人工智能芯片发展现状及趋势

罗亚文[1]

1. 人工智能与自动化学院　人工智能本硕博2001班，学号U202015237

摘　要：人工智能芯片是人工智能技术的重要组成部分，是实现人工智能算法的硬件基础，也是人工智能时代的战略制高点。当前人工智能芯片的市场由英伟达、英特尔、赛灵思、谷歌等国际巨头占据主要份额，国内企业近年来加速布局，在边缘推理芯片等领域发展较快。本文主要围绕智能芯片的“高效能、低能效”，从数据复用技术和高效加速计算的角度来阐述其核心内涵，从数据复用技术、高效并行架构、模型轻量化设计等层面来对当前智能芯片的发展脉络进行梳理，也对人工智能芯片未来的发展方向进行了展望。

关键词：人工智能芯片；高效能；低功耗；数据复用技术；高效并行架构；模型轻量化；

1　引言

人工智能芯片目前尚无准确定义，但可以从广义和狭义两个角度来对其进行阐释：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，人工智能芯片针对人工智能算法做了特殊加速设计的芯片，这也被视为通常意义下对人工智能芯片的定义[1]。

人工智能芯片目前有俩种发展路径。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仍旧发挥着不可替代的作用[2]。另外一种发展路径则是彻底颠覆了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Loihi芯片、IBM公司的TrueNorth芯片等作为代表[3]。

人工智能的发展路线图，可以归纳为三个阶段：短期目标，实现以异构计算为主加速的各类应用算法的落地；中期目标，发展自重构、自适应、自学习、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芯片的终极目标迈进[4]。

虽然摩尔定律放缓，但人工智能芯片作为推动人工智能技术不断进步的硬件基础，未来10年仍将是人工智能芯片发展的重要时期[5]，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　人工智能芯片概述

为了支持多样的AI计算任务和性能要求，理想的AI芯片需要具备高度并行的处理能力，能够支持各种数据长度的按位、固定和浮点运算[5]；比当前大几个数量级的存储器带宽用于存储海量数据；低内存延迟及新颖的架构，以实现计算元件和内存之间灵活而丰富的连接。除此之外，这些连接都需要确保“高效能、低功耗”。

2.1　人工智能芯片分类

GPU：即图形处理单元，是一种由大量运算单元组成的大规模并行计算架构芯片，主要用于处理图形、图像领域的海量数据运算。GPU上集

成了规模巨大的计算矩阵，从而具备了更强大的浮点运算能力和更快的并行计算速度，与CPU 相比，更加适用于解决人工智能算法的训练难题。英伟达的GPU 目前在人工智能计算市场上占据了主导地位。

半定制化 FPGA：即现场可编程门阵列。与GPU的固定电路不同，使用者可以根据不同的应用需求，使用硬件描述语言对 FPGA 芯片上集成的基本门电路和存储器进行重新定义。按照新的定义完成烧录后，FPGA 芯片内部的电路就固化成了实际的连线，从而具备了使用者所需要的功能。将FPGA和CPU 对比可以发现两个特点：1、FPGA没有内存和控制所带来的存储和读取部分速度更快。2、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。

全定制化ASIC：即专用芯片，是一种根据特殊应用场景要求进行全定制化的专用人工智能芯片。与 FPGA 相比，ASIC 芯片无法通过修改电路进行功能扩展；而与 CPU、GPU 等通用计算芯片相比，其性能高、功耗低、成本低，很适合应用于对性能功耗比要求极高的移动设备。

神经拟态芯片：即类脑芯片，是一种对人脑的神经网络结构进行物理模拟的新型芯片架构。通过模拟人脑的神经网络工作机理实现感知和认知等功能。IBM研发的 TrueNorth 芯片就是一种典型的类脑芯片，其逻辑结构颠覆了经典冯·诺依曼架构，把定制化的数字处理内核当作神经元，把内存当作突触，CPU、内存及通信元件等完全集成在本地[6],实现了算存一体，突破了冯·诺依曼架构中CPU与内存之间的内存墙瓶颈。

3 人工智能芯片中的数据复用技术

3.1 数据复用的意义

在设计针对深度学习的人工智能芯片时，功耗、性能、面积是衡量芯片优劣的重要标准。而其中功耗、性能与架构设计时需要完成的数据传输量密切相关。如图1所示，数据在计算单元之间、 芯片内部与芯片外部之间传输所消耗的能量远大于计算的能量。

尽可能减少数据在芯片内部与外部之间的流动，是人工智能芯片架构设计的原则之一。减少数据移动的一种有效方法是数据复用。

数据复用，即在计算过程中对同一数据进行重复的利用，是一种常见的减少存储器重复访问的方法。深度学习领域，普遍存在的一个现象是一份数据会多次参与计算。因此，利用该特点，尽可能使数据只移动一次但被多次使用[7]，就是数据复用方法。

图 1 消耗和计算在不同类型的数据传输上的能量比例关系

3.2 数据复用方法分类

人工智能芯片通常采用三种数据复用模式：输入数据复用，输出数据复用和权重数据复用。下面详细介绍三种复用方式，图2给出了基本网络层的数学模型以方便理解。

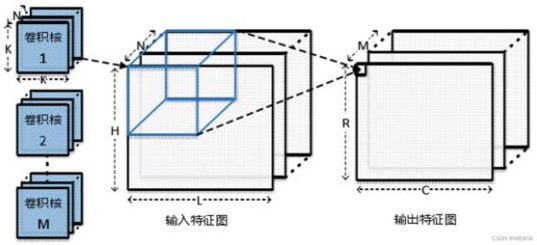

图 2 基本网络层数学模型

输入数据复用：如图3所示，输入数据复用对输入缓存具有最少的访问次数，其可以分为3个步骤：①计算核心把输入特征图读入局部输入寄存器；②计算核心充分复用这些输入数据，更新输出缓存中的所有相关的输出部分和；③更新后的输出部分和会重新写会输出缓存。当新的输入数据被读入计算核心时会重复上述3个步骤。

权重数据复用：如图4所示，权重数据复用对权重具有最少的访问次数，其分为3个步骤：①计算核心读取Tn个输入特征图分块到局部的输入寄存器；②计算核心利用这些输入数据更新Tm个通道的输出部分和；③存储在权重缓存中的Tm个Tn通道的卷积核权重被充分复用，以更新存储

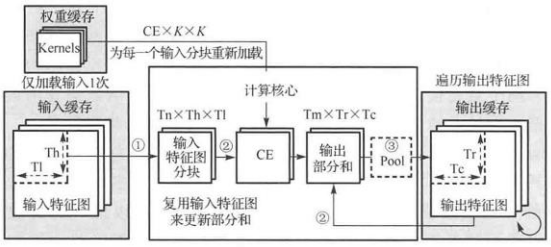

图 3　输入数据复用

在输出缓存中的Tm个通道的R*C输出部分和。重复上述3个步骤以完成整个卷积层的全部计算。

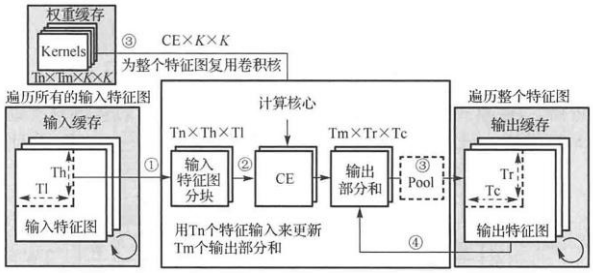

图 4　权重数据复用

输出数据复用：如图5所示，输出数据复用对输出缓存具有最少的访问次数，其分为3个步骤：①计算核心把输入特征图的各通道读入局部的输入缓存器；②存储在计算核心输出寄存器中的输出部分和会被充分复用，以完成3为卷积通道方向上的完全累加；③最终的输出特征图会在池化之后再写入输出缓存。计算过程中不会再有其他对输出缓存的访问，对于剩余的输出特征图计算，会重复上述3个步骤。

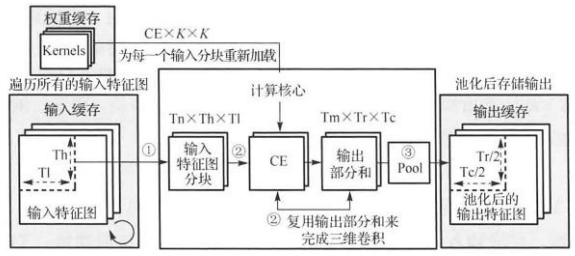

图 5　输出数据复用

混合数据复用：该数据复用模式，将根据每一层单独分配针对该层最优的数据复用模式。因此涉及寻找最优的数据复用模式，需要探索各种分块参数下的不同模式的访存能耗。

3.3　数据复用举例

3.3.1　达芬奇架构

达芬奇架构，是华为自研的面向AI计算特征的全新计算架构，具备高算力、高能效、灵活可裁剪的特性，是实现万物智能的重要基础。具体来说，达芬奇架构采用3D Cube针对矩阵运算做加速，大幅提升单位功耗下的AI算力，每个AI Core可以在一个时钟周期内实现4096个MAC操作，相比传统的CPU和GPU实现数量级的提升。

华为达芬奇架构属于“输出复用”的架构。其架构示意图如图6所示，该架构通过把乘累加（MAC）单元构建为二维的阵列，实现算力的叠加。通过在行方向上和列方向上分别广播左矩阵和右矩阵的元素，并把乘法结果累加到本地，该计算阵列可以高效实现矩阵乘法运算。该架构结合把卷积计算转换为矩阵运算的im2col操作，也可以高效实现二维卷积计算。

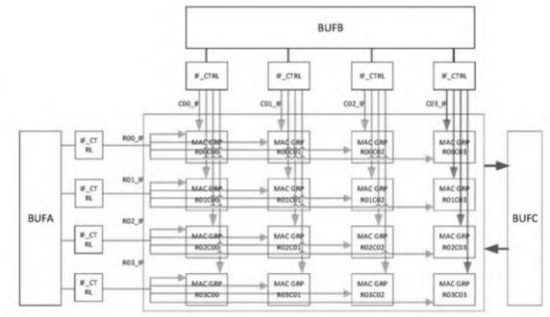

图 6　华为达芬奇架构示意图

3.3.2　ShiDianNao架构

寒武纪公司是国内较早进行针对深度学习的计算阵列研究的公司，其研制的ShiDianNao架构提出利用二维PE阵列处理计算机视觉类深度学习模型，专注于执行卷积神经网络中的二维卷积计算。 ShiDianNao的计算阵列架构如图7所示。根据公开的论文[8]，该架构变换向二维计算阵列输入的特征图和权重，卷积计算的部分和累加在PE内部，属于“输出复用”的架构。其具体架构如图7所示。

3.3.3　脉动阵列架构

谷歌公司利用脉动计算阵列实现了 用于深度学习加速的芯片TPU。TPU芯片的核心计算 单元同样是使用计算阵列的方式实现的，其论文中公开的计算阵列架构如图8所示[9]。

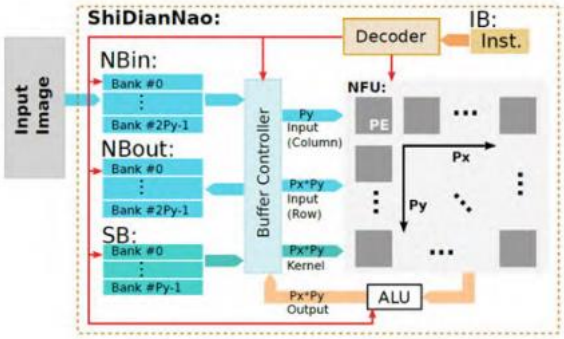

图 7　ShiDianNao架构示意图

与达芬奇架构不同，脉动阵列架构中，权重存储在PE内部不动，乘法计算的部分和并不累加在PE内部，而是不断向下方的相邻PE传递。最终的乘累加结果在整个计算阵列的下方输出。TPU的计算阵列架构属于“权重复用”架构。

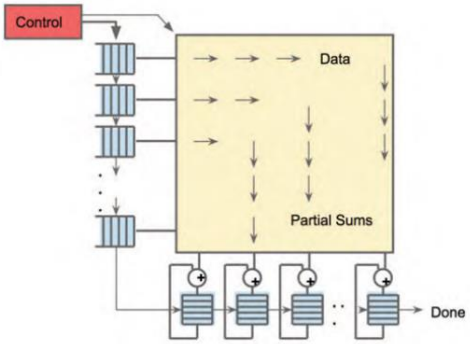

图 8　TPU使用的脉动架构示意图

3.4 硬件上利用数据复用的办法

数据复用在硬件上有四种实现方法：空间上的广播；空间上的归约；时间上的广播；时间上的归约[10]。

广播即为把一份数据传输至多个位置。 空间上的广播即传输一份数据到多个计算单元，从而避免了数据反复的传输，实现数据复用。归约就是指把多个计算结果通过某种方式合并到一起，得到一份数据，同样可以避免数据多次传输，实现数据复用。

空间上的归约一种常用实现方式是加法树，即把多个位置的数据通过加法树合并到一起，得到这些数据的和。时间上的广播指把一份数据在时间维度上传输到多个使用该数据的时机。一种简单的实现方式就是把一份会反复使用的数据存储在寄存器中不动，需要使用时，直接从寄存器中取用即可，避免该数据被反复从片外存储器导入寄存器。

时间上的归约，指在时间维度上把多份数据合并到一起，产生一份数据。时间上的归约的常见实现方式是累加器：通过累加器，把在时间上多次计算得到的结果与累加器的结果累加，把累加结果再存储到累加器中，实现把时间上多次出现的数据合并到一起的效果，同样可以避免数据反复传输[7]。

4　人工智能芯片中的高效并行架构

4.1　并行计算的意义

绝大多数的机器学习算法都涉及到了数据密集型和计算密集型的运算，使用并行计算技术可以大大提高算法的效率和性能，使得模型的训练和推理都更加快速和准确。并行主要有俩类，模型并行和数据并行以及混合并行。

模型并行即为分布式系统中的不同机器负责网络模型的不同部分。例如，神经网络模型的不同网络层被分配到不同的机器，或者同一层内部的不同参数被分配到不同机器。数据并行则为不同的机器有同一个模型的多个副本，每个机器分配到不同的数据，然后将所有机器的计算结果按照某种方式合并。混合并行即为在一个集群中，既有模型并行，又有数据并行。

4.2　并行计算模型

并行计算模型是指基于并行计算技术把一个大型问题划分为若干个小型问题，每个小型问题由一个独立的处理器负责计算，然后将这些小型问题的解合并成为整体问题的解。常见的并行计算模型有SPMD、MPMD、SIMD和MIMD等几种。

SPMD：所有的处理器运行同一个程序，但是处理器所操作的数据是不同的。每个处理器所负责的数据不同，每个处理器就像一个串行处理器一样独立工作，每个处理器都有其独立的内存和寄存器，数据在处理器之间传递。

MPMD：指不同的处理器运行不同的程序，但是处理器所操作的数据相同或者有交叉。每个处

理器所使用的程序不同，但是它们可以访问相同的数据存储区域，每个处理器都有其独立的内存和寄存器，数据在处理器之间传递。

SIMD：指多个处理器执行相同的指令，但是数据不同。所有的处理器都执行同一个指令，并且指令是在同一时刻被执行的，但是每个处理器所处理的数据是不同的。这种模型可以获得很高的计算效率，但是需要求解的问题必须可以分解为相似的并行任务。

MIMD：模型是指多个处理器执行不同的指令，并且数据也不同。每个处理器所执行的指令和数据都独立，处理器之间是没有任何的同步和互动。

4.3 并行计算技术

使用并行计算技术可以极大地提高计算效率，使得模型的训练和推理速度更快、准确度也更高。常见的并行计算技术有分布式计算、GPU并行计算和多核并行计算。

分布式计算：指将一个大任务分解成若干个小任务，然后将这些小任务分布在多个计算机中进行计算。每个计算机都拥有自己的内存和处理器资源，通过网络互相连接起来，通过交换与同步数据和结果来完成任务。分布式计算技术可以增加计算机系统的可扩展性以及可靠性。

GPU并行计算：利用GPU来完成计算任务。GPU可以并行处理大量的矩阵乘法运算、向量加法运算等复杂的计算任务，相比于CPU的串行处理，GPU具有更高的计算效率。GPU并行计算可以利用CUDA、OpenCL等GPU编程平台进行编程。

多核并行计算：利用计算机多个核心来同时处理不同的运算任务。多核并行计算可以增加计算机系统的处理能力和运算速度，提高计算系统的效率和性能。

4.4 卷积层并行计算原理

CNN网络在神经网络中具有举足轻重的地位，其具有5种层次的并行优化。

突触并行：单次卷积窗内的计算并行。K×K个乘法操作无数据依赖，可以并行执行。
一次乘法操作类似神经元的一个突触信号作用在树突上，最大突触并行度为K×K。如图9所示。

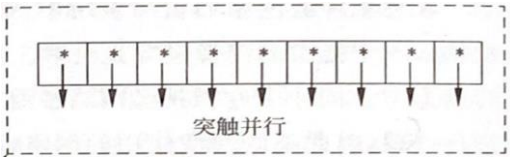

图 9　突触并行

神经元并行：多次卷积窗间的计算并行，多个卷积窗口输入特征值会有重叠，重叠部分的数据具有重用性，充分利用可以减少片上缓存需求和访存次数，从而降低功耗同时提高性能。

IFM并行：多个通道间的计算并行，多个通道的卷积运算之间的数据没有重叠，硬件提供足够的带宽和算力，就能进行IFM的并行计算。

OFM并行：多个卷积核卷积过程的计算并行，多个卷积核运算的权重值没有重合，但IFM数据可以被所有卷积核共享，输入数据有重用性，可减少片外数据访存，降低计算功耗。

批处理并行：输入一个批次的图像进行批处理，任务级别的并行。不同图像使用的网络权重可以复用，充分利用已经搬运到片上的权重数据，避免频繁访存，降低功耗和延时，满足片上海量计算资源的数据需求。

5 模型轻量化设计

5.1 模型轻量化设计的意义

模型的大小不仅影响人工智能芯片的MAC单元、内存等片上资源需求，而且还影响通信带宽需求，与智能计算的性能和功耗直接相关。然而为了实现深度神经网络的精度最大化，神经网络层数不断增加，结构日益复杂，模型越来越大，由此带来巨大的计算和存储代价，最终导致神经网络的应用受到严重阻碍，尤其在一些计算能力相对较弱的边缘设备上，问题更加突出。

为了在有限资源的硬件设备上更好的实现智能计算加速，需要对网络模型压缩和优化，这就设计人工智能芯片的软硬件协同设计技术。目前的人工智能芯片的软硬件协同设计方法主要包括俩类：1、降低操作和操作数位宽的方法，如低位宽神经网络和二值化神经网络。2、减少操作数量和模型大小的方法，如稀疏化神经网络。

5.2 低位宽神经网络

对神经网络进行位宽的量化，是缩减模型的常用方法之一，而模型量化是优化神经网络位宽的首要方法。量化是指将模型数据映射到一组量化级别上，而这组量化级别的数量反映了量化后数据表示需要的位宽。量化过程必然存在误差，然而深度神经网络被过度参数化，进而包含足够的冗余信息，裁剪这些冗余信息并不会导致明显的准确度下降。

神经网络的基本运算是权重和激活值的乘累加，因此模型量化主要针对俩类数据，一是权重数据，二是激活数据。如果将着两种数据之一量化到{-1，1}时，神经网络的乘累加操作就进一步简化为按位操作，这可以极大的降低硬件的计算代价和存储代价。量化的方法很多，通常按照量化级别间距是否相等，可以分为线性量化方法和非线性量化方法。

5.2.1 线性量化

线性量化采用线性映射函数来确定量化级别，如图10所示，其相邻量化级别之间通常是等间距的或者是均匀间隔的。

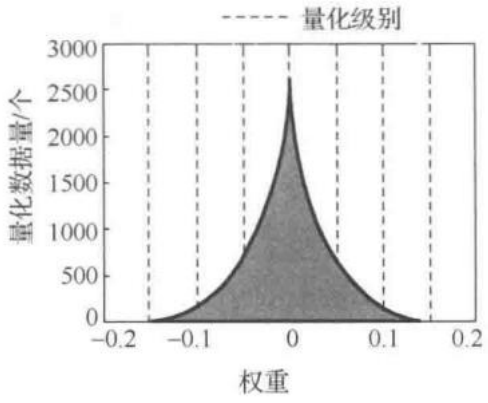

图 10 线性量化方法

5.2.2 非线性量化

研究表明，权重值和激活值的分布通常是不均匀的，因此与量化间隔均匀分布的线性方法相比，采用非线性量化可以使得权重数据和激活数据在各量化级别上的分布更加均匀，每个级别的使用效率更高，量化误差更小，更有利于保证网络精度。非线性量化主要有俩种，对数域量化和权值共享。

对数域量化：该量化方法如图11所示，是指根据对数分布函数确定量化级别的方法。以2为底的对数量化方法不仅可以将乘法运算简化为移位运算，降低运算代价，而且可以显著降低量化带来的精度损失。例如，增量网络量化[11]通过将大、小权重值划分为不同的组，然后对权重进行迭代量化和再训练，可以进一步降低网络精度损失。

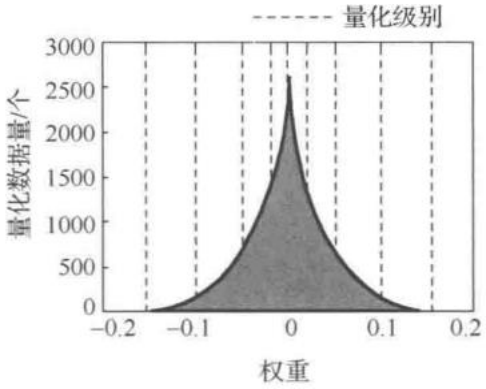

图 11 对数域量化方法

权重共享量化：通过某种映射使得多个权重共享一个值，以减少权重值的数量，这样就可以用更低位宽的索引值来量化权重。

5.3 稀疏化神经网络及其架构设计

除了减少操作数/权重的大小之外，还有大量关于减少操作数和模型大小方法的研究。这些技术大致分为激活稀疏性、网络剪枝和压缩网络架构三类。

5.3.1 利用激活稀疏性

激活函数是神经网络中的非线性映射，它可以极大地增强网络的学习能力和表达能力，是神经元中非常重要的组成部分。

人工智能芯片设计架构的构建可以利用RELU输出的稀疏性，通过数据压缩来降低芯片面积和减少片外通信带宽。除了数据压缩，硬件可以直接跳过0值的MAC操作来降低损耗，提升性能。例如，Eyersis基于这种方法，在以AlexNet为基准测试集的计算中获得了45%的PE功耗节省。

对绝对值较小的激活值进行修剪，还可以进一步提升激活输出的稀疏度。对这种稀疏性的利

用同样可以使得人工智能芯片获得性能的提升和功耗的降低。例如，Minerva[12]通过这种较小激活值剪枝的方法，在MNIST、Forest、Reuters、WebKB和20NG这五个数据集上取得了平均50%的功耗节省。

5.3.2 网络剪枝

在20世纪90年代早期，剪枝技术被提出来，可以在不需要再训练的情况下，将一个训练过的大型网络缩减为一个较小的网络。这使得神经网络可以部署在资源受限的环境中，比如嵌入式系统。

剪枝去除对结果准确性没有显著贡献的冗余参数或神经元。当权重系数为零、接近零或被复制时，可能会出现这种情况。因此，修剪可以降低计算复杂度。如果修剪后的网络被重新训练，它提供了逃避之前的局部最小的概率并进一步提高精度。

根据网络元素类型，可以分为神经元剪枝和连接剪枝；根据剪枝前后网络结构是否是改变，可以分为结构化剪枝和非结构化剪枝；根据在推理阶段是否有剪枝，可以分为静态剪枝和动态剪枝。

5.3.3 压缩网络架构

神经网络的压缩算法旨在将一个庞大而复杂的预训练模型转化为一个精简的小模型。按照压缩过程对网络结构的破坏程度，可以将模型压缩技术分为前端压缩和后端压缩两部分。

前端压缩，是指在不改变原网络结构的压缩技术，主要包括知识蒸馏、轻量级网络（紧凑的模型结构设计）等。后端压缩，是指包括低秩近似、参数量化以及二值网络等，目标在于尽可能减少模型大小，会对原始网络结构造成极大程度的改造。

MobileNet是较为典型的网络架构，其使用深度可分离卷积替代常规卷积，大大降低了计算量和参数数量。

5.4 二值网络

神经网络的数据位宽可以低至1bit。与此相关的网络被称为二值神经网络。BinaryConnect[13]中映入了二值权值的概念，即使用二值权值MAC中 的乘法运算简化为加法和减法运算。

YodaNN是第一个二值权值CNN加速器，其主要关注电路级优化。得益于二值权值无乘法设计，YodaNN实现了很高的能效。

6 人工智能芯片未来趋势

随着以人工智能、物联网、5G 等为核心的新一代信息技术的高速发展，涌现出越来越多新的应用场景和需求。未来物联网领域将需要体积更小、功耗更低、能效比更高的人工智能芯片。

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，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，可以概述为以下几个方面：

1) 可重构计算：随着网络模型的复杂度不断增加，对于计算硬件的要求也越来越高。可重构计算架构允许硬件架构功能随着软件变化而变化，灵活的适应新的算法、架构和任务，兼具有PU/GPU的通用性，ASIC的高性能和低功耗、可按照需求迭代且开发周期短。

2)计算存储一提化架构设计：深度学习负载通常都具有数据密集型的特点，需要大量的存储和各层次存储器之间的数据搬移，导致冯诺依曼架构的内存墙问题更加突出。计算存储一体化架构可以一定程度上弥补计算单元和存储器之间的差距。

3) 软硬件协同的模型压缩：神经网络本身数据量大，对片上资源要求高，在算法层面，如何控制压缩力度的边界，计算量和精确度之间如何折中，都值得深究，如何对压缩后的模型进行高效的片上推理运算也是一个重要的研究方向。

4)AutoML:内涵是通过机器学习技术把参数和结构的挑战交给机器，解决手工调参的问题。

参 考 文 献

[1]徐晨.美国人工智能芯片最新发展[J].电子元器件与信息技术,2022,6(07):19-22.DOI:10.19772/j.cnki.2096-4455.2022.7.006.

[2] 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.

[3] 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.

[4]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(02):14-19.DOI:10.19942/j.issn.2096-5915.2021.2.013.S

[5]尹首一,郭珩,魏少军.人工智能芯片发展的现状及趋势[J].科技导报,2018,36(17):45-51.

[6] 张军阳，王慧丽，郭阳，等 . 深度学习相关研究综述 [J]. 计算机应用研究，2018（7）：1 921-1 928，1 936

[7]林广栋.人工智能芯片的数据复用技术应用[J].集成电路应用,2023,40(04):30-34.DOI:10.19339/j.issn.1674-2583.2023.04.011.

[8] Du Z, Fasthuber R, Chen T, et al. ShiDianNao:Shifting vision processing closer to the sensor[J].Acm Sigarch Computer Architecture News, 2015, 43(3):92-104.

[9] Norman P.Jouppi,Cliff Young,Nishant Patil,David Patterson,et al.In Datacenter Performance Analysis of a Tensor Processing Unit[C].In Proceedings of ISCA '17,Toronto,ON, Canada,2017.

[10] Hyoukjun Kwon,Prasanth Chatarasi, Michael Pellauer, Angshuman Parashar, Vivek Sarkar, Tushar Krishna.2019 Understanding Reuse, Performance,and Hardware Cost of DNN Dataflows:A Data-Centric Approach UsingMAESTRO[C].In The 52nd Annual IEEE/ACM International Symposium onMicroarchitecture(MICRO-52),Columbus, OH,USA,ACM,NY,USA,2019.

[11] Chen W, Wilson J T, Tyree S, et al. Compressing neural networks with the hashing trick[C].ICML, Lille,2015.

[12] Reagen B, Whatmough P, Adolf R, et al.Minerva: Enabling low-power, highly-accurate deep neural network accelerators[C].ISCA, Seoul, 2016.

[13] Courbariaux M, Bengio Y, David J P . Binaryconnect: Training deep neural networks with binary netwokrs during peopogations[C].NIPS, Montreal,2015.

Current Status and Trends of Artificial Intelligence Chip Development

LUO Yawen[1,]

1. *College of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430000, China*

Abstract: Artificial intelligence chips are an important component of artificial intelligence technology, the hardware foundation for implementing artificial intelligence algorithms, and also a strategic high ground in the era of artificial intelligence. At present, the market of artificial intelligence chips is dominated by international giants such as Nvidia, Intel, Xilinx, Google, etc. Domestic enterprises have accelerated their layout in recent years and developed rapidly in such fields as edge reasoning chips. This article mainly focuses on the "high efficiency and low energy efficiency" of smart chips, and elaborates on its core connotation from the perspectives of data reuse technology and efficient accelerated computing. It sorts out the current development trend of smart chips from the perspectives of data reuse technology, efficient parallel architecture, and model lightweight design, and also looks forward to the future development direction of artificial intelligence chips.

Keywords: artificial intelligence chip; high efficiency; low power consumption; data reuse technology; efficient parallel architecture; model lightweight;