嵌入式机器学习调研报告

徐浩博 2020010108

摘 要

伴随机器学习算法的兴起和计算机体系结构的进步，二者的交叉领域也日渐兴盛。然而，机器学习的算法对于资源的要求较高，并且随着研究发展，机器学习模型还将呈现越来越大的特点，因此机器学习算法如何应用于资源受限的嵌入式系统也成为了问题的关键。本文将对嵌入式机器学习的解决方案从软件和硬件两个层面进行系统回顾，除此之外，本文还将对相关应用进行总结并进行了前景展望。

关键词：嵌入式系统；机器学习；资源受限

1 引言

机器学习及其算法已经广泛应用于生产生活中，如目标检测、推荐系统、问答系统等。然而，这些模型和算法大都需要充足的内存、CPU和GPU算力，对于生活中广泛存在的嵌入式系统而言，性能成为了机器学习应用的最大阻碍。

其中一种实现嵌入式智能系统的方案是利用云技术 [11]。具体而言，嵌入式系统的传感器收集数据，并通过互联网将数据转发至云端，云端利用机器学习模型对数据进行推理，并将结果广播转发给嵌入式边缘系统，从而实现嵌入式系统的机器学习应用。然而，该种方案存在局限性。一方面，数据传输过程中的隐私性和安全性问题亟需解决。另一方面，在某些应用之中，实时性的要求不允许云方案——如自动驾驶领域，延迟往往对汽车行驶的安全性造成致命威胁。最后，通讯会导致大量设备功耗，在这种情况下处理数据往往比数据通讯更经济。

从而，如何利用嵌入式设备中资源紧缺的硬件设备进行机器学习，并拓展相关应用成为关键问题。本文将围绕嵌入式的机器学习这一问题，从以下几个维度展开阐述：

* 第2节：从机器学习算法层面优化以适应资源紧缺的嵌入式设备
* 第3节：从嵌入式设备硬件层面适应机器学习算法
* 第4节：嵌入式的机器学习算法相关应用

2 软件层面

**2.1 减小分布式通信开销**

如第1节引言所述，云式机器学习推理可能导致巨大通信开销，在实时性要求较高的领域无法真正推广应用。因此一个软件层面优化的一个思路采用分布式学习推理，并对通信开销进行优化。

DGC [1]进一步探索了梯度交换在分布式学习中的作用，通过大幅度降低梯度交换和进行深度梯度压缩，DGC将通信带宽降低了两个数量级，也压缩了模型大小，同时保证了模型精度。该方法可以实现小型终端设备在图像分类、语音识别等任务上拥有出色表现。

**2.2 机器学习模型优化**

模型优化是基于机器学习模型的特点，对参数或架构进行的优化，常用的方法如网络稀疏化、模型压缩、采用精度更低的参数等。

Protoknn [2]是基于KNN算法的优化算法，实际应用于一些资源稀缺的设备，用来适应小内存和低复杂性的要求。具体来说，ProtoNN的优化思想来源于三个部分：用原型代表整个训练数据集，从而满足小内存的条件；低维稀疏投影降低复杂度；除此之外，显式模型约束也起着重要作用。该算法较少牺牲了预测准确性，但大大降低了内存和计算的要求。

FastRNN / FastGRNN [3]是针对RNN作出的优化算法。门控RNN模型通过增大模型大小，提高预测成本来换取精度上的优势，然而并不适用于小型资源受限的机器。FastRNN / FastGRNN通过对于门控RNN模型进行优化，增加残差连接，并重用低秩、稀疏的RNN矩阵以满足精度的需求，并且比常见门控RNN模型参数小若干数量级，但精度并没有显式差别。最终模型在实际可以放入具有 2 KB RAM 和 32 KB 闪存的 MCU。

**2.3 资源稀缺的机器学习相关工具**

资源稀缺的MCU可能没有浮点计算单元FPU，从而可能会对大多数机器学习算法的实现造成阻碍——大多数机器学习算法大量依赖浮点数运算，缺少浮点数运算可能会严重影响算法实现。SeeDot [4]是一种领域特定语言DSL，目标是克服这一问题，通过将浮点运算转换为高效定点运算，同时还提出一款编译器，能够满足浮点数运算代码在无FPU的MCU上高效运行。

开源软件包为嵌入式设备运行机器学习算法提供更加便利的条件。以FANN-on-MCU [5]为例，它是基于快速神经网络FANN库构建的开源软件包，可以针对低功耗微控制器实现多层感知机MLP并生成具体代码，并进行具体优化。该软件包可在基于 ARM Cortex-M 系列和基于 RISC-V 的新型并行超低功耗 （PULP） 平台的微控制器上运行。

除此之外，仍有许多或企业级别或实验室级别的开源平台，[6]总结了常见的若干开源软件平台，并详细回顾了相关生态系统。

3 硬件优化

**3.1 FPGA**

机器学习的训练和推理往往依靠硬件设备，如CPU、GPU和专用的集成电路ASIC等。对于深度学习，巨大的模型往往使得开发者寻求GPU或专用电路（如TPU）等的帮助。而对于嵌入式设备，体积、能耗功率等限制了GPU或TPU的使用，因此设计FPGA专门对神经网络进行加速是一条解决途径。这些FPGA往往具有低成本、能耗低、可重构等优点，因此较为适宜用于嵌入式设备之中。

举例来说，Qiu et al. [9]提出了一种在嵌入式设备加速大规模卷积神经网络的软硬件协同方法——在软件层面分解重构了CNN的模型架构，在硬件层面设计基于嵌入式FPGA的CNN硬件加速器。该嵌入式FPGA平台可以实现较深的VGG模型，并应用于ImageNet大规模图片分类。更多基于FPGA的设计可以参见[11]。

然而限制基于FPGA的嵌入式设备发展的主要问题在于FPGA很难设计和编程，一个较为通用的相关FPGA设计架构可能有助于缓解这一问题。

**3.2 商业硬件加速器**

对于终端设备上的特定任务，尤其是视觉处理任务，如智能手机上的视频智能分析，各大厂商已有一些成熟的解决方案。我们以Synopsys最新推出的视觉加速处理器EV7x [10]为例进行介绍。

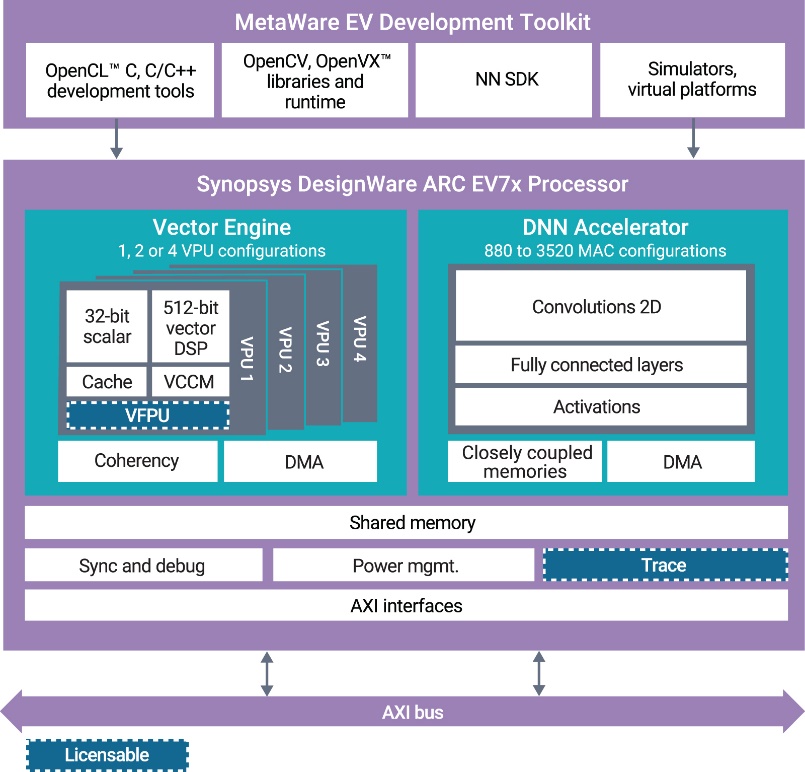


Fig.1 Synopsys EV7x 处理器

Synopsys的EV7x处理器集成了矢量 DSP、矢量 FPU 和神经网络加速器，配备了多达4个增强型矢量处理单元VPU和特殊的DNN加速器。VPU可以针对视频处理任务加速，提高视觉任务处理性能。除此之外，DNN还支持除CNN之外的神经网络模型，如LSTM等，可以适应基于时间序列的特殊任务，如移动行人位置预测。该处理器可以与所有主机处理器配套使用，常见应用场景包括驾驶辅助、视频监控、增强虚拟现实设备等。

4 嵌入式机器学习应用

**4.1 工业应用**

我们以工程学中的结构健康检测为例介绍嵌入式设备在工业生产中的应用。结构健康检测是通过分析各种传感器数据获取建筑等结构健康状况信息的任务，从而降低结构损坏对人身、财产安全造成威胁。将机器学习技术嵌入微型传感器进行结构健康监测是工业界最新的研究方向。Katsidimas et al. [12] 将问题定义为塑料板撞击振动事件的感知，并在 Arduino NANO 33 BLE 上实施并基准测试随机森林和浅层神经网络模型，实现了在10kB大小左右的95%以上的准确率。

**4.2 生活领域应用**

个人医疗设备往往依靠嵌入式设备，如心脏起搏器等。在此类设备中应用机器学习模型有助于预测健康危害。Shoeb et al. [7] 通过将支持向量机SVM嵌入植入式神经刺激器中，可以监控患者脑电数据，并预测患者癫痫发作，比非特异性检测器具有更好的性能和更低功耗。

自动驾驶领域也广泛嵌入智能。汽车具有的多个MCU提供的大量数据涌入中央处理器必然会降低处理效率，而智能传感器中运行机器学习模型自行分类处理，可以减少中央处理器处理时间；同时，微型MCU自身具有处理数据的能力，降低了路面出现网络死区阻碍通讯酿成事故的可能。在视觉和语音检测方面，嵌入智能已经可以实现辅助避障。[8]

除此之外，各大商业芯片厂商已经设计制造出成熟的视觉机器学习加速器，详见3.2节。它们被广泛应用在终端设备中，如视频智能分析、辅助驾驶、VR等。

5 结论和展望

本文系统回顾了在嵌入式设备中应用机器学习算法的相关解决方案，并通过硬件和软件两个层面，从多个角度阐述相关科研进展。之后，本文还简单介绍了嵌入式设备中应用机器学习算法的相关应用。伴随机器学习的兴盛，嵌入式系统和机器学习这一交叉点呈现出越来越蓬勃发展的态势，具有广阔前景，并且已经开始逐渐改变我们的生活。相信随着研究进一步深入和商业公司进一步投入，这一领域还将持续焕发生机，并为人类建构新的生活模态。

参考文献

1. Lin, Y., Han, S., Mao, H., Wang, Y., & Dally, B. Deep Gradient Compression: Reducing the Communication Bandwidth for Distributed Training. *In International Conference on Learning Representations (ICLR)*.
2. Gupta, C., Suggala, A. S., Goyal, A., Simhadri, H. V., Paranjape, B., Kumar, A., ... & Jain, P. (2017, July). Protonn: Compressed and accurate knn for resource-scarce devices. *In International conference on machine learning (ICML)*. PMLR.
3. Kusupati, A., Singh, M., Bhatia, K., Kumar, A., Jain, P., & Varma, M. (2018). Fastgrnn: A fast, accurate, stable and tiny kilobyte sized gated recurrent neural network. *Advances in neural information processing systems (NeurIPS)*, 31.
4. Gopinath, S., Ghanathe, N., Seshadri, V., & Sharma, R. (2019, June). Compiling KB-sized machine learning models to tiny IoT devices. *In Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI)*.
5. Wang, X., Magno, M., Cavigelli, L., & Benini, L. (2020). FANN-on-MCU: An open-source toolkit for energy-efficient neural network inference at the edge of the Internet of Things. *IEEE Internet of Things Journal (IoTJ)*, 7(5), 4403-4417.
6. Sanchez-Iborra, R., & Skarmeta, A. F. (2020). Tinyml-enabled frugal smart objects: Challenges and opportunities. *IEEE Circuits and Systems Magazine*, 20(3), 4-18.
7. Shoeb, A., Carlson, D., Panken, E., & Denison, T. (2009, September). A micropower support vector machine based seizure detection architecture for embedded medical devices. *In 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 4202-4205). IEEE.
8. Chen, J., & Ran, X. (2019). Deep learning with edge computing: A review. *Proceedings of the IEEE*, 107(8), 1655-1674.
9. Qiu, J., Wang, J., Yao, S., Guo, K., Li, B., Zhou, E., ... & Yang, H. (2016, February). Going deeper with embedded fpga platform for convolutional neural network. *In Proceedings of the 2016 ACM/SIGDA international symposium on field-programmable gate arrays (FPGA)*.
10. Synopsys EV7x Vision Processors. <https://www.synopsys.com/dw/ipdir.php?ds=ev7x-vision-processors>
11. Branco, S., Ferreira, A. G., & Cabral, J. (2019). Machine learning in resource-scarce embedded systems, FPGAs, and end-devices: A survey.*Electronics*, 8(11), 1289.
12. Ajani, T. S., Imoize, A. L., & Atayero, A. A. (2021). An overview of machine learning within embedded and mobile devices–optimizations and applications.*Sensors*, 21(13), 4412.
13. Katsidimas, I., Kotzakolios, T., Nikoletseas, S., Panagiotou, S. H., & Tsakonas, C. (2022, November). Smart Objects: Impact localization powered by TinyML. *In Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems (SenSys’ 22)*.