**文章编号**:1007 - 4252(2021)06 - 0006 - 514

# 结合峰度正则化优化存算一体化芯片性能的方法

金健孜[1,3],汤恒松[2],高世凡[3],赵 毅[3,4,*]

(1. 浙江工商大学 萨塞克斯人工智能学院,杭州 310018; 2. 上海闪易半导体有限公司,上海 201210; 3. 浙江大学 信息与电子工程学院,杭州 310027; 4. 中国电子科技南湖研究院,嘉兴 310012)

**摘要**:存算一体化架构通过采用模拟计算,能够极大地提升深度神经网络推理的计算能效。然而,模拟计算的有限精度和神经网络训练平台的高精度之间存在一定的差异,限制了算法在存算一体化芯片的部署。通过在神经网络训练中采用峰度正则化的方式进行算法-电路联合优化,可以增大神经网络权重数据的信息熵,从而充分利用忆阻器单元的模拟特性。在基于可编程线性忆阻器(Programmable linear RAM, PLRAM)的存算一体片上系统中,针对关键词识别任务(6 个分类),引入这一方法最终达到约 97% 的识别准确度,提升识别准确度约 4%。

**关键词**:存算一体化;峰度正则化;算法-电路联合优化

**中图分类号**:TN40 **文献标志码**:A

# Co-optimize the Performance of Computing-in-Memory Chips with the Kurtosis Regularization

JIN Jian-zi[1,3], TANG Heng-song[2], GAO Shi-fan[3], ZHAO Yi[3,4,*]

(1. Sussex Artificial Intelligence Institute, Zhejiang Gongshang University, Hangzhou 310018, China; 2. Flash Billion Semiconductor Co. Ltd., Shanghai 201210, China; 3. College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China; 4. China Nanhu Academy of Eletronics and Informatic Technology, Jiaxing 310012, China)

**Abstract**: The computing-in-memory architecture can greatly improve the computational energy efficiency of deep neural network by analog computation. However, there is a difference between the limited precision of analog computation and the high precision of neural network training platform, which limits the deployment of the algorithm in the computing-in-memory chips. By using the algorithm-chip joint optimization of kurtosis regularization, the information entropy of neural network weight data can be increased. In this way, the memristor accuracy can be fully exploit. The computing-in-memory chips sys-

**收稿日期**:2021-11-10; **修订日期**:2021-12-05

**基金项目**:科技创新 2030-"新一代人工智能"重大项目 (No. 2020AAA0109003); 中央高校基本科研业务费专项资金(No. 2020XZZX005-06).

**作者简介**:金健孜(1998- ), 女, 硕士, 主要研究方向为存算一体化架构 (E-mail:jinjianzi1013@126.com).

**通信作者**:赵 毅(1977- ), 男, 博士生导师, 教授, 主要研究方向为先进集成电路制造工艺和相关器件、基于新器件的先进计算技术、先进器件与芯片测试技术等 (E-mail:yizhao@zju.edu.cn).

tem is based on programmable linear memristor (PLRAM). In this system, the introduction of kurtosis regularization can improve the recognition accuracy by about 4%, and achieve about 97% recognition accuracy for keyword recognition tasks (6 categories).

**Key words**: In-memory computing; Kurtosis regularization; Algorithm-circuit co-optimization

## 0　引言

存算一体化架构能够实现高计算并行度和高计算能效,但同时对器件提出了更高的要求。为实现准确的模拟计算,器件需要具有较高的 $I$-$V$ 线性度,较为准确的模拟状态,以及较好的模拟数据保存特性。针对这些需求,研究者通过采用读写分离的器件结构设计和器件工作原理优化等方式开展了大量工作,并在传统数字存储技术的基础上显著提升了其模拟特性[1-3]。同时,器件侧的问题也可以通过算法与电路的联合优化来得到解决(图 1)。在算法方面,通过设计专用的低精度神经网络,可以降低权重数据对于存储器模拟精度的需求[4]。在电路方面,通过采用时域积分的方式,可以降低计算过程对于器件 $I$-$V$ 线性度的要求[5]。

此外,在算法-器件联合优化方面,还可以通过在神经网络的损失函数中引入正则化项,对神经网络进行系统地调控[6]。具体地,在存算一体化芯片中部署神经网络时,由于器件的精度目前尚无法像数字存储一样进行拓展,因此需要对权重数据进行定点化。由此引入的误差受到权重分布的范围的影响。例如,当浮点数权重分布高度集中时,每个权重间仅存在极小的差异,因而需要将这些小量准确地体现在电路中。对于基于模拟计算的存算一体化芯片,这一要求存在巨大挑战。针对这一问题,本文提出了一种基于权重分布峰度值的正则化方法,并研究了正则化的超参数 $\lambda$ 对该方法有效性的影响。

## 1　存算一体化系统的峰度正则化方法

### 1.1　存算一体化芯片中的浮点数神经网络部署过程

如图 2(a)所示,本文采用基于可编程线性忆阻器(Programmable linear RAM, PLRAM)的存算一体化片上系统进行验证[7-9]。该忆阻器利用简单的差分运放结构,通过双端输入单端输出的电路结构,将模拟信号转换为数字信号输出。这一芯片基于闪存工艺进行开发,能够实现高精度模拟数据存储及计

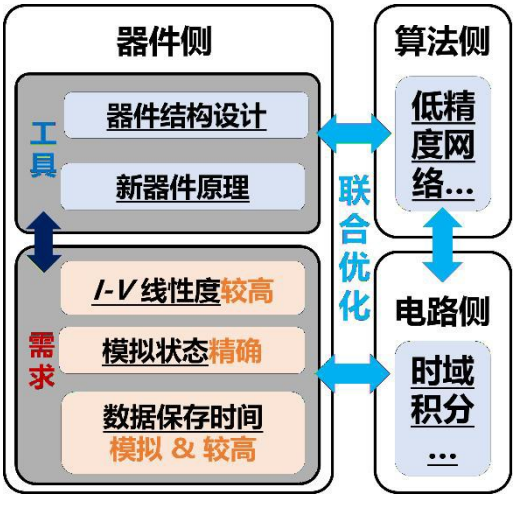


图 1　器件、算法、电路对存算一体化计算系统的联合优化

Fig. 1　Joint optimization for computing-in-memory chips between devices, algorithms and circuits

算,其支持的权重数据为 8 比特带符号定点数。

在浮点数神经网络的部署过程中,首先需要进行算法端近似连续的权重取值的量化(图 2(b))。因为基于 PLRAM 的芯片能够实现较高的精度计算,所以权重取值的量化过程产生的误差成为了瓶颈。具体地,将按照权重中找到绝对值最大值为动态范围,将该范围内的权重按比例映射到 128 的离散数值。在这一过程中,如果有权重分布不均匀,例如部分权重间差异过小,那么这些差异将在定点化过程中被掩盖,并映射为同一个值。这就会导致到硬件端上的模型与原来的模型有出入。我们可以通过信息熵的图像来进一步理解,考察权重分布作为一个随机变量,可以计算其信息熵大小。在较为集中的分布下,每个权重所实际包含的比特数较少。相对而言,均匀分布能够将其信息熵最大化。因此,均一化权重的分布范围,可以使量化过程产生的误差减小。此外,由于模拟写入也存在一定的误差,所以在芯片中权重的实际值将在离散的硬件编码值上重新展开为一个连续分布。由于 PLRAM 阵列能够较好地实现 8 比特的精度,模拟写入阶段所引入的误差较小。因此,本文将主要讨论量化所引入的误差。

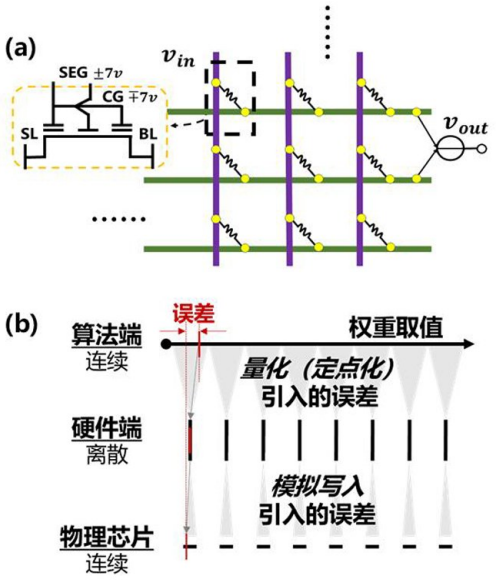


图 2 (a)由 PLRAM 单元组成的计算阵列结构图;(b)算法端、硬件端、物理芯片对于存算一体化系统可能引入的误差

Fig. 2 (a) The calculated array structure consisting of PLRAM units; (b) Errors that algorithm, hardware, and chips may introduce into the computing-in-memory system

### 1.2 深度神经网络中的峰度正则化

在神经网络训练过程中,由于稀疏化等原因,权重中存在着一些特别大的数值。这些较大而又数目较少的权重值会占据相当比例的数据动态范围,从而压缩了其他权重的数值区间。为了解决权重分布中的这一长尾问题,本文采用正则化的方式来调节权重分布。具体地,采用了峰度($K[W]$)这一统计量作为指标,从而训练的校验损失($J$)可以写作:

$$J = J_0 + \lambda * K[W] \tag{1}$$

其中,$\lambda$ 是调节峰度正则化强弱的超参数。峰度在统计学中可以衡量随机变量概率分布的密集程度,如(2)式定义:

$$K[W] = E\left[\left(\frac{W-\mu}{\sigma}\right)4\right] \tag{2}$$

其中 $\mu$ 为均值,$\sigma$ 为标准差。为进一步理解峰度的统计意义,我们可以将(2)式等价变换为如下表达式:

$$K[W] = var\left[\left(\frac{W-\mu}{\sigma}\right)2\right] + 1 \tag{3}$$

由此可以看出峰度具有如下性质[10]:

1. 峰度的最小值将由两点均匀对称的时候产生,所以变量以 0 为中心,呈两点分布的时候,峰度就达到了最小值。

2. 峰度的值变大有两种情况:一种是当数据集中在平均值周围时,偶尔有少量值会远离平均值,此时峰度会变大;第二种是数据集中在分布的尾部,有一个长尾。所以,数据分布越分散,峰度值就越大。

根据峰度的这两个性质可知,通过 $\lambda$ 来调节峰度正则化的强弱,可以实现权重分布在传统的神经网络权重分布和两点 $\delta$ 函数间的混合。研究不同超参数 $\lambda$ 下神经网络片上部署的准确度,是本文研究的重点。

### 1.3 峰度正则化网络在存算一体化系统中的验证

为了验证不同 $\lambda$ 正则化之后对存算一体化芯片性能的影响,我们通过比较语音样本测试集的输出结果与其原始的语音样本数据得到相关系数,并且对输出结果进行分类,得到语音样本分类的准确率。而相关系数跟分类准确率就是评定不同 $\lambda$ 正则化后效果的指标。

本次实验所验证的任务为语音关键词的识别,一共包含 6 个分类,其含义分别是“Power off”、“i_Smart”、“Power_Mode”、“Save_me”、“Help_me” 和“Other”。神经网络训练集包含海量的语音数据,而实验所使用的测试集是原测试集中随机挑选的 300 个语音样本。实验使用的神经网络包含四层全连接层,分别记为 FC1、FC2、FC3 以及输出层 FC4。在中间层我们使用存算一体化芯片和原始算法结果间的相关系数来表征芯片计算的保真度。

## 2 实验结果与分析

如前所述,通过构造较为均一化的权重分布,可以在每个权重中编码更多的数据,从而充分利用器件的模拟特性。我们通过和信息熵类似的相对误差来进一步说明这一问题。为考察量化过程,我们可以将其近似为引入最低位比特的误差,也就是大小为 1。用这个误差值 1 除以原始权重值,可以计算出权重在实际过程中生成的相对误差。如图 3 所示,在不同 $\lambda$ 情况下,FC1 层的训练后权重分布以及其相对误差存在显著差异。随着 $\lambda$ 的增大,权重受到量化过程所引入的相对误差明显减小。

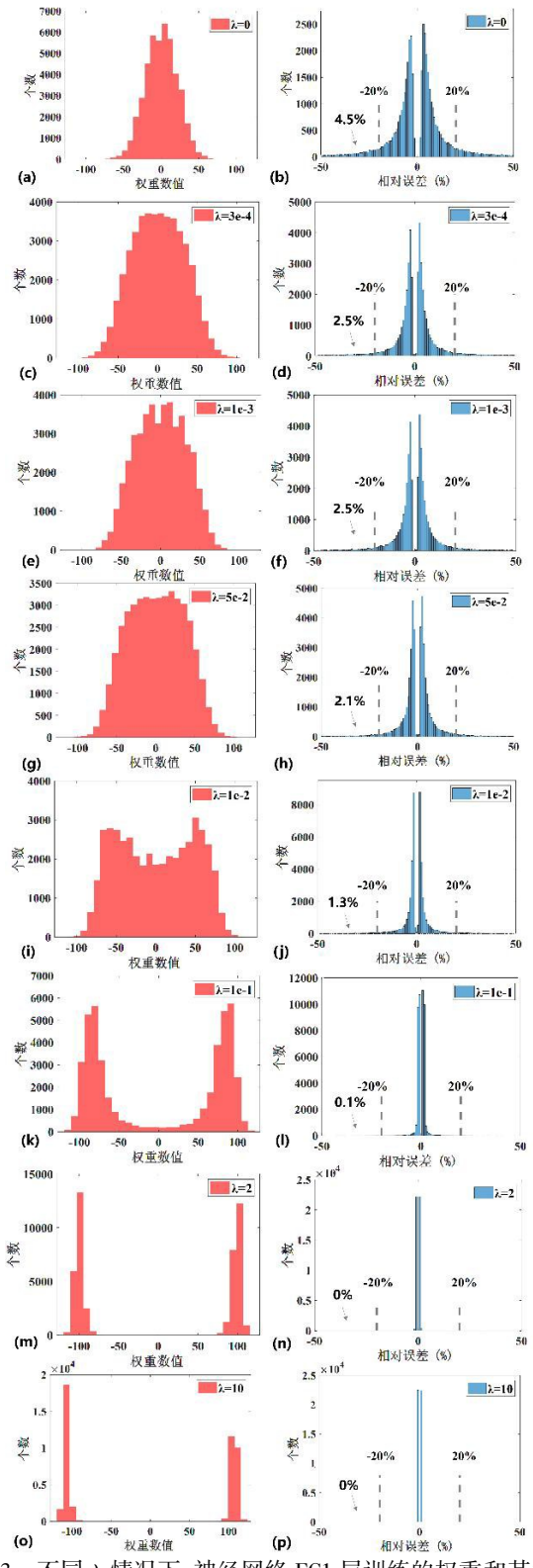


图 3　不同 λ 情况下,神经网络 FC1 层训练的权重和其相对误差的频率分布直方图

Fig. 3 The histograms of the weight of the neural network FC1 layer training and the frequency distribution of its relative error under different conditions

然而,过高的 λ 值将导致神经网络本身的性能出现下降。可以从表 1 中看出,当 λ 大于 2 时,校验精度出现了显著下降。这一趋势可以理解为神经网络在强峰度正则化时出现了二值化的现象,因而其权重所编码的信息量重新开始缩小。

**表 1　不同 λ 情况下,语音样本验证集在神经网络模型上的校验精度与校验损失结果**

Table 1 The results of verification accuracy and the verification loss set on the model under different conditions

| 编号 | λ 值 | 校验精度(验证集在神经网络模型上的分类正确的数量与验证集中的语音样本总数之比) | 校验损失((1)式) |
|---|---|---|---|
| 1 | 0 | 0.9794 | 0.1920 |
| 2 | 3e−4 | 0.9801 | 0.1275 |
| 3 | 1e−3 | 0.9807 | 0.1567 |
| 4 | 5e−2 | 0.9791 | 0.1252 |
| 5 | 1e−2 | 0.9774 | 0.1649 |
| 6 | 1e−1 | 0.9767 | 0.5442 |
| 7 | 2 | 0.9597 | 8.1887 |
| 8 | 10 | 0.9393 | 40.2991 |

进一步地,我们考察神经网络在存算一体化芯片上的部署结果。根据实验的步骤得到相关系数的结果,如图 4。当 λ 在[3e−4,1e−2]范围内时( λ 为 3e−4,1e−3,5e−2,1e−2),FC1−FC3 的输出结果的相关系数是很接近的。到了 FC4 时,λ 为 1e−1、1e−2 和 1e−3 的相关系数都是非常接近的,但是可以看出 λ 为 1e−3 时略高于其他两种情况。而整个芯片的输出,是根据最终 FC4 的输出来决定的。所以,当 λ 为 1e−3 时,芯片的性能是最好的。而当 λ 为 2 和 10 的情况的相关系数结果远低于 λ 为 1e−3 的结果。

如图 5 所示,综合上述结果,可以看到随着峰度正则化强度的增加,芯片结果和算法基线间的差距逐渐缩小。而随着更强的正则化的引入( λ 大于 1e−3),算法基线出现了明显的下滑。因此,在实际应用中,需要进行算法性能和芯片保真度之间的权

衡。

此外,在本文中,所有神经网络层训练权重时均引入相同的 λ 值。在未来的实验中,可以继续研究针对每一层的不同权重数量,引入不同的 λ 值,从而使得芯片的性能得到进一步优化。

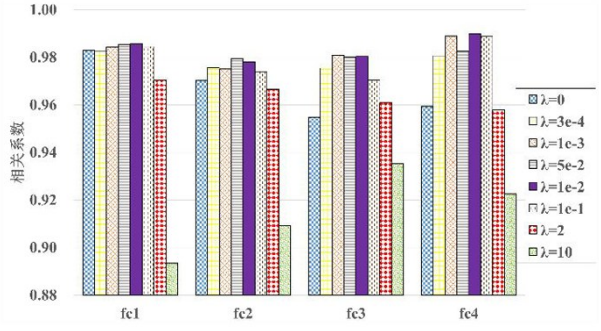


图 4 不同 λ 情况下,神经网络模型四层输出的相关系数的结果

Fig. 4 The correlation coefficients of the four-layer output of the neural network model in different cases

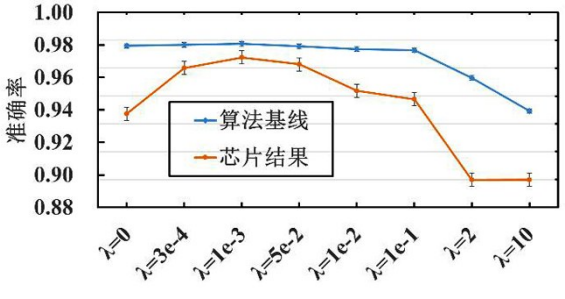


图 5 不同 λ 情况下,验证集与实际神经网络准确率比较结果

Fig. 5 The comparison of the verification setand the actual neural network accuracy in different cases

## 3 结论

在基于 PLRAM 与神经网络结合的存算一体化计算系统训练权重时引入峰度正则化,把训练的权重进行去尾化的操作,使权重的值更加集中,可以减小量化过程带来的误差。在语音关键词识别的任务中,这一方法可以显著缩小芯片和算法间的差距,达到约 1% 的片上部署损失,将准确率提高到 96.8% 。

**参考文献:**

[1] Koelmans W W, Sebastian A, Jonnalagadda V P, et al. Projected phase-change memory devices [J]. Nature Communications, 2015, 6: 8181.

[2] Giannopoulos I, Sebastian A, Gallo M L, et al. 8-bit precision in-memory multiplication with projected phase-change memory [C]// 2018 IEEE International ElectronDevices Meeting (IEDM). San Francisco, USA, 2018.

[3] Wu W, Wu H, Gao B, et al. Improving analog switching in HfOx-based resistive memory with a thermal enhanced layer [J]. IEEE Electron Device Letters, 2017, 38(8): 1019-1022.

[4] Sun X, Wang N, Chen C, et al. Ultra-Low Precision 4-bit Training of Deep Neural Networks [C]// 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Vancouver, Canada, 2020.

[5] Cai F, Correll J M, Lee S H, et al. A fully integrated reprogrammable memristor-CMOS system for efficient multiply-accumulate operations [J]. Nature Electronics, 2019, 2(7): 290-299.

[6] Song C, Liu B, Wen W, et al. A quantization-aware regularized learning method in multilevel memristor-based neuromorphic computing system [C]// 2017 IEEE 6th Non-Volatile Memory Systems and Applications Symposium (NVMSA). 2017, 1-6.

[7] Gao S F, Yang G J, Qiu X, et al. Programmable Linear RAM: A New Flash Memory-based Memristor for Artificial Synapses and Its Application to Speech Recognition System [C]// 2019 IEEE International Electron Devices Meeting (IEDM). San Francisco, CA, 2019.

[8] Zhao L, Gao S, Zhang S, et al. Neural Network Acceleration and Voice Recognition with a Flash-based In-Memory Computing SoC [C]// 2021 IEEE 3rd International Conference on Artificial Intelligence Circuits and Systems (AICAS). IEEE, 2021.

[9] Westfall P H. Kurtosis as peakedness, 1905-2014. RIP [J]. The American Statistician, 2014, 68(3): 191-195.

[10] Moors J J A. The meaning of kurtosis: Darlington reexamined [J]. The American Statistician, 1986, 40(4): 283-284.