【公報種別】特許法第１７条の２の規定による補正の掲載
【部門区分】第６部門第３区分
【発行日】令和3年4月1日(2021.4.1)

【公表番号】特表2021-505993(P2021-505993A)
【公表日】令和3年2月18日(2021.2.18)
【年通号数】公開・登録公報2021-008
【出願番号】特願2020-529245(P2020-529245)
【国際特許分類】
Ｇ０６Ｎ　3/08　(2006.01)
【ＦＩ】
Ｇ０６Ｎ　3/08

【手続補正書】
【提出日】令和3年1月28日(2021.1.28)
【手続補正１】
【補正対象書類名】特許請求の範囲
【補正対象項目名】全文
【補正方法】変更
【補正の内容】
【特許請求の範囲】
【請求項１】
深層学習ニューラル・ネットワーク（ＤＮＮ）の訓練に適応性のある残差勾配圧縮のためのコンピュータ実装方法であって、
複数の学習器のうちの第１の学習器のプロセッサによって、前記ＤＮＮのニューラル・ネットワーク層のための現在の勾配ベクトルを取得することであって、前記現在の勾配ベクトルが、訓練データのミニ・バッチから計算される前記ニューラル・ネットワーク層のパラメータの勾配重みを含む、前記現在の勾配ベクトルを取得することと、
前記プロセッサによって、前記ミニ・バッチのための残差勾配重みを含む現在の残差ベクトルを生成することであって、以前の残差ベクトルと前記現在の勾配ベクトルを合計することを含む、前記現在の残差ベクトルを生成することと、
前記プロセッサによって、前記現在の残差ベクトルの前記残差勾配重みを一様な大きさの複数のビンに分割すること、および前記複数のビンのうちの１つまたは複数のビンの前記残差勾配重みのサブセットを量子化することに少なくとも部分的に基づいて、圧縮された現在の残差ベクトルを生成することであって、前記残差勾配重みの前記サブセットを量子化することが、前記ミニ・バッチのためのスケーリング・パラメータを計算すること、および各ビンの極大値を計算することに少なくとも部分的に基づき、前記ビンの前記一様な大きさが、前記ＤＮＮのハイパー・パラメータである、前記圧縮された現在の残差ベクトルを生成することと、
前記プロセッサによって、前記圧縮された現在の残差ベクトルを前記複数の学習器のうちの第２の学習器に伝送することと
を含む、コンピュータ実装方法。
【請求項２】
前記圧縮された現在の残差ベクトルを生成することが、
前記プロセッサによって、前記ミニ・バッチのためのスケーリングされた残差勾配重みを含むスケーリングされた現在の残差ベクトルを生成することであって、前記現在の勾配ベクトルに前記スケーリング・パラメータを乗じること、および前記乗じた勾配ベクトルと前記以前の残差ベクトルを合計することを含む、前記スケーリングされた現在の残差ベクトルを生成することと、
前記現在の残差ベクトルの前記残差勾配重みを前記一様な大きさの前記複数のビンに分

割することと、
前記複数のビンの各ビンについて、前記ビンの前記残差勾配重みの絶対値の極大値を識別することと、
各ビンの各残差勾配重みについて、前記スケーリングされた残差ベクトルの対応するスケーリングされた残差勾配重みが前記ビンの前記極大値を超過することを判定することと、
各ビンの各残差勾配重みについて、前記スケーリングされた残差ベクトルの前記対応するスケーリングされた残差勾配重みが前記ビンの前記極大値を超過することを識別すると、所与の残差勾配重みに対する量子化値を生成し、前記現在の残差ベクトルの前記残差勾配重みを前記量子化値で置換することによって前記現在の残差ベクトルを更新することとを含む、請求項１に記載のコンピュータ実装方法。
【請求項３】
前記スケーリング・パラメータが、Ｌ２正規化に従って量子化誤差を最小化することによって計算される、請求項２に記載のコンピュータ実装方法。
【請求項４】
前記ＤＮＮが、１つまたは複数の畳み込みネットワーク層を含み、
前記複数のビンの前記大きさが、前記１つまたは複数の畳み込み層に対して５０にセットされる、
請求項２に記載のコンピュータ実装方法。
【請求項５】
前記ＤＮＮが、少なくとも１つまたは複数の完全に接続された層を含み、
前記ビンの前記大きさが、前記１つまたは複数の完全に接続された層に対して５００にセットされる、
請求項２に記載のコンピュータ実装方法。
【請求項６】
深層学習ニューラル・ネットワーク（ＤＮＮ）の訓練に適応性のある残差勾配圧縮のためのシステムであって、複数の学習器を備え、前記複数の学習器のうちの少なくとも１つの学習器が、
前記ＤＮＮのニューラル・ネットワーク層のための現在の勾配ベクトルを取得することであって、前記現在の勾配ベクトルが、訓練データのミニ・バッチから計算される前記ニューラル・ネットワーク層のパラメータの勾配重みを含む、前記現在の勾配ベクトルを取得することと、
前記ミニ・バッチのための残差勾配重みを含む現在の残差ベクトルを生成することであって、以前の残差ベクトルと前記現在の勾配ベクトルを合計することを含む、前記現在の残差ベクトルを生成することと、
前記現在の残差ベクトルの前記残差勾配重みを一様な大きさの複数のビンに分割すること、および前記複数のビンのうちの１つまたは複数のビンの前記残差勾配重みのサブセットを量子化することに少なくとも部分的に基づいて、圧縮された現在の残差ベクトルを生成することであって、前記残差勾配重みの前記サブセットを量子化することが、前記ミニ・バッチのためのスケーリング・パラメータを計算すること、および各ビンの極大値を計算することに少なくとも部分的に基づき、前記ビンの前記一様な大きさが、前記ＤＮＮのハイパー・パラメータである、前記圧縮された現在の残差ベクトルを生成することと、
前記圧縮された現在の残差ベクトルを前記複数の学習器のうちの第２の学習器に伝送することと
を含む方法を行うように構成される、システム。
【請求項７】
深層学習ニューラル・ネットワーク（ＤＮＮ）の訓練に適応性のある残差勾配圧縮のためのコンピュータ・プログラムであって、複数の学習器のうちの少なくとも第１の学習器のプロセッサに、
前記ＤＮＮのニューラル・ネットワーク層のための現在の勾配ベクトルを取得すること

であって、前記現在の勾配ベクトルが、訓練データのミニ・バッチから計算される前記ニューラル・ネットワーク層のパラメータの勾配重みを含む、前記現在の勾配ベクトルを取得することと、
前記ミニ・バッチのための残差勾配重みを含む現在の残差ベクトルを生成することであって、以前の残差ベクトルと前記現在の勾配ベクトルを合計することを含む、前記現在の残差ベクトルを生成することと、
前記現在の残差ベクトルの前記残差勾配重みを一様な大きさの複数のビンに分割すること、および前記複数のビンのうちの１つまたは複数のビンの前記残差勾配重みのサブセットを量子化することに少なくとも部分的に基づいて、圧縮された現在の残差ベクトルを生成することであって、前記残差勾配重みの前記サブセットを量子化することが、前記ミニ・バッチのためのスケーリング・パラメータを計算すること、および各ビンの極大値を計算することに少なくとも部分的に基づき、前記ビンの前記一様な大きさが、前記ＤＮＮのハイパー・パラメータである、前記圧縮された現在の残差ベクトルを生成することと、
前記圧縮された現在の残差ベクトルを前記複数の学習器のうちの第２の学習器に伝送することと
を実行させるためのコンピュータ・プログラム。
【請求項８】
適応性のある残差勾配圧縮を介して深層学習ニューラル・ネットワーク（ＤＮＮ）を訓練するためのコンピュータ実装方法であって、
複数の学習器を備えるシステムによって、１つまたは複数のニューラル・ネットワーク層を使用して前記ＤＮＮの訓練のための訓練データを受信することと、
前記複数の学習器のうちの各学習器において、前記訓練データのミニ・バッチからニューラル・ネットワーク層のための現在の勾配ベクトルを生成することであって、前記現在の勾配ベクトルが、前記ニューラル・ネットワーク層のパラメータの勾配重みを含む、前記現在の勾配ベクトルを生成することと、
前記複数の学習器のうちの各学習器において、前記ミニ・バッチのための残差勾配重みを含む現在の残差ベクトルを生成することであって、以前の残差ベクトルと前記現在の勾配ベクトルを合計することを含む、前記現在の残差ベクトルを生成することと、
前記現在の残差ベクトルの前記残差勾配重みを一様な大きさの複数のビンに分割すること、および前記複数のビンのうちの１つまたは複数のビンの前記残差勾配重みのサブセットを量子化することに少なくとも部分的に基づいて、前記複数の学習器のうちの各学習器において、圧縮された現在の残差ベクトルを生成することであって、前記残差勾配重みの前記サブセットを量子化することが、前記ミニ・バッチのためのスケーリング・パラメータを計算すること、および各ビンの極大値を計算することに少なくとも部分的に基づき、前記ビンの前記一様な大きさが、前記ＤＮＮのハイパー・パラメータである、前記圧縮された現在の残差ベクトルを生成することと
前記複数の学習器の間で、前記圧縮された現在の残差ベクトルを交換することと、
前記複数の学習器のそれぞれにおいて、前記圧縮された現在の残差ベクトルを解凍することと、
前記複数の学習器のそれぞれにおいて、前記ニューラル・ネットワーク層の前記パラメータの前記勾配重みを更新することと
を含む、コンピュータ実装方法。
【請求項９】
適応性のある残差勾配圧縮を介して深層学習ニューラル・ネットワーク（ＤＮＮ）を訓練するためのシステムであって、複数の学習器を備え、
１つまたは複数のニューラル・ネットワーク層を使用して前記ＤＮＮの訓練のための訓練データを受信することと、
前記複数の学習器のうちの各学習器において、前記訓練データのミニ・バッチからニューラル・ネットワーク層のための現在の勾配ベクトルを生成することであって、前記現在の勾配ベクトルが、前記ニューラル・ネットワーク層のパラメータの勾配重みを含む、前

記現在の勾配ベクトルを生成することと、
前記複数の学習器のうちの各学習器において、前記ミニ・バッチのための残差勾配重みを含む現在の残差ベクトルを生成することであって、以前の残差ベクトルと前記現在の勾配ベクトルを合計することを含む、前記現在の残差ベクトルを生成することと、
前記現在の残差ベクトルの前記残差勾配重みを一様な大きさの複数のビンに分割すること、および前記複数のビンのうちの１つまたは複数のビンの前記残差勾配重みのサブセットを量子化することに少なくとも部分的に基づいて、前記複数の学習器のうちの各学習器において、圧縮された現在の残差ベクトルを生成することであって、前記残差勾配重みの前記サブセットを量子化することが、前記ミニ・バッチのためのスケーリング・パラメータを計算すること、および各ビンの極大値を計算することに少なくとも部分的に基づき、前記ビンの前記一様な大きさが、前記ＤＮＮのハイパー・パラメータである、前記圧縮された現在の残差ベクトルを生成することと、
前記複数の学習器の間で、前記圧縮された現在の残差ベクトルを交換することと、
前記複数の学習器のそれぞれにおいて、前記圧縮された現在の残差ベクトルを解凍することと、
前記複数の学習器のそれぞれにおいて、前記ニューラル・ネットワーク層の前記パラメータの前記勾配重みを更新することと
を含む方法を行うように構成される、システム。