【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4年8月8日(2022.8.8)

【公開番号】特開2022-70955(P2022-70955A)
【公開日】令和4年5月13日(2022.5.13)
【年通号数】公開公報(特許)2022-084
【出願番号】特願2022-19764(P2022-19764)
【国際特許分類】
Ｇ０６Ｎ　3/063(2006.01)
【ＦＩ】
Ｇ０６Ｎ　3/063

【手続補正書】
【提出日】令和4年7月28日(2022.7.28)
【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】
ハードウェア集積回路上で実現されるニューラルネットワークを使用してニューラルネットワーク入力のバッチを処理するための方法であって、前記ニューラルネットワークは、有向グラフの状態で配置された複数のニューラルネットワークレイヤを備え、前記方法は、
前記ニューラルネットワーク入力のバッチを受信するステップと、
複数のスーパーレイヤを備えるシーケンスにパーティショニングされる前記ニューラルネットワークのレイヤを識別するステップとを備え、各スーパーレイヤは、２つ以上のニューラルネットワークレイヤを備え、前記有向グラフのパーティションであり、前記方法はさらに、
前記ハードウェア集積回路を使用して前記ニューラルネットワーク入力のバッチを処理するステップを備え、前記ハードウェア集積回路を使用して前記ニューラルネットワーク入力のバッチを処理するステップは、前記複数のスーパーレイヤの各スーパーレイヤについて、
前記バッチ内の前記ニューラルネットワーク入力の各々に対応するそれぞれのスーパーレイヤ入力を取得するステップと、
前記スーパーレイヤにおける前記ニューラルネットワークレイヤのいずれかを使用してスーパーレイヤ入力を処理する前に、前記スーパーレイヤにおける前記レイヤの各々のためのそれぞれのパラメータセットを前記ハードウェア集積回路のパラメータメモリにロードするステップと、
前記パラメータメモリから取得された前記ニューラルネットワークレイヤのための前記それぞれのパラメータセットを使用して、前記スーパーレイヤにおけるニューラルネットワークレイヤを介して前記スーパーレイヤ入力を処理するステップとを備える、方法。
【請求項２】
前記スーパーレイヤにおけるニューラルネットワークレイヤを介した前記スーパーレイヤ入力の前記処理に基づいてスーパーレイヤ出力を生成するステップをさらに備える、請求項１に記載の方法。
【請求項３】
前記スーパーレイヤ出力は、前記シーケンス内の第１のスーパーレイヤの出力であり、前

記方法はさらに、
前記シーケンス内の第２のスーパーレイヤにおけるニューラルネットワークレイヤへのスーパーレイヤ入力として前記スーパーレイヤ出力を受信するステップと、
前記第１のスーパーレイヤの前記スーパーレイヤ出力に対応する前記スーパーレイヤ入力について、前記シーケンス内の前記第２のスーパーレイヤにおけるニューラルネットワークレイヤを介して前記スーパーレイヤ入力を処理するステップとを備える、請求項２に記載の方法。
【請求項４】
前記スーパーレイヤにおける前記ニューラルネットワークレイヤの各々について前記それぞれのパラメータセットをロードするステップは、
前記ハードウェア集積回路および前記パラメータメモリの、外部にあるホストから受信されたデータ値に基づいて、前記それぞれのパラメータセットを前記パラメータメモリに予めロードするステップを備える、請求項１～３のいずれか１項に記載の方法。
【請求項５】
前記ハードウェア集積回路を使用して前記ニューラルネットワーク入力のバッチを処理するステップは、
スケジューリングプロセスに基づいて前記ニューラルネットワーク入力のバッチを処理するステップを備え、前記スケジューリングプロセスは、前記ハードウェア集積回路において実現されるニューラルネットワークモデルのバッチ次元およびレイヤ次元に対してニューラルネットワーク計算のグローバルスケジューリングを実行する、請求項４に記載の方法。
【請求項６】
前記ニューラルネットワーク計算の前記グローバルスケジューリングは、前記ホストを使用して実行される、請求項５に記載の方法。
【請求項７】
前記シーケンス内のスーパーレイヤの各ニューラルネットワークレイヤは、それぞれのワーキングセットに関連付けられ、
前記それぞれのワーキングセットは、部分的に、前記ワーキングセットにおけるスーパーレイヤ入力を処理するために使用される前記ニューラルネットワークレイヤのためのパラメータを格納するのに必要なメモリの量によって定義される、請求項１～６のいずれか１項に記載の方法。
【請求項８】
前記シーケンス内の第１のスーパーレイヤは、前記有向グラフの第１のパーティションを表し、
前記シーケンス内の第２のスーパーレイヤは、前記有向グラフの第２の異なるパーティションを表す、請求項１～７のいずれか１項に記載の方法。
【請求項９】
ハードウェア集積回路上で実現されるニューラルネットワークを使用してニューラルネットワーク入力のバッチを処理するためのシステムであって、前記ニューラルネットワークは、有向グラフの状態で配置された複数のニューラルネットワークレイヤを備え、前記システムは、
前記ハードウェア集積回路と、プロセッサと、命令を格納するための非一時的なコンピュータ読取可能記憶装置とを備え、前記命令は、動作の実行をさせるように前記プロセッサによって実行可能であり、前記動作は、
前記ニューラルネットワーク入力のバッチを受信するステップと、
複数のスーパーレイヤを備えるシーケンスにパーティショニングされる前記ニューラルネットワークのレイヤを識別するステップとを備え、各スーパーレイヤは、２つ以上のニューラルネットワークレイヤを備え、前記有向グラフのパーティションであり、前記動作はさらに、
前記ハードウェア集積回路を使用して前記ニューラルネットワーク入力のバッチを処理す

るステップを備え、前記ハードウェア集積回路を使用して前記ニューラルネットワーク入力のバッチを処理するステップは、前記複数のスーパーレイヤの各スーパーレイヤについて、
前記バッチ内の前記ニューラルネットワーク入力の各々に対応するそれぞれのスーパーレイヤ入力を取得するステップと、
前記スーパーレイヤにおける前記ニューラルネットワークレイヤのいずれかを使用してスーパーレイヤ入力を処理する前に、前記スーパーレイヤにおける前記レイヤの各々のためのそれぞれのパラメータセットを前記ハードウェア集積回路のパラメータメモリにロードするステップと、
前記パラメータメモリから取得された前記ニューラルネットワークレイヤのための前記それぞれのパラメータセットを使用して、前記スーパーレイヤにおけるニューラルネットワークレイヤを介して前記スーパーレイヤ入力を処理するステップとを備える、システム。
【請求項１０】
前記動作は、前記スーパーレイヤにおけるニューラルネットワークレイヤを介した前記スーパーレイヤ入力の前記処理に基づいてスーパーレイヤ出力を生成するステップを備える、請求項９に記載のシステム。
【請求項１１】
前記スーパーレイヤ出力は、前記シーケンス内の第１のスーパーレイヤの出力であり、前記動作はさらに、
前記シーケンス内の第２のスーパーレイヤにおけるニューラルネットワークレイヤへのスーパーレイヤ入力として前記スーパーレイヤ出力を受信するステップと、
前記第１のスーパーレイヤの前記スーパーレイヤ出力に対応する前記スーパーレイヤ入力について、前記シーケンス内の前記第２のスーパーレイヤにおけるニューラルネットワークレイヤを介して前記スーパーレイヤ入力を処理するステップとを備える、請求項１０に記載のシステム。
【請求項１２】
前記スーパーレイヤにおける前記ニューラルネットワークレイヤの各々について前記それぞれのパラメータセットをロードするステップは、
前記ハードウェア集積回路および前記パラメータメモリの、外部にあるホストから受信されたデータ値に基づいて、前記それぞれのパラメータセットを前記パラメータメモリに予めロードするステップを備える、請求項９～１１のいずれか１項に記載のシステム。
【請求項１３】
前記ハードウェア集積回路を使用して前記ニューラルネットワーク入力のバッチを処理するステップは、
スケジューリングプロセスに基づいて前記ニューラルネットワーク入力のバッチを処理するステップを備え、前記スケジューリングプロセスは、前記ハードウェア集積回路において実現されるニューラルネットワークモデルのバッチ次元およびレイヤ次元に対してニューラルネットワーク計算のグローバルスケジューリングを実行する、請求項１２に記載のシステム。
【請求項１４】
前記ニューラルネットワーク計算の前記グローバルスケジューリングは、前記ホストを使用して実現される、請求項１３に記載のシステム。
【請求項１５】
前記シーケンス内のスーパーレイヤの各ニューラルネットワークレイヤは、それぞれのワーキングセットに関連付けられ、
前記それぞれのワーキングセットは、部分的に、前記ワーキングセットにおけるスーパーレイヤ入力を処理するために使用される前記ニューラルネットワークレイヤのためのパラメータを格納するのに必要なメモリの量によって定義される、請求項９～１４のいずれか１項に記載のシステム。
【請求項１６】

前記シーケンス内の第１のスーパーレイヤは、前記有向グラフの第１のパーティションを表し、
前記シーケンス内の第２のスーパーレイヤは、前記有向グラフの第２の異なるパーティションを表す、請求項９～１５のいずれか１項に記載のシステム。
【請求項１７】
ハードウェア集積回路上で実現されるニューラルネットワークを使用してニューラルネットワーク入力のバッチを処理するための命令を格納するように構成された非一時的なコンピュータ読取可能記憶装置であって、
前記ニューラルネットワークは、有向グラフの状態で配置された複数のニューラルネットワークレイヤを備え、前記命令は、動作の実行をさせるようにプロセッサによって実行可能であり、前記動作は、
前記ニューラルネットワーク入力のバッチを受信するステップと、
複数のスーパーレイヤを備えるシーケンスにパーティショニングされる前記ニューラルネットワークのレイヤを識別するステップとを備え、各スーパーレイヤは、２つ以上のニューラルネットワークレイヤを備え、前記有向グラフのパーティションであり、前記動作はさらに、
前記ハードウェア集積回路を使用して前記ニューラルネットワーク入力のバッチを処理するステップを備え、前記ハードウェア集積回路を使用して前記ニューラルネットワーク入力のバッチを処理するステップは、前記複数のスーパーレイヤの各スーパーレイヤについて、
前記バッチ内の前記ニューラルネットワーク入力の各々に対応するそれぞれのスーパーレイヤ入力を取得するステップと、
前記スーパーレイヤにおける前記ニューラルネットワークレイヤのいずれかを使用してスーパーレイヤ入力を処理する前に、前記スーパーレイヤにおける前記レイヤの各々のためのそれぞれのパラメータセットを前記ハードウェア集積回路のパラメータメモリにロードするステップと、
前記パラメータメモリから取得された前記ニューラルネットワークレイヤのための前記それぞれのパラメータセットを使用して、前記スーパーレイヤにおけるニューラルネットワークレイヤを介して前記スーパーレイヤ入力を処理するステップとを備える、非一時的なコンピュータ読取可能記憶装置。
【請求項１８】
前記動作は、前記スーパーレイヤにおけるニューラルネットワークレイヤを介した前記スーパーレイヤ入力の前記処理に基づいてスーパーレイヤ出力を生成するステップをさらに備える、請求項１７に記載の非一時的なコンピュータ読取可能記憶装置。
【請求項１９】
前記スーパーレイヤ出力は、前記シーケンス内の第１のスーパーレイヤの出力であり、前記動作はさらに、
前記シーケンス内の第２のスーパーレイヤにおけるニューラルネットワークレイヤへのスーパーレイヤ入力として前記スーパーレイヤ出力を受信するステップと、
前記第１のスーパーレイヤの前記スーパーレイヤ出力に対応する前記スーパーレイヤ入力について、前記シーケンス内の前記第２のスーパーレイヤにおけるニューラルネットワークレイヤを介して前記スーパーレイヤ入力を処理するステップとを備える、請求項１８に記載の非一時的なコンピュータ読取可能記憶装置。
【請求項２０】
前記スーパーレイヤにおける前記ニューラルネットワークレイヤの各々について前記それぞれのパラメータセットをロードするステップは、
前記ハードウェア集積回路および前記パラメータメモリの外部にあるホストから受信されたデータ値に基づいて、前記それぞれのパラメータセットを前記パラメータメモリに予めロードするステップを備える、請求項１７～１９のいずれか１項に記載の非一時的なコンピュータ読取可能記憶装置。