

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ KỸ THUẬT
THÀNH PHỐ HỒ CHÍ MINH

LÊ MINH

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

LUẬN ÁN TIẾN SĨ
NGÀNH: KỸ THUẬT ĐIỆN TỬ

Tp. Hồ Chí Minh, tháng 12/2025

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ KỸ THUẬT
THÀNH PHỐ HỒ CHÍ MINH

LÊ MINH

NGHIÊN CỨU TỐI ƯU VÀ THỰC THI HỆ ĐIỆN TOÁN MÔ
PHỎNG HỆ THẦN KINH DÙNG ĐIỆN TRỞ NHỚ
NGÀNH: KỸ THUẬT ĐIỆN TỬ - 9520203

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

Phản biện 1:

Phản biện 2:

Phản biện 3:

Tp. Hồ Chí Minh, tháng 12/2025

QUYẾT ĐỊNH GIAO ĐỀ TÀI

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT
THÀNH PHỐ HỒ CHÍ MINH

Số: 260/QĐ-ĐHSPKT

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh phúc

Tp. Hồ Chí Minh, ngày 02 tháng 02 năm 2021

QUYẾT ĐỊNH

Về việc giao đề tài luận án và người hướng dẫn NCS khóa 2020 - 2023

HIỆU TRƯỞNG TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH

Căn cứ Luật Giáo dục đại học ngày 18/6/2012 và Luật sửa đổi, bổ sung một số điều của Luật Giáo dục đại học ngày 19/11/2018;

Căn cứ Nghị định 99/2019/NĐ-CP ngày 30/12/2019 của Chính phủ Quy định chi tiết và hướng dẫn thi hành một số điều của Luật sửa đổi, bổ sung một số điều của Luật giáo dục đại học;

Căn cứ Quyết định số 937/QĐ-TTg ngày 30/6/2017 của Thủ tướng Chính phủ về việc phê duyệt đề án thí điểm đổi mới cơ chế hoạt động của Trường Đại học Sư phạm Kỹ thuật TP. Hồ Chí Minh;

Căn cứ Nghị quyết số 11/NQ-HĐT ngày 08/01/2021 của Hội đồng trường ban hành Quy chế tổ chức và hoạt động của Trường Đại học Sư phạm Kỹ thuật TP. HCM;

Căn cứ Thông tư số 08/2017/TT-BGDĐT ngày 04/4/2017 của Bộ Giáo dục và Đào tạo về việc Ban hành Quy chế tuyển sinh và đào tạo trình độ tiến sĩ;

Theo nhu cầu công tác và khả năng cán bộ;

Theo đề nghị của Trường khoa/Viện quản ngành và Trường phòng Đào tạo,

QUYẾT ĐỊNH:

Điều 1. Giao đề tài luận án tiến sĩ và người hướng dẫn cho:

Nghiên cứu sinh : **Lê Minh**

Ngành : **Kỹ thuật điện tử**

Khoá: **2020 – 2023**

Tên luận án : **Nghiên cứu tối ưu và thực thi hệ điện toán mô phỏng hệ thần kinh dùng điện trở nhớ**

Người HD thứ nhất (HD chính) : **TS. Trương Ngọc Sơn**

Người HD thứ hai :

Thời gian thực hiện : **01/10/2020 đến 30/9/2023**

Điều 2. Giao cho Phòng Đào tạo quản lý, thực hiện theo đúng Quy chế đào tạo trình độ tiến sĩ của Bộ Giáo dục & Đào tạo và Nhà trường đã ban hành.

Điều 3. Trường các đơn vị, phòng Đào tạo, các Khoa/Viện quản ngành tiến sĩ và các Ông (Bà) có tên tại Điều 1 chịu trách nhiệm thi hành quyết định này.

Quyết định có hiệu lực kể từ ngày ký. 

Nơi nhận:

- BGH (để biết);
- Như điều 2, 3;
- Lưu: VT, ĐT (4b).



PGS.TS. Đỗ Văn Dũng

LÝ LỊCH CÁ NHÂN

I. Thông tin cá nhân

Họ và tên: LÊ MINH

Giới tính: Nam

Ngày sinh: 04/11/1986

Nơi sinh: Phú Khánh

II. Quá trình đào tạo

Từ 2004 đến 2009: Sinh viên đại học, khoa Điện – Điện Tử, trường Đại học Sư phạm Kỹ thuật Tp. HCM.

Từ 2009 đến 2012: Học viên cao học, khoa Điện – Điện Tử, trường Đại học Sư phạm Kỹ thuật Tp. HCM.

Từ 2020 đến nay: Nghiên cứu sinh, khoa Điện – Điện tử, trường Đại học Công nghệ Kỹ thuật Tp. HCM.

III. Quá trình công tác

Từ năm 2009 đến nay: Giảng viên, khoa Điện – Điện Tử, trường Đại học Công nghệ Kỹ thuật Tp. HCM.

Tp. Hồ Chí Minh, ngày tháng năm 2025

Lê Minh

LỜI CAM ĐOAN

Tôi cam đoan đây là công trình nghiên cứu của tôi.

Các số liệu, kết quả nêu trong Luận án là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác.

Tp. Hồ Chí Minh, ngày tháng năm 2025

Lê Minh

CẢM TẠ

Trước hết, tôi xin được chân thành cảm ơn Giáo viên hướng dẫn: Thầy PGS.TS Trương Ngọc Sơn. Thầy đã tận tình hướng dẫn, định hướng, truyền đạt cho tôi nhiều kiến thức, kỹ năng và kinh nghiệm nghiên cứu quý báu, đồng thời luôn động viên để tôi từng bước hoàn thành luận án này.

Đồng thời, tôi cũng xin cảm ơn các Thầy, Cô ở Khoa Điện – Điện tử, Viện Sau Đại học của trường Đại học Công nghệ Kỹ thuật Tp.HCM đã hướng dẫn, giúp đỡ tôi trong suốt khóa học.

Tôi cũng xin gửi lời tri ân tới Thầy, Cô, đồng nghiệp ở Bộ môn Kỹ thuật Máy tính – Viễn thông, Khoa Điện – Điện tử, trường Đại học Công nghệ Kỹ thuật Tp.HCM đã chia sẻ, động viên và hỗ trợ tôi về nhiều mặt trong thời gian qua.

Tôi cũng đặc biệt cảm ơn gia đình và bạn bè đã khích lệ tinh thần, chia sẻ và giúp đỡ tôi trong học tập cũng như trong cuộc sống.

Xin được chân thành cảm ơn và kính chúc Thầy Cô, gia đình, đồng nghiệp và các bạn của tôi được nhiều niềm vui và an lành!

TÓM TẮT

Hệ điện toán mô phỏng hệ thần kinh (Neuromorphic Computing System) hướng tới các vi mạch có khả năng tái hiện cơ chế xử lý thông tin của não bộ, cụ thể là thực thi các mạng nơ-ron nhân tạo trực tiếp trên phần cứng thay vì thông qua phần mềm dựa trên kiến trúc CPU hoặc GPU. Các mảng điện trở nhớ nhị phân là một giải pháp công nghệ mới, đầy tiềm năng và khả thi cho việc xây dựng các hệ điện toán mô phỏng hệ thần kinh thực thi các ứng dụng nhận dạng; thay thế cho giải pháp dựa trên các thiết kế mạch VLSI truyền thống vốn dĩ đã đạt tới giới hạn về kích thước và khả năng phát triển công nghệ. Trong thời gian gần đây, một số kiến trúc mảng điện trở nhớ nhị phân đã được đề xuất cho thiết kế các hệ điện toán mô phỏng hệ thần kinh trong thực thi các bài toán nhận dạng, trên cơ sở của phương trình mảng XNOR.

Tuy nhiên, việc triển khai các kiến trúc này, đặc biệt đối với các mảng có kích thước lớn, vẫn còn phải đối mặt với nhiều thách thức như sự biến thiên điện trở, nhiễu trên dữ liệu, sự phụ thuộc vào dữ liệu, lỗi phân tử trong chế tạo và đặc tính phi tuyến của điện trở nhớ. Do đó, việc phân tích các yếu tố ảnh hưởng, nghiên cứu, đề xuất một kiến trúc mảng điện trở nhớ nhị phân tối ưu – nhằm giảm thiểu số lượng điện trở nhớ và kích thước của mảng điện trở nhớ mà vẫn bảo đảm hiệu năng – là một hướng nghiên cứu có ý nghĩa quan trọng và cấp thiết trong bối cảnh tìm kiếm các công nghệ mới thay thế cho công nghệ CMOS truyền thống.

Chính vì vậy, luận án này được thực hiện nhằm mục đích nghiên cứu và đề xuất một thiết kế mảng điện trở nhớ nhị phân tối ưu cho ứng dụng trong thiết kế hệ điện toán mô phỏng hệ thần kinh thực thi nhận dạng ảnh. Căn cứ vào đó, luận án có các mục tiêu, cùng với kết quả và đóng góp mới cụ thể như sau:

Trước hết, luận án phân tích và đánh giá các mảng điện trở nhớ nhị phân ứng dụng trong thực thi hệ điện toán mô phỏng hệ thần kinh, trong đó tập trung chủ yếu vào các kiến trúc nhiều tiềm năng bao gồm: mảng điện trở nhớ bù (complementary crossbar architecture), mảng điện trở nhớ đồng nhất (twin crossbar architecture) và

mảng điện trở nhớ đơn (single crossbar architecture) khi xét về mức độ ảnh hưởng của nhiễu trên dữ liệu và sự biến thiên điện trở đến hiệu suất hoạt động trong thực thi mảng XNOR để nhận dạng. Bằng việc phân tích mô hình toán học và triển khai ứng dụng nhận dạng 10 ảnh xám với các kiến trúc mảng điện trở nhớ dưới sự ảnh hưởng của nhiễu Gauss trên dữ liệu và sự biến thiên điện trở mảng, luận án đã tìm ra được ưu điểm về đáp ứng tốt nhất với nhiễu Gauss và sự biến thiên điện trở của kiến trúc mảng điện trở nhớ đơn, bên cạnh lợi thế về kích thước mảng khi so sánh với kiến trúc mảng bù và kiến trúc mảng đồng nhất. Kết quả này của luận án đã cho thấy việc tiếp tục nghiên cứu chuyên sâu và tối ưu cho mảng điện trở nhớ nhị phân thực thi mảng XNOR trên cơ sở của kiến trúc mảng điện trở nhớ đơn là một hướng nghiên cứu triển vọng và khả thi.

Tiếp theo, luận án tiến hành phân tích ảnh hưởng của mật độ dữ liệu đến hiệu suất hoạt động của các kiến trúc mảng điện trở nhớ nhị phân khi thực thi mảng XNOR để nhận dạng, đồng thời cũng tìm ra được nguyên nhân gây ra các ảnh hưởng này. Kết quả phân tích của luận án đã chỉ ra được rằng sự suy giảm của mật độ dữ liệu sẽ gây ra sự suy giảm đáng kể cho dòng điện ngõ ra cột của kiến trúc mảng điện trở nhớ đơn, từ đó làm giảm hiệu quả nhận dạng của kiến trúc này khi so sánh với kiến trúc mảng điện trở nhớ bù. Đồng thời, luận án cũng tìm ra được nguyên nhân của ảnh hưởng xấu này là do sự giản lược một thành phần hằng số trong phương trình toán học mảng XNOR của kiến trúc mảng điện trở nhớ đơn. Đây là một phát hiện mới, quan trọng của luận án và cho thấy nhu cầu cấp thiết của việc tối ưu hóa kiến trúc mảng điện trở nhớ nhị phân nhằm loại bỏ tác động của mật độ dữ liệu khi thực thi mảng XNOR.

Trên cơ sở của các phân tích và phát hiện đã chỉ ra, luận án thực hiện nghiên cứu tối ưu kiến trúc mảng điện trở nhớ nhị phân để khắc phục ảnh hưởng của mật độ dữ liệu khi thực thi nhận dạng, đảm bảo hoạt động nhận dạng được ổn định và không phụ thuộc vào tập dữ liệu trong khi vẫn giữ ưu điểm của việc chỉ sử dụng một mảng điện trở nhớ duy nhất để thực thi mảng XNOR. Cụ thể hơn, luận án đã thiết kế một mạch bù hằng số dòng điện cột để thực thi tham số $(A' \cdot 1)$ trong khai

triển phương trình toán học của mảng XNOR nhằm ổn định các dòng điện ngõ ra cột và loại bỏ sự ảnh hưởng của sự suy giảm mật độ dữ liệu. Từ đó, luận án đề xuất một thiết kế mạch tối ưu cho kiến trúc mảng điện trở nhớ nhị phân nhằm thực thi phương trình toán học đầy đủ của mảng XNOR chỉ với một mảng điện trở nhớ duy nhất khi nhận dạng ảnh và khắc phục được ảnh hưởng của mật độ dữ liệu đến hoạt động của hệ thống.

Kiến trúc mảng điện trở nhớ tối ưu đề xuất sau đó được ứng dụng để thiết kế một mạng nơ-ron XNOR (XNOR-Net), là một kiến trúc tối ưu được nghiên cứu trong những năm gần đây nhằm cho phép triển khai hiệu quả các ứng dụng học sâu trên các phần cứng có tài nguyên giới hạn về bộ nhớ và tốc độ xử lý. Mạng XNOR-Net dựa trên kiến trúc mảng điện trở nhớ đề xuất đã được thiết kế cho bài toán nhận dạng ký tự viết tay sử dụng tập dữ liệu MNIST và được mô phỏng kiểm chứng trên phần mềm thiết kế mạch Cadence Spectre. Trong kiến trúc mạng XNOR đề xuất, để tối ưu về diện tích và công suất, mạch bù hằng số dòng điện cột tại lớp ngõ ra đã được lược bỏ khi lớp ngõ ra sử dụng mạch Tìm dòng điện lớn nhất (Winner-Take-All) mà không làm ảnh hưởng tới hiệu năng của mạch. Đây cũng chính là một cải tiến của mạch thông qua phân tích để tối ưu kiến trúc mạch, sử dụng kết hợp giữa thiết kế mảng có mạch bù dòng điện cột với thiết kế mảng không có mạch bù dòng điện cột. Mạng nơ-ron XNOR thiết kế có độ chính xác 94% trên tập dữ liệu MNIST. Kết quả mô phỏng mạch cho thấy mạng nơ-ron XNOR dựa trên mảng điện trở nhớ đề xuất có đáp ứng ngõ ra chính xác khi các ký tự từ 0 – 9 được đưa vào ngõ vào của mạng.

ABSTRACT

Neuromorphic computing systems aim to develop nanoscale circuits capable of emulating the information processing mechanisms of the human brain, specifically by implementing artificial neural networks directly on hardware rather than through software based on CPU or GPU. Binary memristor crossbar arrays are novel, promising and practical solutions for neuromorphic computing systems on recognition tasks; replacing solutions on conventional VLSI designs that approached the limits of scaling and technical-development capabilities. In recent years, several binary memristor crossbar array architectures have been proposed for designing neuromorphic computing systems to perform recognition tasks, based on the implementation of the XNOR array.

However, there are challenges associated with large-scale memristor arrays, including memristance variation, noise effect, data dependency, memristor fault due to manufacturing, the non-linear memristance characteristic ... Thus, analyzing the influencing factors and proposing an optimized binary memristor array architecture – aimed at minimizing both the number of memristors and the size of the array while preserving computational performance – constitutes a critical research direction in the pursuit of emerging technologies to replace the conventional CMOS technology.

For this reason, the dissertation aims to research and propose an optimized binary memristor crossbar array architecture applying in neuromorphic computing systems for implementing brain-inspired image recognitions. In specific, the dissertation has objectives along with results and contributions as follows:

Firstly, the dissertation analyzes and evaluates binary memristor array architectures applying neuromorphic computing systems, mainly focus on promising architectures including the complementary crossbar architecture, the twin crossbar architecture, and the single crossbar architecture in terms of impacts of noise in data and resistance variation on the performance when implementing

XNOR arrays. By analyzing mathematical models and applying image recognition of 10 grayscale images with these architectures under the effects of Gaussian noise and memristance variation, the dissertation has found that the single memristor crossbar architecture has the best tolerance of Gaussian-noise input and memristance variation, besides the advantage of array-size, compared to the complementary crossbar architecture and the single crossbar architecture. These results prove that deeper research and optimization of the binary memristor crossbar array implementing XNOR function based on the single memristor array architecture is promising and feasible.

Next, the dissertation analyzes the impact of data density on the performance of binary memristor array architectures when implementing XNOR function to recognize, and points out the reason for the effect. The dissertation reveals that the decrease on data density causes the output column currents of the single memristor crossbar architecture to reduce significantly and therefore degrades the performance of this architecture, in comparison with the complementary memristor crossbar architecture. Furthermore, the dissertation has also found that the reason for the negative impact of data density is the simplification of a constant in the equation of XNOR array. This new and important finding reveals that it is crucial to optimize the binary memristor array architecture to address the influence of data density when implementing the XNOR array.

Basing on the analyses and findings, the dissertation optimizes the binary memristor array architecture to eliminate the negative impact of data density, stabilize the recognition performance without dataset dependency while retaining the advantages of comprising only one memristor array to implement the XNOR array. In particular, the dissertation proposes a column-constant current-generating circuit to implement the parameter $(A' \cdot 1)$ in the XNOR array formulation to stabilize the column output currents and eliminate the negative impact of data density. Consequently, the dissertation proposes the optimized design for the binary memristor crossbar architecture that implements the full XNOR function with only

one memristor array for image recognition, thereby overcoming the negative impact of data density and ensuring stable recognition performance.

The proposed optimizing binary memristor array architecture is subsequently applied to design an XNOR neural network (XNOR-Net), an optimal architecture that has been researched in recent years to enable efficient implementation of deep learning applications on resource-limited devices. The memristor-based XNOR-Net is designed for handwritten digit recognition using MNIST dataset, and evaluated on Cadence Spectre circuit simulation. In the proposed memristor-based XNOR-Net, to optimize the area occupancy and power, the column-constant current-generating circuit at the output layer is excluded without affecting on the network's performance when the output layer is activated by the Winner-take-all circuit. This is also an improvement of the proposed circuit, where the optimized binary memristor crossbar arrays with column-constant current-generating circuits are combined with the counterpart without column-constant current-generating circuit. The proposed XNOR-Net demonstrates a recognition rate of 94% with the MNIST dataset. The simulation results indicate that the memristor crossbar-based XNOR-Net circuit generates the right outputs when digits 0-9 are sequentially applied at the input of the circuit.

MỤC LỤC

LÝ LỊCH CÁ NHÂN	II
LỜI CAM ĐOAN	III
CẢM TẠ	IV
TÓM TẮT	V
ABSTRACT	VIII
MỤC LỤC	XI
DANH SÁCH CÁC CHỮ VIẾT TẮT	XIV
DANH SÁCH CÁC HÌNH	XVI
DANH SÁCH CÁC BẢNG	XX
MỞ ĐẦU	1
CHƯƠNG 1 TỔNG QUAN.....	11
1.1 TỔNG QUAN VỀ HƯỚNG NGHIÊN CỨU VÀ CÁC CÔNG TRÌNH NGHIÊN CỨU CÓ LIÊN QUAN.....	11
1.2 TÍNH CẤP THIẾT VÀ Ý NGHĨA CỦA ĐỀ TÀI	18
1.3 MỤC TIÊU ĐỀ TÀI.....	20
1.4 ĐỐI TƯỢNG NGHIÊN CỨU.....	21
1.5 NHIỆM VỤ VÀ PHẠM VI NGHIÊN CỨU.....	22
1.6 PHƯƠNG PHÁP NGHIÊN CỨU	23
CHƯƠNG 2 CƠ SỞ LÝ THUYẾT	25
2.1 GIỚI THIỆU VỀ ĐIỆN TRỞ NHỚ	26
2.2 MÔ HÌNH HÓA ĐIỆN TRỞ NHỚ.....	28
2.3 PHÂN LOẠI ĐIỆN TRỞ NHỚ DỰA VÀO ĐẶC TÍNH	30
2.4 GIỚI THIỆU HỆ ĐIỆN TOÁN MÔ PHỎNG HỆ THẦN KINH.....	32
2.5 GIỚI THIỆU CÁC KIẾN TRÚC MẢNG ĐIỆN TRỞ NHỚ NHỊ PHÂN.....	37
2.5.1 Kiến trúc mảng điện trở nhớ bù	37
2.5.2 Kiến trúc mảng điện trở nhớ đồng nhất.....	39
2.5.3 Kiến trúc mảng điện trở nhớ đơn.....	41

2.6 KẾT LUẬN CHƯƠNG	44
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	45
3.1 GIỚI THIỆU	45
3.2 NHỮNG THÁCH THỨC KHI XÂY DỰNG CÁC HỆ ĐIỆN TOÁN MÔ PHỎNG HỆ THẦN KINH VỚI MẢNG ĐIỆN TRỞ NHỚ TƯƠNG TỰ	46
3.3 PHÂN TÍCH ẢNH HƯỞNG CỦA NHIỀU 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.....	50
3.3.1 <i>Phân tích mô hình toán học của các kiến trúc mảng điện trở nhớ nhị phân</i> 51	
3.3.2 <i>Thực thi nhận dạng ảnh với các kiến trúc mảng điện trở nhớ nhị phân khác nhau</i>	56
3.3.3 <i>Phân tích ảnh hưởng của nhiều và sự biến thiên điện trở đến hiệu suất hoạt động của các kiến trúc mảng điện trở nhớ nhị phân khi thực thi nhận dạng ảnh</i>	60
3.4 KẾT LUẬN CHƯƠNG	65
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.....	66
4.1 GIỚI THIỆU	66
4.2 KIẾN TRÚC MẢNG ĐIỆN TRỞ NHỚ ĐƠN TRONG ỨNG DỤNG NHẬN DẠNG ẢNH NHỊ PHÂN	68
4.3 MẬT ĐỘ DỮ LIỆU TRONG ẢNH NHỊ PHÂN.....	74

4.4	PHÂN TÍCH SỰ ẢNH HƯỞNG CỦA MẬT ĐỘ DỮ LIỆU ĐẾN HOẠT ĐỘNG CỦA CÁC KIẾN TRÚC MẢNG ĐIỆN TRỞ NHỚ KHI NHẬN DẠNG ẢNH NHỊ PHÂN	77
4.4.1	<i>Ảnh hưởng đối với dòng điện ngõ ra cột</i>	77
4.4.2	<i>Ảnh hưởng đối với hoạt động nhận dạng.....</i>	82
4.5	KẾT LUẬN CHƯƠNG	85
CHƯƠNG 5 THIẾT KẾ KIẾN TRÚC MẢNG ĐIỆN TRỞ NHỚ TỐI ƯU CHO HỆ ĐIỆN TOÁN MÔ PHỎNG HỆ THẦN KINH.....		87
5.1	GIỚI THIỆU	88
5.2	THIẾT KẾ KIẾN TRÚC MẢNG ĐIỆN TRỞ NHỚ TỐI ƯU CHO HỆ ĐIỆN TOÁN MÔ PHỎNG HỆ THẦN KINH.....	89
5.2.1	<i>Thiết kế mạch bù hằng số dòng điện cột để triệt tiêu ảnh hưởng của mật độ dữ liệu.....</i>	89
5.2.2	<i>Thiết kế mạch tối ưu cho kiến trúc mảng điện trở nhớ trong ứng dụng nhận dạng ảnh.....</i>	94
5.2.3	<i>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.....</i>	110
5.2.4	<i>Thiết kế mạng nơ-ron XNOR ứng dụng kiến trúc mảng điện trở nhớ tối ưu</i>	113
5.3	KẾT LUẬN CHƯƠNG	130
CHƯƠNG 6 KẾT LUẬN		132
6.1	KẾT QUẢ ĐẠT ĐƯỢC CỦA LUẬN ÁN	132
6.2	HƯỚNG PHÁT TRIỂN	135
TÀI LIỆU THAM KHẢO.....		137
DANH MỤC CÁC CÔNG TRÌNH ĐÃ CÔNG BỐ		147

DANH SÁCH CÁC CHỮ VIẾT TẮT

<i>Chữ viết tắt</i>	<i>Diễn giải tiếng Anh</i>	<i>Ý nghĩa tiếng Việt</i>
1M	1 Memristor	Cấu trúc khớp nối thần kinh gồm một điện trở nhớ
1P1M	1 layer of Polysilicon and 9 layers of Metal	Công nghệ chế tạo linh kiện bán dẫn với cấu trúc gồm một lớp polysilicon và chín lớp kim loại
1S1M	1 Selector 1 Memristor	Cấu trúc khớp nối thần kinh gồm một bộ lựa chọn và một điện trở nhớ
1T-1M	1 Transistor -1Memristor	Cấu trúc khớp nối thần kinh gồm một transistor và một điện trở nhớ
2T-1M	2 Transistor -1Memristor	Cấu trúc khớp nối thần kinh gồm hai transistor và một điện trở nhớ
3-D	3-Dimension	Ba chiều
ANN	Artificial Neural Network	Mạng nơ-ron nhân tạo
CIFAR-10	Canadian Institute For Advanced Research - 10	Tập dữ liệu cho nhiệm vụ phân loại ảnh, được phát triển bởi Viện Nghiên cứu Nâng cao Canada

CMOS	Complementary Metal-Oxide-Semiconductor	Công nghệ bán dẫn oxit kim loại bù nhau
CPU	Central Processing Unit	Bộ xử lý trung tâm
DCT	Discrete Cosine Transform	Biến đổi cosin rời rạc
GPU	Graphics Processing Unit	Bộ xử lý đồ họa
HRS	High Resistance State	Trạng thái điện trở cao
LO	Linear Optimization	Quy hoạch tuyến tính
LRS	Low Resistance State	Trạng thái điện trở thấp
MB	Main Block	Khối chính
MNIST	Modified National Institute of Standards and Technology	Cơ sở dữ liệu đã sửa đổi của Viện Tiêu chuẩn và Công nghệ quốc gia.
MSCCNN	Memristor-based Sparse Compact Convolutional Neural Network	Kiến trúc mạng nơ-ron tích chập nhỏ gọn dự phòng dựa trên điện trở nhớ
SNR	Signal-to-Noise Ratio	Tỷ số Tín hiệu trên Nhiễu
VLSI	Very-large-scale intergration	Mạch tích hợp quy mô rất lớn
XNOR	Exclusive NOR	Cổng logic Ex-NOR

DANH SÁCH CÁC HÌNH

HÌNH	TRANG
Hình 2.1: Mô hình một điện trở nhớ với 2 lớp màng mỏng [10].	26
Hình 2.2: (a) Mối quan hệ dòng điện và điện áp của điện trở nhớ theo thời gian và (b) đặc tính Dòng-Áp của điện trở nhớ.	28
Hình 2.3: (a) Hàm cửa sổ Joglekar với các giá trị p khác nhau và (b) hàm cửa sổ Biolek với giá trị $p = 2$.	30
Hình 2.4: (a) Cấu trúc điện trở nhớ dựa trên chuyển đổi lớp tiếp giáp [11], (b) Cấu trúc điện trở nhớ dựa trên chuyển đổi sợi.	31
Hình 2.5: Một điện trở nhớ nhị phân được chế tạo thành công theo cơ chế chuyển đổi sợi dùng vật liệu $mSiO_2$ (mesoporous silica) [49].	32
Hình 2.6: (a) Điện trở nhớ được sử dụng để mô hình hóa một khớp nối thần kinh, (b) mô hình một mảng điện trở nhớ hai chiều [11].	35
Hình 2.7: Sơ đồ khối của kiến trúc mảng điện trở nhớ bù.	38
Hình 2.8: Sơ đồ khối của kiến trúc mảng điện trở nhớ đồng nhất.	40
Hình 2.9: Sơ đồ khối của kiến trúc mảng điện trở nhớ đơn.	42
Hình 3.1: Mô hình toán học của kiến trúc mảng điện trở nhớ bù.	51
Hình 3.2: Mô hình toán học của kiến trúc mảng điện trở nhớ đồng nhất.	54
Hình 3.3: Mô hình toán học của kiến trúc mảng điện trở nhớ đơn.	56
Hình 3.4: Mười ảnh xám được sử dụng để nhận dạng.	57
Hình 3.5: Sơ đồ khối của ứng dụng nhận dạng 10 ảnh xám có kích thước 32×32 : (a) với kiến trúc mảng điện trở nhớ bù, (b) với kiến trúc mảng điện trở nhớ đồng nhất, (c) với kiến trúc mảng điện trở nhớ đơn.	59
Hình 3.6: 10 ảnh xám được sử dụng để nhận dạng với nhiễu Gauss có SNR là -10 dB.	61
Hình 3.7: (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.....	62
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.	70
Hình 4.2: Sơ đồ nguyên lý của mạch Tìm dòng điện lớn nhất.	72
Hình 4.3: (a) Sơ đồ khối và (b) sơ đồ mạch nguyên lý của kiến trúc mảng điện trở nhớ bù trong ứng dụng nhận dạng 10 ảnh nhị phân.	73
Hình 4.4: Mười ảnh xám được sử dụng để tạo thành các ảnh nhị phân với sự đa dạng về mật độ dữ liệu.	75
Hình 4.5: 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 và (c) mật độ dữ liệu cao 0,75.	76
Hình 4.6: Các dòng điện ngõ ra cột khi nhận dạng các ảnh nhị phân từ ảnh thứ #0 đến ảnh thứ #9 với kiến trúc mảng điện trở nhớ bù (a) và kiến trúc mảng điện trở nhớ đơn (b).	79
Hình 4.7: Điện áp xả trên các tụ điện, từ VC0 đến VC9 , khi nhận dạng ảnh thứ #6 có mật độ dữ liệu 0,75 với kiến trúc mảng điện trở nhớ đơn.	83
Hình 4.8: Điện áp xả trên các tụ điện, từ VC0 đến VC9 , khi nhận dạng ảnh thứ #2 có mật độ dữ liệu 0,25 với kiến trúc mảng điện trở nhớ đơn.	84
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.	92
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.	95
Hình 5.3: Mười ảnh nhị phân với mật độ dữ liệu khác nhau: (a) mật độ dữ liệu thấp 0,25 (ảnh #0, #1, #2), (b) mật độ dữ liệu trung bình 0,5 (ảnh #3, #4, #5), (c) mật độ dữ liệu cao 0,75 (ảnh #6, #7, #8, #9).	98
Hình 5.4: 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 mảng điện trở nhớ	

đơn nguyên bản, (b) thiết kế mạch tối ưu cho kiến trúc mảng điện trở nhớ với mạch bù hằng số dòng điện cột.....100

Hình 5.5: Điện áp xả của các tụ từ VC0 đến VC9 của thiết kế mạch tối ưu cho kiến trúc mảng điện trở nhớ với mạch bù hằng số dòng điện cột: (a) khi nhận dạng ảnh thứ #2 có mật độ dữ liệu thấp 0,25 và (b) khi nhận dạng ảnh thứ #6 có mật độ dữ liệu cao 0,75.102

Hình 5.6: 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) khi nhận dạng ảnh thứ #2 có mật độ dữ liệu thấp 0,25 và (b) khi nhận dạng ảnh thứ #6 có mật độ dữ liệu cao 0,75.104

Hình 5.7: Điện áp xả của các tụ điện VC0 đến VC9 và các tín hiệu ngõ ra Output khi nhận dạng ảnh thứ #2 có mật độ dữ liệu thấp 0,25 với kiến trúc mảng điện trở nhớ đơn nguyên bản khi không có mạch bù hằng số dòng điện cột đã đề xuất.106

Hình 5.8: 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.107

Hình 5.9: Các mô hình sử dụng 2 điện trở nhớ để mã hóa thông tin cho một trọng số wji nhị phân (-1, +1): (a) mô hình sử dụng 2 mảng điện trở nhớ với nguồn cung cấp lưỡng cực và (b) mô hình sử dụng cột đôi với nguồn cung cấp đơn cực.116

Hình 5.10: Mô hình mạng nơ-ron XNOR ứng dụng kiến trúc mảng điện trở nhớ tối ưu đã đề xuất để thực thi nhận dạng ảnh trong hệ điện toán mô phỏng hệ thần kinh.....118

Hình 5.11: Mạch nguyên lý mạng nơ-ron XNOR ứng dụng kiến trúc mảng điện trở nhớ tối ưu được đề xuất để thực thi nhận dạng ký tự viết tay sử dụng tập dữ liệu MNIST.124

Hình 5.12: 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.

.....129

DANH SÁCH CÁC BẢNG

BẢNG	TRANG
Bảng 2.1: Các trạng thái của vector đầu vào lưỡng cực I	42
Bảng 3.1: Thông số mô phỏng khi thực thi nhận dạng ảnh với các kiến trúc mạng điện trở nhớ nhị phân dưới sự tác động của nhiễu dữ liệu và sự biến thiên điện trở.	60
Bảng 4.1: Thông số mô phỏng của kiến trúc mạng điện trở nhớ bù và kiến trúc mạng điện trở nhớ đơn khi thực thi nhận dạng mười ảnh nhị phân 32x32.	78
Bảng 5.1: Thông số mô phỏng khi thực thi nhận dạng mười ảnh nhị phân 32x32 có mật độ dữ liệu khác nhau với kiến trúc mạng điện trở nhớ tối ưu đề xuất và kiến trúc mạng điện trở nhớ đơn nguyên bản.	99
Bảng 5.2: So sánh công suất tiêu thụ giữa kiến trúc mạng điện trở nhớ đề xuất và kiến trúc mạng điện trở nhớ đơn nguyên bản khi thực thi nhận dạng ảnh nhị phân 32x32.	108
Bảng 5.3: So sánh kiến trúc mạng điện trở nhớ tối ưu đề xuất với các kiến trúc mạng điện trở nhớ nhị phân khác trong thực thi nhận dạng 10 ảnh 32x32.	110
Bảng 5.4: Trạng thái phép XNOR giữa một ngõ vào a_i và một phần tử điện trở nhớ M_{ij} của kiến trúc mạng điện trở nhớ tối ưu đề xuất.	113
Bảng 5.5: Các trạng thái của phép XNOR giữa một ngõ vào x_i và một trọng số w_{ji} trong mạng nơ-ron nhị phân.	115
Bảng 5.6: Thông số thiết kế để thực thi mảng XNOR giữa ngõ vào X và mảng điện trở nhớ M_h ở lớp ẩn H theo phương trình XNOR (5.5).	120
Bảng 5.7: Thông số thiết kế cho thực thi mảng XNOR giữa ngõ vào X và mảng điện trở nhớ M_h của lớp ẩn H theo phương trình XNOR (5.9).	122

MỞ ĐẦU

1. Lý do chọn đề tài

Vào năm 1990, giáo sư Carver A. Mead đã đề xuất khái niệm Hệ điện toán mô phỏng hệ thần kinh (Neuromorphic Computing System) dùng để mô tả các hệ thống phần cứng có khả năng tính toán, xử lý như cách mà một bộ não sinh học thực hiện: các tín hiệu vào song song và được xử lý đồng thời. Kể từ khi được đề xuất, các hệ điện toán mô phỏng hệ thần kinh đã được thiết kế và thực thi trên các hệ thống vi mạch tích hợp cỡ lớn (VLSI). Việc xây dựng các hệ điện toán mô phỏng hệ thần kinh trên các thiết kế VLSI nhằm thực thi các mạng nơ-ron nhân tạo cũng đã đạt được nhiều kết quả nổi bật trong những thập kỷ gần đây. Tuy nhiên, các hệ thống VLSI hiện nay đều được chế tạo dựa trên công nghệ CMOS và đến nay công nghệ này đã đạt tới giới hạn về kích thước, khả năng phát triển cũng như khả năng giảm kích thước công nghệ còn rất thấp. Do đó, để thực thi các hệ điện toán mô phỏng hệ thần kinh, cần có một linh kiện thiết kế vi mạch mới và tiềm năng hơn để thay thế nhằm khắc phục được các giới hạn này.

Được đề xuất vào năm 1971 bởi Giáo sư Leon Chua và được chế tạo thực nghiệm thành công vào năm 2008 bởi Giáo sư R. S. Williams cùng các cộng sự thuộc phòng nghiên cứu HP, điện trở nhớ đã được xem như là linh kiện điện tử thứ tư (bên cạnh tụ điện, điện trở và cuộn dây) và thu hút được nhiều sự quan tâm của các nhà nghiên cứu trên thế giới. Với khả năng lưu lại giá trị điện trở hiện tại cũng như khả năng thay đổi giá trị điện trở bằng điện áp lập trình đặt vào 2 cực, điện trở nhớ có đặc tính rất giống với đặc tính của khớp nối thần kinh trong hệ thần kinh sinh học và đã trở thành một linh kiện tiềm năng để thực hiện mô hình hóa cho các khớp nối thần kinh. Việc thực hiện các phép nhân và các phép cộng tích lũy theo định luật Kirchhoff và định luật Ohm ở cấp độ phần cứng đã giúp điện trở nhớ tạo ra một sự phát triển vượt bậc trong tốc độ thực thi, công suất tiêu thụ cũng như diện tích của các thiết kế phần cứng. Đồng thời, các điện trở nhớ có thể được kết nối với nhau dưới dạng mảng 2-D và các mảng điện trở nhớ cũng có thể được thiết kế theo

kiến trúc 3-D để hiện thực hóa khả năng kết nối cao của hệ thống nơ-ron trong các bộ não sinh học. Các đặc điểm này đã làm cho các mảng điện trở trở thành một giải pháp tiềm năng trong thực thi các hệ điện toán mô phỏng hệ thần kinh, có thể thay thế cho các thiết kế VLSI truyền thống vốn đã đạt tới giới hạn của sự phát triển.

Trong thời gian gần đây, để xây dựng các hệ điện toán mô phỏng hệ thần kinh, các mảng điện trở nhớ tương tự đã được sử dụng để thực thi các mạng nơ-ron nhân tạo như là một giải pháp thiết kế tiềm năng, đạt hiệu quả cao và công suất tiêu thụ thấp. Tuy nhiên, việc xây dựng các hệ điện toán mô phỏng hệ thần kinh bằng các thiết kế trên các mảng điện trở nhớ tương tự vẫn còn nhiều khó khăn và thách thức phải đối mặt. Trước hết, có thể thấy, đặc tính phi tuyến của giá trị điện dẫn của điện trở đã dẫn tới khó khăn trong việc lập trình giá trị các điện trở nhớ tương tự để đạt đến giá trị mong muốn một cách chính xác. Đồng thời, việc 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 khá khó khăn, phức tạp và là một cản trở lớn cho việc thực thi các mạng nơ-ron nhân tạo trên các mảng điện trở nhớ tương tự. Thêm vào đó, sự biến thiên giá trị điện trở của các điện trở nhớ trong mảng, gây ra bởi sự không đồng nhất trong quá trình chế tạo hoặc sự trôi của giá trị điện trở trong quá trình hoạt động, cũng làm giảm đáng kể hiệu suất hoạt động của các ứng dụng sử dụng mảng điện trở nhớ tương tự và gây khó khăn không nhỏ cho việc thực thi các hệ điện toán mô phỏng hệ thần kinh trên các mảng điện trở nhớ tương tự.

Các mảng điện trở nhớ nhị phân, sử dụng điện trở nhớ với 2 mức giá trị: điện trở cao (High Resistance State – HRS) và điện trở thấp (Low Resistance State – LRS), đã mở ra một hướng tiếp cận mới và khả thi cho việc xây dựng các hệ điện toán mô phỏng hệ thần kinh, đặc biệt khi thực thi các mạng nơ-ron nhân tạo có tập trọng số nhị phân. Ở các mảng điện trở nhớ nhị phân, sự biến thiên giá trị điện trở do quá trình chế tạo và hoạt động sẽ ít ảnh hưởng đến hiệu suất hoạt động của hệ thống hơn, khi so sánh với điện trở nhớ tương tự, nhờ vào tỷ lệ HRS/LRS khá cao. Hơn nữa, phép nhân giữa các tín hiệu ngõ vào với các trọng số nhị phân cũng sẽ dễ

dàng được thực thi hơn bằng các phép toán logic như AND, XNOR theo định luật Kirchhoff và định luật Ohm. Như vậy, có thể thấy, các mảng điện trở nhị phân là một hướng tiếp cận mới, đầy triển vọng và khả thi cho việc xây dựng các hệ điện toán mô phỏng hệ thần kinh, đặc biệt là ở các bài toán nhận dạng mẫu như nhận dạng tiếng nói, nhận dạng ảnh.

Trong thời gian qua, đã có một số kiến trúc mảng điện trở nhớ nhị phân được đề xuất để xây dựng các hệ điện toán thần kinh thực thi nhận dạng ảnh như: kiến trúc mảng điện trở nhớ bù, kiến trúc mảng điện trở nhớ đồng nhất, kiến trúc mảng điện trở nhớ đơn. Trong đó, kiến trúc mảng điện trở nhớ đơn đã thể hiện được tính hiệu quả hơn trong công suất tiêu thụ trên mảng cũng như kích thước mảng, khi so sánh với kiến trúc mảng điện trở nhớ bù và kiến trúc mảng điện trở nhớ đồng nhất, nhờ vào việc chỉ sử dụng 1 mạng điện trở nhị phân duy nhất thay vì sử dụng 2 mảng điện trở nhớ nhị phân như các kiến trúc còn lại.

Tuy nhiên, ở các kiến trúc mảng điện trở nhớ này, ảnh hưởng của các yếu tố như độ biến thiên điện trở của mảng điện trở nhớ, nhiễu trên dữ liệu đầu vào vẫn chưa được phân tích cụ thể cũng như đánh giá mức độ ảnh hưởng đến hiệu suất hoạt động của các kiến trúc. Hơn thế nữa, sự ảnh hưởng gây ra bởi sự thay đổi của mật độ dữ liệu (mật độ các bit 1) trong mảng điện trở nhớ tới hiệu suất hoạt động của hệ thống vẫn chưa được phân tích chi tiết, đánh giá cụ thể và có phương án tối ưu để loại bỏ. Như vậy, việc phân tích các yếu tố ảnh hưởng và tối ưu hóa kiến trúc mảng điện trở nhớ là một yêu cầu quan trọng và cấp thiết trong xây dựng hệ điện toán mô phỏng hệ thần kinh với mảng điện trở nhớ nhị phân, để đảm bảo hiệu suất hoạt động ổn định của hệ thống khi thực thi nhận dạng ảnh. Do đó, ở luận án này, tác giả sẽ thực hiện tối ưu hóa kiến trúc mảng điện trở nhớ nhị phân cho thiết kế hệ điện toán mô phỏng hệ thần kinh thực thi nhận dạng ảnh. Thông qua quá trình phân tích, đánh giá cụ thể các yếu tố ảnh hưởng đến hiệu suất hoạt động của các kiến trúc mảng điện trở nhớ như: độ biến thiên điện trở, nhiễu trên dữ liệu, mật độ dữ liệu, luận án sẽ nghiên cứu đề xuất thiết kế tối ưu cho kiến trúc mảng điện trở nhớ nhị phân trong thực thi nhận dạng ảnh nhằm đảm bảo hiệu suất hoạt động ổn định của hệ thống.

Kiến trúc tối ưu của mạng điện trở nhớ sau đó được ứng dụng để thiết kế một mạng nơ-ron XNOR nhiều lớp cho bài toán nhận dạng ký tự viết tay trên tập dữ liệu MNIST. Kết quả của luận án sẽ có ý nghĩa quan trọng trong việc triển khai các hệ điện toán mô phỏng hệ thần kinh thực thi nhận dạng ảnh bằng mạng điện trở nhớ nhị phân. Đồng thời, kết quả của luận án cũng là tiền đề tốt cho các nghiên cứu chuyên sâu hơn trong thực thi các ứng dụng trí tuệ nhân tạo trên nền tảng phần cứng ở các hệ điện toán mô phỏng hệ thần kinh.

2. Mục đích nghiên cứu

Đề tài được thực hiện với mục đích tối ưu hóa kiến trúc mạng điện trở nhớ nhị phân, ứng dụng trong thiết kế hệ điện toán mô phỏng hệ thần kinh cho các bài toán nhận dạng ảnh. Để đạt được mục đích này, trước tiên, đề tài sẽ tiến hành phân tích và đánh giá sự ảnh hưởng của các yếu tố như nhiễu trên dữ liệu, độ biến thiên điện trở, ảnh hưởng của mật độ dữ liệu đến hiệu suất hoạt động của các kiến trúc mạng điện trở nhớ nhị phân khác nhau khi thực thi nhận dạng ảnh nhằm lựa chọn ra được một kiến trúc tiềm năng nhất để phát triển. Dựa trên các kết quả phân tích, luận án sẽ đề xuất một kiến trúc tối ưu nhằm khắc phục hạn chế của mạng điện trở nhớ khi hiệu năng lại phụ thuộc vào mật độ dữ liệu. Kiến trúc tối ưu được đề xuất sau đó được ứng dụng để thiết kế một mạng nơ-ron nhị phân, ứng dụng cho bài toán nhận dạng ký tự viết tay trên tập dữ liệu MNIST. Hệ thống sau đó được mô phỏng trên phần mềm thiết kế vi mạch Cadence Spectre để đánh giá mạng sau huấn luyện. Các kết quả nghiên cứu của đề tài sẽ là tiền đề quan trọng cho việc phát triển các nghiên cứu chuyên sâu hơn trong thiết kế và thực thi các hệ điện toán mô phỏng thần kinh ứng dụng kiến trúc mạng điện trở nhớ tối ưu đã đề xuất.

3. Nhiệm vụ nghiên cứu

Để đạt được các mục tiêu nghiên cứu đã đề ra, đề tài được cụ thể hóa thành các nhiệm vụ nghiên cứu chi tiết. Trước hết, Luận án sẽ tổng hợp và hệ thống hóa các kiến thức nền tảng của lĩnh vực nghiên cứu, bao gồm: điện trở nhớ, mạng điện trở nhớ, hệ điện toán mô phỏng hệ thần kinh và các hướng thực thi hệ điện toán mô

phòng hệ thần kinh. Tiếp đến, Luận án sẽ phân tích các khó khăn của hướng thực thi các hệ điện toán mô phỏng hệ thần kinh trên các mảng điện trở nhớ tương tự. Các khó khăn này sẽ cho thấy nhu cầu cấp thiết của việc phát triển các mảng điện trở nhớ nhị phân, là một hướng nghiên cứu nhiều tiềm năng và khả thi hơn trong thiết kế các hệ điện toán mô phỏng hệ thần kinh. Sau đó, Luận án sẽ phân tích chi tiết mô hình toán và cấu tạo của các mảng điện trở nhớ nhị phân thực thi mảng XNOR, nhằm hướng tới đề xuất một kiến trúc mảng điện trở nhớ nhị phân tối ưu cho thiết kế các hệ điện toán mô phỏng hệ thần kinh ở các bài toán nhận dạng. Kế tiếp, Luận án sẽ phân tích các yếu tố ảnh hưởng đến hiệu suất hoạt động của các kiến trúc mảng điện trở nhớ nhị phân, bao gồm: nhiễu trên dữ liệu, sự biến thiên điện trở của mảng. Sự ảnh hưởng của mật độ dữ liệu đến hoạt động của các kiến trúc mảng điện trở nhớ nhị phân cũng sẽ được Luận án phân tích chi tiết, nhằm đánh giá mức độ và nguyên nhân của sự ảnh hưởng. Trên cơ sở đó, Luận án sẽ tính toán, đề xuất một kiến trúc mảng điện trở nhớ nhị phân tối ưu cho các hệ điện toán mô phỏng hệ thần kinh ở các bài toán nhận dạng. Trong đó, mảng XNOR sẽ được thực thi chỉ với một mảng điện trở nhớ duy nhất, đảm bảo hiệu suất nhận dạng của hệ thống được ổn định và vượt qua được tác động của các yếu tố ảnh hưởng. Kiến trúc mảng điện trở nhớ tối ưu sau đó sẽ được ứng dụng để thiết kế một mạng nơ-ron XNOR để thực thi bài toán nhận dạng ký tự viết tay trên tập dữ liệu MNIST.

4. Phạm vi nghiên cứu

Các nhiệm vụ nghiên cứu trên của đề tài sẽ cần được đặt trong những phạm vi nghiên cứu cụ thể. Đầu tiên, linh kiện điện trở nhớ được sử dụng trong các phân tích của đề tài sẽ được giới hạn ở mô hình mô phỏng Verilog. Tiếp đến, đề tài hướng đến việc thiết kế tối ưu một kiến trúc mảng điện trở nhớ nhị phân cho xây dựng các hệ điện toán mô phỏng hệ thần kinh ở các bài toán nhận dạng và không theo hướng ứng dụng các mảng điện trở nhớ tương tự. Các yếu tố ảnh hưởng đến hiệu suất hoạt động của các kiến trúc mảng điện trở nhớ nhị phân, bao gồm: nhiễu trên dữ liệu, sự biến thiên điện trở, mật độ dữ liệu, sẽ được phân tích nhằm đề xuất một kiến trúc mảng điện trở nhớ nhị phân tối ưu với hiệu suất nhận dạng được đảm bảo ổn định.

Cuối cùng, việc phân tích và đánh giá các thiết kế trong đề tài sẽ được thực hiện bằng các công cụ mô phỏng như Matlab, Cadence.

5. Hướng tiếp cận và phương pháp nghiên cứu

Để hoàn tất các nhiệm vụ nghiên cứu và đạt được mục đích nghiên cứu, đề tài sẽ được tiếp cận theo hướng đi từ phân tích, tổng hợp lý thuyết nền tảng nhằm tìm ra vấn đề còn tồn tại của hướng nghiên cứu, đến tính toán, thiết kế tối ưu cho kiến trúc đề xuất và mô phỏng nhằm phân tích, đánh giá kết quả. Các hướng tiếp cận này sẽ được thực hiện với các phương pháp nghiên cứu cụ thể là: khảo sát lý thuyết; phân tích, tổng hợp lý thuyết; tính toán, thiết kế; mô phỏng kiểm chứng; phân tích, tổng hợp kết quả.

6. Ý nghĩa khoa học và thực tiễn của đề tài

6.1 Ý nghĩa khoa học của đề tài:

Hiện nay, trí tuệ nhân tạo đang được phát triển và kỳ vọng được áp dụng rộng rãi, mang lại hiệu quả trong nhiều lĩnh vực. Tuy nhiên, các mô hình trí tuệ nhân tạo hiện đại thường được thực thi dưới dạng các phần mềm chạy trên các hệ thống máy chủ đủ mạnh, đây là hướng tiếp cận thứ nhất. Một hướng tiếp cận thứ hai là tạo ra phần cứng có khả năng xử lý song song và mô hình hóa được hoạt động của não bộ con người, các hệ thống này được gọi là hệ điện toán mô phỏng theo hệ thần kinh (Neuromorphic Computing System). Đây là một xu hướng đang được các công ty vi mạch bán dẫn hàng đầu thế giới đang đầu tư phát triển. Đề tài này thực hiện theo hướng tiếp cận thứ hai nhằm tạo ra phần cứng có khả năng mô phỏng hoạt động của hệ thần kinh của con người. Đề tài nghiên cứu ứng dụng điện trở nhớ, một linh kiện điện tử tiềm năng đã được nghiên cứu và chế tạo thành công vào năm 2008, để thực thi một hệ điện toán mô phỏng theo hệ thần kinh. Cụ thể, đề tài đã nghiên cứu 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. Kiến trúc mạng điện trở tối ưu sau đó được ứng dụng để thiết kế một mạng nơ-ron XNOR nhiều

lớp, mở ra hướng nghiên cứu tiềm năng cho các thiết kế tối ưu hệ thống mô phỏng hệ thần kinh sử dụng linh kiện điện trở nhớ.

6.2 Ý nghĩa thực tiễn của đề tài:

Bên cạnh các ý nghĩa khoa học trên, đề tài còn có một số ý nghĩa thực tiễn nhất định trong nghiên cứu về vi mạch và trí tuệ nhân tạo. Các ý nghĩa thực tiễn cụ thể của đề tài là:

- 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 công nghệ CMOS vốn đang có xu hướng phát triển chậm do công nghệ này đã đạt tới giới hạn của sự phát triển và khó có khả năng giảm kích thước linh kiện.
- Thiết kế tối ưu của kiến trúc mảng điện trở nhớ nhị phân được đề xuất của đề tài là cơ sở cho việc xây dựng các phần cứng thực thi các ứng dụng trí tuệ nhân tạo cho các nút xử lý tại biên trong ứng dụng Internet của vạn vật.
- Đề tài là góp phần cho những nghiên cứu thực thi hệ thống trí tuệ nhân tạo trên nền tảng phần cứng, cố gắng mô hình hóa hoạt động của não bộ con người ở mức tiếp cận phần cứng.

7. Cấu trúc của Luận án

Quyển Luận án này được trình bày thành 6 chương với nội dung chính như sau:

Chương 1: TỔNG QUAN

Chương này trình bày các phân tích tổng quan về hướng nghiên cứu và các vấn đề còn tồn tại, xác định mục đích nghiên cứu, nhiệm vụ và phạm vi nghiên cứu của đề tài cũng như các phương pháp nghiên cứu của đề tài.

Chương 2: CƠ SỞ LÝ THUYẾT

Ở chương này, tác giả 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

Ở chương này, những khó khăn, thách thức khi xây dựng các hệ điện toán mô phỏng hệ thần kinh với mảng điện trở nhớ tương đã được khảo sát và trình bày cụ thể. Tiếp đến, tác giả phân tích mô hình toán học, cấu tạo và nguyên lý hoạt động của các kiến trúc mảng điện trở nhớ nhị phân khác nhau ứng dụng trong thiết kế hệ điện toán mô phỏng hệ thần kinh thực thi nhận dạng ảnh, bao gồm kiến trúc mảng điện trở nhớ bù, kiến trúc mảng điện trở nhớ đồng nhất và kiến trúc mảng điện trở nhớ đơn. Từ đó, tác giả thực thi ứng dụng nhận dạng ảnh với các kiến trúc mảng điện trở nhớ nhị phân và tiến hành phân tích sự ảnh hưởng của nhiễu và sự biến thiên điện trở đến hiệu suất hoạt động của các kiến trúc này. Những kết quả phân tích ở chương này đã đánh giá được khả năng đáp ứng với nhiễu trên dữ liệu và với sự biến thiên điện trở mảng của các kiến trúc mảng điện trở nhớ nhị phân, làm cơ sở cho việc lựa chọn một kiến trúc mảng điện trở nhớ tiềm năng nhất để tiếp tục phát triển tối ưu cho thiết kế hệ điện toán mô phỏng hệ thần kinh thực thi nhận dạng ảnh.

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

Ở chương này, tác giả trình bày về mật độ dữ liệu trong ảnh nhị phân và phân tích sự ảnh hưởng của mật độ dữ liệu đến hoạt động của các kiến trúc mảng điện trở nhị phân khi thực thi nhận dạng. Cụ thể, sự ảnh hưởng của mật độ dữ liệu đến dòng điện ngõ ra cột và đến hoạt động nhận dạng của mạch tìm dòng điện lớn nhất đã được phân tích chi tiết với kiến trúc mảng điện trở nhớ bù và kiến trúc mảng điện trở nhớ đơn. Các kết quả của chương này đã cho thấy kiến trúc mảng điện trở

nhớ đơn đã chịu sự ảnh hưởng xấu khi mật độ dữ liệu thay đổi, dẫn tới hiệu suất hoạt động của hệ thống không còn ổn định khi mật độ dữ liệu giảm xuống thấp. Những kết quả phân tích ở chương này là một phát hiện quan trọng cho thấy yêu cầu cấp bách và quan trọng của việc nghiên cứu và phát triển một kiến trúc mạng điện trở nhớ tối ưu trên cơ sở của kiến trúc mạng điện trở nhớ đơn nhằm triệt tiêu ảnh hưởng của mật độ dữ liệu và đảm bảo ổn định hoạt động nhận dạng.

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

Ở chương này, tác giả thực hiện nghiên cứu tối ưu kiến trúc mạng điện trở nhớ nhị phân để loại bỏ sự ảnh hưởng của mật độ dữ liệu khi thực thi nhận dạng. Cụ thể, thông qua phân tích phương trình toán học của mạng XNOR, tác giả đã thiết kế một mạch bù hằng số dòng điện cột để kết hợp với kiến trúc mạng điện trở nhớ đơn nhằm triệt tiêu sự ảnh hưởng của mật độ dữ liệu. Lúc này, kiến trúc mạng điện trở nhớ nhị phân sẽ thực thi phương trình toán học đầy đủ của mạng XNOR chỉ với một mạng điện trở nhớ duy nhất và không còn chịu sự ảnh hưởng của mật độ dữ liệu như ở kiến trúc trước đó. Trên cơ sở đó, một thiết kế mạch tối ưu cho kiến trúc mạng điện trở nhớ nhị phân ứng dụng trong thực thi nhận dạng 10 ảnh nhị phân với một mạng điện trở nhớ duy nhất và một mạch bù hằng số dòng điện cột đã được tác giả đề xuất và mô phỏng kiểm chứng. Các kết quả mô phỏng đã cho thấy kiến trúc mạng điện trở nhớ tối ưu đề xuất đã loại bỏ hoàn toàn sự ảnh hưởng của mật độ dữ liệu đến hoạt động nhận dạng trong khi vẫn chỉ sử dụng một mạng điện trở nhớ duy nhất để thực thi mạng XNOR. Kiến trúc mạng điện trở nhớ tối ưu đề xuất đã ổn định được dòng điện ngõ ra cột và đảm bảo được tính chính xác của hoạt động nhận dạng khi mật độ dữ liệu giảm xuống thấp, là điều mà kiến trúc mạng điện trở nhớ đơn trước đó vẫn chưa giải quyết được. Để mở rộng ứng dụng, tác giả cũng đã thiết kế một mạng nơ-ron XNOR nhiều lớp với kiến trúc mạng điện trở nhớ đã tối ưu, giải quyết bài toán nhận dạng ký tự viết tay trên tập dữ liệu MNIST.

Chương 6: KẾT LUẬN

Chương này trình bày tổng kết những kết quả đã đạt được của luận án và đóng góp mang tính cải tiến, tính mới của luận án trong nghiên cứu tối ưu và thực thi hệ điện toán mô phỏng hệ thần kinh cho bài toán nhận dạng với mảng điện trở nhớ. Đồng thời, các hướng phát triển tiếp theo cho đề tài được nêu ra, làm định hướng cho các nghiên cứu chuyên sâu hơn trong thiết kế các hệ điện toán mô phỏng hệ thần kinh thực thi các mô hình trí tuệ nhân tạo ở cấp độ phần cứng với mảng điện trở nhớ nhị phân.

CHƯƠNG 1

TỔNG QUAN

Chương này trình bày tổng quan về đề tài nghiên cứu của Luận án. Trước tiên, tác giả sẽ phân tích tổng quát về hướng nghiên cứu và các công trình nghiên cứu có liên quan, nhằm làm sáng tỏ vấn đề còn tồn tại của hướng nghiên cứu. Từ đó, tính cấp thiết và ý nghĩa của đề tài sẽ được làm rõ. Đồng thời, chương này cũng sẽ xác định rõ các mục tiêu nghiên cứu cũng như đối tượng nghiên cứu của đề tài. Cuối cùng, nhiệm vụ và phạm vi nghiên cứu, phương pháp nghiên cứu để đạt được các mục tiêu cũng sẽ được tác giả trình bày chi tiết.

1.1 TỔNG QUAN VỀ HƯỚNG NGHIÊN CỨU VÀ CÁC CÔNG TRÌNH NGHIÊN CỨU CÓ LIÊN QUAN

Hệ điện toán mô phỏng hệ thần kinh (Neuromorphic Computing System) là một khái niệm được đề xuất vào năm 1990 bởi giáo sư Carver A. Mead thuộc viện Kỹ thuật California [1]. Khái niệm hệ điện toán mô phỏng hệ thần kinh dùng để mô tả các hệ thống phần cứng có khả năng tính toán, xử lý như cách mà một bộ não sinh học thực hiện: các tín hiệu vào song song và được xử lý đồng thời. Có thể hiểu một cách đơn giản, các hệ điện toán mô phỏng hệ thần kinh là các hệ thống thực thi các mô hình trí tuệ nhân tạo trên thiết bị phần cứng, ở đây cụ thể là vi mạch bán dẫn. Kể từ khi được đề xuất, các hệ điện toán mô phỏng hệ thần kinh đã được thiết kế và thực thi trên các hệ thống vi mạch tích hợp cỡ lớn (VLSI). Các thiết kế VLSI để thực thi các mạng nơ-ron nhân tạo nhằm xây dựng các hệ điện toán mô phỏng hệ thần kinh đã đạt được nhiều kết quả nổi bật trong những thập kỷ gần đây [2-5]. Tiêu biểu có thể kể đến như trong một nghiên cứu công bố vào năm 2015, C. Mayr và các cộng sự đã đề xuất một hệ điện toán mô phỏng hệ thần kinh dạng tụ điện – đóng ngắt để thực hiện ghép nối nơ-ron

vòng kín trên công nghệ CMOS 28 nanomet có kích thước 600×600 micromet với 128 ngõ vào, 64 ngõ ra, 8192 khớp nối thần kinh [3]. Vào năm 2023, H. Jiang và các cộng sự đã thiết kế một chip đa lõi cho mạng nơ-ron xung (Spiking Neural Network) dựa trên 60K-ReRAM khớp nối thần kinh và 480 nơ-ron số trong một nút với công nghệ 180 nm, đạt mật độ $20K \text{ bit/mm}^2$ trong mỗi lõi và năng lượng tiêu thụ tối thiểu cho mỗi phép tính toán là $6,6 \text{ pJ}$ [4]. Trong một nghiên cứu khác vào năm 2025, T. Wang và các cộng sự đã thực nghiệm một chip xử lý hệ điện toán thần kinh cho mạng nơ-ron xung, có diện tích hoạt động $7,2 \text{ mm}^2$ với công nghệ chế tạo CMOS 65 nm 1P9M, cho cấu hình mạng nơ-ron xung 4 lớp 256-256-256-256-200 để thực thi nhận dạng trên tập MNIST rút gọn, đạt tốc độ xử lý 802 ảnh/giây và năng lượng tiêu thụ $27 \text{ } \mu\text{J}/\text{ảnh}$ [5].

Tuy nhiên, các công nghệ chế tạo cho các hệ thống VLSI hiện nay đều dựa trên công nghệ CMOS (Complementary Metal Oxide Semiconductor) và công nghệ CMOS đến nay 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, để thực thi các hệ điện toán mô phỏng hệ thần kinh, cần có một linh kiện thiết kế vi mạch thay thế nhằm khắc phục được các giới hạn này.

Được đề xuất vào năm 1971 bởi Giáo sư Leon Chua [9] và được chế tạo thực nghiệm thành công vào năm 2008 bởi Giáo sư R. S. Williams và các cộng sự thuộc phòng nghiên cứu HP [10], điện trở nhớ đã được xem như là linh kiện điện tử thứ tư (bên cạnh tụ điện, điện trở và cuộn dây) và thu hút được nhiều sự quan tâm của các nhà nghiên cứu trên thế giới. Với khả năng lưu lại giá trị điện trở hiện tại cũng như khả năng thay đổi giá trị điện trở bằng điện áp đặt vào 2 cực của rất giống với đặc tính của khớp nối thần kinh trong hệ thần kinh sinh học [11], điện trở nhớ đã được xem như là một linh kiện tiềm năng để thực hiện mô hình hóa cho các khớp nối thần kinh [11, 12]. Với việc thực hiện các phép nhân và các phép cộng tích lũy theo định luật Kirchhoff và định luật Ohm ở cấp độ phân cứng, điện trở nhớ đã tạo ra một sự phát triển vượt bậc trong tốc độ thực thi, công suất tiêu thụ cũng như diện tích của các thiết kế phân cứng [13]. Đồng thời, các điện trở nhớ có thể được kết nối với nhau dưới dạng mảng mà tại mỗi điểm

giao nhau giữa một hàng và một cột sẽ có một điện trở nhớ đóng vai trò như một thiết bị chuyển mạch [14-21]. Tiêu biểu như ở một nghiên cứu được công bố vào năm 2024, các tác giả đã chế tạo thành công một mảng 32×32 điện trở nhớ bằng vật liệu $\text{Ru/HfSiO}_y/\text{Al}_2\text{O}_3/\text{HfSiO}_3/\text{TiN}$ [16]. Trong một nghiên cứu khác vào năm 2024, J. Chen và các cộng sự đã thực nghiệm thành công mảng điện trở nhớ 3×3 bằng cấu trúc vật liệu Ag/ZnO/Ag với công nghệ in hoàn toàn, cho phép sản xuất giá rẻ và quy mô lớn, có thể chuyển trạng thái với dòng điện nhỏ $1 \mu\text{A}$ [20]. Ở một nghiên cứu khác vào năm 2025, S. Jain Smarth và nhóm nghiên cứu của mình đã nghiên cứu chế tạo thành công mảng điện trở nhớ 32×32 theo cấu trúc 1 bộ lựa chọn và 1 memristor (1S1M) với vật liệu Si và HfSe_2 để khắc phục vấn đề dòng rò trong mảng, cho các mạng nơ-ron nhân chập [21]. Hơn thế nữa, các mảng điện trở nhớ cũng có thể được thiết kế theo kiến trúc 3-D để hiện thực hóa khả năng kết nối cao của hệ thống nơ-ron trong các bộ não sinh học [22-26]. Cụ thể, trong một nghiên cứu, nhóm tác giả đã thực thi thành công một mảng 3D các điện trở nhớ bao gồm 3 lớp với vật liệu HfAlO_x [25]. Hoặc trong một nghiên cứu khác vào năm 2025, Y. S. An và các cộng sự đã thiết kế và chế tạo một mảng 3D nguyên khối có dung lượng 1 kbit theo cấu trúc 1 memristor (1M) bằng vật liệu TiO_x và 1 transistor lựa chọn (1 TS), cho khả năng tích hợp để mở rộng mảng lên tới 1,14 Tbit [26]. Các đặc điểm này đã làm cho mảng điện trở nhớ trở thành một giải pháp tiềm năng để thực thi các hệ điện toán mô phỏng hệ thần kinh và thay thế cho các thiết kế VLSI truyền thống vốn đã đạt tới giới hạn.

Trong thời gian gần đây, để xây dựng các hệ điện toán mô phỏng hệ thần kinh, các mảng điện trở nhớ tương tự đã được sử dụng để thực thi các mạng nơ-ron nhân tạo [27-33] như là một giải pháp thiết kế đạt hiệu quả cao và công suất tiêu thụ thấp. Khi thực thi các mạng nơ-ron nhân tạo với mảng điện trở nhớ tương tự, để thực hiện một khớp nối thần kinh với trọng số có dấu (dấu + hoặc dấu -), 2 điện trở nhớ tương tự đã được sử dụng [27, 28, 30, 33-35]. Độ lớn của 2 điện trở nhớ này được điều chỉnh để thu được giá trị và dấu của trọng số tương ứng của một khớp nối thần kinh. Các thiết kế mạng nơ-ron nhân tạo trên các mảng điện trở nhớ tương tự đã đạt được một số kết quả nhất định và có thể xem

như là một giải pháp tiềm năng để xây dựng các hệ điện toán mô phỏng hệ thần kinh nhằm thay thế cho các thiết kế VLSI vốn đã đạt tới giới hạn về kích thước.

Tuy được xem là một giải pháp tiềm năng nhưng việc xây dựng các hệ điện toán mô phỏng hệ thần kinh bằng các thiết kế dùng các mảng điện trở nhớ tương tự để thực thi các mạng nơ-ron nhân tạo vẫn còn nhiều khó khăn và thách thức phải đối mặt. Trước hết, có thể thấy, 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 để đạt đến giá trị mong muốn một cách chính xác [31, 36]. Thêm vào đó, việc 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 khá khó khăn và phức tạp [32, 33, 37] và là một cản trở lớn cho việc thực thi các mạng nơ-ron nhân tạo trên các mảng điện trở nhớ tương tự. Đồng thời, sự biến thiên giá trị điện trở của các điện trở nhớ trong mảng gây ra bởi sự không đồng nhất trong quá trình chế tạo [38-41] hoặc sự trôi của giá trị điện trở [42, 43] trong quá trình hoạt động cũng là một khó khăn không nhỏ cho việc thực thi các mạng nơ-ron nhân tạo trên các mảng điện trở nhớ tương tự. Sự biến thiên giá trị điện trở trong mảng sẽ tạo ra sự sai lệch giữa giá trị thực tế khi hoạt động và giá trị mong muốn trong tính toán của các điện trở nhớ, làm giảm đáng kể hiệu suất hoạt động của các ứng dụng sử dụng mảng điện trở nhớ tương tự [44, 45].

Như vậy, việc xây dựng các hệ điện toán mô phỏng hệ thần kinh bằng cách thực thi các mạng nơ-ron nhân tạo sử dụng các mảng điện trở nhớ tương tự vẫn là một hướng tiếp cận khá khó khăn và còn nhiều thách thức. Thay vì sử dụng điện trở nhớ với các giá trị tương tự, việc sử dụng điện trở nhớ với 2 mức giá trị (điện trở cao và điện trở thấp) mang lại hiệu quả cao hơn, đặc biệt khi thực thi các mạng nơ-ron nhân tạo có tập trọng số nhị phân [24, 44, 46-48]. Với yêu cầu chỉ có 2 mức giá trị, điện trở cao (High Resistance State – HRS) và điện trở thấp (Low Resistance State – LRS), các điện trở nhớ này còn được gọi là các điện trở nhớ nhị phân. Sự biến thiên giá trị do quá trình chế tạo và hoạt động của các điện trở nhớ nhị phân ít ảnh hưởng đến hiệu suất hoạt động của hệ thống hơn khi so với điện trở nhớ tương tự [44] nhờ tỷ lệ HRS/LRS khá cao [25, 49, 50]. Ví dụ

như trong một nghiên cứu vào năm 2021, nhóm tác giả đã thực thi một mảng các điện trở nhớ nhị phân trong đó mỗi điện trở có thể hoạt động được ở một trạng thái điện trở cao (HRS) và bốn trạng thái điện trở thấp (LRS) với tỉ lệ HRS/LRS đạt tối đa là 100 [25]. Ở một nghiên cứu vào năm 2025, S. He và các cộng sự đã chế tạo thành công một mảng điện trở nhớ SRM (self-rectifying memristor) kích thước 16×16 , dựa trên vật liệu $\text{Pt/HfO}_2/\text{Ta}_2\text{O}_{5-x}/\text{Ti}$, cho tỉ lệ HRS/LRS có thể đạt từ 10^2 đến 10^3 [50]. Trong một nghiên cứu khác, nhóm tác giả đã chế tạo một mảng các điện trở nhớ nhị phân có thể đạt được một mức trạng thái điện trở cao ở giá trị hơn $70 \text{ M}\Omega$ và ba mức trạng thái điện trở thấp, thấp nhất khoảng $2 \text{ K}\Omega$, đạt tỉ lệ HRS/LRS lớn hơn 10000 [49]. Phép nhân giữa các tín hiệu ngõ vào với các trọng số nhị phân cũng sẽ dễ dàng được thực thi hơn bằng các phép toán logic như AND, XNOR theo định luật Kirchhoff và định luật Ohm [13, 51]. Do vậy, có thể thấy các mảng điện trở nhớ nhị phân đã mở ra một hướng tiếp cận mới, nhiều tiềm năng và khả thi hơn cho việc xây dựng các hệ điện toán mô phỏng hệ thần kinh, đặc biệt là ở các bài toán nhận dạng mẫu như nhận dạng tiếng nói [44, 52], nhận dạng ảnh [47, 51, 53-55].

Để xây dựng các hệ điện toán mô phỏng hệ thần kinh nhằm 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]. Các kiến trúc này đã được ứng dụng và chứng minh tính khả thi của việc hiện thực hóa các hệ điện toán mô phỏng hệ thần kinh trên nền tảng phần cứng mảng điện trở nhớ nhị phân. Cụ thể, Trong một nghiên cứu, F. L. Aguirre và các cộng sự đã đề xuất và hiện thực một mạng perceptron một lớp với kích thước $n^2 \times 10$ nhằm thực hiện bài toán nhận dạng các ảnh có đặc trưng tương tự bộ dữ liệu MNIST [54]. Trong phương án thiết kế, nhóm tác giả đã áp dụng kiến trúc mảng điện trở nhớ bù, trong đó các linh kiện điện trở nhớ được mô hình hóa theo dạng memdiode, để lưu trữ các trọng số của mạng. Theo cách tiếp cận này, nhằm biểu diễn đầy đủ cả các giá trị trọng số dương và âm, mỗi trọng số được hiện thực bằng hai phần tử điện trở nhớ. Do đó, để lưu trữ một ma trận trọng số

có kích thước $n^2 \times 10$, hệ thống yêu cầu hai mảng điện trở nhớ bù nhau, mỗi mảng có cùng kích thước $n^2 \times 10$. Trong một nghiên cứu khác được công bố năm 2023, C. Y. Han và các cộng sự đã đề xuất và hiện thực một mạng nơ-ron một lớp nhằm thực hiện bài toán nhận dạng ký tự viết tay MNIST, dựa trên kiến trúc mảng điện trở nhớ bù sử dụng linh kiện điện trở nhớ chế tạo từ vật liệu NbOx [55]. Trong thiết kế này, hai mảng điện trở nhớ, $M +$ và $M -$, được sử dụng để lưu trữ các trọng số đã được huấn luyện của mạng nơ-ron, qua đó cho phép biểu diễn các giá trị trọng số dương và âm của ma trận trọng số. Ở một nghiên cứu khác được công bố năm 2024, các tác giả đã ứng dụng một mảng điện trở nhớ để thực thi một mạng nơ-ron thần kinh xung hai lớp, phục vụ bài toán nhận dạng âm thanh [52]. Cụ thể, hệ thống được xây dựng bằng cách kết hợp một mảng điện trở nhớ với một lớp kết nối đầy đủ nhằm thực hiện bài toán nhận dạng. Tuy nhiên, trong nghiên cứu này, mỗi khớp nối thần kinh trong mảng điện trở nhớ được triển khai theo cấu trúc 1T1M (một transistor – một điện trở nhớ), thay vì cấu trúc 1M đơn giản. Việc sử dụng cấu trúc 1T1M thay cho 1M tại mỗi khớp nối sẽ dẫn đến sự gia tăng đáng kể về diện tích phân cứng cũng như độ phức tạp của mảng điện trở nhớ, từ đó ảnh hưởng đến khả năng mở rộng và hiệu quả tích hợp của hệ thống.

Trong các kiến trúc mảng điện trở nhớ nhị phân trên, kiến trúc mảng điện trở nhớ đơn chỉ sử dụng 1 mảng điện trở nhớ nhị phân duy nhất để thực hiện tác vụ nhận dạng, trong khi kiến trúc mảng điện trở nhớ bù và kiến trúc mảng điện trở nhớ đồng nhất yêu cầu 2 mảng điện trở nhớ. Do đó, kiến trúc mảng điện trở nhớ đơn thể hiện ưu thế về kích thước mảng cũng như năng lượng tiêu thụ trên mảng so với hai kiến trúc còn lại.

Ở trong nước, việc nghiên cứu và phát triển các ứng dụng với điện trở nhớ vẫn chưa được phát triển mạnh và vẫn còn ít công trình được công bố. Vào năm 2015, một ứng dụng của mảng điện trở nhớ để thực thi nhận dạng ký tự đã được công bố [56]. Trong nghiên cứu này, thay vì thực hiện phép nhân dựa vào định luật Ohm, nhóm tác giả đã đề xuất phương pháp sử dụng bộ nhân nhị phân. Phương pháp này có khả năng ứng dụng thực tế cao hơn các thiết kế trước, do có

thể làm giảm sai số của quá trình thực hiện phép nhân tương tự sử dụng định luật Ohm ở các thiết kế trước đó [56]. Trong một nghiên cứu khác, nhóm tác giả đã đề xuất một kiến trúc mảng điện trở nhớ cho phép thực thi một mạng nơ-ron nhân tạo, trong đó các điện trở nhớ đóng vai trò là các khớp nối thần kinh nhân tạo [57]. Mảng điện trở nhớ sau đó được huấn luyện để các giá trị trọng số thay đổi theo tập ảnh các ký tự. Sau khi huấn luyện, mảng điện trở nhớ có khả năng nhận dạng các ký tự [57]. Tuy nhiên, các kiến trúc này được thiết kế dựa trên 2 mảng điện trở nhớ. Việc sử dụng 2 mảng điện trở nhớ để thực thi mạng nơ-ron nhân tạo sẽ không tối ưu về công suất tiêu thụ trên mảng và kích thước mảng khi so sánh với các kiến trúc dùng 1 mảng điện trở nhớ. Thêm vào đó, việc chế tạo và lập trình 2 mảng điện trở nhớ như các cấu trúc đề xuất này hầu như ít khả thi đối với các điện trở nhớ được chế tạo trong môi trường phòng thí nghiệm hiện nay. Vào năm 2020, kiến trúc mảng điện trở nhớ đơn, trong đó chỉ sử dụng 1 mảng điện trở nhớ nhị phân, đã được đề xuất để thực thi ứng dụng nhận dạng ảnh [48]. Kiến trúc mảng điện trở nhớ đơn này đã thể hiện được sự tối ưu hơn trong công suất tiêu thụ trên mảng cũng như kích thước của mảng so với các kiến trúc sử dụng 2 mảng điện trở nhớ [48]. Những năm gần đây, trong bối cảnh nghiên cứu về các linh kiện điện trở nhớ phục vụ thực thi các hệ điện toán mô phỏng hệ thần kinh, một số nỗ lực nghiên cứu đã tập trung vào việc phát triển thiết bị điện trở nhớ có khả năng mô phỏng hành vi của các khớp nối thần kinh, tương tự như trong hệ thần kinh sinh học. Tiêu biểu, trong một nghiên cứu công bố năm 2024, các tác giả của nhóm nghiên cứu thuộc Đại học Quốc gia Tp.HCM và Đại học Quốc gia Hà Nội đã chế tạo thành công linh kiện điện trở nhớ tương tự dựa trên các hạt nano ZnO tổng hợp bằng phương pháp xanh [58]. Các linh kiện được xây dựng theo cấu trúc Cr/ZnO/FTO cho thấy đặc tính điện trở chuyển mạch tương tự ổn định, với tỷ số ON/OFF cao, và khả năng tự chỉnh lưu nhờ tiếp giáp Schottky. Khi được kích thích bằng các chuỗi xung điện áp, linh kiện cho thấy khả năng mô phỏng hành vi của khớp nối thần kinh, như tăng cường và suy giảm dẫn điện, cho thấy tiềm năng ứng dụng của linh kiện điện trở nhớ trên vật liệu ZnO trong các hệ điện toán mô phỏng hệ thần kinh. Trong một công trình khác được công

bổ năm 2025, P.-Q. Pham và các cộng sự trong và ngoài nước đã chế tạo và khảo sát thực nghiệm thành công một mảng điện trở nhớ kích thước 16×16 không cần bộ lựa chọn (selector-free), dựa trên hệ vật liệu CrOx/TiO_2 [59]. Các linh kiện điện trở nhớ được chế tạo cho thấy các đặc tính vận hành ổn định, độ đồng nhất cao giữa các phần tử, đồng thời thể hiện khả năng giảm dòng rò hiệu quả (ở mức xấp xỉ 150 nA) – những yếu tố đóng vai trò then chốt trong việc triển khai các mảng điện trở nhớ kích thước lớn. Các kết quả nghiên cứu này đã cho thấy tiềm năng rất lớn trong nước trong khả năng hiện thực hóa các hệ điện toán mô phỏng hệ thần kinh dựa trên mảng điện trở nhớ, đặc biệt đối với các kiến trúc có khớp nối thần kinh với cấu trúc một điện trở nhớ (1M).

Tuy các kiến trúc mảng điện trở nhớ nhị phân, đặc biệt là kiến trúc mảng điện trở nhớ đơn, là rất khả thi cho các thiết kế hệ điện toán mô phỏng hệ thần kinh trên phần cứng, nhưng ở các kiến trúc mảng điện trở nhớ này, ảnh hưởng của các yếu tố như độ biến thiên điện trở của mảng, nhiễu trên dữ liệu đầu vào vẫn chưa được phân tích và đánh giá cụ thể. Hơn nữa, ảnh hưởng của sự thay đổi của mật độ dữ liệu (mật độ các bit 1) trong mảng điện trở nhớ tới hiệu suất hoạt động của hệ thống vẫn chưa được phân tích, đánh giá và có phương án tối ưu để loại bỏ. Do đó, việc phân tích các yếu tố ảnh hưởng và tối ưu hóa kiến trúc mảng điện trở nhớ nhị phân là một yêu cầu quan trọng và cấp thiết. Điều này sẽ đảm bảo hiệu suất nhận dạng của các ứng dụng nhận dạng với kiến trúc mảng điện trở nhớ nhị phân được ổn định hơn, từ đó làm tăng khả năng hiện thực hóa của các hệ điện toán mô phỏng hệ thần kinh khi thiết kế với kiến trúc mảng điện trở nhớ nhị phân.

1.2 TÍNH CẤP THIẾT VÀ Ý NGHĨA CỦA ĐỀ TÀI

Từ các phân tích trên, có thể thấy việc nghiên cứu và phát triển các thiết kế sử dụng một mảng điện trở nhớ nhị phân để xây dựng các hệ điện toán mô phỏng hệ thần kinh thực thi nhận dạng ảnh là một xu hướng triển vọng và cấp thiết để tạo ra các thiết kế mới thay thế cho các hệ điện toán mô phỏng hệ thần kinh trên các thiết kế VLSI truyền thống dựa trên công nghệ CMOS vốn đã đạt tới giới hạn về sự phát triển và khả năng giảm kích thước. Trong đề tài này, thông qua quá

trình phân tích, đánh giá cụ thể các yếu tố ảnh hưởng đến hiệu suất hoạt động của các kiến trúc mảng điện trở nhớ như: độ biến thiên điện trở, nhiễu trên dữ liệu, mật độ dữ liệu, tác giả sẽ nghiên cứu đề xuất 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. 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 khắc phục được ảnh hưởng của mật độ dữ liệu đến hoạt động nhận dạng. So với kiến trúc mảng điện trở nhớ bù ở [44], [54], [55] và kiến trúc mảng điện trở nhớ đồng nhất ở [47], kiến trúc được đề xuất trong Luận án cho thấy mức độ tối ưu cao hơn ở cấu trúc khớp nối thần kinh, khi sử dụng một điện trở nhớ (1M) để thay thế cho cấu trúc hai điện trở nhớ (2M). Cách tiếp cận này cho phép giảm quy mô phần cứng của mảng điện trở nhớ, đồng thời nâng cao hiệu quả sử dụng công suất, so với hai kiến trúc còn lại. Bên cạnh đó, việc triển khai khớp nối thần kinh theo cấu trúc một điện trở nhớ còn cải thiện được khả năng đáp ứng của hệ thống trước nhiễu dữ liệu và sự biến thiên giá trị điện trở, từ đó nâng cao tính ổn định và độ tin cậy của kiến trúc đề xuất so với các kiến trúc mảng điện trở nhớ bù và đồng nhất. So với kiến trúc mảng điện trở nhớ đơn nguyên bản ở [48], kiến trúc được đề xuất trong Luận án khắc phục được ảnh hưởng của sự thay đổi mật độ dữ liệu, qua đó duy trì hiệu suất nhận dạng ổn định trong các điều kiện phân bố mật độ dữ liệu khác nhau. Đây là một hạn chế vẫn còn tồn tại trong nghiên cứu trước đối với kiến trúc mảng điện trở nhớ đơn nguyên bản. Đồng thời, để mở rộng ứng dụng của kiến trúc mảng điện trở nhớ được đề xuất, đề tài cũng sẽ thiết kế một mạng XNOR nhiều lớp ứng dụng kiến trúc mảng điện trở nhớ đã tối ưu cho bài toán nhận dạng ký tự viết tay trên tập dữ liệu MNIST.

Đề tài được thực hiện sẽ tạo ra một số ý nghĩa khoa học và thực tiễn nhất định trong nghiên cứu về vi mạch và trí tuệ nhân tạo. Về mặt khoa học, hiện nay, trí tuệ nhân tạo đang được phát triển, kỳ vọng được áp dụng rộng rãi và mang lại hiệu quả trong nhiều lĩnh vực. Tuy nhiên, các mô hình trí tuệ nhân tạo hiện đại được thực thi dưới dạng các phần mềm chạy trên các hệ thống máy chủ đủ mạnh,

đây là hướng tiếp cận thứ nhất. Một hướng tiếp cận thứ hai là tạo ra phần cứng có khả năng xử lý song song và mô hình hóa được hoạt động của não bộ con người, các hệ thống này được gọi là hệ điện toán mô phỏng theo hệ thần kinh (Neuromorphic Computing System). Đây là một xu hướng đang được các công ty vi mạch bán dẫn hàng đầu thế giới đang đầu tư phát triển [60, 61]. Đề tài này thực hiện theo hướng tiếp cận thứ hai nhằm tạo ra phần cứng có khả năng mô phỏng hoạt động của hệ thần kinh của con người. Đề tài nghiên cứu ứng dụng điện trở nhớ, một linh kiện điện tử tiềm năng đã được nghiên cứu và chế tạo thành công vào năm 2008, để thực thi một hệ điện toán mô phỏng theo hệ thần kinh. Cụ thể, đề tài đã nghiên cứu 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. Kiến trúc mạng điện trở tối ưu được ứng dụng để thiết kế một mạng nơ-ron XNOR nhiều lớp, mở ra hướng nghiên cứu tiềm năng cho các thiết kế tối ưu hệ thống mô phỏng hệ thần kinh sử dụng linh kiện điện trở nhớ. Về mặt thực tiễn, đề tài sẽ là tiền đề cho những nghiên cứu chuyên sâu hơn về các hệ điện toán mô phỏng hệ thần kinh ứng dụng mảng điện trở nhớ thay cho công nghệ CMOS vốn đang có xu hướng phát triển chậm do khó có khả năng giảm kích thước linh kiện. Thiết kế tối ưu đề xuất trong đề tài sẽ là cơ sở cho việc xây dựng các phần cứng thực thi các ứng dụng trí tuệ nhân tạo cho các nút xử lý tại biên trong ứng dụng Internet của vạn vật. Đồng thời, đề tài cũng góp phần cho những nghiên cứu thực thi hệ thống trí tuệ nhân tạo trên nền tảng phần cứng, 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 bao gồm nhiều 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 khi thực thi nhận dạng ảnh, làm cơ sở cho việc phát triển một kiến trúc mảng điện trở nhớ

tối ưu ứng dụng trong thiết kế hệ điện toán mô phỏng hệ thần kinh thực thi nhận dạng ảnh.

- Phân tích và đánh giá sự ảnh hưởng của mật độ dữ liệu đến hoạt động của các kiến trúc mạng điện trở nhớ nhị phân khi thực thi mạng XNOR nhận dạng ảnh và tìm ra nguyên nhân gây ra sự ảnh hưởng. Từ đó, xác định được yêu cầu cần tối ưu của mạng điện trở nhớ nhị phân khi thực thi mạng XNOR trong các bài toán nhận dạng.
- Nghiên cứu và đề xuất một thiết kế mạch tối ưu cho kiến trúc mạng điện trở nhớ nhị phân ứng dụng trong thiết kế hệ điện toán mô phỏng hệ thần kinh thực thi bài toán nhận dạng ảnh. Kiến trúc đề xuất sẽ thực thi được phương trình toán học đầy đủ của mạng XNOR chỉ với một mạng 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 và đề xuất một thiết kế mạng nơ-ron XNOR ứng dụng kiến trúc mạng 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

Với các mục tiêu nghiên cứu trên, đề tài có các đối tượng nghiên cứu cụ thể sau:

- Hệ điện toán mô phỏng hệ thần kinh.
- Điện trở nhớ.
- Mạng điện trở nhớ.
- Vi mạch ứng dụng mạng điện trở nhớ để thực thi các hệ điện toán mô phỏng hệ thần kinh.
- Các kiến trúc mạng điện trở nhớ nhị phân khác nhau ứng dụng trong thiết kế các hệ điện toán mô phỏng hệ thần kinh thực thi các bài toán nhận dạng ảnh.
- Các yếu tố ảnh hưởng đến hiệu suất hoạt động và giải pháp thiết kế tối ưu cho kiến trúc mạng điện trở nhớ nhị phân trong thiết kế hệ điện toán mô phỏng hệ thần kinh thực thi ứng dụng nhận dạng ảnh.

1.5 NHIỆM VỤ VÀ PHẠM VI NGHIÊN CỨU

Để thực hiện được các mục tiêu và đạt được mục đích đã đặt ra, đề tài sẽ có những nhiệm vụ nghiên cứu chi tiết như sau:

- Tìm hiểu về linh kiện điện trở nhớ: cấu tạo, nguyên lý hoạt động và mô phỏng khảo sát các đặc tính của linh kiện này.
- Tìm hiểu các cơ sở lý thuyết nền tảng để thực hiện đề tài, gồm: khái niệm hệ điện toán mô phỏng hệ thần kinh, khảo sát các hướng khác nhau để thực thi hệ điện toán mô phỏng hệ thần kinh với điện trở nhớ.
- Phân tích những thách thức khi ứng dụng các mảng điện trở nhớ tương tự trong xây dựng các hệ điện toán mô phỏng hệ thần kinh thực thi các mạng nơ-ron nhân tạo.
- Phân tích mô hình toán học và cấu tạo các kiến trúc mảng điện trở nhớ nhị phân khác nhau được ứng dụng trong xây dựng hệ điện toán mô phỏng hệ thần kinh thực thi ứng dụng nhận dạng ảnh.
- Phân tích và đánh giá sự ảnh hưởng của các yếu tố: nhiễu trên dữ liệu vào, độ biến thiên điện trở đến hiệu suất hoạt động của các kiến trúc mảng điện trở nhớ nhị phân trong thực thi ứng dụng nhận dạng ảnh.
- Phân tích, đánh giá sự ảnh hưởng và tìm ra nguyên nhân ảnh hưởng của mật độ dữ liệu đến hiệu suất hoạt động của các kiến trúc mảng điện trở nhớ nhị phân nhằm xác định được vấn đề còn tồn tại cần được tối ưu của kiến trúc mảng điện trở nhớ nhị phân khi thực thi mảng XNOR trong nhận dạng ảnh.
- Tính toán và đề xuất 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 thực thi ứng dụng nhận dạng ảnh bằng cách thực thi phương trình toán học đầy đủ của mảng XNOR với một mảng điện trở nhớ, loại bỏ được sự ảnh hưởng của mật độ dữ liệu nhằm đảm bảo hiệu suất hoạt động của hệ thống được ổn định.

- Nghiên cứu, đề xuất một thiết kế mạng nơ-ron XNOR nhiều lớp ứng dụng kiến trúc mảng điện trở nhớ tối ưu cho bài toán nhận dạng ký tự viết tay của tập dữ liệu MNIST.
- Thực hiện mô phỏng để kiểm chứng, phân tích và đánh giá kết quả thực thi nhận dạng ảnh với kiến trúc mảng điện trở nhớ tối ưu đã đề xuất.

Với các nhiệm vụ đã đặt ra, đề tài có phạm vi nghiên cứu cụ thể là:

- Sử dụng mô hình mô phỏng của điện trở nhớ để khảo sát đặc tính của linh kiện này.
- Không tiếp cận theo hướng xây dựng hệ điện toán mô phỏng hệ thần kinh sử dụng các kiến trúc mảng điện trở nhớ tương tự; thay vào đó, đề tài tập trung vào hướng sử dụng kiến trúc mảng điện trở nhớ nhị phân: phân tích và đề xuất thiết kế tối ưu với kiến trúc mảng điện trở nhớ nhị phân để xây dựng hệ điện toán mô phỏng hệ thần kinh thực thi ứng dụng nhận dạng ảnh.
- Đề tài phân tích ảnh hưởng của các yếu tố đến hiệu suất hoạt động của các kiến trúc mảng điện trở nhớ nhị phân, bao gồm: độ biến thiên điện trở, nhiều đầu vào, mật độ dữ liệu; xác định nguyên nhân gây nên sự ảnh hưởng; tính toán, thiết kế và đề xuất giải pháp tối ưu nhằm khắc phục và đảm bảo hiệu quả hoạt động ổn định của hệ thống.
- Việc kiểm chứng, phân tích và tổng hợp các kết quả của các thiết kế mạch đã đề xuất sẽ được thực hiện thông qua sử dụng công cụ mô phỏng: Matlab, Cadence.

1.6 PHƯƠNG PHÁP NGHIÊN CỨU

Để thực hiện đề tài này, tác giả đã sử dụng các phương pháp nghiên cứu sau:

- Phương pháp khảo sát lý thuyết: tìm hiểu thông tin, tổng hợp và hệ thống hóa các cơ sở lý thuyết liên quan tới đối tượng nghiên cứu của đề tài, bao gồm: hệ điện toán mô phỏng hệ thần kinh, điện trở nhớ, mảng điện trở

nhớ, mảng điện trở nhớ nhị phân, thực thi nhận dạng mẫu với mảng điện trở nhớ nhị phân.

- Phương pháp phân tích, tổng hợp lý thuyết: phân tích các công trình nghiên cứu đã công bố trong lĩnh vực thực thi hệ điện toán mô phỏng hệ thần kinh với các mảng điện trở nhớ; tổng hợp và tìm ra những vấn đề còn tồn tại của các kiến trúc mảng điện trở nhớ nhị phân khi thực thi hệ điện toán mô phỏng hệ thần kinh với ứng dụng nhận dạng mẫu.
- Phương pháp tính toán, thiết kế: từ các lý thuyết nền tảng đã khảo sát cũng như vấn đề còn tồn tại của các kiến trúc mảng điện trở nhớ nhị phân đã phát hiện được, tác giả sẽ phân tích tìm nguyên nhân, tính toán các thông số và đề xuất thiết kế mạch tối ưu cho kiến trúc mảng điện trở nhớ nhị phân để thực thi phương trình toán học đầy của mảng XNOR nhằm khắc phục các vấn đề còn tồn tại, đảm bảo hiệu suất hoạt động ổn định cho hệ điện toán mô phỏng hệ thần kinh thực thi ứng dụng nhận dạng ảnh.
- Phương pháp mô phỏng: sử dụng các công cụ mô phỏng: Matlab, Cadance để mô phỏng kiểm chứng hoạt động của thiết kế đã đề xuất, phân tích kết quả và thực hiện hiệu chỉnh để tối ưu cho thiết kế mạch đã đề xuất.
- Phương pháp phân tích, tổng hợp kết quả: phân tích các kết quả đã đạt được của đề tài: các ưu, nhược điểm của thiết kế đã đề xuất; từ đó tổng kết các mục tiêu đạt được của luận án, các đóng góp mang tính mới, tính cải tiến cũng như nêu ra các hướng phát triển chuyên sâu hơn dựa trên kết quả của luận án.

CHƯƠNG 2

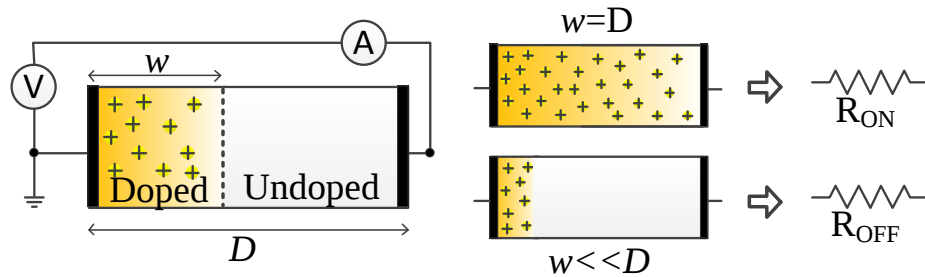
CƠ SỞ LÝ THUYẾT

Chương này trình bày các lý thuyết nền tảng có liên quan tới nội dung của luận án. Các cơ sở lý thuyết được trình bày bao gồm: giới thiệu về cấu tạo, đặc tính, các mô hình và phân loại điện trở nhớ; hệ điện toán mô phỏng hệ thần kinh, các hướng thực thi hệ điện toán mô phỏng hệ thần kinh với mảng điện trở nhớ tương tự và mảng điện trở nhớ nhị phân. Đồng thời, chương này cũng giới thiệu về phương trình toán, mô hình, đặc tính của các kiến trúc mảng điện trở nhớ nhị phân được ứng dụng trong thực thi hệ điện toán mô phỏng hệ thần kinh, bao gồm: kiến trúc mảng điện trở nhớ bù, kiến trúc mảng điện trở nhớ đồng nhất và kiến trúc mảng điện trở nhớ đơn.

Sự ứng dụng trong thực thi các hệ điện toán mô phỏng hệ thần kinh ở các bài toán nhận dạng ảnh, cũng như đặc điểm về công suất tiêu thụ và tính đáp ứng với mật độ dữ liệu, của các kiến trúc mảng điện trở nhớ nhị phân được đề cập trong chương này đã được công bố trong 03 bài báo Hội nghị quốc tế của tác giả. Cụ thể, các công trình đã công bố của Luận án liên quan tới nội dung của chương này bao gồm: (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*

2.1 GIỚI THIỆU VỀ ĐIỆN TRỞ NHỚ

Điện trở nhớ (Memristor) được tìm ra trên cơ sở lý thuyết về mối quan hệ giữa từ thông và điện tích bởi Giáo sư Leon Chua vào năm 1971 [9]. Kể từ đó, điện trở nhớ đã được xem như là phần tử mạch điện thứ 4, bên cạnh 3 phần tử cơ bản cấu thành nên các mạch điện tử là điện trở, tụ điện và cuộn dây. Tuy nhiên, một thời gian dài sau khi được đề xuất, điện trở nhớ cũng chỉ mới tồn tại trên lý thuyết vì chưa tìm ra hợp chất phù hợp để chế tạo. Vào năm 2008, ở phòng nghiên cứu HP, Giáo sư R. S. Williams và các cộng sự đã chế tạo và thử nghiệm được điện trở nhớ đầu tiên dựa trên hợp chất ô-xit titan (TiO_2) [10]. Điện trở nhớ này bao gồm 2 lớp màng mỏng được đặt giữa 2 đầu tiếp giáp bằng kim loại như minh họa ở Hình 2.1 [10].



Hình 2.1: Mô hình một điện trở nhớ với 2 lớp màng mỏng [10].

Trong Hình 2.1, điện trở nhớ có cấu tạo gồm 2 lớp màng mỏng có tổng độ dày là D , trong đó có một lớp giàu hạt dẫn có điện trở nhỏ và một lớp nghèo hạt dẫn có điện trở lớn. Vị trí tiếp giáp giữa 2 lớp có thể thay đổi được và độ dày của lớp giàu hạt dẫn được ký hiệu là w . Khi điện áp dương được đặt vào 2 bản cực, các hạt dẫn ở lớp giàu hạt dẫn bị khuếch tán và di chuyển về bên phía lớp nghèo hạt dẫn. Lúc này có thể xem như giá trị w tiến gần về giá trị D và điện trở của điện trở nhớ trở nên rất nhỏ (R_{ON}). Ngược lại khi điện áp âm được đặt vào 2 bản cực, các hạt dẫn tập trung về phía lớp giàu hạt dẫn làm cho điện trở nhớ trở nên kém dẫn điện; lúc này điện trở nhớ có giá trị điện trở rất lớn (R_{OFF}). Như vậy, giá trị điện trở của một điện trở nhớ có thể thay đổi dựa vào điện áp đặt vào 2 cực

của nó. Tính chất này giúp cho điện trở nhớ khác với các linh kiện thụ động khác và trở nên có nhiều ứng dụng. Giống như linh kiện điện trở, quan hệ giữa dòng và áp tại thời điểm t được thể hiện bởi công thức sau [10]:

$$v(t) = R_X(t)i(t) = \left(R_{ON} \frac{w(t)}{D} + R_{OFF} \left(1 - \frac{w(t)}{D} \right) \right) i(t) \quad (2.1)$$

Phương trình (2.1) về cơ bản trình bày mối quan hệ giữa dòng điện và điện áp nhưng điện trở trong phương trình (2.1) là một hàm phụ thuộc vào biến trạng thái w . Biến trạng thái w là độ dài của lớp giàu hạt dẫn như được mô tả trong Hình 2.1. Có thể xem w là vị trí tiếp giáp giữa 2 lớp: lớp giàu hạt dẫn và lớp nghèo hạt dẫn. Khi w đạt đến giá trị D , vùng giàu hạt dẫn sẽ chiếm toàn bộ không gian trong điện trở nhớ. Như đã giới thiệu ở trên, giá trị w phụ thuộc vào điện áp đặt vào 2 đầu vi điện trở nhớ, điều này đồng nghĩa với w phụ thuộc vào dòng điện chạy qua nó. Sự thay đổi của w theo cường độ dòng điện được minh họa bằng phương trình sau [10]:

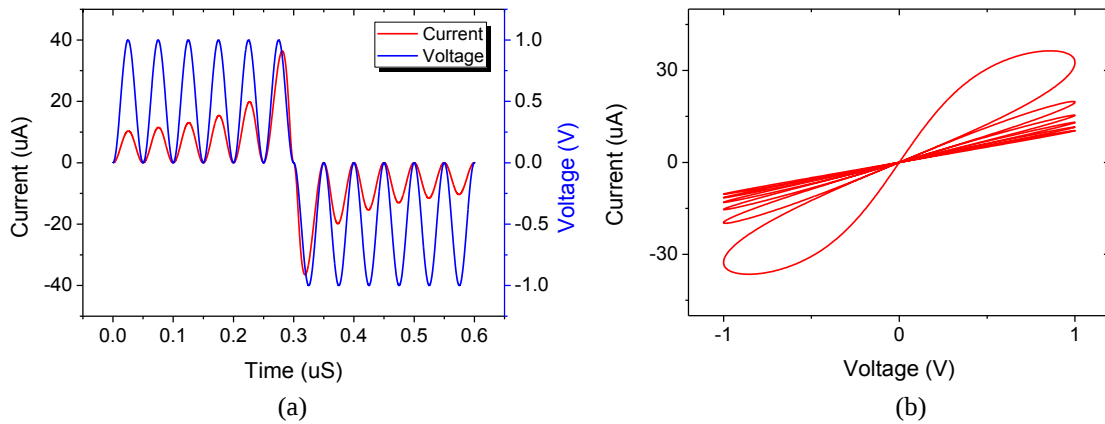
$$\begin{aligned} \frac{dw(t)}{dt} &= \mu_v \frac{R_{ON}}{D} i(t) \\ w(t) &= \mu_v \frac{R_{ON}}{D} q(t) \end{aligned} \quad (2.2)$$

Trong đó μ_v là hệ số thể hiện mật độ các phần tử di động của chất được thêm vào để tạo ra điện trở nhớ. Mức dịch chuyển của biến trạng thái là một hàm theo điện tích. Thay phương trình (2.2) vào phương trình (2.1) và giả định rằng R_{ON} nhỏ hơn R_{OFF} rất nhiều, chúng ta có được điện trở R_X của điện trở nhớ. R_X biểu diễn cho giá trị trở kháng của điện trở nhớ, được gọi là trở kháng nhớ và được đặt tên là $M(q)$ [10].

$$M(q) = R_{OFF} \left(1 - \frac{\mu_v R_{ON}}{D^2} q(t) \right) \quad (2.3)$$

Trở kháng của điện trở nhớ trong phương trình (2.3) là một hàm phụ thuộc vào điện tích, nó bắt đầu từ giá trị R_{OFF} và dần dần giảm xuống mức R_{ON} khi điện

tích liên tục tăng dần lên. Đặc tính của một điện trở nhớ có thể được mô tả như ở Hình (2.2) [10].



Hình 2.2: (a) Mối quan hệ dòng điện và điện áp của điện trở nhớ theo thời gian và (b) đặc tính Dòng-Áp của điện trở nhớ.

Hình 2.2 (a) thể hiện điện áp thay đổi theo thời gian được đặt vào điện trở nhớ và dòng điện tương ứng đi qua điện trở nhớ. Điện áp này bao gồm các pha dương và các pha âm xác định bởi phương trình $\pm v_0 \sin^2(\omega_0 t)$, với $v_0 = 1 \text{ V}$, $f_0 = 20 \text{ MHz}$, $R_{\text{ON}} = 1 \text{ K}\Omega$, $R_{\text{OFF}} = 100 \text{ K}\Omega$. Trong Hình 2.2, điện áp dương được đặt lên điện trở nhớ để làm giảm giá trị điện trở và sau đó giá trị điện áp âm được đặt lên điện trở để tăng điện trở của nó trở lại. Khi điện áp dương được đặt lên 2 đầu, trở kháng của điện trở giảm dần dần về R_{ON} . Điều này được thể hiện bằng sự tăng giá trị của dòng điện đi qua điện trở nhớ (đường màu đỏ trong Hình 2.2(a)). Ngược lại, điện áp âm làm cho trở kháng tăng lên gần với R_{OFF} . Mối quan hệ điện áp và dòng điện trên điện trở nhớ được thể hiện trong Hình 2.2(b). Sự thay đổi điện trở của điện trở nhớ được thể hiện bằng các vòng cánh bướm (Butterfly curve) trong đồ thị điện áp và dòng điện. Các đồ thị dạng cánh bướm là đặc trưng của các điện trở nhớ thể hiện sự thay đổi giá trị điện trở liên tục khi điện áp đặt vào 2 đầu điện trở nhớ đổi dấu.

2.2 MÔ HÌNH HÓA ĐIỆN TRỞ NHỚ

Việc mô hình hóa các điện trở nhớ đóng một vai trò quan trọng trong việc thiết kế ứng dụng dựa trên điện trở nhớ. Mô hình hóa điện trở nhớ giúp chúng ta

dự đoán được hiệu suất hoạt động của mạch cũng như phân tích được ảnh hưởng của các thông số mạch. Vì vậy, sau mô hình thực nghiệm đầu tiên được đề xuất bởi phòng thí nghiệm HP [10] thì một số phương pháp mô hình hóa cho điện trở nhớ đã xuất hiện. Các mô hình của điện trở nhớ hiện nay về mặt toán học được dựa vào mô hình điện trở nhớ của phòng thí nghiệm HP nhưng với biến trạng thái thay đổi một cách phi tuyến do sự trôi không tuyến tính của chất dẫn bên trong linh kiện gây ra [62, 63].

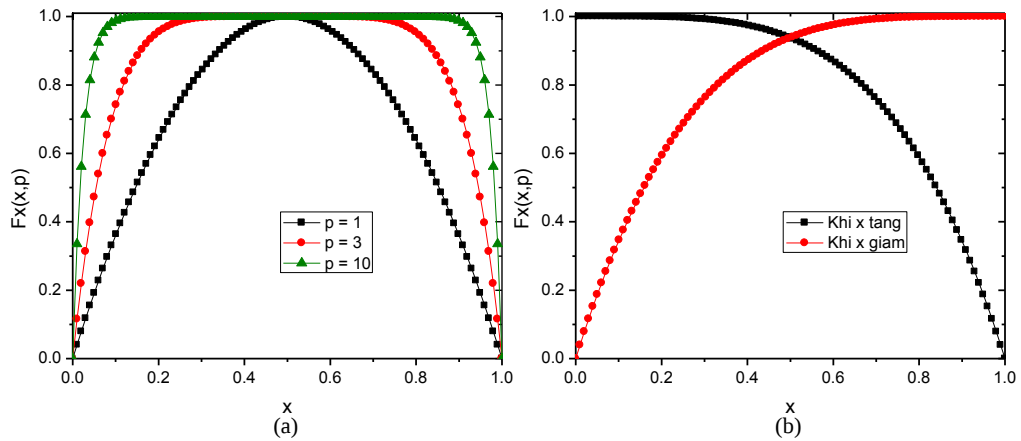
Cụ thể hơn, Y. N. Joglekar và các cộng sự đã đề xuất một hàm cửa sổ biểu diễn sự thay đổi phi tuyến của biến trạng thái với giả định rằng đường phân chia ở giữa hai vùng (vùng giàu hạt dẫn và vùng nghèo hạt dẫn) sẽ di chuyển chậm hơn khi nó ở gần hai cực. Hàm cửa sổ của Joglekar về mặt toán học được mô tả như sau [62]:

$$F(x, p) = 1 - (2x - 1)^{2p} \quad (2.4)$$

Trong đó, $x = w/D$ và x thay đổi trong khoảng từ 0 đến 1. Phương trình (2.4) cho phép ta thay đổi tỉ lệ của cửa sổ bằng cách điều chỉnh tham số p như trong Hình 2.3(a). Từ Hình 2.3(a) ta thấy sự thay đổi của x trở nên chậm hơn khi x tiến tới gần giá trị 0 hoặc 1. Điều này cho phép các mô hình của điện trở nhớ gần giống với thực nghiệm vì các điện trở nhớ thay đổi điện trở của mình chậm hơn khi điện trở gần đạt tới giá trị cao nhất hoặc giá trị thấp nhất.

Một phương pháp khác để mô hình hóa sự dịch chuyển phi tuyến của biến trạng thái đã được đề xuất bởi Z. Biolek và các cộng sự [63]. Sự dịch chuyển của biến trạng thái theo hai đường khác nhau được biểu diễn bằng phương trình (2.5) [63]:

$$F(x, p) = \begin{cases} 1 - x^{2p}, & \text{khi } x \text{ tăng} \\ 1 - (x - 1)^{2p}, & \text{khi } x \text{ giảm} \end{cases} \quad (2.5)$$



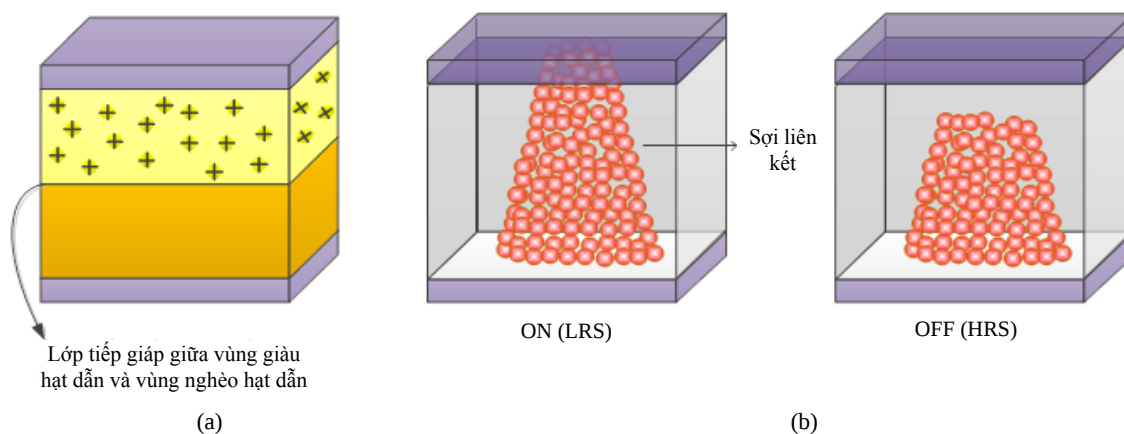
Hình 2.3: (a) Hàm cửa sổ Joglekar với các giá trị p khác nhau và (b) hàm cửa sổ Biolek với giá trị $p = 2$.

Hàm cửa sổ của Biolek được mô tả trong Hình 2.3(b) với $p = 2$. Khi x đang tăng, sự dịch chuyển của x sẽ trở nên chậm hơn khi x tiến gần về giá trị 1. Ngược lại, khi x đang giảm thì sự dịch chuyển của x sẽ trở nên chậm hơn khi x di chuyển ngược về giá trị 0.

2.3 PHÂN LOẠI ĐIỆN TRỞ NHỚ DỰA VÀO ĐẶC TÍNH

Kể từ khi điện trở nhớ đầu tiên dựa trên hợp chất ô-xit titan (TiO_2) được công bố bởi phòng thí nghiệm HP, các điện trở nhớ sau đó được chế tạo và thử nghiệm trên nhiều hợp chất khác nhau với các đặc tính cũng khác nhau. Các điện trở nhớ theo mô hình của HP có điện trở thay đổi lý tưởng từ điện trở cao (R_{OFF}) về điện trở thấp (R_{ON}) và có thể được giữ lại ở một mức điện trở nào đó trong khoảng ($R_{\text{ON}}, R_{\text{OFF}}$). Trong các điện trở nhớ này, lớp tiếp giáp giữa vùng nghèo hạt dẫn và vùng giàu hạt dẫn có thể bị di chuyển và được điều khiển bởi điện áp đặt vào điện trở nhớ. Các điện trở nhớ này được gọi là các điện trở nhớ chuyển đổi lớp tiếp giáp (interface-switching memristor) và hoạt động như các điện trở nhớ tương tự vì có khả năng ghi nhớ nhiều mức điện trở khác nhau [11, 14, 27, 64, 65]. Mô hình của các điện trở nhớ dựa trên chuyển đổi lớp tiếp giáp được trình bày trong Hình 2.4(a). Các điện trở nhớ chuyển đổi lớp tiếp giáp có thể được chế tạo với một số loại vật liệu như: ô-xit Titan [14, 64], hỗn hợp Ag/Si [11],

TiN/PCMO ($\text{Pr}_{0.7}\text{Ca}_{0.3}\text{MnO}_3$) [66], ô-xít Titan kết hợp với linh kiện lựa chọn bằng hợp chất a-ITZO (amorphous In-Sn-Zn-O) [18]Các điện trở nhớ dựa trên chuyển đổi lớp tiếp giáp lý tưởng cho nhiều thiết kế trong đó chủ yếu là bộ nhớ điện trở [67, 68] và mô hình hóa các khớp thần kinh [11, 14, 64, 66].

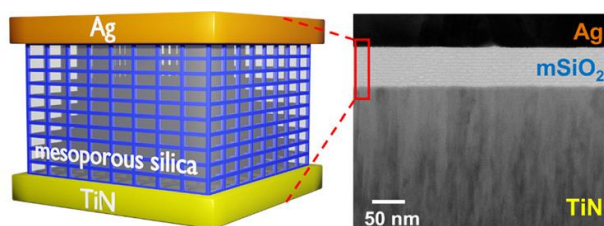


Hình 2.4: (a) Cấu trúc điện trở nhớ dựa trên chuyển đổi lớp tiếp giáp [11], (b) Cấu trúc điện trở nhớ dựa trên chuyển đổi sợi.

Việc chế tạo các điện trở nhớ dựa trên chuyển đổi lớp tiếp giáp thường khó khăn và vật liệu dùng để chế tạo các điện trở nhớ này cũng phức tạp và không phổ biến. Trong khi đó, nhiều phòng nghiên cứu đã lựa chọn giải pháp là sử dụng cơ chế chuyển đổi sợi (filamentary-switching) để chế tạo các điện trở nhớ do vật liệu chuyển đổi sợi phổ biến hơn [17, 44, 47, 69-75]. Hình 2.4(b) minh họa cấu trúc của một điện trở nhớ dựa trên cơ chế chuyển đổi sợi.

Cơ chế chuyển đổi sợi trong các điện trở nhớ đã được chế tạo thành công trên các vật liệu khác nhau như TiO_2 [76], NiO [70, 77], GeS_2 [72], hỗn hợp ZnO và MgO [73], ZnO [74], HfAlO_x (HfO_2 , Al_2O_3) [25], mSiO_2 (mesoporous silica) [49] ... như minh họa ở Hình 2.5. Khi điện áp đặt lên hai cực của thiết bị tăng đủ để điều khiển các ion bên trong sợi thì các sợi sẽ tạo thành một đường dẫn điện giữa hai tiếp điểm kim loại, làm cho thiết bị có trở kháng thấp. Ngược lại, điện áp âm đặt lên hai cực của thiết bị sẽ phá vỡ đường dẫn sợi và giữa hai cực của điện trở nhớ sẽ có trạng thái trở kháng cao. Bằng cách làm này, các điện trở nhớ có thể được lập trình để có trạng thái trở kháng cao hoặc trạng thái trở kháng thấp

như được minh họa trong Hình 2.4 (b). Do vậy, các điện trở nhớ hoạt động dựa trên cơ chế chuyển đổi sợi còn được xem như là các điện trở nhớ nhị phân.



Hình 2.5: Một điện trở nhớ nhị phân được chế tạo thành công theo cơ chế chuyển đổi sợi dùng vật liệu mSiO₂ (mesoporous silica) [49].

Việc nghiên cứu chế tạo các điện trở nhớ nhị phân cũng đã đạt được nhiều thành tựu nổi bật trong thời gian qua. Ở một nghiên cứu được công bố vào năm 2024, các tác giả đã chế tạo thành công một mảng 32×32 điện trở nhớ nhị phân bằng vật liệu Ru/HfSiO_y/Al₂O₃/HfSiO₃/TiN với kích thước chế tạo đạt 2μm×2μm [16]. Vào năm 2025, S. Jain Smarth và nhóm nghiên cứu của mình đã nghiên cứu chế tạo thành công mảng điện trở nhớ 32x32 theo cấu trúc 1 bộ lựa chọn và 1 memristor (1S1M) với vật liệu Si và HfSe₂ để khắc phục vấn đề dòng rò trong mảng, cho các mạng nơ-ron nhân chấp [21]. Trong một nghiên cứu vào năm 2021, nhóm tác giả đã thực thi một mảng 3D các điện trở nhớ nhị phân trên vật liệu HfAlO_x, trong đó mỗi điện trở có thể hoạt động được ở một trạng thái điện trở cao (HRS) và 4 trạng thái điện trở thấp (LRS) với tỉ lệ HRS/LRS đạt tối đa là 100 [25]. Vào năm 2025, S. He và các cộng sự đã nghiên cứu và chế tạo thành công một mảng điện trở nhớ với công nghệ tự chỉnh lưu có kích thước 16x16, dựa trên vật liệu Pt/HfO₂/Ta₂O_{5-x}/Ti, cho tỉ lệ HRS/LRS có thể đạt từ 10² đến 10³ [50]. Trong một nghiên cứu khác, nhóm tác giả đã chế tạo một mảng các điện trở nhị phân với vật liệu mSiO₂ (mesoporous silica) có thể đạt được mức trạng thái điện trở cao ở giá trị hơn 70 MΩ và ba mức trạng thái điện trở thấp, thấp nhất khoảng 2 KΩ, đạt tỉ lệ HRS/RLS lớn hơn 10000 [49].

2.4 GIỚI THIỆU HỆ ĐIỆN TOÁN MÔ PHỎNG HỆ THẦN KINH

Vào năm 1990, giáo sư Carver A. Mead thuộc viện Kỹ thuật California đã đưa ra khái niệm hệ điện toán mô phỏng hệ thần kinh (Neuromorphic Computing

System) [1]. Hệ điện toán mô phỏng hệ thần kinh là một hệ thống phần cứng có khả năng tính toán, xử lý như cách mà bộ não con người thực hiện: các tín hiệu vào song song và được xử lý một cách đồng thời. Các hệ điện toán mô phỏng hệ thần kinh về cơ bản khác với các hệ thống dựa trên trí tuệ nhân tạo thông dụng hiện nay. Các hệ thống trí tuệ nhân tạo hiện nay chủ yếu dựa trên các mạng nơ-ron nhân tạo nhiều lớp được thực thi dựa trên các máy tính có cấu hình mạnh. Do vậy, một hệ thống trí tuệ nhân tạo có thể xem như là một chương trình phần mềm chạy trên một hệ thống máy tính cụ thể mà ở đó các tính toán được xử lý một cách tuần tự. Trong khi đó, một hệ điện toán mô phỏng hệ thần kinh là một mô hình của mạng nơ-ron nhân tạo nhưng được thực thi trên một phần cứng, chẳng hạn như trên một thiết kế vi mạch tích hợp cỡ lớn (Very-large-scale integration – VLSI), với khả năng xử lý song song đồng thời nhiều tín hiệu và phép toán. Như vậy, về cơ bản có thể xem một vi mạch thực thi một mạng nơ-ron nhân tạo là một hệ điện toán mô phỏng hệ thần kinh.

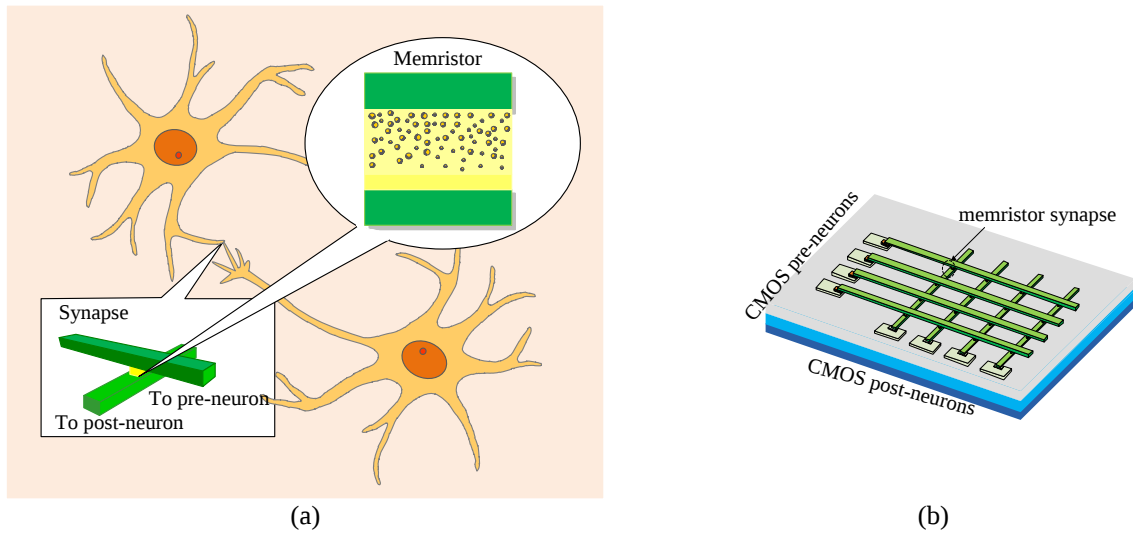
Một nơ-ron sinh học sẽ có mô hình toán bao gồm các phép nhân của các tín hiệu vào với các trọng số được thực hiện đồng thời và các phép cộng để thực hiện cộng các kết quả tác động của tín hiệu ngõ vào. Do đó, khi thực thi mô hình của một nơ-ron sinh học, các tín hiệu vào cần được xử lý đồng thời như cách mà bộ não con người thực hiện. Các mạng nơ-ron nhân tạo trong các hệ thống trí tuệ nhân tạo hiện nay thông thường được thực thi trên các máy tính. Tuy nhiên, máy tính là một đơn vị xử lý tuần tự, nên việc thực thi các mạng nơ-ron nhân tạo trên các máy tính thường không cho tốc độ nhanh và năng lượng tiêu tốn ít như một bộ não sinh học. Mặc dù ngày nay sự ra đời của các công nghệ mới đã cho phép máy tính xử lý song song nhiều tác vụ để tăng hiệu năng khi thực thi các mạng nơ-ron nhân tạo, nhưng cách xử lý của các hệ thống máy tính vẫn còn khác với cách mà một bộ não sinh học xử lý. Ngược lại, các hệ điện toán mô phỏng hệ thần kinh hướng đến các thiết kế phần cứng có thể thực thi hoàn toàn song song các phép toán giữa các tín hiệu vào và trọng số tương ứng, giống với cách mà một bộ não sinh học hoạt động. Do đó, các thiết kế này được xem như là những hệ thống mô phỏng hệ thần kinh sinh học với kỳ vọng thực thi các mạng nơ-ron

nhân tạo với tốc độ nhanh và tiêu tốn ít năng lượng hơn so với khi thực thi trên các máy tính truyền thống.

Sau khi được đề xuất, các hệ điện toán mô phỏng hệ thần kinh đã được thực thi dựa trên các hệ thống vi mạch tích hợp cỡ lớn (VLSI) với mục đích mô hình hóa chức năng của các mạng nơ-ron sinh học. Các thiết kế VLSI thực thi các hệ điện toán mô phỏng hệ thần kinh đã đạt được nhiều kết quả vượt trội trong thời gian qua [2, 3]. Trong một nghiên cứu công bố vào năm 2015, C. Mayr và các cộng sự đã đề xuất một hệ điện toán mô phỏng hệ thần kinh dạng tụ điện – đóng ngắt để thực hiện ghép nối nơ-ron vòng kín trên công nghệ CMOS 28 nanomet có kích thước 600×600 micromet với 128 ngõ vào, 64 ngõ ra, 8192 khớp nối thần kinh [3]. Tuy nhiên, các hệ thống VLSI hiện nay đều được chế tạo dựa vào các công nghệ CMOS và công nghệ này đến nay đã đạt tới giới hạn, khả năng phát triển cũng như khả năng giảm kích thước công nghệ còn rất thấp [6-8]. Do đó, để thực thi các hệ điện toán mô phỏng hệ thần kinh, cần có một linh kiện thiết kế vi mạch mới với tiềm năng phát triển công nghệ và giảm kích thước tốt hơn để thay thế nhằm khắc phục được các giới hạn này.

Kể từ khi được thực nghiệm thành công vào năm 2008 bởi giáo sư R. S. Williams và các cộng sự tại phòng thí nghiệm HP, điện trở nhớ đã được xem như là một linh kiện mô hình hóa hiệu quả cho các khớp nối thần kinh [11, 12] nhờ vào khả năng lưu lại giá trị điện trở hiện tại cũng như khả năng thay đổi giá trị điện trở bằng điện áp lập trình đặt vào 2 cực của nó. Sự thay đổi giá trị điện trở của một điện trở nhớ giống như sự thay đổi giá trị trọng số của một khớp nối thần kinh để cho phép tín hiệu truyền qua khớp nối này là nhiều hay ít [11]. Việc sử dụng điện trở nhớ để thực thi một khớp nối thần kinh trong mạng nơ-ron nhân tạo được minh họa ở Hình 2.6(a) [11]. Đặc biệt, các điện trở nhớ có thể được kết nối với nhau dưới dạng mảng mà tại mỗi điểm giao nhau giữa một hàng và một cột sẽ có một điện trở nhớ đóng vai trò như một thiết bị chuyển mạch [14-19]. Như minh họa ở Hình 2.6(b), mảng điện trở nhớ bao gồm 16 điện trở nhớ, được bố trí tại các giao điểm của 4 hàng và 4 cột của mảng. Mỗi điện trở nhớ sẽ thực thi vai trò của một khớp nối thần kinh, để liên kết một nơ-ron ngõ vào với một

ơ-ron ngõ ra, với giá trị trọng số có thể lập trình và ghi nhớ được thông qua điện áp lập trình.. Hơn thế nữa, các mảng điện trở nhớ có thể được thiết kế theo kiến trúc 3 chiều (3-D) để hiện thực hóa khả năng kết nối cao của hệ thống ơ-ron trong các bộ não sinh học [22-25].



Hình 2.6: (a) Điện trở nhớ được sử dụng để mô hình hóa một khớp nối thần kinh, (b) mô hình một mảng điện trở nhớ hai chiều [11].

Có thể thấy, các mảng điện trở nhớ là giải pháp hữu hiệu để xây dựng các hệ điện toán mô phỏng hệ thần kinh bằng cách thực thi các mạng ơ-ron nhân tạo. Các tín hiệu ngõ vào của một lớp của mạng ơ-ron, được mã hóa dưới dạng điện áp đầu vào, sẽ được đưa vào các hàng của mảng điện trở nhớ, đi qua các điện trở nhớ và được tổng hợp ở các cột ngõ ra của mảng. Lúc này, mỗi điện trở nhớ tại giao điểm của một hàng và một cột trong mảng sẽ thực thi vai trò của một khớp nối thần kinh với trọng số có thể thay đổi nhờ lập trình. Khi đó, phép nhân giữa các tín hiệu ngõ vào với các trọng số của mạng ơ-ron sẽ được thực thi trực tiếp trên phần cứng theo định luật Kirchhoff và định luật Ohm. Do đó, mỗi cột ngõ ra sẽ là tổng hợp của các tín hiệu ngõ vào ở các hàng theo các trọng số khác nhau và mỗi cột ngõ ra sẽ đóng vai trò là một ơ-ron ngõ ra của lớp hiện tại của mạng ơ-ron cần thực thi.

Trong thời gian gần đây, khi xây dựng các hệ điện toán mô phỏng hệ thần kinh, các mảng điện trở nhớ tương tự đã được sử dụng để thực thi các mạng nơ-ron nhân tạo [27-33] như là một giải pháp thiết kế đạt hiệu quả cao và công suất tiêu thụ thấp. Khi thực thi các mạng nơ-ron nhân tạo với mảng điện trở nhớ tương tự, để thực hiện một khớp nối thần kinh với trọng số có dấu (dấu + hoặc dấu -), 2 điện trở nhớ tương tự đã được sử dụng [27, 28, 30, 33-35]. Độ lớn của 2 điện trở nhớ này được điều chỉnh để thu được giá trị và dấu của trọng số tương ứng của một khớp nối thần kinh. Trong một nghiên cứu vào năm 2019, J. Fu và các cộng sự đã thực thi một mạng nơ-ron có cấu trúc gồm 400 nơ-ron ngõ vào, 100 nơ-ron ở 1 lớp ẩn và 10 nơ-ron ngõ ra và đề xuất phương pháp học Quy hoạch Tuyến tính (Linear Optimization – LO) để nhận dạng 10 ký tự viết tay của tập dữ liệu MNIST [31]. Kết quả độ chính xác khi nhận dạng của hệ thống đạt trên 75% với mức độ nhiễu và độ biến thiên giá trị điện trở thấp. Ở một nghiên cứu khác vào năm 2019, O. Krestinskaya và các cộng sự đã xây dựng một mạng ANN gồm 28×28 nơ-ron ngõ vào, 42 nơ-ron lớp ẩn, 10 nơ-ron ngõ ra nhằm thực thi nhận dạng 10 ký tự trong tập MNIST [33]. Trong nghiên cứu này, thay vì chỉ sử dụng một mảng điện trở nhớ lớn, nhóm tác giả đã tiếp cận theo hướng module: chia một mảng điện trở nhớ lớn ra thành nhiều module nhỏ hơn để giảm dòng rò trong mảng và đề xuất mạch học lan truyền ngược tương ứng cho mạng; kết quả độ chính xác khi nhận dạng các ký tự trong tập MNIST của mạng đạt 92%. Trong một nghiên cứu khác vào năm 2020, S. Wen và các cộng sự đã thiết kế một cấu trúc mạng nơ-ron tích chập nhỏ gọn dự phòng dựa trên điện trở nhớ được (MSCCNN- a memristor-based sparse compact convolutional neural network) có cấu trúc gồm 3 khối: 1 lớp nhân chập và theo sau đó là 2 khối song song (Inception blocks) [30]. Trong thiết kế này, nhóm tác giả đã sử dụng các mảng điện trở nhớ để lưu trữ các trọng số có dấu (+ và -) để thực hiện các phép nhân chập với các tín hiệu ngõ vào, cho kết quả độ chính xác khi nhận dạng đạt 67.21% trên tập dữ liệu CIFAR-10 [30].

Có thể thấy, các mảng điện trở nhớ tương tự đã được xem là một giải pháp cho việc xây dựng các hệ điện toán mô phỏng hệ thần kinh bằng việc thực thi các

mạng nơ-ron nhân tạo, thay thế cho việc thực thi trên các kiến trúc mạch phân cứng VLSI vốn đã đạt tới giới hạn về kích thước. Tuy nhiên, việc sử dụng điện trở nhớ như một linh kiện tương tự vẫn còn tồn tại nhiều khó khăn và thách thức cần phải đối mặt và vượt qua liên quan đến đặc tuyến phi tính của linh kiện này.

Một hướng tiếp cận khác, nhiều ưu điểm nổi bật hơn so với sử dụng các mảng điện trở nhớ tương tự, là hướng nghiên cứu sử dụng các mảng điện trở nhớ nhị phân để thực thi các hệ điện toán mô phỏng hệ thần kinh [44, 46-48]. Các điện trở nhớ nhị phân dễ dàng được chế tạo và lập trình hơn so với điện trở nhớ tương tự do chỉ cần lưu trữ hai mức điện trở, điện trở thấp (LRS) và điện trở cao (HRS). Hơn nữa, sự biến thiên giá trị do quá trình chế tạo và hoạt động của điện trở nhớ nhị phân cũng ít ảnh hưởng đến hiệu suất hoạt động của hệ thống hơn so với điện trở nhớ tương tự [44] nhờ tỷ lệ HRS/LRS (R_{OFF} / R_{ON}) khá cao [25, 49, 50]. Đồng thời, phép nhân giữa các tín hiệu ngõ vào với các trọng số nhị phân cũng sẽ dễ dàng được thực thi hơn bằng các phép toán logic như AND, XNOR.

Như vậy, các mảng điện trở nhớ nhị phân đã mở ra một hướng tiếp cận mới, tiềm năng và khả thi hơn cho việc xây dựng các hệ điện toán mô phỏng hệ thần kinh, đặc biệt là ở các bài toán nhận dạng mẫu như nhận dạng tiếng nói, nhận dạng ảnh.

2.5 GIỚI THIỆU CÁC KIẾN TRÚC MẢNG ĐIỆN TRỞ NHỚ NHỊ PHÂN

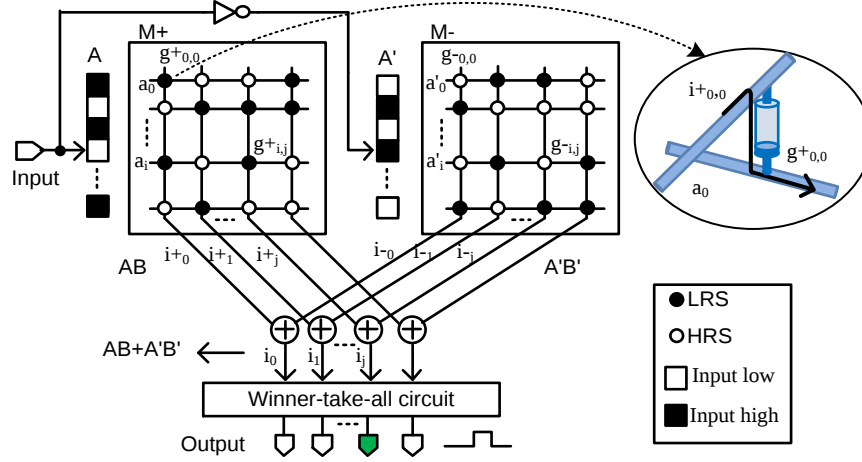
2.5.1 Kiến trúc mảng điện trở nhớ bù

Kiến trúc mảng điện trở nhớ bù được giới thiệu nhằm thực hiện nhận dạng một vector đầu vào A với một tập dữ liệu B bằng cách thực thi hàm XNOR theo phương trình [44]:

$$Y = \overline{A \oplus B} = AB + A'B' = A \cdot (M+) + A' \cdot (M-) \quad (2.6)$$

Ở đây, A là vector dữ liệu đầu vào cần nhận dạng, B là mảng lưu trữ tập dữ liệu. Trong phương trình (2.6), A' là vector đảo của vector đầu vào A , B' là mảng chứa các giá trị đảo của mảng dữ liệu B . Phép XNOR giữa đầu vào A và mảng dữ liệu B được khai triển thành hai phép AND và một phép cộng. Do vậy, để

thực thi phương trình XNOR (2.6), kiến trúc mảng điện trở nhớ bù cần sử dụng 2 mảng điện trở nhớ bù nhau, mảng $M +$ và mảng $M -$ để lưu trữ dữ liệu của B và B' tương ứng. Mô hình của kiến trúc mảng điện trở nhớ bù được thể hiện ở hình 2.7 [48].



Hình 2.7: Sơ đồ khối của kiến trúc mảng điện trở nhớ bù.

Trên Hình 2.7, $M +$ là mảng điện trở nhớ có kích thước $n \times m$, được sử dụng để lưu tập dữ liệu mẫu B . Mỗi mẫu dữ liệu của tập mẫu B sẽ được lưu trên một cột của mảng điện trở nhớ $M +$. Tại vị trí giao nhau của hàng i và cột j của mảng điện trở nhớ $M +$ là một điện trở nhớ có giá trị điện dẫn $g +_{ij}$ tạo thành giá trị điện trở nhớ $M +_{ij} = \frac{1}{g +_{ij}}$. Ở đây, giá trị điện trở nhớ $M +_{ij}$ sẽ được thiết lập ở một trong hai trạng thái: điện trở thấp (LRS) hoặc điện trở cao (HRS). Giá trị điện trở nhớ $M +_{ij}$ sẽ được dùng để mã hóa và lưu thông tin của một bit dữ liệu, giá trị điện trở thấp tương ứng với bit dữ liệu 1 và giá trị điện trở cao tương ứng với bit dữ liệu 0. Mảng điện trở nhớ thứ 2, mảng $M -$, chứa các điện trở nhớ có giá trị bù với các điện trở nhớ ở vị trí tương ứng trong mảng $M +$. Nếu $M +_{ij}$ có giá trị điện trở cao thì phần tử $M -_{ij}$ sẽ có giá trị điện trở thấp và ngược lại. Do đó, mảng $M -$ là mảng bù của mảng $M +$ và kiến trúc này được gọi là kiến trúc mảng điện trở nhớ bù.

Để thực thi nhận dạng với kiến trúc mảng điện trở nhớ bù, vector dữ liệu cần nhận dạng A sẽ được chuyển đổi thành các điện áp ngõ vào, bit dữ liệu 1 tương ứng với điện áp cao và bit dữ liệu 0 tương ứng với điện áp thấp, và được đưa vào mảng điện trở nhớ $M +$ nhằm thực thi phép AND thứ nhất $A \cdot (M +)$ của phương trình (2.6) theo định luật Ohm và định luật Kirchhoff và thu được các dòng điện ngõ ra cột $i +_j$. Đồng thời, A' là vector đảo của A cũng sẽ được đưa vào mảng điện trở nhớ $M -$ để thực hiện phép AND thứ hai $A' \cdot (M -)$ của phương trình (2.6) để thu được các dòng điện ngõ ra cột $i -_j$. Sau đó, các dòng điện ngõ ra cột $i +_j$ và $i -_j$ sẽ được cộng lại để thu được các dòng điện ngõ ra cột cuối cùng i_j là kết quả của việc thực thi hàm XNOR (2.6) giữa vector điện áp ngõ vào A và các mảng điện trở nhớ. Bằng cách thực thi này, nếu vector dữ liệu cần nhận dạng A trùng khớp với mẫu dữ liệu ở cột thứ k của tập dữ liệu B thì dòng điện ngõ ra cột i_k của cột thứ k sẽ là dòng điện có giá trị lớn nhất. Sau đó, các dòng điện ngõ ra cột sẽ được đưa vào mạch Tìm dòng điện lớn nhất (Winner-take-all circuit) để so sánh và tìm ra dòng điện ngõ ra lớn nhất giữa các cột. Nếu mạch Tìm dòng điện lớn nhất phát hiện được rằng dòng điện ngõ ra i_k của cột thứ k là dòng điện lớn nhất, nó sẽ tạo ra một xung ở ngõ ra thứ k để thông báo kết quả vector đầu vào A trùng khớp với dữ liệu lưu ở cột thứ k của các mảng điện trở nhớ. Như vậy, kiến trúc mảng điện trở nhớ bù sẽ sử dụng hai mảng điện trở nhớ bù nhau, $M +$ và $M -$, để thực thi nhận dạng một vector ngõ vào A với các cột của tập dữ liệu B theo phương trình XNOR (2.6) để nhận dạng [78]. Kiến trúc mảng điện trở nhớ bù được ứng dụng trong thực thi các hệ điện toán mô phỏng hệ thần kinh cho các bài toán nhận dạng như nhận dạng tiếng nói [44], nhận dạng ảnh [47, 54, 55, 78].

2.5.2 Kiến trúc mảng điện trở nhớ đồng nhất

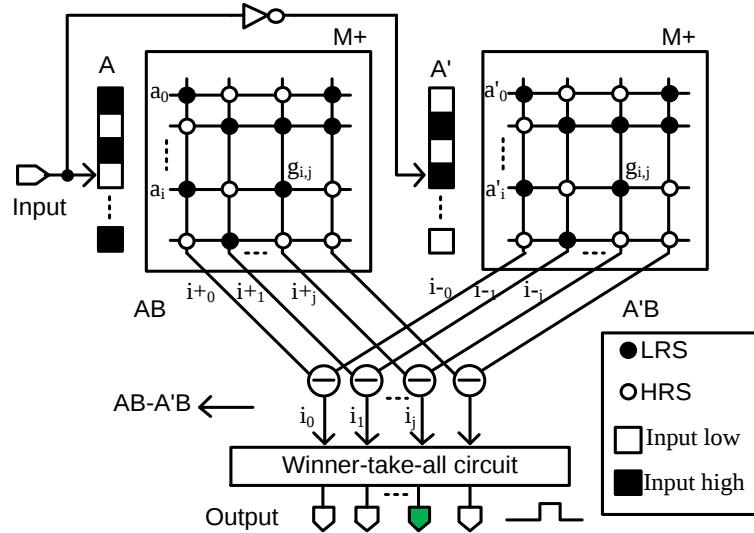
Kiến trúc mảng điện trở nhớ đồng nhất được đề xuất để thực thi nhận dạng một ảnh đầu vào A với tập dữ liệu B bằng hai mảng điện trở nhớ giống nhau nhờ vào việc khai triển phương trình XNOR như sau:

$$\begin{aligned}
Y &= \overline{A \oplus B} = AB + A'B' = AB + A'(1 - B) \\
&= AB - A'B + A'
\end{aligned}
\tag{2.7}$$

Ở phương trình (2.7), phần tử A' là một hằng số và không có sự tương tác với mảng dữ liệu B . Do đó, A' có thể được lược bỏ khi thực hiện phép XNOR giữa vector ngõ vào A và tập dữ liệu B và phương trình 2.7 có thể được viết tương đương như sau [47]:

$$\begin{aligned}
Y &= \overline{A \oplus B} = AB - A'B \\
&= A \cdot (M+) - A' \cdot (M+)
\end{aligned}
\tag{2.8}$$

Lúc này, hàm XNOR giữa ngõ vào A và mảng B chứa tập dữ liệu đã được biến đổi tương đương thành hai phép AND và một phép trừ và cần hai mảng điện trở nhớ giống nhau, $M+$ và $M+$, để thực thi. Mô hình kiến trúc mảng điện trở nhớ đồng nhất thực thi hàm XNOR theo phương trình (2.8) được mô tả chi tiết ở Hình 2.8 [48].



Hình 2.8: Sơ đồ khối của kiến trúc mảng điện trở nhớ đồng nhất.

Như thể hiện ở Hình 2.8, kiến trúc mảng điện trở nhớ đồng nhất sử dụng hai mảng điện trở nhớ giống nhau $M+$ và $M+$ để lưu trữ các ảnh mẫu của tập dữ liệu B nhằm thực thi hàm XNOR với ảnh đầu vào cần nhận dạng A theo phương

trình (2.8). Ở đây, mỗi ảnh mẫu trong tập dữ liệu B sẽ được lưu trên một cột của mảng điện trở nhớ $M +$, bit dữ liệu 1 được mã hóa trong một điện trở nhớ ở trạng thái điện trở thấp (LRS) và bit dữ liệu 0 tương ứng với giá trị điện trở cao (HRS). Mảng điện trở nhớ $M +$ thứ hai sẽ có giá trị hoàn toàn giống với mảng điện trở nhớ $M +$ thứ nhất, hay nói một cách khác là hai mảng điện trở nhớ này là đồng nhất với nhau. Để thực thi nhận dạng, vector ngõ vào A dưới định dạng điện áp, điện áp cao tương ứng với bit dữ liệu 1 và điện áp thấp tương ứng với bit dữ liệu 0, sẽ được đưa vào mảng điện trở nhớ $M +$ thứ nhất để thực thi phép AND thứ nhất $A \cdot (M +)$ trong phương trình (2.8) và thu được các dòng điện ngõ ra cột $i + j$. Đồng thời, vector A' là chứa các giá trị điện áp đảo của vector A cũng được đưa vào mảng điện trở nhớ $M +$ thứ hai để thực thi phép AND thứ hai $A' \cdot (M +)$ và thu được các dòng điện ngõ ra cột $i - j$. Các dòng điện ngõ ra cột cuối cùng, các dòng điện i_j , sẽ được tính toán bằng phép trừ giữa các dòng điện cột $i + j$ và các dòng điện cột $i - j$. Các dòng điện ngõ ra cột cuối cùng này sẽ được đưa vào mạch Tìm dòng điện lớn nhất để so sánh với nhau và tìm ra dòng điện lớn nhất giữa các cột. Nếu mạch Tìm dòng điện lớn nhất xác định được dòng điện ngõ ra cột i_k là lớn nhất, một xung đầu ra sẽ được xuất ở ngõ ra thứ k để thông báo rằng ảnh cần nhận dạng A là trùng khớp với ảnh mẫu lưu ở cột thứ k của mảng điện trở nhớ. Như vậy, kiến trúc mảng điện trở nhớ đồng nhất sử dụng hai mảng điện trở nhớ $M +$ và $M +$ giống nhau để thực thi nhận dạng và đã được chứng minh hiệu quả trong việc giảm công suất tiêu thụ trên mảng điện trở nhớ trong nhận dạng các ảnh đã được biến đổi DCT (Discrete Cosine Transform) [47] và các ảnh có mật độ dữ liệu thấp [79], khi so sánh với kiến trúc mảng điện trở nhớ bù.

2.5.3 Kiến trúc mảng điện trở nhớ đơn

Kiến trúc mảng điện trở nhớ đơn được đề xuất và chứng minh có thể xác định được mức độ giống nhau giữa vector đầu vào A và tập dữ liệu B bằng các triển khai và tinh gọn phương trình của hàm XNOR như sau [48]:

$$\begin{aligned}
Y &= \overline{A \oplus B} = AB + A'B' \\
&= AB + A'(1 - B) \\
&= (A - A')B + A'
\end{aligned} \tag{2.9}$$

Ở phương trình (2.9), phân tử A' là một hằng số và không thực hiện phép tính toán nào với tập dữ liệu B nên có thể được lược bỏ. Do đó, phép XNOR giữa đầu vào A và tập dữ liệu B lúc này được tinh gọn thành một phép AND như sau:

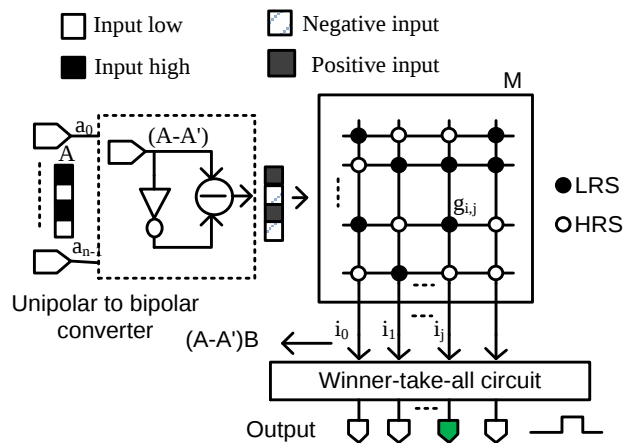
$$Y = \overline{A \oplus B} = (A - A')B = I \cdot M \tag{2.10}$$

Lúc này, $I = (A - A')$ là vector đầu vào lưỡng cực với các giá trị +1 và -1, như được trình bày ở Bảng 2.1, và M là mảng điện trở nhớ duy nhất để lưu các thông tin của tập dữ liệu B .

Bảng 2.1: Các trạng thái của vector đầu vào lưỡng cực I .

A	A'	$I = (A - A')$
1	0	+1
0	1	-1

Từ phương trình XNOR rút gọn (2.10), mô hình kiến trúc mảng điện trở nhớ đơn sẽ được cấu thành từ một bộ chuyển đổi dữ liệu từ đơn cực sang lưỡng cực và một mảng điện trở nhớ M như thể hiện ở Hình 2.9.



Hình 2.9: Sơ đồ khối của kiến trúc mảng điện trở nhớ đơn.

Ở đây, bộ chuyển đổi dữ liệu từ đơn cực sang lưỡng cực sẽ nhận vector đầu vào đơn cực A và thực thi phép trừ $(A - A')$ của phương trình (2.9) để tạo thành vector dữ liệu lưỡng cực I tương ứng với các trạng thái thể hiện ở Bảng 2.1. Một mảng điện trở nhớ M duy nhất sẽ được sử dụng để lưu các ảnh mẫu của tập dữ liệu B , mỗi ảnh được lưu trên một cột của mảng M . Mỗi bit dữ liệu sẽ được lưu trong một điện trở nhớ của mảng M ở trạng thái điện trở thấp nếu là bit dữ liệu 1 và ở trạng thái điện trở cao nếu là bit dữ liệu 0.

Khi thực thi nhận dạng với kiến trúc mảng điện trở nhớ đơn, ảnh cần nhận dạng là vector A dưới định dạng điện áp đơn cực sẽ được đưa vào bộ chuyển đổi dữ liệu đơn cực sang lưỡng cực để thu được dữ liệu lưỡng cực I ở dạng điện áp dương và điện áp âm, tương ứng với dữ liệu +1 và -1. Sau đó, các điện áp lưỡng cực của vector I sẽ được đưa vào mảng điện trở nhớ M để thực hiện phép AND duy nhất $(I \cdot M)$ của phương trình XNOR (2.9) theo định luật Ohm và định luật Kirchhoff và thu được các dòng điện ngõ ra cột. Độ lớn của các dòng điện ngõ ra cột sẽ thể hiện được mức độ giống nhau giữa ảnh cần nhận dạng A với các ảnh của tập dữ liệu được lưu ở các cột của mảng điện trở nhớ M . Sau đó, các dòng điện ngõ ra cột sẽ được đưa vào mạch Tìm dòng điện lớn nhất để so sánh với nhau và tìm ra dòng điện lớn nhất. Khi mạch Tìm dòng điện lớn nhất xác định được dòng điện ngõ ra i_j của cột thứ j là lớn nhất thì mạch sẽ tạo một xung ở ngõ ra thứ j để cho thấy ảnh cần nhận dạng A là trùng khớp với ảnh dữ liệu lưu ở cột thứ j của mảng điện trở nhớ M [80]. Kiến trúc mảng điện trở nhớ đơn đã được chứng minh giảm 50% số lượng điện trở nhớ, có công suất tiêu thụ trên mảng thấp hơn và khả năng chịu lỗi trên phần tử mảng cao hơn khi so sánh với các kiến trúc mảng điện trở nhớ bù và kiến trúc mảng điện trở nhớ đồng nhất vốn sử dụng hai mảng điện trở nhớ để thực thi hàm XNOR [48]. Kiến trúc mảng điện trở nhớ đơn cũng có thể hoạt động hiệu quả dưới sự biến thiên của điện trở nhớ trong mảng [80].

2.6 KẾT LUẬN CHƯƠNG

Ở chương này, các lý thuyết nền tảng phục vụ cho quá trình nghiên cứu và thực hiện của luận án đã được trình bày chi tiết. Trước hết, các đặc tính của linh kiện điện trở nhớ đã được khảo sát và trình bày, bao gồm: cấu tạo, phương trình toán học, đặc tuyến dòng áp, các phương pháp mô hình hóa và phân loại. Đồng thời, các kiến thức nền tảng và các hướng thực thi của hệ điện toán mô phỏng hệ thần kinh cũng đã được giới thiệu và khảo sát chi tiết. Sau đó, phương trình toán và cấu trúc của các kiến trúc mảng điện trở nhớ nhị phân ứng dụng trong thiết kế hệ điện toán mô phỏng hệ thần kinh cũng đã được phân tích, bao gồm: kiến trúc mảng điện trở nhớ bù, kiến trúc mảng điện trở nhớ đồng nhất, kiến trúc mảng điện trở nhớ đơn.

Ở các chương tiếp theo, luận án sẽ tiếp tục phân tích các yếu tố ảnh hưởng đến quá trình thực thi mảng XNOR trong các bài toán nhận dạng của các kiến trúc mảng điện trở nhớ nhị phân, tìm ra ưu điểm cũng như hạn chế của các kiến trúc này. Từ đó, luận án sẽ nghiên cứu và đề xuất một kiến trúc mảng điện trở nhớ nhị phân tối ưu để thực thi mảng XNOR trong thiết kế hệ điện toán mô phỏng hệ thần kinh ở các bài toán nhận dạng, hướng tới sự tối ưu về số lượng và kích thước mảng điện trở nhớ trong khi vẫn đảm bảo về hiệu năng hoạt động của hệ thống.

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

Ở chương này, trước hết, Luận án sẽ phân tích về những thách thức khi xây dựng các hệ điện toán mô phỏng thần kinh với mảng điện trở nhớ tương tự. Những khó khăn và thách thức này là minh chứng cho thấy việc phát triển các hệ điện toán mô phỏng hệ thần kinh với các mảng điện trở nhớ nhị phân, thay vì các mảng điện trở nhớ tương tự, là một yêu cầu cấp thiết và nhiều triển vọng. Tiếp đến, Luận án sẽ phân tích chi tiết các yếu tố ảnh hưởng đến hiệu suất hoạt động của các kiến trúc mảng điện trở nhớ nhị phân, bao gồm nhiều trên dữ liệu và sự biến thiên điện trở, nhằm hướng tới việc nghiên cứu và đề xuất một kiến trúc mảng điện trở nhớ nhị phân tối ưu cho thực thi các hệ điện toán mô phỏng hệ thần kinh trong các bài toán nhận dạng. Các kết quả phân tích sự ảnh hưởng của các yếu tố này đã được công bố trong bài báo tạp chí quốc tế, ở Công trình công bố sau của Luận án: (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).

3.1 GIỚI THIỆU

Hệ điện toán mô phỏng hệ thần kinh là một cách tiếp cận các mô hình học sâu trên phần cứng mà cụ thể là vi mạch. Các mô hình học sâu luôn đòi hỏi một bộ nhớ lớn để lưu số lượng lớn các thông số, đồng thời tài nguyên tính toán phục vụ cho các phép tính trong mô hình trong đó chủ yếu là phép toán nhân, phép toán cộng và các phép toán thực hiện hàm kích hoạt. Thực thi các phép toán này

trên các vi mạch VLSI đòi hỏi nhiều tài nguyên. Một trong những phương pháp thiết kế tiềm năng cho các hệ điện toán mô phỏng hệ thần kinh là sử dụng linh kiện điện trở nhớ có khả năng lưu trữ giá trị thay cho việc sử dụng bộ nhớ dựa trên công nghệ CMOS. Hơn nữa, các phép toán cộng và nhân cũng có thể được thực hiện hiệu quả thông qua mô hình dòng và áp của định luật Ohm và định luật Kirchhoff. Mạng thuần điện trở nhớ, không sử dụng thêm các transistor nối tiếp với điện trở nhớ tại các giao điểm của hàng và cột đã trở thành một kiến trúc lý tưởng cho việc thực thi các mạng neuron nhân tạo trên vi mạch. Trong chương này, luận án sẽ 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) và điện trở nhớ được sử dụng ở 2 mức (nhị phân, hoặc số). Những khó khăn và thách thức này cho thấy nhu cầu cấp thiết và triển vọng của việc phát triển các hệ điện toán mô phỏng hệ thần kinh trên nền tảng các mảng điện trở nhớ nhị phân. Đồng thời, sự ảnh hưởng của các yếu tố bao gồm nhiều trên dữ liệu và sự biến thiên điện trở, đến hiệu suất hoạt động của các kiến trúc mảng điện trở nhớ nhị phân trong bài toán nhận dạng ảnh cũng sẽ được phân tích cụ thể. Các kết quả này sẽ là cơ sở cho việc lựa chọn một kiến trúc mảng điện trở nhớ nhiều tiềm năng để phát triển tối ưu nhằm ứng dụng trong thiết kế các hệ điện toán mô phỏng hệ thần kinh ở các nghiên cứu tiếp theo của Luận án.

3.2 NHỮNG THÁCH THỨC KHI XÂY DỰNG CÁC HỆ ĐIỆN TOÁN MÔ PHỎNG HỆ THẦN KINH VỚI MẢNG ĐIỆN TRỞ NHỚ TƯƠNG TỰ

Tuy việc xây dựng các hệ điện toán mô phỏng hệ thần kinh bằng cách thực thi các mạng nơ-ron nhân tạo với các mảng điện trở nhớ tương tự là một hướng ứng dụng có tiềm năng của điện trở nhớ, nhưng vẫn còn nhiều khó khăn và thách thức đã và đang cản trở việc hiện thực hóa các mạng nơ-ron nhân tạo trên các mảng điện trở nhớ tương tự.

Trước hết, có thể thấy, 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 [31, 36]. Khi thực thi các mạng nơ-

ron nhân tạo, các điện trở nhớ sẽ thực thi chức năng của các khớp nối thần kinh với các giá trị trọng số khác nhau và các trọng số cần được cập nhật giá trị liên tục trong quá trình học bằng cách thay đổi số lượng xung ngõ vào huấn luyện. Tuy nhiên, giá trị điện dẫn thực tế đạt được của điện trở nhớ sau khi áp dụng các xung huấn luyện thường khác với giá trị mong muốn do đặc tính phi tuyến của điện trở nhớ. Do đó, giá trị các trọng số đạt được sau huấn luyện sẽ có sai số so với giá trị trọng số mong muốn và sẽ làm giảm độ chính xác khi nhận dạng của mạng nơ-ron khi thực thi. Trong một nghiên cứu vào năm 2019, J. Fu và các cộng sự đã thực thi một mạng nơ-ron có cấu trúc gồm 400 nơ-ron ngõ vào, 100 nơ-ron ở 1 lớp ẩn và 10 nơ-ron ngõ ra để nhận dạng 10 ký tự viết tay của tập dữ liệu MNIST [31]. Trong nghiên cứu này, để hạn chế sự ảnh hưởng của đặc tính phi tuyến của điện trở nhớ và tăng độ chính xác của các giá trị trọng số sau quá trình huấn luyện, nhóm tác giả đã đề xuất phương pháp Quy hoạch Tuyến tính (Linear Optimization – LO). Với phương pháp Quy hoạch Tuyến tính, đường đặc tính phi tuyến lý tưởng của giá trị điện dẫn của điện trở nhớ theo số lượng xung huấn luyện đã được tuyến tính hóa thành 2, 3 hoặc 4 đoạn tuyến tính. Khi đó, trong quá trình huấn luyện, giá trị điện dẫn của điện trở nhớ sẽ được cập nhật theo các đoạn tuyến tính này bằng các thay đổi độ rộng của chuỗi xung huấn luyện thay vì được cập nhật theo đường đặc tính phi tuyến lý tưởng. Việc này sẽ làm giảm độ sai lệch giữa giá trị trọng số đạt được sau huấn luyện so với giá trị trọng số mong muốn và qua đó làm tăng độ chính xác khi nhận diện của mạng nơ-ron. Với phương pháp LO này, J. Fu và các cộng sự đã chứng minh độ chính xác của mô hình mạng nơ-ron trên khi nhận diện 10 ký tự viết tay của tập dữ liệu MNIST đã tăng thêm trung bình 39,67 % khi thực hiện nhận dạng các ảnh đầu vào với các mức độ nhiễu khác nhau [31]. Tuy nhiên, để giảm hưởng của đặc tính phi tuyến của điện trở nhớ, việc thực thi mạng nơ-ron nhân tạo với phương pháp LO đòi hỏi khối lượng công việc khá lớn và phức tạp như: cần tìm đường đặc tuyến điện dẫn chuẩn hóa của điện trở nhớ sử dụng thông qua 1000 mẫu điện trở nhớ, tìm vị trí các điểm chia đoạn phù hợp trên đường đặc tuyến, tính toán độ rộng của các xung huấn luyện phù hợp với độ dốc của từng đoạn tuyến tính, cần

có bộ nhớ để lưu trữ thông tin của các đoạn tuyến tính, cần so sánh giá trị trọng số đạt được với giá trị tham chiếu sau mỗi lần cập nhật.

Thêm vào đó, 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 lớn cho việc thực thi các mạng nơ-ron nhân tạo trên các mảng điện trở nhớ tương tự. Thông thường, việc học và cập nhật trọng số trực tiếp trên mạch phần cứng khi thực thi các mạng nơ-ron trên các mảng điện trở nhớ tương tự đòi hỏi cần phải có các mạch học và giải thuật học phức tạp. Trong một nghiên cứu vào năm 2015 [37], D. Soudry và các cộng sự đã đề xuất mạch phần cứng để thực hiện phương pháp huấn luyện Giảm dần Độ dốc trực tuyến (Online Gradient Descent) dùng cho các mạng nơ-ron nhiều lớp thực thi trên các mảng điện trở nhớ. Thiết kế này cho phép các trọng số của các khớp nối thần kinh được cập nhật liên tục trên phần cứng và đạt độ chính xác tương đương với khi thực thi trên phần mềm. Tuy nhiên, thiết kế này đòi hỏi mỗi khớp nối thần kinh cần được thực thi bằng một điện trở nhớ và hai transistor CMOS (2 transistor – 1 memristor: 2T-1M) nên chưa tối ưu về diện tích thiết kế cũng như công suất tiêu thụ. Trong một nghiên cứu khác vào năm 2019 [32], Z. Wang cùng các cộng sự đã thực thi giải thuật học tăng cường (reinforcement learning) cho một mạng nơ-ron ba lớp sử dụng các mảng điện trở nhớ tương tự có các kích thước $n_i \times 48$, 96×48 , $48 \times n_o$, trong đó n_i và n_o là số lượng các nơ-ron ngõ vào và ngõ ra và mạng nơ-ron đã được thực nghiệm kiểm chứng hoạt động trong các trò chơi con lắc ngược và xe leo núi. Tuy nhiên, để thực thi giải thuật học tăng cường trên phần cứng, các mảng điện trở nhớ trong nghiên cứu này cần được thiết kế theo kiến trúc 1 transistor – 1 memristor (1T-1M) nên chưa tối ưu về mặt diện tích thiết kế. Ở một nghiên cứu khác [33], O. Krestinskaya và các cộng sự đã đề xuất mạch học lan truyền ngược trực tiếp trên phần cứng cho các mạng nơ-ron. Đồng thời, khi thực thi mạng nơ-ron sử dụng mảng điện trở nhớ tương tự, nhóm tác giả đã không dùng mô hình 1T-1M cho một khớp nối thần kinh mà thay vào đó chỉ sử dụng 1 điện trở nhớ (1M) cho một khớp nối thần kinh và triển khai mảng điện trở nhớ theo hướng module: chia một mảng điện trở lớn ra thành

nhiều module nhỏ hơn để giảm dòng rò trong mảng. Các tác giả đã xây dựng một mạng ANN có cấu trúc gồm 28×28 nơ-ron ngõ vào, 42 nơ-ron lớp ẩn, 10 nơ-ron ngõ ra theo hướng module và sử dụng phương pháp lan truyền ngược đã đề xuất để huấn luyện mạng nhằm thực thi nhận dạng 10 ký tự trong tập MNIST; kết quả độ chính xác khi nhận dạng đạt 92%. Tuy nhiên, để thực thi phương pháp lan truyền ngược trên phần cứng cho mạng nơ-ron, mạch huấn luyện được đề xuất của O. Krestinskaya và cộng sự còn khá phức tạp: cần 4 khối chính (main block – MB) cho mạch học để huấn luyện cho mạng ANN ở trên.

Sự biến thiên giá trị điện trở của các điện trở nhớ trong mảng cũng là một thách thức không nhỏ cho việc thực thi các mạng nơ-ron nhân tạo trên các mảng điện trở nhớ tương tự. Sự biến thiên giá trị điện trở trong mảng sẽ tạo ra sự sai lệch giữa giá trị thực tế khi hoạt động và giá trị mong muốn trong tính toán của các điện trở nhớ. Sự biến thiên điện trở của các điện trở nhớ trong mảng có thể bị gây ra bởi sự không đồng nhất trong quá trình chế tạo [38-41] hoặc sự trôi của giá trị điện trở [42, 43] trong quá trình hoạt động dưới ảnh hưởng của điện áp hoạt động đặt vào 2 cực của điện trở nhớ. Chính sự biến thiên điện trở này sẽ làm giảm đáng kể hiệu suất hoạt động của các ứng dụng sử dụng mảng điện trở nhớ. Trong một nghiên cứu vào năm 2014, nhóm tác giả đã xây dựng ứng dụng nhận dạng giọng nói dùng mảng điện trở nhớ để nhận dạng 5 nguyên âm: ‘a’, ‘i’, ‘e’, ‘u’, ‘o’ [44]. Trong nghiên cứu này, nhóm tác giả đã chứng minh được rằng sự biến thiên điện trở đã làm suy giảm độ chính xác khi nhận dạng của ứng dụng và sự suy giảm trong trường hợp sử dụng mảng điện trở nhớ tương tự sẽ lớn hơn nhiều khi so với khi sử dụng mảng điện trở nhớ nhị phân. Cụ thể, khi độ biến thiên điện trở của mảng là 15% thì độ chính xác khi nhận dạng của ứng dụng đã giảm xuống còn chỉ 9% khi sử dụng mảng điện trở nhớ tương tự nhưng chỉ giảm xuống còn 89,2% khi sử dụng mảng điện trở nhớ nhị phân. Trong một nghiên cứu khác vào năm 2019, một mạng nơ-ron 2 lớp có 196 nơ-ron ngõ vào, 256 nơ-ron ở lớp ẩn và 10 nơ-ron ngõ ra sử dụng mảng điện trở nhớ tương tự để nhận dạng 10 ký tự của tập dữ liệu MNIST [45] đã xây dựng. Trong nghiên cứu này, khi thực thi một khớp nối thần kinh, tác giả đã sử dụng 1 memristor và một phần

tử cố định (constant-term circuit). Tác giả đã khảo sát mức độ ảnh hưởng của sự biến thiên điện trở của mảng điện trở nhớ tới hiệu suất nhận dạng của ứng dụng và cho thấy khi độ biến thiên điện trở của mảng điện trở nhớ tăng từ 0% lên 10% thì độ chính xác của ứng dụng đã giảm từ 97% xuống chỉ còn 17% khi phần tử cố định có giá trị R_B là 30 K Ω .

Việc xây dựng các hệ điện toán mô phỏng hệ thần kinh bằng cách thực thi các mạng nơ-ron nhân tạo trên các mảng điện trở nhớ tương tự vẫn là một hướng tiếp cận khá khó khăn và còn nhiều thách thức. Ngược lại, ứng dụng mảng điện trở nhớ với giá các nhị phân mang lại nhiều lợi ích và tiềm năng trong các hệ toán mô phỏng hệ thần kinh.

3.3 PHÂN TÍCH ẢNH HƯỞNG CỦA NHIỀU 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

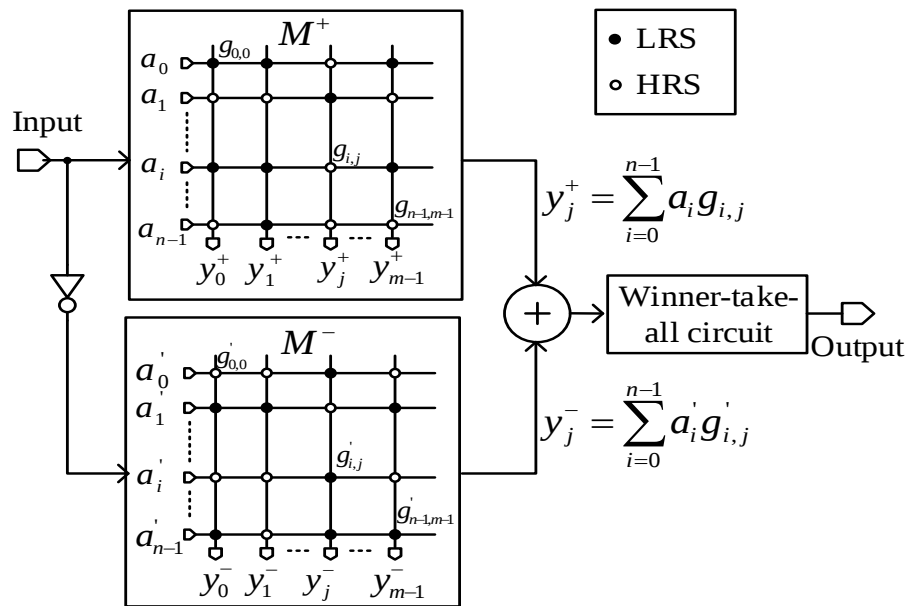
Việc thực thi các hệ điện toán mô phỏng hệ thần kinh sử dụng các điện trở nhớ nhị phân [44, 46-48] đã được nghiên cứu và đã cho thấy nhiều ưu điểm nổi bật hơn so với việc sử dụng điện trở nhớ tương tự. Với yêu cầu lưu trữ hai mức điện trở, điện trở thấp (LRS) và điện trở cao (HRS), các điện trở nhớ nhị phân dễ dàng được chế tạo và lập trình hơn so với điện trở nhớ tương tự. Thêm vào đó, sự biến thiên giá trị do quá trình chế tạo và hoạt động của điện trở nhớ nhị phân cũng ít ảnh hưởng đến hiệu suất hoạt động của hệ thống hơn so với điện trở nhớ tương tự [44] nhờ tỷ lệ HRS/LRS (R_{OFF} / R_{ON}) khá cao, có thể đạt giá trị 100 hoặc cao hơn [25, 49]. Đồng thời, phép nhân giữa các tín hiệu ngõ vào với các trọng số nhị phân cũng sẽ dễ dàng được thực thi hơn bằng các phép toán logic như AND, XNOR.

Các mảng điện trở nhớ nhị phân đã mở ra một hướng tiếp cận mới, tiềm năng và khả thi hơn cho việc xây dựng các hệ điện toán mô phỏng hệ thần kinh, đặc biệt là ở các bài toán nhận dạng mẫu như nhận dạng tiếng nói, nhận dạng ảnh. Để xây dựng các hệ điện toán mô phỏng hệ thần kinh trên các mảng điện trở

nhớ nhị phân nhằm thực thi các ứng dụng nhận dạng mẫu, 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]. Ở phần này, người nghiên cứu sẽ tiến hành phân tích chi tiết các phương trình toán học của các kiến trúc mảng điện trở nhớ nhị phân: kiến trúc mảng điện trở nhớ bù, kiến trúc mảng điện trở nhớ đồng nhất, kiến trúc mảng điện trở nhớ đơn trong thực thi ứng dụng nhận dạng ảnh. Các phân tích này sẽ là cơ sở toán học cho việc xây dựng ứng dụng thực thi nhận dạng ảnh với các kiến trúc mảng điện trở nhớ trên, cũng như phân tích các yếu tố ảnh hưởng đến hiệu suất hoạt động của hệ thống.

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 mẫu được cấu thành từ 2 mảng điện trở nhớ bù nhau và có mô hình toán học như thể hiện ở Hình 3.1 [44]. Trong kiến trúc này, M^+ là mảng điện trở nhớ thứ nhất và M^- là mảng điện trở nhớ thứ 2 chứa các phần tử điện trở nhớ có giá trị điện trở bù với giá trị của các phần tử tương ứng trong mảng M^+ .



Hình 3.1: Mô hình toán học của kiến trúc mảng điện trở nhớ bù.

Trong Hình 3.1, $M +$ là mảng điện trở nhớ có kích thước $n \times m$. Tại điểm giao nhau của hàng thứ i và cột thứ j là một điện trở nhớ có giá trị điện dẫn là g_{ij} tạo thành giá trị trở kháng nhớ, có thể ở trạng thái điện trở thấp (LRS) hoặc điện trở cao (HRS). Ở Hình 3.1, giá trị điện trở thấp được ký hiệu bằng khối tròn màu đen và giá trị điện trở cao được ký hiệu bằng vòng tròn rỗng. Mảng $M -$ chứa các phần tử điện trở nhớ có giá trị bù với phần tử ở vị trí tương ứng trong mảng $M +$. Cụ thể, giá trị điện dẫn g'_{ij} của điện trở nhớ tại giao điểm của hàng thứ i và cột thứ j trong mảng $M -$ là giá trị bù của giá trị g_{ij} trong mảng $M +$. Các mảng $M +$ và $M -$ có thể được biểu diễn dưới dạng ma trận như sau:

$$M+ = \begin{bmatrix} g_{0,0} & g_{0,1} & \cdots & g_{0,(m-1)} \\ g_{1,0} & g_{1,1} & \cdots & g_{1,(m-1)} \\ \vdots & \vdots & \vdots & \vdots \\ g_{(n-1),0} & g_{(n-1),1} & \cdots & g_{(n-1),(m-1)} \end{bmatrix} \quad (3.1)$$

$$M- = \begin{bmatrix} g'_{0,0} & g'_{0,1} & \cdots & g'_{0,(m-1)} \\ g'_{1,0} & g'_{1,1} & \cdots & g'_{1,(m-1)} \\ \vdots & \vdots & \vdots & \vdots \\ g'_{(n-1),0} & g'_{(n-1),1} & \cdots & g'_{(n-1),(m-1)} \end{bmatrix}$$

Để thực hiện nhận dạng với kiến trúc mảng điện trở nhớ bù này, mẫu cần nhận dạng sẽ cần được chuyển đổi thành các vector đầu vào có kích thước $1 \times n$. Ở Hình 3.1, một vector đầu vào $A = [a_0 \ a_1 \ \dots \ a_{n-1}]$ được đưa vào mảng $M +$; đồng thời vector bù của vector A là vector $A' = [a'_0 \ a'_1 \ \dots \ a'_{n-1}]$ sẽ được đưa vào mảng $M -$ để thực hiện việc nhận dạng. Khi đó, mảng M sẽ thực thi phép toán XNOR với vector đầu vào A theo công thức ở phương trình (2.6).

Trong đó, mảng $M +$ đã lưu trữ trước giá trị mẫu của tất cả các vector đầu vào cần nhận dạng. Mẫu nhận dạng của vector đầu vào thứ j , hay ảnh đầu vào thứ j , được lưu ở cột thứ j của mảng $M +$. Tất cả các phần tử của mảng $M -$ sẽ là giá trị bù của các phần tử ở vị trí tương ứng trong mảng $M +$. Hàm XNOR được ứng dụng để tìm mức độ giống nhau giữa vector đầu vào và các mẫu đã được lưu trữ trước bên trong các mảng điện trở nhớ. Vector ngõ ra $Y =$

$[y_0 \ y_1 \ \dots \ y_{m-1}]$, sẽ chứa kết quả là mức độ giống nhau giữa vector đầu vào A và dữ liệu được lưu trữ trước ở các cột của các mảng [44].

Từ phương trình (3.1) và phương trình (2.6), ngõ ra Y của kiến trúc mảng điện trở nhớ bù sẽ được tính toán theo công thức sau:

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

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

Trong công thức (3.2), a_i là điện áp ngõ vào thể hiện cho bit dữ liệu thứ i đầu vào: giá trị của bit dữ liệu là 0 hoặc 1 sẽ được mã hóa thành các mức điện áp thấp hoặc điện áp cao tương ứng, $g_{i,j}$ là giá trị điện dẫn của điện trở nhớ tại giao điểm của hàng thứ i và cột thứ j . Do đó, y_j là dòng điện ngõ ra ở cột thứ j và thể hiện mức độ giống nhau giữa mẫu dữ liệu ngõ vào cần nhận dạng A và mẫu dữ liệu đã được lưu trữ trước ở cột thứ j của các mảng điện trở nhớ. Như vậy, nếu mẫu dữ liệu ngõ vào, được thể hiện bởi vector $A = [a_0 \ a_1 \ \dots \ a_{n-1}]$ trùng khớp với mẫu dữ liệu đã được lưu trữ trước ở cột thứ j của mảng thì dòng điện ngõ ra ở cột thứ j (tức y_j) sẽ có giá trị lớn nhất trong tất cả các dòng điện ngõ ra của các cột. Sau đó, các dòng điện ngõ ra cột, từ y_0 đến y_{m-1} , sẽ được so sánh với nhau bởi mạch Tìm dòng điện lớn nhất (Winner-take-all circuit). Mạch Tìm dòng điện lớn nhất sẽ tìm ra dòng điện y_j là dòng điện ngõ ra lớn nhất và chỉ ra

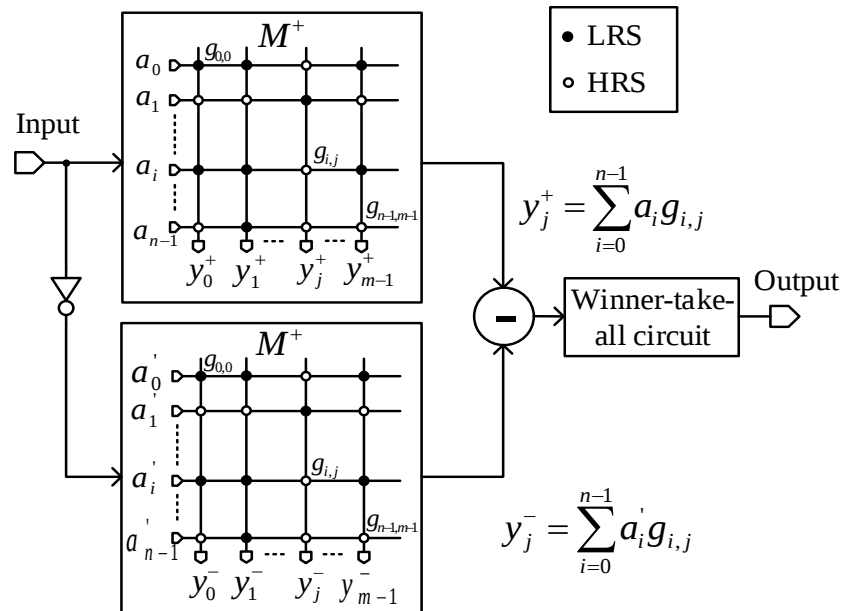
được ảnh dữ liệu ngõ vào A đã trùng khớp với ảnh dữ liệu đã lưu trữ trước ở cột thứ j trong các mảng điện trở nhớ.

Khác với kiến trúc mảng điện trở nhớ bù, kiến trúc mảng điện trở nhớ đồng nhất được cấu thành từ 2 mảng điện trở nhớ $M +$ giống nhau và đã được chứng minh là có cùng chức năng với kiến trúc mảng điện trở nhớ bù [47]. Trong kiến trúc mảng điện trở đồng nhất, hàm XNOR đã được triển khai như ở phương trình (2.7). Trong đó, phần tử A' là một đại lượng hằng số và không có sự tương tác với mảng M . Do đó, A' có thể được lược bỏ khi thực hiện phép XNOR giữa vector ngõ vào và mảng điện trở nhớ. Lúc này, phương trình XNOR của kiến trúc mảng điện trở đồng nhất đã được biến đổi tương đương thành [47]:

$$\begin{aligned}
 Y &= \overline{A \oplus M} = A \cdot (M +) - A' \cdot (M +) \\
 &= [y_0 \quad y_1 \quad \cdots \quad y_{m-1}]
 \end{aligned}
 \tag{3.3}$$

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

Từ phương trình (3.3), kiến trúc mảng điện trở nhớ đồng nhất sẽ có mô hình toán học như thể hiện ở Hình 3.2 [47].



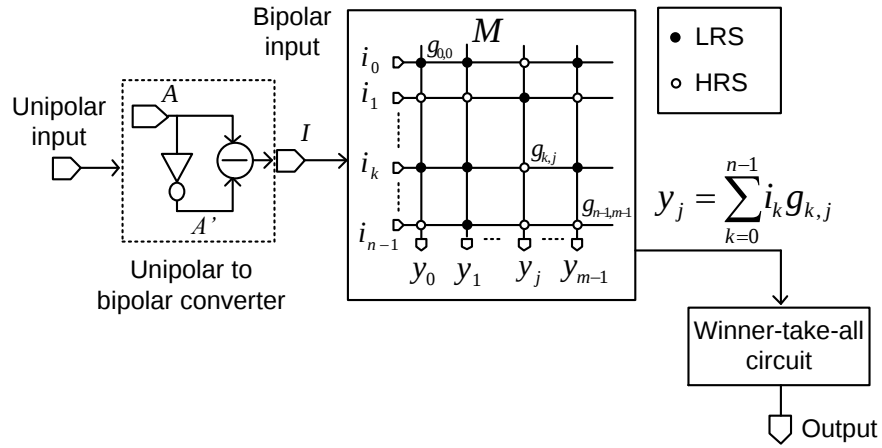
Hình 3.2: Mô hình toán học của kiến trúc mảng điện trở nhớ đồng nhất.

Từ phương trình (3.3) và Hình (3.2), có thể thấy kiến trúc mảng điện trở nhớ đồng nhất sử dụng 2 mảng điện trở nhớ giống nhau, 2 mảng M , để lưu trữ các mẫu dữ liệu là các ảnh cần nhận dạng [47] thay vì 2 mảng điện trở nhớ bù nhau như ở kiến trúc mảng điện trở nhớ bù. Vector ngõ ra Y sau đó cũng được đưa vào mạch Tìm dòng điện lớn nhất (Winner-take-all circuit) để xác định dòng điện ngõ ra cột lớn nhất y_j để chỉ ra rằng vector đầu vào A trùng khớp với mẫu dữ liệu đã được lưu trữ trước ở cột thứ j của các mảng điện trở nhớ.

Kiến trúc mảng điện trở nhớ đơn đã được đề xuất bằng cách rút gọn hàm XNOR ở phương trình (2.7) và được chứng minh là có thể xác định được mức độ giống nhau giữa vector đầu vào A và mảng M [48]. Cụ thể, trong kiến trúc mảng điện trở nhớ đơn, hàm XNOR ở phương trình (2.7) đã được khai triển chi tiết như ở phương trình (2.9) và giản lược thành phần hằng số như ở phương trình (2.10) [48]. Ở phương trình (2.10), $I = (A - A') = [i_0 \ i_1 \ \dots \ i_{(n-1)}]$ là vector đầu vào với kích thước $1 \times n$ chứa dữ liệu lưỡng cực. Ví dụ như, nếu the vector A có giá trị là $A = [1 \ 0 \ 1]$, thì vector I sẽ mang giá trị là $I = [1 \ -1 \ 1]$. Như vậy, phương trình (2.10) có thể được viết chi tiết ở dạng ma trận như sau:

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

Phương trình (3.4) cho thấy kiến trúc mảng điện trở nhớ đơn chỉ sử dụng 1 mảng điện trở nhớ duy nhất, mảng M , để thực hiện phép XNOR với dữ liệu đầu vào lưỡng cực I trong các ứng dụng nhận dạng mẫu. Do đó, kiến trúc mảng điện trở nhớ đơn sẽ có mô hình toán học như thể hiện ở Hình 3.3.



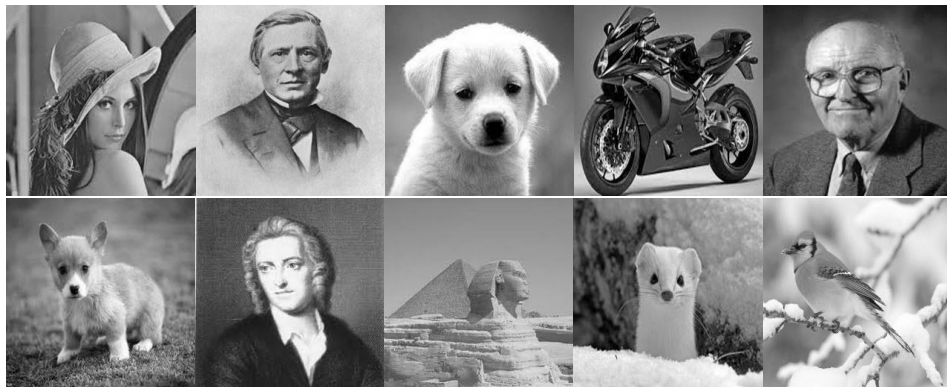
Hình 3.3: Mô hình toán học của kiến trúc mảng điện trở nhớ đơn.

Ở kiến trúc mảng điện trở nhớ đơn, trước hết, vector đầu vào đơn cực A sẽ được đưa vào một bộ chuyển đổi dữ liệu từ đơn cực sang lưỡng cực (Unipolar to bipolar converter) để tạo ra vector đầu vào lưỡng cực I . Vector lưỡng cực I sau đó sẽ được đưa vào 1 mảng điện trở nhớ duy nhất, M , để thực hiện hàm XNOR như phương trình (3.4) và thu được vector ngõ ra Y . Vector ngõ ra Y chứa các dòng điện ngõ ra cột và thể hiện mức độ đồng nhất giữa vector đầu vào A với các mẫu dữ liệu đã được lưu trữ trước ở các cột của mảng điện trở nhớ duy nhất M . Vector ngõ ra Y sau đó được đưa vào mạch mạch Tìm dòng điện lớn nhất (Winner-take-all circuit) để xác định dòng điện ngõ ra cột lớn nhất y_j , tức là chỉ ra rằng vector đầu vào A trùng khớp với mẫu dữ liệu được lưu trữ trước ở cột thứ j của mảng điện trở nhớ M .

3.3.2 Thực thi nhận dạng ảnh với các kiến trúc mảng điện trở nhớ nhị phân khác nhau

Để phân tích và so sánh hiệu suất của các kiến trúc mảng điện trở nhớ nhị phân trong các ứng dụng nhận dạng ảnh trong các hệ điện toán mô phỏng hệ thần kinh, người nghiên cứu đã xây dựng một ứng dụng nhận dạng 10 ảnh mẫu sử dụng 3 kiến trúc mảng điện trở nhớ đã phân tích ở phần trước: kiến trúc mảng điện trở nhớ bù, kiến trúc mảng điện trở nhớ đồng nhất và kiến trúc mảng điện trở nhớ đơn [81]. Trong ứng dụng này, 10 ảnh xám với kích thước 32×32 đã được sử dụng làm ảnh đầu vào để nhận dạng với các kiến trúc mảng điện trở nhớ

nhị phân. Ở đây, dữ liệu ảnh xám sẽ được chọn tương tự như trong các nghiên cứu đã công bố trước đó của các kiến trúc mạng điện trở nhớ để bảo đảm sự đồng nhất cũng như tính khách quan trong việc đánh giá. Khi thực hiện nhận dạng, các kiến trúc mạng điện trở nhớ sẽ thực thi phép toán XNOR giữa vector đầu vào và các ảnh dữ liệu để tính toán mức độ trùng khớp giữa ảnh cần nhận dạng và từng ảnh trong tập dữ liệu, như đã phân tích ở phần trước. Do đó, tất cả các ảnh dữ liệu sẽ được lưu trữ trực tiếp vào các mảng điện trở nhớ, để thực thi phép XNOR giữa ảnh dữ liệu và vector đầu vào, mà không phải là các ma trận trọng số sau quá trình học phức tạp như trong các ứng dụng học sâu. Cụ thể, mỗi ảnh sẽ được lưu trữ trên một cột của mảng điện trở nhớ, bit 1 tương ứng với giá trị điện trở thấp (LRS) và bit 0 tương ứng với giá trị điện trở cao (HSR). Việc đánh giá sẽ tập trung vào tính kiến trúc của hệ thống mà không phải là quá trình huấn luyện và học như trong các ứng dụng học sâu. Do đó, dữ liệu được sử dụng sẽ là 10 ảnh nhị phân đơn giản, thay vì tập dữ liệu lớn như MNIST với sự thay đổi phức tạp giữa các ảnh. Một cách tổng quát, việc thực thi nhận dạng trên các tập dữ liệu khác với các kiến trúc mạng điện trở nhớ nhị phân này cũng sẽ được thực hiện theo nguyên tắc mạng XNOR tương tự như ở tập dữ liệu đang kiểm chứng. Cụ thể, các ảnh dữ liệu sẽ được lưu trữ trực tiếp vào các mảng điện trở nhớ để tính toán mức độ giống nhau giữa vector đầu vào và dữ liệu trên từng cột bằng phép toán XNOR. Cột có dòng điện ngõ ra lớn nhất sẽ là cột có dữ liệu trùng khớp nhất với vector đầu vào cần nhận dạng. Các ảnh dữ liệu được thể hiện chi tiết ở Hình 3.4.



Hình 3.4: Mười ảnh xám được sử dụng để nhận dạng.

Trước hết, mỗi ảnh xám sẽ được chuyển đổi từ định dạng ma trận 32×32 điểm ảnh sang định dạng vector 1×1024 điểm ảnh. Sau đó, mỗi điểm ảnh xám sẽ được số hóa 4 bit [44, 47]. Mỗi 4 bit của điểm ảnh thứ i , $a_i < 0:3 >$, của một ảnh được lưu trữ ở 4 điện trở nhớ tại giao điểm của hàng thứ i với 4 cột khác nhau. Các cột này sẽ có hệ số nhân lần lượt là 8, 4, 2, 1 để thực hiện tính toán dòng điện ngõ ra cột. Tất cả 10 ảnh này được lưu trữ vào các mảng điện trở nhớ ($M +$ và $M -$, $M +$ và $M +$, hoặc M) theo 10 nhóm khác nhau, mỗi nhóm sẽ có 4 cột để tạo thành các mẫu dữ liệu lưu trữ trước ở các mảng điện trở nhớ dành cho việc nhận dạng. 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 1 ảnh đầu vào, trước tiên ảnh này sẽ được chuyển đổi thành vector A với kích thước 1×1024 điểm ảnh. Mỗi điểm ảnh trong vector A 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. Ở kiến trúc mảng điện trở nhớ bù, vector đầu vào A được đưa vào mảng $M +$ và vector bù A' của vector A được đưa vào mảng $M -$ để thực hiện phép XNOR như ở phương trình (3.2). Ở kiến trúc mảng điện trở nhớ đồng nhất, vector đầu vào A được đưa vào mảng $M +$ thứ nhất và vector bù A' được đưa vào mảng $M +$ thứ 2 để thực thi hàm XNOR như đã trình bày ở phương trình (3.3). Với kiến trúc mảng điện trở nhớ đơn, vector đầu vào A được lưỡng cực hóa bởi bộ chuyển đổi dữ liệu từ đơn cực sang lưỡng cực (Unipolar to bipolar converter) để tạo thành vector dữ liệu lưỡng cực I . Vector dữ liệu lưỡng cực I này sau đó sẽ được đưa vào mảng M duy nhất của kiến trúc này để thực hiện phép XNOR như đã phân tích trước đó ở phương trình (3.4). Ở từng kiến trúc, các dòng điện ngõ ra cột của mỗi bit trong nhóm 4 bit $< 0:3 >$ của từng điểm ảnh sẽ được nhân với các hệ số nhân 1, hoặc 2, hoặc 4, hoặc 8 trước khi đi vào khối tính tổng. Dòng điện ngõ ra thứ k , I_k , thể hiện mức độ giống nhau giữa vector đầu vào A và mẫu dữ liệu thứ k đã được lưu trữ trước trong các mảng điện trở nhớ.

A trùng khớp với mẫu dữ liệu thứ k , tức ảnh thứ k , đã được lưu trữ bên trong các mảng điện trở nhớ.

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 khi thực thi nhận dạng ảnh

Tiếp theo, người nghiên cứu sẽ tiến hành khảo sát, phân tích và đánh giá mức độ ảnh hưởng của các yếu tố: nhiễu trên dữ liệu, độ biến thiên điện trở trong mảng đến hiệu suất hoạt động của các kiến trúc mảng điện trở nhớ nhị phân khi thực thi ứng dụng nhận dạng ảnh. Các điều kiện mô phỏng khi thực thi nhận dạng ảnh với các kiến trúc mảng điện trở nhớ nhị phân dưới sự tác động của nhiễu dữ liệu và sự biến thiên điện trở được trình bày trong Bảng 3.1

Bảng 3.1: Thông số mô phỏng khi thực thi nhận dạng ảnh với các kiến trúc mảng điện trở nhớ nhị phân dưới sự tác động của nhiễu dữ liệu và sự biến thiên điện trở.

<i>Phương pháp mô phỏng</i>	<i>Nhiều trên dữ liệu (SNR)</i>	<i>Độ biến thiên điện trở (%)</i>	<i>Mảng điện trở nhớ</i>	<i>Điện áp vào đơn cực</i>	<i>Điện áp vào lưỡng cực</i>
Monte Carlo (2000 lần lặp)	Từ -10 dB đến 4 dB	Từ 0% đến 40%	LRS: 10 K Ω HRS: 1 M Ω	1 V 0 V	+1 V -1 V

Ba kiến trúc mảng điện trở nhớ nhị phân gồm kiến trúc mảng điện trở nhớ bù, kiến trúc mảng điện trở nhớ đồng nhất, kiến trúc mảng điện trở nhớ đơn sẽ được đánh giá bằng phương pháp mô phỏng Monte Carlo khi thực thi nhận dạng 10 ảnh xám với các mức độ nhiễu trên dữ liệu và biến thiên điện trở khác nhau. Tương ứng với mỗi giá trị của mức độ nhiễu dữ liệu và tương ứng với mỗi mức độ biến thiên điện trở (được phân phối theo phân bố Gauss), quá trình nhận dạng sẽ được lặp lại trong 2000 lần để thống kê tỉ lệ nhận dạng đúng trung bình của

mỗi kiến trúc nhằm đảm bảo tính khách quan và tổng quát. Các thông số điện áp vào đơn cực có giá trị là 0 V và 1 V, điện áp vào lưỡng cực có giá trị là -1 V và +1 V. Giá trị LRS và HRS của các điện trở nhớ được giả định là 10 K Ω và 1 M Ω .

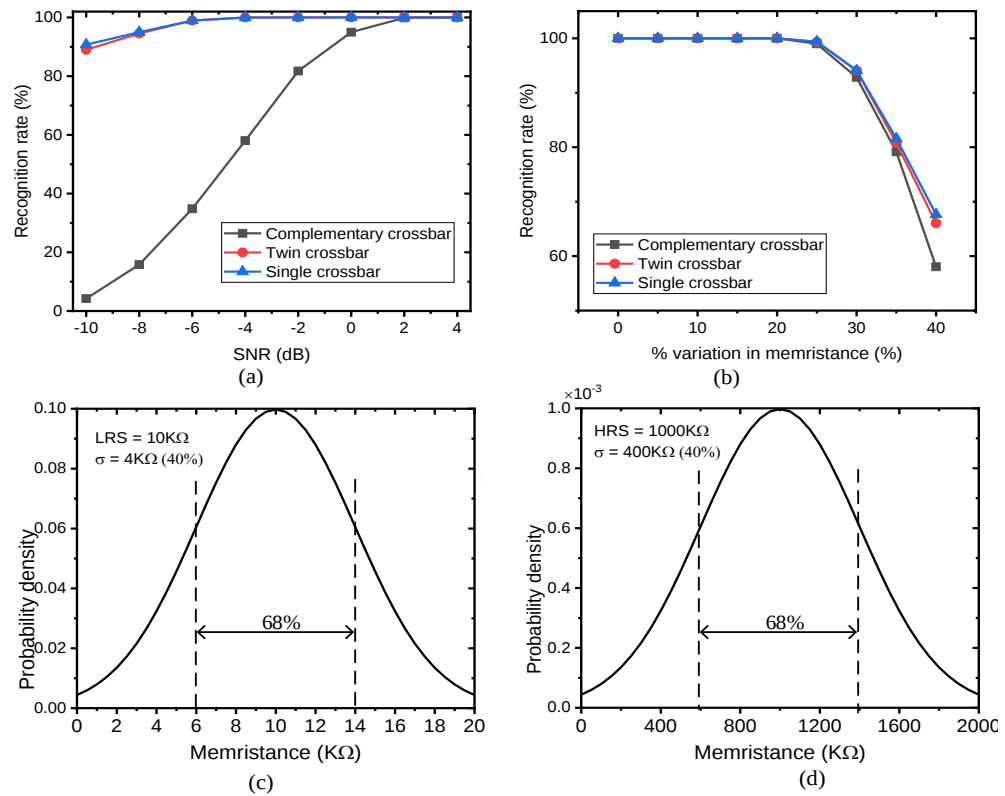
Để đánh giá sự ảnh hưởng của nhiễu trên dữ liệu đến hiệu suất nhận dạng của 3 kiến trúc mạng điện trở nhớ trong ứng dụng nhận dạng ảnh, các ảnh đầu vào sẽ được cộng thêm nhiễu Gauss trước khi đưa vào nhận dạng. Tỉ số tín hiệu trên nhiễu (SNR) của nhiễu Gauss cộng vào các ảnh được thay đổi từ -10 dB đến 4 dB. 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. Các ảnh gốc trước đó đã được lưu trữ trong các mảng điện trở nhớ của cả 3 kiến trúc, như đã phân tích và trình bày ở phần trước.



Hình 3.6: 10 ảnh xám được sử dụng để nhận dạng với nhiễu Gauss có SNR là -10 dB.

Các ảnh đầu vào có nhiễu sau đó được số hóa 4 bit và lần lượt được nhận dạng với 3 kiến trúc mạng điện trở nhớ bằng cách thực thi mảng XNOR giữa ảnh đầu vào và dữ liệu đã lưu trữ trước đó trên các cột của các mảng điện trở nhớ. Kết quả tỉ lệ nhận dạng đúng của 3 kiến trúc trên các ảnh đầu vào với nhiễu Gauss có tỉ số SNR thay đổi từ -10 dB đến 4 dB được so sánh ở Hình 3.7(a). Trong đó, ở mỗi lượt nhận dạng, tỉ lệ nhận dạng sẽ được tính bằng số ảnh nhận dạng đúng trên tổng số 10 ảnh của tập ảnh ở Hình 3.4. Quá trình nhận dạng được thực hiện theo phương pháp mô phỏng Monte Carlo với 2000 lần lặp và tỉ lệ nhận dạng đúng của mỗi kiến trúc tại mỗi giá trị SNR sẽ là tỉ lệ nhận dạng trung bình sau 2000 lần lặp. Ở đây, các kiến trúc mạng điện trở nhớ (kiến trúc mảng

bù, kiến trúc mảng đồng nhất, kiến trúc mảng đơn) thực thi nhận dạng bằng cách tìm mức độ giống nhau thông qua phép XNOR giữa vector đầu vào và tập dữ liệu, vốn đã lưu trữ trực tiếp vào các mảng điện trở nhớ. Do đó, tập dữ liệu được chọn là 10 ảnh cố định, như đề cập ở Hình 3.4. Đây là một điểm khác biệt giữa các kiến trúc mảng điện trở nhớ này với các kiến trúc mạng học sâu, vốn cần được huấn luyện và đòi hỏi tập dữ liệu có tính biến thiên để đảm bảo được sự khái quát hóa và tính thích nghi của các mạng học sâu.



Hình 3.7: (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.

Kết quả phân tích ở Hình 3.7(a) cho thấy rằng kiến trúc mảng điện trở nhớ đơn và kiến trúc mảng điện trở nhớ đồng nhất vẫn duy trì được tỷ lệ nhận dạng đúng khá cao, từ 89% trở lên, trong khi đó tỷ lệ nhận dạng đúng của kiến trúc mảng điện trở nhớ bù giảm rõ rệt khi tỉ số SNR là -10 dB. Nguyên nhân của việc

này là do kiến trúc mảng điện trở nhớ bù sử dụng phép toán cộng khi thực hiện phép XNOR, cũng có nghĩa là dòng điện ngõ ra cột là tổng của dòng điện ngõ ra cột của mảng $M +$ và dòng điện ngõ ra cột của mảng $M -$, như đã trình bày ở phương trình (3.2). Do đó, sự thay đổi trong dòng điện ngõ ra cột gây ra bởi nhiễu trên ảnh đầu vào sẽ tăng lên. Trái lại, kiến trúc mảng điện trở nhớ đồng nhất sử dụng phép toán trừ khi thực thi hàm XNOR, như được trình bày 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 mảng điện trở nhớ đơn được phát triển từ kiến trúc mảng điện trở đồng nhất, như đã trình bày 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. Do vậy, kiến trúc mảng điện trở nhớ đơn cho tỉ lệ nhận dạng đúng có phần tốt hơn so với kiến trúc mảng điện trở nhớ đồng nhất dưới sự thay đổi của tỉ số SNR. Cụ thể, khi tỉ số SNR là -10 dB thì tỉ lệ nhận dạng đúng của kiến trúc mảng điện trở nhớ đơn, kiến trúc mảng điện trở nhớ đồng nhất và kiến trúc mảng điện trở nhớ bù lần lượt là 91%, 89% và 4%. Như vậy, kiến trúc mảng điện trở nhớ đơn có sự kháng nhiễu trên dữ liệu tốt nhất và duy trì được tỉ lệ nhận dạng cao nhất dưới sự ảnh hưởng của nhiễu trên dữ liệu vào khi so sánh với kiến trúc mảng điện trở nhớ bù và kiến trúc mảng điện trở nhớ đồng nhất.

Sự biến thiên của giá trị điện trở nhớ là một trong số các vấn đề gây ra sự sụt giảm trong hiệu quả của các ứng dụng dựa trên mảng điện trở nhớ [38-41, 82]. Trong ứng dụng nhận dạng ảnh này, hiệu suất của 3 kiến trúc mảng điện trở nhớ khi có sự biến thiên trong giá trị điện trở nhớ cũng đã được kiểm chứng. Cụ thể, tỉ lệ biến thiên giá trị điện trở nhớ đã được thay đổi từ 0% đến 40% để so sánh tỉ lệ nhận dạng đúng của 3 kiến trúc mảng điện trở nhớ. Ở đây, độ biến thiên các giá trị điện trở nhớ được thay đổi 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 . Như được thể hiện ở Hình 3.7(c), với giá trị LRS, tỉ lệ biến thiên 40% đồng nghĩa với giá trị điện trở nhớ thay đổi từ 6 k Ω ($\mu - \sigma$) đến 14 k Ω ($\mu + \sigma$) với xác suất 68%. Tương tự, với giá trị HRS, tỉ lệ biến thiên điện trở 40% đồng nghĩa với giá trị điện trở

nhớ thay đổi từ 600 k Ω đến 1400 k Ω với xác suất 68%, như minh họa ở Hình 3.7(d).

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 như đã trình bày ở phương trình (3.3). Vì thế, kiến trúc này có thể triệt tiêu một phần sự thay đổi trong dòng điện ngõ ra cột gây ra bởi sự biến thiên trong giá trị điện trở nhớ. Đây là nguyên nhân giúp kiến trúc mảng điện trở nhớ đồng nhất có tỉ lệ nhận dạng đúng tốt hơn so với kiến trúc mảng điện trở nhớ bù khi tỉ lệ biến thiên giá trị điện trở nhớ lên đến 40%. Khi tỉ lệ biến thiên giá trị điện trở nhớ là 40%, kiến trúc mảng điện trở nhớ đơn có tỉ lệ nhận dạng đúng là 67,8%, đạt tỉ lệ cao nhất khi so sánh với kiến trúc mảng điện trở nhớ bù và kiến trúc mảng điện trở nhớ đồng nhất có tỉ lệ nhận dạng đúng lần lượt là 58% và 66%. Khi tỉ lệ biến thiên giá trị điện trở nhớ lớn hơn 40%, cả ba kiến trúc mảng đều sẽ cho tỉ lệ nhận dạng đúng rất thấp, như có thể suy ra từ Hình 3.7(b). Hơn nữa, các dòng điện ngõ ra cột phụ thuộc chủ yếu vào các tín hiệu đi vào các điện trở nhớ mang giá trị điện trở thấp (LRS) hơn là các điện trở nhớ mang giá trị điện trở cao (HRS). Vì vậy, độ biến thiên giá trị xảy ra trên các điện trở nhớ mang giá trị điện trở thấp (LRS) sẽ ảnh hưởng tới tỉ lệ nhận dạng đúng nhiều hơn so với độ biến thiên xảy ra trên các điện trở nhớ mang giá trị điện trở cao (HRS).

Các kết quả mô phỏng trên cho thấy kiến trúc mảng điện trở nhớ đơn với đầu vào lưỡng cực cho kết quả tỉ lệ nhận dạng đúng cao nhất khi so sánh với kiến trúc mảng điện trở nhớ bù và kiến trúc mảng điện trở nhớ đồng nhất dưới sự ảnh hưởng của nhiễu trên ảnh dữ liệu đầu vào và dưới sự biến thiên trong giá trị điện trở của mảng. Cụ thể, với các ảnh đầu vào có nhiễu với tỉ số SNR -10 dB thì kiến trúc mảng điện trở nhớ đơn cho tỉ lệ nhận dạng đúng cao hơn lần lượt 87% và 2% khi so sánh với kiến trúc mảng điện trở nhớ bù và kiến trúc mảng điện trở nhớ đồng nhất. Khi độ biến thiên của giá trị điện trở nhớ ở mức 40% thì tỉ lệ nhận diện đúng của kiến trúc mảng điện trở nhớ đơn vẫn duy trì ở mức 67,8%, cao hơn 9,8% và 1,8% so với kiến trúc mảng điện trở nhớ bù và kiến trúc mảng điện trở nhớ đồng nhất. Như vậy, bên cạnh ưu điểm về kích thước mảng và công suất

tiêu thụ trên mảng, kiến trúc mảng điện trở nhớ đơn với đầu vào lưỡng cực còn có sự đáp ứng tốt hơn với nhiễu trên ảnh đầu vào và với sự biến thiên giá trị điện trở nhớ của mảng, khi so sánh với kiến trúc mảng điện trở nhớ bù và kiến trúc mảng điện trở nhớ đồng nhất.

3.4 KẾT LUẬN CHƯƠNG

Ở chương này, luận án đã phân tích được sự ảnh hưởng của các yếu tố, bao gồm nhiễu trên dữ liệu và sự biến thiên điện trở nhớ trong mảng, đến hiệu suất nhận dạng của các kiến trúc mảng điện trở nhớ nhị phân trong bài toán nhận dạng ảnh. Cụ thể hơn, chương này của 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 kiến trúc mảng điện trở nhớ đồng nhất. Các kết quả phân tích và phát hiện ở chương này cho thấy 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 nhiệm vụ nghiên cứu quan trọng và tiềm năng trong thiết kế các hệ điện toán mô phỏng hệ thần kinh thực thi nhận dạng ảnh.

Ở các phần tiếp theo của luận án, tác giả sẽ tiến hành phân tích sâu hơn ảnh hưởng của yếu tố mật độ dữ liệu đến hoạt động nhận dạng của các kiến trúc mảng điện trở nhớ, tìm ra nguyên nhân gây ra sự ảnh hưởng và từ đó thực hiện tối ưu kiến trúc mảng điện trở nhớ nhị phân. Việc nghiên cứu và đề xuất một thiết kế tối ưu cho kiến trúc mảng điện trở nhớ nhị phân sẽ là tiền đề vững chắc cho các phát triển tiếp theo trong việc thiết kế các hệ điện toán mô phỏng hệ thần kinh thực thi các ứng dụng nhận dạng.

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

Ở chương này, Luận án sẽ tiếp tục phân tích ảnh hưởng của mật độ dữ liệu đến hoạt động của kiến trúc mảng điện trở nhớ đơn, trong sự so sánh với kiến trúc mảng điện trở nhớ bù. Nội dung của chương sẽ bao gồm việc thiết kế các mạch nguyên lý của kiến trúc mảng điện trở nhớ đơn và kiến trúc mảng điện trở nhớ bù trong bài toán nhận dạng mười ảnh nhị phân; phân tích về sự ảnh hưởng của mật độ dữ liệu đến sự suy giảm của các dòng điện ngõ ra cột và đến tính chính xác trong hoạt động của mạch Tìm dòng điện lớn nhất ở trúc mảng điện trở nhớ đơn, khi so sánh với kiến trúc mảng điện trở nhớ bù. Các kết quả phân tích về sự ảnh hưởng của mật độ dữ liệu ở chương này cũng đã được công bố trong bài báo tạp chí quốc tế, ở Công trình công bố sau của Luận án: (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). Các kết quả phân tích của chương này đã cho thấy yêu cầu cấp thiết của việc nghiên cứu tối ưu nhằm khắc phục ảnh hưởng của mật độ dữ liệu đến hoạt động của kiến trúc mảng điện trở nhớ nhị phân khi thực thi các bài toán nhận dạng, trên cơ sở chỉ sử dụng một mảng điện trở nhớ để thực thi mảng XNOR.

4.1 GIỚI THIỆU

Như đã phân tích ở các chương trước, các kiến trúc mảng điện trở nhớ nhị phân sử dụng các điện trở nhớ ở 2 trạng thái: tổng trở cao và tổng trở thấp, tương

ứng với 2 trạng thái của hệ thống nhị phân. Các kiến trúc mảng điện trở nhớ nhị phân này đã giúp khắc phục được một số hạn chế trong nỗ lực ứng dụng điện trở nhớ nhằm thiết kế các hệ điện toán mô phỏng theo hệ thần kinh. Kiến trúc mảng điện trở nhớ đơn là một kiến trúc cải tiến, được nghiên cứu và đề xuất dựa vào việc tối ưu mô hình toán của mạch XNOR [48]. Cấu trúc cải tiến này giúp kiến trúc mảng điện trở nhớ đơn tiết kiệm được diện tích mảng, công suất tiêu thụ trên mảng, và có hiệu suất tốt hơn khi một số phần tử mảng bị lỗi. Tuy nhiên, kiến trúc này mới chỉ được kiểm chứng trên 1 tập dữ liệu đơn giản. Trong chương này, kiến trúc mảng điện trở nhớ đơn sẽ được tiếp tục phân tích và kiểm chứng trên tập dữ liệu với sự đa dạng về mật độ dữ liệu. Kết quả nghiên cứu và phân tích ở chương này đã cho thấy kiến trúc mảng điện trở nhớ đơn có đáp ứng khác nhau đối với các dữ liệu có mật độ khác nhau. Cụ thể hơn, đối với các dữ liệu có mật độ bit 1 thấp thì kiến trúc mảng điện trở nhớ đơn sẽ gây ra sự suy giảm đáng kể ở các các dòng điện ngõ ra cột, điều này sẽ làm mạch Tìm dòng điện lớn nhất (Winner-take-all) hoạt động không chính xác. Nguyên nhân của sự ảnh hưởng này là do kiến trúc mảng điện trở nhớ đơn đã lược bỏ một thành phần hằng số trong quá trình giản lược mô hình toán. Đây là một phát hiện quan trọng để cho thấy rằng kiến trúc mảng điện trở nhớ đơn cần phải được tối ưu để có thể ổn định hiệu suất hoạt động trên các dữ liệu có mật độ khác nhau. Một cách cụ thể hơn, kiến trúc mảng điện trở nhớ đơn cần phải được thiết kế tối ưu để thực hiện đầy đủ các thành phần trong mô hình toán của phương trình XNOR và tránh việc lược bỏ phần tử, trong khi vẫn phải đảm bảo chỉ sử dụng một mảng điện trở nhớ duy nhất. Nội dung Chương 4 sẽ bao gồm giới thiệu kiến trúc mảng điện trở nhớ đơn từ việc tối ưu mô hình toán của hàm XNOR; tiếp theo là mô phỏng và phân tích dòng điện ngõ ra cột của kiến trúc mảng điện trở nhớ đơn khi nhận dạng ảnh với dữ liệu có mật độ khác nhau trong sự so sánh với kiến trúc mảng điện trở bù, từ đó suy luận và phát hiện ra vấn đề cần giải quyết tiếp theo để thực hiện tối ưu kiến trúc mảng điện trở nhớ nhị phân trong thực thi nhận dạng ảnh.

4.2 KIẾN TRÚC MẢNG ĐIỆN TRỞ NHỚ ĐƠN TRONG ỨNG DỤNG NHẬN DẠNG ẢNH NHỊ PHÂN

Như đã đề cập ở phương trình (2.9) và (2.10), kiến trúc mảng điện trở nhớ đơn trong các ứng dụng nhận dạng ảnh mẫu được đề xuất bằng cách tinh chỉnh việc thực thi hàm XNOR giữa vector ngõ vào A và mảng điện trở nhớ M để chỉ cần sử dụng một mảng điện trở nhớ duy nhất như sau [48]:

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

Trong đó: $I = (A - A')$ là vector dữ liệu lưỡng cực chứa các giá trị 1 hoặc -1 và A' là hằng số đã được lược bỏ.

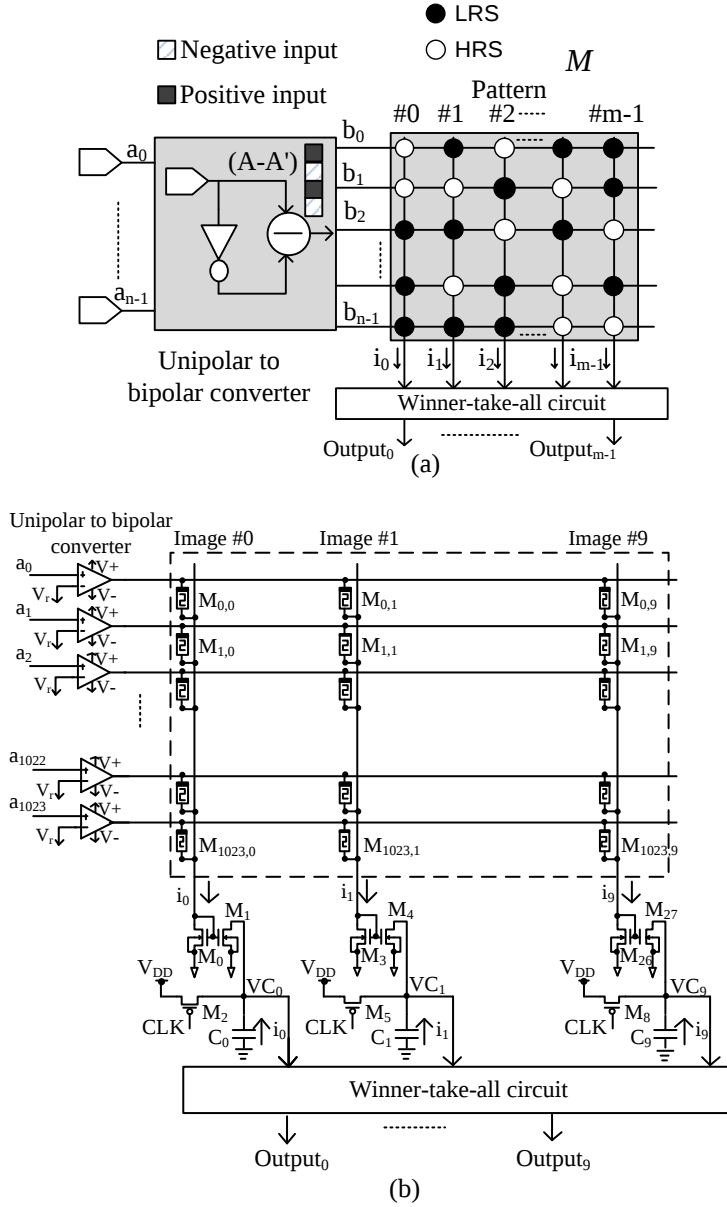
Khi đó, phép XNOR giữa vector ngõ vào A và mảng dữ liệu M được thực hiện bằng phép toán AND giữa dữ liệu lưỡng cực $I = (A - A')$ và một mảng điện trở nhớ M duy nhất. Với việc chỉ sử dụng một mảng điện trở nhớ M duy nhất thay vì sử dụng 2 mảng điện trở nhớ như phép XNOR truyền thống, kiến trúc mảng điện trở nhớ đơn sẽ đem lại sự tối ưu trong kích thước mảng khi so sánh với các kiến trúc sử dụng 2 mảng điện trở nhớ. Tuy việc nhìn nhận đại lượng A' trong phương trình (4.1) như là một hằng số và lược bỏ đi để tối giản phương trình ngõ ra Y là không sai về logic toán học, nhưng việc lược bỏ phần tử A' này sẽ gây ra sự suy giảm một lượng nhất định của dòng điện ngõ ra ở tất cả các cột. Sự suy giảm này sẽ gây ra một sự ảnh hưởng nhất định tới kết quả nhận dạng của kiến trúc mảng điện trở nhớ đơn, đặc biệt là trong ứng dụng nhận dạng ảnh mẫu là các ảnh nhị phân với mật độ bit 1 thấp.

Vì lý do đó, trong chương này tác giả sẽ phân tích sự ảnh hưởng của mật độ dữ liệu (tức mật độ các bit 1) của dữ liệu cần nhận dạng tới dòng điện ngõ ra các cột của kiến trúc mảng điện trở nhớ đơn. Các phân tích này sẽ được thực hiện trong sự so sánh với kiến trúc mảng điện trở nhớ bù nhằm đánh giá mức độ và tìm ra nguyên nhân gây nên ảnh hưởng của mật độ dữ liệu đến hoạt động kiến trúc mảng điện trở nhớ đơn. Các kết quả phân tích và đánh giá của chương này sẽ

là cơ sở quan trọng cho việc đề xuất thiết kế tối ưu cho kiến trúc mảng điện trở nhớ khi thực thi ứng dụng nhận dạng ảnh trong các hệ điện toán mô phỏng hệ thần kinh nhằm triệt tiêu ảnh hưởng của mật độ dữ liệu.

Dữ liệu được sử dụng sẽ có sự đa dạng về mật độ dữ liệu, bao gồm 10 ảnh nhị phân với các mức mật độ dữ liệu (mật độ bit 1) khác nhau. Các ảnh nhị phân này được tạo thành từ 10 ảnh xám kích thước 32×32 . Các ảnh xám được lựa chọn tương tự như trong các nghiên cứu trước của kiến trúc mảng điện trở nhớ bù và kiến trúc mảng điện trở nhớ đơn để bảo đảm sự đồng nhất và khách quan trong việc đánh giá. Các ảnh nhị phân với các mức mật độ dữ liệu khác nhau sẽ được nhận dạng với kiến trúc mảng điện trở nhớ đơn và kiến trúc mảng điện trở nhớ bù nhằm phân tích và đánh giá mức độ ảnh hưởng của mật độ dữ liệu đến dòng điện ngõ ra các cột cũng như đến hoạt động nhận dạng của các kiến trúc. Sơ đồ khối và sơ đồ nguyên lý của ứng dụng nhận dạng 10 ảnh nhị phân với kiến trúc mảng điện trở nhớ đơn được thể hiện ở Hình 4.1.

Ở Hình 4.1(a), kiến trúc mảng điện trở nhớ đơn sử dụng một mảng điện trở nhớ M với kích thước $n \times m$ để thực hiện nhận dạng m ảnh mẫu là m ảnh nhị phân. Mỗi ảnh mẫu sẽ được chuyển đổi thành một vector có kích thước $n \times 1$ và lưu vào một cột của mảng điện trở nhớ M : bit dữ liệu 1 sẽ được lưu vào một điện trở nhớ với giá trị điện trở nhớ thấp (LRS), bit dữ liệu 0 sẽ được lưu vào một điện trở nhớ với giá trị điện trở nhớ cao (HRS). Để nhận dạng, một ảnh nhị phân sẽ được chuyển thành vector ngõ vào A với kích thước $1 \times n$ và đưa vào bộ chuyển đổi dữ liệu từ đơn cực sang lưỡng cực (Unipolar to bipolar converter) để tạo thành vector dữ liệu lưỡng cực $I = (A - A') = [b_0 \quad b_1 \quad \dots \quad b_{(n-1)}]$.



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.

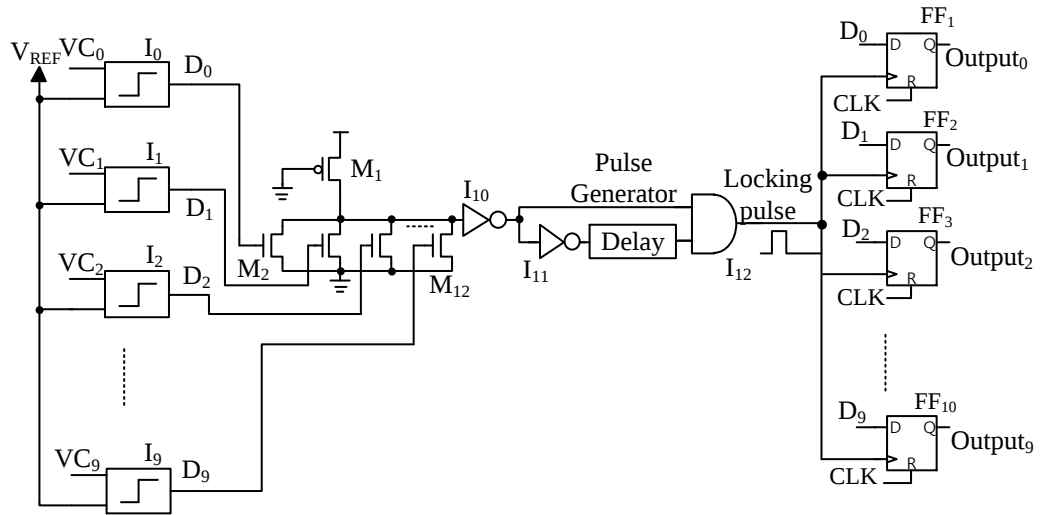
Vector dữ liệu lưỡng cực I sau đó sẽ được đưa vào mảng điện trở nhớ đơn M để thực hiện phép XNOR như đã thể hiện ở phương trình (4.1) và thu được kết quả các dòng điện ngõ ra cột như sau:

$$\begin{aligned}
 Y &= [b_0 \quad b_1 \quad \dots \quad b_{(n-1)}] \cdot \begin{bmatrix} M_{0,0} & M_{0,1} & \dots & M_{0,(m-1)} \\ M_{1,0} & M_{1,1} & \dots & M_{1,(m-1)} \\ \vdots & \vdots & \vdots & \vdots \\ M_{(n-1),0} & M_{(n-1),1} & \dots & M_{(n-1),(m-1)} \end{bmatrix} \\
 &= [i_0 \quad i_1 \quad \dots \quad i_{m-1}]
 \end{aligned} \quad (4.2)$$

Tiếp đó, các dòng điện ngõ ra cột sẽ được đưa vào mạch Tìm dòng điện lớn nhất (Winner-take-all circuit) để xác định dòng điện ngõ ra cột lớn nhất i_k và mạch Tìm dòng điện lớn nhất sẽ tạo ra xung ngõ ra ở ngõ $Output_k$ nhằm chỉ ra rằng vector ngõ vào A trùng khớp với ảnh mẫu được lưu trữ ở cột thứ k của mảng điện trở nhớ M .

Hình 4.1(b) thể hiện sơ đồ mạch nguyên lý của kiến trúc mảng điện trở nhớ đơn khi thực thi ứng dụng nhận dạng 10 ảnh nhị phân có kích thước 32×32 . Các ảnh nhị phân sẽ được chuyển thành các vector có kích thước 1024×1 và lưu vào 10 cột của mảng điện trở nhớ đơn M . Ảnh nhị phân đầu vào cần nhận dạng sẽ được chuyển đổi thành vector ngõ vào A với kích thước 1×1024 , sau đó được đưa qua các bộ so sánh để tạo vector dữ liệu lưỡng cực $I = [b_0 \ b_1 \ \dots \ b_{(n-1)}]$. Vector dữ liệu lưỡng cực I sau đó sẽ được đưa qua mảng điện trở nhớ đơn M để thực hiện phép XNOR và thu được kết quả 10 dòng điện ngõ ra cột $(i_0, i_1, \dots, i_9)$. Các dòng điện ngõ ra cột này, thông qua các gương dòng điện, sẽ làm cho các tụ đã được nạp trước ($C_0, C_1, \dots, C_9$) xả với tốc độ khác nhau. Khi dòng điện của cột thứ k , tức dòng điện i_k , càng lớn thì tụ C_k xả càng nhanh và điện áp xả VC_k của tụ càng giảm nhanh. Sau đó, các điện áp xả của các tụ ($VC_0, VC_1, \dots, VC_9$) sẽ được đưa vào vào mạch Tìm dòng điện lớn nhất (Winner-take-all circuit) để tìm ra điện áp VC_k xả nhanh nhất, tức dòng điện ngõ ra cột i_k lớn nhất, và chỉ ra rằng vector ngõ vào A trùng khớp với ảnh đã được lưu ở cột thứ k của mảng điện trở nhớ đơn M . Mạch Tìm dòng điện lớn nhất có sơ đồ nguyên lý được trình bày chi tiết ở Hình 4.2.

Trong mạch Tìm dòng điện lớn nhất ở Hình 4.2, 10 điện áp xả của các tụ điện ($VC_0, VC_1, \dots, VC_9$) sẽ được đưa vào 10 bộ so sánh ($I_0, I_1, \dots, I_9$) để so sánh với điện áp tham chiếu V_{REF} . Nếu tụ điện C_k có tốc độ xả nhanh nhất thì điện áp xả VC_k sẽ là điện áp đầu tiên giảm về thấp hơn ngưỡng so sánh V_{REF} so với các điện áp xả khác. Lúc này, bộ so sánh I_k sẽ kích hoạt ngõ ra D_k của mình lên mức 1, trong khi các ngõ ra của các bộ so sánh còn lại vẫn giữ nguyên giá trị 0 vì các điện áp xả của các tụ còn lại vẫn còn lớn hơn điện áp ngưỡng so sánh V_{REF} .

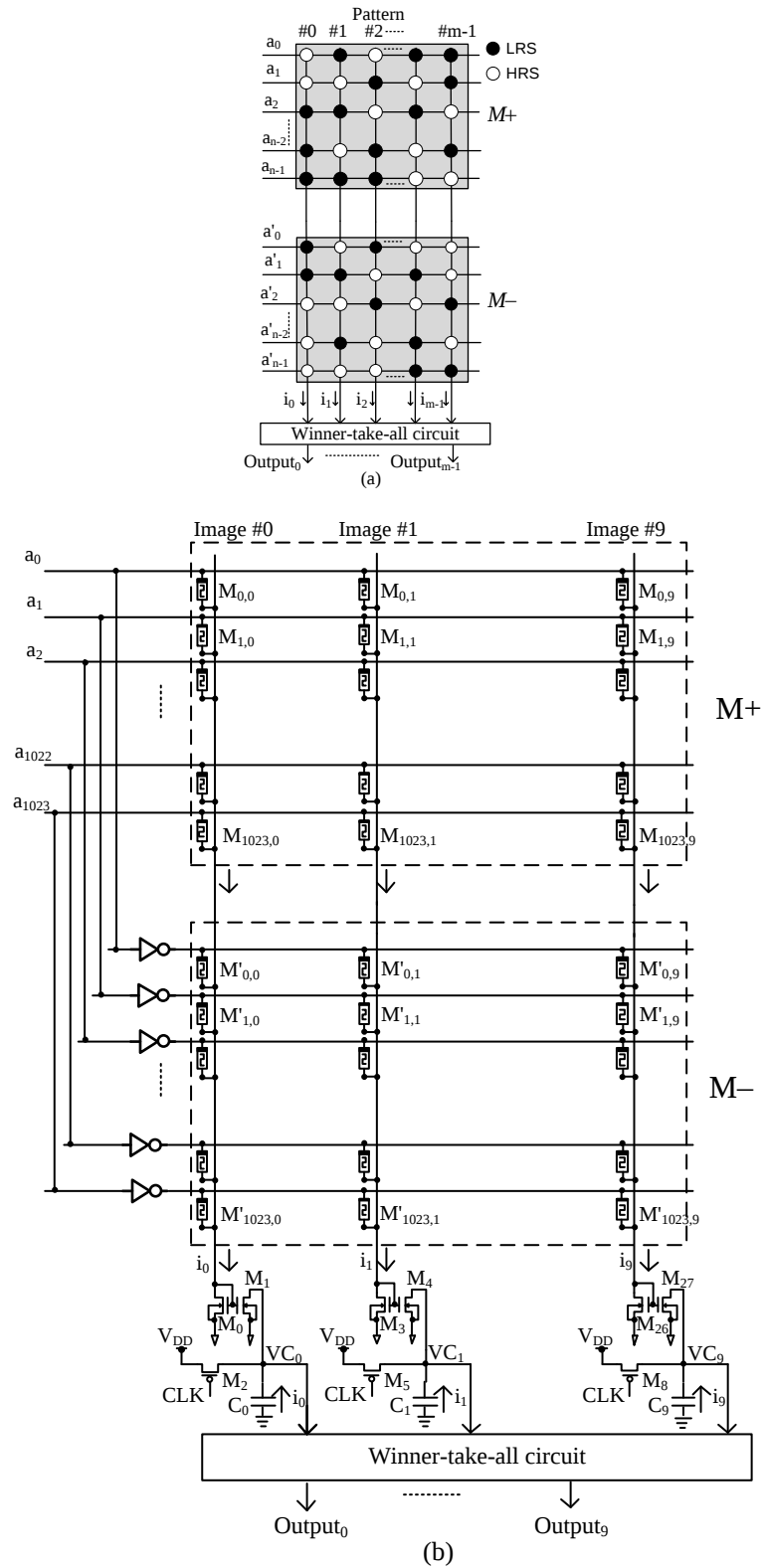


Hình 4.2: Sơ đồ nguyên lý của mạch Tìm dòng điện lớn nhất.

Xung D_k tạo ra bởi bộ so sánh I_k sẽ tiếp tục được đưa qua bộ tạo xung (Pulse Generator) để tạo ra một xung chốt nhằm kích hoạt flip-flop FF_k xuất dữ liệu D_k ở ngõ ra $Output_k$ nhằm thông báo rằng dòng điện ngõ ra cột i_k là dòng điện ngõ ra lớn nhất. Như vậy, khi điện áp xả VC_k của tụ điện thứ k giảm về ngưỡng V_{REF} nhanh nhất thì mạch Tìm dòng điện lớn nhất sẽ kích hoạt ngõ ra $Output_k$ lên mức 1, trong khi các ngõ ra khác vẫn giữ nguyên giá trị 0. Điều này cho thấy hệ thống đã nhận dạng được rằng vector ngõ vào A trùng khớp với ảnh mẫu đã được lưu ở cột thứ k của mảng điện trở nhớ đơn M .

Để làm cơ sở cho việc so sánh và đánh giá khi phân tích sự ảnh hưởng của mật độ dữ liệu của các ảnh tới hoạt động của kiến trúc mảng điện trở nhớ đơn, các ảnh nhị phân này cũng sẽ đồng thời được nhận dạng với kiến trúc mảng điện trở nhớ bù.

Kiến trúc mảng điện trở nhớ bù sẽ sử dụng 2 mảng điện trở nhớ bù nhau, $M+$ và $M-$, để thực hiện phép XNOR như đã đề cập ở Chương 3. Sơ đồ khối và sơ đồ nguyên lý của kiến trúc mảng điện trở nhớ bù cho ứng dụng nhận dạng các ảnh nhị phân được thể hiện ở Hình 4.3.



Hình 4.3: (a) Sơ đồ khối và (b) sơ đồ mạch nguyên lý của kiến trúc mạng điện trở nhớ bù trong ứng dụng nhận dạng 10 ảnh nhị phân.

Để thực hiện nhận dạng với kiến trúc mạng điện trở nhớ bù, mỗi ảnh mẫu sẽ được chuyển đổi thành một vector và lưu vào các cột của mạng điện trở nhớ:

vector ảnh gốc thứ j sẽ được lưu và cột thứ j của mảng điện trở nhớ $M+$, phần bù của vector ảnh gốc thứ j này sẽ được lưu vào cột thứ j của mảng điện trở nhớ $M-$. Như vậy, kiến trúc mảng điện trở nhớ bù sẽ sử dụng 2 mảng điện trở nhớ để lưu trữ các ảnh mẫu: mảng $M+$ lưu trữ các ảnh gốc và mảng $M-$ chứa phần bù của các ảnh gốc, tức 2 mảng điện trở nhớ $M+$ và $M-$ chứa các điện trở nhớ có giá trị bù nhau. Một vector đầu vào A sẽ được nhận dạng với kiến trúc mảng điện trở nhớ bù bằng cách thực hiện phép toán XNOR theo phương trình (2.6) và thu được các dòng điện ngõ ra cột như sau [44]:

$$\begin{aligned} Y &= A \cdot (M+) + A' \cdot (M-) \\ &= [i_0 \quad i_1 \quad \cdots \quad i_{m-1}] \end{aligned} \quad (4.3)$$

Khi được nhận dạng, vector ngõ vào A sẽ được đưa vào mảng điện trở nhớ $M+$ và phần bù của A (vector A') sẽ được đưa vào mảng điện trở nhớ $M-$ để thu được các dòng điện ngõ ra cột. Các dòng điện ngõ ra cột này sẽ tiếp tục được đưa vào mạch Tìm dòng điện lớn nhất để xác định rằng dòng điện ngõ ra cột thứ k , dòng điện i_k , là dòng điện lớn nhất, tức vector đầu vào A trùng khớp với mẫu dữ liệu đã lưu ở cột thứ k của mảng điện trở nhớ.

4.3 MẬT ĐỘ DỮ LIỆU TRONG ẢNH NHỊ PHÂN

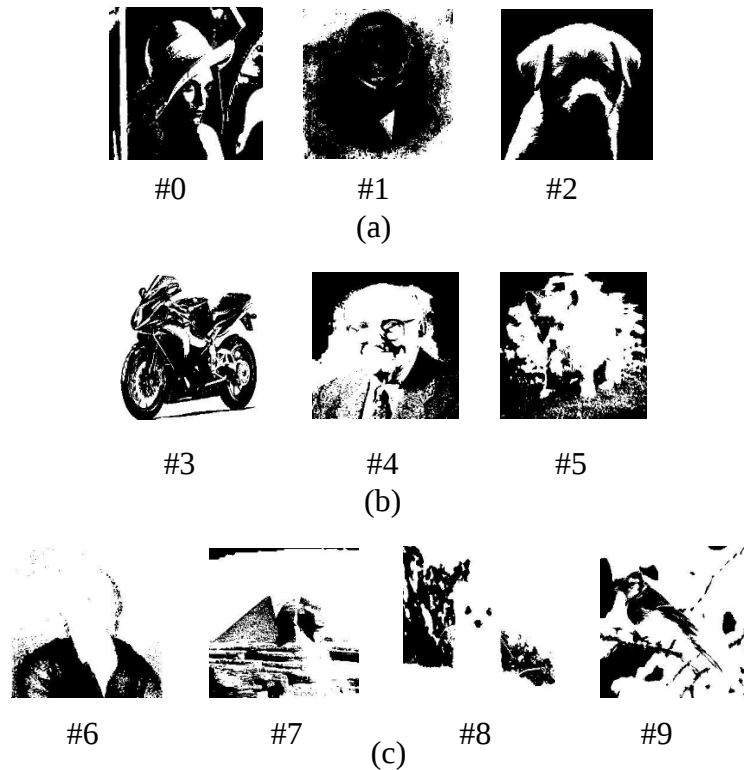
Để phân tích và đánh giá sự ảnh hưởng của mật độ dữ liệu của các ảnh trong tập dữ liệu mẫu đến hoạt động của kiến trúc mảng điện trở nhớ đơn trong ứng dụng nhận dạng ảnh, người nghiên cứu sẽ sử dụng 10 ảnh nhị phân với các mật độ dữ liệu khác nhau để nhận dạng với kiến trúc mảng điện trở nhớ đơn và kiến trúc mảng điện trở nhớ bù. 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. Điều này có nghĩa là một ảnh có mật độ dữ liệu thấp (dưới 0,5) sẽ có số lượng bit 1 ít hơn số lượng bit 0 trong ảnh. Ngược lại, nếu một ảnh có mật độ dữ liệu cao, trên 0,5, sẽ có số lượng bit 1 nhiều hơn số lượng bit 0 trong ảnh. Một ảnh nhị phân với mật độ dữ liệu thấp cũng sẽ được xem là ảnh có dữ liệu phân bố thưa.

Nhằm tạo ra tập mẫu dữ liệu gồm 10 ảnh nhị phân với sự đa dạng về mật độ dữ liệu, 10 ảnh xám có kích thước 32×32 sẽ được sử dụng để nhị phân hóa với 3 tỉ lệ mật độ dữ liệu khác nhau. Ở đây, các kiến trúc mạng điện trở nhớ sẽ thực hiện nhận dạng bằng cách thực thi phép toán XNOR giữa vector ngõ vào và các ảnh dữ liệu để tính toán mức độ trùng khớp giữa ảnh cần nhận dạng và từng ảnh trong tập dữ liệu. Do đó, tất cả các ảnh dữ liệu sẽ được lưu trữ trực tiếp vào các mảng điện trở nhớ, để thực thi phép XNOR với vector ngõ vào mà không cần quá trình học phức tạp như trong các ứng dụng học sâu. Cụ thể, mỗi ảnh sẽ được lưu trữ trên một cột của mảng điện trở nhớ, bit 1 tương ứng với giá trị điện trở thấp (LRS) và bit 0 tương ứng với giá trị điện trở cao (HSR). Việc đánh giá sẽ tập trung vào tính kiến trúc của hệ thống mà không phải là quá trình huấn luyện và học như trong các ứng dụng học sâu. Do đó, dữ liệu được sử dụng sẽ là 10 ảnh nhị phân đơn giản, thay vì tập dữ liệu lớn như MNIST với sự thay đổi phức tạp giữa các ảnh. Dữ liệu ảnh xám sẽ được chọn tương tự như trong các nghiên cứu trước của kiến trúc mạng điện trở nhớ bù và kiến trúc mạng điện trở nhớ đơn để bảo đảm sự đồng nhất cũng như tính khách quan trong đánh giá. Các ảnh xám đã dùng cho việc chuyển đổi được thể hiện ở Hình 4.4.



Hình 4.4: Mười ảnh xám được sử dụng để tạo thành các ảnh nhị phân với sự đa dạng về mật độ dữ liệu.

Mười ảnh xám ban đầu này sẽ được chuyển đổi thành ảnh các ảnh nhị phân với 3 mức mật độ dữ liệu khác nhau: mật độ thấp 0,25 (ảnh thứ #0, #1, #2), mật độ trung bình 0,5 (ảnh thứ #3, #4, #5) và mật độ cao 0,75 (ảnh thứ #6, #7, #8, #9). Các ảnh nhị phân thu được được thể hiện ở Hình 4.5 [83].



Hình 4.5: 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 và (c) mật độ dữ liệu cao 0,75.

Trong Hình 4.5(a), các ảnh có mật độ dữ liệu thấp 0,25 cho thấy số lượng bit 1 chiếm tỉ lệ 25% trong mỗi ảnh, tức mỗi ảnh sẽ có số lượng bit 1 ít hơn so với số lượng bit 0. Các ảnh ở Hình 4.5(b) có mật độ dữ liệu trung bình 0,5 thể hiện rằng số lượng bit 1 và số lượng bit 0 trong mỗi ảnh là bằng nhau. Mật độ dữ liệu của các ảnh ở Hình 4.5(c) là cao 0,75 cho biết các ảnh sẽ có số lượng điểm ảnh trắng nhiều hơn số lượng điểm ảnh đen và số lượng điểm ảnh trắng chiếm tỉ lệ 75% trong ảnh. Các ảnh nhị phân này sẽ được chuyển đổi thành các vector và lưu vào các cột của mảng điện trở nhớ đơn và các cột mảng điện trở nhớ bù, dữ liệu của mỗi ảnh được lưu trên một cột của các mảng điện trở nhớ. Bit dữ liệu 1 sẽ được lưu vào một điện trở nhớ ở trạng thái điện trở nhớ thấp (LRS) và bit dữ liệu 0 sẽ lưu vào điện trở nhớ bằng trạng thái điện trở nhớ cao (HRS), như đã đề cập ở phần 4.1.

4.4 PHÂN TÍCH SỰ ẢNH HƯỞNG CỦA MẬT ĐỘ DỮ LIỆU ĐẾN HOẠT ĐỘNG CỦA CÁC KIẾN TRÚC MẢNG ĐIỆN TRỞ NHỚ KHI NHẬN DẠNG ẢNH NHỊ PHÂN

Như đã đề cập ở phần trước, mật độ dữ liệu của một ảnh nhị phân là tỉ lệ số lượng bit 1 có trong ảnh, do đó sẽ quyết định tỉ lệ số lượng bit dữ liệu 1 được lưu trong mỗi cột của các mảng điện trở nhớ khi thực thi nhận dạng với các kiến trúc mảng điện trở nhớ nhị phân. Khi dữ liệu có mật độ cao thì tỉ lệ số bit 1, tức tỉ lệ số lượng điện trở nhớ ở trạng thái điện trở thấp (LRS), ở mỗi cột của mảng điện trở nhớ sẽ cao. Ngược lại, khi dữ liệu có mật độ thấp thì tỉ lệ số lượng điện trở nhớ ở trạng thái điện trở thấp sẽ giảm đi. Sự thay đổi của mật độ dữ liệu, cũng chính là sự thay đổi của tỉ lệ số lượng điện trở nhớ ở trạng thái điện trở thấp trong mỗi cột của mảng điện trở nhớ, sẽ gây ảnh hưởng tới dòng điện ngõ ra cột của mảng điện trở nhớ. Ở phần này, người nghiên cứu sẽ tiến hành khảo sát và phân tích chi tiết ảnh hưởng của mật độ dữ liệu đến dòng điện ngõ ra các cột cũng như ảnh hưởng đến hoạt động nhận dạng của các kiến trúc mảng điện trở nhớ, làm cơ sở cho việc tối ưu hóa kiến trúc mảng điện trở nhớ trong thực thi bài toán nhận dạng ảnh.

4.4.1 Ảnh hưởng đối với dòng điện ngõ ra cột

Để phân tích và đánh giá sự ảnh hưởng của mật độ dữ liệu trong ảnh tới hoạt động của các kiến trúc mảng điện trở nhớ nhị phân trong ứng dụng nhận dạng ảnh, các ảnh nhị phân ở Hình 4.5 sẽ lần lượt được nhận dạng với kiến trúc mảng điện trở nhớ đơn và kiến trúc mảng điện trở nhớ bù theo sơ đồ đã trình bày ở Hình 4.1 và Hình 4.3. Các thiết kế sẽ được mô phỏng kiểm chứng trên phần mềm thiết kế mạch Cadence Spectre, với các điều kiện mô phỏng và tham số thiết kế được thể hiện chi tiết ở Bảng 4.1.

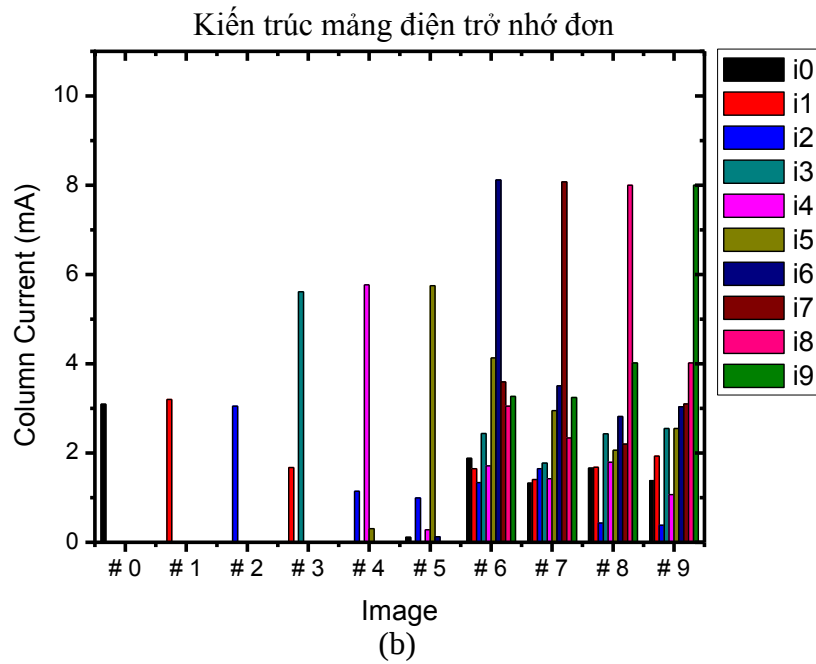
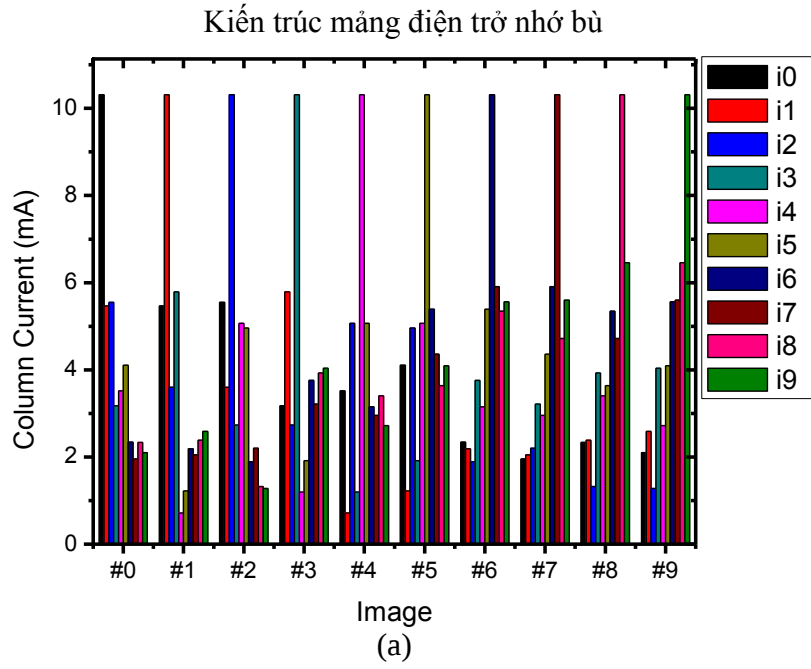
Như được trình bày ở Bảng 4.1, các điện áp ngõ vào đơn cực có giá trị là +1 V và 0 V; các điện áp ngõ vào lưỡng cực có giá trị là +1 V và -1 V tương ứng với các bit dữ liệu 1 và 0 ở ngõ vào. Các mảng điện trở nhớ được thiết kế với mô hình VerilogA. Kiến trúc mảng điện trở nhớ bù sẽ sử dụng 2 mảng điện trở nhớ:

mảng $M +$ có kích thước 1024×10 để lưu trữ 10 ảnh dữ liệu mẫu ở Hình 4.5 và mảng $M -$ có kích thước 1024×10 để lưu các giá trị bù tương ứng của mảng $M +$. Kiến trúc mảng điện trở nhớ đơn sẽ sử dụng một mảng điện trở nhớ duy nhất, mảng M với kích thước 1024×10 để lưu trữ 10 ảnh dữ liệu mẫu. Mỗi ảnh mẫu sau khi chuyển đổi sang định dạng vector sẽ được lưu trữ trong một cột của các mảng điện trở nhớ, các bit dữ liệu 1 và bit dữ liệu 0 được mã hóa vào các điện trở nhớ tương ứng với giá trị $LRS = 100 \text{ K}\Omega$ và $HRS = 10 \text{ M}\Omega$. Các mạch ngoại vi được thiết kế với công nghệ CMOS 130 nm.

Bảng 4.1: Thông số mô phỏng của kiến trúc mảng điện trở nhớ bù và kiến trúc mảng điện trở nhớ đơn khi thực thi nhận dạng mười ảnh nhị phân 32×32 .

	<i>Điện áp vào đơn cực</i>	<i>Điện áp vào lưỡng cực</i>	<i>Mảng điện trở nhớ</i>		<i>Các mạch ngoại vi</i>
Công nghệ			Mô hình VerilogA		CMOS
Tham số	+1 V 0 V	+1 V - 1 V	Kiến trúc mảng bù:	1024×10 1024×10	130 nm
			Kiến trúc mảng đơn:	1024×10	
			LRS:	100 K Ω	
			HRS:	10 M Ω	

Kết quả các dòng điện ngõ ra cột khi thực hiện nhận dạng lần lượt 10 ảnh nhị phân có mật độ dữ liệu khác nhau ở Hình 4.5 với kiến trúc mảng điện trở nhớ đơn và kiến trúc mảng điện trở nhớ bù được thể hiện ở Hình 4.6.



Hình 4.6: 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).

Kết quả cho thấy kiến trúc mạng điện trở nhớ đơn đã bị ảnh hưởng khi khi mật độ dữ liệu trong ảnh giảm đi: các dòng điện ngõ ra cột bị suy giảm khi mật độ dữ liệu giảm. Ngược lại, kiến trúc mạng điện trở nhớ bù không bị ảnh hưởng

bởi sự suy giảm của mật độ dữ liệu: các dòng điện ngõ ra cột có giá trị ổn định và tương đương nhau ở tất cả các mức mật độ dữ liệu.

Cụ thể hơn, như được phân tích và thể hiện ở Hình 4.6(a), khi nhận dạng các ảnh nhị phân có các mức mật độ dữ liệu khác nhau với kiến trúc mạng điện trở nhớ bù thì dòng điện ngõ ra cột lớn nhất là tương đương nhau ở các trường hợp, ở mức xấp xỉ 10,3 mA. Các dòng điện ngõ ra cột khác dao động trong khoảng xấp xỉ 1 mA đến 7 mA. Các dữ liệu dòng điện ngõ ra cột này cho thấy giá trị dòng điện ngõ ra cột lớn nhất tạo ra bởi kiến trúc mạng điện trở nhớ bù là không phụ thuộc vào mật độ dữ liệu của các ảnh nhị phân cần nhận dạng. Nói một cách khác, sự thay đổi trong mật độ dữ liệu của tập dữ liệu ảnh nhị phân không gây ra ảnh hưởng cho các dòng điện ngõ ra cột của kiến trúc mạng điện trở nhớ bù.

Sự không phụ của các dòng điện ngõ ra cột đối với mật độ dữ liệu của tập dữ liệu ở kiến trúc mạng điện trở nhớ bù là nhờ vào đặc điểm sử dụng 2 mảng điện trở nhớ, $M+$ và $M-$, để lưu các ảnh của tập mẫu như đã đề cập ở phương trình (4.3) của kiến trúc này. Khi một ảnh có mật độ dữ liệu thấp được lưu ở cột thứ k của mảng điện trở nhớ thứ nhất, mảng $M+$, thì phần bù của ảnh này, tức là ảnh có mật độ dữ liệu cao, cũng đồng thời sẽ được lưu ở cột thứ k của mảng điện trở nhớ thứ hai, mảng $M-$. Dòng điện ngõ ra cuối cùng ở cột thứ k của kiến trúc này là kết quả tổng của dòng điện ngõ ra ở cột k được tạo ra bởi mảng $M+$ và dòng điện ngõ ra ở cột k được tạo bởi mảng $M-$, như đã trình bày ở phương trình (4.3), và vì thế sẽ không thay đổi khi mật độ dữ liệu của ảnh thay đổi. Đồng thời, khi so sánh với Hình 4.6(b), có thể thấy khi tiến hành nhận dạng các ảnh có mật độ dữ liệu thấp, mật độ 0,25, thì dòng điện ngõ ra cột lớn nhất của kiến trúc mạng điện trở nhớ bù cao hơn xấp xỉ 4 lần so với kiến trúc mạng điện trở nhớ đơn, xấp xỉ 10,3 mA so với xấp xỉ 3 mA. Như vậy, có thể kết luận rằng, sự thay đổi của mật độ dữ liệu sẽ không gây ảnh hưởng trên các dòng điện ngõ ra cột của kiến trúc mạng điện trở nhớ bù nhờ vào đặc tính sử dụng 2 mảng điện trở nhớ bù

nhau, mảng $M +$ và mảng $M -$, để thực thi phương trình toán học XNOR ở 4.3 của kiến trúc này.

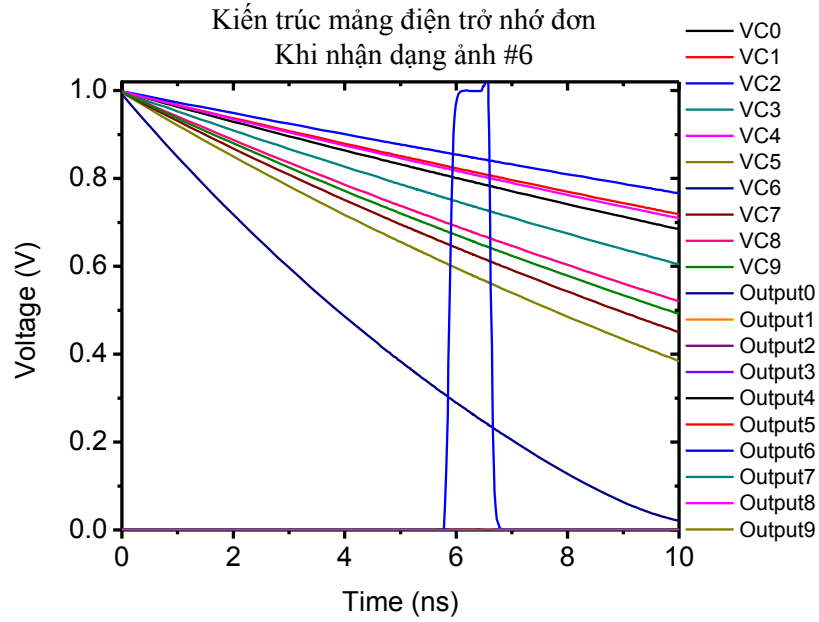
Trong khi đó, khi nhận dạng các ảnh nhị phân có mật độ dữ liệu khác nhau, kiến trúc mảng điện trở nhớ đơn cho kết quả các dòng điện ngõ ra cột bị suy giảm khi mật độ dữ liệu trong ảnh giảm, như thể hiện ở Hình 4.6(b). Cụ thể, khi nhận dạng các ảnh, từ ảnh #0 tới ảnh #2, với mật độ dữ liệu thấp 0,25 thì dòng điện ngõ ra cột lớn nhất chỉ đạt giá trị xấp xỉ 3 mA và các dòng điện ngõ ra cột khác là 0. Trong khi đó, với các ảnh có mật độ dữ liệu trung bình 0,5 (ảnh #3, #4, #5) và các ảnh có mật độ dữ liệu cao hơn 0,75 (ảnh #6, #7, #8, #9) thì dòng điện ngõ ra cột lớn nhất tạo ra bởi kiến trúc mảng điện trở nhớ đơn sẽ cao hơn, đạt giá trị xấp xỉ 5.5 mA và xấp xỉ 8.1 mA. Như vậy, khi mật độ dữ liệu trong ảnh giảm từ giá trị 0,75 xuống 0,25 thì dòng điện ngõ ra cột lớn nhất đã giảm từ 8.1 mA xuống còn 3 mA, tương đương với việc dòng điện ngõ ra cột lớn nhất đã bị giảm đi xấp xỉ 3 lần. Nguyên nhân của hiện tượng dòng điện ngõ ra cột lớn nhất của kiến trúc mảng điện trở nhớ đơn bị suy giảm khi mật độ dữ liệu trong ảnh đầu vào giảm là do phương trình thực thi hàm XNOR của kiến trúc này (phương trình 4.1) đã được tối giản và lược bỏ một đại lượng hằng số A' . Việc lược bỏ đại lượng hằng số A' tuy rằng không sai về logic toán học cho ngõ ra Y của phương trình 4.1, nhưng đã gây ra sự suy giảm một lượng dòng điện nhất định ở tất cả các dòng điện ngõ ra cột. Đặc biệt, khi một vector ảnh đầu vào A có mật độ dữ liệu thấp thì phần bù của A , tức vector A' đã được lược bỏ, 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.

Các kết quả khảo sát và phân tích về dòng điện ở trên đã cho thấy rằng khi có sự thay đổi về mật độ dữ liệu của các ảnh trong tập dữ liệu thì dòng điện ngõ ra cột của kiến trúc mảng điện trở nhớ đơn đã bị ảnh hưởng và suy giảm khi mật độ dữ liệu giảm. Trong khi đó, các dòng điện ngõ ra cột của kiến trúc mảng điện trở nhớ bù vẫn được giữ tương đương nhau và không bị ảnh hưởng bởi sự thay đổi của mật độ dữ liệu. Đây là một cơ sở quan trọng để đánh giá hoạt động nhận

dạng của các kiến trúc mạng điện trở nhớ nhị phân khi có sự thay đổi của mật độ dữ liệu của các ảnh trong tập dữ liệu. Dựa trên các kết quả về sự suy giảm của dòng điện ngõ ra cột này, hoạt động nhận dạng của các kiến trúc mạng điện trở nhớ dưới sự ảnh hưởng của mật độ dữ liệu sẽ tiếp tục được phân tích, đánh giá và trình bày ở phần tiếp theo.

4.4.2 Ảnh hưởng đối với hoạt động nhận dạng

Tiếp đến, hoạt động nhận dạng của kiến trúc mạng điện trở nhớ nhị phân khi có sự thay đổi của mật độ dữ liệu của các ảnh trong tập dữ liệu cũng sẽ được khảo sát và phân tích, trên cơ sở các kết quả dòng ngõ ra cột đã thu được ở phần trước. Như đã thể hiện ở sơ đồ Hình 4.1 và Hình 4.3, các dòng điện ngõ ra cột với giá trị khác nhau sẽ làm cho các tụ điện nạp trước, tụ 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, như đã đề cập ở phần 4.1. 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 nhằm kích hoạt flip-flop thứ k chốt giá trị ngõ ra $Output_k$. Khi nhận dạng với kiến trúc mạng điện trở nhớ đơn, nếu các ảnh đầu vào có sự giảm đi về mật độ dữ liệu thì giá trị dòng điện ngõ ra cột lớn nhất cũng sẽ bị ảnh hưởng và suy giảm, như đã được khảo sát ở Hình 4.6(b) và phân tích ở phần trước. Do đó, tốc độ xả của các tụ và mức độ giảm của các điện áp xả trong cùng một khoảng thời gian cũng sẽ khác nhau ở kiến trúc mạng điện trở nhớ đơn. Do vậy, người nghiên cứu sẽ tiến hành khảo sát điện áp xả theo thời gian của các tụ điện của kiến trúc mạng điện trở nhớ đơn khi nhận dạng các ảnh có mật độ dữ liệu khác nhau để xem xét sự ảnh hưởng của mật độ dữ liệu ảnh đến quá trình nhận dạng. Các tụ điện C_0 đến C_9 được kiểm chứng có giá trị là 50 pF. Lúc này, điện áp xả của các tụ điện từ VC_0 đến VC_9 cùng với tín hiệu xung ngõ ra $Output_6$ khi thực hiện nhận dạng ảnh thứ #6, ảnh có mật độ dữ liệu cao 0,75, với kiến trúc mạng điện trở nhớ đơn được phân tích trong khoảng thời gian 10 ns và thể hiện ở Hình 4.7.

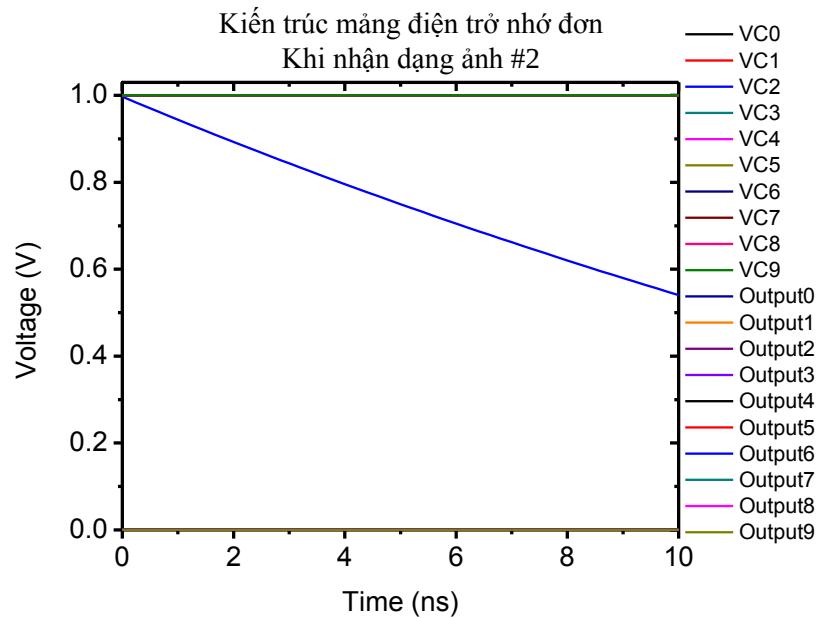


Hình 4.7: Điện áp xả trên các tụ điện, từ VC_0 đến VC_9 , khi nhận dạng ảnh thứ #6 có mật độ dữ liệu 0,75 với kiến trúc mạng điện trở nhớ đơn.

Như được phân tích ở Hình 4.7, với kiến trúc mạng điện trở nhớ đơn, khi nhận dạng ảnh thứ #6 có mật độ dữ liệu cao 0,75 thì điện áp xả tụ điện VC_6 là điện áp giảm nhanh nhất, cũng có nghĩa là tụ điện C_6 là tụ xả nhanh nhất. Cụ thể, sau khoảng thời gian xấp xỉ 4 ns thì điện áp xả VC_6 đã là điện áp giảm nhanh nhất về ngưỡng so sánh $V_{REF} = 0,5$ V của mạch Tìm dòng điện lớn nhất. Trong khi đó, các tụ điện khác xả với tốc độ chậm hơn và sau khoảng thời gian 4 ns thì điện áp của các tụ khác vẫn còn giữ cao ở mức trên 0,7 V, tức là cao hơn ngưỡng so sánh 0,5 V của mạch Tìm dòng điện lớn nhất. Sau đó, mạch Tìm dòng điện lớn nhất đã hoạt động chính xác và chót được xung ngõ ra $Output_6$ để xác định rằng dòng điện ngõ ra cột i_6 là lớn nhất sau khoảng thời gian khoảng 7 ns.

Dựa vào kết quả các điện áp xả tụ điện VC_0 đến VC_9 và tín hiệu xung ngõ ra $Output_6$ khi nhận dạng ảnh thứ #6 có mật độ dữ liệu 0,75, có thể kết luận rằng hoạt động nhận dạng của kiến trúc mạng điện trở nhớ đơn vẫn chính xác và ổn định khi ảnh nhận dạng có mật độ dữ liệu cao. Tuy nhiên, đối với ảnh đầu vào có mật độ dữ liệu thấp, hoạt động nhận dạng của kiến trúc mạng điện trở nhớ đơn cũng cần phải được khảo sát và phân tích để đánh giá được chính xác được sự ảnh hưởng của mật độ dữ liệu đến hoạt động nhận dạng của kiến trúc này.

Do vậy, hoạt động nhận dạng của kiến trúc mạng điện trở nhớ đơn với ảnh đầu vào có mật độ dữ liệu thấp sẽ được khảo sát và phân tích. Như đã trình bày ở Hình 4.6(b) và phân tích ở phần trước, khi kiến trúc mạng điện trở nhớ đơn thực hiện nhận dạng một ảnh đầu vào có mật độ dữ liệu thấp, ảnh thứ #0 đến ảnh thứ #2 có mật độ dữ liệu 0,25, thì dòng điện ngõ ra cột lớn nhất đã bị giảm xuống đáng kể, về mức xấp xỉ 3 mA, và các dòng điện ngõ ra cột khác là 0. Khi dòng điện ngõ ra cột bị giảm đi thì tụ điện sẽ xả chậm hơn và do đó điện áp xả của tụ sẽ giảm chậm hơn. Cụ thể, khi ảnh thứ #2, có mật độ dữ liệu 0,25, được nhận dạng với kiến trúc mạng 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.



Hình 4.8: 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 mạng điện trở nhớ đơn.

Hình 4.8 cho thấy khi nhận dạng ảnh thứ #2 có mật độ dữ liệu thấp 0,25 thì điện áp xả VC_2 của tụ C_2 giảm chậm hơn nếu so sánh với trường hợp ảnh đầu vào có mật độ cao 0,75 ở Hình 4.7. Đồng thời, các điện áp trên các tụ điện khác vẫn giữ nguyên giá trị nạp trước ban đầu là 1 V. Nguyên nhân của hiện tượng này là do khi nhận dạng ảnh #2 có mật độ dữ liệu thấp (0,25) thì các dòng điện ngõ ra cột của kiến trúc mạng điện trở nhớ đơn đã bị suy giảm do sự duy giảm của mật độ dữ liệu trong ảnh. Một cách cụ thể hơn, như đã được phân tích ở Hình 4.6(a),

dòng điện ngõ ra cột i_2 lúc này đã bị giảm về rất thấp thấp (chỉ xấp xỉ 3 mA) và các dòng điện ngõ ra cột khác giảm về 0.

Như vậy, nếu xét trong cùng một khoảng thời gian nhận dạng (7 ns) với ảnh thứ #6 có mật độ dữ liệu cao 0,75 (ở Hình 4.7) thì giá trị điện áp xả tụ VC_2 trong trường hợp này vẫn còn giữ ở mức trên 0,6 V. Điều này cho thấy, khi ảnh đầu vào có mật độ dữ liệu thấp 0,25, tại thời điểm 7 ns, 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 là 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 vị trí cột đã tạo ra dòng điện ngõ ra cột lớn nhất. Điều này dẫn tới hệ quả là quá trình nhận dạng các ảnh có mật độ dữ liệu thấp của kiến trúc mảng điện trở nhớ đơn sẽ không được chính xác. Kết quả điện áp xả tụ điện VC_0 đến VC_9 khi nhận dạng ảnh thứ #2 có mật độ dữ liệu thấp 0,25 đã cho thấy rằng sự suy giảm của mật độ dữ liệu sẽ ảnh hưởng xấu đến hoạt động nhận dạng của kiến trúc mảng điện trở nhớ đơn, 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.

4.5 KẾT LUẬN CHƯƠNG

Với các kết quả phân tích, luận án đã phát hiện được ảnh hưởng xấu và tìm ra được nguyên nhân gây ra sự ảnh hưởng của mật độ dữ liệu đến hoạt động nhận dạng của kiến trúc mảng điện trở nhớ đơn khi thực thi mảng XNOR trong nhận dạng ảnh nhị phân. Cụ thể, kiến trúc mảng điện trở nhớ đơn sẽ hoạt động hiệu quả khi nhận dạng các ảnh có mật độ dữ liệu cao. Điều này đồng nghĩa với việc hiệu suất nhận dạng của kiến trúc mảng điện trở nhớ đơn chỉ đạt ổn định khi nhận dạng dữ liệu có mật độ cao. Ngược lại, khi cần nhận dạng các ảnh có mật độ dữ liệu thấp trong một tập mẫu đa dạng mật độ dữ liệu thì kiến trúc mảng điện trở nhớ đơn sẽ không nhận dạng chính xác, khi xem xét trong cùng một khoảng thời gian nhận dạng. Do đó, hiệu suất nhận dạng của kiến trúc mảng điện trở nhớ đơn sẽ bị suy giảm khi mật độ dữ liệu của ảnh đầu vào giảm đi. Nguyên nhân gây ra sự ảnh hưởng này chính là do sự suy giảm của dòng điện ngõ ra cột gây ra bởi

sự giản lược đại lượng hằng số A' trong phương trình toán học mảng XNOR của kiến trúc mảng điện trở nhớ đơn.

Phát hiện này là một cơ sở rất quan trọng để cho thấy yêu cầu cấp thiết của việc tối ưu hóa kiến trúc mảng điện trở nhớ nhị phân nhằm loại bỏ ảnh hưởng của mật độ dữ liệu và đảm bảo ổn định hiệu suất nhận dạng của hệ thống trên cơ sở chỉ sử dụng một mảng điện trở nhớ duy nhất để thực thi mảng XNOR. Cụ thể hơn, cần tiếp tục nghiên cứu và đề xuất một thiết kế tối ưu cho kiến trúc mảng điện trở nhớ nhị phân để đảm bảo rằng với tập dữ liệu đa dạng về mật độ thì hoạt động nhận dạng của hệ thống khi thực thi hàm XNOR với một mảng điện trở nhớ sẽ không bị ảnh hưởng. Ở chương tiếp theo, tác giả sẽ nghiên cứu để tối ưu kiến trúc mảng điện trở nhớ nhị phân nhằm thực thi được phương trình toán học đầy đủ của mảng XNOR khi nhận dạng ảnh chỉ với một mảng điện trở nhớ duy nhất, khắc phục được ảnh hưởng của mật độ dữ liệu và đảm bảo sự ổn định cho hoạt động nhận dạng của hệ thống.

CHƯƠNG 5

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

Ở chương này, Luận án sẽ nghiên cứu và đề xuất một kiến trúc mảng điện trở nhớ tối ưu cho ứng dụng trong thiết kế các hệ điện toán mô phỏng hệ thần kinh với các bài toán nhận dạng. Kiến trúc mảng điện trở nhớ tối ưu đề xuất sẽ loại bỏ hoàn toàn sự phụ thuộc vào mật độ dữ liệu khi thực thi mảng XNOR, đồng thời duy trì được ưu điểm của việc sử dụng một mảng điện trở nhớ duy nhất. Các nội dung của chương sẽ bao gồm: thiết kế một mạch bù hằng số dòng điện cột để khôi phục thành phần hằng số trong phương trình mảng XNOR; đề xuất một thiết kế mạch tối ưu cho kiến trúc mảng điện trở nhớ nhị phân trong bài toán nhận dạng ảnh bằng cách thực thi phương trình mảng XNOR chỉ với một mảng điện trở nhớ và một mạch bù hằng số dòng điện cột, nhằm loại bỏ sự ảnh hưởng của mật độ dữ liệu đến hoạt động nhận dạng. Kiến trúc mảng điện trở nhớ tối ưu này cũng sẽ được tiếp tục ứng dụng để thiết kế một mạng nơ-ron XNOR cho bài toán nhận dạng ký tự viết tay trên tập dữ liệu MNIST. Các kết quả nghiên cứu của chương đã được công bố trong hai bài báo tạp chí quốc tế, ở các Công trình công bố sau của Luận án: (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*, vol. 25, pp. 598-609, October 2025 (SCI Q3, ISSN: 1598-1657).

5.1 GIỚI THIỆU

Các khảo sát và kết quả nghiên cứu, phân tích được trình bày ở các chương trước đã chỉ ra rằng kiến trúc mảng điện trở nhớ đơn tuy là một kiến trúc tối ưu về kích thước mảng điện trở, độ ổn định dưới sự biến thiên điện trở và khả năng đáp ứng tốt với nhiễu trên dữ liệu đầu vào khi so sánh với kiến trúc mảng điện trở nhớ bù và mảng điện trở nhớ đồng nhất, nhưng vẫn còn chịu sự ảnh hưởng của mật độ dữ liệu đến hoạt động nhận dạng. Cụ thể hơn, với một tập dữ liệu có mật độ dữ liệu không đồng đều, khi mật độ dữ liệu của ảnh đầu vào giảm đi thì dòng điện ngõ ra cột lớn nhất của kiến trúc mảng điện trở nhớ đơn cũng sẽ bị suy giảm, trong khi dòng điện này ở kiến trúc mảng điện trở nhớ bù vẫn được giữ ổn định [83]. Khi thực thi nhận dạng 10 ảnh nhị phân có kích thước 32×32 , với các ảnh đầu vào có mật độ dữ liệu cao 0,75 và các giá trị LRS và HRS được giả định là $100\text{ K}\Omega$ và $10\text{ M}\Omega$, thì dòng điện ngõ ra cột lớn nhất của kiến trúc mảng điện trở nhớ đơn đạt giá trị xấp xỉ 8,1 mA. Tuy nhiên, khi mật độ dữ liệu của ảnh đầu vào giảm xuống còn 0,25 thì dòng điện ngõ ra cột lớn nhất lúc này đã giảm xuống còn xấp xỉ 3 mA. Đặc biệt hơn, khi so sánh với kiến trúc mảng điện trở nhớ bù, khi nhận dạng một ảnh đầu vào có mật độ dữ liệu thấp là 0,25 thì dòng điện ngõ ra cột lớn nhất tạo ra bởi kiến trúc mảng điện trở nhớ đơn đã bị giảm đi xấp xỉ 4 lần. Sự suy giảm trong dòng điện ngõ ra cột lớn nhất sẽ làm cho mạch Tìm dòng điện lớn nhất (Winner-Take-All circuit) không thể xác định được cột đã cho dòng điện lớn nhất, xét trong cùng một khoảng thời gian nhận dạng khi so sánh với kiến trúc mảng điện trở nhớ bù, và dẫn tới sự suy giảm trong hiệu quả nhận dạng của kiến trúc mảng điện trở nhớ đơn.

Do đó, việc phân tích nguyên nhân của ảnh hưởng không tích cực của mật độ dữ liệu đến quá trình nhận dạng của kiến trúc mảng điện trở nhớ đơn và đề xuất thiết kế mạch để khắc phục ảnh hưởng của mật độ dữ liệu là một yêu cầu cấp thiết để tạo ra một kiến trúc mảng điện trở nhớ nhị phân tối ưu cho hệ điện toán mô phỏng hệ thần kinh thực thi nhận dạng ảnh. Cụ thể hơn, mạch tối ưu cho kiến trúc mảng điện trở nhớ nhị phân cần được thiết kế với khả năng đảm bảo ổn định

các dòng điện ngõ ra cột, làm cho dòng điện ngõ ra cột lớn nhất không bị suy giảm khi nhận dạng ảnh có mật độ dữ liệu thấp trong một tập dữ liệu đa dạng về mật độ. Điều này sẽ giúp đảm bảo được tính ổn định của hệ thống khi thực thi nhận dạng trên các tập dữ liệu có sự đa dạng về mật độ.

Chương này sẽ phân tích chi tiết nguyên nhân của sự ảnh hưởng của mật độ dữ liệu đến hoạt động nhận dạng của kiến trúc mạng điện trở nhớ đơn theo phương trình toán học. Đồng thời, Luận án cũng sẽ đề xuất một thiết kế mạch tối ưu cho kiến trúc mạng điện trở nhớ nhị phân trong ứng dụng nhận dạng ảnh trên cơ sở tối ưu phương trình toán học nhằm thực thi phương trình mạng XNOR đầy đủ chỉ với một mảng điện trở nhớ nhằm khắc phục ảnh hưởng của mật độ dữ liệu, là điều mà các nghiên cứu trước vẫn chưa giải quyết được. Một cách cụ thể hơn, kiến trúc mạng điện trở nhớ đề xuất sẽ được tối ưu với mục tiêu đảm bảo độ tin cậy của hệ thống khi tập dữ liệu có sự thay đổi về mật độ dữ liệu, trên cơ sở của việc thực thi mạng XNOR chỉ với một mảng điện trở nhớ. Do đó, chỉ tiêu đánh giá sẽ là sự ổn của tỉ lệ nhận dạng của hệ thống dưới sự thay đổi của mật độ dữ liệu trong phạm vi từ 0,25 đến 0,75. Bên cạnh đó, hệ thống cũng sẽ cần có sự đánh đổi về sự gia tăng công suất tiêu thụ do việc tối ưu sẽ cần thêm mạch phụ trợ cho mảng điện trở nhớ khi thực thi phương trình XNOR với đầy đủ các phần tử toán học. Các kết quả mô phỏng đi kèm cũng sẽ được trình bày và chứng minh cho hiệu quả của thiết kế mạch tối ưu cho kiến trúc mạng điện trở nhớ nhị phân trong ứng dụng nhận dạng ảnh được đề xuất.

5.2 THIẾT KẾ KIẾN TRÚC MẢNG 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

Như đã đề cập ở các Chương 3 và 4, kiến trúc mạng điện trở nhớ đơn thực hiện việc nhận dạng bằng cách thực thi phép toán XNOR giữa dữ liệu đầu vào A và một mảng điện trở nhớ M thay vì sử dụng hai mảng điện trở nhớ như ở các

kiến trúc mảng điện trở nhớ khác. Phép toán XNOR trên kiến trúc mảng điện trở nhớ đơn 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 phương trình toán (5.1) của kiến trúc mảng điện trở nhớ đơn, phép XNOR giữa A và M đã được triển khai và biến đổi tương đương thành một phép toán AND giữa vector ngõ vào lưỡng cực $I = (A - A')$ với mảng điện trở nhớ đơn M và một phép cộng với phần tử A' . Xét về phương diện toán học, A' là một hằng số và tạo ra một đại lượng dòng điện bằng nhau ở tất cả các cột ngõ ra, nên đã được lược bỏ để tạo thành phương trình toán học XNOR tinh gọn mà không ảnh hưởng đến tính logic của phép XNOR trên phương diện toán học.

Sự giản lược phần tử hằng số A' trong phương trình (5.1) là không sai về mặt logic toán học và đã giúp cho kiến trúc mảng điện trở nhớ đơn có phương trình toán đơn giản hơn trong việc thực thi phép XNOR khi so sánh với kiến trúc mảng điện trở nhớ bù và kiến trúc mảng điện trở đồng nhất. Tuy nhiên, ở phương diện mạch phần cứng, sự giản lược phần tử hằng số A' trong phương trình (5.1) đã làm giảm đi một lượng dòng điện ngõ ra bằng nhau ở tất cả các cột. Khi thực hiện phép XNOR theo phương trình (5.1) trên mạch phần cứng, các bit dữ liệu của vector đầu vào sẽ được chuyển đổi thành điện áp: bit 0 tương ứng với điện áp thấp, bit 1 tương ứng với điện áp cao và các bit dữ liệu trong mảng M được mã hóa thành giá trị điện trở nhớ: giá trị điện trở thấp tương ứng với bit 1 và giá trị điện trở cao tương ứng với bit 0. Do đó, khi mật độ dữ liệu của đầu vào A càng thấp, tức là mật độ bit 1 trong vector A càng thấp, cũng đồng nghĩa rằng mật độ bit 1 trong phần tử A' càng cao thì sự suy giảm của dòng điện ngõ ra các cột, gây ra bởi việc lược bỏ phần tử hằng số A' , sẽ càng lớn. Hệ quả của việc suy giảm giá trị dòng điện ngõ ra các cột, bao gồm cả dòng điện ngõ ra cột lớn nhất, là làm cho hiệu quả của việc xác định cột có dòng điện ngõ ra lớn nhất của Mạch Tìm dòng điện lớn nhất sẽ không ổn định, dẫn tới hiệu quả hoạt động của kiến trúc

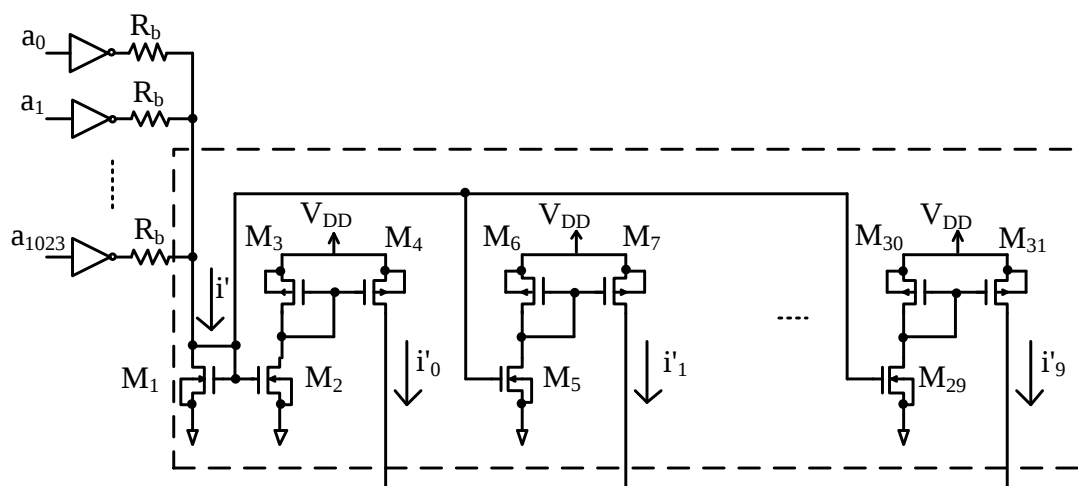
mảng điện trở nhớ đơn không ổn định khi mật độ dữ liệu của ngõ vào giảm xuống, như đã khảo sát, phân tích và trình bày ở Chương 4.

Như vậy, để khắc phục ảnh hưởng của mật độ dữ liệu đến hiệu quả hoạt động của kiến trúc mảng điện trở nhớ đơn, cần có giải pháp tối ưu hóa phần cứng để thực thi việc cộng đại lượng A' vào tất cả các cột ngõ ra nhằm làm cho dòng điện ngõ ra ở tất cả các cột được ổn định và không bị suy giảm khi ngõ vào A có mật độ dữ liệu giảm xuống thấp. Lúc này, kiến trúc mảng điện trở nhớ nhị phân đã được tối ưu sẽ thực thi phép toán XNOR theo phương trình toán đầy đủ mà không bị lược bỏ bất kỳ phần tử nào, trong khi vẫn chỉ sử dụng một mảng điện trở nhớ M , như sau [84]:

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

Ở phương trình (5.2), đại lượng hằng số A' sẽ không bị lược bỏ đi như ở phương trình (5.1) mà thay vào đó, đại lượng hằng số này sẽ được biến đổi tương đương với phép AND ($A' \cdot 1$) để giữ lại. Trong phép AND ($A' \cdot 1$), 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. Từ phương trình (5.2), có thể thấy rằng đại lượng hằng số A' có thể được tạo ra bằng cách lấy giá trị đảo của điện áp đầu vào A đi qua một mảng điện trở nhớ có giá trị điện trở thấp (LRS) để bù thêm một lượng dòng điện là hằng số bằng nhau ở tất cả các cột ngõ ra. Tuy nhiên, việc sử dụng thêm một mảng điện trở nhớ nhị phân thứ 2 sẽ làm thay đổi kiến trúc của mảng điện trở nhớ, vốn dĩ chỉ sử dụng một mảng điện trở nhớ duy nhất để thực thi phép toán XNOR, sẽ dẫn tới làm tăng kích thước cũng như các hiệu ứng phụ của mảng điện trở nhớ. Do đó, 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. Dòng điện tạo ra bởi mạch thực thi phép AND ($A' \cdot 1$) sẽ được sao chép và cộng vào tất cả các cột của mảng điện trở nhớ duy nhất M . Với thiết kế này, dòng điện ngõ ra của mỗi cột của mảng điện trở nhớ sẽ được cộng vào một đại lượng hằng số A' tương

ứng với đầu vào A . Khi đó, kiến trúc mảng điện trở nhớ nhị phân đã được tối ưu này sẽ thực thi được mô hình toán đầy đủ của mạng XNOR như thể hiện ở phương trình (5.2) mà không bị lược bỏ bất kỳ đại lượng nào, trong khi vẫn giữ được các ưu điểm của việc chỉ sử dụng một mảng điện trở nhớ duy nhất thay vì sử dụng hai mảng điện trở nhớ như ở các kiến trúc khác. Việc giữ lại thành phần A' trong phương trình toán (5.1) sẽ giúp cho hoạt động của của kiến trúc mảng điện trở nhớ sẽ được tối ưu và không còn bị ảnh hưởng bởi mật độ dữ liệu của đầu vào A , vì sự suy giảm của dòng điện ngõ ra ở các cột gây ra bởi việc lược bỏ đại lượng A' trong phương trình (5.1) đã được khắc phục. Điều này cũng có nghĩa là hoạt động của kiến trúc mảng điện trở nhớ được tối ưu sẽ tương tự như hoạt động của kiến trúc mảng điện trở nhớ bù dưới sự thay đổi của mật độ dữ liệu đầu vào, trong khi vẫn giữ được ưu điểm của việc chỉ sử dụng một mảng điện trở nhớ duy nhất thay vì sử dụng hai mảng điện trở nhớ. Mạch điện thực thi phép AND ($A' \cdot 1$) để tạo ra dòng điện bù là hằng số cho các cột ngõ ra nhằm tối ưu hóa kiến trúc mảng điện trở nhớ nhị phân khi thực thi hàm XNOR đầy đủ theo phương trình (5.2) được thiết kế và đề xuất ở Hình 5.1 [84].



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.

Hình 5.1 trình bày sơ đồ nguyên lý của mạch phần cứng được đề xuất để thực thi phép AND ($A' \cdot 1$) trong phương trình (5.2) nhằm thực hiện việc bù

hằng số dòng điện cột ngõ ra nhằm tối ưu hóa kiến trúc mảng điện trở nhớ nhị phân khi thực thi nhận dạng 10 ảnh nhị phân có kích thước 32×32 và khắc phục ảnh hưởng của mật độ dữ liệu đến hoạt động nhận dạng. Ở đây, vector đầu vào A có kích thước 1024×1 và mảng điện trở nhớ M sẽ có kích thước 1024×10 , tức có 10 cột, để thực thi nhận dạng 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, dòng điện i' . Dòng điện i' lúc này sẽ được tính toán theo phương trình:

$$i' = \sum_{k=0}^{1023} \frac{a'_k}{R_b} \quad (5.3)$$

Ở đây, điện trở R_b có giá trị bằng với giá trị điện trở thấp (LRS) của một điện trở nhớ, vì để thực thi phép AND ($A' \cdot 1$) trên mạch phần cứng, mỗi bit 1 sẽ cần được chuyển đổi thành một giá trị điện trở thấp để tạo ra dòng điện khi một giá trị điện áp ngõ vào của vector A' được đặt vào. Vì vector $A' = [a'_0 \ a'_1 \ \dots \ a'_{1023}]$ bao gồm 1024 phần tử nên số lượng điện trở R_b sẽ là 1024 điện trở để thực thi được phép AND ($A' \cdot 1$) giữa từng giá trị điện áp ngõ vào a'_i với từng điện trở R_{bi} . Dòng điện tổng được tạo ra sau các điện trở R_b , tức dòng điện i' , 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 $[i'_0 \ i'_1 \ \dots \ i'_9]$ cho các cột của mảng điện trở nhớ M , từ cột thứ 0 đến cột thứ 9. Các dòng điện bù i'_0 đến i'_9 sẽ có giá trị bằng nhau và sẽ được đưa đến các cột của mảng điện trở nhớ M ở phương trình (5.2) nhằm thực hiện việc giữ lại đại lượng hằng số A' . Việc bổ sung các dòng điện bù từ i'_0 đến i'_9 cho các cột ngõ ra của mảng điện trở nhớ M bằng thiết kế mạch đề xuất này đã giúp thực thi được phép AND ($A' \cdot 1$) trong phương trình toán học đầy đủ của hàm XNOR như đã trình bày ở phương trình (5.2) và do đó sẽ giúp làm ổn định dòng điện ngõ ra cột của mảng điện trở nhớ M khi mật độ dữ liệu đầu vào thay đổi. Nói một cách khác, thiết kế mạch bù hằng số dòng điện cột đề xuất đã giúp xây dựng được một kiến

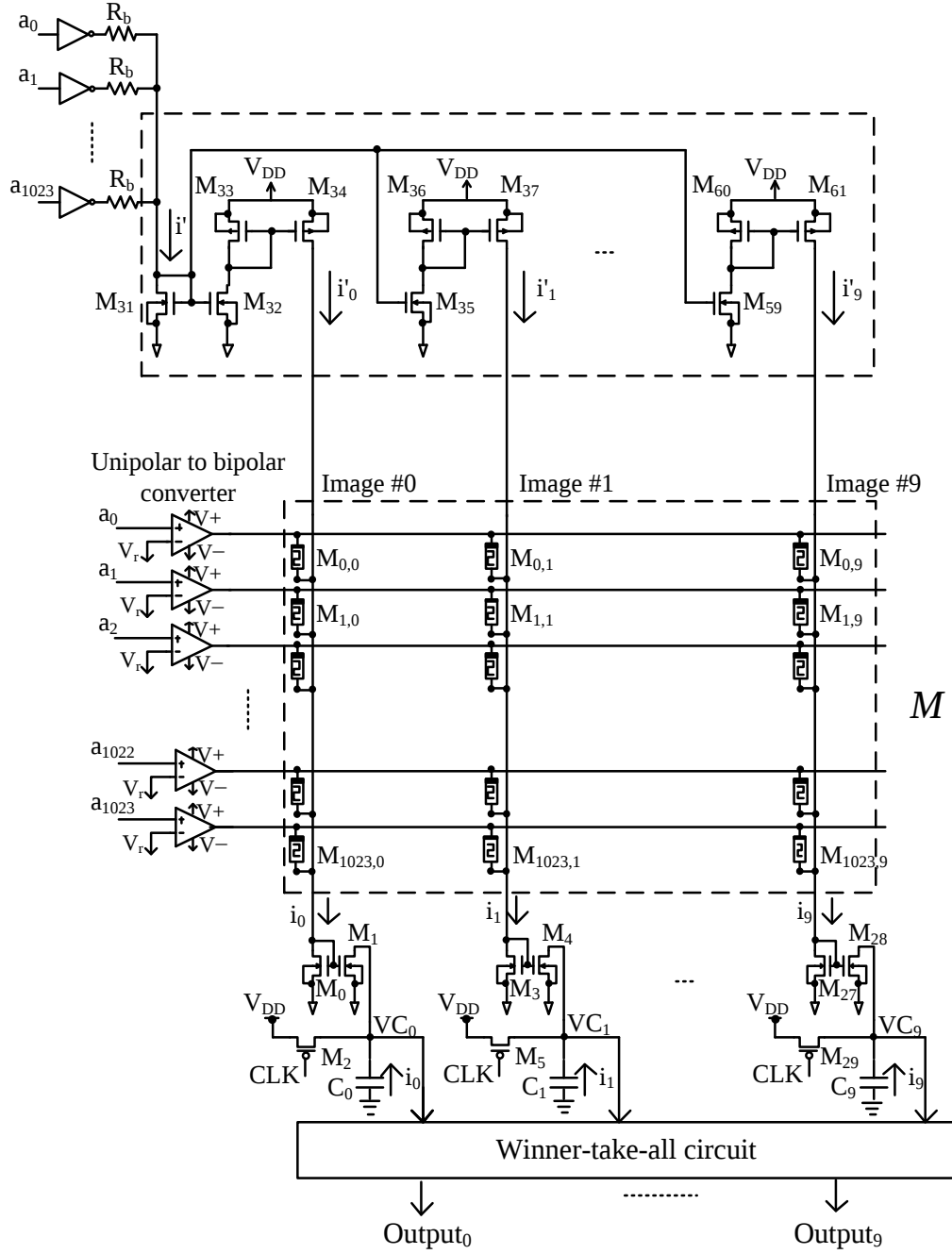
trúc mảng điện trở nhớ nhị phân tối ưu cho thực thi nhận dạng ảnh nhờ vào việc thực hiện phương trình toán học đầy đủ của hàm XNOR ở phương trình (5.2), trong khi vẫn giữ được ưu điểm của việc chỉ sử dụng một mảng điện trở nhớ M duy nhất. Khi đó, sự ảnh hưởng xấu của mật độ dữ liệu đến hoạt động nhận dạng sẽ được loại bỏ, và do đó hiệu suất nhận dạng của hệ thống sẽ được đảm bảo ổn định khi mật độ dữ liệu đầu vào suy giảm. Điều này sẽ được đánh giá và trình bày chi tiết ở phần tiếp theo.

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

Trên cơ sở mạch thực hiện thành phần phép AND ($A' \cdot 1$) trong phương trình (5.2) đã đề xuất ở phần trước, đề tài tiếp tục đề xuất 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 [84]. Thiết kế mạch tối ưu đề xuất được trình bày Hình 5.2. Lúc này, khi thực hiện nhận dạng 10 ảnh nhị phân, kiến trúc mảng điện trở nhớ đã tối ưu sẽ thực thi mảng XNOR giữa vector đầu vào A và mảng điện trở nhớ duy nhất M theo phương trình toán đầy đủ của phép XNOR là:

$$\begin{aligned} Y &= \overline{A \oplus M} = (A - A') \cdot M + A' \cdot 1 \\ &= [i_0 \quad i_1 \quad \cdots \quad i_{m-1}] \end{aligned} \quad (5.4)$$

Ở sơ đồ Hình 5.2, mảng điện trở nhớ M có kích thước 1024×10 để thực hiện lưu trữ 10 ảnh mẫu: mỗi ảnh mẫu là một ảnh nhị phân có kích thước 32×32 . Mỗi ảnh nhị phân sẽ được chuyển đổi thành 1 vector kích thước 1024×1 và được lưu vào một cột của mảng điện trở nhớ M , bit 1 tương ứng với giá trị điện trở thấp và bit 0 tương ứng với giá trị điện trở cao ở một điện trở nhớ. Vector đầu vào $A = [a_0 \quad a_1 \quad \cdots \quad a_{1023}]$ được đưa qua bộ chuyển đổi dữ liệu đơn cực sang lưỡng cực (Unipolar to bipolar converter) và tiếp tục đi qua mảng điện trở nhớ M để thực hiện phép AND thứ nhất $(A - A') \cdot M$ của phương trình (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.

Đồng thời, vector đầu vào A cũng sẽ được đưa vào mạch bù hằng số dòng điện cột đã đề xuất để thực thi phép AND thứ 2, phép $(A' \cdot 1)$, của phương trình (5.4). Việc bổ sung mạch bù hằng số dòng điện cột này sẽ giúp mảng điện trở nhớ M thực thi được phép XNOR theo phương trình toán học đầy đủ (5.4) mà không bị lược bỏ đại lượng hằng số A' như ở phương trình (5.1). Do vậy, các

dòng điện ngõ ra cột của mảng điện trở nhớ M sẽ được nâng lên và đảm bảo ổn định khi có sự suy giảm của mật độ dữ liệu trên dữ liệu đầu vào A . Điều này sẽ được phân tích chi tiết và kiểm chứng ở phần tiếp sau.

Ở mạch bù hằng số dòng điện cột, mỗi điện trở R_b sẽ thực thi chức năng của một bit 1 trong phép AND ($A' \cdot 1$) và có giá trị bằng với một giá trị điện trở thấp (LRS) trong mảng điện trở nhớ M . Vector điện áp đầu vào $A = [a_0 \ a_1 \ \dots \ a_{1023}]$ sẽ được đưa qua các cổng đảo để tạo ra vector A' và tiếp tục đi qua các điện trở R_b để thực thi phép AND ($A' \cdot 1$) và tạo ra dòng điện bù cho một cột của mảng điện trở nhớ M , dòng điện i' , như ở phương trình (5.3). Độ lớn của dòng điện i' này sẽ phụ thuộc vào giá trị cụ thể của vector dữ liệu đầu vào A . Nếu vector dữ liệu đầu vào A có mật độ dữ liệu thấp, thì vector A' sẽ có mật độ dữ liệu cao và do đó dòng điện bù i' sẽ lớn. Ngược lại, nếu vector đầu vào A có mật độ dữ liệu cao, đồng nghĩa rằng vector A' có mật độ dữ liệu thấp, thì dòng điện bù i' sẽ nhỏ đi. Các mạch gương dòng điện cấu thành bởi các transistor, từ M_{31} đến M_{61} , sẽ dùng dòng điện i' này làm dòng điện tham chiếu và sao chép thành các dòng điện từ i'_0 đến i'_9 có giá trị bằng nhau tương ứng với một vector dữ liệu đầu vào A . Lúc này, mạch bù dòng điện sẽ tạo ra các dòng điện bù, từ i'_0 đến i'_9 , cho các cột của mảng điện trở nhớ M . Do vậy, các dòng điện ngõ ra cột cuối cùng của mảng điện trở nhớ M , dòng điện từ i_0 đến i_9 , sẽ được cộng một đại lượng hằng số ($A' \cdot 1$), tức được cộng thêm các dòng điện từ i'_0 đến i'_9 , và đảm bảo ổn định khi có sự thay đổi của mật độ dữ liệu của vector đầu vào A . Việc đảm bảo được sự ổn định của các dòng điện ngõ ra cột của mảng điện trở nhớ M sẽ giúp cho hoạt động nhận dạng được chính xác và hiệu suất nhận dạng được ổn định dưới sự thay đổi của mật độ dữ liệu.

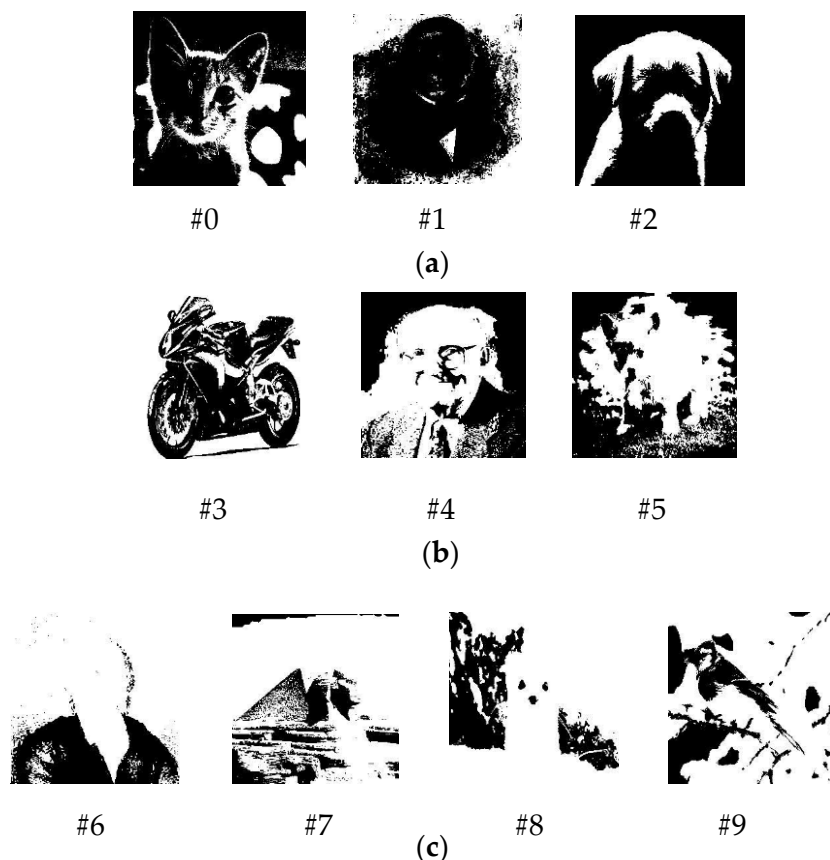
Như vậy, với việc được đề xuất bổ sung mạch bù hằng số dòng điện cột để thực hiện phép AND ($A' \cdot 1$) ở phương trình (5.4), các dòng điện ngõ ra cột của mảng điện trở nhớ M lúc này đã thể hiện đầy đủ các đại lượng trong mô hình toán học của mạng XNOR. Khi mật độ dữ liệu của vector dữ liệu đầu vào A giảm xuống thấp, các dòng điện ngõ ra cột của mảng điện trở nhớ M sẽ không còn bị

suy giảm vì đại lượng hằng số A' , lúc này là đại lượng có mật độ dữ liệu cao, trong phương trình toán của mạng XNOR vẫn được giữ lại. Lúc này, các dòng điện ngõ ra cột của kiến trúc mảng điện trở nhớ đã được tối ưu sẽ có giá trị tương tự như các dòng điện cột tạo ra bởi kiến trúc mảng điện trở nhớ bù, tức các dòng điện ngõ ra cột sẽ được ổn định, không bị suy giảm khi mật độ dữ liệu đầu vào giảm đi và các dòng điện ngõ ra cực đại sẽ có giá trị bằng nhau. Điều này sẽ giúp cho điện áp xả trên các tụ điện C_0 đến C_9 , tức điện áp VC_0 đến VC_9 , không còn bị ảnh hưởng bởi sự suy giảm của mật độ dữ liệu và do đó giúp cho hoạt động của mạch Tìm dòng điện lớn nhất (Winner-take-all circuit) không bị phụ thuộc vào mật độ dữ liệu của vector dữ liệu đầu vào A . Do vậy, thiết kế mạch tối ưu cho kiến trúc mảng điện trở nhớ được đề xuất ở Hình 5.2 đã giúp loại bỏ được ảnh hưởng của mật độ dữ liệu đến hoạt động nhận dạng, vốn dĩ là một vấn đề còn tồn tại trước đó của kiến trúc mảng điện trở nhớ đơn như đã nghiên cứu và trình bày ở chương trước, trong khi vẫn giữ được các ưu điểm của việc chỉ sử dụng một mảng điện trở nhớ nhị phân M cho thực thi nhận dạng ảnh.

Tiếp theo, kiến trúc mảng điện trở nhớ tối ưu đã đề xuất sẽ được kiểm chứng thông qua thực thi nhận dạng 10 ảnh nhị phân có mật độ dữ liệu khác nhau. Các ảnh nhị phân này được chuyển đổi từ các ảnh gốc là ảnh xám nhằm tạo thành tập dữ liệu ảnh nhị phân với ba mức mật độ dữ liệu là: 0,25, 0,5 và 0,75. Dữ liệu ảnh xám sẽ được chọn tương tự như trong các nghiên cứu đã công bố trước đó của các kiến trúc mảng điện trở nhớ nhị phân thực thi nhận dạng bằng mảng XNOR, nhằm bảo đảm sự đồng nhất cũng như tính khách quan trong việc đánh giá. Khi nhận dạng, các kiến trúc mảng điện trở nhớ nhị phân, bao gồm kiến trúc mảng điện trở nhớ tối ưu đề xuất và kiến trúc mảng điện trở nhớ đơn nguyên bản, sẽ thực thi mảng XNOR giữa vector đầu vào cần nhận dạng và tập dữ liệu đã được lưu trữ trực tiếp trong các mảng điện trở nhớ, nhằm xác định được cột dữ liệu nào là trùng khớp nhất với vector đầu vào. Các phân tích ở phần này sẽ tập trung vào việc đánh giá sự ổn định của các kiến trúc mảng điện trở nhớ, trong việc thực thi mảng XNOR và tìm ra cột dữ liệu trùng khớp nhất với vector cần nhận dạng, dưới sự thay đổi của mật độ dữ liệu. Do việc đánh giá sẽ tập trung

vào tính kiến trúc của hệ thống, khác với việc chú trọng vào khả năng đáp ứng sau quá trình học của các mạng học sâu, nên dữ liệu được chọn sẽ là 10 ảnh đơn giản thay vì tập dữ liệu lớn (ví dụ như MNIST) với sự thay đổi phức tạp. Trên các tập dữ liệu khác, kiến trúc mạng điện trở nhớ tối ưu đề xuất cũng sẽ thực thi nguyên lý mạng XNOR giữa vector đầu vào và mạng điện trở nhớ lưu trữ tập dữ liệu cần nhận dạng, tương tự như ở tập dữ liệu này.

Các ảnh nhị phân, được nhị phân hóa từ mười ảnh xám đã chọn, có các mật độ dữ liệu lần lượt là: ảnh #0, #1, #2 có mật độ dữ liệu thấp 0,25; ảnh #3, #4, #5 có mật độ dữ liệu trung bình 0,5; ảnh #6, #7, #8, #9 có mật độ dữ liệu cao 0,75. Các ảnh nhị phân với mật độ dữ liệu khác nhau này được thể hiện ở Hình 5.3.



Hình 5.3: Mười ảnh nhị phân với mật độ dữ liệu khác nhau: (a) mật độ dữ liệu thấp 0,25 (ảnh #0, #1, #2), (b) mật độ dữ liệu trung bình 0,5 (ảnh #3, #4, #5), (c) mật độ dữ liệu cao 0,75 (ảnh #6, #7, #8, #9).

Ở Hình 5.3, một ảnh mật độ dữ liệu thấp 0,25 có nghĩa là số lượng bit 1 trong ảnh sẽ chiếm tỉ lệ 25%. Ngược lại, một ảnh có mật độ dữ liệu cao 0,75 cho thấy số lượng bit 1 trong ảnh chiếm tỉ lệ 75%. Một ảnh có mật độ dữ liệu trung bình 0,5 cho biết rằng số lượng bit 1 và số lượng bit 0 trong ảnh là ngang bằng nhau.

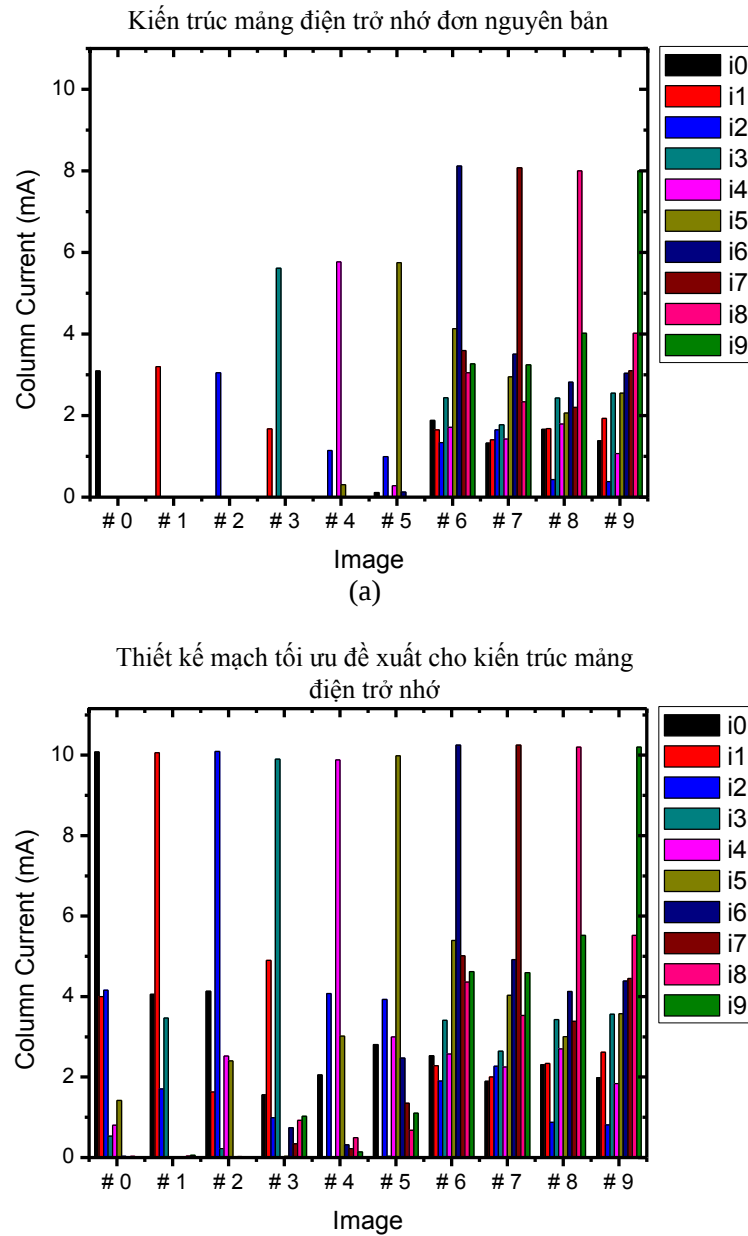
Mạch thiết kế cho kiến trúc mảng điện trở nhớ tối ưu đề xuất ở Hình 5.2 của luận án sẽ được mô phỏng kiểm chứng sử dụng phần mềm thiết kế và mô phỏng mạch Cadence Spectre. Các điều kiện mô phỏng và tham số thiết kế được thể hiện chi tiết ở Bảng 5.1. Mỗi ảnh nhị phân ở Hình 5.3 có kích thước 32×32, sẽ được chuyển đổi thành một vector có kích thước 1024×1 và được lưu trữ vào một cột của mảng điện trở nhớ M . Do đó, mảng điện trở nhớ M sẽ có kích thước là 1024×10 để lưu trữ 10 ảnh ở Hình 5.3. Bit dữ liệu 1 sẽ được lưu vào một điện trở nhớ ở trạng thái điện trở thấp (LRS) và bit dữ liệu 0 sẽ được lưu vào một điện trở nhớ ở trạng thái điện trở cao (HRS). Các trạng thái điện trở thấp và điện trở cao có giá trị lần lượt là 100 K Ω và 10 M Ω .

Bảng 5.1: Thông số mô phỏng khi thực thi nhận dạng mười ảnh nhị phân 32x32 có mật độ dữ liệu khác nhau với kiến trúc mảng điện trở nhớ tối ưu đề xuất và kiến trúc mảng điện trở nhớ đơn nguyên bản.

	<i>Điện áp vào đơn cực</i>	<i>Điện áp vào lưỡng cực</i>	<i>Mảng điện trở nhớ</i>		<i>Các mạch ngoại vi</i>
Công nghệ			Mô hình VerilogA		CMOS
Tham số	+1 V 0 V	+1 V - 1 V	Kích thước:	1024×10	130 nm
			LRS:	100 K Ω	
			HRS:	10 M Ω	

Để được nhận dạng, một ảnh đầu vào sẽ được chuyển đổi thành vector dữ liệu đầu vào $A = [a_0 \ a_1 \ \dots \ a_{1023}]$ có kích thước 1024×1, được đưa vào bộ chuyển đổi dữ liệu đơn cực sang lưỡng cực (Unipolar to bipolar converter) với

điện áp 1 V và -1 V, sau đó tiếp tục đi qua mảng điện trở nhớ M để thực thi phép AND thứ nhất, $(A - A') \cdot M$ của phương trình (5.3). Đồng thời, vector dữ liệu đầu vào A , ở dạng điện áp 0 V và 1 V tương ứng với các dữ liệu là bit 0 và bit 1, cũng sẽ được đi qua mạch bù hằng số dòng điện cột để thực thi phép AND thứ 2, phép $(A' \cdot 1)$, của phương trình (5.3) và thu được các dòng điện ngõ ra cột.



Hình 5.4: 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 mảng điện trở nhớ đơn nguyên bản, (b) thiết kế mạch tối ưu cho kiến trúc mảng điện trở nhớ với mạch bù hằng số dòng điện cột.

Vector dòng điện ngõ ra cột

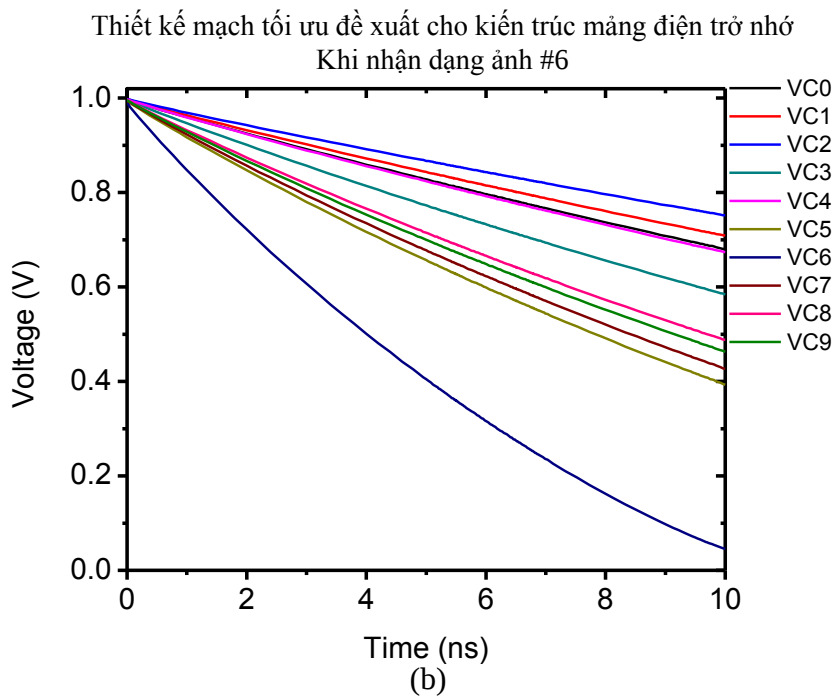
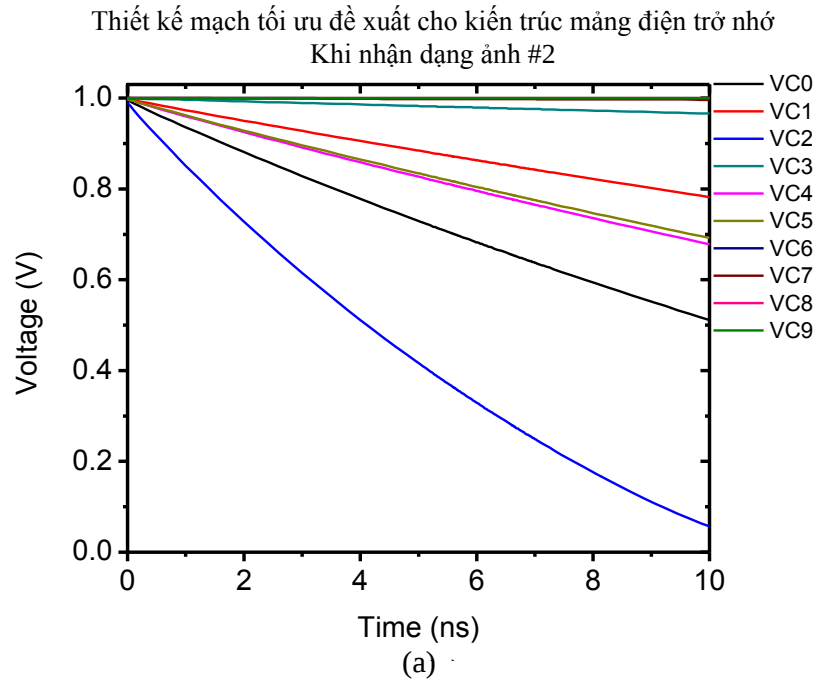
$Y = [i_0 \ i_1 \ \dots \ i_{m-1}]$ là tổng dòng điện ngõ ra của hai phép AND ở phương trình (5.3) nên sẽ được ổn định và không phụ thuộc vào mật độ dữ liệu của ảnh đầu vào A , như đã thảo luận ở phần trước. Giá trị của 10 dòng điện ngõ ra cột từ i_0 đến i_9 khi nhận dạng các ảnh từ #0 đến #9 có mật độ dữ liệu khác nhau với kiến trúc mảng điện trở nhớ đơn nguyên bản và với thiết kế mạch tối ưu cho kiến trúc mảng điện trở nhớ đề xuất được phân tích và thể hiện ở Hình 5.4.

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 thứ #0, #1, #2) và các ảnh có mật độ dữ liệu trung bình 0,5 (ảnh thứ #3, #4, #5) đã được bù thêm, nâng lên, và có giá trị tương đương với các dòng điện ngõ ra cột ở trường hợp các ảnh có mật độ dữ liệu cao 0,75 (ảnh thứ #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 khoảng xấp xỉ 10 mA, khi các ảnh đầu vào có mật độ dữ liệu khác nhau. Các kết quả dòng điện ngõ ra cột này là phù hợp với các phân tích đã trình bày ở các phần trước và cho thấy rằng các dòng điện ngõ ra cột của kiến trúc mảng điện trở nhớ tối ưu lúc này đã được đảm bảo ổn định và không còn chịu sự ảnh hưởng của sự thay đổi mật độ dữ liệu của ảnh đầu vào.

Như vậy, có thể thấy, thiết kế mạch tối ưu đề xuất cho kiến trúc mảng điện trở nhớ nhị phân thực thi nhận dạng ảnh đã giúp ổn định được dòng điện ngõ ra cột trong các điều kiện mật độ dữ liệu vào khác nhau và khắc phục được nhược điểm còn tồn tại trước đó của kiến trúc mảng điện trở nhớ đơn. Sự ổn định của các dòng điện ngõ ra cột sẽ là điều kiện quan trọng để giúp cho mạch Tìm dòng điện lớn nhất hoạt động chính xác và đảm bảo được hiệu suất hoạt động ổn định cho hệ thống khi thực thi nhận dạng ảnh. Điều này sẽ được đánh giá và trình bày cụ thể hơn ở phần tiếp sau.

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 trong thiết kế ở Hình 5.2 xả với tốc độ khác nhau và làm cho các điện áp xả VC_0 đến VC_9 sẽ giảm với tốc độ khác nhau. Ở đây, các tụ điện C_0 đến C_9 được kiểm chứng có giá trị là 50 pF. Hình 5.5 thể hiện điện áp xả của các tụ từ VC_0

đến VC_9 , của thiết kế mạch tối ưu cho kiến trúc mảng điện trở nhớ ở Hình 5.2 khi nhận dạng ảnh thứ #2 có mật độ dữ liệu thấp 0,25 và ảnh thứ #6 có mật độ dữ liệu cao 0,75.



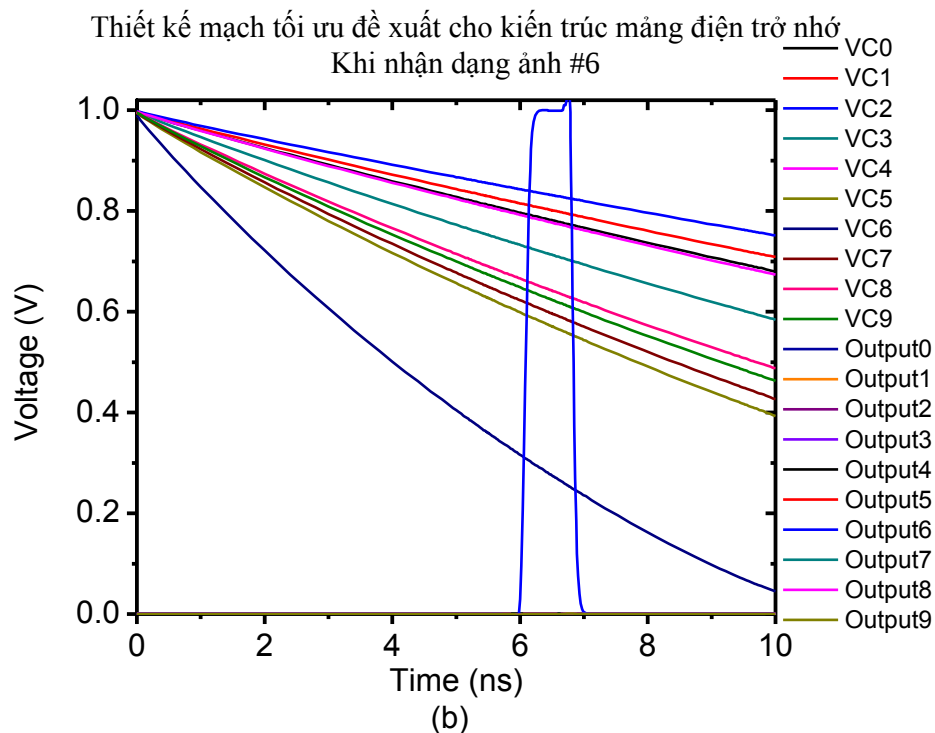
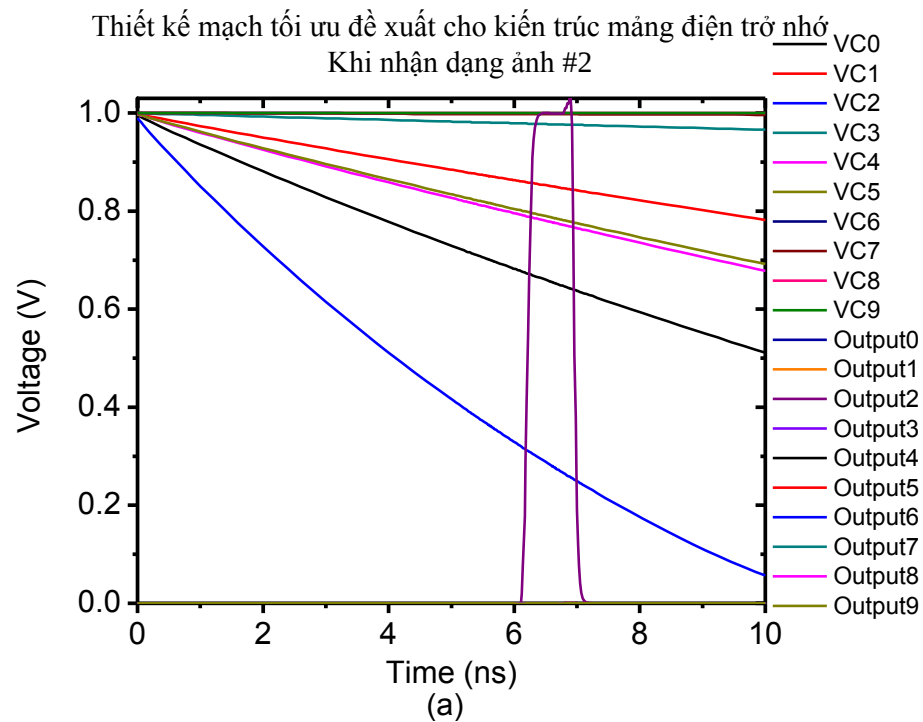
Hình 5.5: Điện áp xả của các tụ từ VC_0 đến VC_9 của thiết kế mạch tối ưu cho kiến trúc mảng điện trở nhớ với mạch bù hằng số dòng điện cột: (a) khi nhận dạng ảnh thứ #2 có mật độ dữ liệu thấp 0,25 và (b) khi nhận dạng ảnh thứ #6 có mật độ dữ liệu cao 0,75.

Hình 5.5(a) cho thấy rằng mặc dù mật độ dữ liệu của ảnh thứ #2 là thấp, mật độ 0,25, nhưng các điện áp xả của các tụ từ C_0 đến C_9 vẫn giảm được về ngưỡng 0,5 V và điện áp xả của tụ C_2 , tức điện áp VC_2 là điện áp giảm nhanh nhất trong khi các điện áp còn lại giảm chậm hơn. Cụ thể hơn, sau khoảng thời gian xấp xỉ 4,2 ns, điện áp xả VC_2 của tụ C_2 đã giảm được về ngưỡng 0,5 V trong khi các điện áp khác vẫn còn giữ ở mức cao trên 0,7 V. Do đó, mạch Tìm dòng điện lớn nhất sẽ chốt xung ngõ ra *Output2* tương ứng với điện áp giảm nhanh nhất, tức điện áp VC_2 , tại thời điểm 6,2 ns như được thể hiện ở Hình 5.6(a), để cho thấy rằng dòng điện ngõ ra cột i_2 là dòng điện lớn nhất và cũng đưa ra kết quả rằng ảnh ngõ vào, ảnh thứ #2, là trùng khớp với dữ liệu được lưu ở cột thứ 2 của mảng điện trở nhớ M . Việc này cho thấy mạch tìm dòng điện ngõ ra đã hoạt động ổn định và chính xác khi ảnh đầu vào có mật độ dữ liệu thấp, mật độ 0,25, nhờ vào việc các dòng điện ngõ ra cột đã được nâng cao và đảm bảo ổn định.

Tương tự, khi ảnh đầu vào là ảnh thứ #6 với mật độ dữ liệu cao (mật độ 0,75) thì Hình 5.5(b) cho thấy rằng các điện áp xả của các tụ vẫn giảm về ngưỡng 0,5 V và điện áp VC_6 của tụ C_6 là điện áp giảm về ngưỡng 0,5 V nhanh nhất, sau khoảng 4 ns. Khi đó, mạch Tìm dòng điện lớn nhất đã chốt ngõ ra *Output6* tương ứng với điện áp VC_6 là điện áp giảm nhanh nhất, tại thời điểm 6,1 ns như được thể hiện ở Hình 5.6(b). Việc xuất xung ngõ ra *Output6* để kết luận ảnh #6 trùng khớp với dữ liệu lưu ở cột 6 của mảng điện trở nhớ đã cho thấy mạch Tìm dòng điện lớn nhất hoạt động chính xác trong trường hợp ảnh đầu vào có mật độ dữ liệu cao 0,75.

Như vậy có thể thấy rằng việc thực hiện đầy đủ phương trình toán học XNOR ở phương trình (5.4) bằng mạch bù hằng số dòng điện cột như đề xuất ở thiết kế mạch ở Hình 5.2 đã giúp tối ưu hóa kiến trúc mảng điện trở nhớ nhị phân khi thực thi nhận dạng ảnh, khắc phục được ảnh hưởng của mật độ dữ liệu đến hoạt động nhận dạng trong khi vẫn giữ được ưu điểm của việc chỉ sử dụng một mảng điện trở nhớ M . Cụ thể, khi các ảnh trong tập dữ liệu có mật độ dữ liệu thấp thì các dòng điện ngõ ra cột đã được bù thêm và nâng lên, dòng điện ngõ ra

cột lớn nhất vẫn được giữ ổn định và ngang bằng với trường hợp ảnh đầu vào có mật độ dữ liệu cao.



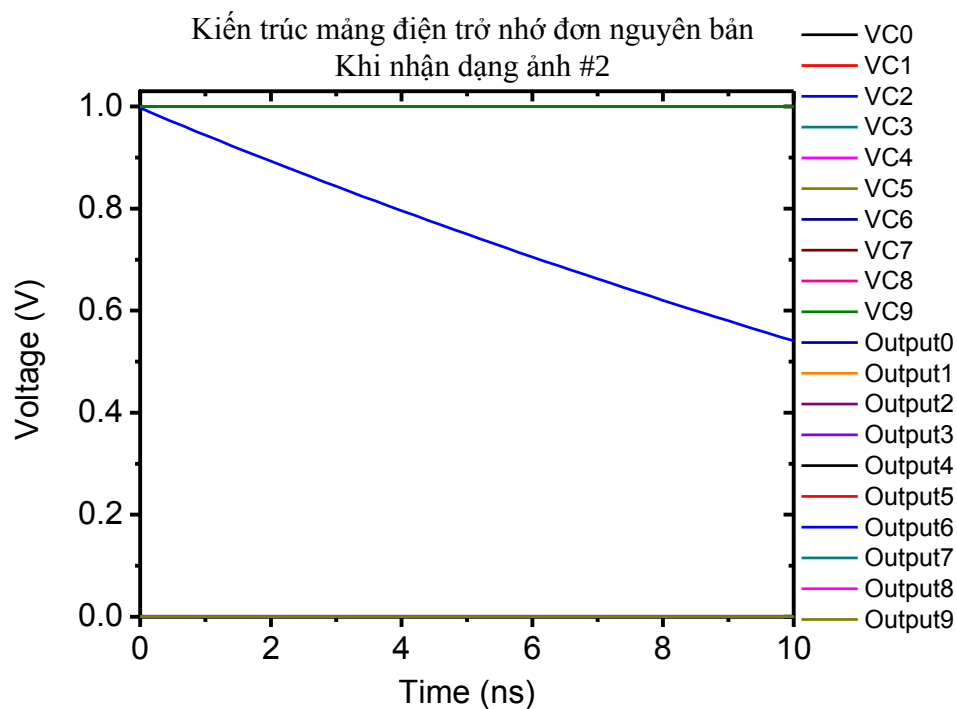
Hình 5.6: Tín hiệu ngõ ra Output tương ứng với điện áp xả của các tụ điện của thiết kế mạch tối ưu cho kiến trúc mảng điện trở nhớ với mạch bù hằng số dòng điện cột: (a) khi nhận dạng ảnh thứ #2 có mật độ dữ liệu thấp 0,25 và (b) khi nhận dạng ảnh thứ #6 có mật độ dữ liệu cao 0,75.

Nói một cách khác, các dòng điện ngõ ra cột của kiến trúc mảng điện trở nhớ tối ưu đề xuất đã được ổn định và không còn bị suy giảm khi mật độ dữ liệu giảm đi. Khi các dòng điện ngõ ra cột được ổn định thì tốc độ xả và điện áp xả của các tụ từ C_0 đến C_9 cũng sẽ được ổn định như đã khảo sát, trình bày và phân tích ở Hình 5.6. Điều này sẽ làm cho mạch Tìm dòng điện lớn nhất hoạt động chính xác, ổn định và chốt được kết quả ngõ ra sau một khoảng thời gian cố định, như được thể hiện và phân tích ở Hình 5.6.

Để phân tích và minh chứng rõ hơn cho hiệu quả khắc phục được ảnh hưởng của mật độ dữ liệu của thiết kế mạch tối ưu cho kiến trúc mảng điện trở nhớ đã đề xuất, sự ảnh hưởng của mật độ dữ liệu đến hiệu quả hoạt động của kiến trúc mảng điện trở nhớ đơn nguyên bản trong trường hợp không có mạch bù hằng số dòng điện cột cũng đã được khảo sát và đánh giá lại bằng việc nhận dạng ảnh thứ #2 có mật độ dữ liệu thấp 0,25. Kết quả điện áp xả của các tụ điện từ VC_0 đến VC_9 khi nhận dạng ảnh thứ #2 có mật độ dữ liệu thấp 0,25 và các tín hiệu ngõ ra Output với kiến trúc mảng điện trở nhớ đơn nguyên bản khi không có mạch bù hằng số dòng điện cột đề xuất được phân tích và thể hiện ở Hình 5.7.

Như có thể thấy ở Hình 5.7, với kiến trúc mảng điện trở nhớ đơn nguyên bản khi không có mạch bù hằng số dòng điện cột đã đề xuất, lúc nhận dạng ảnh thứ #2 có mật độ dữ liệu thấp 0,25, điện áp xả VC_2 là điện áp giảm nhanh nhất trong khi các điện áp xả khác vẫn duy trì ở mức gần 1 V. Tuy nhiên, điện áp xả VC_2 vẫn còn duy trì ở mức cao hơn điện áp ngưỡng tham chiếu 0,5 V của mạch Tìm dòng điện lớn nhất, khi xét trong cùng một khoảng thời gian nhận dạng. Nguyên nhân của việc này là do dưới sự ảnh hưởng của mật độ dữ liệu thấp (mật độ 0,25) của ảnh đầu vào thứ #2, dòng điện ngõ ra cột i_2 đã bị giảm xuống còn xấp xỉ 3 mA (so với khoảng 8 mA khi nhận diện ảnh có mật độ dữ liệu cao 0,75) và các dòng điện ngõ ra cột khác là gần bằng 0 như đã phân tích và thể hiện ở Hình 5.4(a). Việc điện áp xả của các tụ điện còn duy trì cao hơn ngưỡng đã làm cho mạch Tìm dòng điện lớn nhất không còn hoạt động chính xác và do đó không thể tạo xung Output ngõ ra để xuất ra kết quả nhận dạng, xét trong cùng một khoảng thời gian nhận dạng. Điều này sẽ làm suy giảm hiệu suất nhận dạng

của kiến trúc mảng điện trở nhớ đơn nguyên bản, lúc mật độ dữ liệu của ảnh đầu vào giảm xuống thấp.

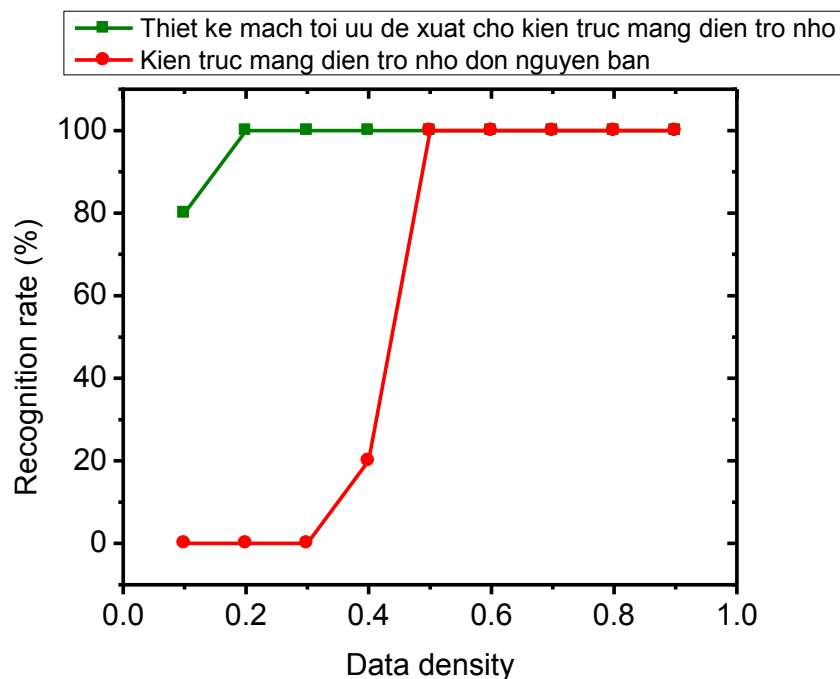


Hình 5.7: Điện áp xả của các tụ điện VC_0 đến VC_9 và các tín hiệu ngõ ra Output khi nhận dạng ảnh thứ #2 có mật độ dữ liệu thấp 0,25 với kiến trúc mảng điện trở nhớ đơn nguyên bản khi không có mạch bù hằng số dòng điện cột đã đề xuất.

Như vậy, khi không có mạch bù hằng số dòng điện cột đã đề xuất thì hiệu suất hoạt động của mảng điện trở nhớ đơn nguyên bản sẽ bị ảnh hưởng dưới sự suy giảm trong mật độ của dữ liệu. Trong khi đó, với thiết kế mạch tối ưu cho kiến trúc mảng điện trở nhớ đã đề xuất ở Hình 5.2, các điện áp xả của các tụ đã được đảm bảo ổn định như đã đánh giá ở Hình 5.5 nhờ vào sự ổn định của các dòng điện ngõ ra cột như đã phân tích ở Hình 5.4, mạch Tìm dòng điện lớn nhất đã hoạt động chính xác và xuất được xung Output ở ngõ ra như đã thể hiện ở Hình 5.6. Như vậy, thiết kế mạch tối ưu cho kiến trúc mảng điện trở nhớ đề xuất đã khắc phục được sự ảnh hưởng xấu của việc suy giảm mật độ dữ liệu và đảm bảo được hiệu quả nhận dạng của hệ thống, trong khi đó kiến trúc mảng điện trở nhớ đơn nguyên bản vẫn còn chịu sự ảnh hưởng của sự thay đổi của mật độ dữ

liệu. Kết quả tối ưu hóa này là một tiền đề quan trọng cho việc tiếp tục phát triển các ứng dụng nhận dạng ảnh dựa trên kiến trúc mạng điện trở nhớ đã tối ưu.

Để kiểm chứng hiệu quả ổn định hiệu suất nhận dạng của thiết kế mạch tối ưu cho kiến trúc mạng điện trở nhớ đã đề xuất, người nghiên cứu tiếp tục tiến hành phân tích và so sánh hiệu suất nhận dạng dưới sự thay đổi của mật độ dữ liệu, xét trong cùng một khoảng thời gian nhận dạng, giữa thiết kế mạch tối ưu cho kiến trúc mạng điện trở nhớ đã đề xuất với kiến trúc mạng điện trở nhớ đơn nguyên bản. Kết quả so sánh hiệu suất nhận dạng dưới sự thay đổi của mật độ dữ liệu giữa kiến trúc mạng điện trở nhớ đơn nguyên bản và thiết kế mạch tối ưu cho kiến trúc mạng điện trở nhớ đã đề xuất được thể hiện ở Hình 5.8.



Hình 5.8: 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 mạng điện trở nhớ đã đề xuất và kiến trúc mạng điện trở nhớ đơn nguyên bản.

Như được thể hiện ở Hình 5.8, 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. Cụ thể hơn, khi mật độ dữ liệu của ảnh đầu vào duy trì cao hơn 0,5, cả

thiết nguyên bản của kiến trúc mảng điện trở nhớ đơn và thiết kế mạch tối ưu cho kiến trúc mảng điện trở nhớ nhị phân đã đề xuất đều có hiệu suất nhận dạng cao bằng nhau và không bị ảnh hưởng. Tuy nhiên, khi mật độ dữ liệu giảm xuống còn 0,4, kiến trúc mảng điện trở nhớ đơn nguyên bản đã bị ảnh hưởng và giảm hiệu suất nhận dạng mạnh, xuống còn 20%. Trong khi đó, thiết kế mạch tối ưu đề xuất đã giúp cho hệ thống vẫn hoạt động ổn định, không bị ảnh hưởng bởi mật độ dữ liệu và duy trì được hiệu suất nhận dạng cao ở mức 100%. Đáng chú ý hơn, khi mật độ dữ liệu giảm xuống còn 0,3, thiết kế nguyên bản của kiến trúc mảng điện trở nhớ đơn đã không thể xuất ra xung *Output* cho kết quả nhận dạng trong cùng một khoảng thời gian nhận dạng và làm cho hiệu suất nhận dạng của hệ thống giảm xuống còn 0%. Hiệu suất nhận dạng được duy trì cao khi mật độ dữ liệu giảm xuống thấp của thiết kế mạch tối ưu cho kiến trúc mảng điện trở nhớ nhị phân đã đề xuất là kết quả của sự ổn định ở các dòng điện ngõ ra cột, nhờ vào mạch bù hằng số dòng điện cột, như đã phân tích và đánh giá chi tiết ở phần trước.

Công suất tiêu thụ trung bình khi thực thi nhận dạng các ảnh nhị phân 32×32 của kiến trúc mảng điện trở nhớ đơn nguyên bản và kiến trúc mảng điện trở nhớ đề xuất cũng được so sánh và trình bày ở Bảng 5.2. Ở đây, công suất tiêu thụ trung bình của mỗi kiến trúc được đo trên phần mềm mô phỏng mạch Cadance khi tiến hành nhận dạng lần lượt các ảnh với mật độ dữ liệu 0,25, 0,5, 0,75 với từng kiến trúc.

Bảng 5.2: So sánh công suất tiêu thụ giữa kiến trúc mảng điện trở nhớ đề xuất và kiến trúc mảng điện trở nhớ đơn nguyên bản khi thực thi nhận dạng ảnh nhị phân 32×32 .

	<i>Kiến trúc mảng điện trở nhớ đơn nguyên bản [48]</i>	<i>Kiến trúc mảng điện trở nhớ đề xuất</i>
Công suất tiêu thụ trung bình (mW)	50.1	95.2

Bảng 5.2 cho thấy, khi thực thi nhận dạng các ảnh nhị phân 32×32 , công suất tiêu thụ trung bình của kiến trúc mảng điện trở nhớ đề xuất sẽ tăng lên khi so

sánh với kiến trúc mảng điện trở nhớ đơn nguyên bản. Khi được tối ưu để cải thiện tỉ lệ nhận dạng dưới sự ảnh hưởng của mật độ dữ liệu, mạch bù dòng điện cột đã được bổ sung vào trong kiến trúc mảng điện trở nhớ đề xuất. Do đó, mạch bù dòng điện cột này là một nguyên nhân làm tăng công suất tiêu thụ trung bình của kiến trúc đề xuất, khi so sánh với kiến trúc mảng điện trở nhớ đơn nguyên bản vốn không có mạch bù dòng điện cột. Một nguyên nhân khác của sự tăng công suất tiêu thụ là do các dòng điện cột của kiến trúc đề xuất đã được bù thêm và nâng lên đồng đều nhau ở tất cả các mức mật độ dữ liệu, để khắc phục sự suy giảm của các dòng điện ngõ ra cột khi mật độ dữ liệu giảm, như đã trình bày và phân tích ở Hình 5.4(b). Điều này sẽ làm tăng công suất tiêu thụ trên mảng điện trở nhớ cũng như trên các mạch gương dòng điện và góp phần vào việc tăng công suất tiêu thụ trung bình của kiến trúc mảng điện trở nhớ đề xuất. Trong khi đó, ở kiến trúc mảng điện trở nhớ đơn nguyên bản, các dòng điện cột ngõ ra sẽ giảm đi đáng kể khi mật độ dữ liệu giảm đi, như đã phân tích và thể hiện ở Hình 5.4(a). Do vậy, công suất tiêu thụ trung bình trên kiến trúc mảng điện trở nhớ đơn nguyên bản sẽ thấp hơn khi so sánh với kiến trúc mảng điện trở nhớ được đề xuất. Đây là một sự đánh đổi cần phải có trong quá trình tối ưu để khắc phục được ảnh hưởng của mật độ dữ liệu. Sự đánh đổi này đã giúp kiến trúc mảng điện trở nhớ đề xuất có thể giữ được hiệu suất nhận dạng cao: 100% so với hiệu suất 0% ở mật độ 0,2 của kiến trúc mảng điện trở nhớ đơn, như đã phân tích và trình bày ở Hình 5.8.

Như vậy, thiết kế mạch tối ưu cho kiến trúc mảng điện trở nhớ nhị phân với mạch bù hằng số dòng điện cột đề xuất đã giúp loại bỏ được ảnh hưởng của mật độ dữ liệu, điều mà kiến trúc mảng điện trở nhớ đơn nguyên bản vẫn chưa khắc phục được, và đảm bảo hiệu suất hoạt động của hệ thống được cao và ổn định khi thực thi nhận dạng ảnh. Mạch thiết kế đề xuất ở đề tài này là một bước tối ưu quan trọng cho kiến trúc mảng điện trở nhớ nhị phân trong xây dựng hệ điện toán mô phỏng hệ thần kinh thực thi nhận dạng ảnh nhằm giúp đảm bảo được hiệu suất hoạt động ổn định cho hệ thống trong khi vẫn giữ được ưu điểm của việc chỉ sử dụng một mảng điện trở nhớ duy nhất. Thiết kế tối ưu cho kiến

trúc mạng điện trở nhớ nhị phân đã đề xuất này sẽ là nền tảng vững chắc cho các bước phát triển tiếp theo trong xây dựng các hệ điện toán mô phỏng hệ thần kinh thực thi nhận dạng ảnh.

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

Kiến trúc mạng điện trở nhớ nhị phân tối ưu đề xuất ở luận án sử dụng một mạng điện trở nhớ kết hợp với một mạch tạo dòng điện cố định đã thực thi một hàm XNOR đầy đủ, giải quyết được nhược điểm của kiến trúc mạng đơn trong bài toán nhận dạng ảnh. Để có cái nhìn tổng thể, luận án đã thực hiện so sánh kiến trúc mạng điện trở nhớ tối ưu đề xuất với các kiến trúc mạng điện trở nhớ khác, bao gồm: kiến trúc mạng điện trở nhớ bù, kiến trúc mạng điện trở nhớ đồng nhất và kiến trúc mạng điện trở nhớ đơn, trong bài toán nhận dạng 10 ảnh 32×32 . Kết quả so sánh được trình bày trong Bảng 5.3.

Bảng 5.3: So sánh kiến trúc mạng điện trở nhớ tối ưu đề xuất với các kiến trúc mạng điện trở nhớ nhị phân khác trong thực thi nhận dạng 10 ảnh 32×32 .

<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>
Kiến trúc mạng điện trở nhớ bù [44], [54], [55]	2	50,7%	100%
Kiến trúc mạng điện trở nhớ đồng nhất [47]	2	58,8%	20%
Kiến trúc mạng điện trở nhớ đơn [48]	1	62,1%	20%
Kiến trúc mạng điện trở nhớ tối ưu đề xuất	1	62,1%	100%

Có thể thấy, khi so sánh về số lượng linh kiện điện trở nhớ được sử dụng trong thực thi nhận dạng 10 ảnh nhị phân 32×32 thì kiến trúc mảng điện trở nhớ tối ưu đề xuất đã cho kết quả giảm 50% số lượng điện trở nhớ khi so sánh với kiến trúc mảng điện trở nhớ bù và kiến trúc mảng điện trở nhớ đồng nhất, và bằng với số lượng điện trở nhớ của kiến trúc mảng điện trở nhớ đơn. Cụ thể, kiến trúc mảng điện trở nhớ tối ưu đề xuất chỉ cần sử dụng 1 điện trở nhớ cho mỗi kết nối khớp thần kinh (synapse), bằng với kiến trúc mảng điện trở nhớ đơn. Trong khi đó, hai kiến trúc còn lại (kiến trúc mảng bù và kiến trúc mảng đồng nhất) phải sử dụng 2 điện trở nhớ cho mỗi kết nối khớp thần kinh.

Khi so sánh về độ chính xác khi nhận dạng của dưới sự biến thiên điện trở của mảng với tỉ lệ 40% thì kiến trúc mảng điện trở nhớ tối ưu đề xuất cho tỉ lệ nhận dạng đúng cao nhất và ngang bằng với kiến trúc mảng điện trở nhớ đơn, đạt 62,1%. Tỉ lệ này đạt cao hơn lần lượt là 11,4% và 3,3% so với kiến trúc mảng điện trở nhớ bù và kiến trúc mảng điện trở nhớ đồng nhất. Các ưu điểm về số lượng điện trở nhớ và sự đáp ứng với độ biến thiên điện trở nhớ là kết quả của sự tối ưu trong thực thi phương trình mảng XNOR: chỉ sử dụng một mảng điện trở nhớ thay vì hai mảng điện trở nhớ, ở kiến trúc mảng điện trở nhớ đề xuất cũng như ở kiến trúc mảng điện trở nhớ đơn. Trong khi đó, kiến trúc mảng điện trở nhớ bù vẫn phải sử dụng hai mảng điện trở nhớ và phép toán cộng, kiến trúc mảng điện trở nhớ đồng nhất cũng phải dùng hai mảng điện trở nhớ và phép toán trừ để thực thi hàm NXOR trong nhận dạng.

Khi so sánh giữa kiến trúc mảng điện trở nhớ tối ưu đề xuất và kiến trúc mảng điện trở nhớ đơn nguyên bản, nếu xét về số lượng linh kiện điện trở nhớ và tỉ lệ nhận dạng dưới sự biến thiên giá trị điện trở thì hai kiến trúc sẽ có hiệu quả ngang bằng nhau. Tuy nhiên, khi đặt dưới sự ảnh hưởng của mật độ dữ liệu thì kiến trúc mảng điện trở nhớ tối ưu đề xuất đã thể hiện được sự vượt trội hơn so với kiến trúc mảng điện trở nhớ đơn nguyên bản. Kiến trúc mảng điện trở nhớ tối ưu đề xuất đã khắc phục được ảnh hưởng của sự suy giảm mật độ dữ liệu: giữ được độ chính xác cao khi nhận dạng trong điều kiện mật độ dữ liệu giảm xuống thấp. Trong khi đó, kiến trúc mảng điện trở nhớ đơn đã bị ảnh hưởng mạnh bởi

sự suy giảm của mật độ dữ liệu, làm độ chính xác bị suy giảm mạnh. Cụ thể, khi mật độ dữ liệu giảm xuống còn 0,4 thì kiến trúc mảng điện trở nhớ tối ưu đề xuất vẫn duy trì được tỉ lệ nhận dạng đúng cao, đạt 100%, cao hơn 5 lần so với kiến trúc mảng đơn. Ngược lại, ở kiến trúc mảng điện trở nhớ đơn, tỉ lệ nhận dạng đúng đã bị suy giảm mạnh, xuống chỉ còn 20%. Đây là kết quả của sự tối ưu trong thiết kế của kiến trúc mảng điện trở nhớ đề xuất: thực thi phương trình mảng XNOR đầy đủ chỉ với một mảng điện trở nhớ duy nhất và một mạch bù hằng số dòng điện cột. Thiết kế này đã giúp cho kiến trúc mảng điện trở nhớ tối ưu đề xuất có đặc tính dòng điện ngõ ra cột giống với kiến trúc mảng điện trở nhớ bù, vốn sử dụng hai mảng điện trở nhớ bù nhau, và đồng thời vẫn chỉ sử dụng một mảng điện trở nhớ duy nhất. Kết quả này là một minh chứng quan trọng cho sự tối ưu của kiến trúc mảng điện trở nhớ nhị phân đề xuất của luận án, giúp đảm bảo sự hoạt động ổn định của kiến trúc mảng điện trở nhớ trên các tập dữ liệu có sự đa dạng về mật độ, là điều mà kiến trúc mảng điện trở nhớ đơn nguyên bản trong nghiên cứu trước đó vẫn chưa làm được.

Như vậy, từ Bảng so sánh 5.3 có thể thấy, kiến trúc mảng điện trở nhớ tối ưu đề xuất của luận án đã đạt được hiệu quả cao nhất so với các kiến trúc mảng điện trở nhớ trong các nghiên cứu trước (kiến trúc mảng điện trở nhớ bù [44], [54], [55], kiến trúc mảng điện trở nhớ đồng nhất [47], kiến trúc mảng điện trở nhớ đơn [48]) khi xét về số lượng điện trở nhớ, sự ảnh hưởng của biến thiên giá trị điện trở và sự ảnh hưởng của mật độ dữ liệu. Cụ thể, khi so sánh về số lượng linh kiện điện trở nhớ thì kiến trúc mảng điện trở nhớ đề xuất đã giảm được 50%, so với kiến trúc mảng điện trở nhớ bù và kiến trúc mảng điện trở nhớ đồng nhất. Số lượng linh kiện điện trở nhớ của kiến trúc đề xuất đạt ngang bằng với kiến trúc mảng điện trở nhớ đơn. Khi so sánh về độ chính xác nhận dạng ở tỉ lệ biến thiên điện trở 40% thì kiến trúc mảng điện trở nhớ tối ưu đề xuất đã có giá trị cao nhất: đạt độ chính xác 62,1%, cao hơn 3,3% và 11,4% so với kiến trúc mảng điện trở nhớ đồng nhất và kiến trúc mảng điện trở nhớ bù, và đạt ngang bằng với độ chính xác nhận dạng của kiến trúc mảng điện trở nhớ đơn. Khi so sánh về độ chính xác nhận dạng ở mật độ dữ liệu thấp, mật độ 0,4, thì kiến trúc mảng điện

trở nhớ tối ưu đề xuất vẫn giữ được tỉ lệ nhận dạng đúng cao nhất: đạt 100%, cao hơn 5 lần so với kiến trúc mảng điện trở nhớ đơn và kiến trúc mảng điện trở nhớ đồng nhất, và ngang bằng với tỉ lệ nhận dạng đúng của kiến trúc mảng điện trở nhớ bù.

Những kết quả này là minh chứng vững chắc cho sự vượt trội của kiến trúc mảng điện trở nhớ đề xuất khi so sánh với các kiến trúc mảng điện trở nhớ ở các nghiên cứu trước. Đồng thời, các kết quả này cũng cho thấy kiến trúc mảng điện trở nhớ đề xuất của Luận án là kiến trúc tối ưu để ứng dụng trong thiết kế các hệ điện toán mô phỏng hệ thần kinh thực thi các bài toán nhận dạng với mảng XNOR cũng như với các mạng nơ-ron XNOR.

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

Trong phần trước, luận án đã thực hiện tối ưu kiến trúc mảng điện trở nhớ nhị phân để khắc phục ảnh hưởng của mật độ dữ liệu đến hoạt động của hệ thống nhằm thực thi mảng XNOR trong ứng dụng nhận dạng ảnh với một mảng điện trở nhớ nhị phân duy nhất M theo phương trình (5.4). Cụ thể, ảnh cần nhận dạng sẽ được chuyển đổi thành một vector đầu vào A , các ảnh của tập dữ liệu cần nhận dạng sẽ được lưu vào một mảng điện trở nhớ nhị phân duy nhất M và sau đó thực thi một mảng XNOR giữa vector A và mảng điện trở nhớ M . Mỗi ngõ vào a_i sẽ thực thi XNOR với một phần tử điện trở nhớ M_{ij} của mảng điện trở nhớ M để thu được kết quả ngõ ra y_j theo Bảng trạng thái 5.4.

Bảng 5.4: Trạng thái phép XNOR giữa một ngõ vào a_i và một phần tử điện trở nhớ M_{ij} của kiến trúc mảng điện trở nhớ tối ưu đề xuất.

a_i	M_{ij}	y_j
0	0	1
0	1	0
1	0	0
1	1	1

Để thực thi Bảng trạng thái 5.4 của mạng XNOR với kiến trúc mảng điện trở nhớ đã tối ưu, mỗi ngõ vào a_i sẽ được chuyển đổi thành điện áp 0 V hoặc V + tương ứng với dữ liệu đầu vào là bit 0 hoặc bit 1, phần tử M_{ij} của mảng điện trở nhớ M sẽ lưu bit dữ liệu 0 hoặc 1 ở trạng thái điện trở cao (HRS) hoặc điện trở thấp (LRS). Sau khi thực thi phương trình (5.4) của mảng XNOR giữa vector ngõ vào A và mảng điện trở nhớ tối ưu M với mạch nguyên lý đã đề xuất ở Hình 5.2, các dòng điện ngõ ra cột i_j sẽ thể hiện mức độ giống nhau giữa vector A và các cột dữ liệu của mảng điện trở nhớ M . Mạch Tìm dòng điện lớn nhất sẽ tìm ra vị trí cột có dòng điện ngõ ra lớn nhất để xuất ra kết quả nhận dạng, như đã phân tích và đánh giá chi tiết ở phần trước của luận án.

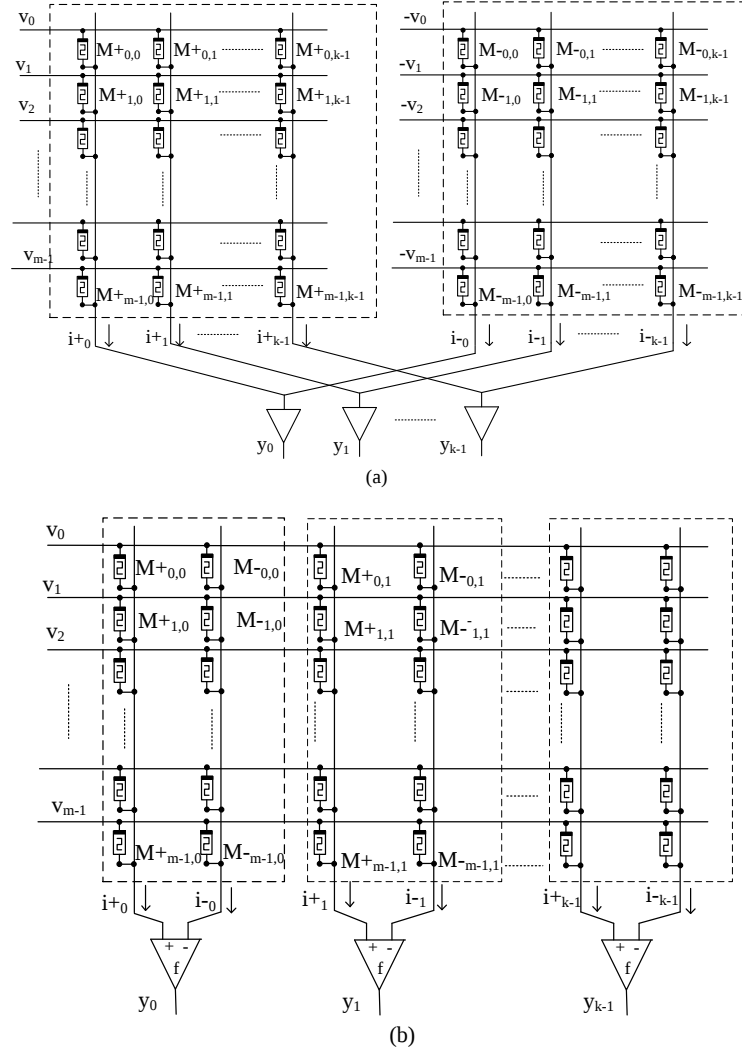
Ở phần này của luận án, tác giả tiếp tục thiết kế và đề xuất một mạng nơ-ron XNOR nhiều lớp ứng dụng kiến trúc mảng điện trở nhớ tối ưu đã đề xuất để thực thi bài toán nhận dạng ảnh trong hệ điện toán mô phỏng hệ thần kinh [85]. Thông thường, để thực thi nhận dạng ảnh, các mạng nơ-ron học sâu nhiều lớp sẽ được thiết kế và sử dụng. Tuy nhiên, với việc sử dụng các tập trọng số và các nơ-ron ở định dạng số thực 32 bit hoặc 64 bit, các mạng nơ-ron học sâu nhiều lớp thường đòi hỏi thiết bị thực thi có cấu hình mạnh và bộ nhớ lớn để lưu trữ cũng như thực hiện phép nhân giữa các nơ-ron và các tập trọng số, ví dụ như các máy tính cấu hình cao. Điều này sẽ là một rào cản cho việc thực thi các mạng nơ-ron nhiều lớp trên các phần cứng có tài nguyên hạn chế hoặc cấu hình thấp, chẳng hạn như các phần cứng tính toán ngoại vi của hệ thống IoT. Do đó, để có thể thực thi trên các phần cứng có tài nguyên hữu hạn, việc tối ưu các mạng nơ-ron nhiều lớp là điều cần được quan tâm nghiên cứu. Trong các phương pháp tối ưu, có thể đề cập đến phương pháp nhị phân hóa mạng nơ-ron nhiều lớp để giảm nhu cầu về bộ nhớ lưu trữ cũng như tài nguyên cho tính toán các phép nhân và phép cộng, dù có thể gây ra sự suy giảm trong tỉ lệ nhận dạng. Cụ thể, thay vì sử dụng các tập trọng số là số thực 32 bit hoặc 64 bit, ở mạng nơ-ron nhị phân, các trọng số sẽ được chuyển về dạng nhị phân (-1, +1) và các nơ-ron ở các lớp cũng được thể hiện dưới dạng nhị phân. Bằng cách sử dụng các giá trị nhị phân, các mạng nơ-ron nhị phân sẽ cần ít bộ nhớ hơn để lưu trữ các tập trọng số so với mạng nơ-ron

nhiều lớp truyền thống. Mạng nơ-ron XNOR với các giá trị trọng số, ngõ vào, ngõ ra nhị phân (-1, +1) có thể giảm được 32 lần kích thước bộ nhớ cần thiết khi so sánh với mạng nơ-ron tính toán trên số thực [86, 87]. Hơn nữa, các phép nhân giữa các nơ-ron vào và các trọng số ở định dạng nhị phân cũng sẽ cần ít tài nguyên và thời gian thực thi hơn so với việc thực thi ở định dạng số thực. Các tính toán trong mạng nơ-ron XNOR với các giá trị nhị phân nhanh hơn 58 lần khi thực thi trên CPU (khi xét về số lượng các phép toán) so với mạng nơ-ron truyền thống sử dụng số thực [86]. Phép nhân ma trận nhị phân khi thực thi trên GPU cũng nhanh hơn 7 lần so với phép nhân ma trận số thực [88]. Lúc này, phép nhân giữa một ngõ vào x_i và một trọng số w_{ji} ở định dạng số nhị phân (-1, +1) sẽ trở thành một phép XNOR, như thể hiện ở Bảng 5.5.

Bảng 5.5: Các trạng thái của phép XNOR giữa một ngõ vào x_i và một trọng số w_{ji} trong mạng nơ-ron nhị phân.

x_i	w_{ji}	$x_i \text{ XNOR } w_{ji}$
-1	-1	+1
-1	+1	-1
+1	-1	-1
+1	+1	+1

Như vậy, có thể xem một mạng nơ-ron nhị phân chính là một mạng nơ-ron XNOR, và mạng XNOR nhiều lớp là một giải pháp tiềm năng để thực thi các mạng nơ-ron nhị phân nhiều lớp trên các phần cứng có tài nguyên hữu hạn, chẳng hạn như các nút phần cứng xử lý ngoại vi của các hệ thống IoT [87]. Với các tập trọng số và các ngõ vào nhị phân, các mạng nơ-ron nhị phân đặc biệt phù hợp để triển khai trên phần cứng là các mảng điện trở nhớ nhị phân. Khi đó, tập trọng số W sẽ được mã hóa và lưu trữ trong các mảng điện trở nhớ nhị phân. Tuy nhiên, như có thể thấy ở Bảng 5.5, mỗi giá trị trọng số w_{ji} có thể mang hai giá trị: +1 hoặc -1, trong khi các điện trở nhớ lại chỉ có giá trị điện trở dương. Do đó, để có thể lưu trữ được dữ liệu +1 hoặc -1 với điện trở nhớ nhị phân, ở các giải pháp truyền thống, mỗi trọng số w_{ji} sẽ được mã hóa bằng 2 điện trở nhớ [19, 89, 90].



Hình 5.9: Các mô hình sử dụng 2 điện trở nhớ để mã hóa thông tin cho một trọng số w_{ji} nhị phân (-1, +1): (a) mô hình sử dụng 2 mảng điện trở nhớ với nguồn cung cấp lưỡng cực và (b) mô hình sử dụng cột đôi với nguồn cung cấp đơn cực.

Các mô hình mảng điện trở nhớ sử dụng 2 điện trở nhớ để mã hóa thông tin cho một trọng số w_{ji} của ma trận trọng số W của mạng nơ-ron nhị phân được minh họa ở Hình 5.9. Ở Hình 5.9(a), khi điện áp đầu vào là điện áp lưỡng cực, mô hình sẽ sử dụng 2 mảng điện trở nhớ M^+ và M^- để lưu một tập trọng số W , mỗi trọng số w_{ji} sẽ được mã hóa ở 2 điện trở nhớ M^+_{ij} và M^-_{ij} trên 2 mảng điện trở nhớ tại vị trí tương ứng [19]. Lúc này, ngõ ra Y sẽ tính theo phương trình $Y = \overline{V \oplus M} = (V^+)(M^+) + (V^-)(M^-)$ của hàm XNOR để thu được giá trị theo Bảng trạng thái 5.3. Với mô hình ở Hình 5.9(b), điện áp đầu vào được dùng là điện áp đơn cực, và mỗi trọng số w_{ji} sẽ được mã hóa ở 2 điện trở nhớ M^+_{ij}

và $M -_{ij}$ trên 2 cột liên kề. Lúc này, ngõ ra ở y_k của cột thứ k sẽ được thực thi theo phương trình $y_k = \sum_{i=0}^{m-1} v_i \times (\frac{1}{M+_{ik}} - \frac{1}{M-_{ik}})$ nhằm đạt được giá trị theo các trạng thái của hàm XNOR ở Bảng 5.5 [89].

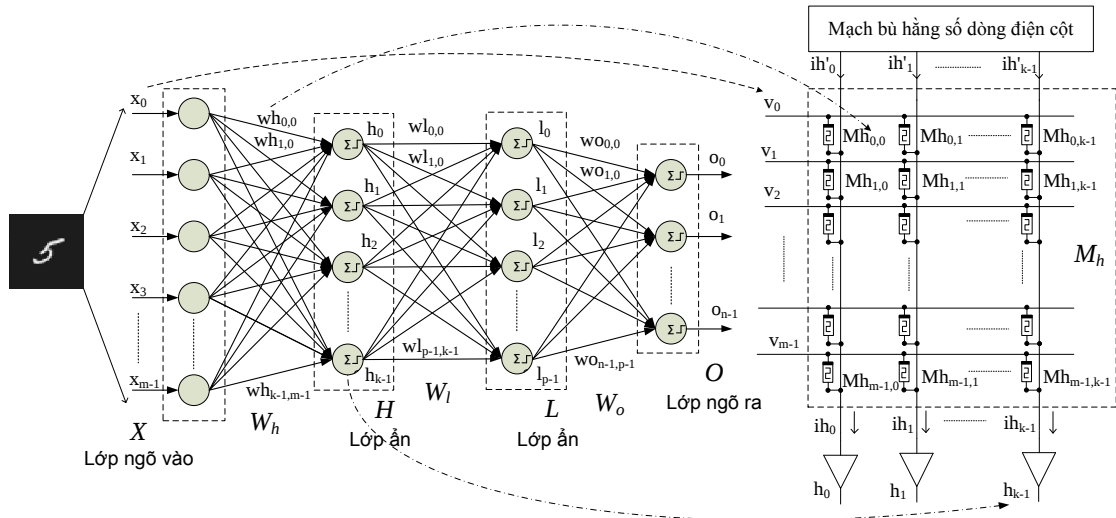
Như vậy, có thể thấy ở các mô hình truyền thống, để thực thi mạng nơ-ron nhị phân với mảng điện trở nhớ, mỗi ma trận trọng số W sẽ cần hai mảng điện trở nhớ để lưu trữ, trong đó mỗi khớp nối thần kinh với trọng số w_{ji} sẽ cần hai điện trở nhớ nhị phân $M +_{ij}$ và $M -_{ij}$ để mã hóa cho các giá trị +1 và -1. Việc sử dụng hai mảng điện trở nhớ để lưu trữ một ma trận trọng số W khi thực thi mạng nơ-ron nhị phân là một giải pháp truyền thống và chưa thực sự tối ưu vì sẽ làm tăng kích thước cũng như các hiệu ứng có liên quan của các mảng điện trở nhớ.

Ở phần trước, luận án đã thực hiện tối ưu kiến trúc mảng điện trở nhớ nhị phân để thực thi một mảng XNOR giữa ngõ vào X và một mảng điện trở nhớ duy nhất M kết hợp với mạch bù hằng số dòng điện cột theo phương trình (5.5) bên dưới:

$$\begin{aligned} Y &= \overline{X \oplus M} = (X - X') \cdot M + X' \cdot 1 \\ &= [i_0 \ i_1 \dots i_{k-1}] \end{aligned} \tag{5.5}$$

Trong đó, X là vector ngõ vào, X' là vector đảo của X và M là mảng điện trở nhớ. Việc sử dụng chỉ một mảng điện trở M và mạch bù hằng số dòng điện cột cho tham số $(X' \cdot 1)$ ở phương trình (5.5) đã giúp kiến trúc mảng điện trở nhớ tối ưu khắc phục được sự ảnh hưởng của mật độ dữ liệu khi thực thi mảng XNOR trong khi vẫn giữ được ưu điểm về kích thước mảng điện trở nhớ so với các kiến trúc khác sử dụng 2 mảng điện trở nhớ, như phân tích ở các phần trước.

Ở phần này, luận án tiếp tục thiết kế một mạng nơ-ron XNOR, ứng dụng kiến trúc mảng điện trở nhớ tối ưu để thực thi nhận dạng ảnh. Mô hình của mạng nơ-ron XNOR ứng dụng kiến trúc mảng điện trở nhớ tối ưu để thực thi nhận dạng ảnh được trình bày ở Hình 5.10.



Hình 5.10: Mô hình mạng nơ-ron XNOR ứng dụng kiến trúc mảng điện trở nhớ tối ưu đã đề xuất để thực thi nhận dạng ảnh trong hệ điện toán mô phỏng hệ thần kinh.

Như thể hiện ở Hình 5.10, mạng nơ-ron thiết kế sẽ bao gồm m nơ-ron ở lớp ngõ vào X , k nơ-ron ở lớp ẩn thứ nhất H , p nơ-ron ở lớp ẩn thứ hai L và n nơ-ron ngõ ra O để thực thi nhận dạng n ảnh nhị phân. Ma trận trọng số W_h thực hiện kết nối các nơ-ron ở lớp ngõ vào X với các nơ-ron lớp ẩn thứ nhất H , ma trận trọng số W_l thực hiện kết nối các nơ-ron lớp ẩn thứ nhất H với các nơ-ron lớp ẩn thứ hai L , và ma trận trọng số W_o để kết nối các nơ-ron lớp ẩn thứ hai L với các nơ-ron của lớp ngõ ra O . Sau khi đã được huấn luyện, ma trận trọng số W_h của lớp ẩn thứ nhất H , chứa các trọng số nhị phân $(-1, +1)$, sẽ được lưu trữ vào một mảng điện trở nhớ M_h sử dụng kiến trúc mảng điện trở tối ưu đã đề xuất, như được minh họa ở Hình 5.10. Tương tự, ma trận trọng số W_l của lớp ẩn thứ hai L và ma trận trọng số W_o của lớp ngõ ra O sau khi được huấn luyện cũng sẽ được lưu trữ vào các mảng điện trở nhớ M_l và M_o ứng dụng kiến trúc mảng điện trở nhớ tối ưu đã đề xuất trong phần trước của luận án. Như vậy, việc triển khai mô hình mạng nơ-ron nhị phân này sẽ chỉ cần ba mảng điện trở nhớ ứng dụng kiến trúc mảng điện trở nhớ tối ưu đã đề xuất ở phần 5.2.2 để lưu trữ kết quả huấn luyện của các ma trận trọng số W_h , W_l và W_o và thực thi nhận dạng.

Khi thực thi nhận dạng theo mô hình mạng nơ-ron XNOR ở Hình 5.10, ảnh cần nhận dạng sẽ được chuyển đổi thành vector đầu vào $X = [x_0 \ x_1 \ \dots \ x_{m-1}]$.

Ma trận trọng số W_h của lớp ẩn thứ nhất H sau khi huấn luyện sẽ được lưu vào mảng điện trở nhớ M_h ứng dụng kiến trúc mảng điện trở nhớ tối ưu với mạch bù hằng số dòng điện cột. Ở thiết kế này, mỗi giá trị trọng số wh_{ji} sẽ chỉ cần một điện trở nhớ để lưu trữ, giá trị trọng số -1 sẽ được mã hóa bởi trạng thái điện trở cao (HRS) và giá trị trọng số $+1$ sẽ được mã hóa ở trạng thái điện trở thấp (LRS) của điện trở nhớ Mh_{ij} tương ứng.

Tại lớp ẩn H , mỗi nơ-ron h_j cần được tính toán là tổng của các phép nhân giữa các ngõ vào x_i và trọng số wh_{ji} theo phương trình sau:

$$h_j = \sum_{i=0}^{m-1} x_i \times wh_{ji} \quad (5.6)$$

Ở đây, mỗi phép nhân $(x_i \times wh_{ji})$ khi thực thi với các giá trị nhị phân $(-1, +1)$ sẽ được triển khai bằng một phép toán XNOR giữa điện áp ngõ vào x_i và điện trở nhớ Mh_{ij} theo phương trình XNOR (5.5) và thu được các dòng điện ngõ ra cột. Nói một cách tổng quát, việc thực hiện phép nhân và cộng giữa ngõ vào X và mảng M_h lưu ma trận trọng số lớp ẩn thứ nhất sẽ được triển khai theo phương trình XNOR tối ưu (5.5) và thu được các dòng điện ngõ ra cột ih_j tại lớp ẩn H . Tiếp đến, các dòng điện ngõ ra cột này sẽ được sử dụng để nhị phân hóa các ngõ ra của lớp ẩn H nhằm đạt được Bảng trạng thái 5.5 cho các ngõ ra h_j của lớp H .

Do đó, việc thực thi mảng XNOR giữa vector ngõ vào X và mảng điện trở nhớ M_h theo phương trình (5.5) sẽ cần được thiết kế nhằm đạt được các thông số ngõ ra tương ứng với các trạng thái của phép XNOR nhị phân $(-1, +1)$ đã thể hiện ở Bảng 5.5. Chi tiết các thông số được thiết kế cho việc thực thi mảng XNOR giữa đầu vào X và mảng điện trở nhớ M_h theo phương trình (5.5) được tính toán và trình bày ở Bảng 5.6.

Như được tính toán ở Bảng 5.6, để thực thi phương trình mảng XNOR (5.5) giữa X và M_h với các giá trị nhị phân $(+1, -1)$, vector đầu vào X sẽ được chuyển hóa thành vector điện áp, với các giá trị $V+$ và $V-$ tương ứng với giá trị nhị

phân đầu vào là +1 và -1. Mạng điện trở nhớ M_h sẽ được sử dụng để lưu các trọng số của tập trọng số W_h của lớp ẩn H sau khi huấn luyện, giá trị trọng số +1 được mã hóa bằng trạng thái điện trở thấp (LRS) và giá trị trọng số -1 được mã hóa bởi trạng thái điện trở cao (HRS) của một điện trở nhớ và giả định HRS lớn hơn nhiều lần so với LRS. Trong tham số $(X' \cdot 1)$ của phương trình (5.5), giá trị 1 sẽ tương đương với giá trị trọng số +1 và được thể hiện bởi hệ số $1/\text{LRS}$.

Bảng 5.6: Thông số thiết kế để thực thi mạng XNOR giữa ngõ vào X và mạng điện trở nhớ M_h ở lớp ẩn H theo phương trình XNOR (5.5).

X	M_h	X'	$(X - X') = 2X$	M_h	$X' \cdot 1$	$(X - X') \cdot M_h + X' \cdot 1$	Y	Y
(logic)	(logic)	(logic)	(điện áp)	(điện trở)	(dòng điện)	(dòng điện)	(dòng điện)	(logic)
-1	-1	+1	$2V -$	HRS	$\frac{V +}{\text{LRS}}$	$\frac{2V -}{\text{HRS}} + \frac{V +}{\text{LRS}}$	$\sim \frac{V +}{\text{LRS}}$	+1
-1	+1	+1	$2V -$	LRS	$\frac{V +}{\text{LRS}}$	$\frac{2V -}{\text{LRS}} + \frac{V +}{\text{LRS}}$	$\frac{V -}{\text{LRS}}$	-1
+1	-1	-1	$2V +$	HRS	$\frac{V -}{\text{LRS}}$	$\frac{2V +}{\text{HRS}} + \frac{V -}{\text{LRS}}$	$\sim \frac{V -}{\text{LRS}}$	-1
+1	+1	-1	$2V +$	LRS	$\frac{V -}{\text{LRS}}$	$\frac{2V +}{\text{LRS}} + \frac{V -}{\text{LRS}}$	$\frac{V +}{\text{LRS}}$	+1

Lúc này, như kết quả ở Bảng 5.6, phần tử $(X - X')$ của phương trình (5.5) khi thực thi với giá trị nhị phân (-1, +1) ở dạng điện áp $V -$ và $V +$ sẽ tạo thành điện áp $2V -$ và $2V +$, tức là tương đương với giá trị $2X$. Hay nói một cách khác, ta có:

$$(X - X') = 2X \quad (5.7)$$

Do đó, phương trình XNOR (5.5) giữa ngõ vào X và ma trận trọng số M_h khi thực thi với giá trị nhị phân $(-1, +1)$ sẽ tương đương với:

$$\begin{aligned} Y = \overline{X \oplus M_h} &= (X - X') \cdot M_h + X' \cdot 1 = 2X \cdot M_h + X' \cdot 1 \\ &= 2(X \cdot M_h + \frac{X' \cdot 1}{2}) \end{aligned} \quad (5.8)$$

Đặt $I = \frac{Y}{2}$ thì phương trình XNOR (5.8) giữa X và M_h lúc này sẽ được biến đổi tương đương thành [85]:

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

Từ đó có thể thấy, các thông số để thực thi mảng XNOR giữa đầu vào X và mảng điện trở nhớ M_h có thể được thiết kế và tính toán theo phương trình (5.9) và vẫn đảm bảo được việc thực thi mảng XNOR như ở phương trình (5.5) với các giá trị nhị phân $(-1, +1)$. Cụ thể, các thông số của phương trình (5.9) được tính toán và trình bày ở Bảng 5.7.

Như vậy, sau khi thực thi mảng XNOR giữa ngõ vào X và mảng điện trở nhớ M_h theo phương trình (5.9), chúng ta sẽ thu được kết quả các dòng điện ngõ ra cột ih_j tại lớp ẩn H là:

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

Sau đó, các dòng điện ngõ ra cột ih_j ở lớp ẩn H sẽ được đưa qua các bộ so sánh dòng điện để kích hoạt nơ-ron h_j của lớp ẩn H thành điện áp bằng phép nhị phân hóa:

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

Bảng 5.7: Thông số thiết kế cho thực thi mảng XNOR giữa ngõ vào X và mảng điện trở nhớ M_h của lớp ẩn H theo phương trình XNOR (5.9).

X	M_h	X'	X	M_h	$\frac{(X' \cdot 1)}{2}$	$\frac{X \cdot M_h}{2} + \frac{X' \cdot 1}{2}$	I	I
(logic)	(logic)	(logic)	điện áp	điện trở	(dòng điện)	(dòng điện)	(dòng điện)	(logic)
-1	-1	+1	$V -$	HRS	$\frac{V +}{2LRS}$	$\frac{V -}{HRS} + \frac{V +}{2LRS}$	$\sim \frac{V +}{2LRS}$	+1
-1	+1	+1	$V -$	LRS	$\frac{V +}{2LRS}$	$\frac{V -}{LRS} + \frac{V +}{2LRS}$	$\frac{V -}{2LRS}$	-1
+1	-1	-1	$V +$	HRS	$\frac{V -}{2LRS}$	$\frac{V +}{HRS} + \frac{V -}{2LRS}$	$\sim \frac{V -}{2LRS}$	-1
+1	+1	-1	$V +$	LRS	$\frac{V -}{2LRS}$	$\frac{V +}{LRS} + \frac{V -}{2LRS}$	$\frac{V +}{2LRS}$	+1

Tương tự, các nơ-ron của lớp ẩn H sẽ tiếp tục thực hiện phép nhân và phép cộng với các trọng số của ma trận trọng số W_l của lớp ẩn L bằng cách thực thi phép XNOR giữa lớp ẩn H với mảng điện trở nhớ M_l (chứa tập trọng số W_l) tương tự như ở phương trình (5.9) và thu được các dòng điện ngõ ra cột. Sau đó, các dòng điện ngõ ra cột của lớp ẩn L cũng sẽ được sử dụng để kích hoạt các nơ-ron l_j của lớp ẩn L thành điện áp $V+$ hoặc $V-$, tương tự như phương trình (5.11).

Tại lớp ngõ ra O , mỗi nơ-ron ngõ ra o_j sẽ được tính toán bằng các phép cộng và phép nhân giữa các nơ-ron ẩn l_i và các trọng số wo_{ji} như sau:

$$o_j = \sum_{i=0}^{p-1} l_i \times wo_{ji} \quad (5.12)$$

Cụ thể, để tính toán các ngõ ra theo phương trình (5.12), lớp ẩn L và ma trận trọng số ngõ ra M_o sẽ thực thi phép XNOR theo phương trình (5.9) với các thông số tương tự như Bảng 5.7 và thu được các dòng điện cột io_j tại lớp ngõ ra:

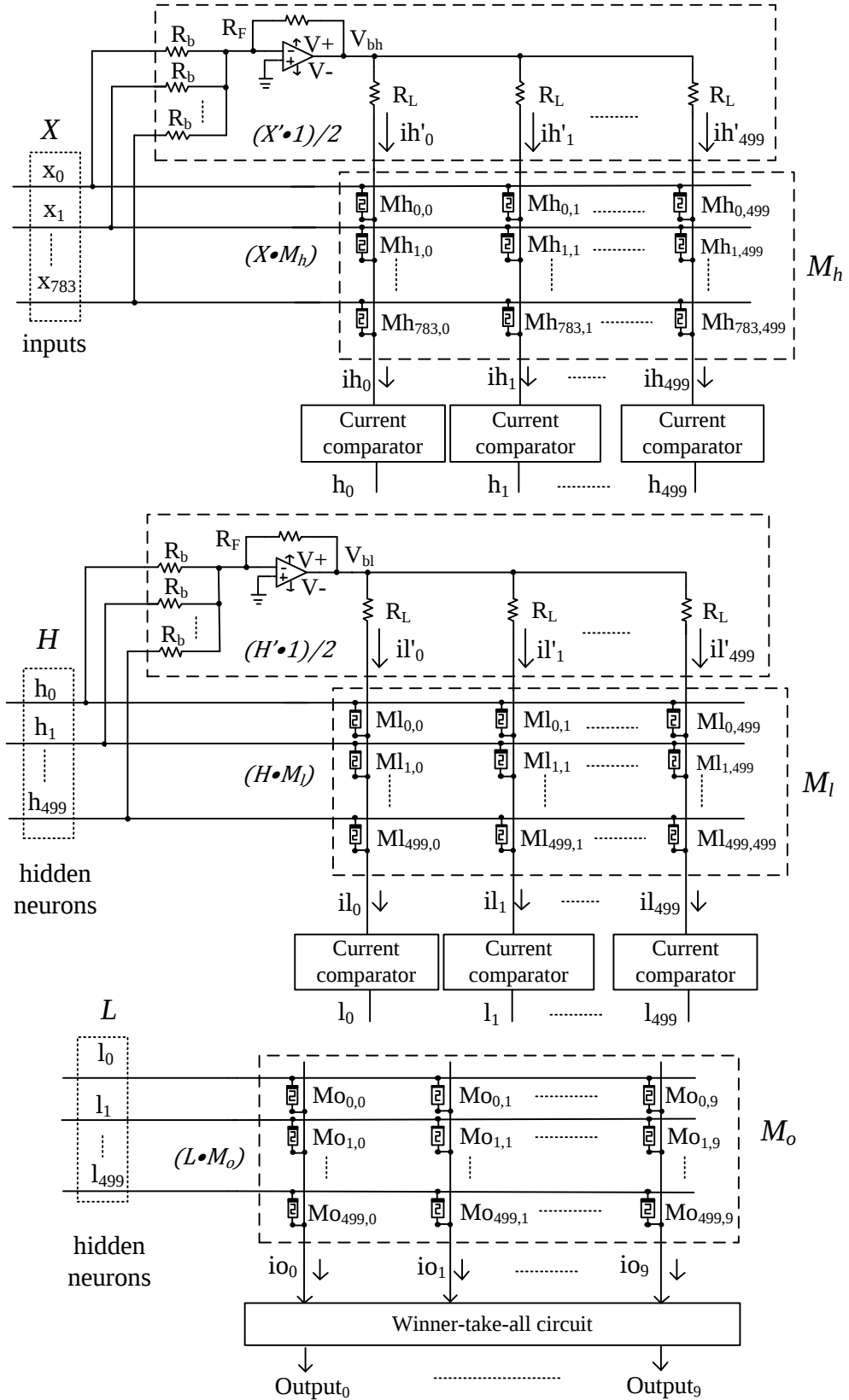
$$io_j = \sum_{i=0}^{p-1} \left(\frac{l_i}{Mo_{i,j}} + \frac{l'_i}{2LRS} \right) = \sum_{i=0}^{p-1} l_i \times \left(\frac{1}{Mo_{i,j}} - \frac{1}{2LRS} \right) \quad (5.13)$$

Các dòng điện cột này sẽ được dùng để kích hoạt ngõ ra o_j và xuất kết quả nhận dạng của mạng.

Từ mô hình mạng tổng quát ở Hình 5.10, phương trình XNOR (5.9) cho giá trị nhị phân (-1, +1) và Bảng thông số thiết kế 5.7, mạch nguyên lý của mạng nơ-ron XNOR ứng dụng kiến trúc mảng điện trở nhớ tối ưu để thực thi nhận dạng ký tự viết tay của tập dữ liệu MNIST được thiết kế và trình bày ở Hình 5.11.

Ở Hình 5.11, mạng nơ-ron XNOR được thiết kế có ngõ vào gồm 784 nơ-ron, hai lớp ẩn với 500 nơ-ron ở mỗi lớp và lớp ngõ ra 10 nơ-ron để nhận dạng 10 ký tự viết tay của tập dữ liệu MNIST. Ngõ vào X có số lượng là 784, tương ứng với kích thước của một ảnh của tập MNIST là 28×28 .

Ma trận trọng số W_h của lớp ẩn thứ nhất sau khi huấn luyện sẽ được lưu vào mảng điện trở nhớ M_h ứng dụng kiến trúc mảng điện trở nhớ tối ưu với mạch bù hằng số dòng điện cột để thực thi phương trình mảng XNOR (5.9) giữa vector ngõ vào X và mảng điện trở nhớ M_h . Mỗi giá trị trọng số wh_{ji} sẽ chỉ cần một điện trở nhớ để lưu trữ, giá trị trọng số -1 sẽ được mã hóa bởi trạng thái điện trở cao (HRS) và giá trị trọng số +1 sẽ được mã hóa bởi trạng thái điện trở thấp (LRS) của điện trở nhớ Mh_{ij} tương ứng.



Hình 5.11: Mạch nguyên lý mạng nơ-ron XNOR ứng dụng kiến trúc mảng điện trở nhớ tối ưu được đề xuất để thực thi nhận dạng ký tự viết tay sử dụng tập dữ liệu MNIST.

Các phép nhân và phép cộng giữa các ngõ vào x_i và các trọng số wh_{ji} sẽ được thực thi bằng phương trình (5.10) giữa các ngõ vào x_i và các điện trở nhớ Mh_{ij} để thu được các dòng điện ngõ ra cột ih_j của lớp ẩn H . Sau đó, các dòng điện ngõ ra cột ih_j sẽ được nhị phân hóa theo phương trình (5.11) bằng các mạch so sánh dòng điện [91] để kích hoạt nơ-ron h_j của lớp ẩn H thành điện áp V+ hoặc V-.

Tương tự, ma trận trọng số W_l của lớp ẩn thứ hai sau khi huấn luyện sẽ được lưu vào mảng điện trở nhớ M_l ứng dụng kiến trúc mảng điện trở nhớ tối ưu với mạch bù hằng số dòng điện cột để thực thi phép XNOR giữa các nơ-ron của lớp ẩn thứ nhất H và mảng điện trở nhớ M_l tương tự như phương trình mảng XNOR (5.9) để thu được các nơ-ron l_i của lớp ẩn thứ hai L .

Tại lớp ngõ ra, ma trận trọng số W_o của lớp ngõ ra sau khi huấn luyện sẽ được lưu vào mảng điện trở nhớ M_o và mạch bù hằng số dòng điện cột để thực thi hằng số $(\frac{L' \cdot 1}{2})$, theo phương trình mảng XNOR (5.9), đã được giản lược mà không làm ảnh hưởng đến kết quả ngõ ra của mạng XNOR được thiết kế. Cụ thể, phần tử hằng số $(\frac{l'_i}{2LRS})$ của phương trình (5.13) có tác dụng tạo ra một hằng số dòng điện giống nhau để cộng vào các cột thứ i của mảng M_o và việc lược bỏ phần tử hằng số này có thể tạo ra một lượng dòng điện suy giảm bằng nhau ở tất cả các cột ngõ ra. Tuy nhiên, tại lớp ngõ ra O , mạch Tìm dòng điện lớn nhất (Winner-take all circuit) sẽ chỉ so sánh các dòng điện ngõ ra cột và tìm ra dòng điện lớn nhất. Hơn nữa, giá trị l'_i trong hằng số $(\frac{l'_i}{2LRS})$ là giá trị đảo của các ngõ ra l_i của lớp ẩn L liền kề trước lớp ngõ ra O nên hằng số $(\frac{l'_i}{2LRS})$ không phụ thuộc vào mật độ dữ liệu của ảnh ngõ vào. Do vậy, việc giản lược mạch bù hằng số dòng điện cột tại lớp ngõ ra O sẽ không ảnh hưởng tới hoạt động của mạch Tìm dòng điện lớn nhất, tức là không ảnh hưởng tới hoạt động của mạng XNOR, và sẽ giúp giảm đi các mạch ngoại vi.

Trong thiết kế đề xuất ở Hình 5.11, mạch bù hằng số dòng điện cột thực thi đại lượng $\frac{(X' \cdot 1)}{2}$ của phương trình (5.9) tại lớp ản H đã được thiết kế bằng một mạch cộng đảo, để tạo thành các hằng số dòng điện cột và bù cho các cột tương ứng của mảng điện trở nhớ M_h . Lúc này, điện áp V_{bh} ở đầu ra của mạch cộng đảo sẽ có giá trị là:

$$V_{bh} = - \sum_{i=0}^{783} \frac{x_i}{R_b} \times R_F \quad (5.14)$$

và dòng điện bù cho cột thứ j của mảng điện trở nhớ M_h sẽ là:

$$ih'_j = \frac{V_{bh}}{R_L} = - \sum_{i=0}^{783} \frac{x_i}{R_b} \times \frac{R_F}{R_L} \quad (5.15)$$

Ở đây, điện trở R_b được chọn có giá trị bằng với giá trị điện trở thấp (LRS) của điện trở nhớ và chọn $R_L = 2R_F$, chúng ta sẽ thu được giá trị dòng điện bù ih'_j cho cột thứ j của mảng điện trở nhớ M_h theo phương trình (5.16) bên dưới và thực thi được phần tử $\frac{(X' \cdot 1)}{2}$ của phương trình (5.9) như đã được tính toán ở Bảng 5.7.

$$ih'_j = - \sum_{i=0}^{783} \frac{x_i}{2LRS} \quad (5.16)$$

Kết quả sau khi thực thi mảng XNOR theo phương trình (5.9) giữa vector đầu vào X và mảng điện trở nhớ M_h sẽ là vector $I_h = [ih_0 \ ih_1 \ ... \ ih_{k-1}]$ chứa các dòng điện ngõ ra cột tại lớp ản H . Cụ thể, dòng điện ngõ ra ih_j tại cột thứ j của lớp ản H sẽ được tính toán theo công thức:

$$ih_j = \sum_{i=0}^{783} \frac{x_i}{Mh_{i,j}} - \sum_{i=0}^{783} \frac{x_i}{2LRS} = \sum_{i=0}^{783} x_i \times \left(\frac{1}{Mh_{i,j}} - \frac{1}{2LRS} \right) \quad (5.17)$$

Ở đầu ra của lớp ẩn H , mỗi dòng điện ngõ ra cột ih_j sẽ được đưa qua một mạch so sánh dòng điện [91] để thực thi kích hoạt nhị phân cho nơ-ron h_j của lớp ẩn H theo phương trình (5.10) thành điện áp $V +$ hoặc $V -$, tương ứng với giá trị +1 hoặc -1.

Sau đó, các nơ-ron của lớp ẩn H sẽ tiếp tục được đưa vào mảng điện trở nhớ M_l , nơi đã lưu ma trận trọng số W_l của lớp ẩn thứ hai L sau huấn luyện, và mạch bù hằng số dòng điện cột để thực thi mảng XNOR giữa vector H và mảng điện trở nhớ M_l theo phương trình:

$$I_l = \overline{H \oplus M_l} = H \cdot M_l + \frac{H' \cdot 1}{2} = [il_0 \ il_1 \ ... \ il_{499}] \quad (5.18)$$

Kết quả thu được là vector I_l chứa các dòng điện ngõ ra cột ở lớp ẩn thứ hai L . Mỗi dòng điện il_j ở cột thứ j của lớp ẩn L sẽ có giá trị bằng:

$$il_j = \sum_{i=0}^{499} \left(\frac{h_i}{Ml_{i,j}} + \frac{h'_i}{2LRS} \right) = \sum_{i=0}^{499} h_i \times \left(\frac{1}{Ml_{i,j}} - \frac{1}{2LRS} \right) \quad (5.19)$$

Các dòng điện ngõ ra cột il_j ở lớp ẩn thứ hai L cũng sẽ được đưa qua các mạch so sánh dòng điện [91] để thực thi kích hoạt nhị phân cho các nơ-ron l_j của lớp ẩn L thành điện áp $V +$ hoặc $V -$, tương ứng với giá trị +1 hoặc -1.

Cuối cùng, các nơ-ron của lớp ẩn L sẽ tiếp tục được đưa vào mảng điện trở nhớ M_o , nơi đã lưu ma trận trọng số W_o của lớp ngõ ra sau huấn luyện để thực thi mảng XNOR giữa vector L và mảng điện trở nhớ M_o theo phương trình:

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

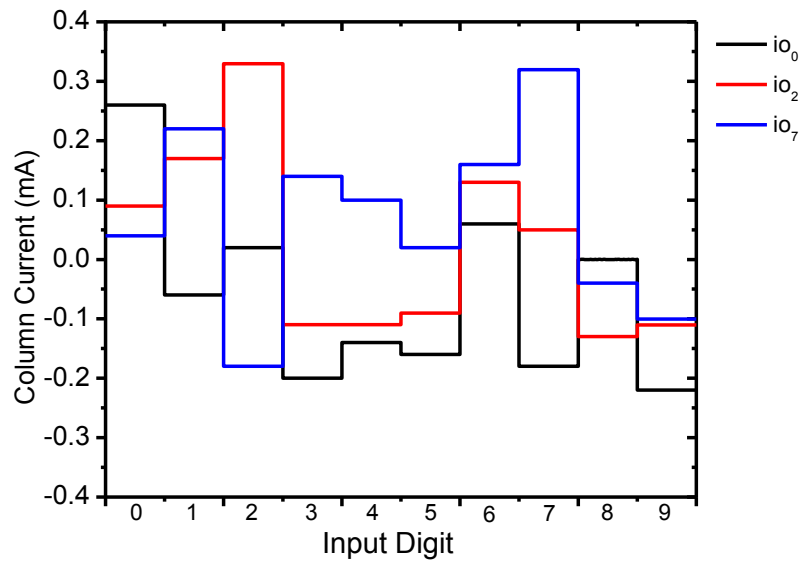
Ở đây, đại lượng hằng số $(\frac{L' \cdot 1}{2})$ là không phụ thuộc vào mật độ dữ liệu của ảnh cần nhận dạng nên đã được giản lược. Mặc dù việc giản lược hằng số $(\frac{L' \cdot 1}{2})$ có thể gây ra một lượng suy giảm dòng điện bằng nhau ở các dòng điện ngõ ra

cột io_j của mảng điện trở nhớ M_O , nhưng sẽ không gây ảnh hưởng tới hoạt động xác định dòng điện lớn nhất của mạch Tìm dòng điện lớn nhất ở phía sau. Do vậy, tại lớp ngõ ra O trong thiết kế ở Hình 5.11, mạch bù hằng số dòng điện cột đã được giản lược, giúp giảm các mạch ngoại vi và không ảnh hưởng đến hoạt động nhận dạng của mạng XNOR đã thiết kế.

Tại lớp ngõ ra của mạng nơ-ron, mạch Tìm dòng điện lớn nhất (Winner-take-all) sẽ thực thi vai trò của hàm kích hoạt ngõ ra. Các dòng điện ngõ ra cột $I_O = [io_0 \ io_1 \ ... \ io_9]$ tại lớp ngõ ra sẽ làm cho các tụ điện nạp trước, từ Co_0 đến Co_9 , xả với tốc độ khác nhau. Mạch Tìm dòng điện lớn nhất sẽ căn cứ vào tốc độ xả của các tụ điện để tìm ra tụ xả nhanh nhất Co_j , tức tìm ra dòng điện ngõ ra cột lớn nhất io_j , để xuất ra kết quả nhận dạng của mạng ở cột thứ j , như đã phân tích và đánh giá ở phần trước.

Tiếp theo, mạch đề xuất sẽ được đánh giá sử dụng phần mềm mô phỏng mạch Cadence Spectre. Mạng nơ-ron XNOR đề xuất được huấn luyện trên máy tính. Các trọng số của các ma trận trọng số sau khi huấn luyện sẽ được mã hóa vào các điện trở nhớ của mảng điện trở nhớ, giá trị trọng số -1 và +1 lần lượt được mã hóa bằng giá trị điện trở cao (HRS) và điện trở thấp (LRS). Kết quả mạng nơ-ron XNOR thiết kế đạt được tỉ lệ chính xác 94% trên tập dữ liệu MNIST.

Hình 5.12 thể hiện kết quả mô phỏng mạch của mạng nơ-ron XNOR đề xuất trên phần mềm mô phỏng mạch Cadence Spectre. Các giá trị ngõ vào -1 và +1 được thực hiện bằng điện áp -1 V và +1 V. Các trạng thái điện trở cao (HRS) và điện trở thấp (LRS) của các điện trở nhớ tương ứng với các trọng số -1 và +1 có giá trị lần lượt là 10 M Ω và 100 K Ω . Các mảng điện trở nhớ được thiết kế bằng mô hình VerilogA và các mạch ngoại vi được thiết kế với công nghệ CMOS 130nm. Các dòng điện ngõ ra cột io_0 , io_2 và io_7 tại lớp ngõ ra O đã được đo đạc 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.



Hình 5.12: Các dòng điện ngõ ra cột i_{o_0} , i_{o_2} và i_{o_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.

Như có thể thấy ở Hình 5.12, khi ký số “0” được đưa vào ngõ vào của mạng, thì dòng điện ngõ ra i_{o_0} của cột thứ 0 là dòng điện đạt giá trị lớn nhất trong các dòng điện ngõ ra cột. Cụ thể, lúc này dòng điện i_{o_0} đạt giá trị 0,26 mA trong khi dòng điện i_{o_2} và i_{o_7} chỉ đạt giá trị xấp xỉ 0,09 mA và 0,04 mA. Điều này cho thấy mạng nơ-ron XNOR thiết kế đã hoạt động đúng khi nhận dạng ký số “0” bằng việc tạo ra dòng điện ngõ ra i_{o_0} ở cột thứ 0 là dòng điện lớn nhất trong các cột ngõ ra. Tương tự, khi ký số “2” được đưa vào nhận dạng thì dòng điện ngõ ra i_{o_2} của cột thứ 2, thể hiện bằng đường màu đỏ, đạt giá trị lớn nhất trong số các cột ngõ ra, ở mức xấp xỉ 0,33 mA. Khi ký số “7” được đưa vào nhận dạng, dòng điện ngõ ra i_{o_7} của cột thứ 7 đạt giá trị lớn nhất trong các cột ngõ ra, ở mức xấp xỉ 0,32 mA. Các kết quả dòng điện ngõ ra cột được đo đạc và phân tích tại lớp ngõ ra O đã chứng minh cho tính chính xác trong hoạt động của mạng nơ-ron XNOR ứng dụng kiến trúc mảng điện trở nhớ tối ưu đề xuất, bằng việc tạo ra dòng điện ngõ ra lớn nhất ở cột thứ j tương ứng khi ký số “ j ” được đưa vào ngõ vào để nhận dạng.

Như vậy, ở phần này, luận án đã thiết kế một mạng nơ-ron XNOR với kích thước $784 \times 500 \times 500 \times 10$ ứng dụng kiến trúc mảng điện trở nhớ tối ưu cho bài toán nhận dạng ký tự viết tay sử dụng tập dữ liệu MNIST. Các ma trận trọng số W_h , W_l của các lớp ẩn sau khi được huấn luyện sẽ được mã hóa vào các mảng điện trở nhớ M_h và M_l với mạch bù hằng số dòng điện cột theo kiến trúc mảng điện trở nhớ đã được tối ưu hóa ở phần 5.3.2. Ma trận trọng số W_o của lớp ngõ ra được mã hóa vào mảng điện trở nhớ M_o và giản lược mạch bù hằng số dòng điện cột để giảm các mạch ngoại vi mà không ảnh hưởng tới hoạt động nhận dạng của mạng. Với thiết kế được đề xuất ở luận án, mỗi ma trận trọng số W của mạng nơ-ron nhị phân chỉ sử dụng một mảng điện trở nhớ M để lưu trữ giá trị và thực thi mảng XNOR với các nơ-ron của lớp phía trước theo phương trình XNOR (5.9). Thiết kế này của luận án sẽ giúp giảm 50% kích thước của các mảng điện trở nhớ, khi so sánh với thiết kế truyền thống của mạng nơ-ron nhị phân với điện trở nhớ vốn sử dụng 2 mảng điện trở nhớ để lưu trữ một ma trận trọng số W . Một đóng góp khác của thiết kế mạng XNOR đề xuất là sự giản lược mạch bù hằng số dòng điện cột ở lớp ngõ ra O , là thiết kế cho phép sử dụng một mảng điện trở nhớ duy nhất mà không cần mạch ngoại vi khi sử dụng kết hợp với mạch Tìm dòng điện lớn nhất ở lớp ngõ ra cuối cùng của mạng nơ-ron XNOR. Điều này sẽ giúp giảm được mạch ngoại vi cho mạng nơ-ron XNOR. Mạng nơ-ron XNOR đề xuất ở luận án sẽ là tiền đề và cơ sở cho các nghiên cứu chuyên sâu hơn về thực thi các mạng nơ-ron học sâu ứng dụng kiến trúc mảng điện trở nhớ trong các hệ điện toán mô phỏng hệ thần kinh.

5.3 KẾT LUẬN CHƯƠNG

Như vậy, ở chương này, luận án đã thực hiện tối ưu cho kiến trúc mảng 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, là điều mà nghiên cứu trước đó chưa giải quyết được. Cụ thể, một mạch bù hằng số dòng điện cột để thực thi tham số $(A' + 1)$ trong khai triển phương trình toán học đầy đủ của mảng XNOR đã được luận án thiết kế và đề xuất; sau đó, luận án cũng đã thiết kế mạch

tối ưu cho kiến trúc mảng điện trở nhớ nhị phân để thực thi phương trình toán học đầy đủ của mảng XNOR với một mảng điện trở nhớ duy nhất để nhận dạng 10 ảnh nhị phân. Kiến trúc mảng điện trở nhớ tối ưu đề xuất đã khắc phục được sự suy giảm của dòng điện ngõ ra cột gây ra bởi sự suy giảm của mật độ dữ liệu và ổn định được hoạt động nhận dạng của hệ thống. Đồng thời, luận án cũng đề xuất một thiết kế mạng nơ-ron XNOR ứng dụng kiến trúc mảng điện trở nhớ đã tối ưu cho bài toán nhận dạng ký tự viết tay của tập dữ liệu MNIST. Thiết kế tối ưu kiến trúc mảng điện trở nhớ nhị phân và thiết kế mạng nơ-ron XNOR ứng dụng kiến trúc mảng đề xuất ở chương này là đóng góp mới và quan trọng của luận án trong thiết kế các hệ điện toán mô phỏng hệ thần kinh thực thi nhận dạng với mảng điện trở nhớ nhị phân.

CHƯƠNG 6

KẾT LUẬN

6.1 KẾT QUẢ ĐẠT ĐƯỢC CỦA LUẬN ÁN

Luận án đã trình bày chi tiết nội dung thực hiện của đề tài trong nghiên cứu tối ưu kiến trúc mạng điện trở nhớ nhị phân ứng dụng trong thiết kế hệ điện toán mô phỏng hệ thần kinh thực thi nhận dạng ảnh. Các kết quả nghiên cứu cho thấy luận án đã hoàn thành được các mục tiêu đặt ra của đề tài với các kết quả và đóng góp mới 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 trong thiết kế hệ điện toán mô phỏng hệ thần kinh thực thi ứng dụng nhận dạng ảnh. Thông qua việc phân tích mô hình toán học và cấu trúc của các kiến trúc mạng điện trở nhớ nhị phân, luận án đã thực thi một ứng dụng nhận dạng 10 ảnh xám với các kiến trúc mạng điện trở nhớ bao gồm kiến trúc mạng bù, kiến trúc mạng đồng nhất và kiến trúc mạng đơn để phân tích sự ảnh hưởng của nhiễu và sự biến thiên điện trở mạng. Với các kết quả nghiên cứu trình bày ở Chương 3 và bài báo công bố, luận án đã tìm ra được ưu điểm về đáp ứng tốt với nhiễu dữ liệu và với sự biến thiên điện trở, bên cạnh lợi thế về kích thước mạng của kiến trúc mạng điện trở nhớ đơn so với các kiến trúc còn lại. Kết quả này cho thấy việc tiếp tục nghiên cứu chuyên sâu và tối ưu cho mạng 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 trong thiết kế các hệ điện toán mô phỏng hệ thần kinh thực thi nhận dạng ảnh.

- Hai là, luận án đã phát hiện được sự ảnh hưởng không tích cực và cũng tìm ra được nguyên nhân gây ra sự ảnh hưởng của mật độ dữ liệu đến hoạt động của kiến trúc mảng điện trở nhớ đơn khi thực thi mảng XNOR để nhận dạng. Cụ thể, bằng việc phân tích và đánh giá sự ảnh hưởng của mật độ dữ liệu đến hoạt động nhận dạng ảnh của các kiến trúc mảng điện trở nhớ nhị phân, luận án đã phát hiện ra rằng sự suy giảm của mật độ dữ liệu đã gây ra sự suy giảm đáng kể dòng điện ngõ ra cột của kiến trúc mảng điện trở nhớ đơn và làm cho mạch Tìm dòng điện lớn nhất không thể hoạt động đúng. Điều này sẽ làm giảm hiệu quả nhận dạng của kiến trúc mảng điện trở nhớ đơn so với kiến trúc mảng điện trở nhớ bù. Đồng thời, luận án cũng đã tìm 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 ở kiến trúc mảng điện trở nhớ đơn. Với các kết quả phân tích được trình bày ở Chương 4 và bài báo công bố, luận án đã có một phát hiện mới và quan trọng cho thấy yêu cầu cấp thiết của việc tối ưu hóa kiến trúc mảng điện trở nhớ nhị phân nhằm loại bỏ ảnh hưởng của mật độ dữ liệu khi thực thi mảng XNOR.
- Ba là, luận án đã thực hiện nghiên cứu tối ưu kiến trúc mảng điện trở nhớ nhị phân trong ứng dụng nhận dạng ảnh để khắc phục ảnh hưởng của mật độ dữ liệu, đảm bảo sự ổn định của hoạt động nhận dạng mà không phụ thuộc vào tập dữ liệu trong khi vẫn sử dụng một mảng điện trở nhớ duy nhất để thực thi mảng XNOR. Cụ thể, luận án đã thiết kế một mạch bù hằng số dòng điện cột để thực thi thành phần $(A' \cdot 1)$ trong khai triển phương trình toán học của mảng XNOR để ổn định các dòng điện ngõ ra cột và loại bỏ sự ảnh hưởng của sự suy giảm mật độ dữ liệu. Từ đó, luận án nghiên cứu đề xuất một thiết kế mạch tối ưu cho kiến trúc mảng điện trở nhớ nhị phân để thực thi phương trình toán học đầy đủ của mảng XNOR chỉ với một mảng điện trở nhớ duy nhất khi nhận dạng ảnh nhằm khắc phục được ảnh hưởng của mật độ dữ liệu đến hoạt động của hệ thống. Các kết quả phân tích ở Chương 5 và bài báo công bố đã cho thấy

với thiết kế mạch tối ưu cho kiến trúc mảng điện trở nhớ nhị phân đã đề xuất của luận án, các dòng điện ngõ ra cột của mảng điện trở nhớ đã được đảm bảo ổn định dưới sự suy giảm của mật độ dữ liệu, đảm bảo mạch tìm dòng điện ngõ ra lớn nhất hoạt động chính xác và tạo ra được xung ngõ ra cho kết quả nhận dạng sau khoảng thời gian 7 ns. Đồng thời, luận án cũng nghiên cứu đề xuất một thiết kế mạng nơ-ron XNOR ứng dụng kiến trúc mảng điện trở nhớ nhị phân tối ưu cho bài toán nhận dạng ký tự viết tay sử dụng tập dữ liệu MNIST. Trong thiết kế mạng nơ-ron XNOR đề xuất, mạch tạo ra hằng số dòng của lớp ngõ ra đã được lược bỏ khi lớp ngõ ra sử dụng mạch Tìm dòng điện lớn nhất (Winner-Take-All) mà không làm ảnh hưởng tới tính chính xác trong hoạt động của mạng nhằm tối ưu về kích thước và công suất mạch. Đây cũng chính là một cải tiến, thông qua phân tích để tối ưu hóa kiến trúc mạch, sử dụng kết hợp giữa thiết kế có sử dụng mạch tạo dòng bù ở các lớp ẩn và thiết kế không có mạch tạo dòng bù ở lớp ngõ ra. Mạng nơ-ron XNOR thiết kế có độ chính xác 94% trên tập dữ liệu MNIST. Kết quả mô phỏng mạch cho thấy mạng nơ-ron XNOR dựa trên mảng điện trở nhớ đề xuất có đáp ứng ngõ ra chính xác khi các ký tự từ 0 – 9 được đưa vào ngõ vào của mạng.

Về ý nghĩa khoa học, các kết quả nghiên cứu của luận án đã góp phần trong việc phát triển các kiến trúc mảng điện trở nhớ tối ưu cho việc thực thi các hệ điện toán mô phỏng hệ thần kinh trên phần cứng. Với nhiều tiềm năng về sự phát triển công nghệ của điện trở nhớ, các phần cứng trên nền tảng các kiến trúc mảng điện trở nhớ tối ưu sẽ là giải pháp thay thế hiệu quả cho các thiết kế mạch truyền thống dựa trên công nghệ CMOS vốn đã tiệm cận đến giới hạn của sự phát triển. Về khả năng ứng dụng, khi các mảng điện trở nhớ được chế tạo rộng rãi và thương mại hóa, kiến trúc mảng điện trở nhớ đề xuất của luận án có thể được ứng dụng để triển khai các mô hình hệ điện toán mô phỏng hệ thần kinh trên các thiết bị xử lý tại biên nhằm đạt được yêu cầu tối ưu về công suất và tài nguyên phần cứng, có thể thay thế cho giải pháp triển khai các nút xử lý tại biên với các mạng nơ-ron nhân tạo xử lý bằng phần mềm trên nền tảng của CPU hoặc GPU.

6.2 HƯỚNG PHÁT TRIỂN

Kết quả nghiên cứu của luận án đã cho thấy kiến trúc mảng điện trở nhớ tối ưu đề xuất đã có thể thực hiện được phương trình toán học đầy đủ của mảng XNOR để nhận dạng chỉ với một mảng điện trở nhớ duy nhất, khắc phục được ảnh hưởng của sự suy giảm mật độ dữ liệu và có đáp ứng tốt với nhiễu và sự biến thiên điện trở. Luận án cũng đã đề xuất một thiết kế mạng nơ-ron XNOR ứng dụng kiến trúc mảng điện trở nhớ tối ưu đề xuất cho bài toán nhận dạng ký tự viết tay sử dụng tập dữ liệu MNIST. Trên cơ sở các kết quả này, các nội dung nghiên cứu của luận án có thể được phát triển hoàn thiện hơn trong tương lai theo các hướng sau:

- Thứ nhất, tiếp tục các nghiên cứu nâng cao hơn cho kiến trúc mảng điện trở nhớ đã đề xuất của luận án nhằm giải quyết các thách thức khác trong ứng dụng mảng điện trở nhớ thực thi hệ điện toán mô phỏng hệ thần kinh. Việc triển khai thực tế của các thiết kế mô phỏng hệ thần kinh trên cơ sở các mảng điện trở nhớ vẫn còn phải đối mặt với nhiều giới hạn kỹ thuật và thách thức như: ảnh hưởng của đặc tính phi tuyến của điện trở nhớ, sai số trong quá trình chế tạo linh kiện, lỗi phân tử mảng, dòng rò trong mảng, sự biến thiên linh kiện điện trở nhớ. Do đó, kiến trúc mảng điện trở nhớ đã đề xuất có thể được nghiên cứu chuyên sâu hơn nữa, để tiếp tục giải quyết các thách thức còn lại nhằm hoàn thiện tính hiệu quả và khả thi của các thiết kế mô phỏng hệ điện toán thần kinh. Cụ thể, tiếp tục nghiên cứu các kỹ thuật giảm sai số do dòng rò (sneak-path leakage) trong mảng, cải thiện ảnh hưởng do tỷ lệ HRS/LRS động, tối ưu hoá kiến trúc mảng, phát triển các thuật toán hiệu chỉnh (calibration) và bù ở cấp độ mạch phân cứng (circuit-level compensation) để bù trừ sai số phi tuyến của linh kiện điện trở nhớ. Đồng thời, có thể thiết kế đa dạng hơn nữa các ứng dụng của mảng điện trở nhớ đã đề xuất, cụ thể như trong các mạng nơ-ron nhị phân/tam phân, mạng nhiều lớp cho ứng dụng nhận dạng ảnh trên các tập dữ liệu đa dạng và phức tạp hơn như CIFAR10, ImageNet ..., bên cạnh tập dữ liệu MNIST.

- 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 các mạng nơ-ron tích chập (CNN) nhị phân cho các bài toán nhận dạng, ứng dụng trên các tập dữ liệu lớn hơn. Kiến trúc mảng điện trở nhớ tối ưu nhằm thực thi phương trình toán học đầy đủ của mảng XNOR được đề xuất ở luận án là một tiền đề tốt để nghiên cứu thiết kế giải pháp phần cứng để thực thi các phép nhân chập trong mạng CNN với giá trị nhị phân, giúp làm giảm yêu cầu về kích thước bộ nhớ lưu trữ và tăng tốc độ thực thi của mạng. Các thiết kế mạng nơ-ron tích chập nhị phân trên nền tảng của kiến trúc mảng điện trở nhớ nhị phân tối ưu sẽ cung cấp thêm giải pháp phần cứng hiệu quả cho các thiết kế hệ điện toán mô phỏng hệ thần kinh ở các bài toán nhận dạng, đặc biệt trong các nút xử lý tại biên của các ứng dụng Internet vạn vật.
- Thứ ba, tiếp tục nghiên cứu hoàn thiện thiết kế cho các mạch ngoại vi để có thể chế tạo một vi mạch thực thi hệ điện toán mô phỏng hệ thần kinh phục vụ cho các ứng dụng học sâu. Điều này sẽ tạo được nền tảng tốt cho các nghiên cứu sâu hơn trong thực thi các ứng dụng trí tuệ nhân tạo trên phần ứng vi mạch, phù hợp với xu hướng đang được các công ty vi mạch bán dẫn đang đầu tư phát triển.

TÀI LIỆU THAM KHẢO

- [1] C. Mead, "Neuromorphic electronic systems," *Proceedings of the IEEE*, vol. 78, pp. 1629-1636, 1990.
- [2] G. Indiveri, E. Chicca, and R. Douglas, "A VLSI array of low-power spiking neurons and bistable synapses with spike-timing dependent plasticity," *IEEE Transactions on Neural Networks*, vol. 17, pp. 211-221, 2006.
- [3] C. Mayr, J. Partzsch, M. Noack, S. Hänzsch, S. Scholze, S. Höppner, *et al.*, "A Biological-Realtime Neuromorphic System in 28 nm CMOS Using Low-Leakage Switched Capacitor Circuits," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 10, pp. 243-254, February 2015.
- [4] H. Jiang, J. Lu, C. Zhang, S. Tang, J. An, L. Cheng, *et al.*, "Multicore Spiking Neuromorphic Chip in 180-nm With ReRAM Synapses and Digital Neurons," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 13, pp. 975-985, December 2023.
- [5] T. Wang, M. Tian, H. Wang, Z. Zhong, J. He, F. Tang, *et al.*, "MorphBungee: A 65-nm 7.2-mm² 27- μ J/Image Digital Edge Neuromorphic Chip With on-Chip 802-Frame/s Multi-Layer Spiking Neural Network Learning," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 19, pp. 209-225, February 2025.
- [6] Y. Taur, "CMOS design near the limit of scaling," *IBM Journal of Research and Development*, vol. 46, pp. 213-222, 2002.
- [7] T. Pesic-Brdanin and B. Dokić, "Strained silicon layer in CMOS technology," *Electronics*, vol. 18, pp. 63-69, December 2014.
- [8] F. Schwierz and J. J. Liou, "Status and Future Prospects of CMOS Scaling and Moore's Law - A Personal Perspective," in *2020 IEEE Latin America Electron Devices Conference (LAEDC)*, San Jose, Costa Rica, 2020, pp. 1-4.
- [9] L. Chua, "Memristor-The missing circuit element," *IEEE Transactions on Circuit Theory*, vol. 18, pp. 507-519, 1971.
- [10] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams, "The missing memristor found," *Nature*, vol. 453, pp. 80-83, May 2008.

- [11] S. Jo, T. Chang, I. Ebong, B. Bhadviya, P. Mazumder, and W. Lu, "Nanoscale Memristor Device as Synapse in Neuromorphic Systems," *Nano letters*, vol. 10, pp. 1297-301, March 2010.
- [12] I. Boybat, M. Le Gallo, S. R. Nandakumar, T. Moraitis, T. Parnell, T. Tuma, *et al.*, "Neuromorphic computing with multi-memristive synapses," *Nature Communications*, vol. 9, p. 2514, June 2018.
- [13] R. Wang, W. Zhang, S. Wang, T. Zeng, X. Ma, H. Wang, *et al.*, "Memristor-Based Signal Processing for Compressed Sensing," *Nanomaterials*, vol. 13, p. 1354, April 2023.
- [14] R. S. Williams, "How We Found The Missing Memristor," *IEEE Spectrum*, vol. 45, pp. 28-35, 2008.
- [15] H. Kim, M. R. Mahmoodi, H. Nili, and D. B. Strukov, "4K-memristor analog-grade passive crossbar circuit," *Nature Communications*, vol. 12, p. 5198, August 2021.
- [16] K. Jeon, J. J. Ryu, S. Im, H. K. Seo, T. Eom, H. Ju, *et al.*, "Purely self-rectifying memristor-based passive crossbar array for artificial neural network accelerators," *Nature Communications*, vol. 15, p. 129, January 2024.
- [17] J. J. Yang, D. B. Strukov, and D. R. Stewart, "Memristive devices for computing," *Nature Nanotechnology*, vol. 8, pp. 13-24, Jan. 2013.
- [18] K. H. Kim, M.-J. Seo, and B. C. Jang, "Amorphous ITZO-Based Selector Device for Memristor Crossbar Array," *Micromachines*, vol. 14, p. 506, February 2023.
- [19] F. Aguirre, A. Sebastian, M. Le Gallo, W. Song, T. Wang, J. J. Yang, *et al.*, "Hardware implementation of memristor-based artificial neural networks," *Nature Communications*, vol. 15, p. 1974, March 2024.
- [20] J. Chen, Q. Xu, Y. Li, J. Cao, X. Liu, J. Qiu, *et al.*, "A Fully Printed ZnO Memristor Synaptic Array for Neuromorphic Computing Application," *IEEE Electron Device Letters*, vol. 45, pp. 1076-1079, April 2024.
- [21] S. Jain, S. Li, H. Zheng, L. Li, X. Fong, and K.-W. Ang, "Heterogeneous integration of 2D memristor arrays and silicon selectors for compute-in-memory hardware in convolutional neural networks," *Nature Communications*, vol. 16, p. 2719, March 2025.

- [22] C. K  geler, M. Meier, R. Rosezin, S. Gilles, and R. Waser, "High density 3D memory architecture based on the resistive switching effect," *Solid-State Electronics*, vol. 53, pp. 1287-1292, Dec. 2009.
- [23] M. M. Shulaker, T. F. Wu, A. Pal, L. Zhao, Y. Nishi, K. Saraswat, *et al.*, "Monolithic 3D integration of logic and memory: Carbon nanotube FETs, resistive RAM, and silicon FETs," in *2014 IEEE International Electron Devices Meeting*, San Francisco, CA, USA, 2014, pp. 27.4.1-27.4.4.
- [24] P. Lin, C. Li, Z. Wang, Y. Li, H. Jiang, W. Song, *et al.*, "Three-dimensional memristor circuits as complex neural networks," *Nature Electronics*, vol. 3, pp. 225-232, April 2020.
- [25] T.-Y. Wang, J.-L. Meng, L. Chen, H. Zhu, Q.-Q. Sun, S.-J. Ding, *et al.*, "Flexible 3D memristor array for binary storage and multi-states neuromorphic computing applications," *InfoMat*, vol. 3, pp. 212-221, February 2021.
- [26] Y. S. An, D. Kim, Y. R. Park, J. S. Eo, M. Kim, D. Kim, *et al.*, "Implementation of monolithic 3D integrated TiOx memristor-based neural network for high-performance in-memory computing," *Nano Energy*, vol. 139, p. 110999, June 2025.
- [27] S. N. Truong and K.-S. Min, "New Memristor-Based Crossbar Array Architecture with 50-% Area Reduction and 48-% Power Saving for Matrix-Vector Multiplication of Analog Neuromorphic Computing," *JSTS:Journal of Semiconductor Technology and Science*, vol. 14, pp. 356-363, Jun. 2014.
- [28] S. Y. Sun, H. Xu, J. Li, Q. Li, and H. Liu, "Cascaded Architecture for Memristor Crossbar Array Based Larger-Scale Neuromorphic Computing," *IEEE Access*, vol. 7, pp. 61679-61688, May 2019.
- [29] S. Wen, S. Xiao, Y. Yang, Z. Yan, Z. Zeng, and T. Huang, "Adjusting Learning Rate of Memristor-Based Multilayer Neural Networks via Fuzzy Method," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 38, pp. 1084-1094, Jun. 2019.
- [30] S. Wen, H. Wei, Z. Yan, Z. Guo, Y. Yang, T. Huang, *et al.*, "Memristor-Based Design of Sparse Compact Convolutional Neural Network," *IEEE Transactions on Network Science and Engineering*, vol. 7, pp. 1431-1440, July 2020.

- [31] J. Fu, Z. Liao, and J. Wang, "Memristor-Based Neuromorphic Hardware Improvement for Privacy-Preserving ANN," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 27, pp. 2745-2754, Jul. 2019.
- [32] Z. Wang, C. Li, W. Song, M. Rao, D. Belkin, Y. Li, *et al.*, "Reinforcement learning with analogue memristor arrays," *Nature Electronics*, vol. 2, pp. 115-124, Mar. 2019.
- [33] O. Krestinskaya, K. N. Salama, and A. P. James, "Learning in Memristive Neural Network Architectures Using Analog Backpropagation Circuits," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 66, pp. 719-732, February 2019.
- [34] Y. Jeong and W. Lu, "Neuromorphic Computing Using Memristor Crossbar Networks: A Focus on Bio-Inspired Approaches," *IEEE Nanotechnology Magazine*, vol. 12, pp. 6-18, Jul. 2018.
- [35] X. Zhang, A. Huang, Q. Hu, Z. Xiao, and P. Chu, "Neuromorphic Computing with Memristor Crossbar," *Physica Status Solidi (A)*, vol. 215, p. 1700875, May 2018.
- [36] S. N. Truong, K. V. Pham, W. Yang, S. Shin, K. Pedrotti, and K.-S. Min, "New pulse amplitude modulation for fine tuning of memristor synapses," *Microelectronics Journal*, vol. 55, pp. 162-168, Aug. 2016.
- [37] D. Soudry, D. D. Castro, A. Gal, A. Kolodny, and S. Kvatinsky, "Memristor-Based Multilayer Neural Networks With Online Gradient Descent Training," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, pp. 2408-2421, Jan. 2015.
- [38] D. Niu, Y. Chen, C. Xu, and Y. Xie, "Impact of process variations on emerging memristor," in *47th Design Automation Conference*, Anaheim, CA, USA, 2010, pp. 877-882.
- [39] M. Hu, H. Li, Y. Chen, X. Wang, and R. E. Pino, "Geometry variations analysis of TiO₂ thin-film and spintronic memristors," in *16th Asia and South Pacific Design Automation Conference (ASP-DAC 2011)*, Yokohama, Japan, 2011, pp. 25-30.
- [40] J. Rajendran, H. Maenm, R. Karri, and G. S. Rose, "An Approach to Tolerate Process Related Variations in Memristor-Based Applications," in *2011 24th International Conference on VLSI Design*, Chennai, India, 2011, pp. 18-23.

- [41] J. Rajendran, R. Karri, and G. S. Rose, "Improving Tolerance to Variations in Memristor-Based Applications Using Parallel Memristors," *IEEE Transactions on Computers*, vol. 64, pp. 733-746, 2015.
- [42] D. Ielmini, S. Lavizzari, D. Sharma, and A. L. Lacaita, "Physical interpretation, modeling and impact on phase change memory (PCM) reliability of resistance drift due to chalcogenide structural relaxation," in *2007 IEEE International Electron Devices Meeting*, Washington, DC, USA, 2007, pp. 939-942.
- [43] M. Suri, D. Garbin, O. Bichler, D. Querlioz, D. Vuillaume, C. Gamrat, *et al.*, "Impact of PCM resistance-drift in neuromorphic systems and drift-mitigation strategy," in *2013 IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH)*, Brooklyn, NY, USA, 2013, pp. 140-145.
- [44] S. N. Truong, S.-J. Ham, and K.-S. Min, "Neuromorphic crossbar circuit with nanoscale filamentary-switching binary memristors for speech recognition," *Nanoscale Research Letters*, vol. 9, p. 629, November 2014.
- [45] S. N. Truong, "Improving the Variation-Tolerance of Memristor Synapse by Selecting the Optimal Memristance," *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 8, pp. 12028-12031, Nov. 2019.
- [46] Y. Zhang, M. Cui, L. Shen, and Z. Zeng, "Memristive Quantized Neural Networks: A Novel Approach to Accelerate Deep Learning On-Chip," *IEEE Transactions on Cybernetics*, vol. 51, pp. 1875-1887, Apr. 2021.
- [47] S. N. Truong, S. Shin, S. Byeon, J. Song, and K. Min, "New Twin Crossbar Architecture of Binary Memristors for Low-Power Image Recognition With Discrete Cosine Transform," *IEEE Transactions on Nanotechnology*, vol. 14, pp. 1104-1111, November 2015.
- [48] S. N. Truong, "Single Crossbar Array of Memristors With Bipolar Inputs for Neuromorphic Image Recognition," *IEEE Access*, vol. 8, pp. 69327-69332, Apr. 2020.
- [49] A. H. Jaafar, L. Shao, P. Dai, T. Zhang, Y. Han, R. Beanland, *et al.*, "3D-structured mesoporous silica memristors for neuromorphic switching and reservoir computing," *Nanoscale*, vol. 14, pp. 17170-17181, November 2022.
- [50] S. He, X. Ye, X. Zhu, Q. Zhong, Y. Liu, G. Li, *et al.*, "High-performance self-rectifying memristor array based on Pt/HfO₂/Ta₂O_{5-x}/Ti structure for flexible electronics," *Nano Research*, vol. 18, p. 94907085, February 2025.

- [51] H. Zhou, S. Li, K. W. Ang, and Y. W. Zhang, "Recent Advances in In-Memory Computing: Exploring Memristor and Memtransistor Arrays with 2D Materials," *Nano-Micro Letters*, vol. 16, p. 121, February 2024.
- [52] X. Wu, B. Dang, T. Zhang, X. Wu, and Y. Yang, "Spatiotemporal audio feature extraction with dynamic memristor-based time-surface neurons," *Science Advances*, vol. 10, p. eadl2767, April 2024.
- [53] Z. Xia, X. Sun, Z. Wang, J. Meng, B. Jin, and T. Wang, "Low-Power Memristor for Neuromorphic Computing: From Materials to Applications," *Nano-Micro Letters*, vol. 17, p. 217, April 2025.
- [54] F. L. Aguirre, S. M. Pazos, F. Palumbo, J. Suñé, and E. Miranda, "Application of the Quasi-Static Memdiode Model in Cross-Point Arrays for Large Dataset Pattern Recognition," *IEEE Access*, vol. 8, pp. 202174-202193, November 2020.
- [55] C. Y. Han, S. L. Fang, Y. L. Cui, W. Liu, S. Q. Fan, X. D. Huang, *et al.*, "Configurable NbOx Memristors as Artificial Synapses or Neurons Achieved by Regulating the Forming Compliance Current for the Spiking Neural Network," *Advanced Electronic Materials*, vol. 9, p. 2300018, June 2023.
- [56] H. H. Huynh and M.-H. Vo, "Memristor-based character recognition using signal multiplier," *Journal of Technical Education Science*, vol. 32, pp. 60-66, 2015.
- [57] N.-T. Le and M.-H. Vo, "Synaptic memristor bridge circuit with pulse width based programmable weights in digit character recognition," *International Journal of Engineering and Science Invention (IJESI)* vol. 5, pp. 11-17, May 2016.
- [58] P.-Q. Pham, T. B. N. Duong, N. Q. N. Le, A. T. T. Pham, T. T. Nguyen, T. B. Phan, *et al.*, "Synaptic behavior in analog memristors based on green-synthesized ZnO nanoparticles," *Ceramics International*, vol. 50, pp. 28480-28489, August 2024.
- [59] P.-Q. Pham, T. A. Tran, T. V. V. Anh, T. D. Nguyen, J. Brugger, Y. Kawazoe, *et al.*, "Selector-Free 16×16 CrOX/TiO₂-Based Memristor Array for Synaptic Dynamics and LTP/ LTD Emulation: Experimental–Computational Correlation," *Advanced Functional Materials*, p. 16695, September 2025.
- [60] M. Davies, N. Srinivasa, T. H. Lin, G. Chinya, Y. Cao, S. H. Choday, *et al.*, "Loihi: A Neuromorphic Manycore Processor with On-Chip Learning," *IEEE Micro*, vol. 38, pp. 82-99, 2018.

- [61] B. Zimmer, R. Venkatesan, Y. S. Shao, J. Clemons, M. Fojtik, N. Jiang, *et al.*, "A 0.32–128 TOPS, Scalable Multi-Chip-Module-Based Deep Neural Network Inference Accelerator With Ground-Referenced Signaling in 16 nm," *IEEE Journal of Solid-State Circuits*, vol. 55, pp. 920-932, 2020.
- [62] Y. N. Joglekar and S. J. Wolf, "The elusive memristor: properties of basic electrical circuits," *European Journal of Physics*, vol. 30, pp. 661-675, May 2009.
- [63] Z. Biolek, D. Biolek, and V. Biolková, "SPICE Model of Memristor with Nonlinear Dopant Drift," *Radioengineering*, vol. 18, pp. 210-214, Jun. 2009.
- [64] K. Seo, I. Kim, S. Jung, M. Jo, S. Park, J. Park, *et al.*, "Analog Memory and Spike-Timing-Dependent Plasticity Characteristics of a Nanoscale Titanium Oxide Bilayer Resistive Switching Device," *Nanotechnology*, vol. 22, p. 254023, 06/24 2011.
- [65] D. Kuzum, R. Jeyasingh, B. Lee, and H. S. P. Wong, "Nanoelectronic Programmable Synapses Based on Phase Change Materials for Brain-Inspired Computing," *Nano letters*, vol. 12, pp. 2179-2186, Jun. 2011.
- [66] S. Park, A. Sheri, J. Kim, J. Noh, J. Jang, M. Jeon, *et al.*, "Neuromorphic speech systems using advanced ReRAM-based synapse," in *2013 IEEE International Electron Devices Meeting*, Washington, DC, USA, 2013, pp. 25.6.1-25.6.4.
- [67] D. Niu, Y. Xiao, and X. Yuan, "Low power memristor-based ReRAM design with Error Correcting Code," in *17th Asia and South Pacific Design Automation Conference*, Sydney, NSW, Australia, 2012, pp. 79-84.
- [68] P. Rabbani, R. deghani, and N. Shahpari, "A multilevel memristor–CMOS memory cell as a ReRAM," *Microelectronics Journal*, vol. 46, pp. 1283-1290, Oct. 2015.
- [69] R. Waser and M. Aono, "Nanoionics-based resistive switching memories," *Nature Materials*, vol. 6, pp. 833-840, Nov. 2007.
- [70] M.-J. Lee, S. Han, S. Jeon, B. Park, B. Kang, s. e. Ahn, *et al.*, "Electrical Manipulation of Nanofilaments in Transition-Metal Oxides for Resistance-Based Memory," *Nano letters*, vol. 9, pp. 1476-1481, Mar. 2009.
- [71] S. Yu and H. S. P. Wong, "Modeling the switching dynamics of programmable-metallization-cell (PMC) memory and its application as synapse device for a

- neuromorphic computation system," in *2010 International Electron Devices Meeting*, San Francisco, CA, USA, 2010, pp. 22.1.1-22.1.4.
- [72] M. Suri, D. Querlioz, O. Bichler, G. Palma, E. Vianello, D. Vuillaume, *et al.*, "Bio-Inspired Stochastic Computing Using Binary CBRAM Synapses," *IEEE Transactions on Electron Devices*, vol. 60, pp. 2402-2409, 2013.
- [73] L. L. Wei, J. X. Wang, Y. S. Chen, D. Shang, Z. Sun, B. Shen, *et al.*, "Pulse-induced alternation from bipolar resistive switching to unipolar resistive switching in the Ag/AgOx/Mg_{0.2}Zn_{0.8}O/Pt device," *Journal of Physics D: Applied Physics*, vol. 45, p. 425303, Oct. 2012.
- [74] C.-H. Huang, J.-S. Huang, C.-C. Lai, H.-W. Huang, S.-J. Lin, and Y.-L. Chueh, "Manipulated Transformation of Filamentary and Homogeneous Resistive Switching on ZnO Thin Film Memristor with Controllable Multistate," *ACS Applied Materials & Interfaces*, vol. 5, pp. 6017-6023, May 2013.
- [75] M.-H. Vo, S. N. Truong, S. Shin, and K.-S. Min, "CMOS-Memristor Hybrid 4-bit Multiplier Circuit for Energy-Efficient Computing," *Journal of IKEEE*, vol. 18, pp. 50-55, Jun. 2014.
- [76] B. J. Choi, D. S. Jeong, S. Kim, C. Rohde, S. Choi, J. Oh, *et al.*, "Resistive Switching Mechanism of TiO₂ Thin Films Grown by Atomic-Layer Deposition," *Journal of Applied Physics*, vol. 98, p. 033715, Aug. 2005.
- [77] D. Kim, S. Seo, s. e. Ahn, D. Suh, M.-J. Lee, B. H. Park, *et al.*, "Electrical Observations of Filamentary Conductions for the Resistive Memory Switching in NiO Films," *Applied Physics Letters*, vol. 88, p. 202102, May 2006.
- [78] 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.
- [79] 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.
- [80] 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.

- [81] 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, Jun. 2021.
- [82] J. Reuben, M. Biglari, and D. Fey, "Incorporating Variability of Resistive RAM in Circuit Simulations Using the Stanford–PKU Model," *IEEE Transactions on Nanotechnology*, vol. 19, pp. 508-518, 2020.
- [83] 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.
- [84] 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*, vol. 25, pp. 598-609, October 2025.
- [85] M. Le and S. N. Truong, "A Memristor Crossbar-Based XNOR-Net Architecture for Image Recognition," *Engineering, Technology & Applied Science Research*, vol. 15, pp. 25127-25132, August 2025.
- [86] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks," in *Computer Vision – ECCV 2016*, Amsterdam, The Netherlands, 2016, pp. 525-542.
- [87] M. Mohaimenuzzaman, C. Bergmeir, and B. Meyer, "Pruning vs XNOR-Net: A Comprehensive Study of Deep Learning for Audio Classification on Edge-Devices," *IEEE Access*, vol. 10, pp. 6696-6707, 2022.
- [88] M. Courbariaux and Y. Bengio, "BinaryNet: Training Deep Neural Networks with Weights and Activations Constrained to +1 or -1," *ArXiv*, vol. abs/1602.02830, 2016.
- [89] Y. Zhao, Y. Wang, R. Wang, Y. Rong, and X. Jiang, "A Highly Robust Binary Neural Network Inference Accelerator Based on Binary Memristors," *Electronics*, vol. 10, p. 2600, 2021.
- [90] F. Jebali, A. Majumdar, C. Turck, K.-E. Harabi, M.-C. Faye, E. Muhr, *et al.*, "Powering AI at the edge: A robust, memristor-based binarized neural network with near-memory computing and miniaturized solar cell," *Nature Communications*, vol. 15, p. 741, January 2024.

- [91] H. Träff, "Novel approach to high speed CMOS current comparators," *Electronics Letters*, vol. 28, pp. 310-311, 30 January 1992.

DANH MỤC CÁC 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*, vol. 25, 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)