



(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 I843108 B

(45)公告日：中華民國 113 (2024) 年 05 月 21 日

(21)申請案號：111119283

(22)申請日：中華民國 111 (2022) 年 05 月 24 日

(51)Int. Cl. : G06N3/06 (2006.01)

G06F17/16 (2006.01)

G06T1/40 (2006.01)

(30)優先權：2021/05/25 美國

17/330,096

(71)申請人：美商應用材料股份有限公司 (美國) APPLIED MATERIALS, INC. (US)

美國

(72)發明人：蘇利 塔密希 SURI, TAMEESH (IN)；莊博超 JUANG, BOR-CHAU (TW)；席
納森尼爾 SEE, NATHANIEL (US)；塞依克 比拉爾沙菲 SHEIKH, BILAL SHAFI
(PK)；札曼 納菲德 ZAMAN, NAVEED (US)；沙克 麥倫 SHAK, MYRON
(US)；丹賈雅各 薩欽 DANGAYACH, SACHIN (IN)；哈曼特 烏戴庫瑪迪里普
洛 HANMANTE, UDAYKUMAR DILIPRAO (IN)

(74)代理人：李世章；彭國洋

(56)參考文獻：

CN 109858575A

US 2017/0316312A1

US 2018/0046916A1

US 2020/0221093A1

審查人員：蘇齊賢

申請專利範圍項數：20 項 圖式數：11 共 52 頁

(54)名稱

神經網路中的動態啟動稀疏性

(57)摘要

一種為神經網路層的輸出引起稀疏性的方法可包括：從神經網路層接收輸出；將輸出分割為複數個分區；識別可以視為具有零值的複數個分區中的第一分區；產生識別在複數個分區中的剩餘第二分區之中的第一分區的位置的編碼；以及將編碼及第二分區發送到神經網路中的後續層。

A method of inducing sparsity for outputs of neural network layer may include receiving outputs from a layer of a neural network; partitioning the outputs into a plurality of partitions; identifying first partitions in the plurality of partitions that can be treated as having zero values; generating an encoding that identifies locations of the first partitions among remaining second partitions in the plurality of partitions; and sending the encoding and the second partitions to a subsequent layer in the neural network.

指定代表圖：

符號簡單說明：

606:輸出張量

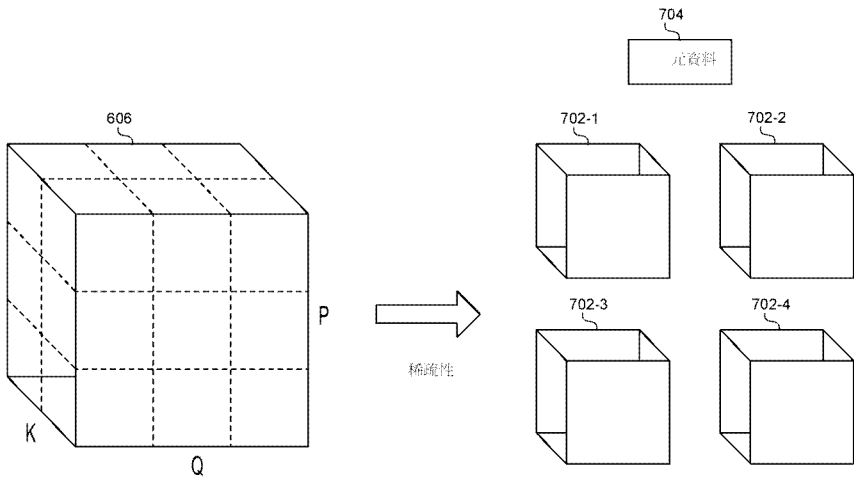
702-1:非稀疏分區

702-2:非稀疏分區

702-3:非稀疏分區

702-4:非稀疏分區

704:元資料



第7圖



I843108

【發明摘要】

【中文發明名稱】神經網路中的動態啟動稀疏性

【英文發明名稱】DYNAMIC ACTIVATION SPARSITY IN NEURAL NETWORKS

【中文】

一種為神經網路層的輸出引起稀疏性的方法可包括：從神經網路層接收輸出；將輸出分割為複數個分區；識別可以視為具有零值的複數個分區中的第一分區；產生識別在複數個分區中的剩餘第二分區之中的第一分區的位置的編碼；以及將編碼及第二分區發送到神經網路中的後續層。

【英文】

A method of inducing sparsity for outputs of neural network layer may include receiving outputs from a layer of a neural network; partitioning the outputs into a plurality of partitions; identifying first partitions in the plurality of partitions that can be treated as having zero values; generating an encoding that identifies locations of the first partitions among remaining second partitions in the plurality of partitions; and sending the encoding and the second partitions to a subsequent layer in the neural network.

【指定代表圖】第（ 7 ）圖。

【代表圖之符號簡單說明】

6 0 6：輸出張量

7 0 2 - 1：非稀疏分區

7 0 2 - 2：非稀疏分區

7 0 2 - 3 : 非 稀 疏 分 區

7 0 2 - 4 : 非 稀 疏 分 區

7 0 4 : 元 資 料

【特徵化學式】

無

【發明說明書】

【中文發明名稱】神經網路中的動態啟動稀疏性

【英文發明名稱】DYNAMIC ACTIVATION SPARSITY IN NEURAL NETWORKS

【技術領域】

【0001】 本申請案主張於2021年5月25日提交並且標題為「DYNAMIC ACTIVATION SPARSITY IN NEURAL NETWORKS」的美國非臨時申請案第17/330,096號的權益及優先權，該申請案的全部內容出於所有目的藉由引用的方式併入本文中。

【0002】 本揭示大體描述了在神經網路計算中引起稀疏性以減少記憶體瓶頸。具體地，本揭示描述了用於分割層輸出並且在每個分區的基礎上引起稀疏性的方法及系統。

【先前技術】

【0003】 神經網路可以大體經定義為識別輸入資料集中的基礎關係的一系列連續操作。神經網路以一方式處理資訊，該方式模型化人類思想操作的方式。由此，神經網路中的中間級可使用被稱為神經元的計算元件。神經元之間的連接如同生物系統中的突觸一般操作以在神經元層之間傳輸中間計算。每個神經元的輸出可使用結合不同突觸輸入的不同類型的函數來計算。突觸可在每個神經元的輸入處進行加權，並且此等權重可使用訓練過程設置。神經網路藉由具有已知結果的處理實例資料來訓練以在輸入與輸

出之間形成概率加權關聯，該等概率加權關聯在網路本身的資料結構內儲存為權重或參數。訓練可以在使用訓練資料的受監督學習環境中進行，或訓練可使用在使用期間接收的輸入資料而不受監督。

【0004】 已經設計計算硬體以經由神經網路函數最佳化輸入資料的處理。例如，神經網路編譯器可接收神經網路的基於代碼的定義，並且為硬體神經網路加速器中的一或多個計算節點產生指令。加速器上的計算節點可包括並行地有效處理神經網路操作的獨立小晶片或其他計算區塊。來自神經網路的每一層的輸出可在已經接收中間結果之後儲存在暫時性緩衝器或晶片上記憶體中，隨後傳遞到神經網路中的後續層。然而，隨著現代神經網路的計算需求及輸入大小持續增加，層之間的記憶體儲存正迅速成為嚴重的瓶頸，並且並行處理的需求正變得難以管理。由此，在此技術中需要改進。

【發明內容】

【0005】 在一些實施例中，一種為神經網路層的輸出引起稀疏性的方法可包括：從神經網路層接收輸出；將輸出分割為複數個分區；識別可以視為具有零值的複數個分區中的第一分區；產生識別在複數個分區中的剩餘第二分區之中的第一分區的位置的編碼；以及將編碼及第二分區發送到神經網路中的後續層。

【0006】 在一些實施例中，一種神經網路加速器可包括經配置為實施神經網路層並且從該層產生輸出的計算節點，

以及經配置為執行包括下列的操作的分割電路：從神經網路層接收輸出；將輸出分割為複數個分區；識別可以視為具有零值的複數個分區中的第一分區；以及產生識別在複數個分區中的剩餘第二分區之中的第一分區的位置的編碼。神經網路加速器亦可包括經配置為儲存編碼及第二分區的記憶體用於神經網路中的後續層。

【0007】 在一些實施例中，一種為神經網路層的輸出引起稀疏性的方法可包括：從神經網路層接收輸出；以及將輸出分割為複數個分區，其中複數個分區的每一者包含複數個輸出。方法亦可包括：識別滿足一準則的複數個分區中的第一分區，該準則指示第一分區中的值可設置為零；產生識別在複數個分區中的剩餘第二分區之中的第一分區的位置的編碼；將編碼及第二分區發送到神經網路中的後續層並且丟棄第一分區；在神經網路中的後續層處接收第二分區；基於編碼來佈置具有零值的第二分區；以及執行神經網路中的後續層。

【0008】 在任何實施例中，以下特徵中的任一者及全部可以任何組合實施並且不作限制。方法/操作亦可包括：在神經網路中的後續層處接收第二分區；以及基於編碼來佈置第二分區。後續層可執行乘法運算，藉此第一分區可以作為乘以零運算而丟棄。輸出可包括來自該層的輸出的三維陣列，其中輸出的陣列包含神經網路中的不同通道的維度。複數個分區可包括輸出的陣列的三維分區。第一分區不需要在複數個分區中係連續的。識別可以視為具有零值

的複數個分區中的第一分區可包括：從設計環境接收準則；以及將準則應用於複數個分區的每一者。準則可包括相對量值函數計算分區中的值的集合，並且若集合小於閾值，則將分區中的值設置為零。準則可以作為運行時間函數從設計環境發送。準則可作為表示神經網路的圖形的部分編碼。神經網路加速器亦可包括複數個小晶片，其中計算節點可在複數個小晶片中的第一小晶片上實施，並且其中後續層可在複數個小晶片中的第二小晶片上實施。神經網路加速器亦可包括經配置為執行包括下列的操作的定序器電路：在神經網路中的後續層處接收第二分區；以及基於編碼來佈置第二分區。神經網路層可包括執行卷積核心。記憶體可包括晶片上靜態隨機存取記憶體（static random-access memory；SRAM）。當訓練神經網路時，不需要使用分割電路。可在訓練神經網路期間決定複數個分區中的分區的數量。識別可以視為具有零值的複數個分區中的第一分區可包括：從設計環境接收準則；以及將準則應用於複數個分區的每一者。輸出可包括來自該層的輸出的三維陣列，其中輸出的陣列可包括神經網路中的不同通道的維度，並且其中複數個分區可包括輸出的陣列的三維分區。

【圖式簡單說明】

【0009】 對各個實施例的性質及優點的進一步理解可藉由參考說明書的剩餘部分及附圖來實現，其中相同的元件符號在若干附圖中使用以代表類似部件。在一些情況中，子

標記與元件符號相關聯以指示多個類似部件中的一者。當參考元件符號而不規定現有的子標記時，其意欲代表所有此種多個類似的部件。

【0010】 第1圖示出了用於不同神經網路架構或模型的計算縮放的圖形。

【0011】 第2圖示出了樣本神經網路中的每個通道的啟動密度分佈的圖表。

【0012】 第3圖示出了根據一些實施例的用於最佳地利用啟動稀疏性的組合的演算法到硬體途徑的圖。

【0013】 第4圖示出了根據一些實施例的通用神經網路加速器。

【0014】 第5圖示出了根據一些實施例的引起稀疏性的改進的神經網路加速器。

【0015】 第6圖示出了根據一些實施例的卷積操作的濾波器可如何產生可以藉由分割電路分割的多維輸出陣列的實例。

【0016】 第7圖示出了輸出張量可如何在任何維度上分割。

【0017】 第8圖示出了根據一些實施例的分割引起的稀疏性提供超過輸出啟動映射中發現的隨機稀疏性的改進。

【0018】 第9圖示出了根據一些實施例的多瓦片或A1小晶片架構。

【0019】 第10圖示出了根據一些實施例的用於為神經網路層的輸出引起稀疏性的方法的流程圖。

【0020】 第 11 圖示出了其中可實施各個實施例的示例性電腦系統。

【實施方式】

【0021】 人工智慧 (Artificial Intelligence; AI) 繼續變得更普遍。隨著 AI 的使用變得更加廣泛，AI 正在啟用先前被認為過於複雜的新使用情況。AI 在許多不同學科中的此增加的採用驅動了 AI 硬體及軟體所需的效能需求。例如，新演算法繼續解決來自電腦視覺 (computer vision; CV) 及自然語言處理 (natural language processing; NLP) 的更複雜的使用情況，並且對計算能力及記憶體儲存的增長的需求正擴展到超出單獨使用習知處理縮放所支援的範圍。對 AI 系統的效率的未來改進將可能導致一起影響技術堆疊的不同位準的創新，而非單獨對硬體、軟體、訓練等的創新。

【0022】 第 1 圖示出了用於不同神經網路架構或模型的計算縮放的圖形 100。此圖形 100 總結了近年來不同 CV 及 NLP 神經網路模型的計算增長。注意到，CV、NLP、及 / 或語音識別的計算需求的增長已經迅速超過了遵循莫耳定律 (Moore's law) 的計算能力的自然增長。當考慮計算需求以甚至更快的速率增長的基於變換器的神經網路時，此種差異變得甚至更加顯著。儘管在第 1 圖中表示的絕對浮點運算 (floating-point operations; FLOPS) 度量具體地關於神經網路訓練，但神經網路執行的訓練及推理計算的總體計算縮放趨勢係相同的。與在資料中心或雲端平臺

上執行的計算相比，當使用具有有限計算能力的智慧邊緣裝置時在第1圖中示出的效能縮放的要求變得甚至更加顯著。

【0023】顯然，傳統計算及記憶體縮放將不能支援將來AL要求的增長及採用率。儘管從神經網路演算法到硬體實施方式對AI堆疊的不同部分進行了持續的努力，但大部分此等努力本質係靜態的。現有的最佳化努力經常以基於參數的模型壓縮途徑為中心，諸如量化或修剪。替代地，最佳化努力專門集中於演算法位準，諸如知識蒸餾或低秩因數分解。儘管此等分離方法獨立地提供記憶體及電腦使用的減少，但歸因於最佳化的過程位準及將此等改進限制為到具體輸入資料集或模型的準確性折衷，總體效率受到限制。

【0024】隨著模型變得更深，更多內層及輸入張量在大小上繼續向上縮放，效能要求可以加劇。例如，ResNet-152模型可包括152個內層，輸入張量可包括高解析度影像，並且輸入可從多個源（諸如多個照相機串流）拼貼在一起。利用此等大的資料集，啟動記憶體大小變成主要瓶頸，並且甚至超過儲存神經網路的權重及參數的參數記憶體大小。如本文使用，參數記憶體指神經網路本身的權重及參數的儲存，而啟動記憶體指流過神經網路的張量的動態輸入/輸出。習知的模型壓縮技術（諸如量化、權重修剪等）僅集中在參數記憶體上而非啟動記憶體上，因此未解決此瓶頸。

【0025】 當前未在神經網路技術中發現用於解決啟動記憶體瓶頸的通用解決方案。具體地，由於大部分神經網路使用某種形式的非線性（例如，**R e L U**、**S i g m o i d**、**T a n h**等）作為每一層的部分，來自每一層的啟動輸出將具有自然發生的稀疏性位準。換言之，隨著執行啟動函數，此等啟動函數趨於將許多值（諸如負值）強制為零。然而，此稀疏性係動態的。與神經網路中的參數權重的稀疏性不同，此稀疏性將隨著每個輸入張量而不同，使在設計時不可能預測此稀疏性的位置。此使得在硬體中利用動態啟動稀疏性非常具有挑戰性，並且習知的硬體加速器不支援此類型的最佳化。

【0026】 第2圖示出了樣本神經網路中的每個通道的啟動密度分佈的圖表200。圖表200中的資料來源於**V G G - 1 6**，其係基於卷積結構的流行影像分類神經網路。Y軸上的每個通道表示唯一神經網路層，並且圖表200上的每個點表示每個通道的密度。可以觀察到，對於跨過神經網路中的大部分層的通道，啟動分佈係高度不規則且不均勻的。換言之，不同通道中的稀疏性係不可預測的，並且很大程度上取決於運行時間輸入。此外，圖表200揭示了由稀疏性的非均勻動態分佈產生的另一挑戰，在本文中稱為「尾工」效應。具體地，尾工效應將總體速度限制到最慢或「尾」工。由於大部分硬體加速器將神經網路層分開或分離為在並行處理元件上並行執行的多個較小內核，所以此導致利用啟動稀疏性來改進效能的上限有限。

【0027】 類似地，啟動輸出中的稀疏性的不可預測分佈限制了可藉由移除零值來實現的記憶體節省。具體地，若從啟動映射中移除稀疏的零值，則仍需要保留所移除的元素的相應編碼。換言之，必須保留指定哪些零元素已經被移除的編碼，使得原始輸出集可以重構為後續層的輸入。此意味著在沒有至少50%稀疏性的情況下將不太可能實現記憶體節省，且低於此閾值的啟動張量可實際上導致記憶體使用及頻寬的增加。

【0028】 本文描述的實施例提出了通用架構框架及整體演算法到硬體的途徑以利用神經網路中的動態啟動稀疏性。此架構在啟動特徵映射（例如，層的輸出）中引入並且引起「結構化稀疏性」，其中藉由在層輸出中創建分區，將稀疏性的結構定製為架構的基礎執行單元。例如，包括SIMD、VLIW、脈動陣列、卷積引擎、MAC操作等的每個執行單元可具有定製的分區類型及大小。此等不同操作的每一者亦可具有用於引起稀疏性並且將整個分區設置為零的個別準則。在演算法及框架位準處經定製為對應執行單元的基礎組織的此結構的使用可產生最佳設計點，目的為最佳化電腦使用、記憶體容量、及互連頻寬。

【0029】 稀疏分區不需要儲存在啟動層之間的記憶體中。除了記憶體節省之外，亦可以消除具有稀疏啟動的計算操作。例如，當將整個輸入張量設置為零時，可以消除將輸入張量乘以具體權重的計算節點的輸入，並且因此此計算操作可以在後續層中完全跳過。此可以導致神經網路中的

顯著計算減少。此外，隨著莫耳定律的緩慢以及採用基於異構小晶片的解決方案來支援AI的不斷增長的計算需求，利用啟動稀疏性的此等實施例可以減輕封裝上互連中的頻寬壓力。此允許對基於小晶片的架構上的AI工作負載進行近單片式縮放，甚至利用封裝上互連及此等設計中固有的減小的密度。

【0030】 第3圖示出了根據一些實施例的用於最佳地利用啟動稀疏性的組合的演算法到硬體途徑的圖300。架構可包括深度學習框架302。深度學習框架可包括允許使用者簡單地構建深度學習模型的使用者介面及程式庫/工具。深度學習框架302的實例可包括TensorFlow®、PyTorch®、Keras®、Sonnet®、及/或其他商業可用工具。深度學習框架可從預先訓練的模型、使用者定義的模型、及/或樣本資料集得出，用於為具體應用開發新神經網路。

【0031】 一些實施例可添加定製程式庫304，在本文中稱為「PartitionDropout」，該定製程式庫可與深度學習框架302整合。PartitionDropout丟棄程式庫可與預先訓練的模型一起使用，或模型可以利用添加到設計中的PartitionDropout來訓練。程式庫304允許神經網路設計者在設計過程期間評估最佳分區大小、計算、記憶體容量及/或頻寬減少折衷。

【0032】 PartitionDropout程式庫可用於添加代碼以配置AI硬體中的額外硬體元件，用於在各層的啟動映射中引

起稀疏性。例如，此程式庫 3 0 4 可允許使用者為來自層的輸出指定各種大小及形狀的分區。此外，程式庫 3 0 4 可允許神經網路設計者指定決定或識別可以視為具有零值的層輸出中的分區的準則或功能。此等兩個參數（亦即，分割方案及準則）可由神經網路設計者經實驗設置或選擇。

【0033】 例如，一些實施例可使用可能分區大小及結構的列表利用神經網路處理樣本資料。所得的模擬輸出隨後可在頻寬、計算、及 / 或記憶體節省方面表徵為與使用其他分區大小 / 結構的模擬結果相比具有準確性的折衷。最佳分區大小 / 結構可隨後從模擬結果中選擇。類似地，可使用不同閾值模擬所使用的準則以識別在準確性與所得硬體效率之間的折衷中的最佳拐點。例如，基於量值的準則可計算分區中的值的集合，並且若集合小於閾值，則將分區中的所有值設置為零。可在模擬期間向上 / 向下調節此閾值以找到最佳值。

【0034】 每個網路或每個層的元資料可能需要與基礎硬體通訊，以便硬體實施如上文描述的深度學習框架中設計的方案。例如，所選的準則及閾值連同分區大小或結構可能需要從深度學習框架 3 0 2 通訊到硬體 3 1 0。架構 3 0 0 提供了用於提供此通訊的多種不同方法。在一些實施例中，編譯器可將分割及 / 或準則整合到神經網路圖形 3 0 6 中，該神經網路圖形傳輸到硬體 3 1 0。編譯的神經網路圖形 3 0 6 可包括在計算層執行之後執行 `PartitionDropout` 層的操作的指令。例如，在神經網路中的層的計算操作之後執行的分割

電路可由編譯器視為神經網路的部分，並且用於產生分區且執行準則以引起稀疏性的指令可實施為神經網路圖形 306 的一部分。替代地，一些實施例可發送包括 PartitionDropout 指令集架構 (instruction set architecture; ISA) 的神經網路運行時間。神經網路運行時間 308 可發送到硬體 310 以分別對 AI 加速器或其他硬體中的分割電路進行程式化。

【0035】 最後，硬體 310 可執行具有如上文描述的 PartitionDropout 分割及 / 或準則的圖形。例如，硬體 310 可包括多瓦片或 AI 小晶片解決方案，其中神經網路或層在不同 AI 瓦片或小晶片上分佈。如下文描述，硬體 310 可包括實施在深度學習框架 310 中指定的準則及 / 或分割功能的電路。可在由硬體 310 中的計算節點實施的任何及 / 或所有層之後包括此等分割電路。

【0036】 第 4 圖示出了根據一些實施例的通用神經網路加速器 400。架構可包括晶片上 SRAM 404 及 / 或晶片上記憶體 402。隨著輸入 / 輸出張量穿過神經網路的各個層傳播，此等記憶體可儲存該等輸入 / 輸出張量。執行單元 406 可執行神經網路的一或多個層的一或多個操作。在此實例中，執行單元 406 可包括內部輸入緩衝器 408，該內部輸入緩衝器從先前計算節點或從神經網路的輸入接收輸入張量。輸入緩衝器 408 可包括具有部分空間維度及通道維度以及一些情況的濾波器。輸入緩衝器 408 可將張量提供到計算核心或計算節點 410，該計算核心或計算節點對從輸入緩衝

器 408 接收的輸入張量執行一或多個操作。例如，計算節點 410 可執行卷積運算並且可使用浮點乘加 (floating-point multiply-add; FMA) 引擎來實施。計算節點 410 的輸出可傳遞到輸出緩衝器 412。輸出緩衝器可累加來自計算節點 410 的卷積結果。藉由計算節點 410 產生的部分和可從輸出緩衝器 410 溢出到晶片上 SRAM 404 中，並且進一步溢出到晶片上記憶體 402 上。

【0037】 第 5 圖示出了根據一些實施例的引起稀疏性的改進的神經網路加速器 500。此神經網路加速器 500 可包括上文針對第 4 圖的神經網路加速器 400 描述的部件。然而，此神經網路加速器 500 亦可包括經配置為在計算節點 410 的輸出中產生稀疏性的分割電路 504，連同經配置為當已經移除稀疏分區之後定序輸入的定序器電路 502。分割電路 504 及定序器電路 502 可使用神經網路圖形及/或使用元資料來程式設計，該元資料來自藉由如上文描述的深度學習框架提供的運行時間。

【0038】 分割電路可從神經網路層接收輸出。此層可藉由計算節點 410 實施，並且可執行不同數學函數，諸如啟動函數、卷積函數等等。來自計算節點 410 的輸出可在輸出緩衝器 412 中接收及/或累加。分割電路 504 可隨後執行數個行動。首先，分割電路 504 可將輸出分割為複數個不同分區。分區結構/大小可在深度學習框架中決定並且傳遞到如上文描述的分割電路 504。在下文提供可如何分割啟動映射張量的實例。注意到，將輸出分割為複數個分區不一

定需要移動或改變任何實際值或記憶體元件。而是，分割電路504可根據預定的分區大小/結構將分區識別為多組值並且可執行準則或以其他方式將每個分區一起處置為單個實體。

【0039】 分割電路亦可識別可以視為具有零值的複數個分區中的分區。此操作可以各種不同方式執行。在一些實施例中，從深度學習框架接收的準則可對每個分區執行。準則的目的可係決定作為整體的分區是否包括足夠小的值，使得分區可視為僅具有零值。例如，若 $2 \times 2 \times 6$ 分區中的值具有小於0.1的集合總數，則分區中的所有值可視為零。注意到，本揭示不限制可使用的準則的類型。準則的一個實例係集合每個分區中的值並且將集合的值與閾值進行比較的準則，若聚合低於閾值，則將分區視為零值。其他實施例可使用不同準則。亦注意到，準則可單獨執行或與其他準則一起作為準則集合來執行。由此，對單個準則的任何引用亦允許在分區上以任何組合執行多個準則。

【0040】 將分區視為具有零值可包括將實際零值（例如，0.0）寫入分區中的儲存位置的每一者中。此操作可覆寫先前儲存為計算節點410的輸出的任何值。注意到，此可能係有損程式，該程式可導致準確性的至少一些損失。然而，神經網路操作可以容忍在中間層處的準確性的小損失。此操作亦可以區別於啟動函數，或其他函數在獨立記憶體位置上一次一個地執行。替代將單個值與閾值進行比較並且將其設置為零，此操作將整個分區的值設置為零（或將其

等視為零)。因此，若用於分區的準則指示零，則單個位置中的相對大的非零值可在該分區中設置為零。

【0041】 在一些實施例中，將分區視為具有零值不需要要求將任何實際零值寫入分區的儲存位置中。而是，分區可視為具有零值。例如，分區可丟棄並且不傳遞到後續層或晶片上 S R A M 4 0 4。無論是否將實際零值寫入分區的記憶體位置，當將輸出儲存到記憶體時，此等分區可丟棄。例如，當將分區儲存到記憶體時，分割電路 5 0 4 可產生識別在整體輸出陣列中視為具有零值的分區的位置的編碼。例如，可產生具有與每個分區相關聯的單個位元的二進制串。0 值可指示分區應當視為具有零值，而 1 值可指示分區應當視為具有在記憶體中儲存的非零值。替代將所有分區儲存到記憶體，視為具有零值的第一組分區（「第一分區」）可丟棄，而具有非零值的第二組分區（「第二分區」）可儲存在記憶體中。此編碼可產生極大的記憶體節省，並且減少由非常大的輸出張量導致的記憶體瓶頸。例如，分為 2 5 個分區的 3 D 輸出陣列可在例如彼等分區中的 1 0 個分區中引起稀疏性。替代儲存充滿值的 2 5 個分區，分割電路 5 0 4 僅需要儲存具有編碼輸出的 2 5 個位元串的 1 5 個分區。

【0042】 一些實施例已在每個層中引起 4 0 % 的平均稀疏性。當如上文描述在分區中引起此稀疏性時，此導致啟動記憶體中的 4 0 % 節省。在對晶片上記憶體資源具有約束的邊緣裝置中，此減少可以直接轉化為非晶片或晶片上記憶體頻寬中的效能節省。此藉由最小化每個操作的記憶體傳

送次數來改進記憶體存取時間並且改進神經網路操作的整體速度。

【0043】 分割電路 504 可將編碼及具有非零值的第二組分區發送到記憶體（例如，晶片上 SRAM 404）。替代地，分割電路 504 可將輸出直接發送到神經網路中的後續層或計算節點的另一輸入緩衝器 408。

【0044】 當後續層從分割電路 504 接收編碼張量時，定序器電路 502 可解碼張量以在正確的位置中提供第二組分區用於處理。稀疏格式的張量可以讀取，並且定序器電路 502 中的控制邏輯可以選擇要發送到此或其他執行單元的不同分區。例如，定序器電路 502 可讀取編碼並且視需要將充滿零值的分區插入輸入張量中。定序器電路 502 可重組張量，使得其具有期望大小，其中非零值出現在輸入張量中的期望佈置次序中。

【0045】 除了節省記憶體頻寬之外，此分割亦可消除藉由神經網路加速器 500 執行的一些計算操作。在一些實施例中，獨立分區可發送到不同執行單元 406。若操作係接收已經設置為零值或否則應當視為具有零值的分區，彼操作可在一些情況下消除。例如，若計算節點處的操作涉及乘法運算，則零分區可導致彼操作的輸出為零。因此，替代實際上執行操作，零輸出可以在不執行乘法運算的情況下產生，並且可消除對應計算級。利用不連續張量，相應輸出緩衝器可基於編碼中的輸入張量結構來選擇。定序器電路 502 中的此控制邏輯可執行此操作。

【0046】 第6圖示出了根據一些實施例的卷積操作的濾波器可如何產生可以藉由分割電路分割的多維輸出陣列的實例。啟動函數的輸入張量602可具有 $H \times W$ （高度 $\times$ 寬度）的空間維度，該等空間維度具有多個輸入通道 C ，因此產生三維輸入陣列。空間卷積可使用複數個濾波器604藉由啟動函數來執行。濾波器的每一者可具有維度 $R \times S$ ，該等維度具有與輸入張量602相同數量的通道 C 。啟動函數可在卷積操作期間應用 K 個不同濾波器。所得的輸出張量606可表徵為 K 個濾波器的每一者的 $P \times Q$ 二維陣列係604。

【0047】 第7圖示出了輸出張量606可如何在任何維度上分割。注意到分區可跨空間及通道維度分離輸出張量606，從而產生2D或3D分區。注意到，第7圖中示出的分區僅藉由實例的方式提供並且不意欲為限制性。可使用分區的任何結果或大小。亦應當注意，當設計不同分區時，神經網路加速器中的不同計算節點之間的通信模式將改變。例如，隨著分區改變，某些分區應當作為神經網路中的區塊發送的位置亦可基於神經網路的獨立設計來改變。此路由資訊亦可從深度學習框架提供到神經網路加速器的硬體部件，使得將分區路由到正確的位置。

【0048】 在對輸出張量606中的各個分區應用準則並且引起稀疏性之後，分割電路可將輸出張量606中的18個分區減少到四個非稀疏分區702。元資料704可儲存編碼，使得原始輸出張量606可以表示/重建，並且非稀疏分區702可

以發送到正確的計算節點。若一些後續層操作需要，則元資料 704 中的編碼亦可用於產生稀疏分區。

【0049】 第 8 圖示出了根據一些實施例的分割引起的稀疏性提供超過輸出啟動映射中發現的隨機稀疏性的改進。儘管一些正規化技術（例如，L1/L2、丟棄等）或修改的啟動函數（例如，FATReLU）已經顯示出增加啟動稀疏性，藉由此等函數引起的稀疏性本質上仍係隨機的並且難以由系統級架構利用，如使用此等標準丟棄技術藉由啟動映射 802 示出。本文引入的新中間層（分割電路及定序器電路）提供了結構化丟棄技術，該技術可以用於強制某一比例的啟動映射完全稀疏。此新層經設計為決定性的並且在訓練及/或推理期間應用。例如，在如上文描述的基於量值的準則中，啟動映射可首先分為跨空間及/或通道維度切割的連續分區的網格，其中每一者可視為具有零值，並且使用分區丟棄技術，基於如由啟動映射 804 示出的啟動量值的秩來將其全部丟棄或保留。儘管此可能會降低準確性，但情況並非如此。在一些情況下，與使用標準稀疏性的啟動映射 802 相比，分區引起的稀疏性已經顯示為獲得較佳驗證準確性。此顯示除了實現上文描述的硬體加速之外，分割的丟棄提供了更有效的正規化。

【0050】 第 9 圖示出了根據一些實施例的多瓦片或 AI 小晶片架構。除了減少記憶體使用並且減少計算使用之外，當跨多個 AI 晶粒、瓦片、或小晶片縮放時，用於神經網路加速器的 Partition Dropout 架構亦可以導致對互連頻寬的

顯著節省。儘管小晶片解決了大的單片晶粒中固有的縮放及成本的問題，其等通常不提供與單片晶粒相同位準的互連密度及功率效率，因此與單片解決方案相比，諸如 AI 加速器的相干區塊的分裂可導致較低計算縮放。然而，本文描述的架構減輕了多個 AI 晶粒、瓦片、或小晶片之間的互連上的頻寬壓力。此亦改進了跨許多不同 AI 小晶片的 AI 計算縮放的效能及功率效率。

【0051】 第 9 圖示出了使用以 2D 網狀拓撲配置的多個 AI 瓦片、小晶片、或晶粒的一個此種實例。在此實例中，每個垂直列可跨上文在第 6 圖至第 7 圖中描述的 K 維度分離。例如，瓦片 (0,0) 可包括 $K = 0 - 15$ 的濾波器，瓦片 (0,1) 可包括濾波器 $K = 16 - 31$ ，並且依此類推。架構中的每個水平行跨 C 維度分離，因此 HCW 0-63 可針對行 0 中的所有列廣播，HCW 64-127 可針對行 1 中的所有列廣播，並且依此類推。此可導致單個列的每個行產生具有相應 K 個分離的部分和。此等均可在單個列內減小，以減少在各個列之中分離的部分輸出張量 $P K Q$ 。因此，每一列的輸出表示總輸出張量的一部分，該部分可級聯以形成完整輸出。

【0052】 在第 9 圖中表示為節點的每個 AI 瓦片、晶粒、或小晶片可實施以使用第 5 圖中的神經網路加速器架構 500。由此，由於分區被視為具有零值並且從穿過瓦片之間的互連傳播丟棄，每個節點的輸出可減少。此導致在輸入及輸出維度上的顯著互連頻寬節省。

【0053】 第 10 圖示出了根據一些實施例的用於為神經網路層的輸出引起稀疏性的方法的流程圖 1000。此方法可藉由在上文第 5 圖中示出的神經網路加速器 500 執行。此外，分區大小 / 結構、所使用的準則、及在實施神經網路加速器的不同節點之間的路由可在如第 3 圖中描述的深度學習環境或框架中程式設計。

【0054】 方法可包括從神經網路層接收輸出 (1002)。輸出可藉由在神經網路的計算層之間添加的層來接收。此額外層可使用上文描述的分割電路及 / 或定序電路來實施。來自層的輸出可直接從計算節點及 / 或從輸出緩衝器來接收，該輸出緩衝器接收及 / 或累加來自計算節點的值。

【0055】 方法亦可包括將輸出分割為複數個分區 (1004)。可使用任何類型、大小、結構、或拓撲的分割。分割可在深度學習框架中定義並且作為神經網路圖形中的編碼或作為程式設計額外層的運行時間元資料傳遞到神經網路加速器。分割可跨空間及 / 或通道維度進行，並且可導致 2D 及 / 或 3D 分區。

【0056】 方法可額外包括識別可以視為具有零值的複數個分區中的第一分區 (1006)。第一分區可藉由對作為整體的每個分區執行準則來識別。例如，準則可係基於量值的並且可將分區內的值的集合與閾值進行比較以決定作為整體的分區中的所有值是否應當視為零。將值視為零可包括將張量中的實際值設置為 0，或丟棄或允許丟棄 (dropout) 視為零的分區而非儲存或傳播到後續層。

【0057】 方法可進一步包括產生識別在複數個分區中的剩餘第二分區之中的第一分區的位置的編碼(1008)。編碼可識別應當視為具有零值的第一分區以及其在輸出張量中的相對位置，該輸出張量具有視為具有非零值的第二分區。編碼可儲存有第二分區及/或傳遞到神經網路中的後續層或計算節點。方法可隨後亦包括將編碼及第二分區發送到神經網路中的後續層(1010)。

【0058】 應當瞭解，在第10圖中示出的具體步驟提供了根據各個實施例的為神經網路層的輸出引起稀疏性的特定方法。步驟的其他序列亦可根據替代實施例執行。例如，替代實施例可以不同次序執行上文概述的步驟。此外，第10圖中示出的獨立步驟可包括多個子步驟，該等子步驟可以適合於獨立步驟的各個序列執行。此外，額外步驟可取決於特定應用而添加或移除。許多變化、修改、及替代亦落入本揭示的範疇內。

【0059】 本文描述的方法的每一者可藉由電腦系統實施。例如，深度學習框架可在計算系統上執行。此等方法的每個步驟可藉由電腦系統自動地執行，及/或可具備涉及使用者的輸入/輸出。例如，使用者可提供輸入用於方法中的每個步驟，並且此等輸入的每一者可回應於請求此種輸入的具體輸出，其中輸出藉由電腦系統產生。每個輸入可回應於對應請求輸出而接收。此外，輸入可從使用者接收、從另一電腦系統作為資料串流接收、從記憶體位置擷取、在網路上擷取、從網服務請求、及/或類似者。同樣，輸出可

提供到使用者、另一電腦系統作為資料串流、保存在記憶體位置中、在網路上發送、提供到網服務、及/或類似者。簡而言之，本文描述的方法的每個步驟可藉由電腦系統執行，並且可涉及任何數量的輸入、輸出、及/或去往及來自電腦系統的請求，該電腦系統可能或可能不涉及使用者。不涉及使用者的彼等步驟可被認為藉由電腦系統自動地執行而無人類介入。由此，將理解，鑒於本揭示，本文描述的每種方法的每個步驟可更改為包括去往及來自使用者的輸入及輸出，或可自動地藉由電腦系統進行而無人類介入，其中任何決定可藉由處理器進行。此外，本文描述的方法的每一者的一些實施例可實施為在有形的非暫時性儲存媒體上儲存的指令集以形成有形軟體產品。

【0060】 第11圖示出了其中可實施各個實施例的示例性電腦系統1100。系統1100可用於實施上文描述的電腦系統的任一者。如圖所示，電腦系統1100包括處理單元1104，該處理單元經由匯流排子系統1102與多個週邊子系統通訊。此等週邊子系統可包括處理加速單元1106、I/O子系統1108、儲存子系統1118及通訊子系統1124。儲存子系統1118包括有形電腦可讀取儲存媒體1122及系統記憶體1110。

【0061】 匯流排子系統1102提供用於使電腦系統1100的各個部件及子系統如意欲彼此通訊的機制。儘管將匯流排子系統1102示意性圖示為單個匯流排，匯流排子系統的替代實施例可利用多個匯流排。匯流排子系統1102可係若干

類型的匯流排結構的任一者，包括使用各種匯流排架構的任一者的記憶體匯流排或記憶體控制器、週邊匯流排、或本端匯流排。例如，此種架構可包括工業標準架構 (Industry Standard Architecture; ISA) 匯流排、微通道架構 (Micro Channel Architecture; MCA) 匯流排、增強的 ISA (EISA) 匯流排、視訊電子標準協會 (Video Electronics Standards Association; VESA) 本端匯流排、及週邊部件互連 (Peripheral Component Interconnect; PCI) 匯流排，其等可以實施為根據 IEEE P1386.1 標準構造的 Mezzanine 匯流排。

【0062】 可以實施為一或多個積體電路（例如，習知微處理器或微控制器）的處理單元 1104 控制電腦系統 1100 的操作。一或多個處理器可包括在處理單元 1104 中。此等處理器可包括單核或多核處理器。在某些實施例中，處理單元 1104 可實施為一或多個獨立處理單元 1132 及 / 或 1134，其中在每個處理單元中包括單核或多核處理器。在其他實施例中，處理單元 1104 亦可實施為藉由將兩個雙核處理器整合到單個晶片中來形成的四核處理單元。

【0063】 在各個實施例中，處理單元 1104 可以回應於程式碼而執行各種程式，並且可以維持多個並發執行的程式或過程。於任何給定時間，待執行的一些或全部程式碼可以駐留在處理器 1104 中及 / 或儲存子系統 1118 中。經由適宜的程式設計，處理器 1104 可以提供上文描述的各種功能性。電腦系統 1100 可額外包括處理加速單元 1106，該處

理加速單元可以包括數位信號處理器(digital signal processor; DSP)、特殊應用處理器、及/或類似者。

【0064】 I/O子系統1108可包括使用者介面輸入裝置及使用者介面輸出裝置。使用者介面輸入裝置可包括鍵盤、指向裝置(諸如滑鼠或軌跡球)、整合到顯示器中的觸控板或觸控螢幕、滾輪、點擊輪、撥號盤、按鈕、開關、小鍵盤、具有語音命令識別系統的音訊輸入裝置、麥克風、及其他類型的輸入裝置。使用者介面輸入裝置可包括例如諸如Microsoft Kinect®運動感測器的運動感測及/或姿勢識別裝置,該等裝置使使用者能夠使用姿勢及口頭命令經由自然使用者介面控制輸入裝置並且與輸入裝置交互,該輸入裝置諸如Microsoft Xbox® 360遊戲控制器。使用者介面輸入裝置亦可包括諸如Google Glass®眨眼偵測器的眼睛姿勢識別裝置,該等裝置偵測來自使用者的眼睛活動(例如,在拍照及/或進行菜單選擇時「眨眼」),並且將眼睛姿勢轉換為到輸入裝置(例如,Google Glass®)中的輸入。此外,使用者介面輸入裝置可包括使使用者能夠經由語音命令與語音識別系統(例如,Siri®導航儀)交互的語音識別感測裝置。

【0065】 使用者介面輸入裝置亦可包括但不限於三維(3D)滑鼠、操縱桿或指示棒、遊戲鍵盤及繪圖平板、以及音訊/視覺裝置,諸如揚聲器、數位相機、數位可攜式相機、可攜式媒體播放機、網路攝像頭、影像掃描儀、指紋掃描儀、條形碼讀取器3D掃描儀、3D印表機、雷射測距儀、

及眼睛注視追蹤裝置。此外，使用者介面輸入裝置可包括例如醫學成像輸入裝置，諸如計算斷層攝影術、磁諧振成像、位置發射斷層攝影術、醫學超聲檢查裝置。使用者介面輸入裝置亦可包括例如音訊輸入裝置，諸如MIDI鍵盤、數位樂器及類似者。

【0066】 使用者介面輸出裝置可包括顯示器子系統、指示器燈、或非視覺顯示器，諸如音訊輸出裝置等。顯示子系統可係陰極射線管(cathode ray tube; CRT)、平板裝置(諸如使用液晶顯示器(liquid crystal display; LCD)或電漿顯示器的平板裝置)、投影裝置、觸控螢幕、及類似者。通常，使用術語「輸出裝置」意欲包括用於將來自電腦系統1100的資訊輸出到使用者或其他電腦的所有可能類型的裝置及機構。例如，使用者介面輸出裝置可包括但不限於視覺地傳達文本、圖形及音訊/視訊資訊的各種顯示裝置，諸如監控器、印表機、揚聲器、頭戴式耳機、自動導航系統、繪圖儀、語音輸出裝置、及數據機。

【0067】 電腦系統1100可包含儲存子系統1118，該儲存子系統包含圖示為當前位於系統記憶體1110中的軟體元件。系統記憶體1110可儲存在處理單元1104上可載入且可執行的程式指令，以及在執行此等程式期間產生的資料。

【0068】 取決於電腦系統1100的配置及類型，系統記憶體1110可係揮發性的(諸如隨機存取記憶體(random access memory; RAM))及/或非揮發性的(諸如唯讀記憶體(read-only memory; ROM)、快閃記憶體等)。

R A M 通常含有可由處理單元 1104 立即存取及 / 或當前正由處理單元 1104 操作及執行的資料及 / 或程式模組。在一些實施方式中，系統記憶體 1110 可包括多種不同類型的記憶體，諸如靜態隨機存取記憶體 (static random access memory; S R A M) 或動態隨機存取記憶體 (dynamic random access memory; D R A M)。在一些實施方式中，含有諸如在開始期間有助於在電腦系統 1100 內的元件之間傳遞資訊的基本常式的基本輸入 / 輸出系統 (basic input/output system; B I O S) 可通常儲存在 R O M 中。舉例而言，並且不作限制，系統記憶體 1110 亦示出應用程式 1112 (可包括客戶端應用、網瀏覽器、中層應用、有關資料庫管理系統 (relational database management system; R D B M S) 等)、程式資料 1114、及作業系統 1116。舉例而言，作業系統 1116 可包括各種版本的 Microsoft Windows®、Apple Macintosh®、及 / 或 Linux 作業系統，各種商業可用的 UNIX® 或類似 UNIX 的作業系統 (包括不作限制各種 GNU/Linux 作業系統、Google Chrome® OS、及類似者) 及 / 或行動作業系統，諸如 iOS、Windows® Phone、Android® OS、BlackBerry® 10 OS、及 Palm® OS 作業系統。

【0069】 儲存子系統 1118 亦可提供用於儲存提供一些實施例的功能性的基本程式設計及資料構造的有形電腦可讀取儲存媒體。當藉由處理器執行時提供上文描述的功能性的軟體 (程式、代碼模組、指令) 可在儲存子系統 1118 中

儲存。此等軟體模組或指令可藉由處理單元 1104 執行。儲存子系統 1118 亦可提供用於儲存根據一些實施例使用的資料的儲存庫。

【0070】 儲存子系統 1100 亦可包括電腦可讀取儲存媒體讀取器 1120，其可以進一步連接到電腦可讀取儲存媒體 1122。一起並且視情況，與系統記憶體 1110 結合，電腦可讀取儲存媒體 1122 可綜合表示遠端、本端、固定、及/或可移除儲存裝置加上用於暫時性及/或更永久地含有、儲存、發送、及擷取電腦可讀取資訊的儲存媒體。

【0071】 含有代碼、或代碼的部分的電腦可讀取儲存媒體 1122 亦可以包括任何適當媒體，包括儲存媒體及通訊媒體，諸如但不限於在任何方法或技術中實施用於儲存及/或傳輸資訊的揮發性及非揮發性、可移除及不可移除媒體。此可以包括有形電腦可讀取儲存媒體，諸如 RAM、ROM、電子可抹除可程式設計 ROM (electronically erasable programmable ROM; EEPROM)、快閃記憶體或其他記憶體技術、CD-ROM、數位多功能光碟 (digital versatile disk; DVD)、或其他光學儲存、盒式磁帶、磁帶、磁碟儲存器或其他磁儲存裝置、或其他有形電腦可讀取媒體。此亦可以包括非有形電腦可讀取媒體，諸如資料信號、資料傳輸、或可以用於發送期望資訊並且可以藉由計算系統 1100 存取的任何其他媒體。

【0072】 舉例而言，電腦可讀取儲存媒體 1122 可包括從不可移除的非揮發性磁性媒體讀取或寫入不可移除的非揮發

性磁性媒體的硬碟驅動器，從可移除的非揮發性磁碟讀取或寫入可移除的非揮發性磁碟的磁碟驅動器，及從可移除的非揮發性光碟（諸如CD ROM、DVD、Blu-Ray®光碟、或其他光學媒體）讀取或寫入可移除的非揮發性光碟的光碟驅動器。電腦可讀取儲存媒體1122可包括但不限於Zip®驅動器、快閃記憶卡、通用串列匯流排(universal serial bus; USB)快閃驅動器、安全數位(secure digital; SD)卡、DVD磁碟、數位視訊磁帶、及類似者。電腦可讀取儲存媒體1122亦可包括：基於非揮發性記憶體的固態驅動器(solid-state drive; SSD)，諸如基於快閃記憶體的SSD、企業快閃驅動器、固態ROM、及類似者；基於揮發性記憶體的SSD，諸如固態RAM、動態RAM、靜態RAM、基於DRAM的SSD、磁阻RAM(MRAM)SSD；及使用DRAM及基於快閃記憶體的SSD的組合的混合SSD。磁碟驅動器及其相關聯的電腦可讀取媒體可提供電腦可讀取指令、資料結構、程式模組、及用於電腦系統1100的其他資料的非揮發性儲存器。

【0073】 通訊子系統1124向其他電腦系統及網路提供介面。通訊子系統1124用作從來自電腦系統1100的其他系統接收資料並且向其他系統發送資料的介面。例如，通訊子系統1124可使電腦系統1100能夠經由網際網路連接到一或多個裝置。在一些實施例中，通訊子系統1124可以包括用於存取無線語音及/或資料網路的射頻(RF)收發機部件（例如，使用蜂巢電話技術，先進資料網路技術，諸如

3 G、4 G 或 E D G E (全 球 演 進 增 強 資 料 率))、W i F i (I E E E 8 0 2 . 1 1 族 標 準、或 其 他 行 動 通 訊 技 術、或 其 任 何 組 合)、全 球 定 位 系 統 (g l o b a l p o s i t i o n i n g s y s t e m ; G P S) 接 收 器 部 件、及 / 或 其 他 部 件。在 一 些 實 施 例 中，除 了 或 替 代 無 線 介 面，通 訊 子 系 統 1 1 2 4 可 以 提 供 有 線 網 路 連 接 性 (例 如，乙 太 網 路)。

【0074】 在 一 些 實 施 例 中，通 訊 子 系 統 1 1 2 4 亦 可 代 表 可 使 用 電 腦 系 統 1 1 0 0 的 一 或 多 個 使 用 者 接 收 呈 結 構 化 及 / 或 非 結 構 化 的 資 料 饋 送 1 1 2 6、事 件 串 流 1 1 2 8、事 件 更 新 1 1 3 0、及 類 似 者 的 形 式 的 輸 入 通 訊。

【0075】 舉 例 而 言，通 訊 子 系 統 1 1 2 4 可 經 配 置 為 從 社 會 網 路 及 / 或 其 他 通 訊 服 務 的 使 用 者 即 時 接 收 資 料 饋 送 1 1 2 6，諸 如 T w i t t e r® 饋 送、F a c e b o o k® 更 新、網 饋 送 (諸 如 豐 富 位 點 摘 要 (R i c h S i t e S u m m a r y ; R S S) 饋 送) 及 / 或 來 自 一 或 多 個 第 三 方 資 訊 源 的 即 時 更 新。

【0076】 此 外，通 訊 子 系 統 1 1 2 4 亦 可 經 配 置 為 接 收 呈 連 續 資 料 串 流 形 式 的 資 料，該 等 資 料 串 流 可 包 括 即 時 事 件 的 事 件 串 流 1 1 2 8 及 / 或 本 質 上 可 係 連 續 或 無 邊 而 無 明 確 結 束 的 事 件 更 新 1 1 3 0。產 生 連 續 資 料 的 應 用 的 實 例 可 包 括 例 如 感 測 器 資 料 應 用、金 融 票 據、網 路 效 能 量 測 工 具 (例 如，網 路 監 控 及 流 量 管 理 應 用)、點 選 串 流 分 析 工 具、自 動 流 量 監 控、及 類 似 者。

【0077】 通 訊 子 系 統 1 1 2 4 亦 可 經 配 置 為 將 結 構 化 及 / 或 未 結 構 化 的 資 料 饋 送 1 1 2 6、事 件 串 流 1 1 2 8、事 件 更 新

1130、及類似者輸出到可與耦接到電腦系統1100的一或多個串流資料源電腦通訊的一或多個資料庫。

【0078】 電腦系統1100可以係各種類型之一，包括手持可攜式裝置（例如，iPhone®蜂巢電話、iPad®計算平板、PDA）、可穿戴裝置（例如，Google Glass®頭戴式顯示器）、PC、工作站、主機、資訊亭(kiosk)、伺服器機架、或任何其他資料處理系統。

【0079】 歸因於電腦及網路的不斷改變的性質，對在圖式中描繪的電腦系統1100的描述僅意欲作為具體實例。具有與圖式中描繪的系統相比更多或更少的部件的許多其他配置係可能的。例如，定製硬體亦可能使用及/或特定元件可能在硬體、韌體、軟體（包括小型應用程式）、或組合中實施。另外，可採用到其他計算裝置（諸如網路輸入/輸出裝置）的連接。基於本文提供的揭示及教示，實施各個實施例的其他方式及/或方法應當係顯而易見的。

【0080】 在以上描述中，出於解釋的目的，闡述數個具體細節以便提供對各個實施例的透徹理解。然而，將顯而易見，可在沒有此等具體細節中的一些細節的情況下實踐一些實施例。在其他實例中，熟知結構及裝置以方塊圖的形式圖示。

【0081】 以上描述僅提供了示例性實施例，並且不意欲限制本揭示的範疇、適用性、或構造。而是，各個實施例的以上描述將提供用於實施至少一個實施例的賦能揭示。應

當理解，可作出元件的功能及佈置的各種改變而不脫離如在隨附申請專利範圍中闡述的一些實施例的精神及範疇。

【0082】 具體細節在以上描述中給出以提供本揭示的透徹理解。然而，將理解，實施例可在沒有此等具體細節的情況下實踐。例如，電路、系統、網路、過程、及其他部件可以方塊圖的形式圖示為部件，以便不以非必要的細節混淆實施例。在其他實例中，熟知的電路、過程、演算法、結構、及技術可在沒有非必要細節的情況下圖示以便避免混淆實施例。

【0083】 此外，注意到，獨立實施例可經描述為過程，該過程被描繪為流程圖 (flowchart)、流程圖 (flow diagram)、資料流圖、結構圖、或方塊圖。儘管流程圖可將操作描述為連續過程，許多操作可以並行或同時執行。此外，可重新佈置操作的順序。過程當其操作完成時可終止，但可以具有圖式中不包括的額外步驟。過程可對應於方法、函數、程序、子常式、子程式等。當過程對應於函數時，其終止可以對應於函數返回到調用函數或主函數。

【0084】 術語「電腦可讀取媒體」包括但不限於可攜式或固定儲存裝置、光學儲存裝置、無線通道及能夠儲存、含有、或攜帶指令及/或資料的各種其他媒體。代碼段或機器可執行指令可表示程序、功能、子程式、程式、常式、子常式、模組、軟體封裝、類別、或指令、資料結構、或程式語句的任何組合。代碼段可藉由傳遞及/或接收資訊、資料、引數、參數、或記憶體內容耦接到另一代碼段或硬體

電路。資訊、引數、參數、資料等可經由任何適宜構件傳遞、轉發、或發送，包括記憶體共用、訊息傳遞、符記傳遞、網路傳輸等。

【0085】 此外，實施例可藉由硬體、軟體、韌體、中介軟體、微代碼、硬體描述語言、或其任何組合來實施。當在軟體、韌體、中介軟體或微代碼中實施時，用於執行必要任務的程式碼或代碼段可儲存在機器可讀取媒體中。處理器可執行必要任務。

【0086】 在以上說明書中，特徵參考其具體實施例描述，但應當認識到，並非所有實施例限於此。一些實施例的各個特徵及態樣可獨立地或聯合地使用。另外，實施例可以在超出本文描述的彼等的任何數量的環境及應用中利用而不脫離說明書的較寬精神及範疇。說明書及附圖由此被認為係說明性而非限制性。

【0087】 此外，出於說明的目的，方法以特定次序描述。應當瞭解，在替代實施例中，方法可以與所描述者不同的次序執行。亦應當瞭解，上文描述的方法可藉由硬體部件執行或可體現在機器可執行指令的序列中，該等指令可用於導致用指令程式設計機器（諸如通用或專用處理器或邏輯電路）以執行方法。此等機器可執行指令可儲存在一或多個機器可讀取媒體上，諸如CD-ROM或其他類型的光碟、軟碟、ROM、RAM、EPROM、EEPROM、磁卡或光卡、快閃記憶體、或適用於儲存電子指令的其他類型的機器可讀取媒體。或者，方法可由硬體及軟體的組合執行。

【符號說明】

【0088】

1 0 0 : 圖 形

2 0 0 : 圖 表

3 0 0 : 圖

3 0 2 : 深 度 學 習 框 架

3 0 4 : 程 式 庫

3 0 6 : 神 經 網 路 圖 形

3 0 8 : 神 經 網 路 運 行 時 間

3 1 0 : 硬 體

4 0 0 : 通 用 神 經 網 路 加 速 器

4 0 2 : 晶 片 上 記 憶 體

4 0 4 : 晶 片 上 S R A M

4 0 6 : 執 行 單 元

4 0 8 : 內 部 輸 入 緩 衝 器

4 1 0 : 計 算 節 點

4 1 2 : 輸 出 緩 衝 器

5 0 0 : 神 經 網 路 加 速 器

5 0 2 : 定 序 器 電 路

5 0 4 : 分 割 電 路

6 0 2 : 輸 入 張 量

6 0 4 : 濾 波 器

6 0 6 : 輸 出 張 量

7 0 2 - 1 : 非 稀 疏 分 區

7 0 2 - 2 : 非 稀 疏 分 區

7 0 2 - 3 : 非 稀 疏 分 區

7 0 2 - 4 : 非 稀 疏 分 區

7 0 4 : 元 資 料

8 0 2 : 啟 動 映 射

8 0 4 : 啟 動 映 射

1 0 0 0 : 流 程 圖

1 0 0 2 : 步 驟

1 0 0 4 : 步 驟

1 0 0 6 : 步 驟

1 0 0 8 : 步 驟

1 0 1 0 : 步 驟

1 1 0 0 : 系 統

1 1 0 2 : 匯 流 排 子 系 統

1 1 0 4 : 處 理 單 元

1 1 0 6 : 處 理 加 速 單 元

1 1 0 8 : I / O 子 系 統

1 1 1 0 : 系 統 記 憶 體

1 1 1 2 : 應 用 程 式

1 1 1 4 : 程 式 資 料

1 1 1 6 : 作 業 系 統

1 1 1 8 : 儲 存 子 系 統

1 1 2 0 : 電 腦 可 讀 取 儲 存 媒 體 讀 取 器

1 1 2 2 : 有 形 電 腦 可 讀 取 儲 存 媒 體

1 1 2 4 : 通 訊 子 系 統

1 1 2 6 : 資 料 饋 送

1 1 2 8 : 事 件 串 流

1 1 3 0 : 事 件 更 新

1 1 3 2 : 處 理 單 元

1 1 3 4 : 處 理 單 元

C : 輸 入 通 道

H : 高 度

W : 寬 度

【生物材料寄存】

國 內 寄 存 資 訊 (請 依 寄 存 機 構 、 日 期 、 號 碼 順 序 註 記)

無

國 外 寄 存 資 訊 (請 依 寄 存 國 家 、 機 構 、 日 期 、 號 碼 順 序 註 記)

無

【發明申請專利範圍】

【請求項 1】 一種為神經網路層的輸出引起稀疏性的方法，
該方法包含以下步驟：

從一神經網路的一層接收輸出；

將該等輸出分割為複數個分區；

識別可以視為具有零值的該複數個分區中的第一分區；

產生識別在該複數個分區中的剩餘第二分區之中的該等第一分區的位置的一編碼；以及

將該編碼及該等第二分區發送到該神經網路中的一後續層。

【請求項 2】 如請求項 1 所述的方法，進一步包含以下步驟：

在該神經網路中的該後續層處接收該等第二分區；以及

基於該編碼來佈置該等第二分區。

【請求項 3】 如請求項 2 所述的方法，其中該後續層執行一乘法運算，藉此該等第一分區可以作為一乘以零運算而丟棄。

【請求項 4】 如請求項 1 所述的方法，其中該等輸出包含來自該層的輸出的一三維陣列，其中輸出的該陣列包含該神經網路中的不同通道的一維度。

【請求項 5】 如請求項 4 所述的方法，其中該複數個分區包含輸出的該陣列的三維分區。

【請求項6】 如請求項1所述的方法，其中該等分區在該複數個分區中係不連續的。

【請求項7】 如請求項1所述的方法，其中識別可以視為具有零值的該複數個分區中的該等第一分區之步驟包含以下步驟：

從一設計環境接收一準則；以及

將該準則應用於該複數個分區的每一者。

【請求項8】 如請求項7所述的方法，其中該準則包含一相對量值函數計算一分區中的該等值的一集合，並且若該集合小於一閾值，則將該分區中的該等值設置為零。

【請求項9】 如請求項7所述的方法，其中作為一運行時間函數從該設計環境發送該準則。

【請求項10】 如請求項7所述的方法，其中作為表示該神經網路的一圖形的部分來編碼該準則。

【請求項11】 一種神經網路加速器，包含：

一計算節點，經配置為實施一神經網路的一層並且從該層產生輸出；

一分割電路，經配置為執行包含下列的操作：

從一神經網路的該層接收輸出；

將該等輸出分割為複數個分區；

識別可以視為具有零值的該複數個分區中的第一分區；以及

產生識別在該複數個分區中的剩餘第二分區之中的

該等第一分區的位置的一編碼；以及

一記憶體，經配置為儲存該編碼及該等第二分區用於該神經網路中的一後續層。

【請求項12】如請求項11所述的神經網路加速器，進一步包含複數個小晶片，其中該計算節點在該複數個小晶片中的一第一小晶片上實施，並且其中該後續層在該複數個小晶片中的一第二小晶片上實施。

【請求項13】如請求項11所述的神經網路加速器，進一步包含經配置為執行包含下列的操作的一定序器電路：
在該神經網路中的該後續層處接收該等第二分區；以及

基於該編碼來佈置該等第二分區。

【請求項14】如請求項11所述的神經網路加速器，其中該神經網路的該層包含執行一卷積核心。

【請求項15】如請求項11所述的神經網路加速器，其中該記憶體包含一晶片上靜態隨機存取記憶體（SRAM）。

【請求項16】如請求項11所述的神經網路加速器，其中當訓練該神經網路時不使用該分割電路。

【請求項17】如請求項11所述的神經網路加速器，其中在訓練該神經網路期間決定該複數個分區中的分區的一數量。

【請求項18】如請求項11所述的神經網路加速器，其中識別可以視為具有零值的該複數個分區中的該等第一分區包含：

從一設計環境接收一準則；以及

將該準則應用於該複數個分區的每一者。

【請求項19】如請求項11所述的神經網路加速器，其中該等輸出包含來自該層的輸出的一三維陣列，其中輸出的該陣列包含該神經網路中的不同通道的一維度，並且其中該複數個分區包含輸出的該陣列的三維分區。

【請求項20】一種為神經網路層的輸出引起稀疏性的方法，該方法包含：

從一神經網路的一層接收輸出；

將該等輸出分割為複數個分區，其中該複數個分區的每一者包含複數個該等輸出；

識別滿足一準則的該複數個分區中的第一分區，該準則指示該等第一分區中的值可設置為零；

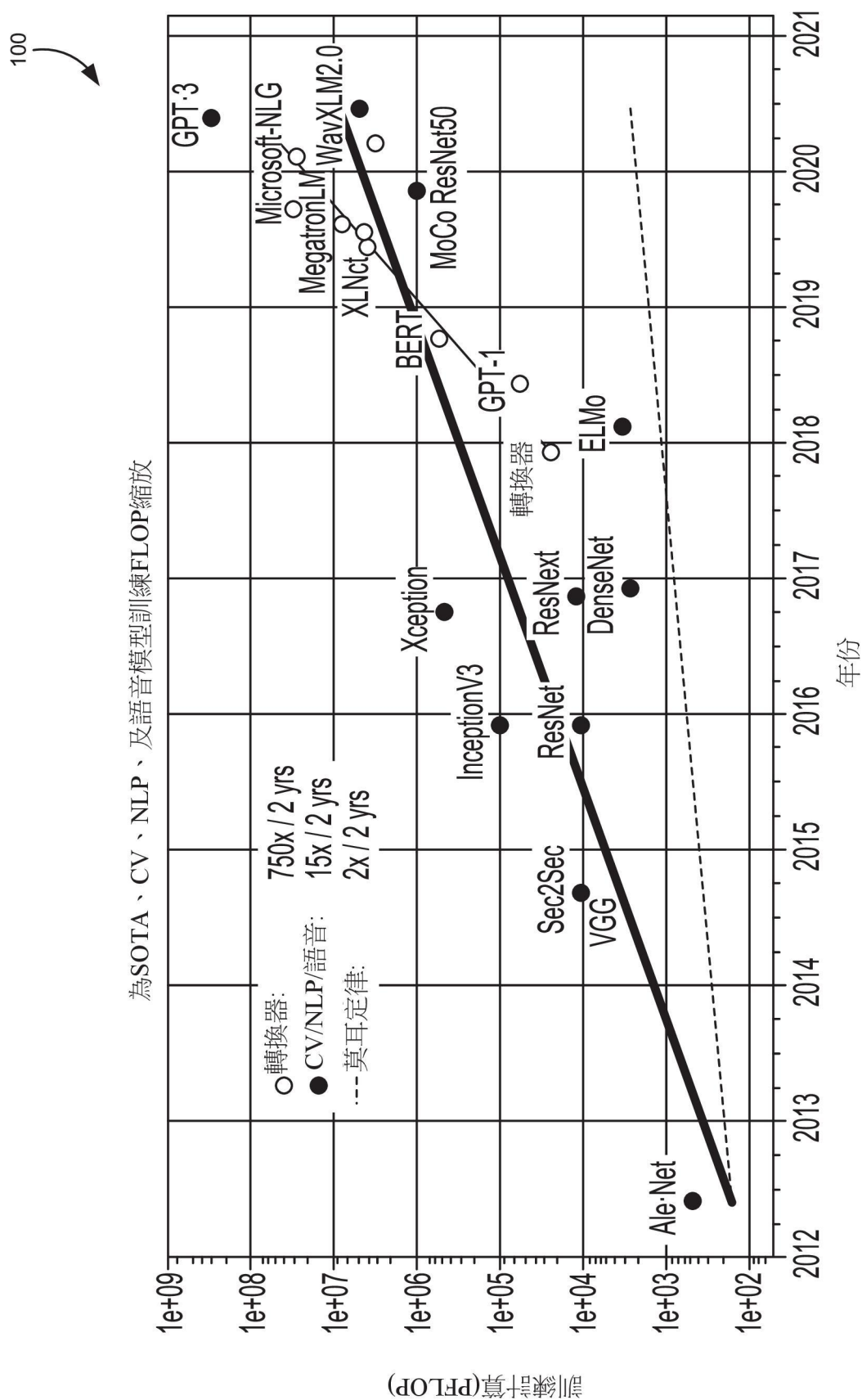
產生識別在該複數個分區中的剩餘第二分區之中的該等第一分區的位置的一編碼；

將該編碼及該等第二分區發送到該神經網路中的一後續層並且丟棄該等第一分區；

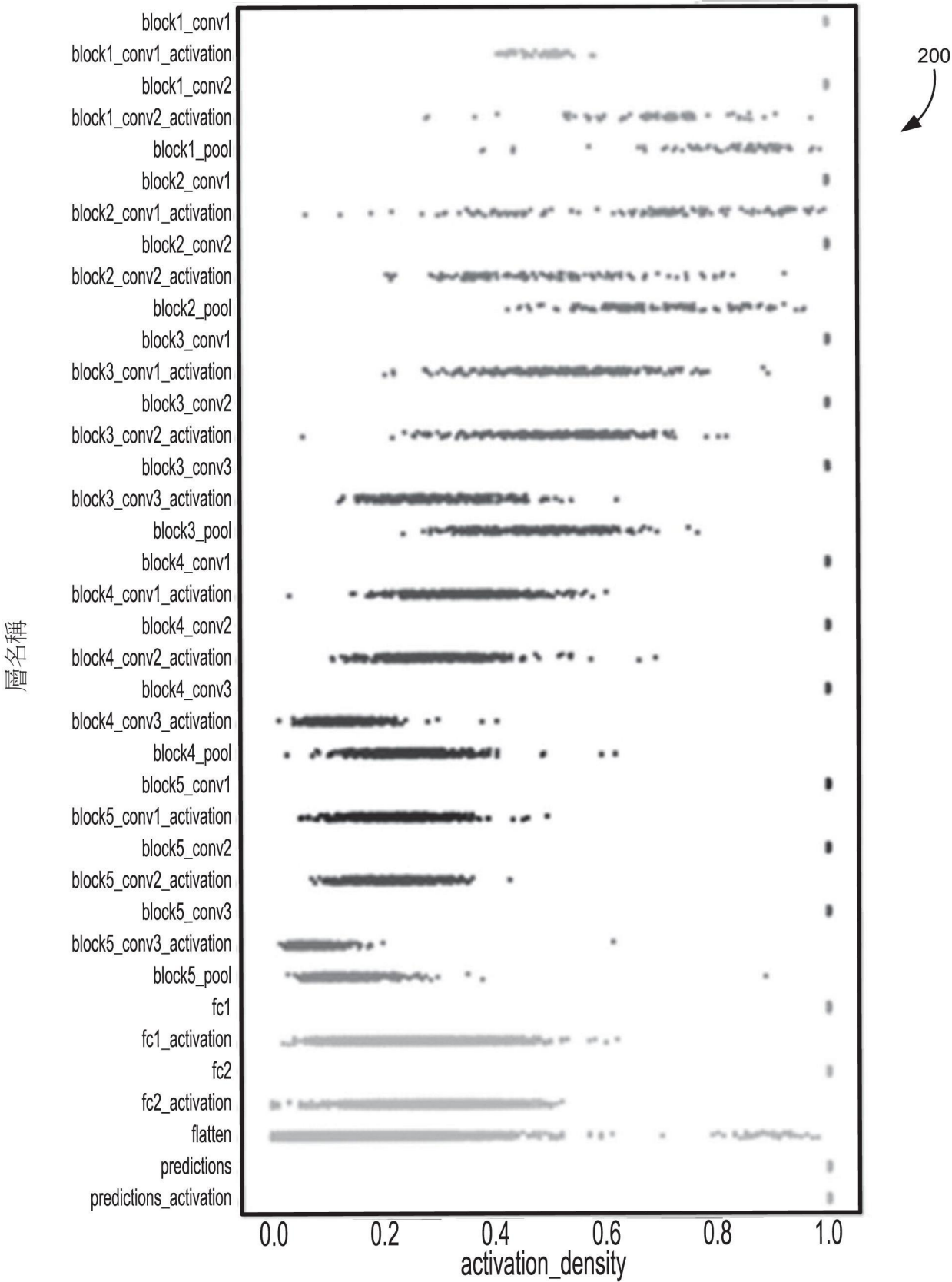
在該神經網路中的該後續層處接收該等第二分區；

基於該編碼來佈置具有零值的該等第二分區；以及

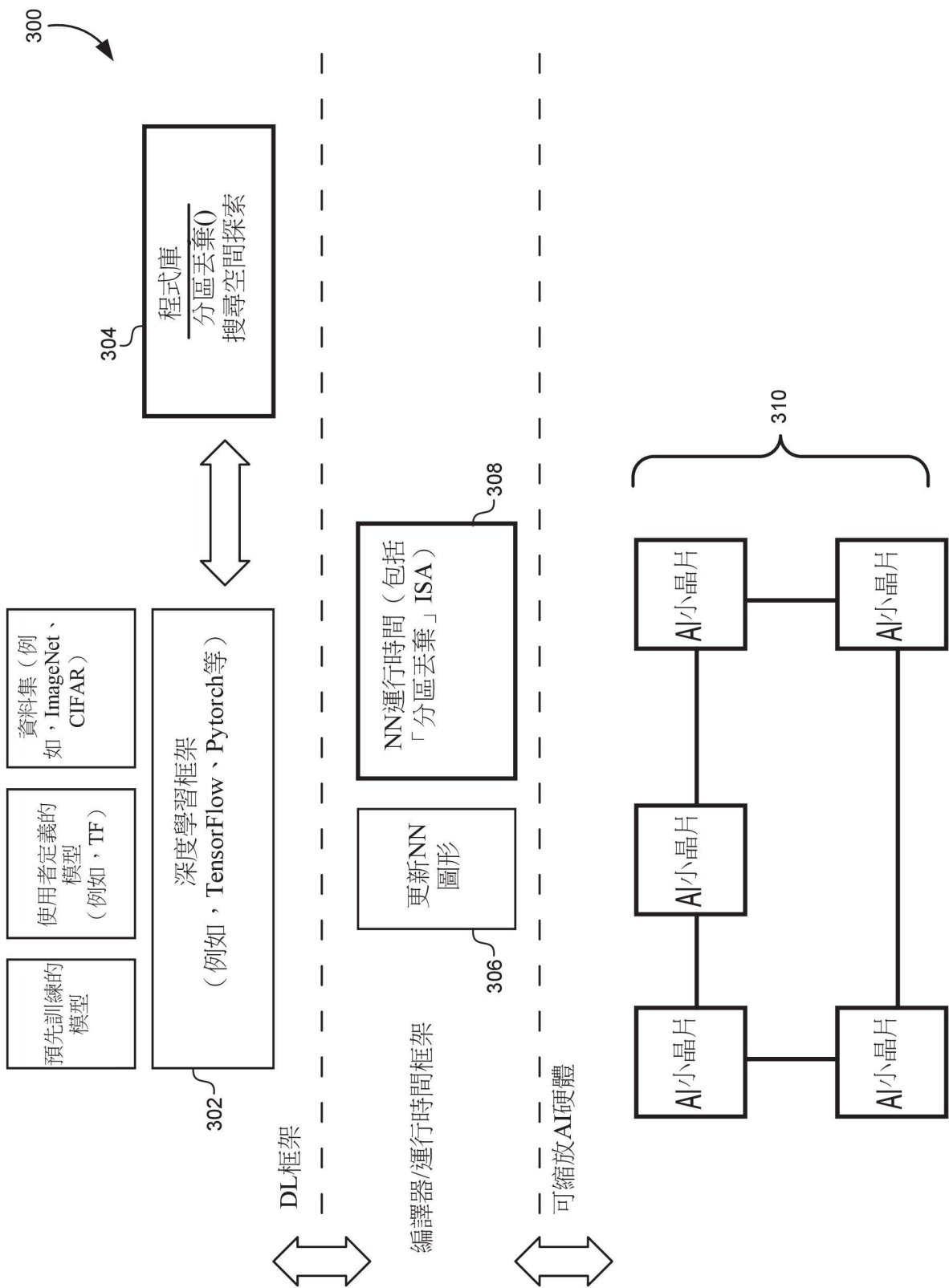
執行該神經網路中的該後續層。



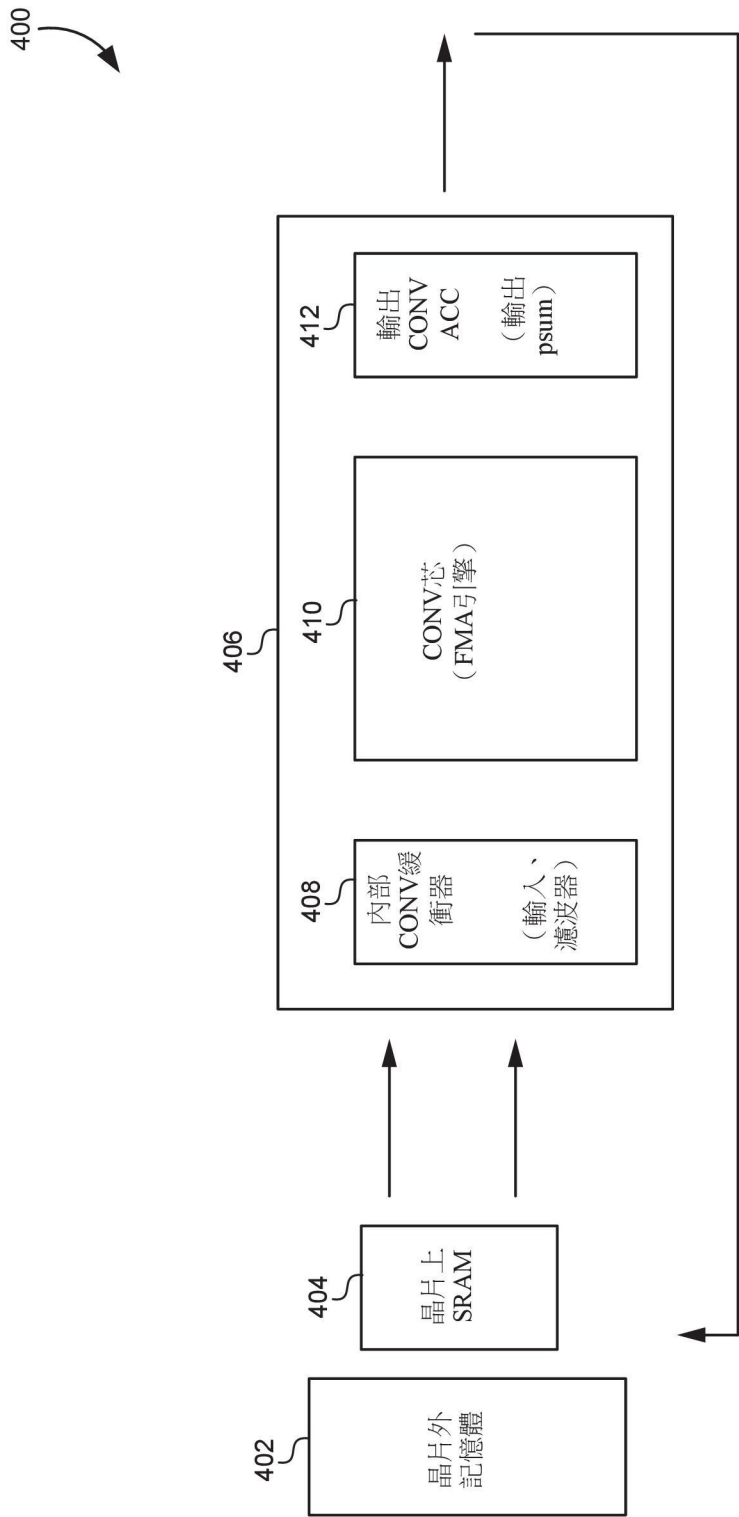
逐通道啟動密度的分佈



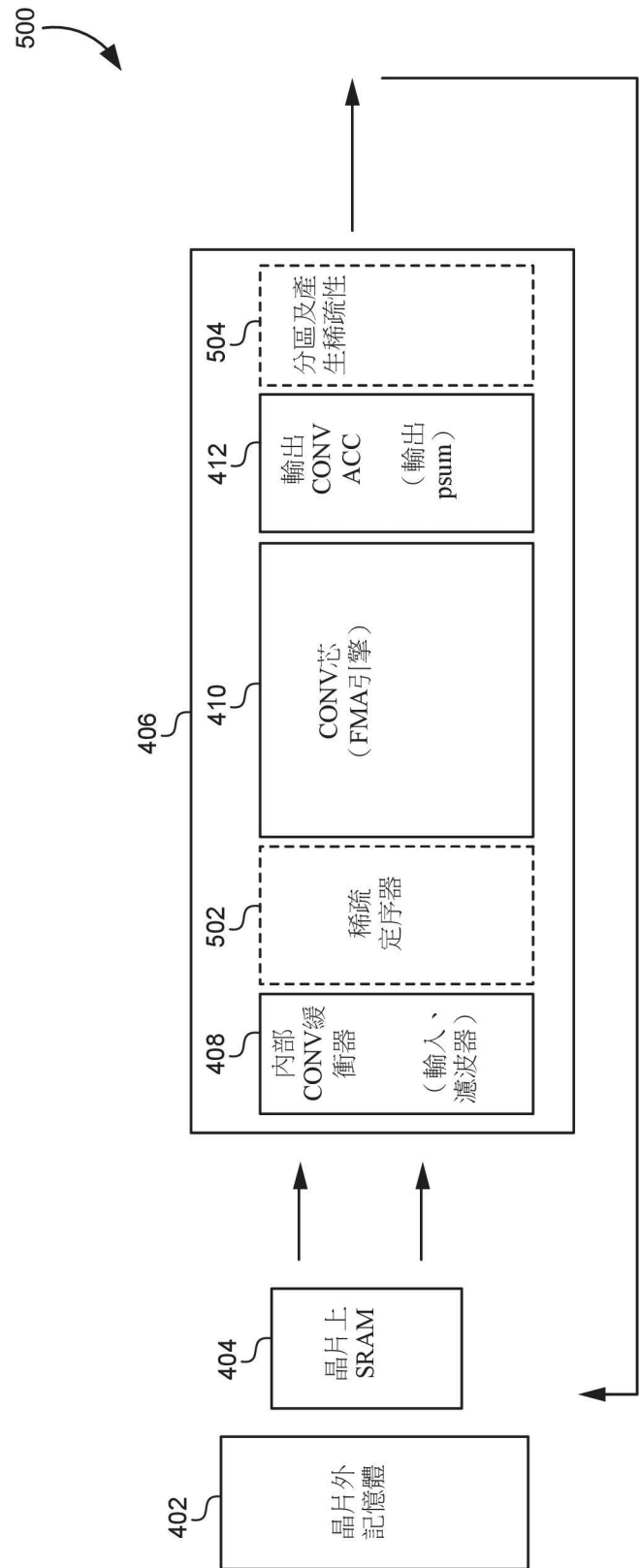
第2圖



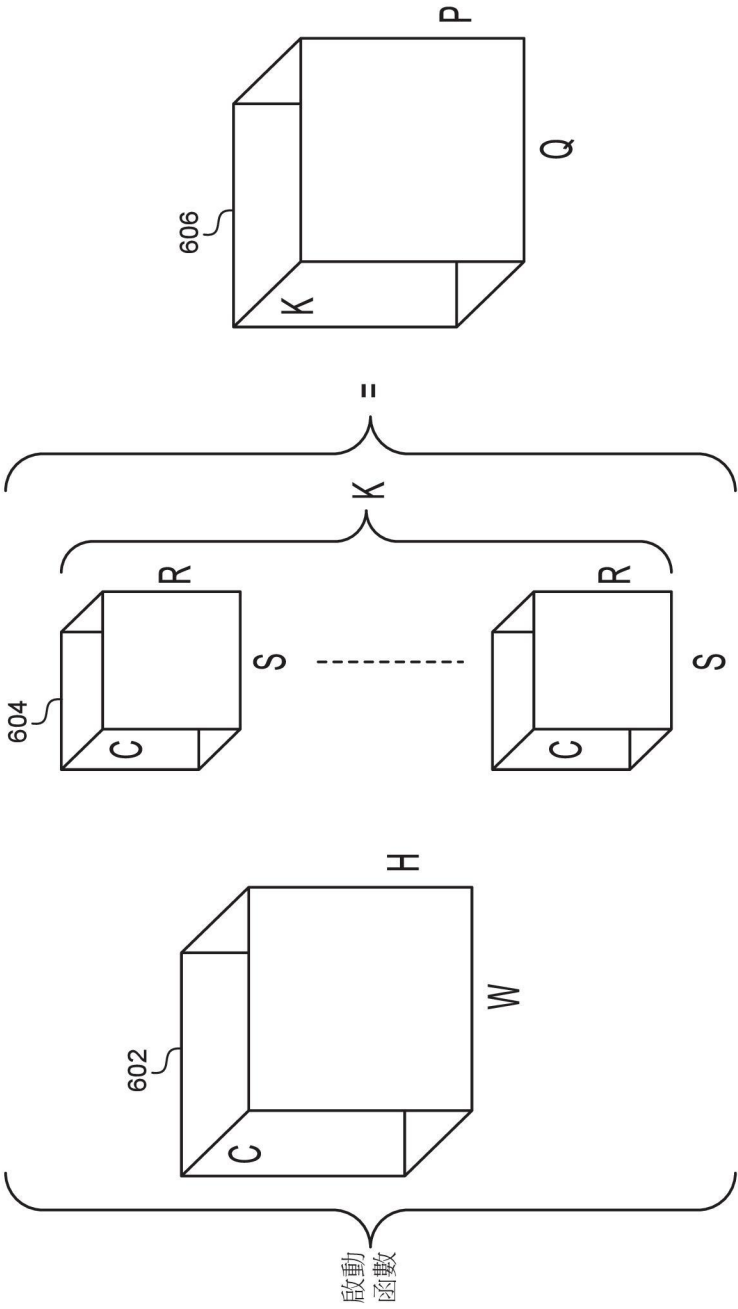
第3圖



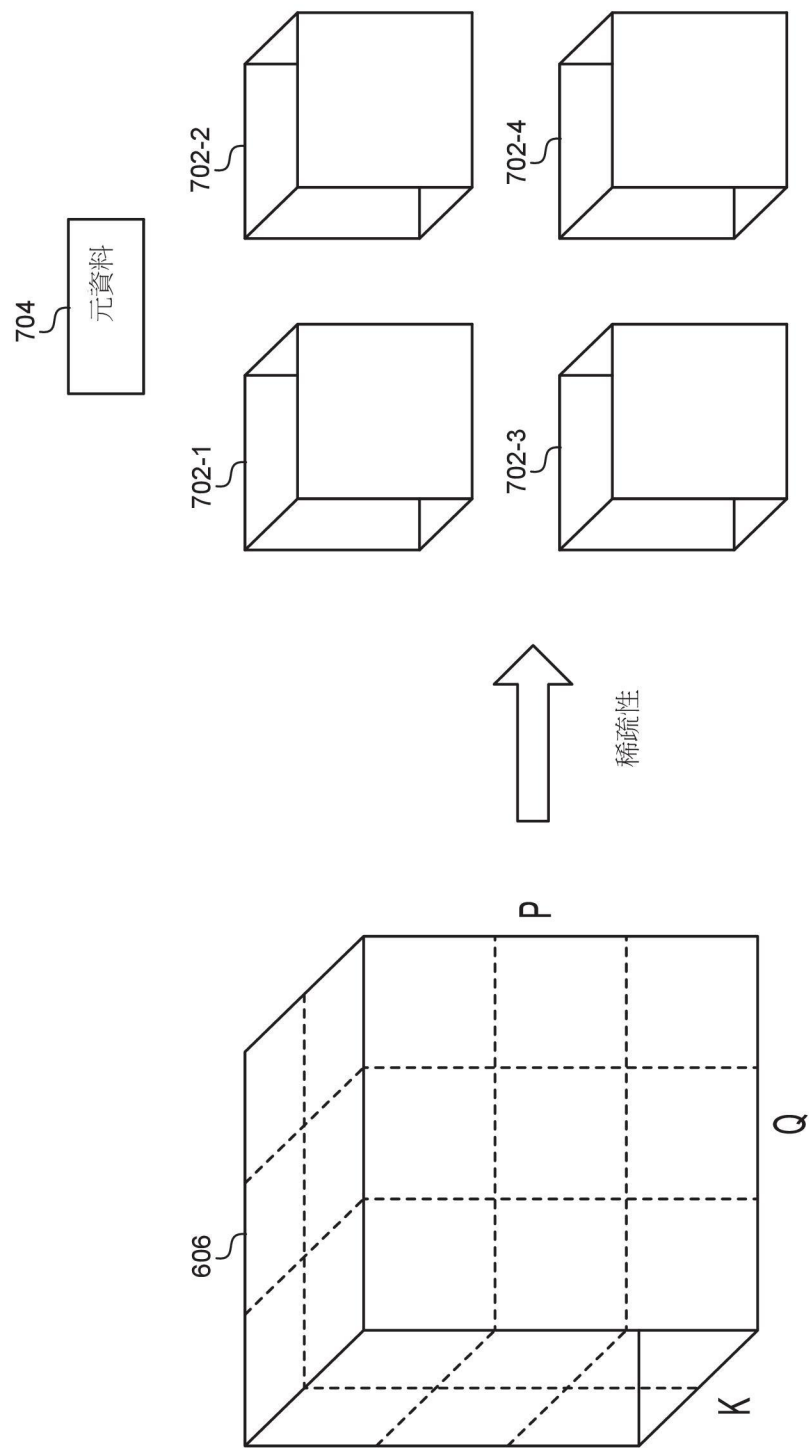
第4圖



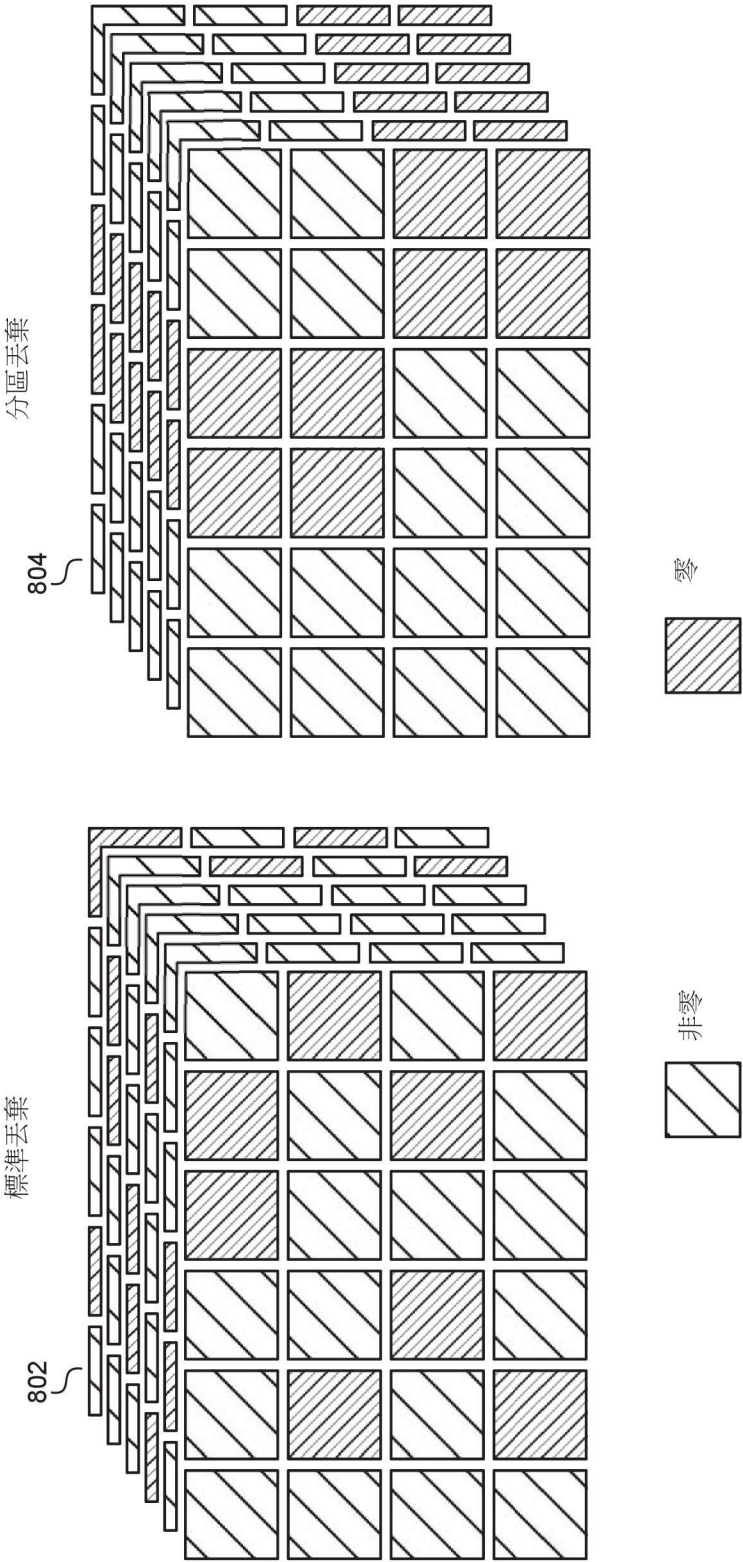
第5圖



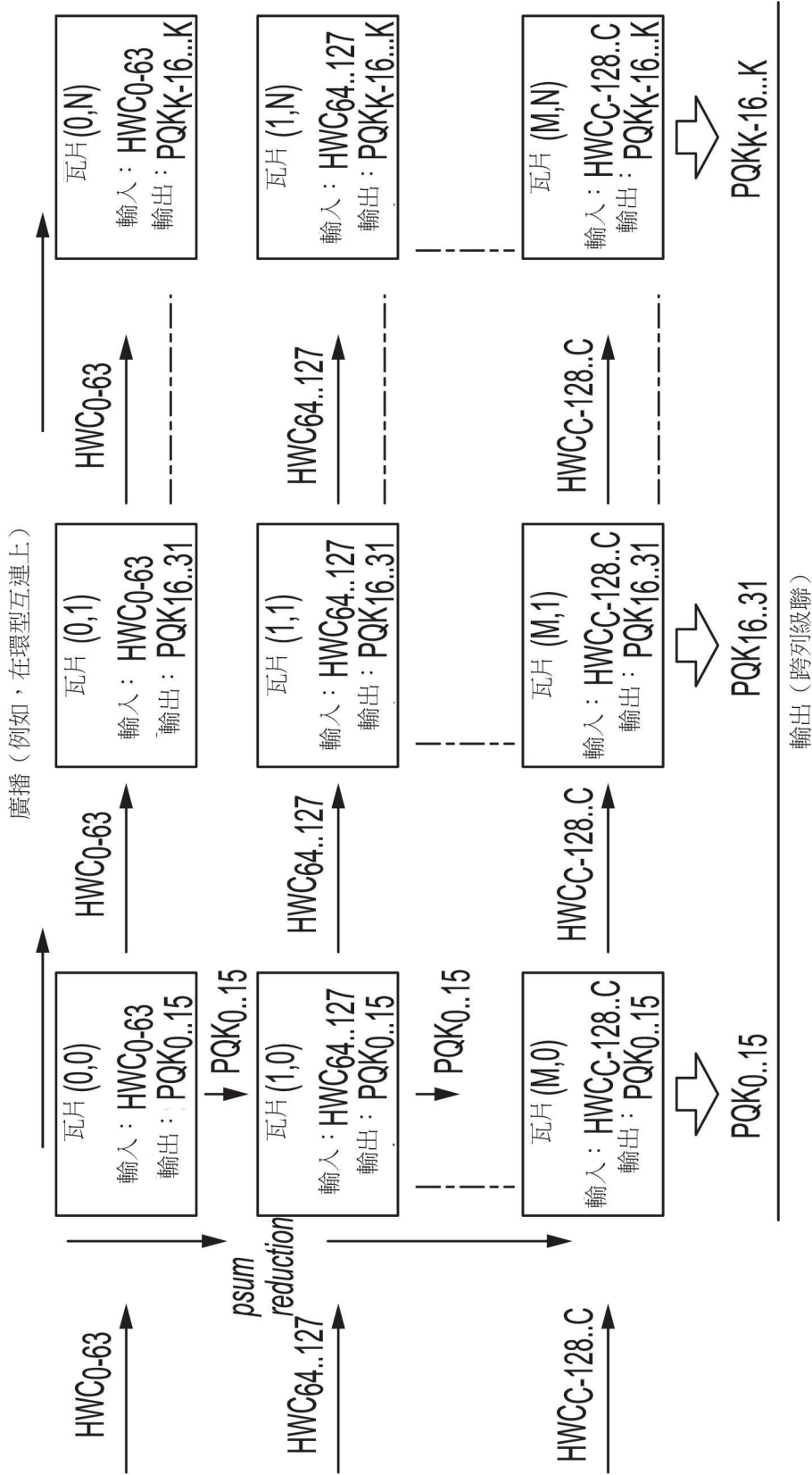
第6圖



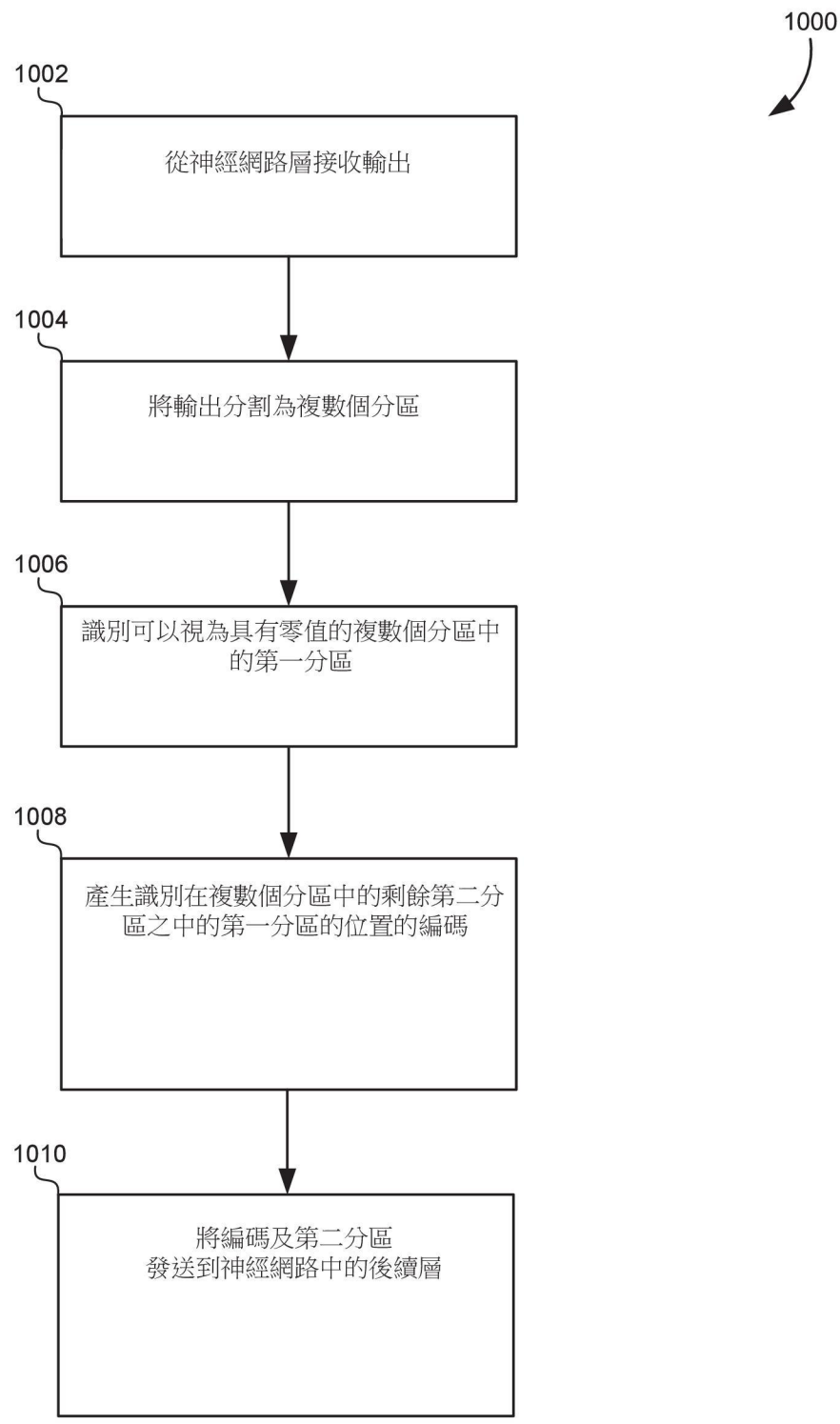
第7圖



第8圖



第9圖



第10圖



第11圖