第5章

数据标注技术

人工智能是机器产生的智能，在计算机领域是指根据对环境的感知做出合理的行动并获得最大收益的计算机程序。人类在认识一个新事物时，首先要形成对该事物的初步印象。例如，要识别飞机，就需要看到相应的图像或者真实物体。数据标注可视为模仿人类学习过程中的经验学习，相当于人类从书中获取已有知识的认知行为。具体操作时，数据标注把需要计算机识别和分辨的图像事先打上标签，让计算机不断地识别这些图像的特征，最终实现计算机能够自主识别。本章将系统地讨论数据标注过程中涉及的技术与方法。

5.1 数据标注的定义与分类

数据标注是大部分人工智能算法得以有效运行的关键环节。人工智能算法是数据驱动型算法，也就是说，如果想实现人工智能，首先需要把人类理解和判断事物的能力教给计算机，让计算机学习到这种识别能力。

目前，学术界尚未对数据标注的概念形成一个统一的定义。蔡莉等归纳了大部分研究者对数据标注的认识，给出如下定义：需要大量的训练数据创建像人类一样行动的人工智能或机器学习模型，必须训练模型理解特定信息以做出决策并采取行动。标注是对未处理的初级数据（包括图像、语音、文本、视频等）进行加工处理，并转换为机器可识别的信息的过程。原始数据一般通过数据采集获得，随后的数据标注相当于对数据进行加工，然后输送到人工智能算法和模型里完成调用。

数据标注的过程是通过人工贴标签的方式为机器系统提供学习的样本。数据标注是把需要机器识别和分辨的数据贴上标签，然后让计算机不断地学习这些数据的特征，最终实现计算机能够自主识别。

数据标注类有 3 种常用的划分方式：

（1）根据标注对象进行分类，包括图像标注、语音标注和文本标注等。

（2）根据标注的构成形式分为结构化标注、非结构化标注和半结构化标注。

（3）根据标注者类型分为人工标注和机器标注。

5.1.1 标注的分类

下面根据标注对象进行分类，以深入地了解不同类型的数据标注。

1. 图像标注

标注图像对于许多用途至关重要，例如涉及计算机视觉、机器人视觉、面部识别以及其他使用机器学习方法破译图像的解决方案的用途。在为学习系统构建训练数据集时，经常使用图像标注。为了在训练中使用图像，需要为其添加信息，例如 ID、标题或关键字。

图像标注涵盖了视频标注，因为视频是由连续播放的图像组成的。图像标注一般要求标注人员使用不同的颜色对不同的目标标记物进行轮廓识别，然后给相应的轮廓打上标签，用标签概述轮廓内的内容，以便让算法模型能够识别图像中的不同标记物。图像标注常用于人脸识别、自动驾驶车辆识别等应用。

有许多应用程序需要大量带标注的照片，例如自动驾驶车辆使用的计算机视觉系统、选择和分类产品的机器以及自动诊断医疗问题的医疗保健应用程序。标注图像是训练这些算法的绝佳方法，可以提高精度和准确度。

图像上的标签数量可能会根据使用场景而增加。就其最基本的形式而言，图像标注可以分为两类：

一类是关于图像的分类。经过带标注图像训练的机器可以通过将图像与一组标签进行比较快速、准确地识别图像的内容。

另一类是关于物体识别和物体检测。它是图像分类的改进版本，可以准确地描述图像中显示的物体的数量和相对位置。与对完整图像进行分类的图像分类不同，对象识别对单个对象进行命名。例如，图像分类需要为图像整体分配“白天”或“夜晚”标签。当使用对象识别处理图像时，多个对象（例如自行车、树或桌子）将被单独分类。

2. 语音标注

语音标注是通过算法模型识别转录后的文本内容并与对应的音频进行逻辑关联。语音标注的应用场景包括自然语言处理、实时翻译等，语音标注的常用方法是语音转写。

3. 文字标注

数据标注对于自然语言处理任务也至关重要。文本标注是指通过添加标签或元数据来添加有关文本数据的相关信息。多种标注（例如情感、意图）甚至查询，都可以应用于文本。文本标注是指根据一定的标准或准则对文字内容进行诸如分词、语义判断、词性标注、文本翻译、主题事件归纳等注释工作，其应用场景有名片自动识别、证照识别等。目前，常用的文本标注任务有情感标注、实体标注、词性标注及其他文本类标注。

4. 情感标注

情感标注依靠高质量的训练数据准确评估人们的感受、想法和观点。通常要由人工标注者收集这些信息。

5. 意图标注

由于人机接口的日益普及，计算机不仅能够理解人类的语言，而且能够理解人类的潜在意图，包括请求、命令、预订、建议和确认等。

6. 语义标注

语义标注可以改进机器学习系统，以识别异常并对其进行充分分类。

7. 命名实体标注

命名实体识别(Named Entity Recognition,NER)系统的训练数据必须广泛且经过人工标注。NER 的主要目标是识别文本中的特定单词或短语并对其进行分类。可以使用 NER 查找诸如人名、地点等内容,具体取决于一组单词的含义。NER 使信息提取、分类变得更加容易。

8. 音频标注

音频标注不仅需要对语音数据和时间戳进行转录,还需要识别语言特征,例如语种、方言和说话者人口统计数据。

9. 视频标注

视频标注与图像标注类似,因为它需要标注视频片段,以便逐帧检测和识别特定对象。机器学习的一个重要组成部分是人工标注的数据。在处理细微差别、细微含义和歧义方面,计算机无法与人类相比。举例来说,需要几个人的意见才能就搜索引擎结果是否相关达成一致。逐帧视频标注采用与图像标注相同的方法,例如边界框或语义分割。该方法对于定位和对象跟踪这两种常见的计算机视觉任务至关重要。

5.1.2 数据标注的应用场景

人工智能的蓬勃兴起促进了数据标注产业的发展,通过大数据和机器学习全面、准确地识别视频、图像及文字内容,并实时返回业务标签,帮助平台实现视频/图像内容的分类、推荐、管理等精细化运营。例如,从图像大数据中需要识别年龄、性别、人脸个数、人脸遮挡、人脸姿态等,可进行人员识别。这些识别的基础工作都是数据标注。这里对数据标注主要的应用场景进行介绍。

(1) 自动驾驶。利用标注数据训练自动驾驶模型,使其能够感知环境并在很少或没有人为控制的情况下移动。自动驾驶中的数据标注涉及行人识别、车辆识别、红绿灯识别、道路识别等内容,可以为相关企业提供精确的训练数据,为智能交通保驾护航。

(2) 智能安防。数据标注扩大了现有安防系统的感知范围,通过融合各种来源的数据并进行协同分析,提高监控和报警的准确性。其对应的标注场景有面部识别、人脸探测、视觉搜索、人脸关键信息点提取以及车牌识别等。

(3) 智慧医疗。人工智能和大数据分析技术应用于医疗行业,可以深入洞察医学知识和数据,帮助医生和患者解决在医学影像、新药研发、肿瘤与基因、健康管理等领域所面临的影像识别困难、药物研发成本巨大、重症治疗效果不佳等难题。其所涉及的标注场景有手术工具标识、处方识别、医疗影像标注、语音标注等。

(4) 工业 4.0。利用标注数据训练和验证机器人应用程序的计算机视觉模型,从而使模型对工业环境内的各类障碍物、机械设备和机器人有更加精确的感知,实现工业智能机器与所处环境中人和物的安全交互。对应的标注场景有机械手臂导航、仓储码垛、自动分拣或抓取、自动焊接等。

(5) 新零售。将人工智能和机器学习应用于新零售行业,可以通过商品销售数据以及用户的真实反馈促进电子商务的销售,提升用户的个性化体验以及预测客户需求,并实现线上货物推荐的精准化。新零售中涉及的标注场景包括超市货架识别、无人超市系统

和电子商务智能搜索与推荐等。

（6）智慧农业。依托精准的数据标注实现对农作物的定位以及对其成熟度和生长状态的识别，实现农作物智能采摘并解决精准农药喷洒问题，从而减少人力消耗并提高农药利用率。目前，智慧农业中有关数据标注的场景有栽培管理、精准水肥和安全监测等。

5.1.3　数据标注的任务

常见的数据标注任务包括分类标注、拉框标注、区域标注、描点标注等。

1. 分类标注

分类标注是从给定的标签集中选择合适的标签分配给被标注的对象。通常，一个图像可以有很多分类/标签，如运动、读书、购物、旅行等。对于文字，又可以标注出主语、谓语、宾语、名词和动词等。此项任务适用于文本、图像、语音、视频等不同的标注对象。以图像的分类标注为例，标注者需要对图像中的不同对象加以区分和识别。Adobe Stock 是 Adobe 公司的一个旗舰产品，它是精选的高质量图库。该图库规模惊人：拥有超过 2 亿项的资产（包括超过 1500 万个视频、3500 万个向量、1200 万个媒体资产、1.4 亿张照片、插图、模板和 3D 资源）。每一项资产都需要尽可能被推送到客户面前。使用数据标注方法，提供精确的训练数据创建模型，这些训练数据帮助 Adobe 公司为其庞大的客户群提供最有价值的、最符合需求的图像。用户无须通过浏览筛选图像，用户想要的图像会主动推送到用户面前。

2. 拉框标注

拉框标注就是从图像中选出要检测的对象，此方法仅适用于图像标注。拉框标注可细分为多边形拉框和四边形拉框两种形式。多边形拉框是将被标注元素的轮廓以多边形的方式勾勒出来。

不同的被标注元素有不同的轮廓，除了同样需要添加单级或多级标签以外，多边形标注还有可能会涉及物体遮挡的逻辑关系，从而实现细线条的种类识别。四边形拉框主要是用特定软件对图像中需要处理的元素（例如人、车、动物等）进行一个拉框处理，同时，用一个或多个独立的标签代表一个或多个需要处理的元素。图 5.1 为道路施工中的交管人工智能辅助摄像头所摄图像，对图像中的车辆进行了四边形拉框标注，生成供人工智能系

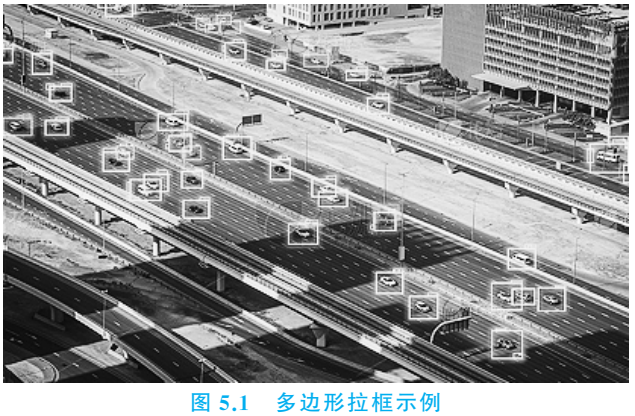

图 5.1　多边形拉框示例

统使用的训练数据。

3. 区域标注

与拉框标注相比，区域标注的要求更加精确，而且边缘可以是柔性的，并仅限于图像标注。其主要的应用场景包括自动驾驶中的道路识别和地图识别。在地图识别中，区域标注的任务是在地图上用曲线将城市中不同行政区域的轮廓勾勒出来，并用不同的颜色加以区分。

4. 描点标注

描点标注是指将需要标注的元素（例如人脸、肢体）按照满足需求的位置进行点位标识，从而实现特定部位关键点的识别。例如，图 5.2 采用描点标注的方法对人脸进行了描点标识。

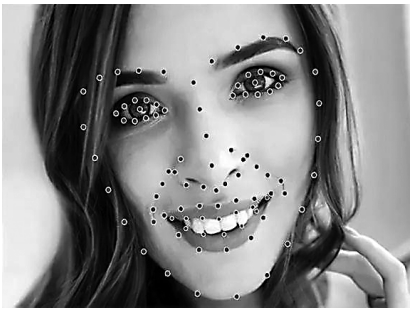

图 5.2　描点标注示例

5. 其他标注

数据标注的任务除了上述 4 种以外，还有很多个性化的标注任务。例如，自动摘要是从新闻事件或者文章中提取出最关键的信息，然后用更加精练的语言写成摘要。自动摘要与分类标注类似，但两者存在一定差异。常见的分类标注有比较明确的界定，例如在对给定图像中的人物、风景和物体进行分类标注时，一般不会产生歧义；而进行自动摘要时需要先对文章的主要观点进行标注，相对于分类标注来说，在标注的客观性和准确性上都没有那么严格，所以自动摘要不属于分类标注。

5.2　数据标注的流程及工具

5.2.1　标注流程

以众包模式下的数据标注为例，蔡莉等提出了一个完整的数据标注流程。首先从标注数据的采集开始，采集的对象包括视频、图像、音频和文本等多种类型和多种格式的数据。由于采集到的数据可能存在缺失值、噪声数据、重复数据等质量问题，故首先需要执行数据清洗任务，以便获得高质量的数据，然后对清洗后的数据进行标注，这是数据标注流程中最重要的一个环节。在具体流程中，管理员会根据不同的标注需求将待标注的数据划分为不同的标注任务。每个标注任务有不同的规范和标注点要求，并且一个标注任务会分配给多个标注员完成。标注员完成标注工作后，将相关数据交给模型训练人员，后

者利用这些标注好的数据训练出需要的算法模型。标注数据的质量主要由审核员检验；审核员进行模型测试，并将测试结果反馈给模型训练人员；而模型训练人员不断地调整参数，以便获得性能更好的算法模型。如果经过参数调整后不能得到最优的算法模型，则说明已标注的数据不满足需求。这时，审核员就会向标注员反馈数据问题，标注员则需要重新标注数据。最后，审核员将最优模型指标发送给产品评估人员使用，并进行上线前的最后评估。

5.2.2　标注内容

无论是开源的标注工具还是商用的数据标注平台，都至少要包含以下内容：

（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。指需要标注的对象，可以根据标注形式进行设计。标注形式一般可以分为单个标注（指对一个对象进行标注）和多个标注（指对多个对象进行标注）两种。

（3）数据导入、导出功能。

（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的数据，检测标注工作的可靠性。

数据的高质量体现在两方面：一是标注的数量多，二是标注的质量高。

（1）图像标注的质量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一像素以内。

（2）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一个语音帧以内。若误差大于一个语音帧，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。

（4）情感标注的质量标准是对标注句子的情感分类级别正确。

5.2.3　标注工具

通常，商用的数据标注工具一般由众包平台提供。数据标注众包平台最早出现在美国，除了亚马逊公司的 Mechanical Turk 平台外，还有 Figure-eight、CrowdFlower、Mighty AI 等初创型标注平台。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也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。这些商业的数据标注平台基本上都能对图像、视频、文本和语音等数据进行标注，但业务方向各有侧重，有的以处理图像见长，有的则更长于视频标注。

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（例如 COCO、Pascal VOC、JSON 等）。常用的开源数据标注工具见表 5.1。

表 5.1 常用的开源数据标注工具

名　称	简　介	运行平台	标注形式
LabelImg	图像标注工具	Windows、Linux、macOS	矩形框
LabelMe	图形界面的标注工具，能够标注图像和视频	Windows、Linux、macOS	多边形、矩形、圆形、多段线、线段、点
RectLabel	图像标注	macOS	多边形、矩形、多段线、线段、点
VOTT	基于 Web 方式本地部署的标注工具，能够标注图像和视频	Windows、Linux、macOS	多边形、矩形、点
LabelBox	适用于大型项目的标注工具，基于 Web，能够标注图像、视频和文本		多边形、矩形、线、点、嵌套分类
VIA	VGG 的图像标注工具，也支持视频和音频标注		矩形、圆、椭圆、多边形、点和线
COCO UI	用于标注 COCO 数据集的工具，基于 Web 方式		矩形、多边形、点和线
Vatic	有目标跟踪能力的视频标注工具，适合目标检测任务	Linux	
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	
DeepDive	非结构化文本标注工具	Linux	
Praat	语音标注工具	Windows、UNIX、Linux、macOS	
精灵标注助手	多功能标注工具	Windows、Linux、macOS	矩形、多边形和曲线

表 5.1 中的数据标注工具除了 COCO UI 和 LabelMe 工具在使用时需要 MIT 许可外，其他工具均为开源使用。大部分开源工具可以运行在 Windows、Linux、macOS 系统上，仅有个别工具是针对特定操作系统开发的，而且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）能够同时标注图像、视频和文本。

5.3 数据标注实例——情感分析

5.3.1 情感分析概述

随着电子商务、社交网络和移动互联网的蓬勃发展，互联网上出现了大量带有情感色彩的文本数据。针对文本数据的情感分析，能够帮助政府及企业更好地理解用户的观点，并及时解决出现的各类问题。因此，情感分析广泛应用在舆情管控、商业决策、观点搜索、信息预测和情绪管理等场景。词语、句子和文章是文本情感分析的 3 个级别：词语级别的情感分析用来确定词语的情感倾向方向和强度；句子级别的情感分析先对句子进行命名实体识别和句法分析，再采用基于词典和机器学习的方法对句子进行情感分析；文章级的情感分析则是分析文章段落的情感倾向方向。情感倾向是主体对某一客体主观存在的

内在评价的一种倾向。它由情感倾向方向和情感倾向度衡量。情感倾向方向也称为情感极性。在情绪文本中，情感倾向方向是用户对客体表达的观点和态度，即支持（正面情感）、反对（负面情感）、中立（中性情感）；情感倾向度是指主体对客体表达正面情感或负面情感时的强弱程度，不同的情感程度往往通过不同的情感词或情感语气等体现。在情感倾向分析研究中，通过对每个情感词赋予不同的权值区分情感倾向度。

5.3.2　情感分析中的数据标注

情绪文本的分析和挖掘涉及文本数据标注中的多项任务，下面对这些任务进行阐述。

1. 中文分词

中文分词是将一个汉字序列切分为一个个单独的词，这是汉语文本处理的基础。例如，要判断下面的句子的情感：

今天是国庆节，可是我们还要加班。

首先要将其切分为一个个词。如果采用自动分词，其结果为

今天/是/国庆节/，/可是/我们/还/要/加班/。

如果采用基于字标注的分词方法，则其结果为

今/B 天/E 是/S 国/B 庆/M 节/E，/S 可/B 是/E 我/B 们/E 还/S 要/S 加/B 班/E。/S

其中，B 表示词首，M 表示词中，E 表示词尾，S 代表单独成词，它们形成了 4 个构词位置。

2. 词性标注

词性标注是将词划分为对应的语法分类，以表达这个词在上下文中的作用。词性主要有名词、动词、形容词、量词、代词、副词、连词、助词等。

3. 情感标注

上面的句子中并没有明确表示情绪的词，不过联系上下文可知，句子表达的情绪是“低落”。为了判断句子所表达的情绪，可以使用一些中文情感极性词典进行分析，例如来源于知网的情感极性字典。但是如果只依靠中文情感极性词典，计算机就很难准确判断句子所反映的真实情绪。因此，事先要采用人工标注的方法对一些带有情绪的语句进行情感标注。通常，人类的基本情绪可以划分为 6 种，即快乐、愤怒、悲伤、恐惧、惊讶和嫉妒。为了正确识别情绪，每一类情绪都要有对应的标注数据，然后利用这些带情绪标注的数据集训练情绪分类模型。情绪分类算法可以采用 KNN、支持向量机、深度置信网络和长短时记忆网络等实现。一旦情绪分类模型训练成功，就能准确地识别句子所表达的情绪。

数据标注工具需要从只支持人工标注逐渐转换为人工标注＋人工智能辅助标注的方法。其基本思路为：基于以往的标注，可以通过人工智能模型对数据进行预处理，然后由标注人员在此基础上做一些校正。以图像标注为例，标注工具首先通过预训练的语义分割模型处理图像，并生成多个图像片段、分类标签及其置信度分数。置信度分数最高的片段用于对标签的初始化，呈现给标注者。标注者可以从机器生成的多个候选标签中为当前片段选择合适的标签，或者对机器未覆盖的对象添加分割段。人工智能辅助标注技术的应用能够极大地降低人力成本并使标注速度大幅提升。目前，已经有一些数据标注公

司开发了相应的半自动化工具，但是从标注比例来看，机器标注占 30%左右，而人工标注占 70%左右。因此，数据标注工具的发展趋势是开发以人工标注为主、以机器标注为辅的半自动化标注工具，同时减小人工标注的比例，并逐步提高机器标注的比例。

人工智能数据标注的终极目标是让人工智能自主学习、自主标记，而不依赖于人类对人工智能的标注与训练。斯坦福大学通过一种编程方式生成训练数据的弱监督范式，并开发了基于弱监督编程范式 Snorkel 的开源框架。将其应用于多任务学习(Multi-Task Learning，MTL)场景，解决了为一个或多个相关任务提供噪声标签的问题。

未来，期待人工智能能够反向作用于数据标注产业，使得人工标注逐渐转变为自动化标注。