

(19)日本国特許庁(JP)

(12)公表特許公報(A)

(11)公表番号

特表2024-536035

(P2024-536035A)

(43)公表日 令和6年10月4日(2024.10.4)

(51)国際特許分類

H 0 4 N 19/90 (2014.01)

H 0 4 N 19/70 (2014.01)

F I

H 0 4 N 19/90

H 0 4 N 19/70

テーマコード (参考)

5 C 1 5 9

審査請求 未請求 予備審査請求 未請求 (全57頁)

(21)出願番号 特願2024-517458(P2024-517458)
(86)(22)出願日 令和4年9月15日(2022.9.15)
(85)翻訳文提出日 令和6年3月19日(2024.3.19)
(86)国際出願番号 PCT/US2022/076496
(87)国際公開番号 WO2023/049655
(87)国際公開日 令和5年3月30日(2023.3.30)
(31)優先権主張番号 17/486,732
(32)優先日 令和3年9月27日(2021.9.27)
(33)優先権主張国・地域又は機関 米国(US)
(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 595020643
クualコム・インコーポレイテッド
QUALCOMM INCORPORATED
アメリカ合衆国、カリフォルニア州 9
2121-1714、サン・ディエゴ、
モアハウス・ドライブ 5775
(74)代理人 110003708
弁理士法人鈴榮特許総合事務所
(72)発明者 ジュ、インハオ
アメリカ合衆国、カリフォルニア州 9
2121、サン・ディエゴ、モアハウス
・ドライブ 5775
(72)発明者 ヤン、ヤン
アメリカ合衆国、カリフォルニア州 9
最終頁に続く

(54)【発明の名称】 メディアの変換コーディングのためのトランスフォーマベースのアーキテクチャ

(57)【要約】

ニューラルネットワークシステムを使用してメディアデータを処理するためのシステム及び技法が本明細書において説明される。例えば、プロセスは、符号化画像データのフレームの潜在表現を取得することと、符号化画像データのフレームの潜在表現を入力として使用して、デコーダサブネットワークの複数のデコーダトランスフォーマ層によって、復号画像データのフレームを生成することと、を含むことができる。複数のデコーダトランスフォーマ層のうちの少なくとも1つのデコーダトランスフォーマ層は、特徴の1つ又は複数のパッチを生成し、1つ又は複数のパッチにわたって適用された1つ又は複数のウィンドウ区分及びシフトウィンドウ区分内で局所的に自己アテンションを決定するための1つ又は複数のトランスフォーマブロックと、1つ又は複数のパッチの各パッチのそれぞれのサイズを減少させるためのパッチ分離エンジンとを含む。

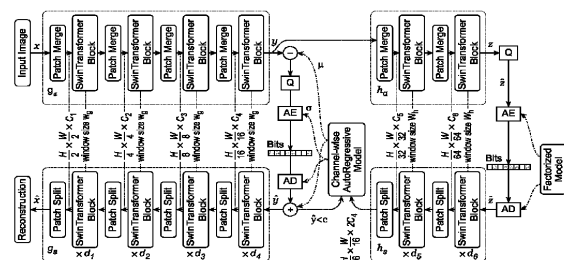


FIG. 5C

【特許請求の範囲】

【請求項 1】

メディアデータを処理するための装置であって、
少なくとも 1 つのメモリと、
前記少なくとも 1 つのメモリに結合された少なくとも 1 つのプロセッサと、を備え、前記少なくとも 1 つのプロセッサが、
符号化画像データのフレームの潜在表現を取得し、
前記符号化画像データのフレームの前記潜在表現を入力として使用して、デコーダサブネットワークの複数のデコーダトランスフォーマ層に基づいて、復号画像データのフレームを生成するように構成されており、前記複数のデコーダトランスフォーマ層のうちの少なくとも 1 つのデコーダトランスフォーマ層が、
特徴の 1 つ又は複数のパッチを生成し、前記 1 つ又は複数のパッチにわたって適用された 1 つ又は複数のウィンドウ区分及びシフトウィンドウ区分内で局所的に自己アテンションを決定するように構成された 1 つ又は複数のトランスフォーマブロックと、
前記 1 つ又は複数のパッチの各パッチのそれぞれのサイズを減少させるように構成されたパッチ分離エンジンと、を含む、装置。

【請求項 2】

前記復号画像データのフレームを生成するために、前記少なくとも 1 つのプロセッサが、
前記複数のデコーダトランスフォーマ層のうちの第 1 のデコーダトランスフォーマ層の第 1 のトランスフォーマブロックを使用して、前記 1 つ又は複数のパッチにわたって適用された 1 つ又は複数の第 1 のウィンドウ区分内で局所的に自己アテンションを決定し、
前記第 1 のデコーダトランスフォーマ層の第 2 のトランスフォーマブロックを使用して、前記 1 つ又は複数のパッチにわたって適用された 1 つ又は複数の第 2 のウィンドウ区分であって、前記 1 つ又は複数の第 1 のウィンドウ区分のうちの隣接したものの間の 1 つ又は複数の境界に重なるようにシフトされている、1 つ又は複数の第 2 のウィンドウ区分内で局所的に自己アテンションを決定し、
前記パッチ分離エンジンを使用して、前記 1 つ又は複数のパッチの各パッチを、重なり合わない複数の分離パッチにセグメント化するように構成されている、請求項 1 に記載の装置。

【請求項 3】

前記少なくとも 1 つのプロセッサが、前記複数の分離パッチを前記複数のデコーダトランスフォーマ層のうちの第 2 のデコーダトランスフォーマ層の第 1 のトランスフォーマブロックに提供するように構成されている、請求項 2 に記載の装置。

【請求項 4】

前記少なくとも 1 つのプロセッサが、
前記第 2 のデコーダトランスフォーマ層のパッチ分離エンジンを使用して、前記複数の分離パッチをセグメント化し、
前記パッチ分離エンジンの出力を前記複数のデコーダトランスフォーマ層のうちの第 3 のデコーダトランスフォーマ層に提供するように構成されている、請求項 3 に記載の装置。

【請求項 5】

前記複数の分離パッチの各分離パッチが均一なパッチサイズを有し、前記パッチ分離エンジンが 2 のパッチサイズ縮小係数を適用する、請求項 2 に記載の装置。

【請求項 6】

前記 1 つ又は複数のパッチの各パッチを前記複数の分離パッチにセグメント化するために、前記少なくとも 1 つのプロセッサが、前記複数の分離パッチの特徴次元を減少させるように構成されている、請求項 2 に記載の装置。

【請求項 7】

前記複数のデコーダトランスフォーマ層が、前記符号化画像データのフレームの前記潜

在表現を入力として受け取り、非線形変換を適用し、復号画像データのフレームを生成するように構成されている、請求項 1 に記載の装置。

【請求項 8】

前記非線形変換が合成変換であり、前記復号画像データのフレームが、前記符号化画像データのフレームに関連付けられた入力画像の復元である、請求項 7 に記載の装置。

【請求項 9】

前記少なくとも 1 つのプロセッサが、レート歪みに少なくとも部分的に基づく損失関数を使用して前記複数のデコーダトランスフォーマ層のうちの 1 つ又は複数のデコーダトランスフォーマ層を訓練するように構成されている、請求項 1 に記載の装置。

【請求項 10】

前記損失関数がレート歪みのためのラグランジュ乗数を含む、請求項 9 に記載の装置。

【請求項 11】

前記少なくとも 1 つのデコーダトランスフォーマ層内に含まれる前記 1 つ又は複数のトランスフォーマブロックの少なくとも部分が同じアーキテクチャを有する、請求項 1 に記載の装置。

【請求項 12】

前記少なくとも 1 つのデコーダトランスフォーマ層内に含まれる前記 1 つ又は複数のトランスフォーマブロックの各々が同じアーキテクチャを有する、請求項 1 に記載の装置。

【請求項 13】

前記符号化画像データのフレームが符号化静止画像を含む、請求項 1 に記載の装置。

【請求項 14】

前記符号化画像データのフレームが符号化ビデオフレームを含む、請求項 1 に記載の装置。

【請求項 15】

前記少なくとも 1 つのプロセッサが、少なくとも第 1 の訓練データセット及び第 2 の訓練データセットで前記複数のデコーダトランスフォーマ層を訓練するように構成されており、前記第 2 の訓練データセットのデータが、前記第 1 の訓練データセットのデータと比べて逆の時間的順序を有する、請求項 1 に記載の装置。

【請求項 16】

前記複数のデコーダトランスフォーマ層が一連の連続したデコーダトランスフォーマ層を含む、請求項 1 に記載の装置。

【請求項 17】

メディアデータを処理するための装置であって、

少なくとも 1 つのメモリと、

前記少なくとも 1 つのメモリに結合された少なくとも 1 つのプロセッサと、を備え、前記少なくとも 1 つのプロセッサが、

フレームを複数のパッチにセグメント化し、

前記複数のパッチを入力として使用して、エンコーダサブネットワークの複数のエンコーダトランスフォーマ層に基づいて、符号化画像データのフレームを生成するように構成されている、装置。

【請求項 18】

前記符号化画像データのフレームを生成するために、前記少なくとも 1 つのプロセッサが、

前記複数のエンコーダトランスフォーマ層のうちの第 1 のエンコーダトランスフォーマ層の第 1 のトランスフォーマブロックを使用して、1 つ又は複数のウィンドウ区分内で局所的に自己アテンションを決定し、

前記第 1 のエンコーダトランスフォーマ層の第 2 のトランスフォーマブロックを使用して、前記 1 つ又は複数のウィンドウ区分に重なる 1 つ又は複数のシフトウィンドウ区分内で局所的に自己アテンションを決定し、

前記第 1 のトランスフォーマブロック及び前記第 2 のトランスフォーマブロックのうち

10

20

30

40

50

の１つ又は複数を使用して、非線形変換を前記セグメント化されたフレームに適用するための特徴の１つ又は複数のパッチを決定し、

パッチ併合エンジンを使用して、前記第１のエンコーダトランスフォーマ層と第２のエンコーダトランスフォーマ層との間でパッチサイズを増大させるように構成されている、請求項１７に記載の装置。

【請求項１９】

前記パッチ併合エンジンが、前記第１のエンコーダトランスフォーマ層からの複数の隣接したパッチを、前記第２のエンコーダトランスフォーマ層に提供される併合パッチに組み合わせるように構成されている、請求項１８に記載の装置。

【請求項２０】

前記第１のエンコーダトランスフォーマ層の前記第２のトランスフォーマブロックの出力が前記第２のエンコーダトランスフォーマ層の入力に結合される、請求項１８に記載の装置。

【請求項２１】

前記少なくとも１つのプロセッサが、

前記複数のパッチを入力として使用して、前記エンコーダサブネットワークの前記複数のエンコーダトランスフォーマ層を使用して、前記セグメント化されたフレームのための階層的特徴マップを生成し、

前記階層的特徴マップから前記符号化画像データのフレームを生成するように構成されている、請求項１８に記載の装置。

【請求項２２】

前記複数のパッチの各パッチが均一なサイズのものであり、前記セグメント化されたフレームの１つ又は複数のピクセルを含む、請求項１８に記載の装置。

【請求項２３】

前記パッチ併合エンジンが、隣接したパッチの１つ又は複数の部分セットから取得された特徴を連結することによって前記パッチサイズを増大させるように構成されており、隣接したパッチの各部分セットが、前記パッチ併合エンジンによって出力される併合パッチに併合される、請求項１８に記載の装置。

【請求項２４】

前記第１のトランスフォーマブロック及び前記第２のトランスフォーマブロックが同じアーキテクチャを有する、請求項１８に記載の装置。

【請求項２５】

前記少なくとも１つのプロセッサが、前記第１のエンコーダトランスフォーマ層の前に前記複数のパッチを前記エンコーダサブネットワークの線形埋め込み層に提供するように構成されている、請求項１８に記載の装置。

【請求項２６】

前記符号化画像データのフレームが画像データの潜在表現である、請求項１７に記載の装置。

【請求項２７】

前記潜在表現が、前記エンコーダサブネットワークの前記複数のエンコーダトランスフォーマ層によって生成された階層的特徴マップである、請求項２６に記載の装置。

【請求項２８】

前記少なくとも１つのプロセッサが、レート歪み損失に基づく損失関数を使用して前記複数のエンコーダトランスフォーマ層のうちの１つ又は複数のエンコーダトランスフォーマ層を訓練するように構成されている、請求項１７に記載の装置。

【請求項２９】

前記損失関数がレート歪みのためのラグランジュ乗数を含む、請求項２８に記載の装置。

【請求項３０】

前記複数のパッチが、静止画像フレーム又はビデオフレームを含む入力からセグメント

10

20

30

40

50

化される、請求項 17 に記載の装置。

【請求項 31】

前記少なくとも 1 つのプロセッサが、因数分解プライアを使用して前記符号化画像データをエントロピコーディングするように構成されている、請求項 17 に記載の装置。

【請求項 32】

メディアデータを処理する方法であって、前記方法が、

符号化画像データのフレームの潜在表現を取得することと、

前記符号化画像データのフレームの前記潜在表現を入力として使用して、デコーダサブネットワークの複数のデコーダトランスフォーマ層によって、復号画像データのフレームを生成することと、を含み、前記複数のデコーダトランスフォーマ層のうちの少なくとも 1 つのデコーダトランスフォーマ層が、

10

特徴の 1 つ又は複数のパッチを生成し、前記 1 つ又は複数のパッチにわたって適用された 1 つ又は複数のウィンドウ区分及びシフトウィンドウ区分内で局所的に自己アテンションを決定するための 1 つ又は複数のトランスフォーマブロックと、

前記 1 つ又は複数のパッチの各パッチのそれぞれのサイズを減少させるためのパッチ分離エンジンと、を含む、方法。

【請求項 33】

前記複数のデコーダトランスフォーマ層によって、前記復号画像データのフレームを生成することが、

前記複数のデコーダトランスフォーマ層のうちの第 1 のデコーダトランスフォーマ層の第 1 のトランスフォーマブロックによって、前記 1 つ又は複数のパッチにわたって適用された 1 つ又は複数の第 1 のウィンドウ区分内で局所的に自己アテンションを決定することと、

20

前記第 1 のデコーダトランスフォーマ層の第 2 のトランスフォーマブロックによって、前記 1 つ又は複数のパッチにわたって適用された 1 つ又は複数の第 2 のウィンドウ区分であって、前記 1 つ又は複数の第 1 のウィンドウ区分のうちの隣接したものの間の 1 つ又は複数の境界に重なるようにシフトされている、1 つ又は複数の第 2 のウィンドウ区分内で局所的に自己アテンションを決定することと、

前記パッチ分離エンジンによって、前記 1 つ又は複数のパッチの各パッチを、重なり合わない複数の分離パッチにセグメント化することと、

30

を含む、請求項 32 に記載の方法。

【請求項 34】

前記複数の分離パッチを前記複数のデコーダトランスフォーマ層のうちの第 2 のデコーダトランスフォーマ層の第 1 のトランスフォーマブロックに提供することを更に含む、請求項 33 に記載の方法。

【請求項 35】

前記第 2 のデコーダトランスフォーマ層のパッチ分離エンジンによって、前記複数の分離パッチをセグメント化することと、

前記パッチ分離エンジンの出力を前記複数のデコーダトランスフォーマ層のうちの第 3 のデコーダトランスフォーマ層に提供することと、

40

を更に含む、請求項 34 に記載の方法。

【請求項 36】

前記複数の分離パッチの各分離パッチが均一なパッチサイズを有し、前記パッチ分離エンジンが 2 のパッチサイズ縮小係数を適用する、請求項 33 に記載の方法。

【請求項 37】

前記 1 つ又は複数のパッチの各パッチを前記複数の分離パッチにセグメント化することが、前記複数の分離パッチの特徴次元を減少させることを更に含む、請求項 33 に記載の方法。

【請求項 38】

前記複数のデコーダトランスフォーマ層によって、前記符号化画像データのフレームの

50

前記潜在表現を入力として受け取り、非線形変換を適用し、復号画像データのフレームを生成することを更に含む、請求項 3 2 に記載の方法。

【請求項 3 9】

前記非線形変換が合成変換であり、前記復号画像データのフレームが、前記符号化画像データのフレームに関連付けられた入力画像の復元である、請求項 3 8 に記載の方法。

【請求項 4 0】

レート歪みに少なくとも部分的に基づく損失関数を使用して前記複数のデコーダトランスフォーマ層のうちの 1 つ又は複数のデコーダトランスフォーマ層を訓練することを更に含む、請求項 3 2 に記載の方法。

【請求項 4 1】

前記損失関数がレート歪みのためのラグランジュ乗数を含む、請求項 4 0 に記載の方法。

【請求項 4 2】

前記少なくとも 1 つのデコーダトランスフォーマ層内に含まれる前記 1 つ又は複数のトランスフォーマブロックの少なくとも部分が同じアーキテクチャを有する、請求項 3 2 に記載の方法。

【請求項 4 3】

前記少なくとも 1 つのデコーダトランスフォーマ層内に含まれる前記 1 つ又は複数のトランスフォーマブロックの各々が同じアーキテクチャを有する、請求項 3 2 に記載の方法。

【請求項 4 4】

前記符号化画像データのフレームが符号化静止画像を含む、請求項 3 2 に記載の方法。

【請求項 4 5】

前記符号化画像データのフレームが符号化ビデオフレームを含む、請求項 3 2 に記載の方法。

【請求項 4 6】

少なくとも第 1 の訓練データセット及び第 2 の訓練データセットで前記複数のデコーダトランスフォーマ層を訓練することを更に含み、前記第 2 の訓練データセットのデータが、前記第 1 の訓練データセットのデータと比べて逆の時間的順序を有する、請求項 3 2 に記載の方法。

【請求項 4 7】

前記複数のデコーダトランスフォーマ層が一連の連続したデコーダトランスフォーマ層を含む、請求項 3 2 に記載の方法。

【請求項 4 8】

メディアデータを処理する方法であって、
フレームを複数のパッチにセグメント化することと、
前記複数のパッチを入力として使用して、エンコーダサブネットワークの複数のエンコーダトランスフォーマ層によって、符号化画像データのフレームを生成することと、
を含む、方法。

【請求項 4 9】

前記符号化画像データのフレームを生成することが、
前記複数のエンコーダトランスフォーマ層のうちの第 1 のエンコーダトランスフォーマ層の第 1 のトランスフォーマブロックによって、1 つ又は複数のウィンドウ区分内で局所的に自己アテンションを決定することと、
前記第 1 のエンコーダトランスフォーマ層の第 2 のトランスフォーマブロックによって、前記 1 つ又は複数のウィンドウ区分に重なる 1 つ又は複数のシフトウィンドウ区分内で局所的に自己アテンションを決定することと、
前記第 1 のトランスフォーマブロック及び前記第 2 のトランスフォーマブロックのうちの 1 つ又は複数によって、非線形変換を前記セグメント化されたフレームに適用するための特徴の 1 つ又は複数のパッチを決定することと、

パッチ併合エンジンによって、前記第1のエンコーダトランスフォーマ層と第2のエンコーダトランスフォーマ層との間でパッチサイズを増大させることと、
を含む、請求項48に記載の方法。

【発明の詳細な説明】

【技術分野】

【0001】

【0001】 本開示は、概して、画像及び／又はビデオの符号化（又は圧縮）と復号（解凍）とを含む、画像及びビデオのコーディングに関する。例えば、本開示の態様は、シフトされた自己アテンションウィンドウを有するトランスフォーマ層を使用して変換コーディング及び非線形変換（non-linear transform）を実行するための技法に関する。

10

【背景技術】

【0002】

【0002】 多くのデバイス及びシステムは、消費のためにビデオデータが処理され出力されることを可能にする。デジタルビデオデータは、消費者及びビデオプロバイダの需要を満たすための大量のデータを含む。例えば、ビデオデータの消費者は、高い忠実度、解像度、フレームレートなどを含む高品質のビデオを望む。その結果、これらの需要を満たすことが必要とされる大量のビデオデータは、通信ネットワーク、及びビデオデータを処理して記憶するデバイスに負担をかける。

【0003】

【0003】 ビデオデータを圧縮するためにビデオコーディング技法が使用され得る。ビデオコーディングの目標は、ビデオ品質に対する劣化を回避又は最小化しながら、より低いビットレートを使用する形式にビデオデータを圧縮することである。絶えず進化するビデオサービスが利用可能になるにつれて、より良いコーディング効率を有する符号化技法が必要とされる。

20

【発明の概要】

【0004】

【0004】 いくつかの例では、トランスフォーマベースのニューラルネットワークアーキテクチャを使用してメディアデータ（media data）をコーディングする（例えば、符号化及び／又は復号する）ためのシステム及び技法が説明される。少なくとも1つの例示的な例によれば、メディアデータを処理するための方法であって、符号化画像データ（encoded image data）のフレーム（frame）の潜在表現（latent representation）を取得することと、符号化画像データのフレームの潜在表現を入力として使用して、デコーダサブネットワーク（decoder sub-network）の複数のデコーダトランスフォーマ層（decoder transformer layer）によって、復号画像データ（decoded image data）のフレームを生成することと、を含み、複数のデコーダトランスフォーマ層のうちの少なくとも1つのデコーダトランスフォーマ層が、特徴（feature）の1つ又は複数のパッチ（patch）を生成し、1つ又は複数のパッチにわたって適用された1つ又は複数のウィンドウ区分（window partition）及びシフトウィンドウ区分（shifted window partition）内で局所的に自己アテンション（self-attention）を決定するための1つ又は複数のトランスフォーマブロック（transformer block）と、1つ又は複数のパッチの各パッチのそれぞれのサイズを減少させるためのパッチ分離エンジン（patch un-merging engine）と、を含む、方法が提供される。

30

40

【0005】

【0005】 別の例では、（例えば、仮想コンテンツデータ、1つ又は複数の画像などの、データを記憶するように構成された）少なくとも1つのメモリと、少なくとも1つのメモリに結合された（例えば、回路において実装された）1つ又は複数のプロセッサとを含む、メディアデータを処理するための装置（apparatus）が提供される。1つ又は複数のプロセッサは、符号化画像データのフレームの潜在表現を取得し、符号化画像データのフレームの潜在表現を入力として使用して、デコーダサブネットワークの複数のデコーダトランスフォーマ層に基づいて、復号画像データのフレームを生成するように構成されて

50

おり、それらを行うことができ、複数のデコーダトランスフォーマ層のうちの少なくとも1つのデコーダトランスフォーマ層が、特徴の1つ又は複数のパッチを生成し、1つ又は複数のパッチにわたって適用された1つ又は複数のウィンドウ区分及びシフトウィンドウ区分内で局所的に自己アテンションを決定するように構成された1つ又は複数のトランスフォーマブロックと、1つ又は複数のパッチの各パッチのそれぞれのサイズを減少させるように構成されたパッチ分離エンジンとを含む。

【0006】

【0006】 別の例では、非一時的コンピュータ可読媒体であって、1つ又は複数のプロセッサによって実行されたとき、1つ又は複数のプロセッサに、符号化画像データのフレームの潜在表現を取得させ、符号化画像データのフレームの潜在表現を入力として使用して、デコーダサブネットワークの複数のデコーダトランスフォーマ層に基づいて、復号画像データのフレームを生成させる命令を記憶しており、複数のデコーダトランスフォーマ層のうちの少なくとも1つのデコーダトランスフォーマ層が、特徴の1つ又は複数のパッチを生成し、1つ又は複数のパッチにわたって適用された1つ又は複数のウィンドウ区分及びシフトウィンドウ区分内で局所的に自己アテンションを決定するように構成された1つ又は複数のトランスフォーマブロックと、1つ又は複数のパッチの各パッチのそれぞれのサイズを減少させるように構成されたパッチ分離エンジンとを含む、非一時的コンピュータ可読媒体が提供される。

【0007】

【0007】 メディアデータを処理するための装置が提供される。装置は、符号化画像データのフレームの潜在表現を取得するための手段と、符号化画像データのフレームの潜在表現を入力として使用して、デコーダサブネットワークの複数のデコーダトランスフォーマ層に基づいて、復号画像データのフレームを生成するための手段と、を含み、複数のデコーダトランスフォーマ層のうちの少なくとも1つのデコーダトランスフォーマ層が、特徴の1つ又は複数のパッチを生成し、1つ又は複数のパッチにわたって適用された1つ又は複数のウィンドウ区分及びシフトウィンドウ区分内で局所的に自己アテンションを決定するように構成された1つ又は複数のトランスフォーマブロックと、1つ又は複数のパッチの各パッチのそれぞれのサイズを減少させるように構成されたパッチ分離エンジンとを含む。

【0008】

【0008】 いくつかの態様では、復号画像データのフレームを生成するために、上述された方法、装置、及びコンピュータ可読媒体は、複数のデコーダトランスフォーマ層のうちの第1のデコーダトランスフォーマ層の第1のトランスフォーマブロックによって、1つ又は複数のパッチにわたって適用された1つ又は複数の第1のウィンドウ区分内で局所的に自己アテンションを決定することと、第1のデコーダトランスフォーマ層の第2のトランスフォーマブロックによって、1つ又は複数のパッチにわたって適用された1つ又は複数の第2のウィンドウ区分であって、1つ又は複数の第1のウィンドウ区分のうちの隣接したものの間の1つ又は複数の境界 (boundary) に重なるようにシフトされている、1つ又は複数の第2のウィンドウ区分内で局所的に自己アテンションを決定することと、パッチ分離エンジンによって、1つ又は複数のパッチの各パッチを、重なり合わない複数の分離パッチ (un-merged patch) にセグメント化することと、を含むことができる。

【0009】

【0009】 いくつかの態様では、上述された方法、装置、及びコンピュータ可読媒体は、複数の分離パッチを複数のデコーダトランスフォーマ層のうちの第2のデコーダトランスフォーマ層の第1のトランスフォーマブロックに提供することを含むことができる。

【0010】

【0010】 いくつかの態様では、上述された方法、装置、及びコンピュータ可読媒体は、第2のデコーダトランスフォーマ層のパッチ分離エンジンによって、複数の分離パッチをセグメント化することと、パッチ分離エンジンの出力を複数のデコーダトランスフォー

マ層のうちの第3のデコーダトランスフォーマ層に提供することと、を含むことができる。

【0011】

【0011】 いくつかの態様では、複数の分離パッチの各分離パッチは均一なパッチサイズを有し、パッチ分離エンジンは2のパッチサイズ縮小係数を適用する。

【0012】

【0012】 いくつかの態様では、1つ又は複数のパッチの各パッチを複数の分離パッチにセグメント化するために、上述された方法、装置、及びコンピュータ可読媒体は、複数の分離パッチの特徴次元 (feature dimension) を減少させることを含むことができる。

【0013】

【0013】 いくつかの態様では、上述された方法、装置、及びコンピュータ可読媒体は、複数のデコーダトランスフォーマ層によって、符号化画像データのフレームの潜在表現を入力として受け取り、非線形変換を適用し、復号画像データのフレームを生成することを含むことができる。

【0014】

【0014】 いくつかの態様では、非線形変換は合成変換 (synthesis transform) であり、復号画像データのフレームは、符号化画像データのフレームに関連付けられた入力画像 (input image) の復元 (reconstruction) である。

【0015】

【0015】 いくつかの態様では、上述された方法、装置、及びコンピュータ可読媒体は、レート歪み (rate-distortion) に少なくとも部分的に基づく損失関数 (loss function) を使用して複数のデコーダトランスフォーマ層のうちの1つ又は複数のデコーダトランスフォーマ層を訓練することを含むことができる。いくつかの場合には、損失関数はレート歪みのためのラグランジュ乗数 (Lagrangian multiplier) を含む。

【0016】

【0016】 いくつかの態様では、少なくとも1つのデコーダトランスフォーマ層内に含まれる1つ又は複数のトランスフォーマブロックの少なくとも部分 (portion) は同じアーキテクチャ (architecture) を有する。

【0017】

【0017】 いくつかの態様では、少なくとも1つのデコーダトランスフォーマ層内に含まれる1つ又は複数のトランスフォーマブロックの各々は同じアーキテクチャを有する。

【0018】

【0018】 いくつかの態様では、符号化画像データのフレームは符号化静止画像 (encoded still image) を含む。

【0019】

【0019】 いくつかの態様では、符号化画像データのフレームは符号化ビデオフレーム (encoded video frame) を含む。

【0020】

【0020】 いくつかの態様では、上述された方法、装置、及びコンピュータ可読媒体は、少なくとも第1の訓練データセット (training data set) 及び第2の訓練データセットで複数のデコーダトランスフォーマ層を訓練することを含むことができ、第2の訓練データセットのデータは、第1の訓練データセットのデータと比べて逆の時間的順序 (reversed temporal order) を有する。

【0021】

【0021】 いくつかの態様では、複数のデコーダトランスフォーマ層は一連の連続したデコーダトランスフォーマ層を含む。

【0022】

【0022】 別の例示的な例によれば、メディアデータを処理するための方法が提供される。本方法は、フレームを複数のパッチにセグメント化することと、複数のパッチを入力

10

20

30

40

50

として使用して、エンコーダサブネットワーク (encoder sub-network) の複数のエンコーダトランスフォーマ層 (encoder transformer layer) によって、符号化画像データのフレームを生成することと、を含む。

【0023】

【0023】 別の例では、(例えば、仮想コンテンツデータ、1つ又は複数の画像などの、データを記憶するように構成された) 少なくとも1つのメモリと、少なくとも1つのメモリに結合された(例えば、回路において実装された) 1つ又は複数のプロセッサとを含む、メディアデータを処理するための装置が提供される。1つ又は複数のプロセッサは、フレームを複数のパッチにセグメント化し、複数のパッチを入力として使用して、エンコーダサブネットワークの複数のエンコーダトランスフォーマ層に基づいて、符号化画像データのフレームを生成するように構成されており、それらを行うことができる。

10

【0024】

【0024】 別の例では、非一時的コンピュータ可読媒体であって、1つ又は複数のプロセッサによって実行されたとき、1つ又は複数のプロセッサに、フレームを複数のパッチにセグメント化させ、複数のパッチを入力として使用して、エンコーダサブネットワークの複数のエンコーダトランスフォーマ層に基づいて、符号化画像データのフレームを生成させる命令を記憶した、非一時的コンピュータ可読媒体が提供される。

【0025】

【0025】 メディアデータを処理するための装置が提供される。装置は、フレームを複数のパッチにセグメント化するための手段と、複数のパッチを入力として使用して、エンコーダサブネットワークの複数のエンコーダトランスフォーマ層に基づいて、符号化画像データのフレームを生成するための手段と、を含む。

20

【0026】

【0026】 いくつかの態様では、符号化画像データのフレームを生成するために、上述された方法、装置、及びコンピュータ可読媒体は、複数のエンコーダトランスフォーマ層のうちの第1のエンコーダトランスフォーマ層の第1のトランスフォーマブロックによって、1つ又は複数のウィンドウ区分内で局所的に自己アテンションを決定することと、第1のエンコーダトランスフォーマ層の第2のトランスフォーマブロックによって、1つ又は複数のウィンドウ区分に重なる1つ又は複数のシフトウィンドウ区分内で局所的に自己アテンションを決定することと、第1のトランスフォーマブロック及び第2のトランスフォーマブロックのうちの1つ又は複数によって、非線形変換をセグメント化されたフレーム (segmented frame) に適用するための特徴の1つ又は複数のパッチを決定することと、パッチ併合エンジン (patch merging engine) によって、第1のエンコーダトランスフォーマ層と第2のエンコーダトランスフォーマ層との間でパッチサイズを増大させることと、を含むことができる。

30

【0027】

【0027】 いくつかの態様では、パッチ併合エンジンは、第1のエンコーダトランスフォーマ層からの複数の隣接したパッチ (adjacent patch) を、第2のエンコーダトランスフォーマ層に提供される併合パッチに組み合わせるように構成されている。

【0028】

【0028】 いくつかの態様では、第1のエンコーダトランスフォーマ層の第2のトランスフォーマブロックの出力が第2のエンコーダトランスフォーマ層の入力に結合される。

40

【0029】

【0029】 いくつかの態様では、上述された方法、装置、及びコンピュータ可読媒体は、複数のパッチを入力として使用して、エンコーダサブネットワークの複数のエンコーダトランスフォーマ層によって、セグメント化されたフレームのための階層的特徴マップ (hierarchical feature map) を生成することと、階層的特徴マップから符号化画像データのフレームを生成することと、を含むことができる。

【0030】

【0030】 いくつかの態様では、複数のパッチの各パッチは均一なサイズのものであり

50

、セグメント化されたフレームの1つ又は複数のピクセルを含む。

【0031】

【0031】 いくつかの態様では、上述された方法、装置、及びコンピュータ可読媒体は、パッチ併合エンジンを使用して、隣接したパッチの1つ又は複数の部分セットから取得された特徴を連結することによってパッチサイズを増大させることを含むことができ、隣接したパッチの各部分セットは、パッチ併合エンジンによって出力される併合パッチ（merged patch）に併合される。

【0032】

【0032】 いくつかの態様では、第1のトランスフォーマブロック及び第2のトランスフォーマブロックは同じアーキテクチャを有する。

【0033】

【0033】 いくつかの態様では、上述された方法、装置、及びコンピュータ可読媒体は、第1のエンコーダトランスフォーマ層の前に複数のパッチをエンコーダサブネットワークの線形埋め込み層（linear embedding layer）に提供することを含むことができる。

【0034】

【0034】 いくつかの態様では、符号化画像データのフレームは画像データの潜在表現である。

【0035】

【0035】 いくつかの態様では、潜在表現は、エンコーダサブネットワークの複数のエンコーダトランスフォーマ層によって生成された階層的特徴マップである。

【0036】

【0036】 いくつかの態様では、上述された方法、装置、及びコンピュータ可読媒体は、レート歪み損失（rate-distortion loss）に基づく損失関数を使用して複数のエンコーダトランスフォーマ層のうちの1つ又は複数のエンコーダトランスフォーマ層を訓練することを含むことができる。いくつかの場合には、損失関数はレート歪みのためのラグランジュ乗数を含む。

【0037】

【0037】 いくつかの態様では、複数のパッチは、静止画像フレーム（still image frame）又はビデオフレーム（video frame）を含む入力からセグメント化される。

【0038】

【0038】 いくつかの態様では、上述された方法、装置、及びコンピュータ可読媒体は、因数分解プライア（factorized prior）で符号化画像データをエントロピコーディングすることを含むことができる。

【0039】

【0039】 いくつかの態様では、装置は、モバイルデバイス（例えば、モバイル電話若しくはいわゆる「スマートフォン」、タブレットコンピュータ、又は他のタイプのモバイルデバイス）、ウェアラブルデバイス、エクステンデッドリアリティデバイス（例えば、仮想現実（VR）デバイス、拡張現実（AR）デバイス、又は複合現実（MR）デバイス）、パーソナルコンピュータ、ラップトップコンピュータ、ビデオサーバ、テレビジョン、車両（又は車両のコンピューティングデバイス）、又は他のデバイスを含む。いくつかの態様では、装置は、1つ又は複数の画像又はビデオフレームをキャプチャするための少なくとも1つのカメラを含む。例えば、装置は、1つ若しくは複数の画像及び／又はビデオフレームを含む1つ若しくは複数のビデオをキャプチャするための1つのカメラ（例えば、RGBカメラ）又は複数のカメラを含むことができる。いくつかの態様では、装置は、1つ又は複数の画像、ビデオ、通知、又は他の表示可能なデータを表示するためのディスプレイを含む。いくつかの態様では、装置は、1つ若しくは複数のビデオフレーム及び／又はシンタックスデータを送信媒体を介して少なくとも1つのデバイスに送信するように構成された送信機を含む。いくつかの態様では、プロセッサは、ニューラル処理ユニット（NPU）、中央処理ユニット（CPU）、グラフィックス処理ユニット（GPU）、又は他の処理デバイス若しくは構成要素を含む。

10

20

30

40

50

【0040】

【0040】 本概要では、特許請求される主題の主要な又は必須の特徴を特定することは意図されず、特許請求される主題の範囲を決定するために独立して使用されることも意図されない。本主題は、この特許の明細書全体、いずれか又は全ての図面、及び各請求項の適切な部分を参照することによって理解されるはずである。

【0041】

【0041】 上記については、他の特徴及び実施形態と共に、以下の明細書、特許請求の範囲、及び添付図面を参照すると、より明らかになるう。

【図面の簡単な説明】

【0042】

【0042】 本出願の例示的な実施形態について、以下の図を参照しながら以下で詳細に説明する。

【図1】【0043】 システムオンチップ（SOC）の例示的な実装形態を示す図である。

【図2A】【0044】 全結合ニューラルネットワークの一例を示す図である。

【図2B】【0045】 局所結合ニューラルネットワークの一例を示す図である。

【図3】【0046】 いくつかの例による、ニューラルネットワークベースのシステムを使用して画像及び／又はビデオコーディング（符号化及び復号）を実行するように動作可能なデバイスを含むシステムの一例を示す図である。

【図4】【0047】 いくつかの例による、赤緑青（RGB）フォーマットを有する入力用のエンドツーエンドニューラルネットワークベースの画像及びビデオコーディングシステムの一例を示す図である。

【図5A】【0048】 いくつかの例による、ニューラルネットワークベースの画像及びビデオコーディングシステムのエンコーダのためのトランスフォーマベースのニューラルネットワークアーキテクチャの一例を示す図である。

【図5B】【0049】 いくつかの例による、ニューラルネットワークベースの画像及びビデオコーディングシステムのデコーダのためのトランスフォーマベースのニューラルネットワークアーキテクチャの一例を示す図である。

【図5C】【0050】 いくつかの例による、ニューラルネットワークベースの画像及びビデオコーディングシステムのためのトランスフォーマベースのエンドツーエンドニューラルネットワークアーキテクチャの一例を示す図である。

【図6A】【0051】 いくつかの例による、シフトウィンドウトランスフォーマブロックの対の例示的なアーキテクチャを示す図である。

【図6B】【0052】 いくつかの例による、1つ又は複数のトランスフォーマベースのニューラルネットワークアーキテクチャを使用するビデオコーディングシステムの一例を示す図である。

【図7A】【0053】 いくつかの例による、エンコーダ又はデコーダニューラルネットワークベースの画像及びビデオコーディングシステムのトランスフォーマ層の間で適用することができるパッチ併合又はパッチ分離プロセスの一例を示す図である。

【図7B】【0054】 いくつかの例による、エンコーダ又はデコーダニューラルネットワークベースの画像及びビデオコーディングシステムの2つの自己アテンション層の間のシフトウィンドウ自己アテンション算出の一例を示す図である。

【図8】【0055】 いくつかの例による、画像及び／又はビデオデータを処理するためのプロセスの一例を示すフローチャートである。

【図9】【0056】 いくつかの例による、画像及び／又はビデオデータを処理するためのプロセスの別の例を示すフローチャートである。

【図10】【0057】 本明細書に記載された様々な技法を実装することができる例示的なコンピューティングデバイスの例示的なコンピューティングデバイスアーキテクチャを示す図である。

【発明を実施するための形態】

【0043】

【0058】 本開示の特定の態様及び実施形態が、以下で提供される。当業者に明らかになるように、これらの態様及び実施形態のうちのいくつかが独立して適用されてもよく、それらのうちのいくつかは組み合わせて適用されてもよい。以下の説明では、説明目的で、本出願の実施形態の完全な理解をもたらすために具体的な詳細が記載される。しかしながら、様々な実施形態がこれらの具体的な詳細なしに実践され得ることは明らかであろう。図及び説明は限定的であることが意図されていない。

【0044】

【0059】 以下の説明は、例示的な実施形態を提供するにすぎず、本開示の範囲、適用可能性、又は構成を限定することを意図しない。そうではなく、例示的な実施形態の以下の説明は、例示的な実施形態を実施することを可能にする説明を当業者に提供する。添付の特許請求の範囲に記載されるような本出願の趣旨及び範囲から逸脱することなく、要素の機能及び構成において様々な変更が加えられてよいことを理解されたい。

【0045】

【0060】 デジタルビデオデータは、特に高品質のビデオデータに対する需要が高まり続けているので大量のデータを含むことができる。例えば、ビデオデータの消費者は、通常、高い忠実度、解像度、フレームレートなどを有する、ますます高い品質のビデオを所望する。しかし、高い要求を満たすために必要とされる大量のビデオデータは、多くの場合、大きな帯域幅及びストレージのニーズに結び付き、通信ネットワーク、並びにビデオデータを処理して記憶するデバイスに大きな負担をかける。

【0046】

【0061】 ビデオデータをコーディングするために様々な技法が使用され得る。例えば、ビデオコーディングは、特定のビデオコーディング規格に従って実行され得る。例示的なビデオコーディング規格としては、ムービングピクチャーエクスパートズグループ（moving picture experts group、MPEG）によって開発された高効率ビデオコーディング（high-efficiency video coding、HEVC）、アドバンスドビデオコーディング（advanced video coding、AVC）、及び多用途ビデオコーディング（versatile video coding、VVC）、並びにアライアンスフォーオープンメディア（Alliance for Open Media、AOM）によって開発されたAOMedia Video 1（AV1）が挙げられる。ビデオコーディングは、しばしば、ビデオ画像又はシーケンス内に存在する冗長構成を利用するインター予測又はイントラ予測などの予測方法を使用する。ビデオコーディング技法の共通の目標は、例えば、圧縮されたビデオデータが解凍される際に、ビデオ品質の劣化を回避又は最小化しながら、より低いビットレートを使用する形式にビデオデータを圧縮することである。ビデオサービスに対する需要が高まり新たなビデオサービスが利用可能になるにつれて、より良好なコーディング効率、性能、及びレート制御を有するコーディング技法が必要とされる。

【0047】

【0062】 1つ又は複数の機械学習（ML）システムを使用して画像及び／又はビデオのコーディングを実行するためのシステム、装置、（方法とも呼ばれる）プロセス、及びコンピュータ可読媒体（「システム及び技法」と総称される）が本明細書に記載される。概して、MLは人工知能（artificial intelligence、AI）の部分セットと考えられ得る。MLシステムは、明示的な命令を使用せずにパターン及び推測に依存することによって様々なタスクを実行するためにコンピュータシステムが使用することができるアルゴリズムと統計モデルとを含むことができる。MLシステムの一例は、人工ニューロン（例えば、ニューロンモデル）の相互接続されたグループを含む場合がある、（人工ニューラルネットワークとも呼ばれる）ニューラルネットワークである。ニューラルネットワークは、とりわけ、画像及び／又はビデオのコーディング、画像解析及び／又はコンピュータビジョンアプリケーション、インターネットプロトコル（IP）カメラ、モノのインターネット（IoT）デバイス、自律車両、サービスロボットなどの様々なアプリケーション及び／又はデバイスに使用される場合がある。

【0048】

【0063】ニューラルネットワーク内の個々のノードは、入力データを取得し、データに対して単純な演算を実行することにより、生体ニューロンをエミュレートすることができる。入力データに対して実行される単純な演算の結果は、他のニューロンに選択的に渡されてよい。ネットワークの中の各ベクトル及びノードに重み値が関連付けられ、これらの値は入力データが出力データにどのように関係するのかを制約する。例えば、各ノードの入力データは、対応する重み値で乗算されてよく、その積が合計されてよい。積の合計は任意選択のバイアスによって調整されてよく、その結果に活性化関数が適用されてよく、ノードの出力信号すなわち「出力活性化」（特徴マップ又は活性化マップと呼ばれることがある）が得られる。重み値は、最初に、ネットワークを通した訓練データの反復フローによって決定され得る（例えば、重み値は、ネットワークが、それらの典型的な入力データ特性によって特定のクラスをどのように特定するかを学習する訓練段階の間に確立される）。

10

【0049】

【0064】とりわけ、畳み込みニューラルネットワーク（convolutional neural network、CNN）、再帰型ニューラルネットワーク（recurrent neural network、RNN）、敵対的生成ネットワーク（generative adversarial network、GAN）、多層パーセプトロン（multilayer perceptron、MLP）ニューラルネットワーク、トランスフォーマニューラルネットワークなどの、異なるタイプのニューラルネットワークが存在する。例えば、畳み込みニューラルネットワーク（CNN）は、フィードフォワード人工ニューラルネットワークのタイプである。畳み込みニューラルネットワークは、各々が受容野（例えば、入力空間の空間的に局在化された領域）を有し、かつ入力空間をまとめてタイリングする、人工ニューロンの集合を含んでよい。RNNは、層の出力をセーブするとともに層の結果を予測する助けとなるようにこの出力を入力にフィードバックするというやり方で働く。GANは、ニューラルネットワークモデルが合理的に元のデータセットからである可能性がある新しい合成出力を生成することができるように、入力データ内のパターンを学習することができる生成ニューラルネットワークの一形態である。GANは、合成された出力を生成する生成ニューラルネットワークと、信頼性について出力を評価する識別ニューラルネットワークとを含む、一緒に動作する2つのニューラルネットワークを含むことができる。MLPニューラルネットワークでは、データは入力層の中に供給されてよく、1つ又は複数の隠れ層が抽象性のレベルをデータに与える。次いで、抽象化されたデータに基づいて、出力層において予測が行われてよい。

20

30

【0050】

【0065】深層学習（Deep learning、DL）は機械学習技法の一例であり、MLの部分セットと考えることができる。多くのDLアプローチは、RNN又はCNNなどの、ニューラルネットワークに基づいており、複数の層を利用する。ディープニューラルネットワークにおける複数の層の使用は、生データの所与の入力から、次第に高くなるレベルの特徴が抽出されることを可能にすることができる。例えば、人工ニューロンの第1の層の出力は人工ニューロンの第2の層への入力になり、人工ニューロンの第2の層の出力は人工ニューロンの第3の層への入力になり、以下同様である。ディープニューラルネットワーク全体の入力と出力との間に配置された層は、しばしば、隠れ層と称される。隠れ層は、最終又は所望の表現がディープニューラルネットワークの最終出力として取得されるまで、前の層からの中間入力を、後の層に提供することができる若干より抽象的及び合成的な表現に変換することを学習する（例えば、訓練される）。

40

【0051】

【0066】CNN（及びいくつかの場合には、RNN）は、画像ベースの入力の文脈では、例えば、コンピュータビジョンを実施するか、又は画像／ビデオコーディングを実行するために、一般的に使用される。例えば、画像／ビデオコーディングの文脈では、既存のアプローチは、エントロピ符号化及び復号を実行するために使用される事前モデルを実施するために、自己回帰型CNNを利用する。いくつかの既存のアプローチは、伝統的なコーデックのものに比肩するか、又はそれをしのぐレート歪み性能を達成することが見受

50

けられたが、自己回帰型CNNによって実行される自己回帰復号の複雑さ及びレイテンシのゆえに、低レイテンシの（例えば、リアルタイム又はほぼリアルタイムの）コーディングは達成することができない。

【0052】

【0067】 画像及びビデオコーディングを低レイテンシをもって正確に、及びより効率的に実行するためのシステム及び技法が必要とされる。ML及びDLベースのアプローチは、レート歪み性能の点では既存のコーデックを上回る理論的改善を示したが、これらの既存のアプローチは、多くの場合、（例えば、上述されたように、自己回帰型CNNによって実行される自己回帰復号に基づく）それらの、低レイテンシをもって実行する能力の欠如によって制約される。更に、画像及びビデオデータの解像度は増大し続けるため、これらの既存のアプローチの計算実行時間はおそらく増大し続ける一方であろう。

10

【0053】

【0068】 いくつかの態様では、本明細書において説明されるシステム及び技法は、CNNベースの変換などの、他のニューラルネットワークベースの画像及び／又はビデオコーディングシステムよりも高速のコーディング（例えば、復号）を使用する低レイテンシの画像及び／又はビデオコーディングを実行することができるトランスフォーマベースの画像及びビデオコーディングシステムを含む。例えば、トランスフォーマベースの変換を使用することで、本明細書において説明されるシステム及び技法は、Kodak画像圧縮データセットを使って、少なくとも7%、及びUltra Video Group (UVG) ビデオ圧縮データセット（低遅延モード）を使って、12%のコーディング効率向上を達成することができる。

20

【0054】

【0069】 トランスフォーマは、入力データの各部分の有意性を異なって重み付けし、長期依存性をモデル化するために、アテンション機構を利用するタイプの深層学習モデルである。トランスフォーマは、多くの場合、順次的な入力データを取り扱うために使用されるが、トランスフォーマは、必ずしも、データが元々受信又は配列された同じ順序でデータを処理するとは限らない。更に、トランスフォーマはアテンションを使用して入力データの小部分の間のコンテキスト関係を決定することができるため、トランスフォーマは、アテンション又は自己アテンションを計算する際などに、小部分の一部又は全てを並列に処理することができる。この並列化は、例えば、同じタスクを実行するために訓練されたRNN、CNN、又は他のニューラルネットワークと比べて、より大きな計算柔軟性をもたすことができる。

30

【0055】

【0070】 いくつかの場合には、本明細書において説明されるトランスフォーマベースの画像及びビデオコーディングシステムは、複数の連続したシフトウィンドウトランスフォーマ層を各々含む、エンコーダサブネットワーク及びデコーダサブネットワークを含む。エンコーダサブネットワークは、入力画像を潜在表現に転換するための変換を適用し、その一方で、デコーダサブネットワークは、潜在表現を、入力画像の復元バージョンである、復元画像に転換するための変換を適用する。入力画像は、静止画像（例えば、写真及び他の種類の静止画像）並びにビデオ画像（例えば、ビデオのフレーム）を含むことができる。いくつかの例では、エンコーダ及びデコーダサブネットワークは、（例えば、CNNベースのアプローチにおいて使用される、計算がより複雑な自己回帰プライアとは対照的に）因数分解プライアと併せて使用され、画像データの低レイテンシ符号化及び復号を可能にする、非線形変換を適用することができる。更に、本明細書において説明されるトランスフォーマベースの画像及びビデオコーディングシステムは、既存のニューラルネットワーク画像及びビデオコーディングシステム（例えば、CNNベースの画像及びビデオコーディングシステム）に付随するレート歪み損失に匹敵する、又はそれを改善するレート歪み損失をもって、この低レイテンシ符号化及び復号を達成することができる。

40

【0056】

【0071】 いくつかの例では、デコーダサブネットワークは、エンコーダサブネットワ

50

ーク内に含まれるパッチ併合エンジンの機能性にとって代わり、それを逆にするパッチ分離エンジン除いて、エンコーダサブネットワークと対称的である。エンコーダ及びデコーダサブネットワークは一連のパッチ（本明細書において「パッチトークン」とも称される）にわたって演算する。いくつかの場合には、一連のパッチは、最初に、エンコーダサブネットワークにおいて、入力画像の重なり合わないセグメント化として形成される。

【0057】

【0072】 いくつかの態様では、エンコーダサブネットワーク及びデコーダサブネットワークは、シフトウィンドウのアプローチを使用する変更された自己アテンション計算を利用する。シフトウィンドウのアプローチは、自己アテンション計算を、重なり合わない局所ウィンドウに限定し、その一方で、また、クロスウィンドウ結合を可能にする自己アテンション層の対（自己アテンション層対とも称される）に基づく。いくつかの例では、クロスウィンドウ結合は、下位の自己アテンション層の複数の別個のアテンションウィンドウに各々またがる（例えば、それらからの要素に基づく）アテンションベクトルを決定するために、より深い自己アテンション層内で利用され得る。例示的な例では、自己アテンション層対の第1の層は第1の区分構成を適用し、パッチのセットを、複数のパッチを各々包含する非オーバーラップウィンドウに分割することができ、その後、自己アテンションが各ウィンドウ内で局所的に計算される。自己アテンション層対の第2の自己アテンション層内では、ウィンドウ区分はシフトされ、第1の自己アテンション層からのウィンドウに重なる新たなウィンドウをもたらす。第2の自己アテンション層のシフトウィンドウ内における自己アテンション計算は第1の自己アテンション層内の以前のウィンドウの境界を横断し、これにより、それらの間の結合をもたらす。

【0058】

【0073】 自己アテンション計算を第1の自己アテンション層の重なり合わない局所ウィンドウに限定することによって、本明細書において説明されるトランスフォーマベースの画像及びビデオコーディングシステムは、低レイテンシの画像及びビデオ符号化及び復号（並びにいくつかの場合には、高速エントロピモデルが使用されるときなどには、リアルタイム若しくはほぼリアルタイムの符号化及び復号）を支援する、より高い効率及び計算性能を達成することができる。例えば、いくつかの態様では、トランスフォーマベースの画像及びビデオコーディングシステムは画像サイズに対して線形の計算複雑性を有する。第2の自己アテンション層のシフトウィンドウ区分によって決定されたクロスウィンドウ結合を導入することによって、本明細書において説明されるトランスフォーマベースの画像及びビデオコーディングシステムは、CNNベースのアプローチ及び他の既存の画像及びビデオコーディングのアプローチに付随するレート歪み損失に匹敵するか、又はそれを改善するレート歪み損失を達成することができる。

【0059】

【0074】 本開示の様々な態様が、図に関して説明される。図1は、本明細書に記載された機能のうちの1つ以上を実行するように構成された中央処理装置（CPU）102又はマルチコアCPUを含む場合があるシステムオンチップ（SOC）100の例示的な実装形態を示す。いくつかある情報の中で特に、パラメータ又は変数（例えば、ニューラル信号及びシナプス荷重）、算出デバイスに関連付けられたシステムパラメータ（例えば、重みを有するニューラルネットワーク）、遅延、周波数ビン情報、タスク情報は、ニューラル処理装置（NPU）108に関連付けられたメモリブロック、CPU102に関連付けられたメモリブロック、グラフィックス処理装置（GPU）104に関連付けられたメモリブロック、デジタル信号プロセッサ（DSP）106に関連付けられたメモリブロック、メモリブロック118に記憶される場合があり、及び／又は複数のブロックにわたって分散される場合がある。CPU102において実行される命令は、CPU102に関連付けられたプログラムメモリからロードされてもよく、又はメモリブロック118からロードされてもよい。

【0060】

【0075】 SOC100はまた、GPU104、DSP106、第5世代（5G）接続

性、第4世代ロングタームエボリューション（4G LTE）接続性、Wi-Fi接続性、USB接続性、Bluetooth接続性などを含んでもよい接続性ブロック110、並びに、例えば、ジェスチャを検出及び認識する場合があるマルチメディアプロセッサ112などの、特定の機能に適合された追加の処理ブロックを含んでもよい。一実装形態では、NPUは、CPU102、DSP106、及び／又はGPU104内に実装される。SOC100はまた、センサープロセッサ114、画像信号プロセッサ（ISP）116、及び／又は全地球測位システムを含んでもよいナビゲーションモジュール120を含んでもよい。

【0061】

【0076】 SOC100は、ARM命令セットに基づいてもよい。本開示の一態様では、CPU102にロードされる命令は、入力値とフィルタ重みの乗算積に対応するルックアップテーブル（LUT）内の記憶された乗算結果を探索するためのコードを備えることができる。CPU102にロードされる命令はまた、乗算積のルックアップテーブルヒットが検出されたとき、乗算積の乗算演算中に乗算器を無効にするためのコードを備えることができる。更に、CPU102にロードされる命令は、乗算積のルックアップテーブルミスが検出されたとき、入力値とフィルタ重みの算出された乗算積を記憶するためのコードを備えることができる。

【0062】

【0077】 SOC100及び／又はその構成要素は、本明細書で説明された本開示の態様に従って機械学習技法を使用して、（ビデオ符号化及び／又は復号とも呼ばれ、ビデオコーディングと総称される）ビデオ圧縮及び／又は解凍を実行するように構成される場合がある。ビデオ圧縮及び／又は解凍を実行するために深層学習アーキテクチャを使用することにより、本開示の態様は、デバイス上でのビデオ圧縮及び／又は解凍の効率を増大させることができる。例えば、記載されたビデオコーディング技法を使用するデバイスは、機械学習ベースの技法を使用してより効率的にビデオを圧縮することができ、別のデバイスに圧縮されたビデオを送信することができ、その別のデバイスは、本明細書に記載された機械学習ベースの技法を使用して、圧縮されたビデオをより効率的に解凍することができる。

【0063】

【0078】 上述のように、ニューラルネットワークは機械学習システムの一例であり、入力層、1つ又は複数の隠れ層、及び出力層を含むことができる。入力層の入力ノードからデータが提供され、1つ又は複数の隠れ層の隠れノードによって処理が実行され、出力層の出力ノードを通じて出力が生成される。深層学習ネットワークは、通常、複数の隠れ層を含む。ニューラルネットワークの各層は、人工ニューロン（又は、ノード）を含むことができる特徴マップ又は活性化マップを含むことができる。特徴マップは、フィルタ、カーネルなどを含むことができる。ノードは、層のうちの1つ又は複数のノードの重要度を示すために使用される1つ又は複数の重みを含むことができる。場合によっては、深層学習ネットワークは一連の多くの隠れ層を有することができ、初期の層は入力単純で低レベルの特性を決定するために使用され、後の層はより複雑で抽象的な特性の階層を構築する。

【0064】

【0079】 深層学習アーキテクチャは、特徴の階層を学習してもよい。視覚データが提示された場合、例えば、第1の層は、入力ストリームにおけるエッジなどの比較的単純な特徴を認識することを学習してもよい。別の実施例では、聴覚データが提示される場合、第1の層は、特定の周波数におけるスペクトル電力を認識することを学習することができる。第1の層の出力を入力として取る第2の層は、視覚データの場合は単純な形状などの特徴の組み合わせ、又は聴覚データの場合は音の組み合わせを認識することを学習してもよい。例えば、上位層は、視覚データにおける複雑な形状又は聴覚データにおける単語を表現することを学習することができる。更に上位の層は、一般的な視覚オブジェクト又は話されたフレーズを認識することを学習してもよい。

【0065】

【0080】 深層学習アーキテクチャは、自然階層構造を有する問題に適用されたとき、特によく機能する場合がある。例えば、電動ビークルの分類は、車輪、フロントガラス、及び他の特徴を認識することを最初に学習することから恩恵を受ける場合がある。これらの特徴は、自動車、トラック、及び飛行機を認識するために、異なる方法で上位層において組み合わせられてもよい。

【0066】

【0081】 ニューラルネットワークは、様々な結合性パターンを用いて設計されてもよい。フィードフォワードネットワークでは、所与の層内の各ニューロンが上位層内のニューロンと通信して、情報が下位層から上位層に渡される。上述したように、フィードフォワードネットワークの連続する層内で階層表現が構築されてもよい。ニューラルネットワークはまた、再帰結合又はフィードバック（トップダウンとも呼ばれる）結合を有してもよい。再帰結合では、所与の層内のニューロンからの出力は、同じ層内の別のニューロンに伝達されてもよい。再帰型アーキテクチャは、シーケンスの中でニューラルネットワークに配信される入力データチャンクのうちの2つ以上にわたるパターンを認識する際に有用であり得る。所与の層内のニューロンから下位層内のニューロンへの結合は、フィードバック（又は、トップダウン）結合と呼ばれる。多数のフィードバック結合を有するネットワークは、高レベルの概念の認識が入力の特定の低レベルの特徴を区別する助けとなり得るときに有用であり得る。

【0067】

【0082】 ニューラルネットワークの層間の結合は、全結合又は局所結合されていてもよい。図2Aは、全結合ニューラルネットワーク202の一例を示す。全結合ニューラルネットワーク202では、第1の層内のニューロンは、その出力を第2の層内のあらゆるニューロンに伝達してもよく、その結果、第2の層内の各ニューロンは、第1の層内のあらゆるニューロンから入力を受信することになる。図2Bは、局所結合ニューラルネットワーク204の一例を示す。局所結合ニューラルネットワーク204では、第1の層内のニューロンは、第2の層内の限られた数のニューロンに結合されていてもよい。より一般的には、局所結合ニューラルネットワーク204の局所結合層は、ある層内の各ニューロンが同じ又は類似の結合性パターンを、しかし異なる値（例えば、210、212、214、及び216）を有してもよい結合強度を伴って、有するように構成することができる。所与の領域内の上位層ニューロンは、訓練を通じてネットワークへの全入力のうちの限定された部分の特性に調整される入力を受信し得るので、局所結合の結合性パターンは、上位層内で空間的に別個の受容野を生じ得る。

【0068】

【0083】 上述のように、デジタルビデオデータは、通信ネットワーク、並びにビデオデータを処理して記憶するデバイスに著しい負担をかける場合がある、大量のデータを含むことができる。例えば、非圧縮ビデオコンテンツを記録することは、一般に、記録されたビデオコンテンツの解像度が高まるにつれて大幅に増大する大規模なファイルサイズをもたらす。例示的な一例では、1080p/24（例えば、毎秒24フレームでキャプチャされる、幅が1920ピクセルかつ高さが1080ピクセルという解像度）で記録される、チャンネル当たり16ビットの非圧縮ビデオは、フレーム当たり12.4メガバイト、すなわち毎秒297.6メガバイトを占有し得る。毎秒24フレームで4K解像度で記録される、チャンネル当たり16ビットの非圧縮ビデオは、フレーム当たり49.8メガバイト、すなわち毎秒1195.2メガバイトを占有し得る。

【0069】

【0084】 ネットワーク帯域幅は、大きいビデオファイルが問題になり得る別の制約である。例えば、ビデオコンテンツは、しばしば、ワイヤレスネットワークを介して（例えば、LTE、LTEアドバンスド、新無線（NR）、Wi-Fi（商標）、Bluetooth（商標）、又は他のワイヤレスネットワークを介して）配信され、消費者のインターネットトラフィックの大部分を構成する可能性がある。ワイヤレスネットワークにおける

利用可能な帯域幅の量の進歩にもかかわらず、これらのネットワークにおいてビデオコンテンツを配信するために使用される帯域幅の量を削減することが依然として望ましい場合がある。

【0070】

[0085] 圧縮されていないビデオコンテンツは、物理記憶用のかなりのメモリと送信用のかなりの帯域幅とを要する場合がある大きいファイルをもたらす可能性があるので、そのようなビデオコンテンツを圧縮し、次いで解凍するためにビデオコーディング技法が利用され得る。

【0071】

[0086] ビデオコンテンツのサイズ、したがってビデオコンテンツを記憶するために要するストレージの量、及びビデオコンテンツを配信する際に要する帯域幅の量を削減するために、とりわけ、H E V C、A V C、M P E G、V V Cなどの特定のビデオコーディング規格に従って、様々なビデオコーディング技法が実行され得る。ビデオコーディングは、しばしば、ビデオ画像又はシーケンス内に存在する冗長構成を利用するインター予測又はイントラ予測などの予測方法を使用する。ビデオコーディング技法の共通の目標は、ビデオ品質の劣化を回避又は最小化しながら、より低いビットレートを使用する形式にビデオデータを圧縮することである。ビデオサービスに対する需要が高まり新たなビデオサービスが利用可能になるにつれて、より良好なコーディング効率、性能、及びレート制御を有するコーディング技法が必要とされる。

【0072】

[0087] 概して、符号化デバイスは、ビデオコーディング規格に従ってビデオデータを符号化し、符号化されたビデオビットストリームを生成する。いくつかの例では、符号化ビデオビットストリーム（又は「ビデオビットストリーム」若しくは「ビットストリーム」）は、一連の1つ又は複数のコーディングされたビデオシーケンスである。符号化デバイスは、各ピクチャを複数のスライスに区分することによってピクチャのコード化表現を生成することができる。スライスは、そのスライスの中の情報が同じピクチャ内の他のスライスからのデータへの依存関係なしにコーディングされるように、他のスライスとは無関係である。スライスは、独立スライスセグメント、及び存在する場合、前のスライスセグメントに依存する1つ又は複数の従属スライスセグメントを含む、1つ又は複数のスライスセグメントを含む。H E V Cでは、スライスは、ルーマサンプル及びクロマサンプルのコーディングツリーブロック（C T B）に区分される。ルーマサンプルのC T B、及びクロマサンプルの1つ又は複数のC T Bは、サンプル用のシンタックスと一緒にコーディングツリーユニット（C T U）と呼ばれる。C T Uは、「ツリーブロック」又は「最大コーディングユニット」（L C U）と呼ばれることもある。C T Uは、H E V C符号化のための基本処理単位である。C T Uは、様々なサイズの複数のコーディングユニット（C U）にスプリットされ得る。C Uは、コーディングブロック（C B）と呼ばれるルーマサンプルアレイ及びクロマサンプルアレイを含む。

【0073】

[0088] ルーマC B及びクロマC Bは、更にスプリットされて予測ブロック（P B）になり得る。P Bは、（利用可能であるか又は使用のために有効化されるとき）インター予測又はイントラブロックコピー（I B C）予測のために同じ動きパラメータを使用するルーマ成分又はクロマ成分のサンプルのブロックである。ルーマP B及び1つ又は複数のクロマP Bは、関連するシンタックスと一緒に、予測ユニット（P U）を形成する。インター予測の場合、動きパラメータのセット（例えば、1つ又は複数の動きベクトル、参照インデックスなど）は、P Uごとにビットストリームにおいてシグナリングされ、ルーマP B及び1つ又は複数のクロマP Bのインター予測のために使用される。動きパラメータは動き情報と呼ばれることもある。C Bはまた、1つ又は複数の変換ブロック（T B）に区分され得る。T Bは、予測残差信号をコーディングするために残差変換（例えば、場合によっては、同じ2次元変換）がそこに適用される色成分のサンプルの正方形ブロックを表す。変換ユニット（T U）は、ルーマサンプル及びクロマサンプルのT B、並びに対応

10

20

30

40

50

するシンタックス要素を表す。以下でより詳細に変換コーディングが説明される。

【0074】

【0089】 H E V C 規格によれば、変換は T U を使用して実行され得る。T U は、所与の C U 内の P U のサイズに基づいてサイズ決定されてよい。T U は、同じサイズであってもよく、又は P U よりも小さくてもよい。いくつかの例では、C U に対応する残差サンプルは、残差四分木 (R Q T) と呼ばれる 4 分木構造を使用して、より小さいユニットに再分割され得る。R Q T のリーフノードは T U に相当し得る。T U に関連するピクセル差分値は、変換係数を生成するために変換され得る。変換係数は、次いで、符号化デバイスによって量子化され得る。

【0075】

【0090】 ビデオデータのピクチャが C U に区分されると、符号化デバイスは予測モードを使用して各 P U を予測する。予測ユニット又は予測ブロックは、次いで、残差（以下で説明する）を得るために元のビデオデータから減算される。C U ごとに、シンタックスデータを使用してビットストリームの内部で予測モードがシグナリングされ得る。予測モードは、イントラ予測（又は、イントラピクチャ予測）又はインター予測（又は、インターピクチャ予測）を含んでよい。イントラ予測は、ピクチャ内で空間的に近隣のサンプル間の相関を利用する。例えば、イントラ予測を使用すると、例えば、P U にとっての平均値を見つけるための D C 予測、平坦面を P U に適合させるための平面予測、隣接データから外挿するための方向予測、又は任意の他の好適なタイプの予測を使用して、同じピクチャの中の隣接する画像データから各 P U が予測される。インター予測は、画像サンプルのブロックに対する動き補償予測を導出するために、ピクチャ間の時間相関を使用する。例えば、インター予測を使用すると、各 P U は、（出力順序で現在のピクチャの前又は後の）1 つ又は複数の参照ピクチャの中の画像データからの動き補償予測を使用して予測される。ピクチャエリアを、インターピクチャ予測を使用してコーディングすべきか、それともイントラピクチャ予測を使用してコーディングすべきかという決定は、例えば、C U レベルにおいて行われてよい。

【0076】

【0091】 イントラ予測及び／又はインター予測を使用して予測を実行した後、符号化デバイスは変換及び量子化を実行することができる。例えば、予測に続いて、符号化デバイスは P U に対応する残差値を計算し得る。残差値は、コーディングされているピクセルの現在ブロック (P U) と現在ブロックを予測するために使用される予測ブロック（例えば、現在ブロックの予測されたバージョン）との間のピクセル差分値を含み得る。例えば、予測ブロックを生成した（例えば、インター予測又はイントラ予測を発行した）後に、符号化デバイスは、現在ブロックから予測ユニットによって生成された予測ブロックを減算することによって残差ブロックを生成することができる。残差ブロックは、現在のブロックのピクセル値と予測ブロックのピクセル値との間の差分を定量化するピクセル差分値のセットを含む。いくつかの例では、残差ブロックは、2 次元ブロックフォーマット（例えば、ピクセル値の 2 次元行列又はアレイ）で表され得る。そのような例では、残差ブロックはピクセル値の 2 次元表現である。

【0077】

【0092】 予測が実行された後に残っている場合がある任意の残差データがブロック変換を使用して変換され、ブロック変換は、離散コサイン変換、離散サイン変換、整数変換、ウェーブレット変換、他の好適な変換機能、又はそれらの任意の組み合わせに基づいてよい。場合によっては、1 つ又は複数のブロック変換（例えば、サイズ 32×32 、 16×16 、 8×8 、 4×4 、又は他の好適なサイズ）が、各 C U における残差データに適用され得る。いくつかの実施形態では、符号化デバイスによって実施される変換プロセス及び量子化プロセスのために T U が使用され得る。1 つ又は複数の P U を有する所与の C U は、1 つ又は複数の T U も含んでよい。以下で更に詳細に説明するように、残差値は、ブロック変換を使用して変換係数に変換されてよく、次いで、T U を使用して量子化及び走査されて、エントロピコーディングのためのシリアル化変換係数を生成し得る。

【0078】

【0093】 符号化デバイスは変換係数の量子化を実行し得る。量子化は、係数を表すために使用されるデータの量を減らすために変換係数を量子化することによって、更なる圧縮をもたらす。例えば、量子化は、係数の一部又は全部に関連付けられたビット深度を低減し得る。一例では、 n ビット値を有する係数は、量子化中に m ビット値に切り捨てられてよく、 n は m よりも大きい。

【0079】

【0094】 量子化が実行されると、コーディングされたビデオビットストリームは、量子化された変換係数、予測情報（例えば、予測モード、動きベクトル、ブロックベクトルなど）、区分情報、及び他のシンタックスデータなど、任意の他の適切なデータを含む。コード化ビデオビットストリームの様々な要素が、次いで、符号化デバイスによってエントロピ符号化され得る。いくつかの例では、符号化デバイスは、既定の走査順序を利用して量子化変換係数を走査して、エントロピ符号化され得るシリアル化ベクトルを生成し得る。いくつかの例では、符号化デバイスは適応走査を実行し得る。量子化変換係数を走査してベクトル（例えば、1次元ベクトル）を形成した後、符号化デバイスはベクトルをエントロピ符号化し得る。例えば、符号化デバイスは、コンテキスト適応型可変長コーディング、コンテキスト適応型バイナリ算術コーディング、シンタックススペースコンテキスト適応型バイナリ算術コーディング、確率間隔区分エントロピコーディング、又は別の好適なエントロピ符号化技法を使用してよい。

【0080】

【0095】 符号化デバイスは、符号化ビデオビットストリームを記憶することができ、かつ／又は通信リンクを介して受信デバイスへ符号化ビデオビットストリームデータを送ることができ、受信デバイスは復号デバイスを含むことができる。復号デバイスは、符号化ビデオデータを構成する1つ又は複数のコード化ビデオシーケンスの要素を（例えば、エントロピデコーダを使用して）エントロピ復号及び抽出することによって、符号化ビデオビットストリームデータを復号し得る。復号デバイスは、次いで、符号化ビデオビットストリームデータを再スケーリングしてよく、符号化ビデオビットストリームデータに対して逆変換を実行してよい。次いで、残差データが復号デバイスの予測ステージに渡される。復号デバイスは、次いで、イントラ予測、インター予測、IBC、及び／又は他のタイプの予測を使用して、ピクセルのブロック（例えば、PU）を予測する。いくつかの例では、予測は、逆変換の出力（残差データ）に追加される。復号デバイスは、復号ビデオをビデオ宛先デバイスに出力してよく、ビデオ宛先デバイスは、復号ビデオデータをコンテンツの消費者に表示するためのディスプレイ又は他の出力デバイスを含んでよい。

【0081】

【0096】 様々なビデオコーディング規格（例えば、上述されたHEVCビデオコーディング技法）によって定義されたビデオコーディングのシステム及び技法は、未加工のビデオコンテンツ内の情報の大部分を保持することができる場合があり、信号処理及び情報理論の概念に基づいてアプリオリに定義される場合がある。しかしながら、場合によっては、機械学習（ML）ベースの画像及び／又はビデオシステムは、エンドツーエンドニューラルネットワークベースの画像及びビデオコーディング（E2E-NNVC）システムなどの、非MLベースの画像及びビデオコーディングシステムよりも利益を提供することができる。上述されたように、多くのE2E-NNVCシステムは、オートエンコーダサブネットワーク（エンコーダサブネットワーク）、及びエントロピコーディングに使用される量子化ラテント上で確率モデルを学習することに関する第2のサブネットワークの組み合わせとして設計される。そのようなアーキテクチャは、変換プラス量子化モジュール（エンコーダサブネットワーク）及びエントロピモデリングサブネットワークモジュールの組み合わせとして見ることができる。

【0082】

【0097】 図3は、E2E-NNVCシステム310を使用して画像及び／又はビデオの符号化及び復号を実行するように構成されたデバイス302を含むシステム300を示

す。デバイス 302 は、カメラ 307 及び記憶媒体 314（例えば、データ記憶デバイス）に結合されている。いくつかの実装形態では、カメラ 307 は、E2E-NNVC システム 310 による符号化のためにプロセッサ 304 に画像データ 308（例えば、ビデオデータストリーム）を提供するように構成されている。いくつかの実装形態では、デバイス 302 は、複数のカメラ（例えば、デュアルカメラシステム、3つのカメラ、又は他の数のカメラ）に結合されることができ、かつ／又はそれらを含むことができる。場合によっては、デバイス 302 は、マイクロフォン並びに／又は他の入力デバイス（例えば、キーボード、マウス、タッチスクリーン及び／若しくはタッチパッドなどのタッチ入力デバイス、並びに／又は他の入力デバイス）に結合され得る。いくつかの例では、カメラ 307、記憶媒体 314、マイクロフォン、及び／又は他の入力デバイスは、デバイス 302 の部分であり得る。

10

【0083】

【0098】 デバイス 302 はまた、1つ若しくは複数のワイヤレスネットワーク、1つ若しくは複数の有線ネットワーク、又はそれらの組み合わせなどの伝送媒体 318 を介して、第2のデバイス 390 に結合されている。例えば、伝送媒体 318 は、ワイヤレスネットワーク、有線ネットワーク、又は有線ネットワークとワイヤレスネットワークの組み合わせによって提供されるチャネルを含むことができる。伝送媒体 318 は、ローカルエリアネットワーク、ワイドエリアネットワーク、又はインターネットなどのグローバルネットワークなどのパケットベースネットワークの一部を形成し得る。伝送媒体 318 は、ルータ、スイッチ、基地局、又はソースデバイスから受信デバイスへの通信を容易にするために有用であり得る任意の他の機器を含み得る。ワイヤレスネットワークは、任意のワイヤレスインターフェース又はワイヤレスインターフェースの組み合わせを含んでよく、任意の好適なワイヤレスネットワーク（例えば、インターネット又は他のワイドエリアネットワーク、パケットベースネットワーク、Wi-Fi（商標）、無線周波数（RF）、UWB、Wi-Fi-Direct、セルラー、ロングタームエボリューション（LTE）、WiMax（商標）など）を含んでよい。有線ネットワークは、任意の有線インターフェース（例えば、ファイバ、イーサネット、電力線イーサネット、同軸ケーブルを介したイーサネット、デジタル信号ライン（DSL）など）を含んでよい。有線ネットワーク及び／又はワイヤレスネットワークは、基地局、ルータ、アクセスポイント、ブリッジ、ゲートウェイ、スイッチなどの様々な機器を使用して実装され得る。符号化ビデオビットストリームデータは、ワイヤレス通信プロトコルなどの通信規格に従って変調され、受信デバイスに送信され得る。

20

30

【0084】

【0099】 デバイス 302 は、メモリ 306、第1のインターフェース（「I/F1」）312、及び第2のインターフェース（「I/F2」）316 に結合された（本明細書では「プロセッサ」と呼ばれる）1つ又は複数のプロセッサ 304 を含む。プロセッサ 304 は、カメラ 307 から、メモリ 306 から、及び／又は記憶媒体 314 から画像データ 308 を受け取るように構成されている。プロセッサ 304 は、第1のインターフェース 312 を介して（例えば、メモリバスを介して）記憶媒体 314 に結合され、第2のインターフェース 316（例えば、ネットワークインターフェースデバイス、ワイヤレストランシーバ及びアンテナ、1つ又は複数の他のネットワークインターフェースデバイス、又はそれらの組み合わせ）を介して伝送媒体 318 に結合される。

40

【0085】

【0100】 プロセッサ 304 は E2E-NNVC システム 310 を含む。E2E-NNVC システム 310 は、エンコーダ部分 362 とデコーダ部分 366 とを含む。いくつかの実装形態では、E2E-NNVC システム 310 は、1つ又は複数のオートエンコーダを含むことができる。エンコーダ部分 362 は、入力データ 370 を受け取り、入力データ 370 に少なくとも部分的に基づいて出力データ 374 を生成するために入力データ 370 を処理するように構成されている。

【0086】

50

【0101】 いくつかの実装形態では、E 2 E - N N V C システム 3 1 0 のエンコーダ部分 3 6 2 は、出力データ 3 7 4 を生成するために入力データ 3 7 0 の不可逆圧縮を実行するように構成され、その結果、出力データ 3 7 4 は入力データ 3 7 0 よりも少ないビットを有する。エンコーダ部分 3 6 2 は、任意の前の表現（例えば、1 つ又は複数の前に復元されたフレーム）に基づく動き補償を使用せずに入力データ 3 7 0（例えば、画像又はビデオフレーム）を圧縮するように訓練され得る。例えば、エンコーダ部分 3 6 2 は、ビデオフレームを、そのビデオフレームからのビデオデータのみを使用し、前に復元されたフレームのいかなるデータも使用せずに圧縮することができる。エンコーダ部分 3 6 2 によって処理されるビデオフレームは、本明細書ではイントラ予測フレーム（I フレーム）と呼ばれ得る。いくつかの例では、I フレームは、従来のビデオコーディング技法を使用して（例えば、H E V C、V V C、M P E G - 4、又は他のビデオコーディング規格に従って）生成され得る。このような例では、プロセッサ 3 0 4 は、H E V C 規格に関して上述されたものなどの、ブロックベースのイントラ予測を実行するように構成されたビデオコーディングデバイス（例えば、符号化デバイス）を含んでもよく、又はそれと結合されてもよい。このような例では、E 2 E - N N V C システム 3 1 0 は、プロセッサ 3 0 4 から除外される場合がある。

10

【0087】

【0102】 いくつかの実装形態では、E 2 E - N N V C システム 3 1 0 のエンコーダ部分 3 6 2 は、前の表現（例えば、1 つ又は複数の前に復元されたフレーム）に基づく動き補償を使用して入力データ 3 7 0（例えば、ビデオフレーム）を圧縮するように訓練され得る。例えば、エンコーダ部分 3 6 2 は、ビデオフレームを、そのビデオフレームからのビデオデータを使用し、前に復元されたフレームのデータを使用して圧縮することができる。エンコーダ部分 3 6 2 によって処理されるビデオフレームは、本明細書ではイントラ予測フレーム（P フレーム）と呼ばれ得る。動き補償は、前に復元されたフレームからのピクセルが残差情報とともに現在フレーム内の新しい位置にどのように移動するかを記述することにより、現在フレームのデータを決定するために使用され得る。

20

【0088】

【0103】 図示されたように、E 2 E - N N V C システム 3 1 0 のエンコーダ部分 3 6 2 は、ニューラルネットワーク 3 6 3 と量子化器 3 6 4 とを含むことができる。ニューラルネットワーク 3 6 3 は、1 つ若しくは複数のトランスフォーマ、1 つ若しくは複数の畳み込みニューラルネットワーク（C N N）、1 つ若しくは複数の全結合ニューラルネットワーク、1 つ若しくは複数のゲート付き再帰型ユニット（gated recurrent unit、G R U）、1 つ若しくは複数の長短期記憶（Long Short-Term Memory、L S T M）ネットワーク、1 つ若しくは複数の C o n v R N N、1 つ若しくは複数の C o n v G R U、1 つ若しくは複数の C o n v L S T M、1 つ若しくは複数の G A N、それらの任意の組み合わせ、及び／又は中間データ 3 7 2 を生成する他のタイプのニューラルネットワークアーキテクチャを含むことができる。中間データ 3 7 2 は量子化器 3 6 4 への入力である。エンコーダ部分 3 6 2 に含まれ得る構成要素の例が、図 5 A 及び図 6 A に示されている。

30

【0089】

【0104】 量子化器 3 6 4 は、出力データ 3 7 4 を生成するために中間データ 3 7 2 の量子化と、場合によってはエントロピコーディングとを実行するように構成されている。出力データ 3 7 4 は、量子化（及び場合によってはエントロピコード化）データを含むことができる。量子化器 3 6 4 によって実行される量子化演算は、中間データ 3 7 2 から、量子化コード（又は E 2 E - N N V C システム 3 1 0 によって生成された量子化コードを表すデータ）の生成をもたらすことができる。量子化コード（又は量子化コードを表すデータ）はまた、（z と表記された）ラテントコード又はラテントと呼ばれ得る。ラテントに適用されるエントロピモデルは、本明細書では「プライア」と呼ばれ得る。いくつかの例では、量子化及び／又はエントロピコーディング演算は、既存のビデオコーディング規格に従ってビデオデータを符号化及び／又は復号するときに実行される既存の量子化及びエントロピコーディング演算を使用して実行され得る。いくつかの例では、量子化及び／

40

50

又はエントロピコーディング演算は、E 2 E - N N V C システム 3 1 0 によって行われ得る。1つの例示的な例では、E 2 E - N N V C システム 3 1 0 は、教師あり訓練を使用して訓練されることができ、訓練中に残差データが入力として使用され、量子化コード及びエントロピコードが既知の出力（ラベル）として使用される。

【0090】

【0105】 E 2 E - N N V C システム 3 1 0 のデコーダ部分 3 6 6 は、（例えば、量子化器 3 6 4 から直接、及び／又は記憶媒体 3 1 4 から）出力データ 3 7 4 を受け取るように構成されている。デコーダ部分 3 6 6 は、出力データ 3 7 4 に少なくとも部分的に基づいて入力データ 3 7 0 の表現 3 7 6 を生成するために、出力データ 3 7 4 を処理することができる。いくつかの例では、E 2 E - N N V C システム 3 1 0 のデコーダ部分 3 6 6 は、1つ若しくは複数のトランスフォーマ、1つ若しくは複数の C N N、1つ若しくは複数の全結合ニューラルネットワーク、1つ若しくは複数の G R U、1つ若しくは複数の L S T M ネットワーク、1つ若しくは複数の C o n v R N N、1つ又は複数の C o n v G R U、1つ若しくは複数の C o n v L S T M、1つ若しくは複数の G A N、それらの任意の組み合わせ、及び／又は他のタイプのニューラルネットワークアーキテクチャを含み得るニューラルネットワーク 3 6 8 を含む。デコーダ部分 3 6 6 に含まれ得る構成要素の例が、図 5 B 及び図 6 A に示されている。

10

【0091】

【0106】 プロセッサ 3 0 4 は、伝送媒体 3 1 8 又は記憶媒体 3 1 4 のうちの少なくとも1つに出力データ 3 7 4 を送るよう構成されている。例えば、出力データ 3 7 4 は、復元データとして入力データ 3 7 0 の表現 3 7 6 を生成するためにデコーダ部分 3 6 6 による後の検索及び復号（又は解凍）のために記憶媒体 3 1 4 において記憶される場合がある。復元データは、出力データ 3 7 4 を生成するために符号化／圧縮されたビデオデータの再生用などの様々な目的に使用され得る。いくつかの実装形態では、出力データ 3 7 4 は、復元データとして入力データ 3 7 0 の表現 3 7 6 を生成するために、（例えば、デバイス 3 0 2 内の、第 2 のデバイス 3 9 0 内の、又は別のデバイス内の）デコーダ部分 3 6 6 に一致する別のデコーダデバイスにおいて復号される場合がある。例えば、第 2 のデバイス 3 9 0 は、デコーダ部分 3 6 6 に一致する（又は実質的に一致する）デコーダを含む場合があり、出力データ 3 7 4 は、伝送媒体 3 1 8 を介して第 2 のデバイス 3 9 0 に送信される場合がある。第 2 のデバイス 3 9 0 は、復元データとして入力データ 3 7 0 の表現 3 7 6 を生成するために、出力データ 3 7 4 を処理することができる。

20

30

【0092】

【0107】 システム 3 0 0 の構成要素は、1つ又は複数のプログラマブル電子回路（例えば、マイクロプロセッサ、グラフィックス処理ユニット（G P U）、デジタル信号プロセッサ（D S P）、中央処理ユニット（C P U）、及び／又は他の好適な電子回路）を含むことができる、電子回路若しくは他の電子ハードウェアを含むことができ、かつ／又はそれらを使用して実装されることが可能であり、並びに／あるいは、本明細書で説明される様々な動作を実行するために、コンピュータソフトウェア、ファームウェア、若しくはそれらの任意の組み合わせを含むことができ、かつ／又はそれらを使用して実装されることが可能である。

40

【0093】

【0108】 システム 3 0 0 はいくつかの構成要素を含むものとして示されるが、システム 3 0 0 は図 3 に示されるものより多数又は少数の構成要素を含むことができることを、当業者は理解するだろう。例えば、システム 3 0 0 はまた、入力デバイスと出力デバイス（図示せず）とを含むコンピューティングデバイスを含むことができるか、又はその一部であり得る。いくつかの実装形態では、システム 3 0 0 は、1つ若しくは複数のメモリデバイス（例えば、1つ若しくは複数のランダムアクセスメモリ（random access memory、R A M）構成要素、リードオンリーメモリ（read-only memory、R O M）構成要素、キャッシュメモリ構成要素、バッファ構成要素、データベース構成要素、及び／若しくは他のメモリデバイス）、1つ若しくは複数のメモリデバイスと通信している、及び

50

／若しくはそれらに電氣的に接続される1つ若しくは複数の処理デバイス（例えば、1つ若しくは複数のCPU、GPU、及び／又は他の処理デバイス）、ワイヤレス通信を実行するための1つ若しくは複数のワイヤレスインターフェース（例えば、各ワイヤレスインターフェースのための1つ若しくは複数のトランシーバ及びベースバンドプロセッサを含む）、1つ若しくは複数のハードワイヤード接続を介して通信を実行するための1つ若しくは複数の有線インターフェース（例えば、ユニバーサルシリアルバス（universal serial bus、USB）入力、lighteningコネクタ、及び／若しくは他の有線インターフェースなどの、シリアルインターフェース）、並びに／又は、図3に示されない他の構成要素を含むコンピューティングデバイスも含んでもよく、又はそれらの一部であることが可能である。

10

【0094】

【0109】いくつかの実装形態では、システム300は、コンピューティングデバイスによって局所的に実装され、及び／又はコンピューティングデバイスに含まれ得る。例えば、コンピューティングデバイスは、モバイルデバイス、パーソナルコンピュータ、タブレットコンピュータ、仮想現実（VR）デバイス（例えば、ヘッドマウントディスプレイ（HMD）若しくは他のVRデバイス）、拡張現実（AR）デバイス（例えば、HMD、ARグラス、若しくは他のARデバイス）、ウェアラブルデバイス、（例えば、サービスとしてのソフトウェア（SaaS）システム若しくは他のサーバベースシステム内の）サーバ、テレビジョン、及び／又は本明細書に記載された技法を実行するリソース能力を有する任意の他のコンピューティングデバイスを含むことができる。

20

【0095】

【0110】一例では、E2E-NNVCシステム310は、プロセッサ304に結合され、プロセッサ304によって実行可能な命令を記憶するように構成されたメモリ306と、アンテナ及びプロセッサ304に結合され、リモートデバイスに出力データ374を伝送するように動作可能なワイヤレストランシーバとを含むポータブル電子デバイスに組み込まれ得る。

【0096】

【0111】図4は、画像及び／又はビデオコーディングのためのハイパープライア（hyperprior）モデルを実施するために畳み込みニューラルネットワーク層を使用するE2E-NNVCシステムの一例を示す図である。図4のE2E-NNVCシステム内の g_a サブネットワーク及び g_s サブネットワークは、それぞれ、エンコーダサブネットワーク（例えば、エンコーダ部分362）及びデコーダサブネットワーク（例えば、デコーダ部分366）に対応する。図4の g_a サブネットワーク及び g_s サブネットワークは、3チャンネルRGB入力向けに設計され、全ての3つのR入力チャンネル、G入力チャンネル、及びB入力チャンネルは、同じニューラルネットワーク層（畳み込み層及び一般化神経活動正規化（generalized divisive normalization、GDN）層）を通過し、それらによって処理される。（図4に示されたものなどの）E2E-NNVCシステムは、RGBデータ（ここで、異なるRチャンネル、Gチャンネル、及びBチャンネルの統計特性は同様である）及び／又はYUVデータなどの同様の統計特性を有する入力チャンネルを対象とすることができる。しかし、上述されたように、画像コーディングを実行するために訓練されたCNNは、（遅い復号時間をもたらす）自己回帰型プライアの使用に基づくなどして、符号化又は復号動作のいずれかの低レイテンシ性能を依然として達成することができない。

30

40

【0097】

【0112】上述されたように、1つ又は複数のトランスフォーマニューラルネットワークを使用して画像及び／又はビデオコーディング（例えば、低レイテンシ符号化及び復号）を実行するためのシステム及び技法が本明細書において説明される。トランスフォーマニューラルネットワークは、例えば、図4のハイパープライアアーキテクチャ及び／又は後述される図6Bのスケールスペースフロー（scale-space flow、SSF）アーキテクチャに従って編成されたトランスフォーマブロック及び／又はトランスフォーマ層を含むことができる。例えば、図4に示される4つの畳み込みネットワーク g_a 、 g_s 、 h_a

50

、及び h_s は、以下において更に深く説明されるように、代わりに、対応する 4 つのトランスフォーマニューラルネットワークとして提供することができる。

【0098】

【0113】 いくつかの例では、本明細書において説明される 1 つ又は複数のトランスフォーマベースのニューラルネットワークは、レート歪みに少なくとも部分的に基づく損失関数を使用して訓練することができる。歪みは、元の画像（例えば、エンコーダサブネットワークへの入力として提供されるであろう画像）と解凍／復号画像（例えば、デコーダサブネットワークによって復元された画像）との間の平均 2 乗誤差（mean square error、MSE）として決定され得る。いくつかの例では、トランスフォーマベースのメディアコーディングニューラルネットワークを訓練する際に使用される損失関数は、ラグランジュ乗数を用いた歪みとレートとの間のトレードオフに基づくことができる。このようなレート歪み損失関数の一例は $L = D + \beta * R$ である。ここで、 D は歪みを表し、 R はレートを表し、異なる β 値は、異なるビットレート及び／又はピーク信号対雑音比（peak-signal-to-noise ratio、PSNR）のために訓練されたモデルを表す。

10

【0099】

【0114】 1 つの例示的な例では、本明細書において説明されるトランスフォーマベースのニューラルネットワークのうちの 1 つ又は複数を訓練するために、教師なし学習技法（監督を行わない）を使用することができる。教師なし訓練プロセスでは、訓練データの部分又はその要素をラベル付け又は分類する必要がなくなり得る。例えば、ニューラルネットワーク、例えば、図 5 A 及び図 5 B にそれぞれ示されるものなどの、エンコーダ及び／又はデコーダサブネットワーク）のノードの重み（及びいくつかの場合には、バイアスなどの他のパラメータ）を調整するために、バックプロパゲーション訓練プロセスを使用することができる。バックプロパゲーションは、フォワードパス、損失関数、バックワードパス、及び重み更新を含む。いくつかの例では、損失関数は、上述されたレート歪みベースの損失関数を含むことができる。フォワードパス、損失関数、バックワードパス、及びパラメータ更新は、1 回の訓練反復の間に実行され得る。プロセスは、エンコーダ又はデコーダサブネットワークのパラメータの重みが正確に調節されるまで、訓練データのセットごとに一定数の反復にわたって繰り返される。

20

【0100】

【0115】 実際の出力値は訓練データ出力とは大きく異なり得るため、最初の訓練データ入力に対して損失（又は、誤差）は高くなり得る。訓練の目標は、予測出力のための損失（例えば、損失関数 $L = D + \beta * R$ を使用するなどする、レート歪み損失）の量を最小化することである。ニューラルネットワークは、どの入力（重み）がニューラルネットワークの損失に最も寄与したのかを決定することによってバックワードパスを実行し、損失が減少し、最終的に最小化されるように重みを調整する。ニューラルネットワークの損失に最も寄与した重みを決定するために、重みに対する損失の導関数（ dL / dW として示され、ここで、 W は特定の層における重みである）が計算される。例えば、重みは、それが勾配の反対方向に変化するように更新される。重み更新は

30

【0101】

【数 1】

$$w = w_i - \eta \frac{dL}{dw}$$

40

【0102】

として示される。ここで、 w は重みを示し、 w_i は初期重みを示し、 η は学習率を示す。学習率は任意の好適な値に設定され、高い学習率はより大きい重み更新を含み、より小さい値はより小さい重み更新を示す。エンコーダ又はデコーダサブネットワーク、あるいは 1 つのこのようなサブネットワークの構成要素のシフトウィンドウトランスフォーマ層又はブロックは、所望の出力が達成されるまでこのような仕方での訓練が続けられる。いくつかの場合には、エンコーダサブネットワーク及び／又はデコーダサブネットワークの

50

構成要素の各々は同様の仕方で訓練される。

【0103】

【0116】 シフトウィンドウトランスフォーマブロックの例示的なアーキテクチャ（図6A）、シフトウィンドウトランスフォーマ層の間のパッチ併合の一例（図7A）、及び所与のトランスフォーマ層内のトランスフォーマブロックの対によって実行されるシフトウィンドウ自己アテンションの一例（図7B）を提供する、図6～図7Bの説明へ進む前に、図5A及び図5Bが以下において簡単に紹介される。上述の例の後に、説明は図5A及び図5Bへ戻り、更に詳細に行われる。

【0104】

【0117】 図5Aは、一連のトランスフォーマ層520、530、540、550を有するエンコーダサブネットワーク500aの例示的なアーキテクチャを示す。いくつかの例では、エンコーダサブネットワーク500aは、入力画像502を潜在表現555aに転換する非線形変換を実行するために訓練され得る。いくつかの場合には、非線形変換はまた、分析変換である。図示のように、各エンコーダトランスフォーマ層は、連続的に配列されたシフトウィンドウトランスフォーマブロック524、534、544、及び554のセット（本明細書において「エンコーダトランスフォーマブロックセット」又は「トランスフォーマブロックセット」とも称される）をそれぞれ含むことができる。いくつかの例では、エンコーダトランスフォーマブロックセット524～554は、図6Aに示されるシフトウィンドウトランスフォーマブロックの例示的な対などの、シフトウィンドウトランスフォーマブロックの1つ又は複数の対を含むことができる。例えば、エンコーダブロックセット524～554からのエンコーダトランスフォーマブロックセット内に提供されるシフトウィンドウトランスフォーマブロックの総数は2の倍数であることができる（例えば、「x2」、「x6」、及び図5Aにおける同様の表記によって示されるように、トランスフォーマブロックセット524内に2つのシフトウィンドウトランスフォーマブロック、トランスフォーマブロックセット534内に2つのシフトウィンドウトランスフォーマブロック、トランスフォーマブロックセット544内に6つのシフトウィンドウトランスフォーマブロック、及びトランスフォーマブロックセット554内に2つのシフトウィンドウトランスフォーマブロック）。2つのシフトウィンドウトランスフォーマブロックを有するトランスフォーマブロックセットの一例が図6Aに示されており、以下において説明される。いくつかの場合には、シフトウィンドウトランスフォーマブロックの対は、（例えば、図6A及び図7Bに関して説明されるように）連続したシフトウィンドウトランスフォーマブロックにわたって自己アテンションを計算するための2つの区分構成の間の使用を交互に行う2ステップの自己アテンション算出プロセスに対応することができる。

【0105】

【0118】 図5Bは、パッチ区分エンジン510b及び一連のトランスフォーマ層560、570、580、590を有するデコーダサブネットワーク500bの例示的なアーキテクチャを示す。パッチ区分エンジン510bはエンコーダサブネットワーク500aのパッチ区分エンジン510aと同様に動作することができる。いくつかの例では、デコーダサブネットワーク500bは、潜在表現555bを復元画像504に転換する非線形変換を実行するために訓練され得る。いくつかの場合には、非線形変換はまた、合成変換である。量子化、エントロピコーディング、及び／又は潜在表現555aをエンコーダサブネットワーク500aからデコーダサブネットワーク500bへ通信するデータ伝送プロセスにおける情報損失又は毀損が存在しない場合などには、潜在表現555bは、エンコーダサブネットワーク500aによって出力された潜在表現555aと同じになることができる。図示のように、各デコーダトランスフォーマ層は、連続的に配列されたシフトウィンドウトランスフォーマブロック564、574、584、及び594のセット（本明細書において「デコーダトランスフォーマブロックセット」又は「トランスフォーマブロックセット」とも称される）をそれぞれ含むことができる。いくつかの例では、デコーダトランスフォーマブロックセット564～594は、図6Aに示されるシフトウィンド

ウトランスフォーマブロックの例示的な対などの、シフトウィンドウトランスフォーマブロックの1つ又は複数の対を含むことができる。例えば、トランスフォーマデコーダブロックセット564～594からのデコーダトランスフォーマブロックセット内に提供されるシフトウィンドウトランスフォーマブロックの総数は2の倍数であることができる（例えば、「x2」、「x6」、及び図5Bにおける同様の表記によって示されるように、トランスフォーマブロックセット564内に2つのシフトウィンドウトランスフォーマブロック、トランスフォーマブロックセット574内に6つのシフトウィンドウトランスフォーマブロック、トランスフォーマブロックセット584内に2つのシフトウィンドウトランスフォーマブロック、及びトランスフォーマブロックセット594内に2つのシフトウィンドウトランスフォーマブロック）。いくつかの場合には、シフトウィンドウトランスフォーマブロックの対は、（例えば、図6A及び図7Bに関するように）連続したシフトウィンドウトランスフォーマブロックにわたって自己アテンションを計算するための2つの区分構成の間の使用を交互に行う2ステップの自己アテンション算出プロセスに対応することができる。

10

20

30

40

50

【0106】

【0119】 いくつかの例では、本明細書においてエンコーダサブネットワーク500aを参照して行われる説明の一部又は全てはデコーダサブネットワーク500bにも適用され得ることが留意される。例えば、いくつかの場合には、デコーダトランスフォーマ層560～590のうちの1つ又は複数の、エンコーダトランスフォーマ層520～550のシフトウィンドウトランスフォーマブロックと同一であるか、又はさもなければ、それらと共通のアーキテクチャを共有するシフトウィンドウトランスフォーマブロックを利用することができる。更に、いくつかの例では、デコーダサブネットワーク500bのアーキテクチャはエンコーダサブネットワーク500aのアーキテクチャと対称であることができ、デコーダサブネットワーク500bはエンコーダサブネットワーク500aのパッチ併合エンジン532、542、552の代わりにパッチ分離エンジン563、573、583を使用する。デコーダサブネットワーク500b及びエンコーダサブネットワーク500aのアーキテクチャの間の対称性は、デコーダ及びエンコーダサブネットワーク内に提供された構成要素のシフトウィンドウトランスフォーマブロックの同じアーキテクチャ配置又は構成の使用を含むことができる。a

【0107】

【0120】 図5Cは、ニューラルネットワークベースの画像及びビデオコーディングシステムのためのトランスフォーマベースのエンドツーエンドニューラルネットワークアーキテクチャの一例を示す図である。いくつかの場合には、図5Cのトランスフォーマベースのエンドツーエンドニューラルネットワークアーキテクチャは図5Aのエンコーダサブネットワーク500a及び図5Bのデコーダサブネットワーク500bを含むことができる。パッチ分割エンジン（図5Cにおいて「パッチ分割」として示される）は、図5Bに示されるパッチ分離エンジンと同様であり、それらと同じ動作を実行することができる。

【0108】

【0121】 次に、図5A及び図5Bに戻る前に、説明は、図6～図7Bに示される例へ進む。前述のように、図6～図7Bは、シフトウィンドウトランスフォーマブロックの例示的なアーキテクチャ（図6A）、シフトウィンドウトランスフォーマ層の間のパッチ併合の一例（図7A）、及び所与のトランスフォーマ層内のトランスフォーマブロックの対によって実行されるシフトウィンドウ自己アテンションの一例を提供する。

【0109】

【0122】 図6Aは、まとめてトランスフォーマブロック対と称される、第1のシフトウィンドウトランスフォーマブロック601及び第2のシフトウィンドウトランスフォーマブロック603の例示的なアーキテクチャ600を示す。図示のように、第1のシフトウィンドウトランスフォーマブロック601は、層ノルム（layer norm）610a、自己アテンション構成要素622（「第1の自己アテンション層」とも称される）、層ノルム612a、及び多層パーセプトロン（MLP）として示される、フィードフォワードニ

ューラルネットワーク構成要素 630 a を含む。第 2 のシフトウィンドウトランスフォーマブロック 603 は、層ノルム 610 b、自己アテンション構成要素 624（「第 2 の自己アテンション層」とも称される）、層ノルム 612 b、及びこの場合も先と同様に、MLP として示される、フィードフォワードニューラルネットワーク構成要素 630 b を含む。

【0110】

【0123】 いくつかの例では、第 1 のシフトウィンドウトランスフォーマブロック 601 は、異なるウィンドウ区分構成を適用する、それらのそれぞれの自己アテンション層 622 及び 624 を除いて、第 2 のシフトウィンドウトランスフォーマブロック 603 と同じであることができる。いくつかの場合には、第 1 の自己アテンション層 622 はウィンドウ化マルチヘッド自己アテンション（windowed multi-head self-attention、W-MSA）を含むことができ、第 2 の自己アテンション層 624 はシフトウィンドウマルチヘッド自己アテンション（shifted window multi-head self-attention、SW-MSA）を含むことができる。

【0111】

【0124】 いくつかの例では、第 1 のシフトウィンドウトランスフォーマブロック 601 の第 1 の自己アテンション層 622 は、シフトサイズ＝0（無シフトに対応する）を有するシフトウィンドウを使用することができる。例えば、ウィンドウサイズが 8 であるとき、第 1 のアテンション層（例えば、第 1 の自己アテンション層 622）はシフトサイズ＝0 を使用することができる、第 2 のアテンション層（第 2 の自己アテンション層 624 などの、第 1 のアテンション層に続くもの）はシフトサイズ＝4 を使用することができる、第 3 のアテンション層（第 2 のアテンション層に続くもの）は、再び、シフトサイズ＝0 を使用することができる、第 4 のアテンション層（第 3 のアテンション層に続くもの）はシフトサイズ＝4 を使用することができる、（例えば、「x2」、「x6」、及び図 5 A 及び図 5 B における同様の表記によって示されるとおりの）所与のトランスフォーマブロックセット内のシフトウィンドウトランスフォーマブロックの個数のために、以下同様である。シフトサイズ値を交互にすること（例えば、シフト＝0 とシフト＝4 とを交互にすること）は、信号をウィンドウ間で伝搬する効果をもたらす。いくつかの場合には、シフトサイズは可変であることができるため、シフト＝0 の後にシフト＝4 を有する必要はない。

【0112】

【0125】 いくつかの例では、第 1 の自己アテンション層 622 は、非オーバーラップウィンドウ区分構成（図 7 B の構成 720 など）を適用し、パッチのセットを、複数のパッチを各々包含する非オーバーラップウィンドウに分割することができる。第 1 の自己アテンション層 622 は、次に、各ウィンドウ内で局所的に自己アテンションを計算することができる。第 1 の自己アテンション層 622 は自己アテンション情報を層ノルム 612 a の層（例えば、ソフトマックス層）に提供することができる。例えば、第 1 の自己アテンション層 622 は、次式のように出力の行列を計算することによって自己アテンション値を計算することができる：

【0113】

【数 2】

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

【0114】

【0126】 ここで、行列 $Q = W_q X$ 、行列 $K = W_k X$ 、及び行列 $V = W_v X$ 、並びに Q 、 K 、 V への入力と同じ X である（及びそれゆえ「自己（self）」である）。 W_q 、 W_k 、及び W_v の項は、入力ベクトル X をクエリ（ Q ）、キー（ K ）、及びバリュー（ V ）行列へ射影又はマップする線形層である。項 d_k はキー k の次元を指し、

【0115】

【数 3】

$$\sqrt{d_k}$$

【0116】

はスケーリングファクタの役割を果たす。softmaxは、自己アテンション値への重みを得るために使用されるソフトマックス関数を指す。層ノルム612aは重みをフィードフォワードニューラルネットワーク構成要素630a（例えば、多層パーセプトロン（MLP）層）に出力することができる。第1のシフトウィンドウトランスフォーマブロック601の出力は、次に、第2のシフトウィンドウトランスフォーマブロック603への入力として提供され得る。

10

【0117】

【0127】 第2の自己アテンション層624内では、ウィンドウ区分はシフトされ、第1の自己アテンション層622のものに重なる新たなウィンドウをもたらす。例えば、シフトウィンドウ区分構成（図7Bの構成730など）が第2の自己アテンション層624によって適用され得る。第2の自己アテンション層624のシフトウィンドウ内における自己アテンション計算は第1の自己アテンション層622内の以前のウィンドウの境界を横断し、層ノルム612bに提供され得るクロスウィンドウ結合をもたらす。層ノルム612bは出力をフィードフォワードニューラルネットワーク構成要素630bに提供することができる。

20

【0118】

【0128】 自己アテンションを局所的に決定することによって、本明細書において説明されるトランスフォーマベースの画像及びビデオコーディングシステムは、低レイテンシ符号化及び復号を支援するより高い効率及び計算性能を達成することができる。例えば、いくつかの態様では、トランスフォーマベースの画像及びビデオコーディングシステムは画像サイズに対して線形の計算複雑性を有する。第2の自己アテンション層を介してクロスウィンドウ結合を得ることによって、トランスフォーマベースの画像及びビデオコーディングシステムは、CNNベースのアプローチ及び他の既存のアプローチに付随するレート歪み損失に匹敵するか、又はそれを改善するレート歪み損失を達成することができる。

【0119】

30

【0129】 図7Aは、エンコーダサブネットワーク500aの文脈では、より深いエンコーダトランスフォーマ層内において（例えば、下から上へ）画像パッチを併合する例示的なプロセスを示し、以下においてより詳細に説明される。図7Bは、ウィンドウ区分構成720及びウィンドウ区分構成730を含む、2つの異なるウィンドウ区分構成の一例を示す。ウィンドウ区分構成720は、パッチトークンのセットにわたって適用された非オーバーラップウィンドウ区分を示し、いくつかの例では、図6Aの第1の自己アテンション層622によって利用され得る。非オーバーラップウィンドウ区分の一例が722において指示されており、その構成要素のパッチトークンのうちの1つの一例が711において指示されている。いくつかの例では、図6Aの第1のシフトウィンドウトランスフォーマブロック601は、自己アテンション構成要素622を使用して非オーバーラップウィンドウ区分構成720を適用することができる。

40

【0120】

【0130】 ウィンドウ区分構成730は、パッチトークンのセットにわたって適用されたシフトウィンドウ区分を示し、いくつかの例では、図6Aの第2の自己アテンション層624によって利用され得る。いくつかの場合には、2つのウィンドウ区分構成720及び730はパッチトークンの同じセットにわたって適用され得る。

【0121】

【0131】 いくつかの例では、非オーバーラップウィンドウ区分構成720は入力パッチトークンのセットを、ここでは、16個のパッチトークンを包含する4×4のウィンドウとして示される、等しいサイズのウィンドウに分割する。ただし、他のウィンドウ幾何

50

構成及び／又はサイズを利用することもできる。シフトウィンドウ区分構成 730 は、重なり合わない区分構成 720 のものに対して変位させられたウィンドウを利用することができる。例えば、シフトウィンドウ 732 及び 734 は、それらが、区分構成 720 の非オーバーラップウィンドウのうちの複数の異なるものに以前に包含されていたトークンのセットを各々包含するように変位させられた。単一のシフトウィンドウは以前の自己アテンション層の複数の非オーバーラップウィンドウからのパッチトークンを包含するため、これにより、上述のクロスウィンドウ結合が導入され得る。図示のように、シフトウィンドウ区分構成 730 は、非オーバーラップウィンドウ区分構成と同じ 4 x 4 のウィンドウサイズを使用し、それがパッチトークンセットの境界を越えて延びる所ではウィンドウサイズの切り落とし又は切り捨てを伴う。しかし、いくつかの例では、シフトウィンドウ部
10
分化構成 730 及び非オーバーラップウィンドウ区分構成 720 は異なるウィンドウサイズを使用することができる。

【0122】

【0132】 次に、図 5 A 及び図 5 B に示されるエンコーダ及びデコーダサブネットワークに戻ると、図 5 A に示されるように、エンコーダサブネットワーク 500 a の第 1 のトランスフォーマ層 520 は 2 つのシフトウィンドウトランスフォーマブロックを含み（例えば、シフトウィンドウトランスフォーマブロック 524 の下の「x2」ラベルによって示される）、トランスフォーマ層 530 は 2 つのシフトウィンドウトランスフォーマブロックを含み、トランスフォーマ層 540 は 6 つのシフトウィンドウトランスフォーマブロックを含み、トランスフォーマ層 550 は 2 つのシフトウィンドウトランスフォーマブ
20
ロックを含む。いくつかの例では、本明細書において説明されるトランスフォーマ層のうちの 1 つ又は複数のは、図 5 A の例において示されるのよりも多数又は少数のシフトウィンドウトランスフォーマブロックを各々含むことができることが留意される。

【0123】

【0133】 トランスフォーマ層 520 ~ 550（及びそれらの構成要素のシフトウィンドウトランスフォーマブロック）は、いくつかの例では、ビジョントランスフォーマに基づくことができる。ビジョントランスフォーマは、画像ベースの入力を使って動作するように構成されるタイプのトランスフォーマアーキテクチャである。画像ベースの入力は、静止画像（例えば、写真又は他の種類の静止画像）並びにビデオ画像（例えば、ビデオのフレーム）を含むことができる。ビジョントランスフォーマアーキテクチャの下では、ま
30
ず、入力画像は複数の非オーバーラップパッチに区分され、次に、それらのパッチは、ビジョントランスフォーマ（複数可）が演算し得るベクトルを得るために線形空間に射影される。

【0124】

【0134】 エンコーダサブネットワーク 500 a は、入力画像 502 を非オーバーラップパッチに分割するパッチ区分エンジン 510 a を含む（図 7 A に、2 つのパッチ区分演算の一例が示されている）。入力画像 502 は、H x W x 3 の次元を有するように指示されている。ここで、H はピクセル単位の高さを指示し、W はピクセル単位の幅を指示し、数 3 は入力画像の次元数（入力画像の色チャネル、例えば、R、G、及び B ごとに 1 つの次元）を指示する。各パッチは、例えば、高さ及び幅次元におけるピクセルの数によって
40
与えられる、同じ又は同様のサイズを有することができる。いくつかの例では、パッチ区分エンジン 510 a によって生成されるパッチのサイズは事前に決定され得る。図示のように、事前に決定されたパッチサイズは 4 x 4 であり、これにより、パッチ区分エンジン 510 a によって生成されるパッチは、辺ごとに 4 つのピクセルの正方形に配列された 16 個のピクセルを含む。ただし、他のサイズ及び／又は高さ－幅比を利用することもできる。加えて、パッチ区分エンジン 510 はエンコーダサブネットワーク 500 a のプロセスフローの部分として入力画像 502 をパッチに区分することができるか、又はあらかじめ入力画像 502 をパッチに区分することができることが留意される。いくつかの例では、パッチ区分エンジン 510 a によって生成されたパッチはトークン（「パッチトークン」とも称される）として扱われ、トークンごとの特徴はパッチの構成要素の生ピクセル
50

R G B 値の連結として設定される。図示の例では、各 4×4 のパッチの特徴次元は、トランスフォーマ層 5 2 0 に提供される入力パッチのセットを詳述するラベル

【0 1 2 5】

【数 4】

$$\frac{H}{4} \times \frac{W}{4} \times 48$$

【0 1 2 6】

によって指示されるように、48（例えば、 $4 \times 4 \times 3 = 48$ ）である。

【0 1 2 7】

【0135】 パッチ区分モデル 5 1 0 a から、入力画像 5 0 2 から生成された非オーバーラップパッチのセットが第 1 のトランスフォーマ層 5 2 0 に提供される。図示のように、第 1 のトランスフォーマ層 5 2 0 は、各パッチの生値の特徴に対して、線形変換を適用することなどによって、それらを任意次元 C に射影するために適用される、線形埋め込み層 5 2 1 を含む。いくつかの例では、線形埋め込み層 5 2 1 は第 1 のトランスフォーマ層 5 2 0 の外部にあるか、又は他の状態でそれと分離していることができ、例えば、パッチ区分エンジン 5 1 0 a と組み合わせて提供され得る。加えて、線形埋め込み層 5 2 1 は第 1 のトランスフォーマ層 5 2 0 内にのみ現れ、次に続くトランスフォーマ層 5 3 0、5 4 0、5 5 0 は線形埋め込み層を欠くが、パッチ併合エンジン 5 3 2、5 4 2、5 5 2 をそれぞれ含むことが留意される。

【0 1 2 8】

【0136】 次に、シフトウィンドウトランスフォーマブロック 5 2 4 がパッチトークンにわたって適用される。前述のように、一連のシフトウィンドウトランスフォーマブロック 5 2 4 は、例えば、図 6 A において見られる 2 つなどの、2 つの連続したシフトウィンドウトランスフォーマブロックを含むように示されている。シフトウィンドウトランスフォーマブロック 5 2 4 は、変更された自己アテンション計算（図 6 及び図 7 A ～図 7 B に関して更に深く説明される）を実行し、特徴変換のために次に使用される 1 つ又は複数のアテンションベクトルを生成する。シフトウィンドウトランスフォーマブロック 5 2 4 の出力は、その後、第 2 のトランスフォーマ層 5 3 0 に入力として提供される。この際には、パッチトークン（例えば、

【0 1 2 9】

【数 5】

$$\frac{H}{4} \times \frac{W}{4}$$

【0 1 3 0】

）及び線形埋め込みの数が両方とも維持される。

【0 1 3 1】

【0137】 入力画像 5 0 2 から階層的表現又は特徴マップを生成するために、トランスフォーマ層 5 3 0、5 4 0、及び 5 5 0 はそれぞれのパッチ併合エンジン 5 3 2、5 4 2、5 5 2 を各々含む。パッチ併合エンジン 5 3 2、5 4 2、5 5 2 は、トランスフォーマ層が演算することになるパッチトークンの数を低減するために適用され得る。図示のように、パッチ併合エンジン 5 3 2 ～5 5 2 は各トランスフォーマ層内のシフトウィンドウトランスフォーマブロック 5 3 4 ～5 5 4 の前に提供される。いくつかの例では、トランスフォーマ層 5 3 0 ～5 5 0 のうちの 1 つ又は複数のは、そのパッチ併合エンジンの出力がその一連のシフトウィンドウトランスフォーマブロックの入力に直接結合されるように構成され得る。

【0 1 3 2】

【0138】 いくつかの例では、パッチ併合エンジン 5 3 2 は、前のトランスフォーマ層 5 2 0 からのパッチのセットを、隣り合ったパッチ又はブロックの重なり合わないグルー

10

20

30

40

50

ブに分割するように構成されている。パッチ併合エンジン532は各グループ内のパッチ特徴を連結することができ、連結されたグループごとに単一の新たなパッチを生成することができる。1つの例示的な例では、パッチ併合エンジン532、542、552によって実行されるパッチ併合演算は、空間データのブロックを深度に再配列することを含むことができる。具体的には、パッチ併合演算は、高さ及び幅次元からの値が深度次元へ移される、入力テンソルのコピーを出力することを含むことができる。block_size x block_size (ここで、block_sizeは、深度／チャンネル次元に再配列されるべき入力ブロック又はパッチの空間サイズを表す)のサイズの非オーバーラップパッチ又はブロックは、(サイズblock_size x block_sizeの各入力パッチは出力内の1つのピクセルに折り畳まれることになるため)ブロック又はパッチの各位置における深度に再配列される。出力テンソルの深度はblock_size * block_size * input_depthである。入力の各ブロック内の鉛直(Y)及び水平(X)座標は出力チャンネル指数の高次成分になる。入力テンソルの高さ及び幅はblock_sizeによって割り切れる。いくつかの例では、エンコーダサブネットワーク500aはパッチ併合演算の後に線形層を適用し、特徴のチャンネルの数を変更することができる。同様の演算がパッチ併合エンジン542及びパッチ併合エンジン552によって実行され得る。したがって、パッチ併合エンジン532～552は、各連続したトランスフォーマ層530～550に提供されるパッチトークンの数を、併合されたグループ当たりのパッチの数に等しい係数で低減することができる。上述されたように、図7Aは、エンコーダサブネットワーク500aの文脈では、より深いエンコーダトランスフォーマ層内において(例えば、下から上へ)画像パッチを併合する例示的なプロセスを示す。

10

20

30

40

【0133】

【0139】 図5Bを参照すると、デコーダサブネットワーク500bのトランスフォーマ層560、570、及び580はそれぞれのパッチ分離エンジン563、573、及び583(パッチ分割エンジンとも称される)を各々含む。パッチ分離エンジン563、573、583は、特徴マップをアップサンプリングする(それらのサイズを増大させる)ためのパッチ分離プロセスを実行することができる。パッチ分離プロセスは、パッチ併合エンジン532～552によって実行されるパッチ併合プロセスの逆のプロセスである。いくつかの例では、パッチ分離エンジン563は、前のブロック又は層(例えば、同じトランスフォーマ層内のシフトウィンドウトランスフォーマブロック)からの隣り合ったパッチの重なり合わないグループを、4つのチャンネルを有する1つのピクセルから、1つのチャンネルを有する4つのピクセルに分割するように構成されている(アップスケール係数が2である場合)。例えば、1つの例示的な例では、パッチ分離エンジン563、573、及び583によって実行されるパッチ分離演算は、形(*, C x r², H, W)のテンソル内の要素を形(*, C, H x r, W x r)のテンソルに再配列することを含むことができる。ここで、rはアップスケール係数である。いくつかの例では、デコーダサブネットワーク500bはパッチ分離演算の後に線形層を適用し、特徴のチャンネルの数を変更することができる。図7Aはまた、デコーダサブネットワーク500bの文脈では、より深いデコーダトランスフォーマ層内において(例えば、上から下へ)画像パッチを分離する例示的なプロセスを示すと見ることもできる。

【0134】

【0140】 第1の層702において、画像データは、パッチの16 x 16の格子に区分されているように表されており、各パッチは元の入力画像502からのいくつかの数の別個のピクセルを包含する。パッチ併合は、第1の層702と第2の層704との間、及び第2の層704と第3の層706との間で適用される。いくつかの例では、層702、704、706は、それぞれ、トランスフォーマ層520、530、540と同じであることができ、上述されたパッチ併合プロセスはパッチ併合エンジン532及び542によって実行され得る。

【0135】

50

【0141】 第1の層702と第2の層704との間では、パッチ併合は、（例えば、第1の層702の）隣り合ったパッチの 2×2 のグループからの特徴を（例えば、第2の層704の）単一のパッチに連結し、連結された特徴に対して線形層を適用することによって実行される。それゆえ、第2の層704内の単一の併合パッチは、第1の層702内の4つの別個のパッチからの連結された特徴を包含する。パッチの総数は4分の1に低減され、解像度は2分の1にダウンサンプリングされ、パッチ併合プロセスの出力次元は $2C$ である。

【0136】

【0142】 第2の層704と第3の層706との間では、同一のパッチ併合プロセスが適用され得る（例えば、 2×2 の隣り合ったパッチのグループが単一のパッチに併合される）。第3の層706内の取得された併合パッチは、第2の層704内の4つの別個のパッチからの連結された特徴を包含するか、又は代替的に、第1の層702内の16個の別個のパッチからの連結された特徴を包含すると捉えることができる。第1の層702と第2の層704との間の場合と同様に、第2の層704と第3の層706との間のパッチ併合プロセスは、この場合も先と同様に、パッチの総数を4分の1に低減し、解像度を2分の1にダウンサンプリングする。したがって、このパッチ併合プロセスの出力次元は $4C$ である。

【0137】

【0143】 したがって、繰り返されるパッチ併合論理が利用されるときには、各パッチ併合エンジンは異なる入力及び出力解像度を有するものの（例えば、パッチ併合エンジン532は

【0138】

【数6】

$$\frac{H}{4} \times \frac{W}{4} \times C$$

【0139】

の入力パッチトークン解像度及び

【0140】

【数7】

$$\frac{H}{8} \times \frac{W}{8} \times 2C$$

【0141】

の出力解像度を有し、パッチ併合エンジン542は

【0142】

【数8】

$$\frac{H}{8} \times \frac{W}{8} \times 2C$$

【0143】

の入力解像度及び

【0144】

【数9】

$$\frac{H}{16} \times \frac{W}{16} \times 4C$$

【0145】

の出力解像度を有し、パッチ併合エンジン552は

【0146】

【数 1 0】

$$\frac{H}{16} \times \frac{W}{16} \times 4C$$

【0 1 4 7】

の入力解像度及び

【0 1 4 8】

【数 1 1】

$$\frac{H}{32} \times \frac{W}{32} \times 8C$$

10

【0 1 4 9】

の出力解像度を有する)、パッチ併合エンジン 5 3 2 ~ 5 5 2 は互いに同一であることができる。

【0 1 5 0】

【0144】 図 6 B は、本明細書において説明される 1 つ又は複数のトランスフォーマベースのニューラルネットワークアーキテクチャが利用され得るビデオコーディングアーキテクチャの一例を示す図である。トランスフォーマベースのニューラルネットワークを使用することは、ビデオコーディングアーキテクチャが低レイテンシをもって動作すること(例えば、低遅延ビデオコーディングアーキテクチャ)を可能にする。図 6 B のビデオコーディングシステムのアーキテクチャは、少なくとも部分的に、システムが不確実性をモデル化することをより良好に可能にするためのスケールパラメータを追加することによって、オプティカルフローを拡張するか、又は一般化する、スケールスペースフロー (SSF) に基づく。図 6 B の文脈では、スケールパラメータはスケールスペースワーピング構成要素に関連付けられている。図示のように、SSF ビデオコーディングアーキテクチャは、動き補償を実行するために、まとめて「フロー AE」又はフローオートエンコーダとラベル付けされた、SSF エンコーダ及び SSF デコーダを使用する。SSF エンコーダ及び SSF デコーダはどちらも 1 つ又は複数の畳み込み変換に基づき得る。同様に、残差エンコーダ及び残差デコーダもどちらも 1 つ又は複数の畳み込み変換に基づき得る。いくつかの場合には、SSF エンコーダ及びデコーダ、並びに残差エンコーダ及びデコーダは CNN として各々実施され得る。いくつかの例では、エンコーダ及びデコーダ構成要素のうちの全ての 4 つは 4 層の CNN として提供され得る。

20

30

【0 1 5 1】

【0145】 図 6 B に示されるように、本明細書において説明される 1 つ又は複数のシフトウィンドウトランスフォーマニューラルネットワークは、いくつかの例では、SSF アーキテクチャ内の CNN 及び/又は畳み込み変換を置換するために使用することができる。例えば、1 つ又は複数のシフトウィンドウトランスフォーマニューラルネットワークを、SSF エンコーダ、SSF デコーダ、残差エンコーダ、及び/又は残差デコーダを置換するために提供することができる。いくつかの例では、以下において更に深く説明されるように、4 つのシフトウィンドウトランスフォーマニューラルネットワークが SSF ビデオコーディングアーキテクチャ内の 4 つの畳み込み構成要素を置換することができる。

40

【0 1 5 2】

【0146】 図 8 は、画像及び/又はビデオデータを処理するためのプロセス 8 0 0 の一例を示すフローチャートである。ブロック 8 0 2 において、プロセス 8 0 0 は、符号化画像データのフレームの潜在表現を取得することを含む。いくつかの例では、符号化画像データのフレームは符号化静止画像を含む。いくつかの例では、符号化画像データのフレームは符号化ビデオフレームを含む。

【0 1 5 3】

【0147】 ブロック 8 0 4 において、プロセス 8 0 0 は、符号化画像データのフレームの潜在表現を入力として使用して、デコーダサブネットワークの複数のデコーダトランス

50

フォーマ層によって、復号画像データのフレームを生成することを含む。いくつかの場合には、複数のデコードトランスフォーマ層は一連の連続したデコードトランスフォーマ層（例えば、図 5 B に示される一連のデコードトランスフォーマ層 5 6 0、5 7 0、5 8 0、及び 5 9 0）を含む。複数のデコードトランスフォーマ層のうちの少なくとも 1 つのデコードトランスフォーマ層は、特徴の 1 つ又は複数のパッチを生成し、1 つ又は複数のパッチにわたって適用された 1 つ又は複数のウィンドウ区分及びシフトウィンドウ区分内で局所的に自己アテンションを決定するための 1 つ又は複数のトランスフォーマブロックを含む。複数のデコードトランスフォーマ層のうちの少なくとも 1 つのデコードトランスフォーマ層は、1 つ又は複数のパッチの各パッチのそれぞれのサイズを減少させるためのパッチ分離エンジンを更に含む。例えば、図 5 B を例示的な例として参照すると、デコードトランスフォーマ層 5 6 0 はシフトウィンドウトランスフォーマブロック 5 6 4 及びパッチ分離エンジン 5 6 3 を含む。いくつかの態様では、少なくとも 1 つのデコードトランスフォーマ層内に含まれる 1 つ又は複数のトランスフォーマブロックの少なくとも部分は同じアーキテクチャを有する。いくつかの態様では、少なくとも 1 つのデコードトランスフォーマ層内に含まれる 1 つ又は複数のトランスフォーマブロックの各々（例えば、全て）は同じアーキテクチャを有する。

10

20

30

40

50

【0154】

【0148】 いくつかの例では、復号画像データのフレームを生成するために、プロセス 8 0 0 は、複数のデコードトランスフォーマ層のうちの第 1 のデコードトランスフォーマ層の第 1 のトランスフォーマブロックによって、1 つ又は複数のパッチにわたって適用された 1 つ又は複数の第 1 のウィンドウ区分内で局所的に自己アテンションを決定することを含むことができる。プロセス 8 0 0 は、第 1 のデコードトランスフォーマ層の第 2 のトランスフォーマブロックによって、1 つ又は複数のパッチにわたって適用された 1 つ又は複数の第 2 のウィンドウ区分内で局所的に自己アテンションを決定することを更に含むことができる。例えば、1 つ又は複数の第 2 のウィンドウ区分は、（例えば、図 7 A 及び図 7 B に示されるように）1 つ又は複数の第 1 のウィンドウ区分のうちの隣接したものの間の 1 つ又は複数の境界に重なるようにシフトされ得る。プロセス 8 0 0 は、パッチ分離エンジンによって、1 つ又は複数のパッチの各パッチを複数の分離パッチにセグメント化することを含むことができる。複数の分離パッチは重なり合わない。いくつかの例では、複数の分離パッチの各分離パッチは均一なパッチサイズを有する。いくつかの場合には、パッチ分離エンジンは 2 又は他の値のパッチサイズ縮小係数を適用する。いくつかの態様では、1 つ又は複数のパッチの各パッチを複数の分離パッチにセグメント化するために、プロセス 8 0 0 は、複数の分離パッチの特徴次元を減少させることを含むことができる。

【0155】

【0149】 いくつかの場合には、プロセス 8 0 0 は、複数の分離パッチを複数のデコードトランスフォーマ層のうちの第 2 のデコードトランスフォーマ層の第 1 のトランスフォーマブロックに提供することを含むことができる。例えば、デコードトランスフォーマ層 5 6 0 のパッチ分離エンジン 5 6 3 は分離パッチをデコードトランスフォーマ層 5 7 0 のトランスフォーマブロック 5 7 4 に提供することができる。いくつかの例では、プロセス 8 0 0 は、第 2 のデコードトランスフォーマ層のパッチ分離エンジンによって、複数の分離パッチをセグメント化することを含むことができる。プロセス 8 0 0 は、パッチ分離エンジンの出力を複数のデコードトランスフォーマ層のうちの第 3 のデコードトランスフォーマ層に提供することを更に含むことができる。例えば、デコードトランスフォーマ層 5 7 0 のパッチ分離エンジン 5 7 3 は分離パッチをデコードトランスフォーマ層 5 8 0 （例えば、デコードトランスフォーマ層 5 8 0 のトランスフォーマブロック 5 8 4）に提供することができる。別の例では、デコードトランスフォーマ層 5 9 0 の線形埋め込みエンジン 5 9 1 は、トランスフォーマブロック 5 9 4 から出力された特徴に線形変換を適用することができ、線形変換された特徴をパッチ区分エンジン 5 1 0 b に出力することができる。

【0156】

【0150】 いくつかの態様では、プロセス800は、複数のデコーダトランスフォーマ層によって、符号化画像データのフレームの潜在表現を入力として受け取ることを含むことができる。プロセス800は、非線形変換を適用し、復号画像データのフレームを生成することを含むことができる。いくつかの場合には、非線形変換は合成変換であり、復号画像データのフレームは、符号化画像データのフレームに関連付けられた入力画像の復元である。

【0157】

【0151】 いくつかの例では、プロセス800は、レート歪みに少なくとも部分的に基づく損失関数を使用して複数のデコーダトランスフォーマ層のうちの1つ又は複数のデコーダトランスフォーマ層を訓練することを含むことができる。いくつかの場合には、損失関数はレート歪みのためのラグランジュ乗数を含む。例えば、本明細書において説明されるように、レート歪み損失関数は $L = D + \beta * R$ を有することができる。ここで、 D は歪みを表し、 R はレートを表し、異なる β 値は、異なるビットレート及び／又はピーク信号対雑音比 (PSNR) のために訓練されたモデルを表す。

10

【0158】

【0152】 いくつかの例では、プロセス800は、少なくとも第1の訓練データセット及び第2の訓練データセットで（例えば、教師なし訓練又は学習を使用して）複数のデコーダトランスフォーマ層を訓練することを含むことができる。このような例では、第2の訓練データセットのデータは、第1の訓練データセットのデータと比べて逆の時間的順序を有する。いくつかの場合には、トランスフォーマブロック内の全ての層ノルム層を除去することができる。いくつかの例では、トランスフォーマブロック内で使用されるウィンドウサイズを8から4に縮小することができる。第1の訓練セットのデータと比べて第2の訓練セットのために逆の時間的順序を使用すること、トランスフォーマブロック内の層ノルム層を除去すること、及び／又はウィンドウサイズを8から4に縮小することは、安定した訓練をもたらすことができる。

20

【0159】

【0153】 図9は、画像及び／又はビデオデータを処理するためのプロセス900の一例を示すフローチャートである。ブロック902において、プロセス900は、フレームを複数のパッチにセグメント化することを含む。いくつかの例では、フレームは静止画像を含む。いくつかの例では、フレームはビデオフレーム（例えば、符号化されていないビデオフレーム）を含む。いくつかの場合には、複数のパッチの各パッチは均一なサイズのものであり、セグメント化されたフレームの1つ又は複数のピクセルを含む。例えば、複数のパッチは、静止画像フレーム又はビデオフレームを含む入力からセグメント化され得る。

30

【0160】

【0154】 ブロック904において、プロセス900は、複数のパッチを入力として使用して、エンコーダサブネットワークの複数のエンコーダトランスフォーマ層によって、符号化画像データのフレームを生成することを含む。いくつかの例では、符号化画像データのフレームは画像データの潜在表現である。いくつかの場合には、潜在表現は、エンコーダサブネットワークの複数のエンコーダトランスフォーマ層によって生成された階層的特徴マップである。いくつかの例では、プロセス900は、複数のパッチを入力として使用して、エンコーダサブネットワークの複数のエンコーダトランスフォーマ層によって、セグメント化されたフレームのための階層的特徴マップを生成することを含むことができる。プロセス900は、階層的特徴マップから符号化画像データのフレームを生成することを含むことができる。

40

【0161】

【0155】 1つの例示的な例では、符号化画像データのフレームを生成するために、プロセス900は、複数のエンコーダトランスフォーマ層のうちの第1のエンコーダトランスフォーマ層の第1のトランスフォーマブロックによって、1つ又は複数のウィンドウ区分内で局所的に自己アテンションを決定することを含むことができる。プロセス900は

50

、第1のエンコーダトランスフォーマ層の第2のトランスフォーマブロックによって、1つ又は複数のシフトウィンドウ区分内で局所的に自己アテンションを決定することを含むことができる。いくつかの態様では、第1のトランスフォーマブロック及び第2のトランスフォーマブロックは同じアーキテクチャを有する。1つ又は複数のシフトウィンドウ区分は1つ又は複数のウィンドウ区分に重なる。プロセス900は、第1のトランスフォーマブロック及び第2のトランスフォーマブロックのうちの1つ又は複数によって、非線形変換をセグメント化されたフレームに適用するための特徴の1つ又は複数のパッチを決定することを更に含むことができる。プロセス900は、パッチ併合エンジンによって、第1のエンコーダトランスフォーマ層と第2のエンコーダトランスフォーマ層との間でパッチサイズを増大させることを含むことができる。本明細書において説明されるように、パッチ併合エンジンは、第1のエンコーダトランスフォーマ層からの複数の隣接したパッチを、第2のエンコーダトランスフォーマ層に提供される併合パッチに組み合わせるように構成されている。いくつかの場合には、第1のエンコーダトランスフォーマ層の第2のトランスフォーマブロックの出力が第2のエンコーダトランスフォーマ層の入力に結合される。

10

【0162】

【0156】 いくつかの場合には、プロセス900は、パッチ併合エンジンを使用して、隣接したパッチの1つ又は複数の部分セットから取得された特徴を連結することによってパッチサイズを増大させることを含むことができる。パッチ併合エンジンは、隣接したパッチの各部分セットを併合パッチ出力に併合することができる。

20

【0163】

【0157】 いくつかの例では、プロセス900は、第1のエンコーダトランスフォーマ層の前に複数のパッチをエンコーダサブネットワークの線形埋め込み層（例えば、図5Aの線形埋め込み層521）に提供することを含むことができる。線形埋め込み層は線形変換を複数のパッチに適用することができる。

【0164】

【0158】 いくつかの態様では、プロセス900は、レート歪み損失に基づく損失関数を使用して複数のエンコーダトランスフォーマ層のうちの1つ又は複数のエンコーダトランスフォーマ層を訓練することを含むことができる。いくつかの場合には、損失関数はレート歪みのためのラグランジュ乗数を含む。上述されたように、レート歪み損失は $L = D + \beta * R$ と表され得る。

30

【0165】

【0159】 いくつかの態様では、プロセス900は、因数分解プライアで符号化画像データをエントロピコーディングすることを含むことができる。

【0166】

【0160】 いくつかの例では、本明細書で説明するプロセス（例えば、プロセス800、プロセス900、及び／又は本明細書で説明する任意の他のプロセス）は、コンピューティングデバイス、装置、又はシステムによって実行されてもよい。一例では、プロセス800及び／又はプロセス900は、図10のコンピューティングデバイスアーキテクチャ1000を有するコンピューティングデバイス又はシステムによって実行され得る。コンピューティングデバイス、装置、又はシステムは、モバイルデバイス（例えば、モバイルフォン）、デスクトップコンピューティングデバイス、タブレットコンピューティングデバイス、ウェアラブルデバイス（例えば、VRヘッドセット、ARヘッドセット、ARグラス、ネットワーク接続ウォッチ若しくはスマートウォッチ、又は他のウェアラブルデバイス）、サーバコンピュータ、自律運転車若しくは自律運転車のコンピューティングデバイス、ロボティクスデバイス、ラップトップコンピュータ、スマートテレビジョン、カメラ、並びに／又は、プロセス800及び／若しくはプロセス900並びに／又は本明細書で説明する任意の他のプロセスを含む、本明細書で説明するプロセスを実行するためのリソース能力を持つ任意の他のコンピューティングデバイスなどの、任意の適切なデバイスを含み得る。場合によっては、コンピューティングデバイス又は装置は、1つ又は複数

40

50

の入力デバイス、1つ又は複数の出力デバイス、1つ又は複数のプロセッサ、1つ又は複数のマイクロプロセッサ、1つ又は複数のマイクロコンピュータ、1つ又は複数のカメラ、1つ又は複数のセンサ、及び／又は本明細書で説明するプロセスのステップを実行するように構成される他の構成要素（複数可）などの、様々な構成要素を含み得る。いくつかの例では、コンピューティングデバイスは、ディスプレイ、データを通信及び／又は受信するように構成されたネットワークインターフェース、それらの任意の組み合わせ、及び／又は他の構成要素（複数可）を含んでよい。ネットワークインターフェースは、インターネットプロトコル（IP）ベースのデータ若しくは他のタイプのデータを通信及び／又は受信するように構成され得る。

【0167】

10

【0161】 コンピューティングデバイスの構成要素は、回路に実装され得る。例えば、構成要素は、1つ又は複数のプログラマブル電子回路（例えば、マイクロプロセッサ、グラフィックス処理ユニット（GPU）、デジタル信号プロセッサ（DSP）、中央処理ユニット（CPU）、及び／又は他の好適な電子回路）を含み得る、電子回路若しくは他の電子ハードウェアを含むことができる、及び／若しくはそれらを使用して実装されることが可能であり、並びに／又は本明細書で説明する様々な動作を実行するために、コンピュータソフトウェア、ファームウェア、若しくはそれらの任意の組み合わせを含むことができる、及び／若しくはそれらを使用して実装されることが可能である。

【0168】

【0162】 プロセス800及びプロセス900は、論理フロー図として図示され、その動作は、ハードウェア、コンピュータ命令、又はそれらの組み合わせで実装され得る動作のシーケンスを表す。コンピュータ命令のコンテキストでは、動作は、1つ又は複数のプロセッサによって実行されたとき、記載された動作を実行する、1つ又は複数のコンピュータ可読記憶媒体上に記憶されたコンピュータ実行可能命令を表す。概して、コンピュータ実行可能命令は、特定の機能を実行するか又は特定のデータタイプを実装するルーチン、プログラム、オブジェクト、構成要素、データ構造などを含む。動作が説明される順序は、限定として解釈されることを意図せず、任意の数の説明される動作は、プロセスを実装するために任意の順序で、及び／又は並列に組み合わせることができる。

20

【0169】

【0163】 追加として、プロセス800、プロセス900、及び／又は本明細書で説明する任意の他のプロセスは、実行可能命令を使用して構成された1つ若しくは複数のコンピュータシステムの制御下で実行されてよく、1つ若しくは複数のプロセッサ上で、ハードウェアによって、又はそれらの組み合わせで、集合的に実行するコード（例えば、実行可能命令、1つ若しくは複数のコンピュータプログラム、又は1つ若しくは複数のアプリケーション）として実装されてよい。上述のように、コードは、例えば、1つ又は複数のプロセッサによって実行可能な複数の命令を備えるコンピュータプログラムの形態で、コンピュータ可読記憶媒体又は機械可読記憶媒体上に記憶されてもよい。コンピュータ可読記憶媒体又は機械可読記憶媒体は、非一時的であってもよい。

30

【0170】

【0164】 図10は、本明細書で説明する様々な技法を実施できる例示的なコンピューティングデバイスの例示的なコンピューティングデバイスアーキテクチャ1000を示す。いくつかの例では、コンピューティングデバイスは、モバイルデバイス、ウェアラブルデバイス、エクステンデッドリアリティデバイス（例えば、仮想現実（VR）デバイス、拡張現実（AR）デバイス、もしくは複合現実（MR）デバイス）、パーソナルコンピュータ、ラップトップコンピュータ、ビデオサーバ、車両（もしくは車両のコンピューティングデバイス）、又は他のデバイスを含むことができる。例えば、コンピューティングデバイスアーキテクチャ1000は、図6Aのシステムを実装することができる。コンピューティングデバイスアーキテクチャ1000の構成要素は、バスなどの接続部1005を使用して、互いに電気通信するように示されている。例示的なコンピューティングデバイスアーキテクチャ1000は、処理ユニット（CPU又はプロセッサ）1010と、リー

40

50

ドオンリーメモリ（ROM）１０２０及びランダムアクセスメモリ（RAM）１０２５などのコンピューティングデバイスメモリ１０１５を含む様々なコンピューティングデバイス構成要素をプロセッサ１０１０に結合するコンピューティングデバイス接続部１００５と、を含む。

【０１７１】

【０１６５】 コンピューティングデバイスアーキテクチャ１０００は、プロセッサ１０１０に直接接続されるか、そのごく近傍にあるか、又はその一部として統合される、高速メモリのキャッシュを含むことができる。コンピューティングデバイスアーキテクチャ１０００は、プロセッサ１０１０による高速アクセスのために、メモリ１０１５及び／又は記憶デバイス１０３０からキャッシュ１０１２にデータをコピーすることができる。このように、キャッシュは、データを待機している間のプロセッサ１０１０の遅延を回避する性能向上をもたらすことができる。これら及び他のエンジンは、様々なアクションを実行するようにプロセッサ１０１０を制御することができ、又はそれを制御するように構成され得る。他のコンピュータデバイスメモリ１０１５も同様に使用のために利用可能であり得る。メモリ１０１５は、異なる性能特性を有する複数の異なる種類のメモリを含むことができる。プロセッサ１０１０は、任意の汎用プロセッサ及びハードウェア、又はプロセッサ１０１０を制御するように構成された、記憶デバイス１０３０の中に記憶されるサービス１ １０３２、サービス２ １０３４、及びサービス３ １０３６などのソフトウェアサービス、並びにソフトウェア命令がプロセッサ設計の中に組み込まれる専用プロセッサを含むことができる。プロセッサ１０１０は、複数のコア又はプロセッサ、バス、メモリコントローラ、キャッシュなどを含む、自己完結型のシステムであってもよい。マルチコアプロセッサは、対称でも非対称でもよい。

【０１７２】

【０１６６】 コンピューティングデバイスアーキテクチャ１０００とのユーザ対話を可能にするために、入力デバイス１０４５は、音声用のマイクロフォン、ジェスチャ又は図形入力用のタッチ感応スクリーン、キーボード、マウス、動き入力、音声などの任意の数の入力機構を表すことができる。出力デバイス１０３５はまた、ディスプレイ、プロジェクタ、テレビ、スピーカデバイス又は複数などの、当業者に知られている多数の出力機構のうちの１つであり得る。いくつかの事例では、マルチモーダルコンピューティングデバイスは、ユーザがコンピューティングデバイスアーキテクチャ１０００と通信するために複数のタイプの入力を提供することを可能にできる。通信インターフェース１０４０は、概して、ユーザ入力及びコンピューティングデバイス出力を統御及び管理することができる。いかなる特定のハードウェア構成上で動作することに対しても制約はなく、したがって、ここでの基本的機能は、改善されたハードウェア又はファームウェア構成が開発されるにつれて、それらと容易に置き換えられ得る。

【０１７３】

【０１６７】 記憶デバイス１０３０は、不揮発性メモリであり、ハードディスク、又は磁気カセット、フラッシュメモリカード、ソリッドステートメモリデバイス、デジタル多用途ディスク、カートリッジ、ランダムアクセスメモリ（RAM）１０２５、リードオンリーメモリ（ROM）１０２０、及びそれらのハイブリッドなどの、コンピュータによってアクセス可能なデータを記憶できる他のタイプのコンピュータ可読媒体であり得る。記憶デバイス１０３０は、プロセッサ１０１０を制御するためのサービス１０３２、１０３４、１０３６を含むことができる。他のハードウェア又はソフトウェアモジュール又はエンジンも企図される。記憶デバイス１０３０は、コンピューティングデバイス接続部１００５に接続され得る。一態様では、特定の機能を実行するハードウェアモジュールは、機能を実行するために、プロセッサ１０１０、接続部１００５、出力デバイス１０３５などの、必須のハードウェア構成要素と接続されたコンピュータ可読媒体の中に記憶されるソフトウェア構成要素を含むことができる。

【０１７４】

【０１６８】 本開示の態様は、１つ又は複数のアクティブ深度検知システムを含むか、又

はそれに結合された（セキュリティシステム、スマートフォン、タブレット、ラップトップコンピュータ、車両、ドローン、又は他のデバイスなどの）任意の適切な電子デバイスに適用可能である。1つの光プロジェクタを有するか、又はそれに結合されたデバイスに関して以下に記載されるが、本開示の態様は、任意の数の光プロジェクタを有するデバイスに適用可能であり、したがって、特定のデバイスに限定されない。

【0175】

【0169】 「デバイス」という用語は、（1つのスマートフォン、1つのコントローラ、1つの処理システムなどの）1つ又は特定の数の物理オブジェクトに限定されない。本明細書で使用されるデバイスは、本開示の少なくともいくつかの部分を実装することができる1つ又は複数の部分を有する任意の電子デバイスであり得る。以下の説明及び例は、本開示の様々な態様を記載するために「デバイス」という用語を使用するが、「デバイス」という用語は、特定の構成、タイプ、又はオブジェクトの数に限定されない。追加として、「システム」という用語は複数の構成要素又は特定の実施形態に限定されない。例えば、システムは、1つ若しくは複数のプリント回路板又は他の基板上に実装されてよく、移動可能又は静的な構成要素を有してよい。以下の説明及び例は、本開示の様々な態様を記載するために「システム」という用語を使用するが、「システム」という用語は、特定の構成、タイプ、又はオブジェクトの数に限定されない。

【0176】

【0170】 本明細書で提供する実施形態及び例の完全な理解を与えるために、上記の説明において具体的な詳細が提供されている。しかしながら、実施形態がこれらの具体的な詳細なしに実践され得ることが当業者によって理解されよう。説明を分かりやすくするために、いくつかの事例では、本技術は、デバイス、デバイス構成要素、ソフトウェアの中で具現される方法におけるステップ若しくはルーチン、又はハードウェアとソフトウェアとの組み合わせを備える機能ブロックを含む、個々の機能ブロックを含むものとして提示されることがある。図中に示され、及び／又は本明細書で説明される構成要素以外の、追加の構成要素が使用されてもよい。例えば、不必要な詳細で実施形態を不明瞭にしないように、回路、システム、ネットワーク、プロセス、及び他の構成要素がブロック図の形態で構成要素として示されてもよい。他の事例では、実施形態を不明瞭にすることを避けるために、よく知られている回路、プロセス、アルゴリズム、構造、及び技法は、不必要な詳細なしに示されてもよい。

【0177】

【0171】 個々の実施形態は、フローチャート、フロー図、データフロー図、構造図、又はブロック図として示されるプロセス又は方法として上記で説明されてもよい。フローチャートは、動作を逐次プロセスとして説明することがあるが、動作の多くは並列に又は同時に実行することができる。加えて、工程の順序は並べ替えられてもよい。プロセスは、その動作が完了するときに終了するが、図に含まれていない追加のステップを有することができる。プロセスは、メソッド、関数、プロシージャ、サブルーチン、サブプログラムなどに対応してもよい。プロセスが関数に対応するとき、その終了は、その関数が呼出し関数又はメイン関数に戻ることに対応することができる。

【0178】

【0172】 上記で説明した例によるプロセス及び方法は、コンピュータ可読媒体に記憶されているか、又はそうでなければコンピュータ可読媒体から入手可能なコンピュータ実行可能命令を使用して実装され得る。そのような命令は、例えば、汎用コンピュータ、専用コンピュータ、又は処理デバイスが、特定の機能又は機能のグループを実施することを引き起こすか、又はそうでなければそれらを実施するように構成する、命令とデータとを含むことができる。使用されるコンピュータリソースの部分は、ネットワークを介してアクセス可能であり得る。コンピュータ実行可能命令は、例えば、バイナリであってよく、アセンブリ言語、ファームウェア、ソースコードなどの中間フォーマット命令であってもよい。

【0179】

【0173】 「コンピュータ可読媒体」という用語は、限定はしないが、ポータブル又は非ポータブルの記憶デバイス、光記憶デバイス、並びに命令（複数可）及び／又はデータを記憶、格納、又は搬送できる様々な他の媒体を含む。コンピュータ可読媒体は、データが記憶され、かつワイヤレスで若しくは有線接続を介して伝搬する搬送波及び／又は一時的な電子信号を含まない、非一時的媒体を含んでもよい。非一時的媒体の例は、限定はしないが、特に、磁気ディスク又は磁気テープ、フラッシュメモリなどの光記憶媒体、メモリ又はメモリデバイス、磁気ディスク又は光ディスク、フラッシュメモリ、不揮発性メモリを用いて設けられるUSBデバイス、ネットワーク化された記憶デバイス、コンパクトディスク（CD）又はデジタル多用途ディスク（DVD）、それらの任意の好適な組み合わせを含んでよい。コンピュータ可読媒体は、プロシージャ、関数、サブプログラム、プログラム、ルーチン、サブルーチン、モジュール、エンジン、ソフトウェアパッケージ、クラス、又は命令、データ構造、若しくはプログラムステートメントの任意の組み合わせを表し得る、コンピュータ可読媒体上に記憶されたコード及び／又は機械実行可能命令を有してもよい。コードセグメントは、情報、データ、引数、パラメータ、又はメモリコンテンツを渡すこと及び／又は受けることによって、別のコードセグメント又はハードウェア回路に結合されてもよい。情報、引数、パラメータ、データなどは、メモリ共有、メッセージパッシング、トークンパッシング、ネットワーク送信などを含む、任意の好適な手段を介して渡され、転送され、又は送信されてもよい。

10

【0180】

【0174】 いくつかの実施形態では、コンピュータ可読記憶デバイス、媒体、及びメモリは、ビットストリームなどを含むケーブル又はワイヤレス信号を含み得る。しかしながら、言及する場合、非一時的コンピュータ可読記憶媒体は、エネルギー、キャリア信号、電磁波、及び信号自体などの媒体を明確に除外する。

20

【0181】

【0175】 これらの開示に従ってプロセス及び方法を実施するデバイスは、ハードウェア、ソフトウェア、ファームウェア、ミドルウェア、マイクロコード、ハードウェア記述言語、又はそれらの任意の組み合わせを含むことができ、様々なフォームファクタのうちのいずれかをとることができる。ソフトウェア、ファームウェア、ミドルウェア、又はマイクロコードにおいて実装されるとき、必要なタスクを実行するためのプログラムコード又はコードセグメント（例えば、コンピュータプログラム製品）は、コンピュータ可読媒体又は機械可読媒体に記憶されてもよい。プロセッサ（複数可）は、必要なタスクを実行してもよい。フォームファクタの典型的な例は、ラップトップ、スマートフォン、携帯電話、タブレットデバイス又は他の小スペース型パーソナルコンピュータ、携帯情報端末、ラックマウントデバイス、スタンドアロンデバイスなどを含む。本明細書で説明した機能はまた、周辺装置又はアドインカードで具現化され得る。そのような機能はまた、更なる例として、異なるチップのうちの回路基板上、又は単一のデバイスにおいて実行する異なるプロセス上で実施され得る。

30

【0182】

【0176】 命令、そのような命令を伝えるための媒体、命令を実行するためのコンピューティングリソース、及びそのようなコンピューティングリソースをサポートするための他の構造は、本開示で説明した機能を提供するための例示的な手段である。

40

【0183】

【0177】 上記の説明では、本出願の態様はそれらの特定の実施形態を参照しながら説明されるが、本出願がそれらに限定されないことを当業者は認識されよう。したがって、本出願の例示的な実施形態が本明細書で詳細に説明されているが、本発明の概念が別のやり方で様々に具現及び採用され得ること、並びに従来技術によって限定される場合を除き、添付の特許請求の範囲がそのような変形を含むものと解釈されることが意図されることを理解されたい。上記で説明した本出願の様々な特徴及び態様は、個別に又は共同で使用され得る。更に、実施形態は、本明細書のより広い趣旨及び範囲から逸脱することなく、本明細書で説明されるもの以外の任意の数の環境及び適用例において利用することができ

50

る。したがって、本明細書及び図面は、限定的ではなく例示的と見なされるべきである。例示の目的のために、方法は特定の順序で説明された。代替実施形態では、方法は、説明された順序とは異なる順序で実行されてもよいことを理解されたい。

【0184】

【0178】 本明細書において使用される、よりも小さい（「<」）及びよりも大きい（「>」）という記号又は用語は、本説明の範囲から逸脱することなく、それぞれ、以下（「≤」）及び以上（「≥」）という記号で置き換えられ得ることを、当業者は理解するであろう。

【0185】

【0179】 構成要素が特定の動作を実行する「ように構成されている」ものとして説明される場合、そのような構成は、例えば、動作を実行するように電子回路若しくは他のハードウェアを設計することによって、動作を実行するようにプログラマブル電子回路（例えば、マイクロプロセッサ、又は他の適切な電子回路）をプログラムすることによって、又はそれらの任意の組み合わせで達成され得る。

【0186】

【0180】 「に結合された」という句は、直接的若しくは間接的のいずれかで別の構成要素に物理的に接続されている任意の構成要素、及び／又は直接的若しくは間接的のいずれかで別の構成要素と通信している（例えば、有線接続若しくはワイヤレス接続及び／又は他の好適な通信インターフェースを介して他の構成要素に接続されている）任意の構成要素を指す。

【0187】

【0181】 集合「のうちの少なくとも1つ」及び／又は集合のうちの「1つ以上」と記載する請求項の文言又は他の文言は、集合の1つのメンバー又は集合の（任意の組み合わせでの）複数のメンバーが請求項を満たすことを示す。例えば、「A及びBのうちの少なくとも1つ」又は「A又はBのうちの少なくとも1つ」を記載する請求項の文言は、A、B、又はA及びBを意味する。別の例では、「A、B、及びCのうちの少なくとも1つ」又は「A、B、又はCのうちの少なくとも1つ」を記載する請求項の文言は、A、B、C、又はA及びB、又はA及びC、又はB及びC、又はA及びB及びCを意味する。集合「のうちの少なくとも1つ」及び／又は集合のうちの「1つ以上」という文言は、集合の中で列挙される項目にその集合を限定しない。例えば、「A及びBのうちの少なくとも1つ」又は「A又はBのうちの少なくとも1つ」と記載する請求項の文言は、A、B、又はA及びBを意味することができ、追加として、A及びBの集合の中で列挙されない項目を含むことができる。

【0188】

【0182】 本明細書で開示する実施形態に関して説明される様々な例示的な論理ブロック、モジュール、エンジン、回路、及びアルゴリズムステップは、電子ハードウェア、コンピュータソフトウェア、ファームウェア、又はそれらの組み合わせとして実装され得る。ハードウェア及びソフトウェアのこの互換性を明確に例示するために、様々な例示的なコンポーネント、ブロック、モジュール、エンジン、回路、及びステップが、概してそれらの機能性の観点から上記に説明されている。そのような機能がハードウェアとして実装されるのか又はソフトウェアとして実装されるのかは、特定の用途及び全体的なシステムに課される設計制約に依存する。当業者は、説明した機能を特定の適用例ごとに様々な方法で実装し得るが、そのような実装の決定は、本出願の範囲から逸脱する原因として解釈されるべきではない。

【0189】

【0183】 本明細書で説明される技法はまた、電子ハードウェア、コンピュータソフトウェア、ファームウェア、又はそれらの任意の組み合わせにおいて実装され得る。そのような技法は、汎用コンピュータ、ワイヤレス通信デバイスハンドセット、又はワイヤレス通信デバイスハンドセット及び他のデバイスにおける適用例を含む複数の用途を有する集積回路デバイスなどの、様々なデバイスのうちのいずれかにおいて実装されてもよい。モ

ジュール又は構成要素として説明した任意の特徴は、集積ロジックデバイスの中で一緒に、又は個別であるが相互動作可能なロジックデバイスとして別々に実装され得る。ソフトウェアで実装される場合、技法は、実行されると、上で説明された方法のうちの1つ以上を実施する命令を含むプログラムコードを備える、コンピュータ可読データ記憶媒体によって少なくとも部分的に実現されてよい。コンピュータ可読データ記憶媒体は、コンピュータプログラム製品の一部を形成してもよく、これはパッケージ材料を含んでもよい。コンピュータ可読媒体は、シンクロナスダイナミックランダムアクセスメモリ (synchronous dynamic random-access memory、SDRAM) などのランダムアクセスメモリ (RAM)、リードオンリーメモリ (ROM)、不揮発性ランダムアクセスメモリ (non-volatile RAM、NVRAM)、電氣的消去可能プログラマブルリードオンリーメモリ (electrically erasable programmable read-only memory、EEPROM)、フラッシュメモリ、磁気若しくは光データ記憶媒体、並びに同様のものなどの、メモリ又はデータ記憶媒体を含んでもよい。技法は、追加又は代替として、命令又はデータ構造の形態でのプログラムコードを搬送又は通信し得る、かつコンピュータによってアクセスされ、読み取られ、及び／又は実行され得る、伝搬される信号又は波などの、コンピュータ可読通信媒体によって少なくとも部分的に実現されてもよい。

10

20

30

40

50

【0190】

【0184】 プログラムコードは、1つ又は複数のデジタル信号プロセッサ (DSP)、汎用マイクロプロセッサ、特定用途向け集積回路 (ASIC)、フィールドプログラマブルロジックアレイ (FPGA)、又は他の同等の集積論理回路若しくは個別論理回路などの1つ又は複数のプロセッサを含み得る、プロセッサによって実行されてもよい。そのようなプロセッサは、本開示に記載された技法のいずれかを実施するように構成されてもよい。汎用プロセッサは、マイクロプロセッサであってもよいが、代わりに、プロセッサは、任意の従来のプロセッサ、コントローラ、マイクロコントローラ、又はステートマシンであってもよい。プロセッサはまた、コンピューティングデバイスの組み合わせ、例えば、DSPとマイクロプロセッサとの組み合わせ、複数のマイクロプロセッサ、DSPコアと連携した1つ又は複数のマイクロプロセッサ、又は任意の他のそのような構成として実装されてもよい。したがって、本明細書で使用する「プロセッサ」という用語は、上記の構造、上記の構造の任意の組み合わせ、又は本明細書で説明した技法の実装に適した任意の他の構造若しくは装置のうちのいずれかを指すことがある。

【0191】

【0185】 本開示の例示的な態様は以下を含む。

【0186】 態様1：メディアデータを処理するための装置であって、少なくとも1つのメモリと、少なくとも1つのメモリに結合された少なくとも1つのプロセッサと、を備え、少なくとも1つのプロセッサが、符号化画像データのフレームの潜在表現を取得し、符号化画像データのフレームの潜在表現を入力として使用して、デコーダサブネットワークの複数のデコーダトランスフォーマ層に基づいて、復号画像データのフレームを生成するように構成されており、複数のデコーダトランスフォーマ層のうちの少なくとも1つのデコーダトランスフォーマ層が、特徴の1つ又は複数のパッチを生成し、1つ又は複数のパッチにわたって適用された1つ又は複数のウィンドウ区分及びシフトウィンドウ区分内で局所的に自己アテンションを決定するように構成された1つ又は複数のトランスフォーマブロックと、1つ又は複数のパッチの各パッチのそれぞれのサイズを減少させるように構成されたパッチ分離エンジンと、を含む、装置。

【0192】

【0187】 態様2：復号画像データのフレームを生成するために、少なくとも1つのプロセッサが、複数のデコーダトランスフォーマ層のうちの第1のデコーダトランスフォーマ層の第1のトランスフォーマブロックを使用して、1つ又は複数のパッチにわたって適用された1つ又は複数の第1のウィンドウ区分内で局所的に自己アテンションを決定し、第1のデコーダトランスフォーマ層の第2のトランスフォーマブロックを使用して、1つ又は複数のパッチにわたって適用された1つ又は複数の第2のウィンドウ区分であって、

1つ又は複数の第1のウィンドウ区分のうちの隣接したものの間の1つ又は複数の境界に重なるようにシフトされている、1つ又は複数の第2のウィンドウ区分内で局所的に自己アテンションを決定し、パッチ分離エンジンを使用して、1つ又は複数のパッチの各パッチを、重なり合わない複数の分離パッチにセグメント化するように構成されている、態様1に記載の装置。

【0193】

【0188】 態様3：少なくとも1つのプロセッサが、複数の分離パッチを複数のデコーダトランスフォーマ層のうちの第2のデコーダトランスフォーマ層の第1のトランスフォーマブロックに提供するように構成されている、態様2に記載の装置。

【0194】

【0189】 態様4：少なくとも1つのプロセッサが、第2のデコーダトランスフォーマ層のパッチ分離エンジンを使用して、複数の分離パッチをセグメント化し、パッチ分離エンジンの出力を複数のデコーダトランスフォーマ層のうちの第3のデコーダトランスフォーマ層に提供するように構成されている、態様3に記載の装置。

【0195】

【0190】 態様5：複数の分離パッチの各分離パッチが均一なパッチサイズを有し、パッチ分離エンジンが2のパッチサイズ縮小係数を適用する、態様2から4のいずれかに記載の装置。

【0196】

【0191】 態様6：1つ又は複数のパッチの各パッチを複数の分離パッチにセグメント化するために、少なくとも1つのプロセッサが、複数の分離パッチの特徴次元を減少させるように構成されている、態様2から5のいずれかに記載の装置。

【0197】

【0192】 態様7：複数のデコーダトランスフォーマ層が、符号化画像データのフレームの潜在表現を入力として受け取り、非線形変換を適用し、復号画像データのフレームを生成するように構成されている、態様1から6のいずれかに記載の装置。

【0198】

【0193】 態様8：非線形変換が合成変換であり、復号画像データのフレームが、符号化画像データのフレームに関連付けられた入力画像の復元である、態様7に記載の装置。

【0199】

【0194】 態様9：少なくとも1つのプロセッサが、レート歪みに少なくとも部分的に基づく損失関数を使用して複数のデコーダトランスフォーマ層のうちの1つ又は複数のデコーダトランスフォーマ層を訓練するように構成されている、態様1から8のいずれかに記載の装置。

【0200】

【0195】 態様10：損失関数がレート歪みのためのラグランジュ乗数を含む、態様9に記載の装置。

【0201】

【0196】 態様11：少なくとも1つのデコーダトランスフォーマ層内に含まれる1つ又は複数のトランスフォーマブロックの少なくとも部分が同じアーキテクチャを有する、態様1から10のいずれかに記載の装置。

【0202】

【0197】 態様12：少なくとも1つのデコーダトランスフォーマ層内に含まれる1つ又は複数のトランスフォーマブロックの各々が同じアーキテクチャを有する、態様1から11のいずれかに記載の装置。

【0203】

【0198】 態様13：符号化画像データのフレームが符号化静止画像を含む、態様1から12のいずれかに記載の装置。

【0204】

【0199】 態様14：符号化画像データのフレームが符号化ビデオフレームを含む、態

10

20

30

40

50

様 1 から 1 3 のいずれかに記載の装置。

【0205】

【0200】 態様 1 5：少なくとも 1 つのプロセッサが、少なくとも第 1 の訓練データセット及び第 2 の訓練データセットで複数のデコーダトランスフォーマ層を訓練するように構成されており、第 2 の訓練データセットのデータが、第 1 の訓練データセットのデータと比べて逆の時間的順序を有する、態様 1 から 1 4 のいずれかに記載の装置。

【0206】

【0201】 態様 1 6：複数のデコーダトランスフォーマ層が一連の連続したデコーダトランスフォーマ層を含む、態様 1 から 1 5 のいずれかに記載の装置。

【0207】

【0202】 態様 1 7：メディアデータを処理するための装置であって、少なくとも 1 つのメモリと、少なくとも 1 つのメモリに結合された少なくとも 1 つのプロセッサと、を備え、少なくとも 1 つのプロセッサが、フレームを複数のパッチにセグメント化し、複数のパッチを入力として使用して、エンコーダサブネットワークの複数のエンコーダトランスフォーマ層に基づいて、符号化画像データのフレームを生成するように構成されている、装置。

【0208】

【0203】 態様 1 8：符号化画像データのフレームを生成するために、少なくとも 1 つのプロセッサが、複数のエンコーダトランスフォーマ層のうちの第 1 のエンコーダトランスフォーマ層の第 1 のトランスフォーマブロックを使用して、1 つ又は複数のウィンドウ区分内で局所的に自己アテンションを決定し、第 1 のエンコーダトランスフォーマ層の第 2 のトランスフォーマブロックを使用して、1 つ又は複数のウィンドウ区分に重なる 1 つ又は複数のシフトウィンドウ区分内で局所的に自己アテンションを決定し、第 1 のトランスフォーマブロック及び第 2 のトランスフォーマブロックのうちの 1 つ又は複数を使用して、非線形変換をセグメント化されたフレームに適用するための特徴の 1 つ又は複数のパッチを決定し、パッチ併合エンジンを使用して、第 1 のエンコーダトランスフォーマ層と第 2 のエンコーダトランスフォーマ層との間でパッチサイズを増大させるように構成されている、態様 1 7 に記載の装置。

【0209】

【0204】 態様 1 9：パッチ併合エンジンが、第 1 のエンコーダトランスフォーマ層からの複数の隣接したパッチを、第 2 のエンコーダトランスフォーマ層に提供される併合パッチに組み合わせるように構成されている、態様 1 8 に記載の装置。

【0210】

【0205】 態様 2 0：第 1 のエンコーダトランスフォーマ層の第 2 のトランスフォーマブロックの出力が第 2 のエンコーダトランスフォーマ層の入力に結合される、態様 1 8 又は 1 9 のいずれかに記載の装置。

【0211】

【0206】 態様 2 1：少なくとも 1 つのプロセッサが、複数のパッチを入力として使用して、エンコーダサブネットワークの複数のエンコーダトランスフォーマ層を使用して、セグメント化されたフレームのための階層的特徴マップを生成し、階層的特徴マップから符号化画像データのフレームを生成するように構成されている、態様 1 8 から 2 0 のいずれかに記載の装置。

【0212】

【0207】 態様 2 2：複数のパッチの各パッチが均一なサイズのものであり、セグメント化されたフレームの 1 つ又は複数のピクセルを含む、態様 1 8 から 2 1 のいずれかに記載の装置。

【0213】

【0208】 態様 2 3：パッチ併合エンジンが、隣接したパッチの 1 つ又は複数の部分セットから取得された特徴を連結することによってパッチサイズを増大させるように構成されており、隣接したパッチの各部分セットが、パッチ併合エンジンによって出力される併

10

20

30

40

50

合パッチに併合される、態様 18 から 22 のいずれかに記載の装置。

【0214】

【0209】 態様 24：第 1 のトランスフォーマブロック及び第 2 のトランスフォーマブロックが同じアーキテクチャを有する、態様 18 から 23 のいずれかに記載の装置。

【0215】

【0210】 態様 25：少なくとも 1 つのプロセッサが、第 1 のエンコーダトランスフォーマ層の前に複数のパッチをエンコーダサブネットワークの線形埋め込み層に提供するように構成されている、態様 18 から 24 のいずれかに記載の装置。

【0216】

【0211】 態様 26：符号化画像データのフレームが画像データの潜在表現である、態様 17 から 25 のいずれかに記載の装置。

【0217】

【0212】 態様 27：潜在表現が、エンコーダサブネットワークの複数のエンコーダトランスフォーマ層によって生成された階層的特徴マップである、態様 26 に記載の装置。

【0218】

【0213】 態様 28：少なくとも 1 つのプロセッサが、レート歪み損失に基づく損失関数を使用して複数のエンコーダトランスフォーマ層のうちの 1 つ又は複数のエンコーダトランスフォーマ層を訓練するように構成されている、態様 17 から 27 のいずれかに記載の装置。

【0219】

【0214】 態様 29：損失関数がレート歪みのためのラグランジュ乗数を含む、態様 28 に記載の装置。

【0220】

【0215】 態様 30：複数のパッチが、静止画像フレーム又はビデオフレームを含む入力からセグメント化される、態様 17 から 29 のいずれかに記載の装置。

【0221】

【0216】 態様 31：少なくとも 1 つのプロセッサが、因数分解プライアを使用して符号化画像データをエントロピコーディングするように構成されている、態様 17 から 30 のいずれかに記載の装置。

【0222】

【0217】 態様 32：メディアデータを処理する方法であって、方法が、符号化画像データのフレームの潜在表現を取得することと、符号化画像データのフレームの潜在表現を入力として使用して、デコーダサブネットワークの複数のデコーダトランスフォーマ層によって、復号画像データのフレームを生成することと、を含み、複数のデコーダトランスフォーマ層のうちの少なくとも 1 つのデコーダトランスフォーマ層が、特徴の 1 つ又は複数のパッチを生成し、1 つ又は複数のパッチにわたって適用された 1 つ又は複数のウィンドウ区分及びシフトウィンドウ区分内で局所的に自己アテンションを決定するための 1 つ又は複数のトランスフォーマブロックと、1 つ又は複数のパッチの各パッチのそれぞれのサイズを減少させるためのパッチ分離エンジンと、を含む、方法。

【0223】

【0218】 態様 33：複数のデコーダトランスフォーマ層によって、復号画像データのフレームを生成することが、複数のデコーダトランスフォーマ層のうちの第 1 のデコーダトランスフォーマ層の第 1 のトランスフォーマブロックによって、1 つ又は複数のパッチにわたって適用された 1 つ又は複数の第 1 のウィンドウ区分内で局所的に自己アテンションを決定することと、第 1 のデコーダトランスフォーマ層の第 2 のトランスフォーマブロックによって、1 つ又は複数のパッチにわたって適用された 1 つ又は複数の第 2 のウィンドウ区分であって、1 つ又は複数の第 1 のウィンドウ区分のうちの隣接したものの間の 1 つ又は複数の境界に重なるようにシフトされている、1 つ又は複数の第 2 のウィンドウ区分内で局所的に自己アテンションを決定することと、パッチ分離エンジンによって、1 つ又は複数のパッチの各パッチを、重なり合わない複数の分離パッチにセグメント化するこ

10

20

30

40

50

とと、を含む、態様 3 2 に記載の方法。

【0224】

【0219】 態様 3 4：複数の分離パッチを複数のデコーダトランスフォーマ層のうちの第 2 のデコーダトランスフォーマ層の第 1 のトランスフォーマブロックに提供することを更に含む、態様 3 3 に記載の方法。

【0225】

【0220】 態様 3 5：第 2 のデコーダトランスフォーマ層のパッチ分離エンジンによって、複数の分離パッチをセグメント化することと、パッチ分離エンジンの出力を複数のデコーダトランスフォーマ層のうちの第 3 のデコーダトランスフォーマ層に提供することと、を更に含む、態様 3 4 に記載の方法。

【0226】

【0221】 態様 3 6：複数の分離パッチの各分離パッチが均一なパッチサイズを有し、パッチ分離エンジンが 2 のパッチサイズ縮小係数を適用する、態様 3 3 から 3 5 のいずれかに記載の方法。

【0227】

【0222】 態様 3 7：1 つ又は複数のパッチの各パッチを複数の分離パッチにセグメント化することが、複数の分離パッチの特徴次元を減少させることを更に含む、態様 3 3 から 3 6 のいずれかに記載の方法。

【0228】

【0223】 態様 3 8：複数のデコーダトランスフォーマ層によって、符号化画像データのフレームの潜在表現を入力として受け取り、非線形変換を適用し、復号画像データのフレームを生成することを更に含む、態様 3 2 から 3 7 のいずれかに記載の方法。

【0229】

【0224】 態様 3 9：非線形変換が合成変換であり、復号画像データのフレームが、符号化画像データのフレームに関連付けられた入力画像の復元である、態様 3 8 に記載の方法。

【0230】

【0225】 態様 4 0：レート歪みに少なくとも部分的に基づく損失関数を使用して複数のデコーダトランスフォーマ層のうちの 1 つ又は複数のデコーダトランスフォーマ層を訓練することを更に含む、態様 3 2 から 3 9 のいずれかに記載の方法。

【0231】

【0226】 態様 4 1：損失関数がレート歪みのためのラグランジュ乗数を含む、態様 4 0 に記載の方法。

【0232】

【0227】 態様 4 2：少なくとも 1 つのデコーダトランスフォーマ層内に含まれる 1 つ又は複数のトランスフォーマブロックの少なくとも部分が同じアーキテクチャを有する、態様 3 2 から 4 1 のいずれかに記載の方法。

【0233】

【0228】 態様 4 3：少なくとも 1 つのデコーダトランスフォーマ層内に含まれる 1 つ又は複数のトランスフォーマブロックの各々が同じアーキテクチャを有する、態様 3 2 から 4 2 のいずれかに記載の方法。

【0234】

【0229】 態様 4 4：符号化画像データのフレームが符号化静止画像を含む、態様 3 2 から 4 3 のいずれかに記載の方法。

【0235】

【0230】 態様 4 5：符号化画像データのフレームが符号化ビデオフレームを含む、態様 3 2 から 4 4 のいずれかに記載の方法。

【0236】

【0231】 態様 4 6：少なくとも第 1 の訓練データセット及び第 2 の訓練データセットで複数のデコーダトランスフォーマ層を訓練することを更に含み、第 2 の訓練データセッ

10

20

30

40

50

トのデータが、第 1 の訓練データセットのデータと比べて逆の時間的順序を有する、態様 3 2 から 4 5 のいずれかに記載の方法。

【0 2 3 7】

【0232】 態様 4 7 : 複数のデコーダトランスフォーマ層が一連の連続したデコーダトランスフォーマ層を含む、態様 3 2 から 4 6 のいずれかに記載の方法。

【0 2 3 8】

【0233】 態様 4 8 : メディアデータを処理する方法であって、フレームを複数のパッチにセグメント化することと、複数のパッチを入力として使用して、エンコーダサブネットワークの複数のエンコーダトランスフォーマ層によって、符号化画像データのフレームを生成することと、を含む、方法。

10

【0 2 3 9】

【0234】 態様 4 9 : 符号化画像データのフレームを生成することが、複数のエンコーダトランスフォーマ層のうちの第 1 のエンコーダトランスフォーマ層の第 1 のトランスフォーマブロックによって、1 つ又は複数のウィンドウ区分内で局所的に自己アテンションを決定することと、第 1 のエンコーダトランスフォーマ層の第 2 のトランスフォーマブロックによって、1 つ又は複数のウィンドウ区分に重なる 1 つ又は複数のシフトウィンドウ区分内で局所的に自己アテンションを決定することと、第 1 のトランスフォーマブロック及び第 2 のトランスフォーマブロックのうちの 1 つ又は複数によって、非線形変換をセグメント化されたフレームに適用するための特徴の 1 つ又は複数のパッチを決定することと、パッチ併合エンジンによって、第 1 のエンコーダトランスフォーマ層と第 2 のエンコーダトランスフォーマ層との間でパッチサイズを増大させることと、を含む、態様 4 8 に記載の方法。

20

【0 2 4 0】

【0235】 態様 5 0 : 1 つ又は複数のプロセッサによって実行されたとき、1 つ又は複数のプロセッサに態様 1 から 1 6 及び態様 3 2 から 4 7 の動作のうちのいずれかを実行させる命令を記憶した、非一時的コンピュータ可読記憶媒体。

【0 2 4 1】

【0236】 態様 5 1 : 態様 1 から 1 6 及び態様 3 2 から 4 7 の動作のうちのいずれかを実行するための手段を備える装置。

【0 2 4 2】

30

【0237】 態様 5 2 : 態様 1 7 から 3 1 の動作のうちのいずれかを実行する方法。

【0 2 4 3】

【0238】 態様 5 3 : 1 つ又は複数のプロセッサによって実行されたとき、1 つ又は複数のプロセッサに態様 1 7 から 3 1 の動作のうちのいずれかを実行させる命令を記憶した、非一時的コンピュータ可読記憶媒体。

【0 2 4 4】

【0239】 態様 5 4 : 態様 1 7 から 3 1 の動作のうちのいずれかを実行するための手段を備える装置。

40

50

【図面】
【図 1】

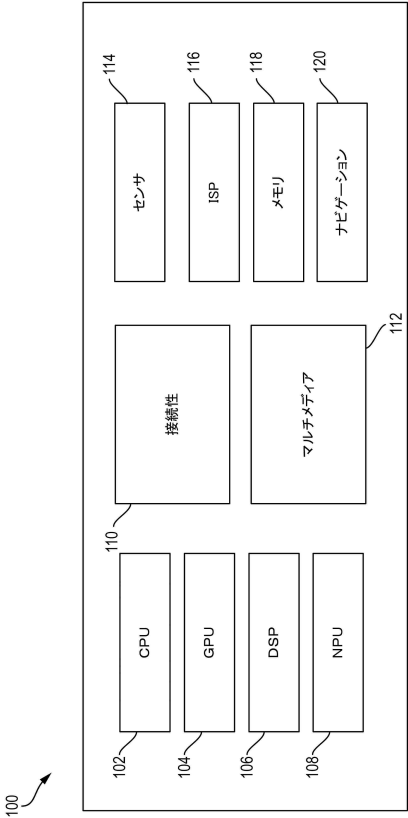


図 1

【図 2 A】

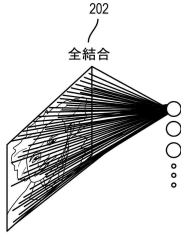


図2A

【図 2 B】

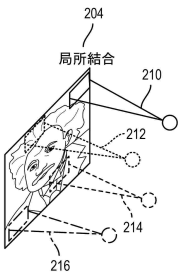


図2B

【図 3】

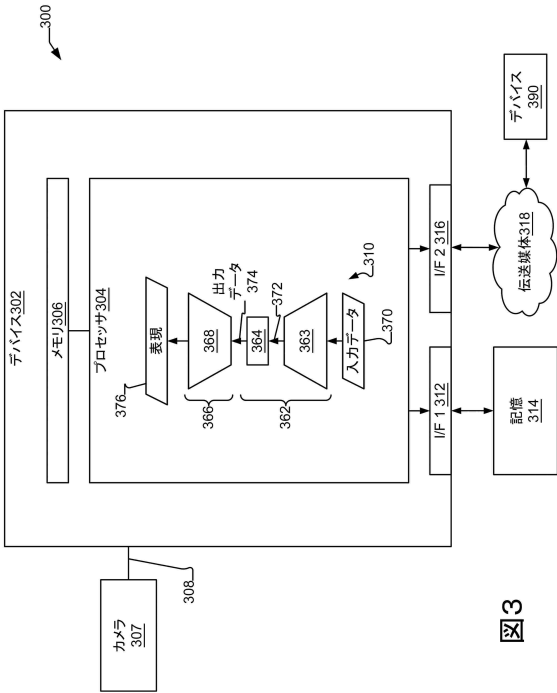


図 3

10

20

30

40

50

【図 4】

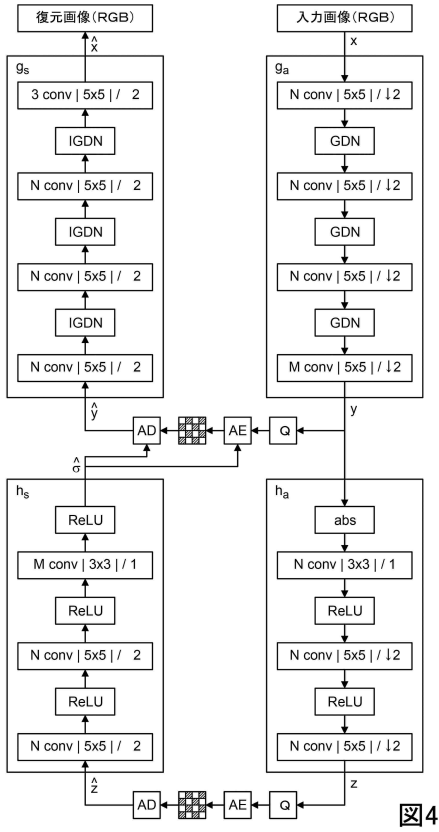


図4

【図 5 A】

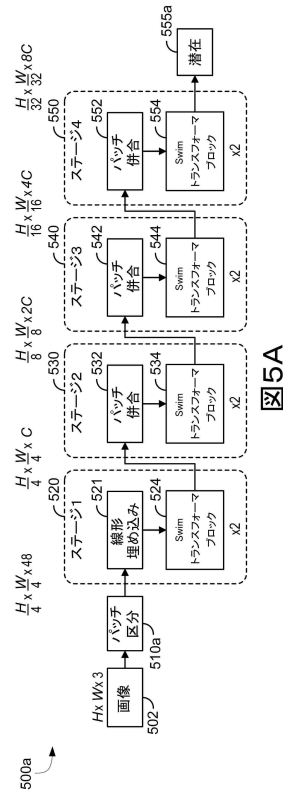


図5A

【図 5 B】

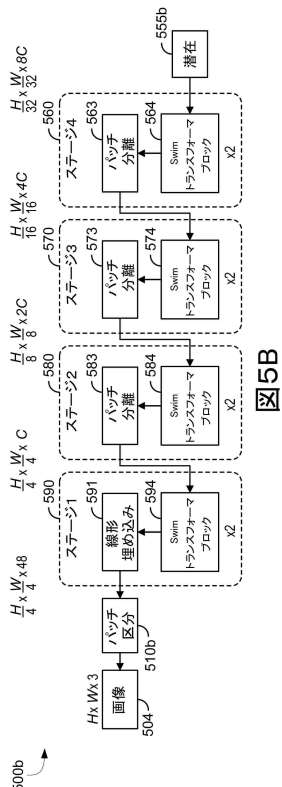


図5B

【図 5 C】

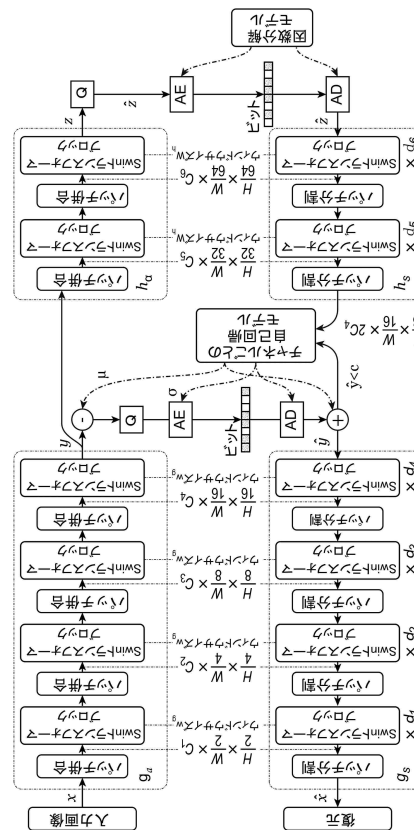


図5C

10

20

30

40

50

【図 8】

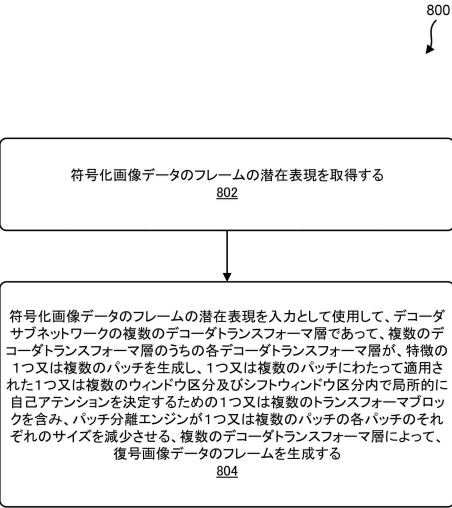


図8

【図 9】

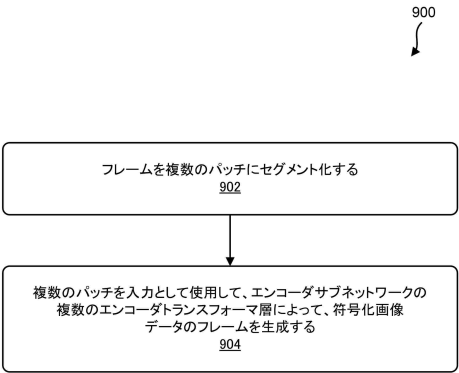


図9

【図 10】

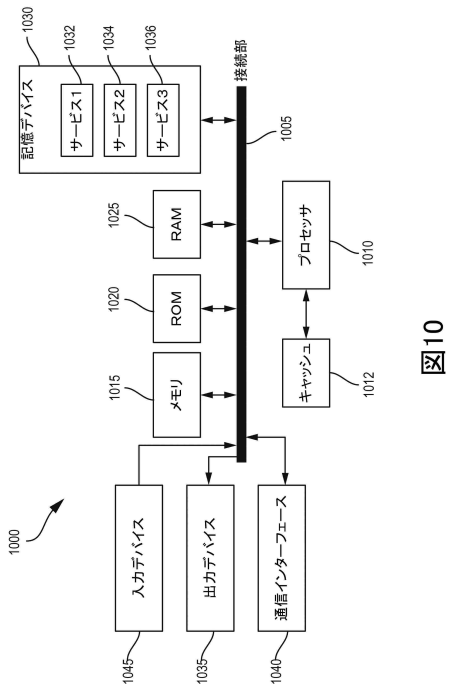


図10

10

20

30

40

50

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2022/076496

A. CLASSIFICATION OF SUBJECT MATTER		
INV.	G06N3/04	H04N19/59 H04N19/51 H04N19/91
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 H04N		
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.
X,P	Zhu Yinhao ET AL: "TRANSFORMER-BASED TRANSFORM CODING", , 29 September 2021 (2021-09-29), XP093005391, Retrieved from the Internet: URL:https://openreview.net/pdf?id=IDwN6xjHnR8 [retrieved on 2022-12-06] the whole document ----- -/--	1-49
☒ Further documents are listed in the continuation of Box C. ☐ 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 8 December 2022		Date of mailing of the international search report 20/12/2022
Name and mailing address of the ISA/ European Patent Office, P.B. 5618 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016		Authorized officer Cakiroglu Garton, S

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

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2022/076496

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	ZE LIU ET AL: "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows", ARKIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 17 August 2021 (2021-08-17), XP091024409, abstract; figures 2,3 paragraphs [0001], [0003] -----	17, 18, 48, 49
X	AILIANG LIN ET AL: "DS-TransUNet: Dual Swin Transformer U-Net for Medical Image Segmentation", ARKIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 12 June 2021 (2021-06-12), XP081988747, abstract; figures 1, 2 paragraphs [000I] - [0III]	1-8, 11-14, 16-27, 30, 32-39, 42-45, 47-49
Y		9, 10, 15, 28, 29, 31, 40, 41, 46
X	HU CAO ET AL: "Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation", ARKIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 12 May 2021 (2021-05-12), XP081962410, abstract; figures 1,2 paragraph [0003] -----	1, 17, 32, 48
Y	DAVID MINNEN ET AL: "Joint Autoregressive and Hierarchical Priors for Learned Image Compression", ARKIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 8 September 2018 (2018-09-08), XP081188741, paragraphs [0001], [0002] abstract -----	9, 10, 28, 29, 31, 40, 41
Y	ZHIHAO SHI ET AL: "Learning for Unconstrained Space-Time Video Super-Resolution", ARKIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 25 February 2021 (2021-02-25), XP081892562, paragraph [IV.A.2] ----- -/--	15, 46

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2022/076496

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	DAVID MINNEN ET AL: "Channel-wise Autoregressive Entropy Models for Learned Image Compression", ARKIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 17 July 2020 (2020-07-17), XP081722259, the whole document -----	1-49
A	FUSHUN ZHU ET AL: "Semi-Supervised Wide-Angle Portraits Correction by Multi-Scale Transformer", ARKIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 14 September 2021 (2021-09-14), XP091057092, the whole document -----	1-49

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,N
E,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,
CV,CZ,DE,DJ,DK,DM,DO,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,GT,HN,HR,HU,ID,IL,IN,IQ,IR,IS,IT,J
M,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,T
H,TJ,TM,TN,TR,TT,TZ,UA,UG,US,UZ,VC,VN,WS,ZA,ZM,ZW

2 1 2 1、サン・ディエゴ、モアハウス・ドライブ 5 7 7 5
(72)発明者 コーエン、タコ・セバスティアン
アメリカ合衆国、カリフォルニア州 9 2 1 2 1、サン・ディエゴ、モアハウス・ドライブ 5 7 7
5
F ターム (参考) 5C159 LA00 PP16 RC11 UA02 UA05