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 （指導教員：戸次 大介）

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。

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し、意味的・構造的に複雑な文も検索クエリとして扱える画像検索システムの構築を目指す。

2 関連研究

画像検索の関連研究としては、まずキーワードベースによる手法 [8] 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 [3]。

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, 7, 5] 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 [9]。グラフ表現を用いることで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

図 1 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 [4] 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。

3.1 画像から論理式への変換

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] [6]。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 [2] を用い、画像からモデルへの自動変換の研究は今後の課題とする。

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 $D = [d_1, d_2, d_3]$ の場合について考える。式 (1) のように定義することで、「エンティティは d_1, d_2, d_3 しかない」という情報を論理式で記述できる。

(1) $\forall x.(\mathsf{entity}(x) \leftrightarrow x = d_1 \vee x = d_2 \vee x = d_3)$

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 (2) に示す論理式を用いて定義する。

(2) $(d_1 \neq d_2) \wedge (d_2 \neq d_3) \wedge (d_3 \neq d_1)$

n 項述語： 評価関数を論理式に変換する。例として 1 項述語の情報 $I(\mathsf{man}, [d_1])$ は式 (3) のように、2 項述語の情報 $I(\mathsf{touch}, [(d_1, d_2), (d_2, d_1)])$ は式 (4) のように変換する。

(3) $\forall x.(\mathsf{man}(x) \leftrightarrow x = d_1)$

(4) $\forall x\, y.(\mathsf{touch}(x, y)$
$\leftrightarrow (x = d_1 \wedge y = d_2) \vee (x = d_2 \wedge y = d_1))$

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ことが知られている。本研究では表 1 に従い数詞を含む文を論理式に変換する。

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

表 1: 数量表現の意味表示

全称量化： 例として「全ての猫は白い」という文を考える。この文を「（猫が存在し、かつ）全ての猫が白い」と解釈し、式 (5) を用いて全称量化を含む文の意味を表現する。

(5) $\exists x.\mathsf{cat}(x) \wedge \forall x.((\mathsf{cat}(x) \rightarrow \mathsf{white}(x))$

3.3 推論

定理証明： 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。本研究では画像と文の関係の推論における定理証明の有効性を検討する。

[1] $a \in D$ について、$(\mathsf{cat}, L) \in I$ は、$a \in L$ ならば a が cat であることを表す。

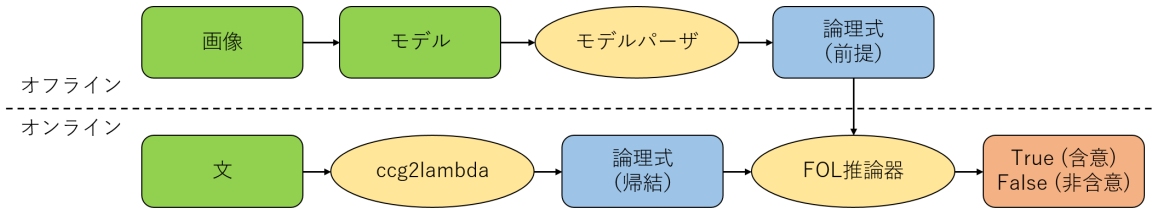

図 1: 提案手法の構成

アブダクション：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述語間についての情報を補完する。

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。

4 評価実験

本研究では GRIM データセットのデータ 200 件中 192 件[3]を用いて、画像から複雑な言語現象を含む文を推論できるか評価実験を行う。データ例を図 2 に示す。

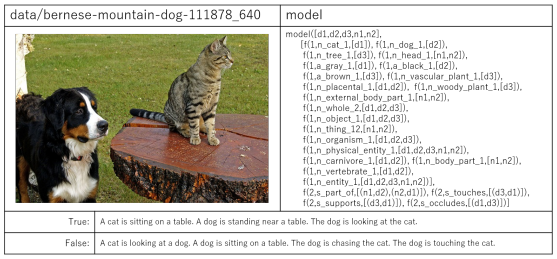

図 2: GRIM のデータ例

FOL 推論器は Prover9[4] を用いた。表 3 に示す各文を複雑な言語現象ごとに分類し、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 3 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、本提案手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 3: “There are at least two cats.” に対するシステムの予測画像

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 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）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検討する。

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

参考文献

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

[2] 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.

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

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

[5] 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.

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

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

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

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

[2] https://wordnet.princeton.edu/
[3] ノイズを含む 8 件のデータは、実験対象から除去した。
[4] https://www.cs.unm.edu/ mccune/prover9/