I765168

(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 I765168 B**

(45)公告日： 中華民國 111 (2022) 年 05 月 21 日

(21)申請案號：108127280 (22)申請日： 中華民國 107 (2018) 年 03 月 09 日

(51)Int. Cl.： ***G06N3/02 (2006.01)*** ***G06F17/16 (2006.01)***

(30)優先權：2017/03/09 美國 15/455,024

(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：楊 雷金納德 克里福德 YOUNG, REGINALD CLIFFORD (US)；厄文 傑佛瑞 IRVING, GEOFFREY (US)

(74)代理人：陳長文

(56)參考文獻：

CN	106022468A	US	5644517A
US	7395251B2	US	2006/0190517A1

審查人員：廖國智

申請專利範圍項數：15 項 圖式數：9 共 40 頁

(54)名稱

用於硬體中之轉置神經網路矩陣之方法、系統及電腦儲存媒體

(57)摘要

本發明揭示包含編碼於一電腦儲存媒體上之電腦程式之方法、系統及設備。在一個態樣中，一種方法包含以下動作：接收用以在具有一矩陣計算單元之一硬體電路上針對一神經網路執行計算之一請求，該請求指定待對一第一神經網路矩陣執行之一轉置運算；及產生在由該硬體電路執行時引起該硬體電路藉由執行第一操作而轉置該第一神經網路矩陣之指令，其中該等第一操作包含重複執行以下第二操作：針對將該第一神經網路矩陣劃分成一或多個當前子矩陣之該第一神經網路矩陣之一當前細分，藉由交換各當前子矩陣之一右上象限與一左下象限而更新該第一神經網路矩陣；及將各當前子矩陣細分成各自新子矩陣以更新該當前細分。

Methods, systems, and apparatus, including computer programs encoded on a computer storage medium. In one aspect, a method includes the actions of receiving a request to perform computations for a neural network on a hardware circuit having a matrix computation unit, the request specifying a transpose operation to be performed on a first neural network matrix; and generating instructions that when executed by the hardware circuit cause the hardware circuit to transpose the first neural network matrix by performing first operations, wherein the first operations include repeatedly performing the following second operations: for a current subdivision of the first neural network matrix that divides the first neural network matrix into one or more current submatrices, updating the first neural network matrix by swapping an upper right quadrant and a lower left quadrant of each current submatrix, and subdividing each current submatrix into respective new submatrices to update the current subdivision.

指定代表圖：

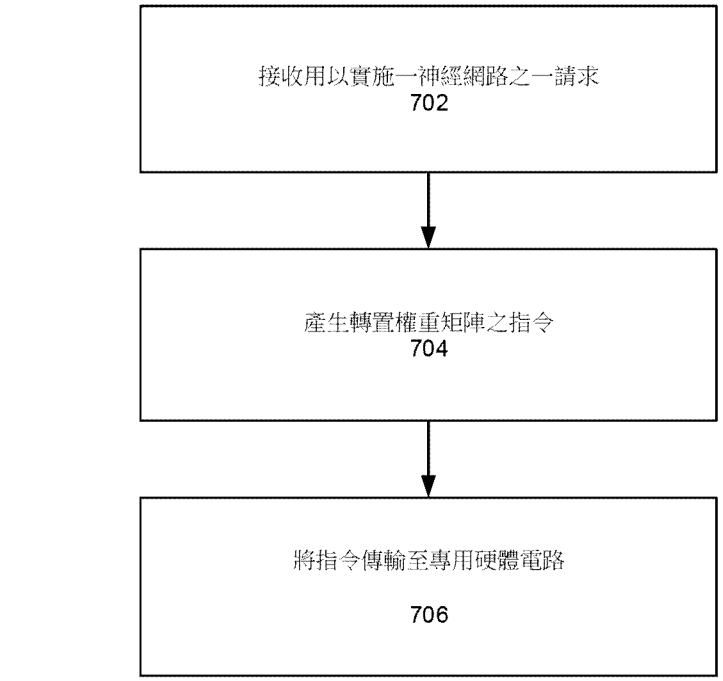
符號簡單說明：

700:程序

702:步驟

704:步驟

706:步驟



【圖7】

公告本

I765168

【發明摘要】

【中文發明名稱】

用於硬體中之轉置神經網路矩陣之方法、系統及電腦儲存媒體

【英文發明名稱】

METHOD, SYSTEM AND COMPUTER STORAGE MEDIUM FOR TRANSPOSING NEURAL NETWORK MATRICES IN HARDWARE

【中文】

本發明揭示包含編碼於一電腦儲存媒體上之電腦程式之方法、系統及設備。在一個態樣中，一種方法包含以下動作：接收用以在具有一矩陣計算單元之一硬體電路上針對一神經網路執行計算之一請求，該請求指定待對一第一神經網路矩陣執行之一轉置運算；及產生在由該硬體電路執行時引起該硬體電路藉由執行第一操作而轉置該第一神經網路矩陣之指令，其中該等第一操作包含重複執行以下第二操作：針對將該第一神經網路矩陣劃分成一或多個當前子矩陣之該第一神經網路矩陣之一當前細分，藉由交換各當前子矩陣之一右上象限與一左下象限而更新該第一神經網路矩陣；及將各當前子矩陣細分成各自新子矩陣以更新該當前細分。

【英文】

Methods, systems, and apparatus, including computer programs encoded on a computer storage medium. In one aspect, a method includes the actions of receiving a request to perform computations for a neural network on a hardware circuit having a matrix computation unit, the request specifying a transpose operation to be performed on a first neural network matrix; and generating instructions that when executed by the hardware circuit cause the hardware circuit to transpose the first

neural network matrix by performing first operations, wherein the first operations include repeatedly performing the following second operations: for a current subdivision of the first neural network matrix that divides the first neural network matrix into one or more current submatrices, updating the first neural network matrix by swapping an upper right quadrant and a lower left quadrant of each current submatrix, and subdividing each current submatrix into respective new submatrices to update the current subdivision.

【指定代表圖】

圖7

【代表圖之符號簡單說明】

700 程序

702 步驟

704 步驟

706 步驟

【發明說明書】

【中文發明名稱】

用於硬體中之轉置神經網路矩陣之方法、系統及電腦儲存媒體

【英文發明名稱】

METHOD, SYSTEM AND COMPUTER STORAGE MEDIUM FOR TRANSPOSING NEURAL NETWORK MATRICES IN HARDWARE

【技術領域】

【先前技術】

本說明書係關於硬體中之轉置神經網路矩陣。

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。除一輸出層之外，一些神經網路亦包含一或多個隱藏層。各隱藏層之輸出用作至網路中之另一層(例如，網路之下一隱藏層或輸出層)之輸入。網路之各層根據一各自參數集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神經網路推理之專用硬體電路。

本說明書中描述之標的之一個發明態樣可體現為包含以下動作之方法：接收用以在具有一矩陣計算單元之一硬體電路上針對一神經網路執行計算之一請求，該請求指定待對與該神經網路相關聯之一第一神經網路矩陣執行之一轉置運算；及產生在由該硬體電路執行時引起該硬體電路藉由執行第一操作而轉置該第一神經網路矩陣之指令，其中該等第一操作包含重複執行以下第二操作：針對將該第一神經網路矩陣劃分成一或多個當前子矩陣之該第一神經網路矩陣之一當前細分，藉由使用該矩陣計算單元交換該當前細分中之各當前子矩陣之一右上象限與一左下象限而更新該第一

神經網路矩陣，及將該當前細分中之各當前子矩陣細分成各自複數個新子矩陣以更新該當前細分，該各自複數個新子矩陣之各者係該當前子矩陣之一各自象限。

此態樣之其他實施例包含各自經組態以執行該等方法之該等動作的對應電腦系統、設備及記錄於一或多個電腦儲存裝置上之電腦程式。一或多個電腦之一系統可經組態以憑藉安裝於該系統上之在操作時可引起該系統執行該等動作之軟體、韌體、硬體或其等之任何組合來執行特定操作或動作。一或多個電腦程式可經組態以憑藉包含在由資料處理設備執行時引起該設備執行該等動作之指令來執行特定操作或動作。

此態樣之實施例可包含以下選用特徵之一或多者。在一些實施方案中，該等第一操作包含：判定該第一神經網路矩陣並非一i x i矩陣，其中i係該硬體電路之一向量長度值；作為回應，藉由在執行該等第二操作之全部迭代之前對該第一神經網路矩陣進行零填補而更新該第一神經網路矩陣以產生一i x i矩陣；及在執行該等第二操作之全部迭代之後，藉由移除在該更新期間填補之零而將該第一神經網路矩陣轉換成其在該更新之前的狀態。在一些實施方案中，該等第一操作進一步包含獲得指示該第一神經網路矩陣之一或多個值係零值之資料；且更新該第一神經網路矩陣包含：防止該矩陣計算單元對包含為零值之該第一神經網路矩陣之該一或多個值之至少一者的一值集執行任何操作。在一些實施方案中，交換該當前子矩陣之該右上象限與各當前子矩陣之該左下象限包含：將該第一神經網路矩陣之各列與一或多個部分特性矩陣相乘以產生各自包含該各自列之一部分之一或多個向量，其中每一各自當前子矩陣之該右上象限及該左下象限之元素經交換；針對該第一神經網路矩陣之各列，組合對應於該第一神經網路

矩陣之每一各自列之一部分之該等向量，其中每一各自當前子矩陣之該右上象限及該左下象限之該等元素經交換；及藉由組合每一各自列而產生經更新第一神經網路矩陣。在一些實施方案中，該矩陣計算單元執行一矩陣乘法運算作為一系列向量乘法運算。在一些實施方案中，該等第二操作進一步包含產生該第一神經網路矩陣之一初始當前細分，其中該初始當前細分含有為該第一神經網路矩陣之一初始子矩陣。在一些實施方案中，該等第一操作進一步包含將該等指令傳輸至該硬體電路。

本說明書中描述之標的之特定實施例可經實施以實現以下優點之一或多者。可藉由一專用硬體電路在硬體中對一矩陣執行一轉置運算，甚至在該硬體電路無法直接執行一矩陣轉置運算之情況下亦如此。藉由使用該專用硬體電路執行該轉置運算，可在未將資料傳回至一主機電腦之情況下(即，未在晶片外執行該計算之至少一部分之情況下)執行對一神經網路操作或指定一轉置運算之其他操作之處理，即使該專用硬體電路無法直接支援此處理。此容許在未修改該專用硬體電路之硬體架構之情況下對一矩陣執行一轉置運算。即，避免由在晶片外、在軟體中或兩者執行該處理之部分所致之處理延遲。

在下文隨附圖式及描述中闡述本說明書之標的之一或多項實施例之細節。自描述、圖式及發明申請專利範圍將明白標的之其他特徵、態樣及優點。

【圖式簡單說明】

圖1展示一例示性神經網路處理系統。

圖2係用於針對一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3展示一例示性神經網路處理系統。

圖4展示包含一矩陣計算單元之一例示性架構。

圖5展示一脈動陣列內部之一胞元之一例示性架構。

圖6展示一向量計算單元之一例示性架構。

圖7係用於實施指定對一矩陣進行之一轉置運算的一神經網路之一例示性程序之一流程圖。

圖8係用於使用一矩陣計算單元交換一子矩陣之右上象限與左下象限之一例示性程序之一流程圖。

圖9係用於對一神經網路矩陣執行一轉置運算之一計算之一實例。

在各個圖式中，相似元件符號及名稱指示相似元件。

【實施方式】

具有多個層之一神經網路可用於計算推理。例如，給定一輸入，神經網路可針對該輸入計算一推理。神經網路藉由透過神經網路之層之各者處理輸入而計算此推理。各層接收一輸入且根據該層之權重集處理該輸入以產生一輸出。

因此，為了自一經接收輸入計算一推理，神經網路接收輸入且透過神經網路層之各者處理該輸入以產生推理，其中來自一個神經網路層之輸出被提供為至下一神經網路層之輸入。至一神經網路層之資料輸入(例如，至神經網路之輸入或在序列上低於該層之層至一神經網路層之輸出)可稱為至該層之激發輸入(activation input)。

在一些實施方案中，神經網路之層以一序列配置。在一些其他實施方案中，該等層配置為有向圖。即，任何特定層可接收多個輸入、多個輸出或兩者。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展示一例示性神經網路處理系統100。神經網路處理系統100係實施為其中可實施下文描述之系統、組件及技術之一或多個位置中之一或多個電腦之一系統之一實例。

神經網路處理系統100係使用一專用硬體電路110執行神經網路計算之一系統。硬體電路110係用於執行神經網路計算之一積體電路且包含在硬體中執行向量-矩陣乘法之一矩陣計算單元120。下文參考圖3更詳細描述一例示性專用硬體電路120。

特定言之，神經網路處理系統100接收用以在專用硬體電路110上實施神經網路之請求、在專用硬體電路110上實施神經網路，且一旦實施一給定神經網路，便使用專用積體電路110處理至神經網路之輸入以產生神經網路推理。

即，神經網路處理系統100可接收指定待用於處理輸入之一神經網路的一神經網路架構之一請求。神經網路架構定義神經網路中之層之數目及組態以及具有參數之層之各者之參數值。

為在專用積體電路110上實施一神經網路，神經網路處理系統100包含一神經網路實施引擎150，其實施為一或多個實體位置中之一或多個電腦上之一或多個電腦程式。

神經網路實施引擎150產生在由專用硬體電路110執行時引起硬體電路110執行由神經網路指定之操作以自一經接收神經網路輸入產生一神經網路輸出之指令。

一旦指令已由神經網路實施引擎150產生且提供至硬體電路110，神經網路處理系統100便可接收神經網路輸入，且可藉由引起硬體電路110

執行所產生指令而使用神經網路處理神經網路輸入。一些神經網路指定對一神經網路矩陣(例如，包含神經網路之一層之權重值的一神經網路矩陣)進行之一轉置運算。例如，一些神經網路可指定對在其等前幾行中比其等在後續行中更密集(即，具有更多有意義的值)之矩陣進行之一轉置運算，以加速對此等矩陣之有意義值之處理。一些神經網路訓練演算法可需要轉置神經網路矩陣(例如，在反向傳播期間)。一些神經網路可需要將矩陣之轉置作為自卷積層(convolutional layer)至完全連接層(或反之亦然)之一過渡之部分。

在積體電路110上執行矩陣運算之主要硬體單元係矩陣計算單元120，其無法直接執行矩陣轉置運算。由於此，積體電路無法直接對一矩陣執行一轉置運算。為實施指定對一矩陣進行之一轉置運算之一神經網路，神經網路實施引擎150產生在由專用硬體電路110在藉由神經網路處理一神經網路輸入之期間執行時引起硬體電路110使用矩陣乘法單元120及向量計算單元140對一矩陣執行一矩陣轉置運算之指令。下文參考圖6至圖9更詳細描述此等指令及其他操作。

圖2係用於使用一專用硬體電路針對一神經網路之一給定層執行一計算之一例示性程序200之一流程圖。為方便起見，將關於具有執行方法200之一或多個電路之一系統描述方法200。可針對神經網路之各層執行方法200以自一經接收輸入計算一推理。

系統針對給定層接收權重輸入集(步驟202)及激發輸入集(步驟204)。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及激發輸入集。在一些實施方案中，可自統一緩衝器接收權重輸入集及激發輸入集兩者。

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206)。在一些實施方案中，累加值係權重輸入集與激發輸入集之點積。即，對於一個權重集(其係層中之全部權重之一子集)，系統可將各權重輸入與各激發輸入相乘並將乘積加總在一起以形成一累加值。接著，系統可計算其他權重集與其他激發輸入集之點積。

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208)。在一些實施方案中，向量計算單元將一激發函數應用於累加值，此將在下文參考圖5進一步描述。層之輸出可儲存於統一緩衝器中以用作至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神經網路之各層處理以針對經接收輸入產生推理時，系統完成處理神經網路。

圖3展示用於執行神經網路計算之一例示性專用硬體電路300。系統300包含一主機介面302。主機介面302可接收包含用於一神經網路計算之參數之指令。參數可包含以下之一或多者：應處理之層之數目、用於模型之各層之對應權重輸入集、一初始激發輸入集(即，將自其計算推理之至神經網路之輸入)、各層之對應輸入及輸出大小、用於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302可將指令發送至一定序器306，該定序器306將指令轉換成控制電路以執行神經網路計算之低階控制信號。在一些實施方案中，控制信號調節電路中之資料流(例如，權重輸入集及激發輸入集如何流動通過電路)。定序器306可將控制信號發送至一統一緩衝器308、一矩陣計算單元312及一向量計算單元314。在一些實施方案中，定序器306亦將控制信號發送至一直接記憶體存取引擎304及動態記憶體310。在一些實施方

案中，定序器306係產生控制信號之一處理器。定序器306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300之各組件。在一些其他實施方案中，主機介面302傳入來自一外部處理器之一控制信號。

主機介面302可將權重輸入集及初始激發輸入集發送至直接記憶體存取引擎304。直接記憶體存取引擎304可將激發輸入集儲存在統一緩衝器308處。在一些實施方案中，直接記憶體存取將權重集儲存至動態記憶體310，該動態記憶體310可為一記憶體單元。在一些實施方案中，動態記憶體310定位成遠離電路。

統一緩衝器308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304之激發輸入集及向量計算單元314之輸出。下文將參考圖6更詳細描述向量計算單元314。直接記憶體存取引擎304亦可自統一緩衝器308讀取向量計算單元314之輸出。

動態記憶體310及統一緩衝器308可分別將權重輸入集及激發輸入集發送至矩陣計算單元312。在一些實施方案中，矩陣計算單元312係一個二維脈動陣列。矩陣計算單元312亦可為一個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312係一通用矩陣處理器。專用硬體電路300可使用矩陣計算單元312來執行一矩陣轉置運算。下文參考圖8至圖10更詳細描述使用矩陣計算單元312執行一矩陣轉置運算。

矩陣計算單元312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314。在一些實施方案中，矩陣計算單元312將輸出之向量發送至統一緩衝器308，該統一緩衝器308將輸出之向量發送至向量計算單元314。向量計算單元314可處理輸出之向量且將經處理輸出之一向

量儲存至統一緩衝器308。經處理輸出之向量可用作至矩陣計算單元312之激發輸入(如，用於神經網路中之一後續層)。下文將分別參考圖4及圖6更詳細描述矩陣計算單元312及向量計算單元314。

圖4展示包含一矩陣計算單元之一例示性架構400。矩陣計算單元係一個二維脈動陣列406。陣列406包含多個胞元404。在一些實施方案中，脈動陣列406之一第一維度420對應於胞元之行，且脈動陣列406之一第二維度422對應於胞元之列。脈動陣列可具有比行更多的列、比列更多的行或相同數目個行及列。

在所繪示實例中，值載入器402將激發輸入發送至陣列406之列，且一權重提取器介面408將權重輸入發送至陣列406之行。然而，在一些其他實施方案中，將激發輸入傳送至陣列406之行且將權重輸入傳送至陣列406之列。

值載入器402可自一統一緩衝器(例如，圖3之統一緩衝器308)接收激發輸入。各值載入器可將一對應激發輸入發送至陣列406之一相異最左側胞元。例如，值載入器412可將一激發輸入發送至胞元414。

權重提取器介面408可自一記憶體單元(例如，圖3之動態記憶體310)接收權重輸入。權重提取器介面408可將一對應權重輸入發送至陣列406之一相異最頂部胞元。例如，權重提取器介面408可將權重輸入發送至胞元414及416。權重提取器介面408進一步能夠自記憶體單元(例如，動態記憶體310)接收多個權重且能夠將多個權重並行發送至陣列406之相異最頂部胞元。例如，權重提取器介面408可同時將不同權重發送至胞元414及416。

在一些實施方案中，一主機介面(例如，圖3之主機介面302)使激發

輸入沿一個維度移位(例如，向右移位)貫穿陣列406，而使權重輸入沿另一維度移位(例如，向底部移位)貫穿陣列406。例如，在一個時脈循環內，胞元414處之激發輸入可移位至胞元416 (其在胞元414右側)中之一激發暫存器。類似地，胞元416處之權重輸入可移位至胞元418 (其在胞元414下方)處之一權重暫存器。

在各時脈循環，各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。亦可將累加輸出傳遞至沿與給定權重輸入相同之維度之相鄰胞元。各胞元亦可處理一給定權重輸入及一給定激發輸入以產生一輸出，而未處理來自一相鄰胞元之一累加輸出。可將輸出傳遞至沿與給定權重輸入及未累加之輸出相同之維度之相鄰胞元。下文參考圖5進一步描述一個別胞元。

可沿與權重輸入相同之行(例如，朝向陣列406中之行之底部)傳遞累加輸出。在一些實施方案中，在各行之底部處，陣列406可包含累加器單元410，該等累加器單元410在運用具有比列更多之激發輸入的層執行計算時儲存並累加來自各行之各累加輸出。在一些實施方案中，各累加器單元儲存多個並行累加。累加器單元410可累加各累加輸出以產生一最終累加值。可將最終累加值傳送至一向量計算單元(例如，圖6之向量計算單元)。在一些其他實施方案中，累加器單元410將累加值傳遞至向量計算單元，而未在處理其中層具有比列更少之激發輸入之層時執行任何累加。

圖5展示一脈動陣列(例如，圖4之脈動陣列406)內部之一胞元之一例示性架構500。

胞元可包含儲存一激發輸入之一激發暫存器506。激發暫存器可取決於胞元在脈動陣列內之位置而自一左側相鄰胞元(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502。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(sum in)暫存器504。總和輸入暫存器504可儲存來自頂部相鄰胞元之一累加值。乘法電路508可用於將來自權重暫存器502之權重輸入與來自激發暫存器506之激發輸入相乘。乘法電路508可將乘積輸出至加總電路510。

加總電路510可加總乘積與來自總和輸入暫存器504之累加值以產生一新累加值。接著，加總電路510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。新累加值可用作用於底部相鄰胞元中之一加總之一運算元。加總電路510亦可接受來自總和輸入暫存器504之一值並將來自總和輸入暫存器504之值發送至一底部相鄰胞元，而未加總來自總和輸入暫存器504之值與來自乘法電路508之乘積。

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，權重路徑暫存器512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506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藉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。控制暫存器可儲存判定胞元是否應將權重輸入或激發輸入移位至相鄰胞元之一控制信號。在一些實施方案中，移位權重輸入或激發輸入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508，或可判定乘法電路508是否對激發輸入及權重輸入操作。亦可例如使用一電線將控

制信號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512中。權重路徑暫存器512可例如自一頂部相鄰胞元接收權重輸入，且基於控制信號將權重輸入傳送至權重暫存器502。權重暫存器502可靜態地儲存權重輸入，使得在多個時脈循環內，在激發輸入例如透過激發暫存器506傳送至胞元時，權重輸入保持在胞元內且未被傳送至一相鄰胞元。因此，可例如使用乘法電路508來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6展示一向量計算單元602之一例示性架構600。向量計算單元602可自一矩陣計算單元(例如，參考圖3描述之矩陣計算單元312或圖4之矩陣計算單元之累加器410)接收累加值之一向量。

向量計算單元602可在激發電路604處處理累加值之向量。在一些實施方案中，激發電路包含將一非線性函數應用於各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。

視情況，向量計算單元602可使用匯集電路(pooling circuitry) 608匯集值(例如，激發值)。匯集電路608可將一彙總函數應用於該等值之一或多者以產生匯集值。在一些實施方案中，彙總函數係返回該等值或該等值之一子集之一最大值、最小值或平均值之函數。

控制信號610可例如藉由圖3之定序器306傳送，且可調節向量計算單元602如何處理累加值之向量。即，控制信號610可調節是否匯集激發值、將激發值儲存於何處(例如，儲存於統一緩衝器308中)，或可以其他方式調節激發值之處置。控制信號610亦可指定激發或匯集函數以及用於處理激發值或匯集值之其他參數(例如，一步幅值)。

向量計算單元602可將值(例如，激發值或匯集值)發送至一統一緩衝器(例如，圖3之統一緩衝器308)。在一些實施方案中，匯集電路608接收激發值或匯集值並將激發值或匯集值儲存於統一緩衝器中。

圖7係用於實施指定對一矩陣進行之一轉置運算的一神經網路之一例示性程序700之一流程圖。一般而言，藉由包含一專用硬體電路(例如，圖1之專用硬體電路110)之一或多個電腦之一系統執行程序700。

系統接收用以在專用硬體電路上實施一神經網路之一請求(步驟702)。特定言之，神經網路包含若干神經網路矩陣，且指定對神經網路矩陣之一第一神經網路矩陣進行之一轉置運算。

系統產生在由專用硬體電路執行時引起專用硬體電路轉置第一神經網路矩陣之指令(步驟704)。指令引起專用硬體電路藉由在各迭代期間更新矩陣之一當前細分之各子矩陣而迭代地轉置矩陣。更新當前細分之各子矩陣包含：使用專用硬體電路中之一矩陣計算單元來交換當前子矩陣之一右上象限與子矩陣之一左下象限。下文參考圖8更詳細描述在各迭代期間更新一當前細分之子矩陣。

一矩陣之一細分係將矩陣劃分成一或多個子矩陣。在各迭代處，指令引起專用硬體電路將矩陣劃分成一或多個(例如，四個)子矩陣以產生矩陣之一當前細分。例如，在第一迭代處，指令引起專用硬體電路110產生包含僅一個當前子矩陣之一初始當前細分。換言之，第一迭代之當前子矩陣包含整個第一神經網路矩陣作為唯一的(one and only)子矩陣。在各後續迭代處，指令引起專用硬體電路藉由將當前細分中之各細分劃分成一或多個(例如，四個)細分而產生一經更新當前細分。

在一些實施方案中，第一神經網路矩陣係一$2^i * 2^i$矩陣，其中i係一

非負整數，且更新矩陣包含：在各迭代處將第一神經網路矩陣劃分成大小為$2^j * 2^j$之子矩陣，且藉由並非垂直或水平鄰近於特定子矩陣而是對角鄰近於特定子矩陣之一個對應子矩陣交換各特定子矩陣。在一些該等實施方案中，j之值在第一迭代中係$(i-1)$且在各迭代中遞減。

迭代繼續，直至當前細分之一子矩陣係第一神經網路矩陣內之一單一值。此時，因為一單一值無法再細分成進一步子矩陣，所以迭代終止。

在一些實施方案中，系統執行矩陣乘法運算作為對具有一最大向量長度之向量進行之向量乘法之一組合。最大向量長度係在一遍次中(即，在未將向量劃分成至矩陣計算單元之多個輸入之情況下)可藉由矩陣計算單元而與一矩陣相乘之一向量之最大長度。例如，若矩陣計算單元係一個一維或二維脈動陣列，則最大向量長度等於單元中之行之數目或單元中之列之數目。

在一些該等實施方案中，系統獲得指示已將零值添加至神經網路矩陣以調整矩陣之尺寸使得可將矩陣劃分成具有最大向量長度之向量之資訊。換言之，已對神經網路矩陣進行零填補以適應系統之架構組態。回應於該資訊，系統可避免執行涉及被識別為已因零填補而添加之值的值乘值乘法運算，此係因為此等運算始終返回零值。因此，系統可減少執行此等向量乘法所需之值乘值乘法運算之數目。

系統將指令傳輸至專用硬體電路(步驟706)。

例如，神經網路實施引擎150可將指令提供至專用硬體電路110，且專用硬體電路110可例如在圖3之主機介面302處接收指令。神經網路實施引擎150亦可提供用於神經網路計算之其他指令及/或參數，其等亦可由主機介面302接收。

圖8係用於使用一專用硬體電路更新一神經網路矩陣之一當前細分之一子矩陣的一例示性程序800之一流程圖。例如，可藉由圖1之專用硬體電路110基於自神經網路實施引擎150接收之指令執行程序800。專用硬體電路藉由使用專用硬體電路110中之一矩陣計算單元交換子矩陣之一右上象限與子矩陣之一左下象限而更新一當前細分之一子矩陣。

專用硬體電路110針對神經網路矩陣之各列產生一向量(802)。

專用硬體電路110針對電路110企圖產生之經交換子矩陣之各值獲得一部分特性矩陣(804)。電路110可使用相同部分特性矩陣來產生經交換子矩陣之兩個或更多個值。

一部分特性矩陣係僅包含「0」及「1」值之一矩陣。一部分特性矩陣中之「1」值經有策略地定位，使得當與包含第一神經網路矩陣之一列中之值之一向量相乘時，乘法之輸出保留向量之特定值而零化(nullify)其他值(即，輸出「0」)。

在一些實施方案中，若含有神經網路矩陣之一列之值的向量具有一尺寸d，則在與該向量相乘時分別在一合成向量之j值及$(j + 1)$值中返回向量之i值及$(i + 1)$值的一部分特性矩陣係在$[i, j]$及$[i+1, j+1]$位置中具有1值且在別處具有零之一$d * d$矩陣。

專用硬體電路110將神經網路矩陣之各列與一或多個部分特性矩陣相乘，以自列獲得更新神經網路矩陣所需之值而交換當前細分中之各子矩陣之右上象限與左下象限(806)。

例如，向量$V_1 = [A\ B]$可包含一神經網路矩陣之第一列之兩個值。為擷取向量之第一值，專用硬體電路將V_1與以下部分特性矩陣I_1相乘：

$$\begin{matrix} 1 & 0 \\ 0 & 0 \end{matrix}$$

輸出為$V_1 * I_1 = [A\ 0]$。因此，保留A之值而零化B之值。

專用硬體電路110組合含有經更新神經網路矩陣之各列之部分的向量以產生該列(808)。例如，V_1可為一神經網路矩陣M_1之第一列：

$$\begin{matrix} A & B \\ C & D \end{matrix}$$

對應於矩陣M_1之經更新矩陣S_1之第一列將包含V_1之第一元素及向量$V_2 = [C\ D]$之第一元素，該向量V_2包含矩陣M_1之第二列之值。

換言之，含有經更新矩陣S_1之第一列之部分的兩個向量係$V_1 * I_1$之輸出及$V_2 * I_1$之輸出。專用硬體電路110可組合該兩個向量以產生經交換子矩陣S_1之第一列。

專用硬體電路110組合神經網路矩陣之各列以產生經更新神經網路矩陣(810)。

因為可使用矩陣乘法單元(例如，使用一系列矩陣-向量乘法及加法，如下文進一步描述)來執行交換各子矩陣之右上象限與左下象限，所以專用硬體電路110可在不具備一直接矩陣轉置運算能力之情況下對一矩陣執行一轉置運算。因而，神經網路實施引擎150可使用硬體電路110來處理指定對一矩陣之轉置之一不相容層。

圖9係用於對一神經網路矩陣執行一轉置運算之一計算之一實例。可使用圖7之程序及圖2之專用硬體電路300來執行圖9之實例。

在圖9之部分(a)中，專用硬體電路藉由產生包含整個矩陣之一子矩陣而形成一4 x 4矩陣之一當前細分。電路產生包含神經網路矩陣之各列之值的一向量。例如，電路產生包含矩陣之第一列之值的一向量Input[0]。

圖9之部分(b)描繪四個部分特性矩陣。各部分特性矩陣具有由矩陣內之「1」值之位置定義之一結構。各部分特性矩陣之結構經有戰略地設計

以自部分(a)中展示之向量擷取特定值而零化該等向量中之其他值。例如，專用硬體電路使用部分特性矩陣W1來擷取一向量之第一值及第二值。

在圖9之部分(c)中，專用硬體電路使用部分(a)中描繪之向量及部分(b)中描繪之部分特性矩陣執行四組計算。電路使用各組計算產生一經更新神經網路矩陣之一列，其包含部分(a)中描繪之神經網路矩陣之元素但其中各子矩陣之右上象限及左下象限經交換。例如，電路使用第一組計算產生[A B I J]，其係神經網路子矩陣之第一列，其中各子矩陣之右上象限及左下象限經交換。

圖9之部分(d)描繪藉由交換神經網路矩陣中之各子矩陣之右上象限與左下象限而更新部分(a)中描繪之神經網路矩陣之輸出。

在部分(e)中，專用硬體電路將部分(d)中描繪之經更新神經網路矩陣劃分成列。

圖9之部分(f)描繪四個部分特性矩陣。部分(f)中描繪之各部分特性矩陣之結構經有戰略地設計以自部分(d)中展示之向量擷取特定值而零化該等向量中之其他值。

在部分(g)中，專用硬體電路使用部分(e)中描繪之向量及部分(f)中描繪之部分特性矩陣執行四組計算。當對部分(e)中描繪之神經網路矩陣執行計算時，該等計算導致更新部分(e)中描繪之神經網路矩陣以在將神經網路矩陣新細分成四個子矩陣時交換各子矩陣之右上象限與左下象限。部分(h)中展示之經更新矩陣係部分(a)中展示之矩陣之一轉置。

在圖9之部分(d)至(g)中執行之操作係在圖9之部分(a)至(c)中執行之操作之一重複。但在部分(g)之後，由部分(e)中描繪之子矩陣形成之新子

矩陣係無法再進一步細分成象限之單一值。因此，將不再重複操作。

本說明書中描述之標的及功能操作之實施例可在以下各者中實施：數位電子電路、有形體現之電腦軟體或韌體、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。本說明書中描述之標的之實施例可實施為一或多個電腦程式，即，編碼於一有形非暫時性程式載體上以由資料處理設備執行或控制資料處理設備之操作的電腦程式指令之一或多個模組。替代地或額外地，可將程式指令編碼於一人工產生的傳播信號(例如，一機器產生之電信號、光學信號或電磁信號)上，該傳播信號經產生以編碼用於傳輸至適合接收器設備之資訊以由一資料處理設備執行。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各種設備、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。設備可包含專用邏輯電路，例如，一FPGA (場可程式化閘陣列)或一ASIC (特定應用積體電路)。除硬體之外，設備亦可包含產生所述電腦程式之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(其亦可被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)可以任何形式之程式設計語言(包含編譯或解譯語言或宣告或程序語言)寫入，且其可部署為任何形式，包含作為一獨立程式或作為一模組、組件、副常式或適用於一計算環境中之其他單元。一電腦程式可(但無需)對應於一檔案系統中之一檔案。一程式可儲存於保存其他程式或資料(例如，儲存於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、儲存於專用於所述程式之一單一檔案中，或儲存於多個協同檔案(例如，儲存一或多個模組、子程式或程式碼之部分之檔案)中。一電腦程式可經部署以在一個電腦上或在定位於一個地點處或跨多個地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一或多個可程式化電腦執行，該一或多個可程式化電腦執行一或多個電腦程式以藉由對輸入資料進行操作且產生輸出而執行功能。亦可藉由專用邏輯電路(例如，一FPGA(場可程式化閘陣列)或一ASIC (特定應用積體電路))執行程序及邏輯流程，且亦可將設備實施為專用邏輯電路。

適於執行一電腦程式之電腦包含例如通用或專用微處理器或兩者或任何其他種類之中央處理單元，或其可基於通用或專用微處理器或兩者或任何其他種類之中央處理單元。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。一電腦之基本元件係用於執行(perform/execute)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，一電腦可嵌入於另一裝置中，該裝置例如(僅列舉幾個)一行動電話、一個人數位助理(PDA)、一行動音訊或視訊播放機、一遊戲控制台、一全球定位系統(GPS)接收器或一可攜式儲存裝置(例如，一通用串列匯流排(USB)快閃隨身碟)。

適於儲存電腦程式指令及資料之電腦可讀媒體包含全部形式之非揮

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磁碟。處理器及記憶體可由專用邏輯電路補充或併入於專用邏輯電路中。

為發送與一使用者之互動，本說明書中描述之標的之實施例可在一電腦上實施，該電腦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)以及一鍵盤及一指向裝置，例如，一滑鼠或一軌跡球，使用者可藉由其發送輸入至電腦。其他種類之裝置亦可用於發送與一使用者之互動；例如，提供至使用者之回饋可為任何形式之感官回饋，例如，視覺回饋、聽覺回饋或觸覺回饋；且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。另外，一電腦可藉由將文件發送至供一使用者使用之一裝置且自該裝置接收文件(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而將網頁發送至該網頁瀏覽器)而與使用者互動。

可在一計算系統中實施本說明書中描述之標的之實施例，該計算系統包含一後端組件(例如，作為一資料伺服器)，或包含一中介軟體組件(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網頁瀏覽器之一用戶端電腦，一使用者可透過其與本說明書中描述之標的之一實施方案互動)或一或多個此後端組件、中介軟體組件或前端組件之任何組合。系統之組件可由數位資料通信之任何形式或媒體(例如，一通信網路)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。一用戶端及伺服器一般彼此遠離

且通常透過一通信網路互動。用戶端與伺服器之關係憑藉在各自電腦上運行且彼此具有一用戶端-伺服器關係之電腦程式引起。

雖然本說明書含有許多具體實施方案細節，但此等細節不應被解釋為限制任何發明或可主張之內容之範疇，而是應解釋為描述可為特定發明之特定實施例所特有之特徵。本說明書中在單獨實施例之內容背景中描述之某些特徵亦可在一單一實施例中組合實施。相反地，在一單一實施例之內容背景中描述之各種特徵亦可單獨地或以任何適合子組合在多個實施例中實施。此外，儘管上文可將特徵描述為以特定組合起作用且即使最初如此主張，然在一些情況中，來自一所主張組合之一或多個特徵可自組合中切除，且所主張組合可關於一子組合或一子組合之變動。

類似地，雖然在圖式中按一特定順序描繪操作，但此不應被理解為要求按所展示之特定順序或循序順序執行此等操作或執行全部所繪示操作以達成期望結果。在某些境況中，多任務處理及並行處理可為有利的。此外，上文描述之實施例中之各種系統模組及組件之分離不應被理解為在全部實施例中皆需要此分離，且應瞭解，所描述程式組件及系統一般可一起整合於一單一軟體產品中或封裝至多個軟體產品中。

已描述標的之特定實施例。其他實施例在以下發明申請專利範圍之範疇內。例如，在發明申請專利範圍中敘述之動作可按一不同順序執行且仍達成期望結果。作為一個實例，在附圖中描繪之程序不一定需要所展示之特定順序或連續順序以達成期望結果。在特定實施方案中，多任務處理及並行處理可為有利的。

【符號說明】

100　　神經網路處理系統

110	專用硬體電路/專用積體電路
120	矩陣計算單元/矩陣乘法單元
140	向量計算單元
150	神經網路實施引擎
200	程序/方法
202	步驟
204	步驟
206	步驟
208	步驟
300	專用硬體電路/系統
302	主機介面
304	直接記憶體存取引擎
306	定序器
308	統一緩衝器
310	動態記憶體
312	矩陣計算單元
314	向量計算單元
400	架構
402	值載入器
404	胞元
406	二維脈動陣列
408	權重提取器介面
410	累加器單元/累加器

412 值載入器

414 胞元

416 胞元

418 胞元

420 脈動陣列之第一維度

422 脈動陣列之第二維度

500 架構

502 權重暫存器

504 總和輸入暫存器

506 激發暫存器

508 乘法電路

510 加總電路

512 權重路徑暫存器

600 架構

602 向量計算單元

604 激發單元

608 匯集電路

610 控制信號

700 程序

702 步驟

704 步驟

706 步驟

800 程序

802 針對神經網路矩陣之各列產生向量

804 獲得部分特性矩陣

806 將神經網路矩陣之各列與一或多個部分特性矩陣相乘，以自列獲得更新神經網路矩陣所需之值而交換當前細分中之各子矩陣之右上象限與左下象限

808 組合含有經更新神經網路矩陣之各列之部分的向量以產生該列

810 組合神經網路矩陣之各列以產生經更新神經網路矩陣

【發明申請專利範圍】

【第1項】

一種用於轉置(transposing)一硬體電路上之神經網路矩陣之方法，該方法包括：

重複地執行以下操作：

將作為一迭代(as of an iteration)之一中間神經網路矩陣(intermediate neural network matrix)劃分成一當前細分(subdivision)中之複數個子矩陣(submatrices)；

產生複數個向量，該等向量之各者對應於該中間神經網路矩陣之各列(row)且包含該中間神經網路矩陣之各列之值；

針對該複數個向量獲得複數個部分特性矩陣(partial identity matrices)，該複數個部分特性矩陣經組態以自該複數個向量之各者擷取一部分值同時零化(nullifying)一剩餘部分數值(remaining portion of values)；

藉由該複數個部分特性矩陣之一或多者將該複數個向量之各者相乘以產生一經更新神經網路矩陣之一列，其中該經更新神經網路矩陣包含該神經網路矩陣之元素(elements)，但調換(swapped)該當前細分中之該複數個子矩陣之右上象限與左下象限；及

組合該等經產生之列以更新該中間神經網路矩陣；及

基於所完成之該等操作之全部迭代，產生該中間神經網路矩陣作為一轉置神經網路矩陣(transposed neural network matrix)。

【第2項】

如請求項1之方法，其進一步包括：

接收針對在該硬體電路上之該神經網路執行計算(computation)之一請求，該請求指定待執行於該神經網路矩陣上之一轉置運算(transpose operation)。

【第3項】

如請求項1之方法，其進一步包括：

判定該神經網路矩陣並非一i x i矩陣，其中i係該硬體電路之一向量長度值；

作為回應，藉由對該神經網路矩陣進行零填補(zero-padding)而更新該神經網路矩陣以產生一i x i矩陣；及

藉由移除在該轉置期間所填補之零，在該轉置之前將該轉置神經網路矩陣轉換成其狀態(condition)。

【第4項】

如請求項1之方法，其進一步包括：

獲得指示該神經網路矩陣之一或多個值係零值之資料；及

防止該硬體電路在一值集合(set of values)上執行任何操作，該值集合包含為零值之該神經網路矩陣之該一或多個值之至少一者。

【第5項】

如請求項1之方法，其中該硬體電路執行一矩陣乘法運算(matrix multiplication operations)作為一系列向量乘法運算。

【第6項】

如請求項2之方法，其進一步包括將該請求傳輸至該硬體電路。

【第7項】

一種包括一或多個電腦及一或多個儲存裝置之系統，該一或多個儲

存裝置儲存當由該一或多個電腦執行時可操作以使得該一或多個電腦執行多個操作之指令，該等操作包括：

在具有一矩陣計算單元之一硬體電路上重複地執行以下操作：

將作為一迭代之一中間神經網路矩陣劃分成一當前細分中之複數個子矩陣；

產生複數個向量，該等向量之各者對應於該中間神經網路矩陣之各列且包含該中間神經網路矩陣之各列之值；

針對該複數個向量獲得複數個部分特性矩陣，該複數個部分特性矩陣經組態以自該複數個向量之各者擷取一部分值同時零化(nullifying)一剩餘部分數值；

藉由該複數個部分特性矩陣之一或多者將該複數個向量之各者相乘以產生一經更新神經網路矩陣之一列，其中該經更新神經網路矩陣包含該神經網路矩陣之元素，但調換該當前細分中之該複數個子矩陣之右上象限與左下象限；及

組合該等經產生之列以更新該中間神經網路矩陣；及

基於所完成之該等操作之全部迭代，產生該中間神經網路矩陣作為一轉置神經網路矩陣。

【第8項】

如請求項7之系統，其中該一或多個電腦執行多個操作進一步包括：

接收針對在該硬體電路上之該神經網路執行計算之一請求，該請求指定待執行於該神經網路矩陣上之一轉置運算。

【第9項】

如請求項7之系統，其中該一或多個電腦執行多個操作進一步包括：

判定該神經網路矩陣並非一i x i矩陣，其中i係該硬體電路之一向量長度值；

作為回應，藉由對該神經網路矩陣進行零填補而更新該神經網路矩陣以產生一i x i矩陣；及

藉由移除在該轉置期間所填補之零，在該轉置之前將該轉置神經網路矩陣轉換成其狀態。

【第10項】

如請求項7之系統，其中該一或多個電腦執行多個操作進一步包括：

獲得指示該神經網路矩陣之一或多個值係零值之資料；及

防止該硬體電路在一值集合上執行任何操作，該值集合包含為零值之該神經網路矩陣之該一或多個值之至少一者。

【第11項】

如請求項7之系統，其中該一或多個電腦執行一矩陣乘法運算作為一系列向量乘法運算。

【第12項】

如請求項7之系統，其中該一或多個電腦執行多個操作進一步包括將該請求傳輸至該硬體電路。

【第13項】

一種編碼有指令之電腦儲存媒體，該等指令在由一或多個電腦執行時引起該一或多個電腦執行包括以下步驟之操作：

在具有一矩陣計算單元之一硬體電路上重複地執行以下操作：

將作為一迭代之一中間神經網路矩陣劃分成一當前細分中之複數個子矩陣；

產生複數個向量，該等向量之各者對應於該中間神經網路矩陣之各列且包含該中間神經網路矩陣之各列之值；

針對該複數個向量獲得複數個部分特性矩陣，該複數個部分特性矩陣經組態以自該複數個向量之各者擷取一部分值同時零化一剩餘部分數值；

藉由該複數個部分特性矩陣之一或多者將該複數個向量之各者相乘以產生一經更新神經網路矩陣之一列，其中該經更新神經網路矩陣包含該神經網路矩陣之元素，但調換該當前細分中之該複數個子矩陣之右上象限與左下象限；及

組合該等經產生之列以更新該中間神經網路矩陣；及

基於所完成之該等操作之全部迭代，產生該中間神經網路矩陣作為一轉置神經網路矩陣。

【第14項】

如請求項13之電腦儲存媒體，其中該等操作進一步包括：

接收針對在該硬體電路上之該神經網路執行計算之一請求，該請求指定待執行於該神經網路矩陣上之一轉置運算。

【第15項】

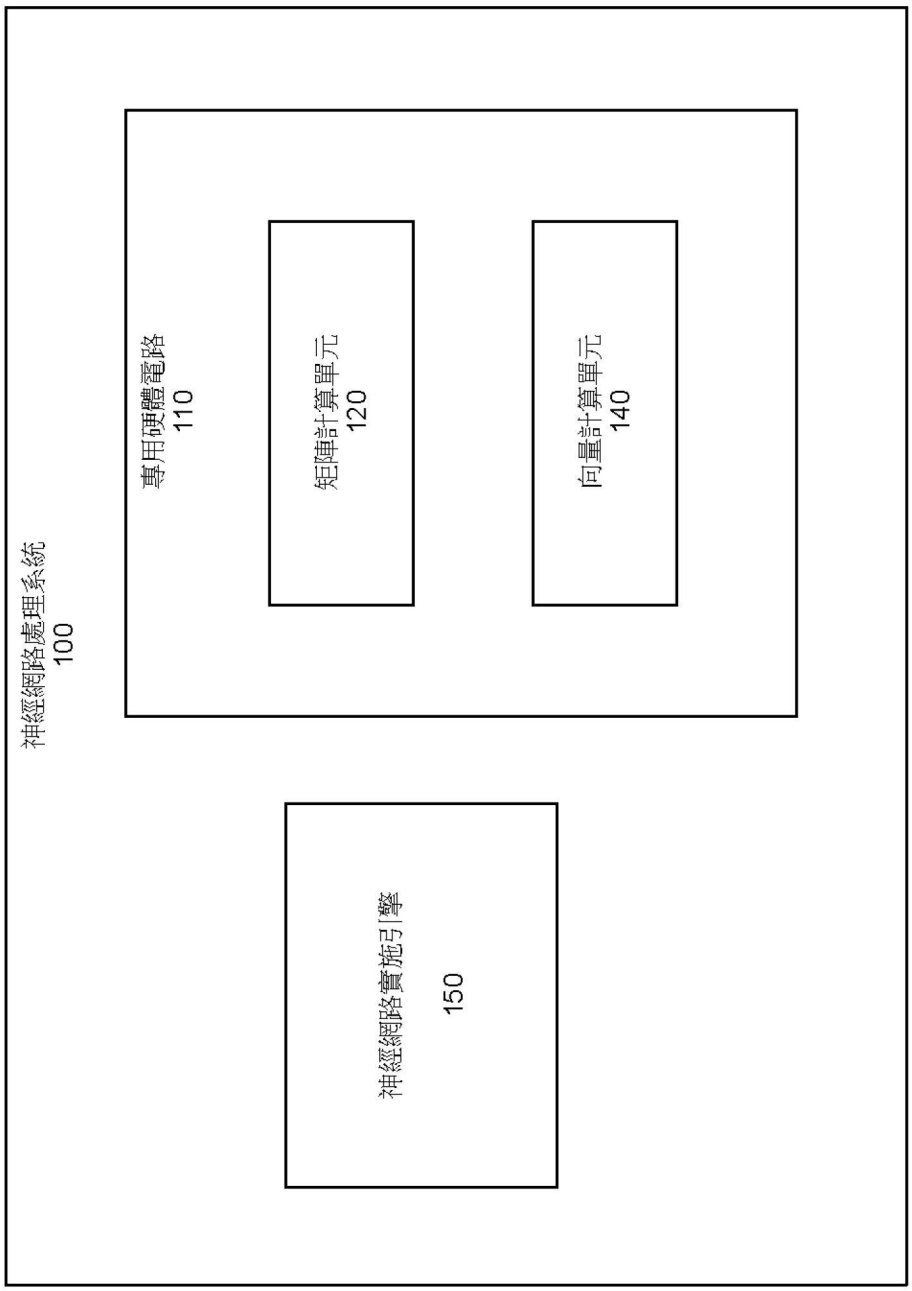
如請求項13之電腦儲存媒體，其中該等操作進一步包括：

判定該神經網路矩陣並非一i x i矩陣，其中i係該硬體電路之一向量長度值；

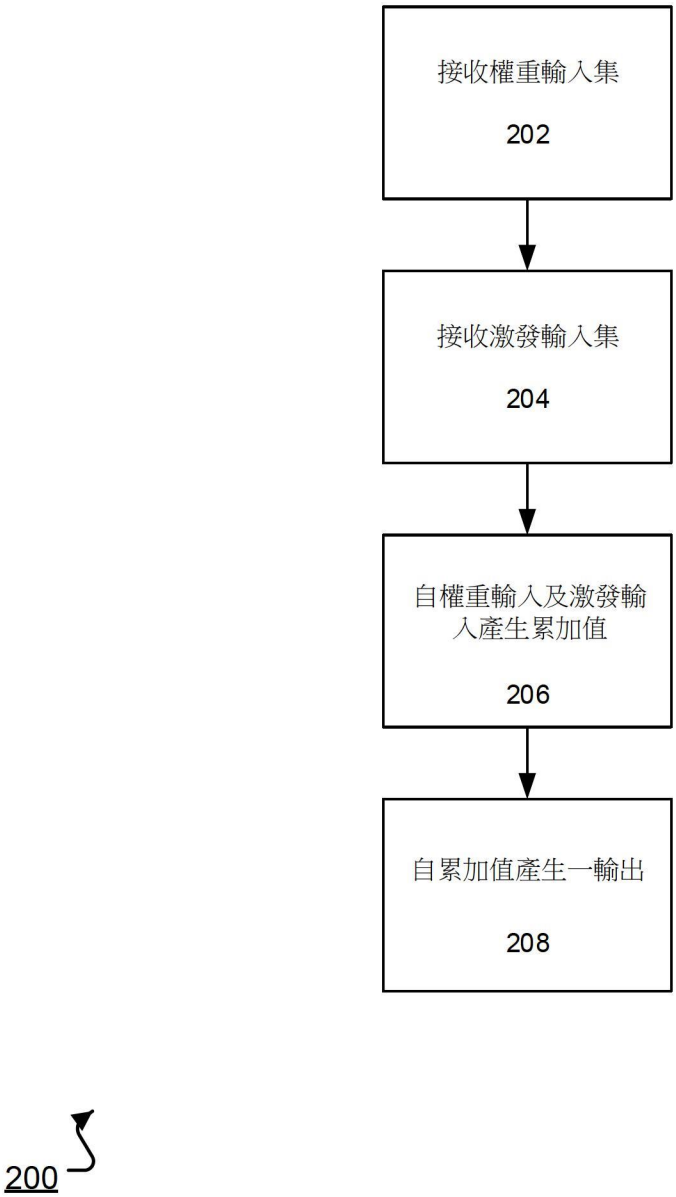
作為回應，藉由對該神經網路矩陣進行零填補而更新該神經網路矩陣以產生一i x i矩陣；及

藉由移除在該轉置期間所填補之零，在該轉置之前將該轉置神經網路矩陣轉換成其狀態。

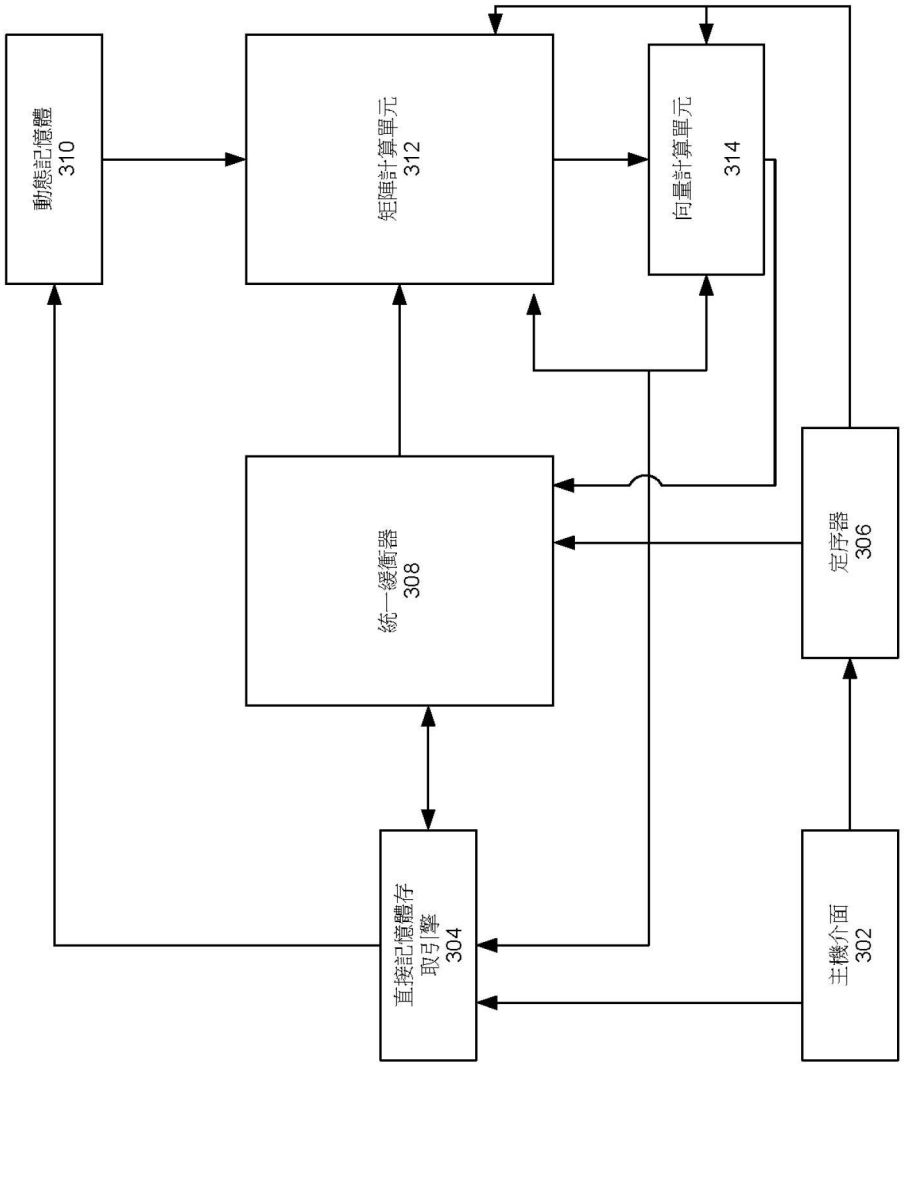
【發明圖式】



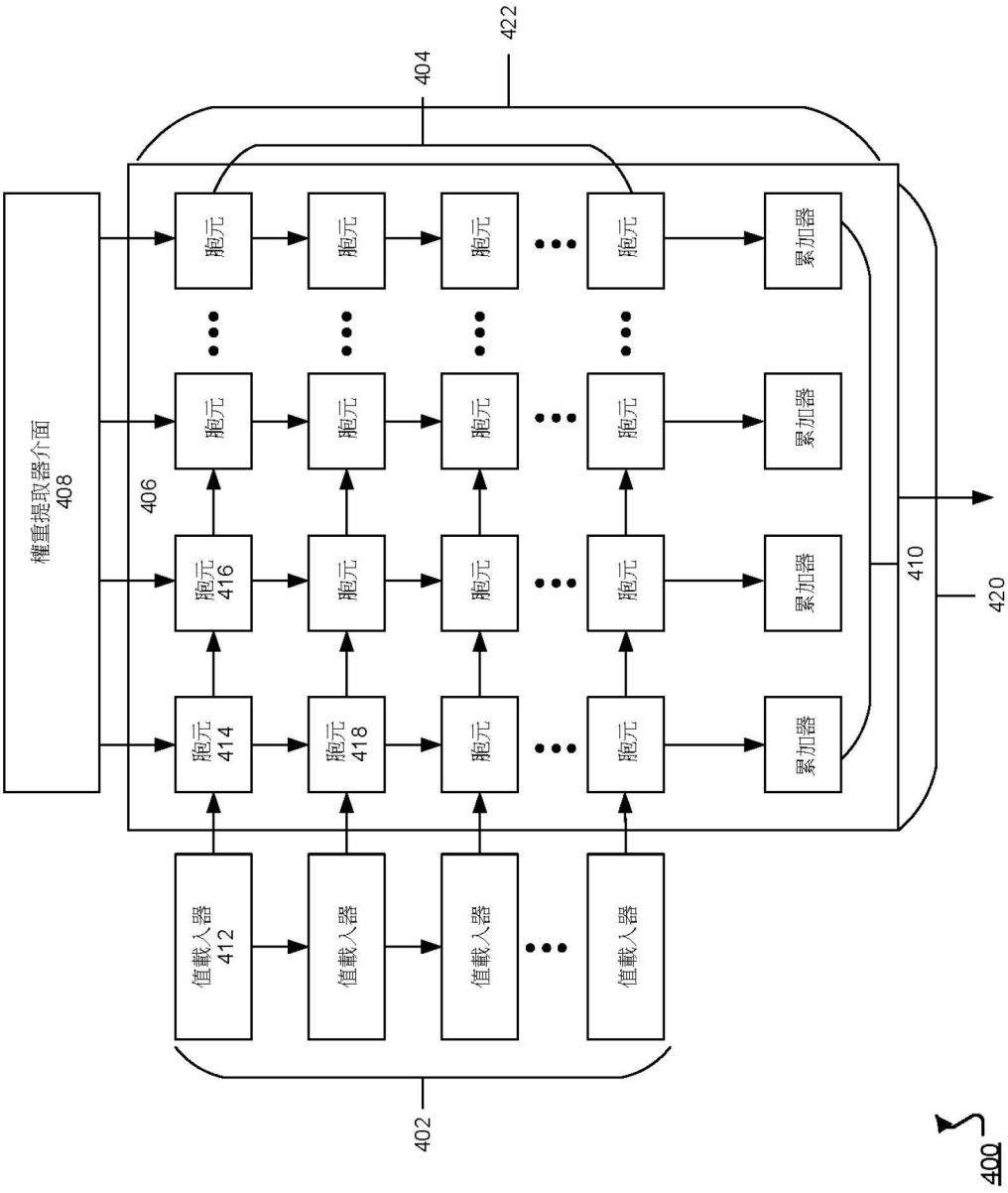
【圖1】


接收權重輸入集
202
接收激發輸入集
204
自權重輸入及激發輸入產生累加值
206
自累加值產生一輸出
208
200

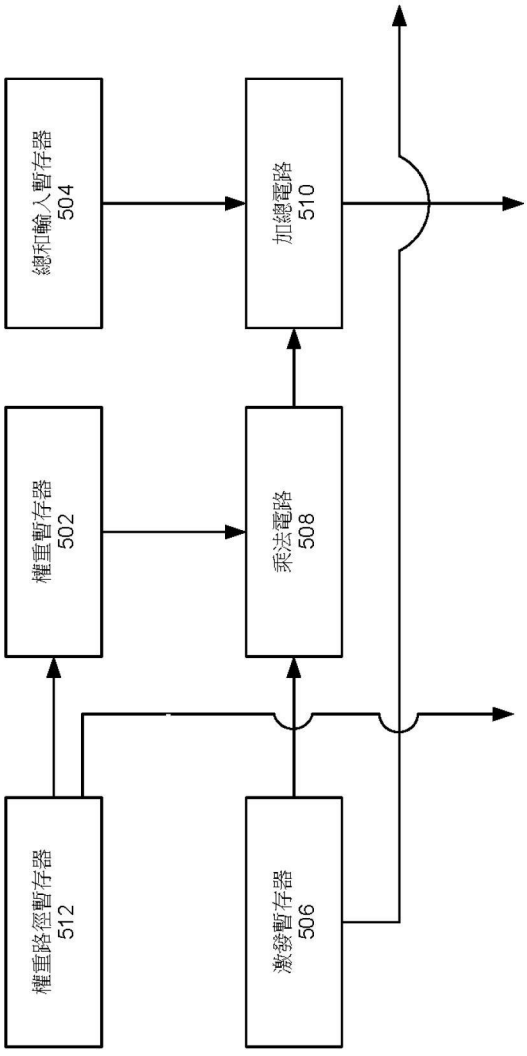
【圖2】


動態記憶體
310
矩陣計算單元
312
向量計算單元
314
統一緩衝器
308
定序器
306
直接記憶體存取引擎
304
主機介面
302
300

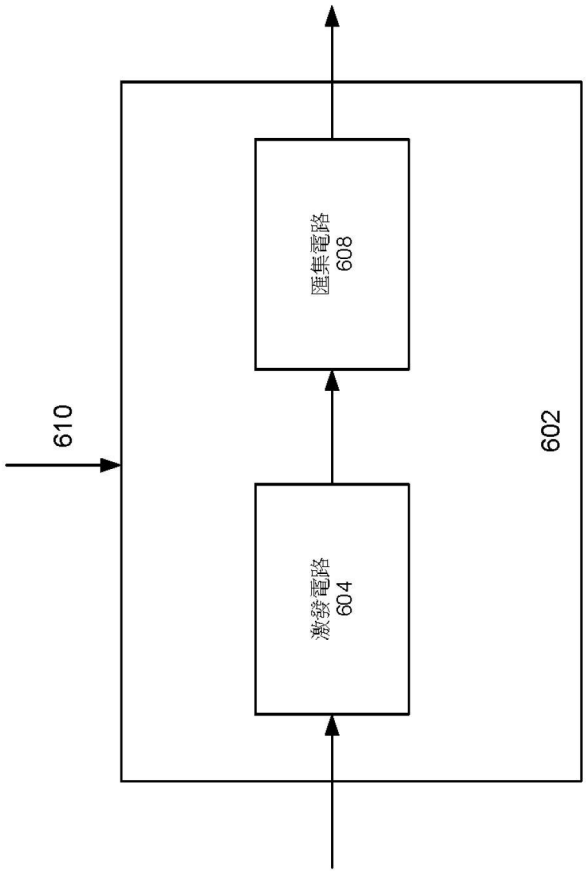
【圖3】


422
404
權重提取器介面
408
406
胞元
累加器
值載入器
412
胞元
416
胞元
414
胞元
418
410
420
402
400

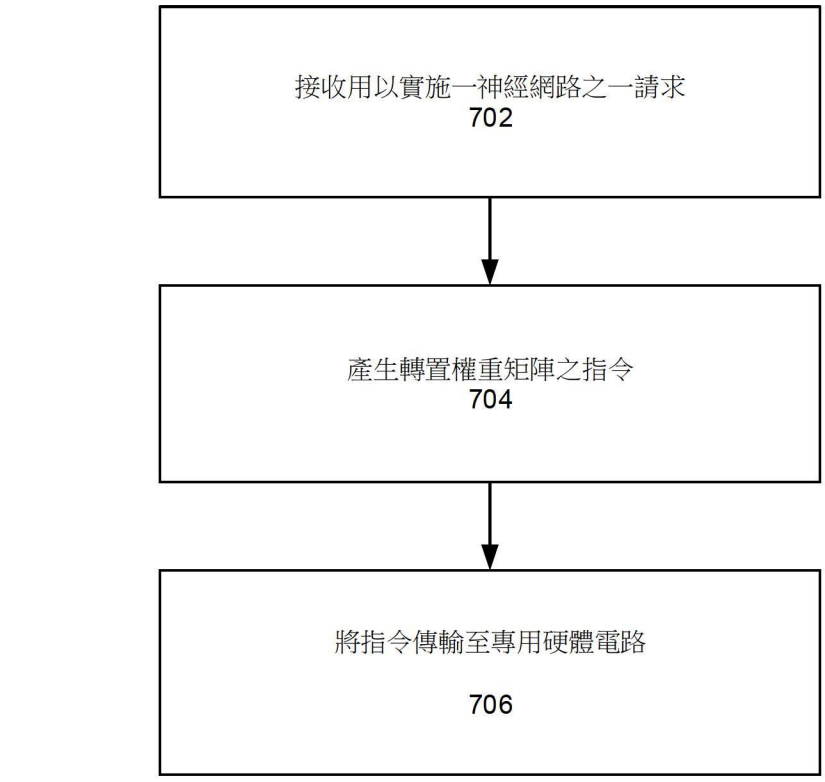
【圖4】


總和輸入暫存器
504
加總電路
510
權重暫存器
502
乘法電路
508
權重路徑暫存器
512
激發暫存器
506
500

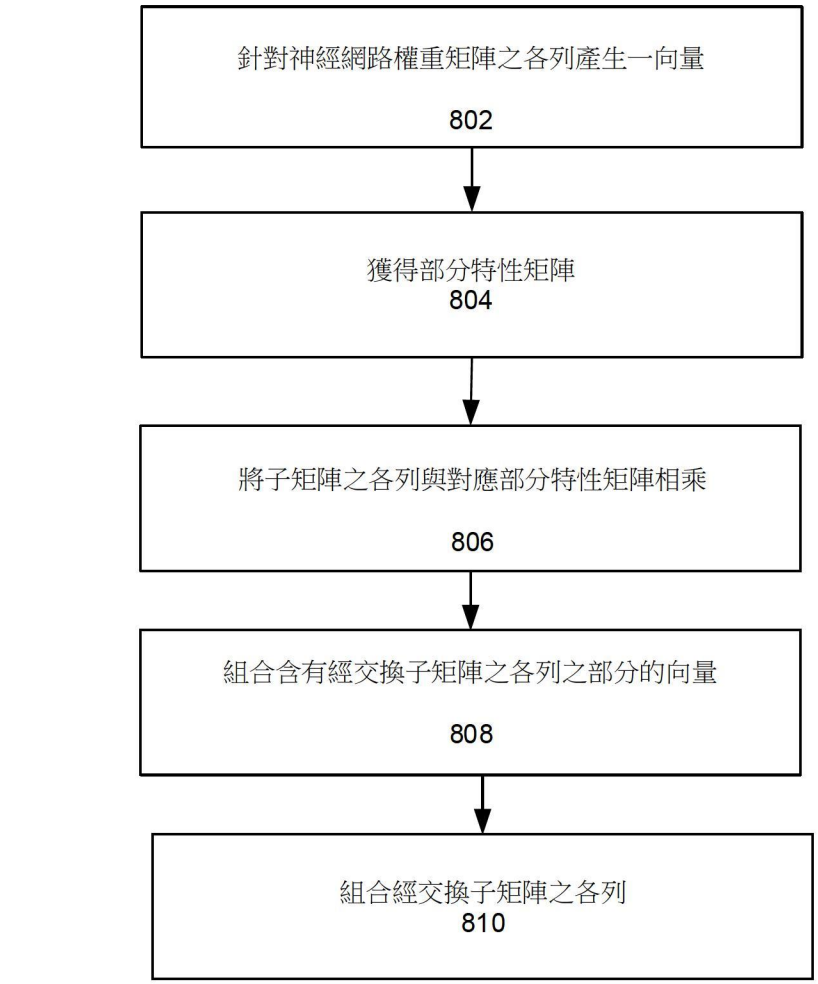
【圖5】



【圖6】


接收用以實施一神經網路之一請求
702
產生轉置權重矩陣之指令
704
將指令傳輸至專用硬體電路
706
700

【圖7】


針對神經網路權重矩陣之各列產生一向量
802
獲得部分特性矩陣
804
將子矩陣之各列與對應部分特性矩陣相乘
806
組合含有經交換子矩陣之各列之部分的向量
808
組合經交換子矩陣之各列
810
800

【圖8】

(a)

Input[0]	A	B	C	D
Input[1]	E	F	G	H
Input[2]	I	J	K	L
Input[3]	M	N	O	P

(b)

1	0	0	0
0	1	0	0
0	0	0	0
0	0	0	0

W1

0	0	0	0
0	0	0	0
1	0	0	0
0	1	0	0

W2

0	0	0	0
0	0	0	0
0	0	1	0
0	0	0	1

W3

0	0	1	0
0	0	0	1
0	0	0	0
0	0	0	0

W4

(c)

(1) (Input[0] * W1) + (Input[2] * W4) = [AB00] + [00IJ] = [ABIJ]

(2) (Input[1] * W1) + (Input[3] * W4) = [EF00] + [00MN] = [EFMN]

(3) (Input[0] * W2) + (Input[2] * W3) = [CD00] + [00KL] = [CDKL]

(4) (Input[1] * W2) + (Input[3] * W3) = [GH00] + [00OP] = [GHOP]

(d)

A	B	I	J
E	F	M	N
C	D	K	L
G	H	O	P

(e)

Input[4]	A	B	I	J
Input[5]	E	F	M	N
Input[6]	C	D	K	L
Input[7]	G	H	O	P

(f)

1	0	0	0
0	0	0	0
0	0	1	0
0	0	0	0

W5

0	1	0	0
0	0	0	0
0	0	0	1
0	0	0	0

W6

0	0	0	0
1	0	0	0
0	0	0	0
0	0	1	0

W7

0	0	0	0
0	1	0	0
0	0	0	0
0	0	0	1

W8

(g)

(1) (Input[4] * W5) + (Input[5] * W6) = AEIM
(2) (Input[4] * W7) + (Input[5] * W8) = BFJN
(3) (Input[6] * W5) + (Input[7] * W6) = CGKO
(4) (Input[6] * W7) + (Input[7] * W8) = DHLP

(h)

A	E	I	M
B	F	J	N
C	G	K	O
D	H	L	P

【圖9】