

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7435793号
(P7435793)

(45)発行日 令和6年2月21日(2024.2.21)

(24)登録日 令和6年2月13日(2024.2.13)

(51)国際特許分類

F I

G O 6 N 20/00 (2019.01)

G O 6 N 20/00

G O 6 N 3/02 (2006.01)

G O 6 N 3/02

請求項の数 6 (全22頁)

(21)出願番号	特願2022-541409(P2022-541409)	(73)特許権者	000004226
(86)(22)出願日	令和2年8月5日(2020.8.5)		日本電信電話株式会社
(86)国際出願番号	PCT/JP2020/030021		東京都千代田区大手町一丁目5番1号
(87)国際公開番号	WO2022/029927	(74)代理人	100098394
(87)国際公開日	令和4年2月10日(2022.2.10)		弁理士 山川 茂樹
審査請求日	令和4年11月25日(2022.11.25)	(74)代理人	100153006
			弁理士 小池 勇三
		(74)代理人	100064621
			弁理士 山川 政樹
		(74)代理人	100121669
			弁理士 本山 泰
		(72)発明者	有川 勇輝
			東京都千代田区大手町一丁目5番1号
			日本電信電話株式会社内
		(72)発明者	坂本 健

最終頁に続く

(54)【発明の名称】 推論処理装置

(57)【特許請求の範囲】

【請求項1】

学習済みニューラルネットワークを用いて、入力データの特徴を推論する推論処理装置であって、
前記入力データを記憶する第1記憶部と、
前記学習済みニューラルネットワークの重みを記憶する第2記憶部と、
受信した前記入力データのうち特定の入力データのみを抽出するデータフィルタ部と、
前記データフィルタ部が抽出した前記特定の入力データ、および前記重みを入力として、
前記学習済みニューラルネットワークの推論演算を実行し、前記入力データの特徴を推論する推論演算部と
を備え、

前記第1記憶部は、複数の異なるデータ発生元から複数の入力データを受信し、記憶するものであり、

前記データフィルタ部は、前記複数の入力データ間の類似性を判断し、前記複数の入力データに類似する入力データがないと判断した場合には、前記複数の入力データを前記推論演算部への入力データとして抽出し、前記複数の入力データに類似するデータがあると判断した場合には、前記複数の入力データのうち、類似しない入力データと、類似する複数の入力データのうちのいずれか一つの入力データを前記推論演算部への入力データとして抽出するように構成される

推論処理装置。

【請求項 2】

学習済みニューラルネットワークを用いて、入力データの特徴を推論する推論処理装置であって、
前記入力データを記憶する第 1 記憶部と、
前記学習済みニューラルネットワークの重みを記憶する第 2 記憶部と、
受信した前記入力データのうち特定の入力データのみを抽出するデータフィルタ部と、
前記データフィルタ部が抽出した前記特定の入力データ、および前記重みを入力として、
前記学習済みニューラルネットワークの推論演算を実行し、前記入力データの特徴を推論する推論演算部と
を備え、

10

前記第 1 記憶部は、複数の異なるデータ発生元から複数の入力データを受信し、記憶するものであり、

前記データフィルタ部は、前記複数の入力データ間の類似性と、受信した前記複数の入力データと以前に推論演算した入力データとの類似性の両方を判断するように構成され、

受信した前記複数の入力データの各々が、前記複数の入力データの他の入力データおよび以前に推論演算した入力データと類似でないと判断した場合には、前記受信した複数の入力データを前記推論演算部への入力データとして抽出し、

前記複数の入力データに類似する入力データがあると判断した場合には、類似する入力データのうちいずれか一つの入力データを抽出し、当該抽出した入力データが、以前に推論演算した入力データと類似でないと判断した場合には、当該抽出した入力データを前記推論演算部への入力データとして抽出するように構成される

20

推論処理装置。

【請求項 3】

請求項 1 または 2 記載の推論処理装置であって、

前記データフィルタ部は、受信した入力データと以前に推論演算した入力データとの類似性を判断し、類似でないと判断した場合には、前記受信した入力データを前記推論演算部への入力データとして抽出し、類似であると判断した場合には、前記受信した入力データを前記推論演算部への入力データとして抽出しないように構成される

推論処理装置。

【請求項 4】

30

請求項 1 から 3 の何れか 1 項に記載の推論処理装置であって、

前記データフィルタ部は、

前記入力データの差分と予め設定した閾値と比較する比較部を備え、前記比較部の比較結果に基づいて、前記類似性の有無を判断するように構成される

推論処理装置。

【請求項 5】

請求項 1 から 4 の何れか 1 項に記載の推論処理装置であって、

前記データフィルタ部は、前記推論演算部の出力データを前記データフィルタ部への入力データとして用いるように構成される

推論処理装置。

40

【請求項 6】

請求項 1 から 5 の何れか 1 項に記載の推論処理装置であって、

前記推論演算部は、前記推論演算部の出力データを前記推論演算部への入力データとして用いるように構成される

推論処理装置。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、推論処理装置に関し、特に、ニューラルネットワークを用いて推論を行う技術に関する。

50

【背景技術】

【0002】

近年、モバイル端末やInternet of Things (IoT) デバイス等のエッジデバイスの増加に伴い、生成されるデータが爆発的に増加している。この膨大なデータから有意義な情報を抽出するには、深層ニューラルネットワーク (Deep Neural Networks: DNN) とよばれる最先端の機械学習技術が優位である。近年のDNNに関する研究の進歩により、データの解析精度は大幅に改善されており、DNNを利用した技術のさらなる発展が期待されている。

【0003】

DNNの処理は学習および推論の2つのフェーズがある。一般に、学習では、大量のデータを必要とするため、クラウドで処理される場合がある。一方、推論では学習済みのDNNモデルを使用し、未知の入力データに対して出力を推定する。

【0004】

より詳細に説明すると、DNNにおける推論処理では、学習済みのニューラルネットワークモデルに時系列データまたは画像データなどの入力データを与えて、入力データの特徴を推論する。例えば、非特許文献1に開示されている具体的な例によると、加速度センサとジャイロセンサとを搭載したセンサ端末を用いて、ゴミ収集車の回転や停止といったイベントを検出することで、ゴミの量を推定している。このように、未知の時系列データを入力として、各時刻におけるイベントを推定するには、予め各時刻におけるイベントが既知である時系列データを用いて学習したニューラルネットワークモデルを用いる。

【0005】

非特許文献1では、センサ端末から取得される時系列データを入力データとして用いており、リアルタイムにイベントを抽出する必要がある。そのため、推論処理をより高速化することが必要となる。従来、処理を実現するFPGAをセンサ端末に搭載し、そのようなFPGAで推論演算を行い、処理の高速化を図っている (例えば、非特許文献2参照)。

【先行技術文献】

【非特許文献】

【0006】

【文献】Kishino, et. al, "Detecting garbage collection duration using motion sensors mounted on a garbage truck toward smart waste management", SPWID17

【文献】Kishino, et. al, "Datafying city: detecting and accumulating spatio-temporal events by vehicle-mounted sensors", BIGDATA 2017

【発明の概要】

【発明が解決しようとする課題】

【0007】

しかし、従来の技術では、全ての入力データに対して推論演算処理を行うため、消費電力に対する制約が大きい小型組込みデバイスへのニューラルネットワーク処理を搭載することが困難であった。そのため、推論演算処理に伴う消費電力を削減しながら、推論演算処理を高速化することが困難であった。

【0008】

本発明は、上述した課題を解決するためになされたものであり、推論演算処理に伴う消費電力を削減しながら、推論演算処理を高速化することができる推論処理技術を提供することを目的とする。

【課題を解決するための手段】

【0009】

上述した課題を解決するために、本発明に係る推論処理装置は、学習済みニューラルネットワークを用いて、入力データの特徴を推論する推論処理装置であって、前記入力デ

10

20

30

40

50

ータを記憶する第1記憶部と、前記学習済みニューラルネットワークの重みを記憶する第2記憶部と、受信した前記入力データのうち特定の入力データのみを抽出するデータフィルタ部と、前記データフィルタ部が抽出した前記特定の入力データ、および前記重みを入力として、前記学習済みニューラルネットワークの推論演算を実行し、前記入力データの特徴を推論する推論演算部とを備える。

【発明の効果】

【0010】

本発明によれば、推論演算処理に伴う消費電力を削減しながら、推論演算処理を高速化することができる。

【図面の簡単な説明】

【0011】

【図1】図1は、第1の実施の形態にかかる推論処理装置の構成を示すブロック図である。

【図2A】図2Aは、第1の実施の形態にかかる推論処理装置におけるデータフィルタ部の構成を示すブロック図である。

【図2B】図2Bは、第1の実施の形態にかかる推論処理装置におけるデータフィルタ部の処理を説明する図である。

【図3】図3は、第1の実施の形態にかかる推論処理装置の他の構成を示すブロック図である。

【図4】図4は、第1の実施の形態にかかる推論処理装置の構成を示すブロック図である。

【図5】図5は、第1の実施の形態にかかる推論処理装置の構成を示すブロック図である。

【図6】図6は、第1の実施の形態にかかる推論処理装置におけるデータフィルタ部の動作を示すフローチャートである。

【図7】図7は、第2の実施の形態にかかる推論処理装置の構成を示すブロック図である。

【図8】図8は、第2の実施の形態にかかる推論処理装置の構成を示すブロック図である。

【図9】図9は、第2の実施の形態にかかる推論処理装置の構成を示すブロック図である。

【図10】図10は、第3の実施の形態にかかる推論処理装置の構成を示すブロック図である。

【図11】図11は、第3の実施の形態にかかる推論処理装置におけるデータフィルタ部の構成を示すブロック図である。

【図12】図12は、第3の実施の形態にかかる推論処理装置の構成を示すブロック図である。

【図13】図13は、第3の実施の形態にかかる推論処理装置の構成を示すブロック図である。

【図14】図14は、第3の実施の形態にかかる推論処理装置におけるデータフィルタ部の動作を示すフローチャートである。

【図15】図15は、第4の実施の形態にかかる推論処理装置の構成を示すブロック図である。

【図16】図16は、第4の実施の形態にかかる推論処理装置の構成を示すブロック図である。

【図17】図17は、第4の実施の形態にかかる推論処理装置の構成を示すブロック図である。

【図18】図18は、第5の実施の形態にかかる推論処理装置の構成を示すブロック図である。

【図19】図19は、第5の実施の形態にかかる推論処理装置におけるデータフィルタ部の動作を示すフローチャートである。

【図20】図20は、本発明の実施の形態に係る推論処理装置のハードウェア構成を示すブロック図である。

【図21】図21は、従来の推論処理装置の構成を示すブロック図である。

【発明を実施するための形態】

【0012】

10

20

30

40

50

以下、本発明の好適な実施の形態について、図面を参照して説明する。本発明は、以下の実施の形態に限定されるものではない。

【0013】

〔第1の実施の形態〕

図1～5を参照して、本発明の第1の実施の形態にかかる推論処理装置の構成について説明する。図1は、第1の実施の形態にかかる推論処理装置の構成を示すブロック図である。図2Aは、第1の実施の形態にかかる推論処理装置におけるデータフィルタ部の構成を示すブロック図である。図2Bは、第1の実施の形態にかかるデータフィルタ部の処理を説明する図である。図3は、第1の実施の形態にかかる推論処理装置の他の構成を示すブロック図である。図4は、第1の実施の形態にかかる推論処理装置の構成を示すブロック図である。図5は、第1の実施の形態にかかる推論処理装置の構成を示すブロック図である。

10

【0014】

本発明の推論処理装置1は、全体として、所定の学習データを用いて重みの値を学習したニューラルネットワークモデルを用いて、未知の入力データに対する推論処理を行う。本発明の推論処理装置1の外部から取得された音声データや言語データなどの時系列データ、または画像データを推論対象の入力データとして用いる。推論処理装置1は、学習済みのニューラルネットワークモデルを用いてニューラルネットワークの演算をバッチ処理し、入力データの特徴を推論する。

【0015】

より詳細には、推論処理装置1は、各時刻におけるイベントが既知である時系列データなどの入力データを用いて予め学習したニューラルネットワークモデルを用いる。推論処理装置1は、未知の時系列データなどの入力データおよび学習済みのニューラルネットワークの重みデータを入力として、各時刻におけるイベントの推定を行う。なお、入力データおよび重みデータは、行列データである。

20

【0016】

例えば、推論処理装置1は、加速度センサとジャイロセンサとを搭載したセンサから取得された入力データを用いて、ごみ収集車の回転や停止といったイベントを検出することで、ゴミの量を推定することができる（非特許文献1参照）。

【0017】

推論処理装置1は、入力データを記憶する第1記憶部10と、学習済みニューラルネットワークの重みを記憶する第2記憶部12と、前記入力データのうち特定のデータのみを抽出して推論演算部13への入力データとするデータフィルタ部11と、前記データフィルタ部11が抽出した入力データ、および前記学習済みニューラルネットワークの重みを入力として、前記学習済みニューラルネットワークの推論演算を実行し、入力データの特徴を推論する推論演算部13と、を備えている。

30

【0018】

第1記憶部10は、入力データを記憶する機能を有している。前記第2記憶部12は、学習済みニューラルネットワークモデル即ち重みデータを記憶する機能を有している。

【0019】

推論演算部13は、前記入力データと、重みデータと、出力データと、を入力として、ニューラルネットワークの演算を行い、その結果を出力する機能を有している。なお、前記入力データが入力されない期間は、当該推論演算部13は、推論演算処理を行わない。

40

【0020】

さらに、この場合、当該推論演算部に対するクロック供給を停止（クロックゲーティング）したり、電源供給を停止（パワーゲーティング）したり、することもあり、消費電力を削減する。また、推論演算処理を行わない期間、当該推論演算部13は、演算処理をせずとも直前の推論結果を上位装置やユーザ装置などの外部に対して出力することもある。

【0021】

データフィルタ部11は、入力データのうち特定のデータのみを抽出して推論演算部1

50

3へ入力する機能を有している。具体的には、入力データと以前に推論演算したデータとの類似性を判断し、類似しない入力データを抽出して推論演算部13に入力する。入力データの類似性を判断して、推論演算処理の結果が同じになる類似する入力データに対する推論演算処理を行わなくとも済むように構成したので、全ての入力データに対して推論演算処理を行う必要がなくなり、推論演算処理に伴う消費電力を削減しながら、推論演算処理を高速化することができる。

【0022】

例えば、図2Aに示すように、直前に推論演算部13にて推論処理し推論結果を出力した際に用いた入力データを保持部120で保持しておき、比較部110において、入力データと直前に推論演算部13にて推論処理し推論結果を出力した際に用いた入力データとを比較し、出力制御部130において、その比較結果に基づいて、入力データを推論演算部13へ出力するかを判断する。

10

【0023】

出力制御部130では、入力データと直前に推論処理し推論結果を出力した際に用いた入力データとの差分が閾値以上である場合、当該入力データを推論演算部13へ入力し、一方、その差分が閾値より小さい場合、当該入力データを推論演算部13へ入力せずに、この場合の当該時刻における推論結果は、直前に推論演算部13にて推論処理して得られた推論結果を用いる。

【0024】

図2Aでは、入力データを直前に推論演算した入力データと比較したが、以前に推論演算した入力データを複数保持しておき、それらの複数の保持データと入力データを比較して、入力データを推論演算部13へ入力するかを判断してもよい。例えば、入力データと複数の保持データの何れか1つのデータとの差分が閾値より小さい場合、当該入力データを推論演算部13へ入力しないように構成することもできる。

20

【0025】

図2Bは、第1の実施の形態にかかるデータフィルタ部の処理を説明する図である。データフィルタ部11は、類似な入力データに対する後段の推論演算処理により得られる結果は変化しないことを利用して、入力データの類似性を用いて、類似するデータについて推論演算部13における推論演算処理を行わなくとも済むように構成することができる。これにより、推論演算処理に伴う消費電力を削減しながら、推論演算処理を高速化できる効果が得られる。

30

【0026】

なお、入力データが複数の要素、例えば、加速度センサとジャイロセンサを搭載したセンサから取得された複数のデータを含む場合には、データフィルタ部11における入力データの比較では、入力データの要素ごとに第1閾値を使って比較し、差分が第1閾値以上の要素の数が第2閾値以上の場合に、入力データを推論演算部13へ入力すると判断し、差分が第1閾値以上の要素の数が第2閾値より少ない場合に、入力データを推論演算部13へ入力しないと判断すればよい。

【0027】

なお、上記の例では、入力データのみを用いて差分の比較を行う例を示したが、比較対象のデータは入力データに限らない。例えば、図3に示すように、入力データと以前のサイクルにて推論演算部13が出力した推論結果をフィードバックとして用いる場合は、当該フィードバックデータ即ち出力データをデータフィルタ部11への入力データとして用いて比較を行うこともある。この場合、各々の比較結果の論理和または論理積をとって差分あり／なしを判断する。

40

【0028】

なお、上記の例では、入力データと重みデータを用いて推論演算を行う例を示したが、推論演算処理の方法はこれに限らない。例えば、図4に示すように、推論演算結果を次サイクルの推論演算処理の入力として用いる、即ち出力フィードバックを行うこともある。

【0029】

50

この場合、推論処理装置 1 はさらに、推論演算部 1 3 からフィードバックされる出力データを保持する第 3 記憶部 1 4 をさらに備える。出力フィードバックを行うことで、文字列や音声・言語処理などの時系列データに適した推論演算を実施できる効果がある。また、図 5 に示すように、出力フィードバックを第 3 記憶部 1 4 に入力するのではなく、直接推論演算部 1 3 の中でフィードバックすることもある、これにより推論処理装置 1 に搭載するメモリ容量を削減できる効果がある。

【0030】

なお、データフィルタ部 1 1 にて入力データに対してサンプリング周期を長周期化することで、データ量を削減することもある。

【0031】

なお、上記の例では、入力データと、直前に推論演算部 1 3 にて推論処理し推論結果を出力した際に用いた入力データと、を比較する例を示したが、比較対象のデータはこれに限らない。例えば、所定のサイクル数前以降に推論演算部 1 3 にて推論処理した推論結果および当該推論処理に用いた入力データと、を記憶しておき、入力データと所定のサイクル数前の入力データとを比較する。

【0032】

なお、図 2 1 に示す従来の推論処理装置 1 との違いは、前記データフィルタ部 1 1 を備える点である。従来の推論処理装置 1 は、全ての入力データに対して、推論演算処理を行うのに対して、本発明の推論処理装置 1 はデータフィルタ部 1 1 にて特定の入力データのみを抽出する点にある。

【0033】

〔第 1 の実施の形態の動作〕

次に、図 6 を参照して、第 1 の実施の形態にかかる推論処理装置 1 におけるデータフィルタ部 1 1 の動作について説明する。図 6 は、第 1 の実施の形態にかかる推論処理装置におけるデータフィルタ部の動作を示すフローチャートである。

【0034】

まず、データフィルタ部 1 1 は、入力データと過去に推論演算処理した入力データとの差分を検出するのに用いる閾値を設定する（ステップ S 1 - 1）。閾値は、初期設定として動作開始時点で予め設定する以外に、動作中に閾値を動的に変更することもある。

【0035】

例えば、ある時刻において用いている閾値に対して、得られた推論処理結果に差がみられない場合、閾値を大きくすることもある。また、推論処理結果に対する推論精度が所望の精度よりも低い場合、閾値を小さくすることで、より多くの入力データに対して推論処理を行うことになるので、精度を改善できることが期待できる。このように、推論演算結果に応じて、入力データの類似性比較に用いる閾値を動的に設定することもある。

【0036】

次に、データフィルタ部 1 1 は、入力データと直前に推論演算した入力データと、を取得して（ステップ S 1 - 2）、以前に推論演算処理した過去の入力データとの差分を計算する（ステップ S 1 - 3）。過去の入力データとしては、例えば、直前に入力され推論処理された入力データを用いることができる。

【0037】

計算した差分が前記閾値と比較して、その差分が閾値以上である場合（ステップ S 1 - 4：Yes）、当該入力データを推論演算部 1 3 へ出力する（ステップ S 1 - 5）。一方、その差分が閾値より小さい場合（ステップ S 1 - 4：No）、当該入力データを推論演算部 1 3 へ出力しない（ステップ S 1 - 6）。この場合の推論結果は、過去の入力データを用いて推論演算部 1 3 にて推論処理して得られた推論結果を用いる。

【0038】

このように、類似な入力データに対する後段の推論演算処理により得られる結果は変化しないため、データフィルタ部 1 1 は、入力データの過去に推論演算を行ったデータに対する類似性を判断して、過去の入力データに類似する入力データを推論演算部 1 3 に出力

10

20

30

40

50

しないように構成することができる。これにより、推論演算部 13 は、過去に推論演算した入力データに類似する入力データに対する推論演算処理を行わなくとも済むので、推論演算処理の高速化や、推論演算処理に伴う消費電力の削減を実現することができる。

【0039】

〔第 1 の実施の形態の効果〕

このように、本実施の形態の推論処理装置は、所定の学習データを用いて重みの値を学習したニューラルネットワークモデルを用いて、未知の入力データに対する推論処理を行うために、第 1 記憶部 10 は入力データを記憶し、第 2 記憶部 12 は、学習済みニューラルネットワークの重みを記憶し、データフィルタ部 11 は、前記入力データのうち特定のデータのみを抽出して推論演算部 13 へ入力データとし、推論演算部 13 は、データフィルタ部 11 が抽出した入力データ、および前記学習済みニューラルネットワークの重みを入力として、前記学習済みニューラルネットワークの推論演算を実行し、前記入力データの特徴を推論する。

10

【0040】

これにより、データフィルタ部 11 は、類似な入力データに対する後段の推論演算処理により得られる結果は変化しないことを利用して、入力データの類似性を判断し、類似するデータについて推論演算処理を行わなくとも済むように構成することができる。本発明の推論処理装置 1 は、全ての入力データに対して推論処理を行う従来の推論処理装置と比較して、推論演算処理を高速化し、推論演算処理に伴う消費電力を削減することが可能となる。また、全ての入力データに対して推論処理をせずともよいため、当該推論処理装置 1 からの推論結果の出力も削減できることから、通信ネットワークの負荷を低減することもできる。

20

【0041】

〔第 2 の実施の形態〕

図 7～9 を参照して、本発明の第 2 の実施の形態にかかる推論処理装置 1 について説明する。図 7 は、第 2 の実施の形態にかかる推論処理装置の構成を示すブロック図である。図 8 は、第 2 の実施の形態にかかる推論処理装置の構成を示すブロック図である。図 9 は、第 2 の実施の形態にかかる推論処理装置の構成を示すブロック図である。

【0042】

第 1 の実施の形態との違いは、記憶部の前段にデータフィルタ部 11 を備えて、入力データのうち特定のデータのみを抽出した後、記憶部に格納する点である。この場合、メモリ制御部は、記憶部に記憶されている入力データ、即ち推論演算処理待ちの入力データの有無を判断し、後段の推論演算部 13 へ入力する。このように、第 1 記憶部 10 の前段にデータフィルタ部 11 を配置する構成とすることで、第 1 記憶部 10 が使用するメモリ量を削減できる効果がある。

30

【0043】

なお、上記の例では、入力データと重みデータを用いて推論演算を行う例を示したが、推論演算処理の方法はこれに限らない。例えば、図 8 に示すように、推論演算結果を次サイクルの推論演算処理の入力として用いる、即ち出力フィードバックを行うこともある。出力フィードバックを行うことで、文字列や音声・言語処理などの時系列データに適した推論演算を実施できる効果がある。また、図 9 に示すように、出力フィードバックを記憶部に入力するのではなく、直接推論演算部 13 の中でフィードバックすることもある、これにより記憶部で消費するメモリ量を削減できる効果がある。

40

【0044】

〔第 2 の実施の形態の効果〕

このように、本実施の形態は、所定の学習データを用いて重みの値を学習したニューラルネットワークモデルを用いて、未知の入力データに対する推論処理を行うために、第 1 記憶部 10 は入力データを記憶し、第 2 記憶部 12 は学習済みニューラルネットワークの重みを記憶し、データフィルタ部 11 は前記入力データのうち特定のデータのみを抽出して前記推論演算部 13 へ入力データとし、推論演算部 13 はデータフィルタ部 11 が抽

50

出した入力データ、および前記学習済みニューラルネットワークの重みを入力として、前記学習済みニューラルネットワークの推論演算を実行し、前記入力データの特徴を推論する。

【0045】

これにより、データフィルタ部11において入力データの類似性を判断して、過去に推論演算した入力データに類似する入力データに対する推論演算処理を行わなくとも済むように構成することができる。本発明の推論処理装置1は、全ての入力データに対して推論処理を行う従来の推論処理装置と比較して、推論演算処理を高速化し、推論演算処理に伴う消費電力を削減することが可能となる。

【0046】

また、第1記憶部10の前段にデータフィルタ部11を配置する構成とすることで、第1記憶部10が使用するメモリ量を削減できる効果がある。

【0047】

また、全ての入力データに対して推論処理をせずともよい場合、当該推論処理装置1からの推論結果の出力も削減できることから、通信ネットワークの負荷を低減できる。

【0048】

〔第3の実施の形態〕

図10～13を参照して、本発明の第3の実施の形態にかかる推論処理装置1の構成について説明する。図10は、第3の実施の形態にかかる推論処理装置の構成を示すブロック図である。図11は、第3の実施の形態にかかる推論処理装置におけるデータフィルタ部の構成を示すブロック図である。図12は、第3の実施の形態にかかる推論処理装置の構成を示すブロック図である。図13は、第3の実施の形態にかかる推論処理装置の構成を示すブロック図である。

【0049】

第1および第2の実施の形態との違いは、複数のデータ発生元からの入力データを受信し、それらの入力データに対する推論演算処理を行い、推論結果を出力する推論処理装置1であって、さらに、同時刻における複数の入力データ間の類似性を検出するデータフィルタ部11を備える点である。

【0050】

データフィルタ部は、複数の入力データのうち特定のデータのみを抽出して推論演算部13へ入力する機能を有している。具体的には、図11に示すように、複数の入力データ同士を比較し、その差分が閾値以下である場合、比較した入力データのうちのいずれか一方の入力データに対してのみ推論演算処理を行う。この場合、推論演算処理を行わなかった入力データの推論結果は、比較した入力データに対して推論演算部13にて推論処理して得られた推論結果と同じものを用いる。一方、差分が閾値より大きい場合、両者の推論演算処理の出力結果は異なるため、両者の入力データに対して推論演算処理を行う。

【0051】

このように、データフィルタ部11は複数の異なる入力データ発生元の入力データ同士の類似性を検出して、類似な入力データに対する後段の推論演算処理により得られる結果は同じであるため、推論演算処理を行わなくとも済む。これにより、推論演算処理を高速化できたり、推論演算処理に伴う消費電力を削減できたりする効果がある。

【0052】

なお、比較する入力データは所定の組合せで行う。所定の組合せは、例えば、入力データの発生元の物理的な距離が最も近い入力データ同士を比較したり、入力データの発生元に付与されている識別子の順番に対応した入力データ同士を比較したりする。

【0053】

なお、入力データ同士を比較する回数は1段に限らず、異なる入力データ同士の組合せにて複数回比較を行うこともある。

【0054】

なお、推論演算部13は、複数の入力データに対する推論処理を並列に行うこともある

10

20

30

40

50

。これにより、推論演算処理を高速化できる効果がある。

【0055】

なお、上記の例では、比較する入力データは所定の入力元からの入力データ同士を比較する例を示したが、必ずしも特定の入力データ同士を比較しなくともよい。すなわち、任意の入力データ同士を比較してもよい。例えば、入力データ発生元の端末等が時間に対して物理的に移動する移動端末である場合、その時刻において、物理的に距離が近い移動端末同士を組合せて、入力データを比較しても良い。

【0056】

なお、上記の例では、推論処理の高速化や低電力化のために入力データを削減する。このため、全ての入力データの組合せに対して網羅的に類似性を探索するのではなく、ある一部の入力データの組合せに対してのみ類似性を比較する例を示したが、必ずしも類似性の検出方法はこれに限らない。

【0057】

例えば、当該推論処理装置1に入力される入力データの発生元に対して全ての組合せで類似性を比較することが、後段の推論処理よりも高速に実行でき、かつ類似性の検出に要する電力が後段の推論処理よりも低い場合、網羅的に類似性を探索して、より高速かつ低電力に推論処理を実現することもある。

【0058】

なお、上記の例では、類似性の検出に用いる閾値を初期設定として与える例を示したが、閾値の設定方法はこれに限らない。例えば、ある時刻において用いている閾値に対して、得られた推論処理結果に差がみられない場合、閾値を大きくすることもある。また推論処理結果に対する推論精度が所望の精度よりも低い場合、閾値を小さくすることで、より多くの入力データに対して推論処理を行うことになるので、精度を改善できることが期待できる。このように、推論演算結果に応じて、入力データの類似性比較に用いる閾値を動的に設定することもある。

【0059】

なお、上記の例では、入力データと重みデータを用いて推論演算を行う例を示したが、推論演算処理の方法はこれに限らない。例えば、図12に示すように、推論演算結果を次サイクルの推論演算処理の入力として用いる、即ち出力フィードバックを行うこともある。出力フィードバックを行うことで、文字列や音声・言語処理などの時系列データに適した推論演算を実施できる効果がある。また、図13に示すように、出力フィードバックを記憶部に入力するのではなく、直接推論演算部13の中でフィードバックすることもある、これにより記憶部で消費するメモリ量を削減できる効果がある。

【0060】

なお、上記の例では、出力データが一つである例を示したが、出力データは複数のこともある。

【0061】

〔第3の実施の形態の動作〕

図14を参照して、第3の実施の形態にかかる推論処理装置1におけるデータフィルタ部11の動作について説明する。図14は、第3の実施の形態にかかる推論処理装置におけるデータフィルタ部の動作を示すフローチャートである。

【0062】

第1～2の実施の形態との違いは、複数のデータ発生元からの入力データを受信し、複数の入力データ間の類似性を判断し、その類似性に基づいて入力データから特定の入力データを抽出する点である。

【0063】

まず、データフィルタ部11は、複数の入力データ間の類似性を検出するのに用いる閾値を設定する（ステップS2-1）。閾値は、初期設定として動作開始時点で予め設定する以外に、動作中に閾値を動的に変更することもある。

【0064】

例えば、ある時刻において用いている閾値に対して、得られた推論処理結果に差がみられない場合、閾値を大きくすることもある。また、推論処理結果に対する推論精度が所望の精度よりも低い場合、閾値を小さくすることで、より多くの入力データに対して推論処理を行うことになるので、精度を改善できることが期待できる。このように、推論演算結果に応じて、入力データの類似性比較に用いる閾値を動的に設定することもある。

【0065】

次に、データフィルタ部11は、複数のデータ発生元からの入力データを取得して（ステップS2-2）、差分を計算する（S2-3）。計算した差分が閾値以上の場合（ステップS2-4：Yes）、複数の入力データに対する推論演算処理の出力結果は異なるため、複数の入力データに対して推論演算処理を行う（ステップS2-5）。

10

【0066】

一方、計算した差分が閾値より小さい場合（ステップS2-4：No）、比較した入力データのうちのいずれか一方の入力データに対してのみ推論演算処理を行う（ステップS2-6）。この場合、推論演算処理を行わなかった他方の入力データの推論結果は、比較した入力データに対して推論演算部13にて推論処理して得られた推論結果と同じものを用いる。

【0067】

このように、データフィルタ部11は、類似な入力データに対する後段の推論演算処理により得られる結果は同じであるため、複数の異なる入力データ発生元の入力データ同士の類似性を判断して、類似する入力データの全てに対して推論演算処理を行わないように構成することができる。これにより、推論演算処理を高速化し、推論演算処理に伴う消費電力を削減することが可能となる。

20

【0068】

〔第3の実施の形態の効果〕

このように、本実施の形態は、所定の学習データを用いて重みの値を学習したニューラルネットワークモデルを用いて、未知の入力データに対する推論処理を行うために、第1記憶部10は複数のデータ発生元からの入力データを記憶し、第2記憶部12は学習済みニューラルネットワークの重みを記憶し、データフィルタ部11は同時刻における複数の入力データ間の類似性を検出し、前記入力データのうち特定のデータのみを抽出して推論演算部13へ入力データとし、推論演算部13はデータフィルタ部11が抽出した入力データ、および前記学習済みニューラルネットワークの重みを入力として、前記学習済みニューラルネットワークの推論演算を実行し、前記入力データの特徴を推論する。

30

【0069】

これにより、データフィルタ部11において複数の異なる入力データ発生元の入力データ同士の類似性を検出して、推論演算処理の結果が同じになる類似な入力データに対する後段の推論演算処理を行わなくとも済む。本発明の推論処理装置1は、全ての入力データに対して推論処理を行う従来の推論処理装置と比較して、推論演算処理を高速化し、推論演算処理に伴う消費電力を削減することが可能となる。

【0070】

また、全ての入力データに対して推論処理をせずともよい場合、当該推論処理装置1からの推論結果の出力も削減できることから、通信ネットワークの負荷を低減できる。

40

【0071】

〔第4の実施の形態〕

図15～17を参照して、本発明の第4の実施の形態にかかる推論処理装置1について説明する。図15は、第4の実施の形態にかかる推論処理装置の構成を示すブロック図である。図16は、第4の実施の形態にかかる推論処理装置の構成を示すブロック図である。図17は、第4の実施の形態にかかる推論処理装置の構成を示すブロック図である。

【0072】

第1～3の実施の形態との違いは、記憶部の前段にデータフィルタ部11を備えて、入力データのうち特定のデータのみを抽出した後、第1記憶部10で格納する点と、複数の

50

データ発生元からの入力データを受信し、それらの入力データに対する推論演算処理を行う推論処理装置 1 であって、さらに、同時刻における複数の入力データ間の類似性を検出する点である。

【0073】

なお、上記の例では、入力データと重みデータを用いて推論演算を行う例を示したが、推論演算処理の方法はこれに限らない。例えば、図 16 に示すように、推論演算結果を次サイクルの推論演算処理の入力として用いる、即ち出力フィードバックを行うこともある。出力フィードバックを行うことで、文字列や音声・言語処理などの時系列データに適した推論演算を実施できる効果がある。また、図 17 に示すように、出力フィードバックを記憶部に入力するのではなく、直接推論演算部 13 の中でフィードバックすることもある、これにより記憶部で消費するメモリ量を削減できる効果がある。

【0074】

〔第 4 の実施の形態の効果〕

このように、本実施の形態は、所定の学習データを用いて重みの値を学習したニューラルネットワークモデルを用いて、未知の入力データに対する推論処理を行うために、第 1 記憶部 10 は複数のデータ発生元からの入力データを記憶し、第 2 記憶部 12 は学習済みニューラルネットワークの重みを記憶し、データフィルタ部 11 は同時刻における複数の入力データ間の類似性を検出し、前記入力データのうち特定のデータのみを抽出して推論演算部 13 へ入力データとし、推論演算部 13 はデータフィルタ部 11 が抽出した入力データ、および前記学習済みニューラルネットワークの重みを入力として、前記学習済みニューラルネットワークの推論演算を実行し、前記入力データの特徴を推論する。

【0075】

これにより、データフィルタ部 11 にて複数の異なる入力データ発生元の入力データ同士の類似性を検出して、推論演算処理の結果が同じになる類似な入力データに対する後段の推論演算処理を行わなくとも済む。本発明の推論処理装置 1 は、全ての入力データに対して推論処理を行う従来の推論処理装置と比較して、推論演算処理を高速化し、推論演算処理に伴う消費電力を削減することができる。

【0076】

また、第 1 記憶部 10 の前段にデータフィルタ部 11 を配置する構成とすることで、第 1 記憶部 10 が使用するメモリ量を削減できる効果がある。また、全ての入力データに対して推論処理をせずともよい場合、当該推論処理装置 1 からの推論結果の出力も削減できることから、通信ネットワークの負荷を低減できる。

【0077】

〔第 5 の実施の形態〕

図 18 を参照して、本発明の第 5 の実施の形態にかかる推論処理装置 1 について説明する。図 18 は、第 5 の実施の形態にかかる推論処理装置におけるデータフィルタ部の構成を示すブロック図である。

【0078】

第 1 ～ 4 の実施の形態との違いは、データフィルタ部 11 において、複数のデータ発生元からの入力データに対して、同時刻における複数の入力データ間の類似性と、入力データと直前に推論演算部 13 にて推論処理し推論結果を出力した際に用いた入力データとの類似性との両方を検出する点である。

【0079】

データフィルタ部 11 は、複数の入力データのうち特定のデータのみを抽出して推論演算部 13 へ入力する機能を有している。具体的には、図 18 に示すように、複数の入力データ同士を比較し、その差分が閾値以下である場合、比較した入力データのうちのいずれか一方の入力データを抽出する。さらに、当該入力データと直前に推論演算部 13 にて推論処理し推論結果を出力した際に用いた入力データとを比較し、その差分が閾値以上である場合、当該入力データを推論演算部 13 へ入力し、当該入力データに対してのみ推論演算処理を行う。

【0080】

この場合、推論演算処理を行わなかった入力データの推論結果は、比較した入力データに対して推論演算部13にて推論処理して得られた推論結果と同じものを用いる。一方、その差分が閾値より小さい場合、当該入力データを推論演算部13へ入力せずに、この場合の当該時刻における推論結果は、直前に推論演算部13にて推論処理して得られた推論結果を用いる。

【0081】

また、複数の入力データ同士を比較し、その差分が閾値より大きい場合、両者の推論演算処理の出力結果は異なるため、両者の入力データに対して推論演算処理を行う。このとき、同様に、当該入力データと直前に推論演算部13にて推論処理し推論結果を出力した際に用いた入力データとを比較し、その差分が閾値以上である場合、当該入力データを推論演算部13へ入力し、当該入力データに対してのみ推論演算処理を行う。この場合、推論演算処理を行わなかった入力データの推論結果は、比較した入力データに対して推論演算部13にて推論処理して得られた推論結果と同じものを用いる。

10

【0082】

このように、データフィルタ部11は複数の異なる入力データ発生元の入力データ同士の類似性を検出して、類似な入力データに対する後段の推論演算処理により得られる結果は同じであるため、推論演算処理を行わなくとも済む。これにより、推論演算処理を高速化し、推論演算処理に伴う消費電力を削減することができる。

【0083】

20

このように、データフィルタ部11は入力データの類似性を検出して、類似な入力データに対する後段の推論演算処理により得られる結果は変化しないため、推論演算処理を行わなくとも済む。これにより、推論演算処理を高速化し、推論演算処理に伴う消費電力を削減することができる。

【0084】

なお、比較する入力データは所定の組合せで行う。所定の組合せは、例えば、入力データの発生元の物理的な距離が最も近い入力データ同士を比較したり、入力データの発生元に付与されている識別子の順番に対応した入力データ同士を比較したりする。

【0085】

なお、入力データ同士を比較する回数は1段に限らず、異なる入力データ同士の組合せにて複数回比較を行うこともある。

30

【0086】

なお、推論演算部13は、複数の入力データに対する推論処理を並列に行うこともある。これにより、推論演算処理を高速化できる効果がある。

【0087】

なお、上記の例では、比較する入力データは所定の入力元からの入力データ同士を比較する例を示したが、必ずしも特定の入力データ同士を比較しなくともよい。すなわち、任意の入力データ同士を比較してもよい。例えば、入力データ発生元の端末等が時間に対して物理的に移動する移動端末である場合、その時刻において、物理的に距離が近い移動端末同士を組合せて、入力データを比較しても良い。

40

【0088】

なお、上記の例では、推論処理の高速化や低電力化のために入力データを削減する。このため、全ての入力データの組合せに対して網羅的に類似性を探索するのではなく、ある一部の入力データの組合せに対してのみ類似性を比較する例を示したが、必ずしも類似性の検出方法はこれに限らない。

【0089】

例えば、当該推論処理装置1に入力される入力データの発生元に対して全ての組合せで類似性を比較することが、後段の推論処理よりも高速に実行でき、かつ類似性の検出に要する電力が後段の推論処理よりも低い場合、網羅的に類似性を探索して、より高速かつ低電力に推論処理を実現することもある。

50

【0090】

なお、上記の例では、類似性の検出に用いる閾値を初期設定として与える例を示したが、閾値の設定方法はこれに限らない。例えば、ある時刻において用いている閾値に対して、得られた推論処理結果に差がみられない場合、閾値を大きくすることもある。また推論処理結果に対する推論精度が所望の精度よりも低い場合、閾値を小さくすることで、より多くの入力データに対して推論処理を行うことになるので、精度を改善できることが期待できる。このように、推論演算結果に応じて、入力データの類似性比較に用いる閾値を動的に設定することもある。

【0091】

なお、上記の例では、入力データと重みデータを用いて推論演算を行う例を示したが、推論演算処理の方法はこれに限らない。例えば、図16に示すように、推論演算結果を次サイクルの推論演算処理の入力として用いる、即ち出力フィードバックを行うこともある。出力フィードバックを行うことで、文字列や音声・言語処理などの時系列データに適した推論演算を実施できる効果がある。

10

【0092】

また、図17に示すように、出力フィードバックを記憶部に入力するのではなく、直接推論演算部13の中でフィードバックすることもある、これにより記憶部で消費するメモリ量を削減できる効果がある。

【0093】

なお、上記の例では、出力データが一つである例を示したが、出力データは複数のこともある。

20

【0094】

なお、前記入力データの比較では、入力データの要素ごとに第1閾値を使って比較し、差分が第1閾値以上の要素が第2閾値以上の場合に差分あり、差分が第1閾値以上の要素が第2閾値より少ない場合に差分なしと判断する。

【0095】

なお、上記の例では、入力データのみを用いて差分の比較を行う例を示したが、比較対象のデータは入力データに限らない。例えば、入力データと以前のサイクルにて推論演算部13が出力した推論結果をフィードバックとして用いる場合は、当該フィードバックデータ即ち出力データに対して比較を行うこともある。この場合、各々の比較結果の論理和または論理積をとって差分あり／なしを判断する。

30

【0096】

なお、上記の例では、入力データの差分を比較することで入力データの類似性を検出する例を示したが、差分の検出方法はこれに限らない。

【0097】

なお、上記の例では、入力データと重みデータを用いて推論演算を行う例を示したが、推論演算処理の方法はこれに限らない。例えば、図4に示すように、推論演算結果を次サイクルの推論演算処理の入力として用いる、即ち出力フィードバックを行うこともある。この場合、推論処理装置1はさらに、推論演算部13からフィードバックされる出力データを保持する第3記憶部14をさらに備える。出力フィードバックを行うことで、文字列や音声・言語処理などの時系列データに適した推論演算を実施できる効果がある。

40

【0098】

また、図5に示すように、出力フィードバックを第3記憶部14に入力するのではなく、直接推論演算部13の中でフィードバックすることもある、これにより推論処理装置1に搭載するメモリ容量を削減できる効果がある。

【0099】

なお、上記の例では、入力データと直前に推論演算部13にて推論処理し推論結果を出力した際に用いた入力データとを比較する例を示したが、比較対象のデータはこれに限らない。例えば、所定のサイクル数前以降に推論演算部13にて推論処理した推論結果および当該推論処理に用いた入力データと、を記憶しておき、入力データと所定のサイクル数

50

前の入力データとを比較する。

【0100】

〔第5の実施の形態の動作〕

図19を参照して、第5の実施の形態にかかる推論処理装置1におけるデータフィルタ部11の動作について説明する。図19は、第5の実施の形態にかかる推論処理装置におけるデータフィルタ部の動作を示すフローチャートである。

【0101】

第1～4の実施の形態との違いは、以前に推論演算した過去の入力データとの類似性の判断と、複数のデータ発生元からの入力データを受信し、同時刻における複数の入力データ間の類似性の判断の両方の判断結果に基づいて、入力データから特定の入力データを抽出する点である。上記の両方の類似性を判断することで、推論演算を行う入力データをさらに削減することが可能となる。

10

【0102】

まず、データフィルタ部11は、複数の入力データ間の類似性を検出するのに用いる閾値と、入力データと過去に推論演算処理した入力データとの差分を検出するのに用いる閾値とを設定する（ステップS3-1）。閾値は、初期設定として動作開始時点で予め設定する以外に、動作中に閾値を動的に変更することもある。

【0103】

例えば、ある時刻において用いている閾値に対して、得られた推論処理結果に差がみられない場合、閾値を大きくすることもある。また推論処理結果に対する推論精度が所望の精度よりも低い場合、閾値を小さくすることで、より多くの入力データに対して推論処理を行うことになるので、精度を改善できることが期待できる。このように、推論演算結果に応じて、入力データの類似性比較に用いる閾値を動的に設定することもある。

20

【0104】

次に、データフィルタ部11は、複数のデータ発生元からの入力データを取得して（ステップS3-2）、差分を計算する（ステップS3-3）。計算した差分が閾値以上の場合（ステップS3-4：Yes）、さらに、以前に推論演算処理した過去の入力データとの差分を計算して（ステップS3-5）、推論演算部へ出力するかを判断する（ステップS3-7）。

【0105】

一方、計算した差分が閾値より小さい場合（ステップS3-4：No）、比較した入力データのうちいずれか一方の入力データに対してのみ以前に推論演算処理した過去の入力データとの差分を計算して（ステップS3-6）、推論演算部13へ出力するかを判断する（ステップS3-7）。

30

【0106】

推論演算処理した過去の入力データとの差分が閾値以上である場合（ステップS3-7：Yes）、当該入力データを推論演算部13へ出力して推論演算を行う（ステップS3-8）。一方、その差分が閾値より小さい場合（ステップS3-7：No）、当該入力データを推論演算部13へ入力せず、推論を行わない（ステップS3-9）。この場合の当該時刻における推論結果は、過去の入力データに対して推論処理して得られた推論結果を用いる。

40

【0107】

このように、データフィルタ部11は、複数の入力データの類似性を判断して、類似な入力データの全てに対して推論演算処理を行わないように構成し、さらに、以前に推論処理を行った過去の入力データとの類似性を判断して、以前に推論処理を行った過去の入力データと類似するデータに対しては、推論演算処理を行わないように構成されている。これにより、類似な入力データに対する後段の推論演算処理を行わなくとも済むので、推論演算処理を高速化し、推論演算処理に伴う消費電力を削減することができる。

【0108】

〔第5の実施の形態の効果〕

50

このように、本実施の形態は、所定の学習データを用いて重みの値を学習したニューラルネットワークモデルを用いて、未知の入力データに対する推論処理を行うために、第1記憶部10は、複数のデータ発生元からの入力データを記憶し、第2記憶部12は、学習済みニューラルネットワークの重みを記憶し、データフィルタ部11は、複数の入力データ間の類似性及び、以前に推論処理を行った過去の入力データとの類似性を判断して、その類似性の判断結果に基づいて、入力データのうち特定のデータのみを抽出して推論演算部13への入力データとし、推論演算部13はデータフィルタ部11が抽出した入力データ、および前記学習済みニューラルネットワークの重みを入力として、前記学習済みニューラルネットワークの推論演算を実行し、入力データの特徴を推論する。

【0109】

これにより、データフィルタ部11にて複数の異なる入力データ発生元の入力データ同士との類似性及び、以前に推論処理を行った過去の入力データとの類似性を判断して、推論演算処理の結果が同じになる類似な入力データに対する後段の推論演算処理を行わなくとも済むように構成することができる。そのため、全ての入力データに対して推論処理を行う従来の推論処理装置と比較して、本発明の推論処理装置1は、推論演算処理を高速化し、推論演算処理に伴う消費電力を削減することができる。

【0110】

また、第1記憶部10の前段にデータフィルタ部11を配置する構成とすることで、第1記憶部10が使用するメモリ量を削減できる効果がある。また、全ての入力データに対して推論処理をせずともよいため、当該推論処理装置1からの推論結果の出力も削減できることから、通信ネットワークの負荷を低減できる。

【0111】

〔推論処理装置のハードウェア構成〕

次に、上述した構成を有する推論処理装置1のハードウェア構成の一例について図20を参照して説明する。

【0112】

図20に示すように、推論処理装置1は、例えば、バス101を介して接続されるプロセッサ102、主記憶装置103、通信インターフェース104、補助記憶装置105、入出力I/O106を備えるコンピュータと、これらのハードウェア資源を制御するプログラムによって実現することができる。推論処理装置1は、例えば、表示装置107がバス101を介して接続され、表示画面に推論結果などを表示してもよい。また、センサ108が入出力I/O106、バス101を介して接続され、推論処理装置1において推論の対象となる音声データなどの時系列データからなる入力データXを測定してもよい。

【0113】

主記憶装置103は、例えば、SRAM、DRAM、およびROMなどの半導体メモリによって実現される。主記憶装置103は、図1等で説明した記憶部を実現する。

【0114】

主記憶装置103には、プロセッサ102が各種制御や演算を行うためのプログラムが予め格納されている。プロセッサ102と主記憶装置103とによって、図1等に示した第1記憶部10、第2記憶部12、データフィルタ部11、推論演算部13を含む推論処理装置1の各機能が実現される。

【0115】

通信インターフェース104は、通信ネットワークNWを介して各種外部電子機器との通信を行うためのインターフェース回路である。推論処理装置1は、通信インターフェース104を介して外部から学習済みニューラルネットワークの重みデータWを受信したり、外部に推論結果Yを送出してもよい。

【0116】

通信インターフェース104としては、例えば、LTE、3G、無線LAN、Bluetooth（登録商標）などの無線データ通信規格に対応したインターフェースおよびアンテナが用いられる。通信ネットワークNWは、例えば、WAN（Wide Area N

10

20

30

40

50

etwork)やLAN(Local Area Network)、インターネット、専用回線、無線基地局、プロバイダなどを含む。

【0117】

補助記憶装置105は、読み書き可能な記憶媒体と、その記憶媒体に対してプログラムやデータなどの各種情報を読み書きするための駆動装置とで構成されている。補助記憶装置105には、記憶媒体としてハードディスクやフラッシュメモリなどの半導体メモリを使用することができる。

【0118】

補助記憶装置105は、推論処理装置1が推論を行うためのプログラムを格納するプログラム格納領域を有する。さらには、補助記憶装置105は、例えば、上述したデータやプログラムやなどをバックアップするためのバックアップ領域などを有していてもよい。補助記憶装置105は、例えば、推論処理プログラムを記憶することができる。

【0119】

入出力I/O106は、表示装置107など外部機器からの信号を入力したり、外部機器へ信号を出力したりするI/O端子により構成される。

【0120】

なお、推論処理装置1は、1つのコンピュータによって実現される場合だけでなく、互いに通信ネットワークNWで接続された複数のコンピュータによって分散されていてもよい。また、プロセッサ102は、FPGA(Field-Programmable Gate Array)、LSI(Large Scale Integration)、ASIC(Application Specific Integrated Circuit)等のハードウェアによって実現されていてもよい。

【0121】

特に、推論演算部13をFPGAなどの書き換え可能なゲートアレイを用いて構成することで、入力データXの構成や使用されるニューラルネットワークモデルに応じて柔軟に回路構成を書き換えることができる。この場合、様々なアプリケーションに対応することが可能な推論処理装置1を実現できる。

【0122】

【実施の形態の拡張】

以上、実施形態を参照して本発明を説明したが、本発明は上記実施形態に限定されるものではない。本発明の構成や詳細には、本発明の範囲内で当業者が理解しうる様々な変更をすることができる。また、各実施形態については、矛盾しない範囲で任意に組み合わせ実施することができる。

【符号の説明】

【0123】

1…推論処理装置、10…第1記憶部、11…データフィルタ部、12…第2記憶部、13…推論演算部13、14…第3記憶部、101…バス、102…プロセッサ、103…主記憶装置、104…通信インターフェース、105…補助記憶装置、106…入出力I/O、107…表示装置、108…センサ。

10

20

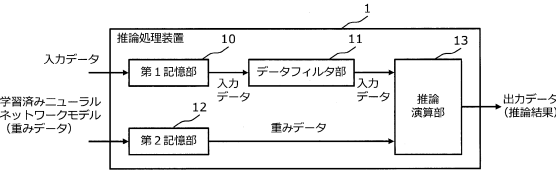
30

40

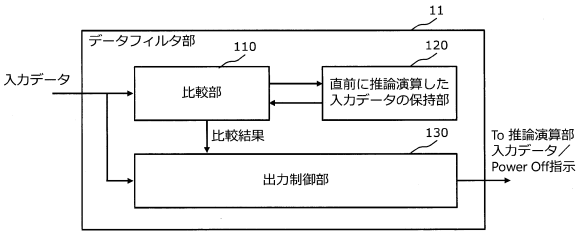
50

【図面】

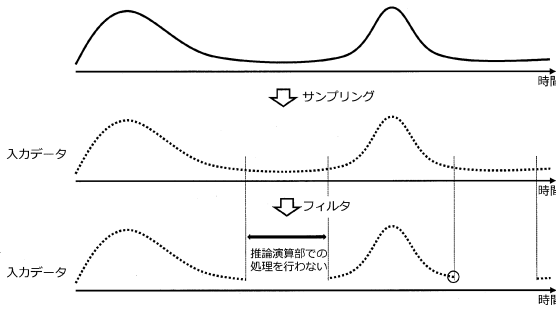
【図 1】



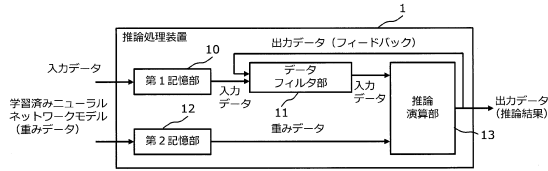
【図 2 A】



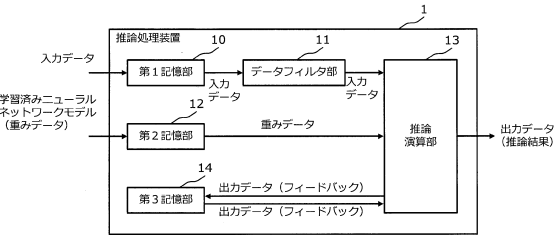
【図 2 B】



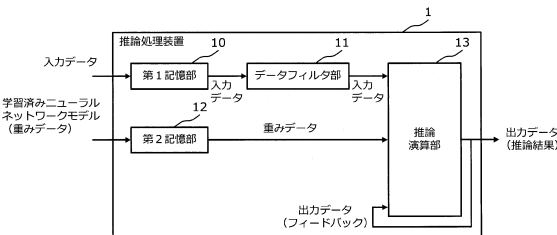
【図 3】



【図 4】



【図 5】



10

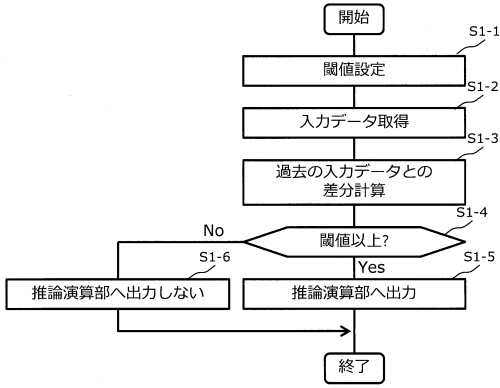
20

30

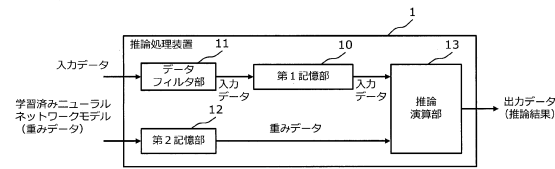
40

50

【図 6】

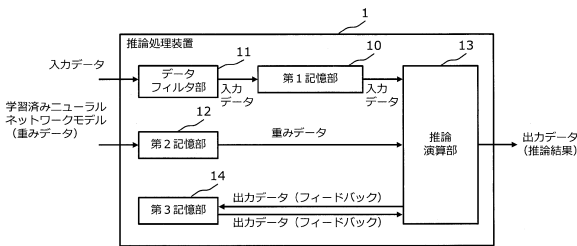


【図 7】

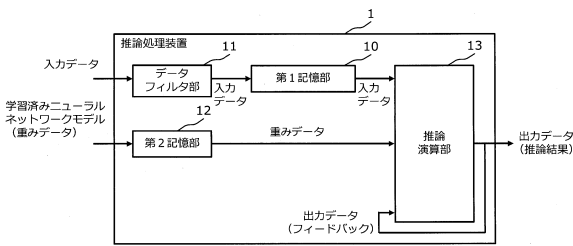


10

【図 8】

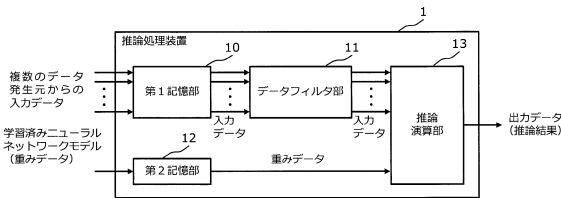


【図 9】

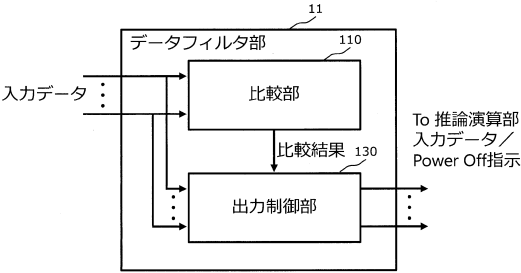


20

【図 10】



【図 11】

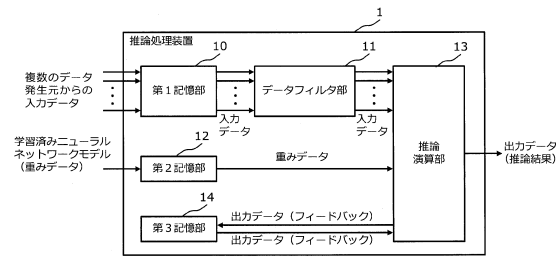


30

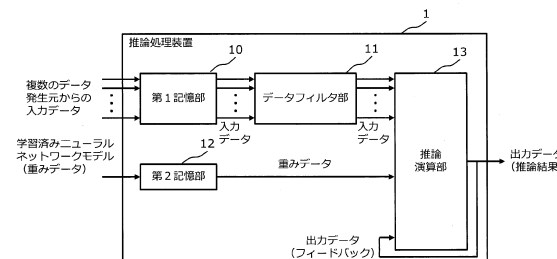
40

50

【図 1 2】

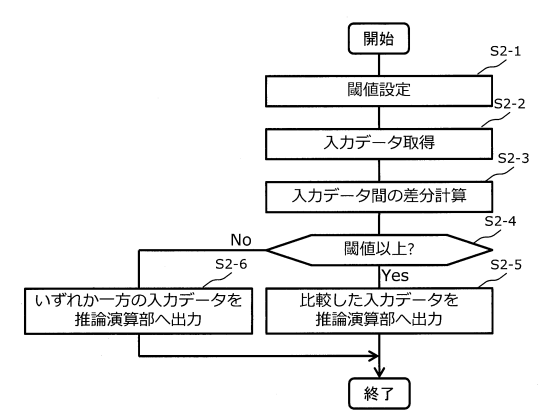


【図 1 3】

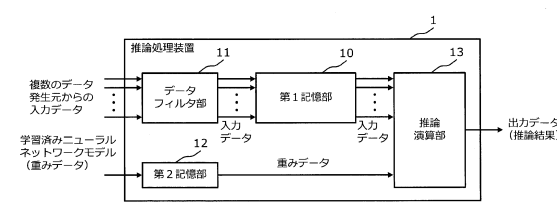


10

【図 1 4】

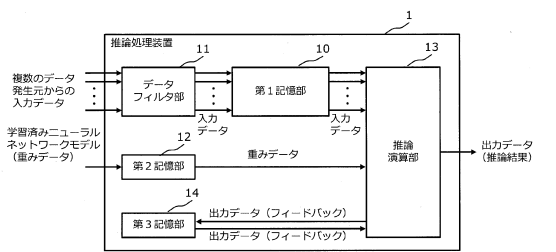


【図 1 5】

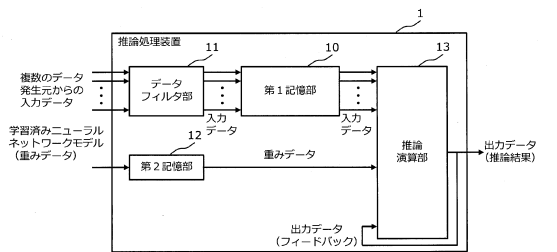


20

【図 1 6】



【図 1 7】

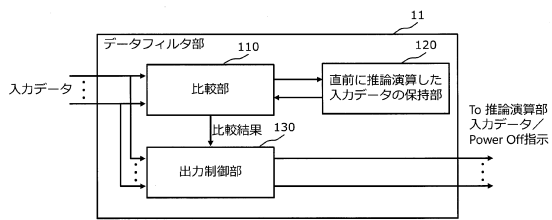


30

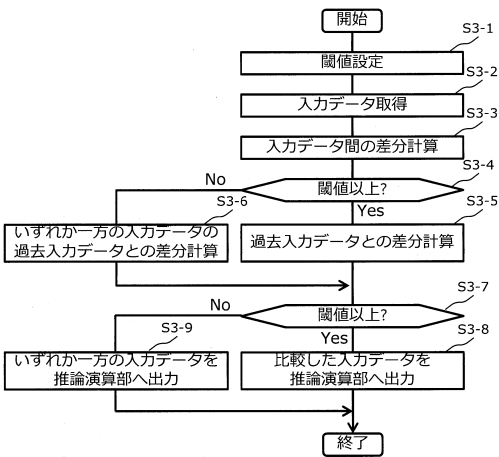
40

50

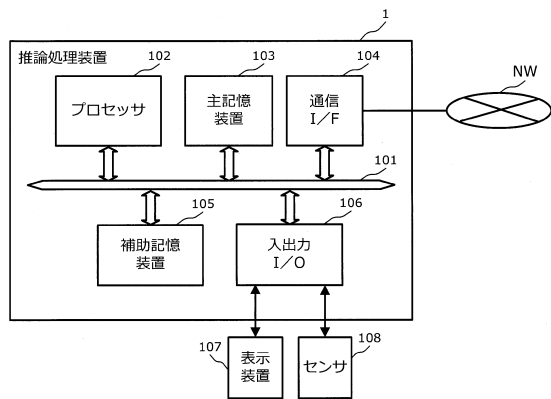
【図 18】



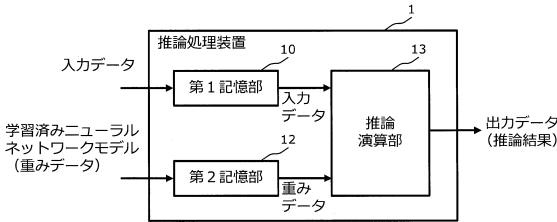
【図 19】



【図 20】



【図 21】



10

20

30

40

50

フロントページの続き

東京都千代田区大手町一丁目5番1号 日本電信電話株式会社内

審査官 藤原 敬利

(56)参考文献 特開2000-089958 (JP, A)

国際公開第2018/179361 (WO, A1)

(58)調査した分野 (Int.Cl., DB名)

G06F 1/00-18/40

G06N 3/00-99/00