上海科技大学
ShanghaiTech University

硕士学位论文

自旋突触和自旋神经元实现具有可训练激活函数的类脑计算网络

作者姓名：辛玥

指导教师：祝智峰 助理教授

上海科技大学信息科学与技术学院

学位类别：工学硕士

一级学科：电子科学与技术

学校/学院名称：上海科技大学信息科学与技术学院

2023 年 6 月

Neuromorphic Computing Networks with Trainable Activation Function by Spin Synapses and Spin Neurons

A thesis submitted to

ShanghaiTech University

in partial fulfillment of the requirement

for the degree of

Master of Science in Engineering

in Electronic Science and Technology

By

Xin Yue

Supervisor: Professor Zhu Zhifeng

School of Information Science and Technology

ShanghaiTech University

June, 2023

上海科技大学

研究生学位论文原创性声明

本人郑重声明: 所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明或致谢。

作者签名: 辛湖

日　　期: 2023.6.9

上海科技大学

学位论文授权使用声明

本人完全了解并同意遵守上海科技大学有关保存和使用学位论文的规定，即上海科技大学有权保留送交学位论文的副本，允许该论文被查阅，可以按照学术研究公开原则和保护知识产权的原则公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存、汇编本学位论文。

涉密及延迟公开的学位论文在解密或延迟期后适用本声明。

作者签名: 辛湖　　导师签名: 祝智峰

日　　期: 2023.6.9　　日　　期: 2023.6.9

摘 要

基于冯·诺依曼架构的传统计算方法，因其系统的能耗较高以及其处理单元和存储单元之间数据交换带宽受限，难以实现更高的计算效率。为了克服这些限制，迫切需要一种更高效的计算架构。类脑计算在能源效率和高性能计算方面的非凡表现，被认为是一种颇具潜力的候选方案。尽管利用传统的电子器件也可以模仿人脑的拓扑结构，但由这些器件搭建的系统需要很高的耗能和较大的面积。以磁性隧道结为代表的自旋电子器件表现出显著的高能效，低功耗，耐用性，非易失性，以及与生物神经系统的相似性，使得自旋电子器件实现类脑计算成为可能。在这项工作中，本文主要围绕基于自旋突触和自旋神经元实现具有可训练激活函数的类脑计算网络这一课题展开研究，结合自旋突触和自旋神经元背后的物理特性，可靠地实现了具有可训练激活函数的自旋神经网络。本文的主要创新研究内容如下：

(1) 第3章通过研究自旋神经元器件的磁化翻转得到了能够产生可调节激活函数和可训练激活函数的自旋神经元模型，搭建了三层自旋神经网络执行推理识别 MNIST 手写数字的任务测试该自旋神经元的功能。

(2) 第3章引入可训练激活函数后，自旋神经网络的推理测试准确率可以从 88% 提高到 91.3%，且不会引入额外的能耗，有效降低了训练的时间成本和系统整体的面积，与基于 CMOS 搭建的神经网络相比，自旋神经网络展现出更低的功耗。

(3) 可训练激活函数的思想类似于批量归一化算法，第3章推导了具有可训练激活函数的自旋神经网络的训练学习过程的算法。

(4) 第4章研究了不同量化策略下自旋突触单元的线性度、对称性和状态数等非理想特性对自旋神经网络的硬件实现的影响。改进的具有可训练激活函数的自旋神经网络的推理测试准确率达到了 95% 以上。

关键词：自旋电子器件，磁性隧道结，类脑计算，自旋神经元，自旋突触

Abstract

The conventional computing method based on the von Neumann architecture is limited by a series of problems such as high energy consumption, finite data exchange bandwidth between processors and storage media, etc., and it is difficult to achieve higher computing efficiency. A more efficient unconventional computing architecture is urgently needed to overcome these problems. Neuromorphic computing has been considered to become the competitive candidate for unconventional computing, due to its extraordinary potential for energy-efficient and high-performance computing. Although conventional electronic devices can mimic the topology of the human brain, these require high power consumption and large area. Spintronic devices represented by magnetic tunnel junctions exhibit remarkable high-energy efficiency, non-volatility, and similarity to biological nervous systems, making them one of the promising candidates for neuromorphic computing. In this work, we mainly focus on the subject of realizing neuromorphic computing networks with trainable activation function based on spin synapses and spin neurons. Combined with the physical properties behind the spin synapses and spin neurons, we reliably realized the spin neural network with trainable activation function. The main innovative research contents of this thesis are as follows:

(1) We exploit a spin neuron model with trainable activation function based on the physics behind the spin neurons in chapter 3. The spin neuron model can generate tunable activation function and trainable activation function, which are obtained by studying the magnetization switching of the spin neuron device. To test the function of the spin neuron, we built a three-layer spin neural network to perform the inference task of MNIST hand-written digit recognition.

(2) The inference test accuracy of the spin neural network with trainable activation function can be improved from 88% to 91.3% in chapter 3, without introducing additional energy consumption, which effectively reduces the training time cost and the overall area of the system. Compared with neural networks based on CMOS, the proposed spin neural network exhibits lower power consumption.

(3) The idea of trainable activation function is similar to the batch normalization algorithm. We derived the algorithm of the learning process of the spin neural network with trainable activation function in chapter 3.

(4) We studied the non-ideal characteristics such as linearity, symmetry, and number of states of spin synaptic units under different quantization strategies when implementing spin neural networks in hardware in chapter 4. The inference test accuracy of the improved spin neural network with trainable activation function has reached more than 95%.

Key Words: Spintronic devices, Magnetic tunnel junction, Neuromorphic computing, Spin neuron, Spin Synapse

目 录

图形列表

表格列表

第 1 章　绪论

1.1　基于自旋电子器件实现类脑计算网络的研究背景及意义

如今的信息社会是构建在数字计算机的基础之上。半个世纪以来，数字计算机取得了迅猛发展，已经能够可靠地执行复杂任务。人工智能 (Artificial Intelligence, AI)、大数据和物联网 (Internet of Things, IoTs) 的空前发展重新定义了计算的概念。计算机的硬件设计不仅需要考虑计算吞吐量、功耗和外形尺寸等严格的要求，还要满足不断增长的对计算性能的需求。然而，传统的 CMOS(Complementary Metal Oxide Semiconductor) 数字计算机采取冯·诺依曼架构，在物理上分离的存储和计算单元导致需要频繁的数据传输，从而产生大量的能耗和时间延迟，这也就是所谓的冯·诺依曼瓶颈。相比而言，人脑采用了一种迥异于数字计算机的架构，可以在非常低的功率下高效地执行复杂的任务。此外，众所周知，随着晶体管的尺寸愈来愈接近其物理极限，缩小晶体管的尺寸也变得低效，这大大增加了提升数字计算系统性能的难度。因此，对计算机的架构和其构建单元进行根本性的变革势在必行。

类脑计算 (Neuromorphic Computing，NC)，也叫神经形态计算，是指受到大脑工作原理的启发而设计的电路，可以高效节能地执行计算任务，有望实现人工智能的同时降低计算能耗的需求。这个跨学科领域始于利用 CMOS 电路实现神经网络的算法，现已发展到通过基于脉冲的编码和事件驱动的硬件架构来实现与智能算法的交互。在过去的十年中，脉冲神经网络 (Spiking Neural Network, SNN) 已成为脑启发式的流行架构之一。

在 SNN 中，信息体现在时序编码的脉冲信号中，神经元之间的通信是通过脉冲完成的。在这样的网络中，脉冲时序依赖可塑性 (Spike timing dependent plasticity, STDP) 机制需要根据神经元传输的脉冲的时间信息对突触的权重进行实时更改。Kaushik Roy 教授的团队基于自旋电子器件进行了多次尝试，探索脉冲神经网络的实现。在这项工作中[1]，创新地利用了基于磁畴壁运动的自旋电子器件实现按照脉冲时序依赖可塑性进行学习的突触，利用自旋轨道力矩的物理机制将读出电流与写入电流的路径分离。该自旋电子器件与传统 CMOS 一起构

建的脉冲神经元和工作在晶体管亚阈值范围内的学习电路一起仿真模拟了脉冲神经网络，执行了模式识别的任务。另外，他们还使用单一的三端磁性隧道结——重金属多层器件来模拟随机神经元和突触的动力学特性[2]，利用自旋电子器件本身具有随机热扰动的特征，研究了脉冲神经元和突触的随机计算模型，探索实现了一个概率神经元和一个随机二元突触。概率随机神经元根据输入电流有条件地发出一个输出 (突触后) 脉冲。随机二元突触以一个可以实现可塑性的概率学习算法为行为指导，来改变二元突触的状态，它的切换概率取决于相应的突触前和突触后脉冲的时间差。不仅仅是常见的磁性隧道结，基于斯格明子这类新型自旋电子器件实现脉冲神经元和突触的核心功能[3]，来模拟全自旋深度脉冲神经网络。突触权重可以通过读取磁性隧道结下的斯格明子数量进行调整，并且通过设置具有不同电导范围的多个分支来提高权重的更新范围。此外，基于斯格明子设计的神经元能够以超低电流切换开关状态，为基于斯格明子的突触提供超低电压操作的可能，也为在低功耗深度脉冲神经架构中基于斯格明子器件构建低功耗深度脉冲神经网络提供了新的可能。

上述提到的磁性隧道结和斯格明子只是自旋电子器件中常见的两类。除了大规模应用于存储[5-6]的传统的磁性隧道结之外，畴壁器件[7-8]、斯格明子器件[9]、自旋波器件[10]和随机器件[11]也进行了大量的应用于计算方面的研究，例如类脑计算[12-14]、逻辑计算[5,15]和随机计算[11,16]。

许多重要的科研成果已经证明了自旋电子器件在计算方面的应用中具有优越性。例如，自旋电子器件能够基于其底层的物理原理以生物启发的方式存储和处理信息，这样的存算方式克服了冯·诺依曼瓶颈，并实现了更高的类脑计算效率[17-21]。对于需要低功耗高精度的布尔逻辑计算，可以将自旋电子器件应用于存内计算，因为它固有的非易失性可以降低能耗，缓解晶体管[22-24]的尺寸限制。此外，根据自旋电子器件的随机性，可以利用它按照一定概率进行随机切换的特质，将它作为一个介于经典位和量子位 (Quantum bit, Q-bit) 之间的概率位 (Probabilistic bit, P-bit)，应用于高能效的随机计算[11,16]。

如图 1.1(a) 所示的基于自旋力矩的自旋电子器件，是可以用于构建神经网络的硬件模块。由上述器件可以实现神经网络基本单元——神经元和突触，神经元与突触相互连接构成了可以执行复杂任务的神经网络。图 1.1(b) 中展示的自旋突触和神经元，它们的开关切换机制均基于自旋力矩。除了自旋转移力矩

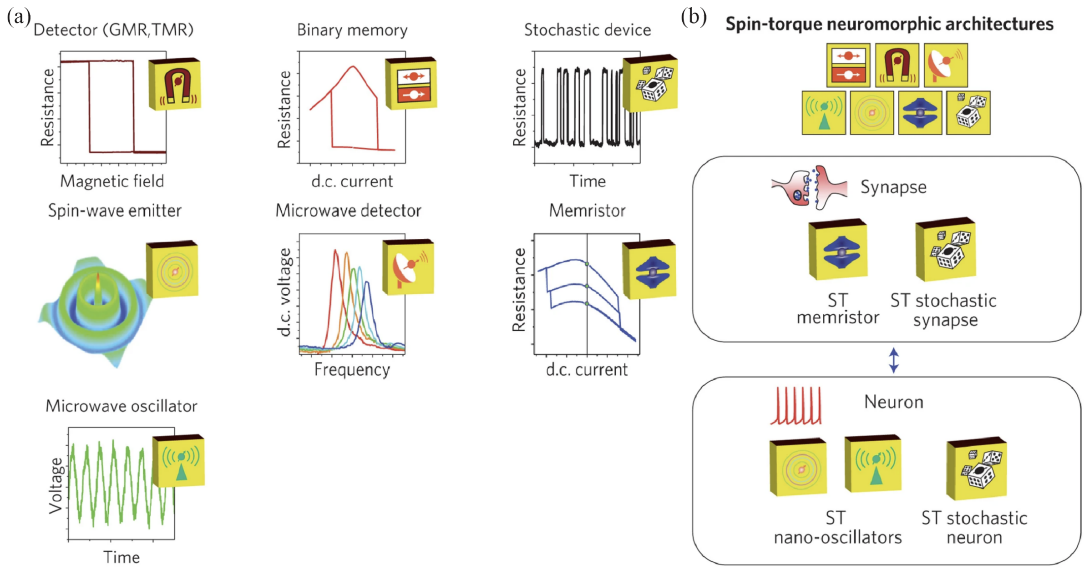

图 1.1　自旋神经网络的构建单元。(a) 基于自旋力矩的自旋电子器件，具有不同的功能，用于构建类脑计算架构的硬件模块。(b) 自旋神经网络的基本单元自旋突触和神经元。自旋力矩忆阻器和随机自旋力矩突触都已实现。不同概念模型的基于自旋力矩的神经元，即自旋力矩纳米振荡器和自旋力矩随机神经元也已实现[4]。

Figure 1.1　The construction unit of spin neural network.(a)Spintronic devices based on spin torque (ST) have different functions and can be assembled to build a new hardware neuromorphic computing architecture. (b)The diagram of the basic units constituting the spin neural network: spin Synapses and spin neurons. Different spin torque based Synapses, ST memristors and ST stochastic Synapses have been proposed. Using ST nano-oscillator or ST stochastic neuron, different conceptual models of neurons based on spin torque have been proposed[4].

(Spin Transfer Torque, STT)[25]外，过去十年中发现的开关机制还包括自旋轨道力矩 (Spin Orbital Torque, SOT)[26]和电压控制的磁各向异性 (Voltage Controlled Magnetic Anisotropy, VCMA)[27]，上述开关机制的具体原理在 2.2节中详细展开。

除了传统的铁磁材料外，亚铁磁[28-30]，反铁磁[31-32]和二维材料[33-34]也已被用于制备自旋电子器件。另外，拓扑绝缘体是一种具有非常特殊性质的材料，它们的内部是绝缘的，而表面具有导电性。这些表面导电态受到强自旋轨道耦合的影响，因此在拓扑绝缘体中可以观察到自旋动量锁定现象。自旋动量锁定是指拓扑绝缘体表面导电态中电子的自旋方向与其动量方向紧密耦合。这种耦合关系导致电子的自旋在特定方向上运动时会发生改变，而在另一特定方向上运动时保持不变。换句话说，这种锁定关系使得电子在沿着不同方向运动时呈现出不同

的自旋极化特性。在拓扑绝缘体中，自旋动量锁定现象的出现与时间反演对称性破缺有关。由于拓扑绝缘体具有非平凡的拓扑不变量，它们的表面态具有特殊的能带结构，即能带交叉点。这些交叉点是由于电子的自旋和动量之间的强耦合所形成的，这种耦合关系使得不同自旋方向的电子在交叉点处具有相反的动量。拓扑绝缘体中的自旋动量锁定现象有许多潜在应用，包括自旋电子学、低耗散电子器件和拓扑量子计算。这些应用的核心思想是利用自旋动量锁定现象实现对电子自旋状态的有效控制和高效输运，从而为未来的电子器件和量子技术提供新的发展方向。

自旋电子纳米器件主要利用了电子的磁性和电学特性，其背后的基本原理符合生物神经系统的标准，为基于自旋电子学的大脑启发式计算奠定了基础。而磁性隧道结 (Magnetic Tunnel Junction, MTJ) 作为类脑计算元件尤其值得关注，因为它们能够与标准集成电路兼容，还支持多种功能。因此，基于 MTJ 的类脑计算可以大大提高计算的能量效率并减少计算架构所需电路的面积，为高性能计算的实现提供了一种变革性的解决方案。

1.2 类脑计算网络的研究现状及发展趋势

现代计算机以冯·诺依曼体系架构为基础，以序列化、确定性、高精度的方式解决数值问题，它经过数十年的广泛发展，仍然是当今信息处理的主流方法。然而，随着计算量和复杂性逐渐增加，大数据的大规模应用，冯·诺依曼计算范式的弊端严重降低了计算效率。因为在处理器和存储单元之间不断传输繁多的信息不可避免地会导致大量的能源消耗和时间成本。因此，在寻求解决冯·诺依曼瓶颈的背景下[35]，需要超越冯·诺依曼架构的新型计算范式，即类脑计算。

与需要给出确定的准确结果的冯·诺依曼范式相反，类脑计算[36]采用充分冗余的计算，并返回满足识别、分类、预测、优化等目标的近似结果。这个新型计算范式有望在涉及大数据处理时实现高性能和高能效计算。它主要的竞争力体现在以下三方面：第一，类脑计算使用许多低精度或概率计算，因此本质上对结果具有一定的容错性；第二，用于类脑计算中的大多数范式是并行的，这将极大程度地有利于加快计算速度；第三，一些与类脑计算相关的架构设计是存算一体的，内存不仅可以存储信息同时可以处理信息，降低了存算之间数据交换的频率，降低了信息传输的能耗和时间延迟。

表 1.1　国内外关于类脑计算研究的发展

Table 1.1　The development of neuromorphic computing

国家	时间/年	相关的类脑计算研究
美国	2004	斯坦福大学研究所研制第一款类脑芯片 Neurogrid 芯片
	2005	美国正式启动 SynAPSE 项目计划
	2008	惠普公司实现 memristor 原型
	2013	启动 BRAIN 计划
	2014	Neuromorphic Engineering 项目中 IBM 公司研制出 TrueNorth 芯片
	2016	脉冲神经元在 IBM 苏黎世研究院诞生
	2017	Intel 公司研发 Loihi 芯片
	2017	Intel 公司研发神经形态计算系统 Pohoiki Beach
	2017	Koniku 的神经元硅芯片 Koniku Kore
	2020	Intel 宣布 Pohoiki Springs 将全面投入使用
	2022	实施 Brain Research through Advancing Innovative Neurotechnologies, BRAIN 计划
日本	2014	正式启动 Brain/MINDS 项目
	2017	Kamitani 团队提出的基于深度学习的神经解码研究
欧盟	2005	海德堡大学牵头研制基于模拟混合信号 (AMS) 的类脑芯片
	2005	IBM 牵头启动了 Blue Brain 项目
	2011	欧盟启动 Brain Scales 项目
	2013	人类脑计划 (Human Brain Project) 提出
	2016	英国曼彻斯特大学的 SpiNNaker 系统对外开放使用
	2022	SynSense 批量生产感算一体动态视觉智能 SoC 芯片 Speck
中国	2015	浙江大学研制出达尔文芯片，支持脉冲神经网络的类脑芯片
	2019	浙江大学研发完成达尔文 2 代类脑芯片
	2019	清华大研发的新型人工智能芯片天机芯 (Tianjic)
	2020	浙江大学 Darwin Mouse 系统平台
	2022	科技创新 2030—“脑科学与类脑研究”重大项目实施

类脑计算由于其大规模并行性、高能效、对复杂输入和瞬时变化的适应性以及对误差的固有容忍度，吸引了各国的关注，开展类脑计算相关内容的研究是大势所趋。自美国的斯坦福大学在 2004 年开发出第一个以大脑为灵感的芯片 Neurogrid 以来[37]，各国系统地、有目的地启动了适合其发展的类脑计算研究项目。如表 1.1所示，类脑计算相关的研究和成果正逐渐在国际范围内出现，美国、日本、欧盟和中国都投入了大量的时间和资源。

相比于欧盟和美国，中国开始类脑计算的研究是谨慎的。Darwin 芯片是浙江大学计算机学院的研究团队于 2015 年推出的一款基于神经元模型的人工智能芯片。该芯片采用类脑计算原理，可以模拟大脑神经元和突触的工作机制，实现人工智能的感知、学习和决策。Darwin 芯片拥有 256 个神经元和 65,536 个突触连接，可以进行图像、语音和行为等多种模式的处理和识别任务。Darwin2 芯片也是由浙江大学计算机学院的研究团队开发的，是 Darwin 系列芯片的升级版。该芯片于 2019 年推出，拥有超过 450 万个可编程神经元和 100 亿个可编程突触连接，是目前世界上最大规模的可重构神经网络芯片之一。

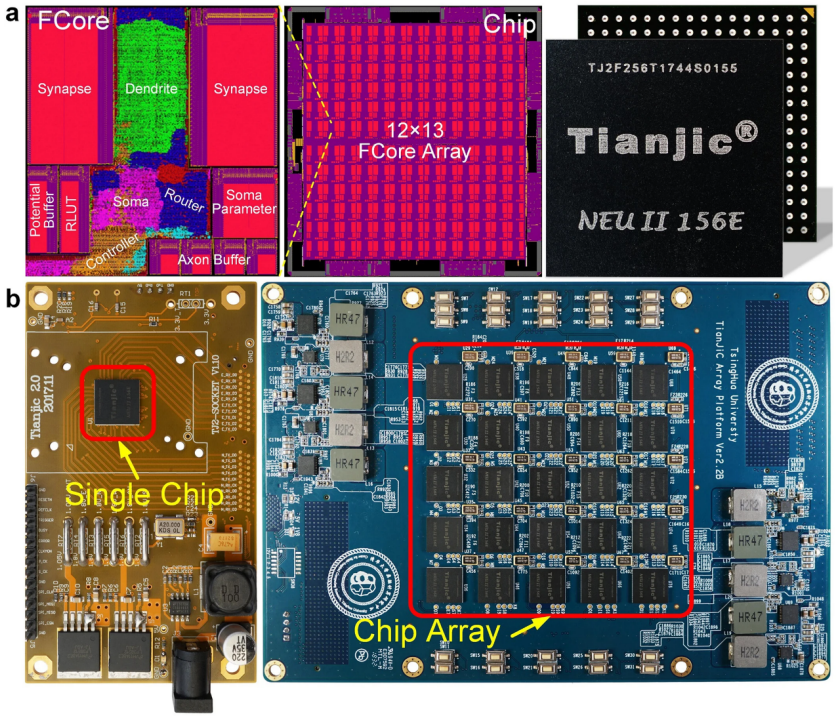

图 1.2　天机芯的芯片布局。(a) 天机芯的芯片布局和封装。(b) 配备了单个天机芯或 5×5 芯片阵列的测试板[38]。

Figure 1.2　The chip layout of Tianjic.(a)The chip layout and packaging of Tianjic. (b)Test board equipped with a single Tianjic or a chip array (5 × 5 size)[38].

与 Darwin1 芯片相比，Darwin2 芯片的规模更大、性能更强。它能够支持更广泛的人工智能应用场景，包括图像和语音的识别、机器翻译、自然语言处理等。Darwin2 芯片也采用了基于神经元模型的类脑计算原理，可以模拟大脑神经元和突触的工作机制，实现人工智能的感知、学习和决策。Darwin2 芯片的研制成功标志着中国在类脑计算领域的研究和发展取得了重要进展。

如图 1.2所示，Tianjic 芯片[38]是清华大学施路平团队提出的一种类脑芯片，它是目前国内研发的一款最为先进的类脑芯片之一。该芯片采用了脉冲神经元模型和基于突触可塑性的学习算法，能够模拟大规模的神经网络。Tianjic 芯片共有 4 个核，每个核可进行 32 亿次每秒的计算，总体计算速度约为 1.28 万亿次每秒。该芯片已经成功应用于图像、语音和机器人控制等领域。

上述介绍的基于 CMOS 的类脑芯片都表现出了强大的计算性能，但是为了硬件实现单个突触或神经元单元，最初通常需要成百上千个晶体管。由于能量和面积的要求，采用大量的晶体管对于类脑计算任务而言是不可取的。近年来，研究团队陆续提出能够模拟突触、神经元功能的单个自旋电子器件。得益于其非易失性、可塑性、随机性和高频振荡等卓越的特性，自旋电子器件成为了实现类脑计算的潜在技术之一[12]。磁性隧道结是自旋电子器件中用于信息存储的一种经典结构，由于其还具有 1.1节中提到的诸多特性，以及出色的耐用性和 CMOS 兼容性，磁性隧道结有望成为高性能类脑计算硬件电路的主要构建模块。

1.3　课题的研究目的及论文内容安排

基于自旋电子器件的计算应用正在快速发展，然而统一集成自旋电子学，电子科学和计算机科学的理论框架还有待开发。

随着深度学习的快速发展，神经网络的硬件实现要求更高的性能和更低的功耗。利用 CMOS 实现人工神经元和突触需要很大的硅面积，并且由于其易失性，它们的能量效率也很低[39-40]。相比之下，自旋电子器件具有非易失性，利用其本身的非易失性可以实现非常低能耗的自旋神经元和自旋突触[41-42,18,43-45]。此外，自旋器件丰富的物理特性使其能够使用简单的器件实现复杂的功能。因此，它们相比于 CMOS 更紧凑。另外，随机磁性隧道结已被提议用于产生不同的激活函数，例如 sigmoid[46-50,18,51-56]、ReLU[47,57]，线性[49]或阶跃[39]函数。然而，在神经网络训练过程中，这些工作中提到的激活函数的形状是固定的，限制

了网络训练过程中权重的更新。由此，一个很直观的想法是，如果激活函数在训练过程中发生变化，神经网络的性能可以得到很大的提升。例如，训练学习刚开始时输出倾向于以更快地变化向期望得到的正确值逼近，因为开始时它们之间存在很大的差异，而随着差异值的减小，变化应该逐渐缓慢而微小。在这种情况下，激活函数应该在开始时有一个陡峭的斜率，然后逐渐变得平滑。此外，机器学习中使用的最重要的算法之一是批量归一化 (Batch Normalization, BN)[58]，其核心想法是在每次迭代后对输入进行重新归一化和偏移。它背后的思想是在神经网络的训练过程中，通过输入的偏移和重新归一化来修正输入分布和适合训练的理想分布之间的偏差。本文提出的可训练激活函数执行与批量归一化类似的作用，本文的工作指出了可以通过与自旋电子器件的物理特性相结合，基于自旋器件来硬件实现具有可训练激活函数的类脑计算网络。

本文主要内容是基于自旋电子器件实现类脑计算网络，通过探索自旋电子器件的物理特性，模拟实现高效的自旋神经元单元和自旋突触单元，并创新性地提出可训练激活函数执行与批量归一化算法类似的作用，讨论了基于自旋神经元单元和自旋突触单元来硬件实现具有可训练激活函数的自旋神经网络。最终实现的集成的自旋神经网络的性能展现出良好的可靠性、较高的推理识别准确率，较低的训练学习损失以及很低的单次写入功耗等。本论文的五个章节内容安排如下：

第一章：绪论。本章主要阐述了基于自旋电子器件实现类脑计算的研究背景和研究意义，整理了国内外关于类脑计算芯片的发展，还介绍了自旋电子器件实现类脑计算的优势。

第二章：自旋突触和自旋神经元背后的物理原理。本章主要阐述了自旋电子器件的存储和处理信息的物理原理，对应于自旋突触的权重更新和自旋神经元的激活功能的实现。自旋转移力矩、自旋轨道力矩和电压控制磁各向异性等效应均是通过只改变输入电信号进而改变自旋磁矩的原理，为实现自旋神经元激活功能和自旋突触写入权重的操作提供了理论基础。隧道磁阻效应也与自旋突触记忆功能息息相关。另外，本章还描述了神经元的非线性激活函数的概念和本文中仿真实验采用的宏自旋模型。

第三章：可训练激活函数的自旋神经元模型。本章提出的自旋神经元模型结合其背后的物理原理实现了可训练激活函数的功能，通过分析自旋神经元的器

件原理发现它可以实现可调节激活函数和可训练激活函数，为进一步研究该模型的性能，搭建了以电脉冲信号调控的三层自旋神经网络，验证了可训练激活函数的引入提升了神经网络的推理测试准确率。通过分析引入的可训练参数，提出了改进的可训练激活函数自旋神经元，并分析测试相应的改进的自旋神经网络的性能。

第四章：非理想特性对自旋神经网络的影响。基于上一章提出的自旋神经网络的理论基础，本章主要研究了具有可训练激活函数的自旋神经网络的实现。本章讨论了自旋突触单元的非理想特性对实现自旋神经网络的影响。接着，通过改变自旋神经元和自旋突触器件中 $CrTe_2$ 层厚度实现具有可调节激活函数的自旋神经网络。最后，通过输入电脉冲调控自旋神经元单元，实现了具有可训练激活函数的自旋神经网络。

第五章：总结与展望。本章回顾了前面具有可训练激活函数的自旋神经网络的内容，列举了自旋神经网络的改进点，展望了基于自旋电子器件实现高性能第三代脉冲神经网络。

第 2 章　自旋突触和自旋神经元背后的物理原理

2.1　神经元的激活函数

大多数人工神经网络都使用 McCulloch-Pitts 神经元模型[59]。在此模型中，神经元 j 的输出如式 2.1所示：

$$y_j = f(\sum_{i=0}^{N} w_{ij} x_i) \quad \ldots (2.1)$$

上式中的 y_j 表示输出值,f 对应于激活函数,N 是神经元 j 的输入个数。$w_{i,j}$ 是神经元 i 到神经元 j 的突触权重，x_i 是神经元 i 的输出值。

在这个神经元模型中，首先，神经元将通过突触整合前神经元的加权输出。接下来，这个线性组合的积分由神经元的激活函数处理，然后将输出发送到下一个神经元。激活函数在数据处理中起着关键作用。而激活函数的选择在很大程度上取决于神经网络的结构和输入数据的特征。不同的激活函数可以通过控制神经元器件的行为来实现。

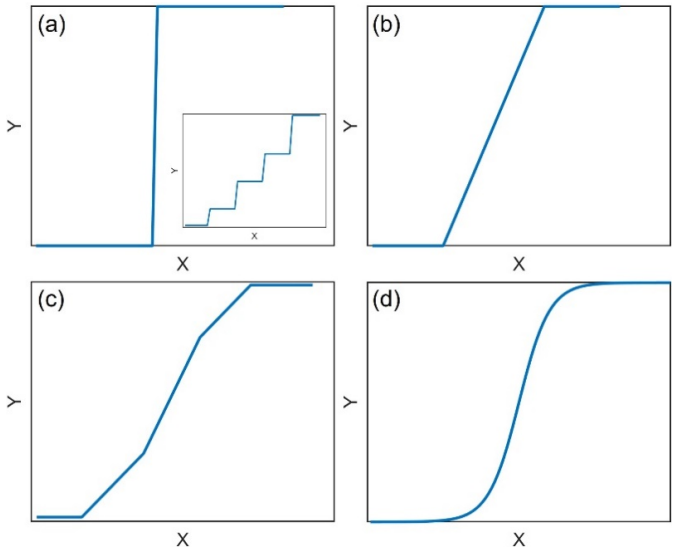

图 2.1　神经元激活函数图。(a) 阶跃函数。插图显示的是在阶跃函数的基础上改进的多级阶梯函数。(b) 斜坡饱和函数。(c) 分段线性函数。(d) 非线性激活函数。

Figure 2.1　Plots of activation functions for neurons. (a) Step function. The illustration shows an improved multi-level staircase function based on the step function.(b) Ramp-saturation function. (c) Piecewise linear function. (d) Nonlinear activation function.

人工神经元的发展过程中，传统的 McCulloch-Pitts 神经元模型有多种实现。通常用于硬件实现的由 CMOS 组成的感知器实现了一个简单的阈值函数[60]。如图 2.1(a) 所示，最简单的激活函数是阶跃函数[61]，此激活功能的神经元硬件需要较少的面积利用，并且不会是计算密集型的。类似地，还有图 (a) 中插图所示的多级阶梯函数。然而，主流的学习算法如反向传播算法都是基于梯度的。由于阶跃函数[39]不可微且不适用于该算法，其他基于硬件的激活函数，包括斜坡饱和函数[62]、线性[63]和分段线性[64]函数，如图 2.1(b) 和 2.1(c) 所示，已经实现以匹配基于梯度的学习算法。随着激活函数复杂性的增加，即从线性函数到非线性函数，学习过程的整体精度增加，如图 2.1(d) 所示的非线性激活函数，例如基本的 sigmoid 函数[65,46-50,18,51-56]、双曲正切函数[66]和 ReLU 函数[47,57]等，可以给出连续变化的导数，为基于梯度的学习算法提供高分辨率。

本文提出的自旋神经网络采用 sigmoid 型激活函数，S 型函数如图 2.1(d) 所示，其数学表达式如式 2.2所示。

$$f(x) = \frac{1}{1+e^{-x}} \qquad \ldots (2.2)$$

在深度学习领域，激活函数是非常关键的组成部分。传统的激活函数，如 Sigmoid 函数、双曲正切函数、ReLU 函数等，都是预先定义好的函数形式。这些激活函数都有它们各自的优点和局限性。例如，ReLU 函数简单且计算效率高，但是它在负数部分完全不激活，可能会导致神经元后续无法激活。Sigmoid 和双曲正切函数在输入值较大或较小的情况下，导数接近于 0，可能会导致梯度消失问题。

为了克服这些问题，研究人员开始尝试使用可训练的激活函数，即其形状可以在训练过程中学习和适应。以下是一些重要的可训练激活函数和相关的研究。

(1) Parametric ReLU (PReLU) 是一种广义的 ReLU 函数，其中负数部分的斜率是可学习的。这是首次尝试在激活函数中引入可训练参数的研究之一，这项研究在 2015 年由何恺明等人提出[67]。

(2) Scaled Exponential Linear Units (SELU) 是在 2017 年由 Klambauer 等人[68]提出的，其形式是在输入小于零时乘以一个可调节的参数。虽然不是所有的参数都是可学习的，但 SELU 是训练深度网络中的重要步骤，因为它自动地归一化输入。

(3) Swish[69]是由 Google 在 2017 年提出的一种自我门控的激活函数。Swish

函数的形式是，其中 β 是一个可训练的参数。这是一种真正意义上的可训练激活函数。

(4) Adaptive Piecewise Linear (APL)[70]是一种分段线性函数，其分段点和斜率都是可学习的参数。这使得 APL 能够以更灵活的方式适应数据。

以上就是一些可训练激活函数的关键研究。这些工作为后续更多的可训练激活函数研究铺平了道路。可训练激活函数之前并未应用于自旋电子器件，计算机科学领域提出并研究了可训练激活函数对神经网络的影响[71]。本文创新性地将可训练激活函数与自旋电子器件的物理特性相结合，从普遍了解的批量归一化算法的角度进一步研究了可训练激活函数对神经网络性能的提升。

2.2 改变自旋磁矩的方法

自旋电子器件是一种新型电子器件，它利用电子的自旋来存储和处理信息。与传统的电子器件不同，自旋电子器件不依赖于电荷来存储信息，而是利用电子的自旋磁矩来实现信息的存储和处理。对于磁性材料，磁化方向来自于相邻区域的交换耦合的电子自旋。从经典物理学的角度来看，磁化方向被认为是由自旋角动量控制的磁矩。由于磁化方向可控，基于磁性材料的器件可以实现信息存储、逻辑计算和其他新颖的功能。在早期阶段，磁化方向由磁场控制，外加磁场在应用上受到了严苛的要求，这种方式也大大增加了电路设计的难度。为了与电路兼容，只依靠改变输入电信号的方式，即利用电子的自旋来操纵磁性材料中的磁化方向的方式受到了广泛的关注。

2.2.1 自旋转移力矩效应

最先被提出的通过电信号改变磁化方向的方法即第 1章所提到的自旋转移力矩。自旋转移力矩[73]效应是通过传导电子的自旋提供有效的磁场，以交换自旋角动量。由于 STT 效应，磁性材料的自旋角动量可以通过注入的自旋极化电子流交换，该自旋极化电子流是由输入电流通过磁体产生的。更具体地说，STT 效应可以通过实验测量如图 2.2(a) 所示的三明治结构，该结构由两个具有大/小饱和磁化的磁性层 (分别称为参考/自由层) 夹着非磁性间隔物实现。被参考层过滤后，流过这个“过滤器”的电子获得了净极化自旋。接下来，当自旋极化电流通过非磁性间隔物并注入自由层时，一方面，只有平行于磁化方向的自旋极化分

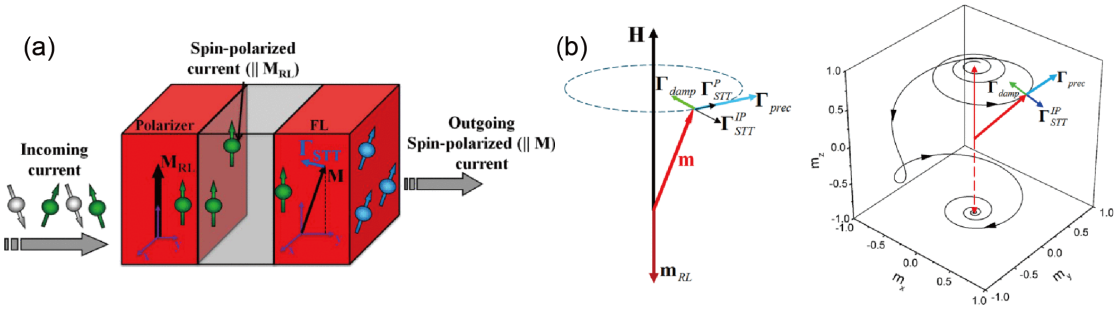

图 2.2　自旋转移力矩效应。(a) 自旋转移力矩的自由电子模型示意图。(b) STT 作用下的磁化翻转轨迹[72]。

Figure 2.2　Spin transfer torque effect.(a)Schematics of free electron model for spin transfer torque.(b)Illustration of STT effect acting on the magnetization and the resulting switching trajectory[72].

量透射且不改变；另一方面，垂直于磁化方向的分量将被吸收，从而由于角动量守恒导致磁化方向翻转。对于这个过程，被称为 STT 翻转磁化方向。也可以简单理解为，当电流通过磁性材料时，电子的自旋与周围的磁矩相互作用，从而产生一个如图 2.2(b) 所示的 STT。电流中的自旋极化电子流通过 STT 作用于磁矩上，可以使磁性材料中的磁矩发生翻转，从而实现数据的写入和存储。得益于低功耗、高速度、高稳定性等优点，STT 效应在磁存储器件和自旋电子学领域中得到了广泛的应用。

接下来，本论文将讨论图 2.2(b) 中 STT 是如何导致磁化方向翻转的。任何磁性层的磁化动力学都可以用朗道-利夫希兹-吉尔伯特 (Landau–Lifshitz–Gilbert, LLG) 方程来描述：

$$\begin{aligned} \frac{\mathrm{d}\mathbf{m}}{\mathrm{d}t} &= \boldsymbol{\Gamma}_{\mathrm{prec}} + \boldsymbol{\Gamma}_{\mathrm{damp}} \\ \boldsymbol{\Gamma}_{\mathrm{prec}} &= -\gamma\mu_0 \mathbf{m}\times\mathbf{H} \\ \boldsymbol{\Gamma}_{\mathrm{damp}} &= -\alpha\gamma\mu_0 \mathbf{m}\times(\mathbf{m}\times\mathbf{H}) \end{aligned} \qquad \ldots (2.3)$$

其中，$\mathbf{m} \equiv \mathbf{m}/M_{\mathrm{S}}$ 是沿着饱和磁化强度为 M_{S} 的磁性层的磁化强度 $\mathbf{m}$ 的归一化矢量，$\mathbf{H}$ 是总有效磁场 (包括各向异性和外加磁场)，α 是吉尔伯特阻尼，γ 是旋磁比，μ_0 是真空磁导率。$\boldsymbol{\Gamma}_{\mathrm{prec}}$ 描述了磁矩在有效磁场 $\mathbf{H}$ 周围的进动运动，而 $\boldsymbol{\Gamma}_{\mathrm{damp}}$ 描述了磁化进动运动朝向有效磁场 $\mathbf{H}$ 的逐渐衰减，从而降低了其总能量。当 STT

作用于磁性层时，LLG 方程中会出现两个附加项：

$$\begin{aligned}\frac{\mathrm{d}\mathbf{m}}{\mathrm{d}t} &= \Gamma_{\mathrm{prec}} + \Gamma_{\mathrm{damp}} + \Gamma_{STT}^{IP} + \Gamma_{STT}^{P} \\ \Gamma_{STT}^{IP} &= \gamma\mu_0\eta\frac{\hbar}{2}\frac{J}{e}\frac{1}{M_{\mathrm{S}}t_{\mathrm{FL}}}\mathbf{m}\times(\mathbf{m}\times\mathbf{m}_{\mathrm{RL}}) \\ \Gamma_{STT}^{P} &= \gamma\mu_0\eta'\frac{\hbar}{2}\frac{J}{e}\frac{1}{M_{\mathrm{S}}t_{\mathrm{FL}}}\mathbf{m}\times\mathbf{m}_{\mathrm{RL}}\end{aligned} \quad \ldots (2.4)$$

其中 J 是电流密度，t_{FL} 是自由层的厚度，e 是电子电荷，$\hbar$ 是约化普朗克常数，η 是自旋转移效率 (与电流自旋极化直接相关)。Γ_{STT}^{IP} 是位于由两个矢量 $\mathbf{m}$ 和 $\mathbf{m}_{\mathrm{RL}}$ 定义的平面内的转矩，因此通常称为面内转矩或类阻尼的转矩 (因为它具有与吉尔伯特阻尼相同的形式，但可以根据电流方向改变符号，因此可以表现为反阻尼转矩) 或 Slonczewski 转矩。Γ_{STT}^{P} 与该平面垂直，因此通常称为垂直旋转转矩或类场转矩。关于 STT 引起的磁化动力学，在最相关的轴对称情况下，当 $\mathbf{m}_{\mathrm{RL}} \parallel \mathbf{H}$ 时，垂直的 STT 具有与进动转矩项相同的作用，即其主要作用是改变磁化进动的频率。相反，平面内 STT 可以增强或减小阻尼转矩，如图 2.2(b) 所示。当面内 STT 与阻尼转矩相反并在数值上克服阻尼转矩时，进动的幅度增加到可以发生磁化方向翻转的点。

2.2.2 自旋轨道力矩效应

然而，通过 STT 控制高阻磁性材料的磁化方向是非常困难的。但磁化方向也可以通过自旋轨道力矩[74]翻转，此时注入电流不需要穿过三层磁性隧道结的结构。SOT 效应起源于自旋轨道耦合 (Spin Orbital Coupling, SOC)[75]，SOC 的原理可以归结为电场产生的有效磁场。由于磁性材料/重金属结构中的不对称自旋散射导致的反演对称性破缺，重金属层中可能会产生净自旋极化，这就是自旋霍尔效应 (Spin Hall Effect, SHE)[76]。然后，自旋极化的电子会聚集在材料的界面处，因此可以被相邻的铁磁体以类阻尼 SOT 的形式吸收。另一种物理解释归因于 Rashba 效应[77]，电子通过具有不对称反转的界面，因此获得自旋极化。极化的电子可以通过交换耦合在相邻的铁磁体上产生转矩。尽管 SHE 和 Rashba 效应在传统铁磁体/重金属异质结构中占主导地位，但这两者并不是 SOT 的唯一来源。其他效应，例如拓扑绝缘体中的量子自旋霍尔效应[78]，也可能产生 SOT，仍在深入研究中。

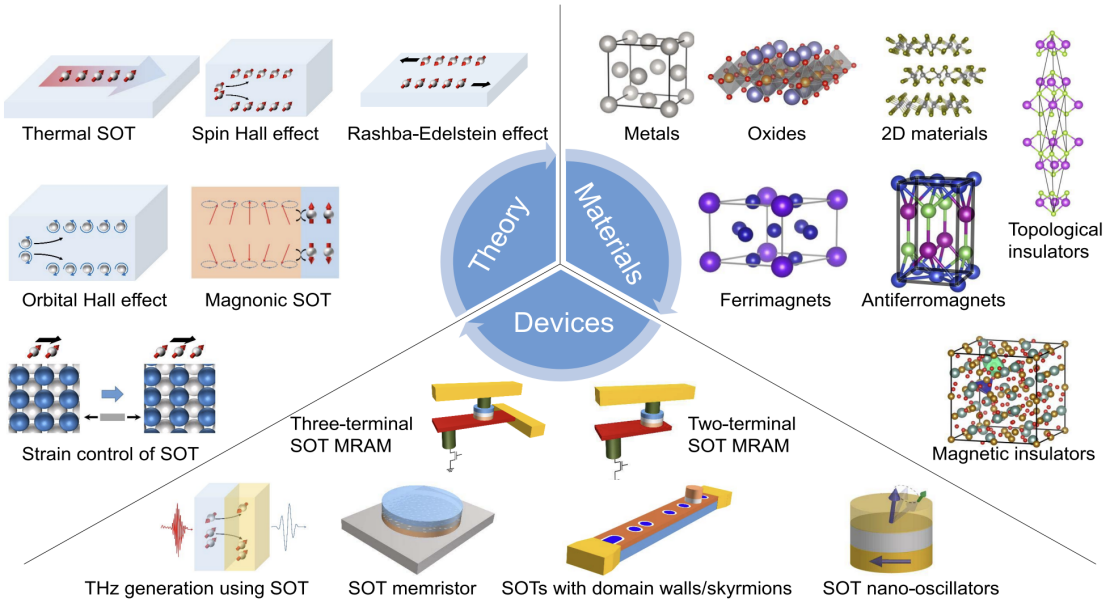

图 2.3　自旋轨道力矩效应的原理，材料和应用器件[79]。

Figure 2.3　Theory, materials and devices of spin orbital torque[79].

2.2.3　电压控制磁各向异性效应

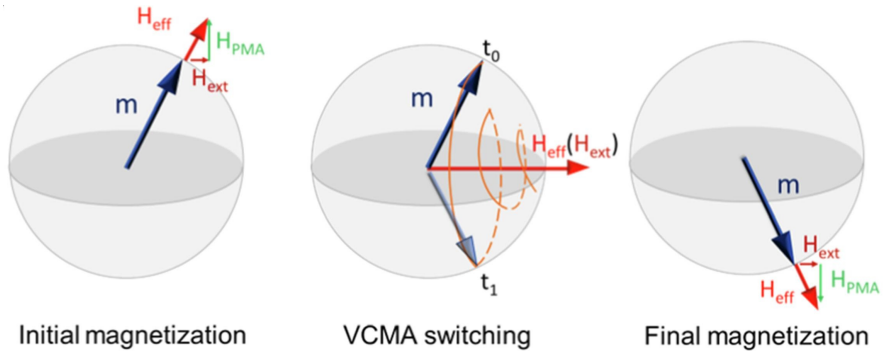

图 2.4　VCMA 动态切换示意图[80]。

Figure 2.4　Schematic of VCMA dynamic switching[80].

除了应用自旋力矩来操纵磁化方向外，还可以利用改变磁体的磁各向异性，从而控制磁化方向。这种效应是电压控制磁各向异性[81]，它是指在电场的控制下，磁性材料的磁各向异性能发生可逆调节的现象。在自旋电子器件中，VCMA 效应可以被用来控制磁性隧道结的磁各向异性，从而调节隧道结中的自旋极化电子的自旋取向。通过改变电场的大小和方向，可以调节自旋极化电子在磁性隧道结中的自旋取向，从而实现对自旋电子器件的高效控制。图 2.4展示了动态 VCMA 切换的示意图。施加的电压消除了垂直磁各向异性 (Perpendicular Magnetic

Anisotropy, PMA)，并导致沿外磁场方向的磁矩进动。在半周期进动后，撤掉电压，PMA 恢复，力矩将在相反方向上稳定。

VCMA 效应的机制与自旋轨道耦合和晶格畸变有关。在晶体结构不对称的材料中，电场可以通过改变晶格畸变的大小和方向来调节自旋轨道耦合和磁各向异性。通过在磁性材料上施加电场，可以控制晶格畸变和自旋轨道耦合，从而调节材料的磁各向异性。VCMA 效应在自旋电子器件中的应用具有广泛的潜力，例如可以用来实现高速低功耗的自旋电子逻辑器件和高密度的自旋电子存储器等。此外，VCMA 效应也为自旋电子器件的设计和优化提供了新的思路和方法。

2.3 隧道磁阻效应

过去二十年自旋电子学技术中的一个重要元素是磁性隧道结，它由两个铁磁层组成，由一层薄非磁性隧道层隔开，如图 2.2(a) 所示。MTJ 的阻值取决于自由层和参考层中磁化的相对方向。

隧道磁阻 (Tunnel Magnetoresistance，TMR) 效应的发现[82]是将基于自旋的器件与 CMOS 技术集成的里程碑之一。自旋电子器件中的 TMR 效应是基于自旋极化电子通过非磁性层的隧穿时发生的一种磁电阻效应。TMR 效应的基本原理是隧穿电流的自旋极化。当电流通过夹在两个磁性层之间的非磁性层时，由于量子隧穿效应，电流会透过非磁性层。在透过非磁性层的过程中，电子的自旋会被非磁性层中的晶格结构影响，从而使电子的自旋朝向更倾向于与其中一个磁性层的自旋方向相同，而与另一个磁性层的自旋方向相反。这样，透过隧道层的电流中就会具有一定的自旋极化度，即电子的自旋朝向在空间中偏向于一个特定的方向。当两个磁性层的自旋方向相同时，电流通过器件时电阻较低；而当两个磁性层的自旋方向相反时，电流通过器件时电阻较高。因此，当 MTJ 中的两个磁性层自旋方向不同时，就会产生 TMR 效应。

TMR 效应可以被广泛应用于磁性存储器、传感器和磁性逻辑器件等领域。TMR 效应是磁性随机存储器 (Magnetoresistive Random Access Memory，MRAM) 这类新型存储器的核心原理之一，因为它可以实现高速度、低功耗、高可靠性的数据读写操作。TMR 提供了非常大的磁阻比，因此它可以为 CMOS 放大器提供足够的信号强度[83]。

2.4　宏自旋模型

本文中自旋电子器件的仿真实验采用宏自旋 (Macrospin) 模型，它是研究磁性材料中磁矩变化的一种简化模型。在这种模型中，磁性材料被视为一个单一的宏自旋，而不是许多微观磁矩的集合。通常情况下，当铁磁体的直径小于 80nm 时，铁磁体的磁化强度在空间上均匀分布，因此可以将磁性材料视为一个大的自旋体。这种简化假设使得研究磁性材料的磁化动力学过程变得更加容易和直观。

本文的自旋电子器件的仿真对应于磁矩变化是由整体磁化翻转引起的情况。在宏自旋模型中，可以使用 Landau-Lifshitz-Gibert-Slonczewski(LLGS) 方程描述磁矩随时间的变化。本文通过原创的 MATLAB 代码求解 LLGS 方程，预测磁性材料在外部磁场、电流等作用下的磁化动力学行为。在 3.1.2中详细阐述了仿真采用的宏自旋模型求解 LLGS 方程式 3.1的过程。

通过求解 LLGS 方程得到自旋电子器件稳定状态下的磁矩。由于 LLGS 方程是微分方程，可以按照微分方程的求解方法计算。本文通过四阶龙格-库塔法 (Fourth-order Runge-Kutta method, RK4)[84]求解显式微分方程，其计算步骤如下：

$$\begin{cases} k_1 = f(y_t, t) \\ k_2 = f(y_t + \frac{h}{2}k_1, t + \frac{h}{2}) \\ k_3 = f(y_t + \frac{h}{2}k_2, t + \frac{h}{2}) \\ k_4 = f(y_t + hk_3, t + h) \\ y_{t+1} = y_t + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4) \end{cases} \quad \ldots (2.5)$$

2.5　本章小结

本章主要阐述了自旋突触和自旋神经元背后的物理原理。首先，明确了神经元的非线性激活函数的概念；其次，为实现自旋神经元的激活功能，研究了三种改变自旋转矩的方法，同时为自旋突触写入操作提供了理论支持；然后，自旋电子器件的隧道磁阻效应与自旋突触读取操作息息相关；最后，本章还描述了自旋电子器件仿真所用的数值模拟模型，宏自旋模型。

第 3 章　可训练激活函数的自旋神经元模型

自旋电子器件已被广泛应用于研究人工神经元的硬件实现。神经元作为大脑构成的基本单元之一，负责接收信息、处理信息和输出信息的任务。神经元的基本行为是当膜电位发生变化时积累电荷，一旦神经元的膜电位超过阈值电压时，它就会释放响应的输出脉冲。神经元通过突触互相连接成神经网络，每个神经元都会接收来自其他神经元的脉冲信号。当它触发至激活状态时，就会向连接的突触后神经元发出响应的脉冲信号。随着科学家对神经元行为认知的提高，神经网络现已发展到第三代。正如第一章提到的，神经网络从开始的只能处理线性可分问题的感知机，现已发展到由事件驱动的脉冲神经网络。本文将其中的神经元对输入脉冲响应至激活状态的行为进行建模，提出了基于磁性隧道结实现可训练激活函数的神经元模型。

在人工神经网络中，神经元的激活行为可以被抽象为非线性的激活函数。自旋力矩驱动的磁性隧道结可以用作随机开关器件，它通常作为随机神经元用于产生类似非线性 sigmoid 函数[46-50,18,51-56]的激活函数。然而，在人工神经网络的学习训练的过程中，这些工作中得到的激活函数的形状是固定的，这会对网络中突触的权重更新产生限制，进而导致整个神经网络的性能提升也受到限制。基于这一问题，本章利用自旋力矩诱导磁性隧道结发生磁化翻转，将其背后的物理原理与可训练激活函数的模型结合，使自旋神经元模型的激活函数能够在训练过程中发生动态变化。

例如，学习训练过程刚开始时，初始的输出值和期望值之间存在很大的差异，因此，希望得到的输出以更快地速度接近期望的正确值。而随着差异值的减小，输出值的变化应该逐渐缓慢，进行微小的变化。在这种情况下，对应的激活函数就应该在开始时是陡峭的，然后逐渐变得平滑。本文提出的可训练激活函数执行与批量归一化[58]类似的作用。批量归一化是机器学习中使用的最重要的算法之一，其核心想法是在每次迭代后对输入进行重新归一化和偏移缩放。它背后的思想是在神经网络的训练过程中，通过引入输入的偏移缩放和重新归一化来修正输入分布和适合训练的理想分布之间的偏差。因此，本文的工作指出了可以通过探索自旋电子器件的物理特性，基于自旋器件来硬件实现重要算法的可能性。

本章首先描述了发生磁化翻转的随机神经元的器件结构和翻转机制。其次，展示了该神经元的电学特性，阐述了该神经元模型可以根据输入电脉冲信号实现可调激活函数。之后，为进一步提升神经网络的性能，通过电脉冲信号控制脉冲宽度和磁各向异性来改变激活函数的斜率。激活函数斜率的改变会使得反向传播算法中梯度的更新能够快速或缓慢的改变。在第 3.5节详细描述了可训练激活函数的实现过程。最后，使用可训练激活函数的神经元搭建的三层自旋神经网络，在不引入额外能量消耗的情况下，对手写数字的识别准确度从 88% 提高到 91.3%。

3.1　自旋神经元的器件原理

本文提出的神经元模型利用自旋转移力矩切换磁性隧道结的状态，使其能够实现按照一定概率切换的随机神经元。本文发现该自旋神经元模型的翻转概率会根据输入脉冲信号的幅值响应出类似非线性 sigmoid 的激活函数的关系曲线。

3.1.1　自旋神经元的器件结构

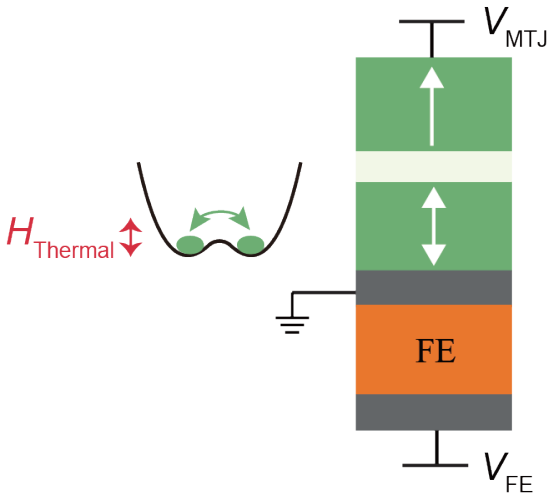

图 3.1　可训练激活函数的自旋神经元。

Figure 3.1　Illustration of the device structure.

因为激活函数需要在学习训练过程中进行动态变化，所以，本文提出的可训练激活函数的神经元模型必须只通过电脉冲信号调控。当使用自旋力矩来切换磁性隧道结的状态时，本文发现其翻转概率曲线的斜率可以通过输入的脉冲信号的脉冲宽度、脉冲幅度或磁各向异性来调节。由此，本文提出了如图 3.1所示

的器件结构，可训练激活函数的自旋神经元是由底部的铁电 (对应图中的 FE) 层和其顶部的磁性隧道结组成，铁电层用于控制自由层的各向异性[85]。本文选择广泛使用的 CoFeB 作为自由层。如图 3.1中左图所示，提出的可训练激活函数的自旋神经元的能量势垒很低，磁性隧道结只需要很低的输入电流就在平行态和反平行态之间进行切换。由于受到热扰动 $\mathbf{H}_{\text{thermal}}$ 的影响，磁性隧道结的翻转具有随机性，所以可训练激活函数的自旋神经元也遵循随机神经元的输出特性。

为了实现可训练激活函数，需要根据输入的脉冲信号响应出具有变化斜率的激活函数。图 3.1所示的器件结构中，在 FE 层上施加电压可以控制 CoFeB/MgO 叠层中的垂直各向异性[85]。由于磁各向异性决定了系统的能量势垒，磁各向异性越小，系统的能量势垒越小，对应的翻转电流就越小。通过电脉冲信号控制磁各向异性的操作，便可以控制翻转概率和输入电脉冲的 S 型关系曲线的偏移量。后面的章节会详细描述可训练激活函数的实现过程。

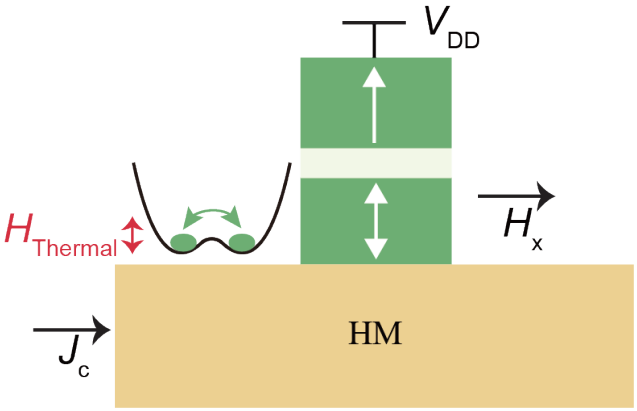

图 3.2　结合 SOT 和 VCMA 的三端磁性隧道结示意图。

Figure 3.2　Illustration of the three terminal device combining SOT and VCMA.

另外，由于可训练激活函数的自旋神经元模型具有普适性，也可以使用如图 3.2所示的器件结构实现功能。该磁性隧道结是一个结合了自旋轨道力矩和电压控制磁各向异性的三端结构。它的自由层的磁化方向可以利用外加场 $\mathbf{H}_x$ 辅助的自旋轨道力矩来翻转。与此同时，相比于图 3.1中利用铁电应变效应改变磁各向异性，图 3.2所示的三端磁性隧道结可以通过电压控制磁各向异性[86-100]在磁性隧道结上施加电脉冲来修改磁各向异性[86]。

正如本节开头提到的，为保证激活函数在学习训练的过程中能够动态变化，本文的神经元模型只通过电脉冲信号调控。为了该模型能够更好与电路集成，不

需要引入额外的外加磁场, 接下来的内容都是基于图 3.1所示的器件开展的。考虑到内容的连贯性, 关于铁电应变效应改变磁各向异性的内容, 可以类比于 2.2.3节中通过电压控制磁各向异性翻转磁性隧道结的磁化方向进行理解，在本章中这两种方法在电脉冲调控神经元模型的激活函数中起到一样的作用。

3.1.2　自旋神经元器件的翻转机制

当施加如图 3.1所示的电压脉冲 (V_{MTJ}) 或电流脉冲 (I_{c})，通过自旋转移力矩, 磁性隧道结的磁化状态会发生翻转。在宏自旋模型中，可以通过求解朗道-利夫希兹-吉尔伯特-斯隆切夫斯基 (LLGS) 方程[101-104]来解释翻转的过程:

$$d\mathbf{m}/dt = -\gamma\mathbf{m}\times\mathbf{H}_{\mathrm{eff}} + \alpha\mathbf{m}\times d\mathbf{m}/dt - \gamma\hbar J_{\mathrm{c}}/(2et_{\mathrm{FL}}M_{\mathrm{S}})\mathbf{m}\times(\mathbf{m}\times\boldsymbol{\sigma}_{\mathrm{STT}}) \quad \ldots (3.1)$$

上式中阻尼常数 $\alpha = 0.0122$，旋磁比 $\gamma = 1.76\times10^{11}\mathrm{rad/(s\cdot T)}$，约化普朗克常数 $\hbar = 6.58\times10^{-16}\mathrm{eV\cdot s}$，电子电荷量 $e = 1.6\times10^{-19}\mathrm{C}$，自由层厚度 t_{FL} = 1.3 nm 和饱和磁化强度 M_{S}=1.58 T[105-106]。

$\mathbf{H}_{\mathrm{eff}}$ 和 $\boldsymbol{\sigma}_{\mathrm{STT}}$ 分别表示有效磁场和自旋极化。$\boldsymbol{\sigma}_{\mathrm{STT}}$ 的方向取决于 V_{MTJ} 的极性，即正值和负值的 V_{MTJ} 分别导致沿 $-\mathbf{z}$ 和 $+\mathbf{z}$ 方向的 $\boldsymbol{\sigma}_{\mathrm{STT}}$。$\mathbf{H}_{\mathrm{eff}}$ 由晶体各向异性场 $\mathbf{H}_{\mathrm{an}}$、退磁场 $\mathbf{H}_{\mathrm{demag}}$ 和热扰动场 $\mathbf{H}_{\mathrm{thermal}}$ 组成。

晶体各向异性场 $\mathbf{H}_{\mathrm{an}}$ 如式 3.2所示:

$$\mathbf{H}_{\mathrm{an}} = 2(K_{\mathrm{bulk}} + K_{\mathrm{i}}/t_{\mathrm{FL}})m_z\hat{\mathbf{z}}/M_{\mathrm{S}} \quad \ldots (3.2)$$

上式中的体各向异性 $K_{\mathrm{bulk}} = 2.245\times10^{5}\mathrm{J/m^3}$，界面各向异性 $K_{\mathrm{i}} = 1.286\times10^{-3}\mathrm{J/m^2}$[105]。

如图 3.3所示，本文已经证实磁各向异性 K_{u} 足以克服形状各向异性，从而使自由层的磁化方向最终平衡后对齐 $\mathbf{z}$ 轴。

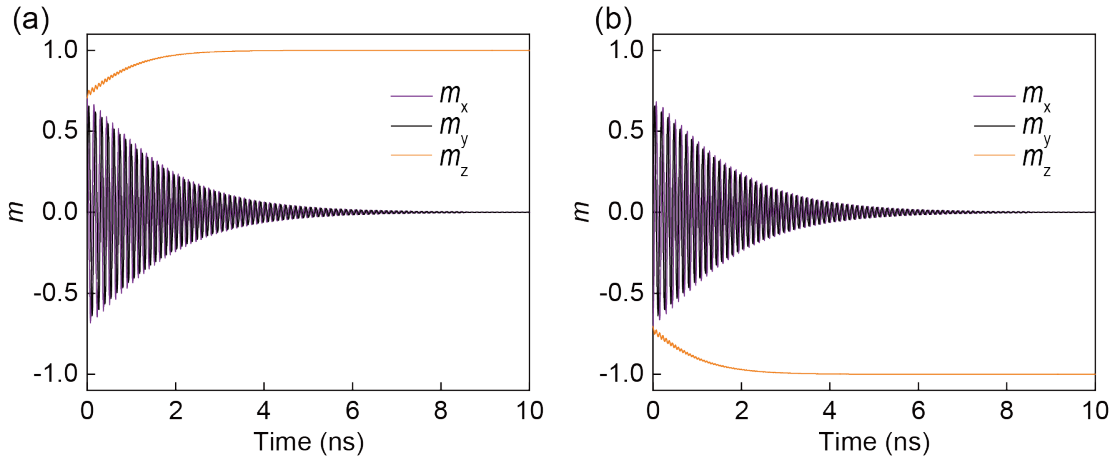

图 3.3　自由层磁化方向弛豫后对齐 z 轴。(a) 自由层磁化方向的初始状态与 +z 方向呈 45 度夹角，最终弛豫至 +z 方向。(b) 自由层磁化方向的初始状态与 +z 方向呈 135 度夹角，最终弛豫至-z 方向。

Figure 3.3　Illustration of relaxation of the magnetization of the free layer.(a)The initial state of the free layer magnetization is at an angle of 45° to the +z direction, and it eventually relaxes to the +z direction.(a)The initial state of the free layer magnetization is at an angle of 135° to the +z direction, and it eventually relaxes to the -z direction.

退磁场 $\mathbf{H}_{\text{demag}}$ 如式 3.3所示:

$$\mathbf{H}_{\text{demag}} = -(N_x m_x \hat{\mathbf{x}} + N_y m_y \hat{\mathbf{y}} + N_z m_z \hat{\mathbf{z}}) \quad \ldots (3.3)$$

上式中的退磁张量 (N_x, N_y 和 N_z) 是基于样品的几何形状计算得出的，图 3.1所示器件自由层的横截面是长度和宽度均为 15nm 的矩形。

热扰动场 $\mathbf{H}_{\text{thermal}}$ 是一个随机场，如式 3.4所示:

$$\mathbf{H}_{\text{thermal}} = N_1(0,u)\hat{\boldsymbol{x}} + N_2(0,u)\hat{\boldsymbol{y}} + N_3(0,u)\hat{\boldsymbol{z}} \quad \ldots (3.4)$$

上式中的 $N_i(0,u)$ 是均值为零，标准差为 $u = \sqrt{2k_{\text{B}}T\alpha/(V_{\text{FL}}M_{\text{S}}\gamma(1+\alpha^2)\Delta t)}$ 的符合正态分布的随机变量。其中玻尔兹曼常数 $k_{\text{B}} = 1.38 \times 10^{-23}$ J/K，温度 T=300 K，V_{FL} 表示磁性隧道结的自由层体积，采用宏自旋模型进行仿真，仿真的时间步长设置为 Δt=5 ps。

根据上述公式利用宏自旋模型求解 LLGS 方程。首先，将输入电脉冲信号 I_{c} 转换为时序的输入电流密度 J_{c} 并代入式 3.1。接着，通过将上述的式 3.2、式 3.3和式 3.4代入式 3.1中，就可以计算得到当前时刻磁性隧道结自由层的磁矩。然后，

按照仿真步长 Δt=5 ps 进行迭代计算，上一时间步长计算得到的状态作为下一时间步长的初始状态。最后，根据自由层最终稳定的磁化方向，就可以得到给定输入电脉冲条件下磁性隧道结的翻转情况。由于式 3.4所示的 $\mathbf{H}_{\text{thermal}}$ 具有随机性，磁性隧道结的磁化方向会发生随机翻转。但由于 $\mathbf{H}_{\text{thermal}}$ 的随机性有限，随着输入电脉冲信号的增强，磁性隧道结的翻转概率呈现出升高的趋势。

3.1.3　自旋神经元模型的仿真结果分析

为了探究磁性隧道结的磁化方向的随机翻转和输入电脉冲之间的关系，本文仿真了 100 次具有相同电脉冲输入的磁性隧道结的随机翻转过程,结果如图 3.4所示。

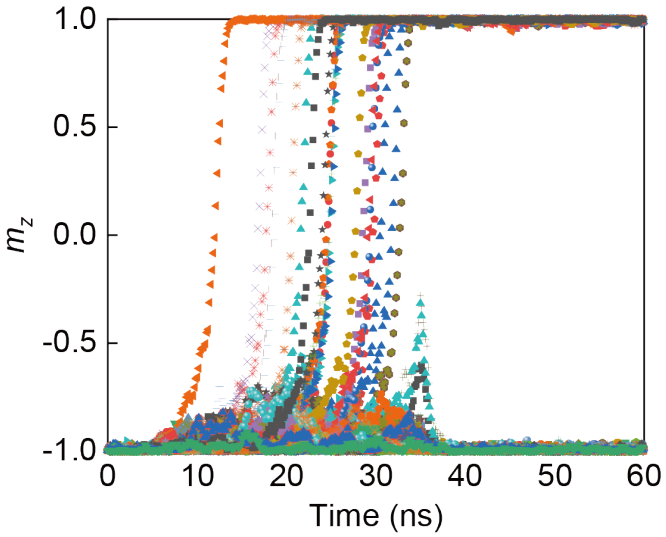

图 3.4　在 $I_c = 11.3\mu$A 和 30 ns 脉冲宽度下独立执行 100 次的翻转轨迹。初始状态 $m_z = -1$，最终状态 $m_z = 1$ 表示自由层的磁化方向翻转成功。

Figure 3.4　The switching trajectories of 100 independent runs at $I_c = 11.3\mu$Aand 30 ns pulse width. The initial state is $m_z = -1$. The final state with $m_z = 1$ indicates successful magnetization switching.

输入电脉冲的脉冲幅度是 $I_c = 11.3\mu$A，脉冲宽度是 30 ns。同一磁性隧道结自由层磁化方向的初始状态均为 $m_z = -1$，在相同的输入电脉冲条件下，该磁性隧道结执行 100 次独立的翻转。当磁性隧道结自由层磁化方向的最终状态为 $m_z = 1$ 时，表示磁化方向发生翻转，磁性隧道结由反平行态切换为平行态。从图 3.4中可以看出，尽管输入电脉冲的脉冲幅度和脉冲宽度完全相同，一部分的磁性隧道结切换至平行态，而仍然有一部分磁性隧道结没有在输入电脉冲的刺

激下发生翻转，经过一段时间的弛豫返回到初始的反平行状态。这样的随机翻转的结果是由随机的热扰动场 $\mathbf{H}_{\text{thermal}}$ 引起的，这和前面的物理原理的解释也是对应的。

虽然磁性隧道结的磁化方向会发生随机翻转，但是自由层磁化方向的翻转概率 P_{sw}[107-120]与输入电脉冲信号 I_{c} 有关。每个输入电脉冲 I_{c} 对应的翻转概率 P_{sw} 都是通过仿真该 I_{c} 输入下磁性隧道结的磁化动力学 100 次，将 100 次独立结果取平均值获得的。图 3.4中 100 次独立结果得到的翻转概率值对应于图 3.5中的 $I_{\text{c}} = 11.3\mu\text{A}$ 时，脉冲宽度是 30 ns 的一个蓝色实心圆点。

此外，我们将仿真中获得的翻转电流与使用解析解获得的翻转电流[121]进行比较。从图 3.5可以看出，翻转电流在 $10\mu\text{A}$ 左右。仿真中使用的参数如下表所示：

表 3.1　仿真参数

Table 3.1　Simulation parameters

参数名称	符号	值
阻尼常数	α	0.0122
自由层体积	V_{FL}	$15 \times 15 \times 1.3 \times 10^{-27}\text{m}^3$
饱和磁化强度	M_{S}	$1.58/(4\pi) \times 10^7\text{A/m}$
磁各向异性	K_{u}	$1.1 \times 10^6\text{J/m}^3$
自旋极化率	P	0.4

根据表 3.1中的参数值，我们可以计算出各向异性场 $\mathbf{H}_{\text{k}}$：

$$\mathbf{H}_{\text{k}} = 2K_{\text{u}}/M_{\text{S}} - 4\pi M_{\text{S}} = 0.169\text{T} \quad \ldots (3.5)$$

因此，我们可以得到解析解中的临界电流 I_{crit} 为

$$I_{\text{crit}} = 2\alpha e M_{\text{S}} V_{\text{FL}} \mathbf{H}_k/(\hbar P) = 5.8\mu\text{A} \quad \ldots (3.6)$$

临界电流 I_{crit} 的值 $5.8\mu\text{A}$ 与仿真得到的 $10\mu\text{A}$ 左右翻转电流非常接近。它们之间的差值是因为解析解得到的临界电流 I_{crit} 是一个粗略的估计，因为热扰动会导致 I_{crit} 降低，而在仿真中使用脉冲宽度较小的输入电脉冲时需要更高的过

驱动电流。此外，自旋力矩效率、形状各向异性和晶体各向异性也在磁化方向发生翻转期间进行动态变化[101]。

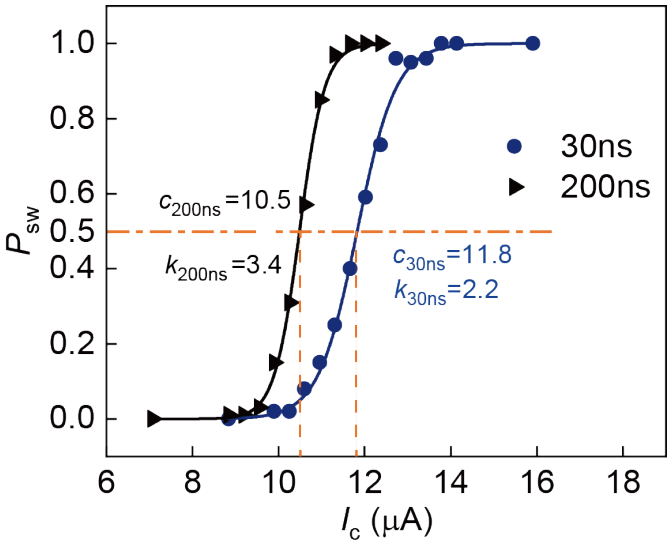

图 3.5　磁性隧道结翻转概率 P_{sw} 与输入电脉冲 I_c 的关系曲线呈 S 型，不同脉冲宽度下的曲线斜率 k 和偏移量 c 的比较。

Figure 3.5　The relationship curve of P_{sw} and I_c is S-shaped. Illustration of k and c at different pulse widths.

正如图 3.5中的蓝线所示，当输入电脉冲的脉冲宽度为 30 ns 时，随着输入脉冲电流值的增加，磁性隧道结的翻转概率 P_{sw} 呈现出 S 型上升的趋势。在 2.1节中提到 sigmoid 激活函数的表达式为 $f(x) = 1/(1 + exp(-x))$。类似地，图 3.5中不同脉冲宽度下的翻转概率 P_{sw} 可以用 $f = 1/(1 + exp(-k(I_c + c)))$ 形式的 S 形函数进行描述。对比脉冲宽度 200 ns 下的黑色 S 型曲线和 30 ns 下的蓝色 S 型曲线，可以发现当输入脉冲的电流值发生变化时，翻转概率曲线的斜率 k 和偏移 c 会发生变化。本文将 k 定义为翻转概率 $P_{sw} = 0.5$ 时的斜率。可以看出，与 30 ns 脉冲时的 $k = 2.2$ 相比，200 ns 的较大脉冲宽度可提供更大的 $k = 3.4$。类似地，本文将 c 定义为 $P_{sw} = 0.5$ 时相对于原点的电流偏移。与 30 ns 脉冲时的 $c = 11.8$ 相比，200 ns 的较大脉冲宽度可提供更大的 $c = 10.5$，可以发现，较大的脉冲宽度会导致较小的电流偏移。因此，通过控制输入电脉冲的脉冲宽度，此磁性隧道结可用作具有可调节 sigmoid 激活函数的神经元。

3.2 可调节激活函数的自旋神经元

本文测试了自旋神经元模型在不同输入电脉冲下的翻转概率。根据表 3.1的参数值，将 $\mu_0 = 1.2567 \times 10^{-6}$T·m/A 代入式 3.7, 可以计算得到神经元模型的能量势垒 E_B:

$$E_B = (K_u - 0.5\mu_0 M_S^2) \times V_{FL} = 7.5 k_B T \quad \ldots (3.7)$$

当输入电脉冲的脉冲宽度从 10 ns 逐渐增加到 200 ns 的条件下，我们得到了如图 3.6所示的翻转概率与输入电脉冲之间的关系。图中的符号点和线分别表示仿真结果和拟合曲线。将图中每一条翻转概率 P_{sw} 与输入脉冲电流 I_c 的关系曲线和 sigmoid 函数表达式相拟合，即把图 3.6中的 S 型曲线表达为函数 $f = 1/(1 + exp(-k(I_c + c)))$ 的形式，就得到了每条 S 型曲线对应的斜率 k 和偏移 c 的参数值。

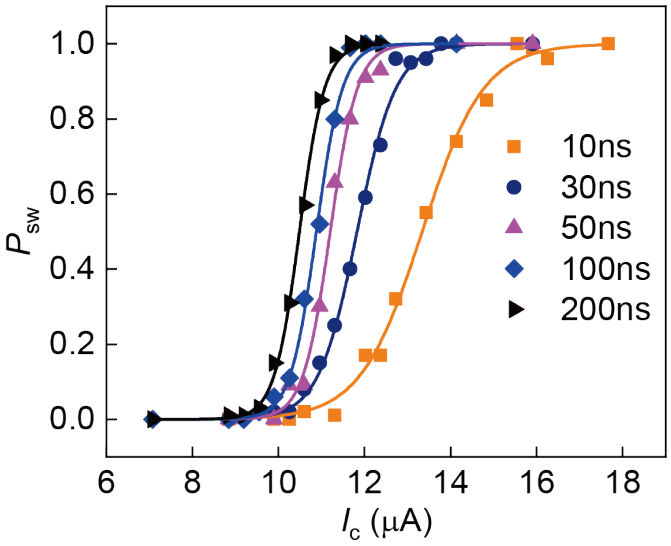

图 3.6 不同脉冲宽度下的磁性隧道结翻转概率 P_{sw} 与输入电脉冲 I_c 的关系曲线图。仿真结果和拟合曲线分别用符号和线表示。

Figure 3.6 P_{sw} as a function of I_c under different pulse widths. The simulation results and fitting are denoted by the symbols and lines, respectively.

从图 3.6中可以看出，随着输入电脉冲的脉冲宽度逐渐增加，翻转概率的 S 型曲线的偏移 c 在逐渐减小。这是因为当输入电脉冲具有较大的脉冲宽度时，翻转磁性隧道结自由层磁化方向所需的翻转电流会减小，对应的 S 型曲线的偏移 c 就会减小，对比相同电流幅度的输入电脉冲会具有更大的概率翻转自由层的磁

化方向。所以，可以通过改变输入电脉冲的脉冲宽度调节自旋神经元的激活函数的偏移 c。

此外，当输入电脉冲的脉冲宽度较大时，会有更多的电子被注入到磁性隧道结的自由层中，导致自旋转移力矩增强，进而导致自由层的磁化方向能够更快地翻转，翻转概率的 S 型曲线会更加陡峭，即 S 型曲线对应的斜率 k 的值也较大。因此，可以通过改变输入电脉冲的脉冲宽度调节自旋神经元的激活函数的斜率 k。

综上，本文提出的自旋神经元模型可以通过控制输入电脉冲的脉冲宽度实现可调节激活函数。类似地，本文仿真验证了也可以通过控制输入电脉冲的脉冲幅度实现可调节激活函数，仿真结果与图 3.6所示的改变脉冲宽度的结果相似，因为改变脉冲宽度或改变脉冲幅度背后的物理原理也大致相同。

为进一步提升自旋神经元模型在神经网络中发挥的作用，我们想到可以将实现的 S 型激活函数的斜率 k 和偏移 c 作为两个额外的自由度添加到神经网络的学习训练过程中。将斜率 k 和偏移 c 作为可训练的参数，使得自旋神经元模型可以实现可训练激活函数。

3.3　可训练激活函数的自旋神经元

根据上一节的结果，直观地认为通过控制额外的自由度斜率 k 和偏移 c 可以提高神经网络的性能。这背后的思想类似于机器学习中广泛使用的批量归一化算法。

批量归一化的算法实现是在每次迭代的过程中将神经元的输入在激活前进行重新归一化和偏移缩放，上述操作可以将神经元的输入限制在高斯范围内进行激活，可以避免神经元在饱和区激活导致的梯度消失或梯度爆炸。通过批量归一化可以拓宽超参数的选择范围，使得神经网络的性能在面对超参数的选择问题上表现地更加稳定。在本文提出的自旋神经元模型中，神经元的输入对应于电脉冲 I_{c}，输入 I_{c} 的偏移操作可以由可训练参数 c 完成，可训练参数 k 可以实现缩放输入的操作，它们类似于批量归一化算法中的可训练参数偏置因子 β 和缩放因子 γ。可训练参数 γ 和 β 在批量归一化算法中的作用是动态调整输入的高斯分布的均值和方差，使批量归一化后的输入与后面的激活过程更匹配，归一化操作的均值和方差不是简单的 0 和 1。

但与批量归一化的算法实现不同的是，斜率 k 和偏移 c 的控制可以和激活函数的实现集合在一个自旋神经元中，不需要在神经网络中增加额外的批量归一化层。此外，由于斜率 k 和偏移 c 都可以通过改变输入电脉冲 I_c 的脉冲宽度进行控制，因此它们可以用作神经网络的可训练参数，并在训练学习过程中进行动态更新。

3.4 三层自旋神经网络

为了评估提出的自旋神经元的性能，本文构建了一个简单的三层神经网络来对 MNIST 数据集[122]中的手写数字执行分类推理任务。MNIST 数据集由 50000 个训练样本和 10000 个测试样本组成。

3.4.1 三层自旋神经网络的搭建

如图 3.7所示的三层自旋神经网络，输入层有 784 个神经元，对应左边手写数字图片中的像素数。三层神经网络的第二层和第三层中的神经元都是前面讨论的的自旋神经元，隐藏层和输出层分别有 25 个和 10 个自旋神经元，它们的激活函数均为 sigmoid 类型。其中，输出层的 10 个自旋神经元对应于手写数字 0~9 共 10 个类别。

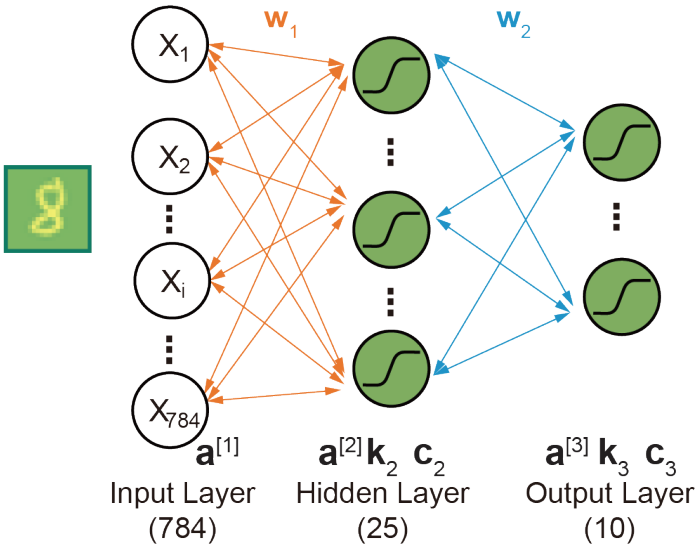

图 3.7　三层自旋神经网络。

Figure 3.7　A three-layer spin neural network.

图 3.7中 $\mathbf{w}_1$ 和 $\mathbf{w}_2$ 表示不同层之间神经元的忆阻器连接，对应于连接神经元

之间的突触。忆阻器的阻值或电导值表示突触的权重值。$\mathbf{a}^{[1]}$,$\mathbf{a}^{[2]}$ 和 $\mathbf{a}^{[3]}$ 分别表示输入层，隐藏层和输出层的神经元的输出响应。$\mathbf{k}_2$ 和 $\mathbf{c}_2$ 分别表示隐藏层的自旋神经元激活函数的斜率和偏移，类似地，$\mathbf{k}_3$ 和 $\mathbf{c}_3$ 分别表示输出层的自旋神经元激活函数的斜率和偏移。

虽然，突触的权重 w 与自旋神经元激活函数的斜率 k 和偏移 c 都是根据梯度下降算法更新的可训练参数。但是，它们在物理实现上是不同的。k 和 c 是局部参数，可以通过给自旋神经元施加输入电脉冲来控制。这与权重不同，权重是表示不同层神经元之间连接强度的非局部参数，可以使用如图 3.8所示的交叉电阻阵列来实现。

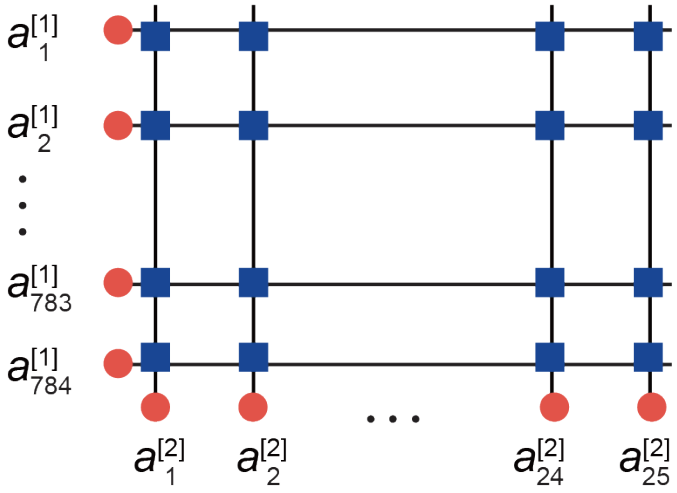

图 3.8　输入层和隐藏层之间的交叉电阻阵列。

Figure 3.8　Illustration of the crossbar resistive array.

图 3.8中蓝色方块的电阻表示权重，红色圆圈表示神经元。最左一列和最下一行的红色圆圈分别表示输入层和隐藏层中的神经元。图 3.7中输入层的每个神经元都连接到隐藏层的所有神经元，权重 $\mathbf{w}_1$ 代表了它们之间的连接强度，所以权重 $\mathbf{w}_1$ 是一个 25×784 的矩阵。从图 3.8中也可以看出，蓝色方块的阵列的大小也是 25×784。因此，每个蓝色方块都可以用来实现输入层和隐藏层之间神经元的连接强度，具体可以使用诸如磁性随机存储器或阻变随机存储器 (Resistive Random Access Memory, RRAM) 之类的忆阻器。对应地，输入层的输入是电流，电流通过交叉电阻阵列，然后产生作用于隐藏层中神经元的电压。

除了物理实现方面的不同，在数值表示方面，权重的维数与神经网络的拓扑结构有关。例如，$\mathbf{w}_1$ 的维度为 25×784，因为它连接输入层 (784 个神经元) 和隐

藏层 (25 个神经元)。相比之下，表示隐藏层中激活函数斜率的 $\mathbf{k}_2$ 的维度为 25×1，与局部神经元的维度相同。

将可训练参数 k 和 c 添加到神经网络后，本文推导了可训练激活函数的反向传播算法，随后 k 和 c 的值在训练过程中进行动态更新。

3.4.2 三层自旋神经网络的算法推导

本文提出了一种具有可训练激活函数的自旋神经元模型，可训练斜率 k 和偏移 c 的 sigmoid 激活函数是通过输入电脉冲控制自旋神经元随机翻转获得的。由于之前的激活函数在神经网络的学习训练过程中斜率和偏移都是固定不变的，所以传统的算法只支持权重 w 的更新。虽然，加入批量归一化层的神经网络的训练算法会引入偏置因子 β 和缩放因子 γ 的更新。但是，由于本文得到的可训练参数 k 和 c 是局部参数，由自旋神经元的行为主导。因此，我们开发了针对提出的可训练激活函数的算法，使 w、k 和 c 遵循该算法在训练学习过程中动态更新，从而提升训练学习的速度和提高识别推理的准确度。

在上一节中提出了如图 3.7所示搭建的自旋神经网络，由输入层、隐藏层和输出层组成。$\mathbf{a}^{[1]}$，$\mathbf{a}^{[2]}$ 和 $\mathbf{a}^{[3]}$ 分别对应输入层，隐藏层和输出层的神经元的输出响应。训练学习的目的就是将给定训练集的 50000 个数字图片转化为对应的像素组输入神经网络，在输出层获得的识别数字和输入的手写数字图片一一对应，即使得图 3.7中神经网络的输出 $\mathbf{a}^{[3]}$ 等于输入的手写数字 8。当训练集中 50000 个数字图片都识别正确后，就认为本文的自旋神经网络已经学习到识别手写数字的方法。接着，为测试完成训练学习的自旋神经网络的性能，本文将与训练集不同的测试集数据输入神经网络，检查神经网络对新的 10000 个手写数字的推理能力。最后，将神经网络对测试集 10000 个数字图片的推理结果与正确值比较，就得到了训练学习后的神经网络对测试集数据进行推理的准确度。

训练学习过程的算法由两部分组成，分别是前向传播和反向传播。

前向传播是将手写数字图片的像素组作为输入层 784 个神经元的输入 X，通过突触逐层传递信号至输出层，然后将输出层的输出 $\mathbf{a}^{[3]}$ 作为神经网络的预测值 h_θ，最后计算预测值 h_θ 与正确值 y 之间的偏差。神经网络通过训练能够学习到识别手写数字图片的方法，要以高准确度识别训练集的数据，即将神经网络预测值 h_θ 与正确值 y 之间的偏差降到很低。传统的神经网络定义了损失函数 J 来反

映预测值 h_θ 与正确值 y 之间的偏差:

$$J = -[ylog\ h_\theta + (1-y)log\ (1-h_\theta)] \quad \ldots (3.8)$$

其中 h_θ 是神经网络通过前向传播得到预测值，y 是正确的期望输出值, 即对应三层神经网络图 3.7中左侧手写数字图的 $y = 8$。

上述的前向传播对应于流程图 3.9中由 $\mathbf{a}^{[1]}$ 传导至 J 的过程。流程图 3.9中部分符号定义如下表所示:

表 3.2　算法流程图中部分符号的定义

Table 3.2　Definition of some symbols in the algorithm flowchart

符号	定义
X	自旋神经网络的输入
$n^{[l]}$	l 层中的神经元数量
$a^{[l]}$	l 层中神经元的激活向量，维度大小为 $(n^{[l]} \times 1)$
$z^{[l]}$	l 层神经元的加权输出向量，维度大小为 $(n^{[l]} \times 1)$
$w^{[l-1]}$	l 层与 $l-1$ 层相关联的权重矩阵，维度大小为 $(n^{[l]} \times n^{[l-1]})$
$g^{[l]}$	l 层神经元输出的激活函数，$a^{[l]} = g^{[l]}(z^{[l]})$
$k^{[l]}$	l 层神经元的激活函数的斜率向量，维度大小为 $(n^{[l]} \times 1)$
$c^{[l]}$	l 层神经元的激活函数的偏移向量，维度大小为 $(n^{[l]} \times 1)$
J	自旋神经网络的损失函数

在输入层，神经元的输出 $a^{[1]} = X$。

在隐藏层，自旋神经元的加权输出 $z^{[2]} = w^{[1]}a^{[1]}$，自旋神经元的激活输出 $a^{[2]} = g^{[2]}(k^{[2]} \cdot (z^{[2]} + c^{[2]}))$。其中，$z^{[2]}$ 表示输入层神经元的输出 $a^{[1]}$ 和突触权重 $w^{[1]}$ 的加权和,$g^{[2]}$ 表示具有可训练参数 $k^{[2]}$ 和 $c^{[2]}$ 的自旋神经元的可训练激活函数。$a^{[2]}$ 表示自旋神经元的输出。

在输出层，自旋神经元的加权输出 $z^{[3]} = w^{[2]}a^{[2]}$，自旋神经元的激活输出 $a^{[3]} = g^{[3]}(k^{[3]} \cdot (z^{[3]} + c^{[3]}))$。其中 $z^{[3]}$ 表示隐藏层神经元的输出 $a^{[2]}$ 和突触权重 $w^{[2]}$ 的加权和,$g^{[3]}$ 表示具有可训练参数 $k^{[3]}$ 和 $c^{[3]}$ 的神经元的自旋神经元的可训练激活函数。$a^{[3]}$ 表示自旋神经元的输出。

然后，神经网络通过前向传播得到预测值 $h_\theta = a^{[3]}$，将 h_θ 作为输入代入损失函数 $J = -[ylog\ h_\theta + (1-y)log\ (1-h_\theta)]$ 中，得到 h_θ 和期望的正确值 y 之间的偏差。

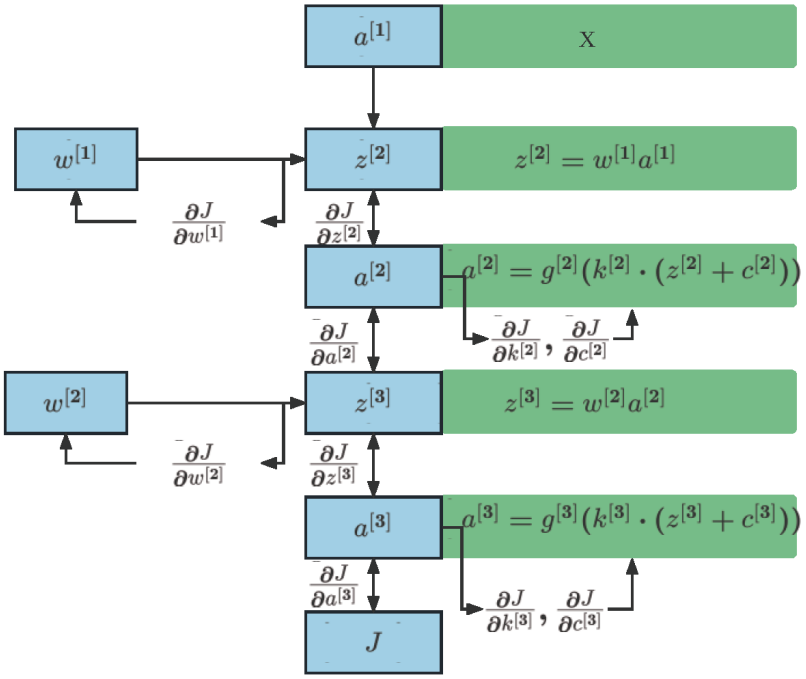

图 3.9　具有可训练的 k 和 c 的神经网络的算法流程图。

Figure 3.9　A flow chart describing the calculation performed in the neural network with trainable k and c.

反向传播是通过梯度下降算法来动态更新神经网络中的可训练参数，最终使得神经网络输出的预测值 h_θ 逼近期望的正确值 y，即通过找到 $w^{[1]}$, $w^{[2]}$, $k^{[2]}$, $k^{[3]}$, $c^{[2]}$ 和 $c^{[3]}$ 的最优值来最小化损失函数 J。这些参数的更新如式 3.9所示。

$$\begin{cases} w^{[2]} = w^{[2]} - \alpha \frac{\partial J}{\partial w^{[2]}} \\ k^{[3]} = k^{[3]} - \alpha \frac{\partial J}{\partial k^{[3]}} \\ c^{[3]} = c^{[3]} - \alpha \frac{\partial J}{\partial c^{[3]}} \\ w^{[1]} = w^{[1]} - \alpha \frac{\partial J}{\partial w^{[1]}} \\ k^{[2]} = k^{[2]} - \alpha \frac{\partial J}{\partial k^{[2]}} \\ c^{[2]} = c^{[2]} - \alpha \frac{\partial J}{\partial c^{[2]}} \end{cases} \quad \ldots (3.9)$$

其中，α 表示学习率，w、k 和 c 的更新值是上一次迭代过程中的参数值减去学习

率和梯度的积。

w、k 和 c 关于 J 的偏导是基于反向传播得到的。反向传播，顾名思义，是从输出层开始逐层计算梯度。首先，我们先计算损失函数 J 关于输出层自旋神经元激活函数的斜率 $k^{[3]}$ 的偏导 $\frac{\partial J}{\partial k^{[3]}}$：

$$\begin{aligned}\frac{\partial \boldsymbol{J}}{\partial k^{[3]}} &= \frac{\partial \boldsymbol{J}}{\partial a^{[3]}} \cdot \frac{\partial a^{[3]}}{\partial k^{[3]}} = -(\frac{y}{a^{[3]}} - \frac{1-y}{1-a^{[3]}}) \cdot \frac{\partial a^{[3]}}{\partial k^{[3]}} \\ &= (-(\frac{y}{a^{[3]}} - \frac{1-y}{1-a^{[3]}})) \cdot *\sigma'(k^{[3]} \cdot (z^{[3]} + c^{[3]})) \cdot *(z^{[3]} + c^{[3]})\end{aligned} \quad \ldots (3.10)$$

接着，与 $\frac{\partial J}{\partial k^{[3]}}$ 类似，损失函数 J 关于输出层自旋神经元激活函数的偏移 $c^{[3]}$ 的偏导 $\frac{\partial J}{\partial c^{[3]}}$：

$$\begin{aligned}\frac{\partial \boldsymbol{J}}{\partial c^{[3]}} &= \frac{\partial \boldsymbol{J}}{\partial a^{[3]}} \cdot \frac{\partial a^{[3]}}{\partial c^{[3]}} = -(\frac{y}{a^{[3]}} - \frac{1-y}{1-a^{[3]}}) \cdot \frac{\partial a^{[3]}}{\partial c^{[3]}} \\ &= (-(\frac{y}{a^{[3]}} - \frac{1-y}{1-a^{[3]}})) \cdot *\sigma'(k^{[3]} \cdot (z^{[3]} + c^{[3]})) \cdot *k^{[3]}\end{aligned} \quad \ldots (3.11)$$

然后，加入矩阵乘法的导数，损失函数 J 关于输出层和隐藏层相关联的权重 $w^{[2]}$ 的偏导 $\frac{\partial J}{\partial w^{[2]}}$：

$$\begin{aligned}\frac{\partial \boldsymbol{J}}{\partial w^{[2]}} &= \frac{\partial \boldsymbol{J}}{\partial z^{[3]}} \cdot a^{[2]^{\boldsymbol{T}}} \\ &= ((-(\frac{y}{a^{[3]}} - \frac{1-y}{1-a^{[3]}})) \cdot *\sigma'(k^{[3]} \cdot (z^{[3]} + c^{[3]})) \cdot *k^{[3]}) \cdot a^{[2]^{\boldsymbol{T}}}\end{aligned} \quad \ldots (3.12)$$

其中，$\frac{\partial J}{\partial z^{[3]}}$ 的计算如下：

$$\begin{aligned}\frac{\partial \boldsymbol{J}}{\partial z^{[3]}} &= \frac{\partial \boldsymbol{J}}{\partial a^{[3]}} \cdot \frac{\partial a^{[3]}}{\partial z^{[3]}} = -(\frac{y}{a^{[3]}} - \frac{1-y}{1-a^{[3]}}) \cdot \frac{\partial a^{[3]}}{\partial z^{[3]}} \\ &= (-(\frac{y}{a^{[3]}} - \frac{1-y}{1-a^{[3]}})) \cdot *\sigma'(k^{[3]} \cdot (z^{[3]} + c^{[3]})) \cdot *k^{[3]}\end{aligned} \quad \ldots (3.13)$$

基于式 3.13计算的 $\frac{\partial J}{\partial z^{[3]}}$，继续反向链式求导，可以得到损失函数 J 关于隐藏层和输入层相关联的权重 $w^{[1]}$ 的偏导 $\frac{\partial J}{\partial w^{[1]}}$：

$$\begin{aligned}\frac{\partial \boldsymbol{J}}{\partial w^{[1]}} &= \frac{\partial \boldsymbol{J}}{\partial z^{[2]}} \cdot a^{[1]^{\boldsymbol{T}}} = \frac{\partial \boldsymbol{J}}{\partial z^{[2]}} \cdot X^{\boldsymbol{T}} \\ &= \frac{\partial \boldsymbol{J}}{\partial a^{[2]}} \cdot \frac{\partial a^{[2]}}{\partial z^{[2]}} \cdot X^{\boldsymbol{T}} = w^{[2]^{\boldsymbol{T}}} \cdot \frac{\partial \boldsymbol{J}}{\partial z^{[3]}} \cdot \frac{\partial a^{[2]}}{\partial z^{[2]}} \cdot X^{\boldsymbol{T}} \\ &= ((w^{[2]^{\boldsymbol{T}}} \cdot \frac{\partial \boldsymbol{J}}{\partial z^{[3]}} \cdot *\sigma'(k^{[2]} \cdot (z^{[2]} + c^{[2]}))) \cdot *k^{[2]}) \cdot X^{\boldsymbol{T}}\end{aligned} \quad \ldots (3.14)$$

类似地，基于式 3.13计算的 $\frac{\partial J}{\partial z^{[3]}}$，可以计算损失函数 J 关于隐藏层自旋神经

元激活函数的斜率 $k^{[2]}$ 的偏导 $\frac{\partial J}{\partial k^{[2]}}$：

$$\begin{aligned}\frac{\partial J}{\partial k^{[2]}} &= \frac{\partial J}{\partial a^{[2]}} \cdot \frac{\partial a^{[2]}}{\partial k^{[2]}} = w^{[2]^T} \cdot \frac{\partial J}{\partial z^{[3]}} \cdot \frac{\partial a^{[2]}}{\partial k^{[2]}} \\ &= (w^{[2]^T} \cdot \frac{\partial J}{\partial z^{[3]}} \cdot *\sigma'(k^{[2]} \cdot (z^{[2]} + c^{[2]}))) \cdot *(z^{[2]} + c^{[2]})\end{aligned} \quad \ldots (3.15)$$

同样，基于式 3.13计算的 $\frac{\partial J}{\partial z^{[3]}}$，可以计算损失函数 J 关于隐藏层自旋神经元激活函数的偏移 $c^{[2]}$ 的偏导 $\frac{\partial J}{\partial c^{[2]}}$：

$$\begin{aligned}\frac{\partial J}{\partial c^{[2]}} &= \frac{\partial J}{\partial a^{[2]}} \cdot \frac{\partial a^{[2]}}{\partial c^{[2]}} = w^{[2]^T} \cdot \frac{\partial J}{\partial z^{[3]}} \cdot \frac{\partial a^{[2]}}{\partial c^{[2]}} \\ &= w^{[2]^T} \cdot \frac{\partial J}{\partial z^{[3]}} \cdot *\sigma'(k^{[2]} \cdot (z^{[2]} + c^{[2]})) \cdot *k^{[2]}\end{aligned} \quad \ldots (3.16)$$

最后，将上述的导数代入梯度更新公式 3.9中更新 $w^{[1]}$, $w^{[2]}$, $k^{[2]}$, $k^{[3]}$, $c^{[2]}$ 和 $c^{[3]}$ 的值。

利用更新后的参数通过前向传播计算更新后的损失函数 J 的值，继续重复前面的训练学习过程，不断降低 J 的值。在经过数次迭代更新之后就可以得到 J 的最小值，至此自旋神经网络的学习训练过程结束，参数 w，k 和 c 固定。

3.4.3 三层自旋神经网络的性能

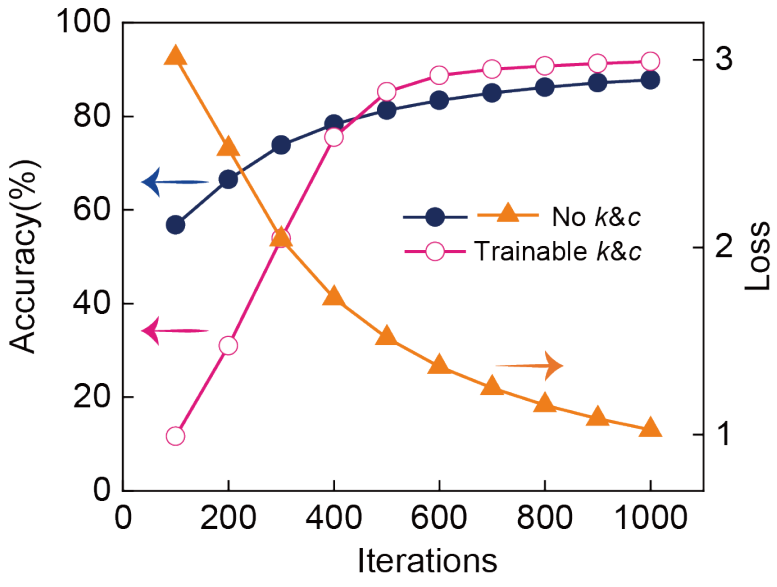

图 3.10　三层神经网络的损失和推理识别准确度。

Figure 3.10　The evolution of loss and accuracy of three-layer neural networks.

上述公式的推导考虑了涉及到的所有可训练参数 w，k 和 c 的梯度更新。对于传统的三层神经网络，只考虑 w 的情况也同样适用。当不考虑可训练激活函

数时，标准的三层神经网络只更新权重 (w_1 和 w_2)，从而逐渐减小损失函数 J 的值。如图 3.10中的实心三角形所示，损失函数 J 的值 Loss 随着训练学习过程的迭代而单调递减，神经网络输出的预测值 h_θ 与期望的正确值 y 的差异在逐渐减小，表明神经网络在不断地学习识别手写数字图片的方法。

接下来，使用测试集来评估已经学习训练过的神经网络的推理能力。因为测试集中 10000 张手写数字图片对于该神经网络来说是全新的输入样本，通过测试训练好的神经网络对新测试样本的识别准确度，可以真实地反映学习训练过的神经网络的推理能力。图 3.10中蓝色的实心圆表明，标准三层神经网络经过 1000 次迭代后，推理的识别准确率达到 88%。图 3.10中粉色的空心圆表明，加入可训练激活函数的自旋三层神经网络经过 1000 次迭代后，执行同样的推理任务，识别准确率达到 91.7%，准确率显著提升。可训练参数 k 和 c 的加入，在不改变神经网络拓扑结构的基础上显著提升了其推理能力，证明了本文提出的可训练激活函数的自旋神经元模型在执行推理任务时可以提供较大的增益。

3.5　改进的可训练激活函数的自旋神经元

在前面的讨论中，我们假设 k 和 c 在训练学习的过程中可以取任意值，这只是理想情况。

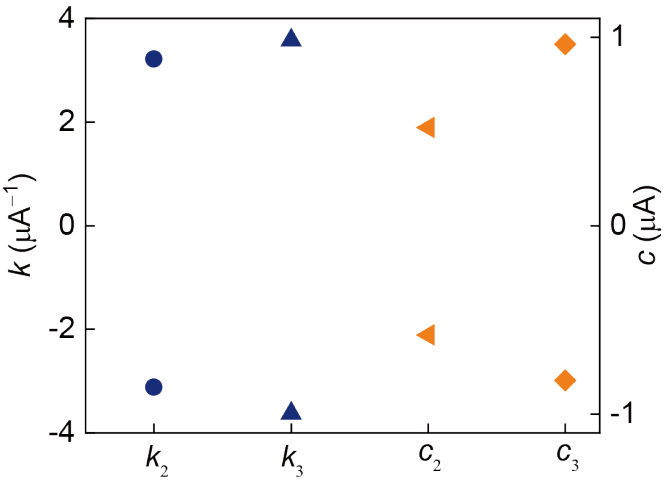

图 3.11　训练学习过程中提取的 k_2，k_3，c_2 和 c_3 的最大范围。

Figure 3.11　Maximum ranges of k_2, k_3, c_2 and c_3 extracted from the training process.

实际上，需要考虑硬件实现时自旋神经元输入的条件和输出的范围。本文将自旋神经网络的训练学习过程和自旋神经元对输入电脉冲的响应相结合，通过

跟踪训练学习过程中可训练参数 k 和 c 的变化，我们发现隐藏层中自旋神经元激活函数的斜率 k_2，偏移 c_2，输出层中自旋神经元激活函数的斜率 k_3 和偏移 c_3 的更新范围分别为 –3.1 到 3.2，–0.6 到 0.5，–3.6 到 3.6 和 –0.8 到 0.9，如图 3.11所示。

3.5.1 可训练参数的取值对训练学习过程的影响

在自旋神经网络的硬件实现中，必须考虑由器件本身物理特性决定的可以提供的 k 和 c 的允许值。我们在前面图 3.6中已经了解到，改变输入电脉冲 I_c 的脉冲宽度会导致 k 和 c 的值发生变化。提取图 3.6中的数据，本文在图 3.12中总结了在不同脉冲宽度条件下 k 和 c 的不同集合。

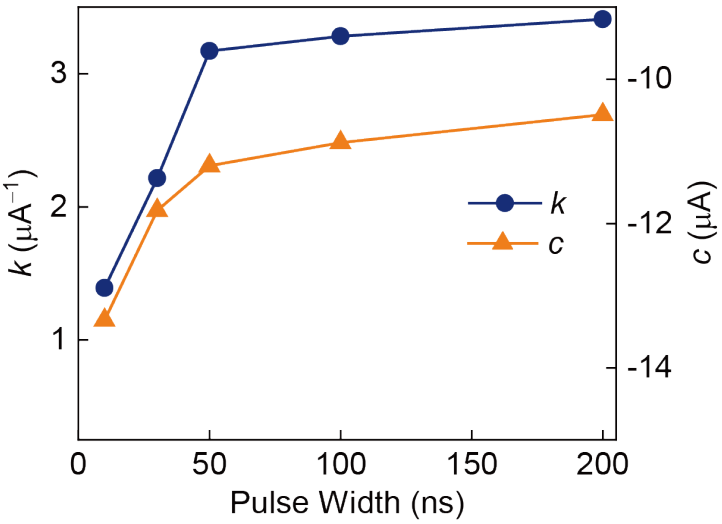

图 3.12　不同脉冲宽度下物理上允许的 k 和 c。

Figure 3.12　Physically allowed k and c under different pulse widths.

从图 3.12中可以看出，k 与脉冲宽度之间的关系呈线性趋势。如图 3.11中的蓝色符号所示，反向传播算法所需的 k_2 和 k_3 的范围小于图 3.12所示的实际器件可以实现的斜率 k 的范围。因此，可以通过改变脉冲宽度来满足训练学习过程中 k 的变化更新。

相比之下，器件允许取到的 c 的值似乎与训练过程中算法所需的值不一致。图 3.11中的橘色符号所示的算法所需的 c 的值小于 1μA，然而，图 3.12所示的自旋电子器件提供的 c 的值大于 10μA，c 取值范围明显不一致。

此外，还需注意，因为输入条件的改变，k 和 c 的响应是同时的，也就意味着 k 和 c 是耦合在一起的。

随着输入电脉冲 I_c 的脉冲宽度增加，磁性隧道结所需的翻转电流会降低并加快自由层磁化方向的翻转，对应的 c 和 k 的值会同时发生变化。这导致硬件实现前面提出的具有两个额外自由度 k 和 c 的可训练激活函数时会出现问题。

具体来说，当调整脉冲宽度以获得反向传播算法所需的 k 时，由于 k 和 c 之间的耦合是由自旋电子器件本身的物理特性确定的，因此 c 也会相应地改变为新值 c_a。但与此同时，与 k 的更新类似，训练学习过程的算法也要求 c 更新为新的值 c_b，很显然，根据前面 c 的取值范围的对比，我们知道 c_b 与 c_a 相同的可能性很小。训练学习过程中算法要求的更新值和器件的物理特性提供的允许值之间的不一致会导致自旋神经网络训练学习失败。

本文设计了仿真实验测试了上述的不一致会导致的影响。仿真实验中我们设定 k 的更新是遵循反向传播算法的要求。同时，c 也会根据 k 和 c 之间的关系随着 k 的更新更改为新值, 如图 3.13(a) 所示。从图 3.12中 k 和 c 分别关于脉冲宽度的值，可以整理得到如图 3.13(a) 所示的 k 和 c 的关系曲线。

对应地，当 k 的值由图 (a) 中的橘色圆点处的 $3.27\mu A^{-1}$ 更新至蓝色三角形处的 $1.38\mu A^{-1}$，c 的值也从相应的橘色圆点处的-10.85μA 变化到对应的蓝色三角形处的-13.29μA。仿真实验的结果表明，训练学习会在几次迭代后失败。

联合图 3.13(a) 和图 3.13(b) 可以理解其训练失败的原因。当 k 从图 3.13(a) 中的橘色圆点处的值变为蓝色三角形处的值，我们将橘色圆点处的 k 和 c 的值定义为旧值，蓝色三角形处的 k 和 c 的值定义为新值。如图 3.13(a) 所示，从旧值到新值，k 值的微小变化将导致 c 值的明显变化，对应于如图 3.13(b) 所示的激活函数从旧的曲线到新的曲线发生明显的偏移。图 3.13(b) 中橘色的旧的曲线对应图 3.13(a) 中的橘色圆点处的 k 和 c 的旧值，而蓝色的曲线对应图 3.13(a) 中的蓝色三角形处的 k 和 c 的新值。

观察图 3.13(b) 中的新旧曲线，可以发现，在相同的输入条件下，输入电脉冲 I_c 等于图 (b) 中虚线处 I_{old} 时，新的曲线在此处的斜率已接近零，输入处在激活函数的饱和区，面临梯度消失的问题，梯度更新受限，这表明权重将不会更改，因此无法继续进行训练学习。

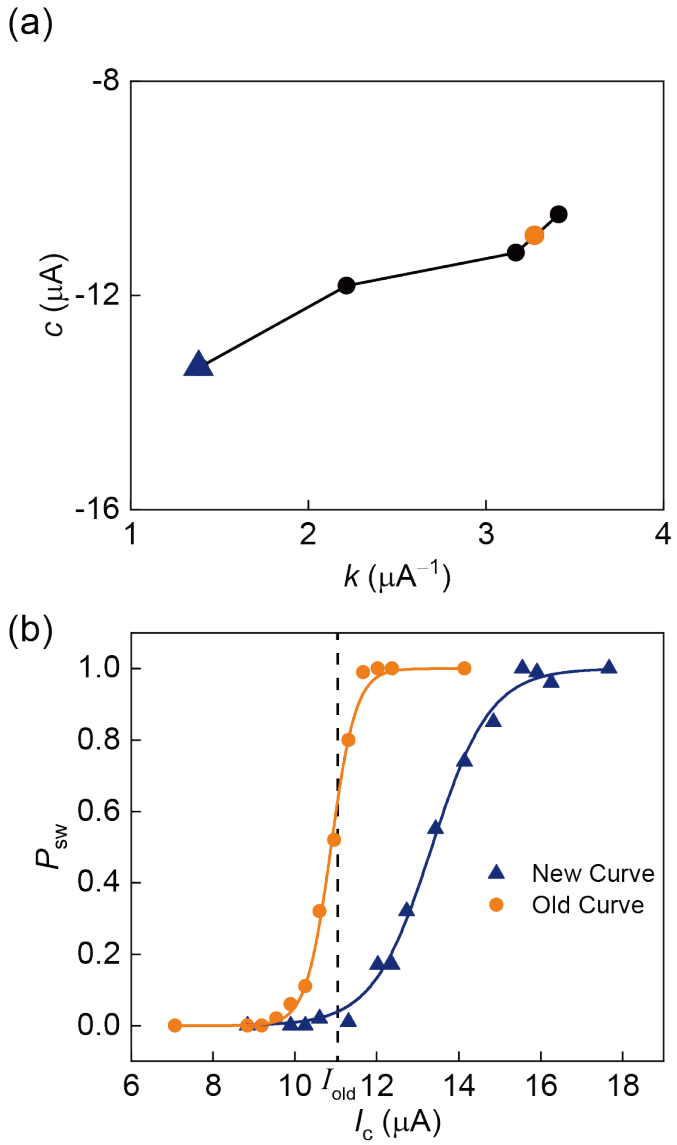

图 3.13　*c* 取值不一致的影响。(a) 从图 3.12中获得的 *k* 和 *c* 之间的关系。(b) 图 (a) 中标记的两点对应的激活函数。

Figure 3.13　The impact of inconsistent *c*.(a) Relation between *k* and *c* obtained from Fig. 3.12. (b) Activation functions corresponding the two points marked in Fig. 3.13(a).

3.5.2　额外单自由度的可训练激活函数

由于 *c* 在器件物理原理决定的允许值和训练学习过程中算法要求的更新值之间不一致，且 *k* 和 *c* 是耦合在一起等原因，导致硬件实现的自旋神经网络的训练学习过程很可能会在几次迭代后失败。

解决方案之一是通过利用器件本身的物理特性来解耦 *k* 和 *c*，并通过添加偏置电流将 *c* 移动到算法所需的值。依照上述解决方案，实际的 *k* 和 *c* 会按照

节 3.4.2中的训练学习过程进行更新，可以预料到按此方案实现的自旋神经网络的性能会如前面图 3.10所示，推理识别精度显著提高。

或者，本文在这里提供一个更容易实现的解决方案。回顾这个问题的产生，它是由两个事实引起的：

(1) k 和 c 耦合在一起，

(2) 算法所需的 c 值与器件提供的值之间不一致。

一个简单而直接的解决方案是在可训练的激活函数中删除只有微小变化的 c，只保留 k 作为可训练参数。改进的可训练激活函数只有一个额外的自由度 k，激活函数的斜率 k 可以在曲线不发生偏移的情况下改变。接下来，本文利用磁化翻转背后的物理原理来实现这一方案。

在 3.1节中详细阐述了提出的自旋神经元模型的器件原理。在前面的章节中也研究了该自旋神经元，在输入电脉冲 I_c 的脉冲宽度改变时，可以实现可调节激活函数。在前面的章节中也提到了该自旋神经元模型可以通过调控输入电脉冲的脉冲幅度实现可调节激活函数。类似于脉冲宽度，脉冲幅度也可以实现相似的调控。当施加幅度较大的电脉冲时，激活函数会向左偏移，曲线也会更陡，斜率也会更大。所以，无论是增加脉冲宽度还是增加脉冲幅度，都会降低翻转电流并加快自由层磁化方向的翻转速度，得到的激活函数也都会向左偏移并变陡。

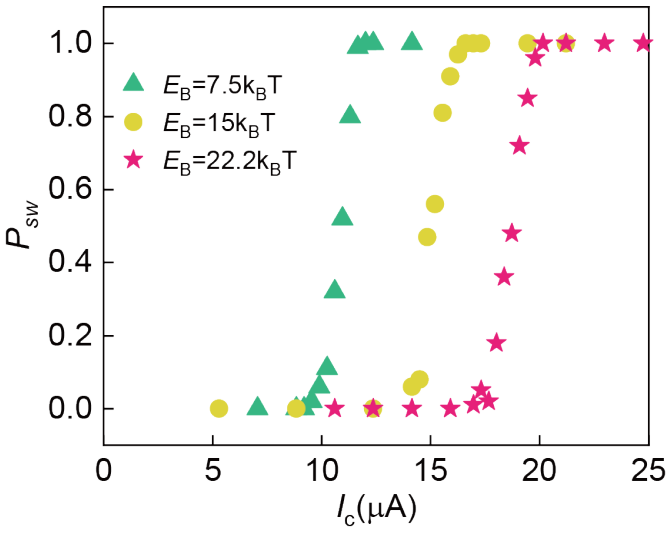

图 3.14　不同 E_B 下，P_{sw} 与 I_c 的关系。

Figure 3.14　P_{sw} as a function of I_c under different E_B.

相反，当通过调控磁各向异性 K_u 来降低自旋神经元的能量势垒 E_B 时，如

图 3.14所示，仅观察到 S 型翻转概率曲线向左偏移而斜率没有变化。图 3.14中绿色符号对应的器件参数以及能量势垒 $E_B = 7.5k_BT$ 的计算见式 3.7，对应的磁各向异性 $K_u = 1.1 \times 10^6 \mathrm{J/m^3}$。当调控磁各向异性 K_u 增加至 $1.21 \times 10^6 \mathrm{J/m^3}$ 时，根据式 3.7，可以计算对应的能量势垒 $E_B = (K_u - 0.5\mu_0 M_S^2) \times V_{FL} = 15k_BT$，其翻转概率曲线如图 3.14中黄色符号所示。

由于磁各向异性 K_u 决定了系统的能量势垒 E_B，翻转具有较小能量势垒 E_B 的系统所需的翻转电流较小，所以较小的磁各向异性 K_u 便对应于较小的翻转电流。因此，图 3.14中随着系统能量势垒 E_B 的降低，翻转概率的曲线会向左偏移。又因为施加的输入电脉冲 I_c 的脉冲宽度均为 30 ns，因此在所有情况下翻转速度都相同，从而导致相同的斜率 k。

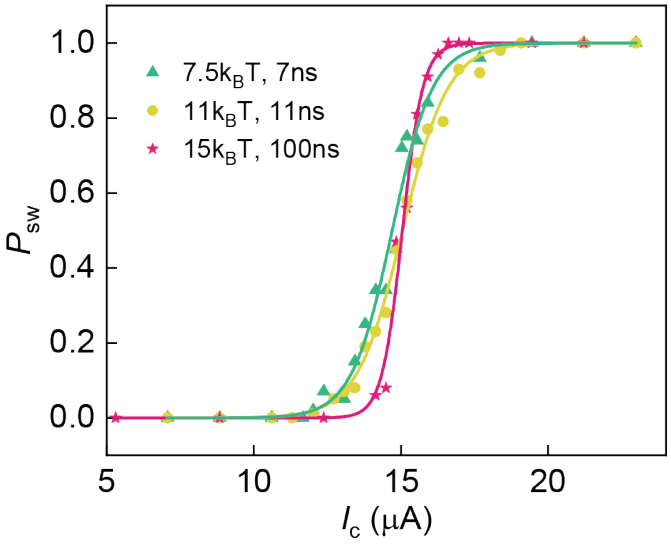

图 3.15　在不同 E_B 和脉冲宽度组合下，P_{sw} 与 I_c 的关系。

Figure 3.15　P_{sw} as a function of I_c under different combination of E_B and pulse widths.

对于在训练学习过程中通过输入电脉冲调控磁各向异性 K_u 的方法，本文在节 3.1中提供了两种实现方法。一种方法是对应于器件图 3.2的利用电压控制磁各向异性效应改变 K_u，另一种方法是对应于器件图 3.1的利用铁电应变效应改变 K_u。无论是图 3.1器件通过在铁电层上施加电压来控制 CoFeB/MgO 叠层中的垂直各向异性，还是图 3.2使用结合自旋轨道力矩和电压控制磁各向异性效应的三端磁性隧道结，利用外场 $\mathbf{H}_x$ 辅助的自旋轨道力矩进行翻转，通过电压控制磁各向异性来调控 K_u，都可以实现如图 3.15的输入电脉冲单独控制可训练参数 k。

综上所述，提出的自旋神经元模型既可以通过控制输入电脉冲 I_c 的脉冲宽

度或幅度来调控 k 和 c, 又可以通过控制磁各向异性 K_u 来调控 c, 因此就可以通过联合控制脉冲宽度和 K_u 来实现可训练参数 k 的单独调控，仿真结果如图 3.15所示。从图 3.15可以观察到，系统的能量势垒 E_B 越大，配合的输入电脉冲 I_c 的脉冲宽度越大，对应的翻转概率 S 型曲线的斜率 k 就可以单独变化。而且，通过联合控制脉冲宽度和 K_u，翻转概率 S 型曲线的偏移 c 很小，基本可以忽略不计。因此，通过联合控制脉冲宽度和 K_u 得到的可训练激活函数只添加了一个额外的自由度 k。

此外，在 K_u 的变化过程中，我们已经确定自由层始终保持沿垂直方向的易轴。例如，图 3.14中的势垒 $E_B = (K_u - 0.5\mu_0 M_S^2) \times V_{FL} = 15k_BT$ 时，K_u 值大于形状各向异性，导致了垂直磁化。前面图 3.3所示的弛豫结果也验证了上述的分析计算，其中磁化强度的初始方向相对 **z** 轴倾斜 45 度，并在 $J_c = 0$ 的情况下自然弛豫到 **z** 轴方向。

仿真实验测试了不同脉冲宽度和磁各向异性 K_u 的组合条件下，得到的改进的可训练激活函数的斜率 k 的值，如图 3.16所示。图 3.16展示了通过改变脉冲宽度和 K_u 来调控 k。仿真结果显示了 k 的变化范围。

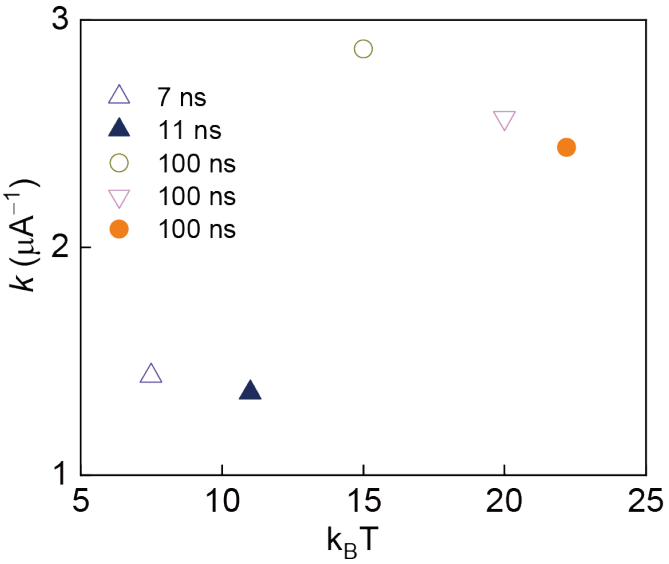

图 3.16　E_B 和脉冲宽度不同组合下的 k。

Figure 3.16　k under different combination of E_B and pulse widths.

3.5.3 改进的可训练激活函数的三层自旋神经网络

改进的自旋神经网络与 3.4节中的三层自旋神经网络的区别只是激活函数的不同。从具有两个额外自由度 k 和 c 的可训练激活函数变为只有一个额外自由度 k 的可训练激活函数。与三层自旋神经网络的学习训练算法类似，通过找到 $w^{[1]}$，$w^{[2]}$，$k^{[2]}$ 和 $k^{[3]}$ 的最优值来最小化损失函数 J。上述参数的更新如式 3.17所示。

$$\begin{cases} w^{[2]} = w^{[2]} - \alpha \frac{\partial J}{\partial w^{[2]}} \\ k^{[3]} = k^{[3]} - \alpha \frac{\partial J}{\partial k^{[3]}} \\ w^{[1]} = w^{[1]} - \alpha \frac{\partial J}{\partial w^{[1]}} \\ k^{[2]} = k^{[2]} - \alpha \frac{\partial J}{\partial k^{[2]}} \end{cases} \quad \ldots (3.17)$$

其中，α 表示神经网络的学习率，w 和 k 的更新值是上一次迭代过程中的参数值减去学习率和反向传播得到的关于 J 的偏导的乘积。

反向传播中只计算 w 和 k 关于 J 的偏导。首先，我们先计算损失函数 J 关于输出层自旋神经元激活函数的斜率 $k^{[3]}$ 的偏导 $\frac{\partial J}{\partial k^{[3]}}$：

$$\begin{aligned} \frac{\partial J}{\partial k^{[3]}} &= \frac{\partial J}{\partial a^{[3]}} \cdot \frac{\partial a^{[3]}}{\partial k^{[3]}} = -(\frac{y}{a^{[3]}} - \frac{1-y}{1-a^{[3]}}) \cdot \frac{\partial a^{[3]}}{\partial k^{[3]}} \\ &= ((-(\frac{y}{a^{[3]}} - \frac{1-y}{1-a^{[3]}})) \cdot *\sigma'(k^{[3]} \cdot z^{[3]})) \cdot *z^{[3]} \end{aligned} \quad \ldots (3.18)$$

其中，c 是可训练激活函数偏移的初始值，在训练学习过程中恒定不变，不可变化更新。然后，加入矩阵乘法的导数，损失函数 J 关于输出层和隐藏层相关联的权重 $w^{[2]}$ 的偏导 $\frac{\partial J}{\partial w^{[2]}}$：

$$\begin{aligned} \frac{\partial J}{\partial w^{[2]}} &= \frac{\partial J}{\partial z^{[3]}} \cdot a^{[2]^T} = \frac{\partial J}{\partial a^{[3]}} \cdot \frac{\partial a^{[3]}}{\partial z^{[3]}} \cdot a^{[2]^T} = -(\frac{y}{a^{[3]}} - \frac{1-y}{1-a^{[3]}}) \cdot \frac{\partial a^{[3]}}{\partial z^{[3]}} \cdot a^{[2]^T} \\ &= (((-(\frac{y}{a^{[3]}} - \frac{1-y}{1-a^{[3]}})) \cdot *\sigma'(k^{[3]} \cdot z^{[3]})) \cdot *k^{[3]}) \cdot a^{[2]^T} \end{aligned} \quad \ldots (3.19)$$

其中，$\frac{\partial J}{\partial z^{[3]}}$ 的计算如下：

$$\begin{aligned} \frac{\partial J}{\partial z^{[3]}} &= \frac{\partial J}{\partial a^{[3]}} \cdot \frac{\partial a^{[3]}}{\partial z^{[3]}} = -(\frac{y}{a^{[3]}} - \frac{1-y}{1-a^{[3]}}) \cdot \frac{\partial a^{[3]}}{\partial z^{[3]}} \\ &= ((-(\frac{y}{a^{[3]}} - \frac{1-y}{1-a^{[3]}})) \cdot *\sigma'(k^{[3]} \cdot z^{[3]})) \cdot *k^{[3]} \end{aligned} \quad \ldots (3.20)$$

基于式 3.20中计算得到的 $\frac{\partial J}{\partial z^{[3]}}$，继续反向链式求导，可以得到损失函数 J 关于隐藏层和输入层相关联的权重 $w^{[1]}$ 的偏导 $\frac{\partial J}{\partial w^{[1]}}$：

$$\begin{aligned}\frac{\partial J}{\partial w^{[1]}} &= \frac{\partial J}{\partial z^{[2]}} \cdot a^{[1]^T} = \frac{\partial J}{\partial a^{[2]}} \cdot \frac{\partial a^{[2]}}{\partial z^{[2]}} \cdot X^T = w^{[2]^T} \cdot \frac{\partial J}{\partial z^{[3]}} \cdot \frac{\partial a^{[2]}}{\partial z^{[2]}} \cdot X^T \\ &= ((w^{[2]^T} \cdot \frac{\partial J}{\partial z^{[3]}} \cdot *\sigma'(k^{[2]} \cdot z^{[2]})) \cdot *k^{[2]}) \cdot X^T\end{aligned} \quad \ldots (3.21)$$

类似地，基于式 3.20计算的 $\frac{\partial J}{\partial z^{[3]}}$，可以计算损失函数 J 关于隐藏层自旋神经元激活函数的斜率 $k^{[2]}$ 的偏导 $\frac{\partial J}{\partial k^{[2]}}$：

$$\begin{aligned}\frac{\partial J}{\partial k^{[2]}} &= \frac{\partial J}{\partial a^{[2]}} \cdot \frac{\partial a^{[2]}}{\partial k^{[2]}} = w^{[2]^T} \cdot \frac{\partial J}{\partial z^{[3]}} \cdot \frac{\partial a^{[2]}}{\partial k^{[2]}} \\ &= (w^{[2]^T} \cdot \frac{\partial J}{\partial z^{[3]}} \cdot *\sigma'(k^{[2]} \cdot z^{[2]})) \cdot *z^{[2]}\end{aligned} \quad \ldots (3.22)$$

最后，将上述的导数代入梯度更新公式 3.17中更新 $w^{[1]}$，$w^{[2]}$，$k^{[2]}$ 和 $k^{[3]}$ 的值。如图 3.17所示，改进的自旋神经网络损失 Loss 降到低于 1，训练学习完成后，w 和 k 固定下来。

3.6　改进的自旋神经网络的性能

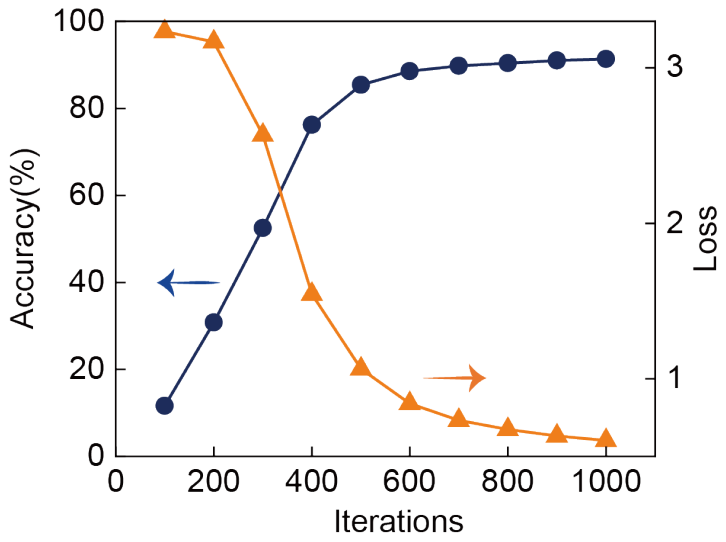

图 3.17　具有可训练 k 的系统损失和识别准确度。

Figure 3.17　The evolution of loss and accuracy in the system with trainable k.

仅将 k 作为一个额外的自由度添加到自旋神经网络中时，训练学习的算法推导与同时具有 k 和 c 的可训练激活函数的情况类似，并且实现了如图 3.17所示的 91.3% 的高识别准确率。

此外，从图 3.17中可以看出，具有可训练 k 的系统仅需 500 次迭代即可达到 85% 的准确率，而在没有可训练 k 的系统中需要 800 次迭代。对比图 3.10和图 3.17，当迭代次数较低时，不同系统的精度差异很大。本文将此归因于使用了不同的参数初始化规则。

3.6.1 参数初始化对准确率的影响

本章中展示的所有的神经网络的性能都是通过对 10 次独立的仿真实验进行平均而获得的。每次仿真实验中的参数都是随机初始化的。在本文的仿真中，按照参考文献[123]初始化参数，即 w 的初始值随机分布在 $[-\varepsilon_{\text{init}}, \varepsilon_{\text{init}}]$ 之间。

$$\varepsilon_{\text{init}} = \frac{\sqrt{6}}{\sqrt{n_{\text{in}} + n_{\text{out}}}} \quad \ldots (3.23)$$

其中 n_{in} 和 n_{out} 分别表示连接突触的输入神经元和输出的神经元的数量。

如图 3.18所示，使用随机初始化的权重进行仿真实验，10 次独立实验得到的平均准确率和平均损失与单次实验相比，差异可以忽略不计。

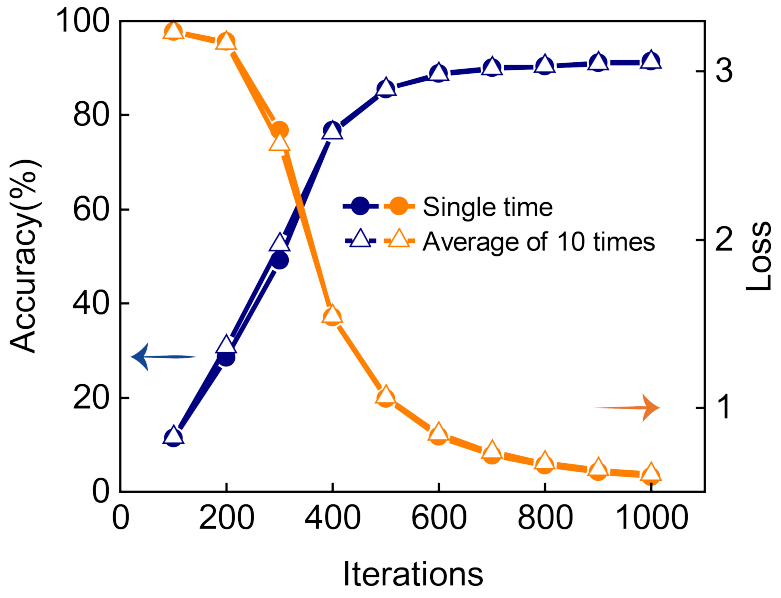

图 3.18　改进的自旋神经网络进行单次实验的结果和十次独立实验结果的平均值的对比。

Figure 3.18　Comparison between the average accuracy and loss of ten independent experiments and a single experiment using an improved spin neural network.

同样地，本文在前面的仿真中可训练参数 k 和 c 的初始值也都遵循上述的随机分布。这导致了在迭代次数较低时，不同激活函数的神经网络推理的准确率显

示出很大的差异。通过检查原始数据，我们发现，在低迭代情况下，具有可训练的 k 和 c 的系统的精度总是低于没有 k 和 c 的系统。

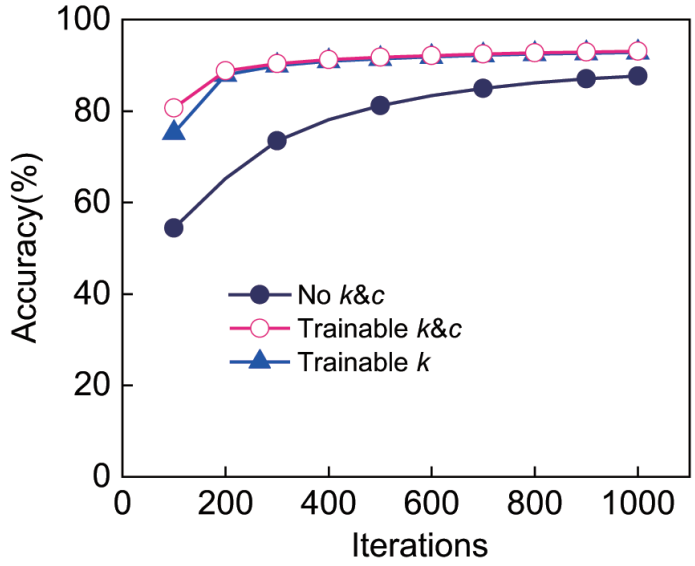

图 3.19　将 $\varepsilon_{\text{init,k}}$ 和 $\varepsilon_{\text{init,c}}$ 分别更改为 [−4，4] 和 [−1，1] 后，自旋神经网络的准确率。每个点都是通过对 10 次独立运行的结果进行平均获得的。

Figure 3.19　The evolution of accuracy after changing $\varepsilon_{\text{init,k}}$ and $\varepsilon_{\text{init,c}}$ to [−4, 4] and [−1, 1], respectively. Each point is obtained by averaging the results in 10 independent runs.

然而，正如前面提到的可训练 k 和 c 与突触权重 w 有根本的不同。因此，对 k 和 c 使用不同的初始化规则时更加合理的。本文在前面对 k 和 c 的分析中已经得到，k 值的变化范围小于 [-4, 4]，c 的变化范围小于 [-1, 1]。因此，本文将 k 和 c 的初始值 $\varepsilon_{\text{init,k}}$ 和 $\varepsilon_{\text{init,c}}$ 分别设置为 4 和 1。相应地，自旋神经网络进行训练学习后测得的准确率如图 3.19所示，低迭代时准确率显著提高。从图 3.19可以看出，引入可训练激活函数的自旋神经网络的性能要远好于没有 k 和 c 的系统。重复仿真实验也显示每次独立实验的准确率都比没有 k 和 c 的系统要好，k 和 c 的初始化对低迭代时的准确率影响显著，适当的参数初始化可以在低迭代时显著提高性能，但仍然显示出可训练激活函数可以提高自旋神经网络推理的准确率。

3.6.2　超参数学习率对准确率的影响

本文在训练过程中跟踪监测了 k 的变化，并确保它落在器件允许值的范围内。另外，本文还研究了超参数学习率 α 对准确率的影响。本文测试了不同的

学习率 (0.1 和 0.3) 的情况，如图 3.20所示。尽管选取不同的学习率会导致训练速度和识别精度产生差异，但这项工作的关键结论并未受到影响，即具有可训练 k 的激活函数的自旋神经元提高了神经网络的性能。后续的工作可以研究这些超参数的影响来进一步优化自旋神经网络。

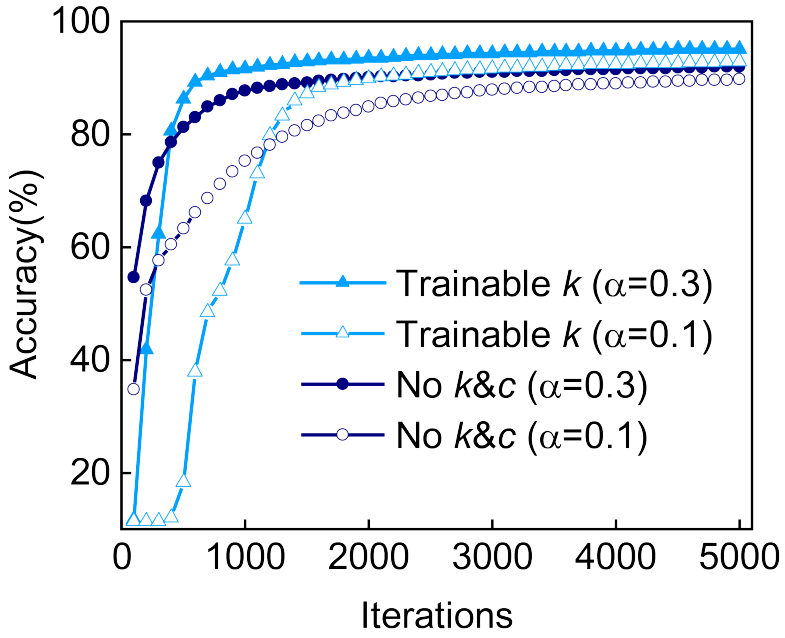

图 3.20　不同学习率的识别准确率比较。

Figure 3.20　Comparison of accuracy with different learning rates.

3.6.3　改进的自旋神经网络的估计功耗

最后，本文简要分析了提出的自旋神经网络的能量消耗。在本章的仿真实验中，将磁性隧道结用作神经元，通过施加输入电脉冲实现一系列操作。能耗的计算主要遵循 I^2Rt。施加的平均电流 $I = 15\mu$A，平均脉冲宽度为 15 ns，典型的磁性隧道结阻值为 R_{MTJ} = 5 KΩ，磁性隧道结近似能量消耗为 E_{MTJ} = 16.9 fJ /神经元。相比之下，实验中 CMOS 实现的单个神经元的能量消耗超 700 fJ[124-125]。其他已发表的关于自旋神经元的工作表明，自旋神经元的能量消耗为 0.3fJ[52]、1fJ[124]、18 至 36fJ[49]和 60fJ[125]。因此，本章提出的可训练激活函数的神经元的能量消耗与其他自旋神经元相当，并且比 CMOS 神经元小得多。需要注意的是，本章提出的自旋神经元中 k 和 c 的更新是通过改变脉冲宽度来实现的，脉冲宽度的能量已经包含在 E_{MTJ} 的计算中。因此，在这项工作中提出的可训练激活函数的自旋神经元可以在不引入额外能量消耗的情况下提高神经网络的性能。

3.7　本章小结

综上所述，本章提出了一种具有可训练激活函数的自旋神经元。通过热效应下控制输入电脉冲翻转磁性隧道结，获得了具有可训练斜率 k 和偏移 c 的 S 形激活函数。在之前的研究中，自旋神经元的激活函数在神经网络训练学习过程中斜率和偏移都是固定的。本章开发了一种算法，使斜率 k 和偏移 c 能够在训练学习过程中动态更新，从而实现更快的训练速度和更高的推理准确率。本文提出的可训练激活函数与批量归一化相类似，后者是深度神经网络中不可或缺的算法。本章还基于仿真实验结果验证了可以使用单个自旋电子器件，通过调控输入电脉冲的脉冲宽度和利用电脉冲改变磁各向异性，实现改进的可训练斜率 k 的激活函数。

本章基于磁性隧道结背后的物理原理的研究，为实现可训练激活函数的自旋神经元，提高自旋神经元性能提供了新的见解。本章提出的硬件级实现的建议为自旋电子器件和机器学习算法之间架起了桥梁，为基于自旋电子器件的神经网络的物理实现铺平道路。

第 4 章　非理想特性对自旋神经网络的影响

在上一章中，已经提出了三层自旋神经网络的架构设想，推导了具有可训练激活函数的自旋神经网络的算法。本章将沿用上一章提出的可训练激活函数的自旋神经网络的架构，并与基于拓扑绝缘体的磁性异质结构的自旋轨道力矩器件的电学表征相结合，研究了非理想特性对自旋神经网络的影响。

根据上一章关于自旋神经网络的研究，了解到自旋神经网络的基本构成单元是自旋突触单元和自旋神经元单元。本章考虑了自旋突触器件权重更新时的非理想特性，例如线性、对称性和稳定状态的数量等，对自旋神经网络执行推理任务的准确率的影响。另外，本章还在此基础上进一步研究了具有可训练激活函数的自旋神经网络，进一步提升了自旋神经网络的性能。

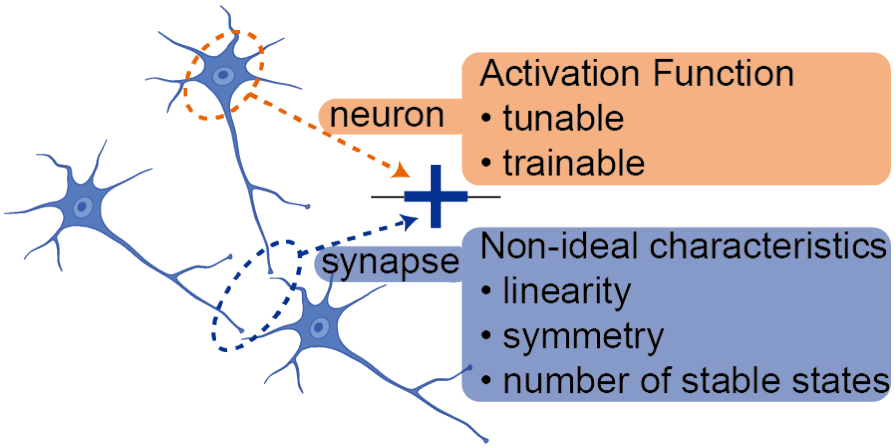

图 4.1　自旋神经网络的拓扑图以及自旋突触和神经元的功能。

Figure 4.1　Illustration of the spin neural networks and functions of spin Synapses and neurons.

如图 4.1所示，大脑中的神经网络是由神经元和突触相互连接形成的。突触是一种连接两个神经元的特殊结构，它允许突触前神经元将电信号或化学信号传递给突触后神经元，突触的权重反映了它们之间的连接强度[126-127]。本章中基于拓扑绝缘体的自旋轨道力矩器件既可以用作自旋突触也可以用作自旋神经元。作为自旋突触单元，它能够形成大量具有很强的热稳定性的中间状态，并且相应的电阻调节是线性和对称的，这些非理想特性的引入丰富了自旋神经网络硬件实现的仿真条件[128-131]。此外，作为自旋神经元单元，与上一章的自旋神经元类

似，可以实现具有可调节激活函数的自旋神经网络和具有可训练激活函数的自旋神经网络[71,132-133]。

虽然通过传统的数字系统实现的神经网络也可以达到很高的识别准确率[134-136]，但使用二进制实现大规模卷积运算会造成大量的能量损耗和时间成本[137-138]。然而，非线性和非易失性自旋动力学可以使磁阻器件具有类似存储器的行为，从而实现节能的自旋神经网络[139-140]。例如，通过自旋纹理或多磁畴结构的自旋电子器件可以实现具有多中间状态变量的人工突触[141,17,142,19,4]。利用自旋转移力矩或自旋轨道力矩，通过施加电脉冲信号调控磁畴，该自旋电子器件就会表现出对称的电阻变化，并且具有高度可编程的线性和持久性，对高精度实现突触功能至关重要[143]。另外，自旋电子器件也可以用于实现具有非线性激活函数的自旋神经元，上一章中详细描述了自旋神经元在电流诱导磁化翻转过程中可以呈现出 S 形的翻转曲线。类似地，基于自旋轨道力矩的自旋电子器件也可以实现自旋神经元的激活功能，而无需在外围电路中添加额外的激活功能模块[141]。

本章首先研究了实际的自旋突触单元的非理想特性对三层自旋神经网络的硬件实现的影响，更加全面地可靠地仿真了三层自旋神经网络的实现。基于自旋突触单元和自旋神经元的电学表征，验证了提出的自旋神经网络与其他已报道的忆阻器[141,17,144]构建的神经网络相比，具备更高的推理准确度。之后，本章提出自旋突触单元的长时程增强和长时程抑制过程的电阻调控和自旋神经元单元的可调节的激活函数都可以通过改变器件的 $CrTe_2$ 层的厚度来进一步优化。基于此实现了具有可调节激活函数的自旋神经网络。最后，实现了上一章提出的具有可训练激活函数的自旋神经网络，基于器件本身物理原理而实现的可训练的激活函数可以提升本章实现的自旋神经网络的性能，以更少的迭代周期获得更好的推理识别准确度。

4.1　自旋神经网络的非理想特性

本章中基于拓扑绝缘体的自旋轨道力矩器件既可以实现自旋突触的权重更新功能，也可以实现自旋神经元的激活功能。图 4.1中蓝色的自旋轨道力矩交叉器件的结构如图 4.2所示。

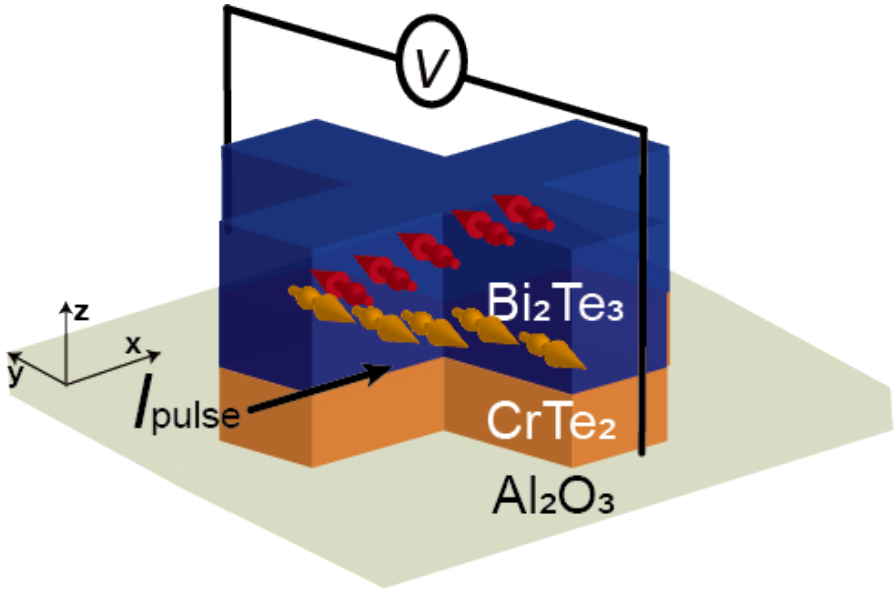

图 4.2　$Bi_2Te_3/CrTe_2$ 异质结构中自旋轨道力矩磁化翻转示意图。

Figure 4.2　Schematics of the Spin-Orbit Torque switching in $Bi_2Te_3/CrTe_2$ heterostructure.

在图 4.2所示的双层堆叠结构中, 拓扑表面态的强自旋轨道耦合 (Strong Spin-Orbit Coupling, SOC) 相关自旋动量锁定机制确保了 Bi_2Te_3 层的有效自旋极化[145-147]。在拓扑绝缘体 (Topological Insulators, TI) 材料中，SOC 引入了一种新型的相互作用，使得拓扑绝缘体表面导电态中电子的自旋方向与其动量方向紧密耦合。自旋动量锁定现象可以简化地理解为：一个电子在移动时，其自旋方向取决于它的动量方向，反之亦然。换句话说，这种锁定关系使得电子在沿着不同方向运动时呈现出不同的自旋极化特性，电子在某个方向上的运动会导致其自旋方向的改变，而在另一个方向上的运动则会使自旋方向保持不变。表面上存在导电态具有很强的自旋极化，即自旋方向与动量方向相互垂直。同时，在不引入额外的 PMA 辅助层 (Perpendicular Magnetic Anisotropy-assisted layer) 的情况下，$CrTe_2$ 固有的垂直磁各向异性 (Perpendicular Magnetic Anisotropy, PMA) 使 $CrTe_2$ 层能够与相邻的 Bi_2Te_3 沟道直接配对，这反过来又大大提高了 SOT 效率 ζ_{SOT}~1.76[148]。因此，基于 Bi_2Te_3(6nm)/$CrTe_2$(21ML) 的交叉器件中可以实现确定性自旋轨道力矩驱动的磁化翻转。在温度 T=120 K 时, 翻转电流密度 J_{SW} 小于 2.9×10^6A/cm^2。

$CrTe_2$ 层本身的多畴性质使得制造的基于 SOT 的 Bi_2Te_3(6 nm)/$CrTe_2$(21 ML) 器件在磁化翻转过程中会产生了多个中间状态变量。由此产生的通过电流调制器件在多个中间状态之间切换的 $V_{xy}-I_{DC}$ 关系如图 4.3所示，图中的小环路表现出两个显著的特征。

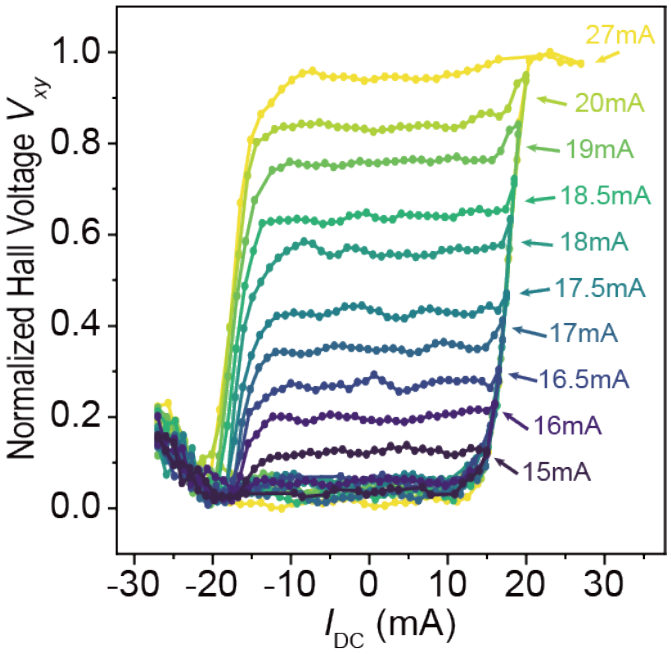

图 4.3　$Bi_2Te_3/CrTe_2$ 中 SOT 驱动的稳定多态翻转。

Figure 4.3　Stable multi-states switching driven by SOT in $Bi_2Te_3/CrTe_2$.

4.1.1　自旋突触单元的电学行为表征

首先，脉冲电流幅度从 10 mA 到 22 mA 连续变化会产生 12 个可重复的霍尔电阻状态，对应于图 4.5中脉冲个数从 1 增加到 12 的过程中产生的从 state #1 到 state #12 的 12 个权重状态，这表明基于 SOT 的 $Bi_2Te_3/CrTe_2$ 器件测得的霍尔电压值 V_{xy} 随着脉冲电流的增加表现出长时程增强 (long-term potentiation, LTP) 的特性，对应于具有 3-bit 权重的自旋突触 (SOT-S) 单元的 LTP 过程。同样地，通过反转训练学习电流脉冲的极性，还会产生如图 4.5中所示的对称的具有线性斜率的长时程抑制 (long-term depression, LTD)。图 4.5中黄色曲线上蓝色星号显示了自旋突触单元 SOT-S 的器件性能，它是由 6 次相同的增强 (或抑制) 的训练学习电流脉冲的实验平均得到的，LTP/LTD 曲线都保持高度线性，计算得到平均线性误差 < 5%。因此，本章的自旋突触 SOT-S 的电阻调制表现出高度线性对称，具有 12 个稳定的状态数。

SOT-S 器件作为自旋突触单元，图 4.3所示的稳定的多态数不仅使其权重值的更新具有低随机性，权重值在读取时也是稳定的。一旦 SOT-S 器件更新了特定的突触权重值，通过输入的读取电流脉冲将 SOT-S 器件的霍尔电阻调控到对应的值，得到的磁畴结构对热波动表现出很强的稳定性。

4.1.2 自旋神经元单元的电学行为表征

同时，基于 SOT 的 $Bi_2Te_3/CrTe_2$ 交叉器件在图 4.3所示的磁化翻转曲线表明，它也可以实现自旋神经元 SOT-N 的激活功能。如图 4.3所示，当施加 [0mA, 27mA] 范围内的输入电流 I_{DC} 时，测得的霍尔电压 V_{xy} 与输入电流之间的关系 $V_{xy}-I_{DC}$ 呈现出类似 sigmoid 激活函数的 S 型曲线。在上一章中， 3.2节阐述了自旋神经元的可调节激活函数的实现。类似地，将归一化后的霍尔电压与输入电流之间的关系 $V_{xy}-I_{DC}$ 和 sigmoid 激活函数相拟合，得到了自旋神经元 SOT-N 的可调节激活函数 $y=1/(1+exp(-k(x-x_c)))$。其中，k 表示 S 型曲线的斜率，x_c 表示 S 型曲线的偏移。当表达式 $y=1/(1+exp(-k(x-x_c)))$ 中斜率 k 和偏移 x_c 分别为 1 和 0 时，S 型曲线的表达式与标准 sigmoid 激活函数一致。图 4.8所示的黄色 S 型曲线对应的斜率 k 和偏移 x_c 的值分别为 0.89 和 17.59。可调节激活函数的参数 (k,x_c) 的值可以通过改变薄膜堆叠结构的厚度或输入电流的范围进行调控。通过调控激活函数的参数 (k,x_c) 可以提升自旋神经网络的性能，会在后面具体讨论。

总之，上述实验测得的器件特性证明了基于 SOT 的 $Bi_2Te_3/CrTe_2$ 交叉器件既可作为自旋突触单元，又可实现自旋神经元的激活功能，可以作为自旋神经网络的主要构建模块。

4.2 不同忆阻器搭建的神经网络的性能比较

利用更多的自旋突触单元和自旋神经元单元可以搭建三层自旋神经网络。通过训练学习后的自旋神经网络执行推理识别标准 MNIST 手写数据集的任务，将推理测试集的准确度作为评估神经网络性能的重要指标。MNIST 数据集由 50000 个训练样本和 10000 个测试样本组成，训练集 50000 个训练样本用于自旋神经网络的训练学习过程，测试集 10000 个测试样本用于推理测试自旋神经网络的识别准确率。

在上一章的 3.4节中详细阐述了三层自旋神经网络的架构和算法推导。本章构建的自旋神经网络与图 3.7大体上一致，区别在于隐藏层的自旋神经元数量由 25 增加至 100。同样地，如图 4.4所示的自旋神经网络，输入层有 784 个神经元，对应 MNIST 数据集中手写数字图片中的像素数 (28×28)。隐藏层和输出层分别有 100 个和 10 个自旋神经元. 其中，输出层的 10 个自旋神经元对应于手写数字

0~9 共 10 个类别。图 4.4中输入层的每个神经元都连接到隐藏层的所有神经元，权重 W_1 代表了它们之间的连接强度，所以权重 W_1 是一个 $100\times28\times28$ 的矩阵。权重 W_2 表示隐藏层神经元和输出层神经元之间的连接强度，对应地，权重 W_2 是一个 $10\times1\times100$ 的矩阵。

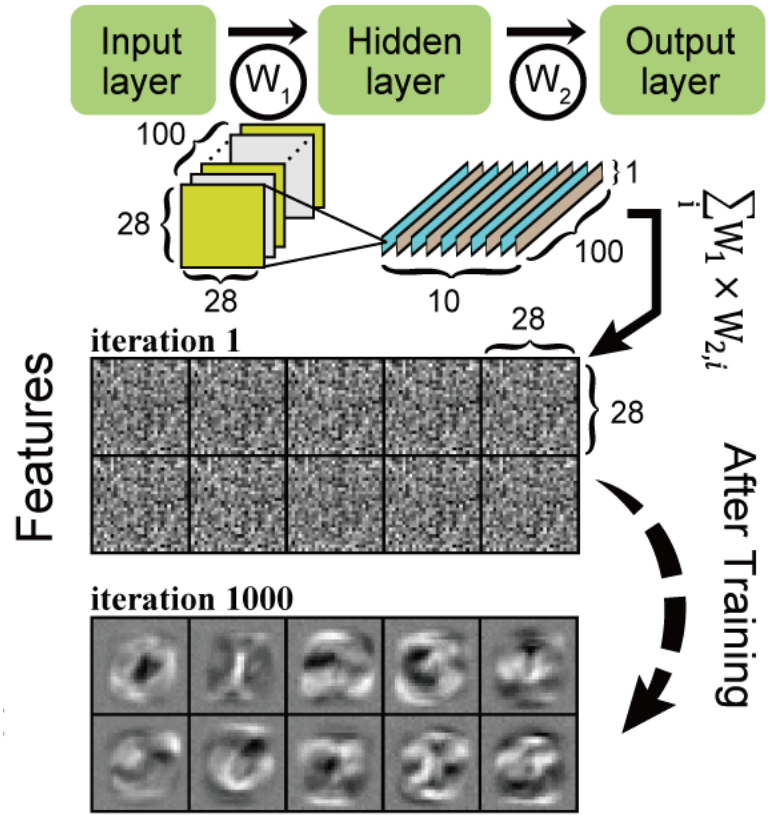

图 4.4　三层自旋神经网络训练学习过程示意图。自旋神经网络在 1000 次迭代之后，提取的输入层和隐藏层之间的权重 (W_1) 与隐藏层和输出层之间的权重 (W_2) 相乘的特征。

Figure 4.4　Schematic of the training process of a three-layer spin neural network.The captured features after 1000 iterations from the multiplication of weights between Input layer and Hidden layer (W_1) and between Hidden layer and Output layer (W_2).

上述的自旋神经网络搭建完成后，以 50000 个训练样本的 MNIST 训练集按照 3.4.2节阐述的训练学习过程进行 1000 次迭代，提取到了如图 4.4底部所示的训练后的特征。为了进一步了解特征提取过程，通过计算 $\sum_{i=1}^{n} W_1 \times W_{2,i}$ 可视化 10 个数字的特征。其中，n 是隐藏层神经元的数量，W_1 是输入层和隐藏层之间的权重数组，W_2 是隐藏层和输出层之间的权重数组。可以发现，在训练迭代 1000 次后，手写数字图片的特征变得越来越明显，表明训练学习过程趋于完成。

然而，在实际硬件实现自旋神经网络的训练学习过程时，突触器件的权重更

新会受到自身电阻调制的非线性、对称性，以及权重位数等影响。如图 4.5所示，通过比较本章提出的自旋突触 SOT-S 和其他先进的忆阻器作为人工突触的电阻调制结果，可以发现提出的自旋突触 SOT-S 与理想情况下的电阻调制最接近，对应图 4.5中黄色的 SOT-S 曲线与线性对称的紫色 Ideal 曲线几乎重叠。

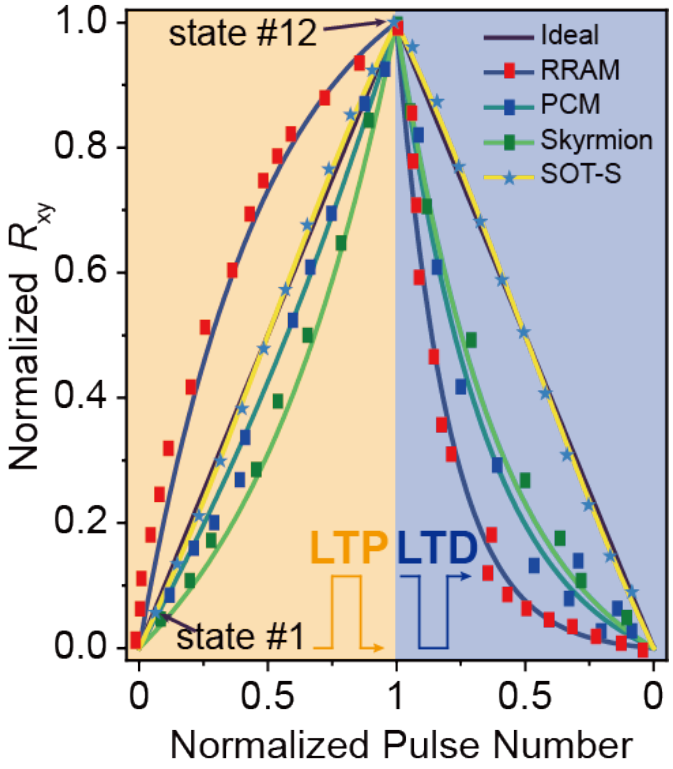

图 4.5　不同忆阻器器件作为人工突触的电阻调制。不同忆阻器对输入电流脉冲的响应，长时程增强和抑制过程的霍尔电阻 R_{xy} 与输入脉冲的归一化的曲线。

Figure 4.5　Resistance modulation of different memristor devices as artificial synapses. The normalized curves of Hall resistance R_{xy} and input pulses in long-term potentiation and depression process through applying input current pulses on different memristors.

在 4.1.1节中，通过固定脉宽为 500μs 的增强/抑制脉冲调制实验测得了如图 4.5图中蓝色星号黄色曲线所示的自旋突触器件 SOT-S 的 LTP/LTD 曲线。图中橘色的 LTP 过程的脉冲个数范围由 [0, 12] 等比例映射到 [0, 1] 范围内，表示 SOT-S 的权重值会随着输入增强型脉冲个数的增加而增加，输入的脉冲个数越多对应的突触权重值越大，自旋突触 SOT-S 处于长时程增强的过程。对应地，图中蓝色的 LTD 过程的脉冲个数范围由 [12, 24] 等比例映射到 [1, 0] 范围内，表示 SOT-S 的权重值会随着输入抑制型脉冲个数的增加而降低，自旋突触 SOT-S 处于长时程抑制的过程。通过归一化的操作，可以将不同忆阻器 LTP 和 LTD 过程

的输入变量统一，方便分析曲线的对称性[129]。通过施加正脉冲和负脉冲来调控突触器件的电阻，进而更新权重值。理想情况下，电阻调制是线性对称的，不会发生变化或产生耐久性下降等问题。然而，硬件实现中的非理想特征会影响在线训练学习和离线推理过程。

具体来说，理想情况下，当权重更新为新值时，只需利用反向传播算法计算出权重关于损失函数的梯度 $\frac{\partial J}{\partial w}$，然后带入权重更新公式 $w = w - \alpha\frac{\partial J}{\partial w}$ 中即可。因为电阻调制具有高线性和对称性，输入电脉冲与权重值一一对应，所以权重改变值 Δw 与期望变化量 $-\alpha\frac{\partial J}{\partial w}$ 一致。

然而，对非线性非对称的电阻调制来说，例如图 4.5中深蓝色的 RRAM 曲线，权重更新的期望值会与实际更新值之间存在差异。已知 t 时刻的权重值为 w_{t}，根据反向传播算法计算出的权重关于损失函数的梯度为 $\frac{\partial J}{\partial w}$，权重值实际改变值 $\Delta w = w_{\mathrm{t+1}} - w_{\mathrm{t}}$ 很难直接调控至期望变化量 $-\alpha\frac{\partial J}{\partial w}$。由于深蓝色的 RRAM 曲线的 LTP 和 LTD 过程中，由上一时刻的电阻值变化为期望的电阻值，需要针对不同的过程对应地输入不同的脉冲序列。所以，实际的非理想特征要求分别考虑对应于 LTP 和 LTD 过程的权重更新情况[128-129]。接下来，本章通过将不同忆阻器对应的 LTP/LTD 过程的实验数据拟合，得到了对应图 4.5中不同忆阻器 LTP/LTD 电阻调制曲线。

图 4.5中深蓝色的 RRAM 的 LTP/LTD 电阻调制曲线对应的梯度更新方程，是由实验数据[141]归一化后拟合得到的非线性非对称函数：

$$\begin{aligned} R_{\mathrm{LTP}} &= \frac{R_{\max} - R_{\min}}{1 - e^{-\upsilon}}(1 - e^{-\upsilon P}) + R_{\min} \\ R_{\mathrm{LTD}} &= R_{\max} - \frac{R_{\max} - R_{\min}}{1 - e^{-\upsilon}}(1 - e^{-\upsilon(1-P)}) \end{aligned} \quad \ldots (4.1)$$

其中，$R_{\max}$ 和 $R_{\min}$ 分别表示电阻调控范围的最大值和最小值，R_{LTP} 和 R_{LTD} 分别代表 LTP 和 LTD 过程响应输入脉冲的电阻值。电阻值 R 对应于神经网络中突触的权重值。υ 是表征非线性的参数。当 υ=0 时，响应是完全线性的。拟合实验测得的 RRAM 数据[141]得到 LTP 过程的 $\upsilon = 2$，LTD 过程的 $\upsilon = 5$。P 表示归一化后的输入脉冲数。

类似地，浅蓝色的 PCM 的 LTP/LTD 电阻调制曲线对应的梯度更新方程：

$$\begin{aligned} R_{\mathrm{LTP}} &= R_{\max} - \frac{R_{\max} - R_{\min}}{1 - e^{-\upsilon}}(1 - e^{-\upsilon(1-P)}) \\ R_{\mathrm{LTD}} &= R_{\max} - \frac{R_{\max} - R_{\min}}{1 - e^{-\upsilon}}(1 - e^{-\upsilon(1-P)}) \end{aligned} \quad \ldots (4.2)$$

其中，拟合实验测得的 PCM 数据[144]得到 LTP 过程的 $\upsilon = 0.6$，LTD 过程的 $\upsilon = 2.8$。

从图 4.5可以看出，绿色的 Skyrmion 的 LTP/LTD 电阻调制曲线与浅蓝色的 PCM 曲线形状相似，其对应的梯度更新方程也与式 4.2的形式一致：

$$\begin{aligned} R_{\mathrm{LTP}} &= R_{\max} - \frac{R_{\max} - R_{\min}}{1 - e^{-\upsilon}}(1 - e^{-\upsilon(1-P)}) \\ R_{\mathrm{LTD}} &= R_{\max} - \frac{R_{\max} - R_{\min}}{1 - e^{-\upsilon}}(1 - e^{-\upsilon(1-P)}) \end{aligned} \quad \ldots (4.3)$$

其中，拟合实验测得的 Skyrmion 数据[17]得到 LTP 过程的 $\upsilon = 1.6$，LTD 过程的 $\upsilon = 2.4$。

黄色的 SOT-S 的 LTP/LTD 电阻调制曲线对应的梯度更新方程：

$$\begin{aligned} R_{\mathrm{LTP}} &= (R_{\max} - R_{\min}) \times \frac{e^{\upsilon} + 1}{e^{\upsilon} - 1} \times \frac{1}{1 + e^{-2\upsilon(P-0.5)}} + R_{\min} - \frac{R_{\max} - R_{\min}}{e^{\upsilon} - 1} \\ R_{\mathrm{LTD}} &= (R_{\max} - R_{\min}) \times \frac{e^{\upsilon} + 1}{e^{\upsilon} - 1} \times \frac{1}{1 + e^{-2\upsilon(P-0.5)}} + R_{\min} - \frac{R_{\max} - R_{\min}}{e^{\upsilon} - 1} \end{aligned} \quad \ldots (4.4)$$

其中，拟合实验测得如图 4.5所示的 SOT-S 数据得到对称的 LTP 和 LTD 过程，非线性的参数相等，υ 均为 1。

为测试上述非理想特性对神经网络性能的影响，将图 4.4中所示的三层神经网络的突触 W_1 和 W_2 换成上述图 4.5所示的非理想特性的突触器件，对应的神经网络经过 1000 次训练迭代后测试其执行推理任务的准确度，结果如图 4.6所示。

在硬件实现三层神经网络的仿真实验中，采用训练后量化 (Post-Training Quantization, PTQ)[149-150]的量化策略。神经网络在完成训练学习过程后，PTQ 通过将模型参数，例如权重和激活函数的输出值从 32 位浮点数量化为低位数值表示，从而实现模型的压缩。这种量化方法可以显著降低模型的存储需求和计算复杂度，压缩硬件实现的电路尺寸，从而提高推理速度和能效。然而，量化可能导致精度损失，因为低位数值表示具有较低的精度和动态范围。本文提出的器件表现出多个稳定的中间状态，可以实现足够的状态数保证自旋神经网络的高准确度[141,17,151,144]。

另外，由于提出的器件既可以作为自旋突触单元 SOT-S，又可以作为自旋神经元单元 SOT-N，所以图 4.6中突触单元的黄色曲线对应于图 4.8中自旋神经元的黄色 S 型曲线，自旋神经网络中激活函数是斜率 k 和偏移 x_c 的值分别为 0.89

和 17.59 的激活函数 $y = 1/(1 + exp(-k(x - x_c)))$。而其他忆阻器器件对应的神经网络只考虑器件的突触的功能，神经元的激活函数与理性情况一致，都是标准 sigmoid 函数 $y = 1/(1 + exp(-x))$。

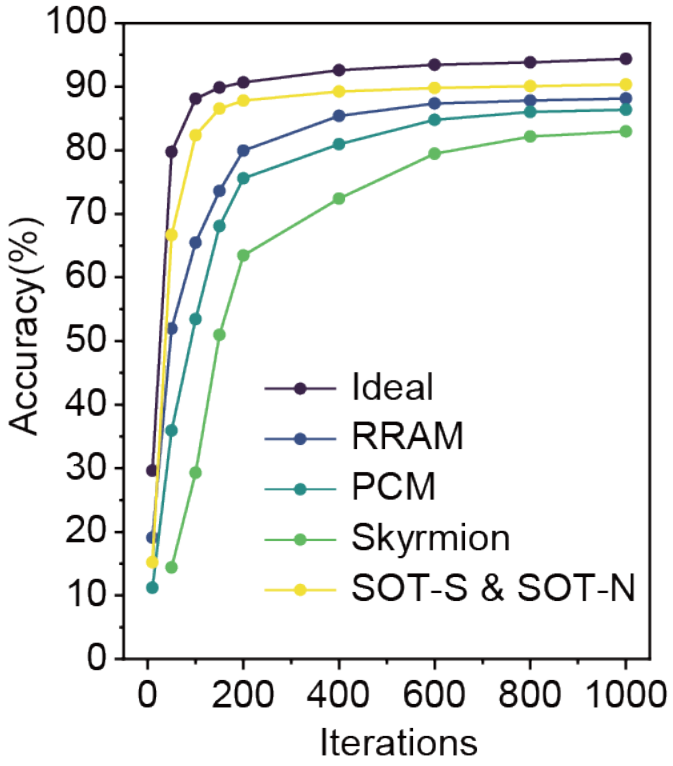

图 4.6　不同忆阻器器件构成的三层神经网络的推理准确度比较。

Figure 4.6　Comparison of testing accuracy of three-layer neural networks composed of different memristors.

通过比较图 4.6所示的推理识别准确度，可以看出，在硬件实现神经网络时，人工突触器件的非理想特性确实会降低神经网络的识别准确度，这也与前面解释的突触权重实际改变量与期望变化量之间的不一致是相符的。此外，图 4.5中本文提出的器件 SOT-S 具有高对称性 (对称误差 < 0.9%)、高线性 (线性误差 < 4.2%) 以及足够的状态数，所以由 SOT-S 和 SOT-N 搭建的三层自旋神经网络实现了 90.34% 的识别准确度，明显优于报道的其他忆阻器[141,17,144]。

综上，本章的基于拓扑绝缘体的自旋轨道力矩器件同时可以实现突触和神经元的功能，而由自旋突触单元 SOT-S 和自旋神经元单元 SOT-N 搭建的三层自旋神经网络可以可靠地执行高准确率的识别推理任务，其良好的对称性和线性等非理想特性使得硬件实现的自旋神经网络的性能接近理性情况下神经网络的性能。

4.3 具有可调节激活函数的自旋神经网络

本文在3.2节详细阐述了自旋神经元模型的可调节激活函数，接下来，将根据本章中的自旋神经元 SOT-N 器件本身的物理特性，实现可调节激活函数的自旋神经网络。同时，本章节还探讨了 $CrTe_2$ 层厚度的变化对自旋突触单元 SOT-S 和自旋神经元单元 SOT-N 性能的影响。

4.3.1 $CrTe_2$ 层厚度和温度的影响

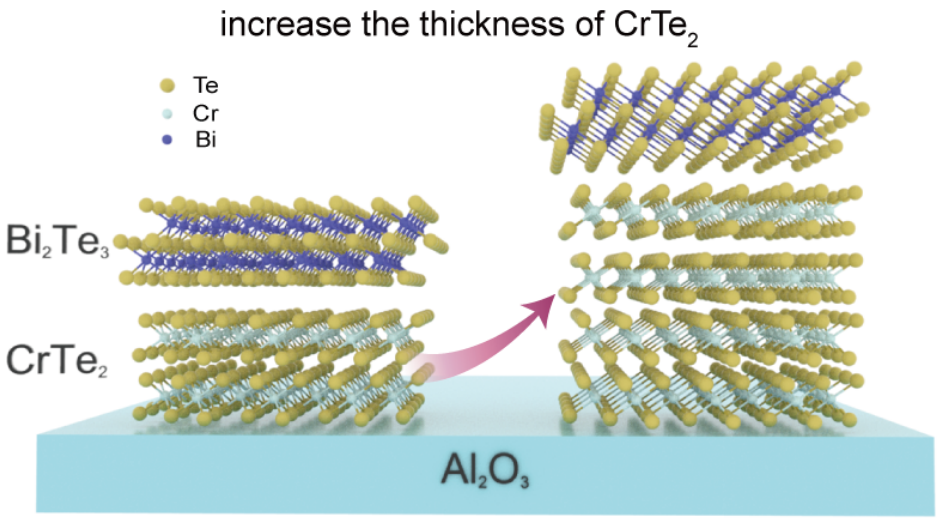

图 4.7 增加 $CrTe_2$ 厚度的异质结构示意图。

Figure 4.7 Schematics of heterostructure with increasing $CrTe_2$ thickness.

已有研究[152-153]表明，在如图 4.7中 $CrTe_2$ 层中发现了一种特有的与厚度相关的铁磁特征，因此可以通过改变 $CrTe_2$ 层的厚度来优化自旋突触单元 SOT-S 和自旋神经元单元 SOT-N 的性能。通过实验比较不同 $CrTe_2$ 层厚度的自旋突触 SOT-S 器件和自旋神经元 SOT-N 器件的磁学和电学行为表征，接着设计仿真实验得到不同 $CrTe_2$ 层厚度样品构建的三层自旋神经网络的准确度，就可以根据自旋神经网络性能的表现选择最优的 $CrTe_2$ 层厚度的自旋突触单元 SOT-S 和自旋神经元单元 SOT-N，达到器件优化的目的。

在前面的章节中，测试研究的基于自旋轨道力矩的 Bi_2Te_3(6 nm)/$CrTe_2$(21 ML) 器件中 $CrTe_2$ 层的厚度是固定的。在本章节中，通过与前面的 Bi_2Te_3(6 nm)/$CrTe_2$(21 ML) 样品同样的制备方法，得到一组不同的 $CrTe_2$ 层厚度的样品，

通过实验测试其磁学行为表征，得到了四种 $CrTe_2$ 层厚度的样品的电阻磁滞回线 (R-H loop)。反常霍尔效应下的电阻 R_{xy} 的 4 条电阻磁滞回线表明 $CrTe_2$ 层厚度 (d) 从 5 ML 到 21 ML 的 4 个样品中存在垂直磁各向异性。此外，矫顽场 H_c 在 120K 的温度条件下，从 12.6 mT (对应 d = 21 ML 的电阻磁滞回线) 单调下降到 7.5 mT (对应 d = 5 ML 的电阻磁滞回线)，垂直磁各向异性强度降低。

随着样品中 $CrTe_2$ 层厚度的减小，矫顽场 H_c 也逐渐减小，进而导致磁性翻转所需的自旋轨道力矩也随之减弱，从而导致翻转电流 I_c 的降低。

通过比较不同温度下的样品的翻转电流 I_c 与 $CrTe_2$ 层厚度的关系，发现在同一温度下，翻转电流 I_c 会随着 $CrTe_2$ 层厚度的减小而降低，而同一 $CrTe_2$ 层厚度的样品的翻转电流 I_c 会随着温度的升高而降低。由此，可以得到不同 $CrTe_2$ 层厚度的样品在指定温度下作为自旋突触单元 SOT-S 和自旋神经元单元 SOT-N 的电学行为表征。

4.3.2　不同 $CrTe_2$ 层厚度的自旋突触单元的性能比较

样品的 $CrTe_2$ 层厚度 5ML、9ML，14ML 以及 21ML 在温度为是 120K 的条件下，作为自旋突触单元 SOT-S，其 LTP 和 LTD 过程的电阻调制结果与图 4.5中类似，都具有高线性度和良好的对称性，但表现出的稳定状态的数量具有明显不同。

$CrTe_2$ 层厚度为 5ML 的样品，高低电阻之间的可调范围是 $CrTe_2$ 层厚度为 21ML 的 10 倍。相比于 $CrTe_2$ 层厚度为 21ML 的 12 个稳定状态，$CrTe_2$ 层厚度为 5ML 的样品在 LTP/LTD 过程中表现出 18 个稳定状态，权重值的更新写入精度提升。另外，相比图 4.5中 12 个状态的自旋突触 SOT-S 器件，$CrTe_2$ 层厚度为 5ML 的样品在相同的输入电流脉冲调控范围内，输出的霍尔电压具有更宽的阈值范围，降低了读取的精度要求。

4.3.3　不同 $CrTe_2$ 层厚度的自旋神经元单元的性能比较

对于自旋神经元器件 SOT-N，$CrTe_2$ 层厚度的可调使其能够实现前面 3.2节和 4.1.2节的可调节激活函数，如图 4.8所示。

在 3.2节中，提出的自旋神经元模型通过控制输入电脉冲的脉冲宽度实现可调节激活函数。通过改变输入电流脉冲的幅度，脉冲宽度固定为 500μs，得到了如

图 4.8所示的 $CrTe_2$ 层厚度为 5ML、9ML，14ML 以及 21ML 的自旋神经元 SOT-N 器件的磁化翻转曲线。

自旋神经元 SOT-N 的可调节激活函数 $y = 1/(1+exp(-k(x-x_c)))$。其中，k 表示 S 型曲线的斜率，x_c 表示 S 型曲线的偏移。当表达式 $y = 1/(1+exp(-k(x-x_c)))$ 中斜率 k 和偏移 x_c 分别为 1 和 0 时，S 型曲线的表达式与标准 sigmoid 激活函数一致。通过改变 $CrTe_2$ 层厚度就可以相应地调控斜率 k 和偏移 x_c。

从图 4.8可以看出，$Bi_2Te_3/CrTe_2$(5 ML) 的自旋神经元 SOT-N 的 S 型曲线具有最大的斜率值，k = 1.076。根据 3.5节中关于斜率 k 和偏移 x_c 对于自旋神经网络推理准确率的影响的分析，理论上斜率 k 值越大的非线性激活函数，可以更有效地提取特征，能够实现更高准确度的自旋神经网络[141]。将 $CrTe_2$ 层厚度 5ML、9ML，14ML 以及 21ML 的器件搭建如图 4.4的三层自旋神经网络，自旋突触 SOT-S 的权重按照 120K 条件下测得的 LTP/LTD 过程的非理想特性的电阻调制进行更新，自旋神经元 SOT-N 的激活功能表达为图 4.8中对应 $CrTe_2$ 层厚度的可调节激活函数。

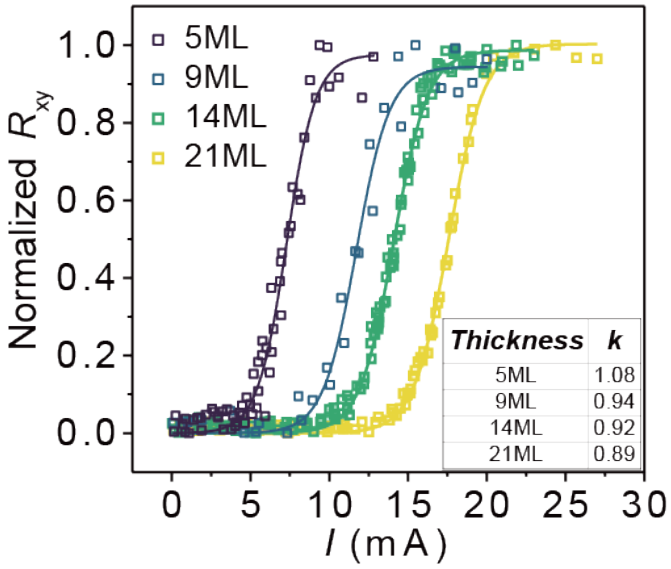

图 4.8　不同 $CrTe_2$ 层厚度的自旋神经元器件 SOT-N 的电学行为表征。

Figure 4.8　Electrical properties of SOT-N with different $CrTe_2$ thicknesses.

4.3.4　具有可调节激活函数的神经网络的性能比较

如图 4.9所示，由 Bi_2Te_3(6 nm)/$CrTe_2$(5 ML) 异质结构制造的自旋突触单元 SOT-S 和自旋神经元单元 SOT-N 器件搭建的三层自旋神经网络的推理准确率最

高，为 93.45%。

图 4.9中分别标注了 $CrTe_2$ 层厚度 5ML、9ML，14ML 以及 21ML 的器件实现的自旋神经网络完成训练学习过程后，通过训练后量化的量化策略得到的关于测试集数据的识别准确率。

造成不同自旋神经网络的推理准确率差异的原因主要有两个。一方面，不同 $CrTe_2$ 层厚度的器件作为自旋突触单元 SOT-S 表现出的电阻调制特性不同，$CrTe_2$ 层厚度为 5ML 时，稳定的中间状态最多，电阻的可调范围最大；另一方面，不同 $CrTe_2$ 层厚度的器件作为自旋神经元单元 SOT-N 测得的可调节激活函数不同，如图 4.8所示，$CrTe_2$ 层厚度为 5ML 时，器件所需的磁化翻转电流最小，磁化翻转也更快，对应的可调节激活函数的斜率 k 也最大。

通过仿真实验研究了自旋神经网络的推理准确率与可调节激活函数的斜率 k 和偏移 x_c 之间的关系，从图 4.9中颜色变化可以看出，准确率会随着可调节激活函数的斜率 k 值的增加而升高，这与前面 3.5节的分析一致。

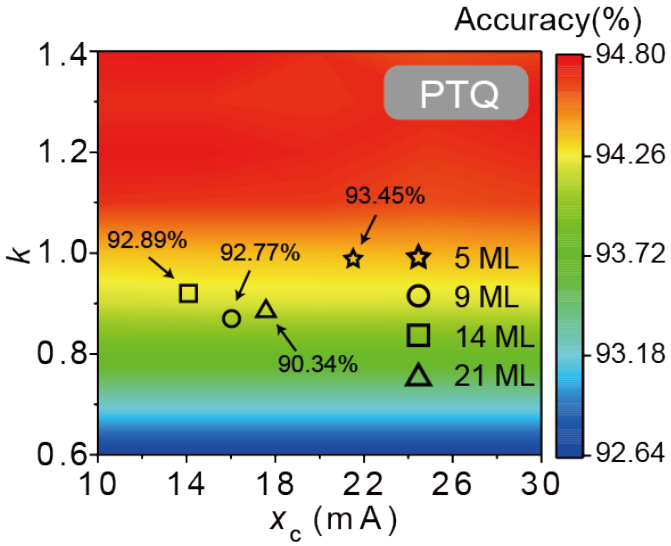

图 4.9　训练后量化策略下，不同斜率 k 和偏移 x_c 的可调节激活函数对应的神经网络识别准确率。

Figure 4.9　Accuracy diagram of neural network with tunable activation function of different slopes*k*and shifts x_c through PTQ.

另外，本章还针对量化感知训练这一量化策略进行了仿真实验。对于不同 $CrTe_2$ 层厚度的 $Bi_2Te_3/CrTe_2$ 器件构成的三层自旋神经网络，有别于上面的训练后量化，采用量化感知训练 (Quantization Aware Training, QAT)[154-156] 的在训练

学习过程中量化的策略，自旋神经网络可以在降低面积成本和功耗的同时保持高精度。QAT 是一种在训练神经网络模型过程中同时进行量化的技术。与 PTQ 相比，QAT 在训练阶段就考虑了量化对模型性能的影响，因此可以在压缩模型的同时减小精度损失。

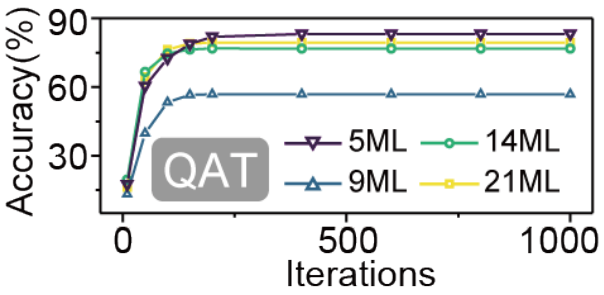

图 4.10　当 n 个 SOT-S 通过量化感知训练组合为一个突触时，四种 $CrTe_2$ 厚度情况的推理准确率。

Figure 4.10　The testing accuracy of four $CrTe_2$ thickness cases when n SOT-S are combined as one synapse through QAT.

与二元突触单元相比，自旋突触单元 SOT-S 的 18 个量化权重提供了更高的集成密度。通过将多个自旋突触单元 SOT-S 组合成一个高精度的自旋突触来满足学习训练过程中量化感知训练的状态数要求。具体来说，当 n 个自旋突触单元 SOT-S 单元连接在一起时，其中，每个自旋突触单元都具有 i 个稳定的中间状态，那么组合后的自旋突触的总状态数便增加到 $N = n \times i$。在图 4.10中可以看到使用量化感知训练的方法训练四个不同 $CrTe_2$ 层厚度的 $Bi_2Te_3/CrTe_2$ 器件构成的三层自旋神经网络，依然是 $Bi_2Te_3/CrTe_2$(5ML) 对应的系统展示出最高的识别准确率。在 1000 次迭代后，通过将 16 个自旋突触单元 SOT-S 封装在一起，从而使总状态数达到 288，具有最大状态数的 $Bi_2Te_3/CrTe_2$(5ML) 样本对应的系统可以达到 83.15% 的准确率。因此，在不同 $CrTe_2$ 层厚度的 $Bi_2Te_3/CrTe_2$ 器件构成的三层自旋神经网络中，不论是采取训练后量化的量化策略，还是量化感知训练，$Bi_2Te_3/CrTe_2$(5ML) 对应的系统面对 10000 组测试集数据都展现出最高的准确率。

此外，本章还对上述自旋神经网络的实现进行了功耗分析，根据公式 4.5计算实验中的能耗，输入脉冲 I_{SW} 的脉冲宽度固定为 500μs。

$$E = {I_{SW}}^2 Rt \quad \ldots (4.5)$$

重复上述计算过程, 分别计算 120K 温度下 $CrTe_2$ 层厚度为 5ML、9ML, 14ML 以及 21ML 的器件和 80K 温度下 $Bi_2Te_3/CrTe_2$(5ML) 的单次写入能耗。在 120K 温度下，当 $CrTe_2$ 层厚度从 21ML 下降至 5ML 时，对应的能耗从 0.201mJ 下降 51.2% 至 0.098mJ，在 120K 温度下 $CrTe_2$ 层厚度为 5ML 的器件的单次写入能耗最低。然而改变温度为 80K 时，$Bi_2Te_3/CrTe_2$(5ML) 单次写入所消耗的能量升高。

4.4　具有可训练激活函数的自旋神经网络

在前面的章节中，可调节激活函数虽然可以提升自旋神经网络的性能，但是一旦器件制备完成 $CrTe_2$ 层厚度便固定下来。3.5节中已经讨论过可训练激活函数对自旋神经网络的提升，基于此本节阐述了具有可训练激活函数的自旋神经网络的实现。在前面已经实现的三层自旋神经网络的基础上，本节进一步研究了自旋神经元单元 SOT-N 产生的激活函数的特性，发现当输入电流扫描范围发生变化时，系统中自旋轨道力矩驱动的磁化翻转回路的轮廓会随之变化。基于此可以实现 3.5节中改进的可训练激活函数，如图 4.11 所示，当输入电流在-32 mA $< I_{neuron} <$ -5 mA 的条件下，得到的可训练激活函数的斜率 k 的调控范围为 [0.37, 1.28]。

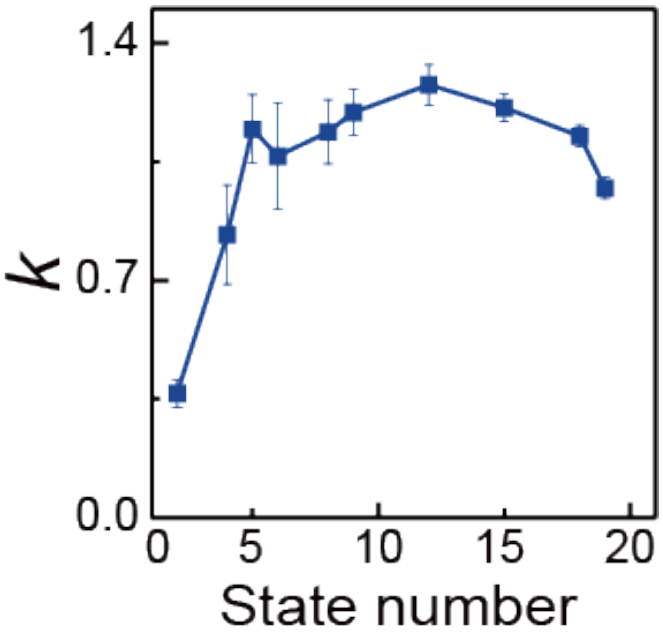

图 4.11　通过改变电流扫描范围得到的可训练激活函数的斜率 *k* 的调控范围。

Figure 4.11　The slope *k* of the trainable activation function by varying the current scanning range.

如图 4.12中左图所示，采用训练后量化的量化策略，比较具有可训练激活函数的自旋神经网络和上一节图 4.9中表现出最高准确率 93.45% 的 Bi_2Te_3(6 nm)/$CrTe_2$(5 ML) 器件搭建的三层自旋神经网络之间的测试推理准确率和训练学习过程的损失。可以发现，斜率 k 的调控范围为 [0.37, 1.28] 的可训练激活函数所对应的三层自旋神经网络的准确率更高，为 95.38%。对应地，其训练学习过程后得到的损失为 0.273，也明显小于上一节中 Bi_2Te_3(6 nm)/$CrTe_2$(5 ML) 器件搭建的自旋神经网络的损失 (0.366)。

此外，相比于理想情况的具有可调节激活函数或可训练激活函数的三层自旋神经网络，本章实现的自旋神经网络的性能仅仅下降了不到 1%，显示出本章实现的自旋神经网络具有很强的可靠性，自旋突触单元 SOT-S 和自选神经元单元 SOT-N 都表现出良好的特性。

如图 4.12中右图所示，类似地，采用量化感知训练的量化策略，比较具有可训练激活函数的自旋神经网络和上一节图 4.9中表现出最高准确率 83.15% 的 Bi_2Te_3(6 nm)/$CrTe_2$(5 ML) 器件搭建的三层自旋神经网络之间的测试推理准确率和训练学习过程的损失。具有可训练激活函数的自旋神经网络进一步将准确率提升至 89.33%，训练学习过程中迭代 1000 次后得到的损失也降至 0.8。

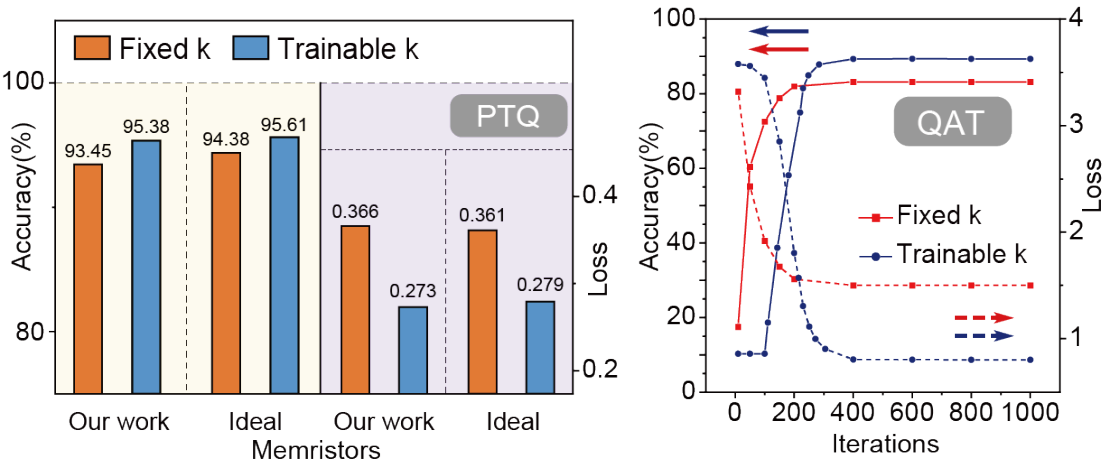

图 4.12　在训练后量化 (左图) 和量化感知训练 (右图) 的量化策略下，比较具有/不具有可训练激活功能的自旋神经网络的准确性。

Figure 4.12　Comparison of the accuracy between the spin neural network with/without the trainable activation function through PTQ (left panel) and QAT (right panel).

综上，上一章提出的具有可训练激活函数的自旋神经网络确实显示出更好的推理准确率和训练学习过程中更低的损失。本节实现的具有可训练激活函数

的自旋神经网络不仅表现出高精确率的推理能力，还具有优良的可靠性。

4.5　本章小结

综上所述，本章提出了一个具有可训练激活函数的自旋神经网络的实现方案。在上一章提出的三层自旋神经网络的基础上，与实际自旋突触单元 SOT-S 的非理想特性和自旋神经元单元 SOT-N 的可调节激活函数或可训练激活函数相结合，搭建了具有可调节激活函数的自旋神经网络和具有可训练激活函数的自旋神经网络。通过比较实现的自旋神经网络的性能，例如推理准确率、训练学习过程的损失以及单次写入能耗等，证实了实现的具有可训练激活函数的自旋神经网络展现出很高的可靠性、执行推理任务的高准确率和低功耗。因此，本章提出的具有可训练激活函数的自旋神经网络推动了高性能类脑计算系统的实现，为基于自旋电子器件本身的物理特性实现类脑计算架构提供了新的视角。

第 5 章　总结与展望

自旋电子器件本身的非易失性、非线性自旋动力学等特性使其用于实现自旋突触和自旋神经元具有天然的优势。但是，相关应用大部分着重于改变器件的结构研究其电学行为表征与神经元或突触的生物表达之间的关系，自旋神经网络多作为器件的应用。一方面，自旋神经网络的实现应该与计算机科学中神经网络相关的算法相结合，将自旋电子器件的特性充分发挥，促进紧凑高效低功耗的类脑计算网络的实现。另一方面，自旋神经网络的实现不应该只是简单地用自旋电子器件替代 CMOS 器件，应该与自旋电子器件本身的物理特性相结合，设计搭建集成的自旋神经网络。

本文针对上述关于实现自旋神经网络的问题，提出了相应的解决方案。基于自旋突触单元和自旋神经元单元背后的物理特性，可靠地实现了具有可训练激活函数的自旋神经网络。第 3章研究了可训练激活函数的自旋神经元模型的原理和应用。磁性隧道结中磁化方向的随机翻转被用于产生 S 形激活函数。然而，在先前的研究中，激活函数的形状在神经网络的训练过程中是固定的。本章创新性地提出在训练过程中也允许激活函数发生变化时，神经网络的性能便可以大大提高。

这项工作利用自旋力矩感应磁化开关切换背后的物理原理，通过添加激活函数的斜率 (k) 和位移 (c) 作为额外的训练学习参数，实现激活函数的动态变化。然而，进一步研究发现反向传播算法所需的 k 和 c 的期望值与自旋电子器件的提供值之间存在差异。通过输入电脉冲调控器件的磁各向异性，将激活函数的 k 和 c 解耦得到了改进的可训练激活函数。之后，改进的自旋神经网络执行推理识别 MNIST 手写数据集的任务，展现出 91.3% 的推理测试精确率。

值得注意的是，本工作中可训练参数 k 和 c 的使用与神经网络中使用的权重有根本不同，即 k 和 c 是神经元的局部参数，而权重是非局部参数。因此，在这项工作中引入 k 和 c 为改进神经网络提供了新的自由度，这在以前的研究中没有讨论过。在这项工作中，自旋神经元不仅仅只是产生激活函数，更多地，基于其本身的物理特性实现可训练激活函数，在没有引入额外能耗的同时提高网络准确率，降低自旋神经网络训练学习过程的时间成本。

第 4章进一步研究了具有可训练激活函数的自旋神经网络的实现。探索了不同的量化策略，训练后量化和量化感知训练，对自旋神经网络的推理测试准确率的影响。另外，比较了不同忆阻器作为突触单元，其线性度、对称性和状态数等非理想特性对神经网络的推理测试准确率的影响，得到自旋神经网络的性能最接近理性情况下的神经网络，其测试准确率为 90.34%。然后，为进一步提升实现的自旋神经网络的性能，本文还研究了具有可调节激活函数的自旋神经网络和具有可训练激活函数的自旋神经网络，证实了实现的具有可训练激活函数的自旋神经网络具有很高的可靠性、执行推理任务的高准确率，更低的训练学习过程的损失和低功耗等优良性能。

具有可训练激活函数的自旋神经网络与深度神经网络中不可或缺的批量归一化算法的思想类似。因此，这项工作表明，类脑计算网络的发展不再局限于软件上算法的实现。事实上，自旋电子器件背后的物理学来可以推动类脑计算网络的发展。因此，本文提出的具有可训练激活函数的自旋神经网络的实现将自旋电子器件和机器学习算法连接起来，为基于自旋突触和自旋神经元实现类脑计算提供了新的见解，为基于自旋电子器件的类脑计算网络的实现铺平道路。

本研究同样存在局限性，本文的自旋神经网络是由输入电脉冲调控的，训练学习过程中的梯度更新遵循反向传播算法。然而，神经网络算法现已发展到第三代由事件驱动的脉冲神经网络，脉冲神经网络更接近人脑的神经系统。同时，脉冲神经网络的高计算效率和低能耗突出了其在高效信息处理方面的巨大潜力，与之相比，本文的自旋神经网络还有很大的进步空间。自旋电子器件具有非易失性和高能效的特点，理论上可以实现脉冲时序依赖可塑性[157-158]的突触学习规则和带泄漏的整合发放神经元模型 (Leaky integrate-and-fir，LIF)[159]。但与所有新兴技术一样，脉冲神经网络的发展存在争议，脉冲神经网络仍面临着实际使用困难、训练和学习困难以及复杂任务的准确性低等问题，设计实现基于真实生物神经系统的类脑计算系统仍是一项巨大的挑战。

参考文献

[1] SENGUPTA A, BANERJEE A, ROY K. Hybrid spintronic-CMOS spiking neural network with on-chip learning: Devices, circuits, and systems[J]. Physical Review Applied, 2016, 6(6): 064003.

[2] SRINIVASAN G, SENGUPTA A, ROY K. Magnetic tunnel junction enabled all-spin stochastic spiking neural network[C]//Design, Automation & Test in Europe Conference & Exhibition (DATE), 2017. IEEE, 2017: 530-535.

[3] CHEN M C, SENGUPTA A, ROY K. Magnetic skyrmion as a spintronic deep learning spiking neuron processor[J]. IEEE Transactions on Magnetics, 2018, 54(8): 1-7.

[4] LOCATELLI N, CROS V, GROLLIER J. Spin-torque building blocks[J]. Nature Materials, 2014, 13(1): 11-20.

[5] DIENY B, PREJBEANU I L, GARELLO K, et al. Opportunities and challenges for spintronics in the microelectronics industry[J]. Nature Electronics, 2020, 3(8): 446-459.

[6] CHAPPERT C, FERT A, VAN DAU F N. The emergence of spin electronics in data storage [J]. Nature materials, 2007, 6(11): 813-823.

[7] PARKIN S, YANG S H. Memory on the racetrack[J]. Nature nanotechnology, 2015, 10(3): 195-198.

[8] ALLWOOD D A, XIONG G, FAULKNER C C, et al. Magnetic domain-wall logic[J]. Science, 2005, 309(5741): 1688-1692.

[9] FERT A, REYREN N, CROS V. Magnetic skyrmions: advances in physics and potential applications[J]. Nature Reviews Materials, 2017, 2(7): 1-15.

[10] CHUMAK A V, VASYUCHKA V I, SERGA A A, et al. Magnon spintronics[J]. Nature physics, 2015, 11(6): 453-461.

[11] BORDERS W A, PERVAIZ A Z, FUKAMI S, et al. Integer factorization using stochastic magnetic tunnel junctions[J]. Nature, 2019, 573(7774): 390-393.

[12] GROLLIER J, QUERLIOZ D, CAMSARI K Y, et al. Neuromorphic spintronics[J]. Nature electronics, 2020, 3(7): 360-370.

[13] MARKOVIĆ D, MIZRAHI A, QUERLIOZ D, et al. Physics for neuromorphic computing [J]. Nature Reviews Physics, 2020, 2(9): 499-510.

[14] FUKAMI S, OHNO H. Perspective: Spintronic synapse for artificial neural network[J]. Journal of Applied Physics, 2018, 124(15): 151904.

[15] MANIPATRUNI S, NIKONOV D E, YOUNG I A. Beyond CMOS computing with spin and polarization[J]. Nature Physics, 2018, 14(4): 338-343.

[16] ZÁZVORKA J, JAKOBS F, HEINZE D, et al. Thermal skyrmion diffusion used in a reshuffler device[J]. Nature nanotechnology, 2019, 14(7): 658-661.

[17] SONG K M, JEONG J S, PAN B, et al. Skyrmion-based artificial synapses for neuromorphic computing[J]. Nature Electronics, 2020, 3(3): 148-155.

[18] 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[J]. Advanced Materials, 2019, 31(23): 1900636.

[19] ROMERA M, TALATCHIAN P, TSUNEGI S, et al. Vowel recognition with four coupled spin-torque nano-oscillators[J]. Nature, 2018, 563(7730): 230-234.

[20] TORREJON J, RIOU M, ARAUJO F A, et al. Neuromorphic computing with nanoscale spintronic oscillators[J]. Nature, 2017, 547(7664): 428-431.

[21] BORDERS W A, AKIMA H, FUKAMI S, et al. Analogue spin−orbit torque device for artificial-neural-network-based associative memory operation[J]. Applied physics express, 2016, 10(1): 013007.

[22] LUO Z, HRABEC A, DAO T P, et al. Current-driven magnetic domain-wall logic[J]. Nature, 2020, 579(7798): 214-218.

[23] MANIPATRUNI S, NIKONOV D E, LIN C C, et al. Scalable energy-efficient magnetoelectric spin−orbit logic[J]. Nature, 2019, 565(7737): 35-42.

[24] LUO Z, DAO T P, HRABEC A, et al. Chirally coupled nanomagnets[J]. Science, 2019, 363(6434): 1435-1439.

[25] BRATAAS A, KENT A D, OHNO H. Current-induced torques in magnetic materials[J]. Nature materials, 2012, 11(5): 372-381.

[26] MANCHON A, ŽELEZNÝ J, MIRON I M, et al. Current-induced spin-orbit torques in ferromagnetic and antiferromagnetic systems[J]. Reviews of Modern Physics, 2019, 91(3): 035004.

[27] MATSUKURA F, TOKURA Y, OHNO H. Control of magnetism by electric fields[J]. Nature nanotechnology, 2015, 10(3): 209-220.

[28] CARETTA L, MANN M, BÜTTNER F, et al. Fast current-driven domain walls and small skyrmions in a compensated ferrimagnet[J]. Nature nanotechnology, 2018, 13(12): 1154-1160.

[29] KIM K J, KIM S K, HIRATA Y, et al. Fast domain wall motion in the vicinity of the angular momentum compensation temperature of ferrimagnets[J]. Nature materials, 2017, 16(12): 1187-1192.

[30] FINLEY J, LIU L. Spin-orbit-torque efficiency in compensated ferrimagnetic cobalt-terbium alloys[J]. Physical Review Applied, 2016, 6(5): 054001.

[31] BALTZ V, MANCHON A, TSOI M, et al. Antiferromagnetic spintronics[J]. Reviews of Modern Physics, 2018, 90(1): 015005.

[32] JUNGWIRTH T, MARTI X, WADLEY P, et al. Antiferromagnetic spintronics[J]. Nature nanotechnology, 2016, 11(3): 231-241.

[33] LIU Y, SHAO Q. Two-dimensional materials for energy-efficient spin–orbit torque devices [J]. ACS nano, 2020, 14(8): 9389-9407.

[34] LIN X, YANG W, WANG K L, et al. Two-dimensional spintronics for low-power electronics[J]. Nature Electronics, 2019, 2(7): 274-283.

[35] MONROE D. Neuromorphic computing gets ready for the (really) big time[Z]. 2014.

[36] HAN J, ORSHANSKY M. Approximate computing: An emerging paradigm for energy-efficient design[C]//2013 18th IEEE European Test Symposium (ETS). IEEE, 2013: 1-6.

[37] 莫宏伟, 丛垚. 类脑计算研究进展[J]. 导航定位与授时, 2021, 8(4): 53-67.

[38] PEI J, DENG L, SONG S, et al. Towards artificial general intelligence with hybrid Tianjic chip architecture[J]. Nature, 2019, 572(7767): 106-111.

[39] SENGUPTA A, ROY K. Spin-Transfer Torque Magnetic neuron for low power neuromorphic computing[C]//2015 International Joint Conference on Neural Networks (IJCNN). Killarney, Ireland: IEEE, 2015: 1-7.

[40] EBONG I E, MAZUMDER P. CMOS and memristor-based neural network design for position detection[J]. Proceedings of the IEEE, 2011, 100(6): 2050-2060.

[41] YU Z, SHEN M, ZENG Z, et al. Voltage-controlled skyrmion-based nanodevices for neuromorphic computing using a synthetic antiferromagnet[J]. Nanoscale Advances, 2020, 2(3): 1309-1317.

[42] ZHENG Q, MI Y, ZHU X, et al. Anticipative Tracking with the Short-Term Synaptic Plasticity of Spintronic Devices[J]. Physical Review Applied, 2020, 14(4): 044060.

[43] ZHANG S, LUO S, XU N, et al. A Spin–Orbit-Torque Memristive Device[J]. Advanced Electronic Materials, 2019, 5(4): 1800782.

[44] LEQUEUX S, SAMPAIO J, CROS V, et al. A magnetic synapse: multilevel spin-torque memristor with perpendicular anisotropy[J]. Scientific Reports, 2016, 6(1): 31510.

[45] SHARAD M, AUGUSTINE C, PANAGOPOULOS G, et al. Spin-Based Neuron Model With Domain-Wall Magnets as Synapse[J]. IEEE Transactions on Nanotechnology, 2012, 11(4): 843-853.

[46] KOBAYASHI K, BORDERS W A, KANAI S, et al. Sigmoidal curves of stochastic magnetic tunnel junctions with perpendicular easy axis[J]. Applied Physics Letters, 2021, 119(13): 132406.

[47] CHEN Y B, YANG X K, YAN T, et al. Voltage-Driven Adaptive Spintronic Neuron for Energy-Efficient Neuromorphic Computing[J]. Chinese Physics Letters, 2020, 37(7): 078501.

[48] DENG J, MIRIYALA V P K, ZHU Z, et al. Voltage-Controlled Spintronic Stochastic Neuron for Restricted Boltzmann Machine With Weight Sparsity[J]. IEEE Electron Device Letters, 2020, 41(7): 1102-1105.

[49] SIDDIQUI S A, DUTTA S, TANG A, et al. Magnetic Domain Wall Based Synaptic and Activation Function Generator for Neuromorphic Accelerators[J]. Nano Letters, 2020, 20(2): 1033-1040.

[50] CAI J, FANG B, ZHANG L, et al. Voltage-Controlled Spintronic Stochastic Neuron Based on a Magnetic Tunnel Junction[J]. Physical Review Applied, 2019, 11(3): 034015.

[51] OSTWAL V, DEBASHIS P, FARIA R, et al. Spin-torque devices with hard axis initialization as Stochastic Binary Neurons[J]. Scientific Reports, 2018, 8(1): 16689.

[52] LIYANAGEDERA C M, SENGUPTA A, JAISWAL A, et al. Stochastic Spiking Neural Networks Enabled by Magnetic Tunnel Junctions: From Nontelegraphic to Telegraphic Switching Regimes[J]. Physical Review Applied, 2017, 8(6): 064017.

[53] SHIM Y, CHEN S, SENGUPTA A, et al. Stochastic Spin-Orbit Torque Devices as Elements for Bayesian Inference[J]. Scientific Reports, 2017, 7(1): 14101.

[54] CAMSARI K Y, FARIA R, SUTTON B M, et al. Stochastic p -Bits for Invertible Logic[J]. Physical Review X, 2017, 7(3): 031014.

[55] SENGUPTA A, PARSA M, HAN B, et al. Probabilistic Deep Spiking Neural Systems Enabled by Magnetic Tunnel Junction[J]. IEEE Transactions on Electron Devices, 2016, 63(7):

2963-2970.

[56] BEHIN-AEIN B, DIEP V, DATTA S. A building block for hardware belief networks[J]. Scientific Reports, 2016, 6(1): 29893.

[57] CAI J, ZHANG L, FANG B, et al. Sparse neuromorphic computing based on spin-torque diodes[J]. Applied Physics Letters, 2019, 114(19): 192402.

[58] IOFFE S, SZEGEDY C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift[C]//Proceedings of the 32nd International Conference on Machine Learning. PMLR, 2015: 448-456.

[59] MCCULLOCH W S, PITTS W. A logical calculus of the ideas immanent in nervous activity [J]. The bulletin of mathematical biophysics, 1943, 5(4): 115-133.

[60] AUNET S, OELMANN B, ABDALLA S, et al. Reconfigurable subthreshold CMOS perceptron[C]//2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541): vol. 3. 2004: 1983-1988 vol.3.

[61] LENNIE P. The Cost of Cortical Computation[J]. Current Biology, 2003, 13(6): 493-497.

[62] BAÑUELOS-SAUCEDO M A, CASTILLO-HERNÁNDEZ J, QUINTANA-THIERRY S, et al. Implementation of a neuron model using FPGAS[J]. Journal of Applied Research and Technology, 2003, 1(3).

[63] JEYANTHI S, SUBADRA M. Implementation of single neuron using various activation functions with FPGA[C]//2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies. 2014: 1126-1131.

[64] HIKAWA H. A digital hardware pulse-mode neuron with piecewise linear activation function[J]. IEEE Transactions on Neural Networks, 2003, 14(5): 1028-1037.

[65] TSAI C H, CHIH Y T, WONG W H, et al. A Hardware-Efficient Sigmoid Function With Adjustable Precision for a Neural Network System[J]. IEEE Transactions on Circuits and Systems II: Express Briefs, 2015, 62(11): 1073-1077.

[66] BAPTISTA D, MORGADO-DIAS F. Low-resource hardware implementation of the hyperbolic tangent for artificial neural networks[J]. Neural Computing and Applications, 2013, 23(3): 601-607.

[67] HE K, ZHANG X, REN S, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1026-1034.

[68] KLAMBAUER G, UNTERTHINER T, MAYR A, et al. Self-normalizing neural networks [J]. Advances in neural information processing systems, 2017, 30.

[69] RAMACHANDRAN P, ZOPH B, LE Q V. Searching for activation functions[J]. arXiv preprint arXiv:1710.05941, 2017.

[70] AGOSTINELLI F, HOFFMAN M, SADOWSKI P, et al. Learning activation functions to improve deep neural networks[J]. arXiv preprint arXiv:1412.6830, 2014.

[71] APICELLA A, DONNARUMMA F, ISGRÒ F, et al. A survey on modern trainable activation functions[J]. Neural Networks, 2021, 138: 14-32.

[72] APALKOV D, DIENY B, SLAUGHTER J M. Magnetoresistive Random Access Memory [J]. Proceedings of the IEEE, 2016, 104(10): 1796-1830.

[73] LI Z, ZHANG S. Magnetization dynamics with a spin-transfer torque[J]. Physical Review B, 2003, 68(2): 024404.

[74] RAMASWAMY R, LEE J M, CAI K, et al. Recent advances in spin-orbit torques: Moving towards device applications[J]. Applied Physics Reviews, 2018, 5(3): 031107.

[75] MAHFOUZI F, MISHRA R, CHANG P H, et al. Microscopic origin of spin-orbit torque in ferromagnetic heterostructures: A first-principles approach[J]. Physical Review B, 2020, 101(6): 060405.

[76] VIGNALE G. Ten Years of Spin Hall Effect[J]. Journal of Superconductivity and Novel Magnetism, 2010, 23(1): 3-10.

[77] KOO H C, KIM S B, KIM H, et al. Rashba Effect in Functional Spintronic Devices[J]. Advanced Materials, 2020, 32(51): 2002117.

[78] BERNEVIG B A, ZHANG S C. Quantum Spin Hall Effect[J]. Physical Review Letters, 2006, 96(10): 106802.

[79] SHAO Q, LI P, LIU L, et al. Roadmap of Spin−Orbit Torques[J]. IEEE Transactions on Magnetics, 2021, 57(7): 1-39.

[80] DAI B, JACKSON M, CHENG Y, et al. Review of voltage-controlled magnetic anisotropy and magnetic insulator[J]. Journal of Magnetism and Magnetic Materials, 2022, 563: 169924.

[81] RANA B, OTANI Y. Towards magnonic devices based on voltage-controlled magnetic anisotropy[J]. Communications Physics, 2019, 2(1): 90.

[82] JULLIERE M. Tunneling between ferromagnetic films[J]. Physics Letters A, 1975, 54(3): 225-226.

[83] ZUO S, FAN H, NAZARPOUR K, et al. A CMOS Analog Front-End for Tunnelling Magnetoresistive Spintronic Sensing Systems[C]//2019 IEEE International Symposium on Circuits and Systems (ISCAS). Sapporo, Japan: IEEE, 2019: 1-5.

[84] CARPENTER M H, KENNEDY C A. Fourth-order 2N-storage Runge-Kutta schemes[R]. NASA-TM-109112. NTRS Author Affiliations: NASA Langley Research Center NTRS Document ID: 19940028444 NTRS Research Center: Legacy CDMS (CDMS). 1994.

[85] 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[J]. Applied Physics Letters, 2015, 106(7): 072402.

[86] CAI K, VAN BEEK S, RAO S, et al. Selective operations of multi-pillar SOT-MRAM for high density and low power embedded memories[C]//2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits). Honolulu, HI, USA: IEEE, 2022: 375-376.

[87] YAMAMOTO T, NOZAKI T, IMAMURA H, et al. Write-Error Reduction of Voltage-Torque-Driven Magnetization Switching by a Controlled Voltage Pulse[J]. Physical Review Applied, 2019, 11(1): 014013.

[88] KANG W, CHANG L, ZHANG Y, et al. Voltage-controlled MRAM for working memory: Perspectives and challenges[C]//Design, Automation & Test in Europe Conference & Exhibition (DATE), 2017. Lausanne, Switzerland: IEEE, 2017: 542-547.

[89] YODA H, SHIMOMURA N, OHSAWA Y, et al. Voltage-control spintronics memory (VoCSM) having potentials of ultra-low energy-consumption and high-density[C]//2016 IEEE International Electron Devices Meeting (IEDM). San Francisco, CA, USA: IEEE, 2016: 27.6.1-27.6.4.

[90] 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[J]. Applied Physics Letters, 2016, 108(1): 012403.

[91] KANAI S, NAKATANI Y, YAMANOUCHI M, et al. Magnetization switching in a CoFeB/MgO magnetic tunnel junction by combining spin-transfer torque and electric field-effect[J]. Applied Physics Letters, 2014, 104(21): 212406.

[92] KANAI S, YAMANOUCHI M, IKEDA S, et al. Electric field-induced magnetization reversal in a perpendicular-anisotropy CoFeB-MgO magnetic tunnel junction[J]. Applied Physics Letters, 2012, 101(12): 122403.

[93] WANG W G, LI M, HAGEMAN S, et al. Electric-field-assisted switching in magnetic tunnel junctions[J]. Nature Materials, 2012, 11(1): 64-68.

[94] ALZATE J G, AMIRI P K, UPADHYAYA P, et al. Voltage-induced switching of nanoscale magnetic tunnel junctions[C]//2012 International Electron Devices Meeting. San Francisco, CA, USA: IEEE, 2012: 29.5.1-29.5.4.

[95] SHIOTA Y, NOZAKI T, BONELL F, et al. Induction of coherent magnetization switching in a few atomic layers of FeCo using voltage pulses[J]. Nature Materials, 2012, 11(1): 39-43.

[96] MARUYAMA T, SHIOTA Y, NOZAKI T, et al. Large voltage-induced magnetic anisotropy change in a few atomic layers of iron[J]. Nature Nanotechnology, 2009, 4(3): 158-161.

[97] NAKAMURA K, SHIMABUKURO R, FUJIWARA Y, et al. Giant Modification of the Magnetocrystalline Anisotropy in Transition-Metal Monolayers by an External Electric Field[J]. Physical Review Letters, 2009, 102(18): 187201.

[98] TSUJIKAWA M, ODA T. Finite Electric Field Effects in the Large Perpendicular Magnetic Anisotropy Surface Pt / Fe / Pt (001) : A First-Principles Study[J]. Physical Review Letters, 2009, 102(24): 247203.

[99] DUAN C G, VELEV J P, SABIRIANOV R F, et al. Surface Magnetoelectric Effect in Ferromagnetic Metal Films[J]. Physical Review Letters, 2008, 101(13): 137201.

[100] WEISHEIT M, FÃHLER S, MARTY A, et al. Electric field-induced modification of magnetism in thin-film ferromagnets[J]. Science, 2007, 315(5810): 349-351.

[101] ZHU Z, DENG J, FONG X, et al. Voltage-input spintronic oscillator based on competing effect for extended oscillation regions[J]. Journal of Applied Physics, 2019, 125(18): 183902.

[102] GILBERT T. Classics in Magnetics A Phenomenological Theory of Damping in Ferromagnetic Materials[J]. IEEE Transactions on Magnetics, 2004, 40(6): 3443-3449.

[103] SLONCZEWSKI J. Current-driven excitation of magnetic multilayers[J]. Journal of Magnetism and Magnetic Materials, 1996, 159(1): L1-L7.

[104] BERGER L. Emission of spin waves by a magnetic multilayer traversed by a current[J]. Physical Review B, 1996, 54(13): 9353-9358.

[105] IKEDA S, MIURA K, YAMAMOTO H, et al. A perpendicular-anisotropy CoFeB−MgO magnetic tunnel junction[J]. Nature Materials, 2010, 9(9): 721-724.

[106] 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]. Journal of Applied Physics, 2009, 105(7): 07D131.

[107] DAS D, FONG X. A Fokker–Planck Approach for Modeling the Stochastic Phenomena in Magnetic and Resistive Random Access Memory Devices[J]. IEEE Transactions on Electron Devices, 2021, 68(12): 6124-6131.

[108] MIRIYALA V P K, FONG X, LIANG G. Influence of Size and Shape on the Performance of VCMA-Based MTJs[J]. IEEE Transactions on Electron Devices, 2019, 66(2): 944-949.

[109] XIE Y, BEHIN-AEIN B, GHOSH A W. Fokker—Planck Study of Parameter Dependence on Write Error Slope in Spin-Torque Switching[J]. IEEE Transactions on Electron Devices, 2017, 64(1): 319-324.

[110] TANIGUCHI T, UTSUMI Y, MARTHALER M, et al. Spin torque switching of an in-plane magnetized system in a thermally activated region[J]. Physical Review B, 2013, 87(5): 054406.

[111] NEWHALL K A, VANDEN-EIJNDEN E. Averaged equation for energy diffusion on a graph reveals bifurcation diagram and thermally assisted reversal times in spin-torque driven nanomagnets[J]. Journal of Applied Physics, 2013, 113(18): 184105.

[112] BUTLER W H, MEWES T, MEWES C K A, et al. Switching Distributions for Perpendicular Spin-Torque Devices Within the Macrospin Approximation[J]. IEEE Transactions on Magnetics, 2012, 48(12): 4684-4700.

[113] TANIGUCHI T, IMAMURA H. Thermally assisted spin transfer torque switching in synthetic free layers[J]. Physical Review B, 2011, 83(5): 054432.

[114] BEDAU D, LIU H, SUN J Z, et al. Spin-transfer pulse switching: From the dynamic to the thermally activated regime[J]. Applied Physics Letters, 2010, 97(26): 262502.

[115] APALKOV D M, VISSCHER P B. Spin-torque switching: Fokker-Planck rate calculation [J]. Physical Review B, 2005, 72(18): 180405.

[116] LI Z, ZHANG S. Thermally assisted magnetization reversal in the presence of a spin-transfer torque[J]. Physical Review B, 2004, 69(13): 134416.

[117] URAZHDIN S, BIRGE N O, PRATT JR W P, et al. Current-driven magnetic excitations in permalloy-based multilayer nanopillars[J]. Physical review letters, 2003, 91(14): 146803.

[118] MYERS E B, ALBERT F J, SANKEY J C, et al. Thermally Activated Magnetic Reversal Induced by a Spin-Polarized Current[J]. Physical Review Letters, 2002, 89(19): 196801.

[119] ALBERT F J, EMLEY N C, MYERS E B, et al. Quantitative Study of Magnetization Reversal by Spin-Polarized Current in Magnetic Multilayer Nanopillars[J]. Physical Review Letters, 2002, 89(22): 226802.

[120] BROWN W F. Thermal Fluctuations of a Single-Domain Particle[J]. Physical Review, 1963, 130(5): 1677-1686.

[121] SUN J Z. Spin-current interaction with a monodomain magnetic body: A model study[J]. Physical Review B, 2000, 62(1): 570-578.

[122] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[123] GLOROT X, BENGIO Y. Understanding the difficulty of training deep feedforward neural networks[C]//Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. JMLR Workshop, 2010: 249-256.

[124] SHARAD M, FAN D, ROY K. Spin-neurons: A possible path to energy-efficient neuromorphic computers[J]. Journal of Applied Physics, 2013, 114(23): 234906.

[125] RAMASUBRAMANIAN S G, VENKATESAN R, SHARAD M, et al. SPINDLE: SPINtronic deep learning engine for large-scale neuromorphic computing[C]//Proceedings of the 2014 international symposium on Low power electronics and design. La Jolla California USA: ACM, 2014: 15-20.

[126] MEHONIC A, KENYON A J. Brain-inspired computing needs a master plan[J]. Nature, 2022, 604(7905): 255-260.

[127] MEAD C. How we created neuromorphic engineering[J]. Nature Electronics, 2020, 3(7): 434-435.

[128] SUN X, YU S. Impact of non-ideal characteristics of resistive synaptic devices on implementing convolutional neural networks[J]. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 2019, 9(3): 570-579.

[129] AGARWAL S, PLIMPTON S J, HUGHART D R, et al. Resistive memory device requirements for a neural algorithm accelerator[C]//2016 International Joint Conference on Neural Networks (IJCNN). IEEE, 2016: 929-938.

[130] ROY S, SRIDHARAN S, JAIN S, et al. Txsim: Modeling training of deep neural networks on resistive crossbar systems[J]. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 2021, 29(4): 730-738.

[131] YU Z, LEROUX N, NEFTCI E. Training-to-Learn with Memristive Devices[C]//2022 International Electron Devices Meeting (IEDM). 2022: 21.1.1-21.1.4.

[132] LI B, LI Y, RONG X. The extreme learning machine learning algorithm with tunable activation function[J]. Neural Computing and Applications, 2013, 22: 531-539.

[133] MIAO P, SHEN Y, XIA X. Finite time dual neural networks with a tunable activation function for solving quadratic programming problems and its application[J]. Neurocomputing, 2014, 143: 80-89.

[134] BISWAS A, CHANDRAKASAN A P. CONV-SRAM: An energy-efficient SRAM with in-memory dot-product computation for low-power convolutional neural networks[J]. IEEE Journal of Solid-State Circuits, 2018, 54(1): 217-230.

[135] VALAVI H, RAMADGE P J, NESTLER E, et al. A 64-tile 2.4-Mb in-memory-computing CNN accelerator employing charge-domain compute[J]. IEEE Journal of Solid-State Circuits, 2019, 54(6): 1789-1799.

[136] WANG P, XU F, WANG B, et al. Three-dimensional NAND flash for vector−matrix multiplication[J]. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 2018, 27(4): 988-991.

[137] SONG J, CHO Y, PARK J S, et al. 7.1 An 11.5 TOPS/W 1024-MAC butterfly structure dual-core sparsity-aware neural processing unit in 8nm flagship mobile SoC[C]//2019 IEEE International Solid-State Circuits Conference-(ISSCC). IEEE, 2019: 130-132.

[138] KECKLER S W, DALLY W J, KHAILANY B, et al. GPUs and the future of parallel computing[J]. IEEE micro, 2011, 31(5): 7-17.

[139] JUNG S, LEE H, MYUNG S, et al. A crossbar array of magnetoresistive memory devices for in-memory computing[J]. Nature, 2022, 601(7892): 211-216.

[140] PATIL A D, HUA H, GONUGONDLA S, et al. An MRAM-based deep in-memory architecture for deep neural networks[C]//2019 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE, 2019: 1-5.

[141] YANG S, SHIN J, KIM T, et al. Integrated neuromorphic computing networks by artificial spin synapses and spin neurons[J]. NPG Asia Materials, 2021, 13(1): 11.

[142] CHANTHBOUALA A, MATSUMOTO R, GROLLIER J, et al. Vertical-current-induced domain-wall motion in MgO-based magnetic tunnel junctions with low current densities [J]. Nature Physics, 2011, 7(8): 626-630.

[143] LIU J, XU T, FENG H, et al. Compensated Ferrimagnet Based Artificial Synapse and Neuron for Ultrafast Neuromorphic Computing[J]. Advanced Functional Materials, 2022, 32(1): 2107870.

[144] SUNG S H, KIM T J, SHIN H, et al. Simultaneous emulation of synaptic and intrinsic plasticity using a memristive synapse[J]. Nature Communications, 2022, 13(1): 2811.

[145] HE Q L, HUGHES T L, ARMITAGE N P, et al. Topological spintronics and magnetoelectronics[J]. Nature materials, 2022, 21(1): 15-23.

[146] FAN Y, KOU X, UPADHYAYA P, et al. Electric-field control of spin−orbit torque in a magnetically doped topological insulator[J]. Nature nanotechnology, 2016, 11(4): 352-359.

[147] WANG Y, YANG H. Spin−Orbit Torques Based on Topological Materials[J]. Accounts of Materials Research, 2022, 3(10): 1061-1072.

[148] SHAO Q, WU H, PAN Q, et al. Room Temperature Highly Efficient Topological Insulator/Mo/CoFeB Spin-Orbit Torque Memory with Perpendicular Magnetic Anisotropy[C]//2018 IEEE International Electron Devices Meeting (IEDM). 2018: 36.3.1-36.3.4.

[149] ZAND R, CAMSARI K Y, DATTA S, et al. Composable probabilistic inference networks using MRAM-based stochastic neurons[J]. ACM Journal on Emerging Technologies in Computing Systems (JETC), 2019, 15(2): 1-22.

[150] JOSHI V, LE GALLO M, HAEFELI S, et al. Accurate deep neural network inference using computational phase-change memory[J]. Nature communications, 2020, 11(1): 2473.

[151] JANG J W, PARK S, BURR G W, et al. Optimization of conductance change in Pr 1−x Ca x MnO 3-based synaptic devices for neuromorphic systems[J]. IEEE Electron Device Letters, 2015, 36(5): 457-459.

[152] HUANG B, CLARK G, NAVARRO-MORATALLA E, et al. Layer-dependent ferromagnetism in a van der Waals crystal down to the monolayer limit[J]. Nature, 2017, 546(7657): 270-273.

[153] DENG Y, YU Y, SONG Y, et al. Gate-tunable room-temperature ferromagnetism in two-dimensional Fe3GeTe2[J]. Nature, 2018, 563(7729): 94-99.

[154] ZHANG D, YANG J, YE D, et al. Lq-nets: Learned quantization for highly accurate and compact deep neural networks[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 365-382.

[155] ZHUANG B, SHEN C, TAN M, et al. Towards effective low-bitwidth convolutional neural

networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7920-7928.

[156] LEE S, JEON J, EOM K, et al. Variance-aware weight quantization of multi-level resistive switching devices based on Pt/LaAlO3/SrTiO3 heterostructures[J]. Scientific Reports, 2022, 12(1): 9068.

[157] NGUYEN V T, TRINH Q K, ZHANG R, et al. STT-BSNN: An In-Memory Deep Binary Spiking Neural Network Based on STT-MRAM[J]. IEEE Access, 2021, 9: 151373-151385.

[158] ABBOTT L F, DEPASQUALE B, MEMMESHEIMER R M. Building functional networks of spiking model neurons[J]. Nature Neuroscience, 2016, 19(3): 350-355.

[159] LIM H, KORNIJCUK V, SEOK J Y, et al. Reliability of neuronal information conveyed by unreliable neuristor-based leaky integrate-and-fire neurons: a model study[J]. Scientific Reports, 2015, 5(1): 9776.

致 谢

翻开我研究生的画卷，故事从我来到上科大的那一天展开，从开始接触科研到发表文章，身边的每一个人都给予了我非常多的帮助。

首先，我要郑重感谢我的导师，祝智峰教授，他带我走进了自旋电子学的大门，在自旋电子学和类脑计算这一交叉学科，我们共同学习，一起进步。在研究期间，是他帮助我进行方案的完善和验证，祝智峰老师总是很耐心，他会在我研一的时候充分支持我略显幼稚的观点。他会和我们谈心，倾听我们的迷茫和困惑，用他的亲身经历鼓舞我们继续在自己的课题深耕。我想我之后总会经常想起他的教诲，他的出现一定是我人生篇章上浓墨重彩的一笔。

我很感谢任杰、龙靖威、张伟岩、张雪、袁正平、徐正德、安丽华、乔怡骁和许焯等实验室的同门们，我们共同钻研课题，探讨学术知识，一起快乐成长。他们每一个人都是我研究生画卷上多彩的颜色。

我特别要感谢和我一起做课题的黄浦阳、蔡葆昉、赫一涵、李宇、安丽华、袁正平和许焯，没有他们，我无法完成现在的这些科研成果。我和他们一起为截稿日期熬过夜，一起为数据结果而头脑风暴，和他们并肩战斗的日子，在未来一定会成为我成长进步的沃土。

我还要感谢信息学院的哈亚军、寇煦丰、刘思廷、林丰涵、杨雨梦、任豪和梁俊睿教授们，他们给予了真诚亲切友好的帮助。还要感谢后摩尔中心的学长们，他们在课业上对我帮助良多，我衷心感谢与他们的相遇。

上科大的夏天很明媚，冬天很温暖，春天和秋天虽然短暂却绚烂多姿，能够在这里遇到大家是我人生中非常奇妙的一个篇章。

作者简历及攻读学位期间发表的学术论文与研究成果

作者简历：

2016 年 9 月至 2020 年 6 月，于上海理工大学光电信息与计算机工程学院获得学士学位。

2020 年 9 月至 2023 年 6 月，于上海科技大学信息科学与技术学院获得硕士学位。

已发表（或正式接受）的学术论文：

Yue Xin, Kang Zhou, Xuanyao Fong, Yumeng Yang, Shenghua Gao, Zhifeng Zhu, ”Electrical Tunable Spintronic Neuron with Trainable Activation Function,” 2022, arXiv preprint arXiv:2211.13391

Puyang Huang#, Xinqi Liu#, Yue Xin#, Liyang Liao, Qi Yao, Peng Chen, Yu Zhang, Weijie Deng, Guoqiang Yu, Zhongkai Liu, Yumeng Yang, Zhifeng Zhu, and Xufeng Kou, ”Integrated Artificial Neural Network with Trainable Activation Function Enabled by Topological Insulator-based Spin-Orbit Torque Devices,” 2022, arXiv preprint arXiv:2209.06001

Baofang Cai#, Yihan He#, Yue Xin#, Zhengping Yuan, Xue Zhang, Zhifeng Zhu, Gengchiau Liang, ”Unconventional Computing Based on Magnetic Tunnel Junction,” Applied Physics A, 2023, https://doi.org/10.1007/s00339-022-06365-4

Lihua An#, Yue Xin#, Zhengping Yuan, Yumeng Yang, and Zhifeng Zhu, ”Study of Write Error Rate in MRAM with Fixed Voltage Input,”in International Conference on Solid State Devices and Materials(SSDM),Chiba, Japan, July 2022.

Jingwei Long, Qi Hu, Zhengping Yuan, Yunsen Zhang, Yue Xin, Jie Ren, Bowen Dong, Gengfei Li, Yumeng Yang, Huihui Li, Zhifeng Zhu, ”Comparative Study of Temperature Impact in Spin-Torque Switched Perpendicular and Easy-Cone MTJs,” Nanomaterials, 2023, https://doi.org/10.3390/nano13020337

Zhengping Yuan, Jingwei Long, Yue Xin, Zhengde Xu, Lihua An, Jie Ren, Xue Zhang,

Yumeng Yang, and Zhifeng Zhu,"Anomalous impact of thermal fluctuations on spin transfer torque induced ferrimagnetic switching," Journal of Applied Physics, 2023, https://doi.org/10.1063/5.0144468

Zhengde Xu#, Jie Ren#, Zhengping Yuan,Yue Xin, Xue Zhang, Shuyuan Shi, Yumeng Yang, Zhifeng Zhu,"Field-free spin-orbit torque switching of an antiferromagnet with perpendicular Néel vector," Journal of Applied Physics, 2023,
https://doi.org/10.1063/5.0138869

Yue Xin, Kang Zhou,Yumeng Yang, Shenghua Gao, Zhifeng Zhu,"Spintronic Neuron with Trainable Activation Function,"in 第十九届全国磁学和磁性材料会议, Hainan, China, Nov 2021.

Zhengde Xu, Jie Ren,Zhengping Yuan, Yue Xin, Xue Zhang, Shuyuan Shi, Yumeng Yang, Zhifeng Zhu," 垂直反铁磁体的无场自旋轨道力矩翻转," in 中国物理学会 2022 秋季学术会议, Shenzhen, China, Oct 2022.

Zhengping Yuan, Yue Xin, Lihua An, Zhengde Xu, Xue Zhang, Jingwei Long, Jie Ren, Zhifeng Zhu," 基于双亚晶格宏自旋模型的自旋力矩驱动亚铁磁动力学研究,"in 中国物理学会 2022 秋季学术会议, Shenzhen, China, Oct 2022.