CCGと定理証明器を用いた画像情報の意味表現と推論の試み

鈴木 莉子 [1,a]　谷中 瞳 [1,2,b]　峯島 宏次 [1,c]　戸次 大介 [1,d]
お茶の水女子大学 [1], 理化学研究所 AIP センター [2]
{g1520544[a], bekki[d]}@is.ocha.ac.jp,
mineshima.koji@ocha.ac.jp[c], hitomi.yanaka@riken.jp[b]

1　はじめに

近年、画像や映像などの非テキストデータとテキストデータといった異なるモーダルの情報を統合的に理解し、新しい知識を獲得するマルチモーダル推論に関する研究が注目を浴びている。画像とテキストを用いた代表的な研究の一つである画像キャプション生成は、自然言語文を検索クエリとした画像検索 [2] への応用も期待されているが、一般に自然言語文は部分的な情報しか表現しないため、キャプションだけで画像情報を全て捉えることは困難である。例として、(1) と (2) の文を検索クエリとして、画像キャプションデータセット MS-COCO [11] にある図 1 の画像を検索できるか否かについて考える。

(1)　少なくとも 2 匹の猫がいる。
(2)　白い猫はいない。

図 1: MS-COCO の画像とそのキャプション

図 1 の画像には猫が 3 匹写っているが、いずれも白い猫ではない。よって、この画像が表す状況で (1) と (2) の文は真である。しかし、これらの情報はキャプションには記述されていないため、クエリ文とキャプションのマッチングによる検索では望まれる検索結果は得られない。こうした意味的に複雑な文による検索を実現にするためには、画像情報をテキスト情報とシームレスな形式で表現し、文から画像、画像から文へと推論できる機構が必要となる。

そこで本論文では、マルチモーダル推論の一つの試みとして、画像から数量表現や否定を含む意味的に複雑な文を推論するシステムを提案する。画像情報を一階述語論理 (FOL) のモデルと論理式を用いて表現することで、画像検索は、画像情報を前提、テキストを結論とした含意関係認識の問題とみなすことが可能である。これにより、構造的に複雑な文も検索クエリとして扱えるシステムを検討したい。

2　関連研究

画像検索の関連研究としては、キーワードベースによる手法 [9] がある。この手法では意味的に複雑なクエリ文を扱うことが困難であり、またキャプションに記述されていない情報を検索クエリとして用いることが難しいという問題が指摘されている [4]。

画像とテキストの意味表現に関する研究としては、画像とキャプション中の単語を共通のベクトル空間に埋め込む multimodal embedding の手法 [1, 8, 6] が活発に研究されている。しかし、これらの手法は画像とキャプションを近いベクトル空間に配置することを目的としているため、キーワードベースによる手法と同様に、キャプションに記述されていない画像情報を表現することは考慮していない。

画像中の物体とその属性、また物体間の関係を表す手法としては、グラフ表現 (Scene Graph) によるものが提案されている [10]。グラフ表現を用いることで、キーワードのみを抽出した画像検索では実現できなかった複雑なクエリ文による画像検索が可能となる。しかし、一般にグラフ表現では物体間の関係に加え、否定や量化などの意味的・構造的に複雑な文の意味を統一的に扱うことは困難である。そのため、グラフ表現よりも表現力の高い意味表現で画像情報を表現する手法が求められる。

3　提案手法

提案手法では、画像情報の意味表現として FOL の論理式と構造（モデル）を採用する。論理式を採用することの利点は二つある。第一に、文を論理式に変換する手法と組み合わせることで、文と画像情報を接合した推論が可能となる点である。特に、画像検索は、画像の意味表現 M を前提、検索クエリ文の意味表現 T を結論とした $M \vdash T$ という含意関係を判定する問題として解くことができる。第二に、論理式の表現力に

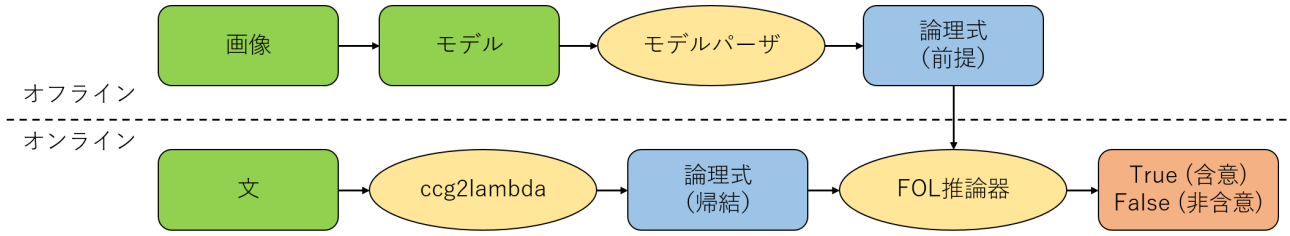

図 2: 提案手法

よって、否定、量化、関係といった表現を含む文も検索クエリとして扱うことが可能になる。

図 2 に提案するシステムの全体像を示す。本システムは大きく分けて 3 つのモジュールに分けられる。1) FOL のモデルとして記述された画像を「モデルパーザ」により論理式に変換する、2) 意味解析・推論システム ccg2lambda [5] を用いて文を論理式に変換する、3) 各画像の論理式を前提、文の論理式を帰結として、その画像が文を含意するか否かを定理証明器を用いて推論する。含意関係が成り立つ場合、その画像は結論の文が成り立つ状況を表すものとみなせる。以下に各モジュールの詳細について述べる。

3.1 画像から論理式への変換

本研究では画像を直接論理式に変換するのではなく、まず画像をモデルとして記述し、モデルパーザを用いてモデルを論理式に変換する方法を用いる。モデル M はドメイン D と評価関数 I からなり、画像に写っている物体や物体間の関係を表現している[1] [7]。本研究の実験では画像に対してモデルを記述した GRIM データセット [3] を用い、画像からモデルへの変換方法の研究は今後の課題とする。この節ではモデル M をどのように論理式に変換するかについて述べる。

エンティティの定義： ドメインの情報からエンティティの定義をする。ここではドメイン D が $[d1, d2, d3]$ の場合について考える。モデルには「モデルの中で記述された物体、関係以外存在しない」という情報が含まれるが、式 (3) のように定義しても「エンティティは $d1, d2, d3$ しかない」という情報は含まれない。

(3) $entity(d1) \wedge entity(d2) \wedge entity(d3)$

そこで式 (4) のように定義することで「エンティティは $d1, d2, d3$ しかない」という情報を扱えるようになる。

(4) $\forall x.(entity(x) \leftrightarrow x = d1 \vee x = d2 \vee x = d3)$

[1] $x \in D, cat \in I$ について $cat(x) = 1$ は x が cat であることを表す。

各エンティティは異なる： 数詞を含む文の意味を記述するためには「各エンティティは異なる」という公理が必要となる。式 (5) に示す論理式を用いて定義する。

(5) $\neg(d1 = d2) \wedge \neg(d2 = d3) \wedge \neg(d3 = d1)$

n 項述語： 次に、評価関数を論理式に変換する。例として 1 項述語の情報 $I(man, [d1])$ は式 (6) のように、2 項述語の情報 $I(touch, [(d1, d2), (d2, d1)])$ は式 (7) のように変換する。

(6) $\forall x.(man(x) \leftrightarrow x = d1)$

(7) $\forall x\, y.(touch(x, y) \leftrightarrow (x = d1 \wedge y = d2) \vee (x = d2 \wedge y = d1))$

3.2 文から論理式への変換

本研究では文から論理式への変換に意味解析・推論システム ccg2lambda を用いる。以下の各表現について画像検索に適した論理式を生成するよう改良を加えた。

数量表現： 一般に「(少なくとも) n の F」「高々 n の F」「ちょうど n の F」という表現は表 1 に示す論理式で表せる。例えば「(少なくとも) 2 匹の猫がいる」「高々 2 匹の猫がいる」「ちょうど 2 匹の猫がいる」という文は式 (8)(9)(10) で表せる。

(8) $\exists x\, y.(cat(x) \wedge cat(y) \wedge \neg(x = y))$

(9) $\forall x\, y\, z.(cat(x) \wedge cat(y) \wedge cat(z)) \rightarrow ((x = y) \vee (y = z) \vee (z = x))$

(10) $\forall x\, y\, z.(cat(x) \wedge cat(y) \wedge cat(z)) \rightarrow ((x = y) \vee (y = z) \vee (z = x)) \wedge \exists x\, y.(cat(x) \wedge cat(y) \wedge \neg(x = y))$

数量表現	論理式
(少なくとも) n の F	$\exists x_1 \ldots x_n F(x_1) \wedge \ldots \wedge F(x_n)$ $\wedge\neg(x_1 = x_2) \wedge \ldots \wedge \neg(x_{n-1} = x_n)$
高々 n の F	$\forall x_1 \ldots x_{n+1}.F(x_1) \wedge \ldots \wedge F(x_{n+1})$ $\rightarrow x_1 = x_2 \vee \ldots \vee x_n = x_{n+1}$
ちょうど n の F	$\exists x_1 \ldots x_n.F(x_1) \wedge \ldots \wedge F(x_n)$ $\wedge\neg(x_1 = x_2) \wedge \ldots \wedge \neg(x_{n-1} = x_n)$ $\wedge\forall x_1 \ldots x_{n+1}.F(x_1) \wedge \ldots \wedge F(x_{n+1})$ $\rightarrow x_1 = x_2 \vee \ldots \vee x_n = x_{n+1}$

表 1: 数量表現の意味表示

全称量化：全称量化の意味表現を考える。例えば「全ての猫は白い」は次のようになる。

(11) $\forall x.(cat(x) \rightarrow white(x))$

しかし式 (11) では猫がいない画像についても True となる。「(猫が存在し、かつ) 全ての猫が白い」の意味を表現するためには (12) のようにすべきである。

(12) $\exists x.cat(x) \wedge \forall x.((cat(x) \rightarrow white(x))$

複合語： “part of” といった複合語は “part_of” のように “_” を用いて一つの述語とする。本研究では複合語のリストを用意し、該当する場合は “_” を用いて 1 つの述語とみなす。

3.3 推論

モデル検査と定理証明： 既存研究 [7] ではモデル検査を用いて画像と文の関係を推論するシステムを開発しているが、本研究では定理証明の有効性を検討する。モデル検査とはモデル M と論理式 A について「モデル M が論理式 A を充足するか ($M \models A$)」を判定する方法である。一方、定理証明とは論理式の集合 Γ と論理式 A について「論理式の集合 Γ が論理式 A を含意するか ($\Gamma \vdash A$)」を判定する方法である。

アブダクション： 一般にモデルは考えている世界についての完全な記述であるが、実験で用いる GRIM データセットでは述語間についての情報が欠けているため保管する必要がある。例えば式 (13) を見てみる。pigeon は bird の一種であるため式 (13) は True となるが、モデルに $I(pigeon, [\ldots])$, $I(bird, [\ldots])$ という情報がない場合、$\forall x.(pigeon(x) \rightarrow bird(x))$ という公理が必要になる。

(13) $\exists x.pigeon(x) \vdash \exists x.bird(x)$

そこでモデルと文に含まれる述語の関係から公理を新たに追加する。本研究では WordNet[2] を用いて 2 語の関係を調べ、表 2 に示す公理の生成規則に従ってアブダクションを行う。

F と G の関係		追加する論理式
F は G の同義語である	$\rightsquigarrow$	$\forall x.F(x) \leftrightarrow G(x)$
F は G の上位語である	$\rightsquigarrow$	$\forall x.G(x) \rightarrow F(x)$, $\neg\exists x.G(x)$
F は G の下位語である	$\rightsquigarrow$	$\forall x.F(x) \rightarrow G(x)$
F は G の反意語である	$\rightsquigarrow$	$\forall x.F(x) \leftrightarrow \neg G(x)$, $\neg\exists x.G(x)$

表 2: アブダクションの生成規則。F はモデルに含まれる述語、G は文に含まれる述語である。

[2]https://wordnet.princeton.edu/

4 評価実験

4.1 データセット

本研究では GRIM を用いて評価実験を行う。GRIM は画像、モデル、2 種類のキャプション（画像に対して真となる文と偽となる文）からなるデータセット[3]であり、物体、属性、物体間の空間情報がモデルで記述されている（図 3）。空間情報を表すために touch、near、support、occlude、part_of の 5 つの 2 項述語が用意されている。

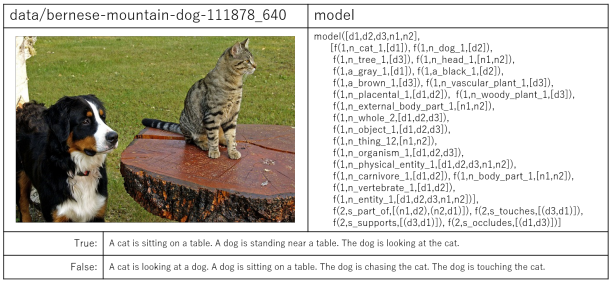

図 3: GRIM のデータ例

4.2 実験

GRIM のデータ 200 件中 192 件[4]を用いて実験を行う。FOL 推論器は Prover9[5]を用いた。まず複雑な言語現象ごとに文のタイプを分類した。現象ラベルとして論理結合子 (Con)、数詞 (Num)、量化詞 (Q)、関係 (Rel) の 4 種類を設けた。表 3 に各文の分類を示す。表 3 の各文に対する GRIM の正解画像を人手でタグ付けした。各分類の F 値は各文ごとの F 値の平均を取り、件数は各文の正解件数を合計した（表 4）。

図 4 に “There are at least two cats.” を入力文としたときのシステムが予測した画像を示す。それぞれ少なくとも二匹の猫がいる画像であり、提案法で意味的に複雑な文に対しても期待通りの画像が得られたことが分かる。

図 4: “There are at least two cats.” に対するシステムの予測画像

[3]http://www.let.rug.nl/bos/comsem/images/
[4]ノイズを含む 8 件のデータは、実験対象から除去した。
[5]https://www.cs.unm.edu/ mccune/prover9/

文	Con	Num	Q	Rel
There is a cat.	✓			
There is no cat.	✓			
There is a white cat.	✓			
There is not a white cat.	✓			
There is a cat and a dog.	✓			
There is a cat or a dog.	✓			
There are two cats.		✓		
There are three cats.		✓		
There are at least two cats.		✓		
Two cats are black.	✓	✓		
Two cats are white or black.	✓	✓		
At least two cats are black.	✓	✓		
Exactly two cats are black.	✓	✓		
All cats are white.	✓		✓	
Every person is touching a bicycle.			✓	✓
A cat is touching a dog.				✓
A cat is touching a head that is part of a dog.				✓
A bicycle is supporting a person.				✓
A person is supporting a bicycle.				✓

表 3: 現象ラベル（Con: 論理結合子, Num: 数詞, Q: 量化詞, Rel: 関係）による各文の分類

分類	F 値	件数
論理結合子	0.79	317
数詞	0.95	22
量化詞	0.73	23
関係	0.90	34

表 4: 複雑な言語現象ごとの F 値と正解画像の件数

4.3 エラー分析

エラー分析の結果、2 つのタイプのエラーが見つかった。1 つは GRIM のアノテーション不備によるエラーである。アノテータは “white cat” の判定を「体の一部が白い猫」が写っている画像を正解としたが、GRIM のデータに体の一部が白い場合でも $white(x)$ がアノテートされていないデータがあった。故に論理結合子の F 値が低くなったと考えられる。

もう 1 つは存在否定表現を含む文の解釈の問題である。本システムでは “There is not a white cat.” の論理式は (14) となり、この式は意味論的に正しく、(14) に対する正解画像は猫が写っていない画像と、猫が写っている場合その猫が白くない画像となる。しかし画像検索タスクにおいては、検索者が猫について検索している状況で、白い猫が写ったものは省きたいと意図する場合がある。その場合 (14) ではなく (15) としたい。これは語用論が関わる問題であり、今後考えていく必要がある。

(14) $\neg\exists x.(cat(x) \wedge white(x))$

(15) $\exists x.cat(x) \wedge \neg\exists y.(cat(y) \wedge white(y))$

5 おわりに

本稿では一階述語論理式を用いることにより画像から文を推論するシステムを提案した。画像を論理式で記述することにより、画像をより詳細に表現することができた。また論理式を用いることで数詞や量化子を含む複雑な文の意味も表現できた。

GRIM には画像、モデルに加えてキャプションも用意されており、今後はキャプションをモデルに変換する方法を試みる。GRIM のモデルには関係を表す述語は空間情報しか付与されていないが、キャプションには “sit” や “look” など、より詳細な情報が記述されている。これらの述語もモデルに加えることで画像検索の幅が広がると考えられる。

謝辞 この研究は、JST CREST「ビッグデータ統合利活用のための次世代基盤技術の創出・体系化」領域「知識に基づく構造的言語処理の確立と知識インフラの構築」プロジェクトの支援を受けたものである。

参考文献

[1] Grzegorz Chrupała, Ákos Kádár, and Afra Alishahi. Learning language through pictures. In *ACL*, 2015.

[2] Jiuxiang Gu, Jianfei Cai, Shafiq R. Joty, Li Niu, and Gang Wang. Look, Imagine and Match: Improving Textual-Visual Cross-Modal Retrieval With Generative Models. In *CVPR*, 2018.

[3] Manuela Hürlimann and Johan Bos. Combining Lexical and Spatial Knowledge to Predict Spatial Relations between Objects in Images. In *Proc. of the Workshop on Vision and Language*, 2016.

[4] Ying Liu, Dengsheng Zhang, Guojun Lu, and Wei-Ying Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 2007.

[5] Koji Mineshima, Pascual Martínez-Gómez, Yusuke Miyao, and Daisuke Bekki. Higher-order logical inference with compositional semantics. In *EMNLP*, 2015.

[6] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg Corrado, and Jeffrey Dean. Zero-Shot Learning by Convex Combination of Semantic Embeddings. In *ICLR*, 2014.

[7] Blackburn Patrick and Bos Johan. *Representation and Inference for Natural Language*. CSLI, 2005.

[8] Andrea Frome *et al.* DeViSE: A Deep Visual-Semantic Embedding Model. In *NIPS*. 2013.

[9] Hao Xu *et al.* Image Search by Concept Map. In *Proc. of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2010.

[10] Justin Johnson *et al.* Image retrieval using scene graphs. In *CVPR*, 2015.

[11] Tsung-Yi Lin *et al.* Microsoft COCO: Common Objects in Context. In *ECCV 2014*, 2014.