特别会议：在非易失性存内计算加速器上可持续部署深度神经网络

Yifan Qin[†⊛] **Zheyu Yan**[†] **Wujie Wen**[‡] **Xiaobo Sharon Hu**[†] **Yiyu Shi**[†*]

[†]University of Notre Dame, [‡]North Carolina State University {[⊛]yqin3, *yshi4}@nd.edu

摘要—非易失性内存（NVM）为基础的存内计算（CIM）加速器因其现场数据处理能力，已成为大幅提高能源效率和减少深度神经网络（DNNs）推理延迟的可持续解决方案。然而，由于 NVM 设备固有的随机性和内在变化，NVCIM 加速器的性能会下降。传统的写入验证操作通过在部署期间迭代写入和验证来增强推理准确性，但在能耗和时间成本方面较高。受负反馈理论启发，我们提出了一种新颖的负优化训练机制，以实现 NVCIM 上鲁棒的 DNN 部署。我们开发了定向变分前向（OVF）训练方法来实现这一机制。实验表明，与现有最先进的技术相比，OVF 在推理准确性方面提高了高达 46.71%，同时减少了认识不确定性。该机制减少了对写入验证操作的依赖，因此有助于 NVCIM 加速器的可持续和实用部署，在保持使用 NVCIM 加速器的可持续计算优势的同时解决了性能下降的问题。

***Index Terms*—内存计算，可持续的，神经网络，加速器**

I. 介绍

深度神经网络（DNNs）已经彻底改变了我们的社会，但它们的加速受到内存和处理单元之间数据传输需求的影响，即冯 · 诺依曼瓶颈 [1]。基于非易失性存储器（NVM）的存内计算（CIM）DNN 加速器 [2] 提供了一个潜在解决方案，通过实现并行原位数据处理，在能效和密度方面超过了基于 CMOS 的同类产品 [1], [3]。这些加速器利用新兴的非易失性存储器设备，如铁电场效应晶体管（FeFETs）[4]、阻变随机存取存储器（RRAMs）[5]、磁阻随机存取存储器（MRAMs）[6] 和相变内存（PCMs）[7]，为 DNN 推理加速提供了可持续的解决方案。尽管它们具有优势，NVM 设备在 NVCIM DNN 加速器中仍受到固有非理想性的困扰，例如设备变化 [5], [8]。这些变化导致编程后设备电导率扰动 [9]，通常会导致呈高斯分布的电导率值 [5]。因此，模型权重受到影响，最终影响 NVCIM DNN 加速器的推理精度 [5], [8], [10]。

确保在不可靠的 NVM 基板上进行可靠的 DNN 推理是一个重大的挑战。在解决方案 [11]–[13] 中，硬件写入验证已作为一种广泛采用的方法用于加速器部署。然而，耗时且能耗高的迭代写入和验证操作阻碍了可持续部署。因此，我们需要更健壮的网络模型以实现更加可持续的部署和加速。

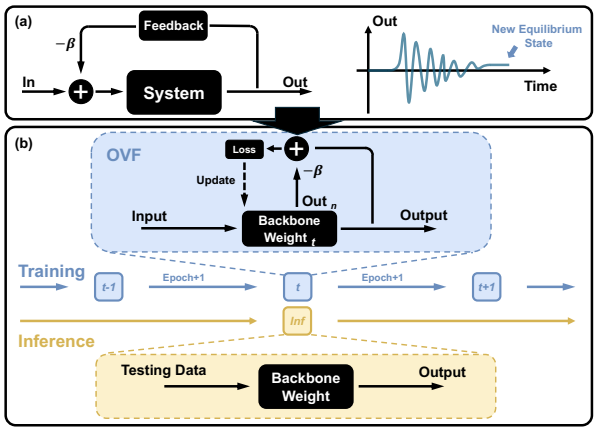

图 1. (a) 经典负反馈系统及其设置到新平衡状态过程的示意图。(b) 负优化训练机制的说明。

噪声注入训练 [8] 广泛用于通过在训练过程中向 DNNs 暴露高斯噪声来增强模型的鲁棒性。这提高了模型对噪声的容忍度。然而，最先进的（SOTA）噪声注入方法存在一些限制，如精度改进有限、认知不确定性 [14] 增加以及难以实现收敛等问题。我们认为这是由于噪声的非确定性和训练的确定性之间的不匹配造成的。具体来说，**第一**，在训练过程中网络仅暴露于有限数量的噪声样本中，这限制了其完全理解噪声模式的能力。**第二**，随机噪声样本提供了多样化的优化方向，但其中一些可能导致错误的状态，增加不确定性并阻碍收敛。

我们相信这种不匹配可以通过在训练过程中获取足够的变异信息来缓解，而不是仅仅依赖最终输出，这是当前最先进的方法中常见的做法。这一假设根植于现代控制理论，在该理论中，稳定性依靠负反馈实现。当系统受到噪声影响时，嘈杂的输出通过负反馈帮助系统抵抗扰动，达到新的平衡状态。图 1(a) 说明了这个过程。反馈是由一部分输出生成，并由负反馈系数 β 调

2832-6474/24/$31.00 © 2024 IEEE
DOI 10.1109/CODES-ISSS60120.2024.00017

制。受此启发，我们引入了一种新颖的**负优化训练机制**，它整合了来自输出的负面贡献，减少了噪声的影响，并帮助神经网络达到更稳健的状态，如图 1(b) 所示。总体而言，整个负优化训练机制可以概括如下：我们使用负优化训练来增强 DNN 主干模型的鲁棒性。训练完成后，移除所有负面贡献组件，留下一个稳健且未改变的 DNN 主干模型。

这里提供了一种称为**定向变分前向（OVF）**训练方法的实现来支撑这一机制。在训练过程中，变分推理输出 Out_n 会使用相同的骨干权重并带有幅度递增的噪声生成，并且在反向传播中以系数 β 负面贡献于目标函数，从而增强 DNN 的鲁棒性。OVF 的“负面”方面减少了目标变化的不良影响同时保持其主体地位，“反馈”组件引入了不同于骨干输出包含的附加噪声信息，对目标做出贡献。OVF 根据噪声本身的影响来约束网络优化，更强的约束对应于更大的扰动。这确保网络不会偏离最优方向太远，在训练过程中有助于稳定地收敛到最优状态。

我们的贡献可概括如下：

- 我们将一种负优化训练机制引入到深度神经网络的训练过程中，以增强稳定性并提高对设备变化的鲁棒性。据我们所知，这是此类方法中的首个尝试。
- 我们提出了一种该机制的新实现方法：定向变分前向（OVF），它从全面的变分性能角度优化网络。
- 我们在具有不同设备变化的 NVCIM DNN 加速器上的模拟展示了 OVF 在减轻敏感性和输出波动方面的有效性。OVF 提高了信心和收敛概率，同时减少了知识不确定性。例如，与最先进的方法相比，它在 DNN 平均推理性能方面实现了高达 46.71%的改进。

II. 提出的的方法论

在本节中，我们介绍了带有 OVF 训练方法的负优化。

我们提出的方法受到了负反馈理论的启发，该理论是系统控制的基础。我们将权重变化视为扰动，并通过负优化抑制这种“噪声”以提高系统的鲁棒性。主要挑战在于构建一个有效的负约束。简单地采用标准负反馈系统中 DNN 的负缩放输出是不够的，因为它仅仅是对损失函数进行缩放而没有改变训练方法。

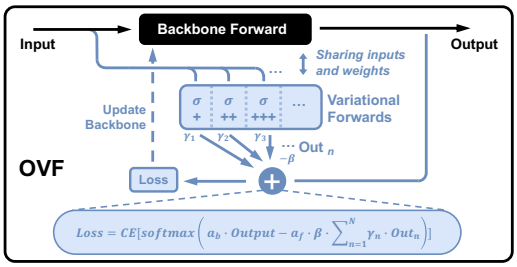

图 2. 负优化实现：定向变分前向训练。

相反，我们要求负约束能够追踪输出的变化，同时与之保持区别。该约束必须满足两个标准：**第一**，它应该由受主干中存在的相同噪声模式影响的组件生成。**第二**，负约束应与主干权重强烈关联，准确反映权重扰动。

如图 2所示，OVF 使用定向变分前向传递中代表性较低的输出 Out_n 来生成约束条件，这些输出涉及的器件变化大于骨干推理中的变化。通过使用负约束条件，OVF 防止骨干偏离最优优化方向。具体来说，在每次训练迭代过程中，我们从高斯分布 $\mathcal{D}ist = \mathcal{N}(0, \sigma^2)$ 中采样一个变分实例 $\Delta \boldsymbol{w}_i$，这与加速器中推理设备的变化相同。这种变化被添加到前馈过程中的骨干权重中，生成骨干输出 $\mathcal{O}_{\text{backbone}}$。与典型的训练直接执行反向传播和权重更新不同，OVF 使用相同的无变化骨干权重并从 $\mathcal{N}(0, \sigma^2)$ 中采样噪音但具有更大的 σ，进行多次定向变分前向操作，收集 N 约束输出 Out_n。总输出 $\mathcal{O}_{\text{total}}$ 由下式给出

$$\mathcal{O}_{\text{total}} = a_b \cdot \mathcal{O}_{\text{backbone}} - a_f \cdot \beta \cdot \sum_{n=1}^{N} \gamma_n \cdot Out_n \qquad (1)$$

其中 β 是负约束系数，γ_n 是衰减因子，而 a_b 和 a_f 影响两部分输出的贡献。反向传播过程仅在获得所有输出后开始。

在选择超参数时，需要注意的是，随着高斯分布的标准差 σ 的增加，噪声实例会向模型及其输出引入更高的熵和不确定性。这导致与目标的偏差更大。因此，在方程 1中，用较大 σ 生成的 Out_n 被赋予更大的衰减因子 γ_n，从而对主干模型施加更强的约束。为了实现这一点，我们将 γ_n 设置为 10^{n-N}。这些变分前向过程从相同的无变异主干权重生成约束，并使用与具有不同参数的主干变分前向相同的高斯噪声模式，从而满足之前概述的标准。

III. 实验

在本节中，我们介绍了权重变化模型和实验设置，然后通过实验结果展示了 OVF 方法的有效性。

A. 模型和设置

不失一般性，我们主要考虑来自编程过程的设备变化，其中 NVM 器件中的编程电导值偏离了期望值。设置具有 M 位的 DNN 权重，量化后的期望权重值 $\bar{\mathcal{W}}_d$ 可以表示为

$$\bar{\mathcal{W}}_d = \frac{\max|\mathcal{W}|}{2^M - 1} \sum_{i=0}^{M-1} m_i \times 2^i \tag{2}$$

其中 $\mathcal{W}$ 代表浮点权重，$\max|\mathcal{W}|$ 表示权重中的最大绝对值，而 $m_i \in \{0, 1\}$ 表示期望权重值的 i^{th} 位的值。对于表示 K 位数据的 NVM 设备，每个权重可以存储在 [1] 设备的 M/K 个中，映射过程由 $\bar{g}_j = \sum_{i=0}^{K-1} m_{j\times K+i} \times 2^i$ 给出，其中 $\bar{g}_j$ 是 j^{th} 设备所需的电导率。需要注意的是，负权重也可以以相同的方式映射到单独的交叉阵列。考虑到器件变化，编程后的实际器件电导率表示为 $g_j = \bar{g}_j + \Delta g$，其中 Δg 表示与期望的电导率值 $\bar{g}_j$ 的偏差，并且遵循高斯分布。因此，由编程 NVM 设备表示的实际权重 $\mathcal{W}_p$ 给出为

$$\mathcal{W}_p = \bar{\mathcal{W}}_d + \frac{\max|\mathcal{W}|}{2^M - 1} \sum_{j=0}^{M/K-1} \Delta g \times 2^{j\times K} \tag{3}$$

在我们的研究中，我们设定了 $K = 2$，而 M 的值则由特定模型配置决定。对于这项研究，我们选择了 $M = 8$，表示单个 DNN 权重为 8 位精度和单个设备电导率为 2 位精度。为了建模器件变化，我们使用了一个具有 $\Delta g \sim \mathcal{N}(0, \sigma_d^2)$ 的高斯分布，其中 σ_d 代表对应于单个器件最大电导率的相对标准差。我们在 σ_d 上设置了限制，将其限定为 $\sigma_d \leq 0.4$。此范围在先前的研究中被认为合理 [4]–[7]，并通过设备层面的优化，包括写入验证技术，可以实现。

我们使用 NVIDIA GPU 上的 PyTorch 环境进行了实验。除非另有说明，报告的结果代表至少五次独立运行的平均值。我们将噪声注入推理的平均准确率作为性能指标，并执行了 200 次运行的蒙特卡洛模拟以确保高精度。我们的实验表明结果具有 95%置信区间 ±0.01，符合中心极限定理。我们比较了 OVF 与两个基线：1）普通的训练（无噪声）和 2）高斯噪声注入训练（有噪声）。我们没有将 OVF 与其他正交方法进行评估，例如基于 NAS 的 DNN 拓扑设计或贝叶斯神经网络，因为可以将其与它们结合使用。

[1]为了简单起见，我们假设 M 是 K 的倍数。

通过我们使用综合数据集和神经网络主干进行的实验，我们发现负约束系数 β 的适当值始终落在集合 $\{1e-1, 1e-2, 1e-3, 1e-4\}$ 内。因此，一个四步搜索就足以确定设置。对于超参数值，我们设 $start = 0$ 和 $end = 2 \times \sigma_d$，评估 OVF 效力在不同 σ_d 值上的表现。我们也设置了贡献因子 $a_b = a_f = 1/(N+1)$，其中 N 表示变分前向的数量。其他训练超参数，如学习率、批量大小和学习率调度器，则遵循训练无噪声模型的最佳实践。

B. 精度提升

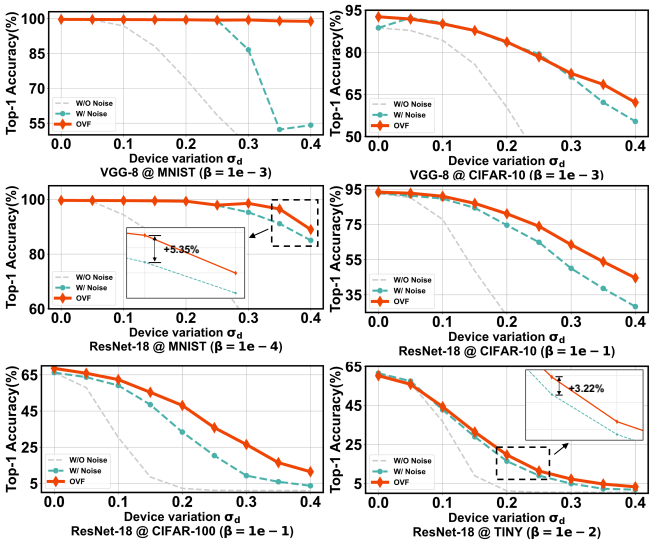

图 3. OVF 的有效性：在不同数据集上，VGG-8 和 ResNet-18 主干模型的平均噪声推理准确率随 σ_d 值的变化。

在我们的实验中，我们使用了 VGG-8 主干网络和 ResNet-18 主干网络，在 MNIST、CIFAR-10、CIFAR-100 和 Tiny ImageNet 数据集上进行了测试。在 OVF 中，我们经验性地将变分前向设置为 $N = 3$，对于 VGG-8 和 ResNet-18 均是如此，并且每次增加量 $\Delta\sigma_d$ 固定为 0.05。此外，在特定数据集上的每个模型的负约束系数 β 是通过四步搜索过程确定的，如第 III-A节所述。

图 3说明了在不同设备变化水平 σ_d 下，使用不同方法训练的模型的 Top-1 推理精度，遵循第 III-A节讨论的噪声模型。OVF 在大多数设备值偏差值上明显超越所有基线，并且在罕见情况下，当设备变化太小而无法

产生显著影响时，表现与基线相似。与高斯噪声注入训练基准相比，OVF 在 VGG-8 对于 MNIST 和 CIFAR-10、ResNet-18 对于 MNIST、CIFAR-10、CIFAR-100 和 Tiny ImageNet 的 Top-1 精度分别提高了 46.71%、6.78%、5.35%、16.30%、17.21% 和 3.22%。OVF 基于整体前向性能来约束网络，特别适合尚未达到其表示能力极限的网络，如 VGG-8 在 MNIST 上的情况。

OVF 方法的有效性突出了负优化训练机制在增强 DNN 对抗设备变异的鲁棒性方面的通用性和实用性，从而有助于 NVCIM 加速器的可持续部署。

C. 不确定性和收敛性

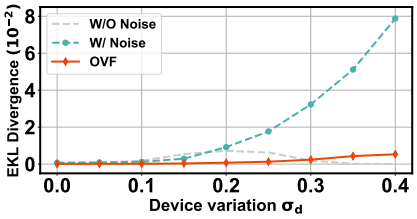

图 4. 正确预测的不同方法的平均 EKL 散度。

器件变化放大了认识不确定性，导致输出不确定性增加。为了量化器件变化对不确定性的的影响，我们采用期望 Kullback-Leibler (EKL) 散度 [14]（较低的值表示更好的性能）。为了公平比较，排除了准确性分析。具体来说，在所有噪声推理中的正确预测中，我们计算每个 softmax 输出与其对应标签之间的 Kullback-Leibler 散度。结果显示在图 4中，代表每个正确预测的平均 EKL 散度。与无噪声基线相比，带噪声的基线提高了准确性，但以增加不确定性为代价。相比之下，我们的 OVF 方法不仅比注入噪声训练基线达到更高的准确性，还保持了低不确定性和高输出置信度。在器件变化过大以至于无法使用普通训练基线进行有效预测的情况下，其 EKL 散度略低于 OVF，因为它生成完全随机且无意义的预测。

对于某些设备和与老化相关的问题，设备变化可能非常显著。OVF 实现的低不确定性也有助于模型收敛。例如，在使用 $\sigma_d = 0.35$ 对 MNIST 进行 VGG-8 的 10 次独立实验运行中，噪声注入训练和 OVF 的非收敛模型数量 [2] 分别为 6 和 0。除了准确率提高之外，这可能部分解释了图 3所示 OVF 在 MNIST 上的 VGG-8 上显著改进的原因。

[2]其中准确率相比多次独立运行的平均准确率下降超过 5%的地方

IV. 结论

总之，所提出的定向变分前向（OVF）方法显著增强了深度神经网络（DNNs）对抗设备变化的鲁棒性，并促进了 NVCIM 加速器的可持续部署。通过保持高精度的同时降低不确定性，OVF 减少了对写入验证操作的依赖，提高了部署时间和能源效率，并实现了相同设备上的芯片具有更高的推理精度。该方法展示了负优化训练机制的通用性和实用性，为开发能够适应 NVM 器件非理想特性的稳健 AI 加速器提供了宝贵的见解，从而促进了 AI 硬件的可持续发展。

致谢

该项研究得到了由香港特别行政区 InnoHK 资助的 ACCESS——新兴智能系统人工智能芯片中心的支持。

参考文献

[1] Y.-H. Chen, J. Emer, and V. Sze, "Eyeriss: A spatial architecture for energy-efficient dataflow for convolutional neural networks," *ACM SIGARCH computer architecture news*, vol. 44, no. 3, pp. 367–379, 2016.

[2] A. Shafiee, A. Nag, N. Muralimanohar, R. Balasubramonian, J. P. Strachan, M. Hu, R. S. Williams, and V. Srikumar, "Isaac: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars," *ACM SIGARCH Computer Architecture News*, vol. 44, no. 3, pp. 14–26, 2016.

[3] W. Zhang, P. Yao, B. Gao, Q. Liu, D. Wu, Q. Zhang, Y. Li, Q. Qin, J. Li, Z. Zhu *et al.*, "Edge learning using a fully integrated neuro-inspired memristor chip," *Science*, vol. 381, no. 6663, pp. 1205–1211, 2023.

[4] D. Reis, M. Niemier, and X. S. Hu, "Computing in memory with fefets," in *Proceedings of the international symposium on low power electronics and design*, 2018, pp. 1–6.

[5] Y. Qin, R. Kuang, X. Huang, Y. Li, J. Chen, and X. Miao, "Design of high robustness bnn inference accelerator based on binary memristors," *IEEE Transactions on Electron Devices*, vol. 67, no. 8, pp. 3435–3441, 2020.

[6] S. Angizi, Z. He, A. Awad, and D. Fan, "Mrima: An mram-based in-memory accelerator," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 39, no. 5, pp. 1123–1136, 2019.

[7] X. Sun, W. Khwa, Y. Chen, C. Lee, H. Lee, S. Yu, R. Naous, J. Wu, T. Chen, X. Bao *et al.*, "Pcm-based analog compute-in-memory: Impact of device non-idealities on inference accuracy," *IEEE Transactions on Electron Devices*, vol. 68, no. 11, pp. 5585–5591, 2021.

[8] Z. Yan, Y. Qin, W. Wen, X. S. Hu, and Y. Shi, "Improving realistic worst-case performance of nvcim dnn accelerators through training with right-censored gaussian noise," in *2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD)*. IEEE, 2023, pp. 1–9.

[9] M. Rizzi, A. Spessot, P. Fantini, and D. Ielmini, "Role of mechanical stress in the resistance drift of ge2sb2te5 films and phase change memories," *Applied Physics Letters*, vol. 99, no. 22, 2011.

[10] Z. Yan, D.-C. Juan, X. S. Hu, and Y. Shi, "Uncertainty modeling of emerging device based computing-in-memory neural accelerators with application to neural architecture search," in *Proceedings of the 26th Asia and South Pacific Design Automation Conference*, 2021, pp. 859–864.

[11] R. Degraeve, A. Fantini, N. Raghavan, L. Goux, S. Clima, B. Govoreanu, A. Belmonte, D. Linten, and M. Jurczak, "Causes and consequences of the stochastic aspect of filamentary rram," *Microelectronic Engineering*, vol. 147, pp. 171–175, 2015.

[12] W. Shim, J.-s. Seo, and S. Yu, "Two-step write–verify scheme and impact of the read noise in multilevel rram-based inference engine," *Semiconductor Science and Technology*, vol. 35, no. 11, p. 115026, 2020.

[13] A. Eldebiky, G. L. Zhang, G. Böcherer, B. Li, and U. Schlichtmann, "Correctnet: Robustness enhancement of analog in-memory computing for neural networks by error suppression and compensation," in *2023 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2023, pp. 1–6.

[14] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher *et al.*, "A survey of uncertainty in deep neural networks," *Artificial Intelligence Review*, vol. 56, no. Suppl 1, pp. 1513–1589, 2023.