核电子学与探测技术
Nuclear Electronics & Detection Technology
ISSN 0258-0934,CN 11-2016/TL

《核电子学与探测技术》网络首发论文

题目：基于 FPGA 的卷积神经网络核素识别硬件加速方法研究
作者：王博，石睿，刘敏俊，曾雄，王洲
收稿日期：2023-11-23
网络首发日期：2024-02-29
引用格式：王博，石睿，刘敏俊，曾雄，王洲．基于 FPGA 的卷积神经网络核素识别硬件加速方法研究[J/OL]．核电子学与探测技术.
https://link.cnki.net/urlid/11.2016.TL.20240229.0925.002

基于 FPGA 的卷积神经网络核素识别硬件加速方法研究

王博，石睿*，刘敏俊，曾雄，王洲

（四川轻化工大学计算机科学与工程学院，四川 宜宾 644005）

摘要：核素识别是核探测领域研究的关键技术之一，传统基于能谱解谱算法的核素识别仪器，实时性差，功耗较高，限制了实际应用中的识别效率，为了加快对放射性核素定性分析，本文提出了一种基于 FPGA 的卷积神经网络核素识别硬件加速方法。首先提出了一种用于核素分类的轻量型一维卷积神经网络模型，再根据模型卷积层、池化层和全连接层的运算特点，利用并行流水线和加法树等硬件加速策略，将模型部署在 Xilinx ZYNQ7020 异构芯片中。实验结果表明，在 FPGA 中，测试集平均识别精度达到 98.41%，单次识别耗时 1.57ms，与桌面端 CPU 相比，该硬件加速方法实现了 64 倍加速效果，功耗仅为 2.115W。在实际测试实验中，^{137}Cs 单源识别精度为 98%，^{137}Cs 与 ^{60}Co 混合源识别精度达到 98.17%。该硬件加速方案满足低延时、低功耗等要求，适合于现场快速核素检测的场景，对便携式核素识别仪器开发具有重要的参考价值。

关键词：能谱数据；核素识别；FPGA；卷积神经网络；硬件加速；

中图分类号：TL81　　文献标志码：A

放射性核素在核能、核医学、核武器等领域具有丰富的应用价值，在使用过程中也伴随核安全事故发生的风险，如放射性物质泄漏，辐射源丢失等事件[1,2]。核素识别是检测放射性物质的重要手段，是核安全、核探测领域重要的研究方向，对保障核安全和防范核与辐射事故具有重要意义。在环境监测、海关安检、核废物处理等场景中，是不可缺少的环节[3]。

传统放射性核素识别算法通过数字化多道能谱仪对核脉冲信号进行采集，统计脉冲幅度形成能谱，对能谱进行光滑、寻峰、能量刻度等分析，提取能谱特征信息，再与核素库特征能量匹配，进行核素识别[4]。传统方法依赖探测器和谱仪性能，识别步骤繁琐，识别时间较长，易受环境影响，在低计数率、本底噪声较大的环境下，能谱解析难度大，在快速核素检测场景中，识别效率不高。为进一步提高识别精度，一些学者尝试利用序贯贝叶斯和支持向量机等传统机器学习算法来解决核素分类问题[5-9]，与能谱分析法相比，机器学习算法能够更快得出核素分类结果，但传统机器学习算法因提取数据特征能力较弱，泛化能力有限等缺点，限制了分类精度的进一步提升。随着机器学习向深度学习的进一步发展，Cao 等人[10]利用人工神经网络，建立了能谱全谱信息与核素类别之间的联系，对单源和混合源的识别率达到了 98.8%和 94.9%，但人工神经网络模型参数量大，计算量较大。为了进一步降低模型体积和提高识别的准确率，Liang 等人[11]构建了卷积神经网络（CNN），通过卷积操作提取能谱特征，降低数据维度，综合识别率可达到 99%以上。Wang 等人[12]也实验证明了 CNN 提取能谱全能峰和康普顿边缘特征的有效性。

综上可知，卷积神经网络为核素识别提供了新的发展思路，简化了识别步骤。随着硬件研究的不断深入，芯片的算力成为算法实际应用的最大瓶颈。深度学习算法常见的部署平台有中央处理器（CPU）、图形处理器（GPU）、专用集成电路（ASIC）、现场可编程逻辑门阵列（FPGA）等。但 CPU 与 GPU 功耗

收稿日期：2023-11-23

基金项目：国家自然科学基金(Nos.42074218, U19A2086, 42204179, 12205210)、四川轻化工大学研究生创新基金项目(No.Y2022171)资助。

作者简介：王博(1998－)，男，四川南充人，在读硕士研究生，主要从事核信息获取与处理研究。

***通讯作者：**石睿(1988－)，男，博士，教授，硕士生导师，主要从事核技术及应用研究，E-mail:shirui@suse.edu.cn。

大，不适用于移动端，ASIC 开发成本高，开发周期较长，难以做到通用性，而 FPGA 是一种可以将代码综合为数字电路的可编程逻辑器件，功耗低，开发方式灵活，开发周期短。一些学者也尝试利用 FPGA 芯片的并行性对 CNN 进行部署。杜煜章等人[13]在 FPGA 中实现了一种心音 CNN 分类算法，将卷积层进行并行计算，相比于 CPU 得到了 14 倍加速。Wang 等人[14]在 FPGA 中实现了一种基于声信号的一维卷积神经网络（1D-CNN）目标识别算法，实验结果表明，在保证精度的前提下，极大地提高了计算速度，保证了识别过程的实时性。在核素识别领域，Huang 等人[15]将脉冲神经网络核素识别算法在 FPGA 中进行部署，实验表明在 FPGA 中以 75mW 的功耗实现了 90.62%的识别精度。前面学者的研究验证了 FPGA 对神经网络推理加速的可行性，考虑到能谱采集设备数字化多道脉冲幅度分析仪内部主要依靠 FPGA 芯片对核脉冲信号进行幅值提取和能量统计[16]，因此选择 FPGA 作为核素识别算法移动端部署平台最具性价比。

本文将在 FPGA 中实现一种基于卷积神经网络的核素识别算法。首先训练一种轻量化 1D-CNN 网络模型；然后在 FPGA 中实现卷积神经网络中卷积层、池化层和全连接层；采用并行流水线和加法树等优化方式，对网络的推理过程进行加速，提高计算的并行度，降低模型推理时间。本文提出的核素识别硬件加速方法，满足低延时，低功耗，低成本的要求，实现了移动端芯片对核素快速，准确的定性分析。

1 基于 1D-CNN 的核素识别模型建立

1.1 γ 能谱数据集

卷积神经网络依赖的数据集主要来自蒙特卡罗程序包 Geant4 所生成的能谱数据，通过构建 NaI(Tl)晶体，以探测距离和发射光子数为变量，每种核素模拟 1500 个能谱，能谱总计数范围在 10^2 至 10^5 之间。测得工业中常见的 ^{60}Co、^{137}Cs 和 ^{152}Eu 三种单源核素以及其构成的混合核素共七种分类，共测得数据 10500 组。数据集按 7:3 划分训练集和测试集。由于 Geant4 模拟并未考虑电离过程的统计涨落，因此需要对统计得到的能谱进行高斯展宽，如公式：

$$FWHM(E_d) = a + b\sqrt{E_d + cE_d^2} \tag{1}$$

式中 $a = 0.1243$、$b = 0.1418$、$c = -0.4388$，均为实验室 NaI(Tl)探测器所测定系数，$FWHM$ 为全能峰的半高宽，E_d 为能量值。为了减少噪声和统计涨落的影响，得到更加清晰可靠的能谱数据，使用数据平滑方法减少能谱中的异常值，常见的平滑方法有多项式平滑法、重心法、小波变换法等[17]。本文采用五点重心平滑，将能谱数据与相邻 4 个数据进行加权求平均，得到新的道址数据，数学表达式为：

$$\bar{Y} = \frac{1}{16}\left(X_{i-2} + 4X_{i-1} + 6X_i + 4X_{i+1} + X_{i+2}\right) \tag{2}$$

式中 X_i 为第 i 道对应的能谱数据，$\bar{Y}$ 为平滑后的数据。平滑后的能谱数据如图 1 所示，道址划分为 2048 道。

1.2 深度可分离卷积

能谱数据是一维向量，因此构建的是一维卷积神经网络。卷积神经网络主要由卷积层、池化层、全连接层组成。卷积层计算量占整个网络模型计算量的 90%以上，为了尽可能的减少计算量，谷歌提出了轻量级神经网络 MobileNet[18]，使用深度可分离卷积代替普通卷积计算。本文将 MobileNet 中二维深度可分离

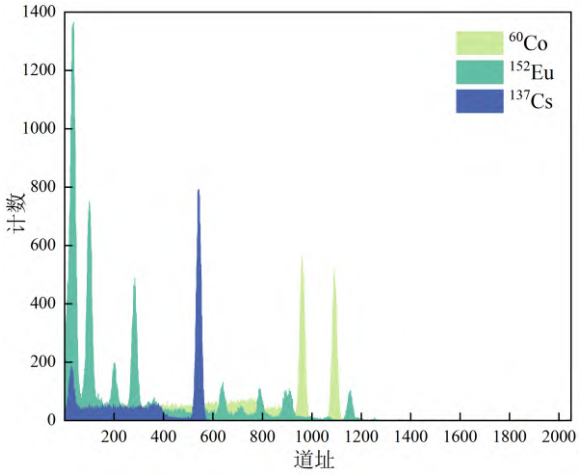

图 1　Geant4 生成的 γ 能谱数据

卷积改进为一维深度可分离卷积。普通卷积中一个卷积核需要联合不同通道进行特征提取。深度可分离卷积将普通的卷积运算分解为逐通道卷积和逐点卷积，逐通道卷积一个卷积核只负责一个通道，一个通道只被一个卷积核卷积，尺寸为 1×1 的逐点卷积核再联合不同通道生成下一层特征图。逐通道卷积计算量为：$D_K \times M \times D_F$。逐点卷积计算量为：$M \times N \times D_F$。相比于普通卷积，深度可分离卷积参数量和计算成本均得到了降低[19]，计算量下降比为：

$$\frac{D_{\mathrm{K}} \times M \times D_{\mathrm{F}} + M \times N \times D_{\mathrm{F}}}{D_{\mathrm{K}} \times M \times N \times D_{\mathrm{F}}} = \frac{1}{N} + \frac{1}{D_{\mathrm{K}}} \tag{3}$$

其中 M、N 分别为卷积核输入、输出通道数量，D_K 为卷积核大小，D_F 为特征图大小。因参数量减少，采用深度可分离卷积的模型更容易泛化到新的数据集中，在训练模型时需要的时间更短，提高了模型的训练速度。在硬件实现模型的正向推理过程时，可减少资源使用率，降低硬件成本。

1.3　1D-CNN 核素识别模型

1D-CNN 核素识别模型结构如图 2 所示，包括三个卷积层，三个池化层和三个全连接层，卷积层使用深度可分离卷积替代。三个卷积层通道数量分别为 4、8、16，padding 填充方式设置为 same，填充大小为 2，卷积层间采用最大池化，激活函数采用 ReLU 函数，学习率设置为 0.001，优化器为随机梯度下降法 (Stochastic Gradient Descent, SGD)。在全连接层间添加 Dropout 操作，随机丢弃部分神经元，防止模型过拟合，输出层采用 Softmax 函数，采用 Pytorch 机器学习库训练。模型总共包含参数 74645 个，模型体积 0.44MB。

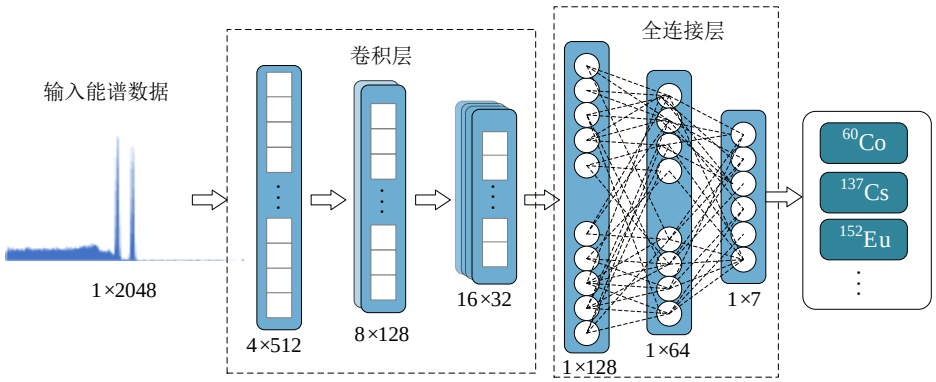

图 2　1D-CNN 核素识别模型

经过 50 个 epoch 后，训练损失值下降小于 0.01%，结束训练。在测试集中，平均准确率达到 98.48%，

各核素分类准确率如表 1 所示。从分类结果中分析，单源的识别准确率多数高于混合源。仔细观察混合源分类错误的样本，模型只能识别出其中一种或者两种核素。分析单源分类错误的部分样本，发现主要是由于能谱所记录的能量太少，全能峰特征不明显，导致分类错误，其中 ^{152}Eu 因特征能量分支较多，易与 ^{60}Co 混淆。

表 1　净峰面积不同的情况下寻峰方法是否可寻到有效峰对比

核素	标签	识别准确率/%
^{60}Co	0	100
^{60}Co+^{137}Cs	1	99.78
^{60}Co+^{137}Cs+^{152}Eu	2	96.22
^{60}Co+^{152}Eu	3	97.56
^{137}Cs	4	98.67
^{137}Cs+^{152}Eu	5	98.45
^{152}Eu	6	98.67

1.4　模型对比

使用深度可分离卷积的 1D-CNN 相对于标准的卷积具有轻量化的特点，为了验证本文所使用的模型，分别使用常见的分类模型进行对比，包括 MobileNet-V1、VGGNet、ResNet18 和 ShuffleNetV2。训练使用的参数与 1.3 小节一致。分类准确率和模型体积如表 2 所示。传统的 VGGNet、ResNet 使用的卷积层数多，模型参数量大，导致模型体积较大，对计算机硬件资源要求更高，在训练时，花费的时间更长。轻量化模型 1D-CNN、MobileNet-V1 和 ShuffleNet-V2 模型体积较小，准确率差距小于 1%，选择体积更小的 1D-CNN，更有利于在嵌入式设备中应用。

表 2　不同模型的准确率和参数量

模型	识别准确率/%	模型体积/MB
1D-CNN	98.48	0.44
ShuffleNet-V2	97.50	4.81
MobileNet-V1	98.90	12.26
ResNet18	99.94	42.62
VGGNet	99.89	128.30

2　FPGA 硬件加速方案研究

2.1　总体方案

CNN 算法的硬件加速实现需要软件和硬件的协同配合，FPGA 中只需实现网络的前向传播过程。整体软硬件实现方案如图 3 所示，主体分为模型建立和硬件实现。在计算机中训练好模型后，导出模型参数。目前 FPGA 实现 CNN 大多采用两种方案，一种是流水线架构[20]，将 CNN 的各层分别封装成 IP（Intellectual Property, IP）核，各层运算时相互独立，互不干扰，可同时计算，每一层都需要单独配置，适用于对识别的连续性较高的场景，该方案多用于视频流的识别。另一种是将 CNN 各层融合到一个 IP 核中，该架构虽没有流水线架构效率高，但简化了实现步骤，适用于识别帧间连续性低的场景，所使用的硬件资源量可以更少。由于能谱数据并不像视频流一样需要将连续的帧送入网络中，且能谱采集需要一定的时间，因此核素识别模型硬件部署不适用流水线架构。本文将 1D-CNN 实现为一个独立的 IP 核，模型参数直接存储在 FPGA 的片上 RAM 中，可以降低芯片功耗，加快计算单元读取数据的速度[21]。

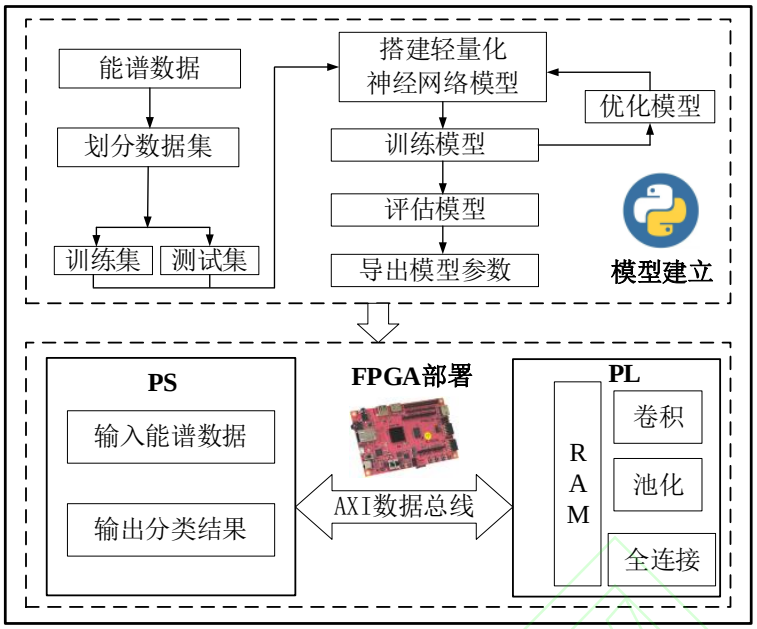

图 3　软硬件总体实施方案

FPGA 没有专用于浮点运算的逻辑单元，需要将模型参数进行定点量化处理。为了尽可能减少模型计算精度的丢失，本文输入数据及特征图采用 16bit 定点数位宽表示，中间计算结果采用 32bit 定点数表示。算法部署采用 Xilinx ZYNQ-7000 系列异构芯片，内部包含处理器单元（PS）和可编程逻辑单元（PL）。PS 与 PL 协同工作，CNN 算法部分主要在 PL 中实现，PS 实现数据流的控制，两者间依靠 AXI 总线传输数据。

2.2　卷积模块设计

卷积运算主要用于提取输入能谱数据特征，将卷积核与输入特征图相乘并进行累加得到输出特征图，其过程主要由大量的乘法和加法构成。卷积计算时各通道之间数据相互独立，无直接联系。在 FPGA 中可采用乘法器对卷积的不同通道，不同卷积核以及卷积核内部的计算循环体进行展开，展开数量根据卷积运算单个循环中乘法数量和 FPGA 乘法器数量而定。一维卷积计算过程伪代码如下所示：

```
一维卷积计算伪代码：
for(n=0;n<N;n++)          //定位输出通道
 for(m=0;m<M;m++)         //定位输入通道
  for(i=0;i<L;i++){       //定位特征图
   sum = 0;
   for(k=0;k<K;k++){      //定位卷积核数据
    sum+= in[m][i+k]*Kw[n][m][k];
   }
    out[n][i] = sum;
  }
```

使用 HLS 中的 PIPELINE 指令对循环进行流水线展开，PIPELINE 可根据 FPGA 芯片内部资源，最大化地增加卷积循环计算并行度。循环展开相当于循环体复制了 *n* 份，单个循环内部对数据操作可分为读取数据、计算数据、存储计算结果三步操作，流水线操作将不同的循环体之间执行顺序间隔一定周期，防止在同一时钟周期内，循环体实现同一操作造成计算瓶颈。如图 4 所示，串行计算中，两个循环操作需要六个时钟周期，并行流水线计算降低为 4 个周期，减少了对硬件资源的依赖性，增加计算吞吐量。

循环流水线增加了对存储在 ROM 中的模型参数访问量，存储在片上的 ROM 只提供了两个读写端口，

为满足并行性要求，使用 ARRAY_PARTITION 指令将数组分割成多个部分并存储在 FPGA 的不同区域，数组分割后有助于并行化数据操作，可降低数据访问延迟，提高运行时的吞吐量。对于不同通道之间，一次卷积计算需要联合不同通道特征图中的同一位置获取数据，在数组分割时，对特征图参数进行重新排序，交替存储。

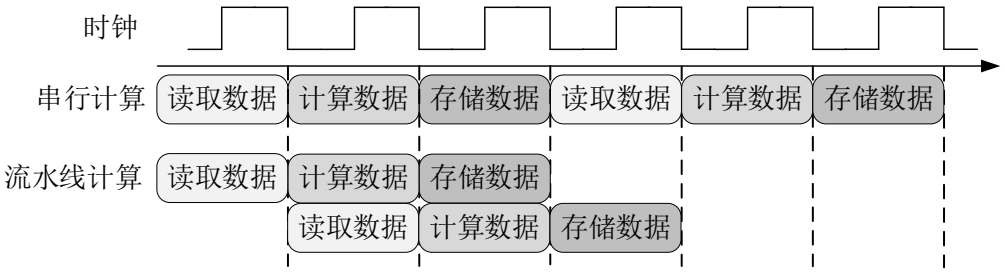

图 4　循环并行流水线执行

2.3　池化模块设计

池化操作在卷积神经网络中用于数据进下采样，其目的是降低数据维度、去除冗余信息，并减少计算复杂度。常见的池化方法包括最大池化和平均池化，其中最大池化是一种消耗资源最少的方式，它只需比较数据中的最大值而不进行其他计算。如图 5 所示，在 FPGA 中，将输入数据分成两组，第一个时钟周期分别比较最大值，在下一个时钟再取上一个周期输出的最大值，依靠三个数据比较器即可实现该功能。池化单元大小设置为 4，两个周期就可以得到池化结果。

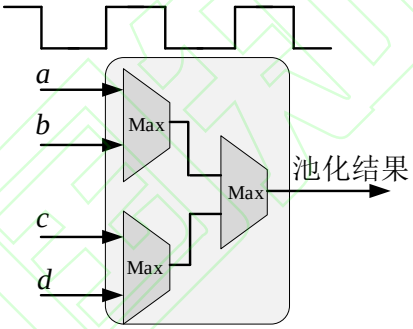

图 5　池化单元电路

2.4　全连接模块设计

全连接层作用是建立能谱数据特征和输出类别之间的联系，逐层降低数据维度进行分类，整个模型参数量主要集中在全连接层，每一层特征数据需要将全部数据映射到下一层神经元中。与卷积计算类似，单层全连接计算中有大量的乘加操作，和卷积层采用同样的策略，使用并行流水线进行加速。为减少单个循环体的计算周期和降低 FPGA 综合时的时序违例，提高工作频率，采用加法树对全连接并行计算进行优化。加法树是一种多级级联的加法器结构，能够同时执行多个加法操作，如图 6 所示，整个结构类似于二叉树，计算时先将权重和特征数据从数据缓冲区中读出，计算乘积使用寄存器保存。在下一个时钟周期内，将寄存器中的乘积值读出再两两相加，上一层相加结果作为下一层输入，依次递归，实现流水。在最后一个累加单元，加上 bias 值。

在训练时，CNN 最后一层通过 Softmax 函数输出各核素分类的概率值，Softmax 将网络的运算结果映射到[0, 1]之间的数，其和为 1，是单调递增函数。但 Softmax 涉及大量的指数运算，为了减少计算时间，在硬件中，使用比较器代替求概率的过程，全连接最后一层输出最大值对应的核素，即为核素识别结果。

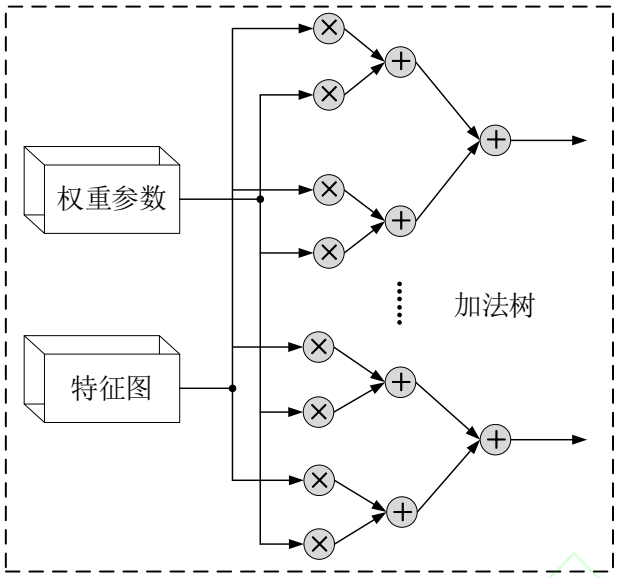

图 6　全连接计算加法树结构

2.5　高层次综合结果

IP 核采用高层次综合（High-Level Synthesis, HLS）的开发方式进行实现。HLS 工具可直接将 C/C++高级语言转化为 Verilog 语言，降低了开发难度，综合为 IP 核，在 FPGA 开发中直接调用。在 HLS 中，使用优化指令，对卷积、池化和全连接计算进行优化。

如图 7 所示，IP 核共提供 6 个接口，采用 AXI_Lite 协议控制寄存器配置，通过 AXI 总线与外界进行数据交互，s_axi_CONTROL 接口为 AXI_Lite 数据传输协议控制接口。ap_clk 为 IP 核工作提供时钟信号。ap_rst_n 为 IP 核提供复位信号，采取的是低电平复位方式，m_axi_FEATURE 接口接收能谱数据，内部计算得出的各核素预测值，通过 m_axi_OUT_r 接口输出。interrupt 输出中断信号。

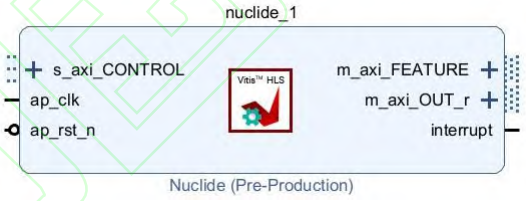

图 7　核素识别 IP 核

3　实验结果及分析

3.1　实验平台

实验采用 Xilinx 公司的 ZYNQ XC7Z020 异构芯片作为部署平台。以 ARM 硬核为主控，通过 AXI Interconnect 控制模块将核素识别 IP 核与 ARM 处理器相连接。开发环境采用 Xilinx Vivado 2022.2 套件，使用的编程语言包括 Python、C/C++、Verilog。CPU 平台为 Intel i5-9600KF，GPU 平台为 NVIDIA GTX2060，ARM 为 Cortex-A9 内核。

3.2　硬件精度评估

利用混淆矩阵给出测试集在 FPGA 中的分类结果。采用准确率(Accuracy, Acc)、召回率(Recall, R)、精确率(Precision, P)和 *F1* 分数(F1-score)指标对模型性能进行评估。其计算公式如下：

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$\mathrm{Recall}(R)=\frac{TP}{TP+FN} \tag{5}$$

$$\mathrm{Precision}(P)=\frac{TP}{TP+FP} \tag{6}$$

$$F1-Score=\frac{2\times P\times R}{P+R} \tag{7}$$

其中，*TP* 指正样本被分类为正样本的数量，*TN* 指负样本被分类为负样本的数量，*FP* 指负样本被分类为正样本的数量，*FN* 表示正样本被分类为负样本的数量。如图 8 所示，由于计算误差，测试集在 FPGA 中准确率为 98.41%，相比下降 0.07%。各类别召回率、精确率和 F1 分数如表 3 所示。由表 3 可知，1D-CNN 模型在 FPGA 中分类精确率均在 95%以上，*F1* 分数均高于 0.97，证明了模型在 FPGA 中进行分类的有效性。

图 8　模型在 FPGA 中分类结果混淆矩阵

表 3　模型在 FPGA 中评估指标

核素	召回率/%	精确率/%	F1分数
^{60}Co	100	96.36	0.9814
^{60}Co+^{137}Cs	99.33	95.10	0.9717
^{60}Co+^{137}Cs+^{152}Eu	96.22	100	0.9807
^{60}Co+^{152}Eu	97.56	98.43	0.9799
^{137}Cs	98.67	99.33	0.9899
^{137}Cs+^{152}Eu	98.44	100	0.9921
^{152}Eu	98.67	100	0.9932

3.3　实际测试

在实验室中搭建测试平台，使用数字化多道谱仪采集数据，通过串口将能谱发送至 FPGA 进行识别。以探测距离为变量，将 ^{137}Cs 单源和 ^{137}Cs+^{60}Co 混合源分别放置在距离探测器 5cm 至 30cm 处，每间隔 5cm 测量 200 组能谱数据进行测试，测量时间在 1s 至 60s 内呈均匀分布，测试结果如图 9 所示。

在 5cm 处，200 组数据均能得到 100%的准确率，随着距离的增加，探测器统计的能量开始减少，CNN 识别性能开始下降，在 30cm 处，^{137}Cs 的识别准确率下降到 93.5%，整体识别率为 98.0%，^{137}Cs+^{60}Co 混合源识别率下降至 95%，整体识别率为 98.17%。随着距离增加，环境本底噪声在能谱中占比提高，也会导致识别准确率下降。

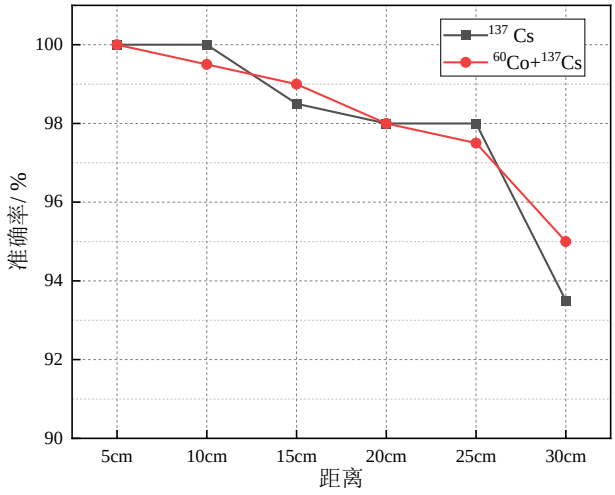

图 9　不同距离的识别效果

3.4　硬件性能评估

在 Zynq 中搭建测试平台，以 PS 为主机，PL 为从机。设置 IP 核工作时钟频率为 100MHz，HLS 综合和上板综合前后 IP 核资源占用率如图 9 所示。硬件加速采用以资源换时间的策略，在 FPGA 中进行布局布线综合后，计算中乘法消耗了 72%的 DSP，片上缓存数据消耗了 74%的 BRAM，时序逻辑电路和组合逻辑电路使用了 29%的查找表(Look Up Table,LUT)。相比 HLS 综合结果，经过 Vivado 编译器优化后的查找表资源下降 40%，由 409 个存储密度更高的 LUT RAM 构成存储器，寄存器(Flip Flop, FF)资源消耗了 12%。Zynq7020 内部资源均能满足实验要求。

对比 1D-CNN 核素识别算法在 FPGA、CPU、GPU 和 ARM 平台的识别效果，分别记录每个平台模型单次推理时间，如表 4 所示，在 FPGA 中 1.57ms 即可完成一次模型的推理，相对于 CPU 有 64 倍的加速效果。使用能效比表示相同时间模型执行效率与能耗的比值，数值与执行效率成正比，结果显示 FPGA 能效比是 GPU 平台的 94 倍。相比之下，ARM 端执行效率远不及 FPGA。表明 FPGA 在保证识别精度的前提下，在移动端中的识别速度和能耗拥有巨大优势，具有一定的现实意义，对便携式核素识别仪器开发具有重要参考价值。

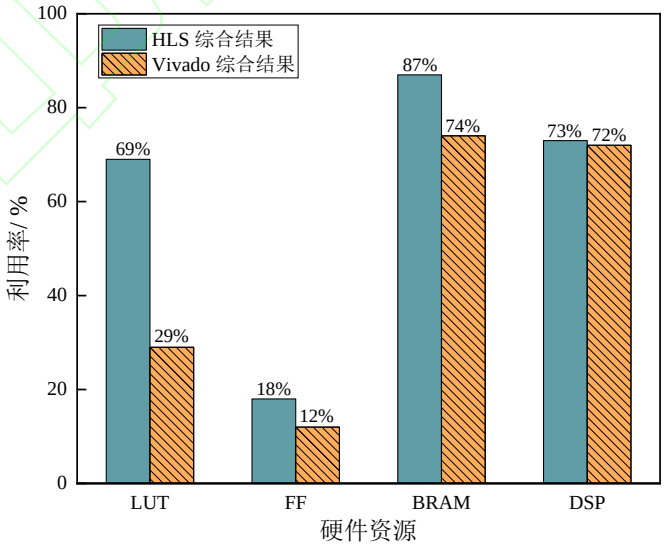

图 10　FPGA 硬件资源使用率

表 4　模型在不同硬件平台性能对比

硬件平台	CPU	GPU	ARM	FPGA

型号	i5-9600kf	GTX2060	Cortex-A9	ZYNQ7020
时钟频率	3.1GHz	1365MHz	666MHz	100MHz
识别耗时/ms	100.70	1.97	2496	1.57
帧率/fps	9.93	507	0.4	636
功耗 /W	95	160	1.9	2.115
能效比	0.1	3.17	0.21	301

4 结论

针对传统核素识别算法分析步骤繁琐，识别时间长等问题，本文基于 1D-CNN 的核素识别算法，利用 FPGA 芯片提出了一种硬件加速方法，经过实验和对比测试，得出以下结论：

（1）针对 γ 放射源的 1D-CNN 轻量化核素识别模型在测试集中准确率达到了 98.48%。可直接输入能谱数据，得出核素种类，简化了核素定性分析的步骤。

（2）在 FPGA 中，采用并行流水线和加法树等方式对模型推理过程进行加速，模型单次推理仅需 1.57ms，相对于 CPU 加速 64 倍，功耗为 2.115W，满足了快速识别，低功耗的要求，适合移动端应用。

本研究为核素识别仪器的研制提供了新的开发思路。在未来，可以结合能谱采集和可视化界面，利用 FPGA 芯片，开发出一种基于深度神经网络的便携式核素识别仪器。

参考文献：

[1] Bernstein A, Bowden N, Goldblum B L, et al. Colloquium: Neutrino detectors as tools for nuclear security[J]. Reviews of Modern Physics, 2020, 92(1): 011003.

[2] Ohba T, Tanigawa K, Liutsko L. Evacuation after a nuclear accident: Critical reviews of past nuclear accidents and proposal for future planning[J]. Environment international, 2021, 148: 106379.

[3] Li C, Liu S, Wang C, et al. A new radionuclide identification method for low-count energy spectra with multiple radionuclides[J]. Applied Radiation and Isotopes, 2022, 185: 110219.

[4] 汤彬, 葛良全, 方方等著. 核辐射测量原理[M]. 哈尔滨: 哈尔滨工程大学出版社, 2011.

[5] Li X Z, Zhang Q X, Tan H Y, et al. Fast nuclide identification based on a sequential Bayesian method[J]. Nuclear Science and Techniques, 2021, 32(12): 143.

[6] Ling Y, Huang T, Yue Q, et al. Improving the estimation accuracy of multi-nuclide source term estimation method for severe nuclear accidents using temporal convolutional network optimized by Bayesian optimization and hyperband[J]. Journal of Environmental Radioactivity, 2022, 242: 106787.

[7] Qi S, Zhao W, Chen Y, et al. Comparison of machine learning approaches for radioisotope identification using NaI (TI) gamma-ray spectrum[J]. Applied Radiation and Isotopes, 2022, 186: 110212.

[8] 张江梅, 任俊松, 李培培, 等. 基于支持向量机的复杂核素能谱识别[J]. 核电子学与探测技术, 2016, 36(08): 856-861.

[9] El_Tokhy M S. Rapid and robust radioisotopes identification algorithms of X-Ray and gamma spectra[J]. Measurement, 2021, 168: 108456.

[10]Van Hiep C, Hung D T, Anh N N, et al. Nuclide Identification Algorithm for the Large-Size Plastic Detectors Based on Artificial Neural Network[J]. IEEE Transactions on Nuclear Science, 2022, 69(6): 1203-1211.

[11]Liang D, Gong P, Tang X, et al. Rapid nuclide identification algorithm based on convolutional neural network[J]. Annals of Nuclear Energy, 2019, 133: 483-490.

[12]Wang Y, Yao Q, Zhang Q, et al. Explainable radionuclide identification algorithm based on the convolutional neural network and class activation mapping[J]. Nuclear Engineering and Technology, 2022, 54(12): 4684-4692.

[13]杜煜章, 潘家华, 宗容等. 基于硬件加速的轻量级网络心音分类器[J]. 计算机工程与应用, 2021, 57(23): 263-269.

[14]Wang W, Zhao X, Liu D. Design and Optimization of 1D-CNN for Spectrum Recognition of Underwater Targets[J]. Integrated Ferroelectrics, 2021, 218(1): 164-179.

[15]Huang X, Jones E, Zhang S, et al. An FPGA implementation of convolutional spiking neural networks for radioisotope identification[C]//2021 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE,2021:1-5.

[16]曾国强, 欧阳晓平, 喻明福, 等. 手持式单板 500MHz 采样率数字化多道设计[J]. 核技术, 2017, 40(03): 31-37.

[17]王瑶,刘志明,万亚平等. 基于长短时记忆神经网络的能谱核素识别方法[J]. 强激光与粒子束, 2020, 32(10): 154-161.

[18]Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017.

[19]Liu B, Zou D, Feng L, et al. An FPGA-based CNN accelerator integrating depthwise separable convolution[J]. Electronics, 2019, 8(3): 281.

[20]Basalama S, Sohrabizadeh A, Wang J, et al. FlexCNN: An End-to-end Framework for Composing CNN Accelerators on FPGA[J]. ACM Transactions on Reconfigurable Technology and Systems, 2023, 16(2): 1-32.

[21]孙敬成, 王正彦, 李增刚. 卷积神经网络数字识别系统的 FPGA 实现[J]. 计算机工程与应用, 2020, 56(13): 181-188.

Hardware acceleration method of convolutional neural network nuclide identification algorithm based on FPGA

WANG Bo，SHI Rui，LIU Minjun，ZENG Xiong，WANG Zhou，

(Sichuan University of Science & Engineering，School of Science and Engineering，Yibin SiChuan 644005，China)

Abstract：Nuclide identification is one of the key techniques researched in the field of nuclear detection. Traditional nuclide identification instruments based on energy spectrum analysis algorithms have poor real-time performance and high power consumption, which limit the identification efficiency in practical applications. This work proposes an FPGA-based convolutional neural network hardware acceleration method for nuclide identification to accelerate the qualitative analysis of radionuclides. Firstly, a lightweight one-dimensional convolutional neural network model for nuclide classification is constructed, and then deployed in a Xilinx ZYNQ7020 heterogeneous chip using hardware acceleration strategies such as parallel pipelines and adder trees according to the computing features of the convolutional, pooling and fully connected layers of the model. The experimental results show that the average recognition accuracy of the model in FPGA reaches 98.41%, and the single recognition takes only 1.57ms. Compared with the desktop CPU, this hardware acceleration method achieves 64 times acceleration effect, and the power consumption is only 2.115W. Measured at a distance of 30 cm, the identification accuracy of ^{137}Cs single source is 98%, and the identification accuracy of ^{137}Cs and ^{60}Co mixed source reaches 98.17%. This hardware acceleration method satisfies the requirements of low latency and low power consumption, and is suitable for the scene of fast nuclide detection, and is of reference value for the development of portable nuclide recognition instruments.

Key words：Energy spectrum data；Nuclide identification；Field programmable gate array；Convolutional neural network；Hardware acceleration