(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7581370号
(P7581370)

(45)発行日　令和6年11月12日(2024.11.12)

(24)登録日　令和6年11月1日(2024.11.1)

(51)国際特許分類
Ｇ０６Ｎ　3/0495(2023.01)

ＦＩ
Ｇ０６Ｎ　3/0495

請求項の数　62　（全59頁）

(21)出願番号　特願2022-562943(P2022-562943)
(86)(22)出願日　令和3年4月13日(2021.4.13)
(65)公表番号　特表2023-522886(P2023-522886A)
(43)公表日　令和5年6月1日(2023.6.1)
(86)国際出願番号　PCT/EP2021/059592
(87)国際公開番号　WO2021/209469
(87)国際公開日　令和3年10月21日(2021.10.21)
審査請求日　令和4年12月14日(2022.12.14)
(31)優先権主張番号　20169502.0
(32)優先日　令和2年4月14日(2020.4.14)
(33)優先権主張国・地域又は機関
欧州特許庁(EP)

(73)特許権者　504315750
フラウンホーファー・ゲゼルシャフト ツア フェルデルング デア アンゲヴァンテン フォルシュング アインゲトラーゲナー フェライン
ドイツ連邦共和国 ８０６８６ ミュンヘン ハンザシュトラーセ ２７ツェー
(74)代理人　100133411
弁理士　山本 龍郎
(74)代理人　100067677
弁理士　山本 彰司
(72)発明者　ヴィーデマン，ジモン
ドイツ連邦共和国 ベルリン １０５８７ アインシュタインウーファー３７ ハーハーイー，ハインリッヒ－ヘルツ－イン

最終頁に続く

(54)【発明の名称】　ニューラルネットワークパラメーターの表現の改良された概念

(57)【特許請求の範囲】

【請求項１】

ＮＮ表現（１１０）を生成する装置（１００）であって、前記ＮＮ表現（１１０）は、ＮＮパラメーター（１３０）を表すために、量子化パラメーター（１４２）及び量子化値（１５２）を含み、前記装置（１００）は、

量子化パラメーター（１４２）から、

前記量子化パラメーター（１４２）によって導出された被除数と精度パラメーター（１４５）によって導出された除数との間の除算の剰余に基づく乗数（１４４）と、

前記除算の商の丸めに基づくビットシフト数（１４６）と、

が導出可能であるように、ＮＮパラメーター（１２０）の前記量子化パラメーター（１４２）及び量子化値（１５２）を決定（１４０）することによって、前記ＮＮパラメーター（１２０）を量子化された値（１３０）に量子化するように構成され、

それにより、前記ＮＮパラメーター（１２０）の前記量子化された値（１３０）は、前記量子化値（１５２）と前記乗数（１４４）に依存する因数（１４８）との間の積であって、前記ビットシフト数（１４６）に依存するビット数だけビットシフトされた積に対応する、装置。

【請求項２】

ＮＮ表現（１１０）からＮＮパラメーターを導出する装置であって、

前記ＮＮ表現（１１０）から量子化パラメーター（１４２）を導出することと、

前記ＮＮ表現（１１０）から量子化値（１５２）を導出することと、

前記量子化パラメーター（１４２）から、
前記量子化パラメーター（１４２）によって導出された被除数と精度パラメーター（１４５）によって導出された除数との間の除算の剰余に基づく乗数（１４４）と、
前記除算の商の丸めに基づくビットシフト数（１４６）と、
を導出することと、
を行うように構成され、
前記ＮＮパラメーター（１３０）は、前記量子化値（１５２）と前記乗数（１４４）に依存する因数（１４８）との間の積であって、前記ビットシフト数（１４６）に依存するビット数だけビットシフトされた積に対応する、装置。
【請求項３】
前記ＮＮ表現（１１０）から前記精度パラメーター（１４５）を導出するように更に構成されている、請求項２に記載の装置。
【請求項４】
前記ＮＮパラメーター（１３０）は、
一対のニューロン（１０）間のニューロン間活性化フィードフォワード（１２）に重み付けする重みパラメーター、
ニューラルネットワーク層のアフィン変換をパラメーター化するバッチノルムパラメーター、及び
所定のニューラルネットワークニューロン（１０）のインバウンドニューロン間活性化フィードフォワード（１２）の和にバイアスをかけるバイアス、
のうちの１つである、請求項２又は３に記載の装置。
【請求項５】
前記ＮＮパラメーター（１３０）は、ＮＮの複数のニューロン間活性化フィードフォワード（１２２）のうちの単一のニューロン間活性化フィードフォワード（１２）に関して前記ＮＮをパラメーター化し、前記装置は、前記複数のニューロン間活性化フィードフォワード（１２２）の各々について、前記ＮＮ表現（１１０）から、対応するＮＮパラメーター（１３０）を、
前記複数のニューロン間活性化フィードフォワード（１２２）の各々（１２）について、
前記ＮＮ表現（１１０）から前記それぞれのニューロン間活性化フィードフォワード（１２）に関連付けられた関連量子化パラメーター（１４２）を導出することと、
前記ＮＮ表現（１１０）から前記それぞれのニューロン間活性化フィードフォワード（１２）に関連付けられた関連量子化値（１５２）を導出することと、
前記関連量子化パラメーター（１４２）から、
前記関連量子化パラメーター（１４２）によって導出された被除数と、前記それぞれのニューロン間活性化フィードフォワード（１２）に関連付けられた関連精度パラメーター（１４５）によって導出された除数との間の除算の剰余に基づいて、前記それぞれのニューロン間活性化フィードフォワード（１２）に関連付けられた関連乗数（１４４）と、
前記除算の商の丸めに基づいて、前記それぞれのニューロン間活性化フィードフォワード（１２）に関連付けられた関連ビットシフト数（１４６）と、を導出することと、
によって導出するように構成され、
前記それぞれのニューロン間活性化フィードフォワード（１２）の前記対応するＮＮパラメーター（１３０）は、前記関連量子化値（１５２）と前記関連乗数（１４４）に依存する因数（１４８）との間の積であって、前記関連ビットシフト数（１４６）に依存するビット数だけビットシフトされた積に対応する、請求項２～４のいずれか一項に記載の装置。
【請求項６】
前記装置は、ＮＮの複数のニューロン間活性化フィードフォワード（１２２）をニューロン間活性化フィードフォワードのサブグループ（１２２ａ、１２２ｂ）に細分するように構成され、それにより、各サブグループ（１２２ａ、１２２ｂ）が、前記ＮＮの関連する一対のＮＮ層に関連付けられ、前記関連する一対のＮＮ層の間のニューロン間活性化フィ

ードフォワードを含み、前記関連する一対の層以外の更なる一対のＮＮ層の間のニューロン間活性化フィードフォワードを除外し、２つ以上のサブグループ（１２２ａ、１２２ｂ）が、所定のＮＮ層に関連付けられ、
前記ＮＮパラメーター（１３０）は、前記ＮＮの前記複数のニューロン間活性化フィードフォワード（１２２）のうちの単一のニューロン間活性化フィードフォワード（１２）に関して前記ＮＮをパラメーター化し、前記装置は、前記複数のニューロン間活性化フィードフォワード（１２２）の各々について、前記ＮＮ表現（１１０）から、対応するＮＮパラメーター（１３０）を、
ニューロン間活性化フィードフォワードの各サブグループ（１２２ａ、１２２ｂ）について、
前記ＮＮ表現（１１０）から前記それぞれのサブグループ（１２２ａ、１２２ｂ）に関連付けられた関連量子化パラメーター（１４２）を導出することと、
前記関連量子化パラメーター（１４２）から、
前記関連量子化パラメーター（１４２）によって導出された被除数と、前記それぞれのサブグループ（１２２ａ，１２２ｂ）に関連付けられた関連精度パラメーター（１４５）によって導出された除数との間の除算の剰余に基づいて、前記それぞれのサブグループ（１２２ａ，１２２ｂ）に関連付けられた関連乗数（１４４）と、
前記除算の商の丸めに基づいて、前記それぞれのサブグループ（１２２ａ、１２２ｂ）に関連付けられた関連ビットシフト数（１４６）と、を導出することと、
前記複数のニューロン間活性化フィードフォワード（１２２）の各々について、
前記ＮＮ表現（１１０）から前記それぞれのニューロン間活性化フィードフォワード（１２）に関連付けられた関連量子化値（１５２）を導出することと、
によって導出するように構成され、
前記それぞれのニューロン間活性化フィードフォワード（１２）の前記対応するＮＮパラメーター（１３０）は、前記それぞれのニューロン間活性化フィードフォワード（１２）が含まれる前記サブグループ（１２２ａ、１２２ｂ）に関連付けられた、前記関連量子化値（１５２）と前記関連乗数（１４４）に依存する因数（１４８）との間の積であって、前記それぞれのニューロン間活性化フィードフォワード（１２）が含まれる前記サブグループ（１２２ａ、１２２ｂ）の前記関連ビットシフト数（１４６）に依存するビット数だけビットシフトされた積に対応する、請求項２～４のいずれか一項に記載の装置。
【請求項７】
前記関連精度パラメーター（１４５）は、前記ＮＮにわたって又は各ＮＮ層内でグローバルに等しく評価される、請求項５又は６に記載の装置。
【請求項８】
前記ＮＮ表現（１１０）から前記関連精度パラメーター（１４５）を導出するように構成されている、請求項５～７のいずれか一項に記載の装置。
【請求項９】
前記ＮＮ表現（１１０）から前記関連量子化パラメーター（１４２）を基準量子化パラメーターとの差分の形態で導出するように構成されている、請求項５～８のいずれか一項に記載の装置。
【請求項１０】
前記量子化パラメーター（１４２）から、前記乗数（１４４）及び前記ビットシフト数（１４６）を、
ｍｕｌ＝ｋ＋ＱＰ％ｋ
【数１】

$$shift = \lfloor QP/k \rfloor$$

に従って導出するように構成され、式中、ｍｕｌは、前記乗数（１４４）であり、ｓｈｉｆｔは、前記ビットシフト数（１４６）であり、ＱＰは、前記量子化パラメーター（１４

２）であり、ｋは、前記精度パラメーター（１４５）であり、
【数２】

⌊ ⌋

は、そのオペランド以下の最大の整数を生成するフロア演算子であり、％は、ｘ％ｙに対して
【数３】

$$x - y \cdot \lfloor x/y \rfloor$$

を生成するモジュロ演算子であり、それにより、前記ＮＮパラメーター（１３０）は、
（ｍｕｌ／ｋ）・２^{ｓｈｉｆｔ}・Ｐ
であり、式中、Ｐは、前記量子化値（１５２）である、請求項２～９のいずれか一項に記載の装置。
【請求項１１】
前記精度パラメーター（１４５）は、２の累乗である、請求項２～１０のいずれか一項に記載の装置。
【請求項１２】
コンテキスト適応型二値算術復号化の使用によって、又は
前記ＮＮ表現（１１０）から前記量子化パラメーター（１４２）を表すビットを直接読み出すことによって、又は
前記装置のコンテキスト適応型二値デコーダーの等確率バイパスモードを介して、前記ＮＮ表現（１１０）から前記量子化パラメーター（１４２）を表すビットを導出することによって、
前記ＮＮ表現（１１０）から前記量子化パラメーター（１４２）を導出するように構成されている、請求項２～１１のいずれか一項に記載の装置。
【請求項１３】
二値化方式を使用してビンストリングを非二値化することによって、前記ＮＮ表現（１１０）から前記量子化パラメーター（１４２）を導出するように構成されている、請求項２～１２のいずれか一項に記載の装置。
【請求項１４】
前記二値化方式は、指数ゴロム符号である、請求項１３に記載の装置。
【請求項１５】
固定小数点表現の形態で前記ＮＮ表現（１１０）から前記量子化パラメーター（１４２）を導出するように構成されている、請求項２～１４のいずれか一項に記載の装置。
【請求項１６】
前記精度パラメーター（１４５）は、２^{ｔ}であり、前記固定小数点表現のビット長は、前記ＮＮに対して一定になるように設定されるか、又は前記ＮＮに対して一定である基底ビット長とｔとの和になるように設定される、請求項１５に記載の装置。
【請求項１７】
整数値シンタックス要素として前記ＮＮ表現（１１０）から前記量子化パラメーター（１４２）を導出するように構成されている、請求項２～１６のいずれか一項に記載の装置。
【請求項１８】
前記ＮＮ表現（１１０）から前記精度パラメーター（１４５）を表すビットを直接読み出すことによって、又は前記装置のコンテキスト適応型二値デコーダーの等確率バイパスモードを介して前記ＮＮ表現（１１０）から前記精度パラメーター（１４５）を表すビットを導出することによって、前記ＮＮ表現（１１０）から前記精度パラメーター（１４５）を導出するように構成されている、請求項２～１７のいずれか一項に記載の装置。

【請求項１９】
固定小数点表現の形態で前記ＮＮ表現（１１０）から前記量子化値（１５２）を導出するように構成されている、請求項２～１８のいずれか一項に記載の装置。
【請求項２０】
二値化方式に従ってビンストリングから前記量子化値（１５２）を非二値化し、コンテキスト適応算術復号化を使用して前記ＮＮ表現（１１０）から前記ビンストリングのビットを復号化することによって、前記ＮＮ表現（１１０）から前記量子化値（１５２）を導出するように構成されている、請求項２～１９のいずれか一項に記載の装置。
【請求項２１】
二値化方式に従ってビンストリングから前記量子化値（１５２）を非二値化し、コンテキスト適応算術復号化を使用して前記ＮＮ表現（１１０）から前記ビンストリングの第１のビットを復号化し、等確率バイパスモードを使用して前記ビンストリングの第２のビットを復号化することによって、前記ＮＮ表現（１１０）から前記量子化値（１５２）を導出するように構成されている、請求項２～２０のいずれか一項に記載の装置。
【請求項２２】
ＮＮ（２０）を使用して推論を実行するデバイス（４００）であって、前記デバイス（４００）は、
前記ＮＮ（２０）をパラメーター化するように構成されたＮＮパラメトライザー（４１０）であって、前記ＮＮパラメトライザー（４１０）は、請求項２～２１のいずれか一項に記載のＮＮ表現（１１０）からＮＮパラメーター（１３０）を導出する装置（３００）を備える、ＮＮパラメトライザー（４１０）と、
前記ＮＮ（２０）を使用してＮＮ入力（４４０）に基づいて推論出力（４３０）を計算するように構成された計算ユニット（４２０）と、
を備える、デバイス。
【請求項２３】
請求項２２に記載のデバイス（４００）であって、
前記ＮＮパラメトライザー（４１０）は、
前記装置（３００）を介して、第１のＮＮパラメーター（１３０）及び第２のＮＮパラメーター（１３０）のうちの少なくとも１つを導出することであって、それにより、前記第１のＮＮパラメーター（１３０）は、第１の量子化値（１５２）と第１の因数（１４８）との積であって、第１のビット数（１４６）だけビットシフトされた積に対応し、前記第２のＮＮパラメーター（１３０）は、第２の量子化値（１５２）と第２の因数（１４８）との積であって、第２のビット数（１４６）だけビットシフトされた積に対応することと、
前記第１のＮＮパラメーター（１３０）の第１の量子化値（１５２）によって形成され、前記第１の乗数（１４４）で重み付けされた第１の加数と、前記第１のＮＮパラメーター（１３０）の第２の量子化値（１５２）によって形成され、前記第２の乗数（１４４）で重み付けされ、前記第１のビット数（１４６）及び前記第２のビット数（１４６）の差だけビットシフトされた第２の加数との間の和を形成することと、前記第１の加数及び前記第２の加数の前記和を、前記第１のビット数（１４６）及び前記第２のビット数（１４６）のうちの１つに依存するビット数だけビットシフトすることと、によって、前記第１のＮＮパラメーター（１３０）及び前記第２のＮＮパラメーター（１３０）に加算を施して、前記ＮＮ（２０）の最終ＮＮパラメーター（１３０）を生成することと、
を行うように構成されている、デバイス。
【請求項２４】
請求項２３に記載のデバイス（４００）であって、
前記第１のＮＮパラメーター（１３０）は、前記ＮＮ（２０）のベース層表現を表し、前記第２のＮＮパラメーター（１３０）は、前記ＮＮのエンハンスメント層表現を表すか、又は
前記第１のＮＮパラメーター（１３０）は、前記ＮＮ（２０）の現在の表現を表し、前

記第２のＮＮパラメーター（１３０）は、前記現在のＮＮ表現（１１０）の更新を表すか、又は
前記第１のＮＮパラメーター（１３０）は、所定のニューラルネットワークニューロン（１０）のインバウンドニューロン間活性化フィードフォワード（１２）の和にバイアスをかけるバイアスを表し、前記第２のＮＮパラメーター（１３０）は、ニューラルネットワーク層のアフィン変換をパラメーター化するバッチノルムパラメーターを表す、デバイス。
【請求項２５】
請求項２２～２４のいずれか一項に記載のデバイス（４００）であって、
前記ＮＮパラメトライザー（４１０）は、
前記装置（３００）を介して、第３のＮＮパラメーター（１３０）及び第４のＮＮパラメーター（１３０）のうちの少なくとも１つを導出するように構成され、それにより、前記第３のＮＮパラメーター（１３０）は、第３の量子化値（１５２）と第３の因数（１４８）との積であって、第３のビット数（１４６）だけビットシフトされた積に対応し、前記第４のＮＮパラメーター（１３０）は、第４の量子化値（１５２）と第４の因数（１４８）との積であって、第４のビット数（１４６）だけビットシフトされた積に対応し、
前記計算ユニット（４２０）は、前記計算を実行する際に、
前記第３のＮＮパラメーター（１３０）及び前記第４のＮＮパラメーター（１３０）に乗算を施して、前記第３のＮＮパラメーター（１３０）の第３の量子化値（１５２）によって形成される第１の因数と、前記第３の乗数（１４４）によって形成される第２の因数と、前記第４のＮＮパラメーター（１３０）の第４の量子化値（１５２）によって形成される第３の因数と、前記第４の乗数（１４４）によって形成される第４の因数との積であって、前記第３のビット数によって形成される第１の加数と前記第４のビット数によって形成される第２の加数とを含む和に対応するビット数だけビットシフトされた積を形成することによって積を生成するように構成される、デバイス。
【請求項２６】
請求項２５に記載のデバイス（４００）であって、
前記第３のＮＮパラメーター（１３０）は、第１のＮＮ層（１１４）の第１のニューロン（１０１）から第２のＮＮ層（１１６）の第２のニューロン（10_2）へのニューロン間活性化フィードフォワード（１２）を重み付けする重みパラメーターを表し、前記第４のＮＮパラメーター（１３０）は、バッチノルムパラメーターを表す、デバイス。
【請求項２７】
請求項２６に記載のデバイス（４００）であって、前記バッチノルムパラメーターは、前記第２のＮＮ層に対して前記第１のニューロン（10_1）の活性化フィードフォワード増幅を調整するものである、デバイス。
【請求項２８】
請求項２２～２７のいずれか一項に記載のデバイス（４００）であって、
活性化について第５の量子化パラメーター（１４２）及び第５の量子化値（１５２）を決定することによって、前記活性化を量子化された値（１３０）に量子化することによって前記ＮＮ入力（４４０）を量子化し、前記第５の量子化パラメーター（１４２）から、前記第５の量子化パラメーター（１４２）によって導出された被除数と前記活性化に関連付けられた精度パラメーター（１４５）によって導出された除数との間の除算の剰余に基づく第５の乗数（１４４）と、前記除算の商の丸めに基づく第５のビットシフト数（１４６）とを導出することにより、前記第５の量子化値（１５２）と前記第５の乗数（１４４）に依存する因数（１４８）との間の積であって、前記第５のビットシフト数（１４６）に依存する第５のビット数だけビットシフトされた積に対応する前記量子化された値（１３０）が得られるように更に構成されている、デバイス。
【請求項２９】
請求項２８に記載のデバイス（４００）であって、
前記ＮＮパラメトライザー（４１０）は、

前記装置（３００）を介して、第６のＮＮパラメーター（１３０）を導出することであって、それにより、前記第６のＮＮパラメーター（１３０）は、第６の量子化値（１５２）と第６の因数（１４８）との積であって、第６のビット数（１４６）だけビットシフトされた積に対応することと、
前記第６のＮＮパラメーター（１３０）及び前記活性化に乗算を施して、前記第６のＮＮパラメーター（１３０）の第６の量子化値（１５２）によって形成される第１の因数と、前記第６の乗数（１４４）によって形成される第２の因数と、前記第５の量子化値（１５２）によって形成される第３の因数と、前記第５の乗数（１４４）によって形成される第４の因数との積であって、前記第６のビット数によって形成される第１の加数と前記第４のビット数（１４６）によって形成される第２の加数と、を含む和に対応するビット数だけビットシフトされた積を形成することによって積を生成することと、
を行うように構成されている、デバイス。
【請求項３０】
ＮＮ（２０）を使用して推論を実行するデバイス（５００）であって、前記デバイス（５００）は、前記ＮＮ（２０）を使用してＮＮ入力（４４０）に基づいて推論出力（４３０）を計算するように構成され、前記ＮＮ（２０）は、一対のＮＮ層と、前記一対のＮＮ層のうちの第１のＮＮ層から前記一対のＮＮ層のうちの第２のＮＮ層へのニューロン間活性化フィードフォワード（１２）とを含み、前記デバイス（５００）は、
前記第１のＮＮ層の前記ニューラルネットワークニューロン（１０）の前記活性化から行列Ｘ（５３２）を形成（５３０）することと、
ｓ・Ｗ’＊Ｘを計算（５４０）することであって、式中、＊は、行列乗算を表し、Ｗ’は、ｎ及びｍ∈Ｎである次元ｎ×ｍの重み行列（５４４）であり、ｓは、長さｎの転置ベクトル（５４６）であり、・は、・の一方の側の行列と・の他方の側の転置ベクトルとの間の列に関するアダマール乗算を示すことと、
によって、前記第１のＮＮ層の前記ニューラルネットワークニューロン（１０）の前記活性化に基づいて、前記第２のＮＮ層の前記ニューラルネットワークニューロン（１０）の活性化を計算するように構成され、
前記デバイス（５００）が、ＮＮ表現（１１０）からＷ’（５４４）を導出するように構成されたＮＮパラメトライザー（４１０）を備え、前記ＮＮパラメトライザー（４１０）は、請求項２～２１のいずれか一項に記載のＮＮ表現（１１０）からＮＮパラメーター（１３０）を導出する装置（３００）を備えている、デバイス。
【請求項３１】
請求項３０に記載のデバイス（５００）であって、ｎビット固定小数点演算を使用して前記行列乗算を計算（５４０）して、内積を生成し、ｍ＞ｎであるｍビット固定小数点演算を使用して前記内積をｓ（５４６）と乗算するように構成されている、デバイス。
【請求項３２】
請求項３０又は３１に記載のデバイス（５００）であって、ｓ（５４６）は、Ｗ’（５４４）を符号化するためのより高い圧縮及び／又はより高い推論忠実度に関するＷ’（５４４）の最適化の結果である、デバイス。
【請求項３３】
請求項３０～３２のいずれか一項に記載のデバイス（５００）であって、前記ＮＮパラメトライザー（４１０）は、Ｗ’（５４４）に関連するＮＮパラメーター（１３０）と比較して異なる量子化パラメーター（１４２）を使用して、前記ＮＮ表現（１１０）からｓ（５４６）を導出するように更に構成されている、デバイス。
【請求項３４】
ＮＮ（２０）のバッチノルム演算子（７１０）のＮＮパラメーターをＮＮ表現（１１０）に符号化する装置（６００）であって、前記バッチノルム演算子（７１０）は、
【数４】

$$\frac{W*X+b-\mu}{\sqrt{\sigma^2+\epsilon}}\cdot\gamma+\beta$$

として定義され、式中、
μ、σ^2、γ、及びβは、バッチノルムパラメーターであり、
Wは、重み行列であり、
Xは、ＮＮ層の活性化から導出される入力行列であり、
ｂは、バイアスを形成する転置ベクトルであり、
εは、ゼロ除算回避のための定数であり、
・は、・の一方の側の行列と他方の側の転置ベクトルとの間の列に関するアダマール乗算を示し、
＊は、行列乗算を示し、
前記装置（６００）は、
ｂ、μ、σ^2又はσ、γ、及びβ（６１０）を受信することと、
【数５】

1) $\beta':=\beta+\frac{(b-\mu)*\gamma}{\sqrt{\sigma^2+\epsilon}}$

2) $\gamma':=\gamma\cdot\frac{\sqrt{\theta+\epsilon}}{\sqrt{\sigma^2+\epsilon}}$

を計算（６２０）することと、
β’及びγ’を前記バッチノルム演算子（７１０）のＮＮパラメーターとして前記ＮＮ表現（１１０）に符号化することであって、前記バッチノルム演算子（７１０）を
【数６】

$$\frac{W*X+b'-\mu'}{\sqrt{\sigma'^2+\epsilon}}\cdot\gamma'+\beta'$$

として定義し、
３）σ'^2：＝θ
４）μ’：＝０
５）ｂ’：＝０であり、ここで、θは所定のパラメーターであることと、
を行うように構成されており、
装置（６００）が、β’及びγ’を量子化して前記ＮＮ表現（１１０）に符号化する、請求項１に記載の装置（１００）を更に備える、装置。
【請求項３５】
請求項３４に記載の装置（６００）であって、前記所定のパラメーターは、１又は１－εである、装置。
【請求項３６】
ＮＮのバッチノルム演算子（７１０）のＮＮパラメーターをＮＮ表現（１１０）に符号化する装置（６００）であって、前記バッチノルム演算子（７１０）は、
【数７】

$$\frac{W*X+b-\mu}{\sqrt{\sigma^2}}\cdot\gamma+\beta$$

として定義され、式中、
μ、σ^2、γ、及びβは、バッチノルムパラメーターであり、
Wは、重み行列であり
Xは、ＮＮ層の活性化から導出される入力行列であり、
bは、バイアスを形成する転置ベクトルであり、
・は、・の一方の側の行列と他方の側の転置ベクトルとの間の列に関するアダマール乗算を示し、
＊は、行列乗算を示し、
前記装置（６００）は、
b、μ、σ^2又はσ、γ、及びβ（６１０）を受信することと、
【数８】

1) $\beta' := \beta + \frac{(b-\mu)*\gamma}{\sqrt{\sigma^2}}$

2) $\gamma' := \gamma \cdot \frac{1}{\sqrt{\sigma^2}}$

を計算（６２０）することと、
β’及びγ’を前記バッチノルム演算子（７１０）のＮＮパラメーターとして前記ＮＮ表現（１１０）に符号化することであって、前記バッチノルム演算子（７１０）を
【数９】

$$\frac{W*X+b'-\mu'}{\sqrt{\sigma'^2}} \cdot \gamma' + \beta'$$

として定義し、
３）σ'^2：＝１
４）μ’：＝０
５）b’：＝０であることと、
を行うように構成され、
装置（６００）が、β’及びγ’を量子化して前記ＮＮ表現（１１０）に符号化する、請求項１に記載の装置（１００）を更に備える、装置。
【請求項３７】
ＮＮのバッチノルム演算子（７１０）のＮＮパラメーターをＮＮ表現（１１０）に符号化する装置（６００）であって、前記バッチノルム演算子（７１０）は、
【数１０】

$$\frac{W*X-\mu}{\sqrt{\sigma^2+\epsilon}} \cdot \gamma + \beta$$

として定義され、式中、
μ、σ^2、γ、及びβは、バッチノルムパラメーターであり、
Wは、重み行列であり、
Xは、ＮＮ層の活性化から導出される入力行列であり、
εは、ゼロ除算回避のための定数であり、
・は、・の一方の側の行列と他方の側の転置ベクトルとの間の列に関するアダマール乗算を示し、
＊は、行列乗算を示し、

前記装置（６００）は、
μ、σ^2又はσ、γ、及びβ（６１０）を受信することと、
【数１１】

1) $\beta' := \beta - \frac{\mu * \gamma}{\sqrt{\sigma^2 + \epsilon}}$

2) $\gamma' := \gamma \cdot \frac{\sqrt{\theta + \epsilon}}{\sqrt{\sigma^2 + \epsilon}}$

を計算（６２０）することと、
β’及びγ’を前記バッチノルム演算子（７１０）のＮＮパラメーターとして前記ＮＮ表現（１１０）に符号化することであって、前記バッチノルム演算子（７１０）を
【数１２】

$$\frac{W * X - \mu'}{\sqrt{\sigma'^2 + \epsilon}} \cdot \gamma' + \beta'$$

として定義し、
３）σ'^2：＝θ
４）μ’：＝０であり、ここで、θは所定のパラメーターであることと、
を行うように構成され、
装置（６００）が、β’及びγ’を量子化して前記ＮＮ表現（１１０）に符号化する、請求項１に記載の装置（１００）を更に備える、装置。
【請求項３８】
請求項３７に記載の装置（６００）であって、前記所定のパラメーターは、１又は１－εである、装置。
【請求項３９】
ＮＮのバッチノルム演算子（７１０）のＮＮパラメーターをＮＮ表現（１１０）に符号化する装置（６００）であって、前記バッチノルム演算子（７１０）は、
【数１３】

$$\frac{W * X - \mu}{\sqrt{\sigma^2}} \cdot \gamma + \beta$$

として定義され、式中、
μ、σ^2、γ、及びβは、バッチノルムパラメーターであり、
Ｗは、重み行列であり、
Ｘは、ＮＮ層の活性化から導出される入力行列であり、
・は、・の一方の側の行列と他方の側の転置ベクトルとの間の列に関するアダマール乗算を示し、
＊は、行列乗算を示し、
前記装置（６００）は、
μ、σ^2又はσ、γ、及びβ（６１０）を受信することと、
【数１４】

1) $\beta' := \beta - \frac{\mu * \gamma}{\sqrt{\sigma^2}}$

2) $\gamma' := \gamma \cdot \frac{1}{\sqrt{\sigma^2}}$

を計算（６２０）することと、
β’及びγ’を前記バッチノルム演算子（７１０）のＮＮパラメーターとして前記ＮＮ表現（１１０）に符号化することであって、前記バッチノルム演算子（７１０）を
【数１５】

$$\frac{W * X - \mu'}{\sqrt{\sigma'^2}} \cdot \gamma' + \beta'$$

として定義し、
３）$\sigma'^2 := 1$
４）μ’：＝０であることと、
を行うように構成され、
装置（６００）が、β’及びγ’を量子化して前記ＮＮ表現（１１０）に符号化する、請求項１に記載の装置（１００）を更に備える、装置。
【請求項４０】
請求項３４～３９のいずれか一項に記載の装置（６００）であって、
前記表現において、σ'^2の全ての成分が互いに等しいこと、及びその値を示し、及び／又は
前記表現において、μ’の全ての成分が互いに等しいこと、及びその値を示し、及び／又は
前記表現において、存在する場合、ｂ’の全ての成分が互いに等しいこと、及びその値を示すように更に構成されている、装置。
【請求項４１】
請求項３４～３９のいずれか一項に記載の装置（６００）であって、２つのバッチノルム符号化モードの間で切り替え可能であるように更に構成され、第１のバッチノルム符号化モードでは、前記装置（６００）は、β’及びγ’の前記計算及び前記符号化を実行するように構成され、第２のバッチノルム符号化モードでは、前記装置（６００）は、前記受信されたμ、σ^2又はσ、γ、及びβ、並びに存在する場合、ｂを符号化するように構成されている、装置。
【請求項４２】
ＮＮ表現（１１０）からＮＮのバッチノルム演算子（７１０）のＮＮパラメーターを復号化する装置（７００）であって、前記バッチノルム演算子（７１０）は、
【数１６】

$$\frac{W * X + b - \mu}{\sqrt{\sigma^2 + \epsilon}} \cdot \gamma + \beta$$

として定義され、式中、
μ、σ^2、γ、及びβは、バッチノルムパラメーターであり、
Ｗは、重み行列であり、
Ｘは、ＮＮ層の活性化から導出される入力行列であり、
ｂは、バイアスを形成する転置ベクトルであり、
εは、ゼロ除算回避のための定数であり、
・は、・の一方の側の行列と他方の側の転置ベクトルとの間の列に関するアダマール乗

算を示し、
　＊は、行列乗算を示し、
前記装置（７００）は、
前記ＮＮ表現（１１０）からγ（７２２）及びβ（７２４）を導出（７２０）することと、
その全ての成分に適用される１つのシグナリング（７３４）によって、
　１）σ'^2：＝θ
　２）μ’：＝０
　３）ｂ’：＝０であり、ここで、θは所定のパラメーターであることを推論又は導出（７３０）することと、
を行うように構成され、
前記ＮＮ表現（１１０）からβ及びγを導出する、請求項２～２１のいずれか一項に記載の装置（３００）を更に備える、装置。
【請求項４３】
請求項４２に記載の装置（７００）であって、前記所定のパラメーターは、１又は１－εである、装置。
【請求項４４】
ＮＮ表現（１１０）からＮＮのバッチノルム演算子（７１０）のＮＮパラメーターを復号化する装置（７００）であって、前記バッチノルム演算子（７１０）は、
【数１７】

$$\frac{W * X + b - \mu}{\sqrt{\sigma^2}} \cdot \gamma + \beta$$

として定義され、式中、
　μ、σ^2、γ、及びβは、バッチノルムパラメーターであり、
　Ｗは、重み行列であり、
　Ｘは、ＮＮ層の活性化から導出される入力行列であり、
　ｂは、バイアスを形成する転置ベクトルであり、
　・は、・の一方の側の行列と他方の側の転置ベクトルとの間の列に関するアダマール乗算を示し、
　＊は、行列乗算を示し、
前記装置（７００）は、
前記ＮＮ表現（１１０）からγ（７２２）及びβ（７２４）を導出（７２０）することと、
その全ての成分に適用される１つのシグナリング（７３４）によって、
　１）σ^2：＝１
　２）μ：＝０
　３）ｂ：＝０であることを推論又は導出（７３０）することと、
を行うように構成され、
前記ＮＮ表現（１１０）からβ及びγを導出する、請求項２～２１のいずれか一項に記載の装置（３００）を更に備える、装置。
【請求項４５】
ＮＮ表現（１１０）からＮＮのバッチノルム演算子（７１０）のＮＮパラメーターを復号化する装置（７００）であって、前記バッチノルム演算子（７１０）は、
【数１８】

$$\frac{W * X - \mu}{\sqrt{\sigma^2 + \epsilon}} \cdot \gamma + \beta$$

として定義され、式中、
μ、σ^2、γ、及びβは、バッチノルムパラメーターであり、
Wは、重み行列であり、
Xは、ＮＮ層の活性化から導出される入力行列であり、
εは、ゼロ除算回避のための定数であり、
・は、・の一方の側の行列と他方の側の転置ベクトルとの間の列に関するアダマール乗算を示し、
＊は、行列乗算を示し、
前記装置（７００）は、
前記ＮＮ表現（１１０）からγ（７２２）及びβ（７２４）を導出（７２０）することと、
その全ての成分に適用される１つのシグナリング（７３４）によって、
１）σ^2：＝θ
２）μ：＝０であり、ここで、θは所定のパラメーターであることを推論又は導出（７３０）することと、
を行うように構成され、
前記ＮＮ表現（１１０）からβ及びγを導出する、請求項２～２１のいずれか一項に記載の装置（３００）を更に備える、装置。
【請求項４６】
請求項４５に記載の装置（７００）であって、前記所定のパラメーターは、１又は１－εである、装置。
【請求項４７】
ＮＮ表現（１１０）からＮＮのバッチノルム演算子（７１０）のＮＮパラメーターを復号化する装置（７００）であって、前記バッチノルム演算子（７１０）は、
【数１９】

$$\frac{W * X - \mu}{\sqrt{\sigma^2}} \cdot \gamma + \beta$$

として定義され、式中、
μ、σ^2、γ、及びβは、バッチノルムパラメーターであり、
Wは、重み行列であり、
Xは、ＮＮ層の活性化から導出される入力行列であり、
・は、・の一方の側の行列と他方の側の転置ベクトルとの間の列に関するアダマール乗算を示し、
＊は、行列乗算を示し、
前記装置（７００）は、
前記ＮＮ表現（１１０）からγ（７２２）及びβ（７２４）を導出（７２０）することと、
その全ての成分に適用される１つのシグナリング（７３４）によって、
１）σ^2：＝１
２）μ：＝０であることを推論又は導出（７３０）することと、
を行うように構成され、
前記ＮＮ表現（１１０）からβ及びγを導出する、請求項２～２１のいずれか一項に記

載の装置（３００）を更に備える、装置。
【請求項４８】
請求項４２～４７のいずれか一項に記載の装置（７００）であって、
前記表現から、σ^2の全ての成分が互いに等しいこと、及びその値を導出し、及び／又は
前記表現から、μ’の全ての成分が互いに等しいこと、及びその値を導出し、及び／又は
存在する場合、前記表現から、ｂ’の全ての成分が互いに等しいこと、及びその値を導出するように更に構成されている、装置。
【請求項４９】
請求項４２～４８のいずれか一項に記載の装置（７００）であって、２つのバッチノルム符号化モード間で切り替え可能であるように更に構成され、第１のバッチノルム符号化モードでは、前記装置（７００）は、前記導出することと、前記推論又は導出することと、を実行するように構成され、第２の第１のバッチノルム符号化モードでは、前記装置（７００）は、μ、σ^2又はσ、γ、及びβ、並びに存在する場合、ｂを前記表現から復号化するように構成されている、装置。
【請求項５０】
ＮＮ表現を生成する方法であって、
前記ＮＮ表現は、ＮＮパラメーターを表すために、量子化パラメーター及び量子化値を含み、
量子化パラメーターから、
前記量子化パラメーターによって導出された被除数と精度パラメーターによって導出された除数との間の除算の剰余に基づく乗数と、
前記除算の商の丸めに基づくビットシフト数と、
が導出可能であるように、ＮＮパラメーターの前記量子化パラメーター及び量子化値を決定することによって、前記ＮＮパラメーターを量子化された値に量子化することであって、
それにより、前記ＮＮパラメーターの前記量子化された値は、前記量子化値と前記乗数に依存する因数との積であって、前記ビットシフト数に依存するビット数だけビットシフトされた積に対応することを含む、方法。
【請求項５１】
ＮＮ表現からＮＮパラメーターを導出する方法であって、
前記ＮＮ表現から量子化パラメーターを導出することと、
前記ＮＮ表現から量子化値を導出することと、
前記量子化パラメーターから、
前記量子化パラメーターによって導出された被除数と精度パラメーターによって導出された除数との間の除算の剰余に基づく乗数と、
前記除算の商の丸めに基づくビットシフト数と、
を導出することと、
を含み、
前記ＮＮパラメーターは、前記量子化値と前記乗数に依存する因数との積であって、前記ビットシフト数に依存するビット数だけビットシフトされた積に対応する、方法。
【請求項５２】
ＮＮを使用して推論を実行する方法であって、
ＮＮ表現からＮＮパラメーターを導出するために請求項５１に記載の方法を使用して、前記ＮＮをパラメーター化することと、
前記ＮＮを使用して、ＮＮ入力に基づいて推論出力を計算することと、
を含む、方法。
【請求項５３】
ＮＮを使用して推論を実行する方法であって、前記ＮＮを使用してＮＮ入力に基づいて推論出力を計算することを含み、前記ＮＮは、一対のＮＮ層と、前記一対のＮＮ層のうちの第１のＮＮ層から前記一対のＮＮ層のうちの第２のＮＮ層へのニューロン間活性化フィードフォワードとを含み、前記方法は、

前記第１のＮＮ層の前記ニューラルネットワークニューロンの前記活性化から行列Ｘを形成することと、
ｓ・Ｗ’＊Ｘを計算することであって、式中、＊は、行列乗算を表し、Ｗ’は、ｎ及びｍ∈Ｎである次元ｎ×ｍの重み行列であり、ｓは、長さｎの転置ベクトルであり、・は、・の一方の側の行列と・の他方の側の転置ベクトルとの間の列に関するアダマール乗算を示すことと、
によって、前記第１のＮＮ層の前記ニューラルネットワークニューロンの活性化に基づいて、前記第２のＮＮ層の前記ニューラルネットワークニューロンの活性化を計算することを含み、
前記方法が、請求項５１に記載のＮＮ表現（１１０）からＮＮパラメーターを導出する方法を用いて、ＮＮ表現からＷ’を導出する、方法。
【請求項５４】
ＮＮのバッチノルム演算子のＮＮパラメーターをＮＮ表現に符号化する方法であって、前記バッチノルム演算子は、
【数２０】

$$\frac{W * X + b - \mu}{\sqrt{\sigma^2 + \epsilon}} \cdot \gamma + \beta$$

として定義され、式中、
μ、σ^2、γ、及びβは、バッチノルムパラメーターであり、
Ｗは、重み行列であり、
Ｘは、ＮＮ層の活性化から導出される入力行列であり、
ｂは、バイアスを形成する転置ベクトルであり、
εは、ゼロ除算回避のための定数であり、
・は、・の一方の側の行列と他方の側の転置ベクトルとの間の列に関するアダマール乗算を示し、
＊は、行列乗算を示し、
前記方法は、
ｂ、μ、σ^2又はσ、γ、及びβを受信することと、
【数２１】

1) $\beta' := \beta + \frac{(b-\mu)*\gamma}{\sqrt{\sigma^2+\epsilon}}$

2) $\gamma' := \gamma \cdot \frac{\sqrt{\theta+\epsilon}}{\sqrt{\sigma^2+\epsilon}}$d

を計算することと、
β’及びγ’を前記バッチノルム演算子のＮＮパラメーターとして前記ＮＮ表現に符号化することであって、前記バッチノルム演算子を
【数２２】

$$\frac{W * X + b' - \mu'}{\sqrt{\sigma'^2 + \epsilon}} \cdot \gamma' + \beta'$$

として定義し、
３）$\sigma'^2 := \theta$

４）μ’：＝０
５）ｂ’：＝０であり、ここで、θは所定のパラメーターであることと、を含み、
前記方法が、β’及びγ’を量子化して前記ＮＮ表現に符号化する、請求項５０に記載の方法を更に備える、方法。
【請求項５５】
ＮＮのバッチノルム演算子のＮＮパラメーターをＮＮ表現に符号化する方法であって、前記バッチノルム演算子は、
【数２３】

$$\frac{W * X + b - \mu}{\sqrt{\sigma^2}} \cdot \gamma + \beta$$

として定義され、式中、
μ、σ^2、γ、及びβは、バッチノルムパラメーターであり、
Ｗは、重み行列であり、
Ｘは、ＮＮ層の活性化から導出される入力行列であり、
ｂは、バイアスを形成する転置ベクトルであり、
・は、・の一方の側の行列と他方の側の転置ベクトルとの間の列に関するアダマール乗算を示し、
＊は、行列乗算を示し、
前記方法は、
ｂ、μ、σ^2又はσ、γ、及びβを受信することと、
【数２４】

1) $\beta' := \beta + \frac{(b-\mu)*\gamma}{\sqrt{\sigma^2}}$

2) $\gamma' := \gamma \cdot \frac{1}{\sqrt{\sigma^2}}$

を計算することと、
β’及びγ’を前記バッチノルム演算子のＮＮパラメーターとして前記ＮＮ表現に符号化することであって、前記バッチノルム演算子を
【数２５】

$$\frac{W * X + b' - \mu'}{\sqrt{\sigma'^2}} \cdot \gamma' + \beta'$$

として定義し、
３）σ'^2：＝１
４）μ’：＝０
５）ｂ’：＝０であることと、を含み、
前記方法が、β’及びγ’を量子化して前記ＮＮ表現に符号化する、請求項５０に記載の方法を更に備える、方法。
【請求項５６】
ＮＮのバッチノルム演算子のＮＮパラメーターをＮＮ表現に符号化する方法であって、前記バッチノルム演算子は、
【数２６】

$$\frac{W * X - \mu}{\sqrt{\sigma^2 + \epsilon}} \cdot \gamma + \beta$$

として定義され、式中、
μ、σ^2、γ、及びβは、バッチノルムパラメーターであり、
Wは、重み行列であり、
Xは、ＮＮ層の活性化から導出される入力行列であり、
εは、ゼロ除算回避のための定数であり、
・は、・の一方の側の行列と他方の側の転置ベクトルとの間の列に関するアダマール乗算を示し、
＊は、行列乗算を示し、
前記方法は、
μ、σ^2又はσ、γ、及びβを受信することと、
【数２７】

1) $\beta' := \beta - \frac{\mu * \gamma}{\sqrt{\sigma^2 + \epsilon}}$

2) $\gamma' := \gamma \cdot \frac{\sqrt{\theta + \epsilon}}{\sqrt{\sigma^2 + \epsilon}}$

を計算することと、
β’及びγ’を前記バッチノルム演算子のＮＮパラメーターとして前記ＮＮ表現に符号化することであって、前記バッチノルム演算子を
【数２８】

$$\frac{W * X - \mu'}{\sqrt{\sigma'^2 + \epsilon}} \cdot \gamma' + \beta'$$

として定義し、
３）σ'^2：＝θ
４）μ’：＝０であり、ここで、θは所定のパラメーターであることと、を含み、
前記方法が、β’及びγ’を量子化して前記ＮＮ表現に符号化する、請求項５０に記載の方法を更に備える、方法。
【請求項５７】
ＮＮのバッチノルム演算子のＮＮパラメーターをＮＮ表現に符号化する方法であって、前記バッチノルム演算子は、
【数２９】

$$\frac{W * X - \mu}{\sqrt{\sigma^2}} \cdot \gamma + \beta$$

として定義され、式中、
μ、σ^2、γ、及びβは、バッチノルムパラメーターであり、
Wは、重み行列であり、
Xは、ＮＮ層の活性化から導出される入力行列であり、
・は、・の一方の側の行列と他方の側の転置ベクトルとの間の列に関するアダマール乗

算を示し、
　＊は、行列乗算を示し、
前記方法は、
μ、σ^2又はσ、γ、及びβを受信することと、
【数３０】

1) $\beta' := \beta - \frac{\mu * \gamma}{\sqrt{\sigma^2}}$

2) $\gamma' := \gamma \cdot \frac{1}{\sqrt{\sigma^2}}$

sを計算することと、
β’及びγ’を前記バッチノルム演算子のＮＮパラメーターとして前記ＮＮ表現に符号化することであって、前記バッチノルム演算子を
【数３１】

$$\frac{W * X - \mu'}{\sqrt{\sigma'^2}} \cdot \gamma' + \beta'$$

として定義し、
　３）σ'^2：＝１
　４）μ’：＝０であることと、を含み、
前記方法が、β’及びγ’を量子化して前記ＮＮ表現に符号化する、請求項５０に記載の方法を更に備える、方法。
【請求項５８】
ＮＮ表現からＮＮのバッチノルム演算子のＮＮパラメーターを復号化する方法であって、
前記バッチノルム演算子は、
【数３２】

$$\frac{W * X + b - \mu}{\sqrt{\sigma^2 + \epsilon}} \cdot \gamma + \beta$$

として定義され、式中、
　μ、σ^2、γ、及びβは、バッチノルムパラメーターであり、
　Ｗは、重み行列であり、
　Ｘは、ＮＮ層の活性化から導出される入力行列であり、
　ｂは、バイアスを形成する転置ベクトルであり、
　εは、ゼロ除算回避のための定数であり、
　・は、・の一方の側の行列と他方の側の転置ベクトルとの間の列に関するアダマール乗算を示し、
　＊は、行列乗算を示し、
前記方法は、
前記ＮＮ表現からγ及びβを導出することと、
その全ての成分に適用される１つのシグナリングによって、
　１）σ'^2：＝θ
　２）μ’：＝０
　３）ｂ’：＝０であり、ここで、θは所定のパラメーターであることを推論又は導出することと、を含み、

前記ＮＮ表現からβ及びγを導出する、請求項５１に記載の方法を更に備える、方法。
【請求項５９】
ＮＮ表現からＮＮのバッチノルム演算子のＮＮパラメーターを復号化する方法であって、
前記バッチノルム演算子は、
【数３３】

$$\frac{W * X + b - \mu}{\sqrt{\sigma^2}} \cdot \gamma + \beta$$

として定義され、式中、
μ、σ^2、γ、及びβは、バッチノルムパラメーターであり、
Ｗは、重み行列であり、
Ｘは、ＮＮ層の活性化から導出される入力行列であり、
ｂは、バイアスを形成する転置ベクトルであり、
・は、・の一方の側の行列と他方の側の転置ベクトルとの間の列に関するアダマール乗算を示し、
＊は、行列乗算を示し、
前記方法は、
前記ＮＮ表現からγ及びβを導出することと、
その全ての成分に適用される１つのシグナリングによって、
１）σ^2：＝１
２）μ：＝０
３）ｂ：＝０であることを推論又は導出することと、を含み、
前記ＮＮ表現からβ及びγを導出する、請求項５１に記載の方法を更に備える、方法。
【請求項６０】
ＮＮ表現からＮＮのバッチノルム演算子のＮＮパラメーターを復号化する方法であって、
前記バッチノルム演算子は、
【数３４】

$$\frac{W * X - \mu}{\sqrt{\sigma^2 + \epsilon}} \cdot \gamma + \beta$$

として定義され、式中、
μ、σ^2、γ、及びβは、バッチノルムパラメーターであり、
Ｗは、重み行列であり、
Ｘは、ＮＮ層の活性化から導出される入力行列であり、
εは、ゼロ除算回避のための定数であり、
・は、・の一方の側の行列と他方の側の転置ベクトルとの間の列に関するアダマール乗算を示し、
＊は、行列乗算を示し、
前記方法は、
前記ＮＮ表現からγ及びβを導出することと、
その全ての成分に適用される１つのシグナリングによって、
１）σ^2：＝θ
２）μ：＝０であり、ここで、θは所定のパラメーターであることを推論又は導出することと、を含み、
前記ＮＮ表現からβ及びγを導出する、請求項５１に記載の方法を更に備える、方法。
【請求項６１】

ＮＮ表現からＮＮのバッチノルム演算子のＮＮパラメーターを復号化する方法であって、
前記バッチノルム演算子は、
【数３５】

$$\frac{W * X - \mu}{\sqrt{\sigma^2}} \cdot \gamma + \beta$$

として定義され、式中、
μ、σ^2、γ、及びβは、バッチノルムパラメーターであり、
Ｗは、重み行列であり、
Ｘは、ＮＮ層の活性化から導出される入力行列であり、
・は、・の一方の側の行列と他方の側の転置ベクトルとの間の列に関するアダマール乗算を示し、
＊は、行列乗算を示し、
前記方法は、
前記ＮＮ表現からγ及びβを導出することと、
その全ての成分に適用される１つのシグナリングによって、
１）σ^2：＝１
２）μ：＝０であることを推論又は導出することと、を含み、
前記ＮＮ表現からβ及びγを導出する、請求項５１に記載の方法を更に備える、方法。
【請求項６２】
コンピュータープログラムがコンピューターで実行される時、請求項５０～６１のいずれか一項に記載の方法を実行するコンピュータープログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明による実施形態は、ニューラルネットワークパラメーターの表現の改善された概念を使用して、ニューラルネットワークパラメーターを符号化又は復号化する装置及び方法に関する。推論及び／又は記憶ビットレート最適化に関する改善を達成することができる。
【背景技術】
【０００２】
ニューラルネットワークは、その最も基本的な形態において、一連のアフィン変換とそれに続く要素ごとの非線形関数とを構成する。図１に示すように、それらは有向非巡回グラフとして表すことができる。各ノードは、エッジのそれぞれの重み値との乗算によって次のノードに順方向伝搬される特定の値を伴う。次に、全ての入力値が単純に集約される。
【０００３】
図１は、フィードフォワードニューラルネットワークのグラフ表現の例を示している。具体的には、この２層ニューラルネットワークは、４次元入力ベクトルを実線に写像する非線形関数である。
【０００４】
数学的には、図１のニューラルネットワークは次の方法で出力を算出する。
$output = L_2(L_1(input))$
ここで、
$L_i(X) = N_i(B_i(X))$
であり、式中、B_iは層ｉのアフィン変換であり、N_iは層ｉの何らかの非線形関数である。
【０００５】
バイアス層

いわゆる「バイアス層」の場合、B_iは、層ｉに関連する重みパラメーター（エッジ重み）W_iと層ｉの入力X_iとの行列乗算と、それに続くバイアスb_iとの和である。

$B_i(X) = W_i * X_i + b_i$

W_iは、次元$n_i \times k_i$を有する重み行列であり、X_iは、次元$k_i \times m_i$を有する入力行列である。バイアスb_iは、長さn_iの転置ベクトルである。演算子＊は、行列乗算を示すものとする。バイアスb_iとの和は、行列の列に対する要素ごとの演算である。より正確には、$W_i * X_i + b_i$は、b_iが$W_i * X_i$の各列に追加されることを意味する。

【０００６】

いわゆる畳み込み層は、非特許文献１に記載されているように、それらを行列－行列積としてキャストすることによって使用することもできる。

【０００７】

以降、所与の入力から出力を算出する手順を推論と称する。また、中間結果を、隠れ層又は隠れ活性化値と称し、これは、例えば上記の第１の内積＋非線形性の算出等、線形変換＋要素ごとの非線形性を構成する。

【０００８】

通常、ニューラルネットワークは、数百万のパラメーターを含むため、その表現のために数百メガバイトを必要とし得る。したがって、その推論手順には、大きな行列間の多くの内積演算の計算が含まれるため、その実行には高い計算リソースが必要となる。したがって、これらの内積を実行する複雑性を低減することが非常に重要である。

【０００９】

バッチノルム層

ニューラルネットワーク層のアフィン変換のより洗練された変形例として、以下のような、いわゆるバイアス及びバッチノルム演算が挙げられる。

式１：

【数１】

$$BN(X) = \frac{B(X) - \mu}{\sqrt{\sigma^2 + \epsilon}} \cdot \gamma + \beta = \frac{W * X + b - \mu}{\sqrt{\sigma^2 + \epsilon}} \cdot \gamma + \beta$$

ここで、μ、σ^2、γ、及びβは、バッチノルムパラメーターを示す。なお、層インデックスｉはここでは無視する。Ｗは、次元ｎ×ｋを有する重み行列であり、Ｘは、次元ｋ×ｍを有する入力行列である。バイアスｂ及びバッチノルムパラメーターμ、σ^2、γ、及びβは、長さｎの転置ベクトルである。演算子＊は、行列乗算を示す。ベクトルを有する行列に対する他の全ての演算（加算、乗算、除算）は、行列の列に対する要素ごとの演算であることに留意されたい。例えば、Ｘ・γは、Ｘの各列がγと要素ごとに乗算されることを意味する。εは、０による除算を避けるために必要な小さなスカラー数（０．００１等）である。ただし、０であってもよい。

【００１０】

ｂの全てのベクトル要素が０に等しい場合、式１はバッチノルム層を指す。

対照的に、ε並びにμ及びβの全てのベクトル要素が０に設定され、γ及びσ^2の全ての要素が１に設定される場合、バッチノルムのない層（バイアスのみ）が処理される。

【００１１】

パラメーターの効率的な表現

パラメーターＷ、ｂ、μ、σ^2、γ、及びβは、集合的に層のパラメーターを示すものとする。それらは通常、ビットストリーム中でシグナリングされる必要がある。例えば、それらは３２ビットの浮動小数点数として表すことができ、又は整数表現に量子化することができる。εは通常、ビットストリーム中でシグナリングされないことに留意されたい。

【００１２】

かかるパラメーターを符号化するための特に効率的な手法では、各値がいわゆる量子化ステップサイズ値の整数倍として表される均一再構成量子化器が用いられる。対応する浮動小数点数は、整数に、通常は単一の浮動小数点数である量子化ステップサイズを乗算することによって再構築することができる。しかしながら、ニューラルネットワーク推論のための効率的な実施態様（すなわち、入力に対するニューラルネットワークの出力の算出）では、可能な限り整数演算が用いられる。したがって、パラメーターを浮動小数点表現に再構成する必要があることは望ましくない場合がある。

【先行技術文献】

【非特許文献】

【００１３】

【文献】“cuDNN: Efficient Primitives for Deep Learning”(Sharan Chetlur, et al.; arXiv: 1410.0759, 2014)

【発明の概要】

【発明が解決しようとする課題】

【００１４】

したがって、ニューラルネットワークパラメーターの効率的な符号化及び／又は復号化をサポートするために、かかるパラメーターの表現の概念を改善することが望まれている。ニューラルネットワークパラメーターが符号化されるビットストリームを低減して、シグナル化コストを低減することが望ましい場合がある。加えて、又は代替として、ニューラルネットワーク推論を改善するために計算リソースの複雑性を低減することが望ましい場合があり、例えば、ニューラルネットワーク推論のための効率的な実施態様を達成することが望ましい場合がある。

【課題を解決するための手段】

【００１５】

これは、本出願の独立請求項の主題によって達成される。

【００１６】

本発明による更なる実施の形態は、本出願の従属請求項の主題によって定義される。

本発明の第１の態様によれば、本出願の発明者らは、ニューラルネットワーク（ＮＮ）表現が直面する１つの問題が、ニューラルネットワークが数百万のパラメーターを含み、したがってその表現のために数百メガバイトを必要とし得るという事実から生じることを認識した。したがって、その推論手順には、大きな行列間の多くの内積演算の計算が含まれるため、その実行には高い計算リソースが必要となる。本出願の第１の態様によれば、この困難は、ＮＮパラメーターの量子化を使用することによって克服され、浮動小数点演算をほとんど又は全く用いずに推論が可能となる。本発明者らは、乗数及びビットシフト数を導出することができる量子化パラメーターを決定することが有利であることを見出した。これは、ビットレートに関して、３２ビット浮動小数点値の代わりに量子化パラメーター及び量子化値（quantization value）のみをシグナリングすることが効率的であるという着想に基づいている。ＮＮパラメーターの量子化された値（quantized value）は、乗数、ビットシフト数、及び量子化値を使用して算出することができるため、浮動小数点領域の代わりに整数領域において、計算、例えばＮＮパラメーターの和及び／又はＮＮパラメーターとベクトルとの乗算を実行することが可能である。したがって、提示するＮＮ表現を用いて、推論の効率的な計算を達成することができる。

【００１７】

したがって、本出願の第１の態様によれば、ＮＮ表現、例えばデータストリームを生成する装置は、量子化パラメーターから乗数及びビットシフト数が導出可能であるように、ＮＮパラメーターの量子化パラメーター及び量子化値を決定することによって、ＮＮパラメーターを量子化された値に量子化するように構成される。生成されたＮＮ表現は、ＮＮ表現、例えばデータストリームからＮＮパラメーター、例えばＮＮパラメーターの量子化された値を導出する装置によって読み出され／復号化され得る。ＮＮパラメーターを導出する装置は、ＮＮ表現から量子化パラメーター及び量子化値を導出し、量子化パラメータ

ーから乗数及びビットシフト数を導出するように構成される。乗数は、量子化パラメーターによって導出された被除数と精度パラメーターによって導出された除数との間の除算の剰余に基づいて量子化パラメーターから導出可能であり、例えば、精度パラメーターは、デフォルト値に設定されてもよく、又は、自然数若しくは２の累乗等の精度パラメーターの複数の異なる整数値が、ＮＮ全体に対して、又は各層等のＮＮの各部分に対して装置によってテストされてもよく、量子化誤差及びビットレートの観点から、そのラグランジュ和に関して最良のものがテストされ、精度パラメーターとして最良の値を取得し、ＮＮ表現でこの選択がシグナリングされる。ビットシフト数は、除算の商の丸めに基づいて量子化パラメーターから導出可能である。ＮＮパラメーターを導出する装置の場合のＮＮパラメーター、又はＮＮ表現を生成する装置の場合のＮＮパラメーターの量子化された値は、（例えば、少なくとも、シフトの場合に符号の別個の処理を伴う量子化された値の絶対値に関して、又はさらに、積、その因数、及びシフトに関してそれぞれ２の補数表現及び２の補数演算を使用する場合等、絶対値と符号の両方に関して）量子化値と乗数に依存する因数との間の積であって、ビットシフト数に依存するビット数だけビットシフトされた積に対応する。デジタルデータは、上述したように、ＮＮパラメーターを表すために、量子化パラメーター及び量子化値を含むＮＮ表現を定義することができる。

【００１８】

ＮＮパラメーターを導出する装置によって導出されるＮＮパラメーターは、ＮＮ表現を生成する装置によって生成されるＮＮパラメーターの量子化された値に対応することに留意されたい。これは、ＮＮパラメーターを導出する装置には、元のＮＮパラメーターが見えないため、ＮＮパラメーターを導出する装置から見て、ＮＮパラメーターの量子化された値をＮＮパラメーターとみなすためである。

【００１９】

一実施の形態は、ＮＮを使用して推論を実行するデバイスに関し、該デバイスは、ＮＮをパラメーター化するように構成されたＮＮパラメトライザーを備える。ＮＮパラメトライザーは、上述したように、ＮＮ表現からＮＮパラメーターを導出する装置を備える。加えて、デバイスは、ＮＮを使用してＮＮ入力に基づいて推論出力を計算するように構成された計算ユニットを備える。上述したように、ＮＮパラメーターは、乗数、ビットシフト数及び量子化値に基づいて導出することができ、そのため、浮動小数点領域の代わりに整数領域において、計算、例えばＮＮパラメーターの和及び／又はＮＮパラメーターとベクトルとの乗算を実行することが可能である。したがって、推論の効率的な計算が、デバイスによって達成され得る。

【００２０】

本発明の第２の態様によれば、本出願の発明者らは、ニューラルネットワーク（ＮＮ）を使用して推論を実行するときに遭遇する１つの問題が、推論に使用される重み行列が量子化誤差を有する可能性があり、そのため、低いレベルの精度しか達成されないという事実に起因することを認識した。本出願の第１の態様によれば、この困難は、重み行列Ｗ’の各列と要素ごとに乗算される転置ベクトルｓ、例えばスケーリングファクターを使用することによって克服される。本発明者らは、算術符号化方法が、重み行列のスケーリングを使用することによってより高い符号化利得をもたらすこと、及び／又は重み行列のスケーリングがニューラルネットワーク性能結果を増加させる、例えば、より高い精度を達成することを見出した。これは、量子化誤差を低減し、それにより量子化されたニューラルネットワークの予測性能を向上させるために、例えば重み行列、例えば量子化された重み行列に依存して、転置ベクトルｓを効率的に適応させることができるという着想に基づいている。さらに、本発明者らは、重みパラメーターを転置ベクトルｓと重み行列Ｗ’との合成として因数分解することで、両者を独立に量子化でき、例えば転置ベクトルｓと重み行列Ｗ’の量子化には、様々な量子化パラメーターを使用できるため、表現の効率を高めることができることを見出した。これは、性能の観点から有益であるが、ハードウェア効率の観点からも有益である。

【００２１】

したがって、本出願の第２の態様によれば、ＮＮを使用して推論を実行するデバイスは、ＮＮを使用してＮＮ入力に基づいて推論出力を計算するように構成される。ＮＮは、一対のＮＮ層と、一対のＮＮ層のうちの第１のＮＮ層から一対のＮＮ層のうちの第２のＮＮ層へのニューロン間活性化フィードフォワードとを含む。デバイスは、第１のＮＮ層のニューラルネットワークニューロンの活性化から行列Ｘを形成し、ｓ・Ｗ’＊Ｘを計算することによって、第１のＮＮ層のニューラルネットワークニューロンの活性化に基づいて第２のＮＮ層のニューラルネットワークニューロンの活性化を計算するように構成される。演算子＊は、行列乗算を示し、Ｗ’は、ｎ及びｍ∈Ｎである次元ｎ×ｍの重み行列であり、ｓは、長さｎの転置ベクトルであり、演算子・は、・の一方の側の行列と・の他方の側の転置ベクトルとの間の列に関するアダマール乗算を示す。

【００２２】

本発明の第３の態様によれば、本出願の発明者らは、バッチノルム層を使用するときに遭遇する１つの問題が、バッチノルム演算子のバッチノルムパラメーター／要素が通常は浮動小数点表現であるという事実から生じることを認識した。しかしながら、ニューラルネットワーク推論のための効率的な実施態様（すなわち、入力に対するニューラルネットワークの出力の算出）では、可能な限り整数演算が用いられる。この困難は、所定の定数値をバッチノルムパラメーター／要素に、例えばｂ及びμ及びσ^2又はσに割り当てることによって克服される。本発明者らは、バッチノルムパラメーター／要素が所定の定数値を有する場合、それらを極めて効率的に圧縮することができることを見出した。これは、全ての要素／パラメーターが所定の定数値を有するかを示す単一のフラグの使用を可能にし、それにより、それらを所定の定数値に設定することができるという着想に基づいている。加えて、バッチノルム演算子の結果は、所定の定数値を使用することによって変更されないことを見出した。

【００２３】

したがって、本出願の第３の態様によれば、第１の実施の形態は、ＮＮのバッチノルム演算子のＮＮパラメーターをＮＮ表現に符号化する装置に関する。バッチノルム演算子は、

【数２】

$$\frac{W * X + b - \mu}{\sqrt{\sigma^2 + \epsilon}} \cdot \gamma + \beta$$

として定義され、式中、

μ、σ^2、γ、及びβは、バッチノルムパラメーター、例えば、各出力ノードについて１つの成分を含む転置ベクトルであり、

Ｗは、重み行列であり、例えば、その各行は１つの出力ノードに対するものであり、それぞれの行の各成分はＸの１つの行に関連付けられており、

Ｘは、ＮＮ層の活性化から導出される入力行列であり、

ｂは、バイアスを形成する転置ベクトル、例えば、各出力ノードに対して１つの成分を含む転置ベクトルであり、

εは、ゼロ除算回避のための定数であり、

・は、・の一方の側の行列と他方の側の転置ベクトルとの間の列に関するアダマール乗算を示し、

＊は、行列乗算を示す。

装置は、ｂ及びμ及びγ及びβ及びσ^2又はσを受信し、

【数３】

$$\beta' := \beta + \frac{(b-\mu)*\gamma}{\sqrt{\sigma^2+\epsilon}}$$

及び
【数４】

$$\gamma' := \gamma \cdot \frac{\sqrt{\theta+\epsilon}}{\sqrt{\sigma^2+\epsilon}}$$

を計算するように構成される。
加えて、装置は、β’及びγ’をＮＮ表現に符号化するように構成され、例えば、出力ノードごとに１つの成分を含む転置ベクトルもバッチノルム演算子のＮＮパラメーターとして符号化するように構成され、バッチノルム演算子を
【数５】

$$\frac{W*X+b'-\mu'}{\sqrt{\sigma'^2+\epsilon}} \cdot \gamma' + \beta'$$

として定義し、
σ’２：＝θ、μ’：＝０、ｂ’：＝０であり、ここで、θは所定のパラメーターである。
【００２４】
ＮＮ表現からＮＮのバッチノルム演算子のＮＮパラメーターを復号化する並列装置は、ＮＮ表現からγ及びβを導出し、その全ての成分に適用される１つのシグナリングによって、σ’２：＝θ及びμ’：＝０及びｂ’：＝０を推論又は導出するように構成され、ここで、θは所定のパラメーターである。装置は、例えば、１つのシグナリング、例えばフラグを読み出し、そこからσ’２：＝θ及びμ’：＝０及びｂ’：＝０を推論又は導出するように構成される。バッチノルム演算子は、第３の態様の第１の実施の形態に関して上記で説明したように定義される。
【００２５】
したがって、本出願の第３の態様によれば、第２の実施の形態は、ＮＮのバッチノルム演算子のＮＮパラメーターをＮＮ表現に符号化する装置に関する。バッチノルム演算子は、
【数６】

$$\frac{W*X+b-\mu}{\sqrt{\sigma^2}} \cdot \gamma + \beta$$

として定義され、式中、
μ、σ２、γ、及びβは、バッチノルムパラメーター、例えば、各出力ノードについて１つの成分を含む転置ベクトルであり、
Ｗは、重み行列であり、例えば、その各行は１つの出力ノードに対するものであり、それぞれの行の各成分はＸの１つの行に関連付けられており、

Ｘは、ＮＮ層の活性化から導出される入力行列であり、
ｂは、バイアスを形成する転置ベクトル、例えば、各出力ノードに対して１つの成分を含む転置ベクトルであり、
・は、・の一方の側の行列と他方の側の転置ベクトルとの間の列に関するアダマール乗算を示し、
＊は、行列乗算を示す。
装置は、ｂ及びμ及びγ及びβ及びσ２又はσを受信し、
【数７】

$$\beta' := \beta + \frac{(b-\mu)*\gamma}{\sqrt{\sigma^2}}$$

及び
【数８】

$$\gamma' := \gamma \cdot \frac{1}{\sqrt{\sigma^2}}$$

を計算するように構成される。
加えて、装置は、バッチノルム演算子のＮＮパラメーターとしてβ’及びγ’をＮＮ表現に符号化するように構成され、バッチノルム演算子を
【数９】

$$\frac{W*X+b'-\mu'}{\sqrt{\sigma'^2}}\cdot\gamma'+\beta'$$

として定義し、
σ’２：＝１、μ’：＝０、ｂ’：＝０である。
【００２６】
ＮＮ表現からＮＮのバッチノルム演算子のＮＮパラメーターを復号化する並列装置は、ＮＮ表現からγ及びβを導出し、その全ての成分に適用される１つのシグナリングによって、σ２：＝１及びμ：＝０及びｂ：＝０を推論又は導出するように構成される。装置は、例えば、１つのシグナリング、例えばフラグを読み出し、そこからσ２：＝１及びμ：＝０及びｂ：＝０を推論又は導出するように構成される。バッチノルム演算子は、第３の態様の第２の実施の形態に関して上記で説明したように定義される。
【００２７】
したがって、本出願の第３の態様によれば、第３の実施の形態は、ＮＮのバッチノルム演算子のＮＮパラメーターをＮＮ表現に符号化する装置に関する。バッチノルム演算子は、
【数１０】

$$\frac{W*X-\mu}{\sqrt{\sigma^2+\epsilon}}\cdot\gamma+\beta$$

として定義され、式中、
μ、σ２、γ、及びβは、バッチノルムパラメーター、例えば、各出力ノードについて１つの成分を含む転置ベクトルであり、

Ｗは、重み行列であり、例えば、その各行は１つの出力ノードに対するものであり、それぞれの行の各成分はＸの１つの行に関連付けられており、
Ｘは、ＮＮ層の活性化から導出される入力行列であり、
εは、ゼロ除算回避のための定数であり、
・は、・の一方の側の行列と他方の側の転置ベクトルとの間の列に関するアダマール乗算を示し、
＊は、行列乗算を示す。
装置は、μ及びγ及びβ及びσ^{2}又はσを受信し、
【数１１】

$$\beta' := \beta - \frac{\mu * \gamma}{\sqrt{\sigma^2 + \epsilon}}$$

及び
【数１２】

$$\gamma' := \gamma \cdot \frac{\sqrt{\theta + \epsilon}}{\sqrt{\sigma^2 + \epsilon}}$$

を計算するように構成される。
加えて、装置は、バッチノルム演算子のＮＮパラメーターとしてβ’及びγ’をＮＮ表現に符号化するように構成され、バッチノルム演算子を
【数１３】

$$\frac{W * X - \mu'}{\sqrt{\sigma'^2 + \epsilon}} \cdot \gamma' + \beta'$$

として定義し、
σ’^{2}：＝θ、μ’：＝０であり、θは所定のパラメーターである。
【００２８】
ＮＮ表現からＮＮのバッチノルム演算子のＮＮパラメーターを復号化する並列装置は、ＮＮ表現からγ及びβを導出し、その全ての成分に適用される１つのシグナリングによって、σ^2：＝θ及びμ：＝０を推論又は導出するように構成され、θは所定のパラメーターである。装置は、例えば、１つのシグナリング、例えばフラグを読み出し、そこからσ^2：＝θ及びμ：＝０を推論又は導出するように構成される。バッチノルム演算子は、第３の態様の第３の実施の形態に関して上記で説明したように定義される。
【００２９】
したがって、本出願の第３の態様によれば、第４の実施の形態は、ＮＮのバッチノルム演算子のＮＮパラメーターをＮＮ表現に符号化する装置に関する。バッチノルム演算子は、
【数１４】

$$\frac{W * X - \mu}{\sqrt{\sigma^2}} \cdot \gamma + \beta$$

として定義され、式中、

μ、σ^2、γ、及びβは、バッチノルムパラメーター、例えば、各出力ノードについて１つの成分を含む転置ベクトルであり、
Ｗは、重み行列であり、例えば、その各行は１つの出力ノードに対するものであり、それぞれの行の各成分はＸの１つの行に関連付けられており、
Ｘは、ＮＮ層の活性化から導出される入力行列であり、
・は、・の一方の側の行列と他方の側の転置ベクトルとの間の列に関するアダマール乗算を示し、
＊は、行列乗算を示す。
装置は、μ及びγ及びβ及びσ^2又はσを受信し、
【数１５】

$$\beta' := \beta - \frac{\mu * \gamma}{\sqrt{\sigma^2}}$$

及び
【数１６】

$$\gamma' := \gamma \cdot \frac{1}{\sqrt{\sigma^2}}$$

を計算するように構成される。
加えて、装置は、バッチノルム演算子のＮＮパラメーターとしてβ’及びγ’をＮＮ表現に符号化するように構成され、バッチノルム演算子を
【数１７】

$$\frac{\boldsymbol{W} * \boldsymbol{X} - \boldsymbol{\mu}'}{\sqrt{\boldsymbol{\sigma}'^2}} \cdot \boldsymbol{\gamma}' + \boldsymbol{\beta}'$$

として定義し、
σ'^2：＝１、μ’：＝０である。
【００３０】
ＮＮ表現からＮＮのバッチノルム演算子のＮＮパラメーターを復号化する並列装置は、ＮＮ表現からγ及びβを導出し、その全ての成分に適用される１つのシグナリングによって、σ^2：＝１及びμ：＝０を推論又は導出するように構成される。装置は、例えば、１つのシグナリング、例えばフラグを読み出し、そこからσ^2：＝１及びμ：＝０を推論又は導出するように構成される。バッチノルム演算子は、第３の態様の第４の実施の形態に関して上記で説明したように定義される。
【００３１】
以下の方法は、上述の原理に従って動作する。
【００３２】
一実施の形態は、ＮＮ表現を生成する方法であって、量子化パラメーターから、量子化パラメーターによって導出された被除数と精度パラメーターによって導出された除数との間の除算の剰余に基づいて乗数が導出可能であり、量子化パラメーターから、除算の商の丸めに基づいてビットシフト数が導出可能であるように、ＮＮパラメーターの量子化パラメーター及び量子化値を決定することによって、ＮＮパラメーターを量子化された値に量子化することを含む方法に関する。量子化パラメーターは、ＮＮパラメーターの量子化された値が、量子化値と乗数に依存する因数との積であって、ビットシフト数に依存するビ

ット数だけビットシフトされた積に対応するように決定される。

【００３３】

一実施の形態は、ＮＮ表現からＮＮパラメーターを導出する方法であって、ＮＮ表現から量子化パラメーター及び量子化値を導出することを含む方法に関する。加えて、方法は、量子化パラメーターから、量子化パラメーターによって導出された被除数と精度パラメーターによって導出された除数との間の除算の剰余に基づいて乗数を導出することと、量子化パラメーターから、除算の商の丸めに基づいてビットシフト数を導出することとを含む。ＮＮパラメーターは、量子化値と乗数に依存する因数との積であって、ビットシフト数に依存するビット数だけビットシフトされた積に対応する。

【００３４】

一実施の形態は、ＮＮを使用して推論を実行する方法であって、ＮＮ表現からＮＮパラメーターを導出するために、ＮＮパラメーターを導出する上述の方法を使用してＮＮをパラメーター化することを含む方法に関する。加えて、推論を実行する方法は、ＮＮを使用してＮＮ入力に基づいて推論出力を計算することを含む。

【００３５】

一実施の形態は、ＮＮを使用して推論を実行する方法であって、ＮＮを使用してＮＮ入力に基づいて推論出力を計算することを含む方法に関する。ＮＮは、一対のＮＮ層と、一対のＮＮ層のうちの第１のＮＮ層から一対のＮＮ層のうちの第２のＮＮ層へのニューロン間活性化フィードフォワードとを含む。方法は、第１のＮＮ層のニューラルネットワークニューロンの活性化から行列Ｘを形成することと、ｓ・Ｗ’＊Ｘを計算することであって、式中、＊は、行列乗算を表し、Ｗ’は、ｎ及びｍ∈Ｎである次元ｎ×ｍの重み行列であり、ｓは、長さｎの転置ベクトルであり、・は、・の一方の側の行列と・の他方の側の転置ベクトルとの間の列に関するアダマール乗算を示すこととによって、第１のＮＮ層のニューラルネットワークニューロンの活性化に基づいて、第２のＮＮ層のニューラルネットワークニューロンの活性化を計算することを含む。

【００３６】

一実施の形態は、ＮＮのバッチノルム演算子のＮＮパラメーターをＮＮ表現に符号化する方法に関し、バッチノルム演算子は、

【数１８】

$$\frac{W*X+b-\mu}{\sqrt{\sigma^2+\epsilon}}\cdot\gamma+\beta$$

として定義され、式中、μ、σ^2、γ、及びβは、バッチノルムパラメーターであり、Ｗは、重み行列であり、Ｘは、ＮＮ層の活性化から導出される入力行列であり、ｂは、バイアスを形成する転置ベクトルであり、εは、ゼロ除算回避のための定数であり、・は、・の一方の側の行列と他方の側の転置ベクトルとの間の列に関するアダマール乗算を示し、＊は、行列乗算を示す。方法は、ｂ、μ、γ、β及びσ^2又はσを受信することと、

【数１９】

$$\beta' := \beta + \frac{(b-\mu)*\gamma}{\sqrt{\sigma^2+\epsilon}}$$

及び

【数２０】

$$\gamma' := \gamma \cdot \frac{\sqrt{\theta + \epsilon}}{\sqrt{\sigma^2 + \epsilon}}$$

を計算することと、を含む。

加えて、方法は、β’及びγ’をバッチノルム演算子のＮＮパラメーターとしてＮＮ表現に符号化することであって、バッチノルム演算子を

【数２１】

$$\frac{\boldsymbol{W} * \boldsymbol{X} + \boldsymbol{b}' - \boldsymbol{\mu}'}{\sqrt{\boldsymbol{\sigma}'^2 + \epsilon}} \cdot \boldsymbol{\gamma}' + \boldsymbol{\beta}'$$

として定義し、σ’2：＝θ、μ’：＝０、及びｂ’：＝０であり、ここで、θは所定のパラメーターであることを含む。

【００３７】

一実施の形態は、ＮＮのバッチノルム演算子のＮＮパラメーターをＮＮ表現に符号化する方法に関し、バッチノルム演算子は、

【数２２】

$$\frac{W * X + b - \mu}{\sqrt{\sigma^2}} \cdot \gamma + \beta$$

として定義され、式中、μ、σ^2、γ、及びβは、バッチノルムパラメーターであり、Ｗは、重み行列であり、Ｘは、ＮＮ層の活性化から導出される入力行列であり、ｂは、バイアスを形成する転置ベクトルであり、・は、・の一方の側の行列と他方の側の転置ベクトルとの間の列に関するアダマール乗算を示し、＊は、行列乗算を示す。方法は、ｂ、μ、γ、β、及びσ^2又はσを受信することと、

【数２３】

$$\beta' := \beta + \frac{(b - \mu) * \gamma}{\sqrt{\sigma^2}}$$

及び

【数２４】

$$\gamma' := \gamma \cdot \frac{1}{\sqrt{\sigma^2}}$$

を計算することと、を含む。加えて、方法は、β’及びγ’をバッチノルム演算子のＮＮパラメーターとしてＮＮ表現に符号化することであって、バッチノルム演算子を

【数２５】

$$\frac{W * X + b' - \mu'}{\sqrt{\sigma'^2}} \cdot \gamma' + \beta'$$

として定義し、σ’2：＝１、μ’：＝０、及びｂ’：＝０であることを含む。

【００３８】
一実施の形態は、ＮＮのバッチノルム演算子のＮＮパラメーターをＮＮ表現に符号化する方法に関し、バッチノルム演算子は、
【数２６】

$$\frac{W*X-\mu}{\sqrt{\sigma^2+\epsilon}}\cdot\gamma+\beta$$

として定義され、式中、μ、σ^2、γ、及びβは、バッチノルムパラメーターであり、Ｗは、重み行列であり、Ｘは、ＮＮ層の活性化から導出される入力行列であり、εは、ゼロ除算回避のための定数であり、・は、・の一方の側の行列と他方の側の転置ベクトルとの間の列に関するアダマール乗算を示し、＊は、行列乗算を示す。方法は、μ、γ、β及びσ^{2}又はσを受信することと、
【数２７】

$$\beta':=\beta-\frac{\mu*\gamma}{\sqrt{\sigma^2+\epsilon}}$$

及び
【数２８】

$$\gamma':=\gamma\cdot\frac{\sqrt{\theta+\epsilon}}{\sqrt{\sigma^2+\epsilon}}$$

を計算することと、を含む。加えて、方法は、β’及びγ’をバッチノルム演算子のＮＮパラメーターとしてＮＮ表現に符号化することであって、バッチノルム演算子を
【数２９】

$$\frac{W*X-\mu'}{\sqrt{\sigma'^2+\epsilon}}\cdot\boldsymbol{\gamma'}+\boldsymbol{\beta'}$$

として定義し、σ$'^2$：＝θ及びμ’：＝０であり、ここで、θは所定のパラメーターであることを含む。
【００３９】
一実施の形態は、ＮＮのバッチノルム演算子のＮＮパラメーターをＮＮ表現に符号化する方法に関し、バッチノルム演算子は、
【数３０】

$$\frac{W*X-\mu}{\sqrt{\sigma^2}}\cdot\gamma+\beta$$

として定義され、式中、μ、σ^2、γ、及びβは、バッチノルムパラメーターであり、Ｗは、重み行列であり、Ｘは、ＮＮ層の活性化から導出される入力行列であり、・は、・の一方の側の行列と他方の側の転置ベクトルとの間の列に関するアダマール乗算を示し、＊は、行列乗算を示す。方法は、μ、γ、β及びσ^{2}又はσを受信することと、
【数３１】

$$\beta':=\beta-\frac{\mu*\gamma}{\sqrt{\sigma^2}}$$

及び
【数３２】

$$\gamma' := \gamma \cdot \frac{1}{\sqrt{\sigma^2}}$$

を計算することと、を含む。加えて、方法は、β’及びγ’をバッチノルム演算子のＮＮパラメーターとしてＮＮ表現に符号化することであって、バッチノルム演算子を
【数３３】

$$\frac{W*X-\mu'}{\sqrt{\sigma'^2}} \cdot \gamma' + \beta'$$

として定義し、σ’２：＝１、及びμ’：＝０であることを含む。
【００４０】
一実施の形態は、ＮＮ表現からＮＮのバッチノルム演算子のＮＮパラメーターを復号化する方法に関し、バッチノルム演算子は、
【数３４】

$$\frac{W*X+b-\mu}{\sqrt{\sigma^2+\epsilon}} \cdot \gamma + \beta$$

として定義され、式中、μ、σ２、γ、及びβは、バッチノルムパラメーターであり、Ｗは、重み行列であり、Ｘは、ＮＮ層の活性化から導出される入力行列であり、ｂは、バイアスを形成する転置ベクトルであり、εは、ゼロ除算回避のための定数であり、・は、・の一方の側の行列と他方の側の転置ベクトルとの間の列に関するアダマール乗算を示し、＊は、行列乗算を示す。方法は、ＮＮ表現からγ及びβを導出することと、その全ての成分に適用される１つのシグナリングによって、σ’２：＝θ、μ’：＝０、及びｂ’：＝０であり、ここで、θは所定のパラメーターであることを推論又は導出することとを含む。
【００４１】
一実施の形態は、ＮＮ表現からＮＮのバッチノルム演算子のＮＮパラメーターを復号化する方法に関し、バッチノルム演算子は、
【数３５】

$$\frac{W*X+b-\mu}{\sqrt{\sigma^2}} \cdot \gamma + \beta$$

として定義され、式中、μ、σ２、γ、及びβは、バッチノルムパラメーターであり、Ｗは、重み行列であり、Ｘは、ＮＮ層の活性化から導出される入力行列であり、ｂは、バイアスを形成する転置ベクトルであり、・は、・の一方の側の行列と他方の側の転置ベクトルとの間の列に関するアダマール乗算を示し、＊は、行列乗算を示す。方法は、ＮＮ表現からγ及びβを導出することと、その全ての成分に適用される１つのシグナリングによって、σ２：＝１、μ：＝０、及びｂ：＝０であることを推論又は導出することとを含む。
【００４２】
一実施の形態は、ＮＮ表現からＮＮのバッチノルム演算子のＮＮパラメーターを復号化する方法に関し、バッチノルム演算子は、
【数３６】

$$\frac{W*X-\mu}{\sqrt{\sigma^2+\epsilon}}\cdot\gamma+\beta$$

として定義され、式中、μ、σ^2、γ、及びβは、バッチノルムパラメーターであり、Wは、重み行列であり、Xは、ＮＮ層の活性化から導出される入力行列であり、εは、ゼロ除算回避のための定数であり、・は、・の一方の側の行列と他方の側の転置ベクトルとの間の列に関するアダマール乗算を示し、＊は、行列乗算を示す。方法は、ＮＮ表現からγ及びβを導出することと、その全ての成分に適用される１つのシグナリングによって、σ^2：＝θ及びμ：＝０であり、ここで、θは所定のパラメーターであることを推論又は導出することとを含む。

【００４３】

一実施の形態は、ＮＮ表現からＮＮのバッチノルム演算子のＮＮパラメーターを復号化する方法に関し、バッチノルム演算子は、

【数３７】

$$\frac{W*X-\mu}{\sqrt{\sigma^2}}\cdot\gamma+\beta$$

として定義され、式中、μ、σ^2、γ、及びβは、バッチノルムパラメーターであり、Wは、重み行列であり、Xは、ＮＮ層の活性化から導出される入力行列であり、・は、・の一方の側の行列と他方の側の転置ベクトルとの間の列に関するアダマール乗算を示し、＊は、行列乗算を示す。方法は、ＮＮ表現からγ及びβを導出することと、その全ての成分に適用される１つのシグナリングによって、σ^2：＝１及びμ：＝０であることを推論又は導出することとを含む。

【００４４】

上述したように、これらの方法は、上述した装置又はデバイスと同じ考察に基づいている。方法は、装置又はデバイスに関しても説明される全ての特徴及び機能で完成され得る。

【００４５】

一実施の形態は、上述したように、ＮＮ表現を生成する方法又は装置によって生成されたＮＮ表現を定義するデジタルデータを含むデジタル記憶媒体に関する。

【００４６】

一実施の形態は、上述の方法のうちの１つを実行するコンピュータープログラムに関する。

【００４７】

一実施の形態は、上述したように、ＮＮ表現を生成する方法又は装置によって生成されるデータストリームに関する。

【００４８】

図面は、必ずしも縮尺通りではなく、代わりに、概して、本発明の原理を図示することに重点が置かれている。以下の説明では、本発明の種々の実施形態が、以下の図面を参照して説明される。

【図面の簡単な説明】

【００４９】

【図１】ニューラルネットワークを示す図である。

【図２】本発明の一実施形態による、ＮＮ表現を生成する装置、ＮＮ表現を定義するデジタルデータ、及びＮＮ表現からＮＮパラメーターを導出する装置を概略的に示す図である。

【図３】フィードフォワードニューラルネットワークを概略的に示す図である。

【図４】本発明の一実施形態による、ＮＮパラメトライザーを使用して推論を実行するデ

バイスを概略的に示す図である。
【図５】本発明の一実施形態による、ベクトル及び行列の合成として重みパラメーターを因数分解することによって推論を実行するデバイスを概略的に示す図である。
【図６】本発明の一実施形態による、ＮＮパラメーターをＮＮ表現に符号化する装置及びＮＮ表現からＮＮパラメーターを復号化する装置を概略的に示す図である。
【図７】行列ＸとＷとの間の可能な関係を概略的に示す図である。
【発明を実施するための形態】
【００５０】
同一若しくは同等の要素、又は同一若しくは同等の機能を有する要素は、異なる図に現れる場合であっても、以下の説明において同一又は同等の参照番号によって示される。
【００５１】
以下の説明では、本発明の実施形態のより完全な説明を提供するために、複数の詳細が記載される。しかしながら、本発明の実施形態がこれらの具体的な詳細なしに実施され得ることは、当業者には明らかであろう。他の例では、本発明の実施形態を不明瞭にすることを回避するために、周知の構造及びデバイスは、詳細にではなくブロック図の形態で示される。加えて、本明細書で後に説明される異なる実施形態の特徴は、特に別段の記載がない限り、互いに組み合わせることができる。
【００５２】
以下では、少数の浮動小数点演算のみを用いた、又は更には浮動小数点演算を全く用いない推論を可能にする層のパラメーターの量子化ステップサイズを表し、シグナリングする効率的な手法を提示する。つまり、ビットレートの点で効率的な表現であり、更に推論の効率的な計算に利用することができる。
【００５３】
図２は、ＮＮ表現１１０を生成する装置１００を示している。装置１００は、量子化パラメーター１４２を決定（１４０）することと、ＮＮパラメーター１２０の量子化値１５２を決定（１５０）することとによって、ＮＮパラメーター１２０を量子化された値１３０に量子化するように構成される。量子化値１５２は、量子化パラメーター１４２に基づいて決定（１５０）することができる。量子化パラメーター１４２の決定（１４０）は、量子化パラメーター決定器によって実行することができる。量子化値１５２の決定（１５０）は、量子化値決定器によって実行することができる。
【００５４】
量子化パラメーター１４２が決定（１４０）され、それにより、量子化パラメーター１４２から、乗数１４４及びビットシフト数１４６が導出可能である。量子化パラメーター１４２の決定（１４０）において、装置１００は、例えば、乗数１４４及びビットシフト数１４６が、決定された量子化パラメーター１４２から導出可能であるかどうかを既にチェックしている場合がある。
【００５５】
任意選択で、装置１００は、量子化パラメーター１４２から乗数１４４を導出し、量子化パラメーター１４２からビットシフト数１４６を導出して、例えば装置１００による量子化された値１３０の決定を可能にするように構成することができる。しかし、量子化された値１３０は量子化パラメーター１４２及び量子化値１５２によって表すことができるので、これは必要ではない。装置１００は、量子化された値１３０を明示的に決定する必要はない。
【００５６】
一実施形態によれば、生成されたＮＮ表現１１０は、決定された量子化パラメーター１４２及び決定された量子化値１５２を含むことができ、それにより、ＮＮパラメーター１２０、すなわちＮＮパラメーター１２０の量子化された値１３０は、ＮＮ表現１１０から導出可能である。例えば、装置１００は、量子化パラメーター１４２及び量子化値１５２をＮＮ表現１１０に符号化するように構成されてもよい。
【００５７】

乗数１４４は、量子化パラメーター１４２によって導出される被除数と、精度パラメーターｋ１４５によって導出される除数との間の除算の剰余に基づいて、量子化パラメーター１４２から導出可能である。

【００５８】

ビットシフト数１４６は、除算の商の丸めに基づいて、すなわち、量子化パラメーター１４２によって導出された被除数と精度パラメーターｋ１４５によって導出された除数との間の除算の商の丸めに基づいて、量子化パラメーター１４２から導出可能である。

【００５９】

量子化パラメーター１４２の決定（１４０）が実行され、それにより、ＮＮパラメーター１２０の量子化された値１３０は、量子化値１５２と乗数１４４に依存する因数１４８との間の積であって、ビットシフト数１４６に依存するビット数だけビットシフトされた積に対応する。ＮＮパラメーター１２０の量子化された値１３０は、例えば、少なくとも、シフトの場合には符号の別個の処理を伴う量子化された値の絶対値に関して、又は更には、積、その因数、及びシフトに対してそれぞれ２の補数表現及び２の補数演算を使用する場合等、絶対値と符号との両方に関して、積に対応する。これは、ユニット１５０に例示的かつ概略的に示されている。

【００６０】

一実施形態によれば、装置１００は、ＮＮパラメーターのための浮動小数点表現を使用してＮＮ２０を訓練することによって、かつ量子化誤差を低減することを目的とする反復最適化方式によってＮＮパラメーターのための量子化パラメーター１４２及び量子化値１５２を決定することによって、ＮＮパラメーター、例えばＮＮパラメーター１２０の量子化された値１３０を提供するように構成される。

【００６１】

装置１００とは別に、図１は、ＮＮ表現１１０を定義するデジタルデータ２００と、ＮＮ表現１１０からＮＮパラメーター、すなわちＮＮパラメーター１２０の量子化された値１３０を導出する装置３００とを示している。デジタルデータ２００及び装置３００がＮＮパラメーターの元の値を見ることがないという事実から、量子化された値１３０は、この文脈ではＮＮパラメーターの値として理解される。このため、ＮＮパラメーターは、デジタルデータ２００及び装置３００の以下の説明のために、１３０として示される。本明細書で説明されるＮＮパラメーターは、ＮＮパラメーターに割り当てられた元の値１２０によって、又は元の値１２０に基づいて決定された量子化された値１３０によって表され得ることが明らかである。したがって、以下では、ＮＮパラメーターを１２０／１３０として示す。これは、例えば、ＮＮパラメーターが元の値１２０と量子化された値１３０のどちらで表されても一般的に適用できる特徴を説明するものである。

【００６２】

デジタルデータ２００は、ＮＮ表現１１０を定義し、ＮＮ表現１１０は、ＮＮパラメーター１３０を表すために、量子化パラメーター１４２及び量子化値１５２を含んでおり、それにより、量子化パラメーター１４２から、量子化パラメーター１４２によって導出された被除数と精度パラメーターｋ１４５によって導出された除数との間の除算の剰余に基づいて乗数１４４が導出可能であり、かつ、量子化パラメーター１４２から、除算の商の丸めに基づいてビットシフト数１４６が導出可能である。ＮＮ表現１１０は、量子化パラメーター１４２及び量子化値１５２を含んでおり、それにより、ＮＮパラメーター１３０は、量子化値１５２と乗数１４４に依存する因数１４８との間の積であって、ビットシフト数１４６に依存するビット数だけビットシフトされた積に対応する。

【００６３】

ＮＮ表現１１０からＮＮパラメーター１３０を導出する装置３００は、例えば、量子化パラメーター導出ユニット３１０を使用して、ＮＮ表現１１０から量子化パラメーター１４２を導出し、例えば、量子化値導出ユニット３２０を使用して、ＮＮ表現１１０から量子化値１５２を導出するように構成される。加えて、装置３００は、量子化パラメーター１４２から、乗数１４４及びビットシフト数１４６を導出するように構成される。装置３

００は、量子化パラメーター１４２によって導出された被除数と精度パラメーター１４５によって導出された除数との間の除算の剰余に基づいて乗数１４４を導出し、除算の商の丸めに基づいてビットシフト数１４６を導出するように構成される。乗数１４４の導出は、乗数導出ユニット３３０を使用して実行されてもよく、ビットシフト数１４６の導出は、ビットシフト数導出ユニット３４０を使用して実行されてもよい。ＮＮパラメーター１３０は、量子化値１５２と乗数１４４に依存する因数１４８との間の積であって、ビットシフト数１４６に依存するビット数だけビットシフトされた積に対応する（図２の装置１００及びユニット１５０についての上記の対応する説明を参照）。ＮＮパラメーター１３０は、例えば、ＮＮパラメーター導出ユニット３５０を用いて導出されてもよい。ＮＮパラメーター導出ユニット３５０は、装置１００のオプションのユニット１５０と同じ特徴及び／又は機能を備えてもよい。

【００６４】

以下では、装置１００及び装置３００の両方に適用可能な実施形態及び例が提示される。

【００６５】

一実施形態によれば、ＮＮパラメーター１２０／１３０は、重みパラメーター、バッチノルムパラメーター、及びバイアスのうちの１つである。重みパラメーター、例えば、Ｗの成分ｗは、一対のニューロンの間のニューロン間活性化フィードフォワードを重み付けするために使用可能であり得るか、又は代替的に言えば、第１のニューロンと第２のニューロンとを接続するエッジに関係し、第２のニューロンのためのインバウンド活性化の和において第１のニューロンの活性化のフォワーディングを重み付けする重みを表し得る。バッチノルムパラメーター、例えば、μ、σ^2、γ、βは、ニューラルネットワーク層のアフィン変換をパラメーター化するために使用可能とすることができ、バイアス、例えば、b_iの成分は、所定のニューラルネットワークニューロンのためのインバウンドニューロン間活性化フィードフォワードの和にバイアスをかけるために使用可能であり得る。

【００６６】

一実施形態によれば、ＮＮパラメーター１２０／１３０は、例えば図１に示すように、ＮＮの複数のニューロン間活性化フィードフォワード１２２のうちの単一のニューロン間活性化フィードフォワード12_i、例えばＷの成分ｗに関して、ＮＮ２０をパラメーター化する。装置１００／装置３００は、複数のニューロン間活性化フィードフォワード１２２の各々について、対応するＮＮパラメーター１２０／１３０をＮＮ表現１１０に符号化／ＮＮ表現１１０から導出するように構成される。対応するＮＮパラメーター１３０は、ＮＮ表現１１０に含まれる。この場合、装置１００は、複数のニューロン間活性化フィードフォワード１２２の各々について、それぞれのニューロン間活性化フィードフォワード12_iに関連付けられた関連量子化パラメーター１４２と、それぞれのニューロン間活性化フィードフォワード12_iに関連付けられた関連量子化値１５２とを決定（１４０）することによって、対応するＮＮパラメーター１２０を対応する量子化された値１３０に量子化するように構成され得る。関連量子化パラメーター１４２の決定（１４０）は、関連量子化パラメーター１４２から、関連量子化パラメーター１４２によって導出された被除数と、それぞれのニューロン間活性化フィードフォワード12_iに関連付けられた関連精度パラメーター１４５によって導出された除数との間の除算の剰余に基づいてそれぞれのニューロン間活性化フィードフォワード12_iに関連付けられた関連乗数１４４と、除算の商の丸めに基づいてそれぞれのニューロン間活性化フィードフォワード12_iに関連付けられた関連ビットシフト数１４６と、を導出することができるように実行される。この場合の対応する装置３００は、複数のニューロン間活性化フィードフォワード１２２の各々について、ＮＮ表現１１０からそれぞれのニューロン間活性化フィードフォワード12_iに関連付けられた関連量子化パラメーター１４２を導出（３１０）し、ＮＮ表現１１０からそれぞれのニューロン間活性化フィードフォワード12_iに関連付けられた関連量子化値１５２を導出（３２０）するように構成される。導出（３１０及び３２０）は、例えばＮＮ表現１１０から復号化することによって実行されてもよく、すなわちエッジごとに１つが復号化されてもよい。加えて、装置３００は、複数のニューロン間活性化フィード

フォワード１２２の各々について、関連量子化パラメーター１４２から、関連量子化パラメーター１４２によって導出された被除数と、それぞれのニューロン間活性化フィードフォワード１２_{i}に関連付けられた関連精度パラメーター１４５によって導出された除数との間の除算の剰余に基づいて、それぞれのニューロン間活性化フィードフォワード１２_{i}に関連付けられた関連乗数１４４と、除算の商の丸めに基づいて、それぞれのニューロン間活性化フィードフォワード１２_{i}に関連付けられた関連ビットシフト数１４６と、を導出するように構成される（３３０及び３４０参照）。導出（３３０及び３４０）は、例えばＮＮ表現１１０から復号化することによって実行されてもよく、すなわちエッジごとに１つが復号化されてもよい。

【００６７】

別の実施形態によれば、装置１００／装置３００は、ＮＮ２０の複数のニューロン間活性化フィードフォワード１２２をニューロン間活性化フィードフォワードのサブグループ１２２ａ、１２２ｂに細分するように構成され、それにより、各サブグループは、ＮＮの関連する一対のＮＮ層に関連付けられ、関連する一対のＮＮ層の間のニューロン間活性化フィードフォワードを含み、関連する一対の層以外の更なる一対のＮＮ層の間のニューロン間活性化フィードフォワードを除外し、２つ以上のサブグループが所定のＮＮ層に関連付けられる（例えば図３を参照）。サブグループ１２２ａは、例えば、ＮＮ２０の関連する一対のＮＮ層１１４及び１１６_{1}に関連付けられ、関連する一対のＮＮ層１１４及び１１６_{1}の間のニューロン間活性化フィードフォワードを含み、関連する一対の層１１４及び１１６_{1}以外の更なる一対のＮＮ層の間、例えば更なる一対のＮＮ層１１６_{1}及び１１６_{2}の間のニューロン間活性化フィードフォワードを除外する。サブグループ１２２ａ及び１２２ｂは、層１１６_{1}に関連付けられる。ＮＮ２０の複数のニューロン間活性化フィードフォワード１２２の細分化は、例えば、ＮＮ２０内の各エッジ／重み１２のインデックスによって、又は他の形態で各層の対の間のエッジ１２をセグメント化することによって実行されてもよい。ＮＮパラメーター１２０／１３０は、ＮＮ２の複数のニューロン間活性化フィードフォワード１２２のうちの単一のニューロン間活性化フィードフォワード１２_{i}に関してＮＮ２０をパラメーター化する。複数のニューロン間活性化フィードフォワード１２２の各々について、対応するＮＮパラメーター１２０／１３０がＮＮ表現１１０に含まれる。装置３００は、例えば、ＮＮ表現から復号化することによって、すなわち、エッジサブグループごとに１つのサブグループを復号化することによって、複数のニューロン間活性化フィードフォワード１２２の各々について、ＮＮ表現１１０から対応するＮＮパラメーター１２０／１３０を導出するように構成される。装置１００／装置３００は、ニューロン間活性化フィードフォワードのサブグループ１２２ａ、１２２ｂごとに、それぞれのサブグループ１２２ａ又は１２２ｂに関連付けられた関連量子化パラメーター１４２を決定（１４０）／導出（３１０）するように構成される。量子化パラメーター１４２は、それぞれのサブグループ１２２ａ又は１２２ｂに関連付けられた関連乗数１４４が、関連量子化パラメーター１４２によって導出された被除数とそれぞれのサブグループに関連付けられた関連精度パラメーター１４５によって導出された除数との間の除算の剰余に基づいて量子化パラメーター１４２から導出可能であるように、装置１００によって決定（１４０）され、量子化パラメーター１４２は、それぞれのサブグループ１２２ａ又は１２２ｂに関連付けられた関連ビットシフト数１４６が、除算の商の丸めに基づいて量子化パラメーター１４２から導出可能であるように、装置１００によって決定（１４０）される。装置３００は、ＮＮ表現１１０から関連乗数１４４及び関連ビットシフト数１４６を導出するように構成される。装置１００／装置３００は、複数のニューロン間活性化フィードフォワード１２２の各々について、ＮＮ表現１１０からそれぞれのニューロン間活性化フィードフォワード１２_{i}に関連付けられた関連量子化値１５２を決定（１５０）／導出（３２０）する（例えば、ＮＮ表現１１０から復号化することによって、すなわち、エッジごとに１つが復号化されることによって、導出（３２０）する）ように構成される。それぞれのニューロン間活性化フィードフォワード１２_{i}の対応するＮＮパラメーター１２０／１３０は、関連量子化値１４２と、それぞれのニューロン間活性化フィードフォ

ワード１２$_i$が含まれるサブグループ、例えば１２２ａ又は１２２ｂに関連付けられた関連乗数１４４に依存する因数１４８との間の積であって、それぞれのニューロン間活性化フィードフォワード１２$_i$が含まれるサブグループ、例えば１２２ａ又は１２２ｂの関連ビットシフト数１４６に依存するビット数だけビットシフトされた積に対応する。

【００６８】

関連精度パラメーター１４５は、例えば、ＮＮ２０にわたって、又は各ＮＮ層１１４、１１６$_1$、及び１１６$_2$内でグローバルに等しく評価される。任意選択で、装置１００／装置３００は、関連精度パラメーター１４５をＮＮ表現１１０に符号化／ＮＮ表現１１０から導出するように構成される。

【００６９】

一実施形態によれば、装置１００／装置３００は、コンテキスト適応型二値算術符号化／復号化を使用することによって、又は量子化パラメーター１４２を表すビットをＮＮ表現１１０に直接書き込む／ＮＮ表現１１０から読み出すことによって、又は装置１００／装置３００のコンテキスト適応型二値エンコーダー／デコーダーの等確率バイパスモードを介して量子化パラメーター１４２を表すビットをＮＮ表現１１０から符号化／導出することによって、量子化パラメーター１４２をＮＮ表現１１０に符号化／ＮＮ表現１１０から導出するように構成される。装置１００／装置３００は、二値化方式を使用してビンストリングを二値化／非二値化することによってＮＮ表現１１０から量子化パラメーター１４２を導出するように構成され得る。二値化方式は、例えば、指数ゴロム符号である。

【００７０】

一実施形態によれば、装置１００は、量子化パラメーター１４２を決定（１４０）し、それを固定小数点表現、例えば２の補数表現の形態でＮＮ表現１１０に符号化するように構成される。装置３００は、固定小数点表現、例えば２の補数表現の形態でＮＮ表現１１０から量子化パラメーター１４２を導出（３１０）するように構成されてもよい。任意選択で、精度パラメーター１４５は2^tであり、固定小数点表現、例えば２の補数表現のビット長は、ＮＮ２０に対して一定になるように設定されるか、又はＮＮ２０に対して一定である基底ビット長とｔとの和になるように設定される。

【００７１】

一実施形態によれば、装置１００／装置３００は、整数値シンタックス要素として量子化パラメーター１４２をＮＮ表現１１０に符号化／ＮＮ表現１１０から導出するように構成される。

【００７２】

一実施形態によれば、装置１００は、量子化値１５２を決定し、それを固定小数点表現、例えば２の補数表現の形態でＮＮ表現１１０に符号化するように構成される。装置３００は、固定小数点表現、例えば２の補数表現の形態でＮＮ表現１１０から量子化値１５２を導出（３２０）するように構成されてもよい。

【００７３】

一実施形態によれば、装置１００／装置３００は、二値化方式に従って量子化値１５２をビンストリングに二値化／ビンストリングから非二値化し、コンテキスト適応算術符号化／復号化を使用してビンストリングのビットを符号化／復号化することによって、量子化値１５２をＮＮ表現１１０に符号化／ＮＮ表現１１０から導出するように構成される。

【００７４】

一実施形態によれば、装置１００／装置３００は、二値化方式に従って量子化値１５２をビンストリングに二値化／ビンストリングから非二値化し、コンテキスト適応算術符号化／復号化を使用してビンストリングの第１のビットを符号化／復号化し、等確率バイパスモードを使用してビンストリングの第２のビットを符号化／復号化することによって、量子化値１５２をＮＮ表現１１０に符号化／ＮＮ表現１１０から復号化するように構成される。

【００７５】

一実施形態によれば、量子化ステップサイズΔ１４９は、装置１００及び／又は装置３

００によって、量子化パラメーターＱＰ１４２で示される符号付き整数及び正の整数パラメーターｋ、すなわち精度パラメーター１４５から、以下の式に従って導出することができる。

ｍｕｌ＝ｋ＋ＱＰ％ｋ

【数３８】

$$shift = \lfloor QP/k \rfloor$$

Δ＝（ｍｕｌ／ｋ）・２^{ｓｈｉｆｔ}

【００７６】

乗数１４４はｍｕｌで表され、ビットシフト数１４６はｓｈｉｆｔで表され、因数１４８はｍｕｌ／ｋで表される。

【００７７】

ＮＮパラメーター１３０は、（ｍｕｌ／ｋ）・２$^{ｓｈｉｆｔ・Ｐ}$であり、ここで、Ｐは量子化値１５２である。

【００７８】

フロア演算子

【数３９】

$$\lfloor \ \rfloor$$

及びモジュロ演算子％は、以下のように定義される。

【数４０】

$$\lfloor x \rfloor$$

は、ｘ以下の最大の整数である。

ｘ％ｙは、

【数４１】

$$x - y \cdot \lfloor x/y \rfloor$$

として定義されるモジュロ演算子である。

【００７９】

任意選択で、装置１００及び／又は装置３００は、精度パラメーターｋ１４５をデフォルト値に設定するように構成することができる。

【００８０】

あるいは、装置１００は、任意選択で、自然数又は２の累乗等の精度パラメーターｋ１４５の複数の異なる整数値をテストすることができる。異なる整数値は、例えば、ＮＮ全体に対して、又は各層等のＮＮの各部分に対してテストされ、量子化誤差及びビットレートに関して、例えばそのラグランジュ和に関して最良の精度パラメーターｋ１４５が選択される。装置１００は、例えば、精度パラメーターｋ１４５を決定して、例えば決定（１４０）において、乗数１４４及びビットシフト数１４６が量子化パラメーター１４２から導出可能であるかどうかをチェックするように構成されてもよい。任意選択で、装置１００によって選択された精度パラメーターｋ１４５は、ＮＮ表現１１０においてシグナリングされ、例えば、ＮＮ表現１１０に符号化される。装置３００は、例えば、ＮＮ表現１１

０から精度パラメーターｋ１４５を導出するように構成される。
【００８１】
一実施形態によれば、精度パラメーター１４５は、２の累乗である。
【００８２】
一実施形態によれば、装置１００／装置３００は、精度パラメーター１４５を表すビットを直接ＮＮ表現１１０に書き込む／ＮＮ表現１１０から読み出すことによって、又は装置１００／装置３００のコンテキスト適応型二値エンコーダー／デコーダーの等確率バイパスモードを介して精度パラメーター１４５を表すビットをＮＮ表現１１０に／ＮＮ表現１１０から導出することによって、精度パラメーター１４５をＮＮ表現１１０に符号化／ＮＮ表現１１０から導出するように構成される。
【００８３】
ビットストリーム、例えばデジタルデータ２００において３２ビット浮動小数点値をシグナリングする代わりに、パラメーターＱＰ１４２及びｋ１４５のみがシグナリングされる必要がある。一部の適用例では、ビットストリーム中でＱＰ１４２をシグナリングし、ｋ１４５を何らかの固定値に設定することで十分な場合もある。
【００８４】
好ましい実施形態において、パラメーターＱＰ’＝ＱＰ－QP_0が、ＱＰ１４２の代わりにビットストリーム内でシグナリングされ、ここで、パラメーターQP_0は、所定の定数値である。換言すれば、一実施形態によれば、装置１００／装置３００は、関連量子化パラメーターＱＰ１４２を、参照量子化パラメーターQP_0との差分の形態で、ＮＮ表現１１０に符号化／ＮＮ表現１１０から導出するように構成される。
【００８５】
別の好ましい実施形態において、ｋ１４５は2^tに設定される。このようにして、Δ１４９の算出は、除算を含まずに以下のように実行することができる。
$\Delta = mul \cdot 2^{shift-t}$
【００８６】
これにより、一部の計算を、以下に例示されるように、浮動小数点領域の代わりに整数領域で実行することが可能となる。
【００８７】
図４は、ＮＮ２０を使用して推論を実行するデバイス４００を概略的に示している。デバイス４００は、ＮＮ２０をパラメーター化するように構成されたＮＮパラメトライザー４１０を備える。ＮＮパラメトライザー４１０は、ＮＮ表現１１０からＮＮパラメーター１３０を導出する装置３００を備える。ＮＮパラメーター１３０を導出する装置３００は、図２の装置３００に関して説明したものと同じ又は同様の特徴を備えることができる。装置３００は、ＮＮパラメーター導出ユニットとして理解され得る。加えて、デバイス４００は、ＮＮ２０を使用して、例えば、ＮＮパラメトライザー４１０によって決定されたＮＮ２０のパラメーター化４５０を使用して、ＮＮ入力４４０に基づいて推論出力４３０を計算するように構成された計算ユニット４２０を備える。
【００８８】
例１：
一実施形態によれば、ＮＮパラメトライザー４１０は、装置３００を介して、第１のＮＮパラメーター及び第２のＮＮパラメーターのうちの少なくとも１つを導出するように構成され、それにより、第１のＮＮパラメーターは、第１の量子化値と第１の因数との間の積であって、第１のビット数だけビットシフトされた積に対応し、第２のＮＮパラメーターは、第２の量子化値と第２の因数との間の積であって、第２のビット数だけビットシフトされた積に対応する。
【００８９】
第１の量子化値及び第２の量子化値は両方とも、図２において１５２で示される量子化値を表す。第１の因数及び第２の因数は両方とも、図２において１４８で示される因数を表す。

【００９０】
例えば、ｔ＝２とし、ｋ＝２^{t}とし、ＱＰ$_{a}$で示される第１のＱＰ、すなわち第１の量子化パラメーター１４２、関連するｓｈｉｆｔ$_{a}$、すなわち第１のビットシフト数１４６、ｍｕｌ$_{a}$、すなわち第１の乗数１４４、及びΔ$_{a}$、すなわち第１の量子化ステップサイズ１４９を定義する。
【００９１】
さらに、ＱＰ$_{b}$で示される第２のＱＰ、すなわち第２の量子化パラメーター１４２、関連するｓｈｉｆｔ$_{b}$、すなわち第２のビットシフト数１４６、ｍｕｌ$_{b}$、すなわち第２の乗数１４４、及びΔ$_{b}$、すなわち第２の量子化ステップサイズ１４９を定義する。
【００９２】
「第１の」パラメーター及び「第２の」パラメーターは、この文脈では同じ参照番号で示されているが、それらが異なる値を有し得ることは明らかである。それらは、それらが図２に示されるどの特徴に属するかを明確にするために、同じ参照番号で示されているのみである。
【００９３】
Ｃ＝Δ$_{a}$・Ｃ$_{a}$が成り立つ第１の量子化行列Ｃ$_{a}$を考える。
Ｄ＝Δ$_{b}$・Ｄ$_{b}$が成り立つ第２の量子化行列Ｄ$_{b}$を考える。
すなわち、Ｃ$_{a}$は、ＱＰ$_{a}$を使用して量子化され、Ｄ$_{b}$は、ＱＰ$_{b}$を使用して量子化されている。
両方の行列は、同じ次元を有する。図２で説明した量子化値１５２は、Ｃ$_{a}$の１つの成分又はＤ$_{b}$の１つの成分を表すことができる。例えば、Ｃ$_{a}$は、複数の第１の量子化値１５２を含むことができ、Ｄ$_{b}$は、複数の第２の量子化値１５２を含むことができる。
【００９４】
さらに、和Ｃ＋Ｄが以下のように算出されると仮定する。
【数４２】

$$C+D=\Delta_a\cdot C_a+\Delta_b\cdot D_b=2^{shift_a-2}\cdot mul_a\cdot C_a+2^{shift_b-2}\cdot mul_b\cdot D_b$$
$$=2^{shift_a-2}\cdot\left(mul_a\cdot C_a+2^{shift_b-shift_a}\cdot mul_b\cdot D_b\right)$$

【００９５】
デバイス４００は、第１のＮＮパラメーターＣの第１の量子化値Ｃ$_{a}$によって形成され、第１の乗数ｍｕｌ$_{a}$で重み付けされた第１の加数、例えばｍｕｌ$_{a}$・Ｃ$_{a}$と、第２のＮＮパラメーターＤの第２の量子化値Ｄ$_{b}$によって形成され、第２の乗数ｍｕｌ$_{b}$で重み付けされ、第１のビット数及び第２のビット数の差、例えば、
【数４３】

$$2^{shift_b-shift_a}$$

だけビットシフトされた第２の加数、例えば、
【数４４】

$$2^{shift_b-shift_a}\cdot mul_b\cdot D_b$$

との間の和を形成することと、第１の加数及び第２の加数の和に、第１のビット数及び第２のビット数の一方に依存する、例えば第１のビットシフト数ｓｈｉｆｔ$_{a}$又は第２のビットシフト数ｓｈｉｆｔ$_{b}$に依存するビット数だけビットシフト

【数４５】

$$2^{shift_a-2}$$

を施すことと、によって、第１のＮＮパラメーターＣ及び第２のＮＮパラメーターＤを加算してＮＮ２０の最終ＮＮパラメーターを生成するように構成される。

【００９６】

任意選択で、この算出／計算は、計算ユニット４２０によって実行することができる。この場合、計算ユニット４２０は、上述したように、第１のＮＮパラメーターＣと第２のＮＮパラメーターＤとを加算してＮＮ２０の最終的なＮＮパラメーターを生成するように構成されている。

【００９７】

式から分かるように、浮動小数点演算を必要とし得るＣ及びＤを導出する必要はない。その代わりに、C_aの要素、すなわち第１の量子化値１５２は、単にmul_a、すなわち第１の乗数１４４と乗算され、D_bの要素、すなわち第２の量子化値１５２は、mul_b、すなわち第２の乗数１４４と乗算され、因数

【数４６】

$$2^{shift_b-shift_a}$$

は、C_aの第１の量子化値１５２、すなわちC_aの成分に関連する第１のビットシフト数$shift_a$１４６と、D_bの第２の量子化値１５２、すなわちD_bの成分に関連する第２のビットシフト数$shift_b$１４６とに依存する単純なビットシフト演算として実施される。ｔ＝２であるので、整数変数mul_a及びmul_bは両方とも値４、５、６、及び７のうちの１つであることに留意されたい。かかる小さい数を有する整数乗算は、ハードウェア又はソフトウェア実施態様において非常に効率的に実施され得る。

【００９８】

一実施形態によれば、第１のＮＮパラメーターはＮＮ２０のベース層表現を表し、第２のＮＮパラメーターはＮＮ２０のエンハンスメント層表現を表す。代替的に、第１のＮＮパラメーターは、例えば、ＮＮ２０の現在の表現を表し、第２のＮＮパラメーターは、現在のＮＮ表現の更新、すなわち、ＮＮ２０の現在の表現の更新を表す。代替的に、例えば、第１のＮＮパラメーターは、所定のニューラルネットワークニューロン１０のインバウンドニューロン間活性化フィードフォワードの和にバイアスをかけるバイアス、すなわちb_iの成分を表し、第２のＮＮパラメーターは、ニューラルネットワーク層１１４、１１$6_1$又は１１$6_2$のアフィン変換をパラメーター化するバッチノルムパラメーター、すなわちμ、σ^2、γ又はβ、例えばｂ＋μを表す。

【００９９】

例２：

一実施形態によれば、ＮＮパラメトライザー４１０は、装置３００を介して、第３のＮＮパラメーター及び第４のＮＮパラメーターのうちの少なくとも１つを導出するように構成され、それにより、第３のＮＮパラメーターは、第３の量子化値と第３の因数との間の積であって、第３のビット数だけビットシフトされた積に対応し、第４のＮＮパラメーターは、第４の量子化値と第４の因数との間の積であって、第４のビット数だけビットシフトされた積に対応する。

【０１００】

第３の量子化値及び第４の量子化値はともに、図２において１５２で示される量子化値を表す。第３の因数及び第４の因数は、両方とも、図２において１４８で示される因数を表す。

【０１０１】
例えば、ｔ＝２とし、ｋ＝２^{t}とし、第１のＱＰ、例えば、ＱＰ$_{a}$で示される第３の量子化パラメーター１４２、関連するｓｈｉｆｔ$_{a}$、すなわち第３のビットシフト数１４６、ｍｕｌ$_{a}$、すなわち第３の乗数１４４、及びΔ$_{a}$、すなわち第３の量子化ステップサイズ１４９を定義する。
【０１０２】
さらに、第２のＱＰ、例えば、ＱＰ$_{b}$で示される第４の量子化パラメーター１４２、関連するｓｈｉｆｔ$_{b}$、すなわち第４のビットシフト数１４６、ｍｕｌ$_{b}$、すなわち第４の乗数１４４、及びΔ$_{b}$、すなわち第４の量子化ステップサイズ１４９を定義する。
【０１０３】
「第３の」パラメーター及び「第４の」パラメーターは、この文脈では同じ参照番号で示されているが、それらが異なる値を有し得ることは明らかである。それらは、それらが図２に示されるどの特徴に属するかを明確にするために、同じ参照番号で示されているのみである。デバイス４００は、第３のパラメーター及び／又は第４のパラメーターのみ、又は、上記の例１で説明したように、更に第１のパラメーター及び／又は第２のパラメーターを導出するように構成されてもよい。
【０１０４】
Ｗ＝Δ$_{a}$・Ｗ$_{a}$が成り立つ量子化行列Ｗ$_{a}$を考える。
γ＝Δ$_{b}$・γ$_{b}$が成り立つ量子化された転置ベクトルγ$_{b}$を考える。
すなわち、Ｗ$_{a}$は、ＱＰ$_{a}$を使用して量子化され、γ$_{b}$は、ＱＰ$_{b}$を使用して量子化されている。
図２で説明した量子化値１５２は、Ｗ$_{a}$の１つの成分又はγ$_{b}$の１つの成分を表すことができる。例えば、Ｗａは複数の量子化値１５２を含むことができ、γ$_{b}$は複数の量子化値１５２を含むことができる。
【０１０５】
さらに、要素ごとの積Ｗ・γが以下のように算出されると仮定する。
【数４７】

$$W \cdot \gamma = \Delta_a \cdot W_a \cdot \Delta_b \cdot \gamma_b = 2^{shift_a - 2} \cdot mul_a \cdot W_a \cdot 2^{shift_b - 2} \cdot mul_b \cdot \gamma_b$$

$$= 2^{shift_a + shift_b - 4} \cdot mul_a \cdot mul_b \cdot W_a \cdot \gamma_b$$

【０１０６】
この算出／計算は、例えば、第３のＮＮパラメーターＷ及び第４のＮＮパラメーターγを乗算に施して、第３のＮＮパラメーターＷの第３の量子化値Ｗ$_{a}$によって形成される第１の因数と、第３の乗数ｍｕｌ$_{a}$によって形成される第２の因数と、第４のＮＮパラメーターγの第４の量子化値γ$_{b}$によって形成される第３の因数と、第４の乗数ｍｕｌ$_{b}$によって形成される第４の因数との積であって、第３のビット数ｓｈｉｆｔ$_{a}$によって形成される第１の加数と第４のビット数ｓｈｉｆｔ$_{b}$によって形成される第２の加数とを含む和に対応するビット数、例えば
【数４８】

$$2^{shift_a + shift_b - 4}$$

だけビットシフトされた積を形成することによって積を得ることによって、演算ユニット４２０によって実行され得る。

【０１０７】
式から分かるように、浮動小数点演算を必要とし得るW及びγを導出する必要はない。代わりに、計算$mul_a \cdot mul_b \cdot W_a \cdot \gamma_b$は、整数乗算のみを伴い、
【数４９】

$$2^{shift_a+shift_b-4}$$

との後続の乗算は、ビットシフトとして実施することができる。ｔ＝２であるので、整数変数mul_a及びmul_bは両方とも値４、５、６、及び７のうちの１つであることに留意されたい。かかる小さい数を有する整数乗算は、ハードウェア又はソフトウェア実施態様において非常に効率的に実施され得る。
【０１０８】
一実施形態によれば、第３のＮＮパラメーターは、第１のＮＮ層１１４の第１のニューロン１０_1から第２のＮＮ層１１６_2の第２のニューロン１０_2へのニューロン間活性化フィードフォワードを重み付けする重みパラメーター、例えば、Wの成分wを表すか、又は代替的に、第３のＮＮパラメーターは、第１のニューロン１０_1と第２のニューロン１０_2とを接続するエッジ１２_iに関連し、第２のニューロン１０_2のインバウンド活性化の和における第１のニューロン１０_1の活性化の転送を重み付けする重みを表す。
【０１０９】
第４のＮＮパラメーターは、例えば、バッチノルムパラメーター、例えば、μ、σ^2、γ又はβを表す。バッチノルムパラメーターは、例えば、第２のＮＮ層１１６_1に対する第１のニューロン１０_1の活性化フィードフォワード増幅を調整するものであり、例えばγである。
【０１１０】
入力Ｘの量子化
一実施形態によれば、デバイス４００は、例えば装置３００を使用して、活性化を量子化された値、例えばＸ’’に量子化することによって、活性化について第５の量子化パラメーターＱＰ、すなわち量子化パラメーター１４２、及び第５の量子化値、例えばＸ’、すなわち量子化値１５２を決定することによって、ＮＮ入力Ｘ４４０を量子化するように構成され、それにより、第５の量子化パラメーターＱＰからの第５の乗数ｍｕｌ、すなわち乗数１４４の導出は、第５の量子化パラメーターによって導出された被除数と、活性化に関連する精度パラメーターｋ、すなわち精度パラメーター１４５によって導出された除数との間の除算の剰余と、除算の商の丸めに基づく第５のビットシフト数ｓｈｉｆｔ、すなわちビットシフト数１４６とに基づいて、第５の量子化値と第５の乗数に依存する因数ｍｕｌ／ｋ、すなわち因数１４８との間の積に対応する量子化された値が、第５のビットシフト数に依存する第５のビット数だけビットシフトされることをもたらす。
【０１１１】
好ましい実施形態において、バイアス層又はバッチ正規化層の入力Ｘ４４０も、本発明の量子化方法を使用して量子化される。図２の装置１００の説明を参照。すなわち、Ｘ’’＝Δ・Ｘ’＝$mul \cdot 2^{shift-t} \cdot X'$が成り立つように、量子化パラメーターＱＰ及び関連する変数、関連するｓｈｉｆｔ、ｍｕｌ、及びΔ（ｔ＝２及び$k = 2^t$）が選択され、ＸがＸ’に量子化される。次に、バイアス層又はバッチノルム層を実行するためにＸを使用する代わりに、Ｘ’’が入力として使用される。Ｘ’は、通常、Ｘよりも極めて少ないビット／要素で表すことができ、これは、効率的なハードウェア又はソフトウェア実施態様のための別の利点であることに留意されたい。
【０１１２】
一実施形態によれば、ＮＮパラメトライザー４１０は、装置３００を介して第６のＮＮパラメーターを導出するように構成され、それにより、第６のＮＮパラメーターは、第６の量子化値と第６の因数ｍｕｌ／ｋとの積であって、第６のビット数だけビットシフトさ

れた積に対する。デバイス４００は、第６のＮＮパラメーター及び活性化に乗算を施して、第６のＮＮパラメーターのための第６の量子化値によって形成される第１の因数と、第６の乗数によって形成される第２の因数と、第５の量子化値によって形成される第３の因数と、第５の乗数によって形成される第４の因数との積を形成することによって積であって、第６のビット数によって形成される第１の加数と第４のビット数によって形成される第２の加数とを含む和に対応するビット数だけビットシフトされた積を生成するように構成される。

【０１１３】

一実施形態によれば、第６のＮＮパラメーターは、入力４４０を重み付けする重みパラメーターＷを表し、それにより、積Ｗ＊Ｘを算出／計算することができる。

【０１１４】

図２に戻って参照すると、以下では、装置１００及び／又は装置３００の更なる任意選択の特徴が説明される。

【０１１５】

パラメーターＱＰの効率的な符号化及び復号化

好ましい実施形態において、パラメーターＱＰ、すなわち量子化パラメーター１４２は、以下の定義に従って、Ｋ次の符号付き指数ゴロム符号を使用して、装置１００／装置３００によってビットストリーム２００内に符号化／ビットストリーム２００から復号化される。

【０１１６】

別の好ましい実施形態は、次数Ｋが０に設定された先の好ましい実施形態と同じである。

【０１１７】

符号なし整数の指数ゴロム符号

符号なし整数の符号なし指数ゴロム符号は、高効率ビデオ符号化（ＨＥＶＣ：High Efficiency Video Coding）規格において定義されているシンタックス要素ｕｅ（ｖ）の復号化仕様に従うものとする。

【０１１８】

この仕様を以下に簡単に説明する。

【０１１９】

次数Ｋの符号なし指数ゴロム符号で符号化された二値表現からの符号なし整数変数「ｄｅｃＮｕｍ」の復号化は、以下の擬似符号に従って定義される。

```
ｌｅａｄｉｎｇＺｅｒｏＢｉｔｓ＝－１
ｆｏｒ（ｂ＝０；！ｂ；ｌｅａｄｉｎｇＺｅｒｏＢｉｔｓ＋＋）
ｂ＝ｒｅａｄ＿ｂｉｔｓ（１）
```

【０１２０】

次に、変数ｃｏｄｅＮｕｍが以下のように割り当てられる。

$$ｄｅｃＮｕｍ＝（２^{ｌｅａｄｉｎｇＺｅｒｏＢｉｔｓ－１}）＊２^{Ｋ}＋ｒｅａｄ＿ｂｉｔｓ（ｌｅａｄｉｎｇＺｅｒｏＢｉｔｓ＋Ｋ）$$

【０１２１】

関数ｒｅａｄ＿ｂｉｔｓ（ｘ）は、ビットストリームからｘビットを読み出し、それらを符号なし整数として返す。読み出されたビットは、最上位ビット（ＭＳＢ）から最下位ビット（ＬＳＢ）に順序付けられる。

【０１２２】

符号付き整数の指数ゴロム符号

符号付き整数の符号なし指数ゴロム符号は、高効率ビデオ符号化（ＨＥＶＣ）規格において定義されているシンタックス要素ｓｅ（ｖ）の復号化仕様に従うものとする。

【０１２３】

この仕様を以下に簡単に説明する。

【０１２４】

符号付き指数ゴロム符号で符号化された二値表現からの符号付き整数「ｓｉｇｎｅｄＤ

ｅｃＮｕｍ」の復号化は、以下の通りである。最初に、符号なし整数が、上記で説明したようにＨＥＶＣのｕｅ（ｖ）シンタックス要素復号化プロセスに従って復号化される。次に、符号なし整数は、以下の式に従って符号付き整数に変換される。

【数５０】

$$signedDecNum = (-1)^{decNum+1} \cdot \lceil decNum/2 \rceil$$

シーリング演算子

【数５１】

$$\lceil x \rceil$$

は、ｘ以上の最小整数を返す。

【０１２５】

更に好ましい実施形態

好ましい実施形態において、パラメーターｋ、すなわち精度パラメーター１４５は、２^{t}に設定され、パラメーターｔは、ｂｉｔｓ＿ｔビットを有する（例えば、ｂｉｔｓ＿ｔ＝３又はｂｉｔｓ＿ｔ＝４を有する）符号なし整数表現を使用して符号化される。

【０１２６】

別の好ましい実施形態において、パラメーターｋ、すなわち精度パラメーター１４５は、２^{t}に設定され、パラメーターｔは、符号なし整数用の指数ゴロム符号を使用して符号化される。

【０１２７】

別の好ましい実施形態において、パラメーターＱＰ、すなわち量子化パラメーター１４２は、符号付き整数用の指数ゴロム符号を使用して符号化される。

【０１２８】

別の好ましい実施形態において、パラメーターｋ、すなわち精度パラメーター１４５は、２^{t}に設定され、パラメーターＱＰは、ｂｉｔｓ＿ｑｐビットを使用して２の補数表現の符号付き整数を使用して符号化される。ｂｉｔｓ＿ｑｐは、例えば、１２又は１３のような一定値に設定されるか、又はｂｉｔｓ＿ｑｐは、ｂｉｔｓ＿ｑｐ０＋ｔに設定され、ｂｉｔｓ＿ｑｐ０は、非ゼロの一定の整数値（例えば、ｂｉｔｓ＿ｑｐ０＝６）である。

【０１２９】

ＣＡＢＡＣ符号化ビットストリーム２００の場合、パラメーターｔ及び／又はＱＰ１４２を表すビットは、（ＣＡＢＡＣのバイパスモードを使用して）バイパスビンとして符号化されるか、又はビットストリーム２００に直接書き込まれるかのいずれかであり得る。

【０１３０】

別の好ましい実施形態において、パラメーターＷ、ｂ、μ、σ^{2}、γ、及びβの各々は、パラメーターの符号化の直前に符号化される個々のＱＰ１４２の値で量子化される。

【０１３１】

別の好ましい実施形態において、第１のＱＰ１４２がビットストリーム２００に符号化され、モデルのパラメーターのサブセットに関連付けられる。このサブセットの各パラメーターｘについて、１つのＱＰオフセットＱＰ$_{x}$がパラメーターごとに符号化され、パラメーターを逆量子化するために使用される有効ＱＰ１４２、すなわちＮＮパラメーター１２０は、ＱＰ＋ＱＰ$_{x}$として与えられる。ＱＰ$_{x}$の二値表現は、好ましくは、ＱＰの二値表現よりも少ないビットを使用する。例えば、ＱＰ$_{x}$は、符号付き整数又は（２の補数表現での）固定数のビットのための指数ゴロムコードを使用して符号化される。

【０１３２】

重みパラメーターの符号化に関する更なる実施形態

図５に示される更なる好ましい実施形態は、重みパラメーターＷ５４５の表現に関する。すなわち、それらをベクトル５４６と行列５４４との合成W→ｓ・Ｗ’として因数分解する。W及びＷ’、すなわち重み行列５４４は、次元ｎ×ｍの行列であり、ｓは、長さｎの転置ベクトル５４６である。ベクトルｓ５４６の各要素は、重み行列Ｗ’５４４の行方向のスケーリングファクターとして使用される。換言すれば、ｓ５４６は、Ｗ’５４４の各列と要素ごとに乗算される。ｓ５４６をローカルスケーリングファクター又はローカルスケール適応（ＬＳＡ：local scale adaptation)と称する。

【０１３３】

図５は、ＮＮ２０を使用して推論を実行するデバイス５００を示している。デバイス５００は、ＮＮ２０を使用してＮＮ入力４４０に基づいて推論出力４３０を計算するように構成される。ＮＮ２０は、一対のＮＮ層１１４及び１１６と、一対のＮＮ層の第１の層１１４からＮＮ層の第２の層１１６へのニューロン間活性化フィードフォワード１２２とを含む。デバイス５００は、例えば、デバイス５００の行列形成ユニット５３０を使用して、第１のＮＮ層１１４のニューラルネットワークニューロン10_1の活性化５２０から行列Ｘ５３２を形成することによって、第１のＮＮ層１１４のニューラルネットワークニューロン10_1の活性化５２０に基づいて第２のＮＮ層１１６のニューラルネットワークニューロン10_2の活性化５１０を計算するように構成される。加えて、デバイス５００は、ｓ・Ｗ’＊Ｘを計算（５４２）することによって、第１のＮＮ層１１４のニューラルネットワークニューロン10_1の活性化５２０に基づいて、第２のＮＮ層１１６のニューラルネットワークニューロン10_2の活性化５１０を計算するように構成され、ここで、＊は、行列乗算を表し、Ｗ’は、ｎ及びｍ∈Ｎである次元ｎ×ｍの重み行列５４４であり、ｓは、長さｎの転置ベクトル５４６であり、・は、・の一方の側の行列と・の他方の側の転置ベクトルとの間の列に関するアダマール乗算を示す、デバイス５００は、計算５４２を実行するように構成された計算ユニット５４０を備えることができる。

【０１３４】

一実施形態によれば、転置ベクトルｓ５４６は、Ｗ’５４４を符号化するためのより高い圧縮及び／又はより高い推論忠実度に関してＷ’５４４を最適化した結果である。

【０１３５】

その論理的根拠は、ＬＳＡが重み行列５４４をスケーリングし、それにより、算術符号化方法がより高い符号化利得をもたらし、及び／又はニューラルネットワーク性能結果を増加させ、例えば、より高い精度を達成することである。例えば、Wの量子化後、ｓ５４６は、入力データ４４０、例えばＸ５３２を使用して又は使用せずに、量子化誤差を低減し、それにより量子化されたニューラルネットワークの予測性能を向上させるために適応され得る。

【０１３６】

したがって、ｓ５４６及びＷ’５４４は、異なる量子化パラメーター、すなわち異なるＱＰを有することができる。これは、性能の観点からだけでなく、ハードウェア効率の観点からも有益であり得る。例えば、Ｗ’５４４は、入力Ｘ５３２との内積が８ビット表現で実行されてもよいが、スケーリングファクターｓ５４６との後続の乗算が１６ビットで実行され得るように量子化されてもよい。デバイス５００は、例えば、内積を得るためにｎビット固定小数点演算を使用して行列乗算Ｗ’＊Ｘを計算し、ｍ＞ｎであるｍビット固定小数点演算を使用して内積をｓ５４６と乗算するように構成される。

【０１３７】

しかしながら、Ｗ’５４４及びｓ５４６が両方ともｎビット表現に量子化される場合であっても、同じ推論精度を得るためにＷ５４５を量子化するのに必要なｎよりも小さいｎで十分な場合がある。同様に、ｓ５４６がＷ’５４４よりも少ないビットの表現に量子化された場合、表現の効率に関する利点を更に達成できる場合がある。

【０１３８】

一実施形態によれば、デバイス５００は、ＮＮ表現１１０からＷ’５４４を導出するよう

に構成されたＮＮパラメトライザー、例えば、図４に示すＮＮパラメトライザー４１０を備える。ＮＮパラメトライザーは、ＮＮ表現１１０からＮＮパラメーターを導出する装置、例えば、図４又は図２に示される装置３００を備える。重み行列Ｗ’５４４は、装置３００によって導出されたＮＮパラメーターであり得る。任意選択で、ＮＮパラメトライザー４１０は、Ｗ’５４４に関連するＮＮパラメーターと比較して異なる量子化パラメーター１４２を使用して、ＮＮ表現１１０からｓ５４６を導出するように更に構成される。

【０１３９】

好ましい実施形態において、重み行列Ｗ５４４の符号化は以下の通りである。第１に、ＬＳＡが使用されるかどうかを示すフラグが符号化される。フラグが１である場合、パラメーターｓ５４６及びＷ’５４４は、ＤｅｅｐＣＡＢＡＣのような現行技術水準のパラメーター符号化方式を使用して符号化される。フラグが０である場合、Ｗ５４５が代わりに符号化される。

【０１４０】

別の好ましい実施形態において、前の好ましい実施形態による、異なるＱＰ値がＷ’５４４及びｓ５４６に使用される。

【０１４１】

バッチノルム圧縮

図６に示す一実施形態は、バッチノルム圧縮を改善することに関する。図６は、ＮＮのバッチノルム演算子７１０のＮＮパラメーター６１０、例えば、μ、σ^2、γ、β、及び任意選択でｂをＮＮ表現１１０に符号化する装置６００と、ＮＮ表現１１０からＮＮのバッチノルム演算子７１０のＮＮパラメーター６１０、例えば、γ７２２及びβ７２４及びパラメーター７３２、すなわち、μ、σ^2、及び任意選択でｂを復号化する装置７００とを示している。４つの実施形態が示されており、第１の実施形態は一般的な場合を説明し、他の実施形態は特別な場合を対象とする。

【０１４２】

概して、バッチノルム演算子７１０_1は、

【数５２】

$$\frac{W * X + b - \mu}{\sqrt{\sigma^2 + \epsilon}} \cdot \gamma + \beta$$

として定義することができ、式中、

μ、σ^2、γ、及びβは、バッチノルムパラメーター、例えば、各出力ノードについて１つの成分を含む転置ベクトルであり、

Ｗは、重み行列であり、例えば、その各行は１つの出力ノードに対するものであり、それぞれの行の各成分はＸの１つの行に関連付けられており、

Ｘは、ＮＮ層の活性化から導出される入力行列であり、

ｂは、バイアスを形成する転置ベクトル、例えば、各出力ノードに対して１つの成分を含む転置ベクトルであり、

εは、ゼロ除算回避のための定数であり、

・は、・の一方の側の行列と他方の側の転置ベクトルとの間の列に関するアダマール乗算を示し、

＊は、行列乗算を示す。

【０１４３】

第２の実施形態において、定数εは０であり、それにより、バッチノルム演算子７１０_2は、

【数５３】

$$\frac{W*X+b-\mu}{\sqrt{\sigma^2+\epsilon}}\cdot\gamma+\beta$$

によって定義される。

【０１４４】

第３の実施形態において、バイアスｂは０であり、それにより、バッチノルム演算子７１０$_3$は、

【数５４】

$$\frac{W*X-\mu}{\sqrt{\sigma^2+\epsilon}}\cdot\gamma+\beta$$

によって定義される。

【０１４５】

第４の実施形態において、バイアスｂ及び定数εは０であり、それにより、バッチノルム演算子７１０$_4$は、

【数５５】

$$\frac{W*X-\mu}{\sqrt{\sigma^2}}\cdot\gamma+\beta$$

によって定義される。

【０１４６】

図６では、バッチノルム演算子７１０の一部のパラメーターは、アポストロフィを有しており、アポストロフィなしのパラメーターによって示される元のパラメーター６１０と、アポストロフィありのパラメーターによって示される修正されたパラメーター７２２、７２４、及び７３２との間の区別を可能にする。元のパラメーター６１０又は修正されたパラメーター７２２、７２４及び７３２のいずれかを、上記で定義されたバッチノルム演算子７１０のうちの１つのパラメーターとして使用できることは明らかである。

【０１４７】

装置６００は、パラメーターμ、γ、β、及びσ^{2}又はσを受信するように構成され（６１０$_1$～６１０$_4$参照）、任意選択でｂを受信するように構成される（６１０$_1$及び６１０$_2$参照）。

【０１４８】

第１の実施形態によれば、装置６００は、

【数５６】

$$\beta':=\beta+\frac{(b-\mu)*\gamma}{\sqrt{\sigma^2+\epsilon}}$$

及び

【数５７】

$$\gamma':=\gamma\cdot\frac{\sqrt{\theta+\epsilon}}{\sqrt{\sigma^2+\epsilon}}$$

を計算するように構成される。

【０１４９】
代替の第２の実施形態によれば、装置６００は、
【数５８】

$$\beta' := \beta + \frac{(b-\mu)*\gamma}{\sqrt{\sigma^2}} \text{ and}$$

及び
【数５９】

$$\gamma' := \gamma \cdot \frac{1}{\sqrt{\sigma^2}}.$$

を計算するように構成される。
【０１５０】
代替の第３の実施形態によれば、装置６００は、
【数６０】

$$\beta' := \beta - \frac{\mu*\gamma}{\sqrt{\sigma^2+\epsilon}}$$

及び
【数６１】

$$\gamma' := \gamma \cdot \frac{\sqrt{\theta+\epsilon}}{\sqrt{\sigma^2+\epsilon}}$$

を計算するように構成される。
【０１５１】
代替の第４の実施形態によれば、装置６００は、
【数６２】

$$\beta' := \beta - \frac{\mu * \gamma}{\sqrt{\sigma^2}}$$

及び
【数６３】

$$\gamma' := \gamma \cdot \frac{1}{\sqrt{\sigma^2}}.$$

を計算するように構成される。
【０１５２】
計算されたパラメーターβ’及びγ’は、バッチノルム演算子７１０のＮＮパラメーター

としてＮＮ表現１１０に符号化され、例えば、同じ（β’及びγ’）が、各出力ノードに対して１つの成分を含む転置ベクトルでもあるようにする。

【０１５３】

したがって、第１の実施形態のバッチノルム演算子７１０$_1$は、

【数６４】

$$\frac{W*X+b'-\mu'}{\sqrt{\sigma'^2+\epsilon}}\cdot\gamma'+\beta'$$

として定義することができ、σ’^{2}：＝θ、μ’：＝０、及びｂ’：＝０であり、ここで、θは所定のパラメーターである。第２の実施形態のバッチノルム演算子７１０$_2$は、

【数６５】

$$\frac{W*X+b'-\mu'}{\sqrt{\sigma'^2}}\cdot\gamma'+\beta'$$

として定義することができ、σ’^{2}：＝１、μ’＝０、及びｂ’＝０である。第３の実施形態のバッチノルム演算子７１０$_3$は、

【数６６】

$$\frac{W*X-\mu'}{\sqrt{\sigma'^2+\epsilon}}\cdot\gamma'+\beta'$$

として定義することができ、σ’^{2}：＝θ及びμ’：＝０であり、ここで、θは所定のパラメーターである。第４の実施形態のバッチノルム演算子７１０$_4$は、

【数６７】

$$\frac{W*X-\mu'}{\sqrt{\sigma'^2}}\cdot\gamma'+\beta'$$

として定義することができ、σ’^{2}：＝１及びμ’：＝０である。

【０１５４】

所定のパラメーターは１又は１－εであり、例えば、ここでもμ’、σ^2’、γ’、及びβ’は各出力ノードに対して１つの成分を含む転置ベクトルであり、Ｗは、重み行列であり、ＸはＮＮ層の活性化から導出される入力行列であり、ｂ’はバイアスを形成する転置ベクトル、例えば、各出力ノードに対して１つの成分を含む転置ベクトルである。

【０１５５】

装置７００は、例えば、装置７００に含まれ得るγ及びβ導出ユニット７２０を使用することによって、ＮＮ表現からγ及びβ、すなわちγ’及びβ’を導出するように構成される。

【０１５６】

第１の実施形態によれば、装置７００は、その全ての成分に適用される１つのシグナリング７３４を介して、σ’^{2}：＝θ、μ’：＝０、及びｂ’：＝０を推論又は導出するように構成され、ここで、θは所定のパラメーターである。

【０１５７】

第２の実施形態によれば、装置７００は、その全ての成分に適用される１つのシグナリ

ング７３４を介して、σ’^{2}：＝１、μ’：＝０、及びｂ’：＝０であると推論又は導出するように構成される。

【０１５８】

第３の実施形態によれば、装置７００は、その全ての成分に適用される１つのシグナリング７３４によって、σ’^{2}：＝θ及びμ’：＝０を推論又は導出するように構成され、ここで、θは所定のパラメーターである。

【０１５９】

第４の実施形態によれば、装置７００は、その全ての成分に適用される１つのシグナリング７３４によって、σ’^{2}：＝１及びμ’：＝０であると推論又は導出するように構成される。

【０１６０】

パラメーターσ’^{2}、μ’、及び任意選択でｂ’のこの導出又は推論は、パラメーター推論／導出ユニット７３０を使用して実行され得る。

【０１６１】

所定のパラメーターは１又は１－εであり、例えば、ここでもμ’、σ^{2}’、γ’、及びβ’は各出力ノードに対して１つの成分を含む転置ベクトルであり、Ｗは、重み行列であり、ＸはＮＮ層の活性化から導出される入力行列であり、ｂ’はバイアスを形成する転置ベクトル、例えば、各出力ノードに対して１つの成分を含む転置ベクトルである。

【０１６２】

図６では、装置７００によって導出又は推論されたパラメーターはアポストロフィによって示されているが、装置７００が元のパラメーター６１０を見ることがないという事実により、装置７００によって導出又は推論されたパラメーターはアポストロフィを使用せずに示されてもよい。装置７００を考慮すると、導出又は推論されたパラメーターは、唯一の既存のパラメーターである。

【０１６３】

任意選択で、装置７００は、例えば推論のために、導出又は推論されたパラメーター７２２、７２４、及び７３２とともにバッチノルム演算子を使用するように構成され得る。バッチノルム演算子計算ユニットは、バッチノルム演算子を使用するように構成され得る。代替的に、推論のデバイス、例えばデバイス４００又はデバイス５００は、バッチノルム演算子７１０のパラメーターを取得する装置７００を備えてもよい。

【０１６４】

定数スカラー値θ、すなわち、例えば、１又は１－εに等しくなり得る所定のパラメーターを導入すると、パラメーターｂ、μ、σ^2、γ、及びβは、ＢＮ（Ｘ）、すなわち、バッチノルム演算子７１０の結果を変更することなく、以下の順序付けられたステップによって修正することができる。

【数６８】

1) $\beta := \beta + \frac{(b-\mu)*\gamma}{\sqrt{\sigma^2+\epsilon}}$

2) $\gamma := \gamma \cdot \frac{\sqrt{\theta+\epsilon}}{\sqrt{\sigma^2+\epsilon}}$

３）σ^2：＝θ

４）μ：＝０

５）ｂ：＝０

【０１６５】

各演算は、転置されたベクトルの要素に対する要素ごとの演算として解釈される。実施形態２～３に例示されるように、ＢＮ（Ｘ）を変更しない更なる修正も可能である。例えば、バイアスｂ及び平均μはβに「積分」され、それにより、ｂ及びμはその後０に設定

される（第３の実施形態を参照）。又は、σ^2は、他のパラメーターがそれに応じて調整されるとき、ＢＮ（Ｘ）における分数の分母を１に等しく設定するために、１－ε（すなわち、θ＝１－ε）に設定され得る。

【０１６６】

それにより、全てのベクトル要素が同じ値を有するので、ｂ、σ^2、μ、及びｂを極めてより効率的に圧縮することができる。

【０１６７】

好ましい実施形態において、パラメーターの全ての要素が所定の定数値を有するかどうかを示すフラグ７３４が符号化される。パラメーターは、例えば、ｂ、μ、σ^2、γ、又はβであってもよい。所定の値は、例えば、０、１、又は１－εであってもよい。フラグが１に等しい場合、パラメーターの全てのベクトル要素は、所定の値に設定される。そうでなければ、パラメーターは、例えばＤｅｅｐＣＡＢＡＣのような現行技術水準のパラメーター符号化方法の１つを使用して符号化される。

【０１６８】

別の好ましい実施形態において、全てのベクトル要素が同じ値を有するかどうかを示すフラグがパラメーターごとに符号化される。全てのベクトル要素が同じ値を有するとき、フラグは１に等しく、その値は、例えばＤｅｅｐＣＡＢＡＣ、又は及び指数ゴロム符号、又は固定長符号のような現行技術水準のパラメーター符号化方法を使用して符号化される。フラグが０である場合、パラメーターのベクトル要素は、例えばＤｅｅｐＣＡＢＡＣのような現行技術水準のパラメーター符号化方法の１つを使用して符号化される。

【０１６９】

一実施形態によれば、装置６００／装置７００は、表現１１０において／表現１１０から、σ'^2の全ての成分（例えば、各成分は、対応する出力ノードを意味するＷの対応する行に対するものである）が互いに等しいこと、及びその値を示す／導出するように構成される。加えて、又は代替として、装置６００／装置７００は、表現１１０において／表現１１０から、μ’の全ての成分（例えば、各成分は、対応する出力ノードを意味するＷの対応する行に対するものである）が互いに等しいこと、及びその値を示す／導出するように構成される。加えて、又は代替として、装置６００／装置７００は、存在する場合、例えば第１の実施形態及び第２の実施形態の場合であるが第３の実施形態及び第４の実施形態の場合ではない場合、表現１１９において／表現１１９から、ｂ’の全ての成分（例えば、各成分は、対応する出力ノードを意味するＷの対応する行に対するものである）が互いに等しいこと、及びその値を示す／導出するように構成される。

【０１７０】

一実施形態によれば、装置６００は、２つのバッチノルム符号化モードの間で切り替え可能であるように更に構成され、第１のバッチノルム符号化モードでは、装置６００は、β’及びγ’の計算及び符号化を実行するように構成され、第２のバッチノルム符号化モードでは、装置は、受信されたμ、σ^2又はσ、γ、及びβ、並びに存在する場合、ｂを符号化するように構成される。換言すれば、受信されたパラメーター６１０は、第２のバッチノルムモードで表現１１０に直接符号化される。並行して、装置７００も、２つのバッチノルム符号化モード間で切り替え可能であるように構成してもよく、第１のバッチノルム符号化モードでは、装置７００は、導出することと、推論又は導出することとを実行するように構成され、第２の第１のバッチノルム符号化モードでは、装置７００は、μ、σ^2又はσ、γ、及びβ、並びに存在する場合、ｂを表現１１０から復号化するように構成される。換言すれば、パラメーター６１０は、第２のバッチノルムモードで表現１１０から直接復号化される。

【０１７１】

一実施形態によれば、装置６００は、β’及びγ’をＮＮ表現１１０に量子化及び符号化する装置１００（図２参照）を備える。例えば、装置６００は、最初に計算６２０を実行し、得られたパラメーターβ’及びγ’を、パラメーターの量子化のために装置１００に渡す。一実施形態によれば、装置７００は、ＮＮ表現１１０からβ及びγを導出する装置３

００（図２参照）を備える。
【０１７２】
理解を容易にするために、Ｘ及びＷと層の対との間の可能な関係が図７に示されており、左は全結合層ｉ＋１であり、右は畳み込み層ｉ＋１である。層のニューロンは円１０で示されている。各層のニューロンは、アレイ位置（ｘ，ｙ）に配置される。各層ｉは、q_i列のニューロン１０とp_i行のニューロン１０とを有する。全結合の場合、X_iは、成分
【数６９】

$$X_{1\ldots p_i \cdot q_i}$$

のベクトルであり、各X_gは、位置
【数７０】

$$\{[g/q_i];g\%q_i+1\}$$

におけるニューロンの活性化で占められ、W_iは、成分
【数７１】

$$W_{1\ldots p_{i+1}\cdot q_{i+1},1\ldots p_i \cdot q_i}$$

の行列であり、各$W_{g,h}$は、位置
【数７２】

$$\{[g/q_{i+1}];g\%q_{i+1}+1\}$$

における層ｉ＋１のニューロン１０と位置
【数７３】

$$\{[h/q_i];h\%q_i+1\}$$

における層ｉのニューロン１０との間のエッジ１２の重みで占められる。畳み込みの場合、X_iは、成分
【数７４】

$$X_{1\ldots r\cdot s,1\ldots p_{i+1}\cdot q_{i+1}}$$

の行列であり、ここで、各$X_{g,h}$は、位置
【数７５】

$$\{\lceil (g+(h-1)*q_i/(q_{i+1}+s-1))/s\rceil ; (g+(h-1)*q_i/(q_{i+1}+s-1))\%s+1\}$$

におけるニューロンの活性化で占められ、Ｗｉは、成分Ｗ１．．．ｒ・ｓのベクトルであり、ここで、各Ｗｇ，ｈは、層ｉにわたって分散されたｐｉ＋１・ｑｉ＋１位置のうちの１つに配置された層ｉ中のサイズｒ×ｓの矩形フィルターカーネル中のニューロンから、カーネル位置に対応する層ｉ＋１中のニューロン位置につながるエッジの重みで占められる。

【０１７３】

一部の態様を装置の文脈で説明してきたが、これらの態様が対応する方法の説明も表すことは明らかであり、ブロック又はデバイスは方法ステップ又は方法ステップの特徴に対応する。同様に、方法ステップの文脈で説明される態様はまた、対応する装置の対応するブロック又は項目又は特徴の説明を表す。方法ステップの一部又は全部は、例えばマイクロプロセッサ、プログラマブルコンピューター又は電子回路のようなハードウェア装置によって（又はそれを使用して）実行されてもよい。一部の実施形態において、最も重要な方法ステップのうちの１つ以上は、かかる装置によって実行されてもよい。

【０１７４】

本発明のＮＮ表現を含む本発明のデジタルデータ、データストリーム、又はファイルは、デジタル記憶媒体上に記憶することができ、又は無線伝送媒体若しくはインターネット等の有線伝送媒体等の伝送媒体上で伝送することができる。

【０１７５】

特定の実装要件に応じて、本発明の実施形態は、ハードウェア又はソフトウェアで実行することができる。実施態様は、それぞれの方法が実行されるようにプログラム可能なコンピューターシステムと協働する（又は協働することができる）電子的可読制御信号が記憶されたデジタル記憶媒体、例えば、フロッピーディスク、ＤＶＤ、Ｂｌｕ－ｒａｙ（登録商標）、ＣＤ、ＲＯＭ、ＰＲＯＭ、ＥＰＲＯＭ、ＥＥＰＲＯＭ又はフラッシュメモリを使用して実行することができる。したがって、デジタル記憶媒体はコンピューター可読であってもよい。

【０１７６】

本発明による一部の実施形態は、電子的可読制御信号を有するデータキャリアを備え、該電子的可読制御信号は、本明細書で説明される方法のうちの１つが実行されるように、プログラム可能なコンピューターシステムと協働することが可能である。

【０１７７】

概して、本発明の実施形態は、プログラムコードを有するコンピュータープログラム製品として実施することができ、プログラムコードは、コンピュータープログラム製品がコンピューター上で実行されるときに方法のうちの１つを実行するように動作可能である。プログラムコードは、例えば、機械可読キャリアに記憶されてもよい。

【０１７８】

他の実施形態は、機械可読キャリア上に記憶された、本明細書で説明される方法のうちの１つを実行するコンピュータープログラムを含む。

【０１７９】

したがって、換言すれば、本発明の方法の一実施形態は、コンピュータープログラムがコンピューター上で実行されるときに、本明細書で説明される方法のうちの１つを実行するプログラムコードを有するコンピュータープログラムである。

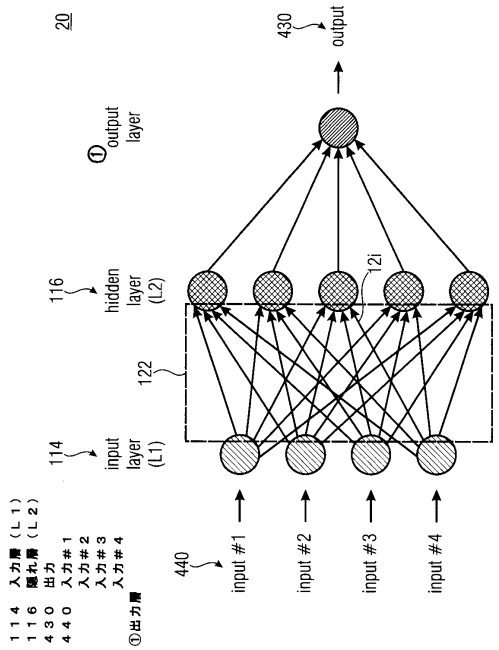
【０１８０】

したがって、本発明の方法の更なる実施形態は、本明細書に記載の方法のうちの１つを実行するコンピュータープログラムを記録したデータキャリア（又はデジタル記憶媒体、又はコンピューター可読媒体）である。データキャリア、デジタル記憶媒体又は記録された媒体は、典型的には有形及び／又は非一時的である。

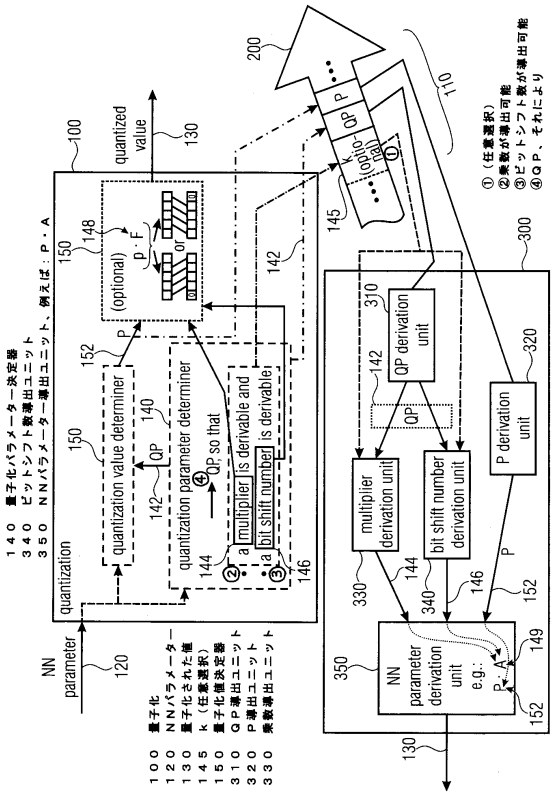
【０１８１】
したがって、本発明の方法の更なる実施形態は、本明細書に記載された方法の１つを実行するコンピュータープログラムを表すデータストリーム又はシグナルのシーケンスである。データストリーム又はシグナルのシーケンスは、例えば、データ通信接続を介して、例えばインターネットを介して転送されるように構成されてもよい。
【０１８２】
更なる実施形態は、本明細書に記載の方法の１つを実行するように構成又は適合された処理手段、例えばコンピューター又はプログラム可能な論理デバイスを含む。
【０１８３】
更なる実施形態は、本明細書に記載の方法のうちの１つを実行するコンピュータープログラムがインストールされたコンピューターを含む。
【０１８４】
本発明による更なる実施形態は、本明細書で説明される方法のうちの１つを実行するコンピュータープログラムを受信機に（例えば、電子的に又は光学的に）転送するように構成された装置又はシステムを含む。受信機は、例えば、コンピューター、モバイルデバイス、メモリデバイス等であってもよい。装置又はシステムは、例えば、コンピュータープログラムを受信機に転送するためのファイルサーバを備えることができる。
【０１８５】
一部の実施形態において、プログラム可能論理デバイス（例えば、フィールドプログラマブルゲートアレイ）が、本明細書に説明される方法の機能性の一部又は全部を行うために使用されてもよい。一部の実施形態において、フィールドプログラマブルゲートアレイは、本明細書に説明される方法のうちの１つを行うために、マイクロプロセッサと協働してもよい。概して、方法は、任意のハードウェア装置によって実行されることが好ましい。
【０１８６】
本明細書で説明される装置は、ハードウェア装置を使用して、又はコンピューターを使用して、又はハードウェア装置とコンピューターとの組み合わせを使用して実施され得る。
【０１８７】
本明細書で説明される装置、又は本明細書で説明される装置の任意の構成要素は、少なくとも部分的にハードウェア及び／又はソフトウェアで実施され得る。
【０１８８】
本明細書で説明される方法は、ハードウェア装置を使用して、又はコンピューターを使用して、又はハードウェア装置とコンピューターとの組み合わせを使用して実行され得る。
【０１８９】
本明細書で説明される方法、又は本明細書で説明される装置の任意の構成要素は、少なくとも部分的にハードウェア及び／又はソフトウェアによって実行され得る。
【０１９０】
上述の実施形態は、本発明の原理を単に例示するものである。本明細書に記載された構成及び詳細の変更及び変形が他の当業者に明らかであることが理解される。したがって、添付の特許請求の範囲によってのみ限定され、本明細書の実施形態の記述及び説明によって提示される特定の詳細によって限定されないことが意図される。

【図面】

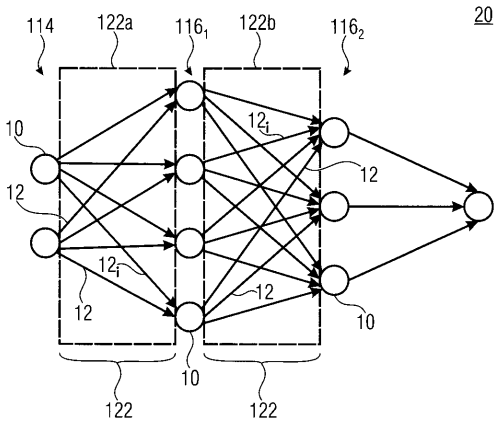
【図１】


20
430 output
① output layer
116 hidden layer (L2)
122
114 input layer (L1)
440 input #1
input #2
input #3
input #4
12i
114 入力層（Ｌ１）
116 隠れ層（Ｌ２）
430 出力
440 入力＃１
入力＃２
入力＃３
入力＃４
①出力層

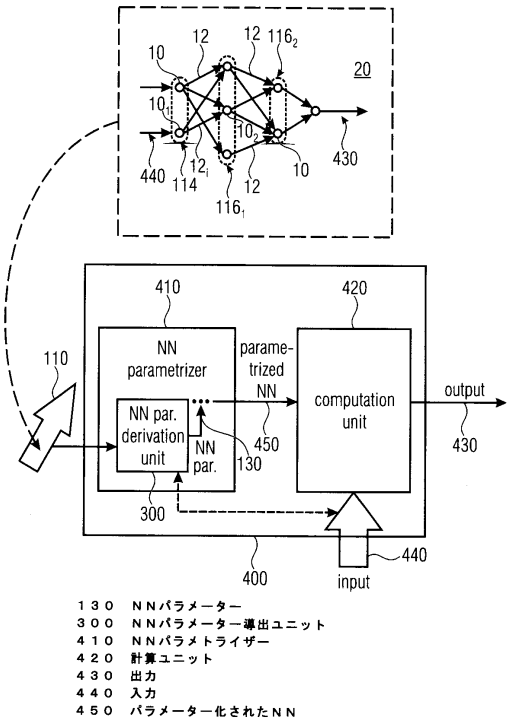
【図２】


200
110
quantized value 130
100
148
(optional)
p · F
or
150
152
quantization value determiner
140
QP
142
quantization parameter determiner
④ QP so that
a multiplier is derivable and
a bit shift number is derivable
144
146
quantization
NN parameter 120
145
QP
k
P
①（任意選択）
②乗数が導出可能
③ビットシフト数が導出可能
④ＱＰ、それにより
300
310 QP derivation unit
320 P derivation unit
multiplier derivation unit
bit shift number derivation unit
330
340
350 NN parameter derivation unit
e.g.: P · A
149
130
140 量子化パラメーター決定器
340 ビットシフト数導出ユニット
350 ＮＮパラメーター導出ユニット
100 量子化
120 ＮＮパラメーター
130 量子化された値
145 ｋ（任意選択）
150 量子化値決定器
310 ＱＰ導出ユニット
320 Ｐ導出ユニット
330 乗数導出ユニット

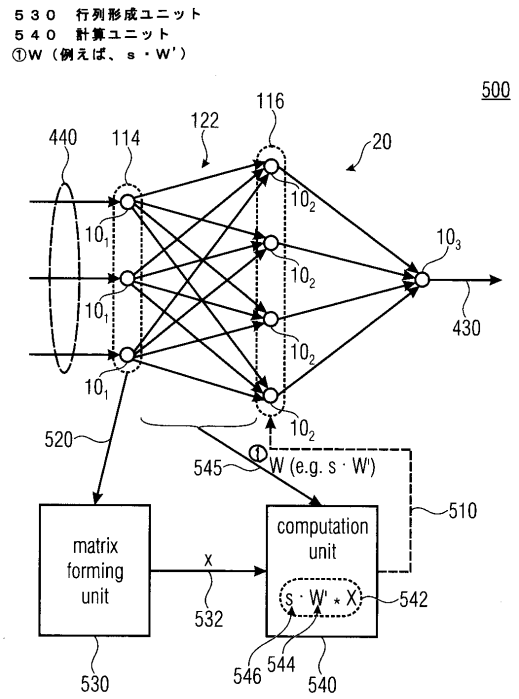
【図３】


20
114
122a
116₁
122b
116₂
10
12
12₁
122
122

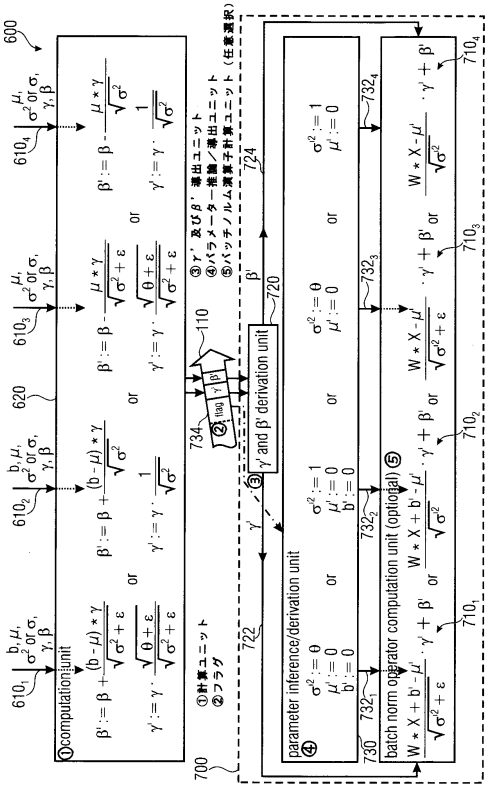
【図４】


20
10
12
116₂
440
114
116₁
430
410
420
110
NN parametrizer
NN par. derivation unit
NN par. 130
parame-trized NN 450
computation unit
output 430
300
400
440
input
１３０　ＮＮパラメーター
３００　ＮＮパラメーター導出ユニット
４１０　ＮＮパラメトライザー
４２０　計算ユニット
４３０　出力
４４０　入力
４５０　パラメーター化されたＮＮ

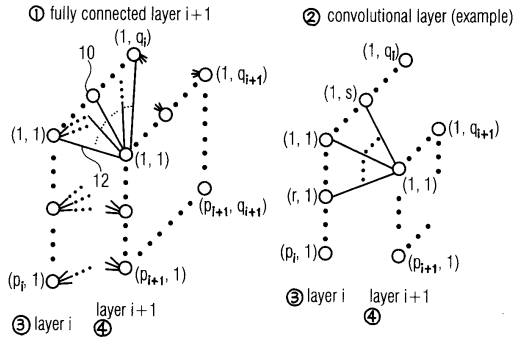
【図５】



【図６】



【図７】



①全結合層ｉ＋１
②畳み込み層（例）
③層ｉ
④層ｉ＋１
⑤B_iは１×n_i行列である
⑥（ｒ・ｓ）×（p_{i+1}・q_{i+1}）行列である
⑦B_iは１×（p_{i+1}・q_{i+1}）行列である

フロントページの続き

スティテュート， ツェー／オー フラウンホーファー－インスティテュート フュア ナッハリヒテンテヒニーク
(72)発明者 マーリンチ，タルマイ
ドイツ連邦共和国 ベルリン １０５８７ アインシュタインウーファー３７ ハーハーイー，ハインリッヒ－ヘルツ－インスティテュート， ツェー／オー フラウンホーファー－インスティテュート フュア ナッハリヒテンテヒニーク
(72)発明者 サメック，ヴォイチェフ
ドイツ連邦共和国 ベルリン １０５８７ アインシュタインウーファー３７ ハーハーイー，ハインリッヒ－ヘルツ－インスティテュート， ツェー／オー フラウンホーファー－インスティテュート フュア ナッハリヒテンテヒニーク
(72)発明者 ハーゼ，パウル
ドイツ連邦共和国 ベルリン １０５８７ アインシュタインウーファー３７ ハーハーイー，ハインリッヒ－ヘルツ－インスティテュート， ツェー／オー フラウンホーファー－インスティテュート フュア ナッハリヒテンテヒニーク
(72)発明者 ミュラー，カーステン
ドイツ連邦共和国 ベルリン １０５８７ アインシュタインウーファー３７ ハーハーイー，ハインリッヒ－ヘルツ－インスティテュート， ツェー／オー フラウンホーファー－インスティテュート フュア ナッハリヒテンテヒニーク
(72)発明者 キルヒホッファー，ハイナー
ドイツ連邦共和国 ベルリン １０５８７ アインシュタインウーファー３７ ハーハーイー，ハインリッヒ－ヘルツ－インスティテュート， ツェー／オー フラウンホーファー－インスティテュート フュア ナッハリヒテンテヒニーク
(72)発明者 マープ，デトレフ
ドイツ連邦共和国 ベルリン １０５８７ アインシュタインウーファー３７ ハーハーイー，ハインリッヒ－ヘルツ－インスティテュート， ツェー／オー フラウンホーファー－インスティテュート フュア ナッハリヒテンテヒニーク
(72)発明者 シュヴァルツ，ハイコ
ドイツ連邦共和国 ベルリン １０５８７ アインシュタインウーファー３７ ハーハーイー，ハインリッヒ－ヘルツ－インスティテュート， ツェー／オー フラウンホーファー－インスティテュート フュア ナッハリヒテンテヒニーク
(72)発明者 ヴィーガント，トーマス
ドイツ連邦共和国 ベルリン １０５８７ アインシュタインウーファー３７ ハーハーイー，ハインリッヒ－ヘルツ－インスティテュート， ツェー／オー フラウンホーファー－インスティテュート フュア ナッハリヒテンテヒニーク

審査官 大倉 崚吾

(56)参考文献
国際公開第２０１９／００８７５２（WO，Ａ１）
JACOB, Benoit et al.，"Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference"，arXiv [online]，2017年12月，［2023年10月11日検索］，インターネット＜ＵＲＬ：https://arxiv.org/abs/1712.05877v1＞，1712.05877v1
FLYNN, David et al.，"G-PCC: Integer step sizes for in-tree geometry quantisation"，MPEG Point Cloud Compression [online]，2020年01月，［2023年10月11日検索］，インターネット＜ＵＲＬ：https://mpeg-pcc.org/index.php/public-contributions/m52522-g-pcc-integer-step-sizes-for-in-tree-geometry-quantisation/＞，m52522
CAI, Yaohui et al.，"ZeroQ: A Novel Zero Shot Quantization Framework"，arXiv [online]，2020年01月，［2023年10月11日検索］，インターネット＜ＵＲＬ：https://arxiv.org/abs/2001.00281v1＞，2001.00281v1

(58)調査した分野 (Int.Cl.，ＤＢ名)
Ｇ０６Ｎ ３／０２－ ３／１０
Ｇ０６Ｎ ２０／００－９９／００