



(12) BẢN MÔ TẢ SÁNG CHẾ THUỘC BẰNG ĐỘC QUYỀN SÁNG CHẾ

(19) Cộng hòa xã hội chủ nghĩa Việt Nam (VN)
CỤC SỞ HỮU TRÍ TUỆ

(11)



1-0048583

(51)^{2020.01} G06N 3/04; G06N 3/08

(13) B

(21) 1-2022-01312

(22) 08/09/2020

(86) PCT/US2020/049754 08/09/2020

(87) WO 2021/050440 A1 18/03/2021

(30) 16/565,308 09/09/2019 US

(45) 25/07/2025 448

(43) 25/07/2022 412A

(73) QUALCOMM INCORPORATED (US)

ATTN: International IP Administration 5775 Morehouse Drive San Diego, California
92121-1714 (US)

(72) WANG, Zhongze (US); TEAGUE, Edward (US); WELLING, Max (NL).

(74) Công ty TNHH Quốc tế D & N (D&N INTERNATIONAL CO.,LTD.)

(54) PHƯƠNG PHÁP VÀ THIẾT BỊ THỰC HIỆN CÁC PHÉP TOÁN XNOR TƯƠNG ĐƯƠNG BẰNG CÁCH ĐIỀU CHỈNH NGUỒN CỘT CỦA MẢNG BỘ NHỚ CÓ TÍNH TOÁN VÀ PHƯƠNG TIỆN BẮT BIẾN ĐỌC ĐƯỢC BẰNG MÁY TÍNH

(21) 1-2022-01312

(57) Sáng chế đề cập đến phương pháp và thiết bị thực hiện các phép toán XNOR tương đương bằng cách điều chỉnh các ngưỡng cột của mảng bộ nhớ có tính toán của mạng nơ-ron nhân tạo và phương tiện bất biến đọc được bằng máy tính. Phương pháp bao gồm bước điều chỉnh ngưỡng kích hoạt được tạo ra cho mỗi cột của mảng bộ nhớ có tính toán dựa trên hàm của giá trị trọng số và giá trị kích hoạt. Phương pháp còn bao gồm bước tính toán tham chiếu dòng điện phân cực chuyển đổi dựa trên giá trị đầu vào từ vectơ đầu vào cho mảng bộ nhớ có tính toán, mảng bộ nhớ có tính toán được lập trình với tập hợp các trọng số. Ngưỡng kích hoạt được điều chỉnh và tham chiếu dòng điện phân cực chuyển đổi được dùng làm ngưỡng để xác định các giá trị đầu ra của mảng bộ nhớ có tính toán.

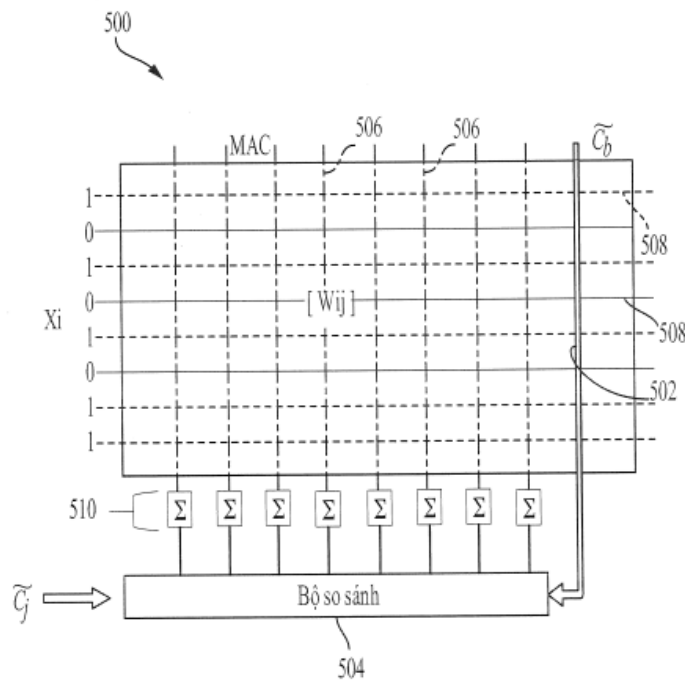


Fig.5

Lĩnh vực kỹ thuật được đề cập

Nói chung các khía cạnh của sáng chế đề cập đến việc thực hiện các phép toán OR loại trừ phủ định (exclusive-negative OR - XNOR) tương đương bằng cách điều chỉnh các ngưỡng cột của mảng bộ nhớ có tính toán của mạng nơron nhân tạo.

Tình trạng kỹ thuật của sáng chế

Các mạng nơron có độ rộng bit rất thấp, như các mạng nơron nhị phân (binary neural network - BNN), là các tiếp cận mới đầy mạnh mẽ trong các mạng nơron sâu (deep neural networking - DNN). Các mạng nơron nhị phân có thể làm giảm đáng kể lưu lượng dữ liệu và tiết kiệm công suất. Ví dụ, bộ nhớ lưu trữ cho các mạng nơron nhị phân được giảm đáng kể vì cả trọng số và kích hoạt nơron đều được nhị phân hóa thành -1 hoặc +1, so với độ chính xác dấu phẩy động/điểm cố định.

Tuy nhiên, quy trình xử lý chất bán dẫn kim loại ôxit bù (complementary metal-oxide-semiconductor - CMOS) kỹ thuật số, sử dụng cơ sở [0,1]. Để thực hiện phương án triển khai nhị phân kết hợp với các mạng nơron nhị phân này, cơ sở [-1, +1] của các mạng nhị phân phải được biến đổi thành cơ sở CMOS [0,1]. Kỹ thuật biến đổi sử dụng phép toán OR loại trừ phủ định (exclusive-negative OR - XNOR) mạnh về mặt tính toán.

Hệ thống bộ nhớ có tính toán có thể thực hiện các mạng nơron có độ rộng bit rất thấp, như các mạng nơron nhị phân (binary neural network - BNN). Hệ thống bộ nhớ có tính toán có bộ nhớ với các khả năng xử lý. Ví dụ, mỗi giao điểm của dòng bit và dòng từ biểu diễn giá trị trọng số của bộ lọc, giá trị này được nhân với giá trị kích hoạt đầu vào trên dòng từ để tạo ra tích. Sau đó, các tích riêng dọc theo mỗi dòng bit được cộng lại để tạo ra các giá trị đầu ra tương ứng của tenxơ đầu ra. Phương án thực hiện này có thể được coi là phép toán nhân tích lũy (multiply accumulate - MAC). Các phép toán MAC này có thể biến đổi cơ sở [-1, +1] của các mạng nhị phân thành cơ sở CMOS [0,1].

Thông thường, việc biến đổi với hệ thống bộ nhớ có tính toán đạt được bằng cách hoàn thành phép toán XNOR tại mỗi ô bit. Kết quả dọc theo mỗi dòng bit sau đó được cộng để tạo ra các giá trị đầu ra tương ứng. Thật không may, việc bao gồm hàm XNOR trong mỗi ô bit làm tổn diện tích lớn và tăng mức tiêu thụ công suất.

Theo phương án thực hiện thông thường, mỗi ô bit bao gồm hàm nhớ cơ sở là đọc và ghi cộng với hàm logic bổ sung của XNOR giữa đầu vào và trạng thái ô. Do đó bao gồm khả năng XNOR, số lượng bóng bán dẫn cho mỗi ô trong bộ nhớ (ví dụ, bộ nhớ truy cập ngẫu nhiên tĩnh (static random-access memory - SRAM)) tăng từ sáu hoặc tám lên mười hai, việc này làm tăng đáng kể kích thước ô và mức tiêu thụ công suất. Mong muốn loại bỏ phép toán XNOR trong khi vẫn có thể biến đổi từ cơ sở $[-1, +1]$ của các mạng nơron nhị phân sang cơ sở CMOS $[0,1]$.

Bản chất kỹ thuật của sáng chế

Theo một khía cạnh của sáng chế, thiết bị bao gồm mảng bộ nhớ có tính toán bao gồm các cột và các hàng. Mảng bộ nhớ có tính toán được tạo cấu hình để điều chỉnh ngưỡng kích hoạt được tạo ra cho mỗi cột của mảng bộ nhớ có tính toán dựa trên hàm của giá trị trọng số và giá trị kích hoạt. Mảng bộ nhớ có tính toán cũng được tạo cấu hình để tính toán tham chiếu dòng điện phân cực chuyển đổi dựa trên giá trị đầu vào từ vector đầu vào cho mảng bộ nhớ có tính toán. Mảng bộ nhớ có tính toán được lập trình với tập hợp các giá trị trọng số. Ngưỡng kích hoạt được điều chỉnh và tham chiếu dòng điện phân cực chuyển đổi được dùng làm ngưỡng để xác định các giá trị đầu ra của mảng bộ nhớ có tính toán.

Khía cạnh khác mô tả phương pháp để thực hiện các phép toán XNOR tương đương bằng cách điều chỉnh các ngưỡng cột của mảng bộ nhớ có tính toán có các hàng và các cột. Phương pháp này bao gồm bước điều chỉnh ngưỡng kích hoạt được tạo ra cho mỗi cột của mảng bộ nhớ có tính toán dựa trên hàm của giá trị trọng số và giá trị kích hoạt. Phương pháp này còn bao gồm bước tính toán tham chiếu dòng điện phân cực chuyển đổi dựa trên giá trị đầu vào từ vector đầu vào cho mảng bộ nhớ có tính toán. Mảng bộ nhớ có tính toán được lập trình với tập hợp các giá trị trọng số. Ngưỡng kích hoạt được điều chỉnh và tham chiếu dòng điện phân cực chuyển đổi được dùng làm ngưỡng để xác định các giá trị đầu ra của mảng bộ nhớ có tính toán.

Theo một khía cạnh khác, phương tiện bất biến đọc được bằng máy tính ghi lại mã chương trình bất biến. Mã chương trình bất biến này, khi được (các) bộ xử lý thực thi, khiến cho (các) bộ xử lý điều chỉnh ngưỡng kích hoạt được tạo ra cho mỗi cột của mảng bộ nhớ có tính toán có các hàng và cột dựa trên hàm của giá trị trọng số và giá trị kích hoạt. Mã chương trình cũng khiến cho (các) bộ xử lý tính toán tham chiếu dòng điện phân cực chuyển đổi dựa trên giá trị đầu vào từ vector đầu cho đến mảng bộ nhớ có tính toán. Mảng bộ nhớ có tính toán được lập trình với tập hợp các giá trị trọng số. Ngưỡng kích hoạt được điều chỉnh và tham chiếu dòng điện phân cực chuyển đổi được dùng làm ngưỡng để xác định các giá trị đầu ra của mảng bộ nhớ có tính toán.

Khía cạnh khác mô tả thiết bị để thực hiện các phép toán XNOR tương đương bằng cách điều chỉnh các ngưỡng cột của mảng bộ nhớ có tính toán có các hàng và các cột. Thiết bị bao gồm phương tiện để điều chỉnh ngưỡng kích hoạt được tạo ra cho mỗi cột của mảng bộ nhớ có tính toán dựa trên hàm của giá trị trọng số và giá trị kích hoạt. Thiết bị này còn bao gồm phương tiện để tính toán tham chiếu dòng điện phân cực chuyển đổi dựa trên giá trị đầu vào từ vector đầu vào cho mảng bộ nhớ có tính toán. Mảng bộ nhớ có tính toán được lập trình với tập hợp các giá trị trọng số. Ngưỡng kích hoạt được điều chỉnh và tham chiếu dòng điện phân cực chuyển đổi được dùng làm ngưỡng để xác định các giá trị đầu ra của mảng bộ nhớ có tính toán.

Tóm lại, việc này đã nêu ra các đặc điểm và ưu điểm kỹ thuật của sáng chế để có thể hiểu rõ hơn phần mô tả chi tiết sau đây. Các tính năng và ưu điểm bổ sung của sáng chế sẽ được mô tả dưới đây. Người có hiểu biết trung bình trong lĩnh vực kỹ thuật này hiểu rằng sáng chế có thể được dùng làm cơ sở để sửa đổi hoặc thiết kế các cấu trúc khác để thực hiện các mục đích tương tự của sáng chế. Người có hiểu biết trung bình trong lĩnh vực kỹ thuật này hiểu rằng các cấu trúc tương đương như vậy không khác với những kiến thức của sáng chế như được nêu trong phần yêu cầu bảo hộ. Các đặc tính mới, được cho là đặc trưng của sáng chế, cả về tổ chức và phương pháp hoạt động của nó, cùng với các đối tượng và ưu điểm khác, sẽ được hiểu rõ hơn từ phần mô tả sau đây khi được xem xét liên quan đến các hình vẽ kèm theo. Tuy nhiên, cần phải hiểu một cách rõ ràng rằng mỗi trong số các hình vẽ được cung cấp chỉ nhằm mục đích minh họa và mô tả và không nhằm mục đích định nghĩa các giới hạn của sáng chế.

Mô tả vắn tắt các hình vẽ

Các đặc điểm, bản chất và ưu điểm của sáng chế sẽ trở nên rõ ràng hơn từ mô tả chi tiết được nêu dưới đây khi được thực hiện cùng với các hình vẽ trong đó các ký tự tham chiếu tương tự xác định toàn bộ phần mô tả tương ứng.

Fig.1 minh họa ví dụ về phương án thực hiện thiết kế các mạng nơron sử dụng hệ thống trên chip (system-on-a-chip - SOC), bao gồm bộ xử lý đa dụng theo một số khía cạnh của sáng chế.

Fig.2A, Fig.2B và Fig.2C là các sơ đồ minh họa mạng nơron theo các khía cạnh của sáng chế.

Fig.2D là sơ đồ minh họa mạng tích chập sâu (deep convolutional network - DCN) làm ví dụ theo các khía cạnh của sáng chế.

Fig.3 là sơ đồ khối minh họa mạng tích chập sâu (DCN) làm ví dụ theo các khía cạnh của sáng chế.

Fig.4 minh họa kiến trúc thể hiện mảng bộ nhớ có tính toán (compute-in-memory - CIM) của mạng nơron nhân tạo, theo các khía cạnh của sáng chế.

Fig.5 minh họa kiến trúc để thực hiện các phép toán XNOR tương đương bằng cách điều chỉnh các ngưỡng cột của mảng bộ nhớ có tính toán của mạng nơron nhân tạo, theo các khía cạnh của sáng chế.

Fig.6 minh họa phương pháp thực hiện các phép toán XNOR tương đương bằng cách điều chỉnh các ngưỡng cột của mảng bộ nhớ có tính toán của mạng nơron nhân tạo, theo các khía cạnh của sáng chế.

Mô tả chi tiết sáng chế

Phần mô tả chi tiết trình bày dưới đây dựa vào các hình vẽ kèm theo được dự định làm phần mô tả về các cấu hình khác nhau và không dự định để chỉ biểu diễn các cấu hình mà trong đó có thể áp dụng các khái niệm được mô tả ở đây. Phần mô tả chi tiết này bao gồm các chi tiết cụ thể để cung cấp sự hiểu biết thấu đáo về các khái niệm khác nhau. Tuy nhiên, hiển nhiên đối với người có hiểu biết trung bình trong lĩnh vực kỹ thuật này là các khái niệm này có thể được thực hiện mà không cần đến những chi tiết cụ thể này. Trong một số trường hợp, các cấu trúc và thành phần đã biết được thể hiện dưới dạng sơ đồ khối để tránh làm tối nghĩa các khái niệm như vậy.

Dựa trên các nguyên lý được đề xuất ở đây, người có hiểu biết trung bình trong lĩnh vực kỹ thuật này có thể thấy rằng phạm vi của sáng chế dự định bao gồm mọi khía cạnh của sáng chế được đề xuất ở đây, cho dù được thực hiện độc lập hay kết hợp với bất kỳ khía cạnh nào khác của sáng chế. Chẳng hạn, thiết bị có thể được thực thi hoặc phương pháp có thể được thực hiện bằng cách sử dụng một số khía cạnh bất kỳ được đề xuất ở đây. Ngoài ra, phạm vi của sáng chế được dự định bao gồm thiết bị hoặc phương pháp được thực hiện bằng cách sử dụng cấu trúc, hàm khác, hoặc cấu trúc và hàm bổ sung hoặc khác với một số khía cạnh của sáng chế được nêu ở đây. Cần phải hiểu rằng khía cạnh bất kỳ của sáng chế bộc lộ ở đây có thể được thực hiện bằng một hoặc nhiều phần tử nêu trong phần yêu cầu bảo hộ.

Thuật ngữ “làm ví dụ” được dùng ở đây có nghĩa là “có vai trò làm ví dụ, mẫu hoặc minh họa”. Khía cạnh bất kỳ được mô tả ở đây là “mang tính ví dụ” không nhất thiết được hiểu là được ưu tiên hoặc có lợi so với các khía cạnh khác.

Mặc dù các khía cạnh cụ thể được mô tả ở đây, nhưng nhiều sửa đổi và hoán vị của các khía cạnh này cũng nằm trong phạm vi của sáng chế. Mặc dù một số lợi ích và ưu điểm của các khía cạnh ưu tiên được đề cập, nhưng phạm vi của sáng chế không được dự định để bị giới hạn ở các lợi ích, cách sử dụng hoặc mục đích cụ thể. Đúng hơn là, các khía cạnh của sáng chế được dự định để có thể áp dụng rộng rãi cho các công nghệ, cấu hình hệ thống, các mạng, và giao thức khác nhau, một số trong đó sẽ được minh họa làm ví dụ trên các hình vẽ và trong phần mô tả chi tiết các khía cạnh ưu tiên dưới đây. Phần mô tả chi tiết và các hình vẽ chỉ mang tính chất minh họa sáng chế chứ không giới hạn, phạm vi sáng chế được xác định bởi các yêu cầu bảo hộ và các dạng tương đương của chúng.

Bộ nhớ có tính toán (CIM) là phương pháp thực hiện phép tính nhân và tích lũy (multiply and accumulate - MAC) trong mảng bộ nhớ. Bộ nhớ có tính toán có thể cải thiện tính song song trong mảng bộ nhớ bằng cách kích hoạt nhiều hàng và sử dụng dòng điện cột tương tự để quản lý các phép toán nhân và tổng. Ví dụ, các ô bit SRAM có thể được tùy chỉnh để kích hoạt XNOR và các phép toán đếm bit cho các mạng nơ-ron nhị phân.

Thông thường, các phương án thực hiện các mạng nơ-ron nhị phân bộ nhớ có tính toán được thực hiện bằng cách hoàn thành các phép toán XNOR tại mỗi ô bit và tính tổng kết quả cho mỗi ô bit. Thêm hàm XNOR trong mỗi ô bit sẽ làm tăng diện tích bố trí và

tăng mức tiêu thụ công suất. Ví dụ, số lượng bóng bán dẫn trong mỗi ô trong bộ nhớ (ví dụ, SRAM) tăng từ sáu hoặc tám lên mười hai.

Các khía cạnh của sáng chế hướng đến việc thực hiện các phép toán XNOR tương đương bằng cách điều chỉnh các ngưỡng cột của mảng bộ nhớ có tính toán của mạng nơron nhân tạo (ví dụ, mạng nơron nhị phân). Theo một khía cạnh, ngưỡng kích hoạt cho mỗi cột của mảng bộ nhớ được điều chỉnh dựa trên hàm của giá trị trọng số và giá trị kích hoạt. Tham chiếu dòng điện phân cực chuyển đổi được tính toán dựa trên giá trị đầu vào từ vector đầu vào.

Theo một khía cạnh, tổng mật độ cư trú dòng bit được so sánh với tổng của tham chiếu dòng điện phân cực chuyển đổi và ngưỡng kích hoạt được điều chỉnh để xác định đầu ra của dòng bit. Tổng mật độ cư trú dòng bit là tổng của mỗi đầu ra của các ô bit tương ứng với dòng bit của mảng bộ nhớ. Ví dụ, tổng của đầu ra (hoặc tổng mật độ cư trú) của mỗi ô bit kết hợp với dòng bit thứ nhất được cung cấp cho bộ so sánh dưới dạng đầu vào thứ nhất. Sau đó, tổng mật độ cư trú được so sánh với tổng của tham chiếu dòng điện phân cực chuyển đổi và ngưỡng kích hoạt được điều chỉnh để xác định đầu ra của dòng bit. Theo một số khía cạnh, ngưỡng kích hoạt nhỏ hơn một nửa số hàng của mảng bộ nhớ. Số hàng tương ứng với kích thước của vector đầu vào. Theo một số khía cạnh, tham chiếu dòng điện phân cực chuyển đổi nhỏ hơn một nửa số hàng của mảng bộ nhớ.

Mạng nơron nhân tạo theo sáng chế có thể là mạng nơron nhị phân, mạng nơron nhiều bit, hoặc mạng nơron có độ rộng bit rất thấp. Các khía cạnh của sáng chế có thể áp dụng được cho các thiết bị (ví dụ, thiết bị biên) xác định xử lý bộ nhớ rất thấp và công suất hoặc các mạng lớn có thể được hưởng lợi từ việc tiết kiệm bộ nhớ do định dạng nhị phân. Các khía cạnh của sáng chế làm giảm kích thước và cải thiện mức tiêu thụ công suất của bộ nhớ bằng cách loại bỏ các phép toán XNOR trong các hệ thống bộ nhớ có tính toán mà sử dụng các mạng nơron nhị phân. Ví dụ, kỹ thuật biến đổi cơ sở được thực hiện để tránh sử dụng hàm XNOR và (các) bóng bán dẫn tương ứng của nó trong mỗi ô bit, do đó làm giảm kích thước của bộ nhớ.

Fig.1 minh họa ví dụ về phương án thực hiện hệ thống trên chip (SOC) 100, hệ thống này có thể bao gồm bộ xử lý trung tâm (central processing unit - CPU) 102 hoặc CPU nhiều lõi được tạo cấu hình để biến đổi các phép toán nhân và tích lũy cho mảng bộ nhớ có tính toán (CIM) của mạng nơron nhân tạo theo một số khía cạnh của sáng chế. Các

biến (ví dụ, tín hiệu noron và các trọng số tiếp hợp), các thông số hệ thống kết hợp với thiết bị tính toán (ví dụ, mạng noron có trọng số), độ trễ, thông tin bin tần số, và thông tin tác vụ có thể được lưu trữ trong khối bộ nhớ kết hợp với bộ xử lý noron (neural processing unit - NPU) 108, trong khối bộ nhớ kết hợp với CPU 102, trong khối bộ nhớ kết hợp với bộ xử lý đồ họa (graphics processing unit - GPU) 104, trong khối bộ nhớ kết hợp với bộ xử lý tín hiệu số (digital signal processor - DSP) 106, trong khối bộ nhớ 118, hoặc có thể được phân tán trên nhiều khối. Các lệnh được thực thi ở CPU 102 có thể được tải từ bộ nhớ chương trình kết hợp với CPU 102 hoặc có thể được tải từ khối bộ nhớ 118.

SOC 100 cũng có thể bao gồm các khối xử lý bổ sung phù hợp với các hàm cụ thể, như GPU 104, DSP 106, khối kết nối 110, khối này có thể bao gồm kết nối thế hệ thứ năm (fifth generation - 5G), kết nối tiến hóa dài hạn thế hệ thứ tư (fourth generation long term evolution - 4G LTE), kết nối Wi-Fi, kết nối USB, kết nối Bluetooth, v.v. và bộ xử lý đa phương tiện 112 có thể, ví dụ, phát hiện và nhận dạng các cử chỉ. Theo một phương án thực hiện, NPU được thực hiện trong CPU, DSP và/hoặc GPU. SOC 100 cũng có thể bao gồm bộ xử lý cảm biến 114, bộ xử lý tín hiệu hình ảnh (image signal processors - ISP) 116 và/hoặc module điều hướng 120, có thể bao gồm hệ thống định vị toàn cầu.

SOC 100 có thể được dựa trên tập lệnh ARM. Theo một khía cạnh của sáng chế, các lệnh được nạp vào bộ xử lý đa dụng 102 có thể bao gồm mã để điều chỉnh ngưỡng kích hoạt cho mỗi cột của mảng dựa trên hàm của giá trị trọng số (ví dụ, ma trận trọng số) và giá trị kích hoạt. Bộ xử lý đa dụng 102 có thể bao gồm thêm mã để tính toán tham chiếu dòng điện phân cực chuyên đổi dựa trên giá trị đầu vào từ vector đầu vào.

Kiến trúc học sâu có thể thực hiện nhiệm vụ nhận dạng đối tượng bằng cách học cách biểu diễn các đầu vào ở các mức trừu tượng hóa cao hơn liên tục trong mỗi lớp, do đó xây dựng dạng biểu diễn tính năng hữu ích của dữ liệu đầu vào. Theo cách này, kỹ thuật học sâu xử lý điểm nghẽn lớn của học máy truyền thống. Trước khi học sâu ra đời, phương pháp học máy đối với vấn đề nhận dạng đối tượng có thể chủ yếu dựa vào các tính năng do con người thiết kế, có thể kết hợp với bộ phân loại nông. Ví dụ, bộ phân loại nông có thể là bộ phân loại tuyến tính hai lớp, trong đó tổng trọng số của các thành phần vector đặc trưng có thể được so sánh với một ngưỡng để dự báo đầu vào thuộc về lớp nào. Các tính năng do con người thiết kế có thể là các mẫu hoặc hạt nhân được các kỹ sư có chuyên môn về miền điều chỉnh cho phù hợp với miền vấn đề cụ thể. Ngược lại, kiến trúc học sâu có

thể học để đại diện cho các tính năng tương tự như các tính năng của kỹ sư con người có thể thiết kế, nhưng qua huấn luyện. Hơn nữa, mạng sâu có thể học để biểu diễn và nhận ra các loại tính năng mới mà con người có thể không xét đến.

Kiến trúc học sâu có thể học hệ phân cấp các tính năng. Ví dụ, nếu được biểu diễn bằng dữ liệu trực quan, lớp thứ nhất có thể học để nhận ra các tính năng tương đối đơn giản, như các biên, trong luồng đầu vào. Theo một ví dụ khác, nếu được biểu diễn với dữ liệu thính giác, lớp thứ nhất có thể học để nhận biết công suất phổ trong các tần số cụ thể. Lớp thứ hai, lấy đầu ra của lớp thứ nhất làm đầu vào, có thể học để nhận ra sự kết hợp các tính năng, như hình dạng đơn giản cho dữ liệu trực quan hoặc kết hợp âm thanh cho dữ liệu thính giác. Ví dụ, các lớp cao hơn có thể học để biểu diễn các hình dạng phức tạp trong dữ liệu trực quan hoặc các từ trong dữ liệu thính giác. Các lớp cao hơn vẫn có thể học để nhận biết các đối tượng trực quan hoặc các cụm từ nói ra chung.

Kiến trúc học sâu có thể thực hiện một cách đặc biệt tốt khi được áp dụng cho các vấn đề có cấu trúc phân cấp tự nhiên. Ví dụ, việc phân loại các phương tiện cơ giới có thể được hưởng lợi từ việc học thứ nhất để nhận biết bánh xe, kính chắn gió, và các đặc điểm khác. Các đặc điểm này có thể được kết hợp ở các lớp cao hơn theo các cách khác nhau để nhận dạng ô tô, xe tải và máy bay.

Các mạng nơron có thể được thiết kế với nhiều mẫu kết nối khác nhau. Trong các mạng truyền thẳng, thông tin được truyền từ các lớp thấp hơn đến các lớp cao hơn, với mỗi nơron trong một lớp nhất định sẽ truyền thông với các nơron trong các lớp cao hơn. Dạng biểu diễn phân cấp có thể được xây dựng trong các lớp kế tiếp của mạng truyền thẳng, như được mô tả ở trên. Các mạng nơron cũng có thể có các kết nối quy lại hoặc hồi tiếp (còn được gọi là từ trên xuống). Trong kết nối quy lại, đầu ra từ một nơron trong một lớp nhất định có thể được truyền thông với nơron khác trong cùng một lớp. Kiến trúc hồi quy có thể hữu ích trong việc nhận ra các mẫu trải trên nhiều hơn một trong các khối dữ liệu đầu vào được phân phối cho các mạng nơron theo trình tự. Kết nối từ một nơron trong một lớp nhất định đến một nơron ở lớp thấp hơn được gọi là kết nối hồi tiếp (hoặc từ trên xuống). Mạng có nhiều kết nối hồi tiếp có thể hữu ích khi việc nhận dạng khái niệm mức cao có thể hỗ trợ việc phân biệt các tính năng mức thấp cụ thể của đầu vào.

Các kết nối giữa các lớp của các mạng nơron có thể được kết nối đầy đủ hoặc kết nối cục bộ. Fig.2A minh họa ví dụ về các mạng nơron được kết nối đầy đủ 202. Trong

mạng nơron được kết nối đầy đủ 202, nơron ở lớp thứ nhất có thể truyền đầu ra của nó cho mọi nơron trong lớp thứ hai, do đó mỗi nơron trong lớp thứ hai sẽ nhận đầu vào từ mọi nơron trong lớp thứ nhất. Fig.2B minh họa ví dụ về mạng nơron được kết nối cục bộ 204. Trong mạng nơron được kết nối cục bộ 204, nơron ở lớp thứ nhất có thể được kết nối với số lượng nơron giới hạn trong lớp thứ hai. Nói chung hơn, lớp kết nối cục bộ của mạng nơron được kết nối cục bộ 204 có thể được tạo cấu hình sao cho mỗi nơron trong lớp sẽ có cùng một mẫu kết nối hoặc mẫu kết nối tương tự, nhưng với cường độ kết nối có thể có các giá trị khác nhau (ví dụ, 210, 212, 214 và 216). Mẫu kết nối được kết nối cục bộ có thể làm phát sinh các trường nhận khác biệt về mặt không gian ở lớp cao hơn, do các nơron của lớp cao hơn trong một vùng nhất định có thể nhận đầu vào được điều chỉnh qua huấn luyện theo các thuộc tính của phân hạn chế trong tổng đầu vào cho mạng.

Một ví dụ về mạng nơron được kết nối cục bộ là mạng nơron tích chập. Fig.2C minh họa ví dụ về mạng nơron tích chập 206. Mạng nơron tích chập 206 có thể được tạo cấu hình sao cho các cường độ kết nối kết hợp với các đầu vào cho mỗi nơron trong lớp thứ hai được dùng chung (ví dụ, 208). Mạng nơron tích chập có thể phù hợp với các vấn đề trong đó vị trí không gian của các đầu vào có ý nghĩa.

Một loại mạng nơron tích chập là mạng tích chập sâu (deep convolutional network - DCN). Fig.2D minh họa ví dụ chi tiết về mạng DCN 200 được thiết kế để nhận dạng các đặc điểm trực quan từ hình ảnh 226 nhập vào từ thiết bị chụp ảnh 230, như camera gắn trên xe hơi. Mạng DCN 200 của ví dụ hiện tại có thể được huấn luyện để nhận dạng các biển báo giao thông và số có trên biển báo giao thông. Tất nhiên, mạng DCN 200 có thể được huấn luyện cho các nhiệm vụ khác, như xác định vạch kẻ làn hoặc xác định đèn giao thông.

Mạng DCN 200 có thể được huấn luyện với kỹ thuật học có giám sát. Trong quy trình huấn luyện, mạng DCN 200 có thể được biểu thị bằng hình ảnh, như hình ảnh 226 của biển báo giới hạn tốc độ, và sau đó trực tiếp có thể được tính toán để tạo ra đầu ra 222. Mạng DCN 200 có thể bao gồm đoạn trích đặc trưng và đoạn phân loại. Khi nhận hình ảnh 226, lớp tích chập 232 có thể áp dụng các nhân tích chập (không được thể hiện trên hình vẽ) cho hình ảnh 226 để tạo ra tập hợp ảnh xạ đặc trưng thứ nhất 218. Ví dụ, nhân tích chập cho lớp tích chập 232 có thể là nhân 5x5 tạo ra các ảnh xạ đặc trưng 28x28. Trong ví dụ hiện tại, do bốn ảnh xạ đặc trưng khác nhau được tạo ra trong tập hợp ảnh xạ đặc trưng

thứ nhất 218, bốn nhân tích chập khác nhau được áp dụng cho hình ảnh 226 tại lớp tích chập 232. Nhân tích chập cũng có thể được gọi là bộ lọc hoặc bộ lọc tích chập.

Tập hợp ánh xạ đặc trưng thứ nhất 218 có thể được lấy mẫu phụ bởi lớp gộp tối đa (không được thể hiện trên hình vẽ) để tạo ra tập hợp ánh xạ đặc trưng thứ hai 220. Lớp gộp tối đa giảm kích thước của tập ánh xạ đặc trưng thứ nhất 218. Tức là, kích thước của tập ánh xạ đặc trưng thứ hai 220, như 14x14, nhỏ hơn kích thước của tập ánh xạ đặc trưng thứ nhất 218, như 28x28. Kích thước rút gọn cung cấp thông tin tương tự cho lớp tiếp theo trong khi giảm mức tiêu thụ bộ nhớ. Tập hợp ánh xạ đặc trưng thứ hai 220 có thể được biến đổi thêm qua một hoặc nhiều lớp tích chập tiếp theo (không thể được thể hiện trên hình vẽ) để tạo ra một hoặc nhiều tập hợp ánh xạ đặc trưng tiếp theo (không được thể hiện trên hình vẽ).

Trong ví dụ trên Fig.2D, tập hợp ánh xạ đặc trưng thứ hai 220 được biến đổi để tạo ra vector đặc trưng thứ nhất 224. Hơn nữa, vector đặc trưng thứ nhất 224 được biến đổi thêm để tạo ra vector đặc trưng thứ hai 228. Mỗi đặc điểm của vector đặc trưng thứ hai 228 có thể bao gồm một số tương ứng với một đặc điểm có thể có của hình ảnh 226, như “biển báo”, “60” và “100”. Hàm softmax (không được thể hiện trên hình vẽ) có thể biến đổi các số trong vector đặc trưng thứ hai 228 thành xác suất. Như vậy, đầu ra 222 của mạng DCN 200 là xác suất của hình ảnh 226 bao gồm một hoặc nhiều đặc điểm.

Trong ví dụ hiện tại, các xác suất trong đầu ra 222 đối với “biển báo” và “60” cao hơn xác suất của các xác suất của các hình ảnh khác của đầu ra 222, như “30,” “40,” “50,” “70,” “80,” “90” và “100”. Trước khi huấn luyện, đầu ra 222 do mạng DCN 200 tạo ra có thể không chính xác. Do đó, sai số có thể được tính giữa đầu ra 222 và đầu ra đích. Đầu ra đích là chân lý chính của hình ảnh 226 (ví dụ, “biển báo” và “60”). Sau đó, các trọng số của mạng DCN 200 có thể được điều chỉnh để đầu ra 222 của mạng DCN 200 được đồng chỉnh chặt chẽ hơn với đầu ra đích.

Để điều chỉnh các trọng số, thuật toán học có thể tính toán vector gradient cho các trọng số. Gradient có thể chỉ báo lượng mà sai số sẽ tăng hoặc giảm nếu trọng số được điều chỉnh. Ở lớp trên cùng, gradient có thể tương ứng trực tiếp với giá trị của trọng số kết nối neuron được kích hoạt ở lớp gần cuối và neuron ở lớp đầu ra. Trong các lớp thấp hơn, gradient có thể phụ thuộc vào giá trị của các trọng số và vào các gradient sai số được tính toán của các lớp cao hơn. Sau đó, các trọng số có thể được điều chỉnh để giảm sai số. Cách điều

chỉnh các trọng số này có thể được gọi là “lan truyền ngược” vì nó liên quan đến “truyền ngược” qua mạng nơron.

Trong thực tế, gradient sai số của các trọng số có thể được tính toán trên số ít ví dụ, do đó gradient được tính toán gần đúng với gradient sai số thực. Phương pháp gần đúng này có thể được gọi là phương pháp giảm dần gradient ngẫu nhiên. Phương pháp giảm dần gradient ngẫu nhiên có thể được lặp đến khi tỷ lệ sai số có thể đạt được của toàn bộ hệ thống ngừng giảm hoặc đến khi tỷ lệ sai số đạt đến mức cuối cùng. Sau khi học, mạng DCN có thể được biểu diễn với các hình ảnh mới (ví dụ, biên giới hạn tốc độ của hình ảnh 226) và việc tính toán trực tiếp qua các mạng có thể tạo ra đầu ra 222 có thể được coi là nhiễu hoặc dự báo của mạng DCN.

Các mạng tin cậy sâu (Deep belief network - DBN) là các mô hình xác suất bao gồm nhiều lớp nút ẩn. Các mạng DBN có thể được dùng để trích dạng biểu diễn phân cấp của các tập dữ liệu huấn luyện. Mạng DBN có thể thu được bằng cách xếp chồng các lớp Máy Boltzmann giới hạn (Restricted Boltzmann Machine - RBM). RBM là loại các mạng nơron nhân tạo có thể học phân phối xác suất trên tập hợp các đầu vào. Vì các RBM có thể học phân phối xác suất trong trường hợp không có thông tin về lớp để mỗi đầu vào cần được phân loại, nên RBM thường được dùng trong quy trình học không giám sát. Bằng cách sử dụng mô hình không được giám sát và giám sát lại, các RBM dưới cùng của mạng DBN có thể được huấn luyện theo cách không giám sát và có thể đóng vai trò là bộ trích đặc trưng, và RBM trên cùng có thể được huấn luyện theo cách được giám sát (dựa vào sự phân phối chung các đầu vào từ lớp trước và các lớp đích) và có thể đóng vai trò như bộ phân loại.

Các mạng tích chập sâu (Deep convolutional network - DCN) là các mạng gồm các mạng tích chập, được tạo cấu hình với các lớp gộp và chuẩn hóa bổ sung. Các mạng DCN đã đạt được hiệu suất tiên tiến trong nhiều nhiệm vụ. Các mạng DCN có thể được huấn luyện bằng cách sử dụng phương pháp học được giám sát, trong đó cả đích đầu vào và đầu ra đều được biết đến với nhiều ví dụ và được dùng để sửa đổi các trọng số của mạng bằng cách sử dụng phương pháp giảm dần gradient.

Mạng DCN có thể là các mạng truyền thẳng. Ngoài ra, như đã mô tả ở trên, các kết nối từ nơron trong lớp thứ nhất của DCN đến nhóm các nơron ở lớp cao hơn tiếp theo được dùng chung qua các nơron trong lớp thứ nhất. Các kết nối truyền thẳng và dùng chung của

các mạng DCN có thể được sử dụng để xử lý nhanh. Ví dụ, gánh nặng tính toán của mạng DCN có thể ít hơn nhiều so với các mạng nơron có kích thước tương tự bao gồm các kết nối quy hoặc kết nối tiếp.

Quy trình xử lý từng lớp của mạng tích chập có thể được coi là mẫu bất biến về mặt không gian hoặc phép chiếu cơ sở. Nếu đầu vào là đầu vào thứ nhất được phân tách thành nhiều kênh, như các kênh màu đỏ, xanh lục và xanh lam của hình ảnh màu, thì các mạng tích chập được huấn luyện trên đầu vào đó có thể được coi là ba chiều, với hai chiều không gian dọc theo các trục của hình ảnh và chiều thứ ba ghi lại thông tin màu sắc. Đầu ra của các kết nối tích chập có thể được xem xét để tạo ra ánh xạ đặc trưng trong lớp tiếp theo, với mỗi phần tử của ánh xạ đặc trưng (ví dụ, 220) nhận đầu vào từ một loạt các nơron ở lớp trước (ví dụ, ánh xạ đặc trưng 218) và từ mỗi trong số nhiều kênh. Các giá trị trong ánh xạ đặc trưng có thể được xử lý thêm với giá trị không tuyến tính, như sự chỉnh lưu, $\max(0, x)$. Các giá trị từ các nơron lân cận có thể được gộp lại, các giá trị này tương ứng với việc lấy mẫu giảm, và có thể tạo ra sự bất biến cục bộ bổ sung và giảm kích thước. Việc chuẩn hóa, tương ứng với việc làm trắng, cũng có thể được áp dụng nhờ ức chế bên giữa các nơron trong ánh xạ đặc trưng.

Hiệu suất của kiến trúc học sâu có thể tăng lên khi có nhiều điểm dữ liệu được gắn nhãn hơn hoặc khi công suất tính toán tăng thêm. Các mạng nơron sâu hiện đại được huấn luyện thường xuyên với tài nguyên tính toán lớn hơn hàng nghìn lần so với các tài nguyên có sẵn cho nhà nghiên cứu điển hình chỉ mười lăm năm trước. Các kiến trúc mới và mô hình huấn luyện có thể thúc đẩy thêm hiệu suất học sâu. Các khối tuyến tính chỉnh lưu có thể giảm vấn đề huấn luyện được gọi là biến mất gradient. Các kỹ thuật huấn luyện mới có thể giảm sự phù hợp quá mức và do đó cho phép các mô hình lớn hơn đạt được khả năng tổng quát hóa tốt hơn. Các kỹ thuật đóng gói có thể tách dữ liệu trong trường để tiếp thu nhất định và tăng thêm hiệu suất tổng thể.

Fig.3 là sơ đồ khối minh họa mạng tích chập sâu 350. Mạng tích chập sâu 350 có thể bao gồm nhiều loại lớp khác nhau dựa trên việc kết nối và dùng chung các trọng số. Như thể hiện trên Fig.3, mạng tích chập sâu 350 bao gồm các khối tích chập 354A, 354B. Mỗi trong số các khối tích chập 354A, 354B có thể được tạo cấu hình với lớp tích chập (CONV) 356, lớp chuẩn hóa (LNorm) 358, và lớp gộp tối đa (MAX POOL) 360.

Các lớp tích chập 356 có thể bao gồm một hoặc nhiều bộ lọc tích chập, các bộ lọc này có thể được áp dụng cho dữ liệu đầu vào để tạo ra ánh xạ đặc trưng. Mặc dù chỉ có hai trong số các khối tích chập 354A, 354B được thể hiện, nhưng sáng chế không quá hạn chế, và thay vào đó, số lượng bất kỳ của các khối tích chập 354A, 354B đều có thể được bao gồm trong mạng tích chập sâu 350 theo ưu tiên thiết kế. Lớp chuẩn hóa 358 có thể chuẩn hóa đầu ra của các bộ lọc tích chập. Ví dụ, lớp chuẩn hóa 358 có thể thực hiện làm trắng hoặc ức chế bên. Lớp gộp tối đa 360 có thể thực hiện tổng hợp lấy mẫu giảm trên không gian để giảm tính bất biến cục bộ và kích thước.

Ví dụ, các giàn bộ lọc song song của mạng tích chập sâu có thể được tải trên CPU 102 hoặc GPU 104 của SOC 100 để đạt được hiệu suất cao và mức tiêu thụ công suất thấp. Theo các phương án thay thế, các giàn bộ lọc song song có thể được tải trên DSP 106 hoặc ISP 116 của SOC 100. Ngoài ra, các mạng tích chập sâu 350 có thể truy cập các khối xử lý khác có thể có mặt trên SOC 100, như bộ xử lý cảm biến 114 và module điều hướng 120, lần lượt được dành riêng cho cảm biến và điều hướng.

Mạng tích chập sâu 350 cũng có thể bao gồm một hoặc nhiều lớp kết nối đầy đủ 362 (FC1 và FC2). Mạng tích chập sâu 350 có thể bao gồm thêm lớp hồi quy logistic (logistic regression - LR) 364. Giữa mỗi lớp 356, 358, 360, 362, 364 của mạng tích chập sâu 350 là các trọng số (không được thể hiện trên hình vẽ) sẽ được cập nhật. Đầu ra của mỗi trong số các lớp (ví dụ, 356, 358, 360, 362, 364) có thể đóng vai trò là đầu vào của một trong số các lớp kế tiếp (ví dụ, 356, 358, 360, 362, 364) trong mạng tích chập sâu 350 để học các dạng biểu diễn đặc trưng phân cấp từ dữ liệu đầu vào 352 (ví dụ, hình ảnh, âm thanh, video, dữ liệu cảm biến và/hoặc dữ liệu đầu vào khác) được cung cấp ở lớp thứ nhất của các khối tích chập 354A. Đầu ra của mạng tích chập sâu 350 là điểm phân loại 366 cho dữ liệu đầu vào 352. Điểm phân loại 366 có thể là tập hợp các xác suất, trong đó mỗi xác suất là xác suất của dữ liệu đầu vào bao gồm một tính năng từ tập hợp các tính năng.

Việc lưu trữ vào bộ nhớ của các mạng nơ-ron nhân tạo (ví dụ, các mạng nơ-ron nhị phân) có thể được giảm đáng kể khi các trọng số và các giá trị kích hoạt nơ-ron được nhị phân hóa thành -1 hoặc +1 (không gian [-1, +1]). Tuy nhiên, logic CMOS số làm việc trong không gian [0,1]. Do đó, việc biến đổi xảy ra giữa các thiết bị CMOS số, sử dụng cơ sở [0,1] và các mạng nơ-ron nhị phân, sử dụng cơ sở [-1, +1] trong phương án thực hiện nhị phân.

Ô nhớ có thể được tạo cấu hình để hỗ trợ hàm XNOR. Ví dụ, các bảng từ 1 đến 3 (ví dụ, các bảng tin cậy) minh họa các ánh xạ của mạng nơron nhị phân trong không gian $[0,1]$ cho phép nhân nhị phân trong không gian $[-1, +1]$ nhị phân hóa. Hàm logic hai đầu vào được minh họa trong các bảng tin cậy từ 1 đến 3.

Bảng 1 minh họa ví dụ về phép nhân nhị phân trong không gian $[-1, +1]$ nhị phân hóa. Ví dụ, phép nhân trong không gian $[-1, +1]$ nhị phân hóa tạo ra đầu ra 4 bit là $-1*-1, -1*+1, +1*-1, +1*+1$ (ví dụ, “1,-1,-1,1”).

Nhân	-1	+1
-1	+1	-1
+1	-1	+1

Bảng 1

Bảng 2 minh họa ví dụ về phương án thực hiện XNOR. Ô nhớ có thể được tạo cấu hình để thực hiện hàm XNOR trên giá trị đầu vào thứ nhất (ví dụ, kích hoạt nơron nhị phân) và giá trị đầu vào thứ hai (ví dụ, trọng số tiếp hợp nhị phân) để tạo đầu ra nhị phân. Ví dụ, hàm XNOR chỉ đúng khi tất cả các giá trị đầu vào là đúng hoặc khi tất cả các giá trị đầu vào là sai. Nếu một số đầu vào là đúng và các đầu vào khác là sai, thì đầu ra của hàm XNOR là sai. Do đó, khi cả hai đầu vào (ví dụ, đầu vào thứ nhất và đầu vào thứ hai) đều sai (ví dụ, đầu vào thứ nhất là 0 và đầu vào thứ hai là 0), như được thể hiện trong Bảng 2, đầu ra là đúng (ví dụ, 1). Khi đầu vào thứ nhất là sai (0) và đầu vào thứ hai là đúng (1), thì đầu ra là sai (0). Khi đầu vào thứ nhất là đúng (1) và đầu vào thứ hai là sai (0), thì đầu ra là sai (0). Khi đầu vào thứ nhất là đúng (1) và đầu vào thứ hai là đúng (1), thì đầu ra là đúng (1). Do đó, bảng tin cậy cho hàm XNOR với hai đầu vào có đầu ra nhị phân là “1,0,0,1”.

Do đó, phép nhân nhị phân trong không gian $[-1, +1]$ được nhị phân ánh xạ lên đầu ra nhị phân của XNOR trong không gian $[0,1]$. Ví dụ, “1s” trong đầu ra 4-bit của không gian $[-1, +1]$ nhị phân ánh xạ lên “1s” trong đầu ra nhị phân của hàm XNOR và “-1s” trong đầu ra 4-bit của không gian $[-1, +1]$ được nhị phân ánh xạ lên các “0” trong đầu ra nhị phân của hàm XNOR.

XNOR	0	1
0	1	0
1	0	1

Bảng 2

Nhân	0	1
0	0	0
1	0	1

Bảng 3

Ngược lại, phép nhân nhị phân trong không gian $[-1, +1]$ được nhị phân không ánh xạ cho phép nhân nhị phân trong không gian $[0, 1]$ được nhị phân trong Bảng 3. Ví dụ, Bảng 3 minh họa phép nhân trong không gian $[0, 1]$ được nhị phân để tạo ra đầu ra 4 bit là “0,0,0,1”, không ánh xạ với đầu ra 4 bit là “1,-1,-1,1” trong không gian $[-1, +1]$ được nhị phân. Ví dụ, đầu ra 4 bit “0,0,0,1” chỉ bao gồm một bit đúng (ví dụ, bit cuối cùng) còn đầu ra 4 bit “1,-1,-1,1” bao gồm hai bit đúng (ví dụ, bit thứ nhất và bit cuối cùng).

Thông thường, các mạng nơon nhị phân được thực hiện với các hệ thống bộ nhớ có tính toán được thực hiện bằng cách tính toán XNOR tại mỗi ô bit và tổng hợp các kết quả dọc theo mỗi dòng bit để tạo ra các giá trị đầu ra. Tuy nhiên, việc thêm hàm XNOR vào mỗi ô bit là tốn kém. Ví dụ, số lượng bóng bán dẫn cho mỗi ô trong bộ nhớ (ví dụ, SRAM) tăng từ sáu hoặc tám lên mười hai, việc này tăng đáng kể kích thước ô và mức tiêu thụ công suất.

Các khía cạnh của sáng chế hướng đến việc giảm kích thước và cải thiện mức tiêu thụ công suất của bộ nhớ bằng cách loại bỏ XNOR khỏi mảng bộ nhớ có tính toán của mạng nơon nhị phân. Theo một khía cạnh, ngưỡng kích hoạt cho mỗi cột của mảng bộ nhớ có tính toán được điều chỉnh để tránh sử dụng hàm XNOR và (các) bóng bán dẫn tương ứng của nó trong mỗi ô bit. Ví dụ, bộ nhớ nhỏ hơn (ví dụ, tám bóng bán dẫn SRAM)

với các ô bit bộ nhớ nhỏ hơn có thể được dùng cho các mạng noron nhị phân bộ nhớ có tính toán.

Fig.4 minh họa kiến trúc làm ví dụ 400 cho mảng bộ nhớ có tính toán (CIM) của mạng noron nhân tạo, theo các khía cạnh của sáng chế. Bộ nhớ có tính toán là cách thực hiện các phép toán nhân và tích lũy trong mảng bộ nhớ. Mảng bộ nhớ bao gồm các dòng từ 404 (hoặc WL_1), 405 (hoặc WL_2)... 406 (hoặc WL_M) cũng như các dòng bit 401 (hoặc BL_1), 402 (hoặc BL_2)... 403 (hoặc BL_N). Các trọng số (ví dụ, các giá trị trọng số tiếp hợp nhị phân) được lưu trữ vào các ô bit của mảng bộ nhớ. Các kích hoạt đầu vào (ví dụ, giá trị đầu vào có thể là vectơ đầu vào) nằm trên các dòng từ. Việc nhân xảy ra ở mỗi ô bit và kết quả của phép nhân được xuất ra qua các dòng bit. Ví dụ, việc nhân bao gồm nhân các trọng số với các kích hoạt đầu vào tại mỗi ô bit. Thiết bị tính tổng (không được thể hiện trên hình vẽ), như thiết bị tính tổng điện áp/dòng điện, kết hợp với các dòng hoặc cột bit (ví dụ, 401) tính tổng đầu ra (ví dụ, điện tích, dòng điện hoặc điện áp) của các dòng bit và chuyển kết quả (ví dụ, đầu ra) cho bộ biến đổi từ tương tự sang số (analog-to-digital converter - ADC). Ví dụ, tổng của mỗi dòng bit được tính toán từ các đầu ra tương ứng của các ô bit của mỗi dòng bit.

Theo một khía cạnh của sáng chế, việc điều chỉnh ngưỡng kích hoạt được thực hiện ở mỗi cột (tương ứng với các dòng bit) thay vì ở mỗi ô bit để cải thiện hiệu suất khu vực.

Thông thường, bắt đầu với phương án thực hiện các mạng noron nhị phân trong đó các trọng số và kích hoạt noron được nhị phân hóa thành -1 hoặc +1 (không gian $[-1, +1]$), các phép toán nhân và tích lũy trở thành phép toán XNOR (trong ô bit) với kết quả XNOR và tổng mật độ cư trú của các kết quả XNOR. Ví dụ, tổng mật độ cư trú của dòng bit (population count of a bit line) bao gồm tổng các kết quả dương (ví dụ, "1") của mỗi ô bit của dòng bit. Hàm ADC tương tự bộ so sánh bằng cách sử dụng ngưỡng kích hoạt. Ví dụ, tổng mật độ cư trú được ADC so sánh với ngưỡng hoặc tiêu chuẩn kích hoạt. Nếu tổng mật độ cư trú lớn hơn ngưỡng kích hoạt, thì đầu ra của dòng bit là "1". Ngược lại, nếu tổng mật độ cư trú nhỏ hơn hoặc bằng ngưỡng kích hoạt, thì đầu ra của dòng bit là "0." Tuy nhiên, mong muốn loại bỏ hàm XNOR và (các) bóng bán dẫn tương ứng của nó trong mỗi ô bit để giảm kích thước của bộ nhớ. Các kỹ thuật cụ thể để điều chỉnh ngưỡng kích hoạt như sau:

Lớp tích chập của mạng noron (ví dụ, mạng noron nhị phân) có thể bao gồm các ô (ví dụ, ô bit) được tổ chức thành mảng (ví dụ, mảng bộ nhớ có tính toán). Các ô bao gồm các thiết bị tạo cổng trong đó mức điện tích trong các thiết bị được tạo cổng biểu diễn trọng số được lưu trữ của mảng. Mạng noron nhị phân XNOR được huấn luyện có mảng M hàng (ví dụ, kích thước của vector đầu vào) và N cột bao gồm $M \times N$ trọng số tiếp hợp nhị phân W_{ij} , trọng số này là trọng số trong giá trị nhị phân $[-1, +1]$, và N ngưỡng kích hoạt C_j . Các đầu vào cho mảng có thể tương ứng với các dòng từ và đầu ra có thể tương ứng với các dòng bit. Ví dụ, kích hoạt đầu vào, X_i , mà là đầu vào trong giá trị nhị phân, $[-1, +1]$, là $X_1, X_2 \dots X_M$. Tổng các tích của các đầu vào có các trọng số tương ứng ($\sum_i^M X_i W_{ij}$) được gọi là tổng có trọng số $Y_j = \sum_i^M X_i W_{ij}$.

Ví dụ, khi tổng trọng số lớn hơn ngưỡng kích hoạt C_j , thì đầu ra bằng một (1). Ngược lại, đầu ra bằng không (0). Mạng noron nhị phân XNOR có thể được ánh xạ thành các mạng noron nhị phân không phải XNOR trong không gian $[0,1]$ đồng thời loại bỏ hàm XNOR và (các) bóng bán dẫn tương ứng của nó trong mỗi ô bit để giảm kích thước bộ nhớ. Theo một khía cạnh, mạng noron nhị phân XNOR có thể được ánh xạ thành mạng noron nhị phân không phải XNOR với sự điều chỉnh trong ngưỡng kích hoạt C_j của mỗi cột, như sau:

Phương trình 1 minh họa mối quan hệ giữa tổng của các tích của các đầu vào có các trọng số tương ứng và ngưỡng kích hoạt C_j đối với mạng noron nhị phân XNOR:

$$Y_j = \sum_i^M X_i W_{ij} > C_j \quad (1)$$

trong đó

- X_i (e.g., $X_1, X_2 \dots X_M$) là đầu vào có giá trị nhị phân $[-1, +1]$
- W_{ij} là trọng số có giá trị nhị phân $[-1, +1]$ (ví dụ, ma trận $M \times N$);
- Y_j (e.g., $Y_1, Y_2 \dots Y_N$) là đầu ra của các dòng bit có giá trị nhị phân $[-1, +1]$

Thông thường, các phương án thực hiện của mạng noron nhị phân bộ nhớ có tính toán đạt được bằng cách hoàn thành các phép toán XNOR tại mỗi ô bit và tính tổng kết quả cho mỗi dòng bit. Tuy nhiên, việc thêm hàm XNOR trong mỗi ô bit làm tăng diện tích bố trí và tăng mức tiêu thụ công suất. Ví dụ, số lượng bóng bán dẫn trong mỗi ô trong bộ nhớ (ví dụ, SRAM) tăng từ sáu hoặc tám lên mười hai. Do đó, các khía cạnh của sáng chế hướng đến việc biến đổi các phép toán nhân và tích lũy cho mảng bộ nhớ có tính toán của

mạng noron nhân tạo (ví dụ, mạng noron nhị phân) từ không gian $[-1, +1]$ cho không gian $[0,1]$ bằng cách sử dụng các điều chỉnh ngưỡng kích hoạt.

Mạng noron nhị phân được biến đổi từ không gian $[-1, +1]$ (trong đó $Y_j = \sum_i^M X_i W_{ij} > C_j$) sang không gian $[0,1]$ và đầu ra của các dòng bit trong không gian $[0,1]$ được so sánh với ngưỡng khác (ví dụ, việc suy ra của nó được mô tả bên dưới), như sau:

$$\sum_i^M \tilde{X}_i \tilde{W}_{ij} > \tilde{C}_j + \tilde{C}_b \quad (2)$$

- $\tilde{X}_i$ là đầu vào có giá trị nhị phân $[0, 1]$
- $\tilde{W}_{ij}$ là trọng số có giá trị nhị phân $[0, 1]$

trong đó $\tilde{C}_j$ tương ứng với kích hoạt điều chỉnh được (hoặc ngưỡng kích hoạt được điều chỉnh) trong không gian $[0,1]$ và $\tilde{C}_b$ đại diện cho độ lệch biến đổi (ví dụ, tham chiếu dòng điện phân cực chuyển đổi) trong không gian $[0,1]$.

Điều chỉnh ngưỡng kích hoạt như được mô tả ở đây cho phép tránh/bỏ phương án thực hiện hàm XNOR trong mỗi ô bit trong khi nhận được kết quả tương tự như khi hàm XNOR được thực hiện. Việc điều chỉnh ngưỡng kích hoạt cho phép sử dụng ô bit nhớ đơn giản hơn và nhỏ hơn (ví dụ, SRAM tám bóng bán dẫn (eight transistor - 8T)).

Các phương trình sau (3 và 4) bao gồm các biến để ánh xạ mạng trong không gian $[-1, +1]$ cho không gian $[0,1]$:

$$X_i = 2\tilde{X}_i - 1 \quad (3)$$

$$W_{ij} = 2\tilde{W}_{ij} - 1, \quad (4)$$

Chèn các giá trị của phương trình ba (3) và bốn (4) vào phương trình 1, ngưỡng kích hoạt đã điều chỉnh có thể được xác định qua biến đổi hoặc biến đổi giữa không gian $[-1, +1]$ và không gian $[0,1]$. Sau đây là nguồn gốc của ngưỡng kích hoạt đã điều chỉnh:

$$Y_j = \sum_i^M X_i W_{ij} = \sum_i^M (2\tilde{X}_i - 1)(2\tilde{W}_{ij} - 1) \quad (5)$$

Mở rộng phương trình 5:

$$Y_j = 4 \sum_i^M (\tilde{X}_i \cdot \tilde{W}_{ij}) - 2 \sum_i^M \tilde{W}_{ij} - 2 \sum_i^M \tilde{X}_i + \sum_i^M 1 \quad (6)$$

So sánh đầu ra của các dòng bit Y_j có giá trị nhị phân $[-1, +1]$ với ngưỡng kích hoạt C_j trong không gian $[-1, +1]$ (như trong phương trình 1)

$$Y_j > C_j$$

Chèn giá trị đầu ra của các dòng bit Y_j trong phương trình 6 vào phương trình 1 để có được tổng mật độ cư trú trên mỗi dòng bit (population count per bit line) trong không gian $[0,1]$ ($\sum_i^M (\tilde{X}_i \cdot \tilde{W}_{ij})$)

$$4 \sum_i^M (\tilde{X}_i \cdot \tilde{W}_{ij}) - 2 \sum_i^M \tilde{W}_{ij} - 2 \sum_i^M \tilde{X}_i + \sum_i^M 1 > C_j \quad (7)$$

$$4 \sum_i^M (\tilde{X}_i \cdot \tilde{W}_{ij}) > 2 \sum_i^M \tilde{W}_{ij} + 2 \sum_i^M \tilde{X}_i - \sum_i^M 1 + C_j \quad (8)$$

$$\sum_i^M (\tilde{X}_i \cdot \tilde{W}_{ij}) > \frac{1}{2} \sum_i^M \tilde{W}_{ij} + \frac{C_j - M}{4} + \frac{1}{2} \sum_i^M \tilde{X}_i \quad (9)$$

Ánh xạ tổng mật độ cư trú trên mỗi dòng bit trong không gian $[0,1]$ ($\sum_i^M (\tilde{X}_i \cdot \tilde{W}_{ij})$) trong phương trình 9 cho tổng mật độ cư trú trên dòng bit trong không gian $[-1, +1]$ ($\sum_i^M X_i W_{ij}$), N ngưỡng kích hoạt C_j trong không gian $[-1, +1]$ ánh xạ lên các kích hoạt điều chỉnh được $\tilde{C}_j$ cũng như độ lệch chuyển đổi $\tilde{C}_b$ trong không gian $[0,1]$. Ví dụ, để thu được mạng noron nhị phân không phải XNOR trong không gian $[0,1]$ trong khi loại bỏ hàm XNOR và (các) bóng bán dẫn tương ứng của nó, các kích hoạt điều chỉnh được $\tilde{C}_j$ cũng như độ lệch chuyển đổi $\tilde{C}_b$ trong không gian $[0,1]$ được sử dụng.

Như được thể hiện trên phương trình 9, hàm $\frac{1}{2} \sum_i^M \tilde{W}_{ij} + \frac{C_j - M}{4}$ được dùng để xác định các kích hoạt điều chỉnh được $\tilde{C}_j$. Ví dụ, phần tử $\frac{C_j - M}{4}$ tương ứng với các giá trị kích hoạt của các kích hoạt điều chỉnh được $\tilde{C}_j$. Các kích hoạt điều chỉnh được $\tilde{C}_j$ không phụ thuộc vào các đầu vào kích hoạt X_i . Các kích hoạt điều chỉnh được $\tilde{C}_j$ cũng được tạo ra từ các thông số xác định trước (ví dụ, các trọng số tiếp hợp nhị phân W_{ij}) mà có thể được thay đổi để điều chỉnh các ngưỡng kích hoạt C_j .

Hàm số $\frac{1}{2} \sum_i^M \tilde{X}_i$ được dùng để xác định độ lệch biến đổi. $\tilde{C}_b$. Độ lệch biến đổi $\tilde{C}_b$ chỉ phụ thuộc vào các kích hoạt đầu vào. Điều này có nghĩa là độ lệch biến đổi liên tục thay đổi khi nhận được các đầu vào mới.

Ví dụ, khi $C_j = 0$; $\sum_i^M \tilde{W}_{ij} = \frac{1}{2} M$; và $\sum_i^M \tilde{X}_i = \frac{1}{2} M$, giá trị $\tilde{C}_j$, $\tilde{C}_b$ và $\sum_i^M (\tilde{X}_i \cdot \tilde{W}_{ij})$ được tính toán như sau đối với phương trình 9:

$$\tilde{C}_j = \frac{1}{2} \sum_i^M \tilde{W}_{ij} + \frac{C_j - M}{4} = \frac{1}{2} * \frac{1}{2} M + \frac{0 - M}{4} = \frac{M}{4} - \frac{M}{4} = 0;$$

$$\tilde{C}_b = \frac{1}{2} \sum_i^M \tilde{X}_i = \frac{1}{2} * \frac{1}{2} M = \frac{1}{4} M; \text{ và}$$

$$\sum_i^M (\tilde{X}_i \cdot \tilde{W}_{ij}) > \tilde{C}_j + \tilde{C}_b = 0 + \frac{1}{4} M = \frac{1}{4} M$$

Tương tự, khi $C_j = \frac{1}{2}M$; $\sum_i^M \widetilde{W}_{ij} = \frac{1}{2}M$; và $\sum_i^M \widetilde{X}_i = \frac{1}{2}M$, các giá trị $\widetilde{C}_j$, $\widetilde{C}_b$ và $\sum_i^M (\widetilde{X}_i \cdot \widetilde{W}_{ij})$ được tính như sau đối với phương trình 9:

$$\widetilde{C}_j = \frac{1}{2} \sum_i^M \widetilde{W}_{ij} + \frac{C_j - M}{4} = \frac{1}{2} * \frac{1}{2}M + \frac{\frac{1}{2}M - M}{4} = \frac{M}{4} - \frac{M}{8} = \frac{1}{8}M;$$

$$\widetilde{C}_b = \frac{1}{2} \sum_i^M \widetilde{X}_i = \frac{1}{2} * \frac{1}{2}M = \frac{1}{4}M; \text{ và}$$

$$\sum_i^M (\widetilde{X}_i \cdot \widetilde{W}_{ij}) > \widetilde{C}_j + \widetilde{C}_b = \frac{1}{8}M + \frac{1}{4}M = \frac{3}{8}M$$

Trong một số khía cạnh, hàm $\sum_i^M \widetilde{X}_i$ có thể được dùng để tạo ra độ lệch biến đổi $\widetilde{C}_b$ từ cột tham chiếu (ví dụ, dòng bit không phải là một phần của N dòng bit, được dùng làm tham chiếu) bằng cách thiết lập trọng số trong giá trị nhị phân $\widetilde{W}_{ij}$ bằng 1. Chèn giá trị $\widetilde{W}_{ij}$ bằng 1 vào phương trình 9, phương trình kết quả như sau:

$$\sum_i^M (\widetilde{X}_i \cdot 1) > \frac{1}{2}M + \frac{C_j - M}{4} + \frac{1}{2} \sum_i^M \widetilde{X}_i \quad (9)$$

Tổng mật độ cư trú bằng $\sum_i^M \widetilde{X}_i \cdot 1$, có thể được dùng để bù dòng kích hoạt. Do đó, chỉ một cột được định rõ để xác định giá trị lệch tham chiếu hoặc độ lệch biến đổi cho toàn bộ mảng bộ nhớ.

Fig.5 minh họa kiến trúc làm ví dụ 500 để thực hiện các phép toán XNOR tương đương bằng cách điều chỉnh ngưỡng cột của mảng bộ nhớ có tính toán của mạng nơ-ron nhân tạo, theo các khía cạnh của sáng chế. Kiến trúc bao gồm mảng bộ nhớ bao gồm cột tham chiếu 502, bộ so sánh 504, các dòng bit 506 (tương ứng với các cột của mảng bộ nhớ), và các dòng từ 508 (tương ứng với các hàng của mảng bộ nhớ). Các kích hoạt đầu vào được nhận qua các dòng từ. Trong một ví dụ, các giá trị nhị phân của các kích hoạt đầu vào X_i là 10101011.

Phép nhân (ví dụ, $X_i W_{ij}$) xảy ra ở mỗi ô bit và kết quả của phép nhân từ mỗi ô bit được xuất ra qua các dòng bit 506. Thiết bị tính tổng 510 kết hợp với mỗi trong số các dòng bit 506 tính tổng mỗi đầu ra của các ô bit của mảng bộ nhớ và chuyển kết quả cho bộ so sánh 504. Bộ so sánh 504 có thể là một phần của bộ biến đổi ADC, được thể hiện trên Fig.4. Theo một khía cạnh, đầu ra của mỗi ô bit kết hợp với dòng bit thứ nhất được tính tổng riêng với đầu ra của mỗi ô bit kết hợp với dòng bit thứ hai.

Việc điều chỉnh ngưỡng kích hoạt xảy ra ở mỗi cột (tương ứng với các dòng bit) thay vì ở mỗi ô bit, để cải thiện hiệu quả diện tích. Ví dụ, tổng đầu ra (hoặc tổng mật độ cư trú) của mỗi ô bit kết hợp với dòng bit thứ nhất được cung cấp cho bộ so sánh 504 làm

đầu vào thứ nhất. Đầu vào thứ hai của bộ so sánh bao gồm tổng kích hoạt điều chỉnh được $\widetilde{C}_j$ và độ lệch biến đổi $\widetilde{C}_b$ (ví dụ, tham chiếu dòng điện phân cực chuyển đổi). Ví dụ, độ lệch biến đổi $\widetilde{C}_b$ có thể được lập trình thành tiêu chuẩn (ví dụ, cột tham chiếu 502) cho mỗi trong số các dòng bit 506. Khi tổng mật độ cư trú lớn hơn tổng kích hoạt điều chỉnh được $\widetilde{C}_j$ và độ lệch biến đổi $\widetilde{C}_b$, đầu ra của bộ so sánh 504, tương ứng với đầu ra của dòng bit thứ nhất, là “1”. Ngược lại, nếu tổng mật độ cư trú nhỏ hơn hoặc bằng tổng của kích hoạt điều chỉnh được $\widetilde{C}_j$ và độ lệch biến đổi $\widetilde{C}_b$, thì đầu ra của bộ so sánh, tương ứng với đầu ra của dòng bit thứ nhất, là “0”. Do đó, mỗi tổng mật độ cư trú dòng bit được so sánh với tổng kích hoạt điều chỉnh được $\widetilde{C}_j$ và độ lệch biến đổi $\widetilde{C}_b$.

Fig.6 minh họa phương pháp 600 để thực hiện các phép toán XNOR tương đương bằng cách điều chỉnh các ngưỡng cột của mảng bộ nhớ có tính toán của các mạng neuron nhân tạo, theo các khía cạnh của sáng chế. Như được thể hiện trên Fig.6, tại khối 602, ngưỡng kích hoạt được tạo ra cho mỗi cột của mảng bộ nhớ có tính toán có thể được điều chỉnh dựa trên hàm của giá trị trọng số và giá trị kích hoạt. Tại khối 604, tham chiếu dòng điện phân cực chuyển đổi được tính toán dựa trên giá trị đầu vào từ vector đầu vào cho mảng bộ nhớ có tính toán, mảng bộ nhớ có tính toán được lập trình với tập hợp các giá trị trọng số. Mỗi trong số ngưỡng kích hoạt được điều chỉnh và tham chiếu dòng điện phân cực chuyển đổi được dùng làm ngưỡng để xác định các giá trị đầu ra của mảng bộ nhớ có tính toán. Mảng bộ nhớ có tính toán có cả cột và hàng.

Theo một khía cạnh khác của sáng chế, thiết bị để thực hiện các phép toán XNOR tương đương bằng cách điều chỉnh các ngưỡng cột của mảng bộ nhớ có tính toán của mạng neuron nhân tạo được mô tả. Thiết bị bao gồm phương tiện để điều chỉnh ngưỡng kích hoạt cho mỗi cột của mảng dựa trên hàm của giá trị trọng số và giá trị kích hoạt. Phương tiện điều chỉnh bao gồm mạng tích chập sâu 200, mạng tích chập sâu 350, lớp tích chập 232, SoC 100, CPU 102, kiến trúc 500, kiến trúc 400, và/hoặc khối tích chập 354A. Thiết bị còn bao gồm phương tiện để tính toán tham chiếu dòng điện phân cực chuyển đổi dựa trên giá trị đầu vào từ vector đầu vào. Phương tiện tính toán bao gồm mạng tích chập sâu 200, mạng tích chập sâu 350, lớp tích chập 232, SoC 100, CPU 102, kiến trúc 500, kiến trúc 400, và/hoặc khối tích chập 354A.

Thiết bị còn bao gồm phương tiện để so sánh tổng mật độ cư trú dòng bit với tổng của tham chiếu dòng điện phân cực chuyển đổi và ngưỡng kích hoạt được điều chỉnh để

xác định đầu ra của dòng bit. Phương tiện so sánh bao gồm bộ so sánh 504 trên Fig.5 và/hoặc bộ biến đổi tương tự sang số (ADC) trên Fig.4. Theo khía cạnh khác, phương tiện nói trên có thể là môđun bất kỳ hoặc thiết bị bất kỳ được tạo cấu hình để thực hiện các hàm thể hiện bởi phương tiện nói trên.

Các phép toán khác nhau của các phương pháp được mô tả trên đây có thể được thực hiện bởi phương tiện thích hợp bất kỳ có khả năng thực hiện các hàm tương ứng. Phương tiện có thể bao gồm (các) thành phần phần cứng và/hoặc phần mềm khác nhau và/hoặc (các) môđun, bao gồm, nhưng không bị giới hạn ở mạch, mạch tích hợp chuyên dụng (application specific integrated circuit - ASIC), hoặc bộ xử lý. Nói chung, nếu có các phép toán được minh họa trên các hình vẽ, thì các phép toán này có thể có các thành phần phương tiện cộng hàm bổ sung tương ứng với cách đánh số tương tự.

Như được dùng ở đây, thuật ngữ “xác định” bao gồm nhiều phép toán khác nhau. Ví dụ, việc “xác định” có thể bao gồm việc tính, tính toán, xử lý, lấy đạo hàm, khảo sát, tra cứu (ví dụ, tra cứu trong bảng, cơ sở dữ liệu hoặc một cấu trúc dữ liệu khác), xác minh và tương tự. Ngoài ra, bước “xác định” có thể bao gồm bước nhận (ví dụ, nhận thông tin), truy cập (ví dụ, truy cập dữ liệu trong bộ nhớ) và tương tự. Ngoài ra, bước “xác định” có thể bao gồm bước giải quyết, chọn lọc, chọn, thiết lập và tương tự.

Như được dùng ở đây, cụm từ đề cập đến “ít nhất một trong số” danh sách các mục đề cập đến kết hợp bất kỳ của các mục đó, bao gồm các thành phần riêng. Ví dụ, “ít nhất một trong số: a, b, hoặc c” dự định bao gồm : a, b, c, a-b, a-c, b-c, và a-b-c.

Các khối logic, môđun và hệ mạch minh họa được mô tả theo sáng chế có thể được thực thi hoặc thực hiện bằng bộ xử lý đa dụng, bộ xử lý tín hiệu số (DSP), mạch tích hợp chuyên dụng (ASIC), mảng cổng lập trình được theo trường (field programmable gate array - FPGA) hoặc thiết bị logic lập trình được (programmable logic device - PLD), cổng rời rạc hoặc logic bóng bán dẫn, các thành phần phần cứng rời rạc hoặc tổ hợp bất kỳ của chúng được thiết kế để thực hiện các hàm được mô tả ở đây. Bộ xử lý đa dụng có thể là bộ vi xử lý, nhưng thay vào đó, bộ xử lý có thể là bất kỳ bộ xử lý có sẵn trên thị trường, bộ điều khiển, bộ vi điều khiển, hoặc máy trạng thái. Bộ xử lý cũng có thể được thực thi dưới dạng tổ hợp của các thiết bị tính toán, ví dụ, tổ hợp của DSP và bộ vi xử lý, nhiều bộ vi xử lý, một hoặc nhiều bộ vi xử lý kết hợp với lõi DSP, hoặc cấu hình như vậy khác bất kỳ.

Các bước của phương pháp hoặc thuật toán được mô tả liên quan đến sáng chế có thể được thể hiện trực tiếp bằng phần cứng, bằng modul phần mềm được thực thi bởi bộ xử lý, hoặc bằng tổ hợp của cả hai. Modul phần mềm có thể nằm ở dạng phương tiện lưu trữ bất kỳ đã biết trong lĩnh vực này. Một số ví dụ về phương tiện lưu trữ có thể được dùng bao gồm bộ nhớ truy cập ngẫu nhiên (random access memory - RAM), bộ nhớ chỉ đọc (read only memory - ROM), bộ nhớ flash, bộ nhớ chỉ đọc lập trình được xóa được (erasable programmable read-only memory - EPROM), bộ nhớ chỉ đọc lập trình được xóa được bằng điện (electrically erasable programmable read-only memory - EEPROM), thanh ghi, đĩa cứng, đĩa di động, CD-ROM, v.v. Modul phần mềm có thể bao gồm một lệnh, hoặc nhiều lệnh, và có thể được phân bố trên một số đoạn mã khác nhau, trong số các chương trình khác nhau và trên nhiều phương tiện lưu trữ. Phương tiện lưu trữ được ghép nối bộ xử lý sao cho bộ xử lý có thể đọc thông tin từ, và ghi thông tin vào phương tiện lưu trữ. Theo cách khác, phương tiện lưu trữ có thể được tích hợp với bộ xử lý.

Các phương pháp được mô tả ở đây bao gồm một hoặc nhiều bước hoặc hành động để thực hiện phương pháp được mô tả. Các bước và/hoặc các hoạt động của phương pháp có thể được hoán đổi cho nhau mà không nằm ngoài phạm vi yêu cầu bảo hộ. Nói cách khác, trừ khi thứ tự cụ thể của các bước hoặc hành động được chỉ rõ, thứ tự và/hoặc việc sử dụng của các bước hoặc hành động cụ thể có thể được thay đổi mà không nằm ngoài phạm vi của phần yêu cầu bảo hộ.

Các hàm được mô tả có thể được thực hiện trong phần cứng, phần mềm, firmware hoặc tổ hợp bất kỳ của chúng. Nếu được thực hiện trong phần cứng, cấu hình phần cứng làm ví dụ có thể bao gồm hệ thống xử lý trong thiết bị. Hệ thống xử lý có thể được thực hiện với kiến trúc bus. Bus có thể bao gồm một số bus và cầu liên kết bất kỳ tùy thuộc vào ứng dụng cụ thể của hệ thống xử lý và các ràng buộc thiết kế tổng thể. Bus có thể liên kết các mạch với nhau bao gồm bộ xử lý, phương tiện đọc được bằng máy, và giao diện bus. Giao diện bus có thể được dùng để kết nối bộ thích ứng mạng, trong số những cái khác, với hệ thống xử lý qua bus. Bộ thích ứng mạng có thể được dùng để thực thi các hàm xử lý. Đối với một số khía cạnh, giao diện người dùng (ví dụ, bàn phím, màn hình hiển thị, chuột, cần điều khiển, v.v.) cũng có thể được kết nối với bus. Bus này cũng có thể liên kết với các mạch khác nhau như các nguồn định thời, thiết bị ngoại vi, bộ điều chỉnh điện áp,

mạch quản lý công suất và tương tự, mà được biết rộng rãi trong lĩnh vực này và do vậy, không được mô tả thêm.

Bộ xử lý có nhiệm vụ quản lý bus và xử lý chung, bao gồm thực thi phần mềm được lưu trữ trên phương tiện đọc được bằng máy. Bộ xử lý có thể được cài đặt với một hoặc nhiều bộ xử lý đa dụng và/hoặc bộ xử lý chuyên dụng. Ví dụ gồm bộ vi xử lý, bộ vi điều khiển, bộ xử lý DSP và mạch khác có thể thực thi phần mềm. Phần mềm được hiểu rộng là các lệnh, dữ liệu hoặc tổ hợp bất kỳ của chúng, dù được gọi là ngôn ngữ mô tả phần mềm, firmware, phần trung gian, vi mã, phần cứng, hoặc các loại khác. Phương tiện đọc được bằng máy có thể bao gồm, ví dụ, bộ nhớ truy cập ngẫu nhiên (RAM), bộ nhớ flash, bộ nhớ chỉ đọc (ROM), bộ nhớ chỉ đọc lập trình được (PROM), bộ nhớ chỉ đọc lập trình được xóa được (EPROM), bộ nhớ chỉ đọc lập trình được xóa được bằng điện (EEPROM), thanh ghi, đĩa từ, đĩa quang, ổ cứng hoặc phương tiện lưu trữ thích hợp khác bất kỳ, hoặc tổ hợp bất kỳ của chúng. Phương tiện đọc được bằng máy có thể được bao gồm trong sản phẩm chương trình máy tính. Sản phẩm chương trình máy tính có thể gồm vật liệu đóng gói.

Theo phương án triển khai phần cứng, phương tiện đọc được bằng máy có thể là một phần của hệ thống xử lý tách biệt với bộ xử lý. Tuy nhiên, người có hiểu biết trung bình trong lĩnh vực kỹ thuật này dễ dàng nhận ra, phương tiện đọc được bằng máy hoặc phần bất kỳ của nó có thể nằm ngoài hệ thống xử lý. Ví dụ, phương tiện đọc được bằng máy có thể gồm đường truyền, sóng mang được điều chế bởi dữ liệu, và/hoặc sản phẩm máy tính tách từ thiết bị, tất cả có thể được truy nhập bởi bộ xử lý qua giao diện bus. Theo cách khác, hoặc theo cách bổ sung, phương tiện đọc được bằng máy hoặc phần bất kỳ của chúng, có thể được tích hợp vào bộ xử lý, như trường hợp có bộ nhớ đệm và/hoặc các tập tin thanh ghi chung. Mặc dù một số thành phần có thể được mô tả là có vị trí cụ thể, như thành phần cục bộ, chúng cũng có thể được tạo cấu hình theo nhiều cách khác nhau, như các thành phần nhất định được tạo cấu hình như một phần của hệ thống tính toán phân tán.

Hệ thống xử lý có thể được tạo cấu hình như hệ thống xử lý đa dụng có một hoặc nhiều bộ vi xử lý cung cấp chức năng xử lý hoặc bộ nhớ ngoài cung cấp ít nhất một phần của phương tiện đọc được bằng máy, tất cả liên kết với mạch hỗ trợ khác qua cấu trúc bus ngoài. Ngoài ra, hệ thống xử lý có thể bao gồm một hoặc nhiều bộ xử lý neuron để thực hiện các mô hình neuron và các mô hình hệ thống neuron được mô tả ở đây. Với giải pháp

thay thế khác, hệ thống xử lý có thể được thực hiện bằng mạch ASIC với bộ xử lý, giao diện bus, giao diện người dùng, mạch hỗ trợ và ít nhất một phần của phương tiện đọc được bằng máy tích hợp vào một chip, hoặc với một hoặc nhiều mảng công lập trình được theo trường (FPGA), thiết bị logic lập trình được (PLD), bộ điều khiển, máy trạng thái, logic tạo cổng, các thành phần phần cứng rời rạc, hoặc mạch thích hợp khác bất kỳ hoặc tổ hợp bất kỳ của các mạch có thể thực hiện các chức năng khác nhau được mô tả trong toàn bộ sáng chế. Người có hiểu biết trung bình trong lĩnh vực kỹ thuật này sẽ nhận ra cách tốt nhất để thực thi chức năng được mô tả đối với hệ thống xử lý tùy thuộc vào ứng dụng cụ thể và các ràng buộc thiết kế tổng thể áp dụng trên toàn hệ thống.

Phương tiện đọc được bằng máy có thể gồm một số module phần mềm. Các module phần mềm gồm các lệnh, khi được bộ xử lý thực thi, khiến cho hệ thống xử lý thực hiện các chức năng khác nhau. Module phần mềm có thể bao gồm module truyền và module nhận. Mỗi module phần mềm có thể nằm trong một thiết bị lưu trữ hoặc được phân tán trên nhiều thiết bị lưu trữ. Ví dụ, module phần mềm có thể được nạp vào RAM từ ổ cứng khi xảy ra sự kiện kích hoạt. Trong khi thực thi module phần mềm, bộ xử lý có thể nạp một số trong số các lệnh vào bộ nhớ sẵn để tăng tốc độ truy cập. Sau đó, một hoặc nhiều đường bộ nhớ đệm có thể được nạp vào tập tin thanh ghi chung để thực thi bằng bộ xử lý. Khi đề cập đến chức năng của module phần mềm dưới đây, cần hiểu rằng chức năng này được thực thi bởi bộ xử lý khi thực hiện các lệnh từ module phần mềm đó. Hơn nữa, cần đánh giá cao rằng các khía cạnh của sáng chế dẫn đến những cải tiến đối với việc thực hiện chức năng của bộ xử lý, máy tính, máy hoặc hệ thống khác thực hiện các khía cạnh như vậy.

Nếu được thực thi trong phần mềm, các chức năng có thể được lưu trữ hoặc truyền dưới dạng một hoặc nhiều lệnh hoặc mã trên phương tiện đọc được bằng máy tính. Phương tiện đọc được bằng máy tính bao gồm cả phương tiện lưu trữ máy tính và phương tiện truyền thông bao gồm phương tiện bất kỳ mà hỗ trợ truyền chương trình máy tính từ vị trí này đến vị trí khác. Phương tiện lưu trữ có thể là phương tiện khả dụng bất kỳ mà có thể được truy cập bởi máy tính. Theo một ví dụ, và không giới hạn, phương tiện lưu trữ đọc được bởi máy tính như vậy có thể bao gồm RAM, ROM, EEPROM, CD-ROM hoặc bộ nhớ đĩa quang khác, bộ nhớ đĩa từ hoặc các thiết bị lưu trữ từ khác, hoặc phương tiện khác bất kỳ mà có thể được dùng để mang hoặc lưu trữ mã chương trình mong muốn dưới dạng các lệnh hoặc cấu trúc dữ liệu và máy tính có thể truy cập được. Ngoài ra, kết nối bất kỳ

được gọi một cách thích hợp là phương tiện đọc được bằng máy tính. Ví dụ, nếu phần mềm được truyền từ trang web, máy chủ hoặc nguồn từ xa khác nhờ sử dụng cáp đồng trục, cáp sợi quang, cặp dây xoắn, đường dây thuê bao số (digital subscriber line - DSL), hoặc các công nghệ không dây như hồng ngoại, sóng vô tuyến, vi sóng, thì cáp đồng trục, cáp sợi quang, cặp dây xoắn, DSL, hoặc các công nghệ không dây như hồng ngoại, sóng vô tuyến, vi sóng này được bao hàm trong định nghĩa về phương tiện. Đĩa từ và đĩa quang, như được dùng ở đây, bao gồm đĩa nén (compact disc - CD), đĩa laze, đĩa quang, đĩa đa dụng kỹ thuật số (digital versatile disc - DVD), đĩa mềm và đĩa Blu-ray®, trong đó các đĩa từ thường tái tạo dữ liệu theo cách từ tính, còn đĩa quang tái tạo dữ liệu theo cách quang học bằng laze. Do đó, theo một số khía cạnh phương tiện đọc được bằng máy tính có thể bao gồm phương tiện bất biến đọc được bằng máy tính (ví dụ, phương tiện hữu hình). Ngoài ra, theo một số khía cạnh khác phương tiện đọc được bằng máy tính có thể bao gồm phương tiện khả biến đọc được bằng máy tính (ví dụ, tín hiệu). Các dạng kết hợp của các loại phương tiện nêu trên cũng được bao gồm trong phạm vi của phương tiện bằng máy tính.

Do vậy, một số khía cạnh có thể bao gồm sản phẩm chương trình máy tính để thực hiện các hoạt động được trình bày theo sáng chế. Ví dụ, sản phẩm chương trình máy tính như vậy có thể bao gồm phương tiện đọc được bởi máy tính có các lệnh được lưu trữ (và/hoặc được mã hóa) trên đó, các lệnh này được thực thi bởi một hoặc nhiều bộ xử lý để thực hiện các hoạt động được mô tả ở đây. Theo một số khía cạnh, sản phẩm chương trình máy tính có thể bao gồm vật liệu đóng gói.

Ngoài ra, cần phải hiểu rõ rằng các môđun và/hoặc các phương tiện phù hợp khác để thực hiện các phương pháp và kỹ thuật được mô tả ở đây có thể được tải xuống và/hoặc ngược lại được thu bằng đầu cuối người dùng và/hoặc trạm gốc nếu có thể áp dụng được. Ví dụ, thiết bị như vậy có thể được nối với máy chủ để hỗ trợ việc truyền của phương tiện thực hiện các phương pháp được mô tả ở đây. Theo cách khác, các phương pháp khác nhau được mô tả ở đây có thể được cung cấp thông qua phương tiện lưu trữ (ví dụ, RAM, ROM, phương tiện lưu trữ vật lý như đĩa nén (CD) hoặc đĩa mềm, v.v.), sao cho đầu cuối người dùng và/hoặc trạm gốc có thể thu được các phương pháp khác nhau khi nối hoặc cung cấp phương tiện lưu trữ cho thiết bị. Ngoài ra, mọi kỹ thuật thích hợp khác để cung cấp các phương pháp và các kỹ thuật được mô tả ở đây cho thiết bị đều có thể được dùng.

Cần phải hiểu rằng yêu cầu bảo hộ không bị giới hạn ở cấu hình và bộ phận chính xác được minh họa trên đây. Các cải biến, thay đổi và biến đổi khác nhau có thể được thực hiện theo sự sắp xếp, hoạt động và chi tiết của các phương pháp và thiết bị được mô tả ở trên mà không nằm ngoài phạm vi của các yêu cầu bảo hộ.

YÊU CẦU BẢO HỘ

1. Thiết bị thực hiện các phép toán XNOR tương đương bằng cách điều chỉnh các ngưỡng cột của mảng bộ nhớ có tính toán có các hàng và các cột, thiết bị này bao gồm:

mảng bộ nhớ có tính toán bao gồm các hàng và cột, mảng bộ nhớ có tính toán được tạo cấu hình để:

điều chỉnh ngưỡng kích hoạt được tạo ra cho mỗi cột của mảng bộ nhớ có tính toán dựa trên hàm của giá trị trọng số và giá trị kích hoạt; và

tính toán tham chiếu dòng điện phân cực chuyển đổi dựa trên giá trị đầu vào từ vector đầu vào cho mảng bộ nhớ có tính toán, mảng bộ nhớ có tính toán được lập trình với tập hợp giá trị trọng số, trong đó ngưỡng kích hoạt được điều chỉnh và tham chiếu dòng điện phân cực chuyển đổi được dùng làm ngưỡng để xác định các giá trị đầu ra của mảng bộ nhớ có tính toán.

2. Thiết bị theo điểm 1, còn bao gồm bộ so sánh được tạo cấu hình để so sánh tổng mật độ cư trú dòng bit (bit line population count) với tổng của tham chiếu dòng điện phân cực chuyển đổi và ngưỡng kích hoạt được điều chỉnh để xác định đầu ra của dòng bit.

3. Thiết bị theo điểm 1, trong đó mạng nơron nhân tạo mà có mảng bộ nhớ có tính toán bao gồm mạng nơron nhị phân.

4. Thiết bị theo điểm 1, trong đó ngưỡng kích hoạt nhỏ hơn một nửa số hàng của mảng bộ nhớ có tính toán, số hàng tương ứng với kích thước của vector đầu vào.

5. Thiết bị theo điểm 1, trong đó tham chiếu dòng điện phân cực chuyển đổi nhỏ hơn một nửa số hàng của mảng bộ nhớ có tính toán, số hàng tương ứng với kích thước của vector đầu vào.

6. Phương pháp do bộ xử lý thực hiện để thực hiện các phép toán XNOR tương đương bằng cách điều chỉnh các ngưỡng cột của mảng bộ nhớ có tính toán có các hàng và các cột, phương pháp này bao gồm các bước:

điều chỉnh ngưỡng kích hoạt được tạo ra cho mỗi cột của mảng bộ nhớ có tính toán có các hàng và cột dựa trên hàm của giá trị trọng số và giá trị kích hoạt;

tính toán tham chiếu dòng điện phân cực chuyển đổi dựa trên giá trị đầu vào từ vector đầu vào cho mảng bộ nhớ có tính toán, mảng bộ nhớ có tính toán được lập trình với tập hợp giá trị trọng số, trong đó ngưỡng kích hoạt được điều chỉnh và tham chiếu dòng điện phân cực chuyển đổi được dùng làm ngưỡng để xác định các giá trị đầu ra của mảng bộ nhớ có tính toán.

7. Phương pháp do bộ xử lý thực hiện theo điểm 6, còn bao gồm bước so sánh tổng mật độ cư trú dòng bit với tổng của tham chiếu dòng điện phân cực chuyển đổi và ngưỡng kích hoạt được điều chỉnh để xác định đầu ra của dòng bit.

8. Phương pháp do bộ xử lý thực hiện theo điểm 6, trong đó mạng nơron nhân tạo mà có mảng bộ nhớ có tính toán bao gồm mạng nơron nhị phân.

9. Phương pháp do bộ xử lý thực hiện theo điểm 6, trong đó ngưỡng kích hoạt nhỏ hơn một nửa số hàng của mảng bộ nhớ có tính toán, số hàng tương ứng với kích thước của vector đầu vào.

10. Phương pháp do bộ xử lý thực hiện theo điểm 6, trong đó tham chiếu dòng điện phân cực chuyển đổi nhỏ hơn một nửa số hàng của mảng bộ nhớ có tính toán, số hàng tương ứng với kích thước của vector đầu vào.

11. Phương tiện bất biến đọc được bằng máy tính ghi trên đó mã chương trình, mã chương trình này được thực thi bởi bộ xử lý và bao gồm:

mã chương trình để điều chỉnh ngưỡng kích hoạt được tạo ra cho mỗi cột của mảng bộ nhớ có tính toán có các hàng và các cột dựa trên hàm của giá trị trọng số và giá trị kích hoạt; và

mã chương trình để tính toán tham chiếu dòng điện phân cực chuyển đổi dựa trên giá trị đầu vào từ vector đầu vào cho mảng bộ nhớ có tính toán, mảng bộ nhớ có tính toán được lập trình với tập hợp giá trị trọng số, trong đó ngưỡng kích hoạt được điều chỉnh và

tham chiếu dòng điện phân cực chuyển đổi được dùng làm ngưỡng để xác định các giá trị đầu ra của mảng bộ nhớ có tính toán.

12. Phương tiện bất biến đọc được bằng máy tính theo điểm 11, phương tiện này còn bao gồm mã chương trình để so sánh tổng mật độ cư trú dòng bit với tổng của tham chiếu dòng điện phân cực chuyển đổi và ngưỡng kích hoạt được điều chỉnh để xác định đầu ra của dòng bit.

13. Phương tiện bất biến đọc được bằng máy tính theo điểm 11, trong đó mạng nơron nhân tạo chịu sự điều chỉnh và tính toán bao gồm mạng nơron nhị phân.

14. Phương tiện bất biến đọc được bằng máy tính theo điểm 11, trong đó ngưỡng kích hoạt nhỏ hơn một nửa số hàng của mảng bộ nhớ có tính toán, số hàng tương ứng với kích thước của vector đầu vào.

15. Phương tiện bất biến đọc được bằng máy tính theo điểm 11, trong đó tham chiếu dòng điện phân cực chuyển đổi nhỏ hơn một nửa số hàng của mảng bộ nhớ có tính toán, số hàng tương ứng với kích thước của vector đầu vào.

16. Thiết bị thực hiện các phép toán XNOR tương đương bằng cách điều chỉnh các ngưỡng cột của mảng bộ nhớ có tính toán có các hàng và các cột, thiết bị này bao gồm:

phương tiện để điều chỉnh ngưỡng kích hoạt được tạo ra cho mỗi cột của mảng bộ nhớ có tính toán có các hàng và các cột dựa trên hàm của giá trị trọng số và giá trị kích hoạt; và

phương tiện để tính toán tham chiếu dòng điện phân cực chuyển đổi dựa trên giá trị đầu vào từ vector đầu vào cho mảng bộ nhớ có tính toán, mảng bộ nhớ có tính toán được lập trình với tập hợp giá trị trọng số, trong đó ngưỡng kích hoạt được điều chỉnh và tham chiếu dòng điện phân cực chuyển đổi được dùng làm ngưỡng để xác định các giá trị đầu ra của mảng bộ nhớ có tính toán.

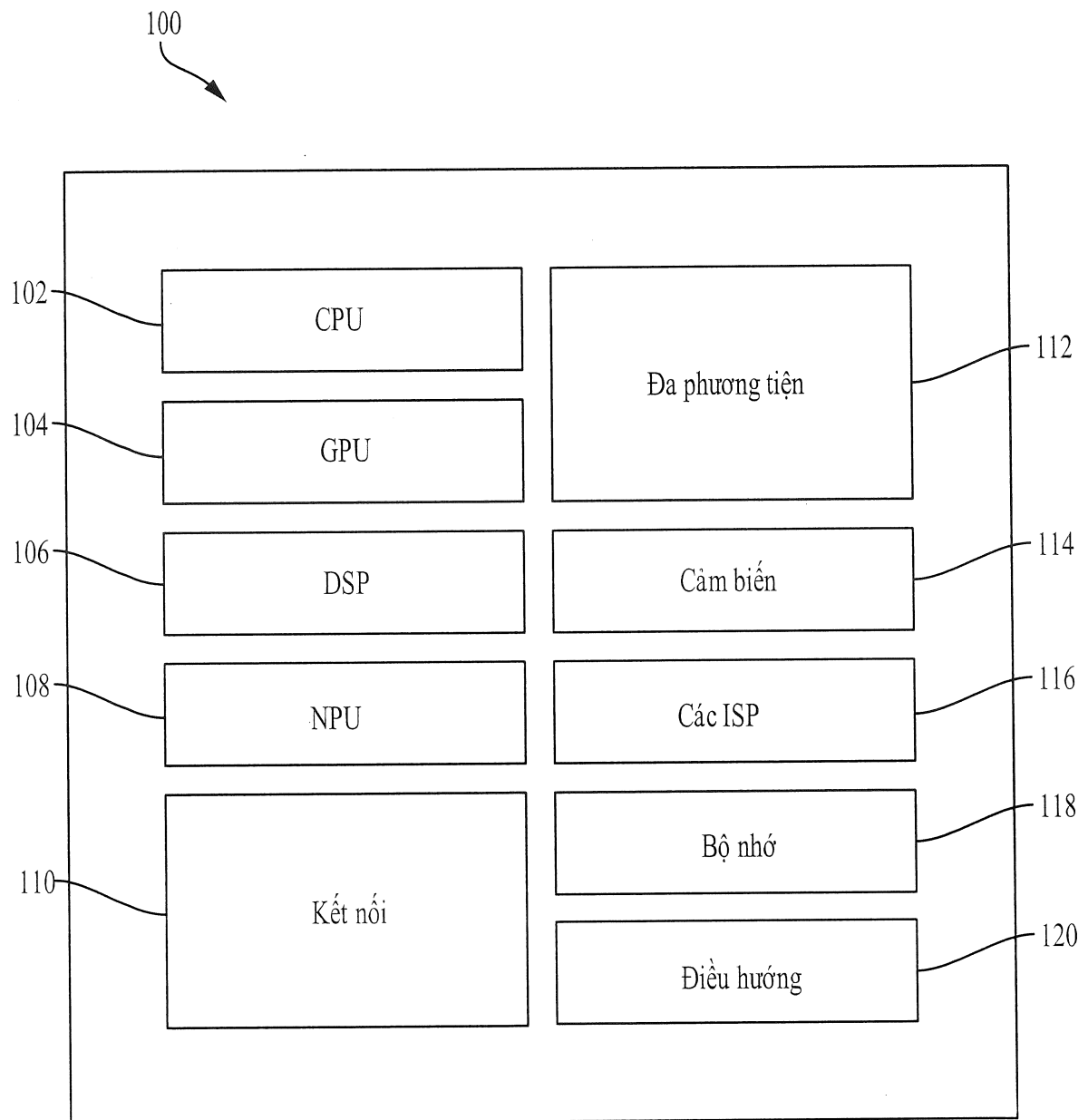
17. Thiết bị theo điểm 16, còn bao gồm phương tiện để so sánh tổng mật độ cư trú dòng bit với tổng của tham chiếu dòng điện phân cực chuyển đổi và ngưỡng kích hoạt được điều chỉnh để xác định đầu ra của dòng bit.

18. Thiết bị theo điểm 16, trong đó mạng nơron nhân tạo mà có mảng bộ nhớ có tính toán bao gồm mạng nơron nhị phân.

19. Thiết bị theo điểm 16, trong đó ngưỡng kích hoạt nhỏ hơn một nửa số hàng của mảng bộ nhớ có tính toán, số hàng tương ứng với kích thước của vector đầu vào.

20. Thiết bị theo điểm 16, trong đó tham chiếu dòng điện phân cực chuyển đổi nhỏ hơn một nửa số hàng của mảng bộ nhớ có tính toán, số hàng tương ứng với kích thước của vector đầu vào.

1/7

**Fig.1**

2/7

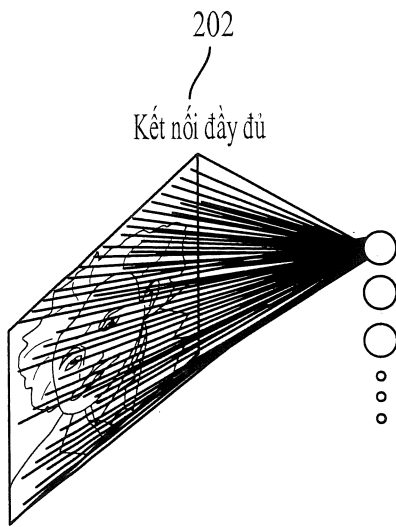


Fig.2A

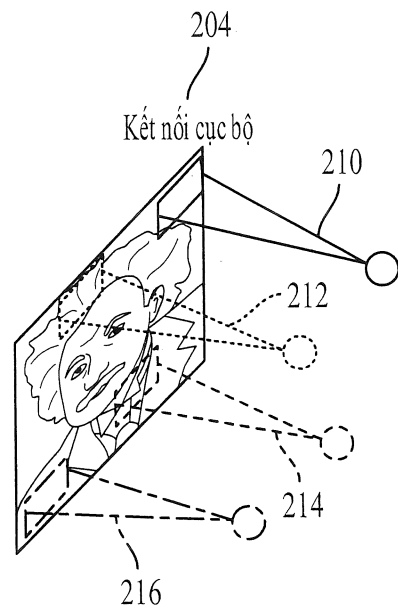


Fig.2B

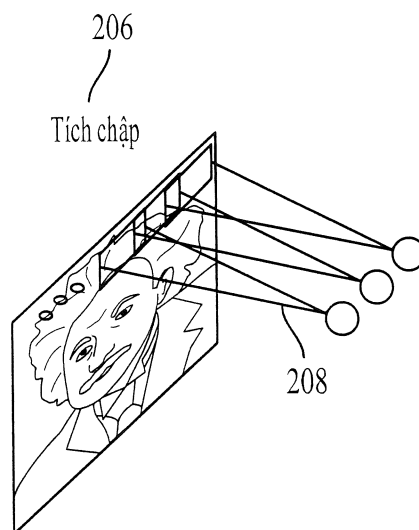
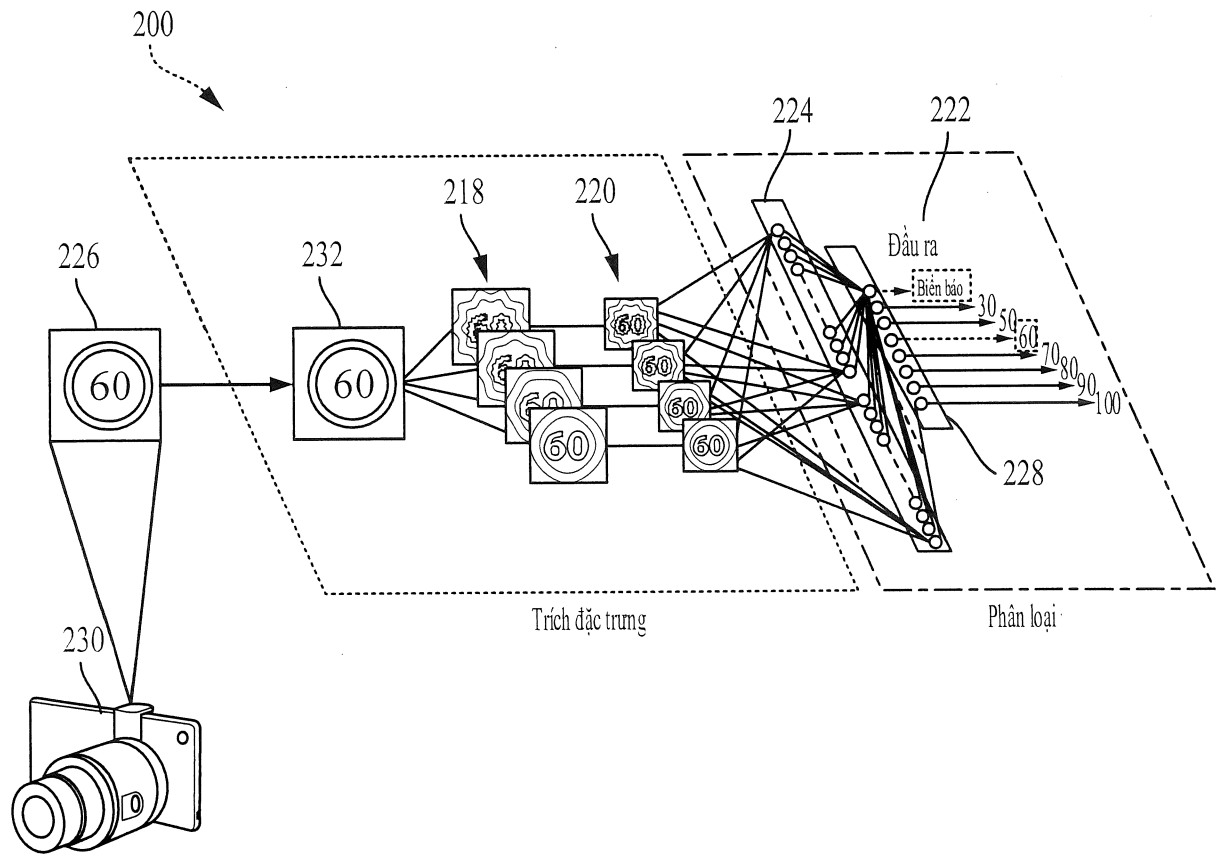
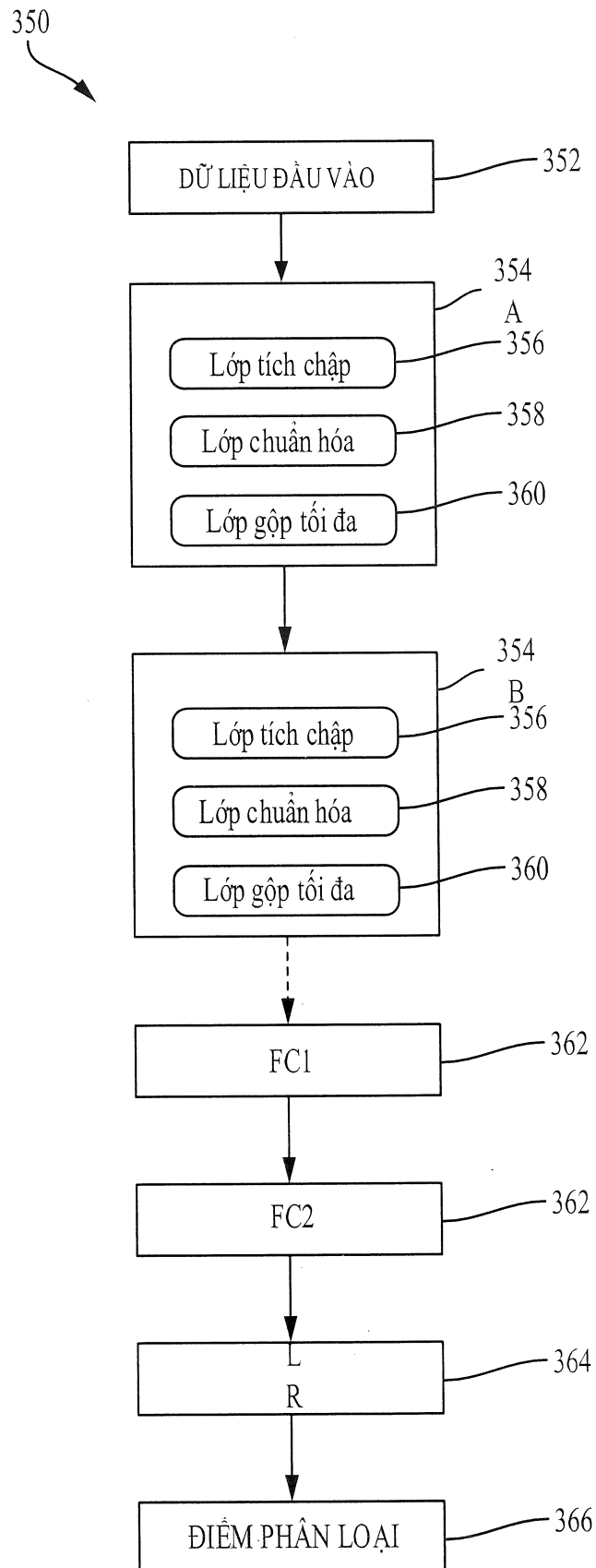


Fig.2C

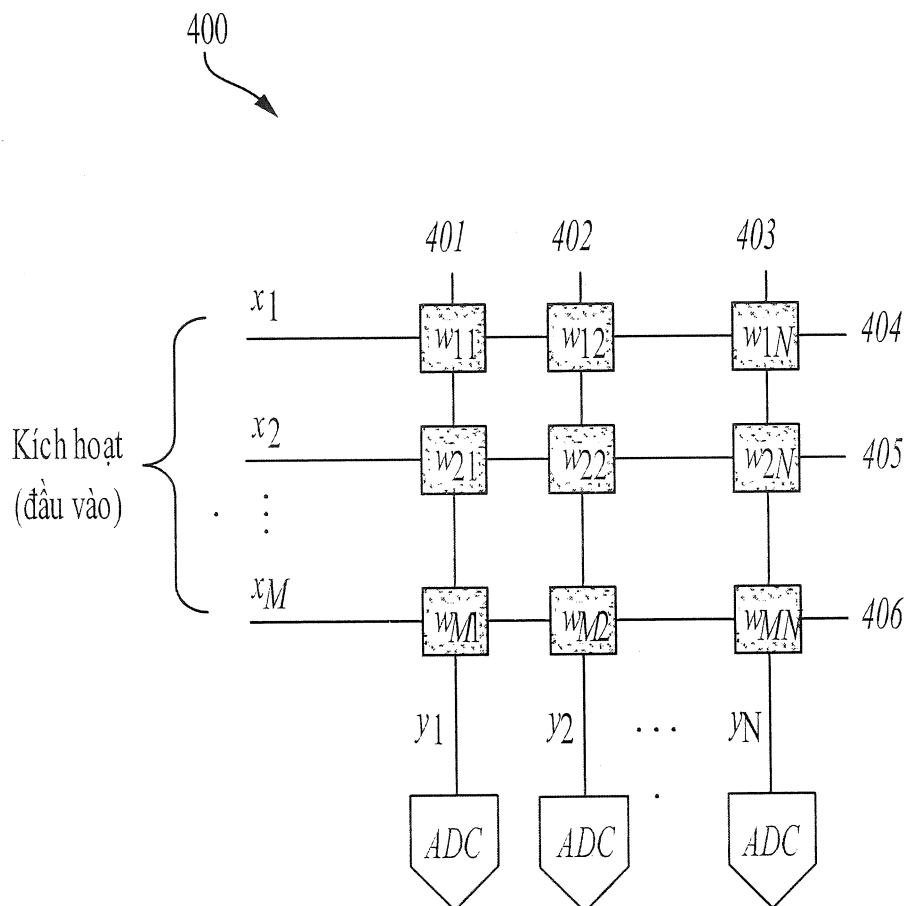
3/7

**Fig.2D**

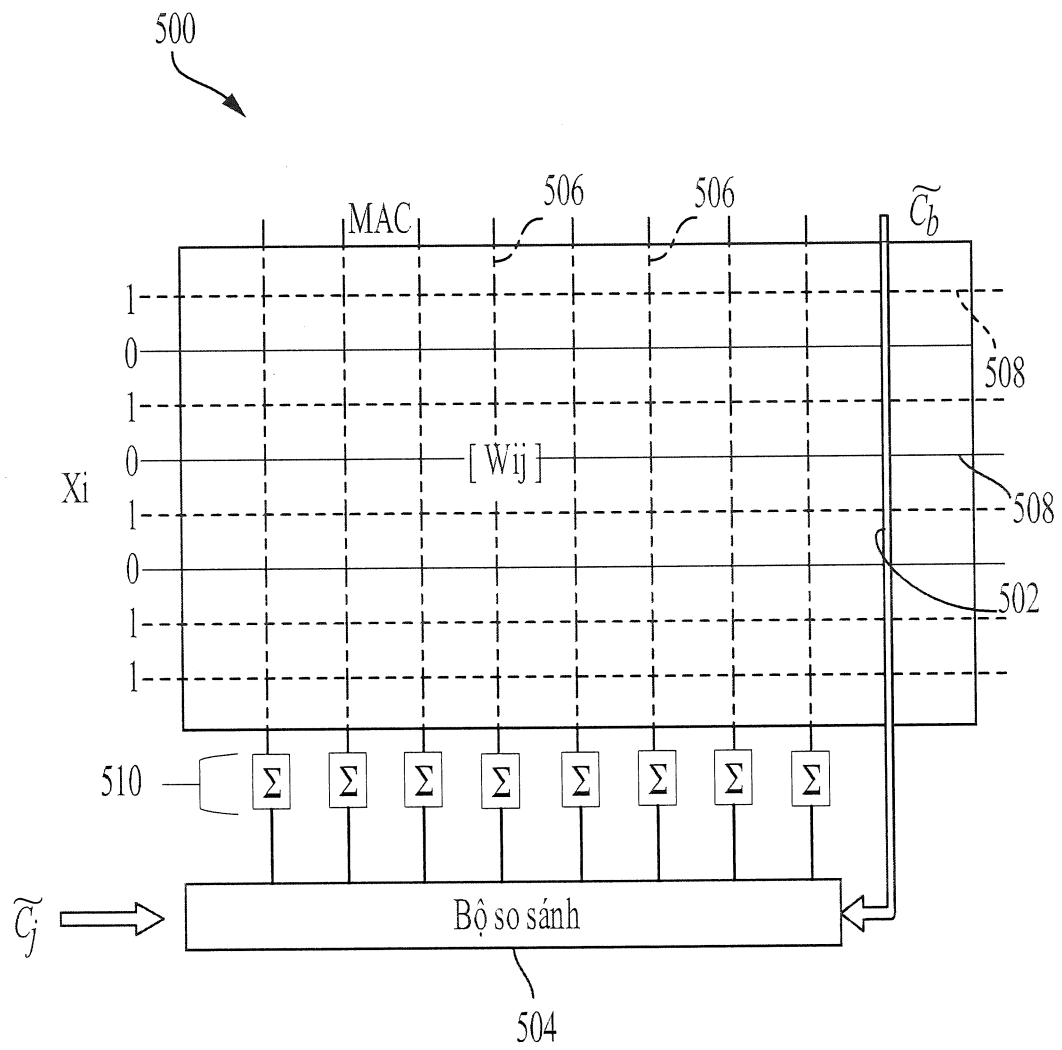
4/7

**Fig.3**

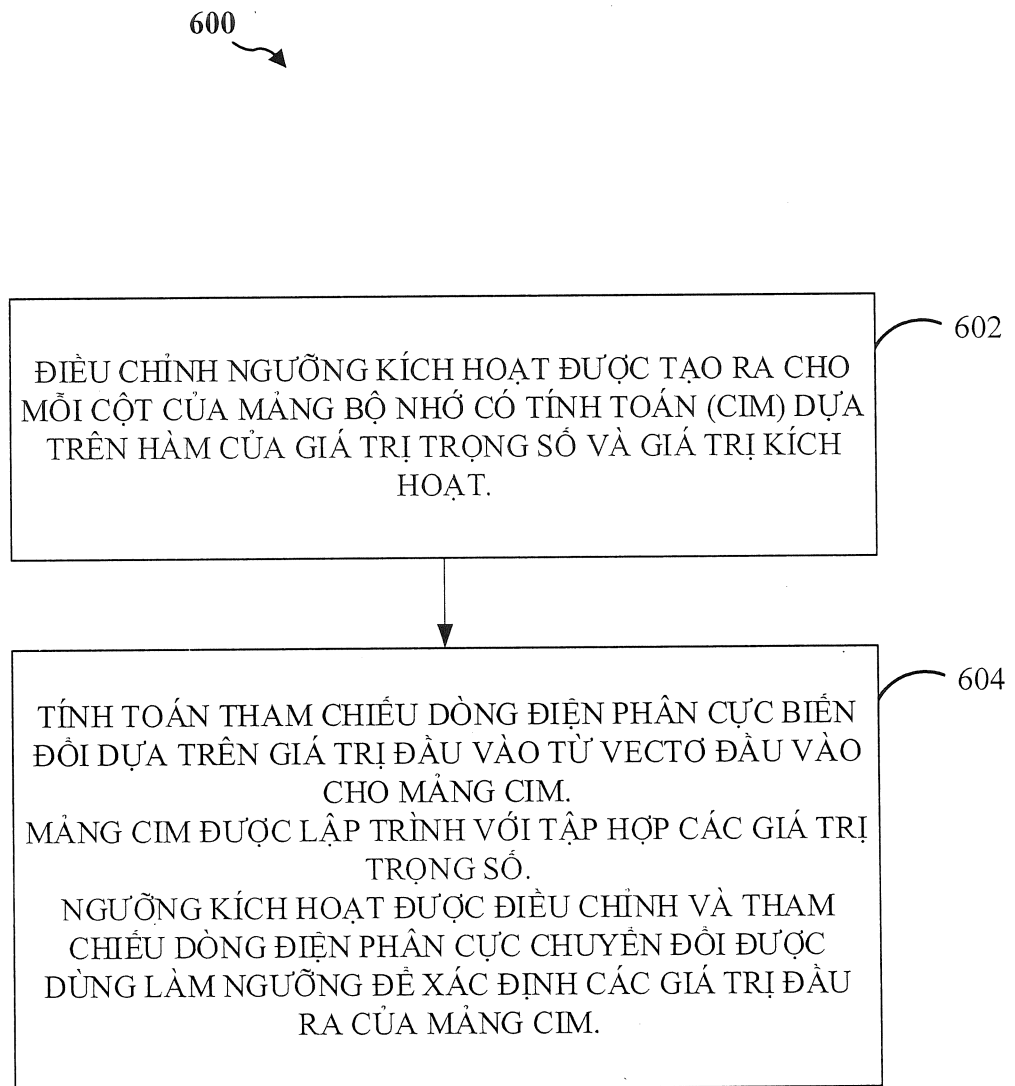
5/7

**Fig.4**

6/7

**Fig.5**

7/7

**Fig.6**