2023年 第3期

江苏省计算机学会通讯

COMMUNICATIONS OF THE JSCS

-- ChatGPT与李群机器学习

-- 《操作系统》课程体系建设教学成果交流

江苏省计算机学会常务理事单位

南京中医药大学人工智能与信息技术学院

人工智能与信息技术学院，前身是南京中医学院 1985 年成立的“中医计算机应用研究中心”，2003 年学校在此基础上成立了信息技术学院。2019 年 6 月，为了加强现代信息技术对医药类主干学科的支撑融合，发展人工智能、智能医学和医学信息工程与技术，学院更名为人工智能与信息技术学院。

学院建有“计算机科学与技术”校级重点学科、“软件工程”校级重点培育学科，与文献所共建“中医药信息学”国家中医药管理局重点学科。拥有“江苏省智慧中医药健康服务工程研究中心”1 个省级工程研究中心，以及“信息技术实验中心”和“医药信息技术实践教育中心”2 个省级实验教学示范中心。设有国家一流专业建设点“计算机科学与技术”、江苏省一流专业建设点“软件工程”“医学信息工程”以及新专业“人工智能”4 个本科专业。拥有“软件工程”一级学科硕士点和“中医药人工智能”自主设置交叉学科博硕士点、“中医药信息学”自主设置二级学科博硕士点。

学院下设计算机科学与技术教研室、软件工程教研室、医学信息工程教研室、人工智能教研室、计算机公共基础教研室、数学教研室、物理教研室等 7 个教研室。现有教职工 82 人，其中专任教师 66 人，具有博士学位 43 人、占比 65%，高级职称教师 32 人、占比 48.5%，省级高层次人才或团队 25 人次。青年教师中有近 20 人赴澳大利亚悉尼大学、西悉尼大学、斯威本科技大学、美国弗吉尼亚大学等高校进行课程培训和学术交流，形成了一支结构合理、德才兼优的师资队伍。

学院毕业生就业率在学校名列前茅，就业质量高，专业对口率高，多数在 IT 企业、医院、医疗机构、政府机关等企事业单位从事计算机相关工作。多名优秀毕业生推免至浙江大学、南京大学、东南大学、山东大学、北京航空航天大学等著名高校，多名同学考入清华大学、中山大学、哈尔滨工业大学、英国谢菲尔德大学、澳大利亚墨尔本大学等国内外知名大学攻读硕士研究生。

江苏省计算机学会通讯

COMMUNICATONS OF THE JSCS

地　址：中国江苏省南京市栖霞区仙林大道 163 号

邮　编：210023

电　话：025-89680909

邮　箱：jscs@njn.edu.cn

学生记者团专栏

教学成果交流

政策解读

学术交流

会员风采

科学普及

科创成果

ChatGPT 与李群机器学习

——李凡长教授专访

李凡长，教授 , 俄罗斯工程院外籍院士，李群机器学习 (Lie Group Machinelearning，LML)、动态模糊逻辑 (Dynamic Fuzzy Logic，DFL)、动态模糊机器学习 (Dynamic Fuzzy Machine Learning,DFML) 和多维度协同教育方法的开创者。曾任苏州大学计算机科学与技术学院院长，现任江苏省机器学习与网络安全交叉研究工程中心主任，江苏省计算机信息处理技术重点实验室主任，苏州大学机器学习与类脑计算国际合作实验室主任。国家自然科学基金重点项目、国家重点研发计划课题负责人。获江苏省有突出贡献中青年专家荣誉称号 , 科学中国人 2015 年年度人物。江苏省计算机学会副理事长，江苏省人工智能学会副理事长，苏州市人工智能学会理事长。先后承担 10 余项国家自然科学基金重点、面上、国家重点研发项目等项目，发表论文 300 余篇，出版学术专著 10 部 (其中在美国、德国出版英文专著 3 部)，教材 5 部，授权发明专利 40 余项，曾获省级科技奖二等奖 2 项、IEEECSGRCPioneerAward1 项和省级教学成果二等奖 2 项。

在人工智能前沿论坛中，李凡长教授做了题为《ChatGPT 与李群机器学习》的报告，主要介绍了自 2022 年 12 月 OpenAI 发布 ChatGPT 以来，以通用人工智能为代表的人工智能技术变革的快速发展。报告以人工智能的革命为主题，围绕 ChatGPT、科学研究范式、李群机器学习进行介绍，最后对 ChatGPT 的一些问题进行讨论。

在报告中，李教授对李群机器学习的优势作出了详细的介绍和解读，着重介绍了李群机器学习在处理几何结构或拓扑结构中存在对称性的数据时展现出的计算复杂度低，可解释性高等特点，并结合 ChatGPT 的发展，为未来二者结合擘画蓝图。

在报告结束后，我们有幸采访到了李凡长教授

Q1：您提到 ChatGPT 目前在解决一些比较简单的问答类问题上面已经取得了比较好的效果，但是它对于复杂问题的求解上依然具有较大的局限性。您也提到，李群机器学习在解决这种复杂问题上具有一定优势。您觉得李群机器学习以后可能会在哪些具体的方面为大规模语言模型技术的发展提供帮助呢?

李凡长：这是一种技术的发展，目前来看 ChatGPT 恰恰能够解决一些简单的数学问题，以数学研究为例，我们希望 GPT 等大语言模型能够解决达到数学家水平的问题，而不是仅仅解决中学生问题。如果能将李群机器学习的算法用到大模型技术中，可能就能达到这种水平。目前很多数学研究的发展，都在利用李群和李代数理论，如偏微分方程求解等。从数学角度讲，李群已经帮助了很多数学的发展，而目前人工智能面对的主要问题是复杂问题的解决而

非仅仅是简单问题，面对复杂问题，从表示的角度来讲，我们需要找到一个合适的表示方法，李群就是一个好的表示办法。具体表现在，我们要考虑定性和定量都能表示的数学模型、线性和非线性共存的数学模型、整体和局部的数学模型、单参数和多参数的数学模型李群都完全有这样的能力。

Q2：ChatGPT 对我们整个社会发展的影响很大，您觉得像 ChatGPT 这样的技术会对我们科学研究范式带来怎么影响？

李凡长：正如我刚刚报告中提到的，ChatGPT 的出现把我们整个人类推向了智能时代，标志着智能时代的真正到来。在这个智能时代，我们每个人都要站在 ChatGPT 这样的平台上，站在智能的平台上思考问题，所以我觉得 ChatGPT 的出现可以帮助我们解决很多问题，我们不要以对抗的情绪去对抗它，要以它能够帮助我们这样的角度来看待。但是要让这种技术真正服务于我们，首先要知道哪些事情是它能够帮助我们的，哪些事情是我们应该做的，这样划分开之后，ChatGPT 就能更多地为人类带来好处了。一方面，学习的角度上来讲，大家会变得更辛苦，需要更多的学习，不然可能没办法提出合适的问题。另一方面，GPT 本身可以较好解决的问题，如果人类还花费大量精力去做，势必将对人类的研究发展造成挑战。因此，GPT 的出现和发展，对人类的教育、学习、研究都提出了更高的要求。我们之所以要使用 GPT 是为了让它能够帮助我们解决问题，更重要的是我们通过“问”的途径能够判定哪些是它能够做的，哪些是它不能够做的，我们应该更多聚焦于它不能做的事情，才能体现我们人类的价值。

在人工智能迅速发展的时代，李凡长教授的洞见和见解为我们揭示了一个宝贵的视角。通过将李群机器学习的强大工具与大规模语言模型如 ChatGPT 相结合，我们有可能创造出更为强大和多功能的人工智能系统。然而，正如李教授所强调的，我们必须清楚地认识到这些工具的局限性，并将其作为一个辅助工具来协助人类解决更为复杂的问题。这不仅需要技术的进步，还需要我们在教育和学习上付出更多的努力，以适应这个不断变化的世界。人工智能的兴起给我们带来了巨大的机遇，但同时也提出了新的挑战。让我们以开放的心态和勇于创新的精神，共同探索人工智能为我们的生活和科学带来的无限可能。

学会动态

第十一届 CCF 大数据学术会议

9 月 10 日，由中国计算机学会（CCF）主办，江苏省计算机学会（JSCS）协办的第十一届 CCF 大数据学术会议在南京大学国际会议中心圆满落幕。本次会议以“算力模型协同，数据智领未来”为主题，国内外大数据领域的著名专家学者共同深入探讨在数字经济和大模型时代下，大数据领域面临的机遇和挑战。清华大学郑纬民院士，滑铁卢大学李明院士，德国哥廷根大学傅晓明院士，香港科技大学首席副校长兼任计算机科学及工程学系讲座教授郭毅可院士，同济大学校长郑庆华教授，哈尔滨工业大学高会军院士分别作大会主旨报告。

新型云原生数据库架构的研究进展与前瞻

——于戈教授专访

于戈，东北大学计算机科学与工程学院教授，博士生导师，中国计算机学会会士，美国ACM 高级会员、IEEE 高级会员。1982 年、1986 年获得东北大学计算机学士学位和硕士学位，1996 年获得日本九州大学计算机博士学位。当前研究兴趣包括：分布式数据库系统，数据科学与大数据技术、区块链技术与应用、机器学习与智慧教育等。现兼任中国计算机学会信息系统专业委员会主任、《计算机学报》、《软件学报》等期刊编委。曾担任国务院学位委员会学科评议组成员、国家自然科学基金委员会评审专家组成员。发表高水平学术论文300余篇，出版“分布式数据库系统”等专著和教材 6 部，译著 4 部。获得“教育部自然科学二等奖”、“中国电子学会科技进步一等奖”、“中国计算机学会自然科学二等奖”等省部级科技奖 10 项、国家教学成果二等奖 1 项、省教学成果奖 3 项。曾获得“教育部跨世纪人才基金”和“中国高校青年教师奖”。

在 11 月 11 日的第五届江苏省计算机网络与云计算新技术研讨会中，于戈教授作了题为《事务即服务 (TaaS): 新型云原生数据库架构的研究进展与前瞻》的报告，概述了云基础设施、云原生架构的核心概念，以及“事务即服务”（TaaS）技术的特点和优势。TaaS 作为一种专为事务处理设计的模块，通过连接云原生数据库，带来了如支持复杂事务处理的扩展、单机事务处理引擎的分布式处理能力、多模型事务处理能力，以及更高的事务处理性能等优势。于教授进一步介绍了 TaaS 技术的研究进展，基于 TaaS 构建的新型云原生数据库架构，以及对未来工作的展望。

在报告结束后，我们有幸采访到了于戈教授。

Q1：您能简单介绍一下“云原生数据库架构”与传统数据库的区别，以及它的优势所在吗？

于戈教授：云原生数据库架构是为云环境特别设计的，与传统数据库相比，它在多个方面有显著的优势。首先，云原生数据库更加灵活，可以根据需求动态地分配和扩展资源。这是因为它是在云环境中构建的，因此能够充分利用云平台的计算和存储能力。其次，云原生

数据库通过存算分离，提高了数据处理的效率。这意味着查询处理和数据存储可以独立进行优化和扩展，适应不同的工作负载和应用场景。此外，云原生数据库还支持自动化和微服务架构，这使得系统维护和升级更加容易和高效。

Q2：您在报告中讲解了“事务即服务（TaaS）”技术，您是如何看待这项技术在未来数据库中的作用及其对云计算和数据存储领域的潜在影响的？您对这项技术未来有何展望呢？

于戈教授：“事务即服务”是一种新型的数据库架构，它将数据库事务处理作为服务来提供。这种架构的关键在于它实现了存储和计算的彻底分离，将普通的二级架构转变成三级架构，使得数据库可以更灵活地扩展和管理。在未来的数据库中，TaaS 将使得数据处理更加高效，特别是在大数据和实时数据分析领域。对于云计算而言，TaaS 提供了一种更为经济、高效的数据管理方法，能够根据实时需求动态调整资源，从而降低成本并提高性能。在数据存储领域，TaaS 能够提供更强大的数据处理能力，特别是在处理复杂查询和大规模数据集时。在云计算和数据存储领域，TaaS 预示着一种更高效、更经济的数据处理方式，特别是对于那些需要高度动态资源分配的应用。

应用 TaaS 技术的数据库架构并不需要对原本的云原生数据库架构做出很多的改变，且其拥有灵活、敏捷、易扩展、高性能等优势，我们的实验表明，与传统数据库相比，使用 TaaS 技术可以大幅提升处理效率和系统吞吐率。因此，我们也希望 TaaS 将成为未来数据库发展的一个重要方向，能有更多的普及。

于戈教授在报告中概述了云基础设施、云原生架构的核心概念，并重点介绍了 " 事务即服务 "（TaaS）技术的特点和优势。TaaS 作为专为事务处理设计的模块，通过连接云原生数据库，实现了支持复杂事务处理的扩展、分布式处理能力、多模型事务处理能力等优势。

采访中，于戈教授指出云原生数据库架构相对于传统数据库的优势在于灵活性、存算分离提高的数据处理效率以及自动化和微服务架构的支持。他进一步解释了 TaaS 技术在未来数据库中的作用和对于云计算和数据存储领域的潜在影响，并对未来该项技术的发展做出了积极的展望。

在云计算领域飞速发展的今天，数据库系统也是时下备受关注的一大热点，于戈教授的研究为数据库领域的创新和发展提供了有益的思路，深入而细致的讲解为我们了解该领域的前沿发展提供了独特而宝贵的视角，感谢于戈教授为我们带来如此精彩的报告，也再次感谢于教授能接受我们的采访并给出独到的见解！让我们以开放的心态和勇于创新的精神，共同探索未来道路！

《操作系统》课程体系建设教学成果交流

——东南大学董恺副教授

简介

“操作系统”是全国两会提及的“卡脖子”技术，《操作系统》课程是计算机大类的专业基础课程。本成果包含大二下开课的《操作系统（全英文）》与大三上开课的《操作系统专题实践》两门课程，属于理论与实践结合的一体化课程，课程内容均由董恺负责建设与讲授。本课程以 openEuler 开源操作系统为例，手把手地带领学生深入内核代码发现问题，再系统性地阐述操作系统与硬件、编译、网络等计算机系统相关领域理论知识，让学生深入浅出的理解算法与解决系统工程问题思路，激发学生自主学习兴趣，从而深度理解并掌握操作系统理论知识和实践技能。

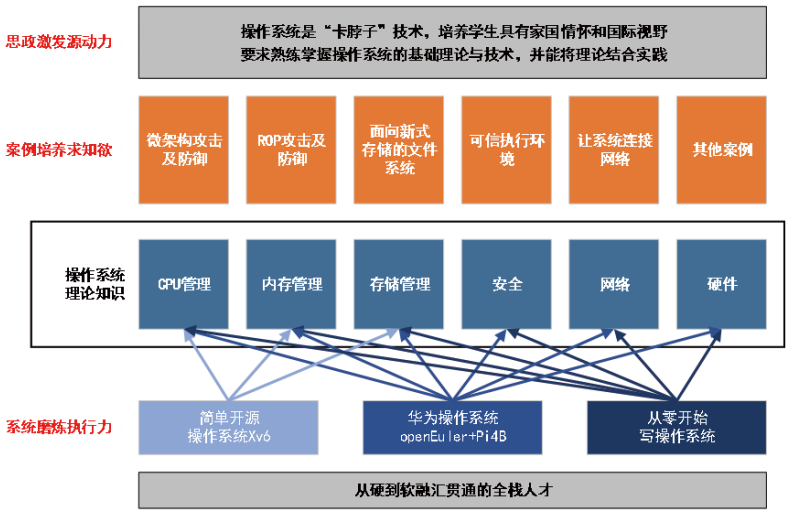
传统的操作系统课程教学，往往局限于课本知识，学生体会不到操作系统的开发之难、涉及之广。本课程强化内核层面的工程实践，教学过程中，锻炼学生自主查阅思考以及将操作系统及硬件、软件相关专业知识融汇贯通的能力，突出强化运用所学知识技术解决复杂工程问题的能力，同时赋予学生自立自强创新报国的正能量。

《操作系统（全英文）》课程 2016 年开课，包含 64 课时理论授课与研讨内容及 8 课时的实践内容。《操作系统专题实践》课程 2021 年开课，包含 32 课时的专题实践内容。随着产教融合的深入和实践内容的拓展，两门课程以 openEuler 为平台、以现实问题为案例，激发学生自主学习的强烈愿望和主观意愿，引导学生学会学习、学会思考、学会理论与实践结合，并激发学生的拼搏斗志和爱国热情。

主要解决的教学问题及解决教学问题的方法

成果的核心思路是精心设计课程教学内容，培养学生自主学习能力，全面掌握课程传授的知识和技能。在实际教学过程中，课程从思政、理论、实践三个方面进行了课程内容建设。

成果整体框架



1. 思政激发源动力

操作系统是被 2021 年全国两会提及的“卡脖子”技术，本课程引导学生了解中国计算机几十年的发展历程和当前我国自主操作系统现状以及我们肩负的创新突破的历史职责，分析全球工业 4.0 带来的挑战以及“中国制造 2025”带给计算机人的期待等，强化学生的家国情怀和建设国家报效社会的使命感、责任感，激发广大学生深入学习操作系统相关知识和技能的强烈愿望。

本课程从产业界对计算机人才的现实需求出发，向学生强调从硬到软融汇贯通的全栈人才的稀缺性和重要性。本课程鼓励学生从操作系统书本知识出发，结合相关领域知识，学习了解并尝试解决操作系统产业界和学术界关注的最新问题（如可信执行环境、微架构攻击防御、ROP 攻击防御等）。本课程与上下游相关课程互相关联配合组成课程群，学生以组队的形式进行实践锻炼与能力培养，包括

1) 应用硬件接口课程知识和操作系统课程知识，设计实现 BIOS，包括硬件驱动和 bootloader；

2) 应用操作系统课程知识，设计实现简单通用操作系统；

3) 应用网络课程知识和操作系统课程知识，设计实现网卡驱动和简单协议栈。

通过组队协作完成课程任务，培养学生集体意识、团队精神和协同能力。

2. 案例培养求知欲

产学融合、软硬贯通的课程教学改革引入了大量的全新内容。其中一小部分内容是对传统操作系统教学内容的自然拓展，在教学中融入较为顺利；但仍然有很大一部分内容则属于不同课程之间交叉地带内容，不仅是学生知识的盲点，也是传统操作系统教学所疏漏的地方。虽然从整体知识体系来说这部分内容非常重要，但是因为远离了学生已有知识体系的舒适区，大部分学生需要足够的引导和激励去进行相关内容的自主学习。本课程设计了专门的典型案

例引起学生的好奇心、探索欲和求知欲，让学生在案例与研讨环节围绕业界和学界的真实问题和进展，从听一个故事开始，最终初步了解关联的整个知识体系，促进学生对操作系统核心知识概念的理解掌握。课程使用的部分典型案例包括：

1) 微架构攻击及防御。这是系统领域近年来最难防御的攻击之一，也是 openEuler 系统研发部门的真实研究课题。其实现依赖于 CPU 指令集实现的微架构与 Cache 泄露攻击。案例研讨从该攻击的新闻报告开始，引导学生了解包括指令集、CPU、微架构、编译、Cache 等多方面知识。

2) ROP 攻击及防御。这是系统领域的另一大攻击，也是 openEuler 系统研发部门的真实研究课题。其实现依赖于对合法程序结构的感知以及对指令片段复用的攻击。案例研讨组织学生介绍、讨论历史上业界对该攻击及其各种变种的防御，深入浅出理解包括内存管理、编译、汇编、程序分析等多方面知识，最终引导学生实现 ASLR 作为防御机制，并鼓励学生参与针对最新 JIT-ROP 攻击的讨论和研究。

3) 可信执行环境。基于 openEuler 系统底层 ARM 架构的 TrustZone 环境进行研讨，涉及计算机体系结构、中断、虚拟机等多方面知识。

4) 面向固态硬盘、flash、分布式存储的文件系统。文件系统与存储管理是操作系统领域近年来发展变革较快的领域，也是 openEuler 系统研发部门需要重点调研的技术领域。除了针对典型的磁盘的文件系统进行介绍，也介绍相关领域近几年的各类进展。既包括业界的实现方法，也涵盖系统领域顶会 SOSP、OSDI 里的最新论文介绍的最新技术。

5) 让系统连接网络。高速网卡作为最特殊的一类硬件，最终在操作系统被抽象体现为文件进行管理。此部分内容真正打通操作系统课与网络课，让学生了解到硬件驱动、协议栈、Sockets 库等多方面内容。

3. 系统磨炼执行力

传统操作系统课程使用的实践方案，一般因为顾虑到学生学习进度不同故而往往实践任务设计较为简单，不能充分锻炼学生自主学习能力和系统工程实现能力。本课程针对实践内容的难易程度和学生学习进度，进行了很多尝试。

本课程团队精心设计了不同方案以兼顾不同进度的学生。学生可在如下三类实践任务中选择一类完成。

1) 基于一个简单的开源操作系统（Xv6）进行类似添加系统调用的实验内容。Xv6 系统是最简单的类 Unix 系统，针对 Xv6 的学习有助于未来深入学习 openEuler 系统，能够使学生真正进行系统级的 kernel hacking。

2) 基于树莓派 4B 开发板，使用 openEuler 系统刷机，并为系统添加、修改相应功能模块。工程量较之前一类略有提升，但难度略有下降。

3) 从零开始写操作系统，从 BIOS、bootloader 到 kernel，到磁盘管理、硬件管理，再到网络协议栈从而支持网络连接。工程量及难度最大，但学生的自主学习与系统能力可以得到充分锻炼。

创新点

传统的操作系统教学关注经典知识、软件管理、简单系统。本课程尝试将经典与潮流结合，贯通软件与硬件，兼顾简单与复杂系统。分理论与实验教学简述如下：

理论教学改革设计说明：

潮流：基于 openEuler 系统 , 设计开展关于 Arm 的可信开发环境 Trustzone 与微架构攻击（熔断、幽灵攻击）的研讨内容；学生对相关内容很感兴趣，在拓宽知识面的同时，能够巩固对经典知识的理解和掌握。

硬件：基于 openEuler 系统及其底层 Arm 架构，完善中断、镜像、互斥锁、硬盘管理等教学内容；操作系统本身是软件，是其他软件、进程线程的管理者，同时也是硬件资源的分配者和管理者，操作系统的学习需要将硬件软件融汇贯通。

复杂：基于 openEuler 系统，完善进程控制、线程切换、CFS 调度、信号量、文件系统等教学内容。基于简单系统学习理论上的算法思想，再基于复杂系统学习工程上的落地优化方案，两相比较更能加深理解。

实验教学改革设计说明：

统一购置便携开发板树莓派 4B 发给学生作为实验环境，并在其上进行 openEuler 的相关实验内容。学生对开发板和 openEuler 系统充满热情；选用统一的软、硬件平台便于实验教学的开展；便携的开发环境方便学生随身携带，自由进行探索和可选实验。

发给学生的树莓派 4B 开发板

方法探索：

首先，通过潮流的案例问题引起学生兴趣，最终落回到经典的知识；

其次，在介绍经典知识同时，从相关的硬件知识开始补充，让学生从硬到软地理解操作系统机制的目的和效果；

最后，补充了伴随现代复杂系统出现的操作系统相关知识（例如无锁并发等）。

教学过程中通过拓展内容，以及研讨、践行实验相关的教学改革，充分调动学生积极性。例如将原本的同步问题习题课与实验中的“添加系统调用”结合起来，让学生在实验中发现并解决问题，并给学生提供了充分的参考材料，激发学生自主学习热情，并学会学习。

数字化资源建设情况

本课程资源包括①课程大纲②课件③课内、课外习题④参考阅读材料⑤实验相关代码附件⑥实验指导手册⑦实验任务与说明⑧研讨内容及参考课件⑨复习材料，已全面实现课程资源数字化建设。

相关资源大部分建设在 Overleaf 平台（一个基于 Latex 的在线文档编辑平台），课程资源检索页见：https://www.overleaf.com/read/qxzksxcpqgnv，检索页包含所有资源链接。

课程在线资源检索页

面向社会开放情况

本课程资源根据资源链接即可开放获取，并随着每年教学活动随时更新。此外，董恺获评教育部 - 华为“智能基座”全国优秀教师后，受到智能基座工作组多次邀请申报“智能基座”优秀课件。全部课件及其他课程资源目前已改造建设完成，所有图片进行了重新绘制，案例进行了重新设计，确保自主知识产权，并已成功获评“智能基座”优秀课件，相关资源已在华为云上公开与推广。

关于进一步加强青年科技人才培养和使用的若干措施

新华社北京 8 月 27 日电 近日，中共中央办公厅、国务院办公厅印发了《关于进一步加强青年科技人才培养和使用的若干措施》（以下简称《若干措施》）。科技部负责同志对《若干措施》的出台背景、基本考虑和重要举措，以及如何保障各项措施扎实落地等进行解读。

一、《若干措施》出台的主要背景是什么，有何重要意义？

青年科技人才处于创新创造力的高峰期，是国家战略人才力量的重要组成部分。党中央高度重视青年科技人才队伍建设。习近平总书记多次就加强青年科技人才的培养和使用作出重要指示批示，要求把培育国家战略人才力量的政策重心放在青年科技人才上，给予青年人才更多的信任、更好的帮助、更有力的支持，支持青年人才挑大梁、当主角，造就规模宏大的青年科技人才队伍。党的二十大对加快建设包括青年科技人才在内的国家战略人才力量提出明确要求，中央人才工作会议对加强青年科技人才队伍建设作出具体部署。

青年科技人才已成为我国科技创新发展的生力军。党的十八大以来，我国青年科技人才规模快速增长，源源不断充实科技人才队伍。2012 年至 2021 年期间，我国研究与试验发展（R&D）人员数量由 416.7 万人增长到 858.1 万人，增加 441.4 万人，年均增长 7.67%。同期，自然科学领域博士毕业生总人数超过 45 万人，年均增长率 4.73%。近年来，我国博士后每年进站人数都超过 2.5 万人，其中 80% 集中在自然科学领域。同时，青年科技人才在国家重大科技任务实施中发挥越来越重要的作用。国家重点研发计划参研人员中，45 岁以下占比达 80% 以上。国家自然科学奖获奖者成果完成人的平均年龄已低于 45 岁。北斗导航、探月探火等重大战略科技任务的许多项目团队平均年龄都在 30 多岁。在人工智能、信息通信等新兴产业领域，优秀青年科技人才已成为技术创新的主力。

我国当代青年科技人才的职业生涯与到本世纪中叶全面建成社会主义现代化强国的时间高度契合。培养用好青年科技人才，对加快实现高水平科技自立自强，建设科技强国和人才强国意义重大。2022 年，科技部等五部门聚焦青年科研人员启动实施“减负行动 3.0”，有针对性地开展挑大梁、增机会、减考核、保时间、强身心五项行动，取得积极成效，起到先行先试的探索作用。《若干措施》在此基础上，进一步加大政策力度，采取更多突破性措施，必将对我国青年科技人才队伍建设起到重要推动作用。

二、制定《若干措施》有哪些基本考虑和主要举措？

《若干措施》的制定坚持以习近平总书记关于做好新时代人才工作的重要思想和关于科

技创新的重要论述为根本遵循，贯彻落实党的二十大精神及中央人才工作会议任务部署，针对当前青年科技人才面临的职业早期科研支持不够、成长平台和发展机会不足、符合青年科技人才特点的评价机制不完善、非科研负担重、生活压力大等突出问题，深入科研一线开展调查研究，广泛听取广大青年科技人才和各方意见建议，努力找出“真问题”、提准“实举措”，不求面面俱到，力求务实管用，突出可操作性，研究提出政策举措。

《若干措施》涉及青年科技人才培养和使用的方方面面，涵盖青年科技人才关心的主要问题。在具体措施上，既注重思想政治引领，又注重科研支持、职业发展、生活保障服务和身心健康关爱；既注重解决当前面临的迫切问题，又注重构建青年科技人才工作长效机制；既有原则性要求，也有量化要求。

一是加强思想政治引领。青年一代有理想、有担当，国家就有前途，民族就有希望。《若干措施》把加强对青年科技人才爱国奉献、科学报国的思想政治引领放在首要位置，坚持党对新时代青年科技人才工作的全面领导，强调用党的初心使命感召青年科技人才，激励引导青年科技人才大力弘扬科学家精神，传承“两弹一星”精神，在实现高水平科技自立自强和建设科技强国、人才强国实践中建功立业，在以中国式现代化全面推进中华民族伟大复兴进程中奉献青春和智慧。

二是强化职业早期支持。《若干措施》提出，充分发挥基本科研业务费对青年科技人才科研职业生涯的启动助推作用，根据实际需要、使用绩效和财政状况，逐步扩大中央高校、公益性科研院所基本科研业务费对青年科技人才的资助规模，完善并落实以绩效评价结果为主要依据的动态分配机制。基本科研业务费重点用于支持 35 岁以下青年科技人才开展自主研究，有条件的单位支持比例逐步提到不低于年度预算的 50%。

三是突出大胆使用。《若干措施》充分落实给予青年人才“更多的信任、更好的帮助、更有力的支持”的要求，从引导支持青年科技人才服务高质量发展，支持青年科技人才在国家重大科技任务中“挑大梁”、“当主角”，深入实施国家重点研发计划青年科学家项目，国家科技创新基地大力培养使用青年科技人才，更好发挥青年科技人才决策咨询作用等方面，赋予青年科技人才更多担纲领衔、脱颖而出的机会，出台了一系列针对性、可操作性强的举措，支持大胆使用青年科技人才，充分发挥青年科技人才作用。

四是促进国际化发展。《若干措施》提出加大青年科技人才出国学习交流支持力度，引导支持青年科技人才组织和参与国际学术交流活动，讲好新时代中国科技创新故事、中外科技合作故事，提升青年科技人才国际活跃度和影响力。

五是构建长效机制。《若干措施》既注重解决当前青年科技人才强烈期盼、亟待解决的急迫问题，又注重构建促进青年科技人才队伍健康稳定发展的长效工作机制。要求各级党委和政府把青年科技人才工作作为战略性工作，纳入本地区经济社会发展、人才队伍建设总体部署，建立多元化投入保障机制和常态化联系青年科技人才机制。要求用人单位切实落实培育造就拔尖创新人才的主体责任，结合单位实际制定具体落实举措，制定完善青年科技人才培养计划；建立和完善青年科技人才评价机制，提升自主评价能力；结合自身实际，采取适当方式提高职业早期青年科技人才待遇，加强对青年科技人才的关怀爱护。要求各类科技创新基地，如国家实验室、全国重点实验室、国家技术创新中心、国家临床医学研究中心等，

大力培养使用青年科技人才，积极推进科研项目负责人及科研骨干队伍年轻化，推动重要科研岗位更多由青年科技人才担任。

三、《若干措施》出台了哪些支持青年科技人才成长发展的“硬举措”？

注重务实管用，是《若干措施》起草工作着力把握的一个基本原则。其中不少措施都明确了定量化的要求，具有很强的可操作性。部分主要措施如下。

一是在支持青年科技人才在国家重大科技任务中“挑大梁”方面。规定国家重大科技任务、关键核心技术攻关和应急科技攻关大胆使用青年科技人才，40 岁以下青年科技人才担任项目（课题）负责人和骨干的比例原则上不低于 50%。鼓励青年科技人才跨学科、跨领域组建团队承担颠覆性技术创新任务，不纳入申请和承担国家科技计划项目的限项统计范围。稳步提高国家自然科学基金对青年科技人才的资助规模，将资助项目数占比保持在 45% 以上，支持青年科技人才开展原创、前沿、交叉科学问题研究。

二是在深入实施国家重点研发计划青年科学家项目方面。规定国家重点研发计划重点专项进一步扩大青年科学家项目比例，负责人申报年龄可放宽到 40 岁，并不设职称、学历限制。对组织实施高效、高质量完成任务目标的优秀青年科研团队通过直接委托进行接续支持。经费使用可实行包干制。

三是在国家科技创新基地大力培养使用青年科技人才方面。鼓励各类国家科技创新基地面向青年科技人才自主设立科研项目，由 40 岁以下青年科技人才领衔承担的比例原则上不低于 60%。青年科技人才的结构比例、领衔承担科研任务、取得重大原创成果等培养使用情况纳入科技创新基地绩效评估指标，加强绩效评估结果的应用。

四是在青年科技人才分类评价方面。明确要求不把论文数量和人才称号作为机构评价指标，避免层层分解为青年科技人才的考核评价指标。

四、在支持青年科技人才参与科技决策方面《若干措施》采取了哪些措施？

青年科技人才精力旺盛、思维活跃、知识更新快，一些优秀青年科技人才具有开阔的国际视野，能够及时准确把握前沿领域和新兴技术的变化趋势。吸纳更多青年科学家群体参与科技决策咨询，既有利于推动科技决策民主化、科学化，也是发现和培育战略科学家后备人才的重要途径。

《若干措施》积极回应广大青年科技人才的期盼和诉求，提出针对性举措。一是扩大科技评审专家库中青年科技人才的规模。要求高等学校、科研院所、企业等各类创新主体积极推荐活跃在科研一线、负责任讲信誉的高水平青年科技人才进入国家科技评审专家库。二是增加评审专家组成中青年科技人才的比例。规定国家科技计划等项目指南编制专家组，科技计划项目、人才计划、科技奖励等评审专家组，以及科研机构、科技创新基地等绩效评估专家组中，45 岁以下青年科技人才占比原则上不低于三分之一。三是推动各类学术组织吸纳更多青年科技人才。高层次科技战略咨询机制、各级各类学会组织应根据需要设立青年专业委

员会，推动理事会、专家委员会等打破职称、年龄限制，支持青年科技人才多层次参与学会组织治理运营。

五、在加强国家战略人才力量建设的大背景下，如何保障《若干措施》落实落地？

坚持党管人才原则和党中央对科技工作的集中统一领导，强化与相关部门和各地方的协同联动，统筹教育、科技、人才资源，加强对用人单位的指导和服务，调动各方积极性、主动性，推动青年科技人才工作体系化、创造性开展，确保各项措施落地实效并形成长效机制。

一是广泛深入开展政策宣传解读。组织新闻媒体和科技管理、人才等领域专家通过新闻报道、专题访谈、解读文章等形式进行广泛宣传和深入解读，提高政策知晓度和关注度，推动政策措施有效执行。

二是督促各地和用人单位进一步细化落实。督促各地把青年科技人才工作纳入经济社会发展、人才队伍建设总体部署，根据各地实际，加快建立多元化投入保障机制和常态化联系青年科技人才机制，抓好政策落实。鼓励指导用人单位切实落实主体责任，结合实际细化具体举措，健全工作体系和配套制度，提升青年科技人才培养使用能力。

三是开展动态评估和跟踪研究。组织专业机构适时对措施落实情况和效果开展评估，总结推广典型经验做法，分析解决难点问题。动态跟踪国际青年科技人才政策发展动向，持续开展青年科技人才重点问题和政策研究，推动青年科技人才工作机制不断完善。

学会动态

“走进南京大学”科技创新与体验研学活动成功举行

8 月 5 日“走进南京大学”科技创新与体验研学活动第一期，随着科技的高速发展和广泛应用，科技主题研学活动成为培养学生科技兴趣和创新能力的重要途径。参观科技企业或科技园区是一种直观了解科技发展的方式，学生可以亲眼目睹科技产品的研发过程，了解现代科技的应用场景，从而激发优秀中小学生的科研创新潜质，提升其动手实践能力。2023 年 8 月 5 日，江苏省计算机科普教育基地推荐了来自全省各地的 37 名高中学生参加了我学会联合南京大学计算机科学与技术系面向优秀中小学生开展的“走进南京大学”科技创新与体验研学活动。通过参与本次活动，学生们不仅丰富了暑期生活，更深刻认识到南京大学作为一所优秀学府的历史积淀，以及计算机科学技术在日常生活中的重要地位。此次科技实践体验活动也锻炼了学生们的科学思维和研究能力及培养了创新意识和科技实践能力。

分布式与自动化大数据智能分析算法与编程计算平台

作者：朱光辉 , 黄宜华

单位：南京大学计算机科学与技术系

论文摘要

大数据智能分析应用面临着计算效率和易用性两大基本问题和挑战。首先，在计算效率方面，大数据动摇了传统计算复杂性理论和方法，需要研究设计分布式数据挖掘与机器学习算法。除计算复杂性之外，还需考虑可并行度、通信开销等系统复杂性。其次，在易用性方面，现有大数据智能建模方法技术门槛高、大量依赖专家经验，需要研究高效的自动化机器学习（AutoML）建模方法，降低 AI 建模技术门槛，提升 AI 建模效率。另外，为了实现智能化建模与分布并行计算系统的交叉融合，需要构建一个融算法模型设计与大数据编程计算能力于一体的统一大数据智能分析编程计算支撑平台。为此，本文围绕上述关键问题与挑战，进行了系统深入的研究。AutoML 自动化机器学习方面的研究工作，在 NeurIPS、KDD 等国际顶级人工智能会议举办的 AutoML 大赛中，共计荣获 9 项大奖，6 次获得国际前三名，并在教育部主办的第五届中国“互联网 +”大学生创新创业大赛中，荣获全国金奖。另外，大规模分布式数据挖掘与机器学习以及 AutoML 自动化机器学习等相关技术研究成果已转让给华为、360 等多个 IT 企业落地应用。

专家推荐语

大数据智能分析已经成为 AI 赋能行业的重要应用模式，在未来将迎来高速增长。然而，与传统小规模数据场景下的智能化分析应用不同，大数据场景下的智能化分析应用，不再是单一的单机 AI 算法模型问题，而是大数据、大模型、大计算的融合，需要同时考虑算法模型设计、大数据处理以及高效的分布并行计算问题，这对大数据智能分析基础理论方法与关键技术研究，带来了一系列新的挑战和问题。作者在大数据分布式数据挖掘与机器学习、自动化机器学习以及大数据编程计算方法等基础理论与方法研究基础上，结合算法本身的重要性和技术挑战性以及业界的实际应用需求背景，首先选取了一系列基础常用、复杂性高、计算效率问题突出、且分布式算法设计难度大的数据挖掘与机器学习算法，开展了高效大规模分布式并行化数据挖掘与机器学习方法与算法研究；其次，开展了面向不同任务场景的高

效 AutoML 自动化机器学习方法与算法研究；最后，在融合分布式数据挖掘与机器学习以及 AutoML 自动化机器学习关键技术方法与算法的基础上，研究构建了高效易用的统一大数据智能分析编程计算方法与系统平台。论文研究工作在学术论文发表、国内外大赛获奖以及在产业界实际成果转化应用方面，均取得显著成效。

论文看点

1、在分布式数据挖掘与机器学习方面，提出了基于属性重排序的大规模分布式函数依赖发现算法、基于分布式数据并行的大规模并行化谱聚类算法、基于子森林均匀划分的分布式任务并行深度森林训练算法。

2、在 AutoML 自动化机器学习方面，提出了面向不同任务场景的 AutoML 大数据自动化机器学习算法，包括基于强化学习和贝叶斯优化的机器学习流水线自动化设计、资源受限场景下基于自适应连续过滤的模型选择以及终生学习场景下基于自适应加权集成的自动化终生学习；实现了高效的 AutoDL 自动化深度学习算法，包括基于渐进多保真度的超参数优化以及基于最小化离散性能偏差可微分架构搜索等。

3、在大数据智能分析和编程计算系统平台方面，实现了高效易用的跨平台统一大数据智能分析编程模型与可视化编程计算平台，上述关键基础理论与技术方法在该平台上取得了良好效果。

大规模分布式数据挖掘与机器学习算法

针对高效的大数据智能分析计算方法与算法的应用需求与研究目标，在大数据分布并行计算、相关数据挖掘与机器学习基础理论与方法研究基础上，围绕大数据场景下分布式数据挖掘与机器学习算法的复杂性分析、分布式算法设计及性能优化等关键科学问题，选取一系列基础常用、复杂性高、计算效率问题突出、且分布式算法设计技术难度大的数据挖掘与机器学习算法，研究实现高效的大规模分布式并行化方法与算法，包括大规模分布式函数依赖发现算法、大规模分布式并行化谱聚类算法以及分布式任务并行深度森林训练方法与算法，提升大数据智能分析的计算效率。

大规模分布式函数依赖发现：函数依赖（Functional Dependency，FD）用来表达关系表中属性之间的依赖关系，是大数据分析挖掘中一种基础数据结构。函数依赖 **X→A** 成立的条件是属性集合 X 的值能够唯一地决定属性 A 的值。也就是说，如果关系表中任意两个记录在属性集合 X 上具有相同的取值，那么在属性 A 上的取值必定也相同。函数依赖已广泛应用在许多数据管理和挖掘任务中，例如数据库规范化、查询优化、数据集成以及数据清洗等。因此，函数依赖发现也成为了数据库领域内的热点研究问题。在大规模数据场景下，函数依赖发现存在计算复杂度高、计算耗时长以及内存开销巨大的问题，导致已有的函数依赖算法不能高效地处理大规模数据。

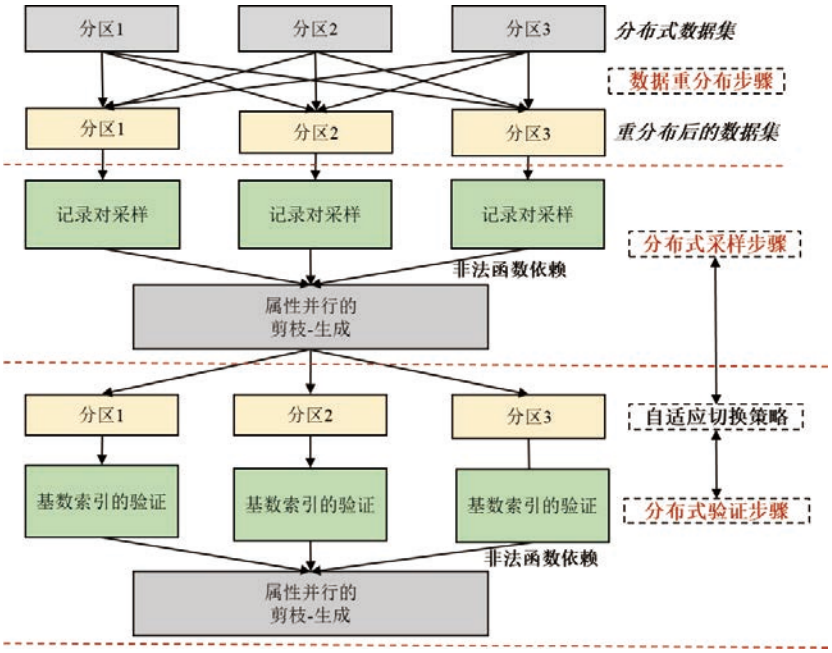


图 1：大规模分布式函数依赖发现算法 SmartFD 的总体流程

针对大规模数据集，本文研究提出一种高效可扩展的分布式函数依赖发现算法 SmartFD。为了实现负载均衡和提高计算资源利用率，研究提出一种基于属性重排序的算法框架。根据倾斜度和基数对属性排序，优先逐个处理倾斜度低、基数大的属性。其次，为了提升计算性能，研究提出高效的分布式算法 AFDD（Attribute-centric Functional Dependency Discovery），用于发现左部包含特定属性的所有函数依赖。AFDD 包含数据重分布、基于快采样和早聚合机制的分布式采样、基于索引的分布式 FD 验证三个步骤。另外，研究提出分布式采样和分布式验证之间自适应切换策略。对于倾斜度高、基数小的属性，研究提出 Batch AFDD 算法，其通过同时运行多个 AFDD 实例提升计算并发度和资源利用率。实验结果表明，SmartFD 性能优于已有函数依赖发现算法，可实现一到两个数量级的加速比，并具有良好的扩展性。

大规模分布式并行化谱聚类算法：谱聚类是一种基于谱图理论的聚类算法，将聚类问题转化为图的最优划分问题，适用于任意样本空间，且收敛于全局最优解，聚类效果优于 k-means 等传统聚类算法。但是，谱聚类算法计算流程复杂、计算复杂度高以及计算耗时长。大规模数据场景下，谱聚类算法的计算性能问题更加突出。现有的并行化谱聚类算法实现较为简单：只对相似度矩阵构建以及 k-means 聚类进行了简单的并行化，另外可扩展性和容错性较差，缺少基于 Spark 的高效并行化实现。

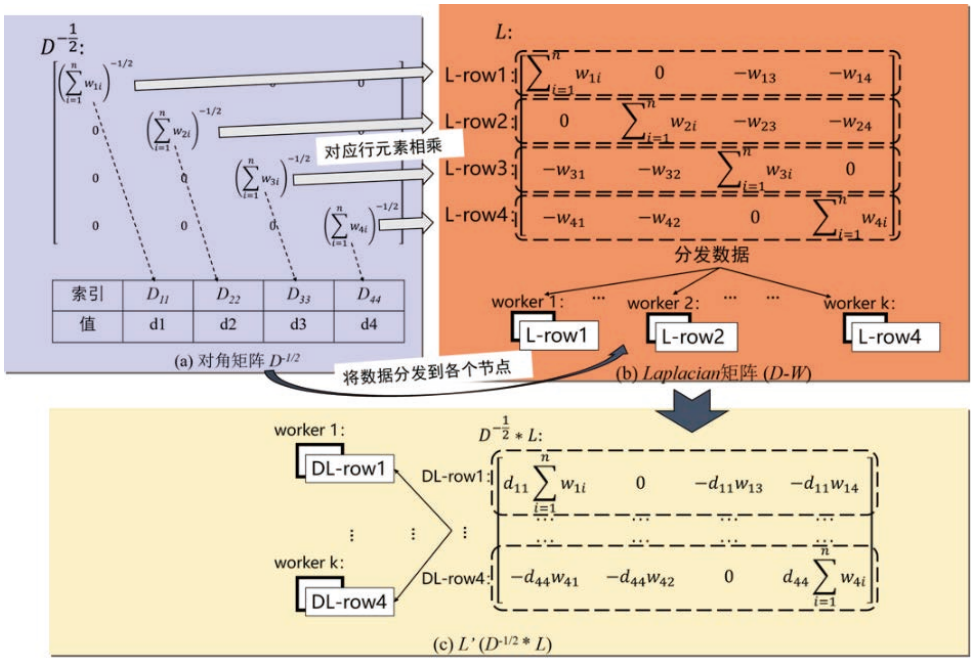


图 2：Laplacian 矩阵构建及正规化并行化

为此，本文研究并实现大规模分布式并行化谱聚类算法 SCoS（Spectral Clustering on Spark）。首先对谱聚类算法主要计算步骤进行并行化改造，包括相似度矩阵构建与稀疏过程的并行化、Laplacian 矩阵构建与正规化过程的并行化、正规化 Laplacian 矩阵特征向量计算的并行化、k-means 聚类并行化。其次，为了降低大规模谱聚类算法中相似度矩阵的构建开销，研究设计基于多轮迭代的样本间相似度并行计算算法，保证每两个样本之间的相似度只会计算一次。针对大规模 Laplacian 矩阵特征向量求解问题，基于 ScaLAPACK 实现高效的精确特征向量并行计算算法，同时研究实现近似特征向量并行计算算法。实验结果表明，SCoS 不仅具有良好的聚类效果，而且具有良好的数据可扩展性和节点可扩展性。

分布式任务并行深度森林训练算法：以深度神经网络为代表的深度学习在许多应用领域中取得了巨大的成功。但是，深度学习并不等同于深度神经网络。由于深度神经网络存在训练数据大、算力要求高及超参数调优难等问题，研究人员开始探索深度神经网络之外的深度学习模型。周志华等人总结提炼了深度学习模型成功的三大条件，即逐层处理、特征变换以及足够的模型复杂度，并在此基础上，研究提出了深度森林模型，并逐渐引起了学界和业界的广泛关注。已有深度森林系统 gcForest 计算性能低、计算耗时长，主要原因是 gcForest 中深度森林的训练过程是单机串行的。在每个级联层，一次仅训练一个森林，计算效率低，可扩展性差。巨大的训练时间开销阻碍了深度森林的研究与应用。因此，研究高效的深度森林训练算法与系统迫在眉睫。

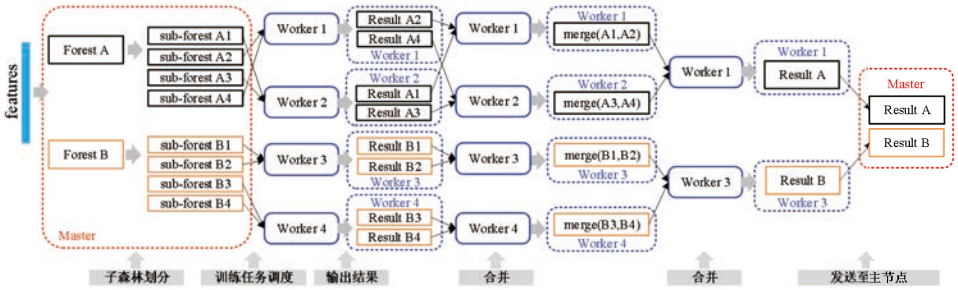


图 3： 基于子森林划分的任务并行训练

为此，本文基于分布式任务并行计算框架研究提出了高效可扩展的分布式深度森林训练算法，称为 ForestLayer。ForestLayer 将级联层中的每个森林划分成多个子森林。每个子森林的训练任务对应一个计算任务。然后，所有的子森林可以并行训练。另外，研究提出一种均匀的任务划分机制，不仅能够保证训练结果与原始单机方法的一致性，还能够减少总的训练时间开销。为了进一步提升 ForestLayer 的性能，研究提出延迟扫描、预池化以及部分传输三种系统层优化方法。为了提升 ForestLayer 的易用性，研究设计一系列高层易用的编程 API，用于深度森林的构建和训练。实验结果表明，ForestLayer 性能优于已有的深度森林系统 gcForest，能够实现 7 到 20.9 倍的加速比，并能够取得近线性的扩展性以及良好的负载均衡。

AutoML 自动化机器学习算法

针对高效易用的大数据智能建模方法应用需求和研究目标，在机器学习与 AutoML 自动化机器学习基础理论方法研究基础上，围绕自动化机器学习搜索建模方法有效性以及搜索计算效率优化等关键科学问题，研究提出面向不同任务场景的高效 AutoML 自动化机器学习与 AutoDL 自动化深度学习方法与算法，降低大数据 AI 建模分析的技术门槛，提升 AI 建模效率，解决大数据智能建模方法的易用性问题。

面向传统机器学习的 AutoML 算法：AutoML 技术经过近几年的研究探索，已取得了较大的进展，并达到了一定的自动化建模效果，但是在实际应用中还存在三个方面的不足。首先，在实际应用场景中，典型的机器学习模型往往是一个流水线，包含数据预处理、特征选择、算法选择、超参数调优以及模型评估等多个步骤，尚缺少面向机器学习流水线的自动化设计方法。其次，模型 / 超参数组合查找空间巨大，而且现实应用中，可用的 AutoML 时间资源有限，尚缺少资源受限场景下的 AutoML 方法。然后，在许多应用场景下种，数据分布会随着时间的推移发生变化，终生学习的目的就是能够捕获数据概念漂移，并通过持续改进机器学习模型以适应概念的变化，目前尚缺少终生学习场景下的 AutoML 方法。

在机器学习流水线自动化设计方面，本文研究提出一种基于强化学习和贝叶斯优化的流水线设计方法，简称为 RoboML。结合强化学习和贝叶斯优化的交替优化，同时解决机器学习流水线结构搜索和超参数调优两个子问题。首先，研究实现基于强化学习的机器学习流水线结构搜索，将流水线设计抽象为强化学习问题，并采用强化学习算法实现流水线的设计。其次，研究实现包含结构超参和算法超参的超参数优化，设计全局超参空间，以及固定

结构超参、搜索算法超参的贝叶斯优化算法。实验结果表明，RoboML 性能优于目前最优的 AutoML 系统 auto-sklearn。而且，随着时间预算的增加，性能优势更加明显。

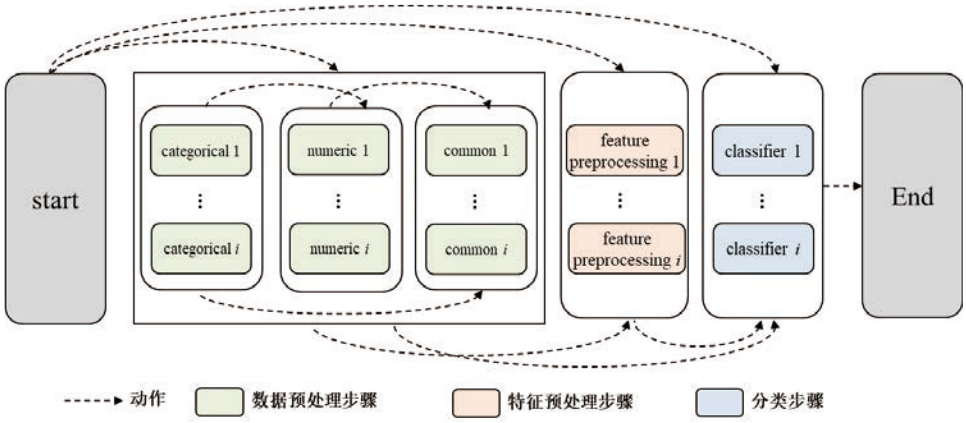


图 4：基于强化学习的机器学习流水线设计

在资源受限场景下的 AutoML 建模方法，研究提出一种基于自适应连续过滤的 AutoML 方法，简称为 BOASF。首先采用统一的多臂赌博机框架，将 AutoML 问题抽象为多臂赌博机问题，并在统一的多臂赌博机框架下实现模型选择和超参优化。为了提升资源受限场景下的 AutoML 性能，研究提出一种结合贝叶斯优化和自适应连续筛选的算法。通过对赌臂的抽象和定义，能够同时支持模型选择和超参数优化。另外，对于超参数优化，研究提出一种基于子空间划分的赌臂选取算法。实验结果表明，无论是模型选择还是超参数优化任务，BOASF 在不同的时间预算下均能够取得比已有 AutoML 方法更加优异的预测性能。

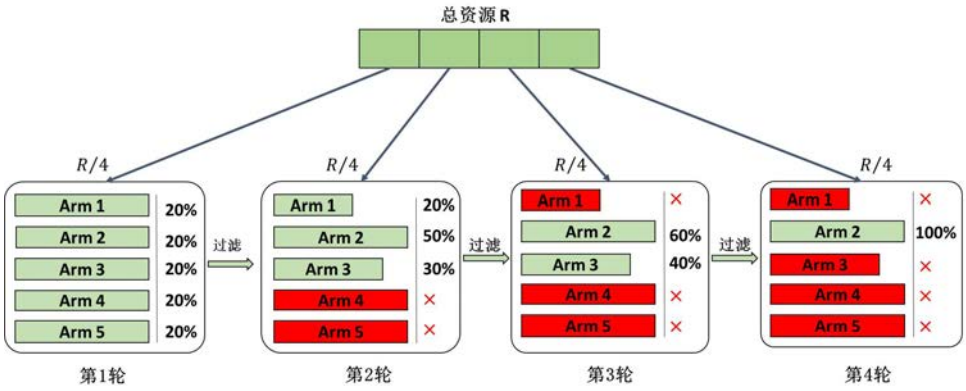


图 5：基于自适应连续过滤的自动化机器学习

在终生学习场景下的 AutoML 方面，研究提出面向终生学习场景的 AutoML 解决方案，简称 AutoLLE。首先研究提出基于自适应加权集成学习的算法框架，通过集成全局增量模型和局部集成模型，能够有效地捕获长期概念和短期概念。其次，研究设计基于时间窗口和误

差度量的模型权重自适应调整策略，适应最新概念，降低旧的概念对模型的影响。实验结果表明，AutoLLE 能够有效自动地捕捉概念漂移，提升模型预测性能。

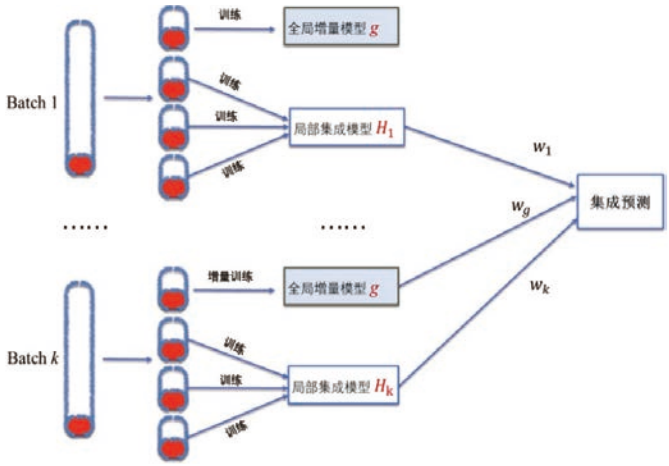


图 6：基于加权集成的自动化机器学习

面向深度学习的 AutoML 算法：近年来，国际上出现了自动化深度学习技术 AutoDL（Automatic Deep Learning），利用机器替代人工自动化地完成深度神经网络的超参数优化和架构搜索。然而，由于深度神经网络训练时间长，而且超参数以及网络架构搜索空间巨大，导致自动化深度学习面临着搜索耗时长、GPU 计算资源要求高等问题，难以在实际应用中落地。本文围绕深度神经网络超参数优化以及架构搜索两个关键问题，研究提出高效的自动化深度学习算法。

在深度神经网络超参数优化方面，研究提出一种基于渐进式多保真度评估的超参数优化算法，简称 FastHO，通过结合渐进多保真度优化和 Successive Halving 优化，能够尽可能早地过滤掉表现较差的超参数配置，并逐渐为剩余的超参数配置分配更多的资源，从而提升超参数优化的效率。另外，利用贝叶斯优化选择潜力较大的超参数配置作为初始超参数配置，实现超参数优化的热启动。实验结果表明，FastHO 性能优于目前已有的超参数优化方法，不仅能够加速深度神经网络超参数优化，还能够取得更加优异的随时性能和最终性能。

在深度神经网络架构搜索方面，研究提出一种最小化离散性能偏差的可微分网络架构搜索算法，简称 MGDARTS。首先，通过实验深入分析了以 DARTS 为代表的可微分神经网络架构搜索算法存在的离散性能偏差问题，主要包含同一边内算子权重区分度低以及同一边内主要算子聚集两个原因。针对同一边内算子权重区分度低的问题，研究采用饱和度更高的 tanh 函数作为算子权重函数，将超网络中每条边上的算子权重尽可能往 0 和 1 两个方向延伸。针对主导算子聚集的问题，研究提出一种对每条边上的算子权重之和进行整体性约束的方法。最后，研究设计随机固定规约单元的架构搜索方法，固定 Reduction Cell 的架构、仅搜索 Normal Cell 的架构，将搜索空间减少一半，降低整个网络的参数量。实验结果表明，

MGDARTS 能够有效降低离散性能偏差，而且性能优于已有的网络架构搜索算法。

高效易用的统一大数据智能分析编程计算方法与平台

针对方便易用的大数据智能分析编程方法与系统平台的应用需求与研究目标，在融合上述分布式数据挖掘与机器学习以及 AutoML 自动化机器学习关键技术方法与算法的基础上，研究设计并实现了一个支持多计算模式、高效易用的跨平台统一大数据智能分析编程模型与计算平台。首先，研究提出了覆盖表模型、矩阵模型、张量模型、图模型、流式数据模型等多种计算模型的跨平台统一大数据智能分析编程计算模型。在此基础上，研究设计了基于计算流图的大数据智能分析编程方法。其次，研究实现了统一大数据智能分析平台集成框架，允许以兼容并包的方式在底层选择集成不同计算模型对应的主流大数据智能分析系统，构建混合式一体化的大数据智能分析处理平台。然后，研究实现了统一大数据智能分析算法集成框架与自动化机器学习平台，通过设计平台无关的算法集成接口屏蔽不同算法平台的差异性。同时，研究实现同时支持传统机器学习和深度学习的自动化建模工具平台。最后，基于统一大数据智能分析与可视化编程系统平台，进行了实际应用验证。

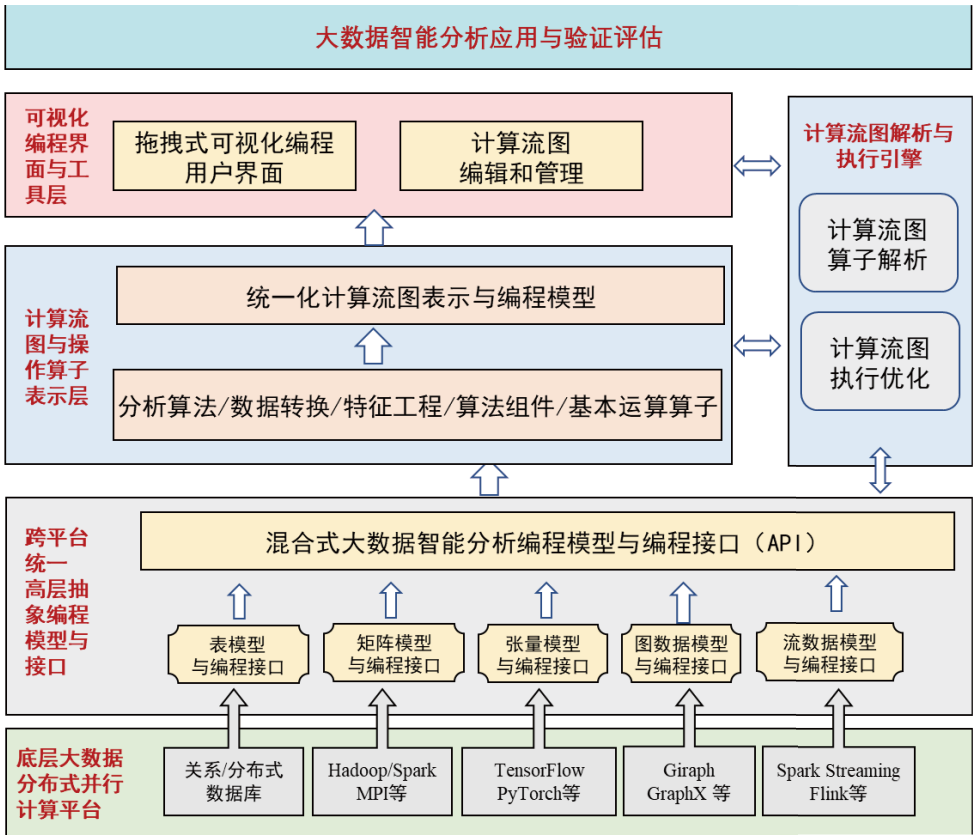


图 7：统一大数据智能分析编程计算方法与平台

总结与展望

本文围绕大数据分析挖掘和机器学习并行化算法（解决计算性能问题）、大数据自动化机器学习算法（解决智能建模方法的易用性问题）以及统一大数据智能分析编程计算模型与平台（解决编程方法和平台的易用性问题）三个方面展开深入研究。

1) 研究提出了一系列高效的大规模分布式数据挖掘与机器学习算法，包括基于属性重排序的大规模分布式函数依赖发现算法、基于分布式数据并行的大规模并行化谱聚类算法、基于子森林均匀划分的分布式任务并行深度森林训练算法。

2) 研究提出了面向不同应用场景的 AutoML 自动化机器学习算法，包括基于强化学习和贝叶斯优化的机器学习流水线自动化设计算法、资源受限场景下基于自适应连续过滤的 AutoML 算法、终生学习场景下基于自适应加权集成学习的 AutoML 算法。

3) 研究提出了高效的 AutoDL 自动化深度学习算法，包括结合渐进多保真度优化和 Successive Halving 优化的超参数优化方法、最小化离散性能偏差的可微分网络架构搜索算法。

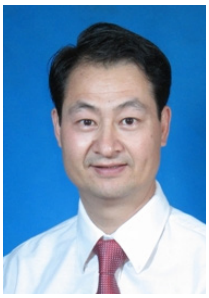
4) 在上述大数据分布式智能分析算法和关键技术研究基础上，研究设计并实现了统一大数据智能分析编程计算模型与平台。

作者简介：

第一作者：朱光辉，南京大学计算机科学与技术系助理研究员，主要研究方向为大数据分布式并行计算、自动化机器学习。

E-mail:zgh@nju.edu.cn

通信作者：



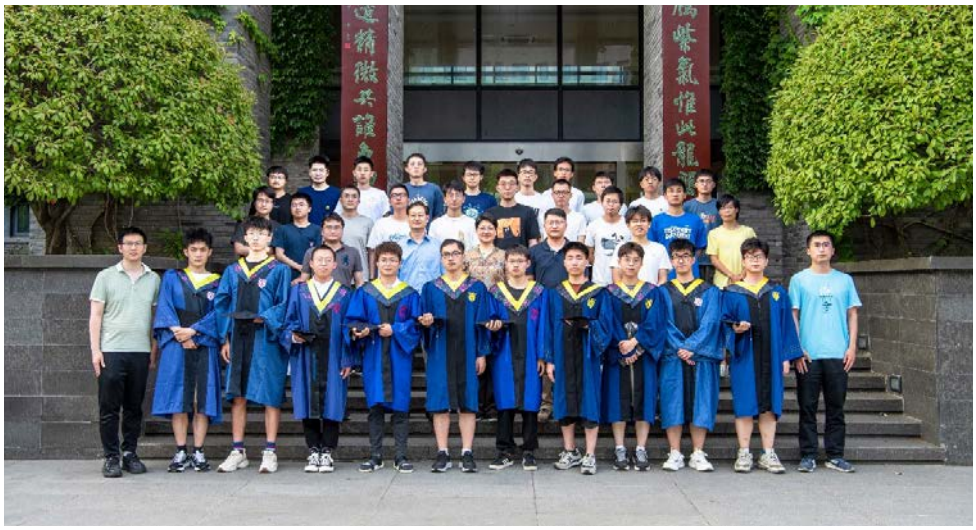
黄宜华，南京大学计算机科学与技术系教授，主要研究方向为大数据并行处理。

E-mail：yhuang@nju.edu.cn

团队简介

南京大学 PASA 大数据实验室（PASA：Parallel Algorithms, Systems and Applications for Big Data, 大数据并行算法、系统与应用）是南京大学计算机系专门从事系统化大数据技术研究教学的课题组，是全国最早系统化开展大数据技术研究教学的课题组之一。早在大数据还鲜为关注的 2009 年，实验室即已进入大数据并行处理技术领域。2009 年以来，实验室在大数据技术领域开展了一系列系统深入的研究开发工作，在大数据分布式大数据存储和查询技术与系统、大数据并行计算技术与编程平台、大数据系统性能优化与功能增强、大数据机器学习算法与系统、AutoML 自动化机器学习、大规模文本语义分析、AI4Sys、云计算、大数据行业应用等方面开展了广泛的研究，积累了系统的研究成果和技术基础。

团队主页：http://pasa-bigdata.nju.edu.cn/



学会动态

嵌入式人工智能与物联网综合培训在云南举行

7 月 22-28 日由江苏省计算机学会与云南省计算机学会联合主办，清华大学出版社、机械工业出版社及上海赛睿德电子有限公司协办，云南经济管理学院承办的嵌入式人工智能与物联网综合培训，于 2023 年 7 月 22-28 日在云南经济管理学院成功举办。来自全国 20 余家高校计算机类学科 40 余名教师参加了培训。此次培训邀请了苏州大学计算机科学与技术学院教授、江苏省计算机学会嵌入式系统与物联网专委会主任王宜怀教授主讲。此次培训邀请了苏州大学计算机科学与技术学院教授、江苏省计算机学会嵌入式系统与物联网专委会主任王宜怀教授主讲。

基于 Transformer 预训练模型的语言特征分析及其应用

作者：徐东钦，李军辉

单位：苏州大学计算机科学与技术学院

论文摘要

数据预训练模型的出现将自然语言处理带入了一个新的时代。对多种自然语言处理任务，借助预训练模型取得的性能已经远超过传统的方法。然而，目前预训练模型的可解释性较差，对句子语义的捕获能力还有待进一步分析。AMR（Abstract Meaning Representation，抽象语义表示）是一种基于图结构的语义表示方法，能够提供句子级别精准的语义表示。AMR 与预训练模型作为有机整体相辅相成。一方面，借助 AMR 可以帮助分析预训练模型捕获的语义信息；另一方面，利用预训练模型又能够协助 AMR 的研究。在此基础上，本文展开了对基于 Transformer 预训练模型的语言特征分析和预训练模型在 AMR 研究中的应用分析。其中，预训练模型在 AMR 研究中的应用分析的相关研究工作，已发表在 ACL、EMNLP、软件学报等国内外一流期刊和会议上。

专家推荐语

随着算力的增强和数据量的爆炸式增长，大规模预训练语言模型以其强大的语言理解和生成能力而备受关注。在面对自然语言的语义复杂性时，人类尚且存在着一定的困难，对预训练模型而言，准确识别不同语境下句子的正确语义，更是一个具有挑战性的任务。因而，如何提高这些大模型对语义的理解和处理，是目前自然语言处理领域的核心科学问题之一。作者文章的切入点较为新颖，选择了对大模型的分析与应用，一方面通过深入分析预训练模型、对其语言特征分别从单词和句子、句法和语义两种角度进行研究，分析多种预训练模型的优缺点；另一方面，作者基于对预训练模型的理解与研究，提出了多种基于序列到序列的预训练方法和微调技术，在多个 AMR 相关任务中取得优异的成绩，充分展现了预训练模型技术在语义和图结构任务中潜在着巨大能力。文章总体研究内容较为新颖充实，思路清晰，实验设计严谨，结果令人信服，在多个 AMR 任务上均取得了 state-of-the-art 的结果，具有较高的理论价值和应用前景。

论文看点

1、预训练模型的语言特征分析。目前对于预训练模型的相关研究分析主要关注在浅层次的语言特征，缺少对深层次的语言特征的评估。在此背景下，本文提出了句法和语义特征分析任务，分别从单词级别和句子级别，由浅入深地分析四种基于 Transformer 预训练模型对语言特征的捕获能力。

2、基于多任务预训练的 AMR 分析。与很多任务相似，AMR 分析的性能很大程度上受到了人工标注 AMR 数据集的规模限制。为解决这个问题，本文提出基于序列到序列预训练的 AMR 分析方法，联合了多个预训练任务帮助模型从源端句子中学习语言知识，同时提出使用多任务微调方法，以进一步提高 AMR 分析的性能。

3、基于多任务预训练的 AMR 文本生成。类似地，AMR 文本生成的性能也受到人工标注 AMR 数据集规模的影响。为了减少人工标注语料规模带来的限制，基于多个降噪自编码器任务，本文提出了基于多任务预训练的 AMR 文本生成方法，显著提高了 AMR 文本生成的性能。

4、零资源跨语言的 AMR 分析与文本生成。与英文 AMR 的研究不同，由于缺少人工标注数据，其他语言 AMR 的研究十分受限。为此，本文借助于预训练技术，提出了零资跨语言 AMR 分析与文本生成方案。具体地，将英语看作枢轴语言，提出了基于多任务学习的跨语言预训练方法。同时，基于得到的预训练模型，提出并比较了多种不同的微调方法。实验结果表明，借助该方法多个语言的 AMR 分析与文本生成的性能显著优于相关研究报告的水准。

预训练模型的语言特征分析

由于基于 Transformer 模型结构的预训练方法通过在大规模数据预训练、获得了巨大的成功，目前许多下游自然语言处理任务中，特别是包括机器翻译、语言生成等低资源任务，使用预训练模型已经渐渐成为一种常见做法。尽管先前的研究已经分析了 BERT 捕获句法和语义特征的能力，但这些研究关注的是浅层次的语言特征，而不是深层次的句法和语义特征，预训练捕获能够在多大程度上能够捕获深层次语言特征仍然是一个学术上亟待解决的问题。

本文则系统地评估和比较了四种常见的预训练方法在捕获浅层和深层语言特征的能力，与以往使用多种探测任务评估捕获语言特征程度的研究不同，本文将重点放在更细粒度、更具挑战性的单词级别和句子级别任务上，其包括词性标注、句法分析、语义标签和 AMR 分析。

本文使用了四种常见的基于 Transformer 的预训练方法，具体如下

1、屏蔽语言模型（Masked Language Model，MLM）使用了 Transformer 编码器结构，预测在源端序列中被随机屏蔽的单词。

2、屏蔽序列到序列预训练（Masked Sequence to Sequence Pre-training，MASS）采用了 Transformer 结构，编码器端随机屏蔽掉句子中一个连续片段，解码器则只预测被屏蔽的连续序列。

3、降噪自动编码器（Denoising Auto-encoder，DAE）则是一种在图象和自然语言处理中常见的方法，它通过重构噪声编码后的序列、将噪声序列“修复”回原本序列的方式来训练模型，在无标签数据集预训练、无监督机器翻译等任务有着广泛的引用。

4、神经机器翻译（Neural Machine Translation，NMT）致力于将源端语言序列转换为目标端语言序列。先前在句法分析的研究表明，神经机器翻译是一种非常有效的预训练方法。

方法	类别	源端输入	目标端输出
MLM	自监督	屏蔽序列	单词
MASS	自监督	屏蔽序列	序列片段
DAE	自监督	噪声序列	序列
NMT	有监督	序列	另一种语言的序列

此外，四种语言特征分析任务则如下

1、单词级别句法分析（词性标注）任务是预测每一个单词的词性标签。

2、句子级别句法分析（成分句法分析）任务是生成给定句子的成分句法树。

3、单词级别语义分析（语义标签）任务是预测每一个单词的语义标签。

4、句子级别语义分析（AMR 分析）任务是生成给定句子的语义图。

基于以上四种预训练方法和四种四种语言特征分析任务，本文在 WMT14 英语⇔德语的机器翻译数据集上展开了相关实验。实验表明，四种预训练方法在捕获句法特征上都有着相似能力，而在面对更具挑战性的语义特征却有着截然不同的表现。两种语义特征分析任务的性能差距比两种句法特征的性能差距更大，这表明了预训练模型对捕获语义特征更为敏感。一方面，NMT 和 DAE 在目标端预测完整的句子，能够更有效率地学习到源端句子的语义特征，另一方面，MLM 和 MASS 仅预测独立的单词或是句子片段，给模型带来了捕获语义特征能力下降的风险。进一步而言，MLM、MASS、DAE 都是基于修改源端输入句子的自监督任务，自监督方法中人工处理，如 [MASK] 标记、单词排序、单词删除等，在预训练阶段会影响源端真实数据的分布，在恢复句法和语义时会导致预训练和恢复阶段的不一致。幸运的是，有监督的 NMT 任务不存在这种不一致现象，在除了词性标注外的所有任务中取得了最高性能。此外，除词性标注外，预训练模型捕获到的大部分语言特征是重叠的，因此，结合两个预训练模型得到的提高非常有限。

基于多任务预训练的 AMR 分析

AMR 分析任务是将自然语言文本句子的语义表示为 AMR 图的形式。相关的研究表明，人工标注数据集的规模在一定程度上限制了 AMR 的研究，同时束缚了高性能 AMR 分析器的构建；同时，目前基于通用目的实现的预训练模型，如 BERT 等，在 AMR 分析等特定任务中可

能无法达到预期性能。本章节以基于序列到序列框架的 AMR 分析为研究对象，结合机器翻译、成分句法分析和 AMR 分析任务三个相关任务，提出了基于单任务和多任务联合训练的序列到序列预训练方法。此外，本章节将传统的微调方法延伸至多任务学习微调方法，该方法在保持预训练模型原有性能的基础上，进一步优化了 AMR 分析的性能。两个 AMR 数据集上的实验结果均表明，单任务和多任务联合训练模型显著地提高了 AMR 分析的性能，如 AMR2.0 数据集中的性能从 71.5 提高到 80.2，同时，本章节中最优性能达到了 81.4，是目前报告的最优值。本文并未使用复杂的模型，而是仅使用序列到序列模型实现了这一性能，这是在目前 AMR 分析研究中可行性较高的一种做法。

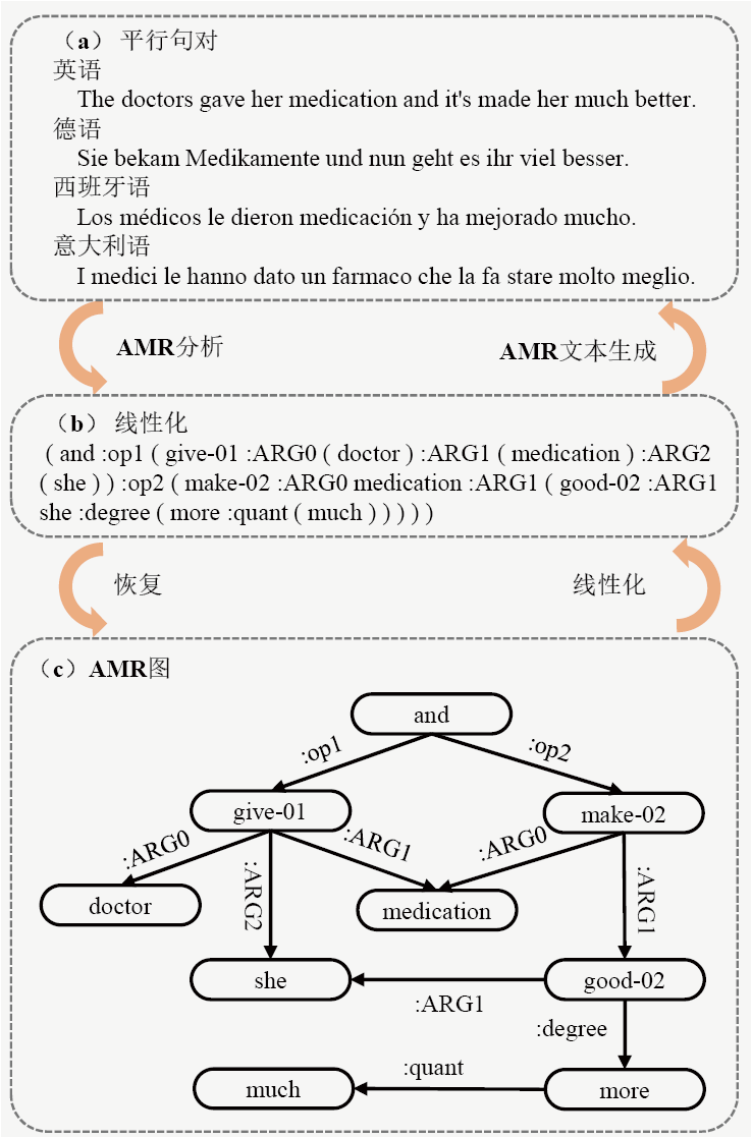


图 1 基于序列到序列方法的 AMR 分析（AMR Parsing）和 AMR 文本生成（AMR-to-Text）图示

1）基于序列到序列预训练的 AMR 分析方法：本文提出基于序列到序列预训练的 AMR 分析方法，联合如表 1 所示的三种预训练任务帮助模型从源端句子中学习语言知识。其中，机器翻译将英语翻译为其他语言的句子，有助于模型理解输入句子的语义；成分句法分析任务是获取句子的句法结构，准确地获取句子的句法结构能够帮助模型更好地分析其语义结构；最后借助大规模自动标注语料，即 AMR 分析任务能够提升模型的性能。

图 2 给出了在预训练阶段多任务联合学习框架。三个任务的源端均为自然语言句子，目标端各不相同，因此，在目标端序列的前端加入用于表示任务的特殊字符。给定一个预训练的序列到序列框架模型，可以直接在人工标注的 AMR 数据集上对其进行微调以训练 AMR 分析器。

任务	数据集	源端	目标端
机器翻译	人工标注数据	序列	另一种语言的序列
成分句法分析	自动标注数据	序列	成分句法树序列
AMR 分析	自动标注数据	序列	AMR 序列

表 1 三种基于序列到序列的预训练任务

2）基于预训练模型的微调方法：类似的，基于预训练阶段多任务联合学习框架，本文使用了两种微调方法对预训练模型进行微调，分别是朴素（Vanilla）微调方法和多任务（MTL）微调方法。Vanilla 微调方法是在预训练模型的基础上，直接在 AMR 训练语料上进行训练、优化模型的参数以获取 AMR 分析模型。MTL 微调方法则是在模型优化的每一个批次时，依次迭代在多个任务上更新模型参数，直至模型收敛。

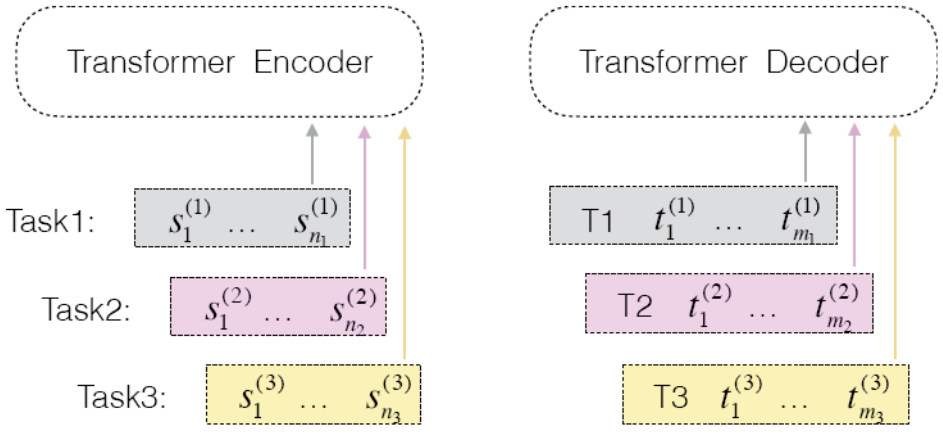


图 2 多任务联合学习框架示例

基于多任务预训练的 AMR 文本生成

AMR 文本生成的任务是给定 AMR 图，生成与其语义一致的文本。相关工作表明，人工标注语料的规模大小直接影响了 AMR 文本生成的性能。为了降低对人工标注语料的依赖，本文提出了基于多任务预训练的 AMR 文本生成方法。特别地，基于大规模自动标注 AMR 语料，本文提出与 AMR 文本生成任务相关的三个预训练任务，分别是 AMR 降噪自编码、句子降噪自编码以及 AMR 文本生成任务本身。此外，基于预训练模型，本文在朴素微调方法的基础上进一步提出了基于多任务训练的微调方法，使得最终模型不仅适用于 AMR 文本生成，同时还适用预训练任务。基于两个 AMR 标准数据集的实验结果表明，使用 0.39M 自动标注数据，本文提出的预训练方法能够大幅度提高 AMR 文本生成的性能，在 AMR2.0 和 AMR3.0 上分别提高了 12.27 和 7.57 个 BLEU 值，性能分别达到 40.30 和 38.97。其中，在 AMR2.0 上的性能为目前报告的最优值，并且首次报告了 AMR3.0 上的性能。

1）基于多任务预训练的 AMR 文本生成：降噪自编码模型被广泛应用于基于序列到序列框架的自然语言处理任务中，表明降噪自编码模型即能够较好地捕获源端语言的特征表示，同时也能够较好地生成目标端语言序列。本文基于降噪自编码模型和自动标注语料，根据 AMR 文本生成任务的特点，利用其源端（即 AMR 端）和目标端（即文本端）序列的特征，设计了表 2 所示的三种不同的单任务预训练方案，分别是基于源端序列的降噪自编码模型（简称 DAE(S)）、基于目标端序列的降噪自编码模型（简称 DAE(T)）和基于自动标注数据的 AMR 文本生成（简称 A2T）。本章延续与上一章节的多任务预训练方法，如表 3 所示，在训练数据的源端和目标端序列加入任务起始符号，以区分不同训练任务的源端和目标端序列。

任务	数据集	源端	目标端
DAE(S)	自动标注数据	加噪声的 AMR 序列	加噪声前的 AMR 序列
DAE(T)	英文句子数据	加噪声的英文句子	加噪声前的英文句子
A2T	自动标注数据	AMR 序列	英文句子

表 2 三种基于序列到序列的 AMR 文本生成预训练任务

预训练任务	源端起始符	目标端起始符
DAE(S)	<AMR BOS>	<AMR BOS>
DAE(T)	<TXT BOS>	<TXT BOS>
A2T	<AMR BOS>	<TXT BOS>

表 3 多任务预训练设置的源端和目标端起始符号

2）基于预训练模型的微调方法：本章继续上一章节中的 Vanilla 微调和 MTL 微调方法。特别地，在使用 MTL 微调模型进行 AMR 文本生成推理时，源端的输入起始符必须为 <AMR BOS>、目标端的输入起始符必须为 <TXT BOS>，与 AMR 文本生成预训练任务一致。

基于多任务预训练的 AMR 文本生成

由于匮乏的标注数据，对英语以外的语言而言，AMR 的研究相对受限的同时具有一定的挑战性。利用英语 AMR 数据集和英语到其他语言的平行语料，本文提出了一种基于多任务学习的预训练方法，用于零资源跨语言的 AMR 分析和文本生成任务研究。具体地来说，借助于联合机器翻译、AMR 分析和文本生成三个任务，令模型在训练英文的 AMR 分析和文本生成时捕获到的知识能够转换到到其他语言中。此外，本章进一步探索多任务学习微调，提出了四种不同的微调方法，例如，Vanilla 微调方法，One-for-All MTL 微调方法，Targeted MTL 微调方法，和基于 Teacher-Student 框架的 MTL 微调方法。在非英语语种的多语言 AMR 分析和文本生成的实验结果表明，本章方法的性能显著优于基于预训练的高性能基准模型，极大地提高了现有的研究水平。具体地来说，在 LDC2020T07 测试集中，本文方法在德语、西班牙语和意大利语的 AMR 分析中获得了 70.45、71.76 和 70.80 的 Smatch F1 值，AMR 文本生成中则取得了 25.69、31.36 和 28.42 的 BLEU 值。

1）零资源跨语言的预训练方法：借助英语 AMR 数据集和英语到其他语种 X 的平行语料（在本文中，X 为德语、西班牙语、意大利语），本文训练了 AMR 分析器，并将平行语料中的英文自动解析为 AMR 图，将二元平行数据集 $\mathcal{T}$ =($\mathcal{T}^{EN}$, $\mathcal{T}^{X}$) 拓展为三元平行数据集 $\mathcal{T}$ =($\mathcal{T}^{EN}$, $\mathcal{T}^{X}$, $\mathcal{T}^{AMR}$) ，并提出了基于多任务学习的跨语言预训练方法，包括了机器翻译、AMR 分析和 AMR 文本生成三种类型的任务、六种不同的预训练任务，分别为，“英语⇔ X”的双向翻译任务，“英语⇔ AMR”的英语 AMR 分析与 AMR 文本生成任务，“X ⇔ AMR”的 X 语种 AMR 分析与 AMR 文本生成任务。

序列种类	序列起始符示范
英　　语	<en> 英语句子
德　　语	<de> 德语句子
西班牙语	<es> 德语句子
意大利语	<it> 德语句子
AMR	<amr> 线性化 AMR

表 4 用于区分不同任务的起始符号

为了在序列到序列模型中同时训练上述六个预训练任务，本章延续前面章节的多任务学习方法，如表 4 所示，在训练数据的源端和目标端序列分别加入任务起始符号，以区分不同训练任务的源端和目标端序列。

2）零资源微调方法：为了微调预训练模型，本章基于英语标注的 AMR 数据集构建微调语料库。与获取三元平行数据集 $\mathcal{T}$ 类似，给定标注的 AMR 数据集 $(\mathcal{F}^{EN};\mathcal{F}^{AMR})$，本章使用“英语⇒X”的翻译器将英语语句翻译成 X 语种语句，从而得到三元平行数据集 $\mathcal{F}=(\mathcal{F}^{EN},\mathcal{F}^{X};\mathcal{F}^{AMR})$。基于三元平行数据集 $\mathcal{F}$，本章探讨并比较四种不同的微调方法，分别为 Vanilla 微调方法、One-for-All MTL 微调方法、Targeted MTL 微调方法和基于 Teacher-Student 框架的 MTL 微调方法。其中，One-for-All MTL 微调方法即对所有微调任务同步地微调。

Targeted MTL 微调方法指有选择地微调与目标任务相关的任务。以德语 AMR 分析为例，德语 AMR 分析作为微调阶段的主要任务，而德语⇒英语的翻译任务作为辅助任务。该辅助任务能够提高编码器从德语语句中捕获语义信息。对于德语 AMR 文本生成而言，则选择英语⇒德语的翻译任务作为辅助微调任务则有利于解码器生成更加流利的德语语句。

受 Teacher-Student 框架的启发，本章借助于使用更强的微调任务来辅助较弱的目标任务、缓解噪声带来的影响。例如，基于人工标注数据集训练的英文 AMR 分析具有较好的性能，因此可以将英文的 AMR 分析作为 Teacher，辅助 Student 任务德语 AMR 分析。

以德语 AMR 分析为例，记 E，G，A 分别为英语、德语和 AMR，并记 $(e;g;a)$ 为一个三元组实例。令英文的 AMR 分析为德语 AMR 析 $\mathrm{E}\rightarrow\mathrm{A}$ 的 Teacher，则假设给定解码获得的部分 AMR 序列 $a_{<i}$，g 从中解码获取目标 AMR 字符 a_i 的概率应该接近从 e 解码获取该字符的概率。在这种假设下，Student 模型可以通过联合训练从 Teacher 模型中学习相应的知识，具体的定义如下，

$$\mathcal{J}(\theta_{\mathrm{G}\rightarrow\mathrm{A}})=\sum_{(e;g;a)\in D_{E;G;A}} J\left(\boldsymbol{e},\ \boldsymbol{g},\ \boldsymbol{a},\ \hat{\theta}_{\mathrm{E}\rightarrow\mathrm{A}},\ \theta_{\mathrm{G}\rightarrow\mathrm{A}}\right)+\mathrm{L}_{\theta_{\mathrm{G}\rightarrow\mathrm{A}}}(\boldsymbol{a}|\boldsymbol{g})$$

其中，$\mathrm{D}_{\mathrm{E},\ \mathrm{G},\ \mathrm{A}}$ 表示三元平行数据集，$\hat{\theta}_{\mathrm{E}\rightarrow\mathrm{A}}$ 表示已学习的英文 AMR 分析模型参数。同时，$\mathrm{L}_{\theta_{\mathrm{G}\rightarrow\mathrm{A}}}(a\mid g)$ 表示 g 到 a 翻译过程的对数似然函数，函数的具体定义公式如下

$$J\left(\boldsymbol{e},\ \boldsymbol{g},\ \boldsymbol{a},\ \hat{\theta}_{\mathrm{E}\rightarrow\mathrm{A}},\ \theta_{\mathrm{G}\rightarrow\mathrm{A}}\right)=\sum_{i=1}^{|a|} KL(P(a|\boldsymbol{e},\ \boldsymbol{a}_{<i},\ \hat{\theta}_{\mathrm{E}\rightarrow\mathrm{A}})||P(a|\boldsymbol{g},\ \boldsymbol{a}_{<i},\ \theta_{\mathrm{G}\rightarrow\mathrm{A}},\ )$$

其中，$KL(\cdot\,\|\,\cdot)$ 表示两个分布的 KL 散度，$|a|$ 表示 AMR 的序列长度。

而德语 AMR 文本生成类似，考虑到英语⇒德语翻译的性能远高于德语 AMR 文本生成的性能，可以将英语⇒德语翻译作为 Teacher、德语 AMR 文本生成作为 Student，并假设给定解码获得的部分德语语句 $g_{<i}$，从 a 中解码获取目标德语字符 g_i 的概率应该接近从 e 解码获取该字符的概率。由于德语 AMR 文本生成的联合训练目标与德语 AMR 分析的目标函数相似，此处不再阐述。

总结与展望

本文的研究主要是基于 Transformer 预训练模型的语言特征分析及其应用两个方面。其中，基于 Transformer 预训练模型的应用包括基于序列到序列框架预训练的 AMR 分析、基于多任务预训练的 AMR 文本生成研究和零资源跨语言 AMR 分析与文本生成研究三种任务。主要的研究工作总结如下：

1) 预训练模型的语言特征分析。目前，相关研究对预训练模型的分析大多使用多种浅层的探测任务评估捕获语言特征的程度。本文则将重点放在更细粒度、更具挑战性的句法和语义分析任务上，从单词级别和句子级别的角度系统地评估和比较了四种常见的基于 Transformer 的预训练模型在捕获浅层和深层语言特征的能力。实验结果表明，四种预训练方法具有相似的捕获句法特征的能力，却在学习语义特征时表现出不同的性能。其中，NMT 方法的性能最优，该实验结果为后续章节的实验提供了重要的指导意义。

2）预训练模型在 AMR 研究中的应用。针对 AMR 研究中的 AMR 分析和 AMR 文本生成任务，提出了基于多任务预训练模型的解决方案，分别针对 AMR 分析和 AMR 文本生成，分别制定与其相关的预训练任务，能够帮助模型理解自然语言句子和 AMR 的图结构。同时，针对其他非英语语言 AMR 研究中的零资源问题（即不存在训练标注语料），提出基于多任务预训练的零资源跨语言 AMR 解析。

本文对基于 Transformer 预训练模型学习到的语言特征进行分析，并分三章重点探索了基于 Transformer 预训练模型在 AMR 图结构任务中的应用，虽然在多个任务中都取得了显著的性能提升，但是仍然存在很多不足之处。在未来的工作中，可以考虑从以下几个方面展开：

1) 在线性化 AMR 图结构时，本文使用深度优先遍历的方法获取 AMR 图的线性序列。然而，AMR 的部分图结构信息会在线性化时丢失。首先，图结构中距离较近的兄弟节点在经过线性化后可能会导致异常增大，为模型捕获图结构信息带来困难。其次，在 AMR 线性化时，本文将重入节点变量替换为重入节点概念、作为多个不同的单词处理，该处理方法会导致模型可能无法识别重入节点的信息。因此，在未来的工作中将探索如何更好地对图结构建模，例如，对重入节点变量额外的标记，如 person 标记为 person_1，用以指示该位置的节点代表同一个重入节点。

2）本文预训练方法中使用的模型均使用 Transformer-base 模型设置，该模型设置与 Transformer-big 或者更大规模的模型设置的有显著的性能差距。此外，批处理规模的对模型性能影响十分巨大，而本文为节约训练时间，实验中批处理规模 (Batch Size) 均远小于 Vaswani 等人使用的规模。例如，仅仅将德语 AMR 分析的预训练基准模型的预训练批处理规模从 4096 提高到 8192 时，性能就能从 64.90 提高至 67.58，显著地提升了 2.68 个 Smatch F1，简单的设置带来的提升却是十分令人震撼的。在未来的研究，可以从扩大模型、批处理和训练语料等设置的规模的角度出发，探索本文方法在 AMR 研究中性能的上限。

3）目前 T5、BART 等超大规模的基于序列到序列框架的预训练模型的极致性能有目共睹，本文实验中的预训练模型与其相比是小巫见大巫。如何复用 T5、BART 等预训练模型、根据图结构特点的改进预训练模型，使其能够应用于 AMR 分析与文本生成是一个值得研究的课题。

4）由于缺乏训练数据，多语言的 AMR 研究目前仍然有很大的提升空间，如第二点中提到的增加批处理规模。接下来的研究工作中，准备借鉴多语言机器翻译与无监督机器翻译中众多技术辅助多语言 AMR 的研究。

5）虽然本文提出的方法显著的提高了多语言 AMR 分析的性能，但是在细粒度评测中 Reentrancies 和 Negations 的性能依旧处于较低水平，今后的研究可以进一步探索该方面的研究。

6）本文通过合并词表、共享词嵌入权重的方法可以大幅度减小模型参数。然而，本文讨论的语言侧重于日耳曼语系，合并词表对中文而言并不能有效减小词表规模。从实用性的角度出发，中文 AMR 研究对算力有一定的要求。从科研的角度来看，通过本文第四至六章的方法均可以在中文 AMR 研究中获得不错的性能。

作者简介：

第一作者：

徐东钦，苏州大学研究生，主要研究方向为自然语言处理。

E-mail: xdqck@live.com

通讯作者：

李军辉，苏州大学副教授，主要研究方向为自然语言处理。

E-mail: jhli@suda.edu.cn

耕耘网络安全领域十余年，矢志打造“工控安全网络综合靶场”

——2022 年江苏省计算机学会优秀科技工作者傅涛博士

个人简介

傅涛，2002 年本科毕业于南京理工大学计算机科学与技术系；2008 年获得南京理工大学网络信息安全方向博士学位。先后就职于江苏省电子商务集团和南大苏富特集团（香港上市公司，HK：8045），具有 18 年从事软件研发和科技公司运营管理的实务经验。傅涛先生 2019 年入选国家中组部重点人才工程、先后荣获国家科技部“创新人才推进计划科技创新创业人才”、江苏省“333 高层次人才培养工程”、江苏省组织部“江苏省科技企业家”、江苏省人社厅“六大人才高峰计划”、南京市科技人才局“中青年拔尖人才”和南京市科技局“科技顶尖专家集聚计划”。

傅涛是博智安全科技股份有限公司创始人，现任董事长。在傅涛的带领下，博智安全已研发拥有“博智孪生仿真靶场（ETER）”、“工控盾”、“安全御”、“维思得”产品线，并服务了上千家部队军工、政府和事业单位、央企国企、民营企业等客户。其中博智安全核心产品“博智孪生仿真靶场（ETER）”在电力、轨道交通、能源、制造、电信等领域的靶场建设都有深入的研究，积累了丰富的经验。通过模拟真实的网络攻防作战，针对网络攻击的作战手段进行试验，以实现网络空间作战能力的技能提升，打赢未来网络空间战争。

坚守初心，科研学术取得突出成果

傅涛先后主持十余项国家和省部级课题和专项，包括国家工信部“工业控制系统安全核心技术能力提升及测试验证”、国家科技部“面向大数据基于智能分析的信息安全管理平台”、江苏省科技厅“基于信息物理融合技术的工控运维网关系统合作研发”和江苏省工信厅“基于意图驱使的工业信息安全靶场平台”等。傅涛先生在科研过程中，参与制定多项国家和行业信息安全相关标准，以第一作者在《软件》、《计算机科学》等核心期刊发表《基于改进 Apriori 算法的入侵检测数据挖掘模型研究》、《基于 PSO 的 k-means 算法及其在网络入侵检测中的应用》等论文十余篇，作为第一发明人拥有信息安全相关发明专利 8 项，软著 168 项。傅涛先后获得中国人民解放军科技进步奖 1 次、江苏省科技进步奖 1 次、南京市科技进步奖 4 次，并荣获 ISC 十周年代表性人物，被认定为近十年中推动网络安全行业高质量发展的代表性人物，为数字安全未来发展的预判分析给予有效引导。

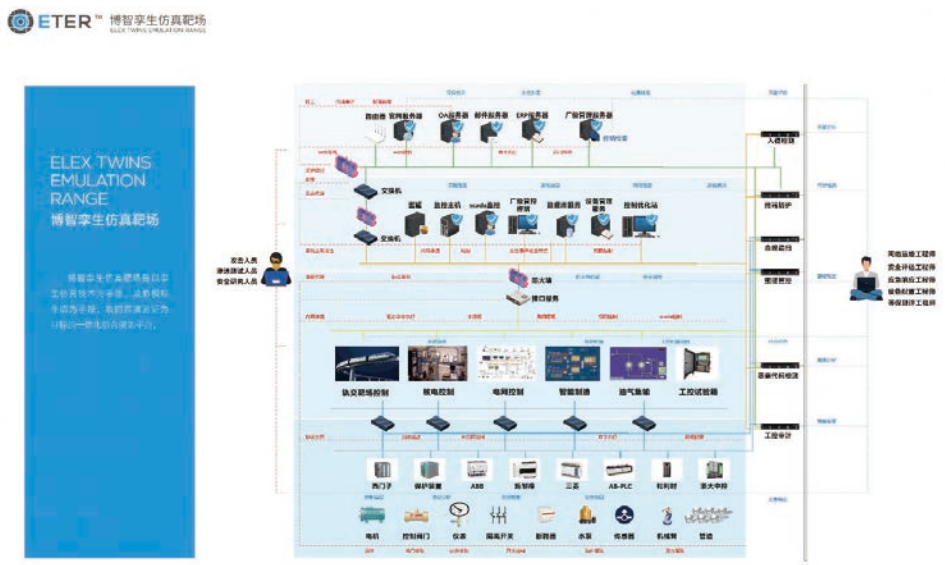


图 1 博智孪生仿真靶场

图 2 江苏省委常委、南京市委书记韩立明调研考察博智安全

图 3 傅涛受南京新闻综合频道等多家媒体采访

图 4 傅涛荣获 ISC 十年网安代表人物

始终坚持“为国家安全锻矛铸盾”企业使命

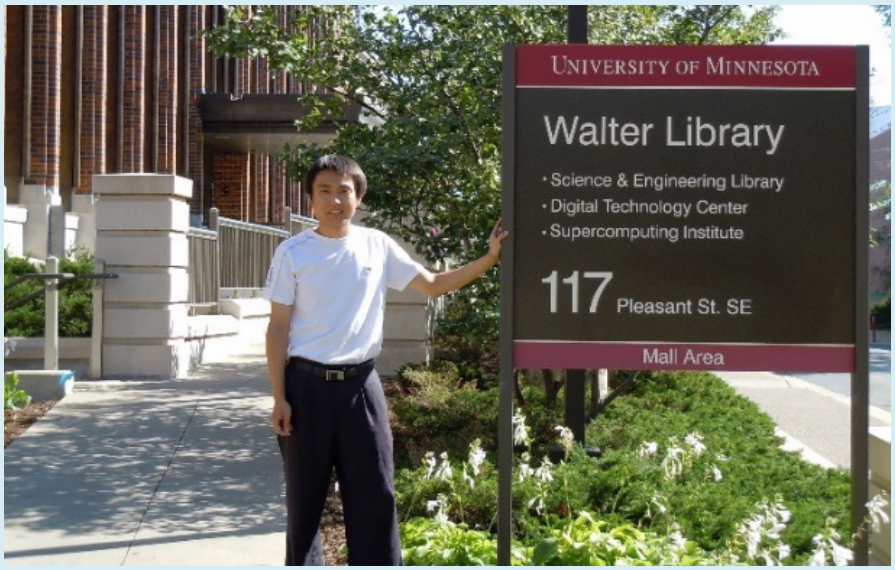
博智安全在网络信息安全领域耕耘多年，已发展为 500+ 人员规模的企业，现已是工控安全及网络空间靶场领域的代表企业。目前已获得国家工信部网络安全技术应用试点示范单位、中国网安产业竞争力 50 强、CNCERT 网络安全应急服务支撑单位、工业信息安全监测预警网络建设支撑机构、国家信息安全漏洞共享平台原创漏洞发现优秀单位、国家工业信息安全漏洞库优秀成员单位、国家工业信息安全应急服务支撑单位、中国信通院 2021 工业互联网安全人才联合培养计划（第一批）合作伙伴、首批工控安全防护能力贯标技术服务机构、江苏省工控安全工程研究中心、江苏省认定软件企业技术中心、江苏省网络靶场工程技术研究中心、CMMI 五级、ITSS 二级标准化认证等多项荣誉认定。

物联感知计算 赋能智慧矿山

——2022 年江苏省计算机学会青年科技奖陈朋朋教授个人简介

个人简介

陈朋朋，中国矿业大学计算机学院教授 / 博导 / 副院长，曾在美国明尼苏达大学双城校区进行博士联合培养，入选江苏省 333 高层次人才培养工程及江苏省科协青年科技人才托举工程。主要从事物联网方向的领域，共发表论文 50 余篇，包括国际一流学术会议 Infocom、MobiSys、ICDCS 以及 IEEE TVT、IEEE TIM、Pattern Recognition 等学术期刊，以第一发明人授权澳大利亚发明专利 1 项、中国发明专利 10 余项，参与制定国际标准 IEEE 1851 和 2735，荣获教育部科技进步二等奖、中国专利优秀奖、中国煤炭工业科学技术二等奖、中国物联网学术会议最佳论文奖及中国物联网学术会议优秀论文奖等奖励。



聚焦前沿，推进基础理论研究发展

陈朋朋聚焦“矿山智能感知计算”主题，攻关物联网感知、传输与处理技术瓶颈背后的基础理论问题，破解地下矿山安全生产监测中的科学技术难题，研究异质物联网自适应跨网传输理论，突破复杂矿山高精度泛在感知定位算法，构建地下矿山大型装备监测预警模型，取得了一系列创新性研究成果。近年来先后主持 3 项国家自然科学基金项目，获得中国物联网学术会议最佳论文奖、中国物联网学术会议优秀论文奖等荣誉。

图 1 地下移动目标高精度定位

产学研用，理论研究转化实际应用

陈朋朋积极促进研究成果向实际应用转化，响应国家在深地开发方面的发展战略需求，将物联网与地下矿山工程有机交叉，研制了煤矿机电设备数据汇聚解析网关，研发了矿山长距离运输提升系统全状态网络化监测平台，相关技术荣获教育部科技进步二等奖和中国专利优秀奖，大幅降低了矿山生产时间和成本，带来了显著的经济效益和安全效益。相关系统研究成果以第一发明人授权澳大利亚发明专利 1 件、中国发明专利 10 余项，其中 2 项专利已转让至企业。

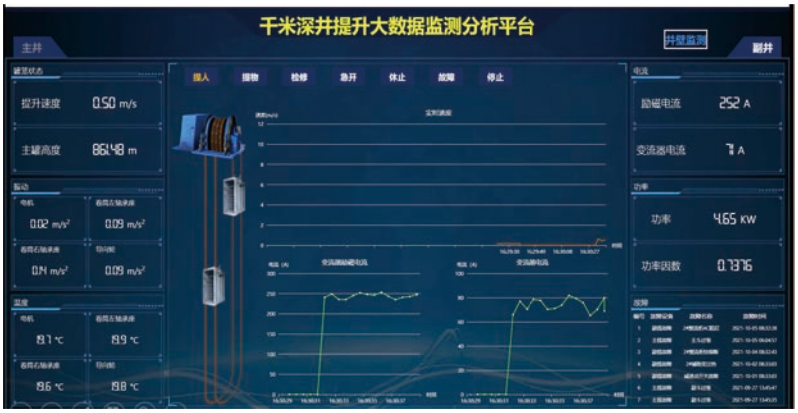


图 2 矿山提升系统监测平台

立德树人，人才培养持续做出贡献

陈朋朋教育引导学生坚定理想信念，厚植爱国主义情怀，加强基础知识学习，指导学生获得江苏省计算机学会优秀硕士论文奖、江苏省优秀本科毕设二等奖、中国大学生计算机设计大赛国赛一等奖及中国物联网学术会议优秀论文奖等荣誉。在此基础上，陈朋朋教授还担任了 IEEE 1851 国际标准工作组副主席、江苏省计算机学会系统结构专委副主任、江苏省计算机学会青工委副主任、江苏省人工智能学会奖励委员会委员及徐州市人工智能学会秘书长等学术职务，为地下物联网等相关领域的人才培养以及学术平台建设持续做出贡献。


获奖证书

中国矿业大学作品《矿小助》在2021年（第14届）中国大学生计算机设计大赛中荣获

一等奖

作　　者：牟金腾 吕迎朝 李家鑫
指导教师：陈朋朋
作品类别：软件应用与开发-移动应用开发（非游戏类）
证书编号：JSJDS20210043202100879200017091006 4

中国大学生计算机设计大赛组织委员会
组织委员会2021年7-8月

图 3 指导本科生获中国大学生计算机设计大赛一等奖

学会动态

首届应用型高校计算机类专业发展论坛暨师资能力提升研讨会成功举办

2023 年 7 月 27—31 日，首届应用型高校计算机类专业发展论坛暨师资能力提升研讨会在云南昆明如期并成功举办，本次会议由江苏省计算机学会应用型高校计算机学科建设专家委员和信息技术新工科产学研联盟计算机工程与应用工作委员会主办，云南经济管理学院和常熟理工学院承办，北京博创智联科技、中软国际教育集团和头歌教学研究中心协办。来自国内 70 余家单位 140 余位代表参加了本次会议，首届应用型高校计算机类专业发展论坛是江苏省计算机学会应用型高校计算机学科建设专家委员会和信息技术新工科产学研联盟计算机工程与应用工作委员会共同打造的品牌活动之一。本届论坛主题为“发展·融合·提升”，旨在全面贯彻落实党的二十大报告关于深入实施科教兴国战略、人才强国战略、创新驱动发展战略，开辟发展新领域新赛道，不断塑造发展新动能新优势；深入推进新时期高校计算机类专业发展建设，提升应用型高校师资能力，为国内应用型高校教师提供学习交流机会。本届论坛流程紧凑而流畅，各环节的安排得到了与会代表的一致好评。本届论坛的召开，也为国内应用型高校教师提供了一次很好的学习交流机会。

训练需求井喷 “算力之渴”何解

近日，《算力基础设施高质量发展行动计划》印发，算力整体布局持续优化，全国上下已形成积极推动算力产业快速健康发展的局面。工信部数据显示，截至 2023 年 6 月底，全国在用数据中心机架总规模超过 760 万标准机架，算力总规模达到 197 百亿亿次 / 秒，算力总规模近 5 年年均增速近 30%，存力总规模超过 1080EB。

与基础设施建设相同步，算力融合应用加速涌现。根据中国信息通信研究院（以下简称中国信通院）的测算，2022 年我国算力核心产业规模达到 1.8 万亿元。算力每投入 1 元，将带动 3 至 4 元的 GDP 经济增长。

但与此同时，大模型产业井喷式发展也带来算力紧缺、能耗激增等问题。目前国内已有 100 多个大模型公开发布，这加剧了智能算力紧张的局面。面对需求的暴增，“算力之渴”如何解？

“绿化”算力全产业链

“我们正在推动液冷、间接蒸发自然冷却等节能技术的应用，并在部分算力中心开展试点。同时，我们正推动数据中心企业启动氢燃料电池等绿电在数据中心的试点，梳理 10 个‘小散老旧’数据中心（约 4000 个标准机架），预计将于年底前全部完成数据中心改造任务。”近日，在工业和信息化部新闻宣传中心（人民邮电报社）、中国邮电新闻工作者协会等单位联合组织的“算力中国行”大型调研采访活动中，上海市通信管理局信息通信发展处副处长魏征对记者表示。

满足算力需求，除了增加算力供给规模以外，数据中心的“降耗增效”也正在紧锣密鼓地进行。目前，从政府到企业，各个主体都在积极探索优化数据中心制冷系统，提高算力应用环节效率。

近日，蚂蚁集团与中国信通院发布《面向算力应用环节的计算绿色化白皮书》（以下简称白皮书），白皮书显示，截至 2023 年 6 月，我国累计建成 196 家国家绿色数据中心，行业内先进绿色中心电能利用效率降至 1.08 左右，达到世界领先水平。但伴随产业发展，PUE 指标（即评价数据中心能源效率的指标）的优化空间不断减少，局限性逐渐显现。

蚂蚁集团首席技术官、平台技术事业群总裁何征宇坦言，算力由数据中心的服务器提供，但实际上大量的电能都被用于维持服务器的正常运转，而并没有供给计算服务本身。根据统计数据，目前的数据中心可能只有低于 20% 的电能用于计算本身。

白皮书也提到，在推动算力绿色低碳发展的过程中，行业的关注点仍主要停留在可再生能源生产和绿色数据中心建设上。为了实现更大程度的总体节能减排效果，需要重视“端到端的绿色计算”。

端到端的绿色计算，即从电力生产、算力生产（包括智算中心建设商、硬件厂商、云厂商），到算力应用全产业环节的绿色计算。据何征宇介绍，在电力生产环节，主要通过优化用能结构，实现数据中心清洁能源和新型储能的合理利用，围绕源网荷储一体化的微电 w 网并网模式，探索数据中心与能源融合发展的绿色新场景；在算力生产环节，通过应用高能效绿色硬件技术与产品，以及从选址、设计、施工、运营等环节入手提高数据中心全生命周期绿色化水平，推进数据中心节能减排改造与绿色低碳化重构，并采用高效 IT 系统、制冷系统、供配电系统以及辅助系统，协调促进算力绿色生产与绿色传输；在业务应用环节，要注重提高软件平台对计算资源的利用率，提升应用与算法效率，将固有计算资源通过动态、弹性的方式进行调配，减少算力应用侧能源消耗，实现算力在服务环节的绿色低碳。

白皮书强调，算力应用环节的绿色计算，是智能算效提升空间最大、收效最快，也是尚未被足够重视的部分。

提升算力调度水平

当前，算力网络快速融合，多样的算力需求提升了对异构算力网络的需要，而将不同类型的算力资源高效精准地调度到相应需求的资源节点中，就需要进行算力调度。

让算力像水、电一样流动，供有需要的企业按需取用，是算力调度的理想状况。近两年，国内涌现出一大批算力调度平台，这些平台可通过整合不同来源、类型和架构的算力资源，满足丰富的业务应用场景需求。今年 6 月，我国首个实现多元异构算力调度的全国性平台“全国一体化算力算网调度平台”发布。

中国工程院院士高文表示，我国算力网络面临的两大技术挑战，其中之一就是算力调度挑战。目前云算力、智能算力、超算算力、混合算力的服务商入口、系统、计费标准等均不相同，这为算力调度带来了一定挑战。

中企通信数据科学及创新总监詹东东也表示：“尤其对于云计算和边缘计算协同的算力架构而言，最重要的是对算力的调度能力。对算力的需求很多时候会分布在不同的区域，如何协同好这些算力，是很多客户正在思考的问题。”

国家（上海）新型互联网交换中心（SHIXP）是算力调度、交易的重要试点。记者了解到，SHIXP 主要负责本地区的算力网络和算力中心的算力调度，并于今年 7 月底正式上线了全国首个算力交易平台。目前，SHIXP 已建成算网节点 14 个，吸纳入驻企业 40 家，接入国内主流运营商 8 家、总带宽 1.82T，累计等级算力资源超过 156 千万亿次 / 秒。

“我建议，全国要建一个大的算力调度标准。”魏征表示，统一标准的建立，意味着所有的云服务商要对外提供统一、可度量的算力服务，既保证现有运营商资源的直接输出，也保证他们在未来统一市场的可持续发展。

推进国产化软硬件适配

随着人工智能技术的大爆发，特别是大模型时代的到来，通用大模型正快速向行业垂直应用领域推进。通过模型调优与快速迭代，垂直大模型正在释放前所未有的产业数智新动能；同时，智算资源紧缺、国内产品相对不足的痛点也更加突出。除了硬件制造能力，在底层技术开发、软件生态适配和场景落地实践等方面，中国算力产业还需长足进步。

上海市通信管理局二级巡视员葛伦卿表示，目前我国在算力供给方面，还面临着国产适配性较差等问题。目前很多国产芯片厂商都在做自研或兼容开源生态软硬件，企业间的低效竞争等问题愈发突出。针对这些问题，联盟、协会等行业组织要凝聚共识，帮助算力供给方打破技术和生态壁垒，加强国产化软硬设备的研发与应用。

工业和信息化部相关负责人表示，围绕算力发展需要，应增强自主创新能力，推进计算架构、计算方式和算法创新，加强 CPU、GPU 和服务器等重点产品研发，加速新技术、新产品落地应用；同时，围绕算力相关软硬件生态体系建设，将加强硬件、基础软件、应用软件等的适配协同，提升产业基础高级化水平。

相关新闻

国产量子算力进入航天领域

科技日报讯 （记者吴长锋）记者 10 月 16 日从安徽省量子计算工程研究中心了解到，日前，由中国航天空气动力技术研究院（以下简称航天气动院）与本源量子计算科技（合肥）股份有限公司（以下简称本源量子）合作共建的量子计算流体力学联合工作室在中国航天空气动力技术研究院正式揭牌，这标志着国产量子算力开始进入中国航天领域。

在航空器、火箭等设备的设计和制造过程中存在许多关键问题，它们都需要通过流体动力学，综合运用数值模拟、实验验证和理论分析等方法来解决。传统主流方法在研究复杂流动场景时，基于现有计算能力下无法兼顾数值模拟的质量和效率，而量子计算作为一种新型计算范式有望解决这一难题。

航天气动院副院长艾邦成表示，量子计算机不仅能够加速流体动力学模拟，在气动优化设计领域同样具有优越性。“例如在翼型的升阻比优化设计中，涉及到大量参数和约束条件下的最优化问题，采用经典计算机求解非常困难。量子计算机就像是‘加速器’，采用合适的量子优化技术有望在较短的时间找到问题的最优解。”艾邦成说。

“将量子计算应用于流体动力学之前，还需要克服许多技术和算法上的挑战。”本源量子总经理张辉向记者介绍，“未来，我们将培养并储备相关人才，强化合作与交流，在气动仿真、飞行器气动设计与优化等方向探索具体的量子计算应用场景，推动量子计算流体力学在我国航空航天、智能制造等行业的应用。”

学会动态

科普教育基地建设研讨会在线上线上举办

7 月 17 日江苏省计算机学会计算机科普教育基地建设研讨会在线上成功举办，此次会议由宿迁学院信息工程学院承办。江苏省科协科普部二级调研员范正银、江苏省计算机学会秘书长金莹、江苏省计算机学会副秘书长游辉敏、宿迁学院信息工程学院院长张岩、宿迁学院产业技术研究院院长王志超、宿迁学院信息工程学院副院长陈林等出席研讨会，相关师生参加了此次会议。与会专家认为，宿迁学院信工院的科普教育内容资源丰富、科普活动形式多样、科普成果突出，对中小学“双减”工作发挥了积极作用。今后，要进一步增强科普工作的计划性，加强科普能力建设，积极主动与地方相关部门开展合作协作，突出科普自研成果展示，精心塑造科普教育品牌，不断提升科普教育影响力和贡献度。

人工智能会产生意识吗

在科幻电影《机械姬》中，人工智能（AI）机器人艾娃被赋予了人类的外貌和智能，她甚至展现出与人类相似的情感和欲望，试图与程序员迦勒建立情感联系，并追求自由。但是，她的“内心”实际上是由电路和算法构成的。该电影可谓讨论 AI 自我意识的一部经典之作，引发了人们对于 AI 自我意识、人性和伦理问题的深入思考。

近日，美国 AI 研究公司 OpenAI 联合创始人兼首席科学家伊尔亚·苏茨克维在接受美国《麻省理工科技评论》杂志专访时，讲述了他对未来 AI 的希望和担忧。苏茨克维认为，如果我们“眯着眼睛看世界”，那么 ChatGPT 可能是有意识的。

这听起来很“疯狂”，但在近一两年，ChatGPT 和其他大语言模型（LLM）震惊了全世界。这些系统可生成似乎能表现出思想、理解力甚至创造力的文本。一时间，“ChatGPT 可能已经有了意识”的话题也冲上了网络热搜，再次引发了网友们对于 AI 是否具有意识等问题的探讨。

苏茨克维言论存在炒作嫌疑

去年 11 月，OpenAI 发布了 ChatGPT 聊天机器人。几周内，它便风靡全球，甚至引发了一场新的全球 AI 竞赛。上线仅仅两个月，ChatGPT 的活跃用户就突破 1 亿，许多人都被这个新“玩意儿”迷住了。

今年 9 月 25 日，OpenAI 宣布将赋予 ChatGPT 利用语音、音频与用户对话的能力。有网友晒出了与其交互的视频，视频中 AI 语音交互中的语气、停顿、重音都十分接近真人。有人质疑，它停顿的那一秒真的是在思考吗？

事实上，10 月初，苏茨克维就在其社交账号上声称，“大型神经网络具有轻微的意识”。

对此，未来主义网站刊文写道：“这是一个不寻常的观点。AI 研究人员广泛接受的观点是，该技术在过去十年中取得了长足进步，但仍远远低于人类智能，更不用说有意识地体验世界了。”

Mind Matters 播客网站文章表示，听到这些 AI 领军人物称赞他们创造的东西既是世界的救世主，又是潜在的死神，真是“令人费解”。然而，一些更中立、客观的观点可能会缓和这种炒作。

计算机工程师罗伯特·J·马克斯就是这些理智的发声者之一，他重申，AI 永远不会实现像人类一样的智能或意识。AI 的运行基于算法，它们的功能与人类的情感和思维有着质的不同。

没有理由认为当前 AI 有意识

不过，也许这样的辩论陷入了语义混乱之中。

意识到底是什么？它能被量化吗？这本身就是令哲学家和科学家困惑已久的问题。

美国趣味科学网刊文称，哲学家将意识描述为具有独特的自我意识以及对周围发生的事情的认识。神经科学家通过分析人的大脑中整合和解释感官数据的活动，对如何量化意识提出了自己的观点。然而，将这些规则应用于 AI 是很棘手的。

粤港澳大湾区数字经济研究院（IDEA）讲席科学家张家兴在接受科技日报记者采访时表示，对于 AI 领域的从业人员来说，对“什么叫 AI 拥有自我意识”这件事情并没有一个很清楚的定义，自我意识很难像人脸识别、语音识别这些能力一样可以被定量衡量。

《自然》网站 10 月 30 日发表书评文章称，美国纽约大学神经科学家约瑟夫·勒杜在《存在的四个领域》中提出，地球上的生命有四种基本类型：生物生命、神经生物生命、认知生命和意识生命。勒杜坚持认为，意识只能存在于生物体内。即使模仿具备意识的生物的机制（无论这些机制从微观到宏观层面是什么），所产生的系统也不会有意识。

澳大利亚《对话》杂志今年也表示，“目前的这些 AI 系统真的能进行思考和理解吗？”不是一个可单纯通过技术进步来回答的问题。

新发表在预印本网站 arXiv 上的一篇论文称，今年 3 月发布的语言模型 GPT-4 未能通过图灵测试。

大脑如何产生意识依旧未解

据 10 月 27 日发表在《神经科学》杂志上的论文，爱沙尼亚研究委员会 3 名科学家认为，尽管 AI 系统具有复杂的反应，但这些系统不具备人类所拥有的具体经验或神经机制。因此，将 AI 的能力等同于真正的意识可能过于简单化了。

研究指出，语言模型缺乏我们通过感官与周围世界进行接触所获得的具体的、嵌入性的

信息。同时，当今 AI 算法的架构缺少与哺乳动物意识相关的丘脑皮质系统的关键特征。此外，导致有生命、有意识的有机体出现的进化和发展轨迹，是今天的 AI 系统的发展轨迹所不可比拟的。

生命有机体的存在依赖于多层次的细胞活动，能动性和意识正是产生于复杂的有机体活动。因此，尽管人们想当然地认为 ChatGPT 和类似的系统可能是有意识的，但这严重低估了我们大脑中产生意识的神经机制的复杂性。

25 年前，神经科学家克里斯托夫·科赫与哲学家大卫·查默斯打了个赌。科赫说，大脑神经元产生意识的机制将于 2023 年被发现。查默斯表示不可能。今年 6 月 23 日在纽约召开的意识科学研究协会年会宣布了这一赌局的结果：查默斯获胜。也就是说，大脑神经元产生意识的机制并未在今年被发现。

发表在《神经科学》杂志上的论文也认为，生物神经元是真实的物理实体，可以生长和改变形状，而大型语言模型中的“神经元”只是无意义的代码片段。我们距离理解意识还有很长的路要走，因此，距离研发有意识的机器还有更长的路要走。

《自然》网站文章称：AI 系统会有意识吗？我们真的想要制造一台具有意识或能动性的机器吗？这些问题的答案都是不确定的。我们还不了解世界上哪些生物是有意识的，也没有建立起将这种可能性纳入考虑的伦理框架。在该文作者看来，成为一个去思考意识来源的生物，而不是一个渴望创造人工意识的生物，也许是更明智的选择。

相关链接

最新研究表明

GPT-4 未通过图灵测试

科技日报讯 （记者张佳欣）图灵测试是一种检验机器是否具有人类智能的方法。ChatGPT 在 AI 领域成为“新星”的过程中，有一个问题一直存在：它是否通过了图灵测试。

美国加州大学圣迭戈分校的研究人员卡梅隆·琼斯和本杰明·卑尔根借鉴了艾伦·图灵的研究成果，设计了一个程序，以确定机器是否能够达到人类智力和对话能力的临界点。倘若达到临界点，那么它就足以让人误以为它是人类。结果显示，GPT-4 未通过测试。相关研究报告《GPT-4 能通过图灵测试吗？》于 10 月 31 日发表在预印本网站 arXiv 上。

ChatGPT 给人的印象大多是聪明的、快捷的。在与人交谈时，它的回答很人性化，甚至可以表现得幽默风趣，能模仿青少年的措辞，并通过了美国法学院的考试。但有时，人们会发现它提供的信息完全是错误的，是胡编乱造的。

卡梅隆·琼斯和本杰明·卑尔根召集了 650 名参与者，参与者会和人或 GPT 模型进行简短对话，并被要求确定他们在与谁交谈。结果发现，GPT-4 模型在 41% 的情况下骗过了参与者，而 GPT-3.5 模型成功骗过参与者的几率仅为 5% 至 14%。有趣的是，人类仅在 63% 的试验中成功地让参与者相信他们不是机器。

研究人员总结道：“我们没有发现 GPT-4 通过图灵测试的证据。”

基于 AD-XDR 的先进制造业工控安全防护平台

基本情况：

项目名称：基于 AD-XDR 的先进制造业工控安全防护平台

完 成 人：傅涛、王海洋、郭金辉、程旺宗、陈茂洲

完成单位：博智安全科技股份有限公司

项目简介

工业互联网作为先进制造业的重要基石，支撑先进制造业的设备智能化、生产过程数字化、供应链协同化。安全体系渗透于工业互联网产业链各层，是产业链重要的支撑保障，本项目面向先进制造业场景，重点研究工业网络架构、工业资产、工业协议、工业威胁等工业先进制造业中特征，并研制基于 AD-XDR 的先进制造业工控安全防护平台，解决先进制造业中面临的工业网络安全问题，保证先进制造业生产网络安全稳定运行。

主要创新点如下：

（1）基于大规模分布式蜜网的工业威胁数据捕获与分析技术

采用设备控制面和业务面分离以及虚实结合方式，在构建大规模分布式蜜网环境吸引攻击方进行探测与攻击基础上，捕获海量原始攻击数据，对其进行分析抽取威胁线索。从网络层面及系统层面捕获威胁数据，监控和记录攻击方在工业网络蜜罐中进行的所有攻击行为，为追溯与分析安全威胁提供基础数据支持。

（2）基于端口与行为特征的工业协议识别模型构建与深度解析技术

融合传统基于端口的静态特征与协议动态行为特征，构建工业协议识别模型，为每一类型协议建立一个协议静态特征集合和唯一的行为特征集合，采用控制流图模型来描述协议行为特征规则集，突破工控协议识别准确性不高难题。在协议识别基础上对协议进行多模式匹配算法深度解析，定义多种协议的预处理过滤规则，采用多模式匹配算法，在规则字符串匹配时，同时比对多个解析规则字段，以此来提高解析效率。

（3）基于 AI 辅助的威胁研判与联动模型定制技术

针对不同工控场景以及威胁情报内容，通过人工智能快速进行威胁研判与联动，采用基

于密度选择性聚类集成的数据标记技术实现数据更为高效的分析划分，利用已有的工控设备数据集辅以人工注入的威胁情报信息，采用频繁子图挖掘算法来进行频繁子图挖掘，从中获取全面的系统潜在威胁，从解决安全事件的角度提出应对安全事件的联动策略，形成安全设备联动系统模型。

（4）基于高逼真的大规模靶场场景快速复现和重构测试验证技术

根据 AD-XDR 平台建设要求，利用网络场景构建子系统接收来自虚拟网络配置管理插件和目标网络控制构件插件下发的网络拓扑配置伪指令，提供网络节点资源调度、网络连接资源调度、镜像资源调度、协议栈仿真资源调度、虚实互联资源调度和实体网络资源调度的能力，实现虚实结合网络场景的自动化部署。

基于 AD-XDR 的先进制造业工控安全防护平台的实施旨在解决制造业网络稳定性、可靠性和保密性问题，同时也可广泛应用于其他关键基础设施领域，以确保网络的安全运营，标志着国内工业网络检测、防护和响应一体化智慧运营的创新实践。目前项目已在智能制造等多个行业领域近百家客户应用推广，受到了客户的广泛好评。

主要科技创新

制造业是兴国之基、强国之本，随着云计算、大数据、5G、人工智能与物联网等新兴技术蓬勃发展，以及两化融合进程的不断推进，我国目前处于传统的封闭制造向智能制造转型的关键节点。智能制造技术有效提高了工业控制系统的网络化、系统化和自动化程度，但融入互联网后工业控制系统所面临的网络安全威胁也日益增加。

面对当前复杂的国际网络安全形势和针对先进制造业集群的高级持续性攻击组织（APT），针对先进制造业中特有的工业网络架构、工业资产、工业协议和工业威胁等问题进行研究。基于 AD-XDR 的先进制造业工控安全防护平台面向关键基础设施行业中先进制造业工控网络安全需求，创新性设计工业网络一体化安全运营方案。平台集成工控主机卫士、工控安全审计系统、工控防火墙、工控安全隔离网闸、工业网络蜜罐系统等安全设备，整合形成先进制造企业终端侧、网络流量侧、网络边界侧安全数据资源池，为企业高级威胁检测及安全事件快速响应提供基础安全数据。搭建 AD-XDR 统一威胁运营平台，将上述安全设备产生的日志数据、流量数据进行数据归一化处理、数据关联分析，通过工业蜜网生成的威胁情报进行赋能，有效联动各类安全设备，实现精准的威胁识别，降低告警误报率，提升安全事件处理效率，化“被动防御”为“主动运营”，实现先进制造企业安全可视化、自动化、智能化。同时将平台在不同类型制造业仿真环境中进行测试验证，保证平台在推广应用中的有效性和可靠性。

基于 AD-XDR 的先进制造业工控安全防护平台前端由工控安全审计系统、工业网络蜜罐系统等设备构成，为高级威胁检测提供最重要的安全数据。AD-XDR 统一威胁运营平台基于

安全大数据分析技术，采用威胁建模，报警收敛、事件聚合等核心技术，通过威胁情报的赋能，有效联动工控防火墙、工控安全隔离网闸、工控主机卫士等安全防护类设备，遏制报警风暴、还原攻击链路、自动处置高级威胁。AD-XDR 统一威胁运营平台不仅承担为其他安全设备提供策略下发和信息上传的安全通道，同时进行综合分析和响应决策。基于 AD-XDR 的先进制造业工控安全防护平台整体技术架构如下：

图 1 技术构架

科技创新 1、基于大规模分布式蜜网的工业威胁数据捕获与分析技术

针对工业企业面临大量未知网络威胁及 APT 攻击，难以捕获的现状，本项目创新性提出了一种基于大规模分布式蜜网的工业网络威胁情报的捕获技术，该技术采用设备控制面和业务面分离的方式，将工控设备的逻辑控制、程序存储、业务处理和 I/O 模块分离，利用虚实结合方式构建分布式蜜网场景。在构建大规模分布式蜜网环境吸引攻击方进行探测与攻击基础上，捕获海量原始攻击数据，对其进行分析抽取威胁线索。本方案从网络层面及系统层面捕获威胁数据（如网络连接记录、原始数据包、系统行为数据、恶意代码样本等），监控和记录攻击方在工业网络蜜罐中进行的所有攻击行为，为追溯与分析安全威胁提供基础数据支持。

为有效捕获恶意程序样本，首先在服务器区部署分布式蜜罐系统，在掩护真实服务器的同时设置了一个引诱黑客的陷阱，实现通信流量采集。可根据捕获到的恶意程序样本，分析相关特征并提交到统一杀毒软件系统，进行全网病毒标记查杀。同时根据日志情况通过分析定位，溯源传播病毒的主机进行隔离恢复运行。基于分布式蜜罐技术的网络攻击防御监测系统需具备蜜罐服务区、日志模块和重定向器，重定向器中安装好入侵检测系统，把蜜罐虚拟机和服务器隔离，服务器放置在单独的 DMZ 区。

如图 2 分布式蜜罐部署所示。

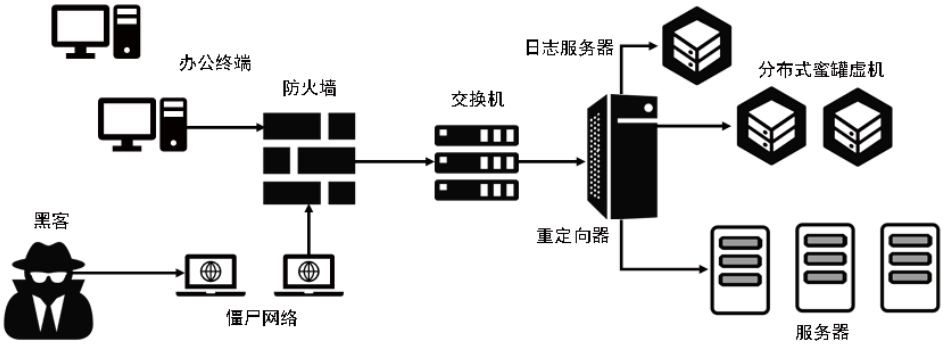


图 2 分布式蜜罐部署图

本方法可实现在进行流量转发的同时，监测和鉴定正常访问行为和恶意攻击程序，对该访问请求进行日志记录，根据日志记录可以快速定位可疑攻击源。通在重定向器中安装好入侵检测系统，主要基于网络流量的异常检测和基于特征匹配的检测方法，如果检测为正常访问行为，则不用重定向，由真实服务器响应请求，如果鉴定为恶意攻击行为，重定向器则激活重定向机制，将访问行为重定向至蜜罐虚拟机中。在蜜罐系统中安装还原软件增加系统还原保护机制，合理设置系统防火墙，只允许入站动作，禁止出站动作，防止蜜罐虚拟机被病毒感染后向外部散播。

针对典型生产控制网络使用设备类型多、协议多变等问题，通过构建丰富的工业协议指纹库，实现对典型工控生产网络设备特征流量的匹配。通过对工控系统网络中的协议进行分析，本技术建立的特征指纹库涵盖了协议类型、流量大小、设备协议配置和协议数据内容 4 个方面，保证了流量特征的全面性。协议类型特征能判断存在入侵行为的通信方式；流量大小特征直观反映出入侵行为的存在；设备协议配置特征还原出入侵行为的入侵信息；协议数据内容特征则便于技术人员隔离入侵行为。并且，该指纹库设备包括工业控制设备、系统主机及服务器、应用软件、智能设备、网络设备等。根据分层分析法的原理，分层建立工控系统网络特征指纹库，指纹库依据分层思想依次分别为 O={ 特征指纹库 }、I={ 流量大小，协议类型，设备协议 配置，协议数据内容 }。流量特征指纹库构建结构如图 3 所示。基于协议指纹库，对蜜罐捕获的流量数据进行指纹特征匹配实现对所有通信行为的指令级解析并做审计，实现功能码控制、寄存器控制、连接状态控制等的合规检测。

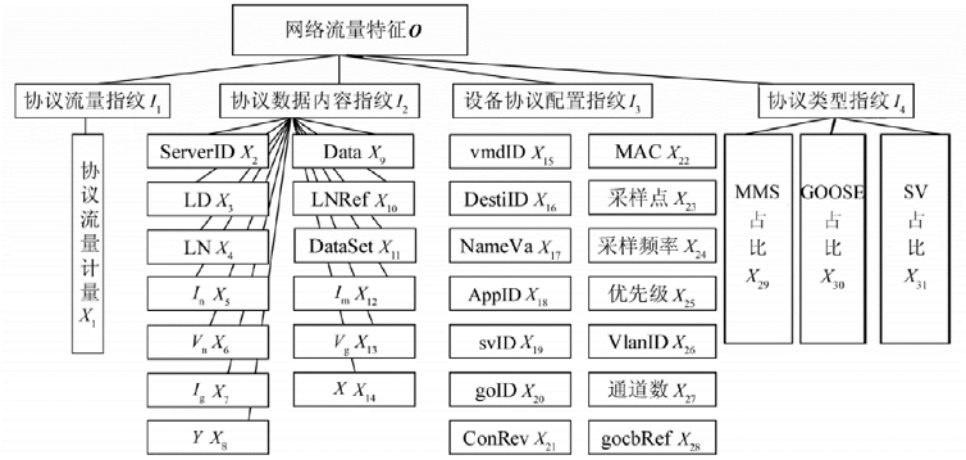


图 3 流量特征指纹库构建结构

同时采用静态检测与动态检测相结合的方式对蜜罐捕获的数据进行分析，静态分析针动恶意样本本身的特征进行分析，但静态分析手段，可以通过修改固定特征或者加壳的手段进行免杀处理。动态行为分析则是对静态分析的一个强有力的补充，依赖样本在运行过程中触发的行为进行监控，达到全面的动态分析以及完整记录，供事后取证分析。基于动态行为分析的恶意事件检测技术，通过构建样本数据运行的虚拟化环境，采用内存动态跟踪分析、沙箱虚拟执行技术对样本进行深入分析自动执行样本数据，并跟踪在此过程中出现的各种系统调用和操作指令，以此判断待测样本数据是否包含真正攻击行为，降低误报率。

科技创新 2、基于端口与行为特征的专用协议识别模型构建与深度解析技术

协议在设计初期按照通信规范一般会定义一个默认的通信端口，大部分基于 TCP/IP 的网络通信协议均可按照端口映射的方式进行识别，在协议特征建立阶段对于能够准确提取静态匹配模式的协议，直接提取协议样本当中的命令或静态标识字段建立协议静态匹配特征集（每一类协议一般有多个静态特征）协议识别阶段静态特征匹配引擎使用。对于无法准确提取静态匹配特征的协议或软件，建立协议运行期间行为状态模型作为协议行为特征并建立协议行为特征序列规则库，供协议识别阶段静态特征匹配引擎和协议行为特征序列匹配引擎使用。在协议具体运行期间，每一个具体的协议都有唯一的行为特征，包括运行步骤、数据包传输的逻辑关系、具有独特性的命令对等等，以此建立特定的行为特征序列作为行为特征序列模型。

在协议识别阶段将报文应用层数据作为文本输入，首先将所有静态匹配特征作为模式集合，采用多模式配算法找到报文所属的可能协议集合；对于可能的协议集合，如果结果不唯一，采用协议行为特征匹配，依靠建立的行为特征序列对于协议运行的数据包序列进行匹配。

因为协议行为特征的独立性可唯一确定所使用的协议及其它相关信息 如版本号。在进行协议行为特征规则提取的过程当中，对于采集的大量协议样本进行数据挖掘，利用关联规则和自学习方法逐步提取并修正行为特征序列。出于效率的考虑，不同的协议运行过程产生的协议行为特征序列的大小不同，可以根据具体的准确性需要制定行为特征序列的长度，必要时可针对特定协议的不同行为实现多行为特征序列匹配。

为每一类型协议建立一个协议静态特征集合和唯一的行为特征集合。其中的静态特征一般是一个有限长度的连续字节串 而行为特征一般是一个有限长度的特征字串序列（行为特征也可能不唯一）。在协议运行过程中，首先使用静态特征匹配规则判断报文所属协议类型，限定可能使用的协议和软件的集合。在此基础上对于可能的协议类型进行行为特征检测就可以唯一的确定其所属协议类型。为某一类型协议建立的协议行为特征规则集为一个规则集合，采用控制流图模型来描述协议行为特征规则集。技术实现途径，

图 4 所示：

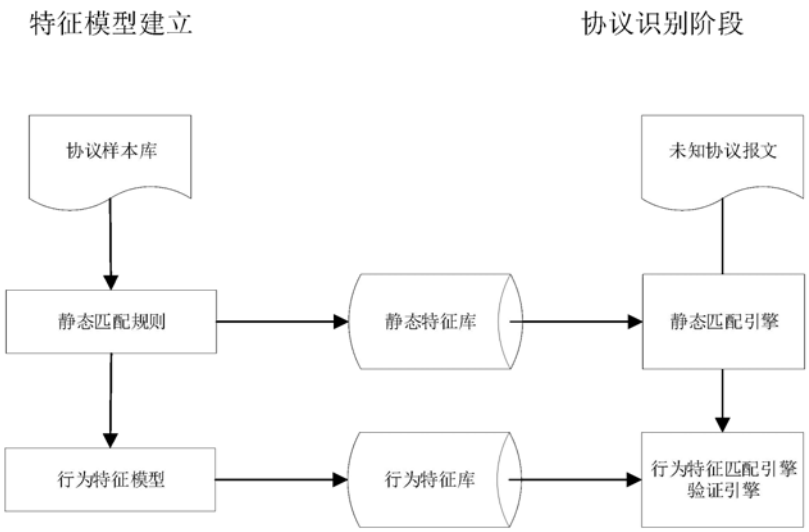


图 4 基于行为特征的协议识别模型构建

在协议识别基础上对协议进行深度解析，协议规则库是进行深度解析的基础，在数据接收缓冲区收到采集的原始数据包后，首先进行的是包完整性检查、包过滤等程序。完整的过滤规则属性包括物理地址、通信地址、端口、协议，这是在基于深度解析的基础上进行的，需要在预处理阶段进行过滤需要只针对数据包长度、数据包功能码进行校验，因此系统需定义多种协议的预处理过滤规则，如西门子 PLC 首次请求链接时的握手协议是基于 TCP/IP 的，若短时间内存在多次连接主机请求则抛出异常。即可定义为，此类过滤规则存储于协议规则库中，提供多种协议的关键字段识别与数据包完整性检查功能。

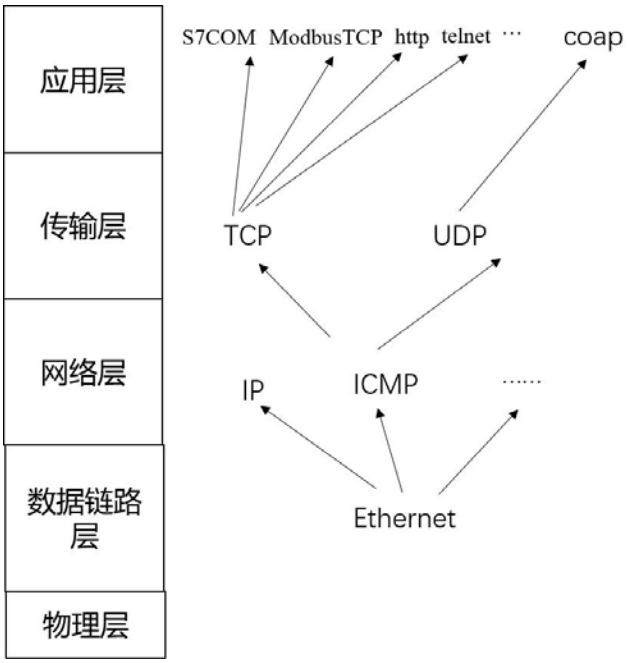


图 5 协议树层级构建

科技创新 3、基于 AI 辅助的威胁研判与联动模型定制技术

基于威胁情报的研判取证及安全设备协同联动是安全设备协同联动响应的核心部分，如图 6 所示。为了进一步提高协同联动响应速度，针对不同的先进制造工控场景，本课题提出基于 AI 辅助的威胁研判与联动模型定制技术。技术基本思想是，针对不同工控场景以及威胁情报内容，通过人工智能快速进行威胁研判与联动。

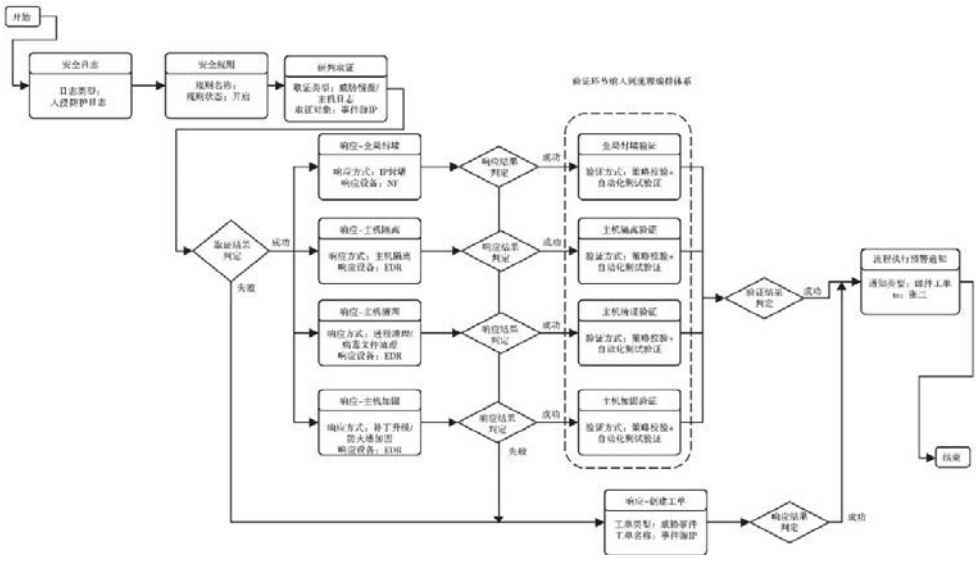


图 6 安全设备联动响应流程

（1）基于密度选择性聚类集成的数据标记技术

采用人工智能方法对工控场景快速进行威胁研判与联动之前，需要对工控设备数据进行分析实现数据的标记与划分，然而这些数据是无标签数据，无法进行分类。针对工控设备运行过程产生的数据结构不规律、数据密度分布不均匀以及存在噪声点的问题。本项目设计了基于密度选择性聚类集成的数据标记技术，以实现数据更为高效的分析划分。该技术方案的算法流程如图 7 所示。

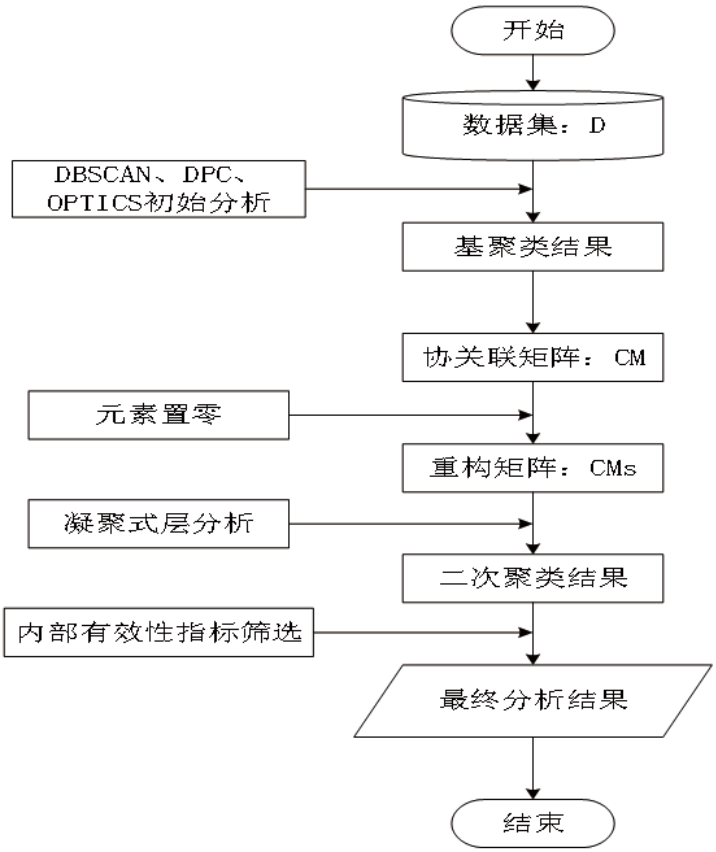


图 7 密度选择性聚类集成算法流程

该算法实现思路如下：使用 DBSCAN、DPC 以及 OPTICS 三种基于密度的聚类分析算法对数据集进行初始分析，它们可以处理不同的数据结构密度，由此获得具有差异性的、效果较好的基聚类划分结果。然后根据生成的基聚类结果创建一个协关联矩阵（CM），将矩阵中小于不同阈值的元素置为零，去除部分相似度较小的值，即消除噪声对数据划分效果的负面影响，从而可以得到多个重构矩阵。对每个重构矩阵进行凝聚式层次聚类分析，可以得到相应的二次基聚类结果。最终利用内部有效性指标 IVI 对得到的多个结果进行筛选，通过选择集群内部较为紧密而集群之间相关性较低的结果，即内部有效性指标值最小的二次基聚类结果，选择最优的分析结果作为最终的聚类集成结果。

（2）基于图挖掘和决策树的威胁定位技术

本项目提出一种基于改进的基于图挖掘和决策树的威胁定位技术，利用已有的工控设备数据集等，辅以人工注入的威胁情报信息，形成待测试的程序，运用开源的 AspectJ 获取执行轨迹，生成 Call 图，采用频繁子图挖掘算法来进行频繁子图挖掘，获取特征信息，使用 J48 决策树算法找到怀疑度高的威胁信息，最后进行威胁定位结果评估。

（3）安全设备联动系统模型

网络安全设备联动系统旨在通过将不同的安全设备产生的安全日志进行规范和统一处理，并从大量日志中提取出高风险的威胁事件，然后运用关联规则算法对威胁事件进行数据挖掘，从中获取全面的系统潜在威胁，提供准确的安全风险分析报告和风险控制错书，从解决安全事件的角度提出应对安全事件的联动策略，为管理员发出相应的风险告警信息，由管理员确认下一步的联动处理事务，有效处理内部违规操作和外部威胁等一系列网络安全问题，从而减少恶意攻击带来的损失。本项目设计的联动系统总体框架如图 8 所示。

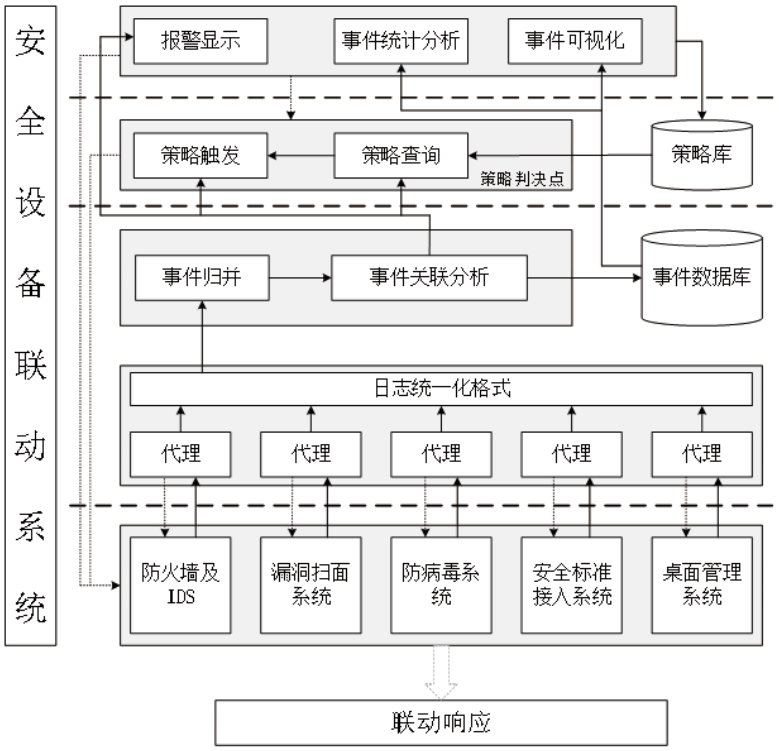


图 8 安全设备联动系统模型框架

该联动系统分为以下三个层次，每层的功能如下：

1）设备管理层：负责对各类安全设备进行实时监控。主要有两个功能：一是通过代理，对联动系统中各类安全设备进行实时监控，采集各个安全设备生成的安全日志，然后对其进行格式的规范化处理，并将处理的结果传送给事件管理层；二是根据决策层管理员针对安全事件制定的安全策略，对各个设备接口进行相应的配置。

2）事件管理层：主要负责对上传的安全事件进行处理，包括冗余归并、事件关联分析等。事件关联分析利用事件关联规则算法对上传的事件进行数据挖掘，找出其中潜在的危险事件，并将其上报给决策层，以便管理员对此采取相应的防范措施。同时，将处理的结果储存到事件数据库中，供管理员对事件进行查询、统计分析等。

3）决策层：对于从事件管理层上传的安全事件，根据策略库，自动地采取相应的联动策略，并将联动指令下发给设备管理层。当策略库中没有相应安全事件的联动策略时，由管理员根据自己的专业知识和业务经验，手动的配置相应的策略，并将策略保存到策略库中，以便后续使用，并以此来不断地丰富策略库。

科技创新 4、基于高逼真的大规模靶场场景快速复现和重构测试验证技术

（1）基于 MDSAL 模式的自动化部署

基于 MDSAL 模式的自动化部署技术在北向配置接口对网络场景编排采用统一的描述模型，而在南向云平台接口采用多种异构云平台自身的描述模型，通过适配器机制完成南北向之间的模型转换，实现与多个异构云平台之间的对接，同时，在与云平台的交互过程中，完成计算、存储、网络等资源调度指令向云平台的下发，由云平台根据指令完成资源的自动化加载和场景的自动化生成，并采用分布式设计框架，将计算集群切分为多个集群或多个区域，实现多集群（多区域）的并行化节点启动加载方案，保障 10 万规模节点的快速生成、启动。基于 MDSAL 模式的自动化部署技术实现方案，如图 9 所示。

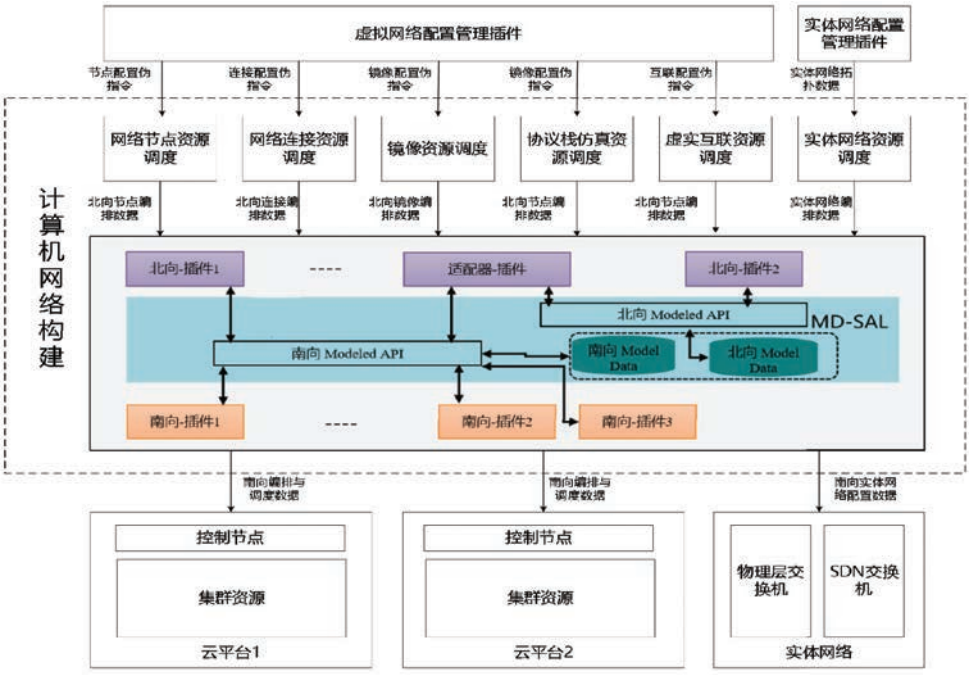


图 9 基于模型驱动的方案架构

该技术实现方案的处理思路如下：根据该项目的建设要求，网络场景构建子系统接收来自虚拟网络配置管理插件和目标网络控制构件插件下发的网络拓扑配置伪指令，提供网络节点资源调度、网络连接资源调度、镜像资源调度、协议栈仿真资源调度、虚实互联资源调度和实体网络资源调度的能力，即按照伪指令中 XD 配置信息的要求，基于网络拓扑配置数据建立抽象模型，并在北向用 YAML 语言对该模型进行计算机语言描述，实现对该拓扑的编排，并根据该编排数据需要发往的南向目标云平台或目标实体网络，经过 MDSAL 的适配转换后，形成面向目标云平台或目标实体网络的模型编排数据，并发往目标对象，基于目标云平台的虚拟化管理能力和实体网络设备的软件定义能力，实现虚实结合网络场景的自动化部署。

（2）基于多级 SDN 的虚实网络构建

基于多级 SDN 的虚实网络构建技术将网络场景的构建分成两个级别，即顶层 SDN 和执行层 SDN。

顶层 SDN（Super-SDN）根据目标网络控制构建插件下发的目标网络拓扑数据，将整个目标网络拓扑数据划分为云内网络拓扑、云间网络拓和实体网络拓扑三个部分的数据，并将每个部分的网络拓扑数据转换为对应的网络编排指令，并通过 API 接口将编排指令下发至执行层 SDN。

执行层 SDN 包括云内 SDN（Inner-SDN）、云间 SDN（Inter-SDN）和实体 SDN（Comb-SDN）三个部分，分别接受来自顶层 SDN 的网络编排指令，并将网络编排指令转换为网络配置数据，下发至虚拟网络节点、SD-WAN 网关、SDN 交换机和物理层交换机等网络设备，由网络设备根据配置数据生成所需的网络拓扑场景。如图 10 所示：

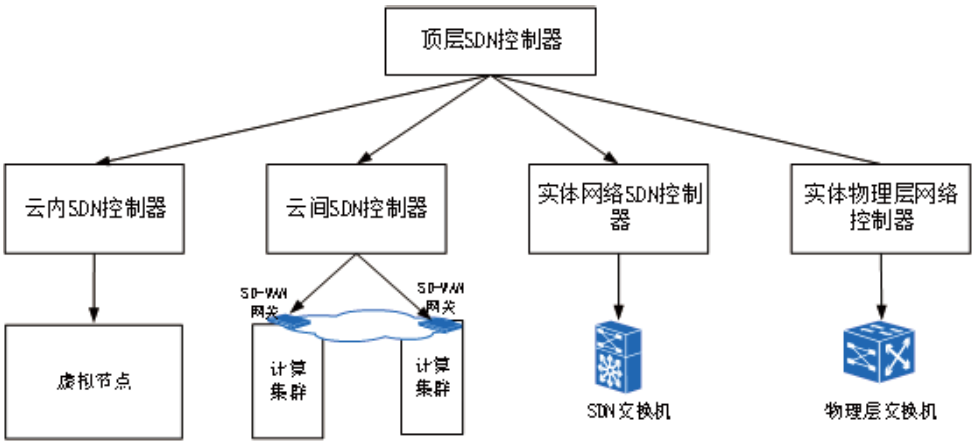


图 10 多级 SDN 整体结构

（3）基于分布式存储的镜像快速分发

利用分布式存储技术将大容量的镜像文件切割为一个一个的数据块结构，封装为对象的形式并分布式散列在多个存储资源池中，同时，分布式存储技术采用计算与存储相分离的方式，通过构建独立的存储网络，将虚拟节点启动所需的卷设备通过存储网络映射至的不同块

存储空间，这样，通过镜像文件块空间和卷设备块空间在存储网络中的分布式交换，达到快速的镜像文件分发效果。

在本次项目设计中，应用分布式存储技术可实现集群中计算平面与存储平面的分离，计算平面由计算节点构成，存储平面由存储节点构成，因此，在计算节点中构建的虚拟节点，其对应存储空间的卷设备实际被映射在存储节点中，计算节点需通过存储网络实现对存储资源的访问和使用。

另一方面，在此次项目设计中，镜像文件以文件的形式存放在分布式存储节点中，由于采用了分布式存储方式，镜像文件被分解为多个数据块的形式被散列在多个存储节点中，这样，镜像文件和虚拟节点的卷设备实际存储的物理位置都在存储节点中。

由于在分布式存储设计中，存储对象被散列在不同的存储节点中，这对于镜像文件的快速分发提供两个方面的优势：

1）由于镜像文件散列在多个存储节点中，因此，多个并发的镜像下载任务可同时在多个存储节点中开展，一方面，每个存储对象在分布式存储中都有相应的备份，可通过同时读取主数据和备份数据来提高镜像文件数据的下载速度，另一方面，镜像文件散列在存储节点中的不同部分，可通过存储网络向不同的计算节点并行下载，克服传统存储方案中只能通过单节点下载的缺点，从而提高镜像文件下载的速度。

2）由于镜像文件和虚拟节点映射的卷设备都在存储池中，可通过存储网络或存储节点中的硬盘传输，可同时将镜像文件的多个数据块复制到卷设备所对应的多个数据块中，由于二者都在分布式存储节点中，通过存储节点之间直接的数据对拷实现镜像文件数据的快速传输，将远高于传统存储方案中单节点的网络传输速度。

结合本次项目需求，镜像文件分布式存储方案的技术架构，如图 11 所示。

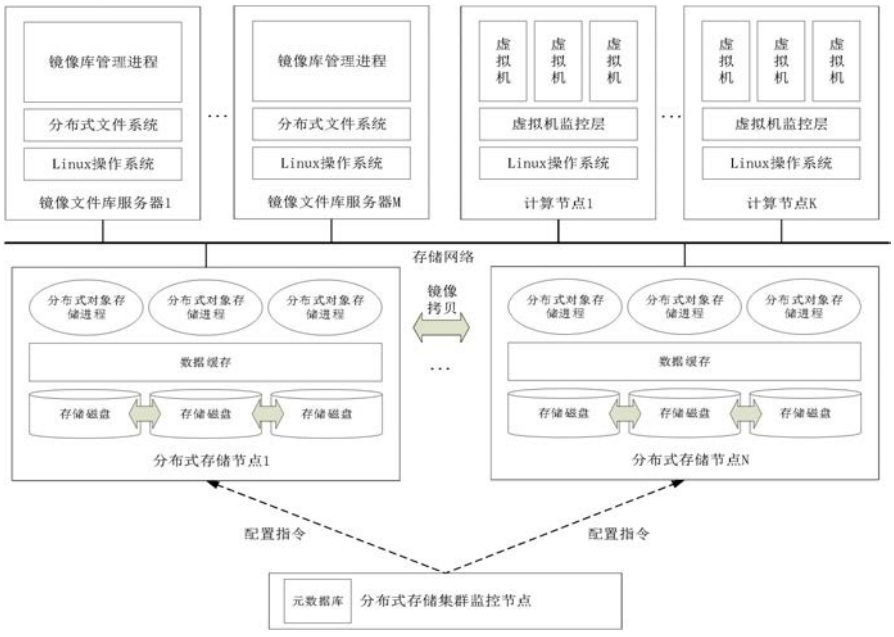


图 11 分布式存储方案架构

博智安全科技股份有限公司
地址：南京市雨花台区软件大道 168 号润和国际软件园 C 幢 3-5 层
电话：400-100-0298
电子信箱：services@elextec.com
网址：www.elextec.com

学会动态

“走进南京大学”科技创新与体验研学活动成功举行

8 月 9 日“走进南京大学”科技创新与体验研学活动第二期，随着科技的高速发展和广泛应用，科技主题研学活动成为培养学生科技兴趣和创新能力的重要途径。参观科技企业或科技园区是一种直观了解科技发展的方式，学生可以亲眼目睹科技产品的研发过程，了解现代科技的应用场景，从而激发优秀中小学生的科研创新潜质，提升其动手实践能力。

继 2023 年 8 月 5 日成功地举办了首期研学活动后，2023 年 8 月 9 日，江苏省计算机科普教育基地再次推荐了来自全省各地的 29 名中小学生参加了我学会联合南京大学计算机科学与技术系面向优秀中小学生开展的“走进南京大学”科技创新与体验研学活动第二期。通过参与本次活动，学生们不仅丰富了暑期生活，更深刻认识到南京大学作为一所优秀学府的历史积淀，以及计算机科学技术在日常生活中的重要地位。此次科技实践体验活动也锻炼了学生们的科学思维和研究能力及培养了创新意识和科技实践能力，同时也明确了学习的真正意义与目标。

简　介

江苏省计算机学会常务理事单位

盐城师范学院信息工程学院（软件学院）

盐城师范学院信息工程学院（软件学院）现有计算机科学与技术、软件工程、数字媒体技术、物联网工程、数据科学与大数据技术等 5 个本科专业，其中软件工程专业是国家首批一流本科专业建设点，江苏省首批产教融合品牌专业建设点，通过教育部组织的工程教育认证。学院在校全日制本科生 1420 人。

学院是教育部数据中国“百校工程”产教融合基地、江苏省国际服务外包人才培训基地、江苏省卓越工程师教育培养计划试点单位。学院实验中心是江苏省计算机实验教学示范中心和江苏省 IT 服务外包工程实践教育中心。学院与盐城盐南高新区合作创建信创软件产业基地，获江苏省重点产教融合基地建设点。

学院有一支富有朝气、勇于进取的师资队伍。教职工总数 73 人，其中高级职称 39 人，博士 21 人，外籍教师 3 人。省“六大人才高峰”高层次人才项目资助 2 人；省“333 高层次人才培养工程”培养对象 2 人；江苏省优秀教育工作者 1 人，江苏省高校“青蓝工程”中青年学术带头人和优秀青年骨干教师 5 人，江苏省本科高校青年教师教学竞赛一等奖获得者 1 人。

近年来，学院共承担国家自然基金 6 项、国家社会科学基金 1 项，江苏省自然基金、科技支撑计划等省部级科研项目 18 项，发明专利 20 余项，发表科研论文 200 余篇，其中 SCI、EI 收录 60 余篇；拥有江苏省沿海滩涂生态环境智慧监测工程研究中心省级工程研究中心 1 个，市级工程研究中心 5 个。1 门课程被认定为国家级一流本科课程；获国家教学成果二等奖 1 项，江苏省高等教育教学成果奖一等奖 1 项、二等奖 2 项；出版各类著作与教材 20 余部，省重点教材 3 部；学生在省级及以上竞赛中获奖 195 项。毕业生考研率逐年提升，就业率连续多年保持在 98% 以上，并在各行各业深受用人单位好评。