

**BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ KỸ THUẬT TP.HCM**

LÊ MINH

**NGHIÊN CỨU TỐI ƯU VÀ THỰC THI HỆ ĐIỆN TOÁN MÔ PHỎNG
HỆ THẦN KINH DÙNG ĐIỆN TRỞ NHỚ**

Chuyên ngành: Kỹ thuật điện tử
Mã số chuyên ngành: 9520203

TÓM TẮT LUẬN ÁN TIẾN SĨ

TP. HỒ CHÍ MINH – NĂM 2025

Công trình được hoàn thành tại **Trường Đại học Công nghệ Kỹ thuật
Tp.HCM**

Người hướng dẫn khoa học 1: PGS. TS. Trương Ngọc Sơn

Phản biện 1:

Phản biện 2:

Phản biện 3:

Luận án sẽ được bảo vệ trước Hội đồng đánh giá luận án Cấp Trường họp tại
Trường Đại học Công nghệ Kỹ thuật Tp.HCM
vào ngày tháng năm

DANH MỤC CÔNG TRÌNH ĐÃ CÔNG BỐ

A. Tạp chí quốc tế:

- (1) **M. Le**, T. K. H. Pham, and S. N. Truong, "Noise and Memristance Variation Tolerance of Single Crossbar Architectures for Neuromorphic Image Recognition," *Micromachines*, vol. 12, p. 690, June 2021 (SCIE Q2, ISSN: 2072-666X).
- (2) **M. Le** and S. N. Truong, "Research on the Impact of Data Density on Memristor Crossbar Architectures in Neuromorphic Pattern Recognition," *Micromachines*, vol. 14, p. 1990, 27 October 2023 (SCIE Q2, ISSN: 2072-666X)
- (3) **M. Le** and S. N. Truong, "A Memristor Crossbar-Based XNOR-Net Architecture for Image Recognition", *Engineering, Technology & Applied Science Research*, vol. 15, no. 4, pp. 25127–25132, August 2025. (Scopus Q2, ISSN: 2241-4487)
- (4) **M. Le** and S. N. Truong, "Optimizing Memristor Crossbar for Neuromorphic Image Recognition by Introducing a Column-Wise Constant Term Circuit", *Journal of Semiconductor Technology and Science*, pp. 598-609, October 2025. (SCI Q3, ISSN: 1598-1657)

B. Hội nghị quốc tế:

- (5) **M. Le** and S. N. Truong, "Neuromorphic Character Recognition using The Single Memristor Crossbar Array," in *2021 International Conference on System Science and Engineering (ICSSE)*, Ho Chi Minh City, Vietnam, 2021, pp. 433-436. (eISBN: 978-1-6654-4848-2)
- (6) **M. Le** and S. N. Truong, "Memristor Crossbar Circuits for Neuromorphic pattern Recognition," in *2021 18th International SoC Design Conference (ISOCC)*, Jeju Island, Korea, 2021, pp. 221-222. (eISBN: 978-1-6654-0174-6)
- (7) **M. Le** and S. N. Truong, "Statistical Study on Data Dependency of Memristor Crossbar Architectures for Neuromorphic Computing," in *2023 8th International Scientific Conference on Applying New Technology in Green Buildings (ATiGB)*, Danang, Vietnam, 2023, pp. 57-61. (eISBN: 979-8-3503-4397-7)

CHƯƠNG 1 GIỚI THIỆU

1.1 Tổng quan về hướng nghiên cứu

Hệ điện toán mô phỏng hệ thần kinh (Neuromorphic Computing System) được đề xuất vào năm 1990 bởi giáo sư Carver A. Mead [1], hướng tới các vi mạch “bắt chước” cách não người xử lý thông tin, hay đơn giản hơn là thực thi các mạng nơ-ron nhân tạo trên phần cứng vi mạch, thay vì phần mềm dựa vào kiến trúc CPU hay GPU. Ban đầu, chúng đã được thiết kế và thực thi trên các hệ thống vi mạch tích hợp cỡ lớn (VLSI) và đạt được các kết quả nổi bật [2-5]. Tuy nhiên, các công nghệ chế tạo cho các hệ thống VLSI hiện nay đều dựa trên công nghệ CMOS vốn đã đạt tới giới hạn về kích thước, khả năng phát triển cũng như khả năng giảm kích thước công nghệ còn rất thấp [6-8]. Do đó, cần có một linh kiện thiết kế vi mạch thay thế nhằm khắc phục được các giới hạn này.

Sau khi được đề xuất vào năm 1971 bởi Giáo sư Leon Chua [9] và thực nghiệm thành công vào năm 2008 bởi Giáo sư R. S. Williams và các cộng sự [10], các mảng điện trở nhớ đã trở thành giải pháp tiềm năng để thực thi các hệ điện toán mô phỏng hệ thần kinh nhờ vào khả năng ghi nhớ và thay đổi giá trị giống với đặc tính các khớp nối thần kinh [11-13], kết nối thành mảng hai chiều [14-21] và khả năng thiết kế theo kiến trúc 3-D để hiện thực hóa sự kết nối cao của hệ thống nơ-ron sinh học [22-26]. Hướng tiếp cận sử dụng mảng điện trở nhớ tương tự thực thi mạng nơ-ron nhân tạo tuy cũng đạt được một số kết quả nhất định [27-33] nhưng vẫn còn nhiều khó khăn và thách thức liên quan tới đặc tính phi tuyến của linh kiện này [31-33, 36-45]. Các mảng điện trở nhớ nhị phân chỉ sử dụng 2 mức điện trở nhớ nên mang lại hiệu quả cao hơn, đặc biệt trong các mạng nơ-ron nhị phân [24, 44, 46-48], ít chịu ảnh hưởng của sự biến thiên giá trị [44] nhờ tỷ lệ HRS/LRS khá cao [25, 49,50] và có thể thực thi các phép toán logic theo định luật Kirchhoff và định luật Ohm [13, 51]. Do vậy, sử dụng các mảng điện trở nhớ nhị phân là hướng nghiên cứu nhiều tiềm năng và khả thi hơn cho các hệ điện toán mô phỏng hệ thần kinh, đặc biệt là ở các bài toán như nhận dạng tiếng nói [44, 52], nhận dạng ảnh [47, 51, 53-55].

Ở trong nước, hiện tại, việc nghiên cứu và phát triển các ứng dụng với điện trở nhớ vẫn chưa được phát triển mạnh. Các nghiên cứu đề xuất để nhận dạng ký tự [56] hoặc thực thi mạng nơ-ron [57] được thiết kế dựa trên 2 mảng điện trở nhớ nên chưa tối ưu về kích thước mảng và công suất tiêu thụ trên mảng so với các kiến trúc dùng 1 mảng điện trở nhớ. Những năm gần đây, việc nghiên cứu chế tạo linh kiện điện trở nhớ và mảng đã đạt được nhiều thành tựu nổi bật [58, 59].

Để thực thi các ứng dụng nhận dạng mẫu dùng các mảng điện trở nhớ nhị phân, đã có một số cấu trúc mảng điện trở nhớ nhị phân được đề xuất như: kiến trúc mảng điện trở nhớ bù [44], kiến trúc mảng điện trở nhớ đồng nhất [47], kiến trúc mảng điện trở nhớ đơn [48]. Tuy nhiên, ở các kiến trúc này, ảnh hưởng của các yếu tố như độ biến thiên điện trở của mảng, nhiễu trên dữ liệu, mật độ dữ liệu đến hiệu quả hoạt động vẫn chưa được phân tích và đánh giá cụ thể. Do đó, việc phân tích và đánh giá các yếu tố ảnh hưởng, nghiên cứu và tìm ra một thiết kế tối ưu của kiến trúc mảng điện trở nhớ nhị phân trong đó giảm thiểu số lượng điện trở nhớ, đảm bảo hiệu quả hoạt động của mảng để thực thi hệ điện toán mô phỏng hệ thần kinh là một nghiên cứu quan trọng, cấp thiết, tiềm năng.

1.2 Tính cấp thiết của đề tài và ý nghĩa của đề tài

Việc nghiên cứu tối ưu kiến trúc mảng điện trở nhớ nhị phân để ứng dụng trong thực thi hệ điện toán mô phỏng hệ thần kinh cho bài toán nhận dạng ảnh là một nghiên cứu cấp thiết, quan trọng và tiềm năng trong việc ứng dụng công nghệ mới thay thế cho công nghệ CMOS truyền thống, vốn đã đạt tới giới hạn của sự phát triển, trong phát triển các vi mạch hệ điện toán mô phỏng hệ thần kinh. Luận án này sẽ nghiên cứu và đề xuất một thiết kế tối ưu cho kiến trúc mảng điện trở nhớ nhị phân ứng dụng trong thiết kế hệ điện toán mô phỏng hệ thần kinh cho bài toán nhận dạng ảnh. Trong đó, một mảng điện trở nhớ nhị phân duy nhất sẽ được sử dụng và được thiết kế tối ưu để thực thi phương trình toán học đầy đủ của mảng XNOR và đảm bảo ổn định hiệu suất nhận dạng.

Về ý nghĩa khoa học, đề tài đã nghiên cứu và đề xuất một thiết kế tối ưu cho kiến trúc mảng điện trở nhớ nhị phân để thực thi nhận dạng ảnh trong hệ điện

toán mô phỏng hệ thần kinh, khắc phục được ảnh hưởng của mật độ dữ liệu đến hoạt động nhận dạng của nghiên cứu trước đó. Thiết kế mạng điện trở nhớ tối ưu này cũng được ứng dụng để thiết kế một mạng XNOR nhiều lớp cho bài toán nhận dạng ký tự viết tay. Về mặt thực tiễn, đề tài sẽ là tiền đề cho những nghiên cứu chuyên sâu hơn về thiết kế các hệ điện toán mô phỏng hệ thần kinh ứng dụng mảng điện trở nhớ thay cho vi mạch CMOS; là cơ sở cho việc xây dựng các phần cứng thực thi các ứng dụng trí tuệ nhân tạo; góp phần cố gắng mô hình hóa hoạt động của não bộ con người ở mức tiếp cận phần cứng.

1.3 Mục tiêu đề tài

Đề tài được thực hiện với các mục tiêu sau:

- Phân tích và đánh giá mức độ ảnh hưởng của các yếu tố: nhiễu trên dữ liệu đầu vào và sự biến thiên điện trở đến hiệu suất hoạt động của các kiến trúc mảng điện trở nhớ nhị phân khác nhau trong nhận dạng ảnh, làm cơ sở cho việc phát triển và đề xuất một kiến trúc mảng điện trở nhớ tối ưu ứng dụng trong thiết kế hệ điện toán mô phỏng hệ thần kinh thực thi ứng dụng nhận dạng ảnh.
- Phân tích và đánh giá sự ảnh hưởng của mật độ dữ liệu đến hoạt động của các kiến trúc mảng điện trở nhớ nhị phân khi thực thi mảng XNOR nhận dạng ảnh; tìm ra nguyên nhân gây ra sự ảnh hưởng và vấn đề cần được tối ưu của mảng điện trở nhớ nhị phân trong thực thi nhận dạng ảnh.
- Thực hiện tối ưu và đề xuất một thiết kế mạch tối ưu cho kiến trúc mảng điện trở nhớ nhị phân ứng dụng trong thiết kế hệ điện toán mô phỏng hệ thần kinh cho bài toán nhận dạng ảnh, thực thi được phương trình toán học đầy đủ của mảng XNOR chỉ với một mảng điện trở nhớ nhị phân duy nhất, khắc phục được sự ảnh hưởng của mật độ dữ liệu và đảm bảo ổn định hoạt động nhận dạng của hệ thống. Từ đó, tiến hành nghiên cứu đề xuất một thiết kế mạng nơ-ron XNOR ứng dụng kiến trúc mảng điện trở nhớ nhị phân đã tối ưu cho bài toán nhận dạng ký tự viết tay trên tập dữ liệu MNIST.

1.4 Đối tượng nghiên cứu

Các đối tượng nghiên cứu của đề tài là: hệ điện toán mô phỏng hệ thần kinh; điện trở nhớ; mảng điện trở nhớ; vi mạch ứng dụng mảng điện trở nhớ để thực thi các hệ điện toán mô phỏng hệ thần kinh; các kiến trúc mảng điện trở nhớ nhị phân khác nhau; các yếu tố ảnh hưởng đến hiệu suất hoạt động và giải pháp thiết kế tối ưu cho kiến trúc mảng điện trở nhớ nhị phân thực thi nhận dạng ảnh.

1.5 Phương pháp nghiên cứu

Các phương pháp nghiên cứu đã được sử dụng là: phương pháp khảo sát lý thuyết; phương pháp phân tích, tổng hợp lý thuyết; phương pháp tính toán, thiết kế; phương pháp mô phỏng; phương pháp phân tích, tổng hợp kết quả.

CHƯƠNG 2 CƠ SỞ LÝ THUYẾT

Chương này trình bày các cơ sở lý thuyết liên quan đến vấn đề nghiên cứu của luận án, bao gồm: cấu tạo điện trở nhớ, mô hình hóa điện trở nhớ, phân loại điện trở nhớ, hệ điện toán mô phỏng hệ thần kinh và các kiến trúc mảng điện trở nhớ nhị phân.

CHƯƠNG 3 PHÂN TÍCH ẢNH HƯỞNG CỦA CÁC YẾU TỐ ĐẾN HIỆU SUẤT HOẠT ĐỘNG CỦA CÁC KIẾN TRÚC MẢNG ĐIỆN TRỞ NHỚ NHỊ PHÂN

3.1 Giới thiệu

Trong chương này, luận án phân tích những thách thức trong việc sử dụng các mảng điện trở nhớ với điện trở nhớ được lập trình nhiều mức (tương tự), điện trở nhớ được sử dụng ở 2 mức (nhị phân, hoặc số); phân tích các yếu tố ảnh hưởng đến hiệu suất hoạt động của các kiến trúc mảng điện trở nhớ nhị phân. Từ đó làm cơ sở lựa chọn một kiến trúc mảng điện trở nhớ tối ưu cho các thiết kế hệ điện toán mô phỏng hệ thần kinh ở các nghiên cứu tiếp theo.

3.2 Những thách thức khi xây dựng các hệ điện toán mô phỏng hệ thần kinh với mảng điện trở nhớ tương tự

Đặc tính phi tuyến của giá trị điện dẫn của điện trở nhớ đã dẫn tới khó khăn trong việc lập trình giá trị các điện trở nhớ tương tự đạt đến giá trị mong muốn

một cách chính xác. Những khó khăn trong thiết kế các mạch học và giải thuật học tương ứng để cho phép quá trình huấn luyện được thực hiện trực tiếp trên phần cứng cũng là một cản trở. Sự biến thiên giá trị điện trở của các điện trở nhớ trong mảng cũng gây ảnh hưởng nhiều đến hiệu suất nhận dạng. Do vậy, việc phát triển một hướng tiếp cận mới, sử dụng linh kiện điện trở nhớ nhị phân để thiết kế các hệ điện toán mô phỏng hệ thần kinh là một nhu cầu cấp bách.

3.3 Phân tích ảnh hưởng của nhiễu và sự biến thiên điện trở đến hiệu suất hoạt động của các kiến trúc mảng điện trở nhớ nhị phân trong hệ điện toán mô phỏng hệ thần kinh

3.3.1 Phân tích mô hình toán học của các kiến trúc mảng điện trở nhớ nhị phân

Kiến trúc mảng điện trở nhớ bù sử dụng trong các ứng dụng nhận dạng được cấu thành từ 2 mảng điện trở nhớ bù nhau $M +$ và $M -$ để nhận dạng vector đầu vào A bằng phép toán XNOR theo phương trình (2.6) [44].

Ngõ ra Y sẽ được tính toán theo phương trình chi tiết sau:

$$\begin{aligned}
 Y &= A \cdot (M +) + A' \cdot (M -) \\
 &= [a_0 \ a_1 \ \dots \ a_{(n-1)}] \cdot \begin{bmatrix} g_{0,0} & g_{0,1} & \dots & g_{0,(m-1)} \\ g_{1,0} & g_{1,1} & \dots & g_{1,(m-1)} \\ \vdots & \vdots & \vdots & \vdots \\ g_{(n-1),0} & g_{(n-1),1} & \dots & g_{(n-1),(m-1)} \end{bmatrix} \\
 &\quad + [a'_0 \ a'_1 \ \dots \ a'_{(n-1)}] \cdot \begin{bmatrix} g'_{0,0} & g'_{0,1} & \dots & g'_{0,(m-1)} \\ g'_{1,0} & g'_{1,1} & \dots & g'_{1,(m-1)} \\ \vdots & \vdots & \vdots & \vdots \\ g'_{(n-1),0} & g'_{(n-1),1} & \dots & g'_{(n-1),(m-1)} \end{bmatrix} \quad (3.2) \\
 &= [y_0 \ y_1 \ \dots \ y_{m-1}] \\
 &\text{trong đó } y_j = y_j^+ + y_j^- = \sum_{i=0}^{n-1} (a_i g_{i,j} + a'_i g'_{i,j})
 \end{aligned}$$

Vector Y sẽ chứa kết quả là các dòng điện ngõ ra cột thể hiện mức độ giống nhau giữa ngõ vào A và các cột của mảng điện trở nhớ.

Ở kiến trúc mảng điện trở nhớ đồng nhất, hàm XNOR ở phương trình (2.6) đã được triển khai như ở phương trình (2.7) [47]:

Trong phương trình (2.7), phần tử A' là một đại lượng hằng số, không có sự tương tác với mảng M và được lược bỏ khi thực hiện hàm XNOR [47]. Lúc này, phương trình(2.7) được được thể hiện tương đương như sau:

$$Y = \overline{A \oplus M} = A \cdot (M +) - A' \cdot (M +) \\ = [y_0 \quad y_1 \quad \cdots \quad y_{m-1}]$$

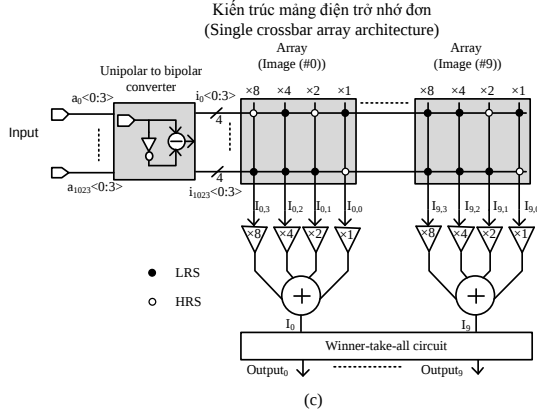
$$trong \text{ đó: } y_j = y_j^+ - y_j^- = \sum_{i=0}^{n-1} (a_i g_{i,j} - a'_i g_{i,j}) \quad (3.3)$$

Do đó kiến trúc mảng điện trở nhớ đồng nhất sử dụng 2 mảng điện trở nhớ giống nhau, 2 mảng $M +$, để lưu trữ các dữ liệu mẫu và thực thi phép XNOR với vector ngõ vào A và thu được vector Y chứa các dòng điện ngõ ra cột.

Ở kiến trúc mảng điện trở nhớ đơn, hàm XNOR được khai triển như ở phương trình (2.9) [48]. Trong phương trình (2.9), A' là một hằng số và có thể được lược bỏ. Do đó, để thực thi phép XNOR, phương trình (2.9) đã được giản lược thành một phép AND như ở phương trình (2.10) [48]. Trong phương trình (2.10), $I = [i_0 \quad i_1 \quad \cdots \quad i_{(n-1)}]$ là vector đầu vào với kích thước $1 \times n$ chứa dữ liệu đầu vào lưỡng cực. Phương trình (2.10) của kiến trúc mảng điện trở nhớ đơn khi biểu diễn ở dạng ma trận sẽ là:

$$Y = \overline{A \oplus M} = (A - A') \cdot M = I \cdot M \\ = [i_0 \quad i_1 \quad \cdots \quad i_{(n-1)}] \cdot \begin{bmatrix} g_{0,0} & g_{0,1} & \cdots & g_{0,(m-1)} \\ g_{1,0} & g_{1,1} & \cdots & g_{1,(m-1)} \\ \vdots & \vdots & \vdots & \vdots \\ g_{(n-1),0} & g_{(n-1),1} & \cdots & g_{(n-1),(m-1)} \end{bmatrix} \\ = [y_0 \quad y_1 \quad \cdots \quad y_{m-1}] \quad (3.4) \\ trong \text{ đó } y_j = \sum_{k=0}^{n-1} i_k g_{k,j}$$

Từ đó, mô hình toán học của kiến trúc mảng điện trở nhớ đơn chỉ sử dụng 1 mảng điện trở nhớ duy nhất để thực hiện phép XNOR và thu được vector Y chứa kết quả là các dòng điện ngõ ra cột.

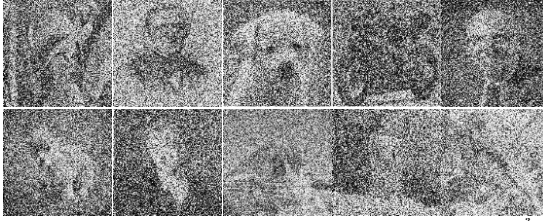


Hình 3.2: Sơ đồ khối của ứng dụng nhận dạng 10 ảnh xám kích thước 32×32 :
(a) với kiến trúc mảng điện trở nhớ bù, (b) với kiến trúc mảng điện trở nhớ đồng nhất và (c) với kiến trúc mảng điện trở nhớ đơn.

Sơ đồ khối của ứng dụng nhận dạng ảnh với 3 kiến trúc mảng điện trở nhớ này được trình bày ở hình 3.5. Để nhận dạng ảnh đầu vào sẽ được chuyển đổi thành vector A với kích thước 1×1024 điểm ảnh. Mỗi điểm ảnh sau đó sẽ được số hóa 4 bit để thực hiện phép XNOR với các mảng điện trở nhớ đã lưu trữ trước 10 ảnh dữ liệu, như ở các như ở phương trình (3.2), (3.3), (3.4). Các dòng điện ngõ ra cột sẽ được đưa vào mạch tìm dòng điện lớn nhất để tìm ra dòng lớn nhất I_k .

3.3.3 Phân tích ảnh hưởng của nhiễu và sự biến thiên điện trở đến hiệu suất hoạt động của các kiến trúc mảng điện trở nhớ nhị phân

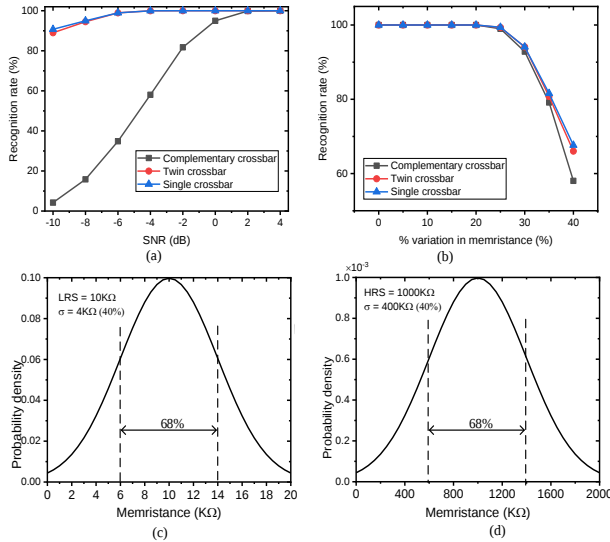
Để đánh giá ảnh hưởng của nhiễu đến hiệu suất nhận dạng của 3 kiến trúc mảng điện trở nhớ, các ảnh đầu vào sẽ được cộng thêm nhiễu Gauss trước khi đưa vào nhận dạng. Hình 3.6 thể hiện các ảnh đầu vào dưới tác động của nhiễu Gauss với tỉ số tín hiệu trên nhiễu (SNR) là -10 dB. Tỉ số tín hiệu trên nhiễu (SNR) của nhiễu Gauss cộng vào các ảnh được thay đổi từ -10 dB đến 4 dB. Kết quả tỉ lệ nhận dạng đúng của 3 kiến trúc trên các ảnh đầu vào có nhiễu Gauss được so sánh ở hình 3.7(a). Kết quả phân tích ở hình 3.7(a) cho thấy kiến trúc mảng điện trở nhớ đơn và kiến trúc mảng điện trở nhớ đồng nhất vẫn duy trì được tỉ lệ nhận dạng đúng khá cao, trong khi đó tỉ lệ nhận dạng đúng của kiến trúc mảng điện trở nhớ bù giảm rõ rệt khi tỉ số SNR là -10 dB.



Hình 3.3: Mười ảnh xám được sử dụng để nhận dạng với nhiễu Gauss có SNR là -10 dB.

Cụ thể, khi tỉ số SNR là -10 dB thì tỉ lệ nhận dạng đúng của kiến trúc mạng điện trở nhớ đơn, kiến trúc mạng điện trở nhớ đồng nhất và

kiến trúc mạng điện trở nhớ bù lần lượt là 91%, 89% và 4%. Các ưu điểm này là nhờ kiến trúc mạng điện trở nhớ đồng nhất sử dụng phép toán trừ khi thực thi hàm XNOR, như ở phương trình (3.3) nên sự thay đổi trong dòng điện ngõ ra cột gây ra bởi nhiễu trên ảnh đầu vào đã được triệt tiêu bớt. Kiến trúc mạng điện trở nhớ đơn được phát triển từ kiến trúc mạng điện trở đồng nhất, như ở phương trình (3.4), nên tín hiệu nhiễu đã được triệt tiêu một phần bởi bộ chuyển đổi dữ liệu từ đơn cực sang lưỡng cực. Như vậy, kiến trúc mạng điện trở nhớ đơn có sự kháng nhiễu trên dữ liệu tốt nhất và duy trì được tỉ lệ nhận dạng cao nhất khi so sánh với hai kiến trúc còn lại.



Hình 3.4: (a) Tỷ lệ nhận dạng đúng của 3 kiến trúc với ảnh đầu vào có nhiễu Gauss, (b) tỷ lệ nhận dạng đúng của 3 kiến trúc khi có sự biến thiên của giá trị điện trở nhớ, (c) phân bố xác suất của giá trị điện trở thấp LRS, (d) phân bố xác suất của giá trị điện trở cao HRS.

Tiếp tục, độ biến thiên các giá trị điện trở nhớ được thay đổi 0% đến 40% theo phân bố Gauss, như thể hiện ở hình 3.7(c), (d). Giá trị LRS và HRS được giả định là 10 K Ω và 1 M Ω . Theo kết quả ở hình 3.7(b), khi tỷ lệ biến thiên giá trị điện trở nhớ là 40%, kiến trúc mảng điện trở nhớ đơn có tỷ lệ nhận dạng đúng là 67.8%, đạt tỷ lệ cao nhất khi so sánh với 58% và 66% của kiến trúc mảng điện trở nhớ bù và kiến trúc mảng điện trở nhớ đồng nhất. Nguyên nhân của điều này là nhờ kiến trúc mảng điện trở nhớ đồng nhất được cấu thành từ 2 mảng điện trở nhớ giống nhau và thực hiện phép toán trừ khi thực thi hàm XNOR, nên có thể triệt tiêu một phần sự thay đổi trong dòng điện ngõ ra cột gây ra bởi sự biến thiên điện trở. Kiến trúc mảng điện trở nhớ đơn chỉ sử dụng một mảng điện trở nhớ để thực thi hàm XNOR, nên có sự đáp ứng tốt nhất với sự biến thiên điện trở của mảng.

3.4 Kết luận chương

Luận án đã tìm ra được ưu điểm về sự đáp ứng tốt với nhiễu trên dữ liệu và sự biến thiên điện trở mảng của kiến trúc mảng điện trở nhớ đơn, bên cạnh lợi thế về kích thước mảng, khi so sánh với kiến trúc mảng điện trở nhớ bù và mảng điện trở nhớ đồng nhất. Do đó, việc tiếp tục nghiên cứu chuyên sâu và phát triển tối ưu cho mảng điện trở nhớ nhị phân trên cơ sở của kiến trúc mảng điện trở nhớ đơn là một nghiên cứu quan trọng và khả thi để thiết kế các hệ điện toán mô phỏng hệ thần kinh.

CHƯƠNG 4 PHÂN TÍCH SỰ PHỤ THUỘC VÀO MẬT ĐỘ DỮ LIỆU NHẪM TỐI ƯU KIẾN TRÚC MẢNG ĐIỆN TRỞ NHỚ CHO HỆ ĐIỆN TOÁN MÔ PHỎNG HỆ THẦN KINH

4.1 Giới thiệu

Chương này sẽ tiếp tục phân tích và đánh giá sự ảnh hưởng của mật độ dữ liệu đến hoạt động của các kiến trúc mảng điện trở nhớ nhị phân, phát hiện ra sự

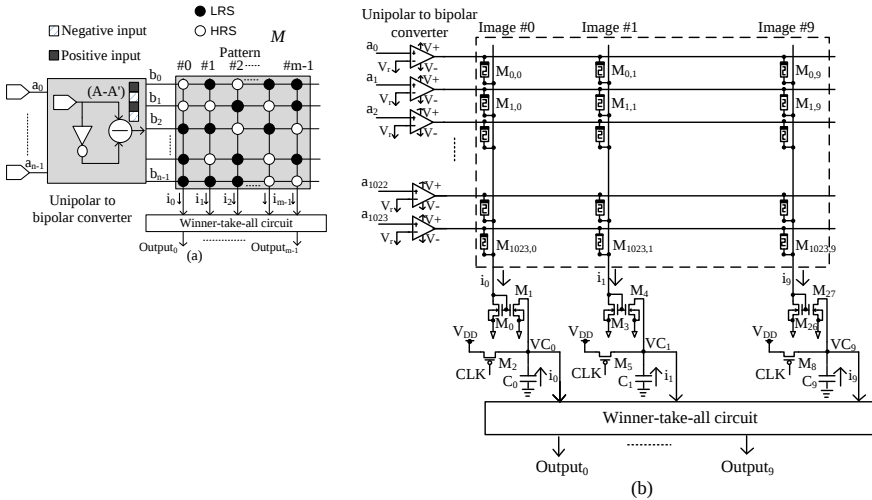
ảnh hưởng và nguyên nhân gây ra sự ảnh hưởng của mật độ dữ liệu, làm cơ sở cho việc thực hiện tối ưu kiến trúc mảng điện trở nhớ nhị phân.

4.2 Kiến trúc mảng điện trở nhớ đơn trong ứng dụng nhận dạng ảnh nhị phân

Kiến trúc mảng điện trở nhớ đơn thực thi hàm XNOR giữa ngõ vào A với một mảng điện trở nhớ duy nhất M theo phương trình sau [48]:

$$\begin{aligned} Y &= \overline{A \oplus M} = AM + A'M' = AM + A'(1 - M) \\ &= (A - A')M = IM \end{aligned} \quad (4.1)$$

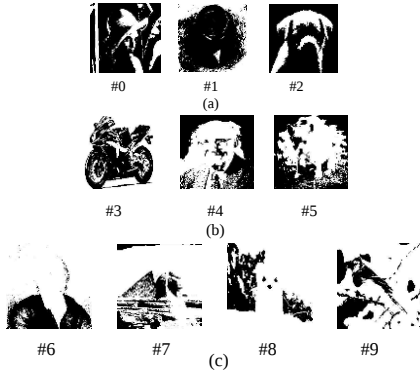
Ở đây, $I = (A - A')$ là vector dữ liệu lưỡng cực, A' là hằng số được giản lược. Sơ đồ khối và sơ đồ nguyên lý của ứng dụng nhận dạng 10 ảnh nhị phân kích thước 32×32 với kiến trúc mảng điện trở nhớ đơn được thể hiện ở hình 4.1.



Hình 4.1: (a) Sơ đồ khối và (b) sơ đồ mạch nguyên lý của kiến trúc mảng điện trở nhớ đơn trong ứng dụng nhận dạng 10 ảnh nhị phân.

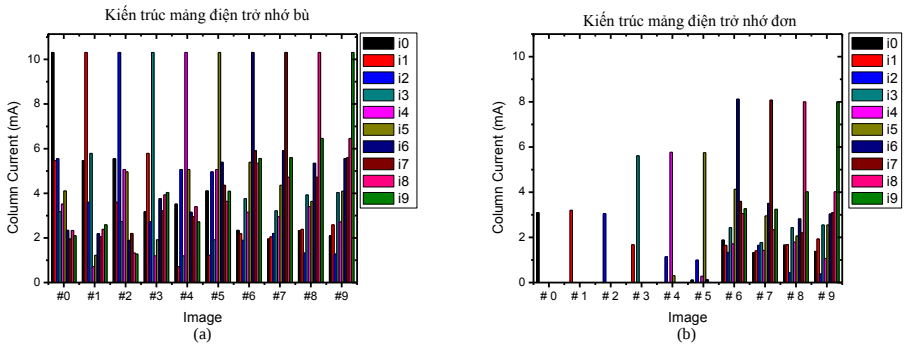
4.3 Mật độ dữ liệu trong ảnh nhị phân

Mật độ dữ liệu của một ảnh nhị phân là tỉ lệ số lượng bit 1, tức các điểm ảnh màu trắng, trong tổng số bit dữ liệu của ảnh. Nhằm tạo ra tập mẫu dữ liệu nhị phân, luận án dùng 10 ảnh xám 32×32 và chuyển đổi thành ảnh trắng đen với 3 tỉ lệ mật độ dữ liệu khác nhau, thể hiện ở hình 4.5 [83].



Hình 4.2: Mười ảnh nhị phân với các mật độ dữ liệu khác nhau: (a) mật độ dữ liệu thấp 0,25, (b) mật độ dữ liệu trung bình 0,5, (c) mật độ dữ liệu cao 0,75.

lần lượt được nhận dạng với kiến trúc mạng điện trở nhớ đơn và kiến trúc mạng điện trở nhớ bù. Với giá trị LRS (bit dữ liệu 1) và HRS (bit dữ liệu 0) được giả định là 100 K Ω và 10 M Ω , các điện áp lưỡng cực là 1 V và -1 V, các điện áp đơn cực là 1 V và 0 V. Kết quả các dòng điện ngõ ra cột của hai kiến trúc được thể hiện ở hình 4.6.



Hình 4.3: Các dòng điện ngõ ra cột khi nhận dạng các ảnh nhị phân từ ảnh thứ #0 đến ảnh thứ #9 với kiến trúc mạng điện trở nhớ bù (a) và kiến trúc mạng điện trở nhớ đơn (b).

Hình 4.6(b) cho thấy kiến trúc mạng điện trở nhớ đơn cho kết quả các dòng điện ngõ ra cột bị suy giảm khi mật độ dữ liệu trong ảnh giảm. Trong khi đó với kiến trúc mạng điện trở nhớ bù thì dòng điện ngõ ra cột lớn nhất là tương đương nhau ở các trường hợp như ở hình 4.6(a). Cụ thể, khi mật độ dữ liệu giảm từ giá

Mật độ dữ liệu lần lượt là 0,25, 0,5 và 0,75 thể hiện tỉ lệ số lượng bit 1 trong ảnh lần lượt là 25%, 50% và 75%.

4.4 Phân tích sự ảnh hưởng của mật độ dữ liệu đến hoạt động của các kiến trúc mạng điện trở nhớ khi nhận dạng ảnh nhị phân

4.4.1 Ảnh hưởng đối với dòng điện ngõ ra cột

Các ảnh nhị phân ở hình 4.5 sẽ

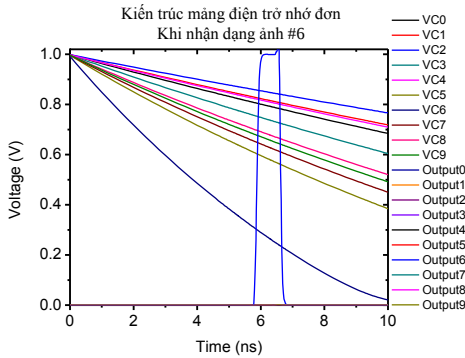
trị 0,75 xuống 0,25 thì dòng điện ngõ ra cột lớn nhất của kiến trúc mảng điện trở nhớ đơn đã giảm từ xấp xỉ 8,1 mA xuống còn xấp xỉ 3 mA, tương đương với việc dòng điện ngõ ra cột lớn nhất đã bị giảm đi xấp xỉ 3 lần. Ngược lại, ở kiến trúc mảng điện trở nhớ bù, dòng điện ngõ ra cột lớn nhất vẫn duy trì ở mức xấp xỉ 10,3 mA và không bị ảnh hưởng.

Nguyên nhân kiến trúc mảng điện trở nhớ đơn bị suy giảm dòng điện ngõ ra cột là do phương trình thực thi hàm XNOR (4.1) đã được tối giản và lược bỏ một đại lượng hằng số A' . Đặc biệt, khi một vector ảnh đầu vào A có mật độ dữ liệu thấp thì vector A' sẽ có mật độ dữ liệu cao và lượng suy giảm của dòng điện cột ngõ ra sẽ lớn. Đồng thời, phép trừ trong phương trình (4.1) có thể tạo ra giá trị âm và các giá trị âm sẽ không tạo ra dòng điện ở ngõ ra cột.

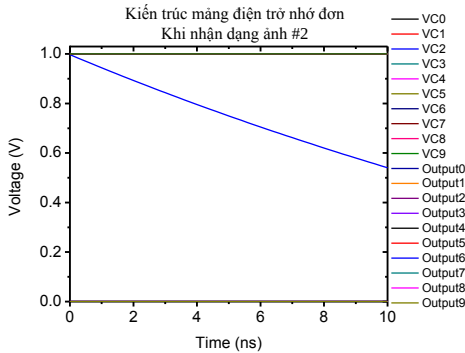
4.4.2 Ảnh hưởng đối với hoạt động nhận dạng

Các dòng điện ngõ ra cột với giá trị khác nhau sẽ làm cho các tụ điện nạp trước từ C_0 đến C_9 xả với tốc độ khác nhau, tức các điện áp xả VC_0 đến VC_9 sẽ giảm với tốc độ khác nhau. Các điện áp VC_0 đến VC_9 sẽ được đưa vào mạch tìm dòng điện lớn nhất để tìm ra điện áp giảm về giá trị ngưỡng nhanh nhất. Khi có một điện áp xả VC_k của tụ thứ k giảm về thấp hơn ngưỡng 0,5 V thì mạch tìm dòng điện lớn nhất sẽ tạo một xung chốt ở ngõ ra $Output_k$ để kết luận.

Điện áp xả của các tụ khi thực hiện nhận dạng ảnh thứ #6, ảnh có mật độ dữ liệu cao 0,75, với kiến trúc mảng điện trở nhớ đơn được phân tích và thể hiện ở hình 4.7. Lúc này, điện áp xả tụ điện VC_6 là điện áp giảm nhanh nhất, sau khoảng thời gian xấp xỉ 4 ns đã giảm nhanh nhất về ngưỡng 0,5 V. Trong khi đó, sau khoảng thời gian 4 ns, điện áp của các tụ khác vẫn còn giữ ở giá trị cao trên 0,7 V, tức là cao hơn ngưỡng 0,5 V. Do đó, mạch tìm dòng điện lớn nhất, với giá trị ngưỡng so sánh V_{REF} là 0,5 V, có thể xác định được dòng điện ngõ ra cột lớn nhất là i_6 trong khoảng thời gian khoảng 7 ns. Có thể kết luận rằng hoạt động nhận dạng của kiến trúc mảng điện trở nhớ đơn vẫn chính xác và ổn định khi ảnh nhận dạng có mật độ dữ liệu cao.



Hình 4.4: Điện áp xả trên các tụ điện, từ VC_0 đến VC_9 , khi nhận dạng ảnh thứ #6 có mật độ dữ liệu 0,75 với kiến trúc mảng điện trở nhớ đơn.



Hình 4.5: Điện áp xả trên các tụ điện, từ VC_0 đến VC_9 , khi nhận dạng ảnh thứ #2 có mật độ dữ liệu 0,25 với kiến trúc mảng điện trở nhớ đơn.

vị trí của cột có dòng điện ngõ ra lớn nhất, dẫn tới quá trình nhận dạng không được chính xác.

4.5 Kết luận chương

L luận án đã phát hiện được rằng sự suy giảm của mật độ dữ liệu có ảnh hưởng xấu đến hoạt động nhận dạng của kiến trúc mảng điện trở nhớ đơn: gây ra sự suy giảm dòng điện ngõ ra cột và làm cho hoạt động nhận dạng của kiến trúc này không còn được chính xác như ở trường hợp ảnh đầu vào có mật độ dữ liệu cao, khi xem xét trong cùng một khoảng thời gian nhận dạng. Phát hiện này là

Khi ảnh thứ #2, có mật độ dữ liệu thấp 0,25, được nhận dạng với kiến trúc mảng điện trở nhớ đơn thì kết quả điện áp xả của các tụ điện VC_0 đến VC_9 được thể hiện ở hình 4.8. Lúc này, điện áp xả VC_2 giảm chậm hơn khi so sánh với trường hợp ảnh đầu vào có mật độ cao và các điện áp trên các tụ còn lại vẫn giữ nguyên giá trị nạp trước là 1 V.

Cụ thể, xét trong cùng một khoảng thời gian nhận dạng 7 ns thì giá trị điện áp xả VC_2 vẫn còn giữ ở mức trên 0,8 V và vẫn chưa có điện áp tụ nào giảm về được dưới mức ngưỡng 0,5 V. Do đó, trong cùng một khoảng thời gian nhận dạng 7 ns, khi nhận dạng một ảnh có mật độ dữ liệu thấp 0,25 thì mạch tìm dòng điện lớn nhất trong kiến trúc mảng điện trở nhớ đơn sẽ không hoạt động đúng và không xác định được

một cơ sở rất quan trọng để cho thấy yêu cầu cấp thiết của việc tối ưu hóa kiến trúc mạng điện trở nhớ nhị phân để ổn định hiệu quả nhận dạng.

CHƯƠNG 5 THIẾT KẾ KIẾN TRÚC MẠNG ĐIỆN TRỞ NHỚ TỐI ƯU CHO HỆ ĐIỆN TOÁN MÔ PHỎNG HỆ THẦN KINH

5.1 Giới thiệu

Chương này sẽ phân tích, đề xuất một thiết kế mạch tối ưu cho kiến trúc mạng điện trở nhớ nhị phân trên cơ sở tối ưu phương trình toán học, nhằm thực thi phương trình XNOR đầy đủ để khắc phục ảnh hưởng của mật độ dữ liệu và ổn định hiệu suất khi thực thi nhận dạng ảnh. Đồng thời, một thiết kế mạng nơ-ron XNOR nhiều lớp ứng dụng kiến trúc mạng điện trở nhớ tối ưu cũng được đề xuất cho bài toán nhận dạng ký tự viết tay trên tập MNIST.

5.2 Thiết kế kiến trúc mạng điện trở nhớ tối ưu cho hệ điện toán mô phỏng hệ thần kinh

5.2.1 Thiết kế mạch bù hằng số dòng điện cột để triệt tiêu ảnh hưởng của mật độ dữ liệu

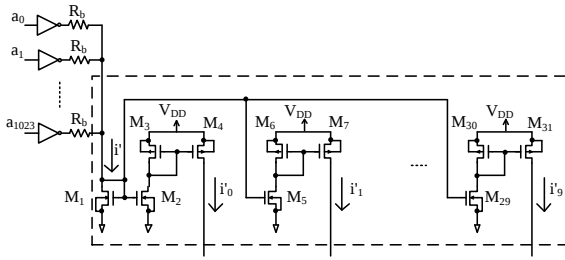
Phép toán XNOR trên kiến trúc mạng điện trở nhớ đơn đã được khai triển theo phương trình toán học như sau [48]:

$$\begin{aligned} Y &= \overline{A \oplus M} = AM + A'M' = AM + A'(1 - M) \\ &= (A - A')M + A' = (A - A')M = IM \end{aligned} \quad (5.1)$$

Trong đó, trên phương diện toán học, A' là một hằng số, nên đã được lược bỏ. Tuy nhiên, ở phương diện mạch phần cứng, sự giản lược hằng số A' đã làm giảm đi một lượng dòng điện ngõ ra bằng nhau ở tất cả các cột. Như vậy, để khắc phục, cần có giải pháp thực thi hàm XNOR đầy đủ:

$$\begin{aligned} Y &= \overline{A \oplus M} = AM + A'M' = AM + A'(1 - M) = (A - A')M + A' \\ &= (A - A') \cdot M + A' \cdot 1 \end{aligned} \quad (5.2)$$

Cụ thể, hằng số A' được biến đổi tương đương với phép AND ($A' \cdot 1$) để giữ lại. Giá trị logic 1 thể hiện cho một điện trở nhớ ở trạng thái giá trị điện trở thấp (LRS) để tạo ra dòng điện.



Hình 5.1: Mạch đề xuất thực thi phép AND ($A' \cdot 1$) ở phương trình (5.2) để bù hằng số dòng điện cột ở các cột ngõ ra nhằm tối ưu hóa kiến trúc mảng điện trở nhớ nhị phân trong nhận dạng 10 ảnh nhị phân.

Mạch thực thi phép AND ($A' \cdot 1$) sẽ được thiết kế riêng bên ngoài như một mạch ngoại vi và sử dụng linh kiện điện trở thay vì điện trở nhớ để không làm thay đổi kiến trúc của mảng điện trở nhớ, được đề

xuất ở hình 5.1 [84].

Các ảnh nhận dạng là 10 ảnh nhị phân có kích thước 32×32 . Vector điện áp đầu vào $A = [a_0 \ a_1 \ \dots \ a_{1023}]$ sẽ được đi qua các cổng đảo để tạo thành vector $A' = [a'_0 \ a'_1 \ \dots \ a'_{1023}]$ và tiếp tục đi qua các điện trở R_b để tạo ra dòng điện bù cho một cột ngõ ra i' theo phương trình:

$$i' = \sum_{k=0}^{1023} \frac{a'_k}{R_b} \quad (5.3) \quad \text{Ở đây, điện trở } R_b = \text{LRS. Dòng điện tổng } i' \text{ sẽ tiếp tục được sử dụng làm}$$

dòng điện tham chiếu và được sao chép bởi các mạch gương dòng điện để tạo ra các dòng điện bù i'_0 đến i'_9 cho các cột của mảng điện trở nhớ M để ổn định dòng điện ngõ ra cột của mảng điện trở nhớ M khi mật độ dữ liệu thay đổi.

5.2.2 Thiết kế mạch tối ưu cho kiến trúc mảng điện trở nhớ trong ứng dụng nhận dạng ảnh

Thiết kế mạch tối ưu cho kiến trúc mảng điện trở nhớ nhị phân nhằm thực thi ứng dụng nhận dạng 10 ảnh nhị phân được đề xuất được ở hình 5.2 [84]. Kiến trúc mảng điện trở nhớ đã tối ưu sẽ thực thi mảng XNOR theo phương trình toán đầy đủ là:

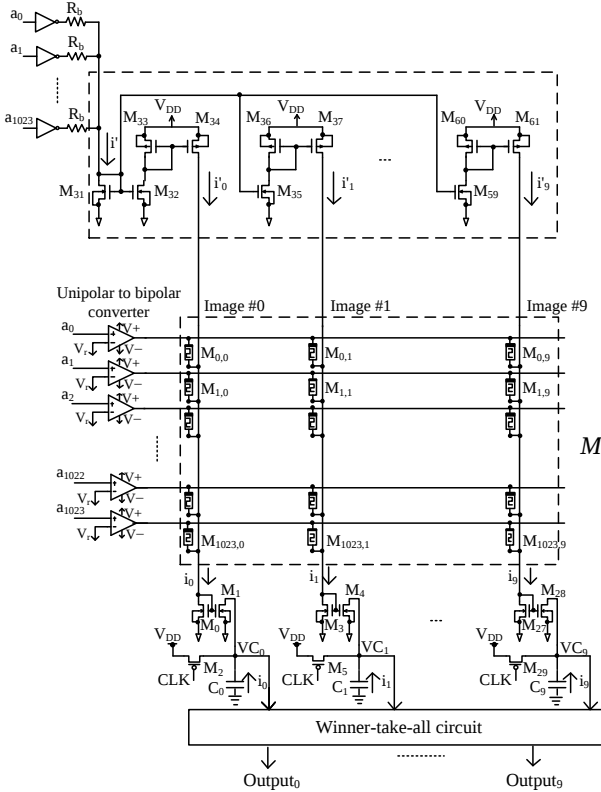
$$Y = \overline{A \oplus M} = (A - A') \cdot M + A' \cdot 1 \quad (5.4)$$

$$= [i_0 \ i_1 \ \dots \ i_{m-1}]$$

Mảng điện trở nhớ M có kích thước 1024×10 để thực hiện lưu trữ 10 ảnh mẫu 32×32 ở 10 cột. Các giá trị LRS và HRS được giả định là $100 \text{ K}\Omega$ và $10 \text{ M}\Omega$.

Vector đầu vào A được đưa qua mảng điện trở nhớ M để thực hiện phép AND thứ nhất $(A - A') \cdot M$ và cũng được đưa vào mạch bù hằng số dòng điện cột đã đề xuất để thực thi phép AND thứ hai $(A' \cdot 1)$ của phương trình (5.4) và thu được các dòng điện ngõ ra cột $[i_0 \ i_1 \ \dots \ i_{m-1}]$.

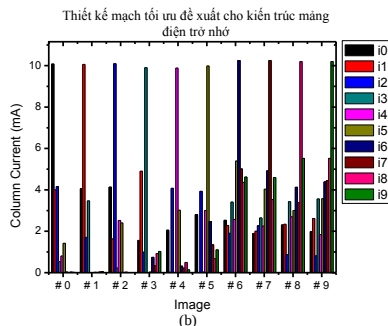
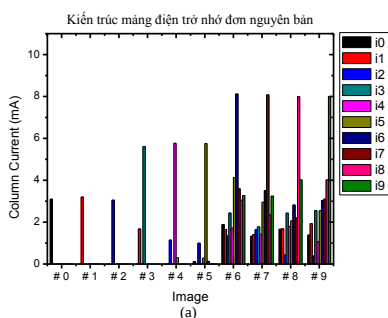
Giá trị của 10 dòng điện ngõ ra cột từ i_0 đến i_9 được thể hiện ở hình 5.4.



Hình 5.2: Thiết kế mạch tối ưu đề xuất cho kiến trúc mảng điện trở nhớ nhị phân thực hiện phương trình toán học đầy đủ (5.3) của mảng XNOR khi thực thi ứng dụng nhận dạng 10 ảnh nhị phân.

khác nhau.

Kết quả ở hình 5.4 cho thấy rằng các dòng điện ngõ ra cột tương ứng với các ảnh có mật độ dữ liệu thấp 0,25 (ảnh #0, #1, #2) và các ảnh có mật độ dữ liệu trung bình 0,5 (ảnh #3, #4, #5) đã được bù thêm, nâng lên, và có giá trị tương đương với trường hợp mật độ dữ liệu cao 0,75 (ảnh #6, #7, #8, #9). Đáng chú ý hơn, các dòng điện ngõ ra cột lớn nhất cũng được ổn định và cao gần bằng nhau, ở giá trị xấp xỉ 10 mA ở các mật độ dữ liệu

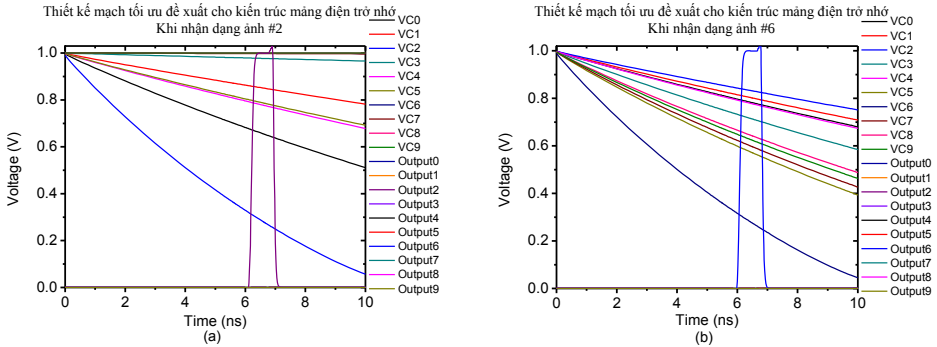


Hình 5.3: Các dòng điện ngõ ra cột khi nhận dạng các ảnh nhị phân, từ ảnh thứ #0 đến ảnh thứ #9, có mật độ dữ liệu khác nhau với (a): kiến trúc mạng điện trở nhớ đơn nguyên bản và (b) thiết kế mạch tối ưu cho kiến trúc mạng điện trở nhớ với mạch bù hằng số dòng điện cột.

Các dòng điện ngõ ra cột, từ i_0 đến i_9 , sẽ làm cho các tụ điện nạp trước C_0 đến C_9 ở hình 5.2 xả và các điện áp xả VC_0 đến VC_9 sẽ giảm với tốc độ khác nhau. Ở đây, các tụ điện được kiểm chứng có giá trị là 50 pF. Hình 5.6 thể hiện điện áp xả của các tụ VC_0 đến VC_9 cùng với xung ngõ ra *Output* của thiết kế mạch tối ưu cho kiến trúc mạng điện trở nhớ ở hình 5.2 tương ứng với ảnh thứ #2 và #6. Hình 5.6(a) cho thấy rằng mặc dù mật độ dữ liệu của ảnh thứ #2 là thấp 0,25 nhưng các điện áp xả của các tụ vẫn giảm được về ngưỡng 0,5 V và điện áp VC_2 giảm nhanh nhất về ngưỡng 0,5 V, sau xấp xỉ 4,2 ns. Mạch tìm dòng điện lớn nhất chốt được xung ngõ ra *Output2* tương ứng với điện áp VC_2 tại thời điểm 6,2 ns. Khi ảnh đầu vào là ảnh thứ #6 với mật độ dữ liệu cao (0,75) thì hình 5.6(b) cho thấy rằng điện áp VC_6 của tụ C_6 là điện áp giảm về ngưỡng 0.5 V nhanh nhất, sau khoảng 4 ns. Mạch tìm dòng điện lớn nhất đã chốt được ngõ ra *Output6* tương ứng với VC_6 là điện áp giảm nhanh nhất, tại 6,1 ns.

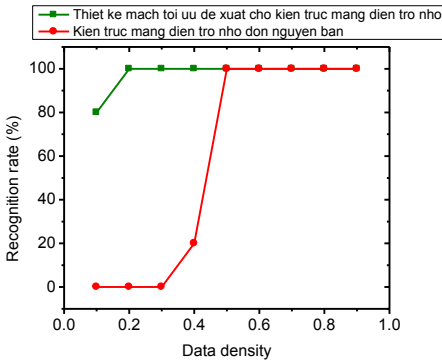
Tiếp theo, hiệu suất nhận dạng dưới sự thay đổi của mật độ dữ liệu của cũng được phân tích ở hình 5.8. Kết quả cho thấy, thiết kế mạch tối ưu cho kiến trúc mạng điện trở nhớ đề xuất đã đạt được hiệu suất nhận dạng cao và ổn định dưới sự thay đổi của mật độ dữ liệu. Trong khi đó, kiến trúc mạng điện trở nhớ đơn nguyên bản có hiệu suất nhận dạng bị suy giảm mạnh khi mật độ dữ liệu của dữ liệu đầu vào giảm đi. Khi mật độ dữ liệu giảm xuống còn 0,4, thiết kế mạch tối ưu đề xuất đã giúp ổn định và duy trì được hiệu suất nhận dạng cao ở mức

100%. Ngược lại, kiến trúc mảng điện trở nhớ đơn nguyên bản đã bị giảm còn 20%.



Hình 5.4: Tín hiệu ngõ ra Output tương ứng với điện áp xả của các tụ điện của thiết kế mạch tối ưu cho kiến trúc mảng điện trở nhớ với mạch bù hằng số dòng điện cột: (a) tương ứng với ảnh thứ #2 có mật độ dữ liệu thấp 0,25 và (b) tương ứng với ảnh thứ #6 có mật độ dữ liệu cao 0,75.

Kiến trúc được đề xuất có công suất tiêu thụ trung bình là 95,2 mW, so với 50,1 mW của kiến trúc mảng điện trở nhớ đơn nguyên bản.



Hình 5.5: Kết quả so sánh hiệu suất nhận dạng dưới sự thay đổi của mật độ dữ liệu giữa thiết kế mạch tối ưu cho kiến trúc mảng điện trở nhớ đã đề xuất và kiến trúc mảng điện trở nhớ đơn nguyên bản.

Như vậy, thiết kế mạch tối ưu cho kiến trúc mảng điện trở nhớ nhị phân đề xuất đã loại bỏ được ảnh hưởng của mật độ dữ liệu, và đảm bảo hiệu suất hoạt động của hệ thống được cao và ổn định khi thực thi nhận dạng ảnh.

5.2.3 So sánh kiến trúc mảng điện trở nhớ tối ưu đề xuất với các kiến trúc mảng điện trở nhớ nhị phân khác ứng dụng trong thiết kế hệ điện toán mô phỏng hệ thần kinh thực thi nhận dạng ảnh

L luận án đã thực hiện một bảng so sánh tổng hợp để so sánh kiến trúc mảng điện trở nhớ tối ưu đề xuất với kiến trúc mảng điện trở nhớ bù, mảng điện trở

đồng nhất và mảng điện trở đơn trong thực thi nhận dạng 10 ảnh 32×32. Kết quả so sánh được trình bày trong bảng 5.3.

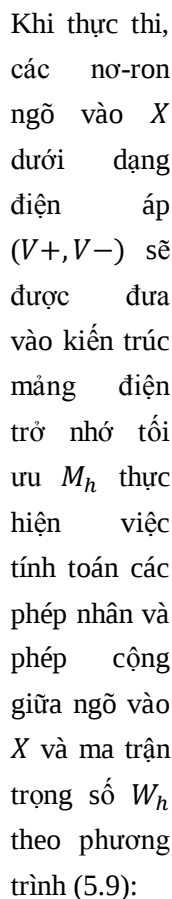
Bảng 5.3: Bảng so sánh các kiến trúc mảng điện trở nhớ nhị phân trong thực thi nhận dạng 10 ảnh 32x32.

<i>Kiến trúc</i>	<i>Số lượng memristor /synapse</i>	<i>Độ chính xác khi tỉ lệ biến thiên điện trở là 40%</i>	<i>Độ chính xác khi mật độ dữ liệu là 0,4</i>
Mảng điện trở nhớ bù [44], [54], [55]	2	50,7%	100%
Mảng điện trở nhớ đồng nhất [47]	2	58,8%	20%
Mảng điện trở nhớ đơn [48]	1	62,1%	20%
Mảng điện trở nhớ tối ưu đề xuất	1	62,1%	100%

Kết quả so sánh cho thấy kiến trúc mảng điện trở nhớ đề xuất là tối ưu nhất so với các kiến trúc còn lại khi xét về số lượng điện trở nhớ/synapse, về khả năng đáp ứng với nhiễu dữ liệu và sự độc lập với mật độ dữ liệu của tập mẫu.

5.2.4 *Thiết kế mạng nơ-ron XNOR ứng dụng kiến trúc mảng điện trở nhớ tối ưu*

Kiến trúc mảng điện trở nhớ đơn tối ưu tiếp tục được ứng dụng để thiết kế mạng nơ-ron XNOR 784×500×500×10 cho bài toán nhận dạng ký tự viết tay của tập dữ liệu MNIST [85]. Sơ đồ nguyên lý của mạng được thiết kế và trình bày ở hình 5.11. Các ma trận trọng số lớp ẩn W_h , W_l và ma trận trọng số lớp ngõ ra W_o sau khi huấn luyện sẽ được lưu vào các mảng M_h , M_l và M_o theo kiến trúc mảng điện trở nhớ tối ưu, giá trị trọng số +1 hoặc -1 được mã hóa bởi giá trị LRS hoặc HRS.


$$I = \overline{X \oplus M_h} = X \cdot M_h + \frac{X' \cdot 1}{2} = [ih_0 \ ih_1 \dots ih_{k-1}] \quad (5.9)$$

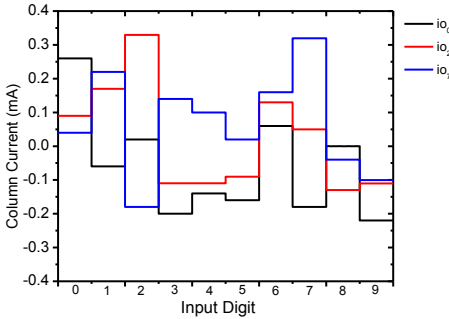
và thu được các dòng điện ngõ ra cột ih_j :

$$ih_j = \sum_{i=0}^{m-1} \left(\frac{x_i}{Mh_{i,j}} + \frac{x'_i}{2LRS} \right) = \sum_{i=0}^{m-1} x_i \times \left(\frac{1}{Mh_{i,j}} - \frac{1}{2LRS} \right) \quad (5.10)$$

Mỗi dòng điện ngõ ra cột ih_j ở lớp ẩn H sẽ được đưa qua một bộ so sánh dòng điện [83] thực thi hàm kích hoạt nhị phân cho nơ-ron h_j của lớp ẩn H như sau:

$$h_j = f(ih_j) = \begin{cases} V+, & \text{nếu } ih_j \geq 0 \\ V-, & \text{nếu } ih_j < 0 \end{cases} \quad (5.11)$$

Tại lớp ẩn L , các phép nhân và phép cộng giữa ngõ vào H và ma trận trọng số W_l được thực hiện tương tự như ở tại lớp ẩn H .



Hình 5.7: Các dòng điện ngõ ra cột io_0 , io_2 và io_7 tại lớp ngõ ra O khi các ký số từ 0 – 9 của tập MNIST lần lượt được đưa vào ngõ vào của mạng nơ-ron XNOR.

Tại lớp ngõ ra O , các nơ-ron l_j của lớp ẩn L dưới định dạng điện áp $V+$ và $V-$ sẽ tiếp tục được đưa vào mảng điện trở nhớ M_O để thực thi các phép nhân và phép cộng với ma trận trọng số ngõ ra W_O và thu được các dòng điện cột của lớp ngõ ra theo phương trình (5.20):

$$I_o = \overline{L \oplus M_O} = L \cdot M_O + \frac{L' \cdot 1}{2} = L \cdot M_O = [io_0 \ io_1 \ ... \ io_9] \quad (5.20)$$

Tại lớp ngõ ra, mạch bù hằng số dòng điện cột cho mảng điện trở nhớ M_O đã được lược bỏ để giảm các mạch ngoại vi mà không ảnh hưởng đến kết quả ngõ ra của mạng XNOR. Mạch tìm dòng điện lớn nhất (Winner-take-all) sẽ thực thi vai trò của hàm kích hoạt ngõ ra và tìm ra dòng điện ngõ ra cột lớn nhất io_j , để xuất ra kết quả nhận dạng ở cột thứ j .

Kết quả mạng nơ-ron XNOR thiết kế đạt được tỉ lệ chính xác 94% trên tập dữ liệu MNIST. Các dòng điện ngõ ra cột io_0 , io_2 và io_7 tại lớp ngõ ra O đã được

đo đạt và thể hiện ở hình 5.12 khi các ký số từ 0-9 của tập dữ liệu MNIST lần lượt được đưa vào ngõ vào của mạng để nhận dạng.

5.3 Kết luận chương

Ở chương này, luận án đã thực hiện tối ưu cho kiến trúc mạng điện trở nhớ nhị phân để có thể thực thi phương trình toán học đầy đủ của mảng XNOR khi nhận dạng ảnh chỉ với một mảng điện trở nhớ và khắc phục được ảnh hưởng của sự suy giảm mật độ dữ liệu đến hoạt động nhận dạng, là điều mà nghiên cứu trước đó chưa giải quyết được. Đồng thời, luận án của đề xuất một thiết kế mạng nơ-ron XNOR ứng dụng kiến trúc mạng điện trở nhớ tối ưu cho bài toán nhận dạng ký tự viết tay trên tập dữ liệu MNIST.

CHƯƠNG 6 KẾT LUẬN

6.1 Kết quả đạt được của luận án

Luận án đã hoàn thành được các mục tiêu đặt ra của đề tài, cụ thể như sau:

- Một là, luận án đã phân tích và đánh giá được mức độ ảnh hưởng của nhiễu trên dữ liệu đầu vào và sự biến thiên điện trở đến hiệu suất hoạt động của các kiến trúc mạng điện trở nhớ nhị phân. Cụ thể, luận án đã tìm ra được ưu điểm về đáp ứng tốt với nhiễu dữ liệu và sự biến thiên điện trở của mảng, bên cạnh lợi thế về kích thước mảng, của kiến trúc mạng điện trở nhớ đơn so với các kiến trúc còn lại. Kết quả này cho thấy việc tiếp tục tối ưu cho mảng điện trở nhớ nhị phân trên cơ sở của kiến trúc mạng điện trở nhớ đơn là một hướng nghiên cứu triển vọng và khả thi.
- Hai là, luận án đã phát hiện được rằng kiến trúc mạng điện trở nhớ đơn chịu sự ảnh hưởng xấu của sự suy giảm mật độ dữ liệu và tìm ra được nguyên nhân gây ra sự ảnh hưởng này, khi so sánh với kiến trúc mảng bù. Cụ thể, sự suy giảm của mật độ dữ liệu đã làm suy giảm đáng kể dòng điện ngõ ra cột của kiến trúc mảng điện trở nhớ đơn, làm cho mạch tìm dòng điện lớn nhất không còn hoạt động đúng và sẽ làm suy giảm hiệu suất nhận dạng của hệ thống. Đồng thời, luận án cũng đã chỉ ra được nguyên nhân của sự ảnh hưởng không tích cực

này là do sự giản lược phần tử hằng số A' trong khai triển phương trình toán học của mảng XNOR. Với các kết quả phân tích và công bố, luận án đã có một phát hiện mới và quan trọng cho thấy yêu cầu cấp thiết tối ưu kiến trúc mảng điện trở nhớ nhị phân để loại bỏ ảnh hưởng của mật độ dữ liệu, đảm bảo hiệu suất nhận dạng ổn định cho hệ thống và chỉ sử dụng một mảng điện trở nhớ để thực thi mảng XNOR.

- Ba là, luận án đã thực hiện nghiên cứu tối ưu kiến trúc mảng điện trở nhớ nhị phân để khắc phục ảnh hưởng của mật độ dữ liệu trong ứng dụng nhận dạng ảnh, đảm bảo sự ổn định của hoạt động nhận dạng mà không phụ thuộc vào tập dữ liệu và chỉ dùng một mảng điện trở nhớ để thực thi mảng XNOR. Các kết quả phân tích đã cho thấy dưới sự suy giảm của mật độ dữ liệu thì dòng điện ngõ ra cột của kiến trúc mảng điện trở nhớ đề xuất vẫn được duy trì ổn định và mạch tìm dòng điện ngõ ra lớn nhất vẫn hoạt động chính xác, tạo ra được xung kết quả ngõ ra sau khoảng thời gian 7 ns. Đồng thời, luận án cũng nghiên cứu đề xuất một thiết kế mạng nơ-ron XNOR nhiều lớp, ứng dụng kiến trúc mảng điện trở nhớ nhị phân đề xuất, cho bài toán nhận dạng ký tự viết tay sử dụng tập dữ liệu MNIST. Trong đó, ở lớp ngõ ra, mạch tạo ra hằng số dòng điện bù đã được lược bỏ mà không làm ảnh hưởng tới hiệu quả của mạng nhằm tối ưu về kích thước và công suất mạch. Mạng nơ-ron XNOR thiết kế có độ chính xác 94% trên tập dữ liệu MNIST.

6.2 Hướng phát triển

Các hướng nghiên cứu sâu hơn và phát triển hoàn thiện hơn trong tương lai là:

-Thứ nhất, tiếp tục các nghiên cứu nâng cao để giải quyết các thách thức khác trong thiết kế hệ điện toán mô phỏng hệ thần kinh ứng dụng mảng điện trở nhớ như: đặc tính phi tuyến của điện trở nhớ, sai số chế tạo, lỗi phần tử, dòng rò.

-Thứ hai, tiếp tục nghiên cứu ứng dụng kiến trúc mảng điện trở nhớ tối ưu đề xuất của luận án để thiết kế các mạng nơ-ron tích chập (CNN) nhị phân.

-Thứ ba, tiếp tục nghiên cứu thiết kế hoàn thiện các mạch ngoại vi để có thể chế tạo một vi mạch thực thi hệ điện toán mô phỏng hệ thần kinh trong học sâu.