2018

人工智能芯片研究报告

AMiner 研究报告第十四期

清华-中国工程院知识智能联合实验室

2018 年 10 月

Contents 目录

图表目录

摘要

2010 年以来，由于大数据产业的发展，数据量呈现爆炸性增长态势，而传统的计算架构又无法支撑深度学习的大规模并行计算需求，于是研究界对 AI 芯片进行了新一轮的技术研发与应用研究。AI 芯片是人工智能时代的技术核心之一，决定了平台的基础架构和发展生态。本报告在此背景下，对人工智能芯片的发展现状进行了简单梳理，包括以下内容：

人工智能芯片概念。首先对人工智能芯片相关概念、技术路线以及各自特点进行介绍，接着对国外、国内 AI 芯片的发展历程及现状进行梳理。

AI 芯片的技术特点及局限性。对 AI 芯片的几个技术流派进行介绍。

AI 芯片厂商介绍。对 AI 芯片领域的国内外代表性厂商进行介绍。

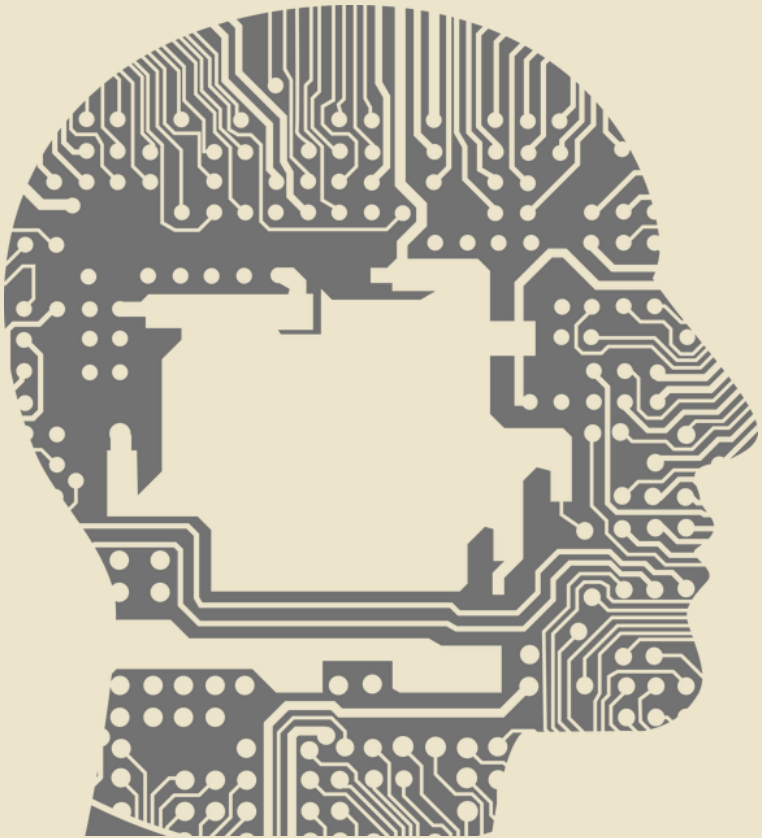
AI 芯片领域专家介绍。通过 AMiner 大数据平台对 AMiner 的人工智能芯片人才库进行数据挖掘，统计分析领域内学者分布及迁徙。同时，介绍了目前 AI 芯片领域的国内外代表性研究学者。

AI 芯片应用领域介绍。AI 芯片已经渗透到日常生活的方方面面，本报告主要对智能手机、ADAS、CV、VR、语音交互设备、机器人等方向的应用进行介绍。

AI 芯片的发展趋势介绍。人工智能的发展历经波折，如今得益于大数据的供给、深度学习算法的革新以及硬件技术的提升，AI 芯片以不可阻挡的势态飞速发展。AI 芯片的算力提高、功耗降低及更合理的算法实现必然是将来的发展趋势。

1 concept

概述篇



1 概述篇

人工智能（Artificial Intelligence，AI）芯片的定义：从广义上讲只要能够运行人工智能算法的芯片都叫作 AI 芯片。但是通常意义上的 AI 芯片指的是针对人工智能算法做了特殊加速设计的芯片，现阶段，这些人工智能算法一般以深度学习算法为主，也可以包括其它机器学习算法。人工智能与深度学习的关系如图 1 所示。

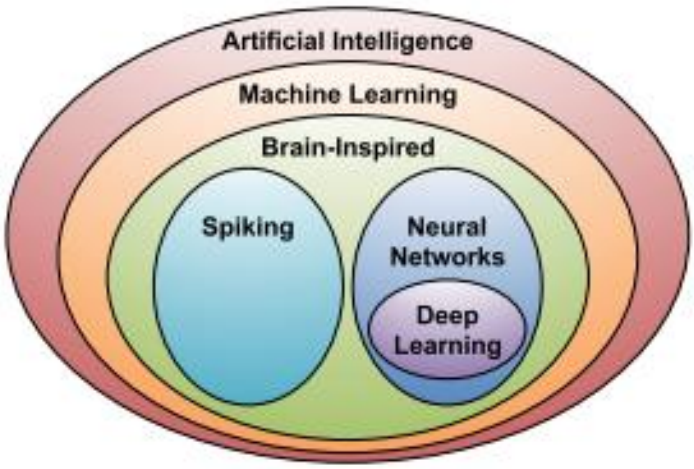


图 1 人工智能与深度学习

深度学习算法，通常是基于接收到的连续数值，通过学习处理，并输出连续数值的过程，实质上并不能完全模仿生物大脑的运作机制。基于这一现实，研究界还提出了 SNN（Spiking Neural Network，脉冲神经网络）模型。作为第三代神经网络模型，SNN 更贴近生物神经网络——除了神经元和突触模型更贴近生物神经元与突触之外，SNN 还将时域信息引入了计算模型。目前基于 SNN 的 AI 芯片主要以 IBM 的 TrueNorth、Intel 的 Loihi 以及国内的清华大学天机芯为代表。

1.1 AI 芯片的分类

（1）AI 芯片按技术架构分类

GPU（Graphics Processing Unit，图形处理单元）：在传统的冯·诺依曼结构中，CPU 每执行一条指令都需要从存储器中读取数据，根据指令对数据进行相应的操作。从这个特点可以看出，CPU 的主要职责并不只是数据运算，还需要执行存储读取、指令分析、分支跳转等命令。深度学习算法通常需要进行海量的数据处理，用 CPU 执行算法时，CPU 将花费大量的时间在数据/指令的读取分析上，而 CPU 的频率、内存的带宽等条件又不可能无限制提高，因此限制了处理器的性能。而 GPU 的控制相对简单，大部分的晶体管可以组成各类专用电路、多条流水线，使得 GPU 的计算速度远高于 CPU；同时 GPU 拥有了更加强大的浮点运算能力，可以缓解深度学习算法的训练难题，释放人工智能的潜能。

但 GPU 无法单独工作，必须由 CPU 进行控制调用才能工作，而且功耗比较高。

半定制化的 FPGA：FPGA（Field Programmable Gate Array）全称“现场可编程门阵列”，其基本原理是在 FPGA 芯片内集成大量的基本门电路以及存储器，用户可以通过更新 FPGA 配置文件来定义这些门电路以及存储器之间的连线。

与 GPU 不同，FPGA 同时拥有硬件流水线并行和数据并行处理能力，适用于以硬件流水线方式处理一条数据，且整数运算性能更高，因此常用于深度学习算法中的推断阶段。不过 FPGA 通过硬件的配置实现软件算法，因此在实现复杂算法方面有一定的难度。将 FPGA 和 CPU 对比可以发现两个特点，一是 FPGA 没有内存和控制所带来的存储和读取部分，速度更快，二是 FPGA 没有读取指令操作，所以功耗更低。劣势是价格比较高、编程复杂、整体运算能力不是很高。目前国内的 AI 芯片公司如深鉴科技就提供基于 FPGA 的解决方案。

全定制化 ASIC：ASICc（Application-Specific Integrated Circuit）专用集成电路，是专用定制芯片，即为实现特定要求而定制的芯片。定制的特性有助于提高 ASIC 的性能功耗比，缺点是电路设计需要定制，相对开发周期长，功能难以扩展。但在功耗、可靠性、集成度等方面都有优势，尤其在要求高性能、低功耗的移动应用端体现明显。谷歌的 TPU、寒武纪的 GPU，地平线的 BPU 都属于 ASIC 芯片。谷歌的 TPU 比 CPU 和 GPU 的方案快 30 至 80 倍，与 CPU 和 GPU 相比，TPU 把控制电路进行了简化，因此减少了芯片的面积，降低了功耗。

神经拟态芯片：神经拟态计算是模拟生物神经网络的计算机制。神经拟态计算从结构层面去逼近大脑，其研究工作还可进一步分为两个层次，一是神经网络层面，与之相应的是神经拟态架构和处理器，如 IBM 的 TrueNorth 芯片，这种芯片把定制化的数字处理内核当作神经元，把内存作为突触。其逻辑结构与传统冯·诺依曼结构不同：它的内存、CPU 和通信部件完全集成在一起，因此信息的处理在本地进行，克服了传统计算机内存与 CPU 之间的速度瓶颈问题。同时神经元之间可以方便快捷地相互沟通，只要接收到其他神经元发过来的脉冲(动作电位)，这些神经元就会同时做动作。二是神经元与神经突触层面，与之相应的是元器件层面的创新。如 IBM 苏黎世研究中心宣布制造出世界上首个人造纳米尺度的随机相变神经元，可实现高速无监督学习。

（2）AI 芯片按功能分类

根据机器学习算法步骤，可分为训练（training）和推断（inference）两个环节：

训练环节通常需要通过大量的数据输入，训练出一个复杂的深度神经网络模型。训练过程由于涉及海量的训练数据和复杂的深度神经网络结构，运算量巨大，需要庞大的计算规模，对于处理器的计算能力、精度、可扩展性等性能要求很高。目前市场上通常使用英伟达的 GPU 集群来完成，Google 的 TPU2.0/3.0 也支持训练环节的深度网络加速。

推断环节是指利用训练好的模型，使用新的数据去“推断”出各种结论。这个环节的计算量相对训练环节少很多，但仍然会涉及到大量的矩阵运算。在推断环节中，除了使用CPU 或 GPU 进行运算外，FPGA 以及 ASIC 均能发挥重大作用。

（3）AI 芯片按应用场景分类

主要分为用于服务器端（云端）和用于移动端（终端）两大类。

服务器端：在深度学习的训练阶段，由于数据量及运算量巨大，单一处理器几乎不可能独立完成一个模型的训练过程，因此，负责 AI 算法的芯片采用的是高性能计算的技术路线，一方面要支持尽可能多的网络结构以保证算法的正确率和泛化能力；另一方面必须支持浮点数运算；而且为了能够提升性能必须支持阵列式结构（即可以把多块芯片组成一个计算阵列以加速运算）。在推断阶段，由于训练出来的深度神经网络模型仍非常复杂，推断过程仍然属于计算密集型和存储密集型，可以选择部署在服务器端。

移动端（手机、智能家居、无人车等）：移动端 AI 芯片在设计思路上与服务器端 AI 芯片有着本质的区别。首先，必须保证很高的计算能效；其次，在高级辅助驾驶 ADAS 等设备对实时性要求很高的场合，推断过程必须在设备本身完成，因此要求移动端设备具备足够的推断能力。而某些场合还会有低功耗、低延迟、低成本的要求，从而导致移动端的 AI 芯片多种多样。

1.2 AI 芯片发展历程

从图灵的论文《计算机器与智能》和图灵测试，到最初级的神经元模拟单元——感知机，再到现在多达上百层的深度神经网络，人类对人工智能的探索从来就没有停止过。上世纪八十年代，多层神经网络和反向传播算法的出现给人工智能行业点燃了新的火花。反向传播的主要创新在于能将信息输出和目标输出之间的误差通过多层网络往前一级迭代反馈，将最终的输出收敛到某一个目标范围之内。1989 年贝尔实验室成功利用反向传播算法，在多层神经网络开发了一个手写邮编识别器。1998 年 Yann LeCun 和 Yoshua Bengio 发表了手写识别神经网络和反向传播优化相关的论文《Gradient-based learning applied to document recognition》，开创了卷积神经网络的时代。

此后，人工智能陷入了长时间的发展沉寂阶段，直到 1997 年 IBM 的深蓝战胜国际象棋大师和 2011 年 IBM 的沃森智能系统在 Jeopardy 节目中胜出，人工智能才又一次为人们所关注。2016 年 Alpha Go 击败韩国围棋九段职业选手，则标志着人工智能的又一波高潮。从基础算法、底层硬件、工具框架到实际应用场景，现阶段的人工智能领域已经全面开花。

作为人工智能核心的底层硬件 AI 芯片，也同样经历了多次的起伏和波折，总体看来，AI 芯片的发展前后经历了四次大的变化，其发展历程如图 2 所示。

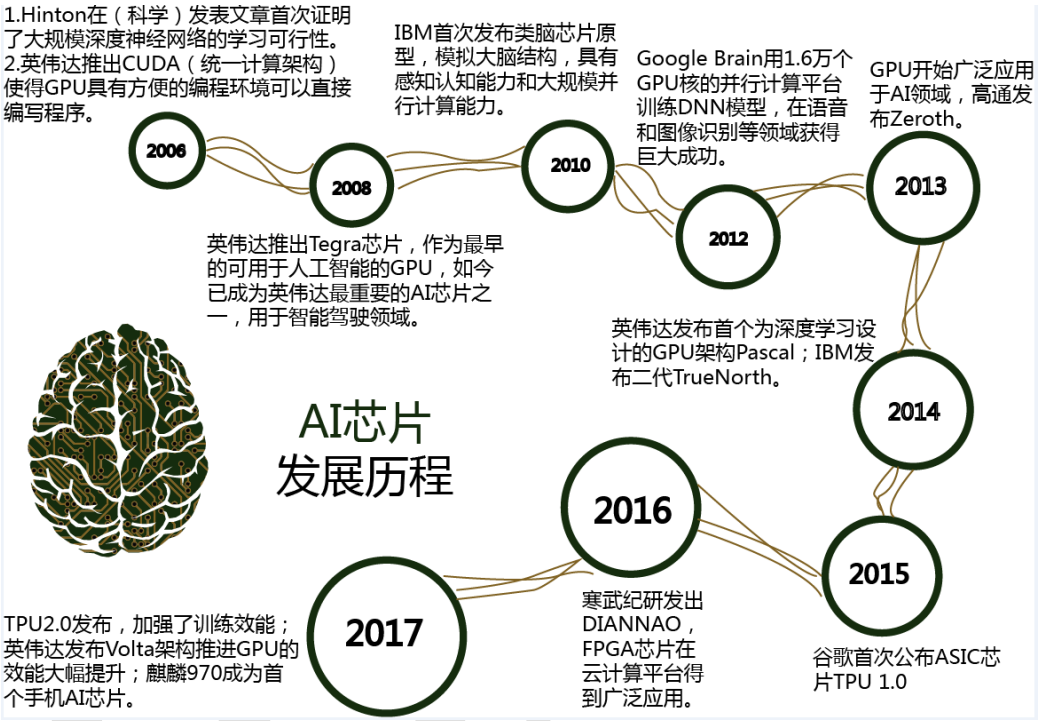


图 2 AI 芯片发展历程

（1）2007 年以前，AI 芯片产业一直没有发展成为成熟的产业；同时由于当时算法、数据量等因素，这个阶段 AI 芯片并没有特别强烈的市场需求，通用的 CPU 芯片即可满足应用需要。

（2）随着高清视频、VR、AR 游戏等行业的发展，GPU 产品取得快速的突破；同时人们发现 GPU 的并行计算特性恰好适应人工智能算法及大数据并行计算的需求，如 GPU 比之前传统的 CPU 在深度学习算法的运算上可以提高几十倍的效率，因此开始尝试使用 GPU 进行人工智能计算。

（3）进入 2010 年后，云计算广泛推广，人工智能的研究人员可以通过云计算借助大量 CPU 和 GPU 进行混合运算，进一步推进了 AI 芯片的深入应用，从而催生了各类 AI 芯片的研发与应用。

（4）人工智能对于计算能力的要求不断快速地提升，进入 2015 年后，GPU 性能功耗比不高的特点使其在工作适用场合受到多种限制，业界开始研发针对人工智能的专用芯片，以期通过更好的硬件和芯片架构，在计算效率、能耗比等性能上得到进一步提升。

1.3 我国 AI 芯片发展情况

目前，我国的人工智能芯片行业发展尚处于起步阶段。长期以来，中国在 CPU、GPU、DSP 处理器设计上一直处于追赶地位，绝大部分芯片设计企业依靠国外的 IP 核设计芯片，在自主创新上受到了极大的限制。然而，人工智能的兴起，无疑为中国在处理器领域实现

弯道超车提供了绝佳的机遇。人工智能领域的应用目前还处于面向行业应用阶段，生态上尚未形成垄断，国产处理器厂商与国外竞争对手在人工智能这一全新赛场上处在同一起跑线上，因此，基于新兴技术和应用市场，中国在建立人工智能生态圈方面将大有可为。

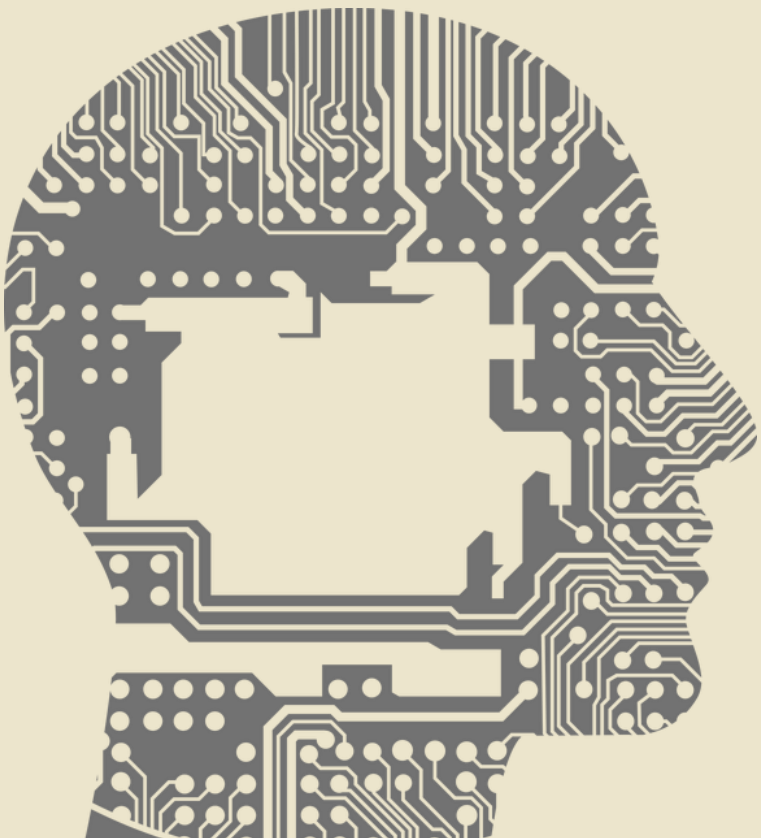
由于我国特殊的环境和市场，国内 AI 芯片的发展目前呈现出百花齐放、百家争鸣的态势，AI 芯片的应用领域也遍布股票交易、金融、商品推荐、安防、早教机器人以及无人驾驶等众多领域，催生了大量的人工智能芯片创业公司，如地平线、深鉴科技、中科寒武纪等。

尽管如此，国内公司却并未如国外大公司一样形成市场规模，反而出现各自为政的散裂发展现状。除了新兴创业公司，国内研究机构如北京大学、清华大学、中国科学院等在 AI 芯片领域都有深入研究；而其他公司如百度和比特大陆等，2017 年也有一些成果发布。

可以预见，未来谁先在人工智能领域掌握了生态系统，谁就掌握住了这个产业的主动权。

2 technology

技術篇



2 技术篇

从概念篇的介绍中我们可以发现，人工智能芯片目前有两种发展路径：一种是延续传统计算架构，加速硬件计算能力，主要以 3 种类型的芯片为代表，即 GPU、FPGA、ASIC，但CPU依旧发挥着不可替代的作用；另一种是颠覆经典的冯·诺依曼计算架构，采用类脑神经结构来提升计算能力，以 IBM TrueNorth 芯片为代表。

2.1 传统的 CPU 及其局限性

计算机工业从 1960 年代早期开始使用 CPU 这个术语。迄今为止，CPU 从形态、设计到实现都已发生了巨大的变化，但是其基本工作原理却一直没有大的改变。通常 CPU 由控制器和运算器这两个主要部件组成。传统的 CPU 内部结构图如图 3 所示，从图中我们可以看到：实质上仅单独的 ALU 模块（逻辑运算单元）是用来完成数据计算的，其他各个模块的存在都是为了保证指令能够一条接一条的有序执行。这种通用性结构对于传统的编程计算模式非常适合，同时可以通过提升 CPU 主频（提升单位时间内执行指令的条数）来提升计算速度。但对于深度学习中的并不需要太多的程序指令、却需要海量数据运算的计算需求，这种结构就显得有些力不从心。尤其是在功耗限制下，无法通过无限制的提升 CPU 和内存的工作频率来加快指令执行速度，这种情况导致 CPU 系统的发展遇到不可逾越的瓶颈。

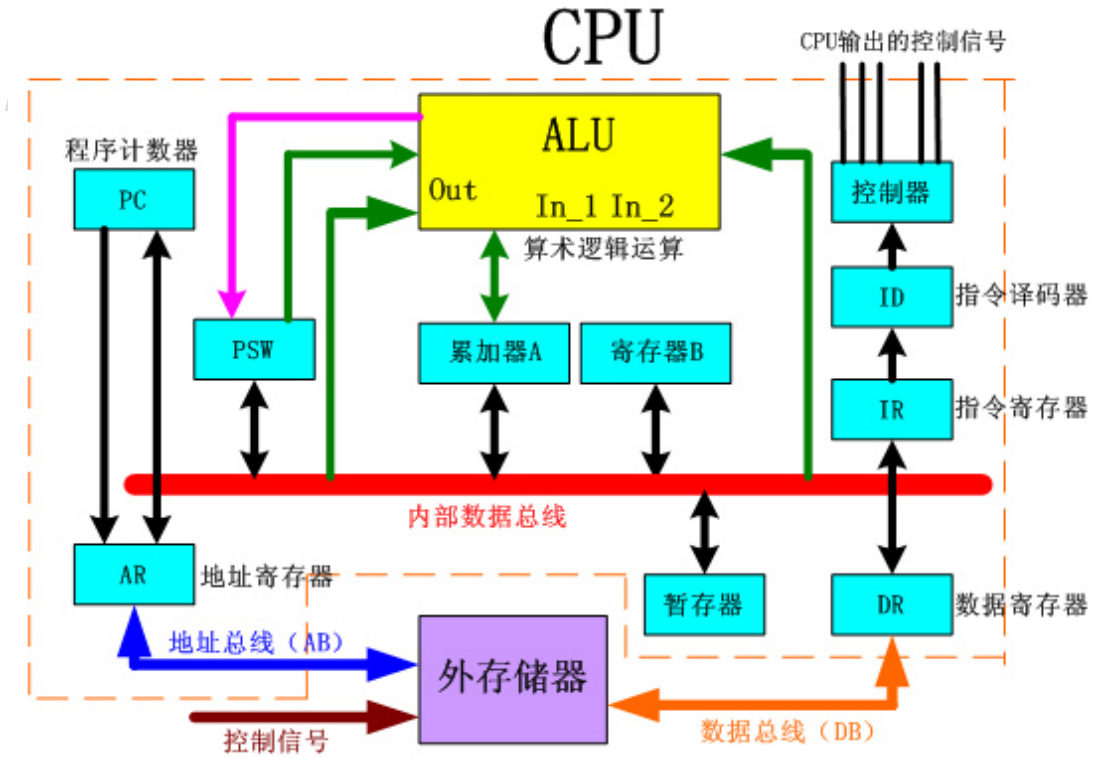


图 3 传统 CPU 内部结构图（仅 ALU 为主要计算模块）

2.2 并行加速计算的 GPU

GPU 作为最早从事并行加速计算的处理器，相比 CPU 速度快，同时比其他加速器芯片编程灵活简单。

传统的 CPU 之所以不适合人工智能算法的执行，主要原因在于其计算指令遵循串行执行的方式，没能发挥出芯片的全部潜力。与之不同的是，GPU 具有高并行结构，在处理图形数据和复杂算法方面拥有比 CPU 更高的效率。对比 GPU 和 CPU 在结构上的差异，CPU 大部分面积为控制器和寄存器，而 GPU 拥有更多的 ALU（ARITHMETIC LOGIC UNIT，逻辑运算单元）用于数据处理，这样的结构适合对密集型数据进行并行处理，CPU 与 GPU 的结构对比如图 4 所示。程序在 GPU 系统上的运行速度相较于单核 CPU 往往提升几十倍乃至上千倍。随着英伟达、AMD 等公司不断推进其对 GPU 大规模并行架构的支持，面向通用计算的 GPU（即 GPGPU，GENERAL PURPOSE GPU，通用计算图形处理器）已成为加速可并行应用程序的重要手段。

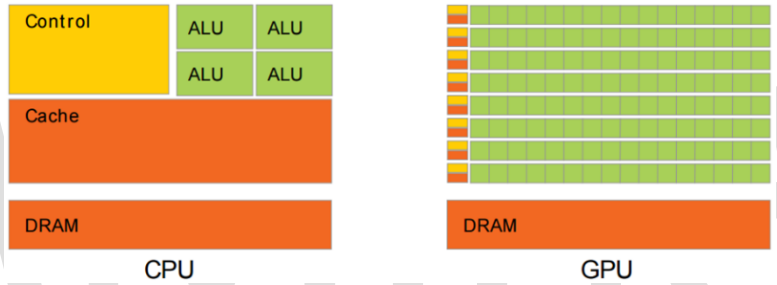


图 4 CPU 及 GPU 结构对比图（引用自 NVIDIA CUDA 文档）

GPU 的发展历程可分为 3 个阶段，发展历程示意图如图 5 所示：

第一代 GPU（1999 年以前），部分功能从 CPU 分离，实现硬件加速，以 GE（GEOMETRY ENGINE）为代表，只能起到 3D 图像处理的加速作用，不具有软件编程特性。

第二代 GPU（1999-2005 年），实现进一步的硬件加速和有限的编程性。1999 年，英伟达发布了“专为执行复杂的数学和几何计算的”GeForce256 图像处理芯片，将更多的晶体管用作执行单元，而不是像 CPU 那样用作复杂的控制单元和缓存，将 T&L（TRANSFORM AND LIGHTING）等功能从 CPU 分离出来，实现了快速变换，这成为 GPU 真正出现的标志。之后几年，GPU 技术快速发展，运算速度迅速超过 CPU。2001 年英伟达和 ATI 分别推出的 GEFORCE3 和 RADEON 8500，图形硬件的流水线被定义为流处理器，出现了顶点级可编程性，同时像素级也具有有限的编程性，但 GPU 的整体编程性仍然比较有限。

第三代 GPU（2006 年以后），GPU 实现方便的编程环境创建，可以直接编写程序。2006 年英伟达与 ATI 分别推出了 CUDA（Compute United Device Architecture，计算统一设备架构）编程环境和 CTM（CLOSE TO THE METAL）编程环境，使得 GPU 打破图形语言的局限成为

真正的并行数据处理超级加速器。

2008 年，苹果公司提出一个通用的并行计算编程平台 OPENCL（OPEN COMPUTING LANGUAGE，开放运算语言），与 CUDA 绑定在英伟达的显卡上不同，OPENCL 和具体的计算设备无关。

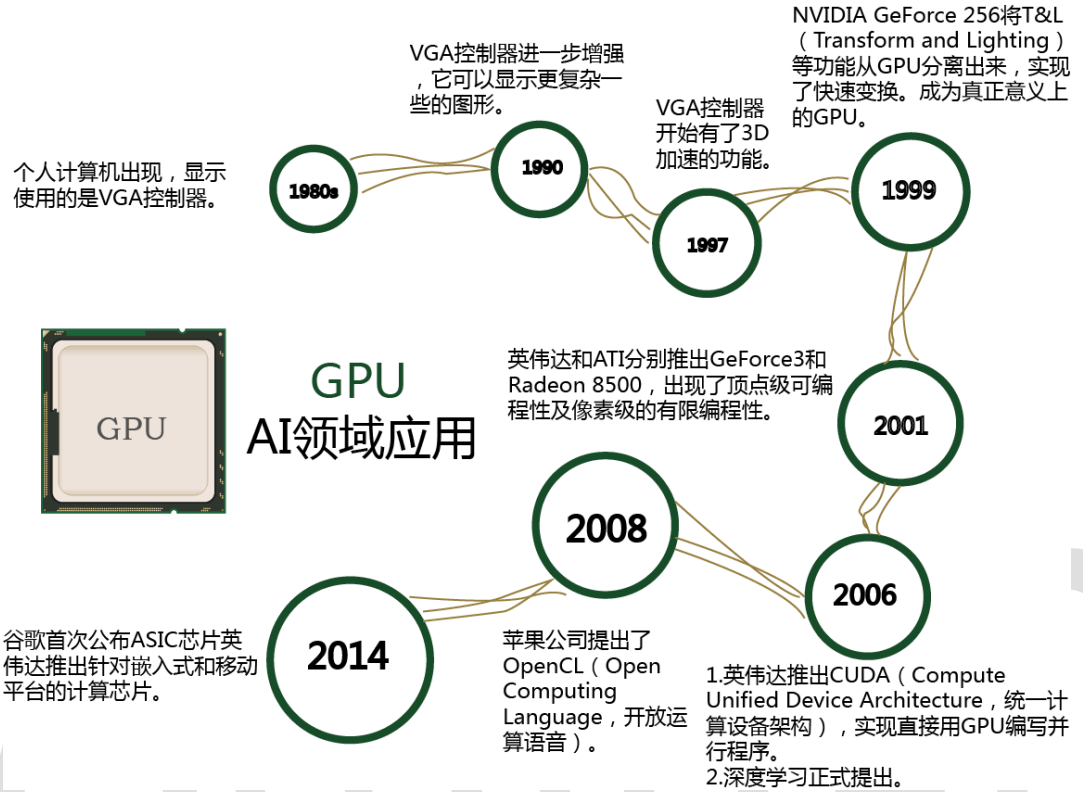


图 5 GPU 芯片的发展阶段

目前，GPU 已经发展到较为成熟的阶段。谷歌、FACEBOOK、微软、TWITTER 和百度等公司都在使用 GPU 分析图片、视频和音频文件，以改进搜索和图像标签等应用功能。此外，很多汽车生产商也在使用 GPU 芯片发展无人驾驶。不仅如此，GPU 也被应用于 VR/AR 相关的产业。

但是 GPU 也有一定的局限性。深度学习算法分为训练和推断两部分，GPU 平台在算法训练上非常高效。但在推断中对于单项输入进行处理的时候，并行计算的优势不能完全发挥出来。

2.3 半定制化的 FPGA

FPGA 是在 PAL、GAL、CPLD 等可编程器件基础上进一步发展的产物。用户可以通过烧入 FPGA 配置文件来定义这些门电路以及存储器之间的连线。这种烧入不是一次性的，比如用户可以把 FPGA 配置成一个微控制器 MCU，使用完毕后可以编辑配置文件把同一个 FPGA 配置成一个音频编解码器。因此，它既解决了定制电路灵活性的不足，又克服了原

有可编程器件门电路数有限的缺点。

FPGA 可同时进行数据并行和任务并行计算，在处理特定应用时有更加明显的效率提升。对于某个特定运算，通用 CPU 可能需要多个时钟周期；而 FPGA 可以通过编程重组电路，直接生成专用电路，仅消耗少量甚至一次时钟周期就可完成运算。

此外，由于 FPGA 的灵活性，很多使用通用处理器或 ASIC 难以实现的底层硬件控制操作技术，利用 FPGA 可以很方便的实现。这个特性为算法的功能实现和优化留出了更大空间。同时 FPGA 一次性成本（光刻掩模制作成本）远低于 ASIC，在芯片需求还未成规模、深度学习算法暂未稳定，需要不断迭代改进的情况下，利用 FPGA 芯片具备可重构的特性来实现半定制的人工智能芯片是最佳选择之一。

功耗方面，从体系结构而言，FPGA 也具有天生的优势。传统的冯氏结构中，执行单元（如 CPU 核）执行任意指令，都需要有指令存储器、译码器、各种指令的运算器及分支跳转处理逻辑参与运行，而 FPGA 每个逻辑单元的功能在重编程（即烧入）时就已经确定，不需要指令，无需共享内存，从而可以极大的降低单位执行的功耗，提高整体的能耗比。

由于 FPGA 具备灵活快速的特点，因此在众多领域都有替代 ASIC 的趋势。FPGA 在人工智能领域的应用如图 6 所示。

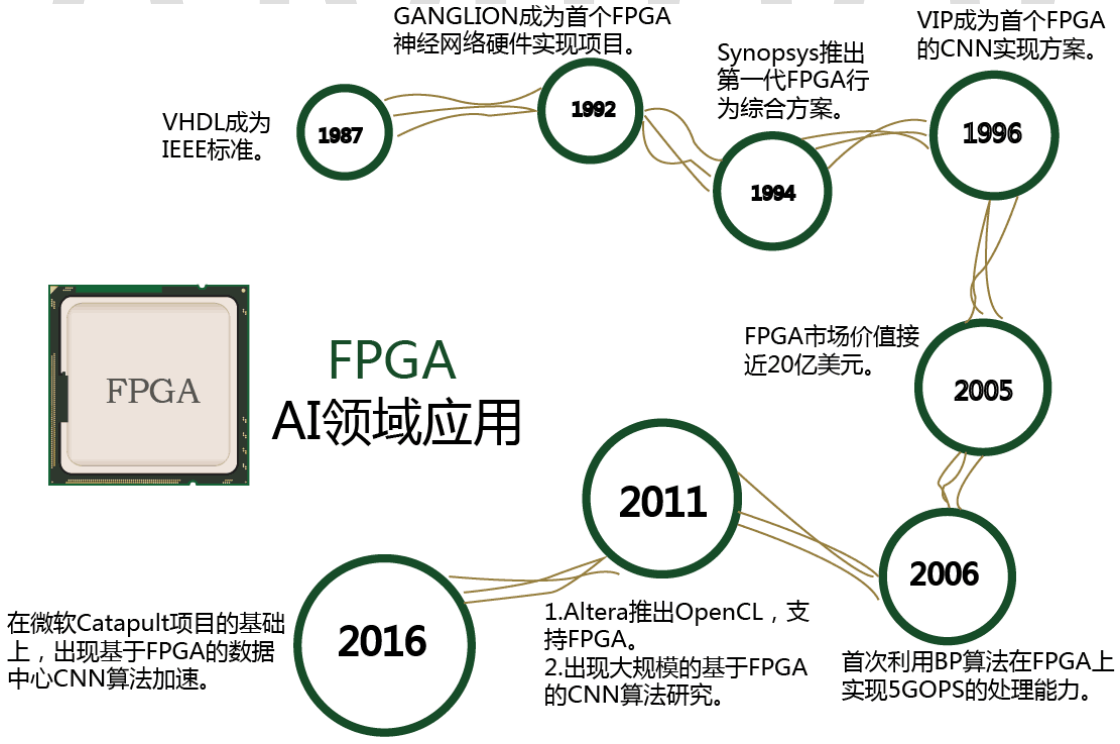


图 6 FPGA 在人工智能领域的应用

2.4 全定制化的 ASIC

目前以深度学习为代表的人工智能计算需求，主要采用 GPU、FPGA 等已有的适合并行计算的通用芯片来实现加速。在产业应用没有大规模兴起之时，使用这类已有的通用芯片可以避免专门研发定制芯片（ASIC）的高投入和高风险。但是，由于这类通用芯片设计初衷并非专门针对深度学习，因而天然存在性能、功耗等方面的局限性。随着人工智能应用规模的扩大，这类问题日益突显。

GPU 作为图像处理器，设计初衷是为了应对图像处理中的大规模并行计算。因此，在应用于深度学习算法时，有三个方面的局限性：第一，应用过程中无法充分发挥并行计算优势。深度学习包含训练和推断两个计算环节，GPU 在深度学习算法训练上非常高效，但对于单一输入进行推断的场合，并行度的优势不能完全发挥。第二，无法灵活配置硬件结构。GPU 采用 SIMT 计算模式，硬件结构相对固定。目前深度学习算法还未完全稳定，若深度学习算法发生大的变化，GPU 无法像 FPGA 一样可以灵活的配制硬件结构。第三，运行深度学习算法能效低于 FPGA。

尽管 FPGA 倍受看好，甚至新一代百度大脑也是基于 FPGA 平台研发，但其毕竟不是专门为了适用深度学习算法而研发，实际应用中也存在诸多局限：第一，基本单元的计算能力有限。为了实现可重构特性，FPGA 内部有大量极细粒度的基本单元，但是每个单元的计算能力(主要依靠 LUT 查找表)都远远低于 CPU 和 GPU 中的 ALU 模块；第二、计算资源占比相对较低。为实现可重构特性，FPGA 内部大量资源被用于可配置的片上路由与连线,；第三，速度和功耗相对专用定制芯片(ASIC)仍然存在不小差距；第四，FPGA 价格较为昂贵，在规模放量的情况下单块 FPGA 的成本要远高于专用定制芯片。

因此，随着人工智能算法和应用技术的日益发展，以及人工智能专用芯片 ASIC 产业环境的逐渐成熟，全定制化人工智能 ASIC 也逐步体现出自身的优势，从事此类芯片研发与应用的国内外比较有代表性的公司如表 1 所示，后续产业篇会做相应的详细介绍。

表 1 人工智能专用芯片（包括类脑芯片）研发情况一览

国家	名称	简介
国外	英伟达 Tesla P100	首个专为深度学习加速计算而设计的图形处理芯片架构
	谷歌 TPU	面向机器学习张量处理的加速芯片
	IBM TrueNorth 芯片	TrueNorth 以分布式、并行的方式来存储处理信息，支持 SNN
	高通 Zeroth 芯片	按照人类神经网络传输信息的方式而设计，支持 SNN
	英特尔神经形态芯片	支持片上学习的 SNN 芯片
	Audience 神经形态芯片	可以模拟人耳抑制噪音，应用于智能手机
国内	中星微	中国首个嵌入式神经网络芯片 NPU
	寒武纪	全球首个提出深度学习处理器芯片指令集
	地平线机器人	专注于人工智能本地化机器学习芯片
	深鉴科技	利用 FPGA 平台打造人工智能芯片 DPU
	灵汐科技	类脑处理芯片，支持 DNN/SNN 混合模式

ASIC 芯片非常适合人工智能的应用场景。首先，ASIC 的性能提升非常明显。例如英伟达首款专门为深度学习从零开始设计的芯片 Tesla P100 数据处理速度是其 2014 年推出 GPU 系列的 12 倍。谷歌为机器学习定制的芯片 TPU 将硬件性能提升至相当于当前芯片按摩尔定律发展 7 年后的水平。正如 CPU 改变了当年庞大的计算机一样，人工智能 ASIC 芯片也将大幅改变如今 AI 硬件设备的面貌。如大名鼎鼎的 AlphaGo 使用了约 170 个图形处理器（GPU）和 1200 个中央处理器（CPU），这些设备需要占用一个机房，还要配备大功率的空调，以及多名专家进行系统维护。而如果全部使用专用芯片，极大可能只需要一个普通收纳盒大小的空间，，且功耗也会大幅降低。

第二，下游需求促进人工智能芯片专用化。从服务器，计算机到无人驾驶汽车、无人机再到智能家居的各类家电，至少数十倍于智能手机体量的设备需要引入感知交互能力和人工智能计算能力。而出于对实时性的要求以及训练数据隐私等考虑，这些应用不可能完全依赖云端，必须要有本地的软硬件基础平台支撑，这将带来海量的人工智能芯片需求。

目前人工智能专用芯片的发展方向包括：主要基于 FPGA 的半定制、针对深度学习算法的全定制和类脑计算芯片 3 个方向。

在芯片需求还未形成规模、深度学习算法暂未稳定，AI 芯片本身需要不断迭代改进的情况下，利用具备可重构特性的 FPGA 芯片来实现半定制的人工智能芯片是最佳选择之一。这类芯片中的杰出代表是国内初创公司深鉴科技，该公司设计了“深度学习处理单元”（Deep Processing Unit，DPU）的芯片，希望以 ASIC 级别的功耗达到优于 GPU 的性能，其第一批产品就是基于 FPGA 平台开发研制出来的。这种半定制芯片虽然依托于 FPGA 平台，但是抽象出了指令集与编译器，可以快速开发、快速迭代，与专用的 FPGA 加速器产品相比，也具有非常明显的优势。

深度学习算法稳定后，AI 芯片可采用 ASIC 设计方法进行全定制，使性能、功耗和面积等指标面向深度学习算法做到最优。

2.5 类脑芯片

类脑芯片不采用经典的冯·诺依曼架构，而是基于神经形态架构设计，以 IBM Truenorth 为代表。IBM 研究人员将存储单元作为突触、计算单元作为神经元、传输单元作为轴突搭建了神经芯片的原型。目前，Truenorth 用三星 28nm 功耗工艺技术，由 54 亿个晶体管组成的芯片构成的片上网络有 4096 个神经突触核心，实时作业功耗仅为 70mW。由于神经突触要求权重可变且要有记忆功能，IBM 采用与 CMOS 工艺兼容的相变非挥发存储器（PCM）的技术实验性的实现了新型突触，加快了商业化进程。

在国内，清华大学类脑计算中心于 2015 年 11 月成功的研制了国内首款超大规模的神经形态类脑计算天机芯片。该芯片同时支持脉冲神经网络和人工神经网络（深度神经网络），

可进行大规模神经元网络的模拟。中心还开发了面向类脑芯片的工具链，降低应用的开发难度并提升效率。第二代 28nm 天机芯片也已问世，在性能功耗比上要优于 Truenorth。

当前，类脑 AI 芯片的设计目的不再仅仅局限于加速深度学习算法，而是在芯片基本结构甚至器件层面上改变设计，希望能够开发出新的类脑计算机体系结构，比如采用忆阻器和 ReRAM 等新器件来提高存储密度。这类芯片技术尚未完全成熟，离大规模应用还有一定的差距，但是长期来看类脑芯片有可能会带来计算机体系结构的革命。

2.6 AI 芯片技术特点比较

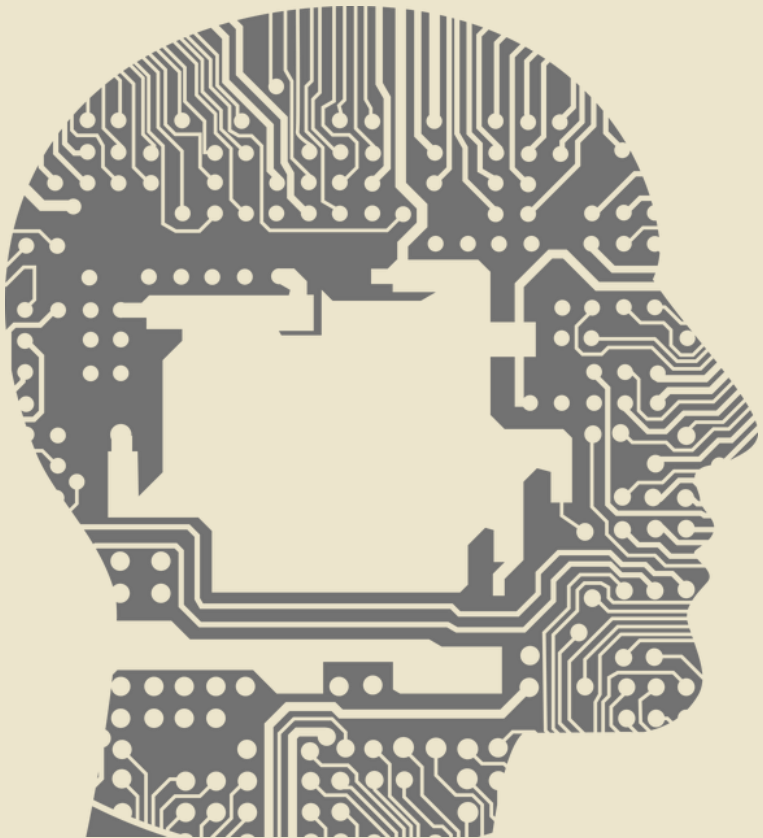
通过以上分析，我们可以总结出以下几个特点。

- CPU 通用性最强，但延迟严重，散热高，效率最低。
- GPU 通用性强、速度快、效率高，特别适合用在深度学习训练方面，但是性能功耗比较低。
- FPGA 具有低能耗、高性能以及可编程等特性，相对于 CPU 与 GPU 有明显的性能或者能耗优势，但对使用者要求高。
- ASIC 可以更有针对性地进行硬件层次的优化，从而获得更好的性能、功耗比。但是 ASIC 芯片的设计和制造需要大量的资金、较长的研发周期和工程周期，而且深度学习算法仍在快速发展，若深度学习算法发生大的变化，FPGA 能很快改变架构，适应最新的变化，ASIC 类芯片一旦定制则难于进行修改。

当前阶段，GPU 配合 CPU 仍然是 AI 芯片的主流，而后随着视觉、语音、深度学习的算法在 FPGA 以及 ASIC 芯片上的不断优化，此两者也将逐步占有更多的市场份额，从而与 GPU 达成长期共存的局面。从长远看，人工智能类脑神经芯片是发展的路径和方向。

3 industry

产业篇



3 产业篇

本篇将介绍目前人工智能芯片技术领域的国内外代表性企业。文中排名不分先后。

人工智能芯片技术领域的国内代表性企业包括中科寒武纪、中星微、地平线机器人、深鉴科技、灵汐科技、启英泰伦、百度、华为等，国外包括英伟达、AMD、Google、高通、Nervana Systems、Movidius、IBM、ARM、CEVA、MIT/Eyeriss、苹果、三星等。

- **中科寒武纪**

寒武纪科技成立于 2016 年，总部在北京，创始人是中科院计算所的陈天石、陈云霁兄弟，公司致力于打造各类智能云服务器、智能终端以及智能机器人的核心处理器芯片。阿里巴巴创投、联想创投、国科投资、中科图灵、元禾原点、涌铧投资联合投资，为全球 AI 芯片领域第一个独角兽初创公司。

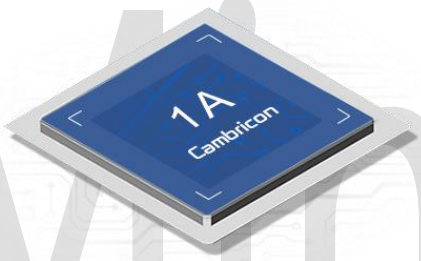


图 7 Cambricon-1A（引用自官网）

寒武纪是全球第一个成功流片并拥有成熟产品的 AI 芯片公司，拥有终端 AI 处理器 IP 和云端高性能 AI 芯片两条产品线。2016 年发布的寒武纪 1A 处理器（Cambricon-1A）是世界首款商用深度学习专用处理器，面向智能手机、安防监控、无人机、可穿戴设备以及智能驾驶等各类终端设备，在运行主流智能算法时性能功耗比全面超越传统处理器。图 7 为寒武纪的 Cambricon-1A 的 AI 芯片。

- **中星微**

1999 年，由多位来自硅谷的博士企业家在北京中关村科技园区创建了中星微电子有限公司，启动并承担了国家战略项目——“星光中国芯工程”，致力于数字多媒体芯片的开发、设计和产业化。

2016 年初，中星微推出了全球首款集成了神经网络处理器（NPU）的 SVAC 视频编解码 SoC，使得智能分析结果可以与视频数据同时编码，形成结构化的视频码流。该技术被广泛应用于视频监控摄像头，开启了安防监控智能化的新时代。自主设计的嵌入式神经网络处理器（NPU）采用了“数据驱动并行计算”架构，专门针对深度学习算法进行了优化，具备高性能、低功耗、高集成度、小尺寸等特点，特别适合物联网前端智能的需求。

图 8 显示了集成了 NPU 的神经网络处理器 VC0616 的内部结构。

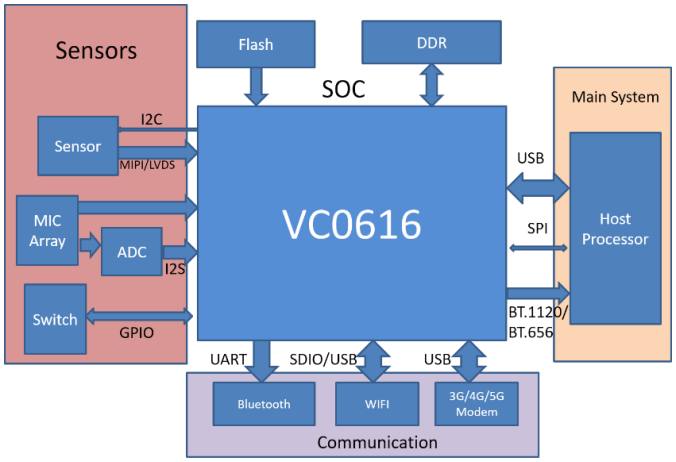


图 8 集成了 NPU 的神经网络处理器（引用自官网）

- **地平线机器人（Horizon Robotics）**

地平线机器人成立于 2015 年，总部在北京，创始人是前百度深度学习研究院负责人余凯。

BPU（BrainProcessing Unit）是地平线机器人自主设计研发的高效人工智能处理器架构 IP，支持 ARM/GPU/FPGA/ASIC 实现，专注于自动驾驶、人脸图像辨识等专用领域。2017 年，地平线发布基于高斯架构的嵌入式人工智能解决方案，将在智能驾驶、智能生活、公共安防三个领域进行应用，第一代 BPU 芯片“盘古”目前已进入流片阶段，预计在 2018 年下半年推出，能支持 1080P 的高清图像输入，每秒钟处理 30 帧，检测跟踪数百个目标。地平线的第一代 BPU 采用 TSMC 的 40nm 工艺，相对于传统 CPU/GPU，能效可以提升 2~3 个数量级（100~1, 000 倍左右）。图 9 为地平线公司公布的 BPU 发展战略图。

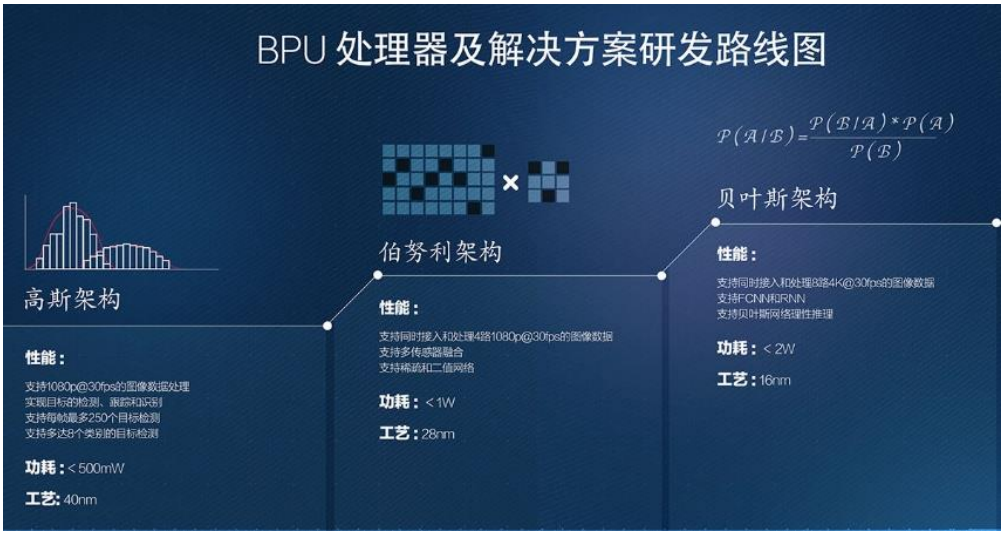


图 9 地平线公布的 BPU 发展战略图（引用自官网）

- **深鉴科技**

深鉴科技成立于 2016 年，总部在北京。由清华大学与斯坦福大学的世界顶尖深度学习硬件研究者创立。深鉴科技于 2018 年 7 月被赛灵思收购。

深鉴科技将其开发的基于 FPGA 的神经网络处理器称为 DPU。到目前为止，深鉴公开发布了两款 DPU：亚里士多德架构和笛卡尔架构，其中，亚里士多德架构是针对卷积神经网络 CNN 而设计；笛卡尔架构专为处理 DNN/RNN 网络而设计，可对经过结构压缩后的稀疏神经网络进行极致高效的硬件加速。相对于 Intel XeonCPU 与 Nvidia TitanX GPU，应用笛卡尔架构的处理器在计算速度上分别提高 189 倍与 13 倍，具有 24,000 倍与 3,000 倍的更高能效。图 10 为深鉴科技的亚里士多德处理器架构图。

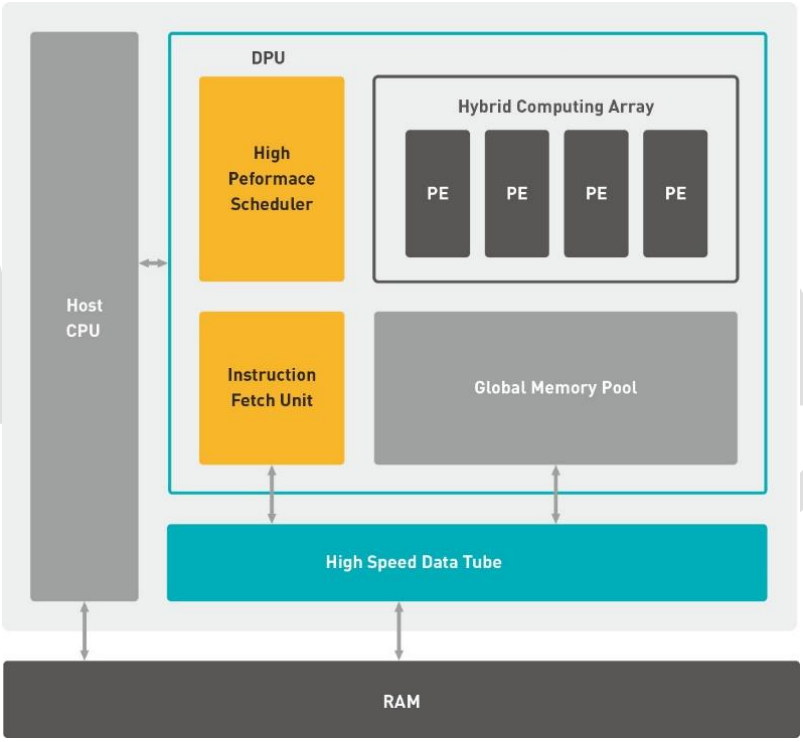


图 10 亚里士多德架构（引用自官网）

- **灵汐科技**

灵汐科技于 2018 年 1 月在北京成立，联合创始人包括清华大学的世界顶尖类脑计算研究者。公司致力于新一代神经网络处理器（Tianjic）开发，特点在于既能够高效支撑现有流行的机器学习算法（包括 CNN，MLP，LSTM 等网络架构），也能够支撑更仿脑的、更具成长潜力的脉冲神经网络算法；使芯片具有高计算力、高多任务并行度和较低功耗等优点。软件工具链方面支持由 Caffe、TensorFlow 等算法平台直接进行神经网络的映射编译，开发友善的用户交互界面。Tianjic 可用于云端计算和终端应用场景，助力人工智能的落地和推广。

- **启英泰伦**

启英泰伦于2015年11月在成都成立，是一家语音识别芯片研发商。启英泰伦的CI1006是基于 ASIC 架构的人工智能语音识别芯片，如图 11 所示，包含了脑神经网络处理硬件单元，能够完美支持 DNN 运算架构，进行高性能的数据并行计算，可极大的提高人工智能深度学习语音技术对大量数据的处理效率。

图 11 CI1006 芯片（引用自官网）

- **百度**

百度 2017 年 8 月 Hot Chips 大会上发布了 XPU，这是一款 256 核、基于 FPGA 的云计算加速芯片。合作伙伴是赛思灵（Xilinx）。XPU 采用新一代 AI 处理架构，拥有 GPU 的通用性和 FPGA 的高效率和低能耗，对百度的深度学习平台 PaddlePaddle 做了高度的优化和加速。据介绍，XPU 关注计算密集型、基于规则的多样化计算任务，希望提高效率和性能，并带来类似 CPU 的灵活性。但目前 XPU 有所欠缺的仍是可编程能力，而这也是涉及 FPGA 时普遍存在的问题。到目前为止，XPU 尚未提供编译器。

- **华为**

麒麟 970 搭载的神经网络处理器 NPU 采用了寒武纪 IP，如图 12 所示。麒麟 970 采用了 TSMC 10nm 工艺制程，拥有 55 亿个晶体管，功耗相比上一代芯片降低 20%。CPU 架构方面为 4 核 A73+4 核 A53 组成 8 核心，能耗同比上一代芯片得到 20%的提升；GPU 方面采用了 12 核 Mali G72 MP12GPU，在图形处理以及能效两项关键指标方面分别提升 20%和 50%；NPU 采用 HiAI 移动计算架构，在 FP16 下提供的运算性能可以达到 1.92 TFLOPs，相比四个 Cortex-A73 核心，处理同样的 AI 任务，有大约具备 50 倍能效和 25 倍性能优势。

图 12 华为麒麟 970 神经网络处理器 NPU

- **英伟达（Nvidia）**

英伟达创立于 1993 年，总部位于美国加利福尼亚州圣克拉拉市。早在 1999 年，英伟达发明了 GPU，重新定义了现代计算机图形技术，彻底改变了并行计算。

深度学习对计算速度有非常苛刻的要求，而英伟达的 GPU 芯片可以让大量处理器并行运算，速度比 CPU 快十倍甚至几十倍，因而成为绝大部分人工智能研究者和开发者的首选。自从 Google Brain 采用 1.6 万个 GPU 核训练 DNN 模型，并在语音和图像识别等领域获得巨大成功以来，英伟达已成为 AI 芯片市场中无可争议的领导者。

- **AMD**

美国 AMD 半导体公司专门为计算机、通信和消费电子行业设计和制造各种创新的微处理器（CPU、GPU、APU、主板芯片组、电视卡芯片等），以及提供闪存和低功率处理器解决方案，公司成立于 1969 年。AMD 致力为技术用户——从企业、政府机构到个人消费者——提供基于标准的、以客户为中心的解决方案。

2017 年 12 月 Intel 和 AMD 宣布将联手推出一款结合英特尔处理器和 AMD 图形单元的笔记本电脑芯片。目前 AMD 拥有针对 AI 和机器学习的高性能 Radeon Instinc 加速卡，开放式软件平台 ROCm 等。

- **Google**

Google 在 2016 年宣布独立开发一种名为 TPU 的全新的处理系统。TPU 是专门为机器学习应用而设计的专用芯片。通过降低芯片的计算精度，减少实现每个计算操作所需晶体管数量的方式，让芯片的每秒运行的操作个数更高，这样经过精细调优的机器学习模型就能在芯片上运行得更快，进而更快地让用户得到更智能的结果。在 2016 年 3 月打败了李世石和 2017 年 5 月打败了柯杰的阿尔法狗，就是采用了谷歌的 TPU 系列芯片。

Google I/O-2018 开发者大会期间，正式发布了第三代人工智能学习专用处理器 TPU 3.0。TPU3.0 采用 8 位低精度计算以节省晶体管数量，对精度影响很小但可以大幅节约功耗、加快速度，同时还有脉动阵列设计，优化矩阵乘法与卷积运算，并使用更大的片上内存，减少对系统内存的依赖。速度能加快到最高 100PFlops（每秒 1000 万亿次浮点计算）。

- **高通**

在智能手机芯片市场占据绝对优势的高通公司，也在人工智能芯片方面积极布局。据高通提供的资料显示，其在人工智能方面已投资了 Clarifai 公司和中国“专注于物联网人工智能服务”的云知声。

而早在 2015 年 CES 上，高通就已推出了一款搭载骁龙 SoC 的飞行机器人——Snapdragon Cargo。高通认为在工业、农业的监测以及航拍对拍照、摄像以及视频新需求上，

公司恰好可以发挥其在计算机视觉领域的能力。此外，高通的骁龙 820 芯片也被应用于 VR 头盔中。事实上，高通已经在研发可以在本地完成深度学习的移动端设备芯片。

- **Nervana Systems**

Nervana 创立于 2014 年，公司推出的 The Nervana Engine 是一个为深度学习专门定制和优化的 ASIC 芯片。这个方案的实现得益于一项叫做 High Bandwidth Memory 的新型内存技术，这项技术同时拥有高容量和高速度，提供 32GB 的片上储存和 8TB 每秒的内存访问速度。该公司目前提供一个人工智能服务“in the cloud”，他们声称这是世界上最快的且目前已被金融服务机构、医疗保健提供者和政府机构所使用的服务。他们的新型芯片将会保证 Nervana 云平台在未来的几年内仍保持最快的速度。

- **Movidius（被 Intel 收购）**

2016 年 9 月，Intel 发表声明收购了 Movidius。Movidius 专注于研发高性能视觉处理芯片。其最新一代的 Myriad2 视觉处理器主要由 SPARC 处理器作为主控制器，加上专门的 DSP 处理器和硬件加速电路来处理专门的视觉和图像信号。这是一款以 DSP 架构为基础的视觉处理器，在视觉相关的应用领域有极高的能耗比，可以将视觉计算普及到几乎所有的嵌入式系统中。

该芯片已被大量应用在 Google 3D 项目的 Tango 手机、大疆无人机、FLIR 智能红外摄像机、海康深眸系列摄像机、华睿智能工业相机等产品中。

- **IBM**

IBM 很早以前就发布过 watson，投入了很多的实际应用。除此之外，还启动了类脑芯片的研发，即 TrueNorth。

TrueNorth 是 IBM 参与 DARPA 的研究项目 SyNapse 的最新成果。SyNapse 全称是 Systems of Neuromorphic Adaptive Plastic Scalable Electronics（自适应可塑可伸缩电子神经系统，而 SyNapse 正好是突触的意思），其终极目标是开发出打破冯·诺依曼体系结构的计算机体系结构。

- **ARM**

ARM 推出全新芯片架构 DynamIQ，通过这项技术，AI 芯片的性能有望在未来三到五年内提升 50 倍。

ARM 的新 CPU 架构将会通过为不同部分配置软件的方式将多个处理核心集聚在一起，这其中包括一个专门为 AI 算法设计的处理器。芯片厂商将可以为新处理器配置最多 8 个核心。同时为了能让主流 AI 在自己的处理器上更好地运行，ARM 还将推出一系列软件库。

- **CEVA**

CEVA 是专注于 DSP 的 IP 供应商，拥有众多的产品线。其中，图像和计算机视觉 DSP 产品 CEVA-XM4 是第一个支持深度学习的可编程 DSP，而其发布的新一代型号 CEVA-XM6，具有更优的性能、更强大的计算能力以及更低的能耗。

CEVA 指出，智能手机、汽车、安全和商业应用，如无人机、自动化将是其业务开展的主要目标。

- **MIT/Eyeriss**

Eyeriss 事实上是 MIT 的一个项目，还不是一个公司，从长远来看，如果进展顺利，很可能孵化出一个新的公司。

Eyeriss 是一个高效能的深度卷积神经网络（CNN）加速器硬件，该芯片内建 168 个核心，专门用来部署神经网路（neural network），效能为一般 GPU 的 10 倍。其技术关键在于最小化 GPU 核心和记忆体之间交换数据的频率（此运作过程通常会消耗大量的时间与能量）：一般 GPU 内的核心通常共享单一记忆体，但 Eyeriss 的每个核心拥有属于自己的记忆体。

目前，Eyeriss 主要定位在人脸识别和语音识别，可应用在智能手机、穿戴式设备、机器人、自动驾驶车与其他物联网应用装置上。

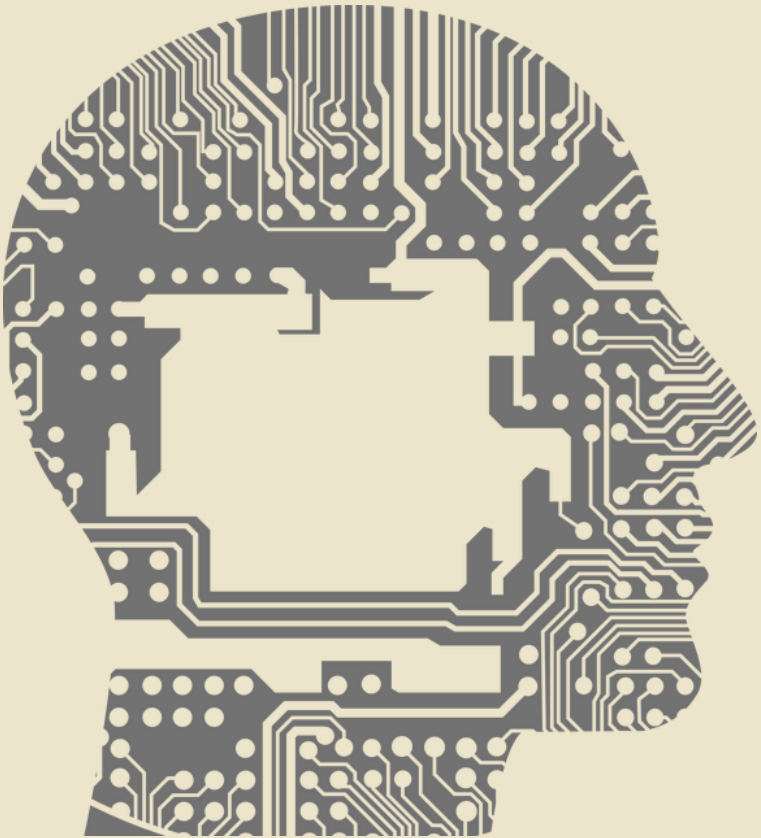
- **苹果**

在 iPhone 8 和 iPhone X 的发布会上，苹果明确表示其中所使用的 A11 处理器集成了一个专用于机器学习的硬件——“神经网络引擎（Neural Engine）”，每秒运算次数最高可达 6000 亿次。这块芯片将能够改进苹果设备在处理需要人工智能的任务时的表现，比如面部识别和语音识别等。

- **三星**

2017 年，华为海思推出了麒麟 970 芯片，据知情人士透露，为了对标华为，三星已经研发了许多种类的人工智能芯片。三星计划在未来三年内新上市的智能手机中都采用人工智能芯片，并且他们还将为人工智能设备建立新的组件业务。三星还投资了 Graphcore、深鉴科技等人工智能芯片企业。

4 talent
人物篇



4 人物篇

本报告 4.1 节通过 AMiner 大数据平台对 AMiner 的人工智能芯片人才库进行数据挖掘，统计分析出领域内学者分布及迁徙。4.2 节介绍了目前人工智能芯片领域的国内外代表性研究学者，文中排名不分先后。

4.1 学者分布及迁徙

通过统计分析 AMiner 的人工智能芯片人才库，我们得到了全球人工智能芯片领域学者分布图，如图 13 所示。从图中可以看到，人工智能芯片领域的学者主要分布在北美洲，其次是欧洲。中国对人工智能芯片的研究紧跟其后，南美洲、非洲和大洋洲人才相对比较匮乏。

图 13 人工智能芯片领域研究学者全球分布

按国家进行统计来看美国是人工智能芯片领域科技发展的核心。英国的人数紧排在美国之后。其他的专家主要分布在中国、德国、加拿大、意大利和日本，如图 14 所示。

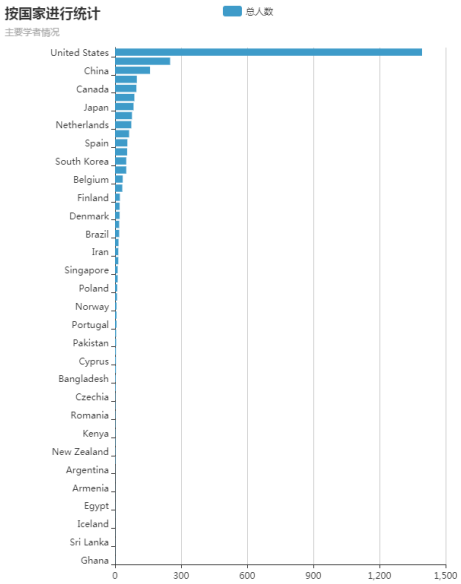


图 14 人工智能芯片领域研究学者全球分布

AMiner 对全球人工智能芯片领域最具影响力的 1000 人的迁徙路径进行了统计分析，得出如图 15 所示的各国人才逆顺差对比图。

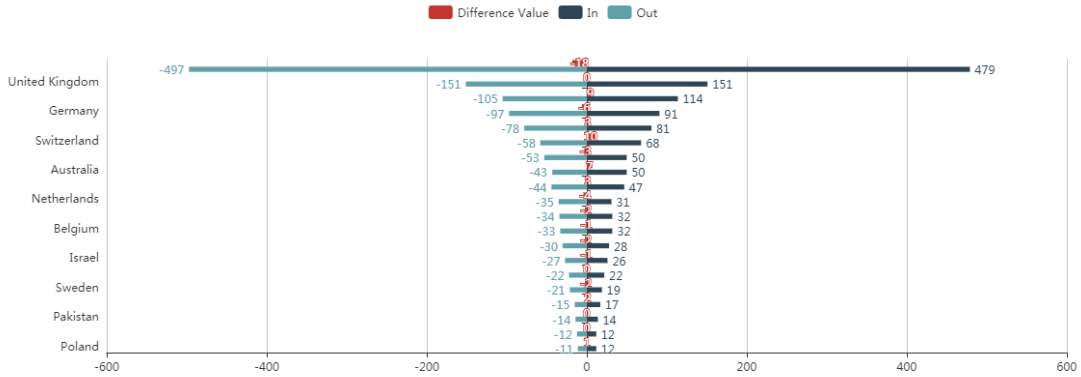


图 15 各国人才逆顺差

由图中可以看出，各国人才的流失和引进是相对比较均衡的，其中美国为人才流动大国，人才输入和输出幅度都大幅度领先。英国、中国、德国和瑞士等国次于美国，但各国之间人才流动相差并不明显。

4.2 代表性研究学者

● Jeff Dean

美国工程院院士，Jeff Dean 是谷歌大脑（Google Brain）、谷歌机器学习开源框架 TensorFlow、谷歌广告系统、谷歌搜索系统等技术的重要创始人之一。

Jeff Dean 在获得华盛顿大学计算机科学博士学位的三年之后（1999 年）加入了谷歌公司，成为了该公司最早的员工之一。在谷歌的成长过程中，他一直是该公司的头面人物——设计和实现了支撑谷歌大部分产品的许多分布式计算基础设施。

■ 主要成就

创建谷歌的广告系统 Adsense——作为谷歌搜索广告，它是如今所有互联网广告的原型。

开发谷歌的检索、索引和搜索系统，利用著名的 Pagerank 搜索算法，一举成为最优秀的搜索引擎公司。

2011 年初，Jeff Dean 与吴恩达主导创建了“谷歌大脑”（Google Brain）这一奠定了谷歌人工智能领先地位的重要部门。

领导开发了谷歌机器学习的标志性软件 TensorFlow、支持谷歌运行的超大规模计算框架 MapReduce 等重要项目。2015 年 11 月，TensorFlow 正式开源发布，目前已经是深度学习领域占据绝对统治地位的框架。

● 黄仁勋

美籍华人，1993 年创办 NVIDIA（全球最大显卡芯片厂商之一）。

黄仁勋于 1984 年在俄勒冈州大学取得电机工程学位，其后在斯坦福大学取得硕士学位。1993 年，创立 NVIDIA。

■ 主要成就

1999 年，英伟达推出了全球第一个图形处理器（GPU）；此后，GPU 成为计算机中独立于 CPU（中央处理器）的另一个重要的计算单元。

2016 年 4 月 5 日，NVIDIA 英伟达宣布推出新的 GPU 芯片 TeslaP100，芯片内置了 150 亿个晶体管，可以用于深度学习，黄仁勋宣称 TeslaP100 是目前为止最大的处理器。

2017 年 5 月 11 日，在 GTC2017 大会上，NVIDIA 发布了 Tesla V100。Tesla V100 采用台积电 12nm 工艺制程，增加了与深度学习高度相关的 Tensor 单元，在 815 平方毫米面积的硅片上集成了 210 亿个晶体管，5210 个 CUDA 核心，其单精度浮点运算性能达到 15 TFLOP/s，双精度浮点运算性能达到 7.5 TFLOP/s。

● Vivienne Sze

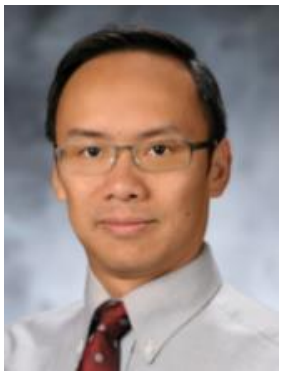
麻省理工学院电子工程和计算机科学系的副教授。研究兴趣包括便携式多媒体应用的节能算法和架构。

Vivienne Sze 于 2010 年，麻省理工学院 EECS 博士。2017 年 7 月至今，在麻省理工学院担任副教授，领导节能多媒体系统团队。

■ 主要成就

Eyeriss 节能加速器的主要研发人员之一。Eyeriss 可重新配置以支持最先进的深层卷积神经网络（CNN）。它专注于最小化加速器和主存储器之间以及加速器的计算结构内的数据传送，与当前的移动 GPU 相比，能实现 10 倍的能效。

● 谢源



加州大学圣芭芭拉分校教授。发表了近 300 篇研究论文，获得多个国际会议的最佳论文奖，以及 NSF CAREER award，中国国家自然科学基金会海外及港澳学者合作研究基金等。2014 年获得 IEEE Fellow 的荣誉。

谢源于 2002 年获得普林斯顿大学电机工程系博士学位。2003 年加入宾夕法尼亚州立大学计算机系，2008 年获得终身教职，2012 年提升正教授。

2012 年到 2013 年期间加入 AMD，负责组建和领导 AMD 北京研发中心的研究部门。

2014 年加入加州大学圣芭芭拉分校电机与计算机工程系（ECE）担任正教授。

■ 主要研究领域

谢源的主要研究领域包括 VLSI 设计，电子设计自动化，计算机架构和嵌入式系统设计。目前的研究项目包括新型内存架构，互连架构和异构系统架构。最近的研究项目侧重于技术驱动和应用驱动的设计/架构创新。技术驱动的研究项目包括新兴存储器技术和 3D 集成电路的 EDA/架构，硬件安全性和 CPU/GPU/FPGA 的异构计算。应用驱动的研究项目包括人工智能（AI）的新型架构，如深度学习神经网络的计算机架构，neuromorphic 计算和 bio-inspired 计算，新应用的硬件加速，如生物信息学应用，图形分析和机器人应用。

● 陈天石

中国科学院计算技术研究所研究员。研究方向为计算机体系结构和计算智能。寒武纪科技创始人兼 CEO。

陈天石于 2010 年在中国科学技术大学计算机学院获得工学博士学位。2016 年至今，中国科学院计算技术研究所 研究员，中科寒武纪科技 CEO，研发出我国首款人工智能芯片寒

武纪。

■ 主要成就

陈天石在 IEEE/ACM Transactions、Theoretical Computer Science、ISCA、HPCA、IJCAI、AAAI、SPAA、DATE 等重要期刊和会议上发表论文 40 余篇。曾先后获得获全国百篇优秀博士论文提名奖、中国计算机学会优秀博士论文奖、中国科学院优秀博士论文奖、中国科学院院长奖、教育部高等学校科学研究优秀成果奖、国家自然科学基金委员会“优青”、Intel 青年学者奖等荣誉。2016 年 3 月，陈天石、陈云霁联合创立了寒武纪科技公司，该公司是全球第一个成功流片并拥有成熟产品的智能芯片公司，拥有终端和服务器两条产品线。

● 施路平

清华大学教授，国家光存储工程研究中心主任。2012 年入选千人计划（A 类），2013 年 3 月入职清华，SPIE Fellow。

施路平于 1992 年在德国科隆大学获得科学博士学位。

1996 年 8 月-2013 年 3 月，任新加坡科学院数据存储研究院资深科学家，光学材料和系统实验室主任，非易失性存储器实验室主任，新加坡科学院人工认知存储器实验室主任。

■ 主要成就

参与创建并领导了新加坡科学院的半导体非易失性存储器，光存储，人工认知存储器研究领域。研究领域包括信息存储，集成光电子学，材料科学，人工认知存储器，自旋电子学，纳米科学与技术等，是人工认知存储器的主要开拓者之一。

2012 年入选千人计划（A 类），2013 年 3 月入职清华，SPIE Fellow。

已发表近 200 多篇学术论文（包括 Science，Nature Photonics，Phys Rev Lett，Advance Mat, Laser&Photonics Review, Scientific Reports），拥有 10 多项专利或专利申请 2004 年获颁新加坡国家科技奖。

● 余凯

前百度研究院副院长，深度学习实验室主任。地平线机器人技术创始人兼 CEO。

余凯在慕尼黑大学获得计算机科学博士学位，曾在微软、西门子和 NEC 工作。

2015 年 5 月 22 日，余凯从百度离职。同年 7 月，创办地平线机器人，致力于“define the brain of things”，打造万物智能时代的“AI Inside”，给人们日常生活的无数设备和产品装上“大脑”。

■ 主要成就

2009 年，在 PASCAL VOC 视觉识别竞赛中获得第一名。

2008 年和 2009 年，美国国家技术与标准局组织的 TRECVID 图像事件检测评比中两次获得多项第一名。

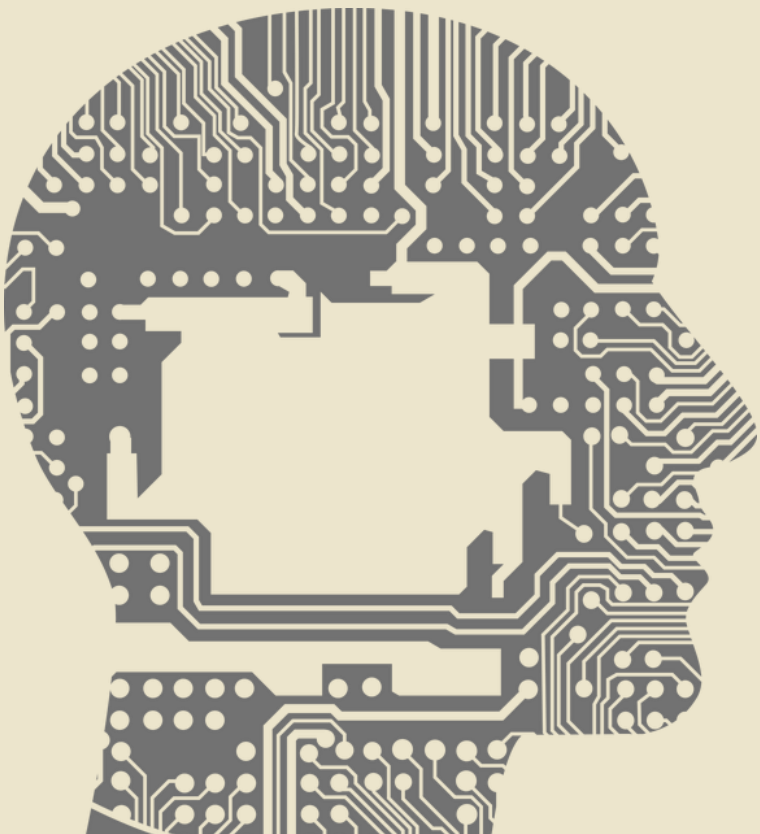
2010 年，带领团队在首届 ImageNet 大规模视觉识别竞赛中获得第一名（Geoffrey Hinton 团队于 2012 年获得第一名）。

余凯在深度学习，特征学习，贝叶斯学习，高斯过程，推荐系统，图像识别，图像检索等领域多有建树，在著名学术会议和杂志上发表了几十篇高质量论文，被同行引用达 7000 次以上。曾获得 1999 中国信号处理学会年会优秀论文奖，第 9 届 PKDD 国际会议最佳论文奖银奖，和第 30 届机器学习国际会议（ICML）的最佳论文奖银奖。

2013 年到 2014 年，余凯所带领的语音技术团队，深度学习技术团队，和图像技术团队，相继 3 次获得业界著名的百万美金“百度最高奖”，创造了百度公司内部各个技术&业务团队的记录。

5 application

应用领域篇



5 应用领域篇

随着人工智能芯片的持续发展，应用领域会随时间推移而不断向多维方向发展，本报告只选目前发展比较集中的几个行业做相关的介绍，如图 16 所示。

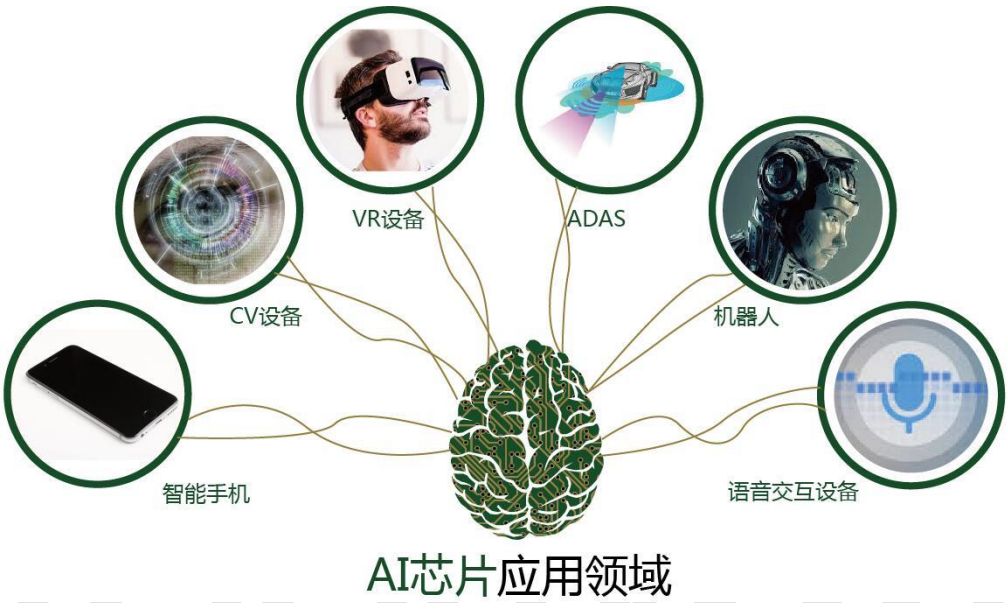


图 16 AI 芯片应用领域

（1）智能手机

2017 年 9 月，华为在德国柏林消费电子展发布了麒麟 970 芯片，该芯片搭载了寒武纪的 NPU，成为“全球首款智能手机移动端 AI 芯片”；2017 年 10 月中旬 Mate10 系列新品（该系列手机的处理器为麒麟 970）上市。搭载了 NPU 的华为 Mate10 系列智能手机具备了较强的深度学习、本地端推断能力，让各类基于深度神经网络的摄影、图像处理应用能够为用户提供更加完美的体验，图 17 显示了 Mate10 的成像效果对比图。

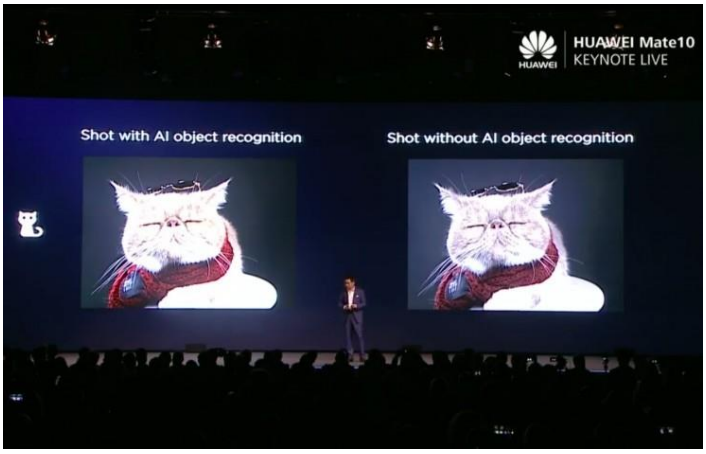


图 17 华为 Mate10 成像效果对比图

2017 年 9 月中旬，苹果发布以 iPhone X 为代表的手机及它们内置的 A11 Bionic 芯片。

A11 Bionic 中自主研发的双核架构 Neural Engine（神经网络处理引擎），它每秒处理相应神经网络计算需求的次数可达 6000 亿次。这个 Neural Engine 的出现，让 A11 Bionic 成为一块真正的 AI 芯片。A11 Bionic 大大提升了 iPhone X 在拍照方面的使用体验，并提供了一些富有创意的新用法。如更具革命性的 FaceID，它能够将传感器数据进行实时 3D 建模，并利用机器学习识别用户容貌改变，在此过程中的大量计算需求，都需要借助 A11 Bionic 和 Neural Engine 来满足，如图 18 所示。除此之外，A11 Bionic 内置了苹果自主设计的第一款 GPU。这款 GPU 是为 3D 游戏和 Metal 2（苹果在 WWDC 2017 上推出的新一代图像渲染技术框架）专门设计的，并且能够与机器学习技术和苹果随 iOS 11 推出的 Core ML（核心机器学习）框架相配合。

图 18 苹果的 Face ID

谷歌，高通同样在随后发布的产品中植入 AI 芯片，或许这将成为业界的一个新趋势，即使因为植入 AI 芯片能为用户带来真正美好体验，还需要等到有足够多的基于深度学习的 APP 出现才可实现。

（2）ADAS（高级辅助驾驶系统）

ADAS 是最吸引大众眼球的人工智能应用之一，它需要处理海量的由激光雷达、毫米波雷达、摄像头等传感器采集的实时数据。ADAS 的中枢大脑——ADAS 芯片市场的主要厂商包括被英特尔收购的 Mobileye、2017 年被高通以 470 亿美元惊人价格收购的 NXP，以及汽车电子的领军企业英飞凌。随着英伟达推出自家基于 GPU 的 ADAS 解决方案 Drive PX2，英伟达也加入到战团之中。

相对于传统的车辆控制方法，智能控制方法主要体现在对控制对象模型的运用和综合

信息学习运用上，包括神经网络控制和深度学习方法等，得益于 AI 芯片的飞速发展，这些算法已逐步在车辆控制中得到应用。

（3）CV（计算机视觉（Computer Vision））设备

需要使用计算机视觉技术的设备，如智能摄像头、无人机、行车记录仪、人脸识别迎宾机器人以及智能手写板等设备，往往都具有本地端推断的需要，如果仅能在联网下工作，无疑将带来糟糕的体验。而计算机视觉技术目前看来将会成为人工智能应用的沃土之一，计算机视觉芯片将拥有广阔的市场前景。

计算机视觉领域全球领先的芯片提供商 Movidius，目前已被英特尔收购，大疆无人机、海康威视和大华股份的智能监控摄像头均使用了 Movidius 的 Myriad 系列芯片。

目前国内做计算机视觉技术的公司以初创公司为主，如商汤科技、阿里系旷视、腾讯优图，以及云从、依图等公司。在这些公司中，未来有可能随着其自身计算机视觉技术的积累渐深，部分公司将会自然而然转入 CV 芯片的研发中，正如 Movidius 走的也是从计算机视觉技术到芯片研发的路径。

（4）VR 设备

VR 设备芯片的代表为 HPU 芯片，是微软为自身 VR 设备 Hololens 研发定制的。这颗由台积电代工的芯片能同时处理来自 5 个摄像头、1 个深度传感器以及运动传感器的数据，并具备计算机视觉的矩阵运算和 CNN 运算的加速功能。这使得 VR 设备可重建高质量的人像 3D 影像，并实时传送到任何地方。

（5）语音交互设备

语音交互设备芯片方面，国内有启英泰伦以及云知声两家公司，其提供的芯片方案均内置了为语音识别而优化的深度神经网络加速方案，实现设备的语音离线识别。稳定的识别能力为语音技术的落地提供了可能；与此同时，语音交互的核心环节也取得重大突破。语音识别环节突破了单点能力，从远场识别，到语音分析和语义理解有了重大突破，呈现出一种整体的交互方案。

语音交互正在悄悄改变人们的家居生活习惯，如居于客厅核心位置的智能电视，越来越多的消费者习惯在沙发上使用语音换台，语音作为智能家居入口将有广阔的想象空间。

（6）机器人

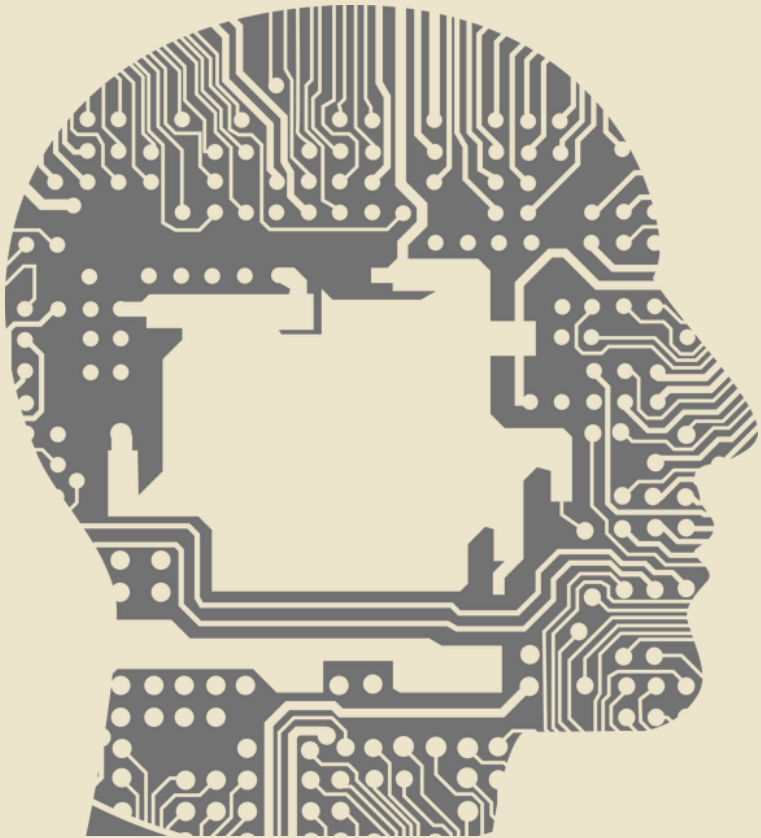
无论是家居机器人还是商用服务机器人均需要专用软件+芯片的人工智能解决方案，这方面典型公司有由前百度深度学习实验室负责人余凯创办的地平线机器人，当然地平线机器人除此之外，还提供 ADAS、智能家居等其他嵌入式人工智能解决方案。

在移动端推断领域，呈现给我们的是一个缤纷的生态。因为无论是 ADAS 还是各类 CV、

VR 等设备领域，人工智能应用仍远未成熟，各人工智能技术服务商在深耕各自领域的同时，逐渐由人工智能软件演进到软件+芯片解决方案是自然而然的路径，因此形成了丰富的芯片产品方案。

6 trend

趋势篇



6 趋势篇

目前主流 AI 芯片的核心主要是利用 MAC（Multiplier and Accumulation，乘加计算）加速阵列来实现对 CNN（卷积神经网络）中最主要的卷积运算的加速。这一代 AI 芯片主要有如下 3 个方面的问题。

（1）深度学习计算所需数据量巨大，造成内存带宽成为整个系统的瓶颈，即所谓“memory wall”问题。

（2）与第一个问题相关，内存大量访问和 MAC 阵列的大量运算，造成 AI 芯片整体功耗的增加。

（3）深度学习对算力要求很高，要提升算力，最好的方法是做硬件加速，但是同时深度学习算法的发展也是日新月异，新的算法可能在已经固化的硬件加速器上无法得到很好的支持，即性能和灵活度之间的平衡问题。

因此，我们可以预见，下一代 AI 芯片将有如下的几个发展趋势。

趋势一：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，如图 19 所示。

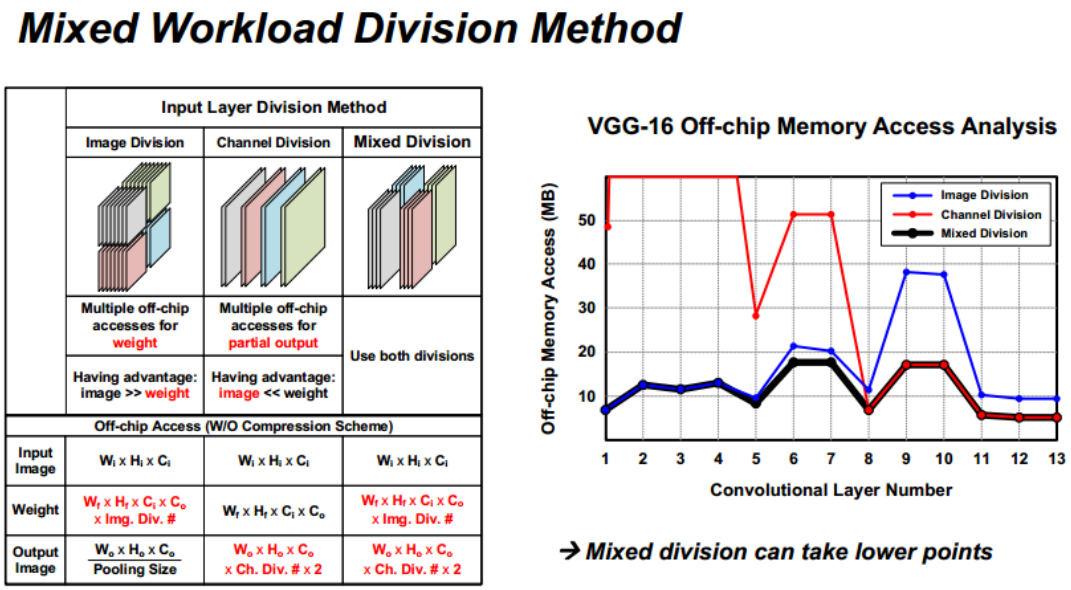


图 19 分解卷积可降低消耗

趋势二：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，如图 20 所示。

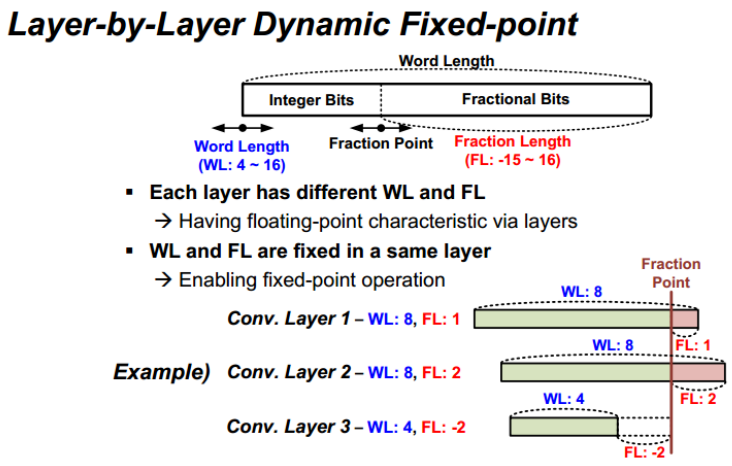


图 20 逐层动态定点方法

趋势三：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。

趋势四：更稀疏的大规模向量实现

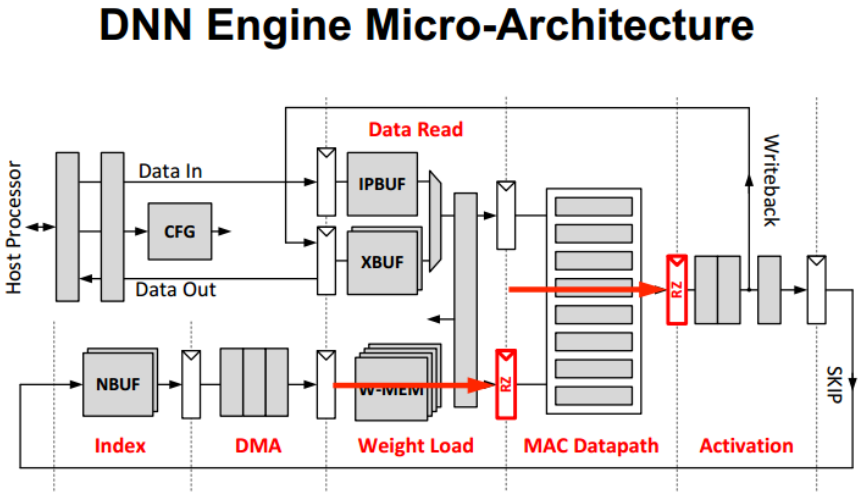


图 21 五级流水线结构

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构，如图 21 所示，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。

趋势五：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。

参考文献

[1] 施羽暇.人工智能芯片技术研究，《电信网技术》，2016，12（12）.

[2] Philipp Gysel, Mohammad Motamedi & Soheil Ghiasi.HARDWARE-ORIENTED APPROXIMATION OF CONVOLUTIONAL NEURAL NETWORKS, 20 Oct 2016.

[3] Martin Thoma.Analysis and Optimization of Convolutional Neural Network Architectures, 31 Jul 2017.

[4] Bohyung Han Computer Vision Lab. Lecture 9: CNN Optimization CSED703R: Deep Learning for Visual Recognition (2017F).

[5] Dongjoo Shin, Jinmook Lee, Jinsu Lee, Juhyoung Lee, and Hoi-Jun Yoo DNPU: An Energy-Efficient Deep Neural Network Processor with On-Chip Stereo Matching Semiconductor System Laboratory School of EE, KAIST.

[6] Paul Whatmough, S. K. Lee, H. Lee, S. Rama, D. Brooks, G.-Y. Wei Harvard University, Cambridge, MA A 28nm SoC with a 1.2GHz 568nJ/ Prediction Sparse Deep-Neural-Network Engine with >0.1 Timing Error Rate Tolerance for IoT Applications.

[7] Paul N. Whatmough S. K. Lee, N. Mulholland, P. Hansen, S. Kodali, D. Brooks, G.-Y. Wei DNN ENGINE: A 16nm Sub-uJ DNN Inference Accelerator for the Embedded Masses.

[8] Vivienne Sze , Yu-Hsin CHen, Tien-Ju Yang, Joel S. Emer Efficient Processing of Deep Neural Networks: A Tutorial and Survey Proceedings of the IEEE Vol. 105, No. 12, December 2017.

[9] 张蔚敏，蒋阿芳，纪学毅.人工智能芯片产业现状.《电信网技术》，2018 年 2 月第 2 期.

[10] 张贝贝.人工智能时代芯片产业迎来大发展产业纵横，2017.09.17.

[11] 曾毅，刘成林，谭铁牛.类脑智能研究的回顾与展望.计算机学 报第 39 卷 第 1 期，2016 年 1 月.

版权声明

AMiner 研究报告版权为 AMiner 团队独家所有，拥有唯一著作权。AMiner 咨询产品是 AMiner 团队的研究与统计成果，其性质是供用户内部参考的资料。

AMiner 研究报告提供给订阅用户使用，仅限于用户内部使用。未获得 AMiner 团队授权，任何人和单位不得以任何方式在任何媒体上（包括互联网）公开发布、复制，且不得以任何方式将研究报告的内容提供给其他单位或个人使用。如引用、刊发，需注明出处为“AMiner.org”，且不得对本报告进行有悖原意的删节与修改。

AMiner 研究报告是基于 AMiner 团队及其研究员认可的研究资料，所有资料源自 AMiner后台程序对大数据的自动分析得到，本研究报告仅作为参考，AMiner团队不保证所分析得到的准确性和完整性，也不承担任何投资者因使用本产品与服务而产生的任何责任。