(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7389575号
(P7389575)

(45)発行日　令和5年11月30日(2023.11.30)

(24)登録日　令和5年11月21日(2023.11.21)

(51)国際特許分類
Ｇ０６Ｎ　3/045(2023.01)

ＦＩ
Ｇ０６Ｎ　3/045

請求項の数　18　（全24頁）

(21)出願番号　特願2019-123945(P2019-123945)
(22)出願日　令和1年7月2日(2019.7.2)
(65)公開番号　特開2021-9622(P2021-9622A)
(43)公開日　令和3年1月28日(2021.1.28)
審査請求日　令和4年7月1日(2022.7.1)

(73)特許権者　000001007
キヤノン株式会社
東京都大田区下丸子３丁目３０番２号
(74)代理人　110003281
弁理士法人大塚国際特許事務所
(72)発明者　舘 俊太
東京都大田区下丸子３丁目３０番２号
キヤノン株式会社内
(72)発明者　チン ソクイ
東京都大田区下丸子３丁目３０番２号
キヤノン株式会社内
審査官　千葉 久博

最終頁に続く

(54)【発明の名称】　データ処理装置、データ処理方法、学習装置、学習方法、ニューラルネットワーク、及びプログラム

(57)【特許請求の範囲】

【請求項１】

ニューラルネットワークの第１の層と接続される第２の層の第３の部分の特徴量データを算出するために参照される前記第１の層の第１の部分と、前記第２の層の第４の部分の特徴量データを算出するために参照される前記第１の層の第２の部分と、をそれぞれ規定する結合パラメータを取得する取得手段と、

前記取得手段によって取得された前記結合パラメータに規定される前記第１の部分の特徴量データから前記第３の部分の特徴量データを算出する第１の算出処理と、前記結合パラメータに規定される前記第２の部分の特徴量データから前記第４の部分の特徴量データを算出する第２の算出処理とを並列に行う演算手段と、を備える

ことを特徴とするデータ処理装置。

【請求項２】

前記第１の層の前記第１の部分と前記第２の部分とは重複しない

ことを特徴とする、請求項１に記載のデータ処理装置。

【請求項３】

前記第１の層の前記第１の部分と前記第２の部分とが部分的に重複する

ことを特徴とする、請求項１に記載のデータ処理装置。

【請求項４】

前記第２の層の前記第３の部分の大きさと前記第４の部分の大きさとが異なる

ことを特徴とする、請求項１から３のいずれか１項に記載のデータ処理装置。

【請求項５】
前記結合パラメータは、チャンネル単位で前記第１の層の前記第１の部分及び前記第２の部分を規定する
ことを特徴とする、請求項１から４のいずれか１項に記載のデータ処理装置。
【請求項６】
前記結合パラメータは、複数のチャンネルを含むブロック単位で前記第１の層の前記第１の部分及び前記第２の部分を規定する
ことを特徴とする、請求項５に記載のデータ処理装置。
【請求項７】
前記演算手段は、前記第１の層の前記第１の部分に含まれる全てのチャンネルの特徴量データを用いて、かつ前記第１の層の前記第２の部分に含まれるチャンネルの特徴量データを用いずに、前記第２の層の第３の部分に含まれるそれぞれのチャンネルの特徴量データを算出する
ことを特徴とする、請求項５又は６に記載のデータ処理装置。
【請求項８】
前記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又は再帰的ニューラルネットワークである
ことを特徴とする、請求項１から７のいずれか１項に記載のデータ処理装置。
【請求項９】
前記取得手段は、前記ニューラルネットワークに入力される学習データに基づいて学習された結合パラメータを取得する
ことを特徴とする、請求項１から８のいずれか１項に記載のデータ処理装置。
【請求項１０】
学習データと、前記学習データに対する処理結果を示す教師データと、を取得する取得手段と、
ニューラルネットワークに前記学習データを入力することにより、前記学習データに対する処理結果を得るデータ処理手段であって、前記ニューラルネットワークの第１の層と接続される第２の層の第３の部分の特徴量データを算出するために参照される前記第１の層の第１の部分と、前記第２の層の第４の部分の特徴量データを算出するために参照される前記第１の層の第２の部分と、をそれぞれ規定する結合パラメータに規定される前記第１の部分の特徴量データから前記第３の部分の特徴量データを算出する第１の算出処理と、前記結合パラメータに規定される前記第２の部分の特徴量データから前記第４の部分の特徴量データを算出する第２の算出処理とを並列に行うデータ処理手段と、
前記学習データに対する処理結果と、前記教師データと、に基づいて、前記結合パラメータ及び前記ニューラルネットワークの階層間の重み係数の学習を行う学習手段と、を備える
ことを特徴とする学習装置。
【請求項１１】
前記学習手段は、第１の結合パラメータに従って前記データ処理手段が得た前記学習データに対する処理結果と、第２の結合パラメータに従って前記データ処理手段が得た前記学習データに対する処理結果と、に基づいて前記結合パラメータの学習を行う
ことを特徴とする、請求項１０に記載の学習装置。
【請求項１２】
前記学習手段は、前記第２の層の前記第１の部分及び前記第２の部分の大きさを学習により決定する
ことを特徴とする、請求項１０又は１１に記載の学習装置。
【請求項１３】
前記取得手段は、さらに、学習により得られる前記結合パラメータに対する制約条件を指示するユーザ入力を取得し、
前記学習手段は、前記制約条件に従って前記結合パラメータの学習を行う

ことを特徴とする、請求項１０から１２のいずれか１項に記載の学習装置。
【請求項１４】
順に接続された前階層、第１の層、及び第２の層を有するニューラルネットワークであって、
前記第１の層は、前記前階層の一部のニューロンセットと結合した第１のニューロンセットと、前記前階層の一部とは異なる前記前階層の他の一部のニューロンセットと結合した第２のニューロンセットと、を有し、
前記第２の層は、結合パラメータにより規定される前記第１のニューロンセットの一部及び前記第２のニューロンセットの一部と結合した第３のニューロンセットと、前記結合パラメータにより規定され前記第１のニューロンセットの一部とは異なる前記第１のニューロンセットの他の一部及び前記第２のニューロンセットの一部とは異なる前記第２のニューロンセットの他の一部と結合した第４のニューロンセットと、を有し、
前記ニューラルネットワークは、前記結合パラメータを取得し、前記結合パラメータに規定される前記第１のニューロンセットの前記一部及び前記第２のニューロンセットの前記一部における特徴量データから前記第３のニューロンセットにおける特徴量データを算出する第１の算出処理と、前記結合パラメータに規定される前記第１のニューロンセットの前記他の一部及び前記第２のニューロンセットの前記他の一部における特徴量データから前記第４のニューロンセットにおける特徴量データを算出する第２の算出処理とを並列に行うよう、コンピュータを機能させることを特徴とするニューラルネットワーク。
【請求項１５】
ニューラルネットワークの第１の層と接続される第２の層の第３の部分の特徴量データを算出するために参照される前記第１の層の第１の部分と、前記第２の層の第４の部分の特徴量データを算出するために参照される前記第１の層の第２の部分と、をそれぞれ規定する結合パラメータを取得する取得工程と、
前記結合パラメータに規定される前記第１の部分の特徴量データから前記第３の部分の特徴量データを算出する第１の算出処理と、前記結合パラメータに規定される前記第２の部分の特徴量データから前記第４の部分の特徴量データを算出する第２の算出処理とを並列に行う演算工程と、
を含むことを特徴とするデータ処理方法。
【請求項１６】
学習データと、前記学習データに対する処理結果を示す教師データと、を取得する取得工程と、
ニューラルネットワークに前記学習データを入力することにより、前記学習データに対する処理結果を得るデータ処理工程であって、前記ニューラルネットワークの第１の層と接続される第２の層の第３の部分の特徴量データを算出するために参照される前記第１の層の第１の部分と、前記第２の層の第４の部分の特徴量データを算出するために参照される前記第１の層の第２の部分と、をそれぞれ規定する結合パラメータに規定される前記第１の部分の特徴量データから前記第２の層の前記第３の部分の特徴量データを算出する第１の算出処理と、前記結合パラメータに規定される前記第２の部分の特徴量データから前記第４の部分の特徴量データを算出する第２の算出処理とを並列に行うデータ処理工程と、
前記学習データに対する処理結果と、前記教師データと、に基づいて、前記結合パラメータ及び前記ニューラルネットワークの階層間の重み係数の学習を行う学習工程と、
を備えることを特徴とする学習方法。
【請求項１７】
コンピュータを、請求項１から９のいずれか１項に記載のデータ処理装置として機能させるためのプログラム。
【請求項１８】
コンピュータを、請求項１０から１３のいずれか１項に記載の学習装置として機能させるためのプログラム。
【発明の詳細な説明】

【技術分野】
【０００１】
本発明は、データ処理装置、データ処理方法、学習装置、学習方法、ニューラルネットワーク、及びプログラムに関し、特に、ニューラルネットワーク演算の処理コストの低減に関する。
【背景技術】
【０００２】
畳み込みニューラルネットワーク（以下ＣＮＮ）のようなニューラルネットワークは複数の層を有しており、各層は複数のニューロンで構成される。通常のニューラルネットワークにおいては、各ニューロンは前階層の全てのニューロンから入力信号を受け取り、後階層の全てのニューロンへと信号を送る。このため、ニューロンとニューロンとの間の重み係数の数は、ニューロンの数の２乗に比例する。
【０００３】
実時間でニューラルネットワークを用いた認識処理を行うためには、十分な処理速度が必要とされる。また、様々な機器でニューラルネットワークを用いた処理を行うためには、処理に必要なメモリ量を削減することが求められることがある。これらの目的のために、階層間の重み係数の数を削減する手法が知られている。
【０００４】
特許文献１は、ニューラルットワークを並列化する方法を開示している。特許文献１では、同じ入力画像が入力される複数の並列ニューラルネットワークが用いられる。それぞれの並列ニューラルネットワークは、相互結合層と非相互結合層とを有している。１つの並列ニューラルネットワークの相互結合層からの出力は全ての並列ニューラルネットワークに入力されるが、非相互結合層からの出力は同じ並列ニューラルネットワークにしか入力されない。
【先行技術文献】
【特許文献】
【０００５】
【文献】米国特許第９８１１７７５号明細書
【発明の概要】
【発明が解決しようとする課題】
【０００６】
特許文献１のように非相互結合層を用いることにより、全ての層において各ニューロンが前階層及び後階層の全てのニューロンと接続している通常の畳み込みニューラルネットワークと比較して、重み係数の数を減らすことができる。
【０００７】
一方で、ニューラルネットワークを用いた認識精度は、それぞれの並列ニューラルネットワークにより抽出された特徴を組み合わせることで向上すると考えられる。このため、相互結合層をニューラルネットワークの後段に配置することにより、認識精度が向上するものと考えられる。ここで、ＣＮＮのような一般的なニューラルネットワークにおいては、後段の層ほどニューロンの数が多い傾向がある。このため、特許文献１の手法に従ってニューラルネットワークの後段に相互結合層を設け、前段に非相互結合層を設けても、重み係数の削減効果があまり大きくないという課題がある。
【０００８】
本発明は、例えばニューラルネットワークに従う処理の速度を向上させる又は処理に必要なメモリ量を削減する等の目的で、ニューラルネットワークで用いられる重み係数の数を削減することを目的とする。
【課題を解決するための手段】
【０００９】
本発明の目的を達成するために、一実施形態に係るデータ処理装置は以下の構成を備える。すなわち、

ニューラルネットワークの第１の層と接続される第２の層の第３の部分の特徴量データを算出するために参照される前記第１の層の第１の部分と、前記第２の層の第４の部分の特徴量データを算出するために参照される前記第１の層の第２の部分と、をそれぞれ規定する結合パラメータを取得する取得手段と、

前記取得手段によって取得された前記結合パラメータに規定される前記第１の部分の特徴量データから前記第３の部分の特徴量データを算出する第１の算出処理と、前記結合パラメータに規定される前記第２の部分の特徴量データから前記第４の部分の特徴量データを算出する第２の算出処理とを並列に行う演算手段と、

を備える。

【発明の効果】

【００１０】

例えばニューラルネットワークに従う処理の速度を向上させる又は処理に必要なメモリ量を削減する等の目的で、ニューラルネットワークで用いられる重み係数の数を削減することができる。

【図面の簡単な説明】

【００１１】

【図１】一実施形態に係るデータ処理装置の基本機能を示す図。
【図２】一実施形態に係るニューラルネットワークの構成を示す図。
【図３】一実施形態に係るデータ処理方法のフローチャート。
【図４】一実施形態に係るデータ処理装置のハードウェア構成例を示す図。
【図５】結合設定部１０５の動作例を示す図。
【図６】様々な結合パラメータの例を示す図。
【図７】一実施形態に係るデータ処理装置の基本機能を示す図。
【図８】一実施形態に係る学習方法のフローチャート。
【図９】結合パラメータの変更方法の例を示す図。
【図１０】一実施形態に係るデータ処理装置による処理方法を説明する図。
【図１１】コンピュータの構成例を示す図。
【図１２】一実施形態に係るデータ処理方法のフローチャート。
【図１３】一実施形態に係るニューロングループの設定例を示す図。
【図１４】遺伝的アルゴリズムを用いた学習方法を説明する図。
【図１５】再帰的ニューラルネットワークへの適用例を説明する図。

【発明を実施するための形態】

【００１２】

以下、添付図面を参照して実施形態を詳しく説明する。なお、以下の実施形態は特許請求の範囲に係る発明を限定するものではない。実施形態には複数の特徴が記載されているが、これらの複数の特徴の全てが発明に必須のものとは限らず、また、複数の特徴は任意に組み合わせられてもよい。さらに、添付図面においては、同一若しくは同様の構成に同一の参照番号を付し、重複した説明は省略する。

【００１３】

［実施形態１］

まず、図１を参照して、本実施形態に係る処理の概要を説明する。本実施形態に係るニューラルネットワークは、階層型のニューラルネットワークであり、第１の層１１１と、第２の層１１２と、を有している。また、第１の層１１１及び第２の層１１２のそれぞれは、第１のニューロングループ１０１と第２のニューロングループ１０２とを有している。第１のニューロングループ１０１は、ニューロン演算１０６ａの対象となるニューロンのグループである。また、第２のニューロングループ１０２は、ニューロン演算１０６ｂの対象となるニューロンのグループである。ここで、第１の層１１１と第２の層１１２との間のニューロン演算１０６ａ，１０６ｂは、同じニューロングループ内のニューロンの間のみを結合する。図１の例において、ニューロングループ１０１，１０２は、第１の層１１１と第２の層１１２の両方にまたがって存在している。以下で説明するように、ニュ

ーロン演算１０６ａ，１０６ｂは、それぞれ別の演算ユニット（１７ａ又は１７ｂ）が行うことができる。

【００１４】

ニューロン演算１０６ａ，１０６ｂは、ニューラルネットワークに従う演算である。ニューロン演算１０６ａ，１０６ｂとしては、畳み込み演算又は活性化関数の適用のような、ニューラルネットワークにおける一般的な処理であってもよい。具体的な演算の種類としては、A. Krizhevsky et al. "ImageNet Classification with Deep Convolutional Neural Networks", NIPS 2012.（以下Krizhevskyと呼ぶ）及びM.D. Zeiler, R. Fergus "Visualizing and Understanding Convolutional Networks", ECCV 2014.（以下Zeilerと呼ぶ）等に挙げられているものを用いることができる。

【００１５】

第１の層１１１の第１のニューロングループ１０１には、入力特徴１１ａと、入力特徴の参照順を示す参照入力特徴１２ａとが模式的に示されている。第２の層１１２の第１のニューロングループ１０１には、出力特徴１３ａが示されている。同様に、第１の層１１１の第２のニューロングループ１０２には入力特徴１１ｂ及び参照入力特徴１２ｂが、第２の層１１２の第２のニューロングループ１０２には出力特徴１３ｂが、それぞれ示されている。図１の例において、入力特徴１１ａには複数の特徴チャンネル（又は特徴面）が含まれており、それぞれの特徴チャンネルが１つのニューロンに相当する。特徴量の演算を行う際には、複数の特徴チャンネルのそれぞれに異なるフィルタ（例えば２×２画素のフィルタなど）が適用される。入力特徴１１ｂ、参照入力特徴１２ａ，１２ｂ、及び出力特徴１３ａ，１３ｂについても同様である。

【００１６】

結合設定部１０５は、第１の層１１１について、ニューロン演算１０６ａで用いられる第１の部分と、ニューロン演算１０６ｂで用いられる第２の部分とを設定する。図１の例において、結合設定部１０５は、ニューロン演算１０６ａ，１０６ｂで用いられる特徴チャンネルを設定している。結合設定部１０５は、結合パラメータに従ってこのような設定を行う。例えば、結合設定部１０５は結合パラメータに従って入力特徴１１ａ，１１ｂの参照順序を変更することができる。結合設定部１０５は、入力特徴１１ａ，１１ｂの参照先アドレスを設定するレジスタ等によって実現することができる。図１においては、結合設定部１０５は、模式的に、特徴チャンネルの順序を入れ換えるユニットとして示されている。参照入力特徴１２ａ，１２ｂは、参照順序に従って並び替えられた入力特徴１１ａ，１１ｂを模式的に示しており、それぞれ、ニューロン演算１０６ａで用いられる第１の部分、及びニューロン演算１０６ｂで用いられる第２の部分に相当する。

【００１７】

このように本実施形態においては、ニューロン演算１０６ａにより、第１の層１１１の参照入力特徴１２ａ（第１の部分の特徴量データ）から、第２の層１１２の出力特徴１３ａ（第３の部分の特徴量データ）が算出される。また、ニューロン演算１０６ｂにより、第１の層１１１の参照入力特徴１２ｂ（第２の部分の特徴量データ）から、第２の層１１２の出力特徴１３ｂ（第４の部分の特徴量データ）が算出される。図１の例においては、参照入力特徴１２ａの全ての特徴チャンネル（ニューロン）と、出力特徴１３ａの全ての特徴チャンネル（ニューロン）と、が互いに結合されている。同様に、参照入力特徴１２ｂの全ての特徴チャンネルと、出力特徴１３ｂの全ての特徴チャンネルと、が互いに結合されている。このように、出力特徴１３ａに含まれるそれぞれのチャンネルの特徴量データは、参照入力特徴１２ａに含まれる全てのチャンネルの特徴量データを用いて、かつ参照入力特徴１２ｂに含まれるチャンネルの特徴量データを用いずに、算出される。

【００１８】

結合設定部１０５が参照する結合パラメータは、結合パラメータ保持部１０４に格納されており、結合パラメータ保持部１０４から取得することができる。図１の例において、結合パラメータ決定部１０３は、結合パラメータ保持部１０４に格納されている結合パラメータを結合設定部１０５に供給することができる。この結合パラメータは、第２の層１

１２の第３の部分（出力特徴１３ａ）の特徴量データを算出するために参照される第１の層１１１の第１の部分（参照入力特徴１２ａ）を規定する。また、この結合パラメータは、第２の層１１２の第４の部分（出力特徴１３ｂ）の特徴量データを算出するために参照される第１の層１１１の第２の部分（参照入力特徴１２ｂ）も規定する。この結合パラメータはさらに、第２の層１１２の第３の部分（出力特徴１３ａ）及び第４の部分（出力特徴１３ｂ）を規定してもよい。

【００１９】

一方で、本実施形態において、結合設定部１０５は入力特徴１１ａ，１１ｂの参照順序を変更し、又は入力特徴１１ａ，１１ｂを並び替えることで参照入力特徴１２ａ，１２ｂを決定している。したがって、第２の層１１２は、入力特徴１１ａ（第１の層の第１のニューロンセット）の一部及び入力特徴１１ｂ（第１の層の第２のニューロンセット）の一部と結合している、出力特徴１３ａ（第２の層の第３のニューロンセット）を有しているといえる。また、第２の層１１２は、入力特徴１１ａ（第１の層の第１のニューロンセット）の一部及び入力特徴１１ｂ（第１の層の第２のニューロンセット）の一部と結合している、出力特徴１３ｂ（第２の層の第４のニューロンセット）を有しているといえる。ここで、入力特徴１１ａ（第１の層の第１のニューロンセット）は、第１の層の前階層の第１のニューロンセットと結合していてもよい。また、入力特徴１１ｂ（第１の層１１１の第２のニューロンセット）は、第１の層１１１の前階層の第２のニューロンセットと結合していてもよい。これらの前階層、第１の層、及び第２の層は、ニューラルネットワークに含まれており、順に接続されている。本明細書において、ニューロンセットとは、１つの階層にある複数のニューロンのサブセットのことを指す。

【００２０】

以上のように、本実施形態において、結合設定部１０５はニューロングループの間で入出力信号を交換している。図１の構成は、入力特徴１１ａと入力特徴１１ｂとの間で特徴量データを交換することにより、参照入力特徴１２ａ及び参照入力特徴１２ｂを設定することと等価である。

【００２１】

図１には、ニューラルネットワークの２層分の構造が示されているが、図２に示されるような、より大きなニューラルネットワークに本実施形態に係る処理を適用することもできる。図２には、４層の多層ＣＮＮが示されており、層間のニューロン演算は畳み込み演算を含んでいる。図２にはまた、ニューラルネットワークにデータを入力するデータ入力部１００が示されている。本実施形態に係るニューラルネットワークが処理対象とするデータの種類は特に限定されず、例えば音声、静止画像、動画、又は文章等であってもよい。図２の例では、入力データ２０として３チャンネルのカラー静止画像が入力されている。以下では、チャンネルのことをｃｈと略すことがある。

【００２２】

図２にはさらに、ニューラルネットワークから出力された演算結果に基づく処理結果を出力する結果出力部１１０も示されている。処理結果は、ニューラルネットワークが処理するタスクに応じて形態が異なる。例えば、意味的領域分割を行う場合、処理結果は、入力画像中の各画素について、Ｃ個のクラスそれぞれについての尤度を示す、Ｃ個のチャンネルを有する尤度マップであってもよい。また、分類を行う場合、処理結果は、Ｃ個の分類カテゴリーそれぞれについての尤度であってもよい。

【００２３】

図２には、第１の層２１１から第４の層２１４の間に、３回のニューロン演算１０６－１ａ，ｂ、１０６－２ａ，ｂ、及び１０６－３ａ，ｂと、２回の結合設定部１０５による処理が示されている。図２には、ニューロングループ１０１とニューロングループ１０２の２つのニューロングループが存在し、ニューロングループ間のニューロン演算による結合は存在しない。また、このニューラルネットワークは、入力層から最終層までにわたって２つのニューロングループに分かれている。一方で、結合設定部１０５により、ニューロンの入出力の参照順序がニューロングループをまたいで変更されている。このため、２

つのニューロングループ間でニューロンの入出力信号を交換することができる。
【００２４】
次に、本実施形態に係るニューラルネットワークに従って処理を行うデータ処理装置について、図３及び図４を参照して説明する。図４は、ニューラルネットワークを用いて認識処理を行う、本実施形態に係るデータ処理装置であるＩＰ等のハードウェア構成例を示す。図４は、ニューラルネットワークを用いて認識処理を行う際の、データ処理装置の動作フローを示す。なお、ＣＮＮの層間の畳み込み演算の重みパラメータは、すでに学習により得られているものとする。また、結合パラメータ保持部１０４はあらかじめ結合パラメータを格納している。結合パラメータの決定方法については後述する。
【００２５】
データ処理装置の動作が開始すると、ステップＳ３０１で演算制御部１０８はデータ入力部１００に制御信号を送る。制御信号に応じてデータ入力部１００は入力データ１０を受け取り、メモリ１０７に入力特徴として保存する。データ入力部１００は、例えば、撮像装置等から入力データ１０を受け取ることができる。以下の例において、入力データ１０は多チャンネルの画像特徴であり、とりわけ６ｃｈのマルチスペクトル画像であるものとする。
【００２６】
次に、ステップＳ３０２～Ｓ３１２のループ処理において、情報処理装置は、メモリ１０７上の入力特徴に対して、ニューラルネットワークの第１層目の処理を行う。図２の例であれば、入力データ２０を用いて、特徴データ２１ａ，２１ｂが算出される。
【００２７】
まずステップＳ３０３で、結合設定部１０５は、結合パラメータ決定部１０３を介して、結合パラメータ保持部１０４から、現在処理中の階層についての結合関係を示す、結合パラメータを読み出す。本実施形態において、結合パラメータ保持部１０４は、結合パラメータを示すテーブルを格納している。
【００２８】
図６（Ａ）は、このようなテーブルの一例を示す。また、図５は、図６（Ａ）に示すテーブル６０１に従う結合設定部１０５の動作を示している。図６（Ａ）に示すテーブル６０１は、チャンネル単位で、参照入力特徴５１２ａ及び参照入力特徴５１２ｂを特定している。すなわち、テーブル６０１は、６ｃｈの入力特徴のそれぞれについて、参照順序を示す６つの値を有している。以下では、このようなテーブルのことを順序指定テーブルと呼ぶ。
【００２９】
後述するステップにおいて、演算部１７ａ，１７ｂは、図５に示されるように、順序指定テーブルが指定する順序に従って特徴チャンネルを参照し、演算処理を行う。図６（Ａ）の例においては、第１のニューロングループ１０１については、入力特徴の［１ｃｈ，２ｃｈ，４ｃｈ］の３つのチャンネルに対応する、参照入力特徴５１２ａに対する畳み込み演算により、出力特徴５１３ａが算出される。また、第２のニューロングループ１０２については、入力特徴の［３ｃｈ，５ｃｈ，６ｃｈ］の３つのチャンネルに対応する、参照入力特徴５１２ｂに対する畳み込み演算により、出力特徴５１３ｂが算出される。すなわち、結合パラメータに従って、入力特徴５１１ａ，５１１ｂのうち、第１のニューロングループ１０１についてのニューロン演算５０６ａに用いられる参照入力特徴５１２ａが設定される。また、同様にニューロン演算５０６ｂに用いられる参照入力特徴５１２ｂも設定される。
【００３０】
ステップＳ３０４～Ｓ３１１までの処理は、ニューロングループ１０１，１０２のそれぞれについて並列に行われる。以下ではニューロングループ１０１についての処理であるステップＳ３０４ａ～Ｓ３１１ａについて説明するが、ニューロングループ１０２についての処理であるステップＳ３０４ｂ～Ｓ３１１ｂも同様に行われる。
【００３１】

ステップＳ３０５ａ～Ｓ３１０ａのループ処理により、出力チャンネルの１チャンネルごとに特徴データが算出される。このようなループ処理を所定回数（Ｎ回）行うことにより、Ｎチャンネルの出力特徴が生成される。

【００３２】

ステップＳ３０６ａで演算部１７ａは、メモリ１０７から畳み込み演算の重み係数（畳み込みカーネルとも呼ばれる）を読み込む。演算部１７ａは、読み込んだパラメータを演算部１７中のレジスタ領域（不図示）にセットする。

【００３３】

ステップＳ３０７ａで参照設定部１６ａは、畳み込み演算に使う入力特徴のアドレスを、演算部１７ａのレジスタ領域（不図示）にセットする。参照設定部１６ａは、ステップＳ３０３ａで読み込まれた順序指定テーブルにより指示される順序で、メモリ１０７上の入力特徴の指定されたチャンネルが参照されるように、アドレスをセットする。

【００３４】

ステップＳ３０８ａで演算部１７ａは、ステップＳ３０７ａでセットされたアドレスに位置する入力特徴を、メモリ１０７から取得する。そして、演算部１７ａは、取得した入力特徴に対して、ステップＳ３０６ａでセットされた重み係数を用いた畳み込み演算を行う。

【００３５】

演算部１７ａは、下式（１）に従って畳み込み演算を行うことができる。

$$F^{OUT}{}_j(x,y) = b_j + \Sigma_i \Sigma_{\Delta x} \Sigma_{\Delta y} W_{ij}(\Delta x, \Delta y) F^{IN}{}_{LUT(i)}(x+\Delta x, y+\Delta y) \quad \cdots\cdots (1)$$

この式において、F^{IN}及びF^{OUT}は入力特徴及び出力特徴を表し、この例では、ともに縦×横×チャンネル方向を有する３次元特徴データである。$F^{IN}{}_k$は、入力特徴F^{IN}のｋ番目のチャンネルを表し、$F^{OUT}{}_j$は出力特徴F^{OUT}のｊ番目のチャンネルを表す。ＬＵＴ（ｉ）は、畳み込み演算の対象となるｉ番目の参照入力特徴のチャンネルの番号に対応する、入力特徴のチャンネルの番号を示す。例えば、図５の例では、ニューロン演算５０６ａでは、参照入力特徴５１２ａの［１ｃｈ，２ｃｈ，３ｃｈ］が用いられ、これらは入力特徴５１１ａ．５１１ｂの［１ｃｈ，２ｃｈ，４ｃｈ］の３つのチャンネルに対応する。すなわち、ｉ＝［１，２，３］の時に、ＬＵＴ（ｉ）は［１，２，４］を表す。b_jはバイアス項である。W_{ij}は重み係数であり、前階層のニューロンｉと後階層のニューロンｊとの間の結合重みを示す。Δｘ，Δｙ（Δｘ，Δｙ∈［－１，０，１］）は、畳み込み範囲を示す変数である。

【００３６】

また、演算部１７ａは、畳み込み演算の結果に対して、活性化演算をさらに行うことができる。演算部１７ａは、下式（２）に従って活性化演算を行うことができる。

$$F^{OUT'}{}_j(x,y) = \theta(F^{OUT}{}_j(x,y)) \quad \cdots\cdots (2)$$

$$\theta(x) = Max(0,x)$$

この式において、θは活性化関数と呼ばれる非線形関数である。なお、演算部１７ａは、最大値プールと呼ばれる演算処理、及び全結合層の演算のような、ＣＮＮにおいて用いられるその他の演算を行うこともできる。

【００３７】

ステップＳ３０９ａで出力部１８は、ステップＳ３０８ａで得られた１チャンネル分の出力特徴を、メモリ１０７上の所定の位置に保存する。以上の処理により、出力特徴のｊ番目のチャンネルの特徴データである$F^{OUT'}{}_j$がメモリ１０７上に生成される。

【００３８】

以上のステップＳ３０４ａ～Ｓ３１１ａの処理を、所定回数（Ｎ回）繰り返すことで、第１のニューロングループ１０１についての出力特徴のＮ個のチャンネルの特徴データがメモリ１０７上に生成される。並行して、以上のステップＳ３０４ａ～Ｓ３１１ａの処理

を、所定回数（Ｎ回）繰り返すことで、第２のニューロングループ１０２についての出力特徴のＮ個のチャンネルの特徴データがメモリ１０７上に生成される。これらの処理により、ニューラルネットワークの１階層分の演算処理が完了する。

【００３９】

このようにしてステップＳ３０２～Ｓ３１２までのループを複数回繰り返すことで、ニューラルネットワークの全階層について同様の処理が行われる。ループの終了時点で、最終層の出力特徴がメモリ１０７上に生成されている。ステップＳ３１３で結果出力部１１０は、メモリ１０７上の最終層の出力特徴を出力する。こうして出力された出力特徴が、ニューラルネットワークによる入力データ１０に対する処理結果である。こうして得られた処理結果に基づいて、分類処理などの様々な認識処理を行うことができる。結果出力部１１０は、処理結果を用いて分類処理などの認識処理を行い、認識結果を出力してもよい。こうして、ニューラルネットワークを用いた認識動作が終了する。

【００４０】

結合パラメータのデータ形式は、図６（Ａ）の順序指定テーブル６０１に限られない。例えば、結合パラメータは、図６（Ａ）に示す結合テーブル６０３又は結合リスト６０２により表されてもよい。なお、図６（Ａ）において、順序指定テーブル６０１、結合リスト６０２、及び結合テーブル６０３は、全て同じ結合関係を示している。結合テーブル６０３は、入力特徴のチャンネルと、出力特徴のチャンネルと、結合の有無を表す。記号○は、チャンネル間の結合が存在することを示す。図６（Ａ）では、入力特徴の［１ｃｈ，２ｃｈ，４ｃｈ］のそれぞれが、出力特徴の［１ｃｈ，２ｃｈ，３ｃｈ，４ｃｈ］のそれぞれと結合していることを示している。図６（Ａ）はまた、次階層のニューロングループが、１ｃｈ～４ｃｈと、５ｃｈ～８ｃｈの２グループに分かれていることも示している。結合リスト６０２は、出力特徴のそれぞれのチャンネルについて、結合している入力特徴のチャンネルの番号を示している。

【００４１】

図６（Ａ）の例では、入力特徴のチャンネルの順序を変更したものが、参照入力特徴に相当する。このため、入力特徴のそれぞれのチャンネルが１回ずつ参照され、畳み込み演算に用いられた。すなわち、結合設定部１０５は、結合パラメータに従って、第１の層について第１の部分及び第２の部分を設定するが、この場合第１の部分と第２の部分とは重複していない。一方で、図６（Ｂ）に示すように、入力特徴の１つのチャンネルが、次階層の複数のニューロングループにより参照されてもよい。すなわち、第１の部分と第２の部分とが部分的に重複していてもよい。図６（Ｂ）の例では、入力特徴の２ｃｈ及び５ｃｈが、第１のニューロングループ（出力特徴の１ｃｈ～４ｃｈを算出）と第２のニューロングループ（出力特徴の５ｃｈ～８ｃｈを算出）の双方から参照されている。図６（Ｂ）においても、順序指定テーブル６１１、結合リスト６１２、及び結合テーブル６１３は、全て同じ結合関係を示している。このように、入力特徴の参照方法の設定は、順序の変更のみに限られず、さまざまな結合関係を設定することができる。

【００４２】

上記の例では、２つのニューロングループが存在していた。しかしながら、ニューロングループの数は２つに限定されない。各階層のニューロンが、任意のｎ個のグループに分けられていてもよい。

【００４３】

また、上記の例では、入力データ２０は６ｃｈのデータであったが、入力データのチャンネル数等の構成は特に限定されない。また、最初の階層においては、通常のＣＮＮと同様に畳み込み処理を行ってもよく、２層目からニューロンを複数のニューロングループに分けてもよい。また、特徴データ２０ａ，２０ｂが、それぞれ同一の入力データ２０であってもよい。例えば、特徴データ２０ａ，２０ｂの参照アドレスを、入力データ２０が格納されている同一のアドレスにすることができる。こうして、第１層目から、同一の入力データ２０が入力される２つのニューロングループについての並列処理を行ってもよい。これらの構成は、３ｃｈのＲＧＢ静止画像データのような、チャンネル数の少ない入力デ

ータに適用されてもよい。

【００４４】

本実施形態によれば、第１の層の第１の部分（例えば参照入力特徴１２ａ）と、第２の層の第３の部分（例えば出力特徴１３ａ）とが結合される。また、第１の層の第２の部分（例えば参照入力特徴１２ｂ）と、第２の層の第４の部分（例えば出力特徴１３ｂ）とが結合される。一方で、第１の層の第１の部分と、第２の層の第４の部分と、の間の結合が省略されるため、この結合に対応する畳み込み演算の演算量、及び重み係数を保持するメモリ量を削減することができる。その一方で、特許文献１のように非相互結合層を設ける場合とは異なり、ニューロングループ間での結合は維持されるため、認識精度の劣化が抑えられることが期待される。

【００４５】

さらに、本実施形態によれば、それぞれのニューロングループについてのニューロン演算を、同様の演算を用いて並列に行うことができる。例えば図４の例では、ニューロン演算１０６ａのための構成（参照設定部１６ａ、演算部１７ａ、及び出力部１８ａ）と、ニューロン演算１０６ｂのための構成（参照設定部１６ｂ、演算部１７ｂ、及び出力部１８ｂ）と、は同様である。すなわち、本実施形態に係るデータ処理装置は、並列に動作する演算部１７ａ（第１の処理部）及び演算部１７ｂ（第２の処理部）を有している。演算部１７ａは、ニューロン演算１０６ａにより、参照入力特徴１２ａから出力特徴１３ａを算出することができる。また、演算部１７ｂは、ニューロン演算１０６ｂにより、参照入力特徴１２ｂから出力特徴１３ｂを算出することができる。

【００４６】

さらには、本実施形態によれば、異なる結合関係を用いたニューロン演算を、同様の構成を用いて実現することができる。例えば図２の例では、結合設定部１０５が読み出した結合パラメータを用いることで、第３の層２１３及び第４の層２１４を算出するためのニューロン演算を、異なる結合関係に従って行うことができる。また、第２の層２１２を算出するためのニューロン演算においては、特徴データ２１ａは特徴データ２０ａから、特徴データ２１ｂは特徴データ２０ａからそれぞれ算出されている。結合パラメータは、このようにニューロングループ間での結合が存在しない結合関係も示すことができる。さらには、一部の階層についての結合パラメータが、前階層の全てのチャンネルと後階層の全てのチャンネルとが互いに結合されることを示してもよい。このような構成によれば、特許文献１のように相互結合層と非相互結合層とが混在する場合と比較して、演算ハードウェアの回路規模を小さくすることができ、並列処理の効率を向上できる。

【００４７】

図２の例では、２つのニューロングループは排他的なグループであり、１つのニューロンは２つのニューロングループのいずれか一方に含まれていた。このような構成は並列処理を容易にするが、ニューロングループの構成はこのような形態には限定されない。例えば、ニューロングループ同士がオーバーラップしていてもよい。例えば、図６（Ｃ）に示す結合テーブル６２３は、出力特徴についてのニューロングループには、１ｃｈ～３ｃｈ、３ｃｈ～５ｃｈ、５ｃｈ～７ｃｈの３つのグループが含まれることを示している。この例では、３ｃｈ及び５ｃｈは複数のニューロングループに含まれている。このように、様々な結合関係を示す結合パラメータを設計することができる。

【００４８】

本実施形態の適用対象は、特定の種別のニューラルネットワークには制限されない。例えば、本実施形態は、再帰的ニューラルネットワーク（Ｒｅｃｕｒｓｉｖｅ Ｎｅｕｒａｌ Ｎｅｔｗｏｒｋ，Byeon et al. "Scene labeling with LSTM recurrent neural networks", CVPR 2015.（以下Byeon）を参照）又はオートエンコーダーのような様々な種類のニューラルネットワークに適用可能である。

【００４９】

［実施形態２］

以下では、実施形態１で説明したニューラルネットワークの結合パラメータの決定方法

について説明する。実施形態２に係る学習装置は、ニューラルネットワークの学習を行うことができる。実施形態２に係る学習装置は、通常の機械学習の手法を用いて、ニューラルネットワークの重み係数の学習を行うことができる。さらに、実施形態２に係る学習装置は、結合パラメータの決定も行うことができる。

【００５０】

本実施形態における学習方法を概説する。学習には、学習データと、学習データに対する処理結果を示す教師データと、が用いられる。実施形態１で説明したように、所定の結合パラメータに従って処理を行うニューラルネットワークに学習データを入力すると、学習データに対する処理結果が得られる。そして、結合パラメータと、学習データに対する処理結果と、教師データと、に基づいて、結合パラメータ及びニューラルネットワークの階層間の重み係数との学習が行われる。結合パラメータの具体的な学習方法としては、学習中に結合パラメータに摂動的な変化を与えることにより、ニューラルネットワークの性能が向上する結合パラメータを探索する方法が挙げられる。例えば、結合パラメータに変化を与えたことによりニューラルネットワーク全体の性能が向上した場合に、この変化を採用するステップを繰り返すことができる。

【００５１】

以下、図７及び図８を参照して本実施形態について説明する。図７は、本実施形態に係る学習装置の機能構成例を示す。図７に示す学習装置は、教師信号提供部１２１及び制約指示部１２２を有することを除き、実施形態１と同様である。図７に示される第１のニューロングループ１０１及び第２のニューロングループ１０２も、図４に示される結合設定部１０５、参照設定部１６ａ，１６ｂ、演算部１７ａ，１７ｂ、及び出力部１８ａ，１８ｂを用いて実現できる。

【００５２】

データ入力部１００は、学習データを取得し、メモリ１０７を介してニューラルネットワークに入力する。また、教師信号提供部１２１は、学習データに対する処理結果を示す教師データを取得し、結果出力部１１０に入力する。教師データは、学習データに対して、ニューラルネットワークが出力する結果の目標値である。このような教師データは予め作成されていてもよい。以下では、教師データを示す信号のことを教師信号と呼ぶ。

【００５３】

制約指示部１２２は、学習により得られるニューラルネットワークについての制約条件を指示する。制約条件に、例えば、ニューラルネットワークの階層数などを含むことができる。また、制約条件に、例えばそれぞれの階層におけるニューロングループの数のような、学習により得られる結合パラメータに対する制約条件を含むことができる。また、制約指示部１２２は、摂動の大きさを決めるパラメータτのような、学習装置による学習時に用いられるパラメータを指示することができる。これらの制約条件及びパラメータは、ユーザが指定することができ、制約指示部１２２は、このような制約条件を指示するユーザ入力を取得することができる。ユーザによる指定がない項目については、制約指示部１２２はデフォルト値を用いることができる。

【００５４】

図８は、本実施形態における学習処理のフローチャートの一例である。学習処理がスタートすると、ステップＳ８０１で制約指示部１２２は、制約情報及びパラメータを示すユーザ入力を取得する。

【００５５】

ステップＳ８０２で演算制御部１０８は、ステップＳ８０１で取得したユーザ入力に従う構成を有するニューラルネットワークについて、全ての重み係数を初期化する。重み係数の初期値としては乱数値を用いることができる。また、結合パラメータ決定部１０３は、結合パラメータを初期化する。本実施形態においては、図６（Ａ）に示すような順序指定テーブル６０１が、結合パラメータを示すために用いられる。結合パラメータの初期値としては、図９の順序指定テーブル９０１のような、順序が変更されていない状態を用いることができる。

【００５６】
ステップＳ８０３でデータ入力部１００は、複数の学習画像をメモリ１０７を介してニューラルネットワークへ入力する。また、教師信号提供部１２１は、複数の学習画像のそれぞれに対応する教師信号を結果出力部１１０に入力する。このように、ニューラルネットワークの学習のためには複数の学習画像を用いることができる。
【００５７】
ステップＳ８０４では演算制御部１０８は、実施形態１と同様にニューラルネットワークを用いた処理を行う。すなわち、結合設定部１０５は、結合パラメータに従って、ニューロン演算において参照される入力特徴のチャンネルを設定する。そして、このような設定に従って、各層のニューロン演算が行われる。こうして、それぞれの学習画像に対するニューラルネットワークによる処理結果が得られる。
【００５８】
ニューラルネットワークを用いた処理が終了すると、結果出力部１１０は、処理結果を教師信号と比較する。そして、結果出力部１１０は、比較結果を示す誤差信号を生成し、ニューラルネットワークに誤差信号を入力することで、ニューラルネットワークの重み係数の学習を行うことができる。ニューラルネットワークの重み係数の学習には、公知の方法を用いることができる。例えば、誤差の算出方法は特に限定されず、Krizhevsky又はZeiler等に記載の方法を用いることができる。具体例としては、学習タスクに応じて、交差エントロピー等の損失関数を用いて誤差を算出することができる。また、こうして算出された誤差に基づいて、例えば誤差逆伝播法を用いることにより、ニューラルネットワークの重み係数の学習を行うことができる。
【００５９】
ステップＳ８０４においては、複数の学習画像に対するニューラルネットワークを用いた処理と、処理結果と教師信号との比較に基づく重み係数の学習とを含む学習ステップが繰り返される。この学習ステップは、所定の回数、又は学習による誤差値の変化が所定の範囲以下に収束するまで、繰り返される。
【００６０】
ステップＳ８０５で結合パラメータ決定部１０３は、ステップＳ８０４終了時の誤差値を変数Ｌｏｓｓ１として記憶する。また、結合パラメータ決定部１０３は、ステップＳ８０４終了時のニューラルネットワークの重み係数を記憶する。
【００６１】
ステップＳ８０６で結合パラメータ決定部１０３は結合パラメータを変更する。例えば、結合パラメータ決定部１０３は、第１の結合パラメータを第２の結合パラメータに変更することができる。図９（Ａ）は、結合パラメータを保持する結合関係テーブルの変更例を示す。結合パラメータ決定部１０３は、例えば、ランダムに選ばれた階層についての結合関係テーブルを結合パラメータ保持部１０４から取得し、ランダムに選ばれたテーブル上のｍ個の値を入れ替えることができる。図９（Ａ）の例では、ｍ＝２である。
【００６２】
ステップＳ８０７で演算制御部１０８は、ステップＳ８０６における変更後の結合パラメータに従ってニューラルネットワークを用いた処理を行う。そして結果出力部１１０は、ステップＳ８０４と同様に、複数の学習画像に対するニューラルネットワークによる処理結果に基づいて、ニューラルネットワークの重み係数の学習を行う。ステップＳ８０６における結合パラメータの変更により、誤差値は一時的に上昇するかもしれないが、ステップＳ８０７においてステップＳ８０４と同様に重みの更新を行うことで、変更後の結合関係に適した重み係数が得られる。
【００６３】
ステップＳ８０８で結合パラメータ決定部１０３は、ステップＳ８０７終了時の誤差値をＬｏｓｓ２として記憶する。
【００６４】
ステップＳ８０９で結合パラメータ決定部１０３は、ステップＳ８０４で第１の結合パ

ラメータに従って得られた処理結果と、ステップＳ８０７で第２の結合パラメータに従って得られた処理結果と、に基づいて結合パラメータの学習を行う。この例において結合パラメータ決定部１０３は、Ｌｏｓｓ１とＬｏｓｓ２とを比較し、この比較に基づいて第１の結合パラメータ又は第２の結合パラメータを採用することができる。例えば、結合パラメータ決定部１０３は、下式（３）に従ってＬｏｓｓ１とＬｏｓｓ２とを比較することができる。

Ｌｏｓｓ２ ＜ Ｌｏｓｓ１ ＋ τ ……（３）

この式が真である場合、処理はステップＳ８１０に進み、偽である場合、処理はステップＳ８１２に進む。上式において、τは摂動の許容度を決めるパラメータであり、ユーザが指定することができる。より大きい値のτは、結合関係をより頻繁に変更することを許容する。

【００６５】

ステップＳ８１０において結合パラメータ決定部１０３は、ステップＳ８０６における変更後の第２の結合パラメータを採用する。さらにステップＳ８１１において結合パラメータ決定部１０３はτの値を小さくする。例えば、結合パラメータ決定部１０３は、τの値にα（α＜１）を乗算することができる。その後、処理はステップＳ８０５に戻り、重み係数及び学習パラメータの学習が継続される。

【００６６】

ステップＳ８１２において結合パラメータ決定部１０３は、ステップＳ８０６における変更後の第２の結合パラメータを棄却し、変更前の第１の結合パラメータを採用する。また、結合パラメータ決定部１０３は、ニューラルネットワークの重み係数を、ステップＳ８０５で記憶された、結合パラメータ変更以前の状態に戻す。ステップＳ８１３において演算制御部１０８は、学習を終了するかどうかを判定する。例えば、これまでの学習の総ステップ数が所定数以上である場合に学習を終了することができる。学習を終了しない場合、処理はステップＳ８０６に戻り、結合パラメータの学習が継続される。以上が、学習処理の流れである。

【００６７】

学習方法は上記の例には限定されない。例えば、図８の例ではランダムに選ばれた層の結合関係が変更されたが、入力に近い低次の階層から高次の階層へと順に結合関係の学習が行われてもよい。また、上記の例では結合パラメータの初期値として順序が変更されていない状態が用いられたが、所定の割合だけランダムに順序が入れ替えられた状態が初期値として用いられてもよい。この割合は、ユーザ入力により指定されていてもよい。

【００６８】

また、図８の例では、結合パラメータを変更し、変更を棄却するか採用するかが判定された。別の学習方法として、複数の結合パラメータを入れ替えながら同じステップ数の学習を行い、誤差が小さくなった結合パラメータを採用することもできる。

【００６９】

さらに、例えば図９（Ｂ）に示すように、結合を追加する摂動を結合パラメータに与えることもできる。図９（Ｂ）の結合リスト９１１は、入力特徴の１～３ｃｈを参照する第１のニューロングループと、入力特徴の４～６ｃｈを参照する第２のニューロングループが分かれていることを示している。一方で、変更後の結合リスト９１２は、第１及び第２のニューロングループが、それぞれ入力特徴の４ｃｈ及び２ｃｈをさらに参照することを示している。このような変更によればニューラルネットワークにおける結合数が増加する。この場合、ニューラルネットワークの規模の増加と、ニューラルネットワークの性能の向上とを比較し、規模の増加と比較して性能が向上している場合に変更を採用することができる。具体例としては、ステップＳ８０９において、下式（４）に従ってＬｏｓｓ１とＬｏｓｓ２とを比較することができる。

－τＬｏｇ（Ｌｏｓｓ１／Ｌｏｓｓ２） ＜ Ｌｏｇ（｜Ｗ１｜／｜Ｗ２｜） ……（４）

この式において、｜Ｗ１｜及び｜Ｗ２｜は、それぞれ結合パラメータの変更前及び変更後におけるニューラルネットワークの結合数（又は重み係数の数）を示す。このように重

み係数の総数の変化を許容する場合、最終的な重み係数の上限を指示するユーザ入力に従って結合パラメータの学習を行ってもよい。

【００７０】

さらに、蒸留学習を用いることもできる。具体的には、目的タスクについて、予め大規模なニューラルネットワークの学習を行うことができる。そして、それぞれ異なる結合パラメータに従うニューラルネットワークの候補を複数用意し、大規模なニューラルネットワークと近似する結果が得られるように、それぞれの候補の転移学習を行うことができる。その結果、もっとも誤差が小さくなった候補を選択することができる。転移学習の具体的な手法としては、例えば、Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." arXiv preprint arXiv:1503.02531 (2015).などに記載の方法を用いることができる。以上のように、機械学習による結合パラメータの学習方法は、特定の形態に限定されない。

【００７１】

このように、機械学習によりニューロングループの結合関係を決定することにより、実施形態１に係るニューラルネットワークの認識精度を向上させることができる。また、上記のように、ニューラルネットワークの演算量（又は重み係数の数）を考慮しながら、ニューロングループの結合関係を決定することもできる。これらの手法によれば、実用的な演算速度及び精度を有するニューラルネットワークを得ることを容易にすることができる。

【００７２】

［実施形態３］

実施形態１，２では、ニューロングループごとに並列処理が行われた。一方でＣＮＮにおいては、後段の階層が多くの特徴チャンネルを持つことが多い。したがって、１つのニューロングループにおけるニューロン演算の出力チャンネルの数も、後段の階層においては膨大になることがある。実施形態３では、１つのニューロングループについても並列処理が行われる。

【００７３】

図１０は、実施形態３におけるニューラルネットワークに従う処理の模式図である。図１０には、２つのニューロングループ１０１，１０２が存在し、それぞれ８ｃｈの入力特徴に対応する８ｃｈの出力特徴を算出する。一方で、図１０には、４つの並列処理ユニット１００１ａ～１００１ｄが存在する。１つの並列処理ユニット１００１ａ～１００１ｄは、それぞれ、１つの処理サイクルにおいて、４ｃｈの入力特徴１０１１ａ～１０１１ｄに対する畳み込み演算を行うことにより、８ｃｈの出力特徴１０１３ａ～１０１３ｃを出力することができる。それぞれの並列処理ユニット１００１ａ～１００１ｄは、図４に示される結合設定部１０５、参照設定部１６ａ、演算部１７ａ、及び出力部１８ａと同様の構成を有することができる。

【００７４】

図１２は、実施形態３における、ニューラルネットワークの１階層についてのニューロン演算のフローチャートである。ステップＳ１２０１では、各並列処理ユニット１００１ａ～１００１ｄの動作が並列に開始される。ステップＳ１２０２において結合設定部１０５は、ステップＳ３０３と同様に結合パラメータを読み出す。ここで、結合パラメータは、各並列処理ユニット１００１ａ～１００１ｄのレジスタに予め保持されていてもよい。

【００７５】

ステップＳ１２０３～Ｓ１２０８において、各並列処理ユニット１００１ａ～１００１ｄは、実施形態１のステップＳ３０５ａ～Ｓ３１０ａと同様に畳み込み演算処理を行う。この結果として、各並列処理ユニット１００１ａ～１００１ｄは、メモリ上に８ｃｈの出力特徴１０１３ａ～１０１３ｄを格納する。ここで、各並列処理ユニット１００１ａ～１００１ｄの処理が終了する。

【００７６】

ステップＳ１２０９において積算ユニット１０２１ａ及び積算ユニット１０２１ｂは、出力特徴１０１３ａ～１０１３ｂ及び１０１３ｃ～１０１３ｄを対応するニューロングル

ープごとに積算することで、出力特徴１０１４ａ及び出力特徴１０１４ｂを生成する。こうして、ニューラルネットワークの１階層分の演算処理が終了する。

【００７７】

なお、結合設定部１０５は、実施形態１と同様に、ニューロングループ間でチャンネルを入れ替えることができる。すなわち、結合設定部１０５は、入力特徴１０１１ａ～１０１１ｂと、入力特徴１０１１ｃ～１０１１ｄと、の間でチャンネルを入れ替えることにより、参照入力特徴１０１２ａ～１０１２ｂを設定することができる。

【００７８】

なお、結合設定部１０５は、特徴のチャンネル単位で結合関係を設定する代わりに、複数のチャンネルを含むブロック単位で結合関係を設定することもできる。すなわち、結合パラメータは、ブロック単位で参照入力特徴（１０１２ａ～１０１２ｄ）を特定していてもよい。例えば、図１０に示すように、並列処理ユニット１００１ａ～１００１ｄが１サイクルで処理する４ｃｈを１つのブロックとしてもよい。例えば、入力特徴１０１１ｂと１０１１ｃの順序を入れ替えてもよい。この場合、参照入力特徴１０１２ａ，１０１２ｂ，１０１２ｃ，１０１２ｄは、それぞれ入力特徴１０１１ａ，１０１１ｃ，１０１１ｂ，１０１１ｄに対応する。このように、チャンネルのブロック単位で結合関係を設定することにより、チャンネルごとに結合関係を設定する場合と比較して、ニューラルネットワークに従う処理を行うデータ処理装置の回路規模を小さくすることができる。

【００７９】

［その他の実施形態］

ニューロングループの設定方法は、上記の例には限定されない。例えば、図６（Ａ）では、複数のニューロンが、出力特徴の前半部分（１ｃｈから４ｃｈ）に対応するニューロングループと、出力特徴の後半部分（５ｃｈから８ｃｈ）に対応するニューロングループとに、ブロック状に分割されていた。しかしながら、例えば、複数のニューロンが、奇数番号のチャンネルに対応するニューロングループと、偶数番号のチャンネルに対応するニューロングループと、に分割されてもよい。このように、チャンネル番号についての巡回的な規則に従って、ニューロングループが構成されていてもよい。ハードウェア又はメモリ等の構成に合わせて、様々なニューロングループを設定することができる。

【００８０】

図９（Ｃ）はこのような一例を示す。図９の例では、入力特徴及び出力特徴のチャンネル数はそれぞれ８ｃｈである。この例において、ニューロンは、入力特徴と出力特徴とのチャンネル間の結合の有無に従って、［１ｃｈ，５ｃｈ］［２ｃｈ，６ｃｈ］［３ｃｈ，７ｃｈ］［４ｃｈ，８ｃｈ］の４つのニューロングループに分かれており、グループ間で信号の交換は行われない。このような例において、結合パラメータの学習を行う際に、入力特徴のチャンネルの入れ替えを摂動として与えることにより、結合テーブル９２１を結合テーブル９２２へと変更することができる。結合テーブル９２２に従って処理を行うと、１番目と２番目のグループ間で信号が交換されることになる。実施形態２と同様に、ニューラルネットワークの性能が向上すればこの入れ替えを採用することができ、演算量を増やさずにニューラルネットワークの性能が向上する結合関係を得ることができる。このように、ニューロングループがブロックに分割されていない場合にも、上述の各実施形態の方法は適用可能である。

【００８１】

また、入力特徴のチャンネルを入れ替える代わりに、出力特徴のチャンネルを入れ替えてもよい。例えば、結合テーブル９２１を結合テーブル９２３へと変更してもよい。このように、結合関係の変更方法としては、様々な方法を用いることができる。

【００８２】

また、べき集合に従ってニューロングループを設定することもできる。例えば、入力特徴がＡ，Ｂ，Ｃ，Ｄの４チャンネルを有し、出力特徴がＷ，Ｘ，Ｙ，Ｚの４チャンネルを有する場合について考える。この場合、出力特徴のそれぞれのチャンネルについての特徴データを算出するために、入力特徴のチャンネルのべき集合を用いるように、ニューロン

グループを設定することができる。例えば、ＷチャンネルはＡＢＣチャンネルから、Ｘチャンネルは ＢＣＤチャンネルが、ＹチャンネルはＣＤＡチャンネルから、ＺチャンネルはＤＡＢチャンネルから、それぞれ算出することができる。また、１つの出力チャンネルに対して、べき集合の一部（例えばべき集合 $x \in \{A, B, C, D\}$ のうちの ${}_4C_3$ の組み合わせ）を初期値として対応づけ、さらに結合パラメータの学習を行ってもよい。

【００８３】

さらに、図１３（Ａ）に示すように、各階層についてニューロングループの数が異なっていてもよい。一般に、より前の階層においては多くの特徴を用いる必要性が低いことが多いため、このような階層においてニューロングループの数を増やしても、性能に対する影響は低いことが期待される。

【００８４】

それぞれのニューロングループのサイズは同じであってもよいし、異なっていてもよい。すなわち、第１のニューロングループが出力特徴の第３の部分の特徴量データを算出し、第２のニューロングループが出力特徴の第４の部分の特徴量データを算出する際に、第３の部分と第４の部分との間で大きさが異なっていてもよい。例えば、第３の部分と第４の部分との間で特徴量データのデータ量が異なっていてもよい。

【００８５】

例えば図１３（Ｂ）の例では、異なるチャンネル数の出力特徴を算出するように、複数のニューロングループが設定されている。すなわち、３２ｃｈの入力特徴から３２ｃｈの出力特徴を生成する層において、以下のようにニューロンを２つのグループに分けてもよい。すなわち、１つのニューロングループは２４ｃｈの入力特徴から２４ｃｈの出力特徴を算出し、もう１つのニューロングループは８ｃｈの入力特徴から８ｃｈの出力特徴を算出する。このような構成によれば、各ニューロングループで行われるニューロン演算の処理の回数は異なることになる。例えば、演算量は入力チャンネル数と出力チャンネル数との積により決まるため、この例では演算量は９対１になる。ハードウェアを用いてこのようなニューラルネットワークに従う演算を行う際には、演算量に応じた数の並列処理ユニットを設けることにより、それの並列処理ユニットによる演算量を均等にすることができる。

【００８６】

図１３（Ｃ）の例では、各ニューロングループで行われるニューロン演算の演算量が等しくなるように、ニューロングループの数がコントロールされている。この例では、第１の層１３１１における８ｃｈの入力特徴から、第２の層１３１２における１６ｃｈの出力特徴が算出される。また、第２の層１３１２における１６ｃｈの入力特徴から、第３の層１３１３における３２ｃｈの出力特徴が算出される。この場合、第１の層１３１１からの入力特徴から第２の層１３１２への出力特徴を与える２つのニューロングループと、第２の層１３１２からの入力特徴から第３の層１３１への出力特徴を与える４つのニューロングループと、を設けることができる。この場合、４ｃｈの入力特徴に対して、８ｃｈの出力特徴を算出する並列処理ユニットが、それぞれのニューロングループに対応する演算を行うことができる。この例では、後ろの階層においてチャンネル数が増えているが、チャンネル数に合わせてニューロングループの数も増えているので、それぞれのニューロングループに対応する演算の演算量は同一となる。このような構成によれば、並列処理を容易にすることができる。

【００８７】

このように、様々なニューロングループの設定方法を採用することができる。学習装置は、ニューロングループの数を学習により決定してもよいし、ニューロングループのサイズを学習により決定してもよい。すなわち、第１のニューロングループが出力特徴の第３の部分の特徴量データを算出し、第２のニューロングループが出力特徴の第４の部分の特徴量データを算出する構成において、第３の部分及び第４の部分の大きさを学習により決定することができる。

【００８８】

ニューロングループの設定を学習により決定する方法は特に限定されない。例えば、適切なニューロングループの設定方法を機械学習により決定するために、前述の蒸留学習を用いることができる。また、実施形態２と同様に、ニューロングループの数をランダムに増減させることにより、結合パラメータの１つであるニューロングループの分割数に摂動的な変化を与えることができる。このような方法により、ニューラルネットワークの性能が向上する分割数を学習により決定することができる。

【００８９】

別の学習方法として、遺伝的アルゴリズムを用いる方法が挙げられる。すなわち、それぞれ異なる結合パラメータに従うニューラルネットワークの学習を行い、性能のより高いニューラルネットワーク同士を組み合わせて新たな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、１つのニューラルネットワークの前半の階層と、他のニューラルネットワークの後半の階層と、を組み合わせることにより、新たなニューラルネットワークを生成することができる。

【００９０】

図１４は、遺伝的アルゴリズムの適用例を示す。図１４には、世代１の候補として３つのニューラルネットワーク１４０１～１４０３が示されている。この中で、認識性能が低いと判定されたニューラルネットワーク１４０３が淘汰（削除）される。そして、性能が高いと判定されたニューラルネットワーク１４０１，１４０２の一部を組み合わせることにより、世代２の候補としてニューラルネットワーク１４０４～１４０６が生成されている。この場合には、式（４）などを用いることにより、ニューラルネットワークの規模を考慮してニューラルネットワークの評価を行うことができる。

【００９１】

さらに、上述の実施形態に係る方法は、再帰的ニューラルネットワークのような様々な形態のニューラルネットワークに適用できる。図１５（Ａ）は、再帰的ニューラルネットワークの一つである、Ｌｏｎｇ　ｓｈｏｒｔ　ｔｅｒｍ　ｍｅｍｏｒｙ（ＬＳＴＭ）ネットワークの構成を示す。Byeonに開示されているように、ＬＳＴＭにおいては、複数のＬＳＴＭニューロンを相互結合することにより、階層的なパターン情報の処理又は二次元情報の処理を行うことができる。それぞれのＬＳＴＭニューロンは、１つのサイクルにおいては独立して並列処理を行う。図１５（Ａ）において、ここでy_i^tは、ｉ番目のＬＳＴＭニューロンへの再帰的入力であり、高次元のベクトルである。図１５（Ｂ）では、ＬＳＴＭニューロンの再帰的結合が展開されている。図１５（Ｂ）は、本質的には図１５（Ａ）と同じ構成を表す。図１５（Ｂ）に示される構成においては、階層ｔの入力特徴から、階層ｔ＋１の出力特徴が生成されていると理解することができる。また、y_i^tはｉ番目のチャンネルの特徴量データであると理解することができる。

【００９２】

図１５（Ｃ）は、上述の実施形態に係る構成を、図１５（Ｂ）に示されるＬＳＴＭネットワークに適用した様子を表す。結合設定部１０５は、各ＬＳＴＭニューロンが生成した再帰的入力y_i^tの一部を混交させ、入力y'_i^tを生成することができる。このような処理は、図１の例において結合設定部１０５が入力特徴１１ａ，１１ｂから参照入力特徴１２を生成した処理と同様に行うことができる。こうして生成された入力y'_i^tは、ＬＳＴＭニューロンへと再帰的に入力される。このような構成においては、これにより位相的に離れたＬＳＴＭニューロンからの信号にも基づいて、各ＬＳＴＭニューロンの状態を変更することができる。このため、各ＬＳＴＭニューロンによる並列処理を可能としながら、また演算処理の増加を抑えながら、図１５（Ａ）の構成と比較してより複雑な演算処理を行うことができる。

【００９３】

本発明の一実施形態に係るデータ処理装置及び学習装置は、例えばネットワークを介して接続された複数の情報処理装置によって構成されていてもよい。

【００９４】

上述の実施形態に係るデータ処理装置及び学習装置は、専用のハードウェアによって実

現される。一方で、データ入力部１００、結合パラメータ決定部１０３、演算制御部１０８、結果出力部１１０、及び制約指示部１２２のような一部又は全部の処理部が、コンピュータにより実現されてもよい。

【００９５】

図１１はコンピュータの基本構成を示す図である。図１１においてプロセッサ１１１０は、例えばＣＰＵであり、コンピュータ全体の動作をコントロールする。メモリ１１２０は、例えばＲＡＭであり、プログラム及びデータ等を一時的に記憶する。コンピュータが読み取り可能な記憶媒体１１３０は、例えばハードディスク又はＣＤ－ＲＯＭ等であり、プログラム及びデータ等を長期的に記憶する。本実施形態においては、記憶媒体１１３０が格納している、各部の機能を実現するプログラムが、メモリ１１２０へと読み出される。そして、プロセッサ１１１０が、メモリ１１２０上のプログラムに従って動作することにより、各部の機能が実現される。図１１において、入力インタフェース１１４０は外部の装置から情報を取得するためのインタフェースである。また、出力インタフェース１１５０は外部の装置へと情報を出力するためのインタフェースである。バス１１６０は、上述の各部を接続し、データのやりとりを可能とする。

【００９６】

また、上述の実施形態で説明されたニューラルネットワークも、本発明の一実施形態である。ニューラルネットワークは、階層構造（例えば階層数及び各階層で行われる演算の種類）を示す情報と、階層間の重み係数を示す情報と、により規定することができる。上述の実施形態に係るニューラルネットワークは、さらに結合パラメータにより規定されている。このようなニューラルネットワークを規定する情報に従って、上述のデータ処理装置又はコンピュータは、入力データに対してニューラルネットワークに従う演算を行うことができる。したがって、このようなニューラルネットワークを規定する情報は、プログラムの一種に相当する。

【００９７】

本発明は、上述の実施形態の１以上の機能を実現するプログラムを、ネットワーク又は記憶媒体を介してシステム又は装置に供給し、そのシステム又は装置のコンピュータにおける１つ以上のプロセッサーがプログラムを読出し実行する処理でも実現可能である。また、１以上の機能を実現する回路（例えば、ＡＳＩＣ）によっても実現可能である。

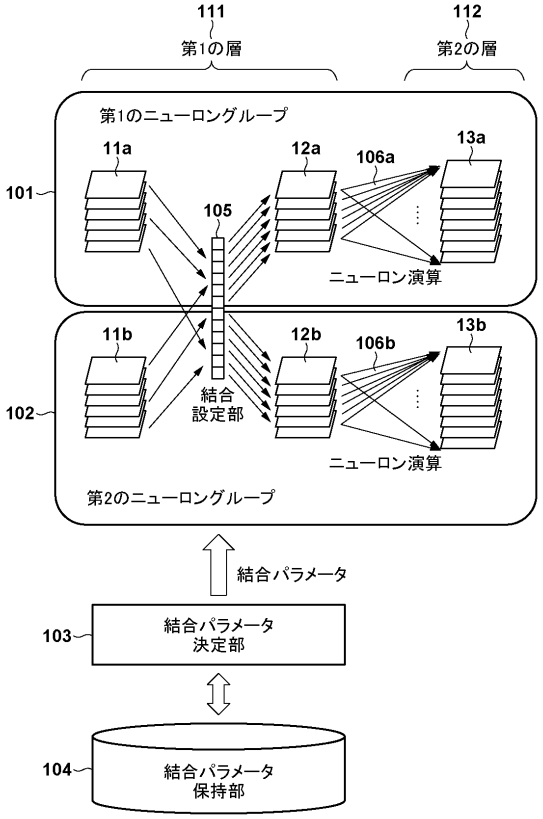
【符号の説明】

【００９８】

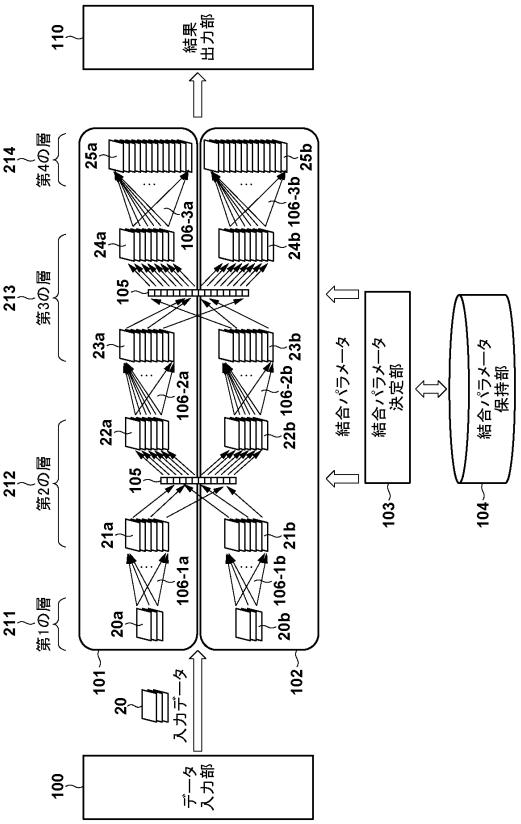
１０１，１０２：ニューロングループ、１０３：結合パラメータ決定部、１０４：結合パラメータ保持部、１０５：結合設定部、１０６ａ，１０６ｂ：ニューロン演算、１１１：第１の層、１１２：第２の層

【図面】

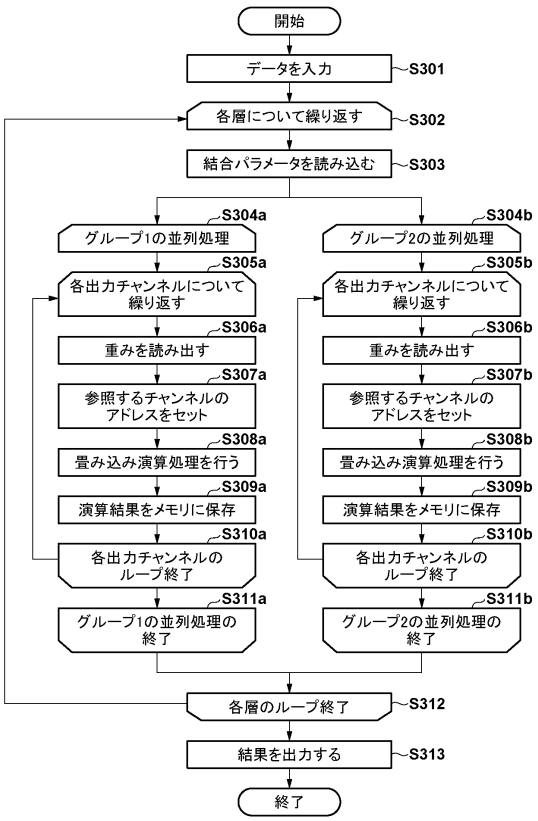
【図１】


111
第1の層
112
第2の層
第1のニューロングループ
11a
12a
106a
13a
101
105
ニューロン演算
11b
12b
106b
13b
102
結合
設定部
ニューロン演算
第2のニューロングループ
結合パラメータ
103
結合パラメータ
決定部
104
結合パラメータ
保持部

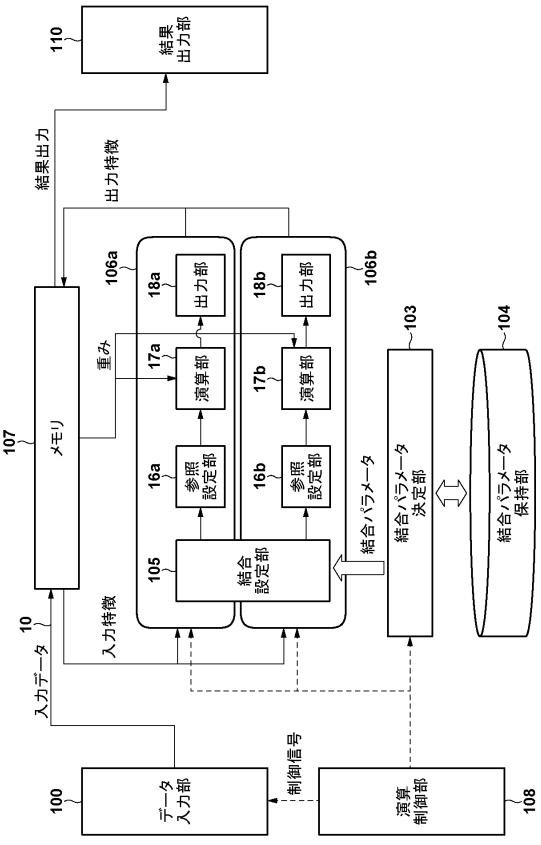
【図２】


110
結果
出力部
214
第4の層
25a
25b
106-3a
106-3b
24a
24b
213
第3の層
105
23a
23b
106-2a
106-2b
22a
22b
212
第2の層
105
21a
21b
106-1a
106-1b
211
第1の層
20a
20b
101
102
20
入力データ
100
データ
入力部
結合パラメータ
結合パラメータ
決定部
結合パラメータ
保持部
103
104

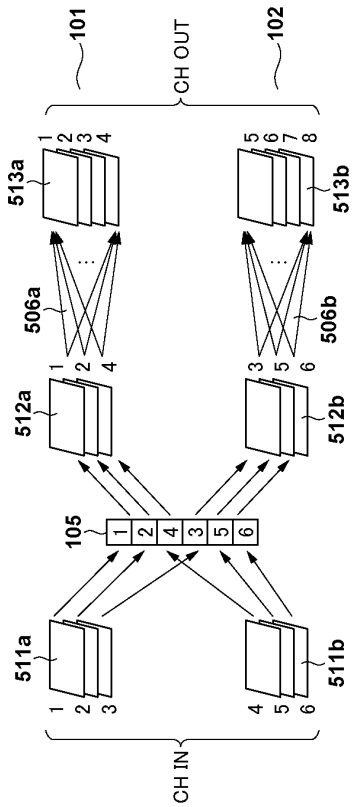
【図３】


開始
データを入力
S301
各層について繰り返す
S302
結合パラメータを読み込む
S303
S304a
グループ1の並列処理
S305a
各出力チャンネルについて
繰り返す
S306a
重みを読み出す
S307a
参照するチャンネルの
アドレスをセット
S308a
畳み込み演算処理を行う
S309a
演算結果をメモリに保存
S310a
各出力チャンネルの
ループ終了
S311a
グループ1の並列処理の
終了
S304b
グループ2の並列処理
S305b
各出力チャンネルについて
繰り返す
S306b
重みを読み出す
S307b
参照するチャンネルの
アドレスをセット
S308b
畳み込み演算処理を行う
S309b
演算結果をメモリに保存
S310b
各出力チャンネルの
ループ終了
S311b
グループ2の並列処理の
終了
各層のループ終了
S312
結果を出力する
S313
終了

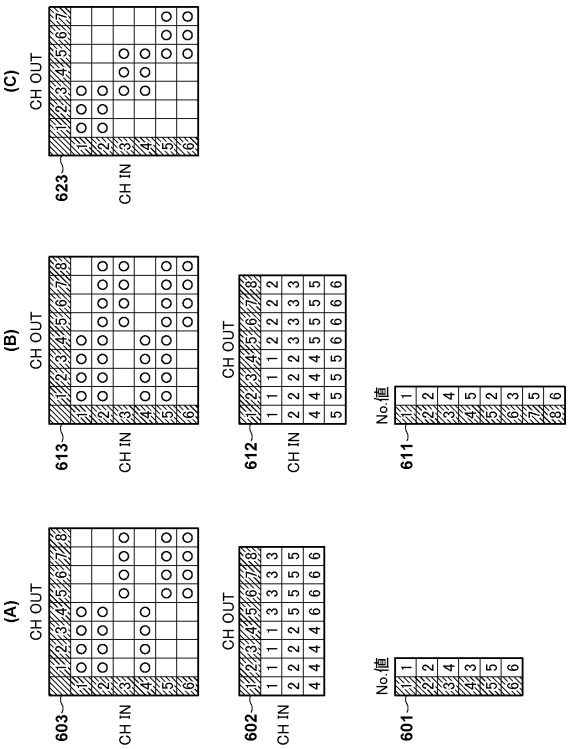
【図４】


110
結果
出力部
結果出力
出力特徴
106a
106b
18a
18b
出力部
出力部
重み
17a
17b
演算部
演算部
107
メモリ
16a
16b
参照
設定部
参照
設定部
105
結合
設定部
結合パラメータ
103
結合パラメータ
決定部
104
結合パラメータ
保持部
入力特徴
10
入力データ
100
データ
入力部
制御信号
演算
制御部
108

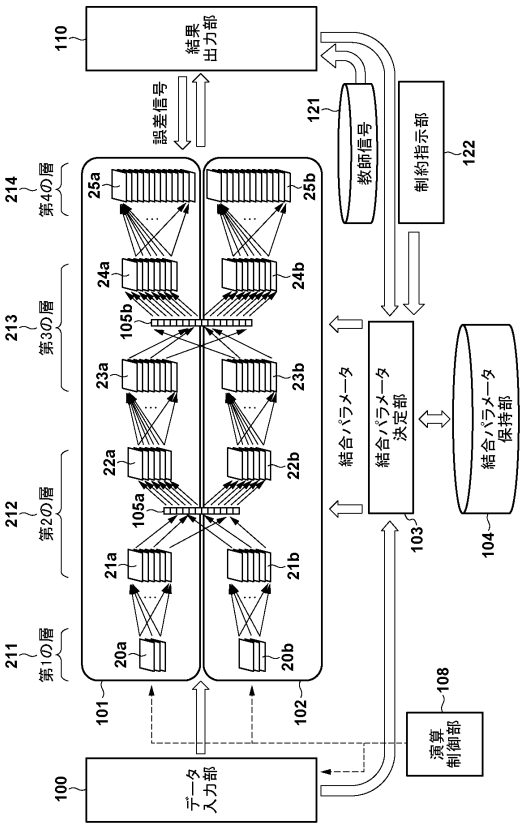
【図５】


101
CH OUT
102
513a
513b
506a
506b
512a
512b
105
511a
511b
CH IN

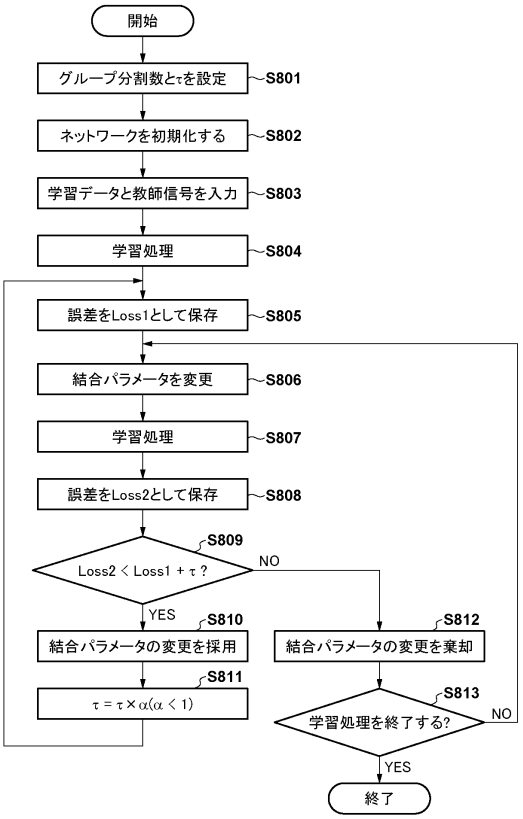
【図６】


(C)
CH OUT
623
CH IN
(B)
CH OUT
613
CH IN
CH OUT
612
CH IN
No.値
611
(A)
CH OUT
603
CH IN
CH OUT
602
CH IN
No.値
601

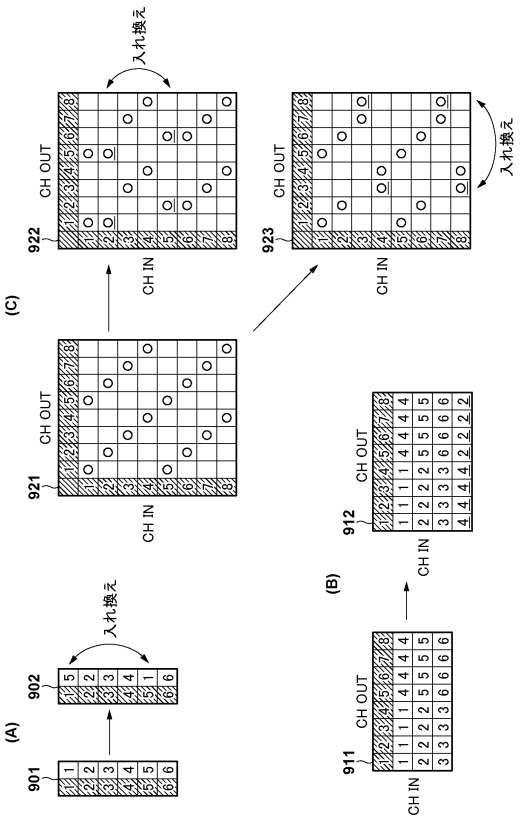
【図７】


110
結果出力部
誤差信号
121
教師信号
制約指示部
122
214
第4の層
213
第3の層
212
第2の層
211
第1の層
25a
25b
24a
24b
105b
23a
23b
22a
22b
105a
21a
21b
20a
20b
101
102
結合パラメータ
結合パラメータ決定部
103
結合パラメータ保持部
104
108
演算制御部
100
データ入力部

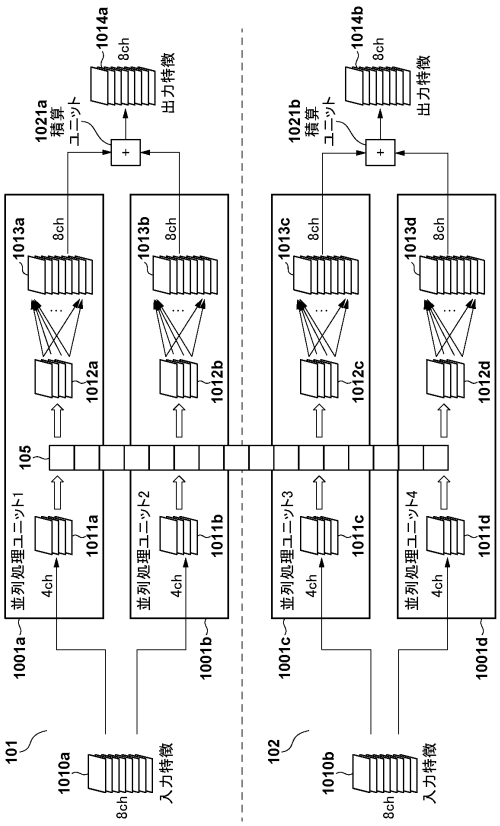
【図８】


開始
グループ分割数とτを設定
S801
ネットワークを初期化する
S802
学習データと教師信号を入力
S803
学習処理
S804
誤差をLoss1として保存
S805
結合パラメータを変更
S806
学習処理
S807
誤差をLoss2として保存
S808
S809
Loss2 < Loss1 + τ ?
NO
YES
S810
結合パラメータの変更を採用
S811
τ = τ×α(α < 1)
S812
結合パラメータの変更を棄却
S813
学習処理を終了する?
NO
YES
終了

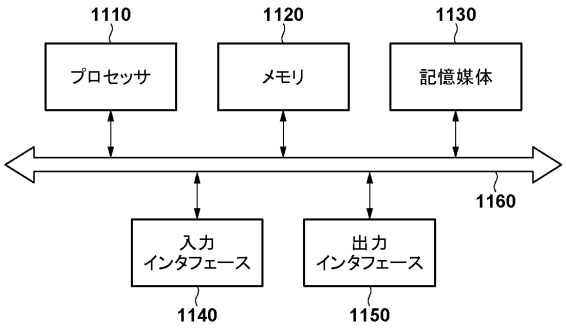
【図９】



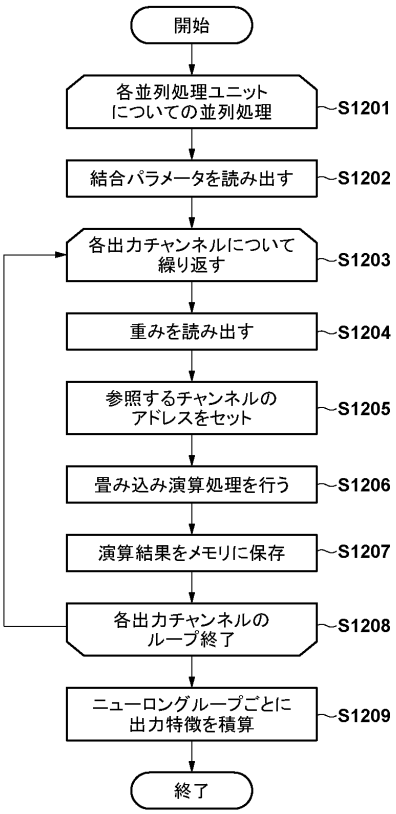
【図１０】



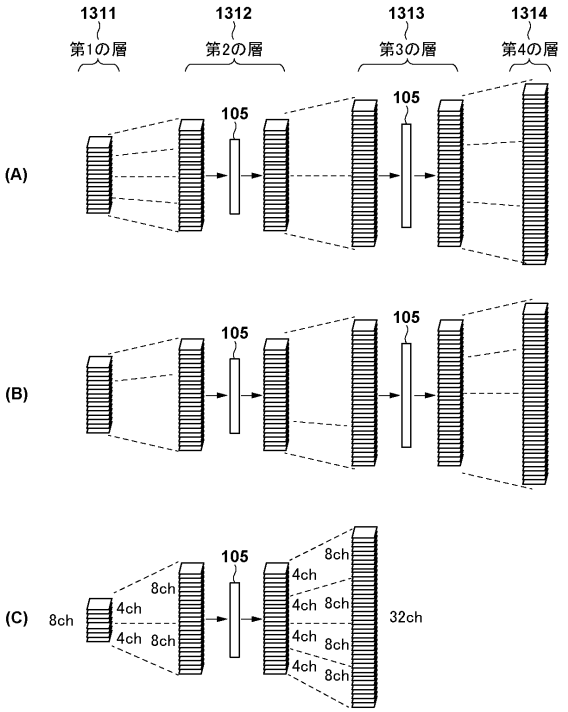
【図１１】



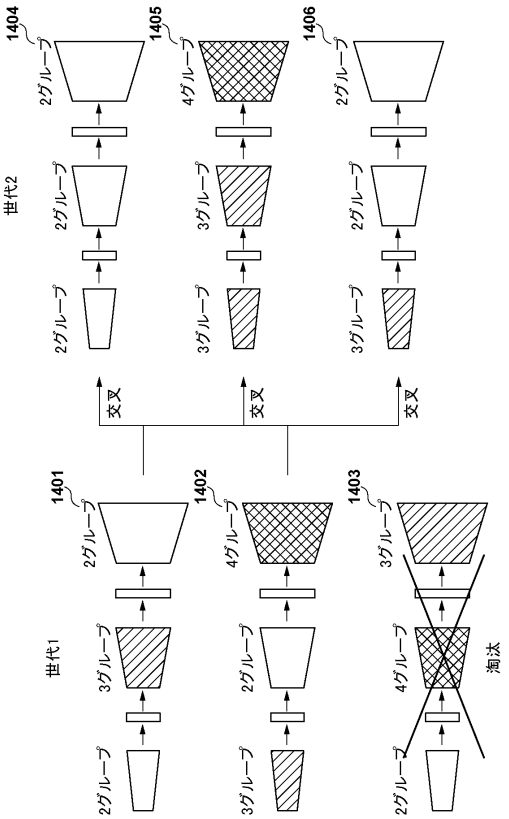
【図１２】



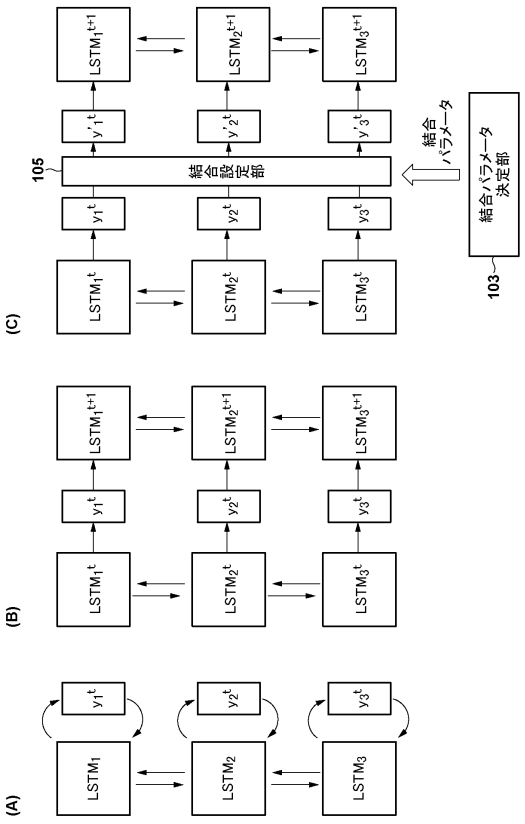
【図１３】


1311
第1の層
1312
第2の層
1313
第3の層
1314
第4の層
(A)
(B)
(C)
105
8ch
4ch
32ch

【図１４】


1404
1405
1406
2グループ
3グループ
4グループ
世代2
交叉
1401
1402
1403
世代1
淘汰

【図１５】


(C)
(B)
(A)
LSTM1
LSTM2
LSTM3
105
結合設定部
結合パラメータ決定部
103

フロントページの続き

(56)参考文献 特開２０１５－２１０７４７（ＪＰ，Ａ）
特開２００７－３０５０７２（ＪＰ，Ａ）
Pablo Barros, 外3名，"A multichannel convolutional neural network for hand posture recognition"，Proceedings of the 24th International Conference on Artificial Neural Networks(ICANN 2014)，2014年09月30日

(58)調査した分野 (Int.Cl.，ＤＢ名)
Ｇ０６Ｎ ３／０４５