基于速率编码的极低延迟深度脉冲神经网络研究

熊志民[1], 陈云华[1†], 冯 忍[1,2], 陈平华[1]

(1. 广东工业大学 计算机学院, 广东 广州 510006; 2. 苏州酷珈科技有限公司, 江苏 苏州 215000)

摘要: 脉冲神经网络(SNN)具有强大的时空信息表征、异步事件处理能力, 但由于脉冲发放过程不具有连续可微性, 其训练是一个难题. 人工神经网络(ANN)转SNN的方法, 能够获得较高推理精度的深度SNN, 但却存在SNN网络延迟和功耗过高的问题. 为了降低网络延迟和功耗, 本文从脉冲信息传递的异步特性入手, 分析了极低延迟下SNN精度损失的主要原因, 提出残余膜电位误差(RMPE)的概念, 并对其进行分析与推导, 建立残余膜电位与初始膜电位和权重之间的关系模型. 基于所建立的残余膜电位模型, 提出一种初始膜电位和权重的分层校准算法, 减少残余膜电位误差, 从而解决脉冲输入序列均匀分布假设与真实分布不一致的问题. 提出一种ANN-SNN的双阶段转化框架, 在第1阶段, 采用带有可训练分层阈值的量化截断激活函数对ANN进行二次训练, 以实现量化误差与截断误差的最优化; 在第2阶段, 对SNN进行微调训练, 以进一步缩小残余膜电位误差, 使得在极低延迟下的ANN-SNN转化也能获得较高的精度. 实验结果表明, 本文方法在推理延迟和功耗方面都优于现有的方法.

关键词: 脉冲神经网络; ANN-SNN转化; 速率编码

引用格式: 熊志民, 陈云华, 冯忍, 等. 基于速率编码的极低延迟深度脉冲神经网络研究. 控制理论与应用, 2025, 42(3): 531 – 540

DOI: 10.7641/CTA.2024.30012

Ultra-low-latency deep spiking neural network based on rate coding

XIONG Zhi-min[1], CHEN Yun-hua[1†], FENG Ren[1,2], CHEN Ping-hua[1]

(1. School of Computer Science and Technology, Guangdong University of Technology, Guangzhou Guangdong 510006, China;
2. Suzhou Kujia Technology Co., Ltd, Suzhou Jiangsu 215000, China)

Abstract: Spiking neural network (SNN) possesses robust capabilities for spatiotemporal information representation and asynchronous event processing. However, training SNN is challenging due to the non-differentiable nature of the spiking process. Converting artificial neural network (ANN) to SNN can yield deep SNN with high inference accuracy, but this approach often results in increased latency and power consumption in SNN. To mitigate network latency and power consumption, we have analyzed the primary reasons behind the loss of SNN accuracy at ultra-low latency, focusing on the asynchronous transfer characteristics of spikes. We introduce the concept of residual membrane potential error (RMPE) to address these issues. We have analyzed and derived the relationship between residual membrane potential and both the initial membrane potential and weights. Based on this understanding, we propose a layer-by-layer calibration algorithm for adjusting initial membrane potential and weights, aiming to reduce residual membrane potential error. This approach resolves the discrepancy between the assumption of a uniform distribution of spike input trains and the actual distribution. We propose a two-stage conversion framework for ANN-to-SNN conversion. In the first stage, we employ a quantization clipping activation function with a trainable stratification threshold. This allows us to train the ANN twice, optimizing both quantization error and clipping error. In the second stage, we further fine-tune the SNN to reduce residual membrane potential errors at ultra-low latency. Experimental results demonstrate that our proposed method outperforms existing methods in terms of inference latency and power consumption.

Key words: spiking neural network; ANN-SNN conversion; rate coding

Citation: XIONG Zhimin, CHEN Yunhua, FENG Ren, et al. Ultra-low-latency deep spiking neural network based on rate coding. *Control Theory & Applications*, 2025, 42(3): 531 – 540

收稿日期: 2023−01−11; 录用日期: 2024−04−11.

†通信作者. E-mail: yhchen@gdut.edu.cn.

本文责任编委: 陈谋.

广东省自然科学基金项目(2021A1515012233, 2016A030313713)资助.

Supported by the Natural Science Foundation of Guangdong Province (2021A1515012233, 2016A030313713).

1 引言

脉冲神经网络 (spiking neural network, SNN) 包含具有时序动力学特性的神经元节点、稳态–可塑性平衡的突触结构、功能特异性的网络环路, 高度借鉴了生物启发的局部非监督、全局弱监督的生物优化方法, 具有强大的时空信息表征、异步事件处理能力. 其被部署在具有神经形态芯片的硬件上时, 将具有深度人工神经网络 (artificial neural network, ANN) 无法企及的低功耗、低延迟和强大的计算力[1–2].

尽管SNN具有上述优点, 但由于脉冲神经元的脉冲发放过程不具有连续可微性, 深度学习中的梯度下降算法不能直接应用于SNN的训练, 因此, SNN的训练仍然是一个难题. 针对这一问题, 有研究者提出了使用时间依赖可塑性(spike-timing-dependent-plasticity, STDP)[3–4]和替代梯度(surrogate gradient, SG)[5–6]方法来训练SNN, 但这两种方法训练得到的SNN的精度较低, 且对梯度消失和梯度爆炸现象更为敏感, 因此很难实现对深层 SNN 的训练. 为了获得更高精度的SNN, Cao等人[7]提出将训练好的深度ANN转化为对应深度SNN的方法, 目前这种基于ANN-SNN转化获得深度SNN的方法已成为构建高精度SNN的首选方法. ANN-SNN转化的基本原理是: 首先, 使用监督学习训练ANN, 然后, 将训练好的ANN参数迁移到具有相同网络结构的SNN. 由于SNN的脉冲神经元与ANN的模拟神经元在信息表示和传递机制方面存在巨大差异, 由训练好的ANN转化得到的SNN与源ANN之间存在近似误差. 因此, 后续研究主要集中在近似误差的消除等方面. 这些方法, 按照神经元信息编码方式的不同可分为两类: 基于时间编码的方法[8–9]和基于速率编码的方法. 其中, 速率编码[10–13]利用神经元的脉冲发放率来表征和识别特征, 相对于时间编码, 具有编码简单和鲁棒性强等优点, 故在ANN-SNN转化中被广泛使用. 这类方法主要通过优化SNN的权重阈值比降低近似误差, 或对ANN-SNN转化误差建模, 找到与之有关的SNN参数, 从而进行针对性的参数优化来降低近似误差.

调整SNN参数以降低近似误差的方法, 主要针对SNN的网络权值、脉冲发放阈值、初始膜电位等参数进行调整, 以使SNN的输出尽可能接近ANN的输出. Diehl等人[11]将ANN中每一层的最大激活值作为归一化尺度因子, 提出权值归一化方法, 优化了SNN的权重阈值比, 从而降低了SNN神经元发放率饱和所产生的近似误差. 但因为异常激活值的存在, 深层次网络神经元的脉冲发放率极低, 使得大量的脉冲神经元处于失活状态. 为此, Rueckauer等人[12]分析了16666个CIFAR-10样本, 发现99.9%的修正线性单元 (rectified linear unit, ReLU) 激活值远远小于最大激活值, 基于此, 提出了99.9百分位阈值方法, 以解决异常值所导致的深层网络脉冲神经元失活的问题. Ho和Ding等人[14–15]在ANN中添加可训练的裁剪层, 利用尺度因子缩放激活值, 然后将缩放后的激活值映射到SNN的阈值, 从而更好平衡网络的精度与延迟性能. 上述方法虽然不同程度上降低了SNN与ANN之间的近似误差, 但由于这些方法仅考虑了激活值的截断误差, 往往需要设置很长的推理时间窗口 (大于1000个时间步), 才能达到与ANN相接近的分类精度, 而这将导致SNN丧失其低延迟优势.

为此, 研究者通过对ANN-SNN转化过程进行深入的研究, 对近似误差进行建模, 以获得与近似误差有关的参数, 从而对相关参数进行针对性地优化. 研究者[16–18]推导出了可分层优化的转化损失函数, 发现量化误差和截断误差与初始膜电位的偏移量有关, 基于SNN输入符合均匀分布这一假设, 通过分层分解近似误差调整初始膜电位的最佳偏移量, 在一定程度上减少了SNN的推理延迟. Bu等人[19]推导估计了SNN的激活函数, 将转化误差分为截断误差、量化误差和不均匀误差. 该方法首先假设剩余膜电位$v^l(T)$在$[0, V_{\rm th})$, 将不均匀误差归为量化误差. 然后, 假设脉冲神经元输入的脉冲序列符合均匀分布, 提出以量化截断移位激活函数替换ReLU激活函数, 在ANN中训练SNN各层阈值的方法. 最后, 在转化得到的SNN中, 将初始膜电位统一设置为$\dfrac{1}{2}V_{\rm th}^l$. 上述方法虽然一定程度上降低了SNN的推理延迟, 但有研究[20]表明: 该类方法所提出的SNN输入符合均匀分布这一假设过于简单, 甚至是错误的, 导致其在极低延迟时(如, $T=4$), 往往会有较大的精度损失. 另一方面, 由脉冲信息传递的异步性可知, 脉冲神经元在时间步T结束之后的残余膜电位$v^l(T)$的值可能为负数, 或者超过阈值, 因此, 将其值限制在$[0, V_{\rm th})$是不合理的. 再者, 直接采用量化截断移位激活函数替换 ReLU 函数进行 ANN 训练[19], 会导致训练得到的ANN存在较大的精度波动.

针对上述问题, 本文分析了极低延迟下 SNN 精度损失的主要原因, 提出残余膜电位误差(residual membrane potential error, RMPE)的概念, 从脉冲神经元输入输出发放率之间的关系, 分析残余膜电位误差与初始膜电位和权重之间的关系, 通过分层校准算法校准初始膜电位和权重来减少残余膜电位误差, 从而解决SNN输入均匀分布假设与真实分布不一致的问题. 此外, 针对直接使用量化截断移位激活函数替代ReLU激活函数, 进行ANN训练导致其精度波动变大的问题, 本文提出一种双阶段转化框架: 第1阶段, 首先基于ReLU激活函数预训练ANN, 然后, 再采用可训练分层阈值的量化激活函数替代ReLU进行ANN的微调, 得到与源ANN精度相当的量化–截断ANN(quanti-

zation-clipping ANN, QC-ANN), 作为预训练SNN. 第2阶段, 通过对预训练SNN的初始膜电位和权重进行分层校准以减少膜电位残余信息误差, 得到最终的SNN. 本文主要贡献总结如下: 1) 分析了极低延迟下SNN精度损失的主要原因, 提出残余膜电位误差(RMPE)的概念, 并对其进行分析与推导, 建立残余膜电位与初始膜电位和权重之间的关系模型; 2) 提出一种初始膜电位和权重的分层校准算法, 减少残余膜电位误差, 从而解决脉冲输入序列均匀分布假设与真实分布不一致的问题; 3) 提出一种ANN-SNN的双阶段转化框架, 实现量化误差与截断误差的最优化; 进一步缩小残余膜电位误差, 使得在极低延迟下的ANN-SNN转化也能获得较高的精度.

2　IF神经元脉冲响应特性及转化误差分析

2.1　脉冲神经元模型及ANN-SNN转化原理

本文选择集成–发放(integrate-and-fire, IF)模型作为SNN神经元模型. 为了避免信息丢失, 本文使用减法复位(reset-by-subtraction)[12]机制. 一旦膜电位$v^l(t)$超过预先设置的阈值$V_{\rm th}^l$, 神经元将发放脉冲并更新膜电位$v^l(t)$. 因此, 膜电位更新的数学形式为

$$v^l(t) = v^{l-1}(t) + W^l x^{l-1} - s^l(t)V_{\rm th}^l, \tag{1}$$

其中: $x^{l-1}(t)$表示$l-1$层突触前神经元在t时刻的加权突触后膜电位; W^l为突触权值; $s^l(t)$表示l层所有神经元在t时刻的输出脉冲, 其值为1表示发放脉冲, 为0则表示不发放脉冲, 脉冲发放函数是Heaviside阶跃函数(Heaviside step function).

基于速率编码的ANN-SNN转化, 其基本思想是假设SNN在一段时间T内的脉冲发放率可以近似等于ANN的激活值, 如式(2)所示:

$$a_i^l \approx r_i^l = \frac{1}{T}\sum_{t=0}^{T} s_i^l(t). \tag{2}$$

2.2　IF神经元脉冲响应特性

本文从SNN的脉冲响应特性出发, 根据脉冲神经元接受输入的增量来显式表示输出值, 假设 ANN 与SNN在网络l层的输入是相同的, 即

$$a^{l-1} = \frac{\sum_{t=1}^{T} s^{l-1}(t)}{T} V_{\rm th}^{l-1},$$

为了方便, 本文记

$$\phi^{l-1}(T) = \frac{\sum_{t=1}^{T} s^{l-1}(t)}{T} V_{\rm th}^{l-1},$$

用来表示SNN的脉冲发放率与阈值的乘积. 假设$v^l(0)=0$, 根据速率编码特性, 每个时间步最多发放一个脉冲, 即$\sum_{t=1}^{T} s^l(t) \in \{0,1,2,\cdots,T\}$, 根据式(2), 理想情况下, 脉冲个数等于膜电位增量向下取整, 如$\sum_{t=1}^{T} s^l(t) = \lfloor \frac{W^l\phi^{l-1}(T)T}{V_{\rm th}^l} \rfloor$, 因此, 可以得到

$$\phi^l(T) = {\rm clip}(\frac{V_{\rm th}^l}{T}\lfloor\frac{W^l\phi^{l-1}(T)T}{V_{\rm th}^l}\rfloor, 0, V_{\rm th}^l) \approx V_{\rm th}^l {\rm clip}(\frac{1}{T}\lfloor\frac{W^l\phi^{l-1}(T)T}{V_{\rm th}^l}\rfloor, 0, 1), \tag{3}$$

$$ {\rm clip}(x,a,b) = \begin{cases} a, & x \leqslant 0, \\ b, & x \geqslant 0, \\ x, & \text{其他}, \end{cases} \tag{4}$$

其中$\lfloor\cdot\rfloor$表示floor函数. 如式(3)所示, SNN的IF神经元脉冲响应特性可以表示为一个阶梯函数(如图1(a)中的IF阶梯曲线).

2.3　ANN-SNN转化误差分析

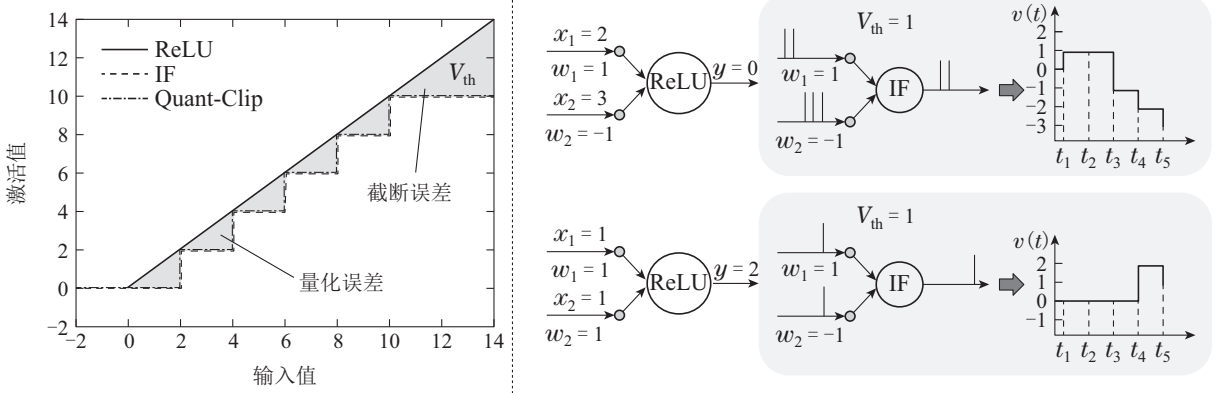
ANN的前向传播方程如图1(a)蓝色曲线所示. 由于SNN的输出是离散值, 并且是floor的形式, 而ANN的输出是连续值, 因此将ANN转化为SNN后导致不可避免的信息损失. 研究表明, ANN激活值与SNN阈值之间存在一个映射关系, 当$\phi^l(T) = \frac{\sum_{t=1}^{T} s^l(t)}{T} V_{\rm th}^l$, SNN输出的范围为$[0, V_{\rm th}^l]$, 然而, ANN输出$a^l$属于范围$[0, a_{\max}^l]$. 如果$V_{\rm th}^l \leqslant a_{\max}^l$, 如图1和式(3)所示, 存在信息表示不对等的误差, 本文称其为截断误差(clipping error, CE). 由于输出脉冲$\sum_{t=1}^{T} s^l(t) \in \{0,1,2,\cdots,T\}$是离散的, 因此式(3)本质上是个量化函数, 量化因子为$\frac{V_{\rm th}^l}{T}$. 如图1所示, 当将激活值a^l量化到$\phi^l(T)$时, 不可避免会产生量化误差(quantization error, QE). 从式(3)可知, 量化误差和截断误差与阈值有关, 且两种误差之间存在一种此消彼长关系, 即当阈值单调增加时, 会减少截断误差, 但会导致量化误差增大. 当阈值单调减少时, 会减少量化误差, 但会导致截断误差增大. 因此, 现有研究[14–15]主要集中在如何调整阈值, 早期的方法, 如Max Norm和Robust Norm[11–12]等, 将ANN中最大的激活值或最大激活值的99.9%位值对应于SNN的阈值, 采用固定阈值的方式, 来实现ANN-SNN的转换. 较近的方法, 如可训练阶段层(trainable clipping layers, TCL)[14]等方法, 在ANN训练的ReLU层增加一层截断操作, 设置一个可训练参数作为最大激活值的缩放因子, 缩放后的最大激活值对应于SNN的阈值, 采用可训练阈值来实现ANN-SNN的转换, 目的是最小化转换误差, 包括截断误差和量化误差, 以使得二者之和最小.

2.4　残余膜电位误差分析与推导

现有研究均未考虑脉冲发放的异步性与ANN数据同步性之间存在差异的问题, 该问题将导致极短时间步内发放的脉冲数与ANN的激活值存在不一致性.

ANN神经元间传递的是实数值, 而SNN神经元间传递的是脉冲序列. ANN神经元的信息处理没有时间概念, 而SNN神经元的信息处理是以时间步为单位, 逐个时间步进行脉冲处理. 脉冲的时序性使得相同频率可能会产生不同顺序的脉冲序列, 从而其所对应的SNN输出也不同. 即在频率编码模式下, 相同的实数值可能被编码为不同的脉冲序列, 其所对应的输出值和原始的ANN实数值之间很可能存在误差. 为此, 本文从脉冲信息传递的异步和时序特性入手, 分析极低延迟下SNN精度损失的主要原因, 以之为基础, 提出残余膜电位误差(RMPE)的概念, 并对其进行分析与推导.

脉冲发放具有时序性, 在较少的时间步内, 极易出现脉冲数和激活值不一致的情况, 会产生更多或更少的误差脉冲数, 这是导致在极低延迟下精度损失的主要原因. 如图1(b)所示, 在第1行的例子中, 假设输入到ANN中的数据$x_1=2$和$x_2=3$, 对应的权重为$w_1=1$和$w_2=-1$, ANN神经元输出的激活值为0. 在第2行的例子中, 假设输入到ANN中的数据为$x_1=1$和$x_2=1$, 对应的权重为$w_1=1$和$w_2=1$, ANN神经元输出的激活值为2. 根据速率编码的特性, 将SNN发放脉冲个数与ANN激活值相对应, 并以某种分布来表示, 本文将时间步和阈值分别设置为$T=5$, $V_{\rm th}=1$.



(a) ANN激活值与SNN输出脉冲($V_{\rm th}^l=10$, $T=5$)之间的转化误差

(b) 残余膜电位信息丢失导致的误差示意图

图 1 ANN-SNN转化误差示意图

Fig. 1 Diagram ofANN-SNN conversion error

图1(b)第1行的例子中, 时间步t_1, t_2达到阈值之后, 即刻发放脉冲, 之后未达到阈值不发放脉冲, 在时间步T累积结束时, 总共发放两个脉冲, 与SNN期望得到的脉冲个数相比, 多发放了两个脉冲. 图1(b)第2行的例子中, 在有限的时间步t_5中, 膜电位包含的信息足以发放两个脉冲, 但实际上只能发放一个脉冲, 与SNN期望得到的脉冲个数相比少发放了一个脉冲. 对于基于速率编码的SNN来说, 上述情况在极少的时间步下是频繁存在的, 其所产生的误差可以在较长的时间步中, 通过各个时间步剩余膜电位的叠加来进行缓解, 反映在整个时间窗口T上, 就是残余膜电位, 因此, 本文称这种误差为残余膜电位误差RMPE, 以期望通过调节与残余膜电位有关的参数来进一步降低误差.

下面, 本文从脉冲神经元输入输出发放率之间的关系, 分析与残余膜电位误差RMPE有关的参数. 本节假设膜电位残余信息误差产生的误差脉冲数为M和发放阈值$V_{\rm th}=1$, 将式(1)从1累加到T, 然后除以T, 得到

$$r^l(T)=\frac{W^l\sum\limits_{t=1}^{T}s^{l-1}(t)}{T}\pm\frac{W^lM}{T}-\frac{v^l(T)-v^l(0)}{T},\tag{5}$$

$$v^l(T)=\sum_{t=1}^{T}W^ls^{l-1}(t)+v^l(0)\pm W^lM-\\ \text{clip}(\lfloor\sum_{t=1}^{T}W^ls^{l-1}(t)\rfloor,0,T),\tag{6}$$

式(5)描述了脉冲神经元输入输出发放率之间的关系, 与ANN前向过程相似. 当模拟时间步T足够长时, $\dfrac{v^l(T)-v^l(0)}{TV_{\rm th}^l}\approx 0$和$\dfrac{(W^lM)}{T}\approx 0$, 然而, 较高的$T$会导致较大的推理延迟. 从式(6)中可以看出, 由残余膜电位$v^l(T)$产生的残余膜电位误差RMPE与W^l, $v^l(0)$有关, 实际上通过调整初始膜电位$v^l(0)$能够调整SNN输入分布, 因此合适的初始膜电位可以使SNN输入分布尽可能满足均匀分布, 并且通过调整初始膜电位$v^l(0)$和权重W^l, 可以减少产生的误差脉冲数M, 并且将残余膜电位$v^l(T)$限制在$[0,V_{\rm th}^l)$. 后续章节会通过调整这些参数校准转化误差.

3　双阶段训练框架与算法

3.1　本文算法训练框架

基于第3章的分析, 为解决对ANN激活值直接进行量化截断导致激活值表示精度不足的问题, 本文提出一种双阶段转化方案, 如图2所示. 第1阶段, 首先, 基于ReLU训练源ANN, 然后, 基于带有可训练阈值的量化截断激活函数训练QC-ANN, 实现量化与截断误差的最小化. 第2阶段, 首先, 将QC-ANN的神经元替换为IF神经元, 实现ANN-SNN的转化, 此时, 需要将网络中BN(batch normalization)层的参数整合到前层中构成预训练SNN, 如式(7)所示:

$$W \leftarrow W\frac{\gamma}{\sigma},\ b \leftarrow \beta + (b-\mu)\frac{\gamma}{\sigma}, \tag{7}$$

其中: W, b为前层的权重和偏置参数; γ, σ, β, μ均为BN层的参数, 且γ, β为需要训练的超参数. 然后, 通过对SNN的初始膜电位$v_i^l(0)$和权重W_i^l进行逐层校准, 以减少残余膜电位误差, 得到最终的SNN.

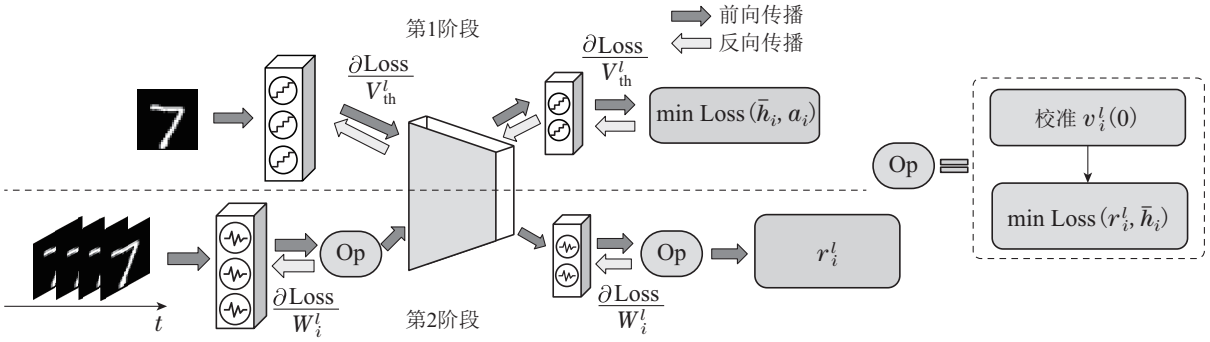


图 2　双阶段训练算法框架

Fig. 2　The framework of the two-stage training algorithm

3.2　一阶段训练

3.2.1　量化截断激活函数

依据第3.3.1节IF神经元脉冲响应特性的分析结果(式(3)), 对ReLU激活函数进行量化, 得到量化截断激活函数$\bar{h}$如下:

$$a^l = \bar{h}(z^l) = \lambda^l \mathrm{clip}(\frac{1}{L}\lfloor\frac{z^l L}{\lambda^l}\rfloor, 0, 1), \tag{8}$$

其中超参数L表示ANN的量化步长, 可训练参数λ^l对应于SNN第l层网络的发放阈值. 采用该激活函数训练得到的 ANN, 本文称之为 QC-ANN, 训练好的 QC-ANN可转为具有相同权值的SNN.

3.2.2　QC-ANN的训练

本文基于梯度下降法训练 QC-ANN, 损失函数Loss为QC-ANN输出层与训练集标签的交叉熵, 即

$$\mathrm{Loss} = -\sum_i y_i \log(a^i), \tag{9}$$

其中: y为训练集的one-hot标签, a^i为QC-ANN输出层第i个神经元输出的激活值. 为了直接训练量化–截断ANN, 使用直通估计器(straight-through estimator)[21]对floor函数进行估计, 即

$$\frac{\partial\lfloor x\rfloor}{\partial x} = 1. \tag{10}$$

3.2.3　IF神经元与QC-ANN神经元之间的误差

假设$\phi^{l-1} = z^{l-1}$, $T = L$, $V_{\mathrm{th}}^l = \lambda^l$, 根据式(3)和式(8), 可得到IF神经元的输出与QC-ANN神经元的输出之间的误差如下:

$$\begin{aligned}\mathrm{Err}^l = \phi^l(T) - a^l = \\ V_{\mathrm{th}}^l \mathrm{clip}(\frac{1}{T}\lfloor\frac{\phi^{l-1}(T)T}{V_{\mathrm{th}}^l}\rfloor, 0, 1) - \\ \lambda^l \mathrm{clip}(\frac{1}{L}\lfloor\frac{z^l L}{\lambda^{l-1}}, 0, 1\rfloor) = 0,\end{aligned} \tag{11}$$

在训练QC-ANN时, 对SNN的各层阈值进行训练, 可以使量化截断激活函数的曲线与IF脉冲神经元特性响应曲线尽可能地接近, 当二者重叠时, 量化误差和截断误之和为最优.

3.3　残余膜电位对应的SNN参数学习与优化

在第1阶段 (图2虚线上部) QC-ANN训练结束后, 将QC-ANN训练的权重和阈值移植到SNN中, 得到不纠正残余膜电位误差RMPE的预训练SNN. 在第2阶段, 如图2(虚线下部)所示, 采用分层校准算法对初始膜电位 $v^l(0)$ (layer-wise membrane potential calibration, LMPC) 进行逐层校准, 并对权重 W^l (layer-wise weights calibration, LWC)进行微调学习, 以减少RMPE.

3.3.1　初始膜电位逐层校准

初始膜电位的设置, 不仅可以用于将残余膜电位控制在$[0, V_{\mathrm{th}}^l)$范围之内, 还可以用于调整脉冲序列的分布, 使其尽可能接近理想的均匀分布. 为了拟合量化截断ANN的激活值$\bar{h}(x)$, 考虑使用N个样本来校准初始膜电位, 在不考虑误差脉冲数的前提下, 依据式(6)可以推导出SNN理想激活值$\bar{h}$与发放率之间的

误差函数, 如下:

$$\min_{v_i^l(0)}\{|\frac{\phi_i^{l-1}(T)}{V_{\rm th}^l}-\frac{v_i^l(T)-v_i^l(0)}{TV_{\rm th}^l}-\bar{h}(z_i^l)|\},\quad(12)$$

在理想状态下, 可以令式(12)等于0, 从而显式推导出初始膜电位$v_i^l(0)$的表示式, 如下:

$$v_i^l(0)=\frac{T}{N}\sum_{i-1}^{N}|\bar{h}(z_i^l)-\phi_i^l(T)|,\quad(13)$$

依据式(13), 可以利用SNN理想输出激活值与QC-ANN实际激活值之间的误差校准初始膜电位. 至此, 即可利用脉冲神经元膜电位在整个时间窗T内(理想的SNN输出激活值)的增量来校准初始膜电位, 控制残余膜电位信息的范围和调整输入脉冲序列的分布, 减少残余膜电位误差RMPE.

3.3.2 权重微调学习

为了更精确的纠正残余膜电位误差RMPE, 本文利用目标激活值a^l与真实脉冲发放率r^l的最小均方误差, 来校准不同时间步T下的权值. 这里采用与Li和Deng等人[16–17]类似的分层优化方法进行权重的微调学习, 即

$$\arg\min_{W^l}{\rm Loss}=\arg\min_{W^l}\{(r^l-\bar{h}^l(x))^2\},\quad(14)$$

真实脉冲发放率r^l需要在时间步T上进行累积, 由于脉冲发放函数是不可微的, 为了解决脉冲函数不可微问题, 本文采用代理梯度$f(v)$[22]替代$\dfrac{\partial H(v)}{\partial v}$和脉冲表示[23]进行时间反向传播(back-propagation through time, BPTT)算法, 即

$$\frac{\partial H(v)}{\partial v}\approx f(v)=\gamma\max\{0,1-|v|\},\quad(15)$$

其中γ是一个常数, 表示梯度的最大值, 然后用BPTT得到的权值的梯度为

$$\frac{\partial{\rm Loss}}{\partial W_{i,j}^l}=\frac{1}{T}\frac{\partial{\rm Loss}}{\partial r_i^l}\sum_{t=1}^{T}\sum_{t'}^{t}\frac{\partial s_i^l(t)}{\partial v_i^l(t')},\quad(16)$$

$$\frac{\partial s_i^l(t)}{\partial v_i^l(t')}=\begin{cases}f(v_i^l(t')),\ s_i^l(t)=1,\\0,\qquad\quad 其他,\end{cases}\quad(17)$$

为了使用较低的成本进行微调, 采用小批量样本(mini-batch)mini-batch=1024, 在SNN层面, 通过利用随机梯度下降算法微调(fine-tuning)权重来减少残余膜电位误差(RMPE). 基于上述校准算法, 可以实现在极少时间步($T=4$)下的高精度ANN-SNN转化.

4 实验

4.1 数据集及相关实验设置

本文训练VGG-16和ResNet-20[24]作为基准ANN模型, 再将其转化为SNN, 并在CIFAR-10和CIFAR-100数据集上验证本文所提出的方法. 本文使用动量参数为0.9的随机梯度下降优化器, 余弦衰减调度器来调整学习速率. 第1阶段: ANN的两次训练: 基于ReLU训练的源ANN的学习率设置为0.1; QC-ANN的学习率设置为1e-5; 第2阶段: 初始膜电位校准和基于BPTT的SNN的微调训练阶段, 在CIFAR-10, CIFAR-100数据集上初始学习率分别设置为1e-4和1e-5. 对于第1层的输入, 本文采用文献[16]中的方法, 将归一化后的像素值直接编码为脉冲序列. 本文经过不同的L对应的QC-ANN, Phase I和Phase I4+Phase II不同阶段训练进行研究, 发现当量化步长$L=4$时, QC-ANN的最终精度较低. 当$L>4$时, 训练得到的QC-ANN精度趋于稳定. 极低延迟下SNN的精度随着L的增加而降低, 但是过小的L会限制QC-ANN及其转换得到的SNN的最高精度. 不同的L对应不同的QC-ANN的精度、量化误差和截断误差, 会影响Phase II校准算法提升精度的幅度. 因此, 本文选取了一个折中的值, 即将L设置为8. 本文实验基于Pytorch深度学习框架实现, 硬件平台: CPU: Intel Xeon E5-2698 v4 2.2 GHz, GPU: Tesla V100.

4.2 消融实验

本文利用VGG-16和ResNet-20网络结构在CIFAR-10和CIFAR-100数据集上训练的ANN精度分别为95.71%, 95.23%和77.10%, 75.07%, 将第1阶段量化截断表示为Phase I, 第2阶段完整校准算法为Phase II, 根据实验结果, 在VGG-16和ResNet-20网络结构上, 基于本文方法将ANN转化为SNN后的分类精度分别为95.75%, 95.23%和77.10%, 76.38%. 在这里, 本文进行消融实验来验证Phase I的量化和Phase II校准方法的有效性. 在4种条件下, 分别在CIFAR-10和CIFAR-100数据集上, 从$T=2$到$T=64$, 测试VGG-16和ResNet-20网络结构在Phase I(量化)、逐层膜电位校准(layer-by-layer calibration of initial membrane potential, LMPC)、逐层权重微调(layer-by-layer weight calibration, LWC)、Phase I+Phase II下的精度.

为了验证第1阶段QC-ANN阈值训练的效果, 本文将阈值训练的结果与Max Norm和Robust Norm等阈值设置方法的结果进行了比较, 如表1所示. 从表1可以看出, 本文的阈值优化方法优于现有的常用阈值优化方法.

进一步验证第2阶段校准方法的有效性, 如图3所示, Phase I(量化)在$T\leqslant 8$取得的转化效果最差, 因为QC-ANN只简单的平衡了截断误差(CE)和量化误差(QE), 而忽略了在极低延迟下的残余膜电位误差导致的损失. 在Phase II中的两种不同的校准方法LMPC和LWC, 对于转化精度都有一定程度的提升, 说明在很好平衡了截断误差(CE)和量化误差(QE)的情况下, 减少了残余膜电位误差RMPE. 在VGG-16网络结构中, LWC校准方法在$T\leqslant 4$时, 得到的结果比LMPC

校准方法的精度更好. 在ResNet-20网络结构中, LWC校准和LMPC校准得到的精度相差不大, 说明通过校准初始膜电位也能达到LWC校准的精度. 图3Phase I+Phase II曲线显示, Phase I+Phase II在所有时间步中均能取得最佳的精度, 说明在平衡了截断误差(CE)和量化(QE)误差的前提下, 极大的减少了残余膜电位误差RMPE.

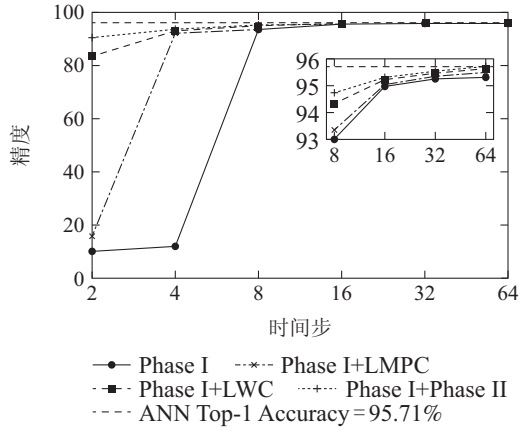
表 1 不同阈值设置方法在不同的模拟时间步长下得到的SNN Top-1精度比较

Table 1 Comparison of SNN Top-1 accuracy obtained by different threshold setting methods under different simulation time steps

阈值设置方法	VGG-16(95.71) on CIFAR-10				
	8	16	32	64	128
Max Norm [11]	10.0	10.11	31.83	91.30	94.53
Robust Norm [12]	14.90	66.67	93.70	95.08	95.31
QC-ANN	**93.04**	**94.09**	**95.24**	**95.30**	**95.35**

根据已有的研究[16, 19], 如果在时间窗口T结束时, 残余膜电位$v^l(T) \in [0, V_{\rm th}^l)$, 表示在时间窗口$T$内没有信息残留. 为了更显式的说明校准RMPE的效果, 本文统计在时间窗口T结束时, 每一层神经元残余膜电位$v^l(T) \notin [0, V_{\rm th}^l)$的神经元数量, 如图4–5所示, VGG-16在CIFAR数据集上, 通过不同校准算法之后得到的SNN具有RMPE神经元的比例在Phase I的基础上有所减少, 说明了第2阶段校准算法的有效性.

本文在CIFAR-100数据集上验证逐层校准初始膜电位与已有校准方法的对比, 如表2所示, 本文的初始膜电位校准方法与已有的方法在不同时间步下的精度比较. 与统一设置初始膜电位为$\frac{1}{2}V_{\rm th}$的OpI[18]方法和每个时间步分配$\frac{1}{2T}V_{\rm th}$偏移量Opt[17]方法相比, 使用本文提出的LMPC膜电位校准方法在低延迟下得到的精度更高, 例如, 在转化VGG-16网络结构时, 本文的初始膜电位校准方法只需要16个时间步就可以获得74.48%的精度.



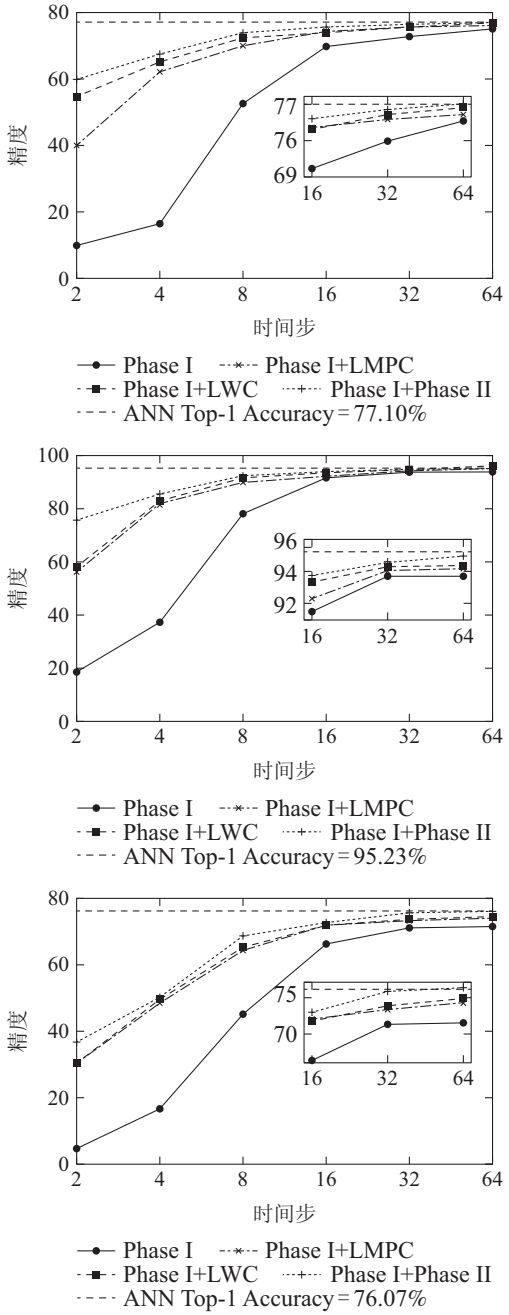


图 3 基于CIFAR-10和CIFAR-100数据集, 在不同时间步下的SNN消融实验精度曲线

Fig. 3 Based on CIFAR-10 and CIFAR-100 datasets, experimental accuracy curves of SNN ablation at different time steps were obtained

4.3 与其它研究工作的对比

在本节中, 为了进一步说明本文工作的有效性, 这里使用VGG-16和ResNet-20网络结构在CIFAR-10和CIFAR-100数据集上进行实验, 并将Top-1的精确度与一些最佳模型进行比较. 如表3所示, 与RateNorm[15]调整阈值的缩放因子相比, RateNorm[15]减少前文分析的截断误差而忽略量化误差, 因为截断误差与量化误差之间存在平衡关系, 仅仅减少截断误差会导致量化误差的增加. 本文在ANN中进行可训练阈值,

实现了截断误差和量化误差的最优平衡. 膜电位阈值优化(optimized potential threshold, Opt)[17]、初始膜电位优比(optimized potential initialization, Opl)[18]假设SNN输入满足均匀分布逐层分析误差, 将初始膜电位设置为理论最优值, 但SNN输入分布会在不同数据集上有所差异, 导致固定设置的初始膜电位在深层次网络中不一定达到最优的结果, 本文利用SNN脉冲神经元实际输入的增量得到的发放率与源ANN激活值之间的差异进行逐层校准初始膜电位, 相较于固定设置的初始膜电位更加精确. 文献[16]利用SNN模拟输出发放率与源ANN激活值直接的误差进行校准, 因为模拟发放率需要在时间步非常长的时候才能达到与实际发放率相似的输出, 导致在极低延迟下的精度很低, 本文从脉冲神经元异步传输特性分析残余膜电位不一定满足$[0, V_{\rm th})$, 并且利用BPTT算法, 在极低延迟下利用实际发放率与源ANN激活值之间的均方误差来校准误差. 量化剪辑激活函数 (quantization clip floor-shift, QCFS)[19] 将集成–发放(IF)神经元脉冲响应特性得到的量化截断移位激活函数作为源ANN的激活函数, 在实验中发现, 对于量化截断移位激活函数训练得到的ANN精度波动较大, 针对这个问题, 本文提出在双阶段训练框架中对ANN进行二次训练算法, 并且QCFS[19]将残余膜电位限制为$[0, V_{\rm th})$, 从而忽略了由于脉冲神经元异步传输特性产生的残余膜电位信息误差, 导致在低延迟下精度有所损失, 本文从脉冲神经元输入输出发放率之间的关系分析出与残余膜电位信息误差有关的参数, 并且提出一套逐层校准算法. 在极低推理延迟$T = 4$时, 在CIFAR-10上训练的VGG-16和ResNet-20精度分别可达93.28%, 85.43%, 在CIFAR-100上训练的VGG-16和ResNet-20精度分别可达67.69%, 50.41%. 为了证明本文模型不需要过多的推理延迟($T \leqslant 128$), 表3还列出了不同时间步长的推理准确率, 并与其他工作进行了比较, 结果表明, 本文的方法优于以往的转化方法.

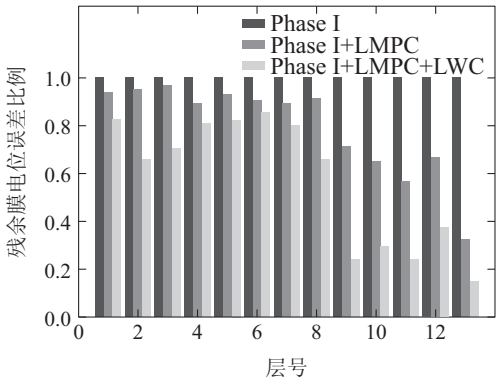


图 4 CIFAR-10上不同校准算法对应的Spiking VGG-16具有的RMPE的比例

Fig. 4 The ratio of RMPE in VGG-16 obtained by different calibration algorithms on CIFAR-10

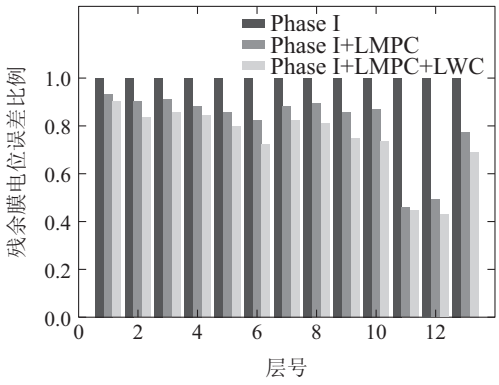


图 5 CIFAR-100上不同校准算法对应的Spiking VGG-16具有的RMPE的比例

Fig. 5 The ratio of RMPE in VGG-16 obtained by different calibration algorithms on CIFAR-100

表 2 不同初始膜电位校准方法, 在不同的模拟时间步下的SNN Top-1精度

Table 2 Accuracy under different simulation time steps

膜电位校准方法	VGG-16(77.10)		ResNet-20(76.07)	
	$T = 4$	$T = 16$	$T = 4$	$T = 16$
Phase I	16.23	69.95	16.56	66.34
Phase I+OpI[18]	42.24	70.67	19.04	67.32
Phase I+Opt[17]	58.19	71.50	24.34	67.16
Phase I+LMPC	**62.42**	**74.48**	**48.47**	**71.99**

4.4 能耗估计

本文对VGG-16网络结构在CIFAR-100数据集上的发放率进行了统计, 时间步$T = 64$时的平均发放率为0.0616, 这反映了脉冲活动的稀疏性. 本文采用已有工作[25]中的能量估计方程进行能耗估计. 本文SNN与ANN能耗之比为

$$\frac{\text{Energy}_{\text{SNN}}}{\text{Energy}_{\text{ANN}}} = \frac{c\alpha + (1 - \dfrac{c}{b})b\beta}{a\alpha}, \tag{18}$$

其中: α表示乘法的能量消耗4.6 pJ[26]; β表示加法的能量消耗0.9 pJ[26]; a, b, c分别表示ANN, SNN和SNN第1层的操作次数. 根据式(18), 计算出当$T = 64$时, 本文提出的模型计算能耗, 只需要ANN能耗的70.06%.

5 总结

针对基于速率编码的SNN推理延迟大、输入符合均匀分布这一简单假设、直接采用量化截断移位激活函数替换ReLU函数进行ANN训练精度波动大等问题, 本文提出了一种ANN-SNN双阶段转化框架与算法, 实现了超低延迟的深度SNN. 本文首先分析了极低延迟下SNN精度损失的主要原因, 提出残余膜电位误差RMPE的概念, 并对其进行分析与推导, 建立残余膜电位与初始膜电位和权重之间的关系模型. 然后,

基于所建立的残余膜电位模型, 提出一种初始膜电位和权重的分层校准算法, 减少残余膜电位误差, 从而解决脉冲输入序列均匀分布假设与真实分布不一致的问题. 最后, 提出一种ANN-SNN的双阶段转化框架, 以实现量化误差与截断误差的最优化, 使得在极低延迟下的ANN-SNN转化也能获得较高的精度. 实验结果表明, 本文方法所构建的SNN在分类精度、推理延迟性能指标上均获得了最好的结果.

表 3 与其他已有的SNN转化算法的比较.

Table 3 Comparison with other existing SNN conversion algorithms

方法	网络结果	ANN Accuracy/%	SNN Accuracy/%						
			$T=2$	$T=4$	$T=8$	$T=16$	$T=32$	$T=64$	$T=128$
CIFAR-10									
RateNorm[15]	VGG16	92.82	—	—	—	57.90	85.40	91.15	92.95
Opt.[17]	VGG16	92.09	—	—	—	—	76.24	90.64	95.73
Calibration[16]	VGG16	95.72	—	—	—	—	93.71	95.14	95.79
OpI.[18]	VGG16	94.57	—	—	90.96	93.38	94.20	94.45	94.55
QCFS*[19]	VGG16	95.52	83.93	91.77	94.45	95.22	95.56	95.74	95.79
Phase I	VGG16	95.71	10.00	11.77	93.04	94.98	95.24	95.30	95.35
Phase I+LWC	VGG16	95.71	83.20	92.93	94.33	95.22	95.47	95.61	95.65
Phase I+LMPC+LWC	VGG16	95.71	89.97	93.28	94.73	95.32	95.53	95.75	95.75
Calibration[16]	ResNet20	95.46	—	—	—	—	94.78	95.30	95.42
QCFS*[19]	ResNet20	93.34	58.67	75.70	87.79	92.14	93.04	93.34	93.24
Phase I	ResNet20	95.23	12.95	37.22	78.19	91.55	93.70	93.72	93.85
Phase I+LWC	ResNet20	95.23	58.01	82.82	91.43	93.32	94.30	94.33	94.35
Phase I+LMPC+LWC	ResNet20	95.23	75.46	85.43	91.98	93.76	94.62	94.99	95.23
CIFAR-100									
Opt.[17]	VGG16	77.89	—	—	—	—	56.16	62.93	77.71
Calibration[16]	VGG16	77.89	—	—	—	—	73.55	76.64	77.87
OpI.[18]	VGG16	76.31	—	—	60.49	70.72	74.84	75.97	76.31
QCFS*[19]	VGG16	76.28	52.46	62.09	70.71	74.83	76.41	76.73	76.74
Phase I	VGG16	77.10	10.00	16.23	52.60	69.85	72.96	75.24	75.95
Phase I+LWC	VGG16	77.10	54.84	65.00	72.63	74.29	75.93	76.83	76.92
Phase I+LMPC+LWC	VGG16	77.10	60.24	67.69	73.78	75.48	76.55	77.10	77.11
Calibration[16]	ResNet20	77.16	—	—	—	—	76.32	77.29	77.73
QCFS*[19]	ResNet20	69.69	19.96	34.14	55.37	67.33	69.82	70.49	70.55
Phase I	ResNet20	76.07	4.51	16.56	45.27	66.34	71.30	71.57	72.02
Phase I+LWC	ResNet20	76.07	30.40	49.78	65.24	71.84	73.78	74.75	75.32
Phase I+LMPC+LWC	ResNet20	76.07	36.65	50.41	68.88	72.95	75.84	76.38	76.36

*: 量化步长$L=8$.

参考文献:

[1] ZHANG Tielin, XU Bo. Research status and prospect of spiking neural network. *Control Theory and Technology*, 2021, 44(9): 1765 – 1785.
(张铁林, 徐波. 脉冲神经网络研究现状及展望. 计算机学报, 2021, 44(9): 1765 – 1785.)

[2] HUANG Tiejun, YU Zhaofei, LI Yuan, et al. Advances in spike vision. *Journal of Image and Graphics*, 2022, 27(6): 1823 – 1839.
(黄铁军, 余肇飞, 李源, 等. 脉冲视觉研究进展.中国图象图形学报, 2022, 27(6): 1823 – 1839.)

[3] AMIRHOSSEIN T, ZACHARY K, ANTHONY M. Training spiking convnets by stdp and gradient descent. *International Joint Conference on Neural Networks*. Riode Janeiro, Brazil: IEEE, 2018: 1 – 8.

[4] AMIRHOSSEIN T, ANTHONY M. BP-STDP: Approximating backpropagation using spike timing dependent plasticity. *Neurocomputing*, 2019, 330: 39 – 47.

[5] LEE J H, DELBRUCK T, PFEIFFER M. Training deep spiking neural networks using backpropagation. *Frontiers in Neuroscience*, 2016, 10: 508.

[6] WU Y J, DENG L, LI G Q, et al. Spatio-temporal backpropagation for training high-performance spiking neural networks. *Frontiers in Neuroscience*, 2018, 12: 331.

[7] CAO Y Q, CHEN Y, KHOSLA D. Spiking deep convolutional neural networks for energy-efficient object recognition. *International Journal of Computer Vision*, 2015, 113(1): 54 – 66.

[8] HAN B, ROY K. Deep spiking neural network: Energy efficiency through time based coding. *European Conference on Computer Vi-*

sion. Cham: Springer, 2020: 388 – 404.

[9] PARK S, KIM S, NA B, et al. T2FSNN: Deep spiking neural networks with time-to-first-spike coding. *ACM/IEEE Design Automation Conference (DAC).* San Francisco, CA, USA: IEEE, 2020: 1 – 6.

[10] CHEN Y H, MAI Y C, FENG R, et al. An adaptive threshold mechanism for accurate and efficient deep spiking convolutional neural networks. *Neurocomputing*, 2022, 469: 189 – 197.

[11] DIEHL P U, NEIL D, BINAS J, et al. Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing. *International Joint Conference on Neural Networks.* Killarney, Ireland: IEEE, 2015: 1 – 8.

[12] RUECKAUER B, LUNGU I, HU Y H. Conversion of continuous-valued deep networks to efficient event-driven networks for image classification. *Frontiers in Neuroscience*, 2017, 11: 682.

[13] MUELLER E, HANSJAKOB J, AUGE D, KNOLL A. Minimizing inference time: Optimization methods for converted deep spiking neural networks. *International Joint Conference on Neural Networks.* Shenzhen, China: IEEE, 2021: 1 – 8.

[14] HO N D, CHANG I J. TCL: An ann-to-snn conversion with trainable clipping layers. *ACM/IEEE Design Automation Conference*. San Francico, CA, USA: IEEE, 2021: 793 – 798.

[15] DING J H, YU Z F, TIAN Y H, et al. Optimal ANN-SNN conversion for fast and accurate inference in deep spiking neural networks. *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence.* Montreal: Morgan Kaufmann, 2021: 2328 – 2336.

[16] LI Y H, DENG S K, DONG X, et al. A free lunch from ANN: Towards efficient, accurate spiking neural networks calibration. *International Conference on Machine Learning.* Virtual: PMLR, 2021: 6316 – 6325.

[17] DENG S K, GU S. Optimal conversion of conventional artificial neural networks to spiking neural networks. *International Conference on Learning Representations.* Vienna, Austria: Open Review, 2021: 1 – 7.

[18] BU T, DING J H, YU Z F, et al. Optimized potential initialization for low-latency spiking neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence.* Palo Alto, California, USA: IEEE, 2022, 36: 11 – 20.

[19] BU T, FANG W, DING J H, et al. Optimal ANN-SNN conversion for high-accuracy and ultra-low-latency spiking neural networks. *International Conference on Learning Representations.* Virtual: Open Review, 2022: 1 – 9.

[20] DATTA G, BEEREL P A. Can deep neural networks be converted to ultra low-latency spiking neural networks? *Design, Automation & Test in Europe Conference & Exhibition*. Antwerp, Belgium: IEEE, 2022: 718 – 723.

[21] YOSHUA B, NICHOLAS L, AARON C. Estimating or propagating gradients through stochastic neurons for conditional computation. *ArXiv Preprint*, 2013, arxiv: 1308.3432.

[22] GUILLAUME B, DARJAN S, ANAND S, et al. Long short-term memory and Learning-to-learn in networks of spiking neurons. *ArXiv Preprinnt*, 2018, arxiv: 1803.09574.

[23] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition. *Conference on Computer Vision and Pattern Recognition.* San Francisco, CA, USA: IEEE, 2016: 770 – 778.

[24] HOROWITZ M. 1.1 Computing's energy problem (and what we can do about it). *International Solid-State Circuits Conference Digest of Technical Papers.* New Orleans, LA, USA: IEEE, 2014: 10 – 14.

[25] MENG Q Y, XIAO M Q, YAN S, et al. Training high-performance low-latency spiking neural networks by differentiation on spike representation. *Conference on Computer Vision and Pattern Recognition.* New Orleans, LA, USA: IEEE, 2022: 12444 – 12453.

[26] RATHI N, ROY K. DIET-SNN: A low-latency spiking neural network with direct input encoding and leakage and threshold optimization. *IEEE Transactions on Neural Networks and Learning Systems*, 2021: 1 – 9.

作者简介:

熊志民 硕士研究生, 主要研究方向为计算机视觉、神经形态计算, E-mail: 2112105296@mail2.gdut.edu.cn;

陈云华 博士, 副教授, 主要研究方向为深度学习、神经形态计算、计算机视觉, E-mail: yhchen@gdut.edu.cn;

冯 忍 硕士研究生, 主要研究方向为计算机视觉、神经形态计算, E-mail: 2112005223@main2.gdut.edu.cn;

陈平华 教授, 硕士, CCF广州委员, 主要研究方向为计算机系统结构、MIS、云计算、大数据、推荐系统, E-mail: phchen@gdut.edu.cn.