(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2018年1月18日(18.01.2018)

WIPO | PCT

(10) 国際公開番号
WO 2018/011842 A1

(51) 国際特許分類:
G06N 3/04 (2006.01)

(21) 国際出願番号： PCT/JP2016/070376

(22) 国際出願日： 2016年7月11日(11.07.2016)

(25) 国際出願の言語： 日本語

(26) 国際公開の言語： 日本語

(71) 出願人:株式会社ＵＥＩ(**UEI CORPORATION**) [JP/JP]; 〒1130034 東京都文京区湯島三丁目１番３号 Tokyo (JP).

(72) 発明者:清水 亮(**SHIMIZU, Ryo**); 〒1130034 東京都文京区湯島３丁目１番３号 株式会社ＵＥＩ内 Tokyo (JP).

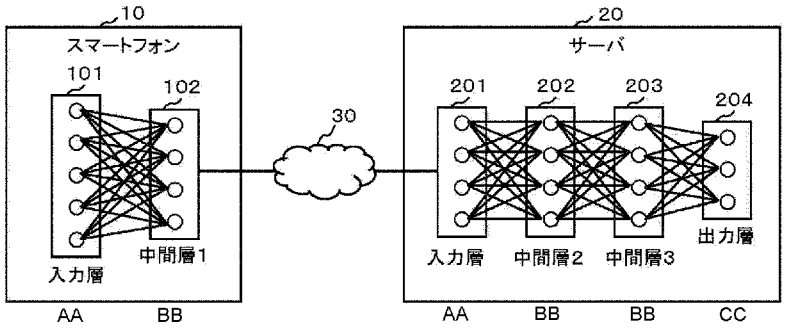
(74) 代理人: 橘 和之 (**TACHIBANA, Kazuyuki**); 〒1020083 東京都千代田区麹町１丁目４番地 半蔵門ファーストビル３階 Tokyo (JP).

(81) 指定国(表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) 指定国(表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS,

(54) **Title:** COMPUTATION SYSTEM USING HIERARCHICAL NETWORK

(54) 発明の名称： 階層ネットワークを用いた演算処理システム



10 Smartphone
20 Server
AA Input layer
BB Intermediate layer
CC Output layer

(57) **Abstract:** A smartphone 10 performs processing for the first and subsequent intermediate layers of a plurality of successive intermediate layers up to and including an intermediate layer 102, and outputs the results to a server 20 as intermediate data. Receiving as an input the intermediate data output from the smartphone 10, the server 20 performs processing for the remaining subsequent intermediate layers 202, 203 of the plurality of successive intermediate layers. This eliminates the need for the smartphone 10 to output the original data to the server 20, thereby ensuring confidentiality of information relating to the privacy of the user in possession of the original data, while causing the server 20, which delivers high computation performance, to perform part of the neural network computation so as to be able to reduce the processing time required to complete a learning computation.

WO 2018/011842 A1

［続葉有］

SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

添付公開書類：

一　国際調査報告（条約第21条(3)）

(57) 要約：スマートフォン１０において、複数の中間層のうち前半の中間層１０２の処理までを実行し、その結果を中間データとしてサーバ２０に出力し、サーバ２０において、スマートフォン１０から出力された中間データを入力として、複数の中間層のうち後半の中間層２０２，２０３の処理を実行することにより、スマートフォン１０からサーバ２０にオリジナルのデータが出力されることをなくして、オリジナルデータを保有するユーザのプライバシーに係る情報の秘匿性を確保するとともに、ニューラルネットワークによる演算の一部が演算処理能力の高いサーバ２０で実行されるようにして、学習処理の演算に要する処理時間を短縮することができるようにする。

明　細　書

発明の名称 ： 階層ネットワークを用いた演算処理システム

技術分野

[0001] 本発明は、階層ネットワークを用いた演算処理システムに関し、特に、複数の処理層が階層的に接続されたニューラルネットワークによる演算を実行する演算処理システムに用いて好適なものである。

背景技術

[0002] 従来、複数の処理層が階層的に接続されたニューラルネットワークによる演算を実行する演算処理装置が知られている（例えば、特許文献１参照）。特に、画像認識を行う演算処理装置においては、いわゆる畳み込みニューラルネットワーク（ＣＮＮ：Convolutional Neural Network）が中核的な存在となっている。

[0003] 畳み込みニューラルネットワークによれば、入力される画像データに対して中間層の処理および全結合層の処理が順次施されることにより、画像に含まれる対象物が認識された最終的な演算結果データが得られる。中間層では、複数の処理層が階層的に接続されており、各処理層において特徴量抽出処理を繰り返すことにより、入力画像データに含まれる特徴量を高次元で抽出し、その結果を中間演算結果データとして出力する。全結合層では、中間層から得られる複数の中間演算結果データを結合して最終的な演算結果データを出力する。

[0004] なお、特許文献１には、中間層を実現する回路を利用して全結合層を実現する回路を構成することにより、ニューラルネットワークによる演算処理を実現する演算処理装置の全体の回路規模を小さくすることが記載されている。

[0005] 近年では、畳み込みニューラルネットワークによる演算を用いたディープラーニング（深層学習）の研究開発が盛んに行われている。ディープラーニングは、大量の入力データをもとに、コンピュータが自ら試行錯誤を繰り返

して高次の特徴量を作り出し、それをもとに画像を分類可能にする「教師なし学習」を行うものである。ディープラーニングによって、これまで人間には認識できなかったものが認識できるようになる可能性が生まれており、産業界の期待を集めている。

[0006] 特許文献1：特開２０１６－０９９７０７号公報

発明の開示

[0007] しかしながら、ディープラーニングの学習処理には多大な演算負荷がかかり、答えを導き出すまでには非常に長い処理時間がかかる。特に、高い演算処理能力を持たないスマートフォンやタブレット等の携帯端末でディープラーニングを行おうとすると、処理に極めて長い時間がかかってしまうという問題があった。

[0008] そこで、この問題の解決方法の１つとして、ディープラーニングに必要な大量のデータを、比較的高い演算処理能力を有するサーバに携帯端末から送信し、サーバにおいて学習処理を実行することが考えられる。例えば、携帯端末で撮影した動画の各フレーム画像や、携帯端末で撮影した多数の写真画像などをサーバに送信し、サーバがこれらの画像を入力データとして学習処理を行うといった使い方が一例として考えられる。

[0009] しかしながら、ユーザの携帯端末で撮影される画像は、そのユーザのプライバシーに係るものであることが多く、これを大量にサーバに送信することに対して抵抗を感じるユーザは多い。

[0010] 本発明は、このような問題を解決するために成されたものであり、プライバシーに係る情報の秘匿性を保ちつつ、学習処理にかかる時間を短くすることができるようにすることを目的とする。

[0011] 上記した課題を解決するために、本発明では、第１の端末と、それよりも演算処理能力の高い第２の端末とに分けてニューラルネットワークによる演算を実行するようにしている。すなわち、第１の端末において、複数の中間層のうち前半の一部の中間層の処理までを実行し、その結果を中間データとして第２の端末に出力し、第２の端末において、第１の端末から出力された

中間データを入力として、複数の中間層のうち後半の一部の中間層の処理を実行するようにしている。

[0012] 上記のように構成した本発明によれば、第１の端末から出力される中間データは、第１の端末に保持されているオリジナルのデータそのものではないので、プライバシーに係る情報の秘匿性を確保することができる。また、ニューラルネットワークによる演算の一部が、演算処理能力の高い第２の端末で実行されるので、学習処理の演算に要する処理時間を短縮することができる。これにより、本発明によれば、プライバシーに係る情報の秘匿性を保ちつつ、学習処理にかかる時間を短くすることができる。

図面の簡単な説明

[0013] [図1]第１の実施形態による階層ネットワークを用いた演算処理システムの全体構成例を示す図である。

[図2]第１の実施形態によるニューラルネットワークの一例を示す図である。

[図3]第１の実施形態によるニューラルネットワークの他の例を示す図である。

[図4]第１の実施形態による階層ネットワークを用いた演算処理システムの機能構成例を示すブロック図である。

[図5]第２の実施形態によるニューラルネットワークの一例を示す図である。

[図6]第２の実施形態による階層ネットワークを用いた演算処理システムの機能構成例を示すブロック図である。

発明を実施するための最良の形態

[0014] （第１の実施形態）

以下、本発明の第１の実施形態を図面に基づいて説明する。図１は、第１の実施形態による階層ネットワークを用いた演算処理システム（以下、単に演算処理システムという）の全体構成例を示す図である。第１の実施形態による演算処理システムは、入力層と、前階層から入力されるデータに含まれる特徴量を抽出する複数の中間層と、出力層とが階層的に接続されたニューラルネットワークによる演算を実行するものである。

[0015] 図１に示すように、第１の実施形態による演算処理システムは、スマートフォン１０とサーバ２０とを備えて構成されている。スマートフォン１０とサーバ２０との間は、例えばインターネット等の通信ネットワーク３０により接続可能に構成されている。スマートフォン１０は、特許請求の範囲に記載した「第１の端末」の一例である。サーバ２０は、特許請求の範囲に記載した「第２の端末」の一例であり、スマートフォン１０よりも高い演算処理能力を有している。

[0016] 図２は、スマートフォン１０およびサーバ２０が行う演算のニューラルネットワークの一例を示す図である。図２に示すように、第１の実施形態では、スマートフォン１０において、入力層１０１に入力されたデータに対し、複数の中間層のうち前半の一部の中間層１０２の処理までを実行し、その結果を中間データとしてサーバ２０に出力する。また、サーバ２０において、スマートフォン１０の中間層１０２より出力される中間データを入力層２０１に対する入力として、複数の中間層のうち後半の一部の中間層２０２，２０３の処理を実行し、その結果を出力層２０４に出力する。

[0017] このように構成した第１の実施形態による演算処理システムでは、入力層１０１に入力されるデータに対して３つの中間層１０２，２０２，２０３の処理を順次実行することにより、各中間層１０２，２０２，２０３において、前階層から入力されるデータに含まれる特徴量を高次元で抽出し、その結果を演算結果データとして出力層２０４に出力する。ここで、スマートフォン１０における中間層１０２の出力データと、サーバ２０における入力層２０１の入力データとは同じものとなる。

[0018] 入力層１０１，２０１、中間層１０２，２０２，２０３、出力層２０４の各層は複数のニューロン（データをセットし、そのデータに対して所定の処理を実行する機能）を含んでいて、隣接する層が備える各ニューロンどうしの間がネットワークにより接続されている（ただし、中間層１０２と入力層２０１との間は通信ネットワーク３０により接続される）。層間の各ネットワークは、データを次の層に伝達する機能を有するものであり、それぞれの

ネットワークには、伝達するデータに対する重み付けが設定されている。

[0019] このようなニューラルネットワークを用いて学習を行う際は、学習対象とする多数のデータを入力層１０１に入力し、正しい答えが出力層２０４から出力されるように、各ネットワークの重み付けを思考錯誤的に変えながら調整する。ここで、出力層２０４から出力されるデータが正解とは異なるたびに重み付けの調整を繰り返すことにより、学習の精度を上げていくことが可能である。一般に、このような学習を演算処理能力の低いスマートフォン１０で行うと、演算に非常に長い時間がかかる。これに対し、第１の実施形態では、演算処理能力が高いサーバ２０と協働して学習を行うことにより、演算時間の短縮を図っている。

[0020] ところで、学習には、入力データと正しい出力データ（正解）とをあらかじめセットで与えて行う「教師あり学習」と、入力データのみを与え、そのデータに潜在する一定のパターンやルールを特徴量として抽出する「教師なし学習」とに大別される。第１の実施形態による演算処理システムは、教師あり学習にも教師なし学習にも適用することが可能である。また、学習処理が完了した後の予測処理にも適用することができることは言うまでもない。予測処理とは、１つのデータを入力し、学習済みのニューラルネットワークを用いて正解を出力する処理のことをいう。

[0021] なお、図２では、中間層の数を３つとし、そのうち最初の中間層１０２の処理のみをスマートフォン１０で実行し、残り２つの中間層２０２，２０３の処理をサーバ２０で実行する例について説明したが、中間層の総数および前後半に分ける位置はこの例に限定されない。ただし、スマートフォン１０は演算処理能力がサーバ２０に比べて低いので、サーバ２０に割り当てる中間層の数よりもスマートフォン１０に割り当てる中間層の数を少なくするのが好ましい。

[0022] 一方、スマートフォン１０に割り当てる中間層の数が少ないと、スマートフォン１０からサーバ２０に出力する中間データの中に、入力層１０１に入力されたオリジナルのデータの特徴が認識できる程度に残っている可能性が

ある。この場合、そのような中間データを外部のサーバ２０に学習のための大量に出力することに対し、スマートフォン１０のユーザが抵抗を感じるかもしれない。そこで、スマートフォン１０に割り当てる中間層の数は、スマートフォン１０での演算量が多くなり過ぎず、かつ、元の入力データの特徴が認識しにくくなる程度に高次元の特徴量が抽出されるような数に設定するのが好ましい。

[0023] あるいは、図３に示すように、中間層１０２の後段に符号化層１０３を追加するようにしてもよい。そして、この符号化層１０３において、不可逆的な符号化処理を行うことによって入力データの特徴が認識できない状態に変換した上で、その変換後の中間データをサーバ２０に出力するようにしてもよい。このようにすれば、中間層１０２より出力される中間データが、入力データの特徴がある程度認識されるようなデータであっても問題はないので、演算量の軽減のみを考慮して、スマートフォン１０に割り当てる中間層の数を少なくすることが可能である。

[0024] なお、入力層１０１にデータを入力して学習処理または予測処理を行う場合、元のデータを復元する必要性は全くない。また、入力データが有する特徴は中間データまで引き継がれているので、その中間データが符号化されたデータは、元データの特徴に対応する固有の特徴を有するデータと言える。さらに、その符号化された中間データを対象としてサーバ２０において特徴量が順次抽出されていくので、最終的に得られる演算結果データは、元データに固有の特徴を引き継いだものとなっている。したがって、一連の畳み込みニューラルネットワークによる演算の途中で不可逆的な符号化処理を行っても問題はない。

[0025] 図４は、第１の実施形態による演算処理システムの機能構成例を示すブロック図である。図４に示す演算処理システムは、階層ネットワークを用いた演算処理の一例として、畳み込みニューラルネットワークによる演算処理を適用した例を示している。また、図４に示す演算処理システムは、図３のようにスマートフォン１０に符号化層１０３を設け、中間層１０２により生成

される中間データに対して符号化層１０３にて不可逆的な符号化処理を行う例を示している。

[0026] 畳み込みニューラルネットワークの場合、入力層に入力されるデータに対して中間層の処理および全結合層の処理を順次実行する。中間層では、複数の特徴量抽出処理層が階層的に接続されており、各処理層において、前階層から入力されるデータに対して畳み込み演算処理、活性化処理、プーリング処理を行う。中間層は、各処理層における処理を繰り返すことで入力データに含まれる特徴量を高次元で抽出し、その結果を中間演算結果データとして全結合層に出力する。全結合層では、中間層から得られる複数の中間演算結果データを結合して最終的な演算結果データを出力する。

[0027] 図４に示すように、第１の実施形態による演算処理システムを構成するスマートフォン１０は、その機能構成として、データ入力部１１、前半中間層処理部１２、変換処理部１３および中間データ出力部１４を備えて構成されている。また、サーバ２０は、その機能構成として、中間データ入力部２１、後半中間層処理部２２、全結合層処理部２３およびデータ出力部２４を備えて構成されている。

[0028] スマートフォン１０の各機能ブロック１１～１４は、ハードウェア、ＤＳＰ（Digital Signal Processor）、ソフトウェアの何れによっても構成することが可能である。例えばソフトウェアによって構成する場合、上記各機能ブロック１１～１４は、実際にはコンピュータのＣＰＵ、ＲＡＭ、ＲＯＭなどを備えて構成され、ＲＡＭやＲＯＭ、ハードディスクまたは半導体メモリ等の記録媒体に記憶されたプログラムが動作することによって実現される。

[0029] また、サーバ２０の各機能ブロック２１～２４も、ハードウェア、ＤＳＰ、ソフトウェアの何れによっても構成することが可能である。例えばソフトウェアによって構成する場合、上記各機能ブロック２１～２４は、実際にはコンピュータのＣＰＵ、ＲＡＭ、ＲＯＭなどを備えて構成され、ＲＡＭやＲＯＭ、ハードディスクまたは半導体メモリ等の記録媒体に記憶されたプログラムが動作することによって実現される。

[0030] データ入力部１１は、学習対象または予測対象のデータを入力する。学習を行う際には、多数のデータをデータ入力部１１より入力する。一方、学習処理が終わった後で予測を行う際には、予測したい１つまたは複数のデータをデータ入力部１１より入力する。このデータ入力部１１の処理は、入力層１０１にデータを入力することに対応する。

[0031] 前半中間層処理部１２は、複数の中間層のうち前半の一部の中間層の処理までを実行し、その結果を中間データとして出力する。図３の例では、前半中間層処理部１２は、データ入力部１１により入力されたデータに対して、１番目の中間層１０２の処理までを実行することに対応する。具体的には、前半中間層処理部１２は、中間層１０２の処理として、データ入力部１１により入力されたデータに対して畳み込み演算処理、活性化処理、プーリング処理を行う。畳み込み演算処理、活性化処理、プーリング処理は何れも、公知の手法を適用してよい。前半中間層処理部１２により処理されたデータは、中間データとしてプーリング層から出力される。

[0032] 変換処理部１３は、前半中間層処理部１２により得られた中間データ（プーリング層の出力データ）に対して不可逆変換処理を行う。不可逆変換処理とは、変換前のデータを完全には復元することができなくする不可逆的な符号化処理のことである。この変換処理部１３による不可逆変換処理は、図３に示す符号化層１０３における符号化処理に対応する。

[0033] ここで、変換処理部１３が行う不可逆変換処理は、不可逆的な符号化処理であればよく、その内容は問わない。一例として、中間層１０２の後段に設けた符号化層１０３を畳み込みニューラルネットワークの全結合層とし、前半中間層処理部１２から得られる複数の中間データ（中間層１０２の各ニューロンから得られる複数のデータ）を結合して出力する全結合処理とすることが可能である。

[0034] このように、前半中間層処理部１２により得られた中間データに対して不可逆変換処理を施すことにより、データ入力部１１により入力された大元のデータの特徴が認識できる程度に中間データの中に残っている場合であって

も、その特徴を認識しにくいデータに変換することができる。また、不可逆変換処理を施した後は、変換前の中間データに復元することが不可能となるので、スマートフォン１０のデータをサーバ２０に提供するユーザのプライバシーを確実に守ることができる。

[0035] なお、上述したように、スマートフォン１０において、大元の入力データの特徴が認識しにくくなる程度まで中間層の処理を実行するように構成する場合は、変換処理部１３を設けることは必須ではない。

[0036] 中間データ出力部１４は、変換処理部１３により不可逆変換処理された中間データをサーバ２０に出力する。サーバ２０の中間データ入力部２１は、スマートフォン１０の中間データ出力部１４より出力された中間データを入力する。中間データ入力部２１が入力する中間データは、図３に示すように、サーバ２０の入力層２０１にセットされるデータである。

[0037] 後半中間層処理部２２は、中間データ入力部２１により入力された中間データに対して、複数の中間層のうち後半の一部の中間層の処理を実行する。図３の例では、後半中間層処理部２２は、中間データ入力部２１により入力された中間データに対して、２番目の中間層２０２および３番目の中間層２０３の処理を実行することに対応する。具体的には、後半中間層処理部２２は、中間層２０２，２０３の処理として、各層において畳み込み演算処理、活性化処理、プーリング処理を順次行う。

[0038] 全結合層処理部２３は、後半中間層処理部２２により得られる複数のデータ（３番目の中間層２０３の各ニューロンから得られる複数のデータ）を結合して出力する。なお、この全結合層処理部２３の処理に対応する処理層は図３には示されていないが、中間層２０３の後段に接続される。データ出力部２４は、全結合層処理部２３により処理されたデータを、最終的な演算結果データとして、出力層２０４から出力する。

[0039] 以上詳しく説明したように、第１の実施形態では、複数階層から成る畳み込みニューラルネットワークによる一連の演算処理を、スマートフォン１０と、それよりも演算処理能力の高いサーバ２０とに分けて実行するようにし

ている。すなわち、スマートフォン１０において、複数の中間層１０２，２０２，２０３のうち前半の一部の中間層１０２の処理までを実行し、その結果を中間データとしてサーバ２０に出力する。そして、サーバ２０において、スマートフォン１０から出力された中間データを入力として、後半の一部の中間層２０２，２０３の処理を実行するようにしている。

[0040] このように構成した第１の実施形態によれば、スマートフォン１０からサーバ２０に出力される中間データは、スマートフォン１０に保持されている元のデータそのものではないので、スマートフォン１０のユーザのプライバシーに係る情報の秘匿性を確保することができる。また、中間データの中に元データの特徴が認識できる程度に残っている可能性を考慮して、中間データに対して不可逆的な符号化処理を行うことにより、ユーザのプライバシーをより強固に守ることができる。

[0041] また、第１の実施形態によれば、ニューラルネットワークによる演算の一部が、演算処理能力の高いサーバ２０で実行されるので、学習処理の演算に要する処理時間を短縮することができる。これにより、第１の実施形態によれば、ユーザのプライバシーに係る情報の秘匿性を保ちつつ、学習処理にかかる時間を短くすることができる。

[0042] （第２の実施形態）

次に、本発明の第２の実施形態を図面に基づいて説明する。上記第１の実施形態では、スマートフォン１０およびサーバ２０において、畳み込みニューラルネットワークによる一連の演算処理を行う例について説明したが、本発明はこれに限定されない。例えば、以下に示す第２の実施形態のように、スマートフォン１０において畳み込みニューラルネットワークによる演算処理を実行し、サーバ２０においてオートエンコーダによる演算処理（自己符号化処理）を実行するようにしてもよい。

[0043] 図５は、サーバ２０において自己符号化処理を行う場合におけるニューラルネットワークの一例を示す図である。図５に示す例では、スマートフォン１０において、入力層１０１に入力されたデータに対し、１番目の中間層１

０２による特徴量抽出処理（畳み込み演算処理、活性化処理、プーリング処理）と、符号化層１０３による不可逆変換処理とを実行し、その結果を中間データとしてサーバ２０に出力する。また、サーバ２０において、スマートフォン１０の符号化層１０３より出力された中間データを入力層２０１に対する入力として、中間層３０２において自己符号化処理を実行し、その結果を出力層３０３に出力する。

[0044] サーバ２０において自己符号化処理を実行する場合、学習処理を行う際に、入力層２０１のデータと同じデータを正解として与える。そして、入力層２０１に中間データを与えたときに、それと同じデータが出力層３０３から出力されるように、入力層２０１の各ニューロンと中間層３０２の各ニューロンとを繋ぐネットワークや、中間層３０２の各ニューロンと出力層３０３の各ニューロンとを繋ぐネットワークに対する重み付けを調整する。

[0045] 図６は、第２の実施形態による演算処理システムの機能構成例を示すブロック図である。なお、この図６において、図４に示した符号と同一の符号を付したものは同一の機能を有するものであるので、ここでは重複する説明を省略する。図６に示すように、サーバ２０は、後半中間層処理部２２および全結合層処理部２３に代えて自己符号化処理部２５を備えている。

[0046] 自己符号化処理部２５は、中間データ入力部２１により入力された入力層２０１の中間データに対して、中間層３０２においてオートエンコーダによる演算処理（自己符号化処理）を実行し、その結果を演算結果データとして出力層３０３に出力する。

[0047] このように、第２の実施形態によれば、スマートフォン１０において実行するニューラルネットワークによる演算処理の内容と、その演算結果である中間データを引き継いでサーバ２０において実行するニューラルネットワークによる演算処理の内容とを異ならせて学習処理や予測処理を実行することができる。このようにすれば、例えば、スマートフォン１０において演算負荷の比較的小さい教師あり学習を行い、演算処理能力の高いサーバ２０において教師なし学習を行うことにより、短時間で高次のディープラーニングを

実現することも可能となる。

[0048] なお、上記第１および第２の実施形態では、サーバ２０に割り当てる中間層の数よりもスマートフォン１０に割り当てる中間層の数を少なくする例について説明したが、本発明はこれに限定されない。例えば、所定数のデータを入力層１０１に与えたときに、スマートフォン１０に割り当てた最終段の中間層に中間データが得られるまでの時間が所定時間以内となるように、スマートフォン１０に所定数の中間層を割り当て、残りの中間層をサーバ２０に割り当てるようにしてもよい。

[0049] 例えば、スマートフォン１０での中間層の処理を１秒以内に終わらせたい場合において、入力層１０１に所定数のサンプルデータを入力したときに、中間層が２階層までであれば１秒以内に処理が終わり、３階層にすると処理が１秒を超えることが分かったとする。この場合、スマートフォン１０に割り当てる中間層の数は、１つまたは２つとする。このようにすれば、スマートフォン１０からサーバ２０に対して学習用に所定数のデータを送信する際に、少なくともスマートフォン１０における処理が所望時間以内に終わるようにすることができる。

[0050] また、上記第１および第２の実施形態では、第１の端末の一例としてスマートフォン１０を用い、第２の端末の一例としてサーバ２０を用いる例について説明したが、本発明はこれに限定されない。第１の端末よりも第２の端末の方が演算処理能力が高い関係を有していれば、第１の端末および第２の端末としてどのようなものを用いてもよい。

[0051] その他、上記第１および第２の実施形態は、何れも本発明を実施するにあたっての具体化の一例を示したものに過ぎず、これらによって本発明の技術的範囲が限定的に解釈されてはならないものである。すなわち、本発明はその要旨、またはその主要な特徴から逸脱することなく、様々な形で実施することができる。

符号の説明

[0052] １０ スマートフォン（第１の端末）

１１　データ入力部

１２　前半中間層処理部

１３　変換処理部

１４　中間データ出力部

２０　サーバ（第２の端末）

２１　中間データ入力部

２２　後半中間層処理部

２３　全結合層処理部

２４　データ出力部

２５　自己符号化処理部

１０１　スマートフォンの入力層

１０２　スマートフォンの中間層

１０３　スマートフォンの符号化層

２０１　サーバの入力層

２０２，３０２　サーバの中間層

２０３　サーバの中間層

２０４，３０３　サーバの出力層

請求の範囲

[請求項1] 入力層と、前階層から入力されるデータに含まれる特徴量を抽出する複数の中間層と、出力層とが階層的に接続されたニューラルネットワークによる演算を実行する演算処理システムであって、

第１の端末において、上記複数の中間層のうち前半の一部の中間層の処理までを実行し、その結果を中間データとして、上記第１の端末より演算処理能力が高い第２の端末に出力し、

上記第２の端末において、上記中間データを入力として、上記複数の中間層のうち後半の一部の中間層の処理を実行するようにしたことを特徴とする階層ネットワークを用いた演算処理システム。

[請求項2] 上記第１の端末は、

上記複数の中間層のうち前半の一部の中間層の処理までを実行し、その結果を中間データとして出力する前半中間層処理部と、

上記前半中間層処理部により得られた上記中間データに対して不可逆変換処理を行う変換処理部と、

上記変換処理部により不可逆変換処理された中間データを上記第２の端末に出力する中間データ出力部とを備えたことを特徴とする請求項１に記載の階層ネットワークを用いた演算処理システム。

[請求項3] 上記変換処理部が行う上記不可逆変換処理は、上記前半中間層処理部から得られる複数の中間データを結合して出力する全結合処理であることを特徴とする請求項２に記載の階層ネットワークを用いた演算処理システム。

[請求項4] 上記第２の端末は、

上記中間データ出力部より出力された上記中間データを入力する中間データ入力部と、

上記中間データ入力部により入力された上記中間データに対して、上記複数の中間層のうち後半の一部の中間層の処理を実行する後半中間層処理部と、

上記後半中間層処理部により得られる複数のデータを結合して出力する全結合層処理部とを備えたことを特徴とする請求項２または３に記載の階層ネットワークを用いた演算処理システム。

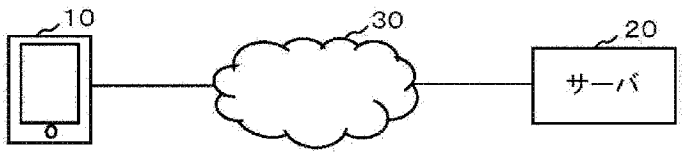
[請求項5] 上記第２の端末は、

上記中間データ出力部より出力された上記中間データを入力する中間データ入力部と、

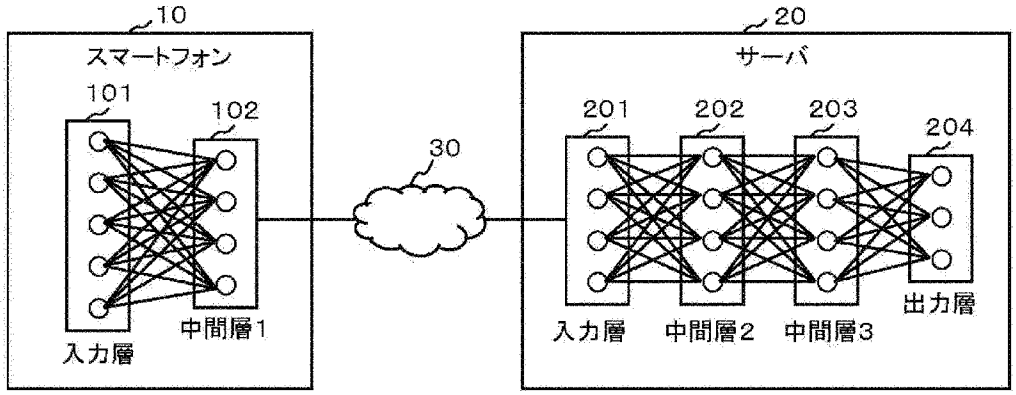
上記中間データ入力部により入力された上記中間データに対して、オートエンコーダによる演算処理を実行する自己符号化処理部とを備えたことを特徴とする請求項２または３に記載の階層ネットワークを用いた演算処理システム。

[請求項6] 上記第１の端末では、畳み込みニューラルネットワークによる演算処理を実行し、上記第２の端末では、上記オートエンコーダによる演算処理を実行することを特徴とする請求項１～３の何れか１項に記載の階層ネットワークを用いた演算処理システム。

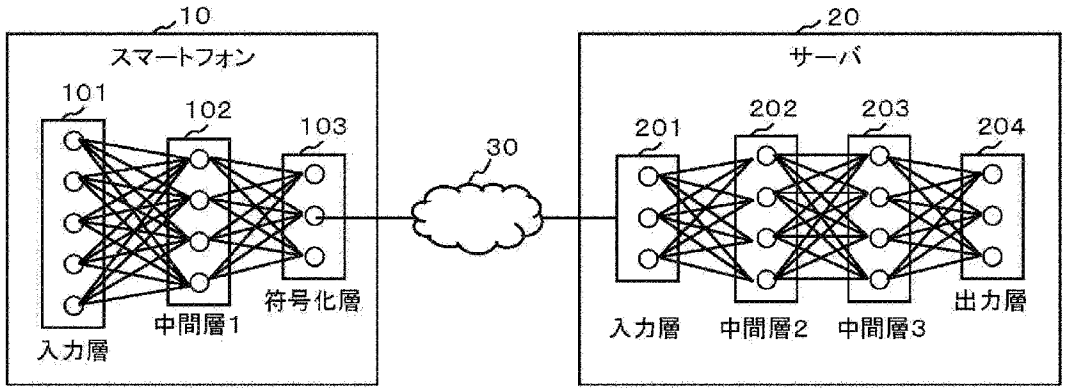
[図1]


10
30
20
サーバ

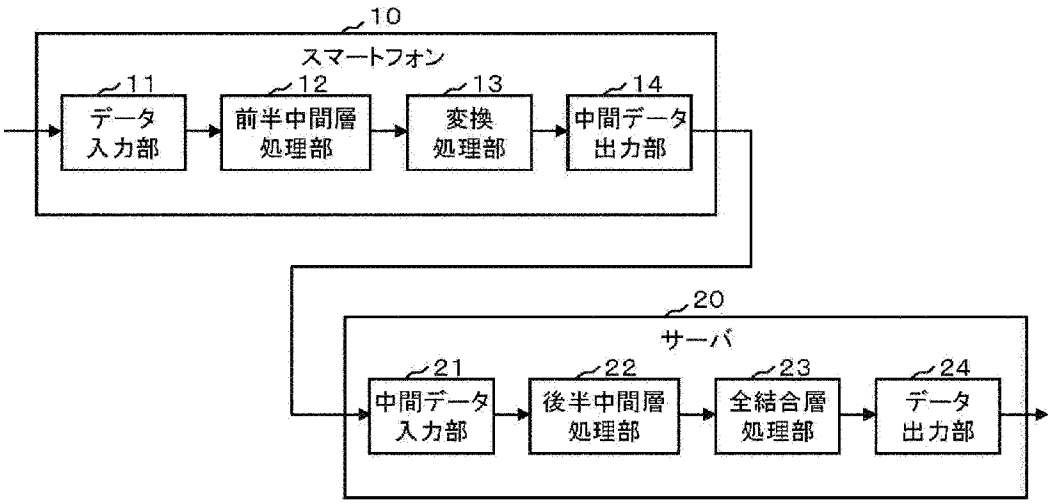
[図2]


10
スマートフォン
101
102
入力層
中間層1
30
20
サーバ
201
202
203
204
入力層
中間層2
中間層3
出力層

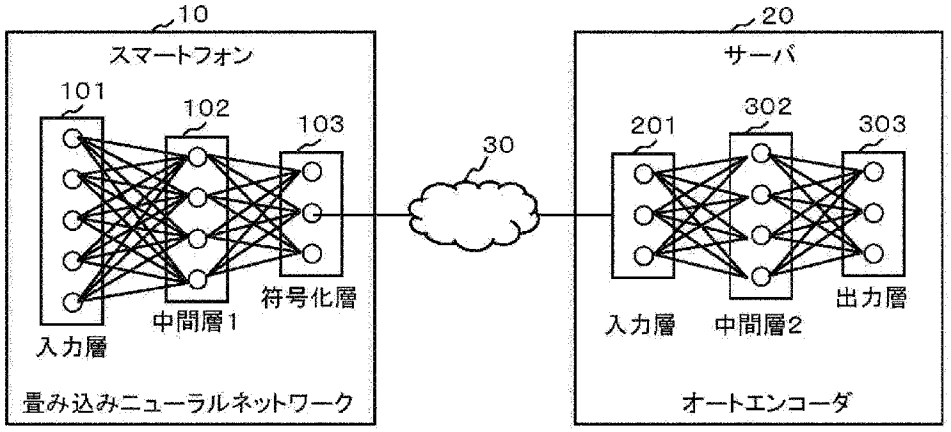
[図3]


10
スマートフォン
101
102
103
30
20
サーバ
201
202
203
204
入力層
中間層1
符号化層
入力層
中間層2
中間層3
出力層

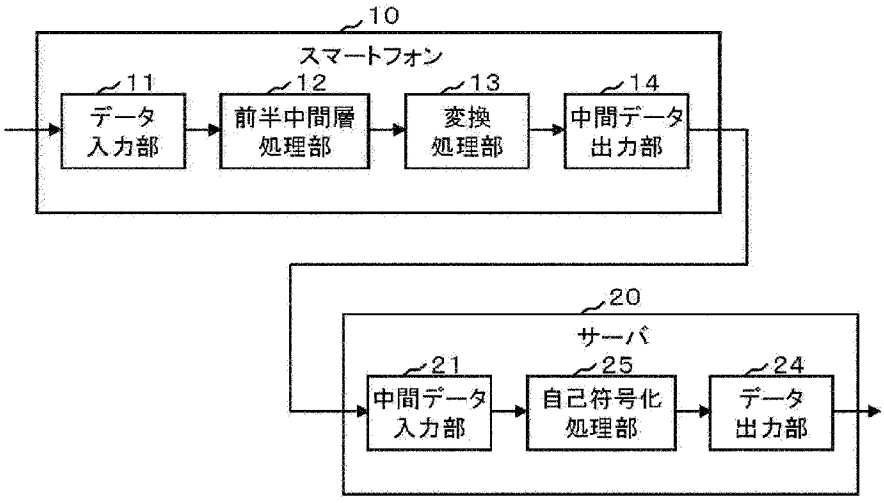
[図4]


10
スマートフォン
11
データ
入力部
12
前半中間層
処理部
13
変換
処理部
14
中間データ
出力部
20
サーバ
21
中間データ
入力部
22
後半中間層
処理部
23
全結合層
処理部
24
データ
出力部

[図5]


10
スマートフォン
101
102
103
入力層
中間層1
符号化層
畳み込みニューラルネットワーク
30
20
サーバ
201
302
303
入力層
中間層2
出力層
オートエンコーダ

[図6]


10
スマートフォン
11
データ
入力部
12
前半中間層
処理部
13
変換
処理部
14
中間データ
出力部
20
サーバ
21
中間データ
入力部
25
自己符号化
処理部
24
データ
出力部

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2016/070376

A. CLASSIFICATION OF SUBJECT MATTER
G06N3/04(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06N3/04

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Jitsuyo Shinan Koho 1922-1996 Jitsuyo Shinan Toroku Koho 1996-2016
Kokai Jitsuyo Shinan Koho 1971-2016 Toroku Jitsuyo Shinan Koho 1994-2016

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
CiNii, IEEE Xplore, THE ACM DIGITAL LIBRARY

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X Y A	Ayae ICHINOSE et al., "Deep Learning Framework Caffe no Bunsan Kankyo eno Tekiyo", DEIM Forum2016, [online], 29 February 2016 (29.02.2016), [retrieval date 04 August 2016 (04.08.2016)], Internet <URL:http://db-event.jpn.org/deim2016/papers/134.pdf> pages 1 to 5	1-2,4 5-6 3
Y A	Kenta MURATA, Dai 3 Sho Shinso Gakushu Nyumon Kaiso ga Fuete Okiru Mondai to sono Kaiketsu Hoho, WEB+DB PRESS, vol.89, 25 November 2015 (25.11.2015), pages 62 to 65	5-6 1-4
A	US 2012/0254086 A1 (MICROSOFT CORP.), 04 October 2012 (04.10.2012), whole document & CN 102737278 A	1-6

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
05 August 2016 (05.08.16)	16 August 2016 (16.08.16)
Name and mailing address of the ISA/ Japan Patent Office 3-4-3,Kasumigaseki,Chiyoda-ku, Tokyo 100-8915,Japan	Authorized officer Telephone No.

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２０１６／０７０３７６

Ａ．　発明の属する分野の分類（国際特許分類（ＩＰＣ））

Int.Cl.　G06N3/04(2006.01)i

Ｂ．　調査を行った分野

調査を行った最小限資料（国際特許分類（ＩＰＣ））

Int.Cl.　G06N3/04

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	１９２２－１９９６年
日本国公開実用新案公報	１９７１－２０１６年
日本国実用新案登録公報	１９９６－２０１６年
日本国登録実用新案公報	１９９４－２０１６年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

CiNii
IEEE Xplore
THE ACM DIGITAL LIBRARY

Ｃ．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
X Y A	一瀬 絢衣 ほか，ディープラーニングフレームワーク Caffe の分散環境への適用,DEIM Forum2016，［オンライン］，2016.02.29，［検索日 2016.08.04］,インターネット <URL:http://db-event.jpn.org/deim2016/papers/134.pdf> pp.1-5	1-2,4 5-6 3
Y A	村田　賢太，第３章　深層学習入門　階層が増えて起きる問題とその解決方法，ＷＥＢ＋ＤＢ　ＰＲＥＳＳ　Ｖｏｌ．８９，2015.11.25，pp.62-65	5-6 1-4

☑　Ｃ欄の続きにも文献が列挙されている。　　☐　パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー
「Ａ」特に関連のある文献ではなく、一般的技術水準を示すもの
「Ｅ」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
「Ｌ」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「Ｏ」口頭による開示、使用、展示等に言及する文献
「Ｐ」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
「Ｔ」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「Ｘ」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Ｙ」特に関連のある文献であって、当該文献と他の１以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「＆」同一パテントファミリー文献

国際調査を完了した日	国際調査報告の発送日
０５．０８．２０１６	１６．０８．２０１６

国際調査機関の名称及びあて先
日本国特許庁（ＩＳＡ／ＪＰ）
郵便番号１００－８９１５
東京都千代田区霞が関三丁目４番３号

特許庁審査官（権限のある職員）　５Ｂ　３４６５
長谷川　篤男
電話番号　０３－３５８１－１１０１　内線　３５４５

Ｃ（続き）．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
A	US 2012/0254086 A1 (MICROSOFT CORPORATION) 2012.10.04, whole document & CN 102737278 A	1-6