# 步骤

## 目标

我们把一个特定场景下的经验和规则，称为专业知识。制作一个实用的图像数据集，即是将特定场景下的知识与该场景下的图像做关联。如同制作一个工具，制作人员事先对于工具的典型使用场景必须有所了解，设计上有对该场景的考虑。如果缺乏特定场景的经验，数据集就无法指导实践。

要评价模型在实际场景中的能力，需要使用特定场景的数据和知识。

在制作数据集的过程中，“用途”作为制作者做取舍的依据，其作用会体现在各个环节、不同层面上。

比如标注“杯子”，杯子形态各异、难以穷举，甚至聚拢手掌也可以是杯子：人是通过“盛水”这一功能去认识杯子的，而不是具体形象——“用途”先于“表相”；而同一个杯子，也可以有不同用途，在使用者眼中有不同的理解方式。 因此标注再多图像，识别效果也未必好。称“用途”，是从工具角度来看；从使用者的角度来看，则称为“任务”。人总是在一定任务背景下去理解事物、操作工具。用途和任务，属于人的认知领域。

## 确认知识点

## 收集

在上一步“确认知识点”的过程中，会先采集少量图像；当知识点确认后，就进入大规模的图像采集。由算法人员采集到的大量图像，将用于第三步的标注。标注好的图像将用于模型训练和评测。要让模型达到识别效果，对每个标签都需要一个最少的训练样本量，例如2000张，这个量同任务和数据都有关系，可以由经验或实验来确定。第二步图像采集的主要挑战是：样本稀缺，即某个标签下的图像过少。

如果始终采不到足够多的样本，可以考虑知识点合并或抛弃。例如，淘宝后台的风格标签中曾有“宫廷风”一项，在实际中“宫廷风”的衣服极少；又如在设计师语言中，袖型有“郁金香袖”一项，实际商品过少，对这些情况我们都做了抛弃。这即是对知识体系的略微修正。

如果某标签很重要，但专家认为不能抛弃，比如某些前瞻性的设计要素，我们还可以为该标签做“悬赏”：在众包平台发布付费任务，由大众来收集图像。使用众包平台来完成数据标注和采集任务，近年来发展迅速，已逐步进入实用。

## 标记

精度和数量的取舍

标记的过程是分类的过程

将采集来的大量图像与知识做关联，就是图像数据标注。知识和数据的关系是一体两面：知识是数据的抽象，数据是知识的载体。这个关系反映在数据集制作过程中，就是：知识会指导数据的采集和标注，而在数据的采集和标注过程中，知识又会被修正，这是一个彼此影响、反复迭代的过程。这个过程按次序可以大致分为四个步骤：

A. 算法人员和专业人士探讨学习，做知识的转译和重整。

B. 算法人员根据知识点采集图像。

C. 标注人员学习标注规则，对图像做标注。

D. 将标注好的图像输入机器，做训练和评测。

标注人员学习规则，对采集好的数据进行标注。如果资源有限，算法人员也往往就是标注人员。

这一步主要要考虑标注人员的学习成本和标注效率。通过标注人员的反馈，算法人员一方面改进规则、补充图例，对标注人员反复出现疑问的地方，考虑知识修正；另一方面，改进标注工具，包括流程、交互、预处理等，以提高标注效率。

### 工具

https://blog.csdn.net/chaipp0607/article/details/79036312

对于监督学习算法而言，数据决定了任务的上限，而算法只是在不断逼近这个上限。世界上最遥远的距离就是我们用同一个模型，但是却有不同的任务。但是数据标注是个耗时耗力的工作，下面介绍几个图像标注工具：

Labelme

Labelme适用于图像分割任务的数据集制作：

它来自下面的项目：https://github.com/wkentaro/labelme

该软件实现了最基本的分割数据标注工作，在save后将保持Object的一些信息到一个json文件中，如下：

https://github.com/wkentaro/labelme/blob/master/static/apc2016\_obj3.json

同时该软件提供了将json文件转化为labelimage的功能：

labelImg

Labelme适用于图像检测任务的数据集制作：

它来自下面的项目：https://github.com/tzutalin/labelImg

其中标签存储功能和“Next Image”、“Prev Image”的设计使用起来比较方便。

该软件最后保存的xml文件格式和ImageNet数据集是一样的。

yolo\_mark

yolo\_mark适用于图像检测任务的数据集制作：

它来自于下面的项目：https://github.com/AlexeyAB/Yolo\_mark

它是yolo2的团队开源的一个图像标注工具，为了方便其他人使用yolo2训练自己的任务模型。在linux和win下都可运行，依赖opencv库。

LabelBox：对于大型标记项目很合适，提供不同类型标记任务的选项。

COCO UI：用于标注 COCO 数据集的工具。

Vatic

Vatic适用于图像检测任务的数据集制作：

它来自下面的项目：http://carlvondrick.com/vatic/

比较特别的是，它可以做视频的标注，比如一个25fps的视频，只需要隔100帧左右手动标注一下物体的位置，最后在整个视频中就能有比较好的效果。这依赖于软件集成的opencv的追踪算法。

Sloth

Sloth适用于图像检测任务的数据集制作：

它来自下面的项目：

https://github.com/cvhciKIT/sloth

https://cvhci.anthropomatik.kit.edu/~baeuml/projects/a-universal-labeling-tool-for-computer-vision-sloth/

在标注label的时候，该软件可以存储标签，并呈现标注过的图片中的bbox列表。

Annotorious

Annotorious适用于图像检测任务的数据集制作：

它来自下面的项目：

http://annotorious.github.io/index.html

代码写的相当规范，提供了相应的API接口，方便直接修改和调用。

RectLabel

RectLabel适用于图像检测任务的数据集制作：

它来自下面的项目：

https://rectlabel.com/

这是一个只适用于Mac OS X的软件，而且可以在apple app store中直接下载。

VoTT

VoTT适用于图像检测任务的数据集制作：

它来自下面的项目：

https://github.com/Microsoft/VoTT/

微软的开源工具，既可以标注视频，也可以标注图片，而且支持已有模型的集成，功能强大。

IAT – Image Annotation Tool

IAT适用于图像分割任务的数据集制作：

它来自下面的项目：

http://www.ivl.disco.unimib.it/activities/imgann/

比较有特色的是，它支持一些基础形状的选择，比如要分割的物体是个圆形的，那么分割时可以直接选择圆形，而不是用多边形选点。

images\_annotation\_programme

images\_annotation\_programme适用于图像检测任务的数据集制作：

它来自下面的项目：

https://github.com/frederictost/images\_annotation\_programme

网页版的哦

除此之外，还有很多类似的工具，与上面的工具相比，并没有什么特色了，我们只给出链接，不详细介绍了：

ImageNet-Utils

https://github.com/tzutalin/ImageNet\_Utils

labeld

https://github.com/sweppner/labeld

VIA

http://www.robots.ox.ac.uk/~vgg/software/via/

ALT

https://alpslabel.wordpress.com/2017/01/26/alt/

FastAnnotationTool

https://github.com/christopher5106/FastAnnotationTool

LERA

https://lear.inrialpes.fr/people/klaeser/software\_image\_annotation

## 随机切分

## 确定性能指标

# 工具

## OS库

## PIL（Python Imaging Library）

是 Python 中最常用的图像处理库，而Image类又是 PIL库中一个非常重要的类，通过这个类来创建实例可以直接载入图像文件，读取处理过的图像和通过抓取的方法得到的图像这三种方法。

# 代码实现

## TFRecords数据集制作

将数据集放在文件夹里读取过于分散,难免导致内存利用率低

TFRecords其实是一种二进制文件，虽然它不如其他格式好理解，但是它能更好的利用内存，更方便复制和移动，并且不需要单独的标签文件(label)。