[引用格式] 葛悦涛，任 彦. 2020 年人工智能芯片技术发展综述[J]. 无人系统技术，2021，4(2)：14–19.

2020 年人工智能芯片技术发展综述

葛悦涛[1]，任 彦[2]

（1. 中国信息通信研究院安全研究所，北京 100191；2. 海鹰航空通用装备有限责任公司，北京 100074）

摘 要：对人工智能芯片领域的研究进行了综合评述，并对未来发展趋势进行了分析。首先分析了当前世界主要科技巨头纷纷布局人工智能芯片的现状；然后重点介绍了边缘侧低功耗人工智能芯片的特点与研发应用现状；最后讨论人工智能芯片发展趋势。综述表明，人工智能芯片在越来越多的场景中表现出广阔的应用前景，低功耗和面向通用人工智能的人工智能芯片研发已成为大趋势，类脑仿生芯片将持续扮演通用人工智能“探路者”角色，人工智能芯片将有助于我国在世界范围的科技竞争中实现弯道超车。

关键词：人工智能芯片；边缘智能；低功耗；神经网络；深度学习；类脑芯片

中图分类号：TP183 **文献标识码：**A **文章编号：**2096–5915(2021)02–14–06

DOI: 10.19942/j.issn.2096–5915.2021.2.013

Survey of Artificial Intelligence Chip in 2020

GE Yuetao[1], REN Yan[2]

(1. Institute of Security, China Academy of Information and Communications Technology, Beijing 100191, China;
2. Hiwing Aviation General Equipment Co., Ltd, Beijing 100074, China)

Abstract: In this paper, the research in the field of Artificial Intelligence (AI) chip is overviewed, and then the future development trends are discussed. This paper first analyzes the current situation of AI chip layout by the world’s major technology giants. Then this paper focuses on the characteristics and research, development and application status of edge-side low-power AI chip. Finally, the development trend of AI chip is described. The survey shows that, recently AI chips have gained broad application prospects in more and more scenes. The research and development of low-power and General AI (GAI) oriented AI chips have become the widely-recognized trends. Moreover, brain-like bionic chips will continue to play the role of “Pathfinder” of GAI, and AI chips will help China to overtake in the world-wide scientific and technological competition.

Key words: Artificial Intelligence Chip; Edge Intelligence; Low-Power; Neural Network; Deep Learning; Brain-Like Chip

1 引 言

当前人工智能各领域的算法和应用处在高速发展和快速迭代的阶段，而人工智能芯片已经成为支持人工智能产业的底层基础，拥有非常广阔的发展前景[1]。关于人工智能芯片的定义，可以从广义和狭义两个角度来阐释：首先，从广义角度，只要能够运行人工智能算法的芯片，都可以被视作人工智能芯片；其次，从狭义角度，人工智能芯片指针对人工智能算法做了特殊加速设计

收稿日期：2021–01–06；修回日期：2021–03–05
基金项目：装备预研联合基金项目（6141B08010102）

的芯片（现阶段的人工智能算法一般以深度学习算法为主，也可以包括其他机器学习算法），这也被视为通常意义下对人工智能芯片的定义[2-4]。此外，可用于人工智能计算任务的各类芯片，通常总称为泛人工智能类芯片。因此，人工智能芯片也称为“人工智能加速器”，即专门用于处理人工智能应用中的大量计算任务的模块（其他非计算任务仍由 CPU 负责）。

人工智能芯片目前有两种发展路径。第一种发展路径延续传统计算架构，旨在对硬件计算能力进行加速，主要以 GPU、FPGA、ASIC 等为代表，但 CPU 依旧发挥着不可替代的作用[5]。另外一种发展路径是彻底颠覆经典的冯・诺依曼计算架构，采用类脑神经结构来提升计算能力，以美国英特尔公司的 Loihi 芯片、美国 IBM 公司的 TrueNorth 芯片等为代表[6]。人工智能芯片发展路线图，可以归纳为如下三个阶段：短期目标，实现以异构计算为主加速各类应用算法的落地；中期目标，发展自重构、自学习、自适应、自组织的异构人工智能芯片来支持人工智能算法的演进和类人智能的升级；长期目标，向设计实现通用人工智能（General Artificial Intelligence，GAI）芯片的终极目标迈进。

虽然当前摩尔定律逐渐放缓，但作为推动人工智能技术不断进步和落地的硬件基础与极优选择，未来十年仍将是人工智能芯片发展的黄金时期——预计到 2021 年，我国的人工智能芯片产值预计将达到 52 亿美元。面对不断增长的市场需求，各类专门针对特定领域人工智能应用的人工智能芯片新颖设计理念和架构创新正在不断涌现、推陈出新。

2　人工智能芯片研究与应用现状

环球市场观察发布报告指出，全球人工智能芯片组的全球市场预计将从 2019 年的 80 亿美元增长到 2026 年的 700 亿美元。集成电路设计行业属于技术密集型行业，而人工智能芯片作为集成电路领域新兴的方向，在集成电路和人工智能方面有着双重技术门槛。目前，泛人工智能类芯片领域中的主要企业分为两类：第一类企业主要包括国际集成电路设计龙头企业以及主要以进行 IP 授权模式经营业务的企业，前者的代表性企业包括英特尔（Intel）、英伟达（Nvidia）、AMD、高通公司（Qualcomm）、NXP、Broadcom、赛灵思（Xilinx）、联发科、华为海思等，后者的代表性企业包括 ARM、Cadence、Synopsys 等。第二类企业主要是专业人工智能芯片设计公司，其代表性企业包括寒武纪、地平线机器人、Graphcore、Wave Computing 等。此外，目前业界重点研发和应用的人工智能芯片，按设计思路主要分为三大类：首先是专用于机器学习和深度学习（以深层神经网络算法为主）的训练和推理用加速芯片；其次是受生物脑启发设计的类脑仿生芯片；最后是可高效计算各类人工智能算法的通用人工智能芯片。

2020 年人工智能芯片行业大型并购频发，在短短两个月时间内便达成了总金额达到约 1000 亿美元的收购交易：9 月，美国英伟达公司宣布将以 400 亿美元现金加股票的形式收购美国 ARM 公司；10 月，美国 AMD 公司宣布将以 350 亿美元收购美国赛灵思公司，同期美国 Marvell 公司宣布将通过股票加现金的方式，以总价约 100 亿美元的价格收购模拟芯片制造商 Inphi 公司；美国英特尔公司延续将其主要精力集中在专用人工智能芯片领域的路线，于 11 月收购致力于创建用于建模和仿真的优化平台的美国 SigOpt 公司以加强其人工智能芯片业务，持续强化在人工智能芯片领域的实力。

英伟达在 2020 年依然扮演着人工智能芯片“领跑者”的重要角色。英伟达公布了其用于超级计算任务的 A100 人工智能芯片（图 1），这款基于第八代 Ampere 架构的芯片所采用的弹性计算技术能将每个芯片分割为多达七个独立实例来执行推理任务，人工智能算力提升 20 倍以上，被业界认为是史上最大性能飞跃。这是人类有史以来首次可以在一个平台上实现对横向扩展以及纵向扩展的负载的加速；此外，A100 人工智能芯片将在提高吞吐量的同时，降低数据中心的成本。

全球知名的数据中心解决方案提供商 VMware 宣布，将在 VMware 数据中心管理软件中首次使用英伟达的人工智能芯片，以提升数据中心效率。紧接着，英伟达发布全球唯一的千万亿级工作组服务器 NVIDIA DGX Station A100，配备四个英伟达 A100 人工智能芯片，具有高达 320GB 的 GPU 内存，加速满足位于全球各地的公司办公室、研究机构、实验室或家庭办公室中的办公团队对于机器学习和数据科学工作负载的强烈需求。

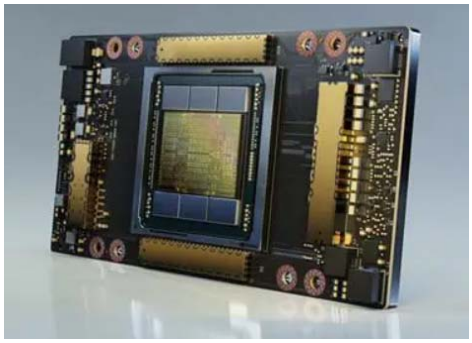

图 1　英伟达 A100 人工智能芯片
Fig. 1　NVIDIA’s A100 AI chip

英特尔则正在成为最贴近军方需求的人工智能硬件供应商。2020 年 10 月，英特尔宣布获批一项与美国军方合作项目的第二阶段合同，旨在帮助美国军方在美国国内生产更先进的人工智能芯片原型，这种封装技术能够将来自不同供应商的“小芯片”集成到一个封装中，从而实现把更多功能整合进一个更小的成品中，同时降低其功耗。

美国谷歌公司持续着力人工智能芯片的硬件优化，使它能更好地支持谷歌的人工智能技术，例如提升谷歌助手的交互体验和长时间保持激活的能力。美国谷歌公司提出通过人工智能程序推进专用人工智能芯片内部开发的设想，旨在建立一种良性循环：人工智能让芯片变得更好，经过改良的人工智能芯片又能增强人工智能算法，依此类推、迭代共进。相关研究已经证明，对于芯片电路布局设计任务，深度神经网络只花了 24 小时就解决了该问题，而人类设计是需要 6~8 周，并且前者的解决方案更好。谷歌宣布联手韩国三星公司共同开发的代号“白教堂”的自研人工智能芯片取得了重大进展，该芯片采用三星的 5 nm 半导体工艺打造、搭载 8 个 ARM 核心，预计 2021 年就可能应用在下一代谷歌手机和笔记本电脑上。

但是“前浪”面临“后浪”的强大挑战与竞争，这些后浪源于以往主营主业不在集成电路与设备制造的科技巨头的“跨界”与“掉转航向”，力求布局人工智能生态全产业链。以美国亚马逊公司为例，其 Alexa 语音助手的计算任务此前是由英伟达的芯片处理，但现在亚马逊 Alexa 语音助手已经在他们自研的 AWS Inferentia 人工智能芯片上运行。Inferentia 人工智能芯片（图 2）由四颗 NeuronCore 组成，每颗 NeuronCore 由以线性独立方式处理数据的大量小型数据处理单元（DPU）组成、实现一个“高性能脉动阵列矩阵乘法引擎”。此外，亚马逊宣布 Alexa 部分计算任务转向自研芯片，预计成本较英伟达 T4 可降低 30%，并降低了 25%的延迟，更低的延迟有助于 Alexa 的开发人员运行更先进的数据分析输入技术、降低用户的等待时间。

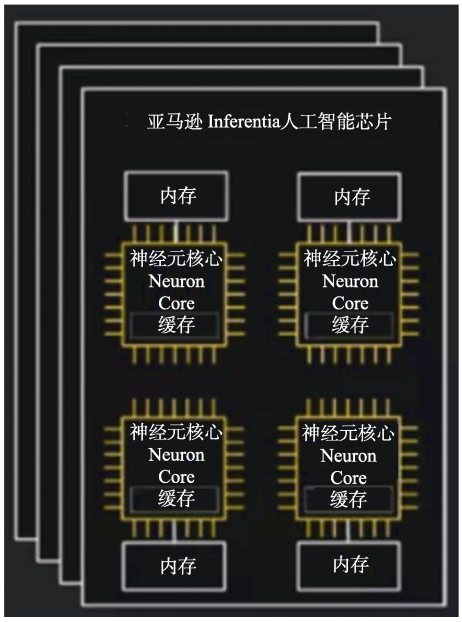

图 2　亚马逊 Inferentia 人工智能芯片
Fig. 2　Amazon’s Inferentia AI chip

3　边缘侧低功耗人工智能芯片研究进展

德勤发布《2020 科技、传媒和电信行业预测》报告，指出到 2024 年，边缘人工智能芯片销量预

计将超过 15 亿片;消费级边缘人工智能芯片市场规模远大于企业市场，但其增长速度可能相对较慢，2020 至 2024 年的复合年均增长率预计将为 18%，如图 3 所示。

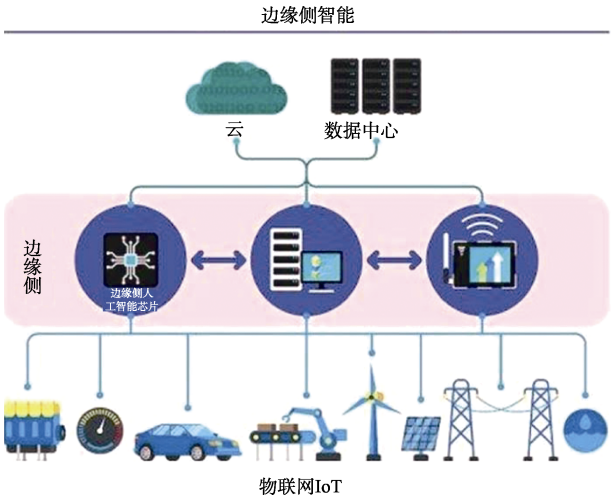

图 3　边缘人工智能芯片应用广泛

Fig. 3　Edge AI chips are widely used

边缘侧人工智能芯片的最成熟战场是自动驾驶领域[7-8]。ABI Research 预计，到 2027 年左右，对先进驾驶辅助系统的需求将增长两倍，但汽车行业已然将目光投向了更长远的全自动驾驶汽车和自动驾驶出租车，从 L2 和 L3 级先进驾驶辅助系统向 L4 和 L5 级全自动驾驶演进的过程中，高性能、低延迟和高能效的结合将是关键所在。目前自动驾驶汽车芯片市场占有领导地位的是英伟达，其在 2015 年就推出了车载计算平台，此后持续迭代，目前在自动驾驶芯片市场已处于优势地位;而以低功耗产品见长的 Imagination 等新兴公司的不断崛起，为边缘人工智能芯片领域注入了新的元素和活力。后者于 2020 年推出的第三代神经网络加速芯片 IMG Series4(其全新的多核架构可提供每秒 600 万亿次操作甚至更高的超高性能) 所面向的市场，由其第二代神经网络加速芯片所聚焦的移动设备和汽车市场，进一步拓展到智能相机监控、消费电子 (尤其是数字电视)、低功耗物联网 (IoT) 智能设备领域。由于以自动驾驶为代表的众多人工智能芯片应用场景都是基于三维数据，因此当前在设计人工智能芯片时，无法回避三维数据的实时高性能计算问题[9-11]。尽管目前世界主要科技巨头企业都对人工智能芯片投入了大量的人力物力,由于三维场景的复杂性，目前这些芯片还远远没有达到令人满意的地步。

人工智能芯片的应用发轫于计算机视觉计算与图像处理，在 2020 年，该领域的边缘人工智能芯片尝试依然产生了瞩目的成果。该领域人工智能芯片无须生成实际图像，而是可以分析所看到的视频并仅提供有关其前面内容的元数据，而不是显示器视野中的内容。由于数据依然在边缘侧、没有被发送到远程服务器，因此黑客截获敏感图像或视频的机会大大减少，这将有助于减轻对隐私的担忧，在智能手机、智能相机、视频安防等领域有着广阔的应用前景[12-13]。据统计，目前边缘人工智能芯片绝大部分将流向高端智能手机，当前在用的所有消费级边缘人工智能芯片中超过 70%均用于智能手机。苹果公司已经证明了结合人工智能和影像技术以通过 iPhone 定制设计的神经引擎处理器支持的 Face ID 生物识别技术创建更安全的系统的功效；华为技术有限公司和美国谷歌公司还在其智能手机中配备了专用的人工智能芯片，以协助图像处理。这些设备上的人工智能芯片所执行的正是“边缘计算”模式：在网络边缘处理复杂的人工智能和机器学习任务，而不是将数据来回发送给服务器。日本索尼公司与美国微软公司达成微型人工智能芯片交易，合作将人工智能功能嵌入索尼最新的成像人工智能芯片中，该新模块的最大优势在于它内置了处理器和内存，可以使用像微软的 Azure 这样的人工智能技术来分析视频，但是在一个独立的系统中，它比现有方法更快、更简单、更安全地操作。美国初创人工智能芯片公司 Kneron 推出专门为边缘计算设备设计的 KL720 AI SoC，该芯片在小区域内可提供高计算性能且功耗低，可以处理 1080P 的 4K 静止图像和视频，并提供面部识别的 3D 传感功能，还为自然语言处理应用程序提供了新的音频识别工具。

边缘人工智能芯片领域的另一个先行者是美国谷歌公司。谷歌的 Edge TPU 边缘人工智能芯

片是专为在边缘运行 TensorFlow Lite ML 模型而设计的 ASIC 芯片，可用于越来越多的工业使用场景，如预测性维护、异常检测、机器视觉、机器人学、语音识别等，可以应用于制造、本地部署、医疗保健、零售、智能空间、交通运输等各个领域，具有体型小、功耗低、性能出色的优势，可以在边缘部署高精度人工智能。2020 年 2 月，谷歌发布首个基于 Edge TPU 人工智能芯片的全球人工智能模型平台——Model Play，这是一款面向全球用户的人工智能模型资源交流与交易平台，为机器学习与深度学习提供丰富多样化的功能模型，兼容多类市场主流的边缘计算人工智能芯片，包括谷歌 CoralEdge TPU、英特尔 Movidius、英伟达 Jetson Nano 等，帮助用户快速创建和部署模型，显著提高模型开发和应用效率，降低人工智能开发及应用门槛。

4 人工智能芯片发展趋势分析

当前，随着全球人工智能产业的蓬勃发展和技术产品的广泛落地，人工智能芯片相较于传统处理器已经成为人工智能算法实现的更优选择，而且只有将人工智能算法与人工智能芯片充分融合与协同，才能够真正推动人工智能技术的商用进程。因此，人工智能芯片被公认为是未来人工智能时代的战略制高点。人工智能芯片未来发展趋势，可以概述为以下三个方面。

一是低功耗人工智能芯片成为“万物互联万物生”的智能物联网时代的标配。近年来，随着物联网技术的发展与产品应用，边缘侧的智能处理所扮演的角色逐渐加码，因此催生了“人工智能物联网（AIoT）”概念[14-15]。通常而言，人工智能物联网是人工智能与物联网的结合，通过将物联网末梢节点（如传感器）采集的海量多源异构数据存储于云端或者边缘侧，并在云端或者边缘侧运行人工智能、大数据、云计算等技术手段形成更高形式的人工智能，实现万物数据化、泛在智联化[16-17]。人工智能物联网应用场景对硬件设备的低功耗要求极高，需要硬件设备兼顾高性能、强智能和低功耗等特点——这已经成为人工智能物联网应用场景下智能硬件的设计和实现的主要要求，而智能硬件的设计过程也必须结合特定物联网场景、从应用需求出发，有针对性、定制化设计人工智能芯片架构与集成方案，才能在保障性能的同时降低功耗。

二是面向通用人工智能的人工智能芯片成为大趋势。目前业界尚没有出现一款通用人工智能芯片，决定了人工智能目前还无法深刻变革人类生活方式，因此设计实现面向通用人工智能的人工智能芯片成为相关研发的终极目标——通用人工智能芯片是指能够支持和加速通用人工智能计算的芯片。在朝通用人工智能芯片前进的道路上所面临的挑战，包括适应人工智能和计算架构通用性、适应人工智能技术的复杂性等，同时需要重点考量摩尔定律的逐渐失效和冯·诺依曼架构的瓶颈所带来的技术挑战与应用难度。

三是类脑仿生芯片将持续扮演通用人工智能“探路者”角色。目前类脑仿生芯片的主流理念是采用神经拟态工程设计的神经拟态芯片。神经拟态芯片采用电子技术模拟已经被证明的生物脑的运作规则，从而构建类似于生物脑的电子芯片，即“仿生电子脑”。神经拟态计算在算法以及芯片的设计上可以实现以低于 1000 倍的功耗去完成同样效果的模型训练。因此，神经拟态芯片是一种环境友好型的芯片，其体积小、功耗低的特点，符合生物进化最本质的优势。2020 年 6 月，Gartner 发布报告预测，到 2025 年神经拟态芯片有望取代 GPU，成为用于人工智能系统的主要芯片之一。神经拟态研究陆续在全世界范围内开展，并且受到了各国政府的重视和支持。受脑结构研究的成果启发，复杂神经网络在计算上具有低功耗、低延迟、高速处理、时空联合等特点。美国苹果公司一直是类脑仿生芯片研发领域的佼佼者：2020 年 8 月，美国苹果公司公布其最新 A14 仿生芯片，该芯片的 CPU 性能相比上一代 A13 仿生芯片提升 40%，GPU 性能相比上一代仿生芯片提升 50%，领先于目前安卓设备搭载的任何处理器技术，包括英特尔芯片；A14 仿生芯片还搭载了定制技术，

这些技术可以驱动速度更快的神经引擎，这将使 iPad Air 在机器学习方面变得更强大。2020 年 11 月，苹果公布 A14X 仿生芯片的 CPU 和 GPU 性能基准，与 A12Z 仿生芯片相比，多核测试的性能提高了 35%。

5　结束语

随着信息化和智能化逐渐渗透进入能源、交通、农业、公共事业等更多行业的商业应用场景中，考虑到智能化任务运算力需求，以及传输带宽、数据安全、功耗、延时等客观条件限制，人工智能芯片在越来越多的场景中展现出广阔的应用前景和旺盛的生命力。根据 ABI Research 数据，人工智能芯片市场规模到 2024 年预计可以达到 100 亿美元。而人工智能芯片市场的持续火热，为我国在全球人工智能科技竞赛中“弯道超车”提供了有利机遇：长期以来，我国在 CPU、GPU、DSP 等处理器设计上一直处于追赶地位，绝大部分芯片设计企业过度依靠欧美的 IP 核设计芯片，在自主创新上受到了极大的限制——智能电子硬件设备及集成电路关键技术面临严峻的“卡脖子”挑战和持续的潜在“断供”风险。“危局中开新局”，人工智能产业的快速兴起与普遍应用，无疑为中国在以人工智能芯片为代表的新兴处理器领域实现“弯道超车”提供了绝佳的历史机遇——人工智能领域的应用目前还处于面向行业应用阶段，生态上尚未形成完全竞争垄断和技术壁垒，因此国产处理器厂商与国外竞争对手在人工智能芯片研发（乃至人工智能）这一全新赛场上处于同一起跑线上。因此，基于新兴技术和应用市场，不仅在突围人工智能芯片领域，中国在建立人工智能生态圈方面也将大有可为。

参 考 文 献

[1] Momose H, Kaneko T, Asai T. Systems and circuits for AI chips and their trends[J]. Japanese Journal of Applied Physics, 2020, 59: 050502.

[2] James A P. Towards strong AI with analog neural chips[C]. 2020 IEEE International Symposium on Circuits and Systems (ISCAS 2020), Seville, Spain, 2020.

[3] 汪鑫. 人工智能芯片的概念和应用分析[J].中国新通信, 2020, 22(20): 112-113.

[4] 尹首一. 人工智能芯片概述[J]. 微纳电子与智能制造, 2019, 1(2): 7-11.

[5] 施羽暇. 人工智能芯片技术体系研究综述[J]. 电信科学, 2019, 35(4): 114-119.

[6] Mashford B S, Jimeno-Yepes A, Kiral-Kornek I, et al. Neural-network-based analysis of EEG data using the neuromorphic TrueNorth chip for brain-machine interfaces[J]. IBM Journal of Research and Development, 2017, 61: 7.

[7] 吴昊, 陈虎, 李俊波. 浅析人工智能技术的发展与应用[J]. 信息系统工程, 2020(6): 69-70.

[8] 冯晓辉, 王哲, 李雅琪. 智能驾驶领域发展态势与展望[J]. 人工智能, 2018(6): 26-36.

[9] Chen J, Bai T. SAANet: Spatial adaptive alignment network for object detection in automatic driving[J]. Image and Vision Computing, 2020, 94: 103873.

[10] 黄漫, 黄勃, 高永彬. 引入深度补全与实例分割的三维目标检测[J]. 传感器与微系统, 2021, 40(01): 129-132.

[11] 田永林, 沈宇, 李强, 等. 平行点云: 虚实互动的点云生成与三维模型进化方法[J]. 自动化学报, 2020, 46(12): 2572-2582.

[12] Mazzia V, Khaliq A, Salvetti F,et al. Real-Time apple detection system using embedded systems with hardware accelerators: An edge AI application[J]. IEEE Access, 2020, 8: 9102-9114.

[13] 李理. 2019 年边缘计算技术发展研究[J]. 无人系统技术, 2020, 3(02): 58-62.

[14] Wu C, He Y, Tsang K F, et al. The IDex case study on the safety measures of AIoT-based railway infrastructures[C]. 2020 IEEE International Symposium on Product Compliance Engineering- Asia (ISPCE-CN 2020), Chongqing, China, November 6-8, 2020.

[15] Ma W, Nian C, Xu H. Application of AIoT in wireless image transmission to rotating machinery[C]. The 3rd IEEE International Conference on Knowledge Innovation and Invention (ICKII 2020), Kaohsiung, Taiwan, China, August 21-23, 2020.

[16] 张辉. 人工智能技术在物联网中的运用探析[J]. 中国设备工程, 2021(1):28-29.

[17] 肖明华, 李琳, 卢镭. 人工智能与计算智能在物联网方面的应用分析[J]. 中小企业管理与科技(上旬刊), 2021(1): 171-173.

作者简介：

葛悦涛（1982–），男，博士，高级工程师，主要研究方向为人工智能、网络与信息安全。

任　彦（1979–），女，硕士，工程师，主要研究方向为科研项目管理、无形资产管理。