DOI:10.3969/j.issn.1003-5060.2024.02.007

基于忆阻器的门控循环单元电路

韩婷婷， 张 章， 陈思锴
(合肥工业大学 微电子学院,安徽 合肥 230601)

摘 要:时间序列数据分析可用于识别长期趋势并进行正确的预测,与人工神经网络(artificial neural network,ANN)相比,门控循环单元(gated recurrent unit,GRU)可以处理时间序列信号,在自然语言处理、语音识别、机器翻译等方面有着广泛的应用。然而,由于参数和模型的复杂性,GRU 模型在硬件实现中遇到了瓶颈。文章构建一个基于忆阻器的 GRU 硬件电路,具有完整的 GRU 功能,而且输入/输出参数更少。仿真结果表明,电路的平均误差为 0.007 5,能够有效地实现 GRU 网络的功能。将设计的 GRU 电路应用在搭建的序列预测模型中,可以预测股票价格变化趋势,且其预测的 R2 分数达到 0.923 4。因此基于忆阻器的 GRU 硬件电路的设计在机器学习和人工智能方面具有一定的应用潜力。

关键词:忆阻器;循环神经网络(RNN);门控循环单元(GRU);序列预测
中图分类号:TP389.1;TN43.1 **文献标志码**:A **文章编号**:1003-5060(2024)02-0189-06

Gated recurrent unit circuit based on memristor

HAN Tingting, ZHANG Zhang, CHEN Sikai
(School of Microelectronics, Hefei University of Technology, Hefei 230601, China)

Abstract: Making accurate predictions and identifying long-term patterns are both possible through the examination of time series data. Gated recurrent unit(GRU), which can process time series signals in comparison to artificial neural network(ANN), is frequently employed in natural language processing, speech recognition, machine translation, etc. The intricacy of the parameters and model, however, has caused a hardware implementation bottleneck for the GRU model. In this study, a memristor-based GRU with full circuit functionality and fewer input/output parameters is built. According to the simulation findings, the average error of the circuit is 0.007 5, effectively realizing the purpose of GRU network. The GRU circuit design is applied to the sequence prediction model, which can forecast the trend of stock price fluctuations, and the anticipated R2 score reaches 0.923 4. The construction of GRU hardware circuits based on memristor offers potential for use in artificial intelligence and machine learning.

Key words: memristor; recurrent neural network(RNN); gated recurrent unit(GRU); sequence prediction

深度神经网络(deep neural network,DNN)模型在解决挑战性任务时的突出表现,使其在语音信号处理[1]、数字图像处理[2]领域受到广泛应用。因为前馈神经网络要求输入和输出维度固定,所以其不适合处理长度不固定的时序数据。循环神经网络(recurrent neural network,RNN)是具有短期记忆能力的一类网络,其具有反馈连接以保留信息的顺序,并且能处理任意长度的输

收稿日期:2022-11-16;**修回日期**:2022-12-12
基金项目:国家自然科学基金区域创新发展联合基金资助项目(U19A2053)
作者简介:韩婷婷(1997—),女,安徽阜阳人,合肥工业大学硕士生;
张 章(1982—),男,安徽淮南人,博士,合肥工业大学教授,博士生导师,通信作者,E-mail:zhangzhang@hfut.edu.cn.

入序列数据,因此被广泛应用在语音识别、语言模型以及自然语言生成等任务上。然而,对 RNN 的训练可能难以进行,因为 RNN 的权重更新是基于梯度算法进行的,所以会有梯度爆炸或消失的问题,这些问题可以通过在传统 RNN 中添加门控机制来解决。长短时记忆网络(long short-term memory,LSTM)[3]和门控循环单元(gated recurrent unit,GRU)[4]是目前广泛应用的2个基于门控的循环神经网络(gated recurrent neural network,Gated RNN)。GRU 是受 LSTM 模型启发,具有更少的参数和更简单的计算步骤。文献[5]对比了 GRU 与 LSTM 在相同任务中的精度、内存消耗和训练时间,发现 GRU 与 LSTM 精度几乎相同,但是 GRU 的内存消耗相比于 LSTM 减少了约 18%,训练速度提升了约 20%,这使得 GRU 在对内存或速度有特殊要求的实际应用中更加流行。

然而,GRU 由于其结构的复杂性和并行性,需要大量的计算资源,传统冯诺依曼体系架构中存在的"冯诺依曼瓶颈"[6]和"内存墙瓶颈"[7]问题导致其实现起来有一定难度。为了推动未来神经形态计算技术的发展,必须找到非冯诺依曼体系架构的解决方案。忆阻器是代表磁通量与电荷之间关系的两端电路器件,是存算一体架构的有力候选者。与冯诺依曼体系架构相比,基于忆阻器的内存计算有效地规避了频繁数据通信带来的巨大能耗和时间开销。

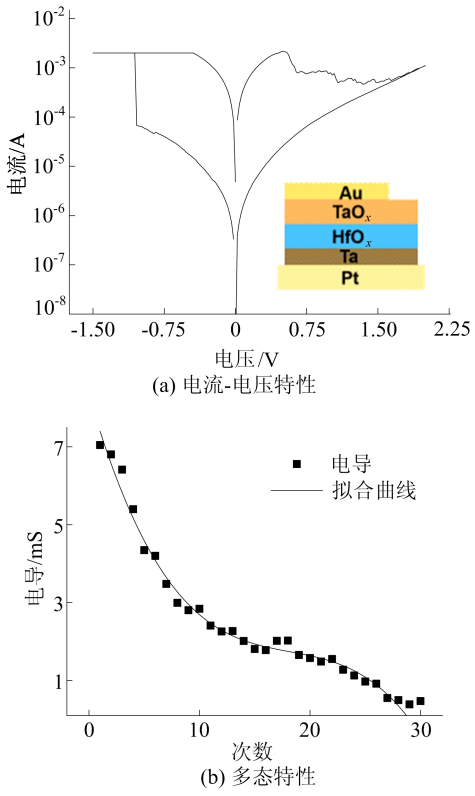
人工神经网络(artificial neural network,ANN)的整体内存计算架构将由多个交叉开关阵列组成,忆阻器可以使用电导值作为突触权重,交叉开关阵列执行向量矩阵乘法(vector matrix multiplication,VMM)。在馈入行的每个输入电压向量和列权重向量之间生成点积向量,每个交叉开关阵列都带有模数转换器和数字电路,然后连接到下一个交叉开关阵列。然而,先前大部分对 GRU 的研究都是基于软件实现的,对 GRU 存算一体的硬件研究介绍很少,并且没有任何使用忆阻器实现 GRU 单元计算的研究。

因此,本文提出一个基于忆阻器的门控循环单元的硬件电路,使用较少的元件以及低精度要求的元件就可以实现门控循环单元的功能。本文首先对搭建的电路进行仿真,结果表明电路的平均误差为 0.007 5;然后将电路应用于 3 层 GRU 模型实现股票的预测,预测的 R2 分数为 0.923 4。本文的设计不仅具有较高的精度,而且具有更小的片上预算和功率消耗。

1 忆阻器性能及其模型

忆阻器为应用非冯诺依曼架构的存内计算集成技术提供了新思路。忆阻器可以用作仿生突触设备,通过动态调整电阻状态来实现权重更新。同时,忆阻器阵列可以根据基尔霍夫定律和欧姆定律实现 VMM 操作,但是使用忆阻器阵列进行 GRU 操作,在每次迭代操作中存在海量数据传输。因此目前还缺乏使用忆阻器来实现 RNN 网络功能的方案。

为实现更切实际的电路仿真,本文利用真实的忆阻器性能进行建模。使用 Au/Ta_2O_5/HfO_2/Pt 叠层用作阻变式存储器(resistive random access memory,RRAM)器件,忆阻器的电流-电压特性和多态特性如图 1 所示。



(a) 电流-电压特性

(b) 多态特性

图 1 忆阻器的电流-电压和多态特性

通过施加 0～－1.5 V 的扫描电压,器件可以从高电阻状态(high resistance state,HRS)变为低电阻状态(low resistance state,LRS)。当施加 0～2.0 V 的扫描电压时,器件可以从 LRS 变为 HRS。此外,本文对 200 个工作电压进行统计,设定电压集中在 0.7～－1.1 V 的范围,复位电压集中在－0.7 ～ 0.4 V 的范围。分布式和集中式

的式工作电压特性有利于后期电路设计。通过改变复位过程的截止电压，可以实现忆阻器的多态调节。本文使用这种方法实现了 64 个状态的连续调节，电导调节的线性关系（图 1b）表明该装置满足实际应用的要求。

对器件的特性进行函数拟合，并将其数学表达式转换为相应的电路元件。将不可转换的部分作为函数方程合并到 Spice 模型中以拟合器件数据，因此新模型具有器件属性。在重置过程中提取一条完整的曲线，如图 2 所示。从图 2 可以看出，器件曲线和模型拟合曲线几乎完全重合。因此，该模型满足多态调控仿真，能够很好地模拟真实忆阻器在后续神经网络构建中的工作状态。

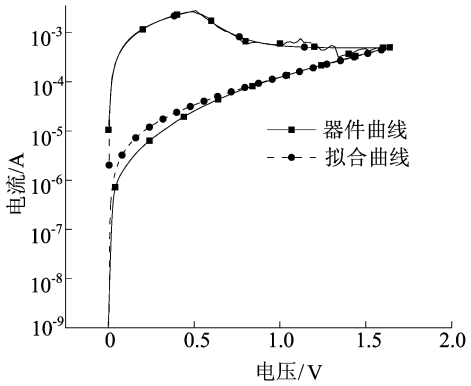


图 2　忆阻器模型的拟合曲线

2　基于忆阻器的门控循环单元电路

2.1　门控循环单元

GRU 是一种结构更简单的门控循环神经网络，其重置门能同时进行遗忘和记忆，并且没有额外的记忆单元，如图 3 所示。从图 3 可以看出，其输入分别为当前时刻输入 $\boldsymbol{x}_t$ 和上一时刻输出的 $\boldsymbol{h}_{t-1}$。输入分别乘以相应的权重，然后通过激活函数输出门控值。sigmoid 激活层后的更新门 $\boldsymbol{z}_t$ 值和重置门 $\boldsymbol{r}_t$ 值的范围都为（0，1）。$\boldsymbol{r}_t$ 和上一时刻的状态 $\boldsymbol{h}_{t-1}$ 相乘用来控制候选状态 $\boldsymbol{h}'$ 对前一时刻 $\boldsymbol{h}_{t-1}$ 的状态进行遗忘。如果重置门近似 0，那么上一个隐藏状态将被丢弃。因此，重置门可以丢弃与预测未来无关的历史信息，通过 tanh 函数过滤可以使输出范围为（－1，1）。$\boldsymbol{z}_t$ 更新门用来控制此刻的时间步 $\boldsymbol{h}_t$ 如何被上一时刻时间步信息 $\boldsymbol{h}_{t-1}$ 和候选激活信息 $\boldsymbol{h}'$ 所更新，若当前时刻输出 $\boldsymbol{h}_t$ 和上一时刻输出 $\boldsymbol{h}_{t-1}$ 之间的关系是线性函数，则这种设计可以应对循环神经网络中的梯度衰减问题，并更好地捕捉时间序列中时间步距离较大的依赖关系。

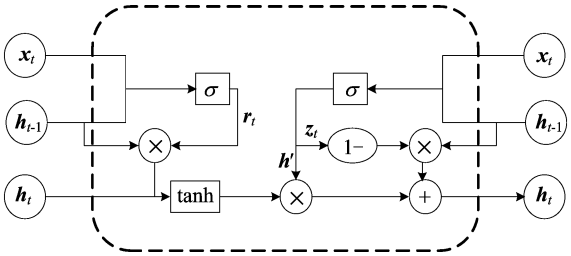


图 3　门控循环单元的结构

GRU 单元操作公式如下：

$$\boldsymbol{z}_t = \sigma(\boldsymbol{W}_z\boldsymbol{x}_t + \boldsymbol{U}_z\boldsymbol{h}_{t-1} + \boldsymbol{b}_z) \tag{1}$$

$$\boldsymbol{r}_t = \sigma(\boldsymbol{W}_r\boldsymbol{x}_t + \boldsymbol{U}_r\boldsymbol{h}_{t-1} + \boldsymbol{b}_r) \tag{2}$$

$$\boldsymbol{h}' = \tanh(\boldsymbol{W}\boldsymbol{x}_t + \boldsymbol{U}(\boldsymbol{r}_t \odot \boldsymbol{h}_{t-1}) + \boldsymbol{b}) \tag{3}$$

$$\boldsymbol{h}_t = \boldsymbol{z}_t \odot \boldsymbol{h}' + (1 - \boldsymbol{z}_t) \odot \boldsymbol{h}_{t-1} \tag{4}$$

其中：$\boldsymbol{x}_t$ 为时间步长为 N 的输入向量；$\boldsymbol{h}_{t-1}$ 为上一时刻的信息输入向量；$\boldsymbol{z}_t$ 为更新门输出矩阵；$\boldsymbol{r}_t$ 为重置门输出矩阵；$\boldsymbol{h}'$ 为中间激活状态输出矩阵；$\boldsymbol{h}_t$ 为当前时刻的输出；$\boldsymbol{W}_z$、$\boldsymbol{U}_z$ 为更新门的权重矩阵；$\boldsymbol{W}_r$、$\boldsymbol{U}_r$ 为重置门的权重矩阵；$\boldsymbol{W}$、$\boldsymbol{U}$ 为中间激活状态的权重矩阵；$\boldsymbol{b}_z$、$\boldsymbol{b}_r$ 分别为更新门和重置门偏置矩阵；σ 为 sigmoid 函数；$\odot$ 为逐元素向量积。

2.2　基于忆阻器的门控循环单元电路

基于忆阻器的门控循环单元的整体电路如图 4所示，硬件电路的设计是基于上述 GRU 单元的详细参数计算过程。门控单元的权重可以由忆阻器电导表示，忆阻器的电导可以通过施加的电压和脉冲时间来控制。另外，通过一个开关和反相电路来控制硬件电路中权重值的正值和负值。当权重值为负时，开关关断，输入通过反相电路用于忆阻器；当权重值为正时，开关断开，输入不经过反相电路用于忆阻器。通过控制输入的方向电路实现负值的乘法结果输出，因此本设计可以通过4 个忆阻器实现更新门和重置门内的乘法。另外，激活函数的硬件实现是设计中具有挑战性的部分。经常使用的激活函数有 sigmoid、tanh、ReLU等，硬件电路中 ReLU 函数可以用一个放大电路近似等效。受此启发，GRU 中使用的 sigmoid和 tanh 函数都可以用一个线性函数去逼近它的线性部分。

基于以上 2 个设计思想，下面介绍 GRU 内部单元的具体实现电路。

权值与电导值之间被设计为 100 倍的线性映射关系，即

$$G = W/100 \tag{5}$$

其中:G 为电导值;W 为权重值,W 的取值范围为(0,0.2]。该设计使得输出电压与理论值之间呈线性关系,将式(5)分别代入式(1)、式(2),可计算复位门和更新门电路的输出电压值。

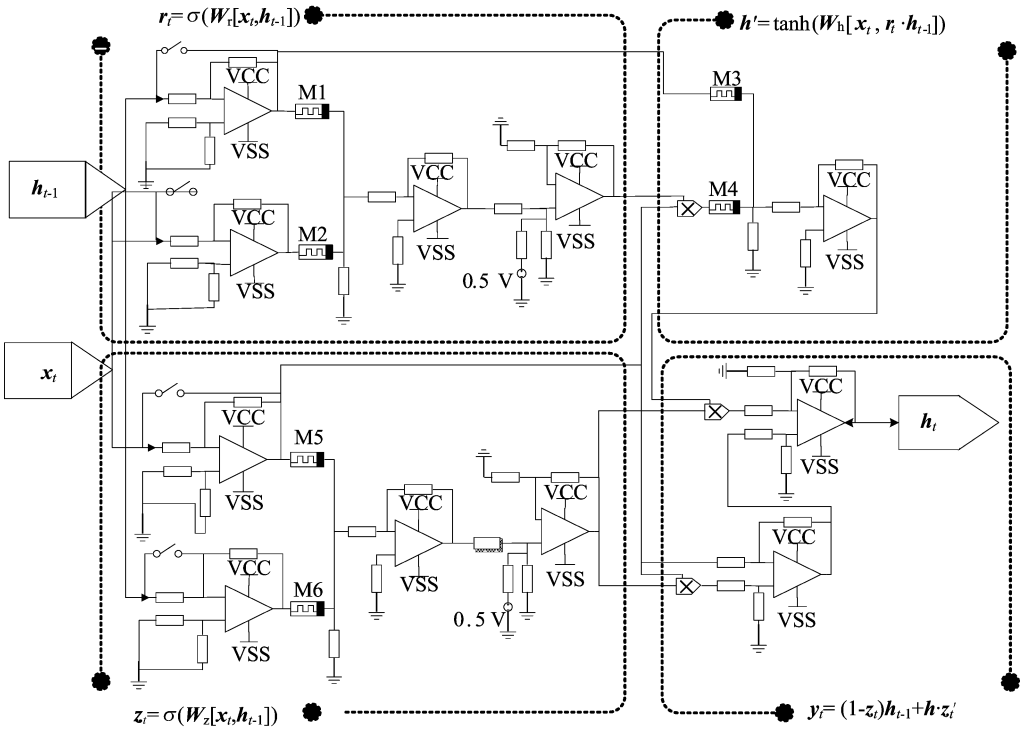


图 4 基于忆阻器的门控循环单元的整体电路

计算公式分别为:

$$\boldsymbol{r}_t=\sigma[\boldsymbol{W}_r/(100\boldsymbol{x}_t)+\boldsymbol{U}_r/(100\boldsymbol{h}_{t-1})] \quad (6)$$

$$\boldsymbol{z}_t=\sigma[\boldsymbol{W}_z/(100\boldsymbol{x}_t)+(\boldsymbol{U}_z/(100\boldsymbol{h}_{t-1})] \quad (7)$$

输出电压与理论值之间存在 100 倍的关系。因为器件电导是 mS 级,所以通过忆阻器的输出电压是 mV 级且在 sigmoid 函数的线性部分,可以在电路中实现一个线性函数替代 sigmoid 函数功能,本文用非零线性函数拟合 sigmoid 函数从 0～0.2的线性部分,线性函数如下:

$$y=0.2x+0.5 \quad (8)$$

根据式(8),sigmoid 函数电路可以简化为求和电路与放大电路的组合。根据式(5)可以看出,电压值与理论值有 100 倍的关系,线性函数的系数为 0.2,根据输入的映射关系,即可将放大电路的反馈电阻阻值设为 40 kΩ、输入电阻阻值设为 1 kΩ,同相求和电路的一端输入设为 0.5 V 直流稳压源,运放同相输入端的 2 个电阻和反向输入端电阻同为 3 kΩ,反馈电阻为 6 kΩ。由此得到 $\boldsymbol{r}_t$ 和 $\boldsymbol{z}_t$ 电路输出值,且其与理论计算为线性关系。

使用同样的方法对 $\boldsymbol{h}'$ 候选激活输出模块进行搭建,将式(5)代入式(3),可得其电路输出公式为:

$$\boldsymbol{h}'=\tanh(\boldsymbol{W}/(100\boldsymbol{x}_t)+(\boldsymbol{U}/100)(\boldsymbol{r}_t\odot\boldsymbol{h}_{t-1})+\boldsymbol{b}) \quad (9)$$

候选激活模块中通过忆阻器后的电压为 mV 级别,而 tanh 函数的−0.5～0.5 部分是线性的,因此为了在电路中实现 tanh 函数,本文使用过零线性函数拟合 tanh 函数,得到的线性拟合函数如下:

$$y=0.8x \quad (10)$$

通过一个放大电路可以简单实现 tanh 函数,放大电路的反馈电阻可以根据拟合的线性函数设置为 80 kΩ,输入电阻为 1 kΩ,即可实现候选激活状态的电路输出。GRU 的当前时刻输出由输出电路完成,其输出方程式(3)根据乘法分配律可以拆分为:

$$\boldsymbol{h}_t=\boldsymbol{h}_{t-1}-\boldsymbol{z}_t\boldsymbol{h}_{t-1}+\boldsymbol{z}_t\boldsymbol{h}' \quad (11)$$

电路的输出还需要电压乘法运算,这并不适合用忆阻器直接实现,因此本文使用 AD633 搭建一个可以实现 mV 级的电压乘法电路。更新门、上一时刻输出、候选激活通过加法电路和减法电路进行组合,可得到当前时刻电路输出。

本文搭建的电路完全从硬件方面实现了 GRU 单元的功能。电路参数只有输入与输出,无需对门信号和候选激活信号进行操作,避免了不

必要的信息传输和外围电路。同时本设计可以通过异地训练方法将权重写入忆阻器，异位方法的主要优点是可以使用任何学习算法进行训练，并且简化了激活函数电路，无需高精度器件就可以实现 GRU 的功能。

3　电路及网络仿真结果

3.1　硬件电路的仿真结果

因为器件的复位电压是 0.5 V，所以电路中的电压范围要控制在－0.5～＋0.5 V 范围内。提取模型中第 1 层 GRU 的权重值、输入和输出值，得到权重值范围为(－0.2，0.2)，输入输出范围均为(－1，1)。因此软件中的参数线性映射为－0.5～0.5 V 的电压值，权重取绝对值后从 0～0.2 线性映射到电导状态 0.000 5 ～ 0.002 0 S，见表 1 所列。

表 1　网络参数与电路参数的映射关系

项目	$\boldsymbol{x}_t$	$\boldsymbol{h}_{t-1}$	$\boldsymbol{W}$
模型	(－1，1)	(1，1)	(－0.2，0.2)
电路	(－0.5，0.5)	(－0.5，0.5)	(0.000 5，0.002 0)

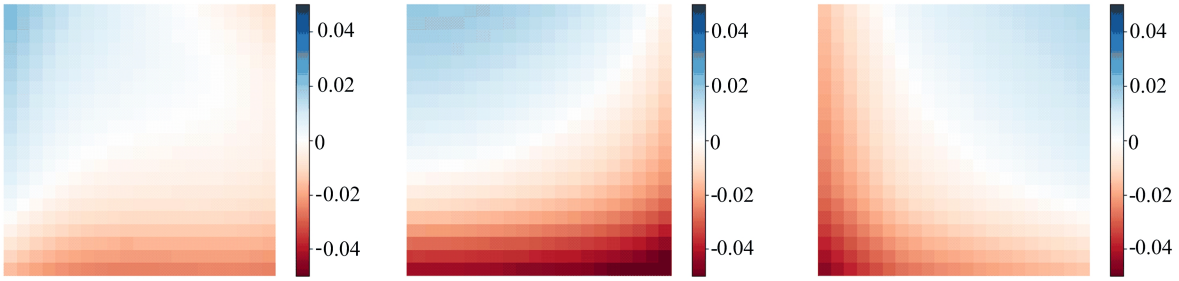
本文将输入 $\boldsymbol{x}_t$ 和 $\boldsymbol{h}_{t-1}$ 从－1 ～ 1 以 0.1 为步长变化进行连续的仿真，以检验仿真模型的运算准确性。由于 GRU 单元中存在 3 次迭代运算，本文提取了迭代运算中的输出结果，以验证模型的实际效果，电路仿真结果混淆矩阵如图 5 所示。因为软件输入是电路值的 2 倍，本文将 sigmoid 激活函数电路输出放大了 200 倍，所以门控单元的输出与理论值的关系式为 $Y=V$，图 5a 的误差计算公式为：

$$E_i = (V-Y)\times 100\% \tag{12}$$

其中：E_i 为第 i 次仿真的误差；V 为电路输出电压值；Y 为理论输出值。软件输出是电路值的 2 倍，即 $Y=2V$，图 5b、5c 的误差计算公式为：

$$E_i = (V-Y/2)\times 100\% \tag{13}$$

由图 5a 可知，复位门和更新门的平均误差值为 0.010 3。由图 5b 可知，候选激活状态的电路输出与理论输出值的平均误差值为 0.014 6。由图 5c 可知，最终电路输出与理论输出值的平均误差值为 0.007 5。1%的误差在很多高精度数学计算中都是不可接受的，但神经网络通常具有较高的鲁棒性，因此对单个神经元的计算准确度要求并不如此苛刻。在神经网络中，考虑到忆阻器硬件电路在执行神经网络过程中相比于传统冯诺依曼架构在能效上的巨大优势，计算精度上的少许损失是完全可以接受的。



(a) 复位门误差　(b) 激活状态误差　(c) 电路输出误差

图 5　电路仿真结果混淆矩阵

3.2　网络仿真结果

通过分析电路输出误差混淆矩阵，获得一个方差为 0.009 7、均值为－0.002 6 的正态分布函数。为了将电路应用于实际情境，本文搭建一个包含 2 层 GRU 和一个全连接层的 3 层神经网络模型，每个 GRU 层具有 64 个单元，全连接层包含 1 个神经元。

在模型的第 1 层 GRU 的输出上，引入一个服从电路误差分布的正态分布随机数，以模拟模型在电路中的运行情况。整个系统的具体操作流程如图 6 所示。

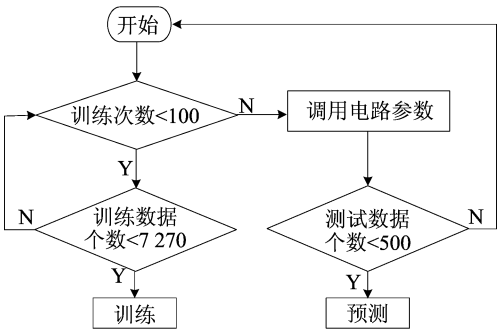


图 6　系统流程

使用 7 270 个训练数据，对模型进行 100 轮的训练。在完成训练后，将电路参数集成到模型

中,并进行 500 个数据的测试。经过训练后,损失值降至 0.000 4,训练 R2 分数达到 0.996 9。为了评估模型性能,对 500 个测试数据集的 R2 分数进行综合分析,得到 0.948 4 的 R2 分数。将电路参数用于模型的前向预测,得到的预测 R2 分数为 0.923 4。

将本文的设计与先前实现 GRU 模型的工作进行对比,结果见表 2 所列。从表 2 可以看出,本文的电路在功耗和速度方面具有显著优势,因此在构建深层网络时可以提供优越的并行度将单元应用到网络中,并且可以快速完成网络的前向传播。

表 2 模型与电路的评估参数对比

设计方法	功耗/mW	执行时间/ns	栅长尺寸/nm	面积/nm^2	模型
文献[8]	732.00	大于 79			GRU
文献[9]		88×10^3	180	$108\ 599\times10^6$	LSTM
文献[10]	155.80		65	10.15×10^{12}	GRU-RNN
本文设计	173.65	45	40	6 284.29	GRU

4 结 论

本文设计了一个基于忆阻器的 GRU 硬件电路,设计的电路可以减少不必要的数据传输和大量外围电路的参与。仿真结果表明,与理论值相比,GRU 电路输出的平均误差低至 0.007 5,而使用搭建的硬件电路实现的股票预测模型的 R2 分数可以达到 0.923 4。与目前的 GRU 硬件实现方式相比,本文设计需要更少的电路组件,具有更快的执行速度和更低的功耗。此外,本文设计能够有效实现 GRU 模型的前向传播。因此,本文设计为未来 GRU 的硬件电路设计提供了一种新的解决方案,并为提高 GRU 网络存算一体的实现速度提供较好的思路。

[参 考 文 献]

[1] 詹新明,黄南山,杨灿.语音识别技术研究进展[J].现代计算机,2008,50(9):43-45.

[2] 龙法宁,朱晓姝,甘井中.基于卷积神经网络的臂丛神经超声图像分割方法[J].合肥工业大学学报(自然科学版),2018,41(9):1191-1195,1296.

[3] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Compution,1997,9(8):1735-1780.

[4] CHUNG J,GULCEHRE C,CHO K,et al. Gated feedback recurrent neural networks[C]//Proceedings of the 32nd International Conference on Machine Learning. [S. l.: s. n.],2015:1-9.

[5] ZHANG X Y,YI F,ZHANG Y M,et al. Drawing and recognizing Chinese characters with recurrent neural network [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2016,40(4):849-862.

[6] 孔祥石.冯诺依曼的计算机理论体系[J].黑龙江广播电视技术,2021(3):106-108.

[7] 韩歌民.系统的性能瓶颈是内存?在“内存墙”的困扰中寻找出路[J].微型计算机,2009,12(34):147-151.

[8] ZAGHLOUL Z S,ELSAVEDl N. The FPGA hardware implementation of the gated recurrent unit architecture[C]//Southeast Con. [S. l.: s. n.],2021:1-5.

[9] ADAM K, SMAGULOVA K, JAMES A P. Memristive LSTM network hardware architecture for time-series predictive modeling problems [C]//2018 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS). [S. l.]: IEEE,2018:459-462.

[10] CHEN C, DING H, PENG H, et al. OCEAN: an on-chip incremental-learning enhanced artificial neural network processor with multiple gated-recurrent-unit accelerators [J]. IEEE Journal on Emerging and Selected Topics in Circuits and Systems,2018,8(3):519-530.

(责任编辑 张 锴)