自然语言处理预训练模型专题导读

专题策划人 >>>

郑纬民

清华大学计算机系教授、中国工程院院士；长期从事高性能计算机体系结构、并行算法和系统研究；提出了可扩展的存储系统结构及轻量并行的扩展机制，发展了存储系统扩展性理论与方法，在中国率先研制并成功应用集群架构高性能计算机，在国产神威太湖之光上研制的极大规模天气预报应用获得ACM Gordon Bell奖；曾获国家科技进步奖一等奖1项、二等奖2项，国家技术发明奖二等奖1项，何梁何利基金科学与技术进步奖，首届中国存储终身成就奖；发表学术论文500余篇，编写和出版相关教材和专著10部。

近年来，预训练语言模型的出现给自然语言处理领域带来了一场变革，成为人工智能技术发展的前沿和热点。大规模预训练可以有效缓解传统技术在特征工程方面面临的压力。通过学习通用语言表示，模型具备了语言理解和生成能力，几乎在所有自然语言处理任务上都取得了突破。因此，各类基准测试任务的效果显著提高，这展示了大规模预训练广阔的应用前景。庞大的参数规模使得模型具备了更强的能力，同时也对模型的构建、训练和应用落地提出了挑战。自然语言处理的关键要素是什么？从多语言、知识和视觉等角度如何提高预训练模型的能力？规模庞大的模型如何进行高效训练？针对预训练语言模型研究中广受关注的问题，本期专题的文章从不同方面论述自然语言处理预训练模型的研究进展及相关成果，希望能对读者有所帮助。

《自然语言处理新范式：基于预训练模型的方法》一文介绍了自然语言处理技术的演化过程，指出自然语言处理主要靠知识、算法和数据来约束形式与意义的映射关系。大模型、大数据和大计算的充分使用，使大规模预训练语言模型在几乎所有自然语言处理任务上的性能都有显著提升。大规模预训练模型仍需解决模型的高效性、易用性、可解释性、鲁棒性以及推理能力等方面的关键问题，将继续沿“同质化”和“规模化”的道路发展。

《知识指导的预训练语言模型》一文提出以预训练语言模型为代表的深度学习仍然面临可解释性不强、鲁棒性差等难题。如何将人类积累的丰富知识引入模型，是改进深度学习性能的重要方向。文章围绕知识表示、知识获取，以及知识在预训练语言模型中的应用，系统地介绍了知识指导的预训练语言模型的最新进展与趋势。

《知识增强预训练模型》一文提出预训练模型主要从海量未标注、无结构化数据中学习，这个过程缺少外部知识指导，模型学习效率、模型效果和知识推理能力受到限制。文章从不同类型知识的引入、融合知识的方法、缓解知识遗忘的方法等角度，介绍了知识增强预训练模型的发展，并以知识增强预训练模型百度文心为例，介绍知识增强预训练模型的原理、方法及应用。

《悟道·文澜：超大规模多模态预训练模型带来了什么?》一文介绍了中国人民大学高瓴人工智能学院研究团队在多模态预训练模型方面的研究进展。针对互联网产生的图文往往只有弱相关语义关系的特点，团队提出了BriVL双塔模型，利用亿级互联网图文数据并通过自监督任务来进行训练。团队还提出了多语言多模态预训练单塔模型MLMM，可以跨语言跨模态学习通用常识。文章还讨论了多模态预训练模型对文本编码、图像生成和图文互检等任务带来的影响。

《鹏程·盘古：大规模自回归中文预训练语言模型及应用》一文介绍了以鹏城实验室为首的团队在鹏城云脑II上训练鹏程·盘古模型的工作。该模型具有2 000亿参数，基于TB级别的中文训练数据，采用自动并行技术将训练任务扩展至4 096个处理器上。该模型在少样本或零样本情况下具有较优性能，在大模型压缩、提示微调学习、多任务学习及持续学习等方面也取得了很好的应用效果。

《超大规模多模态预训练模型M6的关键技术及产业应用》一文介绍了阿里巴巴达摩院在多模态预训练模型方面的

DOI：10.12142/ZTETJ.202202001
收稿日期：2022-02-20

探索，重点聚焦于多模态表示学习和超大规模预训练模型的研究。文章提出了超大规模中文多模态预训练模型M6和参数规模从百亿到十万亿的超大模型，介绍了M6模型的产业化落地情况及其大规模预训练平台。

《高效训练百万亿参数预训练模型的系统挑战和对策》一文介绍了清华大学计算机系研究团队在国产E级高性能计算机上训练上百万亿参数的超大规模预训练模型所采用的系统优化技术，重点讨论了在训练如此规模的预训练模型时遇到的几个关键系统挑战，包括并行策略选取、数据存储方式、数据精度选取，以及负载均衡的实现方式，并总结了针对上述挑战的解决方法。

郑纬民

2022年2月19日