新框架挑战预训练模型

自然语言处理有望“另辟蹊径”

■本报记者 张双虎

预训练模型的兴起给自然语言处理(NLP)带来了“新面貌”。

近年来,谷歌、OpenAI、微软、百度等人工智能(AI)“头部玩家”推出多个颇具影响的预训练模型,并反复迭代出十多个版本。无论学术界还是产业界,人们对大规模预训练模型“热情高涨”。

日前,来自清华大学的一支研究团队提出一种简单高效的NLP学习框架。不同于当下NLP社区主流的“大规模预训练+下游任务微调”的范式,这一框架无需进行大规模预训练,同时将训练效率提升两个数量级,并在多个NLP任务上,实现了比肩甚至超出预训练模型的性能。近日,相关研究以预印本形式在arXiv上发表。

深度学习需要越来越大的数据集,用于充分训练模型参数。 图片来源:unsplash

预训练模型的“内功”

预训练模型在自然语言处理领域蓬勃发展,近年来在多个子方向取得了颠覆性的成果。

“自然语言处理的‘预训练’过程,就像武侠小说中,练武之人‘修炼内功’。”上海对外经贸大学副研究员邵浩说,“一个人要成为武林高手,需要有扎实的‘内功’,内功修炼好后,再去学各种招式就非常容易上手,并能发挥其最大效用。”

随着深度学习的发展,模型参数显著增长,从而需要越来越大的数据集,用于充分训练模型参数。然而,因大部分NLP任务的标注成本极为高昂,尤其是句法和语义相关的任务,构建大规模标注数据集十分困难。

相比较而言,大规模无标注数据集易于构建。为更好地利用海量无标签文本数据,常规做法是首先从这些数据中学到较好的文本表示,然后将其用于其他任务。许多研究表明,在大规模无标注语料中训练的预训练语言模型,可以使多方面NLP任务获得显著的性能提升。

通过海量无标注语料来预训练神经网络模型,可以让AI更利于下游NLP任务的完成。预训练模型的作者已经设计出了基准模型,这样,使用者就可以在NLP数据集上应用该模型,而无需从头开始构建模型来解决类似的问题。尽管后续过程需要进行一些微调,但这为人们节省了大量的时间和计算资源。

2018年,无监督的双向预训练语言模型ELMo被提出,这种上下文相关的文本表示方法在多个典型任务上表现惊艳,能有效处理一词多义问题。紧随其后,GPT、BERT等预训练语言模型相继被提出,预训练模型技术开始在NLP领域大放异彩,并在各种下游任务中遍地开花。

任务驱动模型出场

“预训练语言模型因其强大的性能被广泛关注,基于‘预训练—微调’的范式也成为许多NLP任务的标准方法。”清华大学交叉信息研究院助理教授、RecurrentAI联合创始人杨植麟对《中国科学报》说,“然而,当前通用语言模型的预训练成本极其高昂,这使得只有少数资源充足的研究机构或组织才能够对其展开探索。”

为解决上述问题,杨植麟团队提出一种完全不需要预训练语言模型的高效学习框架。这一框架从通用语料中筛选出与下游任务相关的子集,并将语言建模任务与下游任务进行联合训练。

该论文第一作者、清华大学计算机科学实验班(姚班)大四本科生姚星丞介绍说,提出任务驱动的语言模型的想法源于一个基本的观察:人类可以通过对关键信息的学习,在有限的时间和精力投入情况下,快速掌握某一任务技能。例如,在临近考试时,学生仅根据考纲复习浏览若干相关章节的要点即可应对考试,而不必学习所有可能的知识点。与之类似,预训练语言模型在某一下游任务上的优良表现,“很有可能源自于语料中与下游任务相关的数据”。

基于这一判断,该团队提出任务驱动的语言模型(TLM)。它仅利用从大规模通用语料中提取的少量与下游任务相关的数据,就可以取得与全量数据类似的结果。

“相较于传统的预训练模型RoBERTa(基于BERT的改进模型,使用更大的批次和更多的数据对模型进行更长的训练),TLM仅需要约1%的训练时间与1%的语料,即可在众多NLP任务上,表现出比肩甚至超出预训练模型的性能。”姚星丞说,“我们目前也正在尝试将任务驱动的方法推广到更大规模的模型上,如GPT-3或T5。”

跳出预训练范式

为了从大规模通用语料中抽取关键数据,TLM以任务数据作为查询对象,用基于稀疏特征的BM25算法作为召回算法,对通用语料库进行相似数据的召回。

“除已有的下游任务数据以外,其余的语料均通过BM25算法进行相似性匹配与自动筛选,不需要人工做额外的选择与标记。”姚星丞说,“TLM基于任务数据和召回数据,同时优化任务目标和语言建模目标,从零开始进行联合训练。”

为了测试TLM的性能,研究人员在8项NLP分类任务上,从三个不同规模展开了对比实验。这8项任务涵盖了计算机科学、生物医药、新闻、评论4个领域,包括了训练样本数量小于5000的低资源任务和训练样本数量大于20000的高资源任务,任务类型覆盖了话题分类、情感分类、实体关系抽取等。

测试结果显示,和对应“预训练—微调”基准相比,TLM实现了相同甚至更优的性能。平均而言,TLM减少了两个数量级规模的训练计算量以及训练语料的规模。整体来说,预训练模型以极高的成本学习尽可能多的、和任务无关的知识,而TLM以非常低的成本,针对每个任务学习相关知识。

“当我们有少数特定目标的任务需要完成的时候(例如希望对少量几个数据集进行研究),TLM会是非常高效的。”姚星丞说,“而需要一次性完成大量任务时(例如工业界构建一个NLP平台为多方提供相似的服务),预训练模型仍然具有优势。”

此外,TLM是任务驱动的,所以可以给研究人员更大的自由度,以自定义策略进行标记、调整序列长度、数据表示、超参数的调整等等,从而达到提高性能和效率的目的。

“TLM的提出,让NLP研究跳脱出‘预训练—微调’范式成为可能,这有利于推动NLP研究公平化。”杨植麟解释说,预训练本身严重依赖大量的计算资源,这一限制使大多数NLP研究者只能专注于对微调算法的研究。然而,微调算法的性能上限,很大程度上受预训练模型性能的约束。而TLM可以让大多数研究人员以较低的代价和较高的效率,基于最先进的解决方案对模型架构、损失函数、算法等进一步自由探索。

杨植麟认为,未来会有更多有趣的研究在TLM的基础上展开。例如,如何经济地达到更大规模预训练模型的表现效果;如何提升TLM的通用性与可迁移性;可否利用TLM进行小样本或零样本学习等。此外,还可以将预训练模型和TLM结合,从而在通用性和效率之间实现更好的权衡。

相关论文信息:

https://arxiv.org/pdf/2111.04130.pdf

企业数字化转型的发展趋势

■王志英

数字经济业已上升为我国现阶段的发展战略,包括产业数字化和数字产业化。企业信息化隶属于产业数字化,是企业可持续发展甚至转型升级的必然选择。

信息化与数字化是两个相互包容而又相互独立的概念。人们最初把数字化定义为信息化的高级阶段,而现在又把信息化当作数字化的初级阶段。

企业信息化发展进程

企业信息化的发展一般认为有四个阶段:电算时代、管理信息化时代、信息集成时代和数据应用时代。

电算时代起于20世纪80年代,人们开始用计算机程序辅助完成某一业务模块的工作,如财务电算化、人事管理电算化等;随后进入管理信息化(MIS)时代,即针对某一业务域开发相应的软件“系统”,通常是各专业或部门各自开发系统,结果是“烟囱”林立。这一时期创新最为活跃,系统也频繁换代。在信息集成时代,主要是解决各系统之间的流程贯通和数据共享问题。大部分企业选择了被冠以“最佳管理实践”的企业资源规划(ERP)套装软件。进入数据应用时代,人们开始对信息系统中产生的大量数据加以分析,以优化管理、辅助决策,如商业智能(BI)等。

随着云和大数据技术的出现,数据已成为企业生产的重要资源,进而引发各种数字创新甚或数字化转型。这些新技术的效果和影响是如此之大,以至于在媒体上“数字化”逐渐取代了“信息化”。

总之,企业数字化是信息化的延续、发展和提升。如果说标准化是信息化的基础,那么信息化则是数字化的基础。

衡量企业信息化的三大特征

回顾企业信息化建设过程,能按计划上线并正常运行五年以上的项目少之又少。反观成功的信息化项目,皆同时具备三个条件:切合实际的业务、先进适用的技术、高效创新的管理。取其关键词的首字母,可简称为“BTM准则”。

在“BTM准则”中,“切合实际的业务”是信息化项目开发的基础;“选择先进适用的技术”不仅可降低开发成本,更提供了可持续运行的保证;而“高效创新的管理”既涵盖了在IT加持下的业务流程创新与再造,也体现业务与技术深度融合的程度。

这里必须提到企业级信息化。与业务驱动的传统信息化相比,企业级信息化由企业驱动,并在统一技术架构下开发,从根本上解决了企业信息化的三大顽疾:业务集成、数据共享和可持续发展。

如果采用成熟度模型理论,则一个成熟的信息化企业具有三大特征:全业务接入、全信息对称、全数据集中。

“全业务接入”或“线下无业务”是信息化管理的基本目标,即所有的业务和管理过程均可被记录和测量;同时还包括以消灭系统壁垒和“信息孤岛”为目标的全业务集成。

“全信息对称”指作业层有工具、管理层有数据、决策层有依据。信息化不仅能支持班组所站的作业,而且作业结果可同时被整理分析。这需要建设跨业务的运营管控平台。全信息对称可以最大限度地实现“消灭报表”和“减少会议”。

“全数据集中”简言之即“一个企业一个数据中心”,这是指逻辑上企业数据必须集中,以实现数据“边际效益递增”。

ERP的局限性和“EA+SOA”方案

客观地讲,央企信息化在很大程度上是在外部驱动下开展和推进的,并不完全是企业内生需求。而“完成任务”和“解决问题”这两者所导致的组织行为有很大差别。在这种背景下,国外ERP软件在咨询公司所谓“最佳管理实践”的加持下,几乎占领了所有企业。央企信息化大会往往变成一个ERP实施经验交流会。

然而,即使不论ERP在技术上的局限性,采用ERP之后也至少还有四方面的问题。一是,大多数企业的主营业务系统与ERP的集成问题;二是“必须以ERP为核心”往往牺牲企业的总体技术架构;三是ERP要反向适应现行管理体系;四是ERP的可持续性堪忧。

企业架构(EA)作为一种管理方法论,从上世纪80年代开始,已在很多国家得到应用。EA可以全面准确承接企业战略,并使之逐级实施落地。服务化应用(SOA)作为一种软件开发方法,曾一度被认为是与Internet和Web2.0齐观的技术革命。SOA的主要思路是将软件服务化、模块化,通过总线装配完成功能开发。但SOA的实施高度依赖于具体业务分析(WBS),需要业务专家和IT专家密切协同,而专业服务又很难标准化、产品化,因此发展艰难。

然而,EA与SOA的结合几乎是天衣无缝,能彻底解决IT与业务融合、信息集成共享、敏捷开发诸问题,是具有普适性和可持续性的信息化解决方案。

微服务是一种基于容器技术的软件方法,延续了SOA的思路,采用了最新的技术。可以说,微服务代表着软件技术的发展趋势。但必须强调的是,如果没有架构管控或者离开了统一的业务模型,再先进的技术也是缘木求鱼。

电力企业数字化转型的趋势

首先,就如大数据不是数据大,数字化也不是数码化。因此,数字电网并不依赖于每个元件的数码化。数字化的核心是通过数据应用实现企业价值创新,包括产品/服务创新、效率/质量提升、安全风险降低以及客户体验改善等。如果说信息化是企业管理改进的必需品,那么数字化则是企业价值创造的发动机。换言之,数字化项目必须以价值创造为导向。

其次,数字作为新的生产要素,对每个企业都有挖掘价值的空间,但并不意味着每个企业都有数字化转型的命题。企业转型意味着企业发展战略的重大改变或调整,数字化只是途径和手段。

然而,在“双碳”目标下,电力企业面临战略转型的抉择。这是由3方面的趋势所决定的:电源革命、供电市场重构和电网技术创新。海量分布式绿电预计在2030年将取代火电而占主导地位,新型供电主体的出现将导致电力市场规则和运营模式的重构,特别是电网运行和控制技术必须创新。

这些问题的解决很大程度上都依赖于数字技术,加之,电力企业拥有的电力数据具有广泛覆盖和全时空特性,因此电力企业及数字技术实现战略转型既必然又可行。

过去20年,电力企业通过通信与信息融合,重构了信息基础设施;通过业务与信息融合,建成了企业信息化体系;现在将通过自动化与信息融合,构建一个划时代的新型电力系统。但需要认清的是,电力企业数字化转型是一个长期过程,需要在总结经验中不断创新。

(作者系南方电网改革发展研究中心特聘研究员,本报记者赵广立整理)

生物机器人需在可控前提下研究并关注长远伦理风险

■曾毅

近日,美国佛蒙特大学、塔弗茨大学和哈佛大学威斯研究院的研究人员在美国《国家科学院院刊》上发表论文,描述了世界首批生物机器人Xenobot的自我繁衍方式。这一研究引发了人们的关注。

科学的发展在其萌芽期确实有必要慎重、前瞻地考虑其风险。用人类尚未完全了解的生物组织片段构造人造物的隐忧在玛丽·雪莱1818年发表的作品《弗兰肯斯坦》中已有淋漓尽致的展示。虽然由青蛙的表皮细胞和心肌细胞制造的全球首个“活体机器人”Xenobot还只是由生物细胞自组织堆积且没有意识的生物组织,但是类似于这样的尝试如果没有全周期从伦理视角进行更负责任的研发,一旦这样的“人工生物机器”通过自组织涌现出更复杂的功能,未来的隐忧与《弗兰肯斯坦》中的怪物并无二致。

一切的风险来源于不确定性,以及人类本身对人造生物组织的自组织行为与可能产生的后果缺乏足够的了解。传统机器人输入输出行为基本可控,而利用细胞组织进行人工设计并在此基础上进行自组织衍生则极大降低了可控性,此类风险与实验室内人工培育的类脑组织存在类似的风险。而研究人员对以“缸中之脑”形式存在的类脑组织,以及Xenobot这样的人工生物机器人的研究,需要在对其结构、机理、发育和演化机制实现严格的风险可控前提下进行。

生物组织即使结构非常简单,甚至没有神经系统,都可能具有复杂的功能和行为。例如,黏菌能够通过自组织生长实现最优路径的规划,以及根据环境和能量进行高度自适应的运动。目前对于黏菌这样相对简单的生物组织的机理尚未有效揭示,对于进一步通过人工和自组织衍生的生物人造物则更存在不确定性与风险。目前的科学进展并不能确定人造且由人工智能参与设计的具有一定自组织能力的生物机器人,绝对不具有超越人类预期的行为涌现。

Xenobot具有能量代谢功能,能够回应刺激并有一定自主运动能力,因此可被视为人造生命,但这是未经自然演化生成的生命,没有自然的物竞天择。很难想象这样的生命在被应用过程中,可能产生的“试错过程”的代价会是什么。我也很难认同Xenobot在环境中堆积起的新细胞团在遗传的视角能够称为其后代或是自我复制,因为亲代表达相应性状的基因并没有通过亲代传递给后代。事实上,正因为没有经过典型意义的生物遗传过程,Xenobot的“后代”不确定性更高,不可控性和存在的风险更大。

人类推进科技进步需要时刻保持谨慎的态度,特别是在科学探索过程中保持对自然的反思与敬畏。人工生物机器人是一种极为特殊的生命存在形式,对其技术和伦理风险的研究已应提上日程,并由生命科学、人工智能、医学、社会学、哲学等领域的学者与实践者共同推进。

(作者系中国科学院自动化研究所研究员)

王海峰详解全球首个知识增强千亿大模型鹏城-百度·文心

■本报记者 赵广立

作为当前人工智能(AI)发展的重要方向,预训练模型已成为AI领域的技术新高地。

12月8日,鹏城实验室与百度联合召开发布会,正式发布双方共同研发的全球首个知识增强千亿大模型——鹏城-百度·文心(模型版本号:ERNIE 3.0 Titan)。该模型参数规模达到2600亿,已在60多项任务中取得“最好效果”。

中国工程院院士、鹏城实验室主任高文表示:“预训练模型对整个科学的发展、社会的发展、创新的发展都是非常重要的工具。运用这个工具,可以帮助做很多人工智能的赋能,不局限于某个领域,这对人工智能的发展是一个福音。”

“百度知识增强大模型从大规模知识和海量数据中融合学习,效率更高、效果更好,具有良好的可解释性。”百度首席技术官王海峰介绍。此次发布的鹏城-百度·文心是“全球首个知识增强千亿大模型”,在机器阅读理解、文本分类、语义相似度计算等60多项任务中取得最好效果,并在30余项小样本和零样本任务上刷新基准。

作为“全球首个知识增强千亿大模型”,鹏城-百度·文心有何特点?AI大模型如何应用落地?对此,王海峰接受了《中国科学报》的采访。

首个知识增强千亿大模型鹏城-百度·文心知多少

《中国科学报》:如今大模型大行其道,鹏城实验室与百度共同研发的鹏城-百度·文心有哪些独特之处?

王海峰:该模型是全球首个知识增强千亿大模型,也是目前为止全球最大的中文单体模型,参数规模达到2600亿。

在算法框架上,该模型沿袭了ERNIE 3.0的海量无监督文本与大规模知识图谱的平行预训练算法,模型结构上使用兼顾语言理解与语言生成的统一预训练框架。为提升模型语言理解与生成能力,研究团队进一步设计了可控和可信学习算法。

在训练上,结合百度飞桨自适应大规模分布式训练技术和“鹏城云脑Ⅱ”领先算力集群,解决了超大模型训练中多个公认技术难题。在应用上,首创大模型在线蒸馏框架,大幅降低大模型落地成本。

《中国科学报》:鹏城-百度·文心在具体训练任务中表现如何?

王海峰:目前,该模型已在机器阅读理解、文本分类、语义相似度计算等60多项任务中取得最好效果。

在大模型的场景落地应用中,仅利用少量标注数据甚至无需标注数据,就能解决新场景的任务是AI工业大生产的关键问题。鹏城-百度·文心在30余项小样本和零样本任务上刷新基准,能够实现各类AI应用场景效果的提升,也为产业化规模应用打开了新窗口。

《中国科学报》:鹏城-百度·文心能有如此表现的背后,有哪些支撑加持?

王海峰:鹏城-百度·文心的成功发布,得益于鹏城实验室的算力系统“鹏城云脑Ⅱ”和飞桨深度学习平台的强强联手,解决了超大模型训练的多个公认技术难题,使鹏城-百度·文心训练速度大幅提升,模型效果更优。

“鹏城云脑Ⅱ”是国产自主的首个E级AI算力平台,曾在国际上多个性能测试上获得冠军;飞桨是我国首个自主研发的深度学习开源开放平台,研制了端到端自适应分布式训练框架,实现多硬件支持,并行效率高达90%,有效支持鹏城-百度·文心千亿大模型高效、稳定地训练。

AI预训练模型,越大越好吗?

《中国科学报》:当前,巨量参数的大模型有成为AI领域新一轮“军备竞赛”的趋势,是否参数越大,就代表大模型效果越好?

王海峰:大模型,大是一个相对的概念,可能会不断变得更大。我们也看到百度·文心这几年的发展,随着参数量的增加,效果越来越好。但并非一味地追求参数大,大模型的效果就一定越来越好。或者说百度做大模型不是只考虑参数的大小。

百度·文心大模型的核心特色和优势是知识增强。知识增强大模型从大规模知识和海量数据中融合学习,效率更高,效果更好,具有良好的可解释性。知识增强提升了模型的学习效率,能够在更小的参数规模下,取得更好的效果。

《中国科学报》:超大模型训练、推理需要消耗极其密集和昂贵的资源,因此应用落地成为难题。鹏城-百度·文心在实际应用落地中有何表现?

王海峰:为解决大模型应用落地难题,百度团队首创了大模型在线蒸馏技术,模型参数压缩率可达99.98%,压缩版模型仅保留0.02%参数规模就能与原有模型效果相当,为产业大规模应用打开新窗口。

该技术在鹏城-百度·文心大模型学习的过程中周期性地将知识信号传递给若干个“学生模型”同时训练,达到蒸馏阶段一次性产出多尺寸“学生模型”的目的。相对传统蒸馏技术,该技术极大节省了因大模型额外蒸馏计算以及多个学生的重复知识传递带来的算力消耗问题。

《中国科学报》:目前鹏城-百度·文心大模型有哪些应用落地案例?

王海峰:鹏城-百度·文心大模型近期会在OpenI启智社区开源,依托鹏城云脑Ⅱ对外开放,有助于推动产学研协各方充分挖掘AI大模型的赋能能力,助力科技创新,推动产业发展。

百度·文心产业级知识增强大模型应用于百度搜索、信息流、智能音箱等互联网产品,并通过百度智能云赋能工业、能源、金融、通信、媒体、教育等各行各业。

以金融、保险等行业的合同业务为例,对于任务繁重、人员紧张、工作强度大、准确性和及时性要求高的合同业务处理,人工处理面临着产能和效率的巨大挑战,且需要处理人员具备较高的专业性,企业需付出额外培训成本。

2020年起,百度与金融领域客户联合共建了提供保险合同条款智能解析模型的定制化建模,实现保险合同文本的条款解析分类。通过百度·文心大模型赋能,保险合同条款智能解析模型能够完成一份合同内近40个类目条款的智能分类,使得业务员处理单份合同文本的时长可缩短到1分钟,大大提升了工作效率。目前这套保险合同条款智能解析模型完成了上亿份合同条款的智能分类,实现了降本增效。

再比如,中国联通与百度合作,联手打造了集约化智慧客服,释放人力资源,降低人工成本,让服务更精准、工作更高效。以文心大模型强大的语义表示能力为基础,打造面向对话理解问题的专用预训练模型,从而实现了一套面向场景可定制的对话技术。在保持优异应用效果的同时,该技术对数据标注量的需求降低45%以上,显著提升了智能客服业务铺开的效率。