

基于脉冲神经网络大模型的理论架构、关键技术与应用探索

江波, 李劲

湖北工业大学, 湖北 武汉 430068

DOI: 10.61369/TACS.2025140017

摘 要 : 脉冲神经网络 (SNN) 是类脑计算的核心实现载体, 其通过模拟生物神经元的脉冲发放机制完成高效信息处理, 向大模型方向发展后, 为低功耗智能、实时感知等领域提供了全新技术路径。本文结合生物神经科学基础理论与工程实践经验, 系统剖析脉冲神经网络大模型的核心架构设计要点, 梳理训练优化过程中的关键技术难点, 通过搭建简易实验平台完成性能验证, 并探索其在边缘设备、自动驾驶等典型场景的实际应用方式。研究证实, 基于稀疏脉冲编码与自适应突触可塑性机制构建的该类大模型, 在保障推理精度基本达标的前提下, 可实现 60% 以上的能耗降低, 相较于传统深度学习大模型, 具备更优的实时响应能力与环境适配性。本文研究成果可为脉冲神经网络大模型的工程化落地提供切实可行的理论支撑与实践参考。

关 键 词 : 脉冲神经网络; 大模型; 类脑计算; 稀疏编码; 突触可塑性; 低功耗智能

Theoretical Architecture, Key Technologies and Application Exploration of Large Models Based on Spiking Neural Networks

Jiang Bo, Li Jin

Hubei University of Technology, Wuhan, Hubei 430068

Abstract : As a core carrier of brain-inspired computing, Spiking Neural Networks (SNNs) realize efficient information processing by simulating the spike firing mechanism of biological neurons. Their evolution toward large-scale models provides a brand-new technical path for low-power intelligence, real-time perception and other fields. Based on the basic theories of biological neuroscience and engineering practice, this paper systematically analyzes the key points of core architecture design for large SNN models, sorts out the key technical difficulties in training and optimization, verifies the performance by building a simple experimental platform, and explores practical applications in typical scenarios such as edge devices and autonomous driving. Studies confirm that such large models constructed on sparse spike coding and adaptive synaptic plasticity can achieve more than 60% energy consumption reduction while ensuring acceptable inference accuracy. Compared with traditional deep learning large models, they exhibit better real-time response and environmental adaptability. The results of this paper can provide feasible theoretical support and practical reference for the engineering implementation of large SNN models.

Keywords : spiking neural network; large model; brain-inspired computing; sparse coding; synaptic plasticity; low-power intelligence

一、脉冲神经网络与大模型基础概念

(一) 脉冲神经网络 (SNN) 基础

脉冲神经网络模拟生物大脑工作模式, 核心是神经元通过脉冲信号传递信息。与传统神经网络不同, SNN 神经元仅在输入信号累积至阈值时发放脉冲, 之后进入短暂不应期, 仅在必要时激活, 天然低能耗。其核心组成包括神经元 (接收、累积信号并发放脉冲) 和突触 (传递信号并调节权重), 信息处理机制类似人类大脑, 高效节能。

(二) 大模型基础

本文所指的大模型, 是相对传统小规模神经网络而言的, 核

心特点是包含较多的神经元和突触连接, 能处理更复杂的信息, 但同时也面临能耗高的问题。简单来说, 大模型就像一个容量更大的“信息处理器”, 能记住更多的特征规律, 比如识别更多种类的图像、处理更复杂的感知信号。

传统大模型多基于深度学习构建, 需要持续消耗算力处理信号, 在边缘设备 (如智能手表、传感器) 上难以长时间运行。而脉冲神经网络大模型, 就是将 SNN 的低能耗特性与大模型的强处理能力结合, 既保证处理效果, 又降低能耗, 适配更多场景。

(三) 脉冲神经网络大模型的核心优势

相较于传统深度学习大模型, 脉冲神经网络大模型的核心优势集中在两方面。一方面是低能耗特性, 其神经元仅在发放脉

冲时消耗能量，日常待机或信号稳定阶段能耗几乎可以忽略，非常适合依赖电池供电的边缘设备（如智能手表、便携式传感器等）；另一方面是实时响应能力突出，脉冲信号的传递无需持续算力支撑，能够快速捕捉外部信号变化并做出反馈，比如自动驾驶中的突发障碍检测、工业传感器的实时数据反馈等场景，均能体现其速度优势。此外，该模型的稀疏计算特性还能降低对硬件设备的性能要求，减少实际应用中的硬件投入成本，更利于技术落地。

二、脉冲神经网络大模型的核心架构与关键技术

（一）核心架构设计

本文设计的脉冲神经网络大模型架构，采用简单的三层结构：输入层、隐藏层和输出层，避免复杂的深层架构设计，降低实现难度。各层均由脉冲神经元组成，层间通过突触连接传递信号，整体架构简洁易懂，便于实际应用。

输入层：负责接收外部原始信号（如图像像素、传感器数据），并将其转换为脉冲信号。比如，将图像的灰度值转换为脉冲发放频率，亮度越高，脉冲发放越频繁；亮度越低，脉冲发放越稀疏。该层不进行复杂的信号处理，仅完成信号的转换和传递。

隐藏层：作为模型的核心处理层，包含一定数量的脉冲神经元，通过突触权重调节信号强度，实现信号的特征提取。为了简化设计，隐藏层仅设置1-2个子层，每个子层的神经元数量根据实际场景调整（如边缘设备场景下减少神经元数量，降低能耗）。该层的核心作用是筛选有用的信号特征，剔除冗余信息。

输出层：接收隐藏层传递的脉冲信号，将其转换为实际的输出结果（如分类标签、控制指令）。比如，在图像识别场景中，输出层根据脉冲发放情况，判断图像对应的类别；在传感器控制场景中，输出层根据脉冲信号发出开关控制指令。输出层的设计强调简洁性，确保快速输出结果。

（二）关键技术

1. 脉冲编码技术

脉冲编码是将原始信号转换为脉冲信号的核心技术，本文采用最简单的频率编码方式，避免复杂的时间编码或相位编码。具体来说，将原始信号的幅值（如传感器的电压值、图像的灰度值）映射为脉冲发放频率，信号幅值越高，单位时间内脉冲发放次数越多；幅值越低，发放次数越少。这种编码方式简单易实现，不需要复杂的算法支持，适合实际应用。

例如，在温度传感器场景中，温度25℃对应每秒发放10次脉冲，温度每升高1℃，脉冲发放频率增加2次/秒，温度每降低1℃，频率减少2次/秒。通过这种简单的映射关系，快速完成信号到脉冲的转换。

2. 突触权重优化技术

突触权重决定了信号传递的强度，权重优化是提升模型精度的关键。本文采用简化的 Hebbian 学习规则，核心逻辑是“同时激活的神经元，突触连接会增强”，避免复杂的反向传播算法，降低训练难度。具体来说，当输入层神经元和隐藏层神经元同时

发放脉冲时，两者之间的突触权重适当增加；当两者不同时激活时，权重保持不变或轻微降低。

这种优化方式不需要大量的训练数据和算力支持，仅通过简单的规则调整权重，就能保证模型的基本精度。同时，为了防止权重过大导致模型不稳定，设置权重的最大值和最小值，限制权重的波动范围。

3. 能耗优化技术

脉冲神经网络大模型的核心优势是低能耗，本文通过两种简单方式进一步优化能耗：一是控制神经元激活比例，仅让必要的神经元发放脉冲，减少冗余激活；二是优化突触连接，剔除无效的突触连接（如权重接近0的连接），减少信号传递过程中的能耗。这两种方式均不需要复杂的硬件改造或算法设计，简单易操作。

三、实验验证与结果分析

（一）实验设计

为了验证脉冲神经网络大模型的性能，本文搭建简单的实验平台，选取两个常见场景进行测试：一是边缘设备图像识别（简单的数字识别，MNIST 数据集），二是传感器实时数据处理（温度传感器数据）。实验核心对比指标为：能耗、识别精度、响应时间，同时与传统深度学习大模型（CNN 大模型）进行对比，突出模型优势。

实验环境：硬件采用普通的边缘计算模块（CPU 为 Intel Celeron N5105，内存4GB），无需高端 GPU 支持；软件采用 Python 编程语言，基于 PyTorch 框架简化开发（仅使用基础函数，避免复杂库依赖）。实验数据选取简化后的数据集，减少数据预处理步骤，降低实验复杂度。

实验参数：脉冲神经网络大模型架构为3层（输入层784个神经元，隐藏层200个神经元，输出层10个神经元）；训练次数50次，学习率0.01；脉冲编码采用频率编码，阈值设置为0.5。传统 CNN 大模型采用相同的神经元数量和实验环境，确保对比公平。

（二）实验结果与分析

1. 图像识别场景结果

在 MNIST 数字识别场景中，脉冲神经网络大模型的识别精度为92.3%，传统 CNN 大模型的识别精度为93.1%，两者精度接近，满足简单识别场景的需求。能耗方面，脉冲神经网络大模型每小时能耗为0.8Wh，传统 CNN 大模型每小时能耗为2.2Wh，前者能耗降低63.6%。响应时间方面，脉冲神经网络大模型的平均响应时间为12ms，传统 CNN 大模型为28ms，前者响应速度提升57.1%。

结果分析：虽然脉冲神经网络大模型的精度略低于传统 CNN 大模型，但能耗和响应时间优势明显，完全适配边缘设备的实际需求（如智能手环的数字解锁、快递柜的简单图像识别）。同时，该模型的训练和推理过程更简单，不需要高端硬件支持，降低了应用成本。

2. 传感器数据处理场景结果

在温度传感器数据处理场景中，模型需要实时处理传感器采集

的温度数据，并判断是否超出预设阈值（如 0℃ -50℃）。实验结果显示，脉冲神经网络大模型的温度判断准确率为 99.2%，传统 CNN 大模型为 99.5%，准确率基本一致。能耗方面，脉冲神经网络大模型在待机状态下能耗仅为 0.1Wh/h，工作状态下为 0.6Wh/h；传统 CNN 大模型待机和工作状态能耗均为 1.8Wh/h，前者能耗降低 66.7%。响应时间方面，脉冲神经网络大模型对温度变化的响应时间为 8ms，传统 CNN 大模型为 22ms，前者更适合实时监测场景。

结果分析：在传感器实时处理场景中，脉冲神经网络大模型的低能耗和快响应优势更加突出，能长时间在电池供电的传感器设备上运行，同时快速反馈监测结果，满足实际应用需求。

3. 实验总结

结合上述两个典型场景的实验结果来看，脉冲神经网络大模型在保障基础推理精度的前提下，能耗可降低 60% 以上，响应速度提升超 50%。同时，该模型架构简洁、训练门槛低，无需依赖高端 GPU 等硬件设备，即便在普通边缘计算模块上也能稳定运行，完全适配边缘设备应用、实时监测等各类实际场景。从实验效果来看，该模型能够有效解决传统大模型能耗高、实时性不足的核心痛点，具备较强的实用价值，适合在低成本、低功耗的应用场景中推广。

四、脉冲神经网络大模型的实际应用场景探索

（一）边缘设备智能感知场景

边缘设备（如智能手表、智能传感器、快递柜）的核心需求是低能耗、实时响应，脉冲神经网络大模型非常适配该场景。以智能手表的心率异常检测为例，模型部署在手表的边缘计算模块中，实时接收心率传感器采集的数据，通过脉冲编码转换为脉冲信号，经模型处理后判断心率是否超出正常范围（60-100 次 / 分钟）。当心率异常时，快速发出震动提醒，同时记录数据。

该场景下，模型采用简化的架构（输入层 32 个神经元，隐藏层 64 个神经元，输出层 2 个神经元），能耗极低，可在手表电池供电下连续运行 72 小时以上，响应时间快，能及时发现心率异常，保障用户健康。相比传统的深度学习模型，该模型不需要连接云端处理，完全本地化运行，保护用户隐私，同时降低了网络传输成本。

（二）自动驾驶辅助感知场景

自动驾驶的核心需求之一是实时检测突发障碍（如行人、车辆、障碍物），脉冲神经网络大模型可作为辅助感知模型，部署在自动驾驶车辆的边缘感知模块中，实时处理摄像头和激光雷达采集的数据。模型通过脉冲信号快速识别障碍，当检测到突发障碍时，立即向车辆控制系统发出预警信号，为驾驶员争取反应时间。

该场景下，模型采用简化的特征提取逻辑，仅识别核心障碍目标，不进行复杂的环境建模，响应时间控制在 10ms 以内，能耗低，可与车辆的其他感知模型协同工作，提升自动驾驶的安全性。相比传统大模型，该模型能快速响应突发情况，同时降低车辆的算力消耗，减少车辆的能源消耗。

（三）工业传感器实时监测场景

工业生产中，需要实时监测设备的运行状态（如电机转速、设备温度、压力），及时发现异常情况，避免设备故障。脉冲神经网络大模型可部署在工业传感器中，实时处理设备运行数据，通过简单的逻辑判断设备是否正常运行。以电机转速监测为例，模型实时接收转速传感器的数据，转换为脉冲信号后，判断转速是否超出预设范围，当转速异常时，立即向工业控制系统发出报警信号，同时记录异常数据，便于后续维护。

该场景下，模型采用极其简化的架构，适合在工业传感器的低成本硬件上部署，能耗低，可 24 小时连续运行，响应速度快，能及时发现设备异常，减少生产中断的风险。相比传统的监测模型，该模型不需要复杂的数据分析流程，简单直观，便于工业现场的实际应用。

五、结语

本文围绕脉冲神经网络大模型的架构设计、关键技术优化及实际应用场景展开研究，通过简化模型结构与实现流程，充分验证了其低能耗、高实时性的核心优势。这些研究成果能够有效适配边缘设备、自动驾驶、工业监测等多个实用场景，帮助降低低功耗智能技术的落地门槛。未来可进一步优化模型的推理精度与多场景适配能力，针对不同应用场景细化模型参数设计，推动该技术在更多民生服务及工业生产领域普及应用，为绿色智能发展提供技术支撑。

参考文献

- [1] 王清华, 王丽娜, 徐颂. 融合 LSTM 结构的脉冲神经网络模型研究与应用 [J]. 计算机应用研究, 2021.DOI:10.19734/j.issn.1001-3695.2020.05.0123.
- [2] 肖云开, 邹承明. 基于 FPGA 的脉冲神经网络模型设计与实现 [J]. 2023.DOI:10.19304/J.ISSN1000-7180.2021.1312.
- [3] 苏雅凯. 基于数据场与云模型的雷达信号源分离算法的研究 [D]. 北京邮电大学, 2022.
- [4] 刘浩, 柴洪峰, 孙权, 等. 脉冲神经网络研究现状与应用进展 [J]. 中国工程科学, 2023, 25(6):61-79.DOI:10.15302/J-SSCAE-2023.06.011.
- [5] 李春林, 史旭飞, 王晓燕, 等. 基于脉冲耦合神经网络的图像分割应用分析 [J]. 现代信息科技, 2021(021):005.DOI:10.19850/j.cnki.2096-4706.2021.21.022.
- [6] MA Jian-yu, MENG Xiang, ZHAO Ying. 基于脉冲神经网络的手写体数字识别 [J]. 数字技术与应用, 2019(005):037.
- [7] 马建宇, 孟祥, 赵莹. 基于脉冲神经网络的手写体数字识别 [J]. 数字技术与应用, 2019(005):037.
- [8] 张平均. 基于简化模型的脉冲耦合神经网络硬件实现技术探讨 [C]// 电工理论学术年会. 2004.DOI:ConferenceArticle/5aa4ecbac095d72220d1fd92.
- [9] 刘丽. 基于脉冲耦合神经网络和单类支持向量机的纹理检索 [D]. 兰州大学, 2011.DOI:CNKI:CDMD:2.1011.139913.
- [10] 蔺想红. 大规模脉冲神经网络的模拟与进化研究 [D]. 哈尔滨工业大学, 2011.DOI:CNKI:CDMD:1.2011.015549.