确确实实有数据标注，而且有文本、语音、图片采集项目。

5、百度众包

里面有很多任务与案例，文本、语音、图片都有。

6、阿里众包

图像采集任务？

7、荟萃公司——荟萃-荟集人力之萃

http://huicui.me/?from=singlemessage&isappinstalled=0

图片识别

可智能识别图片内容、属性、分类、是否涉黄等，支持单图多图多种形式。

语音转化

可替您将文字转成语音、文字转成方言（真人语音），识别语音、歌曲等。

视频识别

可以为您完成视频内容收集，字幕识别，视频内容鉴定等内容。

视频创作

为您拍摄或收集某一主题的视频，以小视频形式上传。

网页展示任务

可自定义任意网页在用户端展示时间，如新品推广、广告观看等类型。

自定义任务
抢票？秒杀？联系上下文？只要你脑洞够大，任意H5网页类任务皆可接入。

8、地平线公司
http://www.horizon-robotics.com/index_cn.html

地平线具有世界领先的深度学习和决策推理算法开发能力，将算法集成在高性能、低功耗、低成本的嵌入式人工智能处理器及软硬件平台上。地平线目前提供基于ARM/FPGA等处理器的解决方案，同时开发自主设计研发的Brain Processing Unit (BPU) —一种创新的嵌入式人工智能处理器架构IP，提供设备端上完整开放的嵌入式人工智能解决方案。

公司核心业务面向智能驾驶和智能生活等应用场景，目前已成功推出了面向智能驾驶应用的“雨果”平台及面向智能生活的“安徒生”平台，与国内国际顶尖的汽车Tier 1、OEMs及家电厂商展开了深入的合作，并在成立仅一年多的时间内成功推出量产产品。地平线也正积极搭建开放的嵌入式人工智能产业生态，与产业上下游共同合作发展。

2017年1月6日，地平线与英特尔于CES联合发布了基于单目摄像头和FPGA的最新ADAS系统，可实现在高速公路和市区道路场景下，同时对行人、车辆、车道线和可行驶区域的实时检测和识别。2016年8月1日，地平线与美的联合发布了“智能王”柜机空调，拥有手势控制、智能送风、智能安防三大新功能。

.
.

六、图像数据集

一部分来源：深度学习视觉领域常用数据集汇总

1、LSUN：用于场景理解和多任务辅助（房间布局估计，显着性预测等）。

Category	Training	Validation
Bedroom	3,033,042 images (43 GB)	300 images
Bridge	818,687 images (16 GB)	300 images
Church Outdoor	126,227 images (2.3 GB)	300 images
Classroom	168,103 images (3.1 GB)	300 images
Conference Room	229,069 images (3.8 GB)	300 images
Dining Room	657,571 images (11 GB)	300 images
Kitchen	2,212,277 images (34 GB)	300 images
Living Room	1,315,802 images (22 GB)	300 images
Restaurant	626,331 images (13 GB)	300 images
Tower	708,264 images (12 GB)	300 images
Testing Set	10,000 images (173 MB)	

地址：http://lsun.cs.princeton.edu/2016/

2、行人检测DataSets

（1）.基于背景建模：利用背景建模方法，提取出前景运动的目标，在目标区域内进行特征提取，然后利用分类器进行分类，判断是否包含行人；
（2）.基于统计学习的方法：这也是目前行人检测最常用的方法，根据大量的样本构建行人检测分类器。提取的特征主要有目标的灰度、边缘、纹理、颜色、梯度直方图等信息。分类器主要包括神经网络、SVM、adaboost以及现在被计算机视觉视为宠儿的深度学习。

Caltech行人数据库：http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/

该数据库是目前规模较大的行人数据库，采用车载摄像头拍摄，约10个小时左右，视频的分辨率为640×480，30帧/秒。标注了约250,000帧（约137分钟），350000个矩形框，2300个行人，另外还对矩形框之间的时间对应关系及其遮挡的情况进行标注。数据集分为set00$_{set10}$，其中set00set05为训练集，set06$_{set10}$为测试集（标注信息尚未公开）。性能评估方法有以下三种：（1）用外部数据进行训练，在set06set10进行测试；（2）6-fold交叉验证，选择其中的5个做训练，另外一个做测试，调整参数，最后给出训练集上的性能；（3）用set00$_{set05训练}$，set06set10做测试。由于测试集的标注信息没有公开，需要提交给Pitor Dollar。结果提交方法为每30帧做一个测试，将结果保存在txt文档中（文件的命名方式为I00029.txt I00059.txt ……），每个txt文件中的每行表示检测到一个行人，格式为“[left, top,width, height, score]”。如果没有检测到任何行人，则txt文档为空。该数据库还提供了相应的Matlab工具包，包括视频标注信息的读取、画ROC（Receiver Operatingcharacteristic Curve）曲线图和非极大值抑制等工具。

其他数据集可参考：行人检测：http://www.52ml.net/17004.html

3、人脸数据库UMDFaces等

（1）UMDFaces

http://www.umdfaces.io/

不仅有人脸的目标检测数据，还有关键点的数据，非常适合做训练。

就是比较大，总共有三个文件，一共8000+个类别，总共36W张人脸图片，全都是经过标注的样本，标注信息保存在csv文件中，除了人脸的box，还有人脸特征点的方位信息，强力推荐！

（2）人脸识别数据库

1. 李子青组的 CASIA-WebFace(50万，1万个人). 需申请.Center for Biometrics and Security Research
2. 华盛顿大学百万人脸MegaFace数据集. 邮件申请, 是一个60G的压缩文件. MegaFace
3. 南洋理工 WLFDB. (70万+,6,025). 需申请. WLFDB : Weakly Labeled Faces Database
4. 微软的MSRA-CFW (202792 张, 1583人). 可以直接通过OneDrive下载.MSRA-CFW: Data Set of Celebrity Faces on the Web
5. 汤晓欧实验室的CelebA(20万+), 标注信息丰富. 现在可以直接从百度网盘下载 Large-scale CelebFaces Attributes (CelebA) Dataset
6. FaceScrub. 提供图片下载链接（100,100张，530人）. vintage – resources

作者：疾如风

链接：https://www.zhihu.com/question/33505655/answer/67492825

来源：知乎

4、搜狗实验室数据集：

http://www.sogou.com/labs/dl/p.html

互联网图片库来自sogou图片搜索所索引的部分数据。其中收集了包括人物、动物、建筑、机械、风景、运动等类别，总数高达2,836,535张图片。对于每张图片，数据集中给出了图片的原图、缩略图、所在网页以及所在网页中的相关文本。200多G

```
格式说明：
共包括三个文件：Meta_Data,Original_Pic,Evaluation_Data。其中Meta_Data存储图片的相关元数据；Original_Pic中存储图片的原图；Evaluation_Data是识图搜索结果的人工标注集合。
Meta_Data文件包含所有图片的相关元数据，格式如下：
<PIC>
<PIC_URL>图片在互联网中的URL地址</PIC_URL>
<PAGE_URL>图片所在网页的URL地址</PAGE_URL>
<ALT_TEXT>图片的替换文字</ALT_TEXT>
<ANCHOR_TEXT>以图片为目标的超链接的显示文本</ANCHOR_TEXT>
<SUR_TEXT1>页面中提取的图片上方的文本</SUR_TEXT1>
<SUR_TEXT2>页面中提取的图片下方的文本</SUR_TEXT2>
<PAGE_TITLE>图片所在网页的标题</PAGE_TITLE>
<CONTENT_TITLE>图片所在网页的正文标题</ CONTENT_TITLE>
<WIDTH>图片的宽度</WIDTH>
<HEIGHT>图片的高度</HEIGHT>
<ORIGINAL_PIC_NAME>图片在Original_Pic下的文件名</ ORIGINAL_PIC_NAME>
</PIC>

图片原图存储在Original_Pic文件中，每个图片二进制数据保存成一个单独文件，文件名在Meta_Data的元信息中指明。

Evaluation_Data文件包含所有图片的相关元数据，格式如下：
<PIC>
<QUERY_URL>查询图片在互联网中的URL地址</QUERY_URL>
<RESULT_URL>搜索结果的 PIC_URL，多个分号隔开</RESULT_URL>
</PIC>
```

5、Imagenet数据集

业界标杆

Imagenet数据集有1400多万幅图片，涵盖2万多个类别；其中有超过百万的图片有明确的类别标注和图像中物体位置的标注，具体信息如下：

1）Total number of non-empty synsets: 21841

2）Total number of images: 14,197,122

3）Number of images with bounding box annotations: 1,034,908

4）Number of synsets with SIFT features: 1000

5）Number of images with SIFT features: 1.2 million

Imagenet数据集是目前深度学习图像领域应用得非常多的一个领域，关于图像分类、定位、检测等研究工作大多基于此数据集展开。Imagenet数据集文档详细，有专门的团队维护，使用非常方便，在计算机视觉领域研究论文中应用非常广，几乎成为了目前深度学习图像领域算法性能检验的“标准”数据集。

与Imagenet数据集对应的有一个享誉全球的“ImageNet国际计算机视觉挑战赛(ILSVRC)”，以往一般是google、MSRA等大公司夺得冠军，今年（2016）ILSVRC2016中国团队包揽全部项目的冠军。

Imagenet数据集是一个非常优秀的数据集，但是标注难免会有错误，几乎每年都会对错误的数据进行修正或是删除，建议下载最新数据集并关注数据集更新。

数据集大小：~1TB（ILSVRC2016比赛全部数据）

下载地址：

http://www.image-net.org/about-stats

6、COCO数据集

COCO数据集由微软赞助，其对于图像的标注信息不仅有类别、位置信息，还有对图像的语义文本描述，COCO数据集的开源使得近两三年来图像分割语义理解取得了巨大的进展，也几乎成为了图像语义理解算法性能评价的“标准”数据集。

Google开源的开源了图说生成模型show and tell就是在此数据集上测试的，想玩的可以下下来试试哈。

数据集大小：~40GB

下载地址：http://mscoco.org/

COCO(Common Objects in Context)是一个新的图像识别、分割和图像语义数据集，它有如下特点：

1）Object segmentation

2）Recognition in Context

3）Multiple objects per image

4）More than 300,000 images

5）More than 2 Million instances

6）80 object categories

7）5 captions per image

8）Keypoints on 100,000 people

7、PASCAL VOC

PASCAL VOC挑战赛是视觉对象的分类识别和检测的一个基准测试，提供了检测算法和学习性能的标准图像注释数据集和标准的评估系统。PASCAL VOC图片集包括20个目录：人类；动物（鸟、猫、牛、狗、马、羊）；交通工具（飞机、自行车、船、公共汽车、小轿车、摩托车、火车）；室内（瓶子、椅子、餐桌、盆栽植物、沙发、电视）。PASCAL VOC挑战赛在2012年后便不再举办，但其数据集图像质量好，标注完备，非常适合用来测试算法性能。

数据集大小：~2GB

下载地址：

http://host.robots.ox.ac.uk/pascal/VOC/voc2012/index.html

8、Open Image

过去几年机器学习的发展使得计算机视觉有了快速的进步，系统能够自动描述图片，对共享的图片创造自然语言回应。其中大部分的进展都可归因于 ImageNet 、COCO这样的数据集的公开使用。谷歌作为一家伟大的公司，自然也要做出些表示，于是乎就有了Open Image。

Open Image是一个包含~900万张图像URL的数据集，里面的图片通过标签注释被分为6000多类。该数据集中的标签要比ImageNet（1000类）包含更真实生活的实体存在，它足够让我们从头开始训练深度神经网络。

谷歌出品，必属精品！唯一不足的可能就是它只是提供图片URL，使用起来可能不如直接提供图片方便。

此数据集，笔者也未使用过，不过google出的东西质量应该还是有保障的。

数据集大小：~1.5GB（不包括图片）

下载地址：

https://github.com/openimages/dataset

9、Youtube-8M

Youtube-8M为谷歌开源的视频数据集，视频来自youtube，共计8百万个视频，总时长50万小时，4800类。为了保证标签视频数据库的稳定性和质量，谷歌只采用浏览量超过1000的公共视频资源。为了让受计算机资源所限的研究者和学生也可以用上这一数据库，谷歌对视频进行了预处理，并提取了帧级别的特征，提取的特征被压缩到可以放到一个硬盘中（小于1.5T）。

此数据集的下载提供下载脚本，由于国内网络的特殊原因，下载此数据经常断掉，不过还好下载脚本有续传功能，过一会儿重新连接就能再连上。可以写一个脚本检测到下载中断后就sleep一段时间然后再重新请求下载，这样就不用一直守着了。（截至发文，断断续续的下载，笔者表示还没下完呢……）

数据集大小：~1.5TB

下载地址：https://research.google.com/youtube8m/

10、深度学习数据集收集网站

http://deeplearning.net/datasets/**

收集大量的各深度学习相关的数据集，但并不是所有开源的数据集都能在上面找到相关信息。

11、CoPhIR

http://cophir.isti.cnr.it/whatis.html

雅虎发布的超大Flickr数据集，包含1亿多张图片。

12、MirFlickr1M

http://press.liacs.nl/mirflickr/

Flickr数据集中挑选出的100万图像集。

13、SBU captioned photo dataset

http://dsl1.cewit.stonybrook.edu/~vicente/sbucaptions/

Flickr的一个子集，包含100万的图像集。

14、NUS-WIDE

http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm

Flickr中的27万的图像集。

15、MSRA-MM

http://research.microsoft.com/en-us/projects/msrammdata/

包含100万的图像，23000视频；微软亚洲研究院出品，质量应该有保障。

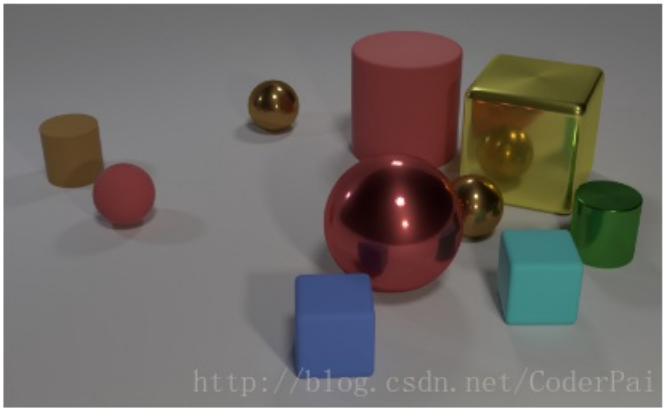
.

16、多物体+关系数据库：HICO & HICO-DET

HICO has images containing multiple objects and these objects have been tagged along with their relationships. The proposed problem is for algorithms to be able to dig out objects in an image and relationship between them after being trained on this dataset. I expect multiple papers to come out of this dataset in future.

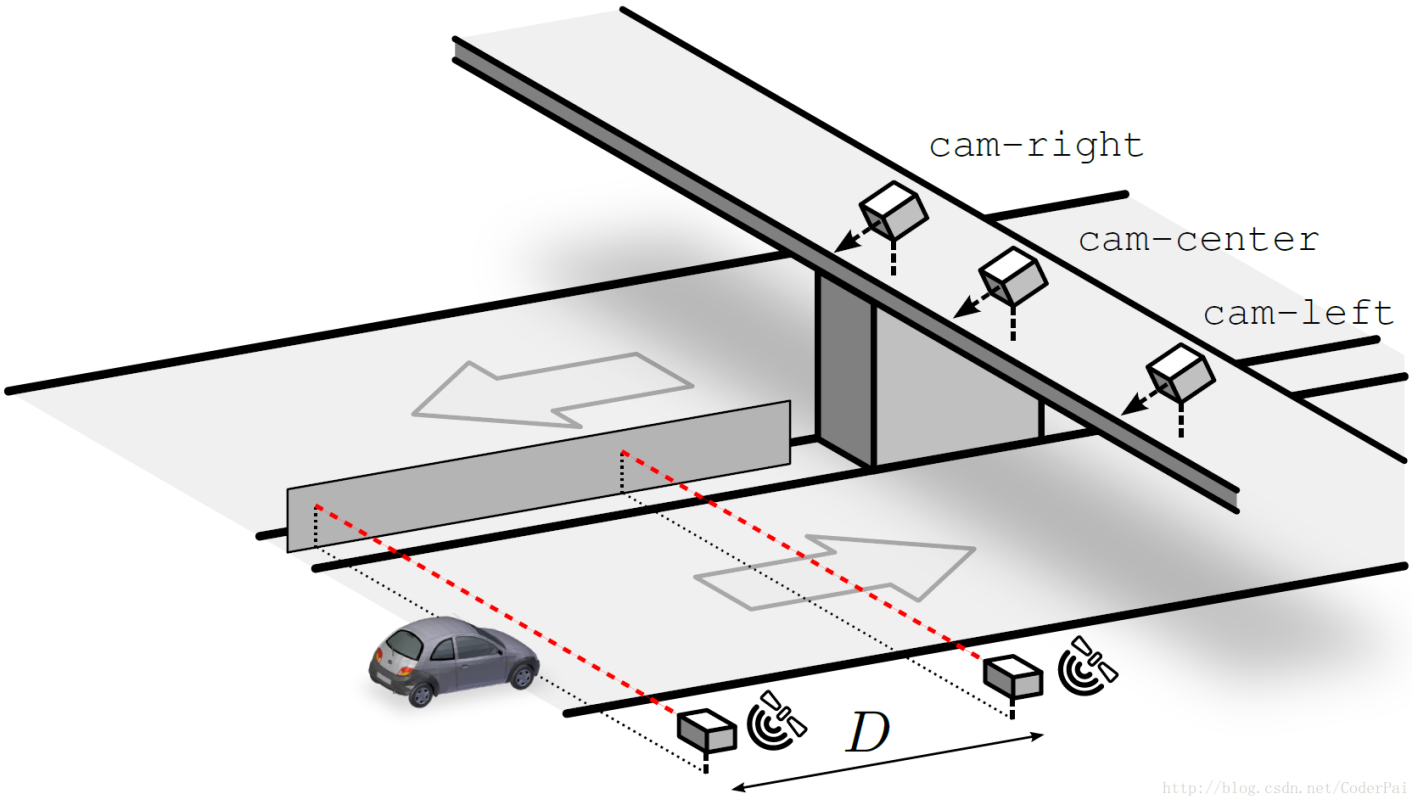
17、QA型图像数据库：CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning

CLEVR is an attempt by Fei-Fei Li's group, the same scientist who developed the revolutionary ImageNet dataset. It has objects and questions asked about those objects along with their answers specified by humans. The aim of the project is to develop machines with common sense about what they see. So for example, the machine should be able to find "an odd one out" in an image automatically. You can download the dataset here.



18、Driver Speed Dataset

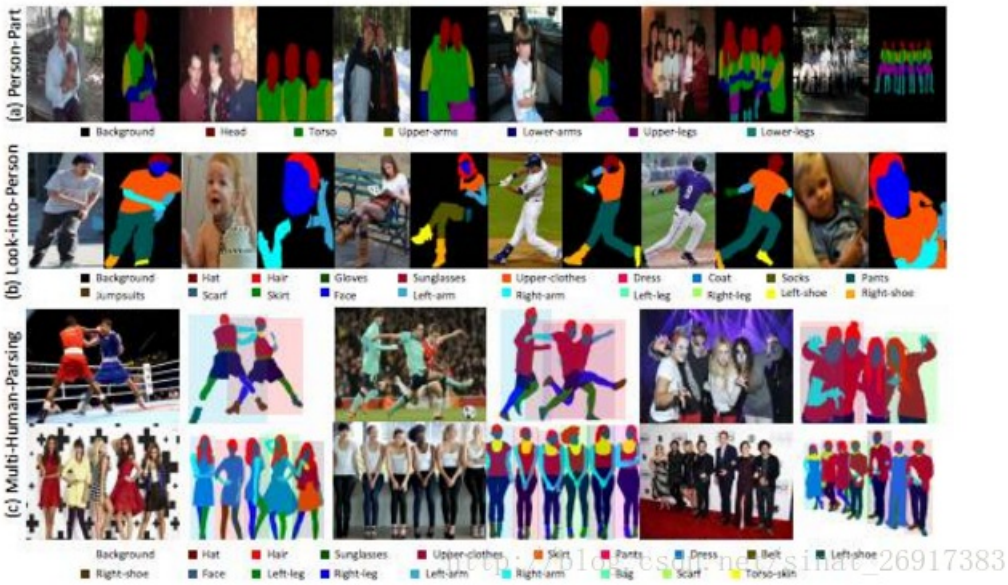
A 200 Gb huge dataset, which is aimed to calculate speed of moving vehicles. Can be downloaded here.



.

19、新加坡国立大学LV实验室发布多人图像解析数据集与模型

为了进一步推进人物解析研究，作者首创多人解析（MHP）数据集，每张图像均包含现实世界场景中的多个人物。 具体而言，MHP数据集的每张图片包含2-16个人物不等，每个人物按照18个语义类别（背景除外）进行像素级别的标注。此外，MHP图像中的人物有多种姿态、不同程度的遮挡以及多样化的交互。为了解决所提出的多人解析这一难题，作者提出了一个新型的多人解析器 (MH-Parser)模型，在针对每个人物进行端到端解析的过程中，同时考虑全局信息与局部信息。实验结果表明，这一模型远优于简单的“检测+解析”方法，使得其作为一个稳定的基准，助推未来在真实场景中人物解析的相关研究。



.

20、300k动作标注视频数据集

DeepMind 最新发布30万 YouTube 视频剪辑的 Kinetics 数据集，包含400类人类动作注释，有助于视频理解机器学习。Kinetics 是一个大规模、高质量的 YouTube 视频URL数据集，包含了各种各样的人类动作标记。我们发布 Kinetics 数据集的目的是助力机器学习社区推进视频理解模型的研究。

Kinetics 数据集包含大约30万个视频剪辑，涵盖400类人类动作，每类动作至少有400个视频剪辑。每个剪辑时长约10秒，并被标记一个动作类别。所有剪辑都经过多轮人工注释，每个剪辑都来自一个单独的 YouTube 视频。这些动作包含了广泛的人类-物体交互的动作，例如演奏乐器，以及人类-人类交互的动作，例如握手和拥抱。

Kinetics 是 ActivityNet 组织的国际人类动作分类竞赛（international human action classification competition）的基础数据集。

官网链接：https://deepmind.com/research/open-source/open-source-datasets/

.

21、MIT 新发布大型数据集 ADE20K：用于场景感知、语义理解等多种任务

每个文件夹包含通过场景范畴进行分类的图像。对于每一张图像，目标和部件分割被存储为两种不同的 png 文件。所有的图像和部件示例都被分别注释。
官方网址：OVERVIEW
Full Dataset，Full-sized images and segmentations

来源机器之心公众号：资源 | MIT 新发布大型数据集 ADE20K：用于场景感知、语义理解等多种任务

22、免费数据集收集网站
各领域公开数据集下载

譬如：

图像数据

综合图像
Visual Genome 图像数据
Visual7w 图像数据
COCO 图像数据
SUFR 图像数据
ILSVRC 2014 训练数据（ImageNet的一部分）
PASCAL Visual Object Classes 2012 图像数据
PASCAL Visual Object Classes 2011 图像数据
PASCAL Visual Object Classes 2010 图像数据
80 Million Tiny Image 图像数据【数据太大仅有介绍】
ImageNet【数据太大仅有介绍】
Google Open Images【数据太大仅有介绍】

场景图像

Street Scences 图像数据

Places2 场景图像数据
UCF Google Street View 图像数据
SUN 场景图像数据
The Celebrity in Places 图像数据
.
23、AVA: 5万+视频/80+动作/21万+标签的视频行为标记数据集

google最新提供了一份5万+视频/80+动作/21万+标签的视频行为标记数据集。

一、视频动作标签类型
stand (45790)
sit (30037)
talk to (e.g., self, a person, a group) (29020)
watch (a person) (25552)
listen to (a person) (21557)
carry/hold (an object) (18381)
walk (12765)
bend/bow (at the waist) (2592)
lie/sleep (1897)
dance (1406)
ride (e.g., a bike, a car, a horse) (1344)
run/jog (1146)
answer phone (1025)
watch (e.g., TV) (993)
grab (a person) (936)
smoke (860)
eat (828)
fight/hit (a person) (707)
sing to (e.g., self, a person, a group) (702)
read (698)
crouch/kneel (678)
touch (an object) (670)
hug (a person) (667)
martial art (624)
open (e.g., a window, a car door) (594)
play musical instrument (545)
give/serve (an object) to (a person) (473)
hand clap (470)
lift/pick up (452)
get up (439)
drink (410)
drive (e.g., a car, a truck) (383)
kiss (a person) (370)
put down (369)
write (340)
close (e.g., a door, a box) (334)
listen (e.g., to music) (290)
catch (an object) (281)
take (an object) from (a person) (257)
hand wave (241)

lift (a person) (201)
pull (an object) (193)
hand shake (179)
jump/leap (151)
dress/put on clothing (130)
push (another person) (122)
text on/look at a cellphone (115)
fall down (114)
throw (99)
sail boat (96)
work on a computer (94)
play with kids (70)
hit (an object) (67)
crawl (61)
enter (58)
take a photo (57)
climb (e.g., a mountain) (57)
push (an object) (56)
play with pets (52)
point to (an object) (45)
cut (43)
shoot (41)
dig (40)
press (38)
play board game (35)
swim (32)
cook (31)
clink glass (30)
fishing (27)
paint (25)
row boat (23)
extract (17)
stir (15)
chop (15)
brush teeth (14)
kick (a person) (13)
kick (an object) (10)
exit (9)
turn (e.g., a screwdriver) (8)

资源地址：https://research.google.com/ava/explore.html
论文地址：https://arxiv.org/abs/1705.08421

七、“稀有”实验室

1、生物识别与安全技术研究中心

CASIA行为分析数据库共有1446条视频数据，是由室外环境下分布在三个不同视角的摄像机拍摄而成，为行为分析提供实验数据。数据分为单人行为和多人交互行为，单人行为包括走、跑、弯腰走、跳、下蹲、晕倒、徘徊和砸车，每类行为有24人参与拍摄，每人4次左右。多人交互行为有抢劫、打斗、尾随、赶上、碰头、会合和超越，每两人1次或2次。

来源：http://www.cbsr.ia.ac.cn/china/Action Databases CH.asp

该实验室拥有的数据库：虹膜数据库，步态数据库，人脸数据库，指纹数据库，掌纹数据库，笔迹数据库，行为分析数据库
该实验室研究成果：
近红外的人脸身份识别技术和系统， 中远距离人脸识别系统， 人脸检测与跟踪， 多目标遮挡跟踪， 目标检测、跟踪与分类，异常动作检测， 人异常行为检测与报警， 交通车辆计数演示， 主从摄像机跟踪，
多摄像机数据融合（全景监控地图）， 交通拥堵检测与报警， 车辆异常行为检测与报警， 夜间跟踪演示， 动态场景下的主动跟踪， 视频图像序列拼接， 人数统计， 视频浓缩

2、中文语言资源联盟
中文语言资源联盟，英文译名Chinese Linguistic Data Consortium，缩写为CLDC。 CLDC是由中国中文信息学会语言资源建设和管理工作委员会发起，由中文语言（包括文本、语音、文字等）资源建设和管理领域的科技工作者自愿组成的学术性、公益性、非盈利性的社会团体，其宗旨是团结中文语言资源建设领域的广大科技工作者，建成代表中文信息处理国际水平的、通用的中文语言语音资源库。

当然，里面的内容都是收费的，而且不便宜，不过毕竟是好东西~

3、中科院自动化研究所 模式识别国家重点实验室

4、北邮模式识别实验室
http://www.pris.net.cn/

图像识别方向的技术有：
高清车牌及车标识别技术、不良图片过滤、图片检索技术

5、中国科学技术大学，图像处理实验室
http://image.ustc.edu.cn/project.html
国家自然基金重点项目：高分辨率SAR图像目标认知模型及高效算法
国家自然科学基金项目：星上原始超光谱图像稀疏编码压缩技术研究
973课题：稀疏微波成像数据压缩及特征理解
果然有钱！！

6、国内高校开源镜像站友情链接
清华大学开源镜像站
http://mirror.tuna.tsinghua.edu.cn/
中国科学技术大学开源镜像站
http://mirrors.ustc.edu.cn
北京交通大学开源镜像站
http://mirror.bjtu.edu.cn/cn/
兰州大学开源镜像站
http://mirror.lzu.edu.cn/
厦门大学开源镜像站
http://mirrors.xmu.edu.cn/
上海交通大学开源镜像站
http://ftp.sjtu.edu.cn/
东软信息学院开源镜像站
http://mirrors.neusoft.edu.cn/

7、网页版呈现各类模型的实现

https://github.com/hwalsuklee/awesome-deep-vision-web-demo

.

八、中文文本语料库

可参考：【语料库】语料库资源汇总
NLP常用信息资源：https://github.com/memect/hao/blob/master/awesome/nlp.md
FudanNLP (FNLP)（FNLP主要是为中文自然语言处理而开发的工具包，也包含为实现这些任务的机器学习算法和数据集。）：https://github.com/FudanNLP/fnlp

(一) 国家语委

1国家语委现代汉语语料库http://www.cncorpus.org/

现代汉语通用平衡语料库现在重新开放网络查询了。重开后的在线检索速度更快，功能更强，同时提供检索结果下载。现代汉语语料库在线提供免费检索的语料约2000万字，为分词和词性标注语料。

2古代汉语语料库http://www.cncorpus.org/login.aspx

网站现在还增加了一亿字的古代汉语生语料，研究古代汉语的也可以去查询和下载。同时，还提供了分词、词性标注软件、词频统计、字频统计软件，基于国家语委语料库的字频词频统计结果和发布的词表等，以供学习研究语言文字的老师同学使用。

(二) 北京大学计算语言学研究所

1《人民日报》标注语料库http://www.icl.pku.edu.cn/icl_res/

《人民日报》标注语料库中一半的语料(1998年上半年)共1300万字已经通过《人民日报》新闻信息中心公开提供许可使用权。其中一个月的语料(1998年1月)近200万字在互联网上公布，供自由下载。

(三) 北京语言大学

汉语国际教育技术研发中心：HSK动态作文语料库http://202.112.195.192:8060/hsk/login.asp

语言研究所：北京口语语料查询系统（B J K Y）http://www.blcu.edu.cn/yys/6_beijing/6_beijing_chaxun.asp

还有很多，可参考：【语料库】语料库资源汇总

百度开源的中文问答语料：

WebQA: A Chinese Open-Domain Factoid Question Answering Dataset
发布的文件有267MB，但对于我们来说，里边的东西貌似有点过多了，因为里边包含了分词结果、序列标注结果、词向量结果，貌似是内部研究小组直接用来做的实验。对于我们来说，显然只需要纯粹的问答语料就行了。
相关介绍可见： 百度的中文问答数据集WebQA

公开语料：
搜集到的一些数据集如下，点击链接可以进入原始地址

dgk_shooter_min.conv.zip
中文电影对白语料，噪音比较大，许多对白问答关系没有对应好

The NUS SMS Corpus
包含中文和英文短信息语料，据说是世界最大公开的短消息语料

ChatterBot中文基本聊天语料
ChatterBot聊天引擎提供的一点基本中文聊天语料，量很少，但质量比较高

Datasets for Natural Language Processing
这是他人收集的自然语言处理相关数据集，主要包含Question Answering，Dialogue Systems， Goal-Oriented Dialogue Systems三部分，都是英文文本。可以使用机器翻译为中文，供中文对话使用

小黄鸡
据传这就是小黄鸡的语料：xiaohuangji50w_fenciA.conv.zip （已分词） 和 xiaohuangji50w_nofenci.conv.zip （未分词）

保险行业语料库
数据集分为两个部分“问答语料”和“问答对语料”。问答语料是从原始英文数据翻译过来，未经其他处理的。问答对语料是基于问答语料，又做了分词和去标去停，添加label。所以，"问答对语料"可以直接对接机器学习任务。如果对于数据格式不满意或者对分词效果不满意，可以直接对"问答语料"使用其他方法进行处理，获得可以用于训练模型的数据。

科学空间：【语料】百度的中文问答数据集WebQA
搜狗：中文词语搭配库(SogouR)
wiki_word2vec：https://github.com/wac81/wiki_word2vec

微博终结者爬虫

这个项目致力于对抗微博的反爬虫机制，集合众人的力量把微博成千上万的微博评论语料爬取下来并制作成一个开源的高质量中文对话语料，推动中文对话系统的研发。
github：https://github.com/jinfagang/weibo_terminater