第3章

数据标注基础知识


微课视频

3.1 数据标注基础

3.1.1 数据标注的概念

人工智能的目标就是机器代替人去认知与思考，我们将某一图片参数设为汽车，计算机搜索到这张图片后就可以知道是汽车图片。但相比人类，机器并不具备思考与联想的能力，换一张同样的图片之后，由于没有设置参数，机器可能就识别不出里面的“汽车”了。那么如何让计算机可以自我学习认识汽车呢？这时数据标注正式出场了。

数据标注就是给机器大量标注好的图片，让机器找到这些图片里汽车的共同特征，那么以后就可以识别出其他汽车了。当前学术界比较认可的数据标注概念是对文本、图像、语音、视频等未处理的初级数据进行归类、整理、编辑、纠错、标记和批注等加工处理，为待标注数据增加标签，生成满足机器学习训练要求的机器可读数据编码的工作，如图 3-1 所示显示了一个图像标注的示例。

标注者需要识别和标注图片中的各类车辆，如卡车、轿车、面包车、皮卡车等各类车的对象。其中需要了解如下概念。

1. 标签

标签主要是标识数据的特征、类别和属性等，可用于建立数据与机器学习训练要求所定义的机器可读数据编码间的联系。

2. 标注任务

标注任务是指按照数据标注规范对数据集进行标注的过程。

图 3-1　数据标注示例（见彩插）

3. 标注工具

标注工具是指数据标注员完成标注任务产生标注结果所需的工具和软件。标注工具按照自动化程序不同，可分为手动标注工具、半自动标注工具和自动标注工具。

综上，数据标注就是通过数据标注员借助标注工具，对人工智能学习数据进行加工的一种行为。随着无人驾驶、智慧医疗、语音交互等各大应用场景的落地和对标注数据需求的扩大，数据标注师职业和数据标注行业也就应运而生。

3.1.2　数据标注行业的特点

数据标注行业是随着人工智能的火爆而兴起的新兴工作，由于发展时间不长，目前该工作还处于摸索阶段，其具有以下几个特点。

1. 劳动密集行业

数据标注工作需要大量的人力完成，因此该行业属于标准的劳动密集型产业，其区位分布特点与传统工厂的分布十分相似，国内主要集中在山东、河南、河北等劳动力丰富且环绕中心一线城市的市县。

2. 准入门槛低

整个市场大大小小共上千家企业和作坊，规模不一，竞争激烈，从而导致利润薄，服务落后。

3. 市场混乱，亟待规范和整治

数据标注行业以外包为主，数据黄牛利用信息差倒卖数据标注资格，从中牟取利益，导致数据标注需求端层层外包，进一步摊薄利润，导致市场混乱，数据质量和服务较差。

4. 从业人员学历普遍较低

数据标注员大多为较低学历者或残疾人，大专为较高学历。其中主要人员包括数据

标注员、数据审核员和标注管理员。

5. 从业人员以兼职为主

国内兼职的数据标注者数量约为全职的 10 倍。

6. 标记质量参差不齐

很多作坊无法保证数据标注的质量和时间，不符合精度和质量要求越来越高的发展趋势。

7. 敏感数据存在安全隐患

由于混乱的市场秩序，极易导致敏感数据泄露，因此，很多需求方会培养内部数据标注员，专门对敏感数据进行标注。

8. 对上游 AI 算法的依赖程度较高

在当前主流算法为有监督学习和半监督学习的大背景下，有大量的数据标注需求，但如果主流算法逐渐转向无监督学习，将不需要对数据进行标注。

3.1.3　数据标注的分类

1. 分类标注

分类标注，就是我们常见的打标签。一般是从既定的标签中选择数据对应的标签，是封闭集合。如图 3-2 所示，一张图就可以有很多分类/标签：树木、猴子、围栏等。对于文字，可以标注主语、谓语、宾语、名词、动词等。

这类分类主要适用于文本、图像、语音、视频，主要应用于脸部识别、情绪识别、性别识别。

图 3-2　分类标注图片示例

2. 标框标注

机器视觉中的标框标注很容易理解，就是框选要检测的对象。如人脸识别，首先要把人脸的位置确定下来。

1）2D 边界框

为那些人类标注器提供图像，并负责在图像中的某些对象周围绘制框。该边框应尽可能地靠近对象的每个边缘。此项工作通常是在不同公司的自定义平台上完成的。如果某个项目有着独特的要求，那么服务公司则可以通过调整其现有平台，以符合此类需求，典型应用是针对汽车自动驾驶的开发。

如图 3-3 所示，标注器需要在捕获到的交通图像内识别车辆、行人和骑车人等实体，并在其周围绘制边界框。开发人员通过为机器学习模型提供带有边界框标注的图像，以

帮助正在进行自动驾驶的车辆，实时地区分出各类实体，并避免触碰到它们。

2）3D 长方体

与边界框非常相似，3D 长方体标注是在立体图像中识别对象，并在其周围绘制边框。与仅描绘长和宽的 2D 边界框不同，3D 长方体则标注了对象的长、宽和近似深度。如图 3-4 所示，使用 3D 长方体标注，人类标注器可以绘制一个框，将感兴趣的对象封装起来，并将锚点放置在对象的每个边缘。如果对象的一个边缘不可见或被图像中的另一个对象所遮挡，那么标注器就会根据该对象的大小、高度以及图像的角度，来估算其边缘的位置。

图 3-3　2D 边界框标注示例

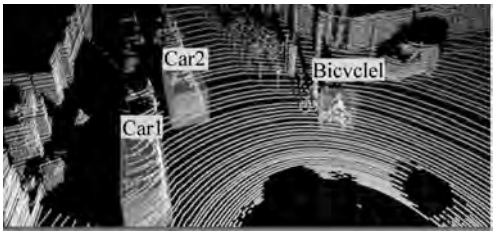

图 3-4　3D 长方体标框示例

这类分类主要适用于图像，主要应用于人脸识别、物品识别。

3. 区域标注

相比于标框标注，区域标注要求更加精确。边缘可以是柔性的。尽管线和样条线可以被用于多种用途，但它们在此主要被用于训练驾驶系统，以识别车道及其边界。顾名思义，标注器将会简单地沿着既定的机器学习方式，去绘制出边界线。通过标注出车行道和人行道，它能够训练自动驾驶系统，了解所处的边界，并保持在某条车道内，以避免压线或转向行驶。此外，如图 3-5 所示的线和样条线也可以被用于训练仓库里的机器人，让它们能够整齐地将箱子挨个摆放，或是将物品准确地放置到传送带上。

图 3-5　区域标注示例

这类分类主要适用于图像，主要应用于自动驾驶中的道路识别。

4. 描点标注

一些对于特征要求细致的应用中常常需要如图 3-6 所示描点标注，如人脸识别、骨骼识别等。

图 3-6　描点标注图片示例(见彩插)

这类分类主要适用于图像，主要应用于人脸识别、骨骼识别。

5. 语义分割

语义分割使用的是和多边形标注类似的平台，能够让标注器在需要标记的一组像素周围绘制线条。和上述主要着眼于绘制对象的外部边缘(或边界)分类不同，语义分割要更加精确和具体一些。它是一个将整个图像中的每个像素与标签相关联的过程。在需要用到语义分割的项目中，通常会为人类标注器提供一系列预定义的标签，以便它能够从中选择需要标记的内容。

在实际应用中，标注器一旦接收到自动驾驶的训练数据，就需要按照道路、建筑物、骑车人、行人、障碍物、树木、人行道以及车辆等，对图像中的所有内容，进行分类分割。而且，人类标注器会使用单独的工具，裁剪掉不属于主体的像素。

语义分割的另一个常见应用场景是医学成像。针对提供的患者照片，标注器将从解剖学角度对不同的身体部位，打上正确的部位名称标签。因此，语义分割可以被用于处理诸如“在 CT 扫描图像中标记脑部病变”之类难度较大的特殊任务。

6. 其他标注

标注的类型除了上面几种常见，还有很多个性化的。根据不同的需求则需要不同的标注。如自动摘要，就需要标注文章的主要观点，这时候的标注严格来说就不属于上面的任何一种了。

3.1.4　数据标注的过程

数据质量是影响人工智能产品准确性的关键所在，一个具有高质量标注的数据集对于模型的提升效果，远远高于算法优化带来的效果。数据标注是通过人工或半自动的方式，将原始数据打上相应的标注，打好标注的原始数据称为标注数据或者训练集数据。

数据标注过程有两个意义：第一，使人类经验蕴含于标注数据之中；第二，使标注数据信息能够符合机器的读取方式。标注数据的难度越高价格越昂贵，以此训练出的模型价值就越高。数据标注的流程如图 3-7 所示，通常分为五个步骤。

图 3-7　标注流程图

1. 业务沟通与业务理解

项目经理与算法工程师要对业务进行理解，明确原始数据的意义与数据标注的价值。业务理解是所有产品工作的基础。

2. 确定原始数据与标注要求

项目经理需要与算法工程师共同确认原始数据及数据标准结果，并确定标注工具。数据标注的结果必须得到算法工程师确认，确保后续建模过程的顺利开展。

3. 编写标注说明

在确认原始数据与标注结果后，项目经理需要编写标注说明。标注说明就好像软件说明书，需要将标注过程按顺序一一列出。标注教程包含 4 个要素：标注软件（平台）、标注要求、标注对象、标注流程。撰写的标注教程同样需要得到算法工程师确认。

4. 进行数据标注

该过程为数据标注过程，项目经理需要不定时进行标注结果抽查。

5. 验收标注结果

项目经理与算法工程师共同对标注结果进行质量验收，验收不合格需要搞清异常原因并重新标注。对于有行业壁垒的数据，标准准确性需要行业专家进行判断。

微课视频

3.2　数据标注的对象

3.2.1　数据标注的人员结构

在数据标注行业流行着一句话，“有多少智能，就有多少人工”。数据标注工作是人工智能领域“入门级”的工种。从工作流程角度看，其技术含量较低，工作量较大，人是这项工作中最大的影响“因素”，因此人们逐渐为数据标注行业贴上了“劳动密集型”的标签。

相较传统的体力工作，数据标注员的工作倒是更轻松体面，因此吸引了众多农民、学生、残疾人群体加入到数据标注大军中，河南、河北、贵州、山西等地的四五线城市相继出现了一些特色的“数据标注村”。在国外，印度同样涌现了不少数据标注村，他们为北美洲、欧洲、大洋洲和亚洲的 AI 公司服务，可见数据标注向劳动力更充足、成本更低的地方

迁移也是全球数据标注行业的发展趋势。这些传统行业的务工者转而成为人工智能浪潮中的参与者,“数据民工”的称谓也由此而来。

但本书作者认为,数据标注其实并不是人工智能行业的“脏活累活”,实际上,真正想高质量地完成数据标注工作并不是随便什么人都可以做到的。AI 本身发展很快,随着应用产品落地,对数据的要求越来越高,对数据采标人员的素质也提出了高要求。

1. 数据标注员

数据标注员是数据标注团队的基石,拥有一批成熟的数据标注员可以让数据标注团队事半功倍,大多数公司对数据标注员的岗位要求如下。

(1) 按照项目的要求,使用标注工具对各类人工智能项目数据(文本、图像、音频、视频)进行标注与质检。

(2) 对不能通过质检的标注结果要进行重新标注。

(3) 理解数据标注规则,根据指导和实际工作要求及时改进工作。

(4) 协助完善标注工具,建立词库定期上交周报和月报,并对工作提出建议。

2. 质检员

质检员一般都是从优秀的数据标注员中挑选出来的。因为数据标注是一个熟能生巧的职业,一个数据标注员接触过的标注对象越多,那么就越有可能熟练掌握各类型项目规则,把质检的任务做好。同时在质检的过程中也会发现问题,把总结出来的经验传达给其他数据标注员,从而提高数据标注质量和效率。

3. 项目经理

项目经理主要就是对团队的各个成员(包括数据标注员和质检员)进行管理和培训,负责组建和培养一批优秀的标注队伍。项目经理需要具备一定的人工智能基础,能够与需求方进行任务对接,把握需求方需求,节约沟通时间,避免导致数据标注员重复返工的情况。标注团队由项目经理、质检员和数据标注员构成,三者之间相互促进,在数据标注过程中分别发挥着重要作用。

3.2.2　数据标注人员的素质要求

数据标注行业的发展越来越趋向于专业化,早期多以中文数据标注为主,现在随着多语种、方言、个性化标注等发展标注需求的增加,对专业化人才的要求也逐步提高,对专业的要求主要表现在如下几个方面。

1. 学习力是数据标注工作的基础要求

学习力是学习的动力、毅力和能力的综合体现,是把知识资源转化为知识资本的能力。学习力包含知识量和知识吸纳的能力。目前数据标注没有统一的规则,有些数据标注项目配备专业的数据标注软件或数据标注平台,但有的数据标注项目只需要用到专业知识或某些大众的数据标注软件。此外,数据标注的需求种类越来越丰富,数据标注的要

求也越来越细致。

因此若想做好数据标注工作，数据标注员需要具备持续的学习能力，不断地学习新规则，开拓专业知识，快速学习掌握行业知识，快速适应数据标注需求，提高各种数据标注软件的操作技能和数据标注能力。

2. 细心是数据标注工作的质量保障

数据标注的终端是人工智能，最终的标注数据是为计算机服务的，所以越精细的标注数据对训练算法越高效（例如，图像标注要求标注误差在 1 个像素点以内，语音标注截取时的误差要控制在 1 个语音帧之内等）。若是标注时不细心，将直接导致数据标注质量不合格，需要打回进行重新标注，这样会浪费很多的时间和人力。态度决定一切，越细心、越认真，标注数据的精细度就越有保证。

在数据标注过程中需要数据标注员细心去找出错误，这样才能不断总结，改进数据标注规则，促进数据标注质量的提升。细心是一个数据标注员具备的基本素质，细心是成为一个合格的标注员最基本的要求。只有细心的数据标注员才能完成数据量极大的数据标注工作。

3. 责任心和耐心是数据标注工作的稳定保证

数据标注在单一的场景中需要重复一个或者几个动作，这种重复的劳动相对比较枯燥，这就要求数据标注员需要有耐心。数据标注员越有耐心，标注数据的稳定性就越有保证。有很多的数据标注项目中标注内容是极其复杂的，如对于车的标注，车辆、人物、指示牌、路灯等都需要标注其类型和属性，每一张图像需要标注很多内容，标注完之后图像会有很多重叠的地方，若是没有耐心就无法完成这类复杂的数据标注项目。

此外，有时候一个场景可能出现多种要标注的元素，这就十分考验数据标注员的耐心，因此具备耐心是一个数据标注员必备的素质。此外，数据标注工作是一份比较枯燥又重复的工作，数据标注员需要重复对一些场景进行标注。具有责任心的员工，会认识到自己的工作在组织中的重要性，把实现组织的目标视为自己的目标。

4. 专注力是数据标注工作的效率保证

专注力是指一个人专心于某一事物或活动时的心理状态。在数据标注过程中，数据标注员需要每天面对大量数据，集中精力进行数据标注，如果没有足够的专注力是做不好数据标注的。

5. 良好的沟通表达力是数据标注工作的有力支撑

沟通表达是将思维所得的成果用语言、语音、语调、表情、行为等方式反映出来的一种行为。很多数据标注项目的数据标注规则可能不是很明确，项目方要充分和需求方进行沟通，表达诉求；并需要将需求方的意思完整传达给数据标注员们。

数据标注员在数据标注过程中可能会遇到一些困难，也需要表达诉求。质检员在质检后指出数据标注错误时也要跟数据标注员说明错误。

3.2.3　数据标注的采集目标

数据标注中的数据来源多种多样，根据当前主流的应用场景可以将数据标注的数据来源归于如下几类。

1. 人脸数据采集

目前对于人脸数据，一方面可通过第三方数据机构购买，另一方面也可自行采集。在采集之前，首先需要根据应用场景，明确采集数据的规格，对包括年龄、人种、性别、表情、拍摄环境、姿态分布等予以准确限定，明确图片尺寸、文件大小与格式、图片数量等要求，并在获得被采集人许可之后，对被采集人进行不同光线、不同角度、不同表情的数据拍摄与收集，并在收集后对数据做脱敏处理。

以下为一个简单的人脸数据采集规格示例。

年龄分布——18～30 岁
性别分布——男：54 人；女：46 人
人种分布——黑种人：50 人；白种人：40 人；黄种人：10 人
表情类型——正常，挑眉，向左看，向右看，向上看，向下看，闭左眼，闭右眼，微张嘴，张大嘴，嘟嘴，微笑，大笑，惊讶，悲伤，厌恶
拍摄环境光线亮的地方，光线暗的地方，光线正常的地方
图片尺寸——1200×160 像素
文件格式——JPG
图片数量——20 000 张
适用领域——人脸识别，人脸检测

2. 车辆数据采集

在对车辆数据的采集中，常见的方式是通过交通监控视频进行图片截取，图片最好包括车牌、车型、车辆颜色、品牌、年份、位置、拍摄时间等车辆信息，并做统一的图片尺寸、文件格式、图片数量规定，同时做脱敏处理（即数据漂白），实时保护隐私和敏感数据。

以下为一个简单的车辆数据采集规格示例。

车型分布——小轿车、SUV、面包车、客车、货车、其他
车辆颜色——白、灰、红、黄、绿
其他拍摄时间——光线亮的时候，光线暗的时候，光线正常的时候
车牌颜色——蓝、白、黄、黑、其他
图片尺寸——1024×768 像素
文件格式——JPG
图片数量——75 000 张
适用领域——自动驾驶、车牌识别

3. 街景数据采集

与车辆数据采集类似，街景数据采集也可通过监控视频进行图片截图与收集，同时可借助车载摄像头、水下相机等进行街景拍摄。例如，谷歌在进行街景拍摄时，通过集采集、定位与数据上传于一体的街景传感器吊舱、街景眼球、街景塔、街景三轮车、街景雪地车、街景水下相机等多种方式进行360°图像采集。采集的街景图片主要包括城市道路、十字路口、隧道、高架桥、信号灯、指示标志、行人与车辆等场景。同时，对于采集的数据同样需要做统一的图片尺寸、文件格式、图片数量规定与脱敏处理。

以下为一个简单的街景数据采集规格示例。

采集环境——城市道路
路况覆盖——十字路口、高架桥、隧道
数据规模——10 000 张
拍摄设备——车载摄像头
图片尺寸——1920×1200 像素
文件格式——PNG
图片数量——15 000 张
适用领域——自动驾驶

4. 语音数据采集

对于语音数据采集，较为直接的方式是语音录制。在录制之前，对采集数量、采集内容、性别分布、录音环境、录音设备、有效时长、是否做内容转写、存储方式、数据脱敏等加以明确，并在征得被采集人的同意之后进行相关录制。由此可建立中文、英语、德语等丰富的语种语料以及方言语音数据。

以下为一个简单的语音数据采集规格示例。

采集数量——500 人
性别分布——男性：200 人；女性：300 人
是否做内容转写——是
录制环境——关窗关音乐，关窗开音乐，开窗开音乐，开窗关音乐
录音语料——新闻句子
录音设备——智能手机
音频文件——WAV
文件数量——200 000 条
适用领域——语音识别

5. 文本数据采集

如前所述，在数据标注中需要建立多种文本语料库，可以通过专业爬虫网页，对定向

数据源进行定向关键词抓取，获取特定主题内容，进行实时文本更新，建立包括多语种语料库、社交网络语料库、知识数据库等，并对词级、句级、段级和篇级等进行说明。在采集之前，对分布领域、记录格式、存储方式、数据脱敏、产品应用等进行明确界定。

以下为一个简单的文本数据采集规格示例。

采集内容——英语、意大利语、法语等语言网络文本语料
文件格式——txt
编码格式——UTF-8
文件数量——50 000 条
适用领域——文本分类、语言识别机译

6. 常用数据标注集

数据集分为图像、视频、文本和语音标注数据集四大类，这些数据集的数据的类别、用途和特性如表 3-1 所示。

表 3-1　常用数据标注集

类别	数据集名称	用途	大小	开放情况
图像标注数据集	ImageNet	图像分类、定位、检测	约 1TB	是
	COCO	图像识别、分割和图像语义	约 40GB	是
	PASCAL VOC	图像分类、定位、检测	约 2GB	是
	OpenImage	图像分类、定位、检测	约 1.5GB	是
	Flickr30k	图片描述	30MB	是
视频标注数据集	YouTube-8M	理解和识别视频内容	1PB	受限
	kinetics	动作理解和识别	约 1.5TB	是
	AVA	人类动作识别	—	是
	UCF101	视频分类、动作识别	6.5GB	是
文本标注数据集	Yelp	文本情感分析	约 2.66GB	是
	IMDB	文本情感分析	80.2MB	是
	Muti-Domain Setiment	文本情感分析	52MB	是
	Setiment140	文本情感分析	80MB	是
语音标注数据集	LibriSpeech	训练声学模型	约 60GB	是
	AudioSet	声学事件检测	80MB	是
	FMA	语音识别	约 1000GB	是

其中常用的主要数据集说明如下。

1）ImageNet 数据集

该数据集拥有专门的维护团队，而且文档详细，几乎成为目前检验深度学习图像领域算法性能的“标准”数据集。

2）COCO 数据集

该数据库是在微软公司赞助下生成的数据集，除了图像的类别和位置标注信息外，还

提供图像的语义文本描述。因此，它也成为评价图像语义理解算法性能的“标准”数据集。

3）YouTube-8M

该数据集是谷歌公司从 YouTube 上采集到的超大规模的开源视频数据集，这些视频共计 800 万个，总时长为 50 万小时，包括 4800 个类别。

4）Yelp 数据集

由美国最大的点评网站提供，包括 70 万条用户评价，超过 15 万条商户信息，20 万张图片和 12 个城市信息。研究者利用 Yelp 数据集不仅能进行自然语言处理和情感分析，还可以用于图片分类和图像挖掘。

5）LibriSpeech 数据集

该数据库是目前最大的免费语音识别数据库之一，由近 1000 小时的多人朗读的清晰音频及其对应的文本组成，是衡量当前语音识别技术最权威的开源数据集。

3.2.4 数据标注平台和工具

1. 数据标注平台

近年来，国内的一些互联网公司、大数据公司和人工智能公司纷纷推出了自己的数据标注众包平台和商用标注工具，如数据堂、百度众测、阿里众包、京东微工等，这些工具至少要包含如下功能。

1）进度条

用于指示数据标注的进度。一方面方便标注人员查看进度，另一方面也利于统计。

2）标注主体

可以根据标注形式进行设计，一般可分为单个标注（指对某一个对象进行标注）和多个标注（指对多个对象进行标注）的形式。

3）数据导入导出功能

可以有效地与外部系统进行数据的交互。

4）收藏功能

针对模棱两可的数据，可以减少工作量并提高工作效率。

5）质检机制

通过随机分发部分已标注过的数据，检测标注人员的可靠性。

2. 开源数据标注工具

在选择数据标注工具时，需要考虑标注对象（如图像、视频、文本等）、标注需求（如画框、描点、分类等）和不同的数据集格式（如 COCO，PASCAL VOC，JSON 等）。常用标注工具如表 3-2 所示。

表 3-2 常用数据标注工具

名称	简介	运行平台	标注形式	导出数据格式
LabelImg	著名的图像标注工具	Windows、Linux、macOS	矩形	XML 格式

续表

名称	简介	运行平台	标注形式	导出数据格式
LabelMe	著名的图像标注工具，能标注图片和视频	Windows、Linux、macOS	多边形、矩形、圆形、多段性、线段、点	VOC 和 COCO 格式
RectLabel	图像标注	macOS	多边形、矩形、多段性、线段、点	YOLO、KITTI、COCO1 与 CSV 格式
VOTT	微软发布，基于 Web，能标注图像和视频	Windows、Linux、macOS	多边形、矩形、点	TFRecord、CSV、VbTT 格式
LabelBox	适用于大型项目的标注工具，能标注图像和视频	—	多边形、矩形、线段、点、嵌套分类	JSON 格式
VIA	VGG 的图像标注工具，也支持音频和视频标注	—	矩形、圆、椭圆、线段、点、多边形	JSON 格式
COCO UI	用于标注 COCO 数据集的工具，基于 Web 方式	—	矩形、线段、点、多边形	COCO 格式
Vatic	带有目标跟踪的视频标注工具，适合目标检测任务	Linux	—	VOC 格式
BRAT	基于 Web 的文本标注工具，主要用于对文本的结构化标注	Linux	—	ANN 格式
DeepDive	处理非结构化文本的标注工具	Linux	—	NLP 格式
Praat	语音标注工具	Windows、Linux、macOS	—	JSON 格式

除了 COCO UI 和 LabelMe 工具在使用时需要 MIT 许可外，其他的工具均为开源使用。大部分的开源工具都可以运行在 Windows、Linux、macOS 系统上，仅有个别工具是针对特定操作系统开发的（如 RectLabel）。

这些开源工具大多只针对特定对象进行标注，只有少部分工具（如精灵标注助手）能同时标注图像、视频和文本。

市场上还有一些特殊功能的标注工具，如人脸数据标注和 3D 点云标注工具。不同标注工具的标注结果会有一些差异，但尚未有研究关注它们的标注效率和标注结果的质量。

3.3　数据标注的质量保证

3.3.1　数据标注质量的地位

蒸汽机释放了人的体力，但是蒸汽机并不是模仿人的体力，汽车比人跑得快，但是汽

车并不是模仿人的双腿。未来的计算会释放人的脑力，但是计算机不是按照人脑一样去思考，计算机必须要有自己的方式去思考。那么如何能让计算机形成一套自主的思考体系呢？我们需要把人类的理解和判断教给计算机，让计算机拥有人类一般的识别能力，数据标注就这样出现了。

数据标注就是人类用计算机能识别的方法，把需要计算机识别和分辨的图片打上特征，让计算机不断地识别这些特征图片，从而最终实现计算机能够自主识别。通俗来讲，想让计算机知道什么是汽车，那么就得在有汽车的图片中，用专业的标注工具按要求标注出来汽车中的各重要元素，计算机通过不断地识别这些特征图片，最终能够自主地识别特征物品。

可见，如果把人工智能看作一个天赋异禀的孩子，数据标注就是这个孩子的启蒙恩师，在传授的过程中，老师讲得越细致，越有耐心，那么孩子成长得也就越稳健。同样，换个角度，如果说人工智能是一条高速公路，那么数据标注就是高速公路的基石，基石越稳固，质量越过硬，那么使用起来就会越放心，越长久。人工智能是一个复杂的过程，但是不论是多复杂的架构，数据标注永远是体系中的养分，通过不断地改变标注内容来适应不断强大的计算机，所以数据标注是人工智能的重中之重。

3.3.2 数据标注质量标准

1. 图像标注的质量标准

图像标注的质量好坏取决于像素点的判定准确性。标注像素点越接近被标注物的边缘像素，标注的质量就越高，标注的难度也越大。如果图像标注要求的准确率为 100%，标注像素点与被标注物的边缘像素点的误差应该在 1 个像素以内。

2. 语音标注的质量标准

语音标注时，语音数据发音的时间轴与标注区域的音标需保持同步。标注于发音时间轴的误差要控制在 1 个语音帧以内。若误差大于 1 个语音帧，很容易标注到下一个发音，造成噪声数据。

3. 文本标注的质量标准

文本标注涉及的任务较多，不同任务的质量标准不同。例如，分词标注的质量标准是标注好的分词与词典的词语一致，不存在歧义；情感标注的标注质量标准是对标注句子的情感分类级别正确。

3.3.3 数据标注质量检验方法

1. 实时检验方法

当标注员对分段数据开始标注时，质检员就可以对标注员进行实时检验，当一个阶段的分段式数据标注完成后，质检员将对该阶段数据标注结果进行检验，如果标注合格就可以放入该标注员已完成的数据集中，如果发现不合格的则可立即让标注员进行返工改正

标注。

实时检验方法的优点如下。

(1) 能够及时发现问题并解决问题。

(2) 能够有效减少标注过程中重复错误的重复出现。

(3) 能够保证整体标注任务的流畅性。

(4) 能够实时掌握数据标注的任务进度。

实时检验方法的缺点主要是对于人员的配备及管理要求较高。

2. 全样检验方法

全样检验是质检员对全部已完成的数据集进行全样检验，通过全样检验合格的数据标注存放在已合格数据集中等待交付，而对于不合格的数据标注，需要标注员进行返工改正标注。

全样检验方法的优点如下。

(1) 能够对数据集做到无遗漏检验。

(2) 可以对数据集进行准确率评估。

全样检验方法的缺点是需要耗费大量的人力精力集中进行。

3. 抽样检验方法

1) 辅助实时检验流程

当标注员完成第一阶段数据标注任务后，质检员会对其第一阶段标注的数据进行检验，如果标注数据全部合格，在第二阶段实时检验时，质检员只需对标注员的 50%进行检验，如果不合格，在第二阶段实时检验时质检员仍需对标注员的数据标注进行全样检验。

2) 辅助全样检验流程

在全样检验完成后，要对标注员的标注数据进行第一轮抽样检验，如果全部检验合格，在第二轮检验中，标注数据量减少 50%，如果第一轮有不合格的标注数据，在第二轮抽样检验中检验的标注数据量较第一轮的增加一倍。

多重抽样检验方法的优点如下。

(1) 能够合理调配质检员的工作重心。

(2) 有效地弥补其他检验方法的疏漏。

(3) 提高数据标注质量检验的准确性。

多重抽样检验方法的缺点是只能辅助其他检验方法，如果单独实施，会出现疏漏。

3.4　自动标注技术的发展

3.4.1　图像自动标注

随着计算机软硬件、互联网、大数据及分布式存储等技术的不断成熟和快速发展，图像数据在数量和内容上呈现爆炸式增长。根据中国互联网络信息中心发布的《中国互联网发展状况统计报告》显示，多媒体形式网页中图片数量已占八成，以数字图像作为载体

也是文化资源数字化的最主要方式。

在数字图像数据保持高速增长的同时，人们对图像数据的利用能力却没有随之增强。究其原因，是计算机难以通过图像的低层视觉特征提取出可供人类理解的高层语义信息，低层视觉特征和高层语义特征之间存在“语义鸿沟”的缺陷。这也导致我们在应对大规模图像数据时缺少有效的检索方案，从而难以获取所需信息，减少“语义鸿沟”的最有效的途径之一是图像自动标注技术。

图像的自动标注是利用人工智能或模式识别等计算机方法对数字图像的低层视觉特征进行分析，从而对图像打上特定语义标签的一个过程。图像的标注框架如图 3-8 所示，总体可分为两个特征提取模型和一个标注模型。

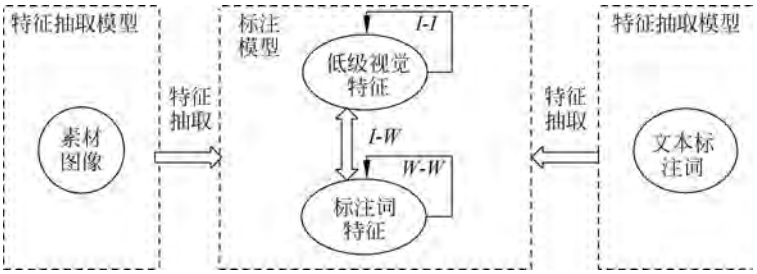

图 3-8　常用自动数据标注模型

两个特征提取模块通过图像的特征提取以及词汇（标签）的特征提取方法可分别得到对应的图像低层视觉特征与标注词特征。图像的标注模型通过需要建立最关键的反映图像和标签关系的 $I(W)$ 映射函数，并通过该映射函数和低层视觉特征矩阵匹配，从而实现对未标注图像进行标签预测。此后再进一步地充分利用分别反映图像之间和标签之间关系的 $I(i)$ 映射函数和 $W(w)$ 映射函数对标注模型进行优化，使其得到鲁棒性较高的标注结果。

1. 标注模型需考虑的问题

图像标注模型最关键的是需要充分利用低层视觉特征和标注词特征，建立起图像之间、图像和标签之间以及标签之间的三种映射函数。但是由于图像训练集本身存在的固有特点、特征提取算法存在差别以及模型对各种特征的适应性不同，所以标注模型还需要考虑如下七个一般性问题。

1）标签的不均衡问题

在图像的训练集中，有少部分标签只存在于较少的图像之中，而另外一些标签则出现频率较高，这种标签分布的不均衡有可能会影响模型的精确性，造成这种问题的原因是在制作训练集时，人们往往倾向用更加广泛和一般性的词汇来进行标注，从而导致标签的频率不尽相同。

2）弱标签问题

这种问题通常在社交领域图像集上出现，即训练集图像中的标注并不能完整地表现图 3-8 中反映的所有语义信息，存在标签缺失或错误的情况。产生这种现象的原因是人

们在标注时的主观性不同。

3）特征的高维度问题

从图像中提取的特征往往维度过高，导致模型计算量增加；此外，维度过高也会产生特征的冗余与噪声。

4）特征内维度不均衡问题

由于标注模型往往需要使用多种低层的图像特征共同作用进行标签预测，而每种特征以及特征内的每一维度对标签预测的贡献程度不一致，影响模型精确性。

5）特征的选择问题

针对特定标注模型所选取、设计的视觉特征，在其他模型上通常表现较差。因此，在设计新的标注模型时需考虑在多种图像特征之中选取具有更加泛化性能的特征。

6）模型优化不足问题

由于在图像之间、图像和标签之间以及标签之间的三种关系中，只需通过最关键的图像和标签间映射函数即可对未标注图像进行标签预测，因此，很多模型忽视了图像之间和标签之间的关系考虑，影响了标注精度的进一步提升。

7）标注模型运行效率问题

图像自动标注模型由于需要大量的计算工作，因此为了标注结果更加精确，模型需要进行大量的关系函数运算，因此需要在标注模型的运行效率和模型精度上找到平衡点，从而可以适用于更多的应用场景。

2. 标注模型类别

图像标注领域自 21 世纪初进入快速发展时期以来，出现了各种不同的方法和模型，但这些图像标注算法和模型依据主要使用的方法可分为如下几类。

1）相关模型

其基本思想为：首先将图像分块，假定分块的图像特征和标签之间存在某种特定的概率；然后建立分块图像的特征和标签之间的联合概率密度；最后根据待标注图像的分块信息，求得其针对每个标签的后验概率。相关模型的代表有 TM 模型、CMRM 模型、CRM 模型以及 MBRM。

2）隐马尔可夫模型

类似于相关模型，同样需要根据图像块和标注词的联合概率密度来求得最终的标注，但不同之处在于隐马尔可夫模型是通过隐马尔可夫链来建立这种相关关系，其代表有 HMM 模型、TSVM-HMM 模型、SHMM 模型以及 HMM-SVM 模型。

3）主题模型

主题模型最早用于对文档的检索，解决了检索问题中的“一词多义”以及“一义多词”问题。在图像标注领域，主题模型同样通过构建隐藏的主题空间，使得具有语义相似度的模态能够映射到同一主题，或者同一主题可被多种模态所表示。

因此，隐藏的主题空间能够较好地建立起图像底层视觉特征同自然语义之间的联系。但由于主题模型依然是通过选取训练集图像中相应的底层视觉特征和标注词汇来进行概率运算，因此其概率分布难以有效描述样本外的情况，泛化性能不高。对于选取何种底层

视觉特征、标注词汇特征，以及对特征的融合利用也是主题模型需要解决的难题。此外，主题模型中需用到 SVD 分解以及 EM 算法等，都需要耗费大量时间以及运算资源。

4）近邻模型

近邻模型在图像标注模型中思想比较简单，其基本原理是具有相似低层视觉特征的图像应该具有相似的语义，因此，利用近邻模型进行图像标注的一般步骤如下。

（1）构建图像低层视觉特征。

（2）通过对低层视觉特征采用某种距离度量策略，选择与待标注图像距离较近的已标注图像。

（3）通过合适的标签扩散方法将已标注图像中的标签应用到待标注图像。

近邻模型的代表模型有 JEC 模型、TagProp 模型、2PKNN 模型、VS-KNN 模型、SNLWL 模型等。

5）图模型

图模型的基本思想是通过图来集成样本间的相似关系，包括样本间视觉特征之间的相似性、标签特征之间的相似性以及视觉特征和标签特征的对应关系，然后再利用相关的图论技术建立图结构中样本以及各种特征的关联模型，从而对标签进行预测。因此，大多基于图的标注模型的区别在于图的构建方式以及选择的图论技术存在差别。

6）相关分析模型

典型相关分析 CCA 模型与 KCCA 模型的本质是用来寻找两组特征变量的最大相关关系，最早被用于基于语义的图像检索领域，其基本思想为：假定图像的视觉特征与对应的标签特征分别为异构的两种特征，则 CCA 模型通过两组对应的基可将两种异构特征分别映射到一个具有最大相关性的可对比隐藏语义空间，进而再通过适当的距离运算或比较模型，获得与图像最相关的标签。由于 KCCA 模型是在 CCA 模型基础上，通过核函数的方式增强了模型的非线性特征，其本质与 CCA 并无区别，因此，本文统一将 CCA 模型和 KCCA 模型都称为 CCA 模型。

7）深度学习模型

近年来，由于硬件运算设备如 GPU、NNU 等运算性能的大幅提升，以深度学习为基础的模型克服了早期存在的运算瓶颈，并且在计算机视觉、文本处理、电子商务等各应用领域，以高泛化性和优异的性能得到了广泛的应用和发展。深度学习模型通过若干层的卷积神经网络（CNN）、非线性激活函数和池化层相连接，直接建立从图像原始像素到图像标签的端到端关系映射。深度学习模型具有以下两个重要优势。

（1）与传统方法中手工设计的图像特征相比，通过预训练的深度学习模型提取出的图像特征具有更高的泛化性以及抽象性。

（2）通过基于深度学习的文本处理模型提取出的标签特征具有高层的语义相关关系。

3.4.2 文本自动标注

Internet 通信技术和大容量存储技术的发展，加速了信息流通的速度，形成了大规模真实文本库。这些文本库具有规模大、实时性强、内容分布广和格式灵活多样等特点。这

些特点导致传统的文本信息处理方法已经无法满足新变化的需要，新式的文本信息处理的词类标注和语义标注工作，无论是在理论、方法还是工具方面都面临着如何适应这些变革。这些变革主要表现为处理对象由少量例句到大规模的真实文本，处理方法由完全语法分析到部分语法分析，处理范围由典型领域到开放的实用领域等。

1. 文本自动标注应采取经验主义和理性主义相结合的方法

1992 年，国际机器翻译会议的主题即为“机器翻译中的经验主义和理性主义方法”。随着对大规模真实文本处理的日益关注，人们已普遍认识到基于语料库的分析方法（即经验主义方法）至少是对基于规则的分析方法（即理性主义方法）的一个重要补充。

众多实验结果表明，基于语料库统计的方法具有很好的一致性和较高的覆盖率，并且可以将一些不确定的知识定量化。但是在这种方法中获取知识的机制与语言学研究中获取知识的机制完全不同，因而所获取的知识很难与现有的语言学成果相结合。同时该类算法的时间和空间复杂度都比较大，随着标记跨段长度的增加以及兼类词标记数目的增大，其实际运行效率将会降低。

基于规则的理性主义方法可以将大量现成的语言学知识形式化，具有较强的概括性，便于引用最新研究成果。因为任何词类都有其内部的共性和区别于其他词类的个性。只要把词类的共性和它外部的个性特征结合起来，词的兼类问题是可能得到妥善解决的。例如，名词的语法个性在于它可以直接受量词的修饰、可以受名词直接修饰、可以做“有”的宾语、可以与名词组成并列结构等。如果某个词具备了上述特征，就可以判定它是名词。例如，“主张”“计划”“建议”本来是动词，根据上述特征判断，它们在“五点主张”“不少计划”“许多建议”的语法环境中则一定是名词。

研究人员在对 50 万汉字语料进行词类标注中，根据词的语法功能这一标准判别兼类词，既具科学性又有可操作性，收到了较好的效果。但是实践表明，基于规则的方法所描述的语言知识的颗粒度太大，难以处理复杂的、不规则的信息，特别是当规则数目增多时，很难使规则全面覆盖某个领域的各种语言现象。

为此，研究人员正在尝试把基于规则的方法和基于统计的方法结合起来使用，使语言知识选择引用和用统计方法建立的语言模型有机地结合起来，使之互相补充，相得益彰。

2. 文本自动标注应同切词过程一体化进行

人们分析和理解自然语言时，其特点和过程是什么样的呢？通过仔细观察和思考，不难发现人脑处理自然语言的特点和过程是将切词和词类识别一体化进行。即边切词、边进行词类或语义识别，二者是不可分离的两个方面。下面以处理兼类词“为”和由“为”构成歧义字段为例，说明切词和词类标注不可分离的性质。例如，“他们以服务社会、报效祖国为人生的第一目标”。

理解这句话的关键是判别兼类词“为”的词性，并处理歧义切分字段“为人生”到底该切分为“为人/生”还是切分为“为/人生”。前者是词性判别，后者是词的切分。句法知识在理解这句话中首先起作用，当我们看/听到介词“以”时，首先查询的是这个介词后面的第一个动词，当兼类词“为”出现时，它的动词词性马上被确认，也就是说，介词的词性同时

被排除，因为汉语中“以……为……”常作为一种固定搭配使用。确定了“为”的词性，歧义切分字段“为人生”的正确分词结果“为/人生”也被随之确定下来，可见句法知识不仅解决了词性的确定，同时也解决了歧义的切分。词类判别和切词是同时进行而不可分离的。

目前把切词和词类标注分离开将带来什么结果呢？还是以《分词规范》为例，它明确规定“场、室、界、力”等字用在某个单位的末尾时，就要一律按“接尾词”单独切分，如“运动/场”“会议/室”“新闻/界”“生产/力”等。因为切词的目的不是为切词而切词，而是要为进一步的句法分析和理解语言服务。那么词性标注就成为下一步不可或缺的工作，但这时上面的分词结果就出现了麻烦。“场、室、界、力”如果是词也只能是名词，可它们是词吗？如果是词，为什么它们从来都不能独立运用单独成词，而只能以附加的成分出现在某些名词性成分之后？语言中真的有粘着的“名词”吗？答案都只能是否定的。这种把构词成分误作“分词单位”切分的做法造成的上述不能自圆其说的窘况，正是脱离词类标注单独切词的结果。

鉴于此，作者觉得应将切词和词类标注作为理解和分析语言材料的两个不可分离的环节进行一体化处理。这样做才真正符合人处理语言和过程的特点，才无愧于“人工智能”，由此而得出的结果才可能达到预期效果。

3. 应加强文本自动语义标注尝试

在中文信息处理中，词汇、句法和语义层面的分析研究都需要借助于词义特征。一词多义形成了词的多义现象，自动语义标注主要是解决词的多义问题。一词多义虽然是自然语言中的常见现象，但是在一定的上下文中一个词一般只能解释为一个义项。所谓自动语义标注就是运用逻辑运算和推理机制，对出现在一定上下文中的词语语义的义项进行正确的判断，确定其正确的语义，并加以标注。

思考题

1. 简述数据标注的概念。
2. 数据标注质量检验方法有哪些？
3. 标注模型需考虑到哪些问题？