201830295

(19)中華民國智慧財產局

(12)發明說明書公開本 (11)公開編號：TW 201830295 A

(43)公開日： 中華民國 107 (2018) 年 08 月 16 日

(21)申請案號：106104512 (22)申請日： 中華民國 106 (2017) 年 02 月 10 日

(51)Int. Cl.： *G06N3/02 (2006.01)*

(71)申請人：耐能股份有限公司 (美國) KNERON, INC. (US)
美國

(72)發明人：李一雷 LI, YI-LEI (CN)；杜源 DU, YUAN (CN)；杜力 DU, LI (CN)；管延城 KUAN, YEN CHENG (TW)；劉峻誠 LIU, CHUN CHEN (TW)

(74)代理人：邱珍元

申請實體審查：有 申請專利範圍項數：20 項 圖式數：7 共 25 頁

(54)名稱

卷積神經網路的池化運算裝置及方法

POOLING OPERATION DEVICE AND METHOD FOR CONVOLUTIONAL NEURAL NETWORK

(57)摘要

一種卷積神經網路的池化運算方法，包括：讀入一池化窗內至少一排的多個新數據；將新數據進行一第一池化運算以產生至少一個排池化結果；將本次排池化結果存於一緩衝器；將緩衝器內的至少一先前排池化結果及本次排池化結果進行一第二池化運算以產生池化窗的一池化結果。

A pooling operation method for convolutional neural network, comprising: reading multiple new data in at least one column of a pooling window; performing a first pooling operation on the new data and generating at least a pooling result in the form of one column; storing the pooling result in a buffer; performing a second pooling operation on the pooling result and at least a preceding pooling result stored in the buffer and generating a pooling result of the pooling window.

指定代表圖：

TW 201830295 A

符號簡單說明：

11、12・・・池化單元

13・・・緩衝器

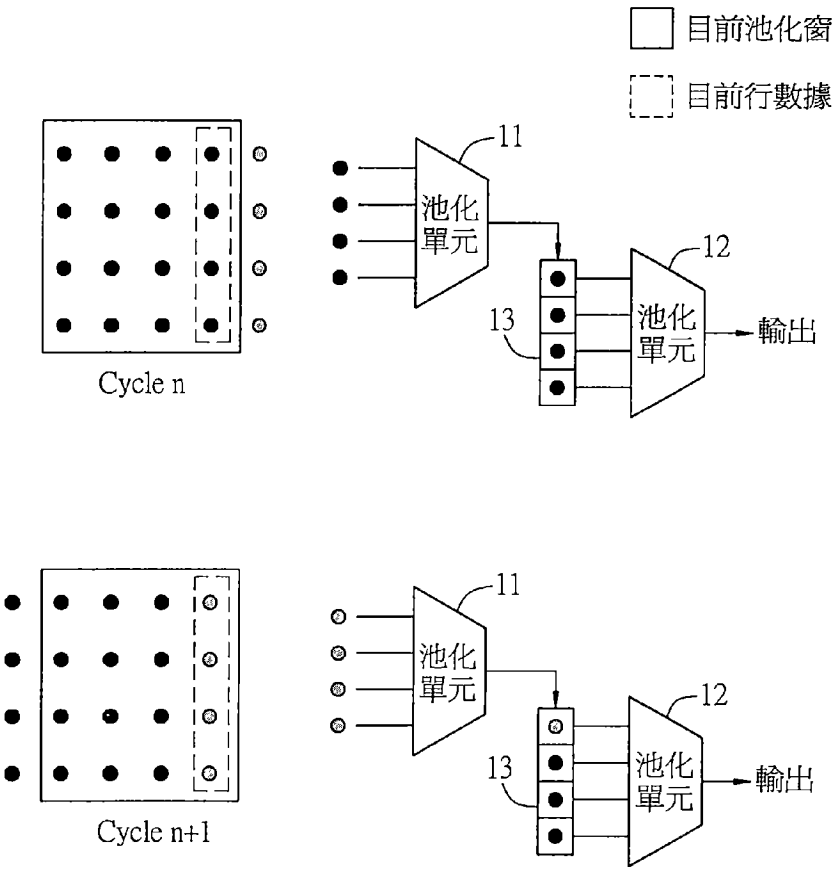

圖 2

發明專利說明書

【發明名稱】 卷積神經網路的池化運算裝置及方法

POOLING OPERATION DEVICE AND METHOD FOR CONVOLUTIONAL NEURAL NETWORK

【技術領域】

【0001】 本發明係關於一種池化運算方法，特別是關於一種執行最大池化運算的裝置及方法。

【先前技術】

【0002】 卷積神經網路（Convolutional Neural Network，CNN）是一種前饋型神經網路，其通常包含多組的卷積層（convolution layer）及池化層（pooling layer）。卷積層能夠提取輸入數據的局部特徵，而池化層可針對輸入數據某個區域上的特定特徵進行最大池化（max pooling）或平均池化（average pooling）運算，以減小參數量及神經網路中的運算，池化運算完的結果可再傳給下一層進行卷積運算，然後再次進行池化運算。然而，傳統上對多個數據進行池化運算時，需等待卷積層的所有數據均完成運算，才會輸入至池化層進行池化運算，故需佔用龐大的讀取頻寬。

【0003】 因此，如何提供一種池化運算方式，以解決讀取頻寬不足的問題，實為當前重要的課題之一。

【發明內容】

【0004】 有鑑於此，本發明之一目的為提供一種池化運算裝置及方法，可降低輸入數據的讀取頻寬，以增進池化運算的效能。

【0005】 為達上述目的，本發明提供一種卷積神經網路的池化運算方法，包括：讀入一池化窗內至少一排的多個新數據；將新數據進行一第一池化運算以產生至少一個排池化結果；將本次排池化結果存於一緩衝器；將緩衝器內的至少一先前排池化結果及本次排池化結果進行一第二池化運算以產生池化窗的一池化結果。

【0006】 在一實施例中，新數據為卷積運算的結果，且新數據在讀入

前未存在於緩衝器。

【0007】 在一實施例中，緩衝器為一先進先出緩衝器，新數據為池化窗的至少一行（column）的新數據。

【0008】 在一實施例中，緩衝器的規模大於或等於池化窗的列（row）數。

【0009】 在一實施例中，池化窗的步幅為 S，當存了 S 個排池化結果在緩衝器後，才進行第二池化運算。

【0010】 在一實施例中，第一池化運算及第二池化運算為最大池化運算。

【0011】 為達上述目的，本發明提供一種卷積神經網路的池化運算裝置，包括：一第一階池化單元、一緩衝器以及一第二階池化單元。第一階池化單元配置來讀入一池化窗內至少一排的多個新數據，並將新數據進行一第一池化運算，以產生至少一排池化結果。緩衝器耦接第一階池化單元，並配置來儲存本次排池化結果及至少一先前排池化結果。第二階池化單元配置來將緩衝器內的先前排池化結果及本次排池化結果進行一第二池化運算，以產生池化窗的一池化結果。

【0012】 在一實施例中，新數據為卷積運算的結果，且新數據在讀入前未存在於緩衝器。

【0013】 在一實施例中，緩衝器為一先進先出緩衝器，新數據為池化窗的至少一行（column）的新數據。

【0014】 在一實施例中，緩衝器的規模大於或等於池化窗的列（row）數。

【0015】 在一實施例中，池化窗的步幅為 S，當存了 S 個排池化結果在緩衝器後，才進行第二池化運算。

【0016】 在一實施例中，第一池化運算及第二池化運算為最大池化運算。

【0017】 為達上述目的，本發明提供一種卷積神經網路的池化運算方法，包括：讀入一池化窗內已備好的至少一數據，其中池化窗內涵蓋尚未備好的數據；將已備好的數據進行池化運算，以產生一部分池化結果；當

池化窗內涵蓋尚未備好的數據變為已備好的數據時，將部分池化結果及變為已備好的數據再進行池化運算，以產生一池化結果。

【0018】 在一實施例中，池化運算為最大池化運算。

【0019】 為達上述目的，本發明提供一種卷積神經網路的池化運算裝置，可進行前述的方法。

【0020】 為達上述目的，本發明提供一種池化運算的裝置，包括多個池化單元以及一緩衝器。多個池化單元各包含多個輸入及一輸出。緩衝器在本輪池化運算中從其中一個池化單元的輸出接收一池化運算結果，並將該池化運算結果於次輪池化運算中輸出到另一個池化單元的其中一個輸入。

【0021】 在一實施例中，其中一個池化單元的其中一個輸入是占位符，另一個池化單元的其中一個輸入是占位符。

【0022】 在一實施例中，池化單元的輸入係從多個卷積運算結果而來，池化單元依據一池化窗在不同位置進行運算，池化窗的步幅為 S，相鄰的池化單元具有 S 個不重複輸入。

【0023】 在一實施例中，其中一個池化單元為尾端的池化單元，另一個池化單元為啟端的池化單元。

【0024】 為達上述目的，本發明提供一種池化運算的方法，包括：在本輪池化運算中，進行多組池化運算；將其中一組池化運算的結果予以暫存；於次輪池化運算中，將暫存的池化運算的結果當作其中另一組池化運算的輸入。

【0025】 承上所述，本發明的運算裝置及運算方法中，以兩級串接的池化單元進行重疊池化（overlapping pooling）運算，第一階池化單元先將首輪運算的行數據輸出結果存放於先進先出（FIFO）緩衝器，待第一階池化單元輸出下一輪的行數據的池化運算結果後，將取代先進先出緩衝器中至少一個首輪運算的行數據輸出結果，最後再將所有儲存於先進先出緩衝器的池化運算結果一起輸入至第二階池化單元，以進行第二次池化運算，藉此得到最終的池化運算結果。此外，當池化窗中有部分列數據未完成卷積運算時，可先對已完成卷積運算的列數據進行池化運算，並將所得到的

部分池化運算結果儲存於列緩衝單元（row buffer），待池化窗中有新的列數據完成卷積運算時，再將列緩衝單元所儲存的部分池化運算結果及新的列數據一併進行池化運算，以得到最終的池化運算結果。因此，本發明的運算裝置及運算方法僅需使用有限的讀取頻寬，即可對大量數據進行池化運算，故可提升池化運算的效能。

【圖式簡單說明】

【0026】

圖 1 為卷積神經網路的示意圖。

圖 2 為依據本發明一實施例執行重疊池化運算的示意圖。

圖 3 為依據本發明一實施例的池化運算的運作示意圖。

圖 4A 及圖 4B 為依據本發明另一實施例的池化運算的運作示意圖。

圖 5 為依據本發明一實施例的一卷積運算裝置的功能方塊圖。

圖 6 為依據本發明一實施例進行最大池化運算的運作示意圖。

圖 7 為圖 6 中進行最大池化運算支援範圍的示意圖。

【實施方式】

【0027】 以下將參照相關圖式，說明依據本發明具體實施例的卷積運算裝置及方法，其中相同的元件將以相同的元件符號加以說明，所附圖式僅為說明用途，並非用於侷限本發明。

【0028】 圖 1 為卷積類神經網路的示意圖。請參閱圖 1 所示，卷積神經網路具有多個運算層，例如卷積層、池化層等，卷積層以及池化層的層數可以是多層，各層的輸出可以當作另一層或後續層的輸入，例如第 N 層卷積層的輸出是第 N 層池化層的輸入或是其他後續層的輸入，第 N 層池化層的輸出是第 N+1 層卷積層的輸入或是其他後續層的輸入，第 N 層運算層的輸出可以是第 N+1 層運算層的輸入。

【0029】 為了提升運算效能，不同層但性質接近的運算可以適當的整合在一起運算，舉例來說，池化層的池化運算是平均持化運算，除法的運算可以整合在次一層運算層中，次一層運算層例如是卷積層，也就是池化層的平均池化的除法是和次一層卷積層的卷積乘法一起運算。另外，池化

層也可以進行移位運算來替代平均計算所需的除法，並將尚未除完的部分整合在次一層運算層中一起計算，也就是池化層的平均池化的除法未能利用移位運算完整替代的部分是和次一層卷積層的卷積乘法一起運算。

【0030】 圖 2 為依據本發明一實施例執行重疊池化運算的示意圖。請參閱圖 2 所示，卷積神經網路的池化運算裝置，包括：池化單元 11、12 及緩衝器 13。緩衝器 13 可為一先進先出（FIFO）緩衝器，並耦接於池化單元 11 與池化單元 12 之間。池化單元 11 用來讀入一池化窗內至少一排的多個新數據，並將所讀入的新數據進行一第一池化運算，以產生至少一排池化結果。舉例來說，池化單元 11 可讀入池化窗內的一行新數據或多行新數據，池化單元 11 亦可讀取池化窗內的一列新數據或多列新數據，以進行第一池化運算。新數據為卷積運算的輸出結果，且通常為具有多行多列的二維數據。

【0031】 池化單元 11 可藉由移動池化窗的步幅（stride），以對卷積運算的輸出結果進行重疊池化（overlapping pooling）運算。緩衝器 13 則用來儲存池化單元 11 所輸出的池化結果，且緩衝器 13 的規模需大於或等於池化窗的列（row）數。池化單元 12 用來將緩衝器 13 內的前一輪池化結果及本輪池化結果進行一第二池化運算，以產生池化窗的最終池化結果。此外，第一池化運算及第二池化運算可為最大池化運算或平均池化運算，以下實施例將以最大池化運算為例進行說明。

【0032】 如圖 2 所示，池化單元 11 為第一階池化單元，池化單元 12 為第二階池化單元，緩衝器 13 耦接於池化單元 11 與池化單元 12 之間。池化窗的大小為 4x4，其步幅為 1，故池化窗於每個運算周期會向右移動一行。於某一個運算周期中，池化單元 11 先讀入池化窗內的一行數據，以進行第一池化運算，並將第一池化運算的輸出結果儲存於緩衝器 13 內。於下一個運算周期中，池化窗向右移動一行，池化單元 11 讀入池化窗中不與前一個運算周期重疊的一行新數據，並將新數據的池化運算結果儲存於緩衝器 13 內，以取代緩衝器 13 內的其中一個前一輪的池化結果。

【0033】 之後，池化單元 12 讀入緩衝器 13 內的本輪池化結果及前一輪的池化結果，並本輪池化結果及前一輪的池化結果進行第二池化運算，

以輸出池化窗的最終池化結果。本實施例中，池化窗的步幅為 1。當池化窗的步幅為 S 時，則需待緩衝器 13 儲存了 S 個新輸入的池化結果後，才進行第二池化運算。舉例來說，當步幅為 2 時，緩衝器 13 儲存了 1 個新輸入的池化結果後，還不會進行第二池化運算；緩衝器 13 儲存了第 2 個新輸入的池化結果後，才進行第二池化運算。也就是每隔 S 個新輸入的池化結果儲存到緩衝器 13 之後，才進行第二池化運算。如果緩衝器 13 每儲存 1 個新輸入的池化結果需要 K 的時脈（K 例如是大於等於 1 的整數），則每隔 S 乘 K 個新輸入的池化結果儲存到緩衝器 13 之後，才進行第二池化運算。

【0034】 圖 3 為依據本發明一實施例的池化運算的運作示意圖。請參閱圖 3 所示，當輸入至卷積層的數據中，僅有部分列數據完成卷積運算時，可採用兩階段進行池化運算，即先對已完成卷積運算的列數據進行池化運算，並將池化運算的結果暫存於列緩衝單元，以做為部分池化結果。待卷積層中剩餘的列數據完成卷積運算之後，再根據列緩衝單元內的部分池化結果及新的列數據計算最終的池化結果。

【0035】 舉例來說，池化單元 15 的池化窗中僅有第一及第二列數據完成卷積運算，第三列數據則尚未完成卷積運算。池化單元 15 首先對池化窗中的第一及第二列數據進行第一階段的池化運算，以輸出部分池化結果，並將部分池化結果暫存於列緩衝單元 14。當池化窗中的第三列數據也完成卷積運算時，則進行第二階段的池化運算，此時池化單元 15 將根據列緩衝單元 14 內的部分池化結果及第三列數據，以計算最終的池化結果。

【0036】 圖 4A 及圖 4B 為依據本發明另一實施例的池化運算的運作示意圖。請參閱圖 4A 所示，池化運算裝置，包括多個池化單元 81~85 以及一列緩衝單元。池化單元 81~85 分別具有多個輸入及一輸出，並對卷積運算的結果進行多次的池化運算。列緩衝單元在每一輪池化運算中從其中一個池化單元的輸出接收一池化運算結果，並將池化運算結果於次輪池化運算中輸出到另一個池化單元的其中一個輸入。在首輪池化運算中，池化運算裝置可進行多組池化運算，並將其中一組池化運算的結果暫存於列緩衝單元。於下一輪池化運算中，再將暫存的池化運算結果當作其中另一組池化運算的輸入。

【0037】 由於一次從記憶體讀取的數據數量通常是數個位元組，多個池化單元所需的輸入數據數量可能會多於一次能從記憶體讀取的數據數量，藉由列緩衝單元可以避免數據需要從記憶體重複讀取。

【0038】 舉例來說，於首輪池化運算中，池化單元 81~85 分別讀入記憶體的第 0 欄至第 3 欄的數據 A0~A7，池化單元 82 對數據 A0~A2 進行池化運算，並將池化運算的輸出結果存放於 A2 的位址上。池化單元 83 對數據 A2~A4 進行池化運算，並將池化運算的輸出結果存放於 A4 的位址上。池化單元 84 則對數據 A4~A6 進行池化運算，並將池化運算的輸出結果存放於 A6 的位址上。池化單元 85 則對數據 A6~A7 及一占位符進行池化運算，並將池化運算的輸出結果暫存於列緩衝單元。於下一輪池化運算中，請參閱圖 4B 所示，暫存於列緩衝單元的池化運算結果將做為池化單元 81 的其中一個輸入。暫存於列緩衝單元的池化運算結果和其他當輪的數據 A8~A15 一起輸入至池化單元 81~85。

【0039】 圖 5 為依據本發明一實施例的的卷積運算裝置的功能方塊圖。請參閱圖 5 所示，卷積運算裝置包括一記憶體 1、一緩衝裝置 2、一卷積運算模組 3、一交錯加總單元 4、一加總緩衝單元 5、一係數擷取控制器 6 以及一控制單元 7。卷積運算裝置可用在卷積神經網路（Convolutional Neural Network，CNN）的應用。

【0040】 記憶體 1 儲存待卷積運算的數據或數據，可例如為影像、視頻、音頻、統計、卷積神經網路其中一層的數據等等。以影像數據來說，其例如是像素（pixel）數據；以視頻數據來說，其例如是視頻視框的像素數據或是移動向量、或是視頻中的音訊；以卷積神經網路其中一層的數據來說，其通常是一個二維陣列數據，以影像數據而言，則通常是一個二維陣列的像素數據。另外，在本實施例中，係以記憶體 1 為一靜態隨機存取記憶體（static random-access memory, SRAM）為例，其除了可儲存待卷積運算的數據或數據之外，也可以儲存卷積運算完成的數據或數據，並且可以具有多層的儲存結構並分別存放待運算與運算完畢的數據，換言之，記憶體 1 可做為如卷積運算裝置內部的快取記憶體（cache memory）。

【0041】 實際應用時，全部或大部分的數據可先儲存在其他地方，例

如在另一記憶體中，另一記憶體可選擇如動態隨機存取記憶體（dynamic random access memory，DRAM）或其他種類之記憶體。當卷積運算裝置要進行卷積運算時，再全部或部分地將數據由另一記憶體載入至記憶體 1 中，然後通過緩衝裝置 2 將數據輸入至卷積運算模組 3 來進行卷積運算。若輸入的數據是串流數據，記憶體 1 隨時會寫入最新的串流數據以供卷積運算。

【0042】 緩衝裝置 2 耦接有記憶體 1、卷積運算模組 3 以及部分加總緩衝單元 5。並且，緩衝裝置 2 也與卷積運算裝置的其他元件進行耦接，例如交錯加總單元 4 以及控制單元 7。此外，對於影像數據或視頻的視框數據運算來說，處理的順序是逐行（column）同時讀取多列（row），因此在一個時序（clock）中，緩衝裝置 2 係從記憶體 1 輸入同一行不同列上的數據，對此，本實施例的緩衝裝置 2 係作為一種行緩衝（column buffer）的緩衝裝置。欲進行運算時，緩衝裝置 2 可先由記憶體 1 擷取卷積運算模組 3 所需要運算的數據，並於擷取後將該些數據調整為可順利寫入卷積運算模組 3 的數據型式。另一方面，由於緩衝裝置 2 也與加總緩衝單元 5 耦接，加總緩衝單元 5 運算完畢後之數據，也將透過緩衝裝置 2 暫存重新排序（reorder）後再傳送回記憶體 1 儲存。換言之，緩衝裝置 2 除了具有行緩衝的功能之外，其還具有類似中繼暫存數據的功能，或者說緩衝裝置 2 可做為一種具有排序功能的數據暫存器。

【0043】 值得一提的是，緩衝裝置 2 還包括一記憶體控制單元 21，當緩衝裝置 2 在進行與記憶體 1 之間的數據擷取或寫入時可經由記憶體控制單元 21 控制。另外，由於其與記憶體 1 之間具有有限的一記憶體存取寬度，或又稱為帶寬或頻寬（bandwidth），卷積運算模組 3 實際上能進行得卷積運算也與記憶體 1 的存取寬度有關。換言之，卷積運算模組 3 的運算效能會受到前述存取寬度而有所限制。因此，如果記憶體 1 的輸入有頻頸，則卷積運算的效能將受到衝擊而下降。

【0044】 卷積運算模組 3 具有多個卷積單元，各卷積單元基於一濾波器以及多個當前數據進行一卷積運算，並於卷積運算後保留部分的當前數據。緩衝裝置 2 從記憶體 1 取得多個新數據，並將新數據輸入至卷積單元，新數據不與當前數據重複。卷積運算模組 3 的卷積單元基於濾波器、保留

的當前數據以及新數據進行次輪卷積運算。交錯加總單元 4 耦接卷積運算模組 3，依據卷積運算的結果產生一特徵輸出結果。加總緩衝單元 5 耦接交錯加總單元 4 與緩衝裝置 2，暫存特徵輸出結果；其中，當指定範圍的卷積運算完成後，緩衝裝置 2 從加總緩衝單元 5 將暫存的全部數據寫入到記憶體 1。

【0045】 係數擷取控制器 6 耦接卷積運算模組 3，而控制單元 7 則耦接緩衝裝置 2。實際應用時，對於卷積運算模組 3 而言，其所需要的輸入來源除了數據本身以外，還需輸入有濾波器（filter）的係數，始得進行運算。於本實施例中所指即為 3x3 的卷積單元陣列之係數輸入。係數擷取控制器 6 可藉由直接記憶體存取 DMA（direct memory access）的方式由外部之記憶體，直接輸入濾波器係數。除了耦接卷積運算模組 3 之外，係數擷取控制器 6 還可與緩衝裝置 2 進行連接，以接受來自控制單元 7 的各種指令，使卷積運算模組 3 能夠藉由控制單元 7 控制係數擷取控制器 6，進行濾鏡係數的輸入。

【0046】 控制單元 7 可包括一指令解碼器 71 以及一數據讀取控制器 72。指令解碼器 71 係從數據讀取控制器 72 得到控制指令並將指令解碼，藉以得到目前輸入數據的大小、輸入數據的行數、輸入數據的列數、輸入數據的特徵編號以及輸入數據在記憶體 1 中的起始位址。另外，指令解碼器 71 也可從數據讀取控制器 72 得到有關濾鏡的種類資訊以及輸出特徵的編號，並輸出適當的空置訊號到緩衝裝置 2。緩衝裝置 2 則根據指令解碼後所提供的資訊來運行，也進而控制卷積運算模組 3 以及加總緩衝單元 5 的運作，例如數據從記憶體 1 輸入到緩衝裝置 2 以及卷積運算模組 3 的時序、卷積運算模組 3 的卷積運算的規模、數據從記憶體 1 到緩衝裝置 2 的讀取位址、數據從加總緩衝單元 5 到記憶體 1 的寫入位址、卷積運算模組 3 及緩衝裝置 2 所運作的卷積模式。

【0047】 另一方面，控制單元 7 則同樣可藉由直接記憶體存取 DMA 的方式由外部之記憶體提取所需的控制指令及卷積資訊，指令解碼器 71 將指令解碼之後，該些控制指令及卷積資訊由緩衝裝置 2 擷取，指令可包含移動窗的步幅大小、移動窗的位址以及欲提取特徵的影像數據行列數。

【0048】 加總緩衝單元 5 耦接交錯加總單元 4，加總緩衝單元 5 包括一部分加總區塊 51 以及一池化單元 52。部分加總區塊 51 暫存交錯加總單元 4 輸出的數據。池化單元 52 對暫存於部分加總區塊 51 的數據進行池化運算。池化運算為最大值池化或平均池化。

【0049】 舉例來說，加總緩衝單元 5 可將經由卷積運算模組 3 卷積計算結果及交錯加總單元 4 的輸出特徵結果予以暫存於部分加總區塊 51。接著，再透過池化單元 52 對暫存於部分加總區塊 51 的數據進行池化（pooling）運算，池化運算可針對輸入數據某個區域上的特定特徵，取其平均值或者取其最大值作為概要特徵提取或統計特徵輸出，此統計特徵相較於先前之特徵而言不僅具有更低的維度，還可改善運算的處理結果。

【0050】 須說明者，此處的暫存，仍係將輸入數據中的部分運算結果相加（partial sum）後才將其於部分加總區塊 51 之中暫存，因此稱其為部分加總區塊 51 與加總緩衝單元 5，或者可將其簡稱為 PSUM 單元與 PSUM BUFFER 模組。另一方面，本實施例之池化單元 52 的池化運算，係可採用前述平均池化（average pooling）的計算方式取得統計特徵輸出。待所輸入的數據全部均被卷積運算模組 3 及交錯加總單元 4 處理計算完畢後，加總緩衝單元 5 輸出最終的數據處理結果，並同樣可透過緩衝裝置 2 將結果回存至記憶體 1，或者再透過記憶體 1 輸出至其他元件。與此同時，卷積運算模組 3 與交錯加總單元 4 仍持續地進行數據特徵的取得與運算，以提高卷積運算裝置的處理效能。

【0051】 卷積運算裝置可包括多個卷積運算模組 3，卷積運算模組 3 的卷積單元以及交錯加總單元 4 係能夠選擇性地操作在一低規模卷積模式以及一高規模卷積模式。在低規模卷積模式中，交錯加總單元 4 配置來對卷積運算模組 3 中對應順序的各卷積運算的結果交錯加總以各別輸出一加總結果。在高規模卷積模式中，交錯加總單元 4 將各卷積單元的各卷積運算的結果交錯加總作為輸出。

【0052】 圖 6 為依據本發明一實施例進行最大池化運算的運作示意圖。請參閱圖 6 所示，同一行（column）的數據同時從卷積運算模組 3 或是從記憶體 1 讀出，這些數據可以是影像的畫素數據。這些數據可視最大

池化的種類（例如 2x2 或 3x3）而輸入到對應的最大池化單元。在本實施例中，加總緩衝單元 5 包括多個池化單元 52，各池化單元 52 包括一組暫存器 REG、一比較器 COMP 以及一輸出開關，比較器 COMP 具有四個輸入以及一個輸出，這組暫存器 REG 具有四個暫存器，各暫存器會將儲存的數據分別輸入到比較器 COMP，其中三個暫存器會接收及儲存從卷積運算模組 3 或是從記憶體 1 讀出的數據，另一個暫存器會接收比較器 COMP 的輸出，這個暫存器會儲存前次比較器 COMP 所輸出的最大值。對於三個輸入的數據以及前次比較的最大值，比較器 COMP 將這些數據比較以輸出最大值。也就是在先前時脈下，比較器 COMP 輸出的最大值會先暫存在暫存中以供在後續時脈下與其他新輸入的數據進行比較，其他新輸入的數據例如是次一行（next column）的數據或是後續行的數據。池化單元 52 運算所需的時脈視池化規模而定，通常約需 2-3 個時脈。池化單元 52 完成一輪的最大池化運算後，一完成訊號 EN 將啟用輸出開關以將比較出的最大值輸出，同時暫存器會重設至最小值以供後續的最大池化運算。接著，例如在次一時脈或後續時脈，輸出開關會被禁用，暫存器會接收及儲存從卷積運算模組 3 或是從記憶體 1 讀出的數據。在這種架構下，如圖 7 所示，對於一行有 9 個數據 D1~D9 的情況，使用 5 個比較器 COMP1~COMP5 就可以進行最大池化運算。

【0053】 綜上所述，本發明的運算裝置及運算方法中，以兩級串接的池化單元進行重疊池化（overlapping pooling）運算，第一階池化單元先將首輪運算的行數據輸出結果存放於先進先出（FIFO）緩衝器，待第一階池化單元輸出下一輪的行數據的池化運算結果後，將取代先進先出緩衝器中至少一個首輪運算的行數據輸出結果，最後再將所有儲存於先進先出緩衝器的池化運算結果一起輸入至第二階池化單元，以進行第二次池化運算，藉此得到最終的池化運算結果。此外，當池化窗中有部分列數據未完成卷積運算時，可先對已完成卷積運算的列數據進行池化運算，並將所得到的部分池化運算結果儲存於列緩衝單元（row buffer），待池化窗中有新的列數據完成卷積運算時，再將列緩衝單元所儲存的部分池化運算結果及新的列

數據一併進行池化運算，以得到最終的池化運算結果。因此，本發明的運算裝置及運算方法僅需使用有限的讀取頻寬，即可對大量數據進行池化運算，故可提升池化運算的效能。

【0054】 上述實施例並非用以限定本發明，任何熟悉此技藝者，在未脫離本發明之精神與範疇內，而對其進行之等效修改或變更，均應包含於後附之申請專利範圍中。

【符號說明】

【0055】

1：記憶體

2：緩衝裝置

3：卷積運算模組

4：交錯加總單元

5：加總緩衝單元

6：係數擷取控制器

7：控制單元

11、12、15、52、81~85：池化單元

13：緩衝器

14：列緩衝單元

21：記憶體控制單元

51：部分加總區塊

71：指令解碼器

72：數據讀取控制器

A0~A15、D1~D9：數據

ADD_IN、RD、WR：線路

COMP：比較器

DMA：直接記憶體存取

EN：完成訊號

REG：暫存器

201830295

發明摘要

※ 申請案號：106104512

※ 申請日： 106/02/10 ※IPC 分類： G06N 3/02 (2006.01)

【發明名稱】卷積神經網路的池化運算裝置及方法

POOLING OPERATION DEVICE AND METHOD FOR CONVOLUTIONAL NEURAL NETWORK

【中文】

一種卷積神經網路的池化運算方法，包括：讀入一池化窗內至少一排的多個新數據；將新數據進行一第一池化運算以產生至少一個排池化結果；將本次排池化結果存於一緩衝器；將緩衝器內的至少一先前排池化結果及本次排池化結果進行一第二池化運算以產生池化窗的一池化結果。

【英文】

A pooling operation method for convolutional neural network, comprising: reading multiple new data in at least one column of a pooling window; performing a first pooling operation on the new data and generating at least a pooling result in the form of one column; storing the pooling result in a buffer; performing a second pooling operation on the pooling result and at least a preceding pooling result stored in the buffer and generating a pooling result of the pooling window.

【代表圖】

【本案指定代表圖】：圖2。

【本代表圖之符號簡單說明】：

11、12：池化單元

13：緩衝器

【本案若有化學式時，請揭示最能顯示發明特徵的化學式】：

無。

申請專利範圍

1、一種卷積神經網路的池化運算的方法，包括：

讀入一池化窗內至少一排的多個新數據；

將該等新數據進行一第一池化運算以產生至少一個排池化結果；

將該排池化結果存於一緩衝器；以及

將該緩衝器內的至少一先前排池化結果及該排池化結果進行一第二池化運算以產生該池化窗的一池化結果。

2、如申請專利範圍第 1 項所述的方法，其中該等新數據為卷積運算的結果，且該等新數據在讀入前未存在於該緩衝器。

3、如申請專利範圍第 1 項所述的方法，其中該緩衝器為一先進先出緩衝器，該等新數據為該池化窗的至少一行（column）的新數據。

4、如申請專利範圍第 3 項所述的方法，其中該緩衝器的規模大於或等於該池化窗的列（row）數。

5、如申請專利範圍第 3 項所述的方法，其中該池化窗的步幅為 S，當存了 S 個該排池化結果在該緩衝器後，才進行該第二池化運算。

6、如申請專利範圍第 1 項所述的方法，其中該第一池化運算及該第二池化運算為最大池化運算。

7、一種卷積神經網路的池化運算的裝置，包括：

一第一階池化單元，配置來讀入一池化窗內至少一排的多個新數據，並將該等新數據進行一第一池化運算以產生至少一排池化結果；

一緩衝器，耦接該第一階池化單元，配置來儲存該排池化結果及至少一先前排池化結果；以及

一第二階池化單元，配置來將該緩衝器內的該先前排池化結果及該排池化結果進行一第二池化運算以產生該池化窗的一池化結果。

8、如申請專利範圍第 7 項所述的裝置，其中該等新數據為卷積運算的結果，且該等新數據在讀入前未存在於該緩衝器。

9、如申請專利範圍第 7 項所述的裝置，其中該緩衝器為一先進先出緩衝器，該等新數據為該池化窗的至少一行（column）的新數據。

10、如申請專利範圍第 9 項所述的裝置，其中該緩衝器的規模大於或等於

該池化窗的列（row）數。

11、如申請專利範圍第 9 項所述的裝置，其中該池化窗的步幅為 S，當存了 S 個該排池化結果在該緩衝器後，才進行該第二池化運算。

12、如申請專利範圍第 7 項所述的裝置，其中該第一池化運算及該第二池化運算為最大池化運算。

13、一種卷積神經網路的池化運算的方法，包括：

讀入一池化窗內已備好的至少一數據，其中該池化窗內涵蓋尚未備好的數據；

將該已備好的數據進行池化運算以產生一部分池化結果；以及

當該池化窗內涵蓋尚未備好的數據變為已備好的數據時，將該部分池化結果及變為已備好的數據再進行池化運算以產生一池化結果。

14、如申請專利範圍第 13 項所述的方法，其中該池化運算為最大池化運算。

15、一種卷積神經網路的池化運算的裝置，進行如申請專利範圍第 13 項或第 14 項所述的方法。

16、一種池化運算的裝置，包括：

多個池化單元，各包含多個輸入及一輸出；以及

一緩衝器，在本輪池化運算中從其中一個池化單元的該輸出接收一池化運算結果，並將該池化運算結果於次輪池化運算中輸出到另一個池化單元的其中一個輸入。

17、如申請專利範圍第 16 項所述的裝置，其中該其中一個池化單元的其中一個輸入是占位符，該另一個池化單元的其中一個輸入是占位符。

18、如申請專利範圍第 16 項所述的裝置，其中該等池化單元的該等輸入係從多個卷積運算結果而來，該等池化單元依據一池化窗在不同位置進行運算，該池化窗的步幅為 S，相鄰的該等池化單元具有 S 個不重複輸入。

19、如申請專利範圍第 16 項所述的裝置，其中該其中一個池化單元為尾端的池化單元，該另一個池化單元為啟端的池化單元。

20、一種池化運算的方法，包括：

在本輪池化運算中，進行多組池化運算；

將其中一組池化運算的結果予以暫存；以及

於次輪池化運算中，將暫存的池化運算的結果當作其中另一組池化運算的輸入。

圖式

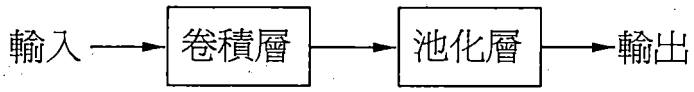

圖 1

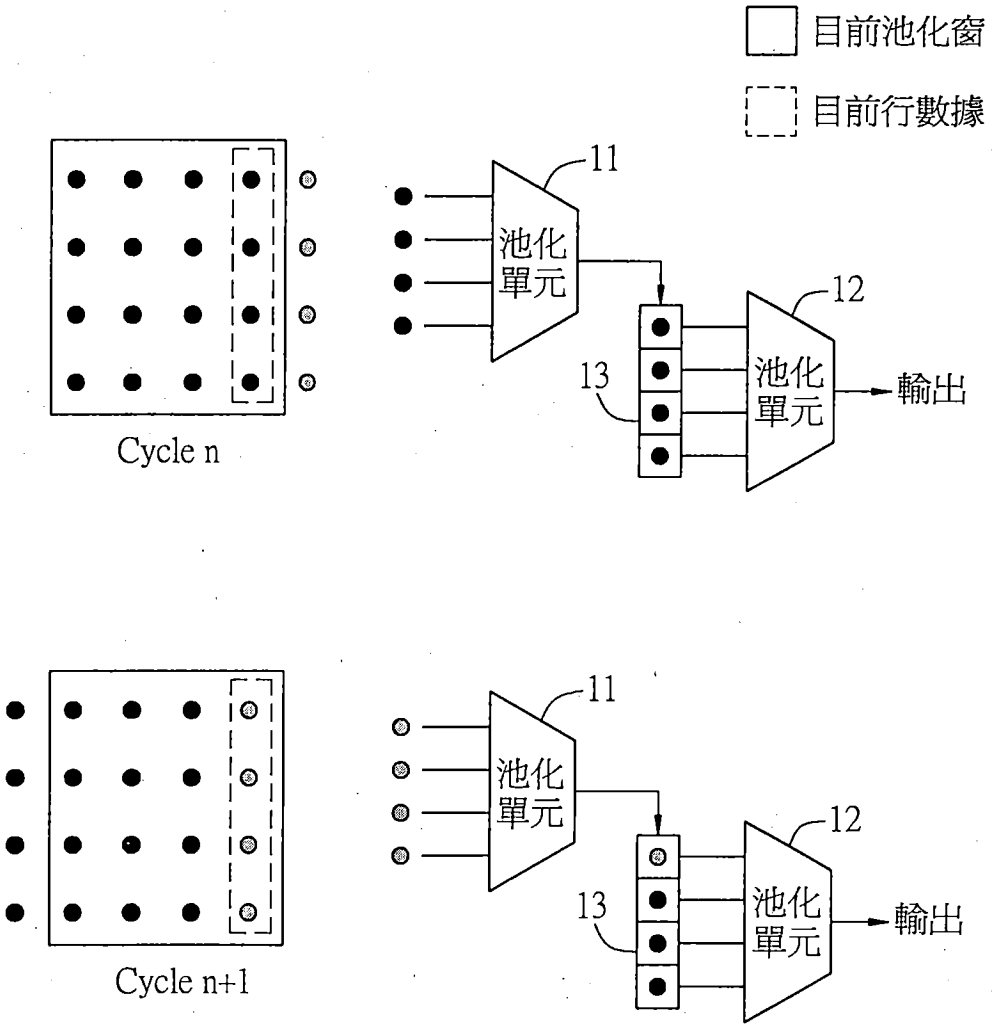
目前池化窗
目前行數據
11
池化
單元
12
池化
單元
13
輸出
Cycle n
11
池化
單元
12
池化
單元
13
輸出
Cycle n+1

圖 2

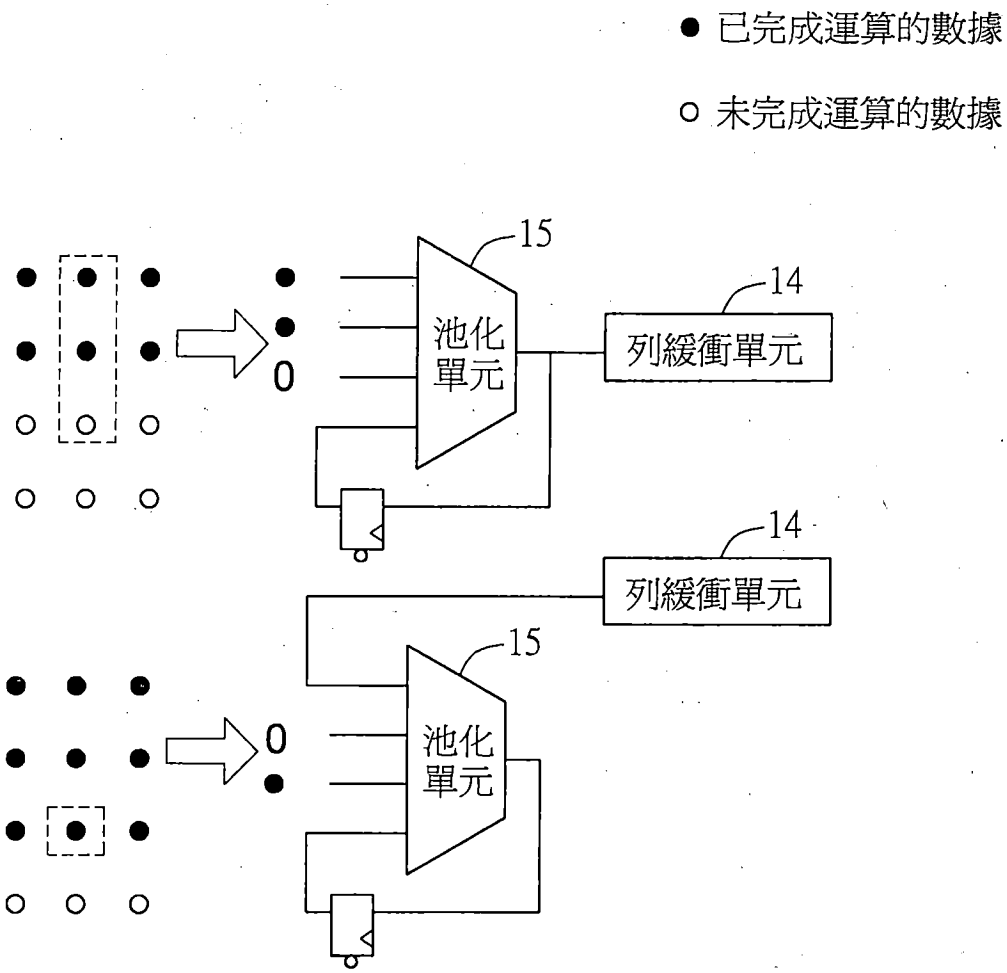

圖 3

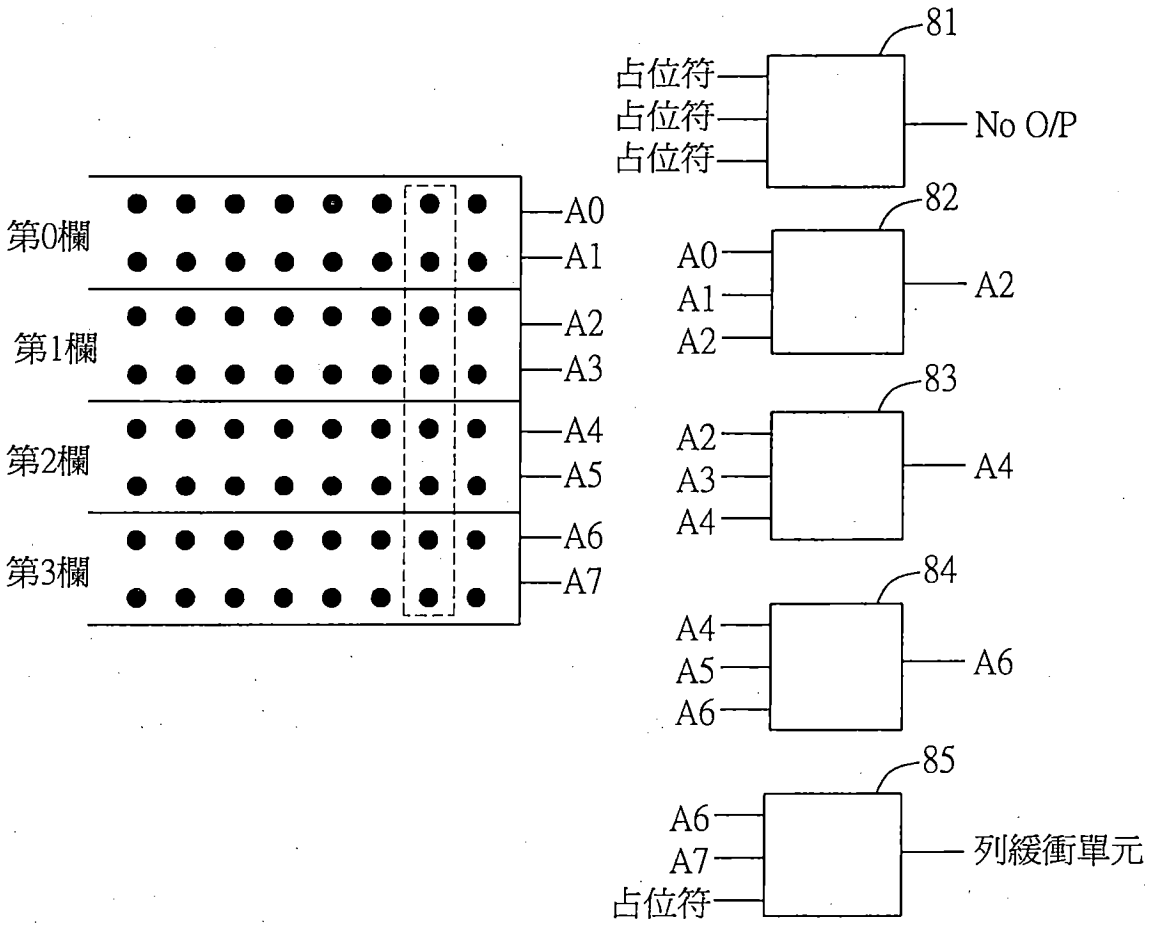
第0欄
第1欄
第2欄
第3欄
A0
A1
A2
A3
A4
A5
A6
A7
81
占位符
占位符
占位符
No O/P
82
A0
A1
A2
A2
83
A2
A3
A4
A4
84
A4
A5
A6
A6
85
A6
A7
占位符
列緩衝單元

圖 4A

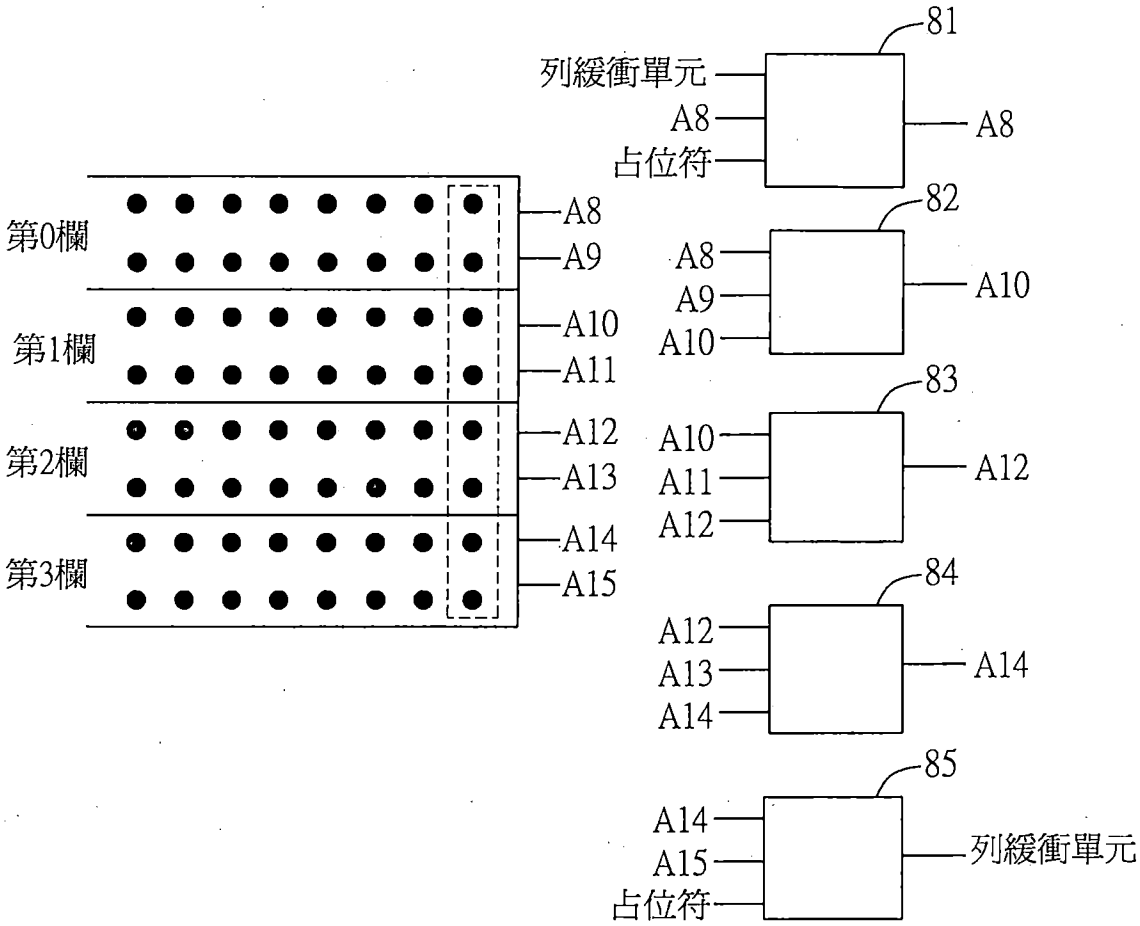
81
列緩衝單元
A8
占位符
A8
82
A8
A9
A10
A10
83
A10
A11
A12
A12
84
A12
A13
A14
A14
85
A14
A15
占位符
列緩衝單元
第0欄
第1欄
第2欄
第3欄
A8
A9
A10
A11
A12
A13
A14
A15

圖 4B

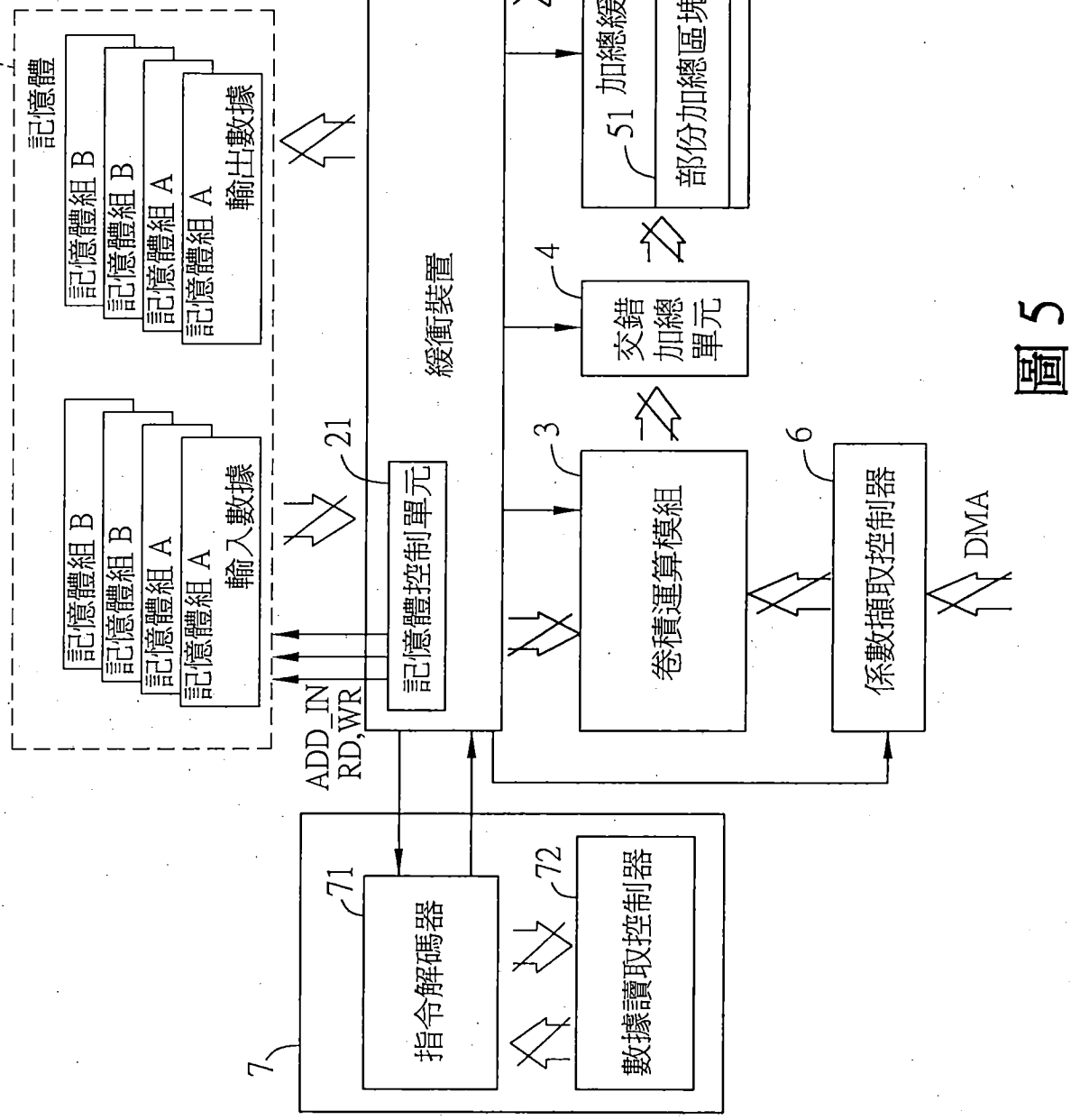
1
記憶體
記憶體組 B
記憶體組 B
記憶體組 A
記憶體組 A
輸入數據
記憶體組 B
記憶體組 B
記憶體組 A
記憶體組 A
輸出數據
2
緩衝裝置
21
記憶體控制單元
ADD_IN
RD,WR
3
卷積運算模組
4
交錯加總單元
5
51 加總緩衝單元
部份加總區塊
52
池化單元
6
係數擷取控制器
DMA
7
71
指令解碼器
72
數據讀取控制器

圖 5

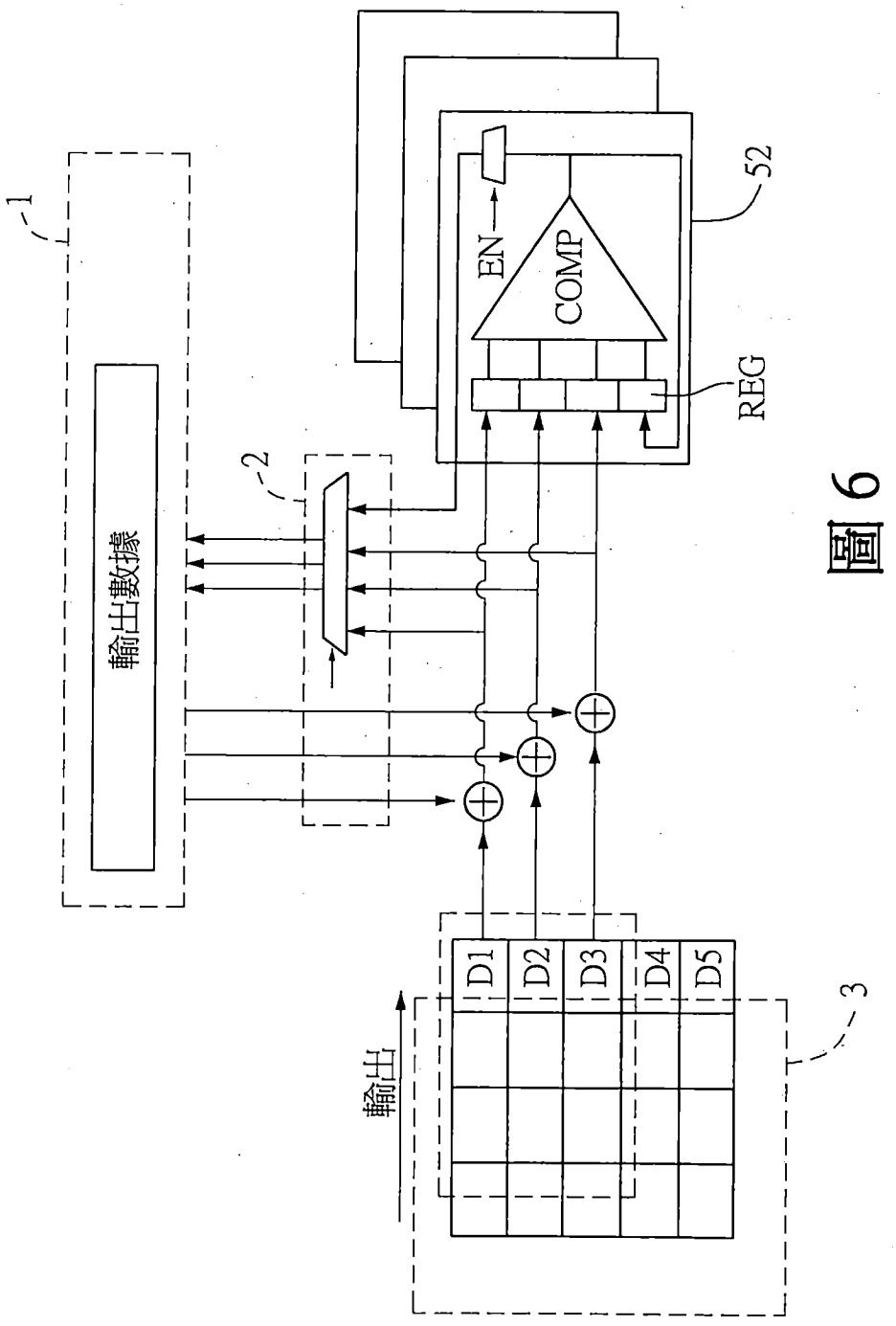

圖 6

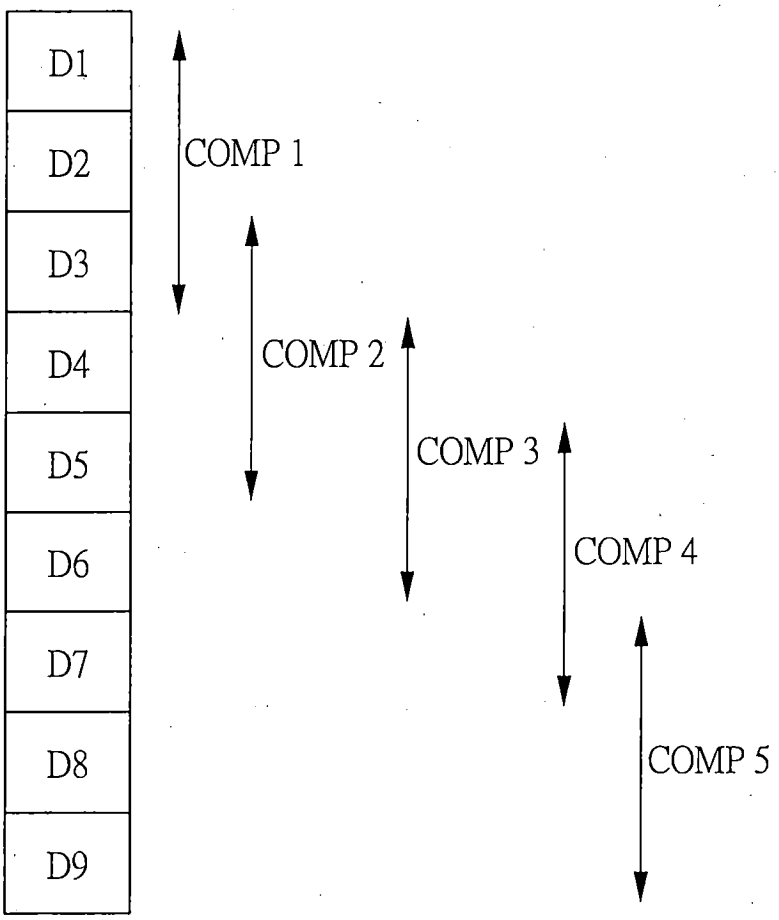

圖 7