



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

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

(11)



1-0053630

(51)^{2021.01} G06N 3/00; G01C 7/00; G06F 15/76

(13) B

(21) 1-2022-05858

(22) 14/09/2022

(45) 25/11/2025 452

(43) 25/11/2022 416A

(73) TRƯỜNG ĐẠI HỌC CÔNG NGHỆ - ĐẠI HỌC QUỐC GIA HÀ NỘI (VN)

Nhà E3, 144 Xuân Thủy, phường Dịch Vọng Hậu, quận Cầu Giấy, thành phố Hà Nội

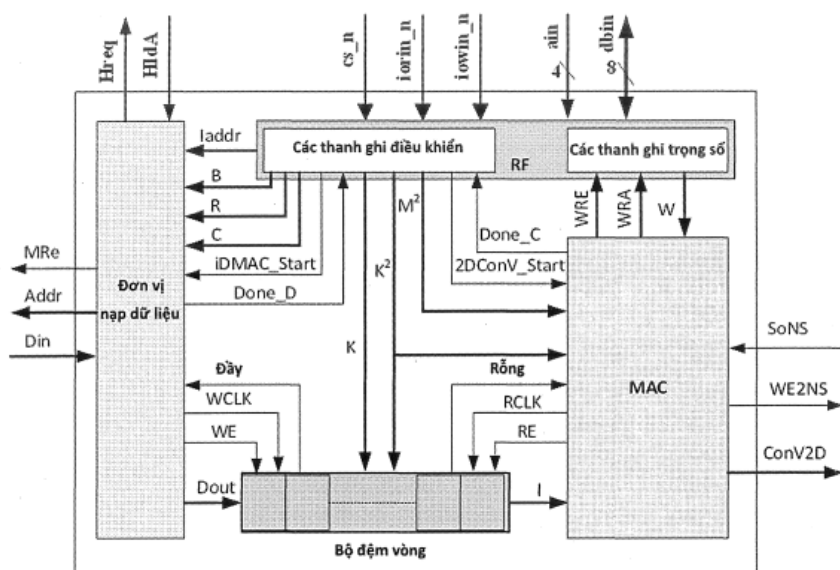
(72) Nguyễn Kiên Hùng (VN); Trần Quốc Long (VN).

(74) Công ty TNHH Sở hữu trí tuệ Bình Minh (SUNRISE IP CO.,LTD.)

(54) BỘ NHÂN CHẬP HAI CHIỀU ĐƯỢC THIẾT KẾ Ở MỨC TRUYỀN THANH GHI
(RTL) DÙNG CHO CÁC MẠNG NÓ-RON NHÂN CHẬP (CNN)

(21) 1-2022-05858

(57) Sáng chế này đề xuất giải pháp thiết kế mức vi cấu trúc RTL (Register Transfer Logic) của một bộ nhân chập 2D (hai chiều) có hiệu quả băng thông cao dùng cho các mạng nơ-ron nhân chập CNN (Convolutional Neural Network) trong học sâu (Deep learning). Bộ nhân chập 2D này có các đơn vị nạp dữ liệu (Loader), bộ đệm dữ liệu vòng (Circle Buffer) và đơn vị tính tổng tích lũy các tích MAC chuyên dụng được thiết kế dưới dạng mạch tích hợp quy mô lớn LSI (Large Scale Integration) có cấu trúc đường ống hai tầng giúp giảm độ trễ, tăng thông lượng xử lý, giảm công suất tiêu thụ. Kiến trúc được đề xuất trong sáng chế loại bỏ được việc nạp lại cả các trọng số cũng như dữ liệu ảnh đầu vào. Bộ nhân chập có khả năng cấu hình để hỗ trợ các phép tính nhân chập 2D với các kích thước bộ lọc và ma trận ảnh đầu vào khác nhau. Kiến trúc có thể giảm thời gian, công suất truy cập bộ nhớ và thời gian thực thi nhờ cơ chế sử dụng lại dữ liệu đầu vào một cách hiệu quả nhưng lại đơn giản trong thực hiện phần cứng.



Hình 4

Lĩnh vực sử dụng sáng chế

Sáng chế đề cập đến thiết kế mức vi cấu trúc RTL (Register Transfer Logic) của một bộ nhân chập trong xử lý hình ảnh, và cụ thể hơn tới bộ nhân chập 2D (hai chiều) dùng cho các mạng nơ-ron nhân chập CNN (Convolutional Neural Network) trong học sâu (Deep learning).

Tình trạng kỹ thuật của sáng chế

Gần đây, trước sự sẵn có của dữ liệu lớn (Big Data) và sự phát triển nhanh chóng của hệ thống tính toán hiệu năng cao, trí tuệ nhân tạo AI (Artificial Intelligence) đã được chú ý và đầu tư mạnh mẽ. Học sâu (Deep Learning) là kỹ thuật được sử dụng rộng rãi trong nhiều ứng dụng trí tuệ nhân tạo AI khác nhau. Trong thập kỷ qua, học sâu đã được áp dụng thành công để giải quyết nhiều vấn đề trong học thuật và trong công nghiệp, chẳng hạn như phân loại hình ảnh, phát hiện và nhận dạng đối tượng, nhận dạng âm thanh, chẩn đoán bệnh và điều khiển ô-tô tự lái.

Mạng nơ-ron nhân chập CNN (Convolutional Neural Networks) là một trong những thuật toán học sâu được sử dụng phổ biến, cung cấp độ chính xác phân loại cao cho phát hiện và nhận diện hình ảnh và đối tượng. Hơn thế, chúng có thể dễ dàng được áp dụng cho các ứng dụng mới. Tuy nhiên, CNN yêu cầu dung lượng bộ nhớ lớn để lưu trữ hàng triệu tham số trong mỗi mô hình CNN. Hơn nữa, CNN rất chuyên sâu về mặt tính toán vì chúng yêu cầu hàng tỷ phép tính trên mỗi hình ảnh. Độ phức tạp tính toán cao kết hợp với tính song song vốn có trong các mô hình này khiến chúng trở thành mục tiêu nghiên cứu để tăng tốc bằng phần cứng. Có hai thách thức chính liên quan đến việc thực hiện phần cứng cho CNN. Thách thức đầu tiên là chi phí tính toán, vì kiến trúc của chúng bao gồm nhiều lớp nhân chập, mỗi lớp lại bao gồm nhiều phép nhân. Thách thức thứ

hai là băng thông truy cập bộ nhớ lớn, trong khi tốc độ nạp dữ liệu từ bộ nhớ thấp hơn nhiều so với tốc độ xử lý. Hai thách thức này đã làm tăng nhu cầu phát triển các kiến trúc phần cứng chuyên dụng để tăng tốc tính toán của CNN trong khi vẫn giữ mức tiêu thụ điện năng ở mức hợp lý cho các ứng dụng nhúng có nguồn năng lượng hạn chế.

Phép tính nhân chập 2D là phép tính cơ bản và quan trọng nhất trong lớp nhân chập của mạng CNN. Theo thống kê, phép tính này tiêu tốn hơn 90% tổng thời gian tính toán của CNN. Hơn nữa, phép tính nhân chập cần một lượng lớn dữ liệu phải nạp từ bộ nhớ hoặc lưu trữ tới bộ nhớ. Do đó, phép tính nhân chập 2D luôn là trọng tâm trong quá trình tối ưu hóa khi thiết kế phần cứng tăng tốc CNN. Nó sẽ có lợi lớn nếu chúng ta có thể rút ngắn thời gian tính toán trong các lớp nhân chập này để đạt được khả năng tăng tốc CNN tốt nhất.

Phương pháp thực hiện quá trình nhân chập 2D có thể tổng hợp lại như sau. Phép tính nhân chập 2D của ma trận I có kích thước $N \times N$ và ma trận hạt nhân W có kích thước $K \times K$ có thể được biểu diễn bằng phương trình (1).

$$O(x, y) = \sum_{m=0}^{K-1} \sum_{n=0}^{K-1} I(x+m, y+n) \times W(m, n) \quad (1)$$

trong đó, $0 \leq x \leq N-K$ và $0 \leq y \leq N-K$.

Đặt $M = N - K + 1$ là kích thước của ma trận đầu ra.

Quá trình tính toán nhân chập được mô tả tường minh hơn bằng thuật toán trong Hình 1.

Theo công thức (1) dữ liệu ma trận đầu vào tổng cộng cần được nạp để hoàn thành tính toán nhân chập là $(K \times K) \times (N - K + 1) \times (N - K + 1)$. Tuy nhiên dữ liệu đầu vào thực sự cần thiết cho tính toán nhân chập chỉ là $N \times N$ (tương ứng kích thước của ma trận đầu vào). Nếu tính lặp lại của dữ liệu giữa các lần tính toán có thể được sử dụng một cách hiệu quả thì có thể giảm băng thông truy cập bộ nhớ đồng thời tăng tốc độ tính nhân chập.

Hình 2 chỉ ra một ví dụ minh họa cho quá trình lấy dữ liệu từ ma trận đầu vào để tính nhân chập giữa một ma trận ảnh đầu vào có kích thước 6×6 (tức $N = 6$) với một ma trận bộ lọc hạt nhân kích thước 3×3 (tức $K = 3$). Lưu ý, trong hình

này các số trong mỗi ô vuông đại diện cho thứ tự của phần tử trong ma trận. Mỗi lần một ma trận con có kích thước 3×3 (màu vàng) được trích từ ma trận đầu vào và nhân chập với ma trận bộ lọc hạt nhân theo công thức (1) để tạo ra một phần tử của ma trận đầu ra. Có 16 ma trận con như vậy. Kết quả là sau quá trình tính toán nhân chập hoàn thành, chúng ta thu được ma trận đầu ra có kích thước 4×4 như chỉ ra trong Hình 3. Lưu ý, trong hình này các số trong mỗi ô vuông đại diện cho giá trị của phần tử trong ma trận. Để ý rằng khi xử lý các ma trận con trong cùng một băng theo thứ tự từ trái qua phải, giữa hai lần tính nhân chập, dữ liệu của ma trận con chỉ khác nhau 3 phần tử. Do đó, 6 phần tử dùng chung giữa 2 lần tính nhân chập nên được lưu lại trong bộ nhân chập để không phải nạp lại từ bộ nhớ bên ngoài.

Do đó, tái sử dụng dữ liệu là một cơ chế quan trọng để nâng cao hiệu năng và giảm băng thông truy cập bộ nhớ trong việc triển khai phần cứng cho CNN.

Có một số phương pháp đã được triển khai để thực hiện phép tính nhân chập. Các phép tính nhân chập có thể được thực hiện thông qua phép nhân ma trận bằng cách chuyển đổi ma trận ảnh đầu vào thành ma trận Toeplitz như cách làm của tác giả Yiran Chen và cộng sự trong bài báo "*A survey of accelerator architectures for deep neural networks*" trên tạp chí Engineering số 6 năm 2020, trang 264-274. Tuy nhiên, phép nhân chập sử dụng phép nhân ma trận đưa vào các dữ liệu dư thừa trong ma trận Toeplitz. Do đó, nhân chập sử dụng phép nhân ma trận mất nhiều thời gian nếu được thực hiện nối tiếp hoặc yêu cầu truy cập bộ nhớ nhiều nếu hoạt động được thực hiện song song. Các phép toán nhân chập có thể được thực hiện thông qua phương pháp cửa sổ trượt trong đó một cửa sổ được trượt trên ma trận ảnh đầu vào để đọc ra ma trận con sẽ được nhân chập với ma trận bộ lọc hạt nhân. Phương pháp cửa sổ trượt truyền thống không có bất kỳ sự tái sử dụng dữ liệu nào dẫn đến việc nạp một số điểm ảnh đầu vào và trọng số nhiều lần. Các giải pháp như giải pháp của tác giả Yu-Hsin Chen và cộng sự trong bài báo "*Eyeriss v2: A flexible accelerator for emerging deep neural networks on mobile devices*" trên tạp chí IEEE Journal on Emerging and Selected Topics in Circuits and Systems số 9 năm 2019, trang 292-308, đề xuất việc tái sử dụng các

trọng số đầu vào để tránh phải nạp lại chúng từ bộ nhớ ngoài chip nhiều lần. Trong bài báo “*A unified hardware architecture for convolutions and deconvolutions in CNN*” trên kỷ yếu hội nghị IEEE International Symposium on Circuits and Systems (ISCAS) năm 2020, trang 1-15, tác giả Lin Bai và cộng sự đã trình bày một kiến trúc nhằm giảm độ trễ của mạng CNN bằng cách thực hiện tái sử dụng các dữ liệu hình ảnh đầu vào. Kiến trúc sử dụng bộ đệm dòng (Line Buffer) để căn chỉnh dữ liệu đầu vào với các hệ số của bộ lọc làm nó chỉ hỗ trợ các ma trận đầu vào và bộ lọc có kích thước cố định đã được xác định tại thời điểm thiết kế. Trong bài báo “*A fast 2-D convolution technique for deep neural networks*” trên kỷ yếu hội nghị IEEE International Symposium on Circuits and Systems (ISCAS) năm 2020, trang 1-5, tác giả Anaam Ansari và cộng sự đã thực hiện một phân tích dữ liệu dư thừa trong phép tính nhân chập 2D và sau đó triển khai một kiến trúc để thực hiện phép nhân chập 2D nhằm tránh đọc nhiều lần cùng một dữ liệu từ bộ nhớ ngoài chip sang bộ nhớ trên chip. Trong bài báo “*Hardware Acceleration of a Generalized Fast 2-D Convolution Method for Deep Neural Networks*” trên tạp chí IEEE Access số 10 năm 2022, trang 16843-16858, họ đã cải tiến công cụ xử lý có thể tính toán kết quả đầu ra của phép nhân chập 2D bằng cách tránh nạp lại bất kỳ dữ liệu đầu vào nào để tính toán bất kỳ tích từng phần nào. Tuy nhiên, kỹ thuật này dẫn đến một thiết kế phần cứng tương đối phức tạp để thực hiện.

Bản chất kỹ thuật của sáng chế

Sáng chế này đề xuất giải pháp thiết kế ở mức vi cấu trúc RTL (Register Transfer Logic) của một bộ nhân chập 2D ứng dụng cho các mạng nơ-ron nhân chập CNN (Convolutional Neural Network). Bộ nhân chập 2D được thiết kế để khai thác hiệu quả dữ liệu dùng chung giữa các lần tính toán để giảm thông lượng truy cập bộ nhớ qua đó giảm công suất tiêu thụ nhưng đồng thời cũng quan tâm đến việc giảm độ phức tạp và chi phí thực thi phần cứng, có các yêu cầu sử dụng bộ nhớ tối thiểu so với các thiết kế khác. Kiến trúc được đề xuất trong sáng chế loại bỏ được việc nạp lại cả các trọng số cũng như dữ liệu ảnh đầu vào. Kiến trúc có thể tính toán tích chập của ma trận đầu vào có kích thước bất kỳ và bộ lọc hạt

nhân có kích thước bất kỳ với độ trễ thấp và thông lượng cao so với các kỹ thuật phổ biến khác. Kiến trúc có thể giảm thời gian, công suất truy cập bộ nhớ và thời gian thực thi nhờ cơ chế sử dụng lại dữ liệu đầu vào một cách hiệu quả.

Trong thiết kế ở mức vi cấu trúc được đề xuất, bộ nhân chập 2D có các đơn vị nạp dữ liệu, bộ đệm dữ liệu và đơn vị tính tổng tích lũy các tích chập chuyên dụng được thiết kế dưới dạng mạch tích hợp quy mô lớn LSI (Large Scale Integration) có cấu trúc đường ống hai tầng giúp giảm độ trễ, tăng thông lượng xử lý và giảm công suất tiêu thụ. Cụ thể hơn, sáng chế này đưa ra một cấu trúc đơn vị nạp có khả năng truy cập trực tiếp tới bộ nhớ và đọc dữ liệu theo một mẫu hình có thể cấu hình. Đơn vị nạp cho phép dữ liệu có thể nạp liên tục và tự động từ bộ nhớ bên ngoài vào bộ nhân chập mà không cần sự can thiệp của đơn vị xử lý trung tâm. Ở một khía cạnh khác, sáng chế đưa ra một cấu trúc bộ đệm dữ liệu chuyên dụng có thể cấu hình theo các kích thước bộ lọc khác nhau trong phép nhân chập. Bộ đệm cho phép lưu lại dữ liệu lặp lại giữa các lần tính toán bằng một cơ chế đệm và quản lý dữ liệu dùng lại hiệu quả nhưng lại đơn giản cho thực thi phần cứng và yêu cầu bộ nhớ tối thiểu. Cơ chế hoạt động không đồng bộ giữa thao tác viết và đọc dữ liệu giúp đơn vị nạp và đơn vị tính tổng tích lũy các tích có thể hoạt động song song ở các tốc độ khác nhau mà không xảy ra xung đột.

Theo một phương án ưu tiên, sáng chế đề xuất bộ nhân chập hai chiều (2D) được thiết kế ở mức truyền thanh ghi (RTL - Register Transfer Logic) dùng cho các mạng nơ-ron nhân chập (CNN - Convolutional Neural Network) hiệu năng cao yêu cầu băng thông truy cập bộ nhớ và chi phí phần cứng thấp, bộ nhân chập được thiết kế dưới dạng mạch tích hợp quy mô lớn LSI (Large Scale Integration) có cấu trúc đường ống hai tầng giúp giảm độ trễ, tăng thông lượng xử lý và giảm công suất tiêu thụ, mà mỗi đơn vị dưới đây được thiết kế chi tiết ở mức RTL (Register Transfer Logic), bộ nhân chập hai chiều này bao gồm:

đơn vị nạp dữ liệu (Loader) được tạo cấu hình để có khả năng truy cập trực tiếp tới bộ nhớ và đọc dữ liệu theo một mẫu hình có thể cấu hình, được thiết kế để đọc dữ liệu hình ảnh đặc trưng từ bộ nhớ bên ngoài theo một mẫu hình quy định trước và nạp vào bộ đệm vòng (Circle Buffer), đơn vị nạp liệu này cho phép

dữ liệu có thể nạp liên tục và tự động từ bộ nhớ bên ngoài vào bộ nhân chập mà không cần sự can thiệp của đơn vị xử lý trung tâm;

bộ đệm vòng (Circle Buffer) có thể cấu hình hoạt động theo kích thước bộ lọc khác nhau trong phép nhân chập, bộ đệm này cho phép lưu lại dữ liệu lặp lại giữa các lần tính toán bằng một cơ chế đệm và quản lý dữ liệu dùng lại hiệu quả nhưng lại đơn giản cho thực thi phần cứng và yêu cầu bộ nhớ tối thiểu, bộ đệm vòng này có cơ chế hoạt động không đồng bộ với với 2 tín hiệu xung nhịp độc lập cho thao tác ghi dữ liệu và đọc dữ liệu giúp đơn vị nạp và đơn vị tính tổng tích lũy các tích có thể hoạt động song song ở các tốc độ khác nhau mà không xảy ra xung đột;

ít nhất một đơn vị tính tổng tích lũy các tích MAC, và tập các thanh ghi RF bao gồm các thanh ghi điều khiển (Control Registers) để lưu các thông tin điều khiển hoạt động của bộ nhân chập, và các thanh ghi trọng số (Weight Registers) để lưu các trọng số của bộ lọc hạt nhân để tính đồng thời nhân chập với nhiều bộ lọc hạt nhân khác nhau, bằng cách này các trọng số bộ lọc chỉ cần nạp một lần ở thời điểm bộ nhân chập bắt đầu hoạt động, để lưu trữ các trọng số của ít nhất một bộ lọc hạt nhân.

Theo một khía cạnh, bộ nhân chập này bao gồm nhiều tập thanh ghi RF và nhiều đơn vị tính toán tổng tích lũy các tích MAC để tính đồng thời nhân chập với nhiều bộ lọc hạt nhân khác nhau.

Theo một khía cạnh, bộ nhân chập này bao gồm nhiều đơn vị nạp và bộ đệm vòng để nạp dữ liệu đồng thời từ nhiều kênh hình ảnh khác nhau.

Mô tả vắn tắt các hình vẽ

Để hiểu rõ hơn về sáng chế, các hình vẽ và sơ đồ được mô tả và giải thích vắn tắt như sau:

Hình 1 là hình vẽ minh họa thuật toán tính nhân chập 2D giữa một ma trận ảnh đầu vào có kích thước $N \times N$ với một ma trận bộ lọc hạt nhân kích thước $K \times K$;

Hình 2 là hình vẽ minh họa một ví dụ minh họa cho quá trình lấy dữ liệu từ ma trận đầu vào để tính nhân chập giữa một ma trận ảnh đầu vào có kích thước 6×6 (tức $N = 6$) với một ma trận bộ lọc hạt nhân kích thước 3×3 (tức $K = 3$);

Hình 3 là hình vẽ minh họa một ví dụ về phép tính nhân chập cho một phần tử của ma trận đầu ra;

Hình 4 là hình vẽ minh họa sơ đồ khối cấu trúc tổng thể của bộ nhân chập 2D;

Hình 5 là hình vẽ minh họa giản đồ thời gian mô tả quy trình hoạt động của bộ nhân chập 2D;

Hình 6 là hình vẽ minh họa cơ chế nạp và đệm dữ liệu của bộ đệm vòng;

Hình 7 là hình vẽ minh họa máy trạng thái mô tả hoạt động của bộ đệm vòng;

Hình 8 là hình vẽ minh họa một số trạng thái điển hình của bộ đệm vòng;

Hình 9 là hình vẽ minh họa cấu trúc của một bộ đệm vòng không đồng bộ với 2 tín hiệu xung nhịp độc lập cho thao tác ghi dữ liệu và đọc dữ liệu;

Hình 10 là hình vẽ minh họa giản đồ thời gian của các tín hiệu điều khiển quá trình ghi dữ liệu từ bên ngoài vào bộ đệm vòng cũng như quá trình lấy dữ liệu từ bộ đệm vòng và viết ra bên ngoài;

Hình 11 là hình vẽ minh họa thuật toán đọc dữ liệu từ bộ nhớ chứa ảnh đầu vào và lưu vào bộ đệm vòng;

Hình 12 là hình vẽ minh họa sơ đồ khối chức năng của đơn vị nạp dữ liệu với khả năng truy cập bộ nhớ trực tiếp;

Hình 13 là hình vẽ minh họa sơ đồ máy trạng thái FSM của đơn vị nạp dữ liệu;

Hình 14 là hình vẽ minh họa thuật toán tính tổng tích lũy các tích MAC;

Hình 15 là hình vẽ minh họa sơ đồ các khối chức năng cấu thành đơn vị tính tổng tích lũy các tích MAC;

Hình 16 là hình vẽ minh họa sơ đồ máy trạng thái FSM của đơn vị tính tổng tích lũy các tích MAC.

Mô tả chi tiết sáng chế

Để hiểu rõ hơn về sáng chế, sáng chế được mô tả chi tiết thông qua các phương án cụ thể. Tuy nhiên, cần phải hiểu rằng, sáng chế không bị giới hạn bởi các phương án cụ thể này.

Trong sáng chế này, bộ nhân chập 2D được thiết kế để giảm băng thông truy cập bộ nhớ bằng cách sử dụng hiệu quả dữ liệu lặp lại giữa các lần tính nhân chập. Cấu trúc của bộ nhân chập 2D được chỉ ra trong Hình 4. Cấu trúc gồm đơn vị nạp dữ liệu (Loader), bộ đệm vòng (Circle Buffer), đơn vị tính tổng tích lũy các tích MAC, và tệp các thanh ghi điều khiển RF (Register File). Trong đó, tệp thanh ghi RF bao gồm các thanh ghi điều khiển (Control Registers) để lưu các thông tin điều khiển hoạt động của bộ nhân chập, và các thanh ghi trọng số (Weight Registers) để lưu các trọng số của bộ lọc hạt nhân. Bằng cách này các trọng số bộ lọc chỉ cần nạp một lần ở thời điểm bộ nhân chập bắt đầu hoạt động. Đơn vị nạp dữ liệu (Loader) được thiết kế để đọc hình ảnh đặc trưng từ bộ nhớ bên ngoài theo một mẫu hình quy định trước và nạp vào bộ đệm vòng. Mỗi đơn vị được thiết kế chi tiết ở mức RTL (Register Transfer Logic) để nhận được hiệu năng tốt nhất và sự tiêu thụ công suất thấp nhất.

Hình 5 trình bày giản đồ thời gian mô tả quy trình hoạt động của bộ nhân chập. Đầu tiên thông tin điều khiển bộ nhân chập cần được ghi vào các thanh ghi điều khiển (Control Registers). Ngay sau đó, đơn vị nạp dữ liệu (Loader) sẽ được kích hoạt hoạt động để nạp dần dữ liệu từ ma trận đầu vào tới bộ đệm vòng (Circle Buffer). Đồng thời trong khoảng thời gian này, ma trận trọng số cũng được nạp vào các thanh ghi trọng số (Weight Registers). Ngay sau khi các thanh ghi trọng số đã nạp xong, lúc này bộ đệm vòng cũng đã có dữ liệu, do đó đơn vị MAC được kích hoạt hoạt động.

Sau đây, mỗi đơn vị thành phần trên đây trong bộ nhân chập 2D sẽ được mô tả một cách chi tiết.

a) Bộ đệm vòng (Circle Buffer)

Hình 6 minh họa cơ chế hoạt động của bộ đệm vòng cho trường hợp kích thước bộ lọc hạt nhân $K = 3$. Bộ đệm vòng được thiết kế để lưu lại 6 phần tử dùng

chung giữa 2 lần tính nhân chấp. Các ô màu xanh biểu thị phần tử dữ liệu cần được nạp từ bộ nhớ ma trận ảnh đầu vào tới bộ đệm. Phần tử được đánh số màu đỏ biểu thị đây là phần tử đầu tiên trong ma trận con cần được tính nhân chấp. Có thể thấy cho lần tính nhân chấp đầu tiên của mỗi băng, vì chưa có dữ liệu trong bộ đệm vòng nên đơn vị nạp cần nạp 9 phần tử từ bộ nhớ ma trận ảnh đầu vào tới bộ đệm vòng. Ở các lần tính tiếp theo, do đã có sẵn 6 phần tử từ lần nạp trước nên đơn vị nạp chỉ cần nạp thêm 3 phần tử mới.

Để quản lý việc viết dữ liệu tới bộ đệm và đọc dữ liệu ra khỏi bộ đệm, bộ đệm vòng sử dụng 3 con trỏ sau:

- *Con trỏ viết WP (Write Pointer)* chỉ ra vị trí trong bộ đệm có thể viết dữ liệu vào. Sau mỗi lần viết dữ liệu con trỏ này sẽ tăng lên 1 đơn vị;
- *Con trỏ đọc RP (Read Pointer)* chỉ ra vị trí bắt đầu trong bộ đệm cho một vòng đọc. Sau mỗi một vòng đọc được hoàn thành thì con trỏ này được tăng lên K đơn vị;
- *Con trỏ vòng CP (Circle Pointer)* chỉ ra vị trí trong bộ đệm có thể đọc cho mỗi lần đọc trong một vòng đọc.

Hình 7 minh họa máy trạng thái mô tả hoạt động của bộ đệm vòng. Hình 8 minh họa một số trạng thái điển hình của bộ đệm vòng. Các trạng thái này được mô tả cụ thể như sau:

- *Trạng thái Reset:* sau khi cấp nguồn hoặc khi tín hiệu Reset ở trạng thái tích cực bộ đệm chuyển vào trạng thái Reset. Ở trạng thái này $WP = RP = CP$;
- *Trạng thái bộ đệm đầy (Full):* xảy ra khi $WP + 1 = RP$;
- *Trạng thái bộ đệm rỗng (Empty):* xảy ra khi $RP = WP$ (trạng thái sau Reset) hoặc $CP = WP$ khi CP chưa di chuyển được một vòng;
- *Trạng thái viết:* ở trạng thái này dữ liệu từ bên ngoài được ghi vào bộ nhớ của bộ đệm tại vị trí được trỏ tới bởi con trỏ WP. Sau khi ghi, WP tăng lên 1 đơn vị để trỏ tới vị trí có thể ghi dữ liệu tiếp theo;

- *Trạng thái đọc*: ở trạng thái này dữ liệu được đọc ra khỏi bộ nhớ của bộ đệm tại vị trí được trỏ tới bởi con trỏ CP. Con trỏ CP di chuyển sang vị trí kế tiếp sau mỗi lần đọc. Quá trình đọc kết thúc khi CP di chuyển được một vòng, tương đương với $R = K \times K$ vị trí. Khi kết thúc đọc, RP được gán bằng $RP + K$ và CP được gán bằng RP.

Cấu trúc của một bộ đệm vòng không đồng bộ với với 2 tín hiệu xung nhịp độc lập cho thao tác ghi dữ liệu và đọc dữ liệu được chỉ ra trong Hình 9. Chức năng của các đơn vị cấu thành bộ đệm vòng như sau:

Bộ nhớ (Dual-port memory): Bộ nhớ được dùng để lưu dữ liệu tạm thời trong bộ đệm vòng. Để hỗ trợ khả năng đọc viết đồng thời, một khối nhớ hai cổng được sử dụng. Dung lượng bộ nhớ được thiết kế bằng $C = K_{\max} \times K_{\max} + 1$. Trong đó $K_{\max}$ là kích thước bộ lọc hạt nhân lớn nhất có thể được tính toán bằng bộ nhân chập 2D.

Đơn vị điều khiển (Control Unit): Đơn vị điều khiển có chức năng tạo ra các tín hiệu điều khiển quá trình ghi dữ liệu từ bên ngoài vào bộ đệm vòng cũng như quá trình lấy dữ liệu từ bộ đệm vòng và viết ra bên ngoài như chỉ ra trong Hình 10.

Các bộ đếm địa chỉ: Có hai bộ đếm địa chỉ (Radd_Cnt và Wadd_Cnt) được sử dụng cho việc tạo ra các địa chỉ điều khiển đọc/viết dữ liệu ra/tới bộ nhớ bên trong của bộ đệm vòng. Các bộ đếm địa chỉ có thể được thiết kế theo phương pháp đếm nhị phân hoặc đếm theo mã Grey, tuy nhiên phương pháp được sử dụng phải là như nhau cho bộ đếm địa chỉ viết và đọc để đảm bảo quá trình đọc và viết diễn ra theo cùng một trật tự.

Mặc dù sáng chế được mô tả ở trên với kích thước bộ lọc $K = 3$, tuy nhiên theo một phương án khác, K tốt nhất có thể nhận giá trị từ 3 đến 11.

b) *Đơn vị nạp dữ liệu (Loader)*

Chức năng của đơn vị nạp dữ liệu (Loader) là đọc dữ liệu của ma trận ảnh đầu vào từ bộ nhớ bên ngoài và lưu vào bộ đệm vòng. Hình ảnh đầu vào kích thước $N \times N$ được chia thành B băng, mỗi băng có độ cao R bằng kích thước K của

bộ lọc hạt nhân và chiều rộng C bằng N (như minh họa trong Hình 2 với $N = 6$ và $K = 3$). Các băng dữ liệu này sẽ được đọc vào bộ nhân chập để xử lý theo thứ tự từ trên xuống dưới. Trong mỗi băng dữ liệu lại được đọc theo cột từ trái sang phải. Quá trình đọc dữ liệu từ bộ nhớ bên ngoài và nạp vào bộ đệm vòng được minh họa trong Hình 6.

Quy trình hoạt động của đơn vị nạp dữ liệu (Loader) được mô tả bằng thuật toán trong Hình 11.

Cấu trúc các khối chức năng của đơn vị nạp dữ liệu (Loader) được chỉ ra trong Hình 12. Chức năng của các đơn vị chính cấu thành đơn vị nạp dữ liệu (Loader) như sau:

Khối điều khiển CU (Control Unit) bao gồm máy trạng thái FSM, các bộ đếm (i_cnt , k_cnt và j_cnt) và các bộ so sánh thực hiện chức năng điều khiển của các vòng lặp trong thuật toán ở Hình 11. Đồng thời khối điều khiển cũng tạo ra các tín hiệu điều khiển quá trình đọc dữ liệu từ bộ nhớ chứa ảnh đầu vào và quá trình viết dữ liệu tới bộ đệm vòng. Hoạt động của máy trạng thái FSM được mô tả bằng quy trình trong Hình 13.

Khối tạo địa chỉ AGU (Address Generating Unit) bao gồm các thanh ghi địa chỉ (Addr và Paddr), các bộ hợp kênh và các bộ cộng để tính địa chỉ của ô nhớ cần được đọc từ bộ nhớ chứa hình ảnh đầu vào.

c) Đơn vị MAC

Đơn vị MAC đọc dữ liệu từ các thanh ghi trọng số và bộ đệm vòng sau đó tính tổng tích lũy các tích theo tham số điều khiển. Hoạt động của MAC được mô tả bằng chương trình giả trong Hình 14. Hình 15 chỉ ra sơ đồ khối chức năng cấu thành lên đơn vị MAC. Các khối chức năng bao gồm:

Khối xử lý (Datapath) được tạo thành từ bộ nhân, bộ cộng và thanh ghi tích lũy. Khối xử lý nhận 2 toán hạng đầu vào X , Y để tính tích $X*Y$ và cộng tích lũy vào Conv2D. Hoạt động của khối xử lý được điều khiển bởi các tín hiệu được tạo bởi khối điều khiển.

Khối điều khiển (Control Unit) tạo ra các tín hiệu để đọc dữ liệu I từ bộ đệm vòng và trọng số W từ tệp thanh ghi trọng số và điều khiển khối xử lý tính

tổng tích lũy ConV2D. Sau khi hoàn thành tính được một giá trị nhân chập, khối điều khiển tạo ra các tín hiệu điều khiển ghi giá trị này tới bộ nhớ bên ngoài. Hoạt động của khối điều khiển tuân theo sơ đồ máy trạng thái trong Hình 16.

Hiệu quả có lợi của sáng chế

Như được mô tả ở trên, chức năng nạp và đệm dữ liệu được thực hiện bởi đơn vị nạp dữ liệu (Loader) và bộ đệm vòng (Circle Buffer) là mẫu chốt để sử dụng lại dữ liệu đã nạp và do đó giảm đi băng thông truy cập bộ nhớ và công suất tiêu thụ.

Số lần dữ liệu được nạp từ ma trận đầu vào khi sử dụng bộ đệm vòng được xác định như sau:

$$(N - K + 1) \times \{K^2 + [(N - K + 1) - 1] \times K\} = (N - K + 1) \times N \times K \quad (2)$$

Như vậy, tỷ lệ tiết kiệm băng thông đạt được khi sử dụng bộ đệm vòng là:

$$r = 1 - \frac{(N-K+1) \times N \times K}{(N-K+1)^2 \times K^2} = 1 - \frac{N}{(N-K+1) \times K} \quad (3)$$

Cho trường hợp $N = 255$ và $K = 11$ thì $r = 90,5\%$.

Mẫu hình dữ liệu cần nạp từ bộ nhớ bên ngoài đơn giản dẫn đến giảm độ phức tạp và chi phí trong thực thi phần cứng. Yêu cầu sử dụng bộ nhớ tối thiểu nhờ bộ nhớ của bộ đệm vòng được sử dụng hiệu quả cho đệm dữ liệu dùng chung giữa các lần tính toán. Điều này giúp giảm thiểu diện tích và năng lượng tiêu thụ cho các thiết bị nhúng, đáp ứng tiêu chí nhỏ gọn và thời gian sử dụng thiết bị dài.

Cấu trúc đường ống hai tầng thông qua bộ đệm vòng cho phép tận dụng phần cứng để thực hiện hoạt động nạp dữ liệu và tính toán song song giúp giảm độ trễ, tăng thông lượng xử lý.

Yêu cầu bảo hộ

1. Bộ nhân chập hai chiều (2D) được thiết kế ở mức truyền thanh ghi (RTL) - Register Transfer Logic) dùng cho các mạng nơ-ron nhân chập (CNN - Convolutional Neural Network) hiệu năng cao yêu cầu băng thông truy cập bộ nhớ và chi phí phần cứng thấp, bộ nhân chập được thiết kế dưới dạng mạch tích hợp quy mô lớn LSI (Large Scale Integration) có cấu trúc đường ống hai tầng giúp giảm độ trễ, tăng thông lượng xử lý và giảm công suất tiêu thụ, mã mỗi đơn vị dưới đây được thiết kế chi tiết ở mức RTL (Register Transfer Logic), bộ nhân chập hai chiều này bao gồm:

đơn vị nạp dữ liệu (Loader) được tạo cấu hình để có khả năng truy cập trực tiếp tới bộ nhớ và đọc dữ liệu theo một mẫu hình có thể cấu hình, được thiết kế để đọc dữ liệu hình ảnh đặc trưng từ bộ nhớ bên ngoài theo một mẫu hình quy định trước và nạp vào bộ đệm vòng (Circle Buffer), đơn vị nạp dữ liệu này cho phép dữ liệu có thể nạp liên tục và tự động từ bộ nhớ bên ngoài vào bộ nhân chập mà không cần sự can thiệp của đơn vị xử lý trung tâm;

bộ đệm vòng (Circle Buffer) có thể cấu hình hoạt động theo kích thước bộ lọc khác nhau trong phép nhân chập, bộ đệm này cho phép lưu lại dữ liệu lặp lại giữa các lần tính toán bằng một cơ chế đệm và quản lý dữ liệu dùng lại hiệu quả nhưng lại đơn giản cho thực thi phần cứng và yêu cầu bộ nhớ tối thiểu, bộ đệm vòng này có cơ chế hoạt động không đồng bộ với với 2 tín hiệu xung nhịp độc lập cho thao tác ghi dữ liệu và đọc dữ liệu giúp đơn vị nạp và đơn vị tính tổng tích lũy các tích MAC có thể hoạt động song song ở các tốc độ khác nhau mà không xảy ra xung đột;

ít nhất một đơn vị tính tổng tích lũy các tích MAC, và tập các thanh ghi RF bao gồm các thanh ghi điều khiển (Control Registers) để lưu các thông tin điều khiển hoạt động của bộ nhân chập, và các thanh ghi trọng số (Weight Registers) để lưu các trọng số của bộ lọc hạt nhân để tính đồng thời nhân chập với nhiều bộ lọc hạt nhân khác nhau, bằng cách này các trọng số bộ lọc chỉ cần nạp một lần ở thời điểm bộ nhân chập bắt đầu hoạt động, để lưu trữ các trọng số của ít nhất một bộ lọc hạt nhân.

2. Bộ nhân chập 2D theo điểm 1, trong đó bộ nhân chập này bao gồm nhiều tệp thanh ghi RF và nhiều đơn vị tính toán tổng tích lũy các tích MAC để tính đồng thời nhân chập với nhiều bộ lọc hạt nhân khác nhau.

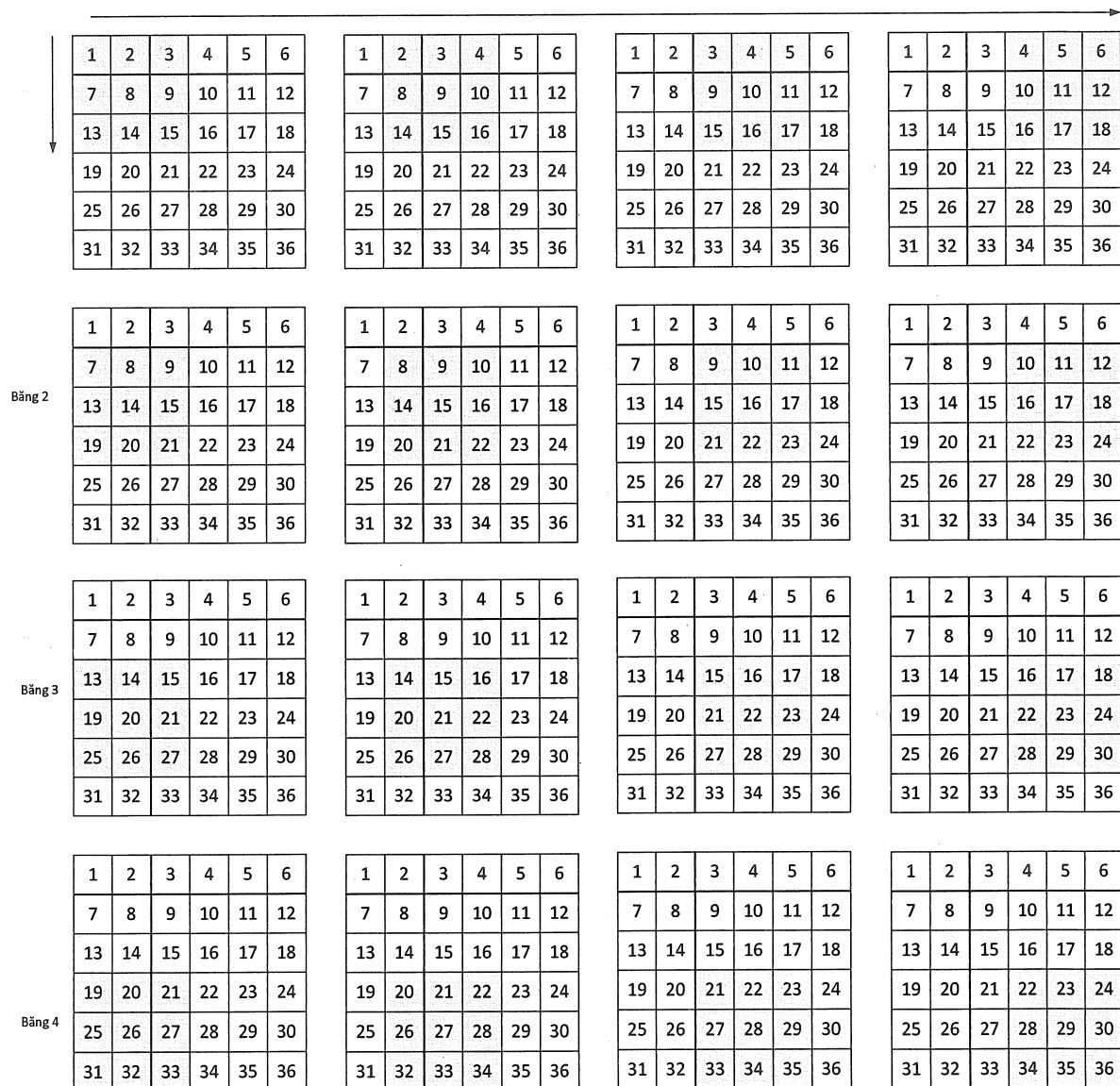
3. Bộ nhân chập 2D theo điểm 1, trong đó bộ nhân chập này bao gồm nhiều đơn vị nạp và bộ đệm vòng để nạp dữ liệu đồng thời từ nhiều kênh hình ảnh khác nhau.

```

1 // duyệt qua từng hàng của ma trận đầu ra
2 For (x = 0; x < M; x++) do
3     // duyệt qua từng cột của ma trận đầu ra
4     For (y = 0; y < M; y++) do
5         // duyệt qua các hàng của bộ lọc hạt nhân
6         For (m = 0; m < K; m++) do
7             // duyệt qua các cột của bộ lọc hạt nhân
8             For (n = 0; n < K; n++) do
9                  $O(x,y) += I(x+m, y+n) \times W(m, n)$ 
10            End for
11        End for
12    End for
13 End for

```

Hình 1



Hình 2

0	1	1 _{x1}	1 _{x0}	0 _{x1}	0
0	0	1 _{x0}	1 _{x1}	1 _{x0}	0
0	0	0 _{x1}	1 _{x0}	1 _{x1}	1
0	0	0	1	1	0
0	0	1	1	0	0
0	1	1	0	0	0

I

*

1	0	1
0	1	0
1	0	1

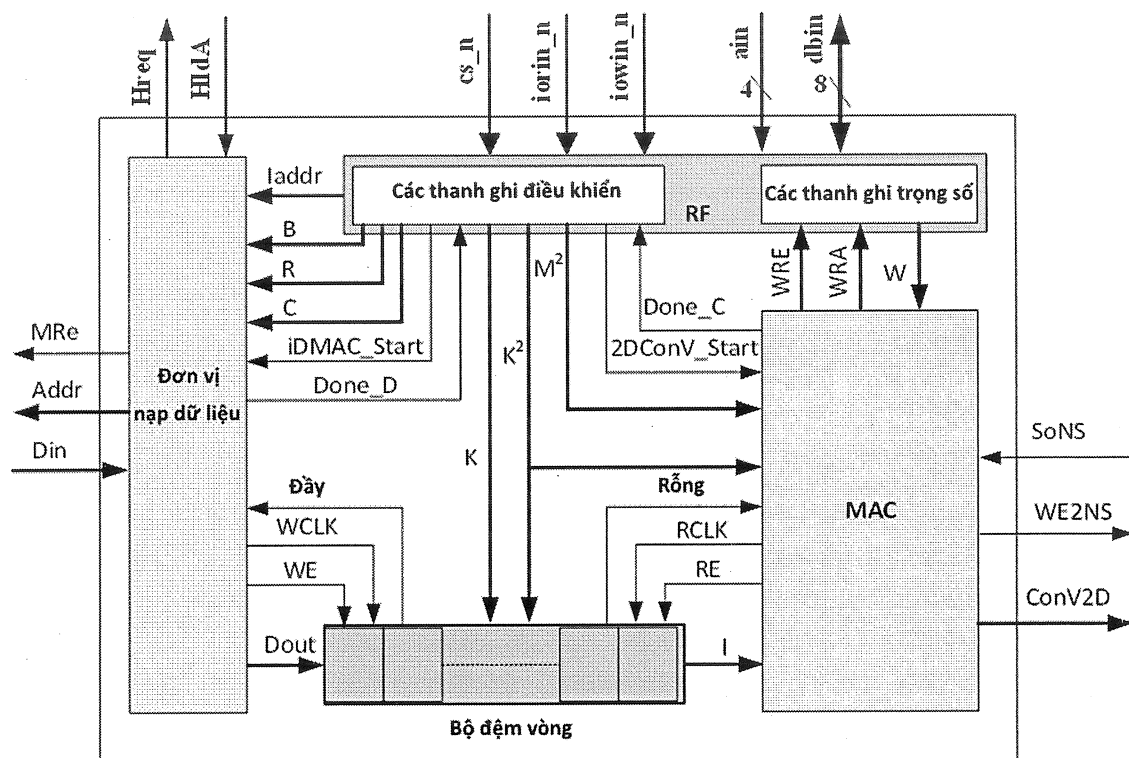
K

=

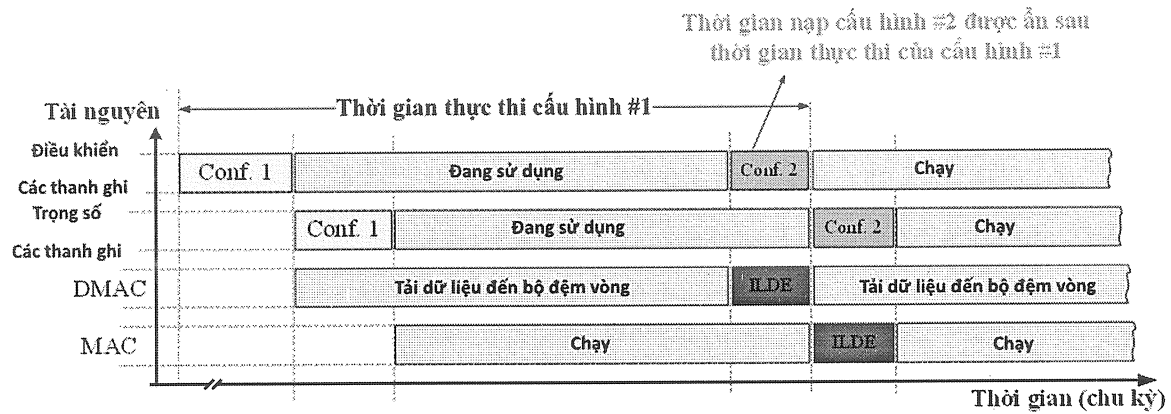
1	4	3	4
1	2	4	3
1	2	3	4
1	3	3	1

I*K

Hình 3



Hình 4



Hình 5

Lần tính #1	1	7	13	2	8	14	3	9	15
Lần tính #2	4	10	16	2	8	14	3	9	15
Lần tính #3	4	10	16	5	11	17	3	9	15
Lần tính #4	4	10	16	5	11	17	6	12	18

Bảng 1

Lần tính #1	7	13	19	8	14	20	9	15	21
Lần tính #2	10	16	22	8	14	20	9	15	21
Lần tính #3	10	16	22	11	17	23	9	15	21
Lần tính #4	10	16	22	11	17	23	12	18	24

Bảng 2

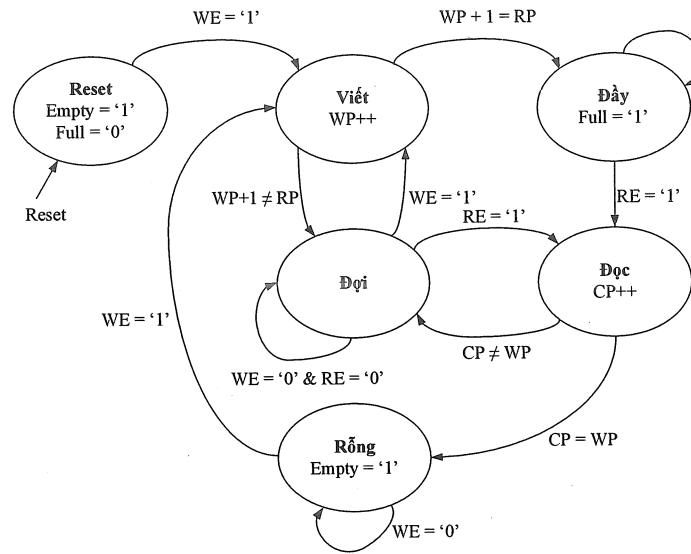
Lần tính #1	13	19	25	14	20	26	15	21	27
Lần tính #2	16	22	28	14	20	26	15	21	27
Lần tính #3	16	22	28	17	23	29	15	21	27
Lần tính #4	16	22	28	17	23	29	18	24	30

Bảng 3

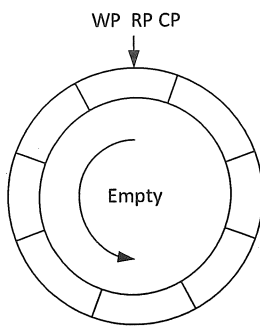
Lần tính #1	19	25	31	20	26	32	21	27	33
Lần tính #2	22	28	34	20	26	32	21	27	33
Lần tính #3	22	28	34	23	29	35	21	27	33
Lần tính #4	22	28	34	23	29	35	24	30	36

Bảng 4

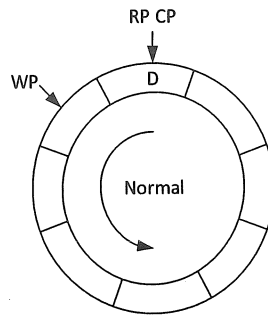
Hình 6



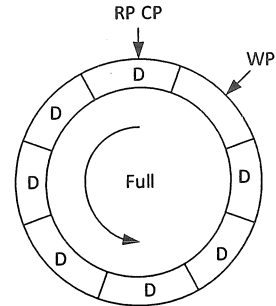
Hình 7



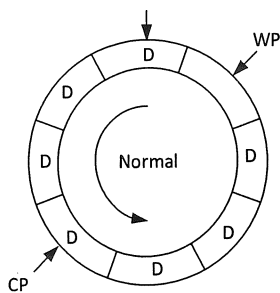
(a) Trạng thái sau Reset



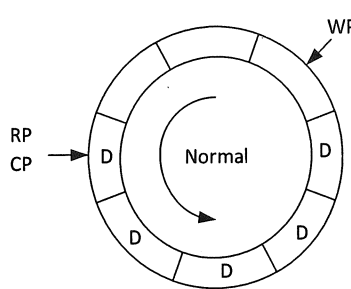
(b) Bắt đầu ghi dữ liệu



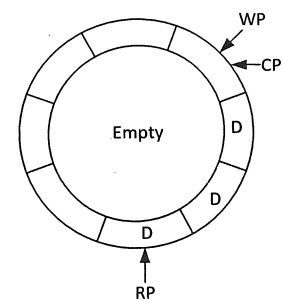
(c) Trạng thái ghi đầy



(d) Đọc dữ liệu cho lần lặp 1

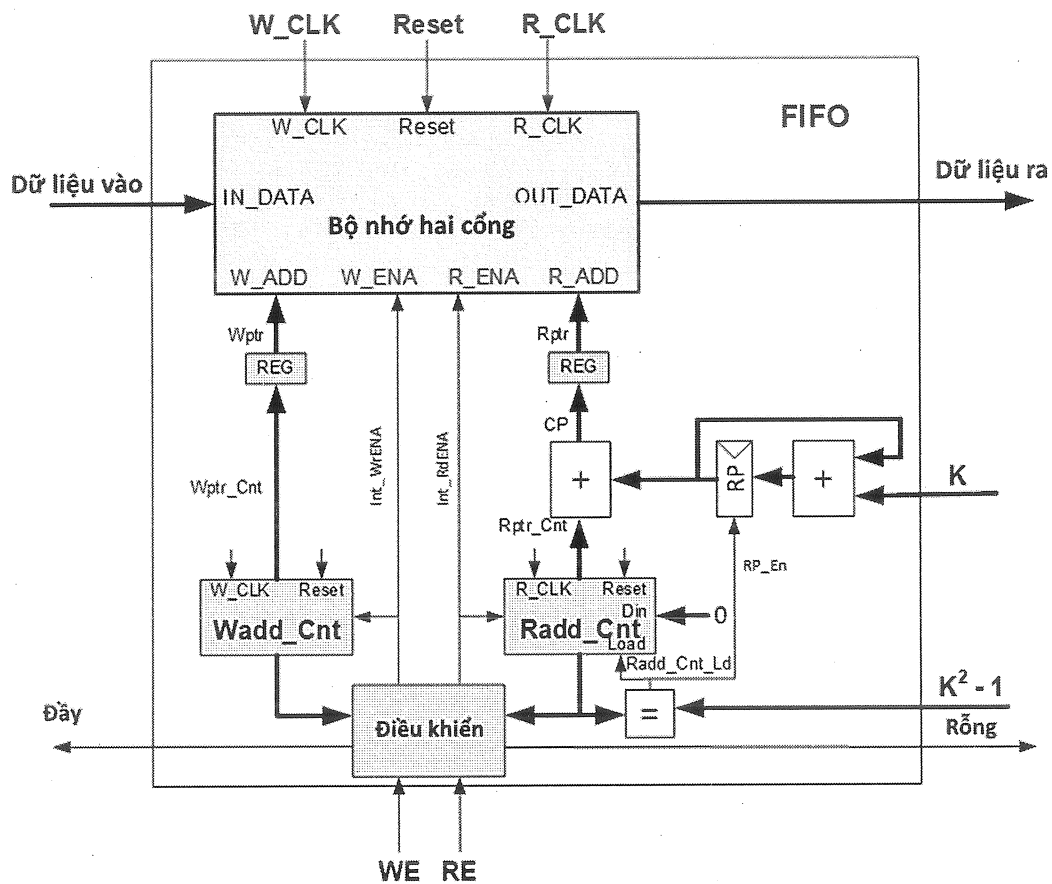


(e) Sau vòng lặp 1



(f) Trạng thái empty ở lần lặp 2

Hình 8



Hình 9

CLK			CLK		
Empty			Full		
RE			WE		
R_CLK			W_CLK		
Dataout			Datain		
Rptr_Cnt	A	A+1	MEM(wptra)		D
Rptr	A	(A+1) + RP	Wptr_Cnt	A	A+1
			Wptr	A	A+1

(a) Đọc bộ đệm

(b) Viết bộ đệm

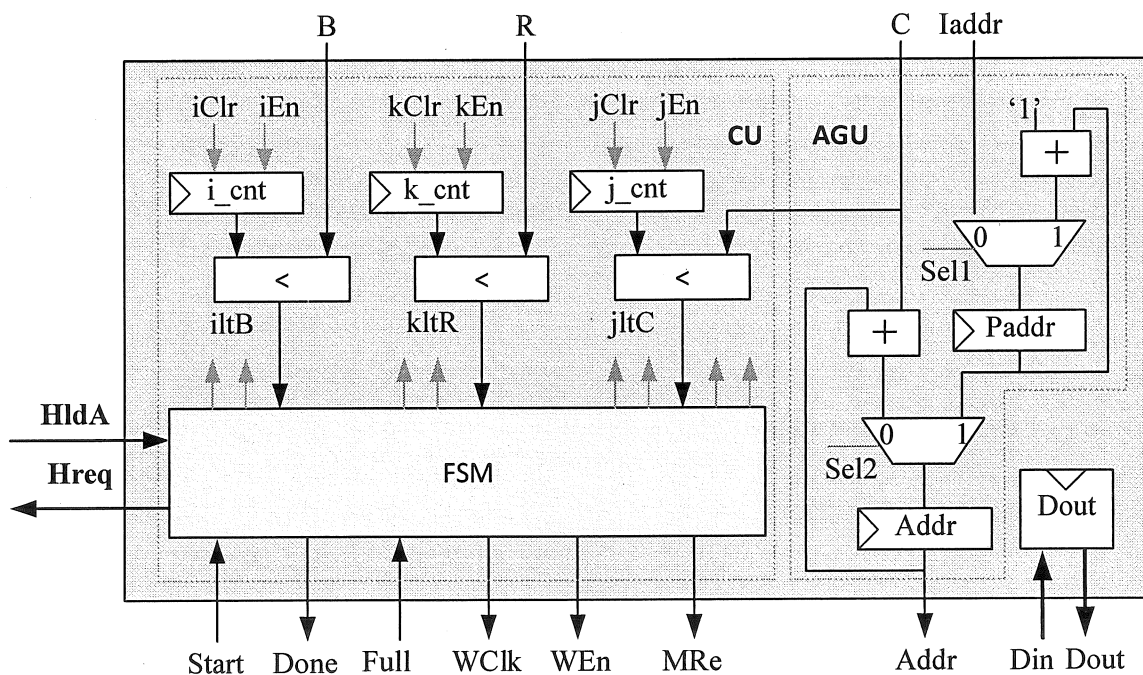
Hình 10

```

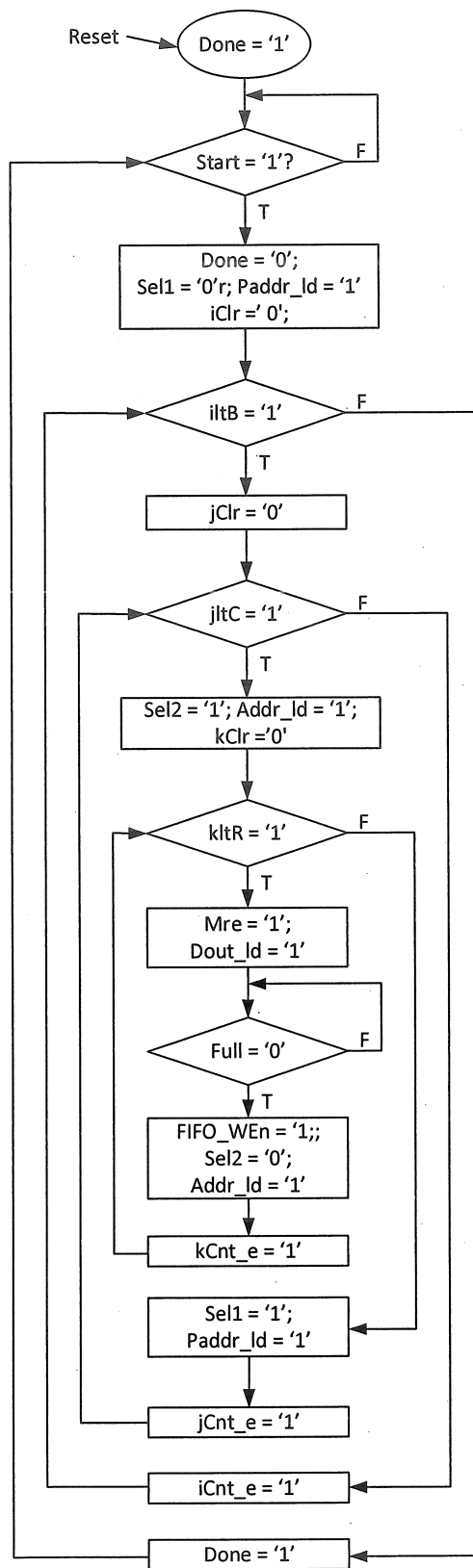
1 //Đầu vào: Iaddr – Địa chỉ bắt đầu của hình ảnh đầu vào trong bộ nhớ
2 //Đầu vào: B – Số băng (Band)
3 //Đầu vào: C – Số cột trong một băng
4 //Đầu vào: R – Số hàng trong một băng
5 //Đầu ra: Dout
6 //Biến: Paddr, Addr
7 While (Start = '1') do
8     Done = '0';
9     Paddr = Iaddr;
10    For i = 0 to B – 1 do
11        For j = 0 to C – 1 do
12            Addr = Paddr
13            For k = 0 to R – 1 do
14                Dout = M(Addr);
15                While (Full_FIFO = '1'); //đợi khi bộ đệm đầy
16                FIFO = Dout;
17                Addr = Addr + C;
18            End for k;
19            Paddr ++;
20        End for j;
21    End for i;
22    Done = '1';
23 End while;

```

Hình 11



Hình 12



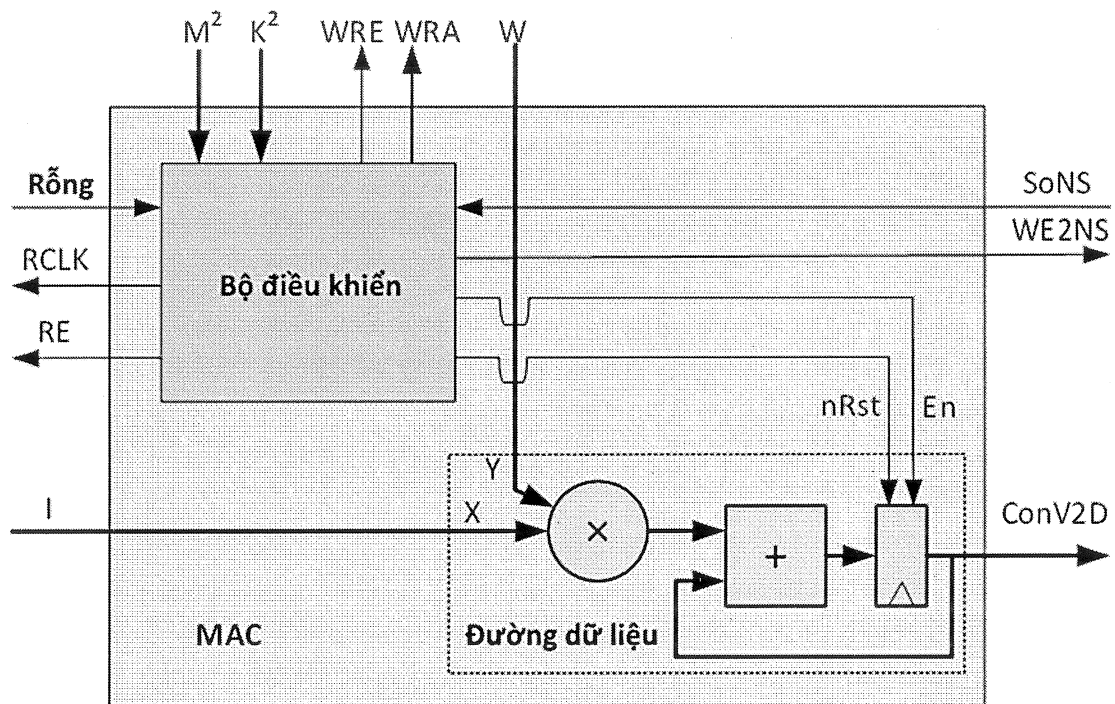
Hình 13

```

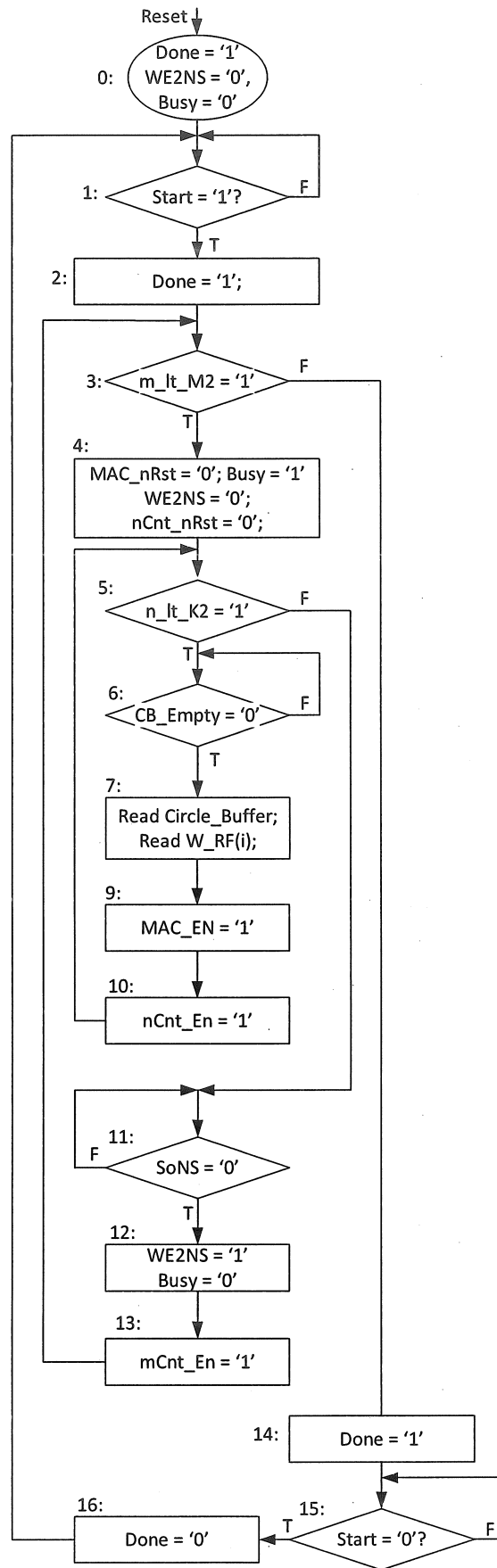
1 While (1) do
2   While (Start = '0'); //đợi đến khi Start = '1'
3   Done = '0';
4   While (m < M×M - 1) do
5     WE2NS = '0';
6     ConV2D = 0;
7     For n = 0 to K×K - 1 do
8       I = Circle_Buffer;
9       W = W_RF(i);
10      ConV2D = MAC(I, W);
11    End for n;
12    While (SoNS = '1'); //đợi khi tầng tiếp theo chưa sẵn sàng
13    WE2NS = '1';
14    m++;
15  End while;
16  Done = '1';
17 End while;

```

Hình 14



Hình 15



Hình 16