201627923

(19)中華民國智慧財產局

(12)發明說明書公開本 (11)公開編號：TW 201627923 A

(43)公開日： 中華民國 105 (2016) 年 08 月 01 日

(21)申請案號：104142524 (22)申請日： 中華民國 104 (2015) 年 12 月 17 日

(51)Int. Cl.： *G06N3/08 (2006.01)* *G06N3/02 (2006.01)*

(30)優先權：2015/01/22 美國 62/106,608

2015/09/04 美國 14/846,579

(71)申請人：高通公司 (美國) QUALCOMM INCORPORATED (US)

美國

(72)發明人：安納普瑞迪凡卡塔斯瑞坎塔瑞迪 ANNAPUREDDY, VENKATA SREEKANTA REDDY (IN)；迪捷克曼丹尼爾亨瑞克司法蘭西克斯 DIJKMAN, DANIEL HENDRICUS FRANCISCUS (NL)；朱利安大衛強納森 JULIAN, DAVID JONATHAN (US)

(74)代理人：李世章

申請實體審查：無 申請專利範圍項數：36 項 圖式數：13 共 71 頁

(54)名稱

模型壓縮和微調

MODEL COMPRESSION AND FINE-TUNING

(57)摘要

壓縮機器學習網路(諸如神經網路)包括用經壓縮層代替神經網路中的一個層以產生經壓縮網路。經壓縮網路可經由更新經壓縮層中的權重值來微調。

Compressing a machine learning network, such as a neural network, includes replacing one layer in the neural network with compressed layers to produce the compressed network. The compressed network may be fine-tuned by updating weight values in the compressed layer(s).

指定代表圖：

TW 201627923 A

符號簡單說明：

1000・・・方法

1002・・・方塊

1004・・・方塊

1006・・・方塊

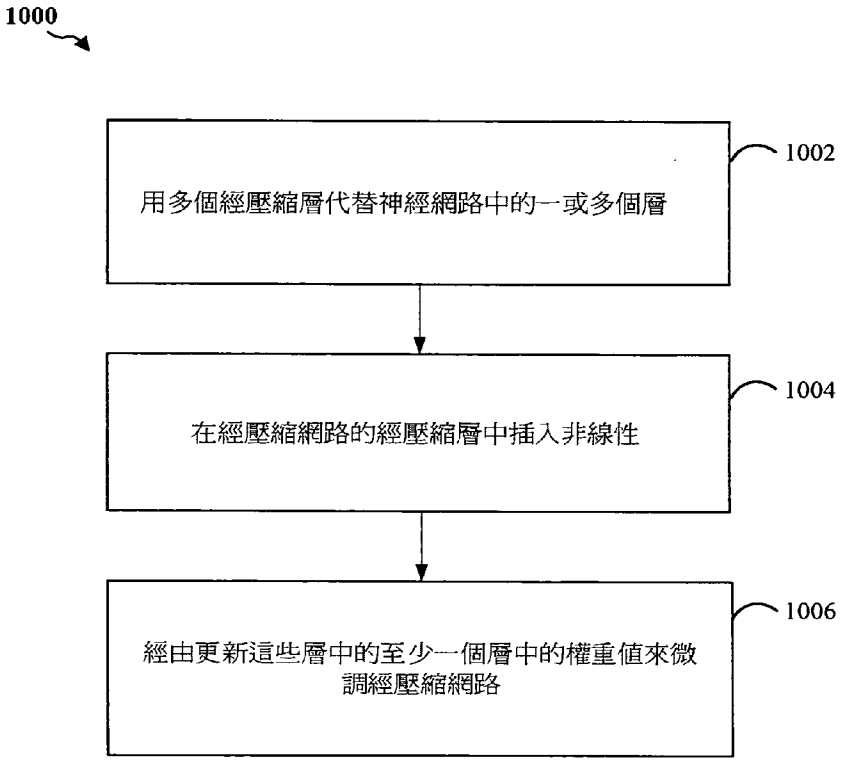

圖10

201627923

發明摘要

※ 申請案號：104142524

※ 申請日： 104 年 12 月 17 日 ※IPC 分類： G06N 3/08 (2006.01)
G06N 3/02 (2006.01)

【發明名稱】（中文/英文）

模型壓縮和微調

MODEL COMPRESSION AND FINE-TUNING

【中文】

壓縮機器學習網路（諸如神經網路）包括用經壓縮層代替神經網路中的一個層以產生經壓縮網路。經壓縮網路可經由更新經壓縮層中的權重值來微調。

【英文】

Compressing a machine learning network, such as a neural network, includes replacing one layer in the neural network with compressed layers to produce the compressed network. The compressed network may be fine-tuned by updating weight values in the compressed layer(s).

【代表圖】

【本案指定代表圖】：第（ 10 ）圖。

【本代表圖之符號簡單說明】：

1000 方法

1002 方塊

1004 方塊

S

1006 方塊

【本案若有化學式時，請揭示最能顯示發明特徵的化學式】：

無

發明專利說明書

（本說明書格式、順序，請勿任意更動）

【發明名稱】（中文/英文）

模型壓縮和微調

MODEL COMPRESSION AND FINE-TUNING

【相關申請案的交叉引用】

【0001】 本專利申請案主張於2015年1月22日提出申請的題爲「MODEL COMPRESSION AND FINE-TUNING（模型壓縮和微調）」的美國臨時專利申請案第62/106,608號的權益，其揭示內容經由援引全部明確納入於本文。

【技術領域】

【0002】 本案的某些態樣一般係關於神經系統工程，並且尤其係關於用於壓縮神經網路的系統和方法。

【先前技術】

【0003】 可包括一群互連的人工神經元（例如，神經元模型）的人工神經網路是一種計算設備或者表示將由計算設備執行的方法。

【0004】 迴旋神經網路是一種前饋人工神經網路。迴旋神經網路可包括神經元集合，其中每個神經元具有感受野並且合而鋪覆一個輸入空間。迴旋神經網路（CNN）具有眾多應用。具體地，CNN已被廣泛使用於模式辨識和分類領域。

【0005】 深度學習架構（諸如，深度置信網路和深度迴旋網路）是分層神經網路架構，其中第一層神經元的輸出變成第

5

二層神經元的輸入，第二層神經元的輸出變成第三層神經元的輸入，以此類推。深度神經網路可被訓練以辨識特徵階層並且因此它們被越來越多地使用於物件辨識應用。類似於迴旋神經網路，這些深度學習架構中的計算可分佈在處理節點群體上，其可被配置在一或多個計算鏈中。這些多層架構可每次訓練一層並可使用反向傳播來進行微調。

【0006】 其他模型亦可用於物件辨識。例如，支援向量機(SVM)是可被應用於分類的學習工具。支援向量機包括分類資料的分離超平面（例如，決策邊界）。該超平面由監督式學習來定義。期望的超平面增加訓練資料的餘量。換言之，該超平面應該具有到訓練實例的最大的最小距離。

【0007】 儘管這些解決方案在數個分類基準上取得了優異的結果，但它們的計算複雜度可能極其高。另外，模型的訓練可能是有挑戰性的。

【發明內容】

【0008】 在一個態樣，揭示一種壓縮機器學習網路（諸如神經網路）的方法。該方法包括用多個經壓縮層代替神經網路中的至少一個層以產生經壓縮神經網路。該方法亦包括在經壓縮網路的經壓縮層之間插入非線性。進一步，該方法包括經由更新這些經壓縮層中的至少一個經壓縮層中的權重值來微調經壓縮網路。

【0009】 另一態樣揭示一種用於壓縮機器學習網路（諸如神經網路）的設備。該設備包括用於用多個經壓縮層代替神經網路中的至少一個層以產生經壓縮神經網路的裝置。該設備

亦包括用於在經壓縮網路的經壓縮層之間插入非線性的裝置。進一步，該設備包括用於經由更新這些經壓縮層中的至少一個經壓縮層中的權重值來微調經壓縮網路的裝置。

【0010】 另一態樣揭示一種用於壓縮機器學習網路（諸如神經網路）的裝置。該裝置包括記憶體以及耦合至該記憶體的至少一個處理器。處理器被配置成用多個經壓縮層代替神經網路中的至少一個層以產生經壓縮神經網路。處理器亦被配置成在經壓縮網路的經壓縮層之間插入非線性。進一步，處理器亦被配置成經由更新至少一個經壓縮層中的權重值來微調經壓縮網路。

【0011】 另一態樣揭示一種非瞬態電腦可讀取媒體。該電腦可讀取媒體上記錄有用於壓縮機器學習網路（諸如神經網路）的非瞬態程式碼。該程式碼在由處理器執行時使處理器用多個經壓縮層代替神經網路中的至少一個層以產生經壓縮神經網路。該程式碼亦使處理器在經壓縮網路的經壓縮層之間插入非線性。進一步，該程式碼亦使處理器經由更新至少一個經壓縮層中的權重值來微調經壓縮網路。

【0012】 在另一態樣，揭示一種用於壓縮機器學習網路（諸如神經網路）的方法。該方法包括用多個經壓縮層代替神經網路中的至少一個層以產生經壓縮神經網路，從而組合的經壓縮層的感受野大小與未壓縮層的感受野相匹配。該方法亦包括經由更新這些經壓縮層中的至少一個經壓縮層中的權重值來微調經壓縮網路。

【0013】 另一態樣揭示一種用於壓縮機器學習網路（諸如神

經網路）的設備。該設備包括用於用多個經壓縮層代替神經網路中的至少一個層以產生經壓縮神經網路，從而組合的經壓縮層的感受野大小與未壓縮層的感受野相匹配的裝置。該設備亦包括用於經由更新這些經壓縮層中的至少一個經壓縮層中的權重值來微調經壓縮網路的裝置。

【0014】 另一態樣揭示一種用於壓縮機器學習網路（諸如神經網路）的裝置。該裝置包括記憶體以及耦合至該記憶體的至少一個處理器。處理器被配置成用多個經壓縮層代替神經網路中的至少一個層以產生經壓縮神經網路，從而組合的經壓縮層的感受野大小與未壓縮層的感受野相匹配。處理器亦被配置成經由更新至少一個經壓縮層中的權重值來微調經壓縮網路。

【0015】 另一態樣揭示一種非瞬態電腦可讀取媒體。該電腦可讀取媒體上記錄有用於壓縮機器學習網路（諸如神經網路）的非瞬態程式碼。該程式碼在由處理器執行時使處理器用多個經壓縮層代替神經網路中的至少一個層以產生經壓縮神經網路，從而組合的經壓縮層的感受野大小與未壓縮層的感受野相匹配。該程式碼亦使處理器經由更新至少一個經壓縮層中的權重值來微調經壓縮網路。

【0016】 在另一態樣，揭示一種壓縮機器學習網路（諸如神經網路）的方法。該方法包括用多個經壓縮層代替神經網路中的至少一個層以產生經壓縮神經網路。該方法亦包括經由應用交替最小化程序來決定經壓縮層的權重矩陣。

【0017】 另一態樣揭示一種用於壓縮機器學習網路（諸如神

經網路）的設備。該設備包括用於用多個經壓縮層代替神經網路中的至少一個層以產生經壓縮神經網路的裝置。該設備亦包括用於經由應用交替最小化程序來決定經壓縮層的權重矩陣的裝置。

【0018】 另一態樣揭示一種用於壓縮機器學習網路（諸如神經網路）的裝置。該裝置包括記憶體以及耦合至該記憶體的至少一個處理器。處理器被配置成用多個經壓縮層代替神經網路中的至少一個層以產生經壓縮神經網路。處理器亦被配置成經由應用交替最小化程序來決定經壓縮層的權重矩陣。

【0019】 另一態樣揭示一種非瞬態電腦可讀取媒體。該電腦可讀取媒體上記錄有用於壓縮機器學習網路（諸如神經網路）的程式碼。該程式碼在由處理器執行時使處理器用多個經壓縮層代替神經網路中的至少一個層以產生經壓縮神經網路。該程式碼亦使處理器經由應用交替最小化程序來決定經壓縮層的權重矩陣。

【0020】 本案的額外特徵和優點將在下文描述。本發明所屬領域中具有通常知識者應該領會，本案可容易地被用作修改或設計用於實施與本案相同的目的的其他結構的基礎。本發明所屬領域中具有通常知識者亦應認識到，此類等效構造並不脫離所附請求項中所闡述的本案的教導。被認爲是本案的特性的新穎特徵在其組織和操作方法兩態樣連同進一步的目的和優點在結合附圖來考慮以下描述時將被更好地理解。然而，要清楚理解的是，提供每一幅附圖均僅用於圖示和描述目的，且無意作爲對本案的限定的定義。

【圖式簡單說明】

【0021】 在結合附圖理解下面闡述的詳細描述時，本案的特徵、本質和優點將變得更加明顯，在附圖中，相同元件符號始終作相應標識。

【0022】 圖1圖示了根據本案的一些態樣的使用片上系統（SOC）（包括通用處理器）來設計神經網路的實例實現。

【0023】 圖2圖示了根據本案的各態樣的系統的實例實現。

【0024】 圖3A是圖示根據本案的各態樣的神經網路的示圖。

【0025】 圖3B是圖示根據本案的各態樣的示例性深度迴旋網路（DCN）的方塊圖。

【0026】 圖4A是圖示根據本案的各態樣的可將人工智慧（AI）功能模組化的示例性軟體架構的方塊圖。

【0027】 圖4B是圖示根據本案的各態樣的智慧手機上的AI應用的執行時操作的方塊圖。

【0028】 圖5A-B和6A-B是圖示根據本案的各態樣的全連接層和經壓縮全連接層的方塊圖。

【0029】 圖7是圖示根據本案的各態樣的示例性迴旋層的方塊圖。

【0030】 圖8A-B和9A-B圖示了根據本案的各態樣的迴旋層的實例壓縮。

【0031】 圖10-13是圖示根據本案的各態樣的用於壓縮神經網路的方法的流程圖。

【實施方式】

【0032】 以下結合附圖闡述的詳細描述意欲作爲各種配置的

描述，而無意表示可實踐本文中所描述的概念的僅有的配置。本詳細描述包括具體細節以便提供對各種概念的透徹理解。然而，對於本發明所屬領域中具有通常知識者將顯而易見的是，沒有這些具體細節亦可實踐這些概念。在一些實例中，以方塊圖形式示出眾所周知的結構和元件以避免湮沒此類概念。

【0033】 基於本教導，本發明所屬領域中具有通常知識者應領會，本案的範疇意欲覆蓋本案的任何態樣，不論其是與本案的任何其他態樣相獨立地還是組合地實現的。例如，可以使用所闡述的任何數目的態樣來實現裝置或實踐方法。另外，本案的範疇意欲覆蓋使用作爲所闡述的本案的各個態樣的補充或者與之不同的其他結構、功能性、或者結構及功能性來實踐的此類裝置或方法。應當理解，所揭示的本案的任何態樣可由請求項的一或多個元素來實施。

【0034】 措辭「示例性」在本文中用於表示「用作實例、例子或圖示」。本文中描述爲「示例性」的任何態樣不必被解釋爲優於或勝過其他態樣。

【0035】 儘管本文描述了特定態樣，但這些態樣的眾多變體和置換落在本案的範疇之內。儘管提到了優選態樣的一些益處和優點，但本案的範疇並非意欲被限定於特定益處、用途或目標。相反，本案的各態樣意欲能寬泛地應用於不同的技術、系統組態、網路和協定，其中一些作爲實例在附圖以及以下對優選態樣的描述中圖示。詳細描述和附圖僅僅圖示本案而非限定本案，本案的範疇由所附請求項及其等效技術方

案來定義。

模型壓縮和微調

【0036】 深度神經網路執行若干人工智慧任務中的現有技術，諸如影像/視訊分類、語音辨識、臉孔辨識等。神經網路模型從大型訓練實例資料庫進行訓練，並且通常而言，較大模型趨向於達成較佳的效能。爲了在行動設備（諸如智慧型電話、機器人和汽車）上部署這些神經網路模型的目的，期望儘可能地減小或最小化計算複雜度、記憶體佔用和功耗。這些亦是雲端應用（諸如，每天處理數百萬影像/視訊的資料中心）的期望屬性。

【0037】 本案的各態樣涉及一種壓縮神經網路模型的方法。經壓縮模型具有顯著更少數目的參數，並且由此具有較小的記憶體佔用。另外，經壓縮模型具有顯著更少的操作，並且由此導致較快的推斷執行時。

【0038】 迴旋神經網路模型可被劃分成層序列。每一層可變換從網路中的一或多個在前層接收的輸入並且可產生可被提供給網路的後續層的輸出。例如，迴旋神經網路可包括全連接層、迴旋層、局部連接層以及其他層。這些不同層中的每一層可執行不同類型的變換。

【0039】 根據本案的各態樣，全連接層、迴旋層和局部連接層可被壓縮。所有這些層可對輸入值應用線性變換，其中線性變換經由權重值來參數化。這些權重值可被表示爲二維矩陣或較高維張量。線性變換結構對於不同層類型可以是不同的。

【0040】 在一些態樣，層可經由用相同類型的多個層代替它來被壓縮。例如，迴旋層可經由用多個迴旋層代替它來被壓縮，並且全連接層可經由用多個全連接層代替它來被壓縮。經壓縮層中的神經元可被配置有相同啓動函數。儘管一層可用多個經壓縮層來代替，但所有經壓縮層一起的總複雜度（和記憶體佔用）可小於單個未壓縮層。

【0041】 然而，模型壓縮可能以神經網路的效能下降爲代價。例如，若神經網路被用於分類問題，則分類準確度可能因壓縮而下降。此準確度下降限制了壓縮程度，因爲較高壓縮導致較高的準確度下降。爲了對抗準確度下降，在一些態樣，經壓縮模型可使用訓練實例來微調。經壓縮模型的微調可收回準確度下降並且由此可在沒有過多分類準確度下降的情況下產生改善的壓縮。經壓縮模型的微調可在逐層基礎上進行。

【0042】 因此，本案的各態樣提供了其中神經網路的層用更多此類層來代替的壓縮技術。因此，結果所得的經壓縮模型是另一迴旋神經網路模型。這是顯著的優勢，因爲可避免開發新的且明確的用於實現經壓縮模型的技術。亦即，若任何神經網路庫與原始未壓縮模型相容，則它亦將與經壓縮模型相容。此外，由於結果所得的經壓縮模型是迴旋神經網路，因此可執行全堆疊微調（例如，反向傳播），而非將微調限於特定層。

【0043】 神經網路由若干層構成。每一層取來自一或多個先前層的啓動向量作爲輸入，對組合輸入向量應用線性/非線性

變換，以及輸出將由後續層使用的啓動向量。一些層用權重來參數化，而一些層並非如此。本案的各態樣關注於以下權重層：1)全連接層，2)迴旋層，以及3)局部連接層。所有三個層皆執行線性變換，但在輸出神經元如何連接至輸入神經元態樣有所不同。

【0044】 圖1圖示了根據本案的某些態樣使用片上系統（SOC）100進行前述壓縮神經網路的實例實現，SOC 100可包括通用處理器（CPU）或多核通用處理器（CPU）102。變數（例如，神經訊號和突觸權重）、與計算設備相關聯的系統參數（例如，帶有權重的神經網路）、延遲、頻率槽資訊、以及任務資訊可被儲存在與神經處理單元（NPU）108相關聯的記憶體塊、與CPU 102相關聯的記憶體塊、與圖形處理單元(GPU）104相關聯的記憶體塊、與數位訊號處理器（DSP）106相關聯的記憶體塊、專用記憶體塊118中，或可跨多個塊分佈。在通用處理器102處執行的指令可從與CPU 102相關聯的程式記憶體載入或可從專用記憶體塊118載入。

【0045】 SOC 100亦可包括爲具體功能定製的額外處理塊(諸如GPU 104、DSP 106、連通性塊110）以及例如可偵測和辨識姿勢的多媒體處理器112，這些具體功能可包括第四代長期進化（4G LTE）連通性、無執照Wi-Fi連通性、USB連通性、藍芽連通性等。在一種實現中，NPU實現在CPU、DSP、及/或GPU中。SOC 100亦可包括感測器處理器114、影像訊號處理器（ISP）、及/或導航120（其可包括全球定位系統）。

【0046】 SOC可基於ARM指令集。在本案的一態樣，載入到

通用處理器102中的指令可包括用於以下操作的代碼：用多個經壓縮層代替機器學習網路（諸如神經網路）中的至少一個層以產生經壓縮網路，在經壓縮網路的經壓縮層之間插入非線性及/或經由更新這些經壓縮層中的至少一個經壓縮層中的權重值來微調經壓縮網路。

【0047】 在本案的另一態樣，載入到通用處理器102中的指令可包括用於以下操作的代碼：用多個經壓縮層代替機器學習網路（諸如神經網路）中的至少一個層以產生經壓縮網路，從而組合的經壓縮層的感受野大小與未壓縮層的感受野相匹配。亦可存在用於經由更新這些經壓縮層中的至少一個經壓縮層中的權重值來微調經壓縮網路的代碼。

【0048】 在本案的又一態樣，載入到通用處理器102中的指令可包括用於以下操作的代碼：用多個經壓縮層代替機器學習網路（諸如神經網路）中的至少一個層以產生經壓縮網路。亦可存在用於經由應用交替最小化程序來決定經壓縮層的權重矩陣的代碼。

【0049】 圖2圖示了根據本案的某些態樣的系統200的實例實現。如圖2中所圖示的，系統200可具有多個可執行本文所描述的方法的各種操作的局部處理單元202。每個局部處理單元202可包括局部狀態記憶體204和可儲存神經網路的參數的局部參數記憶體206。另外，局部處理單元202可具有用於儲存局部模型程式的局部（神經元）模型程式（LMP）記憶體208、用於儲存局部學習程式的局部學習程式（LLP）記憶體210、以及局部連接記憶體212。此外，如圖2中所圖示的，每個

局部處理單元202可與用於爲該局部處理單元的各局部記憶體提供配置的配置處理器單元214對接，並且與提供各局部處理單元202之間的路由的路由連接處理單元216對接。

【0050】 深度學習架構可經由學習在每一層中以逐次更高的抽象程度來表示輸入、藉此構建輸入資料的有用特徵表示來執行物件辨識任務。以此方式，深度學習解決了傳統機器學習的主要瓶頸。在深度學習出現之前，用於物件辨識問題的機器學習辦法可能嚴重依賴人類工程設計的特徵，或許與淺分類器相結合。淺分類器可以是兩類線性分類器，例如，其中可將特徵向量分量的加權和與閾值作比較以預測輸入屬於哪一類。人類工程設計的特徵可以是由擁有領域專業知識的工程師針對具體問題領域定製的模版或核心。相反，深度學習架構可學習去表示與人類工程師可能會設計的相似的特徵，但它是經由訓練來學習的。另外，深度網路可以學習去表示和辨識人類可能亦沒有考慮過的新類型的特徵。

【0051】 深度學習架構可以學習特徵階層。例如，若向第一層呈遞視覺資料，則第一層可學習去辨識輸入流中的簡單特徵（諸如邊）。若向第一層呈遞聽覺資料，則第一層可學習去辨識特定頻率中的頻譜功率。取第一層的輸出作爲輸入的第二層可以學習去辨識特徵組合，諸如對於視覺資料去辨識簡單形狀或對於聽覺資料去辨識聲音組合。更高層可學習去表示視覺資料中的複雜形狀或聽覺資料中的詞語。再高層可學習去辨識常見視覺物件或口語短語。

【0052】 深度學習架構在被應用於具有自然階層結構的問題

時可能表現特別好。例如，機動交通工具的分類可受益於首先學習去辨識輪子、擋風玻璃、以及其他特徵。這些特徵可在更高層以不同方式被組合以辨識轎車、卡車和飛機。

【0053】 神經網路可被設計成具有各種連通性模式。在前饋網路中，資訊從較低層被傳遞到較高層，其中給定層之每一者神經元向更高層中的神經元進行傳達。如前述，可在前饋網路的相繼層中構建階層式表示。神經網路亦可具有回流或回饋（亦被稱爲自頂向下（top-down））連接。在回流連接中，來自給定層中的神經元的輸出被傳達給相同層中的另一神經元。回流架構可有助於辨識跨越不止一個按順序遞送給該神經網路的輸入資料組塊的模式。從給定層中的神經元到較低層中的神經元的連接被稱爲回饋（或自頂向下）連接。當高層級概念的辨識可輔助辨別輸入的特定低層級特徵時，具有許多回饋連接的網路可能是有助益的。

【0054】 參照圖3A，神經網路的各層之間的連接可以是全連接的（302）或局部連接的（304）。在全連接網路302中，第一層中的神經元可將它的輸出傳達給第二層之每一者神經元，從而第二層之每一者神經元將從第一層之每一者神經元接收輸入。替換地，在局部連接網路304中，第一層中的神經元可連接至第二層中有限數目的神經元。迴旋網路306可以是局部連接的，並且被進一步配置成使得與針對第二層中每個神經元的輸入相關聯的連接強度被共用（例如，308）。更一般化地，網路的局部連接層可被配置成使得一層之每一者神經元將具有相同或相似的連通性模式，但其連接強度可具有不

同的值（例如，310、312、314和316）。局部連接的連通性模式可能在更高層中產生空間上相異的感受野，這是由於給定區域中的更高層神經元可接收到經由訓練被調諧爲到網路的總輸入的受限部分的性質的輸入。

【0055】 局部連接的神經網路可能非常適合於其中輸入的空間位置有意義的問題。例如，被設計成辨識來自車載相機的視覺特徵的網路300可發展具有不同性質的高層神經元，這取決於它們與影像下部關聯還是與影像上部關聯。例如，與影像下部相關聯的神經元可學習去辨識車道標記，而與影像上部相關聯的神經元可學習去辨識交通訊號燈、交通標誌等。

【0056】 DCN可以用受監督式學習來訓練。在訓練期間，DCN可被呈遞影像326（諸如限速標誌的經裁剪影像），並且可隨後計算「前向傳遞（forward pass）」以產生輸出322。輸出322可以是對應於特徵（諸如「標誌」、「60」、和「100」）的值向量。網路設計者可能希望DCN在輸出特徵向量中針對其中一些神經元輸出高得分，例如對應於經訓練的網路300的輸出322中所示的「標誌」和「60」的那些神經元。在訓練之前，DCN產生的輸出很可能是不正確的，並且由此可計算實際輸出與目標輸出之間的誤差。DCN的權重可隨後被調整以使得DCN的輸出得分與目標更緊密地對準。

【0057】 爲了調整權重，學習演算法可爲權重計算梯度向量。該梯度可指示若權重被略微調整則誤差將增加或減少的量。在頂層，該梯度可直接對應於連接倒數第二層中的活化神經元與輸出層中的神經元的權重的值。在較低層中，該梯度

可取決於權重的值以及所計算出的較高層的誤差梯度。權重可隨後被調整以減小誤差。這種調整權重的方式可被稱爲「反向傳播」，因爲其涉及在神經網路中的「反向傳遞（backward pass）」。

【0058】 在實踐中，權重的誤差梯度可能是在少量實例上計算的，從而計算出的梯度近似於真實誤差梯度。這種近似方法可被稱爲隨機梯度下降法。隨機梯度下降法可被重複，直到整個系統可達成的誤差率已停止下降或直到誤差率已達到目標水平。

【0059】 在學習之後，DCN可被呈遞新影像326並且在網路中的前向傳遞可產生輸出322，其可被認爲是該DCN的推斷或預測。

【0060】 深度置信網路（DBN）是包括多層隱藏節點的概率性模型。DBN可被用於提取訓練資料集的階層式表示。DBN可經由堆疊多層受限波爾茲曼機（RBM）來獲得。RBM是一類可在輸入集上學習概率分佈的人工神經網路。由於RBM可在沒有關於每個輸入應該被分類到哪個類的資訊的情況下學習概率分佈的，因此RBM經常被用於無監督式學習。使用混合無監督式和受監督式範式，DBN的底部RBM可按無監督方式被訓練並且可以用作特徵提取器，而頂部RBM可按受監督方式（在來自先前層的輸入和目標類的聯合分佈上）被訓練並且可用作分類器。

【0061】 深度迴旋網路（DCN）是迴旋網路的網路，其配置有額外的池化和正規化層。DCN已在許多工上達成現有最先

進的效能。DCN可使用受監督式學習來訓練，其中輸入和輸出目標兩者對於許多典範是已知的並被用於經由使用梯度下降法來修改網路的權重。

【0062】 DCN可以是前饋網路。另外，如前述，從DCN的第一層中的神經元到下一更高層中的神經元群的連接跨第一層中的神經元被共用。DCN的前饋和共用連接可被利用於進行快速處理。DCN的計算負擔可比例如類似大小的包括回流或回饋連接的神經網路小得多。

【0063】 迴旋網路的每一層的處理可被認爲是空間不變模版或基礎投影。若輸入首先被分解成多個通道，諸如彩色影像的紅色、綠色和藍色通道，則在該輸入上訓練的迴旋網路可被認爲是三維的，其具有沿著該影像的軸的兩個空間維度以及捕捉顏色資訊的第三個維度。迴旋連接的輸出可被認爲在後續層318和320中形成特徵圖，該特徵圖（例如，320）之每一者元素從先前層（例如，318）中一定範圍的神經元以及從該多個通道中的每一個通道接收輸入。特徵圖中的值可以用非線性（諸如矯正）max(0,x)進一步處理。來自毗鄰神經元的值可被進一步池化（這對應於降取樣）並可提供額外的局部不變性以及維度縮減。亦可經由特徵圖中神經元之間的側向抑制來應用正規化，其對應於白化。

【0064】 深度學習架構的效能可隨著有更多被標記的資料點變爲可用或隨著計算能力提高而提高。現代深度神經網路用比僅僅十五年前可供典型研究者使用的計算資源多數千倍的計算資源來例行地訓練。新的架構和訓練範式可進一步推升

深度學習的效能。經矯正的線性單元可減少被稱爲梯度消失的訓練問題。新的訓練技術可減少過度擬合（over-fitting）並因此使更大的模型能夠達成更好的推廣。封裝技術可抽象出給定的感受野中的資料並進一步提升整體效能。

【0065】 圖3B是圖示示例性深度迴旋網路350的方塊圖。深度迴旋網路350可包括多個基於連通性和權重共用的不同類型的層。如圖3B所示，該示例性深度迴旋網路350包括多個迴旋塊（例如，C1和C2）。每個迴旋塊可配置有迴旋層、正規化層（LNorm）、和池化層。迴旋層可包括一或多個迴旋濾波器，其可被應用於輸入資料以產生特徵圖。儘管僅圖示兩個迴旋塊，但本案不限於此，相反，根據設計偏好，任何數目的迴旋塊可被包括在深度迴旋網路350中。正規化層可被用於對迴旋濾波器的輸出進行正規化。例如，正規化層可提供白化或側向抑制。池化層可提供在空間上的降取樣聚集以實現局部不變性和維度縮減。

【0066】 例如，深度迴旋網路的平行濾波器組可任選地基於ARM指令集被載入到SOC 100的CPU 102或GPU 104上以達成高效能和低功耗。在替換實施例中，平行濾波器組可被載入到SOC 100的DSP 106或ISP 116上。另外，DCN可存取其他可存在於SOC上的處理塊，諸如專用於感測器114和導航120的處理塊。

【0067】 深度迴旋網路350亦可包括一或多個全連接層（例如，FC1和FC2）。深度迴旋網路350可進一步包括邏輯回歸（LR）層。深度迴旋網路350的每一層之間是要被更新的權重（

未圖示）。每一層的輸出可以用作深度迴旋網路350中後續層的輸入以從在第一迴旋塊C1處提供的輸入資料（例如，影像、音訊、視訊、感測器資料及/或其他輸入資料）學習階層式特徵表示。

【0068】 圖4A是圖示可使人工智慧（AI）功能模組化的示例性軟體架構400的方塊圖。使用該架構，應用402可被設計成可使得SOC 420的各種處理塊（例如CPU 422、DSP 424、GPU 426及/或NPU 428）在該應用402的執行時操作期間執行支援計算。

【0069】 AI應用402可配置成調用在使用者空間404中定義的功能，例如，這些功能可提供對指示該設備當前操作位置的場景的偵測和辨識。例如，AI應用402可取決於辨識出的場景是否爲辦公室、報告廳、餐館、或室外環境（諸如湖泊）而以不同方式配置話筒和相機。AI應用402可向與在場景偵測應用程式設計介面（API）406中定義的庫相關聯的經編譯器代碼作出請求以提供對當前場景的估計。該請求可最終依賴於配置成基於例如視訊和定位資料來提供場景估計的深度神經網路的輸出。

【0070】 執行時引擎408（其可以是執行時框架的經編譯代碼）可進一步可由AI應用402存取。例如，AI應用402可使得執行時引擎請求特定的時間間隔的場景估計或由應用的使用者介面偵測到的事件觸發的場景估計。在使得執行時引擎估計場景時，執行時引擎可進而發送訊號給在SOC 420上執行的作業系統410（諸如Linux核心412）。作業系統410進而可使得

在CPU 422、DSP 424、GPU 426、NPU 428、或其某種組合上執行計算。CPU 422可被作業系統直接存取，而其他處理塊可經由驅動器（諸如用於DSP 424、GPU 426、或NPU 428的驅動器414-418）被存取。在示例性實例中，深度神經網路可被配置成在處理塊的組合（諸如CPU 422和GPU 426）上執行，或可在NPU 428（若存在的話）上執行。

【0071】 圖4B是圖示智慧手機452上的AI應用的執行時操作450的方塊圖。AI應用可包括預處理模組454，該預處理模組454可被配置（例如，使用JAVA程式設計語言被配置）成轉換影像456的格式並隨後對影像458進行剪裁及/或調整大小。經預處理的影像可接著被傳達給分類應用460，該分類應用460包含場景偵測後端引擎462，該場景偵測後端引擎462可被配置（例如，使用C程式設計語言被配置）成基於視覺輸入來偵測和分類場景。場景偵測後端引擎462可被配置成經由縮放（466）和剪裁（468）來進一步預處理（464）該影像。例如，該影像可被縮放和剪裁以使所得到的影像是224圖元x224圖元。這些維度可映射到神經網路的輸入維度。神經網路可由深度神經網路塊470配置以使得SOC 100的各種處理塊進一步借助深度神經網路來處理影像圖元。深度神經網路的結果可隨後被取閾（472）並被傳遞經由分類應用460中的指數平滑塊474。經平滑的結果可接著使得智能手機452的設置及/或顯示改變。

【0072】 在一種配置中，機器學習模型（諸如神經網路）被配置成用於代替網路中的一或多個層、在經壓縮網路的經壓

縮層之間插入非線性及/或微調經壓縮網路。該模型包括代替裝置、插入裝置和調諧裝置。在一個態樣，該代替裝置、插入裝置及/或調諧裝置可以是被配置成執行所敘述的功能的通用處理器102、與通用處理器102相關聯的程式記憶體、記憶體塊118、局部處理單元202、及/或路由連接處理單元216。在另一配置中，前述裝置可以是被配置成執行由前述裝置所敘述的功能的任何模組或任何設備。

【0073】 在另一配置中，機器學習模型（諸如神經網路）被配置成用於用多個經壓縮層代替網路中的一或多個層以產生經壓縮網路，從而組合的經壓縮層的感受野大小與未壓縮層的感受野相匹配。該模型亦被配置成用於經由更新這些經壓縮層中的一或多個經壓縮層中的權重值來微調經壓縮網路。該模型包括代替裝置和調諧裝置。在一個態樣，該代替裝置及/或調諧裝置可以是被配置成執行所敘述的功能的通用處理器102、與通用處理器102相關聯的程式記憶體、記憶體塊118、局部處理單元202、及/或路由連接處理單元216。在另一配置中，前述裝置可以是被配置成執行由前述裝置所敘述的功能的任何模組或任何設備。

【0074】 在又一配置中，機器學習模型（諸如神經網路）被配置成用於代替網路中的一或多個層及/或用於經由應用交替最小化程序來決定經壓縮層的權重矩陣。該模型包括代替裝置和決定裝置。在一個態樣，該代替裝置及/或決定裝置可以是被配置成執行所敘述的功能的通用處理器102、與通用處理器102相關聯的程式記憶體、記憶體塊118、局部處理單元202

、及/或路由連接處理單元216。在另一配置中，前述裝置可以是被配置成執行由前述裝置所敘述的功能的任何模組或任何設備。

【0075】 根據本案的某些態樣，每個局部處理單元202可被配置成基於網路的一或多個期望功能性特徵來決定網路的參數，以及隨著所決定的參數被進一步適配、調諧和更新來使這一或多個功能性特徵朝著期望的功能性特徵發展。

全連接層的壓縮

【0076】 圖5A-B是圖示根據本案的各態樣的全連接層和經壓縮全連接層的方塊圖。如圖5A所示，全連接層（fc）500的每一個輸出神經元504以全體到全體的方式經由突觸連接至每一個輸入神經元502（出於圖示方便起見示出全體到全體連接的子集）。fc層具有n個輸入神經元和m個輸出神經元。將第i個輸入神經元連接至第j個輸出神經元的突觸的權重由w_{ij}標示。連接n個輸入神經元和m個輸出神經元的突觸的所有權重可使用矩陣W來共同表示。輸出啓動向量y可經由將輸入啓動向量x乘以權重矩陣W並加上偏置向量b來獲得。

【0077】 圖5B圖示了示例性經壓縮全連接層550。如圖5B所示，單個全連接層用兩個全連接層代替。亦即，圖5A中名爲'fc'的全連接層用圖5B中所示的兩個全連接層'fc1'和'fc2'代替。儘管僅示出一個額外全連接層，但本案並不被如此限定，並且可使用額外全連接層來壓縮全連接層fc。例如，圖6A-B圖示了全連接層以及具有三個全連接層的經壓縮全連接層。如圖6A-B所示，未壓縮全連接層'fc'用三個全連接層'fc1'、'fc2'

和‘fc3’代替。居間層中的神經元數目r_1和r_2可基於它們達成的壓縮以及結果所得的經壓縮網路的效能來決定。

【0078】 可經由將壓縮前和壓縮後的參數總數進行比較來理解使用壓縮的優勢。壓縮前的參數數目可以等於權重矩陣W中元素數目$=nm$、加上偏置向量中的元素數目（其等於m）。因此，壓縮前的參數總數等於$nm+m$。然而，圖5B中壓縮後的參數數目等於層‘fc1’和‘fc2’中的參數數目之和，其等於$nr+rm$。取決於r的值，可達成參數數目的顯著減少。

【0079】 經壓縮網路所達成的有效變換由下式提供：

$$\boldsymbol{y} = \boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}, \qquad (1)$$

其中W_1和W_2分別是經壓縮全連接層fc1和fc2的權重矩陣。

【0080】 經壓縮層的權重矩陣可被決定以使得有效變換是原始全連接層fc所達成的變換的緊密近似：

$$\boldsymbol{y}=\boldsymbol{W}\boldsymbol{x}+\boldsymbol{b}。 \qquad (2)$$

【0081】 權重矩陣W_1和W_2可被選取成減小近似誤差，近似誤差由下式提供：

$$|\boldsymbol{W} - \boldsymbol{W}_2\boldsymbol{W}_1|^2 \qquad (3)$$

從而由經壓縮層fc1和fc2達成的式1的有效變換近似式2中指定的原始全連接層的變換。

【0082】 類似地，圖6B中壓縮後的參數數目等於$nr_1 + r_1r_2 + r_2m + m$。再次，取決於r_1和r_2的值，可達成參數數目的顯著減少。

【0083】 圖6B中所示的經壓縮網路中的經壓縮層‘fc1’、‘fc2’和‘fc3’一起達成的有效變換爲$\boldsymbol{y} = \boldsymbol{W}_3\boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}$。權重矩陣

W_1、W_2和W_3可被決定以使得有效變換是原始'fc'層（圖6A中所示）所達成的變換$y=Wx+b$的近似。

【0084】 一種用於決定經壓縮層的權重矩陣的方法是重複地應用奇異值分解（SVD）。SVD可被用來計算W的秩近似，可從該秩近似決定權重矩陣W_1和W_2。

【0085】 用於決定經壓縮層的權重矩陣W_1和W_2的另一種方法是應用交替最小化程序。根據交替最小化程序，使用隨機值來初始化矩陣並且使用最小二乘法一次一個地交替地更新經壓縮矩陣，以減小或最小化式3的目標函數並由此改善該近似。

【0086】 在一些態樣，可經由訓練經壓縮網路來決定權重矩陣。例如，可針對一組訓練實例記錄未壓縮層的輸入和輸出。隨後可使用例如梯度下降來初始化經壓縮層以產生對應於未壓縮層的輸出。

【0087】 由於經壓縮網路是原始網路的近似，因此經壓縮網路的端對端行爲可能與原始網路相比會稍微或顯著不同。結果，對於該網路被設計成完成的任務，新的經壓縮網路可能不如原始網路表現那樣好。爲了對抗效能下降，可執行對經壓縮網路的權重的微調和修改。例如，可應用經由所有經壓縮層的反向傳播以修改權重。替換地，可執行經由這些層的子集的反向傳播，或者可執行貫穿從輸出回到輸入的整個模型經由經壓縮層和未壓縮層兩者的反向傳播。可經由用一些訓練實例教導神經網路來達成微調。取決於可用性，可採用用於訓練原始網路的相同訓練實例、或不同訓練實例集合、

或舊訓練實例和新訓練實例的子集。

【0088】 在微調期間，可選擇僅微調新的經壓縮層或者所有層一起（其可被稱爲全堆疊微調）、或經壓縮層和未壓縮層的子集。如關於圖5B和6B中的實例所示的經壓縮層分別是圖5A和6A中所示的全連接層類型的另一實例。因此，可規避用於訓練經壓縮層或用於使用經壓縮網路執行推斷的新技術的設計或實現。取而代之，可以重用可用於原始網路的任何訓練和推斷平臺。

迴旋層的壓縮

【0089】 神經網路的迴旋層亦可被壓縮。用於壓縮迴旋層的辦法在一些態樣類似於以上關於全連接層所描述的。例如，如同全連接層的情形，經壓縮迴旋層的權重矩陣可被選擇以使得經壓縮層所達成的有效變換是對原始未壓縮迴旋層所達成的變換的良好近似。然而，因全連接層與迴旋層之間的架構差異而存在一些差異。

【0090】 圖7是圖示根據本案的各態樣的示例性迴旋層的方塊圖。迴旋層將n個輸入圖連接到m個輸出圖。每個圖可包括對應於影像的空間維度或音訊訊號中的時頻維度等等的神經元集合（例如，二維（2-D）網格或3-D圖）。例如，在一些態樣，該神經元集合可對應於視訊資料或體素數據（例如，磁共振成像（MRI）掃瞄）。該迴旋層將迴旋核心 W^{ij} 應用於第 i 個輸入圖 X^i，並且將結果加至第 j 個輸出圖 Y^j，其中索引 i 從 1 到 n，並且索引 j 從 1 到 m。換言之，第 j 個輸出圖 Y^j 可被表達爲：

$Y^j = \sum_{i=1}^{n} X^i * W^{ij} + B^j$。 (4)

【0091】 圖8A-B圖示了根據本案的各態樣的迴旋層的實例壓縮。參照圖8A-B，名爲'conv'的迴旋層用兩個迴旋層'conv1'和'conv2'代替。

【0092】 如以上關於全連接層所指出的，經壓縮層（代替層）的數目不限於圖8B中所示的兩個，並且可根據設計偏好代替地使用任何數目的經壓縮層。例如，圖9A-B圖示了用三個迴旋層'conv1'、'conv2'和'conv3'代替'conv'層的實例壓縮。在該情形（圖9B）中，居間層中的圖數目r_1和r_2可基於所達成的壓縮以及結果所得的經壓縮網路的效能來決定。

用兩個迴旋層代替

【0093】 圖8B中描述的經壓縮網路中的經壓縮層'conv1'和'conv2'一起達成的有效變換可被表達爲：

$$Y^j = \sum_{k=1}^{r_1} Z^k * W_2^{kj} + B^j = \sum_{i=1}^{n} X^i * \left(\sum_{k=1}^{r_1} W_1^{ik} * W_2^{kj}\right) + B^j, \quad (5)$$

其中Bj爲偏置項，並且Z^k爲添加的神經元層的啓動向量，並且在此實例中$r_1 = r$。實質上，原始網路中的'conv'層的權重矩陣W^{ij}在經壓縮網路中經由'conv1'層的權重矩陣W_1^{ik}和'conv2'層的權重矩陣W_2^{kj}近似爲：

$$W^{ij} \approx \sum_{k=1}^{r_1} W_1^{ik} * W_2^{kj}, \ i = 1,2\ldots,n, \ j = 1,2,\ldots,m。 \quad (6)$$

【0094】 爲了確保經壓縮迴旋層的有效變換（式5）是原始迴旋層的變換的緊密近似，經壓縮層的權重矩陣可被選擇以使得減小或最小化近似誤差，該近似誤差可被指定爲：

$$\sum_{i,j=1,1}^{n,m} \left(W^{ij} - \sum_{k=1}^{r_1} W_1^{ik} * W_2^{kj}\right)^2 \quad (7)$$

【0095】 再次，SVD方法、交替最小化程序或類似方法可被用來決定減小或最小化式7的目標函數的經壓縮層的權重矩陣。

經壓縮迴旋層的核心大小

【0096】 爲了使式7中提及的近似誤差有意義，矩陣 W^{ij} 應當具有與矩陣 $W_1^{ik} * W_2^{kj}$ 相同的維度。出於此目的，經壓縮迴旋層'conv1'和'conv2'的核心大小可被合適地選擇。例如，若 $k_x \times k_y$ 標示未壓縮'conv'層的核心大小，則經壓縮迴旋層'conv1'和'conv2'的核心大小可被選擇爲使得它們滿足：

$$(k_{1x} - 1) + (k_{2x} - 1) = (k_x - 1)$$

以及 (8)

$$(k_{1y} - 1) + (k_{2y} - 1) = (k_y - 1),$$

其中 $k_{1x} \times k_{1y}$ 表示'conv1'層的核心大小，並且 $k_{2x} \times k_{2y}$ 表示'conv2'層的核心大小。

【0097】 因此，具有核心大小 $k_x \times k_y$ 的迴旋層可用分別具有核心大小 $k_x \times k_y$ 和 1×1 的兩個迴旋層來代替，或反之。

【0098】 此外，在一些態樣，具有核心大小 $k_x \times k_y$ 的迴旋層可用分別具有核心大小1×1、$k_x \times k_y$ 和1×1的三個迴旋層來代替。

【0099】 在一些態樣，經壓縮迴旋層'conv1'和'conv2'的相應核心大小 k_1 和 k_2 可被選擇以滿足：

$$k_1 + k_2 - 1 = k, \quad (9)$$

其中k是未壓縮迴旋層'conv'的核心大小。下表1提供了不同的可能核心大小 k_1 和 k_2 的實例，其中未壓縮迴旋層'conv'的核心大小爲k=5。

k_1	k_2	k_1+k_2-1
5	1	5
4	2	5
3	3	5
2	4	5
1	5	5

表1

用多個迴旋層代替

【0100】迴旋層亦可用不止兩個迴旋層來代替。爲了說明方便起見，假定L個迴旋層被用來代替一個迴旋層。則，第l層的權重值W_l被選擇爲減小或最小化目標函數

$$\sum_{i,j=1,1}^{n,m}\left(W^{ij}-\sum_{k_1,k_2\ldots k_L=1,1,\ldots 1}^{r_1,r_2,\ldots,r_L} W_1^{ik_1}*W_2^{k_1k_2}*\cdots*W_L^{k_{L-1}j}\right)^2 \qquad (10)$$

其中變數r_l標示第l層的輸出圖的數目。再次，爲了使目標函數有意義，第l層的核心大小$k_{lx}\times k_{ly}$可被選擇爲使得它們滿足：

$$(k_{1x}-1)+(k_{2x}-1)+\cdots+(k_{Lx}-1)=(k_x-1)$$

以及 (11)

$$(k_{1y}-1)+(k_{2y}-1)+\cdots+(K_{Lx}-1)=(k_y-1)。$$

【0101】相應地，在一些態樣，具有核心大小$k_x\times k_y$的迴旋層用具有核心大小$k_{1x}\times k_{1y}$、$k_{2x}\times k_{2y}$等的多個迴旋層代替，以使得滿足性質$(k_{1x}-1)+(k_{2x}-1)+\cdots=(k_x-1)$和$(k_{1y}-1)+(k_{2y}-1)+\cdots=(k_y-1)$。

交替最小化程序

【0102】在特殊情形中，其中經壓縮迴旋層之一具有核心大

小1×1，並且其他經壓縮層具有與未壓縮層相同的核心大小，則SVD方法可被用來決定最佳化例如在式7和式10中陳述的目標函數的經壓縮權重矩陣。在一個態樣，此特殊情形對應於 $(k_{1x} \times k_{1y} = 1 \times 1$ 和 $k_{2x} \times k_{2y} = k_x \times k_y)$ 或 $(k_{1x} \times k_{1y} = k_x \times k_y$ 和 $k_{2x} \times k_{2y} = 1 \times 1)$。

【0103】在一般情形中，其中經壓縮層均不具有核心大小1×1，則交替最小化方法可被用來決定最佳化式7和式10的經壓縮權重矩陣。交替最小化方法包括以下步驟：

(1)用隨機值來初始化'conv1'層的權重矩陣 W_1^{ik} 和'conv2'層的權重矩陣 W_2^{kj}

(2)以交替方式求解每個經壓縮層的權重：

(a)固定'conv2'層權重，並且求解最小化式7的'conv1'層權重。

(b)固定'conv1'層權重，並且求解最小化式7的'conv2'層權重。

(3)重複步驟2直至收斂、或者達預定數目的反覆運算。

【0104】步驟2a和2b可使用標準最小二乘法程序以封閉形式求解。

【0105】注意，在使用SVD方法或交替最小化方法決定經壓縮權重之後，經壓縮網路可被微調以在一定程度上收回分類效能損耗。

局部連接層的壓縮

【0106】在迴旋層中，貫穿輸入影像圖應用相同迴旋核心，而在局部連接層中，在不同空間位置應用不同權重值。相應

地，關於迴旋層解釋的壓縮辦法可經由在不同空間位置應用壓縮方法（參見以上式4-11）來類似地應用於局部連接層。

設計參數的選擇

【0107】壓縮中涉及的若干設計參數（例如，全連接層情形中居間神經元的數目；迴旋層情形中輸出圖的數目和核心大小）可例如經由經驗式地量測所達成的壓縮以及經壓縮網路的對應功能效能來選擇。

【0108】在一個實例中，居間層中的神經元數目r可使用參數掃掠來選擇。r的值可以按16的增量從16掃掠到$\min(n,m)$。對於每個值，可決定經壓縮網路對於驗證資料集合的分類準確度。可選擇分類準確度下降可接受（例如，低於閾值）的最低值r。

【0109】在一些態樣，可使用'貪婪搜尋'方法來決定設計參數。例如，這些層中展示最大壓縮希望的一或多個層（例如，具有較大計算數目的層）可被壓縮。隨後可選擇性地壓縮供壓縮的附加層。

【0110】在一些態樣，可經由個體地壓縮每一層來決定良好的設計參數。例如，每一層可被壓縮到可能的最大程度，使得功能效能不降到低於閾值。對於每一層個體地學習的這些設計參數隨後可被用來將網路中的多個層一起壓縮。在一起壓縮多個層之後，可經由更新經壓縮層和未壓縮層中的權重值來微調網路以收回功能效能下降。

【0111】在一些態樣，經壓縮網路可針對每個所決定的參數值（例如，r）進行微調或者可僅針對最終選擇的參數值進行

微調。以下偽代碼提供了用於選擇壓縮參數的示例性啓發。

For each fully-connected layer:

max_rank = min(n,m)

for r = 1:max_rank

compress the selected fully-connected layer (leaving others uncompressed)

determine the classification accuracy

choose lowest r such that drop in accuracy is below a threshold

compress all fully-connected layers using the corresponding chosen r values

fine-tune the compressed network.

在經壓縮層之間插入非線性層

【0112】在用多個層代替一層之後，經壓縮層中的神經元可配置有相同啓動。然而，可在經壓縮層之間添加非線性層以改善網路的表示容量。此外，經由添加非線性層，可達成較高的功能效能。

【0113】在一個示例性態樣，可插入矯正器（rectifier）非線性。在此實例中，可使用max(0,x)非線性來向經壓縮層的輸出啓動應用閾值，隨後將它們傳遞給下一層。在一些態樣，矯正器非線性可以是矯正線性單元（ReLU）。亦可使用其他種類的非線性，包括abs(x)、sigmoid(x)和雙曲正切（tanh (x)）或諸如此類。

【0114】在一些態樣，可壓縮神經網路、可插入非線性層及/

或可應用微調。例如，可訓練具有9×9迴旋層的神經網路（例如，使用第一訓練實例集合）。該9×9迴旋層隨後可被壓縮，將其代替爲兩個5×5迴旋層。可在這兩個5×5迴旋層之間添加非線性層。可執行微調。在一些態樣，隨後可經由將每個5×5迴旋層代替爲兩個3×3迴旋層來進一步壓縮該網路。進一步，可在每個3×3迴旋層之間添加非線性層。

【0115】隨後可重複額外的微調和壓縮直至獲得期望效能。在以上實例中，與原始網路相比，最終的經壓縮網路具有四個3×3迴旋層及其間的非線性層，而非一個9×9迴旋層。具有三個額外迴旋層和非線性層的較深網路可達成更佳的功能效能。

【0116】圖10圖示了用於壓縮神經網路的方法1000。在方塊1002，該程序用多個經壓縮層代替神經網路中的一或多個層以產生經壓縮網路。在一些態樣，經壓縮層具有與初始未壓縮網路中的層相同的類型。例如，全連接層可由多個全連接層代替，並且迴旋層可由多個迴旋層代替。類似地，局部連接層可由多個局部連接層代替。

【0117】在方塊1004，該程序在經壓縮網路的經壓縮層之間插入非線性。在一些態樣，該程序經由向經壓縮層的神經元應用非線性啓動函數來插入非線性。非線性啓動函數可包括矯正器、絕對值函數、雙曲正切函數、sigmoid（S形）函數或其他非線性啓動函數。

【0118】此外，在方塊1006，該程序經由更新這些層中的至少一個層中的權重值來微調經壓縮網路。可使用反向傳播或

其他標準程序來執行微調。可經由更新經壓縮神經網路中的所有權重值或者經由更新經壓縮層的子集、未壓縮層的子集、或兩者的混合來執行微調。可使用訓練實例來執行微調。該等訓練實例可以與用於教導原始神經網路的那些訓練實例相同，或者可包括新的實例集合，或者爲兩者的混合。

【0119】在一些態樣，訓練實例可包括來自智慧手機或其他行動設備的資料，包括例如照片圖庫。

【0120】在一些態樣，該程序可進一步包括經由以下操作來初始化神經網路：重複地應用壓縮、插入非線性層、以及微調作爲用於初始化較深神經網路的方法。

【0121】圖11圖示了用於壓縮神經網路的方法1100。在方塊1102，該程序用多個經壓縮層代替神經網路中的至少一個層以產生經壓縮網路，從而組合的經壓縮層的感受野大小與未壓縮層的感受野相匹配。在一些態樣，未壓縮層的核心大小等於感受野大小。

【0122】此外，在方塊1104，該程序經由更新這些經壓縮層中的至少一個經壓縮層中的權重值來微調經壓縮網路。

【0123】圖12圖示了用於壓縮神經網路的方法1200。在方塊1202，該程序用多個經壓縮層代替神經網路中的一或多個層以產生經壓縮網路。在一些態樣，經壓縮層具有與初始未壓縮網路中的層相同的類型。例如，全連接層可由多個全連接層代替，並且迴旋層可由多個迴旋層代替。類似地，局部連接層可由多個局部連接層代替。

【0124】此外，在方塊1204，該程序經由應用交替最小化程

序來決定該多個經壓縮層的權重矩陣。

【0125】在一些態樣，該程序亦經由更新這些經壓縮層中的至少一個經壓縮層中的權重值來微調經壓縮網路。可經由更新經壓縮神經網路中的所有權重值或者經由更新經壓縮層的子集、未壓縮層的子集、或兩者的混合來執行微調。可使用訓練實例來執行微調。該等訓練實例可以與用於教導原始神經網路的那些訓練實例相同，或者可包括新的實例集合，或者爲兩者的混合。

【0126】可在單個階段中執行微調或者可在多個階段中執行微調。例如，當在多個階段中執行微調時，在第一階段，僅對經壓縮層的子集執行微調。隨後，在第二階段，針對經壓縮層和未壓縮層的子集執行微調。

【0127】圖13圖示了根據本案的各態樣的用於壓縮神經網路的方法1300。在方塊1302，該程序表徵機器學習網路，諸如沿多個維度的神經網路。在一些態樣，神經網路的任務效能（P）可被表徵。若神經網路是物件分類器，則任務效能可以是測試集合上被正確分類的物件的百分比。記憶體佔用（M）及/或推斷時間（T）亦可被表徵。記憶體佔用可以例如對應於神經網路的權重和其他模型參數的記憶體儲存要求。推斷時間（T）可以是在新物件被呈現給神經網路之後在該新物件的分類期間逝去的時間。

【0128】在方塊1304，該程序用多個經壓縮層代替神經網路中的一或多個層以產生經壓縮網路。在一些態樣，經壓縮層可具有與初始未壓縮網路中的層相同的類型。例如，全連接

層可由多個全連接層代替，並且迴旋層可由多個迴旋層代替。類似地，局部連接層可由多個局部連接層代替。在示例性配置中，該程序用第二層和第三層代替神經網路的第一層，以使得第三層的大小等於第一層的大小。在第三層被如此指定的情況下，經壓縮神經網路的第三層的輸出可模仿第一層的輸出。該程序可進一步指定第二層的大小（其可被標示爲(r)），以使得經壓縮神經網路（其包括第二層和第三層）的組合記憶體佔用小於未壓縮神經網路的記憶體佔用（M）。作爲記憶體佔用減小的替換或補充，第二層的大小可被選擇以使得經壓縮神經網路的推斷時間可以小於未壓縮神經網路的推斷時間（T）。

【0129】在方塊1306，該程序初始化神經網路的代替層的參數。在一種示例性配置中，投射到經壓縮神經網路的第二層的權重矩陣（W1）以及從第二層投射到經壓縮神經網路的第三層的權重矩陣（W2）可被初始化爲隨機值。在一些態樣，權重參數的初始化可涉及交替最小化程序。在一些態樣，參數的初始化可涉及奇異值分解。權重的初始設置可被稱爲初始化，因爲根據該方法的附加態樣，權重可經歷進一步修改。

【0130】在方塊1308，該程序在經壓縮網路的經壓縮層之間插入非線性。例如，該程序可在經壓縮神經網路的第二層中插入非線性。在一些態樣，該程序經由向經壓縮層的神經元應用非線性啓動函數來插入非線性。非線性啓動函數可包括矯正器、絕對值函數、雙曲正切函數、sigmoid函數或其他非

線性啓動函數。

【0131】在方塊1310，該程序經由更新這些層中的一或多個層中的權重值來微調經壓縮網路。可使用反向傳播或其他標準程序來執行微調。可經由更新經壓縮神經網路中的所有權重值或者經由更新經壓縮層的子集、未壓縮層的子集、或兩者的混合來執行微調。可使用訓練實例來執行微調。該等訓練實例可以與用於教導原始神經網路的那些訓練實例相同，或者可包括新的實例集合，或者爲兩者的混合。在示例性配置中，可對經壓縮神經網路的第二層和第三層執行微調。進而，神經網路中的所有權重可在方塊1312被調整，從而可達成較高水平的任務效能。

【0132】經微調的經壓縮神經網路可沿若干維度被表徵，包括例如任務效能（P）。若經壓縮神經網路的任務效能高於可接受閾值，則該程序可終止。若經壓縮神經網路的任務效能低於可接受閾值（1314：否），則可在方塊1316增大（r）的值（其可以是經壓縮神經網路中的層的大小）。否則（1314：是），該程序在方塊1318結束。（r）的新值可被選擇以使得記憶體佔用（M）及/或推斷時間（T）小於未壓縮神經網路。初始化參數、插入非線性、微調和表徵效能的步驟可被重複直至達到可接受任務效能。

【0133】儘管方法1300圖示了經壓縮層的大小（r）可從較小值增加到較大值直至獲得可接受任務效能的程序，但本案並不被如此限定。根據本案的各態樣，（r）的值可初始被保守地選擇以使得經壓縮神經網路的任務效能很可能獲得可接受

任務效能。該方法由此可產生可代替未壓縮神經網路使用的經壓縮神經網路，從而可實現記憶體佔用或推斷時間優勢。該方法隨後可繼續以（r）的此類越來越小的值進行操作，直至不能獲得可接受任務效能。以此方式，該方法可在執行該方法的程序期間提供越來越高的壓縮水平。

【0134】以上所描述的方法的各種操作可由能夠執行相應功能的任何合適的裝置來執行。這些裝置可包括各種硬體及/或軟體元件及/或模組，包括但不限於電路、特殊應用積體電路（ASIC）、或處理器。一般而言，在附圖中有圖示的操作的場合，那些操作可具有帶相似編號的相應配對手段功能元件。

【0135】如本文所使用的，術語「決定」涵蓋各種各樣的動作。例如，「決定」可包括演算、計算、處理、推導、研究、檢視（例如，在表、資料庫或其他資料結構中檢視）、探知及諸如此類。另外，「決定」可包括接收（例如接收資訊）、存取（例如存取記憶體中的資料）、及類似動作。而且，「決定」可包括解析、選擇、選取、確立及類似動作。

【0136】如本文所使用的，引述一列項目中的「至少一個」的短語是指這些專案的任何組合，包括單個成員。作爲實例，「a、b或c中的至少一者」意欲涵蓋：a、b、c、a-b、a-c、b-c、以及a-b-c。

【0137】結合本案所描述的各種說明性邏輯方塊、模組、以及電路可用設計成執行本文所描述功能的通用處理器、數位訊號處理器（DSP）、特殊應用積體電路（ASIC）、現場可

程式設計閘陣列訊號（FPGA）或其他可程式設計邏輯裝置（PLD）、個別閘門或電晶體邏輯、個別的硬體元件或其任何組合來實現或執行。通用處理器可以是微處理器，但在替換方案中，處理器可以是任何市售的處理器、控制器、微控制器、或狀態機。處理器亦可以被實現爲計算設備的組合，例如DSP與微處理器的組合、複數個微處理器、與DSP核心協同的一或多個微處理器、或任何其他此類配置。

【0138】結合本案所描述的方法或演算法的步驟可直接在硬體中、在由處理器執行的軟體模組中、或在這兩者的組合中體現。軟體模組可常駐在本發明所屬領域所知的任何形式的儲存媒體中。可使用的儲存媒體的一些實例包括隨機存取記憶體（RAM）、唯讀記憶體（ROM）、快閃記憶體、可抹除可程式設計唯讀記憶體（EPROM）、電子可抹除可程式設計唯讀記憶體（EEPROM）、暫存器、硬碟、可移除磁碟、CD-ROM，等等。軟體模組可包括單一指令、或許多數指令，且可分佈在若干不同的程式碼片段上，分佈在不同的程式間以及跨多個儲存媒體分佈。儲存媒體可被耦合到處理器以使得該處理器能從/向該儲存媒體讀寫資訊。替換地，儲存媒體可以被整合到處理器。

【0139】本文所揭示的方法包括用於實現所描述的方法的一或多個步驟或動作。這些方法步驟及/或動作可以彼此互換而不會脫離請求項的範疇。換言之，除非指定了步驟或動作的特定次序，否則具體步驟及/或動作的次序及/或使用可以改動而不會脫離請求項的範疇。

【0140】所描述的功能可在硬體、軟體、韌體或其任何組合中實現。若以硬體實現，則實例硬體設定可包括設備中的處理系統。處理系統可以用匯流排架構來實現。取決於處理系統的具體應用和整體設計約束，匯流排可包括任何數目的互連匯流排和橋接器。匯流排可將包括處理器、機器可讀取媒體、以及匯流排介面的各種電路連結在一起。匯流排介面可用於尤其將網路介面卡等經由匯流排連接至處理系統。網路介面卡可用於實現訊號處理功能。對於某些態樣，使用者介面（例如，按鍵板、顯示器、滑鼠、操縱桿，等等）亦可以被連接到匯流排。匯流排亦可以連結各種其他電路，諸如定時源、周邊設備、穩壓器、功率管理電路以及類似電路，它們在本發明所屬領域中是眾所周知的，因此將不再進一步描述。

【0141】處理器可負責管理匯流排和一般處理，包括執行儲存在機器可讀取媒體上的軟體。處理器可用一或多個通用及/或專用處理器來實現。實例包括微處理器、微控制器、DSP處理器、以及其他能執行軟體的電路系統。軟體應當被寬泛地解釋成意指指令、資料、或其任何組合，無論是被稱作軟體、韌體、仲介軟體、微代碼、硬體描述語言、或其他。作爲實例，機器可讀取媒體可包括隨機存取記憶體（RAM）、快閃記憶體、唯讀記憶體（ROM）、可程式設計唯讀記憶體（PROM）、可抹除可程式設計唯讀記憶體（EPROM）、電可抹除可程式設計唯讀記憶體（EEPROM）、暫存器、磁碟、光碟、硬驅動器、或者任何其他合適的儲存媒體、或其任何

組合。機器可讀取媒體可被實施在電腦程式產品中。該電腦程式產品可以包括包裝材料。

【0142】在硬體實現中，機器可讀取媒體可以是處理系統中與處理器分開的一部分。然而，如本發明所屬領域中具有通常知識者將容易領會的，機器可讀取媒體或其任何部分可在處理系統外部。作爲實例，機器可讀取媒體可包括傳輸線、由資料調制的載波、及/或與設備分開的電腦產品，所有這些皆可由處理器經由匯流排介面來存取。替換地或補充地，機器可讀取媒體或其任何部分可被整合到處理器中，諸如快取記憶體及/或通用暫存器檔可能就是這種情形。儘管所論述的各種元件可被描述爲具有特定位置，諸如局部元件，但它們亦可按各種方式來配置，諸如某些元件被配置成分散式計算系統的一部分。

【0143】處理系統可以被配置爲通用處理系統，該通用處理系統具有一或多個提供處理器功能性的微處理器、以及提供機器可讀取媒體中的至少一部分的外部記憶體，它們皆經由外部匯流排架構與其他支援電路系統連結在一起。替換地，該處理系統可以包括一或多個神經元形態處理器以用於實現本文所述的神經元模型和神經系統模型。作爲另一替換方案，處理系統可以用帶有整合在單塊晶片中的處理器、匯流排介面、使用者介面、支援電路系統、和至少一部分機器可讀取媒體的特殊應用積體電路（ASIC）來實現，或者用一或多個現場可程式設計閘陣列（FPGA）、可程式設計邏輯裝置（PLD）、控制器、狀態機、閘控邏輯、個別硬體元件、或者任

何其他合適的電路系統、或者能執行本案通篇所描述的各種功能性的電路的任何組合來實現。取決於具體應用和加諸於整體系統上的總設計約束，本發明所屬領域中具有通常知識者將認識到如何最佳地實現關於處理系統所描述的功能性。

【0144】機器可讀取媒體可包括數個軟體模組。這些軟體模組包括當由處理器執行時使處理系統執行各種功能的指令。這些軟體模組可包括傳送模組和接收模組。每個軟體模組可以常駐在單個存放裝置中或者跨多個存放裝置分佈。作爲實例，當觸發事件發生時，可以從硬驅動器中將軟體模組載入到RAM中。在軟體模組執行期間，處理器可以將一些指令載入到快取記憶體中以提高存取速度。隨後可將一或多個快取記憶體行載入到通用暫存器檔中以供處理器執行。在以下述及軟體模組的功能性時，將理解此類功能性是在處理器執行來自該軟體模組的指令時由該處理器來實現的。此外，應領會，本案的各態樣導致對實現此類態樣的處理器、電腦、機器、或其他系統的機能的改進。

【0145】若以軟體實現，則各功能可作爲一或多個指令或代碼儲存在電腦可讀取媒體上或藉其進行傳送。電腦可讀取媒體包括電腦儲存媒體和通訊媒體兩者，這些媒體包括促成電腦程式從一地向另一地轉移的任何媒體。儲存媒體可以是能被電腦存取的任何可用媒體。作爲實例而非限定，此類電腦可讀取媒體可包括RAM、ROM、EEPROM、CD-ROM或其他光碟儲存、磁碟儲存或其他磁存放裝置、或能用於攜帶或儲存指令或資料結構形式的期望程式碼且能被電腦存取的任何

其他媒體。另外，任何連接亦被正當地稱爲電腦可讀取媒體。例如，若軟體是使用同軸電纜、光纖電纜、雙絞線、數位用戶線（DSL）、或無線技術（諸如紅外（IR）、無線電、以及微波）從web網站、伺服器、或其他遠端源傳送而來，則該同軸電纜、光纖電纜、雙絞線、DSL或無線技術（諸如紅外、無線電、以及微波）就被包括在媒體的定義之中。如本文中所使用的盤（disk）和碟（disc）包括壓縮光碟（CD）、鐳射光碟、光碟、數位多功能光碟（DVD）、軟碟、和藍光®光碟，其中盤（disk）常常磁性地再現資料，而碟（disc）用鐳射來光學地再現資料。因此，在一些態樣，電腦可讀取媒體可包括非瞬態電腦可讀取媒體（例如，有形媒體）。另外，對於其他態樣，電腦可讀取媒體可包括瞬態電腦可讀取媒體（例如，訊號）。上述的組合應當亦被包括在電腦可讀取媒體的範疇內。

【0146】因此，某些態樣可包括用於執行本文中提供的操作的電腦程式產品。例如，此類電腦程式產品可包括其上儲存（及/或編碼）有指令的電腦可讀取媒體，這些指令能由一或多個處理器執行以執行本文中所描述的操作。對於某些態樣，電腦程式產品可包括包裝材料。

【0147】此外，應當領會，用於執行本文中所描述的方法和技術的模組及/或其他合適裝置能由使用者終端及/或基地台在適用的場合下載及/或以其他方式獲得。例如，此類設備能被耦合至伺服器以促成用於執行本文中所描述的方法的裝置的轉移。替換地，本文所述的各種方法能經由儲存裝置（例

如，RAM、ROM、諸如壓縮光碟（CD）或軟碟等實體儲存媒體等）來提供，以使得一旦將該儲存裝置耦合至或提供給使用者終端及/或基地台，該設備就能獲得各種方法。此外，可利用適於向設備提供本文所描述的方法和技術的任何其他合適的技術。

【0148】將理解，請求項並不被限定於以上所圖示的精確配置和元件。可在以上所描述的方法和裝置的佈局、操作和細節上作出各種改動、更換和變形而不會脫離請求項的範疇。

【符號說明】

【0149】

100 片上系統（SOC）

102 多核通用處理器（CPU）

104 圖形處理單元（GPU）

106 數位訊號處理器（DSP）

108 神經處理單元（NPU）

110 連通性塊

112 多媒體處理器

114 感測器處理器

116 影像訊號處理器（ISP）

118 專用記憶體塊

120 導航

200 系統

202 局部處理單元

204 局部狀態記憶體

206 局部參數記憶體

208 局部（神經元）模型程式（LMP）記憶體

210 用於儲存局部學習程式的局部學習程式（LLP）記憶體

212 局部連接記憶體

214 配置處理器單元

216 路由連接處理單元

300 網路

302 全連接網路

304 局部連接網路

306 迴旋網路

308 連接強度

310 連接強度

312 連接強度

314 連接強度

316 連接強度

318 後續層

320 後續層

322 輸出

326 新影像

350 深度迴旋網路

400 軟體架構

402 AI 應用

404 使用者空間

406 應用程式設計介面（API）

408 執行時引擎

410 作業系統

412 Linux 核心

414 驅動器

416 驅動器

418 驅動器

420 SOC

422 CPU

424 DSP

426 GPU

428 NPU

450 執行時操作

452 智慧手機

454 預處理模組

456 轉換影像

458 影像

460 分類應用

462 場景偵測後端引擎

464 預處理

466 縮放

468 剪裁

470 深度神經網路塊

472 取閾

474 指數平滑塊

500 全連接層（fc）

502 輸入神經元

504 輸出神經元

550 經壓縮全連接層

1000 方法

1002 方塊

1004 方塊

1006 方塊

1100 方法

1102 方塊

1104 方塊

1200 方法

1202 方塊

1204 方塊

1300 方法

1302 方塊

1304 方塊

1306 方塊

1308 方塊

1310 方塊

1312 方塊

1314 方塊

1316 方塊

1318 方塊

【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

【序列表】(請換頁單獨記載)

無

申請專利範圍

1. 一種壓縮一神經網路的方法，包括以下步驟：

用複數個經壓縮層代替該神經網路中的至少一個層以產生一經壓縮神經網路；

在該經壓縮網路的經壓縮層之間插入非線性；及

經由更新該等經壓縮層中的至少一個經壓縮層中的權重值來微調該經壓縮網路。

2. 如請求項1之方法，其中該插入非線性包括向該等經壓縮層的神經元應用一非線性啓動函數。

3. 如請求項2之方法，其中該非線性啓動函數是一矯正器、絕對值函數、雙曲正切函數或一S形函數。

4. 如請求項1之方法，其中該微調是經由更新該經壓縮神經網路中的該等權重值來執行的。

5. 如請求項4之方法，其中該微調包括更新該等經壓縮層的一子集或者未壓縮層的一子集中的至少一者中的權重值。

6. 如請求項4之方法，其中該微調是使用訓練實例來執行的，該等訓練實例包括用來訓練一未壓縮網路的一第一實例集合或一新的實例集合中的至少一者。

S

7. 如請求項1之方法，進一步包括以下步驟：

經由重複地應用壓縮、插入非線性層、以及微調作爲一用於初始化較深神經網路的方法來初始化該神經網路。

8. 一種壓縮一神經網路的方法，包括以下步驟：

用多個經壓縮層代替該神經網路中的至少一個層以產生一經壓縮神經網路，從而組合的該等經壓縮層的一感受野大小與未壓縮層的一感受野相匹配；及

經由更新至少一個經壓縮層中的權重值來微調該經壓縮網路。

9. 如請求項8之方法，其中未壓縮層的一核心大小等於該感受野大小。

10. 如請求項8之方法，其中該代替包括：用具有核心大小 $k_{1x} \times k_{1y}$、$k_{2x} \times k_{2y} \ldots k_{Lx} \times k_{Ly}$的一相同類型的多個經壓縮層來代替該神經網路中具有一核心大小 $k_x \times k_y$的至少一個層以產生該經壓縮網路，其中滿足性質 $(k_{1x}-1)+(k_{2x}-1)+\cdots=(k_x-1)$和 $(k_{1y}-1)+(k_{2y}-1)+\cdots=(k_y-1)$。

11. 如請求項10之方法，其中具有該核心大小$k_x \times k_y$的一迴旋層用分別具有該等核心大小1×1、$k_x \times k_y$和1×1的三個迴旋層代替。

12. 一種壓縮一神經網路的方法，包括以下步驟：

用複數個經壓縮層代替該神經網路中的至少一個層以產生一經壓縮神經網路；及

經由應用一交替最小化程序來決定經壓縮層的權重矩陣。

13. 如請求項12之方法，進一步包括經由更新該等經壓縮層中的至少一個經壓縮層中的權重值來微調該經壓縮網路。

14. 如請求項13之方法，其中該微調包括更新該等經壓縮層的一子集或者未壓縮層的一子集中的至少一者中的權重值。

15. 如請求項13之方法，其中該微調是在多個階段中執行的，其中在一第一階段中針對經壓縮層的一子集執行該微調，並且在一第二階段中針對經壓縮層和未壓縮層的一子集執行該微調。

16. 一種用於壓縮一神經網路的裝置，包括：

一記憶體；及

耦合至該記憶體的至少一個處理器，該至少一個處理器被配置成：

用複數個經壓縮層代替該神經網路中的至少一個層以產生一經壓縮神經網路；

在該經壓縮網路的經壓縮層之間插入非線性；及

經由更新至少一個經壓縮層中的權重值來微調該經壓縮網路。

17. 如請求項16之裝置，其中該至少一個處理器被進一步配置成經由向該等經壓縮層的神經元應用一非線性啓動函數來插入非線性。

18. 如請求項17之裝置，其中該非線性啓動函數是一矯正器、絕對值函數、雙曲正切函數或一S形函數。

19. 如請求項16之裝置，其中該至少一個處理器被進一步配置成經由更新該經壓縮神經網路中的該等權重值來執行該微調。

20. 如請求項19之裝置，其中該至少一個處理器被進一步配置成經由更新經壓縮層的一子集或者未壓縮層的一子集中的至少一者中的權重值來執行該微調。

21. 如請求項19之裝置，其中該至少一個處理器被進一步配置成經由使用訓練實例來執行該微調，該等訓練實例包括用來訓練未壓縮網路的一第一實例集合或一新的實例集合中的至少一者。

22. 如請求項16之裝置，其中該至少一個處理器被進一步配

置成經由重複地應用壓縮、插入非線性層、以及微調作爲一用於初始化較深神經網路的方法來初始化該神經網路。

23. 一種用於壓縮神經網路的裝置，包括：

一記憶體；及

耦合至該記憶體的至少一個處理器，該至少一個處理器被配置成：

用多個經壓縮層代替該神經網路中的至少一個層以產生一經壓縮神經網路，從而組合的該等經壓縮層的一感受野大小與未壓縮層的一感受野大小相匹配；及

經由更新至少一個經壓縮層中的權重值來微調該經壓縮網路。

24. 如請求項23之裝置，其中未壓縮層的一核心大小等於該感受野大小。

25. 如請求項23之裝置，其中該至少一個處理器被進一步配置成：用具有核心大小$k_{1x} \times k_{1y}$、$k_{2x} \times k_{2y} \ldots k_{Lx} \times k_{Ly}$的一相同類型的多個經壓縮層來代替該神經網路中具有一核心大小$k_x \times k_y$的至少一個層以產生該經壓縮網路，其中滿足性質$(k_{1x}-1)+(k_{2x}-1)+\cdots=(k_x-1)$和$(k_{1y}-1)+(k_{2y}-1)+\cdots=(k_y-1)$。

26. 如請求項25之裝置，其中具有該核心大小$k_x \times k_y$的一迴旋

層用分別具有該等核心大小1×1、$k_x\times k_y$和1×1的三個迴旋層代替。

27. 一種用於壓縮一神經網路的裝置，包括：

一記憶體；及

耦合至該記憶體的至少一個處理器，該至少一個處理器被配置成：

用複數個經壓縮層代替該神經網路中的至少一個層以產生一經壓縮神經網路；及

經由應用一交替最小化程序來決定經壓縮層的權重矩陣。

28. 如請求項27之裝置，其中該至少一個處理器被進一步配置成經由更新該等經壓縮層中的至少一個經壓縮層中的權重值來微調該經壓縮網路。

29. 如請求項28之裝置，其中該至少一個處理器被配置成經由更新該等經壓縮層的一子集或者未壓縮層的一子集中的至少一者中的權重值來執行該微調。

30. 如請求項28之裝置，其中該至少一個處理器被進一步配置成在多個階段中執行該微調，其中在一第一階段中針對經壓縮層的一子集執行該微調，並且在一第二階段中針對經壓縮層和未壓縮層的一子集執行該微調。

31. 一種用於壓縮神經網路的設備，包括：

用於用複數個經壓縮層代替該神經網路中的至少一個層以產生一經壓縮神經網路的裝置；

用於在該經壓縮網路的經壓縮層之間插入非線性的裝置；及

用於經由更新該等經壓縮層中的至少一個經壓縮層中的權重值來微調該經壓縮網路的裝置。

32. 一種用於壓縮一神經網路的設備，包括：

用於用多個經壓縮層代替該神經網路中的至少一個層以產生一經壓縮神經網路，從而組合的該等經壓縮層的一感受野大小與未壓縮層的一感受野相匹配的裝置；及

用於經由更新至少一個經壓縮層中的權重值來微調該經壓縮網路的裝置。

33. 一種用於壓縮一神經網路的設備，包括：

用於用複數個經壓縮層代替該神經網路中的至少一個層以產生一經壓縮神經網路的裝置；及

用於經由應用一交替最小化程序來決定經壓縮層的權重矩陣的裝置。

34. 一種其上編碼有用於一壓縮神經網路的程式碼的非瞬態電腦可讀取媒體，該程式碼被一處理器執行並且包括：

用於用複數個經壓縮層代替該神經網路中的至少一個層以產生一經壓縮神經網路的程式碼；

用於在該經壓縮網路的經壓縮層之間插入非線性的程式碼；及

用於經由更新至少一個經壓縮層中的權重值來微調該經壓縮網路的程式碼。

35. 一種其上編碼有用於壓縮一神經網路的程式碼的非瞬態電腦可讀取媒體，該程式碼被一處理器執行並且包括：

用於用多個經壓縮層代替該神經網路中的至少一個層以產生一經壓縮神經網路，從而組合的該等經壓縮層的一感受野大小與未壓縮層的一感受野大小相匹配的程式碼；及

用於經由更新至少一個經壓縮層中的權重值來微調該經壓縮網路的程式碼。

36. 一種其上編碼有用於壓縮一神經網路的程式碼的非瞬態電腦可讀取媒體，該程式碼被一處理器執行並且包括：

用於用複數個經壓縮層代替該神經網路中的至少一個層以產生一經壓縮神經網路的程式碼；及

用於經由應用一交替最小化程序來決定經壓縮層的權重矩陣的程式碼。

圖式

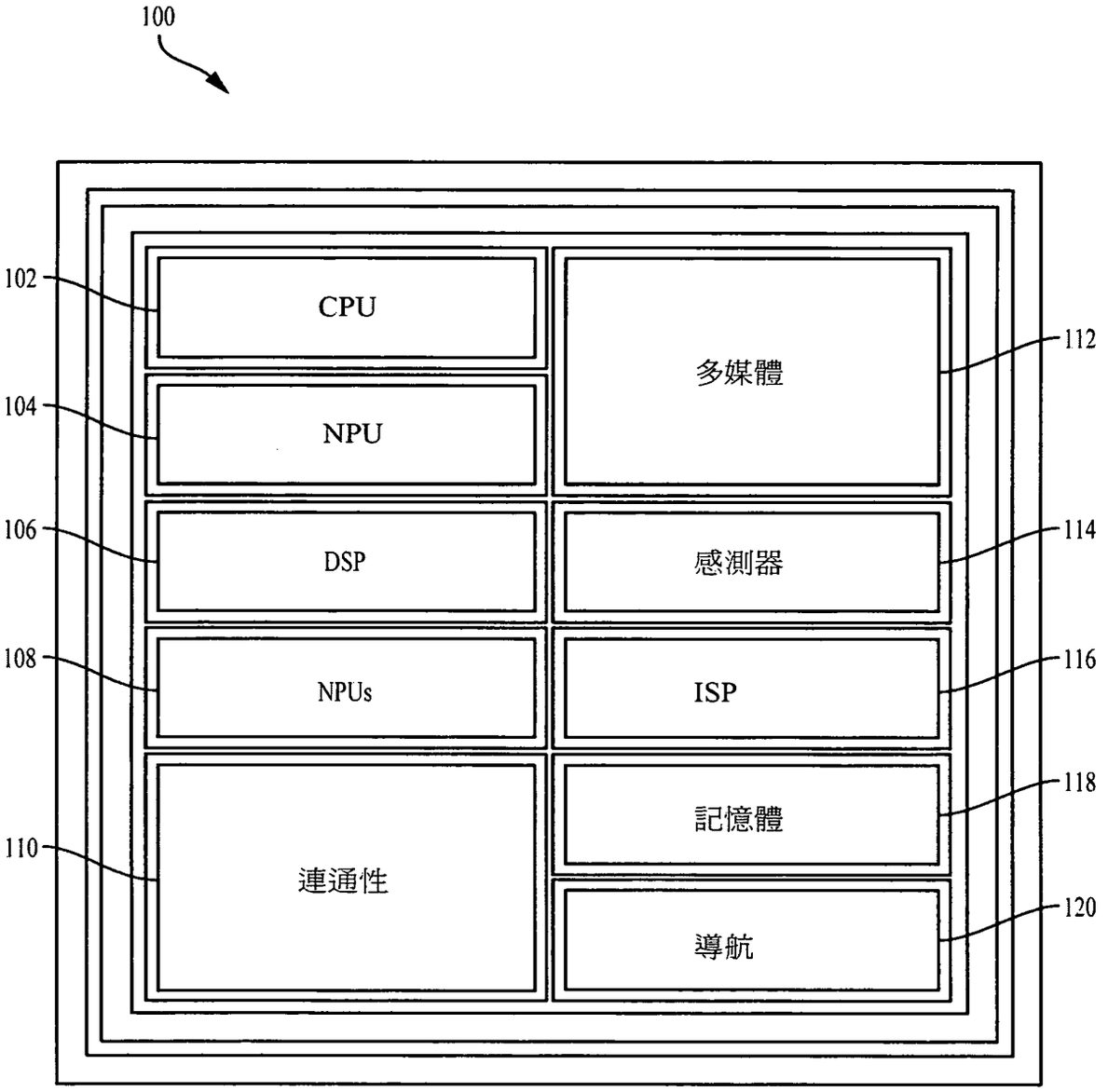
100
102
CPU
104
NPU
106
DSP
108
NPUs
110
連通性
112
多媒體
114
感測器
116
ISP
118
記憶體
120
導航

圖1

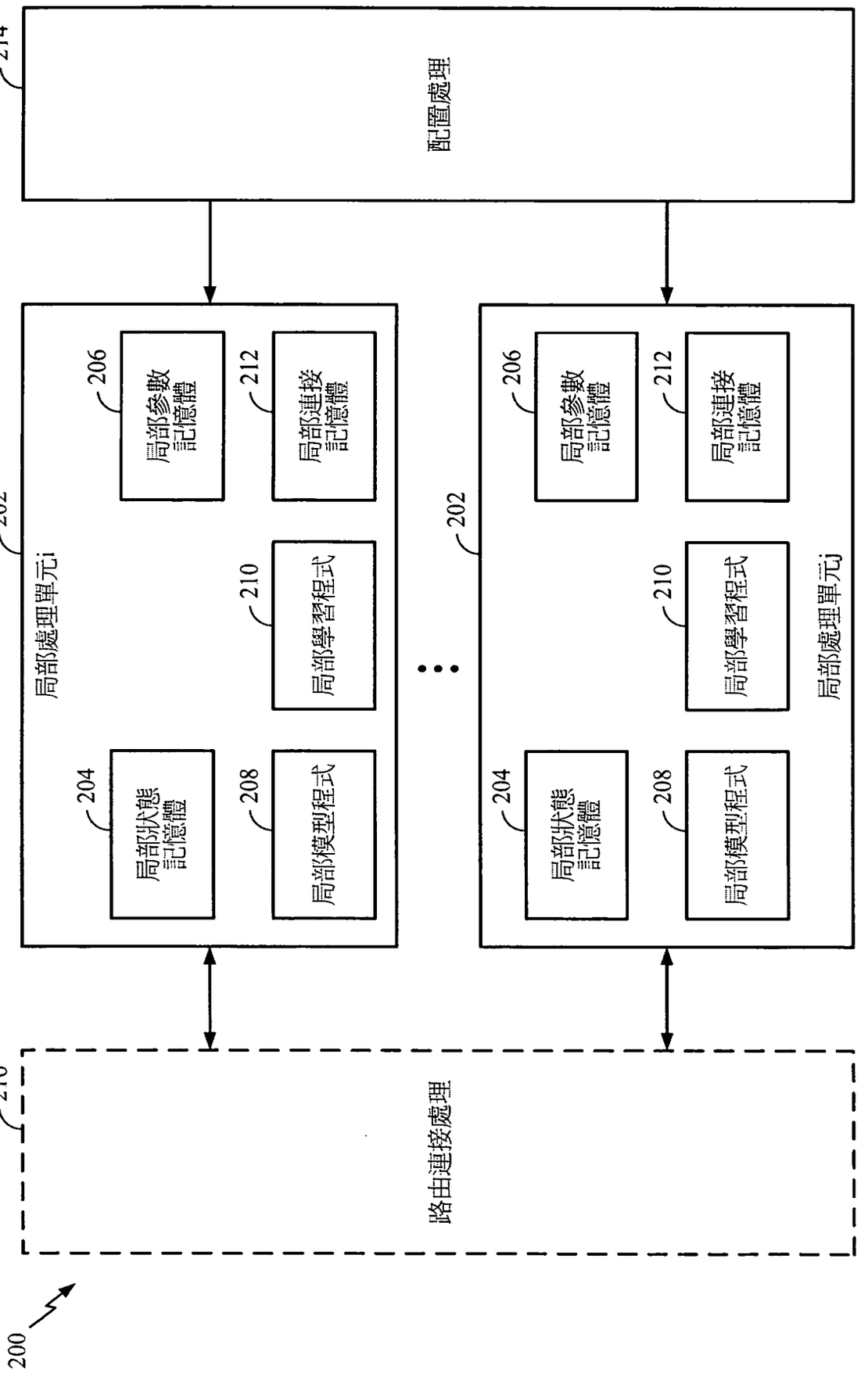
200
214
配置處理
202
局部處理單元i
206
局部參數記憶體
212
局部連接記憶體
210
局部學習程式
204
局部狀態記憶體
208
局部模型程式
202
局部處理單元j
206
局部參數記憶體
212
局部連接記憶體
210
局部學習程式
204
局部狀態記憶體
208
局部模型程式
216
路由連接處理

圖2

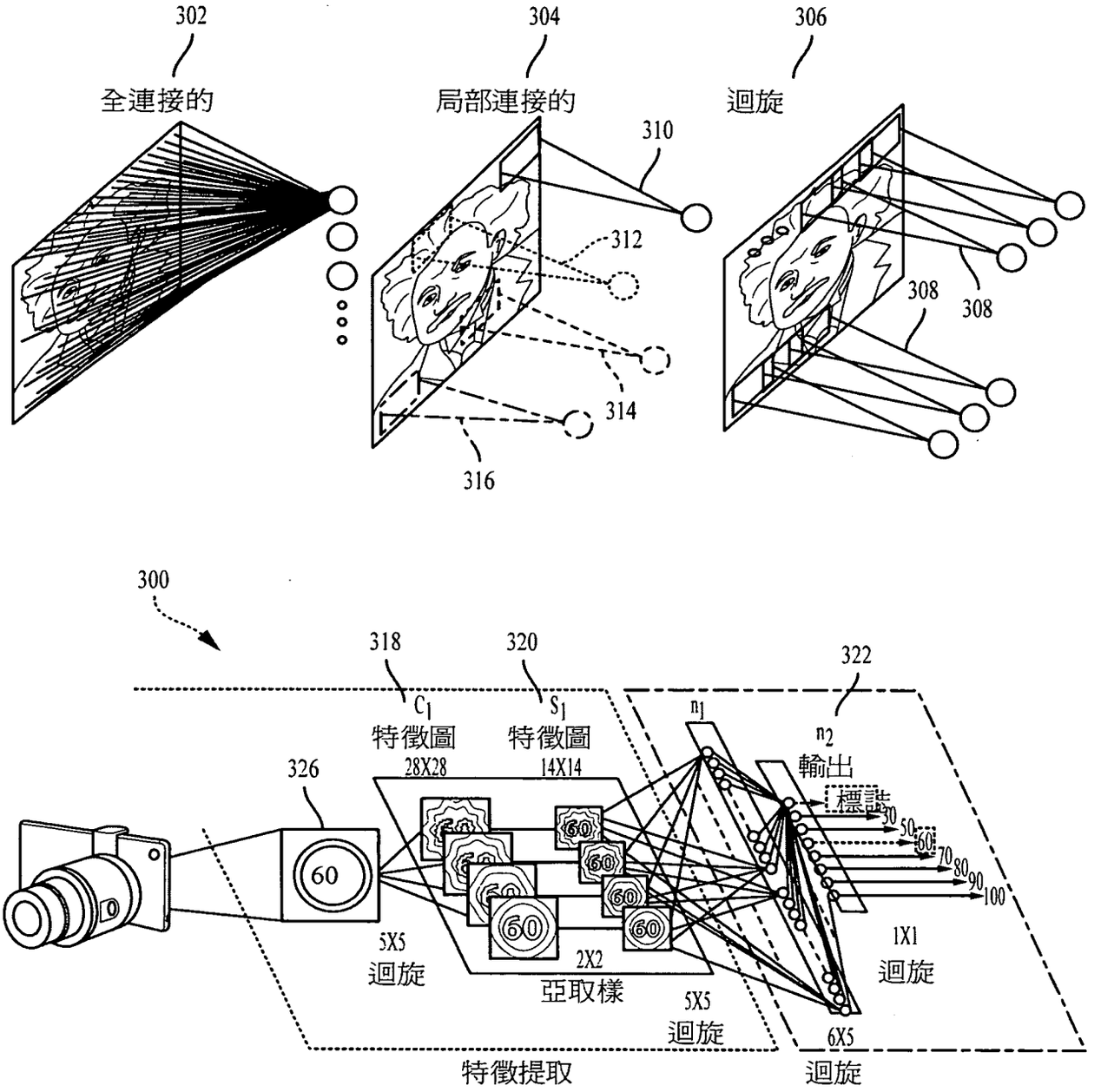
302
全連接的
304
局部連接的
306
迴旋
310
312
314
316
308
308
300
318
320
322
C_1
特徵圖
28X28
S_1
特徵圖
14X14
n_1
n_2
輸出
標誌
326
60
5X5
迴旋
2X2
亞取樣
5X5
迴旋
1X1
迴旋
6X5
迴旋
特徵提取

圖3A

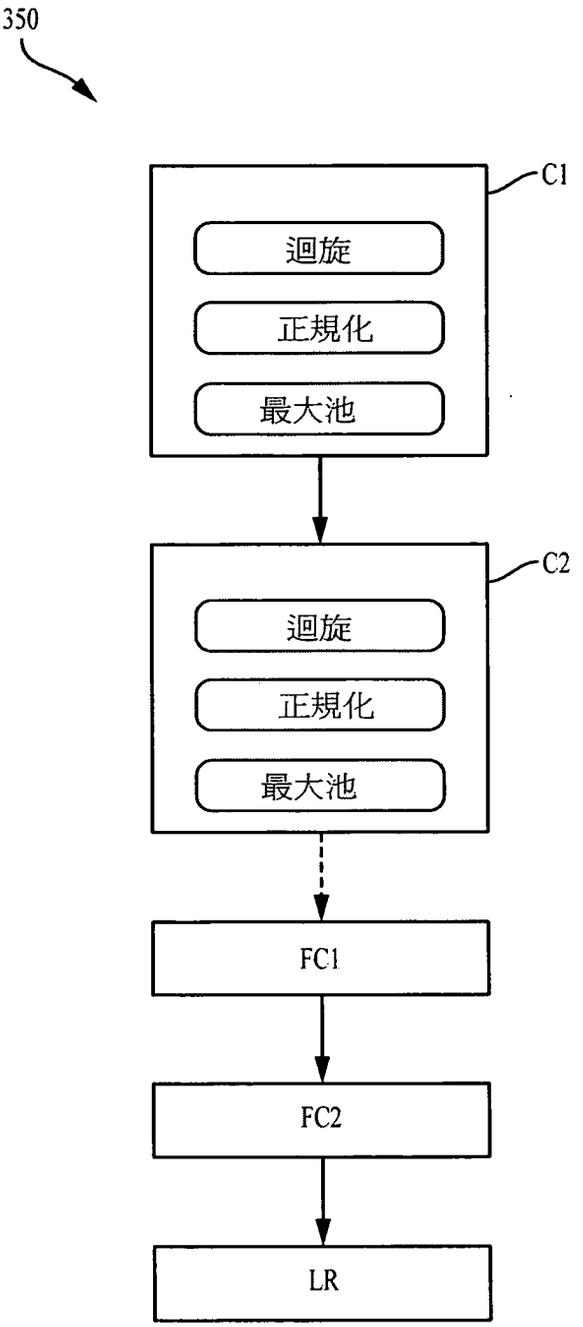

圖3B

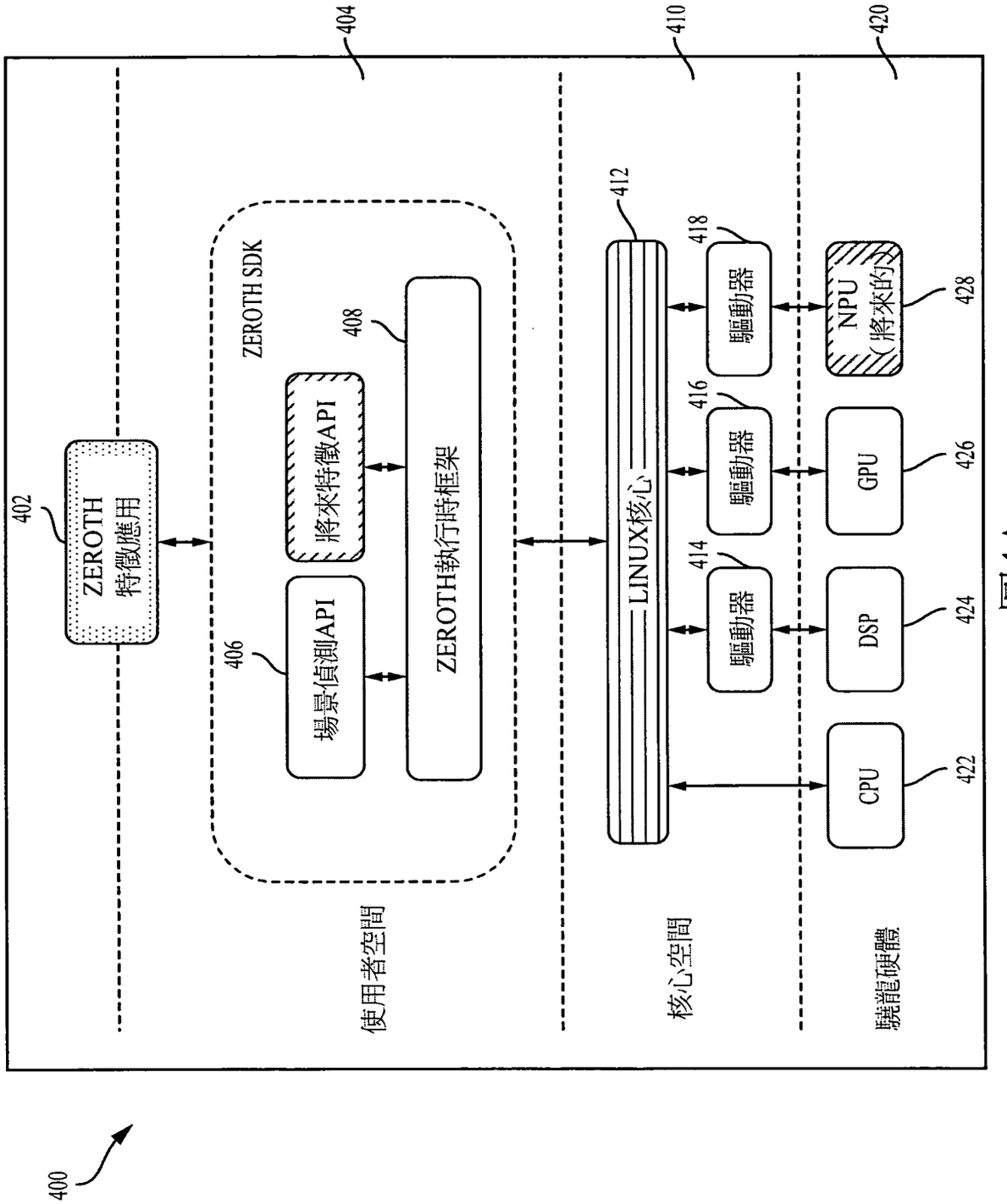
400
404
410
420
402
ZEROTH
特徵應用
ZEROTH SDK
408
將來特徵API
406
場景偵測API
ZEROTH執行時框架
412
LINUX核心
418
驅動器
416
驅動器
414
驅動器
NPU
將來的
428
GPU
426
DSP
424
CPU
422
使用者空間
核心空間
驍龍硬體

圖4A

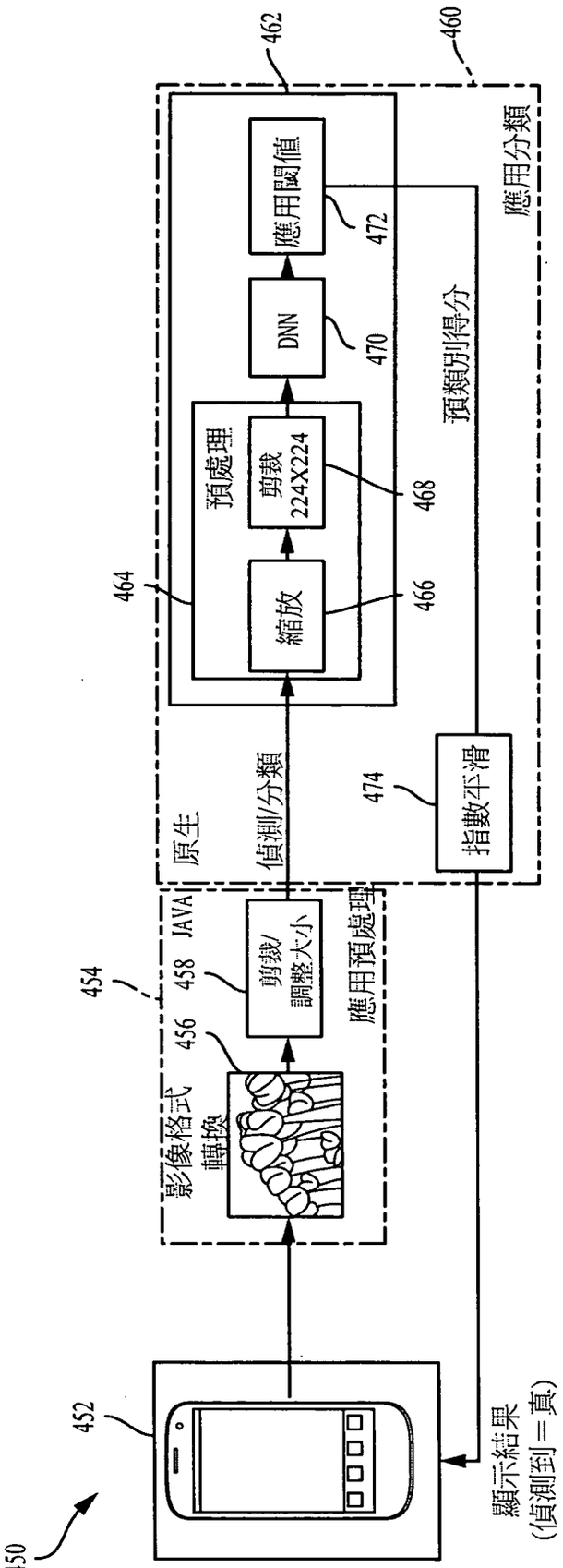
450
452
顯示結果
(偵測到＝真)
影像格式
轉換
456
JAVA
458
剪裁/
調整大小
454
應用預處理
原生
偵測/分類
464
縮放
466
預處理
剪裁
224X224
468
DNN
470
應用閾值
472
462
460
應用分類
預類別得分
474
指數平滑

圖4B

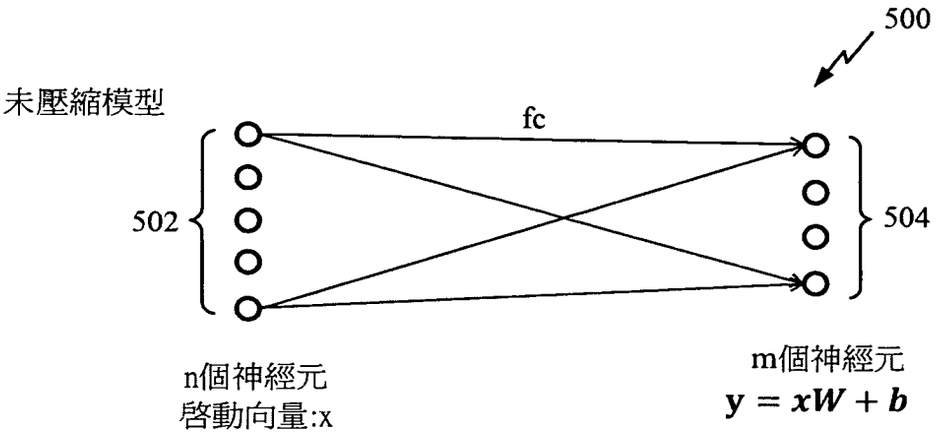
500
未壓縮模型
fc
502
504
n個神經元
啓動向量:x
m個神經元
$\mathbf{y} = \boldsymbol{xW} + \boldsymbol{b}$

圖5A

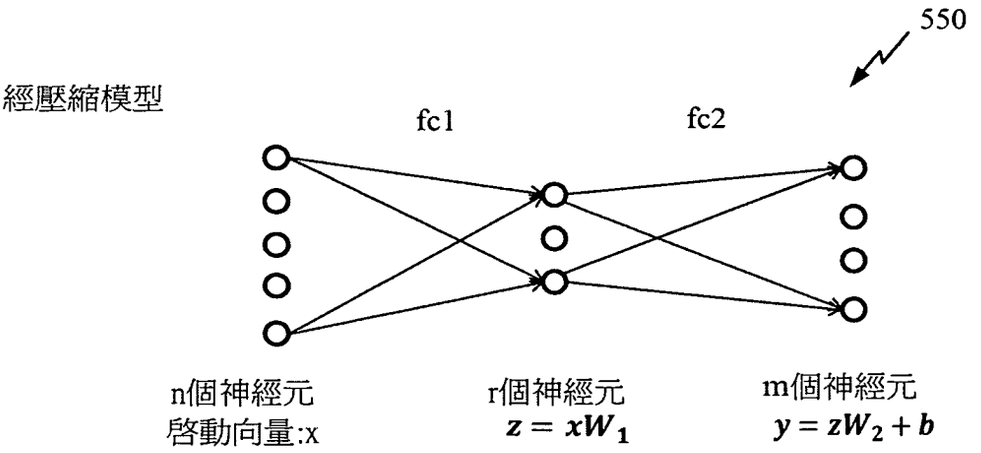
550
經壓縮模型
fc1
fc2
n個神經元
啓動向量:x
r個神經元
$\boldsymbol{z} = \boldsymbol{xW_1}$
m個神經元
$\boldsymbol{y} = \boldsymbol{zW_2} + \boldsymbol{b}$

圖5B

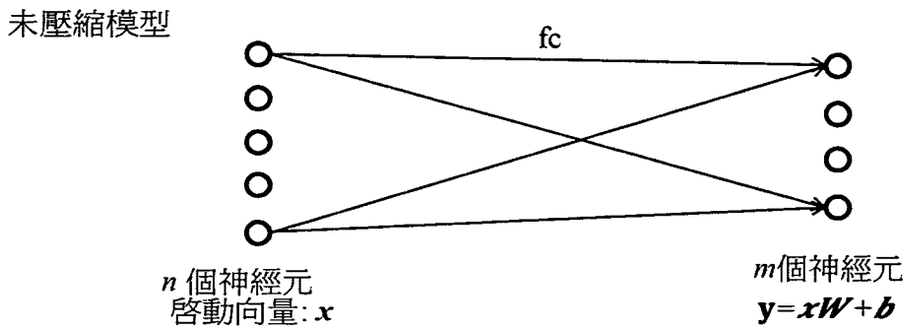
未壓縮模型
fc
n 個神經元
啓動向量: x
m個神經元
y=xW+b

圖6A

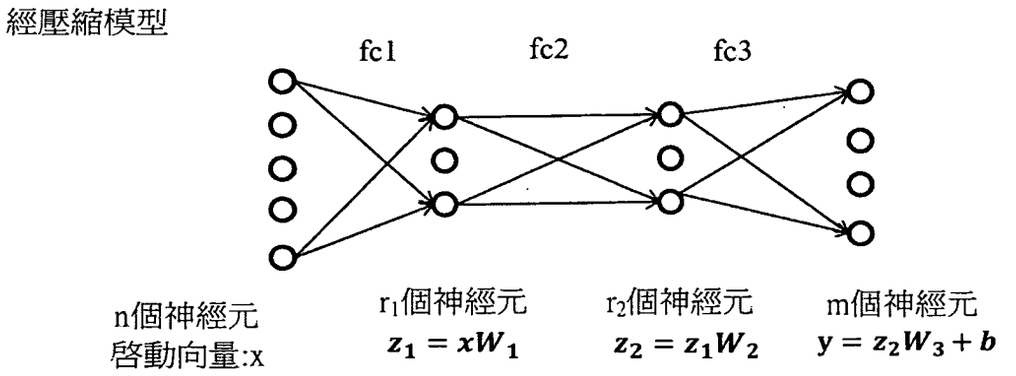
經壓縮模型
fc1
fc2
fc3
n個神經元
啓動向量:x
r1個神經元
z1 = xW1
r2個神經元
z2 = z1W2
m個神經元
y = z2W3 + b

圖6B

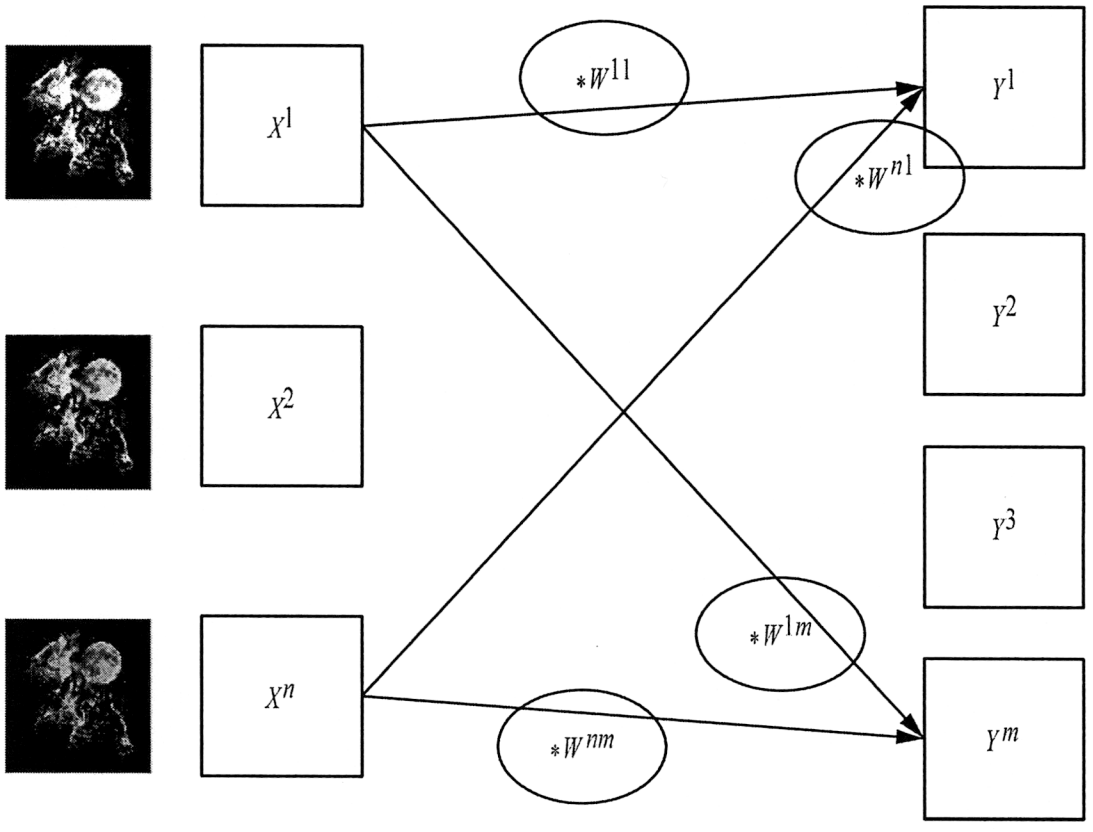

圖7

未壓縮模型

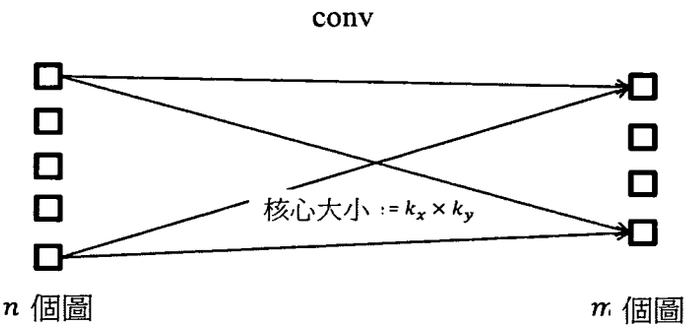

圖8A

經壓縮模型

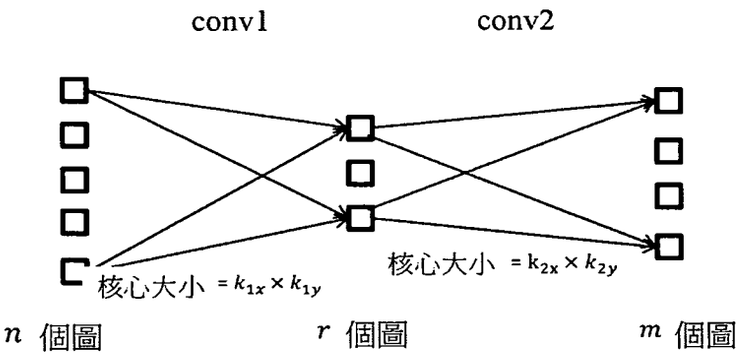

圖8B

未壓縮模型

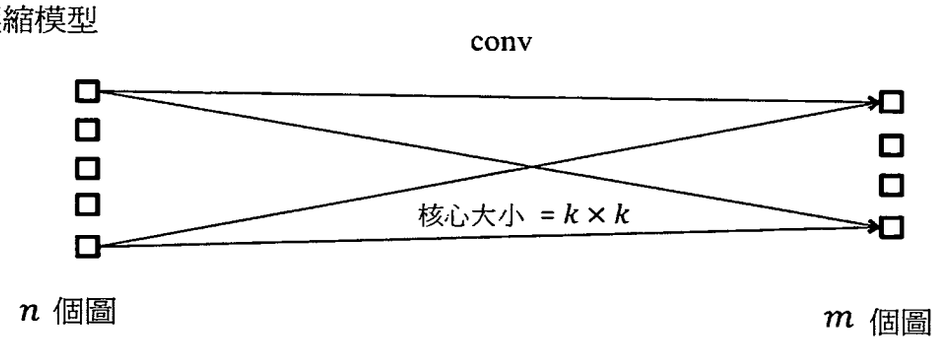

圖9A

經壓縮模型

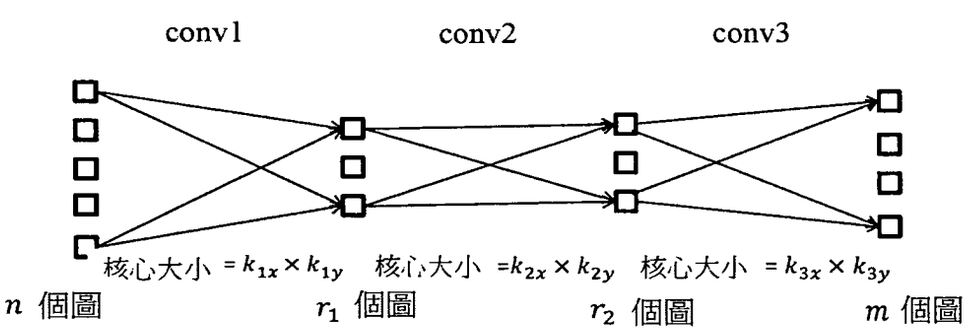

圖9B

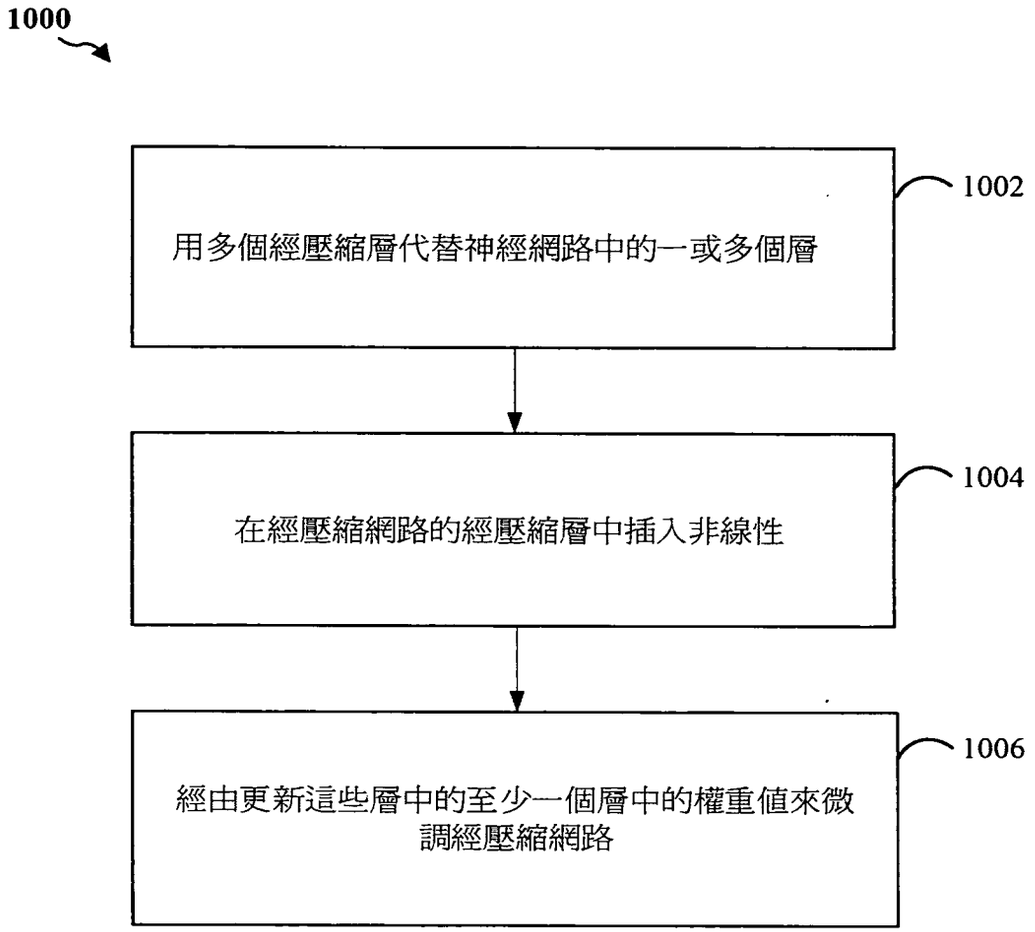
1000
用多個經壓縮層代替神經網路中的一或多個層
1002
在經壓縮網路的經壓縮層中插入非線性
1004
經由更新這些層中的至少一個層中的權重值來微調經壓縮網路
1006

圖10

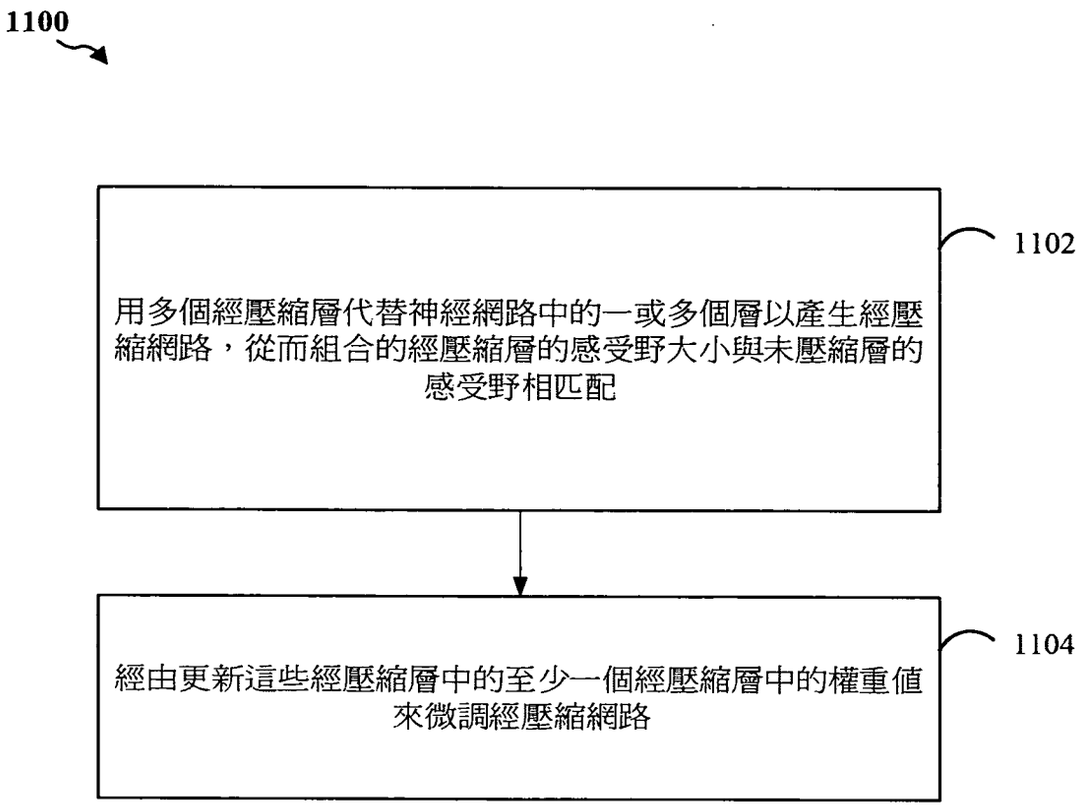
1100
用多個經壓縮層代替神經網路中的一或多個層以產生經壓縮網路，從而組合的經壓縮層的感受野大小與未壓縮層的感受野相匹配
1102
經由更新這些經壓縮層中的至少一個經壓縮層中的權重值來微調經壓縮網路
1104

圖11

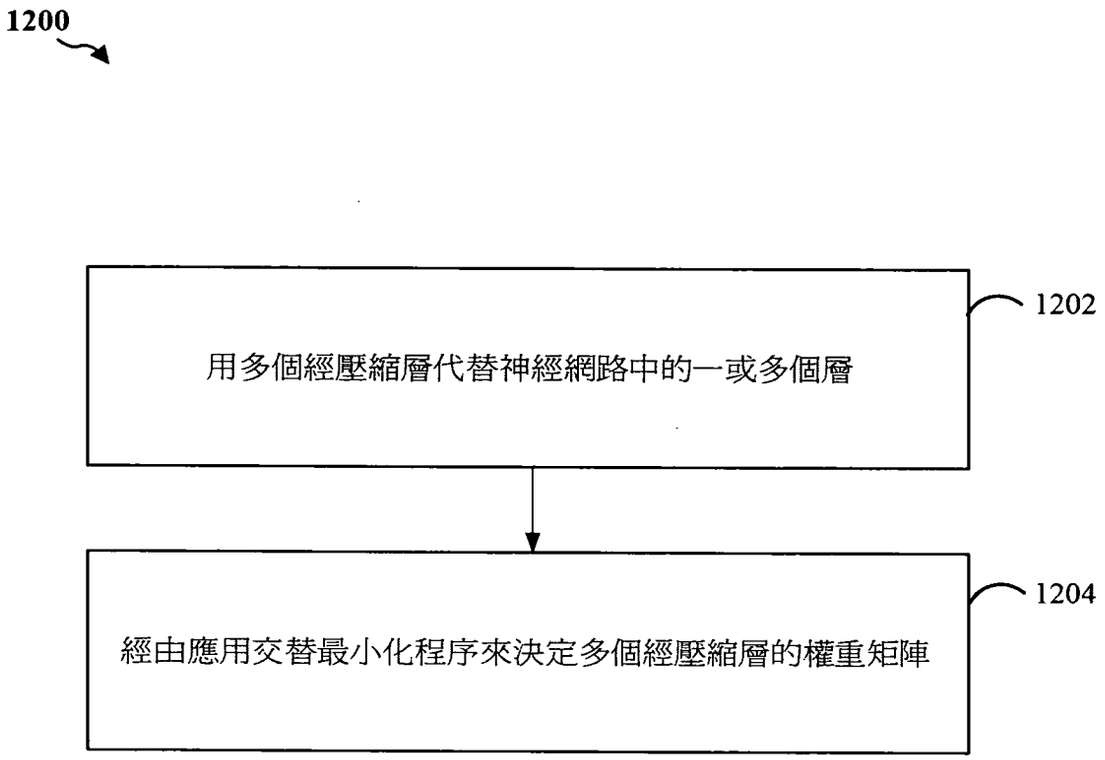
1200
用多個經壓縮層代替神經網路中的一或多個層
1202
經由應用交替最小化程序來決定多個經壓縮層的權重矩陣
1204

圖12

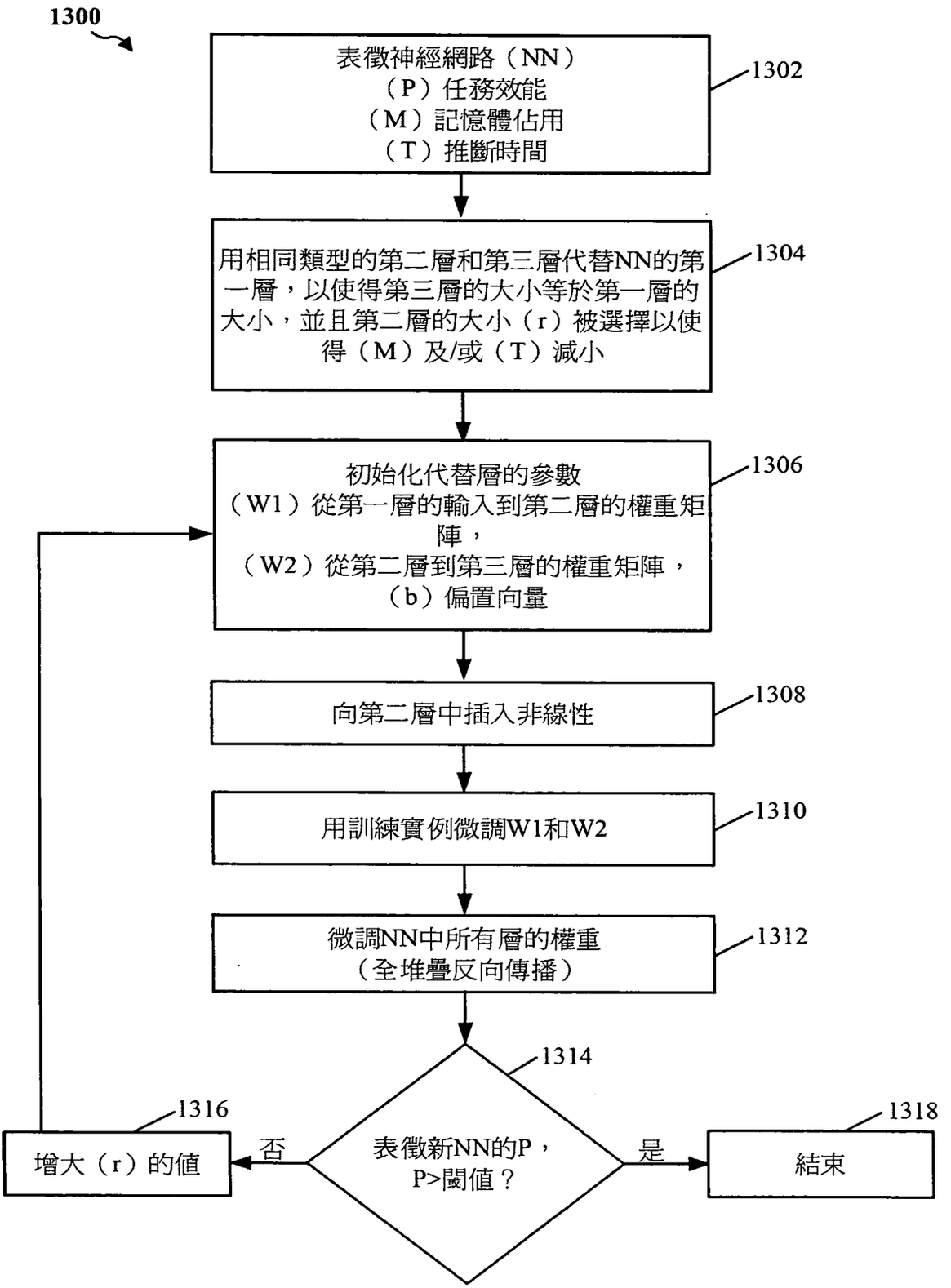
1300
表徵神經網路（NN）
（P）任務效能
（M）記憶體佔用
（T）推斷時間
1302
用相同類型的第二層和第三層代替NN的第一層，以使得第三層的大小等於第一層的大小，並且第二層的大小（r）被選擇以使得（M）及/或（T）減小
1304
初始化代替層的參數
（W1）從第一層的輸入到第二層的權重矩陣，
（W2）從第二層到第三層的權重矩陣，
（b）偏置向量
1306
向第二層中插入非線性
1308
用訓練實例微調W1和W2
1310
微調NN中所有層的權重
（全堆疊反向傳播）
1312
1314
表徵新NN的P，
P>閾值？
否
是
增大（r）的值
1316
結束
1318

圖13