【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】平成31年1月10日(2019.1.10)

【公表番号】特表2018-506785(P2018-506785A)
【公表日】平成30年3月8日(2018.3.8)
【年通号数】公開・登録公報2018-009
【出願番号】特願2017-538296(P2017-538296)
【国際特許分類】
Ｇ０６Ｎ　3/08　(2006.01)
Ｇ０６Ｆ　17/10　(2006.01)
【ＦＩ】
Ｇ０６Ｎ　3/08　１２０
Ｇ０６Ｆ　17/10　Ａ

【手続補正書】
【提出日】平成30年11月20日(2018.11.20)
【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】
ニューラルネットワークを圧縮するコンピュータ実装の方法であって、
圧縮されたニューラルネットワークを生成するために、前記ニューラルネットワーク中の圧縮されていない層を複数の圧縮された層と交換することと、ここにおいて、前記複数の圧縮された層の各々は、前記圧縮されていない層と同じタイプである、
前記圧縮されたネットワークの圧縮された層の間に非線形性を挿入することと、
前記圧縮された層のうちの少なくとも１つにおける重み値を更新することによって、前記圧縮されたネットワークを微調整することと
を備える、方法。
【請求項２】
非線形性を前記挿入することが、前記圧縮された層のニューロンに非線形活性化関数を適用することを備える、請求項１に記載の方法。
【請求項３】
前記非線形活性化関数が、整流関数、絶対値関数、双曲線正接関数、またはシグモイド関数である、請求項２に記載の方法。
【請求項４】
前記微調整することが、前記圧縮された層のサブセットのうちの少なくとも１つにおける重み値、または圧縮されていない層のサブセットにおける重み値を更新することを備える、請求項１に記載の方法。
【請求項５】
前記微調整することが、トレーニング例を使用して実行され、前記トレーニング例が、圧縮されていないネットワークをトレーニングするために使用される例の第１のセット、または例の新しいセットのうちの少なくとも１つを備える、請求項１に記載の方法。
【請求項６】
より深層のニューラルネットワークを初期化するための方法として、圧縮と、非線形層の挿入と、前記微調整することとを繰り返し適用することによって、前記ニューラルネットワークを初期化すること

をさらに備える、請求項１に記載の方法。
【請求項７】
ニューラルネットワークを圧縮するコンピュータ実装の方法であって、
組み合わせられた、複数の圧縮された層の受容野サイズが、圧縮されていない層の受容野サイズに一致するように、圧縮されたニューラルネットワークを生成するために、前記ニューラルネットワーク中の前記圧縮されていない層を前記圧縮された層と交換することと、ここにおいて、前記複数の圧縮された層の各々は、前記圧縮されていない層と同じタイプである、
前記圧縮された層のうちの少なくとも１つにおける重み値を更新することによって、前記圧縮されたネットワークを微調整することと
を備える、方法。
【請求項８】
前記圧縮されていない層のカーネルサイズが、前記受容野サイズに等しい、請求項７に記載の方法。
【請求項９】
前記交換することは、プロパティ（$k_{1x}-1$）＋（$k_{2x}-1$）＋・・・＝（k_x-1）および（$k_{1y}-1$）＋（$k_{2y}-1$）＋・・・＝（k_y-1）が満たされる前記圧縮されたネットワークを生成するために、カーネルサイズ$k_x \times k_y$を有する前記ニューラルネットワーク中の少なくとも１つの層を、前記カーネルサイズ$k_{1x} \times k_{1y}$，$k_{2x} \times k_{2y}$．．．$k_{Lx} \times k_{Ly}$をもつ同じタイプの複数の圧縮された層と交換することを備える、請求項７に記載の方法。
【請求項１０】
前記カーネルサイズ$k_x \times k_y$をもつ畳み込み層が、それぞれ前記カーネルサイズ１×１、$k_x \times k_y$、および１×１をもつ３つの畳み込み層と交換される、請求項９に記載の方法。
【請求項１１】
ニューラルネットワークを圧縮するコンピュータ実装の方法であって、
圧縮されたニューラルネットワークを生成するために、前記ニューラルネットワーク中の圧縮されていない層を複数の圧縮された層と交換することと、ここにおいて、前記複数の圧縮された層の各々は、前記圧縮されていない層と同じタイプである、
１つずつ交互に、それぞれの層の重み行列の重み値に交互最小化プロセスを適用することによって、圧縮された層の前記重み行列を決定することと
を備える、方法。
【請求項１２】
前記圧縮された層のうちの少なくとも１つのための前記重み行列の前記重み値を更新することによって、前記圧縮されたネットワークを微調整することをさらに備える、請求項１１に記載の方法。
【請求項１３】
前記微調整することが、前記圧縮された層のサブセット、または圧縮されていない層のサブセットのうちの少なくとも１つにおける前記重み値を更新すること含む、請求項１２に記載の方法。
【請求項１４】
ニューラルネットワークを圧縮するための装置であって、
圧縮されたニューラルネットワークを生成するために、前記ニューラルネットワーク中の圧縮されていない層を複数の圧縮された層と交換するための手段と、ここにおいて、前記複数の圧縮された層の各々は、前記圧縮されていない層と同じタイプである、
前記圧縮されたネットワークの圧縮された層の間に非線形性を挿入するための手段と、
前記圧縮された層のうちの少なくとも１つにおける重み値を更新することによって、前記圧縮されたネットワークを微調整するための手段と
を備える、装置。

【請求項１５】
ニューラルネットワークを圧縮するためのプログラムコードをその上に符号化した非一時的コンピュータ可読媒体であって、前記プログラムコードが、プロセッサによって実行され、および、請求項１～請求項１３のうちのいずれか一項に記載の方法を実行するためのプログラムコードを備える、非一時的コンピュータ可読媒体。