(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7193252号
(P7193252)

(45)発行日　令和4年12月20日(2022.12.20)

(24)登録日　令和4年12月12日(2022.12.12)

(51)国際特許分類

Ｇ０６Ｔ	*7/00 (2017.01)*	Ｇ０６Ｔ	7/00	３５０Ｃ
Ｇ０６Ｎ	*3/04 (2006.01)*	Ｇ０６Ｎ	3/04	１５４
Ｇ０６Ｎ	*3/08 (2006.01)*	Ｇ０６Ｎ	3/08	

FＩ

請求項の数　12　外国語出願　（全27頁）

(21)出願番号　特願2018-88032(P2018-88032)
(22)出願日　平成30年5月1日(2018.5.1)
(65)公開番号　特開2019-8778(P2019-8778A)
(43)公開日　平成31年1月17日(2019.1.17)
審査請求日　令和3年4月2日(2021.4.2)
(31)優先権主張番号　17305486.7
(32)優先日　平成29年5月2日(2017.5.2)
(33)優先権主張国・地域又は機関
欧州特許庁(EP)

(73)特許権者　500102435
ダッソー システムズ
ＤＡＳＳＡＵＬＴ ＳＹＳＴＥＭＥＳ
フランス国 ７８１４０ ベリジー ビラ
クブレー リュ マルセル ダッソー １０
(74)代理人　110000752
弁理士法人朝日特許事務所
(72)発明者　ニエルス ルバース
フランス国 ヴェリジーヴィラクブレー
マルセルダッソー通り１０ ダッソーシ
ステムズ 郵便番号７８１４０
(72)発明者　マリカ ブルキナフェド
フランス国 ヴェリジーヴィラクブレー
マルセルダッソー通り１０ ダッソーシ
ステムズ 郵便番号７８１４０

最終頁に続く

(54)【発明の名称】　画像の領域のキャプション付加

(57)【特許請求の範囲】

【請求項１】

それぞれが画像と当該画像の領域と当該領域のキャプションとを含む、３つ組のデータセットを提供するステップと、

入力画像と当該入力画像の入力領域とに基づいて出力キャプションを生成するように構成された関数を、前記３つ組のデータセットを用いて学習するステップと

を有することを特徴とする、画像の領域のキャプション付けを行うよう構成された関数を学習するためのコンピュータにより実施される方法であって、

前記関数は、前記入力画像と前記入力領域との組み合わせの署名を決定するように構成された第１の成分と、前記組み合わせの署名に基づいて前記出力キャプションを生成するように構成された第２の成分とを含み、

前記第２の成分は、単語埋め込み空間に基づいて前記出力キャプションを生成するように構成され、

前記第２の成分は、第１の再帰型ニューラルネットワークと、第２の再帰型ニューラルネットワークと、単語埋め込み層とを含んでおり、

前記第１の再帰型ニューラルネットワークは、各入力単語埋め込みベクトルに基づいて出力単語埋め込みベクトルを再帰的に生成するように構成されており、前記第１の再帰型ニューラルネットワークについての前記入力単語埋め込みベクトルは、前記第２の再帰型ニューラルネットワークとそれに続く前記単語埋め込み層との合成物の出力であり、

前記第２の再帰型ニューラルネットワークは、各入力単語埋め込みベクトルと前記組み合

わせの前記署名とに基づいて出力単語埋め込みベクトルを再帰的に生成するように構成されており、前記第２の再帰型ニューラルネットワークについての前記入力単語埋め込みベクトルは前記第１の再帰型ニューラルネットワークの出力であり、
前記単語埋め込み層は、各入力単語埋め込みベクトルに基づいて出力単語埋め込みベクトルを生成するように構成されており、前記単語埋め込み層についての前記生成された単語埋め込みベクトルは、単語埋め込み空間において表現される最も確率の高い語彙の単語に対応する単語埋め込みベクトルである
ことを特徴とする方法。
【請求項２】
前記第２の成分は、前記出力キャプションを反復的に生成するように構成されている
ことを特徴とする請求項１に記載の方法。
【請求項３】
前記第２の成分は、１つまたは複数の再帰型ニューラルネットワークを含む
ことを特徴とする請求項２に記載の方法。
【請求項４】
前記１つまたは複数の再帰型ニューラルネットワークは、１つまたは複数の長短期記憶（Ｌｏｎｇ　Ｓｈｏｒｔ－Ｔｅｒｍ　Ｍｅｍｏｒｙ）ニューラルネットワークを含む
ことを特徴とする請求項３に記載の方法。
【請求項５】
前記方法は、キャプションのデータセットを提供することをさらに含み、
前記学習することは、前記第１の再帰型ニューラルネットワークおよび／または前記単語埋め込み層を、前記キャプションのデータセットを用いて訓練し、次いで前記第２の再帰型ニューラルネットワークを訓練することを含む
ことを特徴とする請求項１ないし４のいずれか一つに記載の方法。
【請求項６】
前記第１の成分は、
前記入力画像の署名を抽出するように構成された成分と、
前記入力領域の署名を抽出するように構成された成分と、
前記入力画像の前記署名を前記入力領域の前記署名と組み合わせるように構成された成分と
を含むことを特徴とする請求項１～５のいずれか１つに記載の方法。
【請求項７】
前記入力画像の署名を抽出するように構成された前記成分と、前記入力領域の署名を抽出するように構成された前記成分とは、それぞれ畳み込みニューラルネットワークである
ことを特徴とする請求項６に記載の方法。
【請求項８】
各畳み込みニューラルネットワークは重みを共有する
ことを特徴とする請求項７に記載の方法。
【請求項９】
前記入力画像の前記署名を前記入力領域の前記署名と組み合わせるように構成された前記成分は、連結成分、または加算成分、および／または全結合層を含む
ことを特徴とする請求項６～８のいずれか一つに記載の方法。
【請求項１０】
前記学習することは、前記第１の成分を学習することと、次いで前記第２の成分の少なくとも一部を学習することとを含む
ことを特徴とする請求項１～９のいずれか１つに記載の方法。
【請求項１１】
コンピュータに、
それぞれが画像と当該画像の領域と当該領域のキャプションとを含む、３つ組のデータセットを提供するステップと、

入力画像と当該入力画像の入力領域とに基づいて出力キャプションを生成するように構成された関数を、前記３つ組のデータセットを用いて学習するステップと
を実行させるためのプログラムであって、
前記関数は、
前記入力画像と前記入力領域との組み合わせの署名を決定するように構成された第１の成分と、前記組み合わせの署名に基づいて前記出力キャプションを生成するように構成された第２の成分とを含み、
前記第２の成分は、さらに単語埋め込み空間に基づいて前記出力キャプションを生成するように構成され、
前記第２の成分は、第１の再帰型ニューラルネットワークと、第２の再帰型ニューラルネットワークと、単語埋め込み層とを含んでおり、
前記第１の再帰型ニューラルネットワークは、各入力単語埋め込みベクトルに基づいて出力単語埋め込みベクトルを再帰的に生成するように構成されており、前記第１の再帰型ニューラルネットワークについての前記入力単語埋め込みベクトルは、前記第２の再帰型ニューラルネットワークとそれに続く前記単語埋め込み層との合成物の出力であり、
前記第２の再帰型ニューラルネットワークは、各入力単語埋め込みベクトルと前記組み合わせの前記署名とに基づいて出力単語埋め込みベクトルを再帰的に生成するように構成されており、前記第２の再帰型ニューラルネットワークについての前記入力単語埋め込みベクトルは前記第１の再帰型ニューラルネットワークの出力であり、
前記単語埋め込み層は各入力単語埋め込みベクトルに基づいて出力単語埋め込みベクトルを生成するように構成されており、前記単語埋め込み層についての前記生成された単語埋め込みベクトルは、単語埋め込み空間において表現される最も確率の高い語彙の単語に対応する単語埋め込みベクトルである
ことを特徴とする、プログラム。
【請求項１２】
請求項１１に記載のプログラムを記録した記憶媒体と、前記記憶媒体に接続されたプロセッサとを備える装置。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、コンピュータプログラムおよびシステムの分野に関し、より具体的には、画像の領域のキャプション付けのために構成された関数の学習に関連する方法、装置、データ構造、およびプログラムに関する。
【背景技術】
【０００２】
オブジェクトの設計、エンジニアリング、製造のため、多数のシステムおよびプログラムが市場に提供されている。ＣＡＤは、コンピュータ支援設計（Ｃｏｍｐｕｔｅｒ－Ａｉｄｅｄ　Ｄｅｓｉｇｎ）の略語であり、例えば、オブジェクトを設計するためのソフトウェア・ソリューションに関する。ＣＡＥは、コンピュータ支援エンジニアリング（Ｃｏｍｐｕｔｅｒ－Ａｉｄｅｄ　Ｅｎｇｉｎｅｅｒｉｎｇ）の略語であり、例えば、将来の製品の物理的挙動をシミュレーションするためのソフトウェア・ソリューションに関する。ＣＡＭは、コンピュータ支援製造（Ｃｏｍｐｕｔｅｒ－Ａｉｄｅｄ　Ｍａｎｕｆａｃｔｕｒｉｎｇ）の略語であり、例えば、製造工程および動作を定義するためのソフトウェア・ソリューションに関する。このようなコンピュータ支援設計システムにおいて、グラフィカル・ユーザ・インターフェースは、技術の効率に関して、重要な役割を果たす。これらの技術は、製品ライフサイクル管理（Ｐｒｏｄｕｃｔ　Ｌｉｆｅｃｙｃｌｅ　Ｍａｎａｇｅｍｅｎｔ：　ＰＬＭ）システムに組み込むことができる。ＰＬＭとは、企業が、拡張エンタープライズの概念全体にわたって、製品データを共有し、共通の工程を適用し、構想に始まり製品寿命の終わりに至る製品開発のための企業知識を活用するのを支援するビジネス戦略を指す。ダッソー・システムズが提供するＰＬＭソリューション（製品名ＣＡＴＩＡ

、ＥＮＯＶＩＡ、ＤＥＬＭＩＡ）は、製品エンジニアリング知識を体系化するエンジニアリング・ハブ、製造エンジニアリング知識を管理する製造ハブ、およびエンジニアリング・ハブと製造ハブの両方に対するエンタープライズ統合と接続を可能にするエンタープライズ・ハブを提供する。全てのシステムは、製品、工程、リソースを結ぶオープンなオブジェクトモデルを提供し、最適化された製品定義、製造準備、生産およびサービスを推進する、動的な知識ベースの製品作成および意思決定支援を可能にする。

【０００３】

こうした背景において、シーン理解および画像のキャプション付けがより重要視されるようになってきている。画像のキャプション付けは、コンピュータビジョンと自然言語処理の交差領域における課題であり、入力画像が与えられると入力画像を説明するキャプションを生成することからなる。領域のキャプション付けは、入力画像と、入力画像内の注目入力領域が与えられると、入力領域を説明するキャプションを生成することからなる、特定の種類の画像キャプション付けである。密なキャプション付けは、さらに進んだアプローチである：画像内の異なる注目領域を自動的に見つけ出し、それぞれに説明を与える。これらの技術は、例えば画像からの３Ｄ体験の自動生成に対して提供することにより、シーン理解アプリケーションにおいて有用であり得る。

以下の論文が画像キャプション付けに関連しており、以降で言及している。

【先行技術文献】

【非特許文献】

【０００４】

【文献】・[1] R. Krishna et al. Visual Genome：Connecting Language and Vision Using Crowdsourced Dense Image Annotations, arXiv 2016
・[2] R. Kiros et al. Unifying Visual−Semantic Embeddings with Multimodal Neural Language Models, ICCV 2015
・[3] R. Lebret et al. Phrase−Based Image Captioning, 2015
・[4] R. Kiros et al. Multimodal Neural Language Models, ICML 2014
・[5] T. Mikolov et al. Distributed Representations of Words and Phrases and their Compositionality, NIPS 2013
・[6] S. Venugopalan et al. Long−term Recurrent Convolutional Networks for Visual Recognition and Description, CVPR 2015
・[7] O. Vinyals et al. Show and Tell：A neural Image Caption Generator, IEEE 2015
・[8] A. Karpathy et al. Deep Visual−Semantic Alignments for Generating Image Descriptions, IEEE 2015
・[9] A. Karpathy et al. DenseCap：Fully Convolutional Localization Networks for Dense Captioning, CVPR 2016
・[10] K. Papineni et al. BLEU：a Method for Automatic Evaluation of Machine Translation, ACL 2002
・[11] M. Denkowski et al. Meteor Universal：Language Specific Translation Evaluation for Any Target Language ACL 2014

・［１２］Ｉ．Ｓｕｔｓｋｅｖｅｒ ｅｔ ａｌ．Ｓｅｑｕｅｎｃｅ ｔｏ Ｓｅｑｕｅｎｃｅ Ｌｅａｒｎｉｎｇ ｗｉｔｈ Ｎｅｕｒａｌ Ｎｅｔｗｏｒｋｓ，ＮＩＰＳ ２０１４

【発明の概要】

【発明が解決しようとする課題】

【０００５】

既存の画像キャプション付け技術は、キャプションを生成するように構成された機械学習モデル（すなわち、関数）を訓練するのに用いられる画像／キャプションの複数の対からなるデータベースに基づいている。そのようなデータベースは、人々が写真を説明するキャプションを書くよう求められるクラウドソーシング・プラットフォームから得てもよい。既存のデータベースには、画像キャプション付け用のＭＳＣＯＣＯと、密なキャプション付け用のＶｉｓｕａｌ Ｇｅｎｏｍｅ［１］とが含まれる。既存のキャプション付けのアプローチは、２つのカテゴリからなる。すなわち、学習されたマルチモーダル空間からの文の取得と、エンコーダ／デコーダフレームワークによる文の生成である。どちらのアプローチでも、モデル内の入力画像が符号化され、画像署名が取得される。次いで、その署名を処理した後、キャプションが取得される。生成されたキャプションの品質の評価は、異なる言語尺度［１０，１１］によって実行されてもよい。

【０００６】

マルチモーダル・アプローチ［２，３，４］では、画像とフレーズ表現のための共通の空間が学習される。このような共通空間は、Ｗｏｒｄ２Ｖｅｃ™を学習するときに［５］で用いられるようなネガティブ・サンプリングなどの技術を用いて学習される、２つの様式、すなわち、画像とテキストのための埋め込み空間のようなものである。そのような空間が学習されると、埋め込み空間における画像照会署名に最も類似した署名を有するキャプションを取得した後、文生成プロセスが実行される。このようなアプローチの課題は、取得したキャプションに対し、既にデータベースに存在するキャプションによる大きなバイアスがかかることである。さらに、最も類似したキャプションの取得は、データベースが大きくなり過ぎた場合、非常に時間がかかり得る動作である。

【０００７】

第２のアプローチでは、文の生成のためにエンコーダ／デコーダフレームワークが使用される［６，７，８］。画像を符号化する第１のステップでは、画像を畳み込みニューラルネットワークを通過させ、より高い全結合層のうちのいくつかの出力を得た後に、画像の署名が取得される。次いで、［１２］において開発された一般的なアプローチのように、文を一語一語生成する再帰型ニューラルネットワークにより、画像署名が復号される。密なキャプション付け処理はまた、画像内の領域のキャプションを生成する際、上述のエンコーダ／デコーダフレームワークを用いる。最先端の方法［９］は、ニューラルネットワークの内部にＬｏｃａｌｉｚａｔｉｏｎ層を統合して、画像内の注目領域を自動的に見つける。これらのアプローチは、訓練したデータベースの品質が十分である限り、画像全体の説明に適している。しかしながら、同じモデルが、画像内の領域のキャプションを生成するのに用いられると、画像全体における場合ほど良好な結果が得られない。したがって、画像の領域のキャプション付けの、改善された解決策が依然として必要とされている。

【課題を解決するための手段】

【０００８】

したがって、本発明では、コンピュータによって実施される、関数を学習するための方法が提供される。本方法は、３つ組のデータセットを提供することを含む。３つ組のそれぞれは、画像と、画像の領域と、領域のキャプションとを含む。本方法はまた、入力画像と当該入力画像の入力領域とに基づいて出力キャプションを生成するように構成された関数を、前記３つ組のデータセットを用いて学習することを含む。前記関数は、それにより画像の領域のキャプション付けを行うよう構成されている。

【０００９】

関数は、入力画像および入力画像の入力領域に基づいて出力キャプションを事前に生成するために予め学習されるので、キャプション付けは比較的高速、例えば実質的にリアル

タイムで実行されてもよい。さらに、従来技術と比較して、前記関数は相対的に高品質、かつ／あるいは、相対的に高いロバスト性を有するキャプション付けを実行するように構成されている。言い換えれば、本方法によって出力されるキャプションは、入力画像の入力領域を、比較的正確に、かつ／あるいは、構文的および／または文法的に比較的正しい言語を用いて説明し、かつ／あるいは、学習に用いられるデータセットの品質に関して比較的高いロバスト性を有する。これは、学習が、画像の領域とそれに対応するキャプションだけでなく、領域自体を含む画像をも含む、３つ組のデータセットを用いて、関数が、入力領域の情報だけでなく、入力領域を含む入力画像の情報にも基づいて出力を生成するように行われるからである。言い換えれば、本方法は、キャプション付けの品質を向上させるために、領域のコンテクスト、すなわちそれが含まれる画像を利用する。

【００１０】

これは、文生成の工程において領域の内側の情報だけを使用し、その周囲の情報を使用しない、既知の領域説明のための方法とは異なる。これは、キャプションを生成する際に、コンテクスト情報に基づかないことを意味する。さらに、既知の訓練されたモデルは、訓練されたデータベースの影響を非常に受けやすい。

【００１１】

本方法は、以下のうちの１つまたは複数を含んでいてもよい。

・前記関数は前記入力画像と前記入力領域との組み合わせの署名を決定するように構成された第１の成分と、前記組み合わせの前記署名に基づき前記出力キャプションを生成するよう構成された第２の成分とを含む。

・前記第２の成分は、前記出力キャプションを反復的に生成するように構成されている。

・前記第２の成分は、１つまたは複数の再帰型ニューラルネットワークを含む。

・前記１つまたは複数の再帰型ニューラルネットワークは、１つまたは複数の長短期記憶（Ｌｏｎｇ　Ｓｈｏｒｔ－Ｔｅｒｍ　Ｍｅｍｏｒｙ）ニューラルネットワークを含む。

・前記第２の成分は、さらに単語埋め込み空間に基づいて前記出力キャプションを生成するように構成されている。

・前記第２の成分は、第１の再帰型ニューラルネットワークと、第２の再帰型ニューラルネットワークと、単語埋め込み層とを含んでおり、前記第１の再帰型ニューラルネットワークは、各入力単語埋め込みベクトルに基づいて出力単語埋め込みベクトルを再帰的に生成するように構成されており、前記第１の再帰型ニューラルネットワークについての前記入力単語埋め込みベクトルは、前記第２の再帰型ニューラルネットワークとそれに続く前記単語埋め込み層との合成物の出力であり、前記第２の再帰型ニューラルネットワークは、各入力単語埋め込みベクトルと前記組み合わせの前記署名とに基づいて出力単語埋め込みベクトルを再帰的に生成するように構成されており、前記第２の再帰型ニューラルネットワークについての前記入力単語埋め込みベクトルは前記第１の再帰型ニューラルネットワークの出力であり、前記単語埋め込み層は各入力単語埋め込みベクトルに基づいて出力単語埋め込みベクトルを生成するように構成されており、前記単語埋め込み層についての前記生成された単語埋め込みベクトルは、単語埋め込み空間において表現される最も確率の高い語彙の単語に対応する単語埋め込みベクトルである。

【００１２】

・本方法は、キャプションのデータセットを提供することを含み、前記学習することは、前記第１の再帰型ニューラルネットワーク、および／または、前記単語埋め込み層を、前記キャプションのデータセットを用いて訓練し、次いで、前記第２の再帰型ニューラルネットワークを訓練することを含む。

前記第１の成分は、前記入力画像の署名を抽出するように構成された成分と、前記入力領域の署名を抽出するように構成された成分と、前記入力画像の前記署名を前記入力領域の前記署名と組み合わせるように構成された成分とを含む。

【００１３】

・前記入力画像の署名を抽出するように構成された前記成分と、前記入力領域の署名を抽出するように構成された前記成分とは、それぞれ畳み込みニューラルネットワークであ

る。
・各畳み込みニューラルネットワークは重みを共有する。
・前記入力画像の前記署名を前記入力領域の前記署名と組み合わせるように構成された前記成分は、連結成分、または加算成分、および／または、全結合層を含む。かつ／あるいは、
・前記学習することは、前記第１の成分を学習することと、次いで前記第２の成分の少なくとも一部を学習することとを含む。
【００１４】
さらには、本方法によって学習可能な関数が提供される。言い換えれば、当該関数は、入力を出力に変換するスキームを構成し、当該スキームは、本方法によって取得可能である。上記関数は、画像の領域のキャプション付けを行うための、コンピュータによって実施される工程において用いられてもよい。当該工程は、例えば、入力画像および入力領域を提供することと、当該入力画像および当該入力領域に基づいて、上記関数を適用して出力キャプションを生成することを含んでいてもよい。入力領域は、例えば、ユーザによって、または他の任意の方法で入力画像が提供された後に、当該入力画像内で識別されてもよい（例えば、密なキャプション付けを構成する工程で行われ、当該識別は、そのような密なキャプション付けにおける任意の古典的な領域識別段階を用いて行う）。当該工程は、例えば同じ関数を用いて異なる入力に繰り返してもよい。当該繰り返しは、密なキャプション付けで識別された同じ画像の異なる領域にわたって実行してもよい。これに代えて、またはこれに加えて、上記入力は、ビデオのフレームのシーケンスを構成し、上記工程は、各フレームについて、リアルタイムで、１つまたは複数の領域のキャプションを出力してもよい。
【００１５】
さらには、前記方法、および／または、前記工程を実行するための命令を含むコンピュータプログラムが提供される。
【００１６】
さらには、前記データセット、前記関数、および／または、前記プログラムを含むデータ構造が提供される。
【００１７】
さらには、前記データ構造を記録したコンピュータ読み取り可能な記憶媒体が提供される。
【００１８】
さらには、前記データ構造を記録したデータ記憶媒体を備える装置が提供される。前記装置は、非一時的コンピュータ読み取り可能媒体を構成してもよい。あるいは、前記装置は、前記データ記憶媒体に接続されたプロセッサを備えていてもよい。前記装置は、そのようにシステムを構成してもよい。前記システムはさらに、前記プロセッサに接続されたグラフィカル・ユーザ・インターフェースを備えていてもよい。
以下、非限定的な例として、本発明の実施の形態を添付の図面を参照しつつ説明する。
【図面の簡単な説明】
【００１９】
【図１】本システムの一例を示す。
【図２】本方法の例を示す。
【図３】本方法の例を示す。
【図４】本方法の例を示す。
【図５】本方法の例を示す。
【図６】本方法の例を示す。
【図７】本方法の例を示す。
【図８】本方法の例を示す。
【図９】本方法の例を示す。
【図１０】本方法の例を示す。

【図１１】本方法の例を示す。
【図１２】本方法の例を示す。
【図１３】本方法の例を示す。
【発明を実施するための形態】
【００２０】
「コンピュータにより実施される」とは、すなわち、ステップ（あるいは略全てのステップ）が少なくとも１つのコンピュータ、または類似の任意のシステムによって実行されることを意味する。よってステップは、コンピュータにより、完全に自動的に、あるいは半自動的に実行される可能性がある。例えば、少なくともいくつかのステップは、ユーザとコンピュータの対話を通じて始動されてもよい。求められるユーザとコンピュータの対話レベルは、想定される自動性のレベルに応じたものであって、ユーザの要望を実装する必要性との間でバランスをとるものとしてもよい。例えば、このレベルは、ユーザが設定し、かつ／あるいは、予め定義されていてもよい。
【００２１】
方法のコンピュータによる実施の典型的な例は、この目的に適したシステムを用いて本方法を実行することである。当該システムは、本方法を実行するための命令を含むコンピュータプログラムを記録したメモリに接続されたプロセッサ、および、グラフィカル・ユーザ・インターフェース（ＧＵＩ）を備えていてもよい。メモリは、データベースを記憶していてもよい。メモリは、そのような記憶に適した任意のハードウェアであり、場合により、物理的に区別可能ないくつかの部分（例えば、プログラム用に１つ、場合によりデータベース用に１つ）を含む。
【００２２】
図１は、本システムの一例を示すものであって、当該システムは、クライアントコンピュータシステム、例えばユーザのワークステーションである。
本例のクライアントコンピュータは、内部通信バス１０００に接続された中央演算処理装置（ＣＰＵ）１０１０、および同じくバスに接続されたランダムアクセスメモリ（ＲＡＭ）１０７０とを備える。クライアントコンピュータは、さらに、バスに接続されたビデオランダムアクセスメモリ１１００と関連付けられたグラフィックス処理装置（ＧＰＵ）１１１０を備える。ビデオＲＡＭ１１００は、当該技術分野において、フレームバッファとしても知られる。大容量記憶装置コントローラ１０２０は、ハードドライブ１０３０などの大容量記憶装置へのアクセスを管理する。コンピュータプログラムの命令及びデータを具体的に実現するのに適した大容量メモリ装置は、例として、ＥＰＲＯＭ、ＥＥＰＲＯＭ及びフラッシュメモリ装置のような半導体メモリ装置、内蔵ハードディスクやリムーバブルディスクなどの磁気ディスク、光磁気ディスク、およびＣＤ－ＲＯＭディスク１０４０を含む、全ての形式の不揮発性メモリを含む。前述のいずれも、特別に設計されたＡＳＩＣ（特定用途向け集積回路）によって補完されてもよいし、組み入れられてもよい。ネットワークアダプタ１０５０は、ネットワーク１０６０へのアクセスを管理する。クライアントコンピュータはまた、カーソル制御装置、キーボードなどの触覚装置１０９０を含んでいてもよい。カーソル制御装置は、ユーザがディスプレイ１０８０上の任意の所望の位置にカーソルを選択的に位置させることを可能にするために、クライアントコンピュータ内で使用される。さらに、カーソル制御装置は、ユーザが様々なコマンドを選択し、制御信号を入力することを可能にする。カーソル制御装置は、システムに制御信号を入力するための多数の信号生成装置を含む。典型的には、カーソル制御装置はマウスであってもよく、マウスのボタンは信号を生成するために使用される。あるいは、または追加的に、クライアントコンピュータシステムは、感知パッドおよび／または感知スクリーンを備えてもよい。
【００２３】
コンピュータプログラムは、コンピュータによって実行可能な命令を含んでいてもよく、命令は、上記システムに本方法を実行させるための手段を含む。プログラムは、システムのメモリを含む任意のデータ記憶媒体に記録可能であってもよい。プログラムは、例え

ば、デジタル電子回路、またはコンピュータハードウェア、ファームウェア、ソフトウェア、またはそれらの組み合わせで実装されてもよい。プログラムは、例えばプログラマブルプロセッサによる実行のための機械読み取り可能な記憶装置に具体的に実現された製品のような装置として実装されてもよい。方法ステップは、プログラム可能なプロセッサが命令のプログラムを実行し、入力データを操作して出力を生成することによって方法の機能を実行することによって実行されてもよい。したがって、プロセッサは、データ記憶システム、少なくとも１つの入力デバイス、および少なくとも１つの出力デバイスからデータおよび命令を受信し、また、それらにデータおよび命令を送信するようにプログラム可能であってもよく、またそのように接続されていてもよい。アプリケーションプログラムは、高水準の手続き型またはオブジェクト指向のプログラミング言語で、または必要に応じてアセンブリ言語または機械語で実装されていてもよい。いずれの場合も、言語はコンパイラ型言語またはインタープリタ型言語であってもよい。プログラムは、フルインストールプログラムまたは更新プログラムであってもよい。いずれの場合も、プログラムをシステムに適用すると、本方法を実行するための命令が得られる。

【００２４】

本方法は、画像の領域のキャプション付けを行うよう構成された関数を学習するためのものである。

画像は、例えばシーン上の、物理的信号の空間分布を表すデータ構造である。空間分布は、任意の次元のものであってよく、例えば２Ｄあるいは３Ｄである。空間分布は、例えばグリッドを形成し、それによってピクセルを定義するなど、任意の形状であってもよく、グリッドは場合により非規則的または規則的である。物理的信号は、画像がＲＧＢ画像またはグレースケール画像となるような、例えば色やグレーレベルなど、任意の信号であってもよい。画像は合成画像であってもよいし、あるいは写真のような自然画像であってもよい。データセットの画像、および／または、関数が適用されることが考えられる画像は、例えばすべてが矩形の２ＤのＲＧＢ画像、あるいはグレースケール画像であるなど、すべて同じタイプであってもよい。あるいは、異なる画像タイプの集合を考えてもよい。

【００２５】

画像の領域とは、画像の任意の部分である。したがって、領域は画像である。当該部分は、コネックス（ｃｏｎｎｅｘ）状、かつ／あるいは、凸状であってもよい。当該部分は矩形であってもよい。関数を適用することが企図されている、データセットの領域、および／または、入力領域は、例えばすべてが矩形状であるなど、すべて同じ形状であってもよい。あるいは、異なる領域形状の集合を考えてもよい。

【００２６】

画像のキャプションとは、画像のコンテンツのテキスト表現である。キャプションは、そのような画像のコンテンツを説明するテキスト表現あるいは文を含むか、またはそれからなっていてもよい。本方法によって学習した関数は、具体的には、入力画像と、当該入力画像の入力領域とに基づいて出力キャプションを生成するように適合されている。言い換えれば、関数は、入力画像と当該入力画像の入力領域とに適用され、さらに言い換えれば、関数は、画像と当該画像の領域とを入力とする。関数は、次いで、入力領域のコンテンツを説明するキャプションを出力し、この出力は、入力画像によって提供される領域のコンテクストに少なくともある程度依存する。

【００２７】

「学習する」とは、出力に関連付けられた入力のデータセットを提供することと、次いで、結果として得られる関数（すなわち学習された関数、つまり、最終的な重みに対応する関数）が所定の基準に応じて当該データセットに最もよく合致するように、重み付き関数（「ニューラルネットワーク」とも呼ぶ）の可変の重みを調整することとからなる機械学習工程を本方法が実施することを意味する。調整は、任意の既知の方法で行ってもよく、例えば、データセットの入力に、重み付き関数を適用し、その結果を、データセット内のこれらの入力に関連付けられた出力と比較することによって評価される、再構成損失を最小化することによって行う。

【００２８】
ここで、本方法の場合、重み付き関数は、画像および画像の領域を含む入力の対に適用して出力キャプションを生成するように設計される。言い換えれば、学習された関数のアーキテクチャは、（従来技術のように単一の入力画像ではなく）入力画像と当該入力画像の入力領域の、両者に適用されるように予め設定される。これに対応して、データセットは、それぞれ、画像、画像の領域、および領域のキャプションを含む３つ組（すなわち順序付けられた３つのデータの集合）からなる。言い換えれば、データセットは、一方の入力画像とその入力領域とを、他方の対応するキャプションに関連付ける。このようなデータセットは、任意の方法で提供すればよく、例えば、１人または複数のユーザが画像領域のキャプションを手動で生成し、かつ／あるいは、データストアから取得することによって得られる。上述のとおり、こうしたデータセットは、そのようなものとして既に存在する。
【００２９】
ここで、本方法の実施例について図２～図１３を参照して説明する。以下に説明する例では、関数のアーキテクチャのオプション的側面と、学習を実行するためのオプション的側面を示しており、このような側面は組み合わせることが可能である。
【００３０】
これらの例において、本方法は、画像内の領域の説明のための改良されたアプローチを構成し、その主な焦点は、画像の署名をその周囲を用いてコンテクスト化することにある。以下に示されるように、本方法は、画像特徴抽出および自然言語処理（ｎａｔｕｒａｌ ｌａｎｇｕａｇｅ ｐｒｏｃｅｓｓｉｎｇ：ＮＬＰ）において顕著なパフォーマンスを示し得るディープラーニング技術に大きく依存していてもよい。これらの例の利点は、エンドツーエンドの学習アーキテクチャ、入力領域画像のサイズに制約がないこと、および特徴のコンテクスト化による改善された説明を含んでいてもよい。
【００３１】
一例において、関数が入力画像と入力領域との組み合わせの署名を決定するように構成された第１の成分Ｃ１と、当該組み合わせの署名に基づき出力キャプションを生成するよう構成された第２の成分Ｃ２とを含むように、関数のアーキテクチャが制約されていてもよい。「成分」という用語は、単に、関数を形成するために他の任意の副関数（群）と合成可能な副関数を指している。したがって、ニューラルネットワークの成分もまた、ニューラルネットワークである。言い換えれば、関数は、Ｃ１と、Ｃ１の出力に適用されるＣ２との合成物を含む。「組み合わせ」という用語は、初期情報について、当該初期情報から導き出された別の情報を指す。組み合わせの署名とは、当該組み合わせを識別して同じタイプの他の情報から区別するベクトルである。したがって、Ｃ１によって決定された署名は、画像領域のコンテンツだけではなく、画像自体のコンテンツも考慮して画像領域のコンテンツを識別する。結果として、２つの異なる画像の領域を形成する同一コンテンツが、異なる署名に関連付けられてもよい。逆に、同一の画像の、異なるコンテンツを有する異なる領域が、異なる署名に関連付けられてもよい。
【００３２】
一例において、Ｃ１は、入力画像の署名を抽出するように構成された成分Ｃ１１（すなわち、他のあらゆる情報から独立している）と、入力領域の署名を抽出するように構成された成分Ｃ１１’（すなわち、他のあらゆる情報から独立している）と、入力画像の署名を入力領域の署名と組み合わせるよう構成された成分Ｃ１２とを含む。このような場合、組み合わせの署名は署名の組み合わせである。
【００３３】
画像の署名を抽出するように構成された任意の成分が、Ｃ１１、および／または、Ｃ１１’に実装されてもよい。Ｃ１１とＣ１１’は等しくてもよい（すなわち、共に同じ工程を適用してもよい）。Ｃ１１および／またはＣ１１’は、それぞれ畳み込みニューラルネットワーク（ＣＮＮ）であってもよい。そのような署名抽出器は、画像、特に規則的なグリッドを形成する画像に適用されるとき、良好で速い結果を提供することが知られている。さ

らに後述する例では、各ＣＮＮが重みを共有してもよい（すなわち、学習は、Ｃ１１およびＣ１１’を構成するＣＮＮが同じアーキテクチャおよび同一の重みを有するように制約され、Ｃ１１とＣ１１’が単に同じ単一のニューラルネットワークのインスタンスであってもよく、これはＣ１１＝Ｃ１１’と書き表せる）。この重み共有は、結果の品質を向上させる。

【００３４】

Ｃ１２は、情報について任意のタイプの組み合わせを実行することができる。例においては、Ｃ１２は、連結成分または加算成分、すなわち、Ｃ１１およびＣ１１’によって出力された署名、言い換えれば、領域および画像の署名を、連結する（例えば、所定の順序で）成分、または加算（すなわち、ベクトル加算、すなわち座標の次元ごとの加算）を行う成分を含んでいてもよい。例において、Ｃ１２は、全結合層をさらに含んでいてもよい。このようなＣ１２の例は、良質な結果をもたらす。

【００３５】

連結成分および全結合層を含むＣ１２の一例は、以下のようなものであってもよい。画像全体の署名x_iと領域の署名x_rとを連結してベクトル（x_i，x_r）とし、次いで全結合層にこれを通し、出力として$y=\sigma(W(x_i, x_r)+b)$を得る。ここでσは非線形性、Ｗは重み行列、ｂはバイアスである。

【００３６】

一例において、学習は、Ｃ１を学習することと、次いでＣ２の少なくとも一部（例えば、Ｃ１の出力に適用する部分）を学習することとを含んでいてもよい。Ｃ１（およびＣ２の少なくとも一部）の学習は、関数の他の成分とは独立して実行される。Ｃ１は、初期データセットの画像および領域を用いて学習されてもよい（例えば、初期データセットのすべてのキャプションを無視する）。Ｃ１は、学習されると、初期データセットの画像および領域の少なくとも一部に適用されて、それぞれがキャプションに（初期データセットのキャプションに応じて）関連付けられた署名の、新しいデータセット（画像および領域の組み合わせのそれぞれ）を作成してもよい。次いで、Ｃ２の少なくとも一部を、この新しいデータセットに基づいて学習してもよい。Ｃ２が学習を必要とする他の部分をも有する場合、そのような学習は、Ｃ１の学習の前、同時（すなわち並行して）、または後の、任意の時点に、かつ／あるいは、初期データセットまたは他のデータセットに基づいて行ってもよい。言い換えれば、学習は分割される。これにより、学習のスピードが向上し、また、良質な結果をもたらす。

【００３７】

これは、Ｃ１の学習の後に学習されたＣ２の部分が、組み合わせの署名を入力とし出力キャプションの反復生成を行う再帰型ニューラルネットワークを含む場合に特に当てはまる。再帰型ニューラルネットワークは、ニューラルネットワークにおける周知のカテゴリである。本方法の場合、再帰型ニューラルネットワークは、キャプションを反復的に生成する際に特に効率的であることがわかる。そのような場合、学習は、Ｃ１を（例えばそのＣＮＮを、例えばそれらが重みを共有するときは一緒に）訓練することと、次いで、そのような再帰型ニューラルネットワークのみを訓練することとを含んでいてもよい。この再帰型ニューラルネットワークは、以下の例で説明する「第２の再帰型ニューラルネットワーク」である。

【００３８】

例において、第２の成分は、組み合わせの署名と、さらには、単語埋め込み空間に基づいて、出力キャプションを生成するように構成されている。単語埋め込み空間は、ベクトル空間であって、そこでは単語がいわゆる単語埋め込みベクトルに対応する。第２の成分は、そのような単語埋め込み空間を用いて出力キャプションを生成してもよい。

【００３９】

例えば、第２の成分は、第１の再帰型ニューラルネットワークと、第２の再帰型ニューラルネットワークと、単語埋め込み層とを含んでいてもよい。第２の成分のこれらの副成分は、図２に示すように、相互作用してもよい。図２は、この構成を満たす本方法によっ

て学習可能な関数の例を示す。
【００４０】
図２の例では、第１および第２の再帰型ニューラルネットワークは、それぞれＬＳＴＭ１およびＬＳＴＭ２と表された長短期記憶（Ｌｏｎｇ Ｓｈｏｒｔ－Ｔｅｒｍ Ｍｅｍｏｒｙ：ＬＳＴＭ）ニューラルネットワークであるが、図２に関する以下の説明は、他のタイプの再帰型ニューラルネットワークにも当てはまる。
【００４１】
また、図２の例では、関数は、上述のように、第１の成分Ｃ１と第２の成分Ｃ２とを含む。Ｃ１は、上述の例で説明したように、Ｃ１が、入力画像２１０の署名を抽出するように構成された成分Ｃ１１と、入力領域２２０の署名を抽出するように構成された成分Ｃ１１’と、入力画像の署名を入力領域の署名と組み合わせるように構成された成分Ｃ１２とを含むことにより、入力画像２１０と当該画像２１０の入力領域２２０との組み合わせの署名を判定するように構成されている。しかしながら、図２に関する以下の説明は、入力画像２１０と入力領域２２０との組み合わせの署名を決定するように構成された他の任意の成分にも当てはまる。この例では、成分Ｃ１１およびＣ１１’はＣＮＮであるが、図２に関する以下の説明は、他の署名抽出成分に同様に当てはまる。成分Ｃ１２は、Ｃ１１およびＣ１１’から入力された署名を連結し、全結合（ＦＣ）層をさらに含むが、図２についての以下の説明は、他のタイプの署名組み合わせ成分にも同様に当てはまる。
【００４２】
ＬＳＴＭ１は、各入力単語埋め込みベクトル２５０に基づいて、出力単語埋め込みベクトル２６０を再帰的に生成するように構成されている。ＬＳＴＭ１についての入力単語埋め込みベクトル２５０は、ＬＳＴＭ２とそれに続く単語埋め込み層との合成物の出力２５０である。言い換えれば、ＬＳＴＭ１は、毎回、過去に生成されたＬＳＴＭ１（反復ループ２８０によって表される）と、ＬＳＴＭ２を入力２３０および２６０に適用し、次いで単語埋め込み層をＬＳＴＭ２の出力２９０に適用することからなる合成関数によって提供される入力単語埋め込みベクトル２５０に基づいて、出力単語埋め込みベクトル２６０を反復的に生成する。
【００４３】
ＬＳＴＭ２は、入力単語埋め込みベクトル２６０と、組み合わせの署名２３０とに基づいて、出力単語埋め込みベクトル２９０を再帰的に生成するように構成されている。ＬＳＴＭ２についての入力単語埋め込みベクトル２６０は、ＬＳＴＭ１の出力２６０である。言い換えれば、ＬＳＴＭ２は、毎回、過去に生成されたＬＳＴＭ２（反復ループ２７０によって表される）と、入力単語埋め込みベクトル２６０と、署名２３０（よってこれは毎回再利用される）とに基づいて、出力単語埋め込みベクトル２９０を反復的に生成する。
【００４４】
ここで、単語埋め込み層は、各入力単語埋め込みベクトル２９０に基づいて、出力単語埋め込みベクトル２５０を生成するように構成されている。単語埋め込み層についての生成された単語埋め込みベクトル２５０は、単語埋め込み空間に表される語彙のうち（入力２９０に応じて）最も確率が高い単語に対応する単語埋め込みベクトルである。
【００４５】
語彙とは、キャプションのような整った表現を生成できる単語の集合である。語彙は単一の言語に属していてもよい。語彙は、キャプションの集合に現れる単語の集合、例えば、初期データセットのキャプションの全部または一部に現れる単語の全部または一部（例えば、発生数の基準に基づいて決定される）であってもよい。語彙という用語は、そのような単語の集合を表す（単語埋め込み層についての）単語埋め込みベクトルの集合を指してもよい。語彙は単語埋め込み空間で表され、語彙の各単語は単語埋め込み層によって既知の各単語埋め込みベクトルに関連付けられ、そのような単語埋め込みベクトルの集合はそれにより予め決定されている。
【００４６】
単語埋め込み層は、そのような情報に基づき、入力２９０に応じて最も高い確率を有す

る集合の要素である、生成された単語埋め込みベクトル２５０を出力する成分である。確率処理は、自然言語処理における古典的な単語埋め込み層において知られている。例えば、単語埋め込み層は、語彙の各単語の生成の確率を出力し、次いで、出力された確率に応じて最も高い確率を有する単語を生成するように構成されていてもよい。

【００４７】

単語埋め込み層によって出力された各単語埋め込みベクトルは語彙の単語に対応し、単語埋め込み層はそれによって語彙の単語を順次決定し、単語のシーケンスはキャプションを形成する。ＬＳＴＭ１に入力するための単語埋め込みベクトル２５０を出力するのと並行して、単語埋め込み層は、各単語が決定される度に対応するキャプションを単語毎に出力するか、またはシーケンス全体が決定された後にまとめて出力することもできる。

【００４８】

単語埋め込み層の一例について、このよく知られたニューラルネットワークを説明するために議論する。単語埋め込み空間における単語をw_1，・・・，w_Nと表し、各ベクトルの次元がｄ＝１２８であり、次元ｄ＝１２８の出力ｈが与えられたとき、単語埋め込み層の出力は、$p_k = exp(<w_k, h>) / \Sigma_i exp(<w_i, h>)$である確率のベクトル$p_1$，・・・，$p_N$であってもよい。訓練中、単語埋め込みにおける異なる単語の表現（すなわち、パラメータw_1，・・・，w_N）が学習される。

【００４９】

図２の例では、一方のＬＳＴＭ１と、他方のＬＳＴＭ２とそれに続く単語埋め込み層との合成物とが、相互作用し、それらの再帰的反復がインターレースされる。言語処理の分野で知られているように、繰り返しの最初については、反復工程の開始時に存在するＬＳＴＭ１およびＬＳＴＭ２のうちの一方（例えばＬＳＴＭ１）は、入力２５０それ自体の代わりに、繰り返しの最初におけるこの開始状況を示す、「文頭（Ｂｅｇｉｎｎｉｎｇ ｏｆ ｓｅｎｔｅｎｃｅ）」あるいは＜ＢＯＳ＞と呼ばれる定数を使用する（なぜなら、この時点では、他方のＬＳＴＭによって提供される出力が利用できないため）。また、この反復工程は、ＬＳＴＭ２が、「文末（ｅｎｄ ｏｆ ｓｅｎｔｅｎｃｅ）」あるいは＜ＥＯＳ＞と呼ばれる、終了を示す定数を出力したときに終了する。

【００５０】

図２によって表される関数は、任意の方法で学習されてもよい。例において、学習は、語彙における各単語の単語埋め込みベクトルを学習することを含んでいてもよい。これらのベクトルは、単語埋め込み層のパラメータである。学習は、キャプションの集合に基づいてＬＳＴＭ１と共に単語埋め込み層を学習することによって実行してもよい。あるいは、単語埋め込み層の学習は、ＬＳＴＭ１の学習とは別に実行してもよく、両者の学習は、キャプションの集合に基づいて実行される。すべての場合において、キャプションの集合は、初期データセットによって提供されるキャプションの一部またはすべてであってもよい。本方法は、そのような、キャプションのデータセットを提供することを含んでいてもよく、学習することは、次いで、キャプションのデータセットを用いてＬＳＴＭ１、および／または、単語埋め込み層を訓練することと、次いで（それぞれが画像だけでなく署名にも関連付けられたキャプションの集合に基づき）ＬＳＴＭ２のみを訓練することとを含む。言い換えれば、ＬＳＴＭ１および／または単語埋め込み層は、ＬＳＴＭ２に対して事前に学習されていてもよい。ＬＳＴＭ１および／または単語埋め込み層の訓練は、ＬＳＴＭ２の訓練の前に、および／またはＣ１の成分（例えば、ＣＮＮであるＣ１１およびＣ１１’および／またはＦＣ層）の訓練の前または同時に実行されてもよい。ＬＳＴＭ２は、Ｃ１の成分（例えば、ＣＮＮであるＣ１１およびＣ１１’および／またはＦＣ層）の訓練の後に訓練されてもよい。

【００５１】

一例において、本方法によって学習される関数は、シーン理解および／または密なキャプション付け処理に含まれていてもよい。密なキャプション付けは、メディア内の注目領域を自動的に検出し、それら領域を文で説明するのに用いられる、すべての技術を集めたものである。本方法は、密なキャプション付けの説明部分に用いてもよく、したがって、

メディア内に提供される領域は、ユーザによって選択されるか、または外部の検出アルゴリズムから取得されると仮定する。

【００５２】

画像キャプション付けに用いられる他の方法と同様、本方法は２つの分野の交差領域におけるものである。

・メディア表現のための特徴抽出：ユーザが提供するメディアは、そこから単語のシーケンスを生成するモデルにとっては、それ自体理解できなくてもよい。そのメディアから、特徴抽出処理によって署名を取得する必要があってもよい。そのような特徴は、記述子から得られた、設計された特徴であってもよいし、機械学習モデルによって学習されてもよい。ディープラーニング、特に画像や動画のようなメディア分析のための畳み込みニューラルネットワークにおける近年の進歩は、特徴学習において良好な結果をもたらす。

・自然言語処理：人間の言語を理解できるアルゴリズムの設計において、多くの進歩が達成されている。これらのモデルのポイントは、人間の言語を理解し、また、関連する文を生成できることである。これは近年まで、文のパーサーを統合し、文の意味表現および依存構造を取得する（すなわち、異なる品詞タグ付けとそれらの関係を区別する）モデルを構築することにより、多くのエンジニアリングによって実現していた。近年のディープラーニングの進歩により、言語の規則を学習することができるディープラーニングモデルを訓練することによって、この工程をすべて自動的に実行することができる。そのようなモデルは、特定の言語の文のデータセットで訓練するだけで、その言語で文を理解し生成することができるようになる。

【００５３】

一例において、本方法は、上述した最近のディープラーニングの進歩を効率的に活用する。

密なキャプション付けの分野に関連する方法同様、本方法は同種のパターンに従うフレームワークの範疇に入る：集中的な演算が実行され得るオフライン段階と、ユーザの待ち時間を最小限に抑えるため、パフォーマンスが重要なオンライン段階とを有する。

【００５４】

オンライン段階は、２つのステップに分けてもよい：

・まず、ユーザによって提供されたメディアが分析されてもよく、第１のプロセスがこのメディアに固有の識別子を出力する。この識別子は、当該技術分野で「署名」または「特徴」と呼ばれる。パフォーマンスのため、署名は通常、メディアのより小さい寸法による表現である。例えば、歌の検索においては、署名は、記録された曲のパワースペクトルにおける最大の高調波の集合であってもよい。

・次いで、署名は、そこから文を生成する第２の工程に転送される。これは文を生成するために訓練された再帰型ニューラルネットワークによって実行される。オンライン段階では、文の生成には、モデルへの入力として与えられるメディアの署名によって、バイアスがかかる。このように生成された文は、メディアのコンテンツに密接に関連し、メディアの説明として解釈される。

【００５５】

密なキャプション付けでは、モデルによって生成された文の品質と関連性は、訓練中に用いられたデータベースの品質に大きく依存する可能性がある。

オフライン段階は、画像と、画像内の領域と、領域の説明とからなる、提供された３つ組のデータセットに基づく、２段階の工程とみなすことができる。

【００５６】

・文の言語に基づき、言語モデルを訓練してもよい。そのために、３つ組のデータベースからのすべての文からなるデータベースを集約してもよい。次に、文をトークン化し、データベース内で頻繁に出現する単語のみを保持することによって、データベースを前処理してもよい：これが学習の語彙となる。言語モデルは、この前処理されたデータベースで訓練され、文を生成することを学習してもよい。文を生成するための学習の工程は、基本的に、文において、直前の単語の後ろに最も可能性の高い単語を生成するようにモデル

を教えることからなっていてもよい。このように、文を一語一語生成することができるモデルを得ることができる。

【００５７】

・メディア用の特徴抽出器を訓練してもよい。この抽出器は、説明を生成するために、メディアから特徴を構築し、その特徴を言語モデルへの入力としてもよい。

一例においては、関数の主な目的は、画像内の領域の説明を言語的な文の形式で与えることである。画像が低次元コードに符号化され、次いで、言語モデルによって復号されて単語のシーケンスを生成するエンコーダ／デコーダフレームワークを用いてもよい。本方法は、エンコーダにおいて２Ｄ特徴抽出のため、効率的な畳み込みニューラルネットワークを活用してもよい。デコーダについては、本方法は、シーケンス生成のため、効率的な再帰型ニューラルネットワークを活用してもよい。

【００５８】

ここで、上述および／または後述の概念について説明する。

密なキャプション付けは、画像内の領域を、人間に理解可能な文（単語のシーケンス）によって説明するという課題を含む。これは、まず領域に関連する署名を抽出し、次いで、その署名から単語シーケンスを生成して対象となる領域を説明する、モデルのパラメータを学習することからなる。

【００５９】

ディープニューラルネットワーク（ＤＮＮ）は、コンピュータが観測データから学習することを可能にする、生物学に着想を得たプログラミングパラダイムであるニューラルネットワーク（論文「Ｒｕｍｅｌｈａｒｔ ｅｔ ａｌ．Ｌｅａｒｎｉｎｇ ｉｎｔｅｒｎａｌ ｒｅｐｒｅｓｅｎｔａｔｉｏｎｓ ｂｙ ｅｒｒｏｒ ｂａｃｋｐｒｏｐａｇａｔｉｏｎ，１９８６」に記載）における学習のための強力な技術の集合である。

【００６０】

物体の認識において、ＤＮＮの成功は、他の画像分類法（ＳＶＭ、Ｂｏｏｓｔｉｎｇ、Ｒａｎｄｏｍ Ｆｏｒｅｓｔなど）で用いられる手作業による低レベルの特徴（Ｚｅｒｎｉｋｅモーメント、ＨＯＧ、Ｂａｇ－ｏｆ－Ｗｏｒｄｓ、ＳＩＦＴなど）とは対照的に、たくさんの中間レベルの２Ｄ画像表現を学習する能力を有するおかげである。より具体的には、ＤＮＮは、未処理のデータに基づくエンドツーエンドの学習に焦点を当てている。言い換えれば、未処理の特徴から始まりラベルで終わるエンドツーエンドの最適化を達成することによって、特徴量エンジニアリングから最大限まで遠く離れる。これは、ディープニューラルネットワークの一例を示す図３に示されている。

畳み込みニューラルネットワーク（論文「ＬｅＣｕｎ ｅｔ ａｌ．Ｃｏｎｖｏｌｕｔｉｏｎａｌ Ｎｅｔｗｏｒｋｓ ｆｏｒ Ｉｍａｇｅｓ，Ｓｐｅｅｃｈ，ａｎｄ Ｔｉｍｅ－Ｓｅｒｉｅｓ」に記載）は、下位層のニューロンが畳み込みフィルタに置き換えられた、ディープニューラルネットワークの特別なケースである。これは、畳み込みフィルタの一例を示す図４に示されている。これらのフィルタは入力のどこにでも適用され、出力として特徴マップが与えられる。この特徴マップは、入力の特定のパターンがフィルタによって認識された、活性化領域を示す。いくつかの畳み込み層を積み重ねるときのディープラーニングの利点は、記述子によって得られる基本的な特徴よりも洗練された、非常に複雑だが強力な特徴を抽出する方法を提供することである。

【００６１】

再帰型ニューラルネットワーク（論文「Ｇｒａｖｅｓ，Ｓｕｐｅｒｖｉｓｅｄ Ｓｅｑｕｅｎｃｅ Ｌａｂｅｌｌｉｎｇ ｗｉｔｈ Ｒｅｃｕｒｒｅｎｔ Ｎｅｕｒａｌ Ｎｅｔｗｏｒｋｓ」に記載）は、時系列処理において興味深い性能を示した特別なタイプのニューラルネットワークである。その成功は、予測を行う際に以前の状態のフィードバックを統合する能力に起因する。これは、ニューロン定義における時間的ループにより実行される。ＲＮＮに関する最近の研究により、ゲートメカニズムがその定義に統合された。これらのモデル（長短期記憶（Ｌｏｎｇ－Ｓｈｏｒｔ Ｔｅｒｍ Ｍｅｍｏｒｙ：ＬＳＴＭ））、ゲート付き再帰型ユニット（Ｇａｔｅｄ Ｒｅｃｕｒｒｅｎｔ Ｕｎｉｔ：ＧＲＵ））は

、メモリの統合と、消失勾配問題に対処する能力とにより、最先端のパフォーマンスを向上させた。

【００６２】

一例において、フレームワークは２つの段階に分解されていてもよい。

第１段階（上述のように「オフライン」段階と呼ぶことができる）は、ディープニューラルネットワーク、特にメディア内の特徴を抽出するための畳み込みニューラルネットワーク、および、それらの特徴から文を生成するための再帰型ニューラルネットワークに大きく依存していてもよい。「オフライン」という用語は、この段階が本方法のユーザにとって透過的であり、大規模な演算が必要な場合であっても演算を行う時間をとることができるという事実を指す。この部分では、これらのネットワークのパラメータが学習されてもよく、これがすべてについて実行されてもよい。学習工程は、２つのステップで実行されてもよい：第１に、言語モデルを得るために再帰型ニューラルネットワークのパラメータが学習され、第２に、どのようにメディアのコンテクストの特徴を抽出し、再帰型ニューラルネットワークの学習を特徴抽出工程に適合させるかを理解するために、畳み込みニューラルネットワークのパラメータが、再帰型ニューラルネットワークと共に学習されてもよい。

【００６３】

第２段階（上述のように「オンライン」段階と呼ぶことができる）は、画像の領域から文を生成する工程中にリアルタイムに実行されるすべてのステップを集約する。

【００６４】

図５～図６は、それぞれ、オフラインおよびオンライン段階の技術的ワークフローの一例を示す。

【００６５】

ここで、言語モデリングの例について説明する。

領域キャプション付け工程の第１の柱は、言語モデリングであってもよい。言語は、画像の領域を説明するキャプションから学習してもよい。これは、生成されたキャプションが正しく、人間によって理解可能であることを確実にするために、可能な限り高い精度で学習してもよい。領域キャプション付け処理の全体的なアーキテクチャはエンドツーエンドのアプローチで学習することができるが、言語モデルだけを学習し、次いでパラメータを微調整して全体的なモデルにすることで、パフォーマンスが向上することがわかった。

【００６６】

最後に、言語モデルの学習は２つのパートで実行してもよい：

１．データセットからのキャプションの前処理

２．生成モードにおける再帰型ニューラルネットワークの訓練

【００６７】

ここで、データセットからのキャプションの前処理の例について説明する。

データセットからのキャプションの前処理は、言語モデルの良好な学習を可能にし得る。

元のデータセットでは、文は文字列の形式であってもよい。トークナイザは、最初にこれらの文をトークン化してもよい。すなわち、異なる語を分離し、それらを小文字にしてもよい。実際、コンピュータの観点では、文字列“Ｃａｒ”と“ｃａｒ”は同じではない可能性がある。さらに、句読点や特別な単語／頭字語を扱うように構成されている可能性があるため、トークナイザを設計するときには多くの作業が行われる可能性がある。例えば、トークナイザは“Ｕ．Ｓ．”と“ＵＳ”を同じトークンとみなす必要がある可能性がある。

【００６８】

文がトークン化されると、言語のための語彙を設計することができる。これは、すべての単語がモデルによって学習されるわけではなく、データセット内の最大頻度の単語のみが学習されることを意味してもよい。実際、モデルが、頻繁には出現しない単語の意味を比較的高い関連性では学習しない方が容易である。

【００６９】

ここで、データセットからのキャプションに関する言語モデルの学習の例について説明

する。
言語モデルの訓練は、再帰型ニューラルネットワークを生成モデルとして考えることによって行ってもよい。
【００７０】
実施例では、長短期記憶（Ｌｏｎｇ Ｓｈｏｒｔ－Ｔｅｒｍ Ｍｅｍｏｒｙ：ＬＳＴＭ）（例えば、論文「Ｓ．Ｈｏｃｈｒｅｉｔｅｒ ｅｔ ａｌ．Ｌｏｎｇ ｓｈｏｒｔ－ｔｅｒｍ ｍｅｍｏｒｙ」に記載）を利用してもよい。ＬＳＴＭは、明示的なメモリ、および予測を行う際にメモリと相互作用するためのゲート機構を有する特定の種類の再帰型ニューラルネットワークである。これらは、時間的シーケンスを示す様々な領域、特に自然言語処理において、効率が良いことが実証された。
【００７１】
ＬＳＴＭユニットにはいくつかの実施例が存在し、ユニット内部の覗き穴結合（ｐｅｅｐｈｏｌｅ ｃｏｎｎｅｃｔｉｏｎｓ）や、ゲートの数が異なり得る。例えば、ゲート付き再帰型ユニット（Ｇａｔｅｄ Ｒｅｃｕｒｒｅｎｔ Ｕｎｉｔ：ＧＲＵ）は、ゲートが2つしかない特定の実施例である。特に効率的な実施例では、本方法は、むしろ、論文「Ｓ．Ｈｏｃｈｒｅｉｔｅｒ ｅｔ ａｌ．Ｌｏｎｇ ｓｈｏｒｔ－ｔｅｒｍ ｍｅｍｏｒｙ」に記載の元のＬＳＴＭの実施を考慮してもよい。本方法は、ＬＳＴＭの出力次元を、言語の単語埋め込み空間の次元に設定してもよい。
【００７２】
以下、単語埋め込みの概念について説明する。
自然言語処理では、ＬＳＴＭにワン・ホット・ベクトル（すなわち、対象となる単語のインデックス以外においてすべてが０の、語彙のサイズのベクトル）で表される単語を扱わせるのは、時間がかかり過ぎ、演算上、効率的ではない。実際、その場合、ベクトルのサイズは語彙のサイズになり、ＬＳＴＭの定義にあまりにも多くのパラメータが必要になる。その代わりに、本方法で行い得ることは、ある埋め込み空間に単語を埋め込むことである。そのような空間は、単語の意味を反映し、その意味によってそれらをクラスタ化するため、興味深い特性を有する。単語の埋め込みという概念は、特にＧｏｏｇｌｅが$Ｗｏｒｄ２Ｖｅｃ^{TM}$を立ち上げたことにより、２０１３年に実際に関心を集めた。これは、スキップ・グラムと呼ばれるネガティブ・サンプリングを用いた具体的なアプローチにより$Ｇｏｏｇｌｅ Ｎｅｗｓ^{TM}$で学習したモデルである（論文「Ｔ．Ｍｉｋｏｌｏｖ ｅｔ ａｌ．Ｄｉｓｔｒｂｉｔｕｔｅｄ Ｒｅｐｒｅｓｅｎｔａｔｉｏｎｓ ｏｆ Ｗｏｒｄｓ ａｎｄ Ｐｈｒａｓｅｓ ａｎｄ ｔｈｅｉｒ Ｃｏｍｐｏｓｉｔｉｏｎａｌｉｔｙ」に記載）。ここで、本方法は、代わりにＬＳＴＭを用いた別のアプローチで単語埋め込みを学習してもよい。
【００７３】
ＬＳＴＭと組み合わせて単語埋め込みを行うと、ＬＳＴＭの文生成工程の解釈に役立つ可能性がある。単語を予測する工程は以下のとおりである。
・ＬＳＴＭの出力は、単語埋め込み空間内の点である。
・このとき、予測された単語は、その埋め込み表現が、予測された点に最も近い単語である。
・予測された単語の埋め込み表現は、次いで、次の単語を予測するためにＬＳＴＭの入力に供給される。
【００７４】
このように、本方法は、訓練中に、単語埋め込みと組み合わせたＬＳＴＭのパラメータを学習してもよい。これらのパラメータは、データセットからの多数のキャプションで学習してもよい。これは、この言語モデリング訓練という特定の処理のための訓練セットであってもよい。トークン化と語彙制限定義により前処理されている可能性があるため、キャプションに追加の制約は必要ない。
【００７５】
ここで、本関数によって行われるコンテクスト特徴抽出について説明する。

領域キャプション付けの第２の基本的な柱は、メディアからの特徴抽出であってもよい。この特徴は、ＬＳＴＭの文予測にバイアスを与え領域を説明する文を生成させるために、後に、ＬＳＴＭへの入力として与えられてもよい。ここで本方法が主に貢献する点は、メディア内の領域の周囲によってコンテクスト化された注目領域の特徴を考慮することであってもよい。このように、本方法は、メディア内の注目領域のコンテクスト的特徴を取得し、また、その領域の周囲の情報を考慮に入れるため、より多くの特徴が得られる。

【００７６】

ここで特徴ベクトルの抽出の例について説明する。

メディアの特徴抽出工程は、取り得るアプローチがいくつかある困難な処理である。ディープラーニングが導入される以前には、これはＳＩＦＴやＳＵＲＦのような、設計された特徴を与える記述子を用いて行われた。もともと、これらの特徴は画像の低レベルの署名を提供していた：記述子は、物体の輪郭のような低レベルの要素が検出された画像内の活性化領域を示す特徴マップを与える。

【００７７】

現在では、ディープラーニングの進歩により、画像から特徴を学習することが可能になっている。また、ニューラルネットワークに積み重ねることができる多数の層により、特徴の定義を非常に深くすることが可能になった。

【００７８】

一実施例において、本方法は、畳み込みニューラルネットワーク（例えば、論文「Ｙ．Ｌｅｃｕｎ ｅｔ ａｌ．Ｂａｃｋｐｒｏｐａｇａｔｉｏｎ ａｐｐｌｉｅｄ ｔｏ ｈａｎｄｗｒｉｔｔｅｎ ｚｉｐ ｃｏｄｅ ｒｅｃｏｇｎｉｔｉｏｎ，Ｎｅｕｒａｌ Ｃｏｍｐｕｔ，１９８９」に記載）を利用することができる。ＣＮＮは、一般的な行列乗算の代わりに畳み込みを使用するニューラルネットワークであり、物体を、雑然とした背景や複数の他の物体を有する画像内の複雑なシーンから、効率よく認識することが実証されている。

【００７９】

より具体的には、本方法は、ＩＬＳＶＲＣ２０１２の分類処理を獲得したＣＮＮを実装することができ、これは１０００個のカテゴリの中から画像に存在するオブジェクトのカテゴリを認識する処理である。これはＡｌｅｘＮｅｔと呼ばれ、論文「Ａｌｅｘ Ｋｒｉｚｈｅｖｓｋｙ，ＩｍａｇｅＮｅｔ Ｃｌａｓｓｉｆｉｃａｔｉｏｎ ｗｉｔｈ Ｄｅｅｐ Ｃｏｎｖｏｌｕｔｉｏｎａｌ Ｎｅｕｒａｌ Ｎｅｔｗｏｒｋｓ，ＮＩＰＳ ２０１２」に記載されている。しかしながら、本方法は、第８の全結合層（文献中のＦＣ８）から上の層を除去することによって、そのようなネットワークを目下の特定の処理に適応させることができる。なぜならその層は１０００個のカテゴリから物体を分類するという課題であるＩＬＳＶＲ Ｃｈａｌｌｅｎｇｅにおいて１０００次元となるように設計されたものであるためである。このように、最終的に、本方法が考慮し得る切り取られたバージョンのＡｌｅｘＮｅｔネットワークは、５つの畳み込みレイヤー（プーリングと正規化を含む）と２つの全結合層からなる。また、切り取られたＡｌｅｘＮｅｔの最後の層（すなわちＦＣ７）には４０９６個のニューロンが含まれているため、画像の特徴ベクトルは４０９６次元のベクトルとなる。

【００８０】

ここで、領域の特徴と周囲の特徴を組み合わせる例について説明する。

領域の署名を設計する際、注目領域内の情報のみを考慮すると、領域のコンテクストによって与えられる多くの情報が失われる。これは、画像内の領域のサイズが小さい場合に特に当てはまる：そのような場合、ＣＮＮは、あまりたくさんの特徴を捉えられない可能性がある。

領域の特徴の品質を改善するため、本方法では、それを、画像内の領域の周囲で抽出されたコンテクストの特徴と組み合わせてもよい。これは、領域の画像から第１のＣＮＮ（領域ＣＮＮとも呼ぶ）によって抽出された特徴と、その領域およびその周囲の領域を含む画像から第２のＣＮＮ（コンテクストＣＮＮとも呼ぶ）によって抽出された特徴とを組み

合わせることによって行われる。
【００８１】
異なる特徴を組み合わせる際、いくつかの方法を用いることができる。それらのうち、本方法は以下を実施してもよい：
・連結：領域の特徴ベクトルと周囲の領域の特徴ベクトルとを連結して、コンテクスト特徴ベクトルを表すベクトルとする。
・加算：領域の特徴ベクトルと周囲の領域の特徴ベクトルとを合計して、コンテクスト特徴ベクトルを表すベクトルとする。
・全結合層の利用：領域の特徴ベクトルと周囲の領域の特徴ベクトルとを連結して１つのベクトルとし、その連結ベクトルの上に全結合層を追加する。この結果、コンテクスト特徴ベクトルを表す別のベクトルが得られる。
【００８２】
図７は、ニューラルネットワークアーキテクチャレベルで特徴を組み合わせる、この工程を示す。
ここで、特徴抽出と言語モデリングの両方についてモデルをエンドツーエンドで訓練する方法の例について説明する。
ネットワークにおける特徴の抽出および組み合わせの訓練は、言語モデリング部分からのＬＳＴＭの訓練と共に実行されてもよい。これはすべてエンドツーエンドで行ってもよい。本方法は、言語モデリング部分だけで学習されたパラメータから、ＬＳＴＭのパラメータを微調整してもよい。したがって、訓練のこの部分では、入力として与えられた領域および周囲のコンテクストからキャプションを生成するために、モデル全体を学習してもよい。
したがって、ネットワークのパラメータは、内部に指定された領域およびそれらの領域を説明するキャプションを有する多数の画像で学習してもよい。そのような３つ組の集合（全体画像、領域画像、キャプション）は訓練データセットを表す。キャプションは、言語モデリングについて説明したのと同様に前処理されているため、ここでは画像のサイズやキャプションの特性を制約する必要はない可能性がある。
【００８３】
ここで、上述したすべての段階の例をまとめる方法の例を、オフライン段階の例と、対応するオンライン段階の例とをそれぞれ表す、図８～図９を参照して説明する。
【００８４】
最終的な工程は、コンテクスト特徴抽出工程と言語モデルを統合することであってもよい。モデル全体の学習は、エンドツーエンドの工程で行ってもよい。しかしながら、言語モデルは、キャプションのデータベースで最初に学習され、次いで、学習された重みがアーキテクチャのＬＳＴＭ部分に合わせて微調整されてもよい。これら２つの工程を組み合わせると、領域およびコンテクストから抽出された特徴の組み合わせのコンテクスト特徴ベクトルが、このベクトルによって与えられる情報による文の予測にバイアスをかけるために、言語モデルへの入力として与えられる。したがって、生成された文は、領域およびそのコンテクストによってもたらされる情報と強く相関し、そして、コンテクストによってもたらされる情報によって強化された、領域の説明を提供する。訓練の詳細のすべてが、図８に示されている。モデルの試験が、図９のオンライン段階に示されている。
【００８５】
図１０は、図２の例に沿ったオンライン段階の簡略化した例を概略的に示している。
Ｔ＝０（繰り返しの１回目）において、文頭を示す単語埋め込みベクトルｈ　＜ＢＯＳ＞がＬＳＴＭ１に入力され、キャプション付け工程が初期化される。ＬＳＴＭ１の出力はＬＳＴＭ２に入力として渡され、ＬＳＴＭ２には、別の入力として、組み合わせの署名２３０も提供される。次いで、ＬＳＴＭ２は、単語埋め込み層に入力された単語埋め込みベクトルを出力し、次に、単語埋め込み層が、語彙の各単語が生成される確率のベクトルを出力する（例を単純にするため、語彙は３つの単語、すなわち「＜ＥＯＳ＞」と、「バス（ｂｕｓ）」と、「駐車（ｐａｒｋｉｎｇ）」とに絞っている）。次いで、単語「バス」

が生成される。なぜなら、これが確率が最も高い単語であるからである。

【００８６】

Ｔ＝１（繰り返しの２回目）では、単語「バス」の単語埋め込みベクトルがＬＳＴＭ１に入力され、ＬＳＴＭ１は、それと、過去の出力（参照符号２８０で表される）に基づき、結果をＬＳＴＭ２に出力し、ＬＳＴＭ２自体は、それと、署名２３０と、過去（参照符号２７０で表される）に基づき、ベクトルを単語埋め込み層に出力し、単語埋め込み層が単語「駐車された（ｐａｒｋｅｄ）」を生成する。

Ｔ＝２（繰り返しの最後）では、同じ工程が実行され、単語埋め込み層の出力が、工程の終了を示す＜ＥＯＳ＞となることが観察される。

こうして、この時点において、図２の画像２１０の領域２２０に対応するキャプション「駐車されたバス（ｐａｒｋｅｄ　ｂｕｓ）」が関数によって生成されると本工程は終了する。

【００８７】

図１１～図１３は、それぞれ、領域３１０～領域３６０を有する画像を示す。以下の表Ｉ～表ＩＩＩは、それらの画像および領域について本方法の実施例により出力されたキャプション（一方は、署名の組み合わせに連結を用いたもの、他方は署名の組み合わせに加算を用いたもの）を、従来技術（ＬＲＣＮ、すなわちコンテクスト化されていない領域署名）の実施例によって出力されたキャプションと比較して示している。図から分かるように、本方法により生成されたキャプションは、より高品質である。

【００８８】

【表１】

	LRCN	コンテクストの特徴：連結	コンテクストの特徴：加算
310	白黒の犬	黒いジャケットを着た男性	黒いジャケットを着た男性
320	白黒の猫	赤い車	白黒の犬
330	タイル張りの床	道路上の白線	道路上の白線
340	葉のない木	背景にある木	青葉のついた木
350	木は裸である	葉のない木	背景にある大きな木
360	大きな窓	大きな建物	大きな時計がある建物

表Ｉ：図１１の画像の領域について生成したキャプション

【００８９】

【表２】

	LRCN	コンテクストの特徴：連結	コンテクストの特徴：加算
310	電車は黒い	歩道	道路上の白線
320	床はタイル張りである	道路上の白線	道路上の白線
330	建物側の壁	白黒の道路標識	白黒写真
340	白黒写真	車道上のバス	道路上のバス
350	大きな建物	大きな建物	大きな建物
360	線路上の電車	白い建物	大きな建物

表ＩＩ：図１２の画像の領域について生成したキャプション

【００９０】

【表３】

	LRCN	コンテクストの特徴： 連結	コンテクストの特徴： 加算
310	白いトイレ	道路は濡れている	建物側の壁
320	白黒写真	道路側に駐車された黒い車	車道上に駐車された車
330	大きな窓	大きな窓	建物の窓
340	葉のない木	青葉のついた木	青葉のついた木
350	背景にある木	背景にある緑の木々	青葉のついた木
360	大きな建物	大きなレンガ造りの建物	大きなレンガ造りの建物

表ＩＩＩ：図１３の画像の領域について生成したキャプション

【００９１】

表ＩＶは、異なる尺度に基づく本方法の異なる実施例（および従来技術ＬＲＣＮの実施例）間の比較を示しており、１つの尺度は前述したものである。

この関数は、データベース「Ｖｉｓｕａｌ　Ｇｅｎｏｍｅ」に基づき、サブ領域＞１００ｐｘに基づいて学習された（表に示されたすべてのモデルがこのデータベースで訓練されテストされた）。

【００９２】

異なるＮＬＰ評価尺度が設定され、生成されたキャプションの品質が、人間によって生成されたテストセットにおける参考キャプションとの比較で評価される。

Ｂｌｅｕ－ｎ（ｎ＝１・・・４）：これは、生成されたキャプションと参考キャプションの両方に現れるｎ－ｇｒａｍの割合である数量を算出する一般的なＮＬＰ評価尺度である。あまりにも短い文にはペナルティを科すペナルティ係数が、この数量に掛け合わされる。

ＭＥＴＥＯＲ：これは言語に固有の評価尺度である。候補文と生成された文とを、単語の同義性を感知する単語ごとのマッチングで比較する。

Ｒｏｕｇｅ－Ｌ：これは、候補文と生成された文とを、これら２つにおける最も長い共通サブシーケンスに着目することにより比較する尺度である。

ＣＩＤＥｒ：これは合意に基づく評価尺度であり、文における単語が原形に変換され、これら変換された文におけるｎ－ｇｒａｍの集合が比較される。

【００９３】

図から分かるように、本方法は、概して先行技術よりも性能が良い。また、画像および領域の署名を抽出するための重みを共有するＣＮＮ、次いで２つの署名の連結、そしてそれに次いで全結合層からなる前述の実施例によって特に良好な結果がもたらされることがわかる。

【００９４】

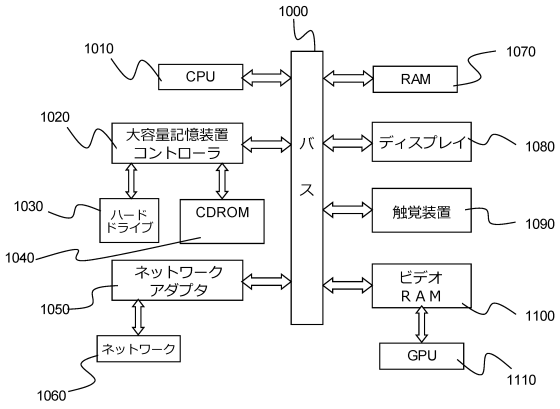
【表４】

尺度	LRCN-2f-fc7	コンテクストの特徴：連結	コンテクストの特徴：加算	コンテクストの特徴：連結＋ＦＣ	重み共有を用いた加算	重み共有を用いた連結＋ＦＣ
Bleu-1	0.187	0.205	0.218	0.238	0.218	0.244
Bleu-2	0.081	0.093	0.104	0.118	0.105	0.123
Bleu-3	0.044	0.051	0.058	0.068	0.059	0.071
Bleu-4	0.030	0.031	0.036	0.041	0.036	0.045
METEOR	0.069	0.082	0.088	0.104	0.091	0.104
ROUGEL	0.175	0.192	0.206	0.228	0.209	0.232
CIDEr	0.333	0.425	0.482	0.625	0.507	0.627

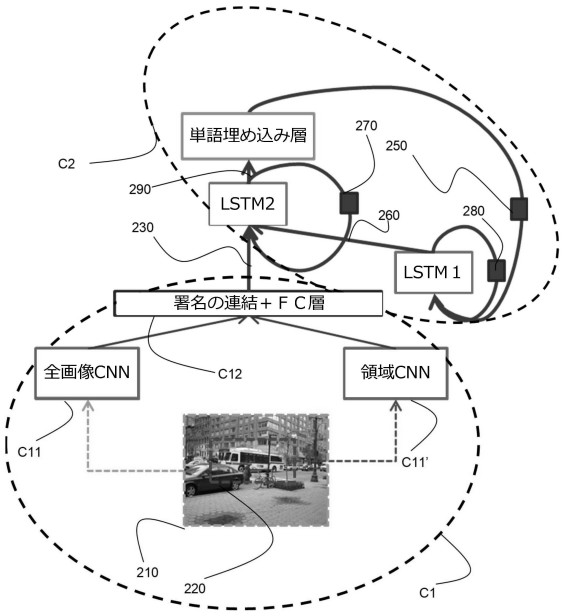
表ＩＶ－異なる実施例の比較

【図面】

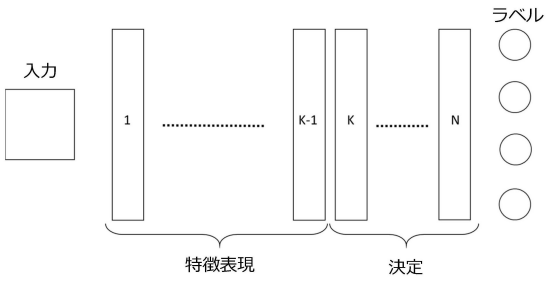
【図１】


1000
1010
CPU
RAM
1070
1020
大容量記憶装置
コントローラ
バス
ディスプレイ
1080
1030
ハード
ドライブ
CDROM
触覚装置
1090
1040
ネットワーク
アダプタ
ビデオ
ＲＡＭ
1050
1100
ネットワーク
GPU
1060
1110

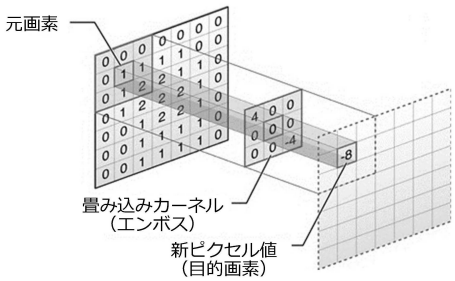
【図２】


単語埋め込み層
270
250
C2
290
LSTM2
260
280
230
LSTM１
署名の連結＋ＦＣ層
全画像CNN
C12
領域CNN
C11
C11'
210
220
C1

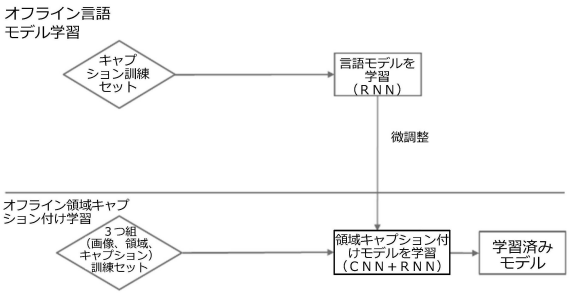
【図３】


ラベル
入力
1
K-1
K
N
特徴表現
決定

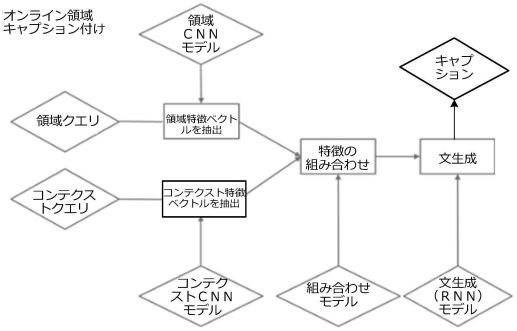
【図４】


元画素
畳み込みカーネル
（エンボス）
新ピクセル値
（目的画素）

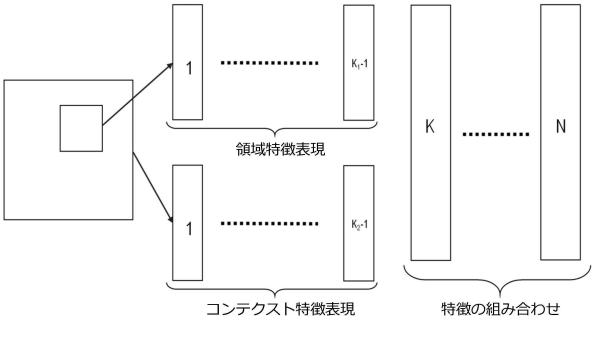
【図５】


オフライン言語
モデル学習
キャプション訓練セット
言語モデルを学習（ＲＮＮ）
微調整
オフライン領域キャプション付け学習
３つ組（画像、領域、キャプション）訓練セット
領域キャプション付けモデルを学習（ＣＮＮ＋ＲＮＮ）
学習済みモデル

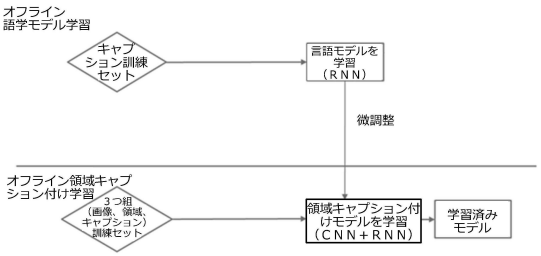
【図６】


オンライン領域
キャプション付け
領域ＣＮＮモデル
キャプション
領域クエリ
領域特徴ベクトルを抽出
特徴の組み合わせ
文生成
コンテクストクエリ
コンテクスト特徴ベクトルを抽出
コンテクストＣＮＮモデル
組み合わせモデル
文生成（ＲＮＮ）モデル

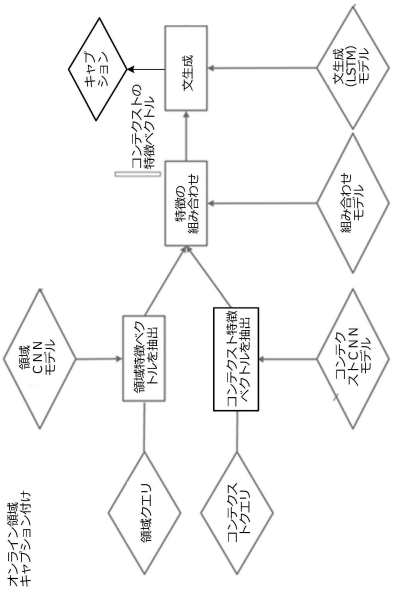
【図７】


1
K1-1
領域特徴表現
1
K2-1
コンテクスト特徴表現
K
N
特徴の組み合わせ

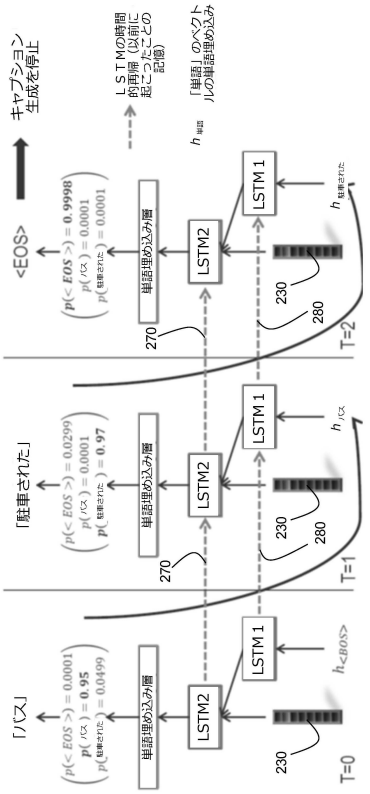
【図８】


オフライン
語学モデル学習
キャプション訓練セット
言語モデルを学習（ＲＮＮ）
微調整
オフライン領域キャプション付け学習
３つ組（画像、領域、キャプション）訓練セット
領域キャプション付けモデルを学習（ＣＮＮ＋ＲＮＮ）
学習済みモデル

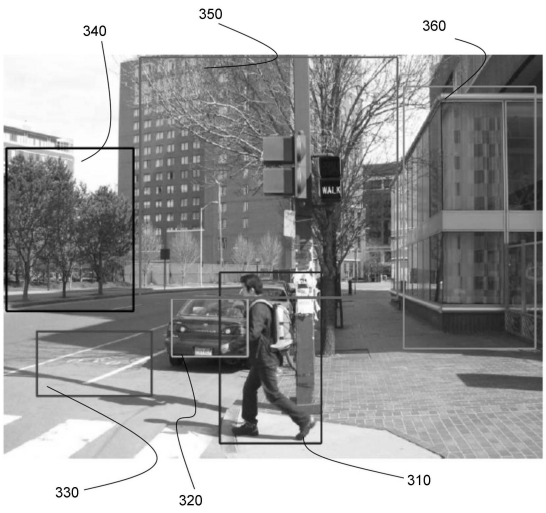
【図９】



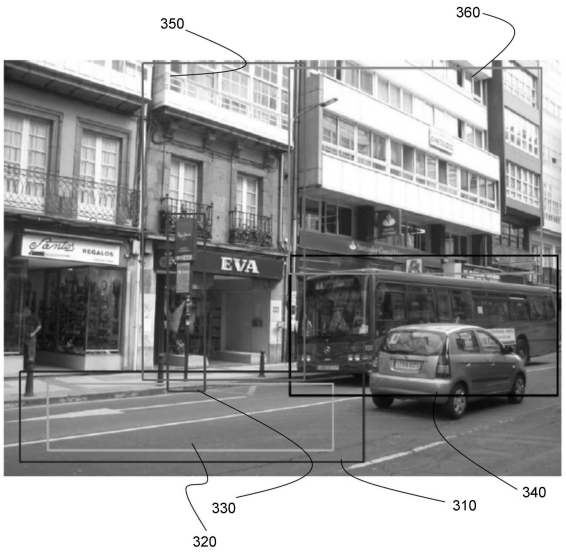
【図１０】



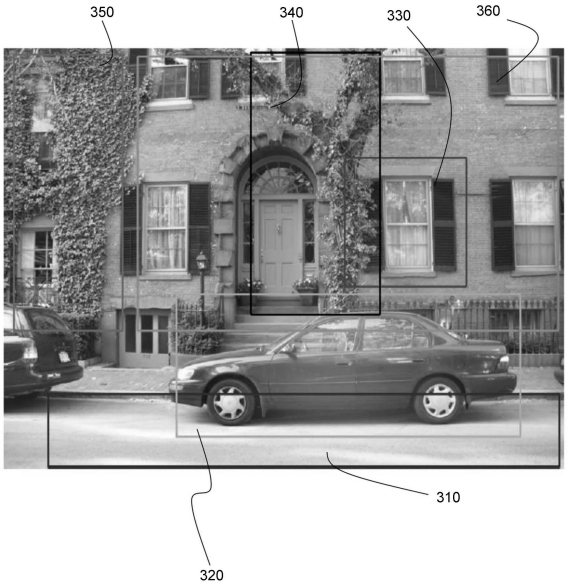
【図１１】



【図１２】



【図１３】



フロントページの続き

審査官　笠田 和宏

(56)参考文献　米国特許出願公開第２０１０／０１１１３７０（ＵＳ，Ａ１）
Justin Johnson，外２名，"DenseCap: Fully Convolutional Localization Networks for Dense Captioning"，2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)，2016年12月12日，pp. 4565-4574
Jeff Donahue，外６名，"Long-Term Recurrent Convolutional Networks for Visual Recognition and Description"，IEEE Transactions on Pattern Analysis and Machine Intelligence，Volume 39，Issue 4，2017年04月01日，pp. 677-691

(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ０６Ｔ　７／００
Ｇ０６Ｎ　３／０４
Ｇ０６Ｎ　３／０８