



基于卷积神经网络的闪存信道检测技术研究

胡海华, 韩国军, 张孝谊

引用本文:

胡海华, 韩国军, 张孝谊. 基于卷积神经网络的闪存信道检测技术研究[J]. 智能系统学报, 2021, 16(6): 1090–1097.

HU Haihua, HAN Guojun, ZHANG Xiaoyi. Research on flash memory channel detection technology based on convolutional neural network[J]. *CAAI Transactions on Intelligent Systems*, 2021, 16(6): 1090–1097.

在线阅读 View online: <https://dx.doi.org/10.11992/tis.202010029>

您可能感兴趣的其他文章

一种基于级联神经网络的飞机检测方法

Cascade convolutional neural networks for airplane detection

智能系统学报. 2020, 15(4): 697–704 <https://dx.doi.org/10.11992/tis.201908028>

一种高效的稀疏卷积神经网络加速器的设计与实现

Design and implementation of an efficient accelerator for sparse convolutional neural network

智能系统学报. 2020, 15(2): 323–333 <https://dx.doi.org/10.11992/tis.201902007>

基于双向消息链路卷积网络的显著性物体检测

Salient object detection based on bidirectional message link convolution neural network

智能系统学报. 2019, 14(6): 1152–1162 <https://dx.doi.org/10.11992/tis.201812003>

基于改进卷积神经网络的多标记分类算法

A multi-label classification algorithm based on an improved convolutional neural network

智能系统学报. 2019, 14(3): 566–574 <https://dx.doi.org/10.11992/tis.201804056>

多层卷积特征的真实场景下行人检测研究

Research on pedestrian detection based on multi-layer convolution feature in real scene

智能系统学报. 2019, 14(2): 306–315 <https://dx.doi.org/10.11992/tis.201710019>

一种多层特征融合的人脸检测方法

Face detection method fusing multi-layer features

智能系统学报. 2018, 13(1): 138–146 <https://dx.doi.org/10.11992/tis.201707018>

微信公众平台



关注微信公众号，获取更多资讯信息

DOI: 10.11992/tis.202010029

网络出版地址: <https://kns.cnki.net/kcms/detail/23.1538.TP.20210901.1055.002.html>

基于卷积神经网络的闪存信道检测技术研究

胡海华, 韩国军, 张孝谊

(广东工业大学 信息工程学院, 广东 广州 510000)

摘要: NAND 闪存信道检测技术直接影响数据存储的可靠性, 本文针对 NAND 闪存信道检测过程中因缺乏信道先验信息而导致检测性能显著降低的问题, 提出了一种基于卷积神经网络 (convolutional neural networks, CNN) 的信道检测器。该检测器通过学习存储单元阈值电压随应用场景的变化特性, 来初始化网络参数, 并通过在系统空闲时间段优化网络参数来实现与信道的匹配。仿真实验结果表明: 在信道先验信息未知的情况下, CNN 检测器可获得比最优阈值检测器更好的检测性能; 与现有的循环神经网络检测器相比, CNN 检测器具有更低的复杂度, 从而能获得更低的检测延时。

关键词: NAND 闪存; 信道检测; 卷积神经网络; 单元间干扰; 信道先验信息; 复杂度; 可靠性; 阈值检测
中图分类号: TN919.5 **文献标志码:** A **文章编号:** 1673-4785(2021)06-1090-08

中文引用格式: 胡海华, 韩国军, 张孝谊. 基于卷积神经网络的闪存信道检测技术研究 [J]. 智能系统学报, 2021, 16(6): 1090-1097.

英文引用格式: HU Haihua, HAN Guojun, ZHANG Xiaoyi. Research on flash memory channel detection technology based on convolutional neural network[J]. CAAI transactions on intelligent systems, 2021, 16(6): 1090-1097.

Research on flash memory channel detection technology based on convolutional neural network

HU Haihua, HAN Guojun, ZHANG Xiaoyi

(School of Information Engineering, Guangdong University of Technology, Guangzhou 510000, China)

Abstract: Channel detection technology for NAND flash memory directly affects the reliability of data storage. Lack of prior information of channel states can lead to significant reduction of detection performances. Aiming at addressing this issue, this paper proposes a channel detector based on convolutional neural networks (CNN). The detector initializes network parameters by learning the varying characteristics with application scenarios of the threshold voltages of the storage cells, and realizes channel matching through optimization of the network parameters during the idle time of the system. The simulation results show that when the channel prior information is unavailable, the CNN detector can achieve better detection performance than the optimal threshold detector. Compared with the existing cyclic neural network detector, CNN detector has lower complexity and therefore reduced detection delay.

Keywords: NAND flash memory; channel detection; convolutional neural network; cell to cell interference; channel prior information; complexity; reliability; threshold detection

近年来, NAND 闪存已成为主流存储介质, 在智能手机、个人电脑和大型数据中心中得到广泛应用^[1-4]。随着制造工艺的进步, 闪存单元的封装尺寸逐渐缩小。一方面, 闪存单元中存储的比特数成倍增加; 另一方面, 闪存内部相邻单元寄生耦合电容效应变大, 单元间干扰 (cell-to-cell inter-

ference, CCI) 增强, 导致数据存储可靠性降低^[5-6]。与此同时, 编程次数和保持时间的增加引发随机电报噪声 (random telegraph noise) 和可持久性噪声 (retention noise), 降低了闪存的使用寿命。

关于以上噪声带来的信道检测可靠性下降问题, 国内外学者提出了一些解决方案^[6-8]。针对单元间干扰, Peng 等^[5]提出 CSAD 和 NIA-CSAD 检测策略; Dong 等^[7]分析了单元间干扰信道模型,

收稿日期: 2020-10-24. 网络出版日期: 2021-09-01.

基金项目: 国家自然科学基金项目 (61871136).

通信作者: 韩国军. E-mail: gjhan@gdut.edu.cn.

并用熵函数来近似求出两种状态分布重叠区的读取参考电压。上述工作需要估计出最优读取参考电压,该方法依赖闪存信道精准的先验信息(相邻单元先验信息、编程和擦除次数、保持时间等相关参数)。为减少对信道先验信息的依赖,文献[8]提出了循环神经网络(recurrent neural network, RNN)检测器。该检测器在未知信道先验信息下能减少噪声对信道可靠性的影响,但是该实验只分析了持久性噪声,没有考虑单元间干扰的影响,并且RNN检测器复杂度较高,被激活会占用大量的系统空闲时间。在实际场景中,闪存信道会同时存在随机电报噪声、可持久性噪声和单元间干扰。信道噪声随保持时间变化,对阈值电压分布扭曲的影响是很难预测的。当前,深度学习技术迅速发展,在信道检测与估计中都表现出优越的性能^[7-12]。本文提出的卷积神经网络(convolutional neural networks, CNN)可以从良好的数据集中学习闪存信道难以预测的变化,并能有效检测闪存单元状态。与现有的大多数检测技术相比,该方法不依赖于精确的信道模型。本文的贡献主要分为3个方面:

1) 通过相关文献对闪存信道特征的测试,可知相邻单元会受到编程单元的干扰^[7-9]。基于闪存相邻单元间特征,本文先在二维坐标下分析了相邻单元编程后阈值电压的相关性,这种相关性表现为相邻单元阈值电压和受干扰单元阈值电压越大,受干扰单元阈值电压越往右偏移。文献[13]揭示了CNN在相关噪声信道中表现出较好的特征提取效果,所以本文挑选CNN网络来检测NAND闪存单元状态。

2) 实验采用以随机电报噪声、可持久性噪声和单元间干扰为主要噪声源的闪存信道^[9],生成受干扰单元和相邻单元的阈值电压序列,组成二维训练数据集。低密度奇偶校验(low-density parity-check, LDPC)解码的单元状态通过独热编码数据预处理作为数据集标签。已训练好的CNN检测器在没有任何信道先验信息下可以直接输出闪存信道的硬判决信息(单元存储状态)和软判决信息(闪存单元属于某种状态的后验概率)。假设信道模型已知,熵函数求解出电压分布相邻状态分布的重叠区间。在重叠区间中用穷举搜索法找到的最优阈值作为性能对比标杆,下面称之为最优阈值检测器(optimal threshold detector)。计算机仿真结果显示:本文提出的CNN检测器与最优阈值检测器对比,在信道先验信息未知时,能有效提升闪存信道的检测性能。

3) 与现有的RNN检测器^[8]相比,分析了两者的浮点运算数(floating-point operations, FLOPs),单层CNN网络复杂度显著小于单层RNN网络。因此,CNN检测器被激活所需的系统空闲时间比RNN要少得多。

1 闪存信道系统

闪存信道系统模型如图1所示,数据先通过擦除与编程操作写入闪存中,随后经过的干扰包括随机电报噪声、可持久性噪声和单元间干扰。

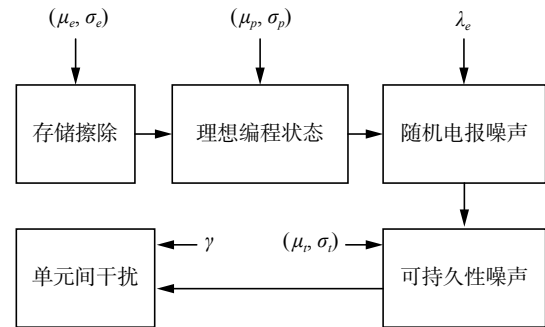


图1 MLC 闪存信道模型

Fig. 1 MLC flash channel model

1.1 编程操作

在NAND闪存中,字线(word-line)由擦除状态单元和编程状态单元组成^[14]。对于MLC型的闪存单元,可以存储2比特位,并将存储类型表示为4个不同的状态:11、10、01和00。11是擦除状态,其余的是编程状态。

在编程之前,必须擦除闪存单元中的数据。擦除单元的阈值电压符合高斯分布^[8],该分布为

$$p_e(x) = \frac{1}{\sigma_e \sqrt{2\pi}} \exp\left(-\frac{(x-\mu_e)^2}{2\sigma_e^2}\right) \quad (1)$$

式中: σ_e 和 μ_e 分别是被擦除单元阈值电压分布的标准差和平均值。

在编程操作中,闪存控制器会向浮栅注入一定量的电荷。编程状态的阈值电压也服从高斯分布^[14]:

$$p_p(x) = \frac{1}{\sigma_p \sqrt{2\pi}} \exp\left(-\frac{(x-\mu_p)^2}{2\sigma_p^2}\right) \quad (2)$$

式中: σ_p 和 μ_p 分别是被编程单元阈值电压分布的标准差和平均值。

1.2 随机电报噪声

编程和擦除次数会损坏浮栅晶体管隧道的氧化物,引起闪存单元阈值电压随机波动。因此,随机电报噪声采用类高斯分布建模为

$$p_r(x) = \frac{1}{2\lambda_e} \exp\left(-\frac{|x|}{\lambda_e}\right) \quad (3)$$

式中: $\lambda_e = K_r \sqrt{PE}$, PE 是编程和擦除循环次数, K_r 为常数^[15]。

1.3 可持久性噪声

可持久性噪声是由编程单元中电荷泄漏引起的阈值电压分布偏移。可持久性噪声遵循高斯分布^[15]:

$$p_i(x) = \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right) \quad (4)$$

平均值 μ_i 和标准差 σ_i 由式 (5)、(6) 给出^[16]:

$$\mu_i = \Delta V_i [A_i(PE)^{a_i} + B_i(PE)^{a_0}] \cdot \log(1+T) \quad (5)$$

$$\sigma_i = 0.5|\mu_i| \quad (6)$$

式中: ΔV_i 为编程前后的闪存单元阈值电压变化; T 为保持时间; PE 为编程和擦除循环次数; a_0 、 a_i 、 A_i 、 B_i 均为常数。

1.4 单元间干扰

单元间干扰是闪存信道中主要噪声之一^[17]。编程单元阈值电压的变化会影响相邻单元的阈值电压。由单元间干扰引起的受干扰单元的总阈值电压影响强度计算公式为^[17]

$$F = \sum_k (\Delta V^{(k)} * \gamma^{(k)}) \quad (7)$$

式中: $\Delta V^{(k)}$ 为第 k 个干扰单元的编程前后阈值电压偏移。如图 2 所示, $\gamma^{(k)}$ 为受干扰单元与干扰单元的耦合比系数, γ_y 表示垂直方向电容耦合系数, γ_{xy} 表示对角线方向的电容耦合系数。对于不同的位线结构, 用于计算 F 的干扰单元数目是不同的。为了简单起见, 本文只考虑全位线结构, 这种结构会降低被干扰单元受到的干扰强度。

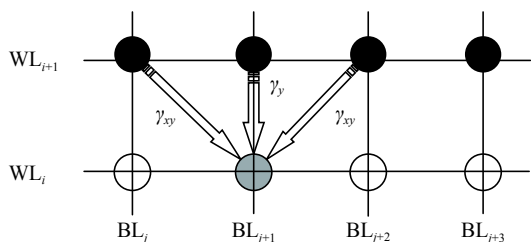


图2 多电平闪存的阵列结构

Fig. 2 MLC flash memory array structure

根据文献 [17], $p_e^{(k)}(x)$ 为单元间干扰的阈值电压分布。在随机电报噪声、单元间干扰和可持久性噪声的干扰下, 阈值电压的条件概率分布函数为^[18]:

$$p^{(k)}(x) = p_p^{(k)}(x) * p_c^{(k)}(x) * p_r^{(k)}(x) * p_t^{(k)}(x) \quad (8)$$

式中: $p^{(k)}(x)$ 表示 k 个状态的阈值电压分布, $k \in \{11, 10, 00, 01\}$ 。为了统计出 $p^{(k)}(x)$, 需要知道式 (5) 中保持时间 T 的大小, 然而在实际场景中闪存控制器没有额外的存储空间来记录保持时间^[19-20]。

参照文献 [8-14], MLC 闪存信道仿真参数的设置如下: $\mu_e=0.14$, $\sigma_e=0.25$, $a_i=0.62$, $a_0=0.3$, $A_i=0.000035$, $B_i=0.000235$, $\gamma_y=0.08s$, $\gamma_{xy}=0.006s$ 。 s 表示电容耦合

强度。

2 CNN 检测器的设计

2.1 相关性

闪存信道参数设置为: PE 次数为 1000, 保持时间 1 年, 电容耦合强度 s 分别为 0、1.4、2.5。由图 3 可以发现不同状态的受干扰单元不是交错呈现, 而是与干扰单元的阈值电压具有正相关关系, 随着耦合强度增大, 受干扰单元的阈值电压向右偏移越明显。这现象显示了单元间的阈值电压具有相关性。当观测的干扰单元越多, 这种相关性越明显, 但是这种现象难以精确建立数学模型。

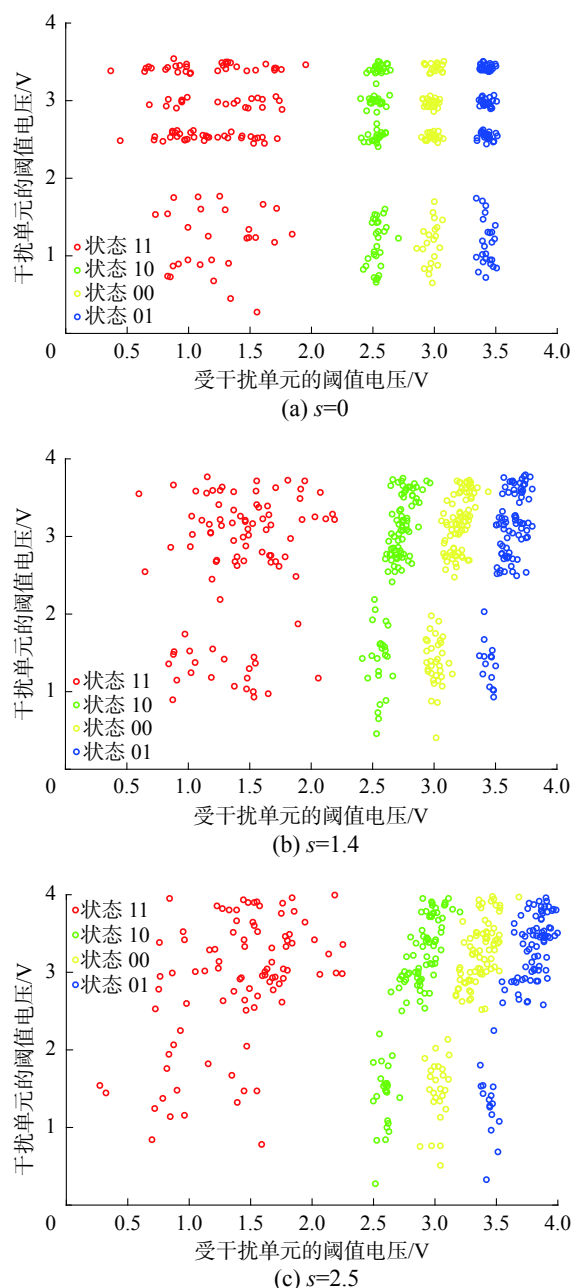


图3 闪存内部阈值电压相关性分析

Fig. 3 Schematic of the correlation analysis of the internal threshold voltage of flash memory

2.2 CNN 用于单元状态检测

本文的主要工作是将受干扰单元字线和干扰单元字线的阈值电压组成序列, 输入到 CNN 中进行提取特征, 再输出闪存单元用于硬判决信息 (单元存储状态) 和软判决信息 (闪存单元属于某种状态的后验概率)。

假设每个输入序列个数为 M (字线长度), 每个向量为 2 维, 序列所有阈值电压向量在矩阵 $V \in \mathbf{R}^{2 \times M}$ 中。

卷积层: 本文挑选是一维卷积神经网络, 输入序列表示为 $[V_1 V_2 \cdots V_M]$ 矩阵, 然后使用卷积核做卷积运算:

$$I_{(1,j)} = \text{RELU}(k_{(1,j)} * V + b_{(1,j)}) \quad (9)$$

式中: $I_{(1,j)}$ 是第一层 CNN 第 j 个特征映射; $k_{(1,j)}$ 为第 j 个卷积核; $b_{(1,j)}$ 为网络偏置值, 挑选的激活函数为整流线性函数 (rectified linear unit, RELU)。下一层的卷积操作的输出为

$$I_{(i,j)} = \text{RELU}(k_{(i,j)} * V + b_{(i,j)}) \quad (10)$$

式中: $I_{(i,j)}$ 是第 i 层 CNN 网络第 j 个特征映射; $k_{(i,j)}$ 是第 i 层 CNN 网络第 j 个卷积核。在本工作中卷积网络池化层会丢弃一些重要的特征信息, 所以本文的卷积网络由过滤器和全连接组成, 并且 CNN 输出长度也为 M 。

假设有 N 个数据样本, 对于输入样本 i , 经过 CNN 网络提取特征后, 形成序列 $Z_i = [Z_i^1 Z_i^2 \cdots Z_i^M]$, 希望输出每个序列节点所属某种类型的概率, 即 $P(y_i^m = j | V_i^m)$ 。以 MLC 为例, 其有 4 种状态, 即 $j \in \{V_{S11}, V_{S10}, V_{S00}, V_{S01}\}$, $m \in \{1, 2, \cdots, M\}$ 。Softmax 函数可以适用于多分类, 其输出为

$$h_\theta(Z_i) = \frac{\begin{bmatrix} P(y_i^1 = 11 | Z_i^1, \theta) \cdots P(y_i^M = 11 | Z_i^M, \theta) \\ P(y_i^1 = 10 | Z_i^1, \theta) \cdots P(y_i^M = 10 | Z_i^M, \theta) \\ P(y_i^1 = 00 | Z_i^1, \theta) \cdots P(y_i^M = 00 | Z_i^M, \theta) \\ P(y_i^1 = 01 | Z_i^1, \theta) \cdots P(y_i^M = 01 | Z_i^M, \theta) \end{bmatrix}}{\sum_{j=1}^k \exp(\theta_j^T Z_i^j)} \begin{bmatrix} \theta_1^T Z_i^1 \cdots \theta_1^T Z_i^M \\ \theta_2^T Z_i^1 \cdots \theta_2^T Z_i^M \\ \theta_3^T Z_i^1 \cdots \theta_3^T Z_i^M \\ \theta_4^T Z_i^1 \cdots \theta_4^T Z_i^M \end{bmatrix} \quad (11)$$

式中: $k=4$; $\theta_1, \theta_2, \theta_3, \theta_4$ 是 Softmax 函数的权重参数; $\theta = [\theta_1 \theta_2 \theta_3 \theta_4]$ 。对于样本 i 中输入序列元素 V_i^m , 属于状态标签类型 j 的概率为

$$p(y_i^m = j | Z_i^m) = \frac{\exp(\theta_j^T Z_i^m)}{\sum_{m=1}^M \exp(\theta_j^T Z_i^m)} \quad (12)$$

3 数据预处理

设计好 CNN 检测器结构后, 要对 CNN 网络

训练, 则需要从本地获得训练数据。本文采用均匀量化策略获得闪存单元中的阈值电压数据。量化数量越多检测越精确, 但是也会带来读取延迟。如表 1, 为了权衡闪存单元电压读取延迟和可靠性, 选择 64-level 均匀感知量化, 在闪存单元状态的符号错误率 (symbol error rate, SER) 为 0.0033 时, 读取延迟和可靠性折中。闪存控制器最小的读取单位是字线, 则需要设置输入网络的序列长度与字线长度相当。

表 1 不同量化区间的 SER 结果

Table 1 Simulation of SER in different quantization intervals

区间	SER	区间	SER
4-level	0.1400	64-level	0.0033
8-level	0.2500	128-level	0.0033
16-level	0.0067	256-level	0.0033
32-level	0.0045	1024-level	0.0032

实验需要采集闪存控制器中读出闪存块中每个单元的量化阈值电压和 LDPC 解码得到的每个单元状态信息。首先, 感应和解码出若干组闪存块。其次, 获得阈值电压 V 和状态标签 x 。其中, 闪存单元的阈值电压在区间 $[V_{\text{ref}}^i, V_{\text{ref}}^{i+1}]$ 中被感知, V_{ref}^i 是第 i 个参考电压, 则闪存单元的量化阈值电压 V 表示为

$$V = \frac{V_{\text{ref}}^i + V_{\text{ref}}^{i+1}}{2} \quad (13)$$

对于 MLC 闪存单元有 4 种不同的状态 $\{V_{S11}, V_{S10}, V_{S00}, V_{S01}\}$, 用独热编码预处理后为

$$\{V_{S11}, V_{S10}, V_{S00}, V_{S01}\} = \left[\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \right]$$

对于字线上的闪存单元 k , 简单来说, 当 Softmax 函数输出的硬信息 $y^k = [1 \ 0 \ 0 \ 0]^T$ 时, 单元 k 对应 V_{S11} ; 以此类推, 当输出 $y^k = [0 \ 0 \ 0 \ 1]^T$, 对应 V_{S01} 。当输出软信息时, Softmax 函数输出每个单元属于各状态的概率, CNN 检测器能自适应地估计出字线中闪存单元状态概率, 然后式 (14) 计算出对数似然比。闪存单元中的比特位 b_{id} 对应的对数似然比 (log-likelihood ratio, LLR) 值可以被计算为

$$L(b_{id}) = \log \frac{\sum_{k \in S_{id}} p(y_i^m = k | Z_i^m)}{\sum_{k \in S} p(y_i^m = k | Z_i^m) - \sum_{k \in S_{id}} p(y_i^m = k | Z_i^m)} \quad (14)$$

式中: S_{id} 代表第 i 个闪存单元中第 d 比特位为 1 的状态集。其中, $S = \{V_{S11}, V_{S10}, V_{S00}, V_{S01}\}$, $S_1 = \{V_{S11}, V_{S10}\}$, $S_2 = \{V_{S11}, V_{S01}\}$ 。

4 实验仿真及分析

实验仿真在 Window 10 操作系统下进行, 采用 Keras 深度学习框架。在训练网络之前, 假设在实验中信道的保持时间未知, 就需要生成失配的验证集来测试已训练好 CNN 检测器的鲁棒性。定义失配时间公式为 $\Delta T = T_{\text{test}} - T_{\text{train}}$, T_{train} 和 T_{test} 分别代表训练集和验证集的保持时间。定义失配编程和擦除循环次数公式为 $\Delta \text{PE} = \text{PE}_{\text{test}} - \text{PE}_{\text{train}}$, PE_{train} 和 PE_{test} 分代表训练集和验证集的编程和擦除循环次数。训练 CNN 检测器步骤:

1) 闪存控制器找到第 i 和第 $i+1$ 字线位置, 即图 2 中的 WL_i 和 WL_{i+1} 。

2) 式 (13) 量化 WL_i 和 WL_{i+1} 中每个单元的感应电压。

3) LDPC 解码 WL_i 中的信息, 若解码成功, 将独热编码预理解码后的状态作为标签。若解码不成功, 跳到下一个字线。返回 1), 直到获取数据集数量满足实验要求。

4) 获取的数据集将用于训练 CNN 检测器。为了提升训练效率, 将数据集分成小批量数据, 分批训练 CNN 检测器。

CNN 结构由卷积核和特征图组成, 可以表示为 $\{L; f_1, f_2, \dots, f_L; k_1, k_2, \dots, k_L\}$ 。这里的 L 表示 CNN 网络层数, f 表示卷积核大小, k 表示特征图数量。在实验中, 通过比较误比特率 (bit error rate, BER) 和误帧率 (frame error rate, FER) 来评估系统性能。先进行单层 CNN 和 RNN 的复杂度分析, 然后用表 2 设置 CNN 结构来测试在原始数据和编码数据下的系统性能。

$$\begin{aligned} \text{误比特率} &= \frac{\text{读取的错误比特数}}{\text{读取的总比特数}} \\ \text{误帧率} &= \frac{\text{读取的错误帧数}}{\text{读取的总帧数}} \end{aligned}$$

表 2 CNN 实验设置
Table 2 CNN experimental setup

名称	参数
CNN 结构	{2; 2, 2; 2, 1}
训练数据	3000 个码字
验证数据	3000 个码字
总比特数	13 824 000
初始方法	Xavier initialization
优化方法	Adam optimization
序列长度	1 152

4.1 复杂度分析

对于文献 [8] 所提出的 RNN 检测器, 在实际

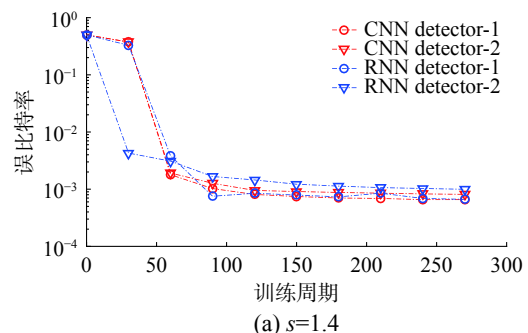
的训练过程中会占用大量的时间, 计算复杂度。在理想条件下, 通过设置 CNN 和 RNN 超参数使两者的复杂度相当, 然后对比它们的性能。可是 RNN 网络要比 CNN 网络复杂多, 在 CNN 网络中有加法和乘法等线性运算, 而 RNN 网络中有许多的 tanh 和 exp 等复杂的非线性运算。为了方便比较, 通过统计 FLOPs 来衡量网络模型的复杂度。FLOPs 计算网络模型所有的加法和乘法次数。GRU 是 RNN 网络模型中复杂度最低的网络模型。若设置一个层 GRU 网络的神经元个数为 L , 输入数据的维度为 D , N 表示层数, 序列为 T 。根据文献 [21], GRU 模型复杂度为

$$O\left(\sum_{i=1}^N T_i L_i (6D_i + 6L_i + 5)\right) \quad (15)$$

根据文献 [22], CNN 网络模型复杂度为

$$O\left(\sum_{i=1}^N 2C_{i-1} K_i^2 T_i C_i\right) \quad (16)$$

式中: C_i 表示第 i 层输出数据的通道数; K_i 表示第 i 层的卷积核大小; T_i 为第 i 层输出的序列。在本文中 $C_0=1$, 采用的二维阈值电压数据作为输入, 则 $D_1=2$, $K_1=(D_1, k)$; 用一层 CNN 网络与一层 GRU 网络对比复杂度, 则 $C_1=1$, k 为第一层卷积核尺寸。Softmax 函数复杂度可以忽略不计。此时, GRU 网络模型复杂度为 $O((6L_1+17L_1^2)T_1)$, CNN 网络模型复杂度为 $O(4kT_1)$ 。如图 4, 设置实验环境 $s \in \{1.4, 1.6\}$, $T_{\text{train}}=1$ 年, $\text{PE}_{\text{train}}=1000$ 。图 4 展示了训练周期对误比特率的影响。CNN detector-1、CNN detector-2、RNN detector-1 和 RNN detector-2 所标记的曲线对应参数 $k=1$ 、 $k=2$ 和 $L_1=1$ 、 $L_2=2$ 时的模型。随着训练周期变化, CNN 检测器和 RNN 检测器趋于稳定。RNN 检测器性能亦可接近于 CNN 检测器, 式 (15) 中得到 RNN 检测器复杂度与 L_1 平方成正比。式 (16) 中 CNN 检测器复杂度与 K 成正比。如果在有单元间干扰情况下要达到相同性能, CNN 检测器复杂度小于 RNN 检测器。



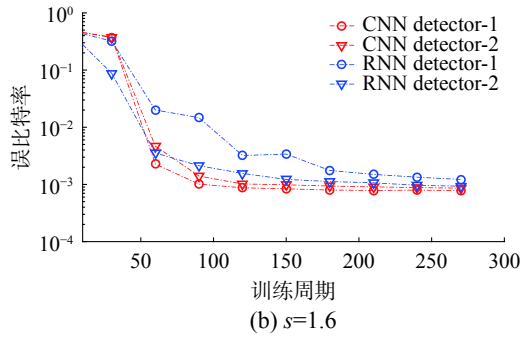


图 4 不同训练周期下 RNN 检测器和 CNN 检测器比特错误率比较

Fig. 4 Comparison of the BERs of the RNN and CNN detectors under different training periods

为了避免 GPU 等硬件带来的功耗等问题^[8], 实验在无 GPU 环境中进行。表 3 是在无 GPU 下参数 FLOPs 和每个训练周期的耗时统计。随着参数的增加, RNN 检测器的 FLOPs 接近倍数增加, 带来的计算消耗和复杂度特别大。从耗时上可以看出, CNN 检测器训练周期相当于 RNN 检测器训练周期的四分之一。MEI 提出在系统空闲时间时激活 RNN 检测器, 但是 RNN 检测器训练耗时长^[8]。与 RNN 检测器相比, CNN 检测器被激活占用的系统空闲时间较短。

表 3 CNN 检测器和 RNN 检测器参数个数及 FLOPs
Table 3 Parameters and FLOPs of the CNN and RNN detectors

模型	参数/个	FLOPs/次	周期/s
CNN detector-1	11	192	0.44
CNN detector-2	13	222	0.48
RNN detector-1	20	361	1.68
RNN detector-2	42	721	1.71

4.2 原始数据分析

为了更好地验证模型检测有效性, LDPC 译码器尽可能收集解码后综合值为 0 的码字数据。实验中选择最优阈值为性能对比标杆。最优阈值检测器通过熵函数得到两种状态分布的重叠区间, 然后在重叠区间中找到最优读取参考电压。本文的 CNN 检测器使用 2 维数据输入网络。然后, 测试了有无单元间干扰下两种检测器误码率对比, $s=0$ 时, 表示没有单元间干扰的信道; $s=1.4$ 时, 表示 3 种噪声同时存在下的信道, 并且单元干扰强度为 1.4。

如图 5 所示, 在实验环境中存在单元间干扰情况下匹配和失配的测试集 ($\Delta PE=2\ 500, 5\ 000$), $T_{\text{train}}=1$ 年, CNN 检测器都取得很好的性能。当 $s=1.4$ 时, CNN 检测器跟最优阈值检测器比提升

了闪存使用寿命。例如, 在误比特率约为 0.01, CNN 检测器可以将闪存编程和擦除循环次数提升大约 11 000 次, 而最优阈值检测器仅在 3 000 次左右。在没有单元间干扰下, 采用 CNN 检测器性能增益较小, 这时的噪声分布是标准 AWGN 闪存信道, 也是理想条件下的信道。此时, CNN 检测器与最优阈值检测器性能相似。由此, 可以得出在有单元间干扰下, CNN 检测器能有效地提取出单元间干扰的特征来检测闪存单元状态, 并且提升了闪存使用寿命。此外, 当用到失配的测试集时, 仿真性能对比结果显示, 在失配编程和擦除循环次数小于 5 000 情况下, 本文 CNN 检测器检测性能也好于最优阈值检测器。

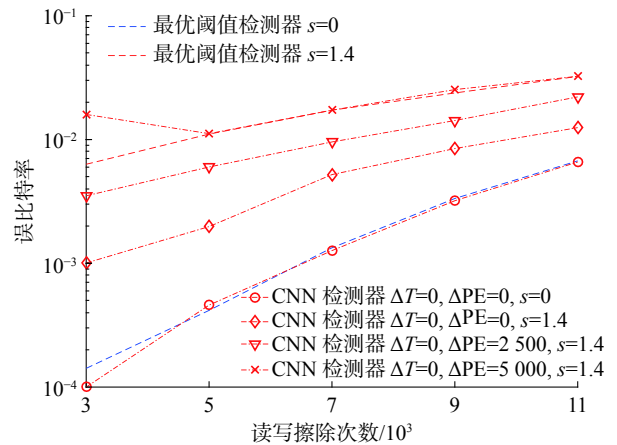


图 5 失配编程和擦除循环次数比较

Fig. 5 Programming and erase cycles by mismatching data

传统保持低误比特率的方法是动态更新读取参考电压。然而在未知信道信息时, 闪存控制器很难记录保持时间 T 。这时就很难通过传统方法获得新读取参考电压。前面的实验结果证明有单元干扰下, CNN 检测器优于最优阈值检测器, 接下来实验通过对失配编程和擦除循环次数固定为 $\Delta PE=1\ 000$, $PE_{\text{train}}=5\ 000$, 测试 CNN 检测器在失配保持时间下的性能。如图 6, 当时间不匹配到 1 个月时, CNN 检测器还是优于最优阈值检测器的性能并且与保持时间匹配的数据接近。当保持时间失配增加到 1 年时, 初期 CNN 检测器与匹配数据的性能略有下降。初期需要短周期更新 CNN 检测器适应信道变化, 后期则可以数年更新一次 CNN 检测器, 就可以接近匹配数据的 CNN 检测器误比特率性能, 也证明了 CNN 检测器鲁棒性强。

4.3 编码数据分析

为了分析 CNN 检测器输出的软信息, 实验选择了码率约 0.889 (2 048, 2 304) 的 LDPC 码写入到闪存单元中。在数据集设置耦合强度因子为 $s=1.4$, 保持时间 $T_{\text{train}}=1$ 年和读写次数 $PE_{\text{train}}=2000$ 。

为了避免验证数产生的读取延迟,实验减少了量化电压的数量,采用 16-level 和 32-level 两种均匀量化读取方法。CNN 检测器评估出闪存单元软信息,然后通过式 (14) 计算出 LLR 值。图 7 给出了本文提出方法与传统方法计算 LLR 的系统性能比较。首先,设置 0~20 最大迭代次数不等。其次,将 CNN 检测器测试的失配保持时间 $\Delta T=5$ 月,失配编程和擦除循环次数 $\Delta PE=2\ 000$ 。本文提出方法比 5 月之前的传统方法计算 LLR 收敛快。也就是说, CNN 检测器评估的 LLR 更可靠,有助于置信传播 (belief propagation, BP) 算法^[23-24] 纠错性能的提升。

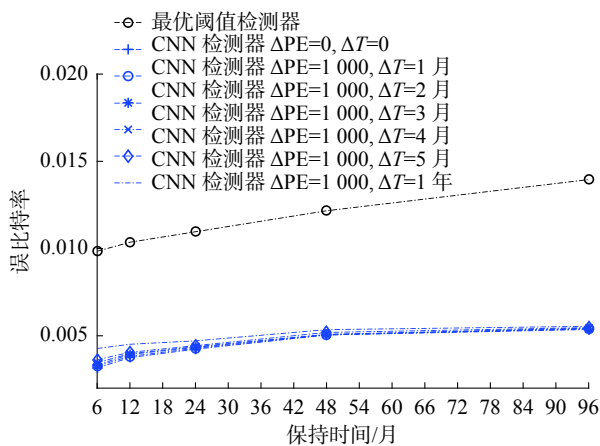


图 6 失配保持时间比较
Fig. 6 Retention time through mismatching data

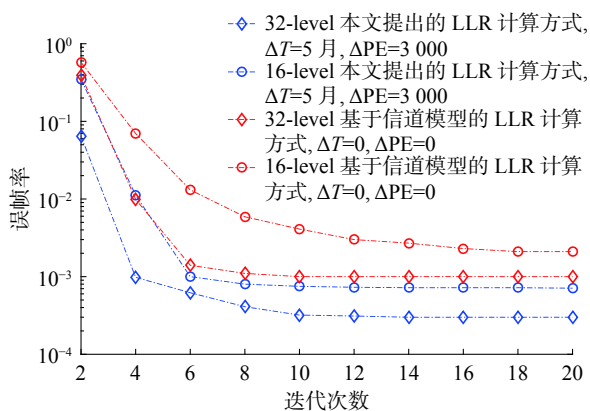


图 7 最大迭代次数比较
Fig. 7 Comparison of maximum iterations

5 结束语

针对 NAND 闪存存储数据可靠性下降的问题,本文提出 CNN 检测器,该检测器不依赖于信道先验信息,通过学习相邻单元阈值电压的相关特征,能有效检测 NAND 闪存单元状态。CNN 检测器具有训练周期短、鲁棒性强等特点,能占用较少的系统空闲时间来适应信道变化。CNN 检

测器还能输出更准确的 LLR 信息,提升 BP 译码算法的纠错性能。闪存信道检测在深度学习中可以看作分类问题,因此本文提出的方法也可以应用到 TLC、QLC 闪存信道中。

参考文献:

- [1] 范正勤, 韩国军. 多电平闪存信道下阈值电压高效检测算法 [J]. 应用科技, 2019, 46(5): 57-62.
FAN Zhengqin, HAN Guojun. High-efficiency detection algorithm for threshold voltage in multi-level cell NAND flash memory [J]. Applied science and technology, 2019, 46(5): 57-62.
- [2] 张旋, 慕建君, 焦晓鹏. 面向 MLC 闪存的比特翻转译码算法研究 [J]. 哈尔滨工程大学学报, 2019, 40(2): 331-337.
ZHANG Xuan, MU Jianjun, JIAO Xiaopeng. Improved bit-flipping decoding algorithm for MLC flash memory [J]. Journal of Harbin Engineering University, 2019, 40(2): 331-337.
- [3] 郑敏华, 韩国军. 融合相邻单元高低页存储可靠性的闪存译码算法 [J]. 应用科技, 2020, 47(3): 17-23.
ZHENG Minhua, HAN Guojun. An improved decoding algorithm for NAND flash memory fusing the storage reliability of upper and lower pages of adjacent cells [J]. Applied science and technology, 2020, 47(3): 17-23.
- [4] LIU Xingcheng, YANG Guojun, CHEN Xuechen. Variable-node-based belief-propagation decoding with message pre-processing for NAND flash memory [J]. IEEE access, 2019, 7: 58638-58653.
- [5] PENG Zishuai, HE Ruiquan, HAN Guojun, et al. Neighbor-a-posteriori information assisted cell-state adaptive detector for NAND flash memory [J]. IEEE communications letters, 2019, 23(11): 1967-1971.
- [6] KIM T, KONG G, XI Weiya, et al. Cell-to-cell interference compensation schemes using reduced symbol pattern of interfering cells for MLC NAND flash memory [J]. IEEE transactions on magnetics, 2013, 49(6): 2569-2573.
- [7] DONG Guiqiang, XIE Ningde, ZHANG Tong. On the use of soft-decision error-correction codes in NAND flash memory [J]. IEEE transactions on circuits and systems I: regular papers, 2011, 58(2): 429-439.
- [8] MEI Zhen, CAI Kui, HE Xuan. Deep learning-aided dynamic read thresholds design for multi-level-cell flash memories [J]. IEEE transactions on communications, 2020, 68(5): 2850-2862.
- [9] FAN Zhengqin, CAI Guofa, HAN Guojun, et al. Cell-state-distribution-assisted threshold voltage detector for NAND flash memory [J]. IEEE communications letters, 2019, 23(4): 576-579.

- [10] SANDELL M, ISMAIL A. Machine learning for LLR estimation in flash memory with LDPC codes[J]. *IEEE transactions on circuits and systems II:express briefs*, 2021, 68(2): 792–796.
- [11] CHOE H, JEE J, LIM S C, et al. Machine-learning-based read reference voltage estimation for NAND flash memory systems without knowledge of retention time[J]. *IEEE access*, 2020, 8: 176416–176429.
- [12] NAKAMURA T, DEGUCHI Y, TAKEUCHI K. Adaptive artificial neural network-coupled LDPC ECC as universal solution for 3-D and 2-D, charge-trap and floating-gate NAND flash memories[J]. *IEEE journal of solid-state circuits*, 2019, 54(3): 745–754.
- [13] LIANG Fei, SHEN Cong, WU Feng. An iterative BP-CNN architecture for channel decoding[J]. *IEEE journal of selected topics in signal processing*, 2018, 12(1): 144–159.
- [14] ASLAM C A, GUAN Yongliang, CAI Kui. Decision-directed retention-failure recovery with channel update for MLC NAND flash memory[J]. *IEEE transactions on circuits and systems I:regular papers*, 2018, 65(1): 353–365.
- [15] DONG Guiqiang, XIE Ningde, ZHANG Tong. Enabling NAND flash memory use soft-decision error correction codes at minimal read latency overhead[J]. *IEEE transactions on circuits and systems I:regular papers*, 2013, 60(9): 2412–2421.
- [16] ASLAM C A, GUAN Yongliang, CAI Kui. Read and write voltage signal optimization for multi-level-cell (MLC) NAND flash memory[J]. *IEEE transactions on communications*, 2016, 64(4): 1613–1623.
- [17] DONG Guiqiang, LI Shu, ZHANG Tong. Using data postcompensation and predistortion to tolerate cell-to-cell interference in MLC NAND flash memory[J]. *IEEE transactions on circuits and systems I:regular papers*, 2010, 57(10): 2718–2728.
- [18] OUYANG Shijie, HAN Guojun, FANG Yi, et al. LLR-distribution-based non-uniform quantization for RBI-MSD algorithm in MLC flash memory[J]. *IEEE communications letters*, 2018, 22(1): 45–48.
- [19] CAI Yu, LUO Yixin, HARATSCH E F, et al. Data retention in MLC NAND flash memory: characterization, optimization, and recovery[C]//2015 IEEE 21st International Symposium on High Performance Computer Architecture (HPCA). Burlingame, USA, 2015: 551–563.
- [20] HE Ruiquan, HU Haihua, XIONG Chunru, et al. Artificial neural network assisted error correction for MLC NAND flash memory[J]. *Micromachines*, 2021, 12(8): 879.
- [21] ZHANG Pengfei, XUE Jianru, LAN Cuiling, et al. EleAtt-RNN: adding attentiveness to neurons in recurrent neural networks[J]. *IEEE transactions on image processing*, 2020, 29: 1061–1073.
- [22] CAO Congzhe, LI Duanshun, FAIR I. Deep learning-based decoding of constrained sequence codes[J]. *IEEE journal on selected areas in communications*, 2019, 37(11): 2532–2543.
- [23] 张旋, 慕建君, 焦晓鹏. 基于错误特征的 MLC 闪存最小和译码算法 [J]. *系统工程与电子技术*, 2019, 41(8): 1887–1895.
- ZHANG Xuan, MU Jianjun, JIAO Xiaopeng. Min-sum decoding algorithm based on error characteristics for MLC flash memory[J]. *Systems engineering and electronics*, 2019, 41(8): 1887–1895.
- [24] 张旋, 周乐, 侯爱华. 一种适用于 MLC 闪存的 CCI 噪声均衡化算法 [J]. *计算机科学*, 2018, 45(S1): 541–544.
- ZHANG Xuan, ZHOU Le, HOU Aihua. CCI noise equalization algorithm for MLC flash memory[J]. *Computer science*, 2018, 45(S1): 541–544.

作者简介:



胡海华, 博士研究生, 主要研究方向为面向数据存储的信号处理技术、差错控制编码技术、机器学习。



韩国军, 教授, 博士生导师, 主要研究方向为面向数据存储的信号处理与差错控制编码技术、车联网及先进辅助驾驶技术。获广东省自然科学二等奖 1 项, 广州市科技进步一等奖 1 项。主持国家自然科学基金面上项目 3 项。获发明专利授权 4 项, 发表学术论文 50 余篇。



张孝谊, 硕士研究生, 主要研究方向为面向数据存储的信号处理技术、差错控制编码技术、机器学习。