忆阻器轻量化门控循环单元网络模型设计

华宏虎 许 佳 张博昊 王 伟 李智炜 刘海军*
(国防科技大学，电子科学学院 长沙 410073)

摘 要：忆阻器门控循环单元 (Gated Recurrent Unit, GRU) 网络对于时序数据处理系统的嵌入式部署提供了新的解决途径，但是由于网络规模大、权值精度高，难以直接部署到嵌入式端侧设备。因此，本文开展了忆阻器轻量化GRU网络模型设计研究，构建了能够部署在有限资源上的GRU网络模型，设计了忆阻器交叉阵列的映射方式，提出了基于性能分析与器件感知的融合量化方法，综合考虑网络性能与权值部署、激活函数计算的不同器件实现方式，使用权值对称量化、激活值非对称量化的策略对忆阻器GRU网络模型进行量化，采用权值加噪的方式提升网络模型对忆阻器件非理想因素的包容性。仿真实验表明，本文所设计的忆阻器GRU网络模型在公开的UrbanSound8K数据集上的分类准确率为93.94%，量化至6 bit后模型分类准确率为92.68%，相比于全精度的Dilated Convolution、LM-MFCC+GRU、TFFS-DNN模型分别高出14.68%、0.68%、3.94%，且权值加噪训练能够有效提升轻量化网络模型对忆阻器件非理想因素的适应能力。此外，还验证了该网络模型在真假轨迹判别任务上的性能，在自建的真假轨迹数据集上的分类准确率为97.35%，量化至6 bit后分类准确率仅下降0.84%。

关键词：门控循环单元；忆阻器；网络模型量化；城市音频分类
中图分类号：TN601; TP183 **文献标识码**：A **文章编号**：1009-5896(2026)00-0001-10
DOI: 10.11999/JEIT260152 **CSTR**: 32379.14.JEIT260152

1 引言

由于芯片制程缩小和“冯·诺依曼”存算分离的限制，传统互补金属氧化物半导体 (Complementary Metal Oxide Semiconductor, CMOS) 器件已逼近物理极限。忆阻器[1]凭借高集成密度、快速开关特性及类脑突触可塑性，结合其交叉阵列实现的向量矩阵乘法 (Vector-Matrix Multiplication, VMM)，为突破计算能效瓶颈提供了有效途径。

随着时序数据处理需求激增，门控循环单元 (Gated Recurrent Unit, GRU)[2]作为循环神经网络 (Recurrent Neural Network, RNN) 的重要变体，在轨迹预测[3]等领域表现卓越，但其传统硬件实现面临数据频繁搬移导致的高能耗问题。忆阻器为GRU硬件优化提供了新方案。例如，Wang等人提出了基于忆阻器GRU的硬件电路设计，应用于锂离子电池电荷状态 (State of Charge, SOC) 估计[4]。Tong等人设计了轻量级忆阻器深度可分离卷积-双向GRU网络，用于脑电信号处理[5]。此外，忆阻器的应用研究也不断向新的网络架构和领域场景拓展，例如面向忆阻器存算一体芯片的神经网络结构搜索框架[6]、利用忆阻非线性动力学构建混沌神经网络以增强数据安全[7]等。

然而，GRU参数量大、权值精度高，而忆阻器阵列规模受限，直接部署需大量器件，导致成本上升与可靠性下降。对此，量化技术成为有效的解决手段[8,9]。例如，Balaskas等人通过联合采用混合粒度剪枝和混合精度量化，以硬件感知的方式压缩深度神经网络模型[8]。Mathieu等人结合硬件感知成本函数与量化感知训练 (Quantization-Aware Training, QAT)，提出了用于潜在贝叶斯量化的神经架构搜索方法[9]。但现有量化策略普遍将权值与激活值进行统一处理，未顾及硬件实现差异：权值映射至忆阻器阵列，其量化方式直接影响写驱动电路复杂度；激活值由外围电路动态计算，更需保留动态范围。此外，忆阻器的电导波动等非理想特性会加剧精度损失，现有方案对此适应性考量不足。现有噪声鲁棒训练多基于全精度模型施加扰动，未与QAT深度融合，模型难以在量化约束下同步学习到对器件波动的鲁棒性。

为此，本文提出一种基于忆阻器的轻量化GRU网络模型，通过融合硬件感知的量化策略与面向器件波动的鲁棒性训练方法，实现算法与硬件约束的协同优化。

2 基于忆阻器的GRU网络模型设计

2.1 GRU网络模型

本文所采用的GRU网络模型及其核心单元的

收稿日期：2026-xx-xx；改回日期：2026-06-02；网络出版：2026-07-04
*通信作者： 刘海军 liuhaijun@nudt.edu.cn
基金项目：国家自然科学基金资助项目 (62074166, 62304254, 62104256, 62404253, U23A20322)
Foundation Items: Project supported by the National Natural Science Foundation of China (62074166, 62304254, 62104256, 62404253, U23A20322)

结构如图1所示，结合GRU层提取时序特征的能力和全连接层预测能力，实现对时序数据的分类。

GRU层作为核心组件负责处理输入的时序数据。其输入为初始时序数据经过预处理后得到的T个时间步的长度为D_{i}的特征向量，输出为最后一个时间步的长度为D_{h}的隐藏层状态。其循环结构如图1(b)所示，第l层的第t个时间步的隐藏层状态$\boldsymbol{h}_t^l$，由该层的上一个时间步的隐藏层状态$\boldsymbol{h}_{t-1}^l$和(当$l>1$时)上一层的同一时间步的隐藏层状态$\boldsymbol{h}_t^{l-1}$或(当$l=1$时)输入层同一时间步的输入特征向量$\boldsymbol{x}_t$共同确定，计算公式如下所示：

$$\begin{cases} \boldsymbol{z}_t^l = \sigma\left(\boldsymbol{W}_{\mathrm{z}}^l \cdot \left[\boldsymbol{h}_{t-1}^l, \boldsymbol{h}_t^{l-1}\right]\right) \\ \boldsymbol{r}_t^l = \sigma\left(\boldsymbol{W}_{\mathrm{r}}^l \cdot \left[\boldsymbol{h}_{t-1}^l, \boldsymbol{h}_t^{l-1}\right]\right) \\ \tilde{\boldsymbol{h}}_t^l = \tanh\left(\boldsymbol{W}_{\mathrm{h}}^l \cdot \left[\boldsymbol{r}_t^l \odot \boldsymbol{h}_{t-1}^l, \boldsymbol{h}_t^{l-1}\right]\right) \\ \boldsymbol{h}_t^l = \left(1-\boldsymbol{z}_t^l\right) \odot \boldsymbol{h}_{t-1}^l + \boldsymbol{z}_t^l \odot \tilde{\boldsymbol{h}}_t^l \end{cases} \tag{1}$$

其中，$\boldsymbol{h}_t^{l-1}$、$\boldsymbol{h}_{t-1}^l$、$\tilde{\boldsymbol{h}}_t^l$和$\boldsymbol{h}_t^l$分别是第$l-1$层在第$t$个时间步的状态信息、第$l$层在第$t-1$个时间步的状态信息、第$l$层在第$t$个时间步的候选状态信息和状态信息；$\boldsymbol{z}_t^l$和$\boldsymbol{r}_t^l$分别是第$l$层的更新门和重置门在第$t$个时间步的输出；$\boldsymbol{W}_{\mathrm{z}}^l$、$\boldsymbol{W}_{\mathrm{r}}^l$和$\boldsymbol{W}_{\mathrm{h}}^l$分别是更新门、重置门和候选状态三部分计算的权值矩阵；$\cdot$为矩阵乘积，$\odot$为矩阵哈达玛乘积，σ和tanh分别表示sigmoid和tanh激活函数。公式(1)为$l>1$时的$\boldsymbol{h}_t^l$计算方法；当$l=1$时，需将$\boldsymbol{h}_t^{l-1}$替换成$\boldsymbol{x}_t$。

2.2 忆阻器映射方法

实现忆阻器GRU网络的关键在于权值与忆阻器阵列的映射。本文采用1T1R (one-Transistor-one-Resistor) 阵列，相较于1R阵列能简化阻值调整的电压配置方案。如图2所示，N维输入特征向量以电压形式施加于行线；权值矩阵中每一列权值映射为交叉阵列中的正、负两列忆阻器单元，以差分方式表示有符号权值；列线电流作为VMM运算结果输出。忆阻器阵列利用器件的可控与非易失特性，实现存算一体，避免了传统CMOS中数据频繁搬移的能耗开销。单个乘累加 (Multiply Accumulate, MAC) 运算基于基尔霍夫定律实现，列电流为输入电压与电导值乘积的按行累加结果：

$$I_j^{(+/-)} = \sum_{i=1}^{N} G_{ij}^{(+/-)} V_i \tag{2}$$

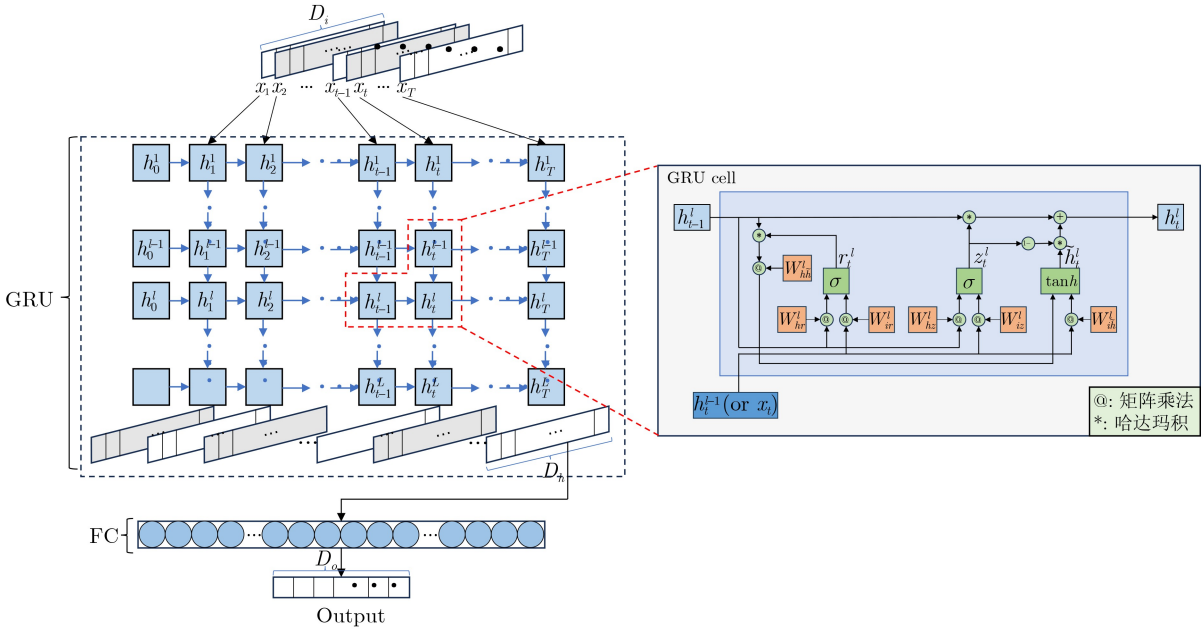
其中，$G_{ij}^{(+/-)}$表示交叉阵列中第i行第j(+)或j(−)列的忆阻器的电导值，V_i为各行线输入电压，$I_j^{(+/-)}$为第j(+)或j(−)列列线的输出电流。

网络权值与忆阻器电导值呈线性映射关系，每个权值矩阵需要映射至正、负两个电导矩阵$\boldsymbol{G}^{(+)}$和$\boldsymbol{G}^{(-)}$。当权值大于(小于)0时，将$\boldsymbol{G}^{(-)}$($\boldsymbol{G}^{(+)}$)对应位置的电导值设置为最小电导值$g_{\min}$；当权值为0时，将$\boldsymbol{G}^{(+)}$和$\boldsymbol{G}^{(-)}$对应位置的电导值均设为$g_{\min}$。$\boldsymbol{G}^{(+)}$与$\boldsymbol{G}^{(-)}$之差为差分电导矩阵$\boldsymbol{G}_{\mathrm{diff}}$：

$$\boldsymbol{G}_{\mathrm{diff}} = \boldsymbol{G}^{(+)} - \boldsymbol{G}^{(-)} \tag{3}$$

$\boldsymbol{G}_{\mathrm{diff}}$与权值矩阵$\boldsymbol{W}$的转换公式如下：

$$\begin{aligned} &\boldsymbol{G}_{\mathrm{diff}} = a \cdot \boldsymbol{W} + g_{\mathrm{b}} \cdot 1, \\ &a = \frac{g_{\max} - g_{\min}}{w_{\mathrm{abs\,max}} - w_{\mathrm{abs\,min}}}, g_{\mathrm{b}} = g_{\min} - a \cdot w_{\mathrm{abs\,min}} \end{aligned} \tag{4}$$



(a) GRU网络模型结构　　(b) GRU层循环单元结构

图 1 GRU网络结构及其核心单元结构

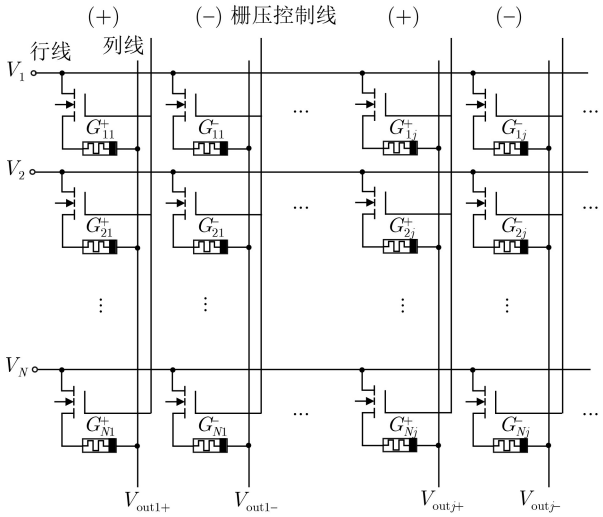


图 2 1T1R忆阻器交叉阵列

其中，a和g_b为线性映射计算因子，g_{max}和g_{min}（$w_{abs\,max}$和$w_{abs\,min}$）分别表示电导(权值的绝对值)的最大值和最小值，1表示与$\boldsymbol{W}$同维的全1矩阵。本文采用异位训练方式，所有训练在通用计算平台上完成，公式(2)–(4)定义的映射关系仅在模型训练和量化收敛后向阵列部署时采用，不参与训练过程。阵列规模为128×128，采用基于固定栅压步进的增量式编程方法，以0.2 V的栅压步进逐步增加栅极电压从而使忆阻器趋近目标电导值[10]，忆阻器电导值编程流程同文献[10]一致。

由于忆阻器态数有限，部署前需对网络权值量化。当网络模型的权值量化精度为N_q bit时，需要用2^{N_q-1}态的忆阻器实现权值部署。量化精度越低，所需阻态数越少，硬件开销越低，但可能带来推理性能下降，因此需权衡推理性能与硬件成本以确定合适的量化位宽。

3　基于性能分析与器件感知的融合量化方法

在对忆阻器GRU网络模型进行轻量化设计时，需根据权值与激活值在硬件实现路径上的差异采用不同量化策略。从硬件实现看，权值训练完成后固化映射至忆阻器交叉阵列，通过多阻态存储权重信息，其量化方式直接影响外围写电路复杂度与阻态调控精度[11]。对称量化因零点恒为零，无需存储和计算零点偏移，写脉冲配置更为规整，利于简化写驱动电路设计。相比之下，激活值在推理中动态产生，其计算通常在阵列外的专用处理器电路中完成[12]，不涉及忆阻器阻态在线编程，因此量化策略应以保留数据分布、最小化量化误差为首要目标。非对称量化能灵活匹配激活值的不对称分布，在相同比特宽度下获得更低量化误差。为此，本文选择权值对称量化、激活值非对称量化的融合量化策略，如图3所示。

量化的核心思想是以低精度数值替代高精度数值，以减少存储与计算开销。本文采用线性量化方法，量化换算公式如下：

$$\boldsymbol{r} = S\left(\boldsymbol{q} - Z\right) \tag{5}$$

$$\boldsymbol{q} = \mathrm{round}\left(\frac{\boldsymbol{r}}{S} + Z\right) \tag{6}$$

其中，$\boldsymbol{r}$为量化前浮点实数，指单层权值或激活值，其最值由统计得到；$\boldsymbol{q}$表示量化后定点整数，其范围由量化精度和方法(对称/非对称)决定；S为缩放因子；Z为零点偏移；round为四舍五入取整函数。

3.1　权值对称量化

权值采用对称量化可简化硬件实现。对称量化中量化范围对称，Z固定为0，S的计算公式如下：

$$S = \frac{\max\left(|r_{min}|, |r_{max}|\right)}{q_{max}} \tag{7}$$

其中，r_{min}、r_{max}为$\boldsymbol{r}$的最小值、最大值，q_{max}为$\boldsymbol{q}$的最大值。若量化位数为n，$\boldsymbol{q}$的范围为$\left[-2^{n-1}, 2^{n-1}-1\right]$。对称量化无需额外存储零点，但若数据分布不对称会导致分辨率浪费。

3.2　激活值非对称量化

激活值采用非对称量化，能够更好地匹配动态范围。激活值非对称量化中量化范围不对称，Z和S的计算公式如下：

$$S = \frac{r_{max} - r_{min}}{q_{max} - q_{min}} \tag{8}$$

$$Z = \mathrm{round}\left(q_{max} - \frac{r_{max}}{S}\right) \tag{9}$$

其中，r_{min}和r_{max}（q_{min}和q_{max}）分别为$\boldsymbol{r}(\boldsymbol{q})$的最小值和最大值。若量化位数为$n$，$\boldsymbol{q}$的范围为$[0, 2^n-1]$。

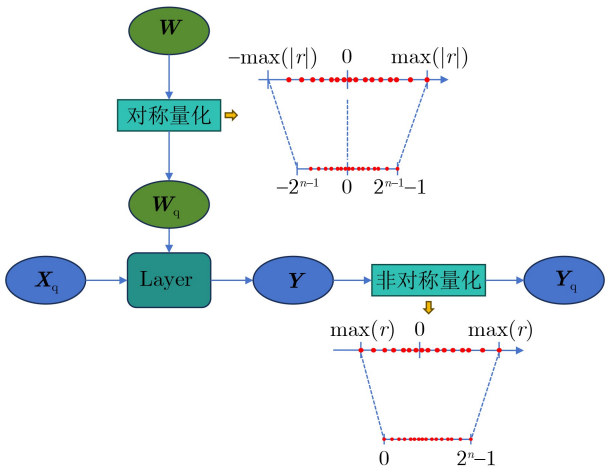


图 3 基于性能分析与器件感知的融合量化方法

非对称量化需存储零点Z，但能适应数据实际分布，量化分辨率更高、误差更小，适合激活值的动态范围匹配。

在QAT的每次训练迭代中，我们执行动态范围统计：在前向传播时，记录该批次数据下各层权值的实际最小值$r_{\min}^{W}$、最大值$r_{\max}^{W}$，以及激活值的实际最小值$r_{\min}^{A}$、最大值$r_{\min}^{A}$。随后，根据公式(7)或公式(8)、(9)更新该层对应的缩放因子S和零点偏移Z。这种每批次更新的策略确保了量化参数能够紧跟权值和激活值分布的变化，从而提升量化精度。

忆阻器的器件间或器件内的离散性、离子运动的随机性等非理想特性，通常表现为阻值波动。仿真中假设器件阻值波动服从正态分布[13]：

$$G_{\mathrm{r}} \sim \mathrm{N}(\mu_0, \sigma_{\mathrm{v}}^2) \tag{10}$$

其中，G_{r}表示阵列中器件的实际电导值，μ_0表示器件的目标电导值，σ_{v}为该正态分布的标准差参数，$\mathrm{N}(\mu_0, \sigma_{\mathrm{v}}^2)$表示均值为$\mu_0$、标准差为$\sigma_{\mathrm{v}}$的正态分布，满足如下关系式：

$$\sigma_{\mathrm{v}} = \mu_0 \cdot r_{\mathrm{v}} \tag{11}$$

其中，r_{v}为相对标准偏差，用于表示器件波动范围，其取值范围为0%-28%[13]。

为分析器件波动性对部署模型性能的影响，本文在仿真中对量化后权值叠加高斯噪声以模拟忆阻器电导波动；为降低该影响，进一步在QAT前向传播中对量化后权值$\boldsymbol{W}_{\mathrm{q}}$施加噪声，使其在训练中提前适应器件波动。加噪强度由器件波动参数r_{v}决定，权值加噪计算公式如下：

$$\begin{aligned} \boldsymbol{W}_{\mathrm{r}} &= \boldsymbol{W}_{\mathrm{q}} + \boldsymbol{N}_{\mathrm{r}}, \boldsymbol{N}_{\mathrm{r}} \sim N(0, \boldsymbol{\sigma}_{\mathrm{n}}^2), \\ \boldsymbol{\sigma}_{\mathrm{n}} &= \boldsymbol{W}_{\mathrm{q}} \cdot r_{\mathrm{v}} \end{aligned} \tag{12}$$

其中，$\boldsymbol{W}_{\mathrm{q}}$和$\boldsymbol{W}_{\mathrm{r}}$分别为加噪前后的权值矩阵，$\boldsymbol{N}_{\mathrm{r}}$为高斯噪声，$\boldsymbol{\sigma}_{\mathrm{n}}$为其标准差，$r_{\mathrm{v}}$与公式(11)中一致。权值加噪训练方法通过主动模拟硬件噪声作为强正则化约束，引导模型参数收敛至更平坦的损失极小值区域，从而学习对权值波动不敏感的鲁棒特征。反向传播时使用直通估计器处理梯度，更新全精度浮点权值，噪声在每次前向传播时动态采样。激活函数通过外围专用处理器电路实现准确输出[12]，因此本文未考虑其非理想特性。

4 仿真实验

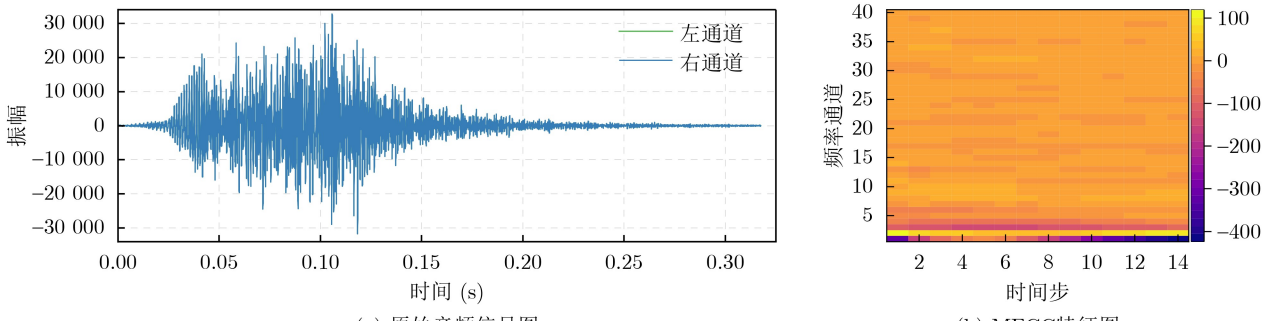
4.1 仿真设置

实验基于AMD Ryzen 9处理器、NVIDIA RTX 4 060显卡的硬件配置和PyTorch框架通过PyCharm集成开发环境完成。忆阻器仿真基于Ti/$\mathrm{AlO_x}$/$\mathrm{TaO_x}$/Pt器件[14]，该器件在1 kΩ至12 kΩ范围内具有超过200个稳定阻态，支持高精度权值映射，并具备长期的权值保持能力(>30 000 s)及高精度的写入调控能力(误差容忍度±1.7%)。

仿真实验基于两个数据集展开：UrbanSound8K公开数据集和真假轨迹自建数据集，分别按8:1:1和6:2:2的比例划分为训练集、验证集和测试集。

UrbanSound8K数据集包含8 732条已标注的声音片段(≤4 s)，涵盖10个类别，广泛用于城市环境声分类研究。实验中利用librosa库提取音频的40维MFCC特征，采样率保持原始44.1 kHz，帧长与帧移采用默认设置(2048和512)，对所有样本填充至最大帧数174帧以便统一推理。狗叫声的可视化示例如图4所示，图4(a)为原始音频信号，图4(b)为预处理后的MFCC特征图。

真假轨迹数据集为自建数据集，用于验证模型在不同时序任务中的可迁移性。基于随机游走模型在16×16网格上生成5帧运动轨迹，通过“相邻运动向量间夹角余弦值是否均大于等于0”判定轨迹真伪(真轨迹方向连续，假轨迹存在突变)。原始二值图像经双线性插值上采样至64×64像素，高斯滤波平滑生成热力图，其中，高斯核标准差为1，像素显示范围为[-0.5, 1.5]。数据集共10 742个样本，每个样本包含5张连续帧图片，按时序构成目标运动轨迹，如图5所示。实验中提取目标在图片中的二维坐标位置作为各时间步特征。



(a) 原始音频信号图　　(b) MFCC特征图

图 4 狗叫声可视化示例

4.2 在城市音频分类任务中的性能

针对UrbanSound8K数据集，GRU网络层数L设定为3，输入维度D_{i}为40，输出维度D_{o}为10。训练采用Adam优化器，损失函数为交叉熵损失，学习率设为0.001，GRU层的随机失活参数设为0.5，训练总轮数设为150。实验考察了隐藏层维度D_{h}对分类性能的影响，结果表明隐藏层维度为300时模型的分类准确率最高，后续实验均采用该设置，模型总参数量约1.4M。

网络训练情况如图6所示，图6(a)为训练集与验证集上的分类准确率随训练轮数的变化曲线，图6(b)为训练损失变化曲线。随着训练轮数增加，准确率先升后稳，损失先降后稳，约40轮后趋于收敛。第136轮训练的模型最佳，其训练损失为0.000 42，训练集准确率为99.99%，验证集准确率为95.30%，测试集分类准确率为93.94%。

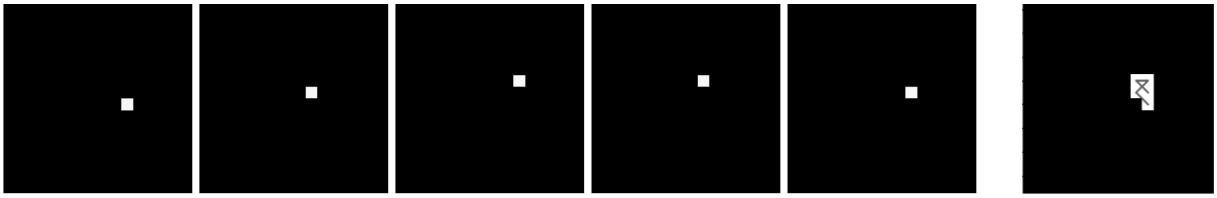
采用融合量化方法对权值和激活值进行量化，量化精度设为2至8 bit及16 bit，模型在测试集上的分类准确率结果如表1所示。分类准确率随量化精度降低而下降，6 bit以上时下降不明显，低于6 bit后显著下降。6 bit时准确率为92.68%，4 bit时降至不足60%，2 bit时仅为12.01%，几乎丧失分类能力。

为了验证网络模型的有效性，将该网络模型与Dilated Convolution[15]、LM-MFCC+GRU[16]、TFCNN[17]、TFFS-DNN[18]、CL-Transformer [19]等模型进行了对比，如表2所示。对比模型均基于相同的UrbanSound8K数据集进行训练、测试，其性能数据均直接引自相应已发表文献，由于部分方法在其原始文献中未提供明确的参数量或网络结构细节，表中以“-”表示。可以看出，全精度(32 bit)本文模型取得93.94%的准确率，优于所有对比模型。6 bit量化后仍保持92.68%的准确率，稍低于全精度TFCNN[17](93.10%)和CL-Transformer[19](92.95%)，但比Dilated Convolution[15]、LM-MFCC+GRU[16]、TFFS-DNN[18]分别高出14.68%、0.68%、3.94%。以6 bit量化为例，权值精度由32 bit降至6 bit，下降81.25%，使模型能够部署于阻态有限的忆阻器阵列。

为了验证模型的泛化能力，在UrbanSound8K数据集中添加不同信噪比 (Signal-to-Noise Ratio, SNR) 水平的高斯噪声，对全精度和6 bit量化模型进行训练测试(迭代20次)，结果如表3所示。6 bit量化模型在各SNR水平下表现出接近或优于全精度模型的鲁棒性，表明融合量化后的模型对带噪输入仍保持良好性能。

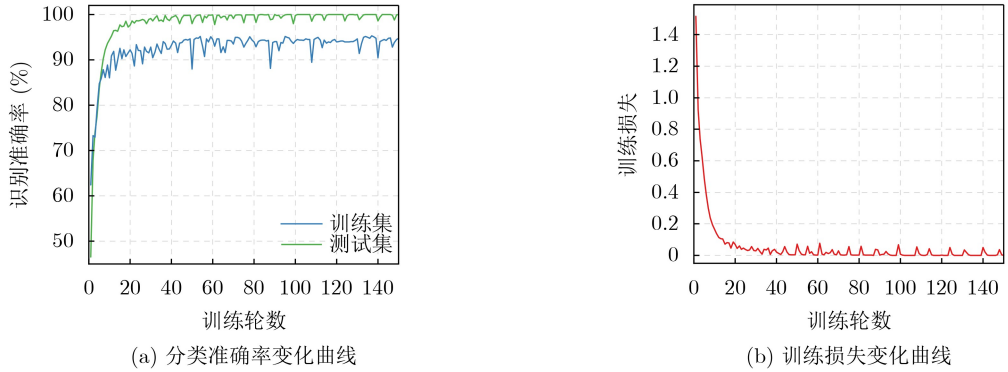
从存储开销、硬件资源和器件可行性三方面分析部署潜力。模型参数量约1.4M, 6 bit量化后存储需求从5.6 MB降至1.05 MB，压缩率81.2%；基于1T1R映射方案，需忆阻器单元2.8M；Ti/AlO_x/TaO_x/Pt忆阻器[14]具备超过200个稳定阻态，远高于6 bit要求的32个阻态。

为分析器件非理想因素影响，对6 bit量化模型进行权值加噪，以2%步长仿真测试器件电导波动



(a) 5帧分步轨迹　　(b) 轨迹整合

图 5 假轨迹示例



(a) 分类准确率变化曲线　　(b) 训练损失变化曲线

图 6 面向城市音频分类任务的GRU网络模型训练情况

在0～28%范围内的分类准确率，实验结果如图7所示。随波动范围增大，分类准确率降低；但权值加噪训练相较未加噪训练显著提升鲁棒性。波动14%时，未加噪训练下分类准确率为82.97%，加噪训练后达91.14%，提升约8%；波动28%时，未加噪训练下分类准确率为54.23%，加噪后达87.01%，提升约33%。表明QAT阶段的权值加噪训练能有效增强轻量化模型对忆阻器件非理想因素的适应能力。

表 1 面向城市音频分类任务的轻量化GRU网络模型分类性能

量化精度 (bit)	分类准确率 (%)
2	12.01
3	36.96
4	51.26
5	78.15
6	92.68
7	93.59
8	93.71
16	93.82

表 2 与其他模型在UrbanSound8K数据集上的性能对比

模型	分类准确率 (%)	权值精度 (bit)	参数量 (M)
Dilated Convolution[15]	78.00	32	-
LM-MFCC+GRU[16]	92.00	32	0.7
TFCNN[17]	93.10	32	1.6
TFFS-DNN[18]	88.74	32	-
CL-Transformer[19]	92.95	32	-
Ours(Before quantification)	93.94	32	1.4
Ours(quantification of 6 bit)	92.68	6	1.4

表 3 在UrbanSound8K数据集上加入不同SNR水平噪声时GRU网络模型的分类性能

SNR (dB)	分类准确率 (%)	
	全精度模型	6 bit量化模型
-10	34.44	38.79
-5	69.91	70.25
0	79.52	85.24
5	84.04	87.30
10	92.11	90.96

4.3 在真假轨迹数据集上的性能

针对真假轨迹数据集，GRU网络层数L设定为1，输入维度D_{i}为2，输出维度D_{o}为2。训练采用Adam优化器，损失函数为交叉熵损失，dropout设为0.5，训练总轮数设为150。实验考察了隐藏层维度D_{h}对分类性能的影响，结果表明隐藏层维度为500时模型的分类准确率最高，后续实验均采用该设置，模型总参数量约0.76M。

为减轻随机性影响，对模型进行30次独立重复训练。图8为30次训练的平均性能曲线，其中，图8(a)为训练集与验证集上的平均分类准确率随训练轮数变化曲线，图8(b)为训练损失变化曲线。准确率随训练轮数增加逐渐升高，初期上升迅速，后期趋于平缓。在第20次重复训练中的第107轮获得最佳模型，其训练损失约0.009 38，训练集准确率约99.98%，验证集准确率约97.86%，测试集分类准确率为97.35%。

采用相同量化策略对权值和激活值进行量化，量化精度设为2至8 bit及16 bit，模型在测试集上的

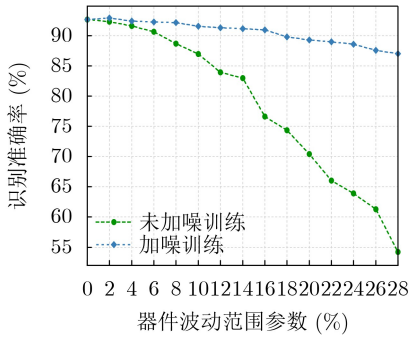
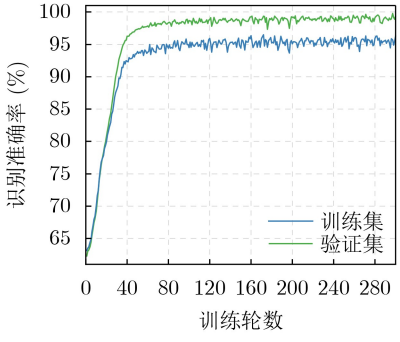
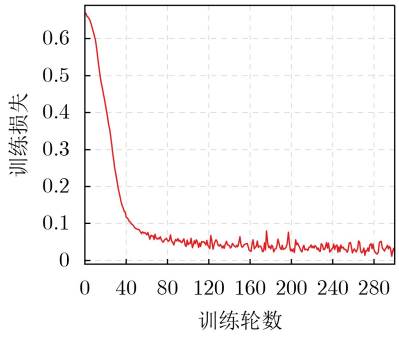


图 7 考虑器件波动性的轻量化GRU网络模型的分类准确率变化情况



(a) 分类准确率变化曲线



(b) 训练损失变化曲线

图 8 面向真假轨迹判别任务的GRU网络模型训练情况

表 4　面向真假轨迹判别任务的轻量化GRU网络模型和Dilated Convolution[15]模型的分类性能

量化精度 (bit)	分类准确率 (%)	
	Ours	Dilated Convolution[15]
2	63.44	65.53
3	63.72	75.53
4	68.28	90.14
5	91.40	90.65
6	96.51	90.56
7	97.16	90.65
8	97.26	90.60
16	97.30	90.98

分类准确率结果如表4所示。分类准确率随量化精度降低而下降，6 bit以上时下降不明显，低于6 bit后迅速下降。6 bit时准确率为96.51%，较全精度仅下降0.84%；低于5 bit后准确率不足70%。考虑到不同模型对自建数据集的适应性，选择Dilated Convolution[15]模型与本模型进行对比，对比结果如表4所示。全精度下，Dilated Convolution准确率为90.70%，分别比6 bit量化的本文模型低5.81%。5至16 bit量化下本文模型准确率均高于Dilated Convolution；2至4 bit量化下则相反。尽管Dilated Convolution在4 bit时仍有90.14%的准确率，但6 bit量化相较于4 bit所需的忆阻器阻态数增加有限，且本文所用忆阻器阻态充足，6 bit时本文模型准确率比4 bit的Dilated Convolution高6.37%，表明轻量化GRU网络仍具有较好的性能优势。

为进一步验证模型在锂离子SOC估计任务上的有效性与扩展能力，并与现有的忆阻器GRU实现方案[20]进行比较。文献[20]基于马里兰大学公共电池数据集[20](含0、25、45 ℃下四种驾驶循环的电流/电压数据)设计了基于忆阻器的GRU硬件电路进行SOC仿真估计。本文采用相同的数据划分方式，其中，BJDST、DST、US06为训练集，FUDS为测试集，并以均方根误差(Root Mean Square Error, RMSE)作为评估指标，实验结果如表5所示。本文提出的全精度模型及6 bit以上量化模型的RMSE均低于文献[20]，其中6 bit模型在0、25、45℃下的RMSE分别为1.48%、0.79%、0.74%，较文献[20]分别降低0.70%、0.57%、0.49%。结果表明，本文模型能够有效迁移至电池SOC估计任务，并在同类忆阻器GRU方案中展现出性能优势。

5　结论

本文设计了一种基于忆阻器的轻量化GRU网络模型，采用1T1R阵列实现了GRU网络模型的权值映射，通过基于性能分析与器件感知的融合量化方法对网络模型的权值和激活值进行量化，采用权值加噪训练的方式提升了网络模型对忆阻器件非理想因素的包容性。仿真实验结果显示，基于忆阻器的GRU网络模型在UrbanSound8K数据集上获得了93.94%的分类准确率，优于现有文献中报道的先进模型结果。采用融合量化方法量化至6 bit后，分类准确率降至92.68%，在模型得到大幅压缩的同时，其性能仍与多种全精度模型的结果具有可比性。权值加噪训练有效提升了轻量化网络模型对忆阻器件非理想因素的适应能力。此外，还在真假轨迹数据集上验证了忆阻器轻量化GRU网络模型的性能。

表 5　不同模型对各种环境温度下FUDS的SOC估计性能RMSE (%)

模型	环境温度 (℃)		
	0	25	45
Memristor-based GRU[20]	2.18	1.36	1.23
Ours (full precision)	1.26	0.58	0.56
Ours (16 bit)	1.57	0.62	0.52
Ours (8 bit)	1.39	0.58	0.52
Ours (7 bit)	1.55	0.73	0.75
Ours (6 bit)	1.48	0.79	0.74
Ours (5 bit)	2.64	1.86	1.76
Ours (4 bit)	13.33	10.50	11.75
Ours (3 bit)	22.62	23.37	24.29
Ours (2 bit)	22.24	22.81	23.82

参 考 文 献

[1] STRUKOV D B, SNIDER G S, STEWART D R, *et al*. The missing memristor found[J]. *Nature*, 2008, 453(7191): 80–83. doi: 10.1038/nature06932.

[2] LECUN Y, BENGIO Y, and HINTON G. Deep learning[J]. *Nature*, 2015, 521(7553): 436–444. doi: 10.1038/nature14539.

[3] 刘淞佐, 王虔, 李磊, 等. 粒子群优化的门控循环单元网络漂流浮标轨迹预测[J]. 电子与信息学报, 2024, 46(8): 3295–3304. doi: 10.11999/JEIT230945.

LIU Songzuo, WANG Qian, LI Lei, *et al*. Gated recurrent unit network of particle swarm optimization for drifting buoy trajectory prediction[J]. *Journal of Electronics & Information Technology*, 2024, 46(8): 3295–3304. doi: 10.11999/JEIT230945.

[4] WANG Jiayang, JI Xiaoyue, DONG Zhekang, *et al*. Circuit design of memristor-based GRU and its applications in SOC estimation[C]. 2023 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, USA, 2023: 1–5. doi: 10.1109/ICCE56470.2023.10043585.

[5] TONG Peiwen, XU Hui, SUN Yi, *et al*. Lightweight and

highly robust memristor-based hybrid neural networks for electroencephalogram signal processing[J]. *Chinese Physics B*, 2023, 32(7): 078505. doi: 10.1088/1674-1056/ac9cbc.

[6] 李源堃, 王泽, 张清天, 等. NAS4CIM: 面向忆阻器存算一体芯片的神经网络结构搜索框架[J]. 电子与信息学报, 2025, 47(12): 4948–4958. doi: 10.11999/JEIT250978.
LI Yuankun, WANG Ze, ZHANG Qingtian, *et al*. NAS4CIM: Tailored neural network architecture search for RRAM-based compute-in-memory chips[J]. *Journal of Electronics & Information Technology*, 2025, 47(12): 4948–4958. doi: 10.11999/JEIT250978.

[7] 蔺海荣, 段晨星, 邓晓衡, 等. 双忆阻类脑混沌神经网络及其在IoMT数据隐私保护中应用[J]. 电子与信息学报, 2025, 47(7): 2194–2210. doi: 10.11999/JEIT241133.
LIN Hairong, DUAN Chenxing, DENG Xiaoheng, *et al*. Dual-memristor brain-like chaotic neural network and its application in IoMT data privacy protection[J]. *Journal of Electronics & Information Technology*, 2025, 47(7): 2194–2210. doi: 10.11999/JEIT241133.

[8] BALASKAS K, KARATZAS A, SAD C, *et al*. Hardware-aware DNN compression via diverse pruning and mixed-precision quantization[J]. *IEEE Transactions on Emerging Topics in Computing*, 2024, 12(4): 1079–1092. doi: 10.1109/TETC.2023.3346944.

[9] PERRIN M, GUICQUERO W, PAILLE B, *et al*. Hardware-aware Bayesian neural architecture search of quantized CNNs[J]. *IEEE Embedded Systems Letters*, 2025, 17(1): 42–45. doi: 10.1109/LES.2024.3434379.

[10] CHEN Junren, WU Huaqiang, GAO Bin, *et al*. Optimization strategy for accelerating multi-bit resistive weight programming on the RRAM array[C]. 2019 IEEE International Workshop on Future Computing (IWOFC), Hangzhou, China, 2019: 1–3. doi: 10.1109/IWOFC48002.2019.9078447.

[11] HONG Haiqiao, DU Zhiyuan, JIANG Mingrui, *et al*. Memristor-based adaptive analog-to-digital conversion for efficient and accurate compute-in-memory[J]. *Nature Communications*, 2025, 16(1): 9749. doi: 10.1038/s41467-025-65233-w.

[12] LI Can, WANG Zhongrui, RAO Mingyi, *et al*. Long short-term memory networks in memristor crossbar arrays[J]. *Nature Machine Intelligence*, 2019, 1(1): 49–57. doi: 10.1038/s42256-018-0001-4.

[13] HUANG Lixing, YU Hongqi, CHEN Changlin, *et al*. A training strategy for improving the robustness of memristor-based binarized convolutional neural networks[J]. *Semiconductor Science and Technology*, 2022, 37(1): 015013. doi: 10.1088/1361-6641/ac31e3.

[14] SUN Yi, XU Hui, WANG Chao, *et al*. A Ti/AlO_x/TaO_x/Pt analog synapse for memristive neural network[J]. *IEEE Electron Device Letters*, 2018, 39(9): 1298–1301. doi: 10.1109/LED.2018.2860053.

[15] CHEN Yan, GUO Qian, LIANG Xinyan, *et al*. Environmental sound classification with dilated convolutions[J]. *Applied Acoustics*, 2019, 148: 123–132. doi: 10.1016/j.apacoust.2018.12.019.

[16] PENG Ning, CHEN Aibin, ZHOU Guoxiong, *et al*. Environment sound classification based on visual multi-feature fusion and GRU-AWS[J]. *IEEE Access*, 2020, 8: 191100–191114. doi: 10.1109/ACCESS.2020.3032226.

[17] MU Wenjie, YIN Bo, HUANG Xianqing, *et al*. Environmental sound classification using temporal-frequency attention based convolutional neural network[J]. *Scientific Reports*, 2021, 11(1): 21552. doi: 10.1038/s41598-021-01045-4.

[18] WU Bo and ZHANG Xiaoping. Environmental sound classification via time–frequency attention and framewise self-attention-based deep neural networks[J]. *IEEE Internet of Things Journal*, 2022, 9(5): 3416–3428. doi: 10.1109/JIOT.2021.3098464.

[19] CHEN Xu, WANG Mei, KAN Ruixiang, *et al*. Improved patch-mix transformer and contrastive learning method for sound classification in noisy environments[J]. *Applied Sciences*, 2024, 14(21): 9711. doi: 10.3390/app14219711.

[20] CHEN Yanan, LUO Wei, CARTER M, *et al*. Organic electrode for non-aqueous potassium-ion batteries[J]. *Nano Energy*, 2015, 18: 205–211. doi: 10.1016/j.nanoen.2015.10.015.

华宏虎：男，国防科技大学电子科学学院博士生，研究方向为忆阻器智能计算架构等.

许 佳：女，国防科技大学电子科学学院硕士生，研究方向为忆阻器智能计算架构等.

张博昊：男，国防科技大学电子科学学院博士生，研究方向为忆阻器智能计算架构等.

王 伟：男，国防科技大学电子科学学院副研究员，研究方向为忆阻器材料、器件和忆阻器类脑芯片等.

李智炜：男，国防科技大学电子科学学院副研究员，研究方向为忆阻器智能计算架构.

刘海军：男，国防科技大学电子科学学院副教授，研究方向为忆阻器智能计算架构、先进集成电路等.

责任编辑：马秀强

Design of Lightweight Gated Recurrent Unit Network Model Based on Memristor

HUA Honghu XU Jia ZHANG Bohao WANG Wei
LI Zhiwei LIU Haijun

(*College of Electronic Science and Technology, National University of Defense Technology, Changsha* 410073, *China*)

Abstract:

Objective With the slowdown of CMOS technology scaling and the inherent memory-computation separation in von Neumann architectures, traditional computing systems face critical bottlenecks in processing increasingly large-scale data. Memristors, which offer high integration density, fast switching speed, and inherent synaptic plasticity, provide a promising pathway to overcome these limitations. Their crossbar arrays naturally support vector-matrix multiplication in the analog domain, enabling energy-efficient in-memory computing. Among sequential data processing models, the Gated Recurrent Unit (GRU) network has emerged as a key recurrent neural network variant, demonstrating superior performance in time-series tasks such as trajectory prediction and audio recognition. However, conventional hardware implementations of GRU networks suffer from frequent data movement between memory and processing units, leading to high energy consumption and low throughput. Although memristor-based GRU implementations offer significant advantages in energy efficiency and computational parallelism, the large parameter size and high weight precision of GRU networks impose substantial hardware costs and reliability challenges when deployed on resource-constrained memristor arrays. Furthermore, device non-idealities, including conductance fluctuations and nonlinear modulation, can substantially degrade model accuracy. Existing memristor-GRU solutions lack comprehensive consideration of these device imperfections, and current quantization methods treat weights and activations uniformly without accounting for their distinct hardware implementation constraints. This paper addresses these challenges through a hardware-algorithm co-design approach.

Methods This paper proposes a lightweight memristor-based GRU network model. A 1T1R (one-transistor-one-resistor) memristor crossbar array is adopted for weight mapping and analog multiply-accumulate (MAC) operations. To accommodate signed weights while memristor conductance values are strictly non-negative, each weight is mapped to a differential pair of positive and negative conductance matrices. The linear mapping between trained weights and memristor conductance values is defined through a transformation formula involving scaling and offset factors. To address the distinct hardware implementation requirements of weights and activations, a fusion quantization method based on performance analysis and device awareness is introduced. Specifically, symmetric quantization is applied to weights mapped to the memristor array, as the zero-centered quantization range simplifies write-driver circuit design by eliminating the need for zero-point storage and computation. In contrast, asymmetric quantization is employed for activation values, which are computed in peripheral circuits without involving online memristor state programming, thereby preserving the dynamic range and minimizing quantization error. To mitigate the impact of memristor conductance fluctuations, a weight noising training mechanism is incorporated into quantization-aware training (QAT). Gaussian noise, with intensity determined by the device variation parameter, is injected into quantized weights during each forward propagation. This approach acts as a strong regularizer, guiding the model to converge to flatter loss landscapes and learn robust features insensitive to weight perturbations. The straight-through estimator is used for gradient backpropagation, enabling updates to full-precision floating-point weights while noise is dynamically sampled in each forward pass.

Results and Discussions On the public UrbanSound8K dataset for urban sound classification, the full-precision proposed model achieves 93.94% classification accuracy. After applying the fusion quantization method, the 6-bit quantized model maintains 92.68% accuracy, with only a 1.26% degradation despite an 81.25% reduction in weight precision (Table 1). This performance surpasses that of comparison models including Dilated

Convolution (78.00%), LM-MFCC+GRU (92.00%), TFFS-DNN (88.74%), TFCNN (93.10%), and CL-Transformer (92.95%) at their full-precision settings (Table 2). When evaluated under noisy input conditions with signal-to-noise ratios ranging from -10 dB to 10 dB, the 6-bit quantized model exhibits comparable or superior robustness compared to its full-precision counterpart, demonstrating the effectiveness of the fusion quantization approach (Table 3). Analysis from storage, hardware, and device feasibility perspectives indicates that 6-bit quantization reduces weight storage from 5.6 MB to 1.05 MB, achieving an 81.2% compression rate, with only 2.8 million memristor cells required based on the 1T1R mapping scheme. Regarding robustness to device non-idealities, weight noising training significantly improves performance under conductance fluctuations (Fig. 7). When the device variation range reaches 14%, noising training improves accuracy from 82.97% to 91.14%; at the worst-case variation of 28%, it improves accuracy from 54.23% to 87.01% (Fig. 7), confirming that the proposed training strategy effectively enhances model adaptability to memristor imperfections. On a self-constructed true/false trajectory dataset, the model achieves 97.35% accuracy at full precision and 96.51% at 6-bit quantization, with only a 0.84% degradation, outperforming the Dilated Convolution baseline at equivalent quantization levels (Table 4). Furthermore, to demonstrate generalization across diverse sequential tasks, the model is evaluated on lithium-ion battery state-of-charge (SOC) estimation using a public dataset. The 6-bit quantized model achieves root mean square errors (RMSE) of 1.48%, 0.79%, and 0.74% at temperatures of 0°C, 25°C, and 45°C, respectively, outperforming the existing memristor-based GRU implementation and showing consistent superiority across all evaluated quantization bit-widths of 6 bits and above (Table 5).

Conclusions This paper presents a lightweight GRU network model tailored for memristor-based hardware deployment. Through device-aware fusion quantization and weight noising training integrated with QAT, the model maintains high classification performance while achieving substantial memory compression and robustness to device non-idealities. Experimental results across multiple datasets and tasks confirm that the 6-bit quantized model retains competitive accuracy and demonstrates stable performance, providing a practical solution for deploying GRU networks on resource-constrained memristor-based edge computing platforms.

Key words: Gated Recurrent Unit (GRU); memristor; network model quantification; urban sound classification