(19)日本国特許庁(JP)　　(12)公表特許公報(A)　　(11)公表番号
特表2024-509670
(P2024-509670A)
(43)公表日　令和6年3月5日(2024.3.5)

(51)国際特許分類		FI		テーマコード（参考）
Ｇ０６Ｎ	*3/08 (2023.01)*	Ｇ０６Ｎ	3/08	５Ｋ０６７
Ｈ０４Ｗ	*28/02 (2009.01)*	Ｈ０４Ｗ	28/02	
Ｇ０６Ｎ	*3/02 (2006.01)*	Ｇ０６Ｎ	3/02	

審査請求　未請求　予備審査請求　未請求　（全43頁）

(21)出願番号　特願2023-544040(P2023-544040)
(86)(22)出願日　令和4年2月3日(2022.2.3)
(85)翻訳文提出日　令和5年8月8日(2023.8.8)
(86)国際出願番号　PCT/EP2022/052633
(87)国際公開番号　WO2022/167547
(87)国際公開日　令和4年8月11日(2022.8.11)
(31)優先権主張番号　21305156.8
(32)優先日　令和3年2月5日(2021.2.5)
(33)優先権主張国・地域又は機関
欧州特許庁(EP)
(81)指定国・地域　AP(BW,GH,GM,KE,LR,LS,MW,MZ,NA,RW,SD,SL,ST,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,RU,TJ,TM),EP(AL,AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,HR,HU,IE,IS,IT,LT,LU,LV,MC,
最終頁に続く

(71)出願人　518341334
インターディジタル・シーイー・パテント・ホールディングス・ソシエテ・パ・アクシオンス・シンプリフィエ
フランス国，７５０１７　パリ，ル　デュ　コロネル　モル　３
(74)代理人　110001243
弁理士法人谷・阿部特許事務所
(74)代理人　110002848
弁理士法人ＮＩＰ＆ＳＢＰＪ国際特許事務所
(72)発明者　クマラスワミ、スレシュ　キルティ
フランス共和国　セソン－セヴィニェ　３５５７６，シーエス　１７６１６，アヴェニュー　デ　シャン　ブラン　９７５，イ
最終頁に続く

(54)【発明の名称】　分割可能なディープニューラルネットワークにおける動的特徴サイズ適応

(57)【要約】
提案される手法は、複数の帯域幅の間で切り替えるための規定を有する分散型ＡＩのための効率的な送信を扱う。エッジデバイスにおける分散推論中、各デバイスは、ＡＩモデルの一部を一度だけロードする必要があるが、それらの間で通信される入力／出力特徴は、動的特徴サイズスイッチ（ＤｙＳｗ）内のノード間の接続を有効化／無効化することによって、利用可能な送信帯域幅に応じて柔軟に構成され得る。所望の圧縮係数を達成するためにいくつかのノードが接続又は切断されるときに、ＤＮＮの他のパラメータは同じままである。すなわち、同じＤＮＮモデルが異なる圧縮係数に使用され、圧縮係数又はネットワーク帯域幅に適合するために新しいＤＮＮモデルをダウンロードする必要はない。
【選択図】図９

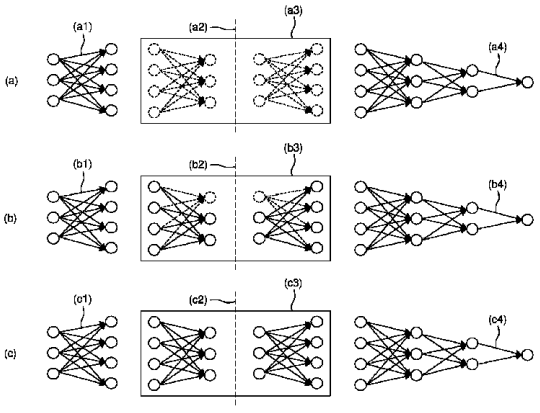


FIG. 9

【特許請求の範囲】
【請求項１】
無線送信／受信ユニット（ＷＴＲＵ）であって、
ディープニューラルネットワーク（ＤＮＮ）モデルの一部であって、前記ＤＮＮモデルの分割点の前にあり、かつ前記ＤＮＮモデルの前記分割点における特徴を圧縮するためのニューラルネットワークを含む、前記ＤＮＮモデルの前記一部を受信するように構成された受信機と、
１つ以上のプロセッサであって、
前記ニューラルネットワークの圧縮係数を取得し、
前記圧縮係数に応答して、前記ニューラルネットワーク内のどのノードが接続されるべきかを決定し、
前記決定に応答して、前記ニューラルネットワークを構成し、かつ
圧縮された特徴を生成するために、前記ＤＮＮモデルの前記一部を用いて推論を実施するように構成されている、１つ以上のプロセッサと、
前記圧縮された特徴を別のＷＴＲＵに送信するように構成された送信機と、を備える、無線送信／受信ユニット（ＷＴＲＵ）。
【請求項２】
前記送信機が、取得された前記圧縮係数の指示を前記別のＷＴＲＵに送るように更に構成されている、請求項１に記載のデバイス。
【請求項３】
無線送信／受信ユニット（ＷＴＲＵ）であって、
ディープニューラルネットワーク（ＤＮＮ）モデルの一部であって、前記ＤＮＮモデルの分割点の後にあり、かつ前記ＤＮＮモデルの前記分割点における特徴を拡張するためのニューラルネットワークを含む、前記ＤＮＮモデルの前記一部を受信するように構成された受信機であって、別のＷＴＲＵから出力された１つ以上の特徴も受信するように構成されている、受信機と、
１つ以上のプロセッサであって、
前記ニューラルネットワークの圧縮係数を取得し、
前記圧縮係数に応答して、前記ニューラルネットワーク内のどのノードが接続されるべきかを決定し、
前記決定に応答して、前記ニューラルネットワークを構成し、かつ
別のＷＴＲＵから出力された前記１つ以上の特徴を前記ニューラルネットワークへの入力として使用して、前記ＤＮＮモデルの前記一部を用いて推論を実施するように構成されている、１つ以上のプロセッサと、を備える、無線送信／受信ユニット（ＷＴＲＵ）。
【請求項４】
前記受信機が、前記圧縮係数を示す信号を受信するように更に構成されている、請求項３に記載のデバイス。
【請求項５】
前記圧縮係数が、推論時に動的に切り替え可能である、複数の圧縮係数から選択される、請求項１～４のいずれか一項に記載のデバイス。
【請求項６】
前記１つ以上のプロセッサが、前記圧縮係数が調整されるときに前記ネットワーク内のどのノードが接続されるべきかを決定するように構成されている、請求項１～５のいずれか一項に記載のデバイス。
【請求項７】
前記１つ以上のプロセッサが、前記ネットワーク内のどのノードが推論時に接続されるべきかを決定するように構成されている、請求項６に記載のデバイス。
【請求項８】
前記ＤＮＮモデルが、複数の分割点を含む、請求項１～７のいずれか一項に記載のデバイス。

【請求項９】
前記ネットワークが、少なくとも畳み込み層及び非線形層を含む、請求項１～８のいずれか一項に記載のデバイス。
【請求項１０】
前記ネットワークが、バッチ正規化層を更に含む、請求項９に記載のデバイス。
【請求項１１】
異なる圧縮係数に対して１つのＤＮＮモデルのみが前記デバイスにロードされる、請求項１～１０のいずれか一項に記載のデバイス。
【請求項１２】
前記分割点及び前記圧縮係数のうちの少なくとも１つが、（１）物理層動作、（２）メディアアクセス制御層動作、（３）無線リソース制御層動作、（４）利用可能な処理リソース、及び（５）制御シグナリングのうちの１つ以上に基づいて適合される、請求項１～１１のいずれか一項に記載のデバイス。
【請求項１３】
前記分割点及び前記圧縮係数のうちの少なくとも１つが、送信データレートに基づいて適合される、請求項１～１２のいずれか一項に記載のデバイス。
【請求項１４】
前記分割点及び前記圧縮係数のうちの少なくとも１つが、前記送信データレートの変化に基づいて適合される、請求項１３に記載のデバイス。
【請求項１５】
無線送信／受信ユニット（ＷＴＲＵ）によって実施される方法であって、
ディープニューラルネットワーク（ＤＮＮ）モデルの一部であって、前記ＤＮＮモデルの分割点の前にあり、かつ前記ＤＮＮモデルの前記分割点における特徴を圧縮するためのニューラルネットワークを含む、前記ＤＮＮモデルの前記一部を受信することと、
前記ニューラルネットワークの圧縮係数を取得することと、
前記圧縮係数に応答して、前記ニューラルネットワーク内のどのノードが接続されるべきかを決定することと、
前記決定に応答して、前記ニューラルネットワークを構成することと、
圧縮された特徴を生成するために、前記ＤＮＮモデルの前記一部を用いて推論を実施することと、
前記圧縮された特徴を別のＷＴＲＵに送信することと、を含む、方法。
【請求項１６】
取得された前記圧縮係数の指示を前記別のＷＴＲＵに送ることを更に含む、請求項１５に記載の方法。
【請求項１７】
方法であって、
ディープニューラルネットワーク（ＤＮＮ）モデルの一部であって、前記ＤＮＮモデルの分割点の後にあり、かつ前記ＤＮＮモデルの前記分割点における特徴を拡張するためのニューラルネットワークを含む、前記ＤＮＮモデルの前記一部を受信することと、
別のＷＴＲＵから出力された１つ以上の特徴を受信することと、
前記ニューラルネットワークの圧縮係数を取得することと、
前記圧縮係数に応答して、前記ニューラルネットワーク内のどのノードが接続されるべきかを決定することと、
前記決定に応答して、前記ニューラルネットワークを構成することと、
別のＷＴＲＵから出力された前記１つ以上の特徴を前記ニューラルネットワークへの入力として使用して、前記ＤＮＮモデルの前記一部を用いて推論を実施することと、を含む、方法。
【請求項１８】
前記圧縮係数を示す信号を受信することを更に含む、請求項１７に記載の方法。
【請求項１９】

前記圧縮係数が、推論時に動的に切り替え可能である、複数の圧縮係数から選択される、請求項１５～１８のいずれか一項に記載の方法。
【請求項２０】
前記ネットワーク内のどのノードが接続されるべきかが、前記圧縮係数が調整されるときに決定される、請求項１５～１９のいずれか一項に記載の方法。
【請求項２１】
前記ネットワーク内のどのノードが接続されるべきかが、推論時に決定される、請求項２０に記載の方法。
【請求項２２】
前記ＤＮＮモデルが、複数の分割点を含む、請求項１５～２１のいずれか一項に記載の方法。
【請求項２３】
前記ネットワークが、少なくとも畳み込み層及び非線形層を含む、請求項１５～２２のいずれか一項に記載の方法。
【請求項２４】
前記ネットワークが、バッチ正規化層を更に含む、請求項２３に記載の方法。
【請求項２５】
異なる圧縮係数に対して１つのＤＮＮモデルのみが前記デバイスにロードされる、請求項１５～２４のいずれか一項に記載の方法。
【請求項２６】
前記分割点及び前記圧縮係数のうちの少なくとも１つが、（１）物理層動作、（２）メディアアクセス制御層動作、（３）無線リソース制御層動作、（４）利用可能な処理リソース、及び（５）制御シグナリングのうちの１つ以上に基づいて適合される、請求項１５～２５のいずれか一項に記載の方法。
【請求項２７】
前記分割点及び前記圧縮係数のうちの少なくとも１つが、送信データレートに基づいて適合される、請求項１５～２６のいずれか一項に記載の方法。
【請求項２８】
前記分割点及び前記圧縮係数のうちの少なくとも１つが、前記送信データレートの変化に基づいて適合される、請求項２７に記載の方法。
【請求項２９】
請求項１５～２８のいずれか一項に記載の方法に従って、ディープニューラルネットワークを適合するための命令を記憶させた、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】
本実施形態は、概して、分割可能なディープニューラルネットワーク（Deep Neural Network、ＤＮＮ）における動的特徴サイズ適応に関する。
【背景技術】
【０００２】
人工知能は、今日の多くの技術分野の重要な機能ブロックである。これは、ディープニューラルネットワーク（ＤＮＮ）の形態のニューラルネットワークの復活によるものである。現代のＤＮＮは、多くの場合、計算集約的であり、したがって、低い処理能力を有するモバイルフォン又は他のエッジデバイス上でＤＮＮ動作を実行することは困難である。これは、多くの場合、モバイルデバイスからクラウドサーバにデータを転送することによって対処され、クラウドサーバにおいて、全ての計算が行われる。
【発明の概要】
【０００３】
一実施形態によれば、デバイスが提示されており、デバイスは、無線送信／受信ユニット（Wireless Transmit/Receive Unit、ＷＴＲＵ）を備え、無線送信／受信ユニッ

ト（ＷＴＲＵ）は、ディープニューラルネットワーク（ＤＮＮ）モデルの一部であって、当該ＤＮＮモデルの分割点の前にあり、かつ当該ＤＮＮモデルの当該分割点における特徴を圧縮するためのニューラルネットワークを含む、当該ＤＮＮモデルの当該一部を受信するように構成された受信機と、１つ以上のプロセッサであって、当該ニューラルネットワークの圧縮係数を取得し、当該圧縮係数に応答して、当該ニューラルネットワーク内のどのノードが接続されるべきかを決定し、当該決定に応答して、当該ニューラルネットワークを構成し、かつ圧縮された特徴を生成するために、当該ＤＮＮモデルの当該一部を用いて推論を実施するように構成されている、１つ以上のプロセッサと、当該圧縮された特徴を別のＷＴＲＵに送信するように構成された送信機と、を備える。

【０００４】

別の実施形態によれば、デバイスが提示されており、デバイスは、無線送信／受信ユニット（ＷＴＲＵ）を備え、無線送信／受信ユニット（ＷＴＲＵ）は、ディープニューラルネットワーク（ＤＮＮ）モデルの一部であって、当該ＤＮＮモデルの分割点の後にあり、かつ当該ＤＮＮモデルの当該分割点における特徴を拡張するためのニューラルネットワークを含む、当該ＤＮＮモデルの当該一部を受信するように構成された受信機であって、別のＷＴＲＵから出力された１つ以上の特徴も受信するように構成されている、受信機と、１つ以上のプロセッサであって、当該ニューラルネットワークの圧縮係数を取得し、当該圧縮係数に応答して、当該ニューラルネットワーク内のどのノードが接続されるべきかを決定し、当該決定に応答して、当該ニューラルネットワークを構成し、かつ別のＷＴＲＵから出力された当該１つ以上の特徴を当該ニューラルネットワークへの入力として使用して、当該ＤＮＮモデルの当該一部を用いて推論を実施するように構成されている、１つ以上のプロセッサと、を備える。

【０００５】

別の実施形態によれば、方法が提示されており、方法は、無線送信／受信ユニット（ＷＴＲＵ）によって実施される方法を含み、方法は、ディープニューラルネットワーク（ＤＮＮ）モデルの一部であって、当該ＤＮＮモデルの分割点の前にあり、かつ当該ＤＮＮモデルの当該分割点における特徴を圧縮するためのニューラルネットワークを含む、当該ＤＮＮモデルの当該一部を受信することと、当該ニューラルネットワークの圧縮係数を取得することと、当該圧縮係数に応答して、当該ニューラルネットワーク内のどのノードが接続されるべきかを決定することと、当該決定に応答して、当該ニューラルネットワークを構成することと、圧縮された特徴を生成するために、当該ＤＮＮモデルの当該一部を用いて推論を実施することと、当該圧縮された特徴を別のＷＴＲＵに送信することと、を含む。

【０００６】

別の実施形態によれば、方法が提示されており、方法は、ディープニューラルネットワーク（ＤＮＮ）モデルの一部であって、当該ＤＮＮモデルの分割点の後にあり、かつ当該ＤＮＮモデルの当該分割点における特徴を拡張するためのニューラルネットワークを含む、当該ＤＮＮモデルの当該一部を受信することと、別のＷＴＲＵから出力された１つ以上の特徴を受信することと、当該ニューラルネットワークの圧縮係数を取得することと、当該圧縮係数に応答して、当該ニューラルネットワーク内のどのノードが接続されるべきかを決定することと、当該決定に応答して、当該ニューラルネットワークを構成することと、別のＷＴＲＵから出力された当該１つ以上の特徴を当該ニューラルネットワークへの入力として使用して、当該ＤＮＮモデルの当該一部を用いて推論を実施することと、を含む。

【０００７】

更なる実施形態は、本明細書に記載の方法を実施するように構成されたシステムを含む。そのようなシステムは、プロセッサと、プロセッサ上で実行されたときに本明細書で説明される方法を実施するように動作する命令を記憶する、非一時的コンピュータ記憶媒体と、を含み得る。

【図面の簡単な説明】

【０００８】
【図１Ａ】１つ以上の開示される実施形態が実装され得る例示的な通信システムを示すシステム図である。
【図１Ｂ】一実施形態による、図１Ａに示される通信システム内で使用され得る例示的な無線送信／受信ユニット（ＷＴＲＵ）を示すシステム図である。
【図２】特徴サイズ圧縮を伴わない２つのデバイス間の分散ＡＩのためのメカニズムを示す。
【図３Ａ】それぞれ、特徴圧縮のための１つ、２つ、及び３つの候補分割を有するＤＮＮを示す。
【図３Ｂ】それぞれ、特徴圧縮のための１つ、２つ、及び３つの候補分割を有するＤＮＮを示す。
【図３Ｃ】それぞれ、特徴圧縮のための１つ、２つ、及び３つの候補分割を有するＤＮＮを示す。
【図４】特徴圧縮のための単一の分割を有するＤＮＮを示す。
【図５Ａ】単一の圧縮係数がサポートされる、帯域幅縮小器（bandwidth-reducer、ＢＷＲ）及び帯域幅拡張器（bandwidth-expander、ＢＷＥ）を使用する、２つのデバイス、デバイス１とデバイス２との間の分散ＡＩのための特徴サイズ圧縮メカニズムを示す。
【図５Ｂ】複数の圧縮係数がサポートされる、特徴サイズ圧縮メカニズムを示す。
【図６Ａ】ＢＷＲ及びＢＷＥを伴わない総推論レイテンシを示す。
【図６Ｂ】ＢＷＲ及びＢＷＲを伴う総推論レイテンシを示し、中間データのサイズが低減され得る。
【図７】一実施形態による、分割と圧縮係数（compression factor、ＣＦ）構成との間で動的に切り替えるプロセスを示す。
【図８Ａ】デバイス１及び２がそれらの計算能力及び送信チャネルを推定することを示す。
【図８Ｂ】デバイスの各々からのＡＩ／ＭＬモデルの受信を示す。
【図８Ｃ】デバイスの推論時間動作を示す。
【図９】一実施形態による、適応特徴圧縮のためのＤＮＮにおける単一分割を伴う方法を示す。
【図１０】サイズ４の入力を縮小及び拡張することができる例示的なＤｙＳｗを示す。
【図１１】図９に示したＤｙＳｗ構成の接続を示す。
【発明を実施するための形態】
【０００９】
図１Ａは、１つ以上の開示された実施形態が実装され得る、例示的な通信システム１００を示す図である。通信システム１００は、音声、データ、ビデオ、メッセージ伝達、ブロードキャストなどのコンテンツを、複数の無線ユーザに提供する、多重アクセスシステムであり得る。通信システム１００は、複数の無線ユーザが、無線帯域幅を含むシステムリソースの共有を通じて、上記のようなコンテンツにアクセスすることを可能にし得る。例えば、通信システム１００は、コード分割多重アクセス（code division multiple access、ＣＤＭＡ）、時分割多重アクセス（time division multiple access、ＴＤＭＡ）、周波数分割多重アクセス（frequency division multiple access、ＦＤＭＡ）、直交ＦＤＭＡ（orthogonal FDMA、ＯＦＤＭＡ）、シングルキャリアＦＤＭＡ（single-carrier FDMA、ＳＣ－ＦＤＭＡ）、ゼロテールユニークワードＤＦＴ－Ｓｐｒｅａｄ　ＯＦＤＭ（zero-tail unique-word DFT-Spread OFDM、ＺＴ　ＵＷ　ＤＴＳ－ｓ　ＯＦＤＭ）、ユニークワードＯＦＤＭ（unique word OFDM、ＵＷ－ＯＦＤＭ）、リソースブロックフィルタ処理ＯＦＤＭ、フィルタバンクマルチキャリア（filter bank multicarrier、ＦＢＭＣ）などの、１つ以上のチャネルアクセス方法を用い得る。
【００１０】

図１Ａに示すように、通信システム１００は、無線送／受信ユニット（ＷＴＲＵ）１０２ａ、１０２ｂ、１０２ｃ、１０２ｄ、ＲＡＮ１０４、ＣＮ１０６、公衆交換電話網（public switched telephone network、ＰＳＴＮ）１０８、インターネット１１０、及び他のネットワーク１１２を含み得るが、開示された実施形態は、任意の数のＷＴＲＵ、基地局、ネットワーク、及び／又はネットワーク要素を想定していることが理解されるであろう。ＷＴＲＵ１０２ａ、１０２ｂ、１０２ｃ、１０２ｄの各々は、無線環境において動作し、かつ／又は通信するように構成された、任意のタイプのデバイスであり得る。例として、それらのいずれも「局」及び／又は「ＳＴＡ」と称され得るＷＴＲＵ１０２ａ、１０２ｂ、１０２ｃ、１０２ｄは、無線信号を送信及び／又は受信するように構成され得、ユーザ機器（user equipment、ＵＥ）、移動局、固定又は移動加入者ユニット、加入ベースのユニット、ページャ、セルラ電話、携帯情報端末（personal digital assistant、ＰＤＡ）、スマートフォン、ラップトップ、ネットブック、パーソナルコンピュータ、無線センサ、ホットスポット又はＭｉ－Ｆｉデバイス、モノのインターネット（Internet of Things、ＩｏＴ）デバイス、ウォッチ又は他のウェアラブル、ヘッドマウントディスプレイ（head-mounted display、ＨＭＤ）、車両、ドローン、医療デバイス及びアプリケーション（例えば、遠隔手術）、工業用デバイス及びアプリケーション（例えば、工業用及び／又は自動処理チェーンコンテキストで動作するロボット及び／又は他の無線デバイス）、家電デバイス、商業用及び／又は工業用無線ネットワークで動作するデバイスなどを含み得る。ＷＴＲＵ１０２ａ、１０２ｂ、１０２ｃ、及び１０２ｄのいずれも、互換的にＵＥと称され得る。

【００１１】

通信システム１００はまた、基地局１１４ａ及び／又は基地局１１４ｂを含み得る。基地局１１４ａ、１１４ｂの各々は、ＣＮ１０６、インターネット１１０、及び／又は他のネットワーク１１２などの１つ以上の通信ネットワークへのアクセスを容易にするために、ＷＴＲＵ１０２ａ、１０２ｂ、１０２ｃ、１０２ｄのうちの少なくとも１つと無線でインターフェース接続するように構成された任意のタイプのデバイスであり得る。例として、基地局１１４ａ、１１４ｂは、基地局トランシーバ（base transceiver station、ＢＴＳ）、ノードＢ、ｅＮｏｄｅＢ、ホームノードＢ、ホームｅＮｏｄｅＢ、ｇＮＢ、ＮＲ　ＮｏｄｅＢ、サイトコントローラ、アクセスポイント（access point、ＡＰ）、無線ルータなどであり得る。基地局１１４ａ、１１４ｂは各々単一の要素として示されているが、基地局１１４ａ、１１４ｂは、任意の数の相互接続された基地局及び／又はネットワーク要素を含み得ることが理解されるであろう。

【００１２】

基地局１１４ａは、基地局コントローラ（base station controller、ＢＳＣ）、無線ネットワークコントローラ（radio network controller、ＲＮＣ）、リレーノードなど、他の基地局及び／又はネットワーク要素（図示せず）も含み得る、ＲＡＮ１０４の一部であり得る。基地局１１４ａ及び／又は基地局１１４ｂは、セル（図示せず）と称され得る、１つ以上のキャリア周波数で無線信号を送信及び／又は受信するように構成され得る。これらの周波数は、認可スペクトル、未認可スペクトル、又は認可及び未認可スペクトルの組み合わせであり得る。セルは、相対的に固定され得るか又は経時的に変化し得る特定の地理的エリアに、無線サービスのカバレッジを提供し得る。セルは、更にセルセクタに分割され得る。例えば、基地局１１４ａと関連付けられたセルは、３つのセクタに分割され得る。したがって、一実施形態では、基地局１１４ａは、３つのトランシーバを、すなわち、セルのセクタごとに１つのトランシーバを含み得る。一実施形態では、基地局１１４ａは、多重入力多重出力（multiple-input multiple output、ＭＩＭＯ）技術を用い得、セルのセクタごとに複数のトランシーバを利用し得る。例えば、ビームフォーミングを使用して、所望の空間方向に信号を送信及び／又は受信し得る。

【００１３】

基地局１１４ａ、１１４ｂは、エアインターフェース１１６を介して、ＷＴＲＵ１０２ａ、１０２ｂ、１０２ｃ、１０２ｄのうちの１つ以上と通信し得るが、このエアインター

フェース１１６は、任意の好適な無線通信リンク（例えば、無線周波数（radio frequency、ＲＦ）、マイクロ波、センチメートル波、マイクロメートル波、赤外線（infrared、ＩＲ）、紫外線（ultraviolet、ＵＶ）、可視光など）であり得る。エアインターフェース１１６は、任意の好適な無線アクセス技術（radio access technology、ＲＡＴ）を使用して確立され得る。

【００１４】

より具体的には、上記のように、通信システム１００は、多重アクセスシステムであり得、例えば、ＣＤＭＡ、ＴＤＭＡ、ＦＤＭＡ、ＯＦＤＭＡ、ＳＣ－ＦＤＭＡなどの、１つ以上のチャネルアクセススキームを用い得る。例えば、ＲＡＮ１０４及びＷＴＲＵ１０２ａ、１０２ｂ、１０２ｃの基地局１１４ａは、広帯域ＣＤＭＡ（wideband CDMA、ＷＣＤＭＡ）を使用してエアインターフェース１１６を確立し得る、ユニバーサル移動体通信システム（Universal Mobile Telecommunications System、ＵＭＴＳ）地上無線アクセス（Terrestrial Radio Access、ＵＴＲＡ）などの無線技術を実装し得る。ＷＣＤＭＡは、高速パケットアクセス（High-Speed Packet Access、ＨＳＰＡ）及び／又は進化型ＨＳＰＡ（ＨＳＰＡ＋）などの通信プロトコルを含み得る。ＨＳＰＡは、高速ダウンリンク（Downlink、ＤＬ）パケットアクセス（High-Speed Downlink Packet Access、ＨＳＤＰＡ）及び／又は高速アップリンクパケットアクセス（High-Speed UL Packet Access、ＨＳＵＰＡ）を含み得る。

【００１５】

一実施形態では、基地局１１４ａ及びＷＴＲＵ１０２ａ、１０２ｂ、１０２ｃは、進化型ＵＭＴＳ地上無線アクセス（Evolved UMTS Terrestrial Radio Access、Ｅ－ＵＴＲＡ）などの無線技術を実装し得、これは、ロングタームエボリューション（Long Term Evolution、ＬＴＥ）及び／又はＬＴＥ－Ａｄｖａｎｃｅｄ（LTE-Advanced、ＬＴＥ－Ａ）及び／又はＬＴＥ－Ａｄｖａｎｃｅｄ　Ｐｒｏ（ＬＴＥ－Ａ　Ｐｒｏ）を使用してエアインターフェース１１６を確立し得る。

【００１６】

一実施形態では、基地局１１４ａ及びＷＴＲＵ１０２ａ、１０２ｂ、１０２ｃは、ＮＲ無線アクセスなどの無線技術を実装することができ、この技術は、新しい無線（New Radio、ＮＲ）を使用してエアインターフェース１１６を確立することができる。

【００１７】

一実施形態では、基地局１１４ａ及びＷＴＲＵ１０２ａ、１０２ｂ、１０２ｃは、複数の無線アクセス技術を実装し得る。例えば、基地局１１４ａ及びＷＴＲＵ１０２ａ、１０２ｂ、１０２ｃは、例えば、デュアルコネクティビティ（dual connectivity、ＤＣ）原理を使用して、ＬＴＥ無線アクセス及びＮＲ無線アクセスを一緒に実装し得る。したがって、ＷＴＲＵ１０２ａ、１０２ｂ、１０２ｃによって利用されるエアインターフェースは、複数のタイプの無線アクセス技術、及び／又は複数のタイプの基地局（例えば、ｅＮＢ及びｇＮＢ）に送られる／そこから送られる送信によって特徴付けられ得る。

【００１８】

他の実施形態では、基地局１１４ａ及びＷＴＲＵ１０２ａ、１０２ｂ、１０２ｃは、ＩＥＥＥ８０２．１１（すなわち、無線フィデリティ（Wireless Fidelity、ＷｉＦｉ）、ＩＥＥＥ８０２．１６（すなわち、ワイマックス（Worldwide Interoperability for Microwave Access、ＷｉＭＡＸ）、ＣＤＭＡ２０００、ＣＤＭＡ２０００　１Ｘ、ＣＤＭＡ２０００　ＥＶ－ＤＯ、暫定規格２０００（ＩＳ－２０００）、暫定規格９５（ＩＳ－９５）、暫定規格８５６（ＩＳ－８５６）、汎欧州デジタル移動電話方式（Global System for Mobile communications、ＧＳＭ）、ＧＳＭ進化型高速データレート（Enhanced Data rates for GSM Evolution、ＥＤＧＥ）、ＧＳＭ　ＥＤＧＥ（ＧＥＲＡＮ）などの無線技術を実装し得る。

【００１９】

図１Ａの基地局１１４ｂは、例えば、無線ルータ、ホームノードＢ、ホームｅノードＢ又はアクセスポイントであり得、事業所、家庭、車両、キャンパス、工業施設、（例えば

、ドローンによる使用のための）空中回廊、道路などの場所などの局所的エリアにおける無線接続を容易にするために、任意の好適なＲＡＴを利用し得る。一実施形態では、基地局１１４ｂ及びＷＴＲＵ１０２ｃ、１０２ｄは、ＩＥＥＥ８０２．１１などの無線技術を実装して、無線ローカルエリアネットワーク（wireless local area network、ＷＬＡＮ）を確立し得る。一実施形態では、基地局１１４ｂ及びＷＴＲＵ１０２ｃ、１０２ｄは、ＩＥＥＥ８０２．１５などの無線技術を実装して、無線パーソナルエリアネットワーク（wireless personal area network、ＷＰＡＮ）を確立し得る。更に別の一実施形態では、基地局１１４ｂ及びＷＴＲＵ１０２ｃ、１０２ｄは、セルラベースのＲＡＴ（例えば、ＷＣＤＭＡ、ＣＤＭＡ２０００、ＧＳＭ、ＬＴＥ、ＬＴＥ－Ａ、ＬＴＥ－Ａ Ｐｒｏ、ＮＲなど）を利用して、ピコセル又はフェムトセルを確立し得る。図１Ａに示すように、基地局１１４ｂは、インターネット１１０への直接接続を有し得る。したがって、基地局１１４ｂは、ＣＮ１０６を介してインターネット１１０にアクセスする必要がない場合がある。

【００２０】

ＲＡＮ１０４は、ＷＴＲＵ１０２ａ、１０２ｂ、１０２ｃ、１０２ｄのうちの１つ以上に、音声、データ、アプリケーション、及び／又はボイスオーバインターネットプロトコル（voice over internet protocol、ＶｏＩＰ）サービスを提供するように構成された任意のタイプのネットワークであり得る、ＣＮ１０６と通信し得る。データは、例えば、異なるスループット要件、待ち時間要件、エラー許容要件、信頼性要件、データスループット要件、モビリティ要件などの、様々なサービス品質（quality of service、ＱｏＳ）要件を有し得る。ＣＮ１０６は、通話制御、ビリングサービス、モバイルロケーションベースのサービス、プリペイド通話、インターネット接続性、映像配信などを提供し、かつ／又はユーザ認証などの高レベルセキュリティ機能を行い得る。図１Ａには示されていないが、ＲＡＮ１０４及び／又はＣＮ１０６は、ＲＡＮ１０４と同じＲＡＴ又は異なるＲＡＴを用いる他のＲＡＮと直接又は間接的に通信し得ることが理解されよう。例えば、ＮＲ無線技術を利用し得るＲＡＮ１０４に接続されることに加えて、ＣＮ１０６はまた、ＧＳＭ、ＵＭＴＳ、ＣＤＭＡ２０００、ＷｉＭＡＸ、Ｅ－ＵＴＲＡ又はＷｉＦｉ無線技術を用いて別のＲＡＮ（図示せず）と通信し得る。

【００２１】

ＣＮ１０６はまた、ＰＳＴＮ１０８、インターネット１１０、及び／又は他のネットワーク１１２にアクセスするために、ＷＴＲＵ１０２ａ、１０２ｂ、１０２ｃ、１０２ｄのゲートウェイとして機能し得る。ＰＳＴＮ１０８は、基本電話サービス（plain old telephone service、ＰＯＴＳ）を提供する公衆交換電話網を含み得る。インターネット１１０は、相互接続されたコンピュータネットワーク及びデバイスのグローバルシステムを含み得るが、これらのネットワーク及びデバイスは、送信制御プロトコル（transmission control protocol、ＴＣＰ）、ユーザデータグラムプロトコル（user datagram protocol、ＵＤＰ）、及び／又はＴＣＰ／ＩＰインターネットプロトコルスイートのインターネットプロトコル（internet protocol、ＩＰ）などの、共通通信プロトコルを使用する。ネットワーク１１２は、他のサービスプロバイダによって所有及び／又は運営される、有線及び／又は無線通信ネットワークを含み得る。例えば、ネットワーク１１２は、ＲＡＮ１０４と同じＲＡＴ又は異なるＲＡＴを用い得る１つ以上のＲＡＮに接続された別のＣＮを含み得る。

【００２２】

通信システム１００におけるＷＴＲＵ１０２ａ、１０２ｂ、１０２ｃ、１０２ｄのいくつか又は全ては、マルチモード能力を含み得る（例えば、ＷＴＲＵ１０２ａ、１０２ｂ、１０２ｃ、１０２ｄは、異なる無線リンクを介して異なる無線ネットワークと通信するための複数のトランシーバを含み得る）。例えば、図１Ａに示されるＷＴＲＵ１０２ｃは、セルラベースの無線技術を用い得る基地局１１４ａ、及びＩＥＥＥ８０２無線技術を用い得る基地局１１４ｂと通信するように構成され得る。

【００２３】

図１Ｂは、例示的なＷＴＲＵ１０２を示すシステム図である。図１Ｂに示すように、ＷＴＲＵ１０２は、とりわけ、プロセッサ１１８、トランシーバ１２０、送信／受信要素１２２、スピーカ／マイクロフォン１２４、キーパッド１２６、ディスプレイ／タッチパッド１２８、非リムーバブルメモリ１３０、リムーバブルメモリ１３２、電源１３４、全地球測位システム（global positioning system、ＧＰＳ）チップセット１３６、及び／又は他の周辺機器１３８を含み得る。ＷＴＲＵ１０２は、一実施形態との一貫性を有したまま、前述の要素の任意の部分的組み合わせを含み得ることが理解されよう。

【００２４】

プロセッサ１１８は、汎用プロセッサ、専用プロセッサ、従来のプロセッサ、デジタル信号プロセッサ（digital signal processor、ＤＳＰ）、複数のマイクロプロセッサ、ＤＳＰコアと関連付けられた１つ以上のマイクロプロセッサ、コントローラ、マイクロコントローラ、特定用途向け集積回路（Application Specific Integrated Circuit、ＡＳＩＣ）、フィールドプログラマブルゲートアレイ（Field Programmable Gate Array、ＦＰＧＡ）回路、任意の他のタイプの集積回路（integrated circuit、ＩＣ）、状態機械などであり得る。プロセッサ１１８は、信号コーディング、データ処理、電力制御、入力／出力処理、及び／又はＷＴＲＵ１０２が無線環境で動作することを可能にする任意の他の機能性を実行し得る。プロセッサ１１８は、送信／受信要素１２２に結合され得るトランシーバ１２０に結合され得る。図１Ｂは、プロセッサ１１８及びトランシーバ１２０を別個の構成要素として示すが、プロセッサ１１８及びトランシーバ１２０は、電子パッケージ又はチップにおいて一緒に統合され得るということが理解されよう。

【００２５】

送信／受信要素１２２は、エアインターフェース１１６を介して基地局（例えば、基地局１１４ａ）に信号を送信するか又は基地局（例えば、基地局１１４ａ）から信号を受信するように構成され得る。例えば、一実施形態では、送信／受信要素１２２は、ＲＦ信号を送信及び／又は受信するように構成されたアンテナであり得る。一実施形態では、送信／受信要素１２２は、例えば、ＩＲ、ＵＶ又は可視光信号を送信及び／又は受信するように構成されたエミッタ／検出器であり得る。更に別の実施形態では、送信／受信要素１２２は、ＲＦ信号及び光信号の両方を送信及び／又は受信するように構成され得る。送信／受信要素１２２は、無線信号の任意の組み合わせを送信及び／又は受信するように構成され得るということが理解されよう。

【００２６】

送信／受信要素１２２は、単一の要素として図１Ｂに示されているが、ＷＴＲＵ１０２は、任意の数の送信／受信要素１２２を含み得る。より具体的には、ＷＴＲＵ１０２は、ＭＩＭＯ技術を用い得る。したがって、一実施形態では、ＷＴＲＵ１０２は、エアインターフェース１１６を介して無線信号を送信及び受信するための２つ以上の送信／受信要素１２２（例えば、複数のアンテナ）を含み得る。

【００２７】

トランシーバ１２０は、送信／受信要素１２２によって送信される信号を変調し、送信／受信要素１２２によって受信される信号を復調するように構成され得る。上記のように、ＷＴＲＵ１０２は、マルチモード能力を有し得る。したがって、トランシーバ１２０は、例えばＮＲ及びＩＥＥＥ８０２．１１などの複数のＲＡＴを介してＷＴＲＵ１０２が通信することを可能にするための複数のトランシーバを含み得る。

【００２８】

ＷＴＲＵ１０２のプロセッサ１１８は、スピーカ／マイクロフォン１２４、キーパッド１２６、及び／又はディスプレイ／タッチパッド１２８（例えば、液晶ディスプレイ（liquid crystal display、ＬＣＤ）表示ユニット若しくは有機発光ダイオード（organic light-emitting diode、ＯＬＥＤ）表示ユニット）に結合され得、これらからユーザが入力したデータを受信し得る。プロセッサ１１８はまた、ユーザデータをスピーカ／マイクロフォン１２４、キーパッド１２６、及び／又はディスプレイ／タッチパッド１２８に出力し得る。加えて、プロセッサ１１８は、非リムーバブルメモリ１３０及び／又はリ

ムーバブルメモリ１３２などの任意のタイプの好適なメモリから情報にアクセスし、かつメモリにデータを記憶し得る。非リムーバブルメモリ１３０は、ランダムアクセスメモリ（random-access memory、ＲＡＭ）、読み取り専用メモリ（read-only memory、ＲＯＭ）、ハードディスク又は任意の他のタイプのメモリ記憶デバイスを含み得る。リムーバブルメモリ１３２は、加入者識別モジュール（subscriber identity module、ＳＩＭ）カード、メモリスティック、セキュアデジタル（secure digital、ＳＤ）メモリカードなどを含み得る。他の実施形態では、プロセッサ１１８は、サーバ又はホームコンピュータ（図示せず）上など、ＷＴＲＵ１０２上に物理的に配置されていないメモリから情報にアクセスし、かつメモリにデータを記憶し得る。

【００２９】

プロセッサ１１８は、電源１３４から電力を受信し得るが、ＷＴＲＵ１０２における他の構成要素に電力を分配し、かつ／又は制御するように構成され得る。電源１３４は、ＷＴＲＵ１０２に電力を供給するための任意の好適なデバイスであり得る。例えば、電源１３４は、１つ以上の乾電池（例えば、ニッケルカドミウム（nickel-cadmium、ＮｉＣｄ）、ニッケル亜鉛（nickel-zinc、ＮｉＺｎ）、ニッケル金属水素化物（nickel metal hydride、ＮｉＭＨ）、リチウムイオン（lithium-ion、Ｌｉ－ｉｏｎ）など）、太陽セル、燃料セルなどを含み得る。

【００３０】

プロセッサ１１８はまた、ＧＰＳチップセット１３６に結合され得、これは、ＷＴＲＵ１０２の現在の場所に関する場所情報（例えば、経度及び緯度）を提供するように構成され得る。ＧＰＳチップセット１３６からの情報に加えて又はその代わりに、ＷＴＲＵ１０２は、基地局（例えば、基地局１１４ａ、１１４ｂ）からエアインターフェース１１６を介して場所情報を受信し、かつ／又は２つ以上の近くの基地局から受信されている信号のタイミングに基づいて、その場所を判定し得る。ＷＴＲＵ１０２は、一実施形態との一貫性を有したまま、任意の好適な位置判定方法によって位置情報を取得し得るということが理解されよう。

【００３１】

プロセッサ１１８は、他の周辺機器１３８に更に結合され得、他の周辺機器１３８には、追加の特徴、機能、及び／又は有線若しくは無線接続を提供する１つ以上のソフトウェア及び／又はハードウェアモジュールが含まれ得る。例えば、周辺機器１３８には、加速度計、電子コンパス、衛星トランシーバ、（写真及び／又はビデオのための）デジタルカメラ、ユニバーサルシリアルバス（universal serial bus、ＵＳＢ）ポート、振動デバイス、テレビトランシーバ、ハンズフリーヘッドセット、Ｂｌｕｅｔｏｏｔｈ（登録商標）モジュール、周波数変調（frequency modulated、ＦＭ）無線ユニット、デジタル音楽プレーヤ、メディアプレーヤ、ビデオゲームプレーヤモジュール、インターネットブラウザ、仮想現実及び／又は拡張現実（Virtual Reality/Augmented Reality、ＶＲ／ＡＲ）デバイス、アクティビティトラッカなどが含まれ得る。周辺機器１３８は、１つ以上のセンサを含み得、センサは、ジャイロスコープ、加速度計、ホール効果センサ、磁力計、方位センサ、近接センサ、温度センサ、時間センサ、ジオロケーションセンサ、高度計、光センサ、タッチセンサ、磁力計、気圧計、ジェスチャセンサ、生体認証センサ、及び／又は湿度センサのうちの１つ以上であり得る。

【００３２】

ＷＴＲＵ１０２は、（例えば、ＵＬ（例えば、送信用）及びダウンリンク（例えば、受信用）の両方のための特定のサブフレームと関連付けられた）信号のいくつか又は全ての送信及び受信が並列及び／又は同時であり得る、全二重無線機を含み得る。全二重無線機は、ハードウェア（例えば、チョーク）又はプロセッサを介した信号処理（例えば、別個のプロセッサ（図示せず）又はプロセッサ１１８を介して）を介して自己干渉を低減し、かつ又は実質的に排除するための干渉管理ユニットを含み得る。一実施形態では、ＷＲＴＵ１０２は、（例えば、ＵＬ（例えば、送信用）又はダウンリンク（例えば、受信用）のいずれかのための特定のサブフレームと関連付けられた）信号のうちのいくつか又は全て

のうちのどれかの送信及び受信のための半二重無線機を含み得る。
【００３３】
ＷＴＲＵは、図１Ａ～図１Ｂに、無線端末装置として説明されているが、特定の代表的な実施形態では、そのような端末装置は、通信ネットワークとの有線通信インターフェースを（例えば、一時的に又は恒久的に）使用し得ることが想定される。
【００３４】
図１Ａ～図１Ｂ、及び図１Ａ～図１Ｂの対応する説明を考慮して、ＷＴＲＵ１０２ａ～ｄ、基地局１１４ａ～ｂ、ｅＮｏｄｅ－Ｂ１６０ａ～ｃ、ＭＭＥ１６２、ＳＧＷ１６４、ＰＧＷ１６６、ｇＮＢ１８０ａ～ｃ、ＡＭＦ１８２ａ～ｂ、ＵＰＦ１８４ａ～ｂ、ＳＭＦ１８３ａ～ｂ、ＤＮ１８５ａ～ｂ、及び／又は本明細書に記載される任意の他のデバイスのうちの１つ以上に関して本明細書に記載される機能のうちの１つ以上又は全部は、１つ以上のエミュレーションデバイス（図示せず）によって実行され得る（図示せず）。エミュレーションデバイスは、本明細書に説明される機能の１つ以上又は全てをエミュレートするように構成された１つ以上のデバイスであり得る。例えば、エミュレーションデバイスを使用して、他のデバイスを試験し、かつ／又はネットワーク及び／若しくはＷＴＲＵ機能をシミュレートし得る。
【００３５】
エミュレーションデバイスは、ラボ環境及び／又はオペレータネットワーク環境における他のデバイスの１つ以上の試験を実装するように設計され得る。例えば、１つ以上のエミュレーションデバイスは、通信ネットワーク内の他のデバイスを試験するために、有線及び／又は無線通信ネットワークの一部として完全に若しくは部分的に実装され、かつ／又は展開されている間、１つ以上若しくは全ての機能を実行し得る。１つ以上のエミュレーションデバイスは、有線及び／又は無線通信ネットワークの一部として一時的に実装／展開されている間、１つ以上若しくは全ての機能を実行し得る。エミュレーションデバイスは、試験を目的として別のデバイスに直接結合され得、かつ／又は地上波無線通信を使用して試験を実行し得る。
【００３６】
１つ以上のエミュレーションデバイスは、有線及び／又は無線通信ネットワークの一部として実装／展開されていない間、全てを含む１つ以上の機能を実行し得る。例えば、エミュレーションデバイスは、１つ以上の構成要素の試験を実装するために、試験実験室での試験シナリオ、並びに／又は展開されていない（例えば、試験用の）有線及び／若しくは無線通信ネットワークにおいて利用され得る。１つ以上のエミュレーションデバイスは、試験機器であり得る。ＲＦ回路（例えば、１つ以上のアンテナを含み得る）を介した直接ＲＦ結合及び／又は無線通信は、データを送信及び／又は受信するように、エミュレーションデバイスによって使用され得る。
【００３７】
上述したように、ＤＮＮ動作の実行は、多くの場合、モバイルデバイスからクラウドサーバにデータを転送することによって対処され、クラウドサーバにおいて、全ての計算が行われる。しかしながら、これは帯域幅を必要とし、（送信レイテンシに起因して）時間集約的であり、データプライバシーの懸念を引き起こす。これを解決することができる１つの方法は、軽量で精度の低いＤＮＮを介してユーザデバイス（例えば、モバイルフォン）上で全ての計算を行うことによるものである。他の方法は、高精度のＤＮＮを介するが、単一／複数のモバイルデバイス及び／又はクラウドにわたって計算を共有することによるものである。
【００３８】
フレキシブルＡＩ方法
ユーザデバイス上でのみＤＮＮモデルを実行するために、モデル圧縮技法が広く活用されている。それらは、モデルメモリフットプリント及びランタイムを低減して、それを特定のデバイスに適合させることを可能にする。しかしながら、モデルがどのデバイス上で実行されるかを前もって知らない場合があり、デバイスが知られている場合であっても、

その利用可能なリソースは、例えば他のプロセスに起因して時間とともに変化する場合がある。これらの問題を克服するために、いわゆるフレキシブルＡＩモデルのファミリーが最近提案された。これらのモデルは、例えば、早期分類終了を可能にすること、モデル幅を適合させること（スリム化）、又は切り替え可能なモデル重み量子化を可能にすることを通じて、利用可能なリソースに即座に適合することができる。

【００３９】

分散ＡＩ方法

いわゆる分散ＡＩ方法のいくつかは、２つ以上のデバイス（すなわち、ＷＴＲＵ）の間で、又はデバイスとクラウド／エッジとの間でモデルを分割する。例えば、図２は、特徴サイズ圧縮を伴わない、２つのデバイス、デバイス１とデバイス２との間の分散ＡＩのためのメカニズムを示している。分散ＡＩでは、非常に高い次元であり得る中間データ（特徴）が送信される必要がある。これは、処理におけるレイテンシを追加し、対応する送信ネットワークの帯域幅制限のために常に可能であるとは限らない。この問題を克服するために、ボトルネックを介して特徴サイズを縮小する方法が提案された。図３Ａは、特徴圧縮のための１つの候補分割を有するＤＮＮを示しており、ここで、ａ１、ａ２、又はａ３を分割点として使用することができる。図３Ｂは、特徴圧縮のための２つの候補分割（例えば、ａ１及びａ２）を有するＤＮＮを示している。図３Ｃは、特徴圧縮のための３つの候補分割（例えば、ｃ１、ｃ２及びｃ３）を有するＤＮＮを示している。

【００４０】

いかなる限定も導入することなく、特徴は、現象を表すために使用され得るデータの個々の測定可能な特性又は特徴として考慮され得る。１つ以上の特徴は、機械学習アルゴリズム、ニューラルネットワーク、及び／又はその層のうちの１つの入力及び／又は出力に関連し得る。例えば、特徴は、ベクトルとして編成されてもよい。例えば、無線ユースケースに関連付けられた特徴は、時間、送信機アイデンティティ、及び基準信号（Reference Signal、ＲＳ）上の測定値を含み得る。

【００４１】

例えば、測位情報を処理するために使用されるアルゴリズムに関連付けられた特徴は、測位ＲＳ（positioning RS、ＰＲＳ）の測定に関連付けられた値、基準信号受信電力（Reference Signal Receive Power、ＲＳＲＰ）などの量、基準信号受信品質（Reference Signal Receive Quality、ＲＳＲＱ）などの量、受信信号強度指示（Received Signal Strength Indication、ＲＳＳＩ）に関連する量、別個のソースの信号に基づく時間差測定に関連する量（例えば、時間ベースの測位方法の場合）、到来角測定に関連する量、ビームの品質に関連する量、及び／又はセンサからの出力（ＷＴＲＵ回転、カメラからの撮像など）に関連付けられた値を含むことができる。

【００４２】

例えば、チャネル状態情報（Channel State Information、ＣＳＩ）を処理するために使用されるアルゴリズムに関連付けられた特徴は、チャネル状態基準信号（Channel State Reference Signal、ＣＳＩ－ＲＳ）、同期信号ブロック（Synchronization Signal Block、ＳＳＢ）、プリコーディング行列指示（Precoding Matrix Indication、ＰＭＩ）、ランクインジケータ（Rank Indicator、ＲＩ）、チャネル品質インジケータ（Channel Quality Indicator、ＣＱＩ）、ＲＳＲＰ、ＲＳＲＱ、ＲＳＳＩなどの受信に関連付けられた量の測定値を含み得る。

【００４３】

例えば、ビーム管理及び選択を処理するために使用されるアルゴリズムに関連付けられた特徴は、ＣＳＩ、送信／受信ポイント（Transmit/Receive Point、ＴＲＰ）識別情報（identity、ＩＤ）、ビームＩＤ、及び／又はビーム障害検出（Beam Failure Detection、ＢＦＤ）に関係する１つ以上のパラメータ（例えば、十分なビーム品質のしきい値決定）を処理するためのものと同様の測定値に関連付けられた量を含み得る。

【００４４】

同様に、本明細書で説明される任意の方法は、ＡＩ／ＭＬ処理の特定の段階、例えば、

訓練又は推論のための機械学習アルゴリズムのために使用されるハイパーパラメータに適用され得るか、又はそのための特定のパラメータ設定を更に含み得る。

【００４５】

図４は、特徴圧縮のための単一の分割（それぞれａ２、ｂ２）を有するＤＮＮを示しており、ここで、特徴サイズは、（ａ）４から２に、かつ（ｂ）４から３に縮小される。具体的には、（ａ３）は、４から２への特徴サイズ縮小を実現するサブネットワークであり、（ｂ３）は、４から３への縮小を実現するサブネットワークである。既存の研究では、ＤＮＮは、各圧縮係数について特徴圧縮器及び拡張器を用いて最初から訓練される。圧縮係数は、圧縮器の出力における特徴サイズと圧縮器への入力における特徴サイズとの比であることに留意されたい。これは、圧縮係数を変更する必要があるときはいつでも、デバイス及びクラウドサーバは、クラウドサーバから新しいモデルを調整し、かつ新たにダウンロードしなければならないことを意味する。図５Ａは、単一の圧縮係数がサポートされる、帯域幅縮小器（ＢＷＲ、５１０）及び帯域幅拡張器（ＢＷＥ、５２０）を使用する、２つのデバイス、デバイス１とデバイス２との間の分散ＡＩのための特徴サイズ圧縮メカニズムを示している。しかしながら、これらの方法では、ボトルネックを異なる送信ネットワーク帯域幅に適合させることができない。

【００４６】

分散ＡＩパラダイムにおける柔軟性を提供するために、我々は、フレキシブル分散ＡＩ（Ｆｌｅｘｉｂｌｅ ａｎｄ Ｄｉｓｔｒｉｂｕｔｅｄ ＡＩ、ＦＤ－ＡＩ）手法を導入する。提案された手法は、ＤＮＮが２つ以上のデバイス間で分割され得るので、分散される。また、提案された手法は、デバイス内の利用可能なリソースに応じて、いくつかの可能な分割点候補の中から分割点を選択することができるので、柔軟である。加えて、各分割点で送信される特徴サイズは、送信のために利用可能なネットワーク帯域幅に適合するように圧縮され得る。

【００４７】

一実施形態では、我々は、ＤＮＮアーキテクチャの一部である切り替え可能なボトルネックサブネットワークを提案する。ボトルネックサブネットワークは、推論時に異なる送信ネットワーク帯域幅に適合することができるので、切り替え可能である。提案された設計では、我々は、特徴サイズを縮小するための層を有する１つのボトルネックサブネットワークと、特徴サイズを元のサイズに戻すための層の他のセットとを有する。これらのボトルネックサブネットワークは、任意の既存のＤＮＮの１つ以上の分割位置に組み込むことができる。簡潔にするために、以下の説明では、我々は、特徴サイズ縮小及び拡張のためのボトルネックサブネットワークの１つのセットを有する単一の分割を有するＤＮＮを検討する。

【００４８】

一実施例では、第１のデバイスは、エッジデバイス又はクラウドサーバのいずれかであり得、第２のデバイスは、エッジデバイス又はクラウドサーバのいずれかであり得る。より一般的には、本明細書で説明する方法は、通信リンクを介してデータを交換する任意のデバイスに適用され得る。そのようなデバイスは、分割ニューラルネットワーク又はオートエンコーダ機能の処理を含み得る。本明細書で説明する方法は、例えば、エンドユーザアプリケーション（例えば、オーディオ、ビデオなど）のための、又はデータの送信及び／若しくは受信のための処理に関係する機能のための、デバイスにおける処理に適用可能であり得る。より一般的には、そのようなデバイスは、モバイル端末、ｇＮＢなどの無線アクセスネットワークノードなどであり得る。そのような通信リンクは、３ＧＰＰ　Ｕｕ、３ＧＰＰサイドリンク、又はＷｉｆｉリンクなどの無線リンク及び／又はインターフェースであり得る。

【００４９】

ボトルネックサブネットワークの特徴サイズ縮小層を有する分割点までのＤＮＮ層は、第１のデバイスにロードされる。残りの部分、すなわち、ボトルネックサブネットワーク拡張器及び分割点の後のＤＮＮの残りは、第２のデバイスにロードされる。我々は、動的特徴サイズスイッチ（Ｄｙｎａｍｉｃ ｆｅａｔｕｒｅ ｓｉｚｅ Ｓｗｉｔｃｈ、ＤｙＳｗ）として、縮小器及

び拡張器を含むボトルネックサブネットワークを参照する。第２のデバイスに送信される特徴は、ＤｙＳｗの中間で抽出される。我々は、これを実現するＤＮＮを動的切り替え可能な機能サイズネットワーク（Dynamic Switchable Feature Size Network、ＤｙＦｓＮｅｔ）と呼ぶ。ＤｙＦｓＮｅｔは、一般に、畳み込みニューラルネットワーク（convolutional neural network、ＣＮＮ）などの任意のＤＮＮアーキテクチャに適用され、設計及び訓練において新規である。ＤｙＦｓＮｅｔにおける推論は単純であり、（分割位置及び利用可能なネットワーク帯域幅に関して）調整可能である。

【００５０】

図５Ｂは、帯域幅縮小器（ＢＷＲ）及び帯域幅拡張器（ＢＷＥ）を使用して、２つのデバイス、デバイス－１とデバイス－２との間の分散ＡＩのための複数の圧縮係数をサポートする特徴サイズ圧縮メカニズムの一実施例を示しており、ここで、K_1、K_2、．．．、K_Nは、推論時に排他的かつ動的に切り替え可能である訓練可能なＢＷＲ（５３０）及びＢＷＥ（５４０）内の圧縮係数を指定する。

【００５１】

より具体的には、デバイス－１及びデバイス－２は、任意選択でサーバとともに、チャネル条件及びデバイス状態を監視し、分割位置における圧縮係数及び特徴サイズを選択する。デバイス－１は、分割位置までのＤＮＮモデルの第１の部分を受信し、デバイス－２は、ＤＮＮモデルの残りの部分を受信する。デバイス－１では、入力から特徴を計算するために推論が実施され、次いでＢＷＲによって圧縮される。図１０及び図１１に関連してより詳細に説明するように、ＢＷＲ（５３０）においてどのノードが接続されるかを制御することによって、異なる圧縮係数を得ることができる。デバイス－２では、圧縮された特徴が受信され、ＢＷＥ（５４０）によって拡張される。ＢＷＲと同様に、ＢＷＥは、ＢＷＥ内のノード接続を制御することによって圧縮係数を制御することができる。次に、デバイス－２は、推論を継続し、最終出力を提供する。

【００５２】

ネットワーク帯域幅制限は、推論全体に追加のレイテンシを導入する。図６Ａは、ＢＷＲ及びＢＷＥなしの全推論レイテンシを示している。図６Ｂは、中間データのサイズが低減され得る、ＢＷＲ及びＢＷＲを伴う総推論レイテンシを示している。

【００５３】

上述したように、我々は、予測の精度をほぼ維持しながら通信ネットワーク上のスループット要件を制限するために、ＤＮＮモデル内の異なる位置における中間データサイズを低減する方法を提案する。図７は、一実施形態による、分割／圧縮係数（ＣＦ）構成間で動的に切り替えるプロセスを示している。

【００５４】

モデル訓練及び分割／ＣＦ推定段階（７１０）の間、ＤｙＦｓＮｅｔモデルは、異なる分割及びＣＦについて訓練される。これは、現在、クラウドサーバにおいてオフラインで行うことができる。訓練されたモデルは、クラウドサーバに保存され、デバイスのダウンロードに利用可能である。（サーバ側の）オーケストレータは、要求に基づいて、訓練されたモデル選択及びエンドデバイスへの送信の調整を管理する。ここでは、帯域幅に関する情報が利用可能であると仮定する。これに基づいて、ＣＦは、特徴サイズと利用可能な帯域幅との比として推定される。

【００５５】

例えば、オーケストレータ又は外部制御システムは、エンドデバイス（例えば、デバイス－１及びデバイス－２）の計算能力に基づいて、ＤＮＮの分割位置を決定する。これは、分割情報に従って処理するためにＤＮＮをロードするデバイスに通信される。

【００５６】

モデル展開段階（７２０）では、訓練された分割モデルがデバイスによって受信される。受信されると、それらは推論のためにデバイスにロードされる。

【００５７】

ネットワーク（例えば、帯域幅）及び／又はデバイス（例えば、利用可能な処理能力）

の状態が監視される（７３０）。デバイスは、それらの間のネットワークチャネルを監視し、それらの間でＣＦを調整する。これは、サーバを伴わずに行われる。

【００５８】

デバイス間でコンセンサスに達すると、ＣＦ選択（７４０）が行われ、したがって、分割位置における特徴サイズに影響を与える。利用可能なＣＦオプションは、分割が実現されるＤＮＮ層のフィルタ内のチャネルの数に依存することに留意されたい。通常、ＣＦは、利用可能な帯域幅にほぼ一致するように選択され、厳密に一致するように選択されない。

【００５９】

分割モデル推論は、第１のデバイス及び第２のデバイス上で実施される（７５０）。例えば、第１のデバイスは、分割までのＤＮＮを用いて中間特徴を計算し、特徴を圧縮し、圧縮された特徴を第２のデバイスに転送する。第２のデバイスは、圧縮された特徴を受信し、それを圧縮せず、ＤＮＮ推論を続ける。デバイスが無線端末デバイスであり、かつ／又はデバイスの通信リンクが無線エアインターフェース（例えば、ＮＲ　Ｕｕ、サイドリンクなど）である一実施形態では、デバイスは、以下のうちの少なくとも１つを実施することができる。

－本明細書で提案される適合を開始する。例えば、デバイスは、分割処理ポイント、特徴次元、圧縮係数、推論レイテンシ、処理要件、機能の精度、又は本明細書で提案される任意の他の態様を適合し得る。

－デバイスは、Ｌ１／物理（physical、ＰＨＹ）層動作に関連して以下のうちの少なくとも１つを決定すると、ＡＩ処理のためのそのような適合をトリガし得る。

○デバイスは、無線特性の変化が発生したと決定し得、ここで、そのような特性は、例えば、セルのアイデンティティの変化、キャリア周波数の変化、帯域幅部分（bandwidth part、ＢＷＰ）の変化、ＢＷＰ及び／又はセルの物理リソースブロック（physical resource block、ＰＲＢ）の数の変化、サブキャリア間隔（sub-carrier spacing、ＳＣＳ）の変化、送信に利用可能なアグリゲートキャリアの数の変化、利用可能な送信電力の変化、測定量の変化などの、インターフェースを介した送信データレートに影響を与え得る。

○デバイスは、制御チャネルリソース（ＣＯＲＥＳＥＴ）又はアイデンティティの変化など、無線インターフェースを介した動作状態の変化が発生したと決定し得、ここで、第１のアイデンティティは、第１のしきい値に関連付けられ得、第２のアイデンティティは、第２のしきい値に関連付けられ得る。

○デバイスは、変化が、チャネル品質の劣化を示す特定の可能な構成されたしきい値を上回ると決定し得、ＡＩ処理に関連付けられたデータレートを低下させる適合を実施し得る。逆に、デバイスは、無線状態の改善を決定し、かつＡＩ処理に関連付けられたデータレートを増加させ得る適合を実施し得る。

例えば、これは、ＣＳＩ自動エンコーディング（autoencoding）のようなデバイスの物理層機能に適用可能である。

－デバイスは、Ｌ２／媒体アクセス制御（Medium Access Control、ＭＡＣ）層動作に関して以下のうちの少なくとも１つを決定すると、ＡＩ処理のためのそのような適合をトリガし得る。

○デバイスは、データ処理、情報ベアラ（例えば、データ無線ベアラ、シグナリング無線ベアラ）の変化が発生したことを決定し得、そのような特性は、論理チャネル優先順位付けパラメータ（例えば、パケット遅延バジェット（Packet Delay Budget、ＰＤＢ）、優先ビットレート（Prioritized Bit Rate、ＰＢＲ）、ＴＴＩ持続時間／ヌメロロジーの変化、関連するＱｏＳフローＩＤの変化、異なるデータレートを可能にするリソースのセットへのマッピング制限など）などの、ＡＩ処理に利用可能なインターフェースを介した送信データレートに影響を及ぼし得る。

○デバイスは、変化が、ＡＩ処理のために利用可能なデータレートの減少を示す特定の可能な構成されたしきい値を上回ると決定し得、ＡＩ処理に関連付けられたデータレー

トを低下させる適合を実施し得る。逆に、デバイスは、利用可能なデータレートの増加を決定し、ＡＩ処理に関連付けられたデータレートを増加させ得る適合を実施し得る。

例えば、これは、デバイスの測位機能のようなシステムレベル機能に適用することができる。例えば、これは、特定のデータ無線ベアラ（data radio bearer、ＤＲＢ）及び／又はＤＲＢタイプ、例えば、ＤＲＢ又はその特性の変化が、関連付けられたアプリケーション層におけるＡＩベースの処理の適合をトリガし得るような特定のＡＩ対応アプリケーションに関連付けられたＤＲＢへの適用であり得る。

－デバイスは、Ｌ３／無線リソース制御（Radio Resource Control、ＲＲＣ）層動作に関して、以下のうちの少なくとも１つを決定すると、ＡＩ処理のためのそのような適合をトリガし得る。

○デバイスは、例えば、利用可能なデータレートを変化し得る上記で説明した態様などの、Ｌ１／Ｌ２構成のうちの１つ以上に影響を与える、構成の変化が発生したと決定し得る。

○デバイスは、例えば、モビリティのための再構成メッセージを受信したこと、及び／又は（例えば、条件付きハンドオーバコマンドのために）適用すべきであることを決定することができ、ここで、メッセージは、ＡＩ処理及び／又はその関連する無線ベアラのための適用可能なデータレートの指示を含むことができる。

○デバイスは、無線リンク障害（radio link failure、ＲＬＦ）などの無線リンク障害が発生したと決定することができる。

○デバイスは、変化が、ＡＩ処理のために利用可能なデータレートの減少を示す特定の可能な構成されたしきい値を上回ると決定し得、ＡＩ処理に関連付けられたデータレートを低下させる適合を実施し得る。逆に、デバイスは、利用可能なデータレートの増加を決定し、ＡＩ処理に関連付けられたデータレートを増加させ得る適合を実施し得る。代替的に、それは、イベント自体が、ＡＩ処理のために利用可能なデータレートの増加（例えば、デバイスの構成の接続性、例えば、デュアル接続性へのセルの追加）又は減少（例えば、ＲＬＦ及び／又はデバイスの構成の接続性へのセルの除去）に関連付けられ得ると決定し得る。

－デバイスは、利用可能な処理リソースに関連して以下のうちの少なくとも１つを決定すると、ＡＩ処理のためのそのような適合をトリガし得る。

○デバイスは、例えば、インスタンス化された及び／又はアクティブなＡＩプロセスの数の変化に基づいて、動的デバイス能力の変化に基づいて、又はＡＩ処理のための処理要件（例えば、推論レイテンシ、精度）の変化に基づいて、利用可能なハードウェア処理の変化が生じたと決定し得る。

○デバイスは、デバイスの電力状態の変化が発生したと決定し得る。例えば、デバイスは、それが第１の状態から第２の状態に遷移したと決定し得、ここで、そのような状態は、ＲＲＣ接続性状態（ＩＤＬＥ、ＩＮＡＣＴＩＶＥ、又はＣＯＮＮＥＣＴＥＤ）、ＤＲＸ状態（アクティブ、非アクティブ）、又はそれらの異なる構成に関係し得る。

○デバイスは、変化が、利用可能な処理リソースの減少を示す、特定の可能な構成されたしきい値を上回ると決定し得る。逆に、デバイスは、利用可能な処理リソースの増加を決定し、ＡＩ処理に関連付けられたデータレートを増加させ得る適合を実施し得る。同様に、特定の状態は、関連付けられた特定のＡＩ処理レベル、分割点構成、及び／又はデータレートに関連付けられ得る。

－デバイスは、以下のうちの少なくとも１つに従って制御シグナリングを受信すると決定すると、ＡＩ処理のためのそのような適合をトリガし得る。

○デバイスは、ＡＩ処理／ＡＩ処理のために利用可能なデータレートの増加又は減少のいずれかを示す、制御情報を受信し得る。これは、シグナリングされた値、並びに／若しくはＬ１、Ｌ２、Ｌ３処理及び／若しくは省電力管理について上記で説明したような値の制御チャネルプロパティの修正に暗黙的に基づいてもよく、又は制御メッセージ中の指示を明示的に使用してもよい。そのような制御情報は、Ｌ１信号、Ｌ１メッセージ、例えば、ＰＤＣＣＨ上のＤＣＩにおいて、Ｌ２　ＭＡＣ制御要素において、又はＲＲＣメッセ

ージにおいて受信され得る。
○制御情報は、所与のＡＩ処理、ハイパーパラメータ設定、ターゲット解像度、ターゲット精度、ターゲット特徴ベクトルなどに適用される特定の分割点構成を含むことができる。
【００６０】
図８Ａ、図８Ｂ、及び図８Ｃは、プロセスの別の図を提供する。図８Ａは、デバイス１及び２（８４０、８６０）が、それらの計算能力及び送信チャネル（８５０）を推定することを示している。それらの推定値は、オペレータ／エッジ／クラウドに伝達され（８２０、８３０）、好適なＡＩ／ＭＬモデル（８１０）が要求される。
【００６１】
図８Ｂには、デバイスの各々からのＡＩ／ＭＬモデルの受信が示されている。オペレータ／クラウド／エッジは、モデルの選択を行い、ネットワークによってモデルを送信し（８３０）、要求されたモデルは、デバイス１及び２によって受信される。
【００６２】
図８Ｃは、デバイスの推論時間動作を示している。デバイス－１は、特徴を計算し、次いで、チャネル条件に基づいて、適切な寸法の特徴サイズがデバイス－２に送信される。デバイス１は、入力データ（８７０）に対して推論を実施する。入力データは、デバイスメモリからの１つ若しくは多くの画像、若しくはデバイスのカメラからライブでキャプチャされた画像、又はデバイスメモリ上のオーディオデータ、若しくはデバイスマイクロフォンからライブでキャプチャされたオーディオデータ、又はＤＮＮによって処理される必要がある任意の他のデータであり得る。デバイス－１は、ＭＳＤＮｅｔタイプのＤＮＮの場合など、ＤＮＮによって処理された中間又は初期出力（８８０）を出力する。特徴の更なる処理に必要な情報がまた、チャネル（８５０）を介してデバイス－２に通信される。デバイス－２は、特徴を受信し、更に、必要に応じて推論及びＣＦの切り替えを継続し、最終出力（８９０）を提供する。追加的に、デバイス－１は、特徴を更に処理するために、制御情報とともに特徴をデバイス－２に送信する。デバイス－２は、特徴及び制御情報を受信し、推論を続ける。
【００６３】
図９は、特徴圧縮のためにＤＮＮにおいて単一の分割を用いる提案された方法を示している。図９（ａ）は、圧縮係数が選択されていない共同訓練されたサブネットワークＤｙＳｗ（ａ３）を示している。図９（ｂ）は、選択された特徴圧縮係数４対２を有する共同訓練されたサブネットワーク（ｂ３）を示している。図９（ｃ）は、選択された特徴圧縮係数４対３を有する共同訓練されたサブネットワーク（ｃ３）を示している。なお、図９（ａ）、図９（ｂ）、及び図９（ｃ）のＤＮＮは、同じ（単一の）ＤＮＮである。
【００６４】
ＤｙＳｗは、ＤＮＮ全体と一緒に訓練され得る。代替的に、ＤｙＳｗを伴わないＤＮＮが事前訓練され、ＤｙＳｗサブネットワークが追加される。この代替解決策では、事前訓練されたＤＮＮは、ＤｙＳｗ（ａ３）サブネットワークで拡張され、訓練は、事前訓練されたＤＮＮ（の重み）を不変（すなわち、固定）に保ちながら、ＤｙＳｗのみに対するものであることに留意されたい。
【００６５】
図９に示すように、ＤｙＳｗは、複数の圧縮係数に適合するように再構成可能である。再構成は、ＤｙＳｗノードの接続詳細を通じて実現される。例えば、図１０に示すようなＤｙＳｗサブネットワークの場合、我々は、図１１に示すようなノード接続を指定するサイズ４×３の行列を維持することができる。行列内の各要素（Ｅ$_{ij}$）は、入力ノードｉが出力ノードｊに接続されているかどうかを表し、「０」は切断されていることを表し、「１」は接続されていることを表す。図１１（ａ）、図１１（ｂ）、及び図１１（ｃ）に示す行列は、それぞれ図９（ａ）、図９（ｂ）、及び図９（ｃ）に対応している。具体的には、図９（ａ）は、いずれの入力ノードもいずれの出力ノードにも接続されないことを指定し、図９（ｂ）は、出力ノードのうちの２つ（出力ノード－２及びノード－３）のみ

が入力ノードに接続されることを指定し、図９（ｃ）は、入力の全てのノードが出力に接続されることを指定する。図１１は、縮小器側の接続を示しており、拡張器は、異なる圧縮係数に対応する行列を維持することができる。一実施例では、拡張器側の行列の形状は（縮小器側のものに対して）転置されるが、全て０の行の数は同じままである。

【００６６】

図８に示されるように、デバイスはＣＦを調整する。一実施形態では、オーケストレータ又は外部制御システムが、利用可能な帯域幅についてデバイス－１に通知する。デバイス－１は、帯域幅に関する情報に基づいて、使用されるＣＦを決定する。次に、デバイス－１は、決定されたＣＦに対応する特徴サイズ圧縮を実現するためにＤｙＳｗを切り替える。デバイス－１はまた、それが使用しているＣＦを通信することができ、したがって、デバイス－２は、通信された情報に適合するようにＤＮＮのその側を切り替える。

【００６７】

一実施形態では、ＣＦが選択された後、デバイス－１は、選択されたＣＦを提供するためにノード間でどの接続が無効にされるべきかを決定し、デバイス－２はまた、拡張を適切に実施するために対応してどの接続が無効にされるべきかを決定する。ＣＦは、いくつの出力ノードが入力ノードに接続されるかを決定するが、その方法及びその数は学習を通して決定される。

【００６８】

上述したように、図１０は、サイズ４の入力を縮小及び拡張することができる例示的なＤｙＳｗを示している。図１０は簡略化のために単一層の「縮小器」ブロックを示しているが、縮小器は単一層に限定されないことに留意されたい。図示されたＤｙＳｗは、４対３、４対２、及び４対１の圧縮、並びに対応する拡張（すなわち、１対４、２対４、及び３対４）が可能である。ＤｙＳｗ設計は、必要に応じて追加の層、例えば、より良好な訓練のためのＢａｔｃｈＮｏｒｍ層を有し得る。ここでは、縮小器（点線の左側に示されたＢＷＲ）及び拡張器（点線の右側に示されたＢＷＥ）のみを示す。非線形性は、層に内在する。ＢａｔｃｈＮｏｒｍ層は、効率的な訓練のために必要とされるオプションの層であり得、したがって、ここでは示されない。

【００６９】

より一般的には、典型的なＤｙＳｗは、４つのタイプの層、すなわち、特徴次元縮小器及び拡張器層、非線形層、並びにバッチ正規化（ＢａｔｃｈＮｏｒｍ）層を備える。これらの層のうち、ＢａｔｃｈＮｏｒｍ層はオプションである。単純なＤｙＳｗを図１０に示す。

【００７０】

ＤＮＮ分類器で使用されるＤｙＳｗは、従来のタスク固有損失、例えば、分類タスクのためのクロスエントロピー損失又は回帰タスクのための平均二乗誤差損失を使用して訓練され得る。ＤｙＳｗは、任意のタスク、すなわち分類、検出、又はセグメント化のために、かつ任意のＤＮＮアーキテクチャ、すなわちＣＮＮ、ＧＡＮ、オートエンコーダなどにおいて使用され得る。ＤｙＳｗを訓練することは、縮小器－拡張器層重み及びバッチ正規化層（「ＢａｔｃｈＮｏｒｍ」としても示される）のパラメータを学習することを伴う。ＢａｔｃｈＮｏｒｍは、訓練のより速い収束のために使用される。

【００７１】

ＤｙＳｗ訓練は、損失目標に対する追加の制約を可能にする。例示として、我々は、ＤｙＳｗにわたる再構成損失の加算を示す。再構成損失は、ＤｙＳｗへの入力とＤｙＳｗへの出力との間の視差にペナルティを課す。ＤｙＳｗは、訓練されたＤＮＮに追加され得る、補助的なオプションのエンティティである。

【００７２】

ＤｙＳｗでは、低減係数は、推論時にオンザフライで切り替え可能である。ＤｙＦｓＮｅｔでは、訓練反復は、以下で更に詳述されるように、複数の低減係数を用いて共有ＤｙＳｗ重みを共学習するように修正される。

【００７３】

ＤｙＳｗの訓練は、クラウド／オペレータ／エッジ上で行われるオフライン若しくはオンラインであり得るか、又はデバイス上の連合訓練であり得る。ここでは、我々は、ＤｙＳｗを有する２つのデバイス間の単一分割の場合の分割ＤＮＮのアーキテクチャ及び訓練について説明する。本明細書で説明される訓練メカニズムは、複数の分割ケースに拡張され得る。以下では、我々は、分割ＤＮＮのアーキテクチャ、ＤｙＳｗ層及びＤｙＦｓＮｅｔ（ＤｙＳｗ層を有するＤＮＮ）のアーキテクチャ、並びに異なる損失関数及びそれらの訓練を詳細に説明する。

【００７４】

デバイス－１が層－１まで処理し、デバイス－２が層ｌ＋１以降で処理する、ｌ番目の層の終わりでの分割を考える。デバイス－１内のＤＮＮの一部を$h_{device1}$とし、同様にデバイス－２内のＤＮＮの一部を$h_{device2}$とする。ＤＮＮへの入力は任意のタイプのデータとすることができるが、ここでは、入力Ｘを、$X \in R^{\{W \times H \times 3\}}$となるようなカラー画像し、式中、Ｗ、Ｈは、幅、高さであり、３は、カラーチャネル（例えば、ＲＧＢ）の数を表す。分割における特徴テンソル（又は単に特徴）は、$y_l \in R^{\{M \times N \times C\}}$であり、式中、Ｍ、Ｎ、及びＣは、その幅、その高さ、及びチャネル数を表す。特徴y_lが無線ネットワークを介してデバイス－２に送信され、デバイス－２は、入力としてy_lを取り、出力Ｙを生成する。したがって、$y_l = h_{device1}(X)$であり、$Y = h_{device2}(y_l)$である。

【００７５】

ＤｙＳｗは、h_{DySw}によって表されるサブネットワークである。h_{DySw}のパラメータは、θ_{DySw}である。ＤｙＳｗの縮小器（第１の部分）及び拡張器（第２の部分）をＢＷＲ及びＢＷＥと呼ぶとすると、そのような縮小器及び拡張器の例示的な実装形態は、以下に要約されるように、畳み込み層、非線形層（ＲｅＬｕ）、及びバッチ正規化層（ＢａｔｃｈＮｏｒｍ）を含むことができる。

【００７６】

【数１】

$$\theta_{\mathrm{DySw}} = [\theta_{DySw_{BWR}}; \theta_{\mathrm{DySw_{BWE}}}]$$

$$\theta_{DySw_{BWR}} = [layer_{conv_{DySwBWR}}; ReLu; BatchNorm]$$

$$\theta_{\mathrm{DySw_{BWE}}} = [layer_{conv_{DySwBWE}}; ReLu; BatchNorm]$$

【００７７】

ＤｙＳｗを有するＤＮＮは、ＤｙＦｓＮｅｔと呼ばれる。ＤｙＦｓＮｅｔをｈのように表す。θはｈのパラメータとする。分割点の前のＤｙＦｓＮｅｔのサブネットワークは、

【００７８】

【数２】

$$h_{device1_{BWR}}$$

である。分割点の後のサブネットワークは、

【００７９】

【数３】

$$h_{device2_{BWE}}$$

である。

【００８０】

ＤｙＳｗは、特徴サイズの様々な圧縮係数（ＣＦ）の間で切り替える。ＣＦ切り替えは、Ｋによってインデックス付けされる。ＤｙＦｓＮｅｔの分割における、Ｋによってイン

デックス付けされる中間出力は、以下の通りである。

【００８１】

【数４】

$$y'_{2_K} = h_{DySw_{BWR}}(y_2)$$

$$\widehat{y_{2_K}} = h_{DySw_{BWE}}(y'_{2_K})$$

$$\hat{Y} = h_{device2}(\widehat{y_{2_K}})$$

ここで

【００８２】

【数５】

$h_{DySw_{BWR}}$ 及び $h_{DySw_{BWE}}$

は、それぞれＢＷＲ及びＢＷＥを行うＤｙＳｗサブネットワークであり、

【００８３】

【数６】

$$y'_{2_K} \in R^{\left\{M \times N \times \frac{C}{K}\right\}},\ \widehat{y_{2_K}} \in R^{\{M \times N \times C\}}$$

及びＤＮＮ分類器について

【００８４】

【数７】

$$\widehat{Y_K} \in R^{\{N_c\}}$$

であり、N_cは、クラスの数であり、下付き文字Ｋは、圧縮係数を表す。

【００８５】

【数８】

$$\widehat{Y_K}$$

は、ＤＮＮの目的によって異なり、それが分類器であるか、回帰器であるか、又は生成器であるかに依存する。一般性を失うことなく、我々は、ここでは分類器の場合を仮定する。

【００８６】

設定は、２つのタイプの監視を提供し、一方のタイプは、グラウンドトゥルースラベル

【００８７】

【数９】

$$_Y_{true} \in B^{\{N_c\}}$$

を介し、他方のタイプは、ＤｙＳｗサブネットワークへの入力とＤｙＳｗサブネットワークの出力との間の再構築損失（例えば、平均二乗誤差の形式）である。追加的に、ＤｙＦｓＮｅｔが事前訓練されたＤＮＮで初期化される場合、事前訓練されたＤＮＮ、Y_{KD}の出力とＤｙＳｗサブネットワークの出力との間の知識蒸留損失を使用することが可能である。利便性のために、我々は、Y_{True} and Y_{KD}で計算された損失を、監視をグローバル損失として、かつＤｙＳｗを通じる再構築損失

【００８８】

【数１０】

$$\left(y_l, \widehat{y_{l_K}}\right)$$

をローカル損失として参照することができる。ネットワークの訓練中に最適化される異なるタイプの損失を以下に示す。

【００８９】

最初から訓練されたＤｙＦｓＮｅｔ：

【００９０】

【数１１】

$$Loss = \lambda * loss_{\{cross-entropy\}}\left(Y_{true}, \widehat{Y_K}\right) + (1-\lambda)\, loss_{reconstruction}\left(y_l, \widehat{y_{l_K}}\right),$$

$$s.t. 0 \leq \lambda \leq 1$$

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\,(Loss)$$

【００９１】

事前訓練された初期化を使用して訓練されたＤｙＦｓＮｅｔ：

【００９２】

【数１２】

$$Loss =$$

$$\lambda_1 * loss_{\{cross-entropy\}}\left(Y_{true}, \widehat{Y_K}\right) + \lambda_2 loss_{reconstruction}\left(y_l, \widehat{y_{l_K}}\right) + \lambda_3\, loss_{KD}\left(Y_{KD}, \widehat{Y_K}\right),$$

$$s.t.\ \lambda_1 + \lambda_2 + \lambda_3 = 1\ and\ 1 \geq\ \lambda_1,\ \lambda_2,\ \lambda_3 \geq 0$$

$$\theta^*_{DySw} = \underset{\theta_{\mathrm{DySw}}}{\operatorname{argmin}}\,(Loss)$$

【００９３】

最初から訓練されたマルチ分割ＤｙＦｓＮｅｔ：

【００９４】

【数１３】

$$Loss = \lambda * loss_{\{cross-entropy\}}\left(Y_{true}, \widehat{Y_K}\right) + (1-\lambda)\sum_{l=1}^{L} loss_{reconstruction}\left(y_l, \widehat{y_{l_K}}\right),$$

$$s.t. 0 \leq \lambda \leq 1$$

$$\theta^*_{DyFsNet} = \underset{\theta_{\mathrm{DyFsNet}}}{\operatorname{argmin}}\,(Loss)$$

【００９５】

事前訓練された初期化から訓練されたマルチ分割ＤｙＦｓＮｅｔ：

【００９６】

【数１４】

$$Loss =$$

$$\lambda_1 * loss_{\{cross-entropy\}}\left(Y_{true}, \widehat{Y_K}\right) + \lambda_2 \sum_{l=1}^{L} loss_{reconstruction}\left(y_l, \widehat{y_{l_K}}\right) + \lambda_3\, loss_{KD}\left(Y_{KD}, \widehat{Y_K}\right),$$

$$s.t.\ \lambda_1 + \lambda_2 + \lambda_3 = 1\ and\ 1 \geq \lambda_1, \lambda_2, \lambda_3 \geq 0$$

$$\theta^*_{DySw} = \underset{\theta_{\mathrm{DySw}}}{\operatorname{argmin}}\ (Loss)$$

【００９７】

ＤｙＦｓＮｅｔ訓練アルゴリズム

（Ｘｉ，Ｙｉ）∈Ｄをデータセットとし、ここで、Ｘｉ及びＹｉは、それぞれデータ及びその監視であり、ｉ∈｛０、１、．．．、Ｎ｝は、インデックスであり、Ｎは、訓練サンプルの数であり、Ｎｕｍ－ｏｆ－ｅｐｏｃｈｓは、訓練エポックの数である。ここで、我々は、グローバル損失、すなわち、クロスエントロピー及びＫＤを使用する分類器のための訓練アルゴリズムを与える。ＫＤベースの損失は、ｉ）圧縮なしのＤｙＳｗ（すなわち、Ｋ＝１を有するＤｙＳｗ）の出力、ｉｉ）目下のより低い圧縮係数を有するＤｙＳｗの出力（すなわち、Ｋ＝Ｋ１を有するＤｙＳｗからＫ＝Ｋ２への蒸留、ここで、Ｋ１＜Ｋ２である）、ｉｉｉ）非圧縮ＤｙＳｗ出力と最も近い圧縮されたＤｙＳｗ出力／ｓとのアフィン結合、又はｉｖ）同じタスクのために十分に訓練された完全に異なるＤＮＮアーキテクチャの出力、からの蒸留による４つのタイプであり得る。

【００９８】

全体のアルゴリズムは以下の通りである。

ａ．ＤｙＳｗの非圧縮構成についてＤｙＦｓＮｅｔの損失を計算する。我々の実施例では、それはクロスエントロピー損失であるが、それだけに限定されない。

ｂ．逆伝播を行い、ＤｙＳｗの非圧縮構成について勾配を累積する。

ｃ．１～Ｃの範囲のＣＦのＮｒ数を選択し、ここで、１は非圧縮を表し、Ｃは最大圧縮を表す。

ｄ．ＣＦ＝２～Ｎｒ：の場合

ｉ．蒸留タイプ（ｉ）、（ｉｉ）、（ｉｉｉ）、又は（ｉｖ）についてＤｙＦｓＮｅｔの損失を計算する。

ｉｉ．逆伝播を行い、ＤｙＳｗの勾配を累積する。

ｅ．累積された勾配を使用して重みを更新する。

【００９９】

一実施例では、以下の擬似コードが使用される。

【０１００】

非圧縮（Ｋ＝１）ＤｙＳｗ出力からのＫＤ：

【０１０１】

【表１】

For n in range(Num-of-epochs) DO:

  For i in range(N) DO:

    // Forward pass through DyFsNet without compression i.e., compression factor K=C

    $\widehat{Y_{K=1}} = h_{device2}(h_{DySw}(h_device1(X_i)))$

    Calculate loss: $\boldsymbol{loss}_{\{cross-entropy\}}(Y_{true}, \widehat{Y_{K=1}})$

    Do loss-Backpropagation and accumulate gradients

    // Sample N_r random numbers where N_r <= C

    // Let S be a set of N_r random numbers

    // Note each element in S represents a compression factor (CF) in ascending order

    S = random (1, C, size= N_r)

    For j in range(N_r) :

      K = S[j]

      $Y_K = h_{device2}(h_{DySw}(h_device1(X_i)))$

      Calculate KD loss: $\boldsymbol{loss}_{\{KD\}}(\widehat{Y_{K=1}}, \widehat{Y_K})$

      Do loss-Backpropagation and accumulate gradients

    END For

    Update weights of DyFsNet (whole or just the DySw if using fixed pretrained weights)

  End For

End For

【０１０２】

Ｋ＝Ｋ１を有するＤｙＳｗの出力からＫ＝Ｋ２を有するＤｙＳｗへのＫＤであり、ここで、Ｋ１＜Ｋ２である：

【０１０３】

【表２】

For n in range(Num-of-epochs) DO:

For i in range(N) DO:

// Sample N_r random numbers where N_r <= C

// Let S be a set of N_r random numbers

// Note each element in S represents a compression factor (CF) in ascending order

S = random (1, C, size= N_r)

// Forward pass through DyFsNet without compression i.e., compression factor K=C

$\widehat{Y_{S[1]}} = h_{device2}(h_{DySw}(h_device1(X_i)))$

Calculate loss: $loss_{\{cross-entropy\}}(Y_{true}, \widehat{Y_{S[1]}})$

Do loss-Backpropagation and accumulate gradients

For j in range (2, N_r*)* :

K = S[j]

$Y_K = h_{device2}(h_{DySw}(h_device1(X_i)))$

Calculate KD loss: $loss_{\{KD\}}(\widehat{Y_{S[j-1]}}, \widehat{Y_K})$

Do loss-Backpropagation and accumulate gradients

END For

Update weights of DyFsNet (whole or just the DySw if using fixed pretrained weights)

End For

End For

【０１０４】

Ｋ＝Ｋ１を有するＤｙＳｗの出力からＫ＝Ｋ２を有するＤｙＳｗへのＫＤであり、ここで、Ｋ１＜Ｋ２である：

【０１０５】

【表３】

For n in range(Num-of-epochs) DO:

For i in range(N) DO:

// Sample N_r random numbers where N_r <= C

// Let S be a set of N_r random numbers

// Note each element in S represents a compression factor (CF) in ascending order

S = random (1, C, size= N_r)

// Forward pass through DyFsNet without compression i.e., compression factor K=C

$\widehat{Y_{K=S[1]}} = h_{device2}(h_{DySw}(h_device1(X_i)))$

Calculate loss: $loss_{\{cross-entropy\}}(Y_{true}, \widehat{Y_{K=S[1]}})$

Do loss-Backpropagation and accumulate gradients

For j in range (2, N_r) :

K = S[j]

$Y_K = h_{device2}(h_{DySw}(h_device1(X_i)))$

Calculate KD loss: $\sum_{t=1}^{j-1} \lambda_t loss_{\{KD\}}(\widehat{Y_{S[t]}}, \widehat{Y_K})$

Do loss-Backpropagation and accumulate gradients

END For

Update weights of DyFsNet (whole or just the DySw if using fixed pretrained weights)

End For

End For

【０１０６】

我々は、周知のＭＳＤＮｅｔモデルを使用して、画像分類タスクについて提案された考え方を試験した。このモデルは、任意のブロックの出力において分類を行うことができるいくつかのＣＮＮブロックを有する。我々は、異なるブロックの終わりでこの大きなネットワークを分割し、対応する特徴を第２のデバイス（又はクラウド）に送信することを望む。ＩｍａｇｅＮｅｔデータセットの各ブロックの終わりのＭＳＤＮｅｔの特徴寸法を表１に示す。

【０１０７】

【表４】

表１

分割点	入力画像（224×224×3）当たりのおおよその出力データサイズ（MByte）	要求されるULデータレート（Mbps）
候補分割点0（クラウドベースの推論）	0.14	34
候補分割点1（ブロック1の後）	2.15	516.72
候補分割点2（ブロック2の後）	1.00	241.17
候補分割点3（ブロック3の後）	0.33	78.95
候補分割点4（ブロック4の後）	0.056	13.63

【０１０８】
ここで、我々は、典型的なＤＮＮにおけるデータレート要件の例示を通じて、特徴サイズ縮小の有用性を示す。画像分類（ＭＳＤＮｅｔ）に使用されるＤＮＮで生成されたサイズ２２４×２２４×３の単一画像に対応する特徴を送信するために必要なデータレートは、１３Ｍｂｐｓから０．５Ｇｂｐｓの範囲である。これは、無線ネットワーク上での送信にとって困難なデータレートである。ＭＳＤＮｅｔモデルを使用する我々のアプローチの予備的な実装において、我々は、精度の最大１％の損失で、特徴サイズを５０％低減することができた。

【０１０９】
以下では、ＤＮＮが７つの位置で分割され、各分割位置における特徴サイズ（各単位は１６ビットである）が表２に示される、ＣＩＦＡＲ－１００のためのＭＳＤＮｅｔにおけるＤｙＳｗの我々の実装について説明する。我々は、１、２、４及び１０の圧縮係数を実現した。

【０１１０】
【表５】
表２

分割位置	特徴サイズ（16ビットのサイズ）
1	10240
2	13312
3	8960
4	12032
5	15104
6	9728
7	12800

【０１１１】
ＭＳＤＮｅｔで帯域幅縮小器－拡張器を追加することの効果を調査する。我々は、ベースライン（帯域幅縮小器－拡張器なし）の場合と、縮小係数１、２、４及び１０の場合について帯域幅縮小器－拡張器ありの場合の結果を表３に示す。縮小係数１、２、４及び１０は、それぞれ元の帯域幅の１００％、５０％、２５％及び１０％に対応する。帯域幅が低減されたＭＳＤＮｅｔの精度は、いかなる低減もないベースラインＭＳＤＮｅｔとほぼ同じであることが分かる。精度は、全ての６つのブロック（０～６）及び全てのスケールの終わりにおける圧縮実装についてのものであることに留意されたい。言い換えれば、各分割点に新しい帯域幅縮小器－拡張器を追加することによって、特徴を大幅に縮小して特徴送信をサポートすることができる一方で、分類精度はほとんど変わらない。

【０１１２】
【表６】
表３

圧縮係数	ブロック－0（上位－1 Acc. %）	ブロック－1（上位－1 Acc. %）	ブロック－2（上位－1 Acc. %）	ブロック－3（上位－1 Acc. %）	ブロック－4（上位－1 Acc. %）	ブロック－5（上位－1 Acc. %）	ブロック－6（上位－1 Acc. %）
ベースライン	65.05	67.7	70.88	71.78	72.35	73.67	73.79
K＝1	65.04	67	69	71	72	72	72.2
K＝2	65.64	69.54	69.91	71.05	71.6	71.34	71.78
K＝4	64.37	68.42	69.99	71.76	72.22	72.65	72.61
K＝10	65.05	67.28	68.15	68.49	69.48	69.75	70.38

【０１１３】
ＣＮＮ重みの精度を参照する切り替え可能な精度ネットワークの方法が存在する。また、切り替え可能な多重幅ＣＮＮに関する研究もあった。しかし、それらとは異なり、我々

は、推論時に異なる特徴帯域幅間で切り替えることができる、切り替え可能な特徴帯域幅ネットワークを提案する。この切り替え可能性は、デバイス又はデバイスクラウド又はそれらの他の組み合わせの間の通信チャネルの帯域幅制約に対処するのに有用である。例えば、このメカニズムは、ＲｅｓＮｅｔ、ＡｌｅｘＮｅｔ、ＤｅｎｓｅＮｅｔ、ＳｏｕｎｄＮｅｔ、ＶＧＧなどの異なる機械学習タスクを実施する既存のモデルとシームレスに使用され得る、ＣＮＮアーキテクチャに依存せずに使用され得る。このメカニズムはまた、重み量子化などの他のタイプの特徴圧縮技法に依存せずに使用され得る。

【０１１４】

提案される手法は、複数の特徴帯域幅の間で切り替えるための規定を有する分散ＡＩのための送信のための効率的な帯域幅を扱う。エッジデバイスにおける分散推論中、各デバイスは、ＡＩモデルの一部を一度だけロードする必要があるが、それらの間で通信される入力／出力特徴は、ＤｙＳｗ内のノード間の接続を有効化／無効化することによって、利用可能な送信帯域幅に応じて柔軟に構成され得る。所望の圧縮係数を達成するためにいくつかのノードが接続又は切断されるときに、ＤＮＮの他のパラメータは同じままである。すなわち、同じＤＮＮモデルが異なる圧縮係数に使用され、圧縮係数又はネットワーク帯域幅に適合するために新しいＤＮＮモデルをダウンロードする必要はない。

【０１１５】

ＡＩ処理は、例えば、ベーシックな携帯電話のカメラで撮影された画像に対して、又はジェスチャ検出を介したＵＩ対話のためにスマートＴＶカメラから撮影された画像に対して使用することができるが、これらに限定されない。提案される手法は、様々なシナリオにおいて使用され得る。例えば、ＡＩモデルは、デバイスとクラウドとの間で分割され得る。以下に、我々は、いくつかの可能な使用シナリオを列挙する。

１．ＡＩモデルは、２つのデバイス間で分割される。例えば、ユーザは、スマートウォッチにおいてキャプチャされたデータを処理することを望み、ここで、処理の一部は、ウォッチ上で行われ、残りはユーザのモバイルフォン上で行われ得る。

２．ＡＩモデルは、複数のデバイスと場合によってはクラウドとの間で分割される。例えば、ユーザは、スマートＣＣＴＶカメラのフィードをカメラ自体で迅速に処理し、詳細な処理をクラウド又はローカルサーバで処理することを望む。

３．ユースケース３と同様であるが、ＣＣＴＶカメラの代わりに計算対応マイクロフォンを用いたスピーチ／オーディオ処理を伴う。

４．医療データ診断室及びクラウドの処理を共有する。

５．無線リンクを介して通信することができる端末デバイスであって、ここで、ＡＩ処理は、無線処理チェーン（例えば、ＣＳＩ圧縮、ＣＳＩ自動符号化、測位決定など）の送信及び／又は受信の機能に関係する。

６．無線リンクを介して通信することができる端末デバイスであって、ここで、ＡＩ処理は、スケジューリング又はデータ処理の機能に関係し、例えば、ＱｏＳ処理（例えば、ユーザプレーンデータレート適応など）に関係する。

【０１１６】

本出願において、様々な数値が使用されている。具体的な値は、例示目的のために提供され、記載の態様は、これらの具体的な値に限定されない。

【０１１７】

特徴及び要素は、特定の組み合わせにおいて上で説明されているが、当業者は、各特徴又は要素が単独で又は他の特徴及び要素との任意の組み合わせで使用され得ることを理解されよう。加えて、本明細書に説明される方法は、コンピュータ又はプロセッサによる実行のためにコンピュータ可読媒体に組み込まれたコンピュータプログラム、ソフトウェア又はファームウェアに実装され得る。非一時的なコンピュータ可読記憶媒体の例としては、読み取り専用メモリ（ＲＯＭ）、ランダムアクセスメモリ（ＲＡＭ）、レジスタ、キャッシュメモリ、半導体メモリデバイス、内部ハードディスク及びリムーバブルディスクなどの磁気媒体、磁気光学媒体及びＣＤ－ＲＯＭディスク及びデジタル多用途ディスク（ｄｉｇｉｔａｌ ｖｅｒｓａｔｉｌｅ ｄｉｓｋ、ＤＶＤ）などの光学媒体が挙げられるが、これらに限定されな

い。ソフトウェアと関連付けられたプロセッサを使用して、ＵＥ、ＷＴＲＵ、端末、基地局、ＲＮＣ又は任意のホストコンピュータにおいて使用するためのビデオエンコーダ、ビデオデコーダ、又はその両方、無線周波数トランシーバを実装し得る。

【０１１８】

更に、上記の実施形態では、処理プラットフォーム、コンピューティングシステム、コントローラ、及びプロセッサを含む他のデバイスが記載されている。これらのデバイスは、少なくとも１つの中央処理装置（「Central Processing Unit、ＣＰＵ」）及びメモリを含み得る。コンピュータプログラミングの技術分野における当業者の慣例によれば、動作、及び演算又は命令の記号表現の言及は、様々なＣＰＵ及びメモリによって実施され得る。そのような動作及び演算又は命令は、「実行」、「コンピュータ実行」、又は「ＣＰＵ実行」されることと呼ばれ得る。

【０１１９】

当該技術分野における通常の技術を有する者には、動作及び記号的に表現された演算又は命令が、ＣＰＵによる電気信号の操作を含むことが理解されるであろう。電気システムは、電気信号の結果的な変換又は減少を引き起こすことができるデータビットを表し、メモリシステムのメモリ位置にデータビットを維持し、それによってＣＰＵの動作及び他の信号の処理を再構成又は別の方法で変更する。データビットが維持されるメモリ位置は、データビットに対応する、又はデータビットを表す特定の電気的特性、磁気的特性、又は光学的特性を有する物理的位置である。例示的な実施形態は、上述したプラットフォーム又はＣＰＵに限定されず、他のプラットフォーム及びＣＰＵが、提供される方法をサポートし得ることを理解されたい。

【０１２０】

データビットはまた、磁気ディスク、光学ディスク、及び任意の他の揮発性（例えば、ランダムアクセスメモリ（「ＲＡＭ」））又はＣＰＵによって読み取り可能な不揮発性（例えば、読み取り専用メモリ（「ＲＯＭ」））大容量記憶システムを含む、コンピュータ可読媒体上に維持され得る。コンピュータ可読媒体は、処理システム上に排他的に存在するか、又は処理システムに対してローカル又はリモートであり得る複数の相互接続された処理システム間で分散された、協調的又は相互接続されたコンピュータ可読媒体を含んでもよい。代表的な実施形態は、上述のメモリに限定されず、他のプラットフォーム及びメモリが、記載された方法をサポートし得るということが理解される。

【０１２１】

例示的な実施形態において、本明細書に記載されている動作、プロセスなどのいずれも、コンピュータ可読媒体に格納されたコンピュータ可読命令として実装されてもよい。コンピュータ可読命令は、移動体、ネットワーク要素、及び／又は任意の他のコンピューティングデバイスのプロセッサによって実行され得る。

【０１２２】

ハードウェア又はソフトウェアの使用は、一般に（常にではないが、特定の状況では、ハードウェアとソフトウェアとの間の選択が大きな意味を持ち得る）、コスト対効率のトレードオフを意味する設計上の選択事項である。本明細書に記載されているプロセス及び／又はシステム及び／又は他の技術が効果的であり得る様々なビークル（例えばハードウェア、ソフトウェア、及び／又はファームウェア）が存在し得、好ましいビークルは、プロセス及び／又はシステム及び／又は他の技術が配備される状況によって変化し得る。例えば、実装者が、速度及び正確性が最重要であると判定した場合、実装者は、主にハードウェア及び／又はファームウェアのビークルを選択することができる。柔軟性が最重要である場合、実装者は、主にソフトウェア実装を選択することができる。代替的に、実装者は、ハードウェア、ソフトウェア、及び／又はファームウェアの何らかの組み合わせを選択してもよい。

【０１２３】

前述の詳細な説明では、ブロック図、フローチャート、及び／又は例の使用を通じて、デバイス及び／又はプロセスの様々な実施形態を示した。そのようなブロック図、フロー

チャート、及び／又は例が１つ以上の機能及び／又は動作を含む限り、そのようなブロック図、フローチャート、又は例の中の各機能及び／又は各動作は、広範なハードウェア、ソフトウェア、ファームウェア、又はそれらの実質的に任意の組み合わせによって、個別にかつ／又は集合的に実装されてよいことが当業者には理解されるであろう。好適なプロセッサとしては、例として、ＧＰＵ（グラフィックスプロセッシングユニット）、汎用プロセッサ、専用プロセッサ、従来型プロセッサ、デジタル信号プロセッサ（ＤＳＰ）、複数のマイクロプロセッサ、ＤＳＰコアと関連する１つ以上のマイクロプロセッサ、コントローラ、マイクロコントローラ、特定用途向け集積回路（ＡＳＩＣ）、アプリケーション特定用途向け標準製品（ＡＳＳＰ）、フィールドプログラマブルゲートアレイ（ＦＰＧＡ）回路、任意のその他のタイプの集積回路ＩＣ）、及び／又は状態機械が挙げられる。

【０１２４】

上記では特徴及び要素が特定の組み合わせにおいて提供されているが、当該技術分野の通常の技術を有する者には、各特徴若しくは各要素を単独で使用する、又は他の特徴及び要素との任意の組み合わせにおいて使用できることが理解されるであろう。本開示は、本出願に記載されている特定の実施形態の観点において限定されるものではなく、これらの実施形態は、様々な態様の例示として意図されるものである。当業者には明らかなように、本発明の趣旨及び範囲から逸脱することなく、多くの修正及び変形を行うことができる。本出願の説明において使用されているいかなる要素、動作、又は指示も、そのように明示的に提示されていない限り、本発明にとって重要又は本質的であると解釈されるべきではない。本明細書に列挙したものに加えて、本開示の範囲内の機能的に等価な方法及び装置が、上述した説明から、当業者には明らかであろう。そのような修正及び変形は、添付の特許請求の範囲に入ることが意図されている。本開示は、添付の特許請求の範囲の条項によってのみ限定されるものであり、かかる特許請求の範囲が権利を有する等価物の完全な範囲とともに、限定されるものである。本開示は、特定の方法又はシステムに限定されないことを理解されたい。

【０１２５】

本明細書で使用される用語は、特定の実施形態のみを説明する目的のためであり、限定することを意図するものではないということも理解されたい。

【０１２６】

特定の代表的な実施形態では、本明細書に記載の主題のいくつかの部分は、特定用途用集積回路（ＡＳＩＣ）、フィールドプログラマブルゲートアレイ（ＦＰＧＡ）、デジタル信号プロセッサ（ＤＳＰ）、及び／又は他の統合フォーマットを介して実装され得る。しかしながら、本明細書に開示されている実施形態のいくつかの態様は、その全体又は一部が、１つ以上のコンピュータ上で動作する１つのコンピュータプログラムとして（例えば１つ以上のコンピュータシステム上で動作する１つ以上のプログラムとして）、１つ以上のプロセッサ上で動作する１つ以上のプログラムとして（例えば１つ以上のマイクロプロセッサ上で動作する１つ以上のプログラムとして）、ファームウェアとして、又はこれらの実質的に任意の組み合わせとして、集積回路において等価的に実施され得ること、並びに、回路を設計すること、及び／又は、ソフトウェア及び／若しくはファームウェアのコードを書くことが、この開示に照らして当業者の技術の範囲内であることが、当業者には認識されるであろう。加えて、本明細書に記載されている主題のメカニズムが、様々な形態のプログラム製品として配布され得ること、及び、本明細書に記載されている主題の例示的な実施形態が、配布を実際に行うために使用される特定のタイプの信号担持媒体にかかわらず適用されることが、当業者には理解されるであろう。信号担持媒体の例としては、フロッピーディスク、ハードディスクドライブ、ＣＤ、ＤＶＤ、デジタルテープ、コンピュータメモリなどの記録可能型媒体、並びに、デジタル及び／又はアナログ通信媒体（例えば光ファイバケーブル、導波管、有線通信リンク、無線通信リンクなど）などの送信型媒体が挙げられ、ただしこれらに限定されない。

【０１２７】

本明細書に記載されている主題は、場合によっては、異なる他の構成要素内に含まれる

か、又は、異なる他の構成要素に接続されている、異なる構成要素を示していることがある。そのような図示されたアーキテクチャは単なる例であり、実際には、同じ機能を達成する他の多くのアーキテクチャが実施され得ることを理解されたい。概念的には、同じ機能を達成するための構成要素の任意の配置は、所望の機能が達成され得るように、効果的に「関連付けられる」。したがって、特定の機能を達成するために本明細書において組み合わされた、任意の２つの構成要素は、アーキテクチャ又は中間構成要素に関係なく、所望の機能が達成されるように、互いに「関連付けられた」として見ることができる。同様に、そのように関連付けられた任意の２つの構成要素は、所望の機能を達成するために互いに「動作可能に接続されている」、又は「動作可能に結合されている」とみなすこともでき、そのように関連付けることができる任意の２つの構成要素は、所望の機能を達成するために互いに「動作可能に結合可能」であるとみなすこともできる。動作可能に結合可能の具体例としては、物理的に嵌合可能かつ／若しくは物理的に相互作用する構成要素、及び／又は、無線で相互作用可能かつ／若しくは無線で相互作用する構成要素、及び／又は、論理的に相互作用するかつ／若しくは論理的に相互作用可能な構成要素が挙げられ、ただしこれらに限定されない。

【０１２８】

本明細書における実質的に任意の複数形及び／又は単数形の用語の使用に関して、当業者は、文脈及び／又は用途に適切であるように、複数形から単数形に、かつ／又は単数形から複数形に変換することができる。本明細書では、明瞭にする目的で、様々な単数形／複数形の並べ換えが明示的に記載され得る。

【０１２９】

一般に、本明細書、特に添付の特許請求の範囲（例えば添付の特許請求の範囲の主文）において使用されている用語は、一般に「非限定」用語として意図されることが当業者には理解されるであろう（例えば、「含んでいる」という用語は、「含んでいるがそれらに限定されない」と解釈するべきであり、「有する」という用語は、「を少なくとも有する」と解釈するべきであり、「含む」という用語は、「含むがそれらに限定されない」と解釈するべきである）。更に、導入された請求項の特定の数の記載が意図される場合、そのような意図は請求項に明示的に記載されており、そのような記載がない場合、そのような意図は存在しないことが、当業者には理解されるであろう。例えば、１つの項目のみが意図される場合、「単一」という用語又は類似する言葉が使用され得る。理解を助けるために、以下の添付の特許請求の範囲及び／又は本明細書の説明は、請求項の記載を導入するために「少なくとも１つの」及び「１つ以上の」という導入句の使用を含み得る。しかしながら、このような句の使用は、不定冠詞「ａ」又は「ａｎ」による請求項の記載の導入が、そのような導入された請求項の記載を含む任意の特定の請求項を、１つのそのような記載のみを含む実施形態に制限することを意味するものと解釈すべきではなく、たとえ同じ請求項に、導入句「１つ以上の」又は「少なくとも１つの」及び「ａ」又は「ａｎ」などの不定冠詞が含まれていても同様である（例えば「ａ」及び／又は「ａｎ」は「少なくとも１つの」又は「１つ以上」を意味するものと解釈すべきである）。請求項の記載を導入するために使用される定冠詞の使用も同様である。加えて、導入された請求項の特定の数の記載が明示的に記載されている場合でも、かかる記載は少なくとも記載された数を意味するものと解釈されるべきであることが、当業者には認識されるであろう（例えば、他の修飾語なしの「２つの記載」という単純な記載は、少なくとも２つの記載、又は２つ以上の記載を意味する）。更に、「Ａ、Ｂ、及びＣのうちの少なくとも１つ」に類似する表記が使用される場合、一般に、そのような構造は、当業者がその表記を理解するであろう意味として意図される（例えば、「Ａ、Ｂ、及びＣのうちの少なくとも１つを有するシステム」は、Ａのみ、Ｂのみ、Ｃのみ、Ａ及びＢを一緒に、Ａ及びＣを一緒に、Ｂ及びＣを一緒に、並びに／又は、Ａ、Ｂ、及びＣを一緒に、有するシステムを含み、ただしこれらに限定されない）。「Ａ、Ｂ、又はＣのうちの少なくとも１つ」に類似する表記が使用される場合、一般に、そのような構造は、当業者がその表記を理解するであろう意味として意図される（例えば、「Ａ、Ｂ、又はＣのうちの少なくとも１つを有するシステム」は、

Ａのみ、Ｂのみ、Ｃのみ、Ａ及びＢを一緒に、Ａ及びＣを一緒に、Ｂ及びＣを一緒に、並びに／又は、Ａ、Ｂ、及びＣを一緒に、有するシステムを含み、ただしこれらに限定されない）。明細書、特許請求の範囲、又は図面のいずれにおいても、２つ以上の代替的な用語を提示する実質的に任意の離接的な語及び／又は句は、用語の一方、用語のいずれか、又は両方の用語を含む可能性を企図するものと理解されるべきであることが、当業者には更に理解されるであろう。例えば、「Ａ又はＢ」という句は、「Ａ」若しくは「Ｂ」又は「Ａ及びＢ」の可能性を含むものと理解されたい。更に、本明細書で使用される、複数の項目のリスト及び／又は複数の項目のカテゴリのリストが後ろに続く「～のいずれか」という用語は、項目及び／又は項目のカテゴリの、「のいずれか」、「の任意の組み合わせ」、「の任意の複数」、及び／又は「の任意の複数の組み合わせ」を、個別に、又は他の項目及び／又は他の項目のカテゴリとの組み合わせにおいて、含むことを意図している。更に、本明細書で使用される場合、「セット／組」又は「グループ／群」という用語は、ゼロを含む任意の数の項目を含むことが意図される。追加的に、本明細書で使用される、「数」という用語は、ゼロを含む任意の数を含むことを意図している。

【０１３０】

加えて、本開示の特徴又は態様がＭａｒｋｕｓｈ群の観点から説明されている場合、当業者には、本開示がそれによってＭａｒｋｕｓｈ群の任意の個々のメンバー又はメンバーのサブグループの観点からも説明されることが認識されるであろう。

【０１３１】

当業者には理解されるように、書面による説明を提供するという観点など、あらゆる目的のために、本明細書に開示される全ての範囲は、その任意の可能な部分範囲及び部分範囲の組み合わせも包含している。任意の列挙された範囲は、同じ範囲が、少なくとも等しい２分の１、３分の１、４分の１、５分の１、１０分の１などに分解されることを十分に説明して可能にするものとして、容易に認識することができる。非限定的な例として、本明細書に記載されている各範囲は、下位３分の１、中央の３分の１、及び上位３分の１などに容易に分解され得る。また、当業者には理解されるように、「まで」、「少なくとも」、「より大きい」、「より小さい」等の全ての言葉は、言及された数を含み、かつ、上述したように更に部分範囲に分解され得る範囲を意味する。最後に、当業者には理解されるように、範囲は個々の各要素を含む。したがって、例えば、１～３個のセルを有するグループは、１個、２個、又は３個のセルを有するグループを指す。同様に、１～５個のセルを有するグループは、１個、２個、３個、４個、又は５個のセルを有するグループを指し、以下同様である。

【０１３２】

更に、請求項は、特にそのように記載されない限り、提供された順序又は提供された要素に限定されるものとして読まれるべきではない。加えて、いかなる請求項においても、「ための手段」という用語の使用は、米国特許法第１１２条、第６項、又はミーンズプラスファンクションの請求項形式に訴えることを意図しており、「ための手段」という用語を有さないいかなる請求項もそのようには意図されていない。

【０１３３】

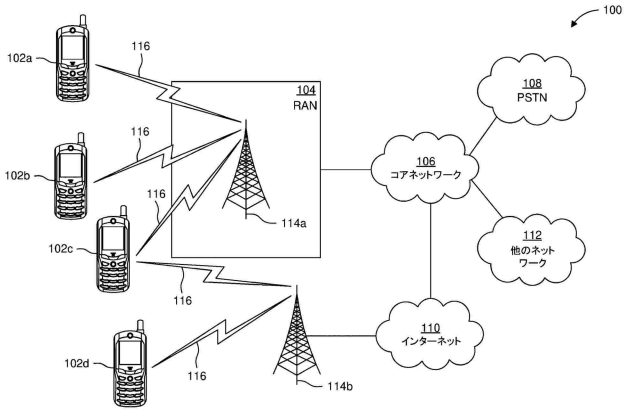
システムは、マイクロプロセッサ／汎用コンピュータ（図示せず）上のソフトウェアで実装され得ることが企図される。特定の実施形態では、様々な構成要素の機能のうちの１つ以上は、汎用コンピュータを制御するソフトウェアに実装され得る。

【０１３４】

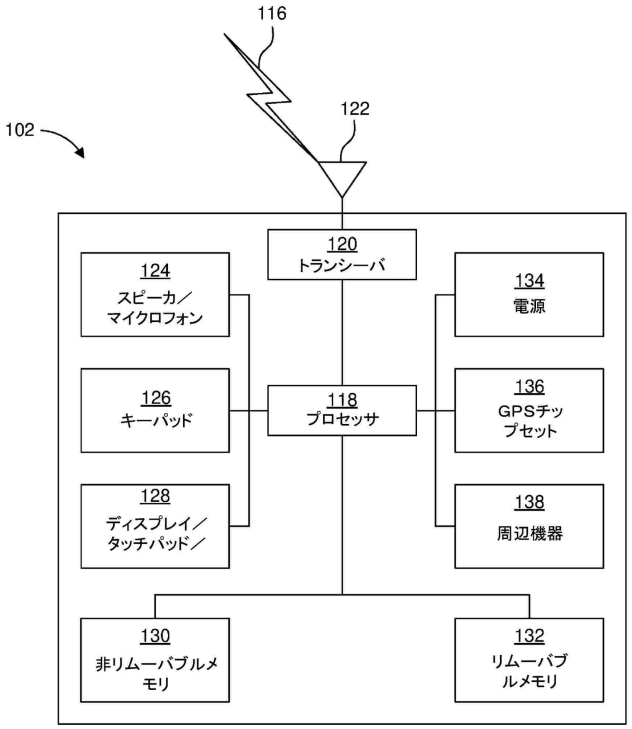
加えて、本発明は、特定の実施形態を参照して本明細書に例示及び説明されるが、本発明は、示された詳細に限定されることを意図していない。むしろ、特許請求の範囲及びその等価物の範囲内において、しかも本発明から逸脱することなく、詳細に様々な修正を行うことができる。

【図面】

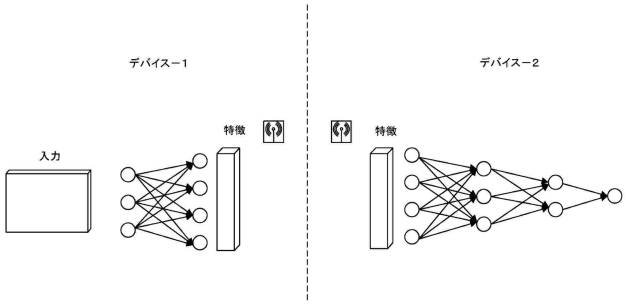
【図１Ａ】



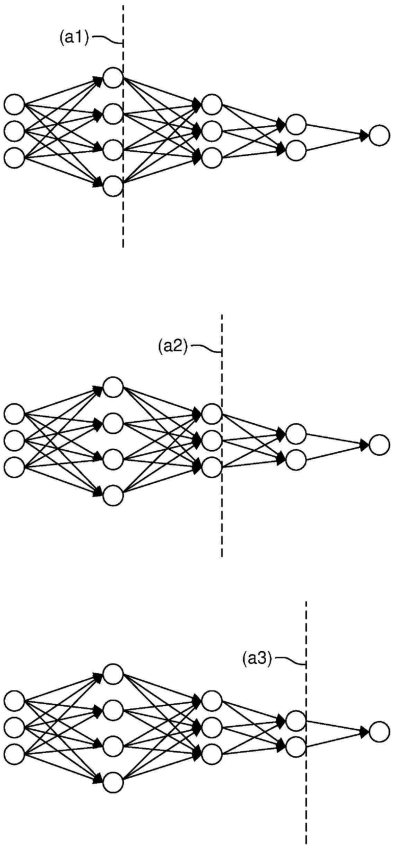
【図１Ｂ】



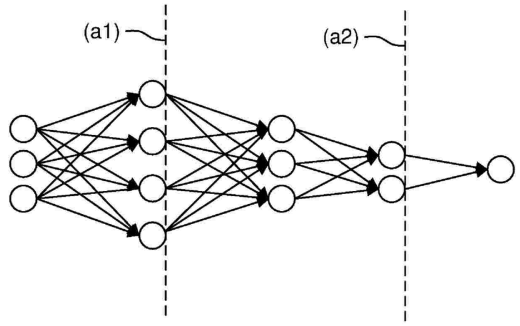
【図２】



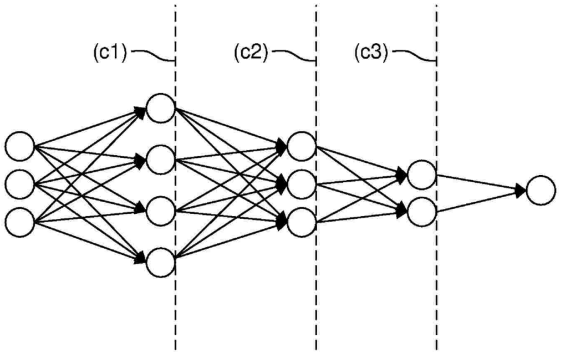
【図３Ａ】



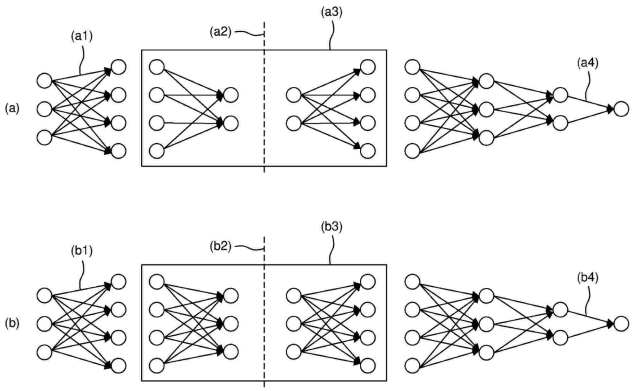
【図３Ｂ】



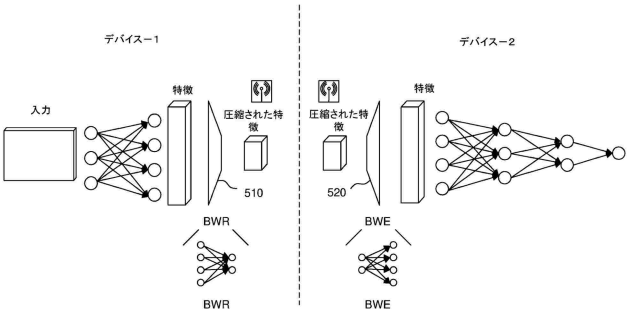
【図３Ｃ】



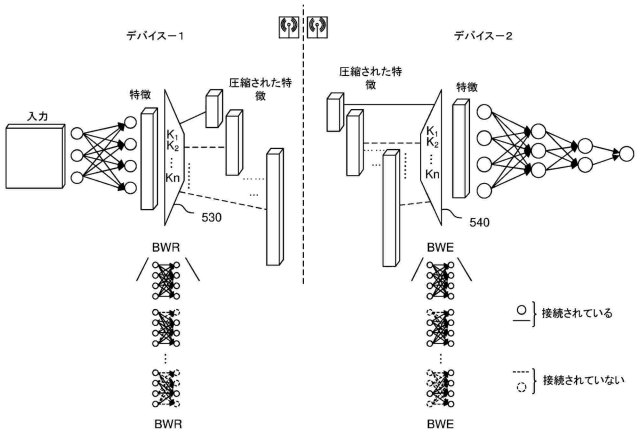
【図４】



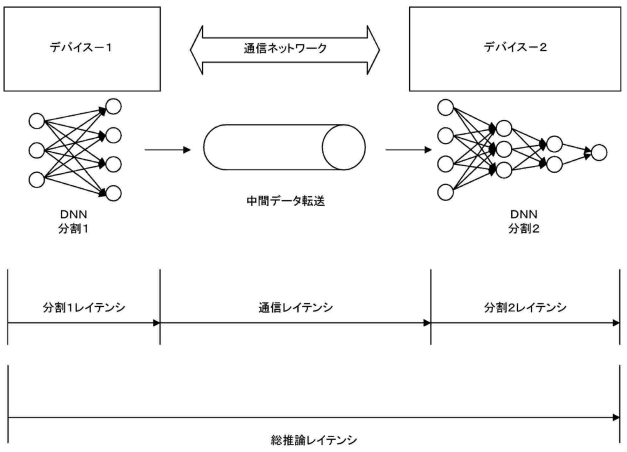
【図５Ａ】



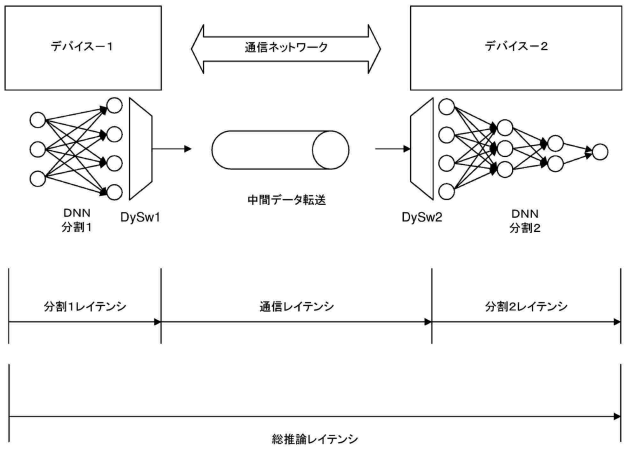
【図５Ｂ】



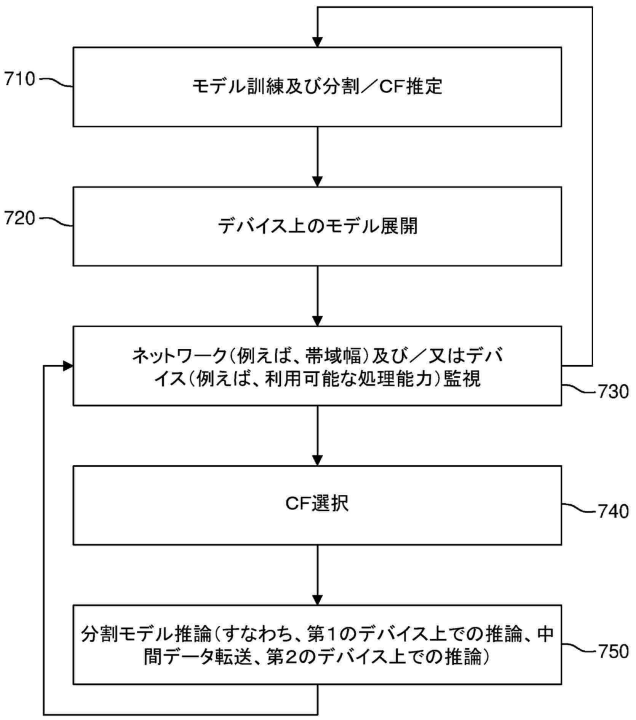
【図６Ａ】



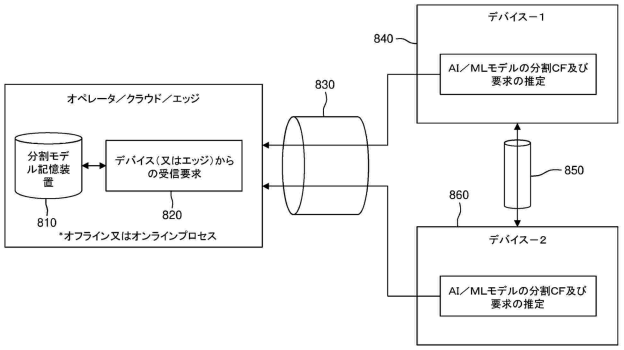
【図６Ｂ】



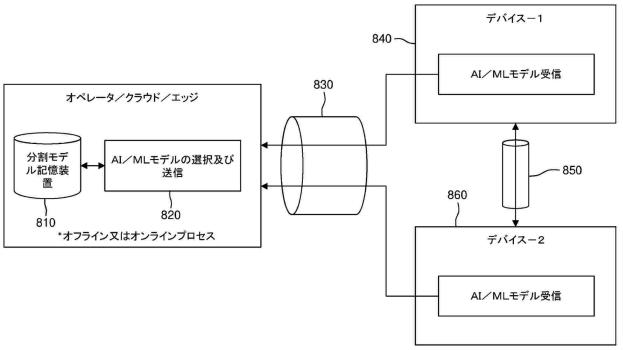
【図７】



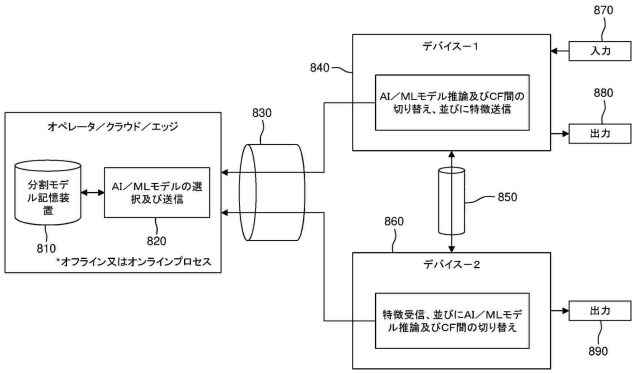
【図８Ａ】



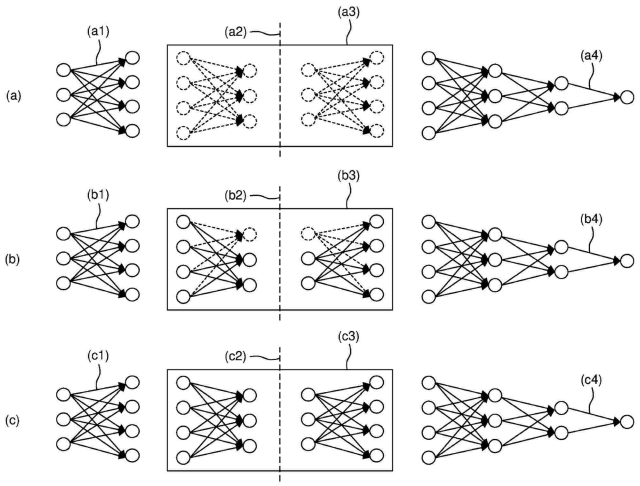
【図８Ｂ】



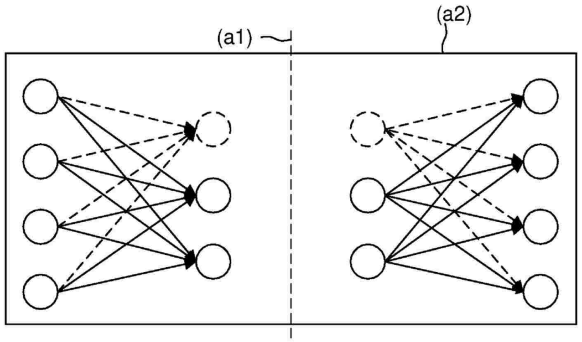
【図８Ｃ】



【図９】



【図１０】



【図１１】

出力ノード →	1	2	3	1	2	3	1	2	3
入力ノード－１	0	0	0	0	1	1	1	1	1
入力ノード－２	0	0	0	0	1	1	1	1	1
入力ノード－３	0	0	0	0	1	1	1	1	1
入力ノード－４	0	0	0	0	1	1	1	1	1
	(a)			(b)			(c)		

【手続補正書】
【提出日】令和5年9月25日(2023.9.25)
【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】
無線送信／受信ユニット（ＷＴＲＵ）であって、
ディープニューラルネットワーク（ＤＮＮ）モデルの一部であって、前記ＤＮＮモデルの分割点の前にあり、かつ前記ＤＮＮモデルの前記分割点における特徴を圧縮するためのニューラルネットワークを含む、前記ＤＮＮモデルの前記一部を受信するように構成された受信機と、
１つ以上のプロセッサであって、
前記ニューラルネットワークの圧縮係数を取得し、
前記圧縮係数に応答して、前記ニューラルネットワーク内のどのノードが接続されるべきかを決定し、
前記決定に応答して、前記ニューラルネットワークを構成し、かつ
圧縮された特徴を生成するために、前記ＤＮＮモデルの前記一部を用いて推論を実施するように構成されている、１つ以上のプロセッサと、
前記圧縮された特徴を別のＷＴＲＵに送信するように構成された送信機と、を備える、無線送信／受信ユニット（ＷＴＲＵ）。
【請求項２】
前記送信機が、取得された前記圧縮係数の指示を前記別のＷＴＲＵに送るように更に構成されている、請求項１に記載のデバイス。
【請求項３】
前記１つ以上のプロセッサが、前記圧縮係数が調整されるときに前記ネットワーク内のどのノードが接続されるべきかを決定するように構成されている、請求項１に記載のデバイス。
【請求項４】
前記分割点及び前記圧縮係数のうちの少なくとも１つが、（１）物理層動作、（２）メディアアクセス制御層動作、（３）無線リソース制御層動作、（４）利用可能な処理リソース、及び（５）制御シグナリングのうちの１つ以上に基づいて適合される、請求項１に記載のデバイス。
【請求項５】
前記分割点及び前記圧縮係数のうちの少なくとも１つが、送信データレートに基づいて適合される、請求項１に記載のデバイス。
【請求項６】
無線送信／受信ユニット（ＷＴＲＵ）によって実施される方法であって、
ディープニューラルネットワーク（ＤＮＮ）モデルの一部であって、前記ＤＮＮモデルの分割点の前にあり、かつ前記ＤＮＮモデルの前記分割点における特徴を圧縮するためのニューラルネットワークを含む、前記ＤＮＮモデルの前記一部を受信することと、
前記ニューラルネットワークの圧縮係数を取得することと、
前記圧縮係数に応答して、前記ニューラルネットワーク内のどのノードが接続されるべきかを決定することと、
前記決定に応答して、前記ニューラルネットワークを構成することと、
圧縮された特徴を生成するために、前記ＤＮＮモデルの前記一部を用いて推論を実施することと、
前記圧縮された特徴を別のＷＴＲＵに送信することと、を含む、方法。

【請求項７】
取得された前記圧縮係数の指示を前記別のＷＴＲＵに送ることを更に含む、請求項６に記載の方法。
【請求項８】
前記ネットワーク内のどのノードが接続されるべきかが、前記圧縮係数が調整されるときに決定される、請求項６に記載の方法。
【請求項９】
異なる圧縮係数に対して１つのＤＮＮモデルのみが前記デバイスにロードされる、請求項６に記載の方法。
【請求項１０】
前記分割点及び前記圧縮係数のうちの少なくとも１つが、（１）物理層動作、（２）メディアアクセス制御層動作、（３）無線リソース制御層動作、（４）利用可能な処理リソース、及び（５）制御シグナリングのうちの１つ以上に基づいて適合される、請求項６に記載の方法。
【請求項１１】
無線送信／受信ユニット（ＷＴＲＵ）であって、
ディープニューラルネットワーク（ＤＮＮ）モデルの一部であって、前記ＤＮＮモデルの分割点の後にあり、かつ前記ＤＮＮモデルの前記分割点における特徴を拡張するためのニューラルネットワークを含む、前記ＤＮＮモデルの前記一部を受信するように構成された受信機であって、別のＷＴＲＵから出力された１つ以上の特徴も受信するように構成されている、受信機と、
１つ以上のプロセッサであって、
前記ニューラルネットワークの圧縮係数を取得し、
前記圧縮係数に応答して、前記ニューラルネットワーク内のどのノードが接続されるべきかを決定し、
前記決定に応答して、前記ニューラルネットワークを構成し、かつ
別のＷＴＲＵから出力された前記１つ以上の特徴を前記ニューラルネットワークへの入力として使用して、前記ＤＮＮモデルの前記一部を用いて推論を実施するように構成されている、１つ以上のプロセッサと、を備える、無線送信／受信ユニット（ＷＴＲＵ）。
【請求項１２】
前記受信機が、前記圧縮係数を示す信号を受信するように更に構成されている、請求項１１に記載のデバイス。
【請求項１３】
前記１つ以上のプロセッサが、前記圧縮係数が調整されるときに前記ネットワーク内のどのノードが接続されるべきかを決定するように構成されている、請求項１１に記載のデバイス。
【請求項１４】
異なる圧縮係数に対して１つのＤＮＮモデルのみが前記デバイスにロードされる、請求項１１に記載のデバイス。
【請求項１５】
前記分割点及び前記圧縮係数のうちの少なくとも１つが、（１）物理層動作、（２）メディアアクセス制御層動作、（３）無線リソース制御層動作、（４）利用可能な処理リソース、及び（５）制御シグナリングのうちの１つ以上に基づいて適合される、請求項１１に記載のデバイス。
【請求項１６】
方法であって、
ディープニューラルネットワーク（ＤＮＮ）モデルの一部であって、前記ＤＮＮモデルの分割点の後にあり、かつ前記ＤＮＮモデルの前記分割点における特徴を拡張するためのニューラルネットワークを含む、前記ＤＮＮモデルの前記一部を受信することと、
別のＷＴＲＵから出力された１つ以上の特徴を受信することと、

前記ニューラルネットワークの圧縮係数を取得することと、
前記圧縮係数に応答して、前記ニューラルネットワーク内のどのノードが接続されるべきかを決定することと、
前記決定に応答して、前記ニューラルネットワークを構成することと、
別のＷＴＲＵから出力された前記１つ以上の特徴を前記ニューラルネットワークへの入力として使用して、前記ＤＮＮモデルの前記一部を用いて推論を実施することと、を含む、方法。
【請求項１７】
前記圧縮係数を示す信号を受信することを更に含む、請求項１６に記載の方法。
【請求項１８】
前記ネットワーク内のどのノードが接続されるべきかが、前記圧縮係数が調整されるときに決定される、請求項１６に記載の方法。
【請求項１９】
異なる圧縮係数に対して１つのＤＮＮモデルのみが前記デバイスにロードされる、請求項１６に記載の方法。
【請求項２０】
前記分割点及び前記圧縮係数のうちの少なくとも１つが、（１）物理層動作、（２）メディアアクセス制御層動作、（３）無線リソース制御層動作、（４）利用可能な処理リソース、及び（５）制御シグナリングのうちの１つ以上に基づいて適合される、請求項１６に記載の方法。

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2022/052633

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06N3/04 G06N3/063 G06N3/08 H03M7/30
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06N H04L H03M

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CONINCK ELIAS DE ET AL: "DIANNE: a modular framework for designing, training and deploying deep neural networks on heterogeneous distributed infrastructure", JOURNAL OF SYSTEMS & SOFTWARE, vol. 141, 1 July 2018 (2018-07-01), pages 52-65, XP055933056, US ISSN: 0164-1212, DOI: 10.1016/j.jss.2018.03.032 page 54, right-hand column - page 55, right-hand column Sections 4.3 and 5; page 58 - page 59 Section 6.3 ----- -/--	1-29

[x] Further documents are listed in the continuation of Box C. [x] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance;; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance;; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
6 July 2022	19/07/2022

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Jacobs, Jan-Pieter

Form PCT/ISA/210 (second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2022/052633

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2020/351509 A1 (LEE JOO-YOUNG [KR] ET AL) 5 November 2020 (2020-11-05) paragraph [0064] - paragraph [0077]; figures 1-3 paragraph [0078] - paragraph [0105]; figures 4-5 paragraph [0173] - paragraph [0195]; figures 11-12	1-29
A	LI EN ET AL: "Edge AI: On-Demand Accelerating Deep Neural Network Inference via Edge Computing", IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE SERVICE CENTER, PISCATAWAY, NJ, US, vol. 19, no. 1, 18 October 2019 (2019-10-18), pages 447-457, XP011766454, ISSN: 1536-1276, DOI: 10.1109/TWC.2019.2946140 [retrieved on 2020-01-07] page 450 - page 455	1-29
A	JIAWEI SHAO ET AL: "BottleNet++: An End-to-End Approach for Feature Compression in Device-Edge Co-Inference Systems", ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 5 June 2020 (2020-06-05), XP081680641, the whole document	1-29

2

Form PCT/ISA/210 (continuation of second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/EP2022/052633

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2020351509 A1	05-11-2020	KR 20190049552 A	09-05-2019
		US 2020351509 A1	05-11-2020

Form PCT/ISA/210 (patent family annex) (April 2005)

フロントページの続き

MK,MT,NL,NO,PL,PT,RO,RS,SE,SI,SK,SM,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,KM,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AO,AT,AU,AZ,BA,BB,BG,BH,BN,BR,BW,BY,BZ,CA,CH,CL,CN,CO,CR,CU,CZ,DE,DJ,DK,DM,DO,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,GT,HN,HR,HU,ID,IL,IN,IR,IS,IT,JO,JP,KE,KG,KH,KN,KP,KR,KW,KZ,LA,LC,LK,LR,LS,LU,LY,MA,MD,ME,MG,MK,MN,MW,MX,MY,MZ,NA,NG,NI,NO,NZ,OM,PA,PE,PG,PH,PL,PT,QA,RO,RS,RU,RW,SA,SC,SD,SE,SG,SK,SL,ST,SV,SY,TH,TJ,TM,TN,TR,TT,TZ,UA,UG,US,UZ,VC,VN,WS,ZA,ZM,ZW

（特許庁注：以下のものは登録商標）

１．ＷＣＤＭＡ

２．３ＧＰＰ

ンターデジタル宛

(72)発明者　ドゥオン、クアン カン ンゴック

フランス共和国 セソン－セヴィニェ ３５５７６，シーエス １７６１６，アヴェニュー デ シャン ブラン ９７５，インターデジタル宛

(72)発明者　オゼロフ、アレクセイ

フランス共和国 セソン－セヴィニェ ３５５７６，シーエス １７６１６，アヴェニュー デ シャン ブラン ９７５，インターデジタル宛

(72)発明者　フォンテーヌ、パトリック

フランス共和国 セソン－セヴィニェ ３５５７６，シーエス １７６１６，アヴェニュー デ シャン ブラン ９７５，インターデジタル宛

(72)発明者　シュニッツラー、フランソワ

フランス共和国 セソン－セヴィニェ ３５５７６，シーエス １７６１６，アヴェニュー デ シャン ブラン ９７５，インターデジタル宛

(72)発明者　ランベール、アンヌ

フランス共和国 セソン－セヴィニェ ３５５７６，シーエス １７６１６，アヴェニュー デ シャン ブラン ９７５，インターデジタル宛

(72)発明者　ペレティエ、ギスラン

カナダ ケベック州 エイチ３エー ３ジー４ モントリオール，シェルブルック　ストリート １００，インターデジタル カナダ リミテー宛

Ｆターム（参考）　5K067　EE02 HH21