如何做一个“实用”的图像数据集

原创 2018-04-24 贾梦雷 et al. 视觉求索

目录

编者序

大家都说这是一个大数据的年代，人工智能的落地需要数据， 深度学习更需要海量数据。于是， 出现了一个流行的口号： “数据就是新的石油 “（Data is the new oil）。这个比喻很形象，但容易把问题简单化。首先，对于人工智能的应用来说，数据顶多只是原油（ crude oil），就是那种黑糊糊的液体， 要变成可以用的汽油， 还需要复杂的取舍和提炼过程。其次，汽油对各种汽车是通用的，而人工智能的需求非常广泛， 任务各异，往往根据不同任务，要精炼不同的汽油。更复杂的是，对于一个行业或者产品来说， 它的任务定义往往是模糊的。那么“数据”和“任务”需要一个长期的迭代过程，这个过程成本是相当高的。

本文作者贾梦雷是中科大毕业的，其带领的阿里巴巴“图像和美”团队开发计算机视觉在时尚领域的应用，花了7年时间迭代数据与任务。非常感谢他与我们分享他们在第一线“炼油”的心路历程， 以及各种洞见。

引言

近年来AI受到各界关注，公司、政府及民众对于AI落地都充满期待。在媒体的描述中，各种AI落地的场景呼之欲出。不过在我们看来，目前AI算法在很多数据集上的成功多是学术意义上的，距离商业落地还有一段较长的路要走。如今众多科研人员从学术界走向工业界，大量在校学生投入AI领域。当前正是时候和大家探讨AI落地中的数据挑战。本文的目的在于分享经验与同行探讨。

我们在阿里巴巴图像和美团队探索将AI用于时尚领域，已经有七个年头。我们希望开发的AI产品对衣服的理解不只限于照片和文字，而是可以理解衣服本身，进而理解时尚穿搭之道、理解流行风向；我们希望这样的理解可以作用在阿里巴巴数以亿计的商品上，从而影响大众、改变行业。

让AI懂得时尚，且不说商业落地，仅从技术上听起来，就有点天方夜谭：时尚是如此主观，人都很难理解，何况机器？其实，做时尚AI的魅力也就在此——“如何客观地看待主观世界”——需要我们把严谨的科研态度和行业洞察力、想象力结合在一起，才能为机器打造一颗“时尚之心”。

让机器理解衣服，核心是制作服饰图像数据集。我们在本文分享“时尚之心”项目中最基础也最有挑战的部分：如何制作一个“实用”的图像数据集？这里的“实用”指的是能够达到商业落地的程度。构建一个图像数据集，即是在一定的“用途”目的下，将“知识”与“图像”做关联，并给出评价算法的“指标”。

文章结构也是按照以上四个关键词来组织的：文章第一部分是对“用途”的探讨，第二、三、四部分围绕着“知识”、“数据”和“指标”来展开，最后是总结。

一、探究数据的“用途”

十几年前我读研究生时，方向是机器人。有朋友问起：“你做的机器人是干什么用的？”我一时语塞，还有点气愤。那时我做机器人，硬件从零做起，用于研究探路算法、发表学术论文。零基础、缺经费、加上学生的目标是纯粹做研究，我压根没想过自己的机器人真能派什么用。做学问嘛，怎么能图“有用”呢？

一方面觉得被拷问“用途”是受辱，一方面又觉得朋友问得对。后来我常拿“做什么用”来问自己，提醒自己这个世界另有期待。毕业后我从事计算机视觉的工作。做图像算法比研究机器人探路更接近现实应用。要让算法走向实用，首先要让数据集走向实用。图像数据集在计算机视觉研究中的的作用，好比实验对象在科研工作中的作用。实验对象的采制是否严谨合理、距离实际有多远，直接决定了科研成果是否可靠、是否能用于实际。可以说，实验对象在相当程度上决定了科研活动的水平。

计算机视觉发展的时间还不长，人们像呵护孩子一样，鼓励新想法、包容不完美。过去学术界对图像数据集的要求实际是比较低的，数据量大一些大家就满意了。如果按一个成熟的科研方向来要求的话，过去二十年业界所出现的数据集，远不能让人满意。绝大部分数据集，内在结构松散，外在用途不明，距离指导算法落地还比较远。

在过去的几年，深度学习的兴起使得计算机视觉的工具有了长足进步。随着媒体热炒、资本涌入、政府重视，人们对于AI落地有了热切的期望。AI算法要落地实用，首先是要数据集能达到落地实用。目前学术界的论文和竞赛所依赖的数据集，距离其所宣称的作用和意义相去甚远。这点也是业界心照不宣的共识。

中国古代用“性、相、用”来分析一个事物，即通过“性质、显现、用途”来认识一个事物。套用在数据集上：“性”是制作数据集的方法和原则，“相”是数据集的具体内容，“用”是数据集的用途。

在过去，论文往往着重介绍数据集的“性”和“相”，即制作方法和具体内容，而对数据集的“用”描述过于简略。也难怪，过去的数据集基本用来验证方法本身（如分类方法、检测方法），是从学者的视角出发，而不是从实际问题出发。业界衡量一个数据集是否成功，往往只用被引用次数、影响力大小，而忽略数据集的内在逻辑结构和外在实用价值，有点像自说自话。

我们关心AI算法的落地，就必须关心数据集的用途。图像中的内容，可分为两大类：“自然的”和“人造的”。自然的如风景、动物，人造的如汽车、文字。内容为自然事物的图像，例如人脸照片，是证件照还是监控摄像头拍照，差异巨大，这是由其使用场景——用途——直接决定的。而对于人造事物，“用途”的重要性更甚：事物的形态往往是其功能的体现，人们是通过“用途”去认识这类东西的。

2017年我去UCLA拜访朱松纯老师时，聊起当年莲花山项目在图像标注上遇到的困难。朱老师举例说，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，这启发他，要解决视觉问题，先要去研究综合各种感官、心理、记忆在一起的认知问题。

可见，强调“用”不仅是出于实用价值，也是加深对研究对象的理解的内在需要。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。接下来我们首先看到的，就是重视“用途”对于数据集中“知识”的影响。

二、梳理专业的“知识”

2.1忽视专业知识，无法做出有用的数据集

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。

例如，LFW Face Database[1]是一个知名的数据集，包括13000张标注好的人脸图像。作者的目的是制作一个“非限制条件下”的数据集，用来评价模型的人脸识别能力。实际上，这批图像主要是采集自网络的欧美名人的正面照片，与摄像头监控、证件识别等实际场景中的照片相去甚远。很多技术团队在此数据集上做激烈的竞争，但这些数字指标对于揭示他们的模型是否能在实际场景中发挥作用，并无太大的参考价值。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。

有的制作者虽然使用了特定领域的数据，但缺乏专业人士的指导，只是沿用学术界惯有的方法，想当然的把一些专有名词与图像做了关联。这样制作出的数据集可能与实际情况有很大偏差。

例如，ChestX-ray8[2]是2017年发布的一个胸部X光数据集。制作者使用自然语言处理的一些手段对X光图像的报告单进行了文本挖掘，得到一系列疾病标签，把这些标签和对应的图像关联起来。专业人士LukeOakden-Rayner医生[3]指出：部分疾病标签并非通过观察医疗影像得出的，而是结合其他诊断信息综合得出的……实际上（报告单的内容是），观察影像的医生在通过影像回答另一位医生的问题，对同一张图像的不同提问，可能有不同、甚至相反的回答。因此，疾病标签和图像的关联很可能不符合实际情况。当然，数据集的制作者也充分认识到了这个局限性。他们基于900张报告单做了一个专家对比实验，实验表明文本挖掘得到的疾病标签准确率远未达到100%。

再举一例。DeepFashion[4]是2016年发布的一个服饰图像数据集，包括了超过80万张时装照片，被归到50个类别里。这50个类别标签来自制作者从两个服饰网站的查询词中抽取的名词，这些标签被声明是互斥的，但实际情况并非如此。例如毛衣（Sweater）和龟领（Turtleneck）这两个标签，毛衣属于“材质”的范畴，而龟领属于“领子设计”的范畴，两个标签在概念上并非平行对等，不能并列作为服饰的两个类别。如图1，“龟领”类别的衣服，同时也是“毛衣”。这类错误在DeepFashion数据集中并不少见。

图1.DeepFashion中的“龟领”标签下服装

显而易见，如果用于指导标注的知识没有被很好的梳理，那么数据集必然质量不佳，很难期望能产出好的模型；即使模型在评测中表现良好，在实际中使用也会很糟糕。

2.2 原有知识体系往往有局限

即使能获取到专门的数据，有专业人士的帮助，数据集制作者在“知识构建”上仍需付出巨大的努力。这是因为，知识的“用途”发生了变化。

一个领域的专业知识，原本只在该领域的专业人士之间流通，是为了人和人的沟通的；而制作数据集的目的是把人的经验传递给机器。直接把原有知识体系照搬到机器学习中来，往往行不通，主要问题就是“不完备”和“二义性”问题。

这里的“完备”，指的是上层概念所覆盖的范围，要能被下层概念完全覆盖。例如“中国人”可以被“南方人”和“北方人”覆盖。如果无法完全覆盖则称为“不完备”。这里的“二义性”，指的是同层级的两个概念，覆盖的范围有一定的重合，例如会有一部分“中国人”归属到“南方人”或“北方人”都说得通，是模棱两可的。专业知识往往来自于人的日常经验，天然具有一定的不完备和二义性。例如医疗影像中的疾病种类，是无法完全枚举的。

人去处理沟通中的“不完备”和“二义性”，问题不大，因为人既有生活常识、也经过一定的背景训练，可以根据经验来纠错。而机器不行，机器就如同白纸一张的婴幼儿，接收到的往往只有标注数据，告诉他什么就是什么，辨别能力或者说容错能力非常低。当然我们可以像训练婴幼儿一样，用多种数据训练一个有一定推理纠错能力的模型，这属于探索性的尝试，对于当下绝大部分的商业应用来说并不现实。因此，有必要对原有知识体系做出修正，减少不完备性和二义性，以适应机器当下的学习能力。

若观察原有知识体系，可以看出它们多是从一个个实例出发、自底向上构建的，因此难以避免不完备和二义性的缺陷。如果换个方式，一开始就注意避免不完备和二义性，自顶向下构建知识体系，这可行吗？答案是否定的。我们虽然可以从逻辑上规避二义性，但实际层面的二义性一点不少。例如，我们把“商品图”分为“模特图”和“非模特图”，简单理解，就是有人的图，和没有人只有商品的图。这在逻辑上没有二义性，但实际情况例如图2，图中是穿在脚上的一双鞋，这算是模特图还是非模特图呢？

图2.穿在脚上的一双鞋

遇到这种情况，需要拆分概念，将“模特图”拆分为“手模图、腿模图、假模图、半身模特图、全身模特图”等等，如图3所示，而拆分又必然导致不完备：这些分类能穷尽模特图的所有情况吗？显然不能。但为了在实际中让标注人员容易理解，又不得不拆分。我们只能根据实际情况，做一个权衡。

图3. 依次为：平铺图、假模图、手模图、腿模图、全身模特图

对于“不完备”，还有一类普遍情况，值得重视。如图4所示，我们要标注裙长，但图中无法展现裙子的全长，这属于“无法判别”的情况。

图4. 裙子被截断的图片

学术界的通常做法是抛弃这类样本，数据集里只保留可以清晰判别的样本。这种减少二义性的措施可以理解，不过，在实用中，这种例子是不能抛弃的。因为在实际中模型必然会碰到这种不可判别情况，没有人会为你挡驾，模型需要具备对这类情况“说不”的能力，准确的说，是打上“无法判别”标签。因此，制作数据集时，我们要保留这部分数据、设立为“无法判别”的分类，这个措施也可看作是为了知识的“完备性”而做的努力。

以上还是专业知识来自单一人群的情况，如果是多类人群对于同一知识点有不同理解，就更复杂了，需要做跨角色的知识重建。

2.3跨越多种角色的知识重建

在实际中，“专业人士”可能并非单一的人群，而是在一件事的不同环节上的多种角色，他们视角不同，使用的知识体系也不同。

例如，服装的“颜色”属性，就有“计算机的颜色空间”、“潘通色卡”、“服装营销色彩”等不同知识体系。在计算机的颜色空间中，一个点可以代表一种颜色，如在“RGB空间”中，一个（R,G,B)三元组就对应一种颜色，这可被计算机理解，但无法用于日常沟通。“潘通色卡”是国际上通用的纺织、印刷、塑胶、绘图等领域的色彩标准语言，其中国际纺织服装的版本包括2310种颜色。这么细的划分，很难被消费者理解，服装商通常会建立一套大众可以理解的颜色标签，即“服装营销色彩”，粗分有8到10种色系，细分有上百种颜色。以“红色”为例，如图5所示。要将算法模型付诸实用，我们就要打通这三套颜色体系、做知识连接。

颜色	RGB	潘通色	营销色	色系
	223,72,79	16-1632, Spiced Coral	石榴花红	红
	204,51,61	16-1626, High Risk Red		
	209,59,64	16-1624, Flame Scarlet		
	195,78,124	15-1922, Fuchsia Rose	桃红色	
	190,76,124	15-1920, Lilac Rose		

图5. 红色在不同颜色体系中的值

以上是简单的例子，更复杂的例子是服饰的“风格”属性。在电商服饰的生产流通中，有“生产商”、“电商平台”、“零售商”三个角色，三者各自有一套“女装风格”体系：

- 生产商的风格体系，是设计师和服饰企划人员用的，用的是工业设计语言，这个体系相对完备和稳定。
- 电商平台的风格体系，是平台运营人员用的，用的是运营语言，用于管理货品、组织卖场，体系的稳定性介于生产商和零售商之间，是二者的桥梁。
- 零售商面对消费者，用的是营销语言，风格体系要根据时尚趋势和消费热点而变化，特点是灵活发散，易于消费者理解和联想。

我们可以通过图6获取三者的直观印象。

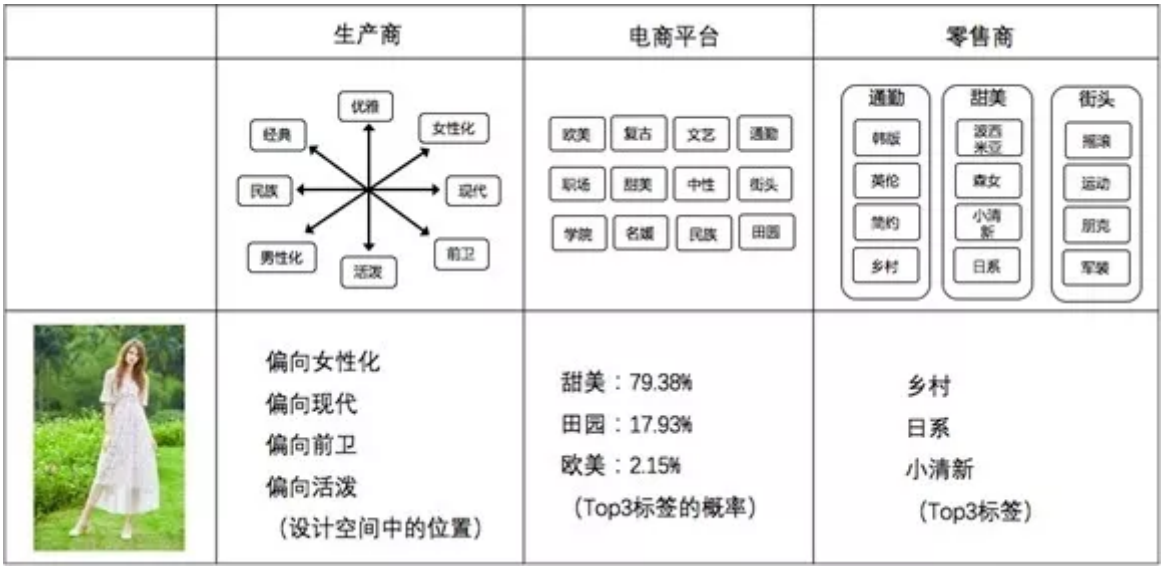

图6. 三种角色各自的风格体系

生产商的风格体系，有四个相对独立的维度，每个维度用代表该维度两极的词汇来表示，例如“男性化，女性化”是其中一个维度，其他三个维度是“经典，前卫”、“民族，现代”、“活泼，优雅”，这构成一个四维的设计空间，在服装企划人员的眼里，每件衣服都对应着这个空间中的一个位置。例如图6中的连衣裙将落在“女性化、现代、前卫、活泼”这个象限里，服装企划人员会给出一个具体的位置。

而在电商平台风格体系，是由“欧美”、“复古”等12种标签组成的一个平铺结构，我们依此训练了模型，来给任意一件女装打标。当一件衣服来的时候，模型会判断衣服与这些标签的匹配程度，按概率大小取前三名（一件衣服可能兼容多种风格）。例如图6中的连衣裙被判为“甜美”的概率为79.38%，其次是“田园”和“欧美”。

零售商的风格体系是在日积月累中形成的，主要出于营销目的。假设最近市场上流行“波西米亚”风，营销人员就找一部分有相似理念的衣服，打上“波西米亚”的标签；过一段“小清新”风盛行，就给相似理念的衣服打上“小清新”标签。这是一种打补丁的标签积累方式，不太注重标签之间的内在逻辑。例如图6中的连衣裙，按营销知识体系会有“乡村”、“日系”、“小清新”等标签。

做一个实用的风格体系，意味着要能贯穿从生产到营销的全过程，我们要建立一个“层次式”的知识体系，如图7所示：以稳定的工业设计知识为基底、平台运营知识为桥梁、大众营销知识为上层，用技术将其打通，使得上层的任一概念可以在下层有投影。这样的体系，对下连接到

海量商品，对上承接消费者需求，把以前分散在各个环节的、主观的、零散的行为，转化为一件系统性的工作。这是理想化的结构，我们也在探索中。

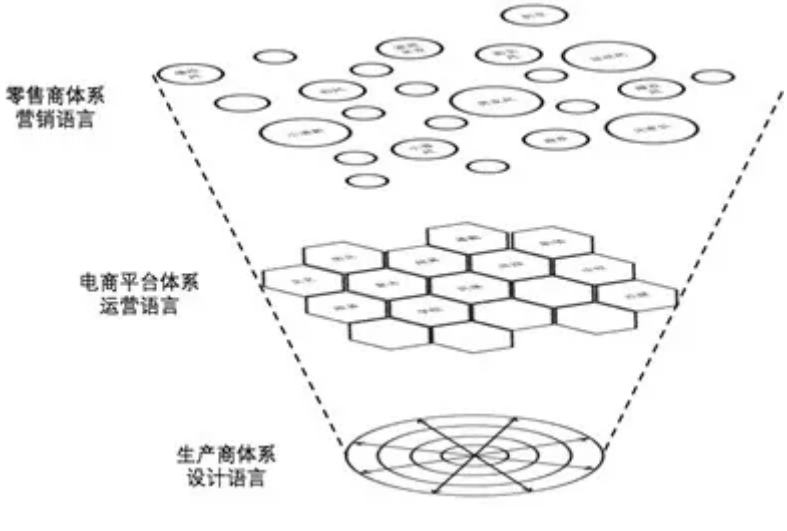

图7. 层次式的风格体系

综上所述，要制作一个实用的数据集，需要制作者在专业人士的指导下做知识重建，这是一个需要制作者亲力亲为、勇于付出的过程，难以讨巧、无法回避。从更大的视角来看，AI落地的过程，实际是一个促使生产的各个环节加强沟通、重建知识体系的过程，是知识从各自的孤岛走向整体的过程。这个过程需要所有环节的人一起努力，而当中AI从业者应肩负起主要责任。

三、数据与知识“迭代”

3.1 数据制作的流程

如上一章介绍，知识需要被重建，因为知识的用途发生了变化：从专业人士之间的沟通到人和机器的沟通。同时，知识的载体——数据——也发生了变化：从日常经验的数据到有组织采集的大量数据。例如，以前服装陈列师对风格的认识，是经年累月从门店、杂志的商品中得来的，而今天算法人员会通过搜索引擎定向收集成千上万的相关图片，这是服装陈列师所没见到过的。

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。

下面按顺序介绍每个步骤中的挑战和应对，我们将看到数据和知识是如何反复迭代的。

3.2 第一步：知识的转译和重整

首先，算法人员要消化专业知识，在专业人士指导下整理出可以标注的规则和图例解释。在这个阶段，挑战主要是：如何对知识点做取舍。

以“领型”为例，圆领的“颈线设计”分为四类，如图8所示。

图8. 四类领线设计

在专业人士眼中，这四类颈线区分很大，但是对算法人员以及标注人员（有时这两种角色是同一个人）来说，很难把握其间的差别。在实际图片中，衣服颈线的圆弧形依照深度和宽度的不同有各种形态，我们看过大量图片也很难选出符合标准定义的样本。考虑到这四类颈线设计对于衣服的整体设计风格影响不大，我们合并这四类颈线为“圆领”。

又如女装的“西装领”，如图9。

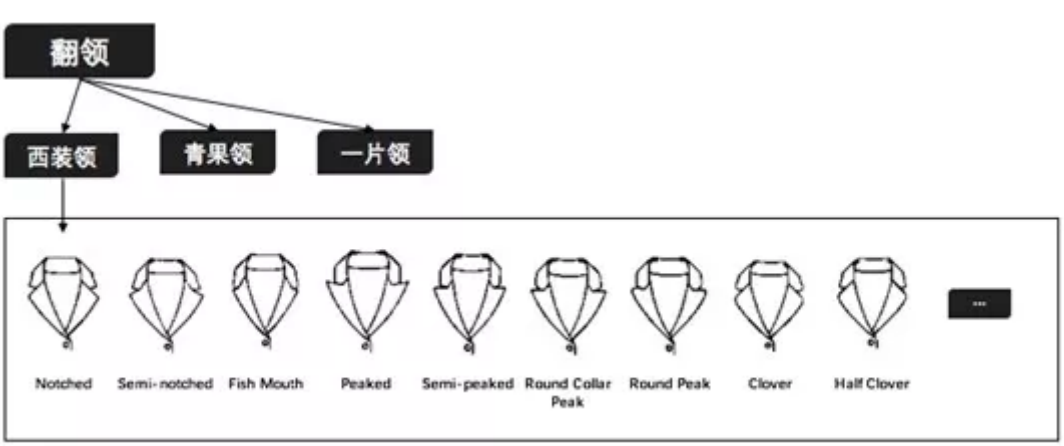

图9. 西装领的分类

“西装领”的子类从视觉上难以区分，标注人员即使努力学习，标注准确率也达不到50%。同时9个子类也意味着，投给机器的训练样本量要增加9倍，一方面是标注成本会高很多，另一方面知识粒度过细还会导致采集不到足够多的图像样本。由于女装样式丰富，视觉刺激点较多，“西装领”子类的差别不太影响整体效果，我们取消子类的划分，都归到“西装领”。

3.3 第二步：根据知识点来收集图像

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。

图10是我们的采集流程。以“深V领”为例，我们用初始查询词“深V领”搜索图片，再由人工筛选出符合标签描述的图像，即“深V领的衣服图片”。如果经人工筛选后，样本充足，就完成“深V领”的采样。如果样本不足，就使用同义词、近义词如“低V领”、“大V领”、“鸡心领”，继续搜索，直到样本充足，或者始终仍无法获取足够多的样本。

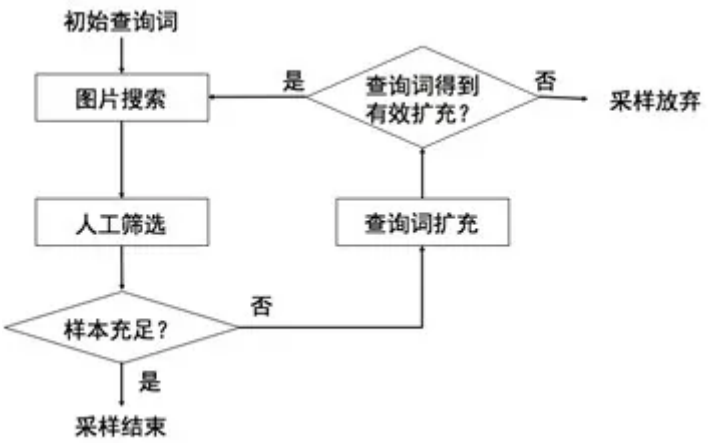

图10. 数据采集流程

在无法获取足够多样本的情况中，有一类是由于标签用语过于专业，不会出现在在图片的日常描述中，对这种情况我们使用“类似描述”来扩充查询词。例如“鼓肩袖”是一个专业术语，在图片描述中很少见，我们会用“肩部折叠”、“袖子褶皱”、“肩部褶皱”、“肩部蓬松”等来发起查询。

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。

当引入各种手段来获取数据时，要警惕一种情况：结构化噪声。

什么是“结构化噪声”呢？要从数据采集说起。数据采集都是“有组织”的获取数据，从信息论来说，“有组织”意味着系统性的引入了新的信号，这种信号可能是噪声。例如某些网站的每张图片都有该网站的Logo，网站Logo对于数据集就是一种“结构化噪声”——称其为“噪声”，是因为这类信号与想要的知识无关；称其“结构化”，是因为噪声信号是由采样方式引入的，是一种带有结构的系统化引入。将带有结构化噪声的数据投给机器，模型会学到错误的相关性，是我们要努力避免的。

例如，斯坦福大学的Novoa博士讨论过一个“肿瘤”和“尺子”的例子[5]，当皮肤科医生在怀疑一种病变是肿瘤时，会借助尺子来准确的测量大小，尺子会留在照片里，见图11；模型会学习到“尺子”和“肿瘤”具有相关性，而这种相关性在实际情况中显然是不存在的。

图11. 包含有尺子的皮肤病变照片（图片来自网络）

结构化噪声的引入并非都显而易见。例如，在购物引擎里搜索“圆领”的衣服，夏天搜到的可能多是T恤，而在冬天搜索得到的多是毛衣，不留意的话，“圆领”标签下就都是同一季的衣服；又如，在使用众包收集图像时，贡献者可能偏向某个特定的网站，该网站图片的特定样式就被带进了数据库。

因此每当引入一种采集数据的手段时，都要小心观察所获取的图像的共性，分析这个共性部分与标签的相关性。如果相关性很强，则不是噪声，例如采集“翻领”时使用“外套”做扩展，因为翻领是在衣服的开襟上设计的，而开襟的衣服一般都是外套，“翻领”和“外套”有强相关性，所以不是噪声。如果相关性很小，例如网站Logo，则显然是噪声，我们可以对图像做处理，去掉Logo区域。如果实在无法去除，可以考虑放弃这种采集方式。

此外，我们还会在第四步建模环节利用模型来检测结构化噪声，将在下文介绍。

样本稀缺还有一种典型的情况，例如文字识别领域的生僻字。全体汉字超过一万个，常用字有3500个，其余的称为生僻字。生僻字在普通语料中极少出现，如果对语料做均匀采样或随机采样，将出现样本量越大、生僻字比例越低的情况，用来训练模型，生僻字的识别能力反而下降。在这种情况下，“采集”的方式已经失效，需要用“生成”的方式：用机器制作生僻字的“人工合成”图像。我们用“生成”方式来制作样本始于三年多前，最初用规则式的生成，即把生僻字的各种形变写成规则由机器来模拟；后来在一年多前开始尝试“对抗生成”[6]。这个方向称为“少样本学习”或“小数据学习”，最近一两年开始被普遍关注，这里不展开讨论。

3.4 第三步：采集好的图像与知识点做关联

在第三步，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。对一些简单任务，上一步中图10里的“人工筛选”就已经完成了实质的标注工作。

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。

经过第一步和第二步，知识体系中不合理的地方已经大部分得到解决。如果在第三步中标注人员仍有困扰，往往困扰的地方可以引发我们深入思考、产生对数据更深的理解。

例如，我们需要判别图12中这件毛衣是“七分袖”、“九分袖”还是“长袖”。仅看最左的平铺图的话，只能得出衣长和袖子的比例，无法判断确切的袖长；而如果单看最右这张模特图，

袖子是撸起来的（这在针织衫中常见），我们也无法判断袖长；直到看到中间这幅图的上身效果，我们才能判断是“长袖”。

图12. 毛衣的三种照片

这启发我们思考商家做如此拍摄的用意：拍摄平铺图是用来展示衣服的物理属性，而拍摄模特图是用来展示穿着方式和穿搭理念——这两方面的知识对于理解衣服都是必要的，模型都要学习到。

我们再审视袖长的命名方式：“七分”、“九分”都是相对胳膊说的，人体是天然的尺子，在人体上才能得到准确测量。我们要给出成对模特图和平铺图，标注人员才能做准确标注。回想上一节中“结构化噪声”的例子，活检照片中的尺子，在那里是噪声，而在这里是合理的：因为衣服是为人服务的。

再举一个例子，见图13，我们要标注这件衣服的“下摆左右端点”和“衣长”。可这是“一件棕色针织衫内搭一件白色衬衣”呢，还是“一件有白色衬衣下摆的棕色针织衫”？

图13. 一件“假两件”的上衣

这种“假两件”衣服并不少见，一度是标注人员的困扰。如上文所述，我们认识到衣服展示有“物理属性”和“穿搭理念”的差别，就“假两件”来说，经过仔细讨论，我们认为这类图片意在表现穿搭理念，应从整体视觉效果考虑，把“假两件”判为一件，衣服下摆的左右点在白色部分，而衣长是“正常”。

经过第三步，我们对知识和数据的理解更深了。

3.5 第四步：利用模型做迭代

好消息是终于走到了最后一步，坏消息是还要走回头路。

在第四步，算法人员把标注好的数据投给机器，做模型的训练和评测。假设算法人员的建模调参的手艺没问题，那模型就该在一定程度上反映数据集的质量好坏、哪里有缺陷，如同一面能隐约成像的镜子。以模型为鉴，就可以迭代改进数据集。如下图所示。

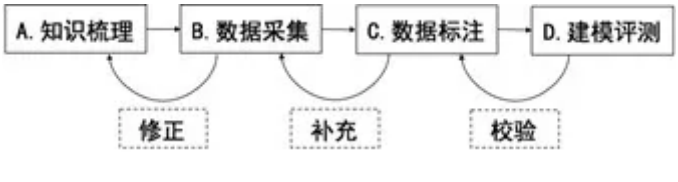

图14.四个步骤的迭代示意图

（1）从D到C的迭代，关键词是“校验”，校验的是标注人员的标注质量。通常我们不会把所有数据都标注完才投给机器去训练，而是分批次标注。假设有10000个样本，我们会分2000、3000、5000三个批次。前一个批次的样本投入训练，如果模型的训练准确率达到满意，说明标注质量合格，才进行下一个批次的标注；否则要总结经验、重新标注。这样可以减少标注的试错成本。

（2）如果标注质量始终不过关，要检查数据采集中的问题。实际上，我们正是利用从D到C再到B的迭代，来应对数据采集中的“样本稀缺”和“结构化噪声”问题。如下图所示。

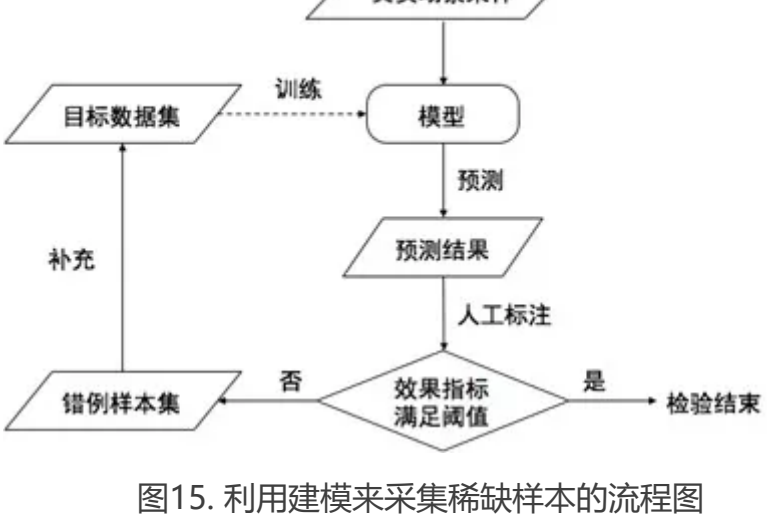

图15. 利用建模来采集稀缺样本的流程图

我们会模拟真实使用场景，进行随机采样，这样得到的样本于环节A和B无关，我们称之为“真实场景采样”。我们把“真实场景采样”放到训练好的模型中去运行一遍，这个过程称为“预测”。预测结果经过人工审核后，如果效果达到满意，就说明模型被训练得不错（即训练数据不错），数据集的“结构化噪声”得到了较好的抑制；如果预测效果不佳，说明数据集中欠缺某方面的训练样本，那把错例补充回数据集，继续训练，并更新真实采样后再做预测，直到效果满意为止。

这个过程既是克服“结构化噪声”的过程，实际上也是一种补充稀缺样本的方法，已成为我们采集数据的常规手段之一。

（3）如果需要回溯到环节A，说明知识体系中的问题跨过了第二步“采集”和第三步“标注”，直到第四步“建模”才暴露出来，这即是人的隐藏很深的认知缺陷，由机器映照出来。这种例子很少，我们在女装“风格”数据集的建设中遇到过。

“风格”是最重要的女装属性维度之一。如上一章介绍，生产商、电商平台、零售商各有其风格体系。我们最初拿到的是平台运营的风格体系，当时就认识到这套体系有诸多不合理、受主观因素和个人影响很大。

这体现在标注过程中，一位服饰专家在第一天标注1000张图，第二天再标注同样的1000张图，结果就差异很大：同一件衣服第一天标“欧美风”，第二天就标“高贵风”。即使身为专家，她也从来没有集中式看过这么大规模有组织的数据；而前后结果的差异表明，当在数据的标注过程中，人的印象被重塑了。

但也没人能告诉我们，合理的风格体系应该长什么样子，只有以这套充满问题的风格体系为起点，采集、标注、建模，通过模型暴露问题，再反馈到专业人士，思考讨论、加深理解，修正体系甚至推倒重建。

从环节D回溯到环节A，这个过程很长，要几个月时间。我们经过了三次大的迭代，花了一年半的时间，才得到一个勉强可用的风格体系。而建设第二章中理想化的层次式风格体系，是大得多的挑战，我们才刚刚开始。

我们的体会是：知识并非生来严谨，而是从混乱中走来。人的认知缺陷，可以通过机器映照出来、加以改正，人和机器在彼此学习，这是AI时代之前不曾见到过的。

四. 确定性能的“指标”

数据集是用来训练和评测模型的。数据集标注好之后，还应有一套用来评测模型的方法，就是“指标”。知识、数据、加上指标，才是一个完整的数据集。好的指标也体现了对于“实用”数据集的追求。

最基础的指标是准确率（P）和召回率（R），常用在搜索和分类任务中。假设模型找回的8个结果中有4个是正例，而数据集里总共有10个正例，则准确率P=4/8=50%, 召回率R=4/10=40%。P和R是一对相互制约的指标，共同刻画模型的能力。

一对（P,R）值对应是模型在固定一组参数时的表现。通过调整模型参数，可以得到一系列的点，就连成一条“P-R曲线”，该曲线可以更全面的体现模型能力，人们用一个值“AveP”来表征，可以把AveP简单理解为“在一个纵轴为P，横轴为R的坐标系里，P-R曲线下方的面积”，面积越大越好。目标检测比赛VOC从2010年后采用的指标就是AveP。

在搜索和分类任务中，识别结果就是一个实例，正例就是识别的标签与标注的标签一致。在有些任务中，如目标检测，识别对象是一个区域，这时要多一个指标IoU。IoU描绘了识别区域与标注区域的面积重合情况，数值上就是二者交集与并集的面积比。IoU高于一个阈值则是正例。业界通常选取IoU>0.5，例如ImageNet比赛[7]就使用IoU>0.5。在我们的一个商用的图像搜索系统中，选择的是IoU>0.7。

当识别对象是一个序列时，如字符串，由于次序本身也是信息的一部分，就需要更精巧的指标。在文字识别和语音识别中，普遍采用“编辑距离”作为指标，即一个字符串经过多少次“增”、“删”、“改”的操作可以变化为另一个字符串。例如，“aboc”和“obac”之间差距为2次“改”，编辑距离为2；“真图像和美”和“图像与美好棒”差距1次“删”、1次“改”、2次“增”，编辑距离为4。如果简单统计字符出现次数的话，“aboc”和“obac”的准确率和召回率都是100%，显然和实际不符。

评测指标还有很多，例如搜索中的R@N，这里不赘述。另一方面还要根据情况划分难度，例如目标检测中按照图像背景复杂度分档，文字识别按照拍照质量分档等。业界在评测方法上有很多经验，也在寻求越来越贴近实际情况的做法。在最近的比赛COCO[8]中，一方面会考察模型取不同IoU阈值时获得的AveP值，另一方面也会根据检测物体大小的不同而区别对待。这些做法使数据集被更合理的结构化了，也可视为数据集建设的一部分。

这里补充一些我们的经验。有的服饰属性维度如“领型”，包括“圆领”、“方领”等标签，识别结果的对和错，就是1和0的关系；而有的属性维度如“袖长”，从短到长有7个标签，加上前文提过的“不可判别”，一共8个标签：“不可见，杯袖，短袖，五分袖，七分袖，九分袖，长袖，超长袖”，如图16所示。我们对
“袖长”的评测方法做了两步细化。

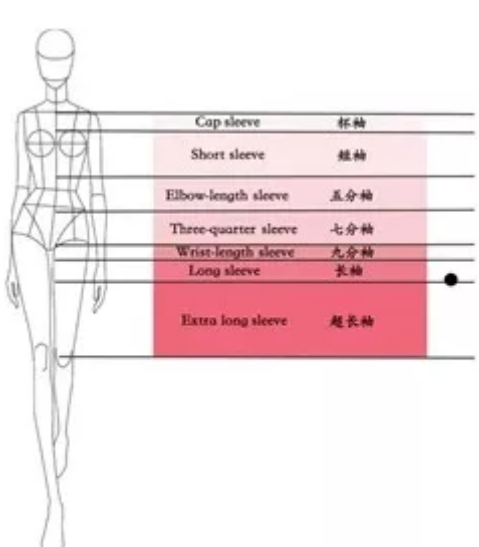

图16.袖长的标注标准示意图

首先，在“是（Y）”和“否（N）”之外，还设立了“模糊（M）”。如果一件衣服出现在图16中黑点的位置，那么对应上面的8个标签，标注结果会是（N,N,N,N,N,N,Y,M）。这使得对边界点的判别更加合理。

进一步观察，把“九分袖”错判为“七分袖”，和错判为“短袖”，错误程度是不同的，应区别对待，我们就引入了标签距离，把标注结果细化为（0,0,0,3,5,7,10,8），这样更贴近实际情

况。

可以看出，指标体系的丰富和细化，其实是知识的一层更精细的表达，数据集要走向实用，要重视这些细节。

五、总结

综上所述，我们介绍了如何做一个“实用”的图像数据集。

我们首先强调了“用途”的重要性：用途是看待事物的视角，是取舍的依据。

其次，我们讨论了数据集建设的三个方面：

- 知识：专业知识的引入是必要的；知识的用途发生了变化，制作者要和专业人士一起来重建知识。
- 数据：从知识到数据，是反复迭代的过程；知识重建贯穿到了采集、标注、建模等所有环节；机器参与到知识重建的过程中来，这是前所未有的新情况。
- 指标：指标可以承载知识的一些更精细的层面，好的指标应在细节上更贴近实际。

制作实用的数据集，不仅是为了AI走向落地，也是计算机视觉自身发展的需要。以我曾接触过的图像技术领域，人们在很多有潜力的议题上浅尝辄止，这其中有工具不得力的原因，另一大原因是研究的基础——数据集——制作不严谨，基础不牢靠，让后来者难以为继。也难怪，倒回去十年，从事计算机视觉的人，吃饭都困难，学生毕业后往往要转行，何谈做一个实用的数据集。

今天情况已经不同，工具发展了，资源丰富了，计算机视觉在走向一门成熟的学科。科研无外乎两件事：1. 制备实验对象、做观测；2. 分析总结、抽象出理论。计算机科学从诞生始，不被视为“科学”，而是“工程”。今天计算机视觉火了，而工程的味道比过去更浓，因为现在正处于工具——深度学习——大发展的时代。长远来看，学科要发展，制作数据集上必然走向更严谨，AI从业者应更有勇气、承担责任，才不负时代的期望。

作者介绍

贾梦雷，1998年至2005年就读于中国科学技术大学，取得本科及硕士学位。毕业后曾任职于微软亚洲研究院和搜狗。于2008年加入淘宝，创立了阿里巴巴最早的图像技术团队，构建了阿里集团内部应用广泛的图像技术基础设施，外部知晓的产品有图片保护产品“八载”、文字识别产品“读光”、以及时尚与AI结合的 “时尚之心” （FashionAI）。

本文谈及的经验基本都来自 “时尚之心” 。大家在年内可以在手机淘宝上用到“时尚之心”支持的应用，并可以去线下体验FashionAI门店。FashionAI全球挑战赛也在如火如荼的举行中，大家可以从以下网址获取FashionAI数据集：

http://fashionai.alibaba.com/datasets

鸣谢

感谢阿里巴巴“图像和美”团队的全体成员，尤其是“时尚之心”项目的同学，包括段曼妮、孔祥衡、曹阳、石克阳、王从德、王永攀等，都参与了写作。此外，感谢香港理工大学的黄伟强教授及邹星星同学，在时尚专业知识方面对项目及本文多有贡献。

文献

[1] LFW人脸数据集：http://vis-www.cs.umass.edu/lfw/

[2] ChestX-ray8医疗影像数据集：https://arxiv.org/abs/1705.02315

[3] Luke Oakden-Rayner的博客：https://lukeoakdenrayner.wordpress.com/2017/12/18/the-chestxray14-dataset-problems/

[4] DeepFashion数据集：http://mmlab.ie.cuhk.edu.hk/projects/DeepFashion.html

[5] “肿瘤”和“尺子”的例子：https://amp.thedailybeast.com/why-doctors-arent-afraid-of-better-more-efficient-ai-diagnosing-cancer

[6] 对抗生成学习：https://arxiv.org/abs/1511.06434

[7] ImageNet数据集：http://www.image-net.org/

[8] MS COCO数据集：http://cocodataset.org/