|  |
| --- |
| [成都信息工程大学](http://www.cuit.edu.cn/) |

本科课程技术报告

**计算机组成原理**

**专 业：\_计算机科学与技术**

**班 级：** **计应173**

**姓 名：**  \_\_**熊术文**

**学 号：** 2017051107

**指导教师：\_\_\_\_**\_\_\_\_张永清\_\_\_\_\_**\_**

**成 绩：\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**成都信息工程大学计算机学院**

**2019年06月19日**

**人工智能芯片浅析**

**（熊术文 成都信息工程大学）**

**[摘要]**人工智能芯片是人工智能时代竞争的战略制高点。人工智能作为信息时代的尖端科技，站在全球科技革命和产业变革的制高点，已经成为推动经济社会发展的新引擎。从人工智能技术发展看，人工智能芯片承载了人工智能所需的计算，是人工智能技术发展的基础。目前，人工智能技术正在面临前所未有的算力挑战，CPU等传统计算架构已无法满足人工智能对于并行计算能力的日益需求，算力的提升必须通过人工智能芯片实现，人工智能芯片成为人工智能技术的突破口。本文将围绕人工智能芯片的发展现状，应用场景，以及发展前景做简单分析。

**[关键词]：**人工智能，芯片，深度学习

**一．引言**

芯片技术在国际上一直都是一个热门话题。世界各国之间，特别是中美两国之间芯片的争端日益严峻。几年前的中兴事件让国人认识到了中国芯片产业的落后局面，并开始思考如何赶超国外。而今年美国和华为在芯片方面的争端，让国人再一次认识到了大力发展芯片的重要性。芯片是20世纪最伟大的工程产品之一、随着科技的不断发展，人类对芯片的要求也越来越高，进入21世纪以来，随着人工智能技术的飞速发展，“人工智能芯片”也引起了越来越多的人的兴趣。所以在当今社会，发展人工智能芯片便显得至关重要。

**二.人工智能芯片的发展现状**

**2.1人工智能芯片技术路径及其发展态势**

目前。主流的人工智能芯片基本以 GPU、FPGA、ASIC以及类脑芯片为主，它们各自发挥优势，多种类型的芯片组合应用，呈现躲技术路径并行发展态势。

深度学习引领了本轮人工智能发展热潮。随着深度学习模型算法对大规模并行计算能力的需求不断激增，GPU、FPGA、等通用芯片在性能、功耗等方面的瓶颈，无法满足人工智能巨大的算力要求。ASIC 等专用芯片应运而生，满足不同应用、不同算法和不同终端的需求。但目前AI算法仍处于快速演进阶段，还不存在适应所有应用的通用人工智能算法，人工智能芯片也就没有确定的架构。这使得定制化的ASIC专用芯片开发周期长，迭代进化快，资金和技术风险都很高。因此，人工智能芯片技术还存在着巨大的探索空间和创新空间，尤其是算法基础理论研究和芯片设计架构创新。

从AI芯片发展的大趋势来看，从确定算法、应用场景的定制化AI加速芯片向具备更高灵活性、适应性的通用智能芯片发展是AI芯片技术发展的必然方向。[1]

**2.2“寒武纪”发布新一代人工智能芯片**

“寒武纪力争在3年后占有中国高性能智能芯片市场30%的份额，并使全世界10亿台以上的智能终端设备集成有寒武纪终端智能处理器。如果这两个目标实现，寒武纪将初步支撑起中国主导的国际智能产业生态。”6日，寒武纪科技发布新一代AI芯片，中科院计算所研究员、寒武纪科技CEO陈天石描述了芯片产品研发路线图。

早在2016年,全球首款商用深度学习处理器IP产品一“寒武纪 1A处理器”问世，达到了传统四核通用CPU 25倍以上的性能和50倍以上的能效。2017年9月，寒武纪1A处理器被搭载于华为Mate 10系列手机中。值得关注的是，中科曙光日前发布的专用AI服务器“Phaneron”，就搭载了自主知识产权的寒武纪芯片。

“寒武纪公 司是中科院计算所在处理器与人工智能交叉领域超前布局的结晶。”中科院计算所所长孙凝晖表示，人工智能经历60余载沉浮，如今迎来收获季节。寒武纪公司在智能芯片领域占据全球领先地位，通过与产业上下游通力合作，有望引领中国人工智能产业“变道超车”。

其间，寒武纪科技发布了面向云端的高性能智能处理器产品线，以及三款全新的智能处理器产品:面向低功耗场景视觉应用的寒武纪1H8，拥有更广泛通用性和更高性能的寒武纪1H16,以及面向智能驾驶领域的寒武纪1M。陈天石介绍,与寒武纪1A相比, 3款新品在功耗、能效比、成本开销等方面进行优化，性能功耗比再次实现飞跃，适用范围覆盖图像识别、安防监控、智能驾驶、无人机、语音识别、自然语言处理等重点应用领域。[2]

**2.2人工智能芯片分类**

**2.2.1 GPU**

图像处理单元Graphics Processing Unit，GPU)最初是为了进行图像处理而被开发出来的芯片。图像处理问题与神经网络有着类似的特点:都需要进行大量的、重复而简单的运算。因此GPU近几年也越来越多地被应用于神经网络问题。

GPU的突出特点是并行、及时处理，逻辑功能弱而计算能力强。由于GPU出色的计算性能和比较通用的计算结构，它常被用于神经网络的训练，当前运用较多的就是NVIDIA公司各种产品。

**2.2.2 ASIC**

GPU虽然在计算性能上已经有了很大的提高，但归根结底仍然是一种泛用化的芯片，在预测问题中很可能存在性能过剩(可以理解为工厂为了加速生产引进了一大批新设备，但在不同的生产任务中需要的设备数量是不-定的，这就使得在某些生产任务中设备无法得到充分使用，造成浪费)，使得芯片发挥不出他应有的性能。同时，这种通用化的结构设计也使得芯片在面对一些特殊结构的神经网络时捉襟见肘。而面向具体应用的集成电路(Application Specific Integrated Circuit，ASIC)就是为了解决这个问题而诞生的。简单来说，当神经网络训练完成后，其结构也就确定了。

可以根据这个结构设计专用的芯片来承载该神经网络，这不仅能够最大效率地发挥芯片性能，更能提升计算速度，使得预测能够进一步加速。但是要注意的是，芯片设计的成本是比较高的，因此如果是为了单次、少量的应用而开发ASIC是得不偿失的。目前得到应用的ASIC芯片有Intel的Nights Mill，Google的TPU以及中国首款嵌入式神经网络处理器芯片星光智能一号等。

**2.2.3 FPGA**

ASIC的突出特点是集成化、特例化，而现场可编程门阵列(Field 一Programmnable Gate Array，FPGA)的特点则是灵活性高。可以将GPGA理解为积木，采取不同的方式搭建，就能够得到不同的结构，进而实现不同的功能。单纯地从性能上来讲，FPGA与ASIC差距甚小，但是FPGA有着致命的缺陷：面积太大。众所周知，集成电路的“集成”二字强调的是“占用面积小”，过大的面积会使得集成电路丧失其本身的意义。因此，对于少量应用或者结构验证等不会大量开发的结构搭建而言，FPGA是性价比最高的一种方式。但对于大规模开发、大规模生产问题而言，ASIC才是成本最低、收益最高的方式。

目前，FPGA在AI芯片上的应用有Xilinx的华为云以及Intel的5G无线和自动驾驶方面等。[3]

**三.人工智能芯片的应用**

**3.1 安全行业的应用**

在近日举行的2018年中国国际社会公共安全产品博览会上注意到，无论是城市还是居家，安全防护早已有了人工智能的“加持”助力:秒速锁定视频目标、秒速发现火险、飞行跟踪发现异常等，人工智能正在让我们的生活更加安全。

人工智能技术已经在安防领域得到应用，产生了一大批出色的产品:芯片、软件、硬件无人机令……人眼花缭乱。国内知名的人工智能芯片企业寒武纪，就在展会上展示了最新研发的MLU100机器学习处理器芯片，单块芯片每秒钟峰值运算能力达128万亿次。技术人员现场演示了这款芯片的强大实力，32路摄像头监控，呈现32幅不同的电视画面。画面里人来人往，车流不息。但随着镜头视角变换，画面里每一个人、每一辆车的信息都能被即时捕获，并同时显示其基本信息。

家庭使用的摄像头也有了人工智能的助力，手机远端查看、遥控画面存储早已是标配。来自中国人民公安大学的专家表示，城市、居家安防引入人工智能技术是未来的趋势，它既可以减少公共资源的消耗，又能提高安防的水平，保护每一位市民的安全。[4]

**3.2无人驾驶领域的应用**

新思科技近日宣布，飞步科技选择了经过硅验证的新思科技DesignWare接口、安全、ARC EM Safety Island处理器以及嵌入式存储器测试与修复IP组合，用于先进汽车应用上的高性能人工智能芯片。飞步科技致力于面向对象和车道检测、交通灯检测、车辆定位、运动分析和环境识别开发智能芯片，于先进的驾驶辅助系统(ADAS)和无人驾驶汽车。

飞步科技选择了新思科技DesignWare IP组合，以满足对专业化处理、存储器性能、实时数据连接和严格安全的人工智能要求。面向汽车应用的新思科技DesignWare IP已获ISO26262ASILB或ASILDReady认证;符合严格的AEC-Q100温度要求;支持汽车质量管理;并且包括全面的汽车安全套件，帮助飞步科技加速芯片级功能安全评估。

飞步科技美国首席技术官 Hang Nguyen示:“飞步科技拥有专有深度学习算法和自定义加速体系结构,正在革新智能信息处理,在能源效率和性能上有了巨大的提高。我们的ADAS和无人驾驶汽车芯片使用新思科技的人工智能和汽车级DesignWare IP组合,帮助芯片集成智能功能,并加速系统级ISO 26262合规之路。

新思科技IP营销副总裁John Koeter表示:“用于汽车和云计算等应用的新时代人工智能芯片需要众多不同的IP来满足不同的处理、内存、连接和安全要求。新思科技与所有细分市场的领先人工智能芯片提供商(如飞步科技合作来降低集成风险,并加快新一代智能系统的上市时间。

**四.人工智能芯片的发展前景**

**4.1更高效的大卷积解构/复用**

在标准 SIMD 的基础上，CNN 由于其特殊的复用机制，可以进一步减少总线上的数据通信。而复用这一概念，在超大型神经网络中就显得格外重要。 如何合理地分解、映射这些超大卷积到有效的硬件上成为了一个值得研究的方向。

**4.2更低的 Inference 计算/存储位宽**

AI 芯片最大的演进方向之一可能就是神经网络参数/计算位宽的迅速减少——从 32 位浮点到 16 位浮点/定点、8 位定点，甚至是 4 位定点。在理论计算领域，2 位甚至 1 位参数位宽，都已经逐渐进入实践领域。

**4.3更多样的存储器定制设计**

当计算部件不再成为神经网络加速器的设计瓶颈时，如何减少存储器的访问延时将会成为下一个研究方向。通常，离计算越近的存储器速度越快，每字节的成本也越高，同时容量也越受限，因此新型的存储结构也将应运而生。

**4.4更稀疏的大规模向量实现**

神经网络虽然大，但是，实际上有很多以零为输入的情况， 此时稀疏计算可以高效的减少无用能效。来自哈佛大学的团队就该问题提出了优化的五级流水线结，在最后一级输出了触发信号。在 Activation层后对下一次计算的必要性进行预先判断，如果发现这是一个稀疏节点，则触发 SKIP 信号，避免乘法运算的功耗，以达到减少无用功耗的目的。

**4.5计算和存储一体化**

计算和存储一体化（process-in-memory）技术，其要点是通过使用新型非易失性存储（如 ReRAM）器件，在存储阵列里面加上神经网络计算功能，从而省去数据搬移操作，即实现了计算存储一体化的神经网络处理，在功耗性能方面可以获得显著提升。

**五.结论**

随着社会科技的发展，特别是人工智能技术的发展，以及当前我国和美国之间的芯片的争端，抓紧人工智能芯片的研发，对我国的发展有着至关重要的作用。要想打破外国对我国的技术封锁，走向芯片技术的独立，就必须把握住这个时代的命脉。而这个命脉正是人工智能芯片。

参考文献：

[1]李丽婷.《人工智能芯片技术进展及产业发展研究报告》. 厦门市科学技术信息研究院. 《厦门科技》·2019年第1期

[3]《光学精密机械》 2017年第4期

[2]缪希辰. 《人工智能芯片分类及反思》. 《科技传播》 2019年第5期

[4]《北京晚报》.《人工智能让我们的生活更安全》. 《共产党员(河北)》 · 2018年第23期

[5]人工智能学家.《深度剖析人工智能芯片的发展现状及趋势》. 《千家网》