

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7695954号
(P7695954)

(45)発行日 令和7年6月19日(2025.6.19)

(24)登録日 令和7年6月11日(2025.6.11)

(51)国際特許分類

F I

H O 4 N 19/13 (2014.01)

H O 4 N 19/196 (2014.01)

H O 4 N 19/91 (2014.01)

H O 4 N 19/13

H O 4 N 19/196

H O 4 N 19/91

請求項の数 59 (全65頁)

(21)出願番号	特願2022-562944(P2022-562944)	(73)特許権者	504315750
(86)(22)出願日	令和3年4月13日(2021.4.13)		フラウンホーファー - ゲゼルシャフト
(65)公表番号	特表2023-522887(P2023-522887 A)		ツア フェルデルング デア アンゲヴァ
(43)公表日	令和5年6月1日(2023.6.1)		ンテン フォルシュング アイングトラ
(86)国際出願番号	PCT/EP2021/059594		ゲナー フェライン
(87)国際公開番号	WO2021/209470		ドイツ連邦共和国 8 0 6 8 6 ミュンヘ
(87)国際公開日	令和3年10月21日(2021.10.21)	(74)代理人	100133411
審査請求日	令和4年12月14日(2022.12.14)		弁理士 山本 龍郎
(31)優先権主張番号	20169498.1	(74)代理人	100067677
(32)優先日	令和2年4月14日(2020.4.14)		弁理士 山本 彰司
(33)優先権主張国・地域又は機関	欧州特許庁(EP)	(72)発明者	ハーゼ, パウル
前置審査			ドイツ連邦共和国 ベルリン 1 0 5 8 7
			アインシュタインウーファー 3 7 ハー
			ハーイー, ハインリッヒ・ヘルツェイン
			最終頁に続く

(54)【発明の名称】 ニューラルネットワークの重みパラメーターを復号化するデコーダー、エンコーダー、方法、及び確率推定パラメーターを使用する符号化表現

(57)【特許請求の範囲】

【請求項1】

ニューラルネットワークの重みパラメーターを復号化するデコーダー（200）であって、

前記デコーダー（200）は、符号化ビットストリーム（202）に基づいて前記ニューラルネットワークの複数のニューラルネットワークパラメーター（204）を取得するように構成され、

前記デコーダー（200）は、コンテキスト依存算術復号化を使用して前記ニューラルネットワークの前記ニューラルネットワークパラメーターを復号化するように構成され、

前記デコーダー（200）は、1つ以上の確率推定パラメーターを使用して、ニューラルネットワークパラメーターの数表現のビンの復号化のための確率推定値を取得するように構成され、

前記確率推定パラメーターは、前記コンテキスト依存算術復号化の前記確率推定値の適応レートに影響を与え、

前記デコーダー（200）は、異なるニューラルネットワークパラメーターの復号化のために異なる確率推定パラメーター値（230）を使用するように、及び／又は異なるコンテキストモデルに関連付けられたビンの復号化のために異なる確率推定パラメーター値（230）を使用するように構成されている、デコーダー。

【請求項2】

ニューラルネットワークの重みパラメーターを復号化するデコーダー（２００）であって、

前記デコーダー（２００）は、符号化ビットストリーム（２０２）に基づいて前記ニューラルネットワークの複数のニューラルネットワークパラメーター（２０４）を取得するように構成され、

前記デコーダー（２００）は、コンテキスト依存算術復号化を使用して前記ニューラルネットワークの前記ニューラルネットワークパラメーターを復号化するように構成され、

前記デコーダー（２００）は、１つ以上の確率推定パラメーターを使用して、ニューラルネットワークパラメーターの数表現のピンの復号化のための確率推定値を取得するように構成され、

前記確率推定パラメーターは、前記コンテキスト依存算術復号化の前記確率推定値の適応レートに影響を与え、

前記デコーダー（２００）は、前記ニューラルネットワークの異なる層に関連付けられたニューラルネットワークパラメーターの復号化のために、異なる確率推定パラメーター値（２３０）を使用するように構成されている、

デコーダー。

【請求項３】

請求項１、２のいずれかに記載のデコーダー（２００）であって、

前記デコーダー（２００）は、異なるニューラルネットワークパラメーターの復号化のコンテキストの適応のための異なる確率推定パラメーター値（２３０）を使用するように構成され、及び／又は

前記デコーダー（２００）は、異なるコンテキストモデルに関連付けられたピンの復号化のために異なる確率推定パラメーター値（２３０）を使用するように構成され、及び／又は

前記デコーダー（２００）は、前記ニューラルネットワークの異なる層に関連付けられたニューラルネットワークパラメーターの復号化のコンテキストの適応のための異なる確率推定パラメーター値（２３０）を使用するように構成されている、

デコーダー。

【請求項４】

請求項１～３のいずれか一項に記載のデコーダー（２００）であって、

前記デコーダー（２００）は、基集合（３００）から、

又は、前記基集合の真の部分集合（３１０、３２０、４１０、４１２、４１４、４２０）から、１つ以上の確率推定パラメーターを選択するように構成されている、

デコーダー。

【請求項５】

請求項１～４のいずれか一項に記載のデコーダー（２００）であって、

前記デコーダー（２００）は、量子化モードに依存して、使用可能なパラメーター値又はパラメーター値の使用可能なタプル（４３０、４４０）の異なる集合（４１０、４１２、４１４）から１つ以上の確率推定パラメーターを選択するように構成され、又は

前記デコーダー（２００）は、量子化モードに依存して、１つ以上の確率推定パラメーターを表す符号化値を１つ以上の確率推定パラメーターにマッピングする異なるマッピング規則（４５０、４６０）を使用するように構成されている、

デコーダー。

【請求項６】

請求項１～５のいずれか一項に記載のデコーダー（２００）であって、

前記デコーダー（２００）は、１つ以上の確率推定パラメーターの一樣量子化が使用される場合に、使用可能なパラメーター値の第１の集合（４１０、４１２）から、又はパラメーター値の使用可能なタプルの第１の集合（４３０）から、前記１つ以上の確率推定パラメーターを抜粋して選択するように構成され、

前記デコーダー（２００）は、１つ以上の確率推定パラメーターの可変量子化が使用さ

10

20

30

40

50

れる場合に、使用可能なパラメーター値の第2の集合（420、414）から、又はパラメーター値の使用可能なタブルの第2の集合（440）から、前記1つ以上の確率推定パラメーターを抜粋して選択するように構成され、又は

前記デコーダー（200）は、1つ以上の確率推定パラメーターの一樣量子化が使用される場合に、1つ以上の確率推定パラメーターを表す符号化値を前記1つ以上の確率推定パラメーターにマッピングする第1のマッピング規則（460）を使用するように構成され、

前記デコーダー（200）は、1つ以上の確率推定パラメーターの可変量子化が使用される場合に、1つ以上の確率推定パラメーターを表す符号化値を前記1つ以上の確率推定パラメーターにマッピングする第2のマッピング規則（450）を使用するように構成され、

10

使用可能なパラメーター値の前記第1の集合（410、412）は、使用可能なパラメーター値の前記第2の集合（414、420）とは異なり、

パラメーター値の使用可能なタブルの前記第1の集合（430）は、パラメーター値の使用可能なタブルの前記第2の集合（440）とは異なり、及び／又は

前記第2のマッピング規則（450）は、前記第1のマッピング規則（450）とは異なる、

デコーダー。

【請求項7】

請求項6に記載のデコーダー（200）であって、

20

平均して、使用可能なパラメーター値の前記第2の集合（414、420）のうちの使用可能なパラメーター値は、使用可能なパラメーター値の前記第1の集合のうちの使用可能なパラメーター値よりも高速な確率推定値の適応を可能にし、又は

平均して、パラメーター値の使用可能なタブルの前記第2の集合（440）のうちのパラメーター値の使用可能なタブルは、パラメーター値の使用可能なタブルの前記第1の集合（430）のうちのパラメーター値の使用可能なタブルよりも高速な確率推定値の適応を可能にする、

デコーダー。

【請求項8】

請求項6又は7に記載のデコーダー（200）であって、

30

使用可能なパラメーター値の前記第2の集合（414、420）は、使用可能なパラメーター値の前記第1の集合（410、412）のうちの使用可能なパラメーター値よりも高速な確率推定値の適応を可能にする使用可能なパラメーター値を含むか、又は

パラメーター値の使用可能なタブルの前記第2の集合（440）は、パラメーター値の使用可能なタブルの前記第1の集合（430）のうちのパラメーター値の使用可能なタブルよりも高速な確率推定値の適応を可能にするパラメーター値の使用可能なタブルを含む、

デコーダー。

【請求項9】

請求項1～8のいずれか一項に記載のデコーダー（200）であって、

40

前記デコーダー（200）は、1つ以上の確率推定パラメーターを選択することであって、前記ニューラルネットワークの層のパラメーターの数に依存して、又は前記選択された1つ以上の確率推定パラメーターを使用して復号化されるニューラルネットワークパラメーターの数に依存して、又は層パラメーターの要素の数に依存して、使用可能なパラメーター値又はパラメーター値の使用可能なタブル（430、440）の異なる集合（410、412、414、420）から選択することを行うように構成され、又は

前記デコーダー（200）は、前記ニューラルネットワークの層のパラメーターの数に依存して、又は前記選択された1つ以上の確率推定パラメーターを使用して復号化されるニューラルネットワークパラメーターの数に依存して、又は層パラメーターの要素の数に依存して、1つ以上の確率推定パラメーターを表す符号化値を1つ以上の確率推定パラメーターにマッピングする異なるマッピング規則（450、460）を使用するように構成

50

されている、
デコーダー。

【請求項 10】

請求項 9 に記載のデコーダー（200）であって、

前記デコーダー（200）は、1つ以上の確率推定パラメータを抜粋して選択することであって、前記ニューラルネットワークの層のパラメータの前記数が閾値を下回る場合、又は前記選択された1つ以上の確率推定パラメータを使用して復号化されるニューラルネットワークパラメータの前記数が閾値を下回る場合、又は前記層パラメータの要素の前記数が閾値を下回る場合、使用可能なパラメータ値の第1の集合（410、412）から、又はパラメータ値の使用可能なタブルの第1の集合（430）から選択することをを行うように構成され、

10

前記デコーダー（200）は、1つ以上の確率推定パラメータを抜粋して選択することであって、前記ニューラルネットワークの層のパラメータの前記数が前記閾値を超える場合、又は前記選択された1つ以上の確率推定パラメータを使用して復号化されるニューラルネットワークパラメータの前記数が前記閾値を超える場合、又は前記層パラメータの要素の前記数が前記閾値を超える場合、使用可能なパラメータ値の第2の集合（414、420）から、又はパラメータ値の使用可能なタブルの第2の集合（440）から選択することをを行うように構成され、又は

前記デコーダー（200）は、前記ニューラルネットワークの層のパラメータの前記数が閾値を下回る場合、又は前記選択された1つ以上の確率推定パラメータを使用して復号化されるニューラルネットワークパラメータの前記数が閾値を下回る場合、又は前記層パラメータの要素の前記数が閾値を下回る場合、1つ以上の確率推定パラメータを表す符号化値を1つ以上の確率推定パラメータにマッピングする第1のマッピング規則（460）を抜粋して使用するよう構成され、

20

前記デコーダー（200）は、前記ニューラルネットワークの層のパラメータの前記数が前記閾値を超える場合、又は前記選択された1つ以上の確率推定パラメータを使用して復号化されるニューラルネットワークパラメータの前記数が前記閾値を超える場合、又は前記層パラメータの要素の前記数が前記閾値を超える場合、1つ以上の確率推定パラメータを表す符号化値を1つ以上の確率推定パラメータにマッピングする第2のマッピング規則（450）を抜粋して使用するよう構成され、

30

使用可能なパラメータ値の前記第2の集合（414、420）は、使用可能なパラメータ値の前記第1の集合（410、412）よりも多くの使用可能なパラメータ値を含み、パラメータ値の使用可能なタブルの前記第2の集合（440）は、パラメータ値の使用可能なタブルの前記第1の集合（430）よりも多くの使用可能なタブルを含み、及び／又は

前記第2のマッピング規則（450）は、前記第1のマッピング規則（450）とは異なる、

デコーダー。

【請求項 11】

請求項 1～10 のいずれか一項に記載のデコーダー（200）であって、

40

前記デコーダー（200）は、前記1つ以上の確率推定パラメータを抜粋して選択することであって、前記選択された1つ以上の確率推定パラメータを使用して復号化されるニューラルネットワークパラメータの数が閾値以上である場合、増加した選択肢から選択することをを行うように構成されている、

デコーダー。

【請求項 12】

請求項 1～11 のいずれか一項に記載のデコーダー（200）であって、

前記デコーダー（200）は、前記1つ以上の確率推定パラメータが、使用可能なパラメータ値のどの集合から、又はパラメータ値の使用可能なタブルのどの集合から選択されるかのシグナリングを評価するよう構成され、又は

50

前記デコーダー（２００）は、１つ以上の確率推定パラメーターを表す符号化値を１つ以上の確率推定パラメーターにマッピングするために、複数のマッピング規則のうちのどのマッピング規則を使用すべきかのシグナリング指示を評価するように構成されている、デコーダー。

【請求項１３】

請求項１～１２のいずれか一項に記載のデコーダー（２００）であって、

前記デコーダー（２００）は、確率推定パラメーター値を記述する、又は複数の確率推定パラメーター値（２３０）を記述する、又は確率推定パラメーター値（２３０）のタブルを記述する１つ以上のインデックス値を復号化するように構成されている、デコーダー。

10

【請求項１４】

請求項１３に記載のデコーダー（２００）であって、前記デコーダー（２００）は、１つ以上のコンテキストモデルを使用して前記１つ以上のインデックス値を復号化するように構成されている、デコーダー。

【請求項１５】

請求項１３又は１４に記載のデコーダー（２００）であって、

前記デコーダー（２００）は、現在考慮されているインデックス値がデフォルト値を取るかどうかを記述する第１のピンを復号化するように構成され、

前記デコーダー（２００）は、前記現在考慮されているインデックス値が前記デフォルト値を取らない場合に、前記現在考慮されているインデックス値又は前記現在考慮されているインデックス値から導出された値を二値表現で表す１つ以上の追加のピンを抜粋して復号化するように構成されている、デコーダー。

20

【請求項１６】

請求項１３又は１４に記載のデコーダー（２００）であって、

前記デコーダー（２００）は、単項コード復号化を使用して、又は短縮単項コード復号化を使用して、又は可変長コード復号化を使用して、前記１つ以上のインデックス値を復号化するように構成されている、デコーダー。

30

【請求項１７】

請求項１～１６のいずれか一項に記載のデコーダー（２００）であって、

前記デコーダー（２００）は、前記１つ以上の確率推定パラメーターを量子化するために使用される量子化モードに応じて、前記１つ以上の確率推定パラメーターを復号化するために使用されるピンの数又はピンの最大数を変更するように構成されている、デコーダー。

【請求項１８】

請求項１～１７のいずれか一項に記載のデコーダー（２００）であって、

前記デコーダー（２００）は、前記ニューラルネットワークの層のパラメーターの数に依存して、又は前記１つ以上の確率推定パラメーターを使用して復号化されるニューラルネットワークパラメーターの数に依存して、又は層パラメーターの要素の数に依存して、前記１つ以上の確率推定パラメーターを復号化するために使用されるピンの数又はピンの最大数を変更するように構成されている、デコーダー。

40

【請求項１９】

請求項１～１８のいずれか一項に記載のデコーダー（２００）であって、

前記デコーダー（２００）は、前記１つ以上の確率推定パラメーターに関連付けられた使用可能なパラメーター値の異なる集合（４１０、４１２、４１４、４２０）の間で、又は複数の確率推定パラメーターに関連付けられた使用可能なパラメーター値のタブルの異なる集合（４３０、４４０）の間で、又は１つ以上の確率推定パラメーターを表す符号化

50

値を1つ以上の確率推定パラメーターにマッピングする異なるマッピング規則（450、460）の間で切り替えるように構成されている、デコーダー。

【請求項20】

請求項19に記載のデコーダー（200）であって、

前記デコーダー（200）は、前記1つ以上の確率推定パラメーターに関連付けられた使用可能なパラメーター値の異なる集合（410、412、414、420）の間、又は複数の確率推定パラメーターに関連付けられた使用可能なパラメーター値のタブルの異なる集合（430、440）の間、又は異なるマッピング規則（450、460）の間の切り替えに従って、選択された確率推定パラメーター又は確率推定パラメーターの選択されたタブルを指定する前記1つ以上の確率推定パラメーターを復号化するために使用されるビンの数又はビンの最大数を変更するように構成されている、デコーダー。

10

【請求項21】

請求項1～20のいずれか一項に記載のデコーダー（200）であって、

前記デコーダー（200）は、1つ以上の状態変数を決定し、前記1つ以上の状態変数を使用して前記確率推定値を導出するように構成されている、デコーダー。

【請求項22】

請求項1～21のいずれか一項に記載のデコーダー（200）であって、

前記デコーダー（200）は、

20

【数1】

$$s_k = \sum_{i=1}^N [s_i^k \cdot d_i^k]$$

及び

【数2】

30

$$p_k = \begin{cases} LUT2[[s_k \cdot a_k]], & [s_k \cdot a_k] \geq 0 \text{ である場合} \\ 1 - LUT2[-[s_k \cdot a_k]], & \text{それ以外の場合} \end{cases}$$

に従って2つの状態変数 s_1^k 、 s_2^k から確率推定値 p_k を導出するように構成されている、デコーダー。

【請求項23】

請求項22に記載のデコーダー（200）であって、

前記デコーダー（200）は、

40

【数3】

$$s_i^k = \begin{cases} s_i^k + [A[z + [s_i^k \cdot m_i^k]] \cdot n_i^k], & \text{復号化されるシンボルが1である場合} \\ s_i^k - [A[z + [-s_i^k \cdot m_i^k]] \cdot n_i^k], & \text{復号化されるシンボルが0である場合} \end{cases}$$

に従って前記状態変数 s_1^k 、 s_2^k を更新するように構成され、

式中、 m_i^k 及び n_i^k は、重み係数であり、

50

A は、ルックアップテーブルであり、
z は、オフセット値である、
デコーダー。

【請求項 24】

請求項 23 に記載のデコーダー（200）であって、

前記デコーダー（200）は、異なるニューラルネットワークパラメーターの復号化のために異なる確率推定パラメーター値（230）を使用するように、及び／又は異なるコンテキストモデルに関連付けられたビンの復号化のために異なる確率推定パラメーター値（230）を使用するように、及び／又は前記ニューラルネットワークの異なる層に関連付けられたニューラルネットワークパラメーターの復号化のために異なる確率推定パラメーター値（230）を使用するように、前記重み係数 n_i^k を変更するように構成されている、

10

デコーダー。

【請求項 25】

請求項 23 又は 24 に記載のデコーダー（200）であって、

前記重み係数 n_i^k と適応パラメーター sh_i^k との関係は、

【数 4】

$$n_i^k = 2^{-sh_i^k + 4}$$

20

に従って定義される、

デコーダー。

【請求項 26】

請求項 25 に記載のデコーダー（200）であって、

前記デコーダー（200）は、前記適応パラメーターを記述する情報を復号化するように構成されている、

デコーダー。

【請求項 27】

ニューラルネットワークの重みパラメーターを符号化するエンコーダーであって、

前記エンコーダーは、前記ニューラルネットワークの複数のニューラルネットワークパラメーターを取得するように構成され、

前記エンコーダーは、コンテキスト依存算術符号化を使用して前記ニューラルネットワークの前記ニューラルネットワークパラメーターを符号化するように構成され、

前記エンコーダーは、1 つ以上の確率推定パラメーターを使用して、ニューラルネットワークパラメーターの数表現のビンの符号化のための確率推定値を取得するように構成され、

前記確率推定パラメーターは、前記コンテキスト依存算術符号化の前記確率推定値の適応レートに影響を与え、

30

前記エンコーダーは、異なるニューラルネットワークパラメーターの符号化のために異なる確率推定パラメーター値（230）を使用するように、及び／又は異なるコンテキストモデルに関連付けられたビンの符号化のために異なる確率推定パラメーター値（230）を使用するように構成されている、

40

エンコーダー。

【請求項 28】

ニューラルネットワークの重みパラメーターを符号化するエンコーダーであって、

前記エンコーダーは、前記ニューラルネットワークの複数のニューラルネットワークパラメーターを取得するように構成され、

前記エンコーダーは、コンテキスト依存算術符号化を使用して前記ニューラルネットワ

50

ークの前記ニューラルネットワークパラメーターを符号化するように構成され、

前記エンコーダーは、1つ以上の確率推定パラメーターを使用して、ニューラルネットワークパラメーターの数表現のビンの符号化のための確率推定値を取得するように構成され、

前記確率推定パラメーターは、前記コンテキスト依存算術符号化の前記確率推定値の適応レートに影響を与え、

前記エンコーダーは、前記ニューラルネットワークの異なる層に関連付けられたニューラルネットワークパラメーターの符号化のために、異なる確率推定パラメーター値（230）を使用するように構成されている、

エンコーダー。

【請求項29】

請求項27、28のいずれかに記載のエンコーダーであって、

前記エンコーダーは、異なるニューラルネットワークパラメーターの符号化のコンテキストの適応のための異なる確率推定パラメーター値（230）を使用するように構成され、及び／又は

前記エンコーダーは、異なるコンテキストモデルに関連付けられたビンの符号化のために異なる確率推定パラメーター値（230）を使用するように構成され、及び／又は

前記エンコーダーは、前記ニューラルネットワークの異なる層に関連付けられたニューラルネットワークパラメーターの符号化のコンテキストの適応のための異なる確率推定パラメーター値（230）を使用するように構成されている、

エンコーダー。

【請求項30】

前記エンコーダーは、基集合（300）から、又は前記基集合の真の部分集合（310、320、410、412、414、420）から、1つ以上の確率推定パラメーターを選択するように構成されている、請求項27～29のいずれか一項に記載のエンコーダー。

【請求項31】

前記エンコーダーは、量子化モードに依存して、使用可能なパラメーター値又はパラメーター値の使用可能なタプル（430、440）の異なる集合（410、412、414）から1つ以上の確率推定パラメーターを選択するように構成されている、請求項27～30のいずれか一項に記載のエンコーダー。

【請求項32】

前記エンコーダーは、1つ以上の確率推定パラメーターの一樣量子化が使用される場合に、使用可能なパラメーター値の第1の集合（410、412）から、又はパラメーター値の使用可能なタプルの第1の集合（430）から、前記1つ以上の確率推定パラメーターを抜粋して選択するように構成され、

前記エンコーダーは、1つ以上の確率推定パラメーターの可変量子化が使用される場合に、使用可能なパラメーター値の第2の集合（414、420）から、又はパラメーター値の使用可能なタプルの第2の集合（440）から、前記1つ以上の確率推定パラメーターを抜粋して選択するように構成され、

使用可能なパラメーター値の前記第1の集合（410、412）は、使用可能なパラメーター値の前記第2の集合（414、420）とは異なり、

パラメーター値の使用可能なタプルの前記第1の集合（430）は、パラメーター値の使用可能なタプルの前記第2の集合（440）とは異なる、

請求項27～31のいずれか一項に記載のエンコーダー。

【請求項33】

平均して、使用可能なパラメーター値の前記第2の集合（414、420）のうちの使用可能なパラメーター値は、使用可能なパラメーター値の前記第1の集合（410、412）のうちの使用可能なパラメーター値よりも高速な確率推定値の適応を可能にし、又は

平均して、パラメーター値の使用可能なタプルの前記第2の集合（440）のうちのパラメーター値の使用可能なタプルは、パラメーター値の使用可能なタプルの前記第1の集

10

20

30

40

50

合（４３０）のうちのパラメーター値の使用可能なタプルよりも高速な確率推定値の適応を可能にする、

請求項３２に記載のエンコーダー。

【請求項３４】

使用可能なパラメーター値の前記第２の集合（４１４、４２０）は、使用可能なパラメーター値の前記第１の集合（４１０、４１２）のうちの使用可能なパラメーター値よりも高速な確率推定値の適応を可能にする使用可能なパラメーター値を含むか、又は

パラメーター値の使用可能なタプルの前記第２の集合（４４０）は、パラメーター値の使用可能なタプルの前記第１の集合（４３０）のうちのパラメーター値の使用可能なタプルよりも高速な確率推定値の適応を可能にするパラメーター値の使用可能なタプルを含む、請求項３２又は３３に記載のエンコーダー。

10

【請求項３５】

前記エンコーダーは、１つ以上の確率推定パラメーターを選択することであって、前記ニューラルネットワークの層のパラメーターの数に依存して、又は前記選択された１つ以上の確率推定パラメーターを使用して符号化されるニューラルネットワークパラメーターの数に依存して、又は層パラメーターの要素の数に依存して、使用可能なパラメーター値又はパラメーター値の使用可能なタプル（４３０、４４０）の異なる集合（４１０、４１２、４１４、４２０）から選択することを行うように構成されている、請求項２７～３４のいずれか一項に記載のエンコーダー。

【請求項３６】

20

前記エンコーダーは、１つ以上の確率推定パラメーターを抜粋して選択することであって、前記ニューラルネットワークの層のパラメーターの前記数が閾値を下回る場合、又は前記選択された１つ以上の確率推定パラメーターを使用して符号化されるニューラルネットワークパラメーターの前記数が閾値を下回る場合、又は前記層パラメーターの要素の前記数が閾値を下回る場合、使用可能なパラメーター値の第１の集合（４１０、４１２）から、又はパラメーター値の使用可能なタプルの第１の集合（４３０）から選択することを行うように構成され、

前記エンコーダーは、１つ以上の確率推定パラメーターを抜粋して選択することであって、前記ニューラルネットワークの層のパラメーターの前記数が前記閾値を超える場合、又は前記選択された１つ以上の確率推定パラメーターを使用して符号化されるニューラルネットワークパラメーターの前記数が前記閾値を超える場合、又は前記層パラメーターの要素の前記数が前記閾値を超える場合、使用可能なパラメーター値の第２の集合（４１４、４２０）から、又はパラメーター値の使用可能なタプルの第２の集合（４４０）から選択することを行うように構成され、

30

使用可能なパラメーター値の前記第２の集合（４１４、４２０）は、使用可能なパラメーター値の前記第１の集合（４１０、４１２）よりも多くの使用可能なパラメーター値を含み、パラメーター値の使用可能なタプルの前記第２の集合（４４０）は、パラメーター値の使用可能なタプルの前記第１の集合（４３０）よりも多くの使用可能なタプルを含む、請求項３５に記載のエンコーダー。

【請求項３７】

40

前記エンコーダーは、前記１つ以上の確率推定パラメーターを抜粋して選択することであって、前記選択された１つ以上の確率推定パラメーターを使用して符号化されるニューラルネットワークパラメーターの数が閾値以上である場合、増加した選択肢から選択することを行うように構成されている、請求項２７～３６のいずれか一項に記載のエンコーダー。

【請求項３８】

前記エンコーダーは、前記１つ以上の確率推定パラメーターが、使用可能なパラメーター値のどの集合（４１０、４１２、４１４、４２０）から、又はパラメーター値の使用可能なタプル（４３０、４４０）のどの集合から選択されるかをシグナリングするように構成されている、請求項２７～３７のいずれか一項に記載のエンコーダー。

50

【請求項 39】

前記エンコーダーは、確率推定パラメーター値を記述する、又は複数の確率推定パラメーター値（230）を記述する、又は確率推定パラメーター値（230）のタブルを記述する1つ以上のインデックス値を符号化するように構成されている、請求項27～38のいずれか一項に記載のエンコーダー。

【請求項 40】

前記エンコーダーは、1つ以上のコンテキストモデルを使用して前記1つ以上のインデックス値を符号化するように構成されている、請求項39に記載のエンコーダー。

【請求項 41】

前記エンコーダーは、現在考慮されているインデックス値がデフォルト値を取る場合、前記現在考慮されているインデックス値が前記デフォルト値を取ることを記述する第1のピンを使用して前記現在考慮されているインデックス値を符号化するように構成され、

前記エンコーダーは、前記現在考慮されているインデックス値が前記デフォルト値を取らない場合、前記現在考慮されているインデックス値が前記デフォルト値を取らないことを記述する第1のピンを使用して、かつ前記現在考慮されているインデックス値又は前記現在考慮されているインデックス値から導出された値を二値表現で表す1つ以上の追加のピンを使用して、前記現在考慮されているインデックス値を符号化するように構成されている、

請求項39又は40に記載のエンコーダー。

【請求項 42】

前記エンコーダーは、単項コードを使用して、又は短縮単項コードを使用して、又は可変長コードを使用して、前記1つ以上のインデックス値を符号化するように構成されている、請求項39又は40に記載のエンコーダー。

【請求項 43】

前記エンコーダーは、前記1つ以上の確率推定パラメーターを量子化するために使用される量子化モードに応じて、前記1つ以上の確率推定パラメーターを符号化するために使用されるピンの数又はピンの最大数を変更するように構成されている、請求項27～42のいずれか一項に記載のエンコーダー。

【請求項 44】

前記エンコーダーは、前記ニューラルネットワークの層のパラメーターの数に依存して、又は前記1つ以上の確率推定パラメーターを使用して符号化されるニューラルネットワークパラメーターの数に依存して、又は層パラメーターの要素の数に依存して、前記1つ以上の確率推定パラメーターを符号化するために使用されるピンの数又はピンの最大数を変更するように構成されている、請求項27～43のいずれか一項に記載のエンコーダー。

【請求項 45】

前記エンコーダーは、前記1つ以上の確率推定パラメーターに関連付けられた使用可能なパラメーター値の異なる集合（410、412、414、420）の間で、又は複数の確率推定パラメーターに関連付けられた使用可能なパラメーター値のタブル（430、440）の異なる集合の間で切り替えるように構成されている、請求項27～44のいずれか一項に記載のエンコーダー。

【請求項 46】

前記エンコーダーは、前記1つ以上の確率推定パラメーターに関連付けられた使用可能なパラメーター値の異なる集合（410、412、414、420）の間、又は複数の確率推定パラメーターに関連付けられた使用可能なパラメーター値のタブルの異なる集合（430、440）の間の切り替えに従って、選択された確率推定パラメーター又は確率推定パラメーターの選択されたタブルを指定する前記1つ以上の確率推定パラメーターを符号化するために使用されるピンの数又はピンの最大数を変更するように構成されている、請求項45に記載のエンコーダー。

【請求項 47】

前記エンコーダーは、1つ以上の状態変数を決定し、前記1つ以上の状態変数を使用し

10

20

30

40

50

て前記確率推定値を導出するように構成されている、請求項 27～46 のいずれか一項に記載のエンコーダー。

【請求項 48】

前記エンコーダーは、

【数 5】

$$s_k = \sum_{i=1}^N [s_i^k \cdot d_i^k]$$

10

及び

【数 6】

$$p_k = \begin{cases} \text{LUT2}[s_k \cdot a_k], & [s_k \cdot a_k] \geq 0 \text{ である場合} \\ 1 - \text{LUT2}[-[s_k \cdot a_k]], & \text{それ以外の場合} \end{cases}$$

に従って 2 つの状態変数 s_1^k 、 s_2^k から確率推定値 p_k を導出するように構成されている、請求項 27～47 のいずれか一項に記載のエンコーダー。

20

【請求項 49】

前記エンコーダーは、

【数 7】

$$s_i^k = \begin{cases} s_i^k + [A[z + [s_i^k \cdot m_i^k]] \cdot n_i^k], & \text{符号化されるシンボルが 1 である場合} \\ s_i^k - [A[z + [-s_i^k \cdot m_i^k]] \cdot n_i^k], & \text{符号化されるシンボルが 0 である場合} \end{cases}$$

30

に従って前記状態変数 s_1^k 、 s_2^k を更新するように構成され、

式中、 m_i^k 及び n_i^k は、重み係数であり、

A は、ルックアップテーブルであり、

z は、オフセット値である、

請求項 48 に記載のエンコーダー。

【請求項 50】

前記エンコーダーは、異なるニューラルネットワークパラメーターの符号化のために異なる確率推定パラメーター値 (230) を使用するように、及び／又は異なるコンテキストモデルに関連付けられたピンの符号化のために異なる確率推定パラメーター値 (230) を使用するように、及び／又は前記ニューラルネットワークの異なる層に関連付けられたニューラルネットワークパラメーターの符号化のために異なる確率推定パラメーター値 (230) を使用するように、前記重み係数 n_i^k を変更するように構成されている、請求項 49 に記載のエンコーダー。

40

【請求項 51】

前記重み係数 n_i^k と適応パラメーター $s_{h_i}^k$ との関係は、

【数 8】

50

$$n_i^k = 2^{-sh_i^k + 4}$$

に従って定義される、

請求項 49 又は 50 に記載のエンコーダー。

【請求項 52】

前記エンコーダーは、適応パラメーターを記述する情報を符号化するように構成されている、請求項 51 に記載のエンコーダー。

【請求項 53】

ニューラルネットワークの重みパラメーターを復号化する方法（600）であって、

前記方法は、符号化ビットストリーム（202）に基づいて前記ニューラルネットワークの複数のニューラルネットワークパラメーター（204）を取得（610）することを含み、

前記方法は、コンテキスト依存算術復号化を使用して前記ニューラルネットワークの前記ニューラルネットワークパラメーターを復号化（620）することを含み、

前記方法は、1つ以上の確率推定パラメーターを使用して、ニューラルネットワークパラメーターの数表現のビンの復号化のための確率推定値を取得（630）することを含み、

前記確率推定パラメーターは、前記コンテキスト依存算術復号化の前記確率推定値の適応レートに影響を与え、

前記方法は、異なるニューラルネットワークパラメーターの復号化のために異なる確率推定パラメーター値（230）を使用（640）すること、及び／又は異なるコンテキストモデルに関連付けられたビンの復号化のために異なる確率推定パラメーター値（230）を使用することを含む、

方法。

【請求項 54】

ニューラルネットワークの重みパラメーターを復号化する方法（700）であって、

前記方法は、符号化ビットストリーム（202）に基づいて前記ニューラルネットワークの複数のニューラルネットワークパラメーター（204）を取得（710）することを含み、

前記方法は、コンテキスト依存算術復号化を使用して前記ニューラルネットワークの前記ニューラルネットワークパラメーターを復号化（720）することを含み、

前記方法は、1つ以上の確率推定パラメーターを使用して、ニューラルネットワークパラメーターの数表現のビンの復号化のための確率推定値を取得（730）することを含み、

前記確率推定パラメーターは、前記コンテキスト依存算術復号化の前記確率推定値の適応レートに影響を与え、

前記方法は、前記ニューラルネットワークの異なる層に関連付けられたニューラルネットワークパラメーターの復号化のために異なる確率推定パラメーター値（230）を使用（740）することを含む、

方法。

【請求項 55】

ニューラルネットワークの重みパラメーターを符号化する方法であって、

前記方法は、前記ニューラルネットワークの複数のニューラルネットワークパラメーター（204）を取得することを含み、

前記方法は、コンテキスト依存算術符号化を使用して前記ニューラルネットワークの前記ニューラルネットワークパラメーターを符号化することを含み、

前記方法は、1つ以上の確率推定パラメーターを使用して、ニューラルネットワークパラメーターの数表現のビンの符号化のための確率推定値を取得することを含み、

10

20

30

40

50

前記確率推定パラメーターは、前記コンテキスト依存算術符号化の前記確率推定値の適応レートに影響を与え、

前記方法は、異なるニューラルネットワークパラメーターの符号化のために異なる確率推定パラメーター値（２３０）を使用すること、及び／又は異なるコンテキストモデルに関連付けられたピンの符号化のために異なる確率推定パラメーター値（２３０）を使用することを含む、

方法。

【請求項５６】

ニューラルネットワークの重みパラメーターを符号化する方法であって、

前記方法は、前記ニューラルネットワークの複数のニューラルネットワークパラメーターを取得することを含み、

前記方法は、コンテキスト依存算術符号化を使用して前記ニューラルネットワークの前記ニューラルネットワークパラメーターを符号化することを含み、

前記方法は、１つ以上の確率推定パラメーターを使用して、ニューラルネットワークパラメーターの数表現のピンの符号化のための確率推定値を取得することを含み、

前記確率推定パラメーターは、前記コンテキスト依存算術符号化の前記確率推定値の適応レートに影響を与え、

前記方法は、前記ニューラルネットワークの異なる層に関連付けられたニューラルネットワークパラメーターの符号化のために異なる確率推定パラメーター値（２３０）を使用することを含む、

方法。

【請求項５７】

コンピュータープログラムであって、前記コンピュータープログラムがコンピューター上で動作するとき、請求項５３～５６のいずれか一項に記載の方法を実行するコンピュータープログラム。

【請求項５８】

ニューラルネットワークの重みパラメーターを復号化するデコーダー（２００）であって、前記デコーダー（２００）は、符号化ビットストリーム（２０２）に基づいて前記ニューラルネットワークの複数のニューラルネットワークパラメーター（２０４）を取得するように構成され、

前記デコーダー（２００）は、コンテキスト依存算術復号化を使用して前記ニューラルネットワークの前記ニューラルネットワークパラメーターを復号化するように構成され、

前記デコーダー（２００）は、１つ以上の確率推定パラメーターを使用して、ニューラルネットワークパラメーターの数表現のピンの復号化のための確率推定値を取得するように構成され、

コンテキスト依存算術復号化の各コンテキストモデルについて、１つ以上の状態変数が維持されて確率値を表し、

前記１つ以上の確率推定パラメーターが、コンテキストモデルの状態変数の初期化値からなり、

前記デコーダー（２００）は、異なるニューラルネットワークパラメーターの復号化のために異なる確率推定パラメーター値（２３０）を使用するように、及び／又は異なるコンテキストモデルに関連付けられたピンの復号化のために異なる確率推定パラメーター値（２３０）を使用するように構成されており、

前記デコーダー（２００）は、前記ニューラルネットワークの異なる層に関連付けられたニューラルネットワークパラメーターの復号化のために、異なる確率推定パラメーター値（２３０）を使用するように構成されている、

デコーダー。

【請求項５９】

ニューラルネットワークの重みパラメーターを符号化するエンコーダーであって、

前記エンコーダーは、前記ニューラルネットワークの複数のニューラルネットワークパ

10

20

30

40

50

ラメーターを取得するように構成され、

前記エンコーダーは、コンテキスト依存算術符号化を使用して前記ニューラルネットワークの前記ニューラルネットワークパラメーターを符号化するように構成され、

前記エンコーダーは、1つ以上の確率推定パラメーターを使用して、ニューラルネットワークパラメーターの数表現のピンの符号化のための確率推定値を取得するように構成され、

コンテキスト依存算術符号化の各コンテキストモデルについて、1つ以上の状態変数が維持されて確率値を表し、

前記1つ以上の確率推定パラメーターが、コンテキストモデルの状態変数の初期化値からなり、

前記エンコーダーは、異なるニューラルネットワークパラメーターの符号化のために異なる確率推定パラメーター値(230)を使用するように、及び／又は異なるコンテキストモデルに関連付けられたピンの符号化のために異なる確率推定パラメーター値(230)を使用するように構成されており、

前記エンコーダーは、前記ニューラルネットワークの異なる層に関連付けられたニューラルネットワークパラメーターの符号化のために、異なる確率推定パラメーター値(230)を使用するように構成されている、

エンコーダー。

【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の重みパラメーターを符号化するエンコーダー、デコーダー、方法、及び確率推定パラメーターを使用する符号化表現を含む。

【背景技術】

【0002】

以下では、本発明のより良い理解のための背景情報が提供される。しかしながら、背景情報に記載された特徴、機能及び詳細は、本発明の任意の実施形態と組み合わせて（個別に及び組み合わせての両方）任意選択で使用されてもよいことに留意すべきである。

【0003】

ニューラルネットワークは、例えば、その最も基本的な形態において、一連のアフィン変換とそれに続く要素ごとの非線形関数とを構成する。例えば、図1に示すように、それらは有向非巡回グラフとして表すことができる。

【0004】

図1は、フィードフォワードニューラルネットワークのグラフ表現の例を示している。具体的には、この2層ニューラルネットワークは、4次元入力ベクトルを実線に写像する非線形関数である。

【0005】

例えば、各ノードは、エッジのそれぞれの重み値との乗算によって次のノードに順方向伝搬される特定の値を伴う。次に、全ての入力値が単純に集約される。

【0006】

数学的には、図1のニューラルネットワークは例えば次の方法で出力を算出する。

$$\text{output} = L_2(L_1(\text{input}))$$

ここで、

$$L_i(X) = N_i(B_i(X))$$

であり、式中、 B_i は、例えば、層*i*のアフィン変換であり、 N_i は、例えば、層*i*の何らかの非線形関数である。 B_i の単純な例は、層*i*に関連付けられた重みパラメーター（エッジ重み） w_i と入力*X*との行列乗算である。

$$B_i(X) = w_i * X$$

演算子*は、行列乗算を示すものとする。

【0007】

10

20

30

40

50

いわゆる畳み込み層は、例えば、非特許文献 1 に記載されているように、それらを行列－行列積としてキャストすることによって使用することもできる。以降、所与の入力から出力を算出する手順を推論と称する。また、中間結果を、隠れ層又は隠れ活性化値と称し、これは、例えば上記の第 1 の内積＋非線形性の算出等、線形変換＋要素ごとの非線形性を構成する。

【0008】

以下では、ニューラルネットワークのバイアス及びバッチノルムについて説明する。ニューラルネットワーク層のアフィン変換のより高度な変形には、例えば以下のような、いわゆるバイアス及びバッチノルム演算が含まれる。

式 1：

【数 1】

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

式中、 b はバイアスを示し、 μ 、 σ^2 、 γ 、及び β はバッチノルムパラメーターを示す。 W は、例えば、次元 $n \times k$ を有する重み行列であり、 X は、例えば、次元 $k \times m$ を有する入力行列である。バイアス b 及びバッチノルムパラメーター μ 、 σ^2 、 γ 、及び β は、例えば、長さ n の転置ベクトルである。演算子 $*$ は、行列乗算を示す。ベクトルを有する行列に対する他の全ての演算（加算、乗算、除算）は、例えば、行列の列に対する要素ごとの演算であることに留意されたい。例えば、 $X \cdot \gamma$ は、 X の各列が γ と要素ごとに乗算されることを意味する。

【0009】

ϵ は、例えば、0 による除算を避けるために必要な（または、有用な）小さなスカラー数（0.001 等）である。ただし、0 であってもよい。 b の全てのベクトル要素が 0 に等しい場合、式 1 はバッチノルム層を指す。対照的に、 ϵ 並びに μ 及び β の全てのベクトル要素が 0 に設定され、 γ 及び σ^2 の全ての要素が 1 に設定される場合、バッチノルムのない層（バイアスのみ）が処理される。

【0010】

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

【0011】

別の結果として、ニューラルネットワークパラメーターの符号化及び／又は復号化は困難である。例えば、ニューラルネットワークの何百万ものパラメーターを伝送するために、大きな伝送レートが必要となる場合がある。

【先行技術文献】

【非特許文献】

【0012】

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

【発明の概要】

【発明が解決しようとする課題】

【0013】

したがって、圧縮、複雑さ、及び計算コストの間の良好な妥協を提供するニューラルネットワークパラメーターを符号化及び／又は復号化する改善された概念が必要である。

【課題を解決するための手段】

【0014】

これは、本出願の独立請求項の主題によって達成される。

【0015】

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

【0016】

本発明による実施の形態は、ニューラルネットワークの重みパラメーターを復号化するデコーダーを含み、デコーダーは、符号化ビットストリームに基づいて、複数のニューラルネットワークパラメーター、例えば、ニューラルネットワークの行列 W の成分 w_i 、 b 、 μ 、 σ^2 、 σ 、 γ 、及び／又は β のうちの少なくとも1つを取得するように構成される。さらに、デコーダーは、コンテキスト依存算術復号化を使用して、例えば、コンテキスト適応型二値算術復号化（CABAC：context-dependent arithmetic decoding）を使用して、ニューラルネットワークパラメーター、例えば、ニューラルネットワークの行列 W の成分 w_i 、又は b 、又は μ 、又は σ^2 、又は σ 、又は γ 、又は β 、例えば、ニューラルネットワークパラメーターの量子化バージョンを復号化するように構成される。任意選択で、ピン値の確率は、異なるコンテキストについて決定されてもよく、例えば、各ピンはコンテキストに関連付けられる。さらに、デコーダーは、1つ以上の確率推定パラメーター、例えば確率推定器パラメーター、例えば N 、 a_i^k 、 b_i^k 、 a_k 、 d_i^k 、 A 、 m_i^k 、 n_i^k 、 sh_i^k 、 $initVal_i^k$ を使用して、例えば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ー又はそのピンに基づいて、ニューラルネットワークパラメーターの数表現のピンの例えば算術復号化のために、例えばコンテキストに関連付けられ得る確率推定値、例えば $P(t)$ 又は p_k を取得するように構成される。加えて、デコーダーは、異なるニューラルネットワークパラメーターの復号化のために異なる確率推定パラメーター値を使用するように、及び／又は異なるコンテキストモデル、例えば c_k に関連付けられたピンの復号化のために異なる確率推定パラメーター値を使用するように構成される。

【0017】

本発明による更なる実施の形態は、ニューラルネットワークの重みパラメーターを復号化するデコーダーを含み、デコーダーは、符号化ビットストリームに基づいて、複数のニューラルネットワークパラメーター、例えば、ニューラルネットワークの行列 W の成分 w_i 、 b 、 μ 、 σ^2 、 σ 、 γ 、及び／又は β のうちの少なくとも1つを取得するように構成される。さらに、デコーダーは、コンテキスト依存算術復号化を使用して、例えば、コンテキスト適応型二値算術復号化（CABAC）を使用して、ニューラルネットワークパラメーター、例えば、ニューラルネットワークの行列 W の成分 w_i 、又は b 、又は μ 、又は σ^2 、又は σ 、又は γ 、又は β 、例えば、ニューラルネットワークパラメーターの量子化バージョンを復号化するように構成される。任意選択で、ピン値の確率は、異なるコンテキストについて決定されてもよく、例えば、各ピンはコンテキストに関連付けられる。さらに、デコーダーは、1つ以上の確率推定パラメーター、例えば確率推定器パラメーター、例えば N 、 a_i^k 、 b_i^k 、 a_k 、 d_i^k 、 A 、 m_i^k 、 n_i^k 、 sh_i^k 、 $initVal_i^k$ を使用して、例えば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ー又はそのピンに基づいて、ニューラルネットワークパラメーターの数表現のピンの例えば算術復号化のために、例えばコンテキストに関連付けられ得る確率推定値、例えば $P(t)$ 又は p_k を取得するように構成される。加えて、デコーダーは、ニューラルネットワークの異なる層に関連付けられたニューラルネットワークパラメーターの復号化のために異なる確率推定パラメーター値を使用するように構成される。

【0018】

本発明による実施の形態は、ニューラルネットワークの重みパラメーターを符号化するエンコーダーを含み、エンコーダーは、複数のニューラルネットワークパラメーター、例えば、ニューラルネットワークの行列 W の成分 w_i 、 b 、 μ 、 σ^2 、 σ 、 γ 、及び／又は β のうちの少なくとも1つを取得するように構成される。さらに、エンコーダーは、コン

テキスト依存算術符号化を使用して、例えば、コンテキスト適応型二値算術符号化（C A B A C）を使用して、ニューラルネットワークパラメーター、例えば、ニューラルネットワークの行列Wの成分 w_i 、又は b 、又は μ 、又は σ^2 、又は σ 、又は γ 、又は β 、例えば、その量子化バージョンを符号化するように構成される。任意選択で、ピン値の確率は、異なるコンテキストについて決定されてもよく、例えば、各ピンはコンテキストに関連付けられる。加えて、エンコーダーは、1つ以上の確率推定パラメーター、例えば確率推定器パラメーター、例えば N 、 a_{j^k} 、 b_{j^k} 、 a_k 、 d_{j^k} 、 A 、 m_{j^k} 、 n_{j^k} 、 sh_{j^k} 、 $initVal_{j^k}$ を使用して、例えば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ー又はそのピンに基づいて、ニューラルネットワークパラメーターの数表現のピンの任意選択で算術符号化のために、例えばコンテキストに関連付けられ得る確率推定値、例えば $P(t)$ 又は p_k を取得するように構成される。さらに、エンコーダーは、異なるニューラルネットワークパラメーターの符号化のために異なる確率推定パラメーター値を使用するように、及び／又は異なるコンテキストモデル、例えば c_k に関連付けられたピンの符号化のために異なる確率推定パラメーター値を使用するように構成される。

【0019】

本発明による更なる実施の形態は、ニューラルネットワークの重みパラメーターを符号化するエンコーダーを含み、エンコーダーは、複数のニューラルネットワークパラメーター、例えば、ニューラルネットワークの行列Wの成分 w_i 、 b 、 μ 、 σ^2 、 σ 、 γ 、及び／又は β を取得するように構成される。さらに、エンコーダーは、コンテキスト依存算術符号化を使用して、例えば、コンテキスト適応型二値算術符号化（C A B A C）を使用して、ニューラルネットワークパラメーター、例えば、ニューラルネットワークの行列Wの成分 w_i 、又は b 、又は μ 、又は σ^2 、又は σ 、又は γ 、又は β 、例えば、その量子化バージョンを符号化するように構成される。任意選択で、ピン値の確率は、異なるコンテキストについて決定されてもよく、例えば、各ピンはコンテキストに関連付けられる。さらに、エンコーダーは、1つ以上の確率推定パラメーター、例えば確率推定器パラメーター、例えば N 、 a_{j^k} 、 b_{j^k} 、 a_k 、 d_{j^k} 、 A 、 m_{j^k} 、 n_{j^k} 、 sh_{j^k} 、 $initVal_{j^k}$ を使用して、例えば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ー又はそのピンに基づいて、ニューラルネットワークパラメーターの数表現のピンの任意選択で算術符号化のために、例えばコンテキストに関連付けられ得る確率推定値、例えば $P(t)$ 又は p_k を取得するように構成される。加えて、エンコーダーは、ニューラルネットワークの異なる層に関連付けられたニューラルネットワークパラメーターの符号化のために異なる確率推定パラメーター値を使用するように構成される。

【0020】

本発明による実施の形態は、ニューラルネットワークの重みパラメーターを復号化する方法を含み、方法は、符号化ビットストリームに基づいて、複数のニューラルネットワークパラメーター、例えば、ニューラルネットワークの行列Wの成分 w_i 、 b 、 μ 、 σ^2 、 σ 、 γ 、及び／又は β のうちの少なくとも1つを取得することを含み、方法は、コンテキスト依存算術復号化を使用して、例えば、コンテキスト適応型二値算術復号化（C A B A C）を使用して、ニューラルネットワークパラメーター、例えば、ニューラルネットワークの行列Wの成分 w_i 、又は b 、又は μ 、又は σ^2 、又は σ 、又は γ 、又は β 、例えば、ニューラルネットワークパラメーターの量子化バージョンを復号化することを含む。任意選択で、ピン値の確率は、異なるコンテキストについて決定され、例えば、各ピンはコンテキストに関連付けられ得る。さらに、本方法は、1つ以上の確率推定パラメーター、例えば確率推定器パラメーター、例えば N 、 a_{j^k} 、 b_{j^k} 、 a_k 、 d_{j^k} 、 A 、 m_{j^k} 、 n_{j^k} 、 sh_{j^k} 、 $initVal_{j^k}$ を使用して、例えば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ー又はそのピンに基づいて、ニューラルネットワークパラメーターの数表現のピンの例えば算術復号化のために、例えばコンテキストに関連付けられ得る確率推定値、例えば $P(t)$ 又は p_k を取得することを含む。加えて、本方法は、異なるニューラルネットワークパラメーターの復号化のために異なる確率推定パラメータ

一値を使用すること、及び／又は異なるコンテキストモデル、例えば c_k に関連付けられたビンの復号化のために異なる確率推定パラメータ値を使用することを含む。

【0021】

本発明による更なる実施の形態は、ニューラルネットワークの重みパラメータを復号化する方法を含み、方法は、符号化ビットストリームに基づいて、複数のニューラルネットワークパラメータ、例えば、ニューラルネットワークの行列 W の成分 w_i 、 b 、 μ 、 σ^2 、 σ 、 γ 、及び／又は β のうちの少なくとも1つを取得することを含む。さらに、本方法は、コンテキスト依存算術復号化を使用して、例えば、コンテキスト適応型二値算術復号化 (CABAC) を使用して、ニューラルネットワークパラメータ、例えば、ニューラルネットワークの行列 W の成分 w_i 、又は b 、又は μ 、又は σ^2 、又は σ 、又は γ 、又は β 、例えば、ニューラルネットワークパラメータの量子化バージョンを復号化することを含む。任意選択で、ビン値の確率は、異なるコンテキストについて決定されてもよく、例えば、各ビンはコンテキストに関連付けられる。さらに、本方法は、1つ以上の確率推定パラメータ、例えば確率推定器パラメータ、例えば N 、 a_j^k 、 b_j^k 、 a_k 、 d_j^k 、 A 、 m_j^k 、 n_j^k 、 sh_j^k 、 $initVal_j^k$ を使用して、例えば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又はそのビンに基づいて、ニューラルネットワークパラメータの数表現のビンの例えば算術復号化のために、例えばコンテキストに関連付けられ得る確率推定値、例えば $P(t)$ 又は p_k を取得することを含む。加えて、本方法は、ニューラルネットワークの異なる層に関連付けられたニューラルネットワークパラメータの復号化のために異なる確率推定パラメータ値を使用することを含む。

10

20

【0022】

本発明による実施の形態は、ニューラルネットワークの重みパラメータを符号化する方法を含み、方法は、複数のニューラルネットワークパラメータ、例えば、ニューラルネットワークの行列 W の成分 w_i 、 b 、 μ 、 σ^2 、 σ 、 γ 、及び／又は β のうちの少なくとも1つを取得することを含む。さらに、本方法は、コンテキスト依存算術符号化を使用して、例えば、コンテキスト適応型二値算術符号化 (CABAC) を使用して、ニューラルネットワークパラメータ、例えば、ニューラルネットワークの行列 W の成分 w_i 、又は b 、又は μ 、又は σ^2 、又は σ 、又は γ 、又は β 、例えば、ニューラルネットワークパラメータの量子化バージョンを符号化することを含む。任意選択で、ビン値の確率は、異なるコンテキストについて決定してもよく、例えば、各ビンはコンテキストに関連付けられ得る。さらに、本方法は、1つ以上の確率推定パラメータ、例えば確率推定器パラメータ、例えば N 、 a_j^k 、 b_j^k 、 a_k 、 d_j^k 、 A 、 m_j^k 、 n_j^k 、 sh_j^k 、 $initVal_j^k$ を使用して、例えば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又はそのビンに基づいて、ニューラルネットワークパラメータの数表現のビンの例えば算術符号化のために、例えばコンテキストに関連付けられ得る確率推定値、例えば $P(t)$ 又は p_k を取得することを含む。加えて、本方法は、異なるニューラルネットワークパラメータの符号化のために異なる確率推定パラメータ値を使用すること、及び／又は異なるコンテキストモデル、例えば c_k に関連付けられたビンの符号化のために異なる確率推定パラメータ値を使用することを含む。

30

40

【0023】

本発明による更なる実施の形態は、ニューラルネットワークの重みパラメータを符号化する方法を含み、方法は、複数のニューラルネットワークパラメータ、例えば、ニューラルネットワークの行列 W の成分 w_i 、 b 、 μ 、 σ^2 、 σ 、 γ 、及び／又は β のうちの少なくとも1つを取得することを含む。さらに、本方法は、コンテキスト依存算術符号化を使用して、例えば、コンテキスト適応型二値算術符号化 (CABAC) を使用して、ニューラルネットワークパラメータ、例えば、ニューラルネットワークの行列 W の成分 w_i 、又は b 、又は μ 、又は σ^2 、又は σ 、又は γ 、又は β 、例えば、ニューラルネットワークパラメータの量子化バージョンを符号化することを含む。任意選択で、ビン値の確率は、異なるコンテキストについて決定してもよく、例えば、各ビンはコンテキストに関

50

連付けられ得る。さらに、本方法は、1つ以上の確率推定パラメーター、例えば確率推定器パラメーター、例えば N 、 a_i^k 、 b_i^k 、 a_k 、 d_i^k 、 A 、 m_i^k 、 n_i^k 、 $s_{h_i}^k$ 、 $initVal_i^k$ を使用して、例えば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ー又はそのピンに基づいて、ニューラルネットワークパラメーターの数表現のピンの例えば算術符号化のために、例えばコンテキストに関連付けられ得る確率推定値、例えば $P(t)$ 又は p_k を取得することを含む。加えて、本方法は、ニューラルネットワークの異なる層に関連付けられたニューラルネットワークパラメーターの符号化のために異なる確率推定パラメーター値を使用することを含む。

【0024】

本発明による実施の形態は、コンピュータプログラムであって、該コンピュータプログラムがコンピュータ上で動作するときに本発明による方法を実行する、コンピュータプログラムを含む。

10

【0025】

本発明による実施の形態は、ニューラルネットワークの重みパラメーターの符号化表現を含み、該符号化表現は、ニューラルネットワークの複数の符号化重みパラメーターと、符号化重みパラメーターの算術復号化のコンテキストの適応のための確率推定の特性を決定する1つ以上の確率推定パラメーターの符号化表現とを含む。

【0026】

本発明の実施の形態の主要な概念をより良く理解するために、以下では、本発明によるニューラルネットワークパラメーターの符号化及び復号化の更なる任意選択の態様が開示される。第1に、とりわけ、実施の形態によるパラメーターの効率的な表現が開示される。以下で説明される詳細は任意選択である。

20

【0027】

パラメーター W 、 b 、 μ 、 σ^2 、 γ 、及び β は、層のパラメーター又は層パラメーターと総称され得る。これらのパラメーターのうちの1つ以上は、前述したように、ニューラルネットワークパラメーターの例であり得る。それらは通常、ビットストリームにおいて（例えば、ニューラルネットワークがビデオデコーダーにおいて使用される場合、例えば、符号化されたビデオ表現において）シグナリングされる必要がある。例えば、それらは32ビットの浮動小数点数として表すことができ、又は例えば、量子化インデックスとしても示される整数表現に量子化することができる。 ε は通常、ビットストリーム中でシグナリングされないことに留意されたい。

30

【0028】

例えば、かかるパラメーターを符号化するための特に効率的な手法では、例えば、各値がいわゆる量子化ステップサイズ値の整数倍として表される、一様再構成量子化器（URQ：uniform reconstruction quantizer）が用いられる。対応する浮動小数点数は、例えば、整数に量子化ステップサイズを乗算することによって再構成することができ、量子化ステップサイズは、通常（必ずしもそうではないが）単一の浮動小数点数である。しかしながら、例えば、ニューラルネットワーク推論のための効率的な実施態様（すなわち、入力に対するニューラルネットワークの出力を計算すること）では、可能な限り整数演算が採用されている。したがって、パラメーターを浮動小数点表現に再構成することを要求することは望ましくない場合がある。

40

【0029】

パラメーターを符号化するための別の効率的な手法では、量子化器のセットが適用され、各値は、例えば、量子化ステップサイズ値の整数倍として表される。通常、例えば、集合内の各量子化器は、適用可能な再構成値として量子化ステップサイズパラメーターの整数倍の互いに素な集合を使用するが、2つ以上の量子化器が1つ以上の再構成値を共有してもよい。適用される量子化器は、例えば、符号化順で前の量子化インデックスの値に依存する。対応する浮動小数点数は、例えば、整数に量子化ステップサイズを乗算することによって再構成することができ、量子化ステップサイズは、通常、例えば、選択された量子化器に依存する浮動小数点数である。かかる量子化器の設計の一例は、トレリス符号化

50

量子化（TCQ：trellis coded quantization）であり、これは、従属量子化（DQ：dependent quantization）とも称される。

【0030】

好ましい実施の形態では、2つの量子化器のセットが使用される。第1の量子化器は、例えば、0を含む量子化ステップサイズの全ての偶数倍を使用し、第2の量子化器は、0を含む量子化ステップサイズの全ての偶数倍を使用する。

【0031】

次に、とりわけ、実施の形態によるエントロピー符号化及び確率推定が開示される。以下で説明される詳細は、任意選択であり、特に、パラメータの効率的な表現に関して、上記で説明された特徴と組み合わせ可能である。

【0032】

例えば上述したような量子化方法によって出力された量子化インデックスは、適切なエントロピー符号化方法を用いてエントロピー符号化されてもよい。

【0033】

かかる量子化インデックスを符号化するための特に適切なエントロピー符号化方法は、CABACとも表されるコンテキストベース適応型二値算術符号化（Context-based Adaptive Binary Arithmetic Coding）である。これにより、各量子化インデックスは、例えば、二値決定のシーケンス、いわゆるピンに分解される。通常、例えば、各ピンは、コンテキストモデルとも称される確率モデルに関連付けられ、該コンテキストモデルは、例えば確率推定方法を使用して、関連付けられたピンの統計をモデル化する。

【0034】

確率推定器は、例えば、確率推定器に関連する既に符号化されたピンに基づいて、ピンが x に等しい確率 $P(t)$ をモデル化する装置であり、ここで、 $x \in \{0, 1\}$ である。 $P(t)$ は、確率推定値の一例であり得る。

【0035】

次に、例えば典型的な推定器の設計について説明する。推定器の設計の詳細は、本発明による実施の形態の任意選択の特徴であり、特に、上記で説明した特徴を含む実施の形態と組み合わせることができる。

【0036】

最初に、ニューラルネットワーク圧縮に適用される典型的な推定器の設計について説明する。

【0037】

例えば、各コンテキストモデル c_k について、1つ以上の状態変数 $s_1^k, \dots, s_N^k$ が $N \geq 1$ で維持される。各状態変数 s_i^k は、例えば、符号付き整数値として実装され、例えば、確率値 $P(s_i^k, i, k) = p_i^k$ を表す。コンテキストモデル c_k の確率推定値 p_k は、例えば、コンテキストモデルの全ての状態変数の確率値 p_i^k の加重和として定義されるものとする。

【0038】

状態変数は、以下の特性を有することが好ましいが、必ずしもそうである必要はない。

1. $s_i^k = 0$ である場合、 $p_i^k = 0.5$ である。
2. s_i^k のより大きな値は、より大きな p_i^k に対応する。
3. $P(-s_i^k, i, k) = 1 - P(s_i^k, i, k)$ 。

【0039】

したがって、負の状態変数は、例えば、 $p_i^k < 0.5$ に対応することができる。概して、各コンテキストモデルの各状態変数に対して異なる関数 $P(\cdot)$ を指定することが可能である。

【0040】

次に、状態変数と確率値とを関連付ける構成例について説明する。状態の関連付けの詳細は、本発明による実施の形態の任意選択の特徴であり、特に、上記で説明した特徴を含む実施の形態と組み合わせることができる。

10

20

30

40

50

【0041】

状態変数を確率値に関連付ける、すなわち $P(\cdot)$ を実装する多くの有用な方法が存在する。例えば、ニューラルネットワーク圧縮で使用される状態表現は、以下の式で達成することができる。

【数2】

$$P(x, i, k) = \begin{cases} 0.5 \cdot \alpha^{|x \cdot \beta_i^k|}, & x \geq 0 \text{ である場合} \\ 1 - 0.5 \cdot \alpha^{|x \cdot \beta_i^k|}, & \text{それ以外の場合} \end{cases}$$

10

β_i^k は重み係数である。

α は、 $0 < \alpha < 1$ のパラメータである。

【0042】

例えば、2つの状態 ($N=2$, s_1^k , s_2^k) を使用する、マルチメディアコンテンツ記述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 規格の現在のドラフトで使用されるものに匹敵する構成を達成するために、 $\alpha = 0.99894079$ 並びに全ての k に対して $\beta_1^k = 16$ 及び $\beta_2^k = 1$ を設定する。

【0043】

20

この例示的な構成は、状態変数がどのように定義され得るかについての何らかの洞察を与えるものである。以下に見られるように、一般に、 $P(\cdot)$ は直接使用されないので定義する必要はない。むしろ、個々の部分の実際の実施態様の結果として得られることが多い。

【0044】

次に、状態変数の初期化について説明する。初期化の詳細は、本発明による実施の形態の任意選択の特徴であり、特に、上記で説明した特徴を含む実施の形態と組み合わせることができる。

【0045】

コンテキストモデルを用いて第1のシンボルを符号化又は復号化する前に、全ての状態変数は、 $initVal_i^k$ として示される同じ値で初期化され、これは、任意選択で、例えば、圧縮アプリケーションに最適化することができる。

30

【0046】

次に、状態変数からの確率推定値の導出について説明する。導出の詳細は、本発明による実施の形態の任意選択の特徴であり、特に、上記で説明した特徴を含む実施の形態と組み合わせることができる。

【0047】

シンボルの符号化又は復号化のために、コンテキストモデルの状態変数から確率推定値が導出される。例として、3つの代替的な手法を以下に提示する。手法1は、手法2及び手法3よりも正確な結果をもたらすが、計算の複雑さも高くなる。

40

【0048】

手法1 (例)

この手法は2つのステップからなる。まず、コンテキストモデルの各状態変数 s_i^k が確率値 p_i^k に変換される。次に、確率推定値 p_k が、確率値 p_i^k の加重和として導出される。

【0049】

ステップ1:

ルックアップテーブル LUT1 は、例えば式(1)に従って、状態変数 s_i^k を対応する確率値 p_i^k に変換するために使用される。

【数3】

50

$$p_i^k = \begin{cases} LUT1[[s_i^k \cdot a_i^k]], & s_i^k \geq 0 \text{である場合} \\ 1 - LUT1[[-s_i^k \cdot a_i^k]], & \text{それ以外の場合} \end{cases} \quad (1)$$

LUT1は、確率値を含むルックアップテーブルである。 a_i^k は、 s_i^k をLUT1のサイズに適応させる重み係数である。

【0050】

ステップ2：

確率推定値 p_k は、例えば次式に従って、確率値 p_i^k から導出される。

【数4】

$$p_k = \sum_{i=1}^N p_i^k \cdot b_i^k \quad (2)$$

b_i^k は、個々の状態変数の影響を制御する重み係数である。

【0051】

手法2（例）

状態変数から確率推定値を導出するための代替的な手法を以下に示す。該手法では、精度の低い結果が得られ、計算の複雑さが低くなる。最初に、状態変数の加重和 s_k が、例えば、次式に従って導出される。

【数5】

$$s_k = \sum_{i=1}^N [s_i^k \cdot d_i^k] \quad (3)$$

d_i^k は、各状態変数の影響を制御する重み係数である。

【0052】

次に、確率推定値 p_k は、状態変数 s_k の加重和から、例えば次式に従って導出される。

【数6】

$$p_k = \begin{cases} LUT2[[s_k \cdot a_k]], & s_k \geq 0 \text{である場合} \\ 1 - LUT2[[-s_k \cdot a_k]], & \text{それ以外の場合} \end{cases} \quad (4)$$

LUT2は、確率推定値を含むルックアップテーブルであり、 a_k は、 s_k をLUT2のサイズに適応させる重み係数である。

【0053】

手法3（例）

状態変数から確率推定値を導出するための更に別の手法を以下に提示する。まず、状態変数の重み付き和 s_k を、例えば手法2のように導出する。次に、確率推定値 p_k は、状態変数 s_k の加重和から、例えば次式に従って導出される。

【数7】

10

20

30

40

50

$$p_k = \begin{cases} LUT2[\lfloor s_k \cdot a_k \rfloor], & \lfloor s_k \cdot a_k \rfloor \geq 0 \text{ である場合} \\ 1 - LUT2[-\lfloor s_k \cdot a_k \rfloor], & \text{それ以外の場合} \end{cases} \quad (5)$$

LUT2は、確率推定値を含むルックアップテーブルである。

【0054】

手法4（例）

更なる手法では、状態値と確率 $P(x, i, k)$ との間の線形関係が用いられる。確率推定値の導出は、例えば、式（2）の手法を使用する。手法4の一例は、汎用ビデオ符号化（VVC：Versatile Video Coding）の現在のドラフトで使用されている確率推定方式である。

【0055】

例えば、マルチメディアコンテンツの記述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規格の現在のドラフトで使用される構成を達成するために、手法3の方法が、例えば、全ての k について $d_1^k = 16$ 、 $d_2^k = 1$ 、及び $a_k = 2^{-7}$ を用いて使用される。確率推定値を含むルックアップテーブルは、例えば、以下の通りである。

【0056】

LUT2 = {0.5000, 0.4087, 0.3568, 0.3116, 0.2721, 0.2375, 0.2074, 0.1811, 0.1581, 0.1381, 0.1206, 0.1053, 0.0919, 0.0803, 0.0701, 0.0612, 0.0534, 0.0466, 0.0407, 0.0356, 0.0310, 0.0271, 0.0237, 0.0207, 0.0180, 0.0158, 0.0138, 0.0120, 0.0105, 0.0092, 0.0080, 0.0070}

【0057】

次に、状態変数の更新について説明する。更新の詳細は、本発明による実施の形態の任意選択の特徴であり、特に、上記で説明した特徴を含む実施の形態と組み合わせることができる。

【0058】

シンボルの符号化又は復号化の後、コンテキストモデルの1つ以上の状態変数は、シンボルシーケンスの統計的挙動を追跡するために更新され得る。

【0059】

更新は、例えば、以下のように行われる。

【数8】

$$s_i^k = \begin{cases} s_i^k + \lfloor A[z + \lfloor s_i^k \cdot m_i^k \rfloor] \cdot n_i^k \rfloor, & \text{符号化されるシンボルが1である場合} \\ s_i^k - \lfloor A[z + \lfloor -s_i^k \cdot m_i^k \rfloor] \cdot n_i^k \rfloor, & \text{符号化されるシンボルが0である場合} \end{cases} \quad (6)$$

Aは、例えば整数値を格納するルックアップテーブルである。 m_i^k 及び n_i^k は、例えば、更新の「アジリティ（agility）」を制御する重み係数である。係数 n_i^k は、例えば、

【数9】

$$n_i^k = 2^{-sh_i^k+4}$$

に従って記述することができ、ここで、 sh_i^k は適応パラメーターとしても示される。 z は、例えば、ルックアップテーブルAが負でない値でのみアクセスされることを保証するオフセットである。

【0060】

ルックアップテーブルA内の値は、例えば、 s_i^k が特定の所与の間隔内に留まるように選択することができる。通常、ルックアップAの値は、例えば更新関数に近似している。代替的に、例えば、状態更新のために関連する更新関数を単に使用することも可能である。

【0061】

例えば、手法4に従うVVCの推定方法は、状態更新のための更新関数を適用し、ビットシフトを使用するものであり、該ビットシフトにより、例えば更新の「アジリティ」が決定される。これは、例えば、上述の適応パラメーターに対応する。本発明の実施の形態（下記参照、例えば、実施の形態の主要な概念に従って後述する）は、同様にそれらに適用され得る。

【0062】

例えば、マルチメディアコンテンツの記述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規格の現在のドラフトで使用される構成を達成するために、パラメーターは、例えば、全ての k に対して、 $m_1^k = 2^{-3}$ 、 $m_2^k = 2^{-7}$ 及び $n_1^k = 2^{-1}$ 、 $n_2^k = 1$ 、並びに $z = 16$ となるように選択される。ルックアップテーブルAは、例えば、 $A = \{157, 143, 129, 115, 101, 87, 73, 59, 45, 35, 29, 23, 17, 13, 9, 5, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 0\}$ である。

【0063】

シンボルを符号化する前に、 s_1^k は、例えば、間隔 $[-127, 127]$ からの値で初期化されるものとし、 s_2^k は、例えば、間隔 $[-2047, 2047]$ からの値で初期化されるものとする。その結果、 s_1^k は、例えば、8ビットの符号付き整数値で実装することができ、 s_2^k は、例えば、12ビットの符号付き整数で実装することができる。

【0064】

上記の例で説明したように、確率推定器は、確率推定器パラメーター又は推定器パラメーター（又は確率推定パラメーター）と称される一部のパラメーターを有し、これらのパラメーターは、確率推定値、例えば適応レートに影響を与える。通常、これらの推定器パラメーターは、適用シナリオ、例えば、ニューラルネットワークパラメーターの符号化に応じて、グローバルに選択される。したがって、例えば、ニューラルネットワーク符号化では、各ニューラルネットワークパラメーターは、推定器パラメーターの同じ集合を適用する。しかし、本発明者らは、現在のニューラルネットワークパラメーターに対して最適化された推定器パラメーターを選択することによって圧縮効率が改善され得ることを見出した。したがって、一態様によれば、基本的な着想は、パラメーターの集合から適切な推定器パラメーターを選択し、次いで、推定器パラメーターがデコーダーにシグナリングされることである。

【0065】

換言すれば、本発明による（例えば、独立請求項によって定義されるような）実施の形態は、ニューラルネットワークパラメーター又はエンティティ、例えば、ニューラルネットワークパラメーターに関連付けられた異なるコンテキストに関連付けられたビンの復号化及びそれぞれの符号化のために異なる確率推定パラメーター値を使用するという着想に基づく。確率推定パラメーター又は確率パラメーター推定値の基集合（base set）から1

10

20

30

40

50

つの固定されたインスタンスを使用する代わりに、確率推定パラメータ値の適格的な、又は例えば最適な個々の選択が実行されてもよい。確率推定値の選択は、任意の適切な基準、例えば、それぞれ、或る特定のニューラルネットワークパラメータの事前に既知の特性、或る特定のニューラルネットワークパラメータのコンテキストモデルの確率的特徴、又はニューラルネットワークパラメータのビンの確率的特徴に基づいて実行され得る。例えば、異なるニューラルネットワークパラメータは、異なる相関特性を含んでもよく、これらの異なる相関特性への確率推定パラメータの適応により、符号化効率を向上させ得ることが認識されている。例えば、ニューラルネットワークによって実現される機能によって、ニューラルネットワークの隣接するブランチ又はエッジ間の相関強度（ニューラルネットワークパラメータ）が異なる場合がある。また、ニューラルネットワークのエッジ（又はブランチ）間の相関関係の拡張も異なり得る。本発明の一態様によれば、確率推定パラメータ値は、ニューラルネットワークの相関特性（相関強度／相関の拡張）に（例えば、動的に、例えば、単一のニューラルネットの異なるニューラルネットワークパラメータの符号化の間で）適応し、それにより、特に効率的な符号化をもたらすことができる。別の例として、ニューラルネットワークの異なる層に関連付けられたニューラルネットワークパラメータの相関特性は、場合によっては実質的に異なることが認識されている。例えば、ニューラルネットワークの異なる層が異なるサイズ（又は幅）の畳み込みを反映する場合、これは、異なるニューラルネットワーク層のニューラルネットワークパラメータの符号化又は復号化のために異なる確率推定パラメータ値を使用することによって達成され、その結果、符号化効率を向上させることができる。

10

20

【0066】

結論として、異なるニューラルネットワークパラメータを符号化又は復号化するとき、又は異なるコンテキストモデルに関連付けられたビンを符号化又は復号化するとき異なる確率推定パラメータを使用すると、符号化効率、複雑さ、及びリソース使用量の間のトレードオフが改善されることが分かっている。例えば、実際に使用される確率推定パラメータ値のシグナリング（又は動的シグナリング）に必要とされ得るオーバーヘッドは、符号化効率の増加によって過剰補償され得ることが認識されている。

【0067】

本発明の実施の形態による第2の態様は、確率推定パラメータの符号化及び／又は復号化が、ニューラルネットワークパラメータの符号化及び／又は復号化と同様のステップ又は同等のステップを含み得ることである。

30

【0068】

本発明者らは、確率推定パラメータ、例えば確率推定器及び／又はコンテキストモデルに関係する全てのパラメータが適応的に選択される場合、符号化／復号化性能が改善され得ることを認識した。確率推定器パラメータは、ニューラルネットワークパラメータと類似のステップを使用して符号化／復号化することができる。例えば、確率推定パラメータ、又は確率パラメータを表す整数インデックス q は、ビンのシーケンス、例えばビン `greaterThan_0`、`greaterThan_1`、...、及びニューラルネットワークパラメータに類似したコンテキストモデルを使用して符号化／復号化されてもよい。

40

【0069】

換言すれば、デコーダーについてはその逆となり、本発明の実施の形態によるデコーダーは、符号化ビットストリームを取得することができ、ビットストリームはニューラルネットワークパラメータを含む。ニューラルネットワークパラメータは、ビンのシーケンスとしてビットストリーム中で符号化され得る。これらのビンは、コンテキスト依存算術復号化、例えば、ニューラルネットワークパラメータを表すビンビットストリームに符号化するために使用されるのと同じコンテキスト依存算術符号化の方法を使用して復号化され得る。したがって、デコーダーは、1つ以上の確率推定パラメータ、例えば、前述の確率推定器パラメータのうちの少なくとも1つ、例えば、 N 、 a_i^k 、 b_i^k 、 a_k 、 d_i^k 、 A 、 m_i^k 、 n_i^k 、 sh_i^k 、 $initVal_i^k$ のうちの少なくとも1つを

50

含む確率推定器を使用して確率推定値を取得するように構成される。デコーダーは、次いで、異なるニューラルネットワークパラメーターの復号化のために異なる確率推定パラメーター値を、及び／又は異なるコンテキストモデル、例えば c_k に関連付けられたビンの復号化のために異なる確率推定パラメーター値を使用してもよい。さらに、デコーダーは、ニューラルネットワークの異なる層に関連付けられたニューラルネットワークパラメーターの復号化のために異なる確率推定パラメーター値を使用することができる。ニューラルネットワークパラメーター、例えば、それらのビンの層又はコンテキストモデルの或る特定の特性に従って確率推定パラメーター値を適応させることによって、復号化及び符号化の効率を向上させることができる。

【0070】

換言すれば、一態様によれば、パラメーター、すなわち、 N 、 a_{i^k} 、 b_{i^k} 、 a_k 、 d_{i^k} 、 A 、 m_{i^k} 、 n_{i^k} 、 sh_{i^k} 、 $initVal_{i^k}$ 及び／又は確率推定器、例えばコンテキストモデルに関係する任意の他のパラメーターは、確率推定器パラメーター又は推定器パラメーター（又は確率推定パラメーター）と総称され得る。

【0071】

通常、例えば、各推定器パラメーターに対して、確率推定器パラメーターの基集合から1つの固定されたインスタンスがネットワーク全体に対して選択される。基集合の値は、適用される状態の数 N に従って、 N 組 (N -tuple) の推定器パラメーターであり得る。本発明の一態様によれば、確率推定、したがって圧縮効率は、例えば、パラメーターが、層（すなわち、 W 、 b 、 μ 、 σ^2 、 γ 、及び β ）及び／又はコンテキストモデル c_k の各パラメーター又はパラメーターの部分集合に対して個々に選択される場合に改善され得る。

【0072】

使用される推定器パラメーターは、例えば、パラメーターの集合のパラメーターの中から決定され、パラメーターの集合は、例えば、基集合又は基集合の任意の部分集合であり得る。この集合の各パラメーターは、例えば、整数インデックス q に関連付けられ得る。例えば、集合の1つのパラメーターがデフォルトパラメーターとして示されてもよい。通常、デフォルトパラメーターは、例えば、0 に等しい整数インデックスに関連付けられる。選択された推定器パラメーターに関連付けられたインデックスは、次いで、例えば、デコーダーにシグナリングされる。

【0073】

本発明の実施の形態によれば、デコーダーは、基集合から、又は基集合の真の部分集合から、1つ以上の確率推定パラメーターを選択するように構成される。任意選択で、基集合は、1つ以上の確率推定パラメーターに関連付けられた複数の使用可能なパラメーター値を含むことができ、又は基集合は、複数の確率推定パラメーターに関連付けられた使用可能なパラメーター値の複数のタプル、例えば、使用可能な対のリスト (sh_{1^k} 、 sh_{2^k}) を含むことができる。

【0074】

確率推定パラメーターの部分集合又は更には真の部分集合を提供することによって、例えば1つ以上の基準に関して選択しやすい、許容可能な確率推定パラメーターの対、タプル及び／又は集合が提供され得る。例えば、ニューラルネットワークの或る特定の層によれば、部分集合は、改善された復号化効率を提供する確率推定パラメーターを選択するために、例えば層 ID を事前に定義されたリストと単に比較することによって、低い計算労力で選択されてもよい。加えて、かかる部分集合の選択は、或る特定の種類の量子化に従って実行され得る。部分集合を提供することによって、多大な割り当て労力を伴うことなく、有利な確率推定パラメーターを提供することができる。複数の条件、例えば各確率推定パラメーターについて1つの条件をチェックする代わりに、1つのチェック、例えば層 ID 又はニューラルネットワークパラメーターのビンのコンテキストモデルと所定のリスト又は基準との前述の比較を行い、それに基づいて確率推定パラメーター又は確率推定パラメーター値の集合全体を選択することができる。また、基集合を提供することによって、例えば、実際のパラメーター値ではなく、基集合のインデックスをシグナリングするこ

10

20

30

40

50

とで十分となり得るため、シグナリングの労力が軽減される。

【0075】

本発明の実施の形態によれば、デコーダーは、量子化モードに応じて、使用可能なパラメーター値の異なる集合又はパラメーター値の使用可能なタプルの異なる集合から1つ以上の確率推定パラメーターを選択するように構成され、又はデコーダーは、量子化モードに応じて、1つ以上の確率推定パラメーターを表す符号化値、例えば符号化インデックス値 q を1つ以上の確率推定パラメーター、例えば $s h_i^k$ にマッピングする、異なるマッピング規則、例えば異なるマッピングテーブルを使用するように構成される。

【0076】

任意選択で、デコーダーは、第1の量子化モード、例えば一様量子化URQが使用される場合に使用可能なパラメーター値の第1の集合又はパラメーター値の使用可能なタプルの第1の集合から1つ以上の確率推定パラメーターを選択することができ、デコーダーは、第2の量子化モード、例えば従属量子化DQが使用される場合に使用可能なパラメーター値の第2の集合から1つ以上の確率推定パラメーターを選択することができる。

【0077】

任意選択で、使用可能なパラメーターの集合により、例えば、符号化ビットストリームに含まれるインデックス値と、関連する確率推定パラメーター又は確率推定パラメーターの関連するタプルとの間の異なる関連付けを定義することができ、例えば、使用可能なパラメーターの第1の集合は、確率推定パラメーターの第1のタプルへの所与のインデックス値のマッピングを定義し、例えば、使用可能なパラメーターの第2の集合は、第1のタプルとは異なる確率推定パラメーターの第2のタプルへの所与のインデックス値のマッピングを定義する。

【0078】

異なる量子化モードは、ニューラルネットワークパラメーターの異なる確率的特性（例えば、相関）を引き起こす可能性があり、したがって、量子化方法に従って確率推定パラメーターの選択を適応させることによって、改善された確率推定（及びその結果として改善された符号化効率）が達成され得ることが認識されている。

【0079】

本発明の実施の形態によれば、デコーダーは、1つ以上の確率推定パラメーターの一様量子化、例えば一様再構成量子化（URQ）又は時不変量子化が使用される場合に、使用可能なパラメーター値の第1の集合から、又はパラメーター値の使用可能なタプルの第1の集合から1つ以上の確率推定パラメーターを抜粋して選択するように構成され、デコーダーは、1つ以上の確率推定パラメーターの可変量子化、例えばトレリス符号化量子化（TCQ）又は依存量子化（DQ）が使用される場合に、使用可能なパラメーター値の第2の集合から、又はパラメーター値の使用可能なタプルの第2の集合から1つ以上の確率推定パラメーターを抜粋して選択するように構成され、ここで、例えば、DQのような可変量子化は、可変量子化の場合にコンテキストモデルごとに復号化されるビンが少なくなり得るように、一様量子化よりも多くのコンテキストモデルを使用することができる。

【0080】

代替的に、デコーダーは、1つ以上の確率推定パラメーターの一様量子化、例えば一様再構成量子化（URQ）又は時不変量子化が使用される場合に、1つ以上の確率推定パラメーターを表す符号化値、例えば符号化インデックス値 q を1つ以上の確率推定パラメーター、例えば $s h_i^k$ にマッピングする第1のマッピング規則、例えば第1のマッピングテーブルを使用するように構成される。

【0081】

加えて、デコーダーは、1つ以上の確率推定パラメーターの可変量子化、例えばトレリス符号化量子化（TCQ）又は依存量子化（DQ）が使用される場合に、1つ以上の確率推定パラメーターを表す符号化値、例えば符号化インデックス値 q を1つ以上の確率推定パラメーターにマッピングする第2のマッピング規則、例えば第2のマッピングテーブルを使用するように構成することができる。任意選択で、DQのような可変量子化は、可変

10

20

30

40

50

量子化の場合にコンテキストモデルごとに復号化されるピンがより少なくなり得るように、一様量子化よりも多くのコンテキストモデルを使用することができる。

【0082】

さらに、使用可能なパラメーター値の第1の集合は、使用可能なパラメーター値の第2の集合とは異なり、パラメーター値の使用可能なタブルの第1の集合は、パラメーター値の使用可能なタブルの第2の集合とは異なり、及び／又は第2のマッピング規則は、第1のマッピング規則とは異なる。

【0083】

したがって、本発明による概念は、任意の形態の量子化に適用することができ、或る特定の種類の量子化に限定されない。さらに、ニューラルネットワークパラメーターの特性に従って、量子化の種類並びに確率推定パラメーターの使用される集合及び／又はマッピングを適応させることができる。したがって、自由度が高いだけでなく、符号化効率も向上させることができる。

【0084】

本発明の更なる実施の形態によれば、平均して、使用可能なパラメーター値の第2の集合のうちの使用可能なパラメーター値、又は例えば第2のマッピング規則の使用可能なパラメーター値は、使用可能なパラメーター値の第1の集合のうちの使用可能なパラメーター値、又は例えば第1のマッピング規則の使用可能なパラメーター値よりも、例えばピン値の頻度の変化に対して、確率推定値のより高速な適応を可能にする。代替的に、平均して、パラメーター値の使用可能なタブルの第2の集合のパラメーター値の使用可能なタブル、又は例えば第2のマッピング規則の使用可能なタブルは、パラメーター値の使用可能なタブルの第1の集合のパラメーター値の使用可能なタブル、又は例えば第1のマッピング規則の使用可能なタブルよりも、例えばピン値の頻度の変化に対して、確率推定値のより高速な適応を可能にする。

【0085】

本実施の形態は、かかる概念により符号化効率が向上するという知見に基づくものである。

【0086】

本発明の更なる実施の形態によれば、使用可能なパラメーター値の第2の集合、又は例えば第2のマッピング規則は、使用可能なパラメーター値の第1の集合、又は例えば第1のマッピング規則の使用可能なパラメーター値、又は例えば全ての使用可能なパラメーター値よりも、例えばピン値の頻度の変化に対して、確率推定値のより高速な適応を可能にする使用可能なパラメーター値を含む。代替的に、パラメーター値の使用可能なタブルの第2の集合は、パラメーター値の使用可能なタブルの第1の集合のパラメーター値の使用可能なタブルよりも、又は例えば全ての使用可能なタブルよりも、例えばピン値の頻度の変化に対して、確率推定値のより高速な適応を可能にするパラメーター値の使用可能なタブルを含む。

【0087】

本実施の形態は、かかる概念により符号化効率が向上するという知見に基づくものである。

【0088】

本発明の更なる実施の形態によれば、デコーダーは、ニューラルネットワークの層のパラメーターの数に応じて、又は選択された1つ以上の確率推定パラメーターを使用して復号化されるニューラルネットワークパラメーターの数に応じて、又は層パラメーターの要素の数、例えば行列Wの成分の数、若しくは（転置）ベクトルbの要素の数等に応じて、使用可能な、例えば許容可能なパラメーター値の異なる集合、又は使用可能な、例えば許容可能なパラメーター値のタブルの異なる集合から1つ以上の確率推定パラメーターを選択するように構成される。

【0089】

代替的に、デコーダーは、ニューラルネットワークの層のパラメーターの数に依存して

10

20

30

40

50

、又は選択された1つ以上の確率推定パラメーターを使用して復号化されるニューラルネットワークパラメーターの数に依存して、又は層パラメーターの要素の数、例えば、行列Wの成分の数、若しくは（転置）ベクトルbの要素の数等に依存して、1つ以上の確率推定パラメーターを表す符号化値、例えば符号化インデックス値qを1つ以上の確率推定パラメーター、例えば $s h_i^k$ にマッピングする、異なるマッピング規則、例えば異なるマッピングテーブルを使用するように構成される。

【0090】

或る特定の値を有するビットの確率値を評価するために、確率推定器の適応パラメーターは、コンテキストモデルを使用して、復号化されるニューラルネットワークパラメーターの数に依存して選択されることが分かっている。例えば、復号化済みのビンの数が多いほど、そのビンの確率を確率的に推定することができる。これにより、符号化効率を向上させることができる。

10

【0091】

本発明の更なる実施の形態によれば、デコーダーは、ニューラルネットワークの層のパラメーターの数が閾値、例えば $X = 1000$ を下回る場合、又は選択された1つ以上の確率推定パラメーターを使用して復号化されるニューラルネットワークパラメーターの数が閾値を下回る場合、又は層パラメーターの要素の数、例えば行列Wの成分の数、若しくは（転置）ベクトルbの要素の数等が閾値を下回る場合、使用可能なパラメーター値の第1の集合から、又はパラメーター値の使用可能なタプルの第1の集合から1つ以上の確率推定パラメーターを抜粋して選択するように構成される。

20

【0092】

加えて、デコーダーは、ニューラルネットワークの層のパラメーターの数が閾値、例えば $X = 1000$ を超える場合、又は選択された1つ以上の確率推定パラメーターを使用して復号化されるニューラルネットワークパラメーターの数が閾値を超える場合、又は層パラメーターの要素の数、例えば行列Wの成分の数、若しくは（転置）ベクトルbの要素の数等が閾値を超える場合、使用可能なパラメーター値の第2の集合から、又はパラメーター値の使用可能なタプルの第2の集合から1つ以上の確率推定パラメーターを抜粋して選択するように構成される。

【0093】

代替的に、デコーダーは、ニューラルネットワークの層のパラメーターの数が閾値、例えば $X = 1000$ を下回る場合、又は選択された1つ以上の確率推定パラメーターを使用して復号化されるニューラルネットワークパラメーターの数が閾値を下回る場合、又は層パラメーターの要素の数、例えば行列Wの成分の数若しくは（転置）ベクトルbの要素の数等が閾値を下回る場合、1つ以上の確率推定パラメーター、例えば $s h_i^k$ 上に1つ以上の確率推定パラメーターを表す符号化値、例えば符号化インデックス値qをマッピングする第1のマッピング規則、例えば第1のマッピングテーブルを抜粋して使用するように構成される。

30

【0094】

さらに、デコーダーは、ニューラルネットワークの層のパラメーターの数が閾値、例えば $X = 1000$ を超える場合、又は選択された1つ以上の確率推定パラメーターを使用して復号化されるニューラルネットワークパラメーターの数が閾値を超える場合、又は層パラメーターの要素の数、例えば行列Wの成分の数若しくは（転置）ベクトルbの要素の数等が閾値を超える場合、1つ以上の確率推定パラメーター、例えば $s h_i^k$ に1つ以上の確率推定パラメーターを表す符号化値、例えば符号化インデックス値qをマッピングする第2のマッピング規則、例えば第2のマッピングテーブルを抜粋して使用するように構成される。

40

【0095】

さらに、使用可能なパラメーター値の第2の集合は、使用可能なパラメーター値の第1の集合よりも多くの使用可能なパラメーター値を含み、パラメーター値の使用可能なタプルの第2の集合は、パラメーター値の使用可能なタプルの第1の集合よりも多くの使用可

50

能なタプルを含み、及び／又は第2のマッピング規則は、第1のマッピング規則とは異なる。

【0096】

或る特定の閾値に従って確率推定パラメータを選択することで、計算量的に安価に確率推定パラメータを適応できる可能性があることが分かっている。前述したように、例えばコンテキストモデルの統計的適応は、多数のパラメータが復号化される場合に、例えば、より良好に実行することができるため、確率推定パラメータは、かかる数に従って調整することができる。また、シグナリングは、現在考慮されているニューラルネットワークによく適した確率推定パラメータ（又は確率推定パラメータの集合）に限定されるので、シグナリングオーバーヘッド（例えば、確率推定パラメータ又は確率推定パラメータの集合のインデックスの符号化に必要なビット数）を低減することができる。

10

【0097】

本発明の更なる実施の形態によれば、デコーダーは、選択された1つ以上の確率推定パラメータを使用して復号化されるニューラルネットワークパラメータの数が閾値、例えば $X = 1000$ 以上である場合、増加した選択肢から1つ以上の確率推定パラメータを抜粋して選択するように構成される。

【0098】

復号化されるニューラルネットワークパラメータの数が増加すると、ニューラルネットワークパラメータの確率的特性、例えばそれらの相関に関する自由度が増え得ることが分かっている。したがって、ニューラルネットワークパラメータの数が増加するにつれて、符号化効率を向上させるために、確率推定パラメータ又は確率推定パラメータ値の増加を考慮することができる。その結果、小規模ニューラルネットワークでは、シグナリングオーバーヘッドが低減される。

20

【0099】

本発明の更なる実施の形態によれば、デコーダーは、例えば専用フラグの形で、例えば符号化ビットストリームに含まれ得るシグナリングを評価するように構成され、そのシグナリングから、使用可能なパラメータ値の集合、例えば小さい集合又は大きい集合、例えば使用可能なパラメータ値の複数の集合のうちの集合であって、基集合の異なる部分集合、場合によっては重複する部分集合であり得る集合から、又はパラメータ値の使用可能なタプルの集合、例えば小さい集合又は大きい集合、例えばパラメータ値の使用可能なタプルの複数の集合のうちの集合であって、基集合の異なる部分集合、場合によっては重複する部分集合から、1つ以上の確率推定パラメータが、例えば復号化される1つ以上のフラグを使用して、例えばフラグ「useSecondSubset」を使用して選択される。

30

【0100】

代替的に、デコーダーは、1つ以上の確率推定パラメータを表す符号化値、例えば符号化インデックス値 q を1つ以上の確率推定パラメータ、例えば $s h_i^k$ にマッピングするために、複数のマッピング規則、例えばマッピングテーブルのうちのどのマッピング規則を使用すべきかの指示であるシグナリングを評価するように構成され、ここで、シグナリングは、例えば符号化ビットストリームに、例えば専用フラグの形態で含まれてもよい。

40

【0101】

確率推定パラメータの対応する集合又はタプルの集合に関する情報を用いて、確率推定パラメータの選択を低い計算コストで実行することができる。加えて、かかるシグナリングを伝送することによって、デコーダーにおいてエンコーダーとは異なる確率推定パラメータを使用するエラー確率を低減することができる。同様に、有益なマッピング規則、例えば、符号化ビットストリームを符号化するために使用されるエンコーダーにおいて使用されるマッピング規則は、かかるシグナリングを介して通信され得る。

【0102】

例えば、かかるシグナリングを提供することによって、符号化は、ニューラルネットワ

50

ークの特性に適応してもよい。例えば、パラメーター値の使用可能なタプルの第1の集合を、第1のタイプのニューラルネットワークのために選択してもよく（例えば、使用可能なパラメーター値の第1の集合が、第1のタイプのニューラルネットワークの統計により良好に適応するため）、パラメーター値の使用可能なタプルの第2の集合を、第2のタイプのニューラルネットワークのために選択してもよい。例えば、ニューラルネットワークに対して使用可能なパラメーター値の集合の選択を1回だけ（又は少なくとも個々のパラメーター値の集合の実際の選択よりもまれに）シグナリングすることで十分であり得る。これにより、符号化効率を向上させることができる。

【0103】

本発明の更なる実施の形態によれば、デコーダーは、確率推定パラメーター値を記述する、又は複数の確率推定パラメーター値を記述する、又は確率推定パラメーター値のタプル、例えばインデックス q を記述する、1つ以上のインデックス値、例えば概して整数値、例えばインデックス q 又は複数のインデックスを復号化するように構成される。

10

【0104】

インデックス値を使用することによって、確率推定パラメーターを容易に圧縮可能な方法で表すことができる。確率推定パラメーターは、ニューラルネットワークパラメーターと同様に符号化され、したがって復号化されるため、符号化／復号化の効率を向上させることができる。

【0105】

本発明の更なる実施の形態によれば、デコーダーは、1つ以上のコンテキストモデルを使用して1つ以上のインデックス値を復号化するように構成され、コンテキストモデルは、例えば、インデックス値を復号化するために使用される1つ以上のビンのビン値の確率を決定することができる。

20

【0106】

本発明の1つの主要な概念を追求すると、ニューラルネットワークパラメーターの符号化／復号化のシーケンスは、確率推定パラメーターに同様に適用することができる。したがって、確率推定パラメーター又は確率推定パラメーターを表すビン、又は確率推定パラメーターを表すインデックス値は、コンテキストモデルに関連付けることができる。コンテキスト依存符号化の利点は、確率推定パラメーターの符号化／復号化により、ニューラルネットワークパラメーターをコンテキスト依存符号化／復号化を行うことで2度利用することができる。

30

【0107】

本発明の更なる実施の形態によれば、デコーダーは、現在考慮されているインデックス値がデフォルト値を取るかどうかを記述する第1のビン、例えば `useNotDefault` ビンを復号化するように構成され、デコーダーは、現在考慮されているインデックス値がデフォルト値を取らない場合、現在考慮されているインデックス値又は現在考慮されているインデックス値から導出された値、例えば $q-1$ を二値表現で表す1つ以上の追加のビンを抜粋して復号化するように構成される。

【0108】

任意選択で、現在考慮されているインデックス値がデフォルト値を取るかどうかを示す第1のビンは、コンテキストを使用して、例えば確率推定値を考慮して復号化され、1つ以上の追加のビンは、ビンごとに1ビットの固定長で復号化される。

40

【0109】

ニューラルネットワークパラメーターが、例えば、第1のビン復号化ステップにおいてデフォルト値として識別される場合、更なる復号化が必要ない場合があるため、かかる第1のビンを復号化することで符号化効率を向上させることができる。第1のビンは常に存在し得るため、例えば、より複雑なコンテキスト依存符号化により、符号化効率を向上させることができる。常に存在するとは限らない追加のビンに対しては、例えば固定された確率を有するより複雑でない符号化を実施することができる。

【0110】

50

本発明の更なる実施の形態によれば、デコーダーは、単項コード復号化 (unary code decoding) を使用して、又は短縮単項コード復号化を使用して、又は可変長コード復号化を使用して、1つ以上のインデックス値を復号化するように構成され、任意選択で、コード長は、異なるインデックス値の発生確率に従って選択される。単項コードを使用することで、プレフィックスフリーなしの自己同期コードが可能となり得る。また、可変長コードでは、出現確率の高いインデックス値の符号長を短くすることができ、出現頻度の高いインデックスの符号長を短くすることでシンボル及び時間を節約することができるため、効率を向上させることができる。

【0111】

本発明の更なる実施の形態によれば、デコーダーは、例えば単項コード又はハフマンコードの場合、1つ以上の確率推定パラメーター、例えば選択された確率推定パラメーター又は確率推定パラメーターの選択されたタプルを指定する整数インデックス q を復号化するために使用されるビンの数又はビンの最大数を、1つ以上の確率推定パラメーターを量子化するために使用される量子化モードに応じて、例えば使用可能なパラメーター値の選択された集合又はパラメーター値の使用可能なタプルの選択された集合に適應するように変更するように構成される。

10

【0112】

或る特定の量子化方法を用いると、確率推定パラメーターの或る特定の正確度又は精度が達成可能であり得る。したがって、量子化又は予想される量子化誤差に関してビンの数又はビンの最大数が選択され得る場合、符号化効率を向上させることができる。

20

【0113】

本発明の更なる実施の形態によれば、デコーダーは、例えば単項コード又はハフマンコードの場合、1つ以上の確率推定パラメーター、例えば選択された確率推定パラメーター又は確率推定パラメーターの選択されたタプルを指定する整数インデックス q を復号化するために使用されるビンの数又はビンの最大数を、ニューラルネットワークの層のパラメーターの数に応じて、又は1つ以上の確率推定パラメーターを使用して復号化されるニューラルネットワークパラメーターの数に応じて、又は層パラメーターの要素の数、例えば行列 W の成分の数、又は (転置) ベクトル b の要素の数等に応じて、例えば使用可能なパラメーター値の選択された集合又はパラメーター値の使用可能なタプルの選択された集合に適應するように変更するように構成される。

30

【0114】

符号化効率は、復号化されるニューラルネットワークパラメーターの数に関して確率推定パラメーターを復号化するために使用されるビンの数を変更することによって向上させることができる。精度と計算コスト及び時間努力との間の良好なトレードオフを実装することができる。加えて、ビンの確率的特性は、ニューラルネットワークパラメーターの数に応じて決定され、例えば、当該コンテキストモデルに依存するニューラルネットワークパラメーターの数が増加するにつれて、コンテキストモデルをより正確に適應させることができる。

【0115】

本発明の更なる実施の形態によれば、デコーダーは、1つ以上の確率推定パラメーターに関連付けられた使用可能なパラメーター値の異なる集合の間で、又は複数の確率推定パラメーターに関連付けられた使用可能なパラメーター値のタプルの異なる集合の間で、又は1つ以上の確率推定パラメーターを表す符号化値、例えば符号化インデックス値 q を1つ以上の確率推定パラメーター、例えば $s h_i^k$ にマッピングする異なるマッピング規則間で切り替えるように構成される。かかる切り替えは、例えば、復号化される別の層のニューラルネットワークパラメーターに切り替えた後に実行されてもよい。この柔軟性により、増加した符号化効率のための復号化のより高度な適應を可能にすることができる。

40

【0116】

本発明の更なる実施の形態によれば、デコーダーは、例えば単項コード又はハフマンコードの場合、1つ以上の確率推定パラメーターに関連付けられた使用可能なパラメーター

50

値の異なる集合の間、又は複数の確率推定パラメーターに関連付けられた使用可能なパラメーター値のタプルの異なる集合の間、又は異なるマッピング規則間の切り替えに従って、選択された確率推定パラメーター又は確率推定パラメーターの選択されたタプルを指定する1つ以上の確率推定パラメーター、例えば整数インデックス q を復号化するために使用されるビンの数又はビンの最大数を変更するように構成される。

【0117】

より良い理解のために、確率推定パラメーターのビン表現に関する前述の実施の形態の態様を、符号化手順を含む実施の形態のコンテキストで以下に説明する。したがって、任意選択の詳細を有する符号化方式が以下で開示される。

【0118】

符号化されるインデックス $q \in [0, q_{MAX}]$ は、例えば、ビンのシーケンスに分解され、次いで、ビンのシーケンスが符号化される。各ピンは、例えば、コンテキストモデルを使用して、又は固定確率を使用して符号化されてもよい。

【0119】

符号化手順は、例えば、以下の方式のうちの1つに従うことができる。

【0120】

1. 第1のピン、例えば `useNotDefault` は、選択される推定器パラメーターがデフォルトパラメーターと異なる（例えば `useNotDefault=1`）か否か（例えば `useNotDefault=0`）を示す。例えば、`useNotDefault=0` である場合、デフォルトパラメーターが選択され、更なるピンは符号化されない。例えば `useNotDefault=1` である限り、ビンのシーケンス、例えば追加のピンが符号化され、これは例えば選択されたパラメーターのインデックスから1を引いたもの（ $q-1$ ）、`indexMinusOne`を示す。インデックスのために符号化されるビンの数は、例えば、

【数10】

$$\lceil \log_2(\text{setLength} - 1) \rceil$$

に等しく、ここで、`setLength`は、集合の要素の数を示す。

【0121】

2. 第2の手順では、単項コードが使用される。第1のピン、例えば `greaterThan_0` は、確率パラメーターに関連するインデックス q が0より大きい（例えば `greaterThan_0=1`）か、又はそうでない（例えば `greaterThan_0=0`）かを示す。例えば `greaterThan_0=0` である場合、更なるピンは符号化されない。例えば `greaterThan_0=1` である場合、別の、例えば追加のピンが符号化され（例えば `greaterThan_1`）、これは、インデックス q が1よりも大きい（例えば `greaterThan_1=1`）か否か（例えば `greaterThan_1=0`）を示す。例えば `greaterThan_1=0` である場合、更なるピンは符号化されない。例えば、`greaterThan_1=0` である場合、フラグ `greaterThan_q` が0に等しくなるまで、更なるピン（`greaterThan_x`）が同様に符号化される。

【0122】

3. この手順は、例えば、 q を符号化するためのインデックスが q_{MAX} に等しい場合を除いて、符号化方法2.において使用される単項コードと同一である、短縮単項コードを適用する。この場合、例えば、ピン `greaterThan_1`（ $q_{MAX}-1$ ）を符号化した後、更なるピンは符号化されない。例えば、デコーダー側では、`greaterThan_1`（ $q_{MAX}-1$ ）が1に等しい場合、 q の値は q_{MAX} であると推論される。

【0123】

4. この手順は、符号長がシンボルの発生確率に従って選択される可変長コード、例えばハフマンコードを使用する。

【0124】

これらの方式のいずれも、本発明の任意の実施の形態において使用され得ることに留意されたい。符号化のための方式が復号化のために適用されてもよく、逆もまた同様であることは、当業者にとって明らかであろう。

【0125】

本発明の更なる実施の形態によれば、デコーダーは、1つ以上の状態変数、例えば s_i^k 又は s_k を決定し、1つ以上の状態変数を使用して、例えば式(1)及び(2)を使用して、又は式(3)及び(4)を使用して、又は式(3)及び(5)を使用して、又は1つ以上の状態変数と確率推定値、例えば $P(x, i, k)$ との間の線形関係を使用して、確率推定値、例えば p_k を導出するように構成される。

10

【0126】

状態変数は、ニューラルネットワークパラメーター又はそのビンの確率を記述する確率モデル、例えばコンテキストモデルの評価のための効率的な手段を提供する。前述したように、状態変数は、確率推定パラメーターの符号化にも使用することができる。確率推定パラメーターは、状態変数を使用して評価されるコンテキストモデルを使用して復号化されてもよく、状態変数は、例えば、既に復号化された、又は例えば最近復号化された確率推定パラメーターに基づいて更新されてもよい。例えば、式(1)に示したようなルックアップテーブルを使用することで、計算コストを更に低減することができる。

20

【0127】

本発明の更なる実施の形態によれば、デコーダーは、

【数11】

$$s_k = \sum_{i=1}^N [s_i^k \cdot d_i^k]$$

及び

【数12】

30

$$p_k = \begin{cases} LUT2[|s_k \cdot a_k|], & |s_k \cdot a_k| \geq \text{である場合} \\ 1 - LUT2[-|s_k \cdot a_k|], & \text{それ以外の場合} \end{cases}$$

に従って2つの状態変数 s_1^k 、 s_2^k から確率推定値 p_k を導出するように構成される。任意選択で、LUT2は、本明細書で説明したようなものであってよく、例えば、全ての k に対して、例えば、 $N=2$ 、 $d_1^k=16$ 、 $d_2^k=1$ 、及び $a_k=2^{-7}$ であり、 k はコンテキストモデルインデックスである。

40

【0128】

本発明の実施の形態によれば、デコーダーは、

【数13】

50

$$s_i^k = \begin{cases} s_i^k + [A[z + [s_i^k \cdot m_i^k]] \cdot n_i^k], & \text{復号化されるシンボルが1である場合} \\ s_i^k - [A[z + [-s_i^k \cdot m_i^k]] \cdot n_i^k], & \text{復号化されるシンボルが0である場合} \end{cases}$$

に従って状態変数 s_1^k 、 s_2^k を更新するように構成される。ここで、 m_i^k 及び n_i^k は重み係数であり、任意選択で確率推定パラメーターを構成し、例えば、 m_1^k は 2^{-3} に等しくてもよく、 m_2^k は 2^{-7} に等しくてもよい。加えて、 A は、例えば本明細書で説明されるように定義され得る、例えば整数値を記憶するルックアップテーブルである。さらに、 z はオフセット値、例えば所定の値であり、例えば 16 に等しくてもよい。任意選択で、 s_i^k は、例えば、本明細書で説明されるように初期化されてもよい。

【0129】

本発明の更なる実施の形態によれば、デコーダーは、異なるニューラルネットワークパラメーターの復号化のために異なる確率推定パラメーター値、例えば n_i^k の異なる値を使用するように、及び／又は異なるコンテキストモデルに関連付けられたビンの復号化のために異なる確率推定パラメーター値を使用するように、及び／又はニューラルネットワークの異なる層に関連付けられたニューラルネットワークパラメーターの復号化のために異なる確率推定パラメーター値を使用するように、重み係数 n_i^k を変更するように構成される。

【0130】

本発明の更なる実施の形態によれば、重み係数 n_i^k と適応パラメーター $s_{h_i}^k$ との間

の関係は、

【数14】

$$n_i^k = 2^{-s_{h_i}^k + 4}$$

に従って定義される。

【0131】

本発明の更なる実施の形態によれば、デコーダーは、適応パラメーターを記述する情報、例えば適応パラメーターのタプルを記述するインデックス q を復号化するように構成され、任意選択で、復号化されたインデックス値 q の意味は、例えば、以下の表2又は表3、又は表4又は表5又は表6又は表7に提供されるように定義されてもよい。これらのテーブルは、マッピング規則又はマッピングテーブルの例である。

【0132】

好ましい実施の形態、例えば上記で説明した実施の形態では、推定器は、例えば、 N 組の適応パラメーター $s_{h_i}^k$ である適応パラメーターの基集合を適用する。次に、基集合の部分集合が選択される。部分集合のうちの1つのパラメーターがシグナリングされ得る。

【0133】

特に好ましい実施の形態では、構成は、例えば、前の好ましい実施の形態と等しいが、マルチメディアコンテンツの記述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規格の現在のドラフトで使用されている推定器と同一であるように構成された推定器が使用され、基集合は、例えば、 $(s_{h_1}^k, s_{h_2}^k)$ について以下の28個の対を含む。

【0134】

10

20

30

40

50

【表 1】

表 1：適応パラメーターの好ましい基集合

0	(0, 0)	1 4	(2, 3)
1	(0, 1)	1 5	(2, 4)
2	(0, 2)	1 6	(2, 5)
3	(0, 3)	1 7	(2, 6)
4	(0, 4)	1 8	(3, 3)
5	(0, 5)	1 9	(3, 4)
6	(0, 6)	2 0	(3, 5)
7	(1, 1)	2 1	(3, 6)
8	(1, 2)	2 2	(4, 4)
9	(1, 3)	2 3	(4, 5)
1 0	(1, 4)	2 4	(4, 6)
1 1	(1, 5)	2 5	(5, 5)
1 2	(1, 6)	2 6	(5, 6)
1 3	(2, 2)	2 7	(6, 6)

10

【0 1 3 5】

サイズ 3 の部分集合は、例えば、層の全てのパラメーターが D Q で量子化される場合に、例えば、表 2 によるインデックス q が割り当てられるように定義され、順序付けられる。インデックス q = 0 を有するパラメーターは、例えば、デフォルトパラメーターとして示される。

20

【0 1 3 6】

【表 2】

表 2：集合サイズ 3 に対する適応パラメーターの好ましい部分集合

q	適応パラメーターの対
0	(1, 4)
1	(0, 1)
2	(2, 6)

30

【0 1 3 7】

例えば、部分集合のうちの 1 つのパラメーターは、例えば、符号化方式 1 に従って q を符号化することによってシグナリングされ、ここで、例えば、ビン `useNotDefault` はコンテキストモデルを使用して符号化され、全ての他のビンはビンごとに 1 ビットの固定長で符号化される。概して、本発明の実施の形態によれば、コンテキスト依存符号化及び／又は他の符号化、例えば可変長又は固定長符号化の任意の混合が、確率推定パラメーター及び／又はニューラルネットワークパラメーターの任意のビンに適用されてもよい。

40

【0 1 3 8】

別の好ましい実施の形態（例）では、構成は、割り当てられた適応パラメーターの対と、選択された部分集合（表 3）のサイズが 5 に等しい以外は、前の好ましい実施の形態と同一である。

【0 1 3 9】

50

【表 3】

表 3：集合サイズ 5 に対する適応パラメーターの好ましい部分集合

q	適応パラメーターの対
0	(1, 4)
1	(0, 0)
2	(0, 6)
3	(1, 1)
4	(2, 6)

10

【0140】

別の好ましい実施の形態（例）では、構成は、割り当てられた適応パラメーターの対（表 4）以外は、前の好ましい実施の形態と同一である。

【0141】

【表 4】

表 4：集合サイズ 5 に対する適応パラメーターの第 2 の好ましい部分集合

q	適応パラメーターの対
0	(1, 2)
1	(0, 0)
2	(0, 5)
3	(2, 5)
4	(3, 4)

20

【0142】

別の好ましい実施の形態（例）では、構成は、割り当てられた適応パラメーターの対と、選択された部分集合（表 5）のサイズが 9 に等しい以外は、前の好ましい実施の形態と同一である。

【0143】

【表 5】

表 5：集合サイズ 9 に対する適応パラメーターの好ましい部分集合

q	適応パラメーターの対
0	(1, 4)
1	(0, 0)
2	(0, 5)
3	(1, 1)
4	(1, 2)
5	(2, 4)
6	(2, 6)
7	(3, 4)
8	(3, 5)

40

【0144】

別の好ましい実施の形態（例）では、構成は、割り当てられた適応パラメーターの対（表 6）以外は、前の好ましい実施の形態と同一である。

【0145】

50

【表 6】

表 6：集合サイズ 9 に対する適応パラメーターの第 2 の好ましい部分集合

d	適応パラメーターの対
0	(1, 3)
1	(0, 0)
2	(0, 5)
3	(1, 1)
4	(1, 6)
5	(2, 4)
6	(2, 6)
7	(3, 5)
8	(4, 4)

10

【0146】

別の好ましい実施の形態（例）では、構成は、割り当てられた適応パラメーターの対（表 7）、選択された部分集合のサイズ（5）、及び U R Q を使用する使用された量子化方法以外は、前の好ましい実施の形態と同一である。

【0147】

【表 7】

表 7：集合サイズ 5 及び U R Q に対する適応パラメーターの好ましい部分集合

q	適応パラメーターの対
0	(1, 4)
1	(0, 6)
2	(1, 1)
3	(2, 6)
4	(3, 4)

20

【0148】

別の好ましい実施の形態（例）では、マルチメディアコンテンツの記述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 M P E G - 7 パート 1 7 規格の現在のドラフトで使用される推定器と同一であるように構成された推定器が使用され、表 1 の基集合が使用される。これを基本構成と称する。

30

【0149】

層パラメーターが D Q を用いて量子化される場合は常に、表 5 中のパラメーターの対の（サイズ 9 の）部分集合を適用することができる。層パラメーターが U R Q を用いて量子化される場合、表 8 の部分集合を使用することができる。

【0150】

40

50

【表 8】

表 8：集合サイズ 9 に対する適応パラメーターの好ましい部分集合

q	適応パラメーターの対
0	(1, 4)
1	(0, 1)
2	(0, 6)
3	(1, 2)
4	(1, 6)
5	(2, 5)
6	(2, 6)
7	(3, 4)
8	(3, 5)

10

【0151】

別の好ましい実施の形態（例）では、前の好ましい実施の形態の基本構成を適用することができる。

【0152】

層パラメーターの要素の数が、例えば $X = 1000$ に設定され得る閾値 X を下回る場合は常に、例えば表 2 において第 1 の部分集合として示されるパラメーターの対のサイズ 3 を有する部分集合を使用することができる。そうではなく、層パラメーターの要素の数が閾値 X 以上である場合、第 2 の部分集合として示される、例えば、表 5 中のサイズ 9 を有する部分集合を使用することができる。

20

【0153】

別の好ましい実施の形態（例）では、構成は、前の好ましい実施の形態と同一であるが、閾値を使用する代わりに、フラグ（例えば、`useSecondSubset`）が符号化され、これは、例えば、使用される部分集合を決定する。例えば、フラグが 0 に等しい場合、第 1 の部分集合が使用され得る。フラグが 1 に等しい場合、第 2 の部分集合が使用され得る。

30

【0154】

本発明の更なる実施の形態によれば、エンコーダーは、基集合から、又は基集合の真の部分集合から、1 つ以上の確率推定パラメーターを選択するように構成される。

【0155】

任意選択で、基集合は、1 つ以上の確率推定パラメーターに関連付けられた複数の使用可能なパラメーター値を含むことができ、又は基集合は、複数の確率推定パラメーターに関連付けられた使用可能なパラメーター値の複数のタプル、例えば、使用可能な対のリスト ($s h_1^k$, $s h_2^k$) を含むことができる。

【0156】

本発明の更なる実施の形態によれば、エンコーダーは、量子化モードに応じて、使用可能なパラメーター値の異なる集合又はパラメーター値の使用可能なタプルの異なる集合から 1 つ以上の確率推定パラメーターを選択するように構成され、例えば、エンコーダーは、第 1 の量子化モード、例えば一様量子化 URQ が使用される場合に使用可能なパラメーター値の第 1 の集合又はパラメーター値の使用可能なタプルの第 1 の集合から 1 つ以上の確率推定パラメーターを選択し、例えば、エンコーダーは、第 2 の量子化モード、例えば従属量子化 DQ が使用される場合に使用可能なパラメーター値の第 2 の集合から 1 つ以上の確率推定パラメーターを選択する。

40

【0157】

本発明の更なる実施の形態によれば、エンコーダーは、1 つ以上の確率推定パラメーターの一様量子化、例えば一様再構成量子化 (URQ) 又は時不変量子化が使用される場合

50

に、使用可能なパラメーター値の第1の集合から、又はパラメーター値の使用可能なタブルの第1の集合から1つ以上の確率推定パラメーターを抜粋して選択するように構成され、エンコーダーは、1つ以上の確率推定パラメーターの可変量子化、例えばトレリス符号化量子化(TCQ)又は依存量子化(DQ)が使用される場合に、使用可能なパラメーター値の第2の集合から、又はパラメーター値の使用可能なタブルの第2の集合から1つ以上の確率推定パラメーターを抜粋して選択するように構成される。任意選択で、DQのような可変量子化は、可変量子化の場合にコンテキストモデルごとに符号化されるビンがより少なくなり得るように、一様量子化よりも多くのコンテキストモデルを使用することができる。

【0158】

加えて、使用可能なパラメーター値の第1の集合は、使用可能なパラメーター値の第2の集合とは異なり、パラメーター値の使用可能なタブルの第1の集合は、パラメーター値の使用可能なタブルの第2の集合とは異なる。

【0159】

本発明の更なる実施の形態によれば、平均して、使用可能なパラメーター値の第2の集合のうちの使用可能なパラメーター値は、使用可能なパラメーター値の第1の集合のうちの使用可能なパラメーター値よりも、例えばビン値の頻度の変化に対して、確率推定値のより高速な適応を可能にする。代替的に、平均して、パラメーター値の使用可能なタブルの第2の集合のパラメーター値の使用可能なタブルは、パラメーター値の使用可能なタブルの第1の集合のパラメーター値の使用可能なタブルよりも、例えばビン値の頻度の変化に対して、確率推定値のより高速な適応を可能にする。

【0160】

本発明の更なる実施の形態によれば、使用可能なパラメーター値の第2の集合は、使用可能なパラメーター値の第1の集合のうちの使用可能なパラメーター値よりも、又は更には例えば全ての使用可能なパラメーター値よりも、例えばビン値の頻度の変化に対して、確率推定値のより高速な適応を可能にする使用可能なパラメーター値を含む。代替的に、パラメーター値の使用可能なタブルの第2の集合は、パラメーター値の使用可能なタブルの第1の集合のパラメーター値の使用可能なタブルよりも、又は更には例えば全ての使用可能なタブルよりも、例えばビン値の頻度の変化に対して、確率推定値のより高速な適応を可能にするパラメーター値の使用可能なタブルを含む。

【0161】

本発明の更なる実施の形態によれば、エンコーダーは、ニューラルネットワークの層のパラメーターの数に応じて、又は選択された1つ以上の確率推定パラメーターを使用して符号化されるニューラルネットワークパラメーターの数に応じて、又は層パラメーターの要素の数、例えば行列Wの成分の数、若しくは(転置)ベクトルbの要素の数等に応じて、使用可能な、例えば許容可能なパラメーター値の異なる集合、又は使用可能な、例えば許容可能なパラメーター値のタブルの異なる集合から1つ以上の確率推定パラメーターを選択するように構成される。

【0162】

本発明の更なる例によれば、エンコーダーは、ニューラルネットワークの層のパラメーターの数が閾値、例えば $X=1000$ を下回る場合、又は選択された1つ以上の確率推定パラメーターを使用して符号化されるニューラルネットワークパラメーターの数が閾値を下回る場合、又は層パラメーターの要素の数、例えば行列Wの成分の数、若しくは(転置)ベクトルbの要素の数等が閾値を下回る場合、使用可能なパラメーター値の第1の集合から、又はパラメーター値の使用可能なタブルの第1の集合から1つ以上の確率推定パラメーターを抜粋して選択するように構成される。

【0163】

加えて、エンコーダーは、ニューラルネットワークの層のパラメーターの数が閾値、例えば $X=1000$ を超える場合、又は選択された1つ以上の確率推定パラメーターを使用して符号化されるニューラルネットワークパラメーターの数が閾値を超える場合、又は層

10

20

30

40

50

パラメーターの要素の数、例えば行列Wの成分の数、若しくは（転置）ベクトルbの要素の数等が閾値を超える場合、使用可能なパラメーター値の第2の集合から、又はパラメーター値の使用可能なタプルの第2の集合から1つ以上の確率推定パラメーターを抜粋して選択するように構成される。

【0164】

さらに、使用可能なパラメーター値の第2の集合は、使用可能なパラメーター値の第1の集合よりも多くの使用可能なパラメーター値を含み、パラメーター値の使用可能なタプルの第2の集合は、パラメーター値の使用可能なタプルの第1の集合よりも多くの使用可能なタプルを含む。

【0165】

本発明の更なる実施の形態によれば、エンコーダーは、選択された1つ以上の確率推定パラメーターを使用して符号化されるニューラルネットワークパラメーターの数が閾値、例えば $X = 1000$ 以上である場合、増加した選択肢から1つ以上の確率推定パラメーターを抜粋して選択するように構成される。

【0166】

本発明の更なる実施の形態によれば、エンコーダーは、使用可能なパラメーター値のどの集合から、例えば、基集合の異なる部分集合、場合によっては重複する部分集合であり得る使用可能なパラメーター値の複数の集合のうちの小さい集合又は大きい集合から、又はパラメーター値の使用可能なタプルのどの集合から、例えば、基集合の異なる部分集合、場合によっては重複する部分集合であり得るパラメーター値の使用可能なタプルの複数の集合のうちの小さい集合又は大きい集合から、1つ以上の確率推定パラメーターが、例えば、符号化された1つ以上のフラグを使用して、例えば、フラグ「useSecondSubset」を使用して選択されるかをシグナリングするように構成される。

【0167】

本発明の更なる実施の形態によれば、エンコーダーは、確率推定パラメーター値を記述する、又は複数の確率推定パラメーター値を記述する、又は確率推定パラメーター値のタプル、例えばインデックスqを記述する、1つ以上のインデックス値、例えば概して整数値、例えばインデックスq又は複数のインデックスを符号化するように構成される。

【0168】

本発明の更なる実施の形態によれば、エンコーダーは、1つ以上のコンテキストモデルを使用して1つ以上のインデックス値を符号化するように構成され、コンテキストモデルは、例えば、インデックス値を符号化するために使用される1つ以上のビンのビン値の確率を決定することができる。

【0169】

本発明の更なる実施の形態によれば、エンコーダーは、現在考慮されているインデックス値がデフォルト値を取ることを記述する第1のビンを使用して、又は、例えば、現在考慮されているインデックス値がデフォルト値を取る場合、第1のビンのみを使用して、現在考慮されているインデックス値を符号化するように構成される。

【0170】

加えて、エンコーダーは、現在考慮されているインデックス値がデフォルト値を取らない場合、現在考慮されているインデックス値がデフォルト値を取らないことを記述する第1のビンを使用して、かつ現在考慮されているインデックス値又は現在考慮されているインデックス値から導出された値、例えば $q-1$ を二値表現で表す1つ以上の追加のビンを使用して、現在考慮されているインデックス値を符号化するように構成される。

【0171】

任意選択で、現在考慮されているインデックス値がデフォルト値を取るかどうかを示す第1のビンは、コンテキストを使用して、例えば確率推定値を考慮して符号化され、1つ以上の追加のビンは、ビンごとに1ビットの固定長で符号化される。

【0172】

本発明の更なる実施の形態によれば、エンコーダーは、単項コードを使用して、又は短

10

20

30

40

50

縮単項コードを使用して、又は可変長コードを使用して、1つ以上のインデックス値を符号化するように構成され、任意選択で、コード長は、異なるインデックス値の発生の確率に従って選択される。

【0173】

本発明の更なる実施の形態によれば、エンコーダーは、例えば単項コード又はハフマンコードの場合、1つ以上の確率推定パラメーター、例えば選択された確率推定パラメーター又は確率推定パラメーターの選択されたタプルを指定する整数インデックスqを符号化するために使用されるビンの数又はビンの最大数を、1つ以上の確率推定パラメーターを量子化するために使用される量子化モードに応じて、例えば使用可能なパラメーター値の選択された集合又はパラメーター値の使用可能なタプルの選択された集合に適応するように変更するように構成される。

10

【0174】

本発明の更なる実施の形態によれば、エンコーダーは、例えば単項コード又はハフマンコードの場合、1つ以上の確率推定パラメーター、例えば選択された確率推定パラメーター又は確率推定パラメーターの選択されたタプルを指定する整数インデックスqを符号化するために使用されるビンの数又はビンの最大数を、ニューラルネットワークの層のパラメーターの数に応じて、又は1つ以上の確率推定パラメーターを使用して符号化されるニューラルネットワークパラメーターの数に応じて、又は層パラメーターの要素の数、例えば行列Wの成分の数、又は（転置）ベクトルbの要素の数等に応じて、例えば使用可能なパラメーター値の選択された集合又はパラメーター値の使用可能なタプルの選択された集合に適応するように変更するように構成される。

20

【0175】

本発明の更なる実施の形態によれば、エンコーダーは、1つ以上の確率推定パラメーターに関連付けられた使用可能なパラメーター値の異なる集合の間で、又は複数の確率推定パラメーターに関連付けられた使用可能なパラメーター値のタプルの異なる集合の間で切り替えるように構成される。

【0176】

本発明の更なる実施の形態によれば、エンコーダーは、例えば単項コード又はハフマンコードの場合、1つ以上の確率推定パラメーターに関連付けられた使用可能なパラメーター値の異なる集合の間、又は複数の確率推定パラメーターに関連付けられた使用可能なパラメーター値のタプルの異なる集合の間の切り替えに従って、選択された確率推定パラメーター又は確率推定パラメーターの選択されたタプルを指定する1つ以上の確率推定パラメーター、例えば整数インデックスqを符号化するために使用されるビンの数又はビンの最大数を変更するように構成される。

30

【0177】

本発明の更なる実施の形態によれば、エンコーダーは、1つ以上の状態変数、例えば s_i^k 又は s_k を決定し、1つ以上の状態変数を使用して、例えば式（1）及び（2）を使用して、又は式（3）及び（4）を使用して、又は式（3）及び（5）を使用して、又は1つ以上の状態変数と確率推定値、例えば $P(x, i, k)$ との間の線形関係を使用して、確率推定値、例えば p_k を導出するように構成される。

40

【0178】

本発明の更なる実施の形態によれば、エンコーダーは、

【数15】

$$s_k = \sum_{i=1}^N [s_i^k \cdot d_i^k]$$

【数 1 6】

$$p_k = \begin{cases} LUT2[\lfloor s_k \cdot a_k \rfloor], & \lfloor s_k \cdot a_k \rfloor \geq 0 \text{ である場合} \\ 1 - LUT2[-\lfloor s_k \cdot a_k \rfloor], & \text{それ以外の場合} \end{cases}$$

に従って2つの状態変数 s_1^k 、 s_2^k から確率推定値 p_k を導出するように構成される。任意選択で、LUT2は、本明細書で説明したようなものであってよく、例えば、全ての k について、例えば、 $N=2$ 、 $d_1^k=16$ 、 $d_2^k=1$ 、及び $a_k=2^{-7}$ であり、ここで、 k はコンテキストモデルインデックスである。

10

【0179】

本発明の更なる実施の形態によれば、エンコーダーは、

【数 1 7】

$$s_i^k = \begin{cases} s_i^k + \lfloor A[z + \lfloor s_i^k \cdot m_i^k \rfloor] \cdot n_i^k \rfloor, & \text{符号化されるシンボルが1である場合} \\ s_i^k - \lfloor A[z + \lfloor -s_i^k \cdot m_i^k \rfloor] \cdot n_i^k \rfloor, & \text{符号化されるシンボルが0である場合} \end{cases}$$

20

に従って状態変数 s_1^k 、 s_2^k を更新するように構成される。ここで、 m_i^k 及び n_i^k は重み係数であり、例えば確率推定パラメータを構成し、例えば、 m_1^k は 2^{-3} に等しくてもよく、 m_2^k は 2^{-7} に等しくてもよい。

【0180】

加えて、 A は、例えば、本明細書で説明されるように定義され得る、例えば、整数値を記憶する、ルックアップテーブルであり、 z はオフセット値、例えば所定の値であり、例えば16に等しくてもよい。任意選択で、 s_i^k は、例えば、本明細書で説明されるように初期化されてもよい。

30

【0181】

本発明の更なる実施の形態によれば、エンコーダーは、異なるニューラルネットワークパラメータの符号化のために異なる確率推定パラメータ値、例えば n_i^k の異なる値を使用するように、及び／又は異なるコンテキストモデルに関連付けられたビンの符号化のために異なる確率推定パラメータ値を使用するように、及び／又はニューラルネットワークの異なる層に関連付けら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に異なる確率推定パラメータ値を使用するように、重み係数 n_i^k を変更するように構成される。

【0182】

本発明の更なる実施の形態によれば、重み係数 n_i^k と適応パラメータ $s_{h_i}^k$ との関係は、

40

【数 1 8】

$$n_i^k = 2^{-s_{h_i}^k + 4}$$

に従って定義される。

【0183】

本発明の更なる実施の形態によれば、エンコーダーは、適応パラメータを記述する情

50

報、例えば適応パラメーターのタプルを記述するインデックス q を符号化するように構成され、例えば、符号化インデックス値 q の意味は、例えば、表2又は表3、又は表4又は表5又は表6又は表7に提供されるように定義されてもよい。

【0184】

デコーダーに関連して上記で説明され、以下で説明される全ての特徴は、それぞれ本発明の実施の形態によるエンコーダーの特徴として理解されるべきである。逆に、エンコーダーのコンテキストで説明される特徴は、それぞれデコーダーの特徴として理解されるべきである。実施の形態によるデコーダー又は復号化方法の特徴は、対応するエンコーダーに対して同一又は同様に適用可能であり、逆もまた同様であることが、当業者には明らかであろう。このコンテキストでは、デコーダーはエンコーダーに対応することができ、ニューラルネットワークを復号化することは、ニューラルネットワークパラメーターを符号化することに対応することができ、前に復号化されたニューラルネットワークパラメーター又はそのピンは、前に符号化されたニューラルネットワークパラメーター又はそのピンに対応することができ、異なるニューラルネットワークパラメーターの復号化及び異なるコンテキストモデルに関連するピンの復号化は、異なるニューラルネットワークパラメーターの符号化及び異なるコンテキストモデルに関連するピンの符号化に対応することができ、ニューラルネットワークの異なる層に関連付けられたニューラルネットワークパラメーターの復号化は、ニューラルネットワークの異なる層に関連付けられたニューラルネットワークパラメーターの符号化に対応することができる。しかし、これらは対応の例にすぎず、本発明の実施の形態によるデコーダーの特徴及び利点は、本発明の実施の形態によるエンコーダーの特徴及び利点と交換可能であり得ることが当業者には明らかであろう。同じことが、実施の形態の詳細な説明における符号化／復号化の方法に関する説明及び図面にも当てはまる。

【0185】

本発明の更なる実施の形態によれば、符号化表現は、異なるニューラルネットワークパラメーターに関連付けられた別個の確率推定パラメーター、及び／又は異なるコンテキストモデルに関連付けられた別個の確率推定パラメーター、及び／又はニューラルネットワークの異なる層に関連付けられた別個の確率推定パラメーターの別個の符号化表現を含む。

【0186】

復号化効率を向上させるために、別個の符号化表現を用いて確率推定パラメーターの適応選択を実行することができる。

【0187】

本発明の更なる実施の形態によれば、符号化表現は、1つ以上の確率推定パラメーターを表す符号化値、例えば符号化インデックス値 q を1つ以上の確率推定パラメーター、例えば $s_{h_i}^k$ にマッピングするために、複数のマッピング規則のうちのどのマッピング規則を使用すべきかを示すフラグ指示を含む。

【0188】

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

【図面の簡単な説明】

【0189】

【図1】 フィードフォワードニューラルネットワークのグラフ表現の一例を示す図である。

【図2】 本発明の実施形態によるデコーダーのブロック概略図である。

【図3】 本発明の実施形態による確率推定パラメーターの選択の一例の概略図である。

【図4】 本発明の実施形態によるデコーダー選択エンティティの一例の概略図である。

【図5】 本発明の実施形態による、符号化ビットストリーム及び確率推定パラメーター値を記述するインデックス値の一例の概略図である。

【図6】 本発明の実施形態による方法の概略ブロック図である。

【発明を実施するための形態】

【0190】

同一若しくは同等の要素、又は同一若しくは同等の機能を有する要素は、異なる図に現れる場合であっても、以下の説明において同一又は同等の参照番号によって示される。

【0191】

以下の説明では、本発明の実施形態のより完全な説明を提供するために、複数の詳細が記載される。しかしながら、本発明の実施形態がこれらの具体的な詳細なしに実施され得ることは、当業者には明らかであろう。他の例では、本発明の実施形態を不明瞭にすることを回避するために、周知の構造及びデバイスは、詳細にではなくブロック図の形態で示される。加えて、本明細書で後に説明される異なる実施形態の特徴は、特に別段の記載がない限り、互いに組み合わせることができる。

10

【0192】

図2は、本発明の実施形態によるデコーダーのブロック概略図を示している。デコーダー200は、コンテキスト依存算術復号化ユニット210と、確率推定器220と、確率推定パラメータ値230とを備える。任意選択で、デコーダー200は、ビットストリーム分解ユニット240及びパラメータ再アセンブリユニット250を備える。デコーダー200は、符号化ビットストリーム202を受信するよう構成され得る。符号化ビットストリーム202は、複数のニューラルネットワークパラメータに関する情報を含み得る。

【0193】

任意選択のビットストリーム分解ユニット240は、符号化ビットストリーム202を、コンテキスト依存算術復号化ユニット210のための処理可能な情報に変換するように構成され得る。この機能は、復号化ユニット210によって提供され得るため、分解ユニット240は、説明の目的でのみ示されている。分解ユニットは、符号化ビットストリーム202を、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情報を含む一部分と、ビットストリームの開始及び／又は誤り訂正若しくは他のオーバーヘッドのためのビットを示す別の部分、例えばフラグとに分解するよう構成することができる。

20

【0194】

復号化ユニット210は、複数の、例えば復号化されたニューラルネットワークパラメータ204を提供するために、符号化されたニューラルネットワークパラメータを復号化するように構成することができる。ニューラルネットワークパラメータを復号化するために、復号化ユニット210は、確率推定器220から確率推定値を受信するように構成される。ニューラルネットワークパラメータは、ビンのシーケンスとしてビットストリーム202中に符号化され得る。1つのビン又は複数のビンは、ニューラルネットワークパラメータを表すことができる。ビンは、例えば、コンテキスト、又は換言すれば確率モデルに関連付けられ得る。確率推定値は、ビンが或る特定の値、例えば1又は0を有する確率を示すことができる。確率推定値は、ビンのコンテキスト、又は換言すればその確率モデルに応じて決定することができる。確率推定器220は、確率推定値を決定するために、確率推定パラメータを含む。

30

【0195】

デコーダー又はコンテキスト依存算術復号化ユニット210は、異なるニューラルネットワークパラメータの復号化のために異なる確率推定パラメータ値を使用するように構成される。したがって、確率推定パラメータを適応させることによって、ニューラルネットワークパラメータの確率的な個々の特性を考慮に入れることができる。加えて、又は代替的に、デコーダー又はコンテキスト依存算術復号化ユニット210は、異なるコンテキストモデルに関連付けられたビンの復号化のために異なる確率推定パラメータ値を使用するように構成することができる。個々のビン、又はビンの集合は、コンテキストモデルに関連付けることができる。コンテキストモデルは、例えば、最近復号化されたニューラルネットワークパラメータ又はそのビンに従って適応され得る。その結果、個々のビン、又はビンに関連付けられた個々のコンテキストモデルについて、確率推定器220のパラメータ化を適応させることができる。任意選択のパラメータ再アセンブリユ

40

50

ニット250は、複数のニューラルネットワークパラメーター204を与えるために、復号化されたピンをニューラルネットワークパラメーターに再アセンブルするように構成することができ、及び／又は復号化ユニット210によって与えられる復号化されたエンティティを解釈するように構成することができる。任意選択で、確率推定器220は、復号化ユニット210の出力から、及び／又は任意選択のパラメーター再アセンブリユニット250から、フィードバック情報を受信することができる。

【0196】

しかしながら、デコーダー200は、代替的に、又は加えて、ニューラルネットワークの異なる層に関連付けられたニューラルネットワークパラメーターの復号化のために異なる確率推定パラメーター値を使用するように構成することができる。或る特定のニューラルネットワークパラメーターの層の情報は、ビットストリームに符号化されてもよく、確率推定パラメーター値230の変更をトリガーしてもよい。

【0197】

図3は、本発明の実施形態による確率推定パラメーターの選択の一例の概略図を示している。図3には、一例として、手法1～4の確率推定パラメーターと、コンテキストモデルを更新するための確率推定パラメーターとが示されている。デコーダーは、1つ以上の確率推定パラメーター、例えばパラメーター310及び／又は320を基集合300から選択することができる。パラメーター310及び320で示されるように、これらのパラメーターは、基集合300の真の部分集合であり得る。

【0198】

結果として、デコーダーは、異なる確率推定パラメーター値だけでなく、異なる確率推定パラメーターも使用するように構成することができる。デコーダーは、手法1～4のいずれかを使用することができるため、基集合300の確率推定パラメーターの部分集合のみを必要とする。

【0199】

加えて、デコーダーは、確率推定パラメーターを選択するだけでなく、その値も選択することができる。例えば、第1の選択ステップにおいて、デコーダーはパラメーター310を選択してもよい。次のステップにおいて、デコーダーは、これらの確率推定パラメーターについて、確率推定パラメーター値の異なる部分集合から選択することもできる。換言すれば、各確率推定パラメーターについて、許容可能な確率推定パラメーター値の複数の集合が存在してもよく、デコーダーは、確率推定パラメーター及び対応する値を選択してもよい。

【0200】

図4は、本発明の実施形態によるデコーダー選択エンティティの一例の概略図を示している。デコーダーは、使用可能なパラメーター値の異なる集合410、420から1つ以上の確率推定パラメーターを選択するように構成することができ、一例として、集合410は、部分集合412（手法3によるパラメーター）及び部分集合414を含む。手法1～4及び状態変数の更新について、ニューラルネットワークパラメーター自体の符号化及び／又は復号化のコンテキストで説明したが、コンテキストモデルの類似又は同等の手法及び更新を、確率推定パラメーターの符号化及び／又は復号化のために実行することができ、したがって、これらのパラメーターがここで示されていることに留意されたい。加えて、又は代替として、デコーダーは、パラメーター値430、440のタプルの使用可能な集合から、及び／又はここでは表5（450）、及び表7（460）の形態で一例として示される異なるマッピング規則から選択してもよい。確率推定パラメーター値、タプル又はマッピングの選択は、量子化モードに基づいて実行され得る。使用可能なパラメーター値の異なる集合は、同じ確率推定パラメーターを含むが、異なる値を有することができることに留意されたい。デコーダーは、第1の集合412と、一例として値 $d_1^k = 15$ 、 $d_2^k = 2$ 、 $a_k = 1.5^{-7}$ 及び $LUT = LUT2$ を含む別の集合414との間で選択することができる。換言すれば、適応は、符号化／復号化に使用されるパラメーターに関して、及びパラメーターのそれぞれの値に関して実行され得る。しかしながら、実施形態

10

20

30

40

50

によれば、値の選択のみ、例えば、集合 4 1 2 と 4 1 4 との間のデコーダー決定が実行されてもよく、パラメーターの選択は、手法 3 の示されたパラメーターに限定されないことに留意されたい。

【0201】

例えば、デコーダーは、1つ以上の確率推定パラメーターの一樣量子化が使用される場合に、使用可能なパラメーター値の第1の集合 4 1 0 から、又はパラメーター値の使用可能なタブルの第1の集合 4 3 0 から、1つ以上の確率推定パラメーターを抜粋して選択することができる。一例として、デコーダーは、一樣量子化の場合、集合 4 1 2 と 4 1 4 との間で選択することができる。一方、デコーダーは、1つ以上の確率推定パラメーターの可変量子化が使用される場合に、使用可能なパラメーター値の第2の集合 4 2 0 から、又はパラメーター値の使用可能なタブルの第2の集合 4 4 0 から、1つ以上の確率推定パラメーターを抜粋して選択することができる。同様に、デコーダーは、1つ以上の確率推定パラメーターの一樣量子化が使用される場合に、1つ以上の確率推定パラメーターを表す符号化値を1つ以上の確率推定パラメーターにマッピングする第1のマッピング規則 4 6 0 を使用することができ、1つ以上の確率推定パラメーターの可変量子化が使用される場合に、1つ以上の確率推定パラメーターを表す符号化値を1つ以上の確率推定パラメーターにマッピングする第2のマッピング規則 4 5 0 を使用することができる。第1の集合、タブル、及びマッピング規則並びに第2の集合、タブル、及びマッピング規則は、互いに異なってもよい。

【0202】

異なる集合、タブル、及びマッピング規則は、例えば、異なる手法 1～4 に従って計算を実行するために、異なる確率推定パラメーターを含んでもよく、又は同じ確率推定パラメーターを含むが、異なる値を有してもよいことに留意されたい。したがって、デコーダーは、任意選択で、最初に計算ルーチンを選択し、次にそのパラメーター化、すなわちその値を選択することができる。一方、デコーダーは、量子化モードに従って確率推定パラメーター値を選択するだけであってもよく、例えば、集合 4 1 2 と集合 4 1 4 との間で選択するか、又は、例えば、同じ確率推定パラメーターを記述するが異なる値を有するタブル 4 3 0 とタブル 4 4 0 との間で選択する。

【0203】

別の任意選択の特徴として、平均して、使用可能なパラメーター値の第2の集合のうちの使用可能なパラメーター値、又は第2のマッピング規則の使用可能なパラメーター値は、使用可能なパラメーター値の第1の集合のうちの使用可能なパラメーター値、又は第1のマッピング規則の使用可能なパラメーター値よりも、例えばビン値の頻度の変化に対して、確率推定値のより高速な適応を可能にすることができる。代替的に、平均して、パラメーター値の使用可能なタブルの第2の集合のパラメーター値の使用可能なタブル、又は第2のマッピング規則の使用可能なタブルは、パラメーター値の使用可能なタブルの第1の集合のパラメーター値の使用可能なタブル、又は第1のマッピング規則の使用可能なタブルよりも、例えばビン値の頻度の変化に対して、確率推定値のより高速な適応を可能にすることができる。

【0204】

任意選択で、使用可能なパラメーター値の第2の集合、又は第2のマッピング規則は、使用可能なパラメーター値の第1の集合のうちの使用可能なパラメーター値、又は第1のマッピング規則の使用可能なパラメーター値よりも、又は更には全ての使用可能なパラメーター値よりも、例えばビン値の頻度の変化に対して、確率推定値のより高速な適応を可能にする使用可能なパラメーター値を含むことができる。代替的に、パラメーター値の使用可能なタブルの第2の集合は、パラメーター値の使用可能なタブルの第1の集合のパラメーター値の使用可能なタブルよりも、又は更には全ての使用可能なタブルよりも、例えばビン値の頻度の変化に対して、確率推定値のより高速な適応を可能にするパラメーター値の使用可能なタブルを含む。

【0205】

さらに、集合 4 1 0 及び 4 2 0 は使用可能な、又は例えば許容可能なパラメーター値であってもよく、タプル 4 3 0、4 4 0 は使用可能な、又は例えば許容可能なタプルであってもよい。かかる集合又はタプルのデコーダー選択は、ニューラルネットワークの層のパラメーターの数に基づいて、若しくはそれらに依存して、又は選択された 1 つ以上の確率推定パラメーターを使用して復号化されるニューラルネットワークパラメーターの数に依存して、又は層パラメーターの要素の数に依存して実行することができる。同様に、異なるマッピング規則 4 5 0、4 6 0 の使用は、選択された 1 つ以上の確率推定パラメーターを使用して復号化されるニューラルネットワークパラメーターの数に依存して、又は層パラメーターの要素の数に依存して、デコーダーによって実行することができる。

【0206】

別の任意選択の特徴として、デコーダーは、ニューラルネットワークの層のパラメーターの数が閾値、例えば $X = 1000$ を下回る場合、又は選択された 1 つ以上の確率推定パラメーターを使用して復号化されるニューラルネットワークパラメーターの数が閾値を下回る場合、又は層パラメーターの要素の数が閾値を下回る場合、使用可能なパラメーター値の第 1 の集合から、又はパラメーター値の使用可能なタプルの第 1 の集合から 1 つ以上の確率推定パラメーターを抜粋して選択するように構成することができる。

【0207】

さらに、デコーダーは、ニューラルネットワークの層のパラメーターの数が閾値を超える場合、又は選択された 1 つ以上の確率推定パラメーターを使用して復号化されるニューラルネットワークパラメーターの数が閾値を超える場合、又は層パラメーターの要素の数が閾値を超える場合、使用可能なパラメーター値の第 2 の集合から、又はパラメーター値の使用可能なタプルの第 2 の集合から 1 つ以上の確率推定パラメーターを抜粋して選択するように構成することができる。

【0208】

代替的に、デコーダーは、ニューラルネットワークの層のパラメーターの数が閾値を下回る場合、又は選択された 1 つ以上の確率推定パラメーターを使用して復号化されるニューラルネットワークパラメーターの数が閾値を下回る場合、又は層パラメーターの要素の数が閾値を下回る場合、1 つ以上の確率推定パラメーターを表す符号化値を 1 つ以上の確率推定パラメーターにマッピングする第 1 のマッピング規則を抜粋して使用するよう構成される。デコーダーは、ニューラルネットワークの層のパラメーターの数が閾値を超える場合、又は選択された 1 つ以上の確率推定パラメーターを使用して復号化されるニューラルネットワークパラメーターの数が閾値を超える場合、又は層パラメーターの要素の数が閾値を超える場合、1 つ以上の確率推定パラメーターを表す符号化値を 1 つ以上の確率推定パラメーターにマッピングする第 2 のマッピング規則を抜粋して使用するよう構成することができる。

【0209】

この場合、使用可能なパラメーター値の第 2 の集合は、使用可能なパラメーター値の第 1 の集合よりも多くの使用可能なパラメーター値を含むことができ、パラメーター値の使用可能なタプルの第 2 の集合は、パラメーター値の使用可能なタプルの第 1 の集合よりも多くの使用可能なタプルを含むことができる。加えて、又は代替的に、第 2 のマッピング規則は、第 1 のマッピング規則とは異なり得る。

【0210】

図 5 は、本発明の実施形態による、符号化ビットストリーム及び確率推定パラメーター値を記述するインデックス値の一例の概略図を示している。一例として、ビットストリーム 202 は、フラグ指示 F、又は換言すればフラグ F の形態のシグナリングを含む。しかしながら、シグナリングは、任意の適切な方法で伝送されてもよい。デコーダーは、1 つ以上の確率推定パラメーターが、使用可能なパラメーター値のどの集合から、又はパラメーター値の使用可能なタプルのどの集合から選択されるかを判定するために、フラグ F を評価し得る。代替的に、デコーダーは、1 つ以上の確率推定パラメーターを表す符号化値を 1 つ以上の確率推定パラメーターにマッピングするために、複数のマッピング規則のう

10

20

30

40

50

ちのどのマッピング規則を使用すべきかのシグナリング指示を評価するように構成され得る。したがって、図4におけるデコーダーによる集合、タプル及び／又はマッピングの選択は、例えば符号化されたフラグFの形のシグナリングに基づき得る。任意選択のビットストリーム分解ユニット240は、例えば、復号化のために使用される集合、タプル、及び／又はマッピングを示すためのフラグにおいて、かつ確率推定パラメータ値に関する情報において、ビットストリーム202を分解することができる。

【0211】

図5に示されているように、ビットストリーム202は、確率推定パラメータ値を記述する、又は複数の確率推定パラメータ値を記述する、又は確率推定パラメータ値のタプルを記述する1つ以上のインデックス値 q_i （ここでは一例として、 $i = 1, 2$ 及び3について示されている）、例えば整数値を含むことができる。換言すれば、インデックス値 q_i は、1つ以上の確率推定パラメータの符号化表現であり得る。

10

【0212】

デコーダーは、例えばシグナリングを使用して、1つ以上のインデックス値 q_i を復号化するように構成されてもよい。加えて、1つ以上のインデックス値 q_i は、1つ以上のコンテキストモデル c_{q_i} に関連付けられ得る。デコーダーは、コンテキストモデル c_{q_i} を使用してインデックス値 q_i を復号化するように構成することができる。

【0213】

インデックス値は、1つ以上のビンによって表すことができる。第1のビンは、例えば、`fb in`で示されるように、現在考慮されているインデックス値がデフォルト値を取るかどうかを記述することができる。インデックス値がデフォルト値を取る場合、インデックス値は第1のビンによって既に決定されているので、インデックス値は1つのビンのみを含み得る。その他の場合、インデックス値は、例えば $j = 1, 2, 3$ で示される一例のように、ビン`ad d b i n j`の形式で、1つ以上の追加のビンで表されてもよい。デコーダーは、第1のビン及び任意選択の追加のビンを復号化するように構成することができる。これらのビンのいずれも、例えば個々に、コンテキストに関連付けられてもよい。任意選択で、インデックス値 q_i のコンテキスト c_{q_i} は、インデックス値の第1のビンに関連付けられてもよく、追加のビンは、ビンごとに固定長で復号化されてもよい。

20

【0214】

加えて、ビットストリーム202は、 $i = 1, 2, 3$ で示される一例のように、ニューラルネットワークパラメータに関連付けられた整数の倍数 r_i を含む。インデックス値 q_i に基づいて、符号化されたニューラルネットワークパラメータの算術復号化のコンテキストの適応を実行することができ、該ニューラルネットワークパラメータは、整数の倍数 r_i によって表される、ニューラルネットワークの符号化重みパラメータであり得る。

30

【0215】

別の任意選択の特徴として、デコーダーは、単項コード復号化を使用して、又は短縮単項コード復号化を使用して、又は可変長コード復号化を使用して、1つ以上のインデックス値を復号化するように構成することができ、例えば、コード長は、異なるインデックス値の発生確率に従って選択される。実施形態によれば、任意の適切な符号化技術を、例えば、異なるインデックス値に対して個別に適用することができ、柔軟性及び符号化効率を向上させることができる。

40

【0216】

加えて、デコーダーは、1つ以上の確率推定パラメータを量子化するために使用される量子化モードに依存して、及び／又はニューラルネットワークの層のパラメータの数に依存して、又は1つ以上の確率推定パラメータを使用して復号化されるニューラルネットワークパラメータの数に依存して、又は層パラメータの要素の数に依存して、1つ以上の確率推定パラメータを復号化するために使用されるビンの数又はビンの最大数を変更するように構成することができる。

【0217】

50

別の任意選択の特徴として、デコーダーは、1つ以上の確率推定パラメーターに関連付けられた使用可能なパラメーター値の異なる集合の間で、又は複数の確率推定パラメーターに関連付けられた使用可能なパラメーター値のタプルの異なる集合の間で、又は1つ以上の確率推定パラメーターを表す符号化値を1つ以上の確率推定パラメーターにマッピングするための異なるマッピング規則の間で切り替えるように構成することができる。デコーダーは、図4に示すように、集合410（及び／又は集合412、414）と420との間で、及び／又はタプル430、440及び／又は異なるマッピング450、460の間で切り替えることができる。

【0218】

さらに、前述したビンの数又はビンの最大数の変更は、異なる集合、タプル及び／又はマッピング規則間の切り替えに従ってデコーダーによって実行することができる。

10

【0219】

加えて、デコーダーは、1つ以上の状態変数、例えば s_i^k 又は s_k を決定し、1つ以上の状態変数を使用して確率推定値、例えば p_k を導出するように構成することができる。

【0220】

さらに、符号化ビットストリーム202は、ニューラルネットワークの重みパラメーターの符号化表現とすることができ、整数の倍数 r_i の形態のニューラルネットワークの複数の符号化重みパラメーターと、1つ以上の確率推定パラメーター、すなわちインデックス値 q_i の符号化表現とを含む。

【0221】

20

図示のように、符号化ビットストリーム202の形態の符号化表現は、別個の確率推定パラメーター、すなわちインデックス値 q_i （ここでは一例として $i = 1, 2$ 及び3について示す）の別個の符号化表現を含み得る。これらのインデックス値 q_i は、異なるニューラルネットワークパラメーター、例えば、 $q_1 \rightarrow r_1$ 、 $q_2 \rightarrow r_2$ 、... に関連付けることができる。代替的に、又は加えて、図示のように、別個の確率推定パラメーター q_i を異なるコンテキストモデル c_{q_i} に関連付けることができる。別の任意選択の特徴として、別個の確率推定パラメーターをニューラルネットワークの異なる層に関連付けることができる。

【0222】

図6は、本発明の実施形態による方法の概略ブロック図を示している。図6は、ニューラルネットワークの重みパラメーターを復号化する方法600、700を示している。方法600及び700は、符号化ビットストリームに基づいて、複数のニューラルネットワークパラメーター、例えば、ニューラルネットワークの行列 W の成分 w_i 、 b 、 μ 、 σ^2 、 σ 、 γ 、及び／又は β のうちの少なくとも1つを取得（610、710）することと、コンテキスト依存算術復号化を使用して、例えば、コンテキスト適応型二値算術復号化（CABAC）を使用して、ニューラルネットワークのニューラルネットワークパラメーター、例えば、その量子化バージョンを復号化（620、720）することとを含む。任意選択で、ビン値の確率は、異なるコンテキストに対して決定され、例えば、各ビンはコンテキストに関連付けられる。方法600、700は、1つ以上の確率推定パラメーター、例えば確率推定器パラメーター、例えば N 、 a_i^k 、 b_i^k 、 a_k 、 d_i^k 、 A 、 m_i^k 、 n_i^k 、 sh_i^k 、 $initVal_i^k$ を使用して、例えば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ー又はそのビンに基づいて、ニューラルネットワークパラメーターの数表現のビンの任意選択で算術復号化のために、例えばコンテキストに関連付けられ得る確率推定値、例えば $P(t)$ 又は p_k を取得（630、730）することを更に含む。

30

40

【0223】

方法600は、異なるニューラルネットワークパラメーターの復号化のために異なる確率推定パラメーター値を使用（640）すること、及び／又は異なるコンテキストモデル、例えば c_k に関連付けられたビンの復号化のために異なる確率推定パラメーター値を使用することを更に含む。

50

【0224】

一方、方法700は、ニューラルネットワークの異なる層に関連付けられたニューラルネットワークパラメーターの復号化のために異なる確率推定パラメーター値を使用（740）することを更に含む。

【0225】

更なる実施形態及び態様

以下では、前述の実施形態のいずれかに組み込まれ得る態様及び特徴を含む更なる実施形態が開示される。

【0226】

パラメーターの効率的な表現（例、詳細は任意である）

パラメーター W 、 b 、 μ 、 σ^2 、 γ 、及び β は、層のパラメーター又は層パラメーターと総称されるものとする。それらは通常、ビットストリームにおいて（例えば、ニューラルネットワークがビデオデコーダーにおいて使用される場合、例えば、符号化されたビデオ表現において）シグナリングされる必要がある。例えば、それらは32ビットの浮動小数点数として表すことができ、又は例えば、量子化インデックスとしても示される整数表現に量子化することができる。 ϵ は通常、ビットストリーム中でシグナリングされないことに留意されたい。

【0227】

例えば、かかるパラメーターを符号化するための特に効率的な手法では、例えば、各値がいわゆる量子化ステップサイズ値の整数倍として表される、一様再構成量子化器（URQ）が用いられる。対応する浮動小数点数は、例えば、整数に量子化ステップサイズを乗算することによって再構成することができ、量子化ステップサイズは、通常（必ずしもそうではないが）単一の浮動小数点数である。しかしながら、例えば、ニューラルネットワーク推論のための効率的な実施態様（すなわち、入力に対するニューラルネットワークの出力を計算すること）では、可能な限り整数演算が採用されている。したがって、パラメーターを浮動小数点表現に再構成することを要求することは望ましくない場合がある。

【0228】

パラメーターを符号化するための別の効率的な手法では、量子化器のセットが適用され、各値は、例えば、量子化ステップサイズ値の整数倍として表される。通常、例えば、集合内の各量子化器は、適用可能な再構成値として量子化ステップサイズパラメーターの整数倍の互いに素な集合を使用するが、2つ以上の量子化器が1つ以上の再構成値を共有してもよい。適用される量子化器は、例えば、符号化順で前の量子化インデックスの値に依存する。対応する浮動小数点数は、例えば、整数に量子化ステップサイズを乗算することによって再構成することができ、量子化ステップサイズは、通常、例えば、選択された量子化器に依存する浮動小数点数である。かかる量子化器の設計の一例は、トレリス符号化量子化（TCQ）であり、これは、従属量子化（DQ）とも称される。

【0229】

好ましい実施形態では、2つの量子化器のセットが使用される。第1の量子化器は、例えば、0を含む量子化ステップサイズの全ての偶数倍を使用し、第2の量子化器は、0を含む量子化ステップサイズの全ての偶数倍を使用する。

【0230】

エントロピー符号化及び確率推定（例、詳細は任意である）

例えば、量子化方法によって出力される量子化インデックスは、適切なエントロピー符号化方法を使用してエントロピー符号化される。

【0231】

かかる量子化インデックスを符号化するための特に適切なエントロピー符号化方法は、CABACとも表されるコンテキストベース適応型二値算術符号化である。これにより、各量子化インデックスは、例えば、二値決定のシーケンス、いわゆるピンに分解される。通常、例えば、各ピンは、コンテキストモデルとも称される確率モデルに関連付けられ、該コンテキストモデルは、例えば確率推定方法を使用して、関連付けられたピンの統計を

10

20

30

40

50

モデル化する。

【0232】

確率推定器は、例えば、確率推定器に関連する既に符号化されたピンに基づいて、ピンが x に等しい確率 $P(t)$ をモデル化する装置であり、ここで、 $x \in \{0, 1\}$ である。

【0233】

例えば、確率推定器は、確率推定器パラメーター又は推定器パラメーター（又は確率推定パラメーター）と称される一部のパラメーターを有し、これらのパラメーターは、確率推定値、例えば適応レートに影響を与える。通常、これらの推定器パラメーターは、例えば、適応シナリオ、例えばニューラルネットワークパラメーターの符号化に応じて、グローバルに選択される。したがって、例えば、ニューラルネットワーク符号化では、各ニューラルネットワークパラメーターは、推定器パラメーターの同じ集合を適用する。

【0234】

しかし、現在のニューラルネットワークパラメーターに対して最適化された推定器パラメーターを選択することによって圧縮効率が改善され得ることが分かっている。したがって、一態様によれば、基本的な着想は、パラメーターの集合から適切な推定器パラメーターを選択し、次いで、推定器パラメーターがデコーダーにシグナリングされることである。

【0235】

典型的な推定器設計（例、詳細は任意である）

最初に、ニューラルネットワーク圧縮に適用される典型的な推定器の設計について説明する。

【0236】

例えば、各コンテキストモデル c_k について、1つ以上の状態変数 $s_1^k, \dots, s_N^k$ が $N \geq 1$ で維持される。各状態変数 s_i^k は、例えば、符号付き整数値として実装され、例えば、確率値 $P(s_i^k, i, k) = p_i^k$ を表す。コンテキストモデル c_k の確率推定値 p_k は、例えば、コンテキストモデルの全ての状態変数の確率値 p_i^k の加重和として定義されるものとする。

【0237】

状態変数は、以下の特性を有することが好ましいが、必ずしもそうである必要はない。

1. $s_i^k = 0$ である場合、 $p_i^k = 0.5$ である。
2. s_i^k のより大きな値は、より大きな p_i^k に対応する。
3. $P(-s_i^k, i, k) = 1 - P(s_i^k, i, k)$ 。

【0238】

したがって、負の状態変数は、例えば、 $p_i^k < 0.5$ に対応することができる。概して、各コンテキストモデルの各状態変数に対して異なる関数 $P(\cdot)$ を指定することが可能である。

【0239】

状態変数と確率値とを対応付けるための構成例（例、詳細は任意である）

状態変数を確率値に関連付ける、すなわち $P(\cdot)$ を実装する多くの有用な方法が存在する。例えば、ニューラルネットワーク圧縮で使用される状態表現は、以下の式で達成することができる。

【数19】

$$P(x, i, k) = \begin{cases} 0.5 \cdot \alpha^{\lfloor x \cdot \beta_i^k \rfloor}, & x \geq 0 \text{ である場合} \\ 1 - 0.5 \cdot \alpha^{\lfloor -x \cdot \beta_i^k \rfloor}, & \text{それ以外の場合} \end{cases}$$

β_i^k は重み係数である。

α は、 $0 < \alpha < 1$ のパラメーターである。

【0240】

例えば、2つの状態 ($N=2$, s_1^k , s_2^k) を使用する、マルチメディアコンテンツ記述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規格の現在のドラフトで使用されるものに匹敵する構成を達成するために、 $\alpha = 0.99894079$ 及び全ての k に対して $\beta_1^k = 16$ 及び $\beta_2^k = 1$ を設定する。

【0241】

この例示的な構成は、状態変数がどのように定義され得るかについての何らかの洞察を与えるものである。一般に、 $P(\cdot)$ は直接使用されないので定義する必要はないが、以下に見られるように、 $P(\cdot)$ の定義も必要である。むしろ、個々の部分の実際の実施態様に起因することが多い。

【0242】

状態変数の初期化 (例、詳細は任意である)。

コンテキストモデルを用いて第1のシンボルを符号化又は復号化する前に、全ての状態変数は、 $initVal_i^k$ として示される同じ値で初期化され、これは、任意選択で、例えば、圧縮アプリケーションに最適化することができる。

【0243】

状態変数からの確率推定値の導出 (例、詳細は任意である)。

シンボルの符号化又は復号化のために、コンテキストモデルの状態変数から確率推定値が導出される。例として、3つの代替的な手法を以下に提示する。手法1は、手法2及び手法3よりも正確な結果をもたらすが、計算の複雑さも高くなる。

【0244】

手法1 (例)

この手法は2つのステップから構成される。まず、コンテキストモデルの各状態変数 s_i^k が確率値 p_i^k に変換される。次に、確率推定値 p_k が、確率値 p_i^k の加重和として導出される。

【0245】

ステップ1:

ルックアップテーブルLUT1は、例えば式(1)に従って、状態変数 s_i^k を対応する確率値 p_i^k に変換するために使用される。

【数20】

$$p_i^k = \begin{cases} LUT1[\lfloor s_i^k \cdot a_i^k \rfloor], & s_i^k \geq 0 \text{ である場合} \\ 1 - LUT1[\lfloor -s_i^k \cdot a_i^k \rfloor], & \text{それ以外の場合} \end{cases} \quad (1)$$

LUT1は、確率値を含むルックアップテーブルである。 a_i^k は、 s_i^k をLUT1のサイズに適應させる重み係数である。

【0246】

ステップ2:

確率推定値 p_k は、例えば次式に従って、確率値 p_i^k から導出される。

【数21】

$$p_k = \sum_{i=1}^N p_i^k \cdot b_i^k \quad (2)$$

b_i^k は、個々の状態変数の影響を制御する重み係数である。

【0247】

手法 2 (例)

状態変数から確率推定値を導出するための代替的な手法を以下に示す。該手法では、精度の低い結果が得られ、計算の複雑さが低くなる。最初に、状態変数の加重和 s_k が、例えば、次式に従って導出される。

【数 2 2】

$$s_k = \sum_{i=1}^N [s_i^k \cdot d_i^k] \quad (3)$$

d_i^k は、各状態変数の影響を制御する重み係数である。

【0 2 4 8】

次に、確率推定値 p_k は、状態変数 s_k の加重和から、例えば次式に従って導出される。

【数 2 3】

$$p_k = \begin{cases} LUT2[|s_k \cdot a_k|], & s_k \geq 0 \text{ である場合} \\ 1 - LUT2[|s_k \cdot a_k|], & \text{それ以外の場合} \end{cases} \quad (4)$$

LUT2 は、確率推定値を含むルックアップテーブルであり、 a_k は、 s_k を LUT2 のサイズに適應させる重み係数である。

【0 2 4 9】

手法 3 (例)

状態変数から確率推定値を導出するための更に別の手法を以下に提示する。まず、状態変数の重み付き和 s_k を、例えば手法 2 のように導出する。次に、確率推定値 p_k は、状態変数 s_k の加重和から、例えば次式に従って導出される。

【数 2 4】

$$p_k = \begin{cases} LUT2[|s_k \cdot a_k|], & [s_k \cdot a_k] \geq 0 \text{ である場合} \\ 1 - LUT2[|s_k \cdot a_k|], & \text{それ以外の場合} \end{cases} \quad (5)$$

LUT2 は、確率推定値を含むルックアップテーブルである。

【0 2 5 0】

手法 4 (例)

更なる手法では、状態値と確率 $P(x, i, k)$ との間の線形関係が用いられる。確率推定値の導出は、例えば、式 (2) の手法を使用する。手法 4 の一例は、汎用ビデオ符号化 (VVC) の現在のドラフトで使用されている確率推定方式である。

【0 2 5 1】

例えば、マルチメディアコンテンツの記述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 規格の現在のドラフトで使用される構成を達成するために、手法 3 の方法が、例えば、全ての k について $d_1^k = 16$ 、 $d_2^k = 1$ 、及び $a_k = 2^{-7}$ を用いて使用される。確率推定値を含むルックアップテーブルは、例えば、以下の通りである。

【0 2 5 2】

LUT2 = {0.5000, 0.4087, 0.3568, 0.3116, 0.2721, 0.2375, 0.2074, 0.1811, 0.1581, 0.1381, 0.1206, 0.1053, 0.0919, 0.08

10

20

30

40

50

0 3, 0. 0 7 0 1, 0. 0 6 1 2,
 0. 0 5 3 4, 0. 0 4 6 6, 0. 0 4 0 7, 0. 0 3 5 6, 0. 0 3 1 0, 0. 0 2
 7 1, 0. 0 2 3 7, 0. 0 2 0 7,
 0. 0 1 8 0, 0. 0 1 5 8, 0. 0 1 3 8, 0. 0 1 2 0, 0. 0 1 0 5, 0. 0 0
 9 2, 0. 0 0 8 0, 0. 0 0 7 0}

【0 2 5 3】

状態変数の更新（例、詳細は任意である）

シンボルの符号化又は復号化の後、コンテキストモデルの1つ以上の状態変数は、シンボルシーケンスの統計的挙動を追跡するために更新され得る。

【0 2 5 4】

更新は、例えば、以下のように行われる。

【数 2 5】

$$s_i^k = \begin{cases} s_i^k + [A[z + [s_i^k \cdot m_i^k]] \cdot n_i^k], & \text{符号化されるシンボルが1である場合} \\ s_i^k - [A[z + [-s_i^k \cdot m_i^k]] \cdot n_i^k], & \text{符号化されるシンボルが0である場合} \end{cases} \quad (6)$$

Aは、例えば整数値を格納するルックアップテーブルである。 m_i^k 及び n_i^k は、例えば、更新の「アジリティ」を制御する重み係数である。係数 n_i^k は、例えば、

【数 2 6】

$$n_i^k = 2^{-sh_i^k + 4}$$

に従って記述することができ、ここで、 sh_i^k は適応パラメーターとしても示される。 z は、例えば、ルックアップテーブルAが負でない値でのみアクセスされることを保証するオフセットである。

【0 2 5 5】

ルックアップテーブルA内の値は、例えば、 s_i^k が特定の所与の間隔内に留まるように選択することができる。通常、ルックアップAの値は、例えば更新関数に近似している。代替的に、例えば、状態更新のために関連する更新関数を単に使用することも可能である。

【0 2 5 6】

例えば、手法4に従うVVCの推定方法は、状態更新のための更新関数を適用し、ビットシフトを使用するものであり、該ビットシフトにより、例えば更新の「アジリティ」が決定される。これは、例えば、上述の適応パラメーターに対応する。本発明（下記参照）は、同様にそれらに適用され得る。

【0 2 5 7】

例えば、マルチメディアコンテンツの記述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規格の現在のドラフトで使用される構成を達成するために、パラメーターは、例えば、全ての k に対して、 $m_1^k = 2^{-3}$ 、 $m_2^k = 2^{-7}$ 及び $n_1^k = 2^{-1}$ 、 $n_2^k = 1$ 、並びに $z = 16$ となるように選択される。ルックアップテーブルAは、例えば、 $A = \{157, 143, 129, 115, 101, 87, 73, 59, 45, 35, 29, 23, 17, 13, 9, 5, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 0\}$ である。

【0 2 5 8】

シンボルを符号化する前に、 s_1^k は、例えば、間隔 $[-127, 127]$ の値で初期

化されるものとし、 s_2^k は、例えば、間隔 $[-2047, 2047]$ の値で初期化されるものとする。その結果、 s_1^k は、例えば、8ビットの符号付き整数値で実装することができ、 s_2^k は、例えば、12ビットの符号付き整数で実装することができる。

【0259】

本発明の態様（詳細は任意である）

以下では、パラメーター、すなわち、 N 、 a_i^k 、 b_i^k 、 a_k 、 d_i^k 、 A 、 m_i^k 、 n_i^k 、 sh_i^k 、 $initVal_i^k$ 、及び確率推定器（コンテキストモデル）に関連する任意の他のパラメーターは、確率推定器パラメーター又は推定器パラメーター（又は確率推定パラメーター）と総称されるものとする。

【0260】

通常、例えば、各推定器パラメーターに対して、確率推定器パラメーターの基集合から1つの固定されたインスタンスがネットワーク全体に対して選択される。基集合の値は、適用される状態の数 N に従って、 N 組の推定器パラメーターであり得る。本発明の一態様によれば、確率推定、したがって圧縮効率は、例えば、パラメーターが、層（すなわち、 W 、 b 、 μ 、 σ^2 、 γ 、及び β ）及び／又はコンテキストモデル c_k の各パラメーター又はパラメーターの部分集合に対して個々に選択される場合に改善され得る。

【0261】

使用される推定器パラメーターは、例えば、パラメーターの集合のパラメーターの中から決定され、パラメーターの集合は、例えば、基集合又は基集合の任意の部分集合であり得る。この集合の各パラメーターは、例えば、整数インデックス q に関連付けられ得る。例えば、集合の1つのパラメーターがデフォルトパラメーターとして示されてもよい。通常、デフォルトパラメーターは、例えば、0に等しい整数インデックスに関連付けられる。選択された推定器パラメーターに関連付けられたインデックスは、次いで、例えば、デコーダーにシグナリングされる。

【0262】

符号化方式（例、詳細は任意である）

符号化されるインデックス $q \in [0, q_{MAX}]$ は、例えば、ビンのシーケンスに分解され、次いで、ビンのシーケンスが符号化される。各ピンは、例えば、コンテキストモデルを使用して、又は固定確率を使用して符号化されてもよい。

【0263】

符号化手順は、例えば、以下の方式のうちの1つに従うことができる。

【0264】

1. 第1のピン、例えば $useNotDefault$ は、選択される推定器パラメーターがデフォルトパラメーターと異なる（例えば $useNotDefault=1$ ）か否か（例えば $useNotDefault=0$ ）を示す。例えば、 $useNotDefault=0$ である場合、デフォルトパラメーターが選択され、更なるピンは符号化されない。例えば $useNotDefault=1$ である限り、ビンのシーケンスが符号化され、これは例えば選択されたパラメーターのインデックスから1を引いたもの（ $q-1$ ）、 $indexMinusOne$ を示す。インデックスのために符号化されるビンの数は、例えば、 $\lfloor \log_2(setLength - 1) \rfloor$

$$\lfloor \log_2(setLength - 1) \rfloor$$

に等しく、ここで、 $setLength$ は、集合の要素の数を示す。

【0265】

2. 第2の手順では、単項コードが使用される。第1のピン、例えば $greaterThan_0$ は、確率パラメーターに関連するインデックス q が0より大きい（例えば $greaterThan_0=1$ ）か、又はそうでない（例えば $greaterThan_0=0$ ）を示す。

= 0) かを示す。例えば $\text{greaterThan_0} = 0$ である場合、更なるピンは符号化されない。例えば $\text{greaterThan_0} = 1$ である場合、別のピンが符号化され（例えば greaterThan_1 ）、これは、インデックス q が 1 よりも大きい（例えば $\text{greaterThan_1} = 1$ ）か否か（例えば $\text{greaterThan_1} = 0$ ）を示す。例えば $\text{greaterThan_1} = 0$ である場合、更なるピンは符号化されない。例えば、 $\text{greaterThan_1} = 0$ である場合、フラグ greaterThan_q が 0 に等しくなるまで、更なるピン (greaterThan_X) が同様に符号化される。

【0266】

3. この手順は、例えば、 q を符号化するためのインデックスが q_{MAX} に等しい場合を除いて、符号化方法 2. において使用される単項コードと同一である、短縮単項コードを適用する。この場合、例えば、ピン $\text{greaterThan_}(q_{\text{MAX}} - 1)$ を符号化した後、更なるピンは符号化されない。例えば、デコーダー側では、 $\text{greaterThan_}(q_{\text{MAX}} - 1)$ が 1 に等しい場合、 q の値は q_{MAX} であると推論される。

【0267】

4. この手順は、符号長がシンボルの発生確率に従って選択される可変長コード、例えばハフマンコードを使用する。

【0268】

好ましい実施形態（例、詳細は任意である）

好ましい実施形態では、推定器は、例えば、 N 組の適応パラメーター sh_i^k である適応パラメーターの基集合を適用する。次に、基集合の部分集合が選択される。部分集合のうちの 1 つのパラメーターがシグナリングされる。

【0269】

特に好ましい実施形態では、構成は、例えば、前の好ましい実施形態と等しいが、マルチメディアコンテンツの記述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 規格の現在のドラフトで使用されている推定器と同一であるように構成された推定器が使用され、基集合は、例えば、 (sh_1^k, sh_2^k) について以下の 28 個の対を含む。

【0270】

【表 9】

表 9：適応パラメーターの好ましい基集合

0	(0, 0)	14	(2, 3)
1	(0, 1)	15	(2, 4)
2	(0, 2)	16	(2, 5)
3	(0, 3)	17	(2, 6)
4	(0, 4)	18	(3, 3)
5	(0, 5)	19	(3, 4)
6	(0, 6)	20	(3, 5)
7	(1, 1)	21	(3, 6)
8	(1, 2)	22	(4, 4)
9	(1, 3)	23	(4, 5)
10	(1, 4)	24	(4, 6)
11	(1, 5)	25	(5, 5)
12	(1, 6)	26	(5, 6)
13	(2, 2)	27	(6, 6)

【0271】

サイズ 3 の部分集合は、例えば、層の全てのパラメーターが DQ で量子化される場合に、例えば、表 2 によるインデックス q が割り当てられるように定義され、順序付けられる

。インデックス $q = 0$ を有するパラメーターは、例えば、デフォルトパラメーターとして示される。

【0272】

【表10】

表10：集合サイズ3に対する適応パラメーターの好ましい部分集合

q	適応パラメーターの対
0	(1, 4)
1	(0, 1)
2	(2, 6)

10

【0273】

例えば、部分集合のうちの1つのパラメーターは、例えば、符号化方式1に従って q を符号化することによってシグナリングされ、ここで、例えば、ピン `useNotDefault` はコンテキストモデルを使用して符号化され、全ての他のピンはピンごとに1ビットの固定長で符号化される。

【0274】

別の好ましい実施形態（例）では、構成は、割り当てられた適応パラメーターの対と、選択された部分集合（表3）のサイズが5に等しい以外は、前の好ましい実施形態と同一である。

20

【0275】

【表11】

表11：集合サイズ5に対する適応パラメーターの好ましい部分集合

q	適応パラメーターの対
0	(1, 4)
1	(0, 0)
2	(0, 6)
3	(1, 1)
4	(2, 6)

30

【0276】

別の好ましい実施形態（例）では、構成は、割り当てられた適応パラメーターの対（表4）以外は、前の好ましい実施形態と同一である。

【0277】

【表12】

表12：集合サイズ5に対する適応パラメーターの第2の好ましい部分集合

q	適応パラメーターの対
0	(1, 2)
1	(0, 0)
2	(0, 5)
3	(2, 5)
4	(3, 4)

40

【0278】

別の好ましい実施形態（例）では、構成は、割り当てられた適応パラメーターの対と、選択された部分集合（表5）のサイズが9に等しい以外は、前の好ましい実施形態と同一である。

【0279】

50

【表 1 3】

表 1 3 : 集合サイズ 9 に対する適応パラメーターの好ましい部分集合

q	適応パラメーターの対
0	(1 , 4)
1	(0 , 0)
2	(0 , 5)
3	(1 , 1)
4	(1 , 2)
5	(2 , 4)
6	(2 , 6)
7	(3 , 4)
8	(3 , 5)

10

【 0 2 8 0】

別の好ましい実施形態（例）では、構成は、割り当てられた適応パラメーターの対（表 6）以外は、前の好ましい実施形態と同一である。

【 0 2 8 1】

【表 1 4】

表 1 4 : 集合サイズ 9 に対する適応パラメーターの第 2 の好ましい部分集合

d	適応パラメーターの対
0	(1 , 3)
1	(0 , 0)
2	(0 , 5)
3	(1 , 1)
4	(1 , 6)
5	(2 , 4)
6	(2 , 6)
7	(3 , 5)
8	(4 , 4)

20

30

【 0 2 8 2】

別の好ましい実施形態（例）では、構成は、割り当てられた適応パラメーターの対（表 7）、選択された部分集合のサイズ（5）、及び U R Q を使用する使用された量子化方法以外は、前の好ましい実施形態と同一である。

【 0 2 8 3】

【表 1 5】

表 1 5 : 集合サイズ 5 及び U R Q に対する適応パラメーターの好ましい部分集合

q	適応パラメーターの対
0	(1 , 4)
1	(0 , 6)
2	(1 , 1)
3	(2 , 6)
4	(3 , 4)

40

【 0 2 8 4】

別の好ましい実施形態（例）では、マルチメディアコンテンツの記述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 M P E G - 7 パート 1 7 規格の現在のドラフトで使用される推定器と同一であるように構成された推定器が使用され、表 1 の基集合が使用される。これを基本構成と称する。

【 0 2 8 5】

層パラメーターが D Q を用いて量子化される場合は常に、表 5 中のパラメーターの対の

50

(サイズ 9 の) 部分集合が適用される。層パラメーターが U R Q を用いて量子化される場合、表 8 の部分集合が使用される。

【 0 2 8 6 】

【表 1 6】

表 1 6 : 集合サイズ 9 に対する適応パラメーターの好ましい部分集合

q	適応パラメーターの対
0	(1 , 4)
1	(0 , 1)
2	(0 , 6)
3	(1 , 2)
4	(1 , 6)
5	(2 , 5)
6	(2 , 6)
7	(3 , 4)
8	(3 , 5)

10

【 0 2 8 7 】

別の好ましい実施形態 (例) では、前の好ましい実施形態の基本構成が適用される。

20

【 0 2 8 8 】

層パラメーターの要素の数が、例えば $X = 1000$ に設定され得る閾値 X を下回る場合は常に、例えば表 2 において第 1 の部分集合として示されるパラメーターの対のサイズ 3 を有する部分集合が使用される。そうではなく、層パラメーターの要素の数が閾値 X 以上である場合、第 2 の部分集合として示される、例えば、表 5 中のサイズ 9 を有する部分集合が使用される。

【 0 2 8 9 】

別の好ましい実施形態 (例) では、構成は、前の好ましい実施形態と同一であるが、閾値を使用する代わりに、フラグ (例えば、`useSecondSubset`) が符号化され、これは、例えば、使用される部分集合を決定する。例えば、フラグが 0 に等しい場合、第 1 の部分集合が使用される。フラグが 1 に等しい場合、第 2 の部分集合が使用される。

30

【 0 2 9 0 】

代替的な実施態様：

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

40

【 0 2 9 1 】

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

【 0 2 9 2 】

本発明による一部の実施形態は、電子的可読制御信号を有するデータキャリアを備え、

50

該電子的可読制御信号は、本明細書で説明される方法のうちの1つが実行されるように、プログラム可能なコンピューターシステムと協働することが可能である。

【0293】

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

【0294】

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

10

【0295】

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

【0296】

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

【0297】

したがって、本発明の方法の更なる実施形態は、本明細書に記載された方法の1つを実行するコンピュータープログラムを表すデータストリーム又はシグナルのシーケンスである。データストリーム又はシグナルのシーケンスは、例えば、データ通信接続を介して、例えばインターネットを介して転送されるように構成されてもよい。

20

【0298】

更なる実施形態は、本明細書に記載の方法の1つを実行するように構成又は適合された処理手段、例えばコンピューター又はプログラム可能な論理デバイスを含む。

【0299】

更なる実施形態は、本明細書に記載の方法のうちの1つを実行するコンピュータープログラムがインストールされたコンピューターを含む。

30

【0300】

本発明による更なる実施形態は、本明細書で説明される方法のうちの1つを実行するコンピュータープログラムを受信機に（例えば、電子的に又は光学的に）転送するように構成された装置又はシステムを含む。受信機は、例えば、コンピューター、モバイルデバイス、メモリデバイス等であってもよい。装置又はシステムは、例えば、コンピュータープログラムを受信機に転送するためのファイルサーバを備えることができる。

【0301】

一部の実施形態において、プログラム可能論理デバイス（例えば、フィールドプログラマブルゲートアレイ）が、本明細書に説明される方法の機能性の一部又は全部を行うために使用されてもよい。一部の実施形態において、フィールドプログラマブルゲートアレイは、本明細書に説明される方法のうちの1つを行うために、マイクロプロセッサと協働してもよい。概して、方法は、任意のハードウェア装置によって実行されることが好ましい。

40

【0302】

本明細書で説明される装置は、ハードウェア装置を使用して、又はコンピューターを使用して、又はハードウェア装置とコンピューターとの組み合わせを使用して実施され得る。

【0303】

本明細書で説明される装置、又は本明細書で説明される装置の任意の構成要素は、少なくとも部分的にハードウェア及び／又はソフトウェアで実施され得る。

【0304】

本明細書で説明される方法は、ハードウェア装置を使用して、又はコンピューターを使

50

用して、又はハードウェア装置とコンピューターとの組み合わせを使用して実行され得る。

【0305】

本明細書で説明される方法、又は本明細書で説明される装置の任意の構成要素は、ハードウェア及び／又はソフトウェアによって少なくとも部分的に実行され得る。

【0306】

特許請求の範囲によって定義されるような任意の実施形態は、本明細書に説明される詳細（特徴及び機能）のうちのいずれかによって補完することができることに留意されたい。

【0307】

また、本明細書に記載される実施形態は、個々に使用することができ、特許請求の範囲に含まれる特徴のいずれかによって補完することもできる。

【0308】

また、本明細書に記載される個々の態様は、個別に又は組み合わせて使用することができることに留意されたい。したがって、当該態様の別の1つに詳細を追加することなく、当該個々の態様の各々に詳細を追加することができる。

【0309】

本開示は、概して、ビデオエンコーダー（入力ビデオ信号の符号化表現を提供するための装置）及びビデオデコーダー（符号化表現に基づいてビデオ信号の復号化された表現を提供するための装置）において、並びにオーディオエンコーダー及びオーディオデコーダーにおいて使用可能な特徴を明示的又は暗示的に説明するものであることにも留意されたい。したがって、本明細書で説明する特徴のいずれも、ビデオエンコーダーのコンテキストにおいて、及びビデオデコーダーのコンテキストにおいて、及びオーディオエンコーダーのコンテキストにおいて、及びオーディオデコーダーのコンテキストにおいて使用され得る。

【0310】

さらに、方法に関して本明細書で開示される特徴及び機能は、（かかる機能を実行するように構成された）装置においても使用され得る。さらに、装置に関して本明細書で開示される任意の特徴及び機能は、対応する方法においても使用され得る。換言すれば、本明細書に開示される方法は、装置に関して説明される特徴及び機能のうちのいずれかによって補完することができる。

【0311】

また、本明細書で説明される特徴及び機能のいずれも、「代替的な実施態様」の節で説明したように、ハードウェア若しくはソフトウェアで、又はハードウェアとソフトウェアの組み合わせを使用して実装され得る。

【0312】

さらに、本明細書で説明される特徴及びシンタックス要素のいずれも、任意選択で、個々に及び組み合わせての両方で、ビデオビットストリームに導入され得る。

【0313】

さらに、エンコーダー又は符号化のコンテキストで説明した全ての特徴、機能、及び詳細は、任意選択で、デコーダー又は復号化のコンテキストでも使用できることに留意されたい。例えば、デコーダーにおけるコンテキスト導出は、エンコーダーにおけるコンテキスト導出に類似していてもよく、復号化された値は、符号化される値の役割を果たしてもよい。典型的には、デコーダーは、デコーダーで使用されるコンテキストがエンコーダーで使用されるコンテキストに対応して、エンコーダーとデコーダーとの同期を保つように設計される。

【0314】

本明細書に記載される実施形態は、本発明の原理を単に例示するものである。本明細書に記載された構成及び詳細の変更及び変形が他の当業者に明らかであることが理解される。したがって、添付の特許請求の範囲によってのみ限定され、本明細書の実施形態の記述及び説明によって提示される特定の詳細によって限定されないことが意図される。

10

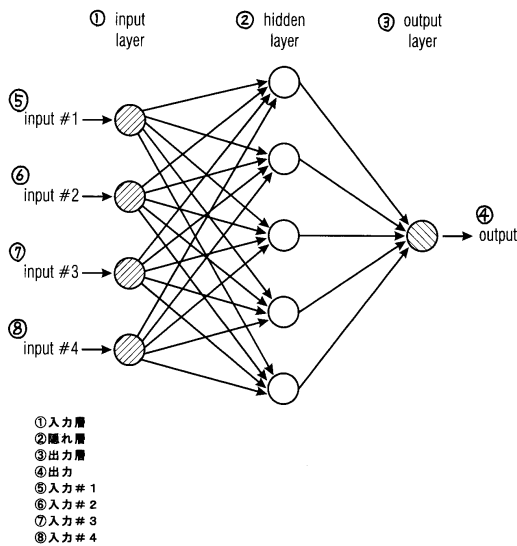
20

30

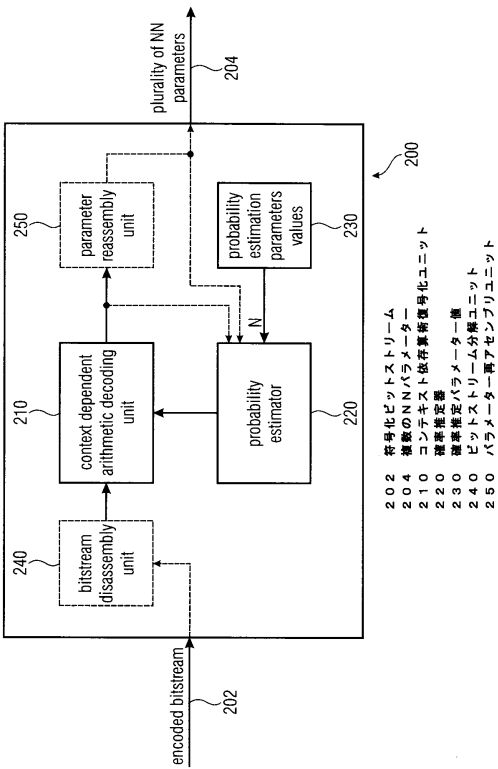
40

50

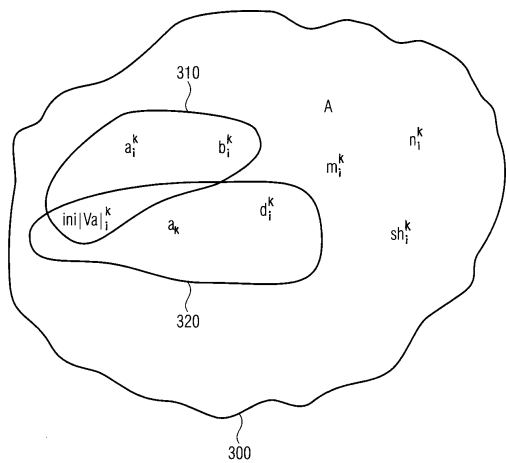
【図面】
【図 1】



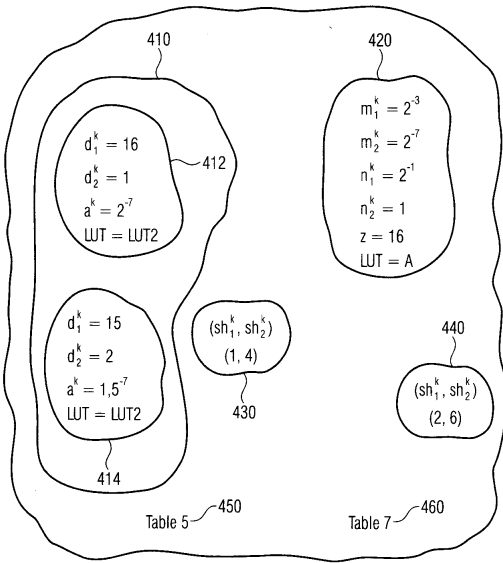
【図 2】



【図 3】



【図 4】



450 表5
460 表7

10

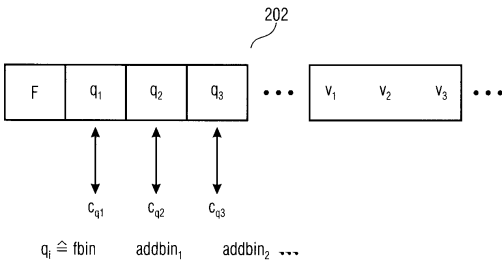
20

30

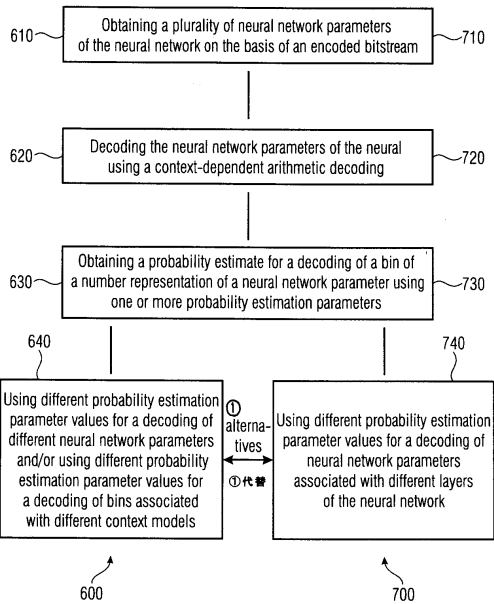
40

50

【図 5】



【図 6】



610 符号化ビットストリームに基づいてニューラルネットワークの複数のニューラルネットワークパラメータを取得する
620 コンテキスト依存算術復号化を使用してニューラルネットワークパラメータを復号化する
630 1つ以上の確率推定パラメータを使用して、ニューラルネットワークパラメータの数表現のビンの復号化のための確率推定値を取得する
640 異なるニューラルネットワークパラメータの復号化のために異なる確率推定パラメータ値を使用し、及び/又は異なるコンテキストモデルに関連付けられたビンの復号化のために異なる確率推定パラメータ値を使用する
740 ニューラルネットワークの異なる層に関連付けられたニューラルネットワークパラメータの復号化のために異なる確率推定パラメータ値を使用する

10

20

30

40

50

フロントページの続き

- スティテュート, ツェー／オー フラウンホーファー－インスティテュート フュア ナッハリヒ
テンテヒニーク
- (72)発明者 ミュラー, カーステン
ドイツ連邦共和国 ベルリン 10587 アインシュタインウーファー37 ハーハーイー, ハイ
ンリッヒ－ヘルツ－インスティテュート, ツェー／オー フラウンホーファー－インスティテュ
ート フュア ナッハリヒテンテヒニーク
- (72)発明者 キルヒホフファー, ハイナー
ドイツ連邦共和国 ベルリン 10587 アインシュタインウーファー37 ハーハーイー, ハイ
ンリッヒ－ヘルツ－インスティテュート, ツェー／オー フラウンホーファー－インスティテュ
ート フュア ナッハリヒテンテヒニーク
- (72)発明者 シュヴァルツ, ハイコ
ドイツ連邦共和国 ベルリン 10587 アインシュタインウーファー37 ハーハーイー, ハイ
ンリッヒ－ヘルツ－インスティテュート, ツェー／オー フラウンホーファー－インスティテュ
ート フュア ナッハリヒテンテヒニーク
- (72)発明者 マーブ, デトレフ
ドイツ連邦共和国 ベルリン 10587 アインシュタインウーファー37 ハーハーイー, ハイ
ンリッヒ－ヘルツ－インスティテュート, ツェー／オー フラウンホーファー－インスティテュ
ート フュア ナッハリヒテンテヒニーク
- (72)発明者 ヴィーガント, トーマス
ドイツ連邦共和国 ベルリン 10587 アインシュタインウーファー37 ハーハーイー, ハイ
ンリッヒ－ヘルツ－インスティテュート, ツェー／オー フラウンホーファー－インスティテュ
ート フュア ナッハリヒテンテヒニーク
- 審査官 田部井 和彦
- (56)参考文献 国際公開第2020/008075 (WO, A1)
特表2021-532698 (JP, A)
国際公開第2016/199330 (WO, A1)
Simon Wiedemann et al., DeepCABAC: A Universal Compression Algorithm for Deep Neural
Networks [online], Published in: IEEE Journal of Selected Topics in Signal Processing (Vol
ume: 14, Issue: 4, May 2020), [2024年1月5日検索], 2020年01月27日, pp.700-714, イ
ンターネット <URL: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8970294>>
4>
Ming Li et al., AHG17 & AHG9: Comments on carriage of coding tool parameters in Adapta
tion Parameter Set [online], JVET-N0065 (JVET-N0065.docx), [2024年1月12日検索],
インターネット <URL: https://jvet-experts.org/doc_end_user/documents/14_Geneva/wg11/JVET-N0065-v1.zip>, 2019年03月11日
- (58)調査した分野 (Int.Cl., DB名)
H04N 19/13
H04N 19/196
H04N 19/91