(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7640552号
(P7640552)

(45)発行日　令和7年3月5日(2025.3.5)

(24)登録日　令和7年2月25日(2025.2.25)

(51)国際特許分類

		FI	
H04N	*19/126 (2014.01)*	H04N	19/126
H04N	*19/13 (2014.01)*	H04N	19/13
H04N	*19/463 (2014.01)*	H04N	19/463

請求項の数　46　（全77頁）

(21)出願番号　特願2022-538077(P2022-538077)
(86)(22)出願日　令和2年12月21日(2020.12.21)
(65)公表番号　特表2023-507502(P2023-507502A)
(43)公表日　令和5年2月22日(2023.2.22)
(86)国際出願番号　PCT/EP2020/087489
(87)国際公開番号　WO2021/123438
(87)国際公開日　令和3年6月24日(2021.6.24)
審査請求日　令和4年10月19日(2022.10.19)
(31)優先権主張番号　19218862.1
(32)優先日　令和1年12月20日(2019.12.20)
(33)優先権主張国・地域又は機関
欧州特許庁(EP)
前置審査

(73)特許権者　591037214
フラウンホッファー－ゲゼルシャフト ツァ フェルダールング デァ アンゲヴァンテン フォアシュンク エー．ファオ
ドイツ連邦共和国 ８０６８６ ミュンヘン ハンザシュトラッセ ２７ツェー
(74)代理人　100079577
弁理士　岡田 全啓
(72)発明者　ハーセ ポール
ドイツ連邦共和国 １０５８７ ベルリン アインシュタインウーファー ３７ フラウンホッファー－インスティチュート フュア ナーハリヒテンテヒニーク ハインリッヒ－ヘルツ－インスティチュート ＨＨＩ 内

最終頁に続く

(54)【発明の名称】　ニューラルネットワークのパラメータを符号化するための概念

(57)【特許請求の範囲】
【請求項１】
ニューラルネットワーク（１０）を定義するニューラルネットワークパラメータ（１３）をデータストリーム（１４）から復号化するための装置であって、前記装置は、
前記データストリーム（１４）から復号化された、以前のニューラルネットワークパラメータのための量子化インデックス（５８）に依存して、現在のニューラルネットワークパラメータ（１３’）について、複数（５０）の再構成レベルセット（５２）の中から再構成レベルセット（４８）を選択（５４）することと、
前記現在のニューラルネットワークパラメータについての選択された前記再構成レベルのセット（４８）のうちの１つの再構成レベルを示す、前記現在のニューラルネットワークパラメータ（１３’）のための量子化インデックス（５６）を前記データストリーム（１４）から復号化することと、
前記現在のニューラルネットワークパラメータのための前記量子化インデックス（５６）によって示される前記選択された再構成レベルセット（４８）のうちの前記１つの再構成レベルに前記現在のニューラルネットワークパラメータ（１３’）を逆量子化（６２）すること
により、前記ニューラルネットワークパラメータ（１３’）を順次復号化するように構成され、
前記装置は、
前記現在のニューラルネットワークパラメータ（１３’）に関連づけられた状態に応じて、

前記現在のニューラルネットワークパラメータ（１３’）について、前記複数（５０）の再構成レベルセット（５２）の中から前記再構成レベルセット（４８）を決定すること、及び、
前記データストリームから復号化された、直前のニューラルネットワークパラメータのための前記量子化インデックス（５８）に応じて、後続のニューラルネットワークパラメータの状態を更新すること
による状態遷移プロセスにより、前記現在のニューラルネットワークパラメータ（１３’）について、前記複数（５０）の再構成レベルセット（５２）の中から前記再構成レベルセット（４８）を選択（５４）し、
前記現在のニューラルネットワークパラメータ（１３’）の状態および以前に復号化されたニューラルネットワークパラメータの前記量子化インデックスに依存する（１２２）確率モデルを用いる算術符号化を用いて、前記データストリーム（１４）から前記現在のニューラルネットワークパラメータ（１３’）のための前記量子化インデックス（５６）を復号化し、
前記現在のニューラルネットワークパラメータ（１３’）に関連付けられた前記状態に依存して、複数の確率モデルの中から確率モデルのサブセットを事前選択し、以前に復号化されたニューラルネットワークパラメータのための前記量子化インデックスに依存して（１２１）、前記確率モデルのサブセットの中から前記現在のニューラルネットワークパラメータのための前記確率モデルを選択し、
前記現在のニューラルネットワークパラメータが関連する部分に隣接する前記ニューラルネットワークの部分に関連する、以前に復号化されたニューラルネットワークパラメータの前記量子化インデックスの値に依存して、前記現在のニューラルネットワークパラメータのための確率モデルを第１、第２および第３の確率モデルを含むサブセットの中から選択する
ように構成され、
ここで、前記以前に復号化されたニューラルネットワークパラメータがゼロよりも小さければ、第１の確率モデルが選択され、
前記以前に復号化されたニューラルネットワークパラメータがゼロよりも大きければ、第２の確率モデルが選択され、
前記以前に復号化されたニューラルネットワークパラメータがゼロに等しければ、前記第３の確率モデルが選択される、
装置。

【請求項２】
前記ニューラルネットワークパラメータ（１３）が、前記ニューラルネットワーク（１０）のニューロン相互接続（１１）の重みに関連する、請求項１に記載の装置。

【請求項３】
前記複数（５０）の再構成レベルセット（５２）のうちの再構成レベルセット（５２）の数は２である、請求項１又は２に記載の装置。

【請求項４】
前記複数（５０）の再構成レベルセット（５２）を既定の量子化ステップサイズ（ＱＰ）によりパラメータ化（６０）し、前記データストリーム（１４）から前記既定の量子化ステップサイズに関する情報を導出するように構成される、請求項１乃至３のいずれかに記載の装置。

【請求項５】
前記ニューラルネットワークは１つ以上のＮＮ層を含み、前記装置は、
それぞれのＮＮ層（ｐ；ｐ－１）について、前記それぞれのＮＮ層のための既定の量子化ステップサイズに関する情報を前記データストリーム（１４）から導出し、
それぞれのＮＮ層について、前記それぞれのＮＮ層について導出された前記既定の量子化ステップサイズを使用して前記複数（５０）の再構成レベルセット（５２）をパラメータ化して、前記それぞれのＮＮ層に属する前記ニューラルネットワークパラメータの逆量

子化のために使用されるようにする
ように構成される、請求項１乃至４のいずれかに記載の装置。
【請求項６】
前記複数（５０）の再構成レベルセット（５２）のうちの再構成レベルセット（５２）の数は２であり、前記複数の再構成レベルセットは、
ゼロと既定の量子化ステップサイズの偶数倍とを含む第１の再構成レベルセット（セット０）と、
ゼロと前記既定の量子化ステップサイズの奇数倍とを含む第２の再構成レベルセット（セット１）と
を含む、請求項１乃至５のいずれかに記載の装置。
【請求項７】
すべての再構成レベルセットのすべての再構成レベルは、既定の量子化ステップサイズの整数倍を表し、前記装置は、
それぞれのニューラルネットワークパラメータについて、前記それぞれのニューラルネットワークパラメータについての前記選択された再構成レベルセットと、前記それぞれのニューラルネットワークパラメータのための前記量子化インデックスをエントロピー復号化したものとに応じて中間整数値を導出すること、及び、
それぞれのニューラルネットワークパラメータについて、前記それぞれのニューラルネットワークパラメータについての前記中間整数値を、前記それぞれのニューラルネットワークパラメータのための前記既定の量子化ステップサイズで乗算すること
により、前記ニューラルネットワークパラメータを逆量子化するように構成される、請求項１乃至６のいずれかに記載の装置。
【請求項８】
前記複数（５０）の再構成レベルセット（５２）のうちの再構成レベルセット（５２）の数は２であり、前記装置は、
前記それぞれのニューラルネットワークパラメータについての前記選択された再構成レベルセットが第１のセットである場合、前記それぞれのニューラルネットワークパラメータのための前記量子化インデックスを２倍して、前記それぞれのニューラルネットワークパラメータについての前記中間整数値を得ること、及び、
それぞれのニューラルネットワークパラメータについての前記選択された再構成レベルセットが第２のセットであり、かつ前記それぞれのニューラルネットワークパラメータのための前記量子化インデックスがゼロに等しい場合、前記それぞれのニューラルネットワークパラメータについての前記中間整数値をゼロに等しく設定すること、及び、
それぞれのニューラルネットワークパラメータについての前記選択された再構成レベルセットが第２のセットであり、かつ前記それぞれのニューラルネットワークパラメータのための前記量子化インデックスがゼロより大きい場合、前記それぞれのニューラルネットワークパラメータのための前記量子化インデックスを２倍し、該乗算の結果から１を引いて前記それぞれのニューラルネットワークパラメータについての前記中間整数値を得ること、及び、
現在のニューラルネットワークパラメータについての前記選択された再構成レベルセットが第２のセットであり、前記それぞれのニューラルネットワークパラメータのための前記量子化インデックスがゼロより小さい場合、前記それぞれのニューラルネットワークパラメータのための前記量子化インデックスを２倍し、該乗算の結果に１を加えて、前記それぞれのニューラルネットワークパラメータについての前記中間整数値を得る
ことにより、それぞれのニューラルネットワークパラメータについての前記中間整数値を導出するように構成される、請求項７に記載の装置。
【請求項９】
前記現在のニューラルネットワークパラメータ（１３’）について、前記データストリーム（１４）から復号化された、以前に復号化されたニューラルネットワークパラメータのための前記量子化インデックス（５８）を２値化したもののＬＳＢ部分又は以前に復号化

されたビンに応じて、前記複数（５０）の再構成レベルセット（５２）の中から前記再構成レベルセット（４８）を選択する（５４）ように構成される、請求項１乃至８のいずれかに記載の装置。

【請求項１０】

前記現在のニューラルネットワークパラメータ（１３’）について、前記データストリーム（１４）から復号化された、以前に復号化されたニューラルネットワークパラメータのための前記量子化インデックス（５８）の２値関数の結果に応じて、前記複数（５０）の再構成レベルセット（５２）の中から前記再構成レベルセット（４８）を選択する（５４）ように構成される、請求項１乃至８のいずれかに記載の装置。

【請求項１１】

前記装置は、前記現在のニューラルネットワークパラメータ（１３’）について、前記データストリーム（１４）から復号化された、以前に復号化されたニューラルネットワークパラメータのための前記量子化インデックス（５８）のパリティに応じて、前記複数（５０）の再構成レベルセット（５２）の中から前記再構成レベルセット（４８）を選択する（５４）ように構成される、請求項１乃至１０のいずれかに記載の装置。

【請求項１２】

前記複数（５０）の再構成レベルセット（５２）のうちの前記再構成レベルセット（５２）の数は２であり、前記装置は、

それぞれのニューラルネットワークパラメータのためのサブセットインデックスを、前記それぞれのニューラルネットワークパラメータについての前記選択された再構成レベルセットと、前記それぞれのニューラルネットワークパラメータのための前記量子化インデックスの２値関数とに基づいて導き出し、前記サブセットインデックスについての４つの可能な値を生じさせ、

前記現在のニューラルネットワークパラメータ（１３’）について、以前に復号化されたニューラルネットワークパラメータのための前記サブセットインデックスに応じて、前記複数（５０）の再構成レベルセット（５２）の中から前記再構成レベルセット（４８）を選択する（５４）

ように構成される、請求項１乃至１１のいずれかに記載の装置。

【請求項１３】

前記装置は、前記現在のニューラルネットワークパラメータ（１３’）について、複数の直前に復号化されたニューラルネットワークパラメータのための前記サブセットインデックスに依存する選択ルールを用いて、前記複数（５０）の再構成レベルセット（５２）の中から前記再構成レベルセット（４８）を選択し（５４）、前記選択ルールを前記ニューラルネットワークパラメータの一部または全部に対して使用するように構成される、請求項１２に記載の装置。

【請求項１４】

前記選択ルールが依存する、前記直前に復号化されたニューラルネットワークパラメータの数は２である、請求項１３に記載の装置。

【請求項１５】

それぞれのニューラルネットワークパラメータのための前記サブセットインデックスは、前記それぞれのニューラルネットワークパラメータについての前記選択された再構成レベルセットと、前記それぞれのニューラルネットワークパラメータのための前記量子化インデックスのパリティとに基づいて導出される、請求項１２乃至１４のいずれかに記載の装置。

【請求項１６】

前記装置は、

前記現在のニューラルネットワークパラメータ（１３’）について、前記現在のニューラルネットワークパラメータ（１３’）に関連づけられた状態に応じて、前記複数（５０）の再構成レベルセット（５２）の中から前記再構成レベルセット（４８）を決定すること、及び、

前記データストリームから復号化された、前記直前のニューラルネットワークパラメータのための前記量子化インデックス（５８）に応じて、後続のニューラルネットワークパラメータの状態を更新すること
による状態遷移プロセスにより、前記現在のニューラルネットワークパラメータ（１３’）について、前記複数（５０）の再構成レベルセット（５２）の中から前記再構成レベルセット（４８）を選択する（５４）ように構成される、請求項１乃至１５のいずれかに記載の装置。
【請求項１７】
前記データストリームから復号化された、前記直前のニューラルネットワークパラメータのための前記量子化インデックス（５８）の２値関数を用いて、前記後続の前記ニューラルネットワークパラメータの状態を更新するように構成される、請求項１６に記載の装置。
【請求項１８】
前記データストリームから復号化された、前記直前のニューラルネットワークパラメータのための前記量子化インデックス（５８）のパリティを使用して、前記後続のニューラルネットワークパラメータの状態を更新するように構成される、請求項１６に記載の装置。
【請求項１９】
前記状態遷移プロセスは、４つ又は８つの可能な状態の間で遷移するように構成される、請求項１６乃至１８のいずれかに記載の装置。
【請求項２０】
前記状態遷移プロセスにおいて偶数の可能な状態の間で遷移し、前記複数（５０）の再構成レベルセット（５２）のうちの再構成レベルセット（５２）の数が２であるように構成され、ここで、前記現在のニューラルネットワークパラメータ（１３’）について、前記現在のニューラルネットワークパラメータ（１３’）に関連づけられた前記状態に応じて、前記再構成レベルセット（５２）のうちの前記再構成レベルセット（４８）を決定することにより、前記状態が前記偶数の可能な状態のうちの前半に属する場合は前記複数（５０）の再構成レベルセット（５２）のうちの第１の再構成レベルセットが決定され、前記状態が前記偶数の可能な状態のうちの後半に属している場合は前記複数（５０）の再構成レベルセット（５２）のうちの第２の再構成レベルセットが決定される、請求項１６乃至１９のいずれかに記載の装置。
【請求項２１】
前記データストリームから復号化された、前記直前のニューラルネットワークパラメータの状態及び前記量子化インデックス（５８）のパリティの組み合わせを、前記後続のニューラルネットワークパラメータに関連づけられた別の状態にマッピングする遷移テーブルによって、前記状態の更新を実行するように構成される、請求項１６乃至２０のいずれかに記載の装置。
【請求項２２】
前記現在のニューラルネットワークパラメータ（１３’）について選択された前記再構成レベルセット（４８）に依存する確率モデル（１２３）を用いる算術符号化を用いて、前記現在のニューラルネットワークパラメータ（１３’）のための前記量子化インデックス（５６）を前記データストリーム（１４）から復号化するよう構成される、請求項１乃至２１のいずれかに記載の装置。
【請求項２３】
前記量子化インデックス（５６）を２値化したもの（８２）の少なくとも１つのビン（８４）についての前記現在のニューラルネットワークパラメータ（１３’）の状態に依存する（１２２）前記確率モデルを用いることにより、２値算術符号化を用いて、前記現在のニューラルネットワークパラメータのための前記量子化インデックス（５６）を前記データストリーム（１４）から復号化するように構成される、請求項１乃至２２のいずれかに記載の装置。
【請求項２４】

前記少なくとも１つのビンが、前記現在のニューラルネットワークパラメータのための前記量子化インデックス（５６）がゼロに等しいか否かを示す有意性ビンを含む、請求項２３に記載の装置。

【請求項２５】

前記少なくとも１つのビンが、前記現在のニューラルネットワークパラメータのための前記量子化インデックス（５６）がゼロより大きいか又はゼロより小さいかを示す符号ビン（８６）を含む、請求項２３または請求項２４に記載の装置。

【請求項２６】

前記少なくとも１つのビンが、前記現在のニューラルネットワークパラメータの前記量子化インデックス（５６）の絶対値がＸより大きいか否かを示す、ｇｒｅａｔｅｒ－ｔｈａｎ－Ｘビンを含み、ここでＸはゼロより大きい整数である、請求項２３乃至２５のいずれかに記載の装置。

【請求項２７】

前記確率モデルの依存性には、前記依存性を用いた、前記ニューラルネットワークパラメータについてのコンテキストセットのうちのコンテキスト（８７）の選択（１０３）を含み、それぞれのコンテキストには既定の確率モデルが関連づけられているように構成される、請求項１乃至２６のいずれかに記載の装置。

【請求項２８】

前記それぞれのコンテキストを用いて算術符号化された前記量子化インデックスに基づいて、前記コンテキストのそれぞれに関連付けられた前記既定の確率モデルを更新するように構成される、請求項２７に記載の装置。

【請求項２９】

前記量子化インデックスを２値化したものの少なくとも１つのビンについて、前記現在のニューラルネットワークパラメータ（１３’）について選択された前記再構成レベルセット（４８）に依存する確率モデルを用いることによって、前記現在のニューラルネットワークパラメータ（１３’）のための前記量子化インデックス（５６）を前記データストリーム（１４）から２値算術符号化を用いて復号化するように構成される、請求項１乃至２８のいずれかに記載の装置。

【請求項３０】

前記少なくとも１つのビンは、前記現在のニューラルネットワークパラメータのための前記量子化インデックス（５６）がゼロに等しいか否かを示す有意性ビンを含む、請求項２９に記載の装置。

【請求項３１】

前記少なくとも１つのビンは、前記現在のニューラルネットワークパラメータのための前記量子化インデックス（５６）がゼロより大きいか又はゼロより小さいかを示す符号ビンを含む、請求項２９又は３０に記載の装置。

【請求項３２】

前記少なくとも１つのビンが、前記現在のニューラルネットワークパラメータのための前記量子化インデックス（５６）の絶対値がＸより大きいか否かを示すｇｒｅａｔｅｒ－ｔｈａｎ－Ｘビンを含み、ここでＸはゼロより大きい整数である、請求項２９乃至３１のいずれかに記載の装置。

【請求項３３】

第１の状態又は再構成レベルセットについて事前選択されたサブセットが、任意の他の状態又は再構成レベルセットについて事前選択されたサブセットと互いに素であるように、前記複数の確率モデルの中から確率モデルの前記サブセットを、前記現在のニューラルネットワークパラメータ（１３’）について選択された再構成レベルの前記状態又は前記再構成レベルセット（４８）に依存して事前選択するように構成される、請求項１乃至３２のいずれかに記載の装置。

【請求項３４】

前記現在のニューラルネットワークパラメータが関連する部分に隣接する前記ニューラ

ルネットワークの部分に関連する、以前に復号化されたニューラルネットワークパラメータの前記量子化インデックスに応じて、前記確率モデルのサブセットの中から前記現在のニューラルネットワークパラメータについての前記確率モデルを選択するように構成される、請求項１乃至３３のいずれかに記載の装置。

【請求項３５】

前記現在のニューラルネットワークパラメータが関連する部分に隣接する前記ニューラルネットワークの部分に関連する、以前に復号化されたニューラルネットワークパラメータのための前記量子化インデックスの特性に依存して、前記確率モデルのサブセットの中から前記現在のニューラルネットワークパラメータについての前記確率モデルを選択するように構成され、前記特性は、

前記現在のニューラルネットワークパラメータが関連する部分に隣接する前記ニューラルネットワークの部分に関連する、以前に復号化されたニューラルネットワークパラメータのゼロでない量子化インデックスの符号と、

前記現在のニューラルネットワークパラメータが関連する部分に隣接する前記ニューラルネットワークの部分に関連する、以前に復号化されたニューラルネットワークパラメータの量子化インデックスの数であって、ゼロでない数と、

前記現在のニューラルネットワークパラメータが関連する部分に隣接する前記ニューラルネットワークの部分に関連する、以前に復号化されたニューラルネットワークパラメータの量子化インデックスの絶対値の合計値と、

前記現在のニューラルネットワークパラメータが関連する部分に隣接する前記ニューラルネットワークの部分に関連する、以前に復号化されたニューラルネットワークパラメータの量子化インデックスの前記絶対値の合計値と、前記現在のニューラルネットワークパラメータが関連する部分に隣接する前記ニューラルネットワークの部分に関連する、以前に復号化されたニューラルネットワークパラメータの量子化インデックスの数であって、ゼロでない数と、の差分と、

のうちの１つ以上を含む、請求項１乃至３３のいずれかに記載の装置。

【請求項３６】

前記以前に復号化されたニューラルネットワークパラメータが前記現在のニューラルネットワークパラメータと同じニューラルネットワーク層に関連するように、前記以前に復号化されたニューラルネットワークパラメータを位置づけるように構成される、請求項３４又は３５に記載の装置。

【請求項３７】

１つ以上の前記以前に復号化されたニューラルネットワークパラメータが、前記現在のニューラルネットワークパラメータが参照するニューロン相互接続が関連するニューロンまたは該ニューロンに隣接する別のニューロンから出現するニューロン相互接続又はこれらのニューロンに向かうニューロン相互接続に関連するように、前記以前に復号化されたニューラルネットワークパラメータのうちの１つ以上のパラメータを位置付けるように構成される、請求項３４又は３５に記載の装置。

【請求項３８】

前記ニューラルネットワークパラメータ（１３）のための前記量子化インデックス（５６）を復号化し、前記ニューラルネットワークパラメータ（１３）間の共通の連続的な順序（１４’）に沿って前記ニューラルネットワークパラメータ（１３）の前記逆量子化を実行するように構成される、請求項１乃至３７のいずれかに記載の装置。

【請求項３９】

前記量子化インデックスを２値化したものの１つ以上のリーディングビン（ｌｅａｄｉｎｇ　ｂｉｎｓ）についての以前に復号化されたニューラルネットワークパラメータに依存する確率モデルを用いること、及び、前記１つ以上のリーディングビンに後続する、前記量子化インデックスを２値化したものの等確率バイパスモードサフィックスビンを用いることによって、前記現在のニューラルネットワークパラメータ（１３’）のための前記量子化インデックス（５６）を前記データストリーム（１４）から２値算術符号化を用いて

復号化するように構成される、請求項１乃至３８のいずれかに記載の装置。
【請求項４０】
前記量子化インデックスを２値化したもののサフィックスビンは、絶対値が前記１つ以上の前記リーディングビンによって表現可能な最大絶対値を超える前記量子化インデックスの値を２値化するためのサフィックス２値化の２値化コードのビンを表し、前記装置は、以前に復号化されたニューラルネットワークパラメータの前記量子化インデックスに応じて前記サフィックス２値化を選択するように構成される、請求項３９に記載の装置。
【請求項４１】
前記ニューラルネットワークパラメータは、前記ニューラルネットワーク（１０）が表現される再構成層のうちの１つの再構成層に関連し、前記装置は、
前記ニューラルネットワークパラメータを、ニューラルネットワークパラメータ単位で、１つ以上の別の再構成層の対応するニューラルネットワークパラメータと組み合わせることによって、前記ニューラルネットワークを再構成するように構成される、
請求項１乃至４０のいずれかに記載の装置。
【請求項４２】
前記現在のニューラルネットワークパラメータに対応する、対応ニューラルネットワークパラメータに依存する確率モデルを用いる算術符号化を用いて、前記現在のニューラルネットワークパラメータ（１３’）のための前記量子化インデックス（５６）を前記データストリーム（１４）から復号化するように構成される、請求項４１に記載の装置。
【請求項４３】
ニューラルネットワークを定義するニューラルネットワークパラメータをデータストリームに符号化するための装置であって、
現在のニューラルネットワークパラメータ（１３’）について、前記データストリーム（１４）に符号化された、以前に符号化されたニューラルネットワークパラメータのための量子化インデックス（５８）に依存して、複数（５０）の再構成レベルセット（５２）の中から再構成レベルセット（４８）を選択（５４）することと、
選択された前記再構成レベルセット（４８）の１つの再構成レベル上に前記現在のニューラルネットワークパラメータ（１３’）を量子化（６４）することと、
前記現在のニューラルネットワークパラメータのための量子化インデックス（５６）が量子化される前記１つの再構成レベルを示す、前記現在のニューラルネットワークパラメータ（１３’）のための前記量子化インデックス（５６）を前記データストリーム（１４）に符号化することと
により、前記ニューラルネットワークパラメータ（１３’）を順次符号化するように構成され、
前記装置は、
前記現在のニューラルネットワークパラメータ（１３’）に関連づけられた状態に応じて、前記現在のニューラルネットワークパラメータ（１３’）について、前記複数（５０）の再構成レベルセット（５２）の中から前記再構成レベルセット（４８）を決定すること、及び、
前記データストリームに符号化された、直前のニューラルネットワークパラメータのための前記量子化インデックス（５８）に応じて、後続のニューラルネットワークパラメータの状態を更新すること
による状態遷移プロセスにより、前記現在のニューラルネットワークパラメータ（１３’）について、前記複数（５０）の再構成レベルセット（５２）の中から前記再構成レベルセット（４８）を選択（５４）し、
前記現在のニューラルネットワークパラメータ（１３’）の状態および以前に符号化されたニューラルネットワークパラメータの前記量子化インデックスに依存する（１２２）確率モデルを用いる算術符号化を用いて、前記データストリーム（１４）から前記現在のニューラルネットワークパラメータ（１３’）のための前記量子化インデックス（５６）を符号化し、

前記現在のニューラルネットワークパラメータ（１３’）に関連付けられた前記状態に依存して、複数の確率モデルの中から確率モデルのサブセットを事前選択し、以前に符号化されたニューラルネットワークパラメータのための前記量子化インデックスに依存して（１２１）、前記確率モデルのサブセットの中から前記現在のニューラルネットワークパラメータのための前記確率モデルを選択し、
前記現在のニューラルネットワークパラメータが関連する部分に隣接する前記ニューラルネットワークの部分に関連する、以前に符号化されたニューラルネットワークパラメータの前記量子化インデックスの値に依存して、前記現在のニューラルネットワークパラメータのための確率モデルを第１、第２および第３の確率モデルを含むサブセットの中から選択する
ように構成され、
ここで、前記以前に符号化されたニューラルネットワークパラメータがゼロよりも小さければ、前記第１の確率モデルが選択され、
前記以前に符号化されたニューラルネットワークパラメータがゼロよりも大きければ、前記第２の確率モデルが選択され、
前記以前に符号化されたニューラルネットワークパラメータがゼロに等しければ、前記第３の確率モデルが選択される、
装置。
【請求項４４】
ニューラルネットワーク（１０）を定義するニューラルネットワークパラメータ（１３）をデータストリーム（１４）から復号化するための方法（４００）であって、
現在のニューラルネットワークパラメータ（１３’）について、前記データストリーム（１４）から復号化された、以前のニューラルネットワークパラメータのための量子化インデックス（５８）に依存して、複数（５０）の再構成レベルセット（５２）の中から再構成レベルセット（４８）を選択する（５４）ことと、
前記現在のニューラルネットワークパラメータについての前記選択された再構成レベルセット（４８）のうちの１つの再構成レベルを示す、前記現在のニューラルネットワークパラメータ（１３’）のための量子化インデックス（５６）を前記データストリーム（１４）から復号化する（４２０）ことと、
前記現在のニューラルネットワークパラメータのための前記量子化インデックス（５６）によって示される前記選択された再構成レベルセット（４８）のうちの前記１つの再構成レベル上に前記現在のニューラルネットワークパラメータ（１３’）を逆量子化する（６２）ことと、
によって、前記ニューラルネットワークパラメータ（１３）を順次復号化するステップを含み、
前記方法はさらに、
前記現在のニューラルネットワークパラメータ（１３’）に関連づけられた状態に応じて、前記現在のニューラルネットワークパラメータ（１３’）について、前記複数（５０）の再構成レベルセット（５２）の中から前記再構成レベルセット（４８）を決定すること、及び、
前記データストリームから復号化された、直前のニューラルネットワークパラメータのための前記量子化インデックス（５８）に応じて、後続のニューラルネットワークパラメータの状態を更新すること
による状態遷移プロセスにより、前記現在のニューラルネットワークパラメータ（１３’）について、前記複数（５０）の再構成レベルセット（５２）の中から前記再構成レベルセット（４８）を選択（５４）するステップと、
前記現在のニューラルネットワークパラメータ（１３’）の状態および以前に復号化されたニューラルネットワークパラメータの前記量子化インデックスに依存する（１２２）確率モデルを用いる算術符号化を用いて、前記データストリーム（１４）から前記現在のニューラルネットワークパラメータ（１３’）のための前記量子化インデックス（５６）を復号

化するステップと、
前記現在のニューラルネットワークパラメータ（１３’）に関連付けられた前記状態に依存して、複数の確率モデルの中から確率モデルのサブセットを事前選択し、以前に復号化されたニューラルネットワークパラメータのための前記量子化インデックスに依存して（１２１）、前記確率モデルのサブセットの中から前記現在のニューラルネットワークパラメータのための前記確率モデルを選択するステップと、
前記現在のニューラルネットワークパラメータが関連する部分に隣接する前記ニューラルネットワークの部分に関連する、以前に復号化されたニューラルネットワークパラメータの前記量子化インデックスの値に依存して、前記現在のニューラルネットワークパラメータのための確率モデルを第１、第２および第３の確率モデルを含むサブセットの中から選択するステップと
を含み、
ここで、前記以前に復号化されたニューラルネットワークパラメータがゼロよりも小さければ、前記第１の確率モデルが選択され、
前記以前に復号化されたニューラルネットワークパラメータがゼロよりも大きければ、前記第２の確率モデルが選択され、
前記以前に復号化されたニューラルネットワークパラメータがゼロに等しければ、前記第３の確率モデルが選択される、
方法。
【請求項４５】
ニューラルネットワークを定義するニューラルネットワークパラメータをデータストリームに符号化するための方法（５００）であって、
現在のニューラルネットワークパラメータ（１３’）について、前記データストリーム（１４）に符号化された、以前に符号化されたニューラルネットワークパラメータのための量子化インデックス（５８）に依存して、複数（５０）の再構成レベルセット（５２）の中から再構成レベルセット（４８）を選択（５４）することと、
前記現在のニューラルネットワークパラメータ（１３’）を、前記選択された再構成レベルセット（４８）の１つの再構成レベル上に量子化する（６４）ことと、
前記現在のニューラルネットワークパラメータのための量子化インデックス（５６）が量子化される前記１つの再構成レベルを示す、前記現在のニューラルネットワークパラメータのための前記量子化インデックス（５６）を前記データストリーム（１４）に符号化する（５３０）ことと、
によって、前記ニューラルネットワークパラメータ（１３）を順次符号化するステップを含み、
前記方法はさらに、
前記現在のニューラルネットワークパラメータ（１３’）に関連づけられた状態に応じて、前記現在のニューラルネットワークパラメータ（１３’）について、前記複数（５０）の再構成レベルセット（５２）の中から前記再構成レベルセット（４８）を決定すること、及び、
前記データストリームに符号化された、直前のニューラルネットワークパラメータのための前記量子化インデックス（５８）に応じて、後続のニューラルネットワークパラメータの状態を更新すること
による状態遷移プロセスにより、前記現在のニューラルネットワークパラメータ（１３’）について、前記複数（５０）の再構成レベルセット（５２）の中から前記再構成レベルセット（４８）を選択（５４）するステップと、
前記現在のニューラルネットワークパラメータ（１３’）の状態および以前に符号化されたニューラルネットワークパラメータの前記量子化インデックスに依存する（１２２）確率モデルを用いる算術符号化を用いて、前記現在のニューラルネットワークパラメータ（１３’）のための前記量子化インデックス（５６）を前記データストリーム（１４）に符号化するステップと、

前記現在のニューラルネットワークパラメータ（１３’）に関連付けられた前記状態に依存して、複数の確率モデルの中から確率モデルのサブセットを事前選択し、以前に符号化されたニューラルネットワークパラメータのための前記量子化インデックスに依存して（１２１）、前記確率モデルのサブセットの中から前記現在のニューラルネットワークパラメータのための前記確率モデルを選択するステップと、
前記現在のニューラルネットワークパラメータが関連する部分に隣接する前記ニューラルネットワークの部分に関連する、以前に符号化されたニューラルネットワークパラメータの前記量子化インデックスの値に依存して、前記現在のニューラルネットワークパラメータのための確率モデルを第１、第２および第３の確率モデルを含むサブセットの中から選択するステップと
を含み、
ここで、前記以前に符号化されたニューラルネットワークパラメータがゼロよりも小さければ、前記第１の確率モデルが選択され、
前記以前に符号化されたニューラルネットワークパラメータがゼロよりも大きければ、前記第２の確率モデルが選択され、
前記以前に符号化されたニューラルネットワークパラメータがゼロに等しければ、前記第３の確率モデルが選択される、
方法。
【請求項４６】
プログラムが１つ以上のコンピュータ上で実行されるときに、請求項４４又は４５に記載の方法を実行するためのコンピュータプログラムを記憶した、非一時的デジタル記憶媒体。
【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の符号化概念に関連する。
【背景技術】
【０００２】
ニューラルネットワークの最も基本的な形態は、アフィン変換の連鎖と、それに続く要素ごとの非線形関数である。図１に示すように、有向非巡回グラフ（ｄｉｒｅｃｔｅｄ ａｃｙｃｌｉｃ ｇｒａｐｈ）として表現することができる。図１は、ニューラルネットワークの一例を示す模式図であり、ここでは例示的に２層フィードフォワードニューラルネットワークである。すなわち、図１は、フィードフォワードニューラルネットワークをグラフで表現したものである。具体的には、この２層構造のニューラルネットワークは、４次元の入力ベクトルを実線にマッピングする非線型関数である。このニューラルネットワークは、ニューラルネットワークの入力となるＩｎｐｕｔ層における４次元の入力ベクトルに応じた４つのニューロン１０ｃと、Ｈｉｄｄｅｎ層における５つのニューロン１０ｃと、ニューラルネットワークの出力を形成するＯｕｔｐｕｔ層における１つのニューロン１０ｃとを含む。ニューラルネットワークは、異なる－又は後続の－層のニューロンを接続する、ニューロン相互接続１１をさらに含む。ニューロン相互接続１１は、重みを関連付けてもよく、重みは、互いに接続されたニューロン１０ｃの間の関係に関連付けられる。特に、重みは、後続の層に転送されるとき、ある層のニューロンの活性化を重み付けし、順次、その後続の層の各ニューロンでインバウンドの重み付き活性化の合計が形成され－線形関数に対応する－続いて、後続の層の各ニューロン／ノードで形成される重み付け合計に非線形スカラー関数が適用され－非線形関数に対応する。したがって、各ノード、例えばニューロン１０ｃは、特定の値を伴い、この値は、エッジ、例えばニューロン相互接続１１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に伝播される。その後、すべての受信値は単純に集計される。
【０００３】

数学的には、図１のニューラルネットワークは、次のように出力を計算することになる。

$$\mathrm{output} = \sigma(W_2 \cdot \sigma(W_1 \cdot \mathrm{input}))$$

ここで、Ｗ２、Ｗ１はニューラルネットワークのパラメータ、例えば、ニューラルネットワークの重みパラメータ（エッジ重み）であり、シグマは何らかの非線形関数である。例えば、［１］にあるように、いわゆる畳み込み層を行列－行列積としてキャストすることで使用することも可能である。以下、与えられた入力から出力を計算する手順を推論と呼ぶことにする。また、例えば上記の第１ドット積＋非直線性の計算のような、線形変換＋要素単位の非直線性を構成する中間結果を隠れ層あるいは隠れ活性化値と呼ぶことにする。

【発明の概要】

【発明が解決しようとする課題】

【０００４】

通常、ニューラルネットワークは数百万個のパラメータを持ち、その表現には数百ＭＢ（例：メガバイト）が必要とされることがある。結果として、それらの推論手順には大きな行列間の多数のドット積演算の計算が含まれるため、実行には高い計算資源が必要となる。そのため、ドット積演算の複雑さを軽減することは非常に重要である。

【０００５】

また、上記の問題点に加えて、ニューラルネットワークの膨大なパラメータを保存する必要があり、例えばサーバからクライアントへ送信する必要がある場合もある。さらに、例えば連合学習環境（ｆｅｄｅｒａｔｅｄ　ｌｅａｒｎｉｎｇ　ｅｎｖｉｒｏｎｍｅｎｔ）において、または特定の受信者が支払った、または推論にニューラルネットワークを使用するときに対処することができる品質の異なる段階でニューラルネットワークのパラメータ化を提供する場合に、ニューラルネットワークのパラメータ化に関する情報を徐々にエンティティに提供することができることが有利である場合もある。

【０００６】

したがって、例えば圧縮の観点からより効率的な、ニューラルネットワークパラメータの効率的な符号化のための概念を提供することが望まれている。さらに、または代替的に、ニューラルネットワークパラメータのためのビットストリーム、ひいては信号化コストを低減することが望まれる。

【０００７】

この目的は、本願の独立請求項の主題によって達成される。

【０００８】

本発明によるさらなる実施形態は、本願の従属請求項の主題によって定義される。

【課題を解決するための手段】

【０００９】

本発明の第１の側面による実施形態は、ニューラルネットワークを定義するニューラルネットワークパラメータをデータストリームから復号化する装置であって、現在のニューラルネットワークパラメータについて、以前のニューラルネットワークパラメータについてデータストリームから復号化された量子化インデックスに応じて複数の再構成レベルセットから再構成レベルセットを選択することによって、ニューラルネットワークパラメータを順次復号化するように構成されている、装置である。さらに、装置は、データストリームから現在のニューラルネットワークパラメータについての量子化インデックスを復号化することによって、ここで量子化インデックスは、現在のニューラルネットワークパラメータについての選択された再構成レベルセットのうちの１つの再構成レベルを示し、及び、現在のニューラルネットワークパラメータについての量子化インデックスによって示される、選択された再構成レベルセットのうちの１つの再構成レベル上に現在のニューラルネットワークパラメータを逆量子化することによって、ニューラルネットワークパラメ

ータを順次復号化するように構成される。

【００１０】

本発明の第１の側面によるさらなる実施形態は、ニューラルネットワークを定義するニューラルネットワークパラメータをデータストリームに符号化する装置であって、現在のニューラルネットワークパラメータについて、以前に符号化されたニューラルネットワークパラメータについてデータストリームに符号化された量子化インデックスに応じて複数の再構成レベルセットの中から再構成レベルセットを選択することによって、ニューラルネットワークパラメータを順次符号化するように構成されている、装置である。さらに、装置は、選択された再構成レベルセットのうちの１つの再構成レベルに現在のニューラルネットワークパラメータを量子化することによって、及び、現在のニューラルネットワークパラメータについての量子化インデックスが量子化される１つの再構成レベルを示す現在のニューラルネットワークパラメータについての量子化インデックスをデータストリームに符号化することにより、ニューラルネットワークパラメータを順次符号化するよう構成される。

【００１１】

本発明の第１の側面による更なる実施形態は、ニューラルネットワークを定義するニューラルネットワークパラメータをデータストリームから復号化するための方法を備える。この方法は、現在のニューラルネットワークパラメータについて、以前のニューラルネットワークパラメータについてのデータストリームから復号化された量子化インデックスに応じて複数の再構成レベルセットの中から再構成レベルセットを選択することによって、ニューラルネットワークパラメータを順次復号化するステップを備える。さらに、この方法は、データストリームから現在のニューラルネットワークパラメータについての量子化インデックスを復号化することによって、ここで、量子化インデックスが、現在のニューラルネットワークパラメータについての選択された再構成レベルセットのうちの１つの再構成レベルを示し、及び、現在のニューラルネットワークパネルについての量子化インデックスが示す選択された再構成レベルセットのうちの１つの再構成レベル上に現在のニューラルネットワークパラメータを逆量子化することによって、ニューラルネットワークパラメータを順次符号化するステップを含む。

【００１２】

本発明の第１の側面による更なる実施形態は、ニューラルネットワークを定義するニューラルネットワークパラメータをデータストリームに符号化するための方法を備える。この方法は、現在のニューラルネットワークパラメータについて、以前に符号化されたニューラルネットワークパラメータについてデータストリームに符号化された量子化インデックスに応じて複数の再構成レベルセットの中から再構成レベルセットを選択することにより、ニューラルネットワークパラメータを順次符号化することを備える。さらに、この方法は、現在のニューラルネットワークパラメータを選択された再構成レベルのうちの１つの再構成レベルに量子化することによって、及び、現在のニューラルネットワークパラメータについての量子化インデックスが量子化される１つの再構成レベルを示す現在のニューラルネットワークパラメータについての量子化インデックスをデータストリームに符号化することによって、ニューラルネットワークパラメータを順次符号化するステップを含む。

【００１３】

本発明の第１の態様による実施形態は、非定数量子化器を使用するが、ニューラルネットワークパラメータの符号化したものの中に同じものを変化させることによって、すなわち、以前のまたはそれぞれ以前に符号化されたニューラルネットワークパラメータのデータストリームから復号された、またはそれぞれ符号化された量子化インデックスに応じて再構成レベルのセットを選択することによって、ニューラルネットワークパラメータをより効率的に圧縮し得るという考え方に基づくものである。したがって、ニューラルネットワークパラメータの順序付けられたセットを参照することができる再構成ベクトルは、Ｎ次元信号空間においてより密に詰め込むことができ、ここで、Ｎは、処理されるサンプル

のセットにおけるニューラルネットワークパラメータの数を示す。このような依存量子化（ｄｅｐｅｎｄｅｎｔ　ｑｕａｎｔｉｚａｔｉｏｎ）は、復号化するための装置による復号化及び逆量子化、又は符号化するための装置による量子化及び符号化にそれぞれ使用されることができる。

【００１４】

本発明の第２の態様による実施形態は、段階、－ニューラル層におけるニューラルネットワークの層構成と区別するために再構成層と呼ばれる－において行われ時に、これらの段階で提供されるパラメータ化を次に、ニューラルネットワークパラメータ単位で組み合わせて、段階のいずれかと比較して改善されたニューラルネットワークのパラメータ化をもたらす場合、より効率のよいニューラルネットワーク符号化が達成できるかもしれないという考え方に基づく。したがって、ニューラルネットワークを定義するニューラルネットワークパラメータを再構成するための装置は、第１のニューラルネットワークパラメータ、例えば、第１の再構成層について第１の再構成層のニューラルネットワークパラメータを導出して、ニューラルネットワークパラメータごとに、第１の再構成層のニューラルネットワークパラメータ値を得ることができる。第１のニューラルネットワークパラメータは、例えば、連合学習プロセス中に以前に送信されている可能性がある。さらに、第１のニューラルネットワークパラメータは、第１の再構成層のニューラルネットワークパラメータ値であってもよい。さらに、装置は、第２のニューラルネットワークパラメータ、例えば第２の再構成層のニューラルネットワークパラメータを、例えば最終ニューラルネットワークパラメータと区別するために、第２の再構成層についてデータストリームから復号化し、ニューラルネットワークパラメータごとに、第２の再構成層ニューラルネットワークパラメータ値をもたらすように構成されている。第２のニューラルネットワークパラメータは、ニューラルネットワーク表現の観点から自己完結した意味を持たず、単に、第１の表現層のパラメータと組み合わせたときに、ニューラルネットワーク表現、すなわち、例えば、最終ニューラルネットワークパラメータをもたらすだけかもしれない。さらに、本装置は、各ニューラルネットワークパラメータについて、第１の再構成層のニューラルネットワークパラメータ値と第２の再構成層のニューラルネットワークパラメータ値とを組み合わせることにより、ニューラルネットワークパラメータを再構成するように構成される。

【００１５】

本発明の第２の側面によるさらなる実施形態は、ニューラルネットワークを定義するニューラルネットワークパラメータを、ニューラルネットワークパラメータごとに、第１の再構成層のニューラルネットワークパラメータ値を含む第１の再構成層についての第１のニューラルネットワークパラメータを用いて符号化するための装置を備える。さらに、この装置は、第２の再構成層についての第２のニューラルネットワークパラメータをデータストリームに符号化するように構成され、第２の再構成層は、ニューラルネットワークパラメータごとに第２の再構成層のニューラルネットワークパラメータ値を含み、ニューラルネットワークパラメータは、ニューラルネットワークパラメータごとに、第１の再構成層のニューラルネットワークパラメータ値と第２の再構成層のニューラルネットワークパラメータ値を組み合わせることによって再構成可能である。

【００１６】

本発明の第２の態様によるさらなる実施形態は、ニューラルネットワークを定義するニューラルネットワークパラメータを再構成するための方法を備える。この方法は、ニューラルネットワークパラメータごとに、第１の再構成層のニューラルネットワークパラメータ値をもたらすために、第１の再構成層について、例えば連合学習プロセス中に以前に送信された可能性があり、例えば第１の再構成層のニューラルネットワークパラメータと呼ばれる可能性がある第１のニューラルネットワークパラメータを導出するステップを含む。

【００１７】

さらに、本方法は、この方法は、例えば最終的な、例えば再構成されたニューラルネットワークパラメータと区別するために第２の再構成層のニューラルネットワークパラメー

タと呼ばれ得る第２のニューラルネットワークパラメータを、データストリームから第２再構成層について復号化して、ニューラルネットワークパラメータごとに、第２の再構成層のニューラルネットワークパラメータ値を生成するステップを含む。本方法は、各ニューラルネットワークパラメータについて、第１の再構成層のニューラルネットワークパラメータ値と第２の再構成層のニューラルネットワークパラメータ値とを組み合わせることによって、ニューラルネットワークパラメータを再構成するステップを含む。第２のニューラルネットワークパラメータは、ニューラル表現に関して自己完結した意味を持たず、単に、第１の表現層のパラメータと組み合わせたときに、ニューラル表現、すなわち、例えば最終ニューラルネットワークパラメータを導くかもしれない。

【００１８】

本発明の第２の側面によるさらなる実施形態は、ニューラルネットワークパラメータごとに、第１の再構成層のニューラルネットワークパラメータ値を含む第１の再構成層についての第１のニューラルネットワークパラメータを用いて、ニューラルネットワークを定義するニューラルネットワークパラメータを符号化する方法を含む。この方法は、第２の再構成層についての第２のニューラルネットワークパラメータをデータストリームに符号化するステップを含み、第２の再構成層は、ニューラルネットワークパラメータごとに、第２の再構成層ニューラルネットワークパラメータ値を含み、ニューラルネットワークパラメータは、ニューラルネットワークパラメータごとに、第１の再構成層のニューラルネットワークパラメータ値と第２の再構成層のニューラルネットワークパラメータ値とを組み合わせることによって再構成可能である。

【００１９】

本発明の第２の態様による実施形態は、例えばニューラルネットワークパラメータによって定義されるニューラルネットワークが、再構成層、例えばベース層およびエンハンスメント層などのサブ層を用いて、例えばビットストリーム内のデータ量が少ない状態で、効率的に圧縮および／または送信され得るという考えに基づいている。再構成層は、ニューラルネットワークパラメータが、各ニューラルネットワークパラメータについて、第１の再構成層のニューラルネットワークパラメータ値と第２の再構成層のニューラルネットワークパラメータ値とを組み合わせることによって再構成可能であるように、定義されることができる。この分布は、ニューラルネットワークパラメータの効率的な符号化、例えば、符号化および／または復号化、および／または送信を可能にする。したがって、第２の再構成層についての第２のニューラルネットワークパラメータは、データストリームに別々に符号化および／または送信されることができる。

【図面の簡単な説明】

【００２０】

図面は必ずしも縮尺通りではなく、代わりに一般的に本発明の原理を説明することに重点が置かれている。以下の説明では、本発明の様々な実施形態が、以下の図面を参照して説明され、その中で、本発明の実施形態が説明される。

【図１】図１は、本発明の実施形態と共に使用され得る２層フィードフォワードニューラルネットワークの例示の概略図である。

【図２】図２は、実施形態による、ニューラルネットワークを定義するニューラルネットワークパラメータをデータストリームから復号化するための装置内で実行される逆量子化のための概念を示す概略図である。

【図３】図３は、実施形態による、ニューラルネットワークパラメータをデータストリームに符号化するための装置内で実行される量子化のための概念を示す概略図である。

【図４】図４は、実施形態による、ニューラルネットワークを定義するニューラルネットワークパラメータを再構成するための装置内で実行される復号化のための概念を示す概略図である。

【図５】図５は、実施形態による、ニューラルネットワークを定義するニューラルネットワークパラメータを再構成するための装置内で実行される符号化のための概念を示す概略図である。

【図６】図６は、本発明による実施形態で使用するための、ニューラルネットワークパラメータについての再構成層を使用する概念の概略図である。
【図７】図７は、本発明の実施形態による均一再構成量子化器のイラストを示す概略図である。
【図８】図８は、本発明の実施形態による２つの重みパラメータの単純なケースについて、許容される再構成ベクトルの位置の例を示す図である。
【図９ａ】図９ａは、本発明の実施形態による単一の量子化ステップサイズΔによって完全に決定される再構成レベルの２つのセットを有する依存量子化のための例を示す図である。
【図９ｂ】図９ｂは、本発明の実施形態による単一の量子化ステップサイズΔによって完全に決定される再構成レベルの２つのセットを有する依存量子化のための例を示す図である。
【図９ｃ】図９ｃは、本発明の実施形態による単一の量子化ステップサイズΔによって完全に決定される再構成レベルの２つのセットを有する依存量子化のための例を示す図である。
【図１０】図１０は、本発明の実施形態による、ニューラルネットワークパラメータの再構成プロセスについての好ましい例を示す疑似コードのための例を示す図である。
【図１１】図１１は、本発明の実施形態による再構成レベルセットを２つのサブセットに分割するための一例を示す図である。
【図１２】図１２は、実施形態による層についてのニューラルネットワークパラメータの再構成プロセスについての好ましい例を示す擬似コードの例である。
【図１３】図１３は、本発明の実施形態による状態遷移表ｓｔｔａｂ及び表ｓｅｔＩｄの好ましい例を示す図であり、これは、状態に関連する量子化セットを指定する。
【図１４】図１４は、本発明の実施形態による状態遷移表ｓｔｔａｂび表ｓｅｔＩｄの好ましい例を示す図であり、これは、状態に関連する量子化セットを指定する。
【図１５】図１５は、本発明の実施形態による、０に等しい量子化インデックスが状態遷移および依存スカラー量子化から除外される、ニューラルネットワークパラメータレベルの代替再構成プロセスを示す擬似コードである。
【図１６】図１６は、本発明の実施形態による、トレリス構造としての依存スカラー量子化における状態遷移の例を示す図である。
【図１７】図１７は、本発明の実施形態による基本トレリスセルの一例を示す図である。
【図１８】図１８は、本発明の実施形態による８つのニューラルネットワークパラメータの依存スカラー量子化のためのトレリス例を示す図である。
【図１９】図１９は、本発明の実施形態による、コスト量（ラグランジュコスト量Ｄ＋λ・Ｒなど、Ｌａｇｒａｎｇｉａｎ　ｃｏｓｔ　ｍｅａｓｕｒｅ）を最小化する量子化インデックスのシーケンス（またはブロック）を決定するために利用することができるトレリス構造例を示す図である。
【図２０】図２０は、本発明の実施形態による、ニューラルネットワークを定義するニューラルネットワークパラメータをデータストリームから復号化する方法のブロック図である。
【図２１】図２１は、本発明の実施形態による、ニューラルネットワークを定義するニューラルネットワークパラメータをデータストリームに符号化するための方法のブロック図である。
【図２２】図２２は、本発明の実施形態による、ニューラルネットワークを定義するニューラルネットワークパラメータを再構成する方法のブロック図である。
【図２３】図２３は、本発明の実施形態による、ニューラルネットワークを定義する、ニューラルネットワークパラメータを符号化するための方法のブロック図である。
【発明を実施するための形態】
【００２１】
等しいまたは同等の機能を有する要素または要素は、異なる図で表されていても、以下

の説明では、等しいまたは同等の参照数字で示される。
【００２２】
以下の説明では、本発明の実施形態のより詳細な説明を提供するために、複数の詳細が記載されている。しかし、本発明の実施形態は、これらの特定の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しないために、周知の構造およびデバイスを詳細ではなくブロック図の形態で示す。加えて、本明細書で後述する異なる実施形態の特徴は、特に断らない限り、互いに組み合わせることができる。
【００２３】
本説明は、本願のいくつかの実施形態の提示から始まる。この説明は、かなり一般的なものであるが、本願の実施形態が基づいている機能性の概要を読者に提供するものである。その後、これらの機能性のより詳細な説明が、実施形態の動機と、それらがどのように上述の効率向上を達成するかと共に提示される。詳細は、現在説明されている実施形態と、個別に、および組み合わせて、組み合わせることが可能である。
【００２４】
図２は、実施形態によるデータストリームからニューラルネットワークを定義するニューラルネットワークパラメータを復号化するための装置内で実行される逆量子化についての概念を示す概略図である。ニューラルネットワークは、例えば、相互接続された層のニューロン間のニューロン相互接続を有する、複数の相互接続されたニューラルネットワーク層を含むことができる。図２は、データストリーム１４における、例えば符号化されたニューラルネットワークパラメータ１３の量子化インデックス５６を示す。ニューラルネットワークパラメータ１３は、したがって、そのニューロン間の重みの観点など、ニューラルネットワークを定義またはパラメータ化することができる。
【００２５】
本装置は、ニューラルネットワークパラメータ１３を順次復号化するように構成されている。この順次プロセスの間、量子化器（再構成レベルセット）が変化させられる。この変化は、より少ない（またはより良いより密でない）レベルを有する量子化器を使用することを可能にし、したがって、より小さい量子化インデックスを符号化することを可能にし、この量子化から得られるニューラルネットワーク表現の品質が、必要な符号化ビットレートと比較して、一定の量子化器を使用するより向上される。詳細は後述する。特に、装置は、現在のニューラルネットワークパラメータ１３’について、以前のニューラルネットワークパラメータについてデータストリーム１４から復号化された量子化インデックス５８に応じて複数（５０）の再構成レベルセット５２（セット０、セット１）のうち再構成レベルセット４８（選択されたセット）を選択５４（再構成レベル選択）することによって、ニューラルネットワークパラメータ１３を順次復号する。
【００２６】
さらに、本装置は、データストリーム１４から現在のニューラルネットワークパラメータ１３’についての量子化インデックス５６を復号化することによって、ここで、量子化インデックス５６が現在のニューラルネットワークパラメータについての再構成レベルの選択されたセット４８のうちの１つの再構成レベルを示し、及び、現在のニューラルネットワークパラメータ１３’を現在のニューラルネットワークパネルについての量子化インデックス５６によって示される再構成レベルの選択されたセット４８の１つの再構成レベルに逆量子化６２することによって、ニューラルネットワークパネル１３を連続的に復号化するよう構成される。
【００２７】
復号化されたニューラルネットワークパラメータ１３は、一例として、行列１５ａで表される。行列は、デシリアライズされた２０ｂ（ｄｅｓｅｒｉａｌｉｚａｔｉｏｎ）ニューラルネットワークパラメータ１３を含んでもよく、これは、ニューラルネットワークのニューロン相互接続の重みに関連してもよい。
【００２８】

任意選択的に、複数（５０）の再構成レベルセット５２の、本明細書で時々量子化器とも呼ばれる再構成レベルセット５２の数は、図２に示すように、例えばセット０とセット１の２つであってよい。

【００２９】

さらに、装置は、例えば△または△ｋで示される所定の量子化ステップサイズ（ＱＰ）によって複数（５０）の再構成レベルセット５２（例えば、セット０、セット１）をパラメータ化６０（ｐａｒａｍｅｔｒｉｚａｔｉｏｎ）し、データストリーム１４から所定の量子化ステップサイズの情報を導出するように構成されることができる。したがって、実施形態による復号化器は、可変ステップサイズ（ＱＰ）に適応することができる。

【００３０】

さらに、実施形態によれば、ニューラルネットワークは１つ以上のＮＮ層を含むことができ、装置は、各ＮＮ層について、データストリーム１４からそれぞれのＮＮ層についての所定の量子化ステップサイズ（ＱＰ）の情報を導出し、それぞれのＮＮ層に属するニューラルネットワークパラメータを逆量子化するために用いられるように、各ＮＮ層について、それぞれのＮＮ層について導出された所定の量子化ステップサイズを用いて複数５０の再構成レベルセット５２をパラメータ化するよう構成され得る。ＮＮ層に関するステップサイズの適応、したがって再構成レベルセット５２の適応は、符号化効率を向上させ得る。

【００３１】

さらなる実施形態によれば、装置は、現在のニューラルネットワークパラメータ１３’について、複数５０の再構成レベルセット５２のうち再構成レベルセット４８を、以前に復号化されたニューラルネットワークパラメータについてのデータストリーム１４から復号化された量子化インデックス５８を２値化したもののＬＳＢ（例えば、最下位ビット）部分または以前に復号化されたビン（例えば、２値決定）に応じて選択５４するよう構成されるとよい。ＬＳＢ比較は、低い計算コストで実行されることができる。特に、状態遷移が用いられることができる。選択５４は、複数５０の再構成レベルセット５２のうちの量子化レベルセット４８のうちの現在のニューラルネットワークパラメータ１３’について、現在のニューラルネットワークパラメータ１３’に関連する状態に応じて複数５０の再構成レベルセット５２のうちの再構成レベルセット４８を現在のニューラルネットワークパラメータ１３’について決定することによって、かつ、直前のニューラルネットワークパラメータについてのデータストリームから復号化された量子化インデックス５８に応じて後続のニューラルネットワークパラメータについての状態を更新することによって、状態推移プロセスにより実行されることができる。また、例えば遷移表を使用することによる状態遷移以外の代替的なアプローチも同様に使用することができ、以下に規定される。

【００３２】

加えて、または代替的に、装置は、例えば、現在のニューラルネットワークパラメータ１３’について、複数５０の再構成レベルセット５２のうち再構成レベルセット４８を、以前に復号化されたニューラルネットワークパラメータについてのデータストリーム１４から復号化された量子化インデックス５８の２値関数の結果に応じて選択５４するように構成されることができる。２値関数は、例えば、量子化インデックス５８が偶数または奇数を表すかどうかを信号化する、ビット単位「ａｎｄ」演算を使用するパリティチェックであってよい。これは、量子化インデックス５８を符号化するために使用される再構成レベルセット４８に関する情報を、したがって、例えば、対応する符号化器で使用される再構成レベルセットの所定の順序のために、現在のニューラルネットワークパラメータ１３’を符号化するために使用される再構成レベルのセットに対して提供し得る。パリティは、前述した状態遷移のために使用されることができる。

【００３３】

さらに、実施形態によれば、装置は、例えば、現在のニューラルネットワークパラメータ１３’について、複数５０の再構成レベルセット５２のうち再構成レベルのセット４８を、以前に復号化されたニューラルネットワークパラメータについてデータストリーム１４

から復号化された量子化インデックス５８のパリティに応じて選択５４するように構成されることができる。パリティチェックは、例えばビット単位「ａｎｄ」演算を用いて、低い計算コストで実行されることができる。

【００３４】

任意選択で、装置は、ニューラルネットワークパラメータ１３についての量子化インデックス５６を復号化し、ニューラルネットワークパラメータ１３の間で共通の連続的な順序１４’に沿ってニューラルネットワークパラメータ１３の逆量子化を実行するように構成されることができる。言い換えれば、両方のタスクに同じ順序が使用されることができる。

【００３５】

図３は、実施形態による、ニューラルネットワークパラメータをデータストリームに符号化するための装置内で実行される量子化の概念を模式的に示す図である。図３は、ニューラルネットワーク層１０ａ、１０ｂを含むニューラルネットワーク（ＮＮ）１０を示し、層はニューロン１０ｃを含み、相互接続された層のニューロンはニューロン相互接続１１を介して相互接続されている。一例として、ＮＮ層（ｐ－１）１０ａおよびＮＮ層（ｐ）１０ｂが示され、ｐはＮＮ層のインデックスであり、１≦ｐ≦ＮＮの層の数である。ニューラルネットワークは、ニューラルネットワークパラメータ１３によって定義またはパラメータ化され、それは任意に、ニューラルネットワーク１０のニューロン相互接続１１の重みに関連し得る。図１の隠れ層のニューロン１０ｃは、図３の層ｐ（Ａ、Ｂ、Ｃ、．．．）のニューロンを表してもよく、図１の入力層のニューロンは、図３に示される層ｐ－１（ａ、ｂ、ｃ、．．．）のニューロンを表すことができる。ニューラルネットワークパラメータ１３は、図１のニューロン相互接続１１の重みに関連づけることができる。

【００３６】

異なる層のニューロン１０ｃの関係は、図１において、ニューラルネットワークパラメータ１３の行列１５ａによって表される。例えば、ネットワークパラメータ１３がニューロン相互接続１１の重みに関連する場合、行列１５ａは、例えば、行列要素が異なる層のニューロン１０ｃ間の重み（例えば、層ｐ－１の場合はａ、ｂ、・・・、層ｐの場合はＡ、Ｂ、・・・）を表すように構成されることができる。

【００３７】

この装置は、例えばシリアル２０ａ（シリアライゼーション、直列化）において、ニューラルネットワークパラメータ１３を順次符号化するように構成されている。この順次処理の間、量子化器（再構成レベルセット）は変化させられる。この変化により、より少ない（またはより良いより密でない）レベルを有する量子化器を使用することができ、したがって、より小さい量子化インデックスを符号化することができ、この量子化から得られるニューラルネットワーク表現の品質が、必要な符号化ビットレートと比較して、一定の量子化器を使用する場合よりも改善される。詳細は後述する。特に、装置は、現在のニューラルネットワークパラメータ１３’について、以前に符号化されたニューラルネットワークパラメータについてデータストリーム１４に符号化された量子化インデックス５８に応じて複数５０の再構成レベルセット５２のうち再構成レベルセット４８を選択５４することによって、ニューラルネットワークパラメータ１３を順次に符号化する。

【００３８】

さらに、装置は、現在のニューラルネットワークパラメータ１３’を選択された再構成レベルセット４８の１つの再構成レベル上に量子化６４（Ｑ）することによって、かつ、現在のニューラルネットワークパラメータについての量子化インデックス５６がデータストリーム１４中に量子化される１つの再構成レベルを示す現在のニューラルネットワークパラメータ１３’についての量子化インデックス５６を符号化することによって、ニューラルネットワークパラメータ１３を順次符号化するよう構成される。任意選択で、複数５０の再構成レベルセット５２のうち、本明細書で時々量子化器とも呼ばれる再構成レベルセット５２の数は、例えばセット０およびセット１を用いて示されるように、２つであることができる。

【００３９】

実施形態によれば、図３に示すように、装置は、例えば、所定の量子化ステップサイズ（ＱＰ）によって複数５０の再構成レベルセット５２をパラメータ化６０し、データストリーム１４に所定の量子化ステップサイズに関する情報を挿入するように構成されていることができる。これにより、例えば、量子化効率を向上させるための適応的量子化が可能となり、ニューラルネットワークパラメータ１３の符号化方法の変化が、所定の量子化ステップサイズに関する情報とともに復号器に伝達されることができる。所定の量子化ステップサイズ（ＱＰ）を使用することにより、情報の伝送のためのデータ量を削減することができる。

【００４０】

さらに、実施形態によれば、ニューラルネットワーク１０は、１つ以上のＮＮ層１０ａ、１０ｂを含んでいることができる。装置は、各ＮＮ層（ｐ；ｐ－１）について、それぞれのＮＮ層についての所定の量子化ステップサイズ（ＱＰ）の情報をデータストリーム１４に挿入し、それぞれのＮＮ層に属するニューラルネットワークパラメータの量子化に使用するように、それぞれのＮＮ層について導かれた所定の量子化ステップサイズを用いて複数５０の再構成レベルセット５２のパラメータ化を行うよう構成されることが可能である。先に説明したように、例えばＮＮ層またはＮＮ層の特性に応じた量子化の適応は、量子化効率を向上させることができる。

【００４１】

任意選択的に、装置は、現在のニューラルネットワークパラメータ１３’について、複数５０の再構成レベルセット５２のうち再構成レベルのセット４８を、以前に符号化されたニューラルネットワークパラメータについてデータストリーム１４に符号化された量子化インデックス５８を２値化したもののＬＳＢ部分または以前に符号化されたビンに応じて選択５４するよう構成されることができる。ＬＳＢ比較は、低い計算コストで実行され得る。

【００４２】

図２で説明した復号化のための装置と同様に、状態遷移が用いられることができる。現在のニューラルネットワークパラメータ１３’について、現在のニューラルネットワークパラメータ１３’に関連する状態に応じて、複数５０の再構成レベルセット５２のうちの再構成レベルのセット４８を決定することによって、及び、直前のニューラルネットワークパラメータについてのデータストリームに符号化された量子化インデックス５８に応じて、後続のニューラルネットワークパラメータについての状態を更新することによって、状態遷移プロセスにより、複数５０の再構成レベルセット５２のうちの量子化レベルセット４８のうち現在のニューラルネットワークパラメータ１３’について選択５４が実行されることができる。例えば遷移表を使用することによる状態遷移以外のアプローチも同様に使用可能であり、以下に規定される。

【００４３】

さらに、または代替的に、装置は、現在のニューラルネットワークパラメータ１３’について、以前に符号化されたニューラルネットワークパラメータについてのデータストリーム１４に符号化された量子化インデックス５８の２値関数の結果に応じて、複数５０の再構成レベルセット５２のうち再構成レベルセット４８を選択５４するよう構成されることができる。２値関数は、例えば、量子化インデックス５８が偶数または奇数を表すかどうかを信号化する、ビット単位「ａｎｄ」演算を使用したパリティチェックであってもよい。これは、量子化インデックス５８を符号化するために使用される再構成レベルセット４８に関する情報を提供してもよく、したがって、例えば、所定の順序のために、対応する復号化器が対応する再構成レベルセット４８を選択できるように、再構成レベルの所定の順序のために、現在のニューラルネットワークのパラメータ１３’についての再構成レベルのセット４８を決定することができる。パリティは、前述した状態遷移のために使用されることができる。

【００４４】

さらに、実施形態によれば、装置は、例えば、現在のニューラルネットワークパラメー

タ１３’について、複数５０の再構成レベルセット５２のうち量子化レベルセット４８を、以前に符号化されたニューラルネットワークパラメータについてのデータストリーム１４に符号化された量子化インデックス５６のパリティに応じて、選択５４するように構成されることができる。パリティチェックは、例えばビット単位「ａｎｄ」演算を使用して、低い計算コストで実行されることができる。

【００４５】

任意選択的に、装置は、ニューラルネットワークパラメータ（１３）についての量子化インデックス（５６）を符号化し、ニューラルネットワークパラメータ（１３）の量子化を、ニューラルネットワークパラメータ（１３）間の共通の連続的な順序（１４’）に沿って実行するように構成されることができる。すなわち、両方のタスクで同じ順序を使用することができる。

【００４６】

図４は、実施形態による量子化されたニューラルネットワークのパラメータを算術復号化するための概念の概略図である。これは、図２の装置内で使用することができる。したがって、図４は、図２の可能な拡張として見ることができる。それは、現在のニューラルネットワークパラメータ１３’についての量子化インデックス５６が、算術符号化、例えば２値算術符号化の使用によって任意例として示すように、図４の装置によって復号化されるデータストリーム１４を示す。例えばあるコンテキストによって定義される確率モデルが使用され、それは、矢印１２３によって示されるように、現在のニューラルネットワークパラメータ１３’について選択された再構成レベルセット４８に依存する。詳細は、本明細書において設定される。

【００４７】

図２に関して説明したように、現在のニューラルネットワークパラメータ１３’に関連する状態に応じて、現在のニューラルネットワークパラメータ１３’について複数５０の再構成レベルセット５２のうちの再構成レベルのセット４８を決定することによって、及び、直前のニューラルネットワークパラメータについてのデータストリームから復号化された量子化インデックス５８に応じて、後続のニューラルネットワークパラメータについての状態を更新することによって、状態遷移プロセスにより、複数５０の再構成レベルセット５２のうちの量子化レベルのセット４８を選択する選択５４が現在のニューラルネットワークパラメータ１３’について実行されることができる。したがって、状態は、現在のニューラルネットワークパラメータ１３’を符号化／復号化するために使用されるべき再構成レベルセット４８へのポインタに準じるものであるが、これは、しかしながら、状態が、過去のニューラルネットワークパラメータまたは過去の量子化インデックスのメモリとして疑似的に機能するように、再構成セットの数に対応する数の状態を区別するだけとして、より細かい要素で更新される。したがって、状態は、ニューラルネットワークパラメータ１３を符号化／復号化するために使用される再構成レベルのセットの順序を規定する。図４によれば、例えば、現在のニューラルネットワークパラメータ（１３’）についての量子化インデックス（５６）は、現在のニューラルネットワークパラメータ（１３’）についての状態に対応する（１２２）確率モデルを用いて、算術符号化を用いてデータストリーム（１４）から復号化される。状態に応じて確率モデルを適応させることにより、確率モデルの推定が良好となり、符号化効率が向上する可能性がある。さらに、状態に応じて適応させることで、少ない追加データの送信で計算効率のよい適応が可能になる場合がある。

【００４８】

さらなる実施形態によれば、装置は、例えば、量子化インデックス５６を２値化８２したものの少なくとも１つのビン８４についての現在のニューラルネットワークパラメータ１３’についての状態に依存１２２する確率モデルを用いて、２値算術コーディングを使ってデータストリーム１４から現在のニューラルネットワークパラメータ１３’についての量子化インデックス５６を復号化するよう構成されることができる。

【００４９】

さらに、または代替的に、装置は、確率モデルの依存性が、依存性を用いるニューラル

ネットワークパラメータについてのコンテキストのセットのうちコンテキスト８７の選択１０３（導出）を含むように構成されてもよく、各コンテキストは、所定の確率モデルが関連づけられるように構成される。使用される確率推定が優れているほど、圧縮の効率は高くなる。確率モデルは、例えば、コンテキスト適応的（２値）算術符号化を用いて、更新されることができる。

【００５０】

任意選択で、装置は、それぞれのコンテキストを用いて算術符号化された量子化インデックスに基づいて、それぞれのコンテキストに関連する所定の確率モデルを更新するように構成されることができる。このように、コンテキストの確率モデルは、実際の統計に適応される。

【００５１】

さらに、装置は、例えば、量子化インデックスを２値化したものの少なくとも１つのビンについて、現在のニューラルネットワークパラメータ１３’に対して選択された再構成レベルセット４８に対応する確率モデルを用いて、２値算術符号化を用いてデータストリーム１４から現在のニューラルネットワークパラメータ１３’についての量子化インデックス５６を復号化するよう構成されることができる。

【００５２】

任意選択で、少なくとも１つのビンは、現在のニューラルネットワークパラメータの量子化インデックス５６がゼロに等しいか否かを示す有意性ビンを含んでいることができる。さらに、または代替的に、少なくとも１つのビンは、現在のニューラルネットワークパラメータの量子化インデックス５６がゼロより大きいか、またはゼロより小さいかを示す符号ビンを含んでもよい。さらに、少なくとも１つのビンは、現在のニューラルネットワークパラメータの量子化インデックス５６の絶対値がＸより大きいか否かを示すｇｒｅａｔｅｒ－ｔｈａｎ－Ｘビンを含んでもよく、ここでＸはゼロより大きい整数である。

【００５３】

以下、図５では、図４を用いて説明した復号化のための概念の対極を説明することができる。したがって、すべての説明および利点は、以下の符号化のための概念の側面に適宜適用することができる。

【００５４】

図５は、実施形態によるニューラルネットワークパラメータを算術符号化するための概念を示す概略図である。これは、図３の装置内で使用され得る。したがって、図５は、図３の可能な拡張として見ることができる。それは、現在のニューラルネットワークパラメータ１３’についての量子化インデックス５６が、算術符号化、例えば任意の例として２値算術符号化の使用によって示されるように図３の装置によって符号化されるデータストリーム１４を示す。例えばあるコンテキストによって定義される確率モデルが使用され、それは、矢印１２３によって示されるように、現在のニューラルネットワークパラメータ１３’に対して選択された再構成レベルセット４８に依存する。詳細は本明細書において設定される。

【００５５】

図３に関して説明したように、現在のニューラルネットワークパラメータ１３’について選択５４が実行される。選択５４は、現在のニューラルネットワークパラメータ１３’について、現在のニューラルネットワークパラメータ１３’に関連する状態に応じて、複数５０の再構成レベルセット５２のうちの量子化レベルセット４８を決定することによって、及び、直前のニューラルネットワークパラメータについてのデータストリームに符号化された量子化インデックス５８に応じて、後続のニューラルネットワークパラメータについての状態を更新することによって、状態遷移プロセスにより複数５０の再構成レベルセット５２のうちの量子化レベルセット４８を選択する。

【００５６】

したがって、状態は、現在のニューラルネットワークパラメータ１３’を符号化／復号化するために使用されるべき再構成レベルセット４８へのポインタに準ずるものであるが、

しかし、状態が疑似的に、過去のニューラルネットワークパラメータまたは過去の量子化インデックスのメモリとして機能するように、再構成セットの数に対応する数の状態を区別するだけとしてより細かい要素で更新されている。したがって、状態は、ニューラルネットワークパラメータ１３を符号化／復号化するために使用される再構成レベルセットの順序を定義する。

【００５７】

さらに、現在のニューラルネットワークパラメータ１３’についての量子化インデックス５６は、現在のニューラルネットワークパラメータ１３’についての状態１２２に対応する確率モデルを用いる算術符号化を用いてデータストリーム１４に符号化されることができる。

【００５８】

例えば図３によれば、量子化インデックス５６は、量子化インデックス５６を２値化８２したものの少なくとも１つのビン８４に対する現在のニューラルネットワークパラメータ１３’に対する状態１２２に対応する確率モデルを用いて、２値算術コーディングを用いてデータストリーム１４へ符号化される。確率モデルは確率モデルの推定に適している可能性があるため、状態に応じて確率モデルを適応させると、符号化効率が向上する可能性がある。さらに、状態に応じた適応は、送信される追加データの量が少なく、計算効率のよい適応を可能にし得る。

【００５９】

さらに、または代替的に、本装置は、確率モデルの依存性が、依存性を用いたニューラルネットワークパラメータのコンテキストのセットのうちコンテキスト８７の選択１０３（導出）を含み、各コンテキストは所定の確率モデルが関連づけられているように構成されることができる。

【００６０】

任意選択で、装置は、それぞれのコンテキストを使用して算術符号化された量子化インデックスに基づいて、それぞれのコンテキストに関連付けられた所定の確率モデルを更新するように構成されることができる。

【００６１】

さらに、装置は、例えば、量子化インデックスを２値化したものの少なくとも１つのビンについて、現在のニューラルネットワークパラメータ１３’について選択された再構成レベルセット４８に対応する確率モデルを用いることによって、２値算術符号化を用いて、データストリーム１４に現在のニューラルネットワークパラメータ１３’の量子化インデックス５６を符号化するよう構成されることができる。２値算術符号化を使用するために、量子化インデックス５６は、２値化（ｂｉｎａｒｉｚａｔｉｏｎ）されることができる。

【００６２】

任意選択で、少なくとも１つのビンは、現在のニューラルネットワークパラメータの量子化インデックス５６がゼロに等しいか否かを示す有意性ビンを含むことができる。さらに、または代替的に、少なくとも１つのビンは、現在のニューラルネットワークパラメータの量子化インデックス５６がゼロより大きいか、またはゼロより小さいかを示す符号ビンを含むことができる。さらに、少なくとも１つのビンは、現在のニューラルネットワークパラメータの量子化インデックス５６の絶対値がＸより大きいか否かを示すｇｒｅａｔｅｒ－ｔｈａｎ－Ｘビンを含むことができ、ここでＸはゼロより大きい整数である。

【００６３】

次に説明する実施形態は、次のような本願の別の態様に集中している。ニューラルネットワークのパラメータ化が段階または再構成層で符号化され、ＮＮパラメータごとに、各ステージからの1つの値は、ニューラルネットワークの改良された／強化された表現を得るために結合される必要があり、少なくとも１つがそれ自体ニューラルネットワークの妥当な表現を表すかもしれないが、低質で貢献する段階のいずれかに強化されているが、後者の可能性は本態様に必須ではないものである。

【００６４】

図６は、本発明による実施形態で使用するためのニューラルネットワークパラメータのための再構成層を使用する概念の概略図である。図６は、例えば第２の再構成層である再構成層ｉ、例えば第１の再構成層である再構成層ｉ－１、および例えば図３からの層１０ｂであるニューラルネットワーク（ＮＮ）層ｐが、例えば図３からの行列１５ａなどのアレイまたは行列の形態で表された層であることを示している。

【００６５】

図６は、ニューラルネットワークを定義するニューラルネットワークパラメータ１３を再構成するための装置３１０の概念を示している。したがって、この装置は、例えば連合学習プロセスの間に以前に送信された可能性があり、例えば、第１の再構成層、例えば再構成層ｉ－１について、第１の再構成層のニューラルネットワークパラメータと呼ばれてもよい第１ニューラルネットワークパラメータ１３ａを導出し、ニューラルネットワークパラメータごと、例えば重みごとまたはニューロン間接続ごとに、第１の再構成層のニューラルネットワークパラメータ値をもたらすよう構成されている。この導出は、そうでなければ、第１のニューラルネットワークパラメータ１３ａを復号化すること、または受け取ることを含むことができる。さらに、装置は、ニューラルネットワークパラメータ１３ごとに第２の再構成層のニューラルネットワークパラメータ値を生成するためのデータストリーム１４から第２の再構成層についての例えば最終ニューラルネットワークパラメータ、例えばパラメータ１３と区別するために第２の再構成層のニューラルネットワークパラメータと呼ばれ得る第２ニューラルネットワークパラメータ１３ｂを復号化３１２するよう構成されている。したがって、第１および第２の再構成層の２つの寄与値が、ＮＮパラメータごとに得られてもよく、第１および／または第２のＮＮパラメータ値の符号化／復号化は、図２および図３に従った依存量子化（ｄｅｐｅｎｄｅｎｔ　ｑｕａｎｔｉｚａｔｉｏｎ）、および／または図４および図５で説明したような量子化インデックスの算術符号化／復号化を使用することができる。第２のニューラルネットワークパラメータ１３ｂは、ニューラル表現の観点から自己完結した意味を持たず、単に、第１の表現層のパラメータと組み合わせたときに、ニューラルネットワーク表現、すなわち最終的なニューラルネットワークパラメータを導くだけかもしれない。

【００６６】

さらに、装置は、各ニューラルネットワークパラメータについて、第１の再構成層のニューラルネットワークパラメータ値と第２再構成層ニューラルネットワークパラメータ値とを、例えば要素単位の加算および／または乗算を用いて組み合わせる（ＣＢ）ことにより、ニューラルネットワークパラメータ１３を再構成３１４するように構成されている。

【００６７】

さらに、図６は、第１の再構成層、例えば再構成層ｉ－１のための第１のニューラルネットワークパラメータ１３ａを用いて、ニューラルネットワークを定義するニューラルネットワークパラメータ１３を符号化するための装置３２０の概念を示している。第１の再構成層は、ニューラルネットワークパラメータ１３ごとに、第１の再構成層のニューラルネットワークパラメータ値を含む。したがって、装置は、第２の再構成層、例えば再構成層ｉのための第２のニューラルネットワークパラメータ１３ｂをデータストリームに符号化３２２するように構成される。第２の再構成層は、ニューラルネットワークパラメータ１３ごとに、第２の再構成層のニューラルネットワークパラメータ値を含む。ニューラルネットワークパラメータ１３は、それぞれのニューラルネットワークパラメータについて、例えば要素単位の加算および／または乗算を用いて、第１の再構成層のニューラルネットワークパラメータ値と第２の再構成層のニューラルネットワークパラメータ値とを組み合わせる（ＣＢ）ことによって再構成可能である。

【００６８】

任意選択的に、装置３１０は、データストリーム１４から、または別個のデータストリームから、第１の再構成層についての第１のニューラルネットワークパラメータを復号化３１６するように構成されることができる。

【００６９】

簡単に言えば、ニューラルネットワークパラメータ１３の分解は、パラメータのより効率的な符号化および／または復号化および送信を可能にし得る。

【００７０】

以下では、特に、ニューラルネットワーク符号化概念を含む、さらなる実施形態が開示される。以下の説明では、上述した実施形態と個別に、および組み合わせて使用することができる更なる詳細を提供する。

【００７１】

まず、本発明の実施形態による依存スカラー量子化（Ｄｅｐｅｎｄｅｎｔ　Ｓｃａｌａｒ　Ｑｕａｎｔｉｚａｔｉｏｎ）を伴うニューラルネットワークのパラメータのエントロピー符号化のための方法を提示する。

【００７２】

依存スカラー量子化を用いたニューラルネットワークパラメータ１３（重み、重みパラメータまたはパラメータとも呼ばれる）のセットのパラメータ符号化方法について説明する。本明細書で提示されるパラメータ符号化は、パラメータ１３の依存スカラ量子化（例えば、図３の文脈で説明したような）および得られた量子化インデックス５６のエントロピー符号化（例えば、図５の文脈で説明したような）から構成される。復号化器側では、量子化インデックス５６のエントロピー復号化（例えば、図４の文脈で説明したように）と、ニューラルネットワークパラメータ１３の依存再構成（例えば、図２の文脈で説明したように）とによって、再構成されたニューラルネットワークパラメータのセットを得ることができる。独立したスカラー量子化およびエントロピー符号化を伴うパラメータ符号化とは対照的に、ニューラルネットワークパラメータ１３についての許容可能な再構成レベルセットは、再構成順序において現在のニューラルネットワークパラメータ１３’に先行する送信済み量子化インデックス５６に依存する。以下に示す提示は、依存スカラ量子化で使用される再構成レベルを指定する量子化インデックスのエントロピー符号化のための方法を追加的に説明する。

【００７３】

本説明は、主にニューラルネットワーク圧縮におけるニューラルネットワークパラメータ層の非可逆符号化を対象としているが、他の分野の非可逆符号化にも適用可能である。

【００７４】

本装置の方法論は、以下のような異なる主要部分に分けられる。

【００７５】

１．量子化
２．ロスレス符号化
３．ロスレス復号化

【００７６】

以下に示す実施形態の主な利点を理解するために、まず、ニューラルネットワークの話題と、パラメータ符号化のための関連する方法について、簡単に紹介する。それにもかかわらず、開示されたすべての側面、特徴、および概念は、本明細書に記載される実施形態と別々にまたは組み合わせて使用することができる。

【００７７】

２　量子化及びエントロピー符号化のための関連する方法

マルチメディアコンテンツの説明及び分析のためのニューラルネットワークの圧縮のためのＭＰＥＧ－７パート１７規格のワーキングドラフト２［２］は、ニューラルネットワークパラメータ符号化に独立スカラー量子化（ｉｎｄｅｐｅｎｄｅｎｔ　ｓｃａｌａｒ　ｑｕａｎｔｉｚａｔｉｏｎ）及びエントロピー符号化を適用している。

【００７８】

２．１　スカラー量子化器

ニューラルネットワークのパラメータは、スカラ量子化器を用いて量子化される。量子化の結果、パラメータ１３の許容値のセットは減少する。言い換えれば、ニューラルネットワークのパラメータは、いわゆる再構成レベルの可算集合（実際には有限集合）にマッ

ピングされる。再構成レベルセットは、可能なニューラルネットワークパラメータ値のセット（集合）の適切なサブセット（部分集合）を表す。以下のエントロピー符号化を単純化するために、許容可能な再構成レベルは、量子化インデックス５６によって表され、これはビットストリーム１４の一部として伝送される。復号化器側では、量子化インデックス５６は、再構成されたニューラルネットワークパラメータ１３にマッピングされる。再構成されたニューラルネットワークパラメータ１３の可能な値は、再構成レベルセット５２に対応する。符号化器側では、スカラー量子化の結果は、１セットの（整数）量子化インデックス５６である。

【００７９】

このアプリケーションでは、均一再構成量子化器（ＵＲＱ）が使用される。その基本設計を図７に示す。図７は、均一再構成量子化器の説明図である。ＵＲＱは、再構成レベルが等間隔に配置されるという特性を持つ。隣接する２つの再構成レベル間の距離Δ（ＱＰ）を量子化ステップサイズと呼ぶ。再構成レベルの１つは０に等しい。したがって、利用可能な再構成レベルの完全なセット、例えば$s'_i, i \in \mathbb{N}_0$,は、量子化ステップサイズΔ（ＱＰ）によって一意に特定される。量子化インデックスｑ５６の再構成された重みパラメータｔ’１３’への復号化器のマッピングは、原理的に、単純な式で与えられる。

$$t' = q \cdot \Delta.$$

【００８０】

このコンテキストでは、「独立スカラー量子化」という用語は、任意の重みパラメータ１３に対する量子化インデックスｑ５６が与えられると、関連する再構成された重みパラメータｔ’１３’が他の重みパラメータに対するすべての量子化インデックスから独立して決定できる、という特性を指す。

【００８１】

２．１．１　符号化器の動作：量子化

ニューラルネットワークの圧縮に関する標準規格は、ビットストリームのシンタックスと再構成プロセスのみを規定している。与えられたオリジナルのニューラルネットワークパラメータ１３のセットと与えられた量子化ステップサイズ（ＱＰ）に対するパラメータ符号化を考える場合、符号化器は多くの自由度を有する。層１０ａ、１０ｂの量子化インデックスq_k５６が与えられると、エントロピー符号化は、データをビットストリーム１４に書き込む（すなわち、算術符号語（コードワード）を構築する）ための一意に定義されたアルゴリズムに従わなければならない。しかし、重みパラメータのオリジナルセット（例えば層）が与えられた量子化インデックスq_k５６を得るための符号化アルゴリズムは、ニューラルネットワーク圧縮の規格の範囲外である。以下の説明では、各ニューラルネットワークパラメータ１３の量子化ステップサイズ（ＱＰ）が既知であると仮定する。それでも、符号化器は、各ニューラルネットワーク（重み）パラメータt_k１３についての量子化器インデックスq_k５６を選択する自由を有する。量子化インデックスの選択は、歪み（または再構成／近似品質）とビットレートの両方を決定するので、使用される量子化アルゴリズムは、生成されるビットストリーム１４のレート歪み性能に実質的な影響を与える。

【００８２】

最も単純な量子化方法は、ニューラルネットワークパラメータｔ$_k$１３を最も近い再構成レベルに丸める（最近傍量子化（ｎｅａｒｅｓｔ　ｎｅｉｇｈｂｏｒ　ｑｕａｎｔｉｚａｔｉｏｎ）とも呼ばれる）。一般的に使用されるＵＲＱについて、対応する量子化インデックスｑ$_k$５６は、以下に従って決定することができる。

$$q_k = \mathrm{sgn}(t_k) \cdot \left\lfloor \frac{|t_k|}{\Delta_k} + \frac{1}{2} \right\rfloor,$$

ここで、ｓｇｎ（）は符号関数、演算子⌊·⌋はその引数と等しいか小さい最大の整数を返す。この量子化法によって、ＭＳＥ歪みは最小になることが保証される。

$$D = \sum_k D_k = \sum_k (t_k - q_k \cdot \Delta_k)^2$$

しかし、結果として得られるパラメータレベル（重みレベル）ｑ$_k$５６を送信するために必要とされるビットレートは完全に無視される。なお、この方法は、ＭＳＥ歪み測定に限定されるものではなく、他の任意の歪み測定、例えば、以下の方法によるＭＡＥ歪みを使用することができる。

$$D^{MAE} = \sum_k D_k^{MAE} = \sum_k |t_k - q_k \cdot \Delta_k|$$

一般的に、丸め方はゼロに偏った方が良い結果が得られる。

$$q_k = \mathrm{sgn}(t_k) \cdot \left\lfloor \frac{|t_k|}{\Delta_k} + a \right\rfloor \quad \text{with} \quad 0 \le a < \frac{1}{2}.$$

【００８３】

量子化プロセスはラグランジュ関数Ｄ＋λ・Ｒを最小化すれば、レート歪み的に良い結果が得られる。ここで、Ｄはニューラルネットワークパラメータセットの歪み（例えば、ＭＳＥ歪み、又は、ＭＡＥ歪み）、Ｒは量子化インデックス５６を伝送するために必要なビット数、λはラグランジュ乗数である。

【００８４】

量子化ステップサイズが与えられると、ラグランジュ乗数λと量子化ステップサイズとの間に以下の関係がしばしば用いられる。

$$\lambda = c_1 \cdot \Delta^2,$$

ここで、c_1はニューラルネットワークパラメータのセットに対する定数ファクタを表す。歪みとレートのラグランジュ関数D＋λ・Rを最小化することを目的とした量子化アルゴリズムは、レート歪み最適化量子化（RDOQ、rate－distortion optimized quantization）とも呼ばれる。歪みをMSEまたは重み付きMSE（またはそれぞれMAE）を用いて測定する場合、重みパラメータのセット（例えば層）に対する量子化指標q_k56は、以下のコスト量が最小となるように決定する必要がある。

$$D + \lambda \cdot R = \sum_k \alpha_k \cdot (t_k - \Delta_k \cdot q_k)^2 + \lambda \cdot R(q_k | q_{k-1}, q_{k-2}, \cdots).$$

【００８５】

このとき、ニューラルネットワークパラメータインデックスkは、ニューラルネットワークパラメータ１３の符号化順序（または走査順序）を指定する。$R(q_k|q_{k-1},q_{k-2},\cdots)$の項は、量子化インデックス$q_k$56を送信するために必要なビット数（またはその推定値）を表している。この条件は、（結合確率または条件付き確率の使用により）特定の量子化インデックスq_kについてのビット数が、典型的には、符号化順序、例えば共通の連続的な順番１４’において先行する量子化インデックスq_{k-1},q_{k-2},等のための選択値に依存していることを示している。上式における係数α_kは、個々のニューラルネットワークパラメータ１３の寄与度を重み付けするために用いることができる。以下では、一般に、すべての重み付け係数α_kが１に等しいと仮定する（ただし、異なる重み付け係数を考慮することができるように、アルゴリズムを端的に修正することができる）。

【００８６】

実際、近傍量子化はλ＝０の些細なケースであり、マルチメディアコンテンツの説明及び分析のためのニューラルネットワークの圧縮に関するMPEG－７ part１７規格のワーキングドラフト２において適用されている。

【００８７】

２．２ エントロピー符号化

前のステップで適用された均一量子化の結果として、重みパラメータはいわゆる再構成レベルの有限集合にマッピングされる。これらは、（整数）量子化器インデックス５６（パラメータレベルまたは重みレベルとも呼ばれる）と量子化ステップサイズ（ＱＰ）によって表すことができ、例えば、全層に対して固定されている場合がある。層のすべての量子化された重みパラメータを復元するために、層のステップサイズ（ＱＰ）および次元は、復号化器によって知られてもよい。これらは、例えば、別々に送信されてもよい。

【００８８】

２．２．１ コンテキスト適応的２値算術符号化（CABAC、Context－ada

ｐｔｉｖｅ ｂｉｎａｒｙ ａｒｉｔｈｍｅｔｉｃ ｃｏｄｉｎｇ）による量子化インデックスの符号化

量子化インデックス５６（整数表現）は、次にエントロピー符号化技術を使用して送信される。したがって、重みの層は、スキャンを使用して量子化された重みレベルのシーケンスにマッピングされる。例えば、行列の最上部の行から始めて、含まれる値を左から右へ符号化する、行ファーストスキャン順序（ｒｏｗ ｆｉｒｓｔ ｓｃａｎ ｏｒｄｅｒ）を使用することができる。この方法では、すべての行が上から下へ符号化される。スキャンは、ニューロン相互接続１１の重みに関連し得るニューラルネットワークパラメータ１３を含む行列１５ａについて、例えば共通の連続的な順番１４’に沿って、図３に示すように実行されてもよい。行列は、重み層、例えば、図３及び図１にそれぞれ示すように、ニューロン相互接続１１の層ｐ－１ １０ａと層ｐ １０ｂとの間の重み、又は隠れ層及び入力層との間の重みを表してもよい。なお、他の任意のスキャンを適用することができる。例えば、行列（例えば、図２又は図３の行列１５ａ）は、行ファーストスキャンを適用する前に、転置され、又は水平及び／又は垂直に反転され、及び／又は左又は右に９０／１８０／２７０度だけ回転されることができる。

【００８９】

図３および図５に関して説明したように、実施形態による装置は、量子化インデックス５６を２値化８２したものの少なくとも１つのビン８４に対する現在のニューラルネットワークパラメータ１３’に対する状態１２２に対応する確率モデルを用いて、２値算術符号化を用いてデータストリーム１４へ現在のニューラルネットワークパラメータ１３’についての量子化インデックス５６を符号化するよう構成されることができる。確率モデルを用いた２値算術符号化は、コンテキスト適応的２値算術符号化（ＣＡＢＡＣ、Ｃｏｎｔｅｘｔ－ａｄａｐｔｉｖｅ ｂｉｎａｒｙ ａｒｉｔｈｍｅｔｉｃ ｃｏｄｉｎｇ）であることができる。

【００９０】

すなわち、実施形態によれば、レベルの符号化のために、ＣＡＢＡＣが使用される。詳細については、［３］を参照されたい。そこで、量子化された重みレベルｑ５６は、一連の２値記号またはシンタックス要素、例えばビン（２値決定）に分解され、その後、２値算術符号化器（ＣＡＢＡＣ）に渡されることがある。最初のステップでは、量子化された重みレベルに対して２値シンタックス要素ｓｉｇ＿ｆｌａｇが導出され、これは対応するレベルがゼロに等しいかどうかを指定する。言い換えれば、図４に示す量子化インデックス５６を２値化８２したものの少なくとも１つのビンは、現在のニューラルネットワークパラメータの量子化インデックス５６がゼロに等しいか否かを示す有意性ビンを含むことができる。

【００９１】

ｓｉｇ＿ｆｌａｇが１に等しい場合、さらなる２値シンタックス要素ｓｉｇｎ＿ｆｌａｇが導き出される。このビンは、現在の重みレベルが正（例えば、ビン＝０）か負（例えば、ビン＝１）かを示す。言い換えれば、図４に示す量子化インデックス５６を２値化８２したものの少なくとも１つのビンは、現在のニューラルネットワークパラメータの量子化インデックス５６がゼロより大きいかまたはゼロより小さいかを示す符号ビン８６を含むことができる。

【００９２】

次に、ビンの単項シーケンスが符号化され、それに続いて、以下のように固定長シーケンスが次のように符号化される。

【００９３】

変数ｋは負でない整数で初期化され、Ｘは１＜＜ｋで初期化される。

【００９４】

量子化された重みレベルの絶対値がＸより大きいことを示すａｂｓ＿ｌｅｖｅｌ＿ｇｒｅａｔｅｒ＿Ｘという一つ以上のシンタックス要素が符号化される。ａｂｓ＿ｌｅｖｅｌ＿ｇｒｅａｔｅｒ＿Ｘが１に等しい場合、変数ｋが更新され（例えば、１だけ増加）、次

に１＜＜ｋがＸに加えられ、さらにａｂｓ＿ｌｅｖｅｌ＿ｇｒｅａｔｅｒ＿Ｘが符号化される。この手順は、ａｂｓ＿ｌｅｖｅｌ＿ｇｒｅａｔｅｒ＿Ｘが０に等しくなるまで続けられる。その後、長さｋの固定長コードで量子化器インデックスの符号化を完了することができる。例えば、変数ｒｅｍ＝Ｘ－｜ｑ｜は、ｋビットを用いて符号化され得る。あるいは、変数ｒｅｍ’は、ｒｅｍ’＝（１＜＜ｋ）－ｒｅｍ－１として定義され、これはｋビットを用いて符号化され得る。また、変数ｒｅｍのｋビットの固定長符号への他のマッピングも使用できる。

【００９５】

言い換えれば、図４に示す量子化インデックス５６を２値化８２したものの少なくとも１つのビンは、現在のニューラルネットワークパラメータの量子化インデックス５６の絶対値がＸより大きいか否かを示すｇｒｅａｔｅｒ－ｔｈａｎ－Ｘビンを含むことができ、ここでＸはゼロより大きい整数である。

【００９６】

各ａｂｓ＿ｌｅｖｅｌ＿ｇｒｅａｔｅｒ＿Ｘの後にｋを１ずつ増加させる場合、このアプローチは、指数ゴロム符号化（ｅｘｐｏｎｅｎｔｉａｌ Ｇｏｌｏｍｂ ｃｏｄｉｎｇ）を適用することと同一である（ｓｉｇｎ＿ｆｌａｇがみなされていない場合）。

【００９７】

また、符号化器側と復号化器側で絶対値の最大値ａｂｓ＿ｍａｘが分かっている場合、次に送信するａｂｓ＿Ｌｅｖｅｌ＿ｇｒｅａｔｅｒ＿Ｘについて、Ｘ＞＝ａｂｓ＿ｍａｘが成り立つとき、ａｂｓ＿Ｌｅｖｅｌ＿ｇｒｅａｔｅｒ＿Ｘシンタックス要素の符号化を終了することができる。

【００９８】

２．２．２　コンテキスト適応的２値算術符号化（ＣＡＢＡＣ）による量子化インデックスの復号化

量子化された重みレベル５６（整数表現）の復号化は、符号化と同様に動作する。復号化器は、まずｓｉｇ＿ｆｌａｇを復号化する。もしそれが１に等しければ、ｓｉｇｎ＿ｆｌａｇとａｂｓ＿ｌｅｖｅｌ＿ｇｒｅａｔｅｒ＿Ｘの単項シーケンスが続く。ここで、ｋの更新（したがってＸの増分）は符号化器と同じルールに従わなければならない。最後に、ｋビットの固定長符号が復号化され、整数値として解釈される（例えば、ｒｅｍまたはｒｅｍ’として。両者のどちらが符号化されたかに応じている）。そして、復号化された量子化された重みレベルの絶対値$|q|$をＸから再構成し、固定長部分を形成することができる。例えば、固定長部分としてｒｅｍが使用された場合、$|q| = X - rem$となる。あるいは、ｒｅｍ’が符号化されていた場合、$|q| = X + 1 + rem' - (1 \ll k)$。最後のステップとして、復号化されたｓｉｇｎ＿ｆｌａｇに対応して$|q|$に符号を適用する必要があり、量子化された重みレベルｑ５６を得ることができる。最後に、量子化された重みレベルｑにステップサイズΔ（ＱＰ）を乗じることにより、量子化された重みｗが再構成される。

【００９９】

言い換えれば、図２および図４に関して説明したように、実施形態による装置は、量子化インデックス５６を２値化８２したものの少なくとも１つのビン８４についての現在のニューラルネットワークパラメータ１３’に対する状態１２２に対応する確率モデルを使用することにより、２値算術符号化を使用してデータストリーム１４から現在のニューラルネットワークパラメータ１３’に対する量子化インデックス５６を復号化するよう構成されることができる。

【０１００】

図５に示す量子化インデックス５６を２値化８２したものの少なくとも１つのビンは、現在のニューラルネットワークパラメータの量子化インデックス５６がゼロに等しいか否かを示す有意性ビンを含むことができる。さらに、または代替的に、少なくとも１つのビ

ンは、現在のニューラルネットワークパラメータの量子化インデックス５６がゼロより大きいか、またはゼロより小さいかを示す符号ビン８６を含むことができる。さらに、少なくとも１つのビンは、現在のニューラルネットワークパラメータの量子化インデックス５６の絶対値がＸより大きいか否かを示すｇｒｅａｔｅｒ－ｔｈａｎ－Ｘビンを含むことができ、ここでＸはゼロより大きい整数である。

【０１０１】

好ましい実施形態では、ｋは０に初期化され、以下のように更新される。各ａｂｓ＿ｌｅｖｅｌ＿ｇｒｅａｔｅｒ＿Ｘが１に等しくなった後、ｋの必要な更新は、以下のルールに従って行われる：Ｘ＞Ｘ’の場合、ｋは１だけ増分され、Ｘ’はアプリケーションに対応する定数である。例えば、Ｘ’は符号化器が導出し、復号化器に通知する数値（例えば、０から１００の間）である。

【０１０２】

２．２．３　コンテキスト・モデリング

ＣＡＢＡＣエントロピー符号化では、量子化された重みレベル５６のほとんどのシンタックス要素は、２値確率モデリングを用いて符号化される。各２値決定（ｂｉｎ）はコンテキストと関連付けられている。コンテキストは、符号化されたビンのクラスに対する確率モデルを表す。２つの可能なビン値のうちの１つに対する確率は、対応するコンテキストで既に符号化されたビンの値に基づいて、各コンテキストについて推定される。アプリケーションに応じて、異なるコンテキストモデリングアプローチを適用することができる。通常、量子化された重み符号化に関連するいくつかのビンについて、符号化に使用されるコンテキストは、既に送信されたシンタックス要素に基づいて選択される。実際のアプリケーションに応じて、例えばＳＢＭＰ　０、またはＨＥＶＣ　０またはＶＴＭ－４．０のものなど、異なる確率推定器が選択され得る。この選択は、例えば、圧縮効率や複雑さに影響を与える。

【０１０３】

言い換えれば、図５、例えばコンテキスト８７に関して説明したような確率モデルは、以前に符号化されたニューラルネットワークパラメータの量子化インデックスに追加的に依存する。

【０１０４】

それぞれ、図４、例えばコンテキスト８７に関して説明したような確率モデルは、さらに、以前に復号化されたニューラルネットワークパラメータの量子化インデックスに依存する。

【０１０５】

広範囲のニューラルネットワークに適合するコンテキストモデリング方式は、以下のように説明される。重み行列（層）の特定の位置（ｘ，ｙ）で量子化された重みレベルｑ５６を復号化するために、ローカルテンプレートが現在の位置に適用される。このテンプレートは、例えば（ｘ－１，ｙ）、（ｘ，ｙ－１）、（ｘ－１，ｙ－１）等のような多数の他の（順序付けられた）位置を含んでいる。各位置に対して、ステータス識別子が導出される。

【０１０６】

好ましい実施形態（Ｓｉ１と表す）において、位置（ｘ，ｙ）に対するステータス識別子$s_{x,y}$は、以下のように導出される。位置（ｘ，ｙ）が行列の外側を指す場合、または位置（ｘ，ｙ）における量子化された重みレベル$q_{x,y}$がまだ復号化されていないかゼロに等しい場合、ステータス識別子$s_{x,y}=0$とする。それ以外の場合、ステータス識別子$s_{x,y}=q_{x,y}<0\,?\,1:2$であるものとする。

【０１０７】

特定のテンプレートに対して、ステータス識別子のシーケンスを導き出し、ステータス

識別子の値の可能な各コンステレーション（ｃｏｎｓｔｅｌｌａｔｉｏｎ）を、使用されるコンテキストを識別するコンテキストインデックスにマッピングする。テンプレートとマッピングは、異なるシンタックス要素に対して異なる場合がある。例えば、（順序）位置（ｘ－１，ｙ）、（ｘ，ｙ－１）、（ｘ－１，ｙ－１）を含むテンプレートから、ステータス識別子$s_{x-1,y}$、$s_{x,y-1}$、$s_{x-1,y-1}$の順序シーケンスが導出される。例えば、このシーケンスは、コンテキストインデックスＣ＝$s_{x-1,y}$＋３＊$s_{x,y-1}$＋９＊$s_{x-1,y-1}$にマッピングされることがある。例えば、コンテキストインデックスＣは、ｓｉｇ＿ｆｌａｇのための多数のコンテキストを識別するために使用されることができる。

【０１０８】

好ましい実施形態（アプローチ１とする）において、位置（ｘ，ｙ）における量子化された重みレベル$q_{x,y}$のｓｉｇ＿ｆｌａｇのため、またはｓｉｇｎ＿ｆｌａｇのためのローカルテンプレートは、１つの位置（ｘ－１，ｙ）（すなわち、左隣）からのみ構成される。関連するステータス識別子$s_{x-1,y}$は、好ましい実施形態Ｓｉ１に従って導出される。

【０１０９】

ｓｉｇ＿ｆｌａｇについては、$s_{x-1,y}$の値に応じて３つのコンテキストのうちの１つが選択され、又はｓｉｇｎ＿ｆｌａｇについては、$s_{x-1,y}$の値に応じて他の３つのコンテキストのうちの１つが選択される。

【０１１０】

別の好ましい実施形態（アプローチ２とする）では、ｓｉｇ　ｆｌａｇのためのローカルテンプレートは、３つの順序付けられた位置（ｘ－１，ｙ）、（ｘ－２，ｙ）、（ｘ－３，ｙ）を含む。ステータス識別子$s_{x-1,y}$，$s_{x-2,y}$，$s_{x-3,y}$の関連するシーケンスは、好ましい実施形態Ｓｉ２に従って導出される。

【０１１１】

ｓｉｇ＿ｆｌａｇについては、コンテキストインデックスＣを以下のように導出する。

【０１１２】

$s_{x-1,y}$≠０ならばＣ＝０，それ以外なら$s_{x-2,y}$≠０ならばＣ＝１，それ以外なら$s_{x-3,y}$≠０ならばＣ＝２、そうでなければ、$s_{x-3,y}$≠０ならば、Ｃ＝２である。そうでなければ、Ｃ＝３。

【０１１３】

これは、次の式で表すこともできる。

$$C = \left(s_{x-1,y} \neq 0\right) ? 0 : \left(\left(s_{x-2,y} \neq 0\right) ? 1 : \left(\left(s_{x-3,y} \neq 0\right) ? 2 : 3\right)\right)$$

【０１１４】

同様に、コンテキストインデックスＣが左側の次のゼロでない重みまでの距離に等しくなる（テンプレートサイズを超えない）ように、左側に隣接するものの数を増減してもよい。

【０１１５】

各ａｂｓ＿ｌｅｖｅｌ＿ｇｒｅａｔｅｒ＿Ｘ　ｆｌａｇは、例えば、２つのコンテキストの独自のセットを適用することができる。そして、２つのコンテキストのうちの１つが、ｓｉｇｎ＿ｆｌａｇの値に応じて選択される。

【０１１６】

好ましい実施形態では、Ｘが予め定義された数Ｘ’より小さいａｂｓ＿ｌｅｖｅｌ＿ｇｒｅａｔｅｒ＿Ｘ　ｆｌａｇについて、異なるコンテキストが、Ｘおよび／またはｓｉｇｎ＿ｆｌａｇの値に応じて区別される。

【０１１７】

好ましい実施形態では、Ｘが予め定義された数Ｘ’より大きいか等しいａｂｓ＿ｌｅｖｅｌ＿ｇｒｅａｔｅｒ＿Ｘ　ｆｌａｇについて、異なるコンテキストは、Ｘに対応してのみ区別される。

【０１１８】
別の好ましい実施形態では、予め定義された数Ｘ’より大きいか等しいＸを有するａｂｓ＿ｌｅｖｅｌ＿ｇｒｅａｔｅｒ＿Ｘ　ｆｌａｇは、１の固定コード長を用いて（例えば、算術符号器のバイパスモードを用いて）符号化される。
【０１１９】
さらに、シンタックス要素の一部または全部は、コンテキストを使用せずに符号化されることもある。その代わり、それらは、例えば、ＣＡＢＡＣのいわゆるバイパスビンを使用して、１ビットの固定長で符号化される。
【０１２０】
別の好ましい実施形態では、固定長の余りｒｅｍは、バイパスモードを使用して符号化される。
【０１２１】
別の好ましい実施形態では、符号化器は、予め定義された数Ｘ’を決定し、Ｘ＜Ｘ’の各シンタックス要素ａｂｓ＿ｌｅｖｅｌ＿ｇｒｅａｔｅｒ＿Ｘに対して、符号に応じて２つのコンテキストを区別し、Ｘ＞＝Ｘ’の各ａｂｓ＿ｌｅｖｅｌ＿ｇｒｅａｔｅｒ＿Ｘに対して、１つのコンテキストを使用する。
【０１２２】
言い換えれば、図５に関して説明したような確率モデル、例えばコンテキスト８７は、現在のニューラルネットワークパラメータが関連する部分に隣接するニューラルネットワークの部分に関連する以前に符号化されたニューラルネットワークパラメータの量子化インデックスに応じて確率モデルのサブセットの中から、現在のニューラルネットワークパラメータのために選択１０３されることができる。
【０１２３】
この部分は、例えば、上記で説明したテンプレートであって、（順序付けられた）位置（ｘ－１，ｙ）、（ｘ，ｙ－１）、（ｘ－１，ｙ－１）を含むテンプレートによって定義されることができる。
【０１２４】
それぞれ、図５に関して説明したような確率モデルは、現在のニューラルネットワークパラメータが関連する部分に隣接するニューラルネットワークの部分に関連する以前に復号化されたニューラルネットワークパラメータの量子化インデックスに応じて、確率モデルのサブセットの中から、現在のニューラルネットワークパラメータについて選択されることができる。
【０１２５】
３　追加の方法
以下では、再構成された層、例えば図６からのニューラルネットワーク層ｐが、例えば別々に伝送されてもよい図６からの再構成層ｉ－１および再構成層ｉのような異なるサブ層の構成である、ニューラルネットワーク１０の圧縮／送信のための追加の、したがって任意の方法について説明する。
【０１２６】
３．１　ベース層とエンハンスメント層の概念
この概念では、ベース層とエンハンスメント層と呼ばれる２種類のサブ層を導入している。そして、再構成プロセス（例えば、すべてのサブ層を追加する）は、サブ層からどのように再構成された層を得ることができるかを定義する。ベース層はベース値を含み、例えば、最初のステップで効率的に表現または圧縮／送信できるように選択することができる。エンハンスメント層は、エンハンスメント情報、（例えばオリジナル層に関する）歪み指標を減少させるために例えば（ベース）層の値に追加される差分値を含む。別の例では、ベース層は（小さなトレーニングセットを用いたトレーニングからの）粗い値を含み、エンハンスメント層は（完全なトレーニングセットまたはより一般的には、別のトレーニングセットに基づく）リファインメント値を含む。サブ層は別々に保存／送信されてもよい。

【０１２７】
好ましい実施形態では、圧縮される層L_R、例えばニューラルネットワークパラメータ、例えば図２及び図３の行列１５ａによって表され得る重みのようなニューラルネットワークの重みは、ベース層L_Bと１つ以上のエンハンスメント層$L_{E,1}$，$L_{E,2}$，・・・，$L_{E,N}$とに分解される。そして、最初のステップでベース層が圧縮／送信され、続くステップでエンハンスメント層$L_{E,1}$，$L_{E,2}$，・・・、$L_{E,N}$が（別個に）圧縮／送信される。
【０１２８】
別の好ましい実施形態では、再構成された層L_Rは、以下のように、すべてのサブ層$L_{S,N}$を（要素単位で）加算することによって得ることができる。

$$L_R = \sum_{i=0}^{N} L_{S,N}$$

【０１２９】
さらなる好ましい実施形態において、再構成された層L_Rは、以下に従って、すべてのサブ層$L_{S,N}$を（要素単位で）乗算することによって得ることができる。

$$L_R = \prod_{i=0}^{N} L_{S,N}$$

【０１３０】
言い換えれば、本発明による実施形態は、再構成された層L_Rの形態で、または例えば再構成された層L_Rを使用して、ニューラルネットワークパラメータごとに、第１の再構成層のニューラルネットワークパラメータ値と第２の再構成層のニューラルネットワークパラメータ値のパラメータ単位の和またはパラメータ単位の積によって、ニューラルネットワークパラメータ１３を再構成するように構成された、装置を含む。
【０１３１】
それぞれ、実施形態によるニューラルネットワークパラメータ１３を符号化するための装置の場合、ニューラルネットワークパラメータ１３は、ニューラルネットワークパラメータごとに、第１の再構成層のニューラルネットワークパラメータ値と第２の再構成層のニューラルネットワークパラメータ値のパラメータ単位の和またはパラメータ単位の積によって再構成可能である。
【０１３２】
さらなる好ましい実施形態では、２．１および／または２．２の方法は、サブセットまたはすべてのサブ層に適用される。
【０１３３】
特に好ましい実施形態では、コンテキストモデリングを使用するエントロピー符号化方式（例えば、２．２．３に同質または類似）が適用されるが、以下のルールのうちの１つ以上に従ってコンテキストモデルの１つ以上のセットを追加する。
【０１３４】
ａ）各サブ層は、独自のコンテキストセットを適用する。言い換えれば、本発明による実施形態は、第１の再構成層についての第１のニューラルネットワークパラメータ１３ａをデータストリームまたは別のデータストリームから／に符号化／復号化し、第２の再構成層についての第２ニューラルネットワークパラメータ１３ｂを第１および第２の再構成層のための別の確率コンテキストを用いたコンテキスト適応的エントロピー符号化によってデータストリームから／に符号化／復号化するように構成された、装置を含む。
【０１３５】
ｂ）　符号化されるエンハンスメント層のパラメータのために選択されたコンテキスト

セットは，符号化順序において先行する層（例えば、ベース層）の同位置のパラメータの値に依存する。コンテキストモデルの第１のセットは、同位置のパラメータがゼロに等しいときは常に選択され、そうでないときは第２のセットが選択される。言い換えれば、本発明による実施形態は、第２の再構成層のニューラルネットワークパラメータ値、例えばエンハンスメント層のパラメータを、第１の再構成層のニューラルネットワークパラメータ値、例えば符号化順序において先行する層（例えばベース層）における同位置のパラメータの値に対応する確率モデルを用いたコンテキスト適応的エントロピー符号化によってデータストリームに符号化するよう構成された、装置を備える。さらなる実施形態は、第１の再構成層のニューラルネットワークパラメータ値に応じて確率コンテキストセットの集合から確率コンテキストセットを選択することによって、及び、第１の再構成層のニューラルネットワークパラメータ値に応じて選択された確率コンテキストセットの中から使用する確率コンテキストを選択することによって、コンテキスト適応的エントロピー符号化によって第２の再構成層のニューラルネットワークパラメータ値をデータストリームに符号化するように構成される装置を含む。それぞれに、実施形態によるニューラルネットワークパラメータ１３を復号化するための装置について、前記装置は、第１の再構成層のニューラルネットワークパラメータ値に対応する確率モデルを用いたコンテキスト適応的エントロピー復号化によってデータストリームから第２の再構成層のニューラルネットワークパラメータ値を復号化するよう構成されることができる。それぞれ、さらなる実施形態は、第１の再構成層のニューラルネットワークパラメータ値に対応する確率コンテキストセットの集合から確率コンテキストセットを選択することによって、及び、第１の再構成層のニューラルネットワークパラメータ値に対応する選択された確率コンテキストセットの中から使用する確率コンテキストを選択することによって、コンテキスト適応的エントロピー復号化によってデータストリームから第２の再構成層のニューラルネットワークパラメータ値を復号化するように構成された、装置を含む。

【０１３６】

ｃ） 符号化されるエンハンスメント層のパラメータについて選択されたコンテキストセットは、符号化順序において先行する層（例えば、ベース層）の同位置のパラメータの値に対応する。コンテキストモデルの第１のセットは、同位置のパラメータがゼロより小さい（負）場合は常に選択され、第２のセットは、同位置のパラメータがゼロより大きい（正）場合は選択され、それ以外は第３のセットが選択される。言い換えれば、本発明による実施形態は、例えば、符号化するための装置を含む。確率コンテキストセットの集合は３つの確率コンテキストセットを含む。装置は、第１の再構成層のニューラルネットワークパラメータ値が負の場合、確率コンテキストセットの集合から第１の確率コンテキストセットを選択された確率コンテキストセットとして選択するように構成され、第１の再構成層のニューラルネットワークパラメータ値が正である場合、確率コンテキストセットの集合から第２の確率コンテキストセットを選択された確率コンテキストセットとして選択するように構成され、第１の再構成層のニューラルネットワークパラメータ値がゼロである場合、確率コンテキストセットの集合から第３の確率コンテキストセットを選択された確率コンテキストセットとして選択するよう構成される。それぞれ、実施形態によるニューラルネットワークパラメータ１３を復号化するための装置については、確率コンテキストセットの集合は、３つの確率コンテキストセットを含むことができ、装置は、第１の再構成層のニューラルネットワークパラメータ値が負の場合、確率コンテキストセットの集合から第１の確率コンテキストセットを選択した確率コンテキストセットとして選択するように構成されることができ、第１の再構成層のニューラルネットワークパラメータ値が正の場合、確率コンテキストセットの集合から第２の確率コンテキストセットを選択した確率コンテキストセットとして選択するように構成されることができ、第１の再構成層のニューラルネットワークパラメータ値がゼロである場合、確率コンテキストセットの集合から第３の確率コンテキストセットを選択した確率コンテキストセットとして選択するよう構成されることができる。

【０１３７】

ｄ）　符号化されるエンハンスメント層のパラメータのために選択されるコンテキストセットは、符号化順序において先行する層（例えば，ベース層）の同位置のパラメータの値に対応する。コンテキストモデルの第１のセットは、同位置にあるパラメータの（絶対）値がＸ（Ｘはパラメータである）より大きいときは常に選択され、そうでないときは第２のセットが選択される。言い換えれば、本発明による実施形態は装置を含み、確率コンテキストセットの集合は、２つの確率コンテキストセットを含み、装置は、第１の再構成層のニューラルネットワークのパラメータ値、例えば、符号化順序において先行する層（例えば、ベース層）における同位置にあるパラメータの値が所定の値、例えばＸ、より大きい場合、確率コンテキストセットの集合から第１の確率コンテキストセットを選択された確率コンテキストセットとして選択するように構成され、第１の再構成層のニューラルネットワークパラメータ値が所定値より大きくない場合、確率コンテキストセットの集合から第２の確率コンテキストセットを選択された確率コンテキストセットとして選択するように構成され、あるいは、第１の再構成層のニューラルネットワークパラメータ値の絶対値が所定値より大きい場合、確率コンテキストセットの集合から第１の確率コンテキストセットを選択された確率コンテキストセットとして選択するように構成され、第１の再構成層のニューラルネットワークパラメータ値の絶対値が所定値より大きくない場合、確率コンテキストセットの集合から第２確率コンテキストセットを選択された確率コンテキストセットとして選択するように構成される。それぞれに、実施形態によるニューラルネットワークパラメータ１３を復号化するための装置について、確率コンテキストの集合は２つの確率コンテキストセットを含むことができ、装置は、第１の再構成層のニューラルネットワークパラメータ値が所定の値、例えばＸ、より大きい場合、確率コンテキストセットの集合から第１の確率コンテキストセットを選択された確率コンテキストセットとして選択するように構成されることができ、第１の再構成層のニューラルネットワークパラメータ値が所定の値より大きくない場合、確率コンテキストセットの集合から第２の確率コンテキストセットを選択された確率コンテキストセットとして選択するように構成されることができ、又は、第１の再構成層のニューラルネットワークパラメータ値の絶対値が所定値より大きい場合、確率コンテキストセットの集合から第１の確率コンテキストセットを選択された確率コンテキストセットとして選択するように構成され、第１の再構成層のニューラルネットワークパラメータ値の絶対値が所定値より大きくない場合、確率コンテキストセットの集合から第２の確率コンテキストセットを選択された確率コンテキストセットとして選択するように構成されることができる。

【０１３８】

４　依存スカラー量子化を伴うニューラルネットワークパラメータ符号化

このセクションでは、図２～図４の文脈で説明したような、本発明による概念および実施形態に対するさらなる任意の側面および特徴が開示される。

【０１３９】

以下では、ニューラルネットワークパラメータ符号化の修正された概念について説明する。先に説明したニューラルネットワークパラメータ符号化に対する主な変更点は、ニューラルネットワークパラメータ１３が独立して量子化および再構成されないということである。その代わりに、ニューラルネットワークパラメータ１３の許容される再構成レベルは、再構成順序において先行するニューラルネットワークパラメータの選択された量子化インデックス５６に対応する。依存スカラー量子化の概念は、ニューラルネットワークパラメータについての確率モデル選択（または、代替的に符号語（コードワード）表選択）が許容される再構成レベルセットに依存する、修正されたエントロピー符号化と組み合わされる。しかしながら、先に説明した実施形態は、以下に説明する特徴のいずれかを別個にまたは組み合わせて使用および／または組み込みおよび／または拡張することができることに留意されたい。

【０１４０】

４．１　関連するニューラルネットワークのパラメータ符号化との比較による優位性

ニューラルネットワークパラメータの依存量子化の利点は、許容される再構成ベクトル

がＮ次元信号空間（ここで、Ｎは、処理されるサンプルのセット、例えば層１０ａ、１０ｂにおけるサンプルまたはニューラルネットワークパラメータ１３の数を表す）において密に詰め込まれることである。ニューラルネットワークパラメータセットの再構成ベクトルは、ニューラルネットワークパラメータセットの順序付けられた再構成されたニューラルネットワークパラメータ（または、代替的に、順序付けられた再構成されたサンプル）を指す。依存スカラー量子化の効果を、２つのニューラルネットワークパラメータの最も単純なケースについて図８で説明する。図８は、２つの重みパラメータの単純な場合について、許容される再構成ベクトルの位置の一例を示す図である。図８（ａ）は独立スカラ量子化の例、図８（ｂ）は依存スカラ量子化の例である。図８ａは、独立スカラ量子化の場合の許容再構成ベクトル２０１（２次元平面上の点を表す）を示している。見て分かるように、第２のニューラルネットワークパラメータt_1’１３に対する許容値のセットは、第１の再構成されたニューラルネットワークパラメータt_0’１３に対する選択された値には依存しない。図８（ｂ）は、依存スカラー量子化の例を示す。独立スカラー量子化とは対照的に、第２のニューラルネットワークパラメータt_1’１３に対して選択可能な再構成値は、第１のニューラルネットワークパラメータt_0’１３に対して選択された再構成レベルに依存することに注意されたい。図８ｂの例では、第２のニューラルネットワークパラメータt_1’１３に対する利用可能な再構成レベルの２つの異なるセット５２が存在する（異なる色で図示されている）。第１のニューラルネットワークパラメータt_0’１３に対する量子化インデックス５６が偶数（．．．，－２，０，２，．．．）であれば、第１のセット（青い点）の任意の再構成レベル２０１ａを第２のニューラルネットワークパラメータt_1’１３に対して選択することが可能である。そして、第１のニューラルネットワークパラメータt_0’に対する量子化インデックス５６が奇数（．．．，－３，－１，１，３，．．．）であれば、第２のセット（赤色点）の任意の再構成レベル２０１ｂを第２のニューラルネットワークパラメータt_1’１３に対して選択することが可能である。この例では、第１セットと第２セットの再構成レベルは、量子化ステップサイズの半分だけシフトされる（第２セットの任意の再構成レベルは、第１セットの２つの再構成レベルの間に位置する）。

【０１４１】

ニューラルネットワークパラメータ１３の依存スカラー量子化は、Ｎ次元単位体積あたりの再構成ベクトル２０１の所定の平均数に対して、ニューラルネットワークパラメータ１３の所定の入力ベクトルと最も近い利用可能な再構成ベクトルとの間の距離の期待値が低減されるという効果を有する。その結果、ニューラルネットワークパラメータの入力ベクトルとベクトル再構成されたニューラルネットワークパラメータとの間の平均歪みは、所与の平均ビット数に対して低減され得る。ベクトル量子化では、この効果をスペースフィリングゲイン（ｓｐａｃｅ－ｆｉｌｌｉｎｇ　ｇａｉｎ）と呼んでいる。ニューラルネットワークパラメータセット１３に対して依存スカラー量子化を用いると、高次元ベクトル量子化の潜在的なスペースフィリングゲインの大部分を利用することができる。そして、ベクトル量子化とは対照的に、再構成プロセス（または復号化プロセス）の実装複雑度は、独立スカラ量子化器を用いた関連するニューラルネットワークパラメータ符号化の実装複雑度に匹敵する。

【０１４２】

４．２　概要

主な変更点は、前述したように依存量子化である。再構成順序インデックスｋ＞０の再構成されたニューラルネットワークパラメータt_k’１３は、関連する量子化インデックスq_k５６だけでなく、再構成順序において先行するニューラルネットワークパラメータの量子化インデックスq_0，q_1・・・，q_{k-1}にも依存することになる。なお、依存量子化では、ニューラルネットワークパラメータ１３の再構成順序を一意に定めなければならない。量子化インデックスq_k５６に関連する再構成レベルセットに関する知識もエントロピー符号化において利用される場合、ニューラルネットワーク符号化全体の性能は、典型的に改善され得る。つまり、ニューラルネットワークパラメータに適用される再構成レベルセ

ットに基づいて、コンテキスト（確率モデル）またはコードワード表を切り替えることが典型的には好ましい。

【０１４３】

エントロピー符号化は、通常、エントロピー復号化プロセスが与えられると一意に規定される。しかし、関連するニューラルネットワークパラメータ符号化と同様に、オリジナルのニューラルネットワークパラメータを与えられた量子化インデックスの選択には多くの自由度が存在する。

【０１４４】

本明細書に記載された実施形態は、層単位ニューラルネットワーク符号化に限定されない。任意の有限のニューラルネットワークパラメータ１３の集合のニューラルネットワークパラメータ符号化にも適用可能である。

【０１４５】

特に、本方法は、ｓｅｃ．３．１で説明したようなサブ層にも適用することができる。

【０１４６】

４．３　ニューラルネットワークパラメータの依存量子化

ニューラルネットワークパラメータ１３の依存量子化とは、ニューラルネットワークパラメータ１３のために利用可能な再構成レベルセットが、再構成順序において（例えば層またはサブ層のようなニューラルネットワークパラメータの同じセットの内部において）先行するニューラルネットワークパラメータに対する選ばれた量子化インデックスに依存する概念をいう。

【０１４７】

好ましい実施形態では、再構成レベルの複数のセットが予め定義され、符号化順序において先行するニューラルネットワークパラメータの量子化インデックスに基づいて、予め定義されたセットのうちの１つが、現在のニューラルネットワークパラメータを再構成するために選択される。言い換えれば、実施形態による装置は、現在のニューラルネットワークパラメータ１３に対して、以前の、例えば先行するニューラルネットワークパラメータに対する量子化インデックス（５８）に応じて、複数５０の再構成レベルセット５２のうち再構成レベルセット４８を選択するように構成されることができる。

【０１４８】

再構成レベルのセットを定義するための好ましい実施形態は、４．３．１節で説明されている。選択された再構成レベルの識別とシグナリングは、４．３．２節に記載されている。４．３．３節では、（再構成順序において先行するニューラルネットワークパラメータの選択された量子化インデックスに基づいた）現在のニューラルネットワークパラメータについての再構成レベルの事前定義されたセットの１つを選択するための好ましい実施形態について説明する。

【０１４９】

４．３．１　再構成レベルセット

好ましい実施形態では、現在のニューラルネットワークパラメータのための許容される再構成レベルセットは、再構成レベルの予め定義されたセット５２の集合（２つ以上のセット、例えば図２および図３からのセット０およびセット１）の中から（符号化順序において先行するニューラルネットワークパラメータの量子化インデックスに基づいて）選択される。

【０１５０】

好ましい実施形態では、パラメータが量子化ステップサイズΔ（ＱＰ）を決定し、すべての（再構成レベルのすべてのセットにおける）再構成レベルが量子化ステップサイズΔの整数倍を表す。しかし、再構成レベルの各セットが量子化ステップサイズΔ（ＱＰ）の整数倍のサブセットのみを含むことに留意されたい。このような依存量子化のための構成は、再構成レベルのすべてのセットについて可能なすべての再構成レベルが量子化ステップサイズ（ＱＰ）の整数倍を表すものであり、均一再構成量子化器（ＵＲＱｓ）の拡張と考えることができる。その基本的な利点は、再構成されたニューラルネットワークパラメ

ータ１３が非常に低い計算複雑度を有するアルゴリズムによって計算され得ることである（より詳細に後述される）。

【０１５１】

再構成レベルセットは完全に不連続であることができるが、１つ以上の再構成レベルが複数のセットに含まれることも可能である（ただし、セットは他の再構成レベルにおいて依然として異なる）。

【０１５２】

好ましい実施形態では、ニューラルネットワークパラメータのための依存スカラー量子化は、再構成レベルのちょうど２つの異なるセット、例えばセット０とセット１とを使用する。そして、特に好ましい実施形態では、ニューラルネットワークパラメータt_k１３に対する２つのセットのすべての再構成レベルは、このニューラルネットワークパラメータ１３に対する量子化ステップサイズΔ_k（ＱＰ）の整数倍を表している。なお、量子化ステップサイズΔ_k（ＱＰ）は、両セットの許容再構成値に対するスケーリングファクタを表しているに過ぎない。再構成レベルの同じ２つのセットは、すべてのニューラルネットワークパラメータ１３に対して使用される。

【０１５３】

図９では、２セットの再構成レベル（セット０とセット１）に対する３つの好ましい構成（（ａ）～（ｃ））が図示されている。図９は、単一の量子化ステップサイズΔ（ＱＰ）により完全に決定される再構成レベルの２つのセットを有する依存量子化の例を示す。再構成レベルの２つの利用可能なセットは、異なる色で強調されている（セット０は青、セット１は赤）。セット内の再構成レベルを示す量子化インデックスの例は、円の下の数字で示されている。中空と塗りつぶしの円は、再構成レベルセット内の２つの異なるサブセットを示し、サブセットは再構成順序において次のニューラルネットワークパラメータの再構成レベルセットを決定するために使用することができる。図には、２つの再構成レベルセットを持つ３つの好ましい構成が示されている：（ａ）２つのセットは不一致でゼロに関して対称である、（ｂ）両方のセットはゼロに等しい再構成レベルを含むがそれ以外は不一致で、セットはゼロの周りで非対称である、（ｃ）両方のセットはゼロに等しい再構成レベルを含むがそれ以外は不一致で、どちらのセットもゼロの周りで対称的である。なお、すべての再構成レベルは、量子化ステップサイズΔの整数倍（ＩＶ）で与えられるグリッド上に存在する。さらに、特定の再構成レベルは両方のセットに含まれる可能性があることに注意する必要がある。

【０１５４】

図９（ａ）に描かれた２つのセットは不連続である。量子化ステップサイズΔ（ＱＰ）の各整数倍は、どちらかのセットにのみ含まれる。第１のセット（セット０）は量子化ステップサイズのすべての偶数整数倍（ＩＶ）を含むが、第２のセット（セット１）は量子化ステップサイズのすべての奇数整数倍を含む。両セットとも、隣接する２つの再構成レベル間の距離は、量子化ステップサイズの２倍である。これらの２つのセットは、通常、高レート量子化、すなわち、ニューラルネットワークパラメータの分散（ｖａｒｉａｎｃｅ、平方偏差）が量子化ステップサイズ（ＱＰ）よりも著しく大きい設定のものに適している。しかし、ニューラルネットワークのパラメータ符号化では、量子化器は通常、低レート領域で動作する。典型的には、多くのオリジナルのニューラルネットワークパラメータ１３の絶対値は、量子化ステップサイズ（ＱＰ）の任意のゼロでない倍数よりもゼロに近い値である。その場合、ゼロが両方の量子化セット（再構成レベルセット）に含まれると、典型的には好ましい。

【０１５５】

図９（ｂ）に示す２つの量子化セットは、いずれもゼロを含む。セット０では、ゼロに等しい再構成レベルとゼロより大きい第１の再構成レベルとの間の距離は、量子化ステップサイズ（ＱＰ）に等しく、隣接する２つの再構成レベル間の他のすべての距離は、量子化ステップサイズの２倍に等しい。同様に、セット１では、ゼロに等しい再構成レベルとゼロより小さい第１の再構成レベルとの間の距離は、量子化ステップサイズに等しく、一

方、２つの隣接する再構成レベル間の他のすべての距離は、量子化ステップサイズの２倍に等しくなる。両再構成セットは、ゼロの周りで非対称であることに注意が必要である。これは符号の確率を正確に推定することが困難になるため、非効率になる可能性がある。

【０１５６】

再構成レベルの２つのセットに関する好ましい構成を図９（ｃ）に示す。第１の量子化セット（図ではセット０と表示）に含まれる再構成レベルは、量子化ステップサイズの偶数整数倍を表す（このセットは、実際には図９（ａ）のセット０と同じであることに注意）。第２の量子化セット（図ではセット１と表示）は、量子化ステップサイズのすべての奇数整数倍を含み、さらに再構成レベルが０に等しい。なお、どちらの再構成セットもゼロについて対称である。ゼロに等しい再構成レベルは両方の再構成セットに含まれ、そうでなければ再構成セットは不連続である。両再構成セットの組み合わせは、量子化ステップサイズのすべての整数倍を含む。

【０１５７】

言い換えれば実施形態によれば、例えばニューラルネットワークパラメータ１３を符号化／復号化するための装置からなり、複数５０の再構成レベルセット５２の数は２（例えばセット０、セット１）であり、複数の再構成レベルセットは、ゼロおよび所定の量子化ステップサイズの偶数倍を含む第１の再構成レベルセット（セット０）、およびゼロおよび所定の量子化ステップサイズの奇数倍を含む第２の再構成レベルセット（セット１）含む。

【０１５８】

さらに、すべての再構成レベルセットのすべての再構成レベルは、所定の量子化ステップサイズ（ＱＰ）の整数倍（ＩＶ）を表してもよく、実施形態による、例えばニューラルネットワークパラメータ１３を復号化するための装置は、各ニューラルネットワークパラメータについて、中間整数値、例えば、それぞれのニューラルネットワークパラメータについて選択された再構成レベルセットおよびそれぞれのニューラルネットワークパラメータ１３’についてのエントロピー復号化量子化インデックス５８に応じた整数倍（ＩＶ）を導出することによって、及び、それぞれのニューラルネットワークパラメータ１３について、それぞれのニューラルネットワークパラメータの中間値をそれぞれのニューラルネットワークパラメータ１３についての所定の量子化ステップサイズに乗じることによって、ニューラルネットワークパネル１３を逆量子化するよう構成されることがある。

【０１５９】

それぞれ、すべての再構成レベルセットのすべての再構成レベルは、所定の量子化ステップサイズ（ＱＰ）の整数倍（ＩＶ）を表してもよく、装置、例えば、実施形態によれば、ニューラルネットワークパラメータ１３を符号化するために、各ニューラルネットワークパラメータについて、それぞれのニューラルネットワークパラメータの選択された再構成レベルセットとそれぞれのニューラルネットワークパラメータのエントロピー符号化量子化インデックスとに応じた中間整数値を導出することによって、及び、それぞれのニューラルネットワークパラメータについて、それぞれのニューラルネットワークパラメータの中間値をそれぞれのニューラルネットワークパラメータの所定の量子化ステップサイズに乗じることによって、同じものが逆量子化可能になるよう、ニューラルネットワークパネルを量子化するよう構成されることができる。

【０１６０】

本明細書で規定する実施形態は、図９に示す構成に限定されるものではない。他の任意の２つの異なる再構成レベルセットを使用することができる。複数の再構成レベルが両方のセットに含まれてもよい。あるいは、両方の量子化セットの組み合わせが、量子化ステップサイズのすべての可能な整数倍を含んでいない可能性がある。さらに、ニューラルネットワークパラメータの依存スカラー量子化には、２つ以上の再構成レベルのセットを使用することが可能である。

【０１６１】

４．３．２　選択された再構成レベルのシグナリング

符号化器が許容される再構成レベルの中から選択する再構成レベルは、ビットストリーム１４の内部で示されなければならない。従来の独立スカラー量子化と同様に、これは、重みレベルとも呼ばれる、いわゆる量子化インデックス５６を用いて実現することができる。量子化インデックス５６（または重みレベル）は、量子化セット５２の内部（すなわち、再構成レベルのセットの内部）で利用可能な再構成レベルを一意に識別する整数値である。量子化インデックス５６は、（任意のエントロピー符号化技術を用いる）ビットストリーム１４の一部として復号化器に送られる。復号化器側では、再構成されたニューラルネットワークパラメータ１３は、再構成レベルの現在のセット４８（これは、符号化／再構成順序において先行する量子化インデックスによって決定される）および現在のニューラルネットワークパラメータ１３’に対する送信された量子化インデックス５６に基づいて一意に計算することが可能である。

【０１６２】

好ましい実施形態では、再構成レベルのセット（または量子化セット）内部の再構成レベルへの量子化インデックス５６の割り当ては、次のルールに従う。説明のために、図９の再構成レベルには、関連する量子化インデックス５６が付けられている（量子化インデックスは、再構成レベルを表す円の下の数字によって与えられる）。再構成レベルセットが０に等しい再構成レベルを含む場合、０に等しい量子化インデックスは０に等しい再構成レベルに割り当てられる。１に等しい量子化インデックスは、０より大きい最小の再構成レベルに割り当てられ、２に等しい量子化インデックスは、０より大きい次の再構成レベル（すなわち、０より大きい２番目に小さい再構成レベル）に割り当てられる、等々である。あるいは、言い換えれば、０より大きい再構成レベルには、その値の昇順において０より大きい整数値で（すなわち、１、２、３などで）ラベル付けされる。同様に、量子化インデックス－１は、０より小さい最大の再構成レベルに割り当てられ、量子化インデックス－２は、０より小さい次の（すなわち、２番目に大きい）再構成レベルに割り当てられる、等である。または、言い換えると、０より小さい再構成レベルには、値の降順において０より小さい整数値（つまり、－１、－２、－３など）がラベル付けされる。図９の例では、（０に等しい再構成レベルを含まない）図９（ａ）のセット１を除くすべての量子化セットについて、説明した量子化インデックスの割り当てを例示している。

【０１６３】

０に等しい再構成レベルを含まない量子化セットに対して、量子化インデックス５６を再構成レベルに割り当てる方法の１つは、以下の通りである。０より大きい全ての再構成レベルには（値の昇順において）０より大きい量子化インデックスを割り当て、０より小さい全ての再構成レベルには（値の降順において）０より小さい量子化インデックスを割り当てる。したがって、量子化インデックス５６の割り当ては、基本的に０に等しい再構成レベルを含む量子化セットと同じ概念に従うが、０に等しい量子化インデックスが存在しない点が異なる（図９（ａ）の量子化セット１のラベルを参照）。量子化インデックス５６のエントロピー符号化では、その点を考慮する必要がある。例えば、量子化インデックス５６は、その絶対値（０からサポートされる最大値までの範囲）を符号化し、０に等しくない絶対値については、量子化インデックス５６の符号を付加的に符号化することによって送信されることが多い。０に等しい量子化インデックス５６がない場合、エントロピー符号化は、絶対レベルから１を引いた値が送信され（対応するシンタックス要素の値は０から最大サポート値までの範囲）、符号が常に送信されるように修正され得る。代替案として、量子化インデックス５６を再構成レベルに割り当てるための割り当てルールが修正され得る。例えば、０に近い再構成レベルの１つは、０に等しい量子化インデックスでラベル付けされ得る。そして、残りの再構成レベルは、以下のルールによってラベル付けされる。０に等しい量子化インデックスを持つ再構成レベルより大きい再構成レベルには、０より大きい量子化インデックスが割り当てられる（量子化インデックスは再構成レベルの値と共に増加する）。そして、０より小さい量子化インデックスは、量子化インデックスが０に等しい再構成レベルより小さい再構成レベルに割り当てられる（量子化インデックスは、再構成レベルの値と共に減少する）。このような割り当ての一つの可能性を

、図９（ａ）の括弧内の数字で示す（括弧内の数字が与えられていない場合は、他の数字が適用される）。

【０１６４】

上述のように、好ましい実施形態では、再構成レベルの２つの異なるセット（これを量子化セットとも呼ぶ）が使用され、両方のセットの内部の再構成レベルは、量子化ステップサイズ（ＱＰ）の整数倍を表す。それは、量子化ステップサイズが、層ベースで（例えば、ビットストリーム１４の内部に層量子化パラメータを送信することによって）、またはニューラルネットワークパラメータ１３の別の有限セット（例えば、ブロック）で（例えば、ビットストリーム１４の内部にブロック量子化パラメータを送信することによって）変更される、場合を含んでいる。

【０１６５】

量子化ステップサイズ（ＱＰ）の整数倍を表す再構成レベルの使用は、復号化器側でのニューラルネットワークパラメータ１３の再構成のための計算量の少ない複雑なアルゴリズムを可能にする。これを図９（ｃ）の好ましい例に基づいて以下に説明する（他の構成、特に図９（ａ）および図９（ｂ）に示す設定についても同様の単純なアルゴリズムが存在する）。図９（ｃ）に示す構成では、第１の量子化セットは量子化ステップサイズ（ＱＰ）のすべての偶数整数倍を含み、第２の量子化セットは量子化ステップサイズのすべての奇数整数倍に０に等しい再構成レベルを加えたもの（これは両方の量子化セットに含まれる）が含まれる。ニューラルネットワークパラメータの再構成プロセスは、図１０の疑似コードで規定されるアルゴリズムと同様に実装され得る。図１０は、ニューラルネットワークパラメータ１３の再構成プロセスに関する好ましい例を示す擬似コードである。ｋは現在のニューラルネットワークパラメータ１３’の再構成順序を指定するインデックスを表し、現在のニューラルネットワークパラメータの量子化インデックス５６はｌｅｖｅｌ［ｋ］２１０で表し、現在のニューラルネットワークパラメータ１３’に適用する量子化ステップサイズΔ_k（ＱＰ）はｑｕａｎｔ＿ｓｔｅｐ＿ｓｉｚｅ［ｋ］で表し、ｔｒｅｃ［ｋ］２２０は再構成したニューラルネットワークのパラメータt_k’の値を表している。変数ｓｅｔＩｄ［ｋ］２４０は、現在のニューラルネットワークパラメータ１３’に適用される再構成レベルのセットを指定する。それは、再構成順序において先行するニューラルネットワークパラメータに基づいて決定される；ｓｅｔＩｄ［ｋ］の可能な値は０及び１である。変数ｎは、量子化ステップサイズ（ＱＰ）の整数係数、例えば中間値ＩＶを規定する。；それは、選択された再構成レベのセット（すなわちｓｅｔＩｄ［ｋ］の値）及び送信された量子化インデックスレベル［ｋ］により与えられる。

【０１６６】

図１０の疑似コードにおいて、ｌｅｖｅｌ［ｋ］は、ニューラルネットワークパラメータt_k１３について送信される量子化インデックス５６を示し、（０または１に等しい）ｓｅｔＩｄ［ｋ］は再構成レベルの現在のセットの識別子を指定する（それは以下でより詳細に説明するように再構成順序において先行する量子化インデックス５６に基づいて決定される）。変数ｎは、量子化インデックスｌｅｖｅｌ［ｋ］とセット識別子ｓｅｔＩｄ［ｋ］によって与えられる量子化ステップサイズ（ＱＰ）の整数倍を表す。ニューラルネットワークパラメータ１３が、量子化ステップサイズΔ_k（ＱＰ）の偶数整数倍を含む再構成レベル（ｓｅｔＩｄ［ｋ］＝＝０）の第１のセットを用いて符号化される場合、変数ｎは、送信された量子化インデックス５６の２倍となる。この場合は、図９（ｃ）の第１の量子化セットＳｅｔ０の再構成レベルで表すことができ、Ｓｅｔ０は量子化ステップサイズ（ＱＰ）の偶数整数倍を全て含んでいる。ニューラルネットワークパラメータ１３が再構成レベル（ｓｅｔＩｄ［ｋ］＝＝１）の第２のセットを用いて符号化される場合、（ａ）ｌｅｖｅｌ［ｋ］が０に等しければ、ｎも０に等しくなり、（ｂ）ｌｅｖｅｌ［ｋ］が０より大きい場合、ｎは量子化インデックスｌｅｖｅｌ［ｋ］の２倍から１を引いたものに等しくなり、（ｃ）ｌｅｖｅｌ［ｋ］が０未満ならｎは量子化インデックスレベル［ｋ］の２倍に１を加えたものに等しくなる、の三つの場合が存在する。これは符号関数を用いて指定することができる。

$$\mathrm{sign}(x)=\begin{cases}1 & : \ x>0\\ 0 & : \ x=0\\ -1 & : \ x<0\end{cases}.$$

【０１６７】

そして、第２の量子化セットを使用する場合、変数ｎは、量子化インデックスｌｅｖｅｌ［ｋ］の２倍から量子化インデックスの符号関数ｓｉｇｎ（ｌｅｖｅｌ［ｋ］）を引いた値に等しくなる。この場合は、図９（ｃ）の第２の量子化セットＳｅｔ１の再構成レベルで表すことができ、Ｓｅｔ１は量子化ステップサイズ（ＱＰ）の奇数整数倍を全て含む。

【０１６８】

（量子化ステップサイズの整数倍を指定する）変数ｎが決定されると、ｎに量子化ステップサイズΔ_kを乗じることによって、再構成されたニューラルネットワークパラメータt_k’が求められる。

【０１６９】

言い換えれば、複数５０の再構成レベルセット５２のうち再構成レベルセット５２の数は２であってもよく、本発明の実施形態による、例えばニューラルネットワークパラメータ１３を復号化および／または符号化するための装置は、以下のようにしてそれぞれのニューラルネットワークパラメータの中間値を導出するように構成され得る。

それぞれのニューラルネットワークパラメータについての選択された再構成レベルセットが第１のセットである場合、それぞれのニューラルネットワークパラメータに対する量子化インデックスを２倍して、それぞれのニューラルネットワークパラメータに対する中間値を求め；及び、

それぞれのニューラルネットワークパラメータについての選択された再構成レベルセットが２番目のセットであり、それぞれのニューラルネットワークパラメータの量子化インデックスがゼロに等しい場合、それぞれのサンプルの中間値をゼロに設定し；及び、

それぞれのニューラルネットワークパラメータについての選択された再構成レベルセットが第２のセットであり、それぞれのニューラルネットワークパラメータに対する量子化インデックスがゼロより大きい場合、それぞれのニューラルネットワークパラメータに対する量子化インデックスを２倍し、その乗算結果から１を引いてそれぞれのニューラルネットワークパラメータに対する中間値を得て、及び、

現在のニューラルネットワークパラメータについての選択された再構成レベルセットが第２のセットであり、それぞれのニューラルネットワークパラメータに対する量子化インデックスがゼロより小さい場合、それぞれのニューラルネットワークパラメータの量子化インデックスを２倍し、その乗算結果に１を加えて、それぞれのニューラルネットワークパラメータの中間値を得る。

【０１７０】

４．３．３　ニューラルネットワークパラメータの依存的再構成

４．３．１節および４．３．２節で説明した再構成レベルのセットの選択に加えて、ニューラルネットワークパラメータ符号化における依存スカラー量子化のもう一つの重要な設計側面は、定義された量子化セット（再構成レベルのセット）間の切り替えに使用するアルゴリズムである。使用されるアルゴリズムによって、ニューラルネットワークパラメータ１３のＮ次元空間（したがって、再構成されたサンプルのＮ次元空間も同様）で達成できる「パッキング密度」が決まる。パッキング密度が高ければ高いほど、最終的に符号化効率が向上する。

【０１７１】

次のニューラルネットワークパラメータのための再構成レベルのセットを決定する好ましい方法は、図１１に示されるように、量子化セットの分割に基づくものである。図１１は、本発明の実施形態による再構成レベルのセットを２つのサブセットに分割するための一例を示す図である。示された２つの量子化セットは、図９（ｃ）の好ましい例の量子化セットである。量子化セット０の２つのサブセットは、「Ａ」及び「Ｂ」を用いてラベル付けされ、量子化セット１の２つのサブセットは、「Ｃ」及び「Ｄ」を用いてラベル付けされている。なお、図１１に示す量子化セットは、図９（ｃ）の量子化セットと同じ量子化セットである。２つ（またはそれ以上）の量子化セットのそれぞれは、２つのサブセットに分割される。図１１の好ましい例では、第１の量子化セット（セット０と表示）は、２つのサブセット（ＡおよびＢと表示される）に分割され、第２の量子化セット（セット１と表示される）も、２つのサブセット（ＣおよびＤと表示される）に分割される。唯一の可能性ではないにしても、各量子化セットに対する分割は、好ましくは、直接隣接する再構成レベル（したがって、隣接する量子化インデックス）が異なるサブセットと関連付けられるように行われる。好ましい実施形態では、各量子化セットは、２つのサブセットに分割される。図９において、量子化セットのサブセットへの分割は、中空および充填された円によって示されている。

【０１７２】

図１１および図９（ｃ）に示される特に好ましい実施形態では、以下の分割ルールが適用される。

－　サブセットＡは、量子化セット０のすべての偶数量子化インデックスから構成される。
－　サブセットＢは、量子化セット０のすべての奇数量子化インデックスから構成される。
－　サブセットＣは、量子化セット１のすべての偶数量子化インデックスから構成される。
－　サブセットＤは、量子化セット１のすべての奇数量子化インデックスで構成される。

【０１７３】

使用されるサブセットは、典型的には、ビットストリーム１４の内部で明示的に示されないことに留意されたい。その代わりに、使用される量子化セット（例えば、セット０またはセット１）および実際に送信された量子化インデックス５６に基づいて導出することができる。図１１に示す好ましい分割の場合、サブセットは、送信された量子化インデックスレベル及び１のビット単位「ａｎｄ」演算によって導出することができる。サブセットＡは、（ｌｅｖｅｌ＆１）が０に等しいセット０のすべての量子化インデックスからなり、サブセットＢは、（ｌｅｖｅｌ＆１）が１に等しいセット０のすべての量子化インデックスからなり、サブセットＣは、（ｌｅｖｅｌ＆１）が０に等しいセット１のすべての量子化インデックスからなり、サブセットＤは、（ｌｅｖｅｌ＆１）が１に等しいセット１のすべての量子化インデックスからなる。

【０１７４】

好ましい実施形態では、現在のニューラルネットワークパラメータ１３’を再構成するために使用される量子化セット（許容再構成レベルのセット）は、最後の２つ以上の量子化インデックス５６に関連付けられるサブセットに基づいて決定される。最後の２つのサブセット（これは最後の２つの量子化インデックスによって与えられる）が使用される例を、表１に示す。この表によって指定される量子化セットの決定は、好ましい実施形態を表す。他の実施形態では、現在のニューラルネットワークパラメータ１３’に対する量子化セットは、最後の３つ以上の量子化インデックス５６と関連するサブセットによって決定される。層の第１のニューラルネットワークパラメータ（またはニューラルネットワークパラメータのサブセット）については、（先行するニューラルネットワークパラメータがないため）先行するニューラルネットワークパラメータのサブセットに関するデータはない。好ましい実施形態では、このような場合に予め定義された値を使用する。特に好ましい実施形態では、利用可能でない全てのニューラルネットワークパラメータについて、サブセットＡを推論する。つまり、第１のニューラルネットワークパラメータを再構成する場合、先行する２つのサブセットを「ＡＡ」（又は、先行する３つのニューラルネットワークパラメータを考慮する場合は「ＡＡＡ」）と推測し、したがって、表１に従って、量子化セット０を使用する。また、２番目のニューラルネットワークパラメータについては、直前の量子化インデックスのサブセットをその値によって決定するが（１番目のニューラルネットワークパラメータについてはセット０が使用されるので、サブセットはＡまたはＢのいずれか）、２番目の最後の量子化インデックス（これは存在しない）についてのサブセットはＡに等しいと推測する。もちろん、存在しない量子化インデックスについてのデフォルト値の推測には他の任意のルールを使用することが可能である。また、存在しない量子化インデックスのデフォルトサブセットを導出するために、他のシンタックス要素を使用することも可能である。さらなる代替案として、初期化のために、ニューラルネットワークパラメータ１３の先行するセットの最後の量子化インデックス５６を使用することも可能である。

【０１７５】

表１：本発明の実施形態による２つの最後の量子化インデックスに関連するサブセットに基づいて、次のニューラルネットワークパラメータに使用される量子化セット（利用可能な再構成レベルのセット）を決定するための例である。サブセットは、左の表の欄に示されており、それらは、（２つの最後の量子化インデックスについての）使用される量子化セットおよび（量子化インデックスのパリティによって決定されてよい）いわゆるパスによって一義的に決定される。量子化セットと、括弧内にはサブセットのためのパスが左から２番目の列にリストアップされている。３列目には、関連する量子化セットを指定する。最後の列には、いわゆる状態変数の値が示されており、量子化セットを決定するためのプロセスを簡略化するために使用することができる。

【０１７６】

【表１】

最後の２つの量子化インデックスのサブセット	最後の２つの量子化インデックスについての量子化セットとパス（括弧内）	現在のニューラルネットワークパラメータに対する量子化セット	状態変数
ＡＡ	0(0), 0(0)	0	0
ＡＢ	0(0), 0(1)	0	0
ＡＣ	0(0), 1(0)	1	1
ＡＤ	0(0), 1(1)	1	1
ＢＡ	0(1), 0(0)	1	1
ＢＢ	0(1), 0(1)	1	1
ＢＣ	0(1), 1(0)	0	0
ＢＤ	0(1), 1(1)	0	0
ＣＡ	1(0), 0(0)	0	2
ＣＢ	1(0), 0(1)	0	2
ＣＣ	1(0), 1(0)	1	3
ＣＤ	1(0), 1(1)	1	3
ＤＡ	1(1), 0(0)	1	3
ＤＢ	1(1), 0(1)	1	3
ＤＣ	1(1), 1(0)	0	2
ＤＤ	1(1), 1(1)	0	2

【０１７７】

なお、量子化インデックス５６のサブセット（Ａ、Ｂ、Ｃ又はＤ）は、使用する量子化セット（セット０またはセット１）と量子化セット内部の使用するサブセット（例えば、セット０ならＡまたはＢ、セット１ならＣまたはＤ）により決定されることに注意する必要がある。量子化セットの内部で選択されたサブセットは、（後述するように、依存量子化プロセスをトレリス構造で表現する場合に、パスを指定するため）パスとも呼ばれる。我々の慣例では、パスは０か１のどちらかに等しい。すると、サブセットＡはセット０におけるパス０に対応し、サブセットＢはセット０におけるパス１に対応し、サブセットＣはセット１におけるパス０に対応し、サブセットＤはセット１におけるパス１に対応する。したがって、次のニューラルネットワークパラメータの量子化セットも、最後の２つ（またはそれ以上）の量子化インデックスに関連付けられた量子化セット（セット０またはセット１）とパス（パス０またはパス１）によって一意に決定される。表１では、関連する量子化セットとパスが２列目に指定されている。

【０１７８】

なお、パスは、例えば２値関数による簡単な演算で決定できることが多い。例えば、図１１のような構成の場合、パスは次式で与えられる。

$$path = (level[k] \& 1),$$

ここで、ｌｅｖｅｌ［ｋ］は量子化インデックス（重みレベル）５６を表し、演算子＆はビット単位「ａｎｄ」（２補整数演算）を指定する。

【０１７９】
言い換えれば、複数５０の再構成レベルセット５２のうちの再構成レベルセット５２の数は、例えばセット０とセット１との２つであってもよく、本発明の実施形態による、例えばニューラルネットワークパラメータ１３を復号化するための装置は、それぞれのニューラルネットワークパラメータについての再構成レベルの選択されたセットとそれぞれのニューラルネットワークパラメータについての量子化インデックスの２値関数とに基づいて、それぞれのニューラルネットワークパラメータについてサブセットインデックスを導き、結果として４つの可能な値、例えばサブセットインデックスについてＡ、Ｂ、Ｃ、またはＤを得ることができ、現在のニューラルネットワークパラメータ１３’について、以前に復号化されたニューラルネットワークパラメータについてのサブセットインデックスに応じて、複数５０の再構成レベルセット５２のうち再構成レベルのセット４８を選択するように構成することができる。

【０１８０】
本発明によるさらなる実施形態は、現在のニューラルネットワークパラメータ１３’について、例えば表１の１列目に示すように、直前に復号化された多数のニューラルネットワークパラメータのサブセットインデックスに対応する選択ルールを使用して、複数５０の再構成レベルセット５）のうち再構成レベルセット４８を選択５４し、ニューラルネットワークパラメータのすべて、または一部について選択ルールを使用するように構成される装置である。

【０１８１】
さらなる実施形態によれば、選択ルールが対応する、直前に復号化されたニューラルネットワークパラメータの数は２つであり、例えば表１に示すように、２つの最後の量子化インデックスのサブセットである。

【０１８２】
追加の実施形態によれば、各ニューラルネットワークパラメータのサブセットインデックスは、それぞれのニューラルネットワークパラメータについての再構成レベルの選択されたセットと、それぞれのニューラルネットワークパラメータについての量子化インデックスの、例えばｐａｔｈ＝（ｌｅｖｅｌ［ｋ］＆１）を用いたパリティに基づいて導出される。

【０１８３】
それぞれ、実施形態によるニューラルネットワークパラメータ１３を符号化するための装置については、複数５０の再構成レベルセット５２の再構成レベルセット５２の数は、例えばセット０とセット１との２つであってよく、装置は、それぞれのニューラルネットワークパラメータのための再構成レベルの選択されたセットとそれぞれのニューラルネットワークパラメータについての量子化インデックスの２値関数とに基づいて、それぞれのニューラルネットワークパラメータのためのサブセットインデックスを導出して、サブセットインデックスについて４つの可能な値、例えば、Ａ、Ｂ、Ｃ、Ｄを生じさせ、現在のニューラルネットワークパラメータ１３’に対して、以前に符号化されたニューラルネットワークパラメータについてのサブセットインデックスに応じて複数５０の再構成レベルセ

ット５２のうちの再構成レベルセット４８を選択５４するように構成することができる。

【０１８４】

本発明によるさらなる実施形態は、現在のニューラルネットワークパラメータ１３’について、例えば表１の第１列に示すように、直前に符号化された多数のニューラルネットワークパラメータのサブセットインデックスに対応する選択ルールを用いて、複数５０の再構成レベルセット５２のうち再構成レベルのセット４８を選択５４し、ニューラルネットワークパラメータのすべて、または一部について選択ルールを用いるよう構成される装置を備える。

【０１８５】

さらなる実施形態によれば、選択ルールが対応する、直前に符号化されたニューラルネットワークパラメータの数は２であり、例えば表１に示すように、２つの最後の量子化インデックスのサブセットである。

【０１８６】

追加の実施形態によれば、各ニューラルネットワークパラメータについてのサブセットインデックスは、それぞれのニューラルネットワークパラメータについての再構成レベルの選択されたセットと、それぞれのニューラルネットワークパラメータの量子化インデックスのパリティ、例えばｐａｔｈ＝（ｌｅｖｅｌ［ｋ］＆１）を使用して、導出される。

【０１８７】

量子化セット５２（セット０とセット１）の間の遷移は、状態変数によってエレガントに表現することも可能である。そのような状態変数の例を表１の最後の列に示す。この例では、状態変数は４つの可能な値（０、１、２、３）を有する。一方、状態変数は、現在のニューラルネットワークパラメータ１３’に使用される量子化セットを指定する。表１の好ましい例では、状態変数が０または２に等しい場合にのみ、量子化セット０が使用され、状態変数が１または３に等しい場合にのみ、量子化セット１が使用される。一方、状態変数は、量子化セット間の可能な遷移も指定する。状態変数を用いることで、表１のルールをより小さな状態遷移表で記述することができる。一例として、表２は、表１で与えられたルールの状態遷移表を規定したものである。これは、好ましい実施形態を表している。現在の状態を指定して、現在のニューラルネットワークのパラメータ（２列目）に量子化セットを規定した。それはさらに、選択された量子化インデックス５６に関連するパスに基づいて状態遷移を指定する（パスは、量子化セットが与えられた場合、使用されるサブセットＡ、Ｂ、Ｃ、またはＤを指定する）。状態変数の概念を使用することによって、実際に選択されたサブセットを追跡する必要がないことに留意されたい。ある層のニューラルネットワークのパラメータを再構築する際には、状態変数を更新し、使用される量子化インデックスのパスを決定すれば十分である。

【０１８８】

表２：本発明の実施形態による、４つの状態を有する構成の状態遷移表の好ましい例。

【表２】

現在の状態	電流係数に対する量子化セット	次の状態	
		パス０	パス１
0	0	0	1
1	1	2	3
2	0	1	0
3	1	3	2

【０１８９】

つまり、実施形態によるニューラルネットワークパラメータを復号化するための装置であって、当該装置は、現在のニューラルネットワークパラメータ１３’に関連する状態に応

じて、複数５０の再構成レベルセット５２のうちの量子化レベルのセット４８を決定することによって、及び、直前のニューラルネットワークパラメータのデータストリームから復号化された量子化インデックス５８に応じて後続のニューラルネットワークパラメータの状態を更新することによって、現在のニューラルネットワークパラメータ１３’について、状態遷移プロセスにより複数５０の再構成レベルセット５２のうちの再構成レベルセット４８を選択５４にように構成されることができる。

【０１９０】

それぞれ、実施形態によるニューラルネットワークパラメータ１３を符号化するための装置について、前記装置は、現在のニューラルネットワークパラメータ１３’に関連する状態に応じて、複数５０の再構成レベルセット５２のうちの再構成レベルのセット４８を決定することによって、及び、直前のニューラルネットワークパラメータのデータストリームに符号化された量子化インデックス５８に応じて、後続のニューラルネットワークパラメータについての状態を更新することによって、現在のニューラルネットワークパラメータ１３’について、状態遷移プロセスによって複数５０の再構成レベルセット５２のうちの再構成レベルのセット４８を選択５４するように構成されてもよい。

【０１９１】

本発明の好ましい実施形態では、パスは、量子化インデックスのパリティによって与えられる。ｌｅｖｅｌ［ｋ］を現在の量子化インデックスとすると、次式に従って決定することができる。

$$path = (level[k] \& 1),$$

ここで、演算子&は、２補整数演算におけるビット単位「ａｎｄ」を表す。

【０１９２】

言い換えれば、実施形態による、例えばニューラルネットワークパラメータを復号化するための装置は、直前のニューラルネットワークパラメータについてのデータストリームから復号化された量子化インデックス５８の２値関数を用いて、例えば表２に従って、後続のニューラルネットワークパラメータの状態を更新するよう構成されることができる。

【０１９３】

さらに、実施形態による装置は、直前のニューラルネットワークパラメータについてのデータストリーム１４から復号化された量子化インデックス５８のパリティ、例えばｐａｔｈ＝（ｌｅｖｅｌ［ｋ］＆１）を使用して、後続のニューラルネットワークパラメータのための状態を更新するように構成されることができる。

【０１９４】

それぞれ、実施形態によるニューラルネットワークパラメータ１３を符号化するための装置について、当該装置は、直前のニューラルネットワークパラメータについてのデータストリームに符号化された量子化インデックス５８の２値関数を用いて、後続のニューラルネットワークパラメータについての状態を更新するように構成されることができる。

【０１９５】

さらに、実施形態による例えばニューラルネットワークパラメータ１３を符号化するための装置は、直前のニューラルネットワークパラメータについてのデータストリームに符号化された量子化インデックス５８のパリティを用いて、後続のニューラルネットワークパラメータについての状態を、例えば表２に従って更新するように構成されることができる。

【０１９６】

好ましい実施形態では、４つの可能な値を有する状態変数が使用される。他の実施形態では、異なる数の可能な値を有する状態変数が使用される。特に興味深いのは、状態変数

の可能な値の数が２の整数乗、すなわち、４、８、１６、３２、６４などを表す状態変数である。（表１および表２で与えられるように）好ましい構成において、４つの可能な値を有する状態変数は、現在の量子化セットが２つの最後の量子化インデックスのサブセットによって決定されるアプローチと同等であることに注意されたい。８つの可能な値を持つ状態変数は、現在の量子化セットが３つの最後の量子化インデックスのサブセットによって決定される同様のアプローチに対応する。１６個の可能な値を持つ状態変数は、現在の量子化セットが最後の４つの量子化インデックスのサブセットによって決定されるアプローチに対応する。一般に、２の整数乗に等しい可能な値の数を有する状態変数を使用することが好ましいにもかかわらず、実施形態は、この設定に限定されない。

【０１９７】

特に好ましい実施形態では、８つの可能な値（０、１、２、３、４、５、６、７）を有する状態変数が使用される。好ましい例の表３では、状態変数が０、２、４または６に等しい場合にのみ、量子化セット０が使用され、状態変数が１、３、５または７に等しい場合にのみ、量子化セット１が使用される。

【０１９８】

実施形態による、８つの状態を有する構成の状態遷移表の好ましい例である。

【表３】

現在の状態	電流係数についての量子化セット	次の状態	
		パス０	パス１
0	0	0	2
1	1	7	5
2	0	1	3
3	1	6	4
4	0	2	0
5	1	5	7
6	0	3	1
7	1	4	6

【０１９９】

すなわち、本発明の実施形態によれば、状態遷移プロセスにおいて、４つ又は８つの可能な状態の間で遷移するように構成される。

【０２００】

さらに、実施形態による、ニューラルネットワークパラメータ１３を復号化／符号化するための装置は、状態遷移プロセスにおいて、偶数の可能な状態の間で遷移し、複数５０の再構成レベルセット５２の数が２であるように構成されることができ、ここで、現在のニューラルネットワークパラメータ１３’について、現在のニューラルネットワークパラメータ１３’に関連する状態に応じて、量子化セット５２のうちの量子化レベルのセット４８を決定することで、状態が偶数の可能な状態の前半に属する場合、複数５０の再構成レベルセット５２のうちの第１の再構成レベルセットが決定され、状態が偶数の可能な状態の後半に属する場合、複数５０の再構成レベルセット５２のうちの第２の再構成レベルセットが決定される。

【０２０１】

さらなる実施形態による、例えばニューラルネットワークパラメータ１３を復号化するための装置は、直前のニューラルネットワークパラメータのデータストリームから復号化された量子化インデックス５８の状態およびパリティの組み合わせを、後続のニューラルネットワークパラメータに関連する別の状態上にマッピングする遷移表によって、状態の

更新を実行するように構成されてもよい。
【０２０２】
それに応じて、実施形態によるニューラルネットワークパラメータ１３を符号化するための装置は、直前のニューラルネットワークパラメータについてのデータストリームに符号化された量子化インデックス５８の状態とパリティとの組み合わせを後続のニューラルネットワークパラメータに関連する別の状態にマッピングする遷移表によって状態の更新を行うように構成されることができる。
【０２０３】
状態遷移の概念を用いると、現在の状態、ひいては現在の量子化セットは、（再構成順序において）以前の状態と以前の量子化インデックス５６によって一意に決定される。しかし、有限集合（例えば層）の第１のニューラルネットワークパラメータ１３については、以前の状態および以前の量子化インデックスが存在しない。したがって、ある層の第１のニューラルネットワークパラメータに対する状態が一意に定義されることが要求される。さまざまな可能性がある。好ましい選択肢は以下の通りである。
・層のための第１の状態は、固定された事前定義された値に常に等しく設定される。好ましい実施形態では、第１の状態は０に等しく設定される。
・第１の状態の値は、ビットストリーム１４の一部として明示的に送信される。これは、可能な状態値のサブセットのみが対応するシンタックス要素によって示され得るアプローチを含む。
・第１の状態の値は、その層の他のシンタックス要素に基づいて導出される。つまり、対応するシンタックス要素（またはシンタックス要素）が復号化器への他の態様のシグナリングに使用されても、それらは、依存スカラー量子化のための第１の状態を導出するために追加的に使用されることを意味する。
【０２０４】
依存スカラー量子化の状態遷移の概念は、復号化器におけるニューラルネットワークパラメータ１３の再構成のための複雑さが少ない実装を可能にする。単層のニューラルネットワークパラメータの再構成プロセスに関する好ましい例を、Ｃ言語形式の擬似コードを用いて図１２に示す。図１２は、本発明の実施形態による層のニューラルネットワークパラメータ１３の再構成プロセスのための好ましい例を示す擬似コードの一例である。なお、量子化インデックスの導出と、例えば量子化ステップサイズを用いた、あるいは代替的にコードブックを用いた再構成値の導出は、次々と別々のループで行われてもよい。すなわち、言い換えれば、「ｎ」の導出と状態更新を第１のループで行い、「ｔｒｅｃ」の導出をもう１つの別の第２のループで行うようにすることができる。アレイレベル（ａｒｒａｙ　ｌｅｖｅｌ）２１０は、その層の送信されたニューラルネットワークパラメータレベル（量子化インデックス５６）を表し、アレイｔｒｅｃ（ａｒｒａｙ　ｔｒｅｃ）２２０は、対応する再構成されたニューラルネットワークパラメータ１３を表す。現在のニューラルネットワークパラメータ１３’に適用される量子化ステップサイズΔｋ（ＱＰ）は、ｑｕａｎｔ＿ｓｔｅｐ＿ｓｉｚｅ［ｋ］で示される。２ｄ表ｓｔｔａｂ（２ｄ　ｔａｂｌｅ　ｓｔｔａｂ）２３０は、例えば表１、表２及び／又は表３のいずれかに従って状態遷移表を指定し、表ｓｅｔＩｄ（ｔａｂｌｅ　ｓｅｔＩｄ）２４０は、状態２５０に関連付けられる量子化セットを指定する。
【０２０５】
図１２の擬似コードにおいて、インデックスｋは、ニューラルネットワークパラメータの再構成順序を指定する。最後のインデックスｌａｙｅｒＳｉｚｅは、最後に再構成されたニューラルネットワークパラメータの再構成インデックスを指定する。変数ｌａｙｅｒＳｉｚｅは、層内のニューラルネットワークパラメータの数と等しく設定されてもよい。各単一のニューラルネットワークパラメータの再構成プロセスは、図１０の例と同じである。図１０の例と同様に、量子化インデックスはｌｅｖｅｌ［ｋ］２１０で表され、関連する再構成されたニューラルネットワークパラメータはｔｒｅｃ［ｋ］２２０で表される。また、状態変数は状態（ｓｔａｔｅ）２１０で表される。図１２の例では、状態は、層

の始めに０に等しく設定されることに留意されたい。しかし、上述したように、他の初期化（例えば、いくつかのシンタックス要素の値に基づいて）も可能である。１ｄ表ｓｅｔＩｄ［］（１ｄ　ｔａｂｌｅ　ｓｅｔＩｄ［］）２４０は、状態変数の異なる値に関連付けられる量子化セットを指定し、２ｄ表ｓｔｔａｂ［］［］（２ｄ　ｔａｂｌｅ　ｓｔｔａｂ［］［］）２３０は、現在の状態（第１の引数）およびパス（第２の引数）を与えられた状態遷移を指定する。この例では、パスは（ビット単位と演算子＆を使用する）量子化インデックスのパリティで与えられるが、他の概念も可能である。Ｃ言語スタイルのシンタックスにおいて、表の例を図１３と図１４に示す（これらの表は表２、表３と同一であり、言い換えれば、表２、表３の表現を提供することができる）。

【０２０６】

図１３は、本発明の実施形態による状態遷移表ｓｔｔａｂ２３０と、状態２５０に関連する量子化セットを指定する表ｓｅｔＩｄ２４０についての好ましい例を示す図である。Ｃスタイルのシンタックスで与えられた表は、表２に規定された表を表す。

【０２０７】

図１４は、本発明の実施形態による状態遷移表ｓｔｔａｂ２３０と、状態２５０に関連する量子化セットを指定する表ｓｅｔＩｄ２４０に関する好ましい例を示す図である。Ｃスタイルのシンタックスで与えられた表は、表３に規定された表を表している。

【０２０８】

別の実施形態では、０に等しいすべての量子化インデックス５６は、状態遷移および依存再構成プロセスから除外される。量子化インデックス５６が０に等しいか等しくないかの情報は、ニューラルネットワークパラメータ１３を０と０でないニューラルネットワークパラメータに分割するために使用されるだけである。依存スカラー量子化のための再構成プロセスは、ゼロでない量子化インデックス５６の順序付けされたセットにのみ適用される。０に等しい量子化インデックスに関連するすべてのニューラルネットワークパラメータは、単に０に等しく設定される。対応する擬似コードを図１５に示す。図１５は、本発明の実施形態による、０に等しい量子化インデックスが状態遷移および依存スカラー量子化から除外される、ニューラルネットワークパラメータレベルの代替再構成プロセスを示す擬似コードである。

【０２０９】

また、依存量子化における状態遷移は、図１６に示されるように、トレリス構造を用いて表現することも可能である。図１６は、本発明の実施形態によるトレリス構造としての依存スカラー量子化における状態遷移の例を示す図である。横軸は、再構成順序において異なるニューラルネットワークパラメータ１３を表している。縦軸は、依存量子化および再構成プロセスにおける異なる可能な状態２５０を表す。示された接続は、異なるニューラルネットワークパラメータに対する状態間の利用可能なパスを指定する。この図に示されるトレリスは、表２に指定される状態遷移に対応する。各状態２５０について、現在のニューラルネットワークパラメータ１３’に対する状態を、再構成順序において次のニューラルネットワークパラメータ１３に対する２つの可能な状態と接続する２つのパスが存在する。パスは、パス０とパス１とでラベル付けされ、この番号は、上で紹介したパス変数に対応する（好ましい実施形態では、そのパス変数は、量子化インデックスのパリティに等しい）。各パスは、量子化インデックスに対するサブセット（Ａ、Ｂ、Ｃ、またはＤ）を一意的に指定することに留意されたい。図１６では、サブセットは括弧で指定されている。初期状態（例えば状態０）が与えられると、トレリスを通るパスは、送信された量子化インデックス５６によって一意的に指定される。

【０２１０】

図１６の例では，状態（０，１，２，３）は次のような性質を持っている。

・状態０：前の量子化インデックスｌｅｖｅｌ［ｋ－１］はセット０の再構成レベルを指定し、現在の量子化インデックスｌｅｖｅｌ［ｋ］はセット０の再構成レベルを指定する。

・状態１：前の量子化インデックスｌｅｖｅｌ［ｋ－１］はセット０の再構成レベルを指定し、現在の量子化インデックスｌｅｖｅｌ［ｋ］はセット１の再構成レベルを指定する。

・状態２：前の量子化インデックスｌｅｖｅｌ［ｋ－１］はセット１の再構成レベルを指定し、現在の量子化インデックスｌｅｖｅｌ［ｋ］はセット０の再構成レベルを指定する。
・状態３：前の量子化インデックスｌｅｖｅｌ［ｋ－１］はセット１の再構成レベルを指定し、現在の量子化インデックスｌｅｖｅｌ［ｋ］はセット１の再構成レベルを指定する。

【０２１１】

トレリスは、いわゆる基本トレリスセルを連結したものである。このような基本トレリスセルについての一例を図１７に示す。図１７は、本発明の実施形態に係る基本トレリスセルの一例を示す図である。本発明は、４つの状態２５０を有するトレリスに限定されないことに留意されたい。他の実施形態では、トレリスは、より多くの状態２５０を有することができる。特に、２の整数乗を表す任意の数の状態が好適である。特に好ましい実施形態では、状態２５０の数は、例えば表３と同様に８に等しい。トレリスが２以上の状態２５０を有する場合でも、現在のニューラルネットワークパラメータ１３’のための各ノードは、典型的には、以前のニューラルネットワークパラメータ１３のための２つの状態および次のニューラルネットワークパラメータ１３の２つの状態と接続される。しかしながら、ノードが、以前のニューラルネットワークパラメータの２つ以上の状態、または次のニューラルネットワークパラメータの２つ以上の状態と接続されることも可能である。完全に接続されたトレリス（各状態２５０は、以前のニューラルネットワークパラメータ１３のすべての状態２５０および次のニューラルネットワークパラメータ１３のすべての状態２５０と接続されている）は、独立したスカラー量子化に対応するであろうことに留意されたい。

【０２１２】

好ましい実施形態では、（この決定を復号化器に送信するために、何らかのサイド情報レートを必要とするため）初期状態を自由に選択することはできない。その代わりに、初期状態は、予め定義された値に設定されるか、またはその値が他のシンタックス要素に基づいて導出されるかのいずれかである。この場合、第１のニューラルネットワークパラメータに対して、すべてのパスと状態２５０が利用できるわけではない。４状態のトレリスの例として、図１８は、初期状態が０に等しい場合のトレリス構造を示す。図１８は、本発明の実施形態による８個のニューラルネットワークパラメータの依存スカラー量子化のためのトレリス例を示す図である。第１の状態（左側）は、初期状態を表し、この例では０に等しく設定されている。

【０２１３】

４．４　エントロピー符号化

依存量子化によって得られた量子化インデックスは、エントロピー符号化方式によって符号化される。これには、任意のエントロピー符号化法が適用可能である。本発明の好ましい実施形態では、コンテキスト適応的２値算術符号化（ＣＡＢＡＣ）を用いた、第２．２節（符号化方法については第２．２．１節、復号化方法については第２．２．２節参照）によるエントロピー符号化方法が適用される。このために、例えば図５に示すように、量子化インデックスを絶対値として送信するために、まず非２値が一連の２値決定（いわゆるビン）上にマッピングされる（２値化）。

【０２１４】

ここで説明したどの概念も、３節における方法と関連する概念（特にコンテキストモデリング（ｃｏｎｔｅｘｔ　ｍｏｄｅｌｌｉｎｇ）に関するもの）と組み合わせることができることに留意する必要がある。

【０２１５】

４．４．１　依存スカラー量子化のコンテキストモデリング

依存スカラー量子化の主な態様は、ニューラルネットワークパラメータ１３のための許容される再構成レベルの異なるセット（量子化セットとも呼ばれる）が存在することである。現在のニューラルネットワークパラメータ１３’に対する量子化セットは、先行するニューラルネットワークパラメータに対する量子化インデックス５６の値に基づいて決定される。図１１の好ましい例を考え、２つの量子化セットを比較すると、ゼロに等しい再構

成レベルと隣接する再構成レベルとの間の距離は、セット０においてセット１よりも大きいことは明らかである。したがって、量子化インデックス５６が０に等しい確率は、セット０が使用される場合により大きく、セット１が使用される場合により小さくなる。好ましい実施形態では、この効果は、現在の量子化インデックスに使用される量子化セット（または状態）に基づいてコードワード表または確率モデルを切り替えることによって、エントロピー符号化で利用される。

【０２１６】

コードワード表または確率モデルの好適な切り替えのために、現在の量子化インデックス（または現在の量子化インデックスの対応する２値決定）をエントロピー復号化するときに、すべての先行する量子化インデックスのパス（使用する量子化セットのサブセットとの関連）が知られていなければならないことに注意されたい。したがって、ニューラルネットワークパラメータ１３が再構成順序で符号化されることが必要である。したがって、好ましい実施形態では、ニューラルネットワークパラメータ１３の符号化順序は、それらの再構成順序に等しい。その態様のほかに、量子化インデックス５６の任意の符号化／再構成順序が可能であり、例えば、２．２．１節で規定した順序は、他の任意の一意的に定義された順序である。

【０２１７】

言い換えれば、本発明による実施形態は、例えば、以前に符号化されたニューラルネットワークパラメータの量子化インデックスに追加的に依存する確率モデルを用いて、ニューラルネットワークパラメータを符号化するための装置を含む。

【０２１８】

それぞれ、本発明による実施形態は、例えば、ニューラルネットワークパラメータを復号化するための装置であって、以前に復号化されたニューラルネットワークパラメータの量子化インデックスに追加的に依存する確率モデルを使用する装置を含む。

【０２１９】

絶対レベルのためのビンの少なくとも一部は、典型的には、適応的な確率モデル（コンテキストとも呼ばれる）を使用して符号化される。本発明の好ましい実施形態では、１つ以上のビンの確率モデルは、対応するニューラルネットワークパラメータの量子化セット（または、より一般的には、対応する状態変数、例えば、表１～３のいずれかによる関係）に基づき選択される。選択された確率モデルは、既に送信された量子化インデックス５６の複数のパラメータまたは特性に依存することができるが、パラメータの１つは、符号化される量子化インデックスに適用される量子化セットまたは状態である。

【０２２０】

言い換えれば、実施形態による、例えばニューラルネットワークパラメータ１３を符号化するための装置は、現在のニューラルネットワークパラメータ１３’に対して選択された再構成レベルの状態またはセット４８に応じて、複数の確率モデルのうち確率モデルのサブセットを事前に選択し、以前に符号化されたニューラルネットワークパラメータの量子化インデックスに応じて１２１確率モデルのサブセットのうち現在のニューラルネットワークパラメータの確率モデルを選択するように構成され得る。

【０２２１】

実施形態による、例えばニューラルネットワークパラメータ１３を復号化するための装置は、現在のニューラルネットワークパラメータ１３’に対して選択された再構成レベルの状態またはセット４８に応じて、複数の確率モデルのうち確率モデルのサブセットを事前選択し、以前に復号化されたニューラルネットワークパラメータの量子化インデックスに応じて１２１、確率モデルのサブセットのうち現在のニューラルネットワークパラメータの確率モデルを選択するように構成されてもよい。

【０２２２】

例えば図９の文脈で説明したような発明的概念と組み合わせて、例えばニューラルネットワークパラメータ１３の符号化および／または復号化のための、本発明による実施形態は、第１の状態または再構成レベルセットに対して事前選択されたサブセットが、任意の

他の状態または再構成レベルセットに対して事前選択されたサブセットと非結合であるように、現在のニューラルネットワークパラメータ１３’に対して選択された再構成レベルの状態またはセット４８に応じて、複数の確率モデルの中から確率モデルのサブセットを事前選択するように構成される装置を含む。

【０２２３】

特に好ましい実施形態では、層の量子化インデックスを送信するためのシンタックスは、量子化インデックスが０に等しいか、または０に等しくないかを指定するビン、例えば前述のｓｉｇ＿ｆｌａｇを含む。このビンの符号化に用いる確率モデルは、２つ以上の確率モデルのセットの中から選択される。使用される確率モデルの選択は、対応する量子化インデックス５６に適用される量子化セット（即ち、再構成レベルのセット）に依存する。本発明の別の実施形態では、使用される確率モデルは、現在の状態変数に依存する（状態変数は、使用される量子化セットを意味する）。

【０２２４】

さらなる実施形態では、層の量子化インデックスを送信するためのシンタックスは、量子化インデックスがゼロより大きいか、ゼロより小さいかを指定するｂｉｎ、例えば前述のｓｉｇｎ＿ｆｌａｇを含む。すなわち、ｂｉｎは、量子化インデックスの符号を示す。使用される確率モデルの選択は、対応する量子化インデックスに適用される量子化セット（即ち、再構成レベルのセット）に依存する。別の実施形態では、使用される確率モデルは、現在の状態変数に依存する（状態変数は、使用される量子化セットを意味する）。

【０２２５】

さらなる実施形態では、量子化インデックスを送信するためのシンタックスは、量子化インデックスの絶対値（ニューラルネットワークのパラメータレベル）がＸより大きいかどうかを指定するビン、例えば前述のａｂｓ＿ｌｅｖｅｌ＿ｇｒｅａｔｅｒ＿Ｘ（詳細はセクション０を参照）を含んでいる。このビンの符号化に用いる確率モデルは、２つ以上の確率モデルのセットの中から選択される。使用される確率モデルの選択は、対応する量子化インデックス５６に適用される量子化セット（即ち、再構成レベルセット）に依存する。別の実施形態では、使用される確率モデルは、現在の状態変数に依存する（状態変数は、使用される量子化セットを意味する）。

【０２２６】

本明細書で論じる実施形態の１つの有利な態様は、ニューラルネットワークパラメータ１３の依存量子化がエントロピー符号化と組み合わされ、量子化インデックスの２値表現の１つ以上のビン（これは量子化レベルとも呼ばれる）に対する確率モデルの選択が、現在の量子化インデックスに対する量子化セット（許容再構成レベルのセット）または対応する状態変数に依存することである。量子化セット５２（または状態変数）は、符号化および再構成順序における先行するニューラルネットワークパラメータの量子化インデックス５６（または量子化インデックスを表すビンのサブセット）により与えられる。

【０２２７】

好ましい実施形態において、確率モデルの記述された選択は、以下のエントロピー符号化の側面のうちの１つ以上と組み合わされる。

・量子化インデックスの絶対値は、適応型確率モデルを用いて符号化される多数のビンと、適応的に符号化されたビンが既に完全に絶対値を指定していない場合、算術符号化エンジンのバイパスモードで符号化されるサフィックス部分（全ビンに対してｐｍｆ（例えば確率質量関数）（０．５、０．５）を有する非適応型確率モデル）と、からなる２値方式を用いて送信される。好ましい実施形態では、サフィックス部分に使用される２値化は、既に送信された量子化インデックスの値に依存する。

・量子化インデックスの絶対値に対する２値化は、量子化インデックスが０に等しくないかどうかを指定する適応的に符号化されたビンを含む。このビンの符号化に用いられる確率モデル（コンテキストと呼ばれる）は、候補確率モデルのセットの中から選択される。選択された候補確率モデルは、現在の量子化インデックス５６に対する量子化セット（許容再構成レベルのセット）または状態変数によって決定されるだけでなく、加えて、その

層に対する既に送信された量子化インデックスによっても決定される。好ましい実施形態では、量子化セット（または状態変数）は、利用可能な確率モデルのサブセット（コンテキストセットとも呼ばれる）を決定し、既に符号化された量子化インデックスの値は、このサブセット（コンテキストセット）内において使用される確率モデルを決定する。

【０２２８】

_実施形態では、コンテキストセット内の使用される確率モデルは、現在のニューラルネットワークパラメータの局所近傍における既に符号化された量子化インデックスの値、例えば、２．２．３で説明したようなテンプレートに基づいて決定される。以下では、局所近傍の量子化インデックスの値に基づいて導出され、その後、事前に決定されたコンテキストセットの確率モデルを選択するために使用することができるいくつかの例示的な量（ｍｅａｓｕｒｅ）をリストアップする。

・局所近傍で０に等しくない量子化インデックスの符号。

・局所近傍領域で０に等しくない量子化インデックスの数。この数は最大値にクリップ（短縮化）される可能性がある。

・局所近傍における量子化インデックスの絶対値の合計。この数値は最大値にクリップされる可能性がある。

・局所近傍における量子化インデックスの絶対値の合計と、局所近傍における０に等しくない量子化インデックスの数との差。この数値は最大値にクリップされる可能性がある。

【０２２９】

_言い換えれば、本発明による実施形態は、例えば、現在のニューラルネットワークパラメータが関連する部分に隣接するニューラルネットワークの部分に関連する以前に符号化されたニューラルネットワークパラメータの量子化インデックスの特性に応じて確率モデルのサブセットから現在のニューラルネットワークパラメータの確率モデルを選択するように構成されたニューラルネットワークパラメータの符号化のための装置を含み、特性は、以下のもののうち１つまたは複数を含む。

現在のニューラルネットワークパラメータが関連する部分に隣接するニューラルネットワークの部分に関連する、以前に符号化されたニューラルネットワークパラメータのゼロでない量子化インデックスの符号。

現在のニューラルネットワークパラメータが関連する部分に隣接するニューラルネットワークの部分に関連する、以前に符号化されたニューラルネットワークパラメータの量子化インデックスの数であって、ゼロでない数

現在のニューラルネットワークパラメータが関連する部分に隣接するニューラルネットワークの部分に関連する、以前に符号化されたニューラルネットワークパラメータの量子化インデックスの絶対値の合計値

現在のニューラルネットワークパラメータが関連する部分に隣接するニューラルネットワークの部分に関連する、以前に符号化されたニューラルネットワークパラメータの量子化インデックスの絶対値の合計と、

現在のニューラルネットワークパラメータが関連する部分に隣接するニューラルネットワークの部分に関連する、以前に符号化されたニューラルネットワークパラメータの量子化インデックスの数であって、ゼロでない数と、

の間の差。

【０２３０】

それぞれ、本発明による実施形態は、例えば、ニューラルネットワークパラメータの復号化のための装置である。装置は、確率モデルのサブセットのうち、現在のニューラルネットワークパラメータのための確率モデルを、現在のニューラルネットワークパラメータが関連する部分に隣接するニューラルネットワークの部分に関連する以前に復号化されたニューラルネットワークパラメータの量子化インデックスの特性に応じて選択するように構成されている。特性は、以下のもののうち１つまたは複数を含む。

現在のニューラルネットワークパラメータが関連する部分に隣接するニューラルネットワークの部分に関連する、以前に復号化されたニューラルネットワークパラメータのゼ

ロでない量子化インデックスの符号。
　現在のニューラルネットワークパラメータが関連する部分に隣接するニューラルネットワークの部分に関連する、以前に復号化されたニューラルネットワークパラメータの量子化インデックスの数であって、ゼロでない数。
　現在のニューラルネットワークパラメータが関連する部分に隣接するニューラルネットワークの部分に関連する、以前に復号化されたニューラルネットワークパラメータの量子化インデックスの絶対値の合計値。
　現在のニューラルネットワークパラメータが関連する部分に隣接するニューラルネットワークの部分に関連する、以前に復号化されたニューラルネットワークパラメータの量子化インデックスの絶対値の合計と、
　現在のニューラルネットワークパラメータが関連する部分に隣接するニューラルネットワークの部分に関連する、以前に復号化されたニューラルネットワークパラメータの量子化インデックスの数であって、ゼロでない数と、
の差。
【０２３１】
・量子化インデックスの絶対値に対する２値化は、量子化インデックスの絶対値がＸより大きいかどうかを指定する適応的に符号化されたビン、例えばａｂｓ＿ｌｅｖｅｌ＿ｇｒｅａｔｅｒ＿Ｘを含む。これらのビンを符号化するために使用される確率モデル（コンテキストと呼ばれる）は、候補確率モデルのセットの中から選択される。選択された確率モデルは、現在の量子化インデックスに対する量子化セット（許容再構成レベルのセット）または状態変数によって決定されるだけでなく、さらに、例えば前述のようなテンプレートを用いて、層に対する既に送信された量子化インデックスによって決定される。好ましい実施形態では、量子化セット（または状態変数）は、利用可能な確率モデルのサブセット（コンテキストセットとも呼ばれる）を決定し、既に符号化された量子化インデックスのデータは、このサブセット（コンテキストセット）内部で使用される確率モデルを決定する、例えば他の言葉で言えば、決定に用いることが可能である。確率モデルの選択には、上述した方法（量子化インデックスが０に等しくないかどうかを指定するビンの場合）のいずれかを使用することができる。
【０２３２】
　さらに、本発明による装置は、以前に符号化されたニューラルネットワークパラメータ１３が、現在のニューラルネットワークパラメータ１３’と同じニューラルネットワーク層に関連するように、以前に符号化されたニューラルネットワークパラメータ１３を位置づけるように構成されることができる。
【０２３３】
　さらに、本発明による例えばニューラルネットワークパラメータを符号化するための装置は、１つ以上の以前に符号化されたニューラルネットワークパラメータが、現在のニューラルネットワークパラメータが参照するニューロン相互接続１１が関連するニューロン１０ｃまたは該ニューロンに隣接する別のニューロンから出現するニューロン相互接続、またはこれらのニューロンに向かうニューロン相互接続に関連するように、以前に符号化されたニューラルネットワークパラメータのうちの１つ以上のパラメータを位置付けるように構成されることができる。
【０２３４】
　さらなる実施形態による装置は、量子化インデックスを２値化したものの１つ以上のリーディングビン（ｌｅａｄｉｎｇ　ｂｉｎ）について以前に符号化されたニューラルネットワークパラメータに対応する確率モデルを使用することによって、及び、１つ以上のリーディンビンに続く量子化インデックスを２値化したものの等確率バイパスモードサフィックスビン（ｅｑｕｉ－ｐｒｏｂａｂｌｅ　ｂｙｐａｓｓ　ｍｏｄｅ　ｓｕｆｆｉｘ　ｂｉｎｓ）を使用することによって、２値算術符号化を使用して、データストリーム１４へ現在のニューラルネットワークパラメータ１３’に対する量子化インデックス５６を符号化するよう構成されることができる。

【０２３５】
量子化インデックスを２値化したもののサフィックスビンは、量子化インデックスの値を２値化するためのサフィックス２値化の２値化コードのビンを表し、その絶対値が１つ以上のリーディングビンで表現できる最大絶対値を超えている場合がある。したがって、本発明の実施形態による装置は、以前に符号化されたニューラルネットワークパラメータ１３の量子化インデックス５６に応じて、サフィックス２値化を選択するように構成されることができる。
【０２３６】
それぞれ、本発明による、例えばニューラルネットワークパラメータを復号化するための装置は、以前に復号されたニューラルネットワークパラメータが、現在のニューラルネットワークパラメータ１３’と同じニューラルネットワーク層に関連するように、以前に復号化されたニューラルネットワークパラメータ１３を位置付けるように構成されてもよい。
【０２３７】
さらなる実施形態によれば、例えば本発明によるニューラルネットワークパラメータを復号化するための装置は、１つ以上の以前に復号化されたニューラルネットワークパラメータが、現在のニューラルネットワークパラメータが参照するニューロン相互接続が関連するニューロン１０ｃまたは該ニューロンに隣接する別のニューロンから出現するニューロン相互接続、またはこれらのニューロンに向かうニューロン相互接続に関連するように、以前に復号化されたニューラルネットワークパラメータ１３のうちの１つ以上のパラメータを位置付けるように構成されることができる。
【０２３８】
さらなる実施形態による装置は、量子化インデックスを２値化したものの１つ以上のリーディングビンについて以前に復号化されたニューラルネットワークパラメータに対応する確率モデルを使用することによって、及び、１つ以上のリーディングビンに続く量子化インデックスを２値化したものの等確率バイパスモードサフィックスビンを使用することによって、２値算術符号化を使用してデータストリーム１４から現在のニューラルネットワークパラメータ１３’に対する量子化インデックス５６を復号化するように構成されることができる。
【０２３９】
量子化インデックスの２値化のサフィックスビンは、量子化インデックスの値を２値化するためのサフィックス２値化の２値化符号のビンを表し、その絶対値が１つ以上のリーディングビンで表現可能な最大絶対値を超える。したがって、実施形態に従う装置は、以前に復号化されたニューラルネットワークパラメータの量子化インデックスに応じてサフィックス２値化を選択するように構成されることができる。
【０２４０】
４．５　符号化の方法例
歪み（再構成品質）とビットレートのトレードオフが非常に良いビットストリームを得るためには、ラグランジュコスト量（Ｌａｇｒａｎｇｉａｎ　ｃｏｓｔ　ｍｅａｓｕｒｅ）が最小となるように量子化インデックスを選択する必要がある。

$$D+\lambda\cdot R=\sum_k D_k+\lambda\cdot R_k=\sum_k \alpha_k\cdot(t_k-t'_k)^2+\lambda\cdot R(q_k|q_{k-1},q_{k-2},\cdots)$$

独立スカラー量子化では、このような量子化アルゴリズム（レート歪み最適化量子化、ｒａｔｅ－ｄｉｓｔｏｒｔｉｏｎ　ｏｐｔｉｍｉｚｅｄ　ｑｕａｎｔｉｚａｔｉｏｎまたはＲＤＯＱと呼ばれる）が２．１．１節で議論された。しかし、独立スカラー量子化と比較して、さらなる難題がある。再構成されたニューラルネットワークのパラメータｔ$_k$’、したがって、その歪み$D_k=|t_k-t'_k|$（または$D_{k,MSE}=(t_k-t'_k)^2$）は、関連する量子化インデックスｑ$_k$５６に依存するだけでなく、符号化順序において先行する量子化インデックスの値にも依存する。

【０２４１】

しかし、４．３．３節で説明したように、ニューラルネットワークのパラメータ１３間の依存関係は、トレリス構造を用いて表現することができる。さらなる説明のために、図１１に与えられた好ましい実施形態を例として用いる。８個のニューラルネットワークパラメータのセットの例に対するトレリス構造を図１９に示す。図１９は、本発明の実施形態による、コスト量（ラグランジュコスト量Ｄ＋λ・Ｒなど）を最小化する量子化インデックスのシーケンス（またはブロック）を決定するために利用することができるトレリス構造の例を示している。トレリス構造は、４つの状態を有する依存量子化の好ましい例を表している（図１８参照）。トレリスは、８つのニューラルネットワークパラメータ（または量子化インデックス）に対して示されている。最初の状態（一番左）は初期状態を表し、０に等しいと仮定する。トレリスを通るパス（左から右へ）は、量子化インデックス５６に対して可能な状態遷移を表す。２つのノード間の各接続は、特定のサブセット（Ａ、Ｂ、Ｃ、Ｄ）の量子化インデックスを表していることに注意が必要である。各サブセット（Ａ、Ｂ、Ｃ、Ｄ）から量子化インデックスq_k５６を選び、対応するレート歪コストを２つのトレリスノード間の関連する接続に割り当てた場合、以下のようになる。

$$J_k = D_k(q_k|q_{k-1}, q_{k-2}, \cdots) + \lambda \cdot R_k(q_k|q_{k-1}, q_{k-2}, \cdots)$$

全体のレート歪みコストＤ＋λ・Ｒを最小化する量子化インデックスのベクトル／ブロックを決定する問題は、トレリスを通る最小コストのパスを見つけることと同等である（図１９において左から右へ）。エントロピー符号化におけるいくつかの依存性を無視すれば、この最小化問題は、よく知られたビタビアルゴリズム（Ｖｉｔｅｒｂｉ　ａｌｇｏｒｉｔｈｍ）を用いて解くことができる。

【０２４２】

言い換えれば、本発明による実施形態は、選択および／または量子化を実行するためにビタビアルゴリズムおよびレート歪みコスト量を使用するように構成された装置を含んでいる。

【０２４３】

層のための適切な量子化インデックスを選択するための例示的な符号化アルゴリズムは、以下の主要なステップで構成され得る。

１．初期状態でのレート歪みコストを０に設定する。

２．符号化順序におけるすべてのニューラルネットワークのパラメータ１３に対して、次のようにする。

ａ．各サブセットＡ、Ｂ、Ｃ、Ｄについて、与えられたオリジナルのニューラルネットワークパラメータ１３に対する歪みを最小化する量子化インデックス５６を決定する。

ｂ．現在のニューラルネットワークパラメータ１３’に対する全てのトレリスノード（０、１、２、３）に対して、以下を行う。

ｉ　先行するニューラルネットワークパラメータ１３の状態と現在の状態とを結ぶ２つのパスのレート歪みコストを計算する。コストは、先行する状態に対するコストと、$D_k + \lambda \cdot R_k$の合計として与えられ、ここで、D_kとR_kは、考慮される接続に関連するサブセット（Ａ、Ｂ、Ｃ、Ｄ）の量子化インデックスを選択するための歪みとレートを表す。

ｉｉ　計算されたコストの最小値を現在のノードに割り当て、最小コストのパスを表さない以前のニューラルネットワークパラメータ１３の状態への接続を取り除く。

注：このステップの後、現在のニューラルネットワークパラメータ１３’に対するすべてのノードは、先行するニューラルネットワークパラメータ１３に対する任意のノードへの単一の接続を有する。

３．（符号化順序における最後のパラメータについての）４つの最終ノードのコストを比較し、コストが最小のノードを選択する。このノードは、トレリスを通るユニークなパスに関連していることに注意（他のすべての接続は以前のステップで取り除かれた）。
４．（最終ノードで指定される）選択したパスを逆順にたどり、トレリスノード間の接続に関連する量子化インデックス５６を収集する。

【０２４４】

ビタビアルゴリズムに基づく量子化インデックス５６の決定は、独立スカラー量子化のためのレート歪み最適化量子化（ＲＤＯＱ）よりも実質的に複雑ではないことに留意されたい。それにもかかわらず、依存量子化のためのより単純な符号化アルゴリズムも存在する。例えば、予め定義された初期状態（または量子化セット）から開始して、量子化インデックス５６は、現在の量子化インデックスの影響のみを考慮する任意のコスト量を最小化することによって、符号化／再構成の順序で決定されることができる。現在のパラメータに対する決定された量子化インデックス（および先行するすべての量子化インデックス）が与えられると、次のニューラルネットワークパラメータ１３に対する量子化セットは既知である。そして、したがって、このアルゴリズムは、符号化順序ですべてのニューラルネットワークパラメータに適用することができる。

【０２４５】

以下の実施形態による方法は、図２０、図２１、図２２、図２３に示されている。

【０２４６】

図２０は、ニューラルネットワークを定義するニューラルネットワークパラメータをデータストリームから復号化するための方法４００のブロック図である。方法４００は、現在のニューラルネットワークパラメータについて、以前のニューラルネットワークパラメータについてデータストリームから復号化された量子化インデックスに応じて、複数の再構成レベルセットから再構成レベルセットを選択５４するステップによって、及び、データストリームから現在のニューラルネットワークパラメータの量子化インデックス４２０を復号化するステップによって、ここで量子化インデックスが現在のニューラルネットワークパラメータのための再構成レベルの選択されたセットのうちの１つの再構成レベルを示し、及び、現在のニューラルネットワークパラメータのための量子化インデックスによって示される再構成レベルの選択されたセットのうちの１つの再構成レベル上に現在のニューラルネットワークパラメータを逆量子化６２するステップによって、ニューラルネットワークパラメータを順次復号化することを含む。

【０２４７】

図２１は、データストリームからニューラルネットワークを定義するニューラルネットワークパラメータを符号化するための方法５００のブロック図である。方法５００は、現在のニューラルネットワークパラメータについて、以前に符号化されたニューラルネットワークパラメータについてデータストリームに符号化された量子化インデックスに応じて複数の再構成レベルセットの中から再構成レベルセットを選択５４するステップによって、及び、現在のニューラルネットワークパラメータを、選択された再構成レベルのセットのうち１つの再構成レベルに量子化６４するステップによって、及び、現在のニューラルネットワークパラメータの量子化インデックスが量子化される１つの再構成レベルを示す現在のニューラルネットワークパラメータの量子化インデックスをデータストリームに符号化５３０するステップによって、ニューラルネットワークパラメータを連続的に符号化することを含む。

【０２４８】

図２２は、本発明の実施形態による、ニューラルネットワークを定義するニューラルネットワークパラメータを再構築する方法のブロック図である。方法６００は、ニューラルネットワークパラメータごとに、第１の再構成層のニューラルネットワークパラメータ値をもたらすために、第１の再構成層についての第１のニューラルネットワークパラメータ６１０を導出するステップを含む。方法６００はさらに、データストリームから第２の再構成層の第２のニューラルネットワークパラメータを復号化６２０（例えば図６の矢印３

１２で示すように）して、ニューラルネットワークパラメータごとに、第２の再構成層のニューラルネットワークパラメータ値をもたらすステップ、及び、ニューラルネットワークパラメータごとに、第１の再構成層のニューラルネットワークパラメータ値および第２の再構成層のニューラルネットワークパラメータ値を組み合わせることによって、ニューラルネットワークパラメータを再構成する６３０（例えば図６の矢印３１４で示す）ステップを含んでいる。

【０２４９】

図２３は、本発明の実施形態による、ニューラルネットワークを定義するニューラルネットワークパラメータを符号化するための方法のブロック図である。方法７００は、ニューラルネットワークパラメータごとに、第１の再構成層のニューラルネットワークパラメータ値を含む第１の再構成層の第１のニューラルネットワークパラメータを使用し、第２の再構成層の第２のニューラルネットワークパラメータをデータストリームに（例えば、図中に矢印３２２で示すように）符号化７１０するステップを含み、第２の再構成層は、ニューラルネットワークパラメータごとに、第２の再構成層ニューラルネットワークパラメータ値を含み、ニューラルネットワークパラメータは、ニューラルネットワークパラメータごとに、第１の再構成層ニューラルネットワークパラメータ値と第２の再構成層ニューラルネットワークパラメータ値とを組み合わせることによって再構成可能である。

【０２５０】

以下では、本発明による追加の実施形態が示される。

quant_tensor(dimensions, maxNumNoRem, entryPointOffset) {	
stateId = 0	997
bitPointer = get_bit_pointer()	998
lastOffset = 0	999
for(i = 0; i < Prod(dimensions); i++) {	1000
idx = TensorIndex(dimensions, i, scan_order)	1001
if(entryPointOffset != -1 && GetEntryPointIdx(dimensions, i, scan_order) != -1) {	1002
lvlCurrRange = 256	1003
j = entryPointOffset + GetEntryPointIdx(dimensions, i, scan_order)	1004
lvlOffset = cabac_offset_list[j]	1005
if(dq_flag)	1006
stateId = dq_state_list[j]	1007
set_bit_pointer(bitPointer + lastOffset + BitOffsetList[j])	1008
lastOffset = BitOffsetList[j]	1009
Invoke initialisation process for probability estimation parameters	1010
}	1011
int_param(idx, maxNumNoRem, stateId)	1012
if(dq_flag) {	1013
nextSt = StateTransTab[stateId][QuantParam[idx] & 1]	1014
if(QuantParam[idx] != 0) {	1015
QuantParam[idx] = QuantParam[idx] << 1	1016
if(QuantParam[idx] < 0)	1017
QuantParam[idx] += stateId & 1	1018
else	1019
QuantParam[idx] += - (stateId & 1)	1020
}	1021
stateId = nextSt	1022
}	
}	
}	

【０２５１】

例えば１０１４行目に示す２次元整数アレイＳｔａｔｅＴｒａｎｓＴａｂ［］［］は、依存スカラー量子化の状態遷移表を指定するもので、次のようになる。

【０２５２】

StateTransTab[][] = { {0, 2}, {7, 5}, {1, 3}, {6, 4}, {2, 0}, {5, 7}, {3, 1}, {4, 6} }

int_param(i, maxNumNoRem, stateId) {	
QuantParam[i] = 0	5997
sig_flag	5998
if(sig_flag) {	5999
QuantParam[i]++	6000
sign_flag	6001
j = -1	6002
do {	6003
j++	6004
abs_level_greater_x[j]	6005
QuantParam[i] += abs_level_greater_x[j]	6006
} while(abs_level_greater_x[j] == 1 && j < maxNumNoRem)	6007
if(j == maxNumNoRem) {	6008
RemBits = 0	6009
j = -1	6010
do {	6011
j++	6012
abs_level_greater_x2[j]	6013
if(abs_level_greater_x2[j]) {	6014
RemBits++	6015
QuantParam[i] += 1 << RemBits	6016
}	6017
} while(abs_level_greater_x2[j] && j < 30)	6018
abs_remainder	6019
QuantParam[i] += abs_remainder	6020
}	6021
QuantParam[i] = sign_flag ? -QuantParam[i] : QuantParam[i]	6022
}	
}	

【０２５３】

このプロセスへの入力は次の通りである。

・復号化されるテンソルの次元を指定する変数　ｔｅｎｓｏｒＤｉｍｓ。

・復号化のためのエントリポイントが存在するかどうか、およびエントリポイントが存在する場合エントリポイントオフセットを示す変数　ｅｎｔｒｙＰｏｉｎｔＯｆｆｓｅｔ。

・コードブックの有無と、コードブックが適用されるかどうか、及びコードブックが適用される場合はどのコードブックを使用するかを示す変数　ｃｏｄｅｂｏｏｋＩｄ。

このプロセスの出力は、ＴＥＮＳＯＲ＿ＦＬＯＡＴ　型の変数ｒｅｃＰａｒａｍであり、次元はｔｅｎｓｏｒＤｉｍｓと等しい。

【０２５４】

変数ｓｔｅｐｓｉｚｅは以下のように導出される。

```
3001   mul = (1 << QpDensity) + ( (qp_value + QuantizationParameter) & ( ( 1
 << QpDensity ) − 1 ) )
3002   shift = (qp_value + QuantizationParameter) >> QpDensity
3003   stepSize = mul * 2shift − QpDensity
Variable recParam is updated as follows:
4001   recParam = recParam * stepSize
```

注－上記の計算から、ｒｅｃＰａｒａｍは常に２進小数で表すことができる。

【０２５５】

シンタックス要素ｓｉｇ＿ｆｌａｇに対して、使用するコンテキストまたは確率推定を示すｃｔｘＩｎｃの導出プロセスについて。

【０２５６】

このプロセスへの入力は、現在のｓｉｇ＿ｆｌａｇの前に復号化されたｓｉｇ＿ｆｌａｇ、状態値ｓｔａｔｅＩｄおよび関連するｓｉｇｎ＿ｆｌａｇ（存在する場合）である。現在のｓｉｇ＿ｆｌａｇの前に復号化されたｓｉｇ＿ｆｌａｇがない場合、それは０と見なされる。また、以前に復号化されたｓｉｇ＿ｆｌａｇに関連するｓｉｇｎ＿ｆｌａｇが復号化されていない場合、それは０と見なされる。

【０２５７】

このプロセスの出力は、変数ｃｔｘＩｎｃである。

変数ｃｔｘＩｎｃは、以下のように導出される。

・ｓｉｇ＿ｆｌａｇが０であれば、ｃｔｘＩｎｃにはｓｔａｔｅＩｄ＊３が設定される。

・そうでなければ、ｓｉｇｎ＿ｆｌａｇが０に等しい場合、ｃｔｘＩｎｃはｓｔａｔｅＩｄ＊３＋１に設定される。

・そうでなければ、ｃｔｘＩｎｃはｓｔａｔｅＩｄ＊３＋２に設定される。

【０２５８】

上記の例は、ニューラルネットワークパラメータ１３をデータストリーム１４へ／から符号化／復号化する概念を示し、ニューラルネットワークパラメータ１３は、ニューラルネットワーク１０のニューロン相互接続１１の重み、例えば重みテンソルの重みに関連し得る。ニューラルネットワークパラメータ１３の復号化／符号化は、順次行われる。テンソルの次元ごとの重みの数の積と同数の重みでテンソルの重みを循環するｆｏｒ－ｎｅｘｔループ１０００を参照されたい。重みは所定の順序ＴｅｎｓｏｒＩｎｄｅｘ（ｄｉｍｅｎｓｉｏｎｓ，ｉ，ｓｃａｎ＿ｏｒｄｅｒ）でスキャンされる。現在のニューラルネットワークパラメータｉｄｘ１３’について、２つの再構成レベルセット５２のうち再構成レベルセットが、１０１８および１０２０において、以前のニューラルネットワークパラメータのデータストリームから復号化された量子化インデックス５８に基づいて連続的に更新される量子化状態ｓｔａｔｅＩｄに応じて選択される。特に、現在のニューラルネットワークパラメータｉｄｘに対する量子化インデックスが１０１２でデータストリームから復号化され、量子化インデックスは、現在のニューラルネットワークパラメータ１３’に対する選択された再構成レベルセットのうち１つの再構成レベルを示している。２つの再構成レベルセットは、１０１６での複製と、それに続く１０１８および１０２０での量子化状態インデックスに応じた１またはマイナス１の加算によって定義される。ここで、１０１８及び１０２０において、現在のニューラルネットワークパラメータ１３’は、現在のニューラルネットワークパラメータ１３’に対する量子化インデックスＱｕａｎｔＰａｒａｍ［ｉｄｘ］によって示される選択された再構成レベルセットのうちの１つの再構成レベル上に実際に逆量子化される。ステップサイズｓｔｅｐＳｉｚｅは、３００１～３００３で再

構成レベルセットをパラメータ化するために使用される。この所定の量子化ステップサイズｓｔｅｐＳｉｚｅに関する情報は、シンタックス要素ｑｐ＿ｖａｌｕｅを介してデータストリームから導出される。後者は、それぞれテンソル全体またはＮＮ層全体について、あるいはＮＮ全体についてデータストリームに符号化されるかもしれない。すなわち、ニューラルネットワーク１０は、１つ以上のＮＮ層１０ａ、１０ｂを含むことができ、各ＮＮ層について、データストリーム１４からそれぞれのＮＮ層について所定の量子化ステップサイズ（ＱＰ）の情報を導出し、その後、それぞれのＮＮ層について、それぞれのＮＮ層に属するニューラルネットワークのパラメータ１３を逆量子化するために用いられるように、それぞれのＮＮ層について導出した所定の量子化ステップサイズを用いて複数の再構成レベルセットのパラメータ化を行うようにしても良い。

【０２５９】

ｓｔａｔｅＩｄ＝０に対する第１の再構成レベルセットは、ここではゼロおよび所定の量子化ステップサイズの偶数倍を含み、１０１８および１０２０で分かるように、ｓｔａｔｅＩｄ＝１に対する第２の再構成レベルセットは、ゼロおよび所定の量子化ステップサイズ（ＱＰ）の奇数倍を含む。各ニューラルネットワークパラメータ１３について、１０１５～１０２１で、それぞれのニューラルネットワークパラメータ１３について選択された再構成レベルセットとそれぞれのニューラルネットワークパラメータについてのエントロピー復号化量子化インデックスＱｕａｎｔＰａｒａｍ［ｉｄｘ］に応じて中間整数値ＱｕａｎｔＰａｒａｍ［ｉｄｘ］（ＩＶ）を導き、次に、各ニューラルネットワークパラメータについて、それぞれのニューラルネットワークパラメータの中間値に、４００１でのそれぞれのニューラルネットワークパラメータの所定の量子化ステップサイズが乗算される。

【０２６０】

２つの再構成レベルセット（例えばセット０、セット１）のうちの再構成レベルセットの、現在のニューラルネットワークパラメータ１３’に対する選択は、１０１４に示すように、以前に復号化されたニューラルネットワークパラメータのデータストリームから復号化された量子化インデックスのＬＳＢ部分に応じて行われ、遷移表が、ｓｔａｔＩｄが既に復号化された量子化インデックス５６の過去のシーケンスに依存するようにＱｕａｎｔＰａｒａｍ［ｉｄｘ］のＬＳＢに応じてｓｔａｔｅＩｄから次の量子化状態ｎｅｘｔＳｔへ遷移される。したがって、状態遷移は、以前に復号化されたニューラルネットワークパラメータのデータストリームから復号化された量子化インデックス５６の２値関数の結果、すなわちそのパリティに依存する。言い換えれば、現在のニューラルネットワークパラメータについて、複数の再構成レベルセットのうちの再構成レベルセットの選択は、現在のニューラルネットワークパラメータについて、１０１８及び１０２０において、現在のニューラルネットワークパラメータに関連付けられた状態ｓｔａｔＩｄに応じて、複数の再構成レベルセットのうちの再構成レベルセットを決定することによって、及び、直前のニューラルネットワークパラメータ、すなわちこれまでｓｔａｔｅＩｄが決定されていたパラメータについてデータストリームから復号化された量子化インデックスに応じて、次に符号化／復号化されるＮＮパラメータとは限らないが、次にｓｔａｔｅＩｄが決定されるべき後続のニューラルネットワークパラメータについて１０１４でｓｔａｔｅＩｄを更新することによって、状態遷移プロセスにより行われる。例えば、ここでは、現在のニューラルネットワークパラメータは、次に符号化／復号化されるべきＮＮパラメータのｓｔａｔｅＩｄをもたらすための更新のために使用される。１０１４における更新は、直前の（現在の）ニューラルネットワークパラメータのデータストリームから復号化された量子化インデックスの２値関数、すなわちそのパリティを使用して行われる。状態遷移プロセスは、８つの可能な状態間を遷移するように構成されている。遷移は、表ＳｔａｔｅＴｒａｎｓＴａｂ［］［］を介して行われる。状態遷移プロセスにおいて、遷移はこれら８つの可能な状態の間で行われ、現在のニューラルネットワークパラメータについて、１０１８および１０２０における、現在のニューラルネットワークパラメータに関連する状態ｓｔａｔｅＩｄに応じた量子化セットのうちの再構成レベルのセットの決定することで、状

態が偶数の可能な状態の前半、すなわち奇数の状態に属する場合には２つの再構成レベルセットのうちの第１の再構成レベルセットが決定され、状態が偶数の可能な状態の後半の状態、すなわち偶数の状態に属する場合には、２つの再構成レベルセットのうちの第２の再構成レベルセットが決定される。状態ｓｔａｔＩｄの更新は、遷移表ＳｔａｔｅＴｒａｎｓＴａｂ［］［］により行われる。遷移表は、直前の（現在の）ニューラルネットワークパラメータのデータストリームから復号化された状態ｓｔａｔＩＤと量子化インデックス（５８）のパリティ、ＱｕａｎｔＰａｒａｍ［ｉｄｘ］＆１の組み合わせを、後続のニューラルネットワークパラメータに関連する別の状態にマッピングする。

【０２６１】

現在のニューラルネットワークパラメータの量子化インデックスは、現在のニューラルネットワークパラメータに対して選択された再構成レベルセット、より正確には量子化状態ｓｔａｔｅＩｄ、すなわち現在のニューラルネットワークパラメータ１３’に対する状態に依存する確率モデルを用いて、算術符号化を用いてデータストリームに符号化し、及びデータストリークから復号化する。１０１２の関数ｉｎｔ＿ｐａｒａｍを呼び出すときの第３のパラメータを参照されたい。特に、現在のニューラルネットワークパラメータに対する量子化インデックスは、量子化インデックスを２値化したものの少なくとも１つのビンに対する現在のニューラルネットワークパラメータについての状態に対応する確率モデルを用いて、２値算術符号化／復号化を用いてデータストリームに符号化され、データストリームから復号化されることができる。ここで、２値化のｓｉｇ＿ｆｌａｇ、ｓｉｇｎ＿ｆｌａｇ（オプション）、ａｂｓ＿ｌｅｖｅｌ＿ｇｒｅａｔｅｒ＿ｘ［ｊ］、ａｂｓ＿ｌｅｖｅｌ＿ｇｒｅａｔｅｒ＿ｘ２［ｊ］、ａｂｓ＿ｒｅｍａｉｎｄｅｒうちビンｓｉｇ＿ｆｌａｇは、現在のニューラルネットワークパラメータの量子化インデックス（５６）がゼロに等しいか否かを示す有意性ビンである。確率モデルの依存性は、依存性を用いたニューラルネットワークパラメータのコンテキストセットのうちコンテキストを選択することを含み、各コンテキストは、所定の確率モデルが関連づけられるように構成される。ここで、ｓｉｇ＿ｆｌａｇのコンテキストは、それぞれが２値確率モデルに関連付けられているコンテキストのリストからコンテキストをインデックス化するためのインデックスのインクリメンターとしてｃｔｘＩｎｃを使用して選択される。モデルは、コンテキストに関連付けられたビンを使用して更新されることができる。すなわち、それぞれのコンテキストに関連付けられた所定の確率モデルは、それぞれのコンテキストを用いて算術符号化された量子化インデックスに基づいて更新されることができる。

【０２６２】

（注）ｓｉｇ＿ｆｌａｇの確率モデルは、さらに、以前に復号化されたニューラルネットワークパラメータの量子化インデックス、すなわち以前に復号化されたニューラルネットワークパラメータのｓｉｇ＿ｆｌａｇ、およびそのｓｉｇｎ＿ｆｌａｇ－その符号を示すもの－に依存することに注意されたい。より正確には、状態ｓｔａｔｅＩｄに応じて、複数の確率モデルのうち、すなわちコンテキストインクリメンター状態０．．．２３のうち、確率モデルのサブセットが予め選択され、すなわち｛０．．．２３｝のうち連続する３つのコンテキストを含むその８つ、ｓｉｇ＿ｆｌａｇに対する確率モデルのサブセットのうち、現在のニューラルネットワークパラメータの確率モデルを、以前に復号されたニューラルネットワークパラメータの量子化インデックスに応じて（１２１）、すなわち以前のＮＮパラメータのｓｉｇ＿ｆｌａｇとｓｉｇｎ＿ｆｌａｇに基づいて選択する。ｓｔａｔｅＩＤの最初の値に対して事前に選択されたサブセットは、ｓｔａｔｅＩＤの他の値に対して事前に選択されたサブセットと不一致である。ｓｉｇ＿ｆｌａｇとｓｇｉｎ＿ｆｌａｇが使用される以前のＮＮパラメータは、現在のニューラルネットワークパラメータが関連する部分に隣接するニューラルネットワークの部分に関連する。

【０２６３】

複数の実施形態が上述されてきた。実施形態の態様および特徴は、個々にまたは組み合わせて使用され得ることに留意されたい。さらに、本発明の第１および第２の側面による実施形態の態様および特徴は、組み合わせて使用されてもよい。

【０２６４】
さらなる実施形態は、装置を構成し、ニューラルネットワークパラメータは、ニューラルネットワーク１０が表される使用する再構成層のうちの１つの再構成層、例えばエンハンスメント層に関連する。装置は、ニューラルネットワークパラメータ単位で、対応する、例えば共通のニューロン相互接続に関連するもの、または率直に言って、異なる表現層におけるＮＮ層の行列表現に併置されるもの、１つ以上の別の再構成層のニューラルネットワークパラメータと組み合わせることによって再構成可能であるように、ニューラルネットワークが構成されてもよい。
【０２６５】
例えばこの実施形態で説明したように、本発明の第１および第２の側面の特徴および態様は、組み合わされてもよい。第２の側面による従属請求項の任意の特徴は、さらなる実施形態をもたらすために、ここにも移転可能であるものとする。
【０２６６】
さらに、本発明の態様による装置は、現在のニューラルネットワークパラメータ１３’に対する量子化インデックス５６を、現在のニューラルネットワークパラメータに対応する対応するニューラルネットワークパラメータに対応する確率モデルを用いた算術符号化によりデータストリーム１４の中に符号化するように構成されてもよい。
【０２６７】
それぞれ、さらなる実施形態は、装置を構成し、ニューラルネットワークパラメータは、ニューラルネットワーク１０が表現される再構成層のうちの１つの再構成層、例えばエンハンスメント層に関連するものである。装置は、ニューラルネットワークパラメータ単位で、ニューラルネットワークパラメータを、対応する例えば共通のニューロン相互接続に関連するもの、率直に言えば、異なる表現層におけるＮＮ層の行列表現において併置されるもの、１つ以上の別の再構成層のニューラルネットワークパラメータと組み合わせることによって、ニューラルネットワークを再構成するように構成されてもよい。
【０２６８】
例えばこの実施形態で説明したように、本発明の第１および第２の側面の特徴および態様は、組み合わされてもよい。第２の側面による従属請求項の任意の特徴は、さらなる実施形態をもたらすために、ここにも移転可能であるものとする。
【０２６９】
さらに、本発明の態様による装置は、現在のニューラルネットワークパラメータに対応する対応するニューラルネットワークパラメータに依存する確率モデルを用いて、算術符号化を用いてデータストリーム１４から現在のニューラルネットワークパラメータ１３’の量子化インデックス５６を復号化するように構成されていてもよい。
【０２７０】
言い換えれば、再構成層のニューラルネットワークパラメータ、例えば説明したような第２のニューラルネットワークパラメータは、それぞれ図３および図５ならびに図２および図４に関して説明した概念に従って符号化／復号化および／または量子化／逆量子化することができる。
【０２７１】
いくつかの態様は装置の文脈で説明されてきたが、これらの態様は、ブロックまたは装置が方法ステップまたは方法ステップの特徴に対応する、対応する方法の説明も表すことが明らかである。同様に、方法ステップの文脈で説明された側面は、対応するブロックまたは項目または対応する装置の特徴の説明も表している。
【０２７２】
本発明データストリームは、デジタル記憶媒体に格納することができ、または、無線伝送媒体またはインターネットなどの有線伝送媒体などの伝送媒体で伝送することができる。
【０２７３】
特定の実装要件に応じて、本発明の実施形態は、ハードウェアで実装することも、ソフトウェアで実装することも可能である。実装は、デジタル記憶媒体、例えばフロッピー（

登録商標）ディスク、ＤＶＤ、ＣＤ、ＲＯＭ、ＰＲＯＭ、ＥＰＲＯＭ、ＥＥＰＲＯＭまたはフラッシュメモリであって、その上に格納された電子的に読み取り可能な制御信号を有し、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システムと協力する（または協力することができる）ものを使用して実行することができる。

【０２７４】

本発明によるいくつかの実施形態は、電子的に読み取り可能な制御信号を有するデータキャリアを備え、このデータキャリアは、本明細書に記載の方法の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システムと協働することが可能である。

【０２７５】

一般に、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、コンピュータプログラム製品がコンピュータ上で実行されるときに、方法の１つを実行するために動作可能である。プログラムコードは、例えば、機械読み取り可能な担体に格納することができる。

【０２７６】

他の実施形態は、本明細書に記載された方法の１つを実行するためのコンピュータプログラムを、機械可読担体に格納することからなる。

【０２７７】

言い換えれば、本発明方法の実施形態は、したがって、コンピュータプログラムがコンピュータ上で実行されるときに、本明細書に記載された方法の１つを実行するためのプログラムコードを有するコンピュータプログラムである。

【０２７８】

したがって、本発明方法のさらなる実施形態は、本明細書に記載の方法の１つを実行するためのコンピュータプログラムをその上に記録してなるデータキャリア（またはデジタル記憶媒体、またはコンピュータ読取可能な媒体）である。

【０２７９】

本発明方法のさらなる実施形態は、したがって、本明細書に記載された方法の１つを実行するためのコンピュータプログラムを表すデータストリームまたは信号のシーケンスである。データストリームまたは信号のシーケンスは、例えば、データ通信接続、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。

【０２８０】

さらなる実施形態は、本明細書に記載された方法の１つを実行するように構成された、または適合された処理手段、例えばコンピュータ、またはプログラマブルロジックデバイスを備える。

【０２８１】

さらなる実施形態は、本明細書に記載された方法の１つを実行するためのコンピュータプログラムをその上にインストールしたコンピュータを具備する。

【０２８２】

いくつかの実施形態では、プログラマブルロジックデバイス（例えば、フィールドプログラマブルゲートアレイ）を使用して、本明細書に記載の方法の機能性の一部または全部を実行してもよい。いくつかの実施形態では、フィールドプログラマブルゲートアレイは、本明細書に記載される方法の１つを実行するためにマイクロプロセッサと協働してよい。一般に、本方法は、好ましくは、任意のハードウェア装置によって実行される。

【０２８３】

上述した実施形態は、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配置および詳細の修正および変形は、当業者には明らかであろうことが理解される。したがって、差し迫った特許請求の範囲の範囲によってのみ限定され、本明細書における実施形態の説明および解説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。

【０２８４】

References

[1] C. W. P. V. J. C. J. T. B. C. E. S. Sharan Chetlur, "cuDNN: Efficient Primitives for Deep Learning," arXiv: 1410.0759, 2014

[2] MPEG, "Working Draft 2 of Compression of neural networks for multimedia content description and analysis", Document of ISO/IEC JTC1/SC29/WG11, w18784, Geneva, Oct. 2019

[3] D. Marpe, H. Schwarz und T. Wiegand, „Context-Based Adaptive Binary Arithmetic Coding in the H.264/AVC Video Compression Standard," *IEEE transactions on circuits and systems for video technology, Vol. 13, No. 7,* pp. 620-636, July 2003.

[4] H. Kirchhoffer, J. Stegemann, D. Marpe, H. Schwarz und T. Wiegand, „JVET-K0430-v3 - CE5-related: State-based probalility estimator," in *JVET*, Ljubljana, 2018.

[5] ITU - International Telecommunication Union, „ITU-T H.265 High efficiency video coding," *Series H: Audiovisual and multimedia systems - Infrastructure of audiovisual services - Coding of moving video,* April 2015.

[6] B. Bross, J. Chen und S. Liu, „JVET-M1001-v6 - Versatile Video Coding (Draft 4)," in *JVET*, Marrakech, 2019.

【図面】

【図１】

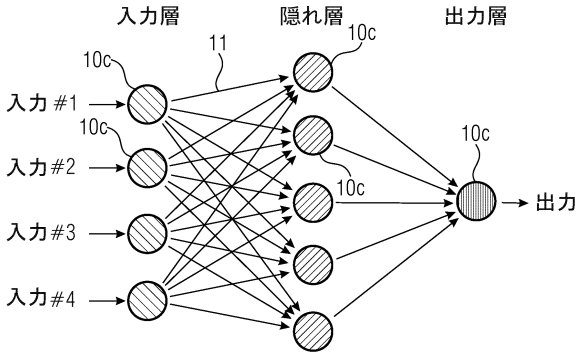

Fig. 1

【図２】

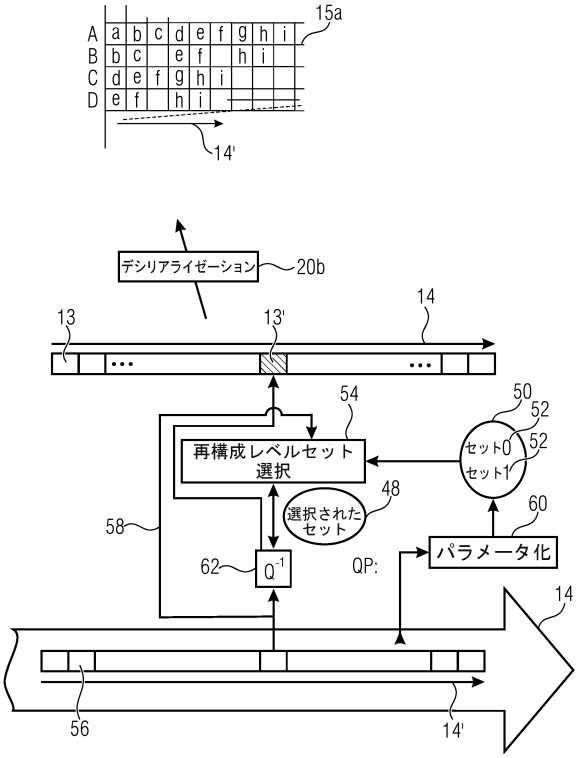

Fig. 2

【図３】

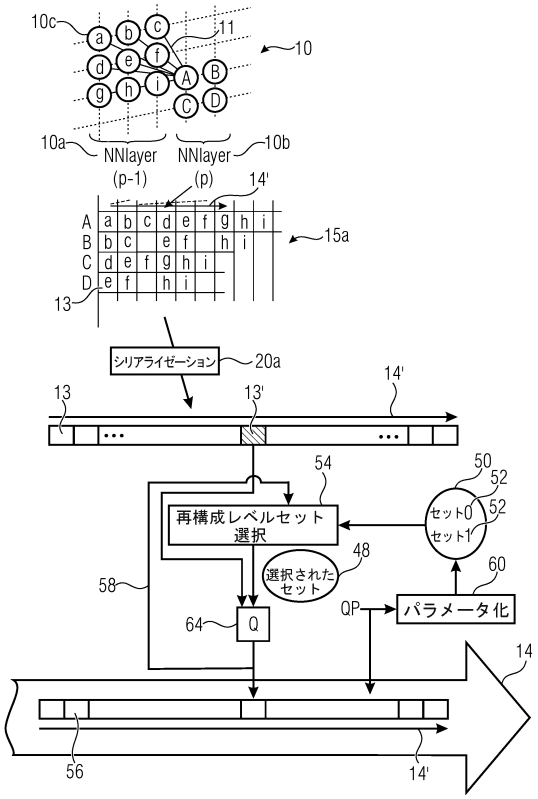

Fig. 3

【図４】

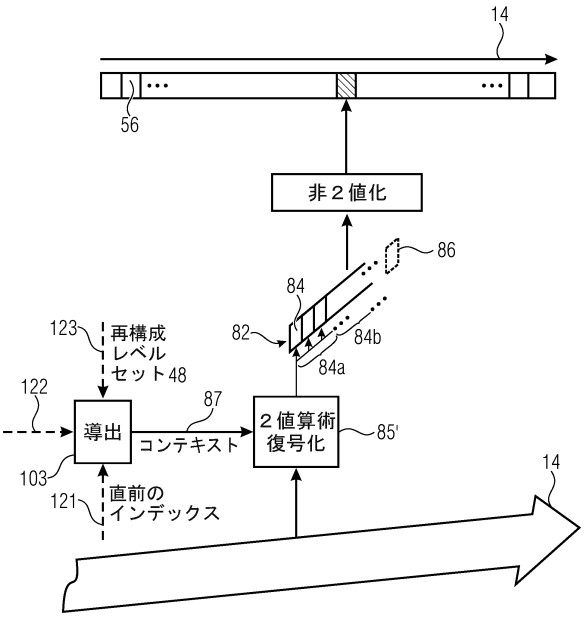

Fig. 4

【図５】

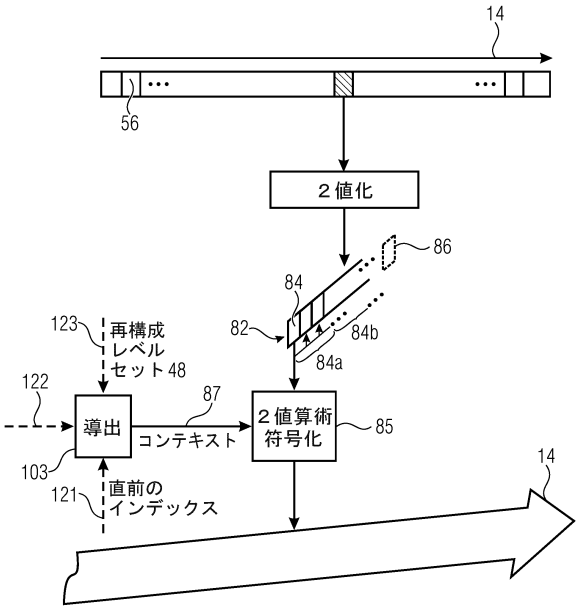

Fig. 5

【図６】

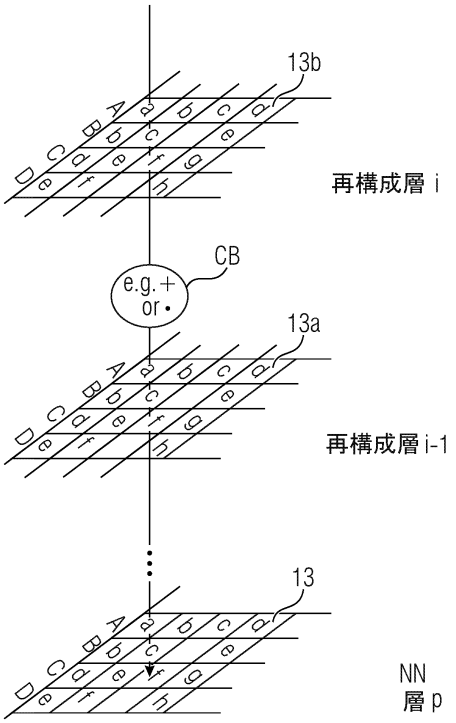

Fig. 6

【図７】

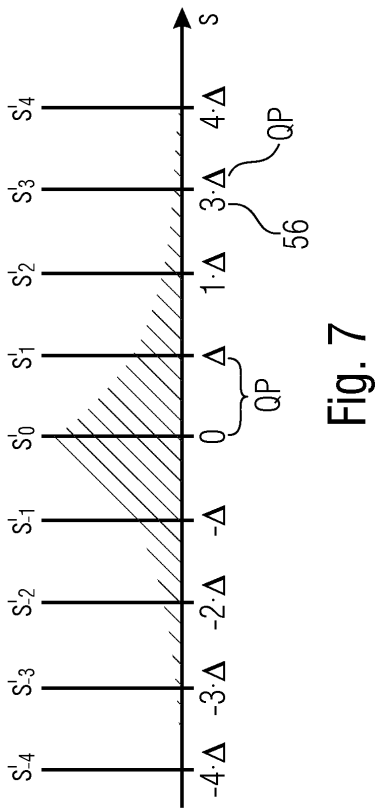

Fig. 7

【図８】

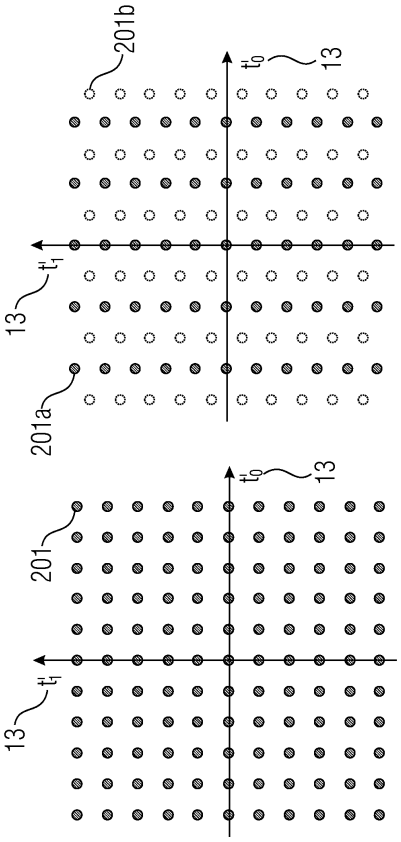

Fig. 8(b)

Fig. 8(a)

【図９ａ】

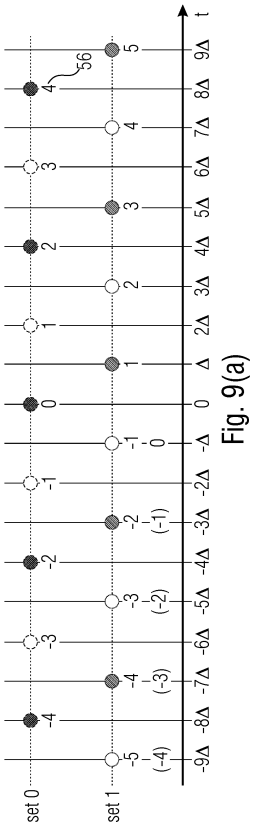

Fig. 9(a)

【図９ｂ】

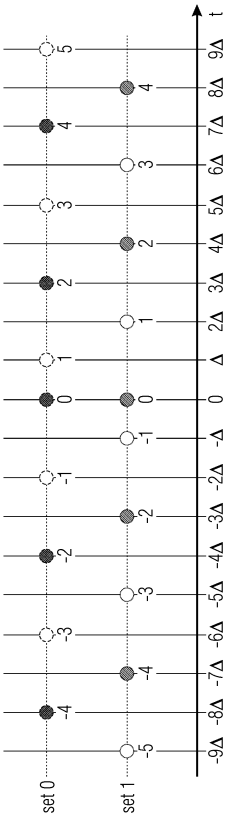

Fig. 9(b)

【図９ｃ】

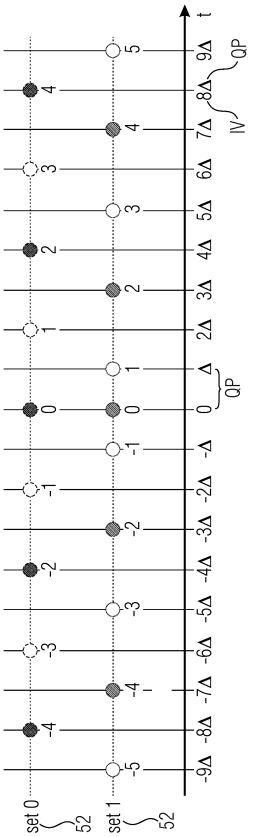

Fig. 9(c)

【図１０】

```
if ( setId [ k ] == 0 )  {
  n = 2 * level [ k ]                    210
} else  {
  n = 2 * level [ k ] - sign ( level [ k ] )
}
trec [ k ] = n * quant_step_size [ k ]
```

240
QP
220

Fig. 10

【図１１】

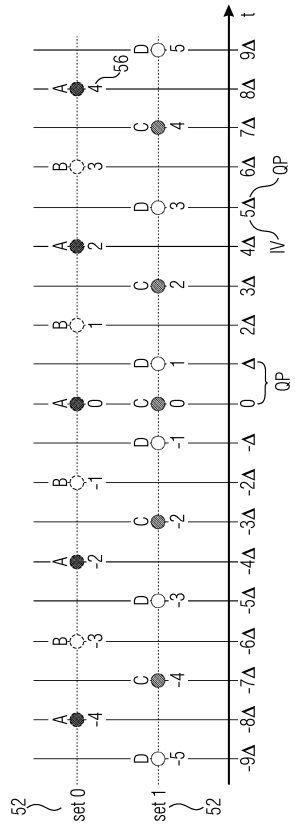

Fig. 11

【図１２】

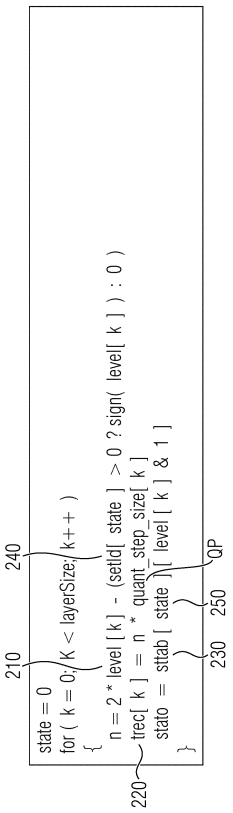

Fig. 12

【図１３】

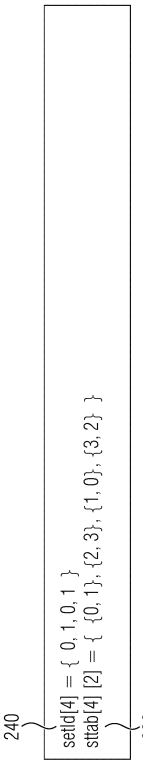

Fig. 13

【図１４】

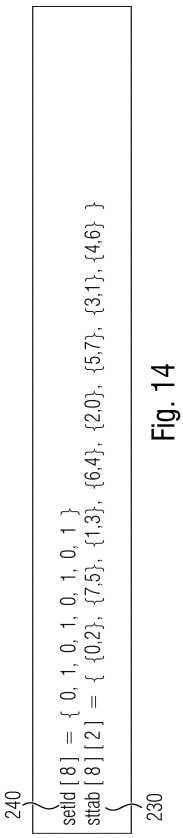

Fig. 14

【図１５】

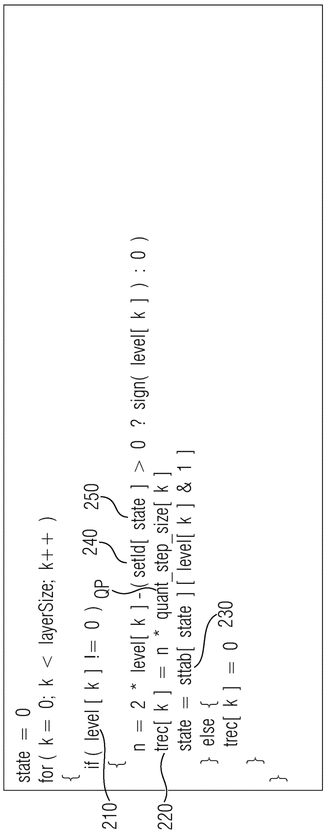

Fig. 15

【図１６】

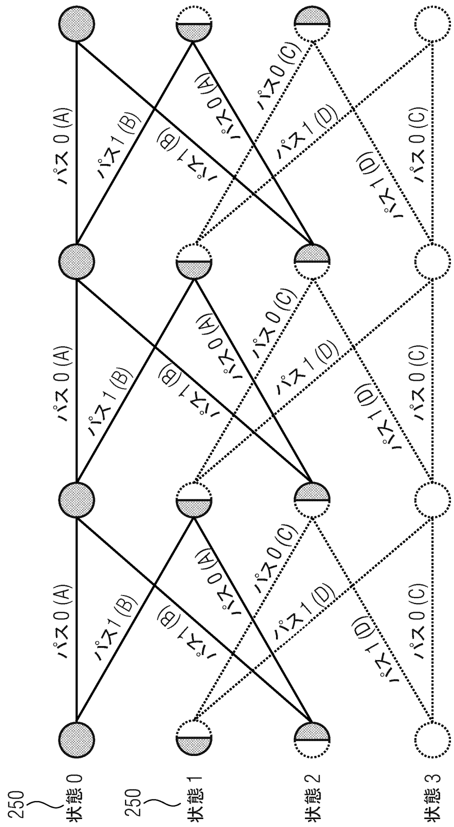

Fig. 16

【図１７】

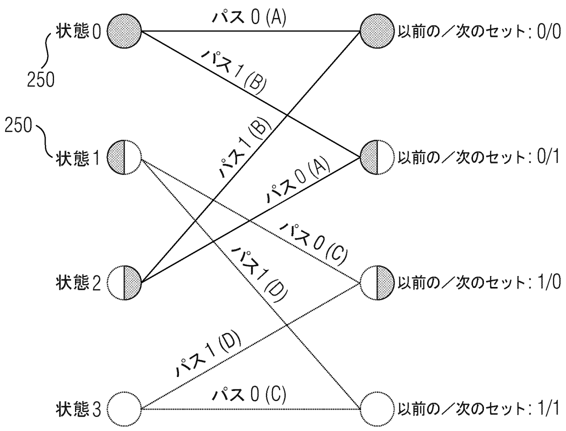

Fig. 17

【図１８】

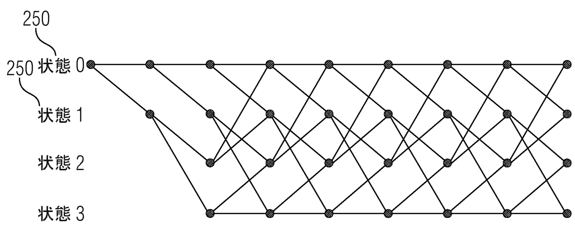

Fig. 18

【図１９】

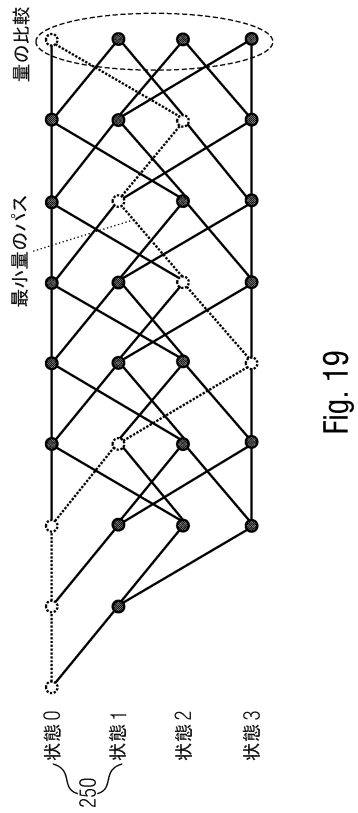

Fig. 19

【図２０】

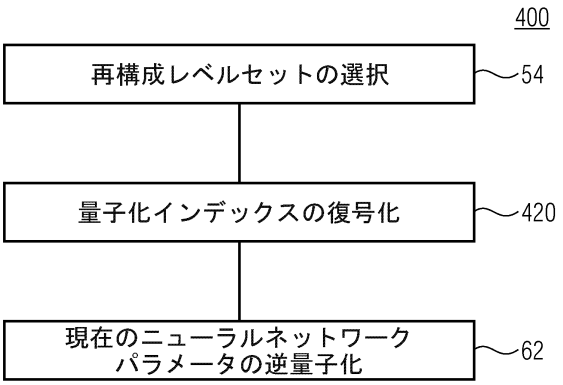

Fig. 20

【図２１】

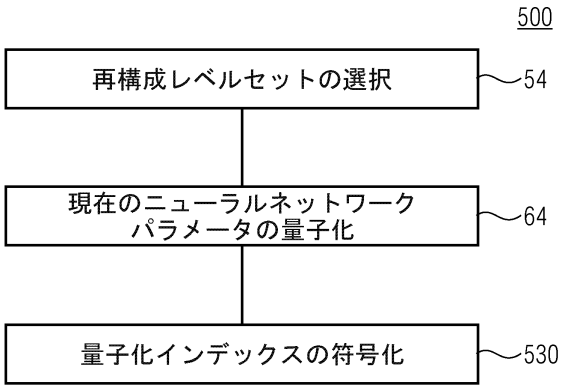

Fig. 21

【図２２】

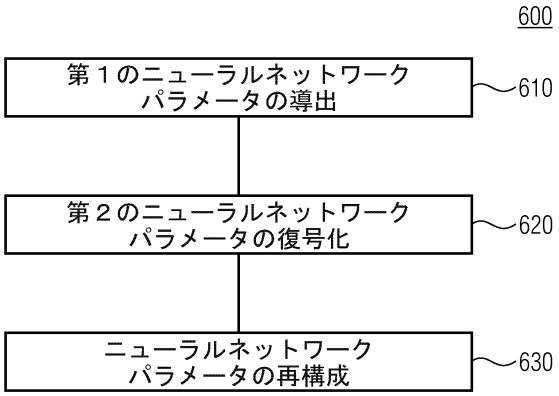

Fig. 22

【図２３】

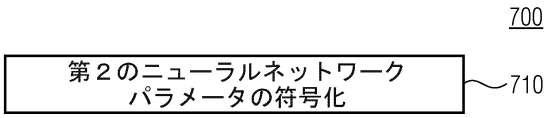

Fig. 23

フロントページの続き

(72)発明者　キルヒホッファー ハイナー
ドイツ連邦共和国 １０５８７ ベルリン アインシュタインウーファー ３７ フラウンホッファー－インスティチュート フュア ナーハリヒテンテヒニーク ハインリッヒ－ヘルツ－インスティチュート ＨＨＩ 内
(72)発明者　シュヴァルツ ハイコ
ドイツ連邦共和国 １０５８７ ベルリン アインシュタインウーファー ３７ フラウンホッファー－インスティチュート フュア ナーハリヒテンテヒニーク ハインリッヒ－ヘルツ－インスティチュート ＨＨＩ 内
(72)発明者　マルペ デトレフ
ドイツ連邦共和国 １０５８７ ベルリン アインシュタインウーファー ３７ フラウンホッファー－インスティチュート フュア ナーハリヒテンテヒニーク ハインリッヒ－ヘルツ－インスティチュート ＨＨＩ 内
(72)発明者　ウィーガント トーマス
ドイツ連邦共和国 １０５８７ ベルリン アインシュタインウーファー ３７ フラウンホッファー－インスティチュート フュア ナーハリヒテンテヒニーク ハインリッヒ－ヘルツ－インスティチュート ＨＨＩ 内
審査官　坂東 大五郎

(56)参考文献　国際公開第２０１９／１８５７６９（ＷＯ，Ａ１）
Simon Wiedemann et al.，DeepCABAC: A universal compression algorithm for deep neural networks，arXiv，2019年07月27日，pp.1-18
Marta Karczewicz et al.，CE8-related: Sign context modelling and level mapping for TS residual coding，Joint Video Experts Team (JVET)，2019年03月21日，[JVET-N0455] (version 3)

(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｈ０４Ｎ　１９／００－１９／９８