様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ５ 月 ２７ 日現在

機関番号：１２６０１
研究種目：基盤研究(B)（一般）
研究期間：2018～2020
課題番号：１８Ｈ０３２６８

研究課題名（和文）言語と視覚をつなぐ形式的意味表現の研究

研究課題名（英文）Formal Semantic Representations to Link Language and Vision

研究代表者
宮尾　祐介（Miyao, Yusuke）

東京大学・大学院情報理工学系研究科・教授

研究者番号：００３４３０９６

交付決定額（研究期間全体）：（直接経費）　　13,200,000円

研究成果の概要（和文）：本研究では、画像などの視覚情報に対して自然言語処理における意味解析技術を応用することを目標として、画像に対する意味表現の研究を行なった。具体的には、画像中のエンティティを認識してデータベース中のIDにリンクするエンティティリンキングと、画像の内容を表す意味表現を構成的に計算する技術の開発を行なった。また、画像と意味表現断片を入力として与えて説明文を生成するタスクを新たに設計し、画像の意味表現の有用性を示した。

研究成果の学術的意義や社会的意義
画像と言語をつなぐ技術は近年数多く研究されているが、そのほとんどは画像と言語を入出力として深層学習モデルを学習する手法である。この手法は大規模な学習データがあれば多くのタスクで高い精度を達成するが、学習データがない場合や、外部知識や推論を必要とする高度なタスクに適用することは難しい。提案手法のように画像に対して意味表現を得ることができれば、意味表現を利用した自然言語処理技術を応用する道が開け、さまざまな技術に発展することが期待できる。

研究成果の概要（英文）：This research explored semantic representations for images with the aim of applying semantic analysis technologies of natural languages to visual information. Specifically, we developed a method for linking entities in an input image into database IDs and a technique for compositionally constructing semantic representations of images. In addition, we designed a new task of generating a caption given an image and a fragment of a semantic representation as input and showed the effectiveness of using semantic representations for images.

研究分野：自然言語処理

キーワード：意味表現　自然言語処理　画像処理

様　式　C－１９、F－１９－１、Z－１９（共通）

１．研究開始当初の背景

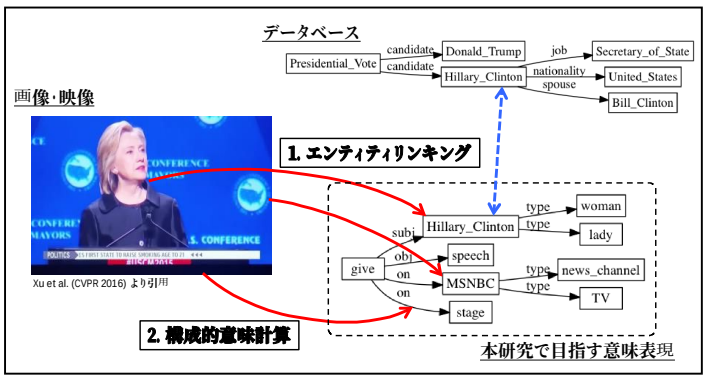
大規模データを利用した機械学習により自然言語処理や画像処理は飛躍的な進歩を遂げ、構文解析、機械翻訳、画像検索、一般物体認識など、個々のタスクにおいては実用レベルの高精度が達成されつつある。この成功のカギは、大量の入出力ペアを学習データとして用いて、入出力の写像関係を学習することにある。特に、近年の深層学習手法によって写像関係をより高精度に学習できるようになり、この傾向はますます強まっている。

しかし、言語や視覚情報の意味理解とは何か、という問題はいまだに解決の緒すら見えていない。むしろ、短期的に実用性の高い技術を開発するためには、意味理解の問題はなるべく回避し、入出力データを大量に作成し、入出力を直接つなぐモデル（End-to-end モデル）を学習する方が早道である。実際、自然言語処理や画像処理の研究の多くはこのアプローチをとっている。

本研究では、言語や視覚情報の意味理解の問題に取り組む。意味理解とは、入力された一次データ（テキストや画像）を、他の情報と結合してさまざまな予測・判断ができる形式（意味表現）に変換することであると考える。例えば、自然言語の意味理解技術として述語論理に基づく意味表現の計算やそれを用いた意味推論が挙げられる。これは、入力テキストからその意味を表す述語論理式を得、言語知識（シソーラスやパラフレーズ）と結合することで、テキスト間の推論関係の計算（テキスト間含意関係認識）を実現する。これは自然言語と言語知識をつなぐ例であるが、このアイディアを他メディアに広げることができれば、より人間に近い意味理解の実現にアプローチすることができる。

２．研究の目的

本研究では、言語と視覚情報、あるいは知識グラフなどのデータベースを相互運用することができる意味理解フレームワークを探究する。特定の応用タスクに特化したモデルを開発するのではなく、意味理解を必要とするさまざまなタスク（質問応答、要約、説明文生成、対話システムなど）において、究極的には学習データを用意しなくても利用できる汎用的な意味表現を考える。具体的には、自然言語の汎用的意味表現として確立されている述語項構造、意味的依存構造、述語論理、あるいはそれを簡略化したものとみなせる統語的依存構造をベースとし、視覚情報を形式的な意味表現の枠組みにのせるための基盤技術として、画像データに対する 1. エンティティリンキングと 2. 構成的意味計算の研究を行う。さらに、意味表現を利用する応用タスクを設計し、本研究の妥当性・有効性を評価する。



Xu et al. (CVPR 2016) より引用

本研究の概略を上図に示す。入力画像に対し、右下のような意味表現を計算することを目指す。この意味表現は、ノードがエンティティや名詞・動詞、エッジがノード間の関係（subj:主語、obj:目的語、など）を表す有向グラフである。エンティティは Freebase や Wikidata などの既存データベースにリンクする。この有向グラフは自然言語処理で意味表現としてよく用いられる述語項構造あるいは依存構造と対応している。

画像に対する説明文生成、質問応答、テキストクエリによる検索など、画像と自然言語をつなぐ研究が多数報告されているが[1,2,6,11,12]、そのほとんどは入力の解析モデルと出力の生成モデルを直接つなげたニューラルネットワークで、end-to-end で学習を行うことで高い精度を達成している。本研究は、特定タスクのためのモデルを構築して精度を追求するのではなく、言語と視覚情報をつなぐインタフェースとなる中間表現として意味表現を導入し、言語・視覚情報それぞれをこの中間表現に写像する、というアプローチを取る。中間表現として自然言語処理で研究されてきた意味表現を用いるため、応用タスクごとに end-to-end の学習データを用意しなくても、意味表現を入力とする既存手法（テキスト間含意関係認識、質問応答、要約、文生成など）はそのまま応用することができる。また、本研究における意味表現は人間が解釈できる表現形式であるため、説明性が高く、またルールベースの処理を行うことも容易である。

３．研究の方法

本研究では、画像データに対する 1. エンティティリンキング、および 2. 構成的意味計算の 2 つの技術について研究を行う。また、画像の意味表現を用いた新たな応用タスクを設計し、提

案手法の有用性を示す。

(1) 画像に対するエンティティリンキング

本研究項目では、画像中のエンティティを認識し、オントロジーやデータベースにリンクする手法の研究を行う。図の例では、映像中の人物やテレビ番組に対し、“Hillary_Clinton”, “MSNBC” といったエンティティ名を認識する。これにより、視覚情報と自然言語、さらには Wikidata などのデータベースを結合することができるようになる。例えば、データベースを使って各エンティティに関する意味情報や、シソーラスなどの言語知識と結合することが可能となり、さまざまな応用への展開が期待される。

画像に対するエンティティリンキングについてはすぐに利用できるリソースが存在しなかったため、画像等のメディアファイルを Wikidata と紐付けられる形で管理している Wikimedia Commons を利用して学習・評価データを開発する。認識手法については、一般物体認識モデルをそのまま適用する手法や、画像特徴量を用いた手法を実装し、比較検討を行う。

(2) 画像に対する構成的意味計算

構成的意味計算とは、単語の意味表現を合成していくことで文全体の意味表現を計算する手法である。述語論理に基づく意味解析でよく使われる手法であり、これを、画像に対して適用することを考える。画像については、単語という単位や単語間の意味合成の方法が自明でなく、自然言語処理の手法をそのまま適用することはできない。

本研究項目では、画像の内容に関する「単語」とその間の関係（述語項関係あるいは依存関係）を認識することで、意味表現を計算する手法の研究を行う。図の例では、Hillary_Clinton を主語 (subj) として、“give a speech”, “on stage” といったアクション・属性や、エンティティ間の関係 (Hillary_Clinton と MSNBC) が認識されている。画像の内容に関する単語とその間の関係を認識することができれば、それをグラフ構造として合成することで、画像の内容を表す意味表現を構成することができる。

(3) 応用タスクの設計と評価、デモシステムの開発

提案手法の有用性を客観的に評価するため、応用タスクを設計し、評価実験を行う。画像処理においては、検索クエリとしてテキスト情報のみがあたえられる画像検索や、画像の内容を説明する文章を生成する画像説明文生成が提案されている。これらを参考とし、複雑な自然言語クエリに答える画像検索、あるいは与えられたクエリにそって画像の内容を要約する画像要約タスクなどを検討する。

４．研究成果

(1) 画像に対するエンティティリンキング

本研究項目では、入力画像に対し、その中の物体をデータベース中のエンティティにリンクする手法について研究を行った。自然言語処理においては、テキスト中の人名や地名を認識し、それが指す内容をデータベースのエンティティ ID にリンクするエンティティリンキングが研究されている。本研究はそのアイディアを参考としているが、自然言語テキストはテキストの周辺情報やデータベース中のテキスト情報（エンティティの説明文など）を利用することができるのに対し、画像についてはそのような情報を用いることは難しい。よって、自然言語テキストのエンティティリンキングとはまったく異なる解析技術が必要である。また、画像に対するエンティティリンキングは確立されたタスクではないため、学習・評価のためのデータセットの構築も新たに必要となる。

まず、画像に対するエンティティリンキングの学習・評価データの構築を行なった。このタスクのためには、画像データに加えて、その内容についてデータベースのエンティティ ID と紐付けされている必要がある。そこで、Wikimedia Commons に大量に収録されている画像データを利用することとした。Wikimedia Commons は、画像や音声等のメディアファイルを一般ユーザがアップロードし、それを再利用可能ライセンスで公開しているプロジェクトである。Wikimedia Commons では、Wikipedia と同様にさまざまな概念（カテゴリ）についてページが用意されており、ユーザはそのカテゴリにあった画像をアップロードする。カテゴリには ID が付与されており、それは大規模データベースである Wikidata と対応づけられている。したがって、まず Wikidata のカテゴリの ID を収集し、それに対応する画像を Wikidmedia Commons からクロールすれば、画像と ID が対応したデータを構築することができる。また、Wikidata ではカテゴリ間の関係が与えられているため、これを利用することであるカテゴリの下位カテゴリを収集することができる。例えば、「犬」についてその下位カテゴリを収集すれば、犬種を表すカテゴリ ID を収集することができる。

Wikimedia 全体の画像データを収集するのは多大なコストがかかるため、本研究では、アメ

リカ大統領、犬、タワー、電子機器、自動車の 5 つのカテゴリについて、エンティティに相当する下位カテゴリ（固有名、種別名あるいは製品名）を収集し、それに対応する画像データを整備した。その結果、それぞれのカテゴリについて 44（アメリカ大統領）から 6,000 以上（自動車）のエンティティおよび 20 万点以上の画像データからなるデータセットを構築した。

次に、本データを学習・評価データとして用いて、入力画像に対してエンティティ ID を認識する手法について研究を行なった。深層学習以前の画像特徴量[10]を用いてエンティティ ID をランキングする手法と、深層学習による一般物体認識モデル[4,9]から得られる特徴量を用いてエンティティ ID をランキングする手法を構築し、評価実験を行なった。実験により、一般物体認識を応用した手法を用いれば、エンティティの曖昧性が少ない場合（アメリカ大統領など）は一定の認識精度が得られるものの、エンティティの候補数が大きい場合（電子機器や自動車）では認識精度が低下することが示された。エンティティ候補数が大きい場合は、エンティティに対応する画像数に大きな偏りがあり、画像が 1 枚しかない場合も多数存在する。したがって、教師あり学習を用いた手法では原理的に限界があると考えられる。

将来研究としては、まず対象カテゴリを拡大し、究極的にはデータベースのあらゆるエンティティについて認識可能な技術を開発することがある。原理的には上述の手法によってデータを構築することができるが、画像データが膨大になるため、効率的な実装が必要となる。また、エンティティの曖昧性が大きいとき、あるいはエンティティに対応する画像データが少ない時に認識精度が低いという問題を解決する必要がある。画像認識においては、少数の学習データあるいは学習データがない時に精度を向上させる手法が研究されており、それを応用することが考えられる。

(2) 画像に対する構成的意味計算

本研究項目では、入力画像に対して単語およびその関係を認識することで意味表現を計算する手法について研究を行なった。画像に対してグラフ構造や論理式を用いて意味を表す既存研究は複数存在するが[5,7]、それらはまず画像中の物体認識を行い、さらに物体間の関係や物体の属性の認識を行う。本研究では、物体に限らず画像の内容を表すのに適切な「単語」（名詞だけでなく動詞、形容詞、副詞を含む）を認識し、さらにそれらの間の依存関係を認識する。これにより、自然言語の意味表現である述語項構造あるいは依存構造に直接対応するグラフ構造を計算する。この手法については本研究開始前に予備的な実験を行なっており、画像説明文生成のデータセットを利用して意味表現のデータを構築し、意味表現の各エッジ（依存関係）を予測するモデルを開発した。そこで、この手法をベースとして、認識精度を向上させる手法を探究した。

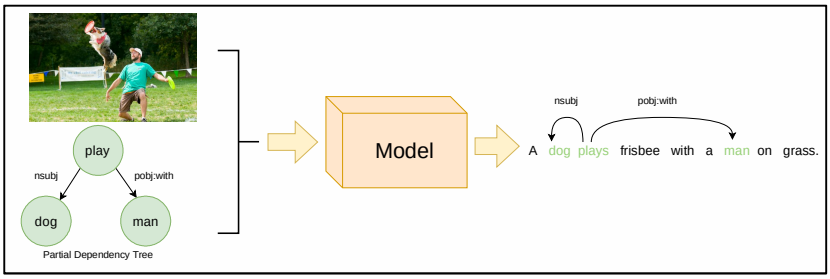
まず、学習・評価データの構築について研究を行なった。画像とその内容を説明した自然言語テキストが対応づけられた画像説明文データセットは、クラウドソーシング等を利用して大規模なデータが開発・公開されている[8]。そこで、このデータセットの自然言語テキストに対して構文・意味解析を行い、意味表現に変換することで、画像と意味表現が対応づけられたデータを構築する。ただし、この手法では自然言語テキストに明示的に現れる単語および依存関係のみが得られるが、実際にはそこに現れない単語・依存関係も正解として考えられる。そこで、シソーラスを利用して、単語に対してその上位語（例えば「チワワ」に対して「犬」）をグラフに追加することで、この問題を部分的に解決する手法を開発した。実際、このようにして構築したデータを用いることで、後述する意味表現認識の精度が向上することを実験において確認した。

次に、以上のように構築した画像・意味表現データセットを用いて、入力画像に対して意味表現を自動認識する手法の研究を行なった。前述のとおり、意味表現の各エッジを認識するモデルをベースとして、これを改善する手法を探究した。まず、上述のように説明文から得られる意味表現グラフを上位語で拡張することで、単語および依存関係の認識精度が向上することを示した。ただし、この手法では説明文に下位語が現れない場合には有効でない（例えば「犬」がうつっている画像に対して「チワワ」といった単語が説明文に現れなければ、「犬」はデータに含まれない）。つまり、このデータセットでは正解ラベルとして意味表現が不足しているということになる。そこで、正解ラベルが完全には付与されていないデータを学習データとする Positive-Unlabeled 学習（PU 学習）[3]を適用する手法を構築した。PU 学習は、学習データにおいて正解ラベルの一部が付与されているが、ラベルが付与されていないデータを負例としては扱わない（Unlabeled データと考える）ことで、不十分な学習データからでもモデルの学習を可能とする手法である。意味表現認識においては、意味表現のエッジをラベルとみなし、PU 学習を適用することができる。評価実験により、PU 学習を適用することで意味表現の認識精度が向上することを示した。

将来課題としては、意味表現認識のさらなる精度向上が挙げられる。一般物体認識の最先端のモデルを応用することである程度の精度の意味表現認識モデルを実現できたが、特に低頻度の単語や依存関係については十分な認識精度が得られていない。これは深層学習を用いた画像処理に共通する問題であるが、転移学習等の利用による精度向上を検討する必要がある。また、研究項目 1 のエンティティリンキングとの統合も将来課題として挙げられる。エンティティリンキングによってデータベース中の知識とつながるため、それを意味表現認識の精度向上に利用することも考えられる。

(3) 応用タスクの設計と評価、デモシステムの開発

本研究項目では、意味表現を活用する新たなタスクを設計し評価実験を行うこと、またデモシステムを開発することを目的とした。自然言語処理においては、意味表現を用いるタスクとしてテキスト間含意関係認識や質問応答などがあるが、画像と言語をつなぐという目的に基づいて新たな問題設定を模索した。画像と言語をつなぐタスクとして画像説明文生成の研究がされているが[11,12]、これをベースとして、意味表現を活用することで新たな価値を生み出すアプリケーションについて検討を行なった。



そこで、入力として画像だけでなく意味表現の断片（依存関係あるいは部分的な依存構造木）を与え、その意味表現を含むような説明文を生成するタスクを設計した。つまり、意味表現の断片によって説明文生成を制御するというアイディアである。具体例を上図に示す。入力として、左にあるような画像に加えて依存構造を与える。そして、説明文生成モデルはこの依存構造を含むような説明文を生成する。例えばこの画像に対しては、A man plays frisbee with a dog.という説明文も正しいが、入力の依存構造により dog が主語、man が修飾語となるような説明文を指定している。画像の内容を説明する文章は多様であり、現在の説明文生成手法はそのうちどれを出力するかは考慮されていないが、本手法により出力すべき意味を指定することができるようになる。

研究項目 2 で構築したデータセットを利用して、本タスクの学習・評価データを構築した。また、このタスクの精度を測定するための手法を新たに開発した。このタスクに対して、入力の依存構造を含む構造を中間表現として生成してから説明文生成を行う手法を提案し、評価実験を行なった。既存の説明文生成手法をそのまま流用して入力の画像と依存構造をそれぞれ埋め込み表現に変換して end-to-end で説明文生成を行う手法をベースラインとして比較実験を行なった結果、提案手法では既存手法と同等の自然さの文を、より正確（入力の依存構造を含む文）に生成できることが示された。

これに加えて、意味表現を用いて画像検索を行うデモシステムを構築した。大規模画像データに対してあらかじめ本研究の意味解析技術を適用し、意味表現のデータベースを構築する。自然言語で検索クエリを与えると、その意味表現とデータベースとを照合し、適合する画像を出力する。このタスクについては、画像と言語を同じベクトル空間に埋め込む手法[6]などが提案されているが、意味表現は離散的かつ解釈可能なデータ構造であり、検索結果の解釈性の高さがデモシステムによって示された。

＜引用文献＞

[1] Antol et al. (2015) VQA: Visual Question Answering. ICCV 2015.
[2] Datta et al. (2008) Image retrieval: Ideas, influences, and trends of the new age. Journal of ACM Computing Surveys.
[3] Elkan et al. (2008) Learning classifiers from only positive and unlabeled data. KDD '08.
[4] He et al. (2016) Deep residual learning for image recognition. CVPR 2016.
[5] Hurlimann et al. (2016) Combining Lexical and Spatial Knowledge to Predict Spatial Relations between Objects in Images. ACL 2016 Workshop on Vision and Language.
[6] Kiros et al. (2014) Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. NIPS 2014 deep learning workshop.
[7] Krishna et al. (2017). Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. International Journal of Computer Vision.
[8] Lin et al. (2014) Microsoft COCO: common objects in context. ECCV 2014.
[9] Simonyan et al. (2015) Very deep convolutional networks for large-scale image recognition. ICLR 2015.
[10] Lowe (2004) Distinctive image features from scale-invariant keypoints. International journal of computer vision 60.
[11] Tran et al. (2016). Rich Image Captioning in the Wild. CVPR 2016.
[12] Vinyals et al. (2015) Show and Tell: A Neural Image Caption Generator. CVPR 2015.

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　0件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 宮尾 祐介	４．巻 34(6)
２．論文標題 多様なデータと自然言語をつなぐ基盤技術	５．発行年 2019年
３．雑誌名 学会誌「人工知能」特集「人間と相互理解できる次世代人工知能技術」	６．最初と最後の頁 811-816
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計1件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 Wenjie Zhong, Yusuke Miyao
２．発表標題 Leveraging Partial Dependency Trees to Control Image Captions
３．学会等名 Proceedings of the Second Workshop on Advances in Language and Vision Research（国際学会）
４．発表年 2021年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

（招待講演）宮尾 祐介．自然言語のグラウンディング研究の概観．計算言語学の現在 2019年12月6日．

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関