(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)公開番号
特開2023-83207
(P2023-83207A)
(43)公開日　令和5年6月15日(2023.6.15)

(51)国際特許分類　FI
G06N 3/04 (2023.01)　G06N 3/04
G06N 3/084(2023.01)　G06N 3/08 140

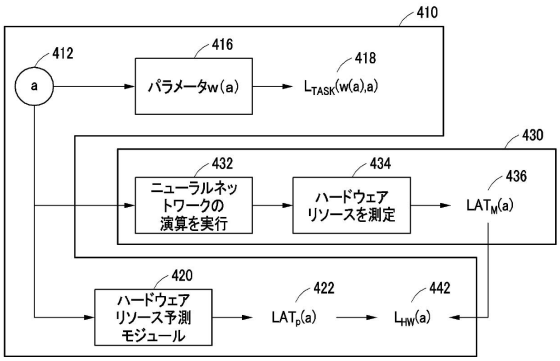
審査請求　未請求　請求項の数　20　OL　（全19頁）

(21)出願番号　特願2022-129509(P2022-129509)
(22)出願日　令和4年8月16日(2022.8.16)
(31)優先権主張番号　10-2021-0171979
(32)優先日　令和3年12月3日(2021.12.3)
(33)優先権主張国・地域又は機関
韓国(KR)

(71)出願人　390019839
三星電子株式会社
Samsung Electronics Co.,Ltd.
大韓民国京畿道水原市霊通区三星路１２９
１２９,Samsung-ro,Yeongtong-gu,Suwon-si,Gyeonggi-do,Republic of Korea
(74)代理人　100107766
弁理士　伊東 忠重
(74)代理人　100070150
弁理士　伊東 忠彦
(74)代理人　100135079
最終頁に続く

(54)【発明の名称】　ニューラルネットワークの最適なアーキテクチャーを探索する装置及び方法

(57)【要約】
【課題】ニューラルネットワークの最適なアーキテクチャーを探索する方法及び装置を提供すること。
【解決手段】ニューラルネットワークの最適なアーキテクチャーを探索する装置は、プロセッサを含み、プロセッサは、ニューラルネットワークの候補アーキテクチャーに対するパラメータに基づいてニューラルネットワーク損失を決定し、候補アーキテクチャーのニューラルネットワークが動作するときに要求されるハードウェアリソースを測定し、ハードウェアリソース予測モジュールを用いて、候補アーキテクチャーのニューラルネットワークが動作するときに要求されるハードウェアリソースを予測し、測定されたハードウェアリソースと予測されたハードウェアリソースに基づいてハードウェアリソース損失を決定し、ニューラルネットワーク損失とハードウェアリソース損失に基づいてニューラルネットワークのターゲットアーキテクチャーを決定することができる。
【選択図】図４



【特許請求の範囲】
【請求項１】
ニューラルネットワークの最適なアーキテクチャーを探索する装置であって、
プロセッサを含み、
前記プロセッサは、
前記ニューラルネットワークの候補アーキテクチャーに対するパラメータに基づいてニューラルネットワーク損失を決定し、
前記候補アーキテクチャーのニューラルネットワークが動作するときに要求されるハードウェアリソースを測定し、
ハードウェアリソース予測モジュールを用いて前記候補アーキテクチャーのニューラルネットワークが動作するときに要求されるハードウェアリソースを予測し、
前記測定されたハードウェアリソースと前記予測されたハードウェアリソースに基づいてハードウェアリソース損失を決定し、
前記ニューラルネットワーク損失と前記ハードウェアリソース損失に基づいて前記ニューラルネットワークのターゲットアーキテクチャーを決定する、
装置。
【請求項２】
前記ハードウェアリソース予測モジュールは、前記候補アーキテクチャーのパラメータを入力とし、前記入力されたパラメータに基づいて、前記候補アーキテクチャーのニューラルネットワークの前記ハードウェアリソースを予測してハードウェアリソース予測値を出力するニューラルネットワークである、請求項１に記載の装置。
【請求項３】
前記プロセッサは、
前記測定されたハードウェアリソースと前記予測されたハードウェアリソースとの間の差に基づいて前記ハードウェアリソース損失を決定し、
前記ハードウェアリソース損失が最小になるように前記候補アーキテクチャーのパラメータをアップデートする、請求項１又は２に記載の装置。
【請求項４】
前記プロセッサは、前記候補アーキテクチャー及び前記候補アーキテクチャーのパラメータによる前記ニューラルネットワーク損失と、前記候補アーキテクチャーのパラメータによる前記ハードウェアリソース損失の加重和を最適化損失として決定し、前記最適化損失を最小にする前記ターゲットアーキテクチャーを決定する、請求項１に記載の装置。
【請求項５】
前記プロセッサは、前記ニューラルネットワーク損失と前記ハードウェアリソース損失を減少させる前記ターゲットアーキテクチャーとターゲットパラメータを決定する、請求項１に記載の装置。
【請求項６】
前記ニューラルネットワークの各レイヤごとに、各レイヤが有する候補演算のいずれか１つの候補演算を選択することによって前記候補アーキテクチャーが決定される、請求項１乃至５のいずれか一項に記載の装置。
【請求項７】
前記ハードウェアリソース予測モジュールには、前記ニューラルネットワークの各レイヤで実行可能な候補演算のうち、前記候補アーキテクチャーを構成する選択された候補演算に関する情報が入力される、請求項１乃至６のいずれか一項に記載の装置。
【請求項８】
前記ハードウェアリソースは、前記候補アーキテクチャーのニューラルネットワークが動作するときの電力消費、メモリ要求量、演算の数、及び処理時間のうち少なくとも１つを含む、請求項１乃至７のいずれか一項に記載の装置。
【請求項９】
前記プロセッサは、

前記ニューラルネットワーク損失と前記ハードウェアリソース損失を含む最適化損失を決定し、
前記候補アーキテクチャーのニューラルネットワークに含まれている各レイヤの候補演算のうち、前記最適化損失を最小にするターゲット演算を選択することによって前記ターゲットアーキテクチャーを決定する、請求項１に記載の装置。
【請求項１０】
前記プロセッサは、前記候補アーキテクチャーのニューラルネットワークが学習データを処理して導き出した結果データと検証データとの間の差に基づいて前記ニューラルネットワーク損失を決定する、請求項１乃至９のいずれか一項に記載の装置。
【請求項１１】
拡張現実提供装置であって、
ターゲットアーキテクチャーを有するニューラルネットワークを用いて処理動作を行うプロセッサを含み、
前記プロセッサは、
前記ニューラルネットワークの候補アーキテクチャーに対するパラメータに基づいてニューラルネットワーク損失を決定し、
前記候補アーキテクチャーのニューラルネットワークが動作するときに要求されるハードウェアリソースを測定し、
ハードウェアリソース予測モジュールを用いて、前記候補アーキテクチャーのニューラルネットワークが動作するときに要求されるハードウェアリソースを予測し、
前記測定されたハードウェアリソースと前記予測されたハードウェアリソースに基づいてハードウェアリソース損失を決定し、
前記ニューラルネットワーク損失と前記ハードウェアリソース損失に基づいて前記ターゲットアーキテクチャーを決定する、
拡張現実提供装置。
【請求項１２】
前記拡張現実提供装置は、映像を撮影するカメラをさらに含み、
前記プロセッサは、前記ターゲットアーキテクチャーを有するニューラルネットワークを用いて前記映像を処理することによって拡張現実コンテンツを生成する、請求項１１に記載の拡張現実提供装置。
【請求項１３】
前記ハードウェアリソース予測モジュールは、前記候補アーキテクチャーのパラメータを入力とし、前記入力されたパラメータに基づいて、前記候補アーキテクチャーのニューラルネットワークの前記ハードウェアリソースを予測してハードウェアリソース予測値を出力するニューラルネットワークである、請求項１１又は１２に記載の拡張現実提供装置。
【請求項１４】
前記プロセッサは、前記候補アーキテクチャー及び前記候補アーキテクチャーのパラメータによる前記ニューラルネットワーク損失と、前記候補アーキテクチャーのパラメータによる前記ハードウェアリソース損失の加重和を最小にする前記ターゲットアーキテクチャーを決定する、請求項１１乃至１３のいずれか一項に記載の拡張現実提供装置。
【請求項１５】
ニューラルネットワークの最適なアーキテクチャーを探索する方法であって、
前記ニューラルネットワークの候補アーキテクチャーに対するパラメータに基づいてニューラルネットワーク損失を決定する動作と、
前記候補アーキテクチャーのニューラルネットワークが動作するときに要求されるハードウェアリソースを測定する動作と、
ハードウェアリソース予測モジュールを用いて、前記候補アーキテクチャーのニューラルネットワークが動作するときに要求されるハードウェアリソースを予測する動作と、
前記測定されたハードウェアリソースと前記予測されたハードウェアリソースに基づい

てハードウェアリソース損失を決定する動作と、
前記ニューラルネットワーク損失と前記ハードウェアリソース損失に基づいて前記ニューラルネットワークのターゲットアーキテクチャーを決定する動作と、
を含む方法。
【請求項１６】
前記ハードウェアリソース予測モジュールは、前記候補アーキテクチャーのパラメータを入力とし、前記入力されたパラメータに基づいて、前記候補アーキテクチャーのニューラルネットワークの前記ハードウェアリソースを予測してハードウェアリソース予測値を出力するニューラルネットワークである、請求項１５に記載の方法。
【請求項１７】
前記ターゲットアーキテクチャーを決定する動作は、前記候補アーキテクチャー及び前記候補アーキテクチャーのパラメータによる前記ニューラルネットワーク損失と、前記候補アーキテクチャーのパラメータによる前記ハードウェアリソース損失の加重和を最小にする前記ターゲットアーキテクチャーを決定する動作を含む、請求項１５又は１６に記載の方法。
【請求項１８】
前記ターゲットアーキテクチャーを決定する動作は、前記候補アーキテクチャーのニューラルネットワークに含まれる各レイヤの候補演算のうちターゲット演算を選択することによって前記ターゲットアーキテクチャーを決定する動作を含む、請求項１５又は１６に記載の方法。
【請求項１９】
前記ハードウェアリソースは、前記候補アーキテクチャーのニューラルネットワークが動作するときの電力消費、メモリ要求量、演算の数及び処理時間のうち少なくとも１つを含む、請求項１５乃至１８のいずれか一項に記載の方法。
【請求項２０】
請求項１５乃至１９のいずれか一項に記載の方法を行うための命令語を含む１つ以上のコンピュータプログラムを格納したコンピュータ読み取り可能な記録媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
以下の開示は、ニューラルネットワークの最適なアーキテクチャーを探索する技術に関する。
【背景技術】
【０００２】
ニューラルアーキテクチャー探索（ｎｅｕｒａｌ　ａｒｃｈｉｔｅｃｔｕｒｅ　ｓｅａｒｃｈ、ＮＡＳ）は、所定の目的のニューラルネットワークにおける最適なアーキテクチャーを自動で探すための方法論のうちの１つである。ＮＡＳは、特定の問題を解決するための最も適切なニューラルネットワークのアーキテクチャーの構造及び形態を、ディープラーニングを介して探索する方法である。ＮＡＳにおけるニューラルネットワークは、探索空間と呼ぶ、予め定義された演算子及び関数で構成されたプリミティブ演算（ｐｒｉｍｉｔｉｖｅ　ｏｐｅｒａｔｉｏｎｓ）を選択及び組み合わせて生成されてもよい。ここで、演算子の例示として、畳み込み（ｃｏｎｖｏｌｕｔｉｏｎ）、プーリング（ｐｏｏｌｉｎｇ）、併合（ｃｏｎｃａｔｅｎａｔｉｏｎ）、スキップ接続（ｓｋｉｐ　ｃｏｎｎｅｃｔｉｏｎ）などが挙げられる。
【発明の概要】
【発明が解決しようとする課題】
【０００３】
本発明の目的は、ニューラルネットワークの最適なアーキテクチャーを探索する技術を提供する。
【課題を解決するための手段】

【０００４】
一実施形態に係るニューラルネットワークの最適なアーキテクチャーを探索する装置は、プロセッサを含み、前記プロセッサは、前記ニューラルネットワークの候補アーキテクチャーに対するパラメータに基づいてニューラルネットワーク損失を決定し、前記候補アーキテクチャーのニューラルネットワークが動作するときに要求されるハードウェアリソースを測定し、ハードウェアリソース予測モジュールを用いて前記候補アーキテクチャーのニューラルネットワークが動作するときに要求されるハードウェアリソースを予測し、前記測定されたハードウェアリソースと前記予測されたハードウェアリソースに基づいてハードウェアリソース損失を決定し、前記ニューラルネットワーク損失と前記ハードウェアリソース損失に基づいて前記ニューラルネットワークのターゲットアーキテクチャーを決定する。
【０００５】
前記ハードウェアリソース予測モジュールは、前記候補アーキテクチャーのパラメータを入力とし、前記入力されたパラメータに基づいて、前記候補アーキテクチャーのニューラルネットワークの前記ハードウェアリソースを予測してハードウェアリソース予測値を出力するニューラルネットワークであってもよい。
【０００６】
前記プロセッサは、前記候補アーキテクチャー及び前記候補アーキテクチャーのパラメータによる前記ニューラルネットワーク損失と、前記候補アーキテクチャーのパラメータによる前記ハードウェアリソース損失の加重和を最適化損失として決定し、前記最適化損失を最小にする前記ターゲットアーキテクチャーを決定することができる。
【０００７】
前記ハードウェアリソースは、前記候補アーキテクチャーのニューラルネットワークが動作するときの電力消費、メモリ要求量、演算の数、及び処理時間のうち少なくとも１つを含むことができる。
【０００８】
前記プロセッサは、前記ニューラルネットワーク損失と前記ハードウェアリソース損失を含む最適化損失を決定し、前記候補アーキテクチャーのニューラルネットワークに含まれている各レイヤの候補演算のうち、前記最適化損失を最小にするターゲット演算を選択することによって前記ターゲットアーキテクチャーを決定することができる。
【０００９】
一実施形態に係る拡張現実提供装置は、ターゲットアーキテクチャーを有するニューラルネットワークを用いて処理動作を行うプロセッサを含み、前記プロセッサは、前記ニューラルネットワークの候補アーキテクチャーに対するパラメータに基づいてニューラルネットワーク損失を決定し、前記候補アーキテクチャーのニューラルネットワークが動作するときに要求されるハードウェアリソースを測定し、ハードウェアリソース予測モジュールを用いて、前記候補アーキテクチャーのニューラルネットワークが動作するときに要求されるハードウェアリソースを予測し、前記測定されたハードウェアリソースと前記予測されたハードウェアリソースに基づいてハードウェアリソース損失を決定し、前記ニューラルネットワーク損失と前記ハードウェアリソース損失に基づいて前記ターゲットアーキテクチャーを決定する。
【００１０】
一実施形態に係るニューラルネットワークの最適なアーキテクチャーを探索する方法は、前記ニューラルネットワークの候補アーキテクチャーに対するパラメータに基づいてニューラルネットワーク損失を決定する動作と、前記候補アーキテクチャーのニューラルネットワークが動作するときに要求されるハードウェアリソースを測定する動作と、ハードウェアリソース予測モジュールを用いて、前記候補アーキテクチャーのニューラルネットワークが動作するときに要求されるハードウェアリソースを予測する動作と、前記測定されたハードウェアリソースと前記予測されたハードウェアリソースに基づいてハードウェアリソース損失を決定する動作と、前記ニューラルネットワーク損失と前記ハードウェア

リソース損失に基づいて前記ニューラルネットワークのターゲットアーキテクチャーを決定する動作とを含む。

【発明の効果】

【００１１】

一実施形態によれば、ハードウェアリソース制限事項を考慮してニューラルネットワークの最適なアーキテクチャーを探索することができ、短い時間で最適化を行うことができる。

【図面の簡単な説明】

【００１２】

【図１】一実施形態に係るニューラルネットワークの最適なアーキテクチャーを探索する探索フレームワークを説明するための図である。

【図２】一実施形態に係るニューラルネットワークの最適なアーキテクチャーを探索する探索装置の構成を示すブロック図である。

【図３】一実施形態に係る各レイヤの候補演算のうち、最適なターゲット演算を選択する過程を説明するための図である。

【図４】一実施形態に係るニューラルネットワークのターゲットアーキテクチャーを決定するための探索過程を説明するための図である。

【図５】一実施形態に係るニューラルネットワークの最適なアーキテクチャーを探索する方法の動作を説明するためのフローチャートである。

【図６】一実施形態に係る電子装置の構成を示す図である。

【発明を実施するための形態】

【００１３】

実施形態に対する特定な構造的又は機能的な説明は、単なる例示のための目的として開示されたものであって、様々な形態に変更されることができる。したがって、実施形態は特定の開示形態に限定されるものではなく、本明細書の範囲は技術的な思想に含まれる変更、均等物ないし代替物を含む。

【００１４】

第１又は第２などの用語を、複数の構成要素を説明するために用いることがあるが、このような用語は１つの構成要素を他の構成要素から区別する目的としてのみ解釈されなければならない。例えば、第１構成要素は第２構成要素と命名することができ、同様に、第２構成要素は第１構成要素にも命名することができる。

【００１５】

いずれかの構成要素が他の構成要素に「連結」されているか「接続」されていると言及されたときには、その他の構成要素に直接的に連結されているか又は接続されているが、中間に他の構成要素が存在し得るものと理解されなければならない。

【００１６】

単数の表現は、文脈上、明白に異なる意味をもたない限り複数の表現を含む。本明細書において、「含む」又は「有する」等の用語は、明細書上に記載した特徴、数字、ステップ、動作、構成要素、部品又はこれらを組み合わせたものが存在することを示すものであって、１つ又はそれ以上の他の特徴や数字、ステップ、動作、構成要素、部品、又はこれを組み合わせたものなどの存在又は付加の可能性を予め排除しないものとして理解しなければならない。

【００１７】

異なるように定義されない限り、技術的又は科学的な用語を含め、ここで用いる全ての用語は、本実施形態が属する技術分野で通常の知識を有する者によって一般的に理解されるものと同じ意味を有する。一般的に用いられる予め定義された用語は、関連技術の文脈上で有する意味と一致する意味を有するものと解釈されなければならず、本明細書で明白に定義しない限り、理想的又は過度に形式的な意味として解釈されることはない。

【００１８】

以下、添付する図面を参照しながら実施形態を詳細に説明する。図面を参照して説明す

る際に、図面符号に拘わらず同じ構成要素は同じ参照符号を付与し、これに対する重複する説明は省略する。

【００１９】

図１は、一実施形態に係るニューラルネットワークの最適なアーキテクチャーを探索する探索フレームワークを説明するための図である。

【００２０】

図１を参照すると、探索フレームワーク１００は、機械学習を介して基本ニューラルネットワーク１２０に対する最適なアーキテクチャー（又は、ニューラルネットワーク構造）を探索するフレームワークである。基本ニューラルネットワーク１２０は、学習される以前のニューラルネットワーク（又は、学習されていないニューラルネットワーク）として、各レイヤの演算及びパラメータ（例えば、連結加重値）などが確定していないニューラルネットワークである。基本ニューラルネットワーク１２０は、複数のニューラルネットワークレイヤ（又は、簡単に「レイヤ」）を含んでもよい。基本ニューラルネットワーク１２０は、深層ニューラルネットワーク（ｄｅｅｐ　ｎｅｕｒａｌ　ｎｅｔｗｏｒｋ、ＤＮＮ）、畳み込みニューラルネットワーク（ｃｏｎｖｏｌｕｔｉｏｎａｌ　ｎｅｕｒａｌ　ｎｅｔｗｏｒｋ、ＣＮＮ）、再帰的ニューラルネットワーク（ｒｅｃｕｒｒｅｎｔ　ｎｅｕｒａｌ　ｎｅｔｗｏｒｋ、ＲＮＮ）、ＲＢＭ（ｒｅｓｔｒｉｃｔｅｄ　ｂｏｌｔｚｍａｎｎ　ｍａｃｈｉｎｅ）、ＤＢＮ（ｄｅｅｐ　ｂｅｌｉｅｆ　ｎｅｔｗｏｒｋ）、ＢＲＤＮＮ（ｂｉｄｉｒｅｃｔｉｏｎａｌ　ｒｅｃｕｒｒｅｎｔ　ｄｅｅｐ　ｎｅｕｒａｌ　ｎｅｔｗｏｒｋ）、深層Ｑ－ネットワーク（ｄｅｅｐ　Ｑ－ｎｅｔｗｏｒｋｓ）又は、このうち２以上の組み合わせの１つであってもよいが、前述した例に限定されない。基本ニューラルネットワーク１２０は、ハードウェア構造及び／又はソフトウェア構造を含んでもよい。

【００２１】

探索フレームワーク１００は、データベース１１０に格納された学習データを用いて基本ニューラルネットワーク１２０に対して機械学習を行う。機械学習は、教師あり学習（ｓｕｐｅｒｖｉｓｅｄ　ｌｅａｒｎｉｎｇ）又は部分的教師あり学習（ｐａｒｔｉａｌ　ｓｕｐｅｒｖｉｓｅｄ　ｌｅａｒｎｉｎｇ）方式により実行され得る。

【００２２】

一実施形態において、探索フレームワーク１００は、教師あり学習を介して基本ニューラルネットワーク１２０を学習させることができる。探索フレームワーク１００は、確率的勾配下降方式（ｓｔｏｃｈａｓｔｉｃ　ｇｒａｄｉｅｎｔ　ｄｅｓｃｅｎｔ）のような調整アルゴリズム及び損失関数を用いて学習を行ってもよい。学習に使用される学習データは、ニューラルネットワークに入力される入力データと、該当入力データに対応する検証データを含んでもよい。基本ニューラルネットワーク１２０は、学習データに含まれる入力データを処理して結果データを出力することができる。探索フレームワーク１００は、基本ニューラルネットワーク１２０から出力された結果データと検証データとの間の比較結果に基づいてニューラルネットワーク損失を決定し、ニューラルネットワーク損失を最小化する最適なアーキテクチャーを探索することができる。

【００２３】

探索フレームワーク１００は、多重目的（ｍｕｌｔｉｐｌｅ　ｏｂｊｅｃｔｉｖｅ）のニューラルアーキテクチャー探索（ＮＡＳ）方式を行って、ターゲットニューラルネットワーク１３０のための最適なアーキテクチャーを探索することができる。探索フレームワーク１００は、基本ニューラルネットワーク１２０のアーキテクチャーをサンプリングするのではなく、基本ニューラルネットワーク１２０の各レイヤごとに様々な候補演算を設定し、この候補演算のうち最も適切な候補演算であるターゲット演算をレイヤごとに選択する方式に基づいて最適なアーキテクチャーを探索することができる。このような探索方法によって探索フレームワーク１００は、早期に最適化を行うことができる。

【００２４】

探索フレームワーク１００は、学習過程を通じた目的（例えば、オブジェクト分類、オ

ブジェクト認識、音声認識など）による最適なアーキテクチャーを有するターゲットニューラルネットワーク１３０を導き出すことができる。最適なアーキテクチャーを探索することは、ニューラルネットワークの各レイヤで実行される演算を決定し、ニューラルネットワークパラメータの最適値を決定することを含む。探索フレームワーク１００は、本明細書で説明されるニューラルネットワークの最適なアーキテクチャーを探索する装置（例えば、図２に示す探索装置２００）によって実行され得る。

【００２５】

探索フレームワーク１００は、ターゲットニューラルネットワーク１３０のための最適なアーキテクチャーを探索することにおいて、ハードウェアリソース制限事項（ｈａｒｄｗａｒｅ　ｒｅｓｏｕｒｃｅ　ｃｏｎｓｔｒａｉｎｔ）を考慮する。探索フレームワーク１００は、ニューラルネットワークが行うタスク（ｔａｓｋ）に対する検証損失だけでなく、ニューラルネットワークが実行されるとき用いられるハードウェアリソースを考慮して最適化することができる。探索フレームワーク１００は、ニューラルネットワークが動作時に必要とするハードウェアリソースを考慮して、ターゲットニューラルネットワーク１３０を探索することができる。ハードウェアリソースは、例えば、電力消費、メモリ要求量、演算の数（ｎｕｍｂｅｒ　ｏｆ　ｏｐｅｒａｔｉｏｎｓ）（例えば、ｍａｕｌｔｉｐｌｙ－ａｃｃｕｍｕｌａｔｅ（ＭＡＣ）演算の数）、処理時間、及びＧＰＵの占有率などであり得る。探索フレームワーク１００は、１つ又は２以上のハードウェアリソースを考慮し、上記の例のハードウェアリソース以外にも数値で観測され得るハードウェアリソースであればいずれも制限されることなく考慮できる。

【００２６】

探索フレームワーク１００は、基本ニューラルネットワーク１２０に対する候補アーキテクチャーが決定されれば、候補アーキテクチャーに対するニューラルネットワーク損失とハードウェアリソースに対するハードウェアリソース損失を決定し、ニューラルネットワーク損失とハードウェアリソース損失を最小化するターゲットアーキテクチャーを探索することができる。ニューラルネットワーク損失とハードウェアリソース損失は、ターゲットアーキテクチャーを決定するための最適化損失を構成する。

【００２７】

探索フレームワーク１００は、ハードウェアリソース損失を決定するとき、候補アーキテクチャーのニューラルネットワークが要求するハードウェアリソースに対する実際の測定値と、ハードウェアリソース予測モジュール（例えば、図４に示すハードウェアリソース予測モジュール４２０）を用いて導き出されたハードウェアリソースに対する予測値に基づいて、ハードウェアリソース損失を決定することができる。下記でより詳しく説明されるが、ハードウェアリソース予測モジュールは、現在の学習の対象である候補アーキテクチャーのニューラルネットワークに対するハードウェアリソースを予測した予測値を提供するモジュールである。ハードウェアリソース予測モジュールは、入力された候補アーキテクチャーのパラメータに基づいて、該当候補アーキテクチャーのニューラルネットワークが必要とするハードウェアリソースに対する予測値を出力するよう学習された、ニューラルネットワークによって実現されることができる。ハードウェアリソース予測モジュールは、微分可能な特性を有し、ハードウェアリソース予測モジュールを介して探索過程における微分可能性が保持され得る、微分可能性が保持されることにより、エンドツーエンド学習（ｅｎｄ－ｔｏ－ｅｎｄ　ｌｅａｒｎｉｎｇ）が可能になる。探索フレームワーク１００は、ハードウェアリソース予測モジュールを介してニューラルネットワークのアーキテクチャーに対するハードウェアリソースを最適化損失に反映する。

【００２８】

上記のように、探索フレームワーク１００は、ハードウェアリソースの制限事項を考慮して、ニューラルネットワークの最適なアーキテクチャーを探索することができ、短い時間で最適化を行うことができる。また、探索フレームワーク１００は、実際の測定したハードウェアリソース測定値を考慮して最適なアーキテクチャーを探索することができる。

【００２９】

図２は、一実施形態に係るニューラルネットワークの最適なアーキテクチャーを探索する探索装置の構成を示すブロック図である。

【００３０】

図２を参照すると、探索装置２００は、ニューラルネットワークに対する最適なアーキテクチャーを探索する装置であって、図１を参照して説明した探索フレームワーク１００を行うことができる。探索装置２００は、アーキテクチャー探索に関連して本明細書で記載されるか、又は、図示した１つ以上の動作を行ってもよい。探索装置２００は、プロセッサ２１０及びメモリ２２０を含む。格納装置２３０は、アーキテクチャー探索のためのデータ（例えば、学習データ）を格納し、学習に使用されるニューラルネットワークを格納することができる。

【００３１】

メモリ２２０は、探索装置２００の構成要素（例えば、プロセッサ２１０）によって用いられる様々なデータを格納することができる。データは、例えば、ソフトウェア、及びこれに関する命令に対する入力データ又は出力データを含んでもよい。メモリ２２０は、揮発性メモリ及び不揮発性メモリのうちの１つ以上を含んでもよい。

【００３２】

プロセッサ２１０は、探索装置２００の全体的な動作を制御し、探索装置２００の動作を行うための命令を実行する。プロセッサ２１０は、例えば、ソフトウェアを実行してプロセッサ２１０に接続された探索装置２００の少なくとも１つの他の構成要素（例えば、ハードウェア、又は、ソフトウェア構成要素）を制御し、様々なデータ処理又は演算を行うことができる。

【００３３】

一実施形態によれば、データ処理又は演算の少なくとも一部として、プロセッサ２１０は、命令又はデータをメモリ２２０に格納し、メモリ２２０に格納された命令又はデータを処理し、結果データをメモリ２２０に格納することができる。プロセッサ２１０は、メインプロセッサ（例えば、中央処理装置又はアプリケーションプロセッサ）、又はこれとは独立的又は共に運営可能な補助プロセッサ（例えば、グラフィック処理装置、神経網処理装置（ＮＰＵ：ｎｅｕｒａｌ　ｐｒｏｃｅｓｓｉｎｇ　ｕｎｉｔ）、イメージ信号処理部、センサハブプロセッサ、又は、コミュニケーションプロセッサ）を含んでもよい。

【００３４】

プロセッサ２１０は、学習データを用いてニューラルネットワーク（例えば、図１に示す基本ニューラルネットワーク１２０）の候補アーキテクチャーに対する学習を行い、ニューラルネットワーク損失を決定することができる。学習される以前のニューラルネットワークは、それぞれ１つ以上の人工ニューロンを含む複数のレイヤを含み、各レイヤは、実行可能な候補演算が予め定義されていることができる。候補演算は、例えば、３×３カーネル（ｋｅｒｎｅｌ）基盤の畳み込み演算、５×５カーネル基盤の畳み込み演算、プーリング演算などを含み得るが、これに限定されることはない。ニューラルネットワークの各レイヤごとに、各レイヤが有する候補演算のいずれか１つの候補演算を選択することで、候補アーキテクチャーが決定されてもよい。プロセッサ２１０は、ニューラルネットワークの候補アーキテクチャーに対するパラメータに基づいてニューラルネットワーク損失を決定することができる。プロセッサ２１０は、候補アーキテクチャーのニューラルネットワークが学習データを処理して導出した結果データと検証データとの間の差に基づいて、ニューラルネットワーク損失を決定することができる。ニューラルネットワーク損失は、予め定義された損失関数によって決定され得る。

【００３５】

プロセッサ２１０は、候補アーキテクチャーのニューラルネットワークが動作するときに要求（又は、使用）されるハードウェアリソースを測定できる。測定されるハードウェアリソースは、例えば、候補アーキテクチャーのニューラルネットワークが動作するときの電力消費、メモリ要求量、演算の数、及び処理時間のうちの１つ又は２以上を含むことができるが、これに限定されることはない。プロセッサ２１０は、ハードウェアリソース

を測定してハードウェアリソース測定値を決定することができる。
【００３６】
プロセッサ２１０は、ハードウェアリソース予測モジュール（例えば、図４に示すハードウェアリソース予測モジュール４２０）を用いて、候補アーキテクチャーのニューラルネットワークが動作するときに要求されるハードウェアリソースを予測できる。ハードウェアリソース予測モジュールは、候補アーキテクチャーのパラメータ（例えば、各レイヤの候補演算のうち選択された候補演算に関する情報）を入力とし、入力されたパラメータに基づいて候補アーキテクチャーのニューラルネットワークのハードウェアリソースを予測し、ハードウェアリソース予測値を出力するニューラルネットワークであり得る。
【００３７】
プロセッサ２１０は、測定されたハードウェアリソースと予測されたハードウェアリソースに基づいてハードウェアリソース損失を決定できる。プロセッサ２１０は、測定されたハードウェアリソースと予測されたハードウェアリソースとの間の差と、予め定義された損失関数に基づいてハードウェアリソース損失を決定することができる。例えば、プロセッサ２１０は、候補アーキテクチャーのニューラルネットワークが動作するとき、実際に測定された処理時間測定値とハードウェアリソース予測モデルから出力された処理時間予想値との間の差を、損失関数に適用してハードウェアリソース損失を決定することができる。
【００３８】
プロセッサ２１０は、ニューラルネットワーク損失とハードウェアリソース損失に基づいて、ニューラルネットワークのターゲットアーキテクチャーを決定できる。プロセッサ２１０は、ニューラルネットワーク損失とハードウェアリソース損失を減少させるターゲットアーキテクチャー及びターゲットパラメータを決定する。プロセッサ２１０は、ハードウェアリソース損失が最小になるよう候補アーキテクチャーのパラメータをアップデートすることができる。プロセッサ２１０は、ニューラルネットワーク損失とハードウェアリソース損失を含む最適化損失を決定し、候補アーキテクチャーのニューラルネットワークに含まれている各レイヤの候補演算のうち、最適化損失を最小にするターゲット演算を選択することで、ターゲットアーキテクチャーを決定することができる。プロセッサ２１０は、ニューラルネットワーク損失とハードウェアリソース損失を最も最小化する各レイヤのターゲット演算を選択し、ニューラルネットワークのパラメータをアップデートすることができる。プロセッサ２１０は、候補アーキテクチャー及び候補アーキテクチャーのパラメータによるニューラルネットワーク損失と、候補アーキテクチャーのパラメータによるハードウェアリソース損失の加重和を最適化損失として決定し、該当最適化損失を最小にするターゲットアーキテクチャーを決定することができる。
【００３９】
以上で説明した探索装置２００により実行される動作は、ウェアラブル装置、スマートフォンのようなモバイル装置だけでなく、埋め込みシステム（ｅｍｂｅｄｄｅｄ　ｓｙｓｔｅｍ）で作動可能なニューラルネットワーク基盤のアルゴリズムに多様に応用できる。
【００４０】
図３は、一実施形態に係る各レイヤの候補演算のうち最適なターゲット演算を選択する過程を説明するための図である。
【００４１】
図３を参照すると、ステップＳ３１０において、学習される以前のニューラルネットワーク（例えば、図１に示す基本ニューラルネットワーク１２０）は、複数のレイヤ３１２，３１４，３１６，３１８を含み、各レイヤ３１２，３１４，３１６，３１８は、実行可能な候補演算が定義されている。図示した実施形態において、ニューラルネットワーク３１０は、各レイヤ３１２，３１４，３１６，３１８に対して３個の候補演算を有するものと定義されている。レイヤ間で実行される候補演算は、互いに異なる演算方式であってもよい。例えば、レイヤ３１２とレイヤ３１４との間で実行される候補演算は、互いに異なる方式の演算であってもよい。

【００４２】
探索装置（例えば、図２に示す探索装置２００）は、学習過程において、この候補演算のうち最適な候補演算であるターゲット演算を選択することができる。ステップＳ３２０において、探索装置は、各レイヤ３１２，３１４，３１６，３１８ごとに、候補演算のいずれか１つの候補演算３２２，３２４，３２６，３２８，３２９を選択し、選択された候補演算３２２，３２４，３２６，３２８，３２９の組み合わせで構成された候補アーキテクチャーに対する最適化損失を決定することができる。探索装置は、探索空間で各レイヤ３１２，３１４，３１６，３１８の候補演算を数回組み合わせ、それぞれの組み合わせにおける最適化損失を算出し、最適化損失を最小にする候補演算（言い換えれば、ターゲット演算）の組み合わせを決定することができる。様々な組み合わせに対する学習過程が完了すれば、各レイヤ３１２，３１４，３１６，３１８ごとに選択されたターゲット演算に基づいて、ターゲットアーキテクチャーを決定することができる。ターゲットアーキテクチャーは、各レイヤのターゲット演算を含む。
【００４３】
図４は、一実施形態に係るニューラルネットワークのターゲットアーキテクチャーを決定するための探索過程を説明するための図である。
【００４４】
図４を参照すると、ニューラルネットワークのターゲットアーキテクチャーの探索過程において、ニューラルネットワークの候補アーキテクチャーａ４１２が与えられれば、候補アーキテクチャーａ４１２のパラメータｗ（ａ）４１６が決定され得る。候補アーキテクチャーａ４１２は、各レイヤごとに選択された候補演算の集合を含むニューラルネットワーク構造を示す。
【００４５】
候補アーキテクチャーａ４１２のパラメータは、ニューラルネットワークの各レイヤの候補演算のうちから選択された候補演算に対するパラメータと、選択された各候補演算に対する演算特性を示すパラメータを含んでもよい。例えば、ニューラルネットワークに含まれるいずれかのレイヤの候補演算が、３×３カーネル基盤の畳み込み演算及び５×５カーネル基盤の畳み込み演算があると仮定すれば、候補アーキテクチャーａ４１２のパラメータは、２つの畳み込み演算のいずれかの畳み込み演算を選択したかを示すパラメータと、選択された畳み込み演算のカーネルパラメータ（ｋｅｒｎｅｌ　ｐａｒａｍｅｔｅｒ）などを含んでもよい。ここで、畳み込み演算は、畳み込みレイヤとして実現されてもよい。
【００４６】
探索装置（例えば、図２に示す探索装置２００）は、パラメータｗ（ａ）４１６に基づいて、候補アーキテクチャーａ４１２のニューラルネットワークの特定タスクに対するニューラルネットワーク損失L_{TASK}（ｗ（ａ）、ａ）４１８を決定することができる。ニューラルネットワーク損失L_{TASK}（ｗ（ａ）、ａ）４１８は、ニューラルネットワークが行うタスクの損失を最小化するための損失である。
【００４７】
探索装置は、候補アーキテクチャーａ４１２を有するニューラルネットワークの演算を実行し（４３２）、該当演算の実行過程で要求される全体ニューラルネットワークのハードウェアリソースを実際に測定する（４３４）。測定されるハードウェアリソースは、例えば、電力消費、メモリ要求量、演算の数、及び処理時間などを含んでもよい。ハードウェアリソースの測定４３４を介して候補アーキテクチャーａ４１２に対するハードウェアリソース測定値LAT_M（ａ）４３６が決定される。実施形態により、マックス演算を介して決定されたアーキテクチャーに対するニューラルネットワーク演算を行うことで、ハードウェアリソースを測定することができる。このような過程を含む過程４３０は、微分可能性が成立されないため、該当過程４３０に対しては、フォワード演算及びバックワード演算を定義することができない。これを解決するために、ハードウェアリソース予測モジュール４２０を使用してもよい。

【００４８】
探索装置は、ハードウェアリソース予測モジュール４２０を用いて、候補アーキテクチャーａ４１２を有するニューラルネットワークがこの演算を行うときに要求するものと予想される、全体ニューラルネットワークのハードウェアリソースを予測することができる。ハードウェアリソース予測モジュール４２０を介して候補アーキテクチャーａ４１２に対するハードウェアリソース予測値ＬＡＴ$_{P}$（ａ）４２２が決定され得る。ハードウェアリソース予測モジュール４２０は、候補アーキテクチャーａ４１２のパラメータを入力とし、入力されたパラメータに基づいて候補アーキテクチャーａ４１２のニューラルネットワークのハードウェアリソースを予測し、ハードウェアリソース予測値ＬＡＴ$_{P}$（ａ）４２２を出力することができる。ハードウェアリソース予測モジュール４２０における演算は、微分可能な演算からなる。

【００４９】
ハードウェアリソース予測モジュール４２０は、ニューラルネットワークのアーキテクチャーのパラメータに基づいて該当アーキテクチャーが要求するか、使用するものと予測されるハードウェアリソースの予測値を出力するよう、学習過程を介して学習されたニューラルネットワークであってもよい。但し、ハードウェアリソース予測モジュール４２０は、ニューラルネットワーク以外に、候補アーキテクチャーａ４１２に基づいてニューラルネットワークのハードウェアリソースを予測できる他の手段により実現されてもよい。

【００５０】
探索装置は、ハードウェアリソース測定値ＬＡＴ$_{M}$（ａ）４３６とハードウェアリソース予測値ＬＡＴ$_{P}$（ａ）４２２に基づいて、候補アーキテクチャーａ４１２に対するリソース損失Ｌ$_{HW}$（ａ）４４２を決定することができる。ハードウェアリソース測定値ＬＡＴ$_{M}$（ａ）４３６とハードウェアリソース予測値ＬＡＴ$_{P}$（ａ）４２２との間の差が最小化されるよう、リソース損失Ｌ$_{HW}$（ａ）４４２が定義される。

【００５１】
一実施形態において、リソース損失Ｌ$_{HW}$（ａ）４４２は、ハードウェアリソース測定値ＬＡＴ$_{M}$（ａ）４３６とハードウェアリソース予測値ＬＡＴ$_{P}$（ａ）４２２との間の差による損失を示すＬ$_{HW1}$（ａ）と、ハードウェアリソースの最適化のための要素（例えば、レイテンシを最小化するための要素）のＬ$_{HW2}$（ａ）に基づいて決定される。Ｌ$_{HW1}$（ａ）とＬ$_{HW2}$（ａ）はそれぞれ次の数式（１）及び数式（２）によって決定され得る。

【数１】

$$L_{HW1}(\mathrm{a}) = (LAT_M(\mathrm{a}) - LAT_P(\mathrm{a}))^2 \qquad (1)$$

【数２】

$$L_{HW2}(\mathrm{a}) = (LAT_M(\mathrm{a}))^2 \qquad (2)$$

リソース損失Ｌ$_{HW}$（ａ）４４２は、例えば、次の数式（３）のようにＬ$_{HW1}$（ａ）とＬ$_{HW2}$（ａ）との間の加重和として決定されてもよい。

【数３】

$$L_{HW}(a) = L_{HW1}(a) + w \times L_{HW2}(a) \qquad (3)$$

ここで、ｗは、Ｌ$_{HW2}$（ａ）に適用される加重値として、例えば、予め設定された定数である。実施形態により、Ｌ$_{HW1}$（ａ）のみに加重値が適用されてもよく、Ｌ$_{HW1}$（ａ）とＬ$_{HW2}$（ａ）にそれぞれ互いに異なる加重値が適用されてもよい。

【００５２】
探索装置は、ハードウェアリソース測定値ＬＡＴ$_{M}$（ａ）４３６とハードウェアリソー

ス予測値ＬＡＴ$_{P}$（ａ）４２２との間の差を予め定義された損失関数に適用し、ハードウェアリソース損失Ｌ$_{HW}$（ａ）４４２を決定することができる。

【００５３】

探索装置は、ニューラルネットワーク損失Ｌ$_{TASK}$（ｗ（ａ）、ａ）４１８とハードウェアリソース損失Ｌ$_{HW}$（ａ）４４２を含む最適化損失を決定し、候補アーキテクチャーａ４１２のニューラルネットワークに含まれている各レイヤの候補演算のうち、最適化損失を最小にするターゲット演算を選択することで、ターゲットアーキテクチャーを決定することができる。例えば、ターゲットアーキテクチャーは、次の数式（４）のように最適化損失Ｌ$_{TASK}$（ｗ（ａ）、ａ）＋λ・Ｌ$_{HW}$（ａ）を最小にする候補アーキテクチャーａ４１２と候補アーキテクチャーａのパラメータｗ（ａ）４１６を探索することによって決定されてもよい。数式（４）の例として、最適化損失は、ハードウェアリソース損失Ｌ$_{HW}$（ａ）４４２に加重値λが適用されたニューラルネットワーク損失Ｌ$_{TASK}$（ｗ（ａ）、ａ）４１８とハードウェアリソース損失Ｌ$_{HW}$（ａ）４４２との間の加重和として決定され得る。

【数４】

$$\min_{a}\min_{w} L_{TASK}(w(a),a)+\lambda\cdot L_{HW}(a) \qquad (4)$$

探索装置は、少ない探索時間でニューラルネットワークのターゲットアーキテクチャーを探索し、ターゲットアーキテクチャーを選定することにおいて、ニューラルネットワークが必要とするハードウェアリソースを最適化制限事項として考慮する。図４において、過程４１０は、微分可能な特性を有し、過程４３０は、微分不可能な（ｎｏｎ－ｄｉｆｆｅｒｅｎｔｉａｂｌｅ）特性を有する。探索装置は、ニューラルネットワークの各レイヤの演算に基づいた全体ニューラルネットワークのハードウェアリソースをニューラルネットワークとして実現できるハードウェアリソース予測モジュール４２０を用いて予測することで、微分可能性を保持することができる。微分可能性を保持することによりエンドツーエンド学習が可能になる。また、上述した探索過程は、全体ニューラルネットワークのハードウェアリソースの制限事項を反映して、ニューラルネットワークのアーキテクチャーを微分可能に最適化することができ、一回の学習過程を介してターゲットアーキテクチャーを探し得るため、最適化時間が短いという長所を有する。そして、ターゲットアーキテクチャーの探索過程で生じるフォワード演算とバックワード演算との間の一貫性も保持され得る。

【００５４】

図５は、一実施形態に係るニューラルネットワークの最適なアーキテクチャーを探索する方法の動作を説明するためのフローチャートである。該当方法の動作は、図２に示す探索装置２００によって実行され得る。

【００５５】

図５を参照すると、動作５１０において、探索装置は、ニューラルネットワーク（例えば、図１に示す基本ニューラルネットワーク１２０）の候補アーキテクチャーを選択する。探索装置は、ニューラルネットワークの各レイヤに対して定義された候補演算のいずれか１つの候補演算を選択することで、候補アーキテクチャーを選択することができる。

【００５６】

動作５２０において、探索装置は、ニューラルネットワークの候補アーキテクチャーに対するパラメータに基づいて、ニューラルネットワーク損失を決定する。探索装置は、学習データを用いてニューラルネットワークの候補アーキテクチャーに対する学習を行い、ニューラルネットワーク損失を決定する。探索装置は、候補アーキテクチャーのニューラルネットワークが学習データを処理して導出した結果データと検証データとの間の差に基づいて、ニューラルネットワーク損失を決定することができる。候補アーキテクチャーのニューラルネットワークから導き出された結果データと目的とする検証データとの間の差が大きくなるほど、ニューラルネットワーク損失は大きくなる。

【００５７】
動作５３０において、探索装置は、候補アーキテクチャーのニューラルネットワークが動作するときに要求されるハードウェアリソースを測定する。測定されるハードウェアリソースは、例えば、候補アーキテクチャーのニューラルネットワークが動作するときの電力消費、メモリ要求量、演算の数、及び処理時間のうちの１つ又は２以上を含むが、これに限定されることはない。
【００５８】
動作５４０において、探索装置は、ハードウェアリソース予測モジュール（例えば、図４に示すハードウェアリソース予測モジュール４２０）を用いて候補アーキテクチャーのニューラルネットワークが動作するときに要求されるハードウェアリソースを予測する。ハードウェアリソース予測モデルには、ニューラルネットワークの各レイヤで実行可能な候補演算のうち、候補アーキテクチャーを構成する選択された候補演算に関する情報が入力され、ハードウェアリソース予測モデルは、入力された情報に基づいて該当ニューラルネットワークを必要とするハードウェアリソースの予測値を提供することができる。
【００５９】
動作５５０において、探索装置は、測定されたハードウェアリソースと予測されたハードウェアリソースに基づいて、ハードウェアリソース損失を決定する。探索装置は、測定されたハードウェアリソースと予測されたハードウェアリソースとの差異と、予め定義された損失関数に基づいてハードウェアリソース損失を決定することができる。
【００６０】
動作５６０において、探索装置は、ニューラルネットワーク損失とハードウェアリソース損失に基づいて、ニューラルネットワークのターゲットアーキテクチャーとターゲットパラメータを決定する。探索装置は、ニューラルネットワーク損失とハードウェアリソース損失を含む最適化損失を最小にするターゲットアーキテクチャーとターゲットパラメータを決定し得る。探索装置は、候補アーキテクチャーのニューラルネットワークに含まれている各レイヤの候補演算のうち、最適化損失を最小にするターゲット演算を選択することでターゲットアーキテクチャーを決定できる。探索装置は、候補アーキテクチャー及び候補アーキテクチャーのパラメータによるニューラルネットワーク損失と、候補アーキテクチャーのパラメータによるハードウェアリソース損失の加重和を最適化損失として決定し、該当最適化損失を最小にするターゲットアーキテクチャーを決定する。
【００６１】
図６は、一実施形態に係る電子装置の構成を示す図である。
【００６２】
図６を参照すると、電子装置６００は、様々な形態の電子装置であり得る。例えば、電子装置６００は、ウェアラブル装置（例えば、ＡＲグラスのような拡張現実提供装置、ＨＭＤ（ｈｅａｄ ｍｏｕｎｔｅｄ ｄｉｓｐｌａｙ））スマートフォン、タブレットコンピュータ、ネットブック、ラップトップ、製品検査装置、パーソナルコンピュータ、又はサーバであってもよいが、これに制限されることはない。
【００６３】
電子装置６００は、プロセッサ６１０、メモリ６２０、カメラ６３０、センサ６４０、出力装置６５０、及び通信装置６６０を含む。電子装置６００の各構成要素のうち少なくとも一部は、周辺機器間の通信インターフェース６７０（例えば、バス、ＧＰＩＯ（ｇｅｎｅｒａｌ ｐｕｒｐｏｓｅ ｉｎｐｕｔ ａｎｄ ｏｕｔｐｕｔ）インターフェース、ＳＰＩ（ｓｅｒｉａｌ ｐｅｒｉｐｈｅｒａｌ ｉｎｔｅｒｆａｃｅ）、又は、ＭＩＰＩ（ｍｏｂｉｌｅ ｉｎｄｕｓｔｒｙ ｐｒｏｃｅｓｓｏｒ ｉｎｔｅｒｆａｃｅ））を介して接続されて信号（例えば、命令又はデータ）を相互間に交換することができる。
【００６４】
プロセッサ６１０は、電子装置６００の全体的な動作を制御し、電子装置６００内で実行するための機能及び命令を実行する。プロセッサ６１０は、図１～図５を参照して前述した探索装置（例えば、図２の探索装置２００）の動作を行う。

【００６５】
メモリ６２０は、プロセッサ６１０によって実行可能な命令と入力／出力されるデータを格納する。メモリ６２０は、ＲＡＭ、ＤＲＡＭ、ＳＲＡＭのような揮発性メモリ及び／又はＲＯＭ、フラッシュメモリのようなこの技術分野で知られた不揮発性メモリを含んでもよい。
【００６６】
カメラ６３０は、映像を撮影する。カメラ６３０は、例えば、カラー映像、白黒映像、グレイ映像、赤外線映像、又は深度映像などを取得してもよい。
【００６７】
センサ６４０は、電子装置６００の作動状態（例えば、電力又は温度）、又は、外部の環境状態（例えば、ユーザ状態）を検出し、検出された状態に対応する電気信号又はデータ値を生成することができる。センサ６４０は、例えば、ジェスチャーセンサ、ジャイロセンサ、気圧センサ、マグネチックセンサ、加速度センサ、グリップセンサ、近接センサ、カラーセンサ、ＩＲ（ｉｎｆｒａｒｅｄ）センサ、生体センサ、温度センサ、湿度センサ、又は、照度センサを含んでもよい。
【００６８】
出力装置６５０は、視覚的、聴覚的、又は触覚的なチャネルを通じてユーザに電子装置６００の出力を提供する。出力装置６５０は、例えば、液晶ディスプレイやＬＥＤ／ＯＬＥＤディスプレイ、マイクロＬＥＤ（ｍｉｃｒｏ ｌｉｇｈｔ ｅｍｉｔｔｉｎｇ ｄｉｏｄｅ、ｍｉｃｒｏＬＥＤ）、タッチスクリーン、スピーカー、振動発生装置又はユーザに出力を提供できる任意の他の装置を含んでもよい。
【００６９】
通信装置６６０は、電子装置６００と外部装置との間の直接（例えば、有線）通信チャネル又は無線通信チャネルの確立、及び確立された通信チャネルを通じた通信実行を支援する。一実施形態によれば、通信装置６６０は、無線通信モジュール（例えば、セルラー通信モジュール、近距離無線通信モジュール、又はＧＮＳＳ（ｇｌｏｂａｌ ｎａｖｉｇａｔｉｏｎ ｓａｔｅｌｌｉｔｅ ｓｙｓｔｅｍ）通信モジュール（例えば、ＬＡＮ（ｌｏｃａｌ ａｒｅａ ｎｅｔｗｏｒｋ）通信モジュール、又は電力線通信モジュール）を含むことができる。無線通信モジュールは、近距離通信ネットワーク（例えば、ブルートゥース（登録商標）、ＷｉＦｉ（ｗｉｒｅｌｅｓｓ ｆｉｄｅｌｉｔｙ）ｄｉｒｅｃｔ又はＩｒＤＡ（ｉｎｆｒａｒｅｄ ｄａｔａ ａｓｓｏｃｉａｔｉｏｎ））又は遠距離通信ネットワーク（例えば、レガシーセルラーネットワーク、５Ｇネットワーク、次世代通信ネットワーク、インターネット、又は、コンピュータネットワーク（例えば、ＬＡＮ又はＷＡＮ））を通じて外部の装置と通信できる。
【００７０】
一実施形態において、電子装置６００は、ニューラルネットワーク基盤のアルゴリズムを使用する拡張現実提供装置（例えば、ＡＲ（ａｕｇｍｅｎｔｅｄ ｒｅａｌｉｔｙ）グラス）であってもよい。拡張現実提供装置は、ユーザの顔面に着用され、ユーザに拡張現実サービス及び／又は仮想現実サービスに関するコンテンツを提供する。プロセッサ６１０は、ターゲットアーキテクチャーを有するニューラルネットワークを用いて処理動作を行う。カメラ６３０は、拡張現実コンテンツの生成のための映像を撮影し、プロセッサ６１０は、ターゲットアーキテクチャーを有するニューラルネットワークを用いて映像を処理することで、拡張現実コンテンツを生成できる。例えば、プロセッサ６１０は、カメラ６３０を介して取得された映像から特定のオブジェクトを認識し、認識したオブジェクト領域又はオブジェクト周辺領域に仮想のコンテンツを重複して表現することで、拡張現実コンテンツを生成することができる。
【００７１】
プロセッサ６１０は、図２及び図５を参照して説明したような過程を介してニューラルネットワークのターゲットアーキテクチャーを決定することができる。例えば、プロセッサ６１０は、ニューラルネットワーク（例えば、図１に示す基本ニューラルネットワーク

１２０）の候補アーキテクチャーに対するパラメータに基づいてニューラルネットワーク損失を決定し、候補アーキテクチャーのニューラルネットワークが動作するときに要求されるハードウェアリソースを測定することができる。プロセッサ６１０は、ハードウェアリソース予測モジュール（例えば、図４に示すハードウェアリソース予測モジュール４２０）を用いて候補アーキテクチャーのニューラルネットワークが動作するときに要求されるハードウェアリソースを予測し、測定されたハードウェアリソースと予測されたハードウェアリソースに基づいてハードウェアリソース損失を決定することができる。プロセッサ６１０は、ニューラルネットワーク損失とハードウェアリソース損失に基づいてターゲットアーキテクチャーを決定する。プロセッサ６１０は、候補アーキテクチャー及び候補アーキテクチャーのパラメータによるニューラルネットワーク損失と、候補アーキテクチャーのパラメータによるハードウェアリソース損失の加重和に基づいて最適化損失を決定し、最適化損失を最小にするターゲットアーキテクチャーと候補アーキテクチャーのパラメータを決定することができる。

【００７２】

以上で説明された実施形態は、ハードウェア構成要素、ソフトウェア構成要素、又はハードウェア構成要素及びソフトウェア構成要素の組み合わせで具現される。例えば、本実施形態で説明した装置及び構成要素は、例えば、プロセッサ、コントローラ、ＡＬＵ（ａｒｉｔｈｍｅｔｉｃ ｌｏｇｉｃ ｕｎｉｔ）、デジタル信号プロセッサ（ｄｉｇｉｔａｌ ｓｉｇｎａｌ ｐｒｏｃｅｓｓｏｒ）、マイクロコンピュータ、ＦＰＡ（ｆｉｅｌｄ ｐｒｏｇｒａｍｍａｂｌｅ ａｒｒａｙ）、ＰＬＵ（ｐｒｏｇｒａｍｍａｂｌｅ ｌｏｇｉｃ ｕｎｉｔ）、マイクロプロセッサー、又は命令（ｉｎｓｔｒｕｃｔｉｏｎ）を実行して応答する異なる装置のように、１つ以上の汎用コンピュータ又は特殊目的コンピュータを用いて具現される。処理装置は、オペレーティングシステム（ＯＳ）及びオペレーティングシステム上で実行される１つ以上のソフトウェアアプリケーションを実行する。また、処理装置は、ソフトウェアの実行に応答してデータをアクセス、格納、操作、処理、及び生成する。理解の便宜のために、処理装置は１つが使用されるものとして説明する場合もあるが、当技術分野で通常の知識を有する者は、処理装置が複数の処理要素（ｐｒｏｃｅｓｓｉｎｇ ｅｌｅｍｅｎｔ）及び／又は複数類型の処理要素を含むことが把握する。例えば、処理装置は、複数のプロセッサ又は１つのプロセッサ及び１つのコントローラを含む。また、並列プロセッサ（ｐａｒａｌｌｅｌ ｐｒｏｃｅｓｓｏｒ）のような、他の処理構成も可能である。

【００７３】

ソフトウェアは、コンピュータプログラム、コード、命令、又はそのうちの１つ以上の組み合わせを含み、希望の通りに動作するよう処理装置を構成するか、独立的又は結合的に処理装置を命令することができる。ソフトウェア及び／又はデータは、処理装置によって解釈されるか、処理装置に命令又はデータを提供するために、いずれかの類型の機械、構成要素、物理的装置、仮想装置、コンピュータ記録媒体又は装置、又は送信される信号波に永久的又は一時的に具体化することができる。ソフトウェアはネットワークに連結されたコンピュータシステム上に分散され、分散した方法で格納されるか実行され得る。ソフトウェア及びデータは１つ以上のコンピュータ読み取り可能な記録媒体に格納され得る。

【００７４】

本実施形態による方法は、様々なコンピュータ手段を介して実施されるプログラム命令の形態で具現され、コンピュータ読み取り可能な記録媒体に記録される。記録媒体は、プログラム命令、データファイル、データ構造などを単独又は組み合わせで含む。記録媒体及びプログラム命令は、本発明の目的のために特別に設計して構成されたものでもよく、コンピュータソフトウェア分野の技術を有する当業者にとって公知のものであり使用可能なものであってもよい。コンピュータ読み取り可能な記録媒体の例として、ハードディスク、フロッピー（登録商標）ディスク及び磁気テープのような磁気媒体、ＣＤ－ＲＯＭ、ＤＶＤのような光記録媒体、フロプティカルディスクのような磁気－光媒体、及びＲＯＭ

、ＲＡＭ、フラッシュメモリなどのようなプログラム命令を保存して実行するように特別に構成されたハードウェア装置を含む。プログラム命令の例としては、コンパイラによって生成されるような機械語コードだけでなく、インタプリタなどを用いてコンピュータによって実行される高級言語コードを含む。

【００７５】

上記で説明したハードウェア装置は、本発明に示す動作を実行するために１つ以上のソフトウェアモジュールとして作動するように構成してもよく、その逆も同様である。

【００７６】

上述したように実施形態を例として限定された図面によって説明したが、当技術分野で通常の知識を有する者であれば、上記の説明に基づいて様々な技術的な修正及び変形を適用することができる。例えば、説明された技術が説明された方法と異なる順に実行され、及び／又は説明されたシステム、構造、装置、回路などの構成要素が説明された方法とは異なる形態に結合又は組み合わされてもよく、他の構成要素又は均等物によって置き換え又は代替されたとしても適切な結果を達成することができる。

【００７７】

したがって、他の具現、他の実施形態及び特許請求の範囲と均等なものも後述する特許請求の範囲に属する。

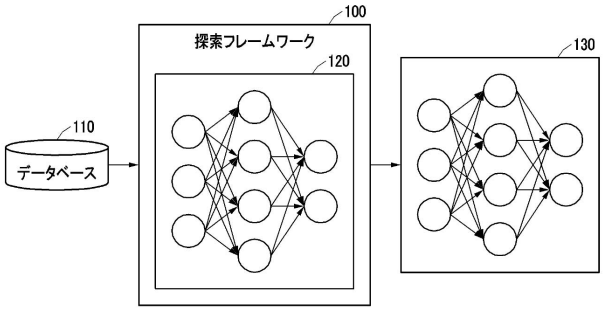
【符号の説明】

【００７８】

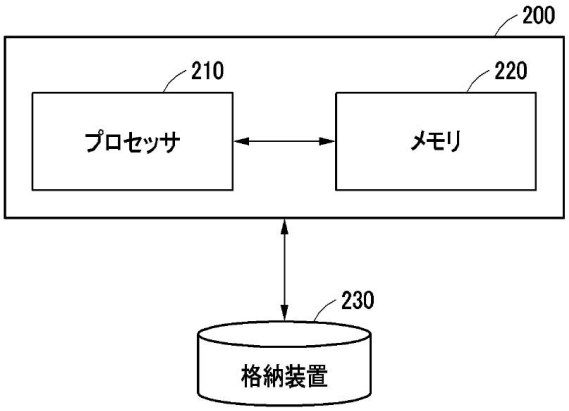
１００：探索フレームワーク
１１０：データベース
１２０：基本ニューラルネットワーク
１３０：ターゲットニューラルネットワーク
２００：探索装置
２１０：プロセッサ
２２０：メモリ
２３０：格納装置
６００：電子装置
６１０：プロセッサ
６２０：メモリ
６３０：カメラ
６４０：センサ
６５０：出力装置
６６０：通信装置
６７０：通信インターフェース

【図面】

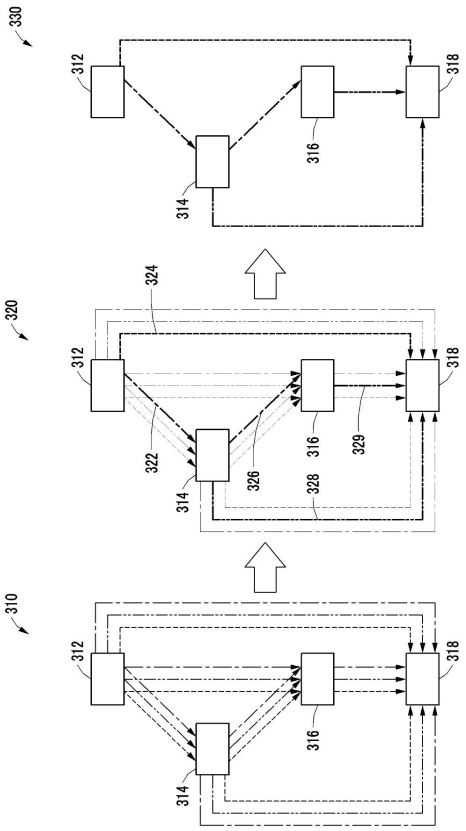
【図１】



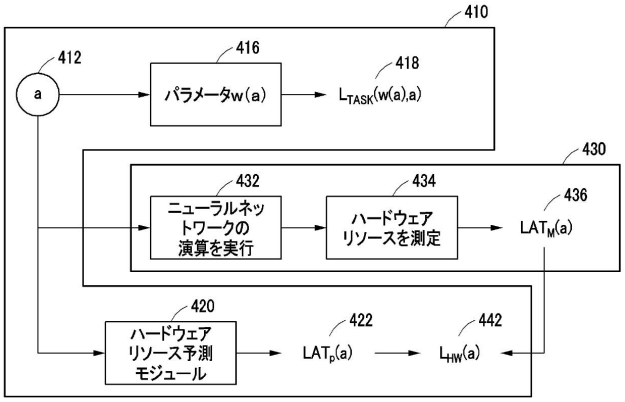
【図２】



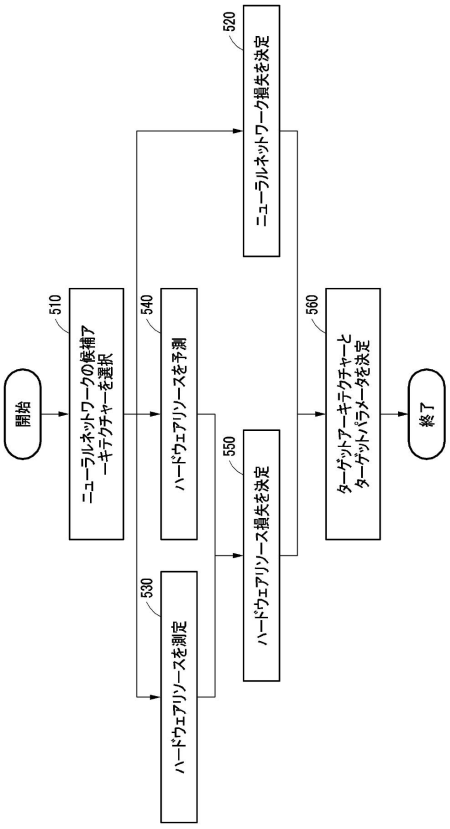
【図３】


330
312
318
316
314
320
324
322
326
328
329
310

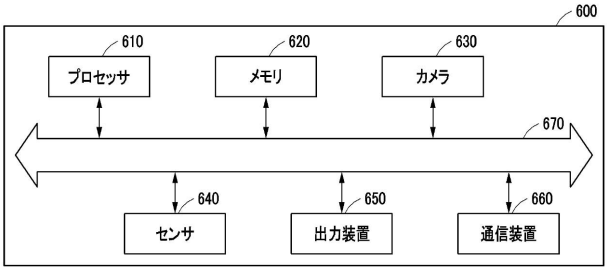
【図４】


410
412
a
416
パラメータw(a)
418
$L_{TASK}(w(a),a)$
430
432
ニューラルネットワークの演算を実行
434
ハードウェアリソースを測定
436
$LAT_M(a)$
420
ハードウェアリソース予測モジュール
422
$LAT_P(a)$
442
$L_{HW}(a)$

【図５】


開始
510
ニューラルネットワークの候補アーキテクチャーを選択
520
ニューラルネットワーク損失を決定
530
ハードウェアリソースを測定
540
ハードウェアリソースを予測
550
ハードウェアリソース損失を決定
560
ターゲットアーキテクチャーとターゲットパラメータを決定
終了

【図６】


600
610
プロセッサ
620
メモリ
630
カメラ
670
640
センサ
650
出力装置
660
通信装置

フロントページの続き

弁理士　宮崎 修
(72)発明者　李 元熙
大韓民国京畿道水原市靈通区三星路１３０ 三星綜合技術院内