重新思考微型机器学习的时序模型：资源受限设备中的长短期记忆网络与一维卷积神经网络之比较

Bidyut Saha Sister Nivedita University，Newtown，印度
bidyut.s@snuniv.ac.in

Riya Samanta Techno India University，Salt Lake，印度
riya.s@technoindiaeducation.com

arXiv:2603.04860v1 [cs.PF] 5 Mar 2026

摘要

时间序列分类是人类活动识别、医疗监测以及物联网领域手势检测等应用的基础。微型机器学习（TinyML）使模型能够直接运行在低功耗微控制器（MCU）上，从而提高效率、保障隐私并降低成本，避免依赖云计算或边缘计算。长短期记忆网络（LSTM）虽广泛用于捕捉时间依赖关系，但其较高的计算与内存需求使其难以在微控制器上实现实时部署。本研究针对五个基准数据集，对长短期记忆网络与一维卷积神经网络（一维卷积神经网络）开展了硬件感知可行性研究。结果表明，一维卷积神经网络始终能达到与长短期记忆网络相当或更高的准确率（≈95% 对比 ≈89%），同时减少 35% 的随机存取存储器用量、约 25% 的闪存用量，并实现实时推理（27.6 毫秒，对比 2038 毫秒）。一维卷积神经网络十分轻量，尤其适合在可穿戴设备及其他低功耗电池供电系统中进行设备端处理，因此是微型机器学习部署中实用且节省资源的选择。

CCS 概念

• **以人为中心的计算** → 普适与移动计算；• **计算方法** → 基于分类的监督学习；• **硬件** → 新兴架构；• **计算机系统组织** → 嵌入式系统。

关键词

微型机器学习、时间序列分类、一维卷积神经网络、长短期记忆网络、资源受限设备、微控制器、端侧机器学习、人类活动识别、高能效人工智能

ACM 参考格式：
Bidyut Saha 和 Riya Samanta。2026。重新思考微型机器学习的时序模型：资源受限设备上的长短期记忆网络与一维卷积神经网络之比较。载于。 ACM，纽约，美国，3 页。https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 引言

时间序列分类为人体活动识别、手势检测和医疗监测等应用提供支持。

表1：一维卷积神经网络与长短期记忆网络在五个基准数据集上、在Float32和INT8量化条件下的性能，报告了在ESP32上的准确率、随机存取存储器、闪存和推理延迟。

数据集	模型	准确率（32/INT8）(%)	RAM (KB)	闪存 (KB)	延迟 (ms)
UCIHAR	一维卷积神经网络	93.34 / 92.03	20.38	95.12	28
	LSTM	92.38 / 88.09	28.01	124.02	2356
PAMAP2	一维卷积神经网络	96.12 / 95.60	13.59	94.67	10
	LSTM	95.81 / 90.80	15.27	123.66	746
WISDM	一维卷积神经网络	94.77 / 94.34	13.59	94.92	20
	LSTM	93.55 / 89.21	15.27	123.65	502
MIT-BIH心律失常数据库	一维卷积神经网络	97.24 / 97.15	24.84	94.28	40
	LSTM	85.21 / 78.22	38.47	123.25	3233
PTB	一维卷积神经网络	97.97 / 97.66	24.84	94.31	40
	LSTM	82.65 / 79.35	36.45	122.71	3354
平均值	一维卷积神经网络	**95.49 / 95.36**	**19.85**	**94.66**	**27.6**
	LSTM	**89.52 / 85.53**	**30.29**	**123.46**	**2038.2**

尽管长短期记忆网络（LSTM）是建模时序依赖的常用选择 [4], ，但其较高的内存和计算开销使其不适合在微控制器（MCU）上进行实时设备端计算。微控制器通常仅提供数百 KB 的静态随机存取存储器（SRAM）和不足 1 MB 的闪存。现有解决方案往往依赖边缘端或云端卸载 [8], ，但这会增加延迟、能耗和基础设施成本 [10], ，带来数据隐私风险 [12] ，并造成对网络连接的依赖 [9]。

微型机器学习（TinyML）支持在低成本、电池供电的微控制器上直接进行设备端推理，兼具高效、保护隐私和价格实惠等优势 [1]。然而，要切实可行，模型必须在准确率与严格的资源限制之间取得平衡。一种颇具前景的替代方案是一维卷积神经网络（1D-CNN）。尽管它最初是为提取空间特征而设计的，但在时间序列任务中也展现出潜力，且资源占用显著更低。

本研究的贡献如下：

（1）针对微型机器学习中的时间序列分类，对长短期记忆网络与一维卷积神经网络架构进行硬件感知、公平的对比。（2）制定以可部署性为重点的评估方案，涵盖准确率、随机存取存储器/闪存占用、推理延迟，以及量化对微控制器的影响。（3）基于五个涵盖人机交互、生活方式和医疗保健领域的基准数据集提供实证证据，表明一维卷积神经网络的准确率始终与长短期记忆网络相当或更高，同时所需资源更少，并支持实时性能。

这些结果表明，一维卷积神经网络轻量且实用，适用于设备端处理，尤其适用于可穿戴设备以及低功耗的电池供电系统。

允许免费制作本作品的全部或部分内容的电子或纸质副本，供个人或课堂使用，前提是这些副本不得用于或分发以获取利润或商业利益，并且副本须附有本声明及首页上的完整引文。对于本作品中非由作者拥有版权的部分，必须尊重其版权。注明出处后允许摘录。以其他方式复制、再版、发布到服务器或重新分发到列表，均须事先获得明确许可和/或支付费用。请向 permissions@acm.org 申请许可。
*Conference'17，2017*年7月，美国华盛顿哥伦比亚特区© 2026 版权归所有者/作者所有。出版权授权给 ACM。ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

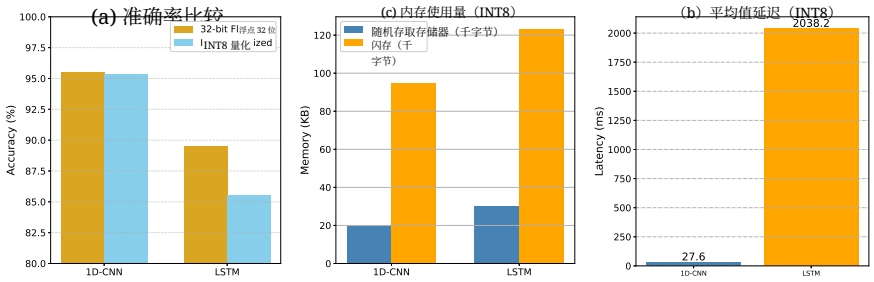


图1：平均值性能 **在 ESP32 上跨 5 个数据集的一维卷积神经网络和长短期记忆网络：（a）准确率、（b）内存使用量，以及（c）推理延迟（毫 s).**

2 方法

数据集与预处理。我们在五个基准数据集上进行评估 . 在人机交互与生活方式领域，我们使用 UCI-HAR [3], PAMAP 2 [6], 和 WISDM [11], , 它们包含用于人体活动识别的惯性测量单元信号。在医疗保健领域，我们纳入 MIT-BIH 心律失常数据库 [5] 和 PTB 诊断性心电图数据库 [2], , 这两个数据库均提供随时间变化的心电图信号。数据预处理遵循 [7] 中描述的流程。

模型类别。 我们选取两种具有代表性的架构——一维卷积神经网络（1D-CNN）和长短期记忆网络（LSTM）——因为它们体现了时间序列数据建模的两种主要方法：通过卷积运算提取局部模式（一维卷积神经网络），以及通过循环建模捕捉长程时间依赖关系（长短期记忆网络）。为清晰起见，本文所用的一维卷积神经网络特指采用深度可分离卷积的实现方式，该方式在保持表征能力的同时提高了计算效率。我们并未穷尽评估各种架构变体，而是聚焦于这些模型的经典容量受控版本。此外，我们还开展了有针对性的消融研究，以证明观察到的权衡在典型设计选择下具有普适性。

在我们的实验中，两种架构的实现方式如下：

(1) **一维卷积神经网络模型：**三个可分离卷积层（滤波器数量分别为 32、48 和 72，卷积核大小为 = 3），每层之后均接最大池化。全局平均池化层用于降低时间维度，随后是一个全连接层（72 个单元，ReLU）、随机失活，以及最终的 Softmax 输出层。

(2) **长短期记忆网络模型：**采用包含 32、48 和 72 个单元的堆叠式长短期记忆网络，以捕捉长程时间依赖。最终输出依次经过全连接层（72 个单元，ReLU）、随机失活和 Softmax 输出层。

训练协议。 两种模型使用相同的优化器、批次大小和训练轮数预算，并根据验证损失进行提前停止。超参数（学习率、卷积核大小/隐藏单元）通过一个小型网格搜索，依据验证性能进行选择，以避免测试集信息泄漏。随机种子固定，以确保结果可复现。

指标。 我们将非量化的 Float32 模型和 INT8 量化模型的分类准确率作为主要性能指标进行报告。此外，我们还评估模型大小（闪存，以千字节计）、推理期间的峰值 RAM 使用量（以千字节计），以及在以 240 兆赫运行的 ESP32 微控制器上测得的设备端延迟（毫秒）。资源占用均统一针对 INT8 量化模型进行报告。

优化与设备端执行。为确保这些模型能够在严格的资源限制下部署于微控制器，我们设计了轻量级架构版本。进一步的优化通过量化实现。具体而言，我们使用*TensorFlow* 模型优化工具包应用 INT8 训练后量化，以减小模型大小并降低计算需求。对于设备端执行，优化后的模型通过*TensorFlow Lite Micro* 框架部署在 ESP32 微控制器平台上。

3 结果与讨论

表1展示了使用一维卷积神经网络和长短期记忆网络模型在全部五个数据集上的详细结果。对于准确率，我们同时报告未量化的 Float32 版本和 INT8 量化版本。如图1(a)所示，图中绘制了平均准确率；无论在 Float32 还是 INT8 设置下，一维卷积神经网络的表现始终优于长短期记忆网络。一个重要发现是，INT8 量化对一维卷积神经网络准确率的影响可以忽略不计，而长短期记忆网络的性能则显著下降，这表明卷积神经网络模型在量化条件下更加稳健。图1(b)进一步展示了其内存优势：与长短期记忆网络相比，一维卷积神经网络所需的随机存取存储器和闪存大幅减少。同样，图1(c)突显了延迟方面的巨大差距——一维卷积神经网络能够保持实时运行（平均 27.6 毫秒），而长短期记忆网络的运行时间则高得难以接受，每次推理都超过 2 秒。

总体而言，表格和图形证据相结合表明，在资源极其受限的微型机器学习环境中（此类环境通常必须采用 INT8量化），一维卷积神经网络比长短期记忆网络实用得多：它在保持更高准确率的同时，延迟更低、内存占用更小。

4 局限性与未来工作。

本研究为未来在分类以外开展预测及其他时间序列应用的探索奠定了坚实基础。所考察的特征形状——时长约为 2 秒，

采样率最高达 50 Hz——代表了许多微型机器学习应用中的实际场景。基于这些结果，未来的工作将把实验扩展到更长的序列、更高频的数据以及更多应用领域，进一步提升所提方法的稳健性和通用性。

参考文献

[1] Colby Banbury、Chuteng Zhou、Igor Fedorov、Ramon Matas、Urmish Thakker、Dibakar Gope、Vijay Janapa Reddi、Matthew Mattina 和 Paul Whatmough。2021 年。Micronets：用于在商用微控制器上部署微型机器学习应用的神经网络架构。机器学习与系统论文集3（2021），517–532。[2] Ralf Bousseljot、Dieter Kreiseler 和 Allard Schnabel。1995 年。通过互联网使用 PTB 的 CARDIODAT 心电信号数据库。（1995）。[3] Reyes Ortiz 等。2012 年。使用智能手机进行人体活动识别。UCI机器学习存储库。DOI: https://doi.org/10.24432/C54S4K。[4] Sepp Hochreiter 和 Jürgen Schmidhuber。1997 年。长短期记忆。神经计算9，8（1997），1735–1780。

[5] George B Moody 和 Roger G Mark。2001 年。MIT-BIH 心律失常数据库的影响。*IEEE* 医学与生物学工程杂志20，3（2001），45–50。[6] Attila Reiss。2012 年。PAMAP2身体活动监测。UCI机器学习存储库。DOI: https://doi.org/10.24432/C5NW2H。[7] Bidyut Saha、Riya Samanta、Soumya K Ghosh 和 Ram Babu Roy。2024 年。Tinytnas：无需 GPU、具有时间约束且硬件感知的神经架构搜索，用于微型机器学习时间序列分类。*arXiv* 预印本 *arXiv:2408.16535*（2024）。[8] Mahadev Satyanarayanan。2017 年。边缘计算的兴起。《计算机》50，1（2017），30–39。[9] Weisong Shi、Jie Cao、Quan Zhang、Youhuizi Li 和 Lanyu Xu。2016 年。边缘计算：愿景与挑战。*IEEE* 物联网期刊3，5（2016），637–646。[10] Blesson Varghese 和 Rajkumar Buyya。2018 年。下一代云计算：新趋势与研究方向。下一代计算机系统79（2018），849–861。[11] Gary Weiss。2019 年。WISDM 智能手机和智能手表活动与生物特征数据集。UCI机器学习存储库。DOI：https://doi.org/10.24432/C5HK59。[12] Zhiyong Zhao、Ruonan Rao、Shaoxiong Tu 和 Jun Shi。2017 年。采用重新定义标签的时间加权长短期记忆模型进行股票趋势预测。载于*2017 年 IEEE 第 29 届国际人工智能工具会议（ICTAI）*。IEEE，1210–1217。