

人工智能基础软硬件架构的关键技术研究^{*}

杨建,傅明驹,徐国强

(江苏自动化研究所 计算机事业部,江苏 连云港 222000)

摘要:针对当前人工智能技术在军用领域的应用问题和难点,研究基于深度学习技术的军用软硬件架构设计方法,并结合业内主流的软硬件基础架构,分析在军事应用中涉及的关键技术和解决思路。通过对深度学习模型压缩转化、数据增强和分布式训练等软硬件架构的关键技术研究,设计构建了一站式人工智能开发平台、嵌入式边缘智能计算平台,对军用人工智能提供端到端的训练、部署和测试支撑。该研究可为未来智能无人作战系统提供软硬协同的智能计算解决方案。

关键词:军用人工智能;模型压缩;数据增强;一站式人工智能开发平台;嵌入式边缘智能计算平台

中图分类号:TP391.7

文献标识码:A

DOI: 10.19358/j.issn.2096-5133.2020.03.009

引用格式:杨建,傅明驹,徐国强. 人工智能基础软硬件架构的关键技术研究[J]. 信息技术与网络安全,2020,39(3): 44-49.

Research on key technologies of basic hardware and software architecture of artificial intelligence

Yang Jian, Fu Mingju, Xu Guoqiang

(BU of Computer, Jiangsu Automation Research Institute, Lianyungang 222000, China)

Abstract: Aiming at the application problems and difficulties of artificial intelligence technology in the military field, the design method of military hardware and software architecture based on deep learning technology is studied. Combined with the mainstream software and hardware infrastructure in the industry, the key technologies and solutions in military applications are analyzed. Through studying the key technologies of software and hardware architecture, including deep learning model compression transformation, data enhancement and distributed training, this paper designs and builds a one-stop artificial intelligence development platform and an embedded edge intelligent computing platform to provide end to end training, department and test support for military AI. It provides a software-hardware collaborative intelligent computing solution for future intelligent unmanned combat system.

Key words: military artificial intelligence; model compression; data enhancement; one-stop artificial intelligence development platform; embedded edge intelligent computing platform

0 引言

未来战争中,战场认知与决策速度将成为战争胜负的决定因素。谁能够更快地处理战场信息(感知)、理解态势(认知)、判断决策(决策)、执行打击(行动)等任务,谁就能压缩OODA任务周期,从而在对战中形成与对手的不对称优势^[1]。而人工智能在军用领域的研究当前还处于初级阶段,属于弱人工智能范畴。其对军用信息系统的智能化分析与辅助能力基本停留在事务处理和业务计算层面,

对所处理的信息(如战场数据、作战分析、态势图生成等)无法实现智能认知以及智能辅助决策^[2],无法满足军用OODA的需求。此外,在实际军事应用中要想取得对抗优势,必须具备强大的计算能力支撑。因此,需要研究人工智能在军用领域中基础软硬件架构所涉及的关键技术,这对解决人工智能在军用领域的落地问题和降低人工智能的应用门槛具有重要的作用。

本文首先描述了基于深度学习技术的军用软硬件架构设计方法,接着介绍了深度学习在当前军用领域面临的问题和挑战。然后借鉴业内主流的

^{*} 基金项目:国防预研课题(31511040102)

软硬件基础架构,分析在军事应用中涉及的关键技术和初步的解决思路。最后提出人工智能开发平台、嵌入式边缘智能计算平台设计的基本架构。本文的主要目的是分析软硬件基础架构设计在军用领域的可行性,提出军用人工智能软硬件架构设计的难点以及解决方法和构想。

1 人工智能软硬件基础架构设计

人工智能是一门交叉性学科,当前主流的研究方向是机器学习和深度学习领域。深度学习又被作为机器学习研究中的一个分支,是由 HINTON G E 等

人于 2006 年提出^[3],其核心实现方式是深度神经网络。深度神经网络的发展需要解决数据、算法和算力问题。首先,高效的深度学习算法主要是依托计算机技术体系架构实现,深度学习算法通过封装成软件框架的方式供开发者使用。因此,软件框架是整个人工智能技术体系的核心,而硬件架构是高效实现人工智能应用的关键。

图 1 所示为基于深度学习技术的人工智能基础软硬件全栈示意图,主要分为以下几部分:

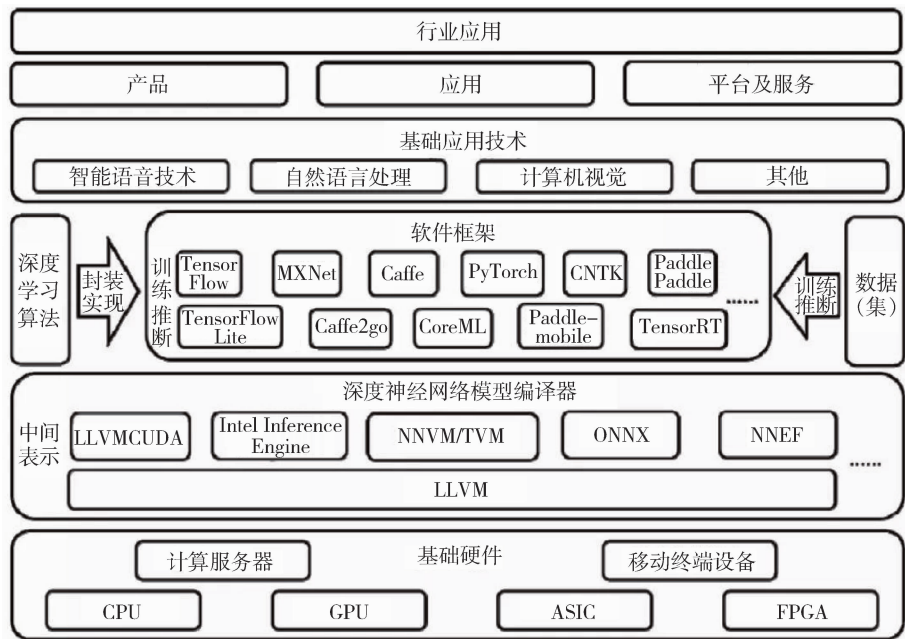


图 1 基于深度学习的人工智能基础软硬件架构图

(1) 基础硬件层

基础硬件层主要是各种架构类型的人工智能芯片。深度学习需要大量重复执行矩阵乘法、激活函数和池化等计算过程。通用 CPU 执行重复计算的性价比比较低,需要采用专用的计算芯片。AI 计算芯片的架构有 GPU、FPGA 或者 ASIC,它们通过把 AI 中常用函数计算硬件化来提升硬件计算速度、降低功耗。其中涉及的关键技术包括指令集、硬件实现方案、异构计算框架和编译器等。考虑到当前军用国产自主安全的需求,可以采用定制化的 ASIC 来实现自主安全的专用计算硬件电路。

(2) 中间表示层

中间表示层主要是深度神经网络模型编译器的实现。中间表示层的核心思想借鉴了 LLVM,它是底层硬件和软件框架,以及不同软件框架之间的

桥梁。此中间表示层是解决模型推理侧运行在不同硬件平台的重要描述方法,主要包括 NNVM/TVM 和 TensorFlow XLA 两大阵营。类似的 ONNX、NNEF,模型交换格式的核心就是对各种中间层表示的定义。由于谷歌的封闭性,可以采用开源的 ONNX 实现解决方案。

(3) 软件框架层

软件框架层是实现深度学习算法封装的关键,为应用开发提供集成软件工具包。一般包括训练和推理框架两部分。决定软件框架易用性的因素是生态和分布式训练。当前生态最好的云端训练框架是 TensorFlow 和 PyTorch 框架,它们同样支持分布式训练。国内百度 PaddlePaddle 以其丰富的生态和齐全的功能,成为国产化深度学习软件的首要选型。

(4) 基础应用技术层

当前人工智能的商业化实现主要是基于计算机视觉、智能语音、自然语言处理等基础应用技术。其中与军用领域最类似的是符合 OODA 环的无人驾驶技术。当前军用领域的应用问题就是要解决智能信息感知、智能信息融合和智能决策等关键技术。

(5) 平台任务层

商业领域的应用任务常见的有人脸识别、人体姿态识别、智能视频分析、智能对话、智能翻译等。军用领域的任务主要是智能辅助决策,包括无人水下扫雷、水下反潜作战、无人战斗机和无人作战艇以及智能作战机器人等^[4]。

2 人工智能软硬件架构的关键技术研究

2.1 深度学习软硬件架构面临的挑战

目前人工智能系统虽然能够完成图像分类、语音识别、自然语言处理和模式识别等任务,但可能会以人预测不到的方式失效。失效的原因是当前的人工智能系统不具备可解释性。这在军用领域,尤其是一些关键任务执行过程是无法容忍的。所以军用领域需要发展稳健和可靠的人工智能系统。当前解决该方法主要分为三类^[5]:在建模之前的可解释性方法;建立本身具备可解释性的模型;在建模之后使用可解释性方法对模型作出解释^[6]。其中,第二种是开发可解释的神经网络模型算法。其他两者是分析数据和特征图,都需要依靠易用的深度学习可视化工具,所以在人工智能软件架构中需要重点构建可视化工具组件来分析军用模型的可靠性和可解释性。该可视化工具的重点应是对数据、特征图和每个神经元权重的可视化和追踪,能够支持对深度神经网络模型进行基于实例的探索和基于子集的探索。此外,它还可以抽象模型结构并显示,使用者可以在此基础上深入分析,在每个模型层或节点这一层面,对神经元激活执行局部检查。为了提升该工具的交互能力,还需要开发基于 Web 的工具链。

目前的人工智能系统需要手工标注非常大的数据集,但当前军用领域却面临样本数量稀少、保密性强、搜集和标记困难等问题^[7]。当前解决小样本问题的方法主要有两种:一是产生更多的训练样本,发展基于样本的数据增强技术;二是发展基于小样本的深度学习算法技术,如迁移学习、few-shot

学习^[8]等。对于数据增强可以开发完整的数据预处理模块,能够支持各种数据清理和数据增强的策略等。而学习方式方面则可以依赖深度学习训练框架发展支持专用的工具包,集成各种算法模块。对于繁琐的样本标记,本文建议发展半自动化或自动化的协同标记流程。

另外,当前军用嵌入式硬件平台异构多样,如何将训练好的模型部署在不同推理框架和硬件之上,并且能保证在异构硬件上达到高效率运行,这是军用人工智能工程落地所必须解决的两个方面,即可移植性和适应性。当前的主流解决方法是发展深度神经网络模型编译器及其上下游的工具集,以解决深度学习模型部署到多种设备时可能存在的适应性和可移植性问题。

最后,在军用硬件平台方面还要解决异构集群的分布式训练问题,这是实现高性能训练的关键。所以要构建异构的计算集群和分布式计算调度系统。这也是当前计算架构最难解决的难题。异构的计算集群可以兼顾主流的 GPU、FPGA、SOC 和 ASIC 等硬件,分布式计算调度系统能支持细粒度的拓扑感知任务部署,能够运行在分布式存储平台之上。基于国产自主安全的需求,本文建议基于新一代人工智能开源开放平台(OpenI 启智)打造异构设备管理框架和资源管理系统。

2.2 人工智能软硬件架构的关键技术

2.2.1 深度学习模型压缩技术

在计算和存储资源受限的嵌入式和移动终端设备上对算法进行高性能的实现,最有效的方式就是进行模型压缩和加速。模型压缩^[9]是指利用数据集对已训练的模型进行精简,进而得到轻量且准确率相当的网络模型,压缩后的模型具有更小的结构和更少的参数,可以显著降低计算和存储开销。

深度学习模型压缩方法主要分为更精细化模型设计、模型裁剪、核的稀疏化、量化、低秩分解、知识蒸馏(迁移学习)等方法^[10],而这些方法又可分为前端压缩和后端压缩;前端压缩是不破坏网络结构的压缩技术,在原有模型上减少 filter 的数量或者网络的深度等;后端压缩的目的是尽可能减少模型大小,会对原始网络结构进行极大的改造。

面向训练模型压缩与优化而开展的自动流程化工具设计,主要包括模型压缩/加速算法组件以及超参数优化组件两部分。以腾讯开源的 Pocket-

Flow 为例,其主要提供通道剪枝组件、权重稀疏化组件、权重量化组件、网络蒸馏组件、多 GPU 训练组件以及超参数优化组件,通过不同算法组件的有效结合,能够实现精度损失更小、自动化程度更高的深度学习模型的压缩与加速。

2.2.2 深度学习编译器构建技术

传统编译器缺少对深度学习算法基础算子的优化,且对人工智能多种异构的计算芯片适配缺失。目前主要是在传统编译器架构的基础上进行扩展,来解决人工智能底层计算芯片与上层软件框架的适配优化问题。例如英伟达的 CUDA 编译器,英特尔的 nGraph 以及华盛顿大学的 NNVM/TVM 编译器都是按照这个思想,参考 LLVM 体系架构设计的。

以 NNVM/TVM 编译器为例,其能够直接从多个深度学习前端框架将工作负载编译成为优化的机器代码,也能在高层图中间表示 (IR) 中表示和优化普通的深度学习工作负载,还能在不同的硬件后端转换计算图、最小化内存占用、优化数据分布、融合计算模式。NNVM 编译器是基于 TVM 堆栈中的两个组件:NNVM 用于生成计算图,TVM 用于映射张量运算。

如图 2 所示,从 TVM 的概览可以看出,其具体工作流程如下:

- (1) TensorFlow、PyTorch 等前端学习框架的模型先转化计算图 IR,主要由 NNVM 组件完成;
- (2) 对原计算图 IR 进行计算图优化(包括操作融合、内存分配优化),得到优化后的计算图;
- (3) 对计算图中的每个 op,获取用张量表达式描述的 Tensor 计算表达式,并针对所需的硬件平台,选择最小计算原语(primitives)生成具体的 schedule(定义怎样调度底层硬件),该步骤由 TVM 组件完成;
- (4) 使用某种基于机器学习的 Automated Optimizer 生成经过优化的 Low Level Loop Program;
- (5) 最后即可生成特定于硬件设备的二进制程序,进一步生成可以部署的 module。

当前主流方案都一致采用了 Graph IR + Tensor IR 两层优化结构。这是解决软件框架移植和硬件优化实现的核心。此外,当前工程应用的训练模型都是按照 IR 层的规定进行表达和存储,IR 是编译器用来表示源码的数据结构或代码。扩充性能的

中间表示层是打通深度学习计算中多种不同前端训练软件框架和多种不同后端表达的桥梁,它可以解决框架间的可移植性问题。当前开源的 ONNX 具有良好的生态,可以作为人工智能中间件的首选。

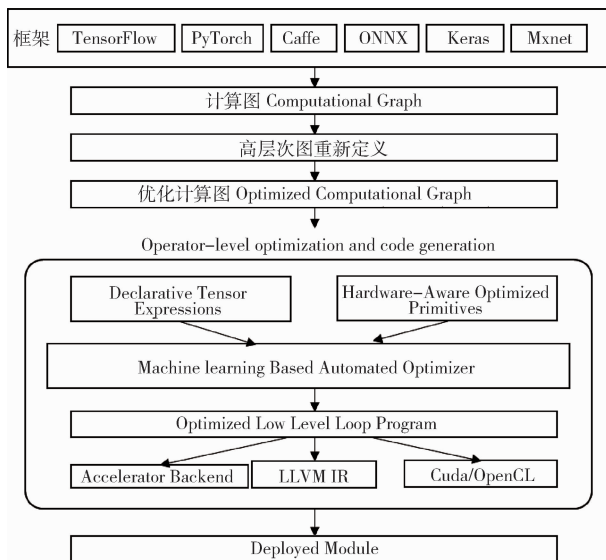


图 2 NNVM/TVM 系统概览图

2.2.3 深度学习样本增强技术

在很多军事场景中,往往难以找到充足的数据来完成训练,只能扩充数据集或对数据集进行增强^[11]。数据增强的目的是让有限的数据产生更多的等价数据。数据增强分为有监督的数据增强和无监督的数据增强两类^[12]。其中有监督的数据增强又可以分为单样本数据增强和多样本数据增强方法,无监督的数据增强分为生成新的数据和学习增强策略两个方向。

因为有监督的数据增强是利用研究者的经验来设计规则,存在两个比较明显的问题:其一,没有考虑不同任务的差异性;其二,数据增强的多样性和质量仍然不够好。因此无监督的数据增强方法逐渐开始被研究者重视,主要包括两类:

(1) 通过模型学习数据的分布,随机生成与训练数据集分布一致的图片,代表方法是生成对抗网络 GAN^[13]。

(2) 通过模型学习出适合当前任务的数据增强方法,代表方法是 Google 的 Auto Augment^[14],基本思路是使用增强学习从数据本身寻找最佳图像变换策略,对于不同的任务学习不同的增强方法。

军用数据增强未来的发展方向就是把目标数据集寻找最有效的数据增强策略这一过程自动化,

并且学习到的策略具有一定的可移植性。

2.3 一站式人工智能开发平台

针对当前对军用数据集搜集困难、训练标记门槛高、开发周期时间长等问题,需要构建一站式人工智能开发平台。该开发平台设计的基本架构如图 3 所示。

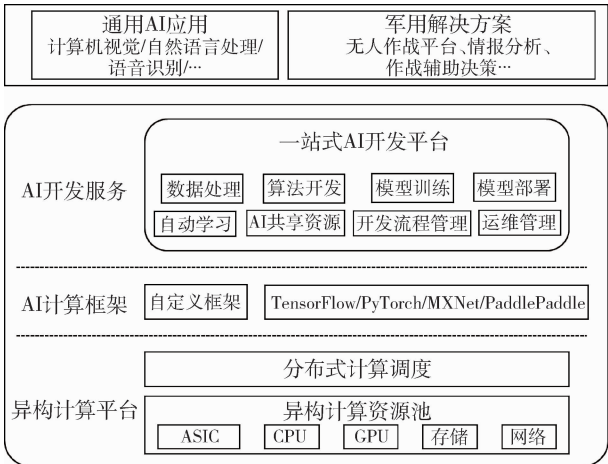


图 3 一站式人工智能开发平台

首先,底层能够支持异构计算平台,可以进行大规模分布式训练;拥有分布式高性能通信库,能够做到高性能吞吐;支持分布式数据,能够同时支持模型和数据并行;可以容器化部署和运维,能够快速移植。

在深度学习框架方面需要支持主流的训练框架,如 TensorFlow、PyTorch 和 PaddlePaddle 等,还可以预留框架接口,添加自定义的训练框架等。

在训练过程中,主要分为数据的预处理、算法开发、模型训练等。在模型推理阶段涉及模型的评测、模型库管理以及模型的部署。

首先数据处理是当前军用领域面临的第一个问题,数据准备与标注会占整个开发流程一半以上的时间,因此需要将其自动或半自动化。数据处理框架包含数据采集、数据筛选、数据增强、数据标注、数据集版本管理功能,支持自动化和半自动化的数据筛选、多种数据增强策略、自动化的数据预标注及辅助自动化标注工具。基于该框架可以实现数据标注全流程处理,大大提高开发效率。

算法开发服务应该支持所有主流的 AI 算法框架,兼容传统机器学习算法,并能够提供在线交互的算法开发环境。除此之外,还要预置常用的机器学习和深度学习算法模板。模型训练阶段最繁琐

的工作就是模型的超参数调优过程,所以应该建立模型参数的自动调优服务,这是实现自动深度学习的关键技术之一。

在模型推理和部署阶段,能够支持模型评测服务,支持模型库的管理功能。在部署方面能支持在线推理服务,结合平台的端云协同能力,也能一键部署到边缘和终端设备上。在部署端侧过程中需要解决上文所提到的深度学习编译器构建和模型压缩等关键技术。

2.4 嵌入式边缘智能开发平台

一站式开发平台主要是以面向数据中心的训练为主,在嵌入式端应用进行模型部署还需要一个面向边缘的智能计算平台。该平台分为轻量级开发和模型推理部署两部分。它首先需要支持一站式开发平台已训练模型的部署,解决软硬件框架的关键就是构建模型中间表示层。

以灵活的 FPGA 为例,硬件架构利用 DNN 加速设计实现神经网络加速计算。软件栈是根据自动化模型压缩流程设计的相关工具集。如图 4 所示,嵌入式边缘智能计算平台核心部分主要是模型的深度压缩工具,包括模型剪枝和优化工具、模型量化工具,提供剪枝和定点化运算;神经网络编译器,主要是将神经网络算法编译为可运行的指令流;运行时工具,是运行时支持环境,负责神经网络的加载、资源管理、调度。

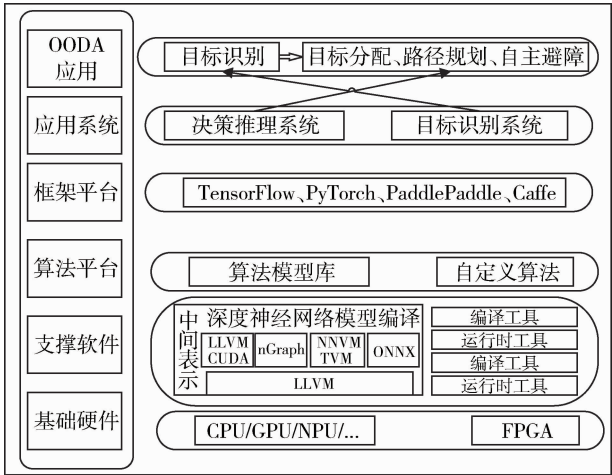


图 4 嵌入式边缘智能计算平台

作为部署和推理平台可以集成到一站式 AI 开发平台上。算法框架层以轻量级推理框架为主,可以看做是训练框架的精简版,一般只用来做推理,不考虑训练。最后,将算法平台集成到应用系统,

支撑智能无人平台的计算需求。

3 结束语

本文首先描述了基于深度学习技术的人工智能基础软硬件架构组成,从数据、算力和算法论证了人工智能软硬件在军用领域的潜在选择。考虑到后期自主安全的要求,在硬件层选用国产指令架构的 ASIC 芯片;在模型中间表示层选择开源并且生态良好的 ONNX 作为模型交换格式;在软件框架层前期使用 TensorFlow 框架为主,后期倾向于使用生态逐渐发展的百度 Paddle 框架进行替换。接着,简要分析了人工智能在军用领域面临的问题及解决思路,分析了软硬件架构中的关键技术和初步解决方案。最后,提出了适合军用领域的一站式开发平台和嵌入式边缘智能计算平台的架构设计思路。结合军用智能化、无人自主化的需求说明建立两者的必要性。人工智能基础软硬件架构需要解决的问题还非常多,但未来军用智能化计算的进程需要以此为基础。

参考文献

- [1] 黄建明,高大鹏.基于OODA环的作战对抗系统动力学模型[J].系统仿真学报,2012,24(3):561-564.
- [2] 韩冯飞.人工智能现状和发展[J].电脑知识与技术,2016,12(24):183-184.
- [3] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006,313(5786):504-507.
- [4] 张斌,付东.智能无人作战系统的发展[J].科技导报,2018,36(12):71-75.
- [5] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks[C]. European Conference on Computer Vision, 2013,8689:818-833.
- [6] MARCUS G. Deep learning: a critical appraisal[J]. arXiv preprint arXiv:1801.00631, 2018.
- [7] 郭圣明,贺筱媛,胡晓峰,等.军用信息系统智能化的挑战与趋势[J].控制理论与应用,2016,33(12):

1562-1571.

- [8] FINN C, ABBEEL P, LEVINE S. Model-agnostic meta-learning for fast adaptation of deep networks[C]. Proceedings of the 34th International Conference on Machine Learning, 2017,70:1126-1135.
- [9] CHENG Y, WANG D, ZHOU P, et al. A survey of model compression and acceleration for deep neural networks[J]. arXiv preprint arXiv:1710.09282, 2017.
- [10] HAN S, MAO H Z, DALLY W J. Deep compression: compressing deep neural networks with pruning, trained quantization and huffman coding[J]. arXiv preprint arXiv:1510.00149, 2015.
- [11] CHEN Z T, FU Y W, ZHANG Y D, et al. Semantic feature augmentation in few-shot learning[J]. arXiv preprint arXiv:1804.05298, 2018.
- [12] PEREZ L, WANG J. The effectiveness of data augmentation in image classification using deep learning[J]. Convolutional Neural Networks Vis. Recognit, 2017.
- [13] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Advances in Neural Information Processing Systems, 2014, 3: 2672-2680.
- [14] CUBUK E D, ZOPH B, MANE D, et al. AutoAugment: learning augmentation policies from data[J]. arXiv:Computer Vision and Pattern Recognition, 2018.

(收稿日期:2019-12-10)

作者简介:

杨建(1989-),男,博士,高级工程师,主要研究方向:图像处理、智能计算、人工智能等。

傅明驹(1986-),男,学士,工程师,主要研究方向:抗恶劣环境计算机设计与工程。

徐国强(1984-),通信作者,男,硕士,高级工程师,主要研究方向:人工智能、新型计算机体系、高性能计算等。E-mail:xuguogiang@jari.cn。

版权声明

经作者授权，本论文版权和信息网络传播权归属于《信息技术与网络安全》杂志，凡未经本刊书面同意任何机构、组织和个人不得擅自复印、汇编、翻译和进行信息网络传播。未经本刊书面同意，禁止一切互联网论文资源平台非法上传、收录本论文。

截至目前，本论文已经授权被中国期刊全文数据库（CNKI）、万方数据知识服务平台、中文科技期刊数据库（维普网）、JST 日本科技技术振兴机构数据库等数据库全文收录。

对于违反上述禁止行为并违法使用本论文的机构、组织和个人，本刊将采取一切必要法律行动来维护正当权益。

特此声明！

《信息技术与网络安全》编辑部
中国电子信息产业集团有限公司第六研究所