

テキスト情報と画像情報を組み合わせた論理推論システムの構築

Towards Building a Logical Inference System for Text and Visual Information

鈴木 莉子^{*1}

Riko Suzuki

吉川 将司^{*2}

Masashi Yoshikawa

谷中 瞳^{*3}

Hitomi Yanaka

峯島 宏次^{*1}

Koji Mineshima

戸次 大介^{*1}

Daisuke Bekki

^{*1}お茶の水女子大学

Ochanomizu University

^{*2}奈良先端科学技術大学院大学

Nara Institute of Science and Technology

^{*3}理化学研究所 AIP センター

RIKEN Center for Advanced Intelligence Project

A large amount of research about multimodal inference across natural language and vision has been recently developed to obtain visually grounded word and sentence representations. In this paper, we use logic-based representations as unified meaning representations for texts and images and present an unsupervised inference system that can effectively prove entailment relations between them. We show that by combining semantic parsing and theorem proving, the system can handle semantically complex queries for image retrieval.

1. 本研究の背景と目的

画像、音声、テキスト等のマルチモーダルデータの蓄積が進み、モダリティの異なるデータ間での推論によって新たな知識を獲得するマルチモーダル推論に関する研究が近年盛んになっている。特にテキストと画像の情報を組み合わせた推論タスクとして、画像間の含意関係を推論する Visual Entailment Task [15] や、与えられた画像とその内容に関する質問に答える Visual Question Answering [14]、画像中の物体の数量に関する複雑な質問に答える Tally QA [7] 等が存在する。

高度なマルチモーダル推論を実現するには、テキスト化された知識と画像に含まれる情報を接合する枠組みが必要である。具体例として次の画像とキャプションについて考えてみよう。



- The man is performing an accordion.
- The woman wears a pink dress.
- There is not a green parasol.

図 1: 画像とキャプションの例

この例では、「男性が女性の隣にいる」という情報は画像からしか獲得することができず、「男性がアコーディオンを演奏している」という情報はキャプションにしか書かれていない。しかし、キャプションからの情報と画像からの情報とを組み合わせることができれば、*The man playing an accordion is next to a woman.* といった新しい知識を獲得できる。

形式意味論に基づく論理的意味表現は、表現同士の推論が可能であり、含意関係認識のタスクにおいて高精度を達成しつつある [9, 3]。画像情報を自然言語の意味表現と接続可能な形式で表すことができれば、自然言語テキスト間の推論と同様の推論を、テキストデータと画像情報の間で行うことができる。

自然言語の意味表現を介してテキストデータと画像を接続することは、高精度な物体関係検出・物体検出の研究の進展と、Visual Genome [10] のような画像とテキストの統合的な大規模データセットの登場により、現実的に可能になりつつある。既存の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、グラフ表現を用いる Scene Graph が提案されている [8]。しかし、一般にグラフ表現では物体間の属性や関係に加えて、数量表現や否定といった意味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。

[17] では、文と画像の情報を一階述語論理 (FOL) のモデル

と論理式を用いて表現し、数量や否定を含む複雑な言語現象を伴う文を画像から推論するシステムを提案した。本論文では、(1) FOL モデルを論理式に変換する方法の改良、(2) キャプションの情報と画像情報の合成という 2 点に着目して、より意味的に複雑な文の処理や効率的な論理推論にも対応可能な推論システムの構築を目指す。文と画像を論理式によって統一的に表現することで、意味表現の評価は、画像情報とキャプションを表す論理式とキャプションを表す論理式間の含意関係認識のタスクの評価に帰着させることができる。評価実験により、本研究が提案する論理式による文と画像の意味表現が複雑な言語現象を伴う推論を正確かつ効率的に行うことが可能であることを示す。

2. 文と画像の意味表現

本研究では、属性・関係表現、数量表現、論理表現を伴う意味的に複雑な推論を扱うために、文と画像の意味表現として等号付き一階述語論理 (FOL) のモデルと論理式を用いる。以下で文と画像の意味表現について順に説明する。

2.1 文の意味表現

本研究では自然言語文から FOL 論理式への変換に組合せ範疇文法 (Combinatory Categorical Grammar, CCG) [5, 16] に基づく意味解析・推論システム ccg2lambda [9, 11]^{*1} を用いる。ccg2lambda は自然言語推論のための高階論理をサポートし、前提文から仮説文を自動推論する含意関係認識システムであるが、本研究では意味表現を FOL 論理式に制限し、また文から論理式へ意味解析部のみを使用する。特に、FOL による数量表現の扱いや *part of* のような複合語の処理など意味解析の拡張を行っている。詳細は [17] を参照されたい。

2.2 一階述語論理のモデルによる画像の表現

論理推論に使うことができる画像情報の意味表現としては、一階述語論理のモデル (FOL モデル) を用いる手法がある [4]。FOL モデル \mathcal{M} は、ドメイン D と解釈関数 I から定義される。ドメインは空でない集合であり、画像中に存在するエンティティの情報を表す。解釈関数は各 n 項述語を D^n の部分集合に対応づける関数であり、ドメイン中のエンティティがもつ属性と関係を表現する。例えば、図 1 の画像情報は、

- ドメイン $D = \{d_1, d_2, d_3, \dots\}$
- 属性 $I(\text{man}) = \{d_1\}$, $I(\text{woman}) = \{d_2\}$,
 $I(\text{accordion}) = \{d_3\}, \dots$
- 関係 $I(\text{play}) = \{(d_1, d_3)\}, \dots$

^{*1} <https://github.com/mynlp/ccg2lambda>

からなるモデル $\mathcal{M} = (D, I)$ によって表現することができる*2。

画像情報を FOL モデルとして表現し、保持することには少なくとも 2 つの利点がある。第一に、FOL モデルによる画像情報の表現は、エンティティ・属性・関係からなるグラフによる画像の意味表現 [8, 13] と明確な対応をもち、グラフがアノテートされた大規模データセット [12] から抽出することが可能である。第二に、自然言語文 S を論理式 A に対応付けることができれば、画像を表すモデル \mathcal{M} と論理式 A との間の充足関係 $\mathcal{M} \models A$ をモデル検査 [2] によりチェックすることで画像と文との対応関係を効率的に調べることができる。

2.3 論理式による画像の表現

画像と文間の含意関係だけでなく、画像と文から得られる情報を組み合わせて別の文を推論し、新たな知識を獲得するというハイブリッドな推論を実現するためには、画像とテキストが表す意味情報を何らかの仕方で結合する必要がある。そこで、画像情報を担う FOL モデルを論理式に変換し、意味的に複雑な文の処理や論理推論にも対応可能な推論システムの構築を目指す。具体的には、(i) FOL モデルを論理式に変換する方法、及び、(ii) 文を論理式に変換した後、その論理式の情報 FOL モデルに変換する方法の 2 つを検討する。以下では (i) の方法について詳述し、2.4 節では (ii) の方法について説明する。

FOL モデルを論理式に変換する方法は、(1) ドメインの変換と (2) 解釈関数の変換の 2 つのパートからなる。

(1) ドメインから論理式への変換

モデルでは、「そのドメインにはそこで記述されたエンティティ以外は存在しない」という否定的な情報が含まれる。この情報は、ドメイン $D = \{d_1, \dots, d_n\}$ を (N1) のように原子論理式の連言に変換することで捉えることはできない。

$$(N1) \quad \text{entity}(d_1) \wedge \dots \wedge \text{entity}(d_n)$$

一方、 D を (C1) のように変換することで「エンティティは d_1, \dots, d_n しかない」という情報が含まれるようになる。

$$(C1) \quad \forall x(\text{entity}(x) \leftrightarrow x = d_1 \vee \dots \vee x = d_n)$$

(2) 解釈関数から論理式への翻訳

次に、解釈関数を論理式に変換する。ドメインの場合と同様に、例えば、 $I(\text{cat}) = \{d_1\}$ を $\text{cat}(d_1)$ に変換するだけでは、「 d_1 だけが猫である」という否定的な情報を表すことができず、そのためには、 $\forall x(\text{cat}(x) \leftrightarrow x = d_1)$ という複雑な翻訳が必要となる。一般に F を属性を表す 1 項述語とすると、 $I(F) = \{d_1, \dots, d_n\}$ のとき、単純な翻訳 (N2) と複雑な翻訳 (C2) の二通りの翻訳を検討する。

$$(N2) \quad F(d_1) \wedge \dots \wedge F(d_n)$$

$$(C2) \quad \forall x.(F(x) \leftrightarrow x = d_1 \vee \dots \vee x = d_n)$$

同様に関係を表す 2 項述語 R に対しても、 $I(R) = \{(d_1, e_1), \dots, (d_n, e_n)\}$ のとき、以下の二通りの変換が考えられる。

$$(N3) \quad R(d_1, e_1) \wedge \dots \wedge R(d_n, e_n)$$

$$(C3) \quad \forall x \forall y (R(x, y) \leftrightarrow (x = d_1 \wedge y = e_1) \vee \dots \vee (x = d_n \wedge y = e_n))$$

以上の否定的な情報を扱う (C1)–(C3) の翻訳は、述語サーカムスクリプション (Predicate Circumscription) [6] の一種とみなせる。これは一般に、画像が伝える情報には、「～だけがその属性 (関係) を満たす」といったドメイン・属性・関係の網羅性 (exhaustivity) にかかわる情報が含まれているためである。サーカムスクリプションに基づく論理式への変換により、この情報を明示的に扱うことで、否定や選言など不確定な情報を伴う文と組み合わせた論理推論が可能になる。

*2 画像情報であるため、論理式に名前は現れないと仮定している。また、以下では $I(\text{cat}) = \{d_1\}$ を $(\text{cat}, \{d_1\}) \in I$ とも書く。

1. A が等式以外の原子論理式のとき、 $A \in \mathcal{P}, \neg A \in \mathcal{N}$.
2. A が等式のとき、 $A, \neg A \in \mathcal{P}$.
3. $A \in \mathcal{P}$ かつ $B \in \mathcal{P}$ のとき、 $A \wedge B, A \vee B \in \mathcal{P}$.
4. $A \in \mathcal{N}$ または $B \in \mathcal{N}$ のとき、 $A \wedge B, A \vee B \in \mathcal{N}$.
5. $A \in \mathcal{N}$ かつ $B \in \mathcal{P}$ のとき、 $A \rightarrow B \in \mathcal{P}$.
6. $A \in \mathcal{P}$ または $B \in \mathcal{N}$ のとき、 $A \rightarrow B \in \mathcal{N}$.
7. $A \in \mathcal{P}$ のとき、 $\forall x A, \exists x A \in \mathcal{P}$.
8. $A \in \mathcal{N}$ のとき、 $\forall x A, \exists x A \in \mathcal{N}$.

表 1: 正の論理式 (\mathcal{P}) と負の論理式 (\mathcal{N})

ただし、(C1)–(C3) は複雑な論理式を伴うため、推論において計算コストを要する。そこで、本研究が提案するシステムでは、推論の結論に現れる論理式に応じて、二種類の翻訳を使い分ける。以下で定義する正 (positive) の論理式に対しては単純な (N1)–(N3) の翻訳を、負 (negative) の論理式に対しては、サーカムスクリプションに基づく (C1)–(C3) の翻訳を用いる。

表 1 の規則によって、各 FOL 論理式 $A \in \mathcal{L}$ を正と負の論理式に分類する。FOL 論理式の集合を \mathcal{L} 、正の論理式の集合を \mathcal{P} 、負の論理式の集合を \mathcal{N} とすると、この定義により、 \mathcal{L} は \mathcal{P} と \mathcal{N} に分割されることがわかる。例えば、 $A \text{ cat touches a dog.}$ と $\text{There are two cats.}$ にそれぞれ対応する $\exists x \exists y (\text{cat}(x) \wedge \text{dog}(y) \wedge \text{touch}(x, y))$ と $\exists x \exists y (\text{cat}(x) \wedge \text{cat}(y) \wedge x \neq y)$ は正の論理式であり、 $\text{No cats are white.}$ と $\text{All cats are white.}$ にそれぞれ対応する $\neg \exists x (\text{cat}(x) \wedge \text{white}(x))$ と $\forall x (\text{cat}(x) \rightarrow \text{white}(x))$ は負の論理式である。

2.4 キャプションの利用

画像の内容を表すキャプションには、視覚・空間情報だけでは判断が難しい物体の状態も記述されている。一方で、文だけでは画像の部分的な情報しか記述できず、画像に含まれる情報を網羅的に記述することは難しい。そこで、キャプションの意味情報を取り出し、FOL モデルに変換することで画像情報を拡充することを試みる。FOL モデルにキャプション情報を集約することで、物体の同一性・網羅性を反映した意味表現が得られると考えられる。

ここでは、特に存在量化と連言からなる論理式に翻訳可能なキャプションを FOL モデルに変換する方法について概略を述べる。否定的情報を含む $\text{All men are wearing caps.}$ のような全称文や $\text{There are no white mice.}$ のような否定文は、FOL モデルに変換せず、論理式として保持する。

キャプションから論理式への変換：まず、`cgg2lambda` を用いてキャプションを論理式に変換する。`cgg2lambda` により生成された CCG の導出木から各単語の品詞の情報を取得し、文とその単語の品詞の情報から、WordNet [1] の synset を参照して、単語の意味の曖昧性解消を行う。

モデルとキャプションの対応づけ：画像から抽出された FOL モデルにキャプションの情報を追加するには、キャプションの表現が画像中のどの物体を指しているか、その指示関係を決定する必要がある。ここでは、指示関係が一意に定まる場合にのみ新しい述語をモデルに追加する方法について述べる。以下では、(1) 新たなエンティティを導入する、(2) エンティティの状態・属性を導入する、(3) エンティティ間の関係を導入する、という 3 つのケースに分けて、具体例に基づいて説明する。

(1) エンティティの導入 キャプションが画像情報のドメイン中に含まれない新しいエンティティを導入する場合、そのエンティティをドメインに追加する。その際、キャプションに含まれる述語、またはその類義語・下位語がモデルに含まれる場合に対応する述語を解釈関数に加える。

具体例を (1) に示す。画像情報として (1a) のモデル $\mathcal{M} = (D, I)$ が与えられたとき、(1b) の論理式によって、 \mathcal{M} は (1c)

のモデル $\mathcal{M}' = (D', I')$ に更新される。

- (1) a. $D = \{d_1, d_2\}, I = \{(\text{man}, \{d_1\}), (\text{street}, \{d_2\})\}$
- b. A man is playing music in the street.
 $\exists x \exists y \exists z (\text{man}(x) \wedge \text{street}(y) \wedge \text{in}(x, y) \wedge \text{music}(z) \wedge \text{play}(x, z))$
- c. $D' := D \cup \{d_3\}, I' := I \cup \{(\text{music}, \{d_3\})\}$

この例では music についてモデル \mathcal{M} には記述されていないため、 d_3 という新たなエンティティを割り当てる。一般に視覚的な属性は画像から抽出しやすく、非視覚的（内包的）な属性はキャプションから抽出しやすい。画像とキャプションによりモデルを構築することで、視覚的な属性と非視覚的な属性をどちらも記述することが可能となる。

(2) 属性・状態の導入 以下の (2a) のモデルでは、ドメインに 1 匹の猫が存在し、(2b) のキャプションは 1 匹の猫は白という情報をもつ。この場合、(2c) のようにモデルを更新する。

- (2) a. $D = \{d_1\}, I = \{(\text{cat}, \{d_1\})\}$
- b. A cat is white.
 $\exists x (\text{cat}(x) \wedge \text{white}(x))$
- c. $D' := D, I' := I \cup \{(\text{white}, \{d_1\})\}$

猫が複数いる場合、同様の更新を行うことはできない。以下の例では、キャプションは「1 匹の猫は白」という意味であるにもかかわらず、(3c) のモデルでは「全ての猫は白」という情報が更新されており、不適切である。

- (3) a. $D = \{d_1, d_2, d_3\}, I = \{(\text{cat}, \{d_1\})\}$
- b. A cat is white.
 $\exists x (\text{cat}(x) \wedge \text{white}(x))$
- c. $D' := D, I' := I \cup \{(\text{white}, \{d_1, d_2, d_3\})\}$

(3) 関係の導入 キャプションが物体間の関係を表す 2 項述語を含む場合、その項となるエンティティがモデル中で一意に定まる場合はモデルを更新する。

- (4) a. $D = \{d_1, d_2\}, I = \{(\text{man}, \{d_1\}), (\text{guitar}, \{d_2\})\}$
- b. A man is playing a guitar.
 $\exists x \exists y (\text{man}(x) \wedge \text{guitar}(y) \wedge \text{play}(x, y))$
- c. $D' := D, I' := I \cup \{(\text{play}, \{d_1, d_2\})\}$

2 つの物体が一意に決まらない場合は WordNet を用いて、キャプションに現れた述語の上位語・下位語の情報を補完する関係のアブダクション [17] を行い、空間情報をもとにモデルへの 2 項述語の追加を行う。

3. 評価実験

3.1 実験設定

本研究では画像に対し、モデル、キャプション 2 文（画像に対し真、偽となる文）を付与した GRIM データセット [4] を実験に用いる。GRIM におけるモデルには物体、属性、物体間の空間情報が記述されており、空間情報を表すために touch, near, support, occlude, part_of の 5 つの関係が用いられている（図 2）。本研究では、GRIM のデータ 200 件のうち、ノイズを含む 6 件を除く 194 件を用いて実験を行う。

モデルや文から論理式への変換法、また実験設定は基本的に [17] に従う。研究 [17] では、複雑な言語現象を含む 19 文に対し、GRIM 上の画像からそれらの文が含意するか否かの情報を付与している。扱われている言語現象として、論理結合子、数詞、量化詞、空間関係の分類が設けられており、評価実験では、画像（に対応する論理式）と文（に対応する論理式）間の含意関係を判定することにより、正解画像の検索能力を F 値で報告している。表 2 に文の例と言語現象の分類を示す。本研究でも同様に、言語現象の分類に基づき F 値を報告する。本研究では研究 [17] の手法を基に、2. 節で提案したモデルの変換方法（サーカムスクリプション）や画像に付与されたキャプション情報を用いた拡張を行った場合の性能改善を評価する。


data/bernese-mountain-dog-111878_640	model
	model([d1,d2,d3,n1,n2], [(f1,n_cat_1,d1)], [(f1,n_dog_1,d2)], [(f1,n_tree_1,d3)], [(f1,n_head_1,n1,n2)], [(f1,a_gray_1,d1)], [(f1,a_black_1,d2)], [(f1,a_brown_1,d3)], [(f1,n_vascular_plant_1,d3)], [(f1,n_placental_1,d1,d2)], [(f1,n_woody_plant_1,d3)], [(f1,n_external_body_part_1,n1,n2)], [(f1,n_whole_2,d1,d2,d3)], [(f1,n_object_1,d1,d2,d3)], [(f1,n_thing_12,n1,n2)], [(f1,n_organism_1,d1,d2,d3)], [(f1,n_physical_entity_1,d1,d2,d3,n1,n2)], [(f1,n_carnivore_1,d1,d2)], [(f1,n_body_part_1,n1,n2)], [(f1,n_vertebrate_1,d1,d2)], [(f1,n_entity_1,d1,d2,d3,n1,n2)], [(f2,s_part_of_1,n1,d2),(n2,d1)], [(f2,s_touches_1,d3,d1)], [(f2,s_supports_1,d3,d1)], [(f2,s_occludes_1,d1,d3)])]
True:	A cat is sitting on a table. A dog is standing near a table. The dog is looking at the cat.
False:	A cat is looking at a dog. A dog is sitting on a table. The dog is chasing the cat. The dog is touching the cat.

図 2: GRIM のデータ例

3.1.1 サーカムスクリプションによる解析性能の評価

2. 節で述べたサーカムスクリプションを用いて GRIM 上のモデルを論理式に変換した場合における、画像検索の精度と速度を評価する。既存研究と同様に F 値と、定理証明にかかった平均時間（秒）を報告する。

また、このときに定理証明器の選択による検索性能への影響を評価する。ベースラインとして [17] 同様に、自動証明器に Prover9^{*3} を用いるが、本研究ではそれに加え、入力 of 定理に対する反例を探索する Mace4^{*3} を同時に実行する場合を評価する。定理によっては反例を見つけるほうが容易であり、Mace4 を用いることで、画像と文が関係ない場合を高速に判断できる。本研究ではまた、高速さで知られる Vampire^{*4} を用いた場合の性能も報告する。Vampire 内部では、定理証明と反例の探索が同時に行われる。

3.1.2 キャプションからモデルへの変換の評価

キャプションから FOL モデルへの変換の正しさを、GRIM のキャプションを用いた含意関係認識を行うことで評価する。具体的には、GRIM の画像に付与されている 2 つのキャプションのうち、真となる文のみを用いてモデルを拡張し、拡張されたモデルから 2 つのキャプションに対応する論理式への含意関係が成り立つかを調べ、F 値を算出する。

本研究ではさらに、キャプション情報の有用性を検証するため、3.1.1 節の実験と同様に画像検索の実験を行う。具体的に、研究 [17] 同様に、表 2 下のような文を人手で 8 件作成し、それらに対して GRIM データセットの 200 件の画像がそれぞれ含意するかをアノテートした。この新たに作成したデータセットを用いて画像検索の実験を行い、その検索性能を報告する。作成した文の特徴として、広範な種類の関係述語を含むことであり、これらの文をクエリとして画像検索を行うには、GRIM のモデルに記述された物体間の位置関係以上の情報をモデルに加える必要がある。

3.2 実験結果と考察

サーカムスクリプションによる解析性能の評価結果を表 3 に示す。サーカムスクリプションにより、特に every や all など量化を含む文による検索の精度の改善が見られた。また論理結合子のうち、not など否定を含む文の精度も大きく改善した。次に、定理証明器の設定による精度と速度の比較を表 4 に示す。いずれの分類においても Prover9 を単独で用いた場合と、Prover9 に Mace4 を組み合わせた場合の F1 は等しいが、速度については後者において著しく改善している。Vampire を用いた場合も同様に、反例モデルの探索により、速度の改善が見られ、特に「論理結合子」に分類される文では高速に証明が得られることがわかる。

キャプションより更新されたモデルを用いて、各画像ごとに真と偽の 2 種類のキャプションとの含意関係を判定した結果は、適合率、再現率、F1 それぞれ 0.91, 0.73, 0.81 であった。

^{*3} <https://www.cs.unm.edu/~mccune/prover9/>

^{*4} <http://www.vprover.org/>

例文	現象
There is a cat or dog.	論理結合子
There are at least two cats.	数詞
Every person is touching a bicycle.	量化、空間関係
A cat is walking .	関係一般
A person is wearing a hat.	関係一般

表 2: 実験で用いた文の例と言語現象の分類。前半の例は研究 [17] によるが、下 2 例は実験で用いるために本研究で GRIM 画像との含意関係を付与した。

手法	F 値 / 速度 (秒)				
	論理結合子	数詞	量化	空間関係	全体
単純	0.68 / 8.9	0.81 / 8.8	0.0 / 10.4	0.73 / 9.3	0.74 / 9.0
Circum.	0.84 / 12.1	0.95 / 9.2	0.76 / 35.0	0.88 / 13.3	0.88 / 11.9

表 3: モデルの翻訳方法の比較。Circum. はサーカムスクリプションの略。定理証明は Prover9 による。

再現率が低い原因として、式 4 のような例において、キャプションに含まれる物体が画像中のどの物体を指すかが一意に決められず、モデルの更新に失敗したケースが見られた。

関係一般を含む文による画像検索の結果を表 5 に示す。キャプション情報を用いないモデルでは空間情報しか記述されていないため、表 2 の下 2 文のような一般の関係を含むクエリ文に対する F 値は 0 になるが、キャプション情報をモデルに加えることでそのような文についても検索できるようになった。再現率が低くなった原因は次の 2 つが考えられる。

第一に、画像情報からは推測されるがキャプションに含まれていない述語は検索に用いることができなかったことが挙げられる。特に物体の状態を表す述語はキャプションにしか記述されておらず、データセットに用意されている約 3 文のキャプションでは画像情報を表現しきれていないことが分かる。今後は Visual Genome [12] のような物体の属性や状態を含む大規模なデータセットを用いた実験を検討する。

第二に、本実験ではアブダクションは行わなかったため、述語間の関係を捉えた検索できなかったことが挙げられる。例えばモデルに *man* がある場合、同時に *person* がいるという推論ができなかった。これは [17] と同様の手法で証明時にアブダクションを行うことで解決可能と考えられる。

There is a cat which is not white. と *A person is riding a bicycle.* に対するシステムの予測画像を図 3 に示す。図 3(a) は、否定表現についてシステムが正しい予測を行った例である。また図 3(b) から *ride* という物体間の関係を表す述語を正しくモデルに変換できていることが分かる。



(a) *There is a cat which is not white.*



(b) *A person is riding a bicycle.*

図 3: システムの予測画像

4. おわりに

本稿では、画像情報とテキスト情報を論理ベースの意味表現を用いて統一的に扱い、論理推論を行うシステムを提案した。

手法	論理結合子	速度 (秒)			
		数詞	量化	空間関係	全体
Prover9	12.1	9.2	35.0	13.3	11.9
+ Mace4	11.0	7.8	12.3	8.4	9.7
Vampire	9.3	9.0	10.3	9.2	9.2

表 4: 定理証明器による証明速度の比較。この実験ではモデルの変換にサーカムスクリプションを用いる。F 値については定理証明器によらず等しい。

適合率	再現率	F1
0.86	0.36	0.49

表 5: キャプションを追加した場合の画像検索の性能。

このシステムは、CCG の意味解析と定理証明を組み合わせることで、文と画像に含まれる属性・関係だけでなく、否定や量化・数量表現を含む意味情報を扱うことが可能である。画像に対してエンティティ・属性・関係の情報がアノテートされたデータセットを用いて、意味的に複雑なクエリを用いた画像検索が実現可能であることを示した。また、キャプションと画像情報を合成する方法を提示した。今後はグラフ表現から FOL モデルの変換をシステムに組み込むことで、より大規模なデータセット [12] を利用したマルチモーダル推論を検討する。

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] Miller George A. WordNet: A Lexical Database for English. *Commun. ACM*, 1995.
- [2] Patrick Blackburn and Johan Bos. *Representation and Inference for Natural Language*. CSLI, 2005.
- [3] Abzianidze Lasha. LangPro: Natural Language Theorem Prover. In *EMNLP*, 2017.
- [4] Hürilman Manuela and Johan Bos. Combining Lexical and Spatial Knowledge to Predict Spatial Relations between Objects in Images. In *Vision and Language Workshop*, 2016.
- [5] Steedman Mark. *The Syntactic Process*. MIT Press, 2000.
- [6] John McCarthy. Applications of circumscription to formalizing common sense knowledge. *Artificial Intelligence*, 1986.
- [7] Acharya Manoj et al. TallyQA: Answering complex counting questions. In *AAAI*, 2019.
- [8] Justin Johnson et al. Image retrieval using scene graphs. In *CVPR*, 2015.
- [9] Koji Mineshima et al. Higher-order logical inference with compositional semantics. In *EMNLP*, 2015.
- [10] Krishna Ranjay et al. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *Comput. Vision*, 2017.
- [11] Pascual Martínez-Gómez et al. ccg2lambda: a compositional semantics system. In *ACL*, 2016.
- [12] Ranjay Krishna et al. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. 2016.
- [13] Schuster Sebastian et al. Generating Semantically Precise Scene Graphs from Textual Descriptions for Improved Image Retrieval. In *Vision and Language Workshop*, 2015.
- [14] Stanislaw Antol et al. VQA: Visual Question Answering. In *ICCV*, 2015.
- [15] Vu Hoa et al. Grounded Textual Entailment. In *COLING*, 2018.
- [16] 戸次大介. 日本語文法の形式理論: 活用体系・統語構造・意味合成. くろしお出版, 2010.
- [17] 鈴木莉子, 谷中瞳, 峯島宏次, 戸次大介. 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. 言語処理学会第 25 回年次大会発表論文集, 2019.