嵌入式智能计算加速技术综述

李欣瑶[1,2*],刘飞阳[1,2],李鹏[1,2]

1. 航空工业西安航空计算技术研究所,陕西 西安 710068
2. 机载、弹载计算机航空科技重点实验室,陕西 西安 710068

摘 要:深度神经网络在处理计算机视觉和语音识别等复杂智能问题方面具有巨大的优势。为满足航空电子系统智能化发展的应用需求,以及机载嵌入式环境对实时性、功耗、体积等方面的制约,本文对适用于嵌入式环境的智能计算加速技术展开研究,重点探索了智能计算芯片架构技术、CNN/RNN 深度神经网络模型压缩技术、智能计算加速平台架构及开发流程等关键技术问题,旨在解决资源受限的嵌入式环境下深度神经网络的部署与优化问题,为下一代智能化航空电子系统提供理论基础和技术支撑。

关键词:智能计算;模型压缩;卷积神经网络;循环神经网络;硬件加速

Survey of Embedded Intelligent Computing Acceleration Technology

Li Xinyao[1,2*], Liu Feiyang[1,2], Li Peng[1,2]

1. *Xi'an Aeronautics Computing Technique Research Institute*, *Xi'an* 710068, *China*
2. *Aviation Key Laboratory of Science and Technology on Ariborne and Missileborne Computer*, *Xi'an* 710068, *China*

Abstract: Deep neural networks have tremendous advantages in dealing with complex computer vision and speech recognition issues. In order to meet the application requirement of the intelligent avionics system and the constraints of real-time, power consumption and volume in the embedded environment, this paper studies the intelligent computing acceleration technologies applicable to embedded environments. This paper focuses on the key technical issues of intelligent computing chip architecture, CNN/RNN model compression, intelligent computing acceleration platform and development process, solving the problem of deep neural network deployment and optimization in the resource-constrained embedded environment. This paper aims at provide theoretical basis and technical support for the design of next generation of intelligent avionics systems.

Key Words: intelligent computing; model compression; convolutional neural network; recurrent neural network; hardware acceleration

航空电子系统是军用飞机机载电子系统的综合,是整个飞机平台信息获取、计算处理、数据共享、任务执行的基础。未来的航空电子系统必将向着综合化、智能化的方向发展[1]。智能化航空电子系统的应用需求主要来源于两个方面:一是根据综合态势信息进行辅助决策,如智能任务规划、自主飞行等;二是目标检测识别跟踪等。美国国防科技公司开发的 X47B 作为人类历史上第一架无须人工干预、完全由电脑操纵的智能化无人驾驶战斗机原型,可实现自主规划航迹与自主空中加油,是美国空军作战武器的革命性创新。英国国防部研发中的最新隐身无人战斗机"雷神"由计算机系统进行智能化自动控制,无须地面指示便可实现自主起降、自主飞行、自主防卫,为英国空军未来进攻性力量的组成提供决策依据。

近年来,在高性能计算、大数据以及人工智能等基础科学技术的推动下,以卷积神经网络 CNN 和循环神

DOI:10.19452/ j.issn1007-5453.2019.S3.153
基金项目:航空科学基金(2018ZC31002)

经网络 RNN 为代表的深度神经网络算法,在图像处理、语音识别以及自动驾驶等多个民用领域均取得了长足的发展。例如,苹果发布的 iPhone XS 系列智能手机中搭载了其自主研发的双核架构神经网络处理引擎(Neural Engine),在机器学习和增强现实方面具有强大性能,每秒计算能力高达 6000 亿次[2];以 4∶1 总分战胜职业九段棋手李世石和 3∶0 总分战胜世界排名第一的围棋冠军柯洁的 AlphaGo,象征着计算机技术已进入人工智能的新信息技术时代。

然而,深度神经网络模型的算法规模、计算复杂度对硬件平台的计算、存储、数据传输能力等均提出了较高的要求。例如,Google 的图片识别模型 GoogleNet 约含 10 亿个连接,在 1000 台机器 16000 个核的计算集群上,使用 1000 万张图片作为训练集,需要三天才能完成训练任务[3]。在将深度神经网络算法应用于航空电子系统时,受限于机载环境计算能力和存储能力的约束,需要利用多种手段进行优化。

在航空领域,智能计算技术具有重要的研究价值和广阔的应用前景,首先其最有可能率先应用于 OODA 任务链的观察、调整与决策环节,提高智能化场景感知/目标感知和智能辅助决策能力。但是,航空电子系统对实时性、功耗、体积等方面有着更为严格的制约,使得设计面向航空电子系统的智能化计算平台面临较高的技术难度。因此,针对机载嵌入式环境对高实时性、高带宽、低功耗、高可靠性的应用需求,研究面向航空电子系统的国产化嵌入式智能计算系统对于我国实现相关技术的自主可控具有重要意义。

本文面向智能化航空电子系统的应用需求,综合考虑卷积神经网络 CNN 与循环神经网络 RNN 两类深度神经网络的计算加速技术。其中,CNN 面向目标识别与态势感知需求;RNN 面向辅助决策需求。本文对嵌入式智能计算加速技术展开研究,针对航空电子系统资源受限的约束,重点探索适用于机载嵌入式环境的智能加速芯片架构技术、深度神经网络模型压缩以及智能计算加速平台架构等关键技术问题,为构建智能化航空电子系统提供理论基础和技术支撑。

1 国内外研究现状

1.1 智能加速芯片的技术架构研究

深度神经网络对硬件平台的数据处理和存储能力有较高要求,目前以 CPU 为基础的通用处理架构难以高效地满足人工智能对硬件资源的应用需求,其并行度较差,且能耗通常较高,不适用于深度神经网络架构。

近年来,高性能计算硬件环境的快速发展为深度神经网络 DNN 的加速提供了基础,基于 GPU、ASIC、FPGA 及类脑芯片等不同硬件平台的加速架构被相继提出以提升 DNN 性能。各种智能计算硬件平台优缺点对比见表 1。

现有研究多采用 GPU 的并行计算能力进行加速,主要应用于算法模型的大规模数据训练,但是在服务器环境下的高性能 GPU 通常功耗较高。如特斯拉 Model S 上的 Autopilot 2.0 自动驾驶计算平台使用的处理器为 GPU TeslaP100,其功耗最高可达 250W[4];新一代 GPU 虽然功耗较低,如英伟达 Jetson TX2 功耗约为 7.5W,可用于嵌入式环境中,但该类 GPU 对软件运行环境和库的依赖性大,难以实现技术自主可控。

ASIC 是为某种特定需求而专门定制的芯片,其计算能力和计算效率都可以直接根据特定算法需要进行定制优化,具有体积小、功耗低和计算效率高等优势。目前国内涌现出以中科寒武纪“MLU100”、地平线“旭日 1.0”以及比特大陆“BM1880”等为代表的 ASIC 智能芯片,其中 MLU100 芯片在平衡模式(主频 1GHz)和高性能模式(主频 1.3GHz)下,峰值速度分别达到 128 万亿次定点运算/166.4 亿次定点运算,功耗分别为 80W/100W。但是 ASIC 相关技术还不成熟,缺乏统一的软硬件开发环境,目前仅有商业试用版应用,尚未考虑复杂恶劣的机载应用环境,短时间内难以应用于航空电子系统的嵌入式环境。此外,ASIC 芯片的开发成本非常高,根据 IBS 的估算数据,65nm 芯片开发费用需 2850 万美元,5nm 芯片开发费用则高达 54220 万美元。

类脑芯片在设计上推翻了经典的冯·诺依曼架构,而是基于仿人类脑神经系统的神经形态计算架构设计。清华大学类脑计算研究中心于 2015 年推出了首款类脑芯片——“天机芯”,将人工神经网络(ANNs)和

脉冲神经网络(SNNs)进行了异构融合,可用于图像处理、语音识别、目标跟踪等多种应用开发[5]。类脑芯片是在芯片基本结构甚至器件层面上进行优化设计,譬如采用忆阻器等新器件来提高存储密度。但是该类芯片相关技术还很不成熟,目前仅停留在实验室研究阶段,其算法模型的准确率仍有待提高,尚无法大规模应用,现阶段不适用于航空机载嵌入式环境。

根据赛灵思于 2018 年 5 月份公布的来源于 Barclays Research 的数据显示,从 2019 年开始,来自于“推理”的需求将会持续快速爆发式增长,“训练”的需求增长将会逐渐放缓,并趋于停滞,而“推理”正是 FPGA 的优势。FPGA 作为一种可编程的半定制芯片,对环境依赖性较低,可实现自主可控;并且 FPGA 在大幅提升能效、降低功耗的同时,还可以降低精度损失,且拥有出色的灵活性和低延时特性,能够满足机载嵌入式应用环境的要求。

表 1 智能加速芯片不同技术架构对比

技术架构	优点	缺点	应用场景
CPU	实现方式简单,网络结构易于更改	并行度较差,能耗通常较高	不适用于 DNN 加速
嵌入式 GPU	并行化程度高,性能好	依赖性大	推理阶段
高性能 GPU	能够并行处理大规模数据流	功耗高,依赖性大	训练阶段
ASIC	体积小,功耗低, 计算效率高	开发周期长,相关技术还不成熟,	推理阶段
类脑芯片	打破冯·诺依曼架构,发展前景好	相关技术还不成熟,尚无法大规模应用	推理阶段
FPGA	对环境依赖性低,可实现自主可控	硬件资源受限,设计复杂	推理阶段

1.2 深度神经网络模型加速技术研究

当前影响深度神经网络模型性能的主要因素包括算法迭代次数多且并行度低、计算复杂度高以及内存开销巨大这三个方面。为解决这三类问题,国内外研究机构主要研究方向有:

(1) 挖掘模型中的可并行部分,结合 FPGA 的特点设计相应的流水线/并行加速结构,增强算法并行性

深度学习算法一般需要进行多次迭代运算,如在 ImageNet 上进行训练时,GoogleNet 大约需要 100 万次迭代,每秒迭代需处理 32 张图片[6]。针对迭代次数多且并行度低的问题,现有研究大都聚焦于提高卷积计算的并行度上。华为的 Ascend 芯片采取了 ARM 核+AI 加速器的模式,其 AI 加速器采用了达芬奇架构,可把卷积计算所需的乘加器按照不同的计算组织成不同的形式,并搭配标准的数据缓存,针对云端应用的 Ascend 910 芯片能在 350W 的功耗上实现 256Tops 半精度浮点数算力或 512Tops 8 位整数算力。

(2) 探究深度神经网络模型压缩技术,减少模型参数量,从而提高计算资源的利用率,降低计算复杂度

常用的 CNN 网络模型在计算量上可达到 10 亿量级,参数量上更是达到了上百兆的量级。针对计算复杂度高的问题,常用的加速技术包括剪枝、数据量化等节省计算资源的模型压缩方法。Tien-Ju Yang 等提出了名为“NetAdapt”的自适应剪枝算法,可自动并逐步简化预训练网络,直到满足资源预算,对 ImageNet 数据集进行图像分类时,与其他最先进的网络简化算法相比,该算法以更高的准确度降低 1.66 倍的延迟[7]。

(3) 通过增加数据重用来减少访存次数,降低模型中中间特征数据的通信开销,达到智能计算加速的目的

巨大的内存开销导致的数据通信问题会严重影响深度神经网络训练和识别的计算效率,AlexNet 模型的参数导致的内存开销就已经达到了 218MB。为减少内存开销,常用的方法为提高数据重用率。Zhuoran Zhao 等提出了一种名为 DeepThings 的轻量级框架,通过融合分区(FTP)的方法,显著减少了内存占用,该方法在不牺牲准确度的情况下可节省至少 68%的内存占用开销,数据调度吞吐量提高 1.7~2.2 倍[8]。知存科技公司的“存算一体化”技术将深度神经网络映射到多个 Flash 阵列中,该 Flash 存储单元可以存储神经网络

的权重参数并完成此权重相关的乘加运算,能够将运算效率提高 10~50 倍[9]。

巨大的内存开销导致的数据通信问题会严重影响深度神经网络训练和识别的计算效率,AlexNet 模型的参数导致的内存开销就已经达到了 218MB。为减少内存开销,常用的方法为提高数据重用率。Zhang C 等设计的 DNN 硬件加速器可根据算法内存分配的特点,设计不同的存储单元存储不同类型的数据,并通过流水线结构提高计算单元利用率,该方法以低于通用处理器约 20 倍的功耗,将计算速度提升了约 117 倍[10]。

由于深度神经网络被广泛应用于解决复杂度越来越高的问题,网络模型的复杂度和可移植性也将直接影响人工智能在多领域中的应用。航空电子系统嵌入式环境具有资源紧张、功耗敏感的特性。深度神经网络如何部署于嵌入式环境中以提高智能化场景感知和决策能力、如何加速嵌入式智能计算技术以更好满足智能化航空电子系统嵌入式应用需求成为亟待解决的问题。

2 智能计算加速技术及架构研究

2.1 面向目标识别的卷积神经网络(CNN)

卷积神经网络(Convolution Neural Network,CNN)是一种前馈神经网络,在图像处理和图像识别领域具有突出表现。

CNN 基本结构如图 1 所示。CNN 模型中具有两种特殊神经元层:卷积层和池化层,可对输入图片进行特征提取。从结构上来看,CNN 模型主要具有局部感知、参数共享和池化的技术特点。

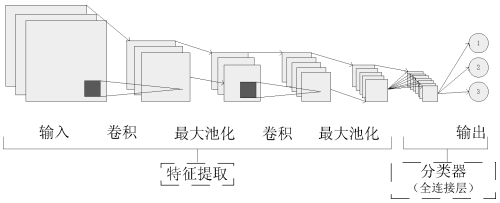

图 1 卷积神经网络基本结构

(1) 局部感知

机器在识别图像时,无须将整张图像按像素全部连接到神经网络中,可采用局部连接的模式,将图像分块连接,每个神经元只感知局部图像区域,局部图像块周边像素联系较为紧密,距离较远的像素联系较弱,然后在更高层将感知不同局部区域的神经元综合起来,得到全局信息。局部感知在不降低识别准确率的前提下,可显著减少模型中相邻层间的连接数。

(2) 参数共享

在同一图像中,从局部图像提取出的特征往往接近于其相邻局部图像中的特征。因此,可将 CNN 网络隐藏层中的每一个神经元连接的局部图像的权值参数,都共享给其它神经元使用,从而大大减少了网络参数量。

(3) 池化

随着网络模型的不断加深和卷积核数量的不断增多,直接用卷积核提取的特征进行训练容易导致过拟合现象。池化操作即在不同位置区域提取出代表性特征,采用取最大值或平均值等方法对卷积提取出的特征进行降维处理,防止过拟合问题的产生。

CNN 自 20 世纪 80 年代提出后,根据应用场景及需求的不同,衍生出多种卷积神经网络的变种模型。较为典型的 CNN 模型有:

(1) AlexNet 网络模型,包括 5 个卷积层、三个全连接层和一个 softmax 层,其中第一层卷积层共有 105705600 个参数,整个网络的权重数为 61M。该模型特点在于使用了非线性激活函数 ReLU、采用了防止

过拟合的方法 Dropout 并且利用多 GPU 进行并行训练。

(2) VGGNet 网络模型,构筑了 16~19 层深的卷积神经网络,证明了增加网络深度能够在一定程度上影响网络最终的性能,大大降低了错误率。目前,该模型仍被广泛用于提取图像特征。

(3) GoogleNet 网络模型,共有 22 层,其最大特点在于采用了 inception 结构,能充分利用网络中的计算资源,在不增加计算负载的情况下,增加网络的宽度和深度,从而提升网络性能。

(4) ResNet 网络模型。ResNet 建立了 34 层、50 层、152 层、1202 层等不同深度的模型,试验发现随网络深度的增加,网络准确度会出现饱和甚至下降的趋势。为解决退化问题,ResNet 网络引入了残差单元,通过给非线性卷积层增加直连边,类似电路"短路"的方法,优化网络性能。残差学习是训练较深网络时的一种重要方法。

(5) R-CNN 网络模型,是利用深度学习进行目标检测的开山之作。该模型预先提取 1000~2000 个可能含检测物体的候选区域,而后对每个候选区域内的图像进行特征提取,再利用分类器进行目标分类,成功将 PASCAL VOC 上的准确率从 35.1%提升到 53.7%。其 2015 年提出的改进型 Fast R-CNN 训练时间由 84h 减少至 9.5h,准确率约 66%左右;Faster R-CNN 在复杂网络检测速度达 5fps,准确率为 78.8%[11]。

2.2 面向智能决策的循环神经网络(RNN)

循环神经网络(Recurrent Neural Network,RNN)常用来分析时间序列数据,例如,以一段文字或语音作为输入,RNN 可以进行语音识别、语义/情感分析等。目前,应用较多且性能较好的两类 RNN 是长短期记忆模型(Long Short-Term Memory,LSTM)和门控循环单元模型(Gated Recurrent unit,GRU)。GRU 模型较为简单,网络参数少,能更快收敛。而面对足够计算力和庞大数据集时,LSTM 模型具有更好的性能。单个 LSTM 模型单元具体结构如图 2 所示。

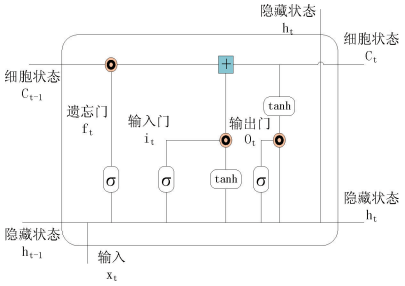

图 2 单个长短期记忆模型单元的具体结构

LSTM 模型的门控结构包括遗忘门(forget gate,f_t)、输入门(input gate,i_t)和输出门(output gate,O_t)三种结构,还有一个特殊的隐藏状态,通常称为细胞状态(Cell State,C_t)。图 2 中输入门 i_t 用于处理当前序列的输入位置,遗忘门 f_t 控制上一层的隐藏细胞状态 C_{t-1}是否被遗忘,这两种门控结构均作用于上一层的隐藏细胞状态 C_{t-1},并产生当前层隐藏细胞状态 C_t,由输出门 O_t 输出。

GRU 模型与 LSTM 模型均通过门控结构来解决标准 RNN 的梯度消失问题,不同之处在于 GRU 将遗忘门和输入门合并为一个独立的更新门(update gate),并将细胞隐藏状态融入其中。图 3 中 z_t 为更新门,可帮助模型决定忽视还是继续传递当前时间步的信息 x_t。r_t 为重置门,用于控制前一时间步隐藏单元 h_{t-1}对 x_t 的影响,若该隐藏单元信息不重要,则 r_t 门打开,从当前信息 x_t 开始表述新的意思而不受 h_{t-1}的影响。GRU 模型通过设计更新门和重置门来控制梯度信息的传播,以此缓解梯度消失的现象。

2.3 嵌入式智能计算加速架构

深度神经网络被成功应用于多种领域中,通过增加网络层数,深度神经网络能够捕获目标更深层次特性,通常可获得更高的准确度,但是其存储占用、内存带宽和计算开销随着网络深度增加呈指数级增长,难以

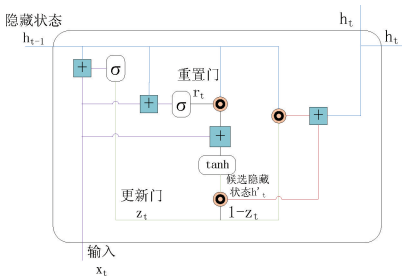

图 3 单个门控循环单元的具体结构

部署于航空电子系统等嵌入式环境中。为了减小计算开销和参数数量以构建低功耗和低存储占用的系统,众多研究集中于研制更好的神经网络架构。设计一个高效的嵌入式深度神经网络架构主要有以下两种途径:

(1) 使用高效的网络架构,优化其内部的操作开销

在优化内部操作开销时,主要采用低维卷积滤波器或稀疏系数等方法来减少内部操作计算复杂度和参数数量,以达到计算加速的目的。如寒武纪提出的 Cambricon-X 架构,利用了稀疏系数的计算架构,在不影响计算精度的前提下,将原网络模型中的许多权重参数去掉,最多可去掉 90%以上。Cambricon-X 加速架构如图 4 所示。

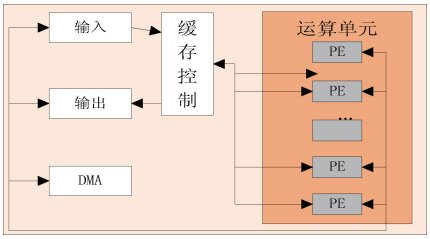

图 4 Cambricon-X 加速架构

图 4 中,缓存控制模块实现将系数为 0 的权重引起的输入数据去掉,具体实现流程为:从输入神经元中筛选出非零权重参数所对应的数据,将该数据按顺序排列好,传输给数据处理单元(Processing Element, PE),由 PE 执行乘加操作。即利用稀疏系数达到加速效果。

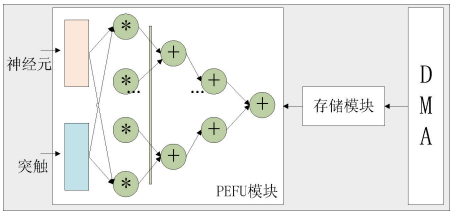

图 5 Cambricon-X 中的 PE 单元

图 5 为 Cambricon-X 架构中的 PE 结构,图中存储模块用于存放有效的权重参数,单个神经元的所有不为 0 的权重参数存储在一起。

在 Cambricon-X 架构中,仅有一列 PE 计算单元,当单个 PE 提前计算完成后,需等待其他 PE 完成计算,才能继续进行下一组输入数据的运算,计算时长将由计算最慢的 PE 决定。在 65nm 工艺下,Cambricon-X 芯片峰值性能可达 0.5Tops/s,性能是高端 GPU 的 10 倍,能耗仅为其 3.4%[12]。

随着 DNN 模型规模的扩大和计算量的剧增,稀疏系数的优势会趋于明显,研究稀疏性矩阵对算力紧缺的嵌入式环境来说是很有必要的。

(2) 使用低精度操作来设计网络,使网络可用硬件高效实现

目前,深度神经网络通常采用浮点计算,对存储空间和计算力的需求颇高。低精度操作便是在尽量不影响准确度的前提下,通过二值化操作等降低部分精度的方法,借助 FPGA 等硬件架构的并行特性,实现智能计算加速。清华大学研发的 Thinker Ⅱ芯片便是应用了二值化操作的典型芯片[13],其加速架构如图 6 所示。

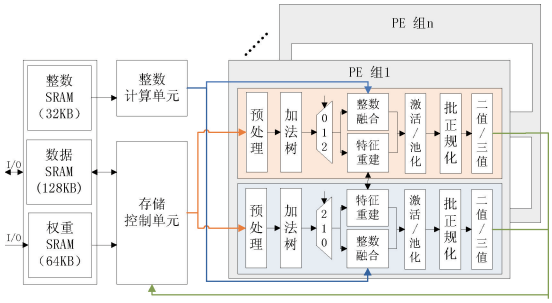

图 6 Thinker Ⅱ架构

Thinker Ⅱ芯片对神经网络低位宽量化方法、计算架构和电路实现进行了系统研究,设计并实现了从浮点转换为二值/三值化的卷积优化计算,大大降低了计算复杂度和功耗。

二值化操作是将权重参数和隐藏层激活值二值化为 0 或 1,使得模型中参数占用的存储空间大大减少(如从 32float 转化为 1bit,理论上内存消耗减少为原来的 1/32 倍)。通过二值/三值化卷积优化算法,Thinker Ⅱ架构可支持低位宽网络高能效计算,大幅降低了算法复杂度。该芯片应用于人脸识别时,能够实现 6ms 人脸识别,准确率超过 98%,运行在 200MHz 时,功耗仅为 10mW。

综上所述,设计面向机载嵌入式环境的智能计算加速平台时,需综合考虑嵌入式环境的应用需求及 CNN、RNN 网络模型的计算特性。由于 CNN 卷积层会以少量的数据占据大量的计算资源,为提高计算效率,可采用低维滤波器或稀疏系数来减少卷积运算的参数量,同时可利用二值/三值化优化计算,减少参数所需的存储空间。

与 CNN 相比,RNN 模型的隐藏层数较少,但数据流向为双向,其输入接收单元需同时接收多个输入矢量,并将这些矢量分配给多个数据处理单元。RNN 中的卷积运算加速可仿照 CNN 加速机制,充分利用 FPGA 的并行特性,达到智能计算加速的目的。

针对下一代智能化航空电子系统的应用需求,设计嵌入式智能计算加速架构。在将深度神经网络部署于机载嵌入式环境时,一方面可优化网络模型,利用多核处理器对深度神经网络模型算法中的高维数据计算进行并行加速,满足航空电子系统处理大规模并行化乘加运算的需求;另一方面可优化存储结构,提升深度神经网络在机载嵌入式环境下的运行效率。

3 机载智能计算加速平台开发流程

机载智能计算技术对下一代智能化航空电子系统的发展起到关键性作用,在智能任务规划、辅助作战人员快速完成信息的处理决策、提升任务的执行效率等方面均可提供技术支撑。

面向智能化航空电子系统的应用需要,需要开发嵌入式智能计算加速平台,用于增强航空电子系统在目标识别和任务规划等方面的智能化水平。面向航空电子系统的嵌入式智能计算加速平台总体开发流程如图 7 所示。

具体来看,开发流程主要分为以下 4 个步骤:

图 7 嵌入式智能计算加速平台开发流程

(1) 针对智能化航空电子系统对目标识别/态势感知及辅助决策的应用需求,对各类 CNN 与 RNN 的运算流程和计算特性进行分析,选取合适的 CNN 与 RNN,建立相应的目标数据集,对选取的算法模型进行训练。

(2) 针对智能化航空电子系统嵌入式环境的计算性能约束,利用模型剪枝、数据量化、Huffman 编码等轻量化技术以及二值/三值化方法来减少算法模型参数量,对算法模型进行优化设计,降低计算复杂度,提高计算资源的利用率。

(3) 针对智能化航空电子系统嵌入式环境的存储性能约束,分析算法的权重参数,优化存储结构并挖掘算法中可并行部分,通过增加数据重用率来减少访存次数,最大化高速存储单元的利用率,降低模型隐藏层特征数据的通信开销。

(4) 针对智能化航空电子系统的架构特性,设计嵌入式智能计算加速平台硬件架构,根据 FPGA 高性能并行计算和低功耗特性,设计流水线结构,简化卷积运算模块结构,并采用片上网络 NoC 实现并行化的 CNN 运算单元阵列和 RNN 运算单元阵列。

片上网络是一种面向多核处理芯片的轻量化分组交换网络,能够提供标准的、可扩展的硬件框架,可以加速模块间的数据交互,满足人工智能的高通信需求,现已被广泛应用于嵌入式智能计算系统的多核处理芯片中。

4 结束语

智能计算加速技术对我国发展自主可控的下一代智能化航空电子系统具有重要意义。该技术可应用于机载嵌入式环境,满足机载嵌入式环境高性能、低功耗的需求;可应用于 OODA 环的观察与决策环节,提高数据分析智能化水平;可解决资源受限环境下深度神经网络的部署与优化问题,大大降低计算复杂度与存储需求。本文探索了应用于机载嵌入式环境的深度神经网络架构技术、CNN/RNN 模型压缩技术以及智能计算加速架构开发流程等关键技术,后续将进一步实现可重构的机载嵌入式智能计算加速平台及综合验证环境,为深度神经网络应用于机载智能计算与信息处理平台提供理论基础和技术支撑。

参考文献

[1] 鲁俊,何锋,熊华钢.航空电子云系统架构与网络[J]. 航空电子技术, 2017, 48(3):1-9.

[2] 冯光顺,应三丛.ZYNQ 的卷积神经网络硬件加速通用平台设计[J].单片机与嵌入式系统应用,2019,19(03):3-6.

[3] 杨天祺,黄双喜.改进卷积神经网络在分类与推荐中的实例应用[J].计算机应用研究,2018,35(04):974-977.
[4] Sze V, Chen Y H, Yang T J, et al. Efficient processing of deep neural networks: A tutorial and survey[J]. Proceedings of the IEEE, 2017, 105(12):2295-2329.
[5] Shi L, Pei J, Deng N, et al. Development of a neuromorphic computing system[C]// Electron Devices Meeting: IEEE, 2016.
[6] Cheng Jian, Wang Peisong. Recent advances in efficient computation of deep convolutional neural networks[J].Frontiers of Information Technology & Electronic Engineering, 2018,19(01):64-77.
[7] Lee H, Battle A, Raina R, et al. Ng. Efficient sparse coding algorithms[Z]. In Nips, 2007.
[8] Zhao Z, Barijough K M, Gerstlauer A. DeepThings: Distributed adaptive deep learning inference on resource-constrained iot edge clusters[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2018, 37(11):2348-2359.
[9] 肖皓,祝永新,汪宁, 等.面向卷积神经网络的 FPGA 硬件加速器设计[J].工业控制计算机,2018,31(6):99-101.
[10] Zhang C, Li P, Sun G, et al. Optimizing FPGA-based accelerator design for deep convolutional neural networks[C]// Acm/sigda International Symposium on Field-programmable Gate Arrays, 2015.
[11] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2017,39(6):1137-1149.
[12] Zhang S, Du Z, Lei Z, et al. Cambricon-X: An accelerator for sparse neural networks[C]// IEEE/ACM International Symposium on Microarchitecture, 2016.
[13] 张洁玉. 基于图像分块的局部阈值二值化方法[J].计算机应用,2017,37(03):827-831.

作者简介

李欣瑶(1994-)女,硕士,助理工程师。主要研究方向:深度神经网络硬件加速。
Tel: 028-89186497
E-mail: 747065580@ qq.com