(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
2022年1月13日(13.01.2022)

WIPO | PCT

(10) 国際公開番号
WO 2022/009433 A1

(51) 国際特許分類:
G06F 3/06 (2006.01)

(21) 国際出願番号： PCT/JP2020/027134

(22) 国際出願日： 2020年7月10日(10.07.2020)

(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: 富士通株式会社 (**FUJITSU LIMITED**) [JP/JP]; 〒2118588 神奈川県川崎市中原区上小田中４丁目１番１号 Kanagawa (JP).

(72) 発明者: 坂井 靖文 (**SAKAI, Yasufumi**); 〒2118588 神奈川県川崎市中原区上小田中４丁目１番１号 富士通株式会社内 Kanagawa (JP).

(74) 代理人: 真田 有，外 (**SANADA, Tamotsu** et al.); 〒1800004 東京都武蔵野市吉祥寺本町１丁目１０番３１号 ＮＭＦ吉祥寺本町ビル８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, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, 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, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

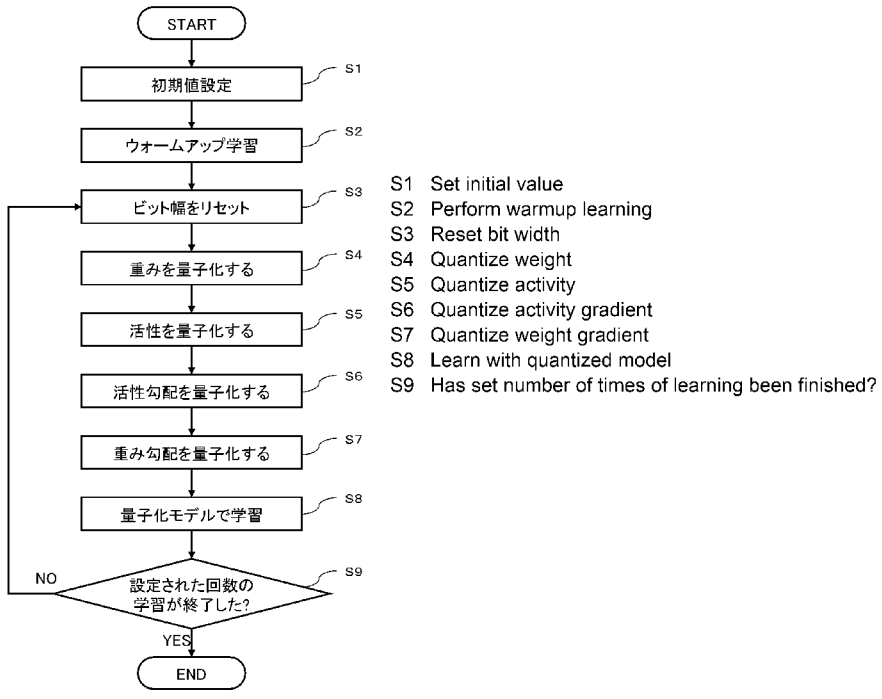
(84) 指定国(表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM,

WO 2022/009433 A1

(54) **Title:** INFORMATION PROCESSING DEVICE, INFORMATION PROCESSING METHOD, AND INFORMATION PROCESSING PROGRAM

(54) 発明の名称： 情報処理装置，情報処理方法および情報処理プログラム

[図9]



(57) **Abstract:** The present invention reduces the execution time of a neural network by uniformly quantizing a plurality of elements included in a gradient vector during back propagation of the neural network.

(57) 要約：ニューラルネットワークの逆伝播時の勾配ベクトルに含まれる複数の要素を一様に量子化することで、ニューラルネットワークの実行時間を短縮する。

［続葉有］

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, 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)）

明　細　書

発明の名称 ：

情報処理装置，情報処理方法および情報処理プログラム

技術分野

[0001] 本発明は、情報処理装置，情報処理方法および情報処理プログラムに関する。

背景技術

[0002] 画像処理などで顕著な成果を示しているニューラルネットワークは、その構成を複雑にすることで高い性能を実現している。

[0003] このように複雑化の傾向にあるニューラルネットワークの実行時間を短縮する手法として量子化が知られている。

[0004] 量子化においては、ニューラルネットワークに使用される重みのデータ型（例えばFP32）をよりデータ容量の小さなデータ型（INT8）に変換することで、計算時間や通信時間を削減する。また、従来の量子化手法においては、重みベクトルを構成する要素毎に量子化の実行判断が行なわれる。

例えば、重みベクトルを構成する要素毎に、量子化誤差を閾値と比較し、量子化誤差が閾値未満である場合にのみ、当該要素に対して量子化が行なわれる。

先行技術文献

特許文献

[0005] 特許文献1：国際公開第２０１９／００８７５２号

[0006] 非特許文献1：S. Khoram, et al, “Adaptive quantization of neural networks”, ICLR, 2018

発明の概要

発明が解決しようとする課題

[0007] しかしながら、上述の如き従来の量子化手法においては、重みだけを量子化するとともに、重みベクトルの要素毎に量子化の実行判断を行なうので、

ニューラルネットワークの学習に当てはめた場合に、実行時間の短縮度合いが低いという課題がある。

１つの側面では、本発明は、ニューラルネットワークの実行時間を短縮できるようにすることを目的とする。

課題を解決するための手段

[0008] このため、この情報処理装置は量子化対象データの量子化を行なう処理を実行するプロセッサを有する情報処理装置であって、前記プロセッサは、ニューラルネットワークの逆伝播時の勾配ベクトルに含まれる複数の要素を一様に量子化する。

発明の効果

[0009] 一実施形態によれば、ニューラルネットワークの実行時間を短縮することができる。

図面の簡単な説明

[0010] [図1]実施形態の一例としての計算機システムのハードウェア構成を例示する図である。

[図2]実施形態の一例としての計算機システムの機能構成を例示する図である。

[図3]ニューラルネットワークの概要を示す図である。

[図4]実施形態の一例としての計算機システムにおけるLossの変異を例示する図である。

[図5]実施形態の一例としての計算機システムにおける活性の閾値を説明するための図である。

[図6]実施形態の一例としての計算機システムにおける勾配の閾値を説明するための図である。

[図7]実施形態の一例としての計算機システムにおいて重み勾配の閾値を設定するための条件式を示す図である。

[図8]実施形態の一例としての計算機システムにおける活性勾配の閾値を説明するための図である。

[図9]実施形態の一例としての計算機システムにおける学習手順を説明するためのフローチャートである。

[図10]実施形態の一例としての計算機システムにおける量子化処理を説明するためのフローチャートである。

[図11]実施形態の一例としての計算機システムによる量子化を行なったニューラルネットワークによる学習結果を、従来手法により量子化した場合および量子化なしで学習させた場合のそれぞれと比較して示す図である。

[図12]実施形態の一例としての計算機システムによる量子化を行なったニューラルネットワークによる学習結果を、従来手法により量子化した場合および量子化なしで学習させた場合のそれぞれと比較して示す図である。

発明を実施するための形態

[0011] 以下、図面を参照して本情報処理装置，情報処理方法および情報処理プログラムにかかる実施の形態を説明する。ただし、以下に示す実施形態はあくまでも例示に過ぎず、実施形態で明示しない種々の変形例や技術の適用を排除する意図はない。すなわち、本実施形態を、その趣旨を逸脱しない範囲で種々変形して実施することができる。また、各図は、図中に示す構成要素のみを備えるという趣旨ではなく、他の機能等を含むことができる。

[0012] （Ａ）構成

図１は実施形態の一例としての計算機システム１のハードウェア構成を例示する図である。

[0013] 計算機システム１は、情報処理装置であって、量子化されたニューラルネットワークを実現する。計算機システム１は、図１に示すように、ＣＰＵ（Central Processing Unit）１０，メモリ１１およびアクセラレータ１２を備える。これらのＣＰＵ１０，メモリ１１およびアクセラレータ１２は、通信バス１３を介して相互に通信可能に接続されている。通信バス１３は、本計算機システム１内のデータ通信を行なう。

[0014] メモリ１１は、ＲＯＭ（Read Only Memory）およびＲＡＭ（Random Access Memory）を含む記憶メモリである。メモリ１１のＲＯＭには、量子化処理に

係るソフトウェアプログラムやこのプログラム用のデータ類が書き込まれている。メモリ１１上のソフトウェアプログラムは、ＣＰＵ１０に適宜読み込まれて実行される。また、メモリ１１のＲＡＭは、一次記憶メモリあるいはワーキングメモリとして利用される。メモリ１１のＲＡＭには、重み，活性，重み勾配，活性勾配等の量子化に用いられるパラメータや各種閾値等も格納される。

アクセラレータ１２は、行列演算などのニューラルネットワークの計算に必要な演算処理を実行する。

[0015] ＣＰＵ１０は、種々の制御や演算を行なう処理装置（プロセッサ）であり、実装されたプログラムに基づき、計算機システム１全体を制御する。そして、このＣＰＵ１０がメモリ１１等に格納された深層学習処理プログラム（図示省略）を実行することで、後述する深層学習処理部１００（図２）としての機能を実現する。

[0016] また、深層学習処理プログラムは情報処理プログラムを含んで構成されてもよい。ＣＰＵ１０がメモリ１１等に格納された情報処理プログラム（図示省略）を実行することで、後述する量子化処理部１０１（図２）としての機能を実現する。

[0017] そして、計算機システム１のＣＰＵ１０が、深層学習処理プログラム（情報処理プログラム）を実行することにより、深層学習処理部１００（量子化処理部１０１）として機能する。

[0018] なお、深層学習処理部１００（量子化処理部１０１）としての機能を実現するためのプログラム（情報処理プログラム）は、例えばフレキシブルディスク，ＣＤ（ＣＤ－ＲＯＭ，ＣＤ－Ｒ，ＣＤ－ＲＷ等），ＤＶＤ（ＤＶＤ－ＲＯＭ，ＤＶＤ－ＲＡＭ，ＤＶＤ－Ｒ，ＤＶＤ＋Ｒ，ＤＶＤ－ＲＷ，ＤＶＤ＋ＲＷ，ＨＤ　ＤＶＤ等），ブルーレイディスク，磁気ディスク，光ディスク，光磁気ディスク等の、コンピュータ読取可能な記録媒体に記録された形態で提供される。そして、コンピュータ（計算機システム１）はその記録媒体からプログラムを読み取って内部記憶装置または外部記憶装置に転送し格

納して用いる。また、そのプログラムを、例えば磁気ディスク，光ディスク，光磁気ディスク等の記憶装置（記録媒体）に記録しておき、その記憶装置から通信経路を介してコンピュータに提供するようにしてもよい。

[0019] 深層学習処理部１００（量子化処理部１０１）としての機能を実現する際には、内部記憶装置（本実施形態ではメモリ１１のＲＡＭやＲＯＭ）に格納されたプログラムがコンピュータのマイクロプロセッサ（本実施形態ではＣＰＵ１０）によって実行される。このとき、記録媒体に記録されたプログラムをコンピュータが読み取って実行するようにしてもよい。

[0020] 図２は実施形態の一例としての計算機システム１の機能構成を例示する図である。

計算機システム１は、図２に示すように、深層学習処理部１００としての機能を備える。深層学習処理部１００は、ニューラルネットワークにおける深層学習を実施する。

[0021] ニューラルネットワークは、ハードウェア回路であってもよいし、ＣＰＵ１０等によりコンピュータプログラム上で仮想的に構築される階層間を接続するソフトウェアによる仮想的なネットワークであってもよい。

[0022] 図３にニューラルネットワークの概要を示す。図３に示すニューラルネットワークは、入力層と出力層との間に複数の隠れ層を含むディープニューラルネットワークである。例えば、隠れ層は、畳み込み層、プーリング層または全結合層等である。各層に示す丸印は、所定の計算をそれぞれ実行するノードを示す。

[0023] ニューラルネットワークは、例えば、画像や音声などの入力データを入力層に入力し、畳み込み層やプーリング層などで構成される隠れ層にて所定の計算を順次実行することで、演算により得られる情報を入力側から出力側に順次伝えるフォーワッド方向の処理（順伝播処理）を実行する。フォーワッド方向の処理の実行後、出力層から出力される出力データと正解データとから得られる誤差関数の値を小さくするために、フォーワッド方向の処理で使用するパラメータを決定するバックワード方向の処理（逆伝播処理）を実行

する。そして、逆伝播処理の結果に基づいて重み等の変数を更新する更新処理が実行される。例えば、逆伝播処理の計算に使用される重みの更新幅を決定するアルゴリズムとして、勾配降下法が使用される。

[0024] 深層学習処理部１００は、量子化処理部１０１を備える。

量子化処理部１０１は、ニューラルネットワークに使用される変数を量子化する。量子化処理部１０１は、図２に示すように、量子化誤差算出部１０２，閾値設定部１０３，量子化実行部１０４および認識率比較部１０５としての機能を備える。

[0025] 本計算機システム１においては、ニューラルネットワークに使用されるテンソル（ベクトル）に含まれる要素のデータ型を全て同じにして量子化する。すなわち、量子化処理部１０１は、テンソル毎に量子化を行なう。これにより、量子化操作数を削減することができる。

[0026] 例えば、図３に例示するニューラルネットワークにおいて、全結合の場合、重みテンソルの要素数は、テンソルが存在する箇所の入出力レイヤのノード数で決まる。テンソルの要素数は以下の式（１）で示すことができる。

テンソルの要素数 = （入力レイヤのノード数）×（出力レイヤのノード数）・・・（１）

[0027] 例えば、入出力レイヤのノード数がたった100であっても、テンソルの要素数は10000にもなってしまう。すなわち、従来手法において要素毎に量子化すると、量子化操作を10000回行なう必要がある。

[0028] 本計算機システム１においては、量子化実行部１０４がテンソルに含まれる要素のデータ型を全て同じにして一様に量子化することで、量子化のための演算行回数を大幅に削減することができる。

[0029] 量子化実行部１０４は、量子化対象の要素について、後述する量子化誤差算出部１０２によって算出される量子化誤差と閾値とを比較し、誤差が閾値よりも小さい場合に、テンソルの量子化ビット幅を決定し、量子化を行なう

。なお、閾値は後述する閾値設定部１０３により設定される。また、量子化ビット幅の決定は既知の手法により実現することができ、その説明は省略する。

[0030] 量子化実行部１０４は、量子化対象の要素の量子化誤差と閾値との比較の結果、量子化誤差が閾値よりも小さい場合に量子化を行ない、量子化誤差が閾値以上である場合には量子化を行なわない。

以下、本実施形態においては、量子化前におけるニューラルネットワークに使用される変数のデータ型がFP32である例について示す。

テンソル毎の閾値を以下の式（２）に示す。

[0031] [数1]

$$\Delta W = \frac{1}{n}\sum_{i=1}^{n}\Delta w_i \leq \sum_{i=1}^{n}\frac{L(FP32) - L(W) - L_{th}}{n \cdot \left|\frac{\partial L(W)}{\partial w_i}\right|} \quad \cdots (2)$$

[0032] L(FP32)： float32に設定したときに得られる損失関数の値

L(W)：設定されたビット幅で量子化したときに得られる損失関数の値

Lth ：外部から設定される損失関数のマージン

∂L(W)/∂w：設定されたビット幅で逆伝搬したときに得られる勾配

n：量子化対象のテンソルに含まれる要素の数

[0033] 上記式（２）においては、テンソルの量子化誤差を、全要素の量子化誤差の合計と定義している。以下、式中における既述の符号と同一の符号は同様の部分を示しているので、その説明は省略する。

量子化実行部１０４は、レイヤ毎に、重み，活性，重み勾配および活性勾配に対してそれぞれ量子化を行なう。

[0034] また、量子化実行部１０４は、後述する認識率比較部１０５による量子化モデルの認識率が、量子化前の認識率よりも劣化したと判定され、パラメータの再設定が行なわれると、再度、量子化を行なう。

ここで、テンソル毎の誤差閾値は以下の式（３）により求められる。

[0035]

[数2]

$$Q_{th} = \sum_{i=1}^{n} \frac{L(FP32) - L(W) - \mathrm{L}_{th}}{n \cdot \left| \frac{\partial L(W)}{\partial w_i} \right|} \quad \cdots (3)$$

[0036] 後述する閾値設定部１０３は、閾値の導出に、FP32におけるLoss（損失）の値（L(FP32)）を使用する。従来手法は蒸留向けであったので、FP32のLossは既知であった。しかしながら、FP32のLossは学習中に変動する。

[0037] そこで、本計算機システム１においては、量子化実行部１０４は、学習時に量子化を行なうたびに、その時点でのLoss(FP32)を順伝播処理により取得する。すなわち、量子化実行部１０４は、変数のデータ型を一時的にFP32にして順伝播を行なうことで、その時点でのLoss(FP32)を取得する。

[0038] 図４は実施形態の一例としての計算機システム１におけるLossの変異を例示する図である。

図４において、横軸は学習時間（Training time）を示し、縦軸はFP32のLossを示す。Loss(FP32)の値は、学習時間に応じて減少する。

[0039] 量子化処理部１０１は、テンソル毎に、ビット幅を変えながら量子化を行ない、量子化誤差を算出する。量子化処理部１０１は、量子化誤差が閾値よりも小さくなる最小のビット幅を求める。量子化処理部１０１は、ビット幅調整時に、FP32のlossを取得する。

[0040] 閾値設定部１０３は、損失関数の値と、ニューラルネットの学習により得られる勾配から閾値を生成する。

閾値設定部１０３は、重み，活性，重み勾配および活性勾配に対して、それぞれ閾値を設定する。

[0041] ［重みの閾値］

ｋ回目の学習時に量子化を行なう際に用いる重みの閾値$Q_{th,\Delta Wk}$を式（４）に示す。

[0042]

[数3]

$$Q_{th,\Delta W_k} = \sum_{i=1}^{n} \frac{\left| L_{full,k} - L_{Q,k} - L_{th} \right|}{n \cdot \left| \frac{\partial L_{Q,k}}{\partial w_{i,k}} \right|} \quad \cdots (4)$$

[0043] なお、$L_{full,k}$は、k回目の学習結果（ネットワークモデル）を量子化せずに導出した、検証データによる損失関数である。また、$L_{Q,k}$は、k回目の学習結果（ネットワークモデル）を量子化して導出した、検証データによる損失関数である。

[0044] ［活性の閾値］

ｋ回目の学習時に量子化を行なう際に用いる活性の閾値$Q_{th,\Delta Xk}$を式（５）に示す。

[0045] [数4]

$$Q_{th,\Delta X_k} = \sum_{i=1}^{n} \frac{\left| L_{full,k} - L_{Q,k} - L_{th} \right|}{n \cdot \left| \frac{\partial L_{Q,k}}{\partial x_{i,k}} \right|} \quad \cdots (5)$$

[0046] 図５は実施形態の一例としての計算機システム１における活性の閾値を説明するための図である。

図５において、符号Ａはニューラルネットワークにおける３つのノード（レイヤ）＃１～＃３を示し、ノード＃１は入力層、ノード＃２は隠れ層、ノード＃３は出力層にそれぞれ相当する。

[0047] 符号Ｂは符号Ａに示すニューラルネットワークのノード＃２におけるフォーワッド時の計算グラフを示す。X_1は入力データであり、X_2，Y_1は活性（中間出力データ）である。W_1，W_2は重みであり、B_1，B_2はバイアスである。これらの、入力データ，活性，重み，バイアスはそれぞれテンソル（ベクトル）である。

[0048] 符号Ｃは符号Ａに示すニューラルネットワークのノード＃２におけるバックワード時の計算グラフを示す。gX_1はX_1の勾配であり、gX_2はX_2の勾配である

。gW_1，gW_2はW_1，W_2の各勾配であり、gB_1，gB_2はB_1，B_2の各勾配である。gY_1はY_1の勾配である。Y_1の閾値とX_2の閾値とは等しく、Y_1の勾配gY_1は、X_2の勾配gX_2と等しい（$gY_1 = gX_2$）。これらの勾配はそれぞれテンソル（ベクトル）である。

活性の閾値においては、順伝播時の活性に対応する活性勾配を感度係数に用いる。

入力データのテンソルＸ$_1$の閾値$Q_{th,\Delta X1}$を式（６）に表す。

[0049] [数5]

$$Q_{th,\Delta X_1} = \sum_{i=1}^{n} \frac{\left|L_{32,k} - L_Q\right|}{n \cdot \left|\frac{\partial L_Q}{\partial x_{1,i}}\right|} \quad \cdots (6)$$

式（６）においては、X_1の勾配gX_1を感度係数に用いている。

[0050] また、中間データのテンソルY_1の閾値$Q_{th,\Delta Y1}$を式（７）に表す。

[0051] [数6]

$$Q_{th,\Delta Y_1} = \sum_{i=1}^{n} \frac{\left|L_{32,k} - L_Q\right|}{n \cdot \left|\frac{\partial L_Q}{\partial x_{2,i}}\right|} \quad \cdots (7)$$

式（７）においては、X_2の勾配gX_2を感度係数に用いている。

[0052] ［勾配の閾値について］

閾値設定部１０３は、勾配について、次回の学習時の損失関数と勾配（感度係数）を用いて閾値を算出する。

ある学習時の損失関数や勾配には、現在の学習時の重み・活性の値しか反映されない。また、勾配の量子化誤差は、現在の学習時の損失関数や勾配に反映されない。

例えばk回目の学習時の損失関数や勾配には、k回目の学習時の重み・活性の値しか反映されず、k回目の学習時に発生する勾配の量子化誤差は、k回目の学習時の損失関数や勾配には反映されない。

[0053] 図６は実施形態の一例としての計算機システム１における勾配の閾値を説明するための図である。

図６において符号Ａはk回目の学習を示し、符号Ｂはk+1回目の学習を示す。

k回目の学習における損失関数は$L_k(W+\Delta W, X+\Delta X)$で表わされる。また、k+1回目の学習における損失関数は$L_{k+1}(W+\Delta W, X+\Delta X)$で表わされる。

閾値設定部１０３は、k+1回目の損失関数と勾配を用いて、k回目の勾配の閾値を決める。

[0054] ［重み勾配の閾値］

k回目の学習時に量子化を行なう際に用いる重み勾配の閾値$Q_{th,\Delta GWk}$を式（８）に示す。

[0055] [数7]

$$Q_{th,\Delta GW_k} = \frac{1}{\eta}\sum_{i=1}^{n}\frac{\left|L_{full,k+1} - L_{Q,k+1} - L_{th}\right|}{n \cdot \left|\frac{\partial L_{Q,k+1}}{\partial w_{i,k+1}}\right|} \quad \cdots (8)$$

ここで、ηは重み更新時に使用される学習率である。k回目の重み勾配を量子化することで、k+1回目の重みに量子化誤差ΔW_{k+1}が含まれる。

[0056] 図７に実施形態の一例としての計算機システム１において重み勾配の閾値を設定するための条件式を示す。

図７において、符号Ａは、k+1回目の重みに誤差が含まれても、その損失関数がk+1回目の理想損失関数よりも小さくなる条件を示す。

[0057] この符号Ａに示す不等式の左辺は、k回目の重み勾配を量子化することで発生した量子化誤差を表す。一方、符号Ａに示す不等式の右辺（$Loss_{FP32,k+1}$）は、量子化を行なわない場合の損失を表す。すなわち、k回目の重み勾配を量子化することで発生した量子化誤差が、量子化を行なわない場合の損失と同等以下であることを条件とする。

[0058] 図７において、符号Ｂに表す不等式は、量子化誤差ΔW_{k+1}を表すものであり、符号Ａで表した不等式を変換することで求められる。このように、k+1回目

の損失関数と勾配とを用いることで、k回目の重み勾配に許容される誤差の閾値が導出される。

[0059] ［活性勾配の閾値］

ｋ回目の学習時に量子化を行なう際に用いる活性勾配の閾値$Q_{th,\Delta GXk}$を式（９）に示す。

[0060] [数8]

$$Q_{th,\Delta GX_k} = \frac{1}{\eta}\sum_{i=1}^{n}\frac{\left|L_{full,k+1} - L_{Q,k+1} - L_{th}\right|}{n \cdot \left|\frac{\partial L_{Q,k+1}}{\partial x_{i,k+1}}\right|} \quad \cdots (9)$$

[0061] 閾値設定部１０３は、バイアス誤差の閾値を、活性誤差の閾値に用いる。テンソルの量子化誤差は、全要素の和と定義する。バイアス勾配と活性勾配の量子化誤差のL1ノルムは同じである。

従って、重み勾配の閾値の導出法の考え方を考慮すると、k回目の活性勾配の閾値は、k+1回目の損失関数と活性勾配で算出できる。

図８は実施形態の一例としての計算機システム１における活性勾配の閾値を説明するための図である。

[0062] 図８において、符号Ａはニューラルネットワークにおける３つのノード（レイヤ）＃１～＃３を示し、ノード＃１は入力層、ノード＃２は隠れ層、ノード＃３は出力層にそれぞれ相当する。

[0063] 符号Ｂは符号Ａに示すニューラルネットワークにおけるフォーワッド時の計算グラフを示す。X_1は入力データのテンソルである。また、符号Ｃは符号Ａに示すニューラルネットワークにおけるバックワード時の計算グラフを示す。

[0064] gx_1はx_1の勾配であり、gx_2はx_2の勾配である。gw_1，gw_2はw_1，w_2の各勾配であり、gb_1，gb_2はb_1，b_2の各勾配である。gx_1はx_1の勾配であり、gy_1はy_1の勾配である。勾配gy_1は勾配gx_2と等しい（$gy_1 = gx_2$）。また、勾配gb_1と勾配gx_2のL1ノルムは等しい（$gb_1 = gx_2$）。また、勾配gy_2および勾配gb_2のL1ノルムは、それぞ

れ損失ΔLのL1ノルムと等しい（gy_2= gb_2=ΔL）。勾配gw_2は、勾配gy_2とx_2との内積に等しい（gw_2＝gy_2・x_2）。勾配gx_2は、勾配gy_2と重みw_2との内積に等しい（gx_2＝gy_2・w_2）。また、勾配gw_2は、勾配gy_2とx_2との内積に等しい（gw_2＝gy_2・x_2）。

gb_nとgy_nのL1ノルムが等しいため、バイアス勾配と活性勾配の閾値とは同じとなる。k回目の活性勾配の閾値は、k+1回目の活性勾配と損失関数とで定義される。

[0065] 認識率比較部１０５は、量子化処理部１０１により、レイヤ毎に重み，活性，重み勾配および活性勾配に対して量子化されたモデル（量子化モデル）の認識率とFP32の認識率とを比較する。

[0066] 比較の結果、量子化モデルの認識率がFP32の認識率と同等でない場合、例えば、量子化モデルの認識率がFP32の認識率と比べて所定値以上低い場合に、認識率比較部１０５は、上述した各閾値が小さくなるようにパラメータを設定する制御を行なう。具体的には、認識率比較部１０５は、各閾値が小さくなるようにLthを増加させる。

[0067] 量子化度合いが大きいと、精度劣化のリスクが高まる。そこで、本計算機システム１において、認識率比較部１０５は、ビット幅決定アルゴリズムの外側に、精度保証するループを備える。

[0068] 本計算機システム１においては、量子化モデルの認識率がFP32の認識率よりも劣化したと判定された場合は、重み，活性，重み勾配および活性勾配の各閾値が小さくなるようにパラメータを再設定したのちに、再度、重み、活性、重み勾配および活性勾配に対して量子化を行なう。

[0069] （Ｂ）動作

上述の如く構成された実施形態の一例としての計算機システム１における学習手順を、図９に示すフローチャート（ステップＳ１～Ｓ９）に従って説明する。

ステップＳ１において、各変数等の初期値が行なわれる。

ステップＳ２において、量子化処理部１０１は、ウォームアップ学習を行

なう。

ステップＳ３において、量子化処理部１０１は、ビット幅をリセットする。

[0070] 量子化処理部１０１は、ニューラルネットワークに使用されるテンソルの量子化誤差を算出する。閾値設定部１０３は、損失関数の値と、ニューラルネットワークの学習により得られる勾配から、重み，活性，重み勾配および活性勾配の各閾値を算出する。

[0071] 量子化実行部１０４は、算出された量子化誤差と閾値とを比較し、テンソルの量子化ビット幅を決定する。量子化実行部１０４は重みを量子化する（ステップＳ４）。また、量子化実行部１０４は、活性を量子化し（ステップＳ５）、活性勾配を量子化し（ステップＳ６）、さらに、重み勾配を量子化する（ステップＳ７）。

[0072] なお、これらのステップＳ４～Ｓ７の処理順序はこれに限定されるものではなく、順序を入れ替える等、適宜変更して実施することができる。また、これらのステップＳ４～Ｓ７における各テンソル毎の量子化処理の詳細は、図１０に示すフローチャートを用いて後述する。

その後、ステップＳ８において、深層学習処理部１００は、量子化したモデル（量子化モデル）を用いた学習を所定の回数だけ行なう。

[0073] ステップＳ９において、深層学習処理部１００は、設定された回数の学習が終了したかを確認する。確認の結果、設定された回数の学習が終了していない場合には（ステップＳ９のＮＯルート参照）、ステップＳ３に戻る。また、設定された回数の学習が終了した場合には（ステップＳ９のＹＥＳルート参照）、処理を終了する。

[0074] 次に、実施形態の一例としての計算機システム１における量子化処理を、図１０に示すフローチャート（ステップＳ１１～Ｓ１７）に従って説明する。

ステップＳ１１において、量子化実行部１０４は、順伝播処理によりFP32の損失関数（Loss(FP32)）を算出する。

[0075] ステップＳ１２において、量子化実行部１０４は、量子化するレイヤを決定する。ステップＳ１３において、量子化実行部１０４は、テンソル毎に量子化誤差の閾値Qthを算出する。

[0076] ステップＳ１４において、量子化実行部１０４は、テンソル毎に、ビット幅を変えながら量子化を行ない、量子化誤差を求める。量子化実行部１０４は、量子化誤差が閾値よりも小さくなる最小のビット幅を求める。

[0077] ステップＳ１５において、量子化実行部１０４は、全レイヤの量子化が終了したかを確認する。確認の結果、全レイヤについて量子化が終了していない場合には（ステップＳ１５のＮＯルート参照）、ステップＳ１２に戻る。一方、全レイヤの量子化が終了した場合には（ステップＳ１５のＹＥＳルート参照）、ステップＳ１６に移行する。

[0078] ステップＳ１６において、認識率比較部１０５は、量子化処理部１０１により量子化された量子化モデルの認識率がFP32の認識率と同等であるかを確認する。例えば、量子化モデルの認識率がFP32の認識率と比べて所定値以上低い場合には、量子化モデルの認識率がFP32の認識率と同等ではない。このような場合に（ステップＳ１６のＮＯルート参照）、ステップＳ１７に移行する。

ステップＳ１７において、認識率比較部１０５は、閾値が小さくなるようにLthを増加させる。その後、処理はステップＳ１１に戻る。

[0079] 一方、ステップＳ１６における確認の結果、量子化モデルの認識率がFP32の認識率と同等である場合には（ステップ１６のＹＥＳルート参照）、処理を終了する。

[0080] （Ｃ）効果

このように、実施形態の一例としての計算機システム１によれば、量子化実行部１０４が、テンソル毎に量子化を実行するとともに、逆伝播時に発生する勾配も量子化する。これにより、ニューラルネットワークの実行時間を短縮することができる。

[0081] 量子化実行部１０４は、量子化対象の要素の量子化誤差と閾値とを比較し

、誤差が閾値よりも小さい場合に量子化を行ない、誤差が閾値以上である場合には量子化を行なわない。これにより、量子化による認識率の劣化を抑止することができる。

[0082] 量子化実行部１０４は、テンソル（重みベクトル，活性ベクトル，活性勾配ベクトル，重み勾配ベクトル）に含まれる複数の要素のデータ型を全て同じにして量子化することで、低負荷で量子化を実現することができ、量子化にかかる処理時間を短縮することができる。

[0083] 閾値設定部１０３が、勾配ベクトルに含まれる複数の要素を量子化する際に、後続して実行された次回学習時（k+1回目）の損失関数を用いた閾値を作成することで、勾配ベクトルに関する量子化の実行可否を確実に判断することができる。

[0084] 図１１および図１２は実施形態の一例としての計算機システム１による量子化を行なったニューラルネットワークによる学習結果を、従来手法により量子化した場合および量子化なしで学習させた場合のそれぞれと比較して示す図である。なお、図１１は認識率を、図１２は実行時間の短縮率を、それぞれ示す。

[0085] これらの図１１および図１２において、従来手法は重みだけを量子化する手法を表す。また、量子化なしの手法は全レイヤをfloat32で学習させたものである。

図１１に示すように、本計算機システム１により量子化したニューラルネットワークは、従来手法および量子化なしの場合に比べて同等の認識率を実現する。

[0086] また、図１２に示すように、実行時間については、量子化を行なわない場合の実行時間を基準として、本計算機システム１による量子化手法によれば、実行時間の短縮率を約７割に増加させることができる。

[0087] また、閾値設定部１０３が、次回の学習を行なうことで量子化対象の勾配向けの誤差閾値を算出する。また、活性勾配の誤差閾値にはバイアス閾値を用いる。これにより、量子化を実行するために好適な閾値を決定することが

できる。

[0088] （Ｄ）その他

そして、開示の技術は上述した実施形態に限定されるものではなく、本実施形態の趣旨を逸脱しない範囲で種々変形して実施することができる。本実施形態の各構成および各処理は、必要に応じて取捨選択することができ、あるいは適宜組み合わせてもよい。

また、上述した開示により本実施形態を当業者によって実施・製造することが可能である。

符号の説明

[0089] １　計算機システム

１０　ＣＰＵ

１１　メモリ

１２　アクセラレータ

１３　通信バス

１００　深層学習処理部

１０１　量子化処理部

１０２　量子化誤差算出部

１０３　閾値設定部

１０４　量子化実行部

１０５　認識率比較部

請求の範囲

[請求項1] 量子化対象データの量子化を行なう処理を実行するプロセッサを有する情報処理装置であって、

前記プロセッサは、

ニューラルネットワークの逆伝播時の勾配ベクトルに含まれる複数の要素を一様に量子化する

ことを特徴とする、情報処理装置。

[請求項2] 前記プロセッサは、

前記勾配ベクトルに含まれる複数の要素のデータ型を全て同じにして量子化する

ことを特徴とする、請求項１記載の情報処理装置。

[請求項3] 前記プロセッサは、

前記勾配ベクトルに含まれる複数の要素を量子化する際に、後続して実行された次回学習時の損失関数に基づく閾値で量子化可否を判断する

ことを特徴とする、請求項１または２に記載の情報処理装置。

[請求項4] 前記プロセッサは、

ニューラルネットワークの順伝播時の活性ベクトルを量子化する

ことを特徴とする、請求項１～３のいずれか１項に記載の情報処理装置。

[請求項5] 前記プロセッサは、

前記活性ベクトルに含まれる複数の要素のデータ型を全て同じにして量子化する

ことを特徴とする、請求項４記載の情報処理装置。

[請求項6] 量子化対象データの量子化を行なう処理を実行するプロセッサを有する情報処理装置であって、

前記プロセッサは、

ニューラルネットワークの順伝播時の活性ベクトルを量子化する

ことを特徴とする、情報処理装置。

[請求項7] 前記プロセッサは、

前記活性ベクトルに含まれる複数の要素のデータ型を全て同じにして量子化する

ことを特徴とする、請求項６記載の情報処理装置。

[請求項8] 量子化対象データの量子化を行なう処理を実行するプロセッサを有する情報処理装置において、

閾値に基づき量子化可否を判断する処理と、

ニューラルネットワークの逆伝播時の勾配ベクトルに含まれる複数の要素を一様に量子化する処理と

を備えることを特徴とする、情報処理方法。

[請求項9] 前記量子化する処理において、

前記勾配ベクトルに含まれる複数の要素のデータ型を全て同じにして量子化する

ことを特徴とする、請求項８記載の情報処理方法。

[請求項10] 前記判断する処理において、

前記勾配ベクトルに含まれる複数の要素を量子化する際に、後続して実行された次回学習時の損失関数に基づく閾値で量子化可否を判断する

ことを特徴とする、請求項８または９に記載の情報処理方法。

[請求項11] 前記量子化する処理において、

ニューラルネットワークの順伝播時の活性ベクトルを量子化する

ことを特徴とする、請求項８～１０のいずれか１項に記載の情報処理方法。

[請求項12] 前記量子化する処理において、

前記活性ベクトルに含まれる複数の要素のデータ型を全て同じにして量子化する

ことを特徴とする、請求項１１記載の情報処理方法。

[請求項13] 量子化対象データの量子化を行なう処理を実行するプロセッサを有する情報処理装置において、

閾値に基づき量子化可否を判断する処理と、

ニューラルネットワークの順伝播時の活性ベクトルを量子化する処理と

を備えることを特徴とする、情報処理方法。

[請求項14] 前記量子化する処理において、

前記活性ベクトルに含まれる複数の要素のデータ型を全て同じにして量子化する

ことを特徴とする、請求項１３記載の情報処理方法。

[請求項15] 量子化対象データの量子化を行なう情報処理装置のプロセッサに、

ニューラルネットワークの逆伝播時の勾配ベクトルに含まれる複数の要素を一様に量子化する

処理を実行させることを特徴とする、情報処理プログラム。

[請求項16] 前記勾配ベクトルに含まれる複数の要素のデータ型を全て同じにして量子化する

処理を前記プロセッサに実行させることを特徴とする、請求項１５記載の情報処理プログラム。

[請求項17] 前記勾配ベクトルに含まれる複数の要素を量子化する際に、後続して実行された次回学習時の損失関数に基づく閾値で量子化可否を判断する

処理を前記プロセッサに実行させることを特徴とする、請求項１５または１６に記載の情報処理プログラム。

[請求項18] ニューラルネットワークの順伝播時の活性ベクトルを量子化する

処理を前記プロセッサに実行させることを特徴とする、請求項１５～17のいずれか１項に記載の情報処理プログラム。

[請求項19] 前記活性ベクトルに含まれる複数の要素のデータ型を全て同じにして量子化する

処理を前記プロセッサに実行させることを特徴とする、請求項１８記載の情報処理プログラム。

[請求項20] 量子化対象データの量子化を行なう情報処理装置のプロセッサに、

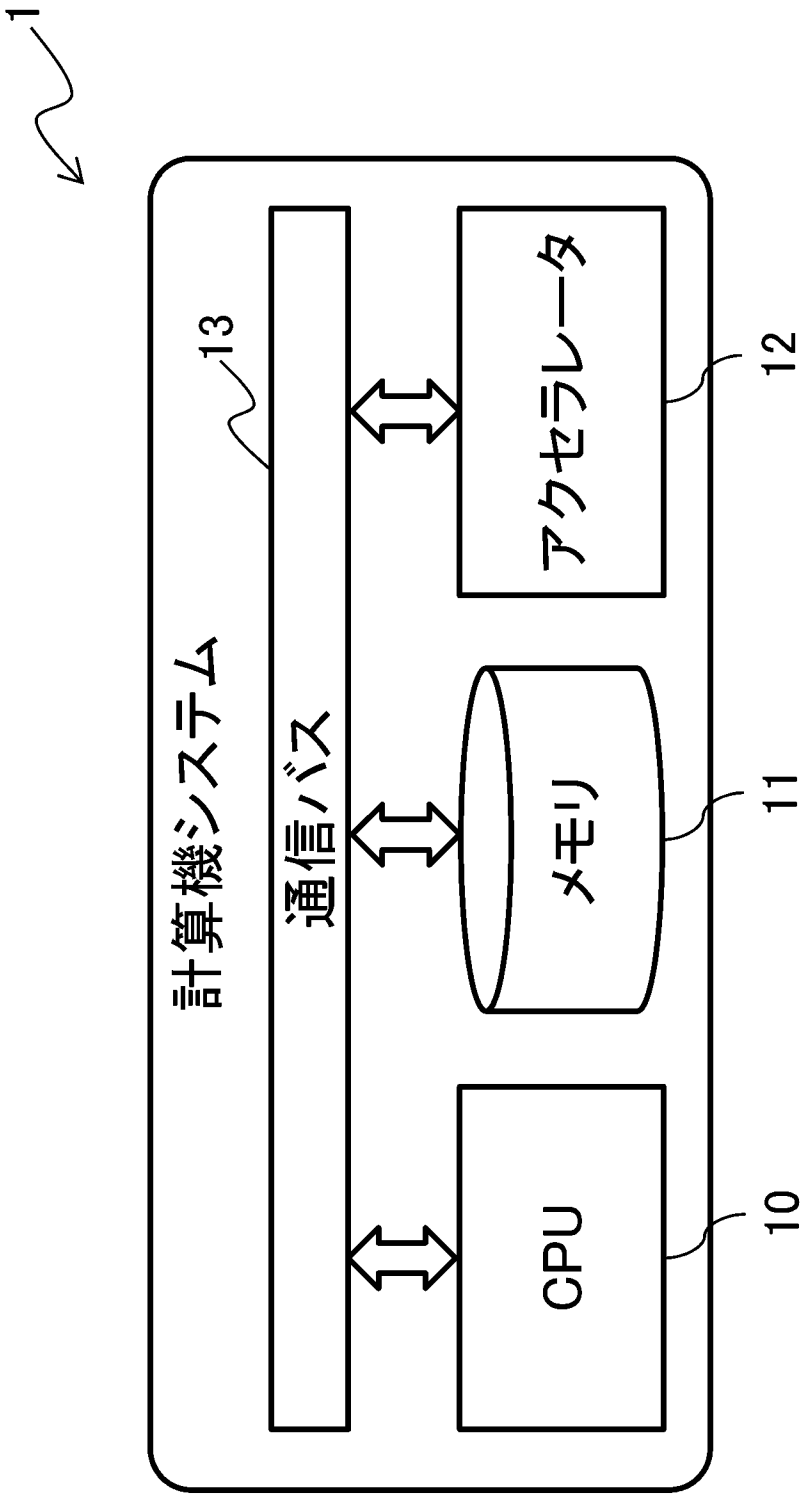
ニューラルネットワークの順伝播時の活性ベクトルを量子化する

処理を実行させることを特徴とする、情報処理プログラム。

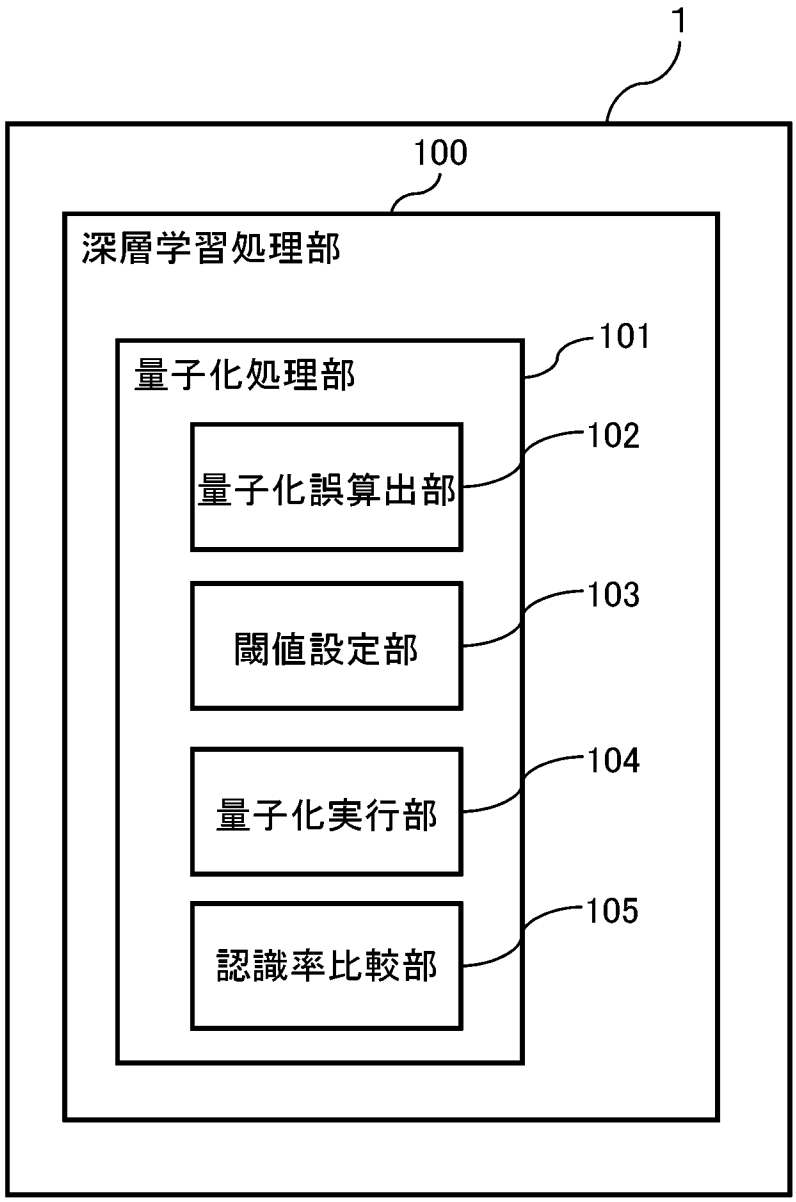
[請求項21] 前記活性ベクトルに含まれる複数の要素のデータ型を全て同じにして量子化する

処理を前記プロセッサに実行させることを特徴とする、請求項２０記載の情報処理プログラム。

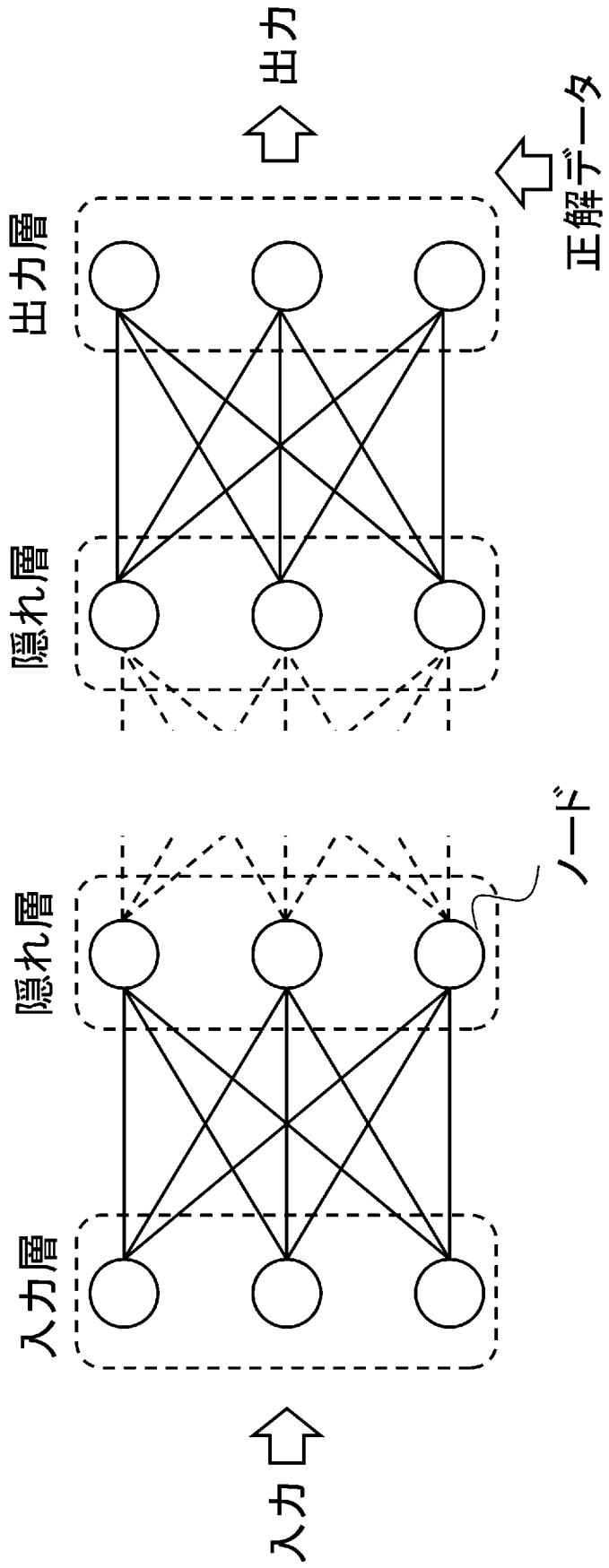
[図1]



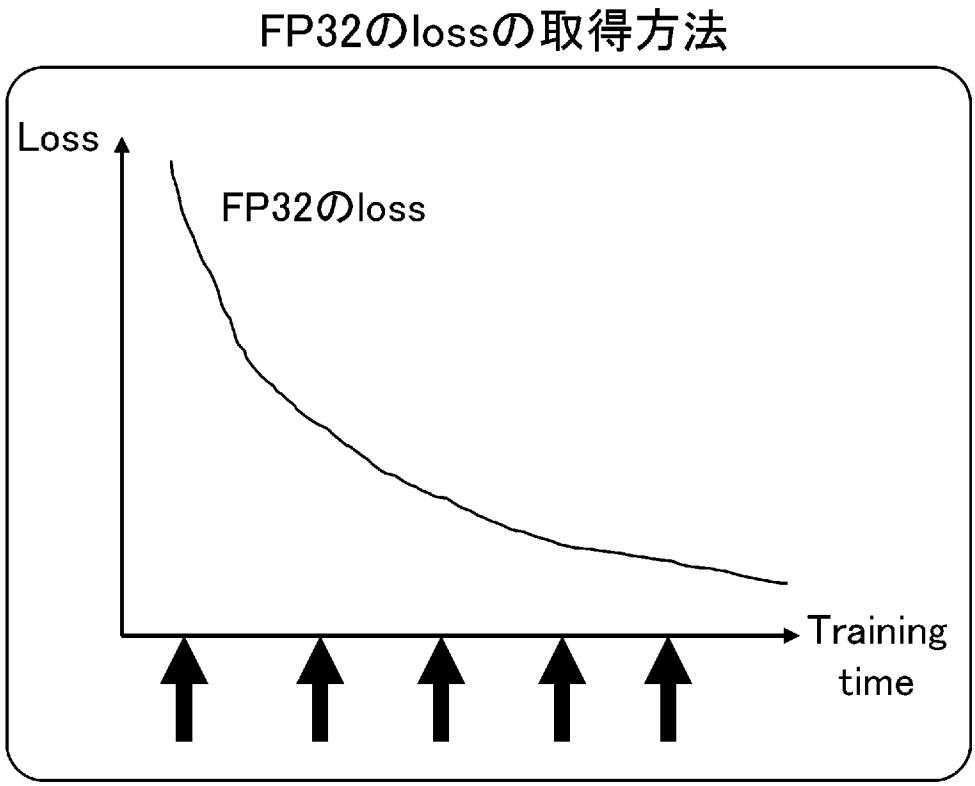
[図2]


1
100
深層学習処理部
101
量子化処理部
102
量子化誤算出部
103
閾値設定部
104
量子化実行部
105
認識率比較部

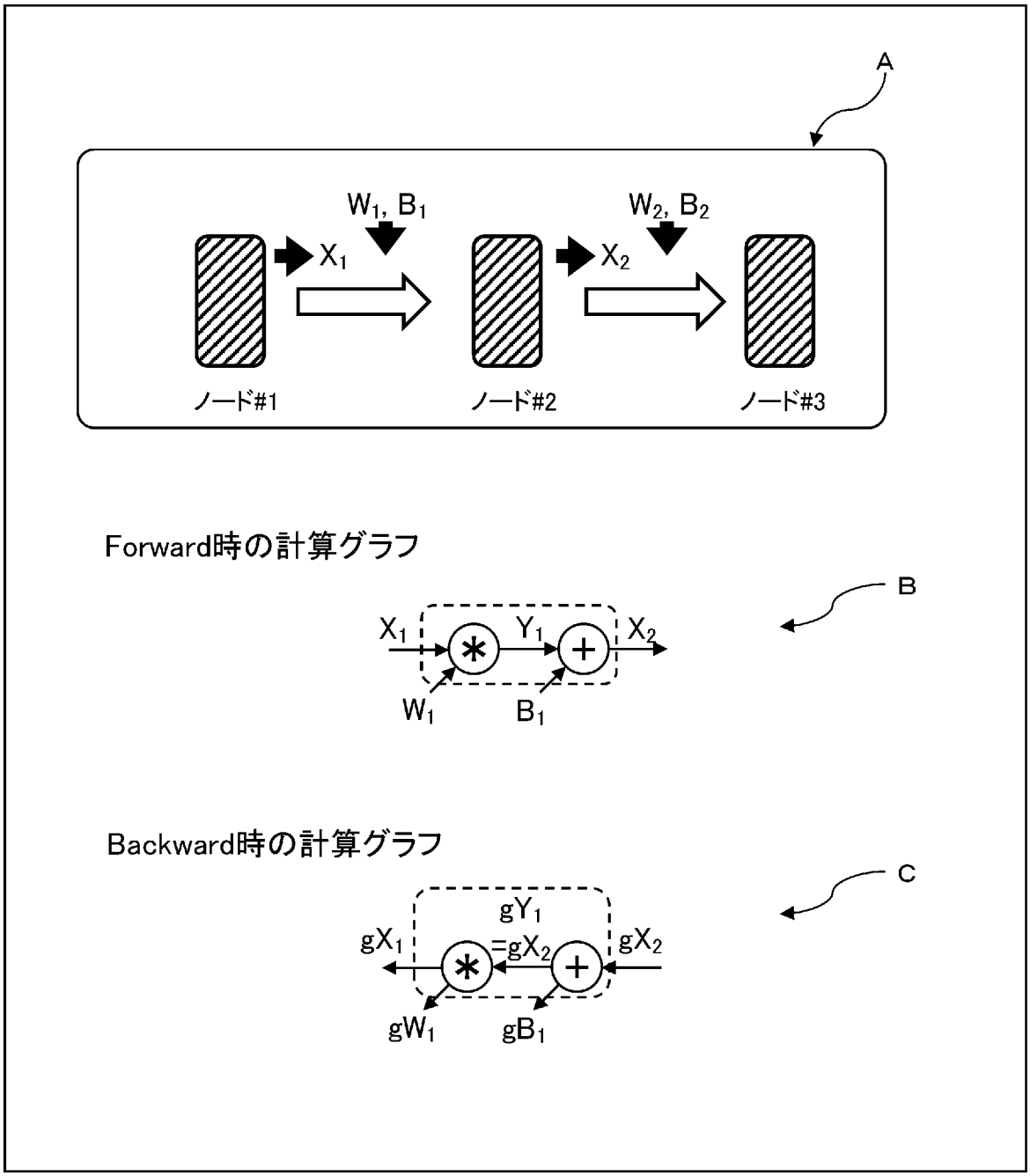
[図3]



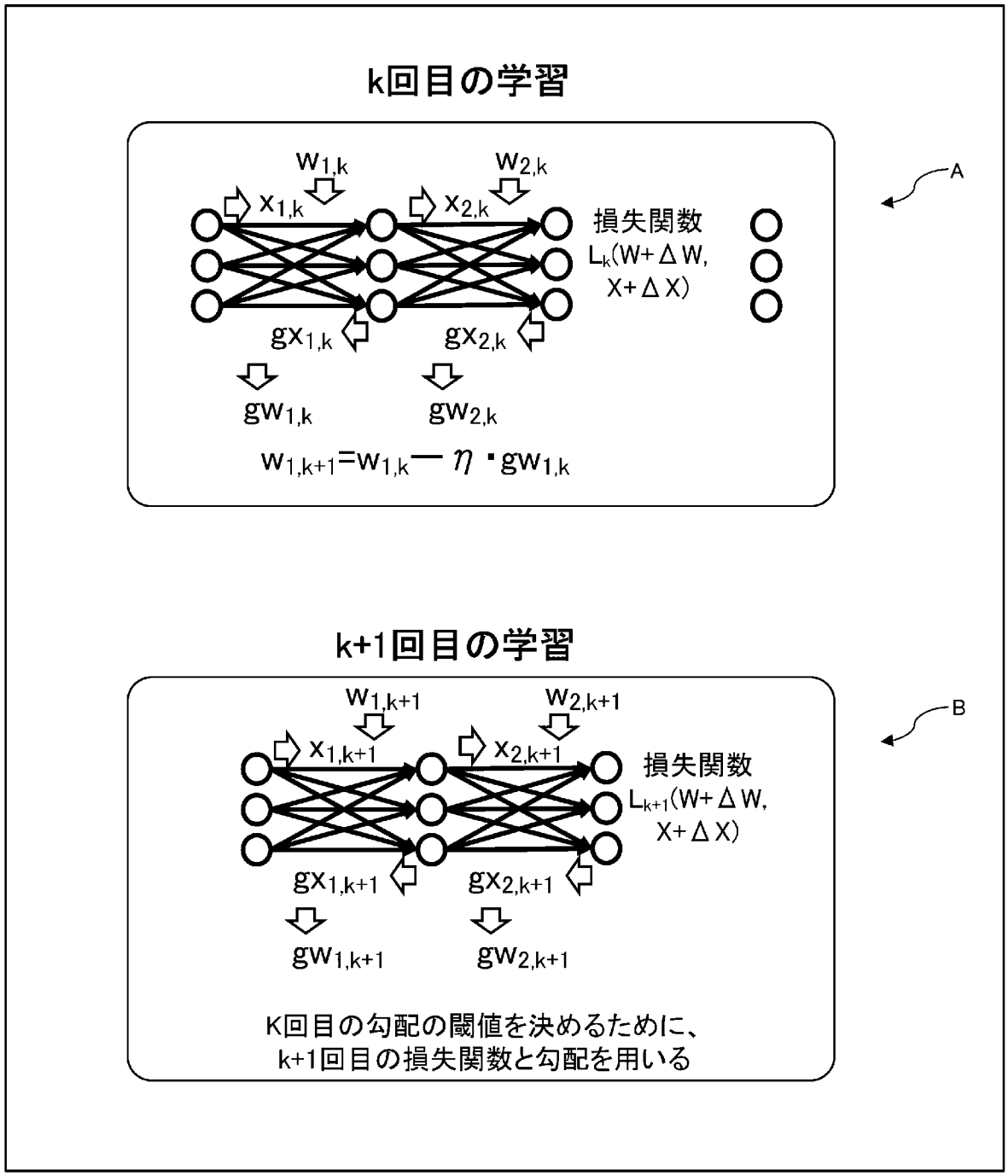
[図4]



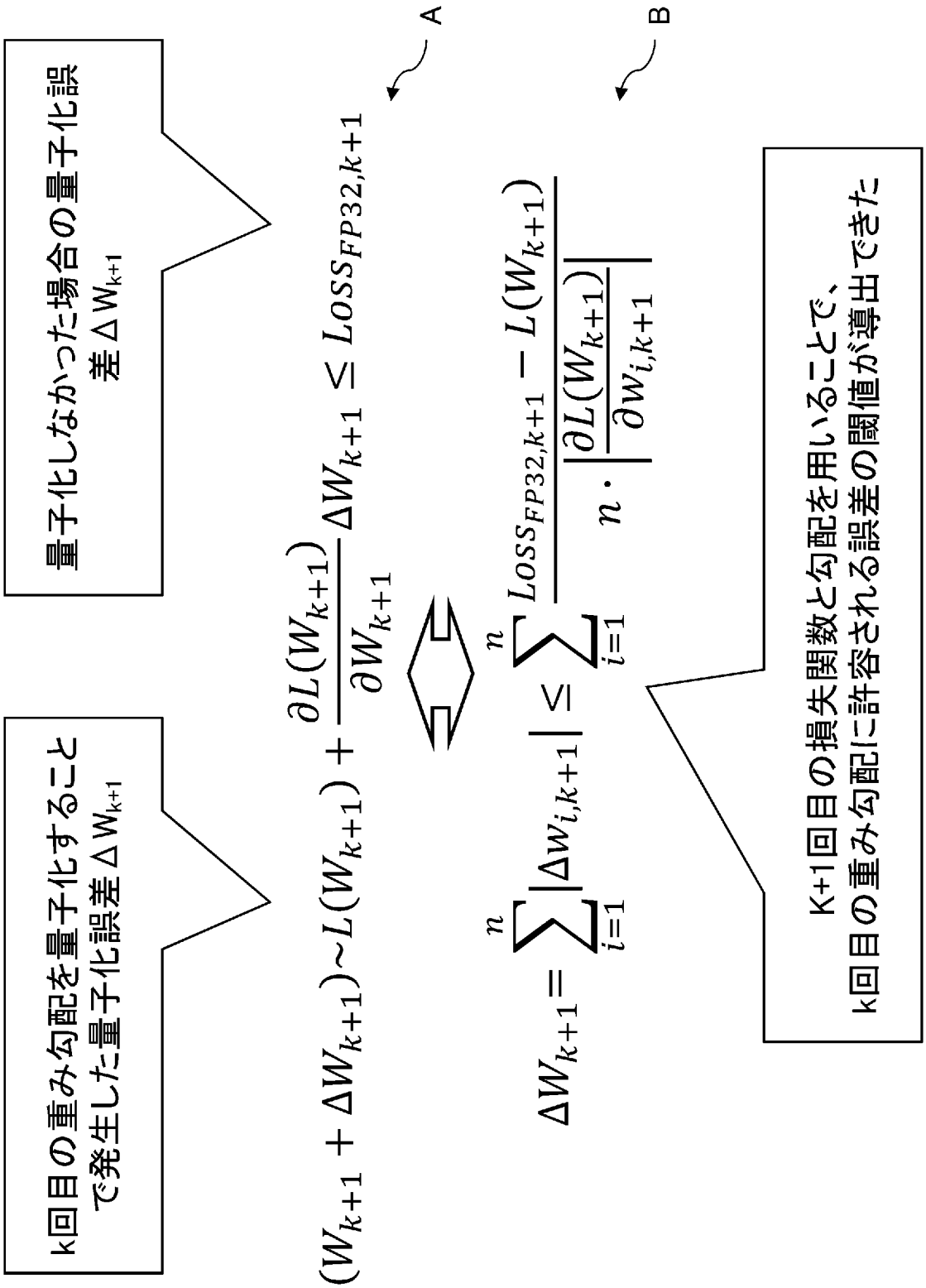
[図5]


A
W1, B1
W2, B2
X1
X2
ノード#1
ノード#2
ノード#3
Forward時の計算グラフ
B
X1
Y1
X2
W1
B1
Backward時の計算グラフ
C
gY1
gX1
=gX2
gX2
gW1
gB1

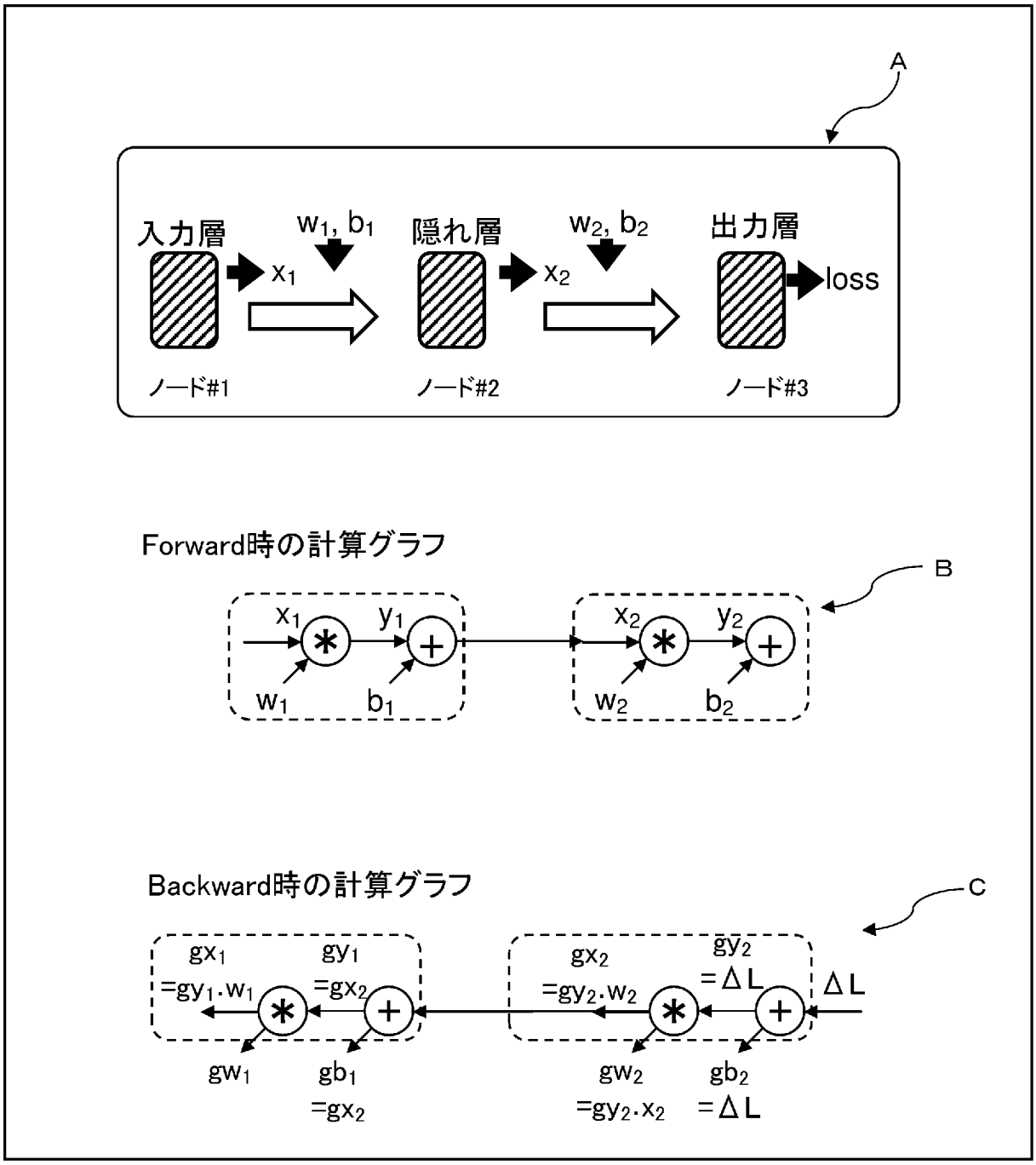
[図6]


k回目の学習
$w_{1,k}$
$w_{2,k}$
$x_{1,k}$
$x_{2,k}$
損失関数
$L_k(W+\Delta W, X+\Delta X)$
$gx_{1,k}$
$gx_{2,k}$
$gw_{1,k}$
$gw_{2,k}$
$w_{1,k+1}=w_{1,k}-\eta \cdot gw_{1,k}$
A
k+1回目の学習
$w_{1,k+1}$
$w_{2,k+1}$
$x_{1,k+1}$
$x_{2,k+1}$
損失関数
$L_{k+1}(W+\Delta W, X+\Delta X)$
$gx_{1,k+1}$
$gx_{2,k+1}$
$gw_{1,k+1}$
$gw_{2,k+1}$
K回目の勾配の閾値を決めるために、
k+1回目の損失関数と勾配を用いる
B

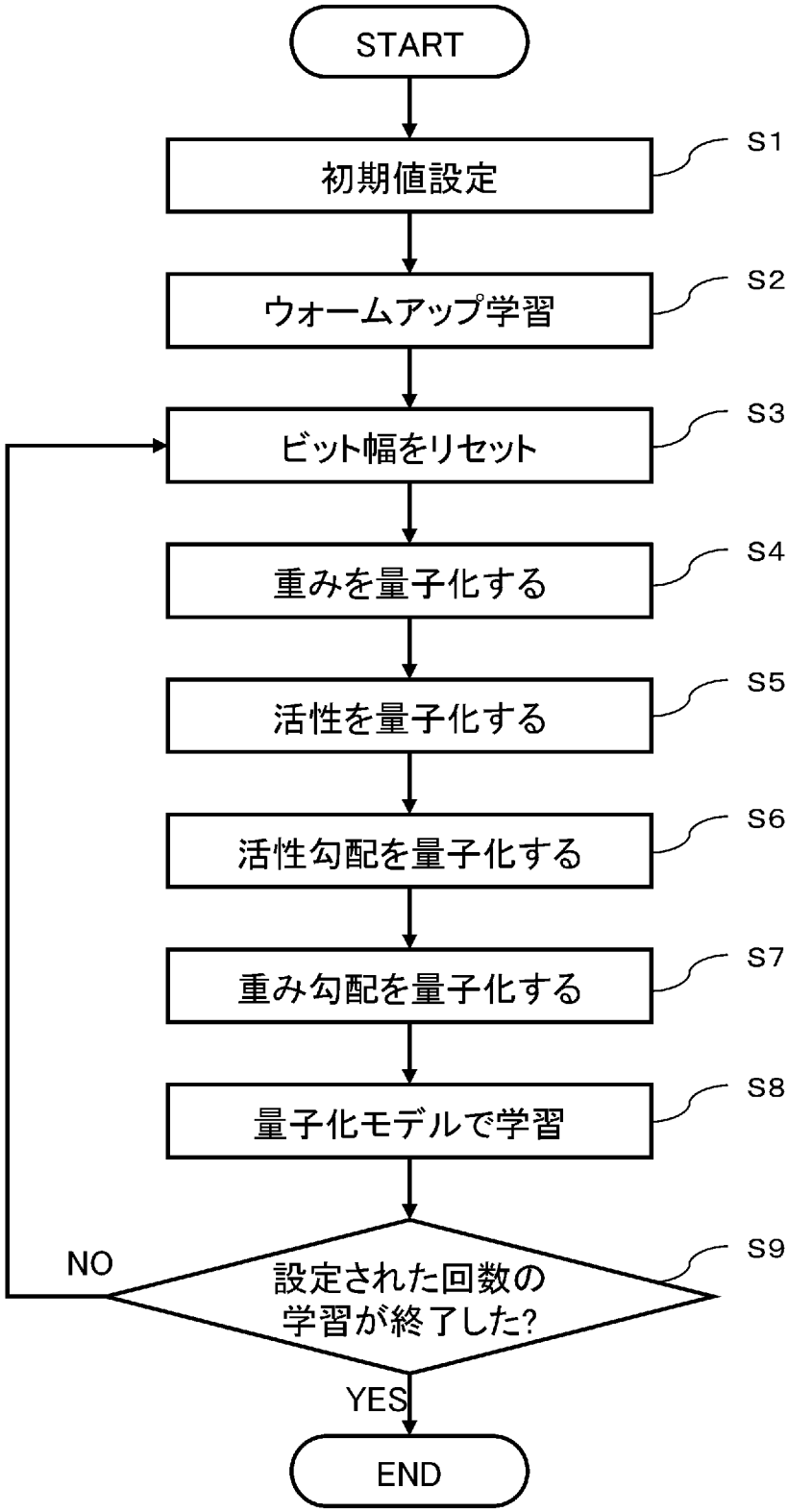
[図7]


量子化しなかった場合の量子化誤差 ΔW_{k+1}
k回目の重み勾配を量子化することで発生した量子化誤差 ΔW_{k+1}
A
B
$L(W_{k+1} + \Delta W_{k+1}) \sim L(W_{k+1}) + \frac{\partial L(W_{k+1})}{\partial W_{k+1}} \Delta W_{k+1} \leq Loss_{FP32,k+1}$
$\Delta W_{k+1} = \sum_{i=1}^{n} |\Delta w_{i,k+1}| \leq \sum_{i=1}^{n} \frac{Loss_{FP32,k+1} - L(W_{k+1})}{n \cdot \left| \frac{\partial L(W_{k+1})}{\partial w_{i,k+1}} \right|}$
K+1回目の損失関数と勾配を用いることで、
k回目の重み勾配に許容される誤差の閾値が導出できた

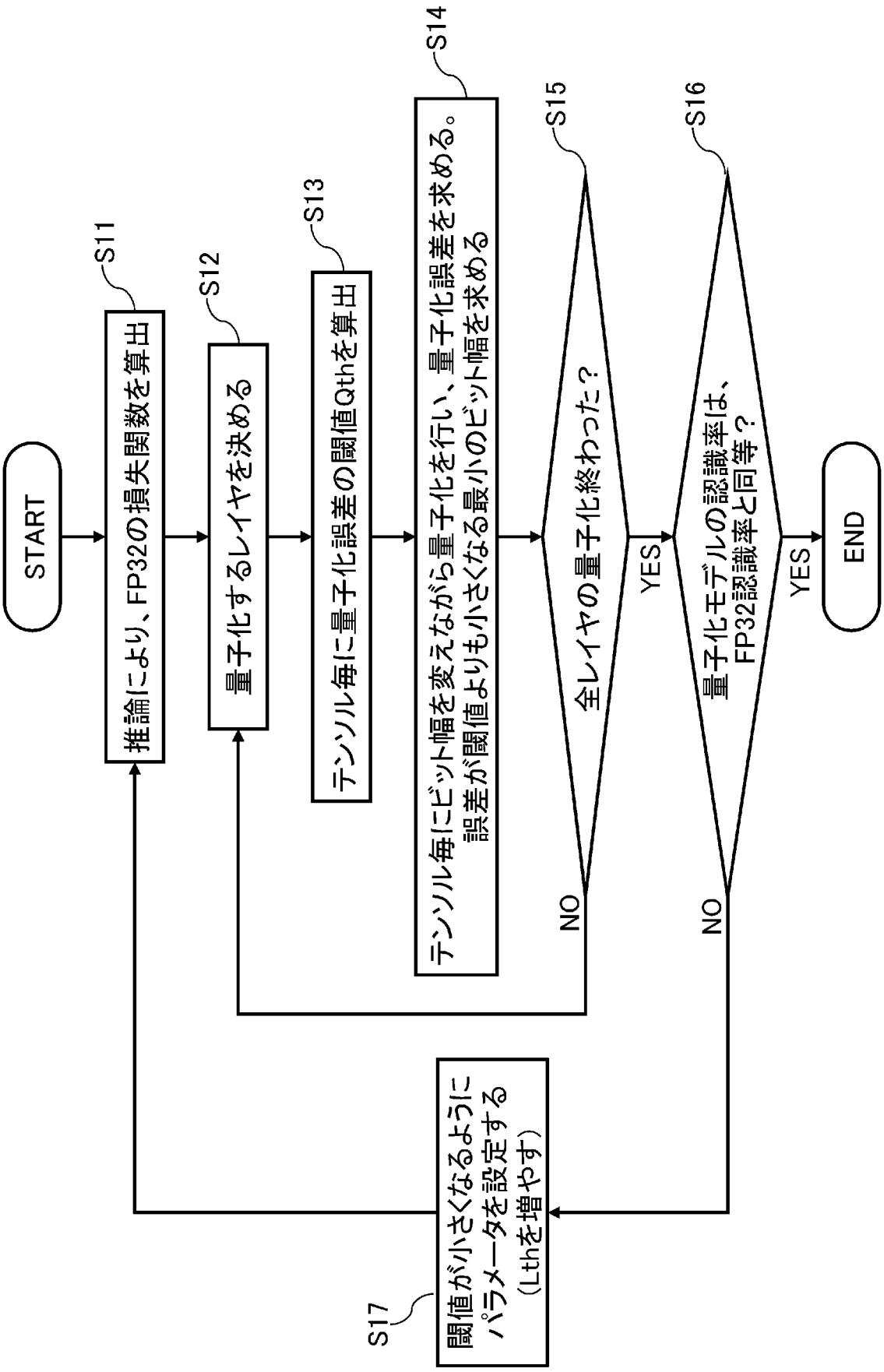
[図8]


A
入力層
w_1, b_1
隠れ層
w_2, b_2
出力層
x_1
x_2
loss
ノード#1
ノード#2
ノード#3
Forward時の計算グラフ
B
x_1
y_1
w_1
b_1
x_2
y_2
w_2
b_2
Backward時の計算グラフ
C
gx_1
$=gy_1 \cdot w_1$
gy_1
$=gx_2$
gw_1
gb_1
$=gx_2$
gx_2
$=gy_2 \cdot w_2$
gy_2
$=\Delta L$
ΔL
gw_2
$=gy_2 \cdot x_2$
gb_2
$=\Delta L$

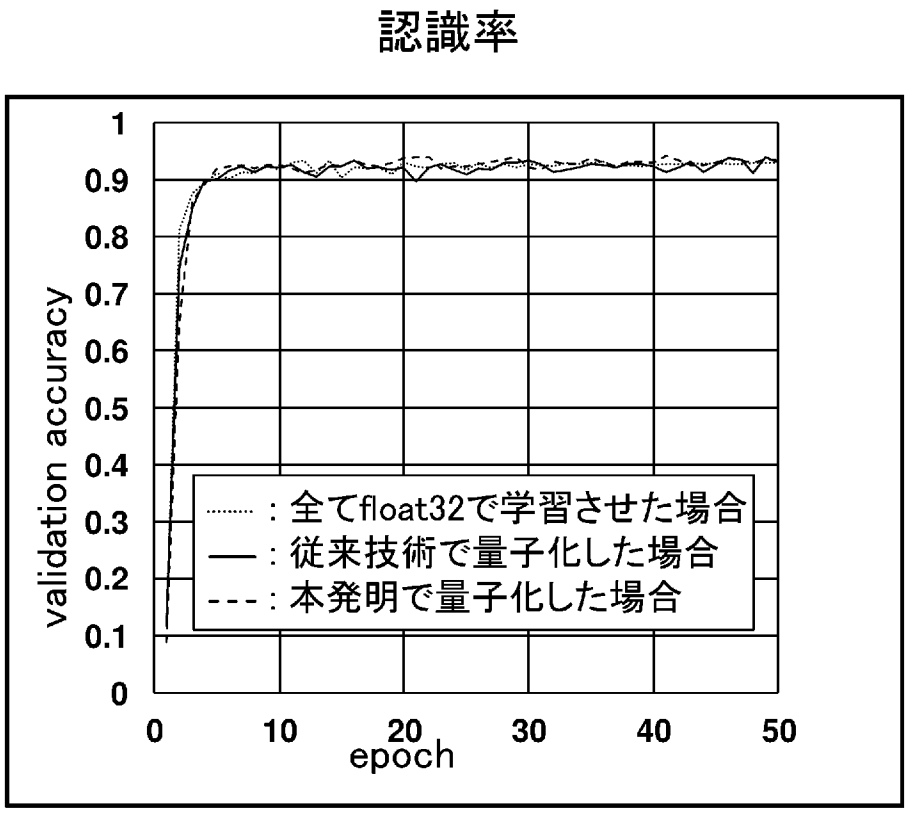
[図9]


START
初期値設定
S1
ウォームアップ学習
S2
ビット幅をリセット
S3
重みを量子化する
S4
活性を量子化する
S5
活性勾配を量子化する
S6
重み勾配を量子化する
S7
量子化モデルで学習
S8
設定された回数の
学習が終了した?
S9
NO
YES
END

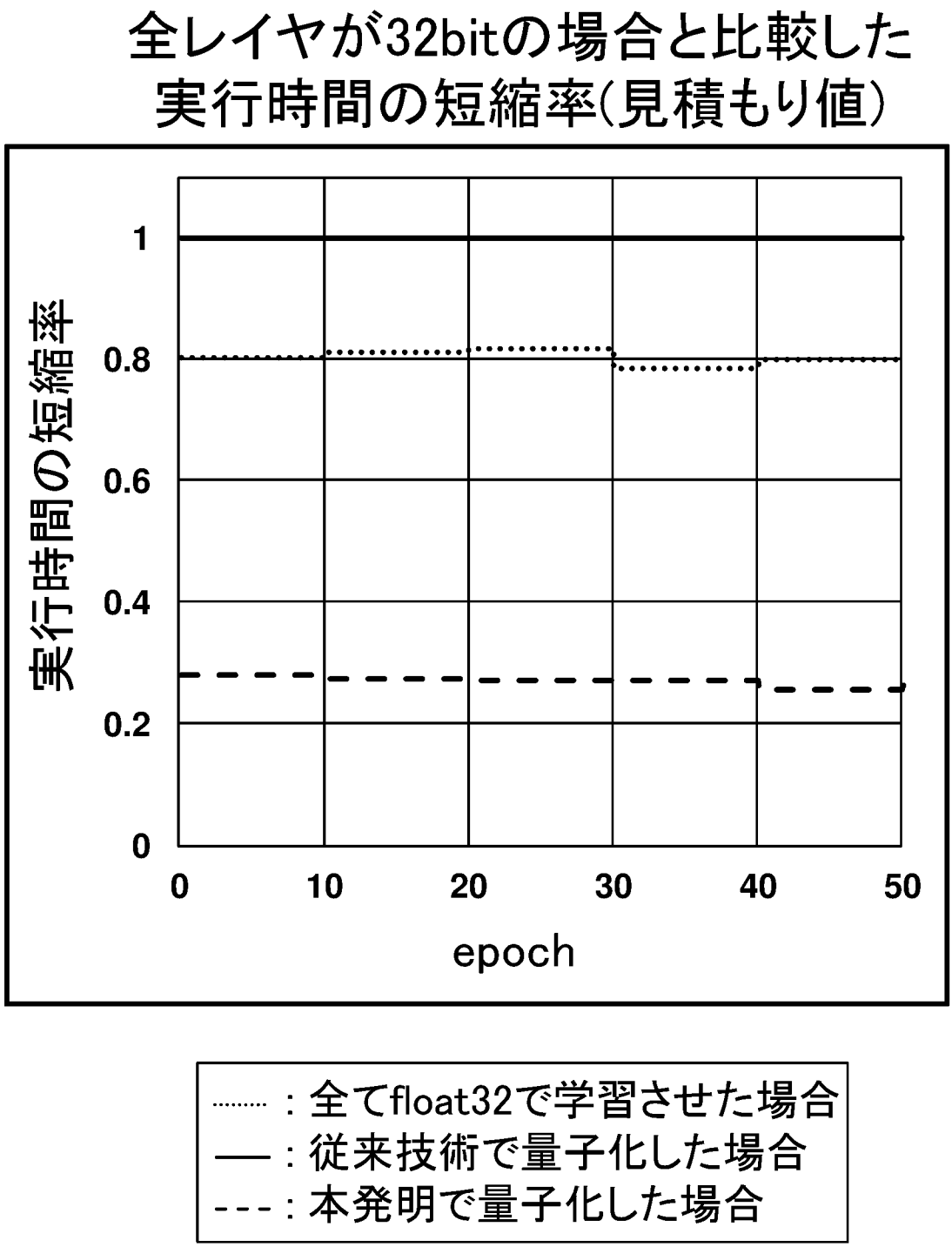
[図10]


START
S11
推論により、FP32の損失関数を算出
S12
量子化するレイヤを決める
S13
テンソル毎に量子化誤差の閾値Qthを算出
S14
テンソル毎にビット幅を変えながら量子化を行い、量子化誤差を求める。
誤差が閾値よりも小さくなる最小のビット幅を求める
S15
全レイヤの量子化終わった？
NO
YES
S16
量子化モデルの認識率は、
FP32認識率と同等？
NO
YES
END
S17
閾値が小さくなるように
パラメータを設定する
（Lthを増やす）

[図11]


認識率
validation accuracy
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
10
20
30
40
50
epoch
：全てfloat32で学習させた場合
：従来技術で量子化した場合
：本発明で量子化した場合
量子化されたNNの認識率は、
FP32の結果と同等

[図12]


全レイヤが32bitの場合と比較した
実行時間の短縮率(見積もり値)
実行時間の短縮率
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
epoch
……：全てfloat32で学習させた場合
——：従来技術で量子化した場合
- - -：本発明で量子化した場合

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2020/027134

A. CLASSIFICATION OF SUBJECT MATTER
Int. Cl. G06F3/06(2006.01)i
FI: G06F3/06

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)
Int. Cl. G06F3/06

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Published examined utility model applications of Japan 1922-1996
Published unexamined utility model applications of Japan 1971-2020
Registered utility model specifications of Japan 1996-2020
Published registered utility model applications of Japan 1994-2020

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X Y	廣瀬一俊、植松瞭太、安藤洗太、折茂健太郎、植吉晃大、高前田伸也、池辺将之、浅井哲也、本村真人，対数量子化による深層ニューラルネットワークのメモリ量削減，電子情報通信学会技術研究報告，15 May 2017, vol. 117, no. 44, pp. 39-44 (CPSY2017-8) ISSN 0913-5685 要約 non-official translation (HIROSE, Kazutoshi, UEMATSU, UEMATSU, Ryota, ANDO, Kota, ORIMO, Kentaro, UEYOSHI, Kodai, TAKAMAEDA, Shinya, IKEBE, Masayuki, ASAI, Tetsuya, Motomura, Masato. Reducing Memory in Deep Neural Networks by Using Quantization. IEICE Technical Report. Abstract.)	1, 8, 15 2-5, 9-12, 16-19
Y	WO 2018/140294 A1 (MICROSOFT TECHNOLOGY LICENSING, LLC) 02 August 2018, paragraphs [0062], [0063], [0065], [0081]	2-5, 9-12, 16-19

☒ 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
07.08.2020	18.08.2020
Name and mailing address of the ISA/ Japan Patent Office 3-4-3, Kasumigaseki, Chiyoda-ku, Tokyo 100-8915, Japan	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2020/027134

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	門倉巧, 中野秀洋, 宮内新, Quantized Neural Networksにおける量子化関数に関する検討, 電子情報通信学会技術研究報告, 22 June 2017, vol. 117, no. 112, pp. 7-10 (CCS2017-2) ISSN 0913-5685, 要約及び「3. Quantized Neural Network(QNN)[9]」の章.	6, 20
Y	(KADOKURA, Takumi, NAKANO, Hidehiro, MIYAUCHI, Arata. Consideration on Functions for Quantization in Quantized Neural Networks. IEICE Technical Report. Abstract and Chapter "3. Quantized Neural Network(QNN)[9]".	4-5, 7, 11, 14, 18-19, 21
Y	JP 2019-079531 A (SAMSUNG ELECTRONICS CO., LTD.) 23 May 2019, paragraphs [0015]-[0017]	5, 7, 12, 14, 19, 21
Y	US 2019/0227893 A1 (IMAGINATION TECHNOLOGIES LIMITED) 25 July 2019, paragraphs [0062]-[0087]	13-14
A	JP 07-084975 A (HITACHI, LTD.) 31 March 1995, abstract	1-21

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/JP2020/027134

Patent Documents referred to in the Report	Publication Date	Patent Family	Publication Date
WO 2018/140294 A1	02.08.2018	CN 108345939 A paragraphs [0079], [0080], [0082], [0101]	
JP 2019-079531 A	23.05.2019	US 2019/0122100 A1 paragraphs [0041], [0043], [0047] EP 3474194 A1 KR 10-2019-0043849 A CN 109685198 A	
US 2019/0227893 A1	25.07.2019	EP 3480690 A1 paragraphs [0048]-[0074] GB 2568084 A CN 110009099 A	
JP 07-084975 A	31.03.1995	(Family: none)	

国際調査報告

国際出願番号
PCT/JP2020/027134

A. 発明の属する分野の分類（国際特許分類（ＩＰＣ））

G06F 3/06(2006.01)i
FI: G06F3/06

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））

G06F3/06

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	１９７１－２０２０年
日本国実用新案登録公報	１９９６－２０２０年
日本国登録実用新案公報	１９９４－２０２０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）

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	廣瀬一俊、植松瞭太、安藤洸太、折茂健太郎、植吉晃大、高前田伸也、池辺将之、浅井哲也、本村真人，対数量子化による深層ニューラルネットワークのメモリ量削減，電子情報通信学会技術研究報告，2017.05.15，Ｖｏｌ．１１７，Ｎｏ．４４，ｐｐ．３９－４４（CPSY2017-8），ISSN 0913-5685 要約	1, 8, 15
Y		2-5, 9-12, 16-19
Y	WO 2018/140294 A1 (MICROSOFT TECHNOLOGY LICENSING, LLC) 02.08.2018 (2018-08-02) 段落[0062]-[0063], [0065]及び[0081]	2-5, 9-12, 16-19
X	門倉 巧、中野秀洋、宮内 新，Quantized Neural Networksにおける量子化関数に関する検討，電子情報通信学会技術研究報告，2017.06.22，Ｖｏｌ．１１７，Ｎｏ．１１２，ｐｐ．７－１０（CCS2017-2），ISSN 0913-5685 要約及び「3. Quantized Neural Network (QNN) [9]」の章	6, 20
Y		4-5, 7, 11-14, 18-19, 21
Y	JP 2019-079531 A (三星電子株式会社) 23.05.2019 (2019-05-23) 段落[0015]-[0017]	5, 7, 12, 14, 19, 21

☑ 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抵触するものではなく、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
07.08.2020	18.08.2020

名称及びあて先
日本国特許庁(ISA/JP)
〒100-8915
日本国
東京都千代田区霞が関三丁目４番３号

権限のある職員（特許庁審査官）
坂庭　剛史 5B 9288
電話番号 03-3581-1101 内線 3545

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
Y	US 2019/0227893 A1 (IMAGINATION TECHNOLOGIES LIMITED) 25.07.2019 (2019-07-25) 段落[0062]-[0087]	13-14
A	JP 07-084975 A (株式会社日立製作所) 31.03.1995 (1995-03-31) 要約	1-21

国際調査報告
パテントファミリーに関する情報

国際出願番号
PCT/JP2020/027134

引用文献	公表日	パテントファミリー文献	公表日
WO 2018/140294 A1	02.08.2018	CN 108345939 A 段落[0079]-[0080], [0082]及び[0101]	
JP 2019-079531 A	23.05.2019	US 2019/0122100 A1 段落[0041], [0043]及び[0047] EP 3474194 A1 KR 10-2019-0043849 A CN 109685198 A	
US 2019/0227893 A1	25.07.2019	EP 3480690 A1 段落[0048]-[0074] GB 2568084 A CN 110009099 A	
JP 07-084975 A	31.03.1995	（ファミリーなし）	