科学通报 2025 年 第 70 卷 第 13 期: 1920 ~ 1928
面向人工智能的物理神经网络专题 论 文
《中国科学》杂志社
SCIENCE CHINA PRESS

利用单个自旋电子神经元实现类批量归一化算法

许焯[1†], 辛玥[1†], 袁正平[1], 周康[1], 杨雨梦[1,2], 高盛华[1], 祝智峰[1,2*]

1. 上海科技大学信息科学与技术学院, 上海 201210
2. 上海高能效与智能定制芯片工程技术中心, 上海 201210
† 同等贡献
* 联系人, E-mail: zhuzhf@shanghaitech.edu.cn

2024-08-19 收稿, 2024-10-21 修回, 2024-10-28 接受, 2024-10-29 网络版发表
国家自然科学基金(12104301)和上海科技大学高性能计算共享服务平台和信息科学与技术学院公共算力与服务平台资助

摘要 近年来, 自旋电子器件已被广泛研究用于实现人工神经元. 传统方法通常使用磁隧道结的随机翻转来产生非线性激活函数(例如Sigmoid). 然而, 这种固定形状的激活函数在神经网络训练过程中限制了权重更新, 进而导致性能的下降. 本文提出利用自旋转矩驱动磁矩翻转, 可以实现激活函数在训练过程中的动态调整. 具体来说, 通过调节电控脉冲宽度和磁各向异性, 改变激活函数的斜率, 从而实现了反向传播算法中输出响应速度的调节. 这一机制类似于机器学习中的批量归一化算法. 研究表明, 该算法不仅可以通过软件实现, 还可以通过单个自旋电子器件在硬件层面实现. 最后, 使用这些可训练的自旋电子神经元, 展示了在不增加额外能耗的情况下, 将手写数字识别的准确率从88%提高到91.3%. 这一成果为自旋电子神经网络的硬件实现提供了新的视角.

关键词 自旋电子神经元, 自旋力矩, 非线性激活函数, 类脑计算

深度学习的快速发展对神经网络硬件提出了更高的性能和能耗要求. 传统CMOS实现的人工神经元和突触存在较大的硅面积和由于易失性导致的低能效等问题[1,2]. 自旋电子学凭借其内禀的非易失性和丰富的物理机制, 能够用简单的器件实现复杂功能, 使得基于自旋电子的神经元和突触能够完成更低能耗和更紧凑的设计[3~8]. 磁隧道结(magnetic tunnel junction, MTJ)近年来已被提出用于生成多种激活函数, 如Sigmoid[5,9~19]、ReLU[17,20]、线性函数[18]和阶跃函数[2]. 然而, 在这些工作中, 激活函数在神经网络训练过程中的形状是固定的, 这限制了权重更新的灵活性. 从直观上分析, 如果允许激活函数在训练过程中发生变化, 神经网络的性能将会得到很大的提高. 例如, 在训练初期, 可能需要输出快速变化以应对较大的误差, 而在训练后期, 则需要输出逐渐平滑以稳定结果. 在这种情况下, 激活函数应在开始时具有陡峭的斜率, 并逐渐变得平滑. 这一过程类似于机器学习中的批量归一化算法[21~23], 通过对输入数据进行偏移和重新归一化, 有效解决了输入分布偏离理想分布的问题, 从而提升了训练的稳定性并加速了收敛过程. 这一机制在神经网络中具有重要作用, 尤其在大规模训练任务中表现出色. 后文展示的结果将证明本文提出的可训练自旋电子神经元与批量归一化的理念类似(见附录S1), 具有动态调整激活函数斜率的能力, 通过在训练过程中灵活调整激活函数的特性, 增强了网络的学习能力和性能[24]. 这一工作不仅证明了自旋电子学在硬件实现关键神经网络算法中的潜力, 还展示了其在人工智能技术中的广阔应用前景.

基于这一设想, 本研究利用磁矩翻转的物理机制来实现具有可训练激活函数的自旋电子神经元. 在神经网络训练过程中, 电学可调性至关重要, 这使得激活函数的形状可以根据需求进行动态调整. 当使用自旋转移力矩(spin transfer torque, STT)翻转磁隧道结时,

引用格式: 许焯, 辛玥, 袁正平, 等. 利用单个自旋电子神经元实现类批量归一化算法. 科学通报, 2025, 70: 1920–1928
Xu Z, Xin Y, Yuan Z, et al. Implementing batch normalization-like algorithm with a single spintronic neuron (in Chinese). Chin Sci Bull, 2025, 70: 1920–1928, doi: 10.1360/TB-2024-0879

© 2024《中国科学》杂志社

我们发现曲线的斜率可以通过脉冲宽度、脉冲幅度或磁各向异性来调节, 这些都是电学调控的表现. 该磁隧道结翻转曲线的斜率作为神经网络中类似权重的参数, 可以通过梯度下降算法进行更新. 然而, 斜率和权重在物理实现层面是不同的. 磁隧道结翻转曲线斜率是一个局部参数, 其状态可以通过在神经元上施加电流来改变; 而权重表示不同神经元之间的连接, 是非局部参数, 通常通过交叉电阻阵列实现. 因此, 本工作中引入的激活函数斜率为神经网络增加了一个新的自由度, 在物理器件中高效实现激活函数的调控是以前的研究中未曾讨论过的. 由于激活函数的斜率可以在训练过程中变化, 我们称之为可训练激活函数[25,26]. 我们对这种自旋电子神经元的性能进行了评估. 以传统神经元作为参照, 我们搭建了神经网络, 发现其识别准确率为88%, 在引入可训练激活函数的神经元之后, 准确率提高到91.3%, 并且所需的学习迭代次数显著减少. 我们的方法在实际硬件实现方面展示了更高的效率和可行性, 显著推进了在物理层面调控激活函数的研究.

1 研究结果与讨论

本文研究的器件由一个位于铁电(ferroelectric, FE)层顶部的磁隧道结组成(图1(a)), 该层将用于控制自由层的各向异性[27]. 自由层材料为广泛使用的CoFeB, 通过施加电压(V_{MTJ})或电流脉冲(I_{c})来翻转磁矩. 磁矩动力学通过求解Landau-Lifshitz-Gilbert-Slonczewski (LLGS)方程得到[28~31], 其公式为$\mathrm{d}\boldsymbol{m}/\mathrm{d}t=-\gamma\boldsymbol{m}\times\boldsymbol{H}_{\mathrm{eff}}+\alpha\boldsymbol{m}\times\mathrm{d}\boldsymbol{m}/\mathrm{d}t-\gamma\hbar J_{\mathrm{c}}/(2et_{\mathrm{FL}}M_{\mathrm{s}})\boldsymbol{m}\times(\boldsymbol{m}\times\boldsymbol{\sigma}_{\mathrm{STT}})$, 包含有效场$\boldsymbol{H}_{\mathrm{eff}}$、阻尼常数$\alpha=0.0122$、磁旋比$\gamma=1.76\times10^{11}$ rad/(s T)、普朗克常数$\hbar=6.58\times10^{-16}$ eV s、电子电荷$e=1.6\times10^{-19}$ C、自由层厚度$t_{\mathrm{FL}}=1.3$ nm、饱和磁化强度$M_{\mathrm{s}}=1.58$ T[32,33]以及自旋极化$\boldsymbol{\sigma}_{\mathrm{STT}}$, 其方向由$V_{\mathrm{MTJ}}$的极性决定, $+V_{\mathrm{MTJ}}$使$\boldsymbol{\sigma}_{\mathrm{STT}}$沿$-\boldsymbol{z}$方向, $-V_{\mathrm{MTJ}}$则使$\boldsymbol{\sigma}_{\mathrm{STT}}$沿$+\boldsymbol{z}$方向. $\boldsymbol{H}_{\mathrm{eff}}$由晶体各向异性场$\boldsymbol{H}_{\mathrm{an}}$、退磁场$\boldsymbol{H}_{\mathrm{demag}}$、热噪声场$\boldsymbol{H}_{\mathrm{thermal}}$和外部磁场$\boldsymbol{H}_{\mathrm{ext}}$组成. $H_{\mathrm{an}}=2\left(K_{\mathrm{bulk}}+\frac{K_{\mathrm{i}}}{t_{\mathrm{FL}}}\right)m_z\hat{z}/M_{\mathrm{s}}$, 其中体各向异性常数$K_{\mathrm{bulk}}=2.245\times10^5$ J/m^3, 界面各向异性常数$K_{\mathrm{i}}=1.286\times10^{-3}$ J/m^2[32]. 我们验证了该K_{bulk}足以克服形状各向异性, 使磁矩在平衡状态下沿$\boldsymbol{z}$轴分布. 基于长度和宽度均为15 nm的矩形样本, 可以计算得到$\boldsymbol{H}_{\mathrm{demag}}=-(N_xm_x\hat{\boldsymbol{x}}+N_ym_y\hat{\mathrm{y}}+N_zm_z\hat{\boldsymbol{z}})$中的形状因子$N_{x,y,z}$. 热噪声场$\boldsymbol{H}_{\mathrm{thermal}}=N_1(0,u)\hat{\boldsymbol{x}}+N_2(0,u)\hat{\mathrm{y}}+N_3(0,u)\hat{\boldsymbol{z}}$是一个随机场, 其中$N_1(0,u)$, $N_2(0,u)$, $N_3(0,u)$是均值为零、标准差为$u=\sqrt{2k_{\mathrm{B}}T\alpha/(V_{\mathrm{FL}}M_{\mathrm{S}}\gamma(1+\alpha^2)\Delta t)}$的正态分布随机函数, 其中$k_{\mathrm{B}}=1.38\times10^{-23}$ J/K, $T=300$ K, V_{FL}, $\Delta t=5$ ps分别是玻尔兹曼常数、温度、自由层体积和仿真的时间步长. 由于热噪声的存在, 磁矩在I_{c}影响下以一定概率(P_{sw})实现翻转[34~40]. 我们对每个电流脉冲下的磁矩动力学进行100次模拟, 精确计算出这一翻转概率(见附录S2). 如图1(b)所示, 不同脉冲宽度下的翻转概率可以很好地用形式为$f=1/(1+\exp(-k(I_{\mathrm{c}}+c)))$的Sigmoid函数描述, 其中$k$和$c$分别表示曲线的斜率和位移(见附录S3). 因此, 通过控制脉冲宽度, 该磁隧道结可以作为电学可调神经元用于产生Sigmoid函数. 为了评估该自旋电子神经元的性能, 本文构建了一个三层神经网络对MNIST数据进行手写数字的分类[41], 该数据集包括50000个训练样本和10000个测试样本. 如图1(c)所示, 输入层有784个神经元, 对应每张图像中的像素数量. 隐藏层和输出层采用了Sigmoid函数, 其中隐藏层有25个神经元, 输出层有10个神经元. 我们对该神经网络进行了训练. 首先, 在前向传播过程中计算损失函数$J=-[y\log(a_3)+(1-y)\log(1-a_3)]$, 其中$a_3$是最后一层的输出, y是正确的输入值(例如图1(c)中的$y=8$). 接下来, 使用反向传播算法更新权重(w_1和w_2), 使得损失函数降低. 重复这两个步骤, 直到损失最小化. 如图1(d)中的实心三角形所示, 随着训练的进行, 损失函数单调降低, 证明该神经网络进行了正确的训练. 随后, 使用该神经网络对未见过的数据进行测试, 以评估网络的识别准确性. 在1000次迭代后, 识别准确率达到88%.

在使用Sigmoid拟合P_{sw}时, 我们发现需要引入两个额外参数, 即k和c. 在较大的脉冲宽度下, 所需的翻转电流减少, 这解释了曲线的位移(c). 同时, 当更多的电子注入自由层时, 自旋转矩增强, 从而实现更快的翻转和更陡的曲线(即更大的k). 因此可以直观地认为, 调整k和c可以优化神经网络的性能. 其思想类似于批量归一化算法, 输入参数会在每次迭代后进行位移和归一化, 进而减小内部协变量偏移. 而在我们提出的系统中, 输入电流的大小通过c实现位移, k进行缩放, 这与归一化过程类似. 与软件实现不同, k和c的控制可以在自旋电子器件中实现, 并且其大小可以通过改变脉冲宽度等方式进行电学调控, 因此它们可以作为神经网络中需要学习的参数, 在训练过程中进行动态更新. 需要注意的是, 本文提出的k和c与神经网络中的权重在

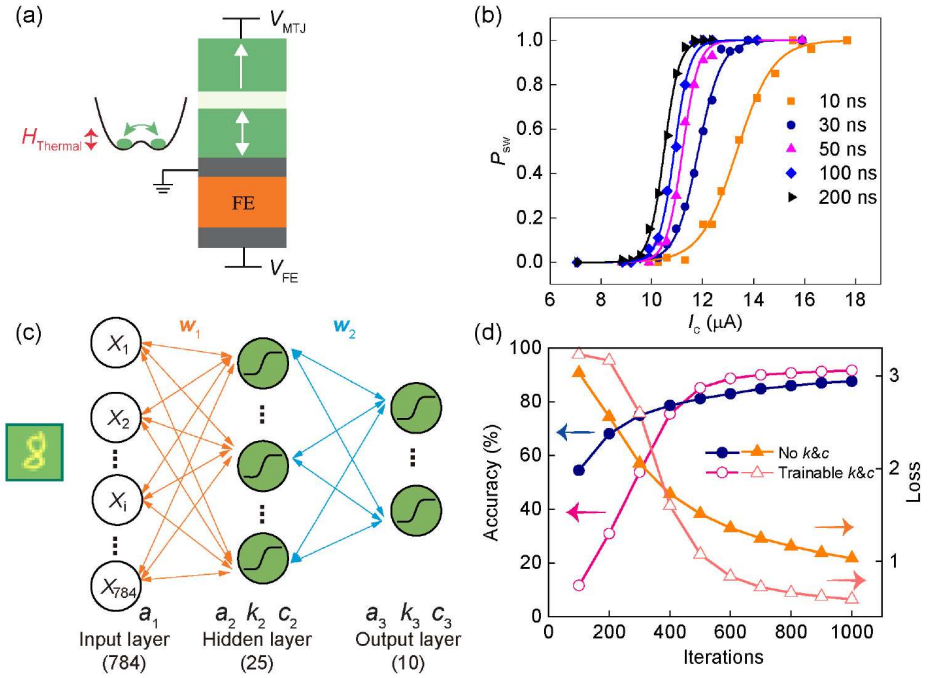


图 1 (网络版彩色)多层神经网络与器件性能关系的示意图. (a) 器件结构示意图. (b) 在E_B=7.5k_BT的系统中, 不同脉冲宽度下P_{sw}与I_c的关系. 模拟结果和拟合结果分别由符号和线条表示. (c) 神经网络示意图. w_1和w_2表示不同层之间神经元的连接. a, k和c的下标分别表示输入层、隐藏层和输出层. (d) 训练过程中识别准确率和损失函数的变化. 不可训练的k和c由实心符号表示, 可训练的k和c由空心符号表示. 对应的源代码发布在https://github.com/zhuzibn/trainable_activation_function/tree/main/2022TrainableNeuon

Figure 1 (Color online) Schematic illustration of the device structure and performance relationship with a multi-layer neural network. (a) Illustration of the device structure. (b) P_{sw} as a function of I_c under different pulse widths in the system with E_B=7.5k_BT. The simulation results and fitting are denoted by the symbols and lines, respectively. (c) A three-layer neural network. w_1 and w_2 denote the resistive connections of the neurons between different layers. The subscript 2 and 3 of a, k and c refer to the input layer, the hidden layer and the output layer, respectively. (d) The evolution of loss and accuracy in the system without trainable k and c are denoted by the filled symbols. The evolution of accuracy in the system with trainable k and c is denoted by the empty symbols. The source code corresponding to the trainable k&c can be accessed at https://github.com/zhuzibn/trainable_activation_function/tree/main/2022TrainableNeuon

本质上是不同的, k和c是体现神经元性质的局部参数, 而权重则是代表不同层之间连接强度的全局参数. 权重的物理实现可利用交叉电阻阵列(见附录S4), 而其数值维度取决于神经网络的拓扑结构. 例如, 本文中w_1的维度为784×25, 表示输入层(784个神经元)和隐藏层(25个神经元)之间的连接. 相比之下, 表示隐藏层中激活函数斜率的k_2的维度与局部神经元的维度相同, 为25×1. 为了能在训练过程中动态更新这两个参数, 本文推导了k和c的反向传播方程(见附录S5). 由于通过训练可获得最合适的k和c, 我们将本文提出的神经元命名为具有可训练激活函数的神经元. 如图1(d)中空心圆点所示, 在引入可训练神经元之后, 该神经网络的准确率提高至91.7%.

在之前的讨论中, 假设k和c的值可以任意变化, 但这只能在软件模拟中实现. 通过跟踪训练过程中k和c的变化, 我们发现k_2、k_3、c_2和c_3的范围分别为−3.1~3.2、−3.6~3.6、−0.6~0.5和−0.8~0.9(图2(a)). 然而, 在硬件实现中, 器件性能参数的可变范围需要根据底层物理原理进行评估. 如图1(b)和2(b)所示, 脉冲宽度的变化会同时影响激活函数的k和c. k和脉冲宽度之间呈线性关系, 通过改变脉冲宽度可以让自旋电子神经元的k_1和k_2处于反向传播算法所需的范围内(见图2(a)中的蓝色符号). 相比之下, 算法中c的范围(1 μA以内)和自旋电子神经元能提供的范围(10 μA以上)表现出较大差别. 此外, k和c之间存在由自旋电子物理产生的耦合关系, 例如, 脉冲宽度的增加会在降低所需的翻转电流的同时加速磁矩翻转, 这对实现可训练的激活函数造成了困难. 当调节脉冲宽度获得反向传播算法所需的k时, 由于k和c之间的耦合关系, c会变为一个新的值c_a. 同时, 类似k的更新, 算法中的c会变为另一个值c_b, 该值通常不等于c_a. 这种算法和物理器件之间的差异会导致训练失败. 我们通过仿真验证了这一点, 其中k是

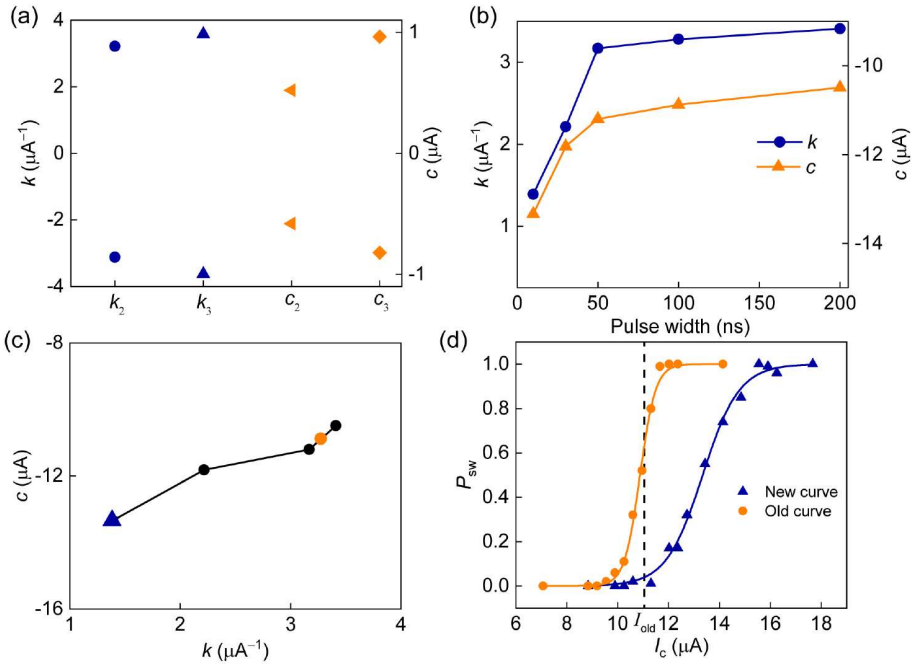


图 2 (网络版彩色)软件训练与物理器件参数关系的示意图. (a) 软件训练过程中得到的k_2, k_3, c_2和c_3的范围; (b) 不同脉冲宽度下, 物理器件能提供的k和c; (c) 从图2(b)中获得的k和c的关系; (d) 对应图2(c)中标记两点的激活函数

Figure 2 (Color online) Schematic illustration of the relationship between software training and physical device parameters. (a) Maximum ranges of k_2, k_3, c_2 and c_3 extracted from the software training; (b) physically allowed k and c under different pulse widths; (c) relation between k and c obtained from Fig. 2(b); (d) activation functions corresponding to the two points marked in Figure 2(c)

反向传播算法所需要的值, c则根据其和k之间关系的耦合进行调整(图2(c)). 模拟结果显示训练在几次迭代后失败(未在本文展示). 这是由于k的微小变化会导致c的大幅度调整(如图2(c)中圆形到三角形的变化), 导致激活函数的偏移(图2(d)). 可以看出, 在相同输入电流(I_{old})下, 新曲线在相应电流下的斜率为零, 这导致权重无法发生变化, 进而使得训练失败.

由此可见, 直接利用自旋电子器件实现可训练神经元会面临两个关键问题: (1) k和c之间存在耦合; (2) 算法要求的c值与自旋电子器件提供的c值不一致. 这两个问题都会导致训练失败. 因此, 我们提出两种解决方案. 第一种是利用器件物理将k和c解耦, 并添加一个偏置电流使c偏移到算法要求的值. 这种解决方案能够提高训练精度, 但需要增加额外的硬件. 另一种更易于实现的方案是去除可训练激活函数中的c, 仅保留k, 使激活函数的斜率可调但无法平移. 这种方案不需要对c进行单独调控, 而是通过控制脉冲宽度来调节k. 具体来说, 脉冲宽度的变化直接影响自旋电子器件的响应速度, 而通过调节脉冲宽度可以实现激活函数的动态变化, 提升训练的适应性. 具体方案将在后文进行详细讨论.

目前为止, 我们仅通过改变脉冲宽度来调节激活函数. 然而, 磁隧道结具有丰富的物理特性, 其他电学方式也同样可以对其性质进行调控. 例如, 较大电流会使翻转曲线向左偏移, 并且斜率变得更大, 这与改变脉冲宽度的效果相同. 此外, 在改变K_u来降低能量势垒(energy barrier, E_B)时, 只观察到翻转曲线的偏移, 其斜率并没有发生变化(图3(a)). 这是由于K_u决定了系统的能量势垒, 而降低K_u可以减小翻转电流. 然而, 尽管不同K_u系统在电流作用下的变化速度相同, 由于输入电流并没有变化, 斜率仍旧保持不变. 在硬件实现过程中K_u的变化可以利用电压控制磁各向异性(voltage-controlled magnetic anisotropy, VCMA)[42~48]或铁电应变效应来完成. 图1(a)所示的器件可以通过在FE层施加电压来调节CoFeB/MgO层中的垂直各向异性[27], 同样的效应也可以在结合自旋轨道力矩(spin orbit torque, SOT)和VCMA的三端器件中实现(见附录S6). VCMA技术通过施加电压调节器件中磁性材料的垂直各向异性. 我们在附录S6中详细探讨了如何在三端器件中结合VCMA和SOT来实现这一目标. 该方案通过同时控制电压和外部磁场, 能够精确调节能量势垒, 从而在硬件

上实现自适应的激活函数. 具体来说, 在外部场$\boldsymbol{H}_x$的辅助下, SOT导致自由层磁矩发生翻转. 同时, 通过给磁隧道结施加电压, VCMA效应将会导致K_u的改变[48]. 至此, 可以通过控制脉冲宽度或输入电流强度来调节k和c, 也可以通过控制K_u来调节c. 综合这两种方式, 本文实现了对k的单独调控(图3(b)). 此外, 在K_u变化过程中, 我们确认自由层磁矩在平衡态时始终保持在垂直方向, 验证了该方法的有效性. 图3(a)中能量势垒的公式为$E_B=(K_u-0.5\mu_0M_s^2)\times V_{FL}=15k_BT$, 其中$K_u$大于形状各向异性的贡献, 从而确保磁矩沿垂直方向. 我们也通过仿真模拟验证了该结果: 初始磁矩被设为沿z轴倾斜45°, 然后在$J_c=0$下弛豫, 最后观察到磁矩能回到z轴.

图4(a)展示了通过改变脉冲宽度和K_u得到的k的范围. 如图4(b)所示, 当把k作为额外的自由度添加到神经网络中时, 能获得91.3%的高识别准确率. 在此过程中, 带斜率可调的自旋电子神经元仅需500次迭代即可达到85%的准确率, 而在传统神经元系统中则需要800次迭代. 对比图1(d)和4(b)可以发现, 在低迭代次数时, 不同系统的准确率差异很大, 这是由于其初始参数不同. 合适的初始参数可以显著改善神经网络在低迭代次数时的性能(见附录S7). 用随机初始化的权重对该训练过程进行了10次计算, 结果显示其平均准确率和损失函数与单次计算相比没有显著差异. 此外, 我们也确认了k在训练过程中的变化始终处于其硬件允许的范围之内. 本工作发现学习率等其他参数也会影响训练结果. 尽管不同的学习率(0.1和0.01)会

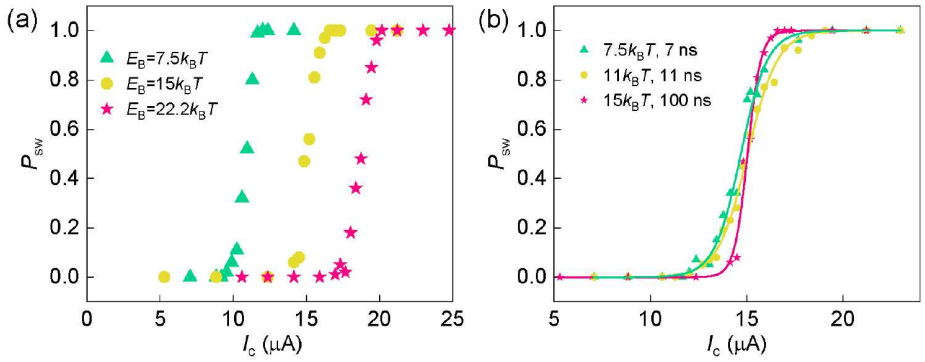


图 3 (网络版彩色)不同E_B与脉冲宽度下P_{sw}与I_c关系的分析. (a) 不同E_B下P_{sw}随I_c的变化, 其中E_B通过改变K_u获得, 其公式为$E_B=(K_u-0.5\mu_0M_s^2)\times V_{FL}$. 在30 ns脉冲下, $E_B=15k_BT$和I_c=15.6 μA的源代码发布在https://github.com/zhuzibn/ODE_LLG_integral/tree/master/2022TrainableNeuron. (b) 不同E_B和脉冲宽度组合下P_{sw}随I_c的变化

Figure 3 (Color online) Analysis of P_{sw} as a function of I_c under different E_B and pulse width combinations. (a) P_{sw} as a function of I_c under different E_B which is obtained by changing K_u that $E_B=(K_u-0.5\mu_0M_s^2)\times V_{FL}$. The source code corresponding to the system with 15k_BT under 30 ns pulse and I_c=15.6 μA can be accessed at https://github.com/zhuzibn/ODE_LLG_integral/tree/master/2022TrainableNeuron. (b) P_{sw} as a function of I_c under different combinations of E_B and pulse widths

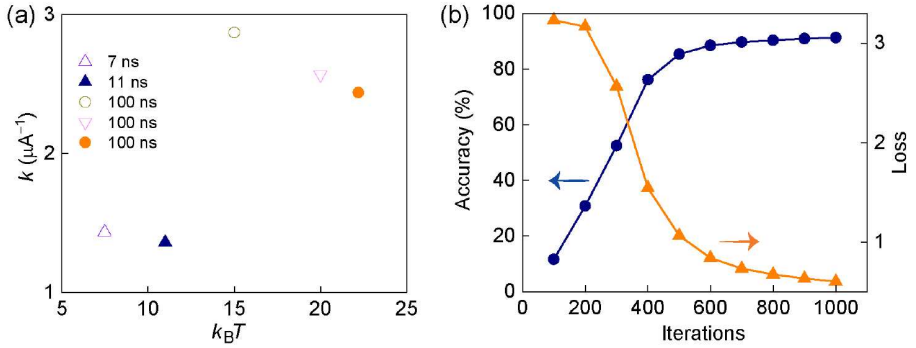


图 4 (网络版彩色)不同E_B与脉冲宽度下k值及可训练k系统性能分析. (a) 不同E_B和脉冲宽度组合下的k值; (b) 可训练k系统中识别准确率和损失函数的变化

Figure 4 (Color online) Analysis of k values and performance in trainable k systems under different E_B and pulse width combinations. (a) k under different combinations of E_B and pulse widths; (b) the evolution of loss and accuracy in the system with trainable k

改变训练速度和识别准确率，但不影响本文提出的关键结论，即具有可训练k的自旋电子神经元能够显著提高神经网络的性能. 最后，我们对硬件实现神经网络的能耗进行比较. 器件在运行过程中产生的热量会影响材料的磁性特性，特别是在小尺寸纳米器件中，温度升高可能会导致磁各向异性的变化，进而影响k和c参数的控制精度. 在实际应用中，大电流输入会增加系统的能量消耗，并产生额外的热量. 热效应还可能导致系统的稳定性下降，缩短器件的使用寿命. 为了解决这个问题，我们研究通过优化脉冲输入方式(如短脉冲调控)来减少能量损耗. 同时，改进器件的散热设计也是提高其可靠性的重要措施之一. 对于电流脉冲激发的磁隧道结神经元，其能耗为I^2Rt. 基于本文使用的平均电流I=15 μA、典型磁隧道结电阻R_{MTJ}=5 kΩ，以及平均脉冲宽度t=15 ns，可得每个神经元磁隧道结的能耗约为E_{MTJ}=16.9 fJ. 相比之下，CMOS实现单个神经元的能耗超过700 fJ[49,50]. 而目前文献中报道的自旋电子神经元的能耗为0.32 fJ[11], 1 fJ[49], 18~36 fJ[18]和60 fJ[50]. 本文提出的可训练神经元的能耗与其他自旋电子神经元相当，且远小于CMOS神经元. 值得注意的是，本文神经元k和c的更新是通过改变脉冲宽度来实现的，其能量计算已包含在E_{MTJ}. 因此，本文提出的可训练神经元在提高神经网络性能的同时不会增加额外的能耗.

2 结论

本文提出一种具有可训练激活函数的自旋电子神经元. 与前期研究不同，我们提出，若允许神经元激活函数的斜率和偏移在训练过程中按照算法要求动态变化，可以加速训练并显著提高识别准确率. 这一原理与深度神经网络中非常重要的批量归一化算法类似. 本文的研究进一步证明，利用自旋电子丰富的物理机制，可以硬件实现这种可训练神经元. 具体来说，通过电流驱动磁隧道结翻转，可以实现斜率和偏移可调的Sigmoid非线性激活函数. 其性质还可以通过脉冲电流宽度和磁各向异性等方式进行电学调控. 本文深入研究了电流驱动磁隧道结翻转的物理机制，提出了将斜率和偏移解耦的方法. 本文的研究为神经网络关键算法的硬件实现提供了思路，对开发基于自旋电子的神经网络具有重要贡献.

参考文献

1 Ebong I E, Mazumder P. CMOS and memristor-based neural network design for position detection. Proc IEEE, 2012, 100: 2050–2060
2 Sengupta A, Roy K. Spin-transfer torque magnetic neuron for low power neuromorphic computing. In: Proceedings of the 2015 International Joint Conference on Neural Networks (IJCNN), 2015
3 Sharad M, Augustine C, Panagopoulos G, et al. Spin-based neuron model with domain-wall magnets as synapse. IEEE Trans Nanotechnol, 2012, 11: 843–853
4 Lequeux S, Sampaio J, Cros V, et al. A magnetic synapse: Multilevel spin-torque memristor with perpendicular anisotropy. Sci Rep, 2016, 6: 31510
5 Kurenkov A, DuttaGupta S, Zhang C, et al. Artificial neuron and synapse realized in an antiferromagnet/ferromagnet heterostructure using dynamics of spin–orbit torque switching. Adv Mater, 2019, 31: 1900636
6 Zhang S, Luo S, Xu N, et al. A spin–orbit-torque memristive device. Adv Elect Mater, 2019, 5: 1800782
7 Zheng Q, Mi Y, Zhu X, et al. Anticipative tracking with the short-term synaptic plasticity of spintronic devices. Phys Rev Appl, 2020, 14: 044060
8 Yu Z, Shen M, Zeng Z, et al. Voltage-controlled skyrmion-based nanodevices for neuromorphic computing using a synthetic antiferromagnet. Nanoscale Adv, 2020, 2: 1309–1317
9 Sengupta A, Parsa M, Han B, et al. Probabilistic deep spiking neural systems enabled by magnetic tunnel junction. IEEE Trans Electron Devices, 2016, 63: 2963–2970
10 Behin-Aein B, Diep V, Datta S. A building block for hardware belief networks. Sci Rep, 2016, 6: 29893
11 Liyanagedera C M, Sengupta A, Jaiswal A, et al. Stochastic spiking neural networks enabled by magnetic tunnel junctions: From nontelegraphic to telegraphic switching regimes. Phys Rev Appl, 2017, 8: 064017
12 Shim Y, Chen S, Sengupta A, et al. Stochastic spin-orbit torque devices as elements for bayesian inference. Sci Rep, 2017, 7: 14101
13 Camsari K Y, Faria R, Sutton B M, et al. Stochastic *p*-bits for invertible logic. Phys Rev X, 2017, 7: 031014
14 Ostwal V, Debashis P, Faria R, et al. Spin-torque devices with hard axis initialization as Stochastic Binary Neurons. Sci Rep, 2018, 8: 16689
15 Cai J, Fang B, Zhang L, et al. Voltage-controlled spintronic stochastic neuron based on a magnetic tunnel junction. Phys Rev Appl, 2019, 11: 034015

16 Deng J, Miriyala V P K, Zhu Z, et al. Voltage-controlled spintronic stochastic neuron for restricted Boltzmann machine with weight sparsity. IEEE Electron Device Lett, 2020, 41: 1102–1105
17 Chen Y B, Yang X K, Yan T, et al. Voltage-driven adaptive spintronic neuron for energy-efficient neuromorphic computing. Chin Phys Lett, 2020, 37: 078501
18 Siddiqui S A, Dutta S, Tang A, et al. Magnetic domain wall based synaptic and activation function generator for neuromorphic accelerators. Nano Lett, 2020, 20: 1033–1040
19 Kobayashi K, Borders W A, Kanai S, et al. Sigmoidal curves of stochastic magnetic tunnel junctions with perpendicular easy axis. Appl Phys Lett, 2021, 119: 132406
20 Cai J, Zhang L, Fang B, et al. Sparse neuromorphic computing based on spin-torque diodes. Appl Phys Lett, 2019, 114: 192402
21 Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of the International Conference on Machine Learning, 2015, 37: 448–456
22 Li Y, Wang N, Shi J, et al. Adaptive Batch Normalization for practical domain adaptation. Pattern Recognition, 2018, 80: 109–117
23 Yao Z, Cao Y, Zheng S, et al. Cross-Iteration Batch Normalization. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019. 12326–12335
24 Yu G L, He X Y, Shi S B, et al. The combined effect of spin-transfer torque and voltage-controlled strain gradient on magnetic domain-wall dynamics: Toward tunable spintronic neuron. Chin Phys Lett, 2024, 41: 057502
25 Wu Y, Zhao M, Ding X. A new kind of neuron model with a tunable activation function and its applications. Sci China Ser E-Technol Sci, 1997, 40: 105–112
26 Bilonoh B, Bodyanskiy Y, Kolchygin B, et al. Tunable activation functions for deep neural networks. In: Proceedings of the Lecture Notes in Computational Intelligence and Decision Making. Cham: Springer, 2022
27 Yu G, Wang Z, Abolfath-Beygi M, et al. Strain-induced modulation of perpendicular magnetic anisotropy in Ta/CoFeB/MgO structures investigated by ferromagnetic resonance. Appl Phys Lett, 2015, 106: 072402
28 Slonczewski J C. Current-driven excitation of magnetic multilayers. J Magn Magn Mater, 1996, 159: L1–L7
29 Berger L. Emission of spin waves by a magnetic multilayer traversed by a current. Phys Rev B, 1996, 54: 9353–9358
30 Gilbert T L. Classics in magnetics a phenomenological theory of damping in ferromagnetic materials. IEEE Trans Magn, 2004, 40: 3443–3449
31 Zhu Z, Deng J, Fong X, et al. Voltage-input spintronic oscillator based on competing effect for extended oscillation regions. J Appl Phys, 2019, 125: 183902
32 Ikeda S, Miura K, Yamamoto H, et al. A perpendicular-anisotropy CoFeB–MgO magnetic tunnel junction. Nat Mater, 2010, 9: 721–724
33 Yakata S, Kubota H, Suzuki Y, et al. Influence of perpendicular magnetic anisotropy on spin-transfer switching current in CoFeB/MgO/CoFeB magnetic tunnel junctions. J Appl Phys, 2009, 105: 07D131
34 Myers E B, Albert F J, Sankey J C, et al. Thermally activated magnetic reversal induced by a spin-polarized current. Phys Rev Lett, 2002, 89: 196801
35 Bedau D, Liu H, Sun J Z, et al. Spin-transfer pulse switching: From the dynamic to the thermally activated regime. Appl Phys Lett, 2010, 97: 262502
36 Brown W F. Thermal fluctuations of a single-domain particle. Phys Rev, 1963, 130: 1677–1686
37 Li Z, Zhang S. Thermally assisted magnetization reversal in the presence of a spin-transfer torque. Phys Rev B, 2004, 69: 134416
38 Butler W H, Mewes T, Mewes C K A, et al. Switching distributions for perpendicular spin-torque devices within the macrospin approximation. IEEE Trans Magn, 2012, 48: 4684–4700
39 Miriyala V P K, Fong X, Liang G. Influence of size and shape on the performance of VCMA-based MTJs. IEEE Trans Electron Devices, 2019, 66: 944–949
40 Das D, Fong X. A Fokker–Planck Approach for modeling the stochastic phenomena in magnetic and resistive random access memory devices. IEEE Trans Electron Devices, 2021, 68: 6124–6131
41 Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proc IEEE, 1998, 86: 2278–2324
42 Weisheit M, Fähler S, Marty A, et al. Electric field-induced modification of magnetism in thin-film ferromagnets. Science, 2007, 315: 349–351
43 Maruyama T, Shiota Y, Nozaki T, et al. Large voltage-induced magnetic anisotropy change in a few atomic layers of iron. Nat Nanotech, 2009, 4: 158–161
44 Duan C G, Velev J P, Sabirianov R F, et al. Surface magnetoelectric effect in ferromagnetic metal films. Phys Rev Lett, 2008, 101: 137201
45 Shiota Y, Nozaki T, Bonell F, et al. Induction of coherent magnetization switching in a few atomic layers of FeCo using voltage pulses. Nat Mater, 2012, 11: 39–43
46 Alzate J G, Amiri P K, Upadhyaya P, et al. Voltage-induced switching of nanoscale magnetic tunnel junctions. In: Proceedings of the 2012 International Electron Devices Meeting, 2012. New York: IEEE, 2012

47 Grezes C, Ebrahimi F, Alzate J G, et al. Ultra-low switching energy and scaling in electric-field-controlled nanoscale magnetic tunnel junctions with high resistance-area product. Appl Phys Lett, 2016, 108: 012403
48 Xin Y, Zhou K, Fong X, et al. Electrical tunable spintronic neuron with trainable activation function. 2022, arXiv: 221113391 arXiv: 221113391
49 Sharad M, Fan D, Roy K. Spin-neurons: A possible path to energy-efficient neuromorphic computers. J Appl Phys, 2013, 114: 234906
50 Ramasubramanian S G, Venkatesan R, Sharad M, et al. SPINDLE: SPINtronic deep learning engine for large-scale neuromorphic computing. In: Proceedings of the 2014 International Symposium on Low Power Electronics and Design, 2014. 15–20

补充材料

附录S1 批量归一化算法的核心思想与单个器件的激活函数调控
附录S2 翻转概率详细计算过程
附录S3 k与c的定义
附录S4 电阻交叉阵列的实现
附录S5 k与c传播公式的推导
附录S6 三端器件
附录S7 k, c和w的初始化

本文以上补充材料见网络版csb.scichina.com. 补充材料为作者提供的原始数据, 作者对其学术质量和内容负责.

Summary for “利用单个自旋电子神经元实现类批量归一化算法”

Implementing batch normalization-like algorithm with a single spintronic neuron

Zhuo Xu[1†], Yue Xin[1†], Zhengping Yuan[1], Kang Zhou[1], Yumeng Yang[1,2], Shenghua Gao[1] & Zhifeng Zhu[1,2*]

[1] *School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China*
[2] *Shanghai Engineering Research Center of Energy Efficient and Custom AI IC, Shanghai 201210, China*
† Equally contributed to this work
* Corresponding author, E-mail: zhuzhf@shanghaitech.edu.cn

Spintronic devices have garnered significant attention for their potential in hardware-based implementations of artificial neurons. One of the widely explored mechanisms in this context is the stochastic switching of magnetic tunnel junctions, driven by spin-transfer torque, which is often utilized to produce a sigmoid-shaped activation function. However, a critical limitation in previous studies is that the shape of the activation function remains static throughout the neural network training process. This restricts the ability to adaptively update the weights, leading to inferior performance of the neural network. In this work, we exploit the physics behind the spin torque induced magnetization switching to enable the dynamic change of the activation function during the training process. Specifically, we demonstrate that the pulse width and magnetic anisotropy of the device can be electrically controlled to modify the slope of the activation function, allowing for a faster or slower change in output as required by the backpropagation algorithm. This dynamic control of the activation function is similar to the widely used concept of batch normalization in machine learning, which adjusts and stabilizes the output of each layer during training, thus accelerating the learning process. Thus, in our approach, the ability to adjust the slope of the activation function provides a more flexible hardware-based solution for neural networks. This overcomes the limitations of previous studies where algorithms were predominantly implemented in software and lacked such adaptability in the hardware fields. This work demonstrates that spintronic devices can perform sophisticated tasks like dynamic adjustment of activation functions, marking a significant step forward in hardware-based neural networks. In addition, trainable spintronic neurons can substantially improve the performance of neural networks. For instance, in a handwritten digit recognition task, the recognition accuracy was improved from 88% to 91.3% by employing these dynamically trainable spintronic neurons. This work not only addresses the constraints of fixed activation functions during training but also opens up new possibilities for the integration of machine learning algorithms with spintronic hardware. By utilizing the physical properties of spin-transfer torque and magnetization switching, the proposed spintronic neurons provide a flexible, energy-efficient, and high-performance solution for hardware neural networks. This breakthrough underscores the potential for spintronics to drive the development of next-generation neuromorphic computing systems, offering both performance improvements and significant energy savings in artificial intelligence applications. As spintronics technology continues to evolve, it is poised to become a cornerstone of future AI hardware innovations, enabling more advanced, efficient, and scalable neural computing architectures.

spintronic neuron, spin torque, nonlinear activation function, neuromorphic computing

doi: 10.1360/TB-2024-0879