I591490

(19)中華民國智慧財產局

(12)發明說明書公告本 **(11)證書號數：TW I591490 B**

(45)公告日： 中華民國 106 (2017) 年 07 月 11 日

(21)申請案號：105115868 (22)申請日： 中華民國 105 (2016) 年 05 月 20 日

(51)Int. Cl.： *G06F15/80 (2006.01)* *G06N3/04 (2006.01)*

(30)優先權：2015/05/21 美國 62/165,022

2015/09/03 美國 14/845,117

(71)申請人：咕果公司 (美國) GOOGLE INC. (US)

美國

(72)發明人：索森 格雷戈里 麥克 THORSON, GREGORY MICHAEL (US)；克拉克 克里斯多福 艾倫 CLARK, CHRISTOPHER AARON (US)；劉 丹 LUU, DAN (US)

(74)代理人：陳長文

(56)參考文獻：

US 8924455B1 US 2007086655A1

US 2011029471A1 US 2014180989A1

US 2014288928A1 US 2014337262A1

審查人員：施佩君

申請專利範圍項數：12 項 圖式數：9 共 39 頁

(54)名稱

類神經網路處理器中之向量運算單元

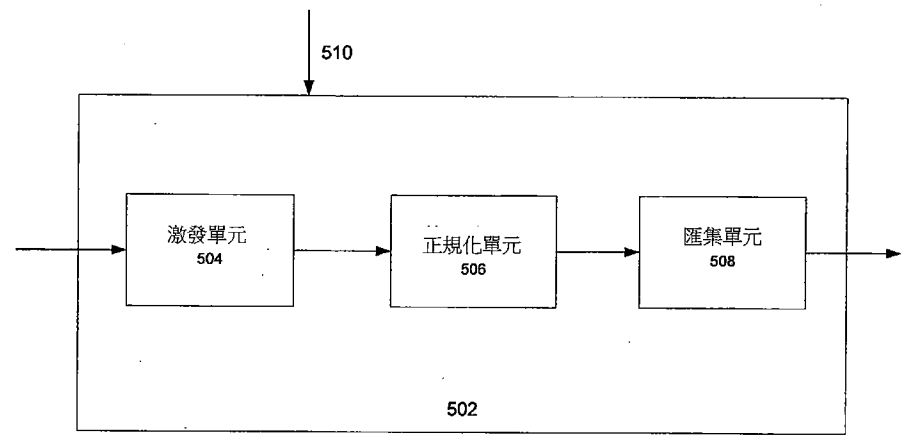
VECTOR COMPUTATION UNIT IN A NEURAL NETWORK PROCESSOR

(57)摘要

本發明揭示一種用於對包括複數個層之一類神經網路執行類神經網路計算之電路，該電路包括：激發電路，其經組態以接收累加值之一向量且經組態以將一函數應用至每一累加值以產生激發值之一向量；及正規化電路，其耦合至該激發電路且經組態以自每一激發值產生一各自正規化值。

A circuit for performing neural network computations for a neural network comprising a plurality of layers, the circuit comprising: activation circuitry configured to receive a vector of accumulated values and configured to apply a function to each accumulated value to generate a vector of activation values; and normalization circuitry coupled to the activation circuitry and configured to generate a respective normalized value from each activation value.

指定代表圖：



500

圖 5

符號簡單說明：

500・・・架構

502・・・向量計算單元

504・・・激發電路

506・・・正規化電路

508・・・匯集電路/匯集單元

510・・・控制信號

I591490

公告本

發明摘要

※ 申請案號：105115868

※ 申請日：105.5.20　　※IPC 分類：G06F 15/80 (2006.01)
3/04 (2006.01)

【發明名稱】

類神經網路處理器中之向量運算單元

VECTOR COMPUTATION UNIT IN A NEURAL NETWORK PROCESSOR

【中文】

本發明揭示一種用於對包括複數個層之一類神經網路執行類神經網路計算之電路，該電路包括：激發電路，其經組態以接收累加值之一向量且經組態以將一函數應用至每一累加值以產生激發值之一向量；及正規化電路，其耦合至該激發電路且經組態以自每一激發值產生一各自正規化值。

【英文】

A circuit for performing neural network computations for a neural network comprising a plurality of layers, the circuit comprising: activation circuitry configured to receive a vector of accumulated values and configured to apply a function to each accumulated value to generate a vector of activation values; and normalization circuitry coupled to the activation circuitry and configured to generate a respective normalized value from each activation value.

【代表圖】

【本案指定代表圖】：第（ 5 ）圖。

【本代表圖之符號簡單說明】：

500　架構

502　向量計算單元

504　激發電路

506　正規化電路

508　匯集電路/匯集單元

510　控制信號

【本案若有化學式時，請揭示最能顯示發明特徵的化學式】：

無

發明專利說明書

（本說明書格式、順序，請勿任意更動）

【發明名稱】

類神經網路處理器中之向量運算單元

VECTOR COMPUTATION UNIT IN A NEURAL NETWORK PROCESSOR

【先前技術】

本說明書係關於計算硬體中之類神經網路推斷。

類神經網路係採用一或多個層以針對一經接收輸入產生一輸出(例如，一分類)之機器學習模型。一些類神經網路除一輸出層之外亦包含一或多個隱藏層。每一隱藏層之輸出用作網路中之下一層(即，網路之下一隱藏層或輸出層)之輸入。網路之每一層根據一各自參數集合之當前值自一經接收輸入產生一輸出。

【發明內容】

一般而言，本說明書描述一種計算類神經網路推斷之專用硬體電路。

一般而言，本說明書中描述之標的物之一個發明態樣可體現在一種用於對包括複數個層之一類神經網路執行類神經網路計算之電路中，該電路包括：激發電路，其經組態以接收累加值之一向量且經組態以將一函數應用至每一累加值以產生激發輸入之一向量；及正規化電路，其耦合至該激發電路且經組態以針對每一激發值產生一各自正規化值。

實施方案可包含以下特徵之一或多者。該激發電路自該電路中之一脈動陣列接收累加值之該向量。該正規化電路包括複數個正規化暫存器行，每一正規化暫存器行包括串聯連接之複數個正規化暫存

器，每一正規化暫存器行經組態以接收一相異激發值，該正規化暫存器行中之一各自正規化單元經組態以計算一各自正規化值。每一正規化單元經組態以將該相異激發值傳遞至一相鄰正規化單元。每一正規化單元經組態以：接收一各自激發值；自該各自激發值產生一各自中間正規化值；及將該各自中間正規化值發送至一或多個鄰近正規化單元。產生該各自中間正規化值包括產生該各自激發值之一平方。每一正規化單元進一步經組態以：自一或多個鄰近正規化單元接收自激發值產生之一或多個中間正規化值；將每一中間正規化值加總以產生一索引；使用該索引以自一查找表存取一或多個值；由該一或多個值及該索引產生一按比例調整因數；及自該按比例調整因數及該各自激發值產生該各自正規化值。匯集電路(pooling circuitry)經組態以接收該等正規化值且經組態以匯集該等正規化值以產生一匯集值。該匯集電路經組態以將該複數個正規化值儲存在複數個暫存器及複數個記憶體單元中，其中該複數個暫存器及該複數個記憶體單元經串聯連接，其中每一暫存器儲存一個正規化值且每一記憶體單元儲存複數個正規化值，其中該匯集電路經組態以在每個時脈循環之後將一給定正規化值移位至一後續暫存器或記憶體單元，且其中該匯集電路經組態以自該等正規化值產生該匯集值。匯集電路經組態以接收該等激發值且經組態以匯集該等激發值以產生一匯集值。該匯集電路經組態以將該複數個激發值儲存在複數個暫存器及複數個記憶體單元中，其中該複數個暫存器及該複數個記憶體單元經串聯連接，其中每一暫存器儲存一個正規化值且每一記憶體單元儲存複數個激發值，其中該匯集電路經組態以在每個時脈循環之後將一給定激發值移位至一後續暫存器或記憶體單元，且其中該匯集電路經組態以自該等激發值產生該匯集值。

本說明書中描述之標的物之特定實施例可經實施以實現以下優點之一或多者。可在一給定時脈循環期間計算用於一類神經網路之每

一類神經網路層之多個激發值。視需要，處理器可在另一給定時脈循環期間自激發值產生多個正規化值。處理器亦可視需要自正規化值或激發值產生匯集值。處理器能夠採用每一時脈循環中之一新累加值並在每一時脈循環中產生一激發、正規化及匯集結果，藉此將計算管線化。

本發明亦提供操作一類神經網路之對應方法。

在以下隨附圖式及描述中陳述本說明書之標的物之一或多項實施例之細節。根據描述、圖式及申請專利範圍將明白標的物之其他特徵、態樣及優點。

【圖式簡單說明】

圖1係用於對一類神經網路之一給定層執行一計算之一例示性方法之一流程圖。

圖2展示一例示性類神經網路處理系統。

圖3展示包含一矩陣計算單元之一例示性架構。

圖4展示一脈動陣列內部之一胞元之一例示性架構。

圖5展示一向量計算單元之一例示性架構。

圖6展示用於正規化電路之一例示性架構。

圖7展示用於具有樣本激發值之正規化電路之另一例示性架構。

圖8展示用於正規化電路內部之一正規化單元之一例示性架構。

圖9展示用於匯集電路之一例示性架構。

各個圖式中之相同元件符號及名稱指示相同元件。

【實施方式】

具有多個層之一類神經網路可用於計算推斷。例如，給定一輸入，類神經網路可計算針對輸入之一推斷。類神經網路藉由透過類神經網路之層之各者處理輸入而計算此推斷。特定言之，類神經網路層係以一序列配置，每一層具有一各自權重集合。每一層接收一輸入並

根據層之權重集合處理輸入以產生一輸出。

因此，為自一經接收輸入計算一推斷，類神經網路接收輸入並透過該序列中之類神經網路層之各者處理該輸入以產生推斷，其中來自一個類神經網路層之輸出被提供為下一類神經網路層之輸入。至一類神經網路層之資料輸入(例如，至類神經網路之輸入或低於該序列中之層的層至一類神經網路層之輸出)可稱作至層之激發輸入。

在一些實施方案中，類神經網路之層依一有向圖予以配置。即，任何特定層可接收多個輸入、多個輸出或兩者。類神經網路之層亦可經配置使得一層之一輸出可作為一輸入發送回至一先前層。

一些類神經網路將來自一或多個類神經網路層之輸出正規化以產生用作後續類神經網路層之輸入之正規化值。將輸出正規化可有助於確保正規化值保留在預期域內用於後續類神經網路層之輸入。此可減少推理計算中的誤差。

一些類神經網路匯集來自一或多個類神經網路層之輸出以產生用作後續類神經網路層之輸入之匯集值。在一些實施方案中，類神經網路藉由判定一輸出群組之最大值或平均值及使用最大值或平均值作為該群組之匯集輸出而匯集該輸出群組。匯集輸出可維持某一空間不變性，因此以各種組態配置之輸出可經處理以具有相同推理。匯集輸出亦可減小後續類神經網路層處接收之輸入之維度，同時維持輸出在匯集之前的所需特性，此可改良效率而不顯著地損及由類神經網路產生的推理品質。

本說明書描述視需要對一或多個類神經網路層之輸出執行正規化、匯集或兩者之專用硬體電路。

圖1係用於使用一專用硬體電路對一類神經網路之一給定層執行一計算之一例示性程序100之一流程圖。為了方便起見，將關於具有執行方法100之一或多個電路之一系統描述方法100。可對類神經網路

之每一層執行方法100以自一經接收輸入計算一推理。

系統接收權重輸入集合(步驟102)及激發輸入集合(步驟104)用於給定層。可分別自專用硬體電路之動態記憶體及一統一緩衝器(unified buffer)接收權重輸入集合及激發輸入集合。在一些實施方案中，可自統一緩衝器接收權重輸入集合及激發輸入集合兩者。

系統使用專用硬體電路之一矩陣乘法單元自權重輸入及激發輸入產生累加值(步驟106)。在一些實施方案中，累加值係權重輸入集合與激發輸入集合之點積。即，對於一個權重集合(其係層中之所有權重之一子集)，系統可將每一權重輸入與每一激發輸入相乘並將乘積加總在一起以形成一累加值。系統接著可計算其他權重集合與其他激發輸入集合之點積。

系統可使用專用硬體電路之一向量計算單元自累加值產生一層輸出(步驟108)。在一些實施方案中，向量計算單元將一激發函數應用至累加值，此將在下文參考圖5進一步描述。層之輸出可經儲存在統一緩衝器中以用作至類神經網路中之一後續層之一輸入或可用於判定推理。當一經接收輸入已透過類神經網路之每一層處理以產生經接收輸入之推理時，系統完成處理類神經網路。

圖2展示用於執行類神經網路計算之一例示性專用積體電路200。系統200包含一主機介面202。主機介面202可接收包含用於一類神經網路計算之參數之指令。參數可包含以下一或多項：應處理的層之數目、用於模型之每一層之對應權重輸入集合、一初始激發輸入集合(即，至類神經網路之輸入(推理由其計算))、每一層之對應輸入及輸出大小、用於類神經網路計算之一步幅值及待處理之層之一類型(例如，一卷積層或一完全連接層)。

主機介面202可將指令發送至一定序器206，該定序器206將指令轉換為低階控制信號，用以控制電路以執行類神經網路計算。在一些

實施方案中，控制信號調節電路中之資料流(例如，權重輸入集合及激發輸入集合如何流動通過電路)。定序器206可將控制信號發送至一統一緩衝器208、一矩陣計算單元212及一向量計算單元214。在一些實施方案中，定序器206亦將控制信號發送至一直接記憶體存取引擎204及動態記憶體210。在一些實施方案中，定序器206係產生控制信號之一處理器。定序器206可使用控制信號之時序以在適當時間將控制信號發送至電路200之每一組件。在一些其他實施方案中，主機介面202接受來自一外部處理器之一控制信號。

主機介面202可將權重輸入集合及初始激發輸入集合發送至直接記憶體存取引擎204。直接記憶體存取引擎204可將激發輸入集合儲存在統一緩衝器208處。在一些實施方案中，直接記憶體存取將權重集合儲存至動態記憶體210，該動態記憶體210可為一記憶體單元。在一些實施方案中，動態記憶體經定位遠離電路。

統一緩衝器208係一記憶體緩衝器。其可用於儲存來自直接記憶體存取引擎204之激發輸入集合及向量計算單元214之輸出。下文將參考圖5更詳細地描述向量計算單元。直接記憶體存取引擎204亦可自統一緩衝器208讀取向量計算單元214之輸出。

動態記憶體210及統一緩衝器208可分別將權重輸入集合及激發輸入集合發送至矩陣計算單元212。在一些實施方案中，矩陣計算單元212係二維脈動陣列。矩陣計算單元212亦可為一維脈動陣列或可執行數學運算(例如，乘法及加法)之其他電路。在一些實施方案中，矩陣計算單元212係一通用矩陣處理器。

矩陣計算單元212可處理權重輸入及激發輸入並將輸出之一向量提供至向量計算單元214。在一些實施方案中，矩陣計算單元將輸出之向量發送至統一緩衝器208，該統一緩衝器208將輸出之向量發送至向量計算單元214。向量計算單元可處理輸出之向量並將經處理輸出

之一向量儲存至統一緩衝器208。經處理輸出之向量可用作至矩陣計算單元212之激發輸入(例如，用於類神經網路中之一後續層)。下文分別參考圖3及圖5更詳細地描述矩陣計算單元212及向量計算單元214。

圖3展示包含一矩陣計算單元之一例示性架構300。矩陣計算單元係二維脈動陣列306。陣列306包含多個胞元304。在一些實施方案中，脈動陣列306之一第一維度320對應於胞元之行，且脈動陣列306之一第二維度322對應於胞元之列。脈動陣列具有的列可多於行、具有的行可多於列或具有的行及列的數目相等。

在經圖解說明之實例中，值載入器302將激發輸入發送至陣列306之列且一權重提取器介面308將權重輸入發送至陣列306之行。然而，在一些其他實施方案中，將激發輸入傳送至陣列306之行且將權重輸入傳送至陣列306之列。

值載入器302可自一統一緩衝器(例如，圖2之統一緩衝器208)接收激發輸入。每一值載入器可將一對應激發輸入發送至陣列306之一相異最左胞元。例如，值載入器312可將一激發輸入發送至胞元314。值載入器亦可將激發輸入發送至一相鄰值載入器，且可在陣列306之另一最左胞元處使用激發輸入。此允許激發輸入移位以在陣列306之另一特定胞元中使用。

權重提取器介面308可自一記憶體單元(例如，圖2之動態記憶體210)接收權重輸入。權重提取器介面308可將一對應權重輸入發送至陣列306之一相異最頂部胞元。例如，權重提取器介面308可將權重輸入發送至胞元314及316。

在一些實施方案中，一主機介面(例如，圖2之主機介面202)使激發輸入沿一個維度移位(例如，移位至右側)貫穿陣列306，同時使權重輸入沿另一維度移位(例如，移位至底部)貫穿陣列306。例如，在

一個時脈循環中，胞元314處之激發輸入可移位至胞元316 (其在胞元314右側)中之一激發暫存器。類似地，胞元314處之權重輸入可移位至胞元318 (其在胞元314下方)處之一權重暫存器。

在每一時脈循環，每一胞元可處理一給定權重輸入、一給定激發輸入及來自一相鄰胞元之一累加輸出以產生一累加輸出。累加輸出亦可被傳遞至沿與給定權重輸入相同之維度之相鄰胞元。下文參考圖4進一步描述一個別胞元。

累加輸出可沿與權重輸入相同之行傳遞(例如，朝向陣列306中之行之底部)。在一些實施方案中，在每一行之底部處，陣列306可包含累加器單元310，其在利用激發輸入多於列之層執行計算時儲存並累加來自每一行之每一累加輸出。在一些實施方案中，每一累加器單元儲存多個平行累加。此將在下文參考圖6進一步描述。累加器單元310可累加每一累加輸出以產生一最終累加值。最終累加值可被傳送至一向量計算單元(例如，圖5之向量計算單元502)。在一些其他實施方案中，累加器單元310將累加值傳遞至向量計算單元而未在利用激發輸入少於列之層處理層或時執行任何累加。

圖4展示一脈動陣列(例如，圖3之脈動陣列306)內部之一胞元之一例示性架構400。

胞元可包含儲存一激發輸入之一激發暫存器406。激發暫存器可取決於胞元在脈動陣列內之位置自一左側相鄰胞元(即，定位於給定胞元左側之一相鄰胞元)或自一統一緩衝器接收激發輸入。胞元可包含儲存一權重輸入之一權重暫存器402。取決於胞元在脈動陣列內之位置，可自一頂部相鄰胞元或自一權重提取器介面傳送權重輸入。胞元亦可包含一總和輸入暫存器404。總和輸入暫存器404可儲存來自頂部相鄰胞元之一累加值。乘法電路408可用於將來自權重暫存器402之權重輸入與來自激發暫存器406之激發輸入相乘。乘法電路408可將乘

積輸出至加總電路410。

加總電路可將乘積與來自總和輸入暫存器404之累加值加總以產生一新累加值。加總電路410接著可將新累加值發送至定位於一底部相鄰胞元中之另一總和輸入暫存器。新累加值可用作底部相鄰胞元中之一加總之一運算元。

胞元亦可將權重輸入及激發輸入移位至相鄰胞元以供處理。例如，權重暫存器402可將權重輸入發送至底部相鄰胞元中之另一權重暫存器。激發暫存器406可將激發輸入發送至右側相鄰胞元中之另一激發暫存器。因此可在一後續時脈循環由陣列中之其他胞元重複使用權重輸入及激發輸入兩者。

在一些實施方案中，胞元亦包含一控制暫存器。控制暫存器可儲存一控制信號，該控制信號判定胞元是否應將權重輸入或激發輸入移位至相鄰胞元。在一些實施方案中，將權重輸入或激發輸入移位花費一或多個時脈循環。控制信號亦可判定是否將激發輸入或權重輸入傳送至乘法電路408或可判定乘法電路408是否對激發及權重輸入操作。控制信號亦可(例如)使用一導線傳遞至一或多個相鄰胞元。

在一些實施方案中，將權重預移位至一權重路徑暫存器412中。權重路徑暫存器412可(例如)自一頂部相鄰胞元接收權重輸入，並基於控制信號將權重輸入傳送至權重暫存器402。權重暫存器402可靜態地儲存權重輸入使得在多個時脈循環中，當激發輸入(例如)透過激發暫存器406傳送至胞元時，權重輸入保留在胞元內且並未被傳送至一相鄰胞元。因此，可(例如)使用乘法電路408將權重輸入施加至多個激發輸入，且可將各自累加值傳送至一相鄰胞元。

圖5展示一向量計算單元502之一例示性架構500。向量計算單元502可自一矩陣計算單元(例如，參考圖2描述之矩陣計算單元)接收累加值之一向量。

向量計算單元502可處理激發單元504處之累加值之向量。在一些實施方案中，激發單元包含將一非線性函數應用至每一累加值以產生激發值之電路。例如，非線性函數可為tanh(x)，其中x係一累加值。

視需要，向量計算單元502可在自激發值產生正規化值之正規化電路506中正規化激發值。

又外視需要，向量計算單元502可使用匯集電路508匯集值(激發值或正規化值)。匯集電路508可將一彙總函數應用至正規化值之一或多者以產生匯集值。在一些實施方案中，彙總函數係傳回正規化值或正規化值之一子集之一最大值、最小值或平均值之函數。

控制信號510可(例如)由圖2之定序器206傳送，且可調節向量計算單元502如何處理累加值之向量。即，控制信號510可調節激發值是否經匯集、正規化或兩者。控制信號510亦可指定激發、正規化或匯集函數以及用於正規化及匯集之其他參數(例如，一步幅值)。

向量計算單元502可將值(例如，激發值、正規化值或匯集值)發送至一統一緩衝器(例如，圖2之統一緩衝器208)。

在一些實施方案中，匯集單元508代替正規化電路506接收激發值，且將匯集值儲存於統一緩衝器中。在一些實施方案中，匯集單元508將匯集值發送至正規化電路506，其產生待儲存於統一緩衝器中之正規化值。

圖6展示用於正規化電路(例如，圖5之正規化電路506)之一例示性架構600。正規化電路可針對每一時脈循環自激發電路602 (例如，圖5之激發電路504)接收激發值之一向量。取決於一系統參數之值，正規化電路可將激發值之向量傳遞至匯集電路(即，未正規化激發值)或自激發值之向量產生正規化值之一向量。例如，若系統參數(例如，由一使用者提供)指示電路將激發值之向量傳遞至匯集電路(例

如，使用者不希望正規化值)，則系統參數可為至一多工器之一信號，該多工器將值直接傳遞至匯集電路並略過正規化電路。

在一些實施方案中，激發值之向量包含藉由將一激發函數應用至基於一權重輸入集合自激發輸入產生之累加值而產生之激發值。

在一些其他實施方案中，用於權重輸入集合之激發值由於在使激發及權重輸入移位時引起的延遲而跨激發值之多個向量交錯。例如，一矩陣計算單元可自來自核心A之一激發輸入集合及一權重輸入集合產生累加值A_0至A_n，自來自核心B之一激發輸入集合及一權重輸入集合產生累加值B_0至B_n，且自來自核心C之一激發輸入集合及一權重輸入集合產生累加值C_0至C_n。累加值A_0至A_n及B_0至B_n可產生於後續時脈循環中，因為權重輸入及激發輸入在計算對應累加值之前跨矩陣計算單元移位，如上文參考圖4描述般。A_0可產生於時脈循環0上，A_1及B_0可產生於時脈循環1上，A_2、B_1及C_0可產生於時脈循環2上，A_n、B_{n-1}及C_{n-2}可產生於時脈循環n上，以此類推。矩陣計算單元可針對時脈循環X產生包含A_0及B_0之累加值之一向量且針對時脈循環X+1產生包含A_1及B_1之累加值之另一向量。因此，一給定核心之累加值(例如，來自核心A之A_0至A_n)可在後續時脈循環中以一交錯方式跨累加值之多個向量展開。

因此，累加值之多個向量可(例如)在藉由圖5之激發電路504處理之後變為激發值之多個向量，且激發值之多個向量之各者可被發送至一相異正規化暫存器行。特定言之，激發電路602可將來自激發值之一向量之每一激發值發送至一相異正規化暫存器行604至610。特定言之，正規化暫存器616至622之各者可接收一各自激發值。一正規化暫存器行可包含串聯連接之一正規化暫存器集合。即，該行中之一第一正規化暫存器之一輸出可作為一輸入發送至該行中之一第二正規化暫存器。在一些實施方案中，每一正規化暫存器儲存一激發值。在一些

其他實施方案中，每一正規化暫存器亦儲存激發值之一平方。在一些實施方案中，正規化電路具有與激發電路或脈動陣列中之行一樣多的正規化暫存器行。

在一些實施方案中，在將激發值之向量提供至正規化暫存器行之前，該電路將向量發送至一平方單元。平方單元可計算每一激發值之一平方以用於計算正規化值，此將在下文進一步描述。平方單元可產生經平方激發值之向量(即，對於激發值之每一向量，具有一經平方激發值之一向量)，且將經平方激發值之向量發送至正規化暫存器行。在一些其他實施方案中，平方單元將激發值之向量及經平方激發值之向量兩者發送至正規化暫存器行。

在一些實施方案中，正規化電路基於一正規化半徑參數形成交錯群組(例如，交錯群組624及628)。正規化半徑參數可指示在計算一正規化值時要使用的來自周圍正規化暫存器之輸出之一數目。輸出之數目可等於正規化半徑參數之兩倍。藉由圖解，交錯群組624及628係由一正規化半徑參數1形成。交錯群組624包含正規化單元632及618，且亦包含零暫存器636。零暫存器636可始終輸出一值0且可在計算正規化電路之邊緣上之正規化值時用作一緩衝器。零暫存器636及638可包含在一零暫存器行612中。下文將參考圖7進一步描述交錯群組內部之值之一實例。

在一些實施方案中，正規化單元(例如，正規化單元626、630)使用來自交錯群組之輸出以產生用以計算一正規化值之一對應分量(例如，交錯群組之暫存器內部之激發值之一平方)。例如，分量可用以產生所有激發值之一平方和。正規化單元可使用平方和以計算正規化值，此將在下文進一步描述。在一些實施方案中，每一交錯群組存在一對應正規化單元。

正規化電路可基於交錯群組產生一激發值之一正規化值。例

如，可將儲存在正規化暫存器632中之一激發值之正規化值儲存在正規化單元626中。特定言之，基於交錯群組624，正規化電路可(例如)使用加總電路計算由交錯群組624內部之正規化暫存器產生之所有平方之總和。該總和可儲存在正規化單元626中。該總和可為對應於一激發值之一正規化值。正規化電路可繼續產生交錯群組628之另一對應正規化值，交錯群組628包含正規化暫存器634、640及620，且對應正規化值可儲存在正規化單元630中。

正規化電路可由經產生正規化值(其可儲存在正規化單元中)產生正規化值之一向量，且可將正規化值之向量發送至匯集電路(若由一類神經網路參數判定)或一統一緩衝器。

圖7展示用於具有正規化暫存器內部之樣本激發值之正規化電路之另一例示性架構700。如交錯群組724及728中證實，正規化半徑參數可為1。特定言之，交錯群組724包含正規化暫存器732及718以及零暫存器736。交錯群組728包含零暫存器738以及正規化暫存器734及740。

正規化暫存器716至720、732、734及740可儲存激發值，例如對應於來自一脈動陣列之行。正規化暫存器740之記法AX,Y (例如，A0,0)表示在時脈循環Y中對應於行X之一激發值。

如圖中證實，以一交錯方式載入激發值。例如，在時脈循環0上，可計算激發值A0,0、A1,0及A2,0，但正規化電路在三個時脈循環中載入三個激發值。在一些實施方案中，以一非交錯方式載入激發值。即，可在一個時脈循環中載入A0,0、A1,0及A2,0。

N0可為針對儲存在正規化暫存器726中之A0,1之一正規化值。N0可基於A0,1及A1,1以及0 (來自零暫存器736)之平方和而計算，此將在下文參考圖8加以描述。類似地，N1可為針對A0,0之一正規化值，其基於A0,0及A1,0以及A2,0 (來自暫存器720)之平方和而計算。

正規化電路可使用一半徑1計算每一激發值之正規化值。其他半徑係可行的。若正規化電路尚未載入一正規化計算所必需的激發值，則正規化電路可將激發值移位至一後續正規化暫存器直至載入必需的激發值。例如，根據一半徑1，針對儲存在正規化暫存器716中之激發值A0,2計算一正規化值需要一激發值A1,2。激發值A1,2可在一後續時脈循環載入至正規化暫存器718中，此時，正規化電路可針對激發值A0,2計算一正規化值。

圖8展示用於正規化電路內部之一正規化單元之一例示性架構800。正規化單元可接收一激發值802。在一些實施方案中，例如當電路判定激發值802處在一錯誤位置中時，激發值802透過一多工器814傳遞至一後續正規化單元(即，激發值需儲存在一後續正規化單元處以進行一正規化計算)。正規化電路可將一控制信號發送至多工器814以通過一特定輸出(例如，一正規化值或一未受影響激發值)。

在一些實施方案中，激發值被傳遞至平方電路804。平方電路804可產生一經平方激發值808，即，將激發值升高至二次冪。平方電路804可將經平方激發值808發送至鄰近正規化單元，例如，正規化單元之相同交錯群組中之其他正規化單元。

在一些實施方案中，經接收激發值在被提供至正規化暫存器行之前已經平方，如上文參考圖6描述。

正規化單元亦可在加總電路806處自鄰近正規化單元接收經平方激發值810。加總電路806可產生經平方激發值808與所接收的經平方激發值810之總和。

該總和可被發送至一記憶體單元812。在一些實施方案中，記憶體單元812包含一查找表及內插單元。正規化單元可使用總和之一部分(例如，總和之一高位元集合)作為一位址以查找由一系統參數提供之一或多個係數。記憶體及內插單元812可基於係數及經平方激發值

之總和產生一正規化按比例調整因數。正規化按比例調整因數可被發送至乘法單元816。

在一些實施方案中，平方和係一12位元值。正規化單元可使用平方和之前4個位元作為查找表之一索引。前4個位元可用以自查找表存取(例如)由一使用者指定之係數。在一些實施方案中，前4個位元存取2個12位元係數：A&B。後8個位元可為在一方程式中使用以計算正規化按比例調整因數之一差量。一例示性方程式係由按比例調整因數=*minimum* (1048575, [A * delta + B * 256 + 2^7]) >> 8給定，其中*minimum*處理兩個引數並傳回具有最小值之引數。

正規化單元可使用乘法單元816將正規化按比例調整因數與激發值802相乘以產生一正規化值。在一些實施方案中，正規化值接著被發送至匯集電路，例如圖5之匯集電路508。

圖9展示用於匯集電路之一例示性架構900。匯集電路可將一彙總函數應用至一或多個正規化或激發值以產生匯集值。藉由圖解，架構900可執行激發或正規化值之一4 x 4集合之一匯集。雖然圖9中所示之匯集具有一正方形區域(即，4 x 4)，但矩形區域亦係可行的。例如，若區域具有n x m之一窗，則架構900可具有n*m個暫存器，即，n行及m列。

匯集電路可(例如)自圖5之正規化電路506接收來自正規化值之向量之一元素序列。例如，該序列可表示一影像之一8 x 8部分之像素，且匯集電路架構900可匯集來自8 x 8部分之一4 x 4子集之值。在一些實施方案中，正規化值一旦由耦合至匯集電路之正規化電路計算便被附加至該序列。在一些實施方案中，類神經網路處理器包含多個平行匯集電路。在每一時脈循環中，每一匯集電路可自正規化電路接收來自正規化值之向量之一各自元素。每一匯集電路可將自正規化電路接收之元素解譯為以光柵順序到達之二維影像。

匯集電路可包含一系列暫存器及記憶體單元。每一暫存器可將一輸出發送至跨儲存在暫存器內部之值應用一彙總函數之彙總電路906。彙總函數可傳回來自一值集合之一最小值、最大值或平均值。

一第一正規化值可被發送至暫存器902並儲存在暫存器902內部。在一後續時脈循環，第一正規化值可移位至一後續暫存器908且儲存在記憶體904中，且一第二正規化值可被發送至暫存器902並儲存在暫存器902內部。

在四個時脈循環之後，四個正規化值儲存在前四個暫存器902、908至912內部。在一些實施方案中，記憶體單元904在先進先出(FIFO)下操作。每一記憶體單元可儲存多達8個正規化值。在記憶體單元904含有一完整像素列之後，記憶體單元904可將一正規化值發送至暫存器914。

在任何給定時刻，彙總電路906可自每一暫存器存取正規化值。暫存器中之正規化值應表示影像之一4 x 4部分之正規化值。

匯集電路可藉由使用彙總電路906自經存取正規化值(例如，一最大正規化值、一最小正規化值或一平均正規化值)產生一匯集值。匯集值可被發送至一統一緩衝器，例如圖2之統一緩衝器208。

在產生第一匯集值之後，匯集電路可繼續藉由透過每一暫存器移位正規化值而產生匯集值使得新正規化值儲存在暫存器中且可由彙總電路906匯集。例如，在架構900中，匯集電路可在4個以上時脈循環中移位正規化值，藉此將記憶體單元中之正規化值移位至暫存器中。在一些實施方案中，匯集電路移位新正規化值直至一新正規化值儲存在一最後最頂部暫存器(例如，暫存器916)中。

彙總電路906接著可匯集儲存在暫存器中之新正規化值。

在一些實施方案中，匯集電路接收激發值之一向量而非接收正規化值之一向量，如上文參考圖5描述。

本說明書中描述之標的物及功能操作之實施例可在數位電子電路、有形體現電腦軟體或韌體、電腦硬體(包含本說明書中揭示之結構及其等結構等效物)或其等之一或多者之組合中實施。本說明書中描述之標的物之實施例可被實施為一或多個電腦程式(即，編碼在一有形非暫時性程式載體上用於由資料處理設備執行或控制資料處理設備之操作之電腦程式指令之一或多個模組)。替代地或此外，可將程式指令編碼在經產生以編碼傳輸至適合接收器設備以由一資料處理設備執行之資訊之一人工產生之傳播信號(例如，一機器產生之電、光學或電磁信號)上。電腦儲存媒體可為一機器可讀儲存裝置、一機器可讀儲存基板、一隨機或串列存取記憶體裝置或其等之一或多者之一組合。

術語「資料處理設備」涵蓋用於處理資料之所有種類的設備、裝置及機器，包含(例如)一可程式化處理器、一電腦或多個處理器或電腦。該設備可包含專用邏輯電路，例如FPGA (場可程式化閘陣列)或ASIC (專用積體電路)。除硬體之外，該設備亦可包含針對討論中的電腦程式產生一執行環境之程式碼，例如，構成處理器韌體、一協定堆疊、一資料庫管理系統、一作業系統或其等之一或多者之一組合之程式碼。

一電腦程式(其亦可稱為或描述為一程式、軟體、一軟體應用程式、一模組、一軟體模組、一指令檔或程式碼)可以任何形式的程式設計語言(包含編譯或解譯語言或宣告或程序語言)寫入且其可以任何形式部署(包含部署為一獨立程式或一模組、組件、子常式或適用於在一計算環境中使用之其他單元)。一電腦程式可(但不一定)對應於一檔案系統中之一檔案。一程式可被儲存在保存其他程式或資料(例如，儲存在一標記語言文件中之一或多個指令檔)之一檔案之一部分中，儲存在專用於討論中的程式之一單個檔案或多個協調檔案(例

如，儲存一或多個模組、子程式或程式碼部分之檔案)中。一電腦程式可經部署以在一個電腦上執行或在定位於一個站點處或跨多個站點分佈且由一通信網路互連之多個電腦上執行。

本說明書中描述之程序及邏輯流程可由執行一或多個電腦程式之一或多個可程式化電腦執行以藉由對輸入資料操作且產生輸出而執行功能。程序及邏輯流程亦可由以下各者執行且設備亦可實施為以下各者：專用邏輯電路，例如，FPGA (場可程式化閘陣列)或ASIC (專用積體電路)。

適用於執行一電腦程式之電腦包含(例如)、可基於通用或專用微處理器或兩者或任何其他種類的中央處理單元。一般而言，一中央處理單元將自一唯讀記憶體或一隨機存取記憶體或兩者接收指令及資料。一電腦之必要元件係用於執行指令之一中央處理單元及用於儲存指令及資料之一或多個記憶體裝置。一般而言，一電腦亦將包含用於儲存資料之一或多個大容量儲存裝置(例如，磁碟、磁光碟或光碟)或可操作地耦合以自該一或多個大容量儲存裝置接收資料或將資料傳送至該一或多個大容量儲存裝置或兩者。然而，一電腦無需具有此等裝置。此外，一電腦可嵌入另一裝置中，例如行動電話、個人數位助理(PDA)、行動音訊或視訊播放器、遊戲控制台、全球定位系統(GPS)接收器或可攜式儲存裝置(例如通用串列匯流排(USB)快閃磁碟機) (僅舉幾例)。

適於儲存電腦程式指令及資料之電腦可讀媒體包含所有形式之非揮發性記憶體、媒體及記憶體裝置，包含(例如)：半導體記憶體裝置，例如，EPROM、EEPROM及快閃記憶體裝置；磁碟，例如內部硬碟或可抽換式磁碟；磁光碟；及CD-ROM及DVD-ROM光碟。處理器及記憶體可由專用邏輯電路補充或併入至專用邏輯電路中。

為發送與一使用者之互動，可在具有用於將資訊顯示給使用者

之一顯示裝置(例如，一CRT (陰極射線管)或LCD (液晶顯示器)監視器)及一鍵盤及使用者可藉由其將輸入發送至電腦之一指標裝置(例如，一滑鼠或一軌跡球)之一電腦上實施本說明書中所描述之標的物之實施例。其他種類之裝置亦可用以發送與一使用者之互動；例如，提供至使用者之回饋可係任何形式之感官回饋，例如視覺回饋、聽覺回饋或觸覺回饋；來自使用者之輸入可以任何形式接收，包含聲學、語音或觸覺輸入。此外，一電腦可藉由將文件發送至由一使用者使用之一裝置或自該裝置接收文件而與該使用者互動；例如，藉由回應於自一使用者之用戶端裝置上之一網頁瀏覽器接收之請求將網頁發送至該網頁瀏覽器。

可在包含一後端組件(例如作為一資料伺服器)或包含一中間軟體組件(例如一應用程式伺服器)或包含一前端組件(例如，具有一圖形使用者介面或一使用者可透過其與本說明書中所描述之標的物之一實施方案互動之一網頁瀏覽器之一用戶端電腦)或一或多個此等後端、中間軟體或前端組件之任何組合之一電腦系統中實施本說明書中所描述之標的物之實施例。系統之組件可藉由數位資料通信(例如，一通信網路)之任何形式或媒體互連。通信網路之實例包含一區域網路(「LAN」)及一廣域網路(「WAN」)，例如，網際網路。

計算系統可包含用戶端及伺服器。用戶端及伺服器通常彼此遠離且通常透過一通信網路互動。用戶端與伺服器之關係由運行於各自電腦上且彼此具有一用戶端-伺服器關係之電腦程式引起。

雖然本說明書含有諸多特定實施方案細節，但不應將此等細節理解為對任何發明或可主張之內容之範疇之限制，而應理解為特定發明之特定實施例所特有之特徵之描述。亦可在一單一實施例中組合實施在本說明書中在單獨實施例之上下文中所描述之特定特徵。相反地，亦可在多項實施例中單獨地實施或以任何適合子組合實施在一單

一實施例之上下文中所描述之各種特徵。此外，儘管在上文可將特徵描述為以特定組合起作用且甚至最初如此主張，然來自一經主張組合之一或多個特徵可在一些情況中自該組合刪除且該經主張組合可關於一子組合或一子組合之變動。

類似地，雖然在圖式中依一特定順序描繪操作，但此不應理解為要求依所展示之特定順序或循序順序執行此等操作，或執行全部經圖解說明之操作以達成所要結果。在某些情況中，多任務處理及平行處理可為有利的。此外，不應將上文所描述之實施例中之各種系統模組及組件之分離理解為在所有實施例中需要此分離，且應理解，通常可將所描述之程式組件及系統一起整合於一單一軟體產品中或封裝至多個軟體產品中。

已描述標的物之特定實施例。其他實施例係在以下申請專利範圍之範疇內。例如，敘述於申請專利範圍中之動作可以一不同順序執行且仍達成所要結果。作為一實例，在附圖中描繪之程序不一定需要所展示之特定順序或循序順序以達成所要結果。在特定實施方案中，多任務及平行處理可係有利的。

【符號說明】

100	程序/方法
102	步驟
104	步驟
106	步驟
108	步驟
200	專用積體電路
202	主機介面
204	直接記憶體存取引擎
206	定序器

208 統一緩衝器

210 動態記憶體

212 矩陣計算單元

214 向量計算單元

300 架構

302 值載入器

304 單元

306 二維脈動陣列

308 權重擷取器介面

310 累加器單元

312 值載入器

314 單元

316 單元

318 單元

320 第一維度

322 第二維度

400 架構

402 權重暫存器

404 求總和輸入暫存器

406 激發暫存器

408 乘法電路

410 加總電路

412 權重路徑暫存器

100 程序/方法

102 步驟

104 步驟

106	步驟
108	步驟
200	專用積體電路
202	主機介面
204	直接記憶體存取引擎
206	定序器
208	統一緩衝器
210	動態記憶體
212	矩陣計算單元
214	向量計算單元
300	架構
302	值載入器
304	胞元
306	二維脈動陣列
308	權重擷取器介面
310	累加器單元
312	值載入器
314	胞元
316	胞元
318	胞元
320	第一維度
322	第二維度
400	架構
402	權重暫存器
404	總和輸入暫存器
406	激發暫存器

408 乘法電路

410 加總電路

412 權重路徑暫存器

500 架構

502 向量計算單元

504 激發電路

506 正規化單元

508 匯集單元

510 控制信號

600 架構

602 激發電路

604 正規化暫存器行

606 正規化暫存器行

608 正規化暫存器行

610 正規化暫存器行

612 零暫存器行

616 正規化暫存器

618 正規化暫存器/正規化單元

620 正規化暫存器

622 正規化暫存器

624 交錯群組

626 正規化單元

628 交錯群組

630 正規化單元

632 正規化單元/正規化暫存器

634 正規化暫存器

636	零暫存器
638	零暫存器
640	正規化暫存器
700	架構
716	正規化暫存器
718	正規化暫存器
720	正規化暫存器
724	交錯群組
726	正規化暫存器
728	交錯群組
732	正規化暫存器
734	正規化暫存器
736	零暫存器
738	零暫存器
740	正規化暫存器
800	架構
802	激發值
804	平方電路
806	加總電路
808	經平方激發值
810	所接收經平方激發值
812	記憶體單元
814	多工器
816	乘法單元
900	匯集電路架構
902	暫存器

904	記憶體
906	彙總電路
908	暫存器
910	暫存器
912	暫存器
914	暫存器
916	暫存器

申請專利範圍

1. 一種用於對包括複數個層之一類神經網路執行類神經網路計算之電路，該電路包括：

激發電路，其經組態以接收累加值之一向量且經組態以將一函數應用至每一累加值以產生激發值之一向量；及

正規化電路，其耦合至該激發電路且經組態以針對每一激發值產生一各自正規化值。

2. 如請求項1之電路，其中該激發電路自該電路中之一脈動陣列接收累加值之該向量。

3. 如請求項1或2之電路，其中該正規化電路包括複數個正規化暫存器行，每一正規化暫存器行包括串聯連接之複數個正規化暫存器，每一正規化暫存器行經組態以接收一各自激發值，其中該正規化電路經組態以在一或多個正規化暫存器周圍形成群組，每一群組對應於一正規化單元，且每一正規化單元經組態以針對該各自激發值計算一各自正規化值。

4. 如請求項3之電路，其中每一正規化暫存器經組態以將該相異激發值傳遞至一相鄰正規化行。

5. 如請求項3之電路，其中每一群組係使用一正規化半徑參數形成。

6. 如請求項3之電路，其中每一正規化單元經組態以：

接收該各自激發值；

自該各自激發值產生一各自中間正規化值；及

將該各自中間正規化值發送至一或多個鄰近正規化單元。

7. 如請求項6之電路，其中產生該各自中間正規化值包括產生該各自激發值之一平方。

8. 如請求項6之電路，其中每一正規化單元進一步經組態以：

自一或多個鄰近正規化單元接收自激發值產生之一或多個中間正規化值；

將每一中間正規化值加總以產生一索引；

使用該索引以自一查找表存取一或多個值；

自該一或多個值及該索引產生一按比例調整因數；及

自該按比例調整因數及該各自激發值產生該各自正規化值。

9. 如請求項1或2之電路，其進一步包括匯集電路，該匯集電路經組態以接收該等正規化值且經組態以匯集該等正規化值以產生一匯集值。

10. 請求項9之電路，其中該匯集電路經組態以將該複數個正規化值儲存在複數個暫存器及複數個記憶體單元中，

其中該複數個暫存器及該複數個記憶體單元經串聯連接，其中每一暫存器儲存一個正規化值且每一記憶體單元儲存複數個正規化值，

其中該匯集電路經組態以在每個時脈循環之後將一給定正規化值移位至一後續暫存器或記憶體單元，且

其中該匯集電路經組態以自該等正規化值產生該匯集值。

11. 如請求項1或2之電路，其進一步包括匯集電路，該匯集電路經組態以接收該等激發值且經組態以匯集該等激發值以產生一匯集值。

12. 如請求項11之電路，其中該匯集電路經組態以將該複數個激發值儲存在複數個暫存器及複數個記憶體單元中，

其中該複數個暫存器及該複數個記憶體單元經串聯連接，其中每一暫存器儲存一個正規化值且每一記憶體單元儲存複數個激發值，

其中該匯集電路經組態以在每個時脈循環之後將一給定激發值移位至一後續暫存器或記憶體單元，且

其中該匯集電路經組態以自該等激發值產生該匯集值。

圖式

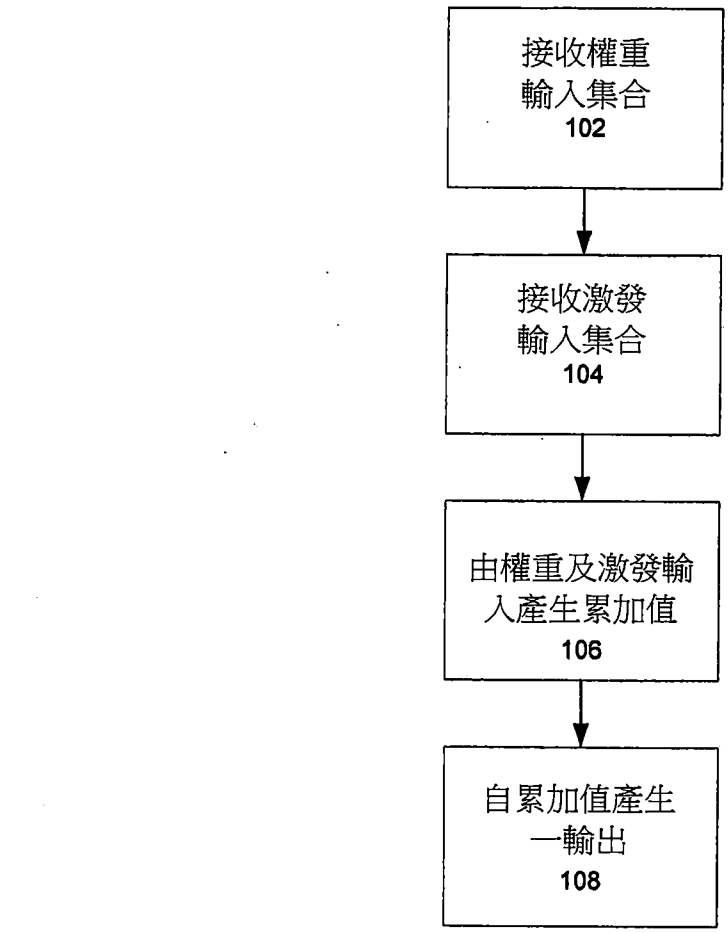

接收權重輸入集合
102
接收激發輸入集合
104
由權重及激發輸入產生累加值
106
自累加值產生一輸出
108
100

圖 1

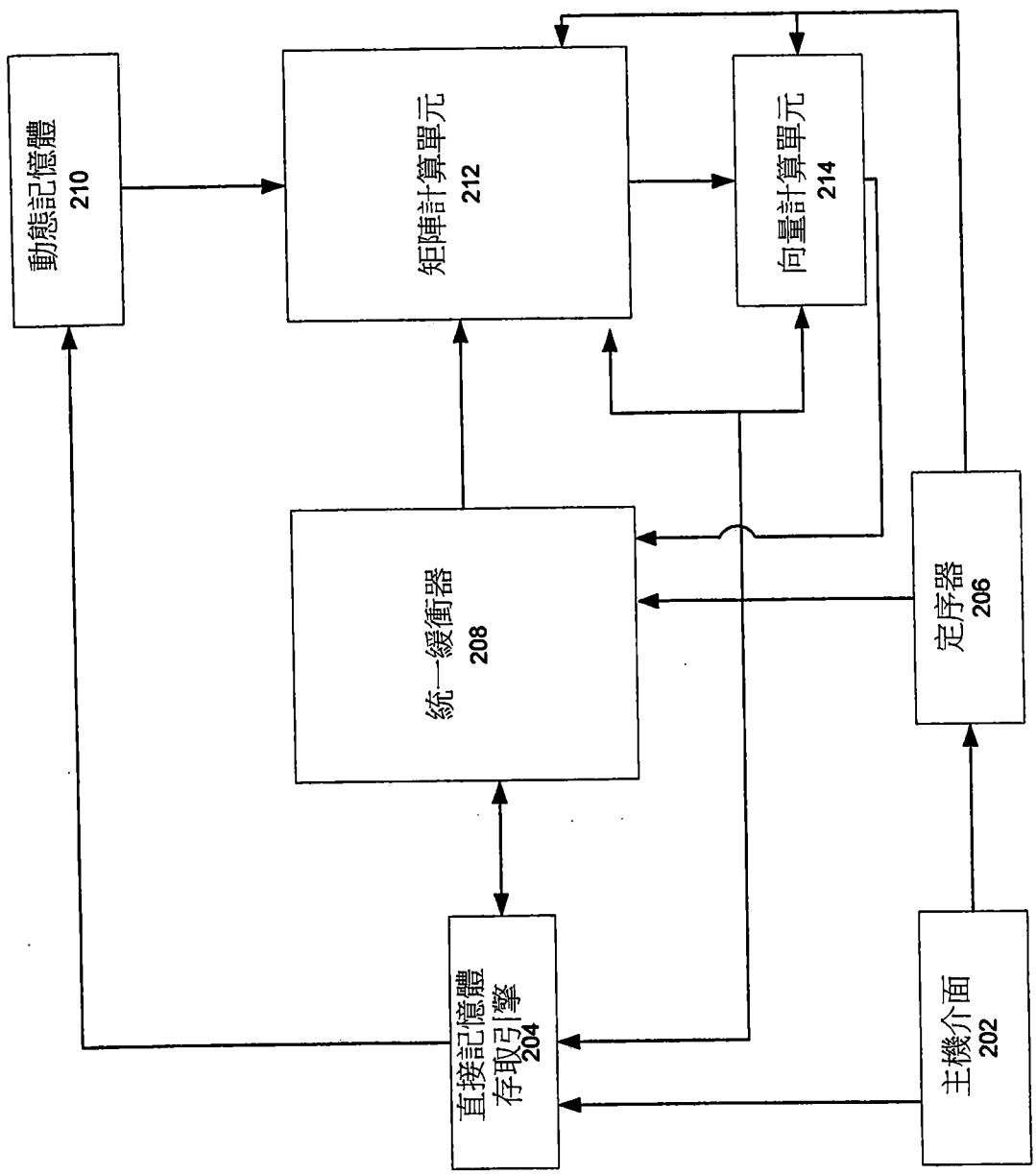

動態記憶體
210
矩陣計算單元
212
向量計算單元
214
統一緩衝器
208
定序器
206
直接記憶體
存取引擎
204
主機介面
202
200

圖 2

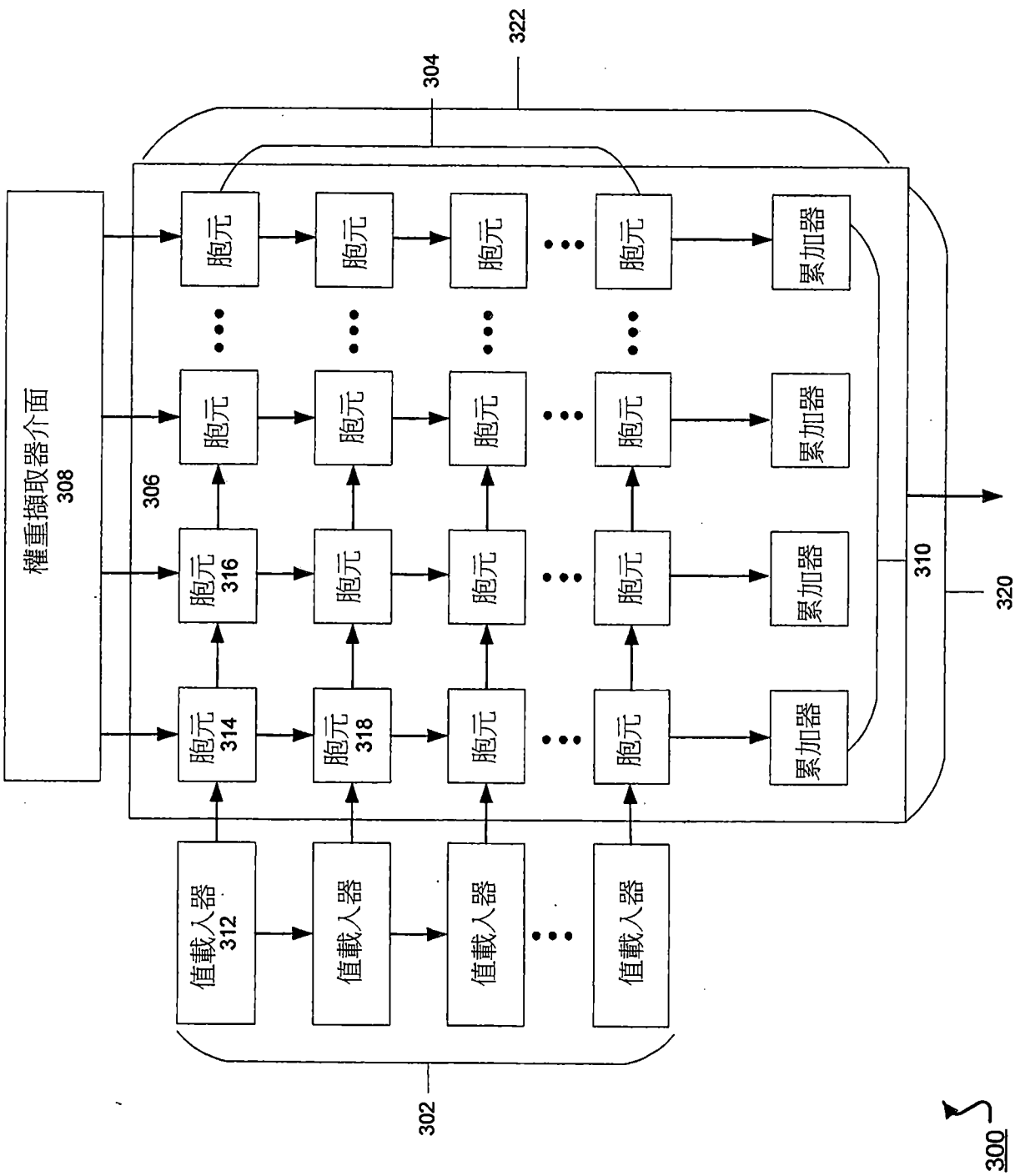

權重擷取器介面
308
值載入器
312
胞元
314
胞元
316
胞元
318
胞元
累加器
302
304
306
310
320
322
300

圖 3

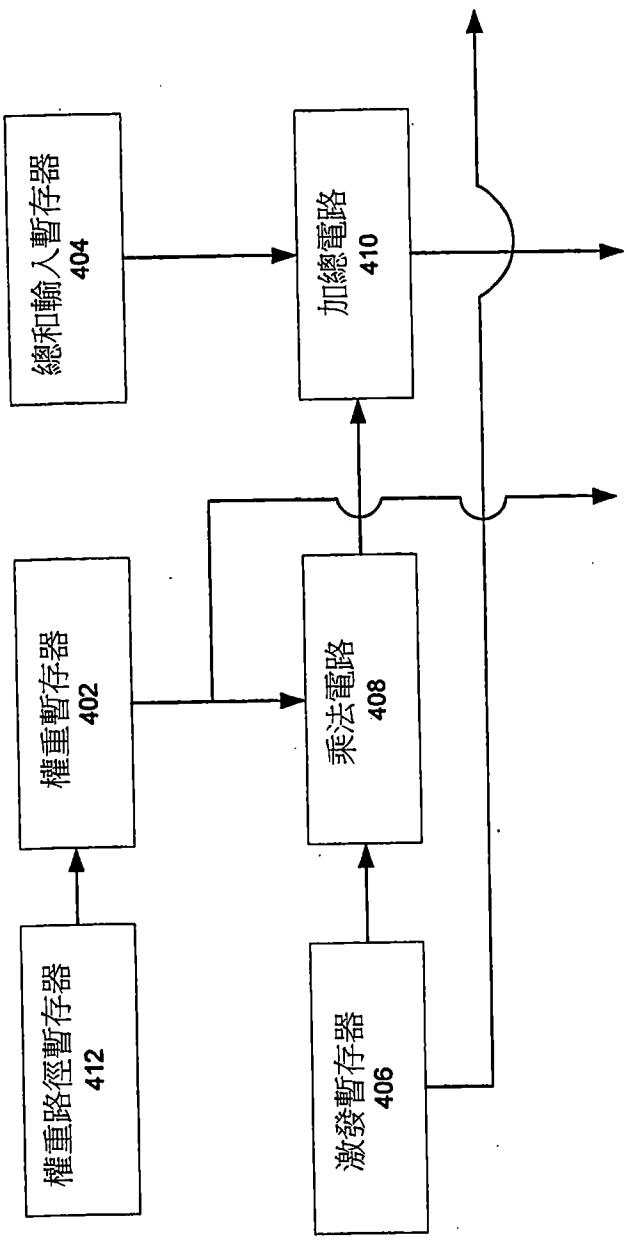

總和輸入暫存器
404
加總電路
410
權重暫存器
402
乘法電路
408
權重路徑暫存器
412
激發暫存器
406
400

圖 4

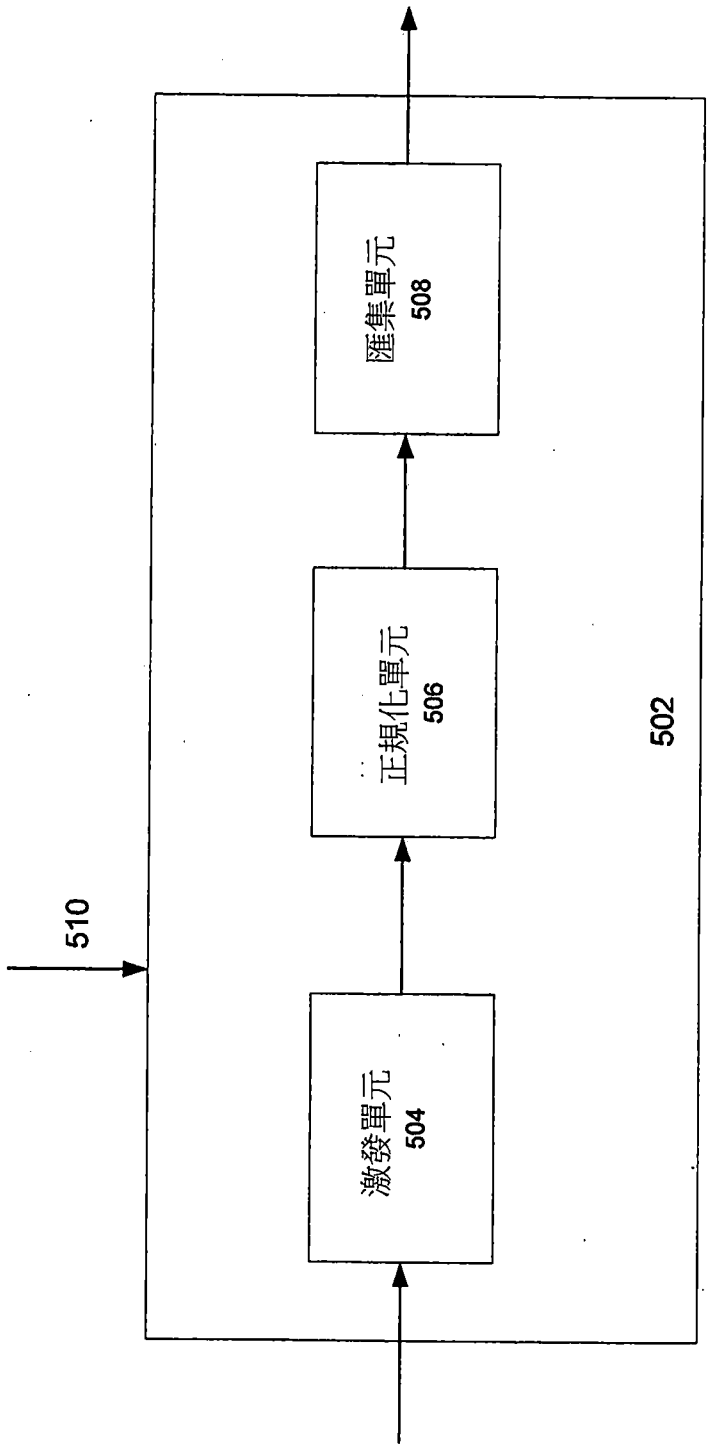


圖 5

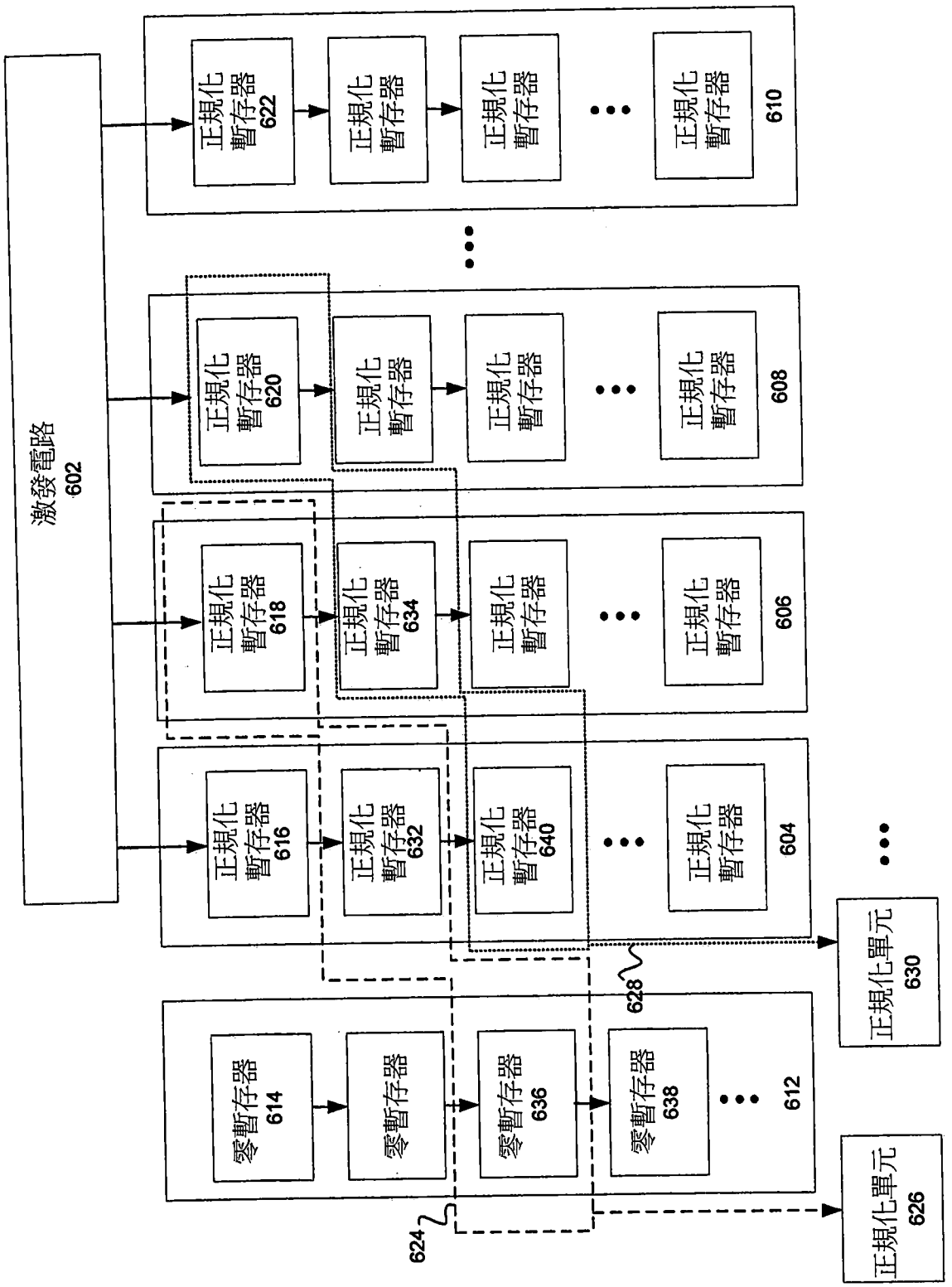

激發電路
602
正規化暫存器
622
正規化暫存器
620
正規化暫存器
618
正規化暫存器
616
正規化暫存器
610
正規化暫存器
608
正規化暫存器
606
正規化暫存器
634
正規化暫存器
632
正規化暫存器
604
正規化暫存器
640
正規化暫存器
630
正規化單元
零暫存器
614
零暫存器
零暫存器
636
零暫存器
638
612
正規化單元
626
624
628
600

圖 6

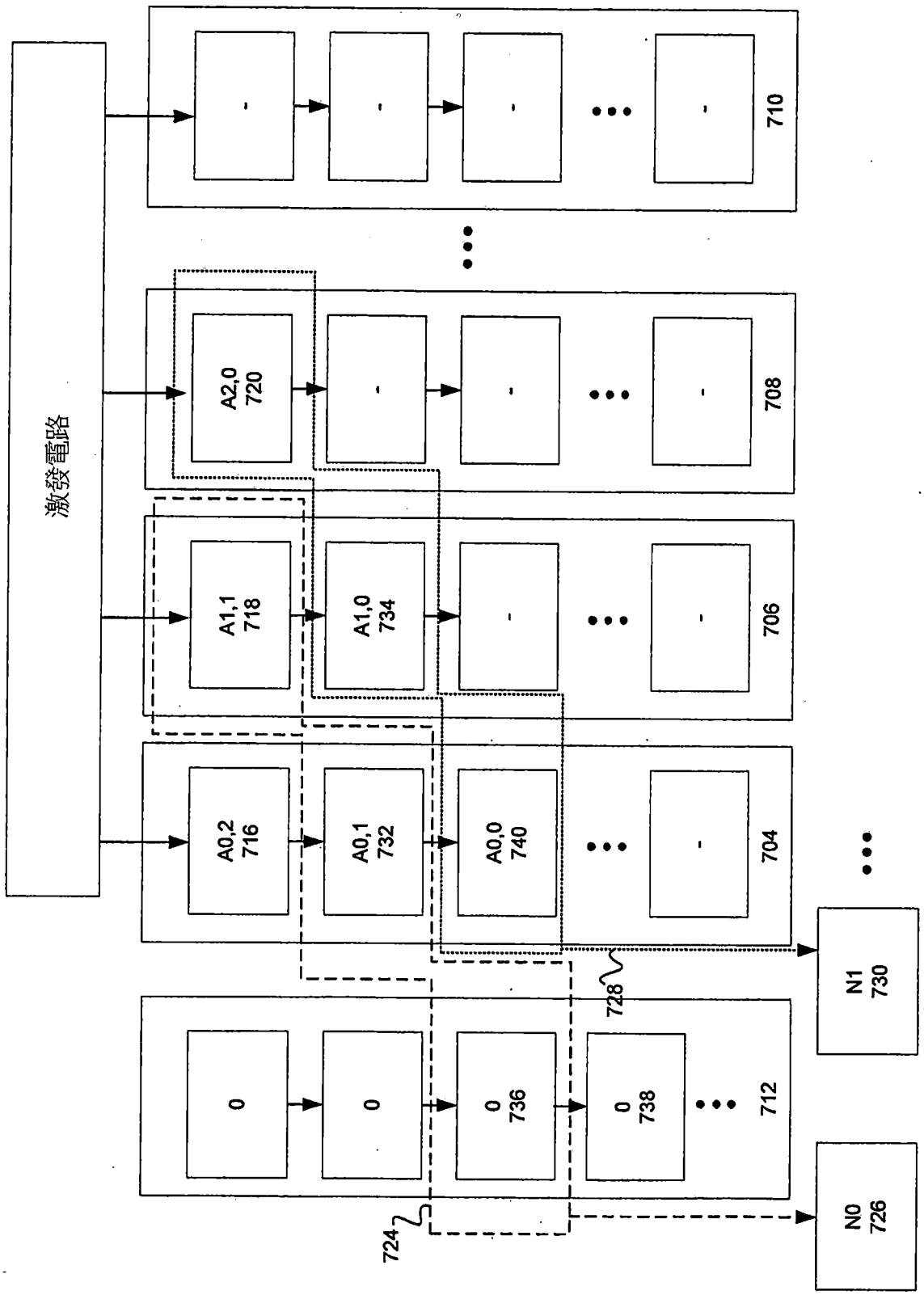

激發電路
A2,0
720
A1,1
718
A1,0
734
A0,2
716
A0,1
732
A0,0
740
0
736
0
738
710
708
706
704
712
728
724
N1
730
N0
726
700

圖 7

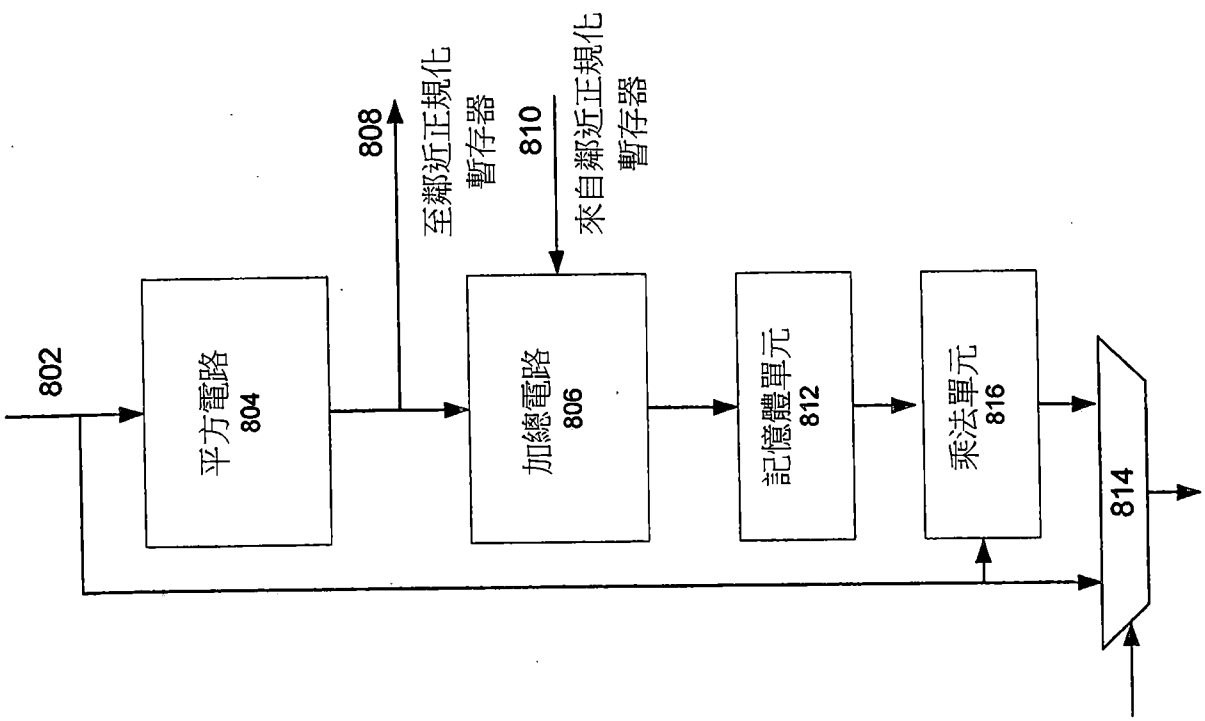

802
平方電路
804
808
至鄰近正規化
暫存器
加總電路
806
810
來自鄰近正規化
暫存器
記憶體單元
812
乘法單元
816
814
800

圖 8

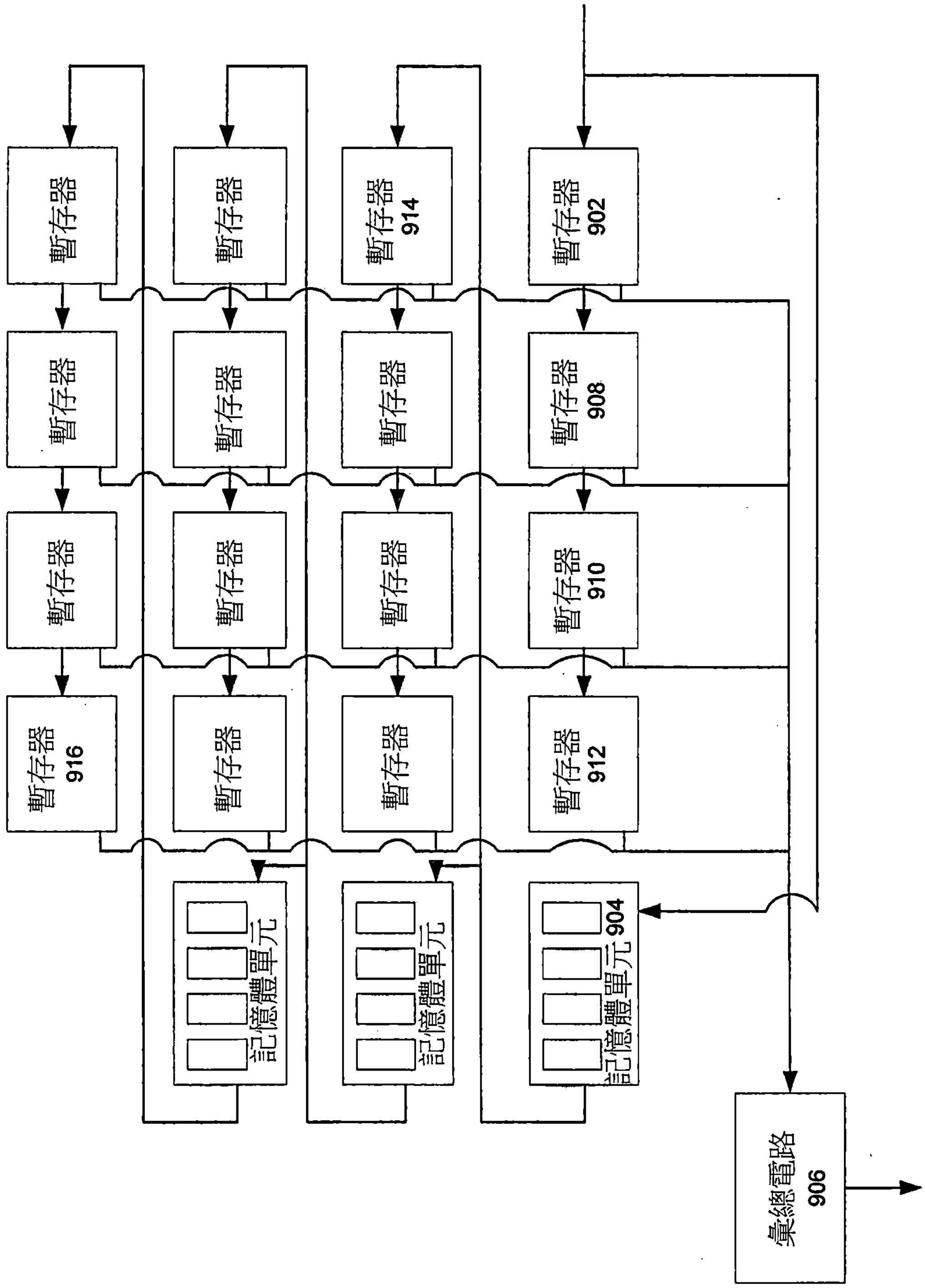

暫存器 902
暫存器 908
暫存器 910
暫存器 912
暫存器 914
暫存器 916
暫存器
記憶體單元 904
記憶體單元
彙總電路 906
900

圖 9