基于 NOR Flash 的卷积计算单元的设计

徐伟民[1],黄　鲁[1],蒋明峰[2]

(1. 中国科学技术大学 微电子学院,安徽 合肥 230026;
2. 中国科学技术大学 信息科学技术学院,安徽 合肥 230026)

摘　要:提出一种基于 NOR Flash 的模拟卷积运算单元,与同类模拟卷积运算单元相比具有高精度、高能耗比、低噪声的特点。该单元采用存算一体架构,将卷积核的权重参数以阈值电压的方式存储在 Flash 中,输入图片经过模拟卷积运算得到输出图片。在 SMIC 65 nm 浮栅工艺下,使用 SOBEL 边缘检测算法评估该单元的性能。仿真结果表明,在 3.3 V 电源电压,100 MHz 时钟下,实现一个 3×3 卷积核的 Flash 阵列的能耗比达到 0.18 TOPS/W,卷积计算结果的峰值信噪比(PSNR)为 39.05 dB。

关键词:NOR Flash;存算一体;卷积加速

中图分类号:TN432　　**文献标识码**:A　　DOI: 10.19358/j.issn.2096-5133.2020.05.013

引用格式:徐伟民,黄鲁,蒋明峰. 基于 NOR Flash 的卷积计算单元的设计[J]. 信息技术与网络安全,2020,39(5):63-68.

Design of convolution calculation unit based on NOR Flash

Xu Weimin[1], Huang Lu[1], Jiang Mingfeng[2]

(1. School of Microelectronics, University of Science and Technology of China, Hefei 230026, China;
2. School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China)

Abstract: An analog convolution operation unit based on NOR Flash is proposed, which has the characteristics of high precision, high energy consumption ratio and low noise compared with similar analog convolution units. The unit adopts a computation in memory architecture, and stores the weight parameters of the convolution kernel in the Flash in the manner of threshold voltage, and the input image is subjected to an analog convolution operation to obtain an output picture. The SOBEL edge detection algorithm was used to evaluate the performance of the cell under the SMIC 65 nm floating gate process. The simulation results show that under a 3.3 V power supply voltage and 100 MHz clock, the energy consumption ratio of a Flash array implementing a 3 × 3 convolution kernel reaches 0.18 TOPS/W, and the peak signal-to-noise ratio (PSNR) of the convolution calculation result is 39.05 dB.

Key words: NOR Flash; computation in memory; convolution acceleration

www.pcachina.com

0　引言

深度学习在人脸识别、音频识别、图像分类等领域中得到广泛应用。深度学习网络具有大量的权重数据和大量的乘累加操作,极大的算力需求和功耗限制使得深度学习应用难以部署在物联网终端设备。而在深度学习网络中,卷积计算占用前向计算 89% 的时间,随之产生巨大的功耗[1]。所以高速、低功耗的卷积计算单元的设计成为迫切的需求。

主流的冯诺依曼架构中,计算单元和内存单元是两个完全分离的单元,计算单元根据指令从内存读取数据,在计算单元完成计算,再存回内存。数据需要在计算单元和存储单元之间进行频繁的移动,因此带来较大的功耗和较低的运算效率。存算一体架构将计算单元与内存单元合二为一,在存储数据的同时完成运算,从而极大地减少了计算过程中数据存取的时间和功耗。实现存算一体化的介质有相变存储 PCM[2],静态随机存储 SRAM[3]、浮栅器件 Flash[4]等。Flash 具有工艺成熟、成本低等特点,因此本设计采用 Flash 作为存算一体的介质。具体做法是将卷积核的权值映射到 Flash 阵列的阈值电压,然后 Flash 阵列进行高速、低功耗的模拟乘累加计算来加速卷积计算过程。

本文的主要内容在于:(1)利用 Flash 的线型区 I/V 特性,设计基于 NOR Flash 的模拟矩阵计算单

元;(2)基于模拟矩阵计算单元,设计了基于 NOR Flash 的模拟卷积计算单元;(3)通过 SOBEL 边缘检测算子评估基于 NOR Flash 的卷积计算单元的性能。

1 基于 NOR Flash 的矩阵计算单元

1.1 基于 Flash 的模拟乘法器

基于 Flash 的模拟乘法器由两个 Flash 单元组成,分别命名为正 Flash 和负 Flash,如图 1 所示。两个 Flash 的栅极(G 端)相连,接固定电位;漏端(D 端)相连,接输入电压 V_{DS};正负 Flash 的源端(S 端)电流相减得到电流输出 I_D。数字输入 A 由数模转换器 DAC 转换为模拟电压 V_{DS},电流输出 I_D 由模数转换器 ADC 转换为数字输出 C。

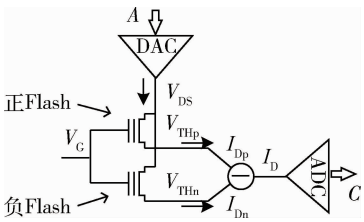


图 1 基于 Flash 的模拟乘法器电路原理图

利用 Flash 在线性区下的 I/V 特性:

$$I_{Dp} = u_n C_{ox} \frac{W}{L'}\left[(V_{GS} - V_{THp})V_{DS} - \frac{1}{2}V_{DS}^2\right] \quad (1)$$

$$I_{Dn} = u_n C_{ox} \frac{W}{L'}\left[(V_{GS} - V_{THn})V_{DS} - \frac{1}{2}V_{DS}^2\right] \quad (2)$$

式(1)、式(2)相减得:

$$I_D = I_{Dp} - I_{Dn} = u_n C_{ox} \frac{W}{L'} \cdot \Delta V_{TH} \cdot V_{DS} \quad (3)$$

式中:I_D 为正负 Flash 的电流差,$u_n C_{ox}$ 为工艺常数,W/L' 为 Flash 的有效宽长比,V_{THp} 为正 Flash 的阈值电压,V_{THn} 为负 Flash 的阈值电压,ΔV_{TH} 为两个 Flash 的阈值电压差,V_{DS} 为两个 Flash 的 D 端电压,V_{GS} 为两个 Flash 的 G 端电压。

假设 C,A,W_{weight} 与 I_D,V_{DS},ΔV_{TH} 的比例关系为:

$$\begin{cases} I_D = k_1 C \\ V_{DS} = k_2 A \\ \Delta V_{TH} = k_3 W_{weight} \end{cases} \quad (4)$$

则:

$$\begin{cases} C = K \cdot A \cdot W_{weight} \\ K = u_n C_{ox} \dfrac{W}{L'} \dfrac{k_2 k_3}{k_1} \end{cases} \quad (5)$$

式中:C 为乘积结果,W_{weight} 为权值,A 为乘数,K 为乘积系数。

上述推导表明利用 Flash 在线性区下的 I/V 特性能构成模拟乘法器。预置权值 W_{weight} 通过 Flash 的阈值电压 V_{TH} 改变,V_{TH} 可以通过 Flash 的隧穿效应或者热电子效应来进行编程和擦除,简而言之,就是通过控制 Flash 的 S、G、D 端的电压,来调节 V_{TH} 的大小。一个 Flash 能够存储 1 ~ 4 bit 数据[5],一对确定阈值电压差 ΔV_{TH} 的 Flash 相当于一个常系数的乘法器。

图 2 为该模拟乘法器在 SMIC 65 nm 浮栅工艺的 SPICE BSIM3 模型下的仿真结果。V_{GS} 固定电压 7 V,V_{DS} 的输入电压范围在 0 ~ 65 mV,选用 17 种不同的 ΔV_{TH}。计算结果 I_D 在 -4 μA ~ 4 μA 之间,I_D 随 V_{DS} 线性变化,最大非线性误差为 3.21%,用 bit 来衡量为 4 bit,所以 Flash 能完成 4 bit 的乘法运算。

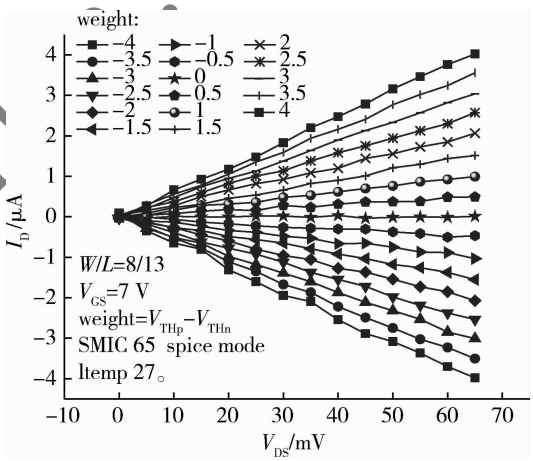


图 2 基于 Flash 的模拟乘法器的 SPICE 仿真

1.2 基于 Flash 的模拟乘累加单元

基于 Flash 的模拟乘累加单元由 n 列个 Flash 模拟乘法器组成,共有 $2 \times n$ 个 Flash,可存储 n 个权重。两行 Flash 的 G 端相连,接固定电位;每列 Flash 的 D 端相连,接模拟输入信号 $[V_1, \cdots, V_n]$;两行 Flash 阵列的 S 端电流差 I_D 为模拟乘累加输出。数字输入 $[a_1, \cdots, a_n]$ 由 DAC 转换成模拟输入 $[V_1, \cdots, V_n]$,电流输出 I_D 由 ADC 转换为数字输出 c,如图 3 所示。

模拟乘累加单元使用 Flash 的线型区 I/V 特性来实现模拟乘法操作,Flash 的 S 端相连完成电流加法操作。相关计算公式如下:

www.pcachi.com

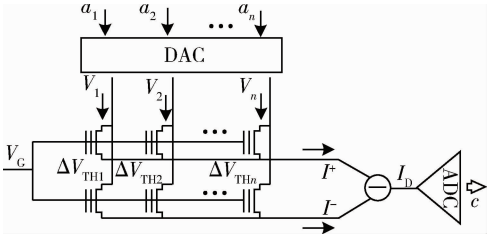


图 3　基于 Flash 的模拟乘累加单元电路原理图

$$I_{\mathrm{D}} = u_{\mathrm{n}} C_{\mathrm{ox}} \frac{W}{L'} \cdot [V_1, \cdots, V_n] \cdot [\Delta V_{\mathrm{TH1}}, \cdots, \Delta V_{\mathrm{THn}}]^{\mathrm{T}} \tag{6}$$

假设使用公式(4)的转换关系,则:

$$c = K \cdot [a_1, a_2, \cdots, a_n] \cdot [w_1, w_2, \cdots, w_n]^{\mathrm{T}} \tag{7}$$

1.3　基于 NOR Flash 的模拟矩阵计算单元

基于 NOR Flash 的模拟矩阵计算单元由 m 行模拟乘累加单元构成,共有 $2m \times n$ 个 Flash,可存储 $m \times n$ 个权重。每个 Flash 以并联方式连接,可以对每个 Flash 进行独立的编程操作。每行乘累加单元的 G 端相连,构成模拟矩阵计算单元的字线 WL 控制端;每列 Flash 的 D 级相连,接入模拟电压信号 $[V_1, \cdots, V_n]$,构成阵列的位线 BL 输入端;m 行乘累加电路输出 m 个模拟运算结果 $[I_{\mathrm{D1}}, \cdots, I_{\mathrm{Dm}}]$。数字输入 $[a_1, \cdots, a_n]$ 由 DAC 转换成模拟输入 $[V_1, \cdots, V_n]$,电流输出 $[I_{\mathrm{D1}}, \cdots, I_{\mathrm{Dm}}]$ 由 ADC 转换为数字输出 $[c_1, \cdots, c_m]$。该单元能够完成向量点乘矩阵计算,如图 4 所示。

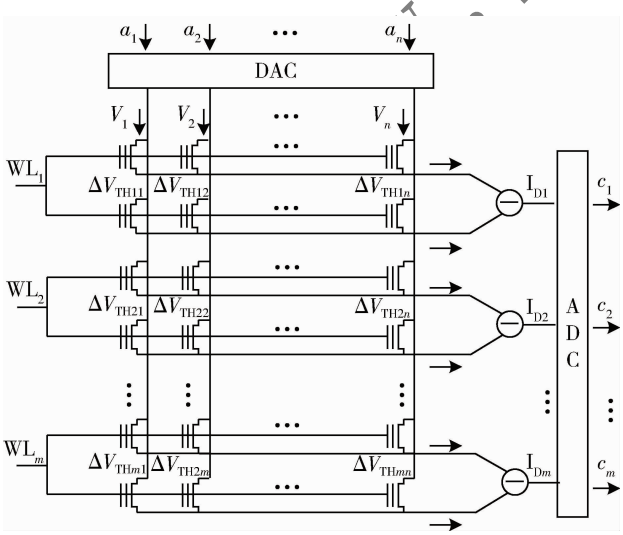


图 4　基于 NOR Flash 的模拟矩阵计算单元

相关的计算公式为:

$$[I_{\mathrm{D1}} \cdots I_{\mathrm{Dm}}] = u_{\mathrm{n}} C_{\mathrm{ox}} \frac{W}{L'} \cdot [V_1 \cdots V_n] \cdot \begin{bmatrix} \Delta V_{\mathrm{TH11}} & \cdots & \Delta V_{\mathrm{TH1}n} \\ \vdots & \ddots & \vdots \\ \Delta V_{\mathrm{TH}m1} & \cdots & \Delta V_{\mathrm{TH}mn} \end{bmatrix}^{\mathrm{T}} \tag{8}$$

代入式(4),得

$$[c_1, \cdots, c_m] = K \cdot [a_1, \cdots, a_n] \cdot \begin{bmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{m1} & \cdots & w_{mn} \end{bmatrix}^{\mathrm{T}} \tag{9}$$

2　基于 NOR Flash 的卷积运算单元

2.1　卷积运算的原理及特点

卷积计算是现代图像处理和深度学习的基础运算。卷积运算从输入图像的左上角开始,开一个卷积核大小的滑动窗口,滑动窗口与卷积核对应元素相乘后相加,用计算结果代替窗口中心数值,滑动窗口经过从左到右从上至下扫描后,得到输出图像。卷积运算的公式为[6]:

$$\boldsymbol{C}(m,n) = \sum_{i=0}^{S-1} \sum_{j=0}^{T-1} \boldsymbol{A}(m-i, n-j) \boldsymbol{B}(i,j) \tag{10}$$

式中 $\boldsymbol{A}(m,n)$ 表示 $M \times N$ 的单通道图片,$\boldsymbol{B}(s,t)$ 表示 $S \times T$ 的卷积核。

卷积计算本质上是矩阵计算。图 5 描述了卷积运算转化为向量点乘矩阵计算的过程。T 个 $k \times k$ 卷积核变换成 $[t, k \times k]$ 的权重矩阵。$m \times n$ 的图像变换成 $(m-k+1) \times (n-k+1)$ 个长度为 $k \times k$ 的输入向量。向量点乘矩阵计算结果为 $(m-k+1) \times (n-k+1)$ 个长度为 t 的输出向量,输出向量变换成 t 个 $(m-1) \times (n-1)$ 输出图片。

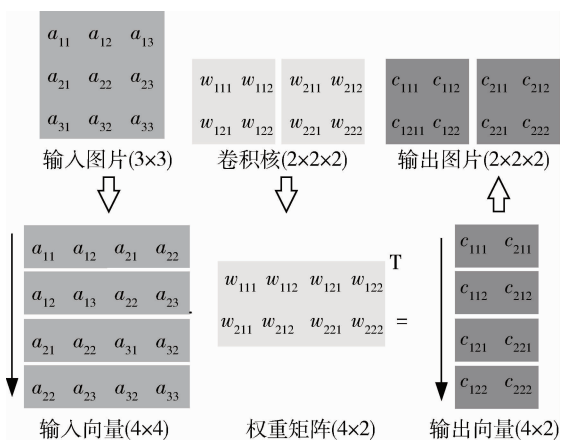


图 5　卷积运算转换为向量点乘矩阵运算

bpachina.com

2.2 基于 NOR Flash 的卷积运算单元

基于 NOR Flash 的卷积运算单元,主要由输入数据缓冲器、数模转换器 DAC、模拟矩阵计算单元、字线控制信号产生器、模拟选通器、模数转换器 ADC、输出数据缓冲器构成。输入数据缓冲器按照卷积计算规律将输入图片转换为多个输入向量,DAC 将输入矩阵转换为模拟电压,模拟矩阵计算单元完成高度并行的乘累加操作,字线控制信号可以控制当前行的乘累加操作是否有效,ADC 将模拟计算结果转换为数字结果,模拟选通器通过切换 S 端与 ADC 的连线实现 ADC 的分时复用,输出数据缓冲器完成输出矩阵到输出图片的转换。卷积运算的系统框图如图 6 所示。

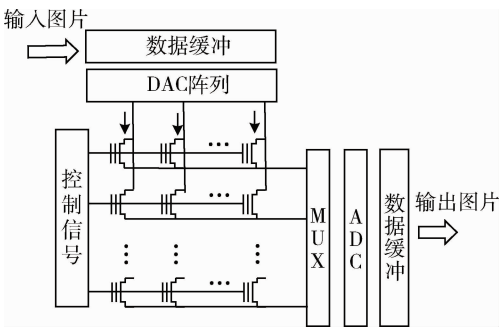


图 6 基于 NOR Flash 的卷积运算的系统框图

数据缓冲单元是将输入图片按照卷积计算规律转换为向量输入的关键电路。图 7 的缓冲单元将大小为 $m \times n$ 输入图片进行相应的向量转换。该单元为 k 行移位寄存器加寄存器的结构,行与行之间串联。每行移位寄存器的长度为$(n-k)$,寄存器的个数为 k。输入图片从第一行的移位寄存器输入,$k \times k$ 个寄存器构成滑动卷积窗口,形成 DAC 阵列的矩阵输入。

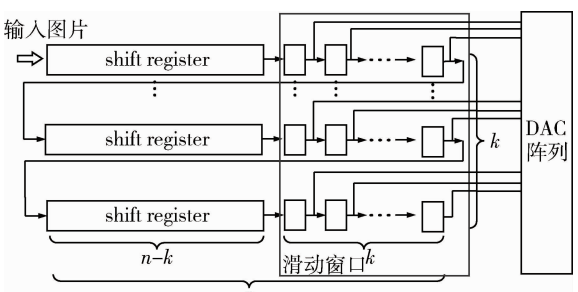


图 7 缓冲单元的结构

基于 NOR Flash 的卷积运算流程分为权重存储和卷积计算两步。第一步进行权重存储,卷积核按照图 5 所示方法展开转换为权重矩阵,然后转换为阈值电压差映射到 Flash 阵列上。第二步进行卷积运算,输入图片通过数据缓冲器转换为数个输入向量,DAC 将其转为模拟电压。模拟矩阵计算单元进行高度并行的模拟计算,输出电流经过 ADC 转换为输出向量。输出向量经过数个时钟周期得到完整的输出图片。图 8 为基于 NOR Flash 的卷积运算的流程图。

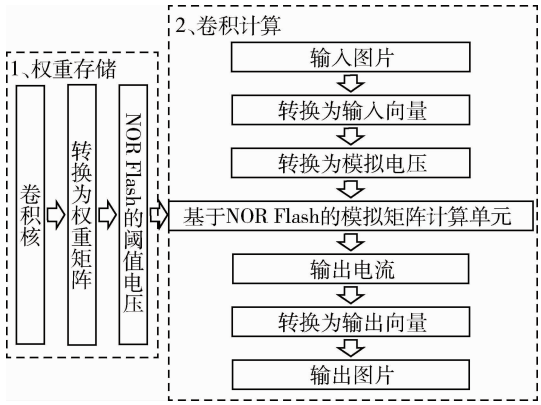


图 8 基于 NOR Flash 的卷积运算的流程图

3 基于 NOR Flash 的边缘检测算法

用 Sobel 边缘检测算法评估基于 NOR Flash 的卷积计算单元的性能。边缘检测可以将周围像素灰度有阶跃变化的像素检测出来,Sobel 算子包括两组卷积核,检测水平边缘的 $\boldsymbol{B}_x$ 算子,检测垂直边缘的 $\boldsymbol{B}_y$ 算子。Sobel 边缘检测主要方法就是将输入图片灰度图片分别经过 $\boldsymbol{B}_x$ 算子,$\boldsymbol{B}_y$ 算子进行卷积运算后得到灰度图[7]。图 9 为 640 × 480 的灰度图片经过 Sobel 边缘检测的 MATLAB 软件仿真。

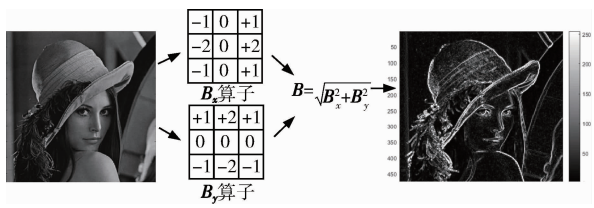


图 9 SOBEL 边缘检测的 MATLAB 软件仿真

图 10 为基于 NOR Flash 的 Sobel 边缘检测验证电路的组成。输入 640 × 480 大小的 4 bit 灰度图经过数据缓冲单元产生输入向量。DAC 精度为 4 bit,最大输出幅值为 65 mV。两行 Flash 的 ΔV_{TH} 作为权重参数,具体分配如表 1 所示。

电流转电压(ITV)电路固定 Flash 阵列的 S 端电压的同时,将微安级别的电流转换为毫伏级别的电压。求差电路(SUB)将两阵列的电压相减,相关公式如下:

www.pachina.

表 1　权重与阈值电压分配

权重	-1	-2	0	1	2
正 Flash	V_{THb}	V_{THb}	V_{THb}	V_{THb}-1	V_{THb}-2
负 Flash	V_{THb}-1	V_{THb}-2	V_{THb}	V_{THb}	V_{THb}

注：VTHb 为 Flash 阈值电压中间值

$$V_o = \left(\frac{R_1 + R_4}{R_1}\right)\left(\frac{R_3}{R_2 + R_3}\right)V_{i2} - \frac{R_4}{R_1}V_{i1} \tag{11}$$

当 $\frac{R_4}{R_1} = \frac{R_2}{R_3} = 1$ 时：

$$V_o = V_{i2} - V_{i1} \tag{12}$$

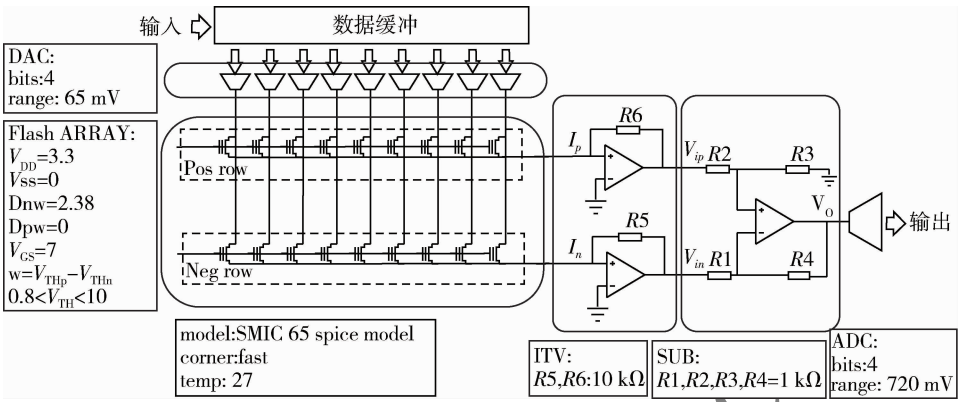


图 10　基于 NOR Flash 的边缘检测的 SPICE 仿真结构

ADC 将模拟计算结果转换为数字结果，分辨率为 4 bit。

使用 SMIC 65 nm 浮栅工艺的 SPICE BSIM3 模型进行 HSPICE 仿真。图 11 为 $\boldsymbol{B}_x$ 算子的卷积运算的部分仿真结果，bl0 ~ bl8 为 9 个 DAC 的数字输入，V_O 为模拟结果输出，out 为数字结果输出，sign 为符号位。将 $\boldsymbol{B}_x$ 算子、$\boldsymbol{B}_y$ 算子的卷积结果合成图片，得到基于 NOR Flash 的 Sobel 边缘检测结果，如图 12(a)所示。图 12(b)为基于 NOR Flash 的 Sobel 边缘检测与标准边缘检测结果相差的噪声。峰值信噪比 PSNR 是一种全参考的图像质量评价指标，能用来评价一幅图片与参考图片相比的质量[7]。本文使用 PSNR 来评估基于 NOR Flash 的 Sobel 边缘检测与标准边缘检测相比的质量。PSNR 值为 39.05 dB，接近 40 dB，说明图像噪声小，非常接近标准计算结果。

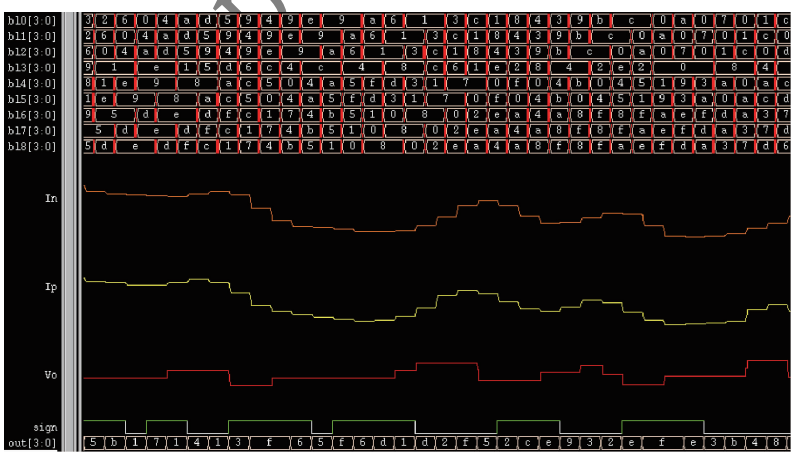
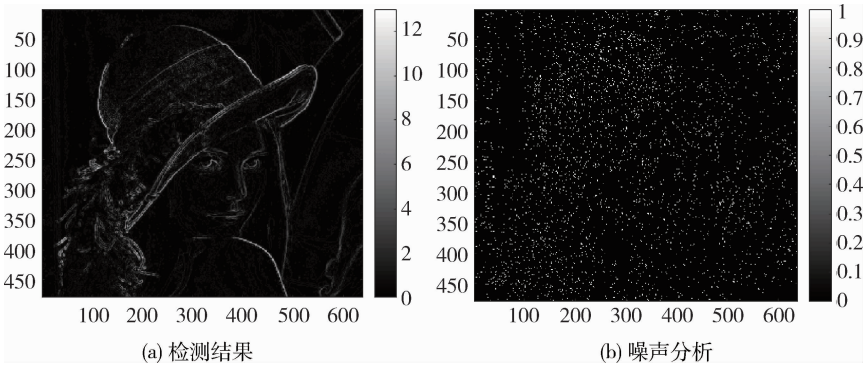


图 11　基于 NOR Flash 的 $\boldsymbol{B}_x$ 算子卷积的部分 SPICE 仿真结果

4　系统评估

表 2 为卷积计算单元的性能比较。文献[8]为使用 Xilinx Zynq-7000 FPGA 实现 Sobel 边缘检测；文献[3]为基于 SRAM 使用存算一体架构来加速卷积神经网络。本设计能够实现输入 4 bit，权重 4 bit 的模拟卷积运算，2 ×9 的 Flash 阵列能够实现一个 3 ×3 卷积核，一个时钟周期能够完成 18 次乘加运算，在 100 MHz 时钟下，算力能够达到 1.8 GOPS，功耗为 9.8 mW，Flash 阵列的能耗比达到 0.18 TOPS/W。可以看到，相对于传统数字架构实现，该架构有较大



(a) 检测结果　　(b) 噪声分析

图 12　基于 NOR Flash 的 Sobel 边缘检测的结果与噪声分析

的性能提升,但是由于此阵列规模较小,并且外围电路占用较多的能耗,因此相对于基于 SRAM 的存算一体架构实现,性能还有提升的余地,但是后续可以通过扩大阵列规模提高算力,减小外围电路的功耗,来达到更高的性能。

表 2　卷积计算单元的性能比较

	文献[8]	文献[3]	本文
工艺/nm	28	65	65
频率/MHz	95	364	100
架构	数字	存算	存算
输入/权重精度/bit	8/4	7/1	4/4
算力/GOPS	3.6	10.7	1.8
功耗/mW	162	0.38	9.8
能耗比/TOPS/W	0.02	28.4	0.18

注:1MAC = 2OPS(ADD + MUL)

5　结论

本文提出一种基于 NOR Flash 的卷积计算单元电路,能够高效率地完成卷积计算。在 SMIC 65 nm 浮栅工艺,100 MHz 时钟,3.3 V 电源电压下,实现一个 3×3 卷积核的 Flash 阵列能耗比能够达到 0.18 TOPS/W。后续将选用合适的卷积神经网络算法部署至该阵列,同时扩大该阵列的规模,提高该单元的性能。该设计对使用 Flash 来实现存算一体具有参考作用。

参考文献

[1] Jia Yangqing. Learning semantic image representations at a large scale[D]. University of California, Berkeley, 2014.

[2] SEBA S, ABU L G, MANUEL B, et al. Tutorial: brain-inspired computing using phase-change memory devices [J]. Journal of Applied Physics, 2018, 124(11): 111101. 1-111101.15.

[3] AVISHEK B, ANANTHA P C. Conv-RAM: an energy-efficient SRAM with embedded convolution computation for low-power CNN-based machine learning applications [C]. IEEE Solid-State Circuits, 2018.

[4] Han Runze, Huang Peng, Xiang Yachen. A novel convolution computing paradigm based on NOR flash array with high computing speed and energy [C]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2019.

[5] CAMPARDO G, MICHELONI R, NOVOSEL D. VLSI-design of non-volatile memories [J]. Springer-Verlag Berlin Heidelberg, 2005.

[6] GONZALEZ R C, WOODS R E. 数字图像处理[J]. 阮秋琦,译. 北京:电子工业出版社,2010.

[7] SIM J Y, PARK J S, KIM M H, et al. A 1.42 TOPS/W deep convolutional neural network recognition processor for intelligent IoE systems [C]. International Solid-State Circuits Conference, 2017.

[8] NGUYEN T K H, CECILE B, TUAN V P. Performance and evaluation sobel edge detection on various methodologies [J]. International Journal of Electronics and Electrical Engineering, 2014, 2(1): 15-20.

(收稿日期:2020-03-05)

作者简介:

徐伟民(1995 -),男,硕士,主要研究方向:存算一体、数字 IC 设计。

黄鲁(1961 -),男,硕士,副教授,主要研究方向:数模混合高速接口集成电路设计。

蒋明峰(1994 -),男,硕士,主要研究方向:存算一体、模拟 IC 设计。

www.pcachina.com

版权声明

经作者授权，本论文版权和信息网络传播权归属于《信息技术与网络安全》杂志，凡未经本刊书面同意任何机构、组织和个人不得擅自复印、汇编、翻译和进行信息网络传播。未经本刊书面同意，禁止一切互联网论文资源平台非法上传、收录本论文。

截至目前，本论文已经授权被中国期刊全文数据库（CNKI）、万方数据知识服务平台、中文科技期刊数据库（维普网）、JST 日本科技技术振兴机构数据库等数据库全文收录。

对于违反上述禁止行为并违法使用本论文的机构、组织和个人，本刊将采取一切必要法律行动来维护正当权益。

特此声明！

《信息技术与网络安全》编辑部

中国电子信息产业集团有限公司第六研究所