近似计算新范式在深度学习加速系统中的应用及研究进展

龚 宇　　王丽萍　　王 佑　　刘伟强*

(南京航空航天大学电子信息工程学院/集成电路学院　南京　211106)

(空天集成电路与微系统工信部重点实验室　南京　211106)

摘　要：深度学习已经成为当前人工智能技术中最为重要的算法之一。随着应用场景不断扩展，深度学习硬件规模越来越大，计算复杂度呈现数量级提升趋势，对加速系统提出了极高能效需求。后摩尔时代，新型计算范式逐渐取代工艺微缩成为提升能效的有效方案，近似计算以牺牲部分精度的代价换取大幅能效提升，成为最具前景的设计方法之一。该文以深度学习加速系统的不同设计层次为切入，首先介绍了深度学习网络模型的算法特征，围绕算法层的近似计算方案介绍了量化方法的研究进展；其次，围绕硬件架构和电路层调研了当前深度学习加速在图像、语音等多个方向采用的近似电路和架构方案，围绕层次化的设计方法调研了当前近似计算的系统设计方法学及EDA领域的关键问题和研究进展；最后，对该领域方向进行展望，旨在推动近似计算新范式在深度学习加速系统中的应用。

关键词：近似计算；计算范式；深度学习；加速系统；研究进展

Application and Research Progress of Approximate Computing as a New Computing Paradigm in AI Acceleration Systems

GONG Yu　　WANG Liping　　WANG You　　LIU Weiqiang

(*College of Electronic and Information Engineering/College of Integrated Circuits, Nanjing University of Aeronautics and Astronautics, Nanjing* 211106, *China*)

(*Key Laboratory of Aerospace Integrated Circuits and Microsystem, Ministry of Industry and Information Technology, Nanjing* 211106, *China*)

Abstract: Deep learning has emerged as one of the most important algorithms in artificial intelligence. With the increasing application scenarios, the hardware scales for deep learning are becoming larger, and the computational complexity has considerably increased, leading to a high demand for energy efficiency in accelerating systems. In the post-Moore's Law era, new computing paradigms are gradually replacing process scaling as an effective solution for improving energy efficiency. One of the most promising design paradigms is approximate computing, which sacrifices some precision to improve energy efficiency. This research focuses on different design layers of deep learning acceleration systems. First, the algorithm characteristics of deep learning network models are introduced, and the research progress on quantization methods is presented in view of the approximate computing scheme at the algorithm layer. Second, approximate circuits and architectures employed in various directions such as image and speech recognition in the circuit-architecture layer are surveyed. Furthermore, the current hierarchical design methods for approximate computing as well as critical issues and research progress in Electronic Design Automation (EDA) are investigated. Finally, the future direction of this field is anticipated to promote the application of a new paradigm of approximate computing in deep learning acceleration systems.

Key words: Approximate computing; Computing paradigm; Deep learning; Acceleration system; Research progress

1　引言

当前，深度学习(Deep Learning, DL)，亦称人工神经网络(Artificial Neural Network, ANN)，因其在语音/图像/视频分类等多个模式识别领域的成功，已经成为人工智能(Artificial Intelligence, AI)领域重要发展方向。2012年AlexNet[1]以高识别精度获得ImageNet冠军，开启了神经网络又一轮研究高潮。而随着算法的不断演进发展，ANN网络规模从兆字节(MB)已经发展到吉字节(GB)[2−5]，海量参数使其计算量巨大，导致传统硬件难以满足网络高能效加速需求[6−8]，迫切需要更大算力、更高能效的加速架构。

后摩尔时代，先进工艺制造价格昂贵、门槛高，使得神经网络加速架构能效的优化逐渐从工艺依赖向新型计算范式转型。当前主流新型计算范式有存内/近存计算[9,10]、随机计算[11−13]以及近似计算[14,15]等。存内/近存计算芯片仍面临着生产良率低、工艺难度大的问题，适用于计算位宽和精度较固定的场景；随机计算在高精度处理时硬件复杂度指数增长，且高昂随机性需求使其设计难度大；近似计算对工艺无额外需求且扩展性更强，相比之下更为经济、易于实现。文献[16]对当前模式识别、计算机视觉处理等12类重要应用场景进行了归纳，其中近83%的运算为可容错计算，对其而言，算得"足够好"比算得"非常准"更重要。可见，传统精确计算非当下系统必选项，在很多应用中算得快更为重要，而近似计算正逐渐广泛采用。

当前，国内外学者对于近似计算在深度学习加速系统中的研究集中在近似计算的单元结构设计、系统结构设计、系统设计方法等方向，据公开论文显示，国内主要研究机构集中在部分高校与科研院所，包括南京航空航天大学[17,18]、东南大学[19−21]、电子科技大学[22,23]、上海交通大学[24−26]、清华大学[27,28]、中国科学院计算所[29−31]等。在先进工艺演进代价越来越高、架构设计相对固定化的情况下，近似计算开始广泛应用于深度学习加速系统的多个设计层次中，以进一步提升系统能效。

本文结合目前深度学习加速系统设计的不同层次，从算法层、电路与架构层、系统层对近几年深度学习加速系统中的近似计算方案及设计方法进行了具体介绍，并为后续研究提供思路。

2　近似计算在深度学习算法中的应用

当前，深度学习在模式识别、自动控制及预测估计等领域获得了广泛应用，但是由于模型尺寸越来越大，其硬件实现更加复杂。为了降低深度学习模型的冗余度，近似计算在其模型简化中得到了深入的应用。本节首先介绍典型的3类计算模型，并进一步对其近似量化方案进行介绍。

2.1　典型神经网络模型

2.1.1　卷积神经网络

卷积神经网络(Convolutional Neural Network, CNN)是一种带有卷积结构的前馈神经网络，被广泛应用于具有网格结构的数据，如图像、音频及视频等目标识别领域。卷积层由2维/3维结构滤波器组成，1个滤波器可以看作由多个卷积核堆叠形成。这些滤波器在输入数据上滑动做卷积运算，从输入数据中提取特征。在多层神经网络中，上层节点的输出和下层节点的输入之间通过激活函数建立联系。

2.1.2　循环神经网络

循环神经网络(Recurrent Neural Networks, RNN)是专用于处理序列信息的神经网络，常用于处理序列数据，如长语音识别及长句语义分析等。相比于前馈神经网络，RNN的隐藏层自身也存在一个回路，以允许信息传递到下一步，因此RNN可以接收上一时间点的隐藏状态，由此具有"记忆"。

2.1.3　全连接神经网络

全连接神经网络(Fully Connected Neural Network, FCNN)是一种多层感知机。CNN, RNN, LSTM等深度学习网络都是基于全连接神经网络设计的。在全连接神经网络中，每一层网络的任一个节点都和下一层所有节点连接。其特征提取更为充分，可作用于大多数场景。但因其权重过多，计算量过大导致其单独应用较少，往往作为深度学习网络的最后几层，用于提取和输出分类结果。

2.2　深度学习模型的量化

在深度学习的硬件加速实现高能效方面，量化(Quantization)是最为有效和广泛使用的近似计算方法之一。量化可以大幅减少数据位宽，进而在硬件设计中可以使用更窄的线宽、更少位数的计算单元。

训练后量化算法即在完成神经网络训练后进行量化的算法。文献[32]和文献[33]分别提出了对称量化和(最小/最大)非对称量化算法实现高精度识别。文献[34]提出通过限制激活值范围减少截断误差、通过设置偏差校正修复量化误差，无需重训练保持识别高精度。文献[35]对网络进行逐层量化，以迭代方式动态调整激活值和权值量化步长。文献[36]通过评估不同位宽和数字格式，实现网络压缩率和精度平衡动态量化。

训练中量化算法即将量化过程融入到模型的训练流程中，前向传递时基于量化模型进行推理，而反向训练时仍采用浮点进行[37]。文献[38]提出了一种基于量化边界的训练算法，在每次训练中将小于阈值的数值进行量化操作，在整个网络模型精度损失可控的情况下实现极低位宽的权值和激活值量化。文献[39]表明，权值和激活值小于2 bit时，仍可以实现高精度的神经网络。文献[40]提出名为DoReFa的训练方法，对除第1层和最后一层外的其他所有隐藏层使用极低的数据位宽，同时在反向传播过程中对梯度也进行量化。

二值化/三值化量化：作为更激进的量化方法，极低功耗场景下出现了采用{0, 1}或{+1, −1}表达的神经网络。文献[41]以1 bit权值和激活值的二值神经网络实现了高精度的识别，能效提升超过25倍。文献[42]将CNN权重量化为1 bit，实现了极低计算功耗，同时计算精度损失在2%以内。文献[43]面向语音识别提出了将权重量化为1 bit位宽、数据8 bit位宽的压缩方式，相比16 bit位宽精度损失3%以内。

可以看出，网络量化在算法层面可大幅提高计算速度，对硬件设计有助于减少计算单元的能量、面积开销，减少片上数据传输带宽压力，为设计高能效深度学习加速系统提供了有力支撑。

3 近似计算在深度学习加速电路与架构的应用

近年来神经网络加速系统已经在FPGA, ASIC, GPU等计算架构中有了大量的实现案例，随着多位宽、高能效近似计算应用逐渐广泛，以新型计算范式优化架构成为后摩尔时代提升芯片能效主流方案之一。

3.1 近似加法器与近似乘法器

由于神经网络的规模越来越大，如何优化基础单元的功耗和面积，如加法器、乘法器，以及核心模块，如乘累加单元，成为提升神经网络加速架构能量以及面积效率的关键因素之一。采用合理的近似计算单元，可以在控制系统精度损失的同时，获取较大的能效、面积提升收益。

3.1.1 近似加法器

一般将加法器分为低延迟加法器(Low Latency Approximate Adder, LLAA)和低功耗加法器(Low Power Approximate Adder, LPAA)。LLAA一般是由多个异构的子加法器构成的并行结构，针对实时性和性能要求高的计算场景，常用的有通用精度可配置近似加法器(Generic Accuracy Configurable Adder, GeAr)[44]、面积-质量最佳低延迟近似加法

器(Quality-area optimal Low-Latency approximate Adder, QuAd)[45]、容错加法器Ⅰ(Error-Tolerant Adder I, ETA-I)[46]、容错加法器Ⅱ(Error-Tolerant Adder II, ETA-II)[47]、近似镜像加法器(Approximate Mirror Adder, AMA)[48]等等；而LPAA一般则为链式的串行结构，面向功耗极为敏感的场景，如行波进位加法器(Ripple Carry Adder, RCA)、进位保存加法器(Carry Save Adder, CSA)[49,50]、超前进位加法器(Carry-Lookahead Adder, CLA)[51,52]等。

面向深度学习加速，尤其是其中对累加结构的大量应用，出现了大量近似加法结构。或门近似加法器通常用于提高硬件加速器的能效[53−55]。文献[54]将或门近似加法器替代低比特位的全加器，精确计算部分使用与门。文献[48]使用近似镜像加法器替代低位或门加法器，使用更多的晶体管实现系统识别精度提高。文献[56]提出了一种精度自适应近似加法器，其系统在不同信噪比下可根据配置信号使用不同近似位宽。

3.1.2 近似乘法器

传统的标准乘法器往往延迟高、功耗高，难以满足深度学习加速系统的低功耗、高能效要求。而近似乘法器具备了低功耗和低延迟的特点，根据近似产生的阶段不同，一般可分为在生成部分积时近似[57−59]、对部分积求和时近似[54,60,61]和近似对数乘法器[62,63]，以下将分别进行介绍。

(1) 生成部分积时近似：文献[57]介绍了两种近似基4 Booth编码器的设计，提出了近似 Booth乘法器。文献[58]通过简化逻辑真值表提出了3种不同的简化方案。与标准乘法器相比，在功耗、延迟和面积方面均有显著降低。文献[59]提出了一种基于近似混合编码的基8 Booth编码近似乘法器。

(2) 对部分积求和时近似：文献[62]提出了一种低功耗乘法器，其中在部分积的求和数组中引入了垂直截断，使乘法器的总功耗降低了58%，但输出精度略有下降。文献[54]引入了水平截断，大大减小了电路面积和功耗。文献[60] 提出通过3个近似累加结构减少部分乘积的行数，但其电路实现相对比较复杂。

(3) 近似对数乘法器：为了提高对数近似乘法器的精度，文献[64]提出了使用截断二进制对数转换器和用于加法的集一加法器的近似对数乘法器。文献[65]中两个连续2次幂之间的输入操作数被分成多段，然后每段分析确定一个误差减少因子，并补偿到基础线性模型的结果。文献[66]通过随机删除精确设计中连接生成近似乘法电路，将多目标笛卡儿遗传规划算法应用于设计空间探索，获得关于MRED的最佳近似电路。

3.2 混合精度的深度学习近似加速系统

基于近似计算的混合精度神经网络处理器已经有了大量的工作发表，对精度的控制方式也出现了多种机制和方案。qcNPE[67]是一款基于近似计算的神经网络加速架构，通过在计算模块中增加精度控制电路进行实时精度重构，并和神经网络的结构相映射以实现高能效。Laika结构[68]在65 nm CMOS工艺下实现了5 μW低功耗永远在线的关键词语音识别芯片，采用了4 bit和8 bit动态切换的近似计算结构，可动态适应网络加速中不同步骤的不同位宽需求。Price等人[69]发表了基于计算宽度可变单元、存储结构可重配置的神经网络芯片，采用动态宽度控制实现不同网络层及不同负载下的资源重配置。

清华大学团队提出的Thinker系列集成了多种混合精度神经网络加速架构[70,71]。Thinker架构[70]中的计算单元可动态配置为不同位宽，重构后加速等不同网络层计算，Thinker-II[71]设计了基于多位宽、可配置加法树实现的位宽可变累加，这一结构可支持1～16 bit不同位宽的累加运算。中国科学院计算所团队设计了多款支持多位宽的智能计算处理器[72,73]，实现了多种位宽下高能效加速和计算精度控制。三星公司在其先进SoC设计中集成了神经网络引擎[74]，采用INT4/INT8/FP16混合的近似乘法器实现多精度MAC单元设计，实现了极高的能量、面积效率表现。南京航空航天大学和东南大学联合研发的面向语音识别的超低功耗近似计算芯片[43]，采用了多种近似加法器、近似乘法器结构，实现了高能效、高精度的关键词语音识别。

4 深度学习中的近似计算协同设计

传统IC设计需要在整个设计流程中尽可能保持数值和变量的等价，即要求前一设计层次的计算结果和后一设计层次的计算结果是相同的，以保证最初的应用开发和最终的芯片输出结果一致。对于近似计算而言，其设计原则是只要计算精度满足系统需求，在IC设计的每个流程中都可以适当引入近似计算，如图1所示。近似计算的核心和关键在于，如何在计算精度和硬件能耗间获得一个较好的均衡，在能效提升同时，要确保最终的执行结果偏差在容忍范围内。近似计算以精度换取系统复杂度降低，但随着近似计算单元设计的不断演进和发展，需要跨越算法、电路、架构多个层次进行系统设计。如何建立跨层设计与系统精度的关联规律以完成近似系统精度和系统复杂度之间的设计取舍、如何协同计算-存储模块的设计以面向具体应用场景实现系统级高能效设计，成为近似系统能效提升的关键性问题。为了解决上述问题，针对近似计算的
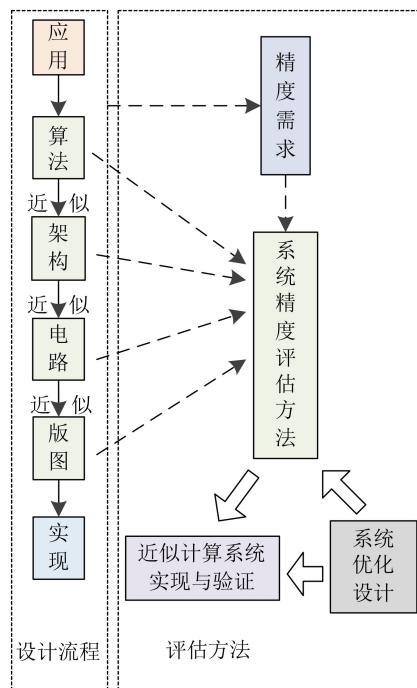

图 1 近似计算设计流程与评估方法

精度建模、精度评估手段不断出现，以提供快速部署的实现方案指导。

4.1 单元误差分析与系统误差评估

4.1.1 近似加法器的误差分析

Mazahir等人[75]提出了针对一系列用于近似加法器均衡设计的概率误差模型，以评估LLAA近似加法器的计算精度。这一模型首先将近似加法器按比特拆分，基于进位出错概率建立子近似加法器精度模型，继而推导出整体误差模型。这一模型可以快速、精确地评估多类基于分块结构的近似加法器。Hanif等人[45]提出了针对低延迟近似加法器精度和复杂度均衡评估的QuAd模型。作为面向设计精度、面积以及延迟均衡的解析表达精度模型，QuAd模型将近似加法器的不同设计参量作为模型变量，可将模型用于LLAA设计空间探索，大大加速了设计初期的参数设置及评估。针对LPAA近似加法器，Ayub等人[76]提出了精度和复杂度均衡的评估方法，通过只考虑正确结果概率进行了模型的压缩和解析简化，大大加速了评估效率。

4.1.2 近似乘法器的误差分析

近似乘法器的精度模型即可看作近似加法器模型的扩展。Zhu等人[77]提出了多分布概率误差模型和评估方法，可高精度预估近似Booth乘法器精度。该工作将近似Booth乘法器的概率误差模型分为编码模块、部分积阵列模块及压缩模块错误概率模型，3个子模型逐层关联，构成最终乘法器精度评估方法。Liu等人[78]提出面向近似计算累加器的

精度评估方案，将累加模块作为整体划分计算流水，抽象为多计算阶段的加法器，最终可以在支持向量机算法上实现接近92%的评估精度以及1.5倍的性能提升。此外，面向神经网络算法中最为重要的乘累加和累加模块，当前基于PMF[77]的分析可以在低比特位宽时实现较高的精度预测。

4.1.3 近似计算的系统误差评估

MACACO模型[79]是较为系统的行为级精度评估方法，这一模型基于时序无关的等价电路结构，创新采用了虚拟误差电路，在特定输入数据或数据集下通过布朗分析、SAT分析、BDD分析等多种方式结合，以蒙卡数据拟合模型，实现了行为级精度评估。Venkataramani等人[67]提出了AxNN框架，专用于神经网络加速系统在近似计算系统中加速时精度、能效均衡设计的建模评估。AxNN在设计阶段增加精度可控的计算单元进行精度调节，提高2.3倍系统性能时仅损失7.5%精度。Liu等人[80]设计了CONCRETE框架用于"算法-架构-电路"逐层地进行神经网络在近似计算系统中的均衡设计及评估，神经网络的结构参数、近似计算单元的位宽参数都纳入精度影响参量，可以评估最终精度，以此优化神经网络设计。Freno等人[81]提出了一套基于机器学习算法的近似计算误差统计模型，通过引入回归分析、随机噪声模型等，参量化并进行定向提取评估近似计算精度，提供优化近似计算系统设计的具体思路。文献[43]面向权重二值化神经网络加速系统，通过近似噪声的概念定义提出了一种系统级误差分析方法，可量化评估算法层、架构层、电路层等不同层次误差的扩散和影响，最终将评估结果用于近似计算的架构设计参数优化。

4.2 深度学习近似加速系统中的设计空间探索

除通用近似计算系统的评估建模，针对神经网络加速架构的设计空间探索也有大量的研究工作展开。面向神经网络的设计空间探索方法主要分为几类：一是朝着更复杂的结构进行探索，如引入近似计算[82-86]、引入新型电路结构[87]、增加存内计算电路[88]；二是面向特定的应用范围，简化设计空间，从而提高设计空间探索的效率[83,84,89]；三是对设计空间探索的收敛速度、优化算法进行研究[90-92]。面向具有海量计算资源的神经网络加速架构，要对多种不同设计方法进行融合，才能快速、准确、有效地实现设计收敛。

5 思考与展望

当前，在深度学习加速系统中近似计算已经有了广泛的研究和应用，但是其更大规模的应用和部署仍存在较为突出的一些挑战，主要集中在近似

计算系统设计中的误差分析与管理、跨层设计与协同优化、近似系统的验证与测试、可靠性与安全性问题。

5.1 误差分析与管理

近似计算以误差为代价换取能效提升，但是系统误差仍是最重要的系统指标之一。由于误差分析必须要考虑具体应用场景，因此一般性的误差分析方法往往集中在小模块上，在系统层仍然缺少普适性的误差分析方法。同时，由于近似计算的设计方案多样，存在确定性误差和非确定性误差两大类。确定性误差可以较容易地确定其边界，而非确定性误差的分析、评估与控制更具有挑战性。

5.2 跨层设计与协同优化

当前，面向近似计算的设计优化多集中在单一的设计层次或设计阶段。而由于误差在系统中是不断传播的，因此，结合多层次的近似计算技术以进行软硬件协同设计，可以更为充分地挖掘近似计算的能效潜力。近似计算在算法层的优化方案如何与电路、架构层的设计有机结合、互相优化，仍需深入研究；同时，将近似计算硬件误差在算法设计中加以考虑、修正，也是提升近似计算系统精度、能效的有效方法。

5.3 近似系统的验证与测试

现有的EDA工具对近似计算系统的验证和测试支持有限，使得近似计算系统的误差范围难以控制在可接受范围之内。充分考虑近似计算的误差来源，重新审视测试向量的生成方法，并对影响误差程度更大的关键电路使用更合适的测试方式，是大幅提升近似系统验证测试效率的重要工具开发方向。

5.4 可靠性与安全性

在面对老化问题时，需要将近似计算引入的额外计算误差纳入系统设计考虑之中。与精确计算不同，近似计算利用了算法、电路、架构的特征，因此电路的可靠性存在着新的设计挑战。此外，安全性也是近似计算面临的重要挑战。由于设计本身引入了计算误差，如何区分硬件/软件木马所引入的误差成为系统的威胁之一。如果将这一技术应用于视频监控、安防、智能电网等业务领域，就必须充分考虑安全防护设计。

6 结束语

本文主要对近似计算在深度学习加速系统不同设计层次的应用进行了总结，包括算法层、电路与架构层和系统层。与传统计算范式相比，近似计算可以在同工艺节点下大幅提高能效，是后摩尔时代最具备潜力的研究方向之一。同时，从误差分析与管理、跨层设计与协同优化、近似系统的验证与测

试、可靠性与安全性4个方面进行了思考与展望，以供学者在后续研究中作参考借鉴。另外，近似计算的设计流程目前仍不完整，难以直接产业应用，仍是其成为主流计算范式所面临的重大挑战和重要研究方向。

参 考 文 献

[1] KRIZHEVSKY A, SUTSKEVER I, and HINTON G E. ImageNet classification with deep convolutional neural networks[J]. *Communications of the ACM*, 2017, 60(6): 84–90. doi: 10.1145/3065386.

[2] SIMONYAN K and ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]. Proceedings of the 3rd International Conference on Learning Representations, San Diego, USA, 2015.

[3] LIN Zichuan, LI Junyou, SHI Jianing, *et al.* JueWu-MC: Playing minecraft with sample-efficient hierarchical reinforcement learning[C]. Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, Vienna, Austria, 2022.

[4] LI Zewen, LIU Fan, YANG Wenjie, *et al.* A survey of convolutional neural networks: Analysis, applications, and prospects[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2022, 33(12): 6999–7019. doi: 10.1109/TNNLS.2021.3084827.

[5] REYNOLDS L and MCDONELL K. Prompt programming for large language models: Beyond the few-shot paradigm[C]. Proceedings of 2021 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 2021, doi: 10.1145/3411763.3451760.

[6] GAWLIKOWSKI J, TASSI C R N, ALI M, *et al.* A survey of uncertainty in deep neural networks[J]. arXiv: 2107.03342, 2021.

[7] LECUN Y. 1.1 Deep learning hardware: Past, present, and future[C]. Proceedings of 2019 IEEE International Solid-State Circuits Conference, San Francisco, USA, 2019: 536–543. doi: 10.1109/ISSCC.2019.8662396.

[8] MENGHANI G. Efficient deep learning: A survey on making deep learning models smaller, faster, and better[J]. *ACM Computing Surveys*, 2023, 55(12): 259. doi: 10.1145/3578938.

[9] CAI Hao, BIAN Zhongjian, HOU Yaoru, *et al.* 33.4 A 28nm 2Mb STT-MRAM computing-in-memory macro with a refined bit-cell and 22.4–41.5 TOPS/W for AI inference[C]. Proceedings of 2023 IEEE International Solid-State Circuits Conference, San Francisco, USA, 2023. doi: 10.1109/ISSCC42615.2023.10067339.

[10] GUO An, SI Xin, CHEN Xi, *et al.* A 28nm 64-kb 31.6-TFLOPS/W digital-domain floating-point-computing-unit and double-bit 6T-SRAM computing-in-memory macro for floating-point CNNs[C]. Proceedings of 2023 IEEE International Solid-State Circuits Conference, San Francisco, USA, 2023. doi: 10.1109/ISSCC42615.2023.10067260.

[11] WANG Shaowei, XIE Guangjun, HAN Jie, *et al.* Highly accurate division and square root circuits by exploiting signal correlation in stochastic computing[J]. *International Journal of Circuit Theory and Applications*, 2022, 50(4): 1375–1385. doi: 10.1002/cta.3219.

[12] XU Wenbing, XIE Guangjun, WANG Shaowei, *et al.* A stochastic computing architecture for local contrast and mean image thresholding algorithm[J]. *International Journal of Circuit Theory and Applications*, 2022, 50(9): 3279–3291. doi: 10.1002/cta.3320.

[13] LIU Shanshan, TANG Xiaochen, NIKNIA F, *et al.* Stochastic dividers for low latency neural networks[J]. *IEEE Transactions on Circuits and Systems I:Regular Papers*, 2021, 68(10): 4102–4115. doi: 10.1109/TCSI.2021.3103926.

[14] LIU Weiqiang and LOMBARDI F. Approximate Computing[M]. Cham: Springer, 2022. doi: 10.1007/978-3-030-98347-5.

[15] LIU Weiqiang, LOMBARDI F, and SCHULTE M. Approximate computing: From circuits to applications[J]. *Proceedings of the IEEE*, 2020, 108(12): 2103–2107. doi: 10.1109/JPROC.2020.3033361.

[16] REDA S and SHAFIQUE M. Approximate Circuits[M]. Cham: Springer, 2019. doi: 10.1007/978-3-319-99322-5.

[17] 高越. 深度神经网络中Softmax函数的近似设计与硬件加速[D]. [硕士论文], 南京航空航天大学, 2021. doi: 10.27239/d.cnki.gnhhu.2021.001639.
GAO Yue. Approximate design and hardware acceleration of Softmax function for deep neural networks[D]. [Master dissertation], Nanjing University of Aeronautics and Astronautics, 2021. doi: 10.27239/d.cnki.gnhhu.2021.001639.

[18] 闫成刚, 赵轩, 徐宸宇, 等. 基于部分积概率分析的高精度低功耗近似浮点乘法器设计[J]. 电子与信息学报, 2023, 45(1): 87–95. doi: 10.11999/JEIT211485.
YAN Chenggang, ZHAO Xuan, XU Chenyu, *et al.* Design of high precision low power approximate floating-point multiplier based on partial product probability analysis[J]. *Journal of Electronics & Information Technology*, 2023, 45(1): 87–95. doi: 10.11999/JEIT211485.

[19] 黄乐朋. 面向低功耗关键词识别的近似计算模块设计[D]. [硕士论文], 东南大学, 2021. doi: 10.27014/d.cnki.gdnau.2021.003060.
HUANG Lepeng. Design of approximate computing module for low power keyword recognition[D]. [Master dissertation], Southeast University, 2021. doi: 10.27014/d.cnki.gdnau.2021.003060.

[20] 朱文涛. 面向多噪声场景低功耗关键词识别的可重构架构及电路实现[D]. [硕士论文], 东南大学, 2021. doi: 10.27014/d.cnki.gdnau.2021.003102.
ZHU Wentao. Design and implementation of a reconfigurable architecture for low-power keyword spotting in multiple noise scenarios[D]. [Master dissertation], Southeast University, 2021. doi: 10.27014/d.cnki.gdnau.2021.003102.

[21] 李焱. 面向二值化权重网络的近似加法单元设计[D]. [硕士论文], 东南大学, 2021. doi: 10.27014/d.cnki.gdnau.2021.003084.
LI Yan. Design of approximate addition unit for binarized weight network[D]. [Master dissertation], Southeast University, 2021. doi: 10.27014/d.cnki.gdnau.2021.003084.

[22] 张子骥. 基于近似计算的高能效电路设计技术研究[D]. [博士论文], 电子科技大学, 2021. doi: 10.27005/d.cnki.gdzku.2021.005070.
ZHANG Ziji. Research on energy-efficient circuit design technology based on approximate calculation[D]. [Ph. D. dissertation], University of Electronic Science and Technology, 2021. doi: 10.27005/d.cnki.gdzku.2021.005070.

[23] 裴浩然. 基于近似计算的自适应滤波器设计[D]. [硕士论文], 电子科技大学, 2021. doi: 10.27005/d.cnki.gdzku.2021.000370.
PEI Haoran. Design of adaptive filter based on approximate computing[D]. [Master dissertation], University of Electronic Science and Technology, 2021. doi: 10.27005/d.cnki.gdzku.2021.000370.

[24] 徐成文. 基于神经网络的近似计算训练算法[D]. [硕士论文], 上海交通大学, 2019. doi: 10.27307/d.cnki.gsjtu.2019.001935.
XU Chengwen. Cost-efficient and quality assured approximate computing framework using nerual network[D]. [Master dissertation], Shanghai Jiao Tong University, 2019. doi: 10.27307/d.cnki.gsjtu.2019.001935.

[25] 武翔宇. 一种新型的近似计算训练框架[D]. [硕士论文], 上海交通大学, 2017.
WU Xiangyu. A novel quality trade-offs method for approximate acceleration by iterative training[D]. [Master dissertation], Shanghai Jiao Tong University, 2017.

[26] 赵越. 基于近似计算的低功耗乘法器设计与实现[D]. [硕士论文], 上海交通大学, 2020. doi: 10.27307/d.cnki.gsjtu.2020.002087.
ZHAO Yue. Design and implementation of low power multiplier based on approximate computing[D]. [Master dissertation], Shanghai Jiao Tong University, 2020. doi: 10.27307/d.cnki.gsjtu.2020.002087.

[27] 季宇, 张悠慧, 郑纬民. 基于忆阻器的近似计算方法[J]. 清华大学学报:自然科学版, 2021, 61(6): 610–617. doi: 10.16511/j.cnki.qhdxxb.2020.22.027.
JI Yu, ZHANG Youhui, and ZHENG Weimin. Approximate computing method based on memristors[J]. Journal of Tsinghua University:Science and Technology, 2021, 61(6): 610–617. doi: 10.16511/j.cnki.qhdxxb.2020.22.027.

[28] 王智慧. 基于近似计算的高能效JPEG加速器设计[D]. [硕士论文], 清华大学, 2018. doi: 10.27266/d.cnki.gqhau.2018.000230.
WANG Zhihui. An energy efficient JPEG encoder with approximate computing paradigm[D]. [Master dissertation], Tsinghua University, 2018. doi: 10.27266/d.cnki.gqhau.2018.000230.

[29] 张士长, 王郁杰, 肖航, 等. 支持CNN与LSTM的二值权重神经网络芯片[J]. 高技术通讯, 2021, 31(2): 122–128. doi: 10.3772/j.issn.1002-0470.2021.02.002.
ZHANG Shichang, WANG Yujie, XIAO Hang, et al. Binary-weight neural network chip supporting CNN and LSTM[J]. Chinese High Technology Letters, 2021, 31(2): 122–128. doi: 10.3772/j.issn.1002-0470.2021.02.002.

[30] 陆维娜, 胡瑜, 叶靖, 等. 面向卷积神经网络加速器吞吐量优化的FPGA自动化设计方法[J]. 计算机辅助设计与图形学学报, 2018, 30(11): 2164–2173. doi: 10.3724/SP.J.1089.2018.17039.
LU Weina, HU Yu, YE Jing, et al. Throughput-oriented automatic design of FPGA accelerator for convolutional neural networks[J]. Journal of Computer-Aided Design & Computer Graphics, 2018, 30(11): 2164–2173. doi: 10.3724/SP.J.1089.2018.17039.

[31] 朱新忠, 程利甫, 吴有余, 等. 基于误差模型的权重二值神经网络近似加速[J]. 上海航天(中英文), 2021, 38(4): 25–30. doi: 10.19328/j.cnki.2096-8655.2021.04.004.
ZHU Xinzhong, CHENG Lifu, WU Youyu, et al. Error model based approximate computing design for binarized weight neural network system[J]. Aerospace Shanghai (Chinese & English), 2021, 38(4): 25–30. doi: 10.19328/j.cnki.2096-8655.2021.04.004.

[32] LIANG Tailin, GLOSSNER J, WANG Lei, et al. Pruning and quantization for deep neural network acceleration: A

survey[J]. *Neurocomputing*, 2021, 461: 370–403. doi: 10.1016/j.neucom.2021.07.045.

[33] BASKIN C, LISS N, SCHWARTZ E, *et al*. UNIQ: Uniform noise injection for non-uniform quantization of neural networks[J]. *ACM Transactions on Computer Systems*, 2019, 37(1/4): 4. doi: 10.1145/3444943.

[34] BANNER R, NAHSHAN Y, and SOUDRY D. Post training 4-bit quantization of convolutional networks for rapid-deployment[C]. Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, Canada, 2019.

[35] VOGEL S, SPRINGER J, GUNTORO A, *et al*. Self-supervised quantization of pre-trained neural networks for multiplierless acceleration[C]. Proceedings of 2019 Design, Automation & Test in Europe Conference & Exhibition, Florence, Italy, 2019: 1094–1099. doi: 10.23919/DATE.2019.8714901.

[36] GYSEL P, PIMENTEL J, MOTAMEDI M, *et al*. Ristretto: A framework for empirical study of resource-efficient inference in convolutional neural networks[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2018, 29(11): 5784–5789. doi: 10.1109/TNNLS.2018.2808319.

[37] NAGEL M, FOURNARAKIS M, AMJAD R A, *et al*. A white paper on neural network quantization[J]. arXiv: 2106.08295, 2021.

[38] JUNG S, SON C, LEE S, *et al*. Learning to quantize deep networks by optimizing quantization intervals with task loss[C]. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 4345–4354. doi: 10.1109/CVPR.2019.00448.

[39] CHOI J, VENKATARAMANI S, SRINIVASAN V, *et al*. Accurate and efficient 2-bit quantized neural networks[C]. Proceedings of Machine Learning and Systems 2019, Stanford, USA, 2019.

[40] ZHOU Shuchang, WU Yuxin, NI Zekun, *et al*. DoReFa-Net: Training low bitwidth convolutional neural networks with low bitwidth gradients[J]. arXiv: 1606.06160, 2018.

[41] RASTEGARI M, ORDONEZ V, REDMON J, *et al*. XNOR-Net: ImageNet classification using binary convolutional neural networks[C]. Proceedings of the 14th European Conference on Computer Vision, Amsterdam, The Netherlands, 2016: 525–542. doi: 10.1007/978-3-319-46493-0_32.

[42] ANDRI R, CAVIGELLI L, ROSSI D, *et al*. YodaNN: An architecture for ultralow power binary-weight CNN acceleration[J]. *IEEE Transactions on Computer-Aided*

Design of Integrated Circuits and Systems, 2018, 37(1): 48–60. doi: 10.1109/TCAD.2017.2682138.

[43] GONG Yu, CAI Hao, WU Haige, *et al*. Quality driven systematic approximation for binary-weight neural network deployment[J]. *IEEE Transactions on Circuits and Systems I:Regular Papers*, 2022, 69(7): 2928–2940. doi: 10.1109/TCSI.2022.3164170.

[44] SHAFIQUE M, AHMAD W, HAFIZ R, *et al*. A low latency generic accuracy configurable adder[C]. Proceedings of the 52nd Annual Design Automation Conference, San Francisco, USA, 2015: 86. doi: 10.1145/2744769.2744778.

[45] HANIF M A, HAFIZ R, HASAN O, *et al*. QuAd: Design and analysis of quality-area optimal low-latency approximate adders[C]. Proceedings of the 54th Annual Design Automation Conference, Austin, USA, 2017: 42. doi: 10.1145/3061639.3062306.

[46] ZHU Ning, GOH W L, ZHANG W, *et al*. Design of low-power high-speed truncation-error-tolerant adder and its application in digital signal processing[J]. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 2010, 18(8): 1225–1229. doi: 10.1109/TVLSI.2009.2020591.

[47] ZHU Ning, GOH W L, and YEO K S. An enhanced low-power high-speed adder for error-tolerant application[C]. Proceedings of the 2009 12th International Symposium on Integrated Circuits, Singapore, 2009: 69–72.

[48] GUPTA V, MOHAPATRA D, RAGHUNATHAN A, *et al*. Low-power digital signal processing using approximate adders[J]. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2013, 32(1): 124–137. doi: 10.1109/TCAD.2012.2217962.

[49] CAMUS V, SCHLACHTER J, and ENZ C. Energy-efficient inexact speculative adder with high performance and accuracy control[C]. Proceedings of 2015 IEEE International Symposium on Circuits and Systems, Lisbon, Portugal, 2015: 45–48. doi: 10.1109/ISCAS.2015.7168566.

[50] CAMUS V, SCHLACHTER J, and ENZ C. A low-power *carry cut-back* approximate adder with fixed-point implementation and floating-point precision[C]. Proceedings of the 53rd Annual Design Automation Conference, Austin, USA, 2016: 127. doi: 10.1145/2897937.2897964.

[51] VERMA A K, BRISK P, and IENNE P. Variable latency speculative addition: A new paradigm for arithmetic circuit design[C]. Proceedings of 2008 Design, Automation and Test in Europe, Munich, Germany, 2008: 1250–1255. doi: 10.1109/DATE.2008.4484850.

[52] MORGENSHTEIN A, YUZHANINOV V, KOVSHILOVSKY A, et al. Full-swing gate diffusion input logic-case-study of low-power CLA adder design[J]. Integration, 2014, 47(1): 62–70. doi: 10.1016/j.vlsi.2013.04.002.

[53] HAN Jie and ORSHANSKY M. Approximate computing: An emerging paradigm for energy-efficient design[C]. Proceedings of 2013 18th IEEE European Test Symposium, Avignon, France, 2013: 1–6. doi: 10.1109/ETS.2013.6569370.

[54] MAHDIANI H R, AHMADI A, FAKHRAIE S M et al. Bio-inspired imprecise computational blocks for efficient VLSI implementation of soft-computing applications[J]. IEEE Transactions on Circuits and Systems I:Regular Papers, 2010, 57(4): 850–862. doi: 10.1109/TCSI.2009.2027626.

[55] JOHN V, SAM S, RADHA S, et al. Design of a power-efficient Kogge–Stone adder by exploring new OR gate in 45nm CMOS process[J]. Circuit World, 2020, 46(4): 257–269. doi: 10.1108/CW-12-2018-0104.

[56] LIU Bo, WANG Ziyu, WANG Xuetao, et al. An efficient BCNN deployment method using quality-aware approximate computing[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2022, 41(11): 4217–4228. doi: 10.1109/TCAD.2022.3197509.

[57] LIU Weiqiang, QIAN Liangyu, WANG Chenghua, et al. Design of approximate radix-4 booth multipliers for error-tolerant computing[J]. IEEE Transactions on Computers, 2017, 66(8): 1435–1441. doi: 10.1109/TC.2017.2672976.

[58] BORO B, REDDY K M, KUMAR Y B N, et al. Approximate radix-8 Booth multiplier for low power and high speed applications[J]. Microelectronics Journal, 2020, 101: 104816. doi: 10.1016/j.mejo.2020.104816.

[59] WARIS H, WANG Chenghua, and LIU Weiqiang. Hybrid low radix encoding-based approximate booth multipliers[J]. IEEE Transactions on Circuits and Systems II:Express Briefs, 2020, 67(12): 3367–3371. doi: 10.1109/TCSII.2020.2975094.

[60] YIN Shouyi, OUYANG Peng, ZHENG Shixuan, et al. A 141 UW, 2.46 PJ/neuron binarized convolutional neural network based self-learning speech recognition processor in 28NM CMOS[C]. Proceedings of 2018 IEEE Symposium on VLSI Circuits, Honolulu, USA, 2018: 139–140. doi: 10.1109/VLSIC.2018.8502309.

[61] ZHAO Yue, LI Tong, DONG Feng, et al. A new approximate multiplier design for digital signal processing[C]. Proceedings of 2019 IEEE 13th international conference on ASIC, Chongqing China, 2019: 1–4. doi: 10.1109/ASICON47005.2019.8983437.

[62] MITCHELL J N. Computer multiplication and division using binary logarithms[J]. IRE Transactions on Electronic Computers, 1962, EC-11(4): 512–517. doi: 10.1109/TEC.1962.5219391.

[63] YIN Peipei, WANG Chenghua, WARIS H, et al. Design and analysis of energy-efficient dynamic range approximate logarithmic multipliers for machine learning[J]. IEEE Transactions on Sustainable Computing, 2021, 6(4): 612–625. doi: 10.1109/TSUSC.2020.3004980.

[64] LIU Weiqiang, XU Jiahua, WANG Danye, et al. Design and evaluation of approximate logarithmic multipliers for low power error-tolerant applications[J]. IEEE Transactions on Circuits and Systems I:Regular Papers, 2018, 65(9): 2856–2868. doi: 10.1109/TCSI.2018.2792902.

[65] SAADAT H, JAVAID H, IGNJATOVIC A, et al. REALM: Reduced-error approximate log-based integer multiplier[C]. Proceedings of 2020 Design, Automation & Test in Europe Conference & Exhibition, Grenoble, France, 2020: 1366–1371. doi: 10.23919/DATE48585.2020.9116315.

[66] MRAZEK V, HRBACEK R, VASICEK Z, et al. EvoApprox8b: Library of approximate adders and multipliers for circuit design and benchmarking of approximation methods[C]. Proceedings of the Design, Automation & Test in Europe Conference & Exhibition, Lausanne, Switzerland, 2017: 258–261. doi: 10.23919/DATE.2017.7926993.

[67] VENKATARAMANI S, RANJAN A, ROY K, et al. AxNN: Energy-efficient neuromorphic systems using approximate computing[C]. Proceedings of 2014 IEEE/ACM International Symposium on Low Power Electronics and Design, La Jolla, USA, 2014: 27–32. doi: 10.1145/2627369.2627613.

[68] GIRALDO J S P and VERHELST M. Laika: A 5uW programmable LSTM accelerator for always-on keyword spotting in 65nm CMOS[C]. Proceedings of 2018 IEEE 44th European Solid State Circuits Conference, Dresden, Germany, 2018: 166–169. doi: 10.1109/ESSCIRC.2018.8494342.

[69] PRICE M, GLASS J, and CHANDRAKASAN A P. 14.4 A scalable speech recognizer with deep-neural-network acoustic models and voice-activated power gating[C]. Proceedings of 2017 IEEE International Solid-State Circuits Conference, San Francisco, USA, 2017: 244–245. doi: 10.1109/ISSCC.2017.7870352.

[70] YIN Shouyi, OUYANG Peng, TANG Shibin, et al. A high

energy efficient reconfigurable hybrid neural network processor for deep learning applications[J]. *IEEE Journal of Solid-State Circuits*, 2018, 53(4): 968–982. doi: 10.1109/JSSC.2017.2778281.

[71] YIN Shouyi, OUYANG Peng, YANG Jianxun, *et al*. An energy-efficient reconfigurable processor for binary-and ternary-weight neural networks with flexible data bit width[J]. *IEEE Journal of Solid-State Circuits*, 2019, 54(4): 1120–1136. doi: 10.1109/JSSC.2018.2881913.

[72] LU Wenyan, YAN Guihai, and LI Xiaowei. AdaFlow: Aggressive convolutional neural networks approximation by leveraging the input variability[J]. *Journal of Low Power Electronics*, 2018, 14(4): 481–495. doi: 10.1166/jolpe.2018.1581.

[73] WANG Ying, HE Yintao, CHENG Long, *et al*. A fast precision tuning solution for always-on DNN accelerators[J]. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2022, 41(5): 1236–1248. doi: 10.1109/TCAD.2021.3089667.

[74] KIM Y D, JEONG W, JUNG L, *et al*. 2.4 A 7nm high-performance and energy-efficient mobile application processor with tri-cluster CPUs and a sparsity-aware NPU[C]. Proceedings of 2020 IEEE International Solid-State Circuits Conference, San Francisco, USA, 2020: 48–50. doi: 10.1109/ISSCC19947.2020.9062907.

[75] MAZAHIR S, HASAN O, HAFIZ R, *et al*. Probabilistic error modeling for approximate adders[J]. *IEEE Transactions on Computers*, 2017, 66(3): 515–530. doi: 10.1109/TC.2016.2605382.

[76] AYUB M K, HASAN O, and SHAFIQUE M. Statistical error analysis for low power approximate adders[C]. Proceedings of the 54th ACM/EDAC/IEEE Design Automation Conference, Austin, USA, 2017: 1–6. doi: 10.1145/3061639.3062319.

[77] ZHU Yiying, LIU Weiqiang, HAN Jie, *et al*. A probabilistic error model and framework for approximate booth multipliers[C]. Proceedings of 2018 IEEE/ACM International Symposium on Nanoscale Architectures, Athens, Greece, 2018: 1–6.

[78] LIU Chang, YANG Xinghua, QIAO Fei, *et al*. Design methodology for approximate accumulator based on statistical error model[C]. Proceedings of the 20th Asia and South Pacific Design Automation Conference, Chiba, Japan, 2015: 237–242. doi: 10.1109/ASPDAC.2015.7059011.

[79] VENKATESAN R, AGARWAL A, ROY K, *et al*. MACACO: Modeling and analysis of circuits for approximate computing[C]. Proceedings of 2011 IEEE/ACM International Conference on Computer-Aided Design, San Jose, USA, 2011: 667–673. doi: 10.1109/ICCAD.2011.6105401.

[80] LIU Zheyu, LI Guihong, QIAO Fei, *et al*. Concrete: A per-layer configurable framework for evaluating DNN with approximate operators[C]. Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton UK, 2019: 1552–1556. doi: 10.1109/ICASSP.2019.8682883.

[81] FRENO B A and CARLBERG K T. Machine-learning error models for approximate solutions to parameterized systems of nonlinear equations[J]. *Computer Methods in Applied Mechanics and Engineering*, 2019, 348: 250–296. doi: 10.1016/j.cma.2019.01.024.

[82] JIANG Weiwen, ZHANG Xinyi, SHA E H M, *et al*. Accuracy vs. efficiency: Achieving both through FPGA-implementation aware neural architecture search[C]. Proceedings of 2019 56th ACM/IEEE Design Automation Conference, Las Vegas, USA, 2019: 1–6.

[83] MRAZEK V, HANIF M A, VASICEK Z, *et al*. autoAx: An automatic design space exploration and circuit building methodology utilizing libraries of approximate components[C]. Proceedings of 2019 56th ACM/IEEE Design Automation Conference, Las Vegas, USA, 2019: 1–6.

[84] ULLAH S, SAHOO S S, and KUMAR A. CLAppED: A design framework for implementing cross-layer approximation in FPGA-based embedded systems[C]. Proceedings of 2021 58th ACM/IEEE Design Automation Conference, San Francisco, USA, 2021: 475–480. doi: 10.1109/DAC18074.2021.9586260.

[85] YU Ye, LI Yingmin, CHE Shuai, *et al*. Software-defined design space exploration for an efficient DNN accelerator architecture[J]. *IEEE Transactions on Computers*, 2021, 70(1): 45–56. doi: 10.1109/TC.2020.2983694.

[86] MEI Linyan, HOUSHMAND P, JAIN V, *et al*. ZigZag: Enlarging joint architecture-mapping design space exploration for DNN accelerators[J]. *IEEE Transactions on Computers*, 2021, 70(8): 1160–1174. doi: 10.1109/TC.2021.3059962.

[87] LI Haitong, BHARGAV M, WHATMOUGH P N, *et al*. On-chip memory technology design space explorations for mobile deep neural network accelerators[C]. Proceedings of 2019 56th ACM/IEEE Design Automation Conference, Las Vegas, USA, 2019: 1–6.

[88] JIANG Weiwen, LOU Qiuwen, YAN Zheyu, *et al*. Device-

circuit-architecture co-exploration for computing-in-memory neural accelerators[J]. *IEEE Transactions on Computers*, 2021, 70(4): 595–605. doi: 10.1109/TC.2020.2991575.

[89] VENKATESAN R, SHAO Y S, WANG Miaorong, *et al.* MAGNet: A modular accelerator generator for neural networks[C]. Proceedings of 2019 IEEE/ACM International Conference on Computer-Aided Design, Westminster, USA, 2019: 1–8. doi: 10.1109/ICCAD45719.2019.8942127.

[90] LIN Yujun, YANG Mengtian, and HAN Song. NAAS: Neural accelerator architecture search[C]. Proceedings of 2021 58th ACM/IEEE Design Automation Conference, San Francisco, USA, 2021: 1051–1056. doi: 10.1109/DAC18074. 2021.9586250.

[91] VENIERIS S I and BOUGANIS C S. fpgaConvNet: Mapping regular and irregular convolutional neural networks on FPGAs[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2019, 30(2): 326–342. doi: 10.1109/TNNLS.2018.2844093.

[92] REAGEN B, HERNÁNDEZ-LOBATO J M, ADOLF R, *et al.* A case for efficient accelerator design space exploration via Bayesian optimization[C]. Proceedings of 2017 IEEE/ACM International Symposium on Low Power Electronics and Design, Taipei, China, 2017: 1–6. doi: 10. 1109/ISLPED.2017.8009208.

龚　宇：男，副研究员，研究方向为近似计算、人工智能系统芯片设计.

王丽萍：女，博士生，研究方向为近似计算、人工智能系统芯片设计.

王　佑：男，副研究员，研究方向为近似计算、数字芯片设计.

刘伟强：男，教授，研究方向为近似计算、集成电路设计等.

责任编辑：于　青