

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号

特許第7366274号

(P7366274)

(45)発行日 令和5年10月20日(2023.10.20)

(24)登録日 令和5年10月12日(2023.10.12)

(51)国際特許分類

G 0 6 N 3/082(2023.01)

F I

G 0 6 N 3/082

請求項の数 24 (全56頁)

(21)出願番号 特願2022-543064(P2022-543064)
(86)(22)出願日 令和3年1月15日(2021.1.15)
(65)公表番号 特表2023-510566(P2023-510566 A)
(43)公表日 令和5年3月14日(2023.3.14)
(86)国際出願番号 PCT/CN2021/072294
(87)国際公開番号 WO2021/143883
(87)国際公開日 令和3年7月22日(2021.7.22)
審査請求日 令和4年8月12日(2022.8.12)
(31)優先権主張番号 202010044288.8
(32)優先日 令和2年1月15日(2020.1.15)
(33)優先権主張国・地域又は機関
中国(CN)

(73)特許権者 503433420
華為技術有限公司
HUAWEI TECHNOLOGIES
CO., LTD.
中華人民共和國 5 1 8 1 2 9 広東省深
▲チェン▼市龍崗区坂田 華為總部▲ベ
ン▼公樓
Huawei Administrat
ion Building, Banti
an, Longgang Distri
ct, Shenzhen, Guang
dong 5 1 8 1 2 9, P. R. C
hina
(74)代理人 100107766
弁理士 伊東 忠重

最終頁に続く

(54)【発明の名称】 ニューラル・ネットワークのための適応的探索方法および装置

(57)【特許請求の範囲】

【請求項1】

コンピュータによって実行される、ニューラル・ネットワークのための適応的探索方法であって、当該方法は：

探索条件セットを受領するステップであって、前記探索条件セットはターゲット・ハードウェア・プラットフォーム情報、ソース・ニューラル・ネットワークのネットワーク構造情報、および一つまたは複数の評価メトリックを含み、前記ソース・ニューラル・ネットワークの前記ネットワーク構造情報は前記ソース・ニューラル・ネットワークを表すために使用される、ステップと；

ターゲット・ニューラル・ネットワークを出力するステップであって、前記ターゲット・ニューラル・ネットワークは、前記一つまたは複数の評価メトリックに基づいて、トレーニングされたスーパーネットワーク上でサブネット探索プロセスを実行することによって得られ、前記トレーニングされたスーパーネットワークは、トレーニングされるべきスーパーネットワーク上でトレーニング・プロセスを実行することによって得られ、前記トレーニングされるべきスーパーネットワークは、前記ソース・ニューラル・ネットワークのネットワーク構造を拡張することによって得られ、ターゲット・ハードウェア・プラットフォーム上で実行される前記ターゲット・ニューラル・ネットワークの評価結果は、前記ターゲット・ハードウェア・プラットフォーム上で実行される前記ソース・ニューラル・ネットワークの評価結果よりよい、ステップとを含む、方法。

10

20

【請求項 2】

前記一つまたは複数の評価メトリックは：

ニューラル・ネットワーク・モデルの精度評価メトリック、前記ニューラル・ネットワーク・モデルの時間オーバーヘッド評価メトリック、前記ニューラル・ネットワーク・モデルの記憶スペース評価メトリック、前記ニューラル・ネットワーク・モデルの電力消費評価メトリック、前記ニューラル・ネットワーク・モデルのテンソル計算ユニットの利用率評価メトリック、および前記ニューラル・ネットワーク・モデルのメモリ読み出し／書き込み速度評価メトリックのうちの一つまたは複数を含み、

前記ニューラル・ネットワーク・モデルの前記精度評価メトリックは、前記ターゲット・ニューラル・ネットワークの精度が精度閾値以上であることを含み；

前記ニューラル・ネットワーク・モデルの前記時間オーバーヘッド評価メトリックは、前記ターゲット・ニューラル・ネットワークの時間オーバーヘッドが時間オーバーヘッド閾値を超えないことを含み；

前記ニューラル・ネットワーク・モデルの前記記憶スペース評価メトリックは、前記ターゲット・ニューラル・ネットワークによって占有される記憶スペースが記憶スペース閾値を超えないことを含み；

前記ニューラル・ネットワーク・モデルの電力消費評価メトリックは、前記ターゲット・ニューラル・ネットワークの電力消費が電力消費閾値を超えないことを含み；

前記ニューラル・ネットワーク・モデルの前記テンソル計算ユニットの前記利用率評価メトリックは、前記ターゲット・ニューラル・ネットワークの前記テンソル計算ユニットの利用率が前記テンソル計算ユニットの利用率閾値以上であることを含み；

前記ニューラル・ネットワーク・モデルの前記メモリ読み出し／書き込み速度評価メトリックは、前記ターゲット・ニューラル・ネットワークのメモリ読み出し／書き込み速度がメモリ読み出し／書き込み速度閾値以上であることを含む、

請求項 1 に記載の方法。

【請求項 3】

前記ターゲット・ニューラル・ネットワークが、前記一つまたは複数の評価メトリックに基づいて、トレーニングされたスーパーネットワーク上でサブネット探索プロセスを実行することによって得られることは：

前記ターゲット・ニューラル・ネットワークが、前記一つまたは複数の評価メトリックに基づいて、ターゲット探索範囲において、一つまたは複数の評価されるべき候補サブネット上で前記サブネット探索プロセスを実行することによって得られることを含み、前記ターゲット探索範囲における前記一つまたは複数の評価されるべき候補サブネットは、第 1 の知識ベースを使用することによって、前記トレーニングされたスーパーネットワークから得られる、

請求項 1 または 2 に記載の方法。

【請求項 4】

前記ターゲット・ニューラル・ネットワークが、前記一つまたは複数の評価メトリックに基づいて、トレーニングされたスーパーネットワーク上でサブネット探索プロセスを実行することによって得られることは：

前記ターゲット・ニューラル・ネットワークが、一つまたは複数の評価されるべき候補サブネットの評価結果（単数または複数）に基づいて、前記トレーニングされたスーパーネットワーク上で前記サブネット探索プロセスを実行することによって得られることを含み、前記一つまたは複数の評価されるべき候補サブネットの評価結果（単数または複数）は、前記一つまたは複数の評価メトリックおよび対応情報に基づいて得られ、前記対応情報は、前記ターゲット・ハードウェア・プラットフォーム、M 個のニューラル・ネットワークのネットワーク構造、N 個の評価メトリック、および P 個の評価結果の間の対応を含み、M、N、P はすべて正の整数であり、前記 M 個のニューラル・ネットワークのネットワーク構造は、前記トレーニングされたスーパーネットワークのネットワーク構造に含まれる、

請求項 1 または 2 に記載の方法。

10

20

30

40

50

【請求項 5】

前記一つまたは複数の評価メトリックは、前記ニューラル・ネットワーク・モデルの前記時間オーバーヘッド評価メトリックを含み、前記ニューラル・ネットワーク・モデルの前記時間オーバーヘッド評価メトリックは、前記ターゲット・ニューラル・ネットワークの前記時間オーバーヘッドが前記時間オーバーヘッド閾値よりも大きくないことを含み；

前記ターゲット・ニューラル・ネットワークが、前記一つまたは複数の評価メトリックに基づいて、トレーニングされたスーパーネットワーク上でサブネット探索プロセスを実行することによって得られることは：

前記ターゲット・ニューラル・ネットワークが、前記ニューラル・ネットワーク・モデルの前記時間オーバーヘッド評価メトリックに基づいて、前記トレーニングされたスーパーネットワークに含まれる一つまたは複数の候補サブネット上でサブネット評価を実行することによって得られることを含む、

請求項 2 に記載の方法。

【請求項 6】

前記トレーニングされるべきスーパーネットワークが、前記ソース・ニューラル・ネットワークのネットワーク構造を拡張することによって得られることは：

前記トレーニングされるべきスーパーネットワークが、前記ソース・ニューラル・ネットワークのネットワーク構造、第 2 の知識ベースにおける拡張規則、および前記ターゲット・ハードウェア・プラットフォーム情報に基づいて、水平拡張モードおよび／または垂直拡張モードで得られることを含む、

請求項 1 ないし 5 のうちいずれか一項に記載の方法。

【請求項 7】

前記トレーニングされるべきスーパーネットワークにおける第 3 演算子セットが、第 1 演算子セットおよび／または派生第 2 演算子セットを含み、前記第 1 演算子セットは、前記ソース・ニューラル・ネットワークに含まれる、請求項 1 ないし 6 のうちいずれか一項に記載の方法。

【請求項 8】

前記一つまたは複数の評価されるべき候補サブネットの評価結果（単数または複数）が、前記一つまたは複数の評価メトリックおよび対応情報に基づいて得られることは：

前記一つまたは複数の評価されるべき候補サブネットの評価結果（単数または複数）が、前記 N 個の評価メトリックの間のプロポーション関係に基づいて、かつ前記トレーニングされたスーパーネットワークにおける第 3 演算子セット内の第 1 演算子セットおよび／または派生第 2 演算子セットに対応する評価結果に基づいて、得られ、前記トレーニングされたスーパーネットワークにおける前記第 3 演算子セットにおける前記第 1 演算子セットおよび／または前記派生第 2 演算子セットに対応する前記評価結果は、前記一つまたは複数の評価メトリックおよび前記対応情報に基づいて得られ、前記第 1 演算子セットは前記ソース・ニューラル・ネットワークに含まれる、

請求項 4 に記載の方法。

【請求項 9】

前記一つまたは複数の評価されるべき候補サブネットの評価結果（単数または複数）が、前記 N 個の評価メトリックの間のプロポーション関係に基づいて、かつ前記トレーニングされたスーパーセットにおける第 3 演算子セット内の第 1 演算子セットおよび／または派生第 2 演算子セットに対応する評価結果に基づいて、得られることは：

前記一つまたは複数の評価されるべき候補サブネットの評価結果（単数または複数）が、r 個の事前探索プロセスに基づいて決定され、前記 r 個の事前探索プロセスにおける各事前探索プロセスは：前記第 1 演算子セットおよび／または前記派生第 2 演算子セットの評価結果の重みの指定された値と、前記 N 個の評価メトリックと、該 N 個の評価メトリック間のプロポーション関係に基づいて、トレーニングされたスーパーネットワークを再トレーニングするステップであって、前記重みは、前記第 1 演算子セットおよび／または前記派生第 2 演算子セットの評価結果に対応する、ステップと；評価されるべき候補サブネットお

10

20

30

40

50

よび該評価されるべき候補サブネットの評価結果を、前記第1演算子セットおよび前記派生第2演算子セットの評価結果に対応する調整された重みに基づいて決定するステップとを実行するために使用され、 r は正の整数である、

請求項 8 に記載の方法。

【請求項 1 0】

前記 N 個の評価メトリックは、前記ニューラル・ネットワーク・モデルの前記精度評価メトリックおよび前記ニューラル・ネットワーク・モデルの前記時間オーバーヘッド評価メトリックを含み；対応して、前記 N 個の評価メトリックの間のプロポーショナル関係は、前記精度評価メトリックと前記時間オーバーヘッド評価メトリックとの間のプロポーショナル関係を含む、請求項 4 が請求項 2 を引用する場合の、請求項 8 または 9 に記載の方法。

10

【請求項 1 1】

前記ターゲット・ニューラル・ネットワークが、一つまたは複数の評価されるべき候補サブネットの評価結果（単数または複数）に基づいて、前記トレーニングされたスーパーネットワーク上で前記サブネット探索プロセスを実行することによって得られることは：

$K \times (n+1)$ 個のターゲット候補サブネットから前記ターゲット・ニューラル・ネットワークを得るステップであって、前記 $K \times (n+1)$ 個のターゲット候補サブネットは、前記一つまたは複数の評価されるべき候補サブネットの評価結果をソートして、最初の K 個の評価結果に対応する K 個の第1のターゲット候補サブネットを得ることによって得られる、ステップと；前記トレーニングされたスーパーネットワークおよび前記 K 個の第1のターゲット候補サブネットに基づく第 n ラウンドの混成および／または変成の後に得られる S 個の第2のターゲット候補サブネットを得るステップであって、 n はあらかじめ設定されたスクリーニング回数であり、 $n \geq 1$ であり、 K および S は正の整数であり、 $S > K$ である、ステップと；前記一つまたは複数の評価メトリックおよび前記対応情報に基づいて前記 S 個の第2のターゲット候補サブネットの評価結果を得て、それらの評価結果をソートして、最初の K 個の評価結果に対応する K 個の第2のターゲット候補サブネットを得るステップとを実行することを含む、

20

請求項 4 に記載の方法。

【請求項 1 2】

前記ソース・ニューラル・ネットワークはソース・ハードウェア・プラットフォーム上で動作するニューラル・ネットワークであり、または前記ターゲット・ニューラル・ネットワークは前記ターゲット・ハードウェア・プラットフォーム上で動作するニューラル・ネットワークであり；前記ソース・ハードウェア・プラットフォームは前記ターゲット・ハードウェア・プラットフォームとは異なる、請求項 1 ないし 1 0 のうちいずれか一項に記載の方法。

30

【請求項 1 3】

前記ソース・ニューラル・ネットワークおよび前記ターゲット・ニューラル・ネットワークは、異なる時間領域において前記ターゲット・ハードウェア・プラットフォーム上で稼働するニューラル・ネットワークである、請求項 1 ないし 1 0 のうちいずれか一項に記載の方法。

【請求項 1 4】

当該方法はさらに：

前記ターゲット・ハードウェア・プラットフォームを有するデバイスにトレーニングされたターゲット・ニューラル・ネットワークを送信するステップを含み、前記トレーニングされたターゲット・ニューラル・ネットワークは、前記ターゲット・ニューラル・ネットワーク上でトレーニング・プロセスを実行することによって得られる、

請求項 1 ないし 1 3 のうちいずれか一項に記載の方法。

40

【請求項 1 5】

前記トレーニングされたターゲット・ニューラル・ネットワークは、ターゲット検出タスクおよび／または画像分類タスクを実行するために前記ターゲット・ハードウェア・プラットフォーム上で実行される、請求項 1 2 に記載の方法。

50

【請求項 16】

ニューラル・ネットワークのための適応的探索装置であって、当該装置は受信ユニットと送信ユニットとを有しており、

前記受信ユニットは、探索条件セットを受領するように構成されており、前記探索条件セットはターゲット・ハードウェア・プラットフォーム情報、ソース・ニューラル・ネットワークのネットワーク構造情報、および一つまたは複数の評価メトリックを含み、前記ソース・ニューラル・ネットワークの前記ネットワーク構造情報は前記ソース・ニューラル・ネットワークを表すために使用され；

前記送信ユニットは、ターゲット・ニューラル・ネットワークを出力するように構成されており、前記ターゲット・ニューラル・ネットワークは、前記一つまたは複数の評価メトリックに基づいて、トレーニングされたスーパーネットワーク上でサブネット探索プロセスを実行することによって得られ、前記トレーニングされたスーパーネットワークは、トレーニングされるべきスーパーネットワーク上でトレーニング・プロセスを実行することによって得られ、前記トレーニングされるべきスーパーネットワークは、前記ソース・ニューラル・ネットワークのネットワーク構造を拡張することによって得られ、ターゲット・ハードウェア・プラットフォーム上で実行される前記ターゲット・ニューラル・ネットワークの評価結果は、前記ターゲット・ハードウェア・プラットフォーム上で実行される前記ソース・ニューラル・ネットワークの評価結果よりよい、装置。

【請求項 17】

一つまたは複数のプロセッサ、
メモリ、および

一つまたは複数のコンピュータ・プログラムを有するサーバーであって、

前記一つまたは複数のコンピュータ・プログラムは前記メモリに記憶されており、前記一つまたは複数のコンピュータ・プログラムは命令を含み、前記命令が前記サーバーによって実行されるとき、前記サーバーは、請求項 1 ないし 15 のうちいずれか一項に記載の、ニューラル・ネットワークのための適応的探索方法を実行できるようにされる、サーバー。

【請求項 18】

コンピュータ命令を含むコンピュータ読み取り可能な記憶媒体であって、前記コンピュータ命令がサーバー上で実行されるとき、前記サーバーは、請求項 1 ないし 15 のうちいずれか一項に記載の、ニューラル・ネットワークのための適応的探索方法を実行できるようにされる、コンピュータ読み取り可能な記憶媒体。

【請求項 19】

コンピュータに請求項 1 ないし 15 のうちいずれか一項に記載の、ニューラル・ネットワークのための適応的探索方法を実行させるためのコンピュータ・プログラム。

【請求項 20】

少なくとも1つのサーバーおよび少なくとも1つの端末装置を含む通信システムであって、前記少なくとも1つのサーバーは、請求項 1 ないし 15 のうちいずれか一項に記載の、ニューラル・ネットワークのための適応的探索方法を実行し；前記少なくとも1つの端末装置は、前記少なくとも1つのサーバーに探索条件セットを送信し、前記探索条件セットは、ターゲット・ハードウェア・プラットフォーム情報、ソース・ニューラル・ネットワークのネットワーク構造情報、および一つまたは複数の評価メトリックを含み；前記ソース・ニューラル・ネットワークの前記ネットワーク構造情報は、前記ソース・ニューラル・ネットワークを表すために使用され；前記少なくとも1つの端末装置は、前記少なくとも1つのサーバーによって送信されるトレーニングされたターゲット・ニューラル・ネットワークを受領する、通信システム。

【請求項 21】

コンピュータによって実行される、ニューラル・ネットワークのための適応的探索方法であって、当該方法は：

少なくとも1つのサーバーに探索条件セットを送信するステップであって、前記探索条件セットは、ターゲット・ハードウェア・プラットフォーム情報、ソース・ニューラル・ネットワークのネットワーク構造情報、および一つまたは複数の評価メトリックを含み、前記ソース・ニューラル・ネットワークの前記ネットワーク構造情報は、前記ソース・ニューラル・ネットワークを表すために使用される、ステップと；

前記少なくとも1つのサーバーによって送信されたトレーニングされたターゲット・ニューラル・ネットワークを受信するステップであって、前記トレーニングされたターゲット・ニューラル・ネットワークは、前記ターゲット・ニューラル・ネットワーク上のトレーニング・プロセスを実行することによって、前記少なくとも1つのサーバーによって取得され、前記ターゲット・ニューラル・ネットワークは、前記一つまたは複数の評価メトリックに基づいてトレーニングされたスーパーネットワーク上でサブネット探索プロセスを実行することによって、前記少なくとも1つのサーバーによって取得され、前記トレーニングされたスーパーネットワークは、トレーニングされるべきスーパーネットワーク上でトレーニング・プロセスを実行することによって、前記少なくとも1つのサーバーによって取得され、前記トレーニングされるべきスーパーネットワークは、前記ソース・ニューラル・ネットワークのネットワーク構造を拡張することによって、前記少なくとも1つのサーバーによって取得され、ターゲット・ハードウェア・プラットフォーム上で動作する前記ターゲット・ニューラル・ネットワークの評価結果は、前記ターゲット・ハードウェア・プラットフォーム上で動作する前記ソース・ニューラル・ネットワークの評価結果よりもよい、ステップとを含む、

方法。

【請求項 2 2】

当該方法は：

前記トレーニングされたターゲット・ニューラル・ネットワークを前記ターゲット・ハードウェア・プラットフォームを有するデバイスに送信するかどうかを決定するステップを含む、

請求項 2 1 に記載の方法。

【請求項 2 3】

ニューラル・ネットワークのための適応的探索装置であって、当該装置は送信ユニットおよび受信ユニットを有しており、

前記送信ユニットは、少なくとも1つのサーバーに探索条件セットを送信するように構成されており、前記探索条件セットは、ターゲット・ハードウェア・プラットフォーム情報、ソース・ニューラル・ネットワークのネットワーク構造情報、および一つまたは複数の評価メトリックを含み、前記ソース・ニューラル・ネットワークの前記ネットワーク構造情報は、前記ソース・ニューラル・ネットワークを表すために使用され、

前記受信ユニットは、前記少なくとも1つのサーバーによって送信されたトレーニングされたターゲット・ニューラル・ネットワークを受信するように構成されており、前記トレーニングされたターゲット・ニューラル・ネットワークは、前記ターゲット・ニューラル・ネットワーク上のトレーニング・プロセスを実行することによって、前記少なくとも1つのサーバーによって取得され、前記ターゲット・ニューラル・ネットワークは、前記一つまたは複数の評価メトリックに基づいてトレーニングされたスーパーネットワーク上でサブネット探索プロセスを実行することによって、前記少なくとも1つのサーバーによって取得され、前記トレーニングされたスーパーネットワークは、トレーニングされるべきスーパーネットワーク上でトレーニング・プロセスを実行することによって、前記少なくとも1つのサーバーによって取得され、前記トレーニングされるべきスーパーネットワークは、前記ソース・ニューラル・ネットワークのネットワーク構造を拡張することによって、前記少なくとも1つのサーバーによって取得され、ターゲット・ハードウェア・プラットフォーム上で動作する前記ターゲット・ニューラル・ネットワークの評価結果は、前記ターゲット・ハードウェア・プラットフォーム上で動作する前記ソース・ニューラル・ネットワークの評価結果よりもよい、

装置。

【請求項 2 4】

当該装置はさらに処理ユニットを有しており、

前記処理ユニットは、前記トレーニングされたターゲット・ニューラル・ネットワークを前記ターゲット・ハードウェア・プラットフォームを有するデバイスに送信するかどうかを決定するように構成されている、

請求項 2 3 に記載の装置。

【発明の詳細な説明】

【技術分野】

【0 0 0 1】

本願は、2020年1月15日に中国国家知識産権局に出願され、「ニューラル・ネットワークのための適応的探索方法および装置」と題された中国特許出願第202010044288.8号に対する優先権を主張するものであり、その全体が参照により本明細書に組み込まれる。

【0 0 0 2】

技術分野

本願は、ニューラル・ネットワークの分野に関し、特に、ニューラル・ネットワークのための適応的探索方法および装置に関する。

【背景技術】

【0 0 0 3】

近年、コンピュータビジョン・アプリケーションにおいてニューラル・ネットワークが急速に発展した。ニューラル・ネットワークの正解率が次第に増加するにつれて、ニューラル・ネットワークのネットワーク構造はますます複雑になり、デバイスの計算資源に対する要求はますます高くなる。よって、デバイスの実行速度は低く、時間オーバーヘッドは高い。

【0 0 0 4】

現在、従来技術のニューラル・ネットワーク探索方法では、一般に、浮動小数点演算毎秒 (floating-point operations per second、FLOPS) に基づいてニューラル・ネットワーク・モデルが探索される。FLOPSは通例、ハードウェア・プラットフォームの実行効率 (モデル・オーバーヘッドとも呼ばれる) を推定するために使用される。FLOPSに基づいてニューラル・ネットワーク・モデルが探索されるとき、異なるハードウェア・プラットフォームについて同じターゲット・ニューラル・ネットワークが見つかる可能性がある。異なるハードウェア・プラットフォーム上の同じターゲット・ニューラル・ネットワークは、同じFLOPS (たとえば、異なるハードウェア・プラットフォーム上のターゲット・ニューラル・ネットワークの加算および乗算の計算量) をもつ。しかしながら、これは、より低いFLOPSがニューラル・ネットワーク・モデルのより短い推論時間を示すことを意味しない。たとえば、ハードウェア・プラットフォーム1 (CPU 1)、ハードウェア・プラットフォーム2 (CPU 2)、ハードウェア・プラットフォーム3 (GPU) の3つの既存のハードウェア・プラットフォームがある。従来の技術では、FLOPSに基づいてニューラル・ネットワークが探索されると、同じターゲット・ニューラル・ネットワークが見出されることがある。これら3つのハードウェア・プラットフォーム上のターゲット・ニューラル・ネットワークのFLOPSは低い。しかしながら、実際の実行プロセスでは、これら3つのハードウェア・プラットフォームの構造が異なるため、ターゲット・ニューラル・ネットワークは、ハードウェア・プラットフォーム1に対してはより良好な実行効果を、ハードウェア・プラットフォーム2に対しては平均的な実行効果を、ハードウェア・プラットフォーム3に対してはより貧弱な実行効果を有することがある。

【0 0 0 5】

あるいはまた、ニューラル・ネットワーク・モデルがFLOPSに基づいて探索される場合、ハードウェア・プラットフォームについてターゲット・ニューラル・ネットワークが見出されてもよい。ハードウェア・プラットフォーム上のターゲット・ニューラル・ネットワークのFLOPSは低い (たとえば、ターゲット・ニューラル・ネットワークの加算および

10

20

30

40

50

乗算の計算量が小さい)。しかしながら、実際には、ハードウェア・プラットフォーム上のターゲット・ニューラル・ネットワークの実行効率は低い(すなわち、モデルオーバーヘッドが高い)。

【0006】

特定のハードウェア・プラットフォームについて、該ハードウェア・プラットフォーム上で効率的に動作するターゲット・ニューラル・ネットワークをどのように見つけるかが、解決されるべき緊急の問題となっていることがわかる。

【発明の概要】

【発明が解決しようとする課題】

【0007】

本願で提供されるニューラル・ネットワークについての適応的探索方法によれば、ターゲット・ハードウェア・プラットフォームに適応するターゲット・ニューラル・ネットワークが高い探索柔軟性を前提として見出されることができる。

【課題を解決するための手段】

【0008】

上記の目的を達成するために、本願では以下の技術的解決策が使用される。

【0009】

第1の側面によれば、本願は、ニューラル・ネットワークのための適応的探索方法を提供する。本方法は：探索条件セットを受領するステップであって、探索条件セットはターゲット・ハードウェア・プラットフォーム情報、ソース・ニューラル・ネットワークのネットワーク構造情報、および一つまたは複数の評価メトリックを含み、ソース・ニューラル・ネットワークのネットワーク構造情報がソース・ニューラル・ネットワークを表すために使用される、ステップと；トレーニング・データセットに基づいてトレーニングされるべきスーパーネットワーク上でトレーニング・プロセスを実行して、トレーニングされたスーパーネットワークを得るステップであって、トレーニングされるべきスーパーネットワークは、ソース・ニューラル・ネットワークのネットワーク構造を拡張することによって得られる、ステップと；前記一つまたは複数の評価メトリックに基づいて、トレーニングされたスーパーネットワーク上でサブネット探索プロセスを実行して、ターゲット・ニューラル・ネットワークのネットワーク構造情報を得るステップであって、ターゲット・ニューラル・ネットワークのネットワーク構造情報が、ターゲット・ニューラル・ネットワークを表すために使用され、ターゲット・ハードウェア・プラットフォーム上で実行されるターゲット・ニューラル・ネットワークの評価結果は、ターゲット・ハードウェア・プラットフォーム上で実行されるソース・ニューラル・ネットワークの評価結果よりよい、ステップとを含んでいてもよい。

【0010】

トレーニングされたスーパーネットワークにおいて見出される一つまたは複数のターゲット・ニューラル・ネットワークがあってもよい。たとえば、最適評価結果を有するニューラル・ネットワークが、前記ターゲット・ハードウェア・プラットフォーム情報および前記一つまたは複数の評価メトリックに基づく前記ターゲット・ニューラル・ネットワークとして使用される。あるいはまた、ソートされた評価結果に基づいて最初にソートされた複数のよりよいニューラル・ネットワークが、ターゲット・ニューラル・ネットワークとして使用される。

【0011】

ターゲット・ハードウェア・プラットフォームは、ニューラル・ネットワークを担持するように構成されたプラットフォームを含む。得られたターゲット・ハードウェア・プラットフォーム情報は、たとえば、属性情報(たとえば、グラフィックス処理ユニット(graphics processing unit、GPU)、ホスト中央処理ユニット(central processing unit、CPU)、およびデジタル信号プロセッサ(digital signal processor、DSP))、モデル、およびハードウェア・プラットフォームの製造者を含んでいてもよい。このように、異なるハードウェア・プラットフォームは、ニューラル・ネットワークの適応的探

10

20

30

40

50

索を実施するために、異なる演算子および／または演算子セットについて異なる実装および最適化状態を有する。

【0012】

評価メトリックは、探索条件、制約条件などと呼ばれることもある。評価メトリックは、最終的に発見されたニューラル・ネットワークが要件を満たすことを確実にするために使用される。たとえば、発見されたニューラル・ネットワークの精度および時間オーバーヘッドは、要件を満たす必要がある。

【0013】

ニューラル・ネットワークのネットワーク構造情報は、たとえば、ニューラル・ネットワークに含まれる演算子および／または演算子セットに関する情報、前記演算子および／または前記演算子セットの配置態様（たとえば、ツリー構造を有するニューラル・ネットワークまたはメッシュ構造を有するニューラル・ネットワーク）、レベル情報、および前記演算子および／または前記演算子セットのパラメータを含んでいてもよい。

【0014】

このように、既存のFLOPSベースのニューラル・ネットワーク探索方法と比較して、本願で提供されるニューラル・ネットワークの適応的探索方法では、ニューラル・ネットワークは、探索条件セットに基づいて、トレーニングされたスーパーネットワーク内で探索される。スーパーネットワークはソース・ニューラル・ネットワークのネットワーク構造を拡張することによって得られる。よって、スーパーネットワークは、演算子／チャネルについての異なるハードウェア・プラットフォームの異なる実装をカバーまたはサポートすることができる。探索条件セットは、柔軟に設定されてもよく、探索条件セットは、少なくともターゲット・ハードウェア・プラットフォーム情報を含む。よって、本願において提供される方法によれば、ターゲット・ハードウェア・プラットフォームについてのターゲット・ニューラル・ネットワークは、演算子／チャネルについての異なるハードウェア・プラットフォームの異なる実装によって引き起こされる異なる演算子／チャネル・オーバーヘッドを考慮することによって、高い探索柔軟性を前提とした適応的探索によって得ることができる。ターゲット・ハードウェア・プラットフォーム上のターゲット・ニューラル・ネットワークの実行効果は、ターゲット・ハードウェア・プラットフォーム上のソース・ニューラル・ネットワークの実行効果よりよい。たとえば、ソース・ニューラル・ネットワークにおける演算子Aは、 1×1 畳み込みアルゴリズムが1000万回の乗算を必要とするというものであり、演算子Bは、 3×3 畳み込みアルゴリズムが1500万回の演算を必要とするというものであると想定される。ソース・ニューラル・ネットワークがターゲット・ハードウェア・プラットフォーム上で動作する場合、演算子Aは $2 \mu s$ 、演算子Bは $3 \mu s$ を必要とする。ターゲット・ニューラル・ネットワーク内の演算子Aがターゲット・ハードウェア・プラットフォーム上で動作する場合、演算子Aは $2 \mu s$ を必要とし、演算子B'（たとえば、演算子Bが最適化された後に得られる演算子）は $1.5 \mu s$ しか必要としない。

【0015】

ある可能な実装において、前記一つまたは複数の評価メトリックは、ニューラル・ネットワーク・モデルの精度評価メトリック、ニューラル・ネットワーク・モデルの時間オーバーヘッド評価メトリック、ニューラル・ネットワーク・モデルの記憶スペース評価メトリック、ニューラル・ネットワーク・モデルの電力消費評価メトリック、ニューラル・ネットワーク・モデルのテンソル計算ユニットの利用率評価メトリック、およびニューラル・ネットワーク・モデルのメモリ読み出し／書き込み速度評価メトリックのうちの一つまたは複数を含む。ニューラル・ネットワーク・モデルの精度評価メトリックは、ターゲット・ニューラル・ネットワークの精度が精度閾値以上であることを含む。ニューラル・ネットワーク・モデルの時間オーバーヘッド評価メトリックは、ターゲット・ニューラル・ネットワークの時間オーバーヘッドが時間オーバーヘッド閾値を超えないことを含む。ニューラル・ネットワーク・モデルの記憶スペース評価メトリックは、ターゲット・ニューラル・ネットワークによって占有される記憶スペースが記憶スペース閾値を超えないことを含む。ニューラル・ネットワーク・モデルの電力消費評価メトリックは、ターゲット・

10

20

30

40

50

ニューラル・ネットワークの電力消費が電力消費閾値を超えないことを含む。ニューラル・ネットワーク・モデルのテンソル計算ユニットの利用率評価メトリックは、ターゲット・ニューラル・ネットワークのテンソル計算ユニットの利用率がテンソル計算ユニットの利用率閾値以上であることを含む。ニューラル・ネットワーク・モデルのメモリ読み出し／書き込み速度評価メトリックは、ターゲット・ニューラル・ネットワークのメモリ読み出し／書き込み速度がメモリ読み出し／書き込み速度閾値以上であることを含む。

【0016】

このようにして、前記一つまたは複数の評価メトリックの組み合わせモードが、異なるユーザー要件、異なる適用シナリオ、または異なるタスクなどの要因に基づいて設定される。たとえば、評価メトリックの組み合わせモードは、ハードウェア・プラットフォームのハードウェアメトリックまたはニューラル・ネットワークの異なるタスクに基づいて決定される。たとえば、精度評価メトリックと時間オーバーヘッド評価メトリックの2つの評価メトリックが構成されてもよい。この場合、精度閾値と時間オーバーヘッド閾値の両方が設定される必要がある。たとえば、シナリオによっては、大量の画像の認識結果を短時間で得る必要があるが、画像認識のための精度の要求は高くない。この場合、見出されるターゲット・ニューラル・ネットワークの認識速度がソース・ニューラル・ネットワークの認識速度よりも速くなるように、より小さな精度閾値およびより小さな時間のオーバーヘッド閾値が設定されてもよい。

【0017】

ある可能な実装では、前記一つまたは複数の評価メトリックに基づいて、トレーニングされたスーパーネットワーク上でサブネット探索プロセスを実行して、ターゲット・ニューラル・ネットワークのネットワーク構造情報を取得することは：第1の知識ベースを使用することによって、トレーニングされたスーパーネットワークからターゲット探索範囲内の一つまたは複数の評価されるべき候補サブネットを決定するステップと；前記一つまたは複数の評価メトリックに基づいて、ターゲット探索範囲内の前記一つまたは複数の評価されるべき候補サブネット上でサブネット探索プロセスを実行して、ターゲット・ニューラル・ネットワークのネットワーク構造情報を得るステップとを含む。

【0018】

第1の知識ベースは、たとえば、専門家の事前知識を含んでいてもよい。第1の知識ベースは、専門家の経験、特定のハードウェア・プラットフォームのための演算子および／または演算子セットの最適化状態などに基づいて決定される、ニューラル・ネットワークの、ネットワーク受容野制約条件を含む。たとえば、最適化状態は：特定のハードウェア・プラットフォームに入力される（ 7×7 畳み込み演算子+プーリング）が最適化され、その結果、（ 7×7 畳み込み演算子+プーリング）の計算量が（ 3×3 畳み込み演算子+プーリング）または（ 5×5 畳み込み演算子+プーリング）の計算量を超えるものの、現在のハードウェア・プラットフォーム上の（ 7×7 畳み込み演算子+プーリング）の時間オーバーヘッドが（ 3×3 畳み込み演算子+プーリング）または（ 5×5 畳み込み演算子+プーリング）の時間オーバーヘッドを下回るようにすることを含んでいてもよい。このようにして、第1の知識ベースに基づいて、より低い時間オーバーヘッドを含むニューラル・ネットワーク（ 7×7 畳み込み演算子+プーリング）が候補サブネットとして選択されることが直接的に決定されてもよい。これは、ターゲット探索範囲を狭め、ニューラル・ネットワーク探索効率を増加させる。

【0019】

このようにして、第1の知識ベースを用いて、ターゲット・ハードウェア・プラットフォーム情報に基づいて、スーパーネットワーク内でターゲット探索範囲が決定されてもよく、サブネット探索プロセスがターゲット探索範囲内で実行される。これは、探索サブネットの探索範囲を狭め、探索時間を短縮し、ニューラル・ネットワーク探索効率を増加させるのに役立つ。

【0020】

ある可能な実装では、前記一つまたは複数の評価メトリックに基づいて、トレーニング

10

20

30

40

50

されたスーパーネットワーク上でサブネット探索プロセスを実行して、ターゲット・ニューラル・ネットワークのネットワーク構造情報を得ることは：ターゲット・ハードウェア・プラットフォーム、M個のニューラル・ネットワークのネットワーク構造、N個の評価メトリック、およびP個の評価結果の間の対応についての情報を得るステップであって、ここで、M、N、Pはすべて正の整数であり、M個のニューラル・ネットワークのネットワーク構造は、トレーニングされたスーパーネットワークのネットワーク構造に含まれる、ステップと；前記一つまたは複数の評価メトリックおよび前記対応情報に基づいて、トレーニングされたスーパーネットワークに含まれる一つまたは複数の評価されるべき候補サブネットの評価結果（単数または複数）を決定するステップと；前記一つまたは複数の評価されるべき候補サブネットの前記評価結果（単数または複数）に基づいて、トレーニングされたスーパーネットワーク上で探索プロセスを実行して、ターゲット・ニューラル・ネットワークのネットワーク構造情報を得るステップとを含む。

10

【0021】

前記M個のニューラル・ネットワークは、トレーニングされたスーパーネットワークから抽出されたサブネット、すなわち評価されるべき候補サブネットである。M個のニューラル・ネットワークのネットワーク構造がトレーニングされたスーパーネットワークのネットワーク構造に含まれることは明らかである。1つの候補サブネットがターゲット・ハードウェア・プラットフォーム上で実行される場合、候補サブネットは、N個の評価メトリックに基づいたN個の評価結果に対応する。M個の候補サブネットがターゲット・ハードウェア・プラットフォーム上で実行されるとき、M個の候補サブネットは、N個の評価メトリックに基づいて、M×Nの評価結果、すなわち、P個の評価結果に対応する。次いで、前記一つまたは複数の候補サブネットが、P個の評価結果に基づいて、ターゲット・ニューラル・ネットワークとして決定される。

20

【0022】

このようにして、探索プロセスにおいて、前記一つまたは複数の評価メトリックのうちの任意の一つまたは複数の選択される場合、候補サブネットの評価結果が、対応に基づいて決定されてもよく、次いで、ターゲット・ニューラル・ネットワークのネットワーク構造情報が決定されてもよい。

【0023】

たとえば、ターゲット・ハードウェア・プラットフォーム1では、評価メトリックは精度であり、候補サブネット1の評価結果はAであり、候補サブネット2の評価結果はBであり、ここで、AのほうがBよりよい。1つのターゲット・ニューラル・ネットワークがある場合、一つまたは複数の評価メトリックおよび対応情報に基づいて、候補サブネット1がターゲット・ニューラル・ネットワークとして決定される。

30

【0024】

ある可能な実装では、前記一つまたは複数の評価メトリックは、ニューラル・ネットワーク・モデルの時間オーバーヘッド評価メトリックを含み、ニューラル・ネットワーク・モデルの時間オーバーヘッド評価メトリックは、ターゲット・ニューラル・ネットワークの時間オーバーヘッドが時間オーバーヘッド閾値よりも大きくないことを含み；対応して、前記一つまたは複数の評価メトリックに基づいて、トレーニングされたスーパーネットワーク上でサブネット探索プロセスを実行するステップは：ニューラル・ネットワーク・モデルの時間オーバーヘッド評価メトリックに基づいて、トレーニングされたスーパーネットワークに含まれる一つまたは複数の候補サブネット上のサブネット評価を実行して、ニューラル・ネットワーク・モデルの時間オーバーヘッドが時間オーバーヘッド閾値よりも小さい一つまたは複数のターゲット・ニューラル・ネットワークを得ることを含む。

40

【0025】

たとえば、ニューラル・ネットワークはハードウェア・プラットフォーム上で動作し、ハードウェア・プラットフォームは画像を処理するように構成され、もとの処理速度は毎秒100画像を処理する。現在、ハードウェア・プラットフォームの画像処理速度を上げる必要がある。たとえば、ハードウェア・プラットフォーム上の最適化されたニューラル・

50

ネットワークの画像処理速度が毎秒150画像に達することを可能にするために、時間オーバーヘッド評価メトリックが設定されてもよく、時間オーバーヘッド閾値が使用されて、発見されたターゲット・ニューラル・ネットワークの画像処理速度が要件を満たすことができることを保証する。

【0026】

ある可能な実装では、本方法は：ソース・ニューラル・ネットワークのネットワーク構造、第2の知識ベースにおける拡張規則、およびターゲット・ハードウェア・プラットフォーム情報に基づいて、トレーニングされるべきスーパーネットワークを水平拡張モードおよび／または垂直拡張モードで構築することをさらに含む。

【0027】

第2の知識ベースは、第1の知識ベースと同じであっても、または異なってもよい。第2の知識ベースにおける拡張規則は、ソース・ニューラル・ネットワークにおける演算子セット（すなわち、下記の第1演算子セット）およびそれらの演算子セットに対応する交換可能演算子セット（すなわち、下記の派生第2演算子セット）が、異なるハードウェア・プラットフォームに基づいて異なる対応を有することを含む。ソース・ニューラル・ネットワーク内に設定された演算子および対応する交換可能演算子セットは、同一または類似の機能を有することが理解されうる。この対応は、ターゲット・ハードウェア・プラットフォームに基づくソース・ニューラル・ネットワークにおける演算子セットと、交換可能な演算子セットとの間の対応を含む。さらに、該対応は、実験データ、専門家の事前知識、およびターゲット・ハードウェア・プラットフォームの設計状態に基づいて確立される。このようにして、拡張規則は、スーパーネットワークのネットワーク・スケールを制限するために使用されうる。たとえば、ソース・ニューラル・ネットワークにおける演算子セットがターゲット・ハードウェア・プラットフォームに対して貧弱な実行効果を有する場合、生成されるスーパーネットワークのネットワーク・スケールを低減するために、演算子セットは、第2の知識ベースに基づいて交換されてもよい（保持されない）。

【0028】

水平拡張モードは、ソース・ニューラル・ネットワークのネットワーク構造情報に基づいて、ソース・ニューラル・ネットワークの1つのレベルに含まれる演算子セットを、スーパーネットワークの1つのレベルに含まれる一つまたは複数の演算子セットに拡張することを含む。垂直拡張モードは、ソース・ニューラル・ネットワークのネットワーク構造情報に基づいて、ソース・ニューラル・ネットワークにおける1つのレベルを複数のレベルに垂直に拡張し、次いで、対応して、垂直拡張後に得られた前記複数のレベルに含まれる演算子セットに基づいて、スーパーネットワークの複数のレベルに含まれる複数の演算子セットを得ることを含む。

【0029】

トレーニングされるべきスーパーネットワークに含まれる得られた演算子セットの関数は、ソース・ニューラル・ネットワークにおける対応する演算子セットの関数と同じであるため、スーパーネットワークに見出されたターゲット・ニューラル・ネットワークは、ソース・ニューラル・ネットワークのタスク、たとえば画像認識における画像分類タスクを実行することができる。

【0030】

ある可能な実装では、トレーニングされるべきスーパーネットワークにおける第3演算子セットは、第1演算子セットおよび／または派生第2演算子セットを含み、第1演算子セットは、ソース・ニューラル・ネットワークに含まれる。

【0031】

スーパーネットワークにおける各第3演算子セットは、ソース・ニューラル・ネットワークの1つのレベルに対応してもよく、第2演算子セットは、知識ベースに基づいて、各レベルにおいて第1演算子セットから導出されてもよい。この場合、第3演算子セットは、第1演算子セットと派生第2演算子セットとを含む。あるいはまた、知識ベースに基づいて、あるレベルにおける派生第2演算子セットが第1演算子セットを置き換えて、対応する第3

10

20

30

40

50

演算子セットが派生第2演算子セットのみを含むようにしてもよいと決定される。むろん、知識ベースに基づいて、あるレベルに派生第2演算子セットが存在せず、対応する第3演算子セットが第1演算子セットのみを含むと判断されてもよい。

【0032】

ある可能な実装では、前記一つまたは複数の評価メトリックおよび対応情報に基づいて、トレーニングされたスーパーネットワークに含まれる一つまたは複数の評価されるべき候補サブネットの評価結果（単数または複数）を決定することは：前記一つまたは複数の評価メトリックおよび対応情報に基づいて、トレーニングされたスーパーネットワークにおける第3演算子セットにおける第1演算子セットおよび／または派生第2演算子セットに対応する評価結果を決定するステップと；N個の評価メトリックの間のプロポーショナル関係に基づいて、および第3演算子セット内の第1演算子セットおよび／または派生第2演算子セットに対応する評価結果に基づいて、前記一つまたは複数の評価されるべき候補サブネットの評価結果（単数または複数）を取得するステップとを含む。

10

【0033】

プロポーショナル関係は、ターゲット・ハードウェア・プラットフォームによって実行されるターゲット・タスク、ターゲット・ハードウェア・プラットフォームの属性などに基づいて決定されてもよい。N個の評価メトリックの間のプロポーショナル関係は、N個の評価メトリックの間の相対的重要度を反映することが理解されうる。ターゲット・ハードウェア・プラットフォーム上の見出されたターゲット・ニューラル・ネットワークの評価効果は、より高い比率の評価メトリックに基づいたほうがよいことが明らかである。たとえば、評価メトリックは、精度と時間オーバーヘッドを含む。ターゲット・ハードウェア・プラットフォームは、郊外のビデオ監視画像を処理するように構成されたハードウェア・プラットフォームである。単位時間あたりに収集される監視画像は、少量の情報を含み、高い画像処理精度が確保される必要がある。しかしながら、低い処理速度が要求される。そのため、プロポーショナル関係においては、高い精度の比率が設定されてもよい。このようにして、より良好なターゲット・ニューラル・ネットワークを見つけることができるよう、ターゲット・ハードウェア・プラットフォーム上でよりよい実行性能を有する候補サブネットが、プロポーショナル関係と評価結果に基づいて得られてもよい。

20

【0034】

ある可能な実装では、N個の評価メトリックの間のプロポーショナル関係に基づいて、および第3演算子セットにおける第1演算子セットおよび／または派生第2演算子セットに対応する評価結果に基づいて、前記一つまたは複数の評価されるべき候補サブネットの評価結果（単数または複数）を取得することは：N個の評価メトリック間のr個のプロポーショナル関係を受領し、N個の評価メトリックの間の各プロポーショナル関係について事前探索を実行するステップであって、rは正の整数である、ステップと；1つの事前探索プロセスにおいて、第1演算子セットおよび／または派生第2演算子セットの評価結果の重みを設定し、N個の評価メトリックと該N個の評価メトリック間のプロポーショナル関係に基づいて、トレーニングされたスーパーネットワークを再トレーニングするステップと；トレーニング・プロセスにおいて、第1演算子セットおよび派生第2演算子セットの評価結果に対応する重みを調整するステップと；評価されるべき候補サブネットおよび評価されるべき候補サブネットの評価結果を、第1演算子セットおよび派生第2演算子セットの評価結果に対応する調整された重みに基づいて決定するステップと；前記r個の事前探索プロセスに基づいて前記一つまたは複数の評価されるべき候補サブネットの評価結果（単数または複数）を決定するステップとを含む。

30

40

【0035】

事前探索プロセスは、ターゲット探索範囲を狭くすることである。候補サブネットの一部は、まず、異なるプロポーショナル関係に基づいて決定され、次いで、候補サブネットの決定された一部は、ターゲット・ニューラル・ネットワークのネットワーク構造情報を決定するために、前記一つまたは複数の評価メトリックに基づいて評価される。

【0036】

50

事前探索プロセスでは、各第3演算子セット内にあり、かつターゲット・ハードウェア・プラットフォームに比較的適用可能である第1演算子セットまたは派生第2演算子セットを決定するために、重みの比率が調整される。このようにして、各事前探索プロセスにおいて、現在のプロポーション関係に基づいて、最適な性能を有する候補サブネットおよび候補サブネットの評価結果が決定されうる。次いで、複数のプロポーション関係に基づいて複数の候補サブネットを得ることができる。このようにして、ターゲット・ニューラル・ネットワークは、複数の候補サブネットから決定されて、ターゲット・ニューラル・ネットワークのネットワーク構造情報を決定してもよい。

【0037】

ある可能な実装では、N個の評価メトリックは、ニューラル・ネットワーク・モデルの精度評価メトリックおよびニューラル・ネットワーク・モデルの時間オーバーヘッド評価メトリックを含み；対応して、N個の評価メトリックの間のプロポーション関係は、精度評価メトリックと時間オーバーヘッド評価メトリックとの間のプロポーション関係を含む。

【0038】

たとえば、ユーザーは、精度評価メトリックと時間オーバーヘッド評価メトリックとの間の第1の比率と第2の比率を入力する。このようにして、ターゲット・ニューラル・ネットワークの精度は、ソース・ニューラル・ネットワークの精度を下回らないか、またはソース・ニューラル・ネットワークの精度と大きく異ならないことを保証することができる。また、ターゲット・ニューラル・ネットワークの精度を確保しつつ、時間オーバーヘッドが低減される、つまり、ターゲット・ハードウェア・プラットフォーム上のターゲット・ニューラル・ネットワークの実行速度が高められ、それにより作業効率を向上させる。

【0039】

ある可能な実装では、前記一つまたは複数の評価されるべき候補サブネットの評価結果（単数または複数）に基づいて、トレーニングされたスーパーネットワーク上で探索処理を実行して、ターゲット・ニューラル・ネットワークのネットワーク構造情報を得ることは、さらに：評価されるべき候補サブネットの評価結果（単数または複数）をソートして、最初のK個の評価結果に対応するK個の第1のターゲット候補サブネットを得るステップと；トレーニングされたスーパーネットワークおよびK個の第1のターゲット候補サブネットに基づく第nラウンドの混成および／または変成の後に得られるS個の第2のターゲット候補サブネットを得るステップであって、nはあらかじめ設定されたスクリーニング回数であり、 $n \geq 1$ であり、KおよびSは正の整数であり、 $S > K$ である、ステップと；前記一つまたは複数の評価メトリックおよび前記対応情報に基づいてS個の第2のターゲット候補サブネットの評価結果を取得するステップと、それらの評価結果をソートして、最初のK個の評価結果に対応するK個の第2のターゲット候補サブネットを得るステップと； $K \times (n + 1)$ 個のターゲット候補サブネットから一つまたは複数のターゲット・ニューラル・ネットワークのネットワーク構造情報を取得するステップとを含む。

【0040】

本願では、混成〔ハイブリダイゼーション〕は、K個の第1のターゲット候補サブネットの一部において対応するレベルにおいて含まれる演算子セットの一部が交換されて、第2のターゲット候補サブネットを得ることを意味する。変成〔ミューテーション〕は、K個の第1のターゲット候補サブネットの一部に含まれる演算子セットの一部が、トレーニングされたスーパーネットワークにおいて対応するレベルにおいて演算子セットの一部で置き換えられて、第2のターゲット候補サブネットを得ることを意味する。

【0041】

たとえば、 $(n + 1)$ ラウンドの逐次反復トレーニングの後、具体的には、nラウンドの混成および／または変成の後、 $K \times (n + 1)$ 個のターゲットサブネットが得られる。 $K \times (n + 1)$ 個のターゲット候補サブネットは、前記一つまたは複数の評価メトリックに基づいて再度ソートされて、前記一つまたは複数のターゲット・ニューラル・ネットワークのネットワーク構造情報を得てもよい。

【0042】

10

20

30

40

50

このようにして、前記一つまたは複数の評価されるべき候補サブネットの評価結果（単数または複数）は、まず、より良い候補サブネットの一部を得るためにソートされ、その結果、ターゲット探索範囲が狭められる。次いで、混成および／または変成方法を用いることにより、よりよい候補サブネットの前記一部に基づいて、より多くの候補サブネットが生成される。逐次反復プロセスにおいて、各逐次反復プロセスにおける、より良い候補サブネットの一部が得られ、前記一つまたは複数の評価メトリックを満たし、よりよい性能を有する一つまたは複数のターゲットサブネットが、それらの候補サブネットから決定される。

【0043】

ある可能な実装では、ソース・ニューラル・ネットワークはソース・ハードウェア・プラットフォーム上で動作するニューラル・ネットワークであり、またはターゲット・ニューラル・ネットワークはターゲット・ハードウェア・プラットフォーム上で動作するニューラル・ネットワークであり；ソース・ハードウェア・プラットフォームはターゲット・ハードウェア・プラットフォームとは異なる。

【0044】

ソース・ハードウェア・プラットフォームは、ターゲット・ハードウェア・プラットフォームとは異なるハードウェア・プラットフォームであり、ソース・ハードウェア・プラットフォームおよびターゲット・ハードウェア・プラットフォームは、異なる属性をもつプロセッサ、または異なるモデルをもつプロセッサ、または同じ属性であるが異なるモデルをもつプロセッサを含む。ソース・ハードウェア・プラットフォームは、ターゲット・ハードウェア・プラットフォームに置き換えられる。ターゲット・ハードウェア・プラットフォームのために使用される必要のあるターゲット・ニューラル・ネットワークは、ソース・ハードウェア・プラットフォーム上で実行されるソース・ニューラル・ネットワークに基づいて探索される。

【0045】

たとえば、ソース・ニューラル・ネットワークは、GPU上で動作するニューラル・ネットワーク1であり、今、GPUがチップAに切り換える必要がある。この場合、ニューラル・ネットワーク1とチップAに基づいてターゲット・ニューラル・ネットワークを探索する必要がある。チップA上でのターゲット・ニューラル・ネットワークの実行効果は、チップA上でのニューラル・ネットワーク1の実行効果よりもよい。

【0046】

ある可能な実装では、ソース・ニューラル・ネットワークおよびターゲット・ニューラル・ネットワークは、異なる時間領域においてターゲット・ハードウェア・プラットフォーム上で稼働するニューラル・ネットワークである。

【0047】

たとえば、ソース・ニューラル・ネットワークは、GPU上で動作するニューラル・ネットワーク1であり、今、GPUの画像処理速度を上げる必要がある。この場合、ニューラル・ネットワーク1は最適化される必要があり、発見されたターゲット・ニューラル・ネットワークは最適化されたニューラル・ネットワーク1である。ニューラル・ネットワーク1とターゲット・ニューラル・ネットワークは、順次実行されるか、またはGPU（ターゲット・ハードウェア・プラットフォーム）上で担持される。

【0048】

ある可能な実装では、本方法は、さらに：ターゲット・ニューラル・ネットワークのネットワーク構造情報に基づいてターゲット・ニューラル・ネットワークを構築するステップと；第2のトレーニング・データセットに基づいて、ターゲット・ニューラル・ネットワーク上でトレーニング・プロセスを実行して、トレーニングされたターゲット・ニューラル・ネットワークを得るステップとを含む。

【0049】

第2のトレーニング・データセットは、第1のトレーニング・データセットと同じか、またはそれと異なるトレーニング・データセットであってよく、ニューラル・ネットワーク

10

20

30

40

50

のトレーニング・プロセスに参加するために使用される。ターゲット・ニューラル・ネットワークのネットワーク構造情報は、前述の方法を用いて見出され、それによりターゲット・ニューラル・ネットワークが構築できる。ターゲット・ニューラル・ネットワークはトレーニング・データセットを用いてトレーニングされ、ターゲット・ニューラル・ネットワーク内のパラメータ情報が調整されて、ターゲット・タスクを実行するためにターゲット・ハードウェア・プラットフォーム上で実行できるニューラル・ネットワークを得る。

【0050】

ある可能な実装では、トレーニングされたターゲット・ニューラル・ネットワークは、ターゲット検出タスクおよび画像分類タスクを実行するために、ターゲット・ハードウェア・プラットフォーム上で動作する。

【0051】

このようにして、得られたトレーニングされたターゲット・ニューラル・ネットワークは、ターゲット・ハードウェア・プラットフォーム上でターゲット・タスクを実行することができる。たとえば、交通信号認識タスク、歩行者・車両検出タスクなどを実行することができる。加えて、トレーニングされたターゲット・ニューラル・ネットワークは、ターゲット・ハードウェア・プラットフォームに対して、より良好な実行効果を有する。たとえば、ソース・ニューラル・ネットワークによるターゲット・ハードウェア・プラットフォーム上での交通信号認識タスクを実行することの時間オーバーヘッドは0.2msであり、トレーニングされたターゲット・ニューラル・ネットワークによるターゲット・ハードウェア・プラットフォーム上での交通信号認識タスクを実行することの時間オーバーヘッドは0.1msである。精度は基本的には変わらないままであることを保証しつつ、時間オーバーヘッドは50%低減される。

【0052】

第2の側面によれば、本願は、ニューラル・ネットワークのための適応的探索装置を提供する。本装置は：受領ユニットと、トレーニング・ユニットと、探索ユニットとを含んでいてもよい。受領ユニットは、探索条件セットを受領ように構成される。探索条件セットはターゲット・ハードウェア・プラットフォーム情報、ソース・ニューラル・ネットワークのネットワーク構造情報、および一つまたは複数の評価メトリックを含み、ソース・ニューラル・ネットワークのネットワーク構造情報はソース・ニューラル・ネットワークを表すために使用される。トレーニング・ユニットは、トレーニング・データセットに基づいて、トレーニングされるべきスーパーネットワーク上でトレーニング・プロセスを実行して、トレーニングされたスーパーネットワークを得るように構成される。トレーニングされるべきスーパーネットワークは、ソース・ニューラル・ネットワークのネットワーク構造を拡張することによって得られる。探索ユニットは、前記一つまたは複数の評価メトリックに基づいて、トレーニングされたスーパーネットワーク上でサブネット探索プロセスを実行して、ターゲット・ニューラル・ネットワークのネットワーク構造情報を得るように構成される。ターゲット・ニューラル・ネットワークのネットワーク構造情報は、ターゲット・ニューラル・ネットワークを表すために使用され、ターゲット・ハードウェア・プラットフォーム上で実行されるターゲット・ニューラル・ネットワークの評価結果は、ソース・ハードウェア・プラットフォーム上で実行されるソース・ニューラル・ネットワークの評価結果よりよい。

【0053】

ある可能な実装において、前記一つまたは複数の評価メトリックは、ニューラル・ネットワーク・モデルの精度評価メトリック、ニューラル・ネットワーク・モデルの時間オーバーヘッド評価メトリック、ニューラル・ネットワーク・モデルの記憶スペース評価メトリック、ニューラル・ネットワーク・モデルの電力消費評価メトリック、ニューラル・ネットワーク・モデルのテンソル計算ユニットの利用率評価メトリック、およびニューラル・ネットワーク・モデルのメモリ読み出し／書き込み速度評価メトリックのうちの一つまたは複数を含む。ニューラル・ネットワーク・モデルの精度評価メトリックは、ターゲット・ニューラル・ネットワークの精度が精度閾値以上であることを含む。ニューラル・ネ

10

20

30

40

50

ットワーク・モデルの時間オーバーヘッド評価メトリックは、ターゲット・ニューラル・ネットワークの時間オーバーヘッドが時間オーバーヘッド閾値を超えないことを含む。ニューラル・ネットワーク・モデルの記憶スペース評価メトリックは、ターゲット・ニューラル・ネットワークによって占有される記憶スペースが記憶スペース閾値を超えないことを含む。ニューラル・ネットワーク・モデルの電力消費評価メトリックは、ターゲット・ニューラル・ネットワークの電力消費が電力消費閾値を超えないことを含む。ニューラル・ネットワーク・モデルのテンソル計算ユニットの利用率評価メトリックは、ターゲット・ニューラル・ネットワークのテンソル計算ユニットの利用率がテンソル計算ユニットの利用率閾値以上であることを含む。ニューラル・ネットワーク・モデルのメモリ読み出し／書き込み速度評価メトリックは、ターゲット・ニューラル・ネットワークのメモリ読み出し／書き込み速度がメモリ読み出し／書き込み速度閾値以上であることを含む。

10

【0054】

ある可能な実装では、探索ユニットは具体的には：第1の知識ベースを使用することによって、トレーニングされたスーパーネットワークからターゲット探索範囲内の一つまたは複数の評価されるべき候補サブネットを決定し；前記一つまたは複数の評価メトリックに基づいて、ターゲット探索範囲内の前記一つまたは複数の評価されるべき候補サブネット上でサブネット探索プロセスを実行して、ターゲット・ニューラル・ネットワークのネットワーク構造情報を得るように構成される。

【0055】

ある可能な実装では、探索ユニットは具体的には：ターゲット・ハードウェア・プラットフォーム、M個のニューラル・ネットワークのネットワーク構造、N個の評価メトリック、およびP個の評価結果の間の対応についての情報を得るステップであって、ここで、M、N、Pはすべて正の整数であり、M個のニューラル・ネットワークのネットワーク構造は、トレーニングされたスーパーネットワークのネットワーク構造に含まれる、ステップと；前記一つまたは複数の評価メトリックおよび前記対応情報に基づいて、トレーニングされたスーパーネットワークに含まれる一つまたは複数の評価されるべき候補サブネットの評価結果（単数または複数）を決定するステップと；前記一つまたは複数の評価されるべき候補サブネットの前記評価結果（単数または複数）に基づいて、トレーニングされたスーパーネットワーク上で探索プロセスを実行して、ターゲット・ニューラル・ネットワークのネットワーク構造情報を得るステップとを実行するように構成される。

20

30

【0056】

ある可能な実装では、前記一つまたは複数の評価メトリックは、ニューラル・ネットワーク・モデルの時間オーバーヘッド評価メトリックを含み、ニューラル・ネットワーク・モデルの時間オーバーヘッド評価メトリックは、ターゲット・ニューラル・ネットワークの時間オーバーヘッドが時間オーバーヘッド閾値よりも大きくないことを含む。対応して、探索ユニットは具体的には、ニューラル・ネットワーク・モデルの時間オーバーヘッド評価メトリックに基づいて、トレーニングされたスーパーネットワークに含まれる一つまたは複数の候補サブネット上のサブネット評価を実行して、ニューラル・ネットワーク・モデルの時間オーバーヘッドが時間オーバーヘッド閾値よりも小さい一つまたは複数のターゲット・ニューラル・ネットワークを得るように構成される。

40

【0057】

ある可能な実装では、本装置は、ソース・ニューラル・ネットワークのネットワーク構造、第2の知識ベースにおける拡張規則、およびターゲット・ハードウェア・プラットフォーム情報に基づいて、トレーニングされるべきスーパーネットワークを水平拡張モードおよび／または垂直拡張モードで構築するように構成された構築ユニットをさらに含む。

【0058】

ある可能な実装では、トレーニングされるべきスーパーネットワークにおける第3演算子セットは、第1演算子セットおよび／または派生第2演算子セットを含み、第1演算子セットは、ソース・ニューラル・ネットワークに含まれる。

【0059】

50

ある可能な実装では、探索ユニットは具体的には：前記一つまたは複数の評価メトリックおよび対応情報に基づいて、トレーニングされたスーパーネットワークにおける第3演算子セットにおける第1演算子セットおよび／または派生第2演算子セットに対応する評価結果を決定するステップと；N個の評価メトリックの間のプロポーション関係に基づいて、および第3演算子セット内の第1演算子セットおよび／または派生第2演算子セットに対応する評価結果に基づいて、前記一つまたは複数の評価されるべき候補サブネットの評価結果（単数または複数）を取得するステップと；前記一つまたは複数の評価されるべき候補サブネットの評価結果（単数または複数）に基づいて、トレーニングされたスーパーネットワーク上で探索プロセスを実行して、ターゲット・ニューラル・ネットワークのネットワーク構造情報を得るステップとを実行するように構成される。

10

【0060】

ある可能な実装では、探索ユニットは具体的には：N個の評価メトリックの間のr個のプロポーション関係を受領し、N個の評価メトリックの間の各プロポーション関係について事前探索を実行するステップであって、rは正の整数である、ステップと；1つの事前探索プロセスにおいて、第1演算子セットおよび／または派生第2演算子セットの評価結果の重みを設定し、N個の評価メトリックと該N個の評価メトリックの間のプロポーション関係に基づいて、トレーニングされたスーパーネットワークを再トレーニングするステップと；トレーニング・プロセスにおいて、第1演算子セットおよび派生第2演算子セットの評価結果に対応する重みを調整するステップと；評価されるべき候補サブネットおよび該評価されるべき候補サブネットの評価結果を、第1演算子セットおよび派生第2演算子セットの評価結果に対応する調整された重みに基づいて決定するステップと；r個の事前探索プロセスに基づいて前記一つまたは複数の評価されるべき候補サブネットの評価結果（単数または複数）を決定するステップとを実行するように構成される。

20

【0061】

ある可能な実装では、N個の評価メトリックは、ニューラル・ネットワーク・モデルの精度評価メトリックおよびニューラル・ネットワーク・モデルの時間オーバーヘッド評価メトリックを含み；対応して、N個の評価メトリックの間のプロポーション関係は、精度評価メトリックと時間オーバーヘッド評価メトリックとの間のプロポーション関係を含む。

【0062】

ある可能な実装では、探索ユニットはさらに：前記一つまたは複数の評価されるべき候補サブネットの評価結果をソートして、最初のK個の評価結果に対応するK個の第1のターゲット候補サブネットを得るステップと；トレーニングされたスーパーネットワークおよびK個の第1のターゲット候補サブネットに基づく第nラウンドの混成および／または変成の後に得られるS個の第2のターゲット候補サブネットを得るステップであって、nはあらかじめ設定されたスクリーニング回数であり、 $n \geq 1$ であり、KおよびSは正の整数であり、 $S > K$ である、ステップと；前記一つまたは複数の評価メトリックおよび前記対応情報に基づいてS個の第2のターゲット候補サブネットの評価結果を取得するステップと、最初のK個の評価結果に対応するK個の第2のターゲット候補サブネットを得るステップと； $K \times (n + 1)$ 個のターゲット候補サブネットから一つまたは複数のターゲット・ニューラル・ネットワークのネットワーク構造情報を取得するステップとを実行するように構成される。

30

40

【0063】

ある可能な実装では、ソース・ニューラル・ネットワークはソース・ハードウェア・プラットフォーム上で動作するニューラル・ネットワークであり、またはターゲット・ニューラル・ネットワークはターゲット・ハードウェア・プラットフォーム上で動作するニューラル・ネットワークであり；ソース・ハードウェア・プラットフォームはターゲット・ハードウェア・プラットフォームとは異なる。

【0064】

ある可能な実装では、ソース・ニューラル・ネットワークおよびターゲット・ニューラル・ネットワークは、異なる時間領域においてターゲット・ハードウェア・プラットフォーム上で稼働するニューラル・ネットワークである。

50

【0065】

ある可能な実装では、構築ユニットさらに：ターゲット・ニューラル・ネットワークのネットワーク構造情報に基づいてターゲット・ニューラル・ネットワークを構築するように構成される。トレーニング・ユニットはさらに、第2のトレーニング・データセットに基づいて、ターゲット・ニューラル・ネットワーク上でトレーニング・プロセスを実行して、トレーニングされたターゲット・ニューラル・ネットワークを得るように構成される。

【0066】

ある可能な実装では、トレーニングされたターゲット・ニューラル・ネットワークは、ターゲット検出タスクおよび画像分類タスクを実行するために、ターゲット・ハードウェア・プラットフォーム上で動作する。

【0067】

第3の側面によれば、本願は、サーバーを提供する。サーバーは、一つまたは複数のプロセッサ、メモリ、および一つまたは複数のコンピュータ・プログラムを含むことができる。前記一つまたは複数のコンピュータ・プログラムはメモリに記憶され、前記一つまたは複数のコンピュータ・プログラムは命令を含む。命令がサーバーによって実行されるとき、サーバーは、第1の側面および第1の側面の可能な実装のうちの任意の1つに従って、ニューラル・ネットワークの適応的探索方法を実行することを可能にされる。

【0068】

第4の側面によれば、本願は、装置を提供する。本装置は、第1の側面および第1の側面の可能な実装のうちの任意の1つに従う、ニューラル・ネットワークの適応的探索方法を実装する機能を有する。この機能は、ハードウェアによって実装されてもよいし、または対応するソフトウェアを実行するハードウェアによって実装されてもよい。ハードウェアまたはソフトウェアは、機能に対応する一つまたは複数のモジュールを含む。

【0069】

第5の側面によれば、本願は、コンピュータ命令を含むコンピュータ読み取り可能な記憶媒体を提供する。コンピュータ命令がサーバー上で実行されるとき、サーバーは、第1の側面および第1の側面の可能な実装のうちの任意の1つに従って、ニューラル・ネットワークの適応的探索方法を実行することを可能にされる。

【0070】

第6の側面によれば、本願は、コンピュータ・プログラム・プロダクトを提供する。コンピュータ・プログラム・プロダクトがサーバー上で実行されるとき、サーバーは、第1の側面および第1の側面の可能な実装のうちの任意の1つに従う、ニューラル・ネットワークの適応的探索方法を実行することを可能にされる。

【0071】

第7の側面によれば、回路システムが提供され、回路システムは、処理回路を含み、処理回路は、第1の側面および第1の側面の可能な実装のうちの任意の1つに従う、ニューラル・ネットワークの適応的探索方法を実行するように構成される。

【0072】

第8の側面によれば、本願のある実施形態は、少なくとも1つのサーバーおよび少なくとも1つの端末装置を含む通信システムを提供する。前記少なくとも1つのサーバーは、第1の側面および第1の側面の可能な実装のうちの任意の1つに従って、ニューラル・ネットワークの適応的探索方法を実行する。前記少なくとも1つの端末装置は、前記少なくとも1つのサーバーに探索条件セットを送信する。ここで、探索条件セットは、ターゲット・ハードウェア・プラットフォーム情報、ソース・ニューラル・ネットワークのネットワーク構造情報、および一つまたは複数の評価メトリックを含み、ソース・ニューラル・ネットワークのネットワーク構造情報は、ソース・ニューラル・ネットワークを表すために使用される。前記少なくとも1つの端末装置は、前記少なくとも1つのサーバーによって送信されたトレーニングされたターゲット・ニューラル・ネットワークを受領する。

【0073】

ある可能な実装では、前記少なくとも1つの端末装置は、トレーニングされたターゲッ

10

20

30

40

50

ト・ニューラル・ネットワークをターゲット・ハードウェア・プラットフォームを含むユーザー装置に送るかどうかを決定するようにさらに構成される。

【図面の簡単な説明】

【0074】

【図1】本願のある実施形態による、ニューラル・ネットワークの適応的探索方法の適用シナリオの概略図である。

【0075】

【図2】本願のある実施形態によるシステムアーキテクチャーの概略図である。

【0076】

【図3】本願のある実施形態によるチップのハードウェア構造の概略図である。

10

【0077】

【図4】本願のある実施形態による、ニューラル・ネットワークの適応的探索方法の概略フローチャート1である。

【0078】

【図5】本願のある実施形態による、ソース・ニューラル・ネットワークのネットワーク構造の概略図1である。

【0079】

【図6】本願のある実施形態による、ソース・ニューラル・ネットワークのネットワーク構造の概略図2である。

【0080】

【図7】本願のある実施形態による、スーパーネットワーク構築プロセスの概略図1である。

20

【0081】

【図8】本願のある実施形態による、スーパーネットワーク構築プロセスの概略図2である。

【0082】

【図9】本願のある実施形態による、スーパーネットワークのネットワーク構造の概略図である。

【0083】

【図10】本願のある実施形態による、ニューラル・ネットワークの適応的探索方法の概略フローチャート2である。

30

【0084】

【図11a】本願のある実施形態による、評価されるべき候補サブネットを得るプロセスの概略図である。

【図11b】本願のある実施形態による、評価されるべき候補サブネットを得るプロセスの概略図である。

【図11c】本願のある実施形態による、評価されるべき候補サブネットを得るプロセスの概略図である。

【0085】

【図12a】本願のある実施形態による、ターゲット候補サブネットを得るプロセスの概略図である。

40

【図12b】本願のある実施形態による、ターゲット候補サブネットを得るプロセスの概略図である。

【図12c】本願のある実施形態による、ターゲット候補サブネットを得るプロセスの概略図である。

【0086】

【図13】本願のある実施形態による、ニューラル・ネットワークの適応的探索方法の概略フローチャート3である。

【0087】

【図14】本願のある実施形態による装置の構造の概略図である。

50

【0088】

【図15】本願のある実施形態による装置のハードウェア構造の概略図である。

【発明を実施するための形態】

【0089】

以下は、添付の図面を参照して、本願の実施形態において提供されるニューラル・ネットワークの適応的探索方法および装置について詳細に説明する。

【0090】

図1は、ニューラル・ネットワークのための適応的探索システムを示す。ニューラル・ネットワークの適応的探索システムは、データ処理装置および端末装置を含む。データ処理装置および端末装置は、有線ネットワークまたは無線ネットワークを介して接続されてもよい。装置間の接続の仕方は、本願のこの実施形態において特に限定されない。

【0091】

端末装置はクライアント装置であり、開発者によって操作される端末装置であってもよく、関連するヒューマン・マシン対話インターフェースを提供し、それによりユーザーは、ニューラル・ネットワークの適応的探索プロセスにおいて必要とされる関連パラメータ、たとえば探索条件セットを入力する。たとえば、探索条件セットは、ターゲット・ハードウェア・プラットフォーム情報（たとえば、ターゲット・ハードウェア・プラットフォームの製造業者、属性、モデル、および実行されるべきターゲット・タスクのタイプ）、ソース・ニューラル・ネットワークのネットワーク構造情報、および一つまたは複数の評価メトリック（たとえば、精度閾値、時間オーバーヘッド閾値、および記憶スペース閾値）を含んでいてもよい。たとえば、端末装置は、デスクトップ・コンピュータ、ラップトップ・コンピュータ、ハンドヘルド・コンピュータ、またはノートブック・コンピュータのような端末装置であってもよい。端末装置の特定の形態は、本願のこの実施形態において特に限定されない。

【0092】

ハードウェア・プラットフォームは、たとえば、グラフィックス処理ユニット（graphics processing unit、GPU）、ホスト中央処理ユニット（central processing unit、CPU）、画像信号プロセッサ（image signal processor、ISP）、ニューラル・ネットワーク処理ユニット（neural network processing unit、NPU）、デジタル信号プロセッサ（digital signal processor、DSP）等のうちの一つまたは複数を含むことができる。実行中または担持されるニューラル・ネットワークによって実現されうるアプリケーションは、画像分類（image classification）、オブジェクト認識（object recognition）、アクション認識（action recognition）、姿勢推定（pose estimation）、およびニューラルスタイル転送（neural style transfer）などのコンピュータビジョン・アプリケーションを含むがこれらに限定されず、または自然言語処理（natural language processing、NLP）などのアプリケーションであってもよい。

【0093】

あるシナリオでは、端末装置内のニューラル・ネットワーク（すなわち、ソース・ニューラル・ネットワーク）を最適化する場合、端末装置は、最適化されたニューラル・ネットワーク（すなわち、ターゲット・ニューラル・ネットワーク）を得るために、ソース・ニューラル・ネットワークのネットワーク構造情報、ソース・ニューラル・ネットワークが担持される特定のハードウェア・プラットフォーム（すなわち、ターゲット・ハードウェア・プラットフォーム）に関する情報などを、処理のためのデータ処理装置に入力することができる。すなわち、ソース・ニューラル・ネットワークとターゲット・ニューラル・ネットワークは、異なる時間領域で同じハードウェア・プラットフォーム上で動作するニューラル・ネットワークである。

【0094】

別のシナリオでは、端末装置内のソース・ニューラル・ネットワークのハードウェア・プラットフォームが、別のハードウェア・プラットフォーム（たとえば、異なる属性を有するハードウェア・プラットフォーム（たとえば、GPU、CPU、NPU、またはDSP）、同

10

20

30

40

50

じ属性であるが異なるモデルを有するハードウェア・プラットフォーム、または同じ属性であるが異なる製造者を有するハードウェア・プラットフォームを含む）に置き換えられる場合、ソース・ニューラル・ネットワークのネットワーク構造情報および置き換え後のハードウェア・プラットフォーム（すなわち、ターゲット・ハードウェア・プラットフォーム）に関する情報が、適応的探索を通じて得られるニューラル・ネットワークを得るために処理のためにデータ処理装置に入力されうる。交換後のハードウェア・プラットフォーム（すなわち、ターゲット・ハードウェア・プラットフォーム）の設計は、ソース・ハードウェア・プラットフォームの設計とは異なる可能性があるため、ソース・ニューラル・ネットワークの実行効果が影響を受ける。あるいは、交換後のハードウェア・プラットフォーム内のいくつかの設計は、ソース・ニューラル・ネットワークに完全には適用されないことがあり、結果として、ネットワーク性能が低下し、ハードウェア・プラットフォームの設計の無駄が生じる。

【0095】

いくつかの他の例において、ニューラル・ネットワークの適応的探索システムは、ユーザー装置をさらに含むことができることが理解されうる。ターゲット・ニューラル・ネットワークが、タスクを実行するために、ユーザー装置内の対応するターゲット・ハードウェア・プラットフォーム上で動作するまたは担持されるよう、ターゲット・ニューラル・ネットワークの得られたネットワーク構造情報をターゲット・ハードウェア・プラットフォームを含むユーザー装置に送信するかどうかを端末装置がさらに決定してもよい。たとえば、ユーザー装置は、携帯電話（mobile phone）、デスクトップ・コンピュータ、ラップトップ・コンピュータ、ハンドヘルド・コンピュータ、ノートブック・コンピュータ、ウルトラ・モバイル・パーソナル・コンピュータ（ultra-mobile personal computer、UMPC）、ネットブック、パーソナルデジタルアシスタント（personal digital assistant、PDA）、拡張現実（augmented reality、AR）/仮想現実（virtual reality、VR）装置、または人工知能（artificial intelligence、AI）端末のような端末装置であってもよい。ユーザー装置の特定の形態は、本開示のこの実施形態では特に限定されない。

【0096】

データ処理装置は、ニューラル・ネットワーク探索機能を有する装置またはサーバー、たとえばクラウドサーバーまたはネットワークサーバーであってもよい。データ処理装置は、対話インターフェースを通じて、端末装置によって送信された探索条件セットを受領し、該探索条件セットは、ターゲット・ハードウェア・プラットフォーム情報、ソース・ニューラル・ネットワークのネットワーク構造情報、前記一つまたは複数の評価メトリック等を含む；次いで、データ処理装置は、メモリに記憶された知識ベースに基づき、プロセッサを用いてスーパーネットワークを構築し、トレーニングし、ターゲット・ニューラル・ネットワークを探索し、探索結果、すなわちターゲット・ニューラル・ネットワークを対応する端末装置に送信する。データ処理装置内のメモリは、一般的な名称であってもよく、ローカル記憶装置と、履歴データを記憶するデータベースとを含む。データベースは、データ処理装置上に位置してもよく、または別のクラウドサーバー上に位置してもよい。

【0097】

データ処理装置はサーバーであってもよく、または複数のサーバーを含むサーバークラスターであってもよく、またはクラウドコンピューティングサービスセンターであってもよいことに留意されたい。

【0098】

たとえば、図1のデータ処理装置が、本願の実施形態におけるニューラル・ネットワークの適応的探索方法を実行することができる。

【0099】

図2は、本願のある実施形態によるシステムアーキテクチャーの一例を示す。

【0100】

10

20

30

40

50

図2のシステムアーキテクチャーに示されるように、入力/出力（input/output、I/O）インターフェース211がデータ処理装置210のために構成され、外部装置とデータを交換するように構成される。ユーザーは、端末装置220を使用して、I/Oインターフェース211にデータを入力してもよい。入力データは、本願の実施形態においては探索条件セットを含んでいてもよい。たとえば、探索条件セットは、ターゲット・ハードウェア・プラットフォーム情報（たとえば、製造業者、属性、モデル、およびターゲット・ハードウェア・プラットフォームの実行されるべきターゲット・タスクのタイプ）、ソース・ニューラル・ネットワークのネットワーク構造情報、および一つまたは複数の評価メトリック（たとえば、精度閾値、時間オーバーヘッド閾値、および記憶スペース閾値）を含んでいてもよい。

【0101】

データ収集装置240は、トレーニング・データを収集するように構成される。たとえば、本願のこの実施形態におけるトレーニング・データは、ニューラル・ネットワークをトレーニングするためのトレーニング・データセットであってもよい。トレーニング・データを収集した後、データ収集装置240は、トレーニング・データをデータベース230に記憶する。データベース230は、さらに、データ収集装置240によって収集された知識ベース231を記憶する。知識ベース231は、ソース・ニューラル・ネットワークに基づくスーパーネットワークを構築するために使用されてもよい。知識ベース231は、ニューラル・ネットワーク内の演算子および／または演算子セットとハードウェア・プラットフォームとの間の整合に関する情報を含んでいてもよい。たとえば、知識ベース231は、第1演算子セットと第3演算子セットとの間の、異なるハードウェア・プラットフォームに基づく対応を含む。第3演算子セットは、第1演算子セットおよび／または一つまたは複数の派生第2演算子セットを含む。各派生第2演算子セットは、その派生第2演算子セットに対応する第1演算子セットと同じ機能を有する。換言すれば、第1演算子セットは、派生第2演算子セットと置き換えられてもよい。この場合、知識ベース231を使用してソース・ニューラル・ネットワークがスーパーネットワークに拡張されると、ソース・ニューラル・ネットワークに含まれる各第1演算子セットは、その第1演算子セットに対応する第3演算子セットで置き換えられる。以下の表1に示されるように、知識ベース231の例が提供される。

【表1】

表 1

	ターゲットハードウェア プラットフォーム	第1 演算子セット	第3 演算子セット
1	モデルが A である GPU	3×3 畳み込み演算子	3×3 畳み込み演算子、 1×1 畳み込み演算子、および 5×5 畳み込み演算子
2	モデルが B である CPU	ストライド=1 の場合の 7×7 畳み 込み演算子+プーリング層演算 子	ストライド=2 の場合の 7×7 畳み込み演算子
3	モデルが C である NPU	7×7 畳み込み演算子	3×3 畳み込み演算子+ 3×3 畳み込み演算子+ 3×3 畳み込み演算子

【0102】

第1演算子セットは、一つまたは複数の演算子を含んでいてもよいことがわかる。各第1演算子セットは、一つまたは複数の派生第2演算子セットに対応する。前記一つまたは複数の派生第2演算子セットおよび／または第1演算子セットが、第1演算子セットに対応する第3演算子セットを構成する。

【0103】

たとえば、表1において、ソース・ニューラル・ネットワークに含まれる第1演算子セットは、 3×3 畳み込み演算子である。知識ベース231に基づいて、モデルがAであるGPU（ターゲット・ハードウェア・プラットフォーム）では、 1×1 畳み込み演算子および 5×5 畳み込み演算子の両方が、 3×3 畳み込み演算子と同じまたは類似の機能を有することがわかる。この場合、派生第2演算子セットは、 1×1 畳み込み演算子および 5×5 畳み込み演算子を含む。スーパーネットワークにおける対応する第3演算子セットは、第1演算子セット（ 3×3 畳み込み演算子）および派生第2演算子セット（ 1×1 畳み込み演算子および 5×5 畳み込み演算子）を含む。

【0104】

別の例として、表1において、ソース・ニューラル・ネットワークに含まれる第1演算子セットは、ストライド（stride）=1の場合の 7×7 畳み込み演算子+プーリング層演算子を含む。知識ベース231に基づいて、モデルがBであるCPU（ターゲット・ハードウェア・プラットフォーム）では、ストライド=2の場合の 7×7 畳み込み演算子は、ストライド=1の場合の 7×7 畳み込み演算子+プーリング層演算子と同じまたは類似の機能を有することがわかる。この場合、派生第2演算子セットは、ストライド=2のときに 7×7 畳み込み演算子を含む。さらに、派生第2演算子セットの評価結果は、第1演算子セットの評価結果よりも良好であり、第1演算子セットは、派生第2演算子セットと置き換えられてもよい。スーパーネットワークにおける対応する第3演算子セットは、派生第2演算子セット（ 7×7 畳み込み演算子）を含む。

【0105】

別の例として、表1において、ソース・ニューラル・ネットワークに含まれる第1演算子セットは、 7×7 畳み込み演算子を含む。知識ベース231に基づいて、モデルがCであるNPU（ターゲット・ハードウェア・プラットフォーム）では、3つの 3×3 畳み込み演算子が、1つの 7×7 畳み込み演算子と同じまたは類似の機能を有する可能性があることがわかる。この場合、派生第2演算子セットは、 3×3 畳み込み演算子+ 3×3 畳み込み演算子+ 3×3 畳み込み演算子を含み、派生第2演算子セットの結果は、第1演算子セットの結果よりも良好である。スーパーネットワークにおける対応する第3演算子セットは、派生第2演算子セット（ 3×3 畳み込み演算子+ 3×3 畳み込み演算子+ 3×3 畳み込み演算子）を含む。

【0106】

表1における演算子間の「+」は、隣接するレベルの演算子セットが組み合わせられること、すなわち、垂直に分布した演算子セットが組み合わせられることを示すために使用されることに留意されたい。

【0107】

知識ベース231が確立されると、第1演算子セットと派生第2演算子セットとの間の対応は、異なるハードウェア・プラットフォーム上で実際に実行される演算子セットのデータ、または、専門家の事前知識に基づいて予測される、演算子セットの実行状態に基づいて決定されうること留意されたい。むろん、知識ベースは、代替的に、異なるハードウェア・プラットフォームのアップグレード状態に基づいて更新されてもよい。知識ベース231を確立するための方法は、本願のこの実施形態において特に限定されない。

【0108】

実際のアプリケーションでは、データベース230内に維持されたトレーニング・データは、全部がデータ収集装置240によって収集されるわけではなく、または別の装置から受領および取得されてもよいことに留意されたい。たとえば、端末装置220によって送信され、知識ベース231を拡張するために使用される情報が受領される。

【0109】

構築モジュール212は、スーパーネットワーク201を構築するように構成される。具体的には、構築モジュール212は、探索条件セット内のソース・ニューラル・ネットワークのネットワーク構造情報に基づいて、データベース230から得られた知識ベース231を使用して、ソース・ニューラル・ネットワークから拡張されたスーパーネットワーク201を構築する。

【0110】

トレーニング・モジュール213は、スーパーネットワーク201をトレーニングして、トレーニングされたスーパーネットワーク202を得るように構成される。具体的には、構築モジュール212によって送信されたスーパーネットワーク201が受領され、該スーパーネットワーク201は、データベース230から得られたトレーニング・データセットに基づいてトレーニングされる。トレーニングされたスーパーネットワーク202は、ニューラル・ネットワーク探索プロセスを実行するために探索モジュール215に送られる。

【0111】

評価結果取得モジュール214は、一つまたは複数の評価メトリックに基づいて、ターゲット・ハードウェア・プラットフォーム上の第1演算子セットおよび／または派生第2演算子セットの評価結果を取得し、ターゲット・ハードウェア・プラットフォーム上の、探索モジュール215によって発見された候補サブネットの評価結果を得るように構成される。

【0112】

探索モジュール215は、ターゲット・ニューラル・ネットワーク203を探索するように構成される。具体的には、トレーニング・モジュール213によって送信されたトレーニングされたスーパーネットワーク202が受領され、トレーニングされたスーパーネットワーク202は、評価結果取得モジュール214により得られたニューラル・ネットワーク評価結果に基づき、ターゲット・ニューラル・ネットワーク203を求めて探索される。さらに、ターゲット・ニューラル・ネットワーク203は、I/Oインターフェース211を通じて端末装置220に出力される。

【0113】

さらに、図2は、本願のこの実施形態によるシステムアーキテクチャの概略図に過ぎないことに留意されたい。図に示した装置、コンポーネント、モジュール等の位置関係は、いかなる限定も構成しない。たとえば、図2において、データベース230は、データ処理装置210に対する外部メモリであり、別の場合には、データベース230は、代替的に、データ処理装置210内に配置されてもよい。

【0114】

本願の実施形態は、ニューラル・ネットワークの大規模な適用に関するもので、理解を容易にするために、以下は、本願の実施形態において使用されうるニューラル・ネットワークに関連する用語および概念を説明する。

【0115】

(1) ニューラル・ネットワーク

【0116】

ニューラル・ネットワークは、ニューロンを含んでいてもよい。ニューロンは、入力として x_s および1の切片を使用する操作ユニットであってもよい。該操作ユニットの出力は以下のようなものでありうる：

【数1】

$$h_{w,b}(x) = f(W^T x) = f\left(\sum_{s=1}^n W_s x_s + b\right)$$

$s=1, 2, \dots$, または n であり、 n は 1 より大きい自然数であり、 W_s は x_s の重みを示し、 b はニューロンのバイアスを示し、 f はニューロンの活性化関数 (activation function) を示す。活性化関数は、ニューロンにおける入力信号を出力信号に変換するために、ニューラル・ネットワークに非線形特性を導入するために使用される。活性化関数の出力信号は、次の畳み込み層への入力として使用されてもよく、活性化関数は、シグモイド関数であってもよい。ニューラル・ネットワークは、複数の単一ニューロンを一緒に接続することによって構成されるネットワークである。具体的には、ニューロンの出力は別のニューロンへの入力であってもよい。各ニューロンへの入力は、前の層の局所的な受容野 (receptive field) に接続されて、局所的な受容野の特徴を抽出してもよい。局所的な受容野は、いく

つかのニューロンを含む領域でありうる。

【0117】

(2) 深層ニューラル・ネットワーク

【0118】

深層ニューラル・ネットワーク (deep neural network、DNN) は、多層ニューラル・ネットワークとも呼ばれ、複数の隠れ層を有するニューラル・ネットワークとして理解することができる。DNNは異なる層の位置に基づいて分割される。DNN内のニューラル・ネットワークは、入力層、隠れ層、出力層の3つのタイプに分類される。一般に、最初の層が入力層であり、最後の層が出力層であり、中間の層が隠れ層である。層は完全に接続される。具体的には、i番目の層の任意のニューロンは、必ず (i+1) 番目の層の任意のニューロンに接続される。

10

【0119】

DNNは複雑に思えるが、各層の作業は実際には複雑ではない。それは、次の線形関係式で簡単に表すことができる。

【数2】

$$\vec{y} = \alpha(W\vec{x} + \vec{b})$$

ここで、ベクトル x (→付きの x) は入力ベクトルを表し、ベクトル y は出力ベクトルを表し、ベクトル b はバイアス・ベクトルを表し、 W は重み行列 (係数とも呼ばれる) を表し、 $\alpha()$ は活性化関数を表す。各層では、出力ベクトル (ベクトル y) を得るために、そのような単純な操作のみが入力ベクトル (ベクトル x) に対して実行される。DNN層が多量に存在するため、係数およびバイアス・ベクトル (ベクトル b) も多量に存在する。これらのパラメータは、DNNでは以下のように定義される：係数 W が例として使用される。3層DNNでは、第2層の第4ニューロンから第3層の第2ニューロンへの線形係数は W_{24}^3 と定義されたとする。上付き添え字3は、その係数が位置する層の順序番号を示し、下付き添え字は、出力についての第3層のインデックス2および入力についての第2層のインデックス4に対応する。

20

【0120】

結論として、第 $(L-1)$ 層の第 k ニューロンから第 L 層の第 j ニューロンへの係数は W_{jk}^L として定義される。

30

【0121】

なお、入力層にはパラメータ W がない。深層ニューラル・ネットワークでは、より多くの隠れ層があれば、ネットワークは、現実の世界における複雑なケースをよりよく記述することができる。理論的には、より多くのパラメータを有するモデルは、より高い複雑さとより大きな「容量」を有する。これは、モデルがより複雑な学習タスクを完了できることを示している。深層ニューラル・ネットワークのトレーニングは、重み行列を学習するプロセスであり、トレーニングの最終目的は、トレーニングされた深層ニューラル・ネットワークのすべての層の重み行列 (多数の層のベクトルによって形成される重み行列 W) を得ることである。

40

【0122】

(3) 損失関数

【0123】

深層ニューラル・ネットワークのトレーニングのプロセスでは、深層ニューラル・ネットワークの出力が実際に予測されると期待される値に可能な限り近いことが期待されるので、現在のネットワークの予測された値と実際に期待されるターゲット値とが比較されてもよく、両者の間の差に基づいてニューラル・ネットワークの各層の重みベクトルが更新される (むろん、通例は最初の更新の前に初期化プロセスがある、具体的には、深層ニューラル・ネットワーク内の各層についてパラメータが事前構成される)。たとえば、ネットワークの予測値がより高い場合、重みベクトルは、より低い予測値を得るために調整さ

50

れる。重みベクトルは、深層ニューラル・ネットワークが、実際に予期されるターゲット値または実際に予期されるターゲット値に非常に近い値を予測できるようになるまで、連続的に調整される。よって、「比較により予測値とターゲット値との間の差をどのように得るか」があらかじめ定義される必要がある。これは損失関数 (loss function) または目的関数 (objective function) である。損失関数と目的関数は、予測値とターゲット値との間の差を測定する重要な式である。損失関数を例に用いる。損失関数の出力値 (損失) が大きいほど、差が大きいことを示す。よって、深層ニューラル・ネットワークのトレーニングは、可能な限り損失を低減するプロセスとなる。

【0124】

(4) 逆伝搬アルゴリズム

【0125】

トレーニング・プロセスにおいて、ニューラル・ネットワークは、ニューラル・ネットワーク・モデルの再構成誤差損失がますます小さくなっていくように、誤差逆伝搬 (back propagation、BP) アルゴリズムを使用することにより、初期ニューラル・ネットワーク・モデルのパラメータ値を補正することができる。具体的には、入力信号は、出力中に誤差損失が発生するまで順方向転送され、初期ニューラル・ネットワーク・モデルのパラメータは、逆伝搬誤差損失情報に基づいて更新され、誤差損失が低減される。逆伝搬アルゴリズムは、主に誤差損失に依存する逆伝搬運動であり、最適なニューラル・ネットワーク・モデルのパラメータ、たとえば重み行列を得ることを目的とする。

【0126】

本願の実施形態において提供される技術的解決策において、演算子および／または演算子セットの一部は、畳み込みニューラル・ネットワーク・モデルの特定の構造に基づいて、かつソース・ニューラル・ネットワークの隠れ層に含まれる諸層における演算子および／または演算子セットの機能および評価結果と組み合わせ、保持または置換されて、ターゲット・ハードウェア・プラットフォームにより適合するターゲット・ニューラル・ネットワークを得ることができる。

【0127】

図3は、本願のある実施形態によるチップのハードウェア構造を示す。チップは、ニューラル・ネットワーク処理ユニット (neural network processing unit、NPU) 300を含む。チップは、図2に示されるデータ処理装置210内に配置されてもよく、図2の構築モジュール212の作業の全部もしくは一部を完了する (たとえば、スーパーネットワーク201を構築する) ように構成され、または、トレーニング・モジュール213の作業の全部もしくは一部を完了する (たとえば、スーパーネットワーク201をトレーニングする) ように構成されてもよく、または探索モジュール215の作業の全部もしくは一部を完了する (たとえば、ターゲット・ニューラル・ネットワーク203を探索する) ように構成されてもよく、または評価結果取得モジュール214の作業の全部もしくは一部を完了する (たとえば、前記一つまたは複数の評価メトリックに基づいて、探索されたニューラル・ネットワークを評価する) ように構成されてもよい。

【0128】

ニューラル・ネットワーク処理ユニットNPU 300は、コプロセッサのはたらきをし、ホスト中央処理ユニット (central processing unit、CPU) (ホストCPU) 上に配置されてもよい。ホストCPUはタスクを割り当てる。NPU 300のコア部分は、動作回路303であり、コントローラ304は、動作回路303を制御して、メモリ (重みメモリまたは入力メモリ) 内のデータを抽出し、演算を実行する。

【0129】

いくつかの実装では、動作回路303は、複数の処理ユニット (プロセス・エンジン、PE) を内部的に含む。いくつかの実装では、動作回路303は、二次元シストリックアレイである。あるいはまた、動作回路303は、一次元シストリックアレイまたは乗算および加算のような数学的演算を実行できる別の電子回路であってもよい。いくつかの実装では、動作回路303は、汎用マトリクスプロセッサである。

10

20

30

40

50

【0130】

たとえば、入力行列A、重み行列B、出力行列Cがあるとする。動作回路303は、重みメモリ302から、行列Bに対応するデータを取り出し、動作回路303の各PE内に該データをバッファする。動作回路303は、アキュムレータ（accumulator）308に、入力メモリ301から取り出された行列Aと行列Bのデータに対して行列演算を行うことにより得られる行列結果または最終行列結果の一部を格納する。

【0131】

ベクトル計算ユニット307は、動作回路303の出力に対して、ベクトル乗算、ベクトル加算、指数演算、対数演算、または値比較などのさらなる処理を実行することができる。たとえば、ベクトル計算ユニット307は、ニューラル・ネットワークの非畳み込み/非FC層において、プーリング（pooling）、バッチ正規化（batch normalization）、またはローカル応答正規化（local response normalization）などのネットワーク計算を実行するように構成されてもよい。

10

【0132】

いくつかの実装では、ベクトル計算ユニット307は、処理された出力ベクトルを統一メモリ306に記憶することができる。たとえば、ベクトル計算ユニット307は、非線形関数を動作回路303の出力に適用してもよい。たとえば、非線形関数は、累積された値のベクトルに適用され、活性化値を生成する。

【0133】

いくつかの実装では、ベクトル計算ユニット307は、正規化された値、組み合わせられた値、またはその両方を生成する。

20

【0134】

いくつかの実装では、処理された出力ベクトルは、動作回路303への活性化入力として使用されることができ、たとえば、処理された出力ベクトルは、ニューラル・ネットワークの後続の層で使用されることができ。

【0135】

統一メモリ306は、入力データおよび出力データを記憶するように構成される。重みデータについては、記憶ユニット・アクセス・コントローラ（direct memory access controller [直接メモリ・アクセス・コントローラ]、DMAC）305は、外部メモリ内の入力データを入力メモリ301および／または統一メモリ306に転送し、外部メモリ内の重みデータを重みメモリ302に記憶し、統一メモリ306内のデータを外部メモリに記憶する。

30

【0136】

バス・インターフェース・ユニット（bus interface unit、BIU）310は、バスを通じて、ホストCPU、DMAC、および命令フェッチ・バッファ309の間の対話を実装するように構成される。

【0137】

コントローラ304に接続された命令フェッチ・バッファ（instruction fetch buffer）309は、コントローラ304によって使用される命令を記憶するように構成される。コントローラ304は、命令フェッチ・バッファ309内にバッファされた命令を呼び出し、動作アクセラレータの作業プロセスを制御するように構成される。

40

【0138】

一般に、統一メモリ306、入力メモリ301、重みメモリ302、および命令フェッチ・バッファ309はそれぞれ、オンチップ（On-Chip）メモリである。前記外部メモリはNPUの外部のメモリである。外部メモリは、二重データレート同期ダイナミックランダムアクセスメモリ（double data rate synchronous dynamic random access memory、DDR SDRAM）、高帯域幅メモリ（high bandwidth memory、HBM）、または他の読み取り可能で書き込み可能なメモリであってもよい。

【0139】

現在、従来技術におけるニューラル・ネットワーク探索方法では、ソース・ニューラル

50

・ネットワークを圧縮し、ターゲット・ニューラル・ネットワークのネットワーク構造情報を得るために、ニューラル・ネットワーク・モデルは一般に、毎秒浮動小数点演算 (floating-point operations per second、FLOPS) に基づいて探索される。FLOPSは、通例、ハードウェア・プラットフォームの実行効率を推定するために使用される。しかしながら、異なるハードウェア・プラットフォームは、異なる演算子について異なる実装および最適化を有する。よって、FLOPSは実際の時間オーバーヘッド（すなわち、時間消費）を常に反映することはできない。

【0140】

たとえば、演算子Aと演算子Bは、同じ機能を実装できるものとする。言い換えると、ニューラル・ネットワークは、演算子Aを用いて演算を実行してもよく、あるいは演算子Bを用いて演算を実行してもよい。演算子Aは 1×1 畳み込みアルゴリズムであり、1000万回の乗算を含み、演算子Bは 3×3 畳み込みアルゴリズムであり、1500万回の乗算を含むとする。演算子AのFLOPSは、演算子BのFLOPSよりも小さく、演算子Aによって消費される時間は、一般に、演算子Bによって消費される時間よりも短いと推定される。

【0141】

実際の応用では、次の場合が存在しうる：演算子Aまたは演算子Bがハードウェア・プラットフォーム1上で動作するとき、演算子Aは実際に $2 \mu s$ を消費し、演算子Bは実際に $3 \mu s$ を消費する。この場合、ニューラル・ネットワークがハードウェア・プラットフォーム1上に配備される場合、主に時間消費を考慮するならば、演算子Aを使用することが好ましい。演算子Aまたは演算子Bがハードウェア・プラットフォーム2上で動作する場合、ハードウェア・プラットフォーム2は演算子Bに対して特定の最適化を行うため、演算子Bが消費する実際の時間は $1.5 \mu s$ であるが、演算子Aが消費する実際の時間はなお $2 \mu s$ である。この場合、ニューラル・ネットワークがハードウェア・プラットフォーム2上に配備される場合、主に時間消費を考慮するならば、演算子Bを使用することが好ましい。

【0142】

よって、ニューラル・ネットワーク・モデルをFLOPSに基づいて探索する場合、ターゲット・ハードウェア・プラットフォーム上のターゲット・ニューラル・ネットワークの得られた評価結果（たとえば、時間オーバーヘッドの評価結果）は、ターゲット・ハードウェア・プラットフォーム上のソース・ニューラル・ネットワークの評価結果よりも必ずしも良好ではない。

【0143】

これに鑑み、本願のある実施形態は、ニューラル・ネットワークのための適応的探索方法を提供する。ニューラル・ネットワークの適応的探索が、ソース・ニューラル・ネットワークから拡張されたスーパーネットワーク上で、ターゲット・ハードウェア・プラットフォームの評価メトリックに基づいて実行されうる。それにより、ターゲット・ハードウェア・プラットフォームに適應するターゲット・ニューラル・ネットワークが高い探索柔軟性を前提として見出されることができ、ターゲット・ハードウェア・プラットフォーム上で実行される見出されたターゲット・ニューラル・ネットワークの評価結果は、ターゲット・ハードウェア・プラットフォーム上で実行されるソース・ニューラル・ネットワークの評価結果よりもよい。

【0144】

以下は、添付の図面を参照しながら、本願のこの実施形態におけるニューラル・ネットワークの適応的探索方法を詳細に説明する。本願のこの実施形態におけるニューラル・ネットワークの適応的探索方法は、図1のデータ処理装置または図2のデータ処理装置210のような装置によって実行されてもよい。

【0145】

図4は、本願のある実施形態による、ニューラル・ネットワークの適応的探索方法の概略フローチャートである。この方法は、S101～S103を含んでいてもよい。

【0146】

S101：探索条件セットを受領する。

【0147】

上述したように、端末装置のハードウェア・プラットフォーム上で動作する（または担持される）ニューラル・ネットワーク（すなわち、ソース・ニューラル・ネットワーク）が最適化される必要がある、またはニューラル・ネットワーク（すなわち、ソース・ニューラル・ネットワーク）が動作する（または担持される）、端末装置のハードウェア・プラットフォームが別のハードウェア・プラットフォームで置き換えられる場合、ソース・ニューラル・ネットワークのネットワーク構造が、ターゲット・ハードウェア・プラットフォームに基づいて最適化される必要がある。一般に、ソース・ニューラル・ネットワークをスーパーネットワークに拡張するために、まず、ソース・ニューラル・ネットワークのネットワーク構造情報に基づいて、水平拡張および／または垂直拡張が実行される。次いで、一つまたは複数の評価メトリックに基づいて、ターゲット・ハードウェア・プラットフォームに適したニューラル・ネットワークが、スーパーネットワークから、ターゲット・ニューラル・ネットワークとして探索される。ターゲット・ニューラル・ネットワークは、ターゲット・ハードウェア・プラットフォーム上で実行されてもよいし、または担持されてもよい。さらに、ターゲット・ハードウェア・プラットフォーム上でのターゲット・ニューラル・ネットワークの評価結果は、ターゲット・ハードウェア・プラットフォーム上で実行されるソース・ニューラル・ネットワークの評価結果よりも良好である。

10

【0148】

まず、ユーザーは、探索条件セットを入力する必要があり、探索条件セットは、ターゲット・ハードウェア・プラットフォーム情報、ソース・ニューラル・ネットワークのネットワーク構造情報、および前記一つまたは複数の評価メトリックを含む。

20

【0149】

たとえば、ターゲット・ハードウェア・プラットフォーム情報は、ターゲット・ハードウェア・プラットフォームの属性、ターゲット・ハードウェア・プラットフォームのモデル、ターゲット・ハードウェア・プラットフォームの製造者、およびターゲット・ハードウェア・プラットフォームの実行されるべきタスクなどの情報を含んでもよい。このようにして、ターゲット・ハードウェア・プラットフォーム情報に基づいて、現在のターゲット・ハードウェア・プラットフォームにより適したターゲット・ニューラル・ネットワークを見つけることができる。同じ属性であるが異なる製造者を有するハードウェア・プラットフォーム、同じ製造者であるが異なる属性を有するハードウェア・プラットフォーム計、または同じ属性であるが異なるモデルを有するハードウェア・プラットフォームのハードウェア設計は異なる可能性があり、同じ演算子または同じ演算子セットを実行することの利点および欠点も異なる可能性がある。よって、異なるハードウェア・プラットフォームが、探索プロセスにおいて考慮される必要がある。さらに、ターゲット・ハードウェア・プラットフォーム上で実行されるターゲット・ニューラル・ネットワークに対応するタスクが異なる場合、異なる評価メトリックが考慮されてもよい。よって、探索プロセスにおいて、ハードウェア・プラットフォーム上で実行されるターゲット・ニューラル・ネットワークに対応するタスクも考慮される必要がある。たとえば、画像検出（detection）タスクは、画像内のオブジェクトを検出し、ボックス内の該オブジェクトの位置を選択することである。画像セグメンテーション（segmentation）タスクは、検出タスクに基づいてバウンディングボックスを洗練し、精密な境界を用いて、認識されたオブジェクトと画像をセグメンテーションすることである。画像セグメンテーション・タスクの精度評価メトリックは、画像検出タスクの精度評価メトリックよりも高いことが明らかである。

30

40

【0150】

端末装置のハードウェア・プラットフォーム上のニューラル・ネットワークが最適化されるシナリオにおいては、ハードウェア・プラットフォームがターゲット・ハードウェア・プラットフォームであることが理解されうる。端末装置のハードウェア・プラットフォームが交換されるシナリオでは、交換後のハードウェア・プラットフォームがターゲット・ハードウェア・プラットフォームとなる。

【0151】

50

ソース・ニューラル・ネットワークのネットワーク構造情報は、ソース・ニューラル・ネットワークを表現するために使用される。たとえば、ソース・ニューラル・ネットワークのネットワーク構造情報は、ソース・ニューラル・ネットワークに含まれる第1演算子セットを含んでいてもよい。各第1演算子セットは、一つまたは複数の演算子を含むことができる。本願のこの実施形態におけるソース・ニューラル・ネットワークは、畳み込みニューラル・ネットワーク (convolutional neural network、CNN)、リカレントニューラル・ネットワーク (recurrent neural network、RNN)、パーセプトロン (perceptron)、フィードフォワードニューラル・ネットワーク (feedforward neural network、FNN)、深層信頼ネットワーク (deep belief network、DBN)、生成的敵対ネットワーク (generative adversarial network、GAN) などであり得る。

10

【0152】

ソース・ニューラル・ネットワークのネットワーク構造情報が、ソース・ニューラル・ネットワークが畳み込みニューラル・ネットワークである例を用いて記述される。

【0153】

図5の (a) は、畳み込みニューラル・ネットワークのネットワーク構造の概略図である。一般に、畳み込みニューラル・ネットワークは、入力層、隠れ層、および出力層を含む。畳み込みニューラル・ネットワークの隠れ層は、一つまたは複数の畳み込み層 (convolutional layer)、一つまたは複数のプーリング層 (pooling layer)、および一つまたは複数の全結合層 (fully-connected layer) を含む。一般に、前記一つまたは複数の畳み込み層の後に、1つのプーリング層が続く。図5の (a) に示されるように、たとえば、1つの畳み込み層は、一つまたは複数の畳み込み演算子を含んでいてもよい。1つの畳み込み演算子は、1つの第1演算子セットに対応してもよく、または複数の畳み込み演算子が1つの第1演算子セットに対応してもよい。1つのプーリング層は、一つまたは複数のプーリング演算子を含んでいてもよく、一つまたは複数の第1演算子セットに対応してもよい。1つの全結合層は、一つまたは複数の全結合層演算子を含んでいてもよく、一つまたは複数の第1演算子セットに対応しうる。

20

【0154】

別の例として、図5の (b) に示されるように、1つの畳み込み層および1つのプーリング層に含まれる複数の演算子は、1つの第1演算子セットを構成しうる。たとえば、ソース・ニューラル・ネットワークの各層が第1演算子セットに対応し、1つの演算子セットが1つの第1演算子セットを含む場合、ソース・ニューラル・ネットワークのネットワーク構造は、入力層、隠れ層、および出力層を含む、図6の (a) に示されるネットワーク構造として表されてもよい。隠れ層の各レベルは、単一の第1演算子を含む1つの第1演算子セットに対応する。別の例では、ソース・ニューラル・ネットワーク内の複数のレベルが1つの第1演算子セットに対応する場合、ソース・ニューラル・ネットワークのネットワーク構造は、入力層、隠れ層、および出力層を含む、図6の (b) に示されるネットワーク構造として表されてもよい。隠れ層は、2つのブロック構造 (複数の演算子の組み合わせ) を含む。各ブロック構造は、1つの畳み込み層の3つの畳み込み演算子と1つのプーリング層の総和演算子を含む。

30

【0155】

評価メトリックは、ニューラル・ネットワーク・モデルを探索するプロセスにおいて、ターゲット・ニューラル・ネットワークが有する必要があるネットワーク性能に対する要件を含んでいてもよい。前記一つまたは複数の評価メトリックは、ニューラル・ネットワーク・モデルの精度評価メトリック、ニューラル・ネットワーク・モデルの時間オーバーヘッド評価メトリック、ニューラル・ネットワーク・モデルの記憶スペース評価メトリック、ニューラル・ネットワーク・モデルの電力消費評価メトリック、ニューラル・ネットワーク・モデルのテンソル計算ユニットの利用率評価メトリック、およびニューラル・ネットワーク・モデルのメモリ読み出し／書き込み速度評価メトリックのうちの一つまたは複数を含む。

40

【0156】

50

ニューラル・ネットワーク・モデルの精度評価メトリックは、ターゲット・ニューラル・ネットワークの精度が精度閾値以上であることを含んでいる。ニューラル・ネットワーク・モデルの時間オーバーヘッド評価メトリックは、ターゲット・ニューラル・ネットワークの時間オーバーヘッドが時間オーバーヘッド閾値を超えないことを含む。ニューラル・ネットワーク・モデルの記憶スペース評価メトリックは、ターゲット・ニューラル・ネットワークによって占有される記憶スペースが記憶スペース閾値を超えないことを含む。ニューラル・ネットワーク・モデルの電力消費評価メトリックは、ターゲット・ニューラル・ネットワークの電力消費が電力消費閾値を超えないことを含む。ニューラル・ネットワーク・モデルのテンソル計算ユニットの利用率評価メトリックは、ターゲット・ニューラル・ネットワークのテンソル計算ユニットの利用率がテンソル計算ユニットの利用率閾値以上であることを含む。ニューラル・ネットワーク・モデルのメモリ読み出し／書き込み速度評価メトリックは、ターゲット・ニューラル・ネットワークのメモリ読み出し／書き込み速度がメモリ読み出し／書き込み速度閾値以上であることを含む。

【0157】

前記一つまたは複数の評価メトリックは、異なるユーザー要件、異なる適用シナリオ、または異なるタスクなどの要因に基づいて設定されてもよいことに留意されたい。

【0158】

たとえば、見出されたターゲット・ニューラル・ネットワークの精度が、ソース・ニューラル・ネットワークの精度を下回らないこと、またはソース・ニューラル・ネットワークの精度のある区間を下回らないこと（たとえば、ソース・ニューラル・ネットワークの精度の99%を下回らないこと）を保証するために、精度閾値が設定されてもよい。

【0159】

別の例では、見出されるターゲット・ニューラル・ネットワークの処理速度を増加させるために、時間オーバーヘッド閾値がターゲット・ニューラル・ネットワークの時間オーバーヘッド閾値未満に設定されてもよい。一般に、ターゲット・ニューラル・ネットワークのより小さな時間オーバーヘッドは、ターゲット・ニューラル・ネットワークのより速い処理速度を示す。たとえば、時間オーバーヘッド閾値は、ソース・ニューラル・ネットワークの時間オーバーヘッドのある区間内（たとえば、ソース・ニューラル・ネットワークの時間オーバーヘッドの80%以内）に設定されてもよい。

【0160】

たとえば、にぎわいのある道路セクションのモニターにおける画像処理モジュールによって単位時間当たりに収集される監視画像は、大量の情報を含み、大量の操作が実行される必要がある。この場合、画像処理モジュール上で担持されるニューラル・ネットワークは、画像をより迅速に処理するために、より低い時間オーバーヘッドを有する必要がある。このようにして、見出されるターゲット・ニューラル・ネットワークが大量の監視画像をより迅速に処理することができるよう、小さな時間オーバーヘッド閾値が設定される必要がある。別の例では、人口の少ない郊外地域におけるモニターにおける画像処理モジュールによって単位時間当たりに収集される監視画像は、少量の情報を含み、少量の操作が実行される。この場合、画像処理モジュール上で担持されるニューラル・ネットワークは、より高い時間オーバーヘッドを有することが許容され、時間オーバーヘッド閾値は、より高いように設定されうる。

【0161】

別の例では、見出されるターゲット・ニューラル・ネットワークが過度に大きな記憶スペースを占有し、ターゲット・ハードウェア・プラットフォームに記憶されることができないケースを避けるために、ターゲット・ハードウェア・プラットフォームの記憶スペース・サイズに基づいて記憶スペース閾値が設定されてもよい。あるいはまた、ターゲット・ニューラル・ネットワークがターゲット・ハードウェア・プラットフォーム上で動作するときに生じる電力消費を低減し、ターゲット・ハードウェア・プラットフォームを動作させるための全体的な電力消費要件を満たすために、ターゲット・ハードウェア・プラットフォームの電力消費要件に基づいて電力消費閾値が設定されてもよい。代替的に、テン

10

20

30

40

50

ソル計算ユニットの利用率閾値とメモリ読み出し／書き込み速度閾値が、見出されるターゲット・ニューラル・ネットワークとターゲット・ハードウェア・プラットフォームとの間の整合の度合いを測定するために使用される。ターゲット・ハードウェア・プラットフォーム上の見出されるターゲット・ニューラル・ネットワークのテンソル計算ユニットのより高い利用率は、見出されるターゲット・ニューラル・ネットワークとターゲット・ハードウェア・プラットフォームとの間のより高い度合いの整合と、より速い全体的な処理速度とを示す。たとえば、画像処理中に、ターゲット・ハードウェア・プラットフォーム上の見出されるターゲット・ニューラル・ネットワークのテンソル計算ユニットのより高い利用率は、より高速な画像処理速度を示す。

【0162】

別の例では、代替的に、2つ以上の評価メトリックが同時に設定されてもよい。たとえば、精度閾値と時間オーバーヘッド閾値が同時に設定されてもよい。たとえば、いくつかのシナリオでは、大量の画像の認識結果を短時間で得る必要があるが、画像認識精度に対する要求は高くない。この場合、見出されるターゲット・ニューラル・ネットワークの認識速度がソース・ニューラル・ネットワークの認識速度よりも速くなるように、小さな精度閾値および小さな時間オーバーヘッド閾値が設定されてもよい。

【0163】

S102：トレーニング・データセットに基づいて、トレーニングされるべきスーパーネットワーク上でトレーニング・プロセスを実行し、トレーニングされたスーパーネットワークが得られる。ここで、トレーニングされるべきスーパーネットワークは、ソース・ニューラル・ネットワークのネットワーク構造を拡張することによって得られる。

【0164】

たとえば、トレーニングされるべきスーパーネットワークは、ソース・ニューラル・ネットワークのネットワーク構造、知識ベースにおける拡張規則、およびターゲット・ハードウェア・プラットフォーム情報に基づいて、水平拡張モードおよび／または垂直拡張モードで構築される。知識ベースにおける拡張規則は、異なるハードウェア・プラットフォームに基づくソース・ニューラル・ネットワークにおける第1演算子セットと派生第2演算子セットとの間に異なる対応があることを含む。第1演算子セットおよび対応する派生第2演算子セットは、同じまたは類似の機能を有することが理解されうる。

【0165】

具体的には、知識ベースにおける拡張規則に基づいて、ソース・ニューラル・ネットワークに含まれる各第1演算子セットに対応し、ターゲット・ハードウェア・プラットフォーム上で動作する（または担持される）派生第2演算子セットが見出される。たとえば、スーパーネットワークが知識ベースに基づいて拡張される場合、ソース・ニューラル・ネットワーク内の第1演算子セットの一部を保持するか、ソース・ニューラル・ネットワーク内の第1演算子セットの一部を第2演算子セットで置き換えるか、または第1演算子セットの一部に対応するレベルをスキップする（skip）するかを決定することができる。

【0166】

たとえば、ソース・ニューラル・ネットワークは、3つの隣接したレベルを含み、各レベルは、 3×3 畳み込み演算子を含む1つの第1演算子セットを含むとする。換言すれば、該3つのレベルは、 3×3 畳み込み演算子 + 3×3 畳み込み演算子 + 3×3 畳み込み演算子として表されうる。ターゲット・ハードウェア・プラットフォーム上で、 5×5 畳み込み演算子が 3×3 畳み込み演算子と同じまたは類似の機能を実装できる場合、3つの隣接するレベルの 3×3 畳み込み演算子の一部が 5×5 畳み込み演算子で置き換えられ、 3×3 畳み込み演算子の一部がスキップで置き換えられてもよい。このようにして、3つのレベルの全体的な時間オーバーヘッドが低減され、交換後の演算子セットが、ソース・ニューラル・ネットワークにおける対応する第1演算子セットと同じタスクを完了することも保証される。たとえば、派生第2演算子セットは、スキップ + 3×3 畳み込み演算子 + 5×5 畳み込み演算子であってもよい。換言すれば、知識ベースにおいて、 3×3 畳み込み演算子は、ターゲット・ハードウェア・プラットフォームに基づく 5×5 畳み込み演算子に対応しうる。この

10

20

30

40

50

場合、ソース・ニューラル・ネットワークのスーパーネットワークが構築される場合、ソース・ニューラル・ネットワークにおける 3×3 畳み込み演算子が 5×5 畳み込み演算子で置き換えられてもよい。

【0167】

別の例では、ソース・ニューラル・ネットワークは、第1演算子セット ($A+B+C$) を含み、ここで、 A 、 B 、および C は異なる演算子である。従来技術から、3つの演算子セット ($A1+B1+C1$)、($A2+B2+C2$)、($A3+B3+C3$) はすべて ($A+B+C$) と同じまたは類似した機能をもつことがわかる。しかしながら、専門家の事前知識に基づいて、ターゲット・ハードウェア・プラットフォーム上での ($A3+B3+C3$) の実装効果は劣ることがわかった。よって、知識ベースが確立されると、第1演算子セット ($A+B+C$) と第2演算子セット ($A3+B3+C3$) の間の対応は、ターゲット・ハードウェア・プラットフォーム上で確立されない。よって、ソース・ニューラル・ネットワークのスーパーネットワークが構築されるとき、第1演算子セット ($A+B+C$) は、知識ベースに基づいて第2演算子セット ($A1+B1+C1$) および第2演算子セット ($A2+B2+C2$) に置き換えられてもよい。

【0168】

知識ベースにおける拡張規則は、水平拡張および／または垂直拡張を含むことに注意する必要がある。水平拡張は、第1演算子セットのレベルは、対応する派生第2演算子セットのレベルと同じであるが、第1演算子セットは、複数の派生第2演算子セットに対応すると理解されうる。たとえば、図7の (a) に示されるように、 7×7 畳み込み演算子を含む第1演算子セットは、現在レベルでの2つの派生第2演算子セットに対応してもよく、ここが、該2つの派生第2演算子セットは、それぞれ 3×3 畳み込み演算子と 5×5 畳み込み演算子を含む。この場合、対応するレベルの配置態様は、図7の (b) に示される。垂直拡張は、スーパーネットワークが構築されるとき、ソース・ニューラル・ネットワーク内のレベルの一部に対して実行される垂直拡張として理解されてもよい。たとえば、ソース・ニューラル・ネットワークにおける4つの隣接するレベルに含まれる第1演算子セットは、 3×3 畳み込み演算子 (64チャンネル) + 3×3 畳み込み演算子 (64チャンネル) + 3×3 畳み込み演算子 (128チャンネル) + 3×3 畳み込み演算子 (128チャンネル) である。同じ畳み込みカーネル・サイズだが、異なるチャンネル数をもつ演算子セットは直接置き換えることができないため、ソース・ニューラル・ネットワークは、スーパーネットワークを構築するために、全体的な時間オーバーヘッドを低減するよう、垂直に拡張されてもよい。たとえば、前述の4つの隣接するレベルにおける演算子セットは、6つの隣接するレベルにおける演算子セットに垂直に拡張され、たとえば、 3×3 畳み込み演算子 (64チャンネル) + 3×3 畳み込み演算子 (64チャンネル) + 3×3 畳み込み演算子 (128チャンネル) + 3×3 畳み込み演算子 (128チャンネル) に垂直に拡張される。この場合、対応する派生第2演算子セットは、 3×3 畳み込み演算子 (64チャンネル) + 3×3 畳み込み演算子 (64チャンネル) + 3×3 畳み込み演算子 (64チャンネル) + 3×3 畳み込み演算子 (128チャンネル) + スキップ + スキップを含んでいてもよい。

【0169】

以下は、従来技術における拡張方法と本発明の実施形態において提供される方法との違いを説明するために、ソース・ニューラル・ネットワークをスーパーネットワークに拡張する2つの例を与える。

【0170】

例1：ソース・ニューラル・ネットワークは、小さな粒度で演算子 (たとえば、畳み込み演算子、プーリング演算子、および活性化関数) を含むチェーン・ニューラル・ネットワークである。換言すれば、ソース・ニューラル・ネットワークに含まれる第1演算子セットは、1つの第1演算子のみを含む。

【0171】

図7の (a) は、ソース・ニューラル・ネットワークの構造の一例を示す。ソース・ニューラル・ネットワークの構造情報は、複数の第1演算子セットを含み、第1演算子セットは

、それぞれ、 7×7 畳み込み演算子、プーリング層演算子、 3×3 畳み込み演算子、PReLU活性化関数などである。

【0172】

図7の(b)は、従来技術を用いて構築されたスーパーネットワークを示す。第1演算子セットと同一または類似の機能を有するすべての第2演算子セットは、第1演算子セットの位置に対応する第3演算子セットに追加されることがわかる。たとえば、ソース・ニューラル・ネットワークにおける 7×7 畳み込み演算子は、 3×3 畳み込み演算子、 7×7 畳み込み演算子、および 5×5 畳み込み演算子で置き換えられる。ソース・ニューラル・ネットワークにおける 3×3 畳み込み演算子は、 1×1 畳み込み演算子、 3×3 畳み込み演算子、および 5×5 畳み込み演算子で置き換えられる。ソース・ニューラル・ネットワークにおけるPReLU活性化関数は、PReLU活性化層およびReLU活性化層で置き換えられる。このように、スーパーネットワークに基づいてニューラル・ネットワークが探索されるとき、ニューラル・ネットワークの探索空間サイズは $3 \times 3 \times 2 = 18$ となる。

10

【0173】

図7の(c)は、本願の実施形態において提供される方法を使用して構築されたスーパーネットワークを示し、具体的には、ターゲット・ハードウェア・プラットフォームに適用可能な演算子セットは、知識ベースに基づいて決定される。たとえば、ソース・ニューラル・ネットワークにおける 7×7 畳み込み演算子に対応する 3×3 畳み込み演算子および 5×5 畳み込み演算子が、知識ベースにおいて、ターゲット・ハードウェア・プラットフォームに対して貧弱な実行効果を有すると判断される場合、 7×7 畳み込み演算子に対応する第3演算子セットは、 7×7 畳み込み演算子のみを含む。別の例として、ソース・ニューラル・ネットワークにおけるPReLU活性化関数が、知識ベースにおいて、ターゲット・ハードウェア・プラットフォームに対して貧弱な実行効果を有すると判定された場合、PReLU活性化関数に対応する第3演算子セットは、ReLU活性化層のみを含む。この場合、ソース・ニューラル・ネットワークに基づいてスーパーネットワークが構築されるとき、図7の(c)に示されるスーパーネットワークが構築される。スーパーネットワークに基づいてニューラル・ネットワークを探索する場合、ニューラル・ネットワークの探索空間サイズは $1 \times 3 \times 1 = 3$ となる。

20

【0174】

本願のこの実施形態では、ターゲット・ハードウェア・プラットフォームに適さない派生演算子セットは、知識ベースを使用することによって除外されることがわかる。これにより、スーパーネットワークのスケールを縮小し、スーパーネットワークのトレーニング・プロセスを加速することを助ける。さらに、スーパーネットワークのスケールが縮小され、ターゲット・ニューラル・ネットワークの探索範囲が小さいため、探索プロセスをさらに短縮することができ、ターゲット・ニューラル・ネットワークを見出すのがより速くできる。

30

【0175】

例2：ソース・ニューラル・ネットワークは、大きな粒度でブロック構造を含むチェーン・ニューラル・ネットワークである。換言すれば、第1演算子セットは、一つまたは複数のブロック構造を含み、各ブロック構造は、複数の演算子を含む。

40

【0176】

図8の(a)は、ソース・ニューラル・ネットワークの構造の例を示す。ソース・ニューラル・ネットワークは、複数の第1演算子セットを含む。各第1演算子セットは、 1×1 畳み込み演算子、 3×3 畳み込み演算子、 1×1 畳み込み演算子、およびジョイント演算のための総和演算子を含むブロック構造であり、ブロック構造1として表記される。

【0177】

図8の(b)は、本願の実施形態において提供される方法を用いて構築されたスーパーネットワークを示す。たとえば、知識ベースに基づいて、ソース・ニューラル・ネットワークのブロック構造1は、ブロック構造2（たとえば、 1×1 畳み込み演算子、 5×5 畳み込み演算子、 1×1 畳み込み演算子、ジョイント演算のための総和演算子を含むブロック構造）

50

およびブロック構造3（たとえば、 1×1 畳み込み演算子、 7×7 畳み込み演算子、 1×1 畳み込み演算子、ジョイント演算のための総和演算子を含むブロック構造）と同じまたは類似した機能を有し、ターゲット・ハードウェアプラットフォームに適用可能であると判断される。この場合、ソース・ニューラル・ネットワークに基づいてスーパーネットワークが構築されるとき、ブロック構造1はブロック構造2およびブロック構造3で置き換えられてもよく、またはブロック構造1は保持されてもよく、換言すれば、図8の（b）に示すスーパーネットワークが構築される。

【0178】

図9は、ソース・ニューラル・ネットワークおよびターゲット・ハードウェア・プラットフォームに基づいて拡張されたスーパーネットワークの一例を示す図である。拡張されたスーパーネットワークは、ソース・ニューラル・ネットワークと同じ機能を実現することができる。しかしながら、この場合、内部パラメータが調整されないため、精度が低いことがあり、ユーザー要求を満たすことができない。よって、拡張されたスーパーネットワークは、トレーニングされるべきスーパーネットワークである。さらに、トレーニングされるべきスーパーネットワークは、トレーニングされたスーパーネットワークを得るために、トレーニング・データセットを使用することによってトレーニングされてもよい。このように、トレーニングされたスーパーネットワークの精度は精度閾に近く、トレーニングされたスーパーネットワークは、その後、ターゲット・ニューラル・ネットワークを探索するための基礎として使用される。

【0179】

いくつかの例では、トレーニング・データセットが得られた場合、トレーニング・データセットのデータ読み取り態様、たとえば、配置態様およびトレーニング・データセットに含まれる画像の読み取り方法がさらに得られる必要がある。他のいくつかの例では、スーパーネットワークをトレーニングするプロセスにおいて必要とされる、指定されたハイパーパラメータ値、たとえば、トレーニング・プロセスにおいてスーパーネットワークをトレーニングするための、事前設定された逐次反復的なトレーニング回数、トレーニング継続時間、またはターゲット精度閾値をさらに得る必要がある。さらに他のいくつかの例では、トレーニングされたスーパーネットワーク内のターゲット・ニューラル・ネットワークを探索するプロセスにおいて必要とされる指定されたハイパーパラメータ値、たとえば、必要とされる逐次反復回数、または探索プロセスにおいて得られる必要のあるサブネットワーク数をさらに得る必要がある。

【0180】

以下は、本願の実施形態において提供されるスーパーネットワークをトレーニングするための2つの方法を詳細に説明する。

【0181】

方法1：加重平均方法を用いてスーパーネットワークがトレーニングされる。

【0182】

いくつかの例では、図9に示される拡張されたスーパーネットワークは、4つの第3演算子セット、すなわち第3演算子セット1ないし第3演算子セット4を含む。言い換えると、1つの破線ボックスに含まれるすべての演算子セットは、1つの第3演算子セットを構成する。第3演算子セット1は、1つの第1演算子セット1と2つの第2演算子セット（第2演算子セット1と第2演算子セット2）を含む。第1演算子セット1、第2演算子セット1、および第2演算子セット2についての重みがそれぞれ設定され、3つの演算子セットの重みの和は1である。いくつかの例では、第1演算子セット1、第2演算子セット1、および第2演算子セット2の重みは、同じであるように、たとえば、約0.33に設定されてもよい。類推により、別の第3演算子セットにおける第1演算子セットおよび／または第2演算子セットの重みが設定される。

【0183】

次いで、トレーニング・データセットは拡張されたスーパーネットワークに入力されて、該スーパーネットワークをトレーニングする。たとえば、トレーニング・データセット

10

20

30

40

50

は第3演算子セット1に入力され、第3演算子セット1に含まれるすべての演算子セットはトレーニング・データセットの計算に参加し、第3演算子セット1に含まれるすべての演算子セットの計算結果に対して重み付け平均が実行された後に、第3演算子セット1によって出力される特徴マップが得られる。計算後に第3演算子セット1によって出力される特徴マップは、第3演算子セット2に入力され、第3演算子セット2に含まれるすべての演算子セットに対して重み付け平均計算が継続して実行される。類推により、最後の第3演算子セットの計算結果が得られた後、最終的な特徴マップが出力される。

【0184】

次いで、最終的な特徴マップが損失関数に入力され、再び、拡張されたスーパーネットワークに対して逐次反復的なトレーニングが実行される。逐次反復的なトレーニング・プロセスでは、拡張されたスーパーネットワークの内部パラメータが調整される。

10

【0185】

最後に、スーパーネットワークをトレーニングするための逐次反復的なトレーニング回数、トレーニング継続時間、またはターゲット精度閾値が要件を満たした後、トレーニングが停止される。この場合、スーパーネットワークは、トレーニングされたスーパーネットワークであり、トレーニングされたスーパーネットワークにおいてターゲット・ニューラル・ネットワークを探索するプロセスが実行されてもよい。

【0186】

方法2：スーパーネットワークは、ワンホット・サンプリング方法を用いてトレーニングされる。

20

【0187】

いくつかの例では、トレーニングされるべきサブネットを構成するために、ワンホット・サンプリング方法を使用することによって、拡張されたスーパーネットワークにおける各第3演算子セットから、第1演算子セットまたは派生第2演算子セットが得られる。トレーニング・データセットは、トレーニングされるべきサブネットに入力され、該サブネットをトレーニングする。

【0188】

たとえば、図9に示される拡張されたスーパーネットワークは、4つの第3演算子セット、すなわち第3演算子セット1ないし第3演算子セット4を含む。言い換えると、1つの破線ボックスに含まれるすべての演算子セットは、1つの第3演算子セットを構成する。第3演算子セット1は、1つの第1演算子セット1と2つの第2演算子セット（第2演算子セット1と第2演算子セット2）を含む。第3演算子セット2は、1つの第1演算子セット2と2つの第2演算子セット（第2演算子セット3と第2演算子セット4）を含む。たとえば、ワンホット・サンプリング方法は、第1演算子セット1を第3演算子セット1から抽出し、第2演算子セット3を第3演算子セット2から抽出するために使用される。トレーニング・データセットは、計算のために第3演算子セット1における第1演算子セット1に入力され、特徴マップが、計算のために第3演算子セット2における第2演算子セット3に出力される。類推により、最後の第3演算子セットの計算結果が得られた後、最終的な特徴マップが出力され、損失関数に入力される。さらに、逐次反復的なトレーニングはサブネット上で実行され、逐次反復的なトレーニング・プロセスでは、トレーニングされるべきサブネットの内部パラメータが調整される。

30

40

【0189】

前述の方法によれば、ワンホット・サンプリング方法を使ってサンプリングが実行されるたびに、1つのトレーニングされるべきサブネットが、トレーニングのために得られてもよく、トレーニングされるべきサブネットの内部パラメータが調整されてもよい。

【0190】

最後に、スーパーネットワークをトレーニングするための、事前設定された逐次反復的なトレーニング回数、トレーニング継続時間、またはターゲット精度閾値が要件を満たした後、トレーニングが停止される。この場合、スーパーネットワークは、トレーニングされたスーパーネットワークであり、ターゲット・ニューラル・ネットワークを探索するプ

50

ロセスは、トレーニングされたスーパーネットワーク内で実行されてもよい。

【0191】

なお、拡張されたスーパーネットワークは、トレーニングされたスーパーネットワークを得るために、上記の方法のいずれかを用いることによりトレーニングされうることに留意されたい。あるいはまた、拡張されたスーパーネットワークは、まず、方法1を使用してトレーニングされうる。次いで、方法2を用いて、拡張されたスーパーネットワークが保持されて、トレーニングされたスーパーネットワークを得る。

【0192】

S103：前記一つまたは複数の評価メトリックに基づいて、トレーニングされたスーパーネットワーク上でサブネット探索プロセスを実行し、ターゲット・ニューラル・ネットワークのネットワーク構造情報を得る。ターゲット・ニューラル・ネットワークのネットワーク構造情報は、ターゲット・ニューラル・ネットワークを表現するために使用される。

【0193】

いくつかの実施形態では、トレーニングされたスーパーネットワークが得られた後、前記一つまたは複数の評価メトリックに基づいて、ターゲット・ハードウェア・プラットフォーム上のトレーニングされたスーパーネットワークにおける第3演算子セット内の第1演算子セットおよび／または派生第2演算子セットの評価結果が、さらに得られる必要がある。1つの第1演算子セットまたは1つの派生第2演算子セットが、トレーニングされたスーパーネットワーク内の各第3演算子セットから抽出されて、評価されるべき候補サブネットを構成する。複数回の抽出を通じて、複数の評価されるべき候補サブネットが得られる。次いで、一つまたは複数のターゲット・ニューラル・ネットワークが、ターゲット・ハードウェア・プラットフォーム上の評価されるべき候補サブネットに含まれる第1演算子セットおよび／または派生第2演算子セットの、前記一つまたは複数の評価メトリックに基づく評価結果に基づいて決定される。ターゲット・ハードウェア・プラットフォーム上で前記一つまたは複数のターゲット・ニューラル・ネットワークが動作する場合、前記一つまたは複数の評価メトリックに基づく全体的な評価結果は、ソース・ニューラル・ネットワークがターゲット・ハードウェア・プラットフォーム上で動作するときに得られる全体的な評価結果よりもよい。

【0194】

言い換えると、トレーニングされたスーパーネットワークが得られた後、ターゲット・ハードウェア・プラットフォーム、M個のニューラル・ネットワークのネットワーク構造、N個の評価メトリック、およびP個の評価結果の間の対応に関する情報が得られる。M、N、およびPはすべて正の整数であり、 $P = M \times N$ である。M個のニューラル・ネットワークはトレーニングされたスーパーネットワークから抽出されたサブネット、すなわち評価されるべき候補サブネットである。M個のニューラル・ネットワークのネットワーク構造がトレーニングされたスーパーネットワークのネットワーク構造に含まれていることは明らかである。このようにして、トレーニングされたスーパーネットワーク内の第3演算子セットにおける第1演算子セットおよび／または派生第2演算子セットに対応する評価結果（それぞれ、N個の評価メトリックに対応する評価結果）は、前述の対応を使用することによって決定されてもよく、評価されるべき候補サブネットの評価結果は、評価されるべき候補サブネットに含まれる第1演算子セットおよび／または派生第2演算子セットの評価結果を追加することによって得られてもよい。1つの候補サブネットがN個の評価メトリックに基づくN個の評価結果に対応し、M個の候補サブネットがN個の評価メトリックに基づく $M \times N$ 個の評価結果、すなわちP個の評価結果に対応することは明らかである。次いで、一つまたは複数の候補サブネットがP個の評価結果に基づいてターゲット・ニューラル・ネットワークとして決定される。

【0195】

たとえば、N個の評価メトリックがニューラル・ネットワーク・モデルの時間オーバーヘッド評価メトリックを含む場合（すなわち、 $N = 1$ ）、ニューラル・ネットワーク・モデルの時間オーバーヘッド評価メトリックは、ターゲット・ニューラル・ネットワークの時

10

20

30

40

50

間オーバーヘッドが時間オーバーヘッド閾値より大きくないことを含む。

【0196】

対応して、トレーニングされたスーパーネットワークに含まれるM個の候補サブネット上で、ニューラル・ネットワーク・モデルの時間オーバーヘッド評価メトリックに基づいてサブネット評価が実行される。まず、ターゲット・ハードウェア・プラットフォーム上のM個の候補サブネットに含まれる、実行中の第1演算子セットおよび／または派生第2演算子セットの時間オーバーヘッド、すなわち、時間オーバーヘッド・メトリックに基づく評価結果が得られる。次いで、ターゲット・ハードウェア・プラットフォーム上のM個の候補サブネット内の各候補サブネットを実行する時間オーバーヘッド、すなわち、時間オーバーヘッド・メトリックに基づく各候補サブネットの評価結果が、その候補サブネットのネットワーク構造に基づいて決定される。時間オーバーヘッド評価メトリックと、ターゲット・ハードウェア・プラットフォーム上でM個の候補サブネットを実行するための時間オーバーヘッドとに基づいて、M個の候補サブネットから、一つまたは複数の候補サブネットが、ターゲット・ニューラル・ネットワークとして決定される。ターゲット・ニューラル・ネットワークの時間オーバーヘッドは、時間オーバーヘッド閾値より大きくない。

10

【0197】

ターゲット・ハードウェア・プラットフォーム上で各候補サブネットを実行するための時間オーバーヘッドは、ターゲット・ハードウェア・プラットフォーム上で、候補サブネットに含まれる第1演算子セットおよび／または派生第2演算子の組み合わせを実行するための時間オーバーヘッドの合計である。

20

【0198】

別の例として、N個の評価メトリックが、ニューラル・ネットワーク・モデルの精度評価メトリックおよび時間オーバーヘッド評価メトリックを含む場合（すなわち、 $N=2$ ）、ニューラル・ネットワーク・モデルの精度評価メトリックは、ターゲット・ニューラル・ネットワークの精度が精度閾値を下回らないことを含む。ニューラル・ネットワーク・モデルの時間オーバーヘッド評価メトリックは、ターゲット・ニューラル・ネットワークの時間オーバーヘッドが時間オーバーヘッド閾値より大きくないことを含む。

【0199】

対応して、トレーニングされたスーパーネットワークに含まれるM個の候補サブネット上で、ニューラル・ネットワーク・モデルの精度評価メトリックと時間オーバーヘッド評価メトリックに基づいて、サブネット評価が実行される。まず、ターゲット・ハードウェア・プラットフォーム上で、M個の候補サブネットに含まれる第1演算子セットおよび／または派生第2演算子セットを実行することの精度および時間オーバーヘッド、すなわち、精度評価メトリックに基づく制度評価結果および時間オーバーヘッド・メトリックに基づく時間オーバーヘッド評価結果が得られる。次いで、ターゲット・ハードウェア・プラットフォーム上のM個の候補サブネット内の各候補サブネットを実行することの精度および時間オーバーヘッド、すなわち、精度評価メトリックに基づく各候補サブネットの精度評価結果と、時間オーバーヘッド・メトリックに基づく各候補サブネットの時間オーバーヘッド評価結果が、候補サブネットのネットワーク構造に基づいて決定される。精度評価メトリックと、時間オーバーヘッド評価メトリックと、ターゲット・ハードウェア・プラットフォーム上でM個の候補サブネットを実行する精度と、ターゲット・ハードウェア・プラットフォーム上でM個の候補サブネットを実行する時間オーバーヘッドとに基づいて、M個の候補サブネットから、一つまたは複数の候補サブネットが、ターゲット・ニューラル・ネットワークとして決定される。ターゲット・ニューラル・ネットワークの精度は精度閾値以上であり、時間オーバーヘッドは時間オーバーヘッド閾値以下である。

30

40

【0200】

前述の説明では、前記一つまたは複数の評価メトリックに基づいて、ターゲット・ハードウェア・プラットフォーム上で各第1演算子セットまたは各派生第2演算子セットの評価結果を得るために、実際の測定方法または予測方法が使用されてもよい。

【0201】

50

この実施形態のいくつかの例において、トレーニングされたスーパーネットワークは、ターゲット・ニューラル・ネットワークが配備される必要がある端末装置に送られてもよく、トレーニングされたスーパーネットワークは、端末装置のターゲット・ハードウェア・プラットフォーム上で動作し、ターゲット・ハードウェア・プラットフォーム上での各第1演算子セットおよび／または各派生第2演算子セットの評価結果を得る、すなわち、実際に測定された評価結果を得る。

【0202】

たとえば、実際の測定中に、任意の1つの第1演算子セットまたは任意の1つの第2演算子セットが、トレーニングされたスーパーネットワーク内の各第3演算子セットから選択され、サブネットを構成してもよい。複数回の抽出を通して、トレーニングされたスーパーネットワークから複数のサブネットが抽出される。複数のサブネットは、ターゲット・ハードウェア・プラットフォーム上で別々に動作し、ターゲット・ハードウェア・プラットフォーム上での複数のサブネットの評価結果を得る。ターゲット・ハードウェア・プラットフォーム上の複数のサブネットおよび該複数のサブネットに含まれる第1演算子セットおよび／または第2演算子セットの評価結果に基づいて、ターゲット・ハードウェア・プラットフォーム上の第1演算子セットおよび第2演算子セットの評価結果が計算を通じて得られる。

【0203】

この実施形態のいくつかの他の例では、トレーニングされたスーパーネットワーク内の第3演算子セット内の第1演算子セットおよび／または派生第2演算子セットは、評価結果予測モデルに入力されて（予測モデルは、データ処理装置上に配備されてもよく、または別の装置上に配備されてもよい）、第1演算子セットおよび／または派生第2演算子セットの評価結果を直接得てもよく、すなわち、予測された評価結果を得てもよい。予測モデルは、ターゲット・ハードウェア・プラットフォーム上での大量の演算子および／または演算子セットの評価結果に基づいて、トレーニングを通して確立されうる。予測モデルのトレーニング・プロセスについては、従来技術を参照されたい。本願のこの実施形態においては、詳細は記載されていない。

【0204】

代替的に、前述のステップS102でトレーニングされるべきスーパーネットワークが得られた後、前記一つまたは複数の評価メトリックに基づいて、ターゲット・ハードウェア・プラットフォーム上での第1演算子セットおよび／または派生第2演算子セットの評価結果を得る前述のステップが実行されてもよいことに留意されたい。このようにして、スーパーネットワークがトレーニングされた後、ターゲット・ニューラル・ネットワークのネットワーク構造情報は、前記一つまたは複数の評価メトリックに基づく、ターゲット・ハードウェア・プラットフォーム上の第1演算子セットおよび／または派生第2演算子のセットの評価結果を直接使用することによって決定されうる。これは、ターゲット・ニューラル・ネットワークを探索するプロセスを加速するのに役立つ。換言すれば、前記一つまたは複数の評価メトリックに基づいてターゲット・ハードウェア・プラットフォーム上での第1演算子セットおよび／または派生第2演算子セットの評価結果を得る機会は、本願のこの実施形態において限定されない。

【0205】

以下は、前記一つまたは複数の評価メトリックに基づいて、トレーニングされたスーパーネットワーク上でサブネット探索を実行するプロセスについて詳細に説明する。

【0206】

図10は、本願のある実施形態による、ニューラル・ネットワークに対するさらに別の適応的探索方法の概略フローチャートである。この方法は：ソース・ニューラル・ネットワークのネットワーク構造情報、ターゲット・ハードウェア・プラットフォーム情報、一つまたは複数の評価メトリックなどを含む探索条件セットを受領することを含む。スーパーネットワークが構築され、ソース・ニューラル・ネットワークのネットワーク構造情報と知識ベースとに基づいてトレーニングされる。関連する内容については、前述の関連する

10

20

30

40

50

ステップの説明を参照されたい。詳細は、ここでは再度説明しない。

【0207】

いくつかの実施形態では、トレーニングされたスーパーネットワーク上でサブネット探索を実行するプロセスにおいて、知識ベースは、ターゲット・ニューラル・ネットワークの探索範囲を狭めるために使用されてもよい。たとえば、まず、知識ベースに基づいて、候補サブネットの一部のネットワーク構造が、ターゲット・ハードウェア・プラットフォームに対して比較的良好な実装効果を有すると判別され、次いで、それらの候補サブネットが、前記一つまたは複数の評価メトリックに基づく評価のために選択され、それにより、候補サブネットの評価結果を得るための不必要な時間を節約する。

【0208】

いくつかの他の実施形態では、前記一つまたは複数の評価メトリックに基づいて、トレーニングされたスーパーネットワーク上でサブネット探索を実行するプロセスにおいて、スーパーネットワークの全体的な構造または部分構造が、前記一つまたは複数の評価メトリックに基づいて評価されて評価結果が得られてもよく、前記評価結果に基づいて、トレーニングされたスーパーネットワークから一つまたは複数のサブネットが探索されて、該一つまたは複数のサブネットをターゲット・ニューラル・ネットワークとして決定してもよい。あるいはまた、評価結果を得るために、まずスーパーネットワークからいくつかのサブネットが抽出されてもよく、抽出されたいくつかのサブネットが前記一つまたは複数の評価メトリックに基づいて評価される。得られた評価結果に基づいて、前記いくつかのサブネットのうちの一つまたは複数のサブネットがターゲット・ニューラル・ネットワークとして決定される。あるいはまた、まずスーパーネットワークからいくつかのサブネットが抽出され、抽出されたいくつかのサブネットが前記一つまたは複数の評価メトリックに基づいて評価され、評価結果が得られ、次いで、評価結果に基づいて、トレーニングされたスーパーネットワークから新しいサブネットが抽出される。新たに抽出されたサブネットは、評価結果を得るために、前記一つまたは複数の評価メトリックに基づいて評価される。次いで、この場合の評価結果に基づいて、再取得したサブネットのうちの一つまたは複数のサブネットがターゲット・ニューラル・ネットワークのネットワーク構造として決定される。換言すれば、前記一つまたは複数の評価メトリックに基づいて、トレーニングされたスーパーネットワークからサブネットを抽出する方法は、本願のこの実施形態に限定されない。

【0209】

以下は、本願のある実施形態による、トレーニングされたスーパーネットワークにおいてサブネット探索プロセスを実行するための2つの方法を詳細に説明する。

【0210】

方法1：サブネット探索プロセスは、ディファレンシャブル法を用いて実行される。

【0211】

いくつかの例では、重み付け平均法を用いてトレーニングされたスーパーネットワークが得られ、サブネット探索プロセスは、ディファレンシャブル法を用いて、トレーニングされたスーパーネットワークにおいて実行される。

【0212】

たとえば、N個の評価メトリックを入力するとき、ユーザーはN個の評価メトリックの間の一つまたは複数のプロポーション関係も入力する。N個の評価メトリックの間のプロポーション関係は、N個の評価メトリックの間の相対的重要度を反映していることが理解されうる。ターゲット・ハードウェア・プラットフォーム上の見出されたターゲット・ニューラル・ネットワークの評価効果は、より高い比率をもつ評価メトリックに基づいたほうがよいことが明らかである。たとえば、高精度を必要とするターゲット・ハードウェア・プラットフォームに適用可能なターゲット・ニューラル・ネットワークが探索される場合、精度メトリック比率を増加させてもよい。たとえば、評価メトリックが精度メトリックと時間オーバーヘッド・メトリックを含む場合、精度メトリックに対応する比率は、時間オーバーヘッド・メトリックに対応する比率より大きい。たとえば、時間オーバーヘッド・メトリックに対する精度メトリックの比率は6:4である。

【0213】

データ処理装置が、ユーザーによって入力される、N個の評価メトリックの間のr個のプロポーショナル関係を受領する場合、データ処理装置は、各プロポーショナル関係について1回の事前探索を実行する。ここで、rは正の整数である。

【0214】

1つの事前探索プロセスにおいて、重みは、トレーニングされたスーパーネットワーク内の第3演算子セット内の第1演算子セットおよび／または派生第2演算子セットの評価結果について、それぞれ設定される。さらに、トレーニングされたスーパーネットワークは、N個の評価メトリックとN個の評価メトリックの間のプロポーショナル関係とに基づいて再トレーニングされる。トレーニング・プロセスでは、第1演算子セットおよび派生第2演算子セットの評価結果に対応する重みが調整される。評価されるべき候補サブネットおよび該評価されるべき候補サブネットの評価結果は、第1演算子セットおよび派生第2演算子セットの評価結果に対応する調整された重みに基づいて決定される。

10

【0215】

一つまたは複数の評価されるべき候補サブネットの評価結果（単数または複数）は、r個の事前探索プロセスに基づいて決定され、前記一つまたは複数の評価されるべき候補サブネットの評価結果（単数または複数）に基づいて、一つまたは複数のターゲット・ニューラル・ネットワークが決定される。一例では、複数の評価されるべき候補サブネットの評価結果は、降順にソートされてもよく、最初にソートされた一つまたは複数の候補サブネットが、ターゲット・ニューラル・ネットワークとして得られる。

20

【0216】

たとえば、N個の評価メトリックは、精度評価メトリックと時間オーバーヘッド評価メトリックを含むとする（すなわち、 $N=2$ ）。対応して、N個の評価メトリックの間のプロポーショナル関係は、精度評価メトリックと時間オーバーヘッド評価メトリックとの間のプロポーショナル関係を含む。たとえば、ユーザーは、精度評価メトリックと時間オーバーヘッド評価メトリックとの間の第1の比率および第2の比率を入力する。

【0217】

図11(a)は、トレーニングされたスーパーネットワークを示す。対応する重みは、スーパーネットワークにおける各第3演算子セットにおける第1演算子セットおよび／または一つまたは複数の派生第2演算子セットについて設定される。1つの第3演算子セットにおける第1演算子セットおよび／または一つまたは複数の派生第2演算子セットの重みの合計は1である。たとえば、図11(a)に示されるスーパーネットワークは、4つの第3演算子セットを含んでおり、言い換えると、1つの破線ボックスに含まれるすべての演算子セットは、それぞれ第3演算子セット1ないし第3演算子セット4である1つの第3演算子セットを構成する。第3演算子セット1は、1つの第1演算子セット1と2つの第2演算子セット（第2演算子セット1と第2演算子セット2）を含む。第1演算子セット1、第2演算子セット1、および第2演算子セット2についてそれぞれ重みが設定され、3つの演算子セットの重みの合計は1である。いくつかの例では、第1演算子セット1、第2演算子セット1、および第2演算子セット2の重みは、同じであるように、たとえば、約0.33に設定されてもよい。類推により、他の第3演算子セットにおける第1演算子セットおよび／または第2演算子セットの重みが設定される。

30

40

【0218】

トレーニング・プロセスでは、トレーニング・データセットが、トレーニングのために、トレーニングされたスーパーネットワークに入力される。たとえば、トレーニング・データセットは、第3演算子セット1に入力され、第3演算子セット1における第1演算子セットおよび／または派生第2演算子セットは、第3演算子セット1によって出力される特徴マップを得るために、トレーニング・データセットに対して計算を実行する。さらに、N個の評価メトリックに基づく、第3演算子セット1の評価結果がさらに計算される必要がある。評価メトリック（たとえば、時間オーバーヘッド）に基づく、第3演算子セット1の評価結果＝前記評価メトリックに基づく、第1演算子セット1の評価結果 $\times W1$ ＋前記評価メ

50

リックに基づく、第2演算子セット1の評価結果 $\times W2$ + 前記評価メトリックに基づく、第2演算子セット2の評価結果 $\times W3$ である。計算後に第3演算子セット1によって出力される特徴マップ出力と、N個の評価メトリックに基づく、第3演算子セット1の評価結果が、第3演算子セット2に入力され、計算を続行する。同様に、N個の評価メトリックに基づく、第3演算子セット2の評価結果 $= N$ 個の評価メトリックに基づく、第1演算子セット2の評価結果 $\times W4$ + N個の評価メトリックに基づく、第2演算子セット3の評価結果 $\times W5$ + N個の評価メトリックに基づく、第2演算子セット4の評価結果 $\times W6$ である。類推により、最後の第3演算子セットの計算結果が得られた後、最終的な特徴マップとN個の評価メトリックに基づく累積評価結果（すなわち、スーパーネットワークの全体的な評価結果）が出力される。

10

【0219】

次いで、最終的な特徴マップとN個の評価メトリックに基づく累積評価結果を損失関数に入力され、トレーニングされたスーパーネットワークに対して逐次反復的なトレーニングが再び実行される。逐次反復的なトレーニング・プロセスにおいて、各第3セットの一つまたは複数の第1演算子セットおよび／または派生第2演算子セットの評価結果（単数または複数）に対応する重み（単数または複数）は、N個の評価メトリックに基づく累積評価結果に基づいて、逆伝搬プロセスにおいて調整される。トレーニングは、再トレーニングされたスーパーネットワークを得るために、損失関数の結果がもはや変化しなくなるか、または事前設定された逐次反復トレーニング回数に達するまで停止される。最大重みを有する第1演算子セットまたは派生第2演算子セットが、ニューラル・ネットワーク内の各第3演算子から選択され、候補サブネットを構成する。

20

【0220】

たとえば、図11（b）は、前述の方法を用いて逐次反復的なトレーニングが実行され、プロポーション関係1に基づいてトレーニングを停止するための条件を満たす、トレーニングされたスーパーネットワークを示す。この場合のスーパーネットワークについては、最大重みをもつ第1演算子セットまたは派生第2演算子セットが、候補サブネットを構成する各第3演算子セットから抽出される。たとえば、図11（b）では、矢印によって接続された実線ボックス内の、選択された、複数の第2演算子セットおよび第1演算子セットが、1つの候補サブネットを構成し、該候補サブネットは、評価されるべき候補サブネット1として示される。同様に、図11（c）は、前述の方法を用いて逐次反復的なトレーニングが行われ、プロポーション関係2に基づいてトレーニングを停止する条件を満たす、トレーニングされたスーパーネットワークを示す。図11（c）では、矢印によって接続された実線ボックス内の、選択された複数の第2演算子セットおよび第1演算子セットが1つの候補サブネットを構成し、該候補サブネットは、評価されるべき候補サブネット2として示されている。

30

【0221】

評価されるべき候補サブネット1および評価されるべき候補サブネット2から、より良好な評価結果をもつニューラル・ネットワークが、見出されたターゲット・ニューラル・ネットワークとして選択される。

【0222】

方法2：進化法を用いてターゲット・ニューラル・ネットワーク探索プロセスが実行される。

40

【0223】

いくつかの例では、ワンホット・サンプリング法を用いてトレーニングされたスーパーネットワークが得られ、進化法を用いて、トレーニングされたスーパーネットワークにおいてサブネット探索プロセスが実行される。あるいはまた、まず重み付け平均法を用いてトレーニングされ、次にワンホット・サンプリング法を用いてトレーニングされたスーパーネットワークが得られ、進化法を用いて、トレーニングされたスーパーネットワークにおいてサブネット探索プロセスが実行される。

【0224】

50

たとえば、一つまたは複数の評価されるべき候補サブネットは、事前設定規則に基づいて、トレーニングされたスーパーネットワークから抽出され、N個の評価メトリックに基づく、評価されるべき候補サブネットのそれぞれの評価結果は、評価されるべき候補サブネットのそれぞれのネットワーク構造と、第3演算子セット内の第1演算子セットおよび／または第2演算子セットの評価結果とに基づいて決定される。評価されるべき候補サブネットは、最初のK個の評価結果に対応するK個の第1のターゲット候補サブネットを得るために、N個の評価メトリックに基づいて得られた評価結果の降順にソートされる。事前設定規則は、ランダムに抽出されてもよく、または好ましくは、より小さなFLOPSを有する第1演算子セットまたは第2演算子セットを選択してもよい。これは、本願のこの実施形態において限定されない。

10

【0225】

たとえば、図12(a)は、トレーニングされたスーパーネットワークを示しており、たとえば、4つの第3演算子セットを含み、言い換えると、1つの破線ボックスに含まれるすべての演算子セットは、それぞれ第3演算子セット1ないし第3演算子セット4である1つの第3演算子セットを構成する。第3演算子セット1は、第1演算子セット1、第2演算子セット1、および第2演算子セット2を含む。まず、1つの第1演算子セットまたは1つの派生第2演算子セットが、各第3演算子セットからランダムに抽出され、ターゲット候補サブネットを構成する。複数回の抽出を通じて、複数のターゲット候補サブネットが得られる。たとえば、図12(b)に示されるように、実線ボックスを使用して選択され、矢印によって接続される、複数の第2演算子セットおよび1つの第1演算子セットを含むニューラル・ネットワークが、抽出された評価されるべき候補サブネットである。別の例として、図12(c)に示されるように、実線ボックスを使用して選択され、矢印によって接続される、複数の第2演算子セットおよび1つの第1演算子セットとを含むネットワークが、もう一つの抽出された評価されるべきサブネットである。

20

【0226】

次いで、第nラウンドの混成および／または変成の後に得られるS個の第2のターゲット候補サブネットが、トレーニングされたスーパーネットワークおよびK個の第1のターゲット候補サブネットに基づいて得られる。ここで、nは事前設定されたスクリーニング回数であり、 $n \geq 1$ であり、KおよびSは正の整数であり、 $S > K$ である。

【0227】

30

たとえば、図12(b)において、矢印によって接続された第2演算子セットおよび第1演算子セットを含むニューラル・ネットワークは、第1のターゲット候補サブネットであり、第1のターゲット候補サブネット1として示される。図12(c)において、矢印によって接続された第2演算子セットおよび第1演算子セットを含むニューラル・ネットワークは、別の第1のターゲット候補サブネットであり、第1のターゲット候補サブネット2として示されている。さらに、第1のターゲット候補サブネット1と第1のターゲット候補サブネット2は、K個の最初のターゲット候補サブネットにおいて最初にソートされた2つの第1のターゲット候補サブネットである。このように、混成は、第1のターゲット候補サブネット1および第1のターゲット候補サブネット2に対して行われる。たとえば、第1のターゲット候補サブネット1における第3演算子セット1における第2演算子セット1は、第1のターゲット候補サブネット2における第3演算子セット1における第1演算子セット1で置き換えられて、第2のターゲット候補サブネット（第1演算子セット1、第2演算子セット4、第2演算子セット6、および第1演算子セット3を含む）が得られる。

40

【0228】

上記の変成は、第2のターゲット候補サブネットを得るために、K個の第1のターゲット候補サブネットにおける第3演算子セットに含まれる第1演算子セットまたは第2演算子セットを、トレーニングされたスーパーネットワークにおける対応する第3演算子セットにおける任意の第1演算子セットまたは第2演算子セットで置き換えることを意味する。

【0229】

たとえば、図12(b)において、矢印によって接続された第2演算子セットおよび第1演

50

算子セットを含むニューラル・ネットワークは、第1のターゲット候補サブネットであり、第1のターゲット候補サブネット1として示される。さらに、第1のターゲット候補サブネット1は、K個の第1のターゲット候補サブネットにおいて最初にソートする第1のターゲット候補サブネットである。このように、第1のターゲット候補サブネット1は、図12(a)に示されるトレーニングされたスーパーネットワークに基づいて変成される。たとえば、第1のターゲット候補サブネット1における第3演算子セット1における第2演算子セット1は、トレーニングされたスーパーネットワークにおける第3演算子セット1に含まれる第2演算子セット2で置き換えられ、第2のターゲット候補サブネット（第2演算子セット2、第2演算子セット4、第2演算子セット6および第1演算子セット3を含む）を得る。

【0230】

K個の第1のターゲット候補サブネットは、S個の第2のターゲット候補サブネットを得るために混成されてもよいことが理解されうる。あるいはまた、K個の第1のターゲット候補サブネットは、S個の第2のターゲット候補サブネットを得るために変成されてもよい。あるいはまた、K個の第1のターゲット候補サブネットは、S個の第2のターゲット候補サブネットを得るために混成および変成される。混成または変成の量的比率は、本願のこの実施形態において特に限定されない。

【0231】

次いで、S個の第2のターゲット候補サブネットの評価結果が得られ、第2のターゲット候補サブネットのネットワーク構造と、N個の評価メトリックに基づく、第3演算子セットにおける第1演算子セットおよび／または第2演算子セットの評価結果とに基づいて降順にソートされて、最初のK個の評価結果に対応するK個の第2のターゲット候補サブネットが得られる。

【0232】

最後に、逐次反復的なトレーニングの(n+1)ラウンドの後、具体的には、混成および／または変成のnラウンドの後、K×(n+1)個のターゲットサブネットが得られる。K×(n+1)個のターゲット候補サブネットは、前記一つまたは複数の評価メトリックに基づいて再度ソートされて、一つまたは複数のターゲット・ニューラル・ネットワークのネットワーク構造情報が得られてもよい。

【0233】

たとえば、S=10およびK=5である場合、10個の第1のターゲット候補サブネットがトレーニングされたスーパーネットワークから初めて選択され、混成および／または変成の各ラウンドの後に10個のターゲット候補サブネットが得られ、5つの最適なターゲット候補サブネットが、各スクリーニング・トレーニングにおいて保持される。混成および／または変成の後、前ラウンドのスクリーニングにおいて選択されたターゲット候補サブネットが消失しうる。よって、トレーニング・プロセスにおいて、消失するターゲット候補サブネットは、最適なネットワーク性能を有するニューラル・ネットワークでありうる。よって、最後に、各回に得られた評価結果が最初にソートされるすべてのターゲット候補サブネットが再度ソートされ、一つまたは複数のより良いターゲット候補サブネットがターゲット・ニューラル・ネットワークとして選択される。n=2、具体的には、逐次反復的なトレーニングが3回実行され、2ラウンドの混成および／または変成が実行されるとすると、最終的に、ターゲット・ニューラル・ネットワークのネットワーク構造情報は、5×(2+1)=15個のターゲット候補サブネットにおいて決定される。

【0234】

これに鑑み、本願のこの実施形態で提供されるニューラル・ネットワークのための適応的探索方法に従う。ターゲット・ハードウェア・プラットフォームの評価基準に基づいて、ニューラル・ネットワークの適応的探索が、ソース・ニューラル・ネットワークから拡張されたスーパーネットワーク上で実行され、それにより、ターゲット・ハードウェア・プラットフォームに適応するターゲット・ニューラル・ネットワークが高い探索柔軟性を前提として見出され、ターゲット・ハードウェア・プラットフォーム上で実行される見出されるターゲット・ニューラル・ニューラル・ネットワークの評価結果が、ターゲット・

ハードウェア・プラットフォーム上で実行されるソース・ニューラル・ネットワークの評価結果よりもよくなる。

【0235】

いくつかの実施形態では、ステップS103の後、図13に示されるように、ニューラル・ネットワークを求める適応的探索方法は、S201およびS202をさらに含んでもよい。

【0236】

S201：ターゲット・ニューラル・ネットワークのネットワーク構造情報に基づいてターゲット・ニューラル・ネットワークを構築する。

【0237】

たとえば、上述したように、ターゲット・ニューラル・ネットワークを表すために使用されるターゲット・ニューラル・ネットワークのネットワーク構造情報、たとえば、含まれる第1演算子セットおよび／または派生第2演算子セット、第1演算子セットおよび／または派生第2演算子セットの配列態様、およびターゲット・ニューラル・ネットワークの層の数に関する情報が得られる。ターゲット・ニューラル・ネットワークは、ターゲット・ニューラル・ネットワークのネットワーク構造情報に基づいて構築される。

【0238】

S202：トレーニングされたターゲット・ニューラル・ネットワークを得るために、第2のトレーニング・データセットに基づいてターゲット・ニューラル・ネットワーク上でトレーニング・プロセスを実行する。

【0239】

第2のトレーニング・データセットは、ステップS102でスーパーネットワークをトレーニングするために使用されるトレーニング・データセットと同じであっても、異なるものであってもよい。

【0240】

いくつかの実施形態では、構築されたターゲット・ニューラル・ネットワークは、トレーニングされたスーパーネットワークに基づいて見出されるニューラル・ネットワークであり、ソース・ニューラル・ネットワークと同じ機能を実装することができる。さらに、ターゲット・ハードウェア・プラットフォーム上の評価結果は、ターゲット・ハードウェア・プラットフォーム上のソース・ニューラル・ネットワークの評価結果よりも良好である。しかしながら、この場合、ターゲット・ニューラル・ネットワークの内部パラメータが最適状態に調整されていないため、ターゲット・ニューラル・ネットワークの実際の適用プロセスにおいて問題が発生する可能性がある。よって、ターゲット・ニューラル・ネットワークは、第2のトレーニング・データセットを使用することによってトレーニングされる必要があり、ターゲット・タスクを実行するために実際に適用されうるトレーニングされたターゲット・ニューラル・ネットワークを得るために、ターゲット・ニューラル・ネットワーク内のパラメータ情報を調整する必要がある。

【0241】

たとえば、プリセット条件が評価メトリックに基づいて設定されてもよく、ターゲット・ニューラル・ネットワークは、第2のトレーニング・データセットを使用することによって逐次反復的にトレーニングされ、ターゲット・ニューラル・ネットワークの内部パラメータは、逐次反復的トレーニング・プロセスにおいて調整される。事前設定されたトレーニング回数に達したとき、または損失関数が減少しないときに得られるニューラル・ネットワークが、実際にターゲット・ハードウェア・プラットフォームに適用できる、トレーニングされたターゲット・ニューラル・ネットワークとして使用される。

【0242】

たとえば、トレーニングされたターゲット・ニューラル・ネットワークは、ターゲット・ハードウェア・プラットフォーム上で、コンピュータビジョン (computer vision、CV) フィールド内のターゲット・タスク、たとえば、ターゲット検出タスクおよび画像分類タスクを実行することができる。具体的には、ターゲット検出タスクは、たとえば、顔検出、車両窓顔検出、人体／顔検出、画像モード検出、オブジェクト分類、およびナンバ

10

20

30

40

50

ープレート検出を含むことができる。画像分類タスクは、たとえば、交通信号認識、車両モデル認識、および特殊車両モデル認識を含んでいてもよい。

【0243】

このようにして、ターゲット・ハードウェア・プラットフォーム上でターゲット・タスクを実際に行うことができるターゲット・ニューラル・ネットワークを得ることができる。

【0244】

本願のある実施形態は、少なくとも1つのサーバーおよび少なくとも1つの端末装置を含む通信システムを提供する。

【0245】

前記少なくとも1つのサーバーは、図4におけるステップS101～ステップS103を実行し、図13におけるステップS201およびステップS202を実行し、および／または本明細書に記載される技術の別のプロセスのために使用される。前記少なくとも1つの端末装置は、前記少なくとも1つのサーバーに探索条件セットを送信し、ここで、この探索条件セットは、ターゲット・ハードウェア・プラットフォーム情報、ソース・ニューラル・ネットワークのネットワーク構造情報、および一つまたは複数の評価メトリックを含み、ソース・ニューラル・ネットワークのネットワーク構造情報は、ソース・ニューラル・ネットワークを表すために使用される。前記少なくとも1つの端末装置は、前記少なくとも1つのサーバーによって送信されたトレーニングされたターゲット・ニューラル・ネットワークを受信する。

【0246】

任意的に、前記少なくとも1つの端末装置は、トレーニングされたターゲット・ニューラル・ネットワークをターゲット・ハードウェア・プラットフォームを含むユーザー装置に送るかどうかを決定するようにさらに構成される。ユーザー装置はターゲット・ニューラル・ネットワークのために使用され、ユーザー装置は、携帯電話 (mobile phone)、デスクトップ・コンピュータ、ラップトップ・コンピュータ、ハンドヘルド・コンピュータ、ノートブック・コンピュータ、ウルトラモバイルパーソナルコンピュータ (ultra-mobile personal computer、UMPC)、ネットブック、パーソナルデジタルアシスタント (personal digital assistant、PDA)、拡張現実 (augmented reality、AR) / 仮想現実 (virtual reality、VR) 装置、または人工知能 (artificial intelligence、AI) 端末などの端末装置であってもよい。ユーザー装置の特定の形態は、本開示のこの実施形態では特に限定されない。

【0247】

図14は、前述の実施形態におけるニューラル・ネットワークのための適応的探索装置の可能な構造の概略図である。この装置は、受領ユニット1401、構築ユニット1402、トレーニング・ユニット1403、および探索ユニット1404を含む。

【0248】

受領ユニット1401は、図4におけるステップS101を実行するよう本装置をサポートするように構成され、および／または本明細書に記載される技術の別のプロセスのために使用される。

【0249】

構築ユニット1402は、図4におけるステップS102、図13におけるステップS201、および図2における構築モジュール212によって実行されるステップを実行するよう本装置をサポートするように構成され、および／または、本明細書に記載される技術の別のプロセスのために使用される。

【0250】

トレーニング・ユニット1403は、図4におけるステップS102、図13におけるステップS202、および図2におけるトレーニング・モジュール213によって実行されるステップを実行するよう本装置をサポートするように構成され、および／または、本明細書に記載される技術の別のプロセスのために使用される。

【0251】

探索ユニット1404は、図4におけるステップS103および図2における探索モジュール215によって実行されるステップを実行するよう本装置をサポートするように構成され、および／または、本明細書に記載される技術の別のプロセスのために使用される。

【0252】

前述の方法実施形態におけるステップのすべての関連する内容は、対応する機能ユニットの機能説明において引用されうる。詳細は、ここでは再度説明しない。

【0253】

図15は、本願のある実施形態による装置のハードウェア構造の概略図である。本装置は、少なくとも1つのプロセッサ1501、通信線1502、メモリ1503、および少なくとも1つの通信インターフェース1504を含む。あるいはまた、メモリ1503は、プロセッサ1501に含まれてもよい。

【0254】

プロセッサ1501は、汎用の中央処理装置（central processing unit、CPU）、マイクロプロセッサ、特定用途向け集積回路（application-specific integrated circuit、ASIC）、または本願の解決策のプログラム実行を制御するように構成された一つまたは複数の集積回路であってもよい。

【0255】

通信線1502は、前述の構成要素間で情報を転送するための経路を含んでいてもよい。

【0256】

通信インターフェース1504は、他の装置と通信するように構成される。本願の実施形態では、通信インターフェースは、モジュール、回路、バス、インターフェース、トランシーバ、または通信機能を実施することができ、他の装置と通信するように構成される他の装置であってもよい。任意的に、通信インターフェースがトランシーバである場合、トランシーバは独立して配置された送信器であってもよく、送信器は別の装置に情報を送信するように構成されてもよい。あるいはまた、トランシーバは、独立して配置された受信器であってもよく、他の装置から情報を受信するように構成される。あるいはまた、トランシーバは、情報を送信および受信する機能を統合するコンポーネントであってもよい。トランシーバの特定の実装は、本願の実施形態において限定されない。

【0257】

メモリ1503は、静的な情報および命令を記憶することができる読み出し専用メモリ（read-only memory、ROM）または情報および命令を記憶することができるランダムアクセスメモリ（random access memory、RAM）または他のタイプの動的記憶装置であってもよく、または電氣的に消去可能なプログラマブル読み出し専用メモリ（electrically erasable programmable read-only memory、EEPROM）、コンパクトディスク読み出し専用メモリ（compact disc read-only memory、CD-ROM）または他のコンパクトディスク記憶、光ディスク記憶装置（コンパクトディスク、レーザーディスク、光ディスク、デジタル汎用ディスク、ブルーレイディスク等を含む）、磁気ディスク記憶媒体または他の磁気記憶装置、または命令またはデータ構造の形で、期待されるプログラムコードを担持または記憶するために使用されることができ、コンピュータによってアクセスすることができる他の媒体であってもよいが、これらに限定されない。メモリは独立して存在してもよく、通信線202を通じてプロセッサに接続される。あるいはまた、メモリは、プロセッサと一体化されてもよい。

【0258】

メモリ1503は、本願の解決策を実装するためのコンピュータ実行可能命令を記憶するように構成され、プロセッサ1501は実行を制御する。プロセッサ1501は、メモリ1503に記憶されたコンピュータ実行可能命令を実行し、本願の以下の実施形態で提供されるニューラル・ネットワークのための適応的探索方法を実施するように構成される。

【0259】

任意的に、本願の実施形態におけるコンピュータ実行可能命令は、アプリケーション・プログラム・コード、命令、コンピュータ・プログラム、または別の名前とも呼ばれる

10

20

30

40

50

。これは、本願の実施形態において特に限定されない。

【0260】

具体的な実装の際、ある実施形態では、プロセッサ1501は、一つまたは複数のCPU、たとえば、図15のCPU 0およびCPU 1を含んでもよい。

【0261】

具体的な実装の際、ある実施形態では、本装置は、複数のプロセッサ、たとえば、図15のプロセッサ1501およびプロセッサ1505を含んでもよい。各プロセッサは、シングルコア（single-core）プロセッサまたはマルチコア（multi-CPU）プロセッサであってもよい。ここで、プロセッサは、データ（たとえば、コンピュータ・プログラム命令）を処理するように構成された一つまたは複数のデバイス、回路、および／または処理コアであつてもよい。

10

【0262】

なお、上記装置は、汎用デバイスであっても、専用デバイスであってもよい。装置のタイプは、本願のこの実施形態では限定されない。本願のこの実施形態において示される構造は、装置に対する特定の限定を構成しない。本願のいくつかの他の実施形態において、装置は、図に示されたものよりも多くのまたは少ない構成要素を含んでもよく、あるいは、いくつかの構成要素が組み合わされてもよく、あるいは、いくつかの構成要素が分割されてもよく、あるいは、異なる構成要素が使用されてもよい。図に示される構成要素は、ハードウェア、ソフトウェア、またはソフトウェアとハードウェアの組み合わせによって実現することができる。

20

【0263】

本願のある実施形態は、コンピュータ読み取り可能な記憶媒体をさらに提供する。コンピュータ読み取り可能な記憶媒体は、コンピュータ命令を記憶する。コンピュータ命令がサーバー上で実行されると、サーバーは、関連する方法ステップを実行して、前述の諸実施形態においてニューラル・ネットワークのための適応的探索方法を実施することが可能にされる。

【0264】

本願のある実施形態は、さらに、コンピュータ・プログラム・プロダクトを提供する。コンピュータ・プログラム・プロダクトがコンピュータ上で実行されると、コンピュータは、前述の実施形態においてニューラル・ネットワークのための適応的探索方法を実施するために、前述の関連するステップを実行することが可能にされる。

30

【0265】

さらに、本願のある実施形態は、装置をさらに提供する。本装置は、具体的には、コンポーネントまたはモジュールであつてもよく、本装置は、互いに接続されたプロセッサおよびメモリを含んでもよい。メモリは、コンピュータ実行可能命令を格納するように構成される。装置が動作するとき、プロセッサは、メモリに記憶されたコンピュータ実行可能命令を実行してもよく、それにより装置は前述の方法実施形態におけるニューラル・ネットワークの適応的探索方法を実行する。

【0266】

本願の実施形態において提供される装置、コンピュータ読み取り可能な記憶媒体、コンピュータ・プログラム・プロダクト、またはチップは、上記で提供される対応する方法を実行するように構成される。よって、本装置、コンピュータ読み取り可能な記憶媒体、コンピュータ・プログラム・プロダクト、またはチップによって達成できる有益な効果については、上述の対応する方法における有益な効果を参照されたい。詳細は、ここでは再度説明しない。

40

【0267】

実装についての前述の説明は、簡便な説明のために、前述の機能モジュールへの分割が例解のための例として使用されることを、当業者が明確に理解することを許容する。実際の適用の際、前述の機能は、異なる機能モジュールに割り当てられ、要件に基づいて実装されうる。すなわち、装置の内部構造は、異なる機能モジュールに分割され、上述の機能

50

の全部または一部を実装する。前述のシステム、装置、およびユニットの詳細な動作プロセスについては、前述の方法実施形態における対応するプロセスを参照されたい。詳細は、ここでは再度説明しない。

【0268】

本願において提供されるいくつかの実施形態において、開示された方法は、他の態様で実装されることが理解されるべきである。たとえば、モジュールまたはユニットへの分割は、単に論理的な機能分割であり、実際の実装の際には他の分割であってもよい。たとえば、複数のユニットまたはコンポーネントは、別のシステムに組み合わされ、または統合されてもよく、あるいはいくつかの特徴が無視され、または実行されなくてもよい。さらに、表示または議論された相互結合または直接結合または通信接続は、いくつかのインターフェースを通じて実装されてもよい。モジュールまたはユニットの間の間接的結合または通信接続は、電子的、機械的、または他の形で実装されうる。

10

【0269】

別個の部品として記載されるユニットは、物理的に別個であってもなくてもよく、ユニットとして表示される部品は、物理的なユニットであってもなくてもよく、1つの位置に位置していてもよく、または複数のネットワーク・ユニット上に分散されていてもよい。ユニットの一部または全ては、実施形態における解決策の目的を達成するために、実際の要件に基づいて選択されてもよい。

【0270】

さらに、本願の実施形態における機能ユニットは、1つの処理ユニットに統合されてもよく、または各ユニットは、物理的に単独で存在してもよく、または2つ以上のユニットが1つのユニットに統合されてもよい。統合されたユニットは、ハードウェアの形で実装されてもよく、またはソフトウェア機能ユニットの形で実装されてもよい。

20

【0271】

統合されたユニットがソフトウェア機能ユニットの形で実装され、独立した製品として販売または使用される場合、統合ユニットは、コンピュータ読み取り可能な記憶媒体に記憶されうる。そのような理解に基づいて、本願の技術的解決策は本質的に、または従来技術に寄与する部分は、または技術的解決策の全部もしくは一部は、ソフトウェア製品の形で実装されてもよい。コンピュータソフトウェア製品は、記憶媒体に記憶され、コンピュータ装置（これは、パーソナルコンピュータ、サーバー、ネットワーク装置等であってもよい）に、本願の実施形態における方法の全部または一部のステップを実行するように命令するためのいくつかの命令を含む。前述の記憶媒体は、プログラム命令を記憶することができる任意の媒体、たとえば、フラッシュメモリ、リムーバブルハードディスク、リードオンリーメモリ、ランダムアクセスメモリ、磁気ディスク、または光ディスクを含む。

30

【0272】

前述の説明は、単に本願の具体的な実装であり、本願の保護範囲を限定することを意図するものではない。本願に開示された技術的範囲内の変成または置換は、本願の保護範囲にはいる。したがって、本願の保護範囲は、特許請求の範囲の保護範囲に従うものとする。

40

【図面】
【図 1】

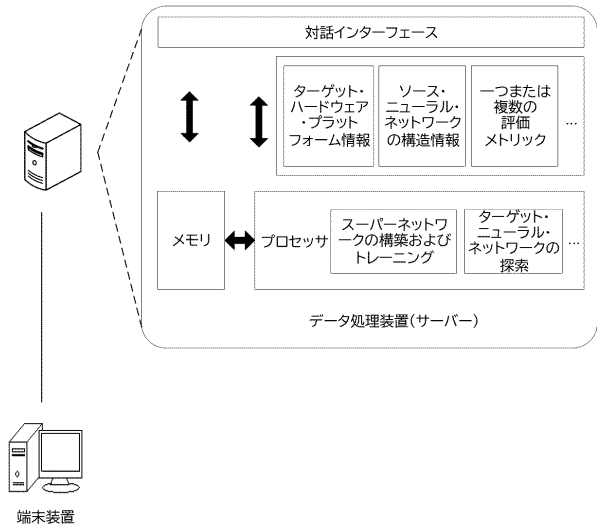


FIG. 1

【図 2】

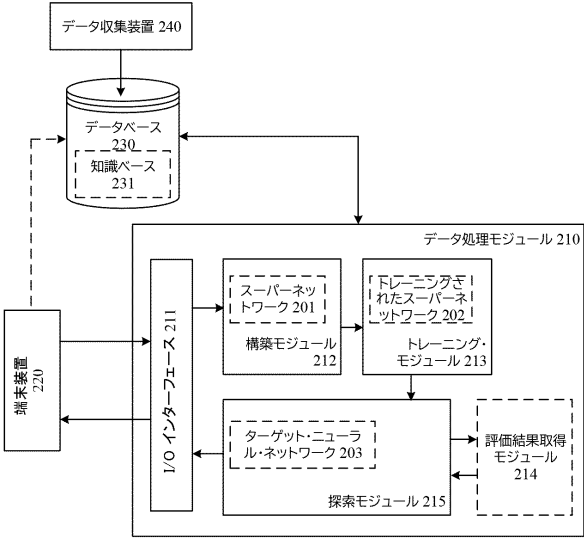


FIG. 2

【図 3】

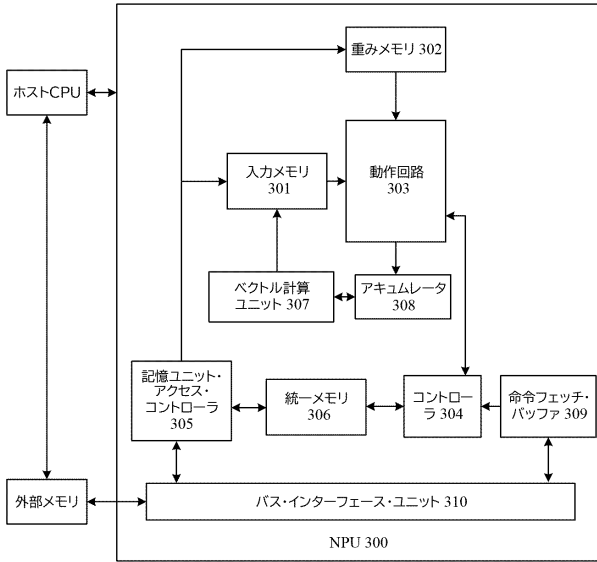


FIG. 3

【図 4】

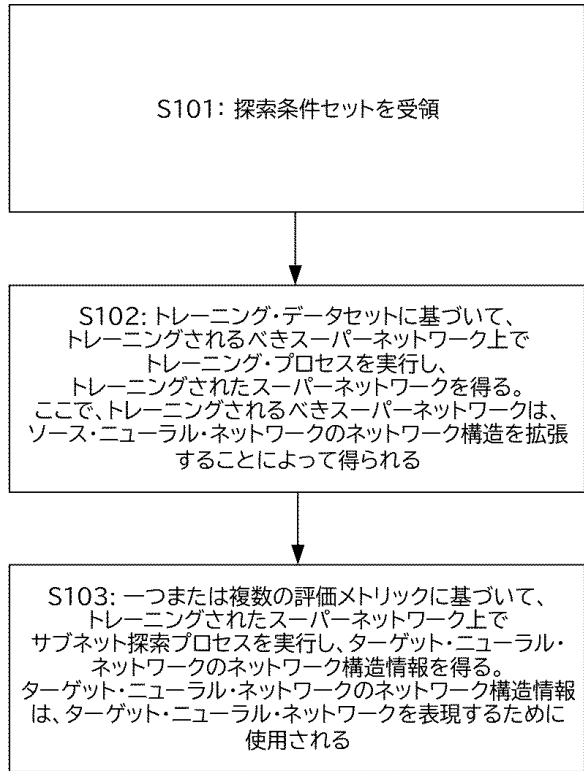
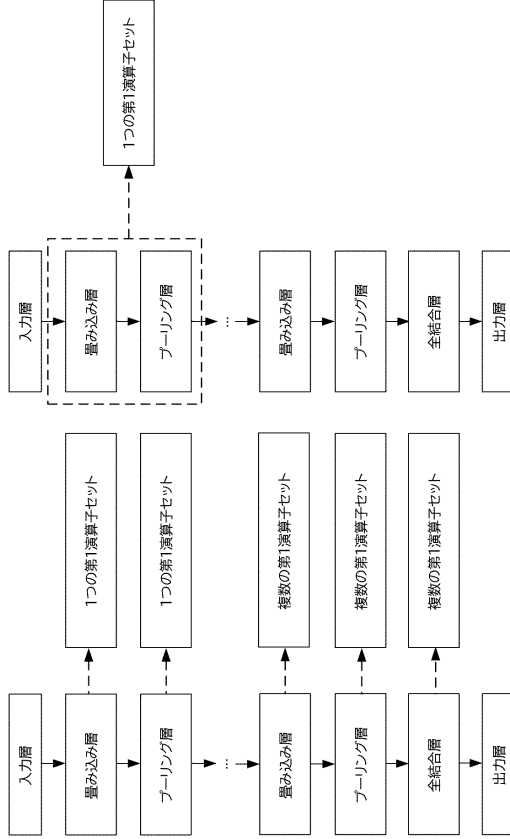
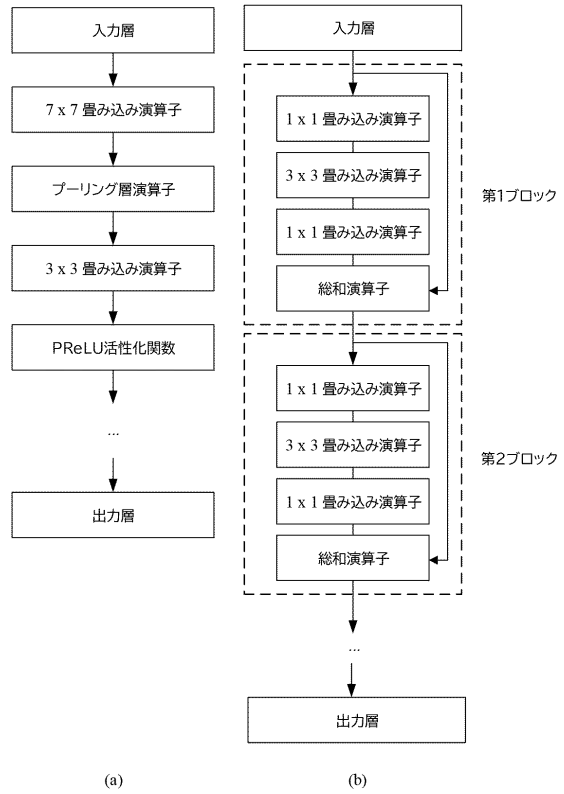


FIG. 4

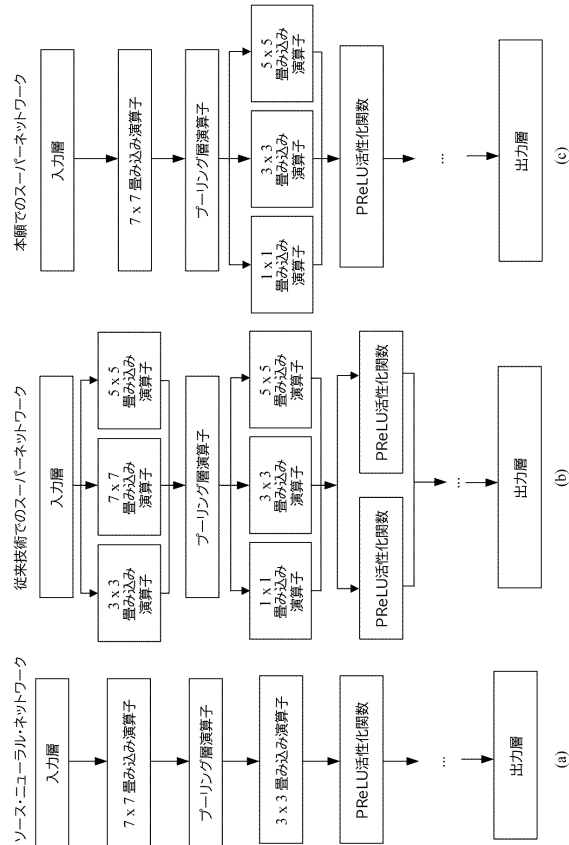
【図 5】



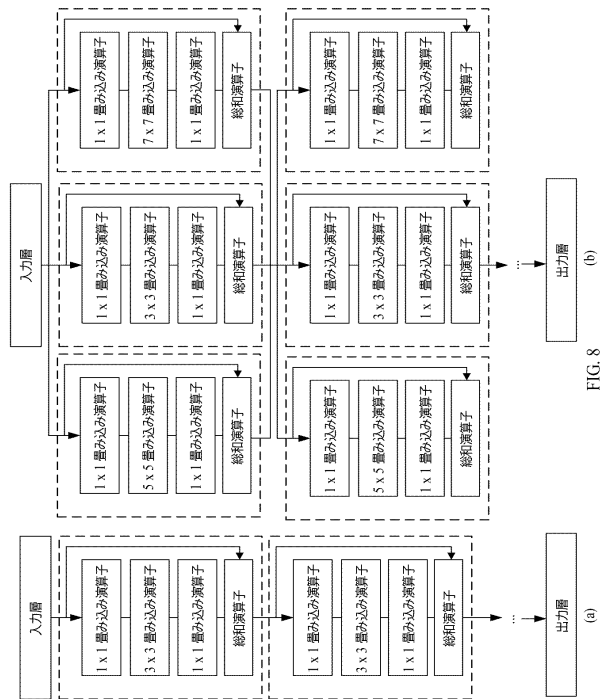
【図 6】



【図 7】



【図 8】



10

20

30

40

50

【図 9】

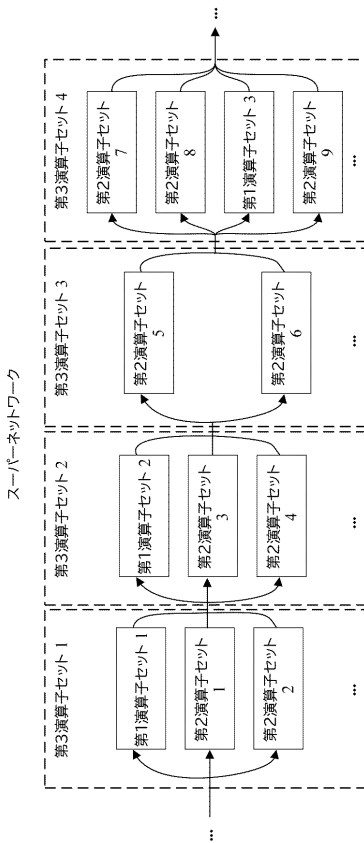


FIG. 9

【図 10】

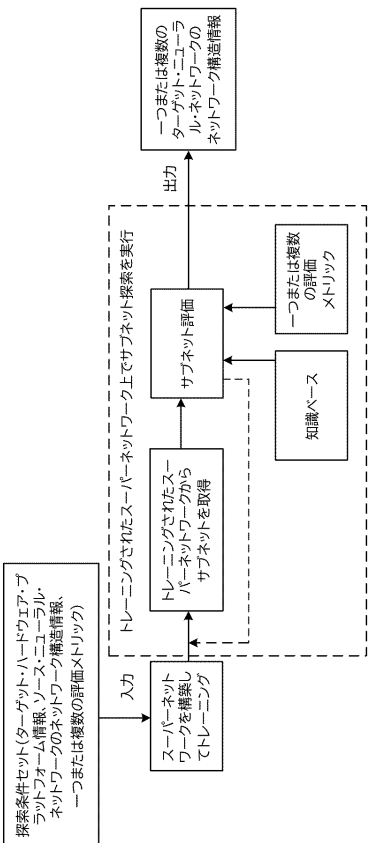


FIG. 10

【図 11 a】

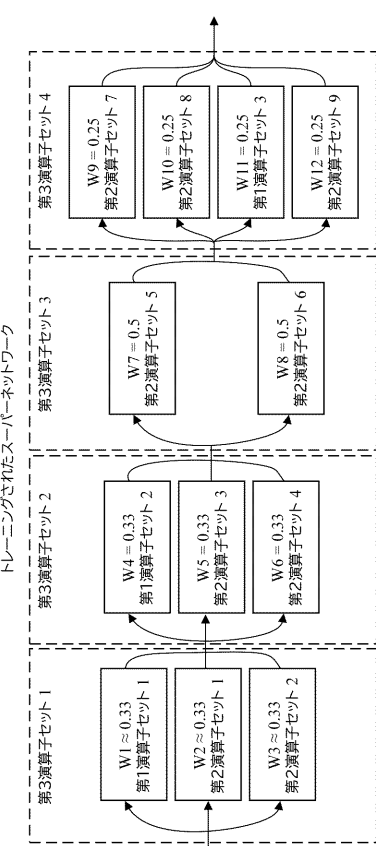


FIG. 11(a)

【図 11 b】

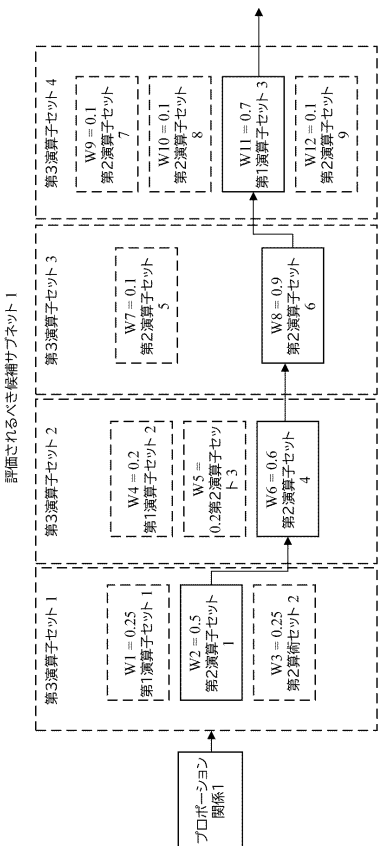


FIG. 11(b)

【図 1 1 c】

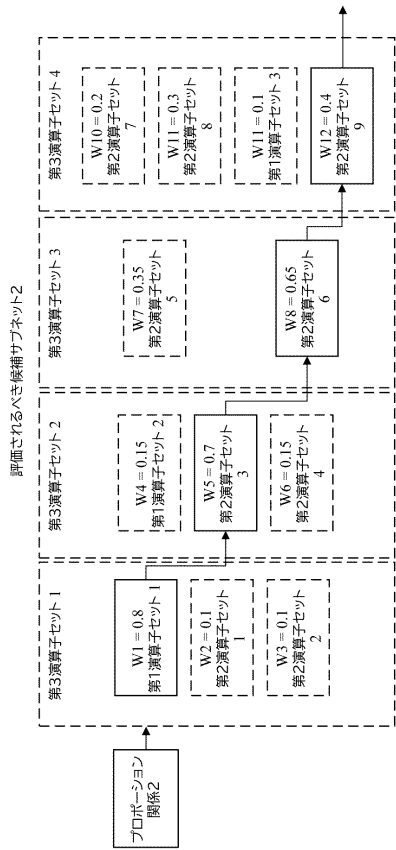


FIG. 11(c)

【図 1 2 a】

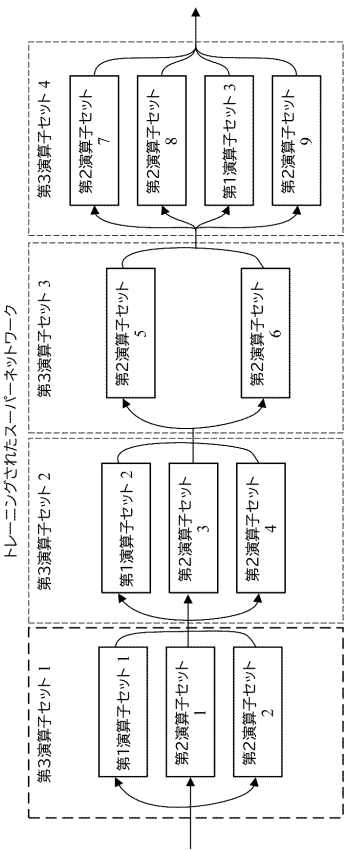


FIG. 12(a)

【図 1 2 b】

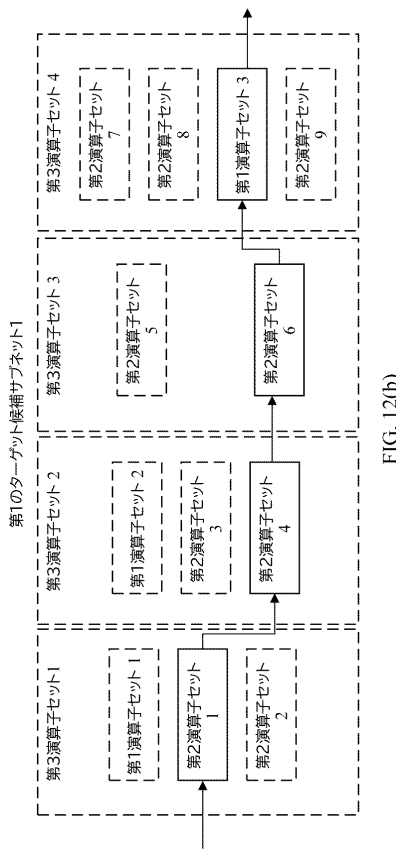


FIG. 12(b)

【図 1 2 c】

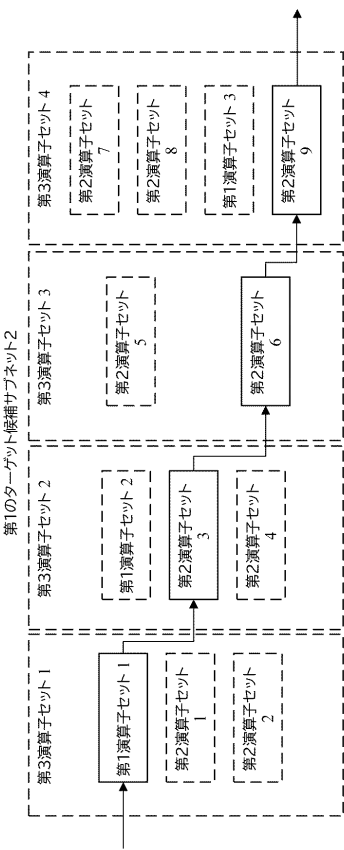


FIG. 12(c)

【図 1 3】

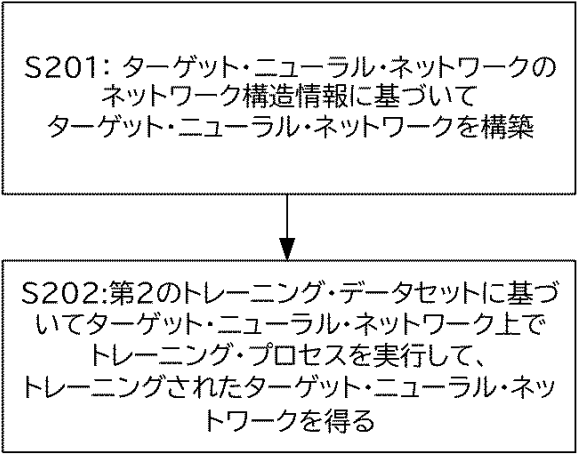


FIG. 13

【図 1 4】

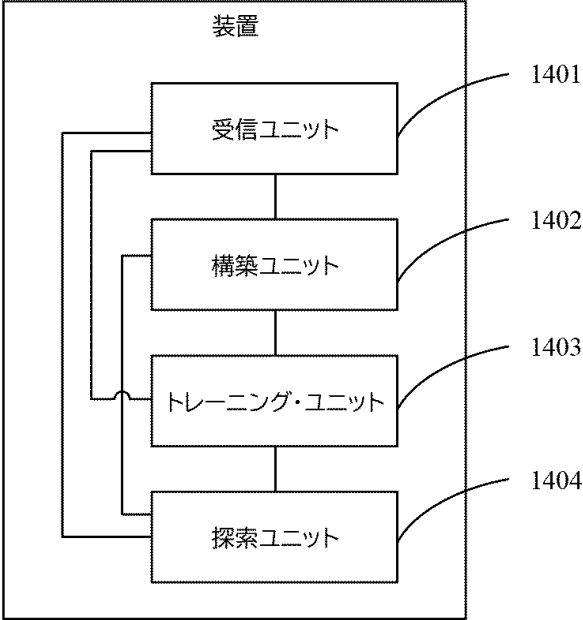


FIG. 14

【図 1 5】

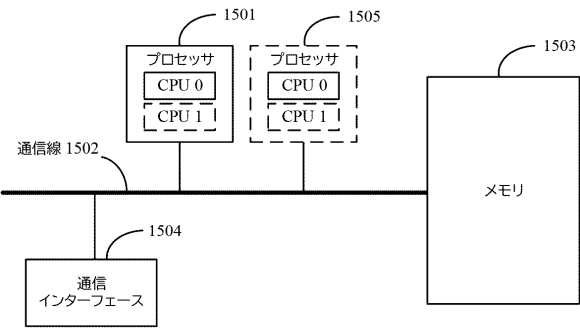


FIG. 15

10

20

30

40

50

フロントページの続き

- (74)代理人 100070150
弁理士 伊東 忠彦
- (74)代理人 100135079
弁理士 宮崎 修
- (72)発明者 チェン, ミンチィ
中国 5 1 8 1 2 9 グァンドン シェンチェン ロンガン・ディストリクト バンティエン ホァウ
エイ・アドミニストレーション・ビルディング
- (72)発明者 デン, チョアンユン
中国 5 1 8 1 2 9 グァンドン シェンチェン ロンガン・ディストリクト バンティエン ホァウ
エイ・アドミニストレーション・ビルディング
- (72)発明者 フウ, シアンユィ
中国 5 1 8 1 2 9 グァンドン シェンチェン ロンガン・ディストリクト バンティエン ホァウ
エイ・アドミニストレーション・ビルディング
- (72)発明者 タン, シャオホア
中国 5 1 8 1 2 9 グァンドン シェンチェン ロンガン・ディストリクト バンティエン ホァウ
エイ・アドミニストレーション・ビルディング
- 審査官 今城 朋彬
- (56)参考文献 特開 2 0 1 5 - 0 1 1 5 1 0 (J P , A)
米国特許出願公開第 2 0 1 5 / 0 0 0 6 4 4 4 (U S , A 1)
中国特許出願公開第 1 1 0 6 5 9 7 2 1 (C N , A)
- (58)調査した分野 (Int.Cl., D B 名)
G 0 6 N 3 / 0 0 - 9 9 / 0 0