



(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 I622939 B

(45)公告日：中華民國 107 (2018) 年 05 月 01 日

(21)申請案號：106111273

(22)申請日：中華民國 105 (2016) 年 05 月 20 日

(51)Int. Cl. : G06N3/02 (2006.01)

(30)優先權：2015/05/21 美國 62/165,020

2015/09/03 美國 14/844,431

(71)申請人：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)

(74)代理人：陳長文

(56)參考文獻：

TW 201232429

TW 201331855

US 2014/0142929A1

US 2015/0066499A1

審查人員：游象甫

申請專利範圍項數：27 項 圖式數：6 共 36 頁

## (54)名稱

用於執行類神經網路之類神經網路計算之方法、系統及電腦可讀取媒體

METHOD, SYSTEM AND COMPUTER-READABLE MEDIUM FOR PERFORMING NEURAL NETWORK COMPUTATIONS FOR A NEURAL NETWORK

## (57)摘要

本發明揭示包含編碼在電腦儲存媒體上之電腦程式用於針對複數個輸入之各者產生一各自類神經網路輸出之方法、系統及設備，該方法包括對於類神經網路層之各者：接收複數個輸入以在類神經網路層處進行處理；自該複數個輸入形成一或多個輸入批次，每一批次具有至多類神經網路層之各自批次大小之一輸入數目；選擇該一或多個輸入批次中之數個輸入批次進行處理，其中該一或多個批次中之該數個批次中之該等輸入之一計數大於或等於序列中之一後續層之該各自相關聯之批次大小；及處理該一或多個輸入批次中之該數個輸入批次以產生該各自類神經網路層輸出。

Methods, systems, and apparatus, including computer programs encoded on computer storage media, for generating a respective neural network output for each of a plurality of inputs, the method comprising, for each of the neural network layers: receiving a plurality of inputs to be processed at the neural network layer; forming one or more batches of inputs from the plurality of inputs, each batch having a number of inputs up to the respective batch size for the neural network layer; selecting a number of the one or more batches of inputs to process, where a count of the inputs in the number of the one or more batches is greater than or equal to the respective associated batch size of a subsequent layer in the sequence; and processing the number of the one or more batches of inputs to generate the respective neural network layer output.

指定代表圖：

符號簡單說明：

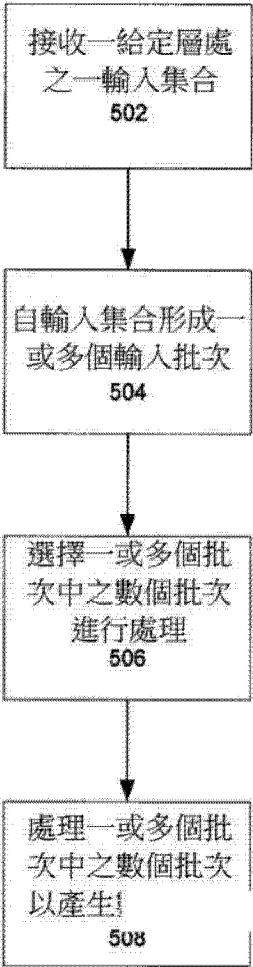
500 . . . 流程圖

502 . . . 步驟

504 . . . 步驟

506 . . . 步驟

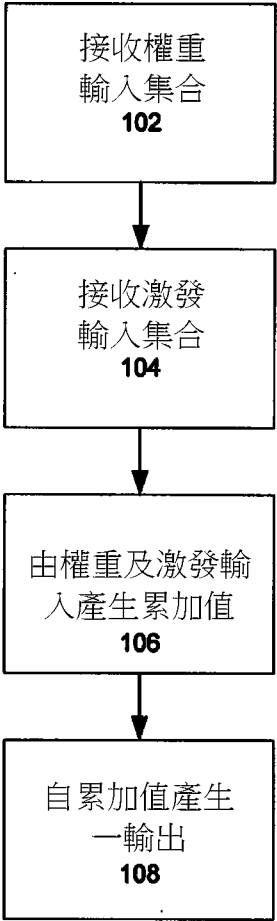
508 . . . 步驟



500 ↗

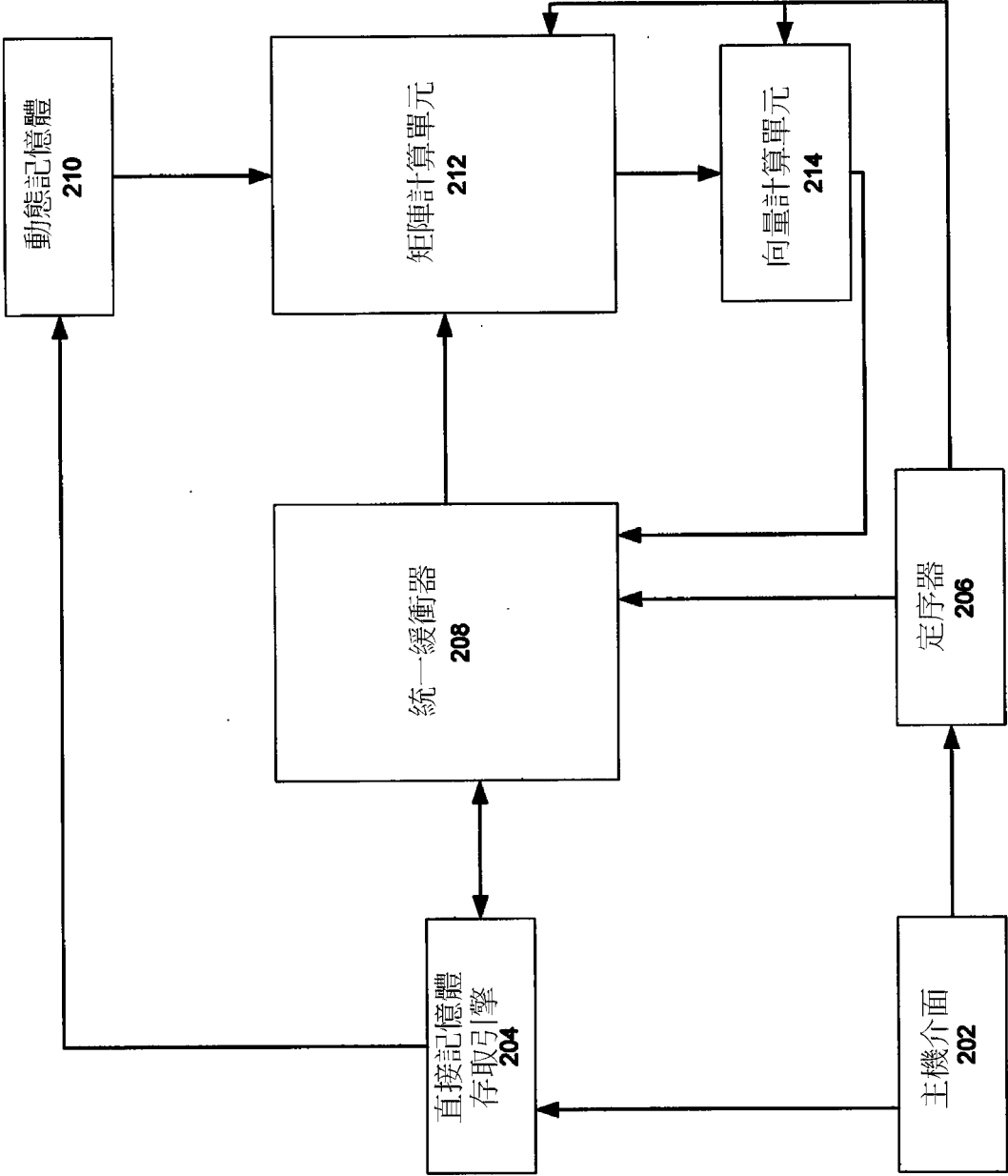
【圖 5】

【發明圖式】



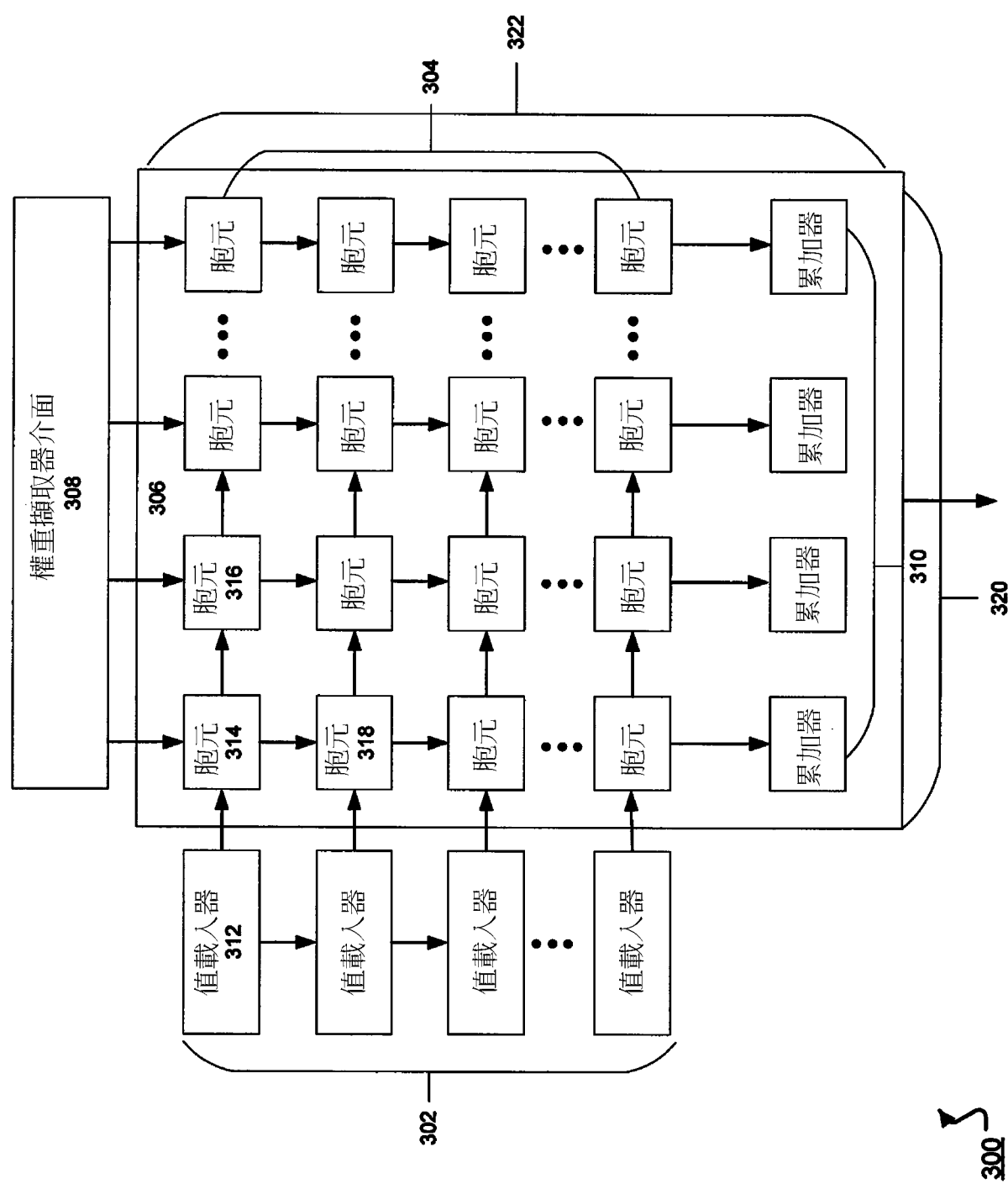
100

【圖 1】

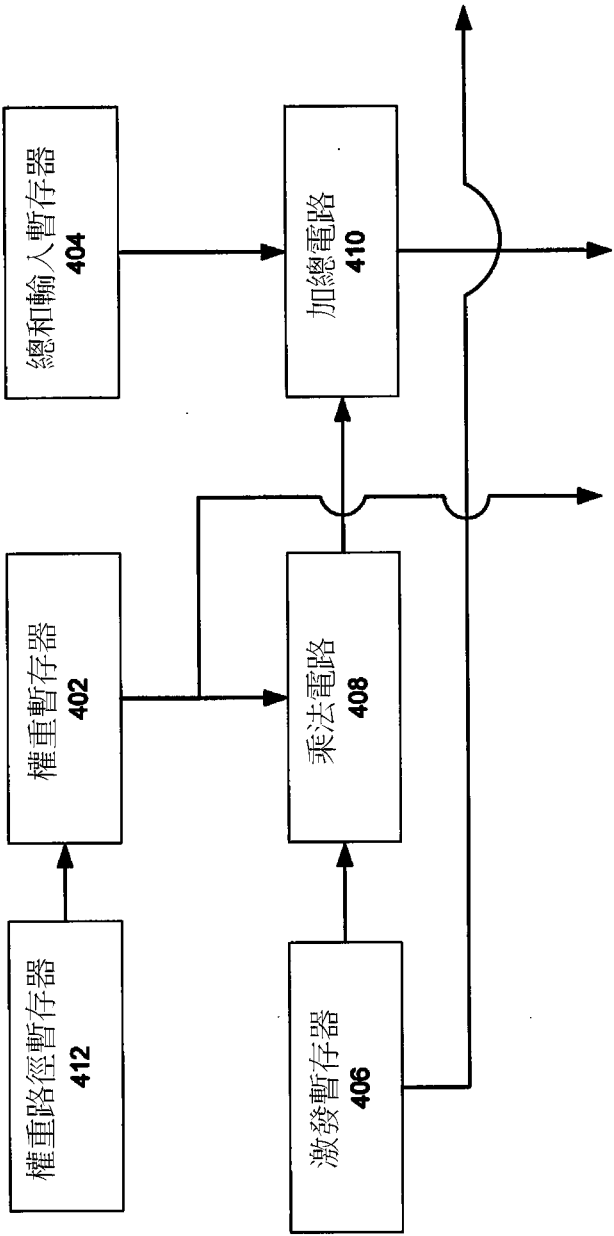


【圖 2】

200

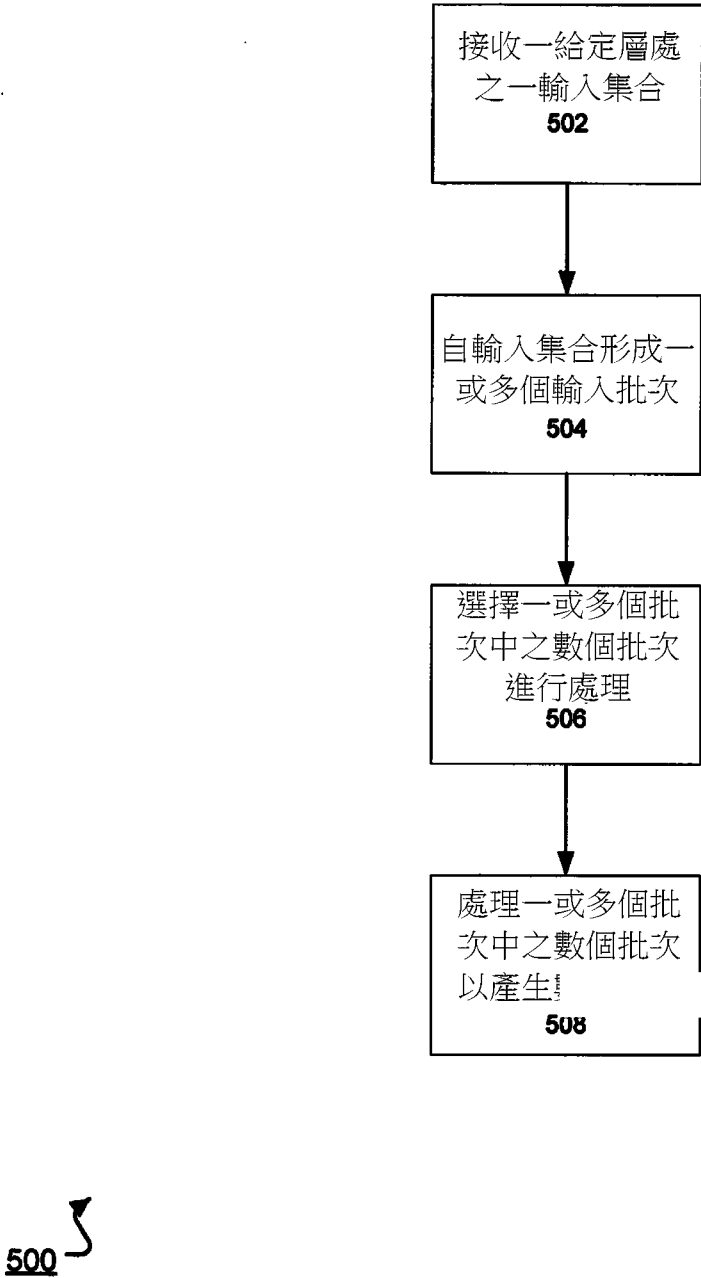


【圖 3】

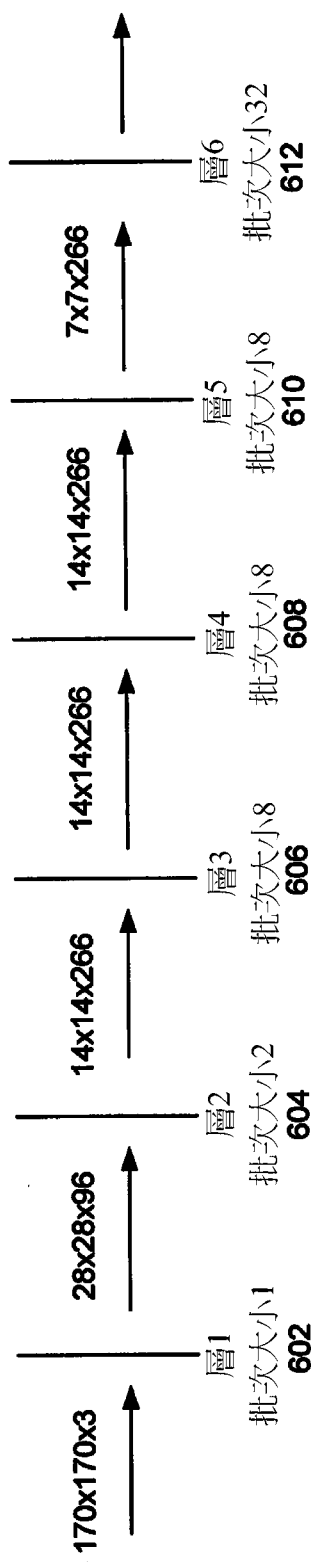


【圖 4】

400



【圖 5】



600

【圖 6】

## 【發明說明書】

### 【中文發明名稱】

用於執行類神經網路之類神經網路計算之方法、系統及電腦可讀取媒體

### 【英文發明名稱】

METHOD, SYSTEM AND COMPUTER-READABLE MEDIUM FOR PERFORMING NEURAL NETWORK COMPUTATIONS FOR A NEURAL NETWORK

### 【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。

類神經網路係採用類神經元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。

傳統上，一些類神經網路系統連續計算推理。即，當針對多個輸入計算推理時，類神經網路系統可透過類神經網路層之各者處理每一輸入以在處理下一輸入之前針對該輸入產生輸出。

### 【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以包含針對複數個輸入之各者產生一各自類神經網路輸出之動作之方法體現，其中該產生包括透過複數個類神經網路層之各者處理每一輸入以針對該輸入產生該各自類神經網路輸出，其中該等類神經網路層以一序列配置，且其中每一類神經網路層具有一各自批次大小，該方法包括對於該等類神經網路層之

各者：接收複數個輸入以在該類神經網路層處進行處理；自該複數個輸入形成一或多個輸入批次，每一批次具有至多該類神經網路層之該各自批次大小之一輸入數目；選擇該一或多個輸入批次中之數個輸入批次進行處理，其中該一或多個批次中之該數個批次中之該等輸入之一計數大於或等於該序列中之一後續層之該各自相關聯之批次大小；及處理該一或多個輸入批次中之該數個輸入批次以產生該各自類神經網路層輸出。

實施方案可包含以下特徵之一或多者。該各自批次大小係至少基於一權重重複使用值，該權重重複使用值表示使在一矩陣計算單元處使用權重輸入計算輸出值之一時間長於自記憶體載入該等權重輸入之一時間而需要重複使用該等權重輸入之一次數。其中該權重重複使用值係至少基於儲存該等權重輸入之該記憶體之一時脈速率。每一批次大小係至少基於該權重重複使用值除以重複使用該各自層之權重輸入之次數。該複數個類神經網路層係在一矩陣處理單元處進行處理，其中處理該一或多個輸入批次中之該數個輸入批次包括使用該矩陣計算單元針對每一輸入計算累加值。該權重重複使用值係基於該矩陣計算單元內部之算術單元之一數目。每一輸入對應於一相異影像資源。自該一或多個層輸出形成一批次以在該後續層處進行處理。針對每一輸出產生一對應推理。

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。專用硬體電路可藉由對多個獨立輸入重複使用一給定層之權重輸入而針對具有多個層之一類神經網路有效地執行計算。特定言之，多次重複使用該等權重輸入使得多次重複使用該等權重輸入之一計算時間大於自記憶體存取新的權重輸入之一提取時間，藉此最大化該電路中之處理量且避免該電路停止運作。即使在每一層處不同次數地重複使用權重輸入，

該電路亦可有效地執行該計算。

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係用於對多個層執行類神經網路計算之一例示性方法之一流程圖。

圖6展示具有多個層及每一層之批次大小之一例示性類神經網路。

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，類神經網路層可以一序列配置，每一層具有一各自權重集合。每一層接收一輸入並根據該層之權重集合處理輸入以產生一輸出。該輸出可用作下一類神經網路層處之一輸入。

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結構予以配置。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(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，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。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組態資訊之指令。組態資訊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將指令轉換為控制電路以執行類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。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。例如，向量計算單元214可將一非線性函數應用至矩陣計算單元之輸出(例如，累加值之一向量)以產生激發值。在一些實施方案中，向量計算單元214產生標準化值、匯集值或兩者。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)。下文參考圖3及圖4更詳細地描述矩陣計算單元212。

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。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。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。

在一些實施方案中，透過類神經網路之層之各者處理一個類神經網路輸入以針對該類神經網路輸入產生一類神經網路輸出。然而，在一些實施方案中，電路同時透過類神經網路之層處理多個類神經網路輸入以針對該等輸入產生類神經網路輸出，藉此增加電路之利用率。

至類神經網路之一特定層之一輸入可稱作一激發輸入集合。因此，一類神經網路輸入可為至一第一層之一激發輸入集合。第一層可產生一輸出以由一第二層處理，且該輸出可稱作至第二層之一激發輸入集合。

在一些實施方案中，電路可處理一第一層處之多個激發輸入集合(例

如，如上文在圖3中所述般針對每一激發輸入集合產生累加值)，接著處理一第二層處之多個激發輸入集合，以此類推，而非電路藉由將來自一第一層之激發輸入發送至類神經網路之一最後層而處理一個激發輸入集合以產生一推理。一給定層處之多個激發輸入集合可構成一或多個輸入批次。在一些實施方案中，在電路處理一後續層之批次之前處理一給定層之批次。電路可藉由針對一批次中之每一輸入(其可為一獨立激發輸入集合)產生累加值而處理該批次。

在一些實施方案中，為最大化處理量同時對一給定類神經網路層執行計算，可引起電路處理多個獨立激發輸入，同時將給定類神經網路層之權重輸入載入脈動陣列中。即，權重輸入可與多個激發輸入一起被重複使用。特定言之，電路藉由使用權重輸入及不同激發輸入(例如，使用圖2之矩陣計算單元212)計算累加值而重複使用權重輸入。藉由圖解，激發輸入可來自多個不同的影像資源或多個音訊樣本。此將在下文參考圖6進一步描述。

電路具有一處理速度及其可存取記憶體之一速度。通常，處理速度快於記憶體存取速度。藉由重複使用權重且藉此利用電路的處理速度，電路可利用脈動陣列之胞元，同時自記憶體並行存取權重輸入以在後續計算中使用。

電路可自該電路之一權重重複使用值而針對類神經網路之每一層判定重複使用權重的次數。此可由電路上針對硬體編譯之軟體或由硬體判定。在一些實施方案中，權重重複使用值係在多次重複使用權重輸入與不同激發輸入之一計算時間大於或等於自記憶體存取新的權重輸入之一提取時間之前電路可重複使用權重輸入之次數。因此，權重重複使用值可指示為了最大化利用率並最小化記憶體存取等待時間電路應重複使用權重輸入之次

數。

權重重複使用值可基於以下一或多項：儲存權重輸入之記憶體(例如，圖2之動態記憶體210)之一時脈速率、電路內部(例如，圖2之矩陣計算單元212內部)之算術單元之數目及記憶體(例如，圖2之統一緩衝器208及動態記憶體210)中之通道數目。權重重複使用值可用於判定類神經網路之每一層之一各自批次大小。批次大小可稱作一有效批次大小，即，允許一層以產生最有效的權重重複使用量之一方式操作之一批次大小。批次大小係待在一層處處理使得電路中之脈動陣列最大化之輸入之數目。在一些實施方案中，一給定層之一批次大小係權重重複使用值除以在該層處重複使用權重輸入之次數之上限值(ceiling)。重複使用權重輸入之次數可直接對應於當執行一累加計算時儲存在電路之一脈動陣列之胞元中之激發輸入之數目。此係因為在載入新的激發輸入集合時，權重輸入可保留在胞元中(即，「暫停」)。當載入新的激發輸入集合且將權重輸入凍結在適當位置中時，電路可再次重複使用權重輸入以憑藉新的激發輸入集合計算卷積。下文將參考圖6進一步描述一實例。

圖5係用於對多個層執行類神經網路計算之一例示性方法之一流程圖500。該方法可由一專用硬體電路(例如，使用圖2之主機202)或由一或多個處理器、韌體、晶片外程序或由經組態以控制電路之某個其他軟體程序執行。

在每一類神經網路層處之處理期間，電路可接收一層輸入集合以供該層處理(步驟502)。該層輸入集合可來自記憶體(例如，圖2之統一緩衝器208)或來自一先前層(即，輸入集合係自網路中之先前層之輸出產生)且使用輸出作為輸入集合可由主機202管理，如上文參考圖2描述。每一輸入可

為一激發輸入集合，且可自一獨立類神經網路輸入產生。

對於一給定類神經網路層，系統可自輸入集合形成一或多個輸入批次(步驟504)。如上所述，每一類神經網路層具有一相關聯之批次大小。該層處之每一經形成批次包含不超過層的批次大小之一輸入數目。每一類神經網路層之批次大小可被判定為類神經網路之一組態。在一些實施方案中，當編譯一類神經網路模型以在電路上執行時，計算每一類神經網路層之批次大小。因為可在處理開始於一層處之前計算一給定層之批次大小，所以可基於給定層之批次大小形成一批次內之輸入數目。

作為一實例，若一類神經網路需要處理24個影像且類神經網路之一第一層具有一批次大小8，則電路可形成8個影像之3個批次，即，將24個影像劃分為8個輸入之3個批次。對於每一批次，電路可藉由重複使用該層之一特定權重輸入集合來處理批次中之8個相異影像。電路接著可(1)處理一後續層處之一或多個批次或(2)使用特定權重輸入集合處理該層處之8個相異影像資源之另一批次。電路基於類神經網路中之其他層之批次大小判定是否進行(1)或(2)，此將在下文以步驟506並參考圖6進一步描述。

對於一給定類神經網路層，電路可選擇一或多個輸入批次中之數個輸入批次進行處理(步驟506)。在一些實施方案中，選擇數目使得一或多個輸入批次中之數個輸入批次中之輸入之一計數大於或等於一後續層之對應批次大小。例如，若一當前層必須處理各具有5個輸入之3個批次且一後續層具有一批次大小10，則電路可選擇3個批次中的2個批次進行處理。該2個選定批次總共具有10個輸入，即， $2*5$ ，其等於後續層之對應批次大小(即，10)。在一些實施方案中，將在電路中之脈動陣列之一後續遍次上處理剩餘第三批次。下文將參考圖6進一步描述又另一實例。

電路可(例如)使用一脈動陣列處理選定數目個輸入批次以針對選定批次中之每一輸入產生一各自層輸出(步驟508)。電路可藉由在對每一輸入計算累加值時重複使用權重輸入來處理每一選定輸入批次。如上所述，將權重輸入施加至多個獨立輸入，因為在載入新的激發輸入集合時，權重輸入可保留在胞元中(即，「暫停」)。

藉由圖解，一批次可具有兩個輸入：一激發輸入集合A1及一激發輸入集合A2。電路可藉由在自記憶體提取一權重輸入集合W2之前施加一權重輸入集合W1至兩個集合A1及A2而重複使用該集合W1，其中集合W2係後續層之一權重集合，或若該層具有更多權重以供電路之一脈動陣列處理，則集合W2係該層之下一權重子集。在另一實例中，若存在各自待處理之5個輸入之8個批次(即，總共40個輸入)且電路選擇4個批次進行處理，則電路可處理4個批次內之輸入(即，總共20個輸入)以產生各自層輸出(即，總共20個層輸出)。此將在下文參考圖6進一步描述。

在處理選定數目個批次之後，電路可處理該層處之第二數目個批次或進行至處理一後續層處之一或多個輸入批次。此將在下文參考圖6進一步描述。

雖然此方法已被描述為實施於處理一類神經網路之一電路上，但此方法亦可實施於一處理器(例如，一中央處理單元(CPU)或一圖形處理單元(GPU))上。例如，處理器可重複使用一權重值達某個次數，同時自記憶體(例如，動態隨機存取記憶體(DRAM))擷取權重。

圖6展示具有多個層及每一層之一批次大小之一示例性類神經網路600。如上所述，一給定層之一批次大小係權重重複使用值除以在該層處重複使用權重輸入之次數之上限值。上限值函數可處理一數字並傳回不小於

該數字之最小整數。重複使用權重輸入之次數可直接對應於儲存在電路之一脈動陣列之胞元中之激發輸入之數目。例如，電路之一權重重複使用值可為1500。層1 602處理一 $170 \times 170$ 輸入。 $170 \times 170$ 激發輸入之各者可儲存在脈動陣列之一胞元中。因此，層1 602之批次大小係 $\text{CEIL} (1500 / (170 * 170)) = 1$ 。層2 604之批次大小係 $\text{CEIL} (1500 / (28 * 28)) = 2$ 。使用相同計算，層3 606至層5 610具有一批次大小8，且層6 612具有一批次大小32。在一些實施方案中，將批次大小捨入至2之一最接近冪次。

藉由圖解，層1具有1個輸入之一批次大小且層2具有2個輸入之一批次大小。電路可(例如)接收10個輸入用於在層1處處理(例如，總共10個相異影像)。

電路可判定應處理層1處之10個批次中之兩個批次，即，總共2個輸入。在一些實施方案中，電路藉由將層2處之批次大小除以層1之批次大小判定此數目，即， $2/1 = 2$ 。藉由處理單個輸入之兩個批次，電路處理兩個輸入，此等於層2之一批次大小，即，2。電路可處理該兩個批次以(例如，藉由計算累加值之兩個向量-一個向量來自每一影像)自兩個輸入產生兩個層輸出。每一層輸出可為輸入至後續層之一輸入。該電路可形成兩個輸入之一批次(例如累加值之兩個向量)以在層2處進行處理。該等批次可儲存在統一緩衝器中。

電路可判定應處理層2處之4個批次。藉由處理大小2之4個批次，該電路處理8個輸入，此等於層3之一批次大小，即，8。電路可處理該4個批次以自8個輸入產生8個層輸出。接著，電路可形成8個輸入之一批次以在層3處進行處理。

電路可繼續直至已處理類神經網路之所有層處之輸入。例如，因為層

3至5具有一批次大小8，所以電路可判定應處理層3至5之各者處之8個輸入之一批次。類似地，電路可判定應處理層5處各具有8個輸入之4個批次，藉此形成32個輸入之一批次，其可在層6處進行處理。對於每一輸入，層6可產生可用於推理之一或多個層輸出。

在一些實施方案中，電路(例如)使用算術電路計算跨類神經網路中之所有層之批次大小之一最小公倍數。電路接著可在處理後續層之前處理每一層處之等於或大於最小公倍數之最小數目個輸入。換言之，電路可基於最小公倍數判定(1)是否處理後續層處之一批次或(2)處理一當前層處之另一批次。例如，層1 602至層6 612之批次大小之最小公倍數係32。在處理層1處之一個批次之後，電路可判定層1僅產生1個輸出，其小於最小公倍數32。因此，電路可判定處理層1處之另一批次，此時已產生總共2個輸出。電路可繼續產生輸出直至在層1處已產生32個輸出(即，最小公倍數)。接著，電路可進行至處理一後續層處之一批次。在一些實施方案中，將每一層輸出儲存在一統一緩衝器(例如，圖2之統一緩衝器208)中。

藉由圖解，電路可在層1處選擇32個批次總計32個輸入用於處理。電路接著可在層2處選擇16個批次總共32個輸入，接著在層3處選擇4個批次總共32個輸入用於處理，接著在層4處選擇4個批次總共32個輸入用於處理，接著在層4處選擇5個批次總共32個輸入用於處理、接著最後在層6處選擇1批次總共32個輸入用於處理。

在一些實施方案中，在一特定層處，電路基於一系統參數(例如，由一使用者提供之一參數)處理大於最小公倍數之數目個輸入。在一些實施方案中，電路處理可除以後續層的批次大小之數目個輸入。在一些其他實施方案中，電路接收指令以處理大量輸入，且電路在進行至產生後續層處之

輸出之前處理輸入批次以最大化一特定層處之經產生輸出之一數目。電路可約束在特定層處產生之輸出之數目使得所產生的輸出之數目可除以一後續層的批次大小。

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/主機   |
| 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 | 流程圖     |
| 502 | 步驟      |
| 504 | 步驟      |
| 506 | 步驟      |
| 508 | 步驟      |
| 600 | 類神經網路   |
| 602 | 層1      |
| 604 | 層2      |
| 606 | 層3      |
| 608 | 層4      |
| 610 | 層5      |
| 612 | 層6      |



申請日: 105/05/20

IPC分類: G06N 3/02 (2006.01)

## 【發明摘要】

## 【中文發明名稱】

用於執行類神經網路之類神經網路計算之方法、系統及電腦可讀取媒體

## 【英文發明名稱】

METHOD, SYSTEM AND COMPUTER-READABLE MEDIUM FOR PERFORMING NEURAL NETWORK COMPUTATIONS FOR A NEURAL NETWORK

## 【中文】

本發明揭示包含編碼在電腦儲存媒體上之電腦程式用於針對複數個輸入之各者產生一各自類神經網路輸出之方法、系統及設備，該方法包括對於類神經網路層之各者：接收複數個輸入以在類神經網路層處進行處理；自該複數個輸入形成一或多個輸入批次，每一批次具有至多類神經網路層之各自批次大小之一輸入數目；選擇該一或多個輸入批次中之數個輸入批次進行處理，其中該一或多個批次中之該數個批次中之該等輸入之一計數大於或等於序列中之一後續層之該各自相關聯之批次大小；及處理該一或多個輸入批次中之該數個輸入批次以產生該各自類神經網路層輸出。

## 【英文】

Methods, systems, and apparatus, including computer programs encoded on computer storage media, for generating a respective neural network output for each of a plurality of inputs, the method comprising, for each of the neural network layers: receiving a plurality of inputs to be processed at the neural network layer; forming one or more batches of inputs from the plurality of inputs, each batch having a number of inputs up to the respective batch size for the neural network layer; selecting a number of the one or more batches of inputs to process, where a count of the inputs in the number of the one or more batches is greater than or equal to the respective associated

batch size of a subsequent layer in the sequence; and processing the number of the one or more batches of inputs to generate the respective neural network layer output.

【指定代表圖】

圖5

【代表圖之符號簡單說明】

- 500        流程圖
- 502        步驟
- 504        步驟
- 506        步驟
- 508        步驟

## 【發明申請專利範圍】

### 【第1項】

一種用於針對具有複數個類神經網路層之一類神經網路來執行類神經網路計算之方法，該方法包含：

獲得將被處理之複數個層輸入；

基於(i)至該複數個類神經網路層之一特定類神經網路層之一層輸入之一大小及(ii)表示一矩陣計算單元重複使用類神經網路計算之權重輸入的一次數之一權重重複使用值，判定該特定類神經網路層之一批次大小，其中該批次大小表示針對該特定類神經網路層藉由該矩陣計算單元將被平行處理之批次之一數目；

藉由該矩陣計算單元且針對該特定類神經網路層，處理層輸入之一或多個批次以產生一或多個層輸出，其中該一或多個批次之每一批次包括對應於該特定類神經網路層之該批次大小之層輸出之一數目。

### 【第2項】

如請求項1之方法，其中該權重重複使用值係至少基於一記憶體儲存權重輸入之一時脈速率。

### 【第3項】

如請求項1之方法，其中該複數個類神經網路層依一有向圖(directed graph)結構予以配置。

### 【第4項】

如請求項3之方法，其進一步包含基於在該有向圖結構中一後續層之一批次大小而判定用於處理之層輸入之該一或多個批次之一數目。

### 【第5項】

如請求項1之方法，其中至少基於該權重重複使用值除以該層輸入之該大小而判定該批次大小。

**【第6項】**

如請求項1之方法，其中處理層輸入之該一或多個批次包含使用該矩陣計算單元針對每一層輸入計算累加值。

**【第7項】**

如請求項6之方法，其進一步包含基於該一或多個層輸出判定一或多個推理(inference)。

**【第8項】**

如請求項1之方法，其中判定該批次大小包含：

針對該複數個類神經網路層，判定跨該複數個類神經網路中之批次大小之一最小公倍數；及

判定包括等於或大於該最小公倍數之一最小數目個層輸入之該批次大小。

**【第9項】**

如請求項1之方法，其中每一層輸入對應於一相異影像資源之一特徵。

**【第10項】**

如請求項1之方法，其中每一層輸入對應於一音訊樣本之一特徵。

**【第11項】**

一種用於針對具有複數個類神經網路層之一類神經網路來執行類神經網路計算之系統，該系統包含：

一或多個處理器；及

電腦可讀媒體，其耦合至該一或多個處理器且其上儲存有指令，該等

指令在由該一或多個處理器執行時引起該一或多個處理器執行包括以下各者之操作：

獲得將被處理之複數個層輸入；

基於(i)至該複數個類神經網路層之一特定類神經網路層之一層輸入之一大小及(ii)表示一矩陣計算單元重複使用類神經網路計算之權重輸入的一次數之一權重重複使用值，判定該特定類神經網路層之一批次大小，其中該批次大小表示針對該特定類神經網路層藉由該矩陣計算單元將被平行處理之批次之一數目；

藉由該矩陣計算單元且針對該特定類神經網路層，處理層輸入之一或多個批次以產生一或多個層輸出，其中該一或多個批次之每一批次包括對應於該特定類神經網路層之該批次大小之層輸出之一數目。

**【第12項】**

如請求項11之系統，其中該權重重複使用值係至少基於一記憶體儲存權重輸入之一時脈速率。

**【第13項】**

如請求項11之系統，其中該複數個類神經網路層依一有向圖(directed graph)結構予以配置。

**【第14項】**

如請求項13之系統，其進一步包含基於在該有向圖結構中一後續層之一批次大小而判定用於處理層輸入之該一或多個批次之一數目。

**【第15項】**

如請求項11之系統，其中至少基於該權重重複使用值除以該層輸入之該大小而判定該批次大小。

**【第16項】**

如請求項11之系統，其中處理層輸入之該一或多個批次包含使用該矩陣計算單元針對每一層輸入計算累加值。

**【第17項】**

如請求項16之系統，其進一步包含基於該一或多個層輸出判定一或多個推理(inference)。

**【第18項】**

如請求項11之系統，其中判定該批次大小包含：

針對該複數個類神經網路層，判定跨該複數個類神經網路中之批次大小之一最小公倍數；及

判定包括等於或大於該最小公倍數之一最小數目個層輸入之該批次大小。

**【第19項】**

一種其上儲存有指令之電腦可讀媒體，該等指令在由一或多個處理器執行時引起該一或多個處理器執行包括以下各者之操作：

針對具有複數個類神經網路層之一類神經網路，獲得將被處理之複數個層輸入；

基於(i)至該複數個類神經網路層之一特定類神經網路層之一層輸入之一大小及(ii)表示一矩陣計算單元重複使用類神經網路計算之權重輸入的一次數之一權重重複使用值，判定該特定類神經網路層之一批次大小，其中該批次大小表示針對該特定類神經網路層藉由該矩陣計算單元將被平行處理之批次之一數目；

藉由該矩陣計算單元且針對該特定類神經網路層，處理層輸入之一

或多個批次以產生一或多個層輸出，其中該一或多個批次之每一批次包括對應於該特定類神經網路層之該批次大小之層輸出之一數目。

**【第20項】**

如請求項19之電腦可讀媒體，其中該權重重複使用值係至少基於一記憶體儲存權重輸入之一時脈速率。

**【第21項】**

如請求項19之電腦可讀媒體，其中該複數個類神經網路層依一有向圖(directed graph)結構予以配置。

**【第22項】**

如請求項21之電腦可讀媒體，其進一步包含基於在該有向圖結構中一後續層之一批次大小而判定用於處理層輸入之該一或多個批次之一數目。

**【第23項】**

如請求項19之電腦可讀媒體，其中至少基於該權重重複使用值除以該層輸入之該大小而判定該批次大小。

**【第24項】**

如請求項19之電腦可讀媒體，其中處理層輸入之該一或多個批次包含使用該矩陣計算單元針對每一層輸入計算累加值。

**【第25項】**

如請求項24之電腦可讀媒體，其中該等操作進一步包含基於該一或多個層輸出判定一或多個推理(inference)。

**【第26項】**

如請求項19之電腦可讀媒體，其中判定該批次大小包含：

針對該複數個類神經網路層，判定跨該複數個類神經網路中之批次

大小之一最小公倍數；及

判定包括等於或大於該最小公倍數之一最小數目個層輸入之該批次大小。

**【第27項】**

如請求項19之電腦可讀媒體，其中每一層輸入對應於一相異影像資源之一特徵。

batch size of a subsequent layer in the sequence; and processing the number of the one or more batches of inputs to generate the respective neural network layer output.

【指定代表圖】

圖5

【代表圖之符號簡單說明】

- 500        流程圖
- 502        步驟
- 504        步驟
- 506        步驟
- 508        步驟