ISSN 2096-742X
CN 10-1649/TP

文献DOI：
10.11871/jfdc.issn.
2096-742X.2020.
01.003

文献PID：
21.86101.2/jfdc.
2096-742X.2020.
01.003

页码：27-37

开放科学标识码
（OSID）

面向人工智能和大数据的高效能计算

李肯立[1,2*]，阳王东[1*]，陈岑[1,2]，陈建国[1]，丁岩[1]

1. 湖南大学信息科学与工程学院，湖南 长沙 410008
2. 国家超级计算长沙中心，湖南 长沙 410008

摘　要：【目的】本文主要分析人工智能和大数据应用随着迅速增大的数据规模，给计算机系统带来的主要挑战，并针对计算机系统的发展趋势给出了一些面向人工智能和大数据亟待解决的高效能计算的若干研究方向。【文献范围】本文广泛查阅国内外在超级计算和高性能计算平台进行大数据和人工智能计算的最新研究成果及解决的挑战性问题。【方法】大数据既为人工智能提供了日益丰富的训练数据集合，但也给计算机系统的算力提出了更高的要求。近年来我国超级计算机处于世界的前列，为大数据和人工智能的大规模应用提供了强有力的计算平台支撑。【结果】而目前以超级计算机为代表的高性能计算平台大多采用CPU+加速器构成的异构并行计算系统，其数量众多的计算核心能够为人工智能和大数据应用提供强大的计算能力。【局限性】由于体系结构复杂，在充分发挥计算能力和提高计算效率方面存在较大挑战。尤其针对有别于科学计算的人工智能和大数据领域，其并行计算效率的提升更为困难。【结论】因此需要从底层的资源管理、任务调度、以及基础算法设计、通信优化，到上层的模型并行化和并行编程等方面展开高效能计算的研究，全面提升人工智能和大数据应用在高性能计算平台上的计算能效。

关键词：超级计算；大数据；高效能计算；人工智能；异构系统

Efficient Computing for Artificial Intelligence and Big Data

Li Kenli[1,2*], Yang Wangdong[1*], Chen Cen[1,2], Chen Jianguo[1], Ding Yan[1]

1. College of Information Science and Engineering, Hunan University, Changsha, Hunan 410008, China
2. National Super-computer Center in Changsha, Changsha ,Hunan 410008, China

Abstract: **[Objective]** This paper mainly analyses the main challenges brought to computer system by the rapid increase of data scale of AI and big data application. In view of the development trend of computer system, some research directions of high-efficiency computing towards AI and big data are given. **[Coverage]** In this paper, the latest research results and challenges of big data and artificial intelligence computing on supercomputing and high performance computing platforms at home and abroad are extensively surveyed. **[Methods]** Big data not only provides an increasingly rich training data set for artificial intelligence, but also puts forward higher requirements for the computing power of computer systems. In recent years, China's supercomputer techniques are

基金项目：国家重点研发计划（2018YFB1003401）；国家杰出青年基金项目（61625202）；国家自然科学基金项目（61872127, 61572175, 61751204, 61472124）；国际交流合作项目（61860206011）

*通讯作者: 李肯立(E-mail: lkl@hnu.edu.cn); 阳王东(E-mail: yangwangdong@163.com)

at the forefront of the world, which provides a powerful computing platform for large-scale applications of big data and artificial intelligence. **[Results]** At present, high-performance computing platforms represented by supercomputers mostly use heterogeneous parallel computing systems composed of CPUs and accelerators, where a large number of computing cores can provide powerful computing power for AI and big data applications. **[Limitations]** However, due to the complex architecture, there are major challenges in making full use of computing power and improving computing efficiency. The parallel computing efficiency is more difficult to improve, especially in the artificial intelligence and big data domains which are different from scientific computing. **[Conclusions]** Therefore, it is required to conduct research on high-performance computing from underlying resource management, task scheduling, basic algorithm design, and communication optimization to the upper level of model parallelization, so that the computational efficiency of artificial intelligence and big data applications on high-performance computing platforms can be improved.

Keywords: artificial intelligence; big data; heterogeneous systems; high efficiency computing; supercomputing

引言

随着互联网、物联网技术的发展，数据量以前所未有的速度爆炸式增长[1]，预计到2020年将达到35 ZB。大数据(Big Data)概念受到越来越多的关注。学术界和产业界关于大数据的认识也在逐步清晰化并形成共识，大数据是现有产业升级与新产业诞生的重要的推动力量[2]。大数据可以由它的四个特征定义，即数据规模庞大（Volume）、数据更新频繁（Velocity）、数据类型多样（Variety）和数据价值（Value），通常称为4V的模型。（1）数据规模庞大（Volume）。大数据最大的特征是数据规模庞大[3]。例如，Flicker公司每天会产生大约3.6 TB的数据。Google公司每天需要处理大约20 000 TB的数据。美国国家安全局报告称，每天会在互联网上聚集大约1.8 PB的数据。（2）数据更新频繁（Velocity）。大数据的数据更新频繁，产生速度快。这对数据的实时处理，响应时间提出了更高的要求。大数据的实时分析对于电子商务等提供在线服务的行业至关重要。（3）数据类型多样（Variety）。大数据另外一个特点是数据种类繁多，数据的存在形式丰富，包括文字，图像，视频，图形等。大多数传统数据采用结构化格式，并且很容易存储在二维表中。但是，超过75%的大数据都是非结构化的。（4）大数据蕴含着巨大的价值，但是价值密度低[4]。

对于大多数应用程序，如工业和医疗，关键是要从大数据中提取有价值的知识。虽然大数据为包括电子商务、工业控制和智能医疗在内的广泛领域提供了很好的机会，但光有数据是不够的，大数据处理技术需要快速、有效的处理数据来满足应用的要求[5]。

人工智能（Artificial Intelligence，AI）技术迅猛发展，它是研究、开发用于模拟、延伸和扩展人的智能的一门新的技术科学。当前以深度学习为代表的机器学习技术在人工智能领域[6]，如图像识别、语音处理和自然语言处理等方面取得较大的突破。与传统的浅层机器学习相比，如支持向量机和朴素贝叶斯，深度学习模型更复杂，可以利用大量的数据样本自动提取高级特征并通过更有效地组合低级输入来学习分层表示。以图像识别应用来说，著名的ImageNet数据集[7]包含了1 500万张高分辨率图片，例如Hinton 2012年用来训练ImageNet的深度神经网络，已包含6 000万参数和65万个神经元。另外，现实中很多新兴人工智能应用需要多模态的跨域协同。如何利用多模态数据进行跨域协同的多模态机器学习是当前人工智能研究的发展趋势与亟待解决的问题之一[8]。大数据的发展从某种程度上说推动了以深度学习为代表的人工智能技术的发展。但是大数据、人工智能处理技术的发展也给计算与存储平台、计

算能力等提出了很高的要求。

另一方面，高性能技术、异构并行处理技术发展迅猛。目前，GPU 的发展成为了通用并行计算设备，它具有数据存储和交换的内存带宽高、基于多核多线程的程序模型技术和超强的计算能力等特点。更为重要的是，近年来，我国采用异构并行体系结构的超级计算机的硬件研制能力已跃居世界前列。“十一五”期间，在国家 863 计划“高效能计算机及网格服务环境”重大项目的支持下，我国先后研制成功若干台百万亿次和千万亿次高性能计算机系统。2008 年，曙光公司研制成功“曙光 5000”百万亿次超级计算机位列全球 TOP500 第十，亚洲第一 [9]；2009 年，国防科技大学研制成功“天河一号”千万亿次计算机，使我国成为继美国之后世界上第二个研制成功千万亿次计算机的国家；2010 年，曙光公司研制成功“星云”千万亿次计算机，性能列世界 TOP500 第二位；而升级后的“天河 -1A”系统也创造了中国高性能计算机全球排名第一的最好成绩 [10]。2013 年，国防科技大学研制的“天河二号”超级计算机再次登顶，计算性能达到每秒 33.86 千万亿次。预计到 2020 年，峰值运算能力还将提升 10 倍以上，达到百万万亿次的规模 [10-11]。与此同时，近年来我国还逐步建设成立了包括广州、天津、长沙、济南、深圳在内的一系列的国家超级计算中心。

在近年来我国高性能计算机连续多年夺得世界第一、我国已成为事实上超算大国的客观环境下，如何利用我国的超级计算机、特别是基于国产自主处理器的异构众核超算系统来面对大数据、人工智能等技术的发展，采用高效能计算技术是迫切需要研究和解决的问题。

1　面向大数据和人工智能的高效能计算所面临的挑战

针对不同数据类型与应用的大数据处理系统和人工智能应用是支持大数据科学研究和人工智能研究的基础平台。对于规模巨大、结构复杂、价值稀疏的大数据，其处理亦面临计算复杂度高、任务周期长、实时性要求强等难题 [12-13]。大数据和人工智能应用的这些难点不仅对高效能计算系统的系统架构、计算框架、处理方法提出了新的挑战，更对高效能计算系统的运行效率及单位能耗提出了苛刻要求，要求高效能计算系统必须具有高效能的特点 [14-16]。对于以高效能为目标的高效能计算的系统架构设计、计算框架设计、处理方法设计和测试基准设计研究，其基础是高效能计算系统的效能评价与优化问题研究。这些问题的解决可奠定高效能计算系统设计、实现、测试与优化的基本准则，是构建能效优化的分布式存储和处理的硬件及软件系统架构的重要依据和基础，因此是大数据分析处理和人工智能应用所必须解决的关键问题。与面向传统业务的高效能计算不同，在大数据和人工智能时代，高效能计算由于其处理对象、计算体系结构、并行处理模式、响应时间和能耗等因素，使得其面临许多新的挑战，具体如以下几方面。

1.1　体系结构

面向大数据和人工智能的高效能计算系统的效能评价与优化问题具有极大的研究挑战性，其解决不但要求理清大数据的复杂性、可计算性与系统处理效率、能耗间的关系，还要综合度量系统中如系统吞吐率、并行处理能力、作业计算精度、作业单位能耗等多种效能因素，更涉及实际负载情况及资源分散重复情况的考虑 [17]。因此，为了解决系统复杂性带来的挑战，人们需要结合大数据的价值稀疏性、访问弱局部性、人工智能计算复杂性的特点，针对能效优化的大数据分布存储和处理的系统架构，以大数据感知、存储与计算融合为大数据的计算准则 [18]。在性能评价体系、分布式系统架构、流式数据计算框架、在线数据处理方法等方面展开基础性研究，并对作为重要验证工具的基准测试程序及系

统性能预测方法进行研究。通过设计、实现与验证的迭代完善，最终实现大数据计算系统的数据获取高吞吐、数据存储低能耗和数据计算高效率。

高性能计算模式更加复杂。虽然任务调度技术已经在传统计算系统中取得了良好的效果，但是它们很难高效地应用于超级计算和高性能计算集群环境。当计算资源越来越庞大并且具有复杂结构，实际应用程序和计算任务越来越复杂时，传统任务调度算法的性能、可执行性和可扩展性都明显下降。迫切需要更加有效的并行任务调度算法来处理高性能计算环境下的大规模复杂计算任务，提供高效可行的调度方案[19]。同时，现有高性能计算技术忽略计算系统的异构性。现有算法主要侧重于考虑系统的整体能量消耗与任务调度长度之间的关系，几乎没有充分考虑各个异构计算节点的计算能力、存储能力、能量消耗等细节。此外，现有算法的评估实验主要采用随机生成应用程序的 DAG 任务图模型进行仿真和模拟，较少采用真实应用程序的任务组合。

1.2 并行处理模式

大数据的涌现使人们处理计算问题时获得了前所未有的大规模样本，但同时也不得不面对更加复杂的数据对象，如前所述，其典型的特性是类型和模式多样、关联关系繁杂、质量良莠不齐。大数据内在的复杂性 (包括类型的复杂、结构的复杂和模式的复杂) 使得数据的感知、表达、理解和计算等多个环节面临着巨大的挑战，导致了传统全量数据计算模式下时空维度上计算复杂度的激增，传统的数据分析与挖掘任务，如检索、主题发现、语义和情感分析等变得异常困难[20-21]。然而，人们对大数据复杂性的内在机理及其背后的物理意义缺乏理解，对大数据的分布与协作关联等规律认识不足，对大数据的复杂性和计算复杂性的内在联系缺乏深刻理解，加上缺少面向领域的大数据处理知识，极大地制约了人们对大数据高效计算模型和方法的设计能力。因此，如何形式化或定量化地描述大数据复杂性的本质特征及其外在度量指标，进而研究数据复杂性的内在机理是个根本问题。通过对大数据复杂性规律的研究有助于理解大数据复杂模式的本质特征和生成机理，简化大数据的表征，获取更好的知识抽象，指导大数据计算模型和算法的设计。为此，需要建立多模态关联关系下的数据分布理论和模型，理清数据复杂度和时空计算复杂度之间的内在联系，通过对数据复杂性内在机理的建模和解析，阐明大数据按需约简、降低复杂度的原理与机制，使其成为大数据计算的理论基石。目前大部分深度学习和人工智能算法都是基于串行程序设计的。这些算法的各个步骤之间、每轮迭代计算之间存在繁重的逻辑依赖和数据依赖关系，难以对这些算法提出有效的并行计算模型，迫切需要探索高效可行的人工智能并行处理技术。

1.3 响应时间

随着并行计算技术的快速发展，传统的数据挖掘和机器学习算法在分布式计算集群和高性能计算机上已经能够快速实现并取得极短的响应时间。然而，面向大数据和人工智能等应用具有计算逻辑异常复杂、处理数据量庞大等特点，所需要的响应时间远远大于传统的数据挖掘和机器学习应用。另一方面，由于市场竞争激烈，各个应用程序服务提供商对大数据应用和人工智能应用程序的响应时间有着严格的要求。每个提供商都希望能够应用高效能计算技术来提供实时响应的大数据服务和人工智能服务[22-23]。因此，如何设计适合于大数据和人工智能的体系结构和并行计算模式，以高效处理这些应用任务，缩短其响应时间，是一项影响着高效性计算技术发展和推广的关键问题。

1.4 能耗

随着高效能计算技术的发展，大规模计算集群系统消耗了越来越多的能量，在运营成本、环境和系统

可用性等方面产生了各种问题[24]。据统计，信息通信技术行业在2016年的总能耗约为8 680亿，能源消耗问题将进一步转化为高碳排放问题。由于大量的能量消耗，计算系统的温度将急剧上升，需要使用各种设备冷却服务，从而产生大量的冷却成本。而且，有证据表明，计算系统的温度每升高10℃，预期的系统失效率就会增加一倍，这将极大地影响系统的可靠性和可用性，并最终损害了系统性能[25]。特别是面向大规模数据处理和人工智能等应用，这些应用需要处理庞大数据量和复杂计算量，从而导致计算集群能耗急剧上升。因此，如何有效降低高效能计算系统的能量消耗是一项关键技术问题，需要引起广泛重视。

2 面向大数据和人工智能的高效能计算的研究方向

随着数据量的急剧增加，大数据挖掘和机器学习训练耗时长的缺点制约了大规模人工智能算法的应用。同时，近年来高性能计算或超级计算呈现出加速发展的趋势，大数据和人工智能的并行算法和并行平台迅速成为国际科研和产业界的热点。面向大数据和人工智能的高效能计算平台的整体框架结构如图1所示。

图1 面向大数据和人工智能的高效能计算平台的整体框架结构

Fig. 1 Overall framework of high efficiency computing platform for large data and artificial intelligence

2.1 面向大数据和人工智能的并行处理体系结构

（1）面向大数据和人工智能的异构众核并行处理平台体系结构

目前的异构计算的计算机体系结构成为了高性能计算和超级计算的主流，其所提供的强大算力为大数据处理和人工智能应用提供了良好的支撑[26-27]。但是异构并行体系结构的设计需要考虑到多模态机器学习的海量数据输入、中间数据存储，各种机器学习算法的混合与协同，多模态数据之间的数据相关性分析等特征融合算法，以及某一模态内部的模型参数更新等需要进行的频繁通信和同步等对高效易用的并行处理平台设计的需求[28-30]。

（2）面向大数据和人工智能的异构众核并行处理框架

在整体体系结构下，根据多模态机器学习算法的建模与分析结构，充分考虑系统的通用性、可扩展性以及并行处理效率的提高，拟借鉴Tensorflow、MXNet等机器学习平台使用参数服务器来处理频繁通信的优点，面向大规模多模态机器学习需求，对其进行改进，设计一种关联网络服务器，以解决多模态机器学习并行处理中的算法协同问题，在此基础上设计并实现支持CPU+GPU/MIC异构结构和自主众核异构系统的并行处理框架[31-33]。

（3）高效的资源管理与任务调度策略

对通用的CPU+GPU/MIC和国产自主异构众核系统的计算能力、存储和网络通信能力分别进行建模，设计相应的资源管理与任务映射机制。同时根据不同机器学习算法的特征，研究在计算节点内以及节点间高效任务映射方式，以充分发挥节点间与节点内各计算设备的计算潜能，提高计算效率[34]。同时，拟根据多模态机器学习过程中的形式化建模结果，探讨面向多模态机器学习的有向图DAG以及循环图DFG的调度理论与算法。

2.2 面向大数据和人工智能的可扩展异构并行算法和模型设计与实现

（1）多模态数据处理和人工智能的异构并行算法设计

根据多模态数据形式化建模过程中体现的数据量大、结构复杂、冗余信息多等特点，对预处理、特征提取、特征融合以及决策过程涉及和重新设计的核心算法 (Tucker，CP，随机梯度，卷积计算及卷积神经网络，循环神经网络等) 的可并行性、并行粒度与规模等进行深入分析，以大规模多模态数据作为输入，分别设计基于 CPU+GPU/MIC 等通用异构众核系统和基于国产自主处理器的异构众核系统的并行算法。

（2）模型的裁剪与压缩

庞大的参数规模是以深度学习为代表的智能学习算法的标志特征，需要占用大量内存。为了降低深度神经网络的参数规模，模型剪枝与压缩是常被采用的两种有效方法 [35-36]。

模型剪枝是通过移除小权重参数或神经元等网络结构降低深度神经网络参数规模的方法 [37-41]。该方法具体可分为细粒度剪枝（又称非结构化剪枝）和粗粒度剪枝（又称结构化剪枝）[35]，如图 2 所示。尽可能训练出多的参数被认为是神经网络学习能力的保证 [42]。因此，在移除部分参数的情况下，并不会对模型性能产生较大的损害 [43]。另外，模型参数的权重大小代表了该特征的重要程度 [39]。因此，移除小权重参数并不会损害模型的学习能力 [44]。模型剪枝的方法简单有效，但也存在一定的局限性。首先，Liu 等人 [35] 的工作显示参数规模和参数权重大小并不能直接代表网络的学习能力和参数的重要程度。其次，非结构化剪枝方法得到的网络参数矩阵是稀疏的，如果没有专用硬件软件支持，则不能有效地实现模型的计算、压缩和加速 [45]。其次，结构化的剪枝方法将导致网络模型某些能力的永久丧失，并大幅度降低网络的性能。第三，结构化的剪枝方法也需要特殊硬件软件的支持 [46]。最后，权重、神经元、网络通道、网络层的重要性并不是静态的，因此永久删除网络中的某个参数或结构会影响网络的泛化能力 [35]。因此，根据任务类型或者输入数据的特点动态剪枝网络方法更加符合未来智能环境下的设备异构、需求多样的应用场景。

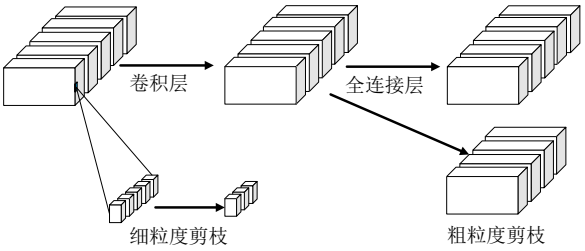


图 2 模型剪枝示例

Fig.2 Model pruning example

模型压缩是通过权重共享或张量分解等技术降低网络的计算开销 [47-49] 和能量损耗的方法 [50-51]。与剪枝方法不同，模型压缩更多关注卷积层的操作，因为卷积计算占据了深度神经网络训练、推断过程中的主要计算开销 [52]。为了满足网络应用实时性的需求，必须加速深度神经网络的训练和推断过程。Denil 等人 [53] 以减少冗余网络为目的，提出了几种 CNN 的压缩方法。Mathieu 等人 [54] 证明傅里叶域卷积运算速度可比普通卷积运算快 2 倍。Denton 等人 [48] 通过奇异值分解（SVD）压缩网络完全连接层的权重矩阵，在没有显着降低模型的预测精度的情况下，提升了网络计算的速度。目前，基于量化 [55]、Hash 技术 [56]、循环投影 [57] 和张量分解 [58] 等方法在网络压缩上均有良好表现。但是，上述工作仅压缩网络中的一层或几层，并不能压缩网络的整体。但对于 ImageNet 这样的复杂任务和网络，压缩整个 CNN 是非常重要的。Zhang 等人 [59] 通过考虑非线性单元加速进而提升卷积效率和网络整体压缩，且该方法可应用于现有的一些神经网络框架中，如 Caffe，Torch 和 Theano。但该方法中涉及的秩选择仍然需要较多的计算开销。由于深度神经网络规模巨大且训练时

间长，压缩网络整体仍然极其具有挑战性。目前的压缩方法大都针对卷积层或全连接层的运算或能耗优化，但在一定程度上影响了网络的性能。因此，未来需要进一步探索能有效降低网络计算和能耗开销，并尽量减少对网络性能影响的模型压缩方法。

2.3 异构众核并行处理算法与平台的稳定性与优化方法

为了提升并行算法和平台的计算性能和稳定性，需要研究以下三个方面：

(1) 大规模多模态机器学习并行处理平台的通信优化方法

大规模多模态机器学习中各模态算法之间的协同以及算法内部的模型更新都需要进行频繁的通信和同步，根据其特点，需研究针对模型压缩的通信优化方法、非规则通信的通信优化方法以及基于张量压缩的通信优化方法等。

(2) 节点内与节点间流水线处理技术

研究并行计算节点内 CPU 与 GPU/MIC 或国产自主超算结点内主处理器与众核协处理器内以及节点间的流水线技术以提升系统的性能。设计节点内与节点间最优的任务重叠方案，尽可能隐藏 CPU 与协处理器间以及节点间数据传输次数和时间。

(3) 运行时容错与稳定性技术

针对大规模异构众核并行系统所具有的规模庞大、软硬件构成复杂、可靠性相对较低等现状，研究面向多模态机器学习的超大规模并行处理系统运行时系统稳定化技术，提出相应的故障预测、恢复和系统重构容错机制和基于检查点的容错机制，以提升系统持续运行能力。研究异构并行系统的容错模型和容错调度算法，通过研究异构计算资源中各主机节点故障和互联网络故障的局部性概率特点，提出以失效率为标准的动态低冗余策略和以主机和网络失效率为核心的容错模型。在此基础上，针对多模态机器学习中具有高可靠性要求的核心算法或模块，且其任务执行时间为已知随机变量的一般任务，设计相应的折衷系统效率和可靠性目标的随机容错调度算法。

3 结束语

利用高性能计算系统，尤其是超级计算系统作为大数据和人工智能的计算平台渐成趋势。各种新型的处理器被应用，不断增加系统的计算能力，从而也促进大数据和人工智能应用向大规模和高深度发展，反过来又对计算系统提出更高要求。单靠提高单处理器的计算性能已经跟不上应用的发展需求，基于多核和从核的大规模异构并行计算机系统成为发展的主流。但是由于异构系统的复杂性，又加上大数据和人工智能应用并行化难度较高，急需提高大规模异构并行计算机系统上执行大数据和人工智能应用的计算效能。计算效能的提升是一个系统工程，需要从底层的资源管理、任务调度、以及基础算法设计、通信优化，到上层的模型并行化和并行编程等方面展开研究，在充分发挥众核处理器的强大并行计算能力的同时，提升计算资源利用率。另外优化算法和模型，能够有效降低能耗开销。

利益冲突声明

所有作者声明不存在利益冲突关系。

参考文献

[1] John Walker S. Big data:A revolution that will transform how we live, work, and think[M]. 2014.

[2] McAfee A, Brynjolfsson E, Davenport T H,et al.. Big data: the management revolution[J]. Harvard business review, 2012, 90(10): 60–68.

[3] Zhang Q, Yang L T, Chen Z, et al.. A survey on deep learning for big data[J]. Information Fusion, 2018, 42: 146-157.

[4] 程学旗,靳小龙,王元卓,郭嘉丰,张铁赢,李国杰. 大数据系统和分析技术综述[J]. 软件学报, 2014, 25(9): 1889-1908.

[5] O'Leary D E. Artificial intelligence and big data[J]. IEEE Intelligent Systems, 2013, 28(2): 96-99.

[6] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. nature, 2015, 521(7553): 436.

[7] Deng J, Dong W, Socher R, et al. .Imagenet: A large-scale hierarchical image database[C]//2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009: 248-255.

[8] Srivastava N, Salakhutdinov R R. Multimodal learning with deep boltzmann machines[C]//Advances in neural information processing systems. 2012: 2222-2230.

[9] top500. https://www.top500.org/list/2008/06/. 2008.

[10] Zhang Y, Sun J, Yuan G, et al.. Perspectives of China's HPC system development: a view from the 2009 China HPC TOP100 list[J]. Frontiers of Computer Science in China, 2010, 4(4): 437–444.

[11] Zhang Y, Sun J, Yuan G, et al. A Brief Introduction to China HPC TOP100: from2002 to 2006[C]. In: Proc of Proceedings of the 2007 Asian technology informationprogram's (ATIP's) 3rd workshop on High performance computing in China: solutionapproaches to impediments for high performance computing. ACM, 2007, 32–36.

[12] J.C. Chaves . Enabling High Productivity Computing through Virtualization[J]. Information Sciences, 2018, 435: 124–149.

[13] 李斌, 周清雷,等. 基于拟态计算的大数据高效能平台设计方法[J]. 计算机应用研究, 2019 (07): 19-25.

[14] 祁琛. 应用于神经网络的高效能计算单元的研究与实现[D]. 南京:东南大学, 2018.

[15] D.H. Jones, A. Powell, C.-S. Bouganis, P.Y.K. Cheung. GPU Versus FPGA for High Productivity Computing[C]. IEEE International Conference on Field Programmable Logic and Applications (FPL). 2010 (06): 112-119.

[16] 张小庆. 高效能云计算虚拟机优化部署策略[J]. 计算机工程与应用, 2016 (04): 28-36.

[17] 王永桂. 流域大尺度水环境模型的高效能集群计算方法研究及其在三峡库区的应用[D]. 武汉: 武汉大学, 2015.

[18] 党林玉. 可重构高效能计算系统中软硬件协同技术研究[D]. 解放军信息工程大学, 2014.

[19] B. Betkaoui, D.B. Thomas, W. Luk. Comparing performance and energy efficiency of FPGAs and GPUs for high productivity computing[C]. IEEE International Conference on Field-Programmable Technology (FPT), 2010 (09): 74-80.

[20] 阮利, 秦广军, 肖利民,等. 基于龙芯多核处理器的云计算节点机[J]. 通信学报, 2013(12): 39-46.

[21] 刘勇鹏. 大规模高效能计算的系统软件关键技术研究[D]. 国防科学技术大学, 2012.

[22] J. Unpingco. User Friendly High Productivity Computational Workflows Using the VISION /HPC Prototype[C]. IEEE International Conference on High-performance Computing, 2018(03): 93-105.

[23] 吴丹. 高效能计算型存储器体系结构关键技术研究与实现[D]. 华中科技大学, 2012.

[24] 李波, 解建仓, 等. 网格环境下的水利高性能计算平行系统及应用[J]. 华中科技大学学报, 2011 (06):73-82.

[25] 王之元. 并行计算可扩展性分析与优化——能耗、可靠性与计算性能[D]. 国防科学技术大学, 2011.

[26] Chu C, Kim S K, Lin Y, et al..Map-reduce for machine learning on multicore[C]//Proceedings of the 20th Annual Conference on Neural Information Processing Systems, Amsterdam, 2007: 281-288.

[27] Gao M, Pu J, Yang X, et al. TETRIS: Scalable and Efficient Neural Network Acceleration with 3D Memory[C] //Proceedings of the Twenty-Second International Conference on Architectural Support for Programming Languages and Operating Systems. ACM,

2017: 751-764.

[28] Jin L, Wang Z, Gu R, et al..Training large scale deep neural networks on the intel xeon phi many- core coprocessor[C] //Proceedings of the International Parallel and Distributed Processing Symposium，Piscataway, 2014: 1622-1630.

[29] Mei K, Dong P, Lei H, et al..A distributed approach for large-scale classifier training and image classification[J]. Neurocomputing, 2014, 144: 304-317.

[30] 杨柳,景丽萍,于剑. 一种异构直推式迁移学习算法[J]. 软件学报,2015,26(11): 2762-2780.

[31] 王岳青,窦勇,吕启,李宝峰,李腾. DLPF:基于异构体系结构的并行深度学习编程框架[J]. 计算机研究与发展,2016,53(06): 1202-1210.

[32] 洪文杰,李肯立,全哲,阳王东,李克勤,郝子宇,谢向辉. 面向神威·太湖之光的PETSc 可扩展异构并行算法及其性能优化[J]. 计算机学报,2017,40(9): 1961-1973.

[33] Chen C, Li K, Ouyang A, et al. .GPU-Accelerated Parallel Hierarchical Extreme Learning Machine on Flink for Big Data[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2017: 1-14.

[34] Xing E P, Ho Q, Dai W, et al. Petuum: A new platform for distributed machine learning on big data[J]. IEEE Transactions on Big Data, 2015, 1(2): 49-67.

[35] Z. Liu, M. Sun, T. Zhou, G. Huang, and T. Darrell. Rethinking the Value of Network Pruning. 2018.

[36] S. Han, H. Mao, and W. J. Dally. A Deep Neural Network Compression Pipeline: Pruning, Quantization, Huffman Encoding. 2015.

[37] Y. L. Cun, J. S. Denker, and S. A. Solla. Optimal brain damage[J]. 1990.

[38] B. Hassibi, and D. G. Stork. Second Order Derivatives for Network Pruning: Optimal Brain Surgeon[J]. Advances in Neural Information Processing Systems, vol. 5, pp. 164-171, 1993.

[39] S. Han, J. Pool, J. Tran, and W. J. Dally. Learning both Weights and Connections for Efficient Neural Networks. 2015.

[40] L. Hao, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf. Pruning Filters for Efficient ConvNets. 2016.

[41] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz. Pruning Convolutional Neural Networks for Resource Efficient Transfer Learning. 2016.

[42] J. H. Luo, J. Wu, and W. Lin. ThiNet: A Filter Level Pruning Method for Deep Neural Network Compression. 2017.

[43] R. Yu, A. Li, C. F. Chen, J. H. Lai, V. I. Morariu, X. Han, M. Gao, C. Y. Lin, and L. S. Davis. NISP: Pruning Networks using Neuron Importance Score Propagation. 2017.

[44] Y. He, X. Zhang, and S. Jian. Channel Pruning for Accelerating Very Deep Neural Networks.2017.

[45] H. Song, X. Liu, H. Mao, P. Jing, A. Pedram, M. A. Horowitz, and W. J. Dally. EIE: Efficient Inference Engine on Compressed Deep Neural Network[J]. Acm Sigarch Computer Architecture News, vol. 44, no. 3, pp. 243-254, 2016.

[46] X. Gao, Y. Zhao, L. Dudziak, R. Mullins, and C. Z. Xu. Dynamic Channel Pruning: Feature Boosting and Suppression. 2018.

[47] E. P. Yong-Deok Kim, Sungjoo Yoo, Taelim Choi, Lu Yang, and Dongjun Shin.COMPRESSION OF DEEP CONVOLUTIONAL NEURAL NETWORKS FOR FAST AND LOW POWER MOBILE APPLICATIONS. in ICLR, 2016.

[48] E. Denton, W. Zaremba, J. Bruna, Y. Lecun, and R. Fergus. Exploiting Linear Structure Within Convolutional Networks for Efficient Evaluation. 2014.

[49] T. Cheng, X. Tong, Z. Yi, X. Wang, and E. Weinan. Convolutional neural networks with low-rank regularization. Computer Science, 2016.

[50] H. Y. ,Y. Z., J. Liu.End-to-End Learning of Energy-

Constrained Deep Neural Networks. in IDLR, 2019.

[51] T. J. Yang, Y. H. Chen, and V. Sze.Designing Energy-Efficient Convolutional Neural Networks Using Energy-Aware Pruning. 2017.

[52] K. Simonyan, and A. Zisserman.Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. Computer Science, 2014.

[53] M. Denil, B. Shakibi, L. Dinh, M. A. Ranzato, and N. D. Freitas. Predicting Parameters in Deep Learning.

[54] M. Mathieu, M. Henaff, and Y. Lecun.Fast Training of Convolutional Networks through FFTs. Eprint Arxiv, 2013.

[55] Y. Gong, L. Liu, Y. Ming, and L. Bourdev.Compressing Deep Convolutional Networks using Vector Quantization[J]. Computer Science, 2014.

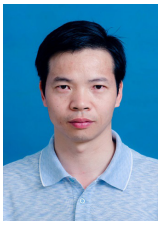
[56] W. Chen, J. T. Wilson, S. Tyree, K. Q. Weinberger, and Y. Chen.Compressing Neural Networks with the Hashing Trick[J].Computer Science, pp. 2285-2294, 2015.

[57] Y. Cheng, F. X. Yu, R. S. Feris, S. Kumar, A. Choudhary, and S. F. Chang.Fast Neural Networks with Circulant Projections. 2015.

[58] A. Novikov, D. Podoprikhin, A. Osokin, and D. Vetrov. Tensorizing Neural Networks. 2015.

[59] X. Zhang, J. Zou, M. Xiang, K. He, and S. Jian.Efficient and Accurate Approximations of Nonlinear Convolutional Networks. 2014.

收稿日期：2019 年 10 月 28 日



李肯立，湖南大学信息科学与工程学院，教授，博士生导师，主要研究方向为高性能计算、并行计算、人工智能。
本文承担工作为：框架的整体结构设计、研究指导。

Li Kenli, Doctor, is a professor of School of Information Science and Engineering, Hunan University. His main research fields are high performance computing, parallel computing and artificial intelligence.
He undertakes the following tasks: the overall structure design and research guidance of the frame.
E-mail: lkl@hnu.edu.cn

阳王东，湖南大学信息科学与工程学院，教授，博士生导师，主要研究方向为高性能计算、并行计算。
本文承担工作为：研究方向的凝练和论文的整合。

Yang Wangdong, Doctor, is a professor of School of Information Science and Engineering, Hunan University. His main research fields are high performance computing, parallel computing.
He undertakes the following tasks: the figure research direction out and the integration of papers.
E-mail: yangwangdong@163.com

陈岑，湖南大学信息科学与工程学院博士后，主要研究方向为大数据处理、并行计算与人工智能。
本文承担工作为：序言撰写和研究问题分析。

Chen Cen, post-doctoral researcher at the School of Information Science and Engineering, Hunan University, focuses on big data processing, parallel computing and artificial intelligence.
He undertakes the following tasks: preface writing and problem analysis.
E-mail: chencen@hnu.edu.cn

陈建国，湖南大学信息科学与工程学院，博士后，主要研究方向为大数据和人工智能。
本文承担工作为：框架的整体结构设计、研究指导。面向大数据和人工智能的高效能计算所面临的挑战分析。

Chen Jianguo, is a post-doctoral researcher at School of Information Science and Engineering, Hunan University. His major research areas include big data and artificial intelligence.

He undertakes the following tasks: being the research director who is responsible for the design of the whole framework and analyzing the challenges of efficient computing for big data and artificial intelligence.

E-mail: jianguochen@hnu.edu.cn

丁岩，湖南大学信息科学与工程学院在读博士生，主要研究方向为边缘计算、数据挖掘。

本文承担工作为：深度神经网络模型剪枝与压缩方法调研。

Ding Yan, a PhD student at College of Information Science and Engineering, Hunan University. His research fields are edge computing and data mining.

He undertakes the following tasks: investigate on pruning and compression methods of deep neural network models.

E-mail: ding@hnu.edu.cn

引文格式：李肯立,阳王东,陈岑,陈建国,丁岩. 面向人工智能和大数据的高效能计算[J].数据与计算发展前沿,2020,2(1):27-37.DOI:10.11871/jfdc.issn.2096-742X.2020.01.003.PID:21.86101.2/jfdc.2096-742X.2020.01.003.

Li kenli, Yang Wangdong, Chen Cen, Chen Jianguo, Ding Yan . Efficient Computing for Artificial Intelligence and Big Data [J].Frontiers of Data & Coputing,2020,2(1):27-37.DOI:10.11871/jfdc.issn.2096-742X.2020.01.003.PID:21.86101.2/jfdc.2096-742X.2020.01.003.