

人造数据“喂养”AI 催生亿级市场

AI 潮水汹涌,数据正成为一门火爆的生意。

为了给人工智能喂上充足的“养料”,从发掘、采集到标注,企业在数据处理的各个环节掘金。到如今,真实的数据已无法满足日渐膨胀的AI“胃口”,企业开始探索AI自产自销的“假”数据——合成数据产业应用而生。

上个月底,国内合成数据公司“光轮智能”宣布完成天使+轮融资;几个月前,新加坡合成数据初创公司 Betterdata 也获得一笔 165 万美元规模的种子轮融资。互联网大厂也开始了布局。微软、英伟达、meta、亚马逊等数得上号的科技巨头中,均有合成数据相关的业务布局、投资或收购举动。

合成数据究竟是“何方神圣”?它有这样的产业价值和风险?会给AI产业带来怎样的颠覆?

“人造”数据崛起

相比于从现实世界中采集或测量的真实数据,合成数据顾名思义是人工合成的“假”数据。由于能够反映原始数据的属性,合成数据可以作为原始数据的替代品来训练、测试和验证AI模型。

但人工合成并不意味着完全凭空捏造。现阶段,大部分合成数据的“根”仍然是真实数据。

Unity中国高级软件工程师钱文亿向记者介绍了其合成数据产品在计算机视觉相关项目中的普遍生成过程:第一步,在现实中找到可识别的对象,通过扫描技术,将物体模型真实地还原在3D场景中;在此基础上,对该物体模型进行打标签,如颜色、大小等,具体标签类型依据训练需求而定;最后,将这些物体放置于各种设定的场景中,随机组合,快速地生成多张图片。

因此,训练同一个AI模型时,使用真实数据也许需要摄像头不断变换地捕捉物体在不同场景、状态下的多张照片,而合成数据则能够通过调整物体位置、角度、所处背景等参数,一分钟内生成上百上千张不同的图片,降低成本,提高数据集生成效率。

事实上,合成数据的概念并不新颖。据说,这一概念早在1993年Donald Rubin的一篇文章中就有雏形。近年来,随着人工智能技术一次次取得突破性发展,真实数据的采集、获取难度也水涨船高,已难以填饱AI训练的庞大“胃口”。

合成数据常常作为真实数据的“平替”而存在。据人工智能初创公司Cohere首席执行官Aiden Gomez在上个月月底透露,由于Reddit、推特等公司的数据采集要价太高,微软、OpenAI和Cohere等公司,已使用合成数据来训练AI模型。Gomez表示,合成数据可以适用于很多训练场景,

只是目前尚未全面推广。

但在广州大学计算机科学与网络工程学院教授王员根看来,价格反而不是选择合成数据最主要的考虑因素。

真实数据涉及大量个人隐私,冒然使用可能引起严重的法律问题,而且并非所有的真实数据都是可用的。互联网上充斥着大量真伪难辨的信息,要从杂乱无章的真实数据中挖掘出可用的信息,需要大量的工人筛选。另外,真实数据还存在分布不均衡的问题。例如,训练人脸识别系统时,从互联网上爬取到的人脸数据中亮皮肤人脸图像占多,而暗皮肤人脸图像偏少,这将导致所训练的模型存在偏见。合成数据恰能在一定程度上人为规避上述问题。

“部分真实数据无法获取,如清晰的水下图像等,通过合成数据技术模拟生成相关数据,能够补充训练数据的完备性。”王员根补充道,尽管现阶段大量合成数据建立在真实数据基础上,但随着技术的进步,未来对真实数据的依赖将逐步减少,目前已有技术能让直接合成的数据“以假乱真”。

但合成数据并非十全十美。在AI训练数据服务商Appen澳鹏官方发布的一篇文章中,就提到合成数据缺乏异常值,而这些异常值自然出现在真实数据中,对于模型精确度至关重要。另外,合成数据的质量通常取决于用于生成的输入数据,输入数据中的偏见很容易传播到合成数据中,因此不能低估使用高质量数据作为起点的重要性。所以,企业需要将合成数据与人工标注的真实数据进行比较,作为额外的输出控制。

合成数据应用多领域突破

目前,合成数据主要应用于哪些领域?

相比于自然语言、音频等形式,合成数据最先在计算机视觉上展露拳脚。受访专家们认为,这与图片处理更加简单直接、人类与环境进行交互时优先通过视觉系统等因素相关。未来,其他领域的合成数据也将得到进一步的发展。

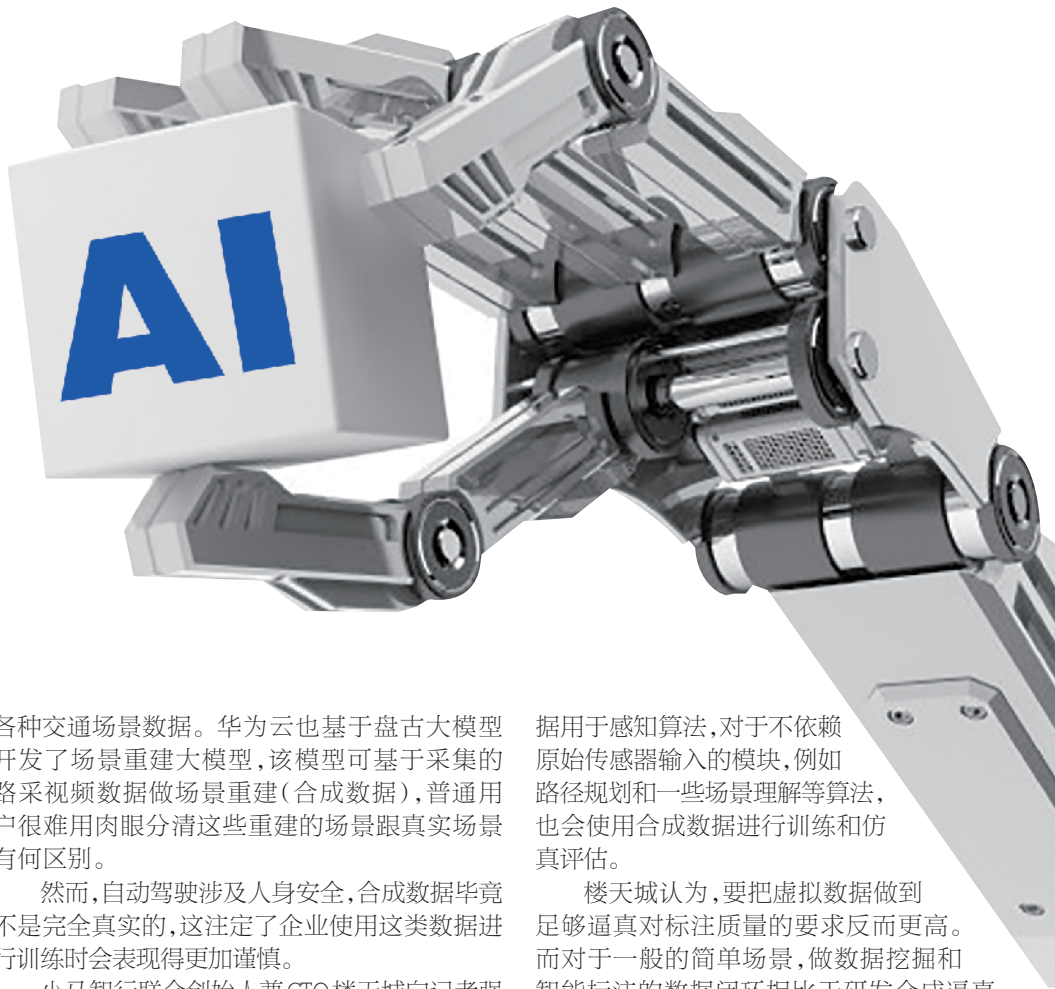
合成数据在自动驾驶、医疗、金融等场景有着广阔的应用前景。这些场景的共同点在于,真实数据敏感,难以获取,但又关乎重大,有的还涉及人身安全,对数据质量要求极高。“哪里最有需要,哪里就会最先得到发展和应用。合成数据技术最有可能在这些敏感场景中取得突破。”王员根表示。

以自动驾驶为例,实际驾驶过程中,车辆可能会碰到各种复杂多变的路况,甚至是极端情况,如严重的交通堵塞、事故、恶劣天气等。尤其是在极端情况下,使用真车冒险测试几乎不可能,极难采集和获取到真实数据。

合成数据可以模拟出这些情景。王员根介绍,“比如,要模拟暴雨天气,我们就用日常能够收集到的普通天气的数据,构建一个物理或网络模型,将‘暴雨’的关键参数输入进去,就能生成相应的场景。模型和参数越准确,场景的逼真程度越高。”如此,能够在保障人员和设备安全的条件下,提升自动驾驶能力。

公开资料显示,许多自动驾驶汽车厂商都在合成数据和模拟方面进行了大量投资。例如,谷歌母公司 Alphabet 旗下的自动驾驶子公司 Waymo 在 2106 年就生成了 25 亿英里的模拟驾驶数据来训练其自动驾驶系统(相比之下,从现实世界收集的驾驶数据仅为 300 万英里)。到 2019 年,这一数字已达到 100 亿英里。

国内,腾讯自动驾驶实验室开发的自动驾驶仿真系统 TADSim 已经可以自动生成无需标注的



各种交通场景数据。华为云也基于盘古大模型开发了场景重建大模型,该模型可基于采集的路采视频数据做场景重建(合成数据),普通用户很难用肉眼分清这些重建的场景跟真实场景有何区别。

然而,自动驾驶涉及人身安全,合成数据毕竟不是完全真实的,这注定了企业使用这类数据进行训练时会表现得更加谨慎。

小马智行联合创始人兼CTO楼天城向记者强调,合成数据既有凭空生成的虚拟数据,也有基于真实数据加以修改得到的数据,目前在L4的感知模块中,小马智行没有使用凭空生成的虚拟数据。主要是因为L4方案依赖于激光雷达,对于如恶劣天气、长尾物体等难度场景,生成激光雷达的虚拟数据与真实数据的分布差异较大,无法用虚拟数据来达到在真实场景下提升的效果。

但小马智行会对真实数据加以修改来合成数

据用于感知算法,对于不依赖原始传感器输入的模块,例如路径规划和一些场景理解等算法,也会使用合成数据进行训练和仿真评估。

楼天城认为,要把虚拟数据做到足够逼真对标注质量的要求反而更高。而对于一般的简单场景,做数据挖掘和智能标注的数据闭环相比于研发合成逼真的虚拟数据的成本还要低不少。目前学术界对使用完全虚拟的数据进行自动驾驶的训练有一些研究,不少公司也在做相关预研。从训练效果来看,从0到80分有帮助,但对90到99分效果一般来,实际落地部署的案例并不普遍。

“我们也在关注合成虚拟数据相关的技术进展并持开放的态度,如果某一天技术足够成熟时也会考虑应用。”楼天城表示。

数据标注产业将被重构？

据咨询公司Gartner预测,到2030年,合成数据将彻底取代真实数据,成为AI模型所使用的数据的主要来源。而美国AI研究机构Cognilytica的数据显示,2021年合成数据市场规模大概在11亿美元,到2027年将达到11.5亿美元。这是一块让不少科技大厂和初创公司垂涎的大蛋糕。

多家的科技巨头均有合成数据相关的业务布局、投资或收购举动。例如,2021年,英伟达发布了用于训练AI的Omniverse Replicator合成数据生成引擎,今年7月,英伟达初创加速计划成员Rendered.ai将Omniverse Replicator集成到其合成数据生成平台,使AI训练变得更加简单易用;亚马逊也在多个场景探索合成数据的应用,例如使用合成数据来训练、调试其虚拟助手Alexa,以避免用户隐私问题等;Meta则直接收购了合成数据创业公司AIReverie,以整合至旗下元宇宙部门Reality Labs。

创业公司方面,合成数据领域的投资并购持续升温。计算机视觉合成数据提供商Datagen于2022年初宣布完成5000万美元B轮融资;今年4月,新加坡合成数据初创公司Betterdata在获得一笔165万美元规模的种子轮融资;7月末,国内合成数据公司“光轮智能”宣布完成天使+轮融资,这家公司刚成立的新公司,已经完成种子轮、天使轮、天使+三轮融资,累计融资金额达数千万元。

钱文亿观察到,“过去几年,全球几乎每年都有好几百甚至上千家新型初创公司建立,为各行各业提供用于算法训练的合成数据产品。”

产业繁荣之下,我国也开始鼓励和引导合成数据产业的发展。今年3月初,中国证监会科技监管局局长姚前曾在《中国金融》杂志撰文称,建议重点发展基于AIGC技术的合成数据产业。以更高效率、更低成本、更高质量为数据要素市场“增量扩容”,助力打造面向人工智能未来发展的数据优势。5月19日,北京发布的“北京市通用人工智能产业创新伙伴计划”,提到谋划建设国家级数据训练基地,也提出了支持发展基于AIGC技术的合成数据新产业。

而在过去很长一段时间里,人工智能对数据的海量需求催生了一批数据标注师的就业群体。如今,当合成数据逐渐成势,数据标注产业的格局是否会因此遭到冲击?

王员根认为,冲击是必然的,但需求仍在。“首先这件事没有那么快到来,其次,标注师们要进行转型。比如,过去需要标注的是原始数据,如今则变成了AI生成的数据;又如在原本数据标注工作的基础上,标注师们还将被要求分辨哪些是AI生成的数据,哪些是自然数据等。即使合成数据越来越多、质量越来越高,也离不开人类的引导和监督,以及修正可能出现的偏差。”

供稿:《21世纪经济报道》

中国（青岛）新媒体基地

China (Qingdao) New Media Development Zone

中国(青岛)新媒体基地
China(Qingdao) New Media Development Zone

办公场地
类型多样

配套设施
功能完善

“管家式”
贴心服务

创业福地 活力社区

📍 山东省青岛市市南区徐州路77号

☎ 0532-58559877 18669858311