面向印刷式多层感知机的硬件感知自动神经网络最小化

Argyris Kokkinis*、Georgios Zervakis§、Kostas Siozios*、Mehdi B. Tahoori‡ 和 J¨org Henkel‡
*希腊塞萨洛尼基亚里士多德大学，§希腊帕特雷大学，‡德国卡尔斯鲁厄理工学院
*{arkokkin、ksiop}@auth.gr，§zervakis@ceid.upatras.gr，‡ {mehdi.tahoori、henkel}@kit.edu

arXiv:2301.11142v1 [cs.AR] 26 Jan 2023

摘要——许多应用领域对灵活性、可拉伸性和多孔性的需求，通常无法通过硅基超大规模集成电路技术满足。印刷电子（PE）作为一种候选解决方案应运而生，既能满足这些要求，又能以极低成本将智能设备集成到消费品上，并实现现场按需制造。然而，印刷电子中的特征尺寸较大，限制了这些努力；受面积和功耗限制，也无法设计复杂的机器学习电路。不过，分类是印刷应用的核心任务。本研究首次考察了神经网络最小化技术与定制电路实现相结合，对印刷式多层感知机分类器面积效率的影响。结果表明，在准确率损失不超过 5% 的情况下，面积最多可缩减至原来的八分之一。

索引词—**近似计算、多层感知机、神经网络最小化、印刷电子**

I. 引言与动机

印刷电子（PE）能够提供柔性、轻量且低成本的设备，这些设备可集成到日常消费品中，从而为低端医疗产品、一次性用品、包装食品、饮料等提供智能服务， [1]。尽管印刷电子具有诸多优势，但其较大的特征尺寸导致所设计的电路面积大且功耗高，难以部署在小型表面上，也难以满足严格的电池供电要求。尤其是，多层感知机（MLP）已被证明难以映射到印刷技术中，同时又不牺牲模型精度或设计的可行性 [1], [2]。有人提出，在印刷电子中设计完全定制的，即定制实现的电路，作为优化面积和功耗开销的一种可能方案 [1], [3]。在定制实现中，模型的系数被硬连线固化在电路中，从而能够针对每个模型进行定制优化。由于印刷电子具有极低的制造成本和非经常性工程成本，这种方法只有在印刷电子中才真正可行。

神经网络最小化技术广泛应用于机器学习（ML）领域和深度神经网络（DNNs）中，旨在压缩神经网络，并在容许一定精度损失的情况下提升其性能。在本研究中，我们首次考察了神经网络最小化技术对印刷式多层感知机分类器设计的影响。具体而言，我们讨论并评估了量化、剪枝和权重聚类对定制多层感知机实现面积效率的影响。结果表明，印刷式多层感知机的面积最多可缩小至原来的八分之一，精度损失不超过 5%。

II. 神经网络最小化技术

本文探讨了量化、剪枝和权重聚类对硬件的影响。前两种技术旨在降低位宽并减少神经网络的权重数量，第三种技术则用于减少定制乘法器的数量。在印刷式定制架构中，结合使用这些技术有望缩小多层感知机的规模，并降低面积和功耗。

A. 量化

量化或精度缩放是神经网络最广泛使用的近似技术之一 [4], ，因为它保留了机器学习模型中常见计算模式的规律性。对于适度的量化，精度损失可以忽略不计；但在精度极低时，精度下降可能会变得显著。如 [2],所示，在定制实现中，机器学习电路的面积直接取决于所选系数。在定制多层感知机中，网络权重的取值以及输入的位表示直接决定所需乘法器的面积需求，进而决定加法器的面积。因此，将网络权重量化为较低位宽，可得到更“硬件友好”的权重，从而构建更小的算术单元，进而实现面积效率更高的神经元。

B. 剪枝

结构化剪枝和非结构化剪枝是压缩深度神经网络、并有望跳过运算以提升性能的成熟方法。结构化剪枝会从模型中移除节点或层，而非结构化剪枝则会减少神经连接的数量。前者因能直接提升性能而更受青睐，尽管后者在稀疏性相近时主要能带来更高的精度。另一方面，定制电路可以无缝受益于非结构化剪枝，因为印刷式多层感知机实现中的权重是硬连线的。移除某个连接后，相应的乘法器便会直接从电路中移除。此外，相应神经元所需的求和操作数也会减少，从而使加法更节省面积。

C. 权重聚类

权重聚类也是一种压缩技术，已被用于降低深度神经网络的内存需求 [5]。对于定制多层感知机，权重聚类可

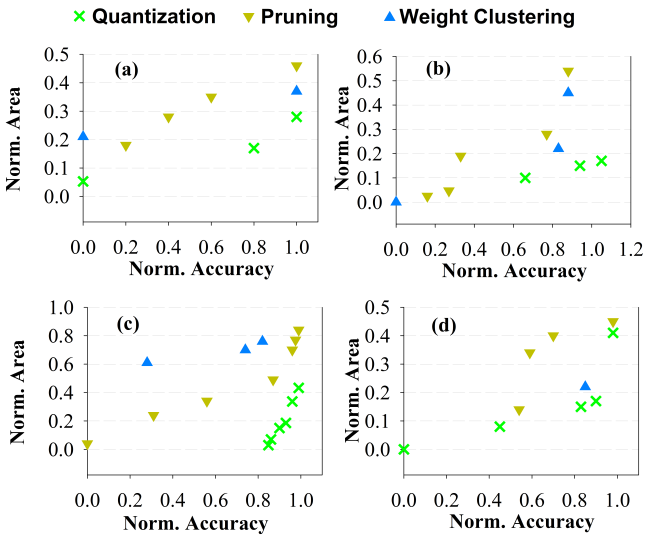

图 1：采用量化、剪枝和权重聚类的印刷多层感知机的面积-准确率权衡 [5]。数值相对于各自的基线多层感知机进行归一化 [1]。分类器：(a) 白葡萄酒，(b) 红葡萄酒，(c) Pendigits，(d) Seeds。

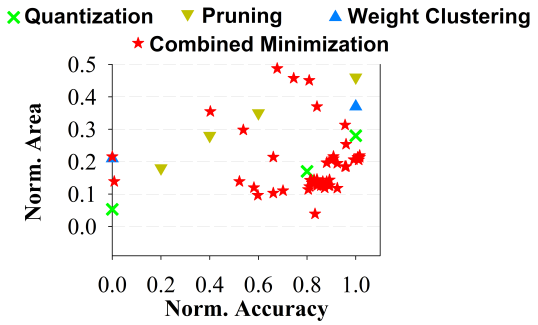

图 2：结合量化、剪枝、权重聚类以及全部三种最小化技术时，白葡萄酒多层感知机分类器的面积-准确率权衡。

用于降低面积需求。通过将处于相同位置（即与相同输入相乘）的权重设为相同值，乘积便可由多个运算共享，从而相应减少所需的乘法器单元数量。我们采用 [5] 中的权重聚类，以在硬件层级实现乘法器共享。

III. 评估分析

本节在四个数据集上评估了上述最小化技术在缩减印刷多层感知机面积方面的效率：UCI 机器学习知识库中的 WhiteWine 、 RedWine 、 Pendig-its 和 Seeds 数据集 [6]。我们使用 Qkeras [7] 对神经网络进行量化，并执行量化感知（再）训练（QAT）。综合和硬件分析采用 Synopsys Design Compiler、PrimeTime 工具以及开源电解质栅控晶体管（EGT）库 [3]。本次评估分析以各自未经最小化的定制多层感知机 [1]为基线。

图 1 展示了将三种技术分别单独应用于所研究分类器时的准确率-面积帕累托前沿。图 1 中的坐标轴相对于基线的面积和准确率数值进行了归一化 [1]。本次评估考察了稀疏度介于 20% 至 60% 之间的非结构化剪枝。量化帕累托曲线是通过评估多个设计生成的，这些设计中分类器量化权重的位精度介于 2 至 7 位之间。最后，权重聚类帕累托点是通过针对选定范围的簇运行算法 [5] 得到的。不出所料，四种最小化技术都能生成更小的分类器，以准确率损失换取面积缩减。量化帕累托前沿优于剪枝和权重聚类：准确率损失不超过 5% 时，平均可将面积缩减 5 倍；在相同准确率约束下，剪枝和权重聚类技术带来的面积缩减分别为 2.8 倍和 3.5 倍。需要注意的是，只有在 RedWine 和 WhiteWine 数据集上，权重聚类方法生成的多层感知机才能满足 5% 的准确率阈值。

最后，图 2 展示了结合全部三种最小化技术时的面积-准确率权衡。为获得这些设计，我们采用了硬件感知遗传算法。如图所示，结合所有技术所得到的设计兼具高准确率和较小面积，优于单独使用各项技术的效果。有趣的是，在 5% 的准确率阈值下，面积最高可缩减至原来的八分之一。

IV. 结论

印刷电子是实现智能服务的一种颇具前景的解决方案，适用于计算技术普及有限的应用领域。然而，较高的硬件开销阻碍了复杂印刷机器学习系统的实现。在本研究中，我们展示了最初为节省内存而设计的神经网络最小化技术，如何为采用多层感知机实现印刷机器学习分类铺平道路。

致谢

本研究得到欧盟委员会资助计划“SERRANO”的支持，该计划由地平线2020资助，项目编号为101017168；同时得到欧盟（欧洲研究委员会，PRICOM，项目编号101052764）的支持，并获得德国研究联合会通过项目“ACCROSS：跨越整个系统栈的近似计算”（HE 2343/16-1）提供的支持。

参考文献

[1] M. H. Mubarik 等，“印刷式机器学习分类器”，载于年度国际微体系结构研讨会（*MICRO*），2020年，第73–87页。[2] G. Armeniakos 等，“印刷式机器学习电路的跨层近似”，载于欧洲设计、自动化与测试会议暨展览会，2022年。[3] N. Bleier 等，“印刷式微处理器”，载于年度国际计算机体系结构研讨会（*ISCA*），2020年6月，第213–226页。[4] J. Henkel 等，“近似计算与高效机器学习之旅”，载于国际计算机辅助设计会议，2022年。[5] S. Han 等，“深度压缩：利用剪枝、训练后量化和霍夫曼编码压缩深度神经网络”，载于国际学习表征会议，*ICLR*，2016年。[6] D. Dua 等，“UCI机器学习库”，2017年。[7] C. N. Coelho 等，“面向粒子探测器边缘端低延迟推理的深度神经网络异构量化自动化”，*arXiv:2006.10159*，2020年。