用于脉冲神经网络仿真、序列化和互操作性的分布式压缩稀疏行格式

Felix Wang 桑迪亚国家实
验室 阿尔伯克基，新墨西哥州，
美国 felwang@sandia.gov

arXiv:2304.05587v1 [cs.DC] 12 Apr 2023

摘要

随着神经形态平台及其相关软件工具的不断发展，以及脉冲神经网络（SNN）模型规模的不断扩大，我们越来越需要具备互操作性和可扩展性的网络状态表示。为此，我们围绕支持大型 SNN 的仿真和序列化，探讨一种广泛使用的稀疏矩阵高效表示格式——压缩稀疏行（CSR）的并行扩展。用于表示图邻接结构的稀疏矩阵非常适合描述 SNN 的连接关系；此前在并行图划分领域的研究已开发出分布式 CSR（dCSR）格式，用于存储和载入大型图。我们认为，将神经元和突触状态等其他网络信息与其邻接关系对齐，并以 dCSR 形式组织起来，可以直接按分区分发网络状态。对于大规模仿真，这意味着每个并行进程只需负责自身分区的状态；当 SNN 的规模超出单个计算节点的内存资源时，这一点尤其有用。对于可能长期运行的仿真，这种方式还支持在并行进程之间基本独立地将网络序列化到磁盘或从磁盘读取（例如，用于检查点/重启容错计算）。我们还提供了一种可行的实现方案，并提议神经计算领域采用该方案。

关键词

脉冲神经网络、分布式数据格式、压缩稀疏行、高性能计算、序列化、互操作性

ACM参考格式：
FelixWang。2023年。用于脉冲神经网络仿真、序列化和互操作性的分布式压缩稀疏行格式。发表于神经启发计算元素会议（*NICE2023*），海报，*2023*年*4*月*11*日至*14*日，圣安东尼奥，德克萨斯州，美国。ACM，纽约，纽约州，美国，共3页。

本文由桑迪亚国家技术与工程解决方案有限责任公司的一名员工撰写，该公司与美国能源部（DOE）签订了合同编号为 DE-NA0003525 的合同。该员工拥有本文的全部权利、所有权及利益，并对本文内容负全部责任。美国政府保留一项非专有、已付费、不可撤销、全球性的许可；出版商接受本文发表即表示承认，美国政府有权为美国政府目的出版或复制本文的已发表形式，或允许他人这样做。美国能源部将根据美国能源部公共访问计划
https://www.energy.gov/downloads/doe-public-access-plan，向公众提供这些联邦资助研究成果的访问权限。*NICE 2023，2023*年*4*月*11*日至*14*日，圣安东尼奥，德克萨斯州，美国© 2023 版权所有归所有者/作者所有。

1 引言

神经计算领域见证了社区可用的软件框架、网络仿真器、硬件平台和工程工具的显著发展与扩展 [1, 3, 5, 7, 10, 18, 20, 21]。尤其是在软件领域，随着高性能计算（HPC）的进步以及图形处理器（图形处理器）等加速器的使用，人们也在持续推动开展更大规模的实验 [8, 14, 16, 23]。然而，伴随框架和工具不断扩展而来的问题是：它们在共享生态系统中的兼容性和互操作性如何。例如，一个常见的讨论主题是如何在不同仿真器和目标神经形态平台之间进行充分比较和基准测试 [6, 15, 22]。

近期的 Fugu 等工作尝试仅采用支持最广泛的神经元和突触模型，解决脉冲神经算法与不同神经形态后端之间的互操作性问题 [1]。相邻的机器学习（ML）领域也有类似的工作，例如开放神经网络交换（开放神经网络交换），它提供了一种开放格式，用于表示 ML 模型（例如 PyTorch 与 TensorFlow 之间的模型）[2]。与本文最相关的工作或许是用于大规模网络模型的可扩展开放网络架构模板（SONATA）数据格式 [4]。尽管它采用了常见的文件格式（如 csv、hdf5、json），但我们注意到，它采用基于层级和群体的分组方式（对神经元和突触均如此），尤其是在数据局部性方面存在一些问题。

本文提出采用一种相对简单直接的数据格式，以便在仿真器与神经形态硬件平台之间实现脉冲神经网络模型的互操作和共享。该格式以广泛使用的压缩稀疏行（CSR）格式为基础，可高效表示稀疏矩阵，并对其进行扩展，以容纳网络状态的并行分区。我们首先在第 2 节对该格式作高层次介绍，第 3 节提供实现的链接，最后在第 4 节进行讨论。

2 概述

压缩稀疏行是一种广泛使用的格式，用于高效表示稀疏矩阵 [19]。尽管存在多种实现方式，但其核心思想是存储矩阵的非零值，并使用表示行和列的索引数组。对于一个 $(n \times n)$ 矩阵，其中包含 m 个非零项，行数组的大小为 $n+1$，其中每个条目提供该行非零列索引数量的前缀或累积和，而行数组中的最后一个条目为 m。列数组的大小为 m，包含给定行列的非零项索引

对。相应的值数组大小也为 m，只需按行优先顺序串联读取得到的非零值即可。

作为对此的扩展，分布式 CSR（dCSR）格式最初主要是在并行图划分算法 [12] 的相关工作中提出，用于存储和读入大型图。该格式额外提供一个大小为 $k+1$ 的索引数组，用于给出行的 k路划分中非零索引数量的前缀和或累积和，其中 $n_1 + n_2 + \cdots + n_k = n$。此外，原始 CSR 值数组和列数组沿这些划分拆分为多个数组，并且 $m_1 + m_2 + \cdots + m_k = m$。

相比之下，对于脉冲神经网络，表示连接关系和整体网络结构的一种自然方式是将其视为有向图 [24]。顶点 $V(G) = \{v_1, v_2、\ldots\ldots、v_n\}$ 对应脉冲神经元，而边 $E(G) = \{e_1, e_2、\ldots\ldots、e_m\}$ 对应突触连接。边 e_l 的方向：$v_i \rightarrow v_j$，对于 $i, j \in 1、\ldots\ldots、n$ 和 $l \in 1、\ldots\ldots、m$，对应事件（例如脉冲）从突触前神经元 v_i 向突触后神经元 v_j 的传播。我们还可以通过拆分顶点来进一步划分此图，使得 $|V_1| + |V_2| + \cdots + |V_k| = |V|$，这称为 k 路划分。在这里，可以根据边的源顶点或目标顶点，将边分配给相应的分区。对于脉冲神经网络的通信和计算模式，通常突触权重会在目标神经元上施加电流，因此，将有向边与其目标顶点放在同一位置更为合理。

我们可以看到，稀疏矩阵的行和列与脉冲神经网络图的顶点和边之间本质上存在同构关系。主要区别在于，对于脉冲神经网络而言，顶点和边通常携带的附加信息要多得多（例如连接延迟、神经元和突触状态），而标准情况下值数组中的单个非零条目所能承载的信息可能远远不足。为了解决这一问题，我们建议进行扩展，允许将值元组与列数组相关联（同时也允许将值元组与行数组相关联）。由于对于任何给定顶点或边而言，必要的唯一状态数量取决于其特定的模型动力学，我们还可以引入一个额外的模型字典来提供元组大小。

3 序列化格式

从这一高层描述过渡到更具体的实现其实相当简单。事实上，我们只需扩展 ParMETIS [13] 所使用的 dCSR 格式即可。这样做的另一个好处是，我们可以直接与现有的图划分工具互操作。尽管将数据序列化为纯文本文件在磁盘上通常比在仿真过程中占用更多内存，但为了便于移植，我们仍选择这种方式。

与其将图邻接关系分别序列化为两个连续的行偏移量和列索引数组，ParMETIS 实现的一种快捷方式是：在给定的 .adjcy.k（邻接）文件中，以行号隐式索引每一行，然后将其列索引为以空格分隔的值。由于处理时必须从磁盘读入整个文件，因此可以在运行时计算行偏移量等数据结构的初始化值。不过，分区偏移量仍然是必需的，这些偏移量通过单独的 .dist（分布）文件提供。

作为对几何分区器的补充信息，我们还初始化一个 .coord.k（坐标）文件，其中记录笛卡尔坐标系内顶点的空间坐标（x，y，z）。当网络规模超过高级分区器所需的内存容量、可能需要退回到简单的基于体素的分区时，这一点尤其有用。

对于主网络的 .state.k（状态）文件，我们提供了一个以空格分隔的、基于字符串的模型标识符和状态数据元组列表。我们选择将顶点和边模型放在一起，从而生成一个以顶点标识符及其状态开头的文件，后面则是每个传入连接的边标识符及状态。这里需要注意的是，由于用于图划分的邻接文件通常是无向的，而不是有向的，因此我们还为这样的边加入了特殊的“none”模型标识符；当存在传出连接但不存在传入连接时，该标识符不关联任何状态。如前所述，我们还引入了一个 .model 文件，用于提供基于字符串的模型标识符与其状态元组大小之间的映射，以及共享模型参数。

还有 .event.k（事件）文件，用于序列化由于连接延迟而尚未在目标顶点上处理的任何“进行中”仿真事件。这些文件包括由空格分隔的元组，依次为源顶点、到达时间、事件类型以及任何相关数据。

这里，我们指出了异步皮层流仿真工具（STACS）仿真器中的这种分布式文件格式实现 [23]。顺带一提，STACS 从零开始构建，基于 Charm++ 并行编程框架进行并行仿真，该框架本质上强制要求对并行对象之间的消息进行打包和解包时进行序列化，以支持容错计算 [11]。实际上，这使得网络生成过程与仿真过程能够通过中间的序列化表示解耦。它还提供了一种高效格式，可用于保存网络状态快照，以便进行检查点保存与重启以及离线分析。

作为一个比较可扩展性的示例，我们构建并序列化了由约 76K个神经元和 0.3B个突触组成的皮层微回路模型 [17]，，在磁盘上约占 12GB（与分区数量无关）。对于神经元数量扩大 2 倍的情况，即154K 个神经元和1.2B 个突触，我们的结果是在磁盘上约占 49GB。这实际上是关于突触数量的线性成本。

4 讨论

扩展dCSR格式尤其具有吸引力的一点在于，它借鉴了现有且广泛使用的格式，这些格式直观易懂。在我们的实现中，所有网络状态基本上都被序列化为四种主要类型的并行文件（邻接、坐标、状态和事件），此外，通过引入少量额外的元数据文件（dist、模型），存储成本的开销也很小。由于其简洁性，与流行的图分析软件包（如NetworkX）及其有向图数据结构进行互操作也变得相对简单 [9]。

除了简单易用之外，我们还认为，基于分区分发网络状态的方式，使 dCSR 能够更直接地适用于

计算并行性，无论目标平台是用于仿真的不同节点，还是神经形态硬件上的不同芯片。此外，由于这种序列化源自图划分器，它也可以直接用于指导脉冲神经网络模型的潜在重新划分，使其能够最佳地适配不同的后端。基于这些原因，我们提出扩展 dCSR 数据格式，供神经计算领域采用。

致谢

桑迪亚国家实验室是一家多任务实验室，由桑迪亚国家技术与工程解决方案有限责任公司管理和运营。该公司是霍尼韦尔国际公司的全资子公司，依据合同 DE-NA0003525 为美国能源部国家核安全管理局提供服务。

本文介绍了客观的技术成果和分析。文中可能表达的任何主观看法或意见，并不一定代表美国能源部或美国政府的观点。

参考文献

[1] James B. Aimone、William Severa 和 Craig M. Vineyard。2019 年。使用 Fugu 组合神经算法。载于国际神经形态系统会议论文集（美国田纳西州诺克斯维尔）（*ICONS '19*）。美国计算机协会，纽约州纽约，美国，文章 3，8 页。https://doi.org/ 10 1145 3354265 3354268 。/ 。Junjie Bai、Fang Lu、Ke Zhang 等。2019 年。开放神经网络交换。https://github.com/onnx/onnx。[3] Romain Brette、Michelle Rudolph、Ted Carnevale、Michael Hines、David Beeman、James M Bower、Markus Diesmann、Abigail Morrison、Philip H Goodman、Frederick C Jr Harris、Milind Zirpe、Thomas Natschläger、Dejan Pecevski、Bard Ermentrout、Mikael Djurfeldt、Anders Lansner、Olivier Rochel、Thierry Vieville、Eilif Muller、Andrew P Davison、Sami El Boustani 和 Alain Destexhe。2007 年。脉冲神经元网络的仿真：工具与策略综述。《计算神经科学杂志》23，3（2007 年 12 月），349–398。[4] Kael Dai 等。2020 年。用于高效描述大规模网络模型的 SONATA 数据格式。《公共科学图书馆·计算生物学》16，2（2020），e1007696。[5] Kael Dai、Sergey L. Gratiy、Yazan N. Billeh、Richard Xu、Binghuang Cai、Nicholas Cain、Atle E. Rimehaug、Alexander J. Stasik、Gaute T. Einevoll、Stefan Mihalas、Christof Koch 和 Anton Arkhipov。2020 年。脑建模工具包：用于脑回路多尺度建模的开源软件套件。《公共科学图书馆·计算生物学》16（2020）。[6] Mike Davies。2019 年。神经形态计算进展基准。《自然·机器智能》1，9（2019），386–388。[7] Andrew Davison、Daniel Brüderle、Jochen Eppler、Jens Kremkow、Eilif Muller、Dejan Pecevski、Laurent Perrinet 和 Pierre Yger。2009 年。PyNN：神经元网络模拟器的通用接口。《神经信息学前沿》2（2009）。https://doi.org/10.3389/neuro.11.011.2008[8] Bruno Golosio、Gianmarco Tiddia、Chiara De Luca、Elena Pastorelli、Francesco Simula 和 Pier Stanislao Paolucci。2021 年。使用图形处理器快速仿真高度连接的脉冲皮层模型。《计算神经科学前沿》15（2021）。[9] Aric A. Hagberg、Daniel A. Schult 和 Pieter J. Swart。2008 年。使用 NetworkX 探索网络结构、动态特性和功能。载于第七届 *Python* 科学会议论文集，Gaël Varoquaux、Travis Vaught 和 Jarrod Millman（编）。美国加利福尼亚州帕萨迪纳，11–15。[10] 英特尔。2022 年。Lava：用于神经形态计算的软件框架。GitHub 代码库。https://github.com/lava-nc/lava [11] Laxmikant V. Kale 和 Sanjeev Krishnan。1993 年。CHARM++：基于 C++ 的可移植并发面向对象系统。载于第八届面向对象编程系统、语言与应用年度会议论文集（美国华盛顿特区）（*OOPSLA '93*）。美国计算机协会，纽约州纽约，美国，91 108 10 1145 165854 165874 –。https://doi.org/ 。/ 。George Karypis 和 Vipin Kumar。。用于多级图划分和稀疏矩阵排序的并行算法。《并行与分布式计算杂志》48, 1 1998 71 95 10 1006 1997 1403 （），–。https://doi.org/ 。/ jpdc。。George Karypis 和 Kirk Schloegel。。*ParMETIS*：并行图划分与稀疏矩阵排序库（第 4.0 版）。明尼苏达大学计算机科学与工程系，美国明尼苏达州明尼阿波利斯。[14] James C. Knight 和 Thomas Nowotny。2021 年。借助程序化连接实现更大规模的图形处理器加速脑模拟。*Nature Computational Science* 1，2（2021 136 142 10 1038 43588 020 00022 7），–。https://doi.org/ 。/ s---S. Kulkarni、M. Parsa、J. P. Mitchell 和 C. D. Schuman。。神经形态模拟器与脉冲神经模拟器性能基准测试。*Neurocomputing* 447（2021），145–160。[16] Christian Pehle 和 Jens Egholm Pedersen。2021 年。Norse——用于脉冲神经网络的深度学习库。https://doi.org/10.5281/zenodo.4422025 文档：https://norse.ai/docs/。[17] Tobias C Potjans 和 Markus Diesmann。2014 年。细胞类型特异性的皮层微回路：在全规模脉冲网络模型中关联结构与活动。*Cereb Cortex* 24，3（2014 年 3 月），785–806。[18] Fredrick Rothganger、Christina Warrender、Derek Trumbo 和 James Aimone。2014 年。N2A：用于从神经元到算法进行建模的计算工具。*Frontiers in Neural Circuits* 8 2014 10 3389 2014 00001 （）。https://doi.org/ 。/fncir。。Yousef Saad。。大型特征值问题的数值方法。工业与应用数学学会。https://doi.org/10.1137/1.9781611970739 arXiv:https://epubs.siam.org/doi/pdf/ 10 1137 1 9781611970739 。/ 。Catherine D. Schuman、Thomas E. Potok、Robert M. Patton、J. Douglas Birdwell、Mark E. Dean、Garrett S. Rose 和 James S. Plank。2017 年。神经形态计算与硬件神经网络综述。*ArXiv* abs/1705.06963（2017）。[21] Marcel Stimberg、Romain Brette 和 Dan FM Goodman。2019 年。Brian 2：直观且高效的神经模拟器。*eLife* 8（2019 年 8 月），e47314。https://doi.org/ 10 7554 47314 。/eLife。Craig M. Vineyard、Sam Green、William M. Severa 和 Çetin Kaya Koç。2019 年。事件驱动神经形态架构的基准测试。载于国际神经形态系统会议论文集（美国田纳西州诺克斯维尔）。美国计算机协会，纽约州纽约，美国，5 页。[23] Felix Wang。2015 年。异步皮层流仿真工具（STACS）：与脉冲神经网络的接口。载于*Procedia Computer Science*（美国加利福尼亚州圣何塞），第 61 卷。322–7。[24] Douglas B. West。2001 年。图论导论（第 2 版）。Prentice Hall。