I851499

(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 I851499 B**

(45)公告日： 中華民國 113 (2024) 年 08 月 01 日

(21)申請案號：112149110 (22)申請日： 中華民國 105 (2016) 年 05 月 20 日

(51)Int. Cl.： *G06N3/02 (2006.01)* *G06N3/04 (2023.01)*
G06N99/00 (2019.01) *G06F15/80 (2006.01)*

(30)優先權：2015/05/21 美國 62/164,931
2015/09/03 美國 14/844,524

(71)申請人：美商谷歌有限責任公司 (美國) GOOGLE LLC (US)
美國

(72)發明人：羅斯 強納森 ROSS, JONATHAN (US)；約皮 諾曼 保羅 JOUPPI, NORMAN PAUL (US)；菲而普斯 安德魯 艾佛列特 PHELPS, ANDREW EVERETT (US)；楊 雷金納德 克里福德 YOUNG, REGINALD CLIFFORD (US)；諾里 湯瑪士 NORRIE, THOMAS (US)；索森 格雷戈里 麥克 THORSON, GREGORY MICHAEL (US)；劉 丹 LUU, DAN (US)

(74)代理人：陳長文；簡秀如；金若芸

(56)參考文獻：

TW	201232429A	TW	201421382A
CN	1333518A	CN	101681450A
CN	104238993A	EP	0422348A2
US	2002/0168100A1	US	2014/0180984A1

審查人員：吳鴻鎮

申請專利範圍項數：20 項 圖式數：6 共 32 頁

(54)名稱

用於執行類神經網路計算之電路、方法及非暫時性機器可讀儲存裝置

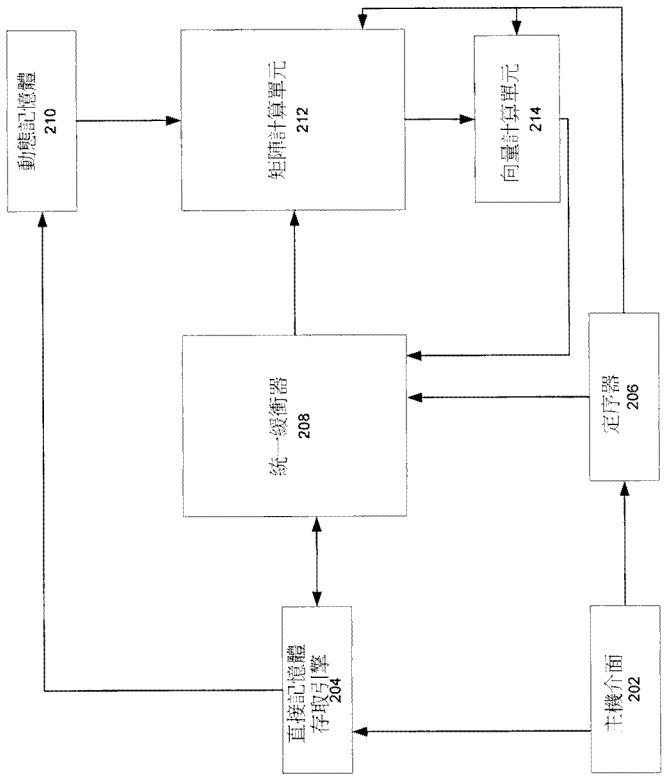
(57)摘要

本發明揭示一種用於對包括複數個類神經網路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：一矩陣計算單元，其經組態以針對該複數個類神經網路層之各者：接收用於該類神經網路層之複數個權重輸入及複數個激發輸入，且基於該複數個權重輸入及該複數個激發輸入產生複數個累加值；及向量計算單元，其通信地耦合至該矩陣計算單元且經組態以針對該複數個類神經網路層之各者：將一激發函數應用至由該矩陣計算單元產生之每一累加值以產生用於該類神經網路層之複數個激發值。

A circuit for performing neural network computations for a neural network comprising a plurality of neural network layers, the circuit comprising: a matrix computation unit configured to, for each of the plurality of neural network layers: receive a plurality of weight inputs and a plurality of activation inputs for the neural network layer, and generate a plurality of accumulated values based on the plurality of weight inputs and the plurality of activation inputs; and a vector computation unit communicatively coupled to the matrix computation unit and configured to, for each of the plurality of neural network layers: apply an activation function to each accumulated value generated by the matrix computation unit to generate a plurality of activated values for the neural network layer.

TW I851499 B

指定代表圖：



【圖2】

符號簡單說明：

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

專利案號：112149110
原申請案號：111106513
I851499
公告本 分割案

【發明摘要】

【中文發明名稱】

用於執行類神經網路計算之電路、方法及非暫時性機器可讀儲存裝置

【英文發明名稱】

CIRCUIT, METHOD AND NON-TRANSITORY MACHINE-READABLE STORAGE DEVICES FOR PERFORMING NEURAL NETWORK COMPUTATIONS

【中文】

本發明揭示一種用於對包括複數個類神經網路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：一矩陣計算單元，其經組態以針對該複數個類神經網路層之各者：接收用於該類神經網路層之複數個權重輸入及複數個激發輸入，且基於該複數個權重輸入及該複數個激發輸入產生複數個累加值；及向量計算單元，其通信地耦合至該矩陣計算單元且經組態以針對該複數個類神經網路層之各者：將一激發函數應用至由該矩陣計算單元產生之每一累加值以產生用於該類神經網路層之複數個激發值。

【英文】

A circuit for performing neural network computations for a neural network comprising a plurality of neural network layers, the circuit comprising: a matrix computation unit configured to, for each of the plurality of neural network layers: receive a plurality of weight inputs and a plurality of activation inputs for the neural network layer, and generate a plurality of accumulated values based on the plurality of

weight inputs and the plurality of activation inputs; and a vector computation unit communicatively coupled to the matrix computation unit and configured to, for each of the plurality of neural network layers: apply an activation function to each accumulated value generated by the matrix computation unit to generate a plurality of activated values for the neural network layer.

【指定代表圖】

圖2

【代表圖之符號簡單說明】

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

【發明說明書】

【中文發明名稱】

用於執行類神經網路計算之電路、方法及非暫時性機器可讀儲存裝置

【英文發明名稱】

CIRCUIT, METHOD AND NON-TRANSITORY MACHINE-READABLE STORAGE DEVICES FOR PERFORMING NEURAL NETWORK COMPUTATIONS

【技術領域】

【先前技術】

本說明書係關於計算硬體中之類神經網路推理。

類神經網路係採用模型之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理之專用硬體電路。

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類神經網路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：一矩陣計算單元，其經組態以針對該複數個類神經網路層之各者：接收用於該類神經網路層之複數個權重輸入及複數個激發輸

入，基於該複數個權重輸入及該複數個激發輸入產生複數個累加值；及一向量計算單元，其通信地耦合至該矩陣計算單元且經組態以針對該複數個類神經網路層之各者：將一激發函數應用至由該矩陣計算單元產生之每一累加值以產生用於該類神經網路層之複數個激發值。

實施方案可包含以下特徵之一或多者。一統一緩衝器(unified buffer)通信地耦合至該矩陣計算單元及該向量計算單元，其中該統一緩衝器經組態以自該向量計算單元接收輸出並儲存該輸出，且該統一緩衝器經組態以將該經接收輸出作為輸入發送至該矩陣計算單元。一定序器經組態以自一主機裝置接收指令並自該等指令產生複數個控制信號，其中該複數個控制信號控制通過該電路之資料流；及一直接記憶體存取引擎通信地耦合至該統一緩衝器及該定序器，其中該直接記憶體存取引擎經組態以將該複數個激發輸入發送至該統一緩衝器，其中該統一緩衝器經組態以將該複數個激發輸入發送至該矩陣計算單元，且其中該直接記憶體存取引擎經組態以自該統一緩衝器讀取結果資料。一記憶體單元經組態以將該多個權重輸入發送至該矩陣計算單元，且其中該直接記憶體存取引擎經組態以將該複數個權重輸入發送至該記憶體單元。該矩陣計算單元被組態為包括複數個胞元之二維脈動陣列。該複數個權重輸入沿該脈動陣列之一第一維度移位穿過第一複數個胞元，且其中該複數個激發輸入沿該脈動陣列之一第二維度移位穿過第二複數個胞元。對於該複數個層中之一給定層，該複數個激發輸入之一計數大於該脈動陣列之該第二維度之一大小，且其中該脈動陣列經組態以：將該複數個激發輸入劃分為部分，其中每一部分具有小於或等於該第二維度之該大小之一大小；針對每一部分產生累加值之一各自部分；及組合累加值之每一部分以針對該給定層產生累加值之一向量。對於該複

數個層中之一給定層，該複數個權重輸入之一計數大於該脈動陣列之該第一維度之一大小，且其中該脈動陣列經組態以：將該複數個權重輸入劃分為部分，其中每一部分具有小於或等於該第一維度之該大小之一大小；針對每一部分產生累加值之一各自部分；及組合累加值之每一部分以各自對該給定層產生累加值之一向量。該複數個胞元中之每一胞元包括：一權重暫存器，其經組態以儲存一權重輸入；一激發暫存器，其經組態以儲存一激發輸入且經組態以將該激發輸入發送至沿該第二維度之一第一相鄰胞元中之另一激發暫存器；一總和輸入(sum-in)暫存器，其經組態以儲存一先前加總值；乘法電路，其通信地耦合至該權重暫存器及該激發暫存器，其中該乘法電路經組態以輸出該權重輸入與該激發輸入之一乘積；及加總電路，其通信地耦合至該乘法電路及該總和輸入暫存器，其中該加總電路經組態以輸出該乘積與該先前加總值之一總和，且其中該加總電路經組態以將該總和發送至沿該第一維度之一第二相鄰胞元中之另一總和輸入暫存器。該複數個胞元中之一或多個胞元各經組態以將各自總和儲存在一各自累加器單元中，其中該各自總和係一累加值。該脈動陣列之該第一維度對應於該脈動陣列之行，且其中該脈動陣列之該第二維度對應於該脈動陣列之列。該向量計算單元將每一激發值正規化以產生複數個正規化值。該向量計算單元匯集(pool)一或多個激發值以產生複數個匯集值。

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。在硬體中實施一類神經網路處理器相對於軟體中之實施方案改良效率，例如增加速度及處理量並減少功率及成本。此可用於推理應用。將類神經網路處理器之組件整合至一個電路中允許計算推理而不招致晶片外通信之懲罰。此外，該電路可處理具有數個輸入(例如，大於該電路內

之一矩陣計算單元之一維度之一大小之數目個權重輸入或激發輸入)之類神經網路層。例如，該電路可處理按類神經網路之每個類神經元之大量權重輸入。

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係用於使用一脈動陣列對激發輸入多於脈動陣列中之列之一給定類神經網路層執行計算之另一例示性程序之一流程圖。

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理。例如，給定一輸入，類神經網路可計算針對輸入之一推理。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理。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理，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理，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。

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，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。參數可包含以下至少一或多項：應處理的層之數目、用於層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係產生時脈信號之一處理器。定序器206可使用時脈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時脈信號。

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。二維脈動陣列306可為一正方形陣列。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。最左胞元可為沿陣列306之一最左行之一胞元。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。最頂部胞元可為沿陣列306之一最頂部列之一胞元。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及一給定激發輸入以產生一累加輸出。累加輸出亦可被傳遞至沿與給定權重輸入相同之維度之一相鄰胞元。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，其在利用權重輸入多於行之層或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處理權重輸入少於行之層或激發輸入少於列之層時執行任何累加。

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一正規化單元506中正規化激發值。

又視需要，向量計算單元502可使用一匯集單元508匯集值(激發值或正規化值)。匯集單元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代替正規化單元506接收激發值，且匯集單元508將匯集值發送至正規化單元506，其產生待儲存於統一緩衝器中之正規化值。

圖6係用於使用一脈動陣列對激發輸入多於脈動陣列中之列之一給定類神經網路層執行計算之例示性程序之一流程圖。為了方便起見，將關於執行程序600之一系統描述程序600。在一些實施方案中，一主機介面或一定序器(例如，分別為圖2之主機介面202或定序器206)執行程序600。在一些其他實施方案中，主機介面自執行程序600之一外部處理器接收指令。

如上文描述，每一層可具有多個激發輸入集合，且每一權重輸入集

合可被傳送至陣列之相異列處之胞元。在一些實施方案中，類神經網路之一些層具有的激發輸入集合多於陣列之列。

系統可(例如)使用一比較器判定給定類神經網路層存在的激發輸入集合是否多於脈動陣列中之列。在一些實施方案中，系統在編譯時間做出判定。一激發輸入集合可對應於被提供至陣列之一單一列之激發輸入。

若列多於激發輸入集合(步驟602)，則系統可如上文在圖3之脈動陣列306中描述般產生累加值(步驟604)。

若待處理之激發輸入集合多於陣列中之列(步驟602)，則系統可將激發輸入集合劃分為部分使得每一部分具有小於或等於陣列中之列之數目之一大小(步驟606)。

系統接著可針對激發輸入之每一部分產生累加值之一部分(步驟608)。一累加值可為至沿一給定行之胞元之激發輸入及權重輸入之乘積之一總和(例如，如圖3之脈動陣列306中所描述)。可將累加值之每一部分儲存在一緩衝器中直至已處理激發輸入之所有部分。緩衝器可為圖3之累加器單元310中之一緩衝器、脈動陣列中之一緩衝器或圖2之統一緩衝器208。

系統接著可將累加值之所有部分組合成累加值之一向量(步驟610)。特定言之，系統可存取累加值之先前儲存部分之緩衝器並(例如)使用圖3之累加器單元310將累加值累加以產生累加值之一向量。系統可將累加值之向量發送至一向量計算單元(例如，圖2之向量計算單元214)。

例如，若陣列中存在256個列且在一給定層處將處理300個激發輸入集合，則系統可自256個激發輸入集合產生256個最終累加值以完全利用脈動陣列並將256個最終累加值儲存在一緩衝器中。系統接著可自44個剩

餘激發輸入集合產生44個最終累加值。最後，系統可組合全部300個最終累加值以形成一向量並將向量發送至向量計算單元。

若權重輸入集合多於陣列之行，則系統可執行類似操作。即，系統可將權重輸入集合劃分為具有的權重輸入集合少於陣列中之行數之部分，針對每一部分產生累加值且將累加值組合成一向量以在向量計算單元中使用。在一些實施方案中，系統可比較累加值之數目與陣列中之行數，而非比較權重輸入集合之數據與陣列中之行數。

雖然已描述其中將權重輸入傳送至陣列之行且將激發輸入傳送至陣列之列之系統，但在一些實施方案中，將權重輸入傳送至陣列之列且將激發輸入被傳送至陣列之行。

雖然硬體被描述為用於計算推理，但硬體可用於以下一或多者：卷積或完全連接類神經網路訓練、線性或邏輯回歸、叢集(例如，k平均值叢集)、視訊編碼及影像處理。

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:胞元

306:二維脈動陣列

308:權重提取器介面

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:激發單元

506:正規化單元

508:匯集單元

510:控制信號

600:程序

602:步驟

604 :步驟

606 :步驟

608 :步驟

610 :步驟

【發明申請專利範圍】

【請求項1】

一種用於執行類神經網路計算之系統，該系統包括一或多個處理器，

其中該一或多個處理器包括一矩陣計算單元，該矩陣計算單元包括一多維脈動陣列，其包括一第一群組胞元(cells)及一第二群組胞元，該第二群組胞元鄰近於該第一群組胞元，該矩陣計算單元經組態以：

藉由該第一群組胞元且沿該矩陣計算單元之一第一維度，接收權重輸入及激發輸入；

藉由該第一群組胞元，使用至少該等權重輸入及該等激發輸入以針對一類神經網路之一類神經網路層執行類神經網路計算之一部分；

藉由該第一群組胞元，使用經接收之該等權重輸入及該等激發輸入處理一累加輸出；

將跨一第二維度之該累加輸出提供至該第二群組胞元；及

使用跨該第一群組胞元及該第二群組胞元之至少該累加輸出產生一最終累加輸出。

【請求項2】

如請求項1之系統，其中該類神經網路計算包括該等權重輸入之一各自權重輸入與該等激發輸入之一各自激發輸入相乘之乘法。

【請求項3】

如請求項1之系統，其中該累加輸出係儲存於複數個累加器單元之一

各自累加器單元中，該各自累加器單元係連接至該第一群組胞元。

【請求項4】

如請求項1之系統，其中：

該第一群組胞元接收沿該第一維度之該等權重輸入及沿該第二維度之該等激發輸入；及

該第一維度係一列維度而該第二維度係一行維度，或者該第一維度係一行維度而該第二維度係一列維度。

【請求項5】

如請求項1之系統，其中該矩陣計算單元進一步經組態以使用至少該第一群組胞元而產生該等激發輸入。

【請求項6】

如請求項5之系統，其中：

該類神經網路層係一第一類神經網路；及

為產生該等激發輸入，該矩陣計算單元經組態以至少部分地執行該類神經網路中在該第一類神經網路之前的一第二類神經網路之類神經網路計算。

【請求項7】

如請求項1之系統，其中各胞元進一步包括一控制暫存器，其經組態以儲存一控制信號以用於判定是否將一權重輸入或一激發輸入移位至一相鄰胞元。

【請求項8】

如請求項7之系統，其中該矩陣計算單元進一步經組態以將該控制信號自該脈動陣列中之一第一胞元傳遞至相鄰於該第一胞元之一第二胞元。

【請求項9】

一種用於執行類神經網路計算之方法，該方法包括：

藉由沿一矩陣計算單元之一第一維度之一第一群組胞元，接收權重輸入及激發輸入；

藉由該第一群組胞元，使用至少該等權重輸入及該等激發輸入以針對一類神經網路之一類神經網路層執行類神經網路計算之一部分；

藉由該第一群組胞元，使用經接收之該等權重輸入及該等激發輸入處理一累加輸出；

將跨一第二維度之該累加輸出提供至鄰近於該第一群組胞元之一第二群組胞元；及

使用跨該第一群組胞元及該第二群組胞元之至少該累加輸出產生一最終累加輸出。

【請求項10】

如請求項9之方法，其中該神經網路計算包括該等權重輸入之一各自權重輸入與該等激發輸入之一各自激發輸入相乘之乘法。

【請求項11】

如請求項9之方法，其中該累加輸出係儲存於複數個累加器單元之一各自累加器單元中，該各自累加器單元係連接至該第一群組胞元。

【請求項12】

如請求項9之方法，其中：

該第一群組胞元接收沿該第一維度之該等權重輸入及沿該第二維度之該等激發輸入；及

該第一維度係一列維度而該第二維度係一行維度，或者該第一維度係一行維度而該第二維度係一列維度。

【請求項13】

如請求項9之方法，其進一步包括藉由該矩陣計算單元使用至少該第一群組胞元而產生該等激發輸入。

【請求項14】

如請求項13之方法，其中：

該類神經網路層係一第一類神經網路；及

為產生該等激發輸入，該矩陣計算單元經組態以至少部分地執行該類神經網路中在該第一類神經網路之前的一第二類神經網路之類神經網路計算。

【請求項15】

如請求項9之方法，其中各胞元進一步包括一控制暫存器，其經組態以儲存一控制信號以用於判定是否將一權重輸入或一激發輸入移位至一相鄰胞元。

【請求項16】

如請求項15之方法，其進一步包括藉由該矩陣計算單元將該控制信號自該矩陣計算單元之一脈動陣列中之一第一胞元傳遞至相鄰於該第一胞元之一第二胞元。

【請求項17】

一或多個非暫時性電腦可讀儲存媒體，其在由包括一矩陣計算單元之一或多個處理器執行時編碼指令，該矩陣計算單元包含一多維脈動陣列，其包含一第一群組胞元及一第二群組胞元，該第二群組胞元鄰近於該第一群組胞元，該等指令引起該矩陣計算單元執行操作，該等操作包括：

藉由該第一群組胞元且沿該矩陣計算單元之一第一維度，接收權重輸入及激發輸入；

藉由該第一群組胞元，使用至少該等權重輸入及該等激發輸入以針對一類神經網路之一類神經網路層執行類神經網路計算之一部分；

藉由該第一群組胞元，使用經接收之該等權重輸入及該等激發輸入處理一累加輸出；

將跨一第二維度之該累加輸出提供至該第二群組胞元；及

使用跨該第一群組胞元及該第二群組胞元之至少該累加輸出產生一最終累加輸出。

【請求項18】

如請求項17之非暫時性電腦可讀儲存媒體，其中該類神經網路計算包括該等權重輸入之一各自權重輸入與該等激發輸入之一各自激發輸入相乘之乘法。

【請求項19】

如請求項17之非暫時性電腦可讀儲存媒體，其中該累加輸出係儲存於複數個累加器單元之一各自累加器單元中，該各自累加器單元係連接至該第一群組胞元。

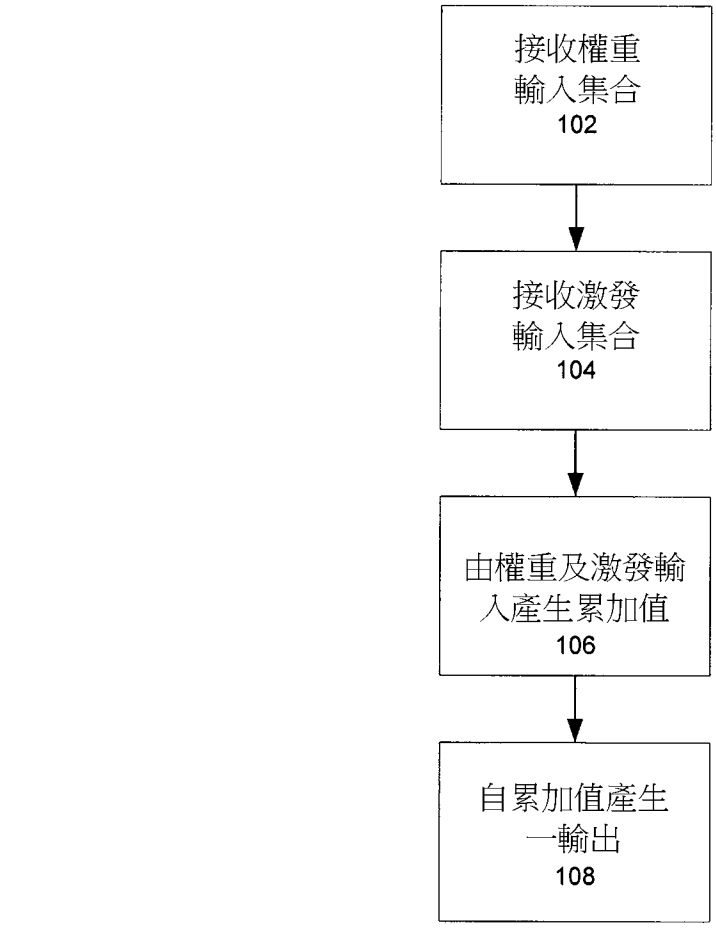
【請求項20】

如請求項17之非暫時性電腦可讀儲存媒體，其中：

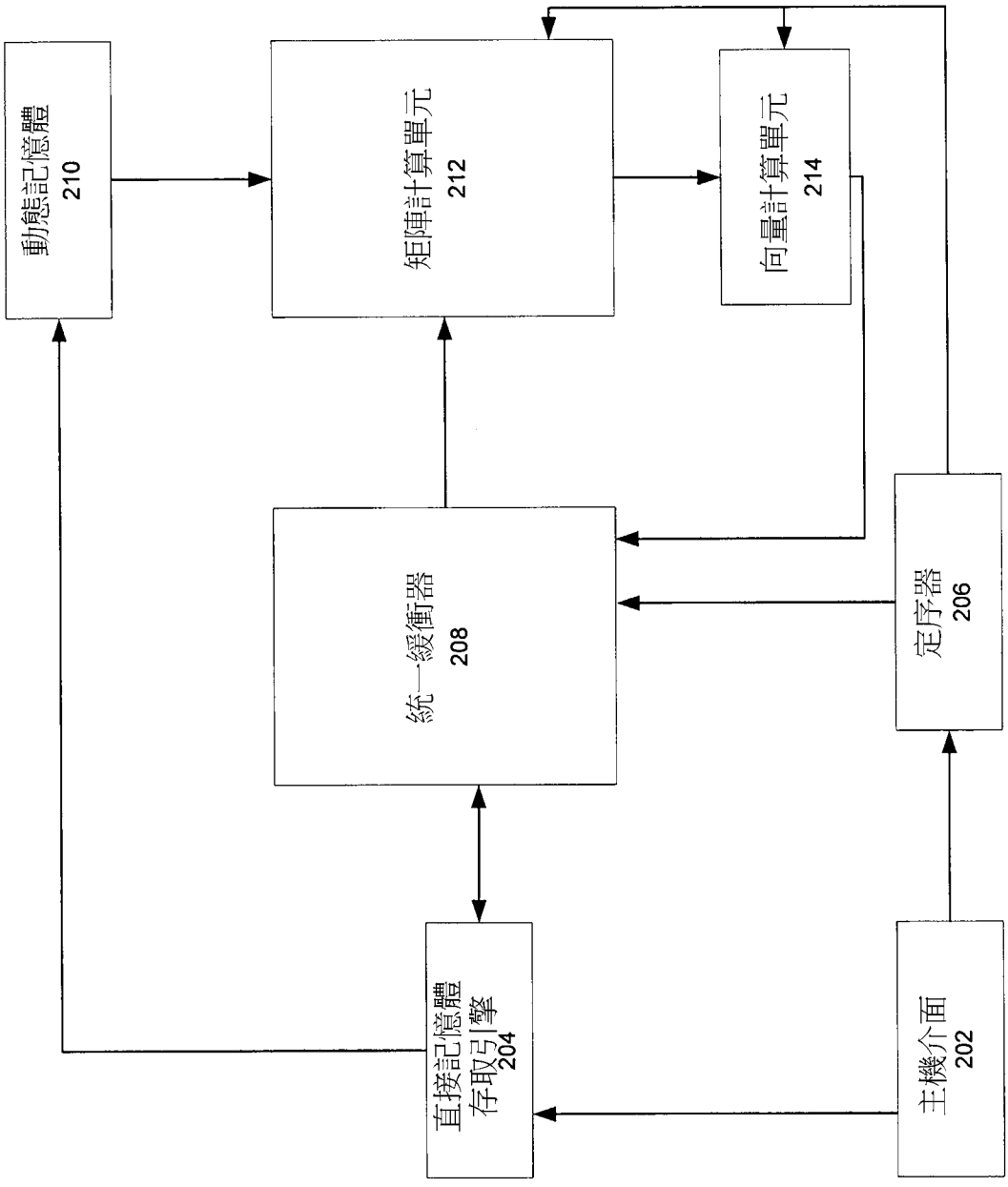
該第一群組胞元接收沿該第一維度之該等權重輸入及沿該第二維度之該等激發輸入；及

該第一維度係一列維度而該第二維度係一行維度，或者該第一維度係一行維度而該第二維度係一列維度。

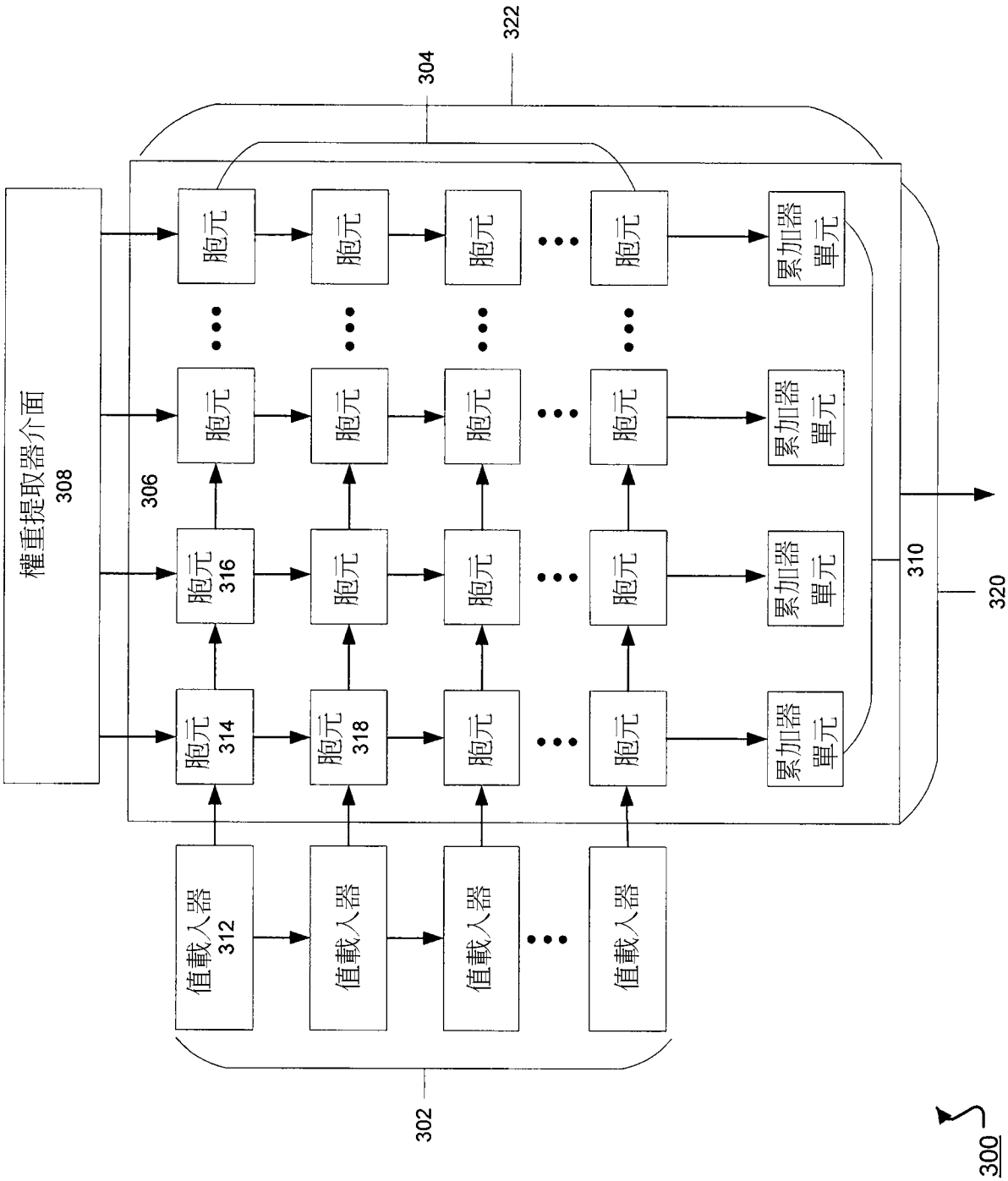
【發明圖式】


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

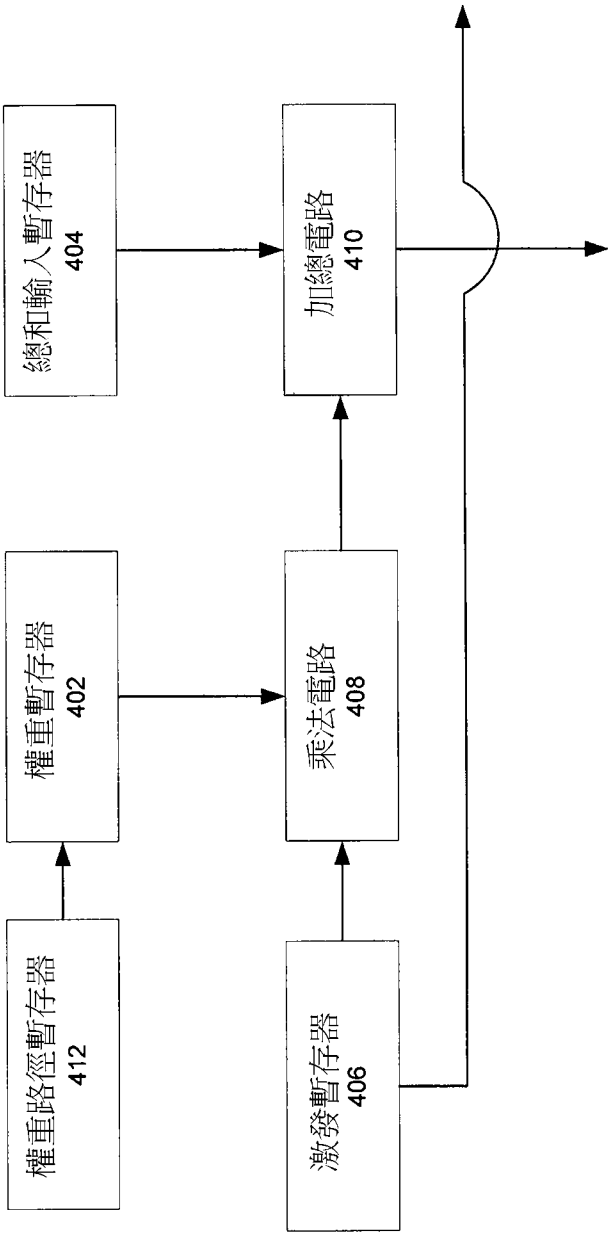
【圖 1】


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

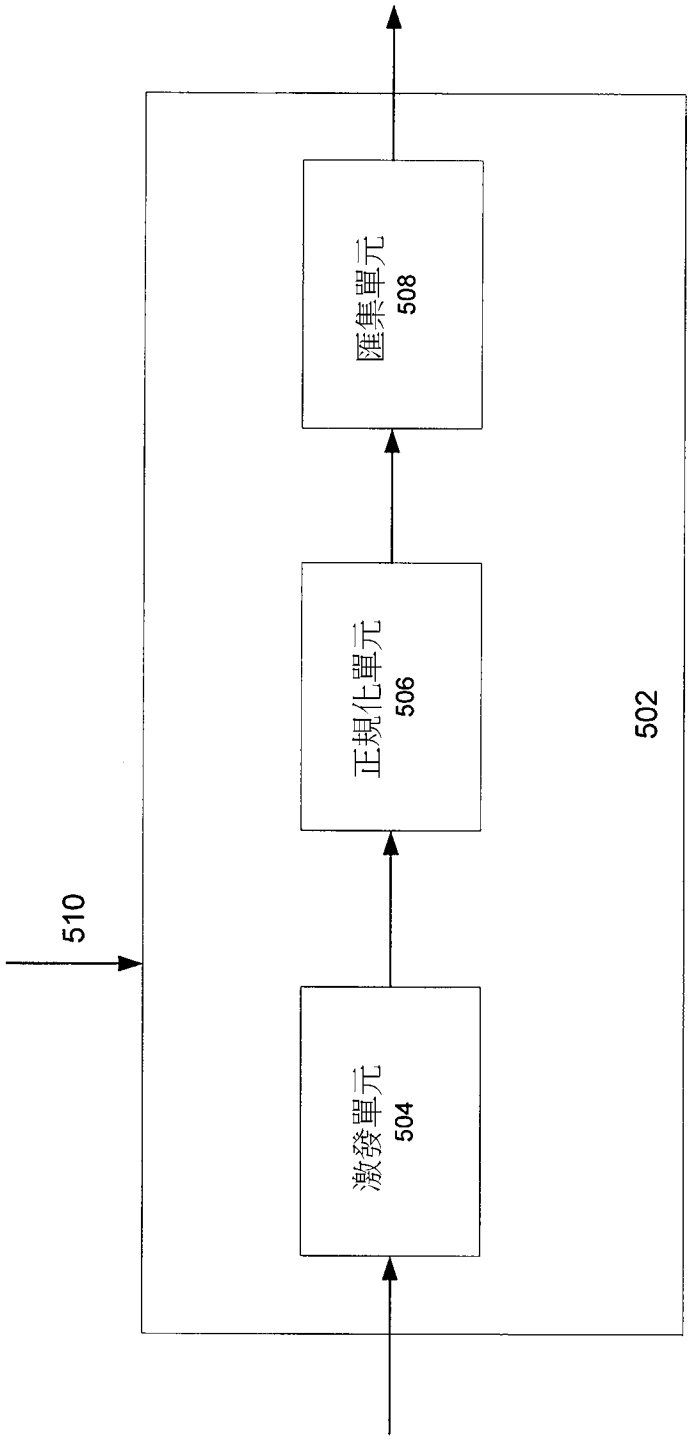
【圖 2】


權重提取器介面
308
306
胞元
胞元 316
胞元 314
胞元 318
累加器單元
值載入器 312
值載入器
304
322
302
310
320
300

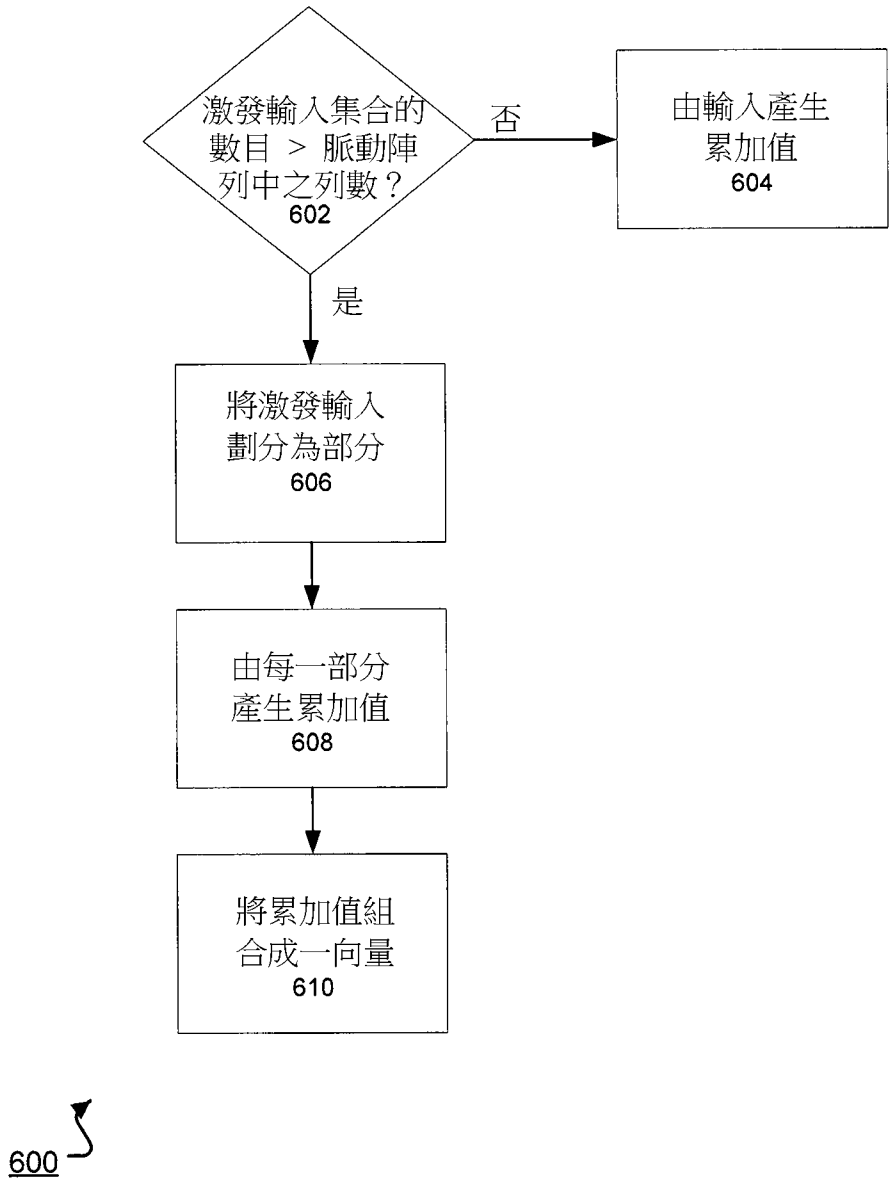
【圖 3】


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

【圖 4】



【圖 5】


激發輸入集合的數目 > 脈動陣列中之列數？
602
否
由輸入產生累加值
604
是
將激發輸入劃分為部分
606
由每一部分產生累加值
608
將累加值組合成一向量
610
600

【圖 6】