

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7674507号
(P7674507)

(45)発行日 令和7年5月9日(2025.5.9)

(24)登録日 令和7年4月28日(2025.4.28)

(51)国際特許分類

F I

G 0 6 N 3/045(2023.01)

G 0 6 T 7/00 (2017.01)

G 0 6 V 10/82 (2022.01)

G 0 6 N 3/045

G 0 6 T 7/00 3 5 0 C

G 0 6 V 10/82

請求項の数 15 (全31頁)

(21)出願番号	特願2023-556541(P2023-556541)	(73)特許権者	502208397
(86)(22)出願日	令和4年5月27日(2022.5.27)		グーグル エルエルシー
(65)公表番号	特表2024-518232(P2024-518232 A)		G o o g l e L L C
(43)公表日	令和6年5月1日(2024.5.1)		アメリカ合衆国 カリフォルニア州 9 4
(86)国際出願番号	PCT/US2022/031466		0 4 3 マウンテン ビュー アンフィシ
(87)国際公開番号	WO2022/251717		アター パークウェイ 1 6 0 0
(87)国際公開日	令和4年12月1日(2022.12.1)		1 6 0 0 Amphitheatre P
審査請求日	令和5年11月1日(2023.11.1)		arkway 9 4 0 4 3 Mounta
(31)優先権主張番号	63/194,926	(74)代理人	in View, CA U. S. A.
(32)優先日	令和3年5月28日(2021.5.28)		100108453
(33)優先権主張国・地域又は機関	米国(US)	(74)代理人	弁理士 村山 靖彦
		(74)代理人	100110364
			弁理士 実広 信哉
		(74)代理人	100133400
			弁理士 阿部 達彦

最終頁に続く

(54)【発明の名称】 エキスパートの混合を使用した画像処理

(57)【特許請求の範囲】

【請求項1】

入力画像を処理し、前記入力画像に関するネットワーク出力を生成するように構成されるニューラルネットワークを含むシステムであって、前記ニューラルネットワークが、1つまたは複数のネットワークブロックのシーケンスを含み、前記1つまたは複数のネットワークブロックが、

前記入力画像の中間表現を表すブロック入力を取得することと、

前記ブロック入力の、または前記ブロック入力の更新された表現の、複数のパッチを決定することであって、各パッチが、前記ブロック入力の、または前記ブロック入力の前記更新された表現の要素の異なる部分集合を含む、ことと、

前記ネットワークブロックの複数のエキスパートモジュールのうちの1つまたは複数のそれぞれのエキスパートモジュールに各パッチを割り振ることと、

前記複数のパッチの各パッチに関して、1つまたは複数のそれぞれのモジュール出力を生成するために1つまたは複数の対応するエキスパートモジュールを使用して前記パッチを処理することと、

前記それぞれのエキスパートモジュールによって生成された前記モジュール出力を組み合わせることによってブロック出力を生成することと、

1つまたは複数のそれぞれのエキスパートモジュールに各パッチを割り振ることの前に、1つまたは複数の自己アテンションニューラルネットワーク層を使用して、各パッチを更新するために前記複数のパッチにわたって自己アテンションメカニズムを適用するように

前記複数のパッチを処理することを含む動作を実行するようにそれぞれ構成される、システム。

【請求項 2】

前記ネットワークブロックの1つまたは複数のそれぞれのエキスパートモジュールに各パッチを割り振ることが、

前記複数のエキスパートモジュールの各エキスパートモジュールに関して、それぞれのルーティングスコアを生成するためにルーティングサブネットワークを使用して前記パッチを処理することと、

前記複数のルーティングスコアのうちの1つまたは複数の最も高いルーティングスコアを決定することと、

前記1つまたは複数の最も高いルーティングスコアの各々に対応する前記エキスパートモジュールに前記パッチを割り振ることを含む、請求項1に記載のシステム。

【請求項 3】

入力画像を処理し、前記入力画像に関するネットワーク出力を生成するように構成されるニューラルネットワークを含むシステムであって、前記ニューラルネットワークが、1つまたは複数のネットワークブロックのシーケンスを含み、前記1つまたは複数のネットワークブロックが、

前記入力画像の中間表現を表すブロック入力を取得することと、

前記ブロック入力の、または前記ブロック入力の更新された表現の、複数のパッチを決定することであって、各パッチが、前記ブロック入力の、または前記ブロック入力の前記更新された表現の要素の異なる部分集合を含む、ことと、

前記ネットワークブロックの複数のエキスパートモジュールのうちの1つまたは複数のそれぞれのエキスパートモジュールに各パッチを割り振ることと、

前記複数のパッチの各パッチに関して、1つまたは複数のそれぞれのモジュール出力を生成するために1つまたは複数の対応するエキスパートモジュールを使用して前記パッチを処理することと、

前記それぞれのエキスパートモジュールによって生成された前記モジュール出力を組み合わせることによってブロック出力を生成することを含む動作を実行するようにそれぞれ構成される、システムであって、

前記複数のエキスパートモジュールの各エキスパートモジュールに関して、それぞれのルーティングスコアを生成するためにルーティングサブネットワークを使用して前記パッチを処理することが、

各エキスパートモジュールのそれぞれの第1の中間ルーティングスコアを生成するために、1つまたは複数の順伝播型ニューラルネットワーク層を使用して前記パッチを処理すること、

各エキスパートモジュールに関して、ノイズ値をサンプリングし、前記エキスパートモジュールの第2の中間ルーティングスコアを生成するために前記エキスパートモジュールの前記第1の中間ルーティングスコアに前記ノイズ値を加えること、

各エキスパートモジュールに関して、前記エキスパートモジュールの前記ルーティングスコアを生成するために、前記エキスパートモジュールの前記第2の中間ルーティングスコアに非線形活性化関数を適用することを含む、システム。

【請求項 4】

入力画像を処理し、前記入力画像に関するネットワーク出力を生成するように構成されるニューラルネットワークを含むシステムであって、前記ニューラルネットワークが、1つまたは複数のネットワークブロックのシーケンスを含み、前記1つまたは複数のネットワークブロックが、

前記入力画像の中間表現を表すブロック入力を取得することと、

前記ブロック入力の、または前記ブロック入力の更新された表現の、複数のパッチを決定することであって、各パッチが、前記ブロック入力の、または前記ブロック入力の前記更新された表現の要素の異なる部分集合を含む、ことと、

10

20

30

40

50

前記ネットワークブロックの複数のエキスパートモジュールのうちの1つまたは複数のそれぞれのエキスパートモジュールに各パッチを割り振ることと、

前記複数のパッチの各パッチに関して、1つまたは複数のそれぞれのモジュール出力を生成するために1つまたは複数の対応するエキスパートモジュールを使用して前記パッチを処理することと、

前記それぞれのエキスパートモジュールによって生成された前記モジュール出力を組み合わせることによってブロック出力を生成することを含む動作を実行するようにそれぞれ構成される、システムであって、

前記複数のパッチの各パッチに関して、1つまたは複数のそれぞれのモジュール出力を生成するために前記1つまたは複数の対応するエキスパートモジュールを使用して前記パッチを処理することが、

10

前記パッチに対応する前記1つまたは複数のエキスパートモジュールの各々に関して、それぞれの初期モジュール出力を生成するために前記エキスパートモジュールを使用して前記パッチを処理することと、

前記パッチに対応する前記1つまたは複数のエキスパートモジュールの各々に関して、前記エキスパートモジュールおよび前記パッチに対応するルーティングスコアに従って、対応する初期モジュール出力をスケールリングすることを含む、システム。

【請求項5】

入力画像を処理し、前記入力画像に関するネットワーク出力を生成するように構成されるニューラルネットワークを含むシステムであって、前記ニューラルネットワークが、1つまたは複数のネットワークブロックのシーケンスを含み、前記1つまたは複数のネットワークブロックが、

20

前記入力画像の中間表現を表すブロック入力を取得することと、

前記ブロック入力の、または前記ブロック入力の更新された表現の、複数のパッチを決定することであって、各パッチが、前記ブロック入力の、または前記ブロック入力の前記更新された表現の要素の異なる部分集合を含む、ことと、

前記ネットワークブロックの複数のエキスパートモジュールのうちの1つまたは複数のそれぞれのエキスパートモジュールに各パッチを割り振ることと、

前記複数のパッチの各パッチに関して、1つまたは複数のそれぞれのモジュール出力を生成するために1つまたは複数の対応するエキスパートモジュールを使用して前記パッチを処理することと、

30

前記それぞれのエキスパートモジュールによって生成された前記モジュール出力を組み合わせることによってブロック出力を生成することを含む動作を実行するようにそれぞれ構成される、システムであって、

前記ニューラルネットワークが、

各エキスパートモジュールに対応するルーティングスコアが中心傾向の同じ測定値を有するように促す重要度損失、または

各エキスパートモジュールがパッチを割り振られるそれぞれの見込みが中心傾向の同じ測定値を持つように促す負荷損失のうちの1つまたは複数を使用して訓練される、システム。

40

【請求項6】

一団の訓練画像Xに関する前記重要度損失 $L_{imp}(X)$ を決定することが、

各エキスパートモジュールiに関して、

【数1】

$$\text{imp}_i(X) = \sum_{x \in X} f_i(x)$$

を計算することであって、式中、 f_i が、前記エキスパートモジュールiの前記ルーティングスコアまたは前記エキスパートモジュールiの中間ルーティングスコアを計算するための、

50

ルーティングサブネットワークによって定義される関数である、計算することと、

各エキスパートモジュール*i*に関して、 $\text{imp}_i(X)$ が中心傾向の同じ測定値を有するように促すために前記重要度損失 $\text{Limp}(X)$ を計算することを含む、請求項5に記載のシステム。

【請求項7】

前記重要度損失 $\text{Limp}(X)$ が、

【数2】

$$\mathcal{L}_{\text{imp}}(X) = \left(\frac{\text{std}(\text{imp}_{1:E}(X))}{\text{mean}(\text{imp}_{1:E}(X))} \right)^2$$

10

であり、式中、*E*が、前記複数のエキスパートモジュールの数である、請求項6に記載のシステム。

【請求項8】

各パッチが、*k*個の異なるエキスパートモジュールに割り振られ、 $k \geq 1$ であり、

一団の訓練画像*X*に関する前記負荷損失 $\text{Lload}(X)$ を決定することが、

前記一団の訓練画像*X*の中の各訓練画像*x*に関して、

【数3】

$$\text{threshold}_k(x) = \max_{(k)} g_i(x)$$

20

を計算することであって、式中、 $g_i(x)$ が、前記エキスパートモジュール*i*の前記ルーティングスコアまたは前記エキスパートモジュール*i*の中間ルーティングスコアを計算するための、ルーティングサブネットワークによって定義される関数である、計算することと、

各訓練画像*x*に関しておよび各エキスパートモジュール*i*に関して、

$$p_i(x) = P(g_i(x) \geq \text{threshold}_k(x))$$

を計算することであって、式中、 $P(e)$ が、事象*e*が発生する見込みである、計算することと、

各エキスパートモジュール*i*に関して、

【数4】

$$\text{load}_i(X) = \sum_{x \in X} p_i(x)$$

30

を計算し、各エキスパートモジュール*i*に関して、 $\text{load}_i(x)$ が中心傾向の同じ測定値を有するように促すために前記負荷損失 $\text{Lload}(X)$ を計算することを含む、請求項5に記載のシステム。

【請求項9】

前記負荷損失 $\text{Lload}(X)$ が、

【数5】

$$\mathcal{L}_{\text{load}}(X) = \left(\frac{\text{std}(\text{load}_{1:E}(X))}{\text{mean}(\text{load}_{1:E}(X))} \right)^2$$

40

であり、式中、*E*が、前記複数のエキスパートモジュールの数である、請求項8に記載のシステム。

【請求項10】

前記複数のパッチの各パッチに関して、前記対応するエキスパートモジュールを使用して前記パッチを処理することが、前記エキスパートモジュールにわたって並列に前記パッ

50

チを処理することを含む、請求項1に記載のシステム。

【請求項11】

前記複数のエキスパートモジュールの数が、前記複数のパッチの数よりも多い、請求項1に記載のシステム。

【請求項12】

ネットワークブロックの前記シーケンスの最後のネットワークブロックの前記ブロック出力が前記ネットワーク出力であるか、またはネットワークブロックの前記シーケンスの前記最後のネットワークブロックの前記ブロック出力が、前記ネットワーク出力を生成するために1つまたは複数のニューラルネットワーク層を使用してさらに処理される、請求項1に記載のシステム。

【請求項13】

前記ネットワーク出力が、複数のカテゴリの各々に対応するそれぞれのスコアを含む分類出力であるか、

前記ネットワーク出力が、前記入力画像内の1つもしくは複数の要素の各々に関して、複数のカテゴリの各々に対応するそれぞれのスコアを含む要素レベルの分類出力であるか、

前記ネットワーク出力が、前記入力画像を特徴付ける1つもしくは複数の連続変数を推定する回帰出力であるか、または

前記ニューラルネットワークが、ビデオフレームを特徴付ける出力を生成するために、前記ビデオフレームである複数の画像を処理するように構成される、請求項1に記載のシステム。

【請求項14】

請求項1から13のいずれか一項に記載のシステムのネットワークブロックによって実行される動作を含む方法。

【請求項15】

1つまたは複数のコンピュータによって実行されるときに前記1つまたは複数のコンピュータに請求項1から13のいずれか一項に記載の動作を実行させる命令を記憶する1つまたは複数の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】

関連出願の相互参照

本出願は、2021年5月28日に出願した米国仮出願第63/194,926号の合衆国法典第35編(米国特許法)第119条の下での優先権の利益を主張するものであり、この仮出願の全体は、参照により本明細書に組み込まれている。

【0002】

本明細書は、ニューラルネットワークを使用して画像を処理することに関する。

【背景技術】

【0003】

ニューラルネットワークは、受け取られた入力に関する出力を予測するために非線形ユニットの1つまたは複数の層を使用する機械学習モデルである。一部のニューラルネットワークは、出力層に加えて1つまたは複数の隠れ層を含む。各隠れ層の出力は、ネットワーク内の次の層、すなわち、次の隠れ層または出力層への入力として使用される。ネットワークの各層は、パラメータのそれぞれの集合の現在の値に従って、受け取られた入力から出力を生成する。

【発明の概要】

【課題を解決するための手段】

【0004】

本明細書は、ニューラルネットワークを使用して画像を処理し、画像を特徴付けるネットワーク出力を生成するように構成される、1つまたは複数の場所の1つまたは複数のコンピュータ上のコンピュータプログラムとして実装されるシステムを説明する。ニューラル

10

20

30

40

50

ネットワークは、画像または画像の中間表現を含むブロック入力を処理し、ブロック出力を生成するようにそれぞれ構成される1つまたは複数のネットワークブロックのシーケンスを含む。

【0005】

ネットワークブロックの少なくとも1つは、ブロック入力の要素の真の部分集合を処理するようにそれぞれ構成される複数の異なるエキスパートモジュールの集合を含むエキスパートネットワークブロックである。特に、エキスパートネットワークブロックは、ブロック入力の要素の複数の「パッチ」を特定し、各パッチに関して、パッチをルーティングするエキスパートモジュールの真の部分集合を特定するためにルーティングモジュールを使用してパッチを処理するように構成される。それから、各エキスパートモジュールは、エキスパートモジュールにルーティングされた各パッチに関して、ルーティングされたパッチに関するそれぞれのエキスパートモジュール出力を生成するためにパッチを処理する。エキスパートネットワークブロックは、各パッチに関して、パッチの処理に応答してそれぞれのエキスパートモジュールによって生成された対応するエキスパートモジュール出力を組み合わせて、それぞれのパッチ出力を生成することができる。そして、エキスパートネットワークブロックは、ブロック入力の各パッチに対応するパッチ出力と一緒に構成してブロック出力を生成することができ、パッチ出力は、ブロック入力において対応するパッチが有していたのと同じ空間的関係をブロック出力において互いに有する。

【0006】

本明細書に記載の主題は、以下の利点のうちの1つまたは複数を実現するように特定の実施形態に実装され得る。

【0007】

一部の既存のシステムは、「高密度(dense)」順伝播型ニューラルネットワーク、すなわち、順伝播型ニューラルネットワークの各ニューラルネットワーク層に関して、ニューラルネットワーク層のあらゆるネットワークパラメータを使用して、ニューラルネットワーク層への層入力のあらゆる要素を処理する順伝播型ニューラルネットワークを使用して画像を処理する。高解像度の画像を処理するように構成された大規模なモデルを用いると、そのような高密度ニューラルネットワークは、単一の画像を処理するために多大な時間および計算リソースを必要とし得る。本明細書において説明される技術を使用すると、システムは、ネットワーク入力に基づいてニューラルネットワークのパラメータの部分集合を選択的にアクティブ化することによって、順伝播型ニューラルネットワークを使用して画像を処理し、画像の処理の時間および計算効率を大幅に改善することができる。さらに、このスパース性(sparsity)を導入することは、任意の所与の入力を処理するためにパラメータの部分集合のみが使用されるので、順伝播型ニューラルネットワークが以前実現可能だったよりもずっと多くのネットワークパラメータを含むことを可能にし得る。

【0008】

本明細書において説明される技術を使用すると、システムは、各エキスパートモジュールが異なる種類の画像(または異なる種類の画像パッチ)を処理するように訓練によって構成されたような複数のエキスパートモジュールを持つエキスパートネットワークブロックを含むニューラルネットワークを実装することができ、モジュールが「特化する」ことを可能にし、ニューラルネットワークの効率および性能をさらに向上させることができる。

【0009】

本明細書において説明されるように、画像を処理するように構成された自己アテンションに基づくニューラルネットワークは、最先端の畳み込みニューラルネットワークと同じ性能を達成するためにはるかに少ない計算を必要とし得る。別の言い方をすれば、決まった計算バジェット(compute budget)に関して、自己アテンションに基づくニューラルネットワークは、画像に対して同じ機械学習タスクを実行するように構成された畳み込みニューラルネットワークよりも優れた(たとえば、テストの正解率、適合率、再現率などによって測定される)性能を発揮する。これは、自己アテンションメカニズムが畳み込みよりも少ない計算で画像の異なる領域に注意を向けることができるので、自己アテンションを適

10

20

30

40

50

用する方が画像全体にカーネルを畳み込むよりも概して計算効率が高いからである。

【0010】

本明細書の主題の1つまたは複数の実施形態の詳細が、添付の図面および以下の説明に記載されている。主題のその他の特徴、態様、および利点は、説明、図面、および特許請求の範囲から明らかになるであろう。

【図面の簡単な説明】

【0011】

【図1A】画像についての予測を生成するように構成される例示的なニューラルネットワークの図である。

【図1B】複数のエキスパートモジュールを含む例示的なニューラルネットワークブロックの図である。

10

【図2】画像パッチにセグメンテーションされた例示的な画像を示す図である。

【図3】複数のエキスパートモジュールを含むネットワークブロックを実行するための例示的なプロセスの流れ図である。

【発明を実施するための形態】

【0012】

様々な図面における同様の参照番号および参照指示は、同様の要素を示す。

【0013】

本明細書は、ジャストインタイムコンパイラおよびコンパイルキャッシュ(compilation cache)を使用してコンピュータプログラムモジュールをコンパイルするように構成される、1つまたは複数の場所の1つまたは複数のコンピュータ上のコンピュータプログラムとして実装されるシステムを説明する。

20

【0014】

図1Aは、例示的なニューラルネットワーク100の図である。ニューラルネットワーク100は、以下で説明されるシステム、コンポーネント、および技術が実装され得る、1つまたは複数の場所の1つまたは複数のコンピュータ上のコンピュータプログラムとして実装されたシステムの一例である。

【0015】

ニューラルネットワーク100は、入力画像102の埋め込みまたは入力画像102についての予測を表すネットワーク出力108を生成するために、訓練されたネットワークパラメータの集合に従って入力画像102を処理するように構成される。ニューラルネットワーク100は、入力画像102を使用して任意の適切な機械学習タスクを実行するように構成され得る。例示的な機械学習タスクは、下で検討される。本明細書において、画像を処理することは、画像の画素の強度(intensity)値を処理することを指す。

30

【0016】

ニューラルネットワーク100は、L個のネットワークブロック110a~lのシーケンスを含み、 $L \geq 1$ である。各ネットワークブロック110a~lは、画像または画像の中間表現を含むブロック入力を処理し、ブロック出力を生成するように構成され得る。

【0017】

図1Aに示されるように、ネットワークブロック110a~lのシーケンスの第1のネットワークブロック110aは、画像102の中間表現である第1のブロック出力104を生成するために画像102自体を処理するように構成される。そして、後続の各ネットワークブロック110b~lは、シーケンスの前のネットワークブロックのブロック出力を処理することができる。すなわち、第(i+1)のネットワークブロック110a~nのブロック入力は、第iのネットワークブロック110a~lのブロック出力である。

40

【0018】

一部のその他の実装形態において、第1のネットワークブロック110aは、入力画像102自体の代わりに、入力画像102の埋め込みを処理するように構成される。たとえば、以下の説明は、概して、ニューラルネットワークが画像102を処理することに言及するが、一部の实装形態において、ニューラルネットワーク100は、画像102の埋め込み、たとえば

50

、画像102を処理することに対応して異なるニューラルネットワークによって生成された埋め込みを入力として受け取るように構成され得る。

【0019】

一部の実装形態において、ニューラルネットワーク100のネットワーク出力108は、ネットワークブロック110a~lのシーケンスの最後のネットワークブロック110lのブロック出力である。たとえば、下でより詳細に説明されるように、最後のネットワークブロック110lのブロック出力は、入力画像102の複数のパッチの各々のそれぞれの更新された表現を含むことができ、したがって、ブロック出力は、集合的に、たとえば、1つまたは複数の下流の機械学習モデルによって記憶されるか、またはさらに処理され得る入力画像102の埋め込みを表し得る。

10

【0020】

一部のその他の実装形態において、シーケンスの最後のネットワークブロック110a~lのブロック出力は、ニューラルネットワーク100のネットワーク出力108を生成するために、1つまたは複数のニューラルネットワーク層を使用してさらに処理される。たとえば、ニューラルネットワーク100が入力画像102を分類するように構成される場合、ニューラルネットワーク100は、候補クラスの集合の各クラスに関して、入力画像102がクラスに属する見込み(likelihood)を表す値を特定するネットワーク出力108を生成するために、最後のロジスティック層(logistic layer)またはソフトマックス層を含む1つまたは複数のニューラルネットワーク層を使用して、シーケンスの最後のネットワークブロック110a~lのブロック出力を処理することができる。

20

【0021】

特定の例として、ニューラルネットワーク100は、それぞれの異なる機械学習タスクのために入力画像102についての予測を生成するために、1つまたは複数のニューラルネットワーク層を使用してシーケンスの最後のネットワークブロック110a~lのブロック出力を処理するようにそれぞれが構成される1つまたは複数の「ヘッド(head)」サブネットワークを含み得る。つまり、ニューラルネットワーク100が訓練された後、(入力画像102の埋め込みとみなされ得る)シーケンスの最後のネットワークブロック110a~lの出力は、複数の異なる機械学習タスクに有用な情報を符号化することができる。

【0022】

一部の実装形態において、ニューラルネットワーク100は、まず、入力画像102を処理して入力画像のパッチの集合を決定し、たとえば、入力画像102の決定されたパッチを表す複数の要素を含む入力シーケンスを決定する。それから、ニューラルネットワーク100は、ネットワークブロック110a~lのシーケンスを使用して、決定されたパッチ(あるいはニューラルネットワーク100の1つもしくは複数の初期ニューラルネットワーク層によって、または異なるニューラルネットワークによって生成された決定されたパッチの埋め込み)を処理することができる。一部のその他の実装形態において、ネットワークブロック110a~lのうちの1つまたは複数の、それぞれ、(入力画像102の中間表現である)ネットワークブロック110a~lへのブロック入力のパッチの集合を決定し、決定されたパッチを処理して、ネットワークブロック110a~lのブロック出力を生成する。ネットワークブロック110a~lの動作は、下でより詳細に説明される。

30

40

【0023】

一部の実装形態において、ニューラルネットワーク100は、入力画像102に加えて、1つまたは複数のその他のネットワーク入力を処理する。たとえば、ニューラルネットワーク100は、入力画像102が属するクラスを特定するデータを処理することができる。別の例として、ニューラルネットワーク100は、入力画像102に対応するテキスト入力、たとえば、入力画像102を特徴付けるテキストを表すテキスト入力を処理することができる。

【0024】

別の例として、たとえば、ニューラルネットワーク100が、ネットワークブロック110a~lのシーケンスによって連続的に更新される入力画像102のパッチ(またはそれらの埋め込み)のシーケンスを決定する実装形態において、ニューラルネットワーク100は、「クラ

50

ス」トークンと呼ばれることがある所定のトークンを処理することができる。すなわち、所定のトークンは、ニューラルネットワーク100によって受け取られた各入力画像102に関して同じであり、たとえば、所定のテンソルであることが可能である。たとえば、所定のトークンは、たとえば、ニューラルネットワーク100の訓練と同時に機械学習され得る。特定の例として、所定のトークンは、ネットワークブロック110a~lのシーケンスの各ネットワークブロックが所定のトークンの更新された表現を生成するように、ニューラルネットワーク100によって処理されるパッチのシーケンスに付加され得る。一部の実装形態において、ネットワーク出力108は、最後のネットワークブロック110lによって生成された所定のトークンの最後の表現であり(またはそのような最後の表現を含み)、最後の表現は、入力画像102の埋め込みとみなされ得る。

10

【0025】

ネットワークブロック110a~lの一部またはすべては、複数の異なるエキスパートモジュールのそれぞれの集合を含む。本明細書において、エキスパートモジュールは、ネットワークブロックへのブロック入力の要素の真の部分集合、すなわち、要素の一部を処理するように構成されるネットワークブロックのコンポーネントである。エキスパートモジュールは、要素の真の部分集合の更新された表現を生成するために、1つまたは複数のニューラルネットワーク層を使用して、ブロック入力の要素の真の部分集合を処理するように構成され得る。本明細書においては、複数のエキスパートモジュールを含むネットワークブロックは、エキスパートネットワークブロックとも呼ばれる。

【0026】

20

たとえば、シーケンスの各ネットワークブロック110a~lが、エキスパートネットワークブロックであることが可能である。別の例として、シーケンスの1つおきのネットワークブロック110a~lが、エキスパートネットワークブロックであることが可能である。別の例として、 $i < l$ として特定のネットワークブロック110iまでの、シーケンスの1つおきのネットワークブロック110a~lが、エキスパートネットワークブロックであることが可能である。

【0027】

たとえば、図1に示されるように、ネットワークブロック110a~lのシーケンスの第jのネットワークブロック110jは、e個のエキスパートモジュール130a~eを含むエキスパートネットワークブロックであり、 $e > 1$ である。以下の説明は、特に第jのネットワークブロック110jに言及するが、同じ技術がニューラルネットワーク100の各エキスパートネットワークブロックのために実施され得ることを理解されたい。

30

【0028】

第jのネットワークブロックは、そのブロック入力112を複数のパッチにセグメンテーションすることができ、各パッチは、ブロック入力112の要素の異なる部分集合を含む。たとえば、ブロック入力112が入力画像102の3次元表現である場合、第jのネットワークブロック110jは、各パッチが次元 $L \times W \times C$ を有すると決定することができ、Cは、入力画像102の表現のチャンネル数を表す(たとえば、ブロック入力112がRGB画像である場合、 $C = 3$ であり、ブロック入力112がRGB画像の中間表現である場合、より大きな数である)。

【0029】

40

一部の实装形態において、ブロック入力112の各パッチは、同じサイズであり、すなわち、同じ数の要素を含み、同じ次元を有する。一部のその他の実装形態において、異なるパッチは、異なるサイズであることが可能であり、たとえば、異なる数の要素を含むことが可能である。

【0030】

一部の实装形態において、ブロック入力112の各要素は、ちょうど1つのパッチ内にある。一部のその他の実装形態において、ブロック入力112の要素の一部またはすべては、複数の異なるパッチ内にあることが可能であり、すなわち、ブロック入力112のパッチが、重なることが可能である。

【0031】

50

一部の実装形態において、ネットワークブロックのシーケンスの各ネットワークブロック110a~lは、入力画像102の同じ集合のパッチを処理する。すなわち、各ブロック入力およびブロック出力は、各ネットワークブロック110a~lによって決定されるパッチの数、パッチのサイズ、およびパッチの空間的關係が同じであるように、同じ次元を有することができる。たとえば、ニューラルネットワーク100は、入力画像102を処理して、入力画像102のパッチの入力シーケンスを決定し、それから、ネットワークブロックのシーケンスの各ネットワークブロック110a~lを使用して、パッチの決定されたシーケンスのそれぞれの更新された表現を処理することができる。

【0032】

ブロック入力112のパッチは、任意の適切な方法で表され得る。たとえば、各パッチは、ブロック入力112の要素を含む2次元テンソル、たとえば、ブロック入力112の要素の空間的關係を維持するテンソルとして表され得る。別の例として、各パッチは、パッチの要素の1次元シーケンスとして表され得る。特定の例として、パッチがブロック入力112の2次元領域である場合、パッチは、たとえば、ブロック入力112の要素がブロック入力112内のそれらのそれぞれの位置に従って1次元テンソル内で順序付けされる、2次元領域の平坦化されたバージョンであることが可能である。別の例として、各初期画像パッチ112a~nは、初期画像パッチ112a~nの画素の無順序集合(unordered set)として表され得る。

【0033】

画像のパッチまたは画像の中間表現を決定するための例示的な技術は、図2を参照して下で検討される。

【0034】

ブロック入力112のパッチを決定した後、第jのネットワークブロック110jは、ルーティングモジュール120を使用して、各パッチを第jのネットワークブロック110jのエキスパートモジュール130a~eのうちの1つまたは複数に割り振ることができる。そして、第jのネットワークブロックは、対応する割り振られたエキスパートモジュール130a~eを使用して各画像パッチを処理して、それぞれのエキスパートモジュール出力を生成し、組合せモジュール140を使用してエキスパートモジュール130a~eのエキスパートモジュール出力を組み合わせて、第jのネットワークブロック110jのブロック出力114を生成することができる。

【0035】

言い換えると、ルーティングモジュール120は、各エキスパートモジュール130a~eに関して、処理のためにエキスパートモジュール130a~eにルーティングされるべきブロック入力112のパッチの対応する集合122a~eを決定するように構成される。それから、各エキスパートモジュール130a~eは、パッチの割り振られた集合122a~eを処理して、エキスパートモジュール出力の対応する集合132a~e、すなわち、それぞれの割り振られたパッチに関するそれぞれのエキスパートモジュール出力を生成する。エキスパートモジュール出力は、単に「モジュール出力」とも呼ばれ得る。

【0036】

パッチをそれぞれのエキスパートモジュール130a~eに割り振るために、ルーティングモジュール120は、1つまたは複数のニューラルネットワーク層を含むルーティングニューラルネットワークを使用して各パッチを処理することができる。特定のパッチの処理に応答して、ルーティングニューラルネットワークは、第jのネットワークブロック110jの各エキスパートモジュール130a~eのそれぞれのルーティングスコアを含む出力を生成することができる。ルーティングモジュール120は、パッチを、パッチの最も高いルーティングスコアに対応する1つまたは複数のエキスパートモジュール130a~eに割り振ると決定することができる。

【0037】

一部の実装形態において、各パッチは、同じ数のエキスパートモジュール130a~eに割り振られる。すなわち、各パッチは、パッチの最も高いルーティングスコアを有するk個の

10

20

30

40

50

エキスパートモジュール130a～eに割り振られ得、kは、すべてのパッチに関して同じである。

【0038】

一部のその他の実装形態において、一部のパッチが、異なる数のエキスパートモジュールに割り振られ得る。たとえば、エキスパートモジュール130a～eおよびパッチのルーティングスコアが所定の閾値を満たす場合、パッチは、エキスパートモジュール130a～eに割り振られ得る。別の例として、各エキスパートモジュール130a～eは、エキスパートモジュールおよびそれぞれのパッチに対応するすべてのルーティングスコアのうち、エキスパートモジュールに対応する最も高いルーティングスコアを有するp個のパッチを割り振ることによって、ちょうどp個のパッチを処理することを保証され得、 $p \geq 1$ である。

10

【0039】

一部の实装形態において、第jのネットワークブロック110jのあらゆるエキスパートモジュール130a～eが、ブロック入力112のパッチを処理するわけではなく、すなわち、一部の入力画像102に関して、第jのネットワークブロック110jの一部のエキスパートモジュール130a～eは、入力画像102から生成されたブロック入力112に関してはアイドルしている。ニューラルネットワーク100の一部の実行中に一部のエキスパートモジュール130a～eがアイドルしていることを可能にすることは、ニューラルネットワーク100のあらゆるパラメータが各入力画像102を処理するために使用されるわけではないので、ネットワーク出力108を生成するために必要とされる計算の数を減らすことによってニューラルネットワーク100の効率を向上させることができる。

20

【0040】

ルーティングモジュール120のルーティングニューラルネットワークは、訓練を通じて、パッチから最も多くの情報を抽出するエキスパートモジュール130a～eに各パッチを割り振るように構成され得る。すなわち、一部の实装形態において、エキスパートモジュール130a～eの少なくとも一部は、特定の種類のパッチ、たとえば、特定のセマンティックオブジェクト(semantic object)またはセマンティックオブジェクトのカテゴリを描写するパッチに「特化」することができる。ニューラルネットワーク100を訓練するための例示的な技術は、下でより詳細に検討される。

【0041】

一部の实装形態において、第jのネットワークブロック110jは、どの1つのエキスパートモジュール130a～eもその他のエキスパートモジュールに比して不釣り合いな量の作業(すなわち、計算)を割り振られ、ひいては、ニューラルネットワーク100の実行のボトルネックになることがないように、各エキスパートモジュール130a～eが処理することができるパッチの数を制限する。たとえば、ネットワークブロック110jは、いずれのエキスパートモジュール130a～eに割り振られ得るパッチの数も

30

【0042】

【数1】

$$P_e = C \frac{KBP}{E}$$

40

【0043】

以下になるように制限することができる。

【0044】

式中、Bは、ニューラルネットワーク100によって処理されている入力画像102の一団(たとえば、推論時にニューラルネットワーク100に提供され、ニューラルネットワーク100によって並列に処理される画像102の一団)の中の画像の数であり、Kは、各パッチが割り振られるエキスパートモジュール130a～eの数であり、Pは、ブロック入力112内のパッチの数であり、Eは、ネットワークブロック110jのエキスパートモジュール130a～eの数である。Cは、各エキスパートモジュール130a～eの能力を表すニューラルネットワー

50

ク100のハイパーパラメータであることが可能であり、 $C=1$ は、各エキスパートモジュール130a～eが、ニューラルネットワーク100の所与の実行中にパッチの

【0045】

【数2】

$$\frac{1}{E}$$

【0046】

を処理することのみを許されることを示す。

10

【0047】

一部の実装形態において、ニューラルネットワーク100は、たとえば、 $C \ll 1$ 、たとえば、 $C=0.5$ 、 0.1 、 0.05 、または 0.01 を設定する。

【0048】

一部の実装形態において、許されたよりも多くの画像パッチがエキスパートモジュール130a～eに割り振られる場合、制限を超えるパッチ(たとえば、エキスパートモジュール130a～eに対応する最も低いルーティングスコアを有するエキスパートモジュール130a～eに割り振られたパッチ)は、たとえば、各パッチに関して、次に高いルーティングスコアを持つエキスパートモジュール130a～eに再度割り振られる。一部のその他の実装形態において、制限を超えるパッチは再度割り振られず、むしろ、パッチがその他のパッチよりも1つ少ないエキスパートモジュール130a～eによって処理されるように「ドロップされ(dropped)」、これは、単に恒等関数(identity function)であるエキスパートモジュールを使用して、ドロップされたパッチを処理することと等価である。

20

【0049】

一部の実装形態において、ニューラルネットワーク100は、ニューラルネットワーク100の異なる実行中に C の異なる値で動作し、つまり、 C の値は、ニューラルネットワーク100が訓練された後に変更され得る。たとえば、ニューラルネットワーク100の訓練中、 C の値は、比較的高いことが可能であり、したがって、比較的少ないパッチがドロップされ、訓練システムが、エキスパートモジュール130a～eのモジュール出力を使用して、エキスパートモジュール130a～eのニューラルネットワーク層のネットワークパラメータの更新を決定することができる。そして、推論時に、 C の値は、ニューラルネットワーク100の時間および計算効率を改善するために下げられ得る。特定の例として、ニューラルネットワーク100は、乏しいリソースを持つ推論環境、たとえば、比較的少ない計算リソースを有するモバイル電話もしくはタブレットなどのデバイス上、またはニューラルネットワーク100が新しい画像102を継続的に受け取り、決まった時間バジェット(time budget)を使用して、対応するネットワーク出力108を生成しなければならない環境にデプロイされ得る。

30

【0050】

一部の実装形態において、ブロック入力112の各パッチに関して、ルーティングモジュール120は、各エキスパートモジュール130a～eのノイズ値を(たとえば、正規分布から)ランダムにサンプリングし、サンプリングされたノイズ値を、パッチの k 個の最も高いルーティングスコアを決定する前に、対応するルーティングスコアに加える。たとえば、ルーティングモジュール120は、ニューラルネットワーク100の訓練中にノイズ値を加えて多様性を促進し、したがって、各エキスパートモジュール130a～eがパラメータの更新を受け取るように促すことができる。ルーティングモジュール120は、高品質のルーティングスコアを割り振るように訓練されたので、推論時には、ノイズ値を加えないように構成され得る。

40

【0051】

一部の実装形態において、ネットワークブロックは、 k 個の最も高いルーティングスコアを決定する前に、非線形活性化関数、たとえば、ソフトマックス、Tanh、またはReLU関

50

数をルーティングスコアに適用することができる。

【0052】

ルーティングニューラルネットワークは、ニューラルネットワーク層の任意の適切な構成を含み得る。

【0053】

たとえば、ルーティングニューラルネットワークは、1つまたは複数の順伝播型ニューラルネットワーク層を含み得る。特定の例として、パッチが次元 $L \times W \times C$ を有する場合、ルーティングモジュール120は、次元 $1 \times (L \cdot W \cdot C)$ を有するようにパッチを再形成し、再形成されたパッチを、順伝播型ニューラルネットワーク層を使用して処理することができる。

【0054】

代わりにまたは加えて、ルーティングニューラルネットワークは、1つまたは複数の畳み込みニューラルネットワーク層を含み得る。特定の例として、パッチが次元 $L_1 \times W_1 \times C$ を有する場合、ルーティングニューラルネットワークは、次元 $L_2 \times W_2 \times C$ を有する畳み込みカーネルを使用してパッチに対して2次元畳み込みを実行することができ、 $L_1 > L_2$ および $W_1 > W_2$ である。

【0055】

代わりにまたは加えて、ルーティングニューラルネットワークは、1つまたは複数の自己アテンション層を含み得る。自己アテンションは、図1Bを参照して下でより詳細に検討される。

【0056】

一部の実装形態において、 e 個のエキスパートモジュール130a~ e の動作は並列に実行され、したがって、ニューラルネットワーク100の時間効率を改善する。

【0057】

一部の实装形態において、各エキスパートモジュール130a~ e は、同じ処理デバイス、たとえば、それぞれの異なるスレッド、たとえば、グラフィックス処理ユニット(GPU)またはテンソル処理ユニット(TPU)上でエキスパートモジュール130a~ e の少なくとも一部を実行するように構成される並列処理ハードウェア上で実行される。一部のその他の実装形態において、それぞれのエキスパートモジュール130a~ e は、異なる処理デバイス、たとえば、複数の異なる並列処理デバイスの分散システムによって実行され得る。エキスパートモジュール130a~ e の少なくとも一部を並列に実行することは、ニューラルネットワーク100の実行のレイテンシを大幅に、たとえば、すべてのエキスパートモジュール130a~ e が異なるデバイス上で実行される場合はおよそ e 分の1に削減することができる。したがって、所与の数のネットワークパラメータに関して、エキスパートモジュールの少なくとも1つのエキスパートネットワークブロックを含むニューラルネットワークは、(i)入力画像の各要素がエキスパートモジュールのネットワークパラメータの真の部分集合によって処理されるため、かつ(ii)一部の实装形態において、高密度ニューラルネットワーク層は並列化され得ないが、エキスパートモジュールの実行は並列化され得るため、高密度ニューラルネットワーク層のみを持つニューラルネットワークに比べて削減されたレイテンシを享受することができる。同様に、同じ理由で、特定の時間バジェットが与えられた場合(すなわち、特定のレイテンシの閾値に関して)、エキスパートモジュールの少なくとも1つのエキスパートネットワークブロックを含むニューラルネットワークは、高密度ニューラルネットワーク層のみを持つニューラルネットワークと比較して、著しく多いネットワークパラメータと、ひいては、増大したネットワーク容量(すなわち、より多くの情報を符号化する能力)とを含むことができる。

【0058】

エキスパートモジュール130a~ e が複数のデバイスにわたって分散される一部の实装形態において、ルーティングモジュール120を実行する処理デバイスは、特定のエキスパートモジュール130a~ e を実行する処理デバイスに、特定のエキスパートモジュール130a~ e に割り振られたパッチの集合122a~ e のみを提供するように構成される。一部のその

10

20

30

40

50

他のそのような実装形態において、ルーティングモジュール120を実行する処理デバイスは、特定のエキスパートモジュール130a～eを実行する処理デバイスに、ブロック入力112のどのパッチが特定のエキスパートモジュール130a～eによって処理されるべきかを特定するデータと一緒に、ブロック入力112全体を提供するように構成される。

【0059】

各パッチが、それぞれのモジュール出力を生成するために、対応する1つまたは複数のエキスパートモジュールによって処理された後、ネットワークブロック110jの組合せモジュール140は、各パッチに対応するモジュール出力を組み合わせ、ネットワークブロック110jのブロック出力114を生成することができる。特に、組合せモジュール140は、ブロック入力112の各パッチに関して、パッチの処理に応答してそれぞれのエキスパートモジュール130a～eによって生成された対応するモジュール出力を組み合わせ、パッチのパッチ出力を生成することができる。それから、組合せモジュール140は、パッチ出力が、ブロック入力112において、対応するパッチが有していたのと同じ空間的関係をブロック出力114において互いに有するようにブロック出力114を生成するためにパッチ出力を再構成することができる。

【0060】

たとえば、一部の实装形態において、パッチ出力は、対応するパッチと同じ次元を有する。したがって、組合せモジュールは、ブロック入力112と同じ次元 $L \times W \times C$ を有するブロック出力114を生成するためにパッチ出力を構成することができ、各パッチ出力は、ブロック入力112において、対応するパッチが占めたのと同じ位置をブロック出力114において占める。

【0061】

一部の实装形態において、組合せモジュール140は、ブロック入力の各パッチに関して、パッチのモジュール出力の合計を計算することによって、対応するモジュール出力を組み合わせる。一部のそのような実装形態において、各パッチのパッチ出力を生成するために、組合せモジュール140は、パッチと、モジュール出力を生成したエキスパートモジュール130a～eとに対応するルーティングスコアを使用して、各モジュール出力を重み付けすることができる。一部のその他の実装形態において、組合せモジュール140は、1つまたは複数のニューラルネットワーク層、たとえば、1つまたは複数の自己アテンション層、あるいは1つまたは複数のリカレントニューラルネットワーク層を使用してモジュール出力を処理することによって、エキスパートモジュール130a～eのモジュール出力を組み合わせる。

【0062】

一部の实装形態において、ブロック入力112のパッチをそれぞれのエキスパートモジュール130a～eに割り振る前に、ネットワークブロック110jは、まず、1つまたは複数のニューラルネットワーク層を使用してブロック入力を処理して、ブロック入力102の更新された表現を生成し、それから、ブロック入力102の更新された表現のパッチをそれぞれのエキスパートモジュール130a～eに割り振る。たとえば、ネットワークブロック110jは、まず、ブロック入力102の各パッチを処理するように構成される1つもしくは複数の自己アテンションニューラルネットワーク層および/または1つもしくは複数の要素ごとの順伝播型ニューラルネットワーク層を使用して、ブロック入力を処理することができる。言い換えると、エキスパートモジュールへの入力、ブロック入力102自体の、ではなく、ブロック入力102の更新された表現の、要素の真の部分集合であることが可能である。概して、本明細書において、我々はブロック入力102のパッチに言及するが、同じ手順がブロック入力102の更新された表現のパッチに適用され得ることを理解されたい。同様に、ルーティングモジュール120に先立つ1つまたは複数のニューラルネットワーク層は、ネットワークブロック110a～lのシーケンスの異なるネットワークブロックの一部であるとみなされ得る。

【0063】

第jのネットワークブロック110jのブロック出力114は、ネットワークブロック110a～

lのシーケンスの後続のネットワークブロックへのブロック入力として提供され得る。

【0064】

上述のように、一部の実装形態において、シーケンスの最後のネットワークブロック110lのブロック出力は、ネットワーク出力108である。たとえば、最後のネットワークブロック110lのブロック出力が、入力画像102の複数のパッチの各々のそれぞれの表現を含む場合、ネットワーク出力108は、各パッチの表現を含み得る。別の例として、ネットワーク出力108は、各パッチの表現から生成された組み合わせられた表現、たとえば、パッチの表現の平均を決定することによって、またはプーリング動作(たとえば、平均プーリング、最大プーリング、もしくはグローバルプーリング(global pooling))を適用することによって生成された組み合わせられた表現を含み得る。

10

【0065】

別の例として、上述のように、一部の实装形態において、ニューラルネットワーク100は、ネットワークブロックのシーケンスの各ネットワークブロックが、パッチの更新された表現に加えて、所定のクラストークンの更新された表現を生成するように、入力画像102と一緒に所定のクラストークン进行处理する。これらの実装形態において、ネットワーク出力108は、最後のネットワークブロック110lによって生成された所定のクラストークンの最後の表現であることが可能である。

【0066】

ニューラルネットワーク100は、任意の適切な技術を使用して訓練され得る。たとえば、訓練システムが、ニューラルネットワーク100を使用して訓練画像を処理してそれぞれの訓練出力を生成し、たとえば、訓練画像に対応するグラウンドトゥルース(ground-truth)出力を使用して訓練出力の誤差を決定することができる。次に、トレーニングシステムは、ニューラルネットワーク100を介して誤差をバックプロパゲートし、勾配降下を使用してニューラルネットワーク100のパラメータの更新を決定することができる。

20

【0067】

代わりにまたは加えて、訓練システムは、ネットワークブロック110a~lのシーケンスの各エキスパートネットワークブロックに関して、エキスパートネットワークブロック内の各エキスパートモジュール130a~eに対応するルーティングスコアが同じ平均を有するように促す「重要度損失(importance loss)」を使用してニューラルネットワーク100を訓練することができる。たとえば、各エキスパートモジュールiに関して、訓練システムは、

30

【0068】

【数3】

$$\text{imp}_i(X) = \sum_{x \in X} f_i(x)$$

【0069】

を計算することができる。

【0070】

40

式中、 f_i は、エキスパートモジュールiのルーティングスコアまたはエキスパートモジュールiの中間ルーティングスコアを計算するための、エキスパートネットワークブロックのルーティングモジュール120のルーティングニューラルネットワークによって定義される関数である。たとえば、 $f_i(x)$ は、ランダムノイズが加えられる前および非線形活性化関数が適用される前の、エキスパートモジュールiおよびパッチxのルーティングスコアであることが可能である。そして、訓練システムは、各エキスパートモジュールiに関して、 $\text{imp}_i(X)$ が同じ平均を有するように促すために重要度損失 $L_{\text{imp}}(X)$ を計算することができる。特定の例として、訓練システムは、

【0071】

【数4】

50

$$\mathcal{L}_{\text{imp}}(X) = \left(\frac{\text{std}(\text{imp}_{1:E}(X))}{\text{mean}(\text{imp}_{1:E}(X))} \right)^2$$

【0072】

を計算することができる。

【0073】

式中、Eは、エキスパートモジュールの数である。

【0074】

代わりにまたは加えて、訓練システムは、各エキスパートネットワークブロックに関し
ておよび各パッチに関して、エキスパートネットワークブロックの各エキスパートモジ
ュール130a～eがパッチを割り振られる見込みが同じ平均を有するように促す「負荷損失(load loss)」を使用してニューラルネットワーク100を訓練することができる。たとえば、
各エキスパートモジュールiに関しておよび各画像xに関して、訓練システムは、

【0075】

【数5】

$$\text{threshold}_k(x) = \max_{(k)} g_i(x)$$

【0076】

を計算することができる。

【0077】

式中、 $g_i(x)$ は、エキスパートモジュールiのルーティングスコアまたはエキスパートモ
ジュールiの中間ルーティングスコアを計算するための、エキスパートネットワークブロッ
クのルーティングモジュール120のルーティングニューラルネットワークによって定義さ
れる関数である。たとえば、 $g_i(x)$ は、ランダムノイズが加えられた後、しかし、非線形活
性化関数が適用される前の、エキスパートモジュールiおよびパッチxのルーティングスコ
アであることが可能である。そして、訓練システムは、各訓練画像xに関しておよび各エキ
スパートモジュールiに関して、

$$p_i(x) = P(g_i(x) \geq \text{threshold}_k(x))$$

を計算することができる。

【0078】

式中、 $P(e)$ は、事象eが発生する見込みである。ランダムノイズの分布が与えられると
、 $p_i(x)$ は、直接計算可能である。それから、各エキスパートモジュールiに関して、訓練
システムは、

【0079】

【数6】

$$\text{load}_i(X) = \sum_{x \in X} p_i(x)$$

【0080】

を計算し、

【0081】

各エキスパートモジュールiに関して、 $\text{load}_i(x)$ が同じ平均を有するように促すために負
荷損失 $L_{\text{load}}(X)$ を計算することができる。特定の例として、訓練システムは、

【0082】

【数7】

10

20

30

40

50

$$\mathcal{L}_{\text{load}}(X) = \left(\frac{\text{std}(\text{load}_{1:E}(X))}{\text{mean}(\text{load}_{1:E}(X))} \right)^2$$

【0083】

を計算することができる。

【0084】

式中、Eは、エキスパートモジュールの数である。

【0085】

一部の実装形態において、ニューラルネットワーク100は、ニューラルネットワーク100のネットワークパラメータの初期値を生成するために、第1の機械学習タスクを使用して事前に訓練され得る。そして、初期値は、第2の機械学習タスクを使用して、たとえば、それぞれの第2の機械学習タスクの最終的な予測を生成するためにネットワーク出力108を処理するようにそれぞれ構成された1つまたは複数のヘッドニューラルネットワークと同時に微調整(すなわち、更新)され得る。たとえば、訓練システムは、たとえば、ラベル付けされていない画像をそれらの類似性に基づいて異なるクラスにセグメンテーションするようにニューラルネットワーク100を訓練するか、またはラベル付けされていない画像の教師なしセマンティックセグメンテーション(semantic segmentation)を実行するようにニューラルネットワーク100を訓練するために、ラベル付けされていない画像を使用して、教師なしまたは自己教師ありの方法でニューラルネットワーク100を事前に訓練することができる。それから、訓練システムは、特定のタスク、たとえば、物体認識、姿勢検出、顔認識、オプティカルフロー追跡、または任意のその他の適切な機械学習タスクを実行するために、ラベル付けされた画像のタスク固有のデータセットを使用して、教師ありの方法で(たとえば、新しい訓練されていないヘッドニューラルネットワークと一緒に)ニューラルネットワーク100のネットワークパラメータを微調整することができる。

【0086】

画像102は、任意の適切な種類の画像であることが可能である。たとえば、画像102は、2次元画像、たとえば、複数のチャンネルを有する2次元画像(たとえば、RGB画像)であることが可能である。別の例として、画像102は、たとえば、画像102の各画素に関して、スペクトル上の分布を特定することによって、波長の連続スペクトルを表すハイパースペクトル画像であることが可能である。別の例として、画像102は、複数の点を含む点群であることが可能であり、各点は、たとえば、3次元またはより高次元の座標空間内のそれぞれの座標を有し、特定の例として、画像102は、LIDARセンサによって生成された点群であることが可能である。別の例として、画像102は、医用撮像デバイスによって生成された医用画像であることが可能であり、特定の例として、画像102は、コンピュータ断層撮影法(CT)画像、磁気共鳴画像法(MRI)画像、超音波画像、X線画像、マンモグラム画像、蛍光透視法画像、または陽電子放射断層撮影法(PET)画像であることが可能である。

【0087】

下の説明は、それぞれが画像102のそれぞれの「画素」を含む画像102の画像パッチを生成することに言及するが、ニューラルネットワークシステム100は、任意の適切な種類の画像102の構成要素を含む画像パッチを生成し得ることを理解されたい。たとえば、画像102が点群である場合、画像102の各画像パッチは、点群の点の部分集合を含み得る。別の例として、画像102が3次元ボクセルグリッド内の複数のボクセルを含むMRI画像である場合、画像102の各画像パッチは、ボクセルグリッド内のボクセルの部分集合を含み得る。

【0088】

ニューラルネットワークは、任意の適切な画像処理タスク、たとえば、分類タスク、回帰タスク、またはそれらの組合せを実行するために画像を処理するように構成され得る。

【0089】

10

20

30

40

50

特定の例として、ニューラルネットワーク100は、複数のカテゴリの各々に対応するそれぞれのスコアを含む分類出力を生成するように構成され得る。あるカテゴリのスコアは、画像がそのカテゴリに属する見込みを示す。場合によっては、カテゴリは、物体のクラス(たとえば、犬、猫、人など)である場合があり、画像は、あるカテゴリに対応する物体クラスに含まれる物体を描写する場合、そのカテゴリに属する場合がある。場合によっては、カテゴリは、大域的な画像特性(たとえば、画像が昼の情景を描写するのかもしれないか、または画像が夏の情景を描写するのかもしれないか、または画像が冬の情景を描写するのかもしれない)を表す場合があり、画像は、カテゴリに対応する大域的な特性を有する場合、そのカテゴリに属する場合がある。

【0090】

10

別の特定の例として、ニューラルネットワーク100は、画像内の各要素に関して、複数のカテゴリの各々に対応するそれぞれのスコアを含む要素レベルの分類出力(たとえば、RGB画像の画素レベルの分類出力またはLIDAR画像の点レベルの分類出力)を生成するように構成され得る。所与の要素(たとえば、所与の画素または点)に関して、あるカテゴリのスコアは、要素がそのカテゴリに属する見込みを示す。場合によっては、カテゴリは、物体のクラスである場合があり、ある要素は、あるカテゴリに対応する物体クラスに含まれる物体の一部である場合、そのカテゴリに属する場合がある。つまり、要素レベルの分類出力は、セマンティックセグメンテーション出力である場合がある。

【0091】

20

別の特定の例として、ニューラルネットワーク100は、画像を特徴付ける1つまたは複数の連続変数(すなわち、無限に多くの可能な数値を想定し得る)を推定する回帰出力を生成するように構成され得る。特定の例において、回帰出力は、画像に描写されたそれぞれの物体を囲むバウンディングボックスの座標を推定する場合がある。バウンディングボックスの座標は、バウンディングボックスの頂点の(x, y)座標によって定義される場合がある。

【0092】

一部の実装形態において、ニューラルネットワーク100は、複数の画像、たとえば、ビデオの複数のフレームを処理するように構成され得る。たとえば、ニューラルネットワークは、ビデオのビデオフレームである複数の画像を受け取ることができ、たとえば、人が特定のアクションを実行するところをビデオフレームが描写するかどうかを特徴付けることによって、ビデオフレームを特徴付ける出力を生成するために、上述のように各ビデオフレームを処理することができる。

30

【0093】

一部のそのような実装形態において、ニューラルネットワーク100は、それぞれの異なる時点で各ビデオフレームを処理して、ビデオフレームに関する予測を特徴付ける各ビデオフレームのそれぞれのネットワーク出力を生成する。たとえば、ニューラルネットワークは、ビデオフレームの分類を予測するネットワーク出力を生成することができる。一部のそのような実装形態において、ニューラルネットワークは、それぞれのビデオフレームに対応する複数のネットワーク出力を組み合わせ、ビデオを特徴付ける最終的なネットワーク出力を生成する。たとえば、ニューラルネットワークは、下流のニューラルネットワーク、たとえば、リカレントニューラルネットワークを使用して、それぞれのネットワーク出力を処理することができる。

40

【0094】

一部のその他の実装形態において、ニューラルネットワーク100は、ビデオを特徴付ける単一のネットワーク出力を生成するために、各ビデオフレームを並列に処理する。特定の例として、システムは、各ビデオフレームに関して入力シーケンス内の1つまたは複数のそれぞれの入力要素を生成することができる。

【0095】

図1Bは、複数のエキスパートモジュール180を含む例示的なニューラルネットワークブロック150の図である。ニューラルネットワークブロック150は、以下で説明されるシス

50

テム、コンポーネント、および技術が実装され得る、1つまたは複数の場所の1つまたは複数のコンピュータ上のコンピュータプログラムとして実装されたシステムの一例である。

【0096】

ネットワークブロック150は、入力画像を特徴付けるネットワーク出力、たとえば、入力画像の埋め込みを表すか、または入力画像についての予測を表すネットワーク出力を生成するために入力画像を処理するように構成されたニューラルネットワークのL個のネットワークブロックのシーケンスのうちの1つであることが可能であり、 $L \geq 1$ である。たとえば、ネットワークブロック150は、図1Aを参照して上で説明されたネットワークブロック110a~lのシーケンスのネットワークブロックのうちの1つであることが可能である。

【0097】

ネットワークブロック150は、入力画像の埋め込まれたパッチ152の集合、すなわち、ネットワークブロックのシーケンスの1つまたは複数の先行するネットワークブロックによって生成された、入力画像から決定されたパッチの更新された表現を処理して、ブロック出力156を生成するように構成される。ネットワークブロック150がシーケンスの最初のネットワークブロックである場合、ネットワークブロック150は、入力画像から決定された元のパッチを処理するように構成され得る。

【0098】

ネットワークブロック150は、1つまたは複数の自己アテンションニューラルネットワーク層を含む。図1Bに示されるように、ネットワークブロック150は、単一の自己アテンションニューラルネットワーク層160を含む。一部の实装形態において、自己アテンションニューラルネットワーク層160は、埋め込まれたパッチ152を取得し、埋め込まれたパッチ152にわたってアテンションメカニズムを適用するように構成される。一部のその他の実装形態において、自己アテンションニューラルネットワーク層160は、埋め込まれたパッチ152のそれぞれの処理されたバージョンを取得し、処理された埋め込まれたパッチ152にアテンションメカニズムを適用するように構成される。たとえば、図1Bに示されるように、ネットワークブロック150は、層正規化層154の出力を自己アテンションニューラルネットワーク層160に提供する前に、まず、埋め込まれたパッチ152に層正規化層154を適用することができる。代わりにまたは加えて、ネットワークブロック150は、自己アテンションニューラルネットワーク層160の前に、1つまたは複数のその他のニューラルネットワーク層、たとえば、1つまたは複数の要素ごとの順伝播型ニューラルネットワーク層を、埋め込まれたパッチ152に適用することができる。

【0099】

たとえば、ネットワークブロック150は、それぞれの埋め込まれたパッチ152に関して、埋め込まれたパッチ152の要素を含む1次元テンソルを生成することができる。それぞれの埋め込まれたパッチ152が次元 $L_1 \times W_1 \times C$ を有する場合、ネットワークブロック150は、次元 $1 \times (L \cdot W \cdot C)$ を有する1次元テンソルを生成することができる。それから、ネットワークブロック150は、1次元テンソルから入力シーケンスを生成することができ、入力シーケンスの入力位置の少なくとも一部は、1次元テンソルであり、ネットワークブロック150は、自己アテンションニューラルネットワーク層160を使用して入力シーケンスを処理することができる。

【0100】

特に、それぞれの特定の入力位置に対応するそれぞれの埋め込まれたパッチ152(またはその処理されたバージョン)に関して、自己アテンションニューラルネットワーク層160は、特定の位置のそれぞれの出力を生成するために、特定の入力位置の埋め込まれたパッチ152から導出された1つまたは複数のクエリ(query)を使用して、入力位置の埋め込まれたパッチ152(すなわち、その他の埋め込まれたパッチ152、および一部の实装形態において、それ自体)にアテンションメカニズムを適用するように構成される。自己アテンションニューラルネットワーク層160の出力は、各入力位置に対応するそれぞれの出力要素を含む出力シーケンスである。

【0101】

10

20

30

40

50

一部の実装形態において、ネットワークブロック150の自己アテンションニューラルネットワーク層(たとえば、図1Bに示された自己アテンションニューラルネットワーク層160)の一部またはすべては、マルチヘッド自己アテンションニューラルネットワーク層である。マルチヘッド自己アテンションニューラルネットワーク層は、 h 個の異なるアテンションメカニズムを並列に適用して出力要素のそれぞれのシーケンスを生成し、次に、出力要素の複数のシーケンスを組み合わせて出力要素の最終的なシーケンスを生成する。

【0102】

一部の実装形態において、ネットワークブロック150の自己アテンションニューラルネットワーク層(たとえば、図1Bに示された自己アテンションニューラルネットワーク層160)の一部またはすべては、入力シーケンスのそれぞれの埋め込まれたパッチ152の位置情報をアテンションメカニズムに組み込む。たとえば、特定の埋め込まれたパッチ152に関してアテンションを適用するとき(すなわち、特定の埋め込まれたパッチ152に対応するそれぞれの出力要素を生成するとき)、自己アテンションニューラルネットワーク層160は、画像に対応する(またはブロック入力内の)埋め込まれたパッチ152の位置を表すアテンション位置埋め込み(attention positional embedding)を特定することができる。

【0103】

本明細書において、埋め込みは、特定の埋め込み空間において入力を表す数値の順序付けられた集まりである。たとえば、埋め込みは、決まった次元を有する、浮動小数点またはその他の数値のベクトルであることが可能である。

【0104】

一部の実装形態において、ネットワークブロック150は、自己アテンションニューラルネットワーク層160の出力を自己アテンションニューラルネットワーク層160への入力と組み合わせる残差接続層162を含む。代わりにまたは加えて、ネットワークブロック150は、自己アテンションニューラルネットワーク層160(または残差接続層162)の出力に層正規化を適用する層正規化層164を含み得る。

【0105】

自己アテンションニューラルネットワーク層160(または残差接続層162または層正規化層164)の出力は、埋め込まれたパッチ152の更新された表現を含む。埋め込まれたパッチ152の更新された表現は、それぞれの埋め込まれたパッチ152のそれぞれの更新された表現に関して、ネットワークブロック150の複数のエキスパートモジュールの集合の1つまたは複数のエキスパートモジュール180に更新された表現を割り振るように構成されるルーティングモジュール170に提供され得る。ルーティングモジュール170は、図1Aを参照して上で説明されたルーティングモジュール120と同様に構成され得る。

【0106】

ネットワークブロックは、各エキスパートモジュール180に関して、それぞれの割り振られた更新された表現に関するそれぞれのエキスパートモジュール出力を生成するプロセスのために、それぞれの埋め込まれたパッチ152の割り振られた更新された表現を提供することができる。エキスパートモジュール180は、図1Aを参照して上で説明されたエキスパートモジュール130a~eと同様に構成され得る。

【0107】

図1Bに示されたように、各エキスパートモジュール180は、1つまたは複数の順伝播型ニューラルネットワーク層を含む多層パーセプトロン(MLP)である。たとえば、エキスパートモジュール180は、活性化関数、たとえば、非線形の要素ごとの活性化関数、たとえば、ReLU活性化関数によってそれぞれ分けられる2つ以上の学習された線形変換を含み得る。

【0108】

そして、エキスパートモジュール出力は、ネットワークブロックのブロック出力156を生成するために、ネットワークブロック150の組合せモジュール190によって組み合わせられ得る。組合せモジュール190は、図1Aを参照して上で説明された組合せモジュール140と同様に構成され得る。

10

20

30

40

50

【0109】

一部の実装形態において、ネットワークブロック150は、組合せモジュール190の出力をルーティングモジュール170への入力と組み合わせる残差接続層192を含む。つまり、ブロック出力156は、組合せモジュール190ではなく残差接続層192の出力であることが可能である。

【0110】

ブロック出力156は、ニューラルネットワークのネットワーク出力を生成するためのさらなる処理のために、ネットワークブロックのシーケンスの1つまたは複数の後続のネットワークブロックに提供され得る。

【0111】

ネットワーク出力を生成した後、ニューラルネットワークは、ネットワーク出力108を1つまたは複数の下流のシステムに提供することができる。たとえば、ニューラルネットワークは、図1Aを参照して上で説明されたように、それぞれの機械学習タスクのための予測を生成するために、ネットワーク出力を1つまたは複数のヘッドニューラルネットワークに提供することができる。別の例として、ニューラルネットワークは、入力画像の埋め込みを表し得るネットワーク出力を、さらなる処理のために、データベースまたは、1つもしくは複数の下流の機械学習モデルに提供することができる。

【0112】

図2は、画像パッチにセグメンテーションされた例示的な画像210、220、230、240、および250を示す。

【0113】

画像210~250は、画像210~250についての予測を生成するために画像210~250を処理するように構成されるニューラルネットワークシステム、たとえば、図1Aを参照して上で説明されたニューラルネットワーク100への入力として提供され得る。ニューラルネットワークシステムは、画像210~250を複数の画像パッチにセグメンテーションし、1つまたは複数のエキスパートネットワークブロックを含むネットワークブロックのシーケンスを使用して画像パッチを処理するように構成され得る。

【0114】

以下の説明は画像のパッチを決定することに言及するが、同じ技術が、画像の中間表現のパッチ、たとえば、図1Aを参照して上で説明されたブロック入力112などのネットワークブロックへのブロック入力のパッチを決定するために使用され得ることを理解されたい。

【0115】

画像210、220、230、240、および250は、画像を画像パッチにセグメンテーションする異なる可能性を示す。特に、図2において、各画像210~250は、それぞれ視覚的に異なる、すなわち、異なる濃淡またはハッチングを使用する複数の画像パッチの集合にセグメンテーションされたものとして描写される。概して、ニューラルネットワークシステムは、すべての受け取られた画像を同じスキーマ(schema)に従ってセグメンテーションするように構成される。つまり、異なるスキーマに従って画像がセグメンテーションされているため、同じニューラルネットワークシステムが、画像210、220、230、240、および250の各々を必ずしも描かれたようにセグメンテーションするとは限らない。

【0116】

第1の画像210に示されるように、一部の实装形態において、ニューラルネットワークシステムは、それぞれが同じサイズおよび形状を有する画像パッチを生成することができ、たとえば、各画像パッチは、長方形であることが可能である。さらに、一部の实装形態において、ニューラルネットワークシステムは、あらゆる画素がちょうど1つの画像パッチの構成要素であるように、第1の画像210をセグメンテーションすることができる。特定の例として、図2に示されたように、画像パッチは、同じサイズの長方形のグリッドを表し得る。別の特定の例として、画像パッチは、同じサイズの六角形のグリッドを表し得る。

【0117】

第2の画像220に示されるように、一部の実装形態において、ニューラルネットワークシステムは、異なるサイズを有する画像パッチを生成することができる。

【0118】

第3の画像230に示されるように、一部の実装形態において、ニューラルネットワークシステムは、一部の画素が複数の異なる画像パッチの構成要素であるように第3の画像230をセグメンテーションすることができる。たとえば、第1の画像パッチは、第2の画像パッチの画素の真上位集合を含むことが可能であり、つまり、第2の画像パッチは、第1の画像パッチ内に完全に包含され得る。特定の例として、画像から決定された複数の第1の画像パッチの各々に関して(たとえば、第1の画像パッチのグリッド内のそれぞれの第1の画像パッチに関して)、ニューラルネットワークシステムは、同じ中心画素を有するがサイズおよび/またはアスペクト比が異なるそれぞれの第2の画像パッチを生成することができる。

10

【0119】

第4の画像240に示されるように、一部の実装形態において、ニューラルネットワークシステムは、一部の画素がいずれの画像パッチの構成要素でもないように第4の画像240をセグメンテーションすることができる。たとえば、ニューラルネットワークシステムは、機械学習モデルを使用して第4の画像240を処理して、1つまたは複数の関心領域を特定することができ、ニューラルネットワークシステムは、それぞれの特定された関心領域のためにそれぞれのパッチを生成することができる。たとえば、機械学習モデルは、1つまたは複数の画素を特定するように構成され得、ニューラルネットワークシステムは、それぞれの特定された画素を中心とするそれぞれのパッチを生成することが可能である。

20

【0120】

第5の画像250に示されるように、一部の実装形態において、ニューラルネットワークシステムは、任意の形状の画像パッチを生成することができる。つまり、画像パッチは、長方形である必要はない。たとえば、ニューラルネットワークシステムは、たとえば、第5の画像250の各画素にそれぞれのクラスを割り振ることによって第5の画像250をセグメンテーションするように構成される機械学習モデルを使用して、第5の画像250を処理することができる。それから、ニューラルネットワークシステムは、機械学習モデルによって同じクラスを割り振られた画素のそれぞれの連続的な集合のためにそれぞれのパッチを生成することができる。

30

【0121】

画像210~250は図2において2次元画像(または複数のチャンネルを持つ2次元の画像、たとえば、RGB画像)として示されているが、概して、ニューラルネットワークシステムは、図1Aを参照して上で説明されたように、任意の種類の画像に関する予測を生成するように構成され得る。

【0122】

図3は、複数のエキスパートモジュールを含むネットワークブロックを実行するための例示的なプロセスの流れ図である。便宜上、プロセス300は、1つまたは複数の場所に置かれた1つまたは複数のコンピュータのシステムによって実行されるものとして説明される。たとえば、本明細書に従って適切にプログラミングされたニューラルネットワークシステム、たとえば、図1Aを参照して上で説明されたニューラルネットワーク100が、プロセス300を実行し得る。

40

【0123】

ネットワークブロックは、入力画像を処理し、入力画像に関するネットワーク出力を生成するように構成されるニューラルネットワークのネットワークブロックのシーケンスの1つのネットワークブロックであることが可能である。

【0124】

システムは、入力画像の中間表現を表すブロック入力を取得する(ステップ302)。

【0125】

システムは、ブロック入力(またはブロック入力の更新された表現、たとえば、図1Bを

50

参照して上で説明された自己アテンションメカニズムを適用することによって生成された更新された表現)の複数のパッチの集合を決定し、各パッチは、ブロック入力(またはブロック入力の更新された表現)の要素の異なる部分集合を含む(ステップ304)。

【0126】

システムは、ネットワークブロックの複数のエキスパートモジュールのうちの1つまたは複数のそれぞれのエキスパートモジュールに各パッチを割り振る(ステップ306)。

【0127】

システムは、複数のパッチの各パッチに関して、1つまたは複数のそれぞれのモジュール出力を生成するために1つまたは複数の対応するエキスパートモジュールを使用してパッチを処理する(ステップ308)。

【0128】

システムは、それぞれのエキスパートモジュールによって生成されたモジュール出力を組み合わせることによって、ネットワークブロックのブロック出力を生成する(ステップ310)。

【0129】

本明細書は、用語「構成される」をシステムおよびコンピュータプログラムのコンポーネントに関連して使用する。1つまたは複数のコンピュータのシステムが特定の動作またはアクションを実行するように構成されることは、システムが、動作中に、システムに動作またはアクションを実行させるソフトウェア、ファームウェア、ハードウェア、またはそれらの組合せをそのシステム上にインストール済みであることを意味する。1つまたは複数のコンピュータプログラムが特定の動作またはアクションを実行するように構成されることは、1つまたは複数のプログラムが、データ処理装置によって実行されるときに、装置に動作またはアクションを実行させる命令を含むことを意味する。

【0130】

本明細書に記載の主題の実施形態および機能的動作は、本明細書において開示された構造およびそれらの構造的均等物を含むデジタル電子回路、有形で具現化されたコンピュータソフトウェアもしくはファームウェア、コンピュータハードウェア、またはそれらのうちの1つもしくは複数の組合せで実装され得る。本明細書に記載の主題の実施形態は、1つまたは複数のコンピュータプログラム、すなわち、データ処理装置による実行のために、またはデータ処理装置の動作を制御するために有形の非一時的記憶媒体上に符号化されたコンピュータプログラム命令の1つまたは複数のモジュールとして実装され得る。コンピュータ記憶媒体は、機械可読ストレージデバイス、機械可読ストレージ基板、ランダムもしくはシリアルアクセスメモリデバイス、またはそれらのうちの1つもしくは複数の組合せであることが可能である。代替的にまたは追加的に、プログラム命令は、データ処理装置による実行のために好適なレシーバ装置に送信するために情報を符号化するように生成される人為的に生成された伝播信号、たとえば、機械によって生成された電氣的信号、光学的信号、または電磁的信号上に符号化され得る。

【0131】

用語「データ処理装置」は、データ処理ハードウェアを指し、例として、1つのプログラミング可能なプロセッサ、1台のコンピュータ、または複数のプロセッサもしくはコンピュータを含む、データを処理するためのすべての種類の装置、デバイス、およびマシンを包含する。装置は、専用論理回路、たとえば、フィールドプログラマブルゲートアレイ(FPGA)または特定用途向け集積回路(ASIC)であることも可能であり、またはそのような専用論理回路をさらに含むことも可能である。任意で、装置は、ハードウェアに加えて、コンピュータプログラムのための実行環境を作成するコード、たとえば、プロセッサのファームウェア、プロトコルスタック、データベース管理システム、オペレーティングシステム、またはそれらのうちの1つもしくは複数の組合せを構成するコードを含み得る。

【0132】

プログラム、ソフトウェア、ソフトウェアアプリケーション、アプリ、モジュール、ソフトウェアモジュール、スクリプト、またはコードと呼ばれる、または記載される場合も

10

20

30

40

50

あるコンピュータプログラムは、コンパイラ型言語もしくはインタープリタ型言語、または宣言型言語もしくは手続き型言語を含む任意の形態のプログラミング言語で記述され得、独立型プログラムとしての形態、またはモジュール、コンポーネント、サブルーチン、もしくはコンピューティング環境内での使用に好適なその他のユニットとしての形態を含む任意の形態でデプロイされ得る。プログラムは、ファイルシステム内のファイルに対応する場合があるが、必ずそうであるとは限らない。プログラムは、その他のプログラムもしくはデータを保持するファイルの一部、たとえば、マークアップ言語のドキュメントに記憶された1つもしくは複数のスクリプト、問題にしているプログラムに専用の単一のファイル、または複数の組織されたファイル、たとえば、1つもしくは複数のモジュール、サブプログラム、もしくはコードの一部を記憶するファイルに記憶され得る。コンピュータプログラムは、1台のコンピュータ上で、または1つの場所に置かれるか、もしくは複数の場所にわたって分散され、データ通信ネットワークによって相互に接続される複数のコンピュータ上で実行されるようにデプロイされ得る。

10

【0133】

本明細書において、用語「データベース」は、データの任意の集まりを指すために広く使用され、データは、いずれかの特定の 방법으로構造化される必要はなく、またはまったく構造化される必要がなく、1つまたは複数の場所のストレージデバイスに記憶され得る。したがって、たとえば、インデックスデータベースが、データの複数の集まりを含むことができ、データの集まりの各々は、異なる方法で編成され、アクセスされる場合がある。

【0134】

20

同様に、本明細書において、用語「エンジン」は、1つまたは複数の特定の機能を実行するようにプログラミングされるソフトウェアベースのシステム、サブシステム、またはプロセスを指すために広く使用される。概して、エンジンは、1つまたは複数の場所の1つまたは複数のコンピュータにインストールされた1つまたは複数のソフトウェアモジュールまたはコンポーネントとして実装される。場合によっては、1つまたは複数のコンピュータが、特定のエンジンに専用であり、その他の場合、複数のエンジンが、同じ1台のコンピュータまたは複数のコンピュータにインストールされ、実行され得る。

【0135】

本明細書に記載のプロセスおよび論理フローは、入力データに対して演算を行い、出力を生成することによって機能を実行するために1つまたは複数のコンピュータプログラムを1つまたは複数のプログラミング可能なコンピュータが実行することによって実行され得る。また、プロセスおよび論理フローは、専用論理回路、たとえば、FPGAもしくはASICによって、または専用論理回路と1つもしくは複数のプログラミングされたコンピュータとの組合せによって実行され得る。

30

【0136】

コンピュータプログラムの実行に好適なコンピュータは、汎用マイクロプロセッサもしくは専用マイクロプロセッサもしくはそれら両方、または任意のその他の種類の中央演算処理装置に基づくことが可能である。概して、中央演算処理装置は、読み出し専用メモリ、またはランダムアクセスメモリ、またはそれら両方から命令およびデータを受け取る。コンピュータの必須の要素は、命令を遂行または実行するための中央演算処理装置、ならびに命令およびデータを記憶するための1つまたは複数のメモリデバイスである。中央演算処理装置およびメモリは、専用論理回路によって補完されるか、または専用論理回路に組み込まれ得る。また、概して、コンピュータは、データを記憶するための1つもしくは複数の大容量ストレージデバイス、たとえば、磁気ディスク、光磁気ディスク、もしくは光ディスクを含むか、またはそれらの大容量ストレージデバイスからデータを受信するか、もしくはそれらの大容量ストレージデバイスにデータを転送するか、もしくはその両方を行うために動作可能なように結合される。しかし、コンピュータは、そのようなデバイスを有していなくてもよい。さらに、コンピュータは、別のデバイス、たとえば、ほんのいくつか例を挙げるとすれば、モバイル電話、携帯情報端末(PDA)、モバイルオーディオもしくはビデオプレーヤー、ゲームコンソール、全地球測位システム(GPS)レシーバ、ま

40

50

たはポータブルストレージデバイス、たとえば、ユニバーサルシリアルバス(USB)フラッシュドライブに組み込まれ得る。

【0137】

コンピュータプログラム命令およびデータを記憶するのに好適なコンピュータ可読媒体は、例として、半導体メモリデバイス、たとえば、EPROM、EEPROM、およびフラッシュメモリデバイス、磁気ディスク、たとえば、内蔵ハードディスクまたはリムーバブルディスク、光磁気ディスク、ならびにCD-ROMディスクおよびDVD-ROMディスクを含むすべての形態の不揮発性メモリ、媒体、およびメモリデバイスを含む。

【0138】

ユーザとのインタラクションを提供するために、本明細書に記載の主題の実施形態は、ユーザに対して情報を表示するためのディスプレイデバイス、たとえば、ブラウン管(CRT)または液晶ディスプレイ(LCD)モニタ、ならびにユーザがコンピュータに入力を与えることができるキーボードおよびポインティングデバイス、たとえば、マウスまたはトラックボールを有するコンピュータ上に実装され得る。その他の種類のデバイスが、ユーザとのインタラクションを提供するためにやはり使用され得、たとえば、ユーザに提供されるフィードバックは、任意の形態の感覚フィードバック、たとえば、視覚フィードバック、聴覚フィードバック、または触覚フィードバックであることが可能であり、ユーザからの入力、音響、スピーチ、または触覚による入力を含む任意の形態で受け取られ得る。加えて、コンピュータは、ユーザによって使用されるデバイスにドキュメントを送信し、そのデバイスからドキュメントを受信することによって、たとえば、ウェブブラウザから受信された要求に回答してユーザのデバイスのウェブブラウザにウェブページを送信することによってユーザとインタラクションすることができる。また、コンピュータは、メッセージングアプリケーションを実行しているパーソナルデバイス、たとえば、スマートフォンにテキストメッセージまたはその他の形態のメッセージを送信し、返報としてユーザから応答メッセージを受信することによってユーザとインタラクションすることができる。

【0139】

機械学習モデルを実装するためのデータ処理装置は、たとえば、機械学習の訓練または生成、すなわち、推論の作業負荷のよくある計算量の多い部分を処理するための専用ハードウェアアクセラレータユニットも含むことが可能である。

【0140】

機械学習モデルは、機械学習フレームワーク、たとえば、TensorFlowフレームワーク、Microsoft Cognitive Toolkitフレームワーク、Apache Singaフレームワーク、またはApache MXNetフレームワークを使用して実装され、デプロイされ得る。

【0141】

本明細書に記載の主題の実施形態は、バックエンドコンポーネントを、たとえば、データサーバとして含むか、またはミドルウェアコンポーネント、たとえば、アプリケーションサーバを含むか、またはフロントエンドコンポーネント、たとえば、ユーザが本明細書に記載の主題の実装形態とインタラクションすることができるグラフィカルユーザインターフェース、ウェブブラウザ、もしくはアプリを有するクライアントコンピュータを含むか、または1つもしくは複数のそのようなバックエンドコンポーネント、ミドルウェアコンポーネント、もしくはフロントエンドコンポーネントの任意の組合せを含むコンピューティングシステムに実装され得る。システムのコンポーネントは、任意の形態または媒体のデジタルデータ通信、たとえば、通信ネットワークによって相互に接続され得る。通信ネットワークの例は、ローカルエリアネットワーク(LAN)および広域ネットワーク(WAN)、たとえば、インターネットを含む。

【0142】

コンピューティングシステムは、クライアントおよびサーバを含み得る。クライアントおよびサーバは、概して互いに離れており、通常は通信ネットワークを通じてインタラクションする。クライアントとサーバとの関係は、それぞれのコンピュータ上で実行されており、互いにクライアント-サーバの関係にあるコンピュータプログラムによって生じる。

一部の実施形態において、サーバは、たとえば、クライアントとして働くデバイスとインタラクションするユーザに対してデータを表示し、そのようなユーザからユーザ入力を受け取る目的でユーザデバイスにデータ、たとえば、HTMLページを送信する。ユーザデバイスにおいて生成されたデータ、たとえば、ユーザインタラクションの結果が、サーバにおいてデバイスから受信され得る。

【0143】

上述の実施形態に加えて、以下の実施形態も、革新的である。

【0144】

実施形態1は、入力画像を処理し、入力画像に関するネットワーク出力を生成するように構成されるニューラルネットワークを含むシステムであって、ニューラルネットワークが、1つまたは複数のネットワークブロックのシーケンスを含み、1つまたは複数のネットワークブロックが、

10

入力画像の中間表現を表すブロック入力を取得することと、

ブロック入力の、またはブロック入力の更新された表現の、複数のパッチを決定することであって、各パッチが、ブロック入力の、またはブロック入力の更新された表現の要素の異なる部分集合を含む、ことと、

ネットワークブロックの複数のエキスパートモジュールのうちの1つまたは複数のそれぞれのエキスパートモジュールに各パッチを割り振ることと、

複数のパッチの各パッチに関して、1つまたは複数のそれぞれのモジュール出力を生成するために1つまたは複数の対応するエキスパートモジュールを使用してパッチを処理することと、

20

それぞれのエキスパートモジュールによって生成されたモジュール出力を組み合わせることによってブロック出力を生成することを含む動作を実行するようにそれぞれ構成される、システムである。

【0145】

実施形態2は、ネットワークブロックの1つまたは複数のそれぞれのエキスパートモジュールに各パッチを割り振ることが、

複数のエキスパートモジュールの各エキスパートモジュールに関して、それぞれのルーティングスコアを生成するためにルーティングサブネットワークを使用してパッチを処理することと、

30

複数のルーティングスコアのうちの1つまたは複数の最も高いルーティングスコアを決定することと、

1つまたは複数の最も高いルーティングスコアの各々に対応するエキスパートモジュールにパッチを割り振ることを含む、実施形態1のシステムである。

【0146】

実施形態3は、複数のエキスパートモジュールの各エキスパートモジュールに関して、それぞれのルーティングスコアを生成するためにルーティングサブネットワークを使用してパッチを処理することが、

各エキスパートモジュールのそれぞれの第1の中間ルーティングスコアを生成するために、1つまたは複数の順伝播型ニューラルネットワーク層を使用してパッチを処理することと、

40

各エキスパートモジュールに関して、ノイズ値をサンプリングし、エキスパートモジュールの第2の中間ルーティングスコアを生成するためにエキスパートモジュールの第1の中間ルーティングスコアにノイズ値を加えることと、

各エキスパートモジュールに関して、エキスパートモジュールのルーティングスコアを生成するために、エキスパートモジュールの第2の中間ルーティングスコアに非線形活性化関数を適用することを含む、実施形態2のシステムである。

【0147】

実施形態4は、複数のパッチの各パッチに関して、1つまたは複数のそれぞれのモジュール出力を生成するために1つまたは複数の対応するエキスパートモジュールを使用してパ

50

ッチを処理することが、

パッチに対応する1つまたは複数のエキスパートモジュールの各々に関して、それぞれの初期モジュール出力を生成するためにエキスパートモジュールを使用してパッチを処理することと、

パッチに対応する1つまたは複数のエキスパートモジュールの各々に関して、エキスパートモジュールおよびパッチに対応するルーティングスコアに従って、対応する初期モジュール出力をスケーリングすることを含む、実施形態2または3のいずれか1つのシステムである。

【0148】

実施形態5は、ニューラルネットワークが、

各エキスパートモジュールに対応するルーティングスコアが中心傾向(central tendency)の同じ測定値(measure)を有するように促す重要度損失、または

各エキスパートモジュールがパッチを割り振られるそれぞれの見込みが中心傾向の同じ測定値を有するように促す負荷損失のうちの1つまたは複数を使用して訓練される、実施形態2から4のいずれか1つのシステムである。

【0149】

実施形態6は、一団の訓練画像Xに関する重要度損失 $L_{imp}(X)$ を決定することが、

各エキスパートモジュールiに関して、

【0150】

【数8】

$$\text{imp}_i(X) = \sum_{x \in X} f_i(x)$$

【0151】

を計算することであって、式中、 f_i が、エキスパートモジュールiのルーティングスコアまたはエキスパートモジュールiの中間ルーティングスコアを計算するための、ルーティングサブネットワークによって定義される関数である、計算することと、

各エキスパートモジュールiに関して、 $\text{imp}_i(X)$ が中心傾向の同じ測定値を有するように促すために重要度損失 $L_{imp}(X)$ を計算することを含む、実施形態5のシステムである。

【0152】

実施形態7は、重要度損失 $L_{imp}(X)$ が、

【0153】

【数9】

$$\mathcal{L}_{imp}(X) = \left(\frac{\text{std}(\text{imp}_{1:E}(X))}{\text{mean}(\text{imp}_{1:E}(X))} \right)^2$$

【0154】

であり、式中、Eが、複数のエキスパートモジュールの数である、実施形態6のシステムである。

【0155】

実施形態8は、

各パッチが、k個の異なるエキスパートモジュールに割り振られ、 $k \geq 1$ であり、

一団の訓練画像Xに関する負荷損失 $L_{load}(X)$ を決定することが、

一団の訓練画像Xの中の各訓練画像xに関して、

【0156】

【数10】

$$\text{threshold}_k(x) = \max_{(k)} g_i(x)$$

10

20

30

40

50

【0157】

を計算することであって、式中、 $g_i(x)$ が、エキスパートモジュール*i*のルーティングスコアまたはエキスパートモジュール*i*の中間ルーティングスコアを計算するための、ルーティングサブネットワークによって定義される関数である、計算することと、

各訓練画像*x*に関しておよび各エキスパートモジュール*i*に関して、

$$p_i(x) = P(g_i(x) \geq \text{threshold}_k(x))$$

を計算することであって、式中、 $P(e)$ が、事象*e*が発生する見込みである、計算することと、

各エキスパートモジュール*i*に関して、

【0158】

【数11】

$$\text{load}_i(X) = \sum_{x \in X} p_i(x)$$

【0159】

を計算し、各エキスパートモジュール*i*に関して、 $\text{load}_i(x)$ が中心傾向の同じ測定値を有するように促すために負荷損失 $L_{\text{load}}(X)$ を計算することを含む、実施形態5から7のいずれか1つのシステムである。

【0160】

実施形態9は、負荷度損失 $L_{\text{load}}(X)$ が、

【0161】

【数12】

$$\mathcal{L}_{\text{load}}(X) = \left(\frac{\text{std}(\text{load}_{1:E}(X))}{\text{mean}(\text{load}_{1:E}(X))} \right)^2$$

【0162】

であり、式中、*E*が、複数のエキスパートモジュールの数である、実施形態8のシステムである。

【0163】

実施形態10は、複数のパッチの各パッチに関して、対応するエキスパートモジュールを使用してパッチを処理することが、エキスパートモジュールにわたって並列にパッチを処理することを含む、実施形態1から9のいずれか1つのシステムである。

【0164】

実施形態11は、複数のエキスパートモジュールの数が、複数のパッチの数よりも多い、実施形態1から10のいずれか1つのシステムである。

【0165】

実施形態12は、実施形態1から11のいずれか1つの動作を含む方法である。

【0166】

実施形態13は、1つまたは複数のコンピュータによって実行されるときに1つまたは複数のコンピュータに実施形態1から11のいずれか1つの動作を実行させる命令を記憶する1つまたは複数のコンピュータ記憶媒体である。

【0167】

本明細書は多くの特定の装形態の詳細を含むが、これらは、いかなる発明の範囲または特許請求される可能性があるものの範囲に対する限定ともみなされるべきでなく、むしろ、特定の発明の特定の装形態に特有である可能性がある特徴の説明ともみなされるべきである。別々の装形態の文脈で本明細書において説明されている特定の装形態が、単一の装形態に組み合わせて装されることも可能である。反対に、単一の装形態の文脈で

10

20

30

40

50

説明されている様々な特徴が、複数の実施形態において別々にまたは任意の好適な部分的組合せで実装されることも可能である。さらに、特徴は、特定の組合せで働くものとして上で説明されている場合があり、最初にそのように主張されてさえいる場合があるが、主張された組合せの1つまたは複数の特徴は、場合によっては組合せから削除され得、主張された組合せは、部分的組合せ、または部分的組合せの変形を対象とする場合がある。

【0168】

同様に、動作が特定の順序で図面に示され、特許請求の範囲に記載されているが、これは、そのような動作が、示された特定の順序で、もしくは逐次的順序で実行されること、または所望の結果を達成するために、示されたすべての動作が実行されることを必要とするものと理解されるべきでない。特定の状況においては、マルチタスクおよび並列処理が有利である場合がある。さらに、上述の実施形態における様々なシステムモジュールおよびコンポーネントの分割は、すべての実施形態においてそのような分割を必要とするものと理解されるべきでなく、説明されたプログラムコンポーネントおよびシステムは、概して、単一のソフトウェア製品と一緒に統合されるか、または複数のソフトウェア製品にパッケージングされ得ることが理解されるべきである。

【0169】

主題の特定の実施形態が説明された。その他の実施形態は、添付の特許請求の範囲の範囲内にある。たとえば、特許請求の範囲に記載のアクションは、異なる順序で実行され得、それでも所望の結果を達成することができる。一例として、添付の図に示されたプロセスは、所望の結果を達成するために、必ずしも、示された特定の順序または逐次的順序を必要としない。場合によっては、マルチタスクおよび並列処理が有利である可能性がある。

【符号の説明】

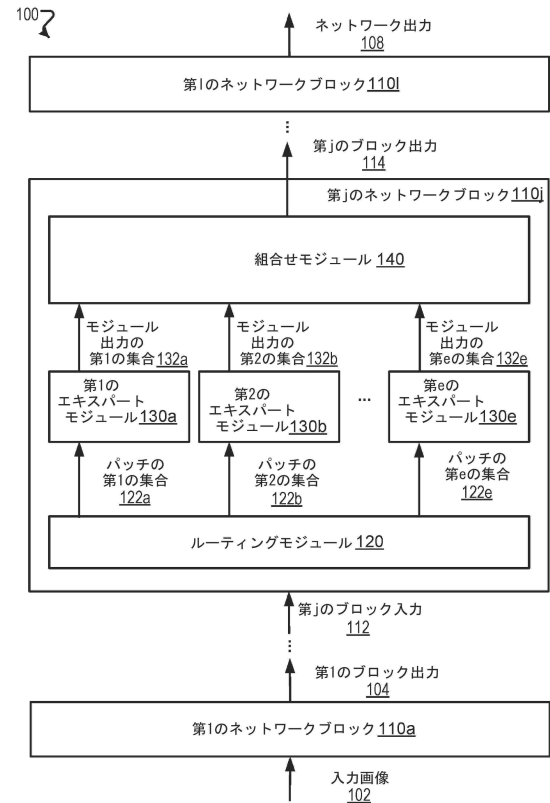
【0170】

- 100 ニューラルネットワーク、ニューラルネットワークシステム
- 102 入力画像
- 104 第1のブロック出力
- 108 ネットワーク出力
- 110a~l ネットワークブロック
- 112 ブロック入力
- 114 ブロック出力
- 120 ルーティングモジュール
- 122a~e パッチの集合
- 130a~e エキスパートモジュール
- 132a~e エキスパートモジュール出力の対応する集合
- 140 組合せモジュール
- 150 ニューラルネットワークブロック、ネットワークブロック
- 152 埋め込まれたパッチ
- 154 層正規化層
- 156 ブロック出力
- 160 自己アテンションニューラルネットワーク層
- 162 残差接続層
- 164 層正規化層
- 170 ルーティングモジュール
- 180 エキスパートモジュール
- 190 組合せモジュール
- 192 残差接続層
- 210 第1の画像
- 220 第2の画像
- 230 第3の画像
- 240 第4の画像

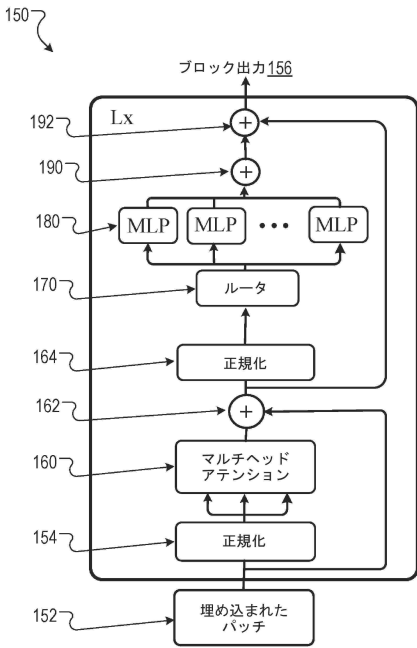
250 第5の画像
300 プロセス

【図面】

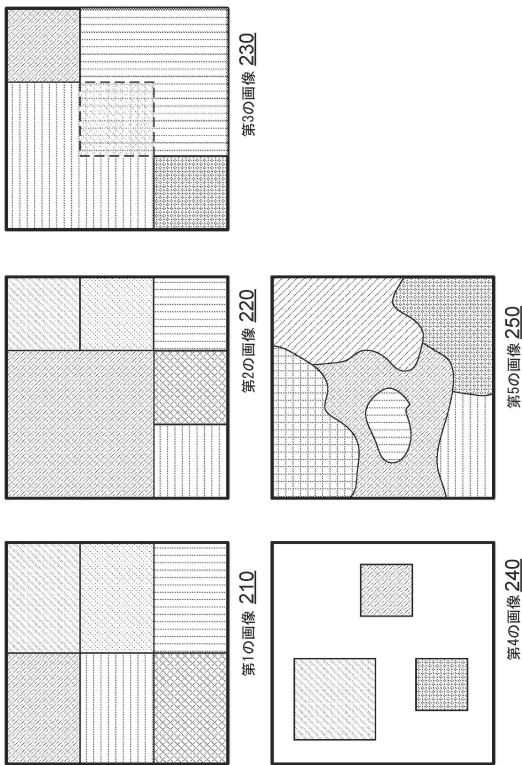
【図1A】



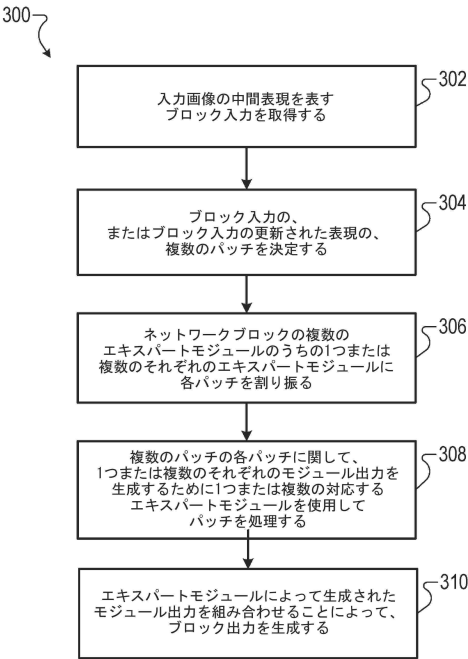
【図1B】



【図2】



【図3】



フロントページの続き

- (72)発明者 カルロス・リケルメ・ルイス
スイス・８００２・チューリッヒ・ブラントシェンケシュトラーセ・１１０
- (72)発明者 アンドレ・スサノ・ピント
スイス・８００２・チューリッヒ・ブラントシェンケシュトラーセ・１１０
- (72)発明者 バジル・ムスタファ
スイス・８００２・チューリッヒ・ブラントシェンケシュトラーセ・１１０
- (72)発明者 ダニエル・エム・カイザース
スイス・８００２・チューリッヒ・ブラントシェンケシュトラーセ・１１０
- (72)発明者 ジョアン・ブイグセルバー・イ・ペレス
スイス・８００２・チューリッヒ・ブラントシェンケシュトラーセ・１１０
- (72)発明者 マキシム・ノイマン
スイス・８００２・チューリッヒ・ブラントシェンケシュトラーセ・１１０
- (72)発明者 ニール・マシュー・ティンマス・ホールズビー
スイス・８００２・チューリッヒ・ブラントシェンケシュトラーセ・１１０
- (72)発明者 ロドルフ・ジェナートン
ドイツ・１０１１７・ベルリン・トゥホルスキーシュトラーセ・２
- 審査官 多賀 実
- (56)参考文献 Hyungtae Lee et al., "ME R-CNN: Multi-Expert R-CNN for Object Detection", IEEE Transactions on Image Processing [online], 米国, IEEE, 2019年09月09日, vol. 29, pp. 1030-1044, [検索日 2024.10.21], インターネット: 〈URL: <https://ieeexplore.ieee.org/abstract/document/8827587>〉, 〈DOI: 10.1109/TIP.2019.2938879〉
- (58)調査した分野 (Int.Cl., DB名)
G 0 6 N 3 / 0 0 - 9 9 / 0 0
G 0 6 T 7 / 0 0 - 7 / 9 0
G 0 6 V 1 0 / 0 0 - 2 0 / 9 0