(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第6996497号
(P6996497)

(45)発行日　令和4年1月17日(2022.1.17)

(24)登録日　令和3年12月20日(2021.12.20)

(51)国際特許分類

		FI		
G06N	*3/02 (2006.01)*	G06N	3/02	
G06N	*3/10 (2006.01)*	G06N	3/10	
G06N	*3/08 (2006.01)*	G06N	3/08	120

請求項の数　8　（全27頁）

(21)出願番号　特願2018-514172(P2018-514172)
(86)(22)出願日　平成29年3月7日(2017.3.7)
(86)国際出願番号　PCT/JP2017/008988
(87)国際公開番号　WO2017/187798
(87)国際公開日　平成29年11月2日(2017.11.2)
審査請求日　令和2年2月17日(2020.2.17)
(31)優先権主張番号　特願2016-91418(P2016-91418)
(32)優先日　平成28年4月28日(2016.4.28)
(33)優先権主張国・地域又は機関
日本国(JP)

(73)特許権者　000002185
ソニーグループ株式会社
東京都港区港南１丁目７番１号
(74)代理人　110002147
特許業務法人酒井国際特許事務所
(72)発明者　大渕 愉希夫
東京都港区港南１丁目７番１号 ソニー株式会社内
(72)発明者　高松 慎吾
東京都港区港南１丁目７番１号 ソニー株式会社内
(72)発明者　福井 啓
東京都港区港南１丁目７番１号 ソニー株式会社内
(72)発明者　井手 直紀

最終頁に続く

(54)【発明の名称】　情報処理装置、及び情報処理方法

(57)【特許請求の範囲】

【請求項１】

ニューラルネットワークが、複数のハードウェアに係る制約を満たすか否か判定を行う判定部と、

前記ニューラルネットワークの設計を制御し、前記判定部により前記制約を満たさないと判定された場合に、前記制約を満たすように、前記ニューラルネットワークにおけるレイヤーの再配置を行う設計制御部と、

を備える情報処理装置。

【請求項２】

ニューラルネットワークが、複数のハードウェアに係る制約を満たすか否か判定を行う判定部と、

前記判定部により前記制約を満たさないと判定された場合に、前記ニューラルネットワークにおいて前記制約が満たされない部分を提示する警告画面を表示させる表示制御部と、

を備える情報処理装置。

【請求項３】

ニューラルネットワークが、複数のハードウェアに係る制約を満たすか否か判定を行う判定部と、

前記制約に基づいて、前記ニューラルネットワークにおけるレイヤーごとに、前記レイヤーに対応付けられたハードウェアに応じた学習を行う学習部と、

を備える情報処理装置。

【請求項４】
ニューラルネットワークが、複数のハードウェアに係る制約を満たすか否か判定を行う判定部と、
前記判定部により前記制約を満たすと判定されたニューラルネットワークから、ネットワーク構造の異なる別のニューラルネットワークを生成する生成部と、
を備え、
前記判定部は、前記生成部により生成されるニューラルネットワークが前記制約を満たすか否か判定を行い、
前記生成部は、前記判定部により前記制約を満たすと判定されるニューラルネットワークが生成されるまで、前記別のニューラルネットワークの生成を繰り返す、
情報処理装置。
【請求項５】
ニューラルネットワークが、複数のハードウェアに係る制約を満たすか否か判定を行うことと、
前記ニューラルネットワークの設計を制御することと、
前記制約を満たさないと判定した場合に、前記制約を満たすように、前記ニューラルネットワークにおけるレイヤーの再配置を行うことと、
を含む情報処理方法。
【請求項６】
ニューラルネットワークが、複数のハードウェアに係る制約を満たすか否か判定を行うことと、
前記制約を満たさないと判定した場合に、前記ニューラルネットワークにおいて前記制約が満たされない部分を提示する警告画面を表示させることと、
を含む情報処理方法。
【請求項７】
ニューラルネットワークが、複数のハードウェアに係る制約を満たすか否か判定を行うことと、
前記制約に基づいて、前記ニューラルネットワークにおけるレイヤーごとに、前記レイヤーに対応付けられたハードウェアに応じた学習を行うことと、
を含む情報処理方法。
【請求項８】
ニューラルネットワークが、複数のハードウェアに係る制約を満たすか否か判定を行うことと、
前記制約を満たすと判定したニューラルネットワークから、ネットワーク構造の異なる別のニューラルネットワークを生成することと、
生成したニューラルネットワークが前記制約を満たすか否か判定することと、
前記制約を満たすと判定されるニューラルネットワークが生成されるまで、前記別のニューラルネットワークの生成を繰り返すことと、
を含む情報処理方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本開示は、情報処理装置、及び情報処理方法に関する。
【背景技術】
【０００２】
近年、脳神経系の仕組みを模したニューラルネットワークが注目されている。また、ニューラルネットワークを開発するための種々の手法が提案されている。例えば、非特許文献１には、ニューラルネットワークによる学習過程をモニタリングするライブラリが開示されている。
【先行技術文献】

【非特許文献】
【０００３】
【文献】M. Abadi、外３９名、「TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems」、２０１５年１１月９日、［Online］、［平成２８年４月２２日検索］、インターネット〈http://download.tensorflow.org/paper/whitepaper2015.pdf〉
【発明の概要】
【発明が解決しようとする課題】
【０００４】
しかし、非特許文献１に記載のライブラリでは、ニューラルネットワークが単一のハードウェア上で実行されることを前提としており、複数のハードウェアによる処理に適合したニューラルネットワークを設計することは困難であった。
【０００５】
そこで、本開示では、複数のハードウェアによる処理に適合したニューラルネットワークをより効率的に設計することが可能な情報処理装置、及び情報処理方法を提案する。
【課題を解決するための手段】
【０００６】
本開示によれば、複数のハードウェアに係る制約を取得する取得部と、ニューラルネットワークが、前記制約を満たすか否か判定を行う判定部と、を備える情報処理装置が提供される。
【０００７】
また、本開示によれば、ニューラルネットワークが複数のハードウェアに係る制約を満たすか否かの判定結果を受信する受信部と、前記判定結果に基づいて処理を行う処理部と、を備える情報処理装置。
【０００８】
また、本開示によれば、複数のハードウェアに係る制約を取得することと、ニューラルネットワークが、前記制約を満たすか否か判定を行うことと、を含む情報処理方法が提供される。
【発明の効果】
【０００９】
以上説明したように本開示によれば、複数のハードウェアによる処理に適合したニューラルネットワークをより効率的に設計することが可能である。
【００１０】
なお、上記の効果は必ずしも限定的なものではなく、上記の効果とともに、または上記の効果に代えて、本明細書に示されたいずれかの効果、または本明細書から把握され得る他の効果が奏されてもよい。
【図面の簡単な説明】
【００１１】
【図１】ニューラルネットワークによる認識処理を複数のハードウェアで実行する例を説明するための説明図である。
【図２】本開示の第一の実施形態に係る情報処理システムの構成例を説明するための説明図である。
【図３】複数のハードウェアに係る制約を入力するための制約入力画面の一例を示す説明図である。
【図４】同実施形態に係るニューラルネットワークの設計を行うための設計画面の一例を示す説明図である。
【図５】同実施形態に係るサーバ２の構成例を説明するための説明図である。
【図６】同実施形態による情報処理システム１０００の処理フロー例を示すフローチャート図である。
【図７】同実施形態に係る情報処理システム１０００の処理フロー例を示すフローチャー

ト図である。
【図８】同実施形態に係るクライアント端末１に表示される画面例を示す説明図である。
【図９】同実施形態に係るクライアント端末１に表示される画面例を示す説明図である。
【図１０】同実施形態に係るクライアント端末１に表示される画面例を示す説明図である。
【図１１】本開示の第二の実施形態に係るサーバ２－２の構成例を説明するための説明図である。
【図１２】同実施形態の動作例を説明するためのフローチャート図である。
【図１３】同実施形態に係るニューラルネットワーク生成に係る処理フローを説明するためのフローチャート図である。
【図１４】同実施形態に係る生成部２１７による突然変異を用いたネットワーク生成を説明するためのフローチャートである。
【図１５】同実施形態に係る生成部２１７による交叉を用いたネットワーク生成を説明するためのフローチャート図である。
【図１６】ハードウェア構成例を示す説明図である。
【発明を実施するための形態】
【００１２】
以下に添付図面を参照しながら、本開示の好適な実施の形態について詳細に説明する。なお、本明細書及び図面において、実質的に同一の機能構成を有する構成要素については、同一の符号を付することにより重複説明を省略する。
【００１３】
＜＜１．第一の実施形態＞＞
＜１－１．背景＞
＜１－２．構成例＞
＜１－３．動作例＞
＜１－４．効果＞
＜＜２．第二の実施形態＞＞
＜２－１．構成例＞
＜２－２．動作例＞
＜２－３．効果＞
＜＜３．変形例＞＞
＜３－１．変形例１＞
＜３－２．変形例２＞
＜＜４．ハードウェア構成例＞＞
＜＜５．むすび＞＞
【００１４】
＜＜１．第一の実施形態＞＞
＜１－１．背景＞
本開示の第一の実施形態に係る情報処理装置について説明する前に、まず、本実施形態の創作に至った背景を説明する。
【００１５】
近年、人間の脳神経回路を模したモデルであり、人間が持つ学習能力をコンピュータ上で実現しようとするニューラルネットワークが注目されている。上述したとおり、ニューラルネットワークは学習能力を有することを特徴の一つとする。ニューラルネットワークでは、シナプスの結合によりネットワークを形成した人工ニューロン（ノード）が、学習によりシナプスの結合強度を変化させることで、問題に対する解決能力を獲得することが可能である。すなわち、ニューラルネットワークは、学習を重ねることで、問題に対する解決ルールを自動的に推論することができる。
【００１６】
ニューラルネットワークによる学習の例としては、画像認識や音声認識が挙げられる。ニューラルネットワークでは、例えば、手書きの数字パターンを繰り返し学習することで、

入力される画像情報を、０～９の数字のいずれかに分類することが可能となる。ニューラルネットワークの有する上記のような学習能力は、人工知能（Ａｒｔｉｆｉｃｉａｌ Ｉｎｔｅｌｌｉｇｅｎｃｅ）の発展を推し進める鍵としても注目されている。また、ニューラルネットワークが有するパターン認識力は、種々の産業分野における応用が期待される。

【００１７】

一方、ニューラルネットワークを設計する際、ニューラルネットワークによる認識処理におけるパフォーマンス（例えば実行時間や消費電力等）は重要な指標となる。また、パフォーマンスを向上させるため、異なる種類の複数のハードウェアを用いてニューラルネットワークによる認識処理を行うことが考えられる。

【００１８】

例えば、ニューラルネットワークによる認識処理が、処理毎に（例えばニューラルネットワークに含まれるレイヤーごと）に異なるハードウェアにより実行されることで、全体としてのパフォーマンスが向上する場合がある。

【００１９】

図１は、ニューラルネットワークによる認識処理を複数のハードウェアで実行する例を説明するための説明図である。図１に示すニューラルネットワークにおいて、入力レイヤー（Input layer）と出力レイヤー（Output layer）の間には処理Ｐ１～Ｐ３が存在する。上記の処理Ｐ１～Ｐ３が、それぞれ異なるハードウェアで実行されてもよく、例えば、それぞれニューロチップ、ＣＰＵ、及びＧＰＵで実行されてもよい。例えば、ＣＰＵは消費電力が大きく、ＧＰＵは消費電力が小さい場合、処理とハードウェアの組み合わせに応じて、パフォーマンスが異なり得る。

【００２０】

しかし、単一のハードウェアで認識処理が実行されることを前提としたニューラルネットワークの設計ツールを用いて、複数のハードウェアで実行されるニューラルネットワークの設計を行うことは困難であった。また、複数のハードウェアで実行されるニューラルネットを設計できたとしても、それぞれのハードウェアには、制約が存在し得るため、例えば設計されたニューラルネットワークが、当該複数のハードウェアの実行に適合しない恐れがあった。例えば、ハードウェアによっては、処理可能なノード数に制約が存在する場合があり、当該ノード数以上のノード数を有するレイヤーが当該ハードウェアに割り当てられてしまうと、認識処理を実行することができない恐れがあった。

【００２１】

そこで、上記事情を一着眼点にして本実施形態を創作するに至った。本実施形態は、複数のハードウェアに係る制約に基づき、ニューラルネットワークが当該制約を満たすか否かを判定し、例えばユーザに判定結果に基づく警告を提供する。本実施形態によれば、複数のハードウェアによる処理に適合したニューラルネットワークをより効率的に設計することが可能である。以下、上記のような効果を実現するための本開示の第一の実施形態の構成例について説明する。

【００２２】

＜１－２．構成例＞

図２を参照しながら、本開示の第一の実施形態の構成例を説明する。図２は、本開示の第一の実施形態に係る情報処理システムの構成例を説明するための説明図である。本実施形態に係る情報処理システム１０００は、ユーザによるニューラルネットワークの設計のための情報処理システムであり、例えばビジュアルプログラミングにより、ニューラルネットワークを設計することが可能なツールを提供してもよい。

【００２３】

本開示において、ビジュアルプログラミングとは、ソフトウェア開発において、プログラムコードをテキストで記述することなく、視覚的なオブジェクトを用いて作成する手法を指す。ビジュアルプログラミングでは、例えば、ＧＵＩ（Ｇｒａｐｈｉｃａｌ Ｕｓｅｒ Ｉｎｔｅｒｆａｃｅ）上で、オブジェクトを操作することで、プログラムを作成することができる。

【００２４】
図２に示すように、本実施形態に係る情報処理システム１０００は、クライアント端末１、サーバ２、及び通信網５を含み、クライアント端末１とサーバ２とは、通信網５を介して互いに通信を行えるように接続される。
【００２５】
クライアント端末１は、ユーザがニューラルネットワークの設計を行うための情報処理装置である。例えば、クライアント端末１は、ビジュアルプログラミングによりニューラルネットワークの設計を行うための設計画面をユーザに提示（表示）してもよい。また、クライアント端末１は、サーバ２からニューラルネットワークが複数のハードウェアに係る制約を満たすか否かの判定結果を受信し、当該判定結果に基づいて、表示制御処理を行ってもよい。
【００２６】
サーバ２は、クライアント端末１にニューラルネットワークの設計を行うための設計画面を提供し、クライアント端末１を介したユーザの入力に基づいて、ニューラルネットワークに係るプログラムを作成する情報処理装置である。また、サーバ２は、複数のハードウェアに係る制約に基づいて、ニューラルネットワークが当該制約を満たすか否か判定を行い、判定結果をクライアント端末１に提供する。また、ニューラルネットワークが当該制約を満たさない場合、サーバ２は当該ニューラルネットワークに係るプログラムを作成しなくてもよい。
【００２７】
通信網５は、通信網５に接続されている装置、またはシステムから送信される情報の有線、または無線の伝送路である。例えば、通信網５は、インターネット、電話回線網、衛星通信網等の公衆回線網や、Ｅｔｈｅｒｎｅｔ（登録商標）を含む各種のＬＡＮ（Ｌｏｃａｌ Ａｒｅａ Ｎｅｔｗｏｒｋ）、WAN（Ｗｉｄｅ Ａｒｅａ Ｎｅｔｗｏｒｋ）等を含んでもよい。また、通信網５は、ＩＰ－ＶＰＮ（Ｉｎｔｅｒｎｅｔ Ｐｒｏｔｏｃｏｌ－Ｖｉｒｔｕａｌ Ｐｒｉｖａｔｅ Ｎｅｔｗｏｒｋ）等の専用回線網を含んでもよい。
【００２８】
本実施形態に係る情報処理システム１０００によれば、ユーザにより設計されるニューラルネットワークが、複数のハードウェアに係る制約を満たすか否かの判定が行われ、判定結果がユーザに提供される。また、本実施形態に係る情報処理システム１０００によれば、複数のハードウェアに係る制約を満たすニューラルネットワークに係るプログラムが作成されるようなニューラルネットワークの設計ツールが提供される。
【００２９】
（クライアント端末）
続いて、本実施形態に係るクライアント端末１について詳細に説明する。図２に示すように、本実施形態に係るクライアント端末１は、制御部１０、通信部１２、表示部１４、及び操作部１６を備える情報処理装置である。なお、クライアント端末１は、例えばＰＣ（Personal Computer）、タブレットＰＣ等であってもよい。
【００３０】
制御部１０は、クライアント端末１の各構成を制御する。例えば、制御部１０は、通信部１２による通信を制御する通信制御部としての機能を有する。係る構成により、通信部１２は、例えば各種画面や、ニューラルネットワークが複数のハードウェアに係る制約を満たすか否かの判定結果をサーバ２から受信することが出来る。
【００３１】
また、制御部１０は、表示部１４による表示の制御処理を行う表示制御部としての機能を有する。例えば、制御部１０は、各種画面を表示部１４に表示させてもよい。図３、図４を参照して、制御部１０が表示部１４に表示させる画面の例を説明する。
【００３２】
図３は、複数のハードウェアに係る制約を入力するための制約入力画面の一例を示す説明図である。図３に示すように、制約入力画面は、例えば、認識処理に用いられるハードウ

ェアを示すハードウェア入力フォームＧ１１～Ｇ１４と、ハードウェア間の通信速度を示す通信速度入力フォームＧ１５～Ｇ１７を含む。ユーザは、ハードウェア入力フォームＧ１１～Ｇ１４を用いて、ハードウェア（ＨＷ：Ｈａｒｄｗａｒｅ）の種類（Ｔｙｐｅ）の選択や、ハードウェアの情報（例えば演算性能等）の入力を行うことが可能である。例えば、ハードウェア入力フォームＧ１１～Ｇ１４によりそれぞれハードウェアＨＷ１～ＨＷ４が選択されることで、予め用意された、ハードウェアに係る制約が設定され得る。また、ユーザがハードウェア入力フォームＧ１１～Ｇ１４を用いて各ハードウェアの情報をさらに入力することで、制約をカスタマイズすることも可能である。

【００３３】

また、ユーザは、通信速度入力フォームＧ１５～Ｇ１７を用いて、ハードウェア間の通信速度の入力（カスタマイズ）を行うことが可能である。なお、ハードウェア入力フォームＧ１１～Ｇ１４の間の接続関係は、ユーザ操作により変更可能であってもよい。

【００３４】

図４は、ニューラルネットワークの設計を行うための設計画面の一例を示す説明図である。図４に示す設計画面は、通信部１２を介して、サーバ２から受信されてもよい。図４に示すように、設計画面は、ニューラルネットワークにおける複数のレイヤーＧ２０～Ｇ３０が配置される。配置されたレイヤーＧ２０～Ｇ３０は、例えばデータの取り込み、データの処理、データの出力をそれぞれ意味していてもよい。ユーザは、図４に示すような設計画面を用いて、レイヤーの追加、削除、配置変更等を行い、ニューラルネットワークを設計することが出来る。

【００３５】

また、入力レイヤーＧ２０と出力レイヤーＧ３０の間のレイヤーＧ２１～Ｇ２９は、配置された順に逐次、または並列に処理されてもよい。例えば、レイヤーＧ２４～Ｇ２６の処理と、レイヤーＧ２７～Ｇ２８の処理とは、並列に処理されてもよい。

【００３６】

また、制御部１０（処理部）は、表示部１４を制御して、サーバ２から受信される判定結果に基づく表示制御処理を行ってもよい。例えば、サーバ２により、ニューラルネットワークが複数のハードウェアに係る制約を満たさないと判定された場合に、制御部１０（表示制御部）は、制約が満たされないことを示す警告画面を表示させてもよい。また、制御部１０が表示させる警告画面は、ニューラルネットワークにおいて当該制約を満たさない部分を提示する画面であってもよい。係る構成により、ユーザは複数のハードウェアに係る制約を満たすようにニューラルネットワークをより効率的に設計することが可能となる。

【００３７】

通信部１２（受信部）は、制御部１０により制御されて、他の装置との間の通信を仲介する通信インタフェースである。通信部１２は、任意の無線通信プロトコルまたは有線通信プロトコルをサポートし、例えば図２に示す通信網５を介して他の装置との間の通信接続を確立する。例えば、通信部１２はニューラルネットワークの設計を行うための設計画面や、ニューラルネットワークが複数のハードウェアに係る制約を満たすか否かの判定結果をサーバ２から受信する。また、通信部１２は、表示部１４に表示される各種画面に対するユーザの入力に係る情報をサーバ２に送信させる。

【００３８】

表示部１４は制御部１０に制御されて、各種画面を表示するディスプレイである。例えば、表示部１４は、上述した制約入力画面、設計画面、警告画面等を表示してもよい。なお、表示部１４は、例えば、ＣＲＴ（Ｃａｔｈｏｄｅ　Ｒａｙ　Ｔｕｂｅ）ディスプレイ装置、液晶ディスプレイ（ＬＣＤ：Ｌｉｑｕｉｄ　Ｃｒｙｓｔａｌ　Ｄｉｓｐｌａｙ）装置、ＯＬＥＤ（Ｏｒｇａｎｉｃ　Ｌｉｇｈｔ　Ｅｍｉｔｔｉｎｇ　Ｄｉｏｄｅ）装置により実現されてもよい。

【００３９】

操作部１６は、ユーザの入力を受け付け、制御部１０に提供する。例えば、ユーザは、操作部１６を操作して、複数のハードウェアに係る制約のカスタマイズや、ニューラルネッ

トワークの設計を行うための入力を行ってもよい。なお、操作部１６は、例えばマウス、キーボード、タッチパネル、ボタン、スイッチ、視線入力装置、ジェスチャ入力装置、音声入力装置などにより実現されてもよい。

【００４０】

（サーバ）

以上、本実施形態に係るクライアント端末１の構成例を説明した。続いて、図５を参照して、本実施形態に係るサーバ２の構成例を説明する。図５は、本実施形態に係るサーバ２の構成例を説明するための説明図である。図５に示すように、サーバ２は、制御部２０、通信部２２、及び記憶部２４を備える情報処理装置である。

【００４１】

制御部２０は、サーバ２の各構成を制御する。また、制御部２０は、図５に示すように、通信制御部２０１、取得部２０２、判定部２０３、設計制御部２０４、学習部２０５、及び認識部２０６としても機能する。

【００４２】

通信制御部２０１は、通信部２２による通信を制御する。例えば、通信制御部２０１は、通信部２２を制御して、ニューラルネットワークの設計画面、判定部２０３による判定結果等を、クライアント端末１に送信させてもよい。また、通信制御部２０１は、通信部２２を制御して、複数のハードウェアに係る制約のカスタマイズや、ニューラルネットワークの設計を行うためのユーザの入力に係る情報を受信してもよい。

【００４３】

取得部２０２は、複数のハードウェアに係る制約を取得する。例えば、取得部２０２は、記憶部２４から、複数のハードウェアに係る制約を取得してもよい。また、取得部２０２は、通信部２２を介して受信される複数のハードウェアに係る制約のカスタマイズを行うためのユーザの入力に基づいて、複数のハードウェアに係る制約を取得してもよい。また、取得部２０２は、ハードウェアの選択に係るユーザの入力と、記憶部２４に予め記憶されたハードウェアに係る制約に基づいて、複数のハードウェアに係る制約を取得してもよい。

【００４４】

以下では、図１を参照して説明した、ニューロチップ、ＣＰＵ、及びＧＰＵに係る制約の例について説明を行うが、ハードウェアに係る制約は以下の例に限定されない。

【００４５】

取得部２０２により取得される、複数のハードウェアに係る制約は、例えば後述する判定部２０３による判定に用いられる制約であってもよい。判定に用いられる制約の例としては、ハードウェア間の接続に係る制約や、ハードウェア間の通信速度に係る制約、ハードウェアの処理能力に係る制約等であってもよい。以下に判定部２０３による判定に用いられる制約の例を示す。

【００４６】

・ニューロチップはセンサ（入力レイヤー）に接続される
・ニューロチップが処理可能なノード数は１０以下である
・ニューロチップが処理可能なレイヤーはコンボリューションレイヤーのみである
・ニューロチップとＣＰＵとの間の通信速度は所定の速度である
・ＣＰＵが利用可能なＲＡＭは所定値以下である
・ＣＰＵとＧＰＵとの間の通信速度は所定の速度である
・ＧＰＵはニューロチップと直接接続できない

【００４７】

また、取得部２０２により取得される、複数のハードウェアに係る制約は、例えば後述する学習部２０５がハードウェアに応じた学習を行うために用いられる制約であってもよい。学習を行うために用いられる制約の例としては、ハードウェアの特性に係る制約や、演算の種類に係る制約等であってもよい。以下に、学習部２０５による学習に用いられる制約の例を示す。

【００４８】
・ニューロチップのニューロンの特性はスパイキングである
・ＣＰＵは整数演算のみ可能（浮動小数点演算不可能）である
【００４９】
判定部２０３は、ニューラルネットワークが、取得部２０２により取得される複数のハードウェアに係る制約を満たすか否か判定を行う。判定部２０３は、例えば、ユーザの入力に基づいて設計制御部２０４により設計されたニューラルネットワークが、上記制約を満たすか否か判定を行う。
【００５０】
なお、判定部２０３は、予め設定される、またはユーザにより入力される、所定の処理時間にさらに基づき、ニューラルネットワークが、取得部２０２により取得される複数のハードウェアに係る制約を満たすか否か判定を行ってもよい。係る場合、判定部２０３は、ニューラルネットワークが取得部２０２により取得される複数のハードウェアに係る制約を満たし、かつ当該所定の処理時間内に処理が完了すると判定された場合に、制約を満たすと判定してもよい。
【００５１】
また、判定部２０３は、ユーザの入力に基づいて、設計制御部２０４によりニューラルネットワークが変更された（例えばレイヤーの追加、削除、配置変更等が発生した）場合に、変更後のニューラルネットワークが上記制約を満たすか否か判定を行ってもよい。なお、判定部２０３は判定結果を通信制御部２０１と設計制御部２０４に提供してもよい。
【００５２】
設計制御部２０４は、通信部２２を介して取得されるユーザの入力に基づいて、ニューラルネットワークの設計を制御する。例えば、設計制御部２０４は、図４を参照して説明した設計画面を生成し、通信部２２を介してクライアント端末１に提供してもよい。また、設計制御部２０４は、ユーザの入力に基づいて、ニューラルネットワークにおけるレイヤーの配置を制御してもよい。
【００５３】
また、設計制御部２０４は、通信部２２を介して取得されるユーザの入力に基づいて、ニューラルネットワークにおけるレイヤーと、ハードウェアの対応付けを行う。対応付けに係るユーザの入力は、ニューラルネットワークの設計画面において範囲選択により行われてもよいし、レイヤーの処理順における順番の指定により行われてもよい。
【００５４】
また、設計制御部２０４は、設計制御部２０４により設計されたニューラルネットワークが、判定部２０３により、制約を満たすと判定された場合に、当該ニューラルネットワークを構築するプログラムを作成してもよい。係る構成によれば、複数のハードウェアに係る制約を満たし、当該複数のハードウェアにより認識を実行可能なニューラルネットワークを構築するプログラムが生成される。
【００５５】
また、設計制御部２０４は、設計制御部２０４により設計されたニューラルネットワークが、判定部２０３により、制約を満たさないと判定された場合に、当該制約を満たすように、ニューラルネットワークの変更に係るレイヤーの再配置を行ってもよい。例えば、設計制御部２０４は、追加されたレイヤーまたは配置変更されたレイヤーを、当該レイヤーが現在対応付けられているハードウェアから通信可能なハードウェアに再配置してもよい。また、設計制御部２０４は、再配置後のニューラルネットワークが制約を満たすか否かの判定結果を判定部２０３から取得して、制約が満たされるまで、再配置を繰り返してもよい。係る構成によれば、より効率的に、複数のハードウェアに係る制約を満たしたニューラルネットワークを設計することが可能となる。
【００５６】
学習部２０５は、設計制御部２０４により設計されたニューラルネットワークの学習を行う。学習部２０５は、例えば取得部２０２により取得される制約に基づいて、ニューラル

根とワークにおけるレイヤーごとに、レイヤーに対応付けられたハードウェアに応じた学習を行ってもよい。係る構成によれば、後述する認識部２０６により行われる認識の実行パフォーマンスが向上し得る。

【００５７】

例えば、ニューロチップと対応付けられたレイヤーに係る学習は、当該ニューロチップの特性に応じた学習手法により行われてもよい。ニューロチップの特性に応じた学習手法は限定されないが、例えばニューロチップの特性がスパイキングである場合には、下記非特許文献２に記載される学習手法を用いることも可能である。

【００５８】

非特許文献２：O. Peter、外４名、「Real-time classification and sensor fusion with a spiking deep belief network」、２０１３年, Neuromorphic Engineering 7: 178.

【００５９】

また、整数演算のみ可能なハードウェアと対応付けられたレイヤーに係る学習は、整数演算のみで処理可能となるような学習手法により行われてもよい。係る学習手法は限定されないが、例えば下記非特許文献３に記載される学習手法を用いることも可能である。

【００６０】

非特許文献３：M. Courbariaux、外２名、「BinaryConnect: Training Deep Neural Networks with binary weights during propagations TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems」、２０１５年１１月１２日、［Online］、［平成２８年４月２１日検索］、インターネット〈http://arxiv.org/pdf/1511.00363.pdf〉

【００６１】

また、浮動小数点演算可能なハードウェアと対応付けられたレイヤーに係る学習は、浮動小数点演算を利用可能な多様な学習手法により行われ得る。

【００６２】

なお、学習部２０５は、記憶部２４に記憶される学習データに基づいて学習を行ってもよいし、通信部２２を介して外部から取得される学習データに基づいて学習を行ってもよい。

【００６３】

認識部２０６は、学習部２０５による学習に基づき、認識を実行する。認識部２０６はニューラルネットワークにおけるレイヤーごとに対応付けられたハードウェアで、フィードフォワード計算を行うことで認識を実行してもよい。なお、認識部２０６は、記憶部２４に記憶されるデータに対して認識を行ってもよいし、通信部２２を介して外部から取得されるデータに対して認識を行ってもよい。

【００６４】

通信部２２は、他の装置との間の通信を仲介する通信インタフェースである。通信部２２は、任意の無線通信プロトコルまたは有線通信プロトコルをサポートし、例えば図２に示した通信網５を介して他の装置との間の通信接続を確立する。それにより、例えば、サーバ２が通信網５に接続されたクライアント端末１からユーザの入力等を受信し、及びクライアント端末１に設計画面、及び判定部２０３による判定結果等を送信することが可能となる。

【００６５】

記憶部２４は、サーバ２の各構成が機能するためのプログラムやパラメータを記憶する。また、記憶部２４は、ハードウェアに係る制約、学習データ、認識用のデータ等を記憶してもよい。

【００６６】

なお、上記ではサーバ２が学習部２０５、認識部２０６を備えて、学習と認識を行う例を説明したが、本実施形態は上記に限定されない。例えば、学習、及び認識は、通信網５に接続される他の装置で行われてもよく、学習、及び認識のそれぞれが異なる装置により行われてもよい。係る場合、取得部２０２は、通信部２２を介して、当該認識を行う装置か

ら、当該認識を行う装置が有する複数のハードウェアに係る制約を取得してもよい。
【００６７】
＜１－３．動作例＞
以上、本実施形態による情報処理システム１０００の構成例について説明した。続いて、本実施形態による情報処理システム１０００の動作例について、図６～１０を参照して説明する。以下では、まず情報処理システム１０００の処理フローについて図６、図７を参照して説明した後、本実施形態においてクライアント端末１に表示される画面遷移例について図８～１０を参照して説明する。
【００６８】
（処理フロー例１）
図６は、本実施形態による情報処理システム１０００の処理フロー例を示すフローチャート図である。図６に示すフローチャートは、本実施形態に係る動作のうち、特にニューラルネットワークの設計に係る処理フローを示す。
【００６９】
まず、ユーザが図３を参照して説明した制約入力画面を用いて、ハードウェアに係る制約を入力し、取得部２０２が制約を取得する（Ｓ１１０１）。
【００７０】
続いて、ユーザにより、ハードウェアとレイヤーの対応に関するハードウェアごとの範囲が設定される（Ｓ１１０２）。なお、ハードウェアごとの範囲の設定は、ニューラルネットワークの設計画面において、表示上の範囲を選択することで行われてもよいし、レイヤーの処理順における順番の指定により行われてもよい。
【００７１】
続いて、ユーザにより、ニューラルネットワークの設計画面を用いたニューラルネットワークの設計が行われる（Ｓ１１０３）。
【００７２】
続いて、判定部２０３により、設計されたニューラルネットワークが、複数のハードウェアに係る制約を満たすか否かの判定が行われる（Ｓ１１０４）。判定部２０３により、ニューラルネットワークが制約を満たさないと判定された場合（Ｓ１１０４：ＮＯ）、制約を満たさない箇所（部分）を提示する警告画面が表示され（Ｓ１１０５）、処理はステップＳ１１０３に戻る。一方、判定部２０３により、ニューラルネットワークが制約を満たすと判定された場合（Ｓ１１０４：ＹＥＳ）、設計処理は終了する。
【００７３】
（処理フロー例２）
図７は、本実施形態による情報処理システム１０００の処理フロー例を示すフローチャート図である。図７に示すフローチャートは、本実施形態に係る動作のうち、特にニューラルネットワークの設計の変更に係る処理フローを示す。例えば以下に説明する処理フローは、図６に示したフローチャートの処理により設計されたニューラルネットワークの設計を変更する際の処理フローであってもよい。
【００７４】
図７に示すステップＳ１２０１、Ｓ１２０２の処理は、図６を参照して説明したステップＳ１１０１、Ｓ１１０２の処理と同様であるため、説明を省略する。
【００７５】
続いて、ユーザにより、ニューラルネットワークの設計の変更（例えば、レイヤーの追加、削除、配置変更）が行われる（Ｓ１２０３）。続いて、判定部２０３により、変更後のニューラルネットワークが、複数のハードウェアに係る制約を満たすか否かの判定が行われる（Ｓ１２０４）。
【００７６】
判定部２０３により、ニューラルネットワークが制約を満たさないと判定された場合（Ｓ１２０４：ＮＯ）、設計制御部２０４による自動的な再配置が可能であるか否かが判定される（Ｓ１２０５）。自動的な再配置が可能ではない場合（Ｓ１２０５：ＮＯ）、処理は

ステップＳ１２０３に戻る。
【００７７】
一方、自動的な再配置が可能である場合、（Ｓ１２０５：ＮＯ）、設計制御部２０４は、自動的に再配置を行う（Ｓ１２０６）。また、自動的な再配置が行われたニューラルネットワークが設計画面に表示される（Ｓ１２０７）
【００７８】
また、ステップＳ１２０４において制約を満たすと判定された場合、設計が変更されたニューラルネットワークが設計画面に表示される（Ｓ１２０７）。
【００７９】
設計が変更されたニューラルネットワーク、または自動的な再配置が行われたニューラルネットワークを設計画面において確認したユーザにより、設計終了の操作入力が行われると（Ｓ１２０８：ＹＥＳ）、処理は終了する。一方、ユーザが続けて設計の変更行う場合（Ｓ１２０８：ＮＯ）、処理はステップＳ１２０３に戻る。
【００８０】
（画面遷移例）
以上、本実施形態による情報処理システム１０００の処理フローを説明した。続いて、図７を参照して説明した処理フローにおいて、クライアント端末１に表示される画面の遷移例を図３、４及び図８～１０を参照して説明する。図８～１０は、本実施形態に係るクライアント端末１に表示される画面例を示す説明図である。なお、図８～１０は図４を参照して説明した設計画面に含まれるニューラルネットワークの設計変更に係る画面遷移の一例である。また、以下では、図７に示した処理ステップを適宜参照しながら説明を行う。
【００８１】
まず、図７のステップＳ１２０１において、図３を参照して説明した制約入力画面が表示される。なお、ハードウェアに係る制約をユーザが変更する必要のない場合には、ユーザの入力なしに、取得部２０２が例えば記憶部２４から制約を取得してもよい。
【００８２】
続いて、図７のステップＳ１２０２において、図８に示す設計画面のように、ハードウェアごとの範囲選択が行われる。ここで、図８に示す範囲Ｇ３１～Ｇ３４に含まれるレイヤーは、設計制御部２０４により、各範囲に対応するハードウェアと対応付けられる。なお、図８に示す範囲Ｇ３１、Ｇ３２、Ｇ３３、Ｇ３４は、例えば、それぞれ図３におけるハードウェアＨＷ１、ＨＷ２、ＨＷ４、ＨＷ３に対応している。
【００８３】
続いて、図７のステップＳ１２０３において、図９に示す画面のように、ニューラルネットワークの変更が行われる。図９に示す例では、レイヤーＧ４２がレイヤーＧ４１とレイヤーＧ４３の間に追加されている。なお、図８を参照すれば、図９において新たに追加されるレイヤーＧ４２は、図８に示す範囲Ｇ３３に含まれるため、ＨＷ４に対応付けられる。
【００８４】
続いて、図７のステップＳ１２０４において、制約を満たすと判定された場合、ステップＳ１２０７において、図１０に示すように、変更によりレイヤーＧ５２が新たに追加されたニューラルネットワークを含む設計画面が表示される。一方、図７のステップＳ１２０４において、制約を満たさないと判定された場合、図４に示した画面のように、変更前のニューラルネットワークを含む設計画面が表示される（表示が図４の設計画面に戻る）。
【００８５】
＜１－４．効果＞
以上説明したように、本開示の第一の実施形態によれば、複数のハードウェアに係る制約に基づき、ニューラルネットワークが当該制約を満たすか否かを判定し、例えばユーザに判定結果に基づく警告画面を提供する。係る構成により、複数のハードウェアによる処理に適合したニューラルネットワークをより効率的に設計することが可能である。また、本実施形態によれば、変更されたニューラルネットワークが当該制約を満たさない場合には、制約を満たすように自動的に再配置が行われることで、制約を満たしたニューラルネッ

トワークの設計を支援することが可能である。
【００８６】
＜＜２．第二の実施形態＞＞
以上、本開示の第一の実施形態を説明した。続いて、本開示の第二の実施形態を説明する。本開示の第二の実施形態は、設計されたニューラルネットワークの評価結果に基づいて、ネットワーク構造の異なる別のニューラルネットワークを生成する。また、本開示の第二の実施形態は、生成されたニューラルネットワークの評価結果に基づいて、評価済のニューラルネットワークに係るパレート最適解を更新する。さらに、本開示の第二の実施形態は、ネットワークの生成とパレート最適解の更新を繰り返すことで、効率の良いネットワーク構造を探索することが可能である。
【００８７】
以下、上記のような効果を実現するための本開示の第二の実施形態の構成例について説明する。なお、以下の説明においては、第一の実施形態との差異について説明し、第一の実施形態と共通するクライアント端末１、及び通信網５については説明を省略する。
【００８８】
＜２－１．構成例＞
図１１は、本開示の第二の実施形態に係るサーバ２－２の構成例を説明するための説明図である。図１１に示すように、サーバ２－２は、制御部２１、通信部２２、及び記憶部２４を備える情報処理装置である。図１１に示すように、本実施形態に係るサーバ２―２は、制御部２１の機能構成が図５の制御部２０の機能構成と一部異なる点で、図５のサーバ２と異なる。なお、図１１に示す各構成のうち、図５に示した各構成と実質的に同様の構成については同一の符号を付してあるため、説明を省略する。以下では、本実施形態に係る制御部２１が有する判定部２１３、生成部２１７、及び評価部２１８、としての機能について説明する。
【００８９】
本実施形態に係る判定部２１３は、第一の実施形態で説明した判定部２０３と同様に、ユーザの入力に基づいて設計制御部２０４により設計された、または変更されたニューラルネットワークが、取得部２０２により取得される制約を満たすか否か判定を行う。また、本実施形態に係る判定部２１３は、後述する生成部２１７により生成されるニューラルネットワークが、取得部２０２により取得される制約を満たすか否か判定を行う。
【００９０】
生成部２１７は、元となるネットワークからネットワーク構造の異なる別のニューラルネットワークを生成する機能を有する。例えば、生成部２１７は、ユーザによる入力に基づいて設計され、判定部２１３により制約を満たすと判定されたニューラルネットワーク（以降、シードネットワークとも呼ぶ）から、ネットワーク構造の異なる別のニューラルネットワークを生成してもよい。また、生成部２１７は、パレート最適解に係るニューラルネットワークから、ネットワーク構造の異なる別のニューラルネットワークを生成してもよい。
【００９１】
本実施形態に係る生成部２１７によるニューラルネットワークの生成は、例えば、突然変異や交叉（または、交差、とも呼ぶ）などを含む遺伝的操作により実現されてもよい。ここで、上記の突然変異とは、生物に見られる遺伝子の突然変異をモデル化したものであってよい。すなわち、本実施形態では、ネットワークを構成する各レイヤーを遺伝子と見立て、レイヤーを突然変異させることで、ネットワーク構造の異なる別のニューラルネットワークを生成することができる。例えば、本実施形態に係る突然変異は、レイヤーの挿入、レイヤーの削除、レイヤー種類の変更、パラメータの変更、グラフ分岐、またはグラフ分岐の削除のうち少なくともいずれか一つを含んでもよい。
【００９２】
また、上記の交叉とは、生物の交配における染色体の部分的交換をモデル化したものであってよい。すなわち、本開示に係る情報処理方法では、２つのネットワークのレイヤー構

成を部分的に交換することで、上記の別のニューラルネットワークを生成することができる。

【００９３】

なお、上記では、遺伝的操作により別のニューラルネットワークを生成する場合を例に説明したが、本実施形態に係るニューラルネットワークの生成方法は、係る例に限定されない。本実施形態に係る別のニューラルネットワークの生成は、例えば、入力されたネットワークのネットワーク構造を変化させるニューラルネットワークを用いて実現されてもよい。ニューラルネットワークの生成には、上記の例を含む種々の方法が適用され得る。

【００９４】

また、生成部２１７は、判定部により制約を満たすと判定されたニューラルネットワークが生成されるまで、別のニューラルネットワークの生成を繰り返してもよい。係る構成によれば、生成部２１７は、複数のハードウェアに係る制約を満たすようなニューラルネットワークを生成することが可能となる。

【００９５】

評価部２１８は、生成されたニューラルネットワークの評価結果を取得する機能を有する。評価部２１８は、例えば、生成されたニューラルネットワークを認識部２０６に実行させ、上記の評価結果を取得してもよい。なお、評価部２１８による評価結果の取得は上記に限定されず、通信網５を介して接続される各種のデバイスに生成されたニューラルネットワークを実行させ、評価結果を取得してもよい。

【００９６】

また、評価部２１８が取得する評価結果には、生成されたニューラルネットワークに係る演算量、及び学習誤差またはヴァリデーション誤差（以下、まとめて誤差と表現することがある）のうち少なくとも一方が含まれてよい。評価部２１８は、生成されたニューラルネットワークのネットワーク構造に基づいて、上記の演算量を取得することができる。なお、本実施形態に係る評価結果は、上記に限定されず、例えば、ハードウェアに係る使用メモリ量、発熱量、消費電力量、演算量から算出されるハードウェアのトータルコストや、サーバ費用などを含むトータルサービスコスト等を含んでもよい。評価部２１８は、予め記憶されたハードウェアやサービスに係る情報を基に、上記の値を算出することができる。

【００９７】

また、評価部２１８は、生成されたニューラルネットワークの評価結果に基づいて、評価済のニューラルネットワークに係るパレート最適解を更新する機能を有する。すなわち、評価部２１８は、生成部２１７が生成したニューラルネットワークの評価結果を取得し、当該評価結果に基づいてパレート最適解の更新を繰り返し実行する。

【００９８】

＜２－２．動作例＞

以上、本実施形態に係るサーバ２－２の構成例について説明した。続いて、本実施形態の動作例について、図１２～１５を参照して説明する。

【００９９】

図１２は、本実施形態の動作例を説明するためのフローチャート図である。図７に示すステップＳ２１００、Ｓ２２００の処理は、図６を参照して説明したステップＳ１１０１、Ｓ１１０２の処理と同様であるため、説明を省略する。

【０１００】

続いて、ユーザによるニューラルネットワークの設計と、判定部２１３による判定が行われる（Ｓ２３００）。なお、ステップＳ２３００の処理は、例えば、図６を参照して説明したステップＳ１１０３～Ｓ１１０５の処理、または図７を参照して説明したステップＳ１２０３～Ｓ１２０８の処理と同様の処理を含んでもよい。

【０１０１】

続いて、生成部２１７は、ステップＳ２３００において、判定部２１３により制約を満たすと判定されたニューラルネットワーク（シードネットワーク）から、ネットワーク構造

の異なる別のニューラルネットワークを生成する（Ｓ２４００）。なお、ステップＳ２４００におけるニューラルネットワークの生成に係る詳細な処理フローについては図１３～１５を参照して後述する。

【０１０２】

続いて、評価部２１８が、生成されたニューラルネットワークの評価結果を取得する（Ｓ２５００）。評価部２１８のネットワーク構造の探索が終了していない場合（Ｓ２６００：ＮＯ）、処理はステップＳ２７００に進む。

【０１０３】

ステップＳ２７００において、評価部２１８は、生成されたニューラルネットワークの評価結果に基づいて、評価済のニューラルネットワークに係るパレート最適解を更新する。続いて、処理はステップＳ２４００に戻り、生成部２１７は、当該パレート最適解に係るニューラルネットワークから、ネットワーク構造の異なる別のニューラルネットワークを生成する。

【０１０４】

上記のステップＳ２４００～Ｓ２７００の処理により、ニューラルネットワークの生成と、パレート最適解の更新が繰り返し実行されて、ネットワーク構造の探索が終了すると（Ｓ２６００：ＹＥＳ）、処理は終了する。

【０１０５】

なお、ネットワーク構造の探索が終了した際、例えば誤差が最小（最高性能）であるニューラルネットワーク、または演算量が最小であるニューラルネットワーク、または中間解であるニューラルネットワークが得られてもよい。中間解の定義は、条件に応じて適宜設計されてよい。なお、上記のようなニューラルネットワークがユーザに提示されて、ユーザがいずれかのニューラルネットワークを選択してもよい。

【０１０６】

以上、本実施形態の動作例の全体的な処理フローについて説明した。続いて、図１２に示すステップＳ２４００のニューラルネットワーク生成に係る処理フローについて、図１３を参照して説明する。図１３はニューラルネットワーク生成に係る処理フローを説明するためのフローチャート図である。

【０１０７】

図１３を参照すると、まず、生成部２１７は、元となるニューラルネットワークに適用する別のニューラルネットワークの生成方法をランダムで決定する（Ｓ２４１０）。この際、元となるニューラルネットワークは、ユーザによる入力に基づいて設計され、判定部２１３により制約を満たすと判定されたシードネットワークであってもよい。また、元となるニューラルネットワークは、評価部２１８が更新したパレート最適解に係るニューラルネットワークから生成部２１７がランダムに選択したネットワークであってもよい。

【０１０８】

続いて、生成部２１７は、ステップＳ２４１０で選択した生成方法に基づいて、元となるニューラルネットワークから、ネットワーク構造の異なる別のニューラルネットワークを生成する。図１３に示す一例を参照すると、本実施形態に係る生成部２１７は、元となるニューラルネットワークを突然変異させることで、上記の別のニューラルネットワークを生成してもよい（Ｓ２４２０）。

【０１０９】

また、生成部２１７は、元となるニューラルネットワークを交叉させることで、上記の別のニューラルネットワークを生成してもよい（Ｓ２４３０）。ステップＳ２４２０及びステップＳ２４３０における突然変異と交叉の詳細な流れについては、それぞれ図１４、図１５を参照して後述する。

【０１１０】

続いて、生成部２１７は、ステップＳ２４２０またはステップＳ２４３０で生成したニューラルネットワークの整合性を判定する（Ｓ２４４０）。この際、生成部２１７は、生成したニューラルネットワークのレイヤー構成にエラーが生じているか否かを判定してもよ

い。生成部２１７は、例えば、Ｍａｘ－Ｐｏｏｌｉｎｇ処理に際し、入力されるデータが小さすぎる場合などに、ネットワークの整合性がない、と判定してよい。このように、生成したニューラルネットワークの整合性がないと判定した場合（Ｓ２４５０：Ｎｏ）、生成部２１７は、生成したニューラルネットワークを破棄し、ステップＳ２４１０に復帰する。

【０１１１】

一方、生成されたニューラルネットワークに整合性が認められる場合（Ｓ２４５０：Ｙｅｓ）、判定部２１３が、生成されたニューラルネットワークが、取得部２０２により取得された制約を満たすか否かを判定する。判定部２１３が、生成されたニューラルネットワークが制約を満たさないと判定した場合（Ｓ２４５０：ＮＯ）、生成部２１７は、生成したニューラルネットワークを破棄し、ステップＳ２４１０に復帰する。

【０１１２】

一方、生成されたニューラルネットワークが制約を満たすと判定した場合、生成部２１７は、生成したニューラルネットワークと、元となるニューラルネットワークと、の入出力が同一であるか否かを判定する（Ｓ２４６０）。ここで、両者の入出力が異なる場合（Ｓ２４６０：Ｎｏ）、想定する認識問題を処理することが困難となるため、生成部２１７は、生成したニューラルネットワークを破棄し、ステップＳ２４１０に復帰する。一方、生成されたニューラルネットワークと、元となるニューラルネットワークとの入出力が同一である場合（Ｓ２４６０：Ｙｅｓ）、生成部２１７は、ネットワーク生成に係る処理を正常に終了する。

【０１１３】

以上、本実施形態に係るニューラルネットワークの生成について説明した。上述したとおり、本実施形態に係る生成部２１７は、シードネットワークやパレート最適解に係るネットワークから、ネットワーク構造の異なる別のニューラルネットワークを生成することが可能である。なお、図１３では、生成部２１７が突然変異または交叉を用いた遺伝的操作により別のニューラルネットワークを生成する場合を例に説明したが、本実施形態に係るネットワークの生成は係る例に限定されない。本実施形態に係る生成部２１７は、入力されたニューラルネットワークのネットワーク構造を変化させるニューラルネットワークを用いて、上記の別のニューラルネットワークを生成してもよい。生成部２１７によるニューラルネットワークの生成には、種々の手法が適用されてよい。

【０１１４】

続いて、本実施形態に係る突然変異によるネットワーク生成の流れについて説明する。図１４は、生成部２１７による突然変異を用いたネットワーク生成を説明するためのフローチャートである。すなわち、図１４に示すフローチャートは、図１３に示したステップＳ２４２０における生成部２１７の詳細な制御を示している。図１４を参照すると、本実施形態に係る突然変異は、レイヤーの挿入、レイヤーの削除、レイヤー種類の変更、パラメータの変更、グラフ分岐、グラフ分岐の削除を含んでよい。

【０１１５】

図１４を参照すると、まず、生成部２１７は、元となるニューラルネットワークに適用する突然変異の手法をランダムで決定する（Ｓ２４２１）。続いて、生成部２１７は、ステップＳ２４２１で選択した手法に基づいて、元となるニューラルネットワークのネットワーク構造を変化させる。

【０１１６】

生成部２１７は、新規レイヤーを挿入する処理を行ってもよい（Ｓ２４２２）。生成部２１７は、例えば、元となるニューラルネットワークに、Ｒｅｌｕなどの活性化関数を新たに挿入することで、ネットワーク構造の異なる別のニューラルネットワークを生成することができる。

【０１１７】

また、生成部２１７は、既存レイヤーを削除する処理を行ってもよい（Ｓ２４２３）。生成部２１７は、例えば、元となるニューラルネットワークから、Ｍａｘ－Ｐｏｏｌｉｎｇ

に係るレイヤーを削除することで、ネットワーク構造の異なる別のニューラルネットワークを生成することができる。

【０１１８】

また、生成部２１７は、既存レイヤーのレイヤー種類を変更する処理を行ってもよい（Ｓ２４２４）。生成部２１７は、例えば、元となるニューラルネットワークに存在する活性化関数を別の活性化関数に置換することで、ネットワーク構造の異なる別のニューラルネットワークを生成することができる。

【０１１９】

また、生成部２１７は、既存レイヤーに係るパラメータを変更する処理を行ってもよい（Ｓ２４２５）。生成部２１７は、例えば、既存するＣｏｎｖｏｌｕｔｉｏｎレイヤーのカーネルシェイプを変更することで、ネットワーク構造の異なる別のニューラルネットワークを生成することができる。

【０１２０】

また、生成部２１７は、新たなグラフ分岐を作成する処理を行ってもよい（Ｓ２４２６）。生成部２１７は、例えば、既存レイヤーの一部をコピーすることでグラフ分岐を作成し、当該グラフ分岐の結合部としてＣｏｎｃａｔｅｎａｔｅレイヤーを挿入することで、別のニューラルネットワークを生成することができる。

【０１２１】

また、生成部２１７は、既存のグラフ分岐を削除する処理を行ってもよい（Ｓ２４２７）。生成部２１７は、例えば、既存するグラフ分岐の１ルートを削除し、当該削除により分岐が消失した場合にはＣｏｎｃａｔｅｎａｔｅレイヤーも削除することで、別のニューラルネットワークを生成することができる。

【０１２２】

以上、本実施形態に係る生成部２１７による突然変異を用いたネットワーク生成について説明した。なお、上記では、生成部２１７がランダムで選択したステップＳ２４２２～Ｓ２４２７の処理を実行する場合を例に説明したが、本実施形態に係る突然変異の制御は、係る例に限定されない。生成部２１７は、ステップＳ２４２２～Ｓ２４２７に係る処理を同時に２つ以上行ってもよいし、ステップＳ２４２２～Ｓ２４２７の実行判断をそれぞれ実施してもよい。また、生成部２１７は、図１４の例に示した以外の処理を実行してもよい。生成部２１７による突然変異の制御は、柔軟に変更され得る。

【０１２３】

続いて、本実施形態に係る交叉によるネットワーク生成の流れについて説明する。図１５は、生成部２１７による交叉を用いたネットワーク生成を説明するためのフローチャート図である。すなわち、図１５に示すフローチャートは、図１３に示したステップＳ２４３０における生成部２１７の詳細な制御を示している。

【０１２４】

図１５を参照すると、まず、生成部２１７は、交叉を実行するために、元となる２つのネットワークを選択する（Ｓ２４３１）。ここで、生成部２１７は、ユーザによる入力に基づいて設計され、判定部２１３により制約を満たすと判定された２つのシードネットワークを選択してもよい。また、生成部２１７は、ユーザによる入力に基づいて設計され、判定部２１３により制約を満たすと判定された１つのシードネットワークと、予め登録された交叉用のネットワークと、を選択することもできる。さらには、生成部２１７は、ユーザによる入力に基づいて設計され、判定部２１３により制約を満たすと判定されたシードネットワークから突然変異により生成した別のニューラルネットワークを選択してもよい。

【０１２５】

続いて、生成部２１７は、ステップＳ２４３１で選択した２つのネットワークを交叉させ、ネットワーク構造の異なる別のニューラルネットワークを生成する（Ｓ２４３２）。この際、生成部２１７は、種々の手法により交叉を実行してよい。生成部２１７は、例えば、一点交叉、二点交叉、多点交叉、一様交叉などにより、上記の別のニューラルネットワークを生成することができる。

【０１２６】
以上、本実施形態に係るニューラルネットワークの生成について説明した。上述したとおり、本実施形態に係る生成部２１７は、突然変異及び交叉を含む遺伝的操作などにより、元となるニューラルネットワークからネットワーク構造の異なる別のニューラルネットワークを生成することができる。すなわち、本実施形態に係る情報処理方法では、生成部２１７が生成したニューラルネットワークの評価結果に基づいてパレート最適解の更新を繰り返すことで、より効率のよいネットワーク構造を探索することが可能となる。
【０１２７】
＜２－３．効果＞
以上説明したように、本開示の第二の実施形態によれば、設計されたニューラルネットワークの評価結果に基づいて、ネットワーク構造の異なる別のニューラルネットワークを生成することができる。また、本実施形態は、ネットワークの生成と、パレート最適解の更新を繰り返すことで、効率の良いネットワーク構造を探索することが可能である。また、本実施形態において、生成されるニューラルネットワークは、判定部２１３により、複数のハードウェアに係る制約を満たすと判定されたニューラルネットワークである。したがって、本実施形態によれば、複数のハードウェアによる処理に適合し、かつ効率の良いネットワーク構造を探索することが可能である。
【０１２８】
＜＜３．変形例＞＞
以上、本開示の実施形態を説明した。以下では、本開示に係る幾つかの変形例を説明する。なお、以下に説明する各変形例は、単独で各実施形態に適用されてもよいし、組み合わせで各実施形態に適用されてもよい。また、各変形例は、上記実施形態で説明した構成に代えて適用されてもよいし、上記実施形態で説明した構成に対して追加的に適用されてもよい。
【０１２９】
＜３－１．変形例１＞
上記実施形態ではビジュアルプログラミングによりニューラルネットワークの設計が行われる例を説明したが、本技術は係る例に限定されない。例えば、本技術に係るニューラルネットワークの設計は、テキストによるプログラミングや、ＣＵＩ（Command User Interface）上の操作により行われてもよい。また、ハードウェアに係る制約や、ハードウェアとレイヤーの対応付けも、上記で説明された例に限定されず、テキスト、またはＣＵＩにより入力されてもよい。
【０１３０】
＜３－２．変形例２＞
また、上記実施形態では、図２、図５、及び図１１を参照してクライアント端末１、サーバ２、及びサーバ２－２が有する機能を説明したが、本技術は係る例に限定されない。上記実施形態で説明したクライアント端末１の機能をサーバ２、またはサーバ２－２が有してもよいし、上記実施形態で説明したサーバ２、またはサーバ２－２の機能をクライアント端末１が有してもよい。
【０１３１】
例えば、クライアント端末１が、図５を参照して説明した取得部２０２、判定部２０３、及び設計制御部２０４の機能を有し、クライアント端末１を用いて設計されたニューラルネットワークの情報が、サーバ２に提供されてもよい。
【０１３２】
＜＜４．ハードウェア構成例＞＞
以上、本開示の各実施形態を説明した。上述した表示制御処理、通信制御処理、取得処理、判定処理、設計制御処理、学習処理、認識処理、ネットワーク生成処理、評価結果取得処理等の情報処理は、ソフトウェアと、クライアント端末１、サーバ２、２－２との協働により実現される。以下では、本実施形態に係るサーバ２のハードウェア構成例について説明する。

【０１３３】
図１６は、サーバ２のハードウェア構成を示した説明図である。図１６に示したように、サーバ２は、ＣＰＵ（Ｃｅｎｔｒａｌ Ｐｒｏｃｅｓｓｉｎｇ Ｕｎｉｔ）２００１と、ＤＳＰ（Ｄｉｇｉｔａｌ Ｓｉｇｎａｌ Ｐｏｃｅｓｓｏｒ）２００２と、ＧＰＵ（Ｇｒａｐｈｉｃｓ Ｐｒｏｃｅｓｓｉｎｇ Ｕｎｉｔ）２００３と、ニューロチップ２００４と、ＲＯＭ（Ｒｅａｄ Ｏｎｌｙ Ｍｅｍｏｒｙ）２００５と、ＲＡＭ（Ｒａｎｄｏｍ Ａｃｃｅｓｓ Ｍｅｍｏｒｙ）２００６と、入力装置２００７と、出力装置２００８と、ストレージ装置２００９と、ドライブ２０１０と、通信装置２０１１とを備える。
【０１３４】
ＣＰＵ２００１は、演算処理装置および制御装置として機能し、各種プログラムに従ってサーバ２内の動作全般を制御する。また、ＣＰＵ２００１は、マイクロプロセッサであってもよい。また、ＤＳＰ２００２、ＧＰＵ２００３、ニューロチップ２００４は、演算処理装置として機能する。例えば、ＣＰＵ２００１、ＤＳＰ２００２、ＧＰＵ２００３、及びニューロチップ２００４は、本開示において、ニューラルネットワークによる認識処理を実行するハードウェアであってもよい。ＲＯＭ２００５は、ＣＰＵ２００１が使用するプログラムや演算パラメータなどを記憶する。ＲＡＭ２００６は、ＣＰＵ２００１、ＤＳＰ２００２、ＧＰＵ２００３、ニューロチップ２００４の実行において使用するプログラムや、その実行において適宜変化するパラメータなどを一時記憶する。これらはＣＰＵバスなどから構成されるホストバスにより相互に接続されている。主に、ＣＰＵ２００１、ＲＯＭ２００５及びＲＡＭ２００６とソフトウェアとの協働により、通信制御部２０１、取得部２０２、判定部２０３、設計制御部２０４、学習部２０５の機能が実現される。
【０１３５】
入力装置２００７は、マウス、キーボード、タッチパネル、ボタン、マイクロフォン、スイッチおよびレバーなどユーザが情報を入力するための入力手段と、ユーザによる入力に基づいて入力信号を生成し、ＣＰＵ２００１に出力する入力制御回路などから構成されている。サーバ２のユーザは、該入力装置２００７を操作することにより、サーバ２に対して各種のデータを入力したり処理動作を指示したりすることができる。
【０１３６】
出力装置２００８は、例えば、液晶ディスプレイ（ＬＣＤ）装置、ＯＬＥＤ（Ｏｒｇａｎｉｃ Ｌｉｇｈｔ Ｅｍｉｔｔｉｎｇ Ｄｉｏｄｅ）装置およびランプなどの表示装置を含む。さらに、出力装置２００８は、スピーカおよびヘッドホンなどの音声出力装置を含む。例えば、表示装置は、撮像された画像や生成された画像などを表示する。一方、音声出力装置は、音声データなどを音声に変換して出力する。
【０１３７】
ストレージ装置２００９は、本実施形態にかかるサーバ２の記憶部２４の一例として構成されたデータ格納用の装置である。ストレージ装置２００９は、記憶媒体、記憶媒体にデータを記録する記録装置、記憶媒体からデータを読み出す読出し装置および記憶媒体に記録されたデータを削除する削除装置などを含んでもよい。このストレージ装置２００９は、ＣＰＵ２００１が実行するプログラムや各種データを格納する。
【０１３８】
ドライブ２０１０は、記憶媒体用リーダライタであり、サーバ２に内蔵、あるいは外付けされる。ドライブ２０１０は、装着されている磁気ディスク、光ディスク、光磁気ディスク、または半導体メモリなどのリムーバブル記憶媒体に記録されている情報を読み出して、ＲＡＭ２００５に出力する。また、ドライブ２０１０は、リムーバブル記憶媒体に情報を書き込むこともできる。
【０１３９】
通信装置２０１１は、通信デバイスなどで構成された通信インタフェースである。また、通信装置２０１１は、無線ＬＡＮ（Ｌｏｃａｌ Ａｒｅａ Ｎｅｔｗｏｒｋ）対応通信装置であっても、ＬＴＥ（Ｌｏｎｇ Ｔｅｒｍ Ｅｖｏｌｕｔｉｏｎ）対応通信装置であっても、有線による通信を行うワイヤー通信装置であってもよい。通信装置２０１１は、サーバ

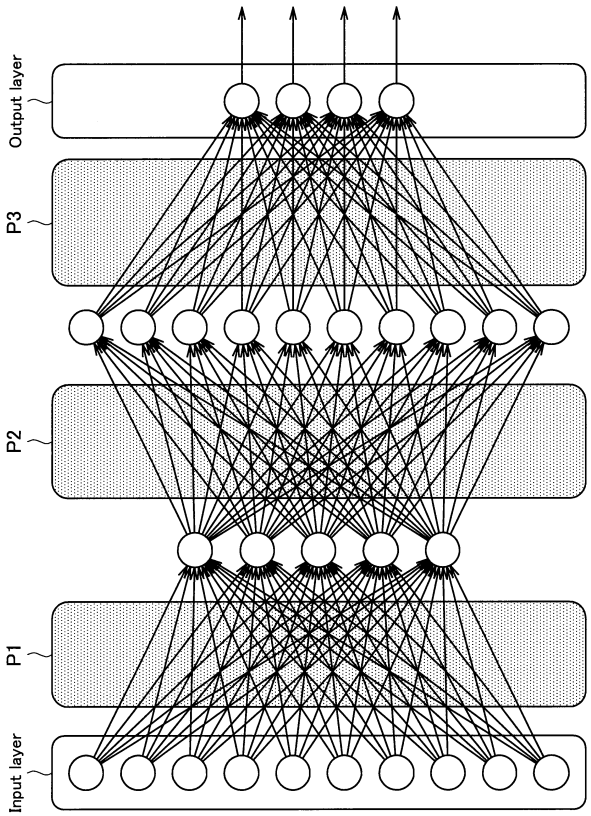
２の通信部２２に対応する。
【０１４０】
なお、上記ではサーバ２のハードウェア構成を説明したが、クライアント端末１及び第二の実施形態によるサーバ２－２も、サーバ２と同様に、ＣＰＵ２００１、ＲＯＭ２０５およびＲＡＭ２０６などに相当するハードウェアを有する。そして、クライアント端末１のハードウェアとソフトウェアとの協働により例えば制御部１０の機能が実現される。また、例えば第二の実施形態によるサーバ２－２、のハードウェアとソフトウェアとの協働により判定部２１３、生成部２１７、及び評価部２１８に相当する機能が実現される。
【０１４１】
＜＜５．むすび＞＞
以上、説明したように、本開示の実施形態によれば、複数のハードウェアによる処理に適合したニューラルネットワークをより効率的に設計することが可能である。
【０１４２】
以上、添付図面を参照しながら本開示の好適な実施形態について詳細に説明したが、本開示の技術的範囲はかかる例に限定されない。本開示の技術分野における通常の知識を有する者であれば、特許請求の範囲に記載された技術的思想の範疇内において、各種の変更例または修正例に想到し得ることは明らかであり、これらについても、当然に本開示の技術的範囲に属するものと了解される。
【０１４３】
例えば、上記実施形態における各ステップは、必ずしもフローチャート図として記載された順序に沿って時系列に処理する必要はない。例えば、上記実施形態の処理における各ステップは、フローチャート図として記載した順序と異なる順序で処理されても、並列的に処理されてもよい。例えば、図６において、ハードウェアに係る制約を入力または取得するステップＳ１１０１の処理、及びハードウェアごとに範囲を設定するステップＳ１１０２の処理は、ニューラルネットワークの設計に係るステップＳ１１０３の処理の後に行われてもよい。
【０１４４】
また、上記実施形態によれば、ＣＰＵ２００１、ＲＯＭ２００５、及びＲＡＭ２００６等のハードウェアを、上述したクライアント端末１、サーバ２、及びサーバ２－２の各構成と同様の機能を発揮させるためのコンピュータプログラムも提供可能である。また、該コンピュータプログラムが記録された記録媒体も提供される。
【０１４５】
また、本明細書に記載された効果は、あくまで説明的または例示的なものであって限定的ではない。つまり、本開示に係る技術は、上記の効果とともに、または上記の効果に代えて、本明細書の記載から当業者には明らかな他の効果を奏しうる。
【０１４６】
なお、以下のような構成も本開示の技術的範囲に属する。
（１）
複数のハードウェアに係る制約を取得する取得部と、
ニューラルネットワークが、前記制約を満たすか否か判定を行う判定部と、
を備える情報処理装置。
（２）
ユーザの入力に基づく前記ニューラルネットワークの設計を制御する設計制御部をさらに備える、前記（１）に記載の情報処理装置。
（３）
前記判定部は、前記設計制御部により前記ニューラルネットワークが変更された場合に、前記判定を行う、前記（２）に記載の情報処理装置。
（４）
前記設計制御部は、前記判定部により、前記制約を満たさないと判定された場合に、前記制約を満たすように、前記変更に係るレイヤーの再配置を行う、前記（３）に記載の情報

処理装置。
（５）
前記設計制御部は、前記判定部により、前記制約を満たすと判定された場合に、前記ニューラルネットワークを構築するプログラムを作成する、前記（２）～（４）のいずれか一項に記載の情報処理装置。
（６）
前記判定部により、前記制約を満たさないと判定された場合に、前記制約が満たされないことを示す警告画面を表示させる、表示制御部をさらに備える、前記（１）～（５）のいずれか一項に記載の情報処理装置。
（７）
前記警告画面は、前記ニューラルネットワークにおいて、前記制約を満たない部分を提示する、前記（６）に記載の情報処理装置。
（８）
前記制約に基づいて、前記ニューラルネットワークにおけるレイヤーごとに、前記レイヤーに対応付けられたハードウェアに応じた学習を行う学習部をさらに備える、前記（１）～（７）のいずれか一項に記載の情報処理装置。
（９）
前記判定部により前記制約を満たすと判定されたニューラルネットワークから、ネットワーク構造の異なる別のニューラルネットワークを生成する生成部と、
生成されたニューラルネットワークの評価結果を取得する評価部と、をさらに備え、
前記評価部は、生成されたニューラルネットワークの評価結果に基づいて、評価済のニューラルネットワークに係るパレート最適解を更新し、
前記生成部は、前記パレート最適解に係るニューラルネットワークから、ネットワーク構造の異なる別のニューラルネットワークを生成する、
前記（１）～（８）のいずれか一項に記載の情報処理装置。
（１０）
前記判定部は、前記生成部により生成されるニューラルネットワークが、前記制約を満たすか否か判定を行い、
前記生成部は、前記判定部により前記制約を満たすと判定されたニューラルネットワークが生成されるまで、前記別のニューラルネットワークの生成を繰り返す、前記（９）に記載の情報処理装置。
（１１）
前記生成部は、遺伝的操作により、前記別のニューラルネットワークを生成する、前記（９）または（１０）に記載の情報処理装置。
（１２）
前記遺伝的操作は、突然変異または交叉のうち少なくとも一方を含む、前記（１１）に記載の情報処理装置。
（１３）
前記突然変異は、レイヤーの挿入、レイヤーの削除、レイヤー種類の変更、パラメータの変更、グラフ分岐、またはグラフ分岐の削除のうち少なくともいずれか一つを含む、前記（１２）に記載の情報処理装置。
（１４）
前記情報処理装置は、前記判定部による判定結果を送信させる通信制御部をさらに備える、前記（１）～（１３）のいずれか一項に記載の情報処理装置。
（１５）
ニューラルネットワークが複数のハードウェアに係る制約を満たすか否かの判定結果を受信する受信部と、
前記判定結果に基づいて処理を行う処理部と、
を備える情報処理装置。
（１６）

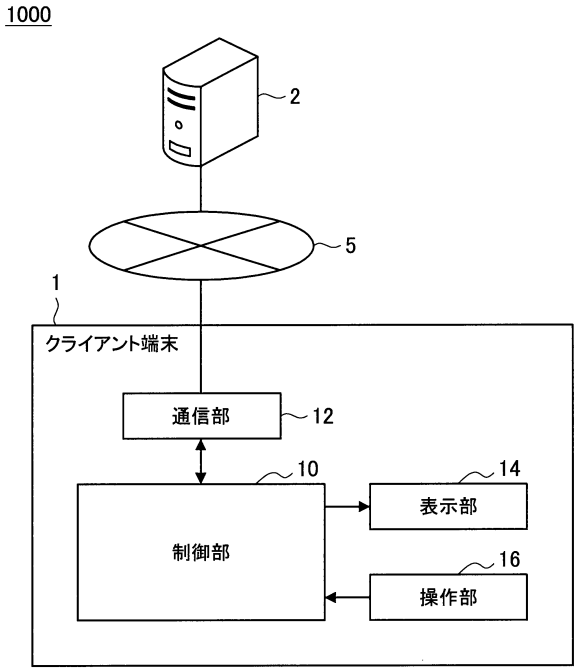
複数のハードウェアに係る制約を取得することと、
ニューラルネットワークが、前記制約を満たすか否か判定を行うことと、
を含む情報処理方法。
【符号の説明】
【０１４７】
１　クライアント端末
２、２－２　サーバ
５　通信網
１０　制御部
１２　通信部
１４　表示部
１６　操作部
２０、２１　制御部
２２　通信部
２４　記憶部
２０１　通信制御部
２０２　取得部
２０３、２１３　判定部
２０４　設計制御部
２０５　学習部
２０６　認識部
２１７　生成部
２１８　評価部
１０００　情報処理システム

【図面】

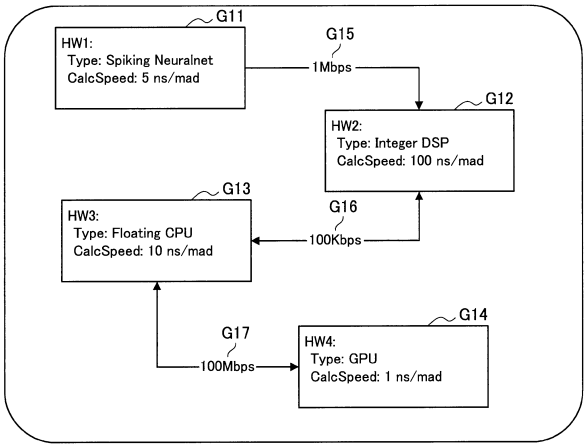
【図１】


Output layer
P3
P2
P1
Input layer

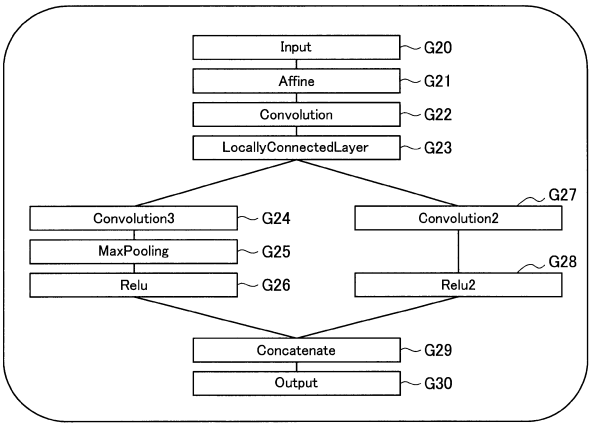
【図２】


1000
2
5
1
クライアント端末
通信部
12
10
14
表示部
制御部
16
操作部

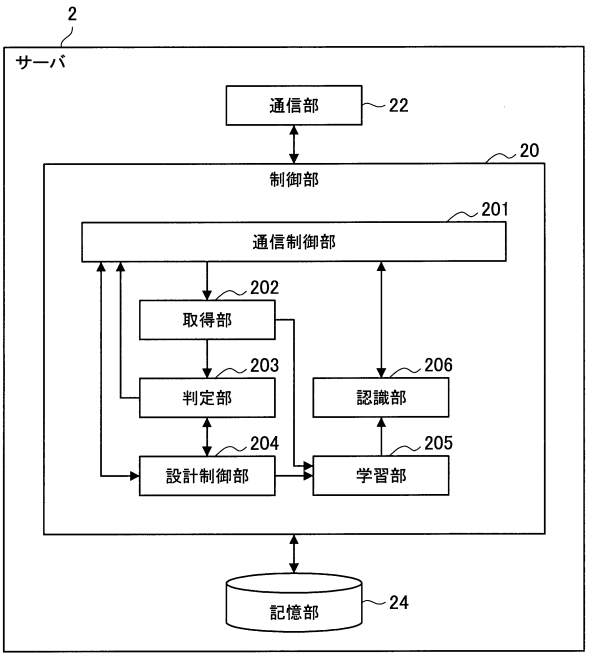
【図３】


G11
HW1:
Type: Spiking Neuralnet
CalcSpeed: 5 ns/mad
G15
1Mbps
G12
HW2:
Type: Integer DSP
CalcSpeed: 100 ns/mad
G13
HW3:
Type: Floating CPU
CalcSpeed: 10 ns/mad
G16
100Kbps
G14
G17
100Mbps
HW4:
Type: GPU
CalcSpeed: 1 ns/mad

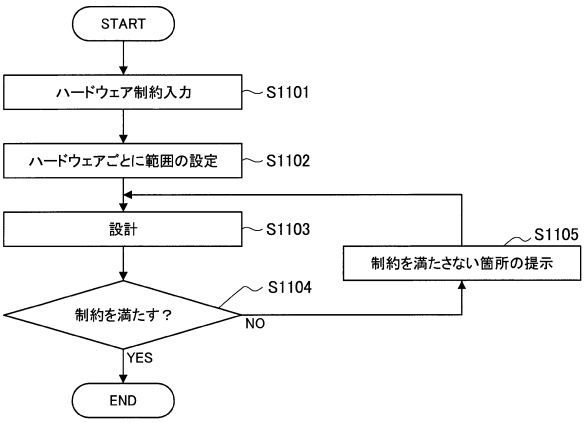
【図４】


Input
G20
Affine
G21
Convolution
G22
LocallyConnectedLayer
G23
G27
Convolution3
G24
Convolution2
MaxPooling
G25
G28
Relu
G26
Relu2
Concatenate
G29
Output
G30

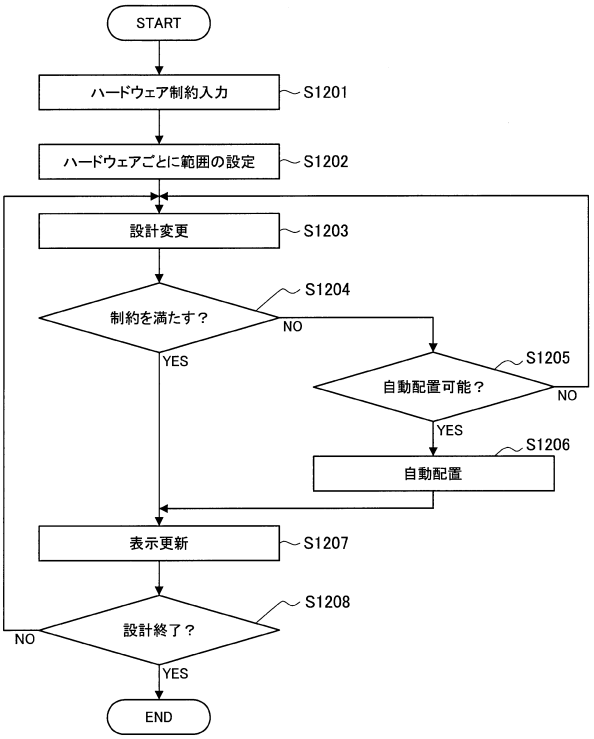
【図５】


2
サーバ
通信部
22
20
制御部
201
通信制御部
202
取得部
203
判定部
206
認識部
204
設計制御部
205
学習部
記憶部
24

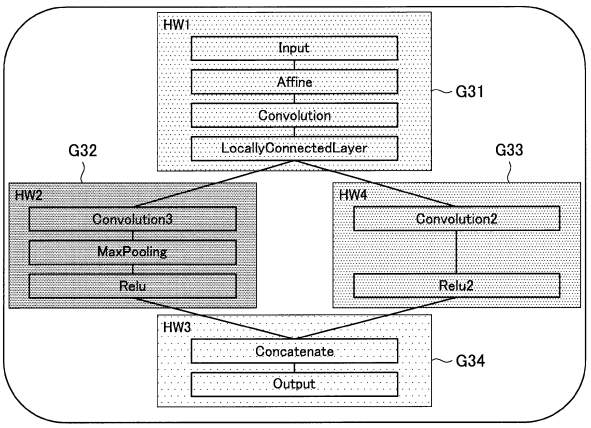
【図６】


START
ハードウェア制約入力
S1101
ハードウェアごとに範囲の設定
S1102
設計
S1103
制約を満たす？
S1104
NO
YES
制約を満たさない箇所の提示
S1105
END

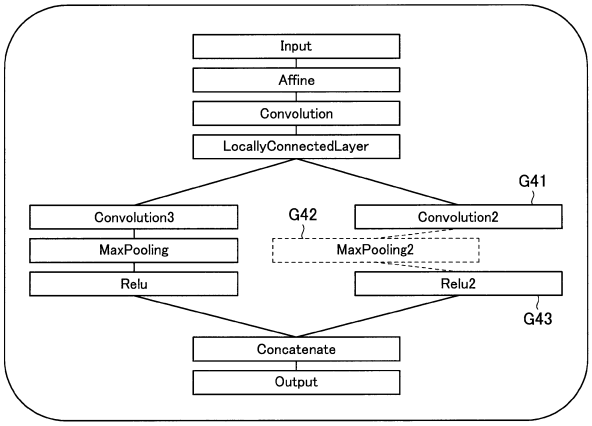
【図７】


START
ハードウェア制約入力
S1201
ハードウェアごとに範囲の設定
S1202
設計変更
S1203
制約を満たす？
S1204
NO
YES
自動配置可能？
S1205
NO
YES
自動配置
S1206
表示更新
S1207
設計終了？
S1208
NO
YES
END

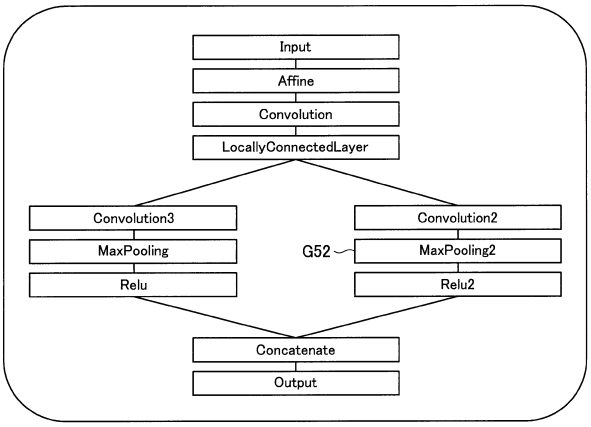
【図８】


HW1
Input
Affine
Convolution
LocallyConnectedLayer
G31
G32
G33
HW2
Convolution3
MaxPooling
Relu
HW4
Convolution2
Relu2
HW3
Concatenate
Output
G34

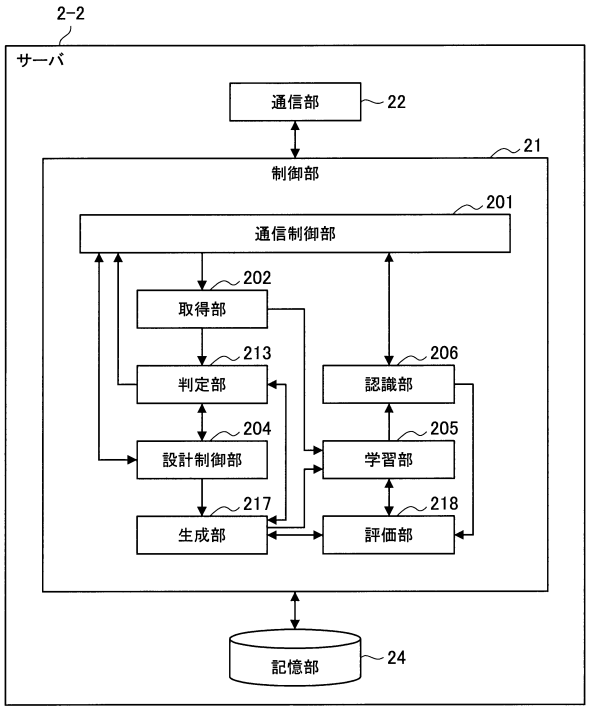
【図９】



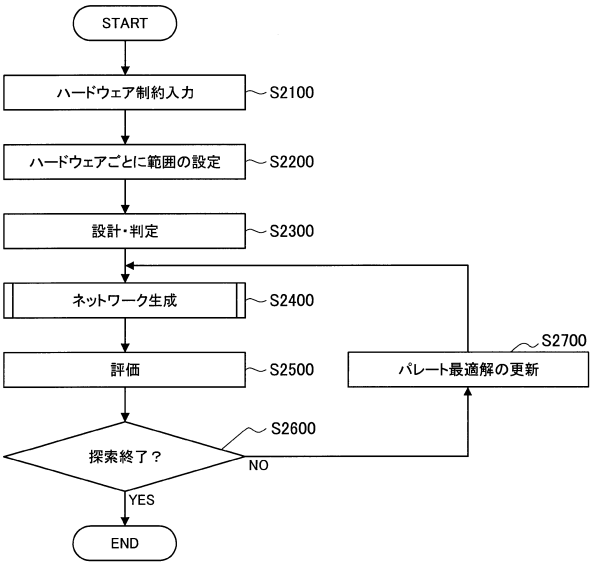
【図１０】



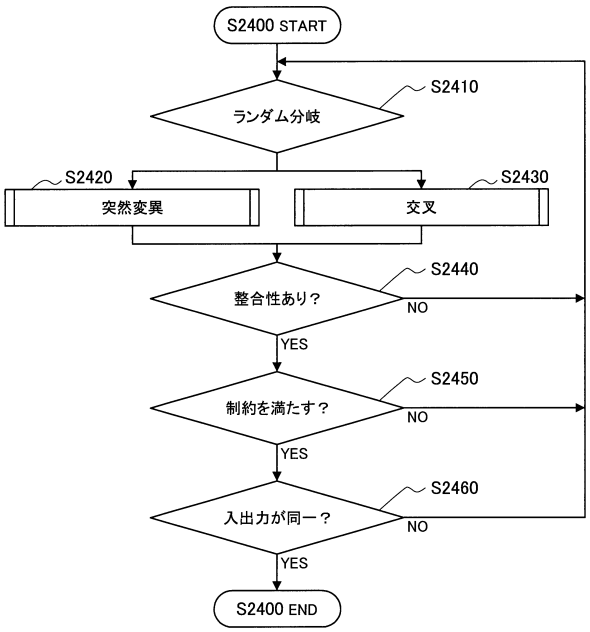
【図１１】



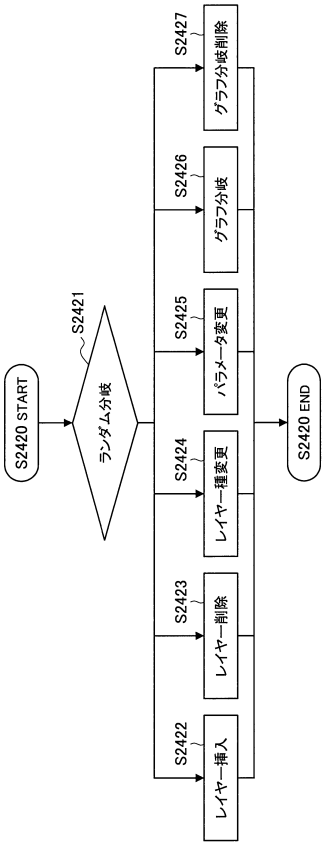
【図１２】



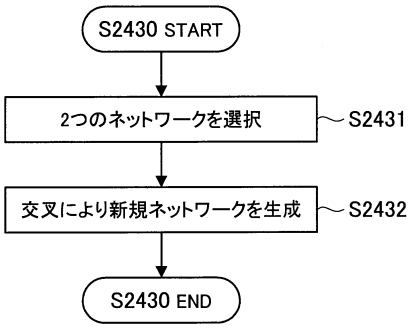
【図１３】



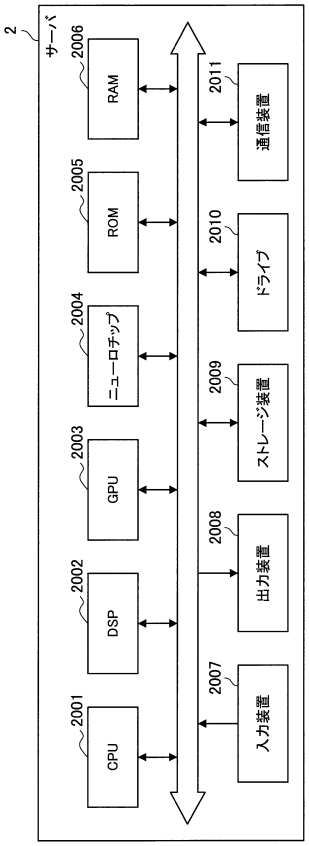
【図１４】



【図１５】



【図１６】



フロントページの続き

東京都港区港南１丁目７番１号 ソニー株式会社内
(72)発明者 小林 由幸
東京都港区港南１丁目７番１号 ソニー株式会社内
審査官 多賀 実
(56)参考文献 特開平０３－２３７５５７（ＪＰ，Ａ）
特開平１０－０９１６７６（ＪＰ，Ａ）
米国特許出願公開第２００８／０３１９９３３（ＵＳ，Ａ１）
片山 立 ほか，「複数評価基準にもとづくファジィモデルとニューロモデルの総合評価」，日本ファジィ学会誌，日本ファジィ学会，1992年10月31日，第4巻, 第5号，pp.942-957
JIN, Yaochu ほか，"Neural network regularization and ensembling using multi-objective evolutionary algorithms"，Proceedings of the 2004 Congress on Evolutionary Computation，米国，IEEE，2004年06月23日，pp.1-8
(58)調査した分野 (Int.Cl.，ＤＢ名)
Ｇ０６Ｆ １１／３４－１１／３６
Ｇ０６Ｆ ３０／００－３０／３９８
Ｇ０６Ｎ ３／０２－３／１２