画像内物体間の視覚的関係の真偽判定データセット

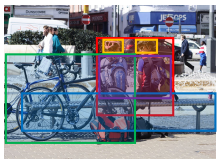
竹林 佑斗 † Chenhui Chu‡ 中島 悠太 ‡
† 大阪大学大学院情報科学研究科
‡ 大阪大学データビリティフロンティア機構
takebayashi.yuto@ist.osaka-u.ac.jp {chu, n-yuta}@ids.osaka-u.ac.jp

1 はじめに

画像内の物体間には様々な視覚的関係が存在する．例えば，図 1 では “two bikers” と “a bench” の関係は “are sitting on” や “are sitting and talking on” であり，“two bikers” と “bike gear” の関係は “in” や “dressed in” である．そのような様々な視覚的関係を認識できれば画像全体の理解に繋がる．画像内物体間の視覚的関係を認識するために，既に Visual Phrases[4]，Scene Graph[2]， Visual Relationship Detection (VRD)[1] と Visual Genome[2] という 4 つのデータセットが構築されている．既存のデータセットでは視覚的関係を (subject; predicate; object) というトリプレットで表す．ここの subject，object は画像内の物体で，predicate は “under” や “in front of” のような位置関係を表す前置詞または “hold” や “ride” のような動詞である．

Visual Phrases の問題点としては視覚的関係の種類が限られていること (13 種類のみ) があげられる．Scene Graph，VRD および Visual Genome は物体間の視覚的関係の種類は多いが，真の視覚的関係[3]しかアノテーションされていない．応用先によってはありえない（偽）関係も重要である．例えば，視覚に基づく言い換え認識 [1] では偽の視覚的関係を言い換えの候補として除く必要がある．また名詞句で表す物体間の視覚的関係の真偽を認識できれば，その名詞句の係受け解析にも役に立つ．

本研究では Flickr30k エンティティ [3] という画像キャプションのデータセットを用いて，キャプションに書かれている 2 つの物体間のあらゆる視覚的関係に対して真偽のアノテーションを行う．まずアノテーションする視覚的関係候補を抽出し，係受け解析を用いて真の関係を抽出する．係受け解析による関係抽出ができなかった残りの候補に対して，クラウドソーシングを行い真偽のアノテーションを行う．これにより，視覚的関係を網羅する上，偽の関係を含むデータセットが構築できた．このデータセットは画像理解を大きく前進させるだけではなく，自然言語処理にも役立つ．



1. **Two bikers** in **bike gear** are sitting on **a bench** .
2. **Two people** rest on **a park bench** next to **their bikes** .
3. **Women** in **bike helmets** take **break** from **long ride** .
4. **Two bikers** are sitting and talking on **a bench** in front of **their bikes** .
5. **A bike riding couple** dressed in **bike gear** and **helmets** take **a minute** to site on **a bench** to talk and park **their bikes** .

図 1: Flickr30k のデータ例．1 画像について 5 文キャプションがつけられている．キャプションの中の物体にはエンティティが付与されており，それぞれ画像内での領域に対応付けられている．アンダースコアで示されている部分は関係である．

2 データセットの作成

データセット作成の流れを図 2 に示す．まず Flickr30k のキャプションから，前処理を行って視覚的関係候補を抽出する．またキャプションからは係受け解析を行い，エンティティ・関係間の係受け関係グラフを構築しておく．係受け関係グラフがあるタイプに従い，かつ，その視覚的関係が前処理で得られた視覚的関係候補に含まれているとき，それを係受け解析結果（真の視覚的関係）として抽出する．これを**タイプ抽出**と定義する．また，視覚的関係候補に含まれていて，タイプ抽出で得られなかった残りの視覚的関係はクラウドソーシングによりアノテーションされる．

2.1 Flickr30k エンティティデータセット

本研究で作成したデータセットは，Flickr30k エンティティデータセット [3] を元にして作成した．Flickr30k とは，画像内の領域とキャプション内のエンティティ間の対応を付けた，大規模な画像とキャプションのデー

[1] https://cs.stanford.edu/people/ranjaykrishna/vrd/
[2] https://visualgenome.org/
[3] 画像と照らし合わせたときに意味をなす視覚的関係を真の視覚的関係，そうでないときを偽の視覚的関係と定義する．

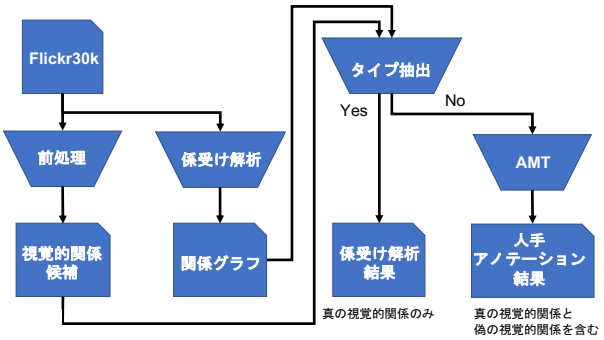


図 2: データセット作成の流れ．係受け解析結果とクラウドソーシング結果が最終的なデータセットとなる．

タセットである．本稿では訓練，開発および評価データをそれぞれ 29,769 画像，1,000 画像および 1,000 画像とした．図 1 に Flickr30k エンティティの画像とキャプションの例を示す．1 つの画像にその画像に言及する 5 文のキャプションが付いており，画像内の領域とキャプション内のエンティティが対応付けられている．この例では画像の領域と文のエンティティが 1 対 1 対応している．イベントやシーンである “break” や “a minute” はキャプション内に現れても画像内に対応する領域はないため，キャプション間の対応のみが得られる．

2.2 前処理による視覚的関係候補の抽出

Flickr30k のキャプションの名詞句にはエンティティが付与されている．以下の例のようにエンティティを 2 つ組み合わせて視覚的関係を作成する．このときエンティティの順番が，文に出現する順と逆にならないという条件のもと，考えられる全てのエンティティペアを組み合わせる．エンティティをペアにする際，2 つのエンティティの間に入る句を関係と呼び，後ろのエンティティの直前に位置する句を関係とする．以下の例では，3 つの視覚的関係候補が抽出される．

[/EN#1/people Two bikers] in [/EN#2/other bike gear] are sitting on [/EN#3/other a bench].

⇓

1. Two bikers in bike gear
2. Two bikers are sitting on a bench
3. bike gear are sitting on a bench

関係が「,」「, and」「and」「while」「 」(空白) の視覚的関係はエンティティを並列関係で接続しているため必ず成り立つ．したがって，これらの関係を持つ視覚的関係は候補から除いた．またエンティティにはその特性を示したタグ（“people” や “scene”）が付与されているが，そのタグが “notvisual” であるものは視覚的関係作成の候補から除いた．前処理によって得られた視覚的関係の統計データを表 1 に示す．

表 1: 前処理によって得られた視覚的関係数．

	訓練	開発	評価
視覚的関係数	249,706	8,425	8,292

2.3 係受け解析による関係抽出

キャプションが画像の視覚的関係を正しく表し係受け解析が正しければ，エンティティの間の係受け関係より真の視覚的関係が抽出できる．この節では係受け解析を用いて真の視覚的関係を抽出する方法を説明する．キャプションから Stanford parser[4]を用いて係受け解析し，2.2 節で得られた視覚的関係候補と照らし合わせ，真の視覚的関係を抽出する．真の視覚的関係抽出は以下のステップで行う．

1. 平文から係受け解析を行い，単語の有向グラフを作成する[5]．
2. Flickr30k のエンティティと関係を元に，それらのノードをマージする．ここで同じノード間に張られたエッジは無視する．
3. エンティティと関係があるタイプに従うとき，その視覚的関係を得る．
4. 得られたものの中で，前処理において得られた視覚的関係候補に含まれるものを抽出する．

ステップ 3，4 がタイプ抽出である．ステップ 3 のあるタイプとは，ペアにしたいエンティティとその関係がどのような係受け関係になっているかである．係受け解析にはエラーが含まれるため，実験的にタイプの組み合わせを決める．2 つのエンティティを EN1・EN2，その関係を RE と表すとき，それらが無関係のノードを間に含まない場合，以下の 7 個の組み合わせが考えられる．ここで EN1 と EN2 が文に出現する順番は区別しないものとする．

- A: RE -> EN1, RE -> EN2
- B: EN1 -> RE, EN1 -> EN2
- C: EN1 -> RE, EN2 -> RE
- D: RE -> EN1, EN2 -> EN1
- E: EN1 -> RE -> EN2
- F: RE -> EN1 -> EN2
- G: EN1 -> EN2 -> RE

[4]https://nlp.stanford.edu/software/lex-parser.shtml
[5]collapsed-dependencies の関係を有向グラフとした

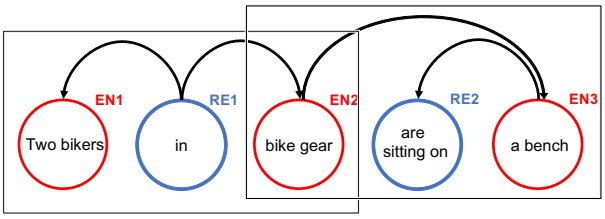


図 3: エンティティと関係別にノードをマージしたグラフ．“Two bikers are sitting on a bench” は抽出できない．

表 2: A〜G の 1 つを用いた場合とそれらを組み合わせた場合の Precision/Recall/F-score．

Type	TP	FP	FN	Precision	Recall	F-score
A	15	1	138	93.8	9.8	17.8
C	5	1	148	83.3	3.3	6.3
E	41	7	112	85.4	26.8	40.8
G	23	1	130	**95.8**	15.0	26.0
AG	37	2	116	94.9	24.2	38.5
AEG	76	8	77	90.5	**49.7**	**64.1**
ACEG	76	8	77	90.5	**49.7**	**64.1**

A と B は共通の親ノードを持つもの，C と D は共通の子ノードを持つものである．

例として “[/EN#1/people Two bikers] in [/EN#2/other bike gear] are sitting on [/EN#3/other a bench].” から係受け解析を行い，エンティティと関係をそれぞれ 1 つのノードにマージしたグラフを図 3 に載せる．この例では A タイプの “Two bikers in bike gear” と G タイプの “bike gear are sitting on a bench” が抽出できる．後者は係受け解析の失敗例である．そのほかに前処理では “Two bikers are sitting on a bench” が候補になるが，ここでは A から G までのどのタイプにも属さないため抽出できない．

A〜G の 1 つを用いた場合とそれらを組み合わせた場合の Precision/Recall/F-score を表 2 に示す．評価には著者自ら開発データからアノテーションした 186 件のデータを使った．このフェーズでは Precision が重視されるため Precision が大きいものから 1 つずつ増やして組み合わせた．タイプ B，D，F は真陽性と偽陽性の個数が共に 0 であったため計算できなかった．結果より Precision が 90.5%，Recall も 49.7%と高いため，本稿では AGE の 3 つを用いることとした．今回実験を行なった 186 件のデータではタイプ C で抽出できる視覚的関係はタイプ E に全て含まれていた．係受け解析によって得られた視覚的関係数を表 3 に示す．

表 3: 係受け解析によって得られた視覚的関係数．

	訓練	開発	評価
視覚的関係数	90,665	3,015	2,933

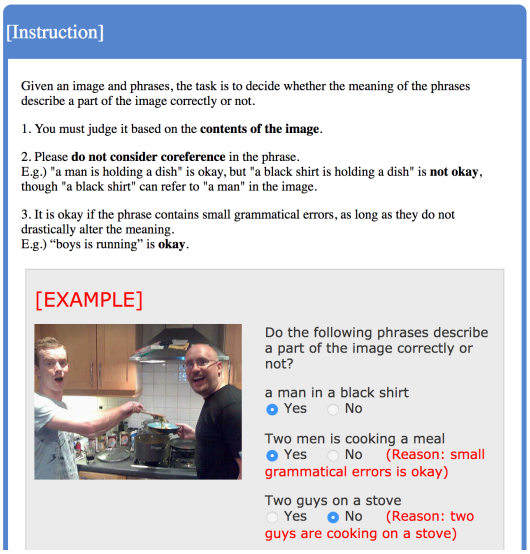


図 4: インストラクションのキャプチャ．

2.4 AMT によるアノテーション

係受け解析で抽出できなかった視覚的関係について，Amazon Mechanical Turk (AMT) を用いてアノテーションを実施した．クラウドソーシングを依頼する際に使用したインターフェースの設計について説明する．インストラクションのキャプチャを図 4 に示す．作業者にはまずこの画面が表示される．一番目には指示を書き，二番目には実際のアノテーションの例を，正解と不正解にそれぞれ理由を付けて載せている．最後にダミー問題を入れること，その結果により却下することを明記してインストラクションを終える．全く関係のない他の画像から視覚的関係を抜き出し加えることで，ダミー問題を作成した．ダミー問題は 1 HIT に 5 個入っており，これの正解率が 0.8 以上なら承認，それよりも低ければ却下した．図 5 の例では 2 つ目の視覚的関係 “man with fish” がダミー問題である．

下にスクロールしていくと，図 5 のようなパネルが約 10 枚表示される．アノテーションされる視覚的関係は，1 枚の画像につき 1〜10 個あり，合計で 50 個になるように HIT を設計した．これがアノテーション対象のデータであり，作業者は Yes/No のラジオボタンを全て押すことで結果を送信できる．AMT の設定としては 1 HIT に対する報酬は$0.2 に設定し，作業者にはマスターを指定した．

アノテーション作業は以下の流れで進めた．

1. AMT にて HIT を発行する．
2. 作業者が AMT 上でアノテーション作業を行う．
3. 作業者の回答が出たのち，ダミー問題を使い，作業者の正解率を測定する．

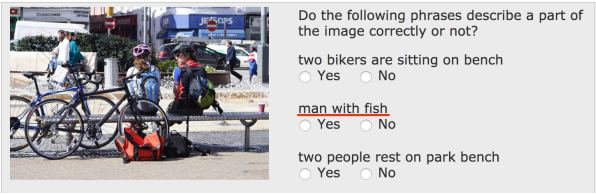
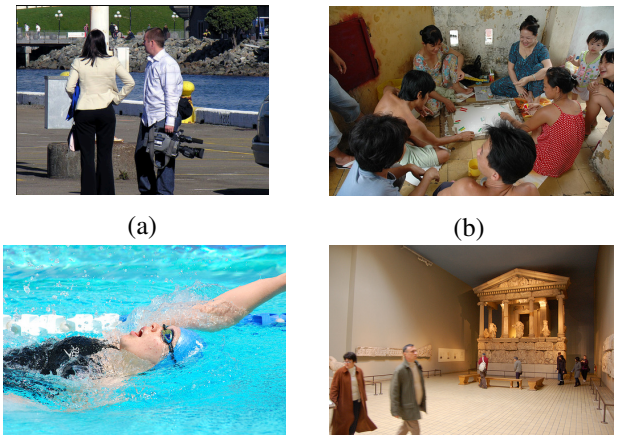


図 5: インターフェースのキャプチャ．



(a) (b)

(c) (d)

図 6: アノテーション失敗例の画像．

4. 正解率が閾値以上であれば承認，下回っていれば却下する．
5. 承認された作業者には報酬が支払われる．却下された作業者には報酬は支払わずその HIT は評価に加えなかった．
6. 作業者の回答を平均し，Yes の数が過半数以上であればその視覚的関係は真，それ以外の場合は偽であるとした．

表 4 に評価データをクラウドソーシングした結果を示す．精度を測るために著者らがアノテーションした，149 件のデータを使用した．結果としては，F-score が 94.9 と高い値を示した．

表 4: AMT でクラウドソーシングをした結果の精度

Accuracy	Precision	Recall	F-score
90.2	96.3	93.5	94.9

3 データセットの分析

図 6 の (a)，(b) に係受け解析での失敗例を示す．画像 (a) の “video camera standing next to woman” は偽の視覚的関係であるが，係受け解析によって真と判定されたしまった．これは係受け解析が誤っている例である．画像 (b) の “group of people playing board game” は真の視覚的関係であるが，係受け解析では抽出できなかった．画像 (a) のように係受け解析が誤って真の視覚的関係として判定した場合，クラウドソーシングに回されないため，そのエラーは訂正できない．この問題への対処は今後の課題である．

図 6 の (c)，(d) にクラウドソーシングでの失敗例を示す．画像 (c) の “backstroke in swimming pool” は “backstroke”（背泳ぎ）という物体は存在しないため偽であるが，100% の作業者が真とした．作業者にはこのような例を偽にするようインストラクション（図 4 のインストラクション 2）で言及しているが，不十分であったと言える．画像 (d) では “museum from antiquity” は画像と一致しないため偽であるが，8 割の作業者が真とした．古いものを展示しているのであり，博物館が古い訳ではない．これは詳細に注意していないためであると考えられる．

4 おわりに

視覚的関係の認識は画像全体の理解に繋がる．本研究では画像内の 2 物体間の視覚的関係の真偽データセットを作成した．Flickr30k のデータから係受け解析と AMT でのクラウドソーシングを行うことで，視覚的関係を網羅する上，偽の関係を含むデータセットの構築ができた．今後，このデータセットを視覚的関係認識のモデル学習や視覚に基づく言い換え認識に使用する予定である．

謝辞

本研究は，JST，ACT-I の支援を受けたものである．

参考文献

[1] Chenhui Chu, Mayu Otani, and Yuta Nakashima. iParaphrasing: Extracting visually grounded paraphrases via an image. In *COLING*, pp. 3479–3492, 2018.

[2] J. Johnson, R. Krishna, M. Stark, L. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei. Image retrieval using scene graphs. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 00, pp. 3668–3678, June 2015.

[3] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, pp. 2641–2649, 2015.

[4] Mohammad Amin Sadeghi and Ali Farhadi. Recognition using visual phrases. In *CVPR*, pp. 1745–1752. IEEE Computer Society, 2011.