I774411

(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 I774411 B**

(45)公告日： 中華民國 111 (2022) 年 08 月 11 日

(21)申請案號：110120608 (22)申請日： 中華民國 110 (2021) 年 06 月 07 日

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

(71)申請人：威盛電子股份有限公司 (中華民國) VIA TECHNOLOGIES INC. (TW)
新北市新店區中正路五三三號八樓

(72)發明人：郭王鼎志 GOUWANG, DING-ZHI (TW)

(74)代理人：吳豐任；戴俊彥

(56)參考文獻：

TW	202004569A	CN	109348707A
CN	111340227A	US	2020/0272905A1

審查人員：李惟任

申請專利範圍項數：12 項 圖式數：4 共 16 頁

(54)名稱

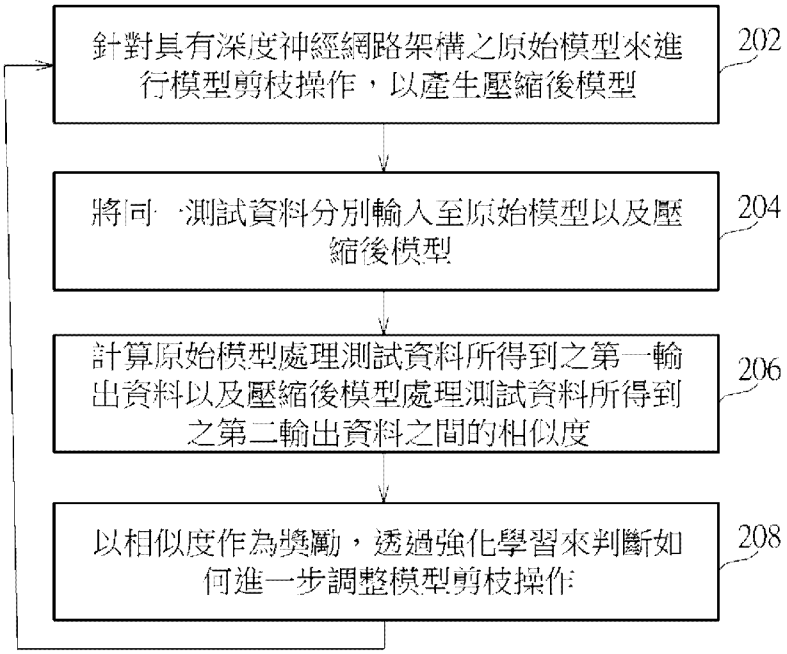
模型壓縮方法以及模型壓縮系統

(57)摘要

一種模型壓縮方法包含：針對具有一深度神經網路架構之一原始模型來進行一模型剪枝操作，以產生一壓縮後模型；將同一測試資料分別輸入至該原始模型以及該壓縮後模型；計算該原始模型處理該測試資料所得到之一第一輸出資料以及該壓縮後模型處理該測試資料所得到之一第二輸出資料之間的相似度；以及以該相似度作為獎勵，透過強化學習來判斷如何進一步調整該模型剪枝操作。

A model compression method includes: performing a model pruning operation upon an original model with deep neural network architecture to generate a compressed model, feeding a same test data into the original model and the compressed model, estimating similarity between a first output data obtained from processing the test data by the original model and a second output data obtained from processing the test data by the compressed model, and determining how to further adjust the model pruning operation through reinforcement learning that uses the similarity as a reward.

指定代表圖：



第2圖

符號簡單說明：

202,204,206,208:步驟

專利案號：110120608

公告本

I774411

【發明摘要】

【中文發明名稱】模型壓縮方法以及模型壓縮系統

【英文發明名稱】MODEL COMPRESSION METHOD AND MODEL COMPRESSION SYSTEM

【中文】

一種模型壓縮方法包含：針對具有一深度神經網路架構之一原始模型來進行一模型剪枝操作，以產生一壓縮後模型；將同一測試資料分別輸入至該原始模型以及該壓縮後模型；計算該原始模型處理該測試資料所得到之一第一輸出資料以及該壓縮後模型處理該測試資料所得到之一第二輸出資料之間的相似度；以及以該相似度作為獎勵，透過強化學習來判斷如何進一步調整該模型剪枝操作。

【英文】

A model compression method includes: performing a model pruning operation upon an original model with deep neural network architecture to generate a compressed model, feeding a same test data into the original model and the compressed model, estimating similarity between a first output data obtained from processing the test data by the original model and a second output data obtained from processing the test data by the compressed model, and determining how to further adjust the model pruning operation through reinforcement learning that uses the similarity as a reward.

【指定代表圖】第（ 2 ）圖。

【代表圖之符號簡單說明】

202, 204, 206, 208:步驟

【特徵化學式】

無

【發明說明書】

【中文發明名稱】模型壓縮方法以及模型壓縮系統

【英文發明名稱】MODEL COMPRESSION METHOD AND MODEL COMPRESSION SYSTEM

【技術領域】

【0001】 本發明係有關於模型壓縮，尤指一種透過以相似度作為獎勵的強化學習機制來調整模型剪枝操作的模型壓縮方法與模型壓縮系統。

【先前技術】

【0002】 在模型壓縮當中有個技術是透過現有的模型(教師模型)去訓練一個更小的模型(學生模型)，教師模型通常參數量較大且不易於部屬在現有設備上，所以透過這種方式來訓練一個能力相似的小模型並部屬到移動設備上，然而這種方式多半還是必須手動設計學生模型的參數，因此亟需一種能夠自動找尋合適的學生模型(亦即壓縮後模型)的方式。

【發明內容】

【0003】 因此，本發明的目的之一在於提出一種透過以相似度作為獎勵的強化學習機制來進行模型剪枝操作的模型壓縮方法與模型壓縮系統。

【0004】 在本發明的一實施例中，揭露一種模型壓縮方法。該模型壓縮方法包含：針對具有一深度神經網路架構之一原始模型來進行一模型剪枝操作，以產生一壓縮後模型；將同一測試資料分別輸入至該原始模型以及該壓縮後模型；計算該原始模型處理該測試資料所得到之一第一輸出資料以及該壓縮後模型處理該測試資料所得到之一第二輸出資料之間的相似度；以及以該相似度作為獎勵，透過強化學習來判斷如何進一步調整該模型剪枝操作。

【0005】　在本發明的另一實施例中，揭露一種模型壓縮系統。該模型壓縮系統包含一儲存裝置以及一處理器。該儲存裝置用以儲存一程式碼。該處理器用以載入並執行該程式碼，以執行以下操作：針對具有一深度神經網路架構之一原始模型來進行一模型剪枝操作，以產生一壓縮後模型；將同一測試資料分別輸入至該原始模型以及該壓縮後模型；計算該原始模型處理該測試資料所得到之一第一輸出資料以及該壓縮後模型處理該測試資料所得到之一第二輸出資料之間的相似度；以及以該相似度作為獎勵，透過強化學習來判斷如何進一步調整該模型剪枝操作。

【0006】　本發明模型壓縮方法採用相似度來作為模型剪枝(模型壓縮)的依據，因此使用者無需提供標記過的資料來作為測試資料，可減少資料標記的成本與時間，另外，使用者無需提供測試原始程式碼，可直接輸入模型來進行壓縮，故能有效推廣模型壓縮的應用，再者，壓縮後留下的泛化特徵也不易產生過擬合。

【圖式簡單說明】

【0007】

第1圖為根據本發明一實施例之模型壓縮系統的示意圖。

第2圖為本發明一實施例之模型壓縮方法的流程圖。

第3圖為第2圖所示之模型壓縮方法的操作示意圖。

第4圖為第3圖所示之原始模型與壓縮後模型所具備之卷積神經網路架構的示意圖。

【實施方式】

【0008】　第1圖為根據本發明一實施例之模型壓縮系統的示意圖。如第1圖所示，模型壓縮系統100包含一處理器102以及一儲存裝置104。儲存裝置104用以儲存一程式碼Code_MC，例如儲存裝置104可以是傳統硬碟、固態硬碟、記憶體等等，但本發明並不以此為限。處理器102可載入並執行程式碼Code_MC，以執行第2圖所示之模型壓縮方法中的各個步驟。

【0009】　請一併參閱第2圖與第3圖。第2圖為本發明一實施例之模型壓縮方法的流程圖。第3圖為第2圖所示之模型壓縮方法的操作示意圖。請注意，假若可以得到相同結果，則模型壓縮方法不一定要完全遵照第2圖所示之步驟來依序執行，此外，根據設計需求及/或應用需求，模型壓縮方法亦可修改來新增其它步驟。於步驟202，針對具有一深度神經網路(deep neural network)架構之一原始模型302來進行一模型剪枝(model pruning)操作，以產生一壓縮後模型(compression model)304。舉例來說，原始模型302可以是基於卷積神經網路(convolution neural network, CNN)架構進行訓練所得到的模型，而模型剪枝的目標是只需保留重要的權重而刪除影響較小的權重，換言之，相較於原始模型302，壓縮後模型304會具有較少的參數數量，故可以降低計算成本與儲存空間，如此一來，便可將壓縮後模型304部署至運算能力有限的產品端，像是手機、邊緣裝置(edge device)等等。此外，本發明的模型剪枝操作也希望能讓壓縮後模型302的輸出能盡量趨近原始模型302的輸出，進一步內容將於後詳述。

【0010】　於步驟204，將同一測試資料308分別輸入至原始模型302以及壓縮後模型304來進行處理，換言之，基於同一測試資料308，原始模型302的輸出以及壓縮後模型304的輸出便可用來評估壓縮後模型304是否與原始模型302相似。

【0011】　於步驟206，計算原始模型302處理測試資料308所得到之輸出資料D1以及壓縮後模型304處理測試資料308所得到之輸出資料D2之間的相似度(similarity)，因此，該相似度的數值可代表修剪後的模型的輸出特徵是否與修剪前的模型的輸出特徵相似。

【0012】　於步驟208，以步驟206所計算得到的相似度作為獎勵(reward)，透過強化學習(reinforcement learning)來判斷如何進一步調整模型剪枝操作。舉例來說，強化學習的主體(agent)可採用深度確定性策略梯度(deep deterministic policy gradient, DDPG)演算法來決定所要採取的動作(action)，其中該動作係用以選擇壓縮的部位，進而達到調整模型剪枝操作的目的。在其他實施例中，亦可使用其他演算法（例如截止自然策略梯度(Truncated Natural Policy Gradient, TNPG)演算法、交叉熵(Cross Entropy Method, CEM)演算法等）來決定所要採取的動作。

【0013】　舉例來說，假設使用者所輸入之原始模型302是一個由3個卷積層(convolution layer)所組成的結構，且所具有的通道尺寸(channel size)分別為[32, 64, 128]。一開始初始化的主體306根據原始模型302的模型信息（例如包括輸入尺寸、各層卷積核尺寸、各層浮點運算次數等等）給出3層各自的初始壓縮率為[60%, 40%, 70%]，因此，本發明模型壓縮方法使用強化學習的方式來對原始模型302進行壓縮，使得壓縮後模型304所具有的通道尺寸分別為[12, 38, 38]，並得到相似度為0.3，接著，相似度為0.3的結果會回饋給主體306，由主體306判斷接下來的壓縮方向（例如調整壓縮的部位），基於該相似度以及該模型信息調整所期望的各層壓縮率，後續模型壓縮操作透過調整後的模型剪枝操作來對原始模型302進行壓縮，使得壓縮後模型304所具有的通道尺寸分別為[14, 32, 64]，並得

到相似度為0.4；上述模型壓縮操作會迭代執行，以透過強化學習的方式來得到具有較高相似度的壓縮後模型304。

【0014】 於本發明之一實施方式中，模型壓縮方法可參照基於自動化機器學習之模型壓縮(AutoML for Model Compression, AMC)的習知架構來實作模型剪枝，但並不以此為限，本領域技術人員可以知道其他多種模型壓縮方法，於此不再贅述。習知模型壓縮架構是將壓縮後模型的輸出作為獎勵，以讓強化學習來判斷如何進一步對原始模型進行壓縮，進一步來說，習知架構採用準確率(accuracy)來作為強化學習之主體的獎勵，為了準確率的計算，需要使用者提供標記過的資料(labeled data)來作為饋入至壓縮後模型的測試資料，以便透過標記所提供的資訊來得知壓縮後模型之輸出的準確率，然而，對使用者而言，資料的標記相當費時費工，此外，計算準確率時，一般會取壓縮後模型之輸出中的最大值來跟標記進行比較，因此，準確率的計算根本不在意壓縮後模型之輸出中除了最大值以外的其它數值，假若輸入資料較難判斷時，壓縮後模型之輸出中的數值會彼此十分接近，單用準確率來作為強化學習之主體的獎勵，可能造成模型過度自信並且損失部分特徵的判斷能力，因此造成類似過擬合(overfitting)的結果或者是與原始模型不相同的輸出，再者，為了得知要採用哪種準確率的算法，習知架構亦需要使用者提供測試原始程式碼(source code)。

【0015】 相較於習知架構採用準確率來作為強化學習之主體的獎勵，本發明模型壓縮方法改用相似度來作為強化學習之主體的獎勵，並透過強化學習(例如DDPG演算法)進行模型剪枝的調整。模型剪枝(模型壓縮)的最主要目的是使壓縮後模型304能跟使用者所提供的原始模型302相似，因此，本發明模型壓縮方法可將原始模型302的輸出資料D1與壓縮後模型304的輸出資料D2進行相似性的

比較，以作為模型剪枝(模型壓縮)的依據。

【0016】 於本發明之一實施例中，相似度可藉由計算原始模型X之輸出與壓縮後模型Y之輸出的皮爾森相關係數(Pearson's correlation coefficient)來得到，例如：

【0017】 Output matrix of X= [1.0, 2.0, 3.0]

【0018】 Output matrix of Y= [2.0, 20.0, 38.0]

【0019】 Pearson's correlation coefficient $\rho(X,Y) = \frac{\sum((x_i - \bar{x}) * (y_i - \bar{y}))}{\sqrt{\sum(x_i - \bar{x})^2} * \sqrt{\sum(y_i - \bar{y})^2}} = 1.0$

【0020】 於本發明之另一實施例中，相似度可藉由計算原始模型X之輸出與壓縮後模型Y之輸出的餘弦相似度(Cosine similarity)來得到，例如：

【0021】 Output matrix of X= [1.0, 2.0, 3.0]

【0022】 Output matrix of Y= [2.0, 20.0, 38.0]

【0023】 Cosine similarity $(X,Y) = \frac{\sum x_i * y_i}{\sqrt{\sum x_i^2} * \sqrt{\sum y_i^2}} = 0.9698612260388879$

【0024】 然而，上述僅作為範例說明之用，並非用來作為本發明的限制條件，實際上，本發明模型壓縮方法亦可根據設計需求及/或應用需求來採用其它適合的相似度算法，這些設計上的變化亦落入本發明的範疇。

【0025】 如上所述，相似度的計算是基於原始模型302與壓縮後模型304各自的輸出資料D1、D2，於本實施例中，假若原始模型302是基於卷積神經網路架構進行訓練所得到的模型，則模型剪枝操作(所要進行壓縮的部位是由強化學習之主體306所採取的動作來選取)僅會施加於卷積層，因此輸出資料D1、D2可以

是卷積神經網路架構中位於卷積層後面之任一層的輸出。第4圖為第3圖所示之原始模型302與壓縮後模型304所具備之卷積神經網路架構的示意圖。如圖所示，卷積神經網路架構400包含有輸入層(input layer)、卷積層404_1～404_N (N≧1)、池化層(pooling layer)406_1～406_N (N≧1)，全連接層(filly-connected layer)408_1～408_M (M≧1)以及輸出層(output layer)410。於本發明之一實施例中，輸出資料D1可以是原始模型302之一全連接層(例如408_i，1≦i≦M)的輸出，以及輸出資料D2可以是壓縮後模型304之同一全連接層(例如408_i，1≦i≦M)的輸出。於本發明之另一實施例中，輸出層410為最後一層，並且會執行 Softmax 函式以使得全連接層408_M之所有節點輸出的機率分佈總和為 1，此外，輸出資料D1可以是原始模型302之最後一層的Softmax函式輸出，以及輸出資料D2可以是壓縮後模型304之最後一層的Softmax函式輸出。請注意，第4圖所示之卷積神經網路架構400僅作為範例說明之用，並非作為本發明的限制條件，實作上，本發明模型壓縮方法亦可適用於其它神經網路架構，這些設計上的變化亦落入本發明的範疇。

【0026】 綜上所述，相似度的計算是基於原始模型與壓縮後模型各自的輸出資料，故無需將壓縮後模型的輸出資料跟測試資料的標記進行比較，換言之，相較於習知架構採用準確率來作為強化學習之主體的獎勵而需要使用者提供標記過的資料來作為測試資料，本發明模型壓縮方法採用相似度來作為強化學習之主體的獎勵而可以不使用標記過的資料來作為測試資料（亦即，測試資料308是未標記過的資料(non-labeled data)），由於測試資料308不用包含標記，故能減少資料標記的成本與時間。此外，本發明模型壓縮方法採用相似度的計算，故使用者無需提供測試原始程式碼，可以直接輸入模型來進行壓縮，因此能有效推廣模型壓縮的應用。再者，本發明模型壓縮方法改用相似度來作為強化學習

之主體的獎勵，故壓縮後留下的泛化特徵也不易產生過擬合。

以上所述僅為本發明之較佳實施例，凡依本發明申請專利範圍所做之均等變化與修飾，皆應屬本發明之涵蓋範圍。

【符號說明】

【0027】

100:模型壓縮系統

102:處理器

104:儲存裝置

202, 204, 206, 208:步驟

302:原始模型

304:壓縮後模型

306:主體

308:測試資料

400:卷積神經網路架構

402:輸入層

404_1, 404_N:卷積層

406_1, 406_N:池化層

408_1, 408_M:全連接層

410:輸出層

Code_MC:程式碼

D1, D2:輸出資料

【發明申請專利範圍】

【請求項1】 一種模型壓縮方法，包含：

針對具有一深度神經網路架構之一原始模型來進行一模型剪枝操作，以產生一壓縮後模型；

將同一測試資料分別輸入至該原始模型以及該壓縮後模型；

計算該原始模型處理該測試資料所得到之一第一輸出資料以及該壓縮後模型處理該測試資料所得到之一第二輸出資料之間的相似度；以及

以該相似度作為獎勵，透過強化學習來判斷如何進一步調整該模型剪枝操作。

【請求項2】 如請求項1所述之模型壓縮方法，其中該測試資料係為未標記過的資料。

【請求項3】 如請求項1所述之模型壓縮方法，其中該相似度係藉由計算該第一輸出資料與該第二輸出資料之一皮爾森相關係數來得到。

【請求項4】 如請求項1所述之模型壓縮方法，其中該相似度係藉由計算該第一輸出資料與該第二輸出資料之一餘弦相似度來得到。

【請求項5】 如請求項1所述之模型壓縮方法，其中該第一輸出資料係為該原始模型之一全連接層的輸出，以及該第二輸出資料係為該壓縮後模型之一全連接層的輸出。

【請求項6】 如請求項1所述之模型壓縮方法，其中該第一輸出資料係為該原始模型之最後一層的Softmax函式輸出，以及該第二輸出資料係為該壓縮後模型

之最後一層的Softmax函式輸出。

【請求項7】 一種模型壓縮系統，包含：

一儲存裝置，用以儲存一程式碼；以及

一處理器，用以載入並執行該程式碼，以執行以下操作：

針對具有一深度神經網路架構之一原始模型來進行一模型剪枝操作，以產生一壓縮後模型；

將同一測試資料分別輸入至該原始模型以及該壓縮後模型；

計算該原始模型處理該測試資料所得到之一第一輸出資料以及該壓縮後模型處理該測試資料所得到之一第二輸出資料之間的相似度；以及

以該相似度作為獎勵，透過強化學習來判斷如何進一步調整該模型剪枝操作。

【請求項8】如請求項7所述之模型壓縮系統，其中該測試資料係為未標記資料。

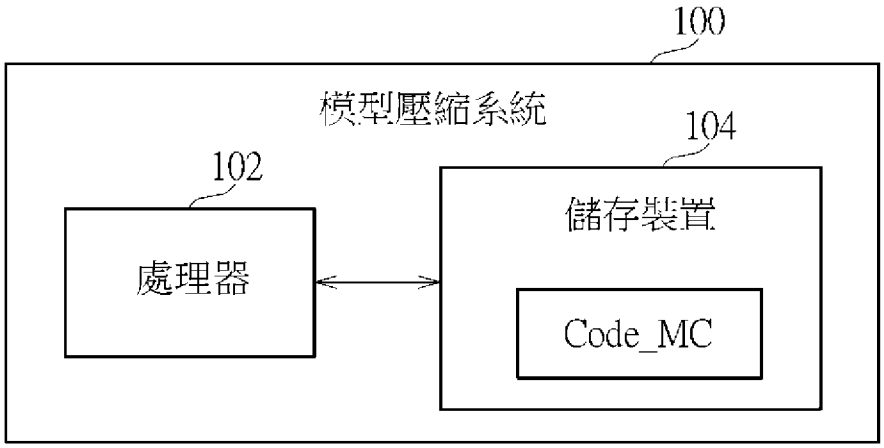
【請求項9】 如請求項7所述之模型壓縮系統，其中該相似度係藉由計算該第一輸出資料與該第二輸出資料之一皮爾森相關係數來得到。

【請求項10】 如請求項7所述之模型壓縮系統，其中該相似度係藉由計算該第一輸出資料與該第二輸出資料之一餘弦相似度來得到。

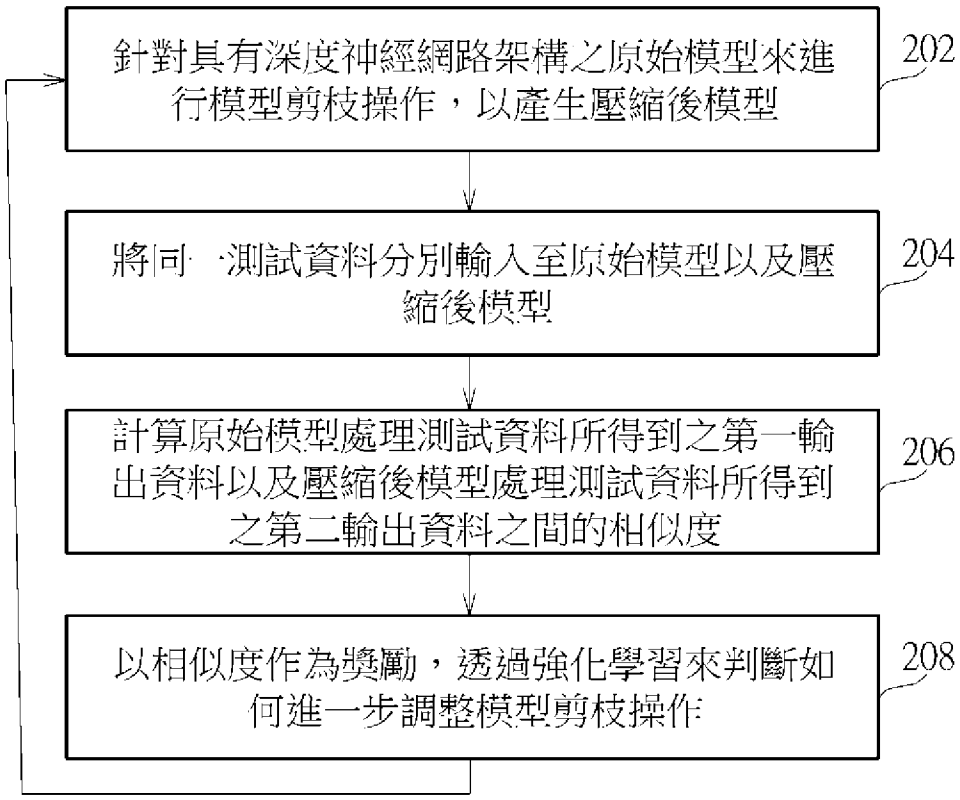
【請求項11】 如請求項7所述之模型壓縮系統，其中該第一輸出資料係為該原始模型之一全連接層的輸出，以及該第二輸出資料係為該壓縮後模型之一全連接層的輸出。

【請求項12】 如請求項7所述之模型壓縮系統，其中該第一輸出資料係為該原始模型之最後一層的Softmax函式輸出，以及該第二輸出資料係為該壓縮後模型之最後一層的Softmax函式輸出。

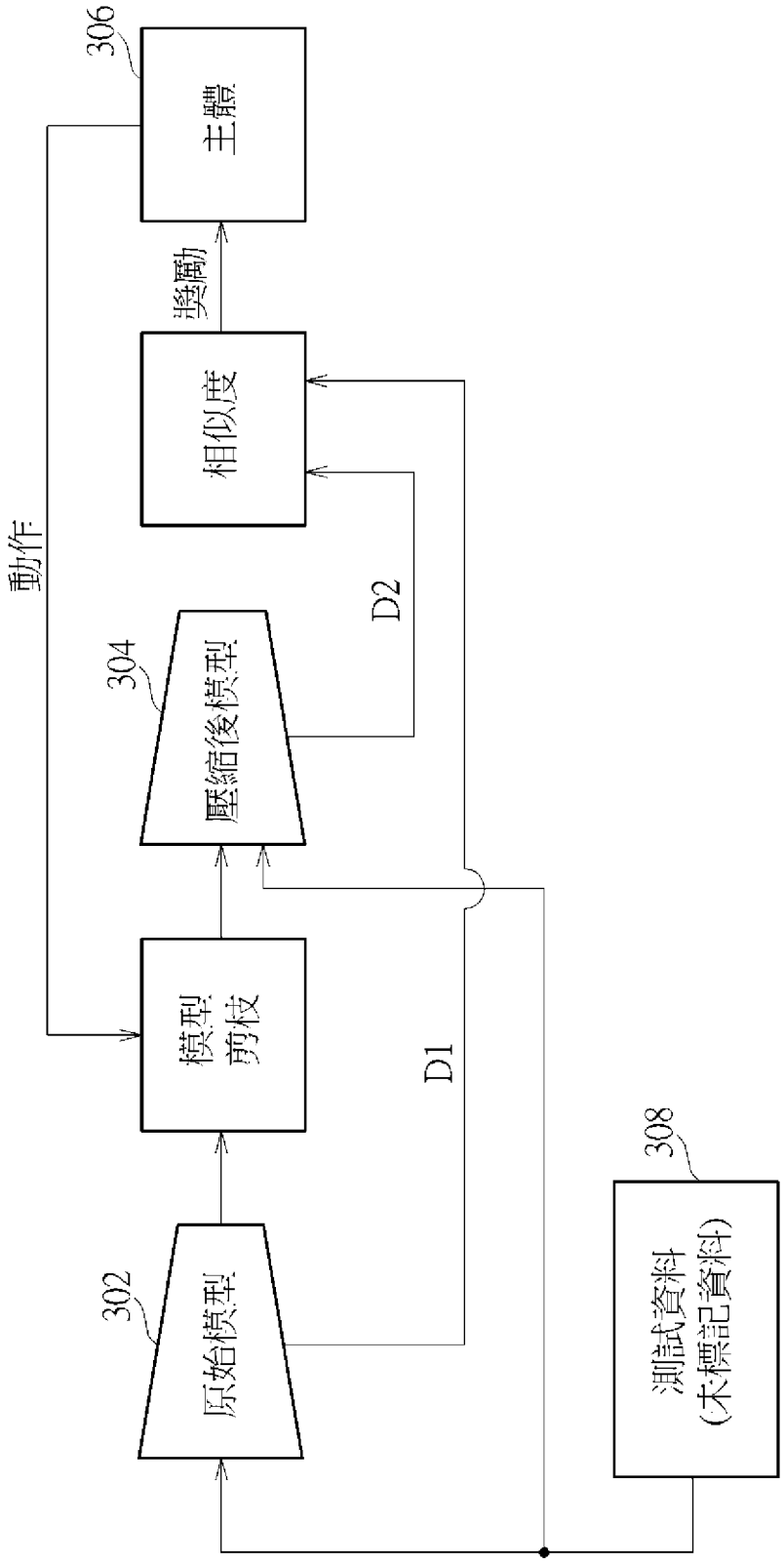
【發明圖式】


100
模型壓縮系統
104
102
儲存裝置
處理器
Code_MC

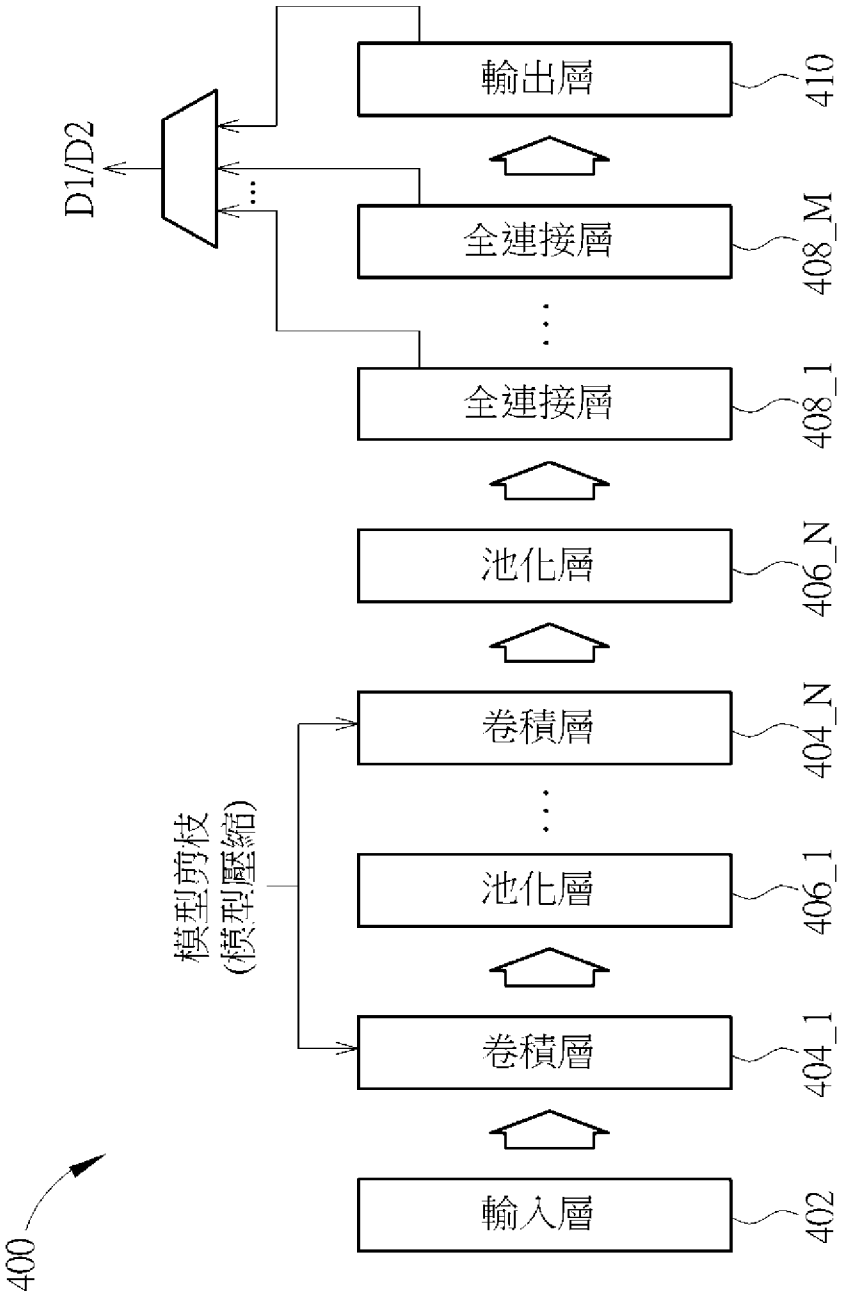
第1圖


針對具有深度神經網路架構之原始模型來進行模型剪枝操作，以產生壓縮後模型
202
將同一測試資料分別輸入至原始模型以及壓縮後模型
204
計算原始模型處理測試資料所得到之第一輸出資料以及壓縮後模型處理測試資料所得到之第二輸出資料之間的相似度
206
以相似度作為獎勵，透過強化學習來判斷如何進一步調整模型剪枝操作
208

第2圖



第3圖


400
輸入層
402
卷積層
404_1
池化層
406_1
卷積層
404_N
池化層
406_N
全連接層
408_1
全連接層
408_M
輸出層
410
D1/D2
模型剪枝
(模型壓縮)

第4圖