(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7488375号
(P7488375)

(45)発行日　令和6年5月21日(2024.5.21)

(24)登録日　令和6年5月13日(2024.5.13)

(51)国際特許分類

		FI	
G06N	*3/0985(2023.01)*	G06N	3/0985
G06N	*3/096(2023.01)*	G06N	3/096
G06N	*3/045(2023.01)*	G06N	3/045
G06N	*3/082(2023.01)*	G06N	3/082

請求項の数　10　（全23頁）

(21)出願番号　特願2022-580186(P2022-580186)
(86)(22)出願日　令和3年6月22日(2021.6.22)
(65)公表番号　特表2023-531538(P2023-531538A)
(43)公表日　令和5年7月24日(2023.7.24)
(86)国際出願番号　PCT/CN2021/101545
(87)国際公開番号　WO2021/259262
(87)国際公開日　令和3年12月30日(2021.12.30)
審査請求日　令和4年12月29日(2022.12.29)
(31)優先権主張番号　202010581487.2
(32)優先日　令和2年6月23日(2020.6.23)
(33)優先権主張国・地域又は機関
中国(CN)

(73)特許権者　511151662
中興通訊股▲ふん▼有限公司
ＺＴＥ　ＣＯＲＰＯＲＡＴＩＯＮ
中華人民共和国広東省深▲せん▼市南山区高新技術産業園科技南路中興通訊大厦
ＺＴＥ　Ｐｌａｚａ，Ｋｅｊｉ　Ｒｏａｄ　Ｓｏｕｔｈ，Ｈｉ－Ｔｅｃｈ　Ｉｎｄｕｓｔｒｉａｌ　Ｐａｒｋ，Ｎａｎｓｈａｎ　Ｓｈｅｎｚｈｅｎ，Ｇｕａｎｇｄｏｎｇ　５１８０５７　Ｃｈｉｎａ
(74)代理人　100112656
弁理士　宮田 英毅
(74)代理人　100089118
弁理士　酒井 宏明
(74)代理人　100114557

最終頁に続く

(54)【発明の名称】　ニューラルネットワークの生成方法、機器及びコンピュータ可読記憶媒体

(57)【特許請求の範囲】

【請求項１】

システムアーキテクチャプラットフォームによって実行されるニューラルネットワークの生成方法であって、

最適マイクロユニットを取得し、前記最適マイクロユニットに基づいて第１ネットワークを構築するステップと、

既定の訓練データセットを用いて前記第１ネットワークを訓練して、第２ネットワークを得るステップと、

マイクロユニットの数が前記第１ネットワークの最適マイクロユニットの数より少ない第３ネットワークを構築し、前記第２ネットワークを用いて前記第３ネットワークの各マイクロユニットを訓練するステップと、

前記第３ネットワーク内の訓練後の前記マイクロユニットに基づいて、ニューラルネットワークモデルを得るステップと、

を含むニューラルネットワークの生成方法。

【請求項２】

前記最適マイクロユニットに基づいて第１ネットワークを構築する前記ステップは、

事前定義のサイズに従って、前記最適マイクロユニットをＮ回スタックして前記第１ネットワークを生成し、ここで、Ｎは２の整数乗であるステップを含む

請求項１に記載のニューラルネットワークの生成方法。

【請求項３】

第３ネットワークを構築する前記ステップは、
マイクロユニットの数はＮ／ＭでＮ／Ｍ＞１を満たす第３ネットワークを構築し、ここで、Ｍは２の整数乗であるステップと、
すべての前記第３ネットワークを初期化するステップと、
を含む請求項２に記載のニューラルネットワークの生成方法。
【請求項４】
前記第３ネットワーク内の訓練後の前記マイクロユニットに基づいて、ニューラルネットワークモデルを得る前記ステップは、
訓練済みの前記マイクロユニットを結合させて、マイクロユニット数の異なる前記ニューラルネットワークモデルを得るステップであって、前記ニューラルネットワークモデルの数の範囲は２～Ｎ個であるステップを含む
請求項３に記載のニューラルネットワークの生成方法。
【請求項５】
前記第２ネットワークを用いて前記第３ネットワークの各マイクロユニットを訓練する前記ステップは、
前記第２ネットワークを前記訓練データセットと組み合わせて用いて、前記第２ネットワークの局所入出力の方式により前記第３ネットワークの各マイクロユニットを訓練するステップを含む
請求項１から４の何れか一項に記載のニューラルネットワークの生成方法。
【請求項６】
各前記ニューラルネットワークモデルをテストまたは評価し、既定条件を満たすニューラルネットワークモデルをターゲットニューラルネットワークとして選択するステップをさらに含む
請求項１に記載のニューラルネットワークの生成方法。
【請求項７】
各前記ニューラルネットワークモデルをテストまたは評価し、既定条件を満たすニューラルネットワークモデルをターゲットニューラルネットワークとして選択する前記ステップは、
既定のテストデータセットを用いて各前記ニューラルネットワークモデルをテストし、各前記ニューラルネットワークモデルのテスト正解率を取得するステップと、
各前記ニューラルネットワークモデルを用いて同じタスクをそれぞれ実行し、各前記ニューラルネットワークモデルの時間遅延を取得するステップと、
ターゲットニューラルネットワークモデルの時間遅延が既定値より低く、且つテスト正解率が最適値であれば、前記ターゲットニューラルネットワークモデルを前記ターゲットニューラルネットワークとして決定するステップと、
を含む請求項６に記載のニューラルネットワークの生成方法。
【請求項８】
最適マイクロユニットを取得する前記ステップは、
ニューラルネットワークアーキテクチャ探索ＮＡＳアルゴリズムに従って、既定の探索空間を用いて前記最適マイクロユニットのネットワーク構造を得るステップを含む
請求項１に記載のニューラルネットワークの生成方法。
【請求項９】
メモリと、プロセッサと、メモリに記憶されて且つプロセッサ上で実行できるコンピュータプログラムとを含む機器であって、
前記プロセッサにより前記コンピュータプログラムが実行された場合、請求項１から８の何れか一項に記載のニューラルネットワークの生成方法を実現する
機器。
【請求項１０】
コンピュータ実行可能な指令を記憶しているコンピュータ可読記憶媒体であって、
前記コンピュータ実行可能な指令は、請求項１から８の何れか一項に記載のニューラル

ネットワークの生成方法を実行するように構成された
コンピュータ可読記憶媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
本願は出願番号が２０２０１０５８１４８７．２で、出願日が２０２０年０６月２３日である中国特許出願に基づいて提出され、その中国特許出願の優先権を主張し、その中国特許出願の全ての内容を参考として本願に援用する。
【０００２】
本願は、コンピュータ技術の分野に関し、特に、ニューラルネットワークの生成方法、機器及びコンピュータ可読記憶媒体に関する。
【背景技術】
【０００３】
近年では、深層学習技術はより強い計算能力のサポートにより大きく発展し、コンピュータ視覚、音声認識、自然言語処理など多くの分野で大きな成功を収めている。研究者は、深層学習モデルを構成する複数の要素のうち、人工ニューラルネットワークの構造の違いがモデルの最終的な性能に莫大な影響を与えることを発見した。通常、特定の問題に対して適切なニューラルネットワークを設計するには、経験が豊富なアルゴリズムエンジニアが膨大な時間と労力を費やしてモデルの調整と検証を繰り返す必要があり、調整の効率が悪く、効果の保証が難しい。そのため、目標問題に合わせてニューラルネットワーク構造を自動的に決定できるニューラルネットワークアーキテクチャ探索技術は、深層学習分野における近年の研究の焦点となっている。
【０００４】
しかしながら、従来のニューラルネットワーク探索技術では、実際の応用時に、具体的な配置の問題が存在し、すなわち、特定の問題を解決するニューラルネットワークは、最終的には、計算能力が大きく異なるさまざまな端末、例えば、各種モデルの携帯電話、タブレットコンピュータ、ノートパソコン、デスクトップなどに配置する必要がある。そのため、様々な推論端末に向けて、サイズの揃ったニューラルネットワークモデルのグループを提供する必要があるが、このようなモデルグループの生成に要する計算能力のコストも莫大なものである。従来、ニューラルネットワークモデルグループの生成スキームとして、大まかに次のようなスキームが挙げられる。（１）マイクロユニットネットワークを構築して、１つずつ訓練する。（２）大量の第３ネットワークを含むハイパーネットワークを定義し、訓練する。二つのスキームのいずれも、計算能力を大量に消費する必要があることが明らかである。そのため、効率的なニューラルネットワークモデル構築方法を提案することは、実際の応用へのニューラルネットワーク探索の配置段階にとって重大な意義を持っている。
【発明の概要】
【発明が解決しようとする課題】
【０００５】
本願は、関連する技術問題の１つを少なくともある程度解決することを目的とする。このために、本願は、ニューラルネットワークの生成方法、機器及びコンピュータ可読記憶媒体を提案する。
【課題を解決するための手段】
【０００６】
第１態様において、本願の実施形態により提供されるニューラルネットワークの生成方法は、最適マイクロユニットを取得し、前記最適マイクロユニットに基づいて第１ネットワークを構築するステップと、既定の訓練データセットを用いて前記第１ネットワークを訓練して、第２ネットワークを得るステップと、マイクロユニットの数が前記第１ネットワークの最適マイクロユニットの数より少ない第３ネットワークを構築し、前記第２ネットワークを用いて前記第３ネットワークの各マイクロユニットを訓練するステップと、前

記第３ネットワーク内の訓練後の前記マイクロユニットに基づいて、ニューラルネットワークモデルを得るステップと、を含む。

【０００７】

第２態様において、本願の実施形態により提供される機器は、メモリ、プロセッサ及びメモリに記憶されて且つプロセッサ上で実行できるコンピュータプログラムを含み、前記プロセッサにより前記コンピュータプログラムを実行した場合、上記の第１態様の実施形態に記載のニューラルネットワークの生成方法を実現する。

【０００８】

第３態様において、本願の実施形態により提供されるコンピュータ可読記憶媒体にはコンピュータ実行可能な指令が記憶されており、前記コンピュータ実行可能な指令は、上記の第１態様の実施形態に記載のニューラルネットワークの生成方法を実行するように構成されている。

【０００９】

本願の他の特徴及び利点は、後の明細書において説明され、明細書から部分的に明らかになるか、または本願を実施することによって理解されるだろう。本願の目的及び他の利点は、明細書、特許請求の範囲及び図面において特別に指摘される構成によって達成し、得ることができる。添付図面は、本願の技術案の更なる理解を提供するものであり、明細書の一部を構成し、本願の実施形態と共に本願の技術案を解釈するために使用され、本願の技術案に対する制限を構成するものではない。

【図面の簡単な説明】

【００１０】

【図１】本願の一実施形態により提供されるシステムアーキテクチャプラットフォームの模式図である。

【図２】本願の一実施形態により提供されるニューラルネットワークの生成方法のフローチャートである。

【図３】本願の別の実施形態により提供されるニューラルネットワークの生成方法のフローチャートである。

【図４】本願の別の実施形態により提供されるニューラルネットワークの生成方法のフローチャートである。

【図５】本願の別の実施形態により提供されるニューラルネットワークの生成方法のフローチャートである。

【図６】本願の別の実施形態により提供されるニューラルネットワークの生成方法のフローチャートである。

【図７】本願の別の実施形態により提供されるニューラルネットワークの生成方法のフローチャートである。

【図８】本願の別の実施形態により提供されるニューラルネットワークの生成方法のフローチャートである。

【発明を実施するための形態】

【００１１】

本願の目的、技術案及び利点をより明らかにするために、以下では、添付図面及び実施形態を組み合わせて本願をさらに詳しく説明する。ここで説明する具体的な実施形態は本願を解釈するためだけに使われるものであって、本願を限定するために使われるものではない。

【００１２】

なお、装置の模式図には機能モジュール分割が示され、フローチャートには論理的順序が示されているが、場合によっては、装置内のモジュール分割またはフローチャート内の順序とは異なるように、図示又は説明されたステップを実行してもよい。明細書、特許請求の範囲または上記図面における用語「第１」、「第２」等は類似の対象を区別するためのものであり、必ずしも特定の順序又は前後の順番を記述するためのものではない。

【００１３】

従来のニューラルネットワークアーキテクチャ探索（Ｎｅｕｒａｌ　Ａｒｃｈｉｔｅｃｔｕｒｅ　Ｓｅａｒｃｈ，ＮＡＳ）においてアーキテクチャ探索ポリシーについての研究は、マクロ探索とマイクロ探索との２種類に大きく分けることができる。マクロ探索とは、探索アルゴリズムがニューラルネットワークを１つのまとまりと見なし、探索ポリシーでは選択的に特定の一層または複数の層を指定して代替的操作（層のタイプまたはパラメータの変更、新しい層の挿入）を実行する方法である。この方法の欠点は、探索空間が膨大すぎて、複雑度の高い問題について、適切なニューラルネットワークアーキテクチャを探し出すのにどれくらいの時間が必要かを判断するのは非常に困難である。そのため、現在注目度の低い研究である。マイクロ探索は人が設計するニューラルネットワークの優れた事例（例えばＲｅｓＮｅｔ、ＩｎｃｅｐｔｉｏｎＮｅｔなど）からの経験を参考して、探索空間を一組のマイクロユニットのスタックと見なして、探索ポリシーとして、相対的に限られた探索空間内で最適なネットワークマイクロアーキテクチャを探し出すだけでよい。具体的な応用時には、異なる数のマイクロユニットをスタックし、複雑度の異なる問題に適合することができる。マイクロ探索ＮＡＳアルゴリズムは、その大きさが合理的な探索空間と柔軟な利用方式から、現在注目度の非常に高い研究となっている。

【００１４】

しかしながら、従来のマイクロ探索ＮＡＳアルゴリズムでは、実際の応用時には、具体的な配置問題が存在する。すなわち、特定の問題を解決するニューラルネットワークは、最終的には、計算能力が大きく異なるさまざまな端末（例えば、各種モデルの携帯電話、タブレットコンピュータ、ノートパソコン、デスクトップなど）に配置する必要がある。そのため、ＮＡＳアルゴリズムの場合、様々な推論端末に向けて、サイズの揃ったニューラルネットワークモデルのグループを提供する必要があるが、このようなモデルグループの生成に要する計算能力のコストも莫大なものであり、大量な計算能力を大量に消費する必要がある。

【００１５】

上記の状況に基づいて、本願の実施形態はニューラルネットワークの生成方法、機器及びコンピュータ可読記憶媒体を提供し、最適マイクロユニットを取得し、該最適マイクロユニットを用いて第１ネットワークを構築することで、第１ネットワークが十分に強力な性能を有するようになり、実際の応用需要を満たすことができる。既定の訓練データセットを用いて第１ネットワークを訓練して、第２ネットワークを得て、マイクロユニットの数が第２ネットワークのマイクロユニットの数より少ない第３ネットワークを構築し、第２ネットワークを用いてすべての第３ネットワークの各マイクロユニットを訓練して、訓練を経た第３ネットワークのマイクロユニットに基づいて構築すると、ニューラルネットワークモデルを得ることができる。こうして、すべての第３ネットワークのマイクロユニット機能が第２ネットワークの機能に対応し、従来のマイクロユニットを一つずつ訓練する方法に比べて、訓練数を減らすことができ、計算能力の需要を効果的に下げることができるため、ニューラルネットワークモデルの生成コストを下げることができるとともに、すべての第３ネットワークのマイクロユニットは並行して訓練することができるため、ニューラルネットワークモデルの構築効率を効果的に高めることができる。

【００１６】

以下では、図面に関連して、本願の技術案を明確且つ完全に説明する。以下で説明される実施形態は本願のすべての実施形態ではなく、本願の一部の実施形態に過ぎないことは明らかである。

【００１７】

図１に示すように、図１は本願の一実施形態により提供される、ニューラルネットワークの生成方法を実行するためのシステムアーキテクチャプラットフォーム１００の模式図である。

【００１８】

図１に示す実施形態において、該システムアーキテクチャプラットフォーム１００はメモリ１２０とプロセッサ１１０とを含み、メモリ１２０とプロセッサ１１０は、バスまた

はその他の方法で接続されてもよく、図１にはバスで接続されている例が示されている。
【００１９】
メモリ１２０は、非一時的なコンピュータ可読記憶媒体として、非一時的なソフトウェアプログラムと、非一時的なコンピュータ実行可能なプログラムとを記憶するために使用することができる。さらに、メモリ１２０は、高速ランダムアクセスメモリを含むことができ、また非一時的なメモリ、例えば少なくとも１つの磁気ディスクメモリ装置、フラッシュメモリ装置、または他の非一時的なソリッドステートメモリ装置を含むことができる。いくつかの実施形態において、メモリ１２０は、プロセッサに対して遠隔地に配置されたメモリを含んでもよく、これらの遠隔メモリ１２０は、ネットワークを介して該システムアーキテクチャプラットフォーム１００に接続することができる。上記のネットワークの実例は、インターネット、社内イントラネット、ローカルエリアネットワーク、移動通信ネットワーク、及びこれらの組み合わせを含むが、これらに限定されない。
【００２０】
本願の実施形態に記載されたシステムアーキテクチャプラットフォーム１００および応用シナリオは、本願の実施形態により提供される技術案に対する限定を構成するものではなく、本願の実施形態の技術案をより明確に説明するためのものである。システムアーキテクチャプラットフォーム１００が進化し、新たな応用シナリオが現れたときに、本願の実施形態により提供される技術案が類似の技術課題に対しても同様に適用できることは、当業者にとって明らかである。
【００２１】
図１に示すシステムアーキテクチャプラットフォーム１００は本願の実施形態に対する限定を構成せず、図示より多い或いは少ない部品を含んでもよく、或いは一部の部品を組み合わせたり、異なる部品配置としたりしてもよいことは、当業者であれば理解できるであろう。
【００２２】
図１に示すシステムアーキテクチャプラットフォーム１００において、プロセッサ１１０は、メモリ１２０に記憶されているデータ処理プログラムを呼び出して、ニューラルネットワークの生成方法を実行することができる。
【００２３】
上述したシステムアーキテクチャプラットフォーム１００に基づいて、本願のニューラルネットワークの生成方法の各実施形態を以下に提案する。
【００２４】
図２を参照し、図２は本願の一実施形態により提供される、ニューラルネットワークの生成方法のフローチャートであり、該ニューラルネットワークの生成方法はステップＳ１００、ステップＳ２００、ステップＳ３００及びステップＳ４００を含むが、これらに限定されない。
【００２５】
ステップＳ１００において、最適マイクロユニットを取得し、最適マイクロユニットに基づいて第１ネットワークを構築する。
【００２６】
一実施形態において、ニューラルネットワークアーキテクチャは大量のマイクロユニットから構成され、ニューラルネットワーク探索は一般的にまず探索空間を定義する。このような探索空間はニューラルネットワークアーキテクチャを定義する基本的な構築モジュールを含む。探索空間が大きいほど、探索アルゴリズムが実行すべき反復回数が多くなり、使用されるグラフィックスプロセッサ（Ｇｒａｐｈｉｃｓ Ｐｒｏｃｅｓｓｉｎｇ Ｕｎｉｔ、ＧＰＵ）の計算能力も高く求められることが理解できる。最適化されたアルゴリズムを利用して最適なマイクロユニットネットワーク構造を自動的に探索することが可能になり、例えば、初期のアルゴリズムは進化的探索であり、良い検証指標が達成されるように、適応度に基づいて絶えず訓練され、突然変異する。また、強化学習に基づく探索アルゴリズムは、知的エージェントにネットワークアーキテクチャの最適構成を発見させ、こ

のような知的エージェントは通常、再帰型ニューラルネットワークにより実現され、より高い報酬を得られるニューラルアーキテクチャを探し出すことを目標とする。本実施形態において、マイクロ探索ＮＡＳアルゴリズムを用いて、探索空間から最適マイクロユニットを探索する。

【００２７】

一実施形態において、得られた最適マイクロユニットに基づいて第１ネットワークを構築し、該第１ネットワークは、最適マイクロユニットを利用して構成される高性能を持つユニットネットワークであると認識してもよく、該ユニットネットワークのサイズは、第１ネットワークが実際の応用処理を満足できる十分に強い性能を有するように、実際の需要に応じてサイズをカスタマイズすることができる。例えば、最適マイクロユニットは、スタック方式で構築されて第１ネットワークを形成し、セルネットワークの迅速な構築を容易にする。スタックされる最適マイクロユニットの数が多いほど、第１ネットワークのサイズが大きくなり、性能も強くなることが理解されるであろう。逆に、スタックされる最適マイクロユニットの数が少ないほど、第１ネットワークのサイズが小さくなる。こうして、実際の応用時の需要に応じて第１ネットワークを構築することで、複雑度の異なる問題に適合することができる。

【００２８】

なお、ステップＳ１００を実行する前に、ユーザは、実際の状況に応じてマイクロユニットネットワーク構造の探索空間を構築し、該探索空間を初期化してから、該探索空間に基づいて最適マイクロユニットを取得し、すなわち、限られた探索空間内で最適マイクロユニットを探し出すことができる。探索空間の合理的なサイズは、計算能力の要求を低くすることおよび処理効率を向上させることに有利である。

【００２９】

ステップＳ２００において、既定の訓練データセットを用いて第１ネットワークを訓練して、第２ネットワークを得る。

【００３０】

一実施形態において、既定の訓練データセットを用いて第１ネットワークを訓練し、訓練データセットは、マイクロユニットを訓練するためのデータ集合と理解してもよい。ステップＳ１００において、最適マイクロユニットを用いて構築された第１ネットワークは、パラメータが定義されていないネットワークモデルであり、第１ネットワークに相応の演算能力を持たせるために、対応する訓練データを用いて、該ネットワークモデルのパラメータに対応して訓練することで、訓練を経た第１ネットワークである第２ネットワークを得る。

【００３１】

深層学習を例として、訓練データセットを用いて深層学習モデルを訓練し、学習モデルが特定の性能指標を満たすようにし、訓練データが増えるにつれて、深層学習モデルの正解率が向上する。これによりわかるように、十分な訓練データリソースさえあれば、モデルを訓練する際のスケールアップによる正解率向上が期待できる。

【００３２】

既定の訓練データセットは、異なるモデル対象に応じて異なるタイプの訓練データセットを選択すること、および異なるモデルの性能要件に応じて異なる訓練データ量を選択することとして理解してもよい。例えば、訓練の対象が画像認識モデルである場合、既定の訓練データセットとして大量の画像を選択することができる。また例えば、訓練の対象が音声認識モデルである場合、既定の訓練データセットとして大量の音声データを選択することができる。

【００３３】

ステップＳ３００において、マイクロユニットの数が第１ネットワークの最適マイクロユニットの数より少ない第３ネットワークを構築し、第２ネットワークを用いて第３ネットワークの各マイクロユニットを訓練する。

【００３４】

一実施形態において、第３ネットワークは、パラメータが定義されていないネットワークモデルであり、構築される第３ネットワークのマイクロユニット数は、第１ネットワークの最適マイクロユニットの数より少ない。当然のことながら、マイクロユニット数の異なる複数の第３ネットワークを構築することが可能で、得られる第３ネットワークのマイクロユニットの数は、第１ネットワーク内の最適マイクロユニットの数より少ない。コンピュータによる演算を容易にするために、第３ネットワークのマイクロユニット数と第１ネットワークの最適マイクロユニット数との関係を倍数、累乗関係等に設定することができる。例えば、第１ネットワークの最適マイクロユニット数を２０、４０、１２０個などとしてもよく、第３ネットワークのマイクロユニット数を２、４、１０個などとしてもよく、マイクロユニット数がそれぞれ２、４、１０個などである第３ネットワークを構築することができる。

【００３５】

第３ネットワークの構築が完了した後、第２ネットワークを用いてすべての第３ネットワークを訓練し、具体的には、第２ネットワークにより、第３ネットワーク内の各マイクロユニットをそれぞれ訓練する。また、各第３ネットワークのマイクロユニット数は、いずれも第２ネットワークのマイクロユニット数より少なく、マイクロユニットの数の多いネットワークモデルに比べて、マイクロユニット数の少ないネットワークモデルを訓練する方が、訓練スピードがもっと速い。また、第３ネットワークのマイクロユニットは並行して訓練することが可能で、すなわち同時に複数のマイクロユニットを訓練することが可能であるため、訓練効率がもっと高く、計算能力の要求を大幅に下げることができる。訓練を経た第３ネットワークは、第２ネットワークと同じ機能を有するようになる。例えば、第２ネットワークのマイクロユニット数が３２個であり、第３ネットワークのマイクロユニット数が４個である場合、該第３ネットワークの４個のマイクロユニットは、それぞれ訓練された後、第２ネットワークの３２個のマイクロユニットと同じ性能を有することができる。

【００３６】

実施形態において、第２ネットワークを教師ネットワークとして定義し、第３ネットワークを学生ネットワークとして定義してもよく、教師ネットワークによりすべての学生ネットワークのマイクロユニットを訓練することができるとともに、すべての学生ネットワークのマイクロユニットは並行して訓練することができるため、訓練時間を削減し、訓練効率を向上させることができる。

【００３７】

ステップＳ４００において、第３ネットワーク内の訓練後のマイクロユニットに基づいて、ニューラルネットワークモデルを得る。

一実施形態において、上記のステップＳ３００に従って、すべての第３ネットワークの各マイクロユニットを訓練する。訓練後、異なる第３ネットワークのマイクロユニットを並べて異なるニューラルネットワークモデルを得る。例えば、第３ネットワークのマイクロユニットは数が４個で、すべて訓練済みであり、該第３ネットワークの４個の訓練済みのマイクロユニットを並べてマイクロユニット数が４の１つのニューラルネットワークモデルを得る。異なる第３ネットワークの訓練済みマイクロユニットを組み合わせて１つのニューラルネットワークモジュールを得ることもできる。例えば、一つの第３ネットワーク内の訓練済みの２つのマイクロユニットと、もう一つの第３ネットワーク内の訓練済みの１つのマイクロユニットとを組み合わせて、マイクロユニット数が３個のニューラルネットワークモデルを得る。このように、ステップＳ４００が完了すると、複数のニューラルネットワークモデルが得られるため、実際の応用時に適切なニューラルネットワークモデルを選択して異なる端末に配置して、計算能力の消費を効果的に削減することができる。

【００３８】

本実施形態は、上述したステップＳ１００、ステップＳ２００、ステップＳ３００及びステップＳ４００のニューラルネットワークの生成方法により、実際の応用需要に応じて計算能力の異なる端末に適したニューラルネットワークモデルを得るため、従来のニュー

ラルネットワーク訓練方法に比べて、訓練回数を効果的に削減し、計算能力の要求を低減することで、ニューラルネットワークモデルの生成コストを下げることができる。

【００３９】

一実施形態において、ステップＳ１００において、最適マイクロユニットに基づいて第１ネットワークを構築する流れは、以下のステップを含むことができるが、これらに限定されない。

【００４０】

ステップＳ１１０において、事前定義のサイズに従って、最適マイクロユニットをＮ回スタックして第１ネットワークを生成し、ここで、Ｎは２の整数乗である。

【００４１】

ここで、事前定義のサイズは、構築される第１ネットワークの、実際の応用需要に応じて設定されるマイクロユニット数として捉えてもよく、最適マイクロユニットはスタック方式で組み合わせられ、最適マイクロユニットのスタックされる数が多いほど、第１ネットワークのサイズが大きくなり、性能が高くなる。実施形態において、最適マイクロユニットをＮ回スタックすることで、第１ネットワークを生成する。Ｎは適合性が十分に強いネットワークを構成するために必要なマイクロユニット数であり、Ｎは３２、６４、１２８など、２の整数乗に設定される。例えば、画像認識のタスクを処理する必要があり、実際の推計によると３２個のマイクロユニットを持つネットワークモデルを構築して初めてマッチングできるとすると、第１ネットワークの最適マイクロユニット数Ｎは３２個、あるいは３２個より大きい数としてもよく、例えばＮの取る値を６４個とすることで、十分に強い性能を有するネットワークモデルを得ることが可能で、複雑度の異なる問題に適合することができる。

【００４２】

一実施形態において、具体的には、トポロジの原理に従って最適マイクロユニットをスタックしてもよい。トポロジ構造の構築原理は、物体の形状や大きさを考慮することなく、物体間の位置関係のみを考慮することである。トポロジでは、最適マイクロユニットをその大きさ、形状に関係のない点として捉えることができる。該トポロジは、ネットワークマクロトポロジまたはシーケンシャルトポロジとすることが可能で、トポロジに応じて最適マイクロユニット数を設定し、すべての最適マイクロユニットを接続して第１ネットワークを構成する。

【００４３】

一実施形態において、ステップＳ１００における最適マイクロユニットを取得する流れは、以下のステップを含むが、これらに限定されるものではない。

【００４４】

ステップＳ１２０において、ＮＡＳアルゴリズムに従って、既定の探索空間を用いて最適マイクロユニットのネットワーク構造を得る。

【００４５】

実施形態において、ＮＡＳアルゴリズムには、特定の問題のために最適なニューラルネットワークアーキテクチャを自動的に見つけるという特徴がある。ＮＡＳアルゴリズムを用いれば、最適マイクロユニットを迅速に得ることができる。ＮＡＳアルゴリズムは、探索空間と呼ばれる候補ニューラルネットワーク構造の集合を与え、探索ポリシーを用いて該探索空間から最適なネットワーク構造を探し出すものであり、ここではＮＡＳアルゴリズムの具体的な原理については説明を省略する。

【００４６】

図３を参照し、一実施形態において、ステップＳ３００における第３ネットワークを構築する流れは、以下のステップを含むことができるが、これらに限定されない。

【００４７】

ステップＳ３１０において、マイクロユニットの数がＮ／ＭでＮ／Ｍ＞1を満たす第３ネットワークを構築し、ここで、Ｍは２の整数乗である。

【００４８】

ステップＳ３２０において、すべての第３ネットワークを初期化する。

【００４９】

ここで、第３ネットワークを学生ネットワークとして定義し、第３ネットワークのマイクロユニット数を、第１ネットワークの最適マイクロユニット数Ｎに応じて決定し、Ｎの取る値は２の整数乗であり、第３ネットワークのマイクロユニット数はＮ／ＭでＮ／Ｍ＞１を満たし、Ｍは２の整数乗である。なお、Ｎ、Ｍの取る値の範囲はいずれも２の整数乗であり、Ｎ／Ｍ＞１はＮの取る値がＭより大きいことを示し、Ｎ／Ｍ＞１の条件を満たすことを前提に、マイクロユニット数の異なる複数の第３ネットワークを構築することが可能で、すなわち、マイクロユニット数がＮ／２、Ｎ／４、Ｎ／８．．．２の学生ネットワークを構築することができる。例えば、第１ネットワークの最適マイクロユニットの数Ｎが３２の場合、Ｍの値を２、４、８、１６とすることができるため、マイクロユニット数が３２／２＝１６、３２／４＝８、３２／８＝４、３２／１６＝２を含む合計４つの学生ネットワークを構築する必要があることが分かる。そして、すべての学生ネットワークを初期化し、各学生ネットワークのモデルのパラメータにデフォルト値を与えることで、マイクロユニット数の異なる学生ネットワークを構築して得ることができる。

【００５０】

なお、学生ネットワークのマイクロユニット数ｎと第１ネットワークの最適マイクロユニット数Ｎとが対数関係にあり、すなわち、ｎ＝ｌｏｇＮであるため、教師ネットワークを用いてすべての学生ネットワークのマイクロユニットを訓練する場合、訓練数はＮ個のネットワーク構造からｌｏｇＮ個のネットワーク構造に減少し、訓練に必要な時間を効果的に下げることができ、訓練スピードは更に速くなる。また、第３ネットワークのマイクロユニットは並行して訓練することが可能で、すなわち同時に複数のマイクロユニットを訓練することが可能であるため、従来のマイクロユニットネットワークを１つずつ訓練するモードに比べて、訓練効率がより高く、計算能力の要求を大幅に下げることができる。

【００５１】

一実施形態において、ステップＳ４００における第３ネットワーク内の訓練後のマイクロユニットに基づいて、ニューラルネットワークモデルを得るステップは、以下のステップを含むことができるが、これらに限定されない。

【００５２】

ステップＳ４１０において、訓練済みのマイクロユニットを結合させて、マイクロユニット数の異なるニューラルネットワークモデルを得て、ニューラルネットワークモデルの数の範囲は２～Ｎ個である。

【００５３】

教師ネットワークを用いて学生ネットワークの各マイクロユニットを訓練し、訓練済みのマイクロユニットを並べて組み合わせるが、具体的には、並べて組み合わせる方式は訓練されるマイクロユニットの数に応じて決定される。例えば、学生ネットワークのマイクロユニット数が４個で、すべて訓練済みであり、該学生ネットワークの４個の訓練済みのマイクロユニットを並べてマイクロユニット数が４の１つのニューラルネットワークモデルを得る。また例えば、一つの学生ネットワーク内の訓練済みの２つのマイクロユニットと、もう一つの学生ネットワーク内の訓練済みの１つのマイクロユニットとを組み合わせて、マイクロユニット数が３のニューラルネットワークモデルを得る。すべての学生ネットワークのマイクロユニットを組み合わせることにより、すべてのサイズを有するニューラルネットワークモデルセットを得ることができる。該すべてのサイズを有するニューラルネットワークモデルセットは、マイクロユニット数の異なるすべてのニューラルネットワークの集合として捉えることができる。

【００５４】

訓練を経たすべての学生ネットワークのマイクロユニットの機能は教師ネットワークのある局所写像に対応しているので、組み合わせにより得られるネットワークは、いずれも直接端末に配置するまたはわずかな微調整をしてから端末に配置することができる。

【００５５】

なお、最適なマイクロユニットがＮ回スタックされて構成された第１ネットワークを例に、第１ネットワークを訓練した後に、マイクロユニット数Ｎの教師ネットワークを得るとともに、マイクロユニット数Ｎ／２、Ｎ／４、Ｎ／８．．．２の学生ネットワークを構築することができる。教師ネットワークを用いてすべての学生ネットワークを訓練し、訓練済みのマイクロユニットを組み合わせて、２～Ｎ個のニューラルネットワークモデルを得る。たとえば、Ｎが３２の場合、マイクロユニット数がそれぞれ２、４、８、１６の学生ネットワークを構築し、４つの学生ネットワークを得る。教師ネットワークは３２個のマイクロユニットを有し、教師ネットワークにより、２つのマイクロユニットを有する学生ネットワークを訓練し、２つのマイクロユニットを有するニューラルネットワークモデルを得るが、該２つのマイクロユニットを有するニューラルネットワークモデルの機能はすべて教師ネットワークに対応する。教師ネットワークにより学生ネットワークを訓練することにより、教師ネットワークのマイクロユニットの機能が学生ネットワークのマイクロユニットにマッピングされることが分かる。また、教師ネットワークにより、４つのマイクロユニットを有する学生ネットワークを訓練し、４つのマイクロユニットを有するニューラルネットワークモデルを得る。このように、学生ネットワークのマイクロユニットを組み合わせることで、異なるニューラルネットワークモデルを得ることができる。

【００５６】

一実施形態において、第２ネットワークを用いて第３ネットワークの各マイクロユニットを訓練するステップは、以下のステップを含むことができるが、これらに限定されない。

【００５７】

ステップＳ３３０において、第２ネットワークを訓練データセットと組み合わせて用いて、第２ネットワークの局所入出力の方式により第３ネットワークの各マイクロユニットを訓練する。

【００５８】

訓練時、訓練データセットに合わせて教師ネットワークを用いて、各学生ネットワークの各マイクロユニットをそれぞれ訓練し、該訓練データセットは教師ネットワークを訓練するための訓練データである。具体的には、教師ネットワークの局所入出力の訓練方式により、各学生ネットワークの各マイクロユニットを訓練する。

【００５９】

【数１】

具体例を挙げて説明すると、第１ネットワークの最適マイクロユニット数はＮであり、表示の便宜上、Ｎを８とし、実際の応用時に、３２、６４またはそれ以上の値に設定すべきである。例えば第１ネットワークは$a\text{-}a\text{-}a\text{-}a\text{-}a\text{-}a\text{-}a\text{-}a$であり、第１ネットワークを訓練すると、教師ネットワーク$a_1\text{-}a_2\text{-}a_3\text{-}a_4\text{-}a_5\text{-}a_6\text{-}a_7\text{-}a_8$を得る。そして、マイクロユニット数がそれぞれ２と４の学生ネットワークを構築し、教師ネットワークを用いて該２つの学生ネットワークを訓練する。マイクロユニット数４の学生ネットワーク$a_1'-a_2'-a_3'-a_4'$を例として説明すると、ある画像データサンプルSに対して、学生ネットワークマイクロユニットa_i'のパラメータ勾配は、教師ネットワークa_1ユニットの入力テンソルおよびa_2ユニットの出力テンソルから以下のように算出できる。

【００６０】

【数２】

$$i_{a_1} = S;$$
$$o_{a_2} = f_{a_2}(W_{a_2}, f_{a_1}(W_{a_1}, i_{a_1}));$$
$$L_{a_1'} = E(f_{a_1'}(W_{a_1'}, i_{a1}), o_{a_2});$$
$$\Delta W_{a_1'} = \frac{\partial L_{a_1'}}{\partial w_{a_1'}};$$
$$W_{a_1'} = W_{a_1'} - \eta \Delta W_{a_1'};$$

【００６１】
【数３】

その中で、i_{a_1} は教師ネットワークa_1ユニットの入力を表し、Sは具体的なデータサンプルを表し、o_{a2} は教師ネットワークマイクロユニットa_2の出力を表し、W_{a1} 、 f_{a1} はマイクロユニットa_1のモデルパラメータ、活性化関数であり、同様に、W_{a2} 、 f_{a2} および $W_{a_1'}, f_{a_1'}$ はマイクロユニットa_2、a_1' のモデルパラメータ、活性化関数であり、$L_{a_1'}$ はマイクロユニットa_1' のデータサンプルＳに基づいて算出される損失関数であり、Ｅは損失関数計算方法を表し、$\Delta W_{a_1'}$ はマイクロユニットa_1' の損失関数$L_{a_1'}$ に基づいて偏導関数を求めて算出される本ユニットモデルのパラメータ勾配であり、η は今回更新されるモデルパラメータの学習率を表す。

【００６２】
【数４】

同様に、学生ネットワーク内の残りの３つのユニットa_2'、a_3'、a_4' のパラメータ勾配計算方式は上記の計算方式と一致する。なお、訓練した後、a_1' で教師ネットワークの局所写像a_1-a_2 を置き換え、a_2' で教師ネットワークの局所写像a_3-a_4 を置き換え、a_3' で教師ネットワークの局所写像a_5-a_6 を置き換え、a_4' で教師ネットワークの局所写像a_7-a_8 を置き換え、得られるニューラルネットワークモデルは$a_1' - a_2' - a_3' - a_4'$ である。

【００６３】
【数５】

ユニット数が２の学生ネットワーク$a_1'' - a_2''$の場合、a_1''で教師ネットワークの局所写像$a_1 - a_3$を置き換え、a_2''で教師ネットワークの局所写像$a_5 - a_7$を置き換え、得られるニューラルネットワークモデルは$a_1'' - a_2''$である。このように、教師ネットワークの局所入出力により学生ネットワークのマイクロユニットを訓練することで、すべての学生ネットワークの各マイクロユニットに対する訓練を行い、さらに、すべての学生ネットワークのマイクロユニットはいずれも並行して訓練できるため、所要時間を大幅に削減し、訓練効率を効果的に向上させることができる。

【００６４】

一実施形態において、ステップＳ４１０において、訓練済みのマイクロユニットを組み合わせて、２～Ｎ個のニューラルネットワークモデルを得る。すなわち、訓練済みのマイクロユニットを並べて組み合わせることにより、すべてのサイズを有するニューラルネットワークモデルセットを得ることができる。例えば、上述したステップＳ３３０の例において、次のようなすべてのサイズを有するニューラルネットワークモデルを得ることができる。

【００６５】

【数６】

２つのマイクロユニット：$a_1'' - a_2''$、

３つのマイクロユニット：$a_1'' - a_3' - a_4'$又は$a_1' - a_2' - a_2''$、

４つのマイクロユニット：$a_1' - a_2' - a_3' - a_4'$、

５つのマイクロユニット：$a_1' - a_2' - a_3' - a_7 - a_8$、

６つのマイクロユニット：$a_1' - a_2' - a_5 - a_6 - a_7 - a_8$、

７つのマイクロユニット：$a_1' - a_3 - a_4 - a_5 - a_6 - a_7 - a_8$、

８つのマイクロユニット：$a_1 - a_2 - a_3 - a_4 - a_5 - a_6 - a_7 - a_8$。

【００６６】

ここで、マイクロユニット数が５、６、７の組み合わせ方は以上に示したモデルに限らず、ここでは繰り返し列挙しない。

【００６７】

図４を参照し、一実施形態において、ニューラルネットワークの生成方法の流れは、以下のステップをさらに含むが、これに限定されない。

【００６８】

ステップＳ５００において、各ニューラルネットワークモデルをテストまたは評価し、既定条件を満たすニューラルネットワークモデルをターゲットニューラルネットワークとして選択する。

【００６９】

実施形態のステップＳ１００、ステップＳ２００、ステップＳ３００およびステップＳ４００が完了すると、すべてのサイズを有するニューラルネットワークモデルが得られる。ニューラルネットワークは、最終的には、計算能力が大きく異なるさまざまな端末に配置する必要があるため、ニューラルネットワークモデルの性能要件や配置対象端末への推

論時間遅延要求などの制約条件に基づいて、すべての組み合せネットワークの中から適切なニューラルネットワークモデルを選択し、そして適切なニューラルネットワークモデルをターゲットニューラルネットワークとして、端末に配置する。

【００７０】

具体的には、各ニューラルネットワークモデルをテストまたは評価する方式により、既定条件を満たすニューラルネットワークモデルをターゲットニューラルネットワークとして絞り込む。ここで、改良されたニューラルネットワークモデルを得るために、テスト方法は、テストデータセットを用いてすべてのニューラルネットワークモデルをテストすることができる。評価方法としては、ニューラルネットワークモデルの１秒あたりの浮動小数点演算回数（ＦＬＯＰｓ）をオフラインで評価してもよく、例えばＦＬＯＰｓが４００Ｍ以下のニューラルネットワークモデルをターゲットニューラルネットワークとして選択する。

【００７１】

図５を参照し、一実施形態において、各ニューラルネットワークモデルをテストまたは評価し、既定条件を満たすニューラルネットワークモデルをターゲットニューラルネットワークとして選択するステップは、以下のステップを含む。

【００７２】

ステップＳ５１０において、既定のテストデータセットを用いて各ニューラルネットワークモデルをテストし、各ニューラルネットワークモデルのテスト正解率を取得する。

【００７３】

ステップＳ５２０において、各ニューラルネットワークモデルを用いて同じタスクをそれぞれ実行し、各ニューラルネットワークモデルの時間遅延を取得する。

【００７４】

ステップＳ５３０において、ターゲットニューラルネットワークモデルの時間遅延が既定値より低く、且つテスト正解率が最適値であれば、ターゲットニューラルネットワークモデルをターゲットニューラルネットワークとして決定する。

【００７５】

ここで、ステップＳ５１０において、既定のテストデータセットは、ニューラルネットワークモデルをテストするためのデータ集合であると理解してもよく、異なる端末性能要件に応じて、相応のテストデータが選択される。各ニューラルネットワークモデルをテストすることで、各ニューラルネットワークモデルのテスト正解率を得ることが可能である。テスト正解率が高いほど、性能指標が高いため、テスト正解率の高いニューラルネットワークモデルが優先的に選択されることが分かる。

【００７６】

ステップＳ５２０は、ターゲット端末上で実行され、各ニューラルネットワークモデルを用いて同じタスクをそれぞれ実行し、各ニューラルネットワークモデルの時間遅延を取得する。このステップは、端末上での各ニューラルネットワークモデルの時間遅延指標をテストするために用いられるが、一般的には、送信時間遅延と伝播時間遅延が主に考慮される指標であり、より高い時間遅延は、より大きいデータ処理の時間間隔とより低いタスク実行効率を示すため、タスク実行時の時間遅延の低いニューラルネットワークモデルを優先的に選択する。

【００７７】

このように、時間遅延とテスト正解率との２つの指標に基づいて各ニューラルネットワークモデルをテストし、テストされたニューラルネットワークモデルの時間遅延が既定値より低く、且つテスト正解率が最適値であれば、該ニューラルネットワークモデルをターゲットニューラルネットワークとして決定し、実施形態において、時間遅延の制限条件を５００ｍｓ以下であるように設定してもよいが、４００ｍｓ以下または他の既定値としてもよい。テストセットの正解率は９５％以上達するようにすることが可能で、実施形態においては９６．３％に達するように選択される。

【００７８】

上述した各実施形態におけるニューラルネットワークの生成方法の具体的なステップの流れをより明確に説明するために、以下に３つの実施形態を用いて説明する。

【００７９】

実施形態１：

図６に示すように、ＮＡＳアルゴリズムの探索空間がマイクロユニット空間であり、マイクロユニットが単入力単出力構造であり、マクロユニットトポロジがシーケンシャル構造であるように定義し、複数のマイクロユニットにより構成されたネットワークモデルは、画像分類問題を解決するために利用され、マルチチャンネル画像データを入力として、画像内容の実際の分類を出力する。ニューラルネットワークの生成方法は以下のようなステップを含む。

【００８０】

ステップＳ６１０において、マイクロユニットネットワーク構造の探索空間を初期化し、マイクロ探索ＮＡＳアルゴリズムを適用して、最適マイクロユニットを取得する。

【００８１】

【数７】

ステップＳ６２０において、シーケンシャルトポロジのマイクロユニット数を設定し、事前定義のサイズを有する第１ネットワーク$a\text{-}a\text{-}a\text{-}a\text{-}a\text{-}a\text{-}a\text{-}a$を構成する。

【００８２】

【数８】

ステップＳ６３０において、訓練データセットの画像データおよび各画像に対応する分類情報を用いて第１ネットワークを訓練し、教師ネットワーク

$a_1\text{-}a_2\text{-}a_3\text{-}a_4\text{-}a_5\text{-}a_6\text{-}a_7\text{-}a_8$を得る。

【００８３】

ステップＳ６４０において、学生ネットワークを構築し、ユニット数４と２の学生ネットワークの各マイクロユニットを、訓練データセットに合わせて教師ネットワークを用いてそれぞれ訓練する。

【００８４】

ステップＳ６５０において、すべての学生ネットワークのマイクロユニットを訓練した後、並べて組み合わせることで、すべてのサイズを有するニューラルネットワークモデルセットを得る。

【００８５】

教師ネットワークにより学生ネットワークを訓練する具体的なステップの流れは、実施形態のステップＳ３３０の流れを参照し、ここでは説明を省略する。該実施形態において、最適マイクロユニットは単入力単出力構造である。

【００８６】

実施形態２：

図７に示すように、ＮＡＳアルゴリズムの探索空間がマイクロユニット空間であり、マイクロユニットが二重入力単出力構造であるように定義し、複数のマイクロユニットにより構成されたネットワークモデルは、画像分類問題を解決するために利用され、マルチチャンネル画像データを入力として、画像内容の実際の分類を出力する。ニューラルネットワークの生成方法は以下のようなステップを含む。

【００８７】

ステップＳ７１０において、マイクロユニットネットワーク構造の探索空間を初期化し、マイクロ探索ＮＡＳアルゴリズムを適用して、最適マイクロユニットを取得する。

【００８８】

【数９】

ステップＳ７２０において、複雑なトポロジネットワークのマイクロユニット数を設定

$$a\text{-}a\text{-}a\text{-}a$$

し、事前定義のサイズを有する第１ネットワーク $a\text{-}a\text{-}a\text{-}a$ を構成する。

【００８９】

【数１０】

ステップＳ７３０において、訓練データセット内の画像データおよび各画像に対応する

$$a_1\text{-}a_3\text{-}a_5\text{-}a_7$$

分類情報を用いて本モデルを訓練し、教師ネットワーク $a_2\text{-}a_4\text{-}a_6\text{-}a_8$ を得る。

【００９０】

ステップＳ７４０において、学生ネットワークを構築し、ユニット数４と２の学生ネットワークの各マイクロユニットを、訓練データセットに合わせて教師ネットワークを用いてそれぞれ訓練する。

【００９１】

ステップＳ７５０において、すべての学生ネットワークのマイクロユニットを訓練した後、並べて組み合わせることで、すべてのサイズを有するニューラルネットワークモデルセットを得る。

【００９２】

【数１１】

マイクロユニット数４の学生ネットワーク $a_1'-a_2'-a_3'-a_4'$ を例として説明すると、ある画像データサンプル S に対して、学生ネットワークマイクロユニット a_2' のパラメータ勾配は、教師ネットワーク a_1、a_2ユニットの出力テンソルおよび a_3、a_4ユニットから出力される合成テンソルから以下のように算出できる。

【００９３】

【数１２】

$$i_{a_1}=S$$

$$o_{a_1}=f_{a_1}(W_{a_1},S,S),o_{a_2}=f_{a_2}(W_{a_2},o_{a_1},S)$$

$$o_{a_3}=f_{a_3}(W_{a_3},o_{a_2},o_{a_1}),o_{a_4}=f_{a_4}(W_{a_1},o_{a_3},o_{a_2})$$

$$L_{a_2'}=E(f_{a_2'}(W_{a_2'},o_{a_2},o_{a_1}),(o_{a_3}\oplus o_{a_4}))$$

$$\Delta W_{a_2'}=\frac{\partial L_{a_2'}}{\partial w_{a_2'}}$$

$$W_{a_2'}=W_{a_2'}-\eta\Delta W_{a_2'}$$

【００９４】

【数１３】

その中で、i_{a_1} は教師ネットワークa_1ユニットの入力を表し、Sは具体的なデータサンプルを表し、o_{a_1} 、o_{a_2} 、o_{a_3} およびo_{a_4} は教師ネットワークマイクロユニットa_1、a_2、a_3、a_4の出力を表し、W_{a1} 、 f_{a1} はマイクロユニットa_1のモデルパラメータ、活性化関数であり、同様に、W_{a2} 、 f_{a2} 、 W_{a3} 、 f_{a3} 、 W_{a4} 、 f_{a4} および$W_{a_2'}, f_{a_2'}$ はマイクロユニットa_2、a_3、a_4、a_2' のモデルパラメータ、活性化関数であり、$L_{a_2'}$ はマイクロユニットa_2' のサンプルＳに基づいて算出される損失関数であり、Ｅは損失関数計算方法を表し、$o_{a_3} \oplus o_{a_4}$ は２つのマイクロユニットの出力の連結を表し、$\Delta W_{a_2'}$ はマイクロユニットa_2'の損失関数$L_{a_2'}$ に基づいて偏導関数を求めて算出される本ユニットモデルのパラメータ勾配であり、η は今回更新されるモデルパラメータの学習率を表す。

【００９５】
【数１４】

同様に、学生ネットワーク内の残りの３つのユニットa_1'、a_3'、a_4' のパラメータ勾配計算方式は上記の計算方式と一致する。マイクロユニット数が２の学生ネットワーク$a_1'' - a_2''$ も同様である。

【００９６】
ステップＳ７５０において、３つのマイクロユニットネットワークを組み合わせる例として、以下の２つの組み合わせ方がある。
【００９７】
【数１５】

１、$a_1'' - a_3' - a_4'$ ：a_1'' で教師ネットワークの局所写像$a_1 - a_3$ $a_2 - a_4$ を置き換え、a_3' でa_5 a_6 を置き換え、a_4' でa_7 a_8 を置き換える。

【００９８】
【数１６】

２、$a_1' - a_2' - a_2''$ ：a_2'' で教師ネットワークの局所写像$a_5 - a_7$ $a_6 - a_8$ を置き換え、a_1' でa_1 a_2 を置き換え、a_2' でa_3 a_4 を置き換える。

【００９９】
該実施形態において、最適マイクロユニットは二重入力単出力構造である。

【０１００】
実施形態３：
図８に示すように、すべてのサイズを有するニューラルネットワークモデルを取得した後、端末を配置する制限条件は、オンライン推論の時間遅延が５００ｍｓを超えてはならないことであり、ニューラルネットワークモデルのテスト流れは、以下のステップを含む。
【０１０１】
ステップＳ８１０において、テストデータセット上でマイクロユニット数の異なるすべてのニューラルネットワークモデルの予測正解率を評価する。
【０１０２】
ステップＳ８２０において、ターゲット端末にマイクロユニット数の異なるすべてのニューラルネットワークモデルを適用して同じ推論タスクをそれぞれ実行し、端末が推論タスクを実行する際の時間遅延指標を記録する。
【０１０３】
ステップＳ８３０において、予測正解率と時間遅延指標とに基づいて、予め設定された推論時間遅延制限条件の下で、推論正解率の最も高いモデルを選択して端末に配置する。
【０１０４】
該実施形態において、端末では５つ以上のマイクロユニットからなる分類ネットワークを実行し、推論時間遅延はいずれも５００ｍｓより大きく、制約条件を満たしていないため、４つ以下のマイクロユニットからなるネットワーク上で性能の最も良いニューラルネットワークモデルを選択することができる。たとえば、条件を満たすモデルのうち正解率の最も良いネットワークの場合、テストセットの正解率が９６．３％に達することができるため、該端末に配置する際に、該ニューラルネットワークモデルを優先的に選択すべきである。
【０１０５】
また、本願の一実施形態はさらに、機器を提供し、該機器は、メモリ、プロセッサ及びメモリに記憶されて且つプロセッサ上で実行できるコンピュータプログラムを含む。プロセッサ及びメモリは、バスまたは他の方法によって接続されてもよい。
【０１０６】
メモリは、非一時的なコンピュータ可読記憶媒体として、非一時的なソフトウェアプログラムと、非一時的なコンピュータ実行可能なプログラムとを記憶するために使用することができる。さらに、メモリは、高速ランダムアクセスメモリを含むことができ、また非一時的なメモリ、例えば少なくとも１つの磁気ディスクメモリ装置、フラッシュメモリ装置、または他の非一時的なソリッドステートメモリ装置を含むことができる。いくつかの実施形態において、メモリは、プロセッサに対して遠隔地に配置されたメモリを含んでもよく、これらの遠隔メモリは、ネットワークを介してこのプロセッサに接続することができる。上記のネットワークの実例は、インターネット、社内イントラネット、ローカルエリアネットワーク、移動通信ネットワーク、及びこれらの組み合わせを含むが、これらに限定されない。
【０１０７】
なお、本実施形態の端末は、図１に示す実施形態におけるシステムアーキテクチャプラットフォーム１００を含むことができ、本実施形態における端末と図１に示す実施形態におけるシステムアーキテクチャプラットフォーム１００とは同一の発明構想に属するため、これらの実施形態は同一の実現原理及び技術的効果を有し、ここでは詳しく説明しない。
【０１０８】
上述した実施形態のニューラルネットワークの生成方法を実現するために必要な非一時的なソフトウェアプログラムおよび指令はメモリに記憶されており、プロセッサにより実行された場合、上述した実施形態におけるニューラルネットワークの生成方法、例えば、上述した図２における方法ステップＳ１００からＳ４００、図３における方法ステップＳ３１０からＳ３２０、図４における方法ステップＳ５００、図５における方法ステップＳ５１０からＳ５３０、図６における方法ステップＳ６１０からＳ６５０、図７における方

法ステップＳ７１０からＳ７５０、図８における方法ステップＳ８１０からＳ８３０を、実行する。

【０１０９】

以上に説明された装置実施形態は、単に例示的なものであり、分離された部品として説明されたユニットは、物理的に分離されていてもよく、そうでなくてもよく、すなわち、一箇所にあってもよく、または複数のネットワークユニットに分散されていてもよい。本実施形態の態様の目的を達成するために、これらのモジュールの一部または全部を実際の必要に応じて選択することが可能である。

【０１１０】

また、本願の一実施形態は、コンピュータ可読記憶媒体をさらに提供し、該コンピュータ可読記憶媒体にはコンピュータ実行可能な指令が記憶されており、該コンピュータ実行可能な指令は、プロセッサまたはコントローラ、例えば、上述した端末実施形態におけるプロセッサによって実行されることで、上述した実施形態におけるニューラルネットワークの生成方法、例えば、上述した図２における方法ステップＳ１００からＳ４００、図３における方法ステップＳ３１０からＳ３２０、図４における方法ステップＳ５００、図５における方法ステップＳ５１０からＳ５３０、図６における方法ステップＳ６１０からＳ６５０、図７における方法ステップＳ７１０からＳ７５０、図８における方法ステップＳ８１０からＳ８３０を、上記プロセッサに実行させることができる。

【０１１１】

本願の実施形態によれば、最適マイクロユニットを取得し、該最適マイクロユニットを用いて第１ネットワークを構築することで、第１ネットワークが十分に強力な性能を有するようになり、実際の応用需要を満たすことができる。既定の訓練データセットを用いて第１ネットワークを訓練して、第２ネットワークを得る。マイクロユニット数が第１ネットワークの最適マイクロユニット数より少ない第３ネットワークを構築し、第２ネットワークを用いてすべての第３ネットワークの各マイクロユニットを訓練する。訓練を経た第３ネットワークのマイクロユニットに基づいて構築すると、ニューラルネットワークモデルを得ることができる。こうして、すべての第３ネットワークのマイクロユニット機能が第２ネットワークの機能に対応し、従来のマイクロユニットを一つずつ訓練する方法に比べて、訓練数を減らすことができ、計算能力の需要を効果的に下げることができるため、ニューラルネットワークモデルの生成コストを下げることができるとともに、すべての第３ネットワークのマイクロユニットは並行して訓練することができるため、ニューラルネットワークモデルの構築効率を効果的に高めることができる。

【０１１２】

当業者であれば、上記で開示された方法のすべてまたはいくつかのステップ、システムは、ソフトウェア、ファームウェア、ハードウェア、及びそれらの適切な組み合わせとして実施できることを理解できるであろう。いくつかの物理的組立体またはすべての物理的組立体は、中央処理装置、デジタルシグナルプロセッサまたはマイクロプロセッサのようなプロセッサによって実行されるソフトウェアとして、あるいはハードウェアとして、あるいは特定用途向け集積回路のような集積回路として実施することができる。そういったソフトウェアは、コンピュータ可読媒体上に分散することができ、コンピュータ可読媒体はコンピュータ記憶媒体（または非一時的な媒体）及び通信媒体（または一時的な媒体）を含むことができる。コンピュータ記憶媒体という用語は、情報（コンピュータ可読指令、データ構造、プログラムモジュール又は他のデータ）を記憶するための任意の方法または技術において実現される、揮発性及び不揮発性、取り外し可能及び取り外し不可能な媒体を含むことは、当業者にとって周知のことである。コンピュータ記憶媒体は、ＲＡＭ、ＲＯＭ、ＥＥＰＲＯＭ、フラッシュメモリまたは他のメモリ技術、ＣＤ－ＲＯＭ、デジタル多用途ディスク（ＤＶＤ）または他の光ディスク記憶装置、磁気カートリッジ、磁気テープ、磁気ディスク記憶装置または他の磁気記憶装置、または所望の情報を記憶するために使用することができ、コンピュータによってアクセスすることができる任意の他の媒体を含むが、これらに限定されない。さらに、通信媒体は通常、コンピュータ可読指令、デ

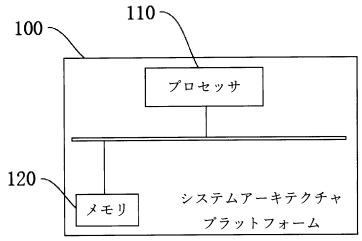
ータ構造、プログラムモジュール、または搬送波または他の伝送メカニズムのような変調データ信号中の他のデータを含み、任意の情報伝送媒体を含むことができることは、当業者にとって周知のことである。

【０１１３】

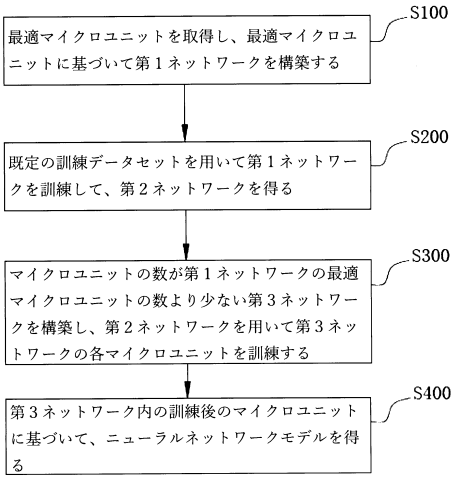
以上では、本願のいくつかの実施形態について具体的に説明したが、本願は上記実施形態に限定されるものではない。当業者であれば、本願の範囲に反することなく、本願の特許請求の範囲に限定された範囲内に含まれる様々な均等的変形又は置換を行うこともできる。

【図面】

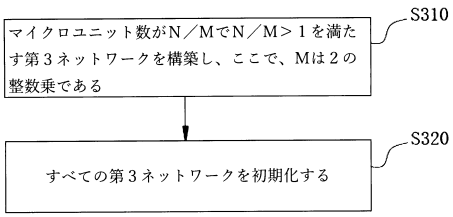
【図１】


100
110
プロセッサ
120
メモリ
システムアーキテクチャ
プラットフォーム

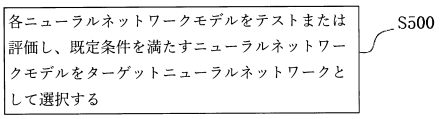
【図２】


最適マイクロユニットを取得し、最適マイクロユニットに基づいて第１ネットワークを構築する
S100
既定の訓練データセットを用いて第１ネットワークを訓練して、第２ネットワークを得る
S200
マイクロユニットの数が第１ネットワークの最適マイクロユニットの数より少ない第３ネットワークを構築し、第２ネットワークを用いて第３ネットワークの各マイクロユニットを訓練する
S300
第３ネットワーク内の訓練後のマイクロユニットに基づいて、ニューラルネットワークモデルを得る
S400

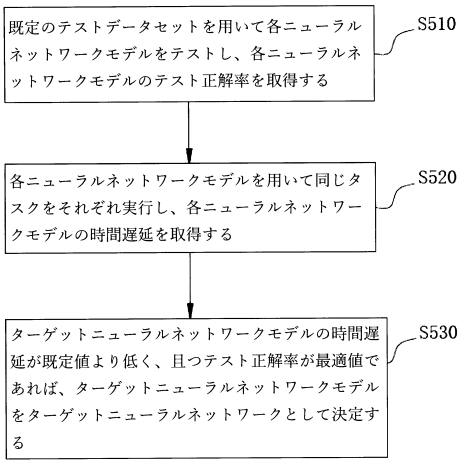
【図３】


マイクロユニット数がN／MでN／M＞１を満たす第３ネットワークを構築し、ここで、Mは２の整数乗である
S310
すべての第３ネットワークを初期化する
S320

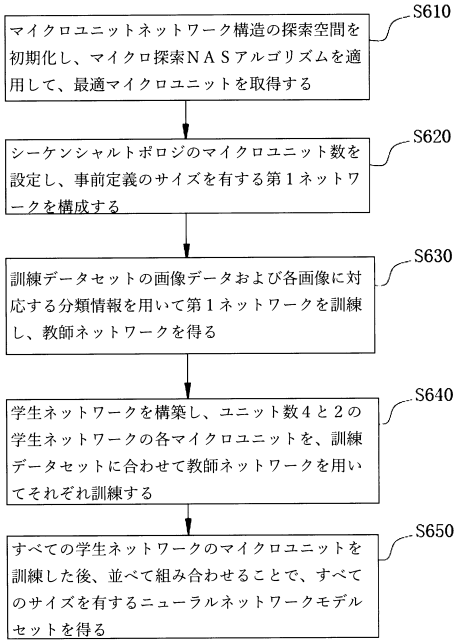
【図４】


各ニューラルネットワークモデルをテストまたは評価し、既定条件を満たすニューラルネットワークモデルをターゲットニューラルネットワークとして選択する
S500

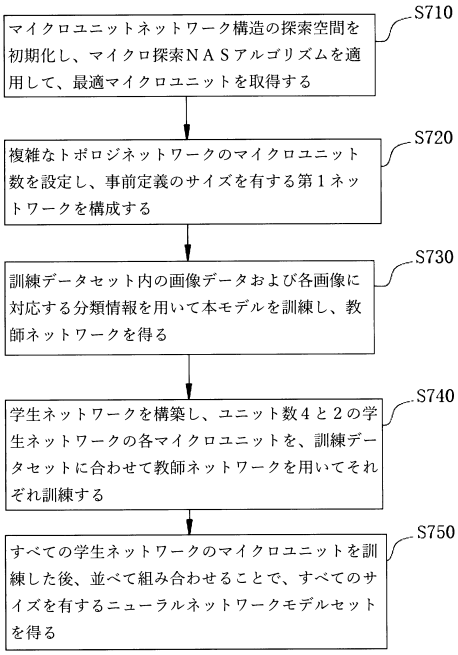
【図５】


既定のテストデータセットを用いて各ニューラルネットワークモデルをテストし、各ニューラルネットワークモデルのテスト正解率を取得する
S510
各ニューラルネットワークモデルを用いて同じタスクをそれぞれ実行し、各ニューラルネットワークモデルの時間遅延を取得する
S520
ターゲットニューラルネットワークモデルの時間遅延が既定値より低く、且つテスト正解率が最適値であれば、ターゲットニューラルネットワークモデルをターゲットニューラルネットワークとして決定する
S530

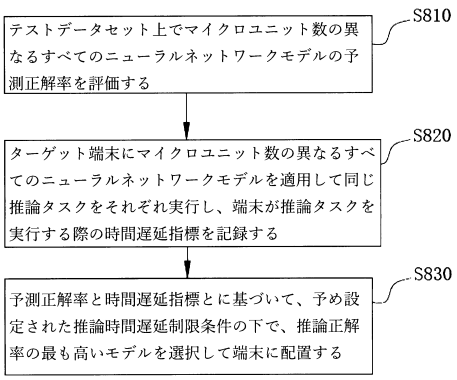
【図６】


マイクロユニットネットワーク構造の探索空間を初期化し、マイクロ探索ＮＡＳアルゴリズムを適用して、最適マイクロユニットを取得する
S610
シーケンシャルトポロジのマイクロユニット数を設定し、事前定義のサイズを有する第１ネットワークを構成する
S620
訓練データセットの画像データおよび各画像に対応する分類情報を用いて第１ネットワークを訓練し、教師ネットワークを得る
S630
学生ネットワークを構築し、ユニット数４と２の学生ネットワークの各マイクロユニットを、訓練データセットに合わせて教師ネットワークを用いてそれぞれ訓練する
S640
すべての学生ネットワークのマイクロユニットを訓練した後、並べて組み合わせることで、すべてのサイズを有するニューラルネットワークモデルセットを得る
S650

【図７】


マイクロユニットネットワーク構造の探索空間を初期化し、マイクロ探索ＮＡＳアルゴリズムを適用して、最適マイクロユニットを取得する
S710
複雑なトポロジネットワークのマイクロユニット数を設定し、事前定義のサイズを有する第１ネットワークを構成する
S720
訓練データセット内の画像データおよび各画像に対応する分類情報を用いて本モデルを訓練し、教師ネットワークを得る
S730
学生ネットワークを構築し、ユニット数４と２の学生ネットワークの各マイクロユニットを、訓練データセットに合わせて教師ネットワークを用いてそれぞれ訓練する
S740
すべての学生ネットワークのマイクロユニットを訓練した後、並べて組み合わせることで、すべてのサイズを有するニューラルネットワークモデルセットを得る
S750

【図８】


テストデータセット上でマイクロユニット数の異なるすべてのニューラルネットワークモデルの予測正解率を評価する
S810
ターゲット端末にマイクロユニット数の異なるすべてのニューラルネットワークモデルを適用して同じ推論タスクをそれぞれ実行し、端末が推論タスクを実行する際の時間遅延指標を記録する
S820
予測正解率と時間遅延指標とに基づいて、予め設定された推論時間遅延制限条件の下で、推論正解率の最も高いモデルを選択して端末に配置する
S830

フロントページの続き

弁理士　河野 英仁
(74)代理人　100078868
弁理士　河野 登夫
(72)発明者　裘 瑞涛
中華人民共和国広東省深▲せん▼市南山区高新技術産業園科技南路中興通訊大厦
(72)発明者　楊 ▲聖▼坤
中華人民共和国広東省深▲せん▼市南山区高新技術産業園科技南路中興通訊大厦
(72)発明者　駱 慶開
中華人民共和国広東省深▲せん▼市南山区高新技術産業園科技南路中興通訊大厦
(72)発明者　韓 炳涛
中華人民共和国広東省深▲せん▼市南山区高新技術産業園科技南路中興通訊大厦
(72)発明者　王 永成
中華人民共和国広東省深▲せん▼市南山区高新技術産業園科技南路中興通訊大厦
(72)発明者　要峰
中華人民共和国広東省深▲せん▼市南山区高新技術産業園科技南路中興通訊大厦

審査官　円子 英紀
(56)参考文献　米国特許出願公開第２０１８／０２６８２９２（ＵＳ，Ａ１）
特開２０２０－０８７２８８（ＪＰ，Ａ）
特開２０２０－０７１８６２（ＪＰ，Ａ）
特表２０１９－５３３２５７（ＪＰ，Ａ）
特表２０１９－５３５０８４（ＪＰ，Ａ）
(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ０６Ｎ　３／００－９９／００
Ｇ０６Ｆ　１８／００－１８／４０