引用格式：程筱舒, 王忆文, 娄鸿飞, 等. 基于位串行卷积神经网络加速器的运动想象脑电信号识别系统 [J]. 电子科技大学学报, 2025, 54(3): 321-332.
CHENG X S, WANG Y W, LOU H F, et al. A motor imagery EEG signal recognition system based on a bit-serial convolutional neural network accelerator[J]. Journal of University of Electronic Science and Technology of China, 2025, 54(3): 321-332.

基于位串行卷积神经网络加速器的运动想象脑电信号识别系统

程筱舒，王忆文*，娄鸿飞，丁玮然，李 平

（电子科技大学 集成电路科学与工程学院（示范性微电子学院），成都 611731）

摘要：准确识别运动想象脑电信号是神经科学和生物医学工程领域的重要挑战。设计了基于位串行卷积神经网络加速器的脑电信号识别系统，充分利用其小体积、低能耗和高实时性的优势。从软件层面，介绍了脑电数据的预处理、特征提取及分类过程，并采用格拉姆角场转换将一维信号映射为二维特征图供网络处理。在硬件层面，提出了列暂存数据流和固定乘数原位串行乘法器等方法，在 FPGA 上实现了位串行卷积神经网络加速器的原型验证。实验表明，基于位串行 LeNet-5 加速器的 FPGA 实现对 BCI 竞赛 IV 数据集 2a 和 2b 的分类平均准确率分别达到 95.68% 和 97.32%，kappa 值分别为 0.942 和 0.946，展现出的优异性为运动想象脑电信号识别的高效实现提供了思路。

关键词：脑机接口；运动想象；卷积神经网络；硬件加速器；格拉姆角场
中图分类号：TN4 **文献标志码：**A DOI: 10.12178/1001-0548.2024145

A motor imagery EEG signal recognition system based on a bit-serial convolutional neural network accelerator

CHENG Xiaoshu, WANG Yiwen*, LOU Hongfei, DING Weiran, and LI Ping

(School of Integrated Circuit Science and Engineering (Exemplary School of Microelectronics),
University of Electronic Science and Technology of China, Chengdu 611731, China)

Abstract: Accurate recognition of motor imagery electroencephalogram (EEG) signals is a significant challenge in neuroscience and biomedical engineering. This paper presents an EEG signal recognition system based on a bit-serial convolutional neural network (CNN) accelerator, leveraging its advantages of compact size, low power consumption, and high real-time performance. The software implementation includes the preprocessing, feature extraction, and classification of EEG data, and utilizes Gramian angular field (GAF) transformation to map one-dimensional signals into two-dimensional feature maps for network processing. On the hardware side, innovative methods such as column-buffering dataflow and fixed-multiplier bit-serial multiplication are proposed, and a prototype of the bit-serial CNN accelerator is successfully implemented on FPGA. The results show that the FPGA implementation of the bit-serial LeNet-5 accelerator achieves average classification accuracies of 95.68% and 97.32% on the BCI Competition IV datasets 2a and 2b, with kappa values of 0.942 and 0.946, respectively. These performances provide an efficient solution for the recognition of motor imagery EEG signals.

Key words: brain computer interface; motor imagery; convolutional neural network; hardware accelerator; Gramian angular field

运动想象（motor imagery, MI）脑电信号（electroencephalograph, EEG）是最为常见的脑机接口（brain computer interface, BCI）应用之一，已被广泛应用于智能医疗保健领域，如中风恢复治疗和移动辅助机器人的开发等方面。MI EEG 信号复杂，具有高维结构。近年来，深度学习对基于 MI EEG 的脑机接口产生了较大影响。因此，需要先进的机器学习和深度学习算法来处理和解码这些复杂的大脑数据。

在软件算法方面，文献 [1] 探讨了使用 CNN 对多类脑电信号进行分类。文献 [2] 提出了基于分层相关性传播（layer-wise relevance propagation,

收稿日期：2024−06−19
作者简介：程筱舒，博士生，主要从事神经网络加速器和脑机接口等方面的研究。
*通信作者 E-mail: yiwen@uestc.edu.cn

LRP）的深度神经网络在脑电数据分析中的应用。通过 LRP，单次试验 DNN 决策被转换成热图，表明每个数据点与决策结果的相关性。DNN 达到了较佳的分类精度，证明了深度神经网络是一个强大的非线性 EEG 分析工具。

在深度学习网络的结构创新方面，文献 [3] 设计了一个单一的 CNN 架构来准确地分类来自不同脑机接口范式的脑电信号，即名为 EEGNet 的神经网络结构，可以在一系列脑机接口任务中学习各种各样的可解释特征。也有从更高维度建立脑电分类框架的工作，如文献 [4] 提出了新的脑电分类框架，包括一种新的脑电三维表示、一种多分支三维 CNN 以及相应的分类策略。该框架在只有 9 个采样电极的情况下也表现出了良好的性能，提高了其实用性。

在硬件实现方面，由于 EEGNet 模型的紧凑性，以及对脑电信号出色的分类预测能力，文献 [5-7] 对其在 FPGA 上进行了硬件实现。但其效率不够，文献 [8]进一步优化了先前的工作，用 ASIC 方法实现了 EEGNet。

本文在软件算法方面，使用格拉姆角场转换，将一维的脑电信号转化成二维的特征图，输入到卷积神经网络中进行运动想象脑电信号的识别与分类；在硬件实现方面，使用列暂存数据流、固定乘数原位串行乘法器等方法实现位串行卷积神经网络加速器；在应用方面，将位串行卷积神经网络加速器在 FPGA 上实现，并对 MI EEG 数据集进行识别与分类。

1 理论基础

1.1 运动想象脑电信号识别

EEG 是一种从人类大脑获取的生物特征数据，能够反映出用户的心理或身体状态，并可用于深入的信息分析。具体的 BCI 系统设计框架如图 1 所示。

BCI 系统将大脑的活动模式解读为可用信息或命令，以此与外部世界沟通。此过程首先包括 EEG 信号的采集，随后对这些信号进行识别，最后将识别结果用于数据分析或传输给控制系统；控制系统的响应会反馈给大脑，从而形成一个新的判断循环，并再次进入信号处理流程。

根据信号采集的方法，运动想象识别技术一般分为侵入式、半侵入式和非侵入式 3 种。这 3 种方法分别通过局部场点位（local field potential, LFP）、皮层脑电图（electrocorticogram, ECoG）和 EEG 来采集信号。EEG 采用非侵入式方法，避免了对用户的身体损害。虽然 EEG 在空间分辨率上可能不如其他两种方法，但它具有较高的时间分辨率。因此，在运动想象识别的应用中，EEG 是最常用的技术。

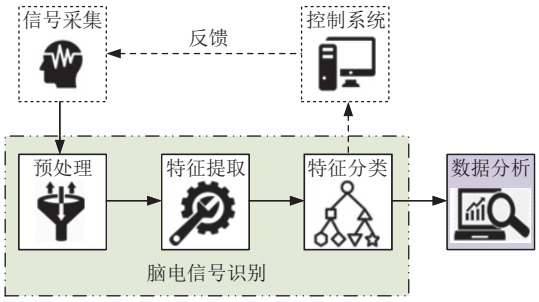


图 1 BCI 系统设计框架

提取和分类 EEG 特征面临多种挑战。首先，EEG 信号的幅度非常微小；其次，EEG 信号具有很强的随机性，是一种非平稳信号，易受到各种生物学和环境因素的干扰[9]；此外，EEG 信号的信噪比（signal to noise ratio, SNR）仅约为 5%[10]。且脑电波主要包含 4 个关键的波段：delta 波、theta 波、alpha 波和 beta 波。这些因素共同构成了 EEG 信号处理和解析的复杂性。

1.2 卷积神经网络

受自然界生物大脑的影响，并在神经网络研究有所进展后，卷积神经网络（convolutional neural networks, CNN）作为先进的深度学习算法，在图像分类、语音处理和对象识别等方面取得了出色的表现，因此受到了广泛关注。相对于多层感知机，CNN 的计算更为复杂，数据存储量巨大，需处理更高维的数据。CNN 的核心原理是通过神经网络找到一种将输入数据映射到预期输出的函数。同时考虑到统计效率和网络的可训练性，CNN 被设计成包含多种特定功能的层，如卷积层用于特征提取，池化层用于降维，以及全连接层用于最终的决策输出[11]。在实际应用中，这些层按照特定的顺序和结构组合在一起，形成一个完整的 CNN 模型。一个典型的 CNN 的基本结构如图 2 所示。

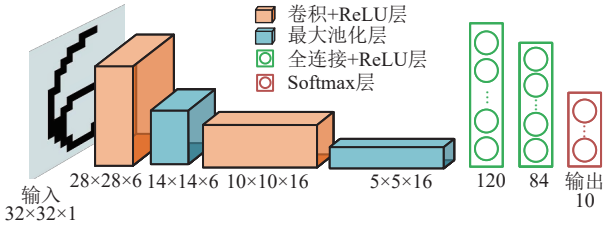


图 2 LeNet-5 卷积神经网络结构

1.3　位串行数字表示及时序

使用补码来表示定点二进制的有符号数据已经是数字系统中的一种常见方法，这种表示法能有效处理正数和负数的运算，使数字的加法和减法操作更为简便。这里只定义两种数据类型：单精度数据和双精度数据。假设有 P 位数据 $X=(x_{P-1}x_{P-2}\cdots x_1x_0)$，其表示为[12]：

$$X=\begin{cases}-x_{P-1}+\sum_{i=0}^{P-2}x_i2^{i-P+1} & \text{（单精度）}\\ -x_{2P-1}+\sum_{i=0}^{2P-2}x_i2^{i-2P+1} & \text{（双精度）}\end{cases} \tag{1}$$

当数据精度 P=4 时，数据位构成和时序如图 3 所示。数据中的小数点的位置位于最高有效位（the most significant bit, MSB）和次高有效位之间。数据的最低有效位（the lowest significant bit, LSB）会优先于其他位进行传输。同时还生成控制输出数据的控制头位。

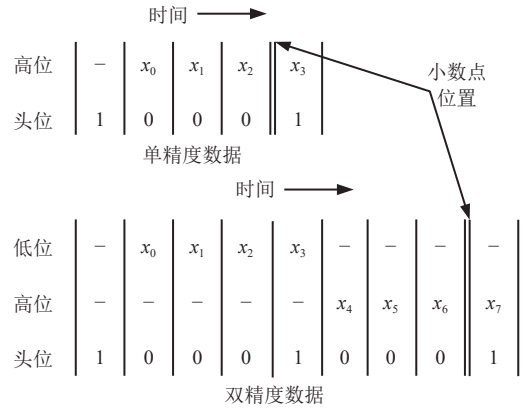


图 3　不同数据类型的数据位构成和时序

2　MI EEG 识别系统

使用的软件开发平台有开源脚本语言 python、Facebook 人工智能研究院研发的开源机器学习框架 PyTorch 和处理脑电信号数据的 MNE 库等，且使用了 Xilinx 公司的 Vivado 设计套件。硬件开发平台为 Xilinx 公司的 Kintex-7 FPGA，核心芯片型号为 XC7K325TFFG900-2。

2.1　MI EEG 识别算法流程

对于图 1 中脑电信号识别部分，主要包括预处理、特征提取和特征分类 3 大步骤。后续的数据分析部分主要包括量化和 FPGA 推理数据集。运动想象脑电信号识别算法流程如图 4 所示。

由于 EEG 信号具有低信噪比的特性，通常在进行神经网络分类前，需要对信号执行一系列预处理以提升其信噪比。这些预处理包括多个环节，如滤波、剔除坏道和选段等。

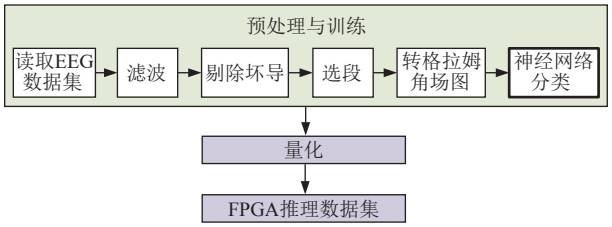


图 4　运动想象脑电信号识别算法流程

特征提取是从输入信号中提取具有区分性特征的过程，依赖于特定领域的知识，通常提取时域特征、频域特征或时频域特征。随着计算机视觉和神经网络技术的快速发展，传统的信息处理方法需要一种能够快速应用于信号处理的新思路，如将一维时间序列映射为二维图像，且这个过程不需要太复杂的优化。这种需求催生了格拉姆角场（Gramian angular field, GAF）方法[13]。GAF 能够将时间序列数据转换为图像数据，既保留了信号的完整信息，又维持了信号对时间的依赖性。将信号数据转换为图像数据后，可利用 CNN 在图像分类和识别方面的优势进行标签分类和其他处理。

特征分类是一种机器学习算法，它的作用是将提取出的特征转化为外部设备能够识别的逻辑控制信号。由于传统的预处理和特征提取方法在处理时间和信息丢失方面存在风险，难以克服低信噪比的问题。如图 5 所示，深度学习算法[14-16] 已被证明比传统的如 SVM 和 KNN 等分类器[17-20] 更为强大。深度学习有两个主要优势：首先，它可以直接作用于原始的大脑信号，从而避免了耗时的预处理和特征提取过程；其次，深度神经网络通过其深层结构能够捕捉到具有代表性的高级特征和潜在的依赖关系。尽管多分类器[21-22] 的平均准确率略有优势，但 CNN 的结构复杂度并不高，同时也能达到与多分类器相当的平均准确率[23]。

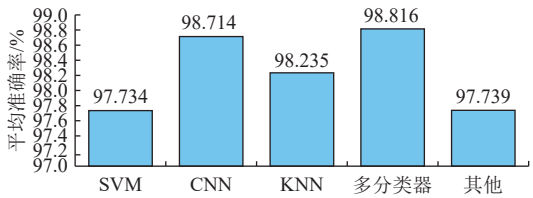


图 5　几种主要的分类器的平均准确率

在保持预处理和特征提取方法一致的基础上对 LeNet-5 神经网络进行量化，并利用 Torch 的 Hook 机制来获取每个层级的权重参数和输入输出特征图的参数。Hook 机制包含两个核心部分：定义一个用于处理特征图的函数和注册 Hook，即指定模型在哪些层上使用这个定义的函数来处理特征图。

为了实现和评估位串行卷积神经网络加速器在运动想象脑电信号识别上的应用效果，在软件方面对数据集进行训练，并在量化后将得到的输入特征图（input feature map, IFM）和权重输出给硬件。

2.2 预处理与训练

预处理部分，对于 BCI 竞赛 IV 数据集 2a，总共有 288 个样本，25 个通道和 751 个采样点。而 BCI 竞赛 IV 数据集 2b，有 120 个样本，6 个通道和 751 个采样点。在调用 MNE 库读取第一个受试者的 EEG 采样信号后，需进行预处理操作，主要是滤波、剔除坏道和选段。由于脑电信号 4 个主要波段在 1～32 Hz，因此采用带通滤波器滤波。对于如眼电这种小而稳定的伪迹，可以采用伪迹拒绝的方法，将眼电信号通道剔除。在采集样本的流程中，只有运动想象阶段对于后续的分类处理有用，因此需要将剩余通道的信号进行选段，截取有用的部分。

特征提取部分，GAF 转换允许将原本主要用于 CNN 等图像处理的强大工具和技术，应用于时间序列数据。这种转换特别适用于 EEG 数据，因为 EEG 数据本质上是按时间顺序记录的一系列数据点。如图 6 所示，转换的基本步骤是：首先将 EEG 数据中的每个数据点映射到坐标空间中的一个点；再计算这些点之间的角度，并将这些角度值转换成像素强度，从而生成能够代表原始 EEG 数据的图像。

对于特征分类部分，由于数据集 2a 和 2b 分别对应的分类数不同，因此需要对原始的 LeNet-5 网络进行结构微调。数据集 2a 是四分类的数据集，所以最后一层的网络输出维度为 4，该层对应的权重维度由原来的 10×84 变为 4×84。对于二分类的数据集 2b，最后一层的输出维度是 2，该层的权重维度为 2×84。将每个样本生成相应的 GFA 图输入 CNN 中进行分类。数据集 2a 的 9 个受试者，每个受试者有 288 个样本，有 4 种不同标签的 GAF 图。数据集 2b 同样是 9 个受试者，每个受试者约有 120 个样本，有两种不同标签的 GAF 图。将生成的每个类别的 GAF 图输入微调后的 LeNet-5 网络中进行分类识别。

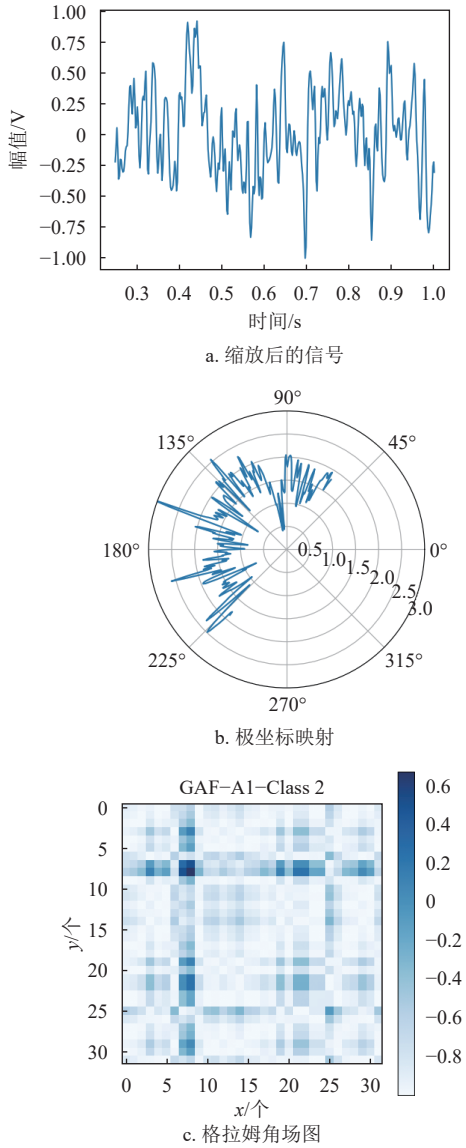


图 6 格拉姆角场变换

2.3 参数量化

量化是一种在较低位宽的条件下对张量进行计算和存储的方法，这不仅有助于减少深度学习模型所需的内存和存储空间，还能提升计算速度，尤其是在模型推理阶段[24-27]。在实现 CNN 模型量化的过程中，选择了使用 PyTorch 提供的一套功能强大的量化应用程序编程接口（application programming interface, API）[28]。PyTorch 作为一个广泛使用的深度学习框架，其量化工具为优化模型提供了一个高效且灵活的方式。与常规的 32 位浮点（32-bit floating point, FP32）模型相比，PyTorch 的量化工具能够将模型转换为 8 位有符号整数

（signed 8-bit integer, INT8）格式，这种转换不仅大幅度减小了模型的存储大小，还使内存带宽的需求减少到原来的四分之一。

在 FPGA 上，为了简化张量累加后的量化过程并使其符合硬件量化的规则，量化比例的值最好是 2 的幂。这样，可通过简单的数据右移操作来替代复杂的除法运算。在此，选择 24 位串行数据的第 23 位作为最高位符号位，以及第 14 至第 8 位的数来构成硬件量化后的数值。因此，量化比例的值被定为 256，其他位数被舍弃。同时，需确保新构成的数据值位于−128～+127 的范围内。这样的处理方式旨在简化数据在 FPGA 上的处理，同时保证量化过程的有效性和准确性。

采用 INT8 格式量化的主要原因是其在硬件实现中具有能耗低、延迟短的优点。为评估其对模型准确度的影响，对比了浮点精度和 INT8 精度的实验结果，发现 INT8 格式的量化在 BCI 竞赛 IV 数据集 2a 和 2b 上的精度下降均不超过 0.5%。此外，由于 EEG 信号具有低频、低动态范围的特性，实验表明 INT8 格式的量化足以表达信号的特征信息。

2.4　FPGA 推理

在确保使用了相同的预处理和特征提取方法的条件下，进行了 FPGA 数据集推理。这个过程首先包括将每层经过量化的权重和特征图参数转换为二进制形式，然后将这些二进制数据再转化成 COE（coefficient）文件，且这些文件被配置到加速器的 RAM 中。在这个过程中，Vivado 工具会解析 COE 文件的格式，并在生成 IP 核时导出相应的 MIF（memory initialization file）格式文件。这些 MIF 文件随后被用于行为级的仿真，以确保整个系统的准确性和有效性。

3　位串行卷积神经网络加速器

3.1　列暂存数据流

为了优化数据传输以更好地适应位串行电路的特性，提出了列暂存（column buffering, CB）数据流来高效计算和减少计算复杂度。具体来说，CB 数据流主要在权重固定（weight stationary, WS）数据流的基础上进行改进，除权重外的 IFM 能够尽可能地被复用。具体实现上，CB 数据流在输入特征图的列缓存中存储一列数据，然后依次用这一列数据进行计算，直到完成该列的所有权重更新或激活函数计算。这一设计特别适用于深度学习位串行硬件加速器。通过减少数据传输和复用输入特征图，CB 数据流不仅减少了数据传输的次数，还降低了缓存失效的风险。

图 7 中，假设有一维的 IFM，一个长度为 K 的一维卷积核在 IFM 上以步长为 1 的距离卷积。根据卷积公式，可得一维的输出特征图（output feature map, OFM），由 S_1、S_2、S_3 等部分构成。一维的处理单元（processing element, PE）由长度为 K 的乘累加器（multiply accumulator, MAC）串联而成。卷积可以由 CB 数据流在一维的 PE 中体现，其伪代码如下。

算法 1　一维列暂存数据流

输入：K 是卷积核的大小，L 是输入特征图大小，C 是输出特征图大小，i 是输入特征图的序列，n 是部分输出特征图序列，C_i 是第 i 列输入特征图

输出：S_i 是第 i 个部分输出特征图。

过程：

$i = n = 1$，$S_n = 0$
for $i = 1, 2, 3, \cdots, L$ do
　$\mathrm{MAC}_j = C_i$，$j \in [1, \min(i, K)]$
　$S_n = S_n + C_i \times W_{i-n+1}$，$n \in [\max(1, i-K+1), i]$
　if i-n+1 = K then
　　output S_i
　end if
end for

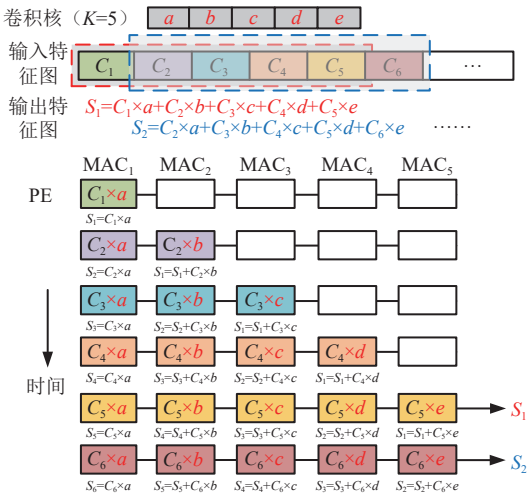


图 7　一维的 CB 数据流

二维 PE 中 CB 数据流的情况是一维的扩展，如图 8 所示。对于 IFM，数据朝列扩展 K 行，而权重按顺序固定在 PE 中。这部分和 S 成为了 IFM 的一列数据 C 分别和权重的乘积之和的累加。二维的 CB 数据流工作步骤如下。

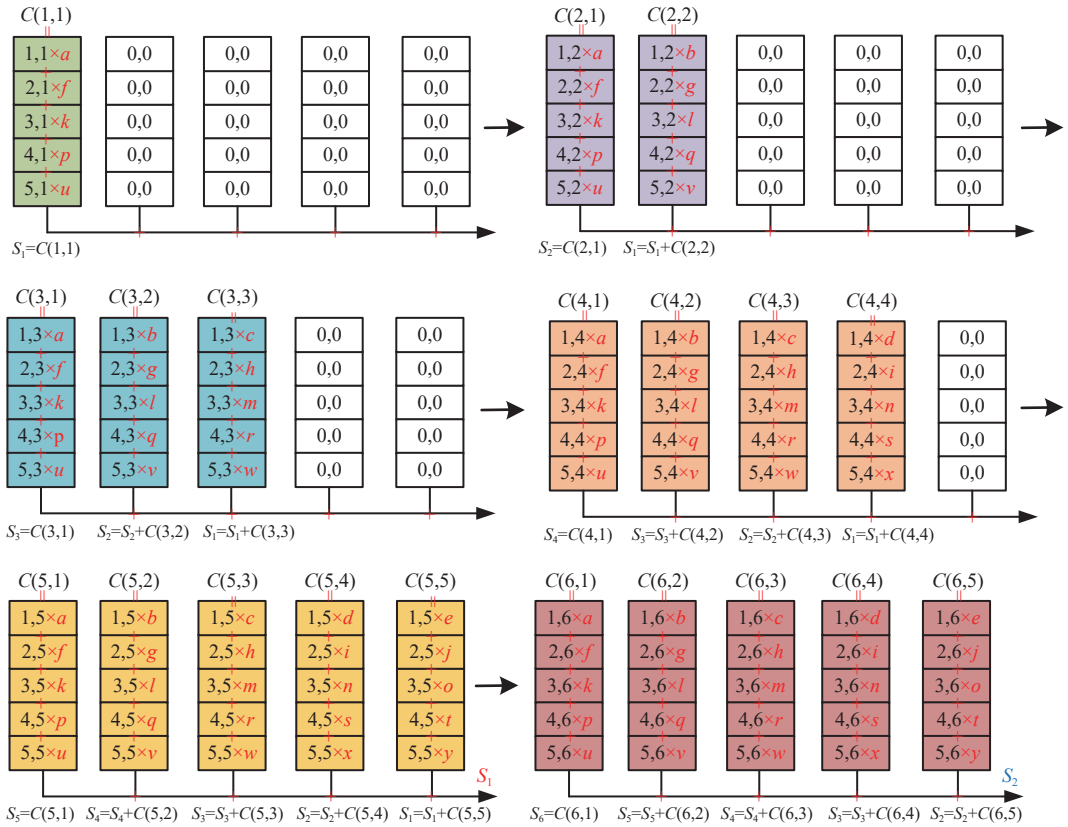


图 8 二维的 CB 数据流

1） IFM 的第 1 列数据 C（1,1）进入第 1 列 MAC，和第 1 列的权重分别相乘后相加，得到第 1 个中间结果 S_1；

2） IFM 的第 2 列数据 C（2,1）和 C（2,2）同时进入第 1 列和第 2 列 MAC，分别和第 1 列和第 2 列的权重相乘后相加，得到第 2 个中间结果 S_2，和第 1 个累加的中间结果 S_1；

3） 按前两个步骤的规律加载第 3、第 4 列数据，得到中间结果 S_1～S_4；

4） IFM 的第 5 列数据 C（5,1）～C（5,5）同时进入第 1 列和第 5 列 MAC，分别和第 1 列～第 5 列的权重相乘后相加，得到中间结果 S_1～S_5。由于 S_1 此时累加次数已经达到 5 次，即完成 OFM 的第 1 个部分和的累加，S_1 输出 PE；

5） IFM 的第 6 列数据 C（6,1）～C（6,5）同时进入第 1 列和第 5 列 MAC，分别和第 1 列～第 5 列的权重相乘后相加，得到中间结果 S_2～S_6。S_2 在 S_1 之后也输出 PE。

3.2 加速器整体架构与工作流程

为了提高神经网络的计算速度和效率，采用了硬件加速方式。本文提出的位串行 CNN 加速器的系统结构如图 9 所示。

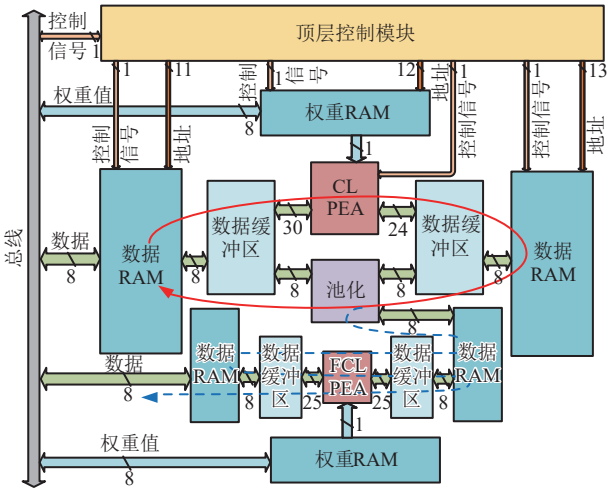


图 9 位串行 LeNet-5 加速器系统结构

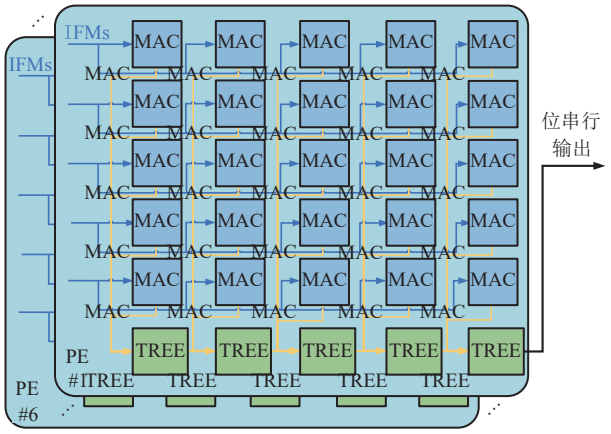
加速器的核心是两个 PE 阵列（processing element array, PEA）计算模块：一个用于处理卷积层（convolutional layer, CL），称为 CL PEA；另一个处理全连接层（fully connected layer, FCL），称为 FCL PEA。CL PEA 由 6 个 PE，即 150 个 MAC 和加法器树构成；FCL PEA 由 25 个 MAC 和加法器树构成。其余则是装有输入特征图、输出特征图和权重的 6 个随机存取存储器（random access

memory, RAM）以及相应的缓冲器、池化模块以及顶层控制模块。总线用于在系统和外部存储单元传输数据。池化模块用于执行池化操作，可以压缩数据量大小。

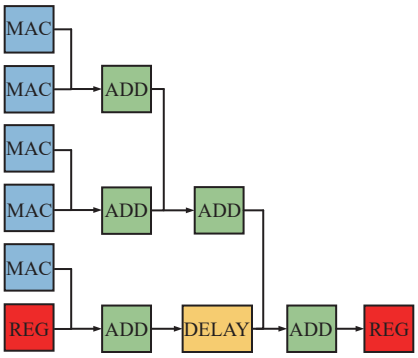
当系统启动时，顶层控制模块首先初始化和配置所有模块。数据通过总线被加载到数据 RAM 和权重 RAM 中。当执行神经网络前向传播时，数据和权重被送到 CL PEA 进行计算。每个 PE 单元并行处理数据，执行乘累加操作。计算完成的中间结果先暂存在数据缓冲区中，然后进行量化和 ReLU 函数处理，将处理完成后的数据送入数据 RAM 中存储，再将数据重新读出，经过池化模块进行池化操作，即图 9 中的红色实线为第 1 层卷积的数据移动通路。最终，第 2 层池化完处理后的数据可以通过内部数据传输线传输到全连接层的数据 RAM，全连接层使用 FCL PEA 对数据矩阵进行处理。最后一层全连接层处理完的数据通过总线输出，即图 9 中的蓝色虚线部分为第 2 层卷积后的 3 层全连接层的数据移动通路。

3.3　PEA 的设计

PEA 的结构如图 10a 所示，由 6 个 PE 构成。PE 由 MAC 单元和 TREE 组成，MAC 单元用于执行乘法，TREE 将乘法出来的结果进行累加操作。输入的 IFM 数据以 K×1 为单位划分列，从第一列开始逐列以 Z 字形更新数据，经过多个 MAC 单元进行处理，具体流程参看对 CB 数据流的描述。然后在 PE 的底部有几个 TREE 结构。这些结构用于汇总从 MAC 单元传出的数据，并将其输出传递到下一个处理阶段。本文中乘法器是串行实现的，在整个 PE 中，都是串行传输数据。对于加法树采用了两两相加的结构，其寄存器 REG 是级联的接口。



a. PEA的结构

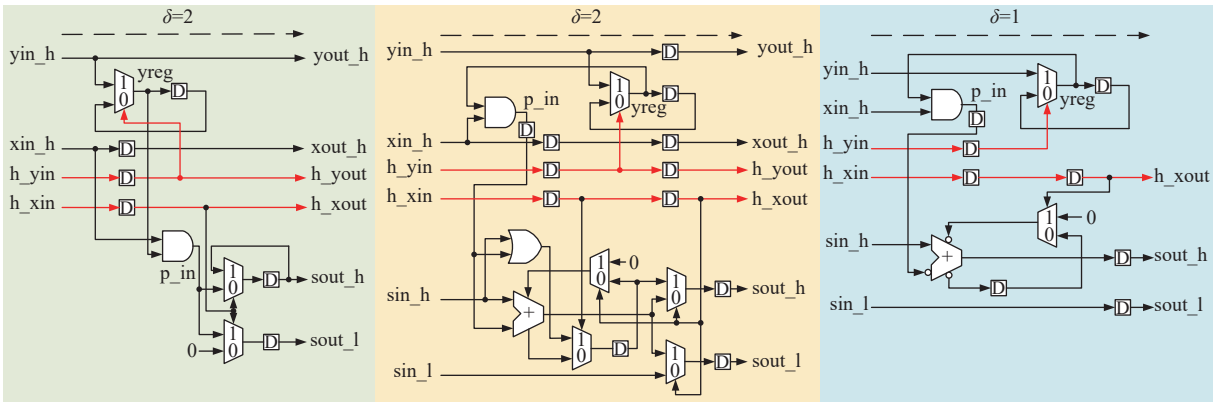


b. 加法器树的结构

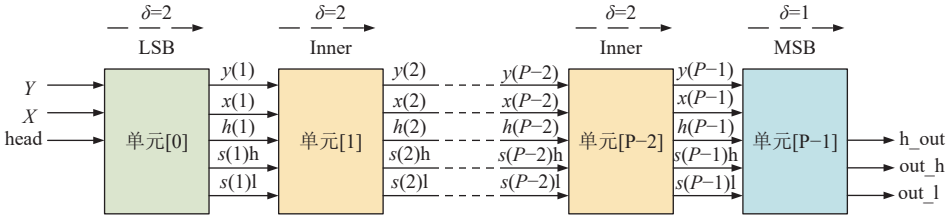
图 10　PE 阵列

3.4　MAC 的设计

为达到减少串并电路的目的，把文献 [12] 的原位串行乘法器输入的头位信号分为 h_yin 和 h_xin，将输入的 IFM 和权重的头位信号分别表示出来。根据 WS 流或 CB 流的规律，以便将权重固定在电路中，而不用一直访存输入，这种乘法器称为固定乘数原位串行乘法器，如图 11 所示。使用这种固定乘数原位串行乘法器和含溢出位的 24 位位串行加法器搭建出如图 12 所示的加速器。



a. 最低有效位乘法器单元　　b. 中间乘法器单元　　c. 最高有效位乘法器单元



d. 乘法器整体结构

图 11　固定乘数原位串行乘法器设计

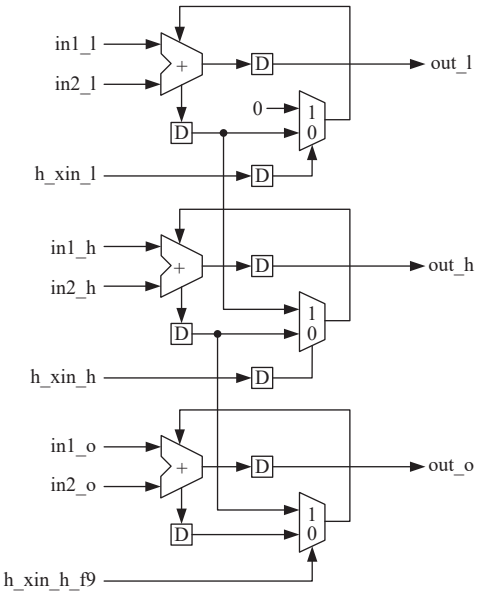


图 12　24 位位串行加法器

3.5　控制电路的设计

顶层控制模块是 1 个控制单元，主要包含有对卷积层和全连接层的控制逻辑，用于管理数据和地址的流动。控制信号用于指示从外部进入数据传输的开始或结束，以及内部数据传输的开始或结束。多个地址线用于定位数据 RAM 中的数据。控制信号在各个状态生成，读写地址视数据量来决定位宽。顶层控制模块对卷积层和池化层的控制状态机分为 10 个状态，具体情况如图 13 所示。其中 IDLE 为初始状态，在此状态加速器等待启动信号有效，以跳转至 CL1 状态，即第 1 层卷积层的输入状态。在 CL1 状态，当第 1 层卷积层的卷积核的行数、列数和个数，以及输入特征图的行数、列数和个数均计数完毕，则进入 WAIT1 状态，即第 1 层卷积层的运算状态，否则继续计数。在 WAIT1 状态，当计数完乘法和加法所需时间后，将跳转至 POOL1 状态，即第 1 层池化层的输入状态，否则继续计数。与卷积层相同，在 POOL1 状态，当第 1 层池化层的池化窗口的行数、列数和个数，以及输入特征图的行数、列数和个数均计数完毕，则进入 WAIT2 状态，即第 1 层池化层的运算状态，否则继续计数。对于第 2 层卷积层的 CL2 和 WAIT3 状态，以及第 2 层池化层的 POOL2 和 WAIT4 状态与第 1 层类似。最后是 READ 状态，即输入特征图从数据 RAM 的读出状态，当计数完成后，将跳转至 IDLE 状态，否则继续计数。以上的 WAIT 状态都留有一定余量，余量的确定主要基于对硬件资源利用率、数据流稳定性和信号处理延迟的综合评估。具体来说，余量设置为信号处理时间的 10%～15%，该设置能够有效避免硬件资源饱和导致的性能下降。依据各维度的计数值，可以生成卷积核、特征图和池化层的权重及特征图数据地址。

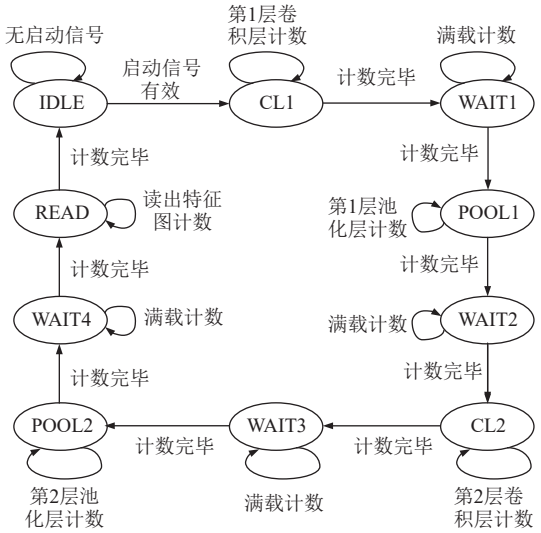


图 13　卷积层和池化层的控制状态机

顶层控制模块对全连接层的控制状态机分为 7 个状态，具体情况如图 14 所示。与卷积层和池化层的控制状态机相似，主要分为初始状态、输入状态和运算状态，控制信号在各个状态生成。在初始 IDLE 状态，加速器在所有的卷积和池化未完成时，等待来自卷积层的 WAIT4 状态的启动信号有效，以跳转至 FCL1 状态，即第 1 层全连接层的输入状态。在 FCL1 状态，当第 1 层全连接层的各个维度计数完毕，则进入 WAIT1 状态，即第 1 层全

连接层的运算状态，否则继续计数。第 2 层和第 3 层全连接层的状态与转移情况亦是如此。在最后的 WAIT3 状态，当计数完毕后会跳转至初始状态。以上的 WAIT 状态都留有一定余量。依据各个全连接层维度的计数值，可以生成相应的权重和特征图数据地址。

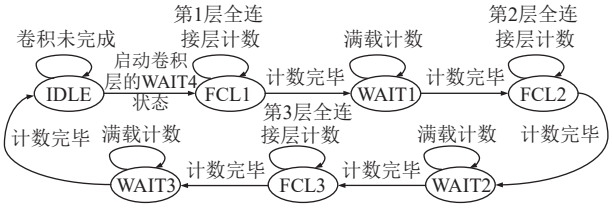
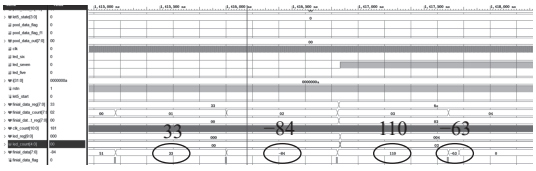


图 14　全连接层的控制状态机

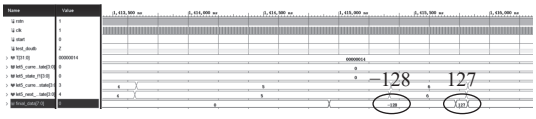
4　实验结果

4.1　仿真波形

取数据集 2a 中受试者 A2，标签为 2 的一个样本，仿真波形如图 15a 所示。从图中可以看到 softmax 层是 4 输出，且输出的个值与软件计算的相同。这 4 个值中第 3 个数最大，因此说明输出为标签 2。取数据集 2b 中受试者 B1，标签为 1 的一个样本，它的仿真波形如图 15b 所示，可见其 softmax 层为 2 输出。这两个值中第 2 个数最大，因此说明输出为标签 1，计算数值与软件计算相同。



a. 数据集2a中标签为2的样本



b. 数据集2b中标签为1的样本

图 15　仿真波形图

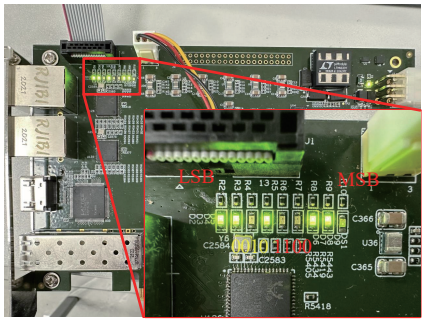
4.2　原型验证

表 1 展示了在不同 FPGA 平台上实施 LeNet-5 模型的性能对比。与文献 [29] 和文献 [30] 的研究成果相比较，本文研究采取了串行位处理方式，而非两者的并行方式。就精度而言，本文研究与文献 [29] 都使用了 8 位固定精度计算，而文献 [30] 则采用了 16 位。在 FCL 的支持上，本文工作及文献 [29] 均提供支持，而文献 [30] 则没有。本文工作的 FPGA 以 500 MHz 的速度运行，超过了文献 [29] 的 136 MHz 和文献 [30] 的 88.07 MHz。在资源利用效率方面，本文研究在 DSP、LUT、FF、BRAM 等方面的利用远少于文献 [29]，表明了更高的资源效率。尽管本研究的功率消耗 865 mW，略高于文献 [30] 的 616 mW，但它在峰值吞吐量和能效上显著优于文献 [29]，处理时间仅为 284.13 μs，远小于文献 [29] 的 4 530 μs 和文献 [30] 的 734 μs。总体来看，本工作尽管在功率消耗上略高于文献 [29]，但在硬件资源利用率、峰值吞吐量、能效和处理时间上均表现出色。

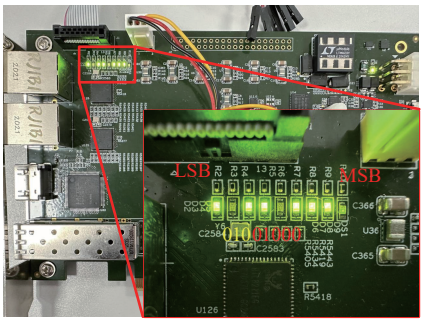
表 1　不同 FPGA 平台实现 LeNet-5 加速器的比较

指标	文献[29]	文献[30]	本文
数据类型	并行	并行	位串行
实现平台	Pynq-z2	Altera Cyclone II 2C70	Xilinx Kintex-7
数据精度	8位定点	16位定点	8位定点
是否支持全连接层	是	否	是
频率/MHz	136	88.07	500
DSP	91	142	2
LUT	34 643	9 399	8 981
FF	18 272	NA	19 050
BRAM	139	22	21
功耗/mW	616	143	865
吞吐量/GOPS	4.46	N/A	7.87
能效/GOPS·W^{-1}	7.24	N/A	9.10
耗时/μs	4 530	734	284.13

将数据集 2a 中标签为 2 的样本的权重和特征图参数以 COE 文件的形式输入 FPGA 板上后，通过 testbench 测试位串行卷积神经网络加速器的推理准确率。如图 16a 所示，LED 灯为低电平使能，MSB 在右边，而 LSB 在左边。4 位标黄的数据指示了是第几个标签。其值为 0100，说明是第 3 个标签，即标签 2。而右边 4 位代表编码后的数据，为 0011，也指示标签 2。对于数据集 2b 中标签为 1 的样本，如图 16b 所示，3 位标黄的数据指示是第几个标签。其值为 010，说明是第 2 个标签，即指示标签 1。而右边 5 位代表编码后的数据，为 00010，也指示标签 1。



a. 数据集2a中标签为2的样本



b. 数据集2b中标签为1的样本

图 16 原型验证

4.3 推理准确率对比

表 2 展示了 FPGA 推理数据集 2a 的准确率和 kappa 值。本文研究的 9 位受试者的平均准确率为 95.68%，平均 kappa 值为 0.942。说明了本研究的位串行 LeNet-5 加速器在数据集 2a 的格拉姆图识别任务中具有很高的准确性和一致性，在推理数据集 2a 上表现良好。

FPGA 推理数据集 2b 的准确率如表 3 所示，本文研究的 9 位受试者的平均准确率为 97.32%，平均 kappa 值为 0.946，也说明了本研究的位串行 LeNet-5 加速器在数据集 2b 的格拉姆图分类任务上具有极高的性能和可靠性。

表 2 在 BCI 竞赛 IV 的 2a 数据集上的识别结果对比

被试者编号	准确率/%			
	文献[31]	文献[32]	文献[33]	本文
A1	91.72	91	97.23	98.61
A2	88.48	89	94.77	99.31
A3	91.72	93	94.12	94.10
A4	88.95	89	95.22	96.88
A5	88.31	92	93.25	95.49
A6	89.12	92	92.36	96.53
A7	89.53	92	94.80	97.92
A8	91.78	93	95.75	90.63
A9	93.75	95	96.38	91.67
平均	90.37	92	94.88	95.68

表 3 在 BCI 竞赛 IV 的 2b 数据集上的识别结果对比

被试者编号	准确率/%		
	文献[34]	文献[35]	本文
B1	87.20	93.4	100
B2	79.79	88.7	97.50
B3	84.19	89.1	93.33
B4	96.32	95.3	99.17
B5	94.06	95.1	92.50
B6	89.27	93.8	99.17
B7	82.98	94.7	96.67
B8	90.63	96.8	97.50
B9	92.80	96.7	100
平均	88.58	93.7	97.32

为了验证 GAF 转换在系统性能中的作用，设计并进行了消融实验，比较了不同特征提取与分类方法在 BCI 竞赛 IV 数据集 2a 和 2b 上的分类性能。具体对比的 3 种方法为共空间模式（common spatial pattern, CSP）结合 FCL 方法，CSP 结合 CNN 方法，以及 GAF 结合 CNN 方法。CSP 结合 FCL 方法是指使用共空间模式对脑电信号进行特征提取，并采用全连接层进行识别与分类的方法，其余两种方法的含义类似。

从表 4 可以看出，GAF 结合 CNN 方法在两个数据集上的分类性能均显著优于其他方法。在数据集 2a 上，GAF 结合 CNN 方法的平均准确率达到 95.68%，比 CSP 结合 FCL 方法和 CSP 结合 CNN 方法分别高出 19.45% 和 10.05%；kappa 值达到 0.942，显著高于 CSP 结合 FCL 方法的 0.683 和 CSP 结合 CNN 方法的 0.808。在数据集 2b 上，GAF 结合 CNN 方法的平均准确率和 kappa 值分别为 97.32% 和 0.946，也显著优于 CSP 结合 FC 方法和 CSP 结合 CNN 方法。

表 4 不同方法在 BCI 竞赛 IV 数据集 2a 和 2b 上的识别结果对比

数据集	方法	平均准确率/%	平均kappa值
2a	CSP+FCL	76.23	0.683
	CSP+CNN	85.63	0.808
	GAF+CNN	95.68	0.942
2b	CSP+FCL	80.09	0.602
	CSP+CNN	86.08	0.714
	GAF+CNN	97.32	0.946

实验结果表明，CNN 在处理 CSP 提取的特征时，能够更充分地学习复杂的非线性模式和空间信息，从而提高分类精度和鲁棒性。相较于传统的全连接层，CNN 通过卷积操作更擅长捕获特征图中的局部结构信息，提升分类的准确性和稳定性。GAF 转换在特征提取中起到了重要作用。通过将一维脑电信号映射为二维特征图，GAF 有效捕捉了信号的时间序列特征和空间分布信息，使 CNN 能够更充分地提取特征并实现更高的分类性能。

这些结果证明了 GAF 结合 CNN 方法在运动想象脑电信号识别中的显著优势，为优化特征提取与分类方法提供了思路。

5 结束语

本文主要探讨了位串行卷积神经网络加速器在脑机接口系统中，特别是在运动想象脑电信号识别领域的应用。

在未来的工作中，可以从系统的架构和实现两个方面进行优化。在系统架构方面，列暂存数据流是针对特定的权重固定数据流优化的，这种设计在性能和资源利用率上具有显著优势，但灵活性略有限制。将探索可配置的硬件架构，如动态的可重配置模块，以支持不同模型权重和数据流的灵活加载，同时在架构中保留列暂存数据流的高效性，从而实现性能与灵活性的平衡。在系统实现方面，计划探索混合精度的量化方法，在关键层保留更高准确度，以进一步提高精度表达能力。此外，复杂环境下的实时性需要进一步验证。通过引入更复杂的数据采集系统，如多用户的并发信号处理场景，并测试该系统在不同信号噪声水平和硬件资源约束下的实际性能，以更全面地评估本系统在复杂环境的实时性。并且需进一步验证系统在更大规模和多样化数据集上的性能，增强模型对复杂场景的适应能力，从而进一步提高系统的泛化能力。

参 考 文 献

[1] YANG H J, SAKHAVI S, ANG K K, et al. On the use of convolutional neural networks and augmented CSP features for multi-class motor imagery of EEG signals classification[C]//Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. New York: IEEE, 2015: 2620-2623.

[2] STURM I, LAPUSCHKIN S, SAMEK W, et al. Interpretable deep neural networks for single-trial EEG classification[J]. Journal of Neuroscience Methods, 2016, 274: 141-145.

[3] LAWHERN V J, SOLON A J, WAYTOWICH N R, et al. EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces[J]. Journal of Neural Engineering, 2018, 15(5): 056013.

[4] ZHAO X Q, ZHANG H M, ZHU G L, et al. A multi-branch 3D convolutional neural network for EEG-based motor imagery classification[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering: A Publication of the IEEE Engineering in Medicine and Biology Society, 2019, 27(10): 2164-2177.

[5] TSUKAHARA A, ANZAI Y, TANAKA K, et al. A design of EEGNet-based inference processor for pattern recognition of EEG using FPGA[J]. Electronics and Communications in Japan, 2021, 104(1): 53-64.

[6] HERNANDEZ-RUIZ A C, ENÉRIZ D, MEDRANO N, et al. Motor-imagery EEGNet-based processing on a low-spec SoC hardware[C]//Proceedings of the IEEE Sensors. New York: IEEE, 2021: 1-4.

[7] FENG L C, YANG L Y, LIU S B, et al. An efficient EEGNet processor design for portable EEG-Based BCIs[J]. Microelectronics Journal, 2022, 120: 105356.

[8] FENG L C, SHAN H W, ZHANG Y Q, et al. An efficient model-compressed EEGNet accelerator for generalized brain-computer interfaces with near sensor intelligence[J]. IEEE Transactions on Biomedical Circuits and Systems, 2022, 16(6): 1239-1249.

[9] ABDULKADER S N, ATIA A, MOSTAFA M S M. Brain computer interfacing: Applications and challenges[J]. Egyptian Informatics Journal, 2015, 16(2): 213-230.

[10] SAMEK W, MÜLLER K R, KAWANABE M, et al. Brain-computer interfacing in discriminative and stationary subspaces[C]//Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society. New York: IEEE, 2012: 2873-2876.

[11] XUE C B, CAO S, JIANG R K, et al. A reconfigurable pipelined architecture for convolutional neural network acceleration[C]//Proceedings of the IEEE International Symposium on Circuits and Systems. New York: IEEE, 2018: 1-5.

[12] ISSHIKI T. High-performance bit-serial datapath implementation for large-scale configurable systems[D]. Santa Cruz: University of California, 1996.

[13] WANG Z, OATES T. Imaging time-series to improve classification and imputation[EB/OL]. [2023-12-21]. http://arxiv.org/abs/1506.00327.

[14] RASHED-AL-MAHFUZ M, ALI MONI M, UDDIN S, et al. A deep convolutional neural network method to detect seizures and characteristic frequencies using epileptic electroencephalogram (EEG) data[J]. IEEE Journal of Translational Engineering in Health and Medicine, 2021, 9: 2000112.

[15] YILDIZ A, ZAN H S, SAID S. Classification and analysis of epileptic EEG recordings using convolutional neural network and class activation mapping[J]. Biomedical Signal Processing and Control, 2021, 68: 102720.

[16] DISSANAYAKE T, FERNANDO T, DENMAN S, et al. Deep learning for patient-independent epileptic seizure prediction using scalp EEG signals[J]. IEEE Sensors Journal, 2021, 21(7): 9377-9388.

[17] ZHANG T, CHEN W Z. LMD based features for the automatic seizure detection of EEG signals using SVM[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering: A Publication of the IEEE Engineering in Medicine and Biology Society, 2017, 25(8): 1100-1108.

[18] LU X J, ZHANG J Q, HUANG S F, et al. Detection and classification of epileptic EEG signals by the methods of nonlinear dynamics[J]. Chaos, Solitons & Fractals, 2021, 151: 111032.

[19] CHOUBEY H, PANDEY A. A combination of statistical parameters for the detection of epilepsy and EEG classification using ANN and KNN classifier[J]. Signal, Image and Video Processing, 2021, 15(3): 475-483.

[20] SYED R S, NAJUMNISSA J D, KAJA MOHIDEEN S. Detection of epilepsy seizure in adults using discrete wavelet transform and cluster nearest neighborhood classifier[J]. Iranian Journal of Science and Technology, Transactions of Electrical Engineering, 2021, 45(4): 1103-

1115.
[21] ECH-CHOUDANY Y, SCIDA D, ASSARAR M, et al. Dissimilarity-based time-frequency distributions as features for epileptic EEG signal classification[J]. Biomedical Signal Processing and Control, 2021, 64: 102268.
[22] OMIDVAR M, ZAHEDI A, BAKHSHI H. EEG signal processing for epilepsy seizure detection using 5-level Db4 discrete wavelet transform, GA-based feature selection and ANN/SVM classifiers[J]. Journal of Ambient Intelligence and Humanized Computing, 2021, 12(11): 10395-10403.
[23] AL-HAMZAWI A A, AL-SHAMMARY D, HAMMADI A H. A survey on healthcare EEG classification-based ML methods[C]//Mobile Computing and Sustainable Informatics. Singapore: Springer, 2022: 923-936.
[24] RUOSPO A, SANCHEZ E, TRAIOLA M, et al. Investigating data representation for efficient and reliable convolutional neural networks[J]. Microprocessors and Microsystems, 2021, 86: 104318.
[25] MITSCHKE N, HEIZMANN M, NOFFZ K H, et al. A fixed-point quantization technique for convolutional neural networks based on weight scaling[C]//Proceedings of the IEEE International Conference on Image Processing. New York: IEEE, 2019: 3836-3840.
[26] YOUNG S I, ZHE W, TAUBMAN D, et al. Transform quantization for CNN compression[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(9): 5700-5714.
[27] SEO S, KIM J. Efficient weights quantization of convolutional neural networks using kernel density estimation based non-uniform quantizer[J]. Applied Sciences, 2019, 9(12): 2559.
[28] BRIAN G. Quantization — PyTorch 2.0 documentation [EB/OL]. [2023-10-02]. https://pytorch.org/docs/stable/quantization.html.
[29] YANAMALA R M R, PULLAKANDAM M. A high-speed reusable quantized hardware accelerator design for CNN on constrained edge device[J]. Design Automation for Embedded Systems, 2023, 27(3): 165-189.
[30] DE FRANÇA A B Z, OLIVEIRA F D V R, GOMES J G R C, et al. Hardware designs for convolutional neural networks: Memoryful, memoryless and cached[J]. Integration, 2024, 94: 102074.
[31] HU Y, LIU Y, ZHANG S Q, et al. A cross-space CNN with customized characteristics for motor imagery EEG classification[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering: A Publication of the IEEE Engineering in Medicine and Biology Society, 2023, 31: 1554-1565.
[32] KHADEMI Z, EBRAHIMI F, KORDY H M. A transfer learning-based CNN and LSTM hybrid deep learning model to classify motor imagery EEG signals[J]. Computers in Biology and Medicine, 2022, 143: 105288.
[33] XIE Y, ONIGA S. Classification of motor imagery EEG signals based on data augmentation and convolutional neural networks[J]. Sensors, 2023, 23(4): 1932.
[34] ZHANG C, KIM Y K, ESKANDARIAN A. EEG-inception: An accurate and robust end-to-end neural network for EEG-based motor imagery classification[J]. Journal of Neural Engineering, 2021, 18(4): 046014.
[35] ROY A M. An efficient multi-scale CNN model with intrinsic feature integration for motor imagery EEG subject classification in brain-machine interfaces[J]. Biomedical Signal Processing and Control, 2022, 74: 103496.

编 辑 税 红