中文大模型让AI更“接地气”

目前成熟的生成式AI模型大多基于英文数据进行训练,在国内各行各业的应用环境中,中文大模型显然更“接地气”。通过中文或英文数据训练出来的大模型,差异比较大,中文的上下文理解和语义的多解性要大于英文。大模型首先要理解人类意图,因此对于国内用户来说,用中文去训练的大模型比较适用。

“请讯飞星火认知大模型模仿梁晓声先生笔下的小说《人世间》的风格,续写一小段文章。”5月20日,在第七届世界智能大会闭幕式上,主持人蒋昌建向讯飞星火认知大模型发问。短短几秒钟,续写文字便“跃然纸上”。原作者梁晓声认为,续写内容简练、文字有一定温度,从传达的情感和思想来看几乎“无可挑剔”。

在本次世界智能大会上,生成式人工智能毫无疑问成为大家关注的焦点。近期,国内各大厂商纷纷加快开展生成式AI核心技术的研发,无论是讯飞星火认知大模型展现出的雄厚“中文功底”,还是国家超级计算天津中心发布的基于国产天河超级算力、智能算力和汇集构建中文大数据集研发训练的天河天元大模型,都让大众对我国自主研发的中文生成式AI大模型充满期待。

开发适合国人的中文大模型

“AI大模型是基于海量多源数据打造的预训练模型,是对原有算法模型的技术升级和产品迭代。”国家超级计算天津中心数据智能部部长康波介绍,预训练大模型在基于海量数据的自监督学习阶段完成了“通识”教育,再借助“预训练+精调”等模式,在共享参数的情况下,根据具体应用场景的特性,用少量数据进行相应微调,即可高水平完成任务。

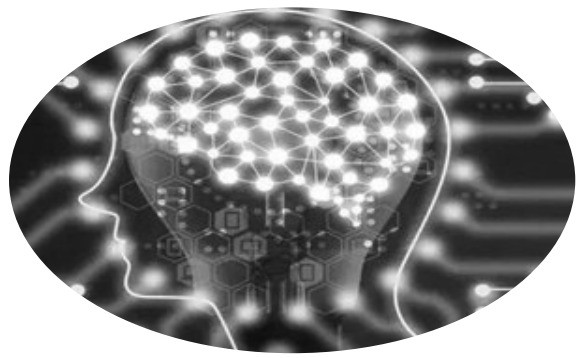
AI大模型能够理解人类的自然语言表达,并通过庞大的网络结构实现具有针对性的内容输出。

从效果上看,生成式AI表现为“无所不知、无所不能”,其具备了逻辑推理、上下文理解、文字创作、知识提取、代码生成等非常多元化的强大能力。

不过,目前成熟的生成式AI大模型大多基于英文数据进行训练。“通过中文或英文数据训练出来的大模型,差异还是比较大的,中文的上下文理解和语义的多解性要大于英文。大模型首先要理解人类意图,因此对于国内用户来说,用中文去训练的大模型比较适用。”康波说。

此外,生成式AI正一步步向生产工具方向发展,为产业深度赋能,或将成为人工智能与实体经济深度融合的重要力量。那么作为数据驱动的AI大模型,其训练数据来源的可靠性和安全性,便成为推动科技创新的关键。因此,自主研发中文大模型成为越来越多科技巨头的首要选择。

三月以来,国内大模型领域已进入“混战”模式,各路玩家纷纷入局,其中有不少都“相中”了研发中文大模型。



“抢抓通用人工智能的发展机遇有几个基本要素。”科大讯飞董事长刘庆峰认为,第一,必须要在自主可控的平台上;第二,必须要同时做中文和英文,不只学习中国的“智慧”,还要向世界学习;第三,在“硬碰硬”的科技对比上,不仅要学习,还要想办法赶超。

例如,阿里推出了首个中文AI模型社区,社区首批上架超300个模型,其中中文模型超过100个,覆盖了视觉、语音、自然语言处理、多模态等AI主要领域,覆盖主流任务超过60个,且均全面开源并开放使用。360公司推出的“360智脑”背后的360GPT大模型,在海量的中文文本数据上进行了预训练和微调,从而具备了强大的语言理解和生成能力。据悉,该模型目前已经达到了100亿参数规模,并且还在不断扩展中。

中文大语言模型数据集稀缺

生成式人工智能是人工智能发展到一定阶段的产物。就像ImageNet数据集推动了残差网络等计算机视觉算法的成熟,openslr等开源数据集的发布催生了长短期记忆神经网络等自然语言神经网络的发展,图形处理器的大量使用使得模型参数从百万级发展到千亿级(ChatGPT使用了上万块A100显卡开展训练)。可以看出,生成式人工智能的快速成长,离不开算力和数据的支撑。

“大模型是大数据、大算力驱动的结果,两者缺一不可。”超级计算天津中心首席科学家孟祥飞博士强调。

一方面,中文大模型的理解能力来自于数据,它需要用海量数据来学习,通过自注意力和多头注意力机制来建立知识之间的联系。这就意味着,更多、更高质量的数据供给,将会带来模型网络中知识之间关系的完善性和贯通性。当用户提问到深层次或者冷门问题时,数据质量越高,AI大模型回答出正确答案的概率就越大。

“但目前中文大语言模型的数据集非常稀缺。”孟祥飞介绍,为了解决这个问题,天津超算中心搜集整理了全域的网页数据,并从中提取处理高质量的中文数据做成数据集,同时采集纳入各种开源训练数据、中文小说数据、古文数据、百科数据、新闻数据以及专业领域的诸如医学、法律等多种数据集,训练数据集总token数达到3500亿,训练打造了中文语言大模型——天河天元大模型。

另一方面,算力的供应是大模型的基础保障。大模型发端于自然语言处理领域,以谷歌的BERT、Open AI的ChatGPT和百度文心一言等大模型为代表,参数规模逐步提升至千亿、万亿,同时用于训练的数据量级也显著提升,带来了模型能力的提高,这也代表着算力需求的指数级上升。

“而超级计算可以说是算力中的战斗机。”孟祥飞说,为了保证大模型的训练顺利,天津超算中心充分利用了天河新一代超级计算机的双精度、单精度、半精度融合计算输出能力,构建基于自主E级算力体系架构的智能计算引擎,建设人工智能大规模训练与应用系统支撑环境,特别是在中文处理方面构建了中文大模型数据处理的工作流技术体系,从而保障了训练任务的顺利开展。

技术成果广泛应用于多领域

在此次世界智能大会上,随着讯飞星火认知大模型一起展示的还有多款搭载了大模型的行业应用成果。

康波认为,人工智能是驱动新一轮科技革命和产业变革的巨大力量,应将大模型作为产业智能化升级的基座,用专业数据集打造更贴合行业领域的智能化高水平“专家”。

以讯飞星火认知大模型为例,该大模型的整体布局为“1+N”体系。其中“1”是指通用认知智能大模型,“N”就是大模型在教育、办公、汽车、人机交互等各个领域的应用。例如在教育领域,作为全球首款搭载认知大模型的学习机,科大讯飞推出的学习机可像真人教师一样与3岁至18岁的学生进行互动式辅学;在办公领域,基于大模型能力升级的产品具备语篇规整、会议纪要、一键成稿等功能。

康波认为,在各行各业的应用中,中文大模型显然更“接地气”。他举例说,天津超算中心综合实现了文本、语音、视频等多模态的大模型生成能力,从而形成了“一平台三能力”的基础架构,实现了更广泛的产业融合能力。基于其自然语言的理解和表达能力,与医疗结合,学习医学指南等专业规范,可以迅速地掌握对应的专业知识。其中,中文大模型可以解决“同词不同义”在医疗上的歧义性,实现精准的输出,为医疗辅助诊断提供更为全面的支撑能力。

同样,在工业检测和流程控制方面,大模型基于多元化输出能力,可以进行规范辅导、缺陷检测、流程指令生成一系列操作,降低错误率,提升生产效率。其中,中文大模型可以更好地理解复杂的专业术语以及流程指令逻辑,让输出更准确、严谨。

“在大模型通用性、泛化性以及降低人工智能应用门槛的优势推动下,人工智能也将会加快落地,形成新的机遇。”康波表示。 来源:科技日报

科技让博物馆有一颗现代的心

故宫文物的数字化采集与利用又有新进展。国际博物馆日当天,故宫博物院向社会发布2万件院藏文物高清数字影像。而截至目前,“数字文物库”文物总数超过10万件,浏览量超3300万次,是故宫博物院官网上最受公众欢迎的数字产品。同日,依托互联网公司数字孪生、虚拟演播、云音视频创作等下一代互联网技术,“故宫·腾讯联合创新实验室”正式成立。创新实验室将一体化采集文物的多维度数据,加速文物数字资源采集、加工、展示的全流程智能化管理,提升数字化质量和流程效率。

来自国际博物馆协会的报告显示,新冠疫情加快了古老文博拥抱前沿科技的速度,与疫情之前相比,博物馆的线上活动明显增加,主要体现在线上藏品、线上展览、活动线上直播和社交媒体的使用上。

近日,公众在手机上登录“云游敦煌”小程序,就可以进入高清还原的数字藏经洞中,近距离观赏洞窟里的壁画、彩塑和碑文等细节。不仅如此,公众还可以通过人物角色的扮演,“穿越”到不同历史节点,与多位历史人物展开互动,“亲历”藏经洞的前世今生。三星堆博物馆再次“上新”,通过三维扫描技术将3号坑出土的顶尊跪坐人像与8号坑出土的青铜神兽成功“拼对”,让人们在跨越千年之后重见文物合体的模样。

博物馆作为承载、传播文化历史的主要平台和工具,成为了解一座城市、一个地域乃至一个国家历史文化的“百科全书”。在裸眼3D、全息投影展示、VR虚拟现实、体感互动等数字化技术的加持下,博物馆在文化传播与共享、增强公众互动性和体验感、提升公共服务效率等方面都有了长足的进步,前沿科技与传统文化在博物馆上实现了深度的融合。

数字化的多年实践,使得数字化与博物馆已经超越了体与用的关系。古人将历史写在竹简上、写在丝绸上、写在纸张上,但所有的当代史写在数据中。勒石未必长存,铸鼎未必传世,今天不管是历史的记录方式、保存方式甚至记录的深度与广度,都已经彻底改变了。一些在文博领域世代恪守的准则被打破,灰飞烟灭不再成为必然,永久保存、永续利用成为现实。

但博物馆的数字化意义仅在于此吗?我们是否可以展望更深远的未来?在几年以前,我们曾感慨科技发展让古老文博有了一张现代的脸,而今天数字化则让博物馆有了现代的心、数据的心,反而是线下的展品变为了数据的具身。从内到外,焕然一新的博物馆将带给人们什么样的惊喜与体验,让人期待。

这样的变化不是孤证。以文学与影视这对关系为例,随着影视媒介的崛起和视觉文化的兴盛,文学的审美形态发生了变化,由阅读带来审美想象转变为凸显视觉感官,从而反过来使得文字具有更鲜明的视觉冲击力。一些影视上使用的方法,促使文学创作特别是小说创作不断进行内在的调整,蒙太奇等创作思维和视听语言等叙事手段被嫁接、融合到小说的写作中,文学具有了鲜明的影像化书写的风格,文学在影视艺术的推动下,从平面的状态转变为动态的、纵向的、充满画面感、参与感、融入情节中的、多元的、超维度的阅读方式。这样互相影响、互相增长的变化也正发生在文博数字化的过程中。 来源:光明日报