(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7493380号
(P7493380)

(45)発行日　令和6年5月31日(2024.5.31)

(24)登録日　令和6年5月23日(2024.5.23)

(51)国際特許分類		FI	
G06N	*3/0495(2023.01)*	G06N	3/0495
G06N	*3/0464(2023.01)*	G06N	3/0464

請求項の数　12　外国語出願　（全14頁）

(21)出願番号　特願2020-81985(P2020-81985)
(22)出願日　令和2年5月7日(2020.5.7)
(65)公開番号　特開2020-184341(P2020-184341A)
(43)公開日　令和2年11月12日(2020.11.12)
審査請求日　令和5年2月6日(2023.2.6)
(31)優先権主張番号　10 2019 206 621.6
(32)優先日　令和1年5月8日(2019.5.8)
(33)優先権主張国・地域又は機関
ドイツ(DE)
早期審査対象出願

(73)特許権者　390023711
ローベルト ボツシユ ゲゼルシヤフト ミツト ベシユレンクテル ハフツング
ＲＯＢＥＲＴ ＢＯＳＣＨ ＧＭＢＨ
ドイツ連邦共和国 シユツツトガルト （番地なし）
Ｓｔｕｔｔｇａｒｔ，Ｇｅｒｍａｎｙ
(74)代理人　100114890
弁理士　アインゼル・フェリックス＝ラインハルト
(74)代理人　100098501
弁理士　森田 拓
(74)代理人　100116403
弁理士　前川 純一
(74)代理人　100135633

最終頁に続く

(54)【発明の名称】　機械学習システム、並びに、機械学習システムを構成する方法、コンピュータプログラム及び装置

(57)【特許請求の範囲】

【請求項１】

量子化されたフィルタ係数を有する少なくとも１つのフィルタ（２１）を含む、コンピュータ実装されて学習する機械学習システム（１０）であって、

前記機械学習システム（１０）は、畳み込み層を備えたディープニューラルネットワークであり、

前記機械学習システム（１０）は、前記フィルタ（２１）を使用して入力変数（１２）に依存して出力変数を算出するように構成されており、

前記入力変数（１２）は、画像の時間的シーケンスを含み、

前記機械学習システム（１０）は、すべてが前記画像のそれぞれ同一座標上に存在する又は前記機械学習システム（１０）の中間結果のそれぞれ同一座標上に存在する画素の時間的シーケンスに、前記フィルタ（２１）を適用するように構成されている、機械学習システム（１０）において、

前記フィルタ（２１）の前記フィルタ係数及び前記機械学習システム（１０）の前記中間結果は、最大で８ビットで量子化されており、前記機械学習システム（１０）は、前記フィルタ（２１）の量子化された前記フィルタ係数を使用して学習するように構成されている、

機械学習システム（１０）。

【請求項２】

前記フィルタ（２１）は、多次元であり、前記機械学習システム（１０）は、前記フィ

ルタ（２１）を、すべてが前記画像のそれぞれ同一座標上に存在する又は前記中間結果のそれぞれ同一座標上に存在するセクションのシーケンスに適用するように構成されている、請求項１に記載の機械学習システム（１０）。

【請求項３】

前記フィルタ係数及び／又は前記中間結果は、２進数又は３進数である、

請求項１に記載の機械学習システム（１０）。

【請求項４】

前記フィルタ（２１）は、前記フィルタ（２１）の使用の際に、前記画像又は中間結果の個々の画素を省略するように構成されている、

請求項１に記載の機械学習システム（１０）。

【請求項５】

前記機械学習システム（１０）は、前記フィルタ（２１）の直後に配置されているプーリング演算（pooling）を含む、

請求項１に記載の機械学習システム（１０）。

【請求項６】

前記画像は、予め定められた順序で相前後して配置されており、前記フィルタ（２１）は、それぞれが同一座標上に存在する画素又はセクションのシーケンスに適用され、前記シーケンスの画素又はセクションは、前記順序で配置されている、

請求項１に記載の機械学習システム（１０）。

【請求項７】

前記機械学習システム（１０）は、車両に組み込まれており、前記車両は、制御ユニットを含み、前記制御ユニットは、前記機械学習システム（１０）からの算出された前記出力変数を使用して前記車両のアクチュエータを制御するように構成されている、

請求項１に記載の機械学習システム（１０）。

【請求項８】

コンピュータ実装されて学習する機械学習システム（１０）を動作させる方法であって、

フィルタ係数を有する少なくとも１つのフィルタ（２１）を含む機械学習システム（１０）を準備するステップであって、前記機械学習システム（１０）は、畳み込み層を備えたディープニューラルネットワークであり、前記機械学習システム（１０）は、前記フィルタ（２１）を使用して入力変数（１２）に依存して出力変数を算出するように構成されており、前記入力変数（１２）は、画像の時間的シーケンスを含み、前記機械学習システム（１０）は、すべてが前記画像のそれぞれ同一座標上に存在する又は前記機械学習システム（１０）の中間結果のそれぞれ同一座標上に存在する画素の時間的シーケンスに、前記フィルタ（２１）を時間軸に沿って適用するように構成されている、機械学習システム（１０）を準備するステップと、

前記画像のシーケンスを、前記機械学習システム（１０）の前記入力変数（１２）に対してグループ化するステップと、

前記機械学習システム（１０）によって、前記フィルタ（２１）を前記時間軸に沿って使用して前記出力変数を前記入力変数（１２）に依存して算出するステップと、

前記フィルタ（２１）の前記フィルタ係数及び前記機械学習システム（１０）の前記中間結果を、最大で８ビットで量子化するステップと、

前記フィルタ（２１）の量子化された前記フィルタ係数を使用して、前記機械学習システム（１０）を学習させるステップと、

を含む方法。

【請求項９】

算出された前記出力変数を使用して車両のアクチュエータを制御するステップをさらに含む、

請求項８に記載の方法。

【請求項１０】

前記画像は、時間的に順次連続して検出され、次いで、前記出力変数が算出された場合

、新たに検出された画像が前記入力変数（１２）に追加され、前記入力変数（１２）に含まれている画像であって、時間的に最も長く含まれていた画像が前記入力変数（１２）から削除され、前記機械学習システム（１０）によってさらなる出力変数が、前記フィルタ（２１）を使用して前記入力変数に依存して算出される、
請求項８に記載の方法。
【請求項１１】
コンピュータ実装されて学習する機械学習システム（１０）をトレーニングする方法であって、前記機械学習システム（１０）は、フィルタ係数を有する少なくとも１つのフィルタ（２１）を含み、前記機械学習システム（１０）は、畳み込み層を備えたディープニューラルネットワークであり、前記機械学習システム（１０）は、前記フィルタ（２１）を使用して入力変数（１２）に依存して出力変数を算出するように構成されており、前記入力変数（１２）は、画像の時間的シーケンスを含み、前記機械学習システム（１０）は、すべてが前記画像のそれぞれ同一座標上に存在する又は前記機械学習システム（１０）の中間結果のそれぞれ同一座標上に存在する画素の時間的シーケンスに、前記フィルタ（２１）を適用するように構成されている、方法において、
少なくとも複数の画像のシーケンスと割り当てられたトレーニング出力変数とを含むトレーニングデータを提供するステップと、
それぞれ少なくとも１つの出力変数を、前記画像のシーケンスの各々に対して算出するステップと、
損失関数を、算出された前記出力変数と前記トレーニング出力変数とに依存して算出するステップと、
前記フィルタ（２１）の前記フィルタ係数を含む、前記機械学習システム（１０）のパラメータを、前記損失関数に依存して適応化させ、それによって、前記損失関数が予め設定可能な基準に関して最適にされるステップと、
前記適応化の後に、前記機械学習システム（１０）の前記フィルタ（２１）の前記フィルタ係数及び前記機械学習システム（１０）の前記中間結果を、最大で８ビットで量子化するステップと、
前記フィルタ（２１）の量子化された前記フィルタ係数を使用して、前記機械学習システム（１０）を学習させるステップと、
を含む方法。
【請求項１２】
コンピュータにより実行されるときに、当該コンピュータに以下のステップ、即ち、
フィルタ係数を有する少なくとも１つのフィルタ（２１）を含む、コンピュータ実装されて学習する機械学習システム（１０）を準備するステップであって、前記機械学習システム（１０）は、畳み込み層を備えたディープニューラルネットワークであり、前記機械学習システム（１０）は、前記フィルタ（２１）を使用して入力変数（１２）に依存して出力変数を算出するように構成されており、前記入力変数（１２）は、画像の時間的シーケンスを含み、前記機械学習システム（１０）は、すべてが前記画像のそれぞれ同一座標上に存在する又は前記機械学習システム（１０）の中間結果のそれぞれ同一座標上に存在する画素の時間的シーケンスに、前記フィルタ（２１）を時間軸に沿って適用するように構成されている、機械学習システム（１０）を準備するステップと、
前記画像のシーケンスを、前記機械学習システム（１０）の前記入力変数（１２）に対してグループ化するステップと、
前記機械学習システム（１０）によって、前記フィルタ（２１）を前記時間軸に沿って使用して前記出力変数を前記入力変数（１２）に依存して算出するステップと、
前記フィルタ（２１）の前記フィルタ係数及び前記機械学習システム（１０）の前記中間結果を、最大で８ビットで量子化するステップと、
前記フィルタ（２１）の量子化された前記フィルタ係数を使用して、前記機械学習システム（１０）を学習させるステップと、
を実施させるためのコンピュータプログラムであって、前記機械学習システム（１０）を

動作させるためのコンピュータプログラムが記憶されている機械可読メモリ要素。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、量子化されたパラメータを有する機械学習システムに関する。さらに、本発明は、機械学習システムを構成する方法及びコンピュータプログラム並びに装置に関する。
【背景技術】
【０００２】
背景技術
未公開の独国特許出願第１０２０１７２１８８８９．８号明細書は、１つ以上の入力変数を、内部処理チェーンによって１つ以上の出力変数に処理するように構成されている人工知能モジュールを開示している。この内部処理チェーンは、１つ以上のパラメータによって確定されている。ここでは、少なくとも１つの統計的分布に依存してパラメータを算出するように構成されている分布モジュールが想定されている。
【０００３】
未公開の独国特許出願第１０２０１８２１６４７１．１号明細書は、量子化されたニューラルネットワークを開示している。
【０００４】
量子化ニューラルネットワークを、大幅に強化した量子化によって構成することは、例えば、Ｈｕｂａｒａらによる文献「Quantized neural networks:Training neural networks with low precision weights and activations: The Journal of Machine Learning Research, 2017, 18. Jg., Nr.1, S.6869-6898」に示されているように可能である。
【先行技術文献】
【特許文献】
【０００５】
【文献】独国特許出願第１０２０１７２１８８８９．８号明細書
【文献】独国特許出願第１０２０１８２１６４７１．１号明細書
【非特許文献】
【０００６】
【文献】Ｈｕｂａｒａら著、「Quantized neural networks:Training neural networks with low precision weights and activations: The Journal of Machine Learning Research, 2017, 18. Jg., Nr.1, S.6869-6898」
【発明の概要】
【発明が解決しようとする課題】
【０００７】
リカレントニューラルネットワーク（recurrent neural networks；略してＲＮＮ）は、センサデータの時間的次元に含まれた情報も処理することができるという利点を有する。ただし、ＲＮＮの欠点は、量子化を強化することができないことにある。
【０００８】
しかしながら、ＲＮＮを、モバイルコンピュータユニット上でより効率的に動作させることができるようにするために、ＲＮＮのパラメータを量子化することが望ましい。
【課題を解決するための手段】
【０００９】
独立請求項１によれば、ＲＮＮの代わりに、時間的畳み込みによるニューラルネットワークの使用が提案される。これには、次のような利点がある。即ち、これらのニューラルネットワークが、複雑な計算の畳み込みにもかかわらず自身のパラメータの量子化の際に、上記のＲＮＮの利点を維持しつつも同等のＲＮＮよりも驚くほどエネルギーの節約ができるようになるということである。
【００１０】

発明の開示

第１の態様においては、逐次処理／時系列処理のための特にメモリ効率及びエネルギー効率に優れた機械学習システムが提案される。

【００１１】

請求項１に記載の特徴を有する機械学習システム、特にニューラルネットワークは、少なくとも１つのフィルタを含み、ここでは、少なくともフィルタのフィルタ係数が量子化されている。フィルタ係数の他に、機械学習システムのさらなるパラメータが量子化されていることも考えられる。代替的又は付加的に、少なくとも活性化が量子化されてもよいであろう。この活性化の量子化は、活性化の数値が機械学習システムのパラメータ（例えば、重み）の数値を上回った場合に有意になり得る。

【００１２】

この機械学習システムは、フィルタを使用して入力変数に依存して出力変数を算出するように構成されている。入力変数は、画像の時間的シーケンスを含む。代替的に、時間的に順次連続して検出された複数の変数、例えば一次元センサデータなどを使用することができる。機械学習システムは、さらに、フィルタを、すべてが画像のそれぞれ同一座標上に存在する、又は、機械学習システムの中間結果の、特に中間処理された画像のそれぞれ同一座標上に存在する画素のシーケンスに適用するように構成されている。画像の代わりに検出された変数が使用されるケースについては、フィルタは、それぞれ同一の位置に存在する、検出された変数又は中間結果のデータ点に適用される。これらの中間結果は、例えば、ニューラルネットワークの活性化層である。

【００１３】

また、フィルタ係数及び好適には機械学習システムのさらなるパラメータが、それぞれ、予め定められ量子化された値範囲からの値を有することも考えられる。この予め定められ量子化された値範囲は、量子化レベルの予め定められた数によって、及び／又は、ビットの予め設定された数によって定められてもよい。

【００１４】

量子化とは、フィルタ係数が１０進表記の量子化値であること、好ましくは、浮動小数点、固定小数点又は整数の記数形式の量子化値のみであることを想定することができることを意味するものと理解されたい。好適には、フィルタ係数は、最大で３２ビット、特に最大で１６ビット又は８ビット、好適には４ビット又は１ビットで量子化されている。さらに、量子化とは、フィルタ係数が、予め定められた複数の量子化レベルのすぐ次の量子化レベルに切り上げられたこと又は切り捨てられたことを意味するものと理解されたい。したがって、フィルタ係数は、それぞれ複数の量子化レベルの値を有し、この場合、これらの値はメモリに格納されている。これについて、好適には、量子化レベルは、メモリ効率に優れた手法で格納することができる。量子化レベルの数は、量子化に使用されたビットの数に依存している。代替的に、そのパラメータを学習の際に完全な分解能で適応化させ、これに続いて量子化させるように学習させた機械学習システムで量子化することも可能である。

【００１５】

予め設定可能に量子化される値範囲は、リストであってもよい。フィルタ係数が低減された分解能を有することも考えられ、その場合、分解能は、量子化を用いて低減されている。

【００１６】

解像度は、それぞれがパラメータ及び／又は中間結果を表す、種々異なって可能な複数の特に量子化された値を用いて特徴付けることが可能である。

【００１７】

フィルタの適用とは、すべてが同一座標を有するデータ点又はセクションを用いて数学的に畳み込みされることを意味するものと理解されたい。

【００１８】

第１の態様による機械学習システムは、量子化によって、フィルタ係数の記憶のために

より少ないビットしか必要なく、結果として、より少ないメモリスペースしか必要なくなるという利点を有する。これにより、ハードウェアにおける機械学習システムのよりコンパクトな構造形式を可能にする、より小さなメモリを使用することができるようになる。パラメータ及び活性化の記憶と読み取りのためのエネルギー消費量は、ビットの数に比例し、乗算についてはパラメータ及び活性化の分解能（ビットの数）に対して少なくとも二次的であるため、特に量子化によるビットの数の低減は、特にコンピュータのリソース効率に優れる。機械学習システムがソフトウェアで実現されている場合であっても、より少ないメモリしか必要ないため、これらの利点が活用される。

【００１９】

例えば、画像などの高分解能の入力変数の他に、フィルタは、エネルギー消費量に対して最大の影響を有しているので、フィルタの量子化は有利である。というのも、その結果として生じるコンピュータリソースに優しい取り扱いにより、この機械学習システムは、据え置き型の適用（例えば、中央コンピュータなど）にも移動型の適用（例えば、モバイルデータ処理装置又は組み込みシステムなど）にも使用可能である。この種の量子化された機械学習システムは、特に、センサ近傍又はセンサ自体において測定されたデータを処理するのに適している。これにより、測定されたデータを帯域幅効率に優れた手法により圧縮及び伝送することができる。

【００２０】

画像は、Ｌｉｄａｒ画像、レーダ画像、超音波画像、又は、カメラ画像であるものとしてもよい。

【００２１】

入力変数は、複数の検出された変数又は画像のシーケンスを含むテンソルであることが提案される。検出された変数又は画像は、テンソルの一次元に沿って配置されている。次いで、フィルタは、少なくとも一次元に沿って適用される。それゆえ、時間的シーケンスを特徴付ける一次元に沿ったフィルタの適用は、時間的畳み込みとも称される。

【００２２】

この機械学習システムの利点は、時間軸に沿って含まれている情報も、時間的フィルタリングによってともに処理されることにある。これにより、この機械学習システムは、回帰特性を有するが、ただし、典型的には量子化を強化することができない回帰型結合は不要である。それにもかかわらず、量子化されたフィルタにより、ＲＮＮを用いた場合よりも効率に優れた入力変数のシーケンスの処理が達成される。

【００２３】

さらに、フィルタは、多次元であり、機械学習システムはさらに、フィルタを、すべてが画像のそれぞれ同一座標上に存在する、又は、中間結果のそれぞれ同一座標上に存在するセクションのシーケンスに適用するように構成されていることが提案される。

【００２４】

１つのセクションは、検出された変数又は画像又は中間結果のそれぞれ座標上に存在する複数のデータ点又は画素を含む。これらのセクションは、一次元／多次元であってもよく、それに応じてフィルタの次元の数を選択する必要がある。それにより、すべてが一次元に沿って、特にそれに平行に存在する複数の画素又はデータ点は、一次元に沿ったフィルタリングの際に付加的に考慮される点に留意されたい。

【００２５】

例えば、セクションは、線又は矩形であってもよい。画像の内部においては、セクションの様々な配向、例えば、垂直線又は水平線又は対角線などが考えられる。

【００２６】

また、機械学習システムは、さらに付加的に、フィルタを、それぞれが画像の内部の異なる座標上に存在する又はそれぞれが中間結果の内部の異なる座標上に存在するさらなるセクションのシーケンスに適用するように構成されていることが提案される。

【００２７】

ここでの利点は、フィルタの付加的な適用を用いることにより、付加的情報がそれぞれ

検出された変数に沿って出力変数に導入される点にある。即ち、このフィルタは、テンソルのさらなる次元に適用される。このことは、時間的畳み込み及び空間的畳み込みとも称される。

【００２８】

さらに、フィルタ係数の他に、機械学習システムの中間結果、特に活性化も量子化されていることが提案される。

【００２９】

さらに、中間結果が、予め定められ量子化された値範囲又はさらなる予め定められ量子化された値範囲からのそれぞれ１つの値を有することも考えられる。

【００３０】

さらに、フィルタ係数及び／又は中間結果は、２進数又は３進数であることも提案される。この提案のもとでは、それぞれのフィルタ係数及び／又は中間結果は、最大で２つ又は３つの異なる値をとり得ることが考えられる。好適には、これらの異なる値は、最大で１つ又は２つのビットを用いて表される。理想的には１つのビットによってである。なぜなら、そうすることによって、特に計算効率及びメモリ効率に優れたバイナリニューラルネットワークが存在するからである。

【００３１】

さらに、フィルタは、個々のデータ点を、フィルタリングの際に、特に時間的次元に沿って省略することが提案される。このことは、より少ないフィルタ係数を用いて、より大きなフィルタの視野を達成することができ、ひいてはさらに過去に遡ってみることができるという利点を有する。

【００３２】

さらに、機械学習システムは、フィルタの直後に配置されているプーリング演算（pooling）を含むことが提案される。

【００３３】

このプーリング演算とは、個々のフィルタリングされたデータ点又はフィルタリングされた画素が無視されること、特に拒否されることを意味するものと理解されたい。好適には、クラスターの内部のデータ点が無視され、それらの値は、ゼロ近傍に存在する局限的プーリング（local pooling）が使用される。代替的に、平均的プーリング（average pooling）も使用することができる。

【００３４】

さらに、画像又は検出された変数は、予め設定可能な順序で相前後して、特に画像又は検出された変数のそれぞれ検出された時点に依存して配置されていることが提案される。好適には、この配置は、検出の時点の後に上昇／下降するように行われる。

【００３５】

機械学習システムは、専らハードウェアにおいても、ソフトウェアにおいても、又は、ソフトウェア及びハードウェアからなる混合形態においても、実現することができる点に留意されたい。これにより、それぞれパラメータを格納するためのわずかなメモリを準備すればよく、そのため、本発明の第１の態様による機械学習システムは、最小の技術システムにも組み込んで使用することが可能である。

【００３６】

本発明の第２の態様においては、本発明の第１の態様の機械学習システムを動作させる、特にコンピュータで実現される方法が提案される。この方法は、複数の検出された変数を、機械学習システムの入力変数に対してグループ化するステップと、続いて、フィルタを使用して出力変数を入力変数に依存して算出するステップと、を含む。ここでは、機械学習システムのフィルタが使用され、特に、入力変数又は中間結果の時間的フィルタリングが実施される。

【００３７】

本発明の第２の態様については、画像は、それぞれ直接後続する時点で相前後して検出されることが提案される。次いで、出力変数が算出された場合、新たに検出された画像が

入力変数に追加され、含まれている画像のうちの１つ、特に時間的に最も長く含まれていた画像が入力変数から削除される。

【００３８】

好適には、入力変数の画像は、テンソルの一次元に沿った位置だけシフトされ、ここで、新たに検出された画像は空きになった箇所に追加され、入力変数の最も過去に検出された画像はテンソルから削除される。

【００３９】

さらに、変数、特に画像は、それぞれの変数が検出された時点のシーケンスのそれぞれ１つの時点に割り当てられており、ここで、各時点において新たな変数が検出され、この新たに検出された変数は、直接先行する時点において検出された変数の他に、当該入力変数に追加され、最も過去の時点に割り当てられている検出された変数が当該入力変数から削除される。

【００４０】

本発明のさらなる態様においては、第１の態様の機械学習システムを学習させる方法が提案されている。この学習は、少なくともトレーニングデータを提供するステップを含む。これらのトレーニングデータは、少なくとも複数の画像シーケンス又は検出された変数の時系列を含む。これに続いて、それぞれ少なくとも１つの出力変数を、画像シーケンス又は時系列の各々に対して算出するステップが追従する。これに続いて、損失関数を、算出された出力変数及びトレーニングデータに依存して算出するステップが追従する。引き続き、機械学習システムのパラメータを、損失関数に依存して適応化させ、それによって、損失関数が予め設定可能な基準に関して最適にされるステップが追従する。これに続いて、パラメータ、特にフィルタ係数を量子化するステップが追従する。

【００４１】

好適には、量子化は最初から使用され、これを考慮に入れて、機械学習システムは、学習する。

【００４２】

さらなる態様においては、第１の態様の機械学習システム又は第２の態様に従って生成された機械学習システムは、検出されたセンサ変数、特にセンサの画像に依存して、出力変数を算出することができ、当該出力変数は、これに続いて制御ユニットを用いた制御変数の算出に用いられる。

【００４３】

この制御変数は、技術システムのアクチュエータの制御に使用することができる。この技術システムは、例えば、少なくとも半自律型の機械、少なくとも半自律型の車両、ロボット、ツール、工作機械、又は、ドローンなどの自律的な飛行物体であってもよい。入力変数は、例えば、検出されたセンサデータに依存して算出することができ、機械学習システムに提供することができる。これらのセンサデータは、技術システムのセンサ、例えば、カメラなどによって検出することができ、代替的に外部から受信することができる。

【００４４】

さらなる態様においては、コンピュータプログラムが提案される。このコンピュータプログラムは、本発明の第２の態様の前述した方法の１つを実施するように構成されている。このコンピュータプログラムは、当該コンピュータプログラムがコンピュータ上で実行されるときに、コンピュータにこれらの前記方法の１つをそのすべてのステップについて実施させるための命令を含む。さらに、コンピュータプログラムが記憶されている機械可読メモリモジュールが提案される。

【００４５】

上述の態様の実施例は、添付の図面に示されており、以下に続く明細書において、より詳細に説明される。

【図面の簡単な説明】

【００４６】

【図１】例えばそれぞれ時系列を処理するための２つの異なるニューラルネットワークに

対する量子化分解能に関してプロットされたエネルギー消費量の概略図。
【図２】時間的畳み込み用のフィルタを含むディープニューラルネットワークの概略図。
【図３】時間的畳み込みの概略図。
【図４】時間的畳み込み用のフィルタを含むニューラルネットワークを動作させる方法のフローチャートの概略図。
【図５】ロボット用の時間的畳み込み用のフィルタを含むニューラルネットワークの使用の概略図。
【発明を実施するための形態】
【００４７】
図１は、時間的フィルタリング（ＮＮ－ｔｅｍｐ）を用いてニューラルネットワークの測定されたエネルギー消費量Ｅ及びＲＮＮの測定されたエネルギー消費量Ｅが、量子化分解能Ｑに関してプロットされている線図の概略図を示している。ここでは、２つのネットワークが同一のデータセット上において類似の結果を得るように留意した。ｙ軸は、エネルギー消費量Ｅを表し、このエネルギー消費量Ｅは、線図の原点において最も低く、原点からの距離の増加に伴って増加する。ｘ軸には、線図の原点の方向において増加する量子化分解能Ｑがプロットされている。即ち、原点に近いほど、量子化分解能、特にパラメータを表すビットの数が高くなる。以下では、高い量子化分解能用の時間的フィルタリング（ＮＮ－ｔｅｍｐ）を備えたニューラルネットワークが、同等の量子化分解能のもとでの同等のＲＮＮよりも大幅に増加したエネルギー消費量を有している例示的なケースを検討する。
【００４８】
２つのニューラルネットワークのパラメータの分解能が低減すると、２つのニューラルネットワークのエネルギー消費量もそれぞれ減少する。２つのニューラルネットワークの結果の品質が実質的に不変のまま維持されるという条件のもとで、時間的フィルタリング（ＮＮ－ｔｅｍｐ）を用いたニューラルネットワークの最小の量子化分解能を、ＲＮＮの最小の量子化分解能と比較すると、時間的フィルタリング（ＮＮ－ｔｅｍｐ）を用いたニューラルネットワークの最小の量子化分解能は、ＲＮＮの最小の量子化分解能よりも大幅に低減していることが顕著である。このことは、時間的フィルタリング（ＮＮ－ｔｅｍｐ）を用いたニューラルネットワークが、強化された量子化に基づいて、ＲＮＮよりも少ないエネルギーを使用することに結び付く。
【００４９】
図２は、機械学習システム１０、特にディープニューラルネットワークの概略図を示している。この機械学習システム１０は、予め設定可能な順序で互いに接続された複数の層１１を含む。この機械学習システム１０は、入力変数１２に依存して出力変数ｙを算出する。
【００５０】
機械学習システム１０は、出力変数ｙとして、入力変数１２の画像の１つ又は最も新しく検出された画像の１つの分類、セグメント化を出力することができる。代替的に、出力変数ｙは、回帰型であってもよい。
【００５１】
入力変数１２は、この実施例においては、テンソルである。このテンソルは、相前後して検出された複数の画像１３を含む。これらの画像は、ここでは、テンソルの一次元に沿って配置されている。
【００５２】
これらの層は、完全に接続された（fully connected）層であるものとしてもよく、ここで、これらの層の少なくとも１つは、時間的フィルタリングを実施するように構成されている。機械学習システムによる入力変数１２の伝播の際には、時間的フィルタリングが使用される。この時間的フィルタリングは、以下に続く図３において、より詳細に説明される。
【００５３】

図３は、時間的畳み込み、換言すれば時間的フィルタリングの概略図を示している。

【００５４】

図３に示されているテンソル２０は、機械学習システム１０の入力変数１２又は中間変数であるものとしてもよい。この中間変数は、例えば、層１１のうちの１つの出力変数である。

【００５５】

テンソル２０の左下には、座標系２３が概略的に示されている。この座標系２３は、例示的に３つの次元ｈ，ｔ，ｗを含み、これらの次元ｈ，ｔ，ｗに沿ってテンソル２０は延在する。画像１３は、「時間的」次元ｔに沿って相前後してテンソル２０内に配置されている点に留意されたい。次元ｈ及びｗは、例示的な空間的次元であり、これらの次元内に画像１３は延在する。

【００５６】

ここでは、フィルタ２１を用いて、テンソル２０が時間的にフィルタリングされる。このことは、「時間的」次元ｔに沿ったフィルタが、テンソル２０に沿ってフィルタリングを実施することによって行われる。テンソル２０を用いたフィルタ２１の時間的畳み込み、特に時間的畳み込みの例示的な経過は、破線の矢印２２に基づいて概略的に示されている。

【００５７】

時間的フィルタリングは、代替的に、さらなるフィルタ２１ａ，２１ｂ，２１ｃのうちの１つを用いて実施することができる。さらなるフィルタは、それらが複数の異なる次元を有しているという点で互いに異なる。例えば、さらなるフィルタ２１ｃは、３次元フィルタである。テンソル２０を用いた３次元フィルタ２１ｃの畳み込みの可能な経過は、３次元フィルタ２１ｃに隣接する破線の矢印２２ｂによって例示的に示されている。テンソル２０を用いた３次元フィルタ２１ｃのフィルタリングは、畳み込みが破線の矢印２２ｂに応じて実施される場合、時間的フィルタリングも空間的フィルタリングも表す点に留意されたい。

【００５８】

図４は、機械学習システム１０を動作させる方法の概略的フローチャートを示している。

【００５９】

第１の実施形態においては、この方法は、ステップＳ４０において開始される。このステップにおいては、入力変数１２が、機械学習システム１０に提供される。

【００６０】

それに続くステップＳ４１においては、提供された入力変数１２が機械学習システム１０によって伝搬される。この伝搬の際に、入力変数１２か又は機械学習システム１０の中間変数が、フィルタ２１を用いて時間的にフィルタリングされる（図３に対する説明を参照）。このフィルタは、量子化フィルタ係数を有する。好適には、機械学習システム（１０）のすべてのさらなるパラメータも同様に量子化されている。

【００６１】

それに続くステップＳ４２においては、機械学習システムの出力変数が出力される。

【００６２】

これに続いて、後続のステップＳ４３においては、新たな画像が検出され、入力変数１２に追加され、入力変数１２の含まれる画像が入力変数１２から削除される。

【００６３】

これに続いて、この方法は、そのすべてのステップＳ４０乃至Ｓ４３を繰り返し実施する。

【００６４】

機械学習システム１０を動作させる方法のさらなる実施形態においては、ステップＳ４０乃至Ｓ４３は、機械学習システム１０の学習のために使用される。この学習の際、ステップＳ４３の後で付加的に、損失関数（loss function）に依存して機械学習システムのパラメータが最適化される。この損失関数は、この場合、機械学習システム１０の算出さ

れた出力変数と、それぞれ使用された入力変数の割り当てられたトレーニング出力変数との間の偏差を特徴付ける。出力変数の算出の際には、既に量子化されたパラメータを使用することができる。機械学習システムのパラメータを介した損失関数の最適化は、例えば「逆伝搬（スルー時間）」などの例えば勾配降下法を用いて実施することができる。この場合、この勾配降下法は、パラメータに依存して損失関数を最小化又は最大化するために使用される。パラメータが最適化された後においては、当該パラメータ、特にフィルタ係数が量子化される。好適には、パラメータの最適化のために、少なくとも３２ビットのパラメータの分解能が使用される。次いで、量子化の際に分解能は、例えば、８ビット又は１ビットにまで低減される。

【００６５】

機械学習システム１０を動作させる方法の第１の実施形態の発展形態においては、ステップＳ４３の終了後、任意選択的に、機械学習システム１０の出力変数を、この出力変数に依存して少なくとも半自律型のロボットを駆動制御するために使用することができる。この少なくとも半自律型のロボットは、以下の図５において例示的に示されている。

【００６６】

図５は、第１の実施例において少なくとも半自律型の車両１００によって提供されている少なくとも半自律型のロボットの概略図を示している。さらなる実施例においては、少なくとも半自律型のロボットは、サービス、取り付け、又は、定常的な製造ロボット、代替的には、例えば、ドローンなどの自律的な飛行物体であってもよい。

【００６７】

少なくとも半自律型の車両１００は、検出ユニット３０を含み得る。この検出ユニット３０は、例えば、車両１００の周辺環境を検出するカメラであってもよい。検出ユニット３０は、機械学習システム１０に接続されてもよい。機械学習システム１０は、例えば、検出ユニット３０によって提供された入力変数に依存して、及び、機械学習システム１０の複数のパラメータに依存して出力変数を算出する。これらの出力変数は、制御ユニット４０に転送することができる。

【００６８】

制御ユニット４０は、機械学習システム１０の出力変数に依存してアクチュエータを制御し、好適には、これによって車両１００が衝突のない操縦を実施するようにアクチュエータを制御する。第１の実施例においては、アクチュエータは、車両１００のモータ又はブレーキシステムであってもよい。

【００６９】

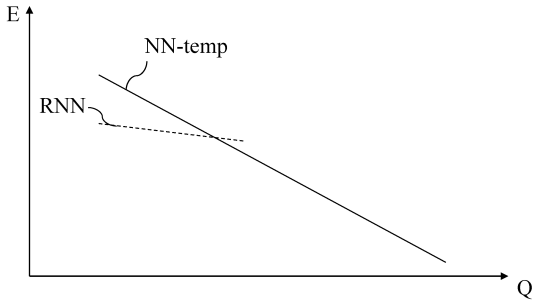
さらなる実施例においては、半自律型のロボットは、工具、工作機械又は製造ロボットであるものとしてもよい。ワークピースの材料は、機械学習システム１０を用いて分類することができる。アクチュエータは、ここでは、例えば、研削ヘッドを駆動するモータであってもよい。

【００７０】

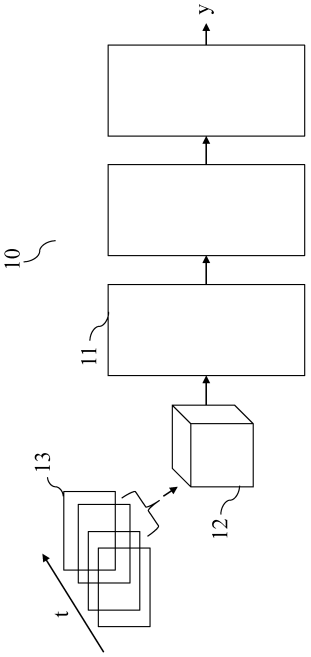
さらに、車両１００、特に半自律型のロボットは、計算ユニット５０及び機械可読メモリ要素６０を含む。このメモリ要素６０上には、計算ユニット５０上において命令が実行されるときに、当該計算ユニット５０を用いて機械学習システム１０を動作させるための命令を含むコンピュータプログラムが記憶されているものとするとよい。

【図面】

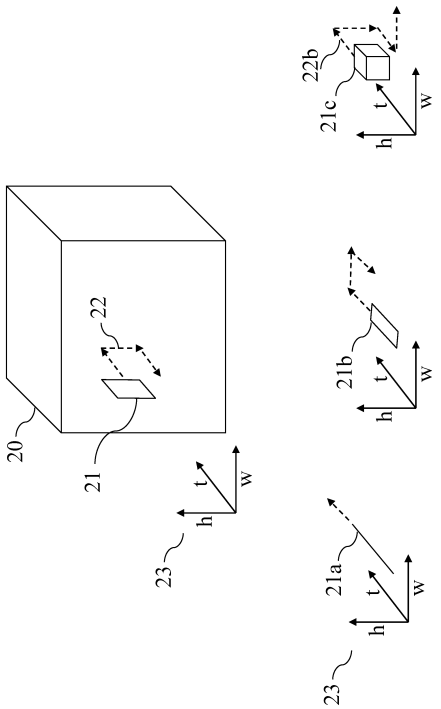
【図１】


E
NN-temp
RNN
Q

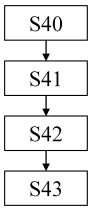
【図２】


y
10
11
13
12
t

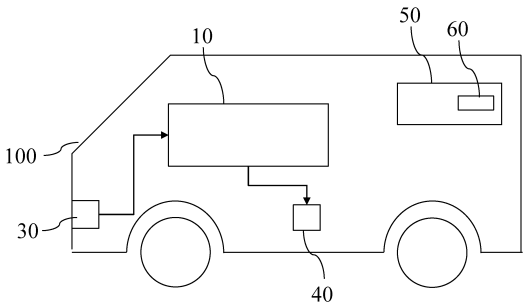
【図３】


22b
21c
t
w
h
22
20
21
21b
23
21a

【図４】


S40
S41
S42
S43

【図５】



フロントページの続き

弁理士　二宮 浩康
(74)代理人　100162880
弁理士　上島 類
(72)発明者　トーマス プファイル
ドイツ連邦共和国 レニンゲン アンナートイラーシュトラーセ １４
審査官　児玉 崇晶
(56)参考文献　特開平１０－３２６２６５（ＪＰ，Ａ）
Hyunhoon LEE et al.，“ Fixed-Point Quantization of 3D Convolutional Neural Networks for Energy-Efficient Action Recognition ”，2018 International SoC Design Conference (ISOCC)，2018年11月，DOI: 10.1109/ISOCC.2018.8649987
Md Zahangir Alom et al.，“ Effective Quantization Approaches for Recurrent Neural Networks ”，arXiv，2018年02月07日，DOI: 10.48550/arXiv.1802.02615
Okan Kopuklu et al.，“ Resource Efficient 3D Convolutional Neural Networks ”，arXiv，2021年10月18日，DOI: 10.48550/arXiv.1904.02422
Shuiwang JI et al.，“ 3D Convolutional Neural Networks for Human Action Recognition ”，IEEE Transactions on Pattern Analysis and Machine Intelligence，2013年01月，Vol. 35, No. 1，pp.221-231，DOI: 10.1109/TPAMI.2012.59
(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ０６Ｎ　３／０２
Ｇ０６Ｔ　７／００
Ｇ０６Ｎ　３／０４９５
Ｇ０６Ｎ　３／０４６４