



信息通信领域产学研合作特色期刊 十佳皖刊
第三届国家期刊奖百种重点期刊 中国科技核心期刊

ISSN 1009-6868
CN 34-1228/TN

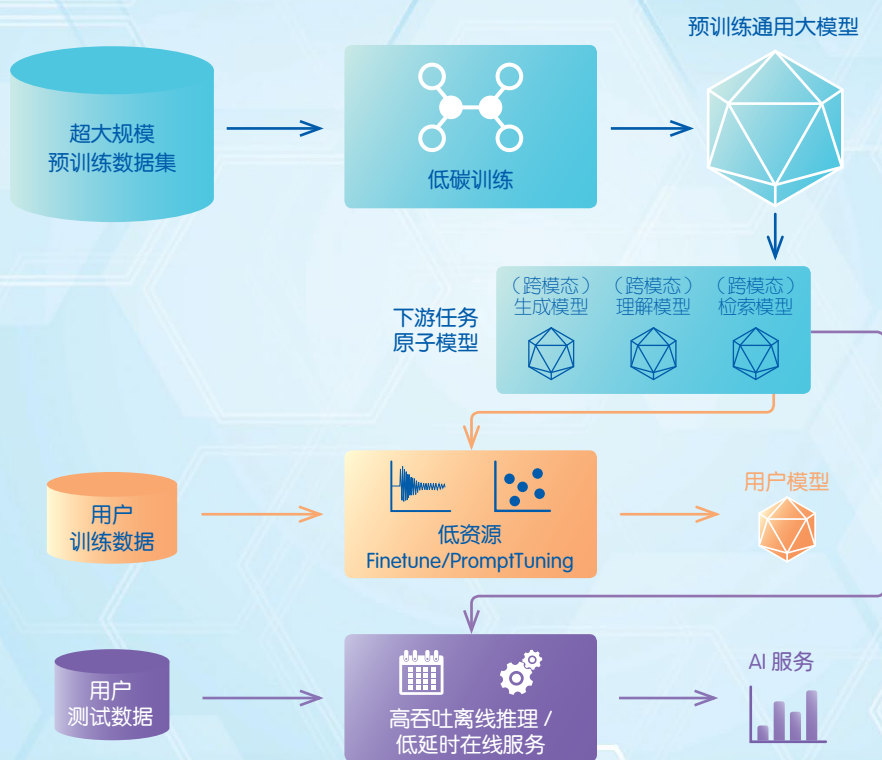
中兴通讯技术

ZTE TECHNOLOGY JOURNAL

<http://tech.zte.com.cn>

2022 年 4 月 · 第 2 期

专题：自然语言处理预训练模型



《中兴通讯技术》第9届编辑委员会成员名单

顾问 侯为贵(中兴通讯股份有限公司创始人) 钟义信(北京邮电大学教授)
陈锡生(南京邮电大学教授) 糜正琨(南京邮电大学教授)

主任 陆建华(中国科学院院士)

副主任 李自学(中兴通讯股份有限公司董事长) 李建东(西安电子科技大学教授)

编委 (按姓名拼音排序)

陈建平	上海交通大学教授	唐宏	中国电信IP领域首席专家
陈前斌	重庆邮电大学教授、副校长	唐雄燕	中国联通研究院副院长
段晓东	中国移动研究院副院长	陶小峰	北京邮电大学教授
葛建华	西安电子科技大学教授	王文博	北京邮电大学教授、副校长
管海兵	上海交通大学教授	王文东	北京邮电大学教授
郭庆	哈尔滨工业大学教授	王喜瑜	中兴通讯股份有限公司执行副总裁
洪波	中兴发展股份有限公司总裁	王翔	中兴通讯股份有限公司高级副总裁
洪伟	东南大学教授	王耀南	中国工程院院士
黄宇红	中国移动研究院院长	卫国	中国科学技术大学教授
纪越峰	北京邮电大学教授	吴春明	浙江大学教授
江涛	华中科技大学教授	邬贺铨	中国工程院院士
蒋林涛	中国信息通信研究院科技委主任	向际鹰	中兴通讯股份有限公司首席科学家
金石	东南大学首席教授、副校长	肖甫	南京邮电大学教授
李尔平	浙江大学教授	解冲锋	中国电信研究院教授级高工
李红滨	北京大学教授	徐安士	北京大学教授
李厚强	中国科学技术大学教授	徐子阳	中兴通讯股份有限公司总裁
李建东	西安电子科技大学教授	续合元	中国信息通信研究院副总工
李乐民	中国工程院院士	薛向阳	复旦大学教授
李融林	华南理工大学教授	薛一波	清华大学教授
李少谦	电子科技大学教授	杨义先	北京邮电大学教授
李自学	中兴通讯股份有限公司董事长	叶茂	电子科技大学教授
林晓东	中兴通讯股份有限公司副总裁	易芝玲	中国移动研究院首席科学家
刘健	中兴通讯股份有限公司高级副总裁	张宏科	中国工程院院士
刘建伟	北京航空航天大学教授	张平	中国工程院院士
隆克平	北京科技大学教授	张钦宇	哈尔滨工业大学(深圳)副校长
陆建华	中国科学院院士	张卫	复旦大学教授
马建国	浙江大学教授	张云勇	中国联通云南分公司党委书记、总经理
毛军发	中国科学院院士	赵慧玲	工业和信息化部通信科技委信息通信网络专家组组长
孟洛明	北京邮电大学教授	郑纬民	中国工程院院士
任品毅	西安交通大学教授	钟章队	北京交通大学教授
石光明	鹏程实验室副主任、西安电子科技大学教授	周亮	南京邮电大学教授
孙知信	南京邮电大学教授	朱近康	中国科学技术大学教授
谈振辉	北京交通大学教授、原校长	祝宁华	中国科学院院士

目次

中兴通讯技术 (ZHONGXING TONGXUN JISHU)
总第 163 期 第 28 卷 第 2 期 2022 年 4 月

信息通信领域产学研合作特色期刊 第三届全国期刊奖百种重点期刊 中国科技核心期刊 工信部优秀科技期刊 十佳期刊 中国五大文献数据库收录期刊 1995 年创刊

热点专题►

自然语言处理预训练模型

- 01 专题导读 郑纬民
- 03 自然语言处理新范式:基于预训练模型的方法 车万翔, 刘挺
- 10 知识指导的预训练语言模型 韩旭, 张正彦, 刘知远
- 16 知识增强预训练模型 王海峰, 孙宇, 吴华
- 25 悟道·文澜:超大规模多模态预训练模型带来了什么? 卢志武, 金琴, 宋睿华, 文继荣
- 33 鹏程·盘古:大规模自回归中文预训练语言模型及应用 曾炜, 苏腾, 王晖, 田永鸿, 高文
- 44 超大规模多模态预训练模型 M6 的关键技术及产业应用 林俊扬, 周畅, 杨红霞
- 51 高效训练百万亿参数预训练模型的系统挑战和对策 马子轩, 翟季冬, 韩文弢, 陈文光, 郑纬民

专家论坛►

- 59 自然语言处理技术发展 王海宁

企业视界►

- 65 数字基础设施建设的思考与实践 王喜瑜
- 68 5G 行业虚拟专网能力提升与实践 陆平, 欧阳新志, 高雯雯

综合信息►

- 09 新增编委介绍

2022 年第 1—6 期专题计划及策划人

1. 新型网络技术

中国联通研究院副院长 唐雄燕

3. 智能超表面技术

中兴通讯技术预研总工 赵亚军
北京理工大学教授 费泽松

5. 通信感知一体化

中国科学技术大学教授 卫国

2. 自然语言处理预训练模型

中国工程院院士 郑纬民

4. 多频段协同通信

电子科技大学教授 李少谦
中国联通研究院副院长 唐雄燕
中兴通讯首席科学家 向际鹰

6. 网络内生安全

北京航空航天大学教授 刘建伟

MAIN CONTENTS

ZTE TECHNOLOGY JOURNAL
Vol. 28 No. 2 Apr. 2022

Special Topic ►

Pre-Trained Models for Natural Language Processing

01	Editorial	ZHENG Weimin
03	New Paradigm of Natural Language Processing: A Method Based on Pre-Trained Models	CHE Wanxiang, LIU Ting
10	Knowledge-Guided Pre-Trained Language Models ...	HAN Xu, ZHANG Zhengyan, LIU Zhiyuan
16	Knowledge-Enhanced Pre-Trained Models	WANG Haifeng, SUN Yu, WU Hua
25	WuDao-WenLan: What Do Very-Large Multimodal Pre-Training Models Bring?	LU Zhiwu, JIN Qin, SONG Ruihua, WEN Jirong
33	Pengcheng-PanGu: Large-Scale Autoregressive Pre-Trained Chinese Language Model with Auto-Parallel Computation and Its Application	ZENG Wei, SU Teng, WANG Hui, TIAN Yonghong, GAO Wen
44	Key Technologies and Applications of Extremely Large-Scale Multimodal Pre-Trained Model M6	LIN Junyang, ZHOU Chang, YANG Hongxia
51	Challenges and Measures for Efficient Training of Trillion-Parameter Pre-Trained Models	MA Zixuan, ZHAI Jidong, HAN Wentao, CHEN Wenguang, ZHENG Weimin
59	Development of Natural Language Processing Technology	WANG Haining
65	Reflections and Practice on Digital Infrastructure Constructions	WANG Xiyu
68	Capacity Improvement and Practice of 5G Industry Virtual Private Network	LU Ping, OUYANG Xinzhì, GAO Wenwen

Expert Forum ►

Enterprise View ►

期刊基本参数: CN 34-1228/TN*1995*b*16*74*zh*P*¥20.00*6500*11* 2022-04

敬告读者

本刊享有所发表文章的版权, 包括英文版、电子版、网络版和优先数字出版版权, 所支付的稿酬已经包含上述各版本的费用。未经本刊许可, 不得以任何形式全文转载本刊内容; 如部分引用本刊内容, 须注明该内容出自本刊。

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

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

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。

郑伟民

2022 年 2 月 19 日

自然语言处理新范式: 基于预训练模型的方法



New Paradigm of Natural Language Processing: A Method Based on Pre-Trained Models

车万翔/CHE Wanxiang, 刘挺/LIU Ting

(哈尔滨工业大学, 中国 哈尔滨 150001)
(Harbin Institute of Technology, Harbin 150001, China)

DOI: 10.12142/ZTETJ.202202002

网络出版地址: <https://kns.cnki.net/kcms/detail/34.1228.TN.20220408.1732.008.html>

网络出版日期: 2022-04-11

收稿日期: 2022-02-26

摘要: 以BERT和GPT为代表的、基于超大规模文本数据的预训练语言模型能够充分利用大模型、大数据和大计算,使几乎所有自然语言处理任务性能都得到显著提升,在一些数据集上达到甚至超过人类水平,已成为自然语言处理的新范式。认为未来自然语言处理,乃至整个人工智能领域,将沿着“同质化”和“规模化”的道路继续前进,并将融入多模态数据、具身行为数据、社会交互数据等更多的“知识”源,从而为实现真正的通用人工智能铺平道路。

关键词: 人工智能; 自然语言处理; 预训练语言模型; 同质化

Abstract: Pre-trained language models based on super-large-scale raw corpora, represented by BERT and GPT, can make full use of big models, big data, and big computing, which have significantly improved the performances of almost all-natural language processing tasks. The performances have reached or exceeded the human level on some datasets. Pre-trained language models have become a new paradigm for natural language processing. It is believed that in the future, natural language processing and even the entire field of artificial intelligence will continue to move forward along the path of “homogenization” and “scale”, and will integrate more sources of “knowledge”, such as multi-modal data, embodiment data, and social interaction data. Consequently, these methods will pave the way for achieving true general artificial intelligence.

Keywords: artificial intelligence; natural language processing; pre-trained language model; homogenization

1 自然语言处理的背景

自然语言通常指的是人类语言(本文特指文本符号,而非语音信号),是人类思维的载体和交流的基本工具,也是人类区别于动物的根本标志,更是人类智能发展的外在表现形式之一。自然语言处理(NLP)主要研究用计算机来理解和生成自然语言的各种理论和方法,属于人工智能领域的一个重要甚至核心的分支。人工智能应用领域的快速拓展对自然语言处理提出了巨大的应用需求。同时,自然语言处理研究为人们更深刻地理解语言的机理和社会的机制提供了一条重要的途径,因此具有重要的科学意义。

目前,人们普遍认为人工智能的发展先后经历了运算智能、感知智能、认知智能3个阶段。其中,运算智能关注的是机器的基础运算和存储能力。在这方面,机器已经完胜人类。感知智能则强调机器的模式识别能力,如语音的识别和

图像的识别,目前机器在感知智能上的水平基本达到甚至超过了人类的水平。然而,在涉及自然语言处理以及常识建模和推理等研究的认知智能上,机器与人类还有很大的差距。

为什么计算机在处理自然语言时会如此困难呢?这主要是因为自然语言具有高度的抽象性、近乎无穷变化的语义组合性、无处不在的歧义性和持续的进化性,并且理解语言通常需要背景知识和推理能力等。由于面临以上问题,自然语言处理已成为目前制约人工智能取得更大突破和更广泛应用瓶颈之一。包括图灵奖得主在内的多位知名学者都很关注自然语言处理,甚至图灵本人,也将验证机器是否具有智能的手段(即“图灵测试”)应用在自然语言处理上。因此,自然语言处理又被誉为“人工智能皇冠上的明珠”。

2 自然语言处理问题的解决之道

自然语言处理的本质是形式与意义的多对多映射关系。

其中,形式和意义的空间都近乎无限。那么,如何才能找到正确的映射关系呢?利用“知识”进行约束是唯一有效的办法。因此,如何获取和利用“知识”成为解决自然语言处理问题的关键科学问题。

应用于自然语言处理的知识来源主要有三大类:狭义知识、算法和数据。表1对这三大类知识来源进行了总结。

第一大类知识是狭义知识,即通过规则或词典等形式由人工定义的显性知识,也就是人们通常所理解的知识类型。具体来讲,狭义知识又包括3类,即语言知识、常识知识和世界知识。其中,语言知识是指对语言的词法、句法或语义进行的定义或描述。例如,WordNet^[1]是由普林斯顿大学编写的英文语义词典,其主要特色是定义了同义词集合。每个同义词集合由具有相同意义的词组成。此外,WordNet还为每个同义词集合提供简短的释义,同时不同同义词集合之间还具有一定的语义关系。常识知识是指人们基于共同经验而获得的基本知识。常识往往是不言自明的,并没有记录为文字,所以很难从文本中挖掘到。著名的Cyc项目试图将上百万条知识编码成机器可用的形式,用以表示人类常识。世界知识包括实体、实体属性、实体之间的关系等。这类知识往往通过知识图谱的形式加以描述和存储。

第二大类知识是算法。算法的本质是解决问题的过程或者方法,它也是一种知识类型。机器学习算法则可以看作人为定义的函数。虽然这种函数的参数是由机器自动学习获得的,但是函数的类型仍然由人类来定义。这在某种程度上反映了设计者对待解决问题的认知,具有一定的归纳偏执性。例如,卷积神经网络(CNN)就利用了识别对象的平移不变性质。面向各种自然语言处理问题的算法更是和语言知识密切相关。与狭义知识相比,算法知识具有更好的灵活性和动态性。

▼表1 自然语言处理中的三大类知识来源

知识类别	细分类	特点	举例
狭义知识	语言知识	词典、规则库	WordNet
	常识知识	很难从文本中挖到	Cyc项目
	世界知识	可以从文本中挖到	知识图谱
算法	浅层机器学习	人工提取特征	SVM、CRF
	深度学习	自动归纳特征	MLP、RNN、CNN
	NLP算法	利用语言知识	CKY、MST
数据	有标注	专家标注、众包	Penn TreeBank
	无标注	大量原始语料	预训练语言模型
	伪数据	与目标任务近似	数据增广

CKY: Cocke-Younger-Kasami 算法
CNN: 卷积神经网络
CRF: 条件随机场
MLP: 多层感知机
MST: 最大生成树
NLP: 自然语言处理
RNN: 循环神经网络
SVM: 支持向量机

第三大类知识是数据。数据是机器学习算法的基础。机器学习算法通常会依赖有标注的数据,需要借助人工方式来为每个输入标注出相应的输出结果。由于并没有具体指明如何从输入转换到输出,因此数据是一种隐性的知识。然而,由人工进行数据标注的方式费时又费力,导致标注数据量往往较小,不足以训练一个性能优异的机器学习算法。为了解决这一问题,预训练语言模型可直接利用大量未标注的原始语料,将语言模型作为训练的目标,即根据历史的词序列来预测下一个单词是什么,或者根据周围的词来预测当前的词是什么。由于未标注数据量近乎无限,因此可以训练一个性能较好的语言模型,并将该模型的参数作为下游任务模型的初始参数。这样便可以减少模型对标注数据量的依赖,大幅提高下游任务的准确率。

3 自然语言处理技术的发展历史

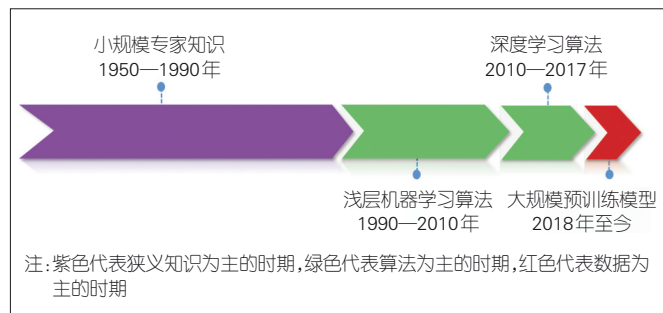
自然语言处理技术自从诞生以来经历了以狭义知识、算法和数据为主的3个时期(如图1所示)。

3.1 狭义知识为主时期

早期的自然语言处理(20世纪50年代到90年代)主要采用基于小规模专家知识的方法(规则、词典等狭义知识),通过专家总结的符号逻辑知识来处理通用的自然语言。然而,由于自然语言的复杂性,基于理性主义的规则方法在实际应用场景中仍存在一些不足。

3.2 算法为主时期

20世纪90年代开始,计算机运算速度和存储容量的快速增加,以及统计学习(浅层机器学习)算法的大量普及,均使得基于小规模语料库的浅层机器学习算法在自然语言处理领域得以大规模应用。由于语料库中包含了一些关于语言的知识,基于浅层机器学习算法的自然语言处理方法能够更加客观、准确、细致地捕获语言规律。这一时期,词法分



▲图1 自然语言处理的发展历史

析、句法分析、信息抽取、机器翻译、自动问答等领域的研究均取得了一定程度的突破。

尽管如此,基于浅层机器学习算法的自然语言处理技术仍存在明显的局限性,即需要事先利用经验性规则将原始的自然语言输入转化为机器能够处理的向量形式。这一转化过程(也称为特征提取)需要细致的人工操作和一定的专业知识,因此也被称为特征工程。

2010年以后,基于深度神经网络的表示学习方法(也称深度学习方法)逐渐兴起,可以直接端到端地完成各种自然语言处理任务,不再依赖人工设计。这里,表示学习是指机器能根据输入自动发现可用于识别或分类等任务的表示。具体来讲,深度学习模型在结构上通常包含多个处理层。最底层的处理层会接收原始输入,并对原始输入进行抽象处理,然后该层后面的每一层都会在前一层的结果上进行更深层次的抽象处理。最后一层的抽象处理结果即为输入的一个表示,以用于最终的目标任务。其中,抽象处理是由模型内部的参数来控制的,而参数的更新值则是由反向传播算法根据训练数据中模型的表现来学习得到的。由此可以看出,深度学习可以有效避免统计学习方法中的人工特征提取操作,自动地发现对目标任务有效的表示。

深度学习方法还能打破不同任务之间的壁垒。传统浅层机器学习方法需要为不同任务设计不同的特征,而这些特征往往是不通用的。然而,深度学习方法却能够将不同任务在相同的向量空间内进行表示,从而具备跨任务迁移的能力。此外,深度学习方法还可以实现跨语言甚至跨模态的迁移,可以综合利用多项任务、多种语言、多个模态的数据,使得人工智能向更通用的方向迈进一步。

同样,得益于深度学习技术的快速发展,自然语言处理的另一个主要研究方向——自然语言生成也取得了长足进步。长期以来,自然语言生成的研究几乎处于停滞状态:除了使用模板生成一些简单的语句外,并没有获得其他有效的解决办法。随着基于深度学习的序列到序列生成框架的提出,这种逐词文本生成方法全面提升了生成技术的灵活性和实用性,完全革新了机器翻译、文本摘要、人机对话等技术范式。

虽然深度学习技术能够大幅提高自然语言处理系统的准确率,但是基于深度学习的算法仍有一个致命的缺点:过度依赖大规模标注数据。对于语音识别、图像处理等感知类任务,标注数据相对容易获得。例如,在图像处理领域,人们已经为上百万幅图像标注了相应的类别(如ImageNet数据集)。用于语音识别的“语音和文本”平行语料库标注的时间也有几十万小时。然而,自然语言处理具有主观性特

点,它所面对的任务和领域又众多。这些均使得大规模语料库标注的时间和人力成本变得很高。因此,自然语言处理的标注数据往往不够充足,很难满足深度学习模型训练的需要。

3.3 数据为主时期

早期的静态词向量预训练模型和后来的动态词向量预训练模型,特别是自2018年以来以BERT、GPT为代表的超大规模预训练语言模型,都很好地弥补了自然语言处理标注数据不足的缺点。这些模型大大促进了自然语言处理技术的发展,使得包括阅读理解在内的几乎所有自然语言处理任务性能都得到了大幅提高,在有些数据集上的性能表现达到甚至超过了人类水平。

模型预训练是指,首先在一个源任务上训练一个初始模型,然后在下游任务(也称目标任务)上继续对该模型进行精调,从而达到提高下游任务准确率的目的。模型预训练本质上是迁移学习思想的一种应用。然而,由于同样需要人工标注,源任务标注数据的规模往往非常有限。那么,如何获得更大规模的标注数据呢?

实际上,文本自身的顺序就是一种天然的标注数据。通过若干连续出现的词语来预测下一个词语(又称语言模型)就可以构成一项源任务。由于图书、网页等文本数据的规模近乎无限,因此模型可以非常容易地获得超大规模的预训练数据。有人将这种不需要人工标注数据的预训练学习方法称为无监督学习方法。其实,这种叫法并不准确。这是因为这种方法的学习过程仍然是有监督的。更准确的叫法应该是自监督学习。

为了能够刻画大规模数据中复杂的语言现象,深度学习模型的容量需要足够大。基于自注意力机制的Transformer模型能够显著提升自然语言建模能力,是近年来具有里程碑意义的进展之一。要想在可容忍的时间内在如此大规模的数据上训练一个超大规模的Transformer模型,就离不开以图形处理器(GPU)、张量处理器(TPU)为代表的现代并行计算硬件。可以说,超大规模预训练语言模型完全依赖“蛮力”,在大数据、大模型和大计算资源的加持下,使自然语言处理取得了长足的进步。例如,OpenAI推出的GPT-3拥有1750亿个参数,无须接受任何特定任务的训练,便可通过小样本学习来完成10余种文本生成任务(如风格迁移、网页生成等)。

目前,预训练模型已经成为了自然语言处理的新范式。它甚至影响了整个人工智能的研究和应用,开启了人工智能领域“大规模预训练模型”时代的大门。

4 几种具有代表性的预训练语言模型

4.1 词嵌入预训练语言模型

在基于浅层机器学习的自然语言处理时期,人们使用高维、离散的向量来表示自然语言。其中,每个词用独热向量来表示,向量维度表示词的大小(只有一位为1,其余均为0)。然而,这种独热向量表示方法无法解决“多词一义”的问题。也就是说,即便两个词含义相近,它们的表示也是截然不同的。例如,“马铃薯”和“土豆”会使用两个不同的独热向量表示。假如训练数据中只出现“土豆”,那么当测试系统中出现“马铃薯”时,模型就无法进行正确加权。

语言学家J. R. FIRTH在1957年指出,通过一个词周围的词便可理解该词的含义^[2],即“观其伴知其义”。例如,“马铃薯”和“土豆”的周围经常出现“吃”“烹饪”“种植”等,所以这两个词就比较相似。因此,可以将一个词周围出现过的词收集起来,构建一个相对更稠密的向量,然后用该向量来表示这个词。当然,还可以使用降维等技术,用更低维、更稠密的向量来表示词。

2003年,图灵奖得主Y. BENGIO首次提出词嵌入的概念^[3],即直接使用一个低维、稠密、连续的向量来表示词。那么,如何获得(即预训练)一个好的词嵌入表示呢?对此,可以通过其在下游任务上表现,对向量每一维的数值进行自动设置。除了需要一个下游任务外,还需要针对该任务的大规模训练数据,以保证模型能覆盖足够多的语言现象。然而,由于自然语言的主观性,很难获得大规模的标注数据。比如,情感分析数据最多也就几万条。好在语言具有“观其伴知其义”的性质,因此可以通过一个词周围的词,来预测当前的词,这样就自然构成了一个“下游任务”。具体的任务可以分为两类:一类是通过历史词序列来预测下一个词,这类任务又被称作“语言模型”任务;另一类是利用周围的词来预测中间的词,这类任务类似于“完形填空”任务。各种电子文档、图书乃至整个互联网上的文本数据,都可以作为训练数据,从而大大增强了词嵌入表示的学习能力。虽然Y. BENGIO等早在2003年便提出了词向量概念,并通过语言模型任务对词向量进行了预训练,但是直到2013年谷歌的T. MIKOLOV等提出Word2vec模型^[4]后,该思想才开始普及。

4.2 上下文相关词嵌入预训练语言模型

虽然词嵌入表示可以处理“多词一义”现象,但是其本身仍然存在一个致命的缺点,即无法处理“一词多义”现象。词嵌入的一个基本假设是:每一个词都对应唯一一个词

嵌入表示。如果一个词有多种词义,那么用哪个词义的向量来表示这个词呢?这里我们仍然以“土豆”这个词为例进行说明。作为一种蔬菜时,“土豆”应该和“马铃薯”等词的表示相似;而作为一个视频网站时,“土豆”又应该和“爱奇艺”等词的表示相似。那么,最终“土豆”的词嵌入表示必将是个“四不像”。

为解决上述问题,AllenNLP于2018年提出了ELMo模型^[5]。该模型的核心思想是将语言模型的输出作为词向量表示。这种表示是上下文相关的。例如,在句子“我喜欢吃土豆”中,“土豆”的表示应该和“马铃薯”相似;而在句子“我在土豆上看电影”中,“土豆”的表示则应该和“爱奇艺”相似。将ELMo输出的词向量作为特征,大大提高了下游任务的性能。

4.3 大规模预训练语言模型

在ELMo模型提出后不久,OpenAI便提出了第1代GPT模型^[6],正式将自然语言处理技术带入“预训练”时代。和ELMo一样,GPT也把语言模型任务作为预训练任务。总的来说,GPT模型主要有三大创新点:首先,它使用了性能更强大的Transformer模型;其次,它在目标任务上精调整个模型,而不是只将模型的输出结果作为固定的词向量特征;最后,由于预训练模型自身非常复杂,因此接入的下游任务模型可以非常简单,这极大降低了自然语言处理的技术门槛。

在GPT提出后不久,谷歌便提出了著名的BERT模型^[7]。与GPT相比,BERT最大的改进在于它能利用词两边的上下文来预测中间的词,即使用“完形填空”作为预训练任务。由于使用了更为丰富的上下文,因此BERT能够获得更好的预训练效果。BERT一问世便刷新了各大自然语言处理任务记录,在有些任务上的表现甚至超越了人类。

随后,微软也建造了自己的超大规模人工智能计算平台,并同OpenAI联合训练了GPT-3模型^[8]。

GPT-3含有1750亿个超大规模参数。由于模型参数太大,研究人员无法再对它进行精调。为了满足不同的任务需求,模型需要针对不同任务提供相应的“提示语”。例如,只输入任务描述“Translate English to French: cheese=>”,GPT-3就能够直接输出翻译结果。如果在输入任务描述之后再给出一个或几个示例,那么任务完成的效果会更好。这种技术也被称为“提示学习”,并被认为是自然语言处理的一种新技术范式。

在GPT、BERT模型提出以后,各种预训练模型如雨后春笋般涌现,并从各个方面提升了预训练模型的效果,例如

更大规模的预训练模型、多语言多模态预训练模型、面向各种领域的预训练模型等,其中也包括新的预训练任务、各种预训练模型压缩与加速方法。文献[9-10]对此做了详细描述。

5 自然语言处理的未来展望

5.1 预训练模型亟待解决的关键技术问题

目前,大规模预训练模型的发展势头非常强劲。模型规模不断扩大,模型渗透的领域也在不断增加。因此,短期内自然语言处理仍将沿着大规模预训练模型的道路继续前进。不过,若要取得更好的效果并实现模型的应用落地,在开展大规模预训练模型研究时仍需要解决以下几个关键的研究问题:

(1) 模型的高效性。大规模预训练模型的训练和部署都需要消耗大量的计算资源。考虑到大规模预训练模型在训练时产生的大量碳排放对环境的影响,研制计算效率更高的模型将是未来研究的重要方向。另外,还可以通过蒸馏、剪枝等技术将大模型压缩为规模更小的模型,以便于模型实现更大规模的部署应用。

(2) 模型的易用性。自然语言处理任务和应用领域层出不穷。为了能够满足新任务和新领域的需求,预训练模型还需要解决小样本甚至零样本学习问题。另外,还需要构建大规模预训练模型的工程化开发能力,建设通用的开发工作流,减少专家干预及人为调整参数,构建一整套数据、代码、模型、应用程序接口(API)等服务的平台,从而支撑工业、医疗、城市、金融、物流、科学研究等领域,拓展人工智能的应用范围,并对人类生产和生活产生更广泛的影响。

(3) 模型的可解释性。深度学习模型一直存在可解释性差的问题,而预训练模型也并没有解决这一问题。医疗诊断、法律判案等需要证据的应用场合仍无法直接利用该技术。即便在一些不需要提供证据的应用中,如垃圾邮件识别,模型如果能够解释自身是如何做出预测的,那么这将对提高模型的可信性大有裨益。

(4) 模型的鲁棒性。虽然在很多数据集上,预训练模型已经取得不错的性能突破,在有些方面甚至已经超过人类,但有时只要测试数据稍加变动,即便语义不发生变化,之前能够被正确预测的数据也可能会获得错误的预测结果。这就是目前模型遇到的典型的鲁棒性(也称健壮性)问题,导致模型很容易被别有用心者攻击。此外,由于预训练模型极度依赖大规模未标注数据,如果所收集的数据中存在错误或陈

旧的信息,甚至被人为地植入后门数据,模型可能会被误导或误用。

(5) 模型的推理能力。目前预训练模型拥有的强大性能主要来自对数据的记忆能力。模型能够很容易地回答曾经见过的知识,但是对于不曾见过尤其需要多步推理才能解答的问题,往往不具有很好的解决能力。推理能力恰恰是人类解决问题的重要手段,是智能的重要体现形式,因此也是预训练模型需要重点解决的问题。

那么,自然语言处理是否会沿着预训练模型这条路一直发展下去呢?对此,本文首先分析一下人工智能的发展趋势。

5.2 自然语言处理的历史发展趋势

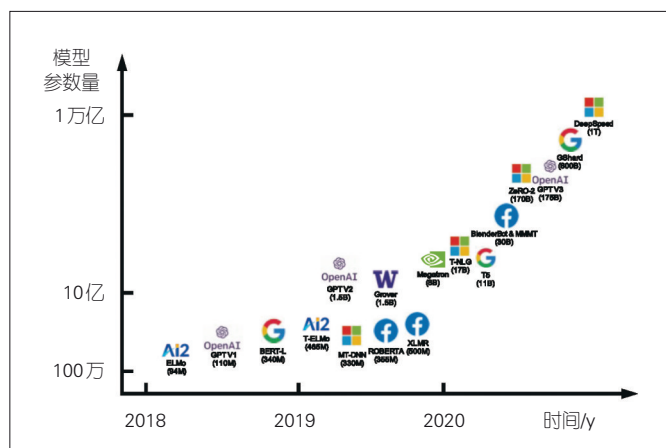
经过60余年的发展,自然语言处理经历了小规模专家知识、浅层机器学习算法、深度学习算法、基于大规模预训练模型的方法等阶段,呈现了明显的“同质化”和“规模化”两个相辅相成的发展趋势。

5.2.1 模型“同质化”的趋势明显

自然语言处理的发展呈现出明显的“同质化”趋势。早期利用专家知识的自然语言处理系统需要针对不同的任务编写不同的规则,因此不具有通用性和可移植性。后来,浅层机器学习算法需要根据不同的任务来编写特定的逻辑,以便将原始文本(也可以是声音、图像等)转换为更高级别的特征,然后使用相对“同质化”的机器学习算法(如支持向量机)进行结果预测。此后,深度学习技术能够使用更加“同质化”的模型架构(包括卷积神经网络、循环神经网络等),直接将原始文本作为学习模型的输入,并在学习的过程中自动“涌现”出用于预测的更高级别的特征。大规模预训练模型“同质化”的特性更加明显。例如,几乎所有新的自然语言处理模型都源自少数大规模预训练模型(比如BERT、RoBERTa、BART、T5等)。GPT-3模型只需要进行一次预训练就可以直接(或仅使用极少量的训练样本)完成特定的下游任务。“同质化”还体现在跨数据模态上。基于Transformer的序列建模方法和预训练模型在被成功应用于自然语言处理后,现已在图像、视频、语音、表格数据、蛋白质序列、有机分子等模态数据上取得优异的效果。这使得未来构建一个能够统一各种模态的大规模预训练模型成为可能。

5.2.2 “规模化”是智能涌现的必要条件

虽然预训练模型只是迁移学习的简单应用,但是它涌现



▲图2 大规模预训练模型模型参数的发展趋势

出了令人惊讶的“智能”。其中“规模化”是必不可少的条件。“规模化”需要的3个必要前提目前皆已成熟。

(1) 计算机硬件的升级。例如，GPU 吞吐量和存储容量在过去4年中增加了10倍。

(2) Transformer 模型架构的发明。该模型能直接对序列中的远程依赖关系进行建模，还能充分利用硬件的并行性。

(3) 更多可用的数据。过去，使用人工标注的数据进行有监督模型训练是标准的做法。然而，较高的标注成本限制了模型优势的发挥。预训练模型能够充分利用超大规模的未标注数据进行自监督学习，从而比在有限标注数据上进行训练的模型能获得更好的泛化性能。

正是由于“规模化”的重要性，越来越多的科研机构不断推出规模越来越大的预训练模型。例如，与 GPT-2 的 15 亿个参数相比，OpenAI 的 GPT-3 模型参数规模达到了惊人的 1 750 亿。谷歌、微软、北京智源、华为、阿里、鹏城实验室等也相继推出了同等甚至更大规模的预训练模型，如图 2 所示。

5.3 自然语言处理的未来技术趋势

基于自然语言处理的历史发展趋势来判断,自然语言处理将沿着“同质化”和“规模化”的道路继续前进。

首先，以Transformer为代表的自注意力模型具有非常好的“同质化”性质。也就是说，该类模型不会对所处理的问题进行约束和限制，因此适用于自然语言、图像、语音等各类数据的处理。未来，也许会出现性能更优异的模型，但是该模型一定是更加“同质化”的。

其次，模型规模的发展速度已经远远超过摩尔定律限制的硬件发展速度。然而，无论是神经元还是它们之间连接的数量，都远远不及人脑。因此，“规模化”的发展趋势仍不

会改变。期待新的能够突破摩尔定律的硬件形态的出现。

最后，人类习得语言所需的知识并非仅仅是规则、算法以及文本数据这3种类型，还包括大量其他模态的知识，如声音、视频、图像等。多模态预训练模型（如文本、图像、视频、音频之间的联合预训练）已成为目前研究的热点。此外，如要实现真正的自然语言处理，甚至通用人工智能，那么智能体就需要从物理世界中获得反馈，这样才能真正理解“冷暖”“软硬”等概念，即拥有具身的能力。另外，语言作为一种人类交流的工具，具有极强的社会属性。因此，智能体还需要与其他人进行交流，在应用中真正习得语言。在未来，自然语言处理模型一定需要融合这些更广义的知识。“同质化”和“规模化”的模型也为融合这些知识提供了必要的支撑条件。

6 结束语

在大数据、大模型和大算力的加持下，基于预训练的模型完全革新了自然语言处理的研究范式。在未来，自然语言处理，乃至整个人工智能领域，仍将沿着“同质化”和“规模化”的道路继续前进，并将融入更多的“知识”源，包括多模态数据、具身行为数据、社会交互数据等，从而实现真正的通用人工智能。

参考文献

- [1] CHRISTIANE F, MILLER G A. WordNet: an electronic lexical database [M]. Cambridge: MIT Press, 1998
- [2] FIRTH J R. A synopsis of linguistic theory 1930–55. [J]. Studies in linguistic analysis, 1957:1–32
- [3] BENGIO Y, DUCHARME R, VINCENT P, et al. A neural probabilistic language model [J]. Journal of machine learning research, 2003, 3: 1137–1155
- [4] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space [EB/OL]. [2022–02–25]. <http://export.arxiv.org/pdf/1301.3781>
- [5] PETERS M, NEUMANN M, IYER M, et al. Deep contextualized word representations [EB/OL]. [2022–02–25]. <https://arxiv.org/abs/1802.05365>
- [6] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training [EB/OL]. [2022–02–25]. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf
- [7] DEVLIN J, CHANG M, LEE K, et al. Pre-training of deep bidirectional transformers for language understanding [EB/OL]. [2022–02–25]. <https://arxiv.org/abs/1810.04805>
- [8] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners [EB/OL]. [2022–02–25]. <https://arxiv.org/abs/2005.14165>
- [9] QIU X P, SUN T X, XU Y G, et al. Pre-trained models for natural language processing: a survey [J]. Science China technological sciences, 2020, 63 (10): 1872–1897. DOI: 10.1007/s11431-020-1647-3
- [10] KALYAN K S, RAJASEKHARAN A, SANGEETHA S. AMMUS: a survey of transformer-based pretrained models in natural language processing [EB/OL]. [2022–02–25]. <https://arxiv.org/abs/2108.05542>

作者简介



车万翔, 哈尔滨工业大学计算学部教授、人工智能研究院副院长, 教育部“青年长江学者”, 斯坦福大学访问学者, 现任中国中文信息学会理事、计算语言学专业委员会副主任兼秘书长, 国际计算语言学学会亚太分会(AACL)执委兼秘书长, 中国计算机学会高级会员; 主要研究方向为自然语言处理、人机对话系统; 2030“新一代人工智能”重大项目课题负责人, 承担国家自然科学基金项目3项; 获黑龙江省青年科技奖、谷歌专注研究奖、黑龙江省科技进步一等奖、中国中文信息学会“钱伟长”中文信息处理科学技术奖一等奖、首届汉王青年创新奖、AAAI 2013最佳论文提名奖等; 发表学术论文100余篇, 论文累计被引用近6 000次(Google Scholar数据), H-index值为40, 出版教材4部, 翻译著作2部。



刘挺, 哈尔滨工业大学教授、计算学部主任兼计算机学院院长, 国家“万人计划”科技创新领军人才, “十四五”国家重点研发计划先进计算与新兴软件重点专项专家组成员, 教育部人工智能科技创新专家组成员, 中国计算机学会会士, 中国中文信息学会副理事长、社交媒体处理专委会(SMP)主任, 黑龙江省中文信息处理重点实验室主任, 黑龙江省“人工智能头雁”团队带头人, 国家重点研发项目、国家自然科学基金重点项目负责人, 多次担任国家“863”重点项目总体组专家、国家自然科学基金委评审专家, 曾任国际顶级会议ACL、EMNLP领域主席; 主要研究方向为人工智能、自然语言处理和社会计算; 主持研制的“语言技术平台LTP”“大词林”等科研成果被业界广泛使用; 曾获国家科技进步奖二等奖、黑龙江省科技进步奖一等奖、中国中文信息学会“钱伟长”中文信息处理科学技术奖一等奖等奖项。

新增编委介绍



金石

东南大学副校长、首席教授、博士生导师、教育部“长江学者奖励计划”特聘教授、国家自然科学基金杰出青年科学基金获得者、国家“万人计划”科技创新领军人才、江苏省特聘教授、中国通信学会会士、全国工程专业学位研究生教育指导委员会委员、民盟中央青年工作委员会委员、民盟江苏省委青年工作委员会副主任; 长期从事移动通信的教学和研究工作, 围绕蜂窝移动通信理论与关键技术、物联网理论与关键技术, 以及人工智能在移动通信中的应用等领域开展研究工作, 在多天无线传输理论与关键技术、智能通信理论与方法、智能超表面无线通信等方面取得系列创新成果; 研究成果获省部级科学技术奖一等奖3项、二等奖1项, IEEE通信学会莱斯奖, IEEE信号处理学会青年作者最佳论文奖, 《China Communications》最佳论文奖, 《Electronics Letters》最佳论文奖, 《National Science Review》最佳论文奖, 《Journal of Communications and Information Networks》青年作者最佳论文奖, 以及IEEE ICC/GLOBECOM等10余个国际重要学术会议最佳论文奖, 2014年至今连续入选爱思唯尔中国高被引学者, 2019年至今连续入选科睿唯安全球高被引学者; 共发表学术论文400余篇, 授权国际/国家发明专利50余件, 出版专著2部、教材1本。

知识指导的预训练语言模型



Knowledge-Guided Pre-Trained Language Models

韩旭/HAN Xu, 张正彦/ZHANG Zhengyan, 刘知远/LIU Zhiyuan

(清华大学, 中国 北京 100084)
(Tsinghua University, Beijing 100084, China)

DOI: 10.12142/ZTETJ.202202003

网络出版地址: <https://kns.cnki.net/kcms/detail/34.1228.tn.20220410.1321.002.html>

网络出版日期: 2022-04-12

收稿日期: 2022-02-20

摘要: 作为典型的数据驱动工具, 预训练语言模型 (PLM) 仍然面临可解释性不强、鲁棒性差等难题。如何引入人类积累的丰富知识, 是改进预训练模型性能的重要方向。系统介绍知识指导的预训练语言模型的最新进展与趋势, 总结知识指导的预训练语言模型的典型范式, 包括知识增强、知识支撑、知识约束和知识迁移, 从输入、计算、训练、参数空间等多个角度阐释知识对于预训练语言模型的重要作用。

关键词: 自然语言处理; PLM; 知识图谱

Abstract: As a typical data-driven method, pre-trained language models (PLMs) still face challenges such as poor interpretability and robustness. Hence, it is important to introduce human knowledge into these models for better performance. The latest progress and trend of knowledge-guided PLMs are introduced and the paradigm of knowledge-guided PLMs is summarized, including knowledge augmentation, knowledge support, knowledge regularization, and knowledge transfer.

Keywords: natural language processing; PLMs; knowledge graphs

1 知识的重要作用

20世纪90年代前, 研究人员将大量的精力投入到语法规则^[1-2]和专家系统^[3-4]的研究中。无论是语法规则中的语言规则还是专家系统中的知识库, 其背后的核心思想均为使用符号体系来表示语言理解所需的各类知识。这些离散稀疏的符号系统有利于抽象丰富的人类知识, 并通过人为设计的精密规则实现语言理解中的知识推理。

近些年来, 陆续构建的大型知识图谱 (知识库), 诸如 Wikidata、YAGO 和 DBpedia, 就采用了结构化的符号形式来存储海量的世界知识, 并在语言理解中发挥重要作用。近些年的研究也证明, 大规模知识图谱中的丰富知识可以有力驱动一系列人工智能和自然语言处理的应用, 例如问答系统、对话系统、文本检索和推荐系统。

符号知识的一大痛点在于难以发挥机器所擅长的数值计算优势。此外, 早期的语法规则与专家系统在泛化性上也存在问题。这就需要一套基于数值计算且具有一定泛化性的知识表示框架。统计学习^[5-6]也由此被应用于自然语言处理任务中。20世纪90年代后, 支持向量机^[7]、决策树^[8]、条件随机场^[9]的诸多经典统计模型被广泛应用, 在各类自然语言处理任务上取得了一系列突破。这些统计方法用模型参数来隐式地表示各类知识, 并基于概率计算来进行推理。相对于符号知识的“人类友好”, 这种连续数值化的模型知识更加“机器友好”。

统计模型拉开了从符号知识到模型知识的序幕, 开启了用数值表示知识的新纪元, 但统计模型本身的性能是十分有限的。近年来, 神经网络蓬勃发展, 它为数值化的知识表示及语义理解提供了更强大的工具。浅层神经网络首先被应用于知识表示中。分布式词向量表示旨在利用低维连续向量来表示词汇相关的语言知识, 并通过海量无标签文本的自监督学习来学习词向量^[10]。得益于分布式词向量中蕴含的丰富语言知识, 词的向量化表示已经成为当前完成各类自然语言处理任务的标准范式, 也有效地填补符号知识与数值计算间的鸿沟。

随着神经网络的深度与参数量的增加, 大规模预训练语言模型 (PLM) 被提出, 这推动了一系列自然语言处理任务的发展。预训练语言模型的主要特点在于其两阶段的构建方法: 第1阶段, 与分布式词向量表示类似, 在海量无标签文本上进行自监督学习, 以学习通用的语言特征和规则 (即预训练); 第2阶段, 将预训练模型在具体的自然语言处理任务上进行小规模、有标注数据的二次训练 (即微调), 以快速提升模型在这些任务中的性能, 最终形成可部署应用的模型。研究表明, 在自监督学习过程中, 预训练语言模型可以捕捉到丰富的词法知识、句法知识、语义知识、世界知识, 并通过庞大的参数将这些知识存储起来。这样一来, 微调模型的参数可以有效地将模型知识迁移到具体的任务上。

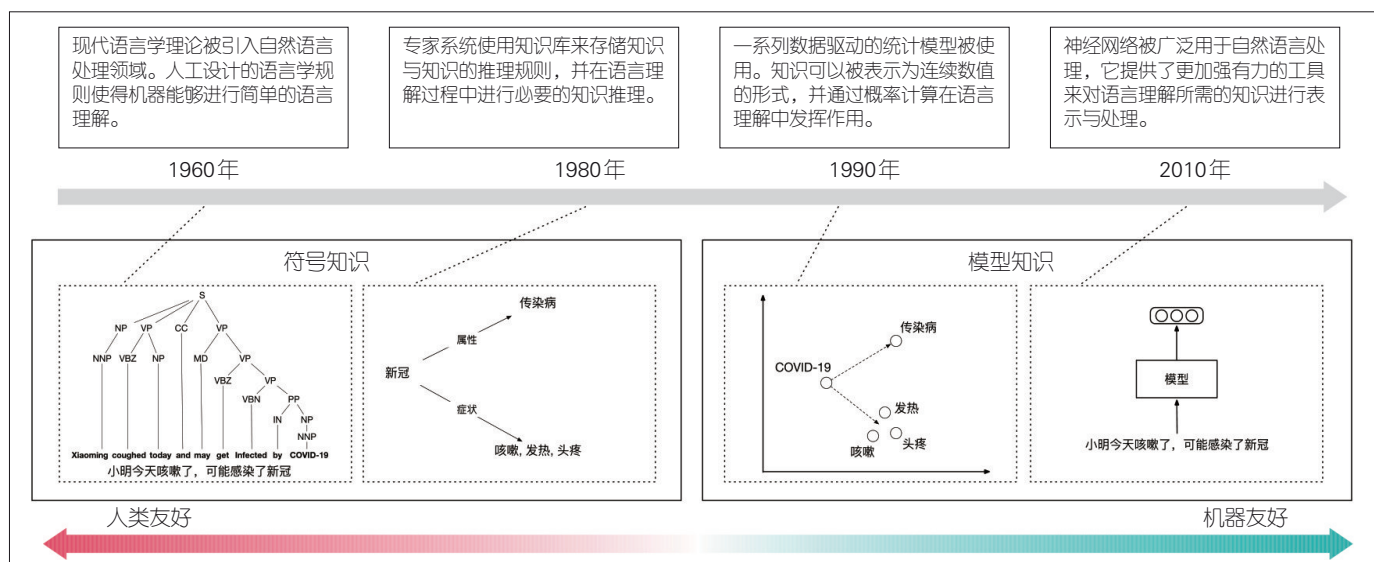
图1显示了自然语言处理技术的发展脉络, 清晰地表明

了各个时期知识是如何表示的, 以及如何被运用于语言理解的。在使用上, 符号知识与模型知识也各有优势。尽管预训练语言模型已经在当前诸多自然语言处理任务上取得了很好的效果, 但大量数据驱动下的预训练语言模型依然在可解释性、鲁棒性上存在不足。数据驱动的预训练语言模型具有善于学习的语义特征, 同时符号表示的结构化知识有着善于认知推理的特征。综合发挥以上两个优势, 形成知识指导的预训练语言模型, 对于揭示自然语言处理机理, 实现智能语言理解, 具有重要的理论意义与实用价值。

2 知识指导的预训练语言模型范式

对于如何将知识有效地应用在预训练语言模型中, 我们已在文献[11]中做了简要介绍。本文中我们进一步扩展并提出了知识指导的预训练语言模型。如图2所示, 一般来讲, 预训练语言模型有4个要素: 模型输入、模型架构、训练目标和参数空间。

- 对模型输入而言, 知识是输入的重要补充, 为文本中的关键词句提供更加有效的语义解释和语义特征;
- 对模型架构而言, 知识可以引入先验指导模型内部的特征处理流程, 进而提升模型性能;



▲图1 自然语言处理技术发展脉络^[11]

	无知识学习	知识增强	知识支撑	知识约束	知识迁移
结构风险	$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^N \mathcal{L}(y_i, f(x_i)) + \lambda \mathcal{J}(f)$	$f(x_i) \rightarrow f(x_i, k)$	$f(x_i) \rightarrow f(k(x_i))$ $f(x_i) \rightarrow k(f(x_i))$	$\mathcal{L}(y_i, f(x_i)) \rightarrow \mathcal{L}(y_i, f(x_i)) + \lambda \mathcal{L}_k(k, f(x_i))$ $\mathcal{L}(y_i, f(x_i)) \rightarrow \mathcal{L}(k(y_i), f(x_i))$	$f \in \mathcal{F} \rightarrow f \in \mathcal{F} \cap \mathcal{K}$
训练目标					
模型架构					
模型输入					
参数空间	$\begin{pmatrix} 0.1, \dots, 0.2 \\ \dots, 0.5, \dots \\ \dots, \dots, 0.4 \end{pmatrix}$	$\begin{pmatrix} 0.1, \dots, 0.2 \\ \dots, 0.5, \dots \\ \dots, \dots, 0.4 \end{pmatrix}$	$\begin{pmatrix} 0.1, \dots, 0.2 \\ \dots, 0.5, \dots \\ \dots, \dots, 0.4 \end{pmatrix}$	$\begin{pmatrix} 0.1, \dots, 0.2 \\ \dots, 0.5, \dots \\ \dots, \dots, 0.4 \end{pmatrix}$	$\begin{pmatrix} 0.1, \dots, 0.2 \\ \dots, 0.5, \dots \\ \dots, \dots, 0.4 \end{pmatrix}$

▲图2 知识指导的预训练语言模型范式^[11]

- 在训练目标上,知识可用于构造新的训练任务,提供更加丰富的训练目标,促进预训练语言模型能力的多样化;

- 在参数空间里,相比于随机初始化,用引入知识的方式来约束参数空间可以提供一个更好的参数空间初始点,有利于加速收敛,优化出更好的模型参数。

正如图2所示,知识可被应用于其中任意一部分,以起到强化预训练模型性能的作用。接下来,我们将介绍这个框架的具体内容。在图中,我们给出了结构风险函数在知识指导前后的变化。其中, x 、 y 是样本的输入输出, k 是引入的知识信息或者知识驱动的模块, f 是预训练语言模型本身, \mathcal{F} 、 \mathcal{K} 分别是参数空间、知识约束的参数空间。

2.1 知识增强

在语言表达过程中,人们习惯省略一些众所周知的背景知识。这并不影响人类对语言的理解,却不利于机器对语言的理解。知识增强旨在将这部分背景知识显式地作为补充输入,丰富上下文信息,以帮助模型更好地进行文本理解。

知识增强的方式主要有两种。第一种是直接将知识转换成文本形式,并拼接到已有文本中作为输入。最简单的做法就是将相关的结构化图谱信息转换为文本内容^[12]。在此过程中,如何找到和输入相关的知识就是一个主要挑战。基于信息检索的预训练语言模型是一个有效的解决方案,例如REALM^[13]和RAG^[14]。其预训练一个文本检索器,用于构建输入文本和背景知识文本的关联,使用时再将检索到的知识文本与输入文本拼接起来,给模型提供更加丰富的信息。

知识增强的另一种方式则是通过设计特定的知识融合模块,将文本的表示向量和相关知识向量融合在一起^[15]。这与上述文本拼接有明显不同:知识不再以符号形式进行表达,而是被蕴含在模型参数中。ELMo^[16]是该方向的代表性工作。由于ELMo是一个在超大规模语料上训练的语言模型,其表示向量可以提供丰富的语言知识,解决一词多义等问题。人们通常使用ELMo来代替传统词向量,以提升模型的基本文本理解能力。更进一步地,不少工作^[17-20]将知识图谱中的实体与关系表示为向量,并将这些向量输入到预训练语言模型以进行知识融合,这也是非常有效的知识增强方法。

2.2 知识支撑

知识支撑可以利用大量已有的知识来构建更好的结构先验。具体而言,在模型底层,知识支撑可以作为一种数据预处理模块;而在模型顶层,知识支撑可以指导模型的预测。

知识记忆网络^[21]是数据预处理模块的代表技术。根据输入特征,底层的网络结构会动态调整,以连接对应的记忆区

域,从而将记忆模块中的知识注入到模型的推理计算中。在此过程中,知识的表示形式通常为低维稠密向量,也就是所谓的模型知识。采用了记忆机制的预训练语言模型^[22-23]在多跳推理、长文本处理等需要长距离语义关系处理的任务上有显著效果。

当知识支撑作为顶层的预测指导模块时,其目标是借助知识的先验信息,构建答案之间的关联,更好地对备选答案进行筛选。在此过程中,知识的表示形式通常是符号化、层次化的。结构化知识库支撑的语言模型是该方向具有代表性的研究工作^[24-26]。在生成句子的过程中,语言模型可以利用知识库信息生成更加适合当前语境的词。

2.3 知识约束

对于知识约束,我们既可以基于已有输入数据并结合相关知识来构建训练目标,也可以直接使用外部知识来构建新数据和新目标。

知识蒸馏是一种代表性的知识约束方法^[27],也是知识结合已有输入数据来构建训练目标的典型案例。知识蒸馏能够利用大模型对已有数据进行预测,从而提供新的监督信号,帮助小模型学习取得更好的效果。具体而言,知识蒸馏要求小模型的中间计算结果和大模型的中间计算结果尽可能保持一致,包括隐层表示以及预测的标签分布。相比于单一的人工标注标签,知识蒸馏能提供更加丰富的模型知识信息。知识蒸馏已被广泛用于预训练语言模型以提升其计算效率与模型表现^[28-31]。

远程监督是另一种具有代表性的知识约束方法^[32],能够根据已有知识图谱和无监督文本自动生成大量新训练数据。远程监督在信息抽取领域获得广泛应用,大大降低了数据标注成本,显著提升了模型性能。我们给出了一个远程监督的简单示例:给定知识图谱中的三元组(包含头实体、尾实体及其关系),找出同时包含头尾实体的文本,并将其标注为该关系类型的样例。基于上述启发式规则,我们可以自动获取大量知识相关的文本分类数据来训练预训练语言模型。尽管这种自动标注方式存在噪音,如标注的样例可能并不反映头尾实体间的标注关系,但不少工作表明^[17-18,33-35],远程监督数据依然能够有效地帮助模型的训练。这些使用远程监督数据增强的预训练语言模型被验证具有强大的实体关系理解能力。

2.4 知识迁移

知识迁移的目的在于利用知识进行参数空间的约束,以降低参数空间的搜索代价,提升最终模型的性能。知识迁移

技术已被广泛应用于自然语言处理。迁移学习和自监督学习都是知识迁移的重要研究方向^[36]。各种预训练语言模型的微调阶段本身就是一种知识迁移，旨在将预训练阶段获取的丰富模型知识迁移到具体任务上。

对于预训练过程而言，最近的一些工作尝试以已有的预训练语言模型为基底来训练新的预训练模型。部分工作^[30,37]侧重于利用较小的预训练语言模型的模型知识，来降低大规模预训练模型的训练代价；而另一些工作^[38-39]则基于已有预训练语言模型的通用知识，来指导更多垂直领域的知识。

无论是对于预训练语言模型的预训练还是下游任务适配，充分迁移已有的模型知识相较于毫无基础的重新学习，在计算效率和模型效果上均有显著优势。

总之，我们从预训练语言模型的模型输入、模型架构、训练目标和参数空间4个方面入手，构建了全面的知识指导的预训练语言模型框架。在该框架下，符号知识和模型知识均可以得到充分利用，有效提升预训练模型的学习能力和模型表现。

3 预训练语言模型的知识激发

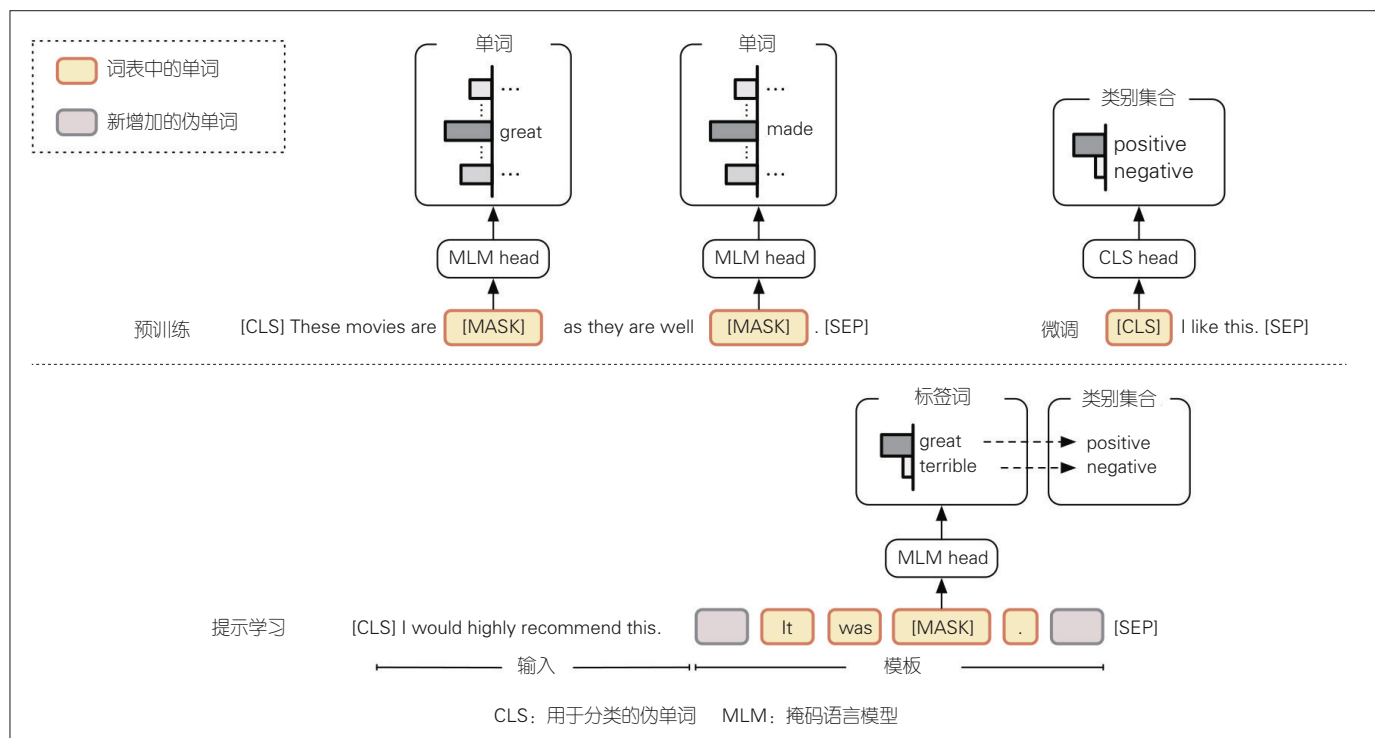
在上一章节中，我们关注的是如何将知识注入预训练语言模型之中。在这一章节中，我们将简单介绍如何激发预训练模型中的知识。这对于应用知识指导的预训练语言模型具

有重要意义。

预训练语言模型能够通过微调显著提升下游任务性能，却仍然面临着两个重要挑战：（1）预训练和微调之间的任务形式存在较大差别，预训练只考虑语言建模，但下游任务目标形式可能各有不同，这种差别会显著影响知识迁移的效能。（2）随着预训练模型参数规模迅速增加，即使进行模型微调，也需要大量技术资源。为了解决这些问题，最近学术界提出了一种新的微调技术，即提示学习（Prompt Tuning）。该技术能够有效利用大规模的模型知识，日益获得广泛关注。

提示学习的目的是将下游任务转化为类似于预训练目标的填空任务。采用相同的优化目标有利于在下游任务中更好地激发预训练模型中的知识。以情感分类的提示学习为例（图3），模型的输入由两部分组成：输入数据以及提示学习所需的提示模板。基于该输入，预训练语言模型在一组标签词中选择概率最高的词进行填空，再将预测的词映射到相应的分类标签上。图3中，提示模板为“It was [Mask]”，“[Mask]”代表需要进行填空的位置。标签词为“great”和“terrible”，“great”对应正向情感，“terrible”对应负向情感。提示微调也在一系列自然语言处理任务上取得了成效，包括文本分类^[40-43]、序列标注^[44-45]、文本生成^[46-47]等任务。

为了在下游任务上取得成功，提示模板和标签词（提示



▲图3 预训练、微调、提示学习示意图

语)需要进行精细的设计和选择。为了避免费力而复杂的提示语设计,自动搜索高质量的提示语成为目前工作的一个重点:研究者探索使用梯度优化来搜索最佳提示语^[48],或使用生成模型来提供多个候选提示语^[42],然后逐一评估其有效性,以选择最佳提示语。目前,自动搜索提示语的成本仍然很高,这限制了这些自动方法的使用场景。为此,也有研究者提出用逻辑规则指导提示学习^[49]。这种方法将先验知识编码到提示语中,降低搜索以及训练难度,使模型知识可以更好地为下游任务服务。为了避免复杂的提示设计,一些工作^[50-52]采用了可学习的提示向量来驱动预训练语言模型进行提示微调,无须变动预训练模型的任何参数,只须调整提示向量即可。

不少知识探测工作^[53-55]表明,通过设计提示模板,预训练语言模型甚至可以补全结构化知识信息。上述研究表明,除了知识模型的性质外,预训练语言模型也有一定的符号知识特性。输入提示能充分激发出预训练语言模型中各个层面丰富的知识信息,以解决具体问题。预训练语言模型在推动自然语言处理中模型知识的使用方面有着重要作用。从某种程度上而言,预训练模型也将影响自然语言处理中符号知识的使用范式。尽管预训练语言模型仍需符号知识进行强化,但其本身也是一种符号知识的优秀载体,有利于符号知识与模型知识的融合与统一。

4 结束语

在文章中,我们围绕知识对于自然语言处理的重要性、知识指导的预训练范式、预训练语言模型的知识激发3个方面,介绍了知识指导的预训练语言模型的相关技术。在各个方向上,尽管目前均已获得一些成果,但仍有许多尚未解决的重要问题。这需要研究者进一步努力,以取得突破。

致谢

清华大学姚远、李涓子和孙茂松在文章的撰写过程中,给出了宝贵的建议,在此表示感谢。

参考文献

- [1] CHOMSKY N. Syntactic structures [M]. Germany: Walter de Gruyter, 1957
- [2] NOAM C. Aspects of the theory of syntax [M]. USA: The MIT Press, 1969
- [3] WILSON S, BARR A, COHEN P R, et al. The handbook of artificial intelligence [J]. Leonardo, 1984, 17(4): 299. DOI: 10.2307/1575114
- [4] ROTH H F, WATERMAN A D. Building expert system [M]. USA: Addison-Wesley, 1983
- [5] LANDAUER T K, DUMAIS S T. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge [J]. Psychological review, 1997, 104(2): 211-240. DOI: 10.1037/0033-295x.104.2.211
- [6] PENNINGTON J, SOCHER R, MANNING C. Glove: global vectors for word representation [C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2014: 1532-1543. DOI: 10.3115/v1/d14-1162
- [7] BOSER B E, GUYON I M, VAPNIK V N. A training algorithm for optimal margin classifiers [C]//COLT'92: Proceedings of the Fifth Annual Workshop on Computational Learning Theory. ACM, 1992: 144-152. DOI: 10.1145/130385.130401
- [8] BREIMAN L, FRIEDMAN J, STONE C J, et al. Classification and regression trees [M]. USA: CRC press, 1984
- [9] LAFFERTY J D, McCALLUM A, FERNANDO C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C]//Proceedings of the 18th International Conference on Machine Learning (ICML 2001). ICML, 2001: 282-289
- [10] MIKLOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [C]//Proceedings of the 26th International Conference on Neural Information Processing Systems (NeurIPS 2013). NIPS, 2013: 3111-3119
- [11] HAN X, ZHANG Z, LIU Z. Knowledgeable machine learning for natural language processing [J]. Communications of the ACM, 2021, 64(11): 50-51
- [12] LIU W J, ZHOU P, ZHAO Z, et al. K-BERT: enabling language representation with knowledge graph [J]. Proceedings of the AAAI conference on artificial intelligence, 2020, 34(3): 2901-2908. DOI: 10.1609/aaai.v34i03.5681
- [13] GUU K, LEE K, TUNG Z. REALM: integrating retrieval into language representation models [EB/OL]. [2022-01-10]. <https://arxiv.org/abs/2005.11401>
- [14] LEWIS P, PEREZ E, PIKUS A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks [EB/OL]. (2021-04-12) [2022-01-10]. <https://arxiv.org/abs/2005.11401>
- [15] LIU Z H, XIONG C Y, SUN M S, et al. Entity-duet neural ranking: understanding the role of knowledge graph semantics in neural information retrieval [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2018: 2395-2405. DOI: 10.18653/v1/p18-1223
- [16] PETERS M, NEUMANN M, IYER M, et al. Deep contextualized word representations [C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Association for Computational Linguistics, 2018: 2227-2237. DOI: 10.18653/v1/n18-1202
- [17] ZHANG Z Y, HAN X, LIU Z Y, et al. ERNIE: enhanced language representation with informative entities [EB/OL]. (2019-06-04) [2022-01-10]. <https://arxiv.org/abs/1905.07129>
- [18] WANG X Z, GAO T Y, ZHU Z C, et al. KEPLER: a unified model for knowledge embedding and pre-trained language representation [J]. Transactions of the association for computational linguistics, 2021, 9: 176-194. DOI: 10.1162/tacl_a_00360
- [19] SU Y S, HAN X, ZHANG Z Y, et al. CokeBERT: Contextual knowledge selection and embedding towards enhanced pre-trained language models [J]. AI open, 2021, 2: 127-134. DOI: 10.1016/J.AIOPEN.2021.06.004
- [20] PETERS M E, NEUMANN M, LOGAN R, et al. Knowledge enhanced contextual word representations [EB/OL]. (2019-10-31) [2022-01-10]. <https://arxiv.org/abs/1909.04164>
- [21] WESTON J, CHOPRA S, BORDES A. Memory networks [EB/OL]. (2014-10-15) [2022-01-10]. <https://arxiv.org/abs/1410.3916>
- [22] DING M, ZHOU C, CHEN Q B, et al. Cognitive graph for multi-hop reading comprehension at scale [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2019. DOI: 10.18653/v1/p19-1259
- [23] RAE W J, POTAPENKO A, JAYAKUMAR S M, et al. Compressive transformers for long-range sequence modelling [EB/OL]. (2019-11-13) [2022-01-12]. <https://arxiv.org/abs/1911.05507>
- [24] LOGAN R, LIU N F, PETERS M E, et al. Barack's wife Hillary: using knowledge graphs for fact-aware language modeling [EB/OL]. (2019-06-17) [2022-01-10]. <https://arxiv.org/abs/1906.07241>
- [25] AHN S, CHOI H, PARNAMAA T, et al. A neural knowledge language model [EB/OL]. (2017-03-02) [2021-12-12]. <https://arxiv.org/pdf/1608.00318.pdf>
- [26] HAYASHI H, HU Z C, XIONG C Y, et al. Latent relation language models [EB/OL]. (2017-03-02) [2022-01-12]. <https://arxiv.org/abs/1908.07690>

- [27] HINTON G E, VINYALS O, DEAN J. Distilling the knowledge in a neural network [EB/OL]. (2015-03-09) [2022-01-10]. <https://arxiv.org/abs/1503.02531>
- [28] SUN S Q, CHENG Y, GAN Z, et al. Patient knowledge distillation for BERT model compression [EB/OL]. (2015-03-09) [2022-01-10]. <https://arxiv.org/abs/1908.09355>
- [29] RASHID A, LIOUTAS V, REZAGHOLIZADEH M. MATE-KD: masked adversarial TExt, a companion to knowledge distillation [EB/OL]. (2021-05-12) [2022-01-10]. <https://arxiv.org/abs/2105.05912v1>
- [30] QIN Y, LIN Y, YI J, et al. Knowledge inheritance for pre-trained language models [EB/OL]. (2021-05-28) [2022-01-12]. <https://arxiv.org/abs/2105.13880v1>
- [31] JIAO X Q, YIN Y C, SHANG L F, et al. TinyBERT: distilling BERT for natural language understanding [EB/OL]. (2019-09-23) [2022-01-10]. <https://arxiv.org/abs/1909.10351v4>
- [32] MINTZ M, BILLS S, SNOW R, et al. Distant supervision for relation extraction without labeled data [EB/OL]. [2022-01-10]. <https://paperswithcode.com/paper/distantly-supervised-ner-with-partial>
- [33] BALDINI SOARES L, FITZGERALD N, LING J, et al. Matching the blanks: distributional similarity for relation learning [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2019. DOI: 10.18653/v1/p19-1279
- [34] SUN Y, WANG S H, LI Y K, et al. ERNIE: enhanced representation through knowledge integration [EB/OL]. (2019-04-19) [2022-01-10]. <https://arxiv.org/abs/1904.09223v1>
- [35] SUN Y, WANG S H, FENG S K. Ernie 3.0: large-scale knowledge enhanced pre-training for language understanding and generation [EB/OL]. (2019-04-19) [2022-01-10]. <https://arxiv.org/abs/2107.02137>
- [36] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 2019. DOI: 10.18653/v1/n19-1423
- [37] GU X T, LIU L Y, YU H K, et al. On the transformer growth for progressive BERT training [C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021. DOI: 10.18653/v1/2021.naacl-main.406
- [38] GURURANGAN S, MARASOVIĆ A, SWAYANDIPTA S, et al. Don't stop pretraining: adapt language models to domains and tasks [EB/OL]. (2020-04-23) [2022-01-10]. <https://arxiv.org/abs/2004.10964v2>
- [39] PFEIFFER J, RÜCKLÉ A, POTH C, et al. AdapterHub: a framework for adapting transformers [EB/OL]. (2020-10-06) [2022-01-10]. <https://arxiv.org/abs/2007.07779>
- [40] LIU X, ZHENG Y N, DU Z X, et al. GPT understands, too [EB/OL]. (2021-03-18) [2022-01-11]. <https://arxiv.org/abs/2103.10385v1>
- [41] LIU X, JI K X, FU Y C, et al. P-tuning v2: prompt tuning can be comparable to fine-tuning universally across scales and tasks [EB/OL]. (2021-10-18) [2022-01-10]. <https://arxiv.org/abs/2110.07602v2>
- [42] GAO T Y, FISCH A, CHEN D Q. Making pre-trained language models better few-shot learners [EB/OL]. (2021-06-02) [2022-01-12]. <https://arxiv.org/abs/2012.15723v2>
- [43] HU S D, DING N, WANG H, et al. Knowledgeable prompt-tuning: incorporating knowledge into prompt verbalizer for text classification [EB/OL]. (2021-08-04) [2022-01-11]. <https://paperswithcode.com/paper/knowledgeable-prompt-tuning-incorporating>
- [44] DING N, CHEN Y, HAN X, et al. Prompt-learning for fine-grained entity typing [EB/OL]. (2021-08-24) [2022-01-10]. <http://121.199.17.194/paper/1430587541732179968?adv>
- [45] MA R, ZHOU X, GUI T, et al. Template-free Prompt Tuning for few-shot NER [EB/OL]. (2021-09-28) [2022-01-10]. <https://paperswithcode.com/paper/template-free-prompt-tuning-for-few-shot-ner>
- [46] DATHATHRI S, MADOTTO A, LAN J, et al. Plug and play language models: a simple approach to controlled text generation [EB/OL]. [2022-01-10]. <https://paperswithcode.com/paper/plug-and-play-language-models-a-simple>
- [47] ZOU X, YIN D, ZHONG Q Y, et al. Controllable generation from pre-trained language models via inverse prompting [C]//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. ACM, 2021. DOI: 10.1145/3447548.3467418
- [48] SHIN T, RAZEGHI Y, LOGAN R L, et al. AutoPrompt: eliciting knowledge from language models with automatically generated prompts [C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2020. DOI: 10.18653/v1/2020.emnlp-main.346
- [49] HAN X, ZHAO W L, DING N, et al. PTR: prompt tuning with rules for text classification [EB/OL]. (2021-05-24) [2022-01-10]. <https://paperswithcode.com/paper/ptr-prompt-tuning-with-rules-for-text>
- [50] LESTER B, AL-RFOU R, CONSTANT N. The power of scale for parameter-efficient prompt tuning [C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021. DOI: 10.18653/v1/2021.emnlp-main.243
- [51] GU Y X, HAN X, LIU Z Y, et al. PPT: pre-trained prompt tuning for few-shot learning [EB/OL]. (2021-09-09) [2022-01-12]. <https://paperswithcode.com/paper/ppt-pre-trained-prompt-tuning-for-few-shot>
- [52] VU T, LESTER B, CONSTANT N, et al. SPoT: better frozen model adaptation through soft prompt transfer [EB/OL]. (2021-10-15) [2022-01-12]. <https://paperswithcode.com/paper/spot-better-frozen-model-adaptation-through>
- [53] PETRONI F, ROCKTASCH T, RIEDEL S, et al. Language models as knowledge bases? [EB/OL]. [2022-01-12]. <https://paperswithcode.com/paper/language-models-as-knowledge-bases>
- [54] PETRONI F, LEWIS P, PIKTUS A, et al. How context affects language models' factual predictions [EB/OL]. (2021-10-15) [2022-01-10]. <https://paperswithcode.com/paper/spot-better-frozen-model-adaptation-through>
- [55] JIANG Z B, XU F F, ARAKI J, et al. How can we know what language models know? [J]. Transactions of the association for computational linguistics, 2020, 8: 423-438. DOI: 10.1162/tacl_a_00324

作者简介



韩旭, 清华大学计算机系2017级博士研究生; 研究方向为预训练语言模型及知识图谱; 已在ACL、EMNLP等会议及期刊发表论文50余篇, 出版专著1部。



张正彦, 清华大学计算机系在读博士研究生; 研究方向为预训练语言模型及其加速; 发表论文20余篇, 出版专著1部。



刘知远, 清华大学计算机系副教授、博士生导师, 北京智源人工智能研究院青年科学家; 研究方向为知识图谱、预训练模型等; 获得多项国家自然科学基金资助; 曾获中文信息学会青年创新奖, 入选国家青年拔尖人才支持计划、中国科协青年人才托举工程; 发表论文80余篇, 出版专著5部, Google Scholar统计引用超过1万次。

知识增强预训练模型



Knowledge-Enhanced Pre-Trained Models

王海峰/WANG Haifeng, 孙宇/SUN Yu, 吴华/WU Hua

(北京百度网讯科技有限公司, 中国 北京 100193)
(Beijing Baidu Netcom Science and Technology Co., Ltd., Beijing 100193, China)

DOI: 10.12142/ZTETJ.202202004

网络出版地址: <https://kns.cnki.net/kcms/detail/34.1228.TN.20220408.1423.006.html>

网络出版日期: 2022-04-08

收稿日期: 2022-02-16

摘要: 预训练模型主要从海量未标注、无结构化的数据中学习, 但缺少外部知识指导, 存在模型学习效率不高、模型效果不佳和知识推理能力受限等不足。如何在预训练模型中引入语言知识、世界知识等外部知识, 提升模型效果以及知识记忆和推理能力是一个难题。本文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, 详细探讨知识增强预训练模型的原理和应用。

关键词: 自然语言处理; 预训练模型; 知识增强

Abstract: Pre-trained models can automatically learn from massive unmarked and unstructured data. Nevertheless, the lack of guidance from external knowledge has dramatically hindered the learning efficiency, model effect, and reasoning capacity. There are still challenges in incorporating external supervision such as linguistics and world knowledge to improve pre-trained models' ability of knowledge memorization and reasoning. This paper provides a comprehensive review of knowledge-enhanced pre-trained models from various perspectives, such as multi-source knowledge incorporation, knowledge fusion, and knowledge forgetting alleviation. Here we take Baidu ERNIE as an example to describe the principles and applications of knowledge-enhanced pre-trained models.

Keywords: natural language processing; pre-trained model; knowledge-enhanced

自然语言处理中的预训练模型与语言模型的建立密切相关。语言模型是自然语言处理的一个重要分支。早期的语言模型能够对由单词组成的文本序列进行概率建模^[1-2], 并计算句子的联合概率。该模型技术被广泛应用于自然语言处理任务中, 例如语音识别^[3]、机器翻译^[4]等。

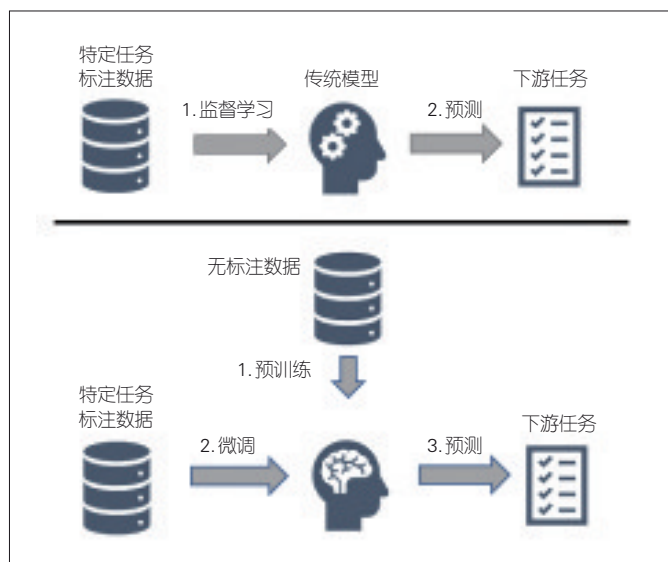
2003年, 随着深度学习技术的发展, Y. BENGIO等提出神经网络语言模型 NNLM^[5]。该模型被用来学习词的分布式表示以解决词表示的维数灾难问题。2013年, 词表示训练技术 Word2Vec^[6]被提出。该技术可使用词的上下文来对当前词进行建模, 从而学习单词的分布式向量表示。随后, 一系列词表示技术如雨后春笋般涌现, 例如基于词汇共现矩阵的 GloVe^[7]、基于字符级别 N-Gram 的 FastText^[8]等。词表示技术的提出是深度学习在自然语言处理方向应用的一座里程碑。这种技术极大地加速了自然语言处理领域的发展进程。

由于 Word2Vec^[6]词表示技术仅能将语言中的词语映射到一个静态的、与上下文无关的语义表示空间上, 因此该技术无法解决语言中的一词多义问题。2018年, ELMo 模型^[9]采用了双向长短期记忆 (LSTM) 网络对文本序列的上下文进行建模。该模型能够将双向语言模型中的不同层表示进行融

合, 并计算上下文相关的词表示, 在一定程度上解决了一词多义问题。紧接着, BERT 模型^[10]使用双向 Transformer 网络^[11]对文本序列进行建模, 并采用预训练-微调方法一举刷新众多自然语言理解任务的基准纪录。预训练模型技术的成熟进一步推动了自然语言处理的发展。

与传统监督学习方法不同, 基于自监督学习方法的预训练-微调首先对大规模无标注数据进行学习, 然后再对小规模任务标注数据进行微调, 如图 1 所示。由于能同时对未标注文本和标注文本进行学习, 预训练-微调方法取得了远超传统监督学习的效果, 并且显著缩小了任务标注数据的规模^[10,12]。因此, 预训练-微调方法逐渐成为自然语言处理领域的应用范式。近期, 基于超大规模预训练模型的预训练-提示方法^[13]取得了能够与预训练-微调方法相媲美的效果, 并逐渐成为自然语言处理领域的又一范式。该方法可将下游任务改造为自然语言表达形式, 使下游任务的建模形式更接近预训练模型的学习过程, 从而挖掘出预训练模型强大的零样本和小样本学习能力。

得益于深度学习技术和硬件算力的飞速发展, 以 BERT^[10]、GPT-3^[12]、ERNIE 3.0^[14]为代表的预训练语言模型



▲图1 传统监督学习(上)与预训练-微调(下)的对比

在自然语言理解、语言生成、机器翻译、人机对话等领域取得了突破性进展。预训练模型的出现使得人们对自然语言处理领域的研究重点从过去的结构工程转移到目标工程上,即从设计不同的网络结构并引入相应的归纳偏置,转移到基于统一的Transformer模型来设计启发式的预训练目标。预训练模型凭借自监督学习方法和预训练-微调应用方法,已逐步占据自然语言处理领域的主导地位。

当前的预训练模型主要依赖大量无结构化数据的学习。由于缺少外部知识指导,这些模型存在学习效率不高、模型效果不佳和知识推理能力受限等问题。因此,如何使用知识来增强预训练模型的表示能力,是预训练模型研究和应用的难点之一。目前,主流的知识增强预训练模型主要分为两类。一类模型可通过弱监督方法,对文本中蕴含的知识进行标注,然后设计知识类预训练任务,以便对文本中的知识进行学习。例如,ERNIE 1.0^[15]通过对数据中的短语和实体进行标注并掩码,来学习文本中的知识。文献[16]对实体知识进行替换,使语言模型能够根据上下文信息对知识图谱中的实体和关系进行推断,从而加强对文本序列知识的学习。另一类模型可对构建好的结构化知识库和无结构化文本进行联合预训练学习,例如K-BERT^[17]、CoLAKE^[18]和ERNIE 3.0^[14]。通过对结构化知识和海量无结构化数据的联合学习,知识增强的预训练模型可以很好地提升知识记忆能力和推理能力^[14]。

1.3 种知识增强预训练模型

根据融合知识的类型和作用,本文将预训练模型分为3

类:融合语言知识的预训练模型、融合世界知识的预训练模型和融合领域知识的预训练模型。

1.1 融合语言知识的预训练模型

语言知识是理解自然语言的基础,主要包含词法知识、句法结构知识、语义知识等。预训练模型对语言知识的融合方法有两种:一种是通过自动标记无标注文本中的语言知识来指导预训练模型的学习,另外一种融合人工构建的语言知识库。ERNIE-Gram^[19]通过构建基于N-Gram的多粒度掩码语言模型,可同时学习N-Gram内部和N-Gram之间的语义关系,使模型能够同时捕获细粒度和粗粒度语言知识,显著提升了模型的语义表示能力。除了融合语言粒度知识外,也有工作研究如何学习句子中的语义关系。通过在预训练的过程中对指代消解进行建模,CorefBERT^[20]增强了模型对语义知识的学习能力。其中,“指代”是自然语言表达中的常见现象。基于在一段文本中多次出现的命名实体是同一个事物的假设,CorefBERT提出提及指代预测算法。通过预测文本中被掩盖的、重复出现的命名实体,该算法提升了模型对指代关系的建模能力。

上述方法主要对无标注数据中蕴含的人类知识进行标注,让模型通过学习标注信息来融合语言知识。此外,也有研究将人工构建的语言知识库融合到预训练模型中。其中,WordNet^[21]和HowNet^[22]是具有代表性的语言知识库。这些知识库含有丰富的语言知识。以WordNet为例,它将不同词性的单词各自组成一个同义词集合。每个同义词集合各表示一个基本的语义概念。WordNet利用语义关系将这些集合连接成网络。其中,每个词语均有对应的解释和例句。SenseBERT^[23]融合了WordNet中的超义等概念知识。通过还原被掩盖的词并预测其对应的超义,该模型可以显式学习词语在给定语境下的语义信息。SenseBERT在词义消歧等任务上的效果取得了显著提升。LIBERT^[24]利用WordNet中词语与词语间的同义关系和上下位关系设计了词汇关系分类预训练任务过程,增强了预训练模型对语义信息的建模能力,在大部分自然语言处理任务上的效果均有提升。

1.2 融合世界知识的预训练模型

人类在认识世界的过程中产生了大量的世界知识。其中,部分知识可以利用实体以及实体之间的关系进行描述,比如“安徒生”创作了“《夜莺》”。研究者通过知识图谱^[25]来表达这些世界知识。在知识图谱中,实体表示网络中的一个节点,实体间的关系则表示对应节点间的边。利用知识图谱存储世界知识,并让模型显式学习人类对世界的认

知，是融合世界知识的预训练模型采用的重要方法。KEPLER^[26]将预训练上下文编码器与知识模型相结合，使得预训练模型不仅可以将图谱三元组中的事实知识更好地融合到模型中，而且还可以通过丰富的实体描述，有效地学习实体和关系的知识表示。不同于KEPLER，有的模型将语言 and 知识进行统一表示。CoLAKE^[18]将文本序列视为一个全链接的词图，并以每个实体为锚点，将文本中实体所对应的知识图谱中的子图进行连接，以构成一个同时包含词语、实体和关系的词语-知识图。通过学习词语-知识图，模型能够同时融合训练语料中的语言知识和图谱中的世界知识。然而，CoLAKE主要侧重实体在知识图谱中的建模，却忽视了实体在训练语料中的表述。为此，ERNIE 3.0^[14]提出知识图谱与文本平行预训练的方法，使用文本来表述知识。ERNIE 3.0突破了异构结构化知识表示与无结构文本表示难以统一建模的瓶颈。

1.3 融合领域知识的预训练模型

人工智能行业应用存在着丰富的、由众多行业专家积累的专业知识。当前的预训练模型主要依赖互联网数据进行训练。数据中缺乏行业相关的领域知识，导致预训练模型在专业领域的自然语言处理任务上的表现不佳。以医疗领域为例，CBLUE^[27]的应用表明，通用预训练模型处理该类任务的效果差于人类。为了增强预训练模型在专业领域的应用效果，研究者们对如何将领域知识融入到预训练模型进行了探索。BioBERT^[28]是一个生物医学领域的预训练模型。实验表明，在生物医学语料库上的预训练可以显著提高模型在生物医疗领域任务上的性能。针对领域知识的预训练方法，ERNIE-Health利用医疗实体掩码算法对专业术语等实体知识进行学习。同时，通过医疗问答匹配任务，该模型能对病状描述与医生专业治疗方案的对应关系进行学习，可获得医疗实体知识之间的内在联系，在包含医学信息抽取、医学术语归一化等中文医疗文本处理任务上的效果取得了显著提升。进一步地，结合世界知识和领域知识的学习方法，BERT-MK^[29]基于医疗知识图谱的子图进行学习，提高了预训练模型在医疗领域任务上的应用效果。

为了充分地融合领域知识，以FLAN^[30]、ExT5^[31]和T0^[32]为代表的模型分别收集了60、107、171个领域的任务数据，并针对每项任务设计了任务模板。将多种多样的任务转化为由文本至文本生成的统一格式，使模型在预训练阶段就能融合并使用多领域、多任务的知识，可显著提高模型的通用能力与泛化性能。PPT^[33]延续了将多种任务通过模板转化为统一格式的方式，在预训练阶段就可对连续提示词进行领域知

识的学习，提升了模型在训练样本匮乏的下游任务上的少样本迁移能力。

知识增强预训练模型通过融合多种类型的外部知识来显著提升自身性能。然而，在学习知识的过程中，模型通常存在知识遗忘问题，即在学习新的知识后会忘记之前学过的知识。因此，如何解决知识遗忘问题显得非常重要。为了避免知识遗忘，ERNIE 2.0^[34]构建了持续预训练的框架。在该框架下，每当引入新任务时，该框架可在学习该任务的同时仍记住之前学过的知识。此外，K-ADAPTER^[35]通过不同的适配器来学习世界知识和语言知识。在下游任务中，该方法能够将不同适配器产生的特征表示进行拼接，并生成同时具有各种知识的表示，从而将多种知识同时应用到任务中，有效解决了知识遗忘问题。

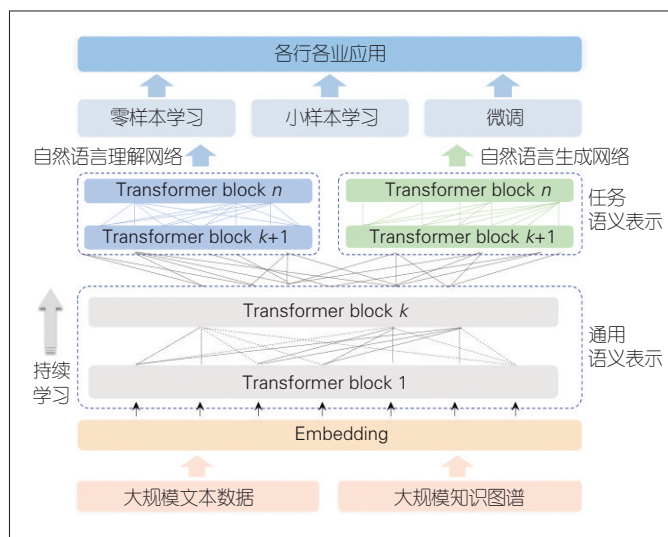
2 文心知识增强预训练模型

本文中，我们将以百度文心（ERNIE）知识增强预训练模型为例，详细阐述知识增强预训练模型的模型结构、知识融合方法，以及该模型在知识增强跨语言预训练模型、知识增强跨模态预训练模型上的扩展。文心是最早探索预训练模型融入知识的工作^[15]之一，并在文献[14]和文献[34]等工作中逐步迭代。其中，最新的ERNIE 3.0 Titan模型^[36]使用2 600亿个参数，在海量的未标注文本数据和大规模知识图谱中持续学习，突破了多源异构数据难以统一表示与学习的瓶颈，在60余项任务上的表现是最好的。

2.1 模型结构

文心使用了一种通用语义表示与任务语义表示相结合的模型框架，如图2所示。该框架融合了自编码和自回归等不同的任务语义表示网络。因此，文心既可以同时完成语言理解和语言生成任务，又能进行无标注数据的零样本学习和有标注数据的微调训练。该模型结构共包括两层：第1层是通用语义表示网络，该网络主要学习数据中的基础知识和通用知识；第2层是任务语义表示网络，该网络可基于通用语义表示来学习与任务相关的知识。不同任务语义表示网络可通过自编码结构或者自回归结构来实现。底层共享有助于这些任务语义表示网络实现交互和增强。在学习过程中，任务语义表示网络只学习对应类别的预训练任务，而通用语义表示网络则学习所有的预训练任务。

文心将Transformer^[11]作为基础的模型结构，通过多层统一的自注意力机制，采用并行计算的方式来获得词与词之间的关系权重，并根据所得到的权重来生成每个词在整段语义单元的动态词表示。为了增强模型对长距离语义知识的建模



▲图2 文心模型结构

能力，文心引入了递归性记忆单元^[37]，并在此基础上形成了一种增强记忆力机制^[38]，使模型能够对超长文本进行建模。

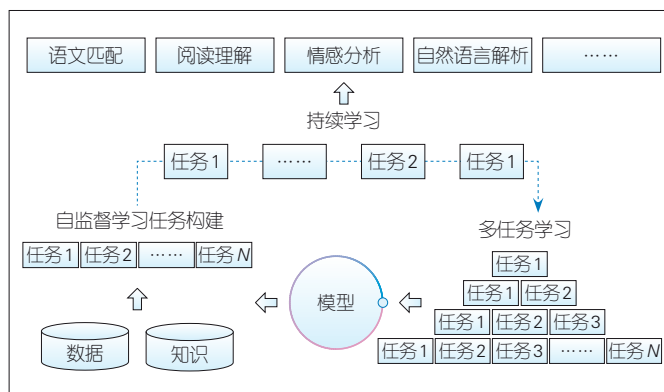
2.2 持续学习方法

ERNIE 2.0^[34]拥有一种持续学习的预训练框架，可增量学习海量数据中的知识，持续提升语义理解效果。如图3所示，知识可通过预训练任务的形式加入训练框架。每当引入新的预训练任务时，该框架可在学习新任务的同时学习之前的任务。新任务与旧任务之间通过多任务进行学习可避免知识遗忘。基于该框架，模型可以快速学习词法、结构、语义层面的语言知识、实体-关系世界知识等。模型的通用能力可得到大幅提升。ERNIE 2.0将这种学习方式与传统的持续学习及多任务学习进行对比，结果证明了该方法的有效性。

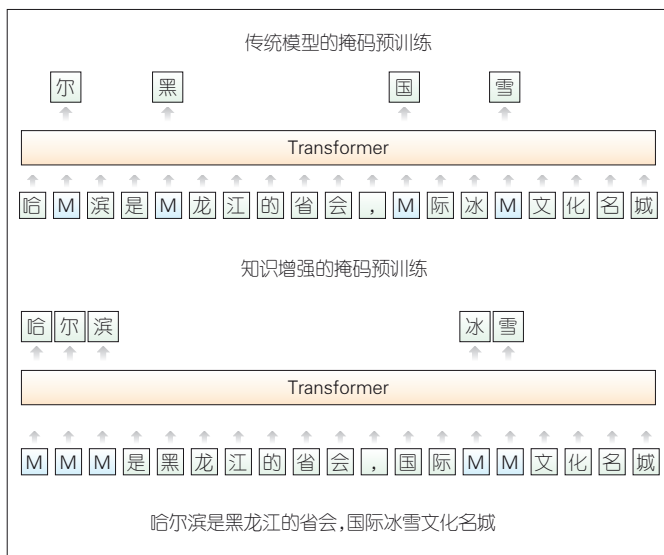
2.3 知识融合方法

2.3.1 语言知识融合方法

ERNIE 1.0^[15]模型提出了知识增强的预训练方法，即知识掩码预训练方法。该模型通过对海量数据中的字、词、实体等不同语言单元和知识进行建模，来学习不同粒度语言知识的完整语义。图4给出了传统预训练模型和ERNIE 1.0学习方法的对比。在预测还原过程中，传统预训练模型通过诸如“哈尔滨”“黑龙江”等短距离固定记忆对被掩码的字进行还原，难以学习到“哈尔滨”“黑龙江”等命名实体的完整语义。而在ERNIE 1.0的学习过程中，只有学习到“哈尔滨”“黑龙江”等命名实体的关系，“哈尔滨”这一命名实体的属性才能正确预测被掩盖的知识。ERNIE 1.0本身可基



▲图3 文心模型中的持续学习语义理解框架



▲图4 文心语言知识学习方法

于字特征输入完成建模，在应用时不需要依赖其他信息，具有很强的通用性和可扩展性。例如，在对红色、绿色、蓝色等表示颜色的词语进行建模时，ERNIE 1.0通过相同字的语义组合可以学习词之间的语义关系。

在语义知识融合方面，短句中的连词往往准确地表示了它们的细分逻辑语义关系。例如，在“因为人们的滥砍乱伐，所以今年以来洪涝不断”中，“人们的滥砍乱伐”和“近年来洪涝不断”就是因果关系；“尽管风雨交加，但是同学们还是坚持按时到校上课”中的“风雨交加”和“同学们还是坚持按时到校上课”之间就是转折关系。为了能够实现短句间的逻辑关系建模，文心构建了逻辑关系知识：首先将具有逻辑关系的句子挖掘出来，然后再将句子中的连词去掉，最后让模型进行无监督的逻辑关系分类。

2.3.2 世界知识融合方法

ERNIE 3.0在引入蕴含丰富世界知识的大规模知识图谱

后，实现了海量无监督文本与大规模知识图谱的平行预训练。以图 5 为例，ERNIE 3.0 在训练过程中会将文本端信息和知识端信息同时输入到模型中进行训练。知识端信息会输入图谱中的三元组。例如，“安徒生”“作品”“《夜莺》”三元组代表了《夜莺》是安徒生的作品这一世界知识。文本端就会使用三元组中的“安徒生”和“《夜莺》”在海量文本中检索出与之相关的句子。ERNIE 3.0 在训练过程中使用联合掩码进行训练。训练过程主要包括两个方面：在知识端方面，由于知识图谱中的世界知识片段会被掩盖，模型需要通过文本中的信息对知识端被掩盖的信息进行推理；在文本端方面，由于无标注文本的语言知识片段也会被掩盖，模型需要通过图谱中的结构化信息对文本端被掩盖的信息进行还原。这种方式促进了结构化的知识和无结构文本之间的信息共享，大幅提升了模型对知识的记忆和推理能力。

与 CoLAKE^[18]、K-BERT^[17]、KG-BART^[39]、KnowBert^[40] 等融入知识图谱的工作原理不同，ERNIE 3.0 利用知识图谱中三元组文本表述和对应的文本信息，在统一的空间同时对知识端和文本端进行平行学习。而先前的知识增强方法在融合知识与文本时使用了不同的编码结构，使得知识与文本只能在不同的表示空间中被学习。大部分研究工作只强调知识对文本的增强，却忽略文本对知识的作用，致使文本与知识的交互不充分。ERNIE 3.0 增强了结构化知识与无结构文本间的双向交互，提升了模型对知识的理解与推理能力。

2.4 文心模型效果分析

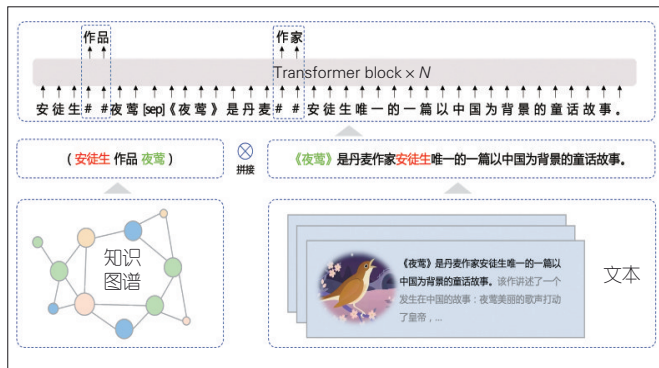
文心所使用的一系列知识增强方法显著提升了模型效果和学习效率，增强了知识推理能力。

知识增强预训练模型显著提升了下游任务效果。通过知识融合，相对于其他预训练模型，ERNIE 3.0^[14]模型在包括情感分析、信息抽取、对话生成、数学计算、阅读理解等 21 类 54 个自然语言理解和生成数据集上的效果是最好的。表 1 表明，在语义匹配、文本摘要等任务上，只用 3% 的参数量，知识增强预训练模型就可以达到甚至超过百亿参数非知识增强预训练模型^[41-42]的效果。同时，百亿参数的知识增强预训练模型效果可以得到进一步提升。

知识增强预训练模型的知识推理能力也得到了进一步提升。图 6 给出了 ERNIE 3.0 Titan 模型和 GPT-3 模型在知识问答数据集上的对比效果^[36]。其中，ERNIE 3.0 Titan 的准确率比 GPT-3 高 8%。

2.5 知识增强预训练模型扩展

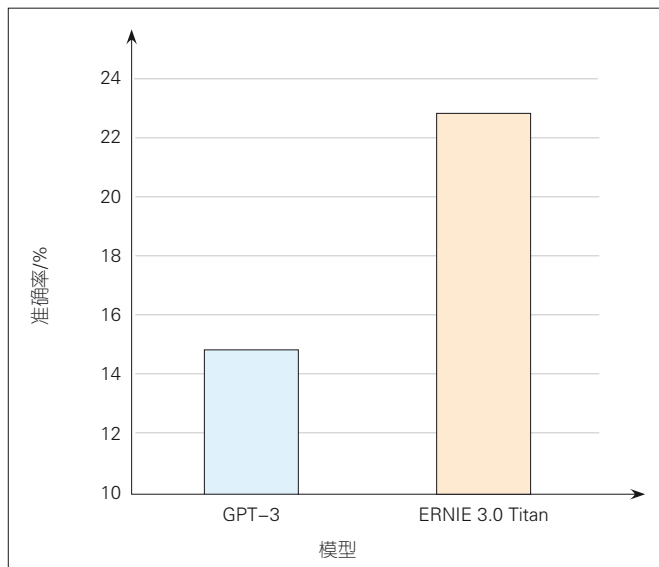
在单语言理解与生成预训练模型的基础上，为了融合更



▲图 5 文心中的文本与知识平行预训练

▼表 1 传统模型与知识增强模型效果对比

模型	是否知识增强	参数规模/亿	语义匹配准确率/%	文本摘要指标 (Rouge-L)
mT5-XXL	否	110	88.3	34.8
CPM-2	否	130	89.2	35.9
文心	是	3	89.1	41.4
		100	90.4	48.4



▲图 6 GPT-3 和 ERNIE 3.0 Titan 知识问答效果

多维度的知识，文心进一步衍生出知识增强跨语言模型和知识增强跨模态模型。

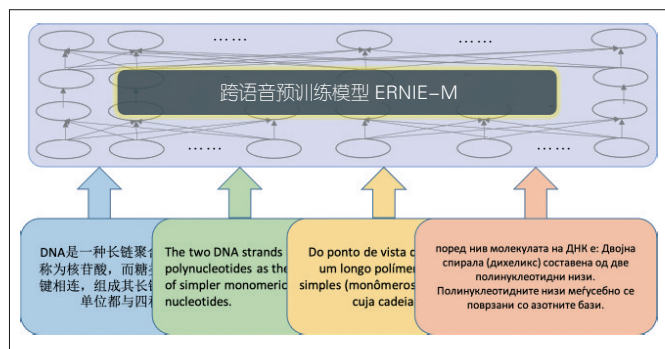
2.5.1 知识增强跨语言预训练模型

不同语言中的语料蕴含了不同地区的人们在历史发展过程中收集的不同知识。受限于语料的不完备性，模型从单一语言的语料中难以完全学到跨语言知识。因此，我们需要探索将多种语言数据中的知识进行融合的方法，以提升模型能力，解决单一语言数据的知识稀疏性问题。

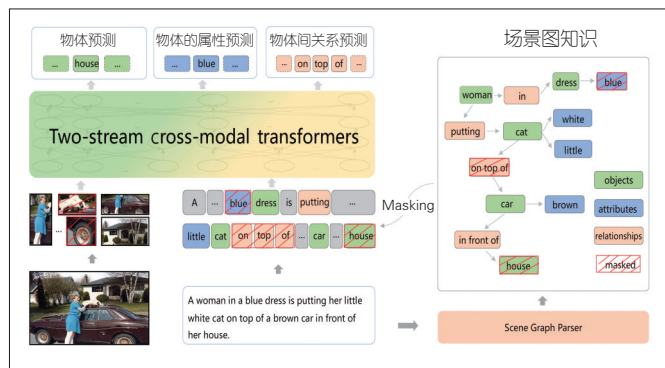
知识增强跨语言预训练模型^[43]实现了从多种语言数据中进行跨语言知识学习的目标。在预训练过程中, ERNIE-M会使用统一的模型同时对海量未标注的多语言数据进行建模,从而统一学习跨语言知识和跨语言语义表示。如图7所示,对于DNA这一知识,不同语言的语料蕴含了不同的信息。因此,模型可以从不同语言中学到跨语言知识的不同侧面。在跨语言预训练模型使用某种语言的任务数据进行训练后,其他语言的相同任务无须进行进一步训练,即可实现跨语言迁移。这种跨语言迁移方式能够解决低资源语言任务数据稀疏性问题,有助于实现任务知识在不同语言间的迁移。从单语语料中学习多语间的隐式语义对齐知识的方法,能够突破双语平行语料规模对跨语言模型的限制。ERNIE-M对96种语言进行统一建模,并在5项跨语言任务中取得了最好的效果^[43]。

2.5.2 知识增强跨模态模型

跨模态表示学习的目标是,通过对齐语料学习跨模态的通用联合表示,将各个模态之间的语义对齐信号融合到联合表示中,从而提升下游任务效果。目前的视觉-语言跨模态预训练方法,例如ViLBERT^[44]等,在预训练过程中无法区分普通词和与场景相关的词,学到的联合表示也无法实现模态



▲图7 知识增强跨语言模型ERNIE-M



▲图8 跨模态知识增强模型ERNIE-ViL

间细粒度语义(如物体、物体属性、物体间关系)的对齐。

ERNIE-ViL^[45]将包含细粒度语义信息的场景图先验知识融入视觉-语言跨模态预训练过程中,如图8所示。基于场景图的结构化知识,ERNIE-ViL创建物体预测、属性预测、关系预测3个预训练任务,在预训练过程中更加关注细粒度语义的跨模态对齐,从而可以学习到能够刻画更好跨模态语义对齐信息的联合表示,并提升自身在视觉问答、视觉常识推理、引用表达式理解、跨模态文本-图像检索等5个多模态典型任务上的应用效果。

3 知识增强预训练模型应用

随着预训练技术的快速发展,知识增强预训练模型有着非常广阔的应用场景,例如搜索引擎、推荐系统、智能创作、人机对话、文档分析、金融风控、智慧医疗等。这里,我们将从搜索引擎、人机对话、行业领域应用3个方面,详细阐述知识增强预训练模型的应用。

3.1 搜索引擎应用

搜索引擎通过对网页内容和用户查询请求进行分析和理解,让用户可以在海量的互联网数据中查询到所需的信息。通用的预训练模型很好地提升了搜索引擎效果,例如:谷歌在BERT问世一年之际宣布将预训练模型应用到搜索引擎中,并称BERT比以往任何技术都能更好地理解用户搜索意图;微软将Turing-NLG模型应用在必应搜索方案中,使得搜索引擎在搜索框内即可辅助用户完成查询词的输入;在中文搜索引擎中,百度将知识增强的文心模型运用到搜索引擎的不同检索阶段,包括端到端的大规模语义索引系统^[46]、精细化语义相关性建模^[47]、智能问答等。得益于基于大规模文本和大规模知识的自监督训练,文心模型可以帮助搜索引擎更加准确地理解网页内容和用户查询语句,从而提升搜索结果的准确性。传统的搜索引擎通过文章中的词语建立倒排索引,并通过统计相同词语的个数等方式来计算查询词与网页的相关性。这种方式只能为用户返回字面上匹配的内容。基于知识增强预训练模型的搜索引擎,通过查询请求和网页内容的统一语义表示,实现了基于语义理解与匹配的搜索,使搜索效果显著提升。

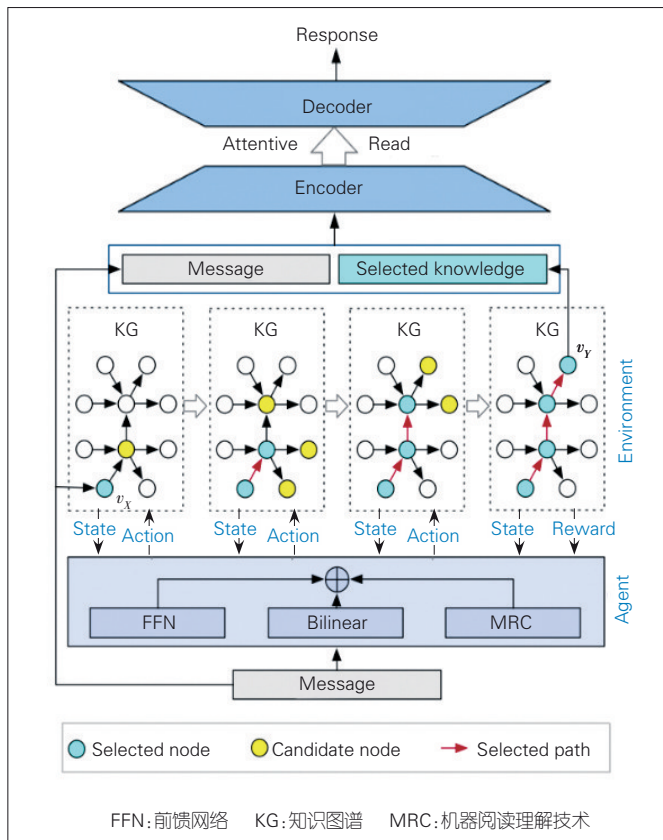
除了应用于搜索引擎的检索阶段和排序阶段之外,文心知识增强模型也能对用户搜索查询的意图进行分析与识别。用户搜索意图识别的准确性将直接影响用户使用搜索引擎的满意度。传统的用户意图识别方法多基于监督学习方法,受限于标注数据的覆盖度,对冷门知识信息搜索查询的识别准确率并不高。而基于文心的用户搜索意图识别方法,能够学

习大量的数据和知识，具备更强的泛化性，使得冷门知识信息搜索意图准确率比传统方法高12%。

3.2 人机对话应用

让机器像人一样有逻辑、有知识、有情感地与人对话，是人机交互的重要发展方向之一。知识增强的对话预训练模型通过对海量无标注数据和大规模知识的学习，使人机对话系统可以更容易模仿人与人的交互方式，让人使用更加自然的方式与机器交流。典型的应用包括智能音箱、智能客服、智能车载等。

文心系列模型包含了基于知识增强的对话预训练模型PLATO^[48-50]。基于PLATO模型，我们探索了知识内化和知识外用两种知识增强技术，如图9所示。知识内化是指，在训练阶段，模型将知识信息内化到模型参数中。通过多阶段的模型训练方式来引入大规模通用领域问答知识，可使PLATO融入生成问答能力，进而将问答准确率从3.2%提升至90%。知识外用是指，在推理阶段，模型动态地引入外部知识以指导回复生成。这两种方式能够有效提升PLATO多轮对话的内容丰富度和主题连贯性。



▲图9 知识增强的对话预训练模型

3.3 行业领域应用

知识增强预训练模型在医疗、金融、媒体等人工智能行业中表现出极大的应用价值。

在医疗行业中，中国的医疗卫生事业存在医疗资源不平衡、医生人力短缺等问题。基于知识增强预训练模型构建的临床医疗辅助技术是解决这些问题的关键技术之一。知识增强的医疗语义理解与图推理模型^[51-52]，可实现医学知识的计算，并通过患者场景化子图推断，实现可循证的医学决策。该技术突破了以往数据驱动的深度学习方法不可解释的局限，大幅提升了推理决策效果，具备贴合医学临床诊疗思维的优点，改善了临床辅助决策和智能诊前助手等场景应用效果，提高了医护人员临床工作效率。

在金融行业中，知识增强的文心模型被用于金融文本分析，提高了企业对金融信息的处理与决策效率。金融行业需要处理大量的文本信息，例如企业新闻、行业报道、招股书、财报、合同等。在传统模式下，金融从业人员很难从海量文本中获得有效信息。而基于文心模型构建的金融知识计算引擎能够帮助他们从海量的金融文本中快速查找有用的关键信息。例如，文心模型能够对保险合同中的条款文本进行解析，可实现39个维度的关键信息抽取，使单份合同的处理时间从30 min降低到1 min，能显著提升金融从业人员的工作效率和决策能力。

在媒体行业中，知识增强的文心模型对语言、知识和创作成果进行持续学习，能够实现智能辅助创作。在文章撰写的过程中，基于文心模型的智能创作引擎会对全网热点资讯进行系统分析与计算，为撰稿人提供素材推荐、智能纠错、标题生成、用词润色、文章审校等全方位的帮助。除了自动创作文本外，知识增强的跨模态文心模型实现了以文生图。文心模型可根据文章的文字内容输出具有原创性和艺术性的图片，并将其作为文章的配图使用，进一步丰富内容创作。在知识增强预训练模型的帮助下，智能创作平台将人类从重复劳动中解放出来，有效提升了内容生产的效率和效果。

4 结束语

本文系统阐述了知识增强预训练模型的发展脉络，分析了现有知识增强预训练模型对语言知识、世界知识、领域知识等知识的融合方法，重点介绍了文心知识增强预训练模型的原理、方法和应用效果。通过搜索引擎、人机对话、行业应用3个方面详细介绍了知识增强预训练模型的应用。

知识增强预训练模型已经取得长足发展，但诸多研究方向依然面临巨大挑战。例如，由于知识的稀疏性，现有知识增强预训练模型依旧难以解决逻辑、常识等问题；由于模型

是基于深度神经网络方法来建立的,模型的可解释性、可靠性和可控性仍然较差。因此,如何使模型更具常识性,如何提升模型的可解释性和可靠性,以及如何将跨模态知识、符号化知识与深度学习进行深度融合,都是知识增强预训练模型未来发展的重要方向。

参考文献

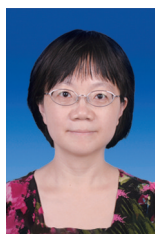
- [1] SUEN C Y. N-gram statistics for natural language understanding and text processing [J]. IEEE transactions on pattern analysis and machine intelligence, 1979, 1(2): 164–172. DOI: 10.1109/tpami.1979.4766902
- [2] BROWN P, PIETRA D, MERCER R. Class-based N-gram models of natural language [J]. Computational linguistics, 1992, 18(4): 467–480
- [3] ODELL J J, VALTCHEV V, WOODLAND P C, et al. A one pass decoder design for large vocabulary recognition [C]//Proceedings of the Workshop on Human Language Technology–HLT '94. Association for Computational Linguistics, 1994: 405–410. DOI: 10.3115/1075812.1075905
- [4] MARIÑO J B, BANCHS R E, CREGO J M, et al. N-gram-based machine translation [J]. Computational linguistics, 2006, 32(4): 527–549. DOI: 10.1162/coli.2006.32.4.527
- [5] BENGIO Y, DUCHARME D, VINCENT P, et al. A neural probabilistic language model [J]. Journal of machine learning research, 2003, 3(6): 932–938
- [6] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space [EB/OL]. [2022–02–15]. <https://arxiv.org/abs/1301.3781>
- [7] PENNINGTON J, SOCHER R, MANNING C. Glove: global vectors for word representation [C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2014: 1532–1543. DOI: 10.3115/v1/d14-1162
- [8] BOJANOWSKI P, GRAVE E, JOULIN A, et al. Enriching word vectors with subword information [J]. Transactions of the association for computational linguistics, 2017, 5: 135–146. DOI: 10.1162/tacl_a_00051
- [9] PETERS M, MAUMANN M, IYYER M, et al. Deep contextualized word representations [EB/OL]. [2022–02–15]. <https://aclanthology.org/N18-1202.pdf>
- [10] DEVLIN J, CHANG M, LEE K, et al. Pre-training of deep bidi-rectional transformers for language understanding [EB/OL]. [2022–02–15]. <https://aclanthology.org/N19-1423.pdf>
- [11] VASWANI A, SHAZEER N M, PARMAR N, et al. Attention is all you need [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. ACM, 2017: 6000–6010
- [12] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners [EB/OL]. [2022–02–15]. <https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf>
- [13] SCHICK T, SCHÜTZE H. Exploiting cloze-questions for few-shot text classification and natural language inference [C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2021: 255–269. DOI: 10.18653/v1/2021.eacl-main.20
- [14] SUN Y, WANG S H, FENG S, et al. Ernie 3.0: large-scale knowledge enhanced pre-training for language understanding and generation [EB/OL]. [2022–02–15]. <https://arxiv.org/abs/2107.02137>
- [15] SUN Y, WANG S H, LI Y K, et al. ERNIE: enhanced representation through knowledge integration [EB/OL]. [2022–02–15]. <https://arxiv.org/abs/1904.09223v1>
- [16] XIONG W H, DU J F, WANG W Y, et al. Pretrained encyclopedia: weakly supervised knowledge-pretrained language model [EB/OL]. [2022–02–15]. <https://arxiv.org/abs/1912.09637>
- [17] LIU W J, ZHOU P, ZHAO Z, et al. K-BERT: enabling language representation with knowledge graph [EB/OL]. [2022–02–15]. <https://arxiv.org/abs/1909.07606>
- [18] SUN T, SHAO Y, QIU X, et al. CoLAKE: contextualized language and knowledge embedding [EB/OL]. [2022–02–15]. <https://arxiv.org/abs/2010.00309>
- [19] XIAO D L, LI Y K, ZHANG H, et al. ERNIE-gram: pre-training with explicitly N-gram masked language modeling for natural language understanding [EB/OL]. [2022–02–15]. <https://arxiv.org/abs/2010.12148>
- [20] YE D, LIN Y, DU J, et al. Coreferential reasoning learning for language representation [EB/OL]. [2022–02–15]. <https://aclanthology.org/2020.emnlp-main.582.pdf>
- [21] MILLER G. Wordnet: a lexical database for English [J]. Communications of the ACM, 1995, 38(11):39–41
- [22] DONG Z D, DONG Q. HowNet: a hybrid language and knowledge resource [C]//Proceedings of International Conference on Natural Language Processing and Knowledge Engineering. IEEE, 2003: 820–824. DOI: 10.1109/NLPKE.2003.1276017
- [23] LEVINE Y, LENZ B, DAGAN O, et al. SenseBERT: driving some sense into BERT [EB/OL]. [2022–02–15]. <https://arxiv.org/abs/1908.05646>
- [24] LAUSCHER A, VULIC I, PONTI E, et al. Specializing unsupervised pretraining models for word-level semantic similarity [EB/OL]. [2022–02–15]. <https://arxiv.org/abs/1909.02339>
- [25] JI S X, PAN S R, CAMBRIA E, et al. A survey on knowledge graphs: representation, acquisition, and applications [J]. IEEE transactions on neural networks and learning systems, 2022, 33(2): 494–514. DOI: 10.1109/TNNLS.2021.3070843
- [26] WANG X Z, GAO T Y, ZHU Z C, et al. KEPLER: a unified model for knowledge embedding and pre-trained language representation [J]. Transactions of the association for computational linguistics, 2021, 9: 176–194. DOI: 10.1162/tacl_a_00360
- [27] ZHANG N, BI Z, LIANG X, et al. Cblue: a Chinese biomedical language understanding [EB/OL]. [2022–02–15]. <https://arxiv.org/abs/2106.08087>
- [28] LEE J, YOON W, KIM S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining [J]. Bioinformatics, 2019, 36(4): 1234–1240. DOI: 10.1093/bioinformatics/btz682
- [29] HE B, ZHOU D, XIAO J H, et al. BERT-MK: integrating graph contextualized knowledge into pre-trained language models [C]//Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, 2020: 2281–2290. DOI: 10.18653/v1/2020.findings-emnlp.207
- [30] WEI J, BOSMA M, ZHAO V Y, et al. Finetuned language models are zero-shot learners [EB/OL]. [2022–02–15]. <https://arxiv.org/abs/2109.01652>
- [31] ARIBANDI V, TAY Y, SCHUSTER T, et al. Ext5: Towards extreme multi-task scaling for transfer learning [EB/OL]. [2022–02–15]. <https://arxiv.org/abs/2111.10952>
- [32] SANH V, WEBSON W, RAFFEL C, et al. Multitask prompted training enables zero-shot task generalization. [EB/OL]. [2022–02–15]. <https://arxiv.org/abs/2110.08207>
- [33] GU Y, HAN X, LIU Z, et al. PPT: pre-trained prompt tuning for few-shot learning [EB/OL]. [2022–02–15]. <https://arxiv.org/abs/2109.04332>
- [34] SUN Y, WANG S H, LI Y K, et al. ERNIE 2.0: a continual pre-training framework for language understanding [EB/OL]. [2022–02–15]. <https://arxiv.org/abs/1907.12412>
- [35] WANG R Z, TANG D Y, DUAN N, et al. K-adaptor: infusing knowledge into pre-trained models with adapters [EB/OL]. [2022–02–15]. <https://arxiv.org/abs/2002.01808>
- [36] WANG S H, SUN Y, XIANG Y, et al. ERNIE 3.0 titan: exploring larger-scale knowledge enhanced pre-training for language understanding and generation [EB/OL]. [2022–02–15]. <https://arxiv.org/abs/2112.12731>
- [37] DAI Z H, YANG Z L, YANG Y M, et al. Transformer-XL: attentive language models beyond a fixed-length context [EB/OL]. [2022–02–15]. <https://arxiv.org/abs/1901.02860>
- [38] DING S Y, SHANG J Y, WANG S H, et al. ERNIE-doc: a retrospective long-document modeling transformer [EB/OL]. [2022–02–15]. <https://arxiv.org/abs/2012.15688>
- [39] LIU Y, WAN Y, HE L F, et al. KG-BART: knowledge graph-augmented BART for generative commonsense reasoning [EB/OL]. [2022–02–15]. <https://arxiv.org/abs/2009.12677>
- [40] PETERS M E, NEUMANN M, LOGAN R, et al. Knowledge enhanced contextual word representations [C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, 2019: 43–54. DOI: 10.18653/v1/d19-1005

- [41] XUE L T, CONSTANT N, ROBERTS A, et al. MT5: a massively multilingual pre-trained text-to-text transformer [EB/OL]. [2022-02-15]. <https://arxiv.org/abs/2010.11934>
- [42] ZHANG Z Y, GU Y X, HAN X, et al. CPM-2: large-scale cost-effective pre-trained language models [EB/OL]. [2022-02-15]. <https://arxiv.org/abs/2106.10715>
- [43] OUYANG X, WANG S H, PANG C, et al. ERNIE-M: enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora [EB/OL]. [2022-02-15]. <https://arxiv.org/abs/2012.15674>
- [44] LU J S, BATRA D, PARIKH D, et al. ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks [EB/OL]. [2022-02-15]. <https://arxiv.org/abs/1908.02265>
- [45] YU F, TANG J J, YIN W C, et al. ERNIE-ViL: knowledge enhanced vision-language representations through scene graph [EB/OL]. [2022-02-15]. <https://arxiv.org/abs/2006.16934>
- [46] LIU Y D, LU W X, CHENG S Q, et al. Pre-trained language model for web-scale retrieval in Baidu search [C]//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. ACM, 2021: 3365-3375. DOI: 10.1145/3447548.3467149
- [47] ZOU L X, ZHANG S Q, CAI H Y, et al. Pre-trained language model based ranking in Baidu search [C]//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. ACM, 2021: 4014-4022. DOI: 10.1145/3447548.3467147
- [48] BAO S Q, HE H, WANG F, et al. PLATO: pre-trained dialogue generation model with discrete latent variable [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020: 85-96. DOI: 10.18653/v1/2020.acl-main.9
- [49] BAO S, HE H, WANG F, et al. PLATO-XL: exploring the large-scale pre-training of dialogue generation [EB/OL]. [2022-02-15]. <https://arxiv.org/abs/2109.09519>
- [50] LIU Z B, NIU Z Y, WU H, et al. Knowledge aware conversation generation with explainable reasoning over augmented graphs [EB/OL]. [2022-02-15]. <https://arxiv.org/abs/1903.10245>
- [51] YUAN Q, CHEN J, LU C, et al. The graph-based mutual attentive network for automatic diagnosis [C]//Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization, 2020: 3393-3399. DOI: 10.24963/ijcai.2020/469
- [52] CHEN J, YUAN Q, LU C, et al. A novel sequence-to-subgraph framework for diagnosis classification [EB/OL]. [2022-02-15]. <https://www.ijcai.org/proceedings/2021/0496.pdf>

作者简介



王海峰，百度首席技术官、深度学习技术及应用国家工程实验室主任、教授级高工、国际计算机语言学学会（ACL）50 余年来首位华人主席、ACL 亚太分会创始主席、ACL Fellow、IEEE Fellow、CAAI Fellow、国际欧亚科学院院士；长期从事自然语言处理技术研究及产业化工作；突破知识增强深度语义理解技术，主持研制了产业级深度学习开源开放平台飞桨，实现了智能搜索、机器翻译等大规模产业应用；以第一完成人获国家技术发明奖二等奖 1 项、国家科技进步奖二等奖 1 项、中国专利金奖 1 项等，获光华工程科技奖、全国创新争先奖、首个吴文俊人工智能杰出贡献奖等；发表论文近 200 篇，获授权专利 150 余项。



孙宇，百度杰出研发架构师；主要研究领域包括语义理解、语义表示与计算、大规模预训练模型等；曾获国家技术发明奖二等奖 1 项、中国电子学会科技进步奖一等奖 1 项、世界人工智能大会 SAIL 奖等；发表顶级学术会议论文 10 余篇，获授权专利 30 余项。



吴华，百度技术委员会主席、教授级高工；主要从事自然语言处理技术研究及产业化工作；曾获国家技术发明奖二等奖 1 项、国家科技进步奖二等奖 1 项、中国专利金奖 1 项，荣获杰出工程师、青年北京学者等奖项；发表论文 100 余篇，获授权专利 100 余项。

悟道·文澜: 超大规模多模态预训练模型带来了什么?



WuDao-WenLan: What Do Very-Large Multimodal Pre-Training Models Bring?

卢志武/LU Zhiwu¹, 金琴/JIN Qin², 宋睿华/SONG Ruihua¹, 文继荣/WEN Jirong^{1,2}

(1. 中国人民大学高瓴人工智能学院, 中国 北京 100872;

2. 中国人民大学信息学院, 中国 北京 100872)

(1. Gaoling School of Artificial Intelligence, Renmin University of China, Beijing 100872, China;

2. School of Information Renmin University of China, Beijing 100872, China)

DOI: 10.12142/ZTETJ.202202005

网络出版地址: <https://kns.cnki.net/kcms/detail/34.1228.tn.20220411.1207.002.html>

网络出版日期: 2022-04-12

收稿日期: 2022-02-20

摘要: 提出了悟道·文澜的BriVL双塔模型。该模型利用6.5亿对互联网图文数据, 通过自监督的任务来训练, 是目前最大的中文通用图文预训练模型。同时, 还提出了悟道·文澜的多语言多模态预训练单塔模型—MLMM。实验结果证明, 这两个模型在多个国际公开数据集上均取得了最佳性能。设计了实验并讨论超大规模多模态预训练模型对文本编码、图像生成和图文互检带来的影响, 以及文澜模型的落地应用与学科交叉成果。

关键词: 多模态预训练; 多语言预训练; 双塔模型; 单塔模型

Abstract: A multimodal pre-training two-tower model called WuDao-WenLan BriVL is proposed, which is trained through self-supervised learning over 650 million image-text pairs crawled from the Web. This is the largest open-sourced Chinese image-text pre-training model. Moreover, a multi-lingual pre-training single-tower model called WuDao-WenLan MLMM is also proposed. Extensive experiments show that these two models achieve the new state-of-the-art performance on multiple public benchmark datasets. In addition, experiments are conducted to discuss what very-large multimodal pre-training models bring to text encoding, text-to-image generation, and image-text retrieval, as well as in what applications WenLan can be applied in multiple fields.

Keywords: multimodal pre-training; multi-lingual pre-training; two-tower model; single-tower model

人脑是一个复杂的系统, 能够处理多种感官模态例如视觉、听觉、嗅觉等的信息。这使得人们能够准确、有效地完成感知、理解和决策任务。为了模仿人类的这些核心认知能力, 人工智能模型利用大规模多模态数据来进行训练。如何利用从互联网上爬取的大规模多模态数据进行模型训练, 成为近期业界的研究热点。如何能有效地利用这些爬取数据是一个巨大的挑战, 因为我们无法对其进行详细的人工标注。另外, 这些数据不可避免地存在一定量的数据噪声。如图1所示, 学术界数据集多为由人工编写的强相关文本, 如“水果蛋糕上有一些蜡烛在燃烧”, 规模多为几万到百万图文对。与此不同的是, 从互联网上搜集到的图像的周边文本通常与内容弱相关。

多模态预训练的目标是对齐不同模式的大规模数据, 从而可以将所学知识迁移到各种下游任务中, 并最终接近通用人工智能。目前, 多模态预训练模型已经在广泛的多模态任

务中取得了巨大成功。然而, 学术界往往只重视在有限规模的标注数据集上取得更好的效果, 因此多采用单塔模型, 并在英文数据集上进行训练。这使得其应用场景被规模、性能



▲图1 两种不同的图文数据

和语言所局限。在北京智源研究院悟道项目的支持下，文继荣教授带领中国人民大学卢志武教授、宋睿华长聘副教授、金琴教授等师生团队搜集了6.5亿对中文图文数据，率先提出图文弱相关是更为现实的假设，并利用跨模态对比学习来自监督地训练超大规模图像-文本多模态预训练模型文澜 BriVL。另外，我们认为：不同模态和不同语言都有可能表示相同的语义信息。如图2所示，中文单词“狗”、英文单词“dog”或是一张狗的视觉图像，都能表示狗这一动物。因此，我们研究了如何通过预训练来捕捉视觉与语言在语义上的共通点，提供更好的视觉和语言特征，以支持不同的多语言多模态下游任务；同时提出文澜多语言多模态预训练模型 MLMM。实验证明，两个模型均能在多项下游任务中获得国际最佳性能。

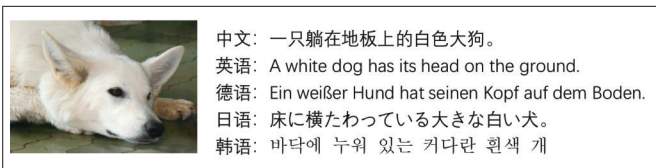
此外，我们还着重讨论了超大规模多模态预训练带来的影响，包括对文本编码、图像生成和图文互检的影响。总之，多模态预训练带来的改变才刚刚开始，它在人工智能方面有着巨大的潜力。

1 文澜 BriVL 超大规模图文预训练模型

1.1 相关工作

自2018年以来，单模态预训练模型（如BERT^[1]、GPT^[2]、ViT^[3]等）的出现，极大地促进了相关领域的发展。人们也在持续探索具有更强通用性的多模态预训练模型，具有代表性的工作有UNITER^[4]、OSCAR^[5]等。然而，由于视觉数据集的标注需要的成本高昂，多模态数据集往往维持在百万的数据量级，因此，难以在此基础上训练出具备良好通用性与泛化性的多模态模型。多模态预训练模型根据其框架可分为两类：单塔和双塔。

最近的UNITER^[4]、Oscar^[5]、M6^[6]、VisualBERT^[7]、Unicoder-VL^[8]、VL-BERT^[9]等模型都采用单塔网络，它们利用一个特征融合模块（例如Transformer）来得到图像-文本对的嵌入。其中，一些单塔模型还使用对象检测器来检测图像区域，并将这些区域与相应的单词进行匹配。UNITER作为单塔模型的代表，对560万图文对进行遮挡语言建模（MLM）、遮挡区域建模（MRM）和图像文本匹配（ITM）的联合训练，从而学到通用的图像文本表示。Oscar将语义相同的对象（名词）作为图像和文本对齐的基础，从而简化图像和文本语义对齐的学习任务，即使用快速目标检测器（Fast R-CNN）就可以将检测到的对象标签与文本中的单词建立关联。现有单塔结构通常依赖于强相关的图文对数据，而这一强相关假设对大规模网络数据集来说通常是无效的。



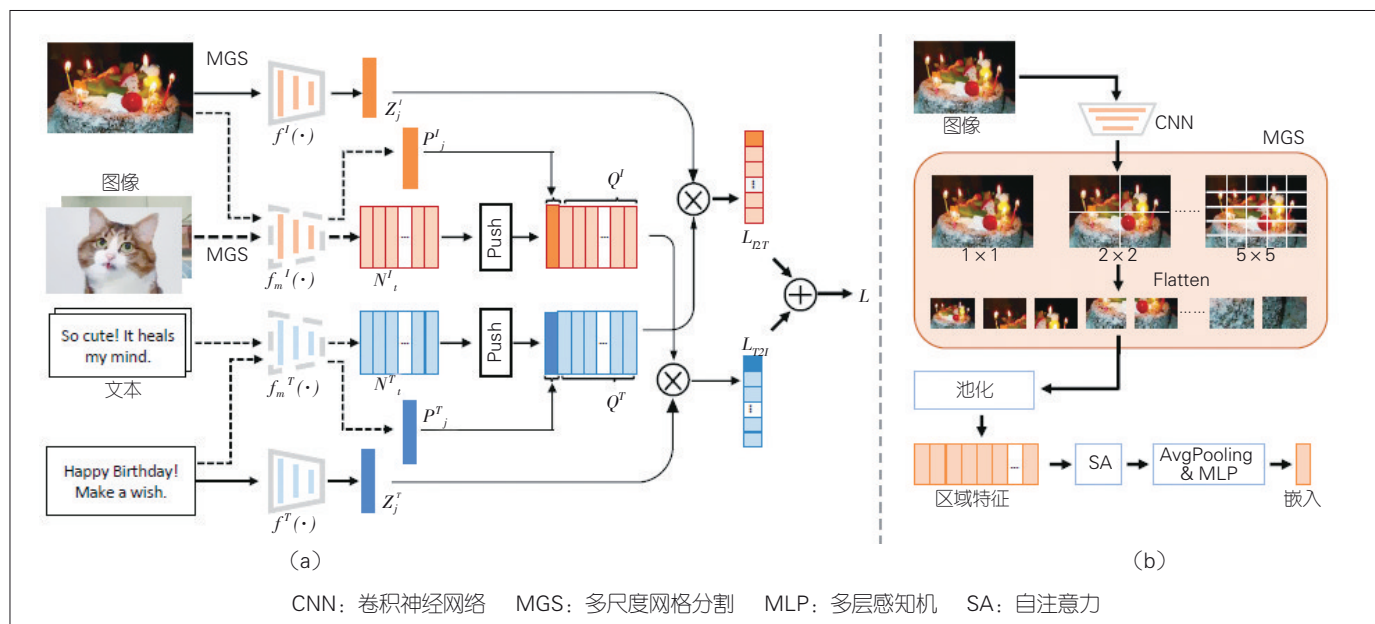
▲图2 不同语言和模态能够表达相同的语义

此外，单塔模型在推理阶段需要较高的计算成本。例如，需要将查询内容（图像或文本）输入到单塔模型中，计算它和所有候选对象的匹配分数。

相比之下，采用双塔结构的多模态预训练模型使用单独的图像和文本编码器，分别对图像和文本进行编码，然后进行图文对匹配来完成检索任务。这种模式的检索效率更高，但由于缺乏更深层次的图像-文本交互（即图像区域与单词的交互），通常只能达到次优性能。最近的双塔工作，如LightningDot^[10]，通过重新设计目标检测过程来应对这一挑战；CLIP^[11]、ALIGN^[12]、WenLan 1.0^[13]和WenLan 2.0^[14]则放弃了昂贵的对象检测器，利用跨模态对比学习任务来进行模型训练。

1.2 模型介绍

文澜 BriVL 模型在预训练数据的选择上，不再遵循强相关语义假设，而是转向弱相关假设；在网络架构上，选择双塔结构而不是单塔结构；使用了更加节约计算资源的跨模态对比学习算法来进行预训练。具体来说：（1）在弱相关语义假设下，图文数据不再需要任何人工标注，互联网上的海量多模态数据成为文澜 BriVL 模型的预训练数据来源。相比于人工标注的几百上千万强语义相关图文数据，文澜 BriVL 模型使用的预训练数据全部爬取自互联网，规模达到了6.5亿对。更重要的是，弱语义相关数据包含了复杂、抽象的人类情感和想法，能够帮助我们把文澜 BriVL 模型训练成一个更具认知能力的模型。（2）文澜 BriVL 模型不再需要耗时的目标检测器，使用的双塔网络架构在应用时也有明显的效率优势。双塔包含两个独立的编码器：一个用于图片，另一个用于文本。因此，在跨模态检索时，候选的图片或者文本可以提前计算出嵌入表示并做好索引，以满足现实应用的效率需求。（3）受到单模态对比学习算法 MoCo 的启发，文澜 BriVL 模型在使用跨模态对比学习的同时也引入 Momentum 机制以及动态维护负样本队列（如图3所示）。这样就解构了batch大小与负样本数量，从而在相对较小的batch下（即较少的图形处理器资源）就可以得到性能较好的预训练模型。



▲图3 文澜 BriVL 的网络架构图与图像编码器

1.3 实验分析

我们在图像零样本分类、文本零样本分类两个下游任务上进行实验，以验证文澜 BriVL 模型的迁移能力。

(1) 下游任务 1: ImageNet 的零样本分类

我们利用文澜 BriVL 的图文编码器，可以直接在 ImageNet 数据集的 200 类图像子集上进行零样本分类。这需要提前将这 200 个类名翻译成中文。ImageNet 200 类挑选的原则为：英文类名在翻译成中文时无明显错误。OpenAI CLIP 则直接在英文数据集上进行测试。从表 1 可以发现，文澜 BriVL 2.0 的零样本图片分类准确率要高于 CLIP。这说明我们的模型具有更好的泛化能力。

(2) 下游任务 2: 中文学科的零样本分类

我们利用文澜 BriVL 1.0 以及 2.0 的文本编码器，在中文学科分类数据集 (CSLDCP) 上进行小样本分类。我们采用被广泛使用的 prompt-tuning 方法来为 1-shot 分类。针对文澜 BriVL 模型，我们同时利用了视觉和文本两个模态的信息来进行 prompt-tuning。对比实验考虑了单模态预训练的 RoBERTa-base 和 RoBERTa-large。从表 2 可以发现，相比于单模态预训练模型 RoBERTa，文澜 BriVL 模型具有更好的中文小样本分类能力。这说明多模态预训练在纯粹的 NLP 下游任务中也发挥了重要的作用。

1.4 模型可视化

文澜 BriVL 模型的可视化流程为：

- (1) 给定一个文本，输入一张随机噪声图像；

▼表 1 ImageNet 200 类的零样本分类结果

模型	ImageNet 200 类	
	Top-1 准确率/%	Top-5 准确率/%
OpenAI CLIP	82.5	95.2
文澜 BriVL 1.0 (RoBERTa-base)	69.6	88.9
文澜 BriVL 2.0 (RoBERTa-large)	85.0	96.0

▼表 2 中文学科的 1-shot 小样本分类结果

模型	CSLDCP
单模态 RoBERTa-base	33.63
单模态 RoBERTa-base (在文澜 1.0 的文本数据上微调)	33.40
文澜 BriVL 1.0 (RoBERTa-base)	35.59
单模态 RoBERTa-large	38.24
文澜 BriVL 2.0 (RoBERTa-large)	46.47

CSLDCP: 中文学科分类数据集

- (2) 通过模型的文本编码器得到文本的特征表示；
- (3) 多模态神经元可视化的目标函数为：让当前输入图像的视觉特征表示逼近文本特征；
- (4) 固定文澜的所有参数，通过反向传播来更新输入的噪声图像。

总之，算法收敛后，得到的图像是文澜 BriVL 认为的对输入文本最为接近的可视化处理结果。如图 4 所示，大规模多模态预训练后的神经网络已经能够理解古诗句的意境，展示了强大的中文理解能力。

2 文澜 MLMM 多语言多模态预训练模型

2.1 相关工作

目前,在多语言多模态的语义学习方面,已有一些工作陆续开展。M3P^[15]首次采用了预训练来学习多语言多模态知识,以多任务学习的方式轮流将英文的图像描述数据和单模态的多语言语料输入到模型中,以进行预训练;UC2^[16]使用机器翻译对现有的图像描述数据集进行多语言扩充,同时遮蔽两种语言相同意义的词来迫使模型根据图像内容进行还原。文献[17]采用英文图像描述数据和平行语料进行预训练,将Unicoder^[18]扩展到多语言多模态上。

这些工作虽然取得了一定的成果,但其预训练规模仍局限于Conceptual Caption 3M数据集。较小规模的预训练使得模型的零样本跨语言迁移能力较弱。因此,我们致力于利用更大规模、更加开放领域的数据进行预训练,以获得更加通用、更加强大的多语言多模态预训练模型。

2.2 模型介绍

我们设计的MLMM模型的整体结构如图5所示。我们首先使用在Visual Genome数据集上预训练的Faster R-CNN目标检测器来提取图像中的区域特征,并将这些特征与相应的多语言文本Token一同输入到Transformer Encoder中。

为了捕获不同层次的视觉与语言特征,MLMM采用4个任务进行预训练:

(1) ITM。为了建模图像与多语言文本的全局语义信息,我们使用ITM任务对MLMM模型进行预训练。该任务的目标是,判断输入的图像和多语言文本是否是语义匹配的。在ITM任务中,模型需要理解输入图像和多语言文本的全局语义信息,进而做出判断。

(2) MLM。我们采用MLM任务来建模多语言文本的细粒度语义信息。MLM的目标是根据图像区域信息和文本上下文,让模型来预测被遮蔽的多语言文本单词。

(3) 图像区域回归(MRFR)。为了增强模型对图像的细致建模能力,MRFR任务要求模型根据文本和其他图像区域还原被遮蔽的图像区域特征。

(4) 图像区域分类(MRC)。为了让模型能够细粒度地识别图像语义,我们实施了MRC任务,因此让模型来预测被遮蔽图像区域所属类别。虽然数据集中没有区域语义的标注信息,但是目标检测器检测得到的类别可以作为该任务的伪标注。目标检测器预测的类别并不是完美的,我们将目标检测器在目标类别上的分布作为软标签,通过计算MLMM预测分布与目标检测器软标签的KL divergence,来优化整个



图4 文澜 BrVL 对诗句的神经元可视化

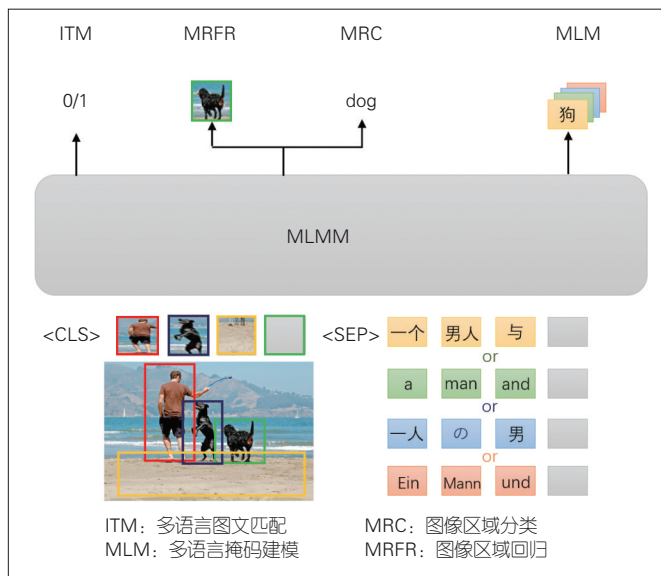


图5 MLMM 模型结构图

模型。

我们使用的多语言多模态预训练数据集涵盖汉语、英语、德语、法语、捷克语、日语、韩语7种语言和与语义相匹配的图像,包含2.1亿对多语言图文数据。该数据集在以下两个数据集的基础上通过机器翻译进行构建:

(1) 英文图文数据集 Conceptual Caption 3M+12M。该数据集是目前图文预训练的通用数据集,约有1 500万图文对。数据集中的文本具体描述了图像中所包含的内容。针对该数据集,我们采用4种预训练任务进行训练。

(2) 中文图文数据集 RUC-CAS-WenLan。该数据集是我们构建的,涵盖新闻、百科、微博、微信等领域,文本内容与对应的图像呈弱相关关系。我们选取其中的1 500万图文对进行预训练。针对该数据集的特点,我们仅训练ITM任务。

2.3 实验分析

我们在多语言图文检索、多语言视觉问答两个下游任务中进行了实验,以验证MLMM的多语言多模态能力。

(1) 下游任务1:多语言图文检索

多语言图文检索任务为:给定一段多语言文本,模型可以从数据库中找到与之语义最相关的一张图像,或通过一张图片找到与之最相关的多语言文本。对于多语言图文检索,我们在两个常用的多语言图文数据集 Multi30K 和 MSCOCO 上进行评测。Multi30K 是英文图文数据集 Flickr30K 的扩展,支持英语、德语、法语和捷克语 4 种语言;文献[19–20]分别将最初的英文 MSCOCO 数据集扩展到中文和日文。通常,多语言图文检索评测包含以下几个设定:

- Finetune on en。只使用英文下游数据对模型进行微调,然后测试模型在其他语言上的表现,以衡量模型在多语言上的扩展性。
- Finetune on each。使用多种语言的下游数据,分别对预训练模型进行微调,以衡量模型的单语言能力
- Finetune on all。同时使用多种语言的下游数据对一个预训练模型进行微调,以衡量模型的多语言容量。

与 M3P 和 UC2 相同,我们采用平均召回率,即图像检索文本、文本检索图像两个检索方向上的 Recall@1、5、10 的平均值,来衡量模型的检索效果。3 种微调设定下的实验结果如表 3 所示。

从表 3 中可以看出,在 3 种设定上,MLMM 都超过了现有最好的多语言预训练模型 M3P 和 UC2,达到当前最佳性能。尤其在英文上进行微调时,英文与其他语言之间的性能差距明显小于现有的工作中两者间的性能差距。这说明得益于更大规模的预训练,MLMM 能够表现出很强的跨语言迁移能力。

(2) 下游任务 2: 多语言视觉问答

给定一张图像和一个与图像内容相关的特定语言上的提问,多语言视觉问答任务要求模型能够给出正确的答案。我们采用 VQA 2.0 和 VQA VG Japanese 两个数据集进行多语言

视觉问答的实验。其中,VQA 2.0 是英文视觉问答数据集,而 VQA VG JA 则是日文视觉问答数据集。与 UC2 相同,MLMM 将视觉问答任务视为多标签分类任务,即模型从一个固定的候选池中选择问题的答案。对于 VQA 2.0 数据集,我们选择最常见的 3 129 个回答作为答案候选池;对于 VQA VG Japanese,我们选择最常见的 3 000 个回答作为答案候选池。表 4 展示了 MLMM 在多语言视觉回答上的实验结果。

从表 4 中可以看出,MLMM 在多语言图文检索上超越了目前的预训练模型,在两个多语言视觉问答数据集上同样表现出色。这验证了通过大规模的预训练,MLMM 能够轻松适配各种多语言多模态的下游任务。

2.4 可视化分析

我们对 MLMM 学习到的跨语言跨模态的通用知识进行了可视化。我们将语义相匹配的多语言文本和图像输入到 MLMM 中,将最后一层 Transformer Encoder 的文本对图像区域的注意力权重进行可视化,如图 6 所示。对于中文和英文相同语义的单词,其注意力权重在图像区域上的分布基本一致。这说明通过大规模的预训练,MLMM 学习到了多语言单词之间以及和图像区域之间的语义对应关系。

3 超大规模多模态预训练模型带来的影响

3.1 多模态信息对文本编码的影响

当图像信息通过文澜预训练模型影响文本编码时,到底发生了怎样的改变?给定一个词,我们将该词在两个空间中的 K 近邻的词集合分别表示为 N_w^1 和 N_w^2 ,然后用笛卡尔相似度来计算该词在两个空间表示的相似性:

▼表 3 多语言图文检索平均召回率

数据集	评测语言 模型	Multi30K				MSCOCO			
		En	De	Fr	Cs	En	Zh	Ja	
En	M3P	87.4	58.8	46.0	36.8	88.6	53.8	56.0	
	UC2	87.2	74.9	74.0	67.9	88.1	82.0	71.7	
	MLMM	91.9	86.7	86.9	85.6	90.6	90.3	86.6	
Each	M3P	87.4	82.1	67.3	65.0	88.6	75.8	80.1	
	UC2	87.2	83.8	77.6	74.2	88.1	84.9	87.3	
	MLMM	91.9	88.1	85.3	83.8	90.6	89.0	90.9	
All	M3P	87.7	82.7	73.9	72.2	88.7	87.9	86.2	
	UC2	88.2	84.5	83.9	81.2	88.1	89.8	87.5	
	MLMM	92.0	88.7	88.2	87.4	90.8	92.4	91.2	

▼表 4 多语言视觉问答准确率

数据集模型	VQA 2.0 test-dev	VQA VG Japanese
UNITER	71.22	22.70
UC2	71.48	34.20
MLMM	73.21	35.40



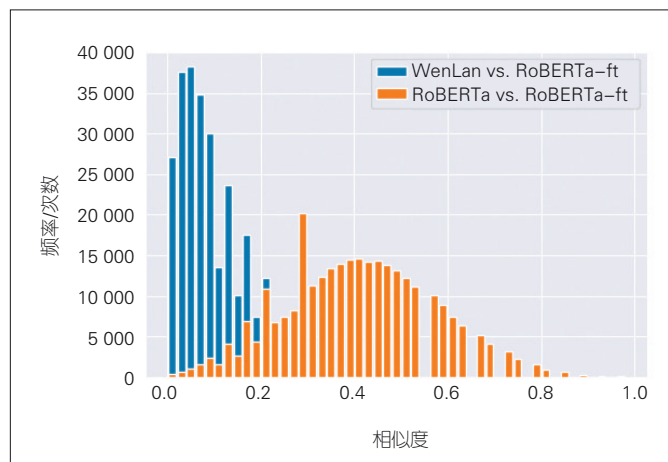
▲图 6 MLMM 模型在多语言图文检索中的注意力权重可视化

$$\text{Jaccard Similarity} = \frac{|N_w^1 \cap N_w^2|}{|N_w^1 \cup N_w^2|}$$

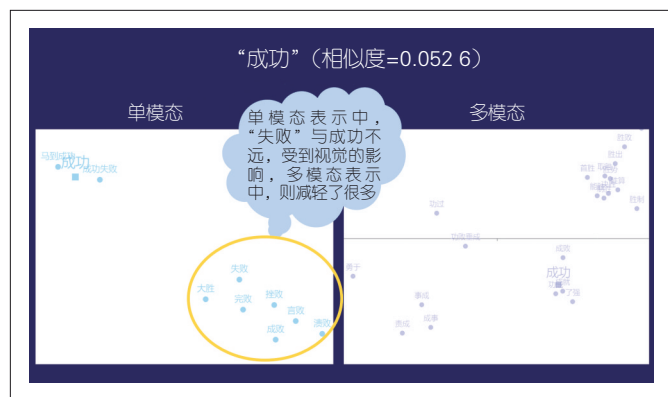
公平起见,哈尔滨工业大学的车万翔老师团队使用文澜的图文训练集中的所有文字,对RoBERTa进行了微调。在17万的词表上进行统计的结果如图7所示。和微调后的RoBERTa相比,RoBERTa看上去是一个相似度均值在0.4附近的正态分布;但和微调后的RoBERTa相比,WenLan的相似度明显变低,大部分样本集中在0.1以下。这说明图像对文本词向量有着显著的影响。

我们在查看了相似度较低的词语后发现了一些共同点:

(1) 如图8所示,在单模态语言模型中,由于上下文类似,反义词的词嵌入向量经常会非常相似。例如,在图8的左部分中,当RoBERTa微调后,离“成功”不远的地方有一组与“失败”相关的词语;经过文澜多模态预训练,“成功”周围则以“成功”为主了(如图8右部分所示)。这可能是因为与“成功”和“失败”相关联的图像在色调和内容上相差较大。



▲图7 同样的词在两个空间中的词向量相似性分布



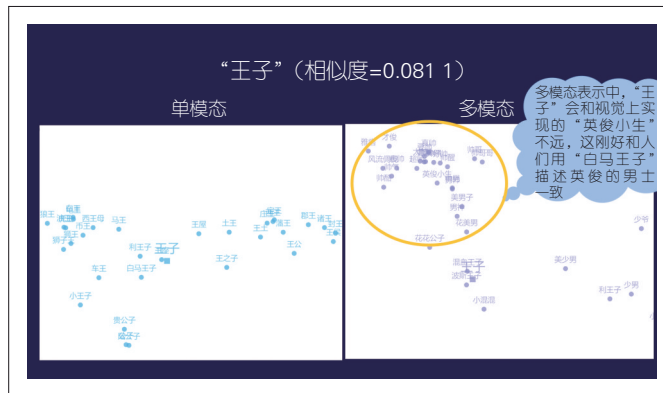
▲图8 “成功”在单模态RoBERTa微调模型与多模态文澜模型中所对应的空间上的邻近词语

(2) 视觉上相似的词语会被拉近距离。以图9为例,RoBERTa微调模型会把“王子”与“王公”“狮子王”“贵公子”等语义上比较相近的词语拉近。多模态预训练模型会将“王子”和“美男子”“帅哥”“英俊小生”等词语拉近。这些概念在人们的印象中确实有很强的视觉语义相关性。

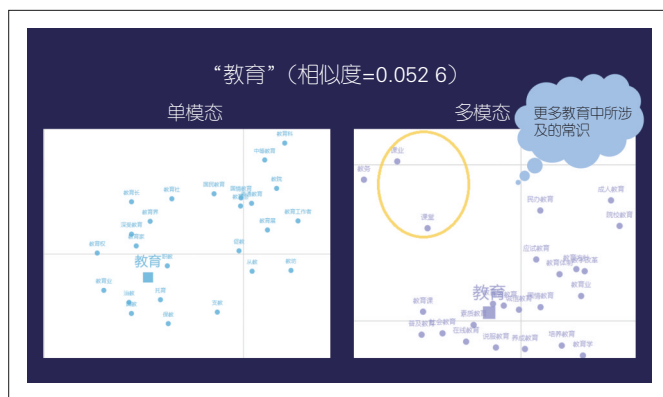
(3) 同一情境的词语被拉近。如图10所示,RoBERTa微调模型通常会找到和“教育”同层次的近义词语,如“保育”“国民教育”“教育界”等;文澜模型则会找到一些“课业”“课堂”等词语,这些词语可能出现在类似的图片周围,并通过跨模态之间的对比学习拉近距离。

3.2 多模态预训练对图像生成的影响

基于单模态预训练生成模型的主要问题是,输入句子嵌入是由在单一模态中预先训练的文本编码器提取的,这在语义上与图像模态不一致。因此,单模态预训练生成模型需要学习、处理视觉和自然语言的不同统计特性,以便生成与给定文本对齐的真实图像。为此,现有方法采用了对比学习,并仔细设计了基于注意的单词和区域自我调节,以便更好地进行训练,这种方式是相当耗时的。在跨模态生成中(如文



▲图9 “王子”在单模态RoBERTa微调模型与多模态文澜模型中所对应的空间上的邻近词语



▲图10 “教育”在单模态RoBERTa微调模型与多模态文澜模型中所对应的空间上的邻近词语

本生成图像), 高效地弥补这两种模态之间的差距非常具有挑战性。

与以往方法不同, 我们可以利用多模态预训练模型对图像和文本进行编码。例如, 借助 VQGAN inversion, 可以实现基于文澜 BriVL 的文生成图。具体地, 给定一个文本, 输入一张随机噪声图像, 通过文澜 BriVL 的文本编码器就可以得到文本的特征表示。VQGAN inversion 的目标函数为: 当前输入图像经过 VQGAN 后输出的图像, 其视觉特征(通过文澜图像编码器得到)必须逼近输入文本的特征。固定 VQGAN 和文澜模型的所有参数, 通过反向传播可以更新输入的噪声图像。算法收敛后, 最终得到的图像即可看作关于给定文本的文生成图结果。如图 11 所示, 借助 VQGAN, 文澜 BriVL 模型能够生成更贴近自然的图像。

这里的关键之处在于, 由多模态预训练模型提取的文本嵌入可以自然地与图像模态对齐, 这避免了之前方法中的额外复杂架构。总之, 多模态预训练模型给文生成图任务带来了新的研究思路。

3.3 多模态预训练对文本-图像检索的影响

当文澜模型将图像和文本映射到同一空间时, 文本与图像的互检就变得非常容易。当文本检索图像时, 不再需要图像周围的文字作为桥梁, 因此文澜模型可以匹配图像周围文字并没有描述的意境。图像检索文本也成为可能, 不仅能识别出物体、场景或情感等类别标签, 还可以和任意的句子、段落进行多模态共享语义空间上的匹配。这首次跨越了图文的语义鸿沟, 实现了真正的跨模态检索。

基于文澜 BriVL 模型, 文澜团队实现了多个在线演示系统, 具体见图 12。

4 结束语

我们尝试了利用亿级的、来自互联网的图文对数据来训练多模态双塔模型 BriVL 和多语言多模态单塔模型 MLMM。这两个预训练模型均在多个下游任务中获得了国际最佳性能。通过实验, 我们发现多模态预训练模型将更多视觉相似或在同一场景中的词语拉近; 能为文生成图提供统一的语义基础, 提升图像生成的泛化能力和效果; 能让文字和图像可以在映射到同一空间后实现真正的跨模态检索。目前, 文澜 BriVL 1.0 已开源, 可以通过以下网址访问或者申请下载:

- 文澜 BriVL 1.0 源码下载: <https://github.com/BAAI-WuDao/BriVL>

- 文澜 BriVL 1.0 模型申请: <https://wudaoai.cn/model/detail/BriVL>



▲图 11 借助 VQGAN inversion 得到的文澜文生成图结果



▲图 12 基于文澜模型开发的 3 款跨模态检索小应用

- 文澜 BriVL 1.0 在线 API: <https://github.com/chuhaojin/WenLan-api-document>

自 2021 年 3 月发布以来, 文澜受到了腾讯、酷我音乐、爱奇艺、网易等多家企业的关注。与长城汽车合作, 文澜完成了由图像检索金句的“欧拉喵语”小应用, 并在上海和成都车展以及 ChinaJoy 上与参观者进行现场的品牌互动; 与 OPPO 合作, 文澜模型实现了为视障人士读取收集图片的功能, 践行科技向善的理念。

文澜模型的强大能力也产生了一些跨学科研究成果。由中国人民大学新闻学院和高瓴人工智能学院合作的《空间漫游与想象生产——线上影像策展中的网红城市建构: 基于视觉·语言多模态预训练模型的计算传播研究》, 获得了 2021 年计算传播学会学生论文三等奖。中国人民大学艺术学院师生与上海大学教师组成的“云端艺术”团队, 将文澜融合到他们的微信程序“红色夏天智能航宇”作品中, 获得 2021 年上海图书馆开放数据竞赛优秀设计奖。

最后, 如何平衡单双塔的有效性和效率是未来的重要问题, 目前主要方法有两种: (1) 对于单塔模型, 可以在跨模态融合模块之前放置双塔体系结构, 以减少巨大的检索延迟, 同时尽可能保持高性能优势; (2) 对于双塔模式, 可以

考虑建立更精细/更紧密的模式相关性的学习目标,以提高其性能,同时保持高效率的优势。

参考文献

- [1] NI M, HUANG H, SU L, et al. Learning universal representations via multitask multilingual multimodal pre-training [EB/OL]. [2022-01-12]. <https://paperswithcode.com/paper/m3p-learning-universal-representations-via>
- [2] ZHOU M, ZHOU L, WANG S, et al. UC2: universal cross-lingual cross-modal vision-and-language pre-training [EB/OL]. [2022-01-10]. <https://paperswithcode.com/paper/uc2-universal-cross-lingual-cross-modal>
- [3] FEI H, YU T, LI P. Cross-lingual cross-modal pretraining for multimodal retrieval [EB/OL]. [2022-01-10]. <https://paperswithcode.com/paper/cross-lingual-cross-modal-pretraining-for>
- [4] HUANG H, LIANG Y, DUAN N, et al. Unicoder: a universal language encoder by pre-training with multiple cross-lingual tasks [EB/OL]. [2022-01-10]. <https://paperswithcode.com/paper/unicoder-a-universal-language-encoder-by-pre>
- [5] LI X, XU C, WANG X, et al. COCO-CN for cross-lingual image tagging, captioning, and retrieval [EB/OL]. [2022-01-12]. <https://paperswithcode.com/paper/coco-cn-for-cross-lingual-image-tagging>
- [6] YOSHIKAWA Y, SHIGETO Y, TAKEUCHI A. Stair captions: Constructing a large-scale Japanese image caption dataset [EB/OL]. [2022-01-10]. <https://paperswithcode.com/paper/stair-captions-constructing-a-large-scale>
- [7] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [EB/OL]. [2022-01-10]. <https://paperswithcode.com/paper/bert-pre-training-of-deep-bidirectional>
- [8] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners [EB/OL]. [2022-01-10]. <https://paperswithcode.com/paper/language-models-are-unsupervised-multitask>
- [9] DOSOVITSKIY A, BEYER L, KOLESNIKOV W, et al. An image is worth 16x16 words: transformers for image recognition at scale [EB/OL]. [2022-01-10]. <https://paperswithcode.com/paper/an-image-is-worth-16x16-words-transformers-1>
- [10] CHEN Y, LI L, YU L, et al. Uniter: universal image-text representation learning [EB/OL]. [2022-01-10]. <https://paperswithcode.com/paper/uniter-learning-universal-image-text-1>
- [11] LI X, YIN X, LI C, et al. Oscar: object-semantic aligned pre-training for vision-language tasks [EB/OL]. [2022-01-12]. <https://paperswithcode.com/paper/oscar-object-semantics-aligned-pre-training>
- [12] LIN J, MEN R, YANG A, LIN J, et al. M6: A Chinese multimodal pretrainer [EB/OL]. [2022-01-12]. <https://paperswithcode.com/paper/oscar-object-semantics-aligned-pre-training>
- [13] LI L, YATSKAR M, YIN D, et al. Visualbert: a simple and performant baseline for vision and language [EB/OL]. [2022-01-10]. <https://paperswithcode.com/paper/oscar-object-semantics-aligned-pre-training>
- [14] LI G, DUAN N, FANG Y, et al. Unicoder-vl: a universal encoder for vision and language by cross-modal pre-training [EB/OL]. [2022-01-10]. <https://paperswithcode.com/paper/unicoder-vl-a-universal-encoder-for-vision>
- [15] SU W, ZHU X, GAO Y, et al. Vi-bert: pre-training of generic visual-linguistic representations [EB/OL]. [2022-01-10]. <https://paperswithcode.com/paper/unicoder-vl-a-universal-encoder-for-vision>
- [16] SUN S, CHEN Y, LI L, et al. Lightningdot: pre-training visual-semantic embeddings for real-time image-text retrieval [EB/OL]. [2022-01-10]. <https://paperswithcode.com/paper/lightningdot-pre-training-visual-semantic>
- [17] RADFORD A, KIM J, HALLACY C, et al. Learning transferable visual models from natural language supervision [EB/OL]. [2022-01-10]. <https://paperswithcode.com/paper/lightningdot-pre-training-visual-semantic>
- [18] JIA C, YANG Y, XIA Y. Scaling up visual and vision-language representation learning with noisy text supervision [EB/OL]. [2022-01-10]. <https://paperswithcode.com/paper/lightningdot-pre-training-visual-semantic>
- [19] HUO Y, ZHANG M, LIU G, et al. Wenlan: bridging vision and language by large-scale multi-modal pre-training [EB/OL]. [2022-01-10]. <https://paperswithcode.com/paper/wenlan-bridging-vision-and-language-by-large>
- [20] FEI F, LU Z, GAO Y, et al. Wenlan 2.0: make ai imagine via a multimodal foundation model [EB/OL]. [2022-01-10]. <https://paperswithcode.com/paper/wenlan-bridging-vision-and-language-by-large>

作者简介



卢志武, 中国人民大学高瓴人工智能学院教授、博士生导师; 主要研究方向为机器学习、计算机视觉等; 设计了首个公开的中文通用图文预训练模型文澜 BriVL; 发表论文 90 余篇。



金琴, 中国人民大学信息学院教授、博士生导师; 主要研究方向为多媒体智能计算、人机交互; 在多媒体情感计算、视觉描述生成、跨模态交互等研究与应用中取得了突出成果, 蝉联多项国际权威竞赛冠军, 包括: 2017—2021 年 TRECVID 视频描述 (VTT) 评测冠军、2018—2020 年 CVPR ActivityNet Dense Video Captioning 竞赛冠军、2017—2019 年 ACM Multimedia Audio-Visual Emotion Challenge (AVEC) 竞赛冠军, 获得 2019 年之江杯全球人工智能大赛视频内容描述生成冠军; 发表论文 100 余篇。



宋睿华, 中国人民大学高瓴人工智能学院院长聘副教授, 曾任微软亚洲研究院主管研究员和微软小冰首席科学家, 担任 SIGIR 2021 短文的主席、EMNLP 2021 的资深区域主席、《Information Retrieval Journal》的主编; 提出的算法完成了人类史上第一本人工智能创作的诗集《阳光失了玻璃窗》, 参与完成了首个公开的中文通用图文预训练模型文澜 BriVL; 发表论文 90 余篇, 申请国际专利 25 项。



文继荣, 中国人民大学长聘教授, 现任高瓴人工智能学院执行院长和信息学院院长、大数据管理与分析方法研究北京市重点实验室主任, 并担任北京智源人工智能研究院首席科学家、北京市第十三届政协委员、中央统战部党外知识分子建言献策专家组专家, 入选首批“北京高校卓越青年科学家计划项目”, 还担任 AIRS 2016 大会名誉主席、CCIR 2017 大会主席、SIGIR 2018 领域主席、SIGIR 2020 程序委员会主席、WWW 2021 领域主席、《ACM Transactions on Information Systems》《IEEE Transactions on Knowledge and Data Engineering》的编委; 长期从事大数据和人工智能领域的研究; 发表论文 200 余篇。

鹏程·盘古:大规模自回归中文预训练语言模型及应用



Pengcheng-PanGu: Large-Scale Autoregressive Pre-Trained Chinese Language Model with Auto-Parallel Computation and Its Application

曾炜/ZENG Wei^{1,2}, 苏腾/SU Teng³, 王晖/WANG Hui¹,
田永鸿/TIAN Yonghong^{1,2}, 高文/GAO Wen¹

(1. 鹏城实验室, 中国 深圳 518055;

2. 北京大学, 中国 北京 100871;

3. 华为技术有限公司, 中国 杭州 310052)

(1. Pengcheng Laboratory, Shenzhen 518055, China;

2. Peking University, Beijing 100871, China;

3. Huawei Technologies Co., Ltd., Hangzhou 310052, China)

DOI: 10.12142/ZTETJ.202202006

网络出版地址: <https://kns.cnki.net/kcms/detail/34.1228.tn.20220408.1733.010.html>

网络出版日期: 2022-04-12

收稿日期: 2022-02-26

摘要: 在鹏城云脑 II 上训练了全球首个拥有全开源 2 000 亿参数的自回归中文预训练语言大模型——鹏程·盘古。鹏程·盘古模型基于 1.1 TB 高质量中文训练数据, 采用全场景人工智能计算框架 MindSpore 自动并行技术实现了五维并行训练策略, 从而可将训练任务高效扩展到 4 096 个处理器上。对比实验表明, 在少样本或零样本情况下, 鹏程·盘古模型在多个中文自然语言理解或生成任务上都具有较优的性能。在此基础上, 鹏程·盘古模型在大模型压缩、提示微调学习、多任务学习以及持续学习等方面也取得了很好的应用效果。

关键词: 大规模预训练语言模型; 鹏城云脑 II; 大规模分布式训练; 中文理解与生成; 提示微调学习

Abstract: The world's first large-scale autoregressive pre-trained Chinese language model named Pengcheng-PanGu with up to 200 billion parameters is presented. Pengcheng-PanGu is developed under the Pengcheng cloud brain II. 1.1 TB high-quality Chinese data from a wide range of domains to pre-train the model are collected. The training parallelism strategy is implemented based on all-scenarios artificial intelligence computing framework MindSpore Auto-parallel, which composes five parallelism dimensions to scale the training task to 4 096 processors efficiently. The experimental results demonstrate the superior capabilities of Pengcheng-PanGu in performing various natural language understanding and natural language generation tasks under few-shot or zero-shot settings. On this basis, Pengcheng-PanGu model has also achieved better application results in large model compression, prompt fine-tuning, multi-task, and continuous learning.

Keywords: large-scale pre-trained language models; Pengcheng cloud brain II; large-scale distributed training; Chinese language understanding and generation; tip fine-tuning learning

近年来, 有关大规模预训练语言模型 (PLM) 的研究^[1-9]在自然语言处理 (NLP) 领域取得了巨大的突破。通过自监督方式从大规模语料库中学习文本的上下文表示, 预训练语言模型在完成自然语言理解和自然语言生成 (NLG) 等任务时所表现的性能已达到国际先进水平。A. RADFORD^[10]等首次提出基于自回归语言模型 (ALM) 的预训练模型——GPT。通过在大规模文本数据上进行无监督预训练, 并针对不同有监督任务进行微调, GPT 模型的性能在各种 NLP 任务上均获得了显著提升。

2020 年, 美国 OpenAI 团队推出 GPT 系列模型的最新版本 GPT-3^[11]。其中, 最大的 GPT-3 模型包含 1 750 亿个参数, 能使用 570 GB 的文本数据进行训练。除了具有高质量的文本生成能力外, 在没有进行特定任务微调的情形下, GPT-3 模型小样本学习和零样本学习的性能会随着模型参数的增加而稳步提升。有些任务的性能甚至达到了当前最高水平。GPT-3 模型的提出是革命性的, 它减轻了人们为新任务标记更多示例和再次训练模型的负担, 成为模拟人类小样本学习能力的新范式, 为探索通用人工智能 (AI) 开辟了新途径。

目前, GPT-3 模型主要是基于英文语料数据训练出来的, 且只能通过 OpenAI 应用程序接口 (API) 进行有限度访问。为了促进中文预训练语言模型的研究和应用, 以鹏城实

基金项目: 广东省重点领域研发计划“新一代人工智能”重大专项 (2021B0101400002)

验室为首的联合团队在基于昇腾910芯片的E级智能算力平台(鹏城云脑II)上训练了全球首个全开源2 000亿参数的自回归中文预训练语言大模型——鹏程·盘古。

当模型规模超过100亿时,模型越大,模型训练的难度就越高。其中,模型训练面临的技术挑战主要包括以下几个方面:

(1) 模型设计。随着模型规模的扩大,训练过程中可能会出现收敛缓慢甚至发散的问题。在前期工作的基础上,鹏程·盘古模型将基于Transformer的ALM作为基础架构,并在Transformer层之上增加了Query层以诱导模型的预期输出。实验证明,该架构具有很好的扩展性,能够有效支持2 000亿参数规模的模型训练。

(2) 训练语料库。训练语料对一个强大、可扩展的预训练模型至关重要。一方面,语料的数据量应该足以满足一个预训练大模型的需求;另一方面,语料数据应是高质量和多样性的,以确保PLM的通用性。为了覆盖广泛的中文语料库,鹏城团队从Common Crawl、电子书、百科全书等资源中收集大量数据,并在此基础上,对数据进行多重过滤和清洗,以确保语料数据满足高质量和多样性需求。

(3) 分布式训练。2 000亿参数规模的鹏程·盘古模型对内存的需求远远超出了目前普通多机多卡集群。因此,模型需要在大规模AI集群上进行基于模型切分的并行训练。然而,在大规模AI集群上保持高资源利用率的同时,模型很难获得较大的端到端吞吐量。当涉及硬件拓扑结构时,这个问题变得更具挑战性。通过将多维并行与精心设计的并行策略结合起来,鹏城团队在2 048个Ascend 910处理器^[12]大集群上,基于昇腾处理器的异构计算架构(CANN)完成了鹏程·盘古模型的高效并行训练。

鹏城团队在1.1 TB高质量中文文本语料库上训练了鹏程·盘古2.6B、鹏程·盘古13B和鹏程·盘古200B 3个模型,并评估了鹏程·盘古2.6B、鹏程·盘古13B两个模型在16个NLP下游任务上的小样本学习能力。实验结果表明,随着模型参数规模的扩大,鹏程·盘古模型在各种下游任务上的性能表现会更优异。

然而,大模型如何赋能实际应用仍然面临很大挑战。例如,当模型太大时,如何通过有效的模型压缩来赋能边缘应用场景?如何将应用任务转化为大模型的原始任务,并通过提示微调学习技术来实现NLP模型训练新范式?如何针对新的数据集和任务,在大模型基础上开展持续学习,并构建高效持续演化的大模型生态?针对这些挑战,我们进一步研发了鹏程·盘古增强版模型。该模型在大模型压缩、提示微调学习、多任务学习以及持续学习等方面均表现出很好的

效果。

1 模型架构

鹏程·盘古是一个基于海量文本语料库进行预训练得到的大规模ALM。该模型的训练语料绝大部分是中文。该模型会对语料库中所有Token的生成过程进行建模。一个序列中的一个Token的生成取决于它前面的所有Token。假设一个序列 $X = \{x_1, x_2, \dots, x_N\}$ 由 N 个Token组成,那么训练目标可以表述为最大化对数似然:

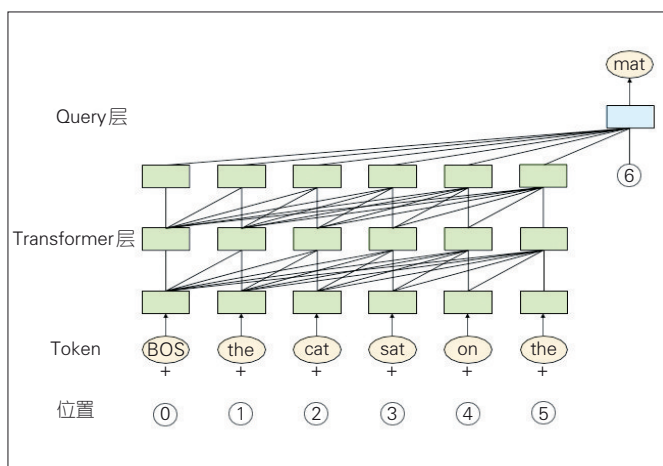
$$\mathcal{L} = \sum_{n=1}^N \log p(x_n | x_1, \dots, x_{n-1}; \theta), \quad (1)$$

其中, $p(x_n | x_1, \dots, x_{n-1}; \theta)$ 是指,在知道前 $n-1$ 个Token $x_{1:n-1}$ 的情况下,观察到第 n 个Token x_n 的概率; θ 表示模型参数。如图1所示,鹏程·盘古保留了Transformer^[13]架构,在Transformer层之上还构建了Query层。Query层可用来预测下一个Token。

1.1 Transformer 层

(1) 多头注意力。第 l 层的自注意网络由4个投影矩阵组成: $\mathbf{W}_h^k, \mathbf{W}_h^q, \mathbf{W}_h^v, \mathbf{W}_h^m$,且它们均属于集合 $\mathbb{R}^{d \times d/N_h}$ 。其中, d 为隐藏维度, h 为头的索引, N_h 为头的数量。根据前一层的输出 $H_{l-1} \in \mathbb{R}^{N \times d}$,我们可以计算出当前层的状态。该状态包含3个主要部分:Query $Q_h = H_{l-1} \mathbf{W}_h^q$ 、Key $K_h = H_{l-1} \mathbf{W}_h^k$ 和Value $V_h = H_{l-1} \mathbf{W}_h^v$ 。注意力函数的计算方法为:

$$\begin{aligned} A_h &= Q_h K_h^\top = H_{l-1} \mathbf{W}_h^q \mathbf{W}_h^{k^\top} H_{l-1}^\top, \\ \text{Attention}_h(H_{l-1}) &= \\ \text{Softmax}\left(\frac{A_h}{\sqrt{d}}\right) V_h &= \text{Softmax}\left(\frac{A_h}{\sqrt{d}}\right) H_{l-1} \mathbf{W}_h^v. \end{aligned} \quad (2)$$



▲图1 鹏程·盘古模型结构

在计算完包含多个头的注意力后,输出就可以变成:

$$\begin{aligned} \text{MHA}(H_{l-1}) &= \sum_{h=1}^{N_h} \text{Attention}_h(H_{l-1})W_h^m, \\ H_l^{\text{MHA}} &= H_{l-1} + \text{MHA}(\text{LayerNorm}(H_{l-1})). \end{aligned} \quad (3)$$

(2) 前馈神经网络 (FFN)。FFN 由两个线性层组成,相关参数为 $W^1 \in \mathbb{R}^{d \times d_f}$ 、 $b^1 \in \mathbb{R}^{d_f}$ 、 $W^2 \in \mathbb{R}^{d_f \times d}$ 、 $b^2 \in \mathbb{R}^d$,其中 d_f 是内层维度。如果将多头注意力层 (MHA) 输出作为输入,那么 FFN 层的输出为:

$$\begin{aligned} \text{FFN}(H_l^{\text{MHA}}) &= \text{GeLU}(H_l^{\text{MHA}}W^1 + b^1)W^2 + b^2, \\ H_l &= H_l^{\text{MHA}} + \text{FFN}(\text{LayerNorm}(H_l^{\text{MHA}})). \end{aligned} \quad (4)$$

对于 MHA 和 FFN,本文采取了 Pre-layer Normalization 方案。该方案可以使 Transformer 模型训练变得更加简单、高效^[14]。

1.2 Query 层

模型在 Transformer 层之上堆叠了一个 Query 层,目的是输出一个明确的引导。在 ALM 的预训练阶段,Query 层可被用来预测下一个 Token。Query 层的结构与 Transformer 层类似。在计算注意力机制的时候,Query 层会对表示下一个位置的位置嵌入 $p_n \in \mathbb{R}^d$ 做 Query 向量处理。具体来说,假设 H_l 是最上层 Transformer 层的输出,则 Query 层的注意力向量可以表示为:

$$a_h = p_n W_h^q W_h^{kT} H_L^T. \quad (5)$$

随后, MHA 和 FFN 的计算方式仍与原始 Transformer 相同。如果把最终的输出表示为 o_n ,则下一个 Token 的负对数似然就可以写为:

$$\text{CrossEntropy}(x_n, \text{Softmax}(o_n W^o + b^o)), \quad (6)$$

其中, x_n 表示真实 Token, W^o 、 b^o 是任务相关的额外参数。

1.3 模型配置

为了评估鹏程·盘古模型的扩展能力,本文训练了3个参数不断增加的模型,即鹏程·盘古2.6B、鹏程·盘古13B和鹏程·盘古200B。表1展示了这3个模型的详细配置,包括参数总数量、Token 的隐藏维度、前馈层的内层维度和注意力的头数。

2 数据集

超大规模高质量中文语料数据集对训练千亿级参数规模

的鹏程·盘古模型至关重要。目前已有3个100 GB以上规模的中文语料数据集,它们分别是:(1)从 Common Crawl 抽取得到的 CLUECorpus2020^[15],该模型的数据量为100 GB;(2)阿里巴巴集团发布的 M6^[16]中文多模态模型,该模型使用300 GB 语料;(3)北京智源研究院面向合作者发布的包含300 GB 高质量中文语料 WuDaoCorpus。然而,与目前同等规模参数量的英文预训练模型所使用的数据量相比,上面这些中文语料数据仍然不能满足2 000亿中文预训练语言模型的训练数据需求。

尽管像 SogouT 和 Common Crawl 等原始网页数据已经包含大量的中文语料数据,但是构建一个可满足2 000亿参数模型训练需求的大规模语料数据集仍需要解决诸多问题。这些问题包括:(1)原始网页数据质量参差不齐,语料预处理流程繁琐复杂;(2)海量原始语料数据处理缺少大规模存储和计算能力的支撑;(3)缺乏一个有效准确的数据质量评估方法。

为解决上述问题,我们搭建了一个大规模中文语料数据处理平台,以提升海量数据采集、清洗、过滤等处理效率,并以此构建了一个1.1 TB 的高质量中文语料数据集。在数据集的构建过程中,我们采用人工评估与模型评估相结合的方法为数据集的清洗、过滤以及训练数据集的选择提供指导。

2.1 数据集构建

为了构建一个大规模高质量中文语料数据集,我们收集了包含开放数据集、Common Crawl 原始网页数据、百科数据、新闻数据、电子书籍等近80 TB 的原始数据。如图2所示,数据集构建流程包括3个主要步骤:基于规则的数据清洗、基于模型的过滤、数据去重。我们通过人工和模型分别对数据质量进行评估,并且通过不断迭代前两个步骤来提升数据质量。整个数据集的构建过程是基于 Spark/Hadoop 搭建的大数据处理平台完成的。该平台使数据处理效率得到了明显提升。

2.1.1 数据清洗和过滤

在图2所示的5种数据来源中,Common Crawl 的数据占

▼表1 鹏程·盘古模型的规模和参数

模型	参数总数量/亿	层数	内层维度	FFN大小	头数
鹏程·盘古 2.6B	26	32	2 560	10 240	32
鹏程·盘古 13B	131	40	5 120	20 480	40
鹏程·盘古 200B	2 070	64	16 384	65 536	128

FFN:前馈网络

比虽然最大,但是它包含了大量低质量的网页数据。因此,我们首先采用如下规则对 Common Crawl 的原始数据进行清洗。

- 去除中文字符低于 60% 或者字符数小于 150 的数据 (仅有网页名称的数据也会被去除);
- 去除特殊字符,并去除在一个网页中重复出现的段落;
- 通过广告关键词去除包含大量广告的网页数据;
- 将所有繁体中文转换为简体中文;
- 识别并去除网页导航页。

在完成原始数据清洗后,我们采用 3 个过滤器来进一步过滤数据中的敏感词、广告、低质量段落等信息。

- 关键词过滤。构建一个包含 724 个敏感词的词库,并通过敏感词库去除包含 3 个以上敏感词的网页数据。
- 基于模型的过滤。为了进一步去除垃圾广告和垃圾邮件数据,我们通过人工标注数据来训练一个 FastText 文本分类模型。负样本为从 Common Crawl 数据中人工挑选的 1 万条垃圾文本数据,正样本为从高质量中文语料数据中抽样得到的数据。基于 FastText 的文本分类模型可以对语料进行垃圾过滤处理。
- 低质量文本过滤。借鉴 GPT-3 的数据处理策略,我们训练了一个数据质量评分模型。该模型可去除得分较低的文本 (详见 GPT-3 附录-A^[1])。

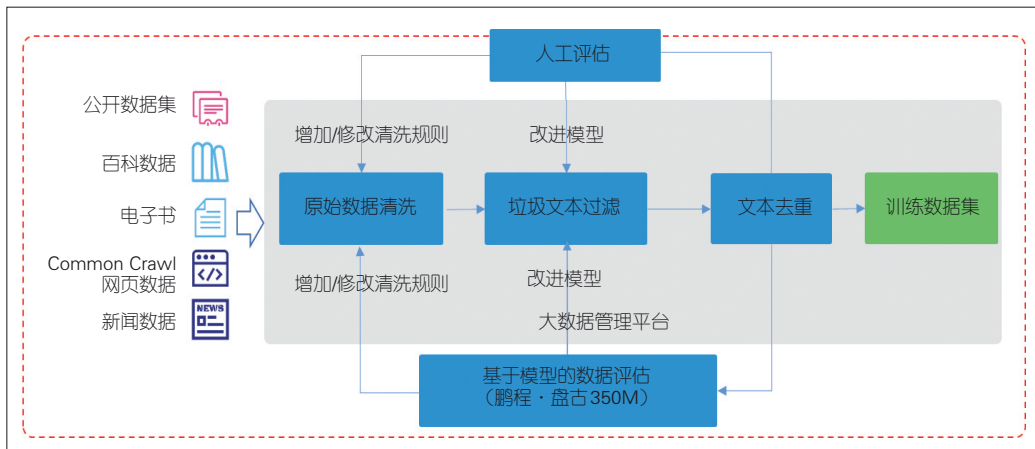
2.1.2 文本去重

由于全量数据太大,基于 Spark 的 MinHashLSH 算法在对 200 MB 数据进行去重时需要消耗至少 8 h 的时间,去重效率较低。为了加速文本数据去重过程,我们设计了一种分布式海量文本数据重复检测和去重算法。针对 500 GB 语料数据的去重任务,该算法能够将原本 20 000 h 的时间缩短至 3.5 h,极大地提升了去重效率。

2.1.3 数据质量评估

数据清洗过程中的一个重要问题是如何确定清洗规则或数据过滤阈值。对此,我们提出人工和模型相结合的数据质量评估方法,在每一轮数据清洗和过滤后,对数据清洗和过滤的效果进行评估,并通过清洗过滤、质量评估的多轮迭代来提升数据质量。其中,人工评估数据原则主要从句子通顺性、文本低质量内容占比 (如广告短语、重复短句、敏感词等) 两个维度进行评估。

人工评估虽然有效,但是对于大规模语料数据来说,能够评估的语料占比太小,不能充分反应数据集的整体质量。为了提高数据评估准确度,我们从待评估全量数据中抽取 30 GB 的数据来训练鹏程·盘古 350M 模型,并通过该模型在高质量数据集中的困惑度 (PPL) 指标来评估数据质量。模型在高质量数据中的 PPL 越小,数据集清洗和过滤所采用的清洗规则和过滤模型就越好。



▲图2 鹏程·盘古模型的训练数据处理流程

▼表2.1.1 TB中文语料数据组成

数据来源	大小/GB	数据来源	数据处理步骤
开放数据集	27.9	15 个开放数据集,如 DuReader、BaiduQA、CAIL2018、Sogou-CA 等	数据格式转换、文本去重
百科数据	22.0	百度百科、搜狗百科等百科类数据	文本去重
电子书籍	299.0	不同主题的电子书籍,如小说、历史、诗歌、古文等	敏感词过滤、基于模型的文本过滤
Common Crawl	714.9	2018 年 1 月—2020 年 12 月的 Common Crawl 网页数据	数据清洗、过滤、去重等所有数据处理步骤
新闻数据	35.5	1992—2011 年的新闻数据	文本去重

2.2 数据采样策略

通过图 2 中的数据处理流程,我们从 5 种来源的近 80 TB 原始数据中清洗并构建了一个 1.1 TB 的高质量中文语料数据集。数据集组成和数据处理方法如表 2 所示。基于由上述步骤构建的语料数据,采样形成的两个训练数据集被用于训练鹏程·盘古 2.6B、鹏程·

盘古 13B 和鹏程·盘古 200B 模型。两个训练数据集的数据量分别是 100 GB 和 1 TB。如表 3 所示, 训练数据集对每个数据源的数据进行采样。采样比例和重复次数越大, 数据源的质量就越好。在两个训练集的词 Token 数量分布方面, 100 GB 训练集和 1 TB 训练集的平均段落长度分别是 239、405 个 Token。可以看出, 1 TB 的平均段落长度更长。这是因为 1 TB 训练集中 Common Crawl 数据的占比更大。值得注意的是, 训练数据段落长短与模型生成效果有关。当训练样本平均长度较短时, 模型倾向于生成更短的句子, 从而有利于模型处理下游任务中需要生成短句的任务; 反之, 当训练样本平均长度较长时, 模型会倾向于生成更长的句子。

3 并行训练系统

鹏程·盘古 200B 的模型训练将面临巨大挑战。比如, 鹏程·盘古 200B 的内存存储需求就高达 750 GB。由于梯度和优化器状态对参数更新也很重要, 因此训练如此庞大的模型所消耗的内存会比参数存储要高好几倍。相比之下, 现代 AI 处理器 (如图形处理器、Ascend 910 AI 处理器^[12]) 的内存约为 30~40 GB。因此, 将模型切分到设备 (处理器) 的集群中是不可避免的。为此, 我们需要应对两个方面的技术挑战。首先, 多个不同的并行功能应该结合起来, 以使模型获得较高的端到端性能。然而, 由于策略空间巨大, 寻找最佳的策略组合是一个挑战。其次, 并行训练应满足易用性与高效性的双重需求, 底层与并行相关的处理逻辑应该与模型定义的处理逻辑相解耦。

在基于昇腾 910 芯片的 E 级智能算力平台 (鹏城云脑 II) 上, 我们使用 MindSpore 自动并行技术来应对上述两个方面的挑战, 从而最大限度地提高计算通信比。该自动并行技术支持五维度的并行能力, 并使用拓扑感知调度将切片的模型映射到集群上, 以获得较高的端到端性能。此外, 该自动并行技术只需要对单机代码进行最少的代码修改, 就可以实现快捷高效的超大模型并行训练。

(1) 五维并行和拓扑感知调度

最常用的并行方式是数据并行, 它在设备之间划分训练

的批次大小, 并在执行迭代优化命令之前与来自不同设备的梯度信息保持同步, 如图 3 (a) 所示。模型并行有 3 种方式。第 1 种是算子级并行^[17-23], 它对每个算子所涉及的张量进行切分。如图 3 (b) 所示, 算子级并行通过对参数和显存进行切片来减少显存消耗, 同时通过通信优化来使连续算子之间的分布式张量状态保持一致。第 2 种是流水并行^[24-28], 它将总的模型层划分为不同阶段, 然后将不同阶段的模型层放置到不同的设备上, 如图 3 (c) 所示。每台设备只拥有模型层次的一部分, 可大大节省显存占用, 并使通信只发生在不同状态的边界上。第 3 种机制是优化器并行^[29], 其作用是减少由数据并行所导致的优化器内存冗余和计算消耗。图 3 (d) 中前向运算阶段的一些中间结果要在显存中驻留相当长的时间, 以加速后向阶段的梯度计算。如图 3 (e) 所示, 重计算前向运算结果可以释放部分中间结果, 以减少整个训练阶段显存消耗。需要指出的是, 每个维度的并行都要通过计算 (或通信) 开销来换取显存 (或吞吐量) 收益。因此, 为了获得最大的端到端吞吐量, 我们需要在多维并行之间找到一个最佳组合平衡点。而设备集群中的异构带宽使这变得更具挑战性。

(2) 混合并行训练

图 4 展示了鹏程·盘古 200B 模型的混合并行方案。首先, 将模型总层次 (64 层) 划分成 16 个状态, 每个状态包含 4 层。每一层会为每个算子切分所需要的参数和张量。具体来说, Query(Q)、Key(K) 和 Value(V) 算子相关的参数被切分为 8 片。我们将这 3 个算子的输入张量划分为 16 个切片, 并以此确定优化器并行的维度。该层中其他算子的并行策略也以同样的方式进行配置。每层算子都首先被切分, 然后再执行下发命令。这有效降低了额外的计算开销。在本方案中, 我们总共使用了 2 048 个来自鹏城云脑 II 的 Ascend 910 AI 处理器。

鹏程·盘古 200B 模型具体的混合并行策略为: 数据并行 8 路、算子级并行 8 路、流水并行 16 路, 在数据并行的同时叠加优化器并行。模型会将通信量大的并行方式 (算子级并行) 放置在服务器内部的多卡之间, 将通信量较小的并行

▼表 3 鹏程·盘古模型训练数据采样策略

数据来源	鹏程·盘古 200B			鹏程·盘古 2.6B 和鹏程·盘古 13B	
	Token 数量/亿	数据占比/%	训练完成时重复次数	Token 数量/亿	数据占比/%
开放数据集	258	10.23	3.65	70	27.99
电子书籍	309	12.23	0.41	56	18.00
Common Crawl	1 762	62.81	0.85	25	10.00
新闻数据	198	7.83	2.20	56	22.00
百科数据	58	6.90	3.00	58	23.00

方式（流水并行）放置在同一机架内的服务器之间，将部分数据并行（叠加优化器并行）放置在不同机架之间。因此，通信可以与计算同时进行，对带宽要求较低。

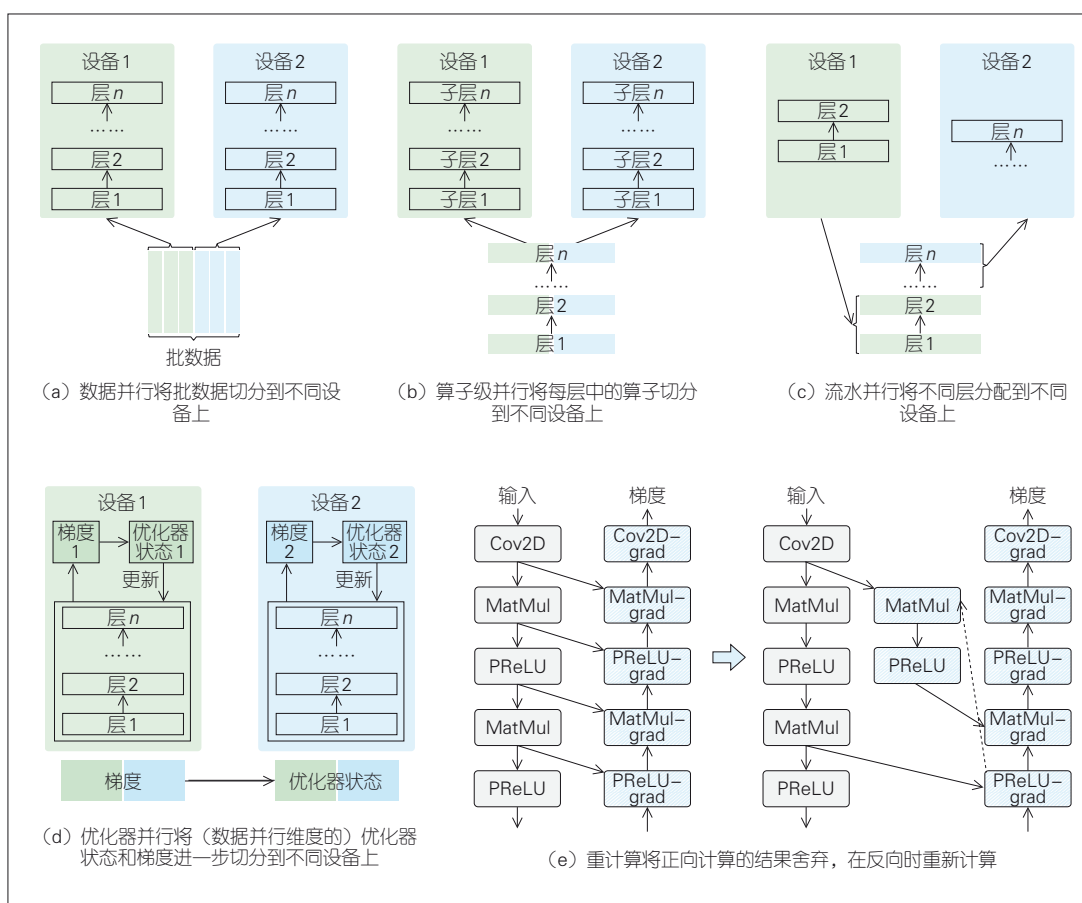
4 实验

4.1 训练细节

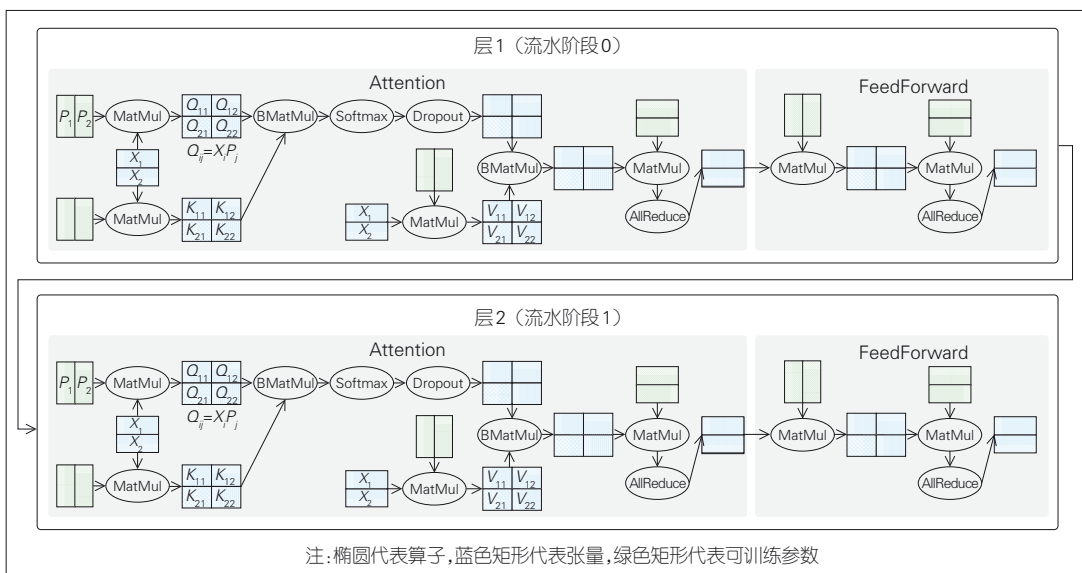
鹏程·盘古模型是基于华为 Mindspore 框架开发的，它采用由 2 048 块 Ascend910 AI 处理器组成的集群进行训练，并最终扩展到全机 4 096 块 Ascend910 AI 处理器集群上。模型的详细配置如表 4 所示。在训练鹏程·盘古 200B 模型时，我们首先采用 2 048 块处理器，然后将其切换到 1 024 块上继续进行训练。实验将字节对编码（BPE）作为分词器，词表的规模为 40 000，并且所有模型均采用 1 024 的序列长度。

鹏程·盘古模型的训练损失曲线如图 5 所示。因为鹏程·盘古 200B、鹏程·盘古 13B 和鹏程·盘古 2.6B 模型训练的批量大小不同，所以我们用

Token 数作为 X 轴。由图 5 可以看出，鹏程·盘古 200B、鹏程·盘古 13B 和鹏程·盘古 2.6B 的模型训练损失分别收敛在 2.49、2.58 和 2.64，并且在训练结束时训练损失仍然在下降。

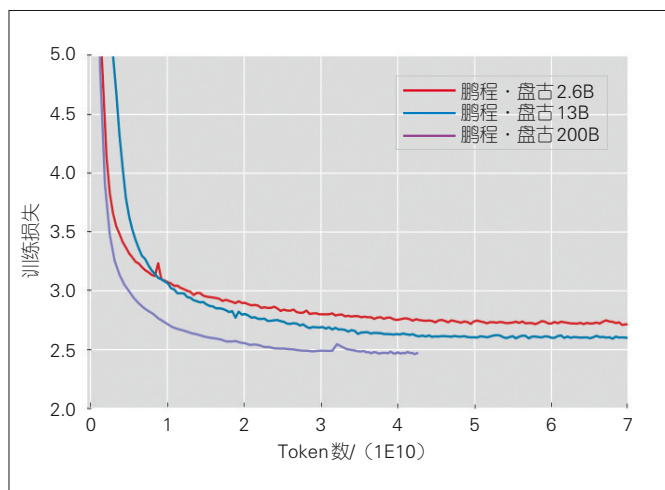


▲图 3 5 种并行方式及其优化显存和吞吐量的过程



▲图 4 一个简化的鹏程·盘古并行策略

这表明模型精度仍有提升的潜力。本文在验证集上评估了模型的 PPL 性能。其中，验证集是从 Common Crawl 数据集中随机抽取的。评估结果表明，模型越大，PPL 就越低，模型



▲图5 不同参数规模下鹏程·盘古的训练曲线

▼表4 鹏程·盘古的详细训练配置

模型	训练步长	处理器个数	Adam Betas	学习率	权重衰减
鹏程·盘古 2.6B	0~70 000	512	$\beta_1=0.9$ $\beta_2=0.999$	$1E-4$	0.01
鹏程·盘古 13B	0~84 000	1 024	$\beta_1=0.9$ $\beta_2=0.980$	$5E-5$	0.01
鹏程·盘古 200B	0~130 000 130 000~260 000	2 048 1 024	$\beta_1=0.9$ $\beta_2=0.950$	$2E-5$	0.10

性能也就越优。

4.2 任务描述

本文在多种自然语言处理下游任务的基础上来评估模型的性能。与GPT-3^[11]类似,实验采用3种不经任务微调的配置:零样本学习、单样本学习和小样本学习。如果能获取到测试集,每个下游任务就会在测试集上进行评估。参与的16个下游任务包含7个类别:完形填空与补全、阅读理解、闭卷问答、指代消解、常识推理、自然语言推理、文本分类。

4.3 评估细节

由于测试方式不同,本文将所有任务分为两大类:生成类任务和分类任务。

(1) 生成类任务的评测方法

生成类任务包含词级别和句子级别的生成任务。鹏程·盘古模型天然具备强大的文本生成能力,能够采用模型自然生成文本的方式生成此类任务的答案。对于中文上下文词语预测数据集(WPLC)、中文填空型阅读理解(PD&CFT)和阅读理解评测(CMRC2017)这类完形填空与补全任务,上下文可作为提示被放置在待预测位置的前面。而对于阅读理

解和闭卷问答任务,模型则能够根据需要设计相应的提示模板。例如,阅读理解任务可将样本填充到“Reading document: \$Document Question: \$Question Answer:”模板中,并将其作为提示输入到模型中。类似于GPT-3,小样本学习采用上下文学习的方式,即把K个提示相互拼接。其中,前K-1个提示均包含答案,最后一个提示的答案则通过模型预测来获得。

(2) 分类任务的评测方法

分类任务主要采用基于PPL的评测方法。针对每组<段落,标签>数据对,该方法会根据预设计模板自动生成输入。由模板生成的序列将被输入到模型中,同时模型将计算出相应的PPL值。具有最小PPL值的标签将被作为该段落的预测结果。与生成类任务评测类似,分类任务也采用上下文学习策略来完成小样本学习任务。

4.4 实验结果

本文对比了鹏程·盘古2.6B模型和CPM2.6B模型^[3]在16个中文下游任务上的表现。鹏程·盘古2.6B模型在11个零样本学习任务、12个单样本学习任务、14个小样本学习任务上的表现均超越CPM 2.6B模型。实验结果表明,相比于CPM2.6B模型,鹏程·盘古2.6B模型具有更强的上下文学习能力(尤其在小样本学习和生成方面)。在生成任务方面,鹏程·盘古2.6B模型要比CPM2.6B模型平均高出6个百分点。具体地,在阅读理解任务和闭卷问答任务上,鹏程·盘古2.6B模型比CPM2.6B模型高出5个百分点;在无选项完形填空任务上,鹏程·盘古2.6B模型比CPM2.6B模型高出7个百分点。在PPL任务方面,鹏程·盘古2.6B模型与CPM2.6B模型相当,而在TNEWS和IFLYTEK分类任务上的表现则不如CPM2.6B模型。造成这种现象的主要原因是CPM2.6B模型和鹏程·盘古2.6B模型的训练语料具有差异性。

我们同时对对比了鹏程·盘古13B和鹏程·盘古2.6B在16个中文NLP下游任务上的表现。鹏程·盘古13B在所有生成式任务和绝大多数PPL任务上的表现,均明显优于鹏程·盘古2.6B模型。在CMRC2018、DRCD和WebQA任务上,鹏程·盘古13B小样本学习的性能比零样本学习高10个百分点。这说明鹏程·盘古13B模型具有极强的上下文学习能力。鹏程·盘古13B在16个下游任务中的表现比鹏程·盘古2.6B高出近3个百分点。具体地,鹏程·盘古13B模型在阅读理解和闭卷问答任务上的表现比鹏程·盘古2.6B模型高出近4个百分点,在无选项完形填空任务上的表现比鹏程·盘古2.6B高出近2个百分点。在自然语言推理(NLI)任务上,鹏程·盘古13B模型的表现则不如鹏程·盘古

2.6B, 这与 GPT-3 实验结果是一致的。总之, 鹏程·盘古 13B 模型和鹏程·盘古 2.6B 模型的对比实验表明: 更大规模的预训练模型的性能通常能在小样本学习任务上取得提升。

5 大模型应用

5.1 模型压缩

虽然鹏程·盘古模型具备强大的能力, 但超大规模的模型参数量却限制了它的应用。我们通常希望应用端能够在保持几乎同等性能的条件就可得到较小参数规模的模型, 以便提升应用效率。因此, 我们研究了鹏程·盘古的模型压缩技术, 并采用量化与参数共享的方法实现了鹏程·盘古 13B 模型和鹏程·盘古 2.6B 模型在单张 Ascend 910 卡上的应用。其中, 量化是指借助低精度类型加载模型, 用 FP16 代替大部分 FP32 类型参数, 同时对量化噪声和数值溢出进行处理; 参数共享是指将部分层的参数进行共享, 例如将输出层参数与嵌入层参数进行共享。这种压缩技术使显存占用降低 50%, 系统性能波动仅为 2% 左右。为了评估压缩技术对模型性能的影响, 实验测试了部分下游任务在压缩前后的性能指标。结果表明, 在闭卷问答任务 (WebQA) 上, 压缩后鹏程·盘古 13B 模型的 F1 值比压缩前小 0.01; 在代词消歧任务 (CLUEWSC2020) 中, 压缩后鹏程·盘古 13B 模型的精度仅比压缩前下降 1 个百分点。

5.2 模型移植

为了便于更多用户使用鹏程·盘古模型, 我们将该模型从 Mindspore 框架成功移植到 PyTorch 框架下。移植流程主要包括 3 个步骤。

第 1 步: 在 PyTorch 框架上复现鹏程·盘古。复现工作是基于开源分布式 Transformer 的 Megatron 框架实现的, 即在 Transformer 的 decoder 解码结构上加一层 Query 层。

第 2 步: 手动转换模型文件。由于鹏程·盘古的部分算子不能转成开放神经网络交换 (ONNX) 通用模型格式, 所以我们需要对模型文件进行手动转换。手动转换模型文件的流程包括: (1) 提取 Mindspore 模型中的参数数据和参数名称, 并将提取的参数保存为数组数据; (2) 手动对齐 Mindspore 和 PyTorch 模型的参数名称和参数维度, 并将其保存为 PyTorch 模型文件类型。

第 3 步: 对齐 PyTorch 实现版本和 Mindspore 实现版本的并行策略。在进行分布式训练时, 我们需要把隐藏态分割到不同的设备上。然而, 基于两个不同框架实现的分割方案存在一定的差异。图 6 展示了 Mindspore 实现版本和 PyTorch 实

现版本的模型切割策略。可以看出, 左边 Mindspore 的切割策略是直接从中将隐藏态分成两份, 然后为每个设备分配一份; 而右边 PyTorch 的切割策略则是先将隐藏态分成 3 份, 然后再把每份数据平均分配到不同的设备上。为了保证最后输出结果的一致性, 我们需要手动调整移植后模型文件的权重。

目前, 移植到 PyTorch 框架后的鹏程·盘古代码和模型文件已经在 OpenI 社区开源共享。

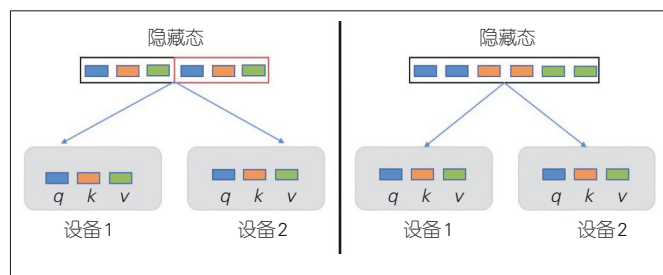
5.3 基于增量推理的在线体验服务加速

为了让更多的用户体验鹏程·盘古模型的强大功能, 我们设计并开放了在线体验服务, 目前已处理上万条用户请求。鹏程·盘古模型一次完整的在线推理可以包含多个 Tokens 的生成。模型需要根据上文输入来预测下一个 Token, 然后将预测的 Token 追加至输入内容结尾, 以便让模型继续生成下一个 Token。

通常, 在对输入序列 Pad 补到固定长度 (如 1024) 后再让模型进行自回归生成的方式, 会明显引入冗余计算。这将极大降低模型的推理能力。对此, 我们采用状态复用的改进算法 (增量推理), 来提高模型的推理能力。对于不同步的输入, 前部分序列的内容完全相同。当计算索引为 i 的位置时, 前 $0 \sim i-1$ 位置对应的状态在上一步中已进行计算。因此, 在推理过程中, 第 i 步可以复用第 $i-1$ 步的状态, 并将其和当前推理得到的状态进行拼接, 以便作为第 i 步的完整状态。系统在得到第 i 步输出的 Token 时, 即可省掉这些重复计算。这将极大提升模型的推理能力。测试结果表明, 增量推理可使模型性能提升 5 倍以上 (评估的方法是: 输入一段话, 预测下一个词的平均输出时间)。基于增量推理的在线体验服务网址为 <https://pangu-alpha.openi.org.cn/>。

5.4 鹏程·盘古增强版

鹏程·盘古增强版能针对多个下游任务进行持续的提示微调训练, 其主要创新包括:



▲图6 鹏程·盘古的模型切割策略

• 创新应用多任务学习、任务统一格式、提示微调和持续学习技术，对基本版模型进行能力扩展和增强，使模型性能得到大幅提升；

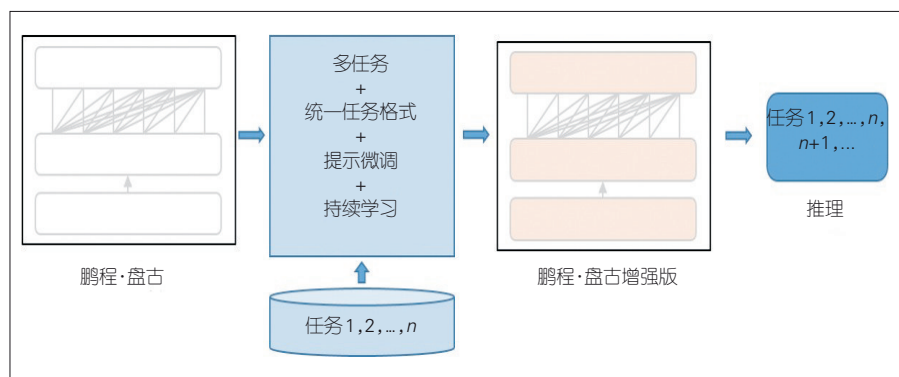
• 形成基于鹏程·盘古模型进行提示微调和持续学习的应用新范式，更好地识别用户的任务说明，同时能尽量保持模型的原始生成能力；

• 参数量为26亿规模，在中英文翻译、开放域知识问答、文本分类、摘要生成等方面的能力提升显著，在一张V100 GPU卡上就可以完成多路并行推理。

(1) 技术方案

图7显示了构建鹏程·盘古增强版模型的技术方案。该方案采用多任务、统一任务格式、提示微调和持续学习等技术方法来增强模型的任务处理能力，同时提升对任务描述的泛化能力。

(2) 统一任务格式



▲图7 鹏程·盘古增强版技术原理

我们设计了统一的任务数据格式。该统一格式旨在减少下游任务之间的差异，提高知识转移和任务描述的泛化能力。借助统一格式，我们在18个任务上构建了50多个提示，并通过提示微调技术来训练盘古增强版模型。统一任务格式的具体细节可从<https://git.openi.org.cn/PCL-Platform/Intelligence/PanGu-Alpha-Evolution>网页中查询了解。

(3) 实验性能

我们开展了大量实验来比较鹏程·盘古增强版和鹏程·盘古基本版在自然语言理解任务、自然语言生成任务中的性能。对于每个任务来说，如果能获取到测试集则在测试集上进行评估，否则就在验证集上进行评估。为了降低计算资源消耗，部分任务会从数据中随机采样部分子集来评估。性能对比结果如表5，表中“Δ”是指鹏程·盘古增强版相对鹏程·盘古基本版提升的绝对值，“相对提升”是指鹏程·盘古增强版相对鹏程·盘古基本版提升的百分比。因为鹏程·盘古不具备翻译能力，所以我们不计算这方面的相对提升百分比。结果表明，鹏程·盘古增强版在各项任务上的表现均远远优于鹏程·盘古基本版（平均相对提升高达1 064.70%）。人工评估表明，鹏程·盘古增强版具有与鹏程·盘古基本版相同的文本生成能力。

6 结束语

本文详细介绍了大规模自回归中文

▼表5 鹏程·盘古增强版优越的性能

序号	任务	指标	鹏程·盘古2.6B基本版	鹏程·盘古2.6B增强版	差值(绝对值)	相对提升/%
1	CMRC2018	Em	1.21	64.00	62.79	5 189.26
2	DRCD	Em	0.80	66.00	65.20	8 150.00
3	Dureader	Rouge-L	21.07	65.57	44.50	211.20
4	CHID	Acc.	68.73	74.00	5.27	7.67
5	PD&CFT	Acc.	38.47	88.00	49.53	128.75
6	CMRC2017	Acc.	37.83	96.00	58.17	153.77
7	CMNLI	Acc.	50.20	80.00	29.80	59.36
8	TNEWS	Acc.	60.95	88.00	27.05	44.38
9	AFQMC	Acc.	59.29	66.00	6.71	11.32
10	CSL	Acc.	50.50	52.00	1.50	2.97
11	WebQA.v1.0	Em	6.00	48.00	42.00	700.00
12	CLUEWSC2020	Acc.	73.36	84.00	10.64	14.50
14	C3	Acc.	53.42	66.00	12.58	23.55
15	LCSTS	Rouge-L	11.21	34.65	23.44	209.04
16	Wmt20 enzh	Bleu-4	0	9.84	9.84	
17	Wmt20 zhen	Blen-4	0	18.52	18.52	

预训练语言模型鹏程·盘古,并探索了该模型的具体应用。大规模语言模型虽然是当前的研究热点,但仍存在很多开放性的问题。

(1) 大规模语言模型在NLP任务上表现出较好的小样本学习能力,但目前对大规模语言模型的系统性研究仍然比较缺乏。如何训练出大规模PLM并使模型生成的文本更加规范安全、更加鲁棒、更加符合常识(或知识)仍是最具挑战性的问题。

(2) 超大规模语言模型的训练、推理和维护成本非常高。如何高效地训练出一个大模型并使模型具有持续演化能力?大规模PLM的绿色生态、持续学习演化等是一个有趣的探索方向。

(3) 大规模语言模型被认为是通向通用人工智能的重要途径。它的自监督预训练模式、小样本学习能力以及单模型多任务的适配能力都具有很好的应用前景。由于目前仍缺乏兼具逻辑推理、常识和认知能力的大模型,人们只能使用巨量参数来拟合并训练语料中长尾分布的记忆“巨兽”。如何大模型它具备人类推理、思考和认知能力仍然任重而道远。

致谢

感谢鹏城实验室提供鹏城云脑支撑本文研究。感谢鹏城实验室的王晖、张艳、颜达森、蒋芳清、易泽轩、陶恒韬、王进,华为公司的苏腾、任晓哲、廖亿、蒋欣、王志伟等为本研究做了大量工作。

参考文献

- [1] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners. [EB/OL]. [2022-02-25]. <https://arxiv.org/abs/2005.14165>
- [2] DEVLIN J, CHANG M, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [EB/OL]. [2022-02-25]. <https://arxiv.org/abs/1810.04805>
- [3] ZHANG Z Y, HAN X, ZHOU H, et al. CPM: a large-scale generative Chinese pre-trained language model [J]. AI open, 2021, 2: 93-99. DOI: 10.1016/J.AIOOPEN.2021.07.001
- [4] YANG Z, DAI Z, YANG Y, et al. XLNet: generalized autoregressive pretraining for language understanding. [EB/OL]. [2022-02-25]. <https://paperswithcode.com/paper/xlnet-generalized-autoregressive-pretraining>
- [5] LIU Y H, OTT M, GOYAL N, et al. RoBERTa: a robustly optimized BERT pretraining approach [EB/OL]. [2022-02-25]. <https://arxiv.org/abs/1907.11692>
- [6] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer [EB/OL]. [2022-02-25]. <https://arxiv.org/abs/1910.10683>
- [7] SUN Y, WANG S H, LI Y K, et al. ERNIE: enhanced representation through knowledge integration [EB/OL]. [2022-02-25]. <https://arxiv.org/abs/1904.09223>
- [8] ZHANG Z, HAN X, LIU Z, et al. ERNIE: enhanced language representation with informative entities. [EB/OL]. [2022-02-25]. <https://arxiv.org/abs/1905.07129>
- [9] WEI J Q, REN X Z, LI X G, et al. NEZHA: neural contextualized representation for Chinese language understanding [EB/OL]. [2022-02-25]. <https://arxiv.org/abs/1909.00204>
- [10] RADFORD A, NARASIMHAN K. Improving language understanding by generative pre-training [EB/OL]. [2022-02-25]. <https://paperswithcode.com/paper/improving-language-understanding-by-generative-pre-training>
- [11] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners [EB/OL]. [2022-02-25]. <https://paperswithcode.com/paper/language-models-are-unsupervised-multitask>
- [12] LIAO H, TU J J, XIA J, et al. DaVinci: a scalable architecture for neural network computing [C]//Proceedings of 2019 IEEE Hot Chips 31 Symposium. IEEE, 2019: 1-44. DOI: 10.1109/HOTCHIPS.2019.8875654
- [13] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need. [EB/OL]. [2022-02-25]. <https://paperswithcode.com/paper/attention-is-all-you-need>
- [14] XIONG R, YANG Y, HE D, et al. On layer normalization in the transformer architecture [EB/OL]. [2022-02-25]. <https://paperswithcode.com/paper/on-layer-normalization-in-the-transformer-1>
- [15] XU L, ZHANG X W, DONG Q Q. CLUECorpus2020: a large-scale Chinese corpus for pre-training language model [EB/OL]. [2022-02-25]. <https://arxiv.org/abs/2003.01355>
- [16] LIN J, MEN R, ZHOU C, et al. M6: A chinese multimodal pretrainer [EB/OL]. [2022-02-25]. <https://paperswithcode.com/paper/m6-a-chinese-multimodal-pretrainer>
- [17] SHAZEER N, CHENG Y L, PARMAR N, et al. Mesh-TensorFlow: deep learning for supercomputers [EB/OL]. [2022-02-25]. <https://paperswithcode.com/paper/mesh-tensorflow-deep-learning-for>
- [18] JIA Z H, LIN S N, QI C R, et al. Exploring hidden dimensions in parallelizing convolutional neural networks [EB/OL]. [2022-02-25]. <https://paperswithcode.com/paper/exploring-hidden-dimensions-in-accelerating>
- [19] WANG M J, HUANG C C, LI J Y. Supporting very large models using automatic dataflow graph partitioning [C]//Proceedings of the Fourteenth EuroSys Conference 2019. ACM, 2019. DOI: 10.1145/3302424.3303953
- [20] LEPIKHIN D, LEE H, XU Y Z, et al. GShard: scaling giant models with conditional computation and automatic sharding [EB/OL]. [2022-02-25]. <https://paperswithcode.com/paper/gshard-scaling-giant-models-with-conditional>
- [21] SHOEYBI M, PATWARY M, PURI R, et al. Megatron-LM: training multi-billion parameter language models using GPU model parallelism [EB/OL]. [2022-02-25]. <https://paperswithcode.com/paper/megatron-lm-training-multi-billion-parameter>
- [22] SONG L H, CHEN F, ZHUO Y W, et al. AccPar: tensor partitioning for heterogeneous deep learning accelerators [C]//2020 IEEE International Symposium on High Performance Computer Architecture (HPCA). IEEE, 2020
- [23] JIA Z H, ZAHARIA M, AIKEN A. Beyond data and model parallelism for deep neural networks [EB/OL]. [2022-02-25]. <https://arxiv.org/abs/1807.05358>
- [24] HUANG Y P, CHENG Y L, BAPNA A, et al. TensorPipe: easy scaling with micro-batch pipeline parallelism [EB/OL]. [2022-02-25]. <https://paperswithcode.com/paper/gpipe-efficient-training-of-giant-neural>
- [25] NARAYANAN D, HARLAP A, PHANISHAYEE A, et al. PipeDream: generalized pipeline parallelism for DNN training [C]//Proceedings of the 27th ACM SOSP. ACM, 2019: 1-15
- [26] FAN S Q, RONG Y, MENG C, et al. DAPPLE: a pipelined data parallel approach for training large models [EB/OL]. [2022-02-25]. <https://cs.paperswithcode.com/paper/dapple-a-pipelined-data-parallel-approach-for>
- [27] TARNAWSKI J, PHANISHAYEE A, DEVANUR N R, et al. Efficient algorithms for device placement of DNN graph operators [EB/OL]. [2022-02-25]. <https://paperswithcode.com/paper/efficient-algorithms-for-device-placement-of>
- [28] PARK J H, YUN G, YI C M, et al. HetPipe: enabling large DNN training on (whimpy) heterogeneous GPU clusters through integration of pipelined model parallelism and data parallelism [EB/OL]. [2022-02-25]. <https://arxiv.org/abs/2005.14038>
- [29] RAJBHANDARI S, RASLEY J, RUWASE O, et al. ZeRO: memory optimizations toward training trillion parameter models [C]//Proceedings of SC20: International Conference for High Performance Computing, Networking, Storage and Analysis. IEEE, 2020. DOI: 10.1109/sc41405.2020.00024

作 者 简 介



曾炜, 北京大学视频与视觉技术国家工程研究中心副研究员; 长期从事人工智能、智能计算系统、图像处理、计算机视觉、模式识别、多媒体领域的研究工作; 曾获 10 余项国家和省部级科研项目资金资助; 发表论文 50 余篇, 申请专利 20 余项。



苏腾, 华为技术有限公司分布式并行计算技术专家、MindSpore 副首席专家; 长期从事分布式并行计算系统的设计与开发工作; 主导华为分布式任务计算框架、分布式强一致性 K-V 存储框架、容器化资源管理与调度框架的设计与开发, 在大规模分布式系统方面拥有丰富的实践经验。



王晖, 鹏城实验室网络智能部开源所研究员, 曾为国防科技大学系统工程学院信息系统工程国家重点实验室教授、博士生导师; 在 NLP 大模型、分布式机器学习、联邦学习等领域开展关键技术及应用研究工作; 获得军队科技进步奖一等奖 2 项、二等奖 3 项。



田永鸿(通信作者), 北京大学博雅特聘教授、IEEE Fellow、鹏城实验室网络智能部副主任兼云脑研究所所长、2018 年国家杰出青年科学基金获得者、首届“高校计算机专业优秀教师奖励计划”获奖者; 主要研究方向为分布式机器学习、神经形态视觉和视频大数据; 累计主持国家、省部级与企业合作项目 40 余项; 获国际期刊和会议最佳论文奖 2 次、国家与省部级奖 4 次; 发表论文 280 余篇, 拥有中国和美国发明专利 90 项。



高文, 北京大学博雅讲座教授、中国工程院院士、ACM Fellow、IEEE Fellow、鹏城实验室主任、北京大学信息与工程科学部主任, 曾任中国科学院计算技术研究所研究员、副所长、所长, 中国科学技术大学副校长, 中国科学院研究生院常务副院长; 主要从事人工智能应用和多媒体技术、计算机视觉、模式识别与图像处理、虚拟现实方面的研究; 获得国家技术发明奖一等奖 1 次、二等奖 1 次, 国家科技进步奖二等奖 5 次, 获得 2005 中国十大教育英才称号和中国计算机学会王选奖。

超大规模多模态预训练模型 M6 的关键技术及产业应用



Key Technologies and Applications of Extremely Large-Scale Multimodal Pre-Trained Model M6

林俊阳/LIN Junyang, 周畅/ZHOU Chang,
杨红霞/YANG Hongxia

(阿里巴巴达摩院, 中国 杭州 311100)
(Alibaba DAMO Academy, Hangzhou 311100, China)

DOI: 10.12142/ZTETJ.202202007

网络出版地址: <https://kns.cnki.net/kcms/detail/34.1228.TN.20220410.2248.004.html>

网络出版日期: 2022-04-11

收稿日期: 2022-02-25

摘要: 阿里巴巴达摩院研发了超大规模中文多模态预训练模型 M6, 并陆续推出了百亿、千亿、万亿和十万亿参数规模的预训练模型, 实现了高效低碳的预训练, 推动超大规模预训练模型的产业化应用。同时, 推出了 M6 服务化平台, 帮助广大用户快速使用大模型。未来, 大模型在产业领域的应用将更加丰富。

关键词: 多模态预训练; 大规模预训练; 图像生成; 文本生成

Abstract: The extremely large-scale Chinese multimodal pre-trained model M6 is proposed by Alibaba DAMO Academy, and the 10 B, 100 B, 1 T, and 10 T versions of M6 are released. M6 has been trained efficiently with low carbon emission, and it has been deployed in multiple scenarios, which leads to the creation of new products as well as performance improvement. Also, to provide better services, the easy-to-use M6 platform for users to leverage large-scale pre-trained models is released by DAMO Academy.

Keywords: multimodal pre-training; large-scale pre-training; image generation; text generation

近年来, 预训练技术的诞生与迅速崛起成为人工智能(AI)发展史的一大标志。基于无监督学习和弱监督学习的预训练具备强大的迁移能力, 可以充分利用海量无标注数据, 因此能够应用于多种不同类型的下游任务中。此外, 研究人员通过扩大模型容量、扩大训练数据、降低人工标注的依赖等方式让模型取得更好的效果及通用性。随着模型规模和数据规模的不断扩大, 模型效果也会显著提升。预训练大模型的研究具有深远的学术意义, 并有着广泛的应用前景。

传统的预训练多集中于单模态数据, 且多数的预训练工作均在英文数据上实现。在很长一段时间里, 中文数据都缺乏大规模预训练模型和多模态预训练模型。自 2020 年以来, 阿里巴巴达摩院认识到这一问题的重要性, 提出了超大规模中文多模态预训练的课题。基于多模态表示学习以及超大规模预训练模型的研究, 达摩院掌握了基于超大规模多模态预训练的核心技术, 于 2021 年提出了超大规模中文多模态预训练模型 M6。在之后的一年内, 达摩院陆续发布了百亿、千亿、万亿和十万亿参数规模的超大模型。这些工作推动了

低碳 AI 的发展, 同规模的 M6 耗电量不到 GPT-3 的 1/100。达摩院还积极推进 M6 大模型的产业化落地, 这包括手机淘宝推荐、支付宝搜索推荐等 100 余种算法场景。同时, M6 利用其能力支持多个行业实现创新产品的孵化, 如 AI 服饰设计能力可以支持服饰制造行业。同时, M6 还推出了大规模预训练平台, 使得大模型的应用以服务化的形式对外提供服务。该平台也是当前下游任务覆盖最广的预训练平台。平台化使得大模型同时服务于学界和产业界, 大幅降低了大模型的门槛, 并让 AI 大模型简单易用。

1 M6 技术进展与突破

此前的研究^[1]指出, 随着数据规模、模型规模和计算资源的不断增长, 模型能力也会不断提高。在过去的几年里, 学习了海量无监督数据的预训练模型的规模实现了指数级增长。2018 年, 最大规模的 BERT^[2]和 GPT^[3]参数规模仅约为 3 亿。2019—2020 年, 具有 15 亿参数规模的 GPT-2^[4]、83 亿参数规模的 Megatron^[5]、110 亿参数规模的 T5^[6]、170 亿参数规模的图灵自然语言生成 (Turing-NLG)^[7], 以及史无前例

的 1 750 亿参数规模的 GPT-3^[8]陆续出现。随着模型规模的增大,模型学习大规模数据的能力逐渐增强,并展现出强大的小样本和零样本学习的能力。

以上工作大多集中于纯文本的预训练,多模态预训练的工作规模较小,且主要针对理解类的任务。另外,还缺少成熟的中文多模态预训练模型。针对以上问题,达摩院联合阿里云机器学习平台及清华大学在 2021 年 1 月提出了首个中文领域的超大规模多模态预训练模型 M6^[9],并在 KDD2021 (2021 年数据挖掘顶会)发表了相应的论文。M6 在超大规模的中文多模态数据上做预训练,兼容多模态及单模态的理解与生成能力。M6 的下游任务包括视觉问答、视觉描述、跨模态检索、基于文本的图像生成和图像编辑、文本摘要、诗歌生成等,覆盖领域广。同时,针对超大模型训练效率低的问题,达摩院联合阿里云机器学习平台实现了基于图形处理器 (GPU) 的混合专家 (MoE) 机制的开发,这不仅是中国首个基于 MoE 的超大规模预训练的实践,也是全球首个基于 MoE 的、最大规模的多模态预训练模型。随后,达摩院针对 MoE 机制做了细致的分析和优化,提出专家分组机制^[10],并通过一系列的优化,提升了训练效率,降低了资源消耗,在 480 个 GPU 上实现了万亿参数规模的 M6 预训练。该 MoE 机制是多模态领域的首个万亿参数规模的预训练模型。相比于同为万亿参数规模的 MoE 模型 Switch Transformer^[11] (使用了 2 048 个 TPU),M6-T 更为低碳和高效 (仅使用 480 个 GPU)。考虑到不断增长的参数规模,为了实现更为绿色环保的模型训练,达摩院开始研究极限参数规模即十万亿参数的 M6 预训练^[12],提出共享解除的训练机制,实

现了 512 个 GPU 的十万亿参数规模 M6 模型的预训练,助力绿色环保的人工智能的发展。

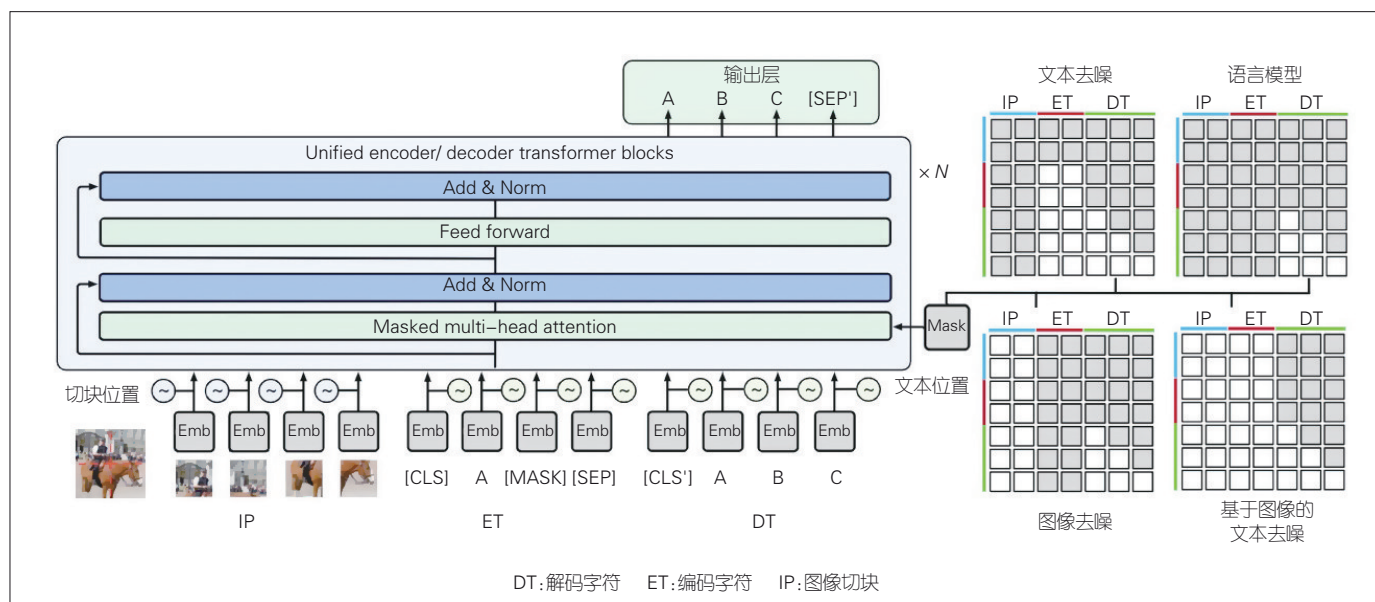
1.1 百亿参数规模 M6

M6 系列模型以自然语言预处理 (NLP) 和多模态领域中最为主流的 Transformer 架构为主体,针对不同模态的数据设计了特定的模块及预训练任务 (如图 1 所示)。在多模态预训练模型中,研究人员针对图像数据和文本数据的差异,对图像数据进行特征提取的预处理。不同于物体检测提取特征的方式,M6 采用了图像切块、backbone 模型 (如 ResNet) 等提取特征的方式,并根据块的位置提供位置表示。在主体架构上,M6 依然采用 Transformer block 堆叠的方式,具体架构如图 1 所示。

在预训练任务上,为了让模型兼具多模态及单模态的理解和生成能力,研究人员设计了有无图像信息条件的文本去噪和语言模型的任务。这样一来,模型通过学习便可根据上下文还原和续写文本,从而掌握跨模态的理解和生成能力,并可以便捷地迁移到多种类型的下游任务中。

传统的分布式训练因其有限的显存,无法支持百亿参数规模的模型训练。为了打破显存的限制,研究人员在数据并行的基础上,增加了重计算的机制,并采用优化器状态分片和梯度分片的策略,在单台机器 8 个 GPU 的条件下即可训练一个百亿参数规模的 M6 模型。

结合上述模型架构、训练任务及大模型训练机制,达摩院采集并处理了超过 2 TB 的中文图像及接近 300 GB 的中文文本数据,然后使用这份大规模数据对 M6 进行预训练,并



▲图1 M6 模型及训练任务示意图

将这些数据迁移到多种类型的下游任务中（包括视觉问答、视觉描述、跨模态检索、基于文本的图像生成和图像编辑、文本摘要、诗歌生成、故事生成、自然语言理解等）。实验结果证明，M6大模型在中文视觉问答和视觉描述中均能取得最优的效果。同时，M6大模型具备极强的小样本学习能力，在多个小样本任务评测上超出同期的中文预训练大模型（CPM）^[13]。

1.2 千亿参数规模 M6-MoE

在攻克了百亿参数规模 M6 的难关后，研究人员使用阿里巴巴自研框架 Whale，实现了专家并行机制，并将其和 M6 模型相结合，在 GPU 集群上训练出首个千亿参数规模的多模态预训练模型。具体而言，研究人员使用 Whale 框架的算子拆分功能，将多个专家网络分配到多个 GPU 上，并使用 all-to-all 通信机制实现输入信息的分配和聚合。在此基础上，研究人员添加了峰值显存优化、通信优化和混合精度优化的一系列策略，在 128 个 A100 上达到 1 440 个样本/s 的训练效率。实验结果表明，相较于百亿参数规模 M6，M6-MoE 模型的参数规模虽然增长了 10 倍，但训练效率依然高于 M6。M6-MoE 的语言模型困惑度评测具有较大优势。

1.3 万亿参数规模 M6-T

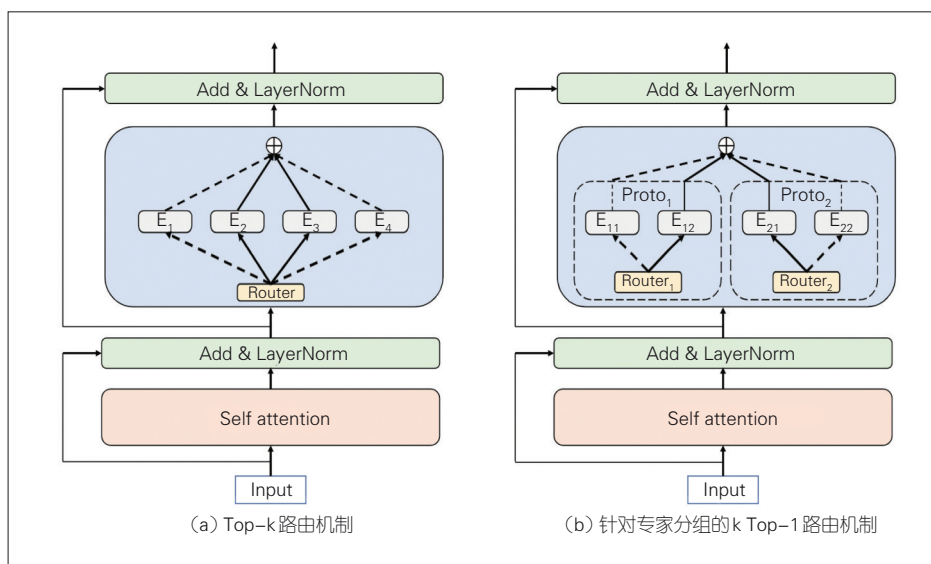
研究人员在研究混合专家机制与大模型训练的结合并分析其中的问题后发现，辅助损失的必要性比较低，并且其中的 top-k 路由机制中 k 的大小对模型效果具有决定性的影响。随着 k 的增大，模型性能逐渐提升，但同时也会出现边际效应递减及模型的训练效率显著下降的情况。针对以上问题，研究人员设计了专家分组机制^[10]（如图 2 所示），将专家网络分成多组后再以并行的方式对每组进行 top-k 路由。实验表明，分组机制能够显著提升模型的训练效率；当 k 值较大时，分组机制的模型表现也优于传统方法；在上下游的语言模型困惑度评测上，分组后的模型均显著优于传统方法。

研究人员将此方法应用于万亿参数规模的 M6-T 模型的训练，并优化了大模型中的显存占用，即在 480 个 V100-32G 上仅用约 3 天的时间便实现了万亿参数规模 M6-T 模型的预训练。实验表明，相较于基线模型，结合了专家分组机

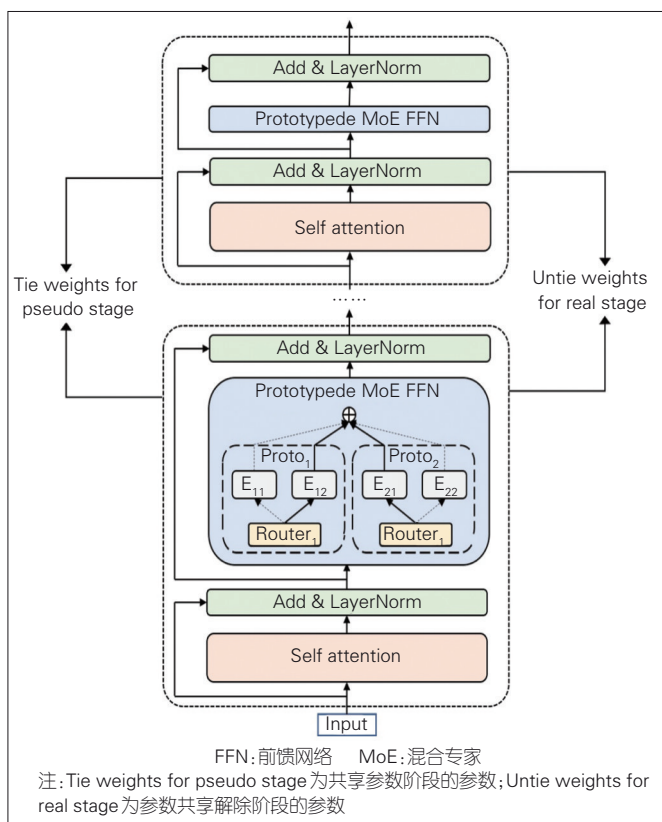
制的 M6-T 收敛速度显著加快，同时损失也更少。

1.4 十万亿参数规模 M6-10T

经过研究，达摩院提出了更加低碳的共享解除的训练机制（具体如图 3），并且设计了粒度可控的 CPU offload（中央处理器负载迁移）^[12]，成功地用 10 天左右的时间在 512 个



▲图2 专家分组机制示意图



▲图3 共享解除机制示意图

GPU上完成十万亿参数规模的M6-10T的预训练。

实验证明，M6-10T方案在收敛和下游迁移的过程中具有有效性。同时，它在十万亿参数规模的M6-10T模型上做出了成功实践，仅用10天左右的时间便取得了非常突出的收敛效果。十万亿模型达到相同预训练损失所需的样本量仅为万亿模型的40%，这充分显示出该机制的优势。

2 M6的产业化应用

超大规模多模态预训练取得的成功也意味着它将成为社会发展的重要基础设施，为各类下游任务提供支持。目前，针对互联网生态中各种复杂的业务场景，M6做出了相应的优化，在服装设计、自动文案、金融服务、搜索推荐等业务场景中实现商业落地，产生了巨大的商业价值以及社会价值。

此外，千亿参数和万亿参数M6大模型的研发，大力推动了低碳大模型的发展，并助力绿色环保AI的发展，响应了中国的碳中和战略部署。相比于传统方法，M6元生款数智制造结合犀牛环保面料的研发应用，能在全链路中减少30%以上的碳排放。每卖出一件元生款链路生产的服装，就能减排0.35 kg二氧化碳。也就是说，卖出50件就相当于种下一棵树。

2.1 AI图像生成在服饰制造等行业的应用

M6具有的一项重要能力是基于文本的图像生成和图像编辑。为了充分利用Transformer架构对大数据的高效处理和泛化能力，研究人员在文到图的生成架构上，选用了两阶段模型：第一阶段，需要将图像进行离散编码；而在第二阶段，则利用预训练模型M6来建模文本和图像离散编码的关系。为了进一步提升图像生成的清晰度和细节丰富度，研究人员将序列长度从1 024延长至4 096，并加入了稀疏注意力，成功将生成图像分辨率提升至1 024×1 024。在服饰制造行业，M6模型能够生成具有高清晰度和丰富细节的图片（如图4），并利用AI服饰设计的能力在服装制造行业实现落地。

为了进一步提升M6在图像生成过程中的可控性和效率，研究人员提出了基于M6的自非回归的图像生成模型^[14]，实现在不同控制条件下（包括文本、图像、风格等）的图像生成，使得模型具备了图像编辑的能力。

基于以上能力，M6以AI设计师的身份参与到服饰制造等行业中。对该行业传统企划链路来说，从样式规划到最终的生产上架，往往需要耗时半年。例如，一件冬季的羽绒服，需要在初夏时就决定其款式并开始漫长的人工设计和反

复的打样修改。2021年，达摩院智能计算实验室与阿里巴巴犀牛智造深度合作，借助数字化的能力，自动化地捕捉流行机会，并结合M6的生成能力，为商家提供敏捷高效的设计和丰富的体验。

目前，M6生成的服装图片已通过质检并达到商家的质量标准。首期文到图生成的人工质检优质率约为10%，比人工设计师的效率高10~20倍。研究人员会根据人工反馈的结果不断优化模型的生成质量，提升优质率。研究人员不断优化模特试衣算法，协助犀牛智造为商家提供更多的模特试穿效果图。

为了进一步验证并应用M6的超强图像生成以及创新能力，达摩院和某车厂合作产出概念车型以及未来车型，以辅助汽车设计师进行车型设计，并和阿里云LOGO服务团队合作产出种类丰富的LOGO配图供客户挑选（如图5所示）；另外，还和蚂蚁花呗团队合作生成宠物头像图作为宠物唯一身份认证，通过少样本或者文本描述即可生成符合要求的图像。

2.2 工业级文案生成

M6的另一个重要能力便是文本生成。该能力能够运用于视觉描述、视觉问答、文本摘要、问答、对话、文案创作等。目前M6的文本生成已经达到工业级标准。在训练语料较少的情况下，M6的优势更加明显，仅以此前5%的数据再

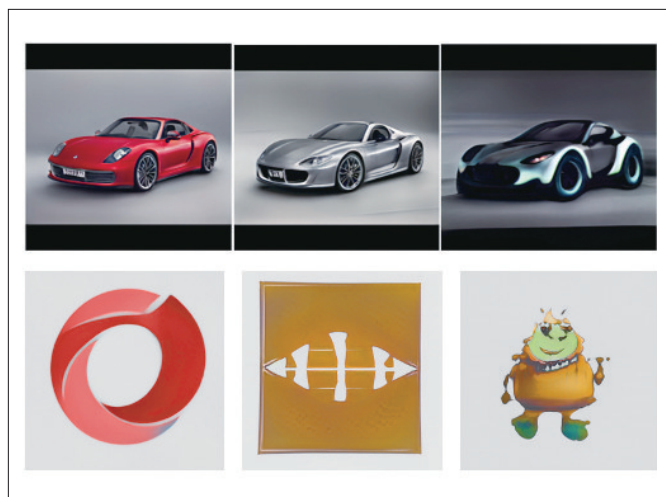


▲图4 羊羔绒毛衣生成示例

通过微调便可达到更优的效果，人工审核通过率最高可达85%。目前M6的文本生成已经成功应用于UC畅聊中台、手机淘宝和支付宝Push文案、阿里小蜜智能客服、阿里小蜜虚拟直播、支付宝腰封文案中。在UC畅聊平台中，算法团队利用畅聊评论中台沉淀的历史用户回复数据对M6模型进行微调，批量构造机器人自动回复，并通过M6模型每天的自动总结和生成推荐理由，覆盖社会、娱乐、体育、汽车等10个内容类目，引发用户回复率显著提升（约10%）。

虚拟直播中，在不足1 000条标注数据的条件下，M6结合非自回归生成方法（SANA）^[15]实现了在双语评估替换（BLEU）评测上超过10%的大幅提升，人工审核通过率超70%，大大提升了智能剧本的生成效率（如图6所示）。

在支付宝的腰封文案场景中，M6利用其生成能力为该业务产出主副标题。相较于基线模型，M6产出的文案在线上实现了超5%的点击率。



▲图5 概念车LOGO配图示例

ASR原始文本	M6输出
那宝宝们用我们这款燕麦爽肤水真的是特别的合适啊。在我们77号链接这款燕麦爽肤水呢，它里面是顾名思义就是蕴含了一个燕麦成分，对不对？我们这个燕麦成分呢，它不是一般的燕麦，它是北美进口的微米级别的燕麦添加在里面，然后形成了一个燕麦黄铜的成分。燕麦黄铜呢是可以有效的去强韧到我们肌肤的角质层，然后呢也可以给我们肌肤建立一个隐形的保护膜，这个保护膜呢是可以有效的隔绝住外界环境啊，天气对我们肌肤带来的一个损伤。然后也可以把补充到基底的一个水分养分牢牢的去锁起来啊，宝宝们，对，而且我们这个基底呢也是。可也是可以是给我们基底带来一个密集性的保湿补水锁水的一个功效，让我们肌肤的每一个角落喝饱水。因为它里面呢是蕴含了一个透明质酸成分，再加角鲨烷的一个成分，一起打到底底扩散出来，给肌肤做到一个密集性的保湿补水啊。	宝宝们这款燕麦爽肤水真的是特别的合适。这款燕麦爽肤水呢它里面是顾名思义就是蕴含了一个燕麦成分，我们这个燕麦成分呢，它不是一般的燕麦，它是北美进口的微米级别的燕麦添加在里面，然后形成了一个燕麦黄铜的成分。燕麦黄铜呢是可以有效的去强韧到我们肌肤的角质层，然后呢也可以给我们肌肤建立一个隐形的保护膜，这个保护膜呢是可以有效的隔绝住外界环境天气对我们肌肤带来的一个损伤。然后也可以把补充到基底的一个水分养分牢牢的去锁起来啊，而且我们这个基底也是给我们基底带来一个密集性的保湿补水锁水的一个功效，让我们肌肤的每一个角落喝饱水。因为它里面呢是蕴含了一个透明质酸成分，再加角鲨烷的一个成分，一起打到底底扩散出来，给肌肤做到一个密集性的保湿补水啊。

▲图6 生成剧本示例

2.3 表示提取功能助力搜索推荐等应用

在支付宝搜索场景中，结合经典的双塔结构，M6在场景语料上进行微调，然后应用于向量召回中。更加关注语义和知识信息的预训练模型能够有效提升检索的准确率。超大规模模型强大的小样本学习和理解能力则助力冷启场景和长尾查询词的召回，从而实现了搜索页点击率（PVCTR）的显著提升。

在更加依赖内容表示的内容导购轻应用中，M6的作用更加明显。在家居场景中，相比于传统的图像单模态表示，M6的多模态表示包含更加丰富的信息，在召回和排序场景均能实现效果的提升，在风格一致性上的表现尤其突出。这体现了多模态表示学习的特点。

在搜索推荐以外的场景中，达摩院联合斑马汽车，使用M6来提升车辆检索准确率。在该场景中，系统需要根据用户指令检索出最相关的车辆，而M6提供的用户指令表示能够更加准确地反映用户意图，帮助下游的车辆检索模型指代最相关车辆。这能够将准确率提升5%。

3 M6平台化服务

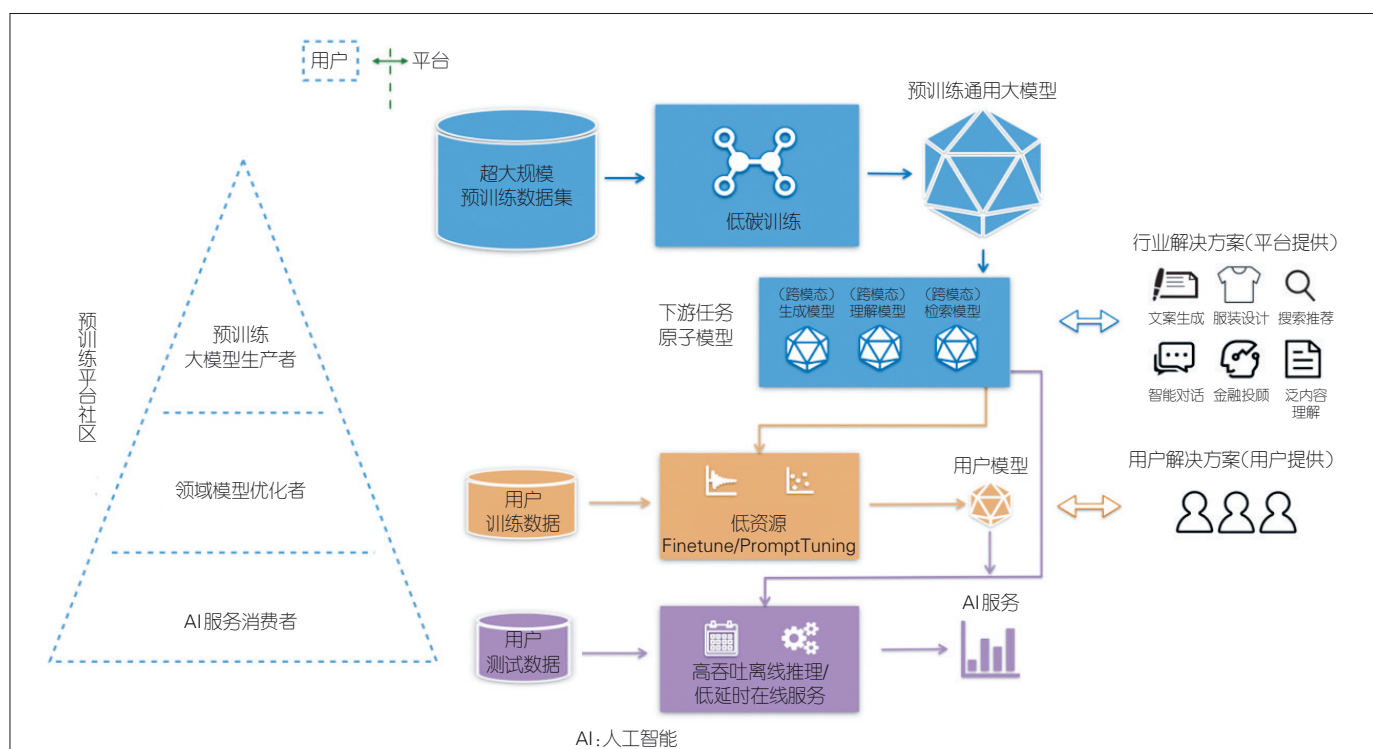
超大规模多模态预训练的应用意味着大模型开始实现对各行各业的支持。为了让大模型更加简便易用，达摩院推出了M6预训练平台，以服务化的形式将大模型的能力运用到各行各业中。凭借平台的易用性、训练高效低碳、下游任务覆盖广泛等特点，M6服务化平台服务了阿里巴巴集团内部各类业务。同时，M6预训练平台也已经通过阿里云对外发布。

M6多模态预训练服务化平台具有3个核心能力点：

- 任务形式覆盖广。该平台是目前下游任务覆盖最广泛的预训练平台，覆盖多模态输入输出的常见任务超20个。
- 高性能&简单易用。无须关注分布式训练、数据输入输出、数据并行、底层实现以及训练评估流程等细节，准备好输入数据，再对参数进行简单修改，即可实现多机多卡的训练或推理任务。
- 下游任务内源+支持自定义模型改造。用户可以在M6的上层实现自定义模型，无须关注过多大模型细节。

M6平台集成了多种微调形式，并提供了高效&低碳的分布式训练、低延迟的模型服务、统一的数据&模型管理、一键式模型部署方案。用户可以根据自身数据情况，选择使用软件开发工具包（SDK）调用、微调、自定义模型等方式来灵活支持自身的下游任务。M6平台的整体框架如图7所示。

相较于传统小作坊式的AI服务，集中式的数据、算力



▲图7 平台整体架构示意图

开发模式有着更好的平台粘性，因此我们希望能将内部的成功经验以服务化平台的方式让更多外部用户获益。同时，基于阿里云强大的基础设施，M6预训练平台能提供更完善的全链路服务，实现AI普惠化目标。

4 结束语

自2020年以来，达摩院深入研究超大规模多模态预训练关键技术，在2021年陆续发布百亿、千亿、万亿、十万亿参数规模的M6模型。并且，达摩院持续发力，解决大模型应用落地难的问题，并专注于大模型的产业化落地，在服饰制造、工业级文案生产、搜索推荐等场景实现了应用。在此基础上，达摩院将M6的能力以服务化的形式集成到M6预训练平台中并对外发布，帮助行业、企业和个体快速使用大模型，推动普惠AI发展。

未来，大模型在产业领域的应用将更加丰富。前景主要包括：推动传统产业智能化转型，催生基于智能模型的新产业，以及改变人类社会的生产和管理模式等。预训练大模型还有亟待突破的几个难题：

(1) 目前的主流实践是先通过训练大模型得到参数规模大、精度高的模型，再基于下游任务数据，通过剪枝、微调的方法将模型的体积压缩，在基本不损失精度的情况下减轻部署压力。目前，业界还没找到通用的、直接训练小型模型

就能得到较满意精度的办法。

(2) 训练千亿、万亿模型动辄需要上千个GPU卡，这对大模型的推广和普惠带来了很大的挑战。

(3) 因为参数量大，目前预训练模型主要针对大量非结构化数据。如何与知识等结构化数据进行结合，更加有效地进行认知推理，也是一个非常大的挑战。

以上难题使得大模型参数竞赛进入“冷静期”，而大小模型在云边端协同进化则带来了新的突破可能性。云边端协同使小模型更容易获取通用的知识与能力。小模型专注于在特定场景做极致优化，从而获得性能与效率的提升。大小模型的协同进化可以更好地服务于更加复杂的新场景，例如元宇宙、数字人等。同时，该体系更有利于保护用户数据隐私。

参考文献

- [1] KAPLAN J, MCCANDLISH S, HENIGHAN T, et al. Scaling laws for neural language models [EB/OL]. (2019-05-10) [2022-01-12]. <https://paperswithcode.com/paper/mutual-information-scaling-and-expressive>
- [2] DEVLIN J, CHANG M, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [EB/OL]. [2022-01-10]. <https://>

- paperswithcode.com/paper/bert-pre-training-of-deep-bidirectional
- [3] RADFORD A, NARASIMHAN K. Improving language understanding by generative pre-training [EB/OL]. [2022-01-12]. <https://paperswithcode.com/paper/improving-language-understanding-by-generative-pre-training>
- [4] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners [EB/OL]. [2022-01-10]. <https://paperswithcode.com/paper/language-models-are-unsupervised-multitask>
- [5] SHOHEYBI M, PATWARY M, PURI R, et al. Megatron-LM: training multi-billion parameter language models using GPU model parallelism [EB/OL]. [2022-01-10]. https://www.researchgate.net/publication/335908286_Megatron-LM_Training_Multi-Billion_Parameter_Language_Models_Using_GPU_Model_Parallelism
- [6] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer [EB/OL]. [2022-01-10]. <https://paperswithcode.com/paper/exploring-the-limits-of-transfer-learning-with-a-unified-text-to-text-transformer>
- [7] ROSSET C. Turing-NLG: a 17-billion-parameter language model by Microsoft [EB/OL]. [2020-02-13] [2022-01-10]. <https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/>
- [8] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners [EB/OL]. [2022-01-10]. <https://paperswithcode.com/paper/language-models-are-few-shot-learners>
- [9] LIN J Y, MEN R, YANG A, et al. M6: a Chinese multimodal pretrainer [EB/OL]. [2021-03-01] [2022-01-12]. <https://paperswithcode.com/paper/m6-a-chinese-multimodal-pretrainer>
- [10] YANG A, LIN J, MEN R, et al. Exploring sparse expert models and beyond [EB/OL]. [2021-03-31] [2022-01-12]. <https://arxiv.org/abs/2105.15082v2>
- [11] FEDUS W, ZOPH B, SHAZEER N, et al. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity [EB/OL]. [2021-11-11] [2022-01-12]. <https://arxiv.org/abs/2101.03961>
- [12] LIN J, YANG A, BAI J, et al. M6-10T: a sharing-delinking paradigm for efficient multi-trillion parameter pretraining [EB/OL]. [2021-10-08] [2022-01-12]. <https://paperswithcode.com/paper/m6-10t-a-sharing-delinking-paradigm-for-efficient-multi-trillion-parameter-pretraining>
- [13] ZHANG Z, HAN X, ZHOU H, et al. CPM: a large-scale generative Chinese pre-trained language model [EB/OL]. [2021-10-08] [2022-01-12]. <https://paperswithcode.com/paper/cpm-a-large-scale-generative-chinese-pre-trained-language-model>
- [14] ZHANG Z, MA J, ZHOU C, et al. M6-UFC: unifying multi-modal controls for conditional image synthesis [EB/OL]. [2021-10-08] [2022-01-12]. <https://paperswithcode.com/paper/ufc-bert-unifying-multi-modal-controls-for-conditional-image-synthesis>
- [15] WANG P, LIN J, YANG A, et al. Sketch and refine: towards faithful and informative table-to-text generation [EB/OL]. [2021-10-08] [2022-01-12]. <https://paperswithcode.com/paper/sketch-and-refine-towards-faithful-and-informative-table-to-text-generation>

作者简介



林俊阳，阿里巴巴达摩院智能计算实验室算法专家；主要研究领域包括自然语言处理及多模态表征学习（侧重于多模态预训练及其应用），并参与研究大模型的低破训练、提示学习、轻量化应用等问题；参与研究的全球最大规模的十万亿参数多模态预训练模型M6广泛应用于下游应用（如图像生成、自然语言生成、跨模态检索等）；发表论文30余篇。



周畅，阿里巴巴达摩院高级算法专家，并担任NeurIPS/ICML/KDD/WWW等会议PC member；主要研究领域包括表征学习、推荐系统、多模态预训练；带领团队研发了全球最大参数规模的多模态预训练模型M6，大幅优化了大规模预训练模型的计算效率，取得了中国首个大模型商业化落地成果，并作为算法负责人研发了大规模GNN训练平台AliGraph，获得中国电子学会科学技术进步奖一等奖；发表文章30余篇。



杨红霞，阿里巴巴达摩院人工智能科学家，曾担任IBM全球研发中心Watson研究员、雅虎首席数据科学家等；主导阿里巴巴下一代人工智能突破性技术——认知智能技术的发展与场景应用落地；曾获2019世界人工智能大会最高奖卓越人工智能引领者（Super AI Leader，简称SAIL奖）、2020年国家科学技术进步奖二等奖、杭州市领军型创新团队、2021年电子学会科学技术进步奖一等奖；发表文章80余篇，获得中国和美国专利近20项。

高效训练百万亿参数预训练模型的系统挑战和对策



Challenges and Measures for Efficient Training of Trillion-Parameter Pre-Trained Models

马子轩/MA Zixuan, 翟季冬/ZHAI Jidong, 韩文弢/HAN Wentao, 陈文光/CHEN Wenguang, 郑伟民/ZHENG Weimin

(清华大学, 中国 北京 100083)
(Tsinghua University, Beijing 100083, China)

DOI: 10.12142/ZTETJ.202202008

网络出版地址: <https://kns.cnki.net/kcms/detail/34.1228.tn.20220418.1725.002.html>

网络出版日期: 2022-04-19

收稿日期: 2022-02-20

摘要: 随着预训练模型规模的急剧增长, 训练此类模型需要海量的计算和存储能力。为此, 本工作在新一代国产高性能计算机上训练了一个174万亿参数的超大规模预训练模型, 模型参数量可与人脑中的突触数量相媲美。重点讨论在训练这一超大规模预训练模型中遇到的几个关键系统挑战: 如何选取高效并行策略, 如何进行高效数据存储, 如何选取合适的数据精度, 以及如何实现动态负载均衡, 并总结了针对上述挑战的一些解决方法。

关键词: 人工智能; 超级计算机; 混合专家; 异构系统

Abstract: As the size of pre-trained artificial intelligence models grows dramatically each year, training such models requires massive computing and memory capabilities. To this end, an unprecedentedly large-scale pre-trained model with 174 trillion parameters on an entire supercomputer is proposed, which rivals the number of synapses in a human brain. The key challenges encountered in such large-scale model training, including deciding efficient parallel strategy, performing efficient data storage, deciding appropriate data precision, and dynamic load balancing are proposed. Then the solutions to the above challenges are summarized.

Keywords: artificial intelligence; supercomputer; mixture of experts; heterogeneous architecture

1 大规模预训练模型的发展背景

近年来, 深度学习在计算机视觉 (CV)、自然语言处理 (NLP)、推荐系统、决策模型等各个领域都发挥了重要作用。同时, 大规模预训练模型技术已经成为深度学习领域最先进的技术, 并在多个领域的下游应用中表现出优秀的性能。近年来, 谷歌的推荐系统、搜索引擎, 阿里巴巴的推荐系统、图像生成等任务均采用了预训练模型。而在预训练模型方面, 学术界已达成共识: 模型规模和模型准确度有明显的正相关关系^[1-6]。表1展示了近年来预训练模型的发展趋势: 从最早的1.1亿参数量的第1代预训练GPT发展到最新的万亿参数量的GPT-3、Switch Transformer等模型。探索更大参数量的模型具有重要的科学意义。

从计算的角度看, 随着模型规模的增长, 训练模型的数据会变多。此时, 单机已经无法满足大规模模型训练的计算需求。在模型训练的过程中, 计算分为3个步骤: 前向、反向、更新。在前向过程中, 模型使用输入数据和参数进行计

算, 得到预测结果并和标注数据进行对比, 从而计算出该次预测的损失; 在反向过程中, 模型对损失进行传播, 计算所有参数的梯度; 在更新阶段, 模型再使用计算出的梯度来更新参数。该过程在训练中不断迭代, 最终达到模型的收敛状态。当模型规模较大时, 单机串行执行已无法满足模型的计算需求。这时, 我们需要使用一些并行策略来辅助训练, 如数据并行和模型并行。

顾名思义, 使用数据并行策略时, 模型对输入的数据进

▼表1 近年发布的预训练模型

预训练模型	模型参数量	时间/年
GPT ^[7]	1.1亿	2018
BERT-Large ^[2]	3.4亿	2018
GPT-2 ^[4]	15亿	2019
GShard ^[8]	6 000亿	2020
GPT-3 ^[1]	1 750亿	2020
Switch Transformer ^[9]	1.6万亿	2021
BaGuaLu	174万亿	2021

行拆分。如图1所示，数据被划分到不同的节点后再进行计算，每个节点上都保存一份完整的模型。模型在正向和反向的过程中，均不会进行通信；而在正向结束计算梯度以及更新时，则会引入全局的All-Reduce通信。

使用模型并行策略时，模型对模型参数本身进行拆分，这可以理解成对参数矩阵进行切分。在执行过程中，模型执行被切分的计算时，会根据切分方式的不同引入不同的通信方式。如图1所示，在两种不同的模型并行执行模式中，不同的切分策略产生了两种不同的通信模式：All-Reduce和All-Gather。

无论是数据并行还是模型并行，在执行过程中，都需要通过对数据进行切分来减少每个节点的计算量和存储量，从而达到并行训练的目的。在模型训练中，参数量和训练速度是两个重要的衡量指标。经验证明，越大规模的模型所需的训练样本数就越多^[10]。因此，大模型训练需要同时扩展参数量和训练吞吐量。在传统的并行模式中，通过对输入进行切分，数据并行可以有效扩展训练吞吐量。而由于数据并行需要在每个节点上都存储一份参数，这种并行模式不能扩展参数量。模型并行可以对参数进行切分，从而有效扩展参数量。切分后多个机器使用同一组输入数据，因此不能有效扩展训练吞吐量。传统的大模型训练通常使用混合数据并行和模型并行的方法进行训练^[11]。

混合专家（MoE）模型是近年来一种新的预训练模型^[8-9]。这种模型将一个模型切分为多个小模型（专家）。在训练过程中，数据先通过一个小规模网络（一般称之为网关）进行路由，再选择适合的一部分专家进行计算，最后加权求和得到输出。

对MoE模型而言，一个整体的大模型由多个小模型构成。这样一来，这类大模型的参数量可以和传统的稠密大模型一样大或更大。同时，由于每次运算只是从离散模型中稀疏地选取一部分来激活，MoE模型的整体计算量与稠密小模型相当。因此，MoE模型可以在扩展模型参数、增大模型

规模的同时，在相同的训练时间内，达到优于稠密大模型的准确度。谷歌最新的Switch Transformer^[9]就采用了MoE的架构。

在MoE大模型中，并行训练会引入两种划分：数据划分和专家划分。在训练过程中，每一组数据需要选择合适的专家，在不同节点间交换数据。这个过程会引入全局的All-to-All通信。MoE的并行训练很好地满足了大模型训练对吞吐量和模型规模的需求。在MoE并行训练中，不同机器存储不同的专家，有效扩展了模型的参数规模。同时，由于不同机器选取不同的输入，该方法也有效扩展了模型训练的吞吐量。因此，MoE模型非常适合对大模型进行扩展。

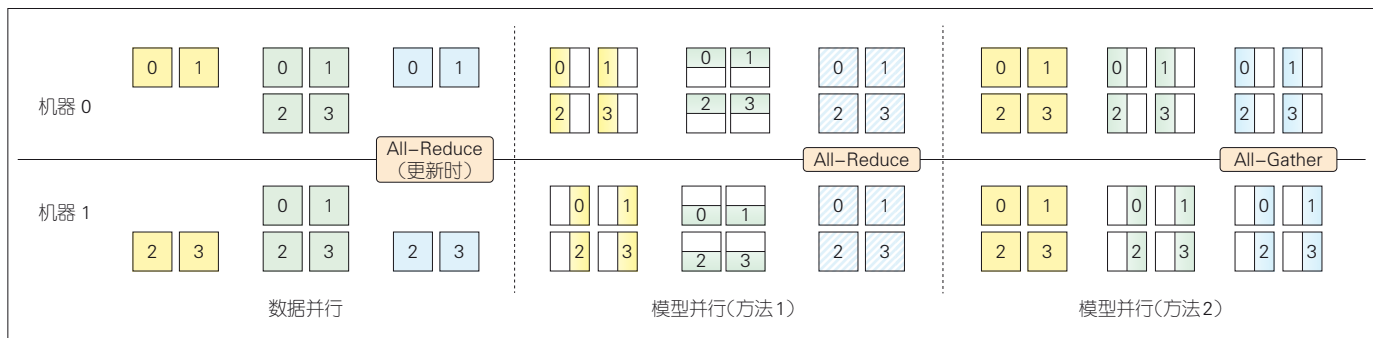
2 大规模预训练模型面临的挑战

当模型扩展到超大规模时，会遇到多方面的挑战。例如，如何选取高效的并行策略，如何进行高效数据存储，如何选择合适数据精度以及如何实现动态负载均衡。

2.1 如何选取高效的并行策略

在大模型训练中，有多种并行策略可供选择，如数据并行、模型并行、流水线并行以及MoE并行等各种并行模式。这些并行策略将大模型训练按照不同的方式进行划分，从而使任务可以很好地在多机上执行。而不同的划分方法，对底层有不同的计算与通信需求。例如，一类并行策略的通信需求虽然小，但可能会引入更多的计算；另一类并行策略虽然不引入额外计算，但可能会引入更多的通信量。在训练过程中，如何选择合适的并行策略是一个重要的挑战。

以16个节点为例，这些节点在训练过程中需要被切分的部分包括输入数据与模型参数。如果使用简单的数据并行，那么模型参数并不被划分，每台机器上都有一份完整复制的模型。那么，数据并行的核心是把输入数据简单地切分成16份。与此类似，模型并行把模型切成16份。如果是MoE并行，则有16个专家，每个节点1个专家。在此基础



▲图1 数据并行与模型并行示意图

上,不同并行策略之间还可以进行混合,例如数据并行加模型并行的混合方式。

这些策略的组合给并行设计带来了巨大的选择空间。在此基础上,在很多系统中,网络拓扑会带来不同节点间通信性能的不同。例如,在胖树结构中,通常会存在网络裁剪的问题,这导致一部分节点之间的通信延迟更高、带宽更低。这种差异导致相同并行策略在不同情况下的性能有所不同,从而使并行策略选择问题更加复杂。因此,如何选择一个高效的并行策略具有很大的挑战性。

2.2 如何高效存储和划分数据

以万亿参数量的模型为例,如果模型精度是32位,模型参数本身就会占用4 TB的空间。与此同时,模型的梯度需要占用4 TB的空间,优化器的更新参数需要占用8 TB的空间,计算中间值需要约8 TB的空间。为了保证模型的训练性能,这些数据需要存储在内存中。如果使用NVIDIA Tesla V100训练该模型,仅在显存中存储这些数据就需要768块显卡。在这种情况下,不同的划分方式对底层的计算和通信也会产生不同的影响。因此,用高效划分数据来支持高效训练同样非常具有挑战性。

2.3 如何选取合适的存储精度

在计算过程中,使用更低的精度(如16位)训练时,模型需要的内存量更小且性能更好,但准确度有所下降,训练轮数有所增加;使用更高精度(如32位或64位)训练时,模型需要的内存量更大,计算的要求更高,但模型准确度高,训练轮数少。因此,如何在模型中选择合适的精度进行计算,以平衡模型精度与模型训练时间,给模型扩展带来了

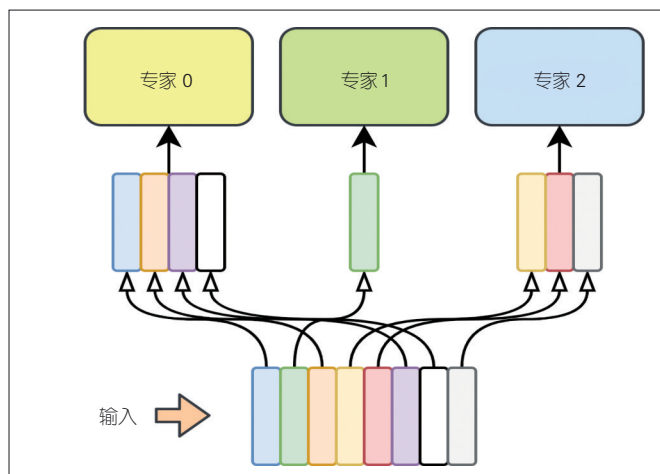
新的挑战。

2.4 如何实现动态负载均衡

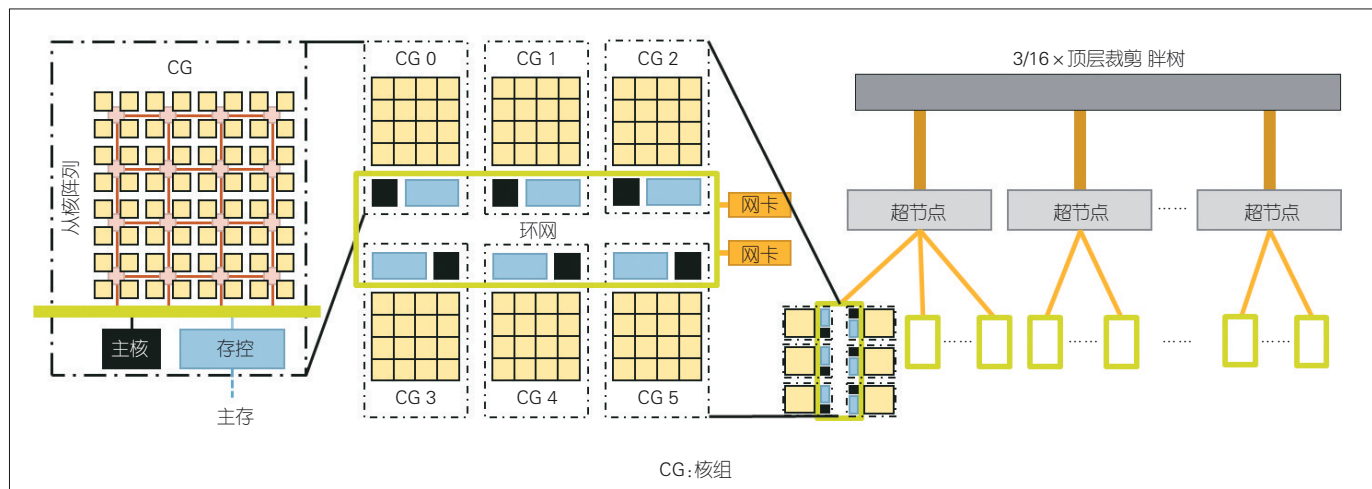
MoE模型会通过网关对不同的输入样本选择不同的专家进行计算。在实际运行过程中,样本的选择存在严重的负载不均衡问题。例如,在语言处理中,常用语言元素出现的频率高,对应的专家负载更高,如图2所示的专家0。此时,如果训练规模非常大,部分节点的高负载就会降低整体系统的执行效率。因此,如何动态改善负载均衡,以提高模型的训练效率成为大模型训练的重要挑战。

3 新一代国产超级计算机

新一代国产神威超级计算机^[13]采用新一代神威处理器芯片——申威26010-Pro,其架构如图3所示。其中,每个节点包括6个核组(CG),每个核组包含1个主核(计算控制



▲图2 混合专家模型中的负载均衡问题



▲图3 新一代国产超级计算机架构

核心)和64个从核(计算核心CPE)。加速计算主要使用从核阵列。如果使用中央处理器-图形处理器(CPU-GPU)系统来做类比,我们可以认为主核相当于其中的CPU,从核阵列相当于GPU。神威处理器将6个核组通过1个环形网络连接起来,并封装到1个芯片中。同时,每个节点上包含2个自主研发的高速网络芯片,所有的节点通过网络再进行连接。神威网络的架构是一个双层胖树^[12],最下面的一层把256个节点组织成一个超节点。超节点的内部节点间完全互连;超节点之间通过顶层网络互联,跨超节点的通信需要经过顶层网络。

出于成本考虑,两层的网络拓扑会带来网络裁剪。当通信跨越超节点时,带宽会降低到超节点内通信带宽的3/16。正如前文所述,因为训练对数据和模型的需求不同,将并行模式扩展到超大规模的异构系统时,需要考虑如何实现高效的映射。

大模型对内存计算和通信都有非常大的需求,新一代的超级计算机有PB级的内存空间、强大的计算能力、自主可控的高速网络以及灵活的通信接口。因此,新一代的国产超级计算机为大规模模型训练奠定了基础。

4 在国产超算平台上训练大模型

根据国产超算平台的特点,我们将大模型训练扩展到了千万核心规模。我们的工作核心在于验证当模型做到一定规模时,并行加速还面临哪些挑战,以及如何应对这些挑战。当前,我们的训练参数规模达到了174万亿,已达到人脑神经元突触数量的规模。针对前文所述的4个挑战,我们提出了一些解决方案。尽管这些策略并不一定是最优的,但仍可以为接下来进一步的优化扩展提供参考。

4.1 根据通信带宽采用合适的并行策略

新一代神威超级计算机的网络拓扑分为两层,其中上层是3/16的裁剪网络。此时,采用单一的并行策略,如数据并行或模型并行,会带来全局通信。受限于顶层带宽和参与通信的高进程数,全局通信带宽性能相比于小规模(如一个超节点以内)有显著的下降。那么,如何选择合适的并行策略来规避这类问题呢?

本质上,该问题属于高性能计算的基本问题。我们将应用程序的通信模式与底层的网络拓扑相结合,构建了一个性能模型,并通过性能模型判断并行策略在大规模下的执行效率。最终,如图4所示,我们在全机规模下的解决方案是在超节点内(256个节点)做数据并行,在跨超节点做MoE并行。相比于对称的方案(即超节点内做MoE并行,超节点

间做数据并行),该方案可以将性能提高1.6倍。

4.2 实现高效不重复的参数存储和划分

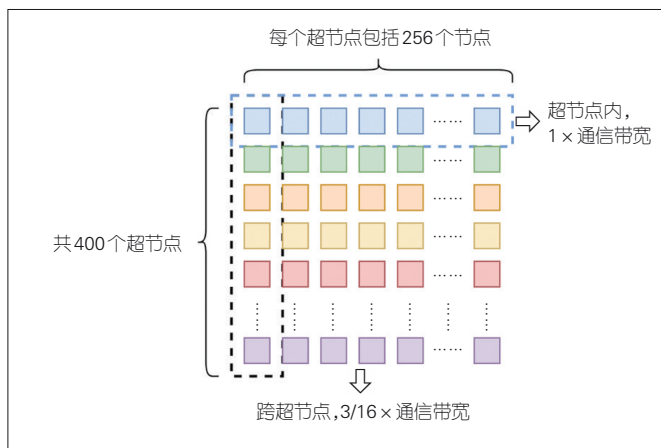
前文提到的数据存储问题,在本质上为如何将参数进行高效划分。在训练过程中,参数主要分为3部分:模型参数、梯度参数和更新参数。其中,模型参数指训练需要的计算参数;梯度参数指反向传播时每个模型参数的梯度;更新参数指优化器所需的参数,例如Adam优化器^[14]中需要的参数。这些参数的规模都和模型参数的大小相当。此外,在混合精度训练中,我们还需要考虑混合精度策略需要存储的额外参数,如主参数与主梯度。因此,如何对这些参数做切分是一个复杂的问题。

我们以APEX O2^[15]的混合精度训练为例。在训练过程中,模型的更新参数会占用75%的空间。如果简单地采用数据并行,模型参数在参与数据并行的节点中将被重复存储,浪费了宝贵的存储资源。微软于2019年提出的工作ZerO^[16-17]就对数据并行的参数存储进行了优化,其核心思想是把参数分布在不同节点来进行更新,并在不影响通信量的前提下,将参数分布式存储到不同机器中。如图5所示,分布式参数更新有效降低了模型参数的内存。

我们对ZerO的方法进行了深度扩展。在裁剪网络中,采用分层训练策略,并针对全局和部分通信,扩展了通信算法,从而在不影响通信量的前提下,在分层并行策略中实现了分布式参数更新。

4.3 设计合适的混合精度策略

在新一代神威超级计算机中,神威CPU同时支持双精度和半精度计算。单精度计算使用双精度计算模拟,因此半精度性能为单精度的4倍。在机器学习训练中,广泛使用的



▲图4 全机规模并行策略设计

是单精度与半精度类型。此时,如何针对国产超级计算机设计合适的混合精度策略,成为影响性能优化的重要因素。

以 Multibox SSD^[18]模型为例,对该模型参数分布的分析^[19]如图6所示,横坐标表示参数数值的指数分布。红色区域表示FP16可以表示的范围。如果该模型的参数直接采用半精度存储,那么大部分参数会失真或无法表示。因此,我们采用了混合精度训练中常用的动态损失缩放技术^[20]:在正向计算结束后对损失进行缩放,即乘以一个系数,然后在反向计算结束后再进行反缩放,并动态调整缩放系数,保证在不发生上溢出的情况下,数值尽量大。理想情况下,缩放后的表示范围如图6中绿色区域所示。此时,大部分参数都可以被半精度表示。

NVIDIA的APEX^[15]系统是针对NVIDIA GPU的混合精度框架。APEX提供了4种训练策略,分别是O0(单精度训练)、O1(只优化计算)、O2(优化除更新外的部分)、O3

(全部优化)。在NVIDIA平台中,使用较多的是O2策略。我们将APEX的策略移植到了神威平台。遗憾的是,这4种策略并不能直接在神威平台的大模型训练中使用。测试结果表明:O0和O1策略可以正常收敛,但性能差;O2、O3性能好,但不能收敛。

因此,我们提出了一种分层的混合精度策略,如图7所示。通过对不同层进行小规模验证,我们发现不同层对精度的要求不同。例如,训练中前馈网络(FFN)层和注意力层对精度不敏感,可以使用半精度计算、单精度更新;而其他层对精度敏感,因此可以使用单精度训练。这种策略在保证收敛的前提下,可以获得较高的训练性能。

此外,混合精度训练还包含更多的探索方向,例如,训练时间对训练精度的敏感性等。另外,在大的科学计算程序中,不同迭代使用不同精度。在迭代的开始、中间和最后,迭代都可以选择不同的混合精度策略。以动态的方式使计算

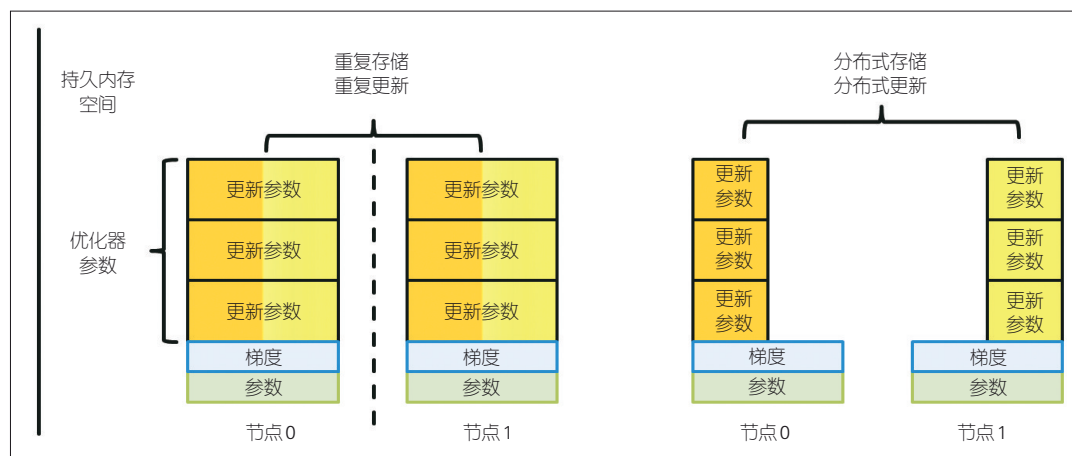
和内存达到最高效率也是一个比较前沿的研究方向。

4.4 实现高效均衡负载

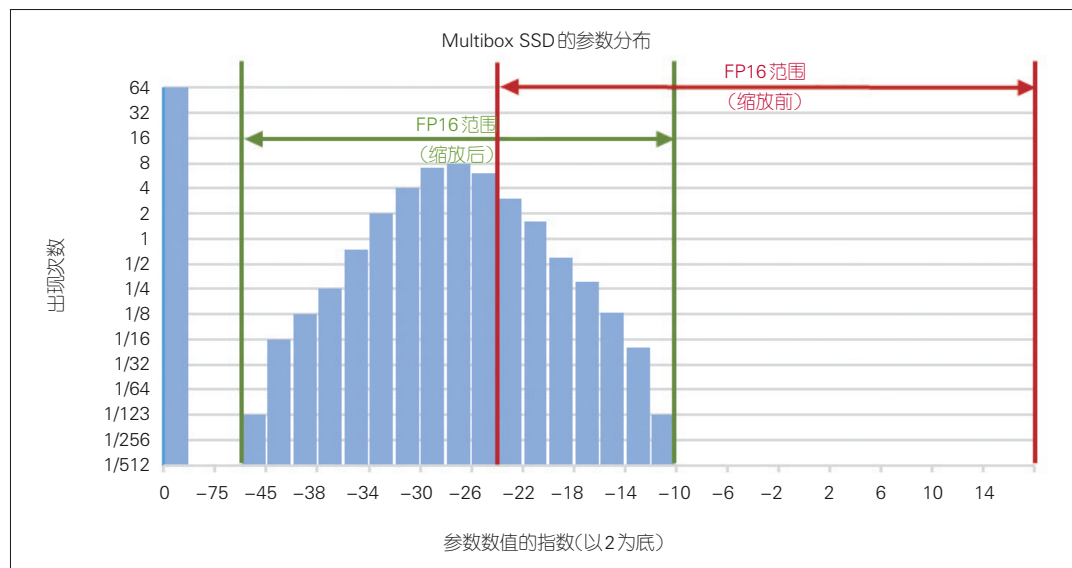
为了解决MoE模型负载不均衡的问题,我们提出了一个算法SWIPE。算法的核心思想是当一个专家非常受欢迎,且样本数多于均值时,就将多余样本分配给其他专家,以保证全局的绝对均衡。这个策略虽然简单直接,但验证结果显示模型训练能够有效收敛。

4.5 国产系统底层优化

此外,为了适配新一代神威超级计算机,我们针对神威平台的软件系统、算子库、计算框架和基础设施等进行了大量的优化工作。例如,针对神威系统动态模式下内存分配的性能问题,



▲图5 分布式参数更新方法



▲图6 Multibox SSD模型的参数分布示意

根据机器学习应用的特点，我们设计了高效内存分配器 SWAlloc；针对神威平台，实现了一套高性能算子库 SW-Tensor；同时深度重构了 SWPyTorch，以支持各类自定义算子表示。该部分工作包含约 60 000 行 C/C++ 代码，实现了 100 余个定制算子，有效地支持了模型训练。

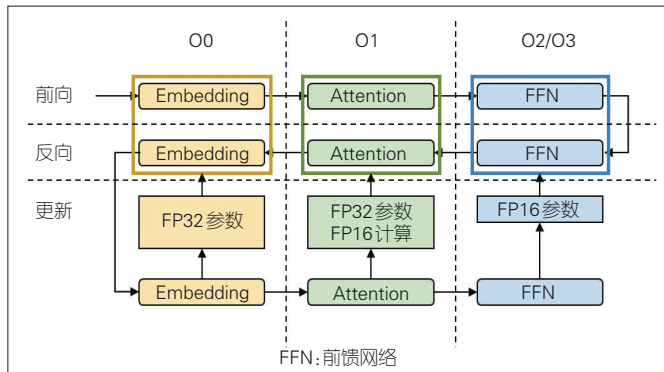
5 实验结果

最终，我们在新一代神威平台上整合了上述优化方法，促成了百万亿参数模型的训练。

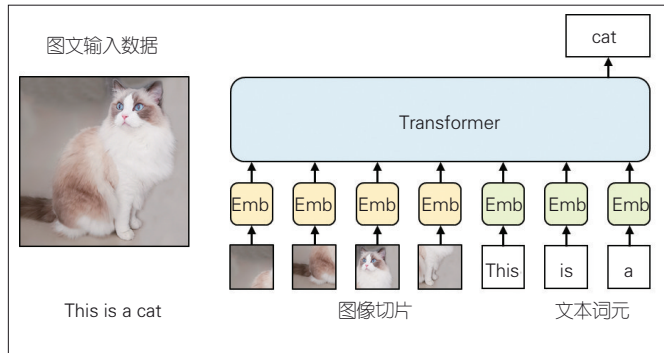
我们的模型基于阿里巴巴的中文预训练处理模型 M6^[21]，模型输入包含文本和图像数据。如图 8 所示，模型将输入图像拆分为多个块，并使用 ResNet^[22] 等预训练模型来提取特征，然后将图像特征与词向量连接起来形成一个序列，再由 Transformer 模型处理并生成高级表示。

该模型在中文最大的多模态数据集 M6-Corpus^[21] 上进行训练。数据来源包括百科全书、电子商务平台和其他类型的网页。最终处理的数据集的详细统计情况如表 2 所示，其中，其中，图像部分提取特征后大小为 16 TB。

我们共测试了 3 个不同的参数量模型的性能，模型超参数如表 3 所示。其中， d_{model} 表示模型的隐藏层规模， d_{ff} 表示 FFN 内部层规模。在训练的 3 个模型中，MoDa-174T 规模达到了百万亿参数量（我们获得的最大规模模型）。



▲图 7 分层的混合精度策略



▲图 8 模型结构示意图

▼表 2 预训练数据集规模

数据集	文本规模/GB	图片规模/TB
百科	15.0	0.1
网页	70.0	1.5
电子商务	12.2	0.3
总计	97.2	1.9

▼表 3 实验用到的测试模型超参数

模型	参数量/ 万亿	层数	头数	d_{model}	d_{ff}	专家数
MoDa-1.93T	1.93	12	8	4 096	$4\,096 \times 12$	400
MoDa-14.5T	14.50	10	8	4 096	$4\,096 \times 18$	2 400
MoDa-174T	173.90	3	8	4 096	$4\,096 \times 18$	96 000

训练过程与性能实验均在新一代神威超级计算机上进行。其中，通过对 PyTorch 的前向、反向和更新时间的检测，可以得到时间数据。浮点运算次数（FLOPs）通过神威系统的性能计数器得到，主核和从核分别计数，累加结果作为总浮点运算次数。

性能测试结果如表 4 所示。在 1.93 万亿参数规模时，模型训练的计算性能达到了 1.18 EFLOPS（百亿亿次浮点计算每秒）；当模型增大至 14.5 万亿时，混合精度性能达到 1.00 EFLOPS；当模型扩展到 174 万亿时，由于模型规模太大，模型训练无法进行混合并行，并行策略只有一维，因此通信性能会有明显下降，只达到 0.230 EFLOPS。

图 9 和图 10 分别展示了 MoDa-1.93T 模型的弱扩展性与强扩展性测试结果。在实验中，我们将问题规模定义为专家个数和样本个数。在弱扩展性测试中，两者随测试规模的变化发生变化，即在测试规模为全系统的 $1/n$ 的测试中，模型规模和样本个数均为全系统测试的 $1/n$ 。因此，每个计算节点处理的样本数和模型规模均保持不变。测试结果表明，我们的系统可以做到接近线性的可扩展性。

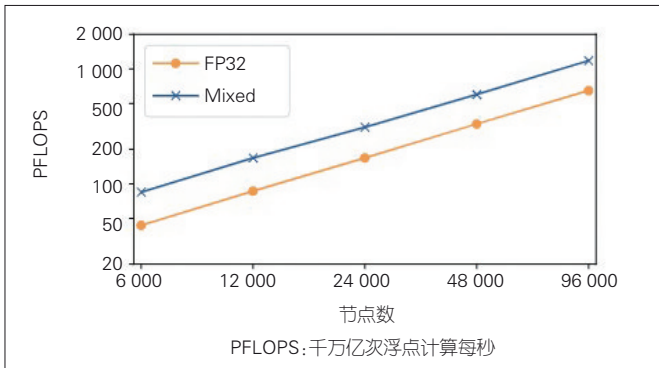
而在强扩展性测试中，我们固定专家个数和样本个数。此时，在测试规模为全系统的 $1/n$ 时，每个节点处理的样本为全系统测试中每个节点的 n 倍。测试结果表明，从全系统 $1/16$ 扩展到整机时，性能提升约为 12 倍。我们的系统表现出很好的扩展性。

为了验证系统的正确性，我们将 MoDa-1.93T 模型训练了 500 步，收敛曲线如图 11 所示。损失在 500 步后下降到 3.46，根据参考文献[21]，我们认为训练接近收敛，并且可以验证系统的正确性。

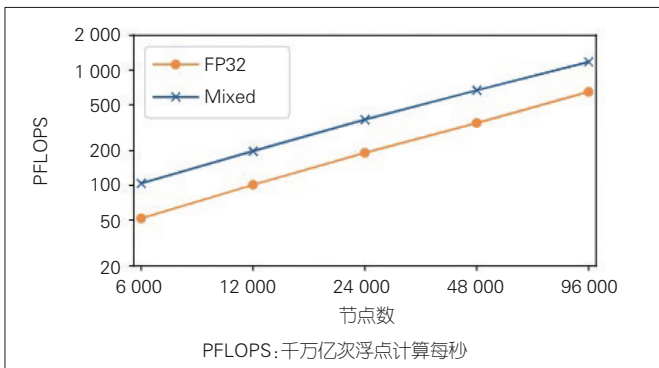
▼表4 性能测试结果

模型规模 (参数量)/万亿	单精度性能/ PFLOPS	单精度时间 (每轮迭代)/s	混合精度性能/ EFLOPS	混合精度时间 (每轮迭代)/s
1.93	647	14.50	1.18	7.78
14.50	525	18.70	1.00	10.20
174	198	13.10	0.23	10.80

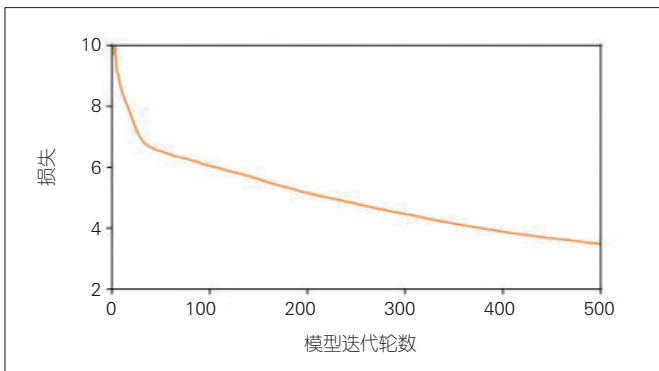
PFLOPS: 千万亿次浮点计算每秒 EFLOPS: 百亿亿次浮点计算每秒



▲图9 弱扩展性测试结果



▲图10 强扩展性测试结果



▲图11 MoDa-1.93T收敛曲线

6 结束语

我们把目前的工作命名为“八卦炉”^[23]。“八卦炉”是一个高性能计算（HPC）和人工智能（AI）结合得较好的例子。在系统方面，结合以上提到的优化策略，我们进行了预训练模型加速。在此过程中，我们发现HPC存在许多挑战，

例如网络裁剪等。本研究为将来探索新的大模型训练系统提供了宝贵的经验。

总的来说，该项研究主要针对国产系统的大型模型加速训练，并结合研究中遇到的问题提出了一系列解决方案。目前，这些方案仍处于研究阶段，希望有更多学者能一起参与讨论。

参考文献

- [1] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners [EB/OL]. [2022-01-10]. <https://paperswithcode.com/paper/language-models-are-few-shot-learners>
- [2] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding [EB/OL]. [2022-01-10]. <https://paperswithcode.com/paper/bert-pre-training-of-deep-bidirectional>
- [3] LIU Y H, OTT M, GOYAL N, et al. RoBERTa: a robustly optimized BERT pretraining approach [EB/OL]. [2022-01-10]. <https://paperswithcode.com/paper/roberta-a-robustly-optimized-bert-pretraining>
- [4] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners [EB/OL]. [2022-01-10]. <https://paperswithcode.com/paper/language-models-are-unsupervised-multitask>
- [5] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer [EB/OL]. [2022-01-10]. <https://paperswithcode.com/paper/exploring-the-limits-of-transfer-learning>
- [6] YANG Z L, DAI Z H, YANG Y M, et al. XLNet: generalized autoregressive pretraining for language understanding [EB/OL]. [2022-01-10]. <https://paperswithcode.com/paper/xlnet-generalized-autoregressive-pretraining>
- [7] RADFORD A, NARASIMHAN K. Improving language understanding by generative pre-training [EB/OL]. [2022-01-10]. <https://paperswithcode.com/paper/improving-language-understanding-by>
- [8] LEPIKHIN D, LEE H, XU Y Z, et al. GShard: scaling giant models with conditional computation and automatic sharding [EB/OL]. [2022-01-10]. <https://paperswithcode.com/paper/gshard-scaling-giant-models-with-conditional>
- [9] FEDUS W, ZOPH B, SHAZEER N M. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity [EB/OL]. [2022-01-10]. <https://paperswithcode.com/paper/switch-transformers-scaling-to-trillion>
- [10] KAPLAN J, MCCANDLISH S, HENIGHAN T, et al. Scaling laws for neural language models [EB/OL]. [2022-01-10]. <https://paperswithcode.com/paper/mutual-information-scaling-and-expressive>
- [11] SHOEYBI M, PATWARY M, PURI R, et al. Megatron-LM: training multi-billion parameter language models using GPU model parallelism [EB/OL]. (2019-09-17)[2022-01-10]. <https://arxiv.org/abs/1909.08053v2>
- [12] LEISERSON C E. Fat-trees: universal networks for hardware-efficient supercomputing [J]. IEEE transactions on computers, 1985, 100(10): 892-901. DOI:10.1109/TC.1985.6312192
- [13] FU H, LIAO J, YANG J, et al. The Sunway TaihuLight supercomputer: system and applications [J]. Science China information sciences, 2016, 59(7): 1-16
- [14] KINGMA D P, BA J. Adam: a method for stochastic optimization [EB/OL].

- (2019-09-17)[2022-01-10]. <https://paperswithcode.com/paper/adam-a-method-for-stochastic-optimization>
- [15] NVIDIA. Apex (A PyTorch Extension) [EB/OL]. [2022-01-10]. <https://nvidia.github.io/apex/>
- [16] RAJBHANDARI S, RASLEY J, RUWASE O, et al. ZeRO: memory optimizations toward training trillion parameter models [C]//Proceedings of SC20: International Conference for High Performance Computing, Networking, Storage and Analysis. IEEE, 2020: 1-16. DOI: 10.1109/SC41405.2020.00024
- [17] RASLEY J, RAJBHANDARI S, RUWASE O, et al. DeepSpeed: system optimizations enable training deep learning models with over 100 billion parameters [C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2020: 3505-3506. DOI: 10.1145/3394486.3406703
- [18] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot MultiBox detector [EB/OL]. [2022-01-10]. <https://paperswithcode.com/paper/ssd-single-shot-multibox-detector>
- [19] NVIDIA. Training with mixed precision [EB/OL]. [2021-01-10]. <https://docs.nvidia.com/deeplearning/performance/mixed-precision-training/index.html>
- [20] ZHAO R Z, VOGEL B, AHMED T. Adaptive loss scaling for mixed precision training [EB/OL]. [2022-01-10]. <https://paperswithcode.com/paper/adaptive-loss-scaling-for-mixed-precision>
- [21] LIN J Y, MEN R, YANG A, et al. M6: a Chinese multimodal pretrainer [EB/OL]. [2022-01-10]. <https://paperswithcode.com/paper/m6-a-chinese-multimodal-pretrainer>
- [22] HE K M, ZHANG X Y, REN S Q, et al. Identity mappings in deep residual networks [EB/OL]. [2022-01-10]. <https://paperswithcode.com/paper/identity-mappings-in-deep-residual-networks>
- [23] MA Z, HE J, QIU J, et al. BaGuaLu: targeting brain scale pretrained models with over 37 million cores [C]//Proceedings of the 27th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming. ACM, 2022: 192-204

作者简介



马子轩，清华大学计算机科学与技术系在读硕士研究生；主要研究方向为高性能计算、编译优化。



翟季冬，清华大学计算机科学与技术系长聘副教授、博士生导师，担任多个国际学术期刊编委；主要研究方向为高性能计算、性能评测和编译优化等；获中国计算机学会科学技术奖自然科学奖一等奖、《IEEE TPDS》杰出编委奖、CCF-IEEE CS 青年科学家奖等。



韩文强，清华大学计算机科学与技术系讲师，担任中国计算机学会 NOI 科学委员会副主席、中国科协“英才计划”计算机学科工作委员会委员等职务；主要研究方向为并行与分布式计算，特别是大数据处理系统和机器学习系统；发表论文 10 余篇。



陈文光，清华大学计算机科学与技术系教授，现为中国计算机学会会士、杰出讲者、副秘书长、青年科技论坛荣誉委员，并担任 ACM 中国理事会常务理事、北京计算机学会副理事长等职务；主要研究方向为操作系统、程序设计语言与并行计算；获国家科技进步奖二等奖 1 次，部级科技一等奖 2 次。



郑伟民，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。

自然语言处理技术发展



Development of Natural Language Processing Technology

王海宁/WANG Haining

(英特尔(中国)有限公司, 中国 北京 100013)
(Intel China Ltd., Beijing 100013, China)

DOI: 10.12142/ZTETJ.202202009

网络出版地址: <https://kns.cnki.net/kcms/detail/34.1228.TN.20220408.1420.004.html>

网络出版日期: 2022-04-08

收稿日期: 2022-02-26

摘要: 基于神经网络和深度学习的预训练语言模型为自然语言处理技术带来了突破性发展。基于自注意力机制的Transformer模型是预训练语言模型的基础。GPT、BERT、XLNet等大规模预训练语言模型均基于Transformer模型进行堆叠和优化。认为目前依赖强大算力和海量数据的大规模预训练语言模型存在实用问题, 指出轻量预训练语言模型是未来重要的发展方向。

关键词: 自然语言处理; 预训练语言模型; Transformer; GPT; BERT; XLNet; 模型优化

Abstract: The pre-trained language model based on neural network and deep learning has brought breakthrough development for natural language processing technology. The Transformer model based on self-attention mechanism is the basis of the pre-trained language model. Large-scale pre-trained language models such as GPT, BERT, XLNet, etc. are based on the Transformer model or its optimization. However, the current large-scale pre-training language models that rely on powerful computing resources and massive data have practical problems. It is pointed out that lightweight pre-trained language models are an important development direction in the future.

Keywords: natural language processing; pre-trained language model; Transformer; GPT; BERT; XLNet; model optimization

自然语言处理(NLP)是基于自然语言理解和自然语言生成的信息处理技术^[1]。这里的自然语言是指任何一种人类语言, 例如中文、英语、西班牙语等, 并不包括形式语言(如Java、Fortran、C++等)。

自然语言处理的历史可以追溯到17世纪。那时莱布尼茨等哲学家对跨越不同语言的通用字符进行探索^[2], 认为人类思想可以被归约为基于通用字符的运算。虽然这一观点在当时还只是理论上的, 但却为自然语言处理技术的发展奠定了基础。

作为人工智能的一个重要领域, 当代自然语言处理技术与人工智能技术的兴起和发展是一致的。1950年, 图灵提出了著名的基于人机对话衡量机器智能程度的图灵测试^[3]。这不仅是人工智能领域的开端, 也被普遍认为是自然语言处理技术的开端。20世纪50年代至90年代, 早期自然语言处理领域的发展主要基于规则和专家系统, 即通过专家从语言学角度分析自然语言的结构规则, 来达到处理自然语言的目的。

从20世纪90年代起, 伴随着计算机运算速度、存储容量的快速发展, 以及统计学习方法的成熟, 研究人员开始使用统计机器学习方法来处理自然语言任务。然而, 此时自然语言的特征提取仍然依赖人工, 同时受限于各领域经验知识的积累。

深度学习算法于2006年被提出之后, 不仅在图像识别领域取得了惊人的成绩, 也在自然语言处理领域得到了广泛应用。不同于图像的标注, 自然语言的标注领域众多并具有很强的主观性。因此, 自然语言处理领域不容易获得足够多的标注数据, 难以满足深度学习模型训练对大规模标注数据的需求。

近年来, GPT^[4]、BERT^[5]等预训练语言模型可以很好地解决上述问题。基于预训练语言模型的方法本质上是一种迁移学习方法, 即通过在容易获取、无需人工标注的大规模文本数据基础上依靠强大算力进行预先训练, 来获得通用的语言模型和表示形式, 然后在目标自然语言处理任务上结合任务语料对预训练得到的模型进行微调, 从而在各种下游自然语言处理任务中快速收敛以提升准确率。因此, 预训练语言模型自面世以来就得到了迅速发展和广泛应用, 并成为当前各类自然语言处理任务的核心技术。

1 语言表示的发展

自然语言处理涉及众多任务。从流水线的角度上看, 我们可以将这些任务划分为3类: 完成自然语言处理之前的语言学知识建设和语料库准备任务; 对语料库开展分词、词性标注、句法分析、语义分析等基本处理任务; 利用自然语言处理结果完成特定目标的应用任务, 如信息抽取、情感分

析、机器翻译、对话系统、意图识别等。其中，将自然语言转变为计算机可以存储和处理的形式（即文本的表示）是后续各类下游自然语言处理任务的基础和关键。

字符串是最基本的文本表示方式，即符号表示。这种表示方式主要应用在早期基于规则的自然语言处理方式中。例如，基于预定义的规则对句子进行情感分析：当出现褒义词时，句子表达正向情感；当出现贬义词时，句子表达负向情感。显然，这种使用规则的方式只能对简单的语言进行分析处理，在遇到矛盾的情况下系统很可能无法给出正确的结论。

以向量的形式表示词语，即词向量，是广泛应用于目前自然语言处理技术中的表示方式。词向量的表示有多种方式。其中，最简单的是基于词出现次数统计的独热表示和词袋表示。这类表示方式的主要缺点在于，不同的词需要用完全不同的向量来表示，维度高并且缺乏语义信息的关联，同时存在数据稀疏问题。

另外一大类词向量表示是基于分布式语义假设（上下文相似的词，其语义也相似）的分布式表示。这种词向量表示具体又可以分为3类：

（1）基于矩阵的词向量表示。该方法基于词共现频次构建体现词与上下文关系的（词-上下文）矩阵。矩阵每行表示一个词向量 w_i 。第 j 个元素 w_{ij} 的取值可以是 w_i 与上下文的共现次数，也可以由基于其共现概率进行的点互信息（PMI）、词频-逆文档频率（TF-IDF）、奇异值分解（SVD）等数学处理来获得。这种方法更好地体现了高阶语义相关性，可解决高频词误导计算等问题。其中，上下文可以是整个文档，也可以是每个词。此外，我们也可以选取 w_i 附近的 N 个词作为一个 N 元词窗口。

（2）基于聚类的词向量表示。这类方法通过聚类手段构建词与上下文之间的关系。例如，布朗聚类是一种基于 N -gram模型和马尔可夫链模型的自底向上的分层聚类算法。在这种算法中，每个词都在且仅在唯一的一个类中。在初始的时候，每个词均被独立分成一类，然后系统将其中的两类进行合并，使得合并之后的评价函数（用以评估 n 个连续的词序列能否组成一句话的概率）达到最大值。系统将不断重复上述过程，直至获得期望的类数量为止。

（3）基于神经网络的语言模型，也称为词嵌入表示。这类方法将词向量中的元素值作为模型参数，采用神经网络结合训练数据学习的方式来获得语言模型参数值。基于神经网络的语言模型具体又包括静态语言模型和动态语言模型。这两种语言模型的区别在于：静态语言模型通过一个给定的语料库得到固定的表示，不随上下文的变化而变化，例如

Word2vec、GloVe和FastText模型；动态语言模型由上下文计算得到，并且随上下文的变化而变化，例如CoVe、ELMo、GPT和BERT模型。其中，基于神经网络的语言模型充分利用了文本天然的有序性和词共现信息的优势，无需人工标注也能够通过自监督学习从文本中获取语义表示信息，是预训练语言模型的重要基础，也是目前词表示研究与应用的热点。

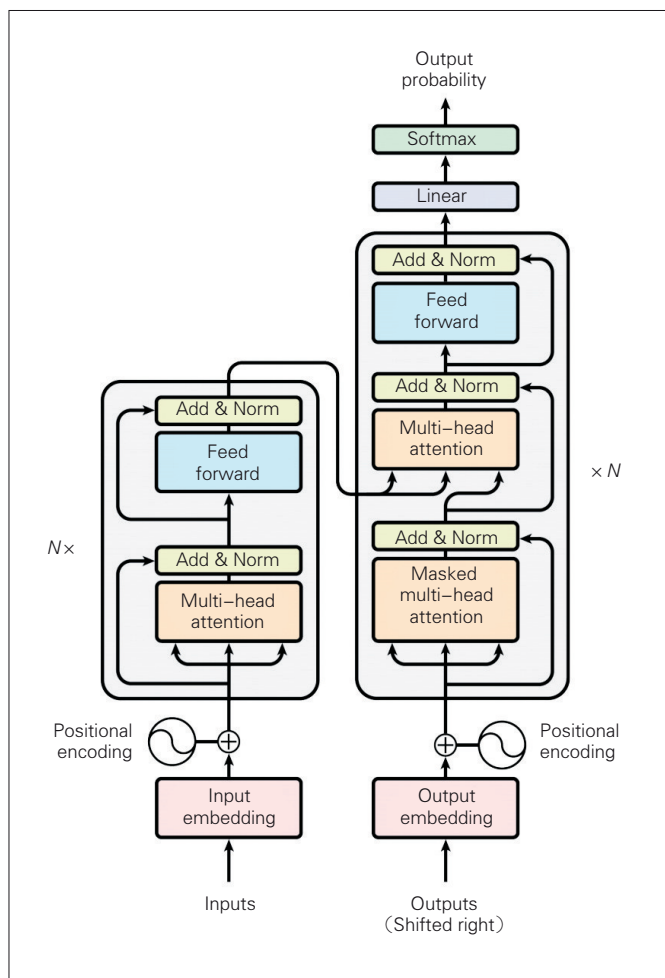
2 预训练语言模型

2.1 预训练语言模型基础

2003年，Y. BENGIO首次提出神经网络语言模型^[6]。2017年之前，在进行自然语言处理时人们常用多层感知机（MLP）、卷积神经网络（CNN）和循环神经网络（RNN），包括长短期记忆（LSTM）网络，来构建神经网络语言模型。由于每层都使用全连接方式，MLP难以捕捉局部信息。CNN采用一个或多个卷积核依次对局部输入序列进行卷积处理，可以比较好地提取局部特征。由于适用于高并发场景，较大规模的CNN模型经过训练后可以提取更多的局部特征。然而，CNN却难以捕获远距离特征。RNN将当前时刻网络隐含层的输入作为下一时刻的输入。每个时刻的输入经过层次递归后均对最终输出产生影响，这就像网络有了历史记忆一样。RNN可以解决时序问题和序列到序列问题，但是这种按照时序来处理输入的方式使得RNN很难充分利用并行算力来加速训练。LSTM是一种特殊的RNN，它对隐含层进行跨越连接，减少了网络的层数，从而更容易被优化。

2017年，来自谷歌的几位工程师在不使用传统CNN、RNN等模型的情况下，完全采用基于自注意力机制的Transformer模型，取得了非常好的效果^[7]。在解决序列到序列问题的过程中，他们不仅考虑前一个时刻的影响，还考虑目标输出与输入句子中哪些词更相关，并对输入信息进行加权处理，从而突出重要特征对输出的影响。这种对强相关性的关注就是注意力机制。Transformer模型是一个基于多头自注意力机制的基础模型，不依赖顺序建模就可以充分利用并行算力处理。在构建大模型时，Transformer模型在训练速度和长距离建模方面都优于传统的神经网络模型。因此，近年来流行的GPT、BERT等若干超大规模预训练语言模型基本上都是基于Transformer模型构建的。Transformer模型整体架构如图1所示。

自注意力机制的本质是学习序列中的上下文相关程度和深层语义信息。然而，随着输入序列长度的增加，学习效率会降低。为了更好地处理长文本序列，Transformer模型又衍



▲图1 Transformer 模型架构^[7]

生出一些“变种”，例如 Transformer-XL^[8]。Transformer-XL 采用段级循环和相对位置编码的优化策略，将 Transformer 中固定长度的输入片段进一步联系起来，具备更强的长文本处

理能力。

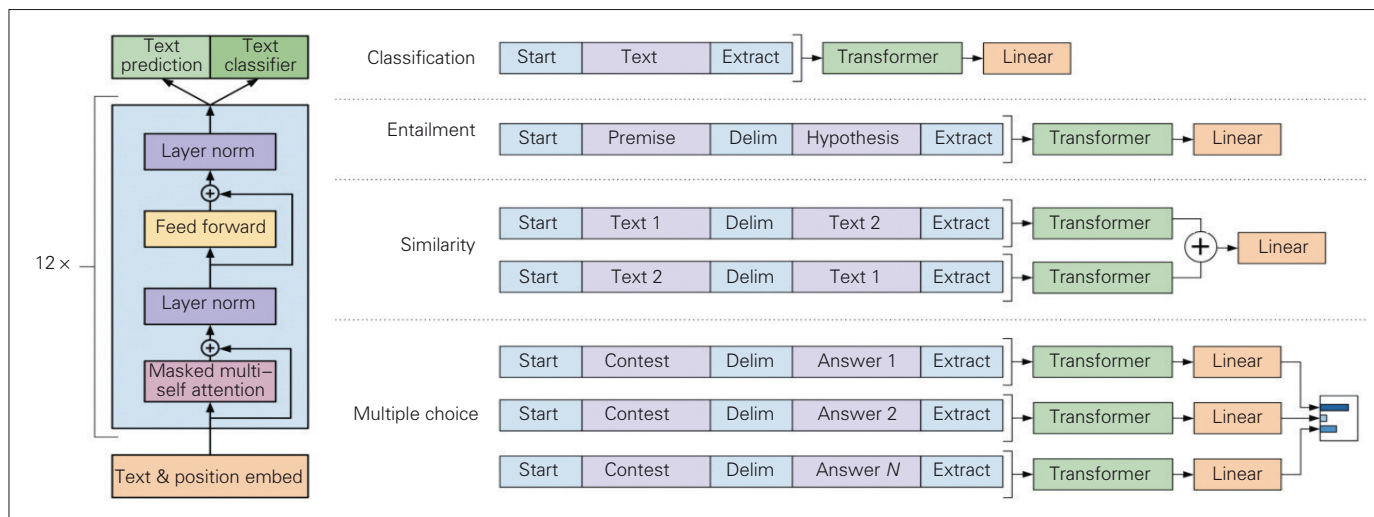
2.2 大规模预训练语言模型

广义预训练语言模型泛指经过提前训练得到的语言模型。各类神经网络语言模型在理论上都可以做预训练处理。而目前自然语言处理领域常涉及的预训练语言模型，通常是指一些参数数量过亿甚至超千亿的大规模语言模型。这些模型的训练依赖强大算力和海量数据。典型的大规模预训练语言模型包括 GPT 系列、BERT、XLNet 等。此外，这些模型的各种改进模型也层出不穷。

2.2.1 GPT 系列

2018年6月，OpenAI 公司提出初代 GPT 模型^[4]，开启了具有“基于大量文本学习高容量语言模型”和“对不同任务使用标注数据来进行微调”两个阶段的自然语言处理预训练模型大门。GPT 模型基于 12 层 Transformer 基础模型构建了单向解码器，约有 1.17 亿个参数。具体解码器结构、训练目标和针对不同下游任务的输入转换如图 2 所示。

OpenAI 公司在 2019 年 2 月进一步提出 GPT 模型的升级版本，即 GPT-2^[9]。由于担心该技术可能会被恶意利用，研究团队并没有对外发布预训练好的 GPT-2 模型，而是发布了一个小规模模型。GPT-2 保留了 GPT 的网络结构，直接进行规模扩张，即堆叠更多层的 Transformer 模型，并使用 10 倍于 GPT 模型的数据集进行训练，参数数量超过 15 亿。随着规模的增加，GPT-2 也获得了更好的泛化功能，包括生成前所未有的高质量合成文本功能。虽然在部分下游任务上尚未超过当时的最优水平，但是 GPT-2 证明了大规模预训练词向量模型在迁移到下游任务时，可以超越使用特定领域数



▲图2 Transformer 解码器结构和训练目标(左)及针对不同下游任务的输入转换(右)^[4]

据进行训练的语言模型，并且在拥有大量（未标注）数据和具备足够算力时，使下游任务受益于无监督学习技术。

GPT-3^[10]模型于2020年5月被提出，是目前最强大的预训练语言模型之一。GPT-3在GPT-2的基础上进一步进行了规模扩张，使用高达45 TB的数据进行训练，参数数量高达1 750亿。正是这样巨大的网络规模，才使得GPT-3模型在不进行任何微调的情况下，可以仅利用小样本甚至零样本就能在众多下游任务中超越其他模型。OpenAI公司虽未开源GPT-3模型，但是提供了多种应用程序接口（API）服务以供下游任务调用。

2.2.2 BERT

BERT^[5]是由谷歌公司于2018年10月提出的。与单向的GPT模型不同，BERT基于Transformer模型构建了多层双向编码器。

BERT模型包括两个训练任务：一个是掩码语言模型（MLM），另一个是下一句预测（NSP）。MLM可以很好地解决双向建模时逆序信息泄露的问题；NSP则可以很好地理解两段文本之间的关系，适用于完成阅读理解或文本蕴含类任务。BERT的每个下游任务都采用相同的预训练模型架构并使用预训练模型的参数来进行初始化。BERT的预训练和微调过程如图3所示。

BERT的设计团队按照模型规模的大小将BERT分为含有1.1亿个参数的BERT_{BASE}和含有3.4亿个参数的BERT_{LARGE}，并与其他模型（包括GPT）进行对比。对比结果表明，BERT模型在GLUE^[11]、SQuAD^[12]、SWAG^[13]的11项NLP任务评估中全面刷新了最佳成绩纪录，甚至在SQuAD测试中超越了

人类。

BERT模型是近年来NLP领域发展的一大里程碑。BERT陆续衍生出了许多优化的模型。例如，显著增强了长文本理解能力的XLNet^[14]、占用更少存储空间 of ALBERT^[15]、具备更强大文本生成能力的BART^[16]、能够学习视频知识的VideoBERT^[17]等。这些模型推动了NLP的快速发展。

2.2.3 XLNet

由于在预训练的输入数据中人为地引入了掩码，BERT模型忽略了被掩码信息之间的依赖性。这将导致预训练数据与微调阶段使用的真实数据之间产生微小差异。针对上述问题，卡内基梅隆大学和谷歌公司于2019年6月进一步提出了一种基于Transformer-XL的自回归语言模型，即XLNet模型^[14]。

通过置换语言建模（PLM），XLNet对序列中输入信息

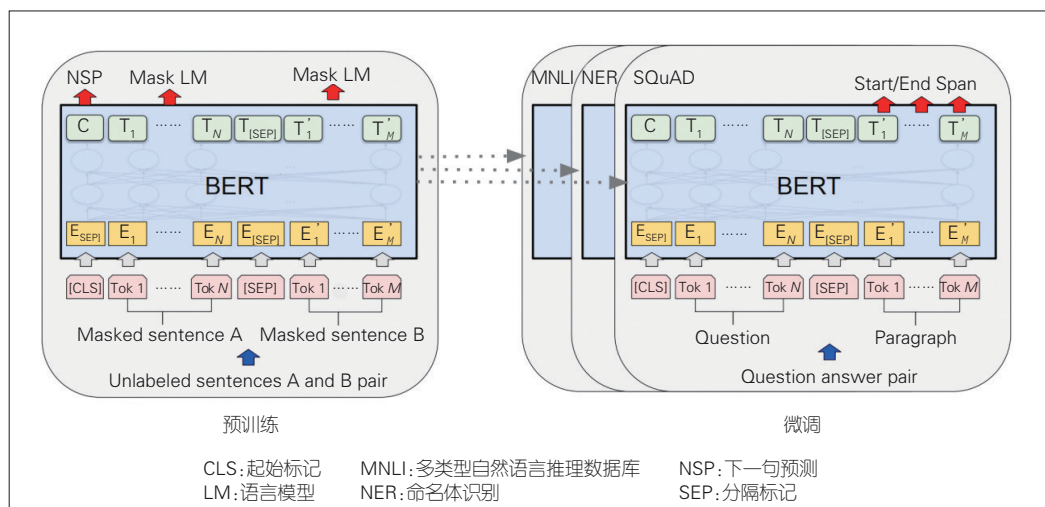


图3 BERT的预训练和微调过程^[5]

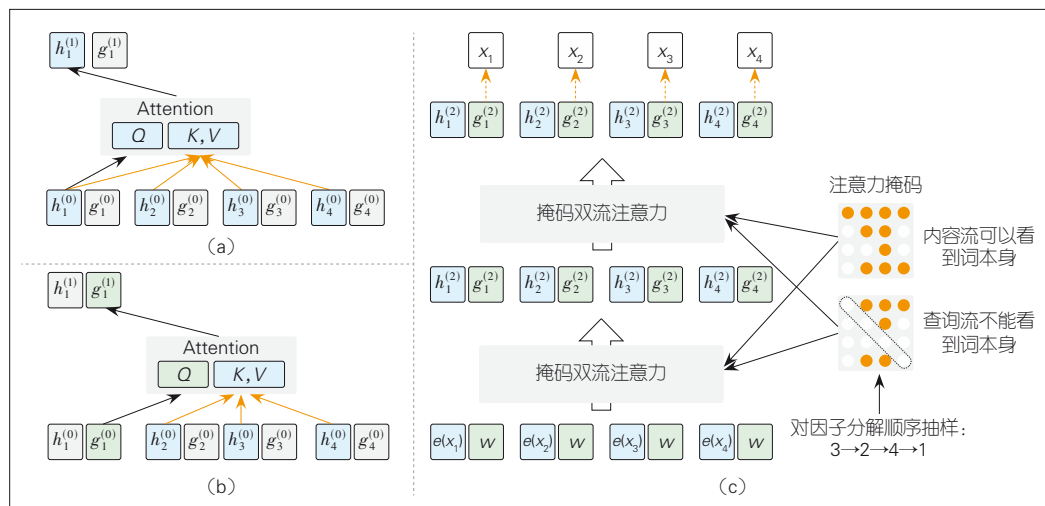


图4 XLNet双流自注意力机制^[14]

进行排列重组,可实现双向上下文的建模,并形成双流自注意力机制,以解决由PLM重新排列所引入的位置信息混淆问题。如图4所示,XLNet用内容流和查询流两种不同的掩码矩阵来进行预测。其中,内容流用于保留词的语义信息,可以看到词本身;查询流不能看到词本身,用于保留词的位置信息,仅在预训练阶段使用。

此外,由于XLNet使用Transformer-XL来替代Transformer,并将其作为特征提取器,因此XLNet拥有比BERT更强的长文本理解能力。

3 预训练语言模型优化方向

预训练语言模型在各类NLP任务中的效果是显而易见的。随着参数规模的扩大和训练数据的增加,预训练语言模型可以获得更好的准确性和泛化性。然而,这是以巨大算力支持为前提的,只有少数大公司才能够承担起这种高昂的算力成本。这个问题在GPT-3模型的研发过程中表现得尤为突出。据报道,为了训练GPT-3模型,微软在Azure云上构建了一个包含1万个GPU、28.5万个CPU内核和400 Gbit/s网络连接的超级计算系统。其中,GPT-3训练一次的费用约为460万美元^[18-19]。在这种情况下,进一步发现、验证和解决模型的潜在问题都非常困难。对此,微软研发团队也认为,当系统出现Bug时,他们也无法对模型进行再训练。

相应地,预训练模型在应用时也需要较大算力和内存支持,往往需要多块高端人工智能芯片或者服务器集群来支撑模型的部署。为了降低预训练模型的部署门槛,业界往往采用量化、剪枝、蒸馏等方法对模型进行压缩,以形成更加轻量化的预训练模型。

(1) 量化是指将模型参数转换为更少比特数来存储和运算,即将模型的精度降低。虽然量化损失了一定的精度,但是它在可接受的准确率范围内能大大提升模型的训练和推理速度。例如,BF16是一种专为加速深度学习训练而设计的16位数字精度格式,在保留FP32(32位浮点数)指数位数的同时减少了16位尾数位。将模型参数从FP32转换为BF16后,模型可以在维持相近准确率的同时实现训练速度的数倍提升^[20-21]。

(2) 剪枝是指去掉模型参数中冗余或者不重要的部分,即减少模型参数。具体来说,剪枝包括元素剪枝和结构剪枝两种方式。其中,元素剪枝是指去掉单个绝对值过小或者对模型影响过小的参数;结构剪枝是指去掉整块模型结构,例如减少多头注意力的数量,或者减少堆叠的Transformer块数量等。

(3) 蒸馏是指较小规模的模型(称为学生模型)从较大

规模的模型(称为教师模型)中学习知识,并替代学生模型从训练数据中学习知识的过程。典型的蒸馏模型包括DistilBERT^[22]、TinyBERT^[23]、MobileBERT^[24]等。这些模型与BERT_{BASE}模型的对比如表1所示。

▼表1 蒸馏模型效果对比

模型	参数量对比/%	推理速度对比	GLUE性能对比/%
BERT _{BASE}	100	1	100
DistilBERT	60.0	1.6	97.0
TinyBERT ₄ (4层)	13.3	9.4	96.8
MobileBERT	23.3	5.5	99.2

在上述优化方法中,量化和剪枝是比较常用的方法。此外,还有其他比较成熟的优化工具,例如TensorFlow Model Optimization、TensorFlow Lite、TensorRT、OpenVINO、PaddleSlim等。由于蒸馏的压缩比更大,它可以和量化、剪枝叠加使用。

4 结束语

自然语言处理技术经历了近百年的发展。机器翻译、智能客服、信息检索与过滤、情感分析和文本生成等,在教育、医疗、司法、互联网等行业中得到了广泛的应用。近年来,预训练语言模型的提出和算力的快速提升,将自然语言处理技术的发展推向了新的高度,使自然语言处理技术在某些领域达到甚至超越了人类水平。然而,目前大规模预训练语言模型仍需要极大的算力支持,训练模型所需的成本仍然较高,能源消耗和碳排放也并不经济,距离落地应用尚有距离。因此,研发出更加轻量的预训练语言模型,是未来重要的发展方向。

参考文献

- [1] ISO/IEC. Information technology—artificial intelligence—artificial intelligence concepts and terminology: ISO/IEC TR 24372:2021(E) [S]. 2021
- [2] 段德智. 莱布尼茨语言哲学的理性主义实质及其历史地位研究[J]. 武汉大学学报(人文科学版), 2013, 66(5): 54-63
- [3] TURING A M. Computing machinery and intelligence [J]. Mind, 1950, 49: 433-460. DOI: 10.1093/mind/lix.236.433
- [4] RADFORD A, NARASIMHAN K. Improving language understanding by generative pre-training [EB/OL]. [2022-02-25]. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
- [5] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [EB/OL]. [2022-02-25]. <https://aclanthology.org/N19-1423.pdf>
- [6] BENGIO Y, DUCHARME R, VINCENT P, et al. A neural probabilistic language model [J]. Journal of machine learning research, 2003, 3: 1137-1155
- [7] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. ACM, 2017: 6000-6010

- [8] DAI Z H, YANG Z L, YANG Y M, et al. Transformer-XL: attentive language models beyond a fixed-length context [EB/OL]. [2022-02-25]. <https://arxiv.org/abs/1901.02860v3>. DOI: 10.18653/v1/p19-1285
- [9] RADFORD A, JEFFREY W, CHILD R, et al. Language models are unsupervised multitask learners [EB/OL]. [2022-02-25]. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
- [10] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners [EB/OL]. [2022-02-25]. <https://arxiv.org/abs/2005.14165>
- [11] WANG A, SINGH A, MICHAEL J, et al. GLUE: a multi-task benchmark and analysis platform for natural language understanding [C]//Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Association for Computational Linguistics, 2018: 353-355. DOI: 10.18653/v1/w18-5446
- [12] RAJPURKAR P, ZHANG J, LOPYREV K, et al. SQuAD: 100 000+ questions for machine comprehension of text [C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2016: 2383-2392. DOI: 10.18653/v1/d16-1264
- [13] ZELLERS R, BISK Y, SCHWARTZ R, et al. SWAG: a large-scale adversarial dataset for grounded commonsense inference [C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2018: 93-104. DOI: 10.18653/v1/d18-1009
- [14] YANG Z L, DAI Z H, YANG Y M, et al. XLNet: generalized autoregressive pretraining for language understanding [EB/OL]. [2022-02-25]. <https://arxiv.org/abs/1906.08237>
- [15] LAN Z Z, CHEN M D, GOODMAN S, et al. ALBERT: a lite BERT for self-supervised learning of language representations [EB/OL]. [2022-02-25]. <https://arxiv.org/abs/1909.11942>
- [16] LEWIS M, LIU Y H, GOYAL N, et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension [EB/OL]. [2022-02-25]. <https://arxiv.org/abs/1910.13461>
- [17] SUN C, MYERS A, VONDRICK C, et al. VideoBERT: a joint model for video and language representation learning [C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2019: 7463-7472. DOI: 10.1109/ICCV.2019.00756
- [18] DEUTSCHER M. OpenAI makes GPT-3 more broadly available to developers [EB/OL]. [2022-02-25]. <https://siliconangle.com/2021/11/18/openai-makes-gpt-3-broadly-available-developers/>
- [19] DICKSON B. The untold story of GPT-3 is the transformation of OpenAI [EB/OL]. [2022-02-25]. <https://bdtechtalks.com/2020/08/17/openai-gpt-3-commercial-ai/#:~:text=According%20to%20one%20estimate%2C%20training%20GPT-3%20would%20cost,tuning%20that%20would%20probably%20increase%20the%20cost%20several-fold>
- [20] HENRY G, TANG P T P, HEINECKE A. Leveraging the bfloat16 artificial intelligence datatype for higher-precision computations [C]//Proceedings of 2019 IEEE 26th Symposium on Computer Arithmetic. IEEE, 2019: 69-76. DOI: 10.1109/ARITH.2019.00019
- [21] Intel. Code sample: Intel® deep learning boost new deep learning instruction bfloat16 - intrinsic functions [EB/OL]. [2022-02-25]. <https://www.intel.cn/content/www/cn/zh/developer/articles/technical/intel-deep-learning-boost-new-instruction-bfloat16.html?wapkw=BF16>
- [22] SANH V, DEBUT L, CHAUMOND J, et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter [EB/OL]. [2022-02-25]. <https://arxiv.org/abs/1910.01108>
- [23] JIAO X Q, YIN Y C, SHANG L F, et al. TinyBERT: distilling BERT for natural language understanding [C]//Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, 2020: 4163-4174. DOI: 10.18653/v1/2020.findings-emnlp.372
- [24] SUN Z Q, YU H K, SONG X D, et al. MobileBERT: a compact task-agnostic BERT for resource-limited devices [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020: 2158-2170. DOI: 10.18653/v1/2020.acl-main.195

作者简介



王海宁，英特尔（中国）有限公司人工智能技术政策和标准总监、中关村高端领军人才、正高级工程师、北京邮电大学兼职教授，担任 ETSI ISG ENI 副主席、CCSA SP1 NFV 特设项目组副组长、CCSA TC610 网络人工智能应用工作组组长等职务；主要研究方向为 4G/5G 网络技术、SDN/NFV、人工智能；2017 年获北京市委组织部青年骨干个人项目资助；主持编制数十项国际标准和行业标准，发表文章多篇，拥有授权专利 30 余项。

数字基础设施建设的思考与实践



Reflections and Practice on Digital Infrastructure Constructions

王喜瑜/WANG Xiyu

(中兴通讯股份有限公司, 中国 深圳 518057)
(ZTE Corporation, Shenzhen 518057, China)

DOI:10.12142/ZTETJ.202202010

网络出版地址: <https://kns.cnki.net/kcms/detail/34.1228.TN.20220409.1854.002.html>

网络出版日期: 2022-04-11

收稿日期: 2022-02-16

摘要: 在通往万物智联的道路上, 高效的数字基础设施和可交易的数字化能力, 正在成为数字经济的核心源动力。构建高效的数字基础设施, 需要“从建好网到用好网”“从连接到算力”的演进; 构建可交易的数字化能力, 则需要“从消费到产业”“从工具到交易”的数字化转型。在当前复杂的产业环境下, 数字经济要实现可持续的高质量发展, 必须培育具备多样性、系统性和开放性的创新生态。

关键词: 数字基础设施; 数字化能力; 连接; 算力; 生态; 交易

Abstract: On the path toward artificial intelligent Internet of Everything (AIoE), highly efficient digital infrastructure and transferable digital capabilities are the core driving forces of the digital economy. Building highly efficient digital infrastructure requires the evolution from reliable network to effective application and from connectivity to computing power; while developing transferable digital capabilities calls for the shift from consumers to industries and from tools to transactions. Under the complex industrial environment, the digital economy can sustain high-quality growth only when a diverse, systematic, and open innovation ecosystem is established across industries.

Keywords: digital infrastructure; digital capability; connection; computing power; ecology; transaction

1 全球数字经济前景

得益于新一代信息技术的发展, 全球数字经济规模持续扩大。2020年, 中国、美国、德国等47个主要国家的数字经济增加值规模达到32.6万亿美元, 占GDP比重为43.7%。在此过程中, 各国采取不同的路径来促进数字经济的发展。例如, 欧盟以数字治理规则的领先探索和数字单一市场的建设为双轮驱动, 来打造数字经济生态; 英国以数字政府为龙头, 来引领数字化转型; 中国则立足于产业基础, 并发挥市场活力, 以适度超前的基础设施建设加速“虚实”经济融合, 引领社会经济高质量发展。在通往万物智联的道路上, 高效的数字基础设施和可交易的数字化能力, 正在成为数字经济的核心源动力。

2 数字基础设施的迭代演进

(1) 从建好网到用好网

以5G和千兆光网为代表的“双千兆”网络初具规模, 为新型基础设施奠定了坚实的基础。网络已从建设期进入发展期, 之后将更加聚焦于以下两个方面: 如何更加精准、

经济地完善网络建设, 即“建好网”; 如何更加高效、规模化地拓展不同场景应用, 即“用好网”。

首先, “好”网络的评价重点已从技术指标过渡到客户感知。在面向个人 (ToC) 和面向家庭 (ToH) 领域, 相比于4G时代, 市场差异主要体现在: 疫情带来的工作和生活方式的变化、短视频及直播等应用的爆发、移动互联市场的快速下沉、扩展现实 (XR) 及元宇宙等差异化应用。因此, 想保持良好的用户感知, 需要5G网络的覆盖广度和深度进一步提升。在面向企业 (ToB) 领域, 信息与通信技术 (ICT) 向运营技术 (OT) 的纵深拓展和贯通融合, 对网络的性能、经济便捷性和安全可靠性等提出了更高的期望。

其次, 在实现路径上, 需要考虑不同场景下性能组合、建维成本、部署便捷等因素的平衡, 精细化的系列方案会成为更优的选择。例如, 杆站等小型化设备的使用, 可以实现5G零占地快速部署; 大规模多输入多输出单边带 (MMSSB1)+X/天线权值自适应 (AAPC) 自优化软件的使用, 使得5G高层覆盖提升了20%~30%; 软件优化实现数字室分和传统分布式天线系统 (DAS) 联合部署, 可以大幅度降低室内覆盖部署成本; 700 MHz/900 MHz等低频频分双工

(FDD)的使用,可以实现5G低成本广域覆盖;Wi-Fi 6实现“真千兆”业务体验,可以支撑新型智慧家庭的全方位需求;5G+无源光纤局域网(POL)、5G时间敏感型网络(TSN),以及5G局域网(LAN)的引入,可以更好地保障企业应用的差异化要求。

建好网的同时还要用好网,而终端和应用是“用”好网的关键。目前,5G已经有千元机,低成本的终端无疑将加速5G渗透率的提升;高清直播、赛事直播等视频应用内容的丰富,更易于发挥双千兆网络差异化优势;依托云网融合的软硬件架构、端到端的一站式管理服务以及工业互联网平台等,ToB领域的应用创新层出不穷,并向“商用化交付、规模化复制”迈进。这里需要特别强调,PC的“wintel”平台生态促进了互联网的蓬勃发展,苹果操作系统(iOS)平台和Android平台催生了移动互联网时代。在网络带宽与算力无限丰富的背景下,可以预料,云电脑将与PC共存,混合现实(MR)技术将成为移动端新的生态。其中,以元宇宙数字孪生、云游戏等为代表的应用越来越多地进入个人消费领域,并将进一步推动5G业务的爆发式成长。

(2) 从连接到算力

互联网数据中心(IDC)的数据显示,过去10年,全球数据量的年均复合增长率(CAGR)接近50%。随着万物智联时代的到来,CAGR的增幅曲线将更加陡峭。与此同时,摩尔定律和尼尔森定律依然发挥作用,但表现出此消彼长的状态,即网络带宽增速已大大超越中央处理器(CPU)性能增速。在数据洪流对端、边、云的冲击之下,分布式和异构计算应运而生,网络和算力相辅相成,体现出更加紧密的关系和更加模糊的边界,以实现海量数据的存储、交换和处理的全局效益最优。更为重要的是,碳中和已经成为全球、全人类共同的价值观和目标,中国陆续出台“新基建”“双碳”“东数西算”等政策指引,加速了绿色低碳的进程。

数字化和低碳化正加速驱动算网进入发展的新阶段。“从连接到算力”的演进,其根基是融合新型基础设施及服务体系,算力网络也因此应运而生。产业界携手创新,在切片、TSN、云网融合、网络智能自主进化、算力网络等技术方向上共同发力,以期达到“网络无所不达、算力无所不在、智能无所不及”的目标。

打造高速泛在、天地一体、云网融合、智能敏捷、绿色低碳、安全可控的高效“数字底座”,需要单一纵深突破和立体协同融合并举,需要软、硬、芯协同,需要ABCDNET(智、链、云、数、网、边、端)贯通。在芯片方面,特定领域架构(DSA)、封装和架构创新延续了摩尔红利,已经成熟应用的基于现场可编程门阵列(FPGA)、图形处理器

(DPU)等硬件加速技术,大幅提升了性价比和边缘效率。在网络方面,持续追求更高频谱和光谱效率的同时,加速向基于新一代IP承载协议(SRv6)的软件定义广域网(SD-WAN)推进,实现网络资源的跨域高效协同编排。在“云”的方面,新型模块化数据中心可以有效降低数据中心总能耗(PUE),满足低碳节能建设要求;可以融合高效的云平台,适配异构资源,支持资源灵活分配、弹性伸缩;可以有效支撑边缘轻量化部署和低成本创新试错,进而探索算力度量、算力感知、算力路由和算力编排等技术。在“智”的方面,除了网络性能优化和自主进化,未来还将构建“算网能”高阶编排大脑,采用统一的应用程序编程接口(API)管控,屏蔽多厂家网络设备、多云环境的差异,为普通用户和垂直行业用户提供“连接+计算+数智能力”的融合服务。

3 数字化能力的纵横拓展

(1) 从消费到产业

随着泛5G等新型数字基础设施建设的推进,数字化应用创新也从个人消费向产业转型和社会治理转变。与消费者相对类似的需求不同,产业数字化转型的核心诉求是降本提效和生存发展。由于行业场景和企业发展阶段的差异,产业的数字化应用明显呈现出碎片化的特征。同时,由于产业数字化依然处于拓展期,必须经历创新试错、商业模式探索和生态孵化等过程,客户也普遍希望数字化资源和能力能够按需部署、灵活扩展、安全可靠、经济便捷。因此,在产业数字化领域,数字化的实现需要聚焦组件化和服务化,围绕场景和关键业务,低成本起步,快速迭代,持续创新。

2021年举办的第四届“绽放杯”5G应用征集大赛吸引了近7000家企业申报12281个行业应用项目,其中包括很多具有商业价值和规模推广潜力的项目。云南神火铝业的平台接入感知数据源已过万,其生产、控制、管理系统初步实现数字贯通,实现了传送带裂纹在线检测、电解槽漏液实时监测、天车远程实时操控等一系列数字化应用,使得阳极组装合格率提升15%,检修皮带空转减少80%,天车单车作业效率提升60%,每年可节约生产用电9000万度以上。山能集团通过轻量化井下5G本质安全型基站、站点的算力引擎NodeEngine、5G网关等部署,使得井下工作面视频传输时延降低50%以上,并实现了掘进机、挖煤机等综采设备的实时远程操控,井下作业人员减少50%,大幅提升了安全生产水平。中国石化石油物探技术研究院通过使用小型化核心网i5GC、拉远型5G基站等车载一体化方案,实现了野外的灵活快速布网,使得先导勘探综合作业效率提升约500%,勘探工期缩短50%以上,人力成本节约50%以上。在天津

港,智能理货、岸桥远程实时操控、无人驾驶电动集卡等泛5G数字化应用不仅带来20%~30%的工作效率提升,而且支撑疫情防控措施的落实。在南京滨江“5G制造5G”生产基地,16类场景60余种5G应用实现了产线柔性化、仓储物流自动化、多场景控制远程化,使得人力成本节省28%,周转效率提升15%,生产效率提升40%。还有广州高铁、商业综合体、医疗等众多创新应用,在为企业降本提效的同时,实现了科技惠民。综上所述,以冶金、钢铁、矿山为代表的大工业园区场景,已经成为垂直行业数字化转型的先锋。

(2) 从工具到交易

除了运营技术(OT)在生产域的深化拓展,企业自身的数字化转型升级也是一个重要课题。数据贯通、流程可控、内外高效协同和泛在智能等应用,有助于打造敏捷且兼具柔性和韧性的企业。从作业流程工具化、自动化的角度来看,数字化转型实质上治标不治本。企业的本质就是交易,而企业数字化转型的本质就是利用数字化手段构建成本最低、体验最优的交易架构,缩短与用户及供应商之间的距离,提升交易效率,进而创造经济效益和社会效益。

以获得2021年“拉姆·查兰管理实践奖”的中兴通讯为例。自2016年启动数字化变革以来,中兴通讯从企业交易架构出发,基础设施先行,通过高效的数字底座,实现统一入口和团队孪生协同;要事优先,从作业人员的角度打造极致体验的“局部工具”;以场景驱动来进行公司系统的改造和数据治理,尤其是最靠近客户的一线员工的场景;综合考虑各领域业务数据的低成本消费诉求,在企业内建立普遍数据思维,以“最靠近客户的一线场景”推动端到端“全域数据”治理。目前,通过iCenter线上沟通与自动化办公软件的使用,全员远程办公效率达95%;通过DevOps研发云,实现30 000多研发人员跨地实时在线协同。通过“从工具到交易”的数字化转型,企业实现数字化的连点成线、连线成面。这不仅能为企业带来效益及效率上的提升,还发挥了企业的交易边界、能力边界和价值圈边界的拓展潜力。

4 建立数字产业创新生态的企业实践

在当前复杂的产业环境下,数字经济要实现可持续的高质量发展,形成全球领先的竞争力与生产力,必须培育具备多样性、系统性和开放性的产业创新生态,需要数字运营体、各行业大型企业、设备商、中小企业在新生态中发挥各自优势,立足产业链中的定位,协同发展,蓬勃生长,形成共赢的生态链。

从企业实践来说,中兴通讯成立了冶金钢铁项目和矿山项目特战队,进一步加速面向场景的能力整合和组织响应。随着5G确定性网络能力的持续提升及5G与视频的深度融合,港口、电子制造、新媒体等行业的5G应用场景也将逐步成熟,并进入规模复制阶段。

中兴通讯坚持“数字经济筑路者”的定位,坚持“开放共赢”的理念,一方面定位于基础设施设备与技术提供商,以用户场景和体验为驱动,提供全球领先的云、网、边、端设备,并积极开放自身核心能力,助力数字运营体及大型企业转型升级;另一方面,以自身能力带动“隐形冠军”等中小企业,坚持与生态伙伴共生、共赢、共智。在底层源动力技术领域,中兴通讯面向算力与网络多样化需求,针对不同业务组合场景,着力DSA、封装和架构创新,持续深化芯片、算法和架构的软硬件协同优化;在产业赋能领域,中兴通讯提供云网基础能力和积木式组件,贡献5G行业标杆经验;在自身实践方面,中兴通讯在数字化研发、数字化办公、数字化生产领域坚定转型,将数字化融入企业血液并与产业分享。

数字经济是世界经济发展的新动能,新一代信息通信技术是数字基础设施建设的强支撑。数字产业生态的完善需要从战略部署向政策落地加速推进,需要新技术、新模式和新业态的突破,需要政产学研用金的合作与协同。高速泛在、集成互联、智能绿色、安全可靠的新型数字基础设施,将充分赋能经济社会的数字化转型升级。

参考文献

- [1] 中国信息通信研究院. 2021年全球数字经济白皮书[EB/OL]. [2022-01-25] http://www.caict.ac.cn/kxyj/qwfb/bps/202108/t20210802_381484.htm
- [2] 杨伊静. 国家发展改革委 中央网信办 工业和信息化部等部门印发《关于加快构建全国一体化大数据中心协同创新体系的指导意见》[J]. 中国科技产业, 2021(2): 31-33. DOI: 10.16277/j.cnki.cn11-2502/n.2021.02.015

作者简介



王喜瑜,中兴通讯股份有限公司执行副总裁、CTO、移动网络和移动多媒体技术国家重点实验室学术委员会主任,中兴通讯技术杂志社总编,教授级高工;1998年入职中兴通讯,先后担任无线研究院院长、技术规划部部长,现全面负责中兴通讯系统产品规划及研发;曾获国家科学技术进步二等奖、广东省科学技术进步一等奖、中国通信协会科技进步一等奖等多项荣誉。

5G行业虚拟专网能力提升与实践



Capacity Improvement and Practice of 5G Industry Virtual Private Network

陆平/LU Ping, 欧阳新志/OUYANG Xinzhi,
高雯雯/GAO Wenwen

(中兴通讯股份有限公司, 中国 深圳 518057)
(ZTE Corporation, Shenzhen 518057, China)

DOI: 10.12142/ZTETJ.202202011

网络出版地址: <https://kns.cnki.net/kcms/detail/34.1228.TN.20220408.1420.002.html>

网络出版日期: 2022-04-08

收稿日期: 2022-02-20

摘要: 作为5G行业应用融合发展的必经之路, 5G行业虚拟专网将持续增强行业应用专网能力, 推动行业应用创新、落地及规模复制。中兴通讯从可靠性提升、服务质量(QoS)保证、网络互联、安全性提升、易运维5个方面来提升5G行业虚拟专网能力, 并总结了行业应用落地的问题与经验, 以进一步指导其他试点应用, 推动5G行业应用形成规模复制的发展态势。

关键词: 5G行业虚拟专网; 专网能力提升; 产业数字化转型

Abstract: As the only way for the integrated development of 5G industry applications, 5G industry virtual private network continues to enhance the capacity of industry application private network and promote the innovation, landing, and scale replication of industry applications. ZTE Corporation mainly improves the virtual private network capability of 5G industry from five aspects: reliability improvement, quality of service (QoS) assurance, network interconnection, security improvement, and easy operation and maintenance. The problems and experience of the implementation of industrial applications are summarized, so as to further guide other pilot applications and promote 5G industrial applications to form the development trend of large-scale replication.

Keywords: 5G industry virtual private network; improvement of 5G private network capacity; industrial digital transformation

1 5G行业虚拟专网逐浪前行

1.1 5G支撑产业数字化转型

数字经济已上升为国家战略, 并成为拉动中国经济增长的重要引擎。发展数字经济意义重大。数字经济是把握新一轮科技革命和产业变革新机遇的战略选择。其中, 产业数字化是数据经济发展的主战场。作为新一代信息通信技术领域的引领性技术, 5G网络是赋能数字经济的关键新型基础设施^[1-2]。

5G行业虚拟专网^[3]是实现5G行业融合发展的必由之路。5G行业虚拟专网与传统运营技术(OT)的融合, 可连通底层设备与企业信息“大脑”, 提供高速率、高可靠、低时延的信息传输通道, 有助于打通企业内部信息纵向流通通道, 构建企业之间、企业与客户之间的横向数据传输通道。通过与大数据、人工智能、云计算等技术的进一步结合, 5G行业虚拟专网能够有效使能行业应用, 推进5G网络与行业的

融合, 助力行业数字化转型。

习近平总书记多次做出重要指示, 要求加快5G网络等新型基础设施建设, 丰富5G技术应用场景。2021年7月, 工业和信息化部会同中央网络安全和信息化委员会办公室、国家发展和改革委员会等部门联合发布《5G应用“扬帆”行动计划(2021—2023年)》(以下简称《行动计划》)。《行动计划》确定了未来3年5G应用的发展路径, 指明了前进的新方向, 同时部署了具体的新任务。

1.2 5G行业虚拟专网的价值与现状

实际上, 诸多行业已经借助以太网、Wi-Fi等网络技术开发多种生产、经营等活动。在3G、4G时代, 交通运输、钢铁、石化化工等领域早已使用专网。然而, 传统专网通信难以适用智慧交通、工业互联网等移动性强的场景, 无法满足多终端、大带宽等网络需求。在工业领域中, 以太网技术的应用存在工业协议不统一、布线繁复等问题。Wi-Fi网络的可靠性、稳定性、移动性都相对较差, 难以承载港口、矿山等复杂环境下的信息传输业务。

在5G网络建设和运营方面, 企业希望在获得网络可控能力的前提下, 进一步降低5G网络的使用成本, 即在

基金项目: 国家重点基础研究发展计划(“973”计划)(2010CB328200、2010CB328201); 国家高技术研究发展计划(“863”计划)(2006AA01Z257); 国家自然科学基金(60602058、60572120); 国家科技重大专项(2009ZX03003-002-02)

获得5G网络运营权的同时,降低企业自身的网络运营成本。5G虚拟专网基于公网资源进行逻辑划分,使部分网元实现共享,并根据行业需求独享部分网络资源,从而降低5G网络的使用和运营成本。无论是各个垂直行业对传输速率、网络时延、安全性的高要求,还是降低网络建设和运营成本的迫切需求,均使得5G行业虚拟专网成为当下的最佳选择。

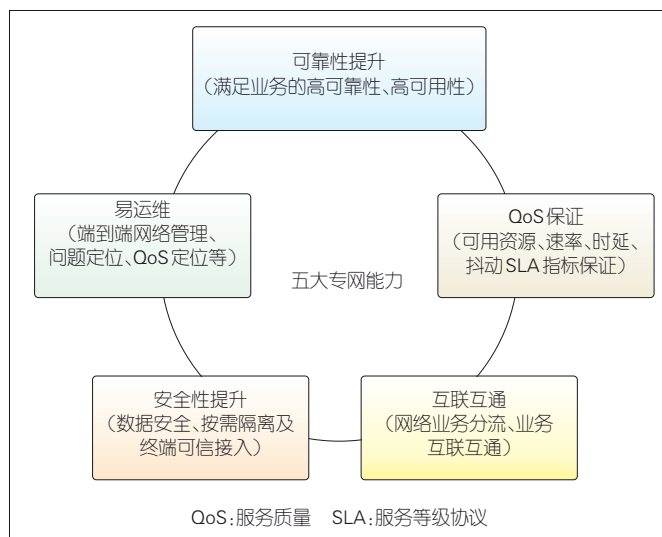
据工业和信息化部统计,截至2021年底,中国累计建成并开通的5G基站数量达到142.5万个,5G终端连接数超过4.9亿个。工业园区、港口和医院等重点区域已建成的5G行业虚拟专网数量超过2300个。这些均加快了5G网络体系的形成。

2 专网能力提升,促进5G行业应用融合发展

随着5G网络与行业应用的不断融合,重点行业和典型应用场景逐步明确。然而,5G应用要实现规模发展仍需要解决一些问题。目前,5G网络技术与行业既有业务的融合仍处于初级阶段。5G主要应用于辅助生产类的业务和信息管理类业务,尚未实现行业核心业务的承载。满足行业业务需求是所有融合工作的前提。

行业在引入5G时,主要考虑以下5个需求:(1)多域多类业务承载。5G网络需要满足企业园区内广域、局域场景下多类型业务的差异化网络需求,如园区生产及管理业务、园区之间的通信等。(2)业务隔离和保障数据安全。5G网络要能够保障公网业务和企业业务安全隔离,使核心业务不出园区。(3)高可靠网络。5G网络要满足企业的高可靠要求,提供连续性通信服务,保证生产安全和生产效率,降低由网络故障导致的损失。(4)网络专享并且建设成本足够低。企业希望在网络专享的前提下,尽量降低5G网络建设和使用成本。(5)简单自主运维5G网络。企业希望拥有内部5G网络的运营管理权和简化的网络运营运维。

为满足行业需求,中兴通讯5G行业虚拟专网主要从5个方面对面向企业(ToB)的网络做专网能力提升,如图1所示。(1)可靠性提升:根据行业用户对网络的可靠性要求,提供最高99.999%的网络可靠性分级保障;(2)服务质量(QoS)保证:能够进一步满足业务服务等级协议(SLA)要求,在满足业务隔离需求的前提下,增强网络上行容量并减少全维度时延,实现核心生产业务承载;(3)互联互通:满足企业园区内二层网络的互通需求,同时园区内的终端与后端服务器能实现互相访问;(4)安全性提升:虚拟专网通过网络切片、非公共网络(NPN)、接入认证增强等多种手段实现园区业务的安全隔离和独立传输;(5)易运维:提供简单运维服务能力,开放企业自服务门户,将5G网络的网



▲图1 中兴通讯5G专网能力提升

络监测、网络参数修改统一提供给行业用户,使得行业用户能自主、便捷地管理园区网络。

2.1 增强网络可靠性,保障端到端高可靠传输

出于生产安全和生产效率的考虑,企业要求网络能够提供不中断的服务。即使出现网络故障,在故障后网络也能够快速恢复,并使网络数据能够稳定、安全地传输。本文将从系统组网可靠性和业务可用性两个维度来讨论网络可靠性提升的技术手段。

(1) 系统组网可靠性

基于双客户前置设备(CPE)的链路冗余、无线双连接冗余和端到端双会话冗余,能够建立端到端双重连接,即建立两个独立、冗余的协议数据单元(PDU)会话。但是并联链路的冗余将大大消耗系统资源并增加系统成本,仅适用于那些需要最高级别保证端到端业务的场景。

结合接入路由器(AR)双发选收功能,当某条链路出现异常时,基于双CPE的链路冗余可以无缝切换到另一条链路上,无须重新建立握手,即可实时计算链路的往返时间(RTT)值,从而选择最低时延的链路。这保证了控制指令下达的实时性和可靠性。

无线双连接冗余通过主备基站与同一个用户面功能网元(UPF)进行相连。PDU会话中的QoS流冗余传输由会话管理功能网元(SMF)来决定。

端到端双会话冗余是基于N3接口的冗余传输实现的。5G无线接入网(NG-RAN)复制上行数据包,并通过两条冗余链路(N3接口)通道发送不同的UPF。这些UPF将与同一个数据网络(DN)相连接。

针对矿区、园区等公网专用的场景，我们可以通过部署基站级算力引擎（NodeEngine）或极简行业5G核心网（5GC），来开通业务断链保活功能，保障企业生产系统的高可用性，进而实现“网断业不断”。

在部署NodeEngine后，当N2、N3断链时，在线业务流仍可保持24 h不断开。在极简行业5GC的实际部署中，我们可以选择仅将UPF下沉，也可以选择将5GC整体下沉。在仅将UPF下沉的情况下，当N2、N4接口断链时，本地UPF与基站仍保持连接，同时在线业务流也可保持24 h不断开；在5GC整体下沉的情况下，当N2、N4断链时，系统将自动切换到5GC应急备份控制面，同时终端可重连至5GC并快速恢复业务。

（2）业务可用性

分组数据汇聚协议（PDCP）复制^[4]是指，在载波聚合和双连接场景下，对PDCP层数据进行复制（包括传输和增强）。PDCP实体在两个无线链路上传输相同的数据，以消除无线环境恶化带来的影响。

在相同的信道条件下，高可靠调制与编码策略（MCS）表格的自适应编码调制更保守。通过更低的调制阶数和编码率，MCS可提高业务的抗干扰能力，降低误码率，增强空口信道的容错性。

2.2 提升业务SLA保证，服务企业核心生产域

为了深化5G与行业应用的融合，5G网络将进一步承载企业核心生产业务。这将对上行带宽、网络时延等SLA指标提出更高的要求。在保证业务隔离的前提下，虚拟专网对上行容量和网络时延进行优化，以满足不同生产业务的承载需求。

（1）资源预留和保障

除了基于5QI资源保障的逻辑软切片外，虚拟专网还可以通过物理资源块（PRB）预留的硬切片实现资源保障。PRB可以按需采取专用预留、优先预留、共享模式进行资源划分，以满足行业用户的差异化SLA需求。

（2）上行容量增强

企业园区会涉及视频监控、机器视觉、可编程逻辑控制器（PLC）远程控制等应用，需要获取高清图像视频数据。这对上行容量有更高的要求。5G上行容量的增强技术主要包括以下几个方面。

在一些特殊场景，如封闭隔离的矿山或4.9 GHz专网场景，引入1D3U帧结构，可大大提升频谱上行占比，从而使上行容量得到提升。

局部的4.9 GHz频段、毫米波频段覆盖，有助于提升行

业室内的网络容量。文献[5]对4.9 GHz频段的上下行容量进行了分析。在鞍钢集团智慧炼钢的应用^[6]中，上行带宽高达750 Mbit/s。由于毫米波频段能提供更大的信号带宽，因此在超高容量区域中，我们可以增加毫米波站点的吸收容量。

此外，针对室内小站密集部署的场景，超级多输入多输出（SuperMIMO）小区+空分多用户多输入多输出（MU-MIMO）技术结合，能够发挥分布式天线的优势。这种组合技术能根据用户设备（UE）分布位置进行多UE空分配对，可实现资源复用，在解决干扰问题同时提升小区容量。

（3）网络时延优化

远程控制、电网差动保护等低时延、高可靠业务对时延有极高的要求，需要为用户提供毫秒级的端到端时延SLA保证。当行业用户要求端到端时延小于20 ms时，我们可以通过以下技术来实现。

频分双工（FDD）频段可重耕为新空口（NR），并在叠加Mini-Slot帧结构或1 ms单周期DS帧结构（D表示全下行时隙，S则包含保护间隔和上下行转换符号）后，使空口时延进一步降低。此时，超可靠低时延通信（URLLC）将占用增强移动宽带（eMBB）的部分时频资源，以保障关键数据的低时延传输。

智能预调度功能使得基站在发送下行数据时主动触发上行预调度，加快上行传输反馈信息的发送速度，从而降低整体数据传输的时延。使用小的上行调度请求（SR）周期，可减少终端发送上行数据的时间间隔，从而达到降低时延的目的。

混合自动重传请求（HARQ）反馈允许系统在一个时隙内的多个物理上行链路控制信道（PUCCH）上反馈混合自动重传请求确认（HARQ-ACK），降低了最大时延和丢包率。

此外，保守调度将有助于避免引入重传。即使发生重传，系统也可以通过降低MCS阶数，来提升重传接收的准确率。

2.3 打通园区二层通信，满足行业应用需求

企业园区内终端之间、服务器与终端之间均需要打通二层网络通信，以实现园区内的互联互通，为实现园区核心生产业务提供基础条件。园区之间的二层通信有助于实现园区之间的业务协同，如智慧物流、供应链协同等。

NodeEngine的eBridge服务功能可以为终端与终端之间、园区业务服务器与终端之间建立一条不受移动网内部地址变化影响的通信路径。在不改变终端原有配置和工作机制的前提下，替代园区有线/Wi-Fi内网，可实现终端与园区服务之

间的互访、园区专网内不同终端之间的互通访问。

5G局域网（LAN）广泛应用于各类工业场景，如PLC远控、机器视觉、电网差动保护等。这种网络可满足企业多园区子网隔离、跨域私网互联、远程终端接入内网等需求。5G LAN能够实现与企业内网的融合组网和无损改造。

2.4 提高业务安全性,免除行业后顾之忧

随着信息技术的发展,企业的网络系统安全和数据安全将受到更多威胁。数据丢失、损坏或泄露,都会给企业造成巨大的经济损失,甚至会影响企业的生存。安全是行业引入5G的前提。虚拟专网将通过NPN专网、接入认证增强来进一步保障企业业务安全。

(1) 基于闭合接入组(CAG)的NPN专网

公网集成的非公共网络(PNI-NPN)^[7]是基于网络切片+CAG的专用网络。该网络与运营商共享无线接入网(RAN)、频谱和核心网的控制面,甚至整个核心网。

UE负责签约CAG,并完成CAG信息的配置。基站小区广播CAG指示和它所支持的CAG身份标识(ID)列表。接入管理功能(AMF)则根据小区CAG能力和UE的CAG签约来判断UE的接入控制和移动性管理。

(2) 基于二次鉴权的接入认证增强

5G网络的二次鉴权会涉及终端、AMF和SMF。在发起PDU会话建立的请求消息中,终端会把与鉴权相关的参数发送至网络侧的AMF。随后,AMF会对终端接入和移动过程进行管理,并与终端进行对应的控制信令交互,同时起到网络附属存储(NAS)信令中转站的作用。SMF与终端进行NAS层的信令交互。NAS信令从UE获取二次鉴权所需的所有信息,从而发起与数据网络认证授权计费(DN-AAA)服务器之间的认证过程。如果这些信息是正确的,那么在DN-AAA服务器返回认证成功的结果以后,SMF会把该认证成功的结果再次反馈给终端。

2.5 简化行业运维,赋能企业自运维

行业用户希望拥有企业内部5G网络的运营管理权,支持5G网络的可视、可管、可控,既能自主开户,查看网络和终端设备的状态,也能根据不同业务要求调整网络,进行网络优化,以最大限度地保证企业生产的连续性和安全性。在5G网络的企业用户自运维方面,中兴通讯提供端到端QoS监测和ToBeEasy行业统一运维两种运维辅助手段,以降低企业网络运维的难度和成本。

(1) 端到端QoS监测

行业应用通过能力开放平台下发端到端SLA测量请求。

UPF和基站根据标记的时间戳来分别计算核心网侧时延和空口时延,并将测量结果通过能力开放平台上报给行业应用或网络编排与管理系统(UME)。在网络及业务上线后,端到端QoS监测将持续提供网络优化服务,支持SLA可视、可管、可控。

(2) ToBeEasy行业统一运维

如图2所示,ToBeEasy行业统一运维包括ElasticNet UME R88专业集中运维系统、ElasticNet产业数字操作系统(IDOS)企业运维门户、ZXeLMT本地运维工具、端到端定位方案以及其他网规网优工具箱。这些能力组合应用可以为各种ToB场景提供匹配度最高的管理维护方案。

基于云原生技术,IDOS企业运维门户面向企业连接业务,采用业务、连接、切片、终端、专业网络以及行业应用程序(APP)等多种运维方式剖析网络和业务状态,实时监测业务健康度和SLA指标劣化趋势,提前预测风险,保障ToB园区业务连接的正常运行,并用信息技术(IT)化服务的方式为运营商和企业提供直观、简单、智能的运维和能力开放服务。IDOS运维门户主要部署在园区内部。

3 5G行业虚拟专网的实践

自5G行业虚拟专网商用两年以来,在行业应用方面,基础电信企业和垂直行业企业共同探索5G应用试点,并在冶金、港口、矿山、电力等重点行业进行了5G应用的技术和场景验证。中兴通讯总结行业应用落地交付阶段的问题与经验,对5G行业虚拟专网能力和技术体系进行补充和完善,进一步指导其他试点应用,推动5G行业应用形成规模复制的发展态势。

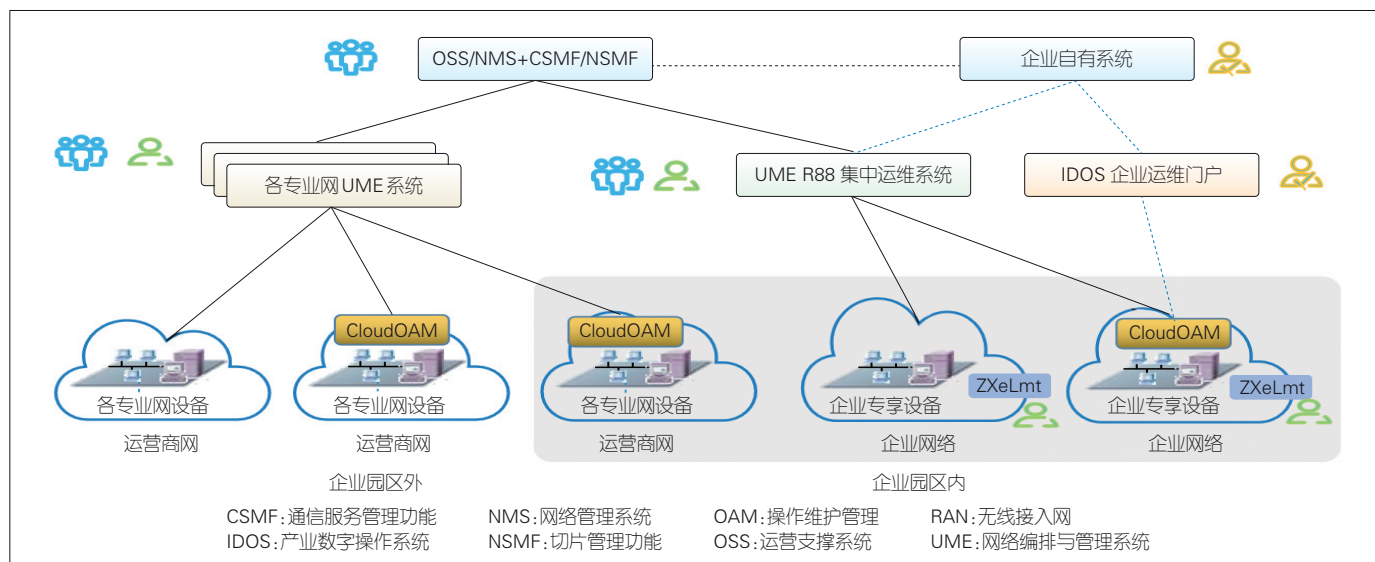
3.1 特殊场景覆盖优化

随着无线环境的日益复杂、5G行业应用场景的持续丰富以及用户需求的不断提升,5G网络覆盖不断面临新的挑战。通过细分场景和需求,虚拟专网灵活采用多种方式实现覆盖,可满足特殊场景的覆盖,如低空、矿山井下、近海海面等。这大大拓展了5G网络的应用场景。

(1) 低空覆盖

随着民用无人机技术的迅猛发展,无人机在安防^[8]、电力^[9]、气象、石油管线巡查、应急通信、野生动物保护等领域均产生了显著的经济效益。利用5G网络为无人机提供数据链路,可构建低空覆盖网络,将极大提升无人机的服务能力,丰富相关应用场景。

基于大规模天线阵列的数据联合接收和发送,低空场景使用主波束完成面向地面用户的覆盖,并使用波束旁瓣在垂



▲图2 ToBeEasy行业统一运维

直维度完成低空区域覆盖，形成低空-地面一体的立体覆盖能力。

在天津港项目中，无人机在北港池海域上空进行大范围巡航，整个航线长度约20 km。在5G低空覆盖的场景下，通过内嵌5G模组，无人机拍摄的4K高清视频可以直接通过5G网络来传输。在该项目中，无人机实现了低时延下的广域飞行、视频实时回传、状态实时监测等。该项目使用3台无人机，在中国联通的3.5 GHz频段下，采用广播波束单边带（SSB）1+X覆盖方案，实现了整个北港池区域的覆盖。无人机飞行高度为120 m，数据传输速率为48 Mbit/s。根据覆盖要求，我们进行链路预算，并结合地对空的传播模型，进行站点规划。

在该项目中，建立单个站点即可实现整个北港池的网络覆盖。该项目采用基站天线挂高，以及广播波束SSB 1+X天线方案，即1个水平宽波束和X个垂直波束旁瓣（ $X=\{1,2,3,4\}$ ）。这种方案使用最少的水平SSB波束和按需垂直波束实现了三维全空间覆盖，可适应复杂环境下的各种场景。在水平覆盖方面，我们针对时域轮发的特点，为不同小区配置不同的SSB波束，以错开邻区之间SSB波束的发送时间，从而规避邻区之间的干扰；在垂直覆盖方面，根据覆盖需求，我们使用更多的波束来实现空间覆盖，从而达到5G网络最优覆盖的效果。SSB功率增强功能可以提升水平波束的增益能力，使SSB 1波束增益提高6 dBm。因此，水平SSB 1波束与水平SSB 8波束覆盖相当，同时预留出更多波束的位置以用于垂直覆盖。仿真测试表明，满足上行速率要求的覆盖百分比为97.94%，满足相关设计要求。

（2）超远覆盖

超远覆盖可实现大于普通宏站的大范围覆盖，覆盖半径范围为15~100 km。相关应用场景包括海面、沙漠、农村、草原、港口等。超远覆盖主要解决两大问题：覆盖距离和终端正常接入。

超远覆盖的覆盖距离主要体现在视距传播场景中，其有效覆盖范围取决于多种因素：频段、传播模型、天线高度、基站功率、天线选型等。频段越低，天线海拔高度越高，基站功率就越大，小区覆盖距离也就越远。

超远距离无线覆盖的目标是保证覆盖范围内的终端能够正常接入网络。在电磁波速率确定的条件下，覆盖距离越远，无线信号传播的时延就越大，无线覆盖的蜂窝小区所能够容忍的传播时延也就越大。根据第3代合作伙伴计划（3GPP）协议，超远覆盖适合采用长序列（序列长度为839）的物理随机接入信道（PRACH）格式。这是因为这种格式中的循环前缀（CP）和保护间隔（GP）时延更长，即在接入阶段小区所能容忍的时延更长。

在福建移动海域覆盖项目中，我们采用移动700 MHz的PRACH Format 1格式，使接入距离长达100 km；在天线选用上，采用高增益平板天线和透镜天线；在站址选择上，优先考虑高海拔站点。海面传播模型要求天线挂高与覆盖目标之间有良好的无线传播环境。目前，移动700 MHz的频段虽然实现了海面超远覆盖，但是仍存在同频连续组网小区间干扰、超远覆盖站点对内陆同频公网干扰等问题。因此，相关方案需要做进一步验证与优化。

（3）矿山井下巷道覆盖

井下巷道是一个狭长并且封闭的空间,巷道宽度、高度不一,空间受限严重,无线传播环境复杂。针对这种典型的线状覆盖场景,我们在每个5G皮基站的左右两个方向上分别布置一个定向天线。井下基站的覆盖半径会随着频段选择、巷道环境、业务要求的不同而发生变化。一般来说,2.6 GHz/3.5 GHz的覆盖半径范围为100~300 m。与之相比,700 MHz频段的覆盖范围更广,建站成本更低,绕射能力和穿透能力更强,在相同发射功率下信号传输效率高、损耗小,适用于井下粉尘大、电磁信号干扰强、金属设备多的场景。在中煤大海则煤矿项目中,中兴通讯与合作伙伴首次把700 MHz频段引入井下,实现了700 MHz和2.6 GHz的混合组网和全矿井的网络覆盖。测试表明,700 MHz的覆盖半径是2.6 GHz/3.5 GHz的3~4倍。

3.2 视频业务卡顿问题优化

视频业务对网络带宽、时延、抖动等SLA指标有一定的要求。在项目实施过程中,多个项目现场出现了视频卡顿的情况。产生这种现象的原因主要有3种。(a)网络抖动:网络抖动会导致客户端收到的数据时间忽长忽短,同时那些不允许在客户端进行缓存的高实时性业务也会放大网络抖动产生的影响。(b)网络丢包:视频数据在传输过程中会出现数据包丢失的现象。例如,当用户数据报协议(UDP)突发的高带宽超过网络传输门限时,系统在解码时无法将帧数据还原,从而导致网络丢包。(c)带宽不足:当视频播放数据所需的带宽大于实际带宽时,客户端需要等数据完成缓冲以后再进行播放,从而造成业务卡顿。

下面我们以天津港视频回传和岸桥远控业务为例,分析当多路视频I帧冲突和最大传输单元(MTU)超过1398时乱序被引入的情况。

(1) 多路视频I帧冲突

网络摄像头在传输数据时会进入突发模式。当多路摄像头在同一个网络中进行数据传输时,同一时刻可能会有多个摄像头的I帧并发传输。单个摄像头需要4 Mbit/s的带宽,I帧大小约为150 kB。当同时有48路摄像头接入的时候,就有可能发生5路I帧碰撞。短时带宽约为301 Mbit/s,5路I帧有750 kB。这种尖峰的UDP流量在网络中传输时除了会造成拥塞以外,还会在缺少缓存的点位上产生丢包。

为解决上述问题,我们首先考虑降低摄像头端的数据量,即通过调整摄像头端的相关参数来降低网络需求;其次在中间设备中增加缓存,以缓冲突发流量,避免丢包,同时“熨平”流量波动;最后对无线网络侧进行扩容,以提升带宽门限。

在天津港岸桥远程控制场景中,岸桥上所有摄像头的数据都需要通过CPE统一上传。前端交换机、eBridge类似于网络交换设备,会将所有处理后的数据包线速转发到后端。这里的缓冲点是流量上行的汇聚点,它承载着非常大的数据压力,能够在CPE侧应对大流量冲击。

(2) MTU值超过1398而引入乱序

在网络传输中造成视频卡顿的另一原因是,网络中传输数据包的包长超过了独立网元的MTU。当包长超过MTU时,路由设备会对数据报文进行分片或者丢弃处理。

在天津港项目中,主要业务消息的终端MTU值超过了1398。当项目现场采用传输控制协议(TCP)来传输视频业务时,服务端出现了明显的业务卡顿现象。经抓包确认,产生这种现象的主要原因是网络中存在数据乱序。当终端侧MTU值为1398时,数据包传输eBridge将增加44字节头+14字节MAC,无线将增加44字节。此时,无线侧MTU值则为1500。因此,路由设备会对数据报文进行分片或丢弃处理。

对此,我们可以通过修改网络传输路径上各个节点的MTU值,包括CPE侧MTU、无线侧MTU、核心网UPF的MTU等,来保证大数据包的通行,从而避免分片或者丢包。

3.3 PLC业务操控中断问题优化

PLC业务的高度实时性处理和相关编程模型,均要求数据网络传输精确到单个数据包。在多个工业项目中,出现了PLC业务操控中断的问题。对此,我们可以从网络架构、无线参数、PLC程序等方面进行优化。

以天津港岸桥远控PLC业务为例,PLC链路经常出现业务断链的情况。在无线通信情况下,对业务影响最大的是人机接口(HMI)主动断链。在HMI主动断开链接以后,系统会先等待2060 ms再重新发起建链请求。由于在中间这段时间里链路上不会有数据传输存在,所以一旦出现这种情况,在HMI服务端上的心跳监测界面就会出现数据跳变。

通过分析HMI断链的数据包,我们梳理了4种场景:岸桥PLC数据包未同时到达服务端、数据包到达服务端但出现乱序丢包、同时发送的数据包未同时到达服务器、上行数据丢包重传后业务无法恢复。

对此,我们可以借助多种技术手段来解决PLC操控中断问题。在网络结构调整方面,PLC数据的传输方式由与视频传输共用CPE变为独立使用CPE。在无线参数调整方面,我们在PLC数据无线侧提升QoS保障,同时PLC SIM卡配置的5QI为6,普通用户5QI为9。此外,我们也可以对PLC程序做优化处理,调整MTU值,即把原先的1436调整为1380。借助以上手段进行网络优化可以最大限度地避免PLC链路出

现业务断链现象。

3.4 基于二层隧道协议(L2TP)实现5G CPE下多终端与服务端的双向互通

在使用5G网络替代原本的有线接入时,企业需要解决终端之间双向互通访问的关键问题。在5G独立组网(SA)的环境下,基于L2TP实现虚拟专有拨号网络(VPDN)双向路径的业务互通,可同时避免5GC开启用户静态签约服务,为现场应用提供可行方案。这不仅简化了系统配置,还能实现多台现场设备与服务器的相互访问,满足PLC远控和视频回传的业务场景需求。

然而,南京滨江、云南神火等项目中普遍存在3个问题:一是现场多项目存在单CPE下挂多终端需求,而终端桥接模式的单个CPE只能下挂单个终端;二是现场多项目当前采用eBridge进行组网,而单个终端接入就需要至少3台eBridge设备,这使系统配置变得复杂;三是当前CPE仅可单向访问核心网设备。因此,使用L2TP LAC-Auto-Initiated进行二层通道打通,可解决以上问题,促进了项目交付进度和相关方案推广。

在终端侧,5G CPE通过L2TP与后端建立VPDN隧道。MAC地址绑定可确保下挂设备地址的固定。MC801A终端可作为L2TP客户端以及远端侧的DHCP服务器。网络侧开启L2TP网络服务器(LNS)服务。皖通ISG1800-2S多业务路由可作为LNS服务端以及近端侧的DHCP服务器。进一步地,MC801A和ISG1800-2S的VPDN配置均需要做到:在MC801A上完成LAC、L2TP以及DHCP配置,在ISG1800-2S上完成LNS配置。

相关测试表明:L2TP通道可达,801A CPE下挂多终端设备可行,远程桌面及文件传输协议(FTP)业务能够开展。因此,L2TP实现了5G CPE下挂多终端与服务端的双向互通,有助于实现PLC远控、视频回传业务的部署,进一步促进了相关方案的推广。

4 结束语

随着3GPP 5G国际标准研制的持续推进,技术标准的演化将持续推动行业应用的发展。R17版本在持续提升通信基础能力的同时,将支持更广泛的行业应用。5G行业虚拟专网将重点关注技术的试点验证和商用落地,进一步丰富5G应用场景,助力行业应用实现从“1”到“N”的复制。行业应用目前已完成从“0”到“1”的突破。工业制造、能源

矿山等多个先导行业正在打造应用标杆。部分先导场景应用,如机器视觉质检、无人巡检、远程掘进等,将实现规模复制。

参考文献

- [1] 陈亿根,尹晓峰,邵黎勋. 5G+工业互联网应用实践[J]. 中兴通讯技术, 2020, 26(6): 2-6. DOI: 10.12142/ZTETJ.202006002
- [2] LIU Z, GAO Y, LI D, et al. Enabling energy efficiency in 5G network[J]. ZTE Communications, 2021, 19(1): 20-29. DOI: 10.12142/ZTECOM. 202101004
- [3] 中国信息通信研究院. 5G行业虚拟专网网络架构白皮书[R/OL]. [2022-01-28]. http://www.caict.ac.cn/kxyj/qwfb/bps/202007/t20200728_287336.htm
- [4] 马瑞涛,王光全,任驰,等. 3GPP R16 5G核心网标准及关键技术研究[J]. 电子技术应用, 2020, 46(11): 30-35+40. DOI: 10.16157/j.issn.0258-7998.200993
- [5] 陈锦浩. 基于4.9 GHz频段的5G专网覆盖和容量能力研究[J]. 广东通信技术, 2021, 41(5): 50-53, 60. DOI: 10.3969/j.issn.1006-6403.2021.05.011
- [6] 中兴技术. 鞍钢集团、中国移动、中兴通讯联合发布全球首个5G工业专网在鞍钢智慧炼钢中的应用[EB/OL]. (2020-09-21) [2022-01-21]. <https://mp.weixin.qq.com/s/bKZkDc4YjqDrsY1loY4kg>
- [7] 蒋峥,刘胜楠,田树一,等. 5G非公共网络技术分析[J]. 移动通信, 2020, 44(4): 12-18. DOI: 10.3969/j.issn.1006-1010.2020.04.003
- [8] 范永飞. 警用无人机的应用特点与发展趋势[J]. 中国安防, 2018(6): 42-44. DOI: 10.3969/j.issn.1673-7873.2018.06.009
- [9] 陈熙. 无人机在输电线路巡检中的应用及发展前景[J]. 电力系统装备, 2019(1): 203-204

作者简介



陆平,中兴通讯股份有限公司副总裁、移动网络和移动多媒体技术国家重点实验室副主任;研究方向为云计算、大数据、增强现实、基于多媒体服务的技术等;主持和参与多个国家科技重大专项、国家科技支撑项目;发表论文多篇,撰写《物联网能力开发与应用》《云计算中的大数据技术与应用》等多部著作。



欧阳新志,中兴通讯股份有限公司5G行业产品线工业总经理、中兴通讯技术专家委员会委员;曾负责短消息、WAP网关、数据中心、云计算、大数据等技术的规划及研发,主要从事5G产品行业应用的整体规划和应用推广;多次荣获省、部级科技进步奖。



高雯雯,中兴通讯股份有限公司产业数字化方案部行业综合方案经理;主要研究领域为5G行业专网及其应用;2021年,参与第4届“绽放杯”5G应用征集大赛全国赛,所负责的项目获工业互联网专题赛道一等奖、全国总决赛二等奖。

《中兴通讯技术》杂志（双月刊）投稿须知

一、杂志定位

《中兴通讯技术》杂志为通信技术类学术期刊。通过介绍、探讨通信热点技术，以展现通信技术最新发展动态，并促进产学研合作，发掘和培养优秀人才，为振兴民族通信产业做贡献。

二、稿件基本要求

1. 投稿约定

- (1) 作者需登录《中兴通讯技术》投稿平台：tech.zte.com.cn/submission，并上传稿件。第一次投稿需完成新用户注册。
- (2) 编辑部将按照审稿流程聘请专家审稿，并根据审稿意见，公平、公正地录用稿件。审稿过程需要 1 个月左右。

2. 内容和格式要求

- (1) 稿件须具有创新性、学术性、规范性和可读性。
- (2) 稿件需采用 WORD 文档格式。
- (3) 稿件篇幅一般不超过 6 000 字（包括文、图），内容包括：中、英文题名，作者姓名及汉语拼音，作者中、英文单位，中文摘要、关键词（3 ~ 8 个），英文摘要、关键词，正文，参考文献，作者简介。
- (4) 中文题名一般不超过 20 个汉字，中、英文题名含义应一致。
- (5) 摘要尽量写成报道性摘要，包括研究的目的、方法、结果 / 结论，以 150 ~ 200 字为宜。摘要应具有独立性和自明性。中英文摘要应一致。
- (6) 文稿中的量和单位应符合国家标准。外文字母的正斜体、大小写等须写清楚，上下角的字母、数据和符号的位置皆应明显区别。
- (7) 图、表力求少而精（以 8 幅为上限），应随文出现，切忌与文字重复。图、表应保持自明性，图中缩略词和英文均要在图中加中文解释。表应采用三线表，表中缩略词和英文均要在表内加中文解释。
- (8) 所有文献必须在正文中引用，文献序号按其在文中出现的先后次序编排。常用参考文献的书写格式为：
 - 期刊 [序号] 作者. 题名 [J]. 刊名, 出版年, 卷号 (期号): 引文页码. 数字对象唯一标识符
 - 书籍 [序号] 作者. 书名 [M]. 出版地: 出版者, 出版年: 引文页码. 数字对象唯一标识符
 - 论文集中析出文献 [序号] 作者. 题名 [C]// 论文集编者. 论文集名 (会议名). 出版地: 出版者, 出版年 (开会年): 引文页码. 数字对象唯一标识符
 - 学位论文 [序号] 作者. 题名 [D]. 学位授予单位所在城市名: 学位授予单位, 授予年份. 数字对象唯一标识符
 - 专利 [序号] 专利所有者. 专利题名: 专利号 [P]. 出版日期. 数字对象唯一标识符
 - 国际、国家标准 [序号] 标准名称: 标准编号 [S]. 出版地: 出版者, 出版年. 数字对象唯一标识符
- (9) 作者超过 3 人时，可以感谢形式在文中提及。作者简介包括：姓名、工作单位、职务或职称、学历、毕业于何校、现从事的工作、专业特长、科研成果、已发表的论文数量等。
- (10) 提供正面、免冠、彩色标准照片一张，最好采用 JPG 格式（文件大小超过 100 kB）。
- (11) 应标注出研究课题的资助基金或资助项目名称及编号。
- (12) 提供联系方式，如：通讯地址、电话（含手机）、Email 等。

3. 其他事项

- (1) 请勿一稿多投。凡在 2 个月（自来稿之日算起）以内未接到录用通知者，可致电编辑部询问。
- (2) 为了促进信息传播，加强学术交流，在论文发表后，本刊享有文章的转摘权（包括英文版、电子版、网络版）。作者获得的稿费包括转摘酬金。如作者不同意转摘，请在投稿时说明。
- (3) 编辑部地址：安徽省合肥市金寨路 329 号凯旋大厦 1201 室，邮政编码：230061。
- (4) 联系电话：0551-65533356，联系邮箱：magazine@zte.com.cn。
- (5) 本刊只接受在线投稿，欢迎访问本刊投稿平台：tech.zte.com.cn/submission。

中兴通讯技术

(ZHONGXING TONGXUN JISHU)

办刊宗旨：

以人为本，荟萃通信技术领域精英
迎接挑战，把握世界通信技术动态
立即行动，求解通信发展疑难课题
励精图治，促进民族信息产业崛起

产业顾问（按姓名拼音排序）：

段向阳、高 音、胡留军、刘新阳、
陆 平、史伟强、王会涛、熊先奎、
赵志勇、朱 方、朱晓光

双月刊 1995 年创刊 总第 163 期
2022 年 4 月 第 28 卷 第 2 期

主管：安徽出版集团有限责任公司
主办：时代出版传媒股份有限公司
深圳航天广宇工业有限公司
出版：安徽科学技术出版社
编辑、发行：中兴通讯技术杂志社

总编辑：王喜瑜
主编：蒋贤骏
执行主编：黄新明
编辑部主任：卢丹
责任编辑：徐烨
编辑：杨广西、朱莉、任溪溪
设计排版：徐莹
发行：王萍萍
编务：王坤

《中兴通讯技术》编辑部
地址：合肥市金寨路 329 号凯旋大厦 1201 室
邮编：230061
网址：tech.zte.com.cn
投稿平台：tech.zte.com.cn/submission
电子信箱：magazine@zte.com.cn
电话：(0551)65533356

发行方式：自办发行
印刷：合肥添彩包装有限公司
出版日期：2022 年 4 月 20 日
中国标准连续出版物号：ISSN 1009-6868
CN 34-1228/TN
定价：每册 20.00 元