【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2年8月27日(2020.8.27)

【公表番号】特表2020-521195(P2020-521195A)
【公表日】令和2年7月16日(2020.7.16)
【年通号数】公開・登録公報2020-028
【出願番号】特願2019-552217(P2019-552217)
【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2年6月23日(2020.6.23)
【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】
方法であって、
ハードウェア回路上のニューラルネットワークを使用して処理される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を備え、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レイヤを有し、各レイヤは、それぞれのパラメータセットを有し、前記方法はさらに、
前記ニューラルネットワークのレイヤの、スーパーレイヤのシーケンスへのパーティショニングを決定するステップを備え、各スーパーレイヤは、１つ以上のレイヤを含む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、前記シーケンスにおける各スーパーレイヤについて、
前記スーパーレイヤにおける前記レイヤのための前記それぞれのパラメータセットを前記ハードウェア回路のメモリにロードするステップと、
前記バッチ内の各ニューラルネットワーク入力について、前記ハードウェア回路の前記メモリ内の前記パラメータセットを使用して、前記スーパーレイヤの各レイヤを介して前記ニューラルネットワーク入力に対応するスーパーレイヤ入力を処理して、前記ニューラルネットワーク入力のためのスーパーレイヤ出力を生成するステップとを備える、方法。
【請求項２】
前記シーケンスにおける第１のスーパーレイヤでは、前記ニューラルネットワーク入力に対応する前記スーパーレイヤ入力が前記ニューラルネットワーク入力である、請求項１に記載の方法。
【請求項３】
前記第１のスーパーレイヤ出力の後の各スーパーレイヤへの前記スーパーレイヤ入力は、前記シーケンスにおける先行するスーパーレイヤによって生成されたスーパーレイヤ出力である、請求項２に記載の方法。
【請求項４】
前記ハードウェア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、各スーパーレイヤについて、
前記バッチ内の第２のニューラルネットワーク入力に対応するスーパーレイヤ入力が前記スーパーレイヤの各レイヤを介して後に処理される前に前記バッチ内の第１のニューラルネットワーク入力のための前記スーパーレイヤ入力が前記スーパーレイヤの各レイヤを介して処理されるように、前記スーパーレイヤの各レイヤを介して前記ニューラルネットワーク入力のバッチに対応する前記スーパーレイヤ入力をシーケンシャルに処理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】
スーパーレイヤのそれぞれのレイヤは、ワーキングセットに関連付けられ、各ワーキングセットは、少なくとも
ｉ）前記ハードウェア回路上の前記ニューラルネットワークを使用して処理される前記ニューラルネットワーク入力のバッチの１つ以上の入力または前記スーパーレイヤの先行するレイヤの１つ以上の出力、および
ｉｉ）前記スーパーレイヤの各レイヤを介して前記１つ以上の入力を処理するのに必要なメモリの量を示すサイズパラメータ
によって定義される、請求項１～４のいずれか１項に記載の方法。
【請求項６】
前記ニューラルネットワークのレイヤの、スーパーレイヤのシーケンスへの前記パーティショニングを決定するステップは、
ｉ）少なくとも１つのワーキングセットのための特定のサイズパラメータを決定するステップと、
ｉｉ）前記ハードウェア回路の前記メモリの特定の集約パラメータ容量を決定するステップと、
ｉｉｉ）前記少なくとも１つのワーキングセットのための前記特定のサイズパラメータまたは前記ハードウェア回路の前記メモリの特定の集約パラメータ容量のうちの少なくとも１つに基づいて、前記ニューラルネットワークのレイヤの、スーパーレイヤのシーケンスへの前記パーティショニングを決定するステップとを備える、請求項５に記載の方法。
【請求項７】
前記ハードウェア回路の前記メモリは、閾値記憶容量を有し、前記ニューラルネットワークのレイヤの、スーパーレイヤのシーケンスへの前記パーティショニングを決定するステップは、
前記ハードウェア回路の前記メモリの前記閾値記憶容量に基づいて、前記ニューラルネットワークのレイヤをスーパーレイヤのシーケンスにパーティショニングするステップを備える、請求項１～６のいずれか１項に記載の方法。
【請求項８】
前記ニューラルネットワークのレイヤは、前記ハードウェア回路が前記ニューラルネットワーク入力のバッチを処理する際に前記メモリの前記閾値記憶容量を超えないようにスーパーレイヤのシーケンスにパーティショニングされる、請求項７に記載の方法。
【請求項９】
前記ニューラルネットワーク入力のバッチおよび前記それぞれのパラメータセットは、前記ハードウェア回路の外部のソースから受信され、前記スーパーレイヤの各レイヤを介して前記ニューラルネットワーク入力に対応する前記スーパーレイヤ入力を処理するステップは、前記外部のソースから追加のパラメータを受信することなく前記スーパーレイヤ入力を処理するステップを備える、請求項１～８のいずれか１項に記載の方法。
【請求項１０】
コンピューティングシステムであって、
前記コンピューティングシステムに配設されたハードウェア回路を備え、前記ハードウェア回路は、１つ以上の処理装置を含み、前記コンピューティングシステムはさらに、
動作を実行するように前記１つ以上の処理装置によって実行可能な命令を格納するための１つ以上の機械読取可能記憶装置を備え、前記動作は、

ハードウェア回路上のニューラルネットワークを使用して処理される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を備え、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レイヤを有し、各レイヤは、それぞれのパラメータセットを有し、前記動作はさらに、

前記ニューラルネットワークのレイヤの、スーパーレイヤのシーケンスへのパーティショニングを決定するステップを備え、各スーパーレイヤは、１つ以上のレイヤを含む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、前記シーケンスにおける各スーパーレイヤについて、

前記スーパーレイヤにおける前記レイヤのための前記それぞれのパラメータセットを前記ハードウェア回路のメモリにロードするステップと、

前記バッチ内の各ニューラルネットワーク入力について、前記ハードウェア回路の前記メモリ内の前記パラメータを使用して、前記スーパーレイヤの各レイヤを介して前記ニューラルネットワーク入力に対応するスーパーレイヤ入力を処理して、前記ニューラルネットワーク入力のためのスーパーレイヤ出力を生成するステップとを備える、コンピューティングシステム。

【請求項１１】

前記シーケンスにおける第１のスーパーレイヤでは、前記ニューラルネットワーク入力に対応する前記スーパーレイヤ入力が前記ニューラルネットワーク入力である、請求項１０に記載のコンピューティングシステム。

【請求項１２】

前記第１のスーパーレイヤ出力の後の各スーパーレイヤへの前記スーパーレイヤ入力は、前記シーケンスにおける先行するスーパーレイヤによって生成されたスーパーレイヤ出力である、請求項１１に記載のコンピューティングシステム。

【請求項１３】

前記ハードウェア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、各スーパーレイヤについて、

前記バッチ内の第２のニューラルネットワーク入力に対応するスーパーレイヤ入力が前記スーパーレイヤの各レイヤを介して後に処理される前に前記バッチ内の第１のニューラルネットワーク入力のための前記スーパーレイヤ入力が前記スーパーレイヤの各レイヤを介して処理されるように、前記スーパーレイヤの各レイヤを介して前記ニューラルネットワーク入力のバッチに対応する前記スーパーレイヤ入力をシーケンシャルに処理するステップを備える、請求項１０～１２のいずれか１項に記載のコンピューティングシステム。

【請求項１４】

スーパーレイヤのそれぞれのレイヤは、ワーキングセットに関連付けられ、各ワーキングセットは、少なくとも

ｉ）前記ハードウェア回路上の前記ニューラルネットワークを使用して処理される前記ニューラルネットワーク入力のバッチの１つ以上の入力または前記スーパーレイヤの先行するレイヤの１つ以上の出力、および

ｉｉ）前記スーパーレイヤの各レイヤを介して前記１つ以上の入力を処理するのに必要なメモリの量を示すサイズパラメータ

によって定義される、請求項１０～１３のいずれか１項に記載のコンピューティングシステム。

【請求項１５】

前記ニューラルネットワークのレイヤの、スーパーレイヤのシーケンスへの前記パーティショニングを決定するステップは、

ｉ）少なくとも１つのワーキングセットのための特定のサイズパラメータを決定するステップと、

ｉｉ）前記ハードウェア回路の前記メモリの特定の集約パラメータ容量を決定するステ

ップと、
ｉｉｉ）前記少なくとも１つのワーキングセットのための前記特定のサイズパラメータまたは前記ハードウェア回路の前記メモリの特定の集約パラメータ容量のうちの少なくとも１つに基づいて、前記ニューラルネットワークのレイヤの、スーパーレイヤのシーケンスへの前記パーティショニングを決定するステップとを備える、請求項１４に記載のコンピューティングシステム。
【請求項１６】
前記ハードウェア回路の前記メモリは、閾値記憶容量を有し、前記ニューラルネットワークのレイヤの、スーパーレイヤのシーケンスへの前記パーティショニングを決定するステップは、
前記ハードウェア回路の前記メモリの前記閾値記憶容量に基づいて、前記ニューラルネットワークのレイヤをスーパーレイヤのシーケンスにパーティショニングするステップを備える、請求項１０～１５のいずれか１項に記載のコンピューティングシステム。
【請求項１７】
前記ニューラルネットワークのレイヤは、前記ハードウェア回路が前記ニューラルネットワーク入力のバッチを処理する際に前記メモリの前記閾値記憶容量を超えないようにスーパーレイヤのシーケンスにパーティショニングされる、請求項１６に記載のコンピューティングシステム。
【請求項１８】
前記ニューラルネットワーク入力のバッチおよび前記それぞれのパラメータセットは、前記ハードウェア回路の外部のソースから受信され、前記スーパーレイヤの各レイヤを介して前記ニューラルネットワーク入力に対応する前記スーパーレイヤ入力を処理するステップは、前記外部のソースから追加のパラメータを受信することなく前記スーパーレイヤ入力を処理するステップを備える、請求項１０～１７のいずれか１項に記載のコンピューティングシステム。
【請求項１９】
動作を実行するように１つ以上の処理装置に命令を実行させるプログラムであって、前記動作は、
ハードウェア回路上のニューラルネットワークを使用して処理される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を備え、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レイヤを有し、各レイヤは、それぞれのパラメータセットを有し、前記動作はさらに、
前記ニューラルネットワークのレイヤの、スーパーレイヤのシーケンスへのパーティショニングを決定するステップを備え、各スーパーレイヤは、１つ以上のレイヤを含む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、前記シーケンスにおける各スーパーレイヤについて、
前記スーパーレイヤにおける前記レイヤのための前記それぞれのパラメータセットを前記ハードウェア回路のメモリにロードするステップと、
前記バッチ内の各ニューラルネットワーク入力について、前記ハードウェア回路の前記メモリ内の前記パラメータセットを使用して、前記スーパーレイヤの各レイヤを介して前記ニューラルネットワーク入力に対応するスーパーレイヤ入力を処理して、前記ニューラルネットワーク入力のためのスーパーレイヤ出力を生成するステップとを備える、プログラム。
【請求項２０】
前記ハードウェア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、各スーパーレイヤについて、
前記バッチ内の第２のニューラルネットワーク入力に対応するスーパーレイヤ入力が前記スーパーレイヤの各レイヤを介して後に処理される前に前記バッチ内の第１のニューラ

ルネットワーク入力のための前記スーパーレイヤ入力が前記スーパーレイヤの各レイヤを介して処理されるように、前記スーパーレイヤの各レイヤを介して前記ニューラルネットワーク入力のバッチに対応する前記スーパーレイヤ入力をシーケンシャルに処理するステップを備える、請求項１９に記載のプログラム。