DOI:10.19651/j.cnki.emt.2415890

时空图卷积网络的骨架识别硬件加速器设计*

谭会生 严舒琪 杨 威
(湖南工业大学轨道交通学院 株洲 412000)

摘 要: 随着人工智能技术的不断发展,神经网络的数据规模逐渐扩大,神经网络的计算量也迅速攀升。为了减少时空图卷积神经网络的计算量,降低硬件实现的资源消耗,提升人体骨架识别时空图卷积神经网络(ST-GCN)实际应用系统的处理速度,利用现场可编程门阵列(FPGA),设计开发了一个基于时空图卷积神经网络的骨架识别硬件加速器。通过对原网络模型进行结构优化与数据量化,减少了 FPGA 实现约 75%的计算量;利用邻接矩阵稀疏性的特点,提出了一种稀疏性矩阵乘加运算的优化方法,减少了约 60%的乘法器资源消耗。经过对人体骨架识别实验验证,结果表明,在时钟频率 100 MHz 下,相较于 CPU,FPGA 加速 ST-GCN 单元,加速比达到 30.53;FPGA 加速人体骨架识别,加速比达到 6.86。

关键词: 人体骨架识别;时空图卷积神经网络(ST-GCN);硬件加速器;现场可编程门阵列(FPGA);稀疏矩阵乘加运算硬件优化

中图分类号: TN791 **文献标识码:** A **国家标准学科分类代码:** 510.4030

Hardware accelerator design for skeleton recognition in spatio-temporal graph convolutional networks

Tan Huisheng Yan Shuqi Yang Wei
(College of Railway Transportation, Hunan University of Technology, Zhuzhou 412000, China)

Abstract: With the continuous advancement of artificial intelligence technology, the scale of data in neural networks is gradually expanding, leading to a rapid increase in computational complexity. In order to reduce the computational load of SpatioTemporal Graph Convolutional Neural Networks (ST-GCN), decrease hardware resource consumption, and improve processing speed in practical applications of human skeleton recognition systems, a hardware accelerator based on ST-GCN was designed and developed using Field Programmable Gate Arrays (FPGA). By optimizing the structure of the original network model and quantifying the data, the computational load of FPGA implementation is reduced by about 75%. Based on the sparsity of adjacency matrix, an optimization method for multiplicative and additive operation of sparsity matrix is proposed, which reduces the multiplier resource consumption by about 60%. Experimental validation on human skeleton recognition demonstrated that compared to CPUs, FPGA-accelerated ST-GCN units achieved a speedup of 30.53 at a clock frequency of 100 MHz. The FPGA acceleration for human skeleton recognition achieved a speedup of 6.86.

Keywords: human skeleton recognition; spatiotemporal graph convolutional neural network (ST-GCN); hardware accelerator; field programmable gate array (FPGA); hardware optimization of sparse matrix multiplication and addition

0 引 言

卷积神经网络(convolutional neural network, CNN)在提取文本、图像等欧几里得结构数据的特征方面取得了很大的成功。然而,随着越来越多的应用数据产生自非欧几里得空间,比如图(Graph)这一类典型的非欧数据结构的出现,卷积神经网络变得难以处理[1],为了解决这一问题,图神经网络(graph neural network, GNN)随之出现,并表现出了优异的性能。

GNN 通过在图数据上运行,能够捕捉到节点间的复杂关系和依赖性,这使得 GNN 广泛应用于节点分类、风评评估、推荐系统[1-4]等多个领域。常见的图神经网络可分为卷

收稿日期:2024-04-20
* 基金项目:湖南省学位与研究生教学改革研究项目(2022JGYB183)资助

积图神经网络、循环图神经网络、图自编码器和时空图神经网络[5-7]。其中图卷积神经网络(graph convolutional network, GCN)被广泛应用到图像分类、动作识别等计算机视觉领域,交通预测、天气预测等实况预测领域,文本分类、关系提取等自然语言处理领域[8-9]。

随着人工智能技术的不断发展,神经网络的数据规模逐渐扩大,计算量迅速攀升,增长了处理时间,且在计算机视觉、实况预测等应用领域对实时性要求较高,因此对神经网络进行硬件加速研究具有重要意义。

缪丹丹等[10]针对 CNN 计算量大卷积运算复杂的问题,提出了硬件通用卷积加速器设计方案,根据卷积计算特点,优化了乘加运算及缓存机制,该加速器对所有卷积神经网络的卷积层适用。秦文强等[11]利用卷积分块、并行卷积计算、数据流优化的方法,提高了卷积运算效率。但卷积神经网络加速结构,主要是面向静态计算和规则访存进行设计与优化,而对于图神经网络计算加速结构,由于存在大量动态计算与不规则访存,因此卷积神经网络加速结构不能直接用于图神经网络的加速计算。

针对应用于社交网络分析、推荐系统等特定领域的图神经网络加速的研究已有一些研究,比如:2020 年,图神经网络专用加速结构 HyGCN[12]发表,HyGCN 为图聚合和图更新分别设计了 Aggregation 和 Combination 加速引擎,进行混合结构的协调控制。Zeng 等[13-14]在 CPU 和现场可编程门阵列(field programmable gate array, FPGA)的异构系统上执行图神经网络,使 CPU 和 FPGA 都能够在所擅长的执行行为上进行加速。但在计算机视觉和自然语言处理等领域,图神经网络加速器结构更复杂、实时性更高,目前还未有研究。

骨架识别是计算机视觉领域的重要任务,广泛应用于姿态估计、行为分析等领域。在人体骨架识别方面,Yan 等[15]首次提出的时空图卷积神经网络(spatio temporal graph convolutional neural network, ST-GCN),对人体进行动态骨架建模以关节为节点,人体结构和时间上的自然连通性作为边,构建时空图,提高动作识别任务的准确性。刘斌斌等[16]为解决 ST-GCN 网络在建模关节特征时,缺乏建模全局关节间依赖关系的问题,设计了时空卷积 Transformer 网络对空间和时间关节特征进行建模。虽然基于 ST-GCN 的骨架识别研究已经取得了一些成果,但目前的研究主要集中在网络结构的优化方面,对于基于 FPGA 的硬件加速尚未有研究。

针对图神经网络加速的一些研究主要是应用于文本分类和推荐系统等特定领域,而基于 ST-GCN 的骨架识别的硬件加速尚未有研究,并且 ST-GCN 图神经网络加速计算量更大、消耗的资源更多、并且需要解决计算过程存在的大量动态计算与不规则访存等问题,本文对基于时空图卷积神经网络的骨架识别硬件加速器进行了设计研究。为了减少时空图卷积神经网络的计算量,降低硬件实现的资源消耗,提升人体骨架识别系统的处理速度,本文的主要贡献有:

1) 通过对时空图卷积神经网络模型进行轻量化处理与数据量化,减少计算量,降低计算复杂度。

2) 利用图卷积神经网络邻接矩阵稀疏性的特点,提出了一种根据网络模型的训练结果分布图进行稀疏性矩阵乘加运算的优化方法,减少乘法器资源消耗。

3) 采用流水线结构和并行处理的方式提高 ST-GCN 神经网络硬件处理速度;

4) 设计开发了一个基于时空图卷积神经网络的骨架识别硬件加速器,实验验证硬件加速器的可行性。

1 时空图卷积神经网络骨架识别模型与优化

1.1 时空图卷积神经网络原理

图卷积神经网络是一种用于处理图结构数据的神经网络模型。图结构数据是一种复杂的数据结构,由节点和边构成,其中节点可以表示特定信息,而边则表示各节点间的关系。与传统的矩阵结构不同,图结构数据中节点和边的数量以及类型可能会发生变化。

对于图 $\boldsymbol{g}=\{v,e,\boldsymbol{A}\}$ 上的图卷积,主要由节点集合 v、边集合 e 和邻接矩阵 $\boldsymbol{A}$ 组成。若节点 i 与节点 j 连接,存在一条边,则 $\boldsymbol{A}(i,j)$ 表示边的权重,对于未加权图,则 $\boldsymbol{A}(i,j)=1$。邻接矩阵 $\boldsymbol{A}$ 的度矩阵为 $\boldsymbol{D}$,其中 $\boldsymbol{D}(i,j)=\sum_{j=1}^{n}\boldsymbol{A}(i,j)$。拉普拉斯矩阵为 $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{A}$,对称归一化矩阵为 $\tilde{\boldsymbol{L}}=\boldsymbol{I}-\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}$,$\boldsymbol{I}$ 为单位矩阵。基于上述内容,图卷积运算如式(1)所示。

$$\boldsymbol{g}=\sigma\left(\sum_{j=0}^{k}\boldsymbol{a}_j\tilde{\boldsymbol{L}}\boldsymbol{f}\right) \tag{1}$$

式中:σ 为激活函数,k 为卷积核的接收域,$\boldsymbol{a}_j$ 为卷积核,$\tilde{\boldsymbol{L}}$ 为对称归一化的拉普拉斯矩阵,$\boldsymbol{f}$ 为特征向量,$\boldsymbol{g}$ 为输出特征向量。

时空图卷积神经网络可以有效地分析人体关节在空间中的位置关系,同时捕捉随时间变化的动作模式,从而实现准确的动作识别和姿势跟踪,时空图卷积神经网络运算如式(2)所示。

$$\boldsymbol{g}=\sum_{V_{tj}\in B(V_{ti})}\boldsymbol{M}_{st}(V_{ti},V_{tj})\circ\boldsymbol{L}(V_{ti},V_{tj})\cdot \boldsymbol{f}_{in}(\boldsymbol{p}(V_{ti},V_{tj}))\cdot\boldsymbol{w}(V_{ti},V_{tj}) \tag{2}$$

式中:$\boldsymbol{L}(V_{ti},V_{tj})$ 表示拉普拉斯矩阵,$\boldsymbol{f}_{in}(\boldsymbol{p}(V_{ti},V_{tj}))$ 为输入特征,$\boldsymbol{w}(V_{ti},V_{tj})$ 为权重函数,o 表示哈达玛积,$\boldsymbol{M}_{st}(V_{ti},V_{tj})$ 为加权矩阵。

1.2 时空图卷积神经网络骨架识别模型优化

1)网络结构改进

图 1 为基于图卷积神经网络的人体骨架识别示意图,首先对视频进行姿态估计,并在骨架序列上构建时空图。然后应用至多层时空图卷积网络,并逐渐在图上生成更高层次的特征图。最后,它将被标准 Softmax 分类器分类到

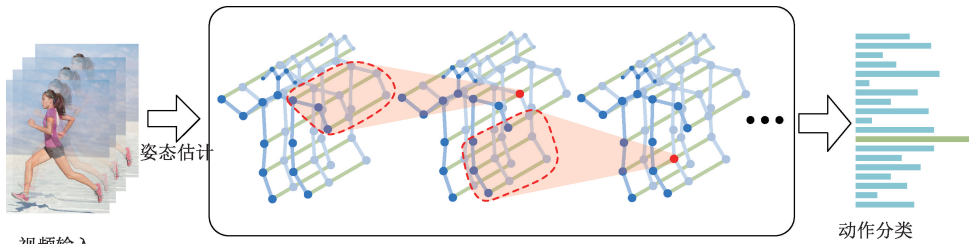

图 1　基于图卷积神经网络的人体骨架识别示意图

Fig. 1　Schematic diagram of human skeleton recognition based on graph convolutional neural network

相应的动作类别。

基于 ST-GCN 的人体骨架动作识别模型如图 2 所示，由 10 层时空图卷积神经网络单元组成。轻量后的网络如图 3 所示，由 8 层时空图卷积算组成，并将通道数减少至原来 1/4。

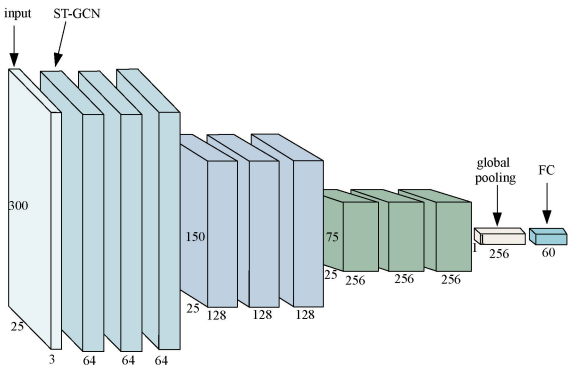

图 2　基于 ST-GCN 的人体骨架动作识别模型

Fig. 2　Human skeleton action recognition model based on ST-GCN

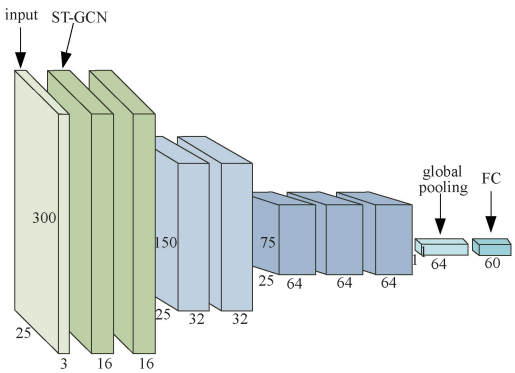

图 3　轻量化后人体骨架动作识别模型

Fig. 3　Human skeleton action recognition model after lightweight

经过测试，结果如图 4 所示，准确率如表 1 所示，其中 Top1 指识别结果与实际结果相符的准确率，Top5 指排名前五识别结果与实际结果相符的准确率。将通道数减少至原网络的 1/4 并去除两层时空图卷积算子，Top1 的准确度仅下降 1.33%，Top5 的准确度仅下降 0.26%，在保证网络精度的同时，使得 FPGA 的减少约 75%计算量。

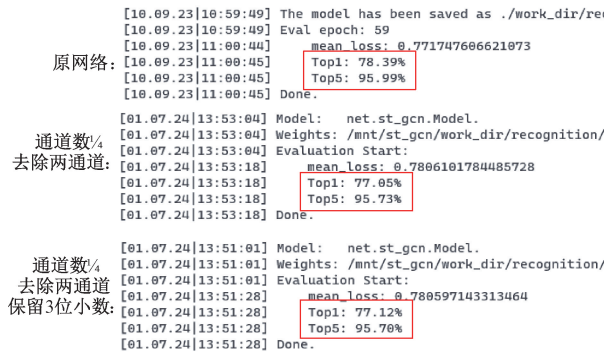

图 4　轻量化后测试结果

Fig. 4　Test results after lightweight

表 1　网络结构优化前后准确率对比

Table 1　Comparison of accuracy before and after network structure optimization　%

模型	原网络结构	网络结构优化 1	网络结构优化 2
Top1	78.39	78.16	77.05
Top5	95.99	95.85	95.73

注：原网络结构如图 3 所示；网络结构优化 1 指将通道数减少至原网络结构的 1/2 并去除一个 16 层的时空图卷积算子；网络结构优化 2 指将通道数减少至原网络结构的 1/4 并去除一个 32 层的时空图卷积算子。

2）数据量化

神经网络中的权重通常以浮点数形式表示，但在 FPGA 中使用浮点运算会消耗大量资源和能耗。因此，将网络中的权重数据进行量化，转化为定点数表示，在减少计算复杂度和硬件资源消耗的同时，尽量减小对网络精度的影响。

为更好进行网络数据量化，通过仿真选择合适数据位数。仿真结果如图 5 所示，数据为双精度浮点数时，Top1 和 Top5 的准确度分别为 77.05%和 95.73%。仅保留 3 位小数时，Top1 的准确度为 77.12%，相较于双精度浮点数，准确度上升 0.06%，Top5 的准确度为 95.70%，仅下降 0.03%。小数位越少，即可用更少定点数位数进行量化，因此保留 3 位小数，既能保证模型识别准确度，也能节省

硬件资源开销。

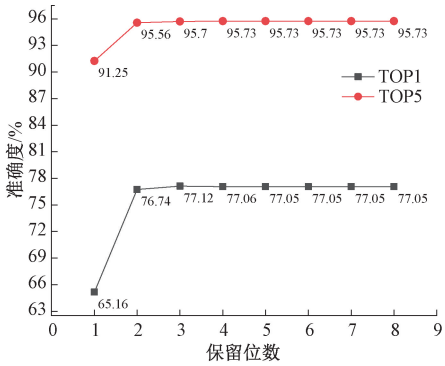

图 5 数据量化准确率对比

Fig. 5 Comparison of data quantization accuracy

将网络结构优化 2 和保留 3 位小数的网络权重数据导出,归一化人体骨架数据,将数据量化为 16 位有符号定点后加载到硬件加速器中。

2 ST-GCN 硬件加速器设计

本文使用 Vivado 2017.4 作为 FPGA 设计开发平台,以 Verilog HDL 作为硬件设计描述语言进行加速器开发。模型数据均采用 16 位有符号定点数表示,其中低 12 位为小数位,最高位为符号位,初始化权重值预先储存在单端口 ROM 中。

2.1 硬件加速器总体设计

根据人体骨架动作识别模型,硬件加速器主要由 ST-GCN 模块、Pooling 池化模块、Softmax 输出模块和输入输出缓存模块组成。

其中加速器的核心为 ST-GCN 模块,由空间图卷积 GCN 单元和时间图卷积 TCN 单元组成,如图 6 所示。GCN 模块由更新单元和聚合单元构成,更新单元为 1×1 卷积计算,聚合单元为哈达玛积运算。TCN 模块为 BN 归一化计算单元和 1×9 卷积计算单元。根据网络结构和 ST-GCN 单元计算原理,ST-GCN 网络整体设计如图 7 所示。

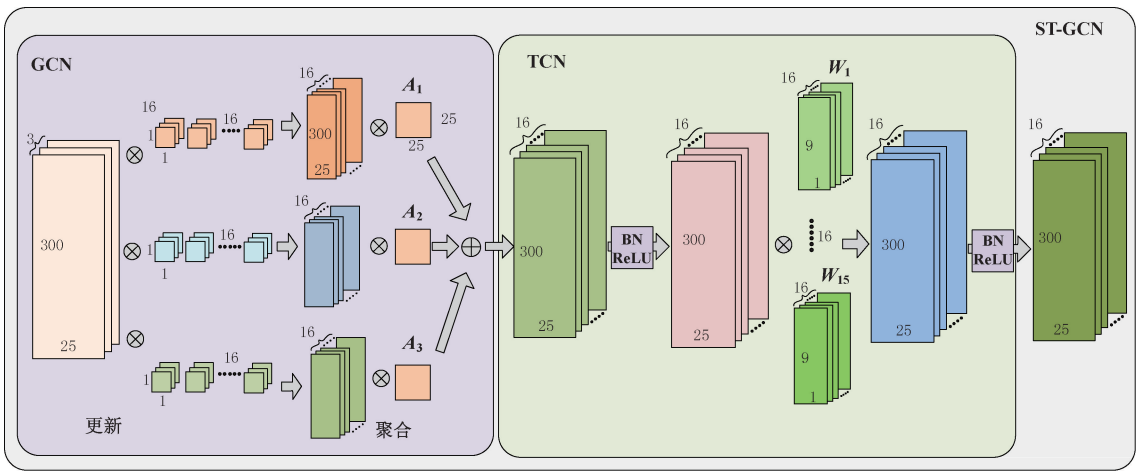

图 6 ST-GCN 单元计算原理

Fig. 6 ST-GCN cell calculation principle

在 ST-GCN 模块中,先经过 GCN 模块处理:首先经过 3 次更新与聚合的处理,每次更新和聚合操作采用并行处理,提高系统的处理速度;然后,数据通过一个同步模块来确保 3 次更新与聚合操作的输出同步;最后进行加法操作。将 GCN 输出数据送入 TCN 模块,经过 BN0 模块后对数据进行缓存,再送入卷积模块和 BN1 模块计算输出数据,作为下一层网络的输入。经过 8 层 ST-GCN 模块后,进入池化模块 Pooling 和最大值激活模块 Softmax 计算出最终分类结果。

2.2 GCN 模块设计

1)更新模块

GCN 模块的更新阶段主要为 1×1 卷积操作,模块设计如图 8 所示。当输入信号有效时,计数器 0 和计数器 1 开始计数。计数器 0 用于产生权值 ROM 读取的控制信号,包括权重 ROM 的读取地址信号 rd_addr 和读使能信号 rd_en。计数器 1 进行数据计数,用于计数输入数据个数,使能寄存器 Reg 累加。完成所有乘累加运算后,加上偏置参数 bias。

2)聚合模块

聚合阶段主要是进行哈达玛积运算,也称为逐元素乘积或元素级乘积,是一种对应位置相乘的运算。

因人体骨架定义了 25 个骨骼节点,因此本模型的邻接矩阵为 25×25 的矩阵。当第一行的第一个数据输入后,需要与 A 中第一列数据逐一相乘,以此类推,计算出第一行 25 个数据的乘积后相加,最后将 3 次聚合的结果相加得到第一个输出。为实现流水线操作,每一个聚合模块均需要 25 个乘法器,3 次聚合则需要 75 个乘法器,将消耗大量资源。

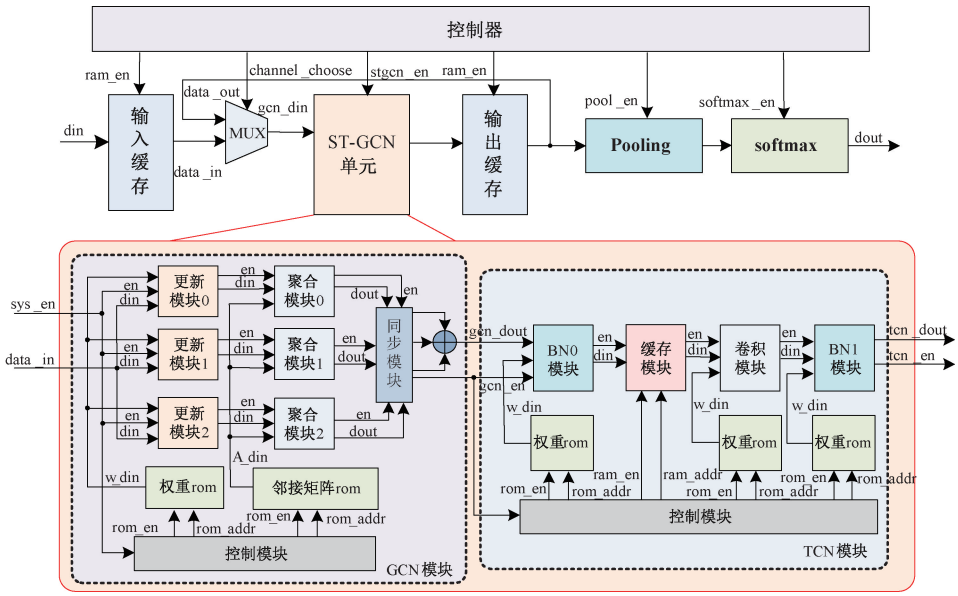

图 7　ST-GCN 网络整体设计

Fig. 7　Overall design of ST-GCN network

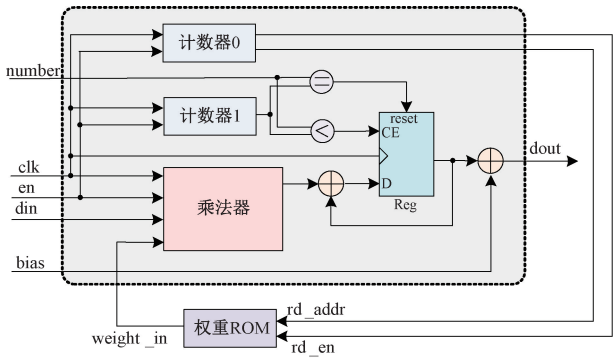

图 8　GCN 卷积模块设计

Fig. 8　GCN convolutional module design

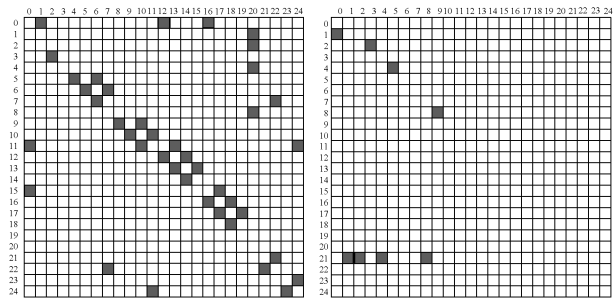

图 9　部分邻接矩阵数据分布图

Fig. 9　Partial adjacency matrix data distribution diagram

因此本文考虑到图神经网络邻接矩阵稀疏性的特点，提出了一种根据网络模型的训练结果，进行稀疏性矩阵乘加运算的优化方法。根据空间划分方法，本网络模型训练出的部分邻接矩阵如图 9 所示，为 25×25 的矩阵，图中灰色部分为非零值，其他空白部分均为 0。在计算过程中，与零相乘为零，因此可以跳过大量的零元素，减少不必要的计算量，加快计算速度。

本文对 3 个聚合模块进行设计。以聚合模块 2 为例，其数据分布如图 9 的邻接矩阵 2 所示。当第一个数据输入时，只需要与坐标为(0,1)的非零数据相乘，因此只需一个乘法器进行计算即可完成第一个数据计算，同理整个邻接矩阵 2 总共需要 5 个乘法器，配合选择器使能特定乘法器，即可完成整个聚合模块 2 的计算。

聚合模块的优化设计如图 10 所示，完成乘加运算后，将数据通过并转串模块串行输出。聚合模块 0、1 也采用同样设计方法，通过计算，整个聚合模块共使用 30 个乘法器，较之前 75 个乘法器的使用，节约了 60%的乘法器。

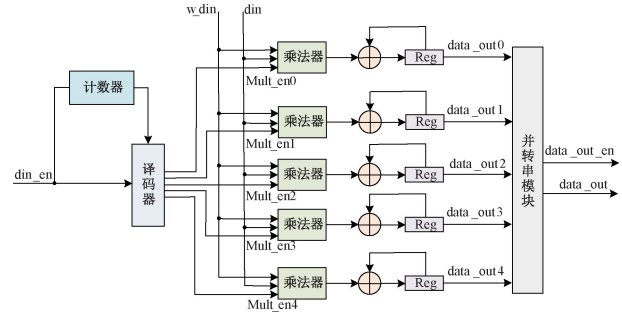

图 10　聚合模块的优化设计

Fig. 10　Optimal design of aggregation module

2.3　TCN 模块设计

TCN 模块的核心模块为 1×9 的 2D 卷积模块。进行 1×9 的 2D 卷积运算，需要获得一个 1×9 的窗口，这里采用输入缓存的方法。如图 11 所示，卷积模块采用 9 个 FIFO，实现 9 行数据缓存，可实现 1×9 卷积滑动窗操作，

同时送入乘法器进行乘加运算。

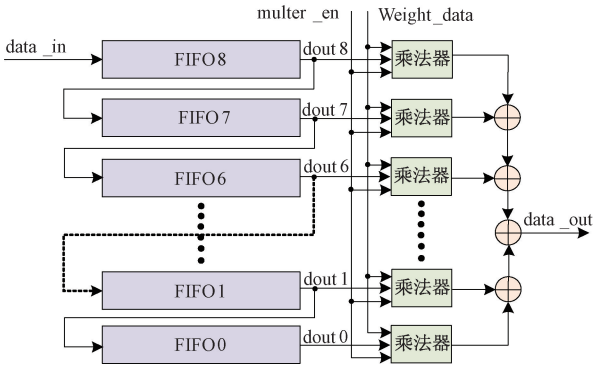

图 11　1×9 卷积模块设计

Fig. 11　1×9 convolutional module design

3　ST-GCN 硬件加速器的仿真与测试

ST-GCN 硬件加速器采用 Vivado 2017.4 进行综合，利用 Xilinx Vivado Simulator 进行仿真，FPGA 型号为 XC7Z100FFG900-2I。

3.1　Python 软件系统测试

在 Intel Core-i7 CPU 工作频率 2.3 GHz 情况下使用 Python 进行软件测试。输入 NTU RGB-D 数据集中的人体骨架动作数据，输出识别结果前五的动作，输出结果如图 12 所示。

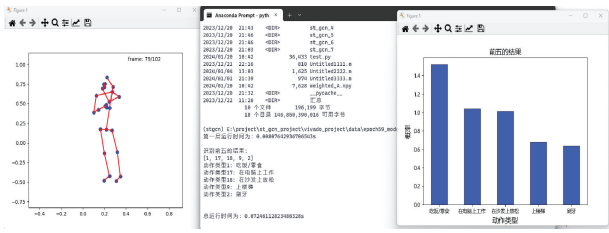

图 12　软件识别结果

Fig. 12　Software recognition result

运行十次识别结果，记录第一层以及总消耗时间，计算平均消耗时间，软件运行结果及时间消耗如表 2 所示，第一层平均消耗时间为 9.16 ms，总消耗时间为 69.88 ms。

3.2　FPGA 硬件模块仿真

在仿真中，将训练好的权值导入 ROM 中，选择 NTU RGB-D 数据集中的一个骨架序列作为输入，其标签为“喝水”，数字标签为“0”。GCN 仿真结果如图 13 所示。在 100 MHz 时钟条件下，数据输入在 225.11 μs 处完成输入，仿真正确；因采用流水线技术，因此 GCN 模块计算花费时间较少，经过 0.36 μs 后处理完成输出，总花费时间 225.47 μs。

TCN 仿真结果如图 14 所示，GCN 模块输出为 TCN 模块输入，数据经缓存后进行卷积处理，数据在 225.46 μs 处开始输入，约 75.71 μs 后全部输出完成。因此，两个模块即 ST-GCN 模块总共花费时间 301.17 μs。

表 2　软件运行结果及时间消耗

Table 2　Software running results and time consumption

数字标签	动作类型	第一层消耗时间/ms	总消耗时间/ms	识别结果
0	喝水	9.43	65.35	喝水
1	吃饭	8.08	72.46	吃饭
2	刷牙	9.25	73.82	刷牙
3	梳头	9.84	66.10	梳头
4	掉落	8.60	66.92	掉落
5	拾起	9.73	68.14	拾起
6	扔	9.97	66.27	玩手机
7	坐下	8.88	76.29	坐下
8	站起	8.27	79.94	站起
9	上楼梯	9.57	63.52	上楼梯
平均耗时	—	9.16	69.88	—

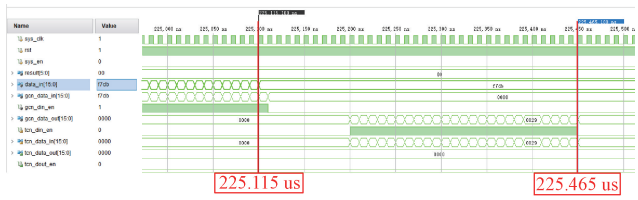

图 13　GCN 仿真结果

Fig. 13　GCN simulation results

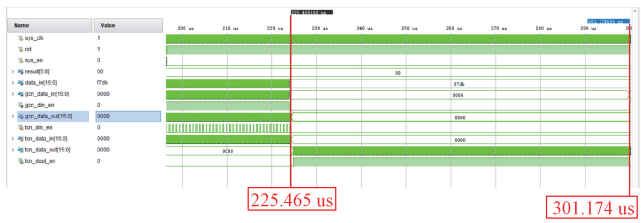

图 14　TCN 仿真结果

Fig. 14　TCN simulation results

3.3　FPGA 硬件加速器系统仿真

人体骨架识别的 ST-GCN 模型为 8 层，在进行整个系统的仿真时，需要将 ST-GCN 单元的第一次计算的输出数据作为 ST-GCN 单元下一次计算的输入数据，共进行 8 次 ST-GCN 单元的计算，然后输入 pooling 模块和 softmax 模块。最后输出概率最大的数字标签，作为整个系统的输出结果。

ST-GCN 系统仿真结果如图 15 所示，识别结果为“1”，由于在 FPGA 中，数字标签从 1 开始，因此标签“1”表示动作“喝水”，仿真正确，在 100 MHz 时钟条件下，总消耗时间为 10.185 ms。

3.4　FPGA 硬件测试结果

利用 ILA 在线逻辑分析仪对 FPGA 系统的输出进行观测。因 ILA 采集深度有限，无法对整体耗时进行检测，

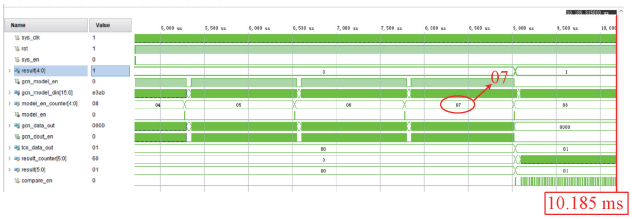

图 15　系统仿真结果

Fig. 15　System simulation result

因此仅对识别结果进行观测。当 *result_counter* = 60 时开始采集识别数据，如图 16 所示。FPGA 识别结果如图 17 所示，用 LED 灯进行二进制显示，FPGA 识别结果为“1”即动作“喝水”，与 ILA 采集结果一致，与仿真结果一致，识别结果正确。

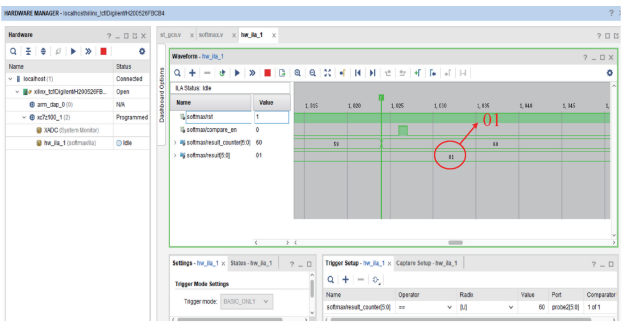

图 16　ILA 在线逻辑分析仪结果

Fig. 16　ILA online logic analyzer results

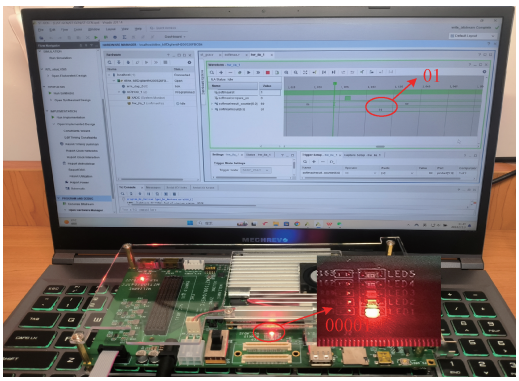

图 17　FPGA 硬件实现

Fig. 17　FPGA hardware realization

ST-GCN 神经网络 FPGA 系统实现资源消耗如表 3 所示。因权重数据及输入均存在 ROM 中，且系统存在 RAM 的缓存模块，因此对 BRAM 资源消耗较大，且各通道之间并行处理，对 DSP 资源也消耗较大。

表 4 给出了软硬件执行时间对比。其中软件时间为 CPU 工作在 2.3 GHz 情况下的执行时间，硬件时间为在 FPGA 时钟频率 100 MHz 时的执行时间，加速比为软件执行时间与硬件执行时间之比。在 FPGA 中，ST-GCN 单元的处理时间仅为 CPU 的 3.27%，人体骨架识别模型的处理时间仅为 CPU 的 14.58%，加速效果明显。

表 3　FPGA 资源消耗

Table 3　FPGA resource consumption

Resource	Utilization	Available	Utilization%
LUT	124 256	277 400	44.79
LUTRAM	2 304	108 200	2.13
FF	40 944	554 800	7.38
BRAM	323	755	42.78
DSP	418	2 020	20.69
IO	8	362	2.21

表 4　软硬件执行时间对比

Table 4　Comparison of software and hardware execution time

参数	CPU	FPGA	加速比
芯片	i7-12700H	Zynq-7000	—
系统时钟	2.3 GHz	100 MHZ	—
数据量化	Float 32	Int 16	—
ST-GCN 总耗时	9.16 ms	0.30 ms	30.53
骨架识别总耗时	69.88 ms	10.19 ms	6.86

4　结　　论

为了减少时空图卷积神经网络的计算量，降低硬件实现的资源消耗，提升人体骨架识别实际应用系统的处理速度，设计开发了一个基于时空图卷积神经网络的骨架识别硬件加速器。通过对原网络模型进行轻量化处理，在保证识别准确度的同时，减少了 FPGA 实现约 75%的计算量；采用定点数计算、流水线结构和并行处理，在减少复杂度的同时提高了计算速度；利用图卷积神经网络邻接矩阵稀疏性的特点，提出了一种根据网络模型的训练结果进行稀疏性矩阵乘加运算的优化方法，减少了 GCN 模块约 60%的乘法器资源消耗。实验结果表明，在时钟频率 100 MHz 下，FPGA 在实现 ST-GCN 神经网络单元的时间约为 301.17 μs，CPU 实现时间为 9.16 ms，加速比为 30.53。FPGA 实现人体骨架识别模型的时间约为 10.18 ms，CPU 实现时间为 69.88 ms，加速比为 6.86，在提高的 ST-GCN 神经网络处理速度的同时降低了资源消耗。

参考文献

[1]　GUI C Y, ZHENG L, HE B S, et al. A survey on graph processing accelerators: Challenges and opportunities[J]. Journal of Computer Science and Technology, 2019, 34(2): 339371.

[2]　KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[J]. ArXiv preprint arXiv: 1609.02907, 2016.

[3]　DAI H J, KOZAREVA Z, DAI B, et al. Learning

steady-states of iterative algorithms over graphs[C]. ProcofInt Confon Machine Learning, New York: Association for Computing Machinery,2018.

[4] 吴相帅,孙福振,张文龙,等.基于图注意力的异构图社交推荐网络[J].计算机应用研究,2023,40(10):3076-3081,3106.
WU X S, SUN F Z, ZHANG W L, et al. GAT based heterogeneous graph neural network for social recommendation [J]. Application Research of Computer, 2023,40(10):3076-3081,3106.

[5] KIPF T N, WELLING M. Variational graph auto-encoders[J]. ArXiv preprint arXiv:1611.07308,2016.

[6] JAIN A, ZAMIR A R, SAVARESE S, et al. Structural-RNN: Deep learning on spatiotemporal graphs[C]. Proc of the IEEE Conf on Computer Vision and Pattern Recognition, Piscataway, NJ: IEEE, 2016.

[7] ZHOU J, CUI G Q, HU S D, et al. Graph neural networks: Areview of methods and applications[J]. ArXiv preprint arXiv:1812.08434,2018.

[8] 蒋玉英,陈心雨,李广明,等.图神经网络及其在图像处理领域的研究进展[J].计算机工程与应用,2023,59(7):15-30.
JIANG Y Y, CHEN X Y, LI G M, et al. Graph neural network and its research progress in field of image processing [J]. Computer Engineering and Applications,2023,59(7):15-30.

[9] 李涵,严明玉,吕征阳,等.图神经网络加速结构综述[J].计算机研究与发展,2021,58(6):1204-1229.
LI H, YAN M Y, LV Z Y, et al. Survey on graph neural network acceleration architectures[J]. Journal of Computer Research and Development,2021,58(6):1204-1229.

[10] 缪丹丹,张鹏,张鑫宇,等.基于 ZYNQ 平台的通用卷积加速器设计[J].国外电子测量技术,2022,41(11):72-77.
LIAO D D, ZHANG P, ZHANG X Y, et al. Generalized convolutional accelerator design based on ZYNQ platform[J]. Foreign Electronic Measurement Technology,2022,41(11):72-77.

[11] 秦文强,吴仲城,张俊,等.基于异构平台的卷积神经网络加速系统设计[J].计算机工程与科学,2024,46(1):12-20.
QIN W Q, WU ZH C, ZHANG J, et al. Design of convolutional neural network acceleration system based on heterogeneous platform[J]. Computer Engineering & Science, 2024,46(1):12-20.

[12] YAN M Y, DENG L, HU X, et al. HyGCN: A GCN accelerator with hybrid architecture[C]. Proc of 2020 IEEE Int Symp on High Performance Computer Architecture(HPCA), Piscataway, NJ:IEEE, 2020.

[13] ZENG H Q, PRASANNA V. GraphACT: Accelerating GCN training on CPU-FPGA heterogeneous platforms[C]. Proc of the 2020 ACM/SIGDA Int Symp on Field-Programmable Gate Arrays, New York: ACM, 2020.

[14] ZHANG B Y, ZENG H Q, PRASANNA V. Hardware acceleration of largescale GCN inference[C]. Proc of 2020 IEEE 31st Int Conf on Application-Specific Systems, Architectures and Processors (ASAP), Piscataway, NJ: IEEE, 2020.

[15] YAN S, XIONG Y, LIN D. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]. Thirty-second AAAI Conference on Artificial Intelligence, 2018.

[16] 刘斌斌,赵宏涛,王田,等.用于骨架行为识别的时空卷积 Transformer 网络[J].电子测量技术,2024,47(1):169-177.
ZHANG B B, ZHAO H T, WANG T, et al. Spatial temporal convolutional Transformer network for skeleton-based action recognition [J]. Electronic Measurement Technology, 2024,47(1):169-177.

作者简介

谭会生(通信作者),硕士,教授,主要研究方向为 EDA/SOPC 技术、VLSI 数字系统、嵌入式系统和功率半导体器件。
E-mail:huisheng21nd@163.com

严舒琪,硕士研究生,主要研究方向为 FPGA 应用,神经网络硬件加速。
E-mail:shuqiyan1999@163.com

杨威,硕士研究生,主要研究方向为 EDA 技术、图神经网络应用。
E-mail:yw143hq@163.com