ソーシャルメディアのための柔軟なビジュアルグラウンディング

Yongmin Kim[1] Chenhui Chu[1] 加藤 圭造[1,2] 黒橋 禎夫[1]
[1] 京都大学
[2] 富士通研究所
{kim, chu, kuro}@nlp.ist.i.kyoto-u.ac.jp
kato.keizo@fujitsu.com

1 はじめに

ビジュアルグランディングとは画像の中にあるクエリに対応する領域を探すタスクである [1]．画像・言語のマルチモーダル学習の分野において，基礎的な研究分野の一つである．このタスクにはデータセットによってクエリの形態が違う．例えば，Flickr30k entities データセット [2] は一枚の画像に複数キャプションがあり，そのキャプションの中にある名詞句と画像の中の領域の対応関係がアノテーションされている．一方，RefCOCO+[3] データセットのクエリは画像の中に存在する物体を説明するものであり，Visual7W[4] データセットのクエリは画像内のある領域に関する質問からなっている．

上記のような人為的に作られたデータセットには当然ながら，クエリに対して必ず対応する領域が画像の中に存在する．しかし，実世界のデータはクエリに対応する画像中の領域がない場合が多く存在する．このようなデータの一つとしてソーシャルメディアデータが挙げられる．画像とそれに関するテキストからなるソーシャルメディアデータでは，テキスト中の名詞句がグランウンディングできる場合もあるが（図 1 の左），できない場合もある（図 1 の右）．

図 1 のようにクエリがグラウンディングできないことを扱うためにはクエリがグラウンディングできるかないかを判断し，できる場合はグラウンディングし，できない場合はできないことを柔軟に答えることが必要である．しかし，従来のビジュアルグラウンディング手法 [1, 2, 5] では入力クエリに対して，必ずグラウンディングする領域を選択しなければならないので，グラウンディング不可能な場合を取り扱うことができない．

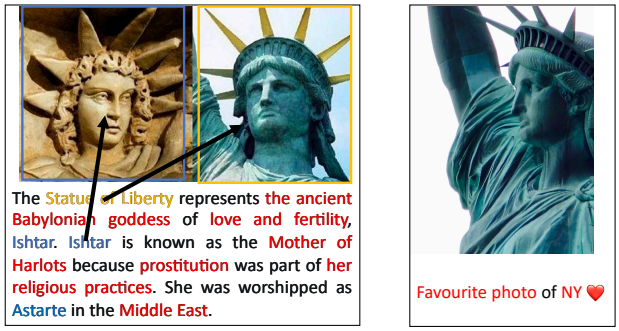

図 1 ソーシャルメディアデータにおけるビジュアルグランディングの例．左のクエリの一部はグラウンディングできるが，右のクエリはグランディングできない．

本研究では，グラウンディング不可能な場合に対して，擬似的な領域を候補領域として追加することで対応する手法を提案する．グラウンディング不可能な学習データを RefCOCO+から擬似的にグラウンディングできない擬似データセットを作成して，モデルを学習する．その結果，グラウンディングできないデータを高い精度で検出できることが確認された．

また，Twitter からデータを収集し，ビジュアルグラウンディングのデータセットを構築する．提案手法をソーシャメディアで検証するためには用いたモデルの学習データである RefCOCO+のような画像である必要がある．従って，画像認識モデル [6, 7, 8] を用いて，収集した tweet をフィルタリングを行い，RefCOCO+のような検証用の小規模なデータセットを作成し，提案手法の有効性の検証を行った．

2 提案手法

従来のビジュアルグラウンディングモデルは図 2 に示されるように，まず画像から各候補領域に対して，特徴ベクトル $f_v \in \mathbb{R}^{d_v}$ と空間ベクトル $f_s \in \mathbb{R}^5$ を抽出する．各候補領域とそれの特徴ベクトルは

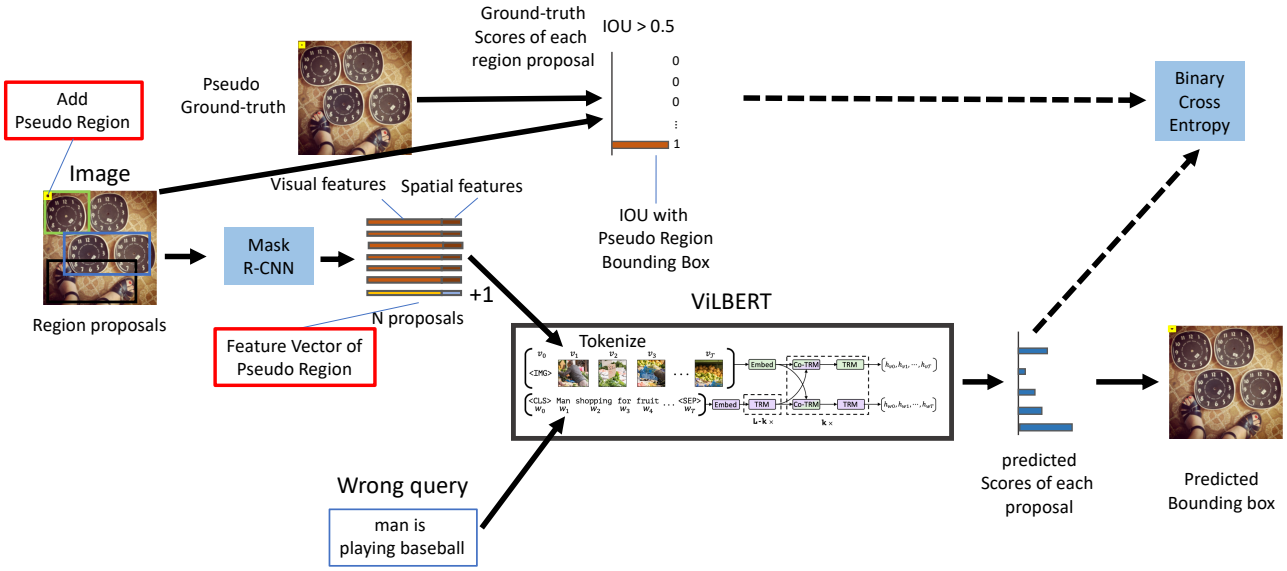

図 2 提案手法の概要．グラウンディングできないクエリに対する擬似的な候補領域（pseudo region）を追加し，その領域を選択されるようにモデルを学習させ，グラウンディングできないクエリに対応する．

Mask R-CNN[9] のような物体検出モデルを用いて，抽出する．空間ベクトルは式 1 のように正規化された各領域の 4 つの頂点の座標とその幅と縦の長さで構成されている．

$$f_s = \left[\frac{w_{tl}}{W}, \frac{y_{tl}}{H}, \frac{x_{br}}{W}, \frac{y_{br}}{W}, \frac{wh}{WH}\right] \tag{1}$$

(w_{tl}, y_{tl}) と (y_{tl}, x_{br}) はそれぞれ領域の左上と右下の座標である．次に，このように抽出された候補領域の画像のベクトルとクエリを比較し，類似スコアを出力して，最終的にスコアが最も高い領域を選択する．ViLBERT の場合，モデルが i 番目の候補領域の画像のベクトル v_i とクエリ Q から，最終的に出力する表現ベクトル h_{vi} を $W_r \in \mathbb{R}^{d\times 1}$ 線形投影し，類似スコアを出す．

$$\text{ViLBERT}(v_i, Q) = W_r h_{vi} \tag{2}$$

学習する際にはこの各候補領域から生成される表現ベクトルが，類似スコアを予測するように学習される．

本研究では候補領域に容易にグラウンディングできないクエリに対する擬似的な領域を入れ込むことで，グラウンディングできないクエリに対し，この擬似領域を出力して，対応できるような手法を提案する．この擬似領域は特徴ベクトル f_v の全ての値を 1 にして，空間ベクトル f_s の値を全部 0 に設定した．すなわち，擬似領域は画像の左上に位置して，広さが 0 であり，特徴ベクトルの全ての値が 1 となる．

3 ソーシャルメディアデータの構築

3.1 クローリング

ソーシャルメディアデータセットは Twitter からあるキーワードを検索し，クローリングした．キーワードはクローリングしたデータが学習データである RefCOCO と類似性を持つために, RefCOCO のカテゴリから取得した．

3.2 画像データのフィルタリング

上記のクローリングで多くの tweet が得られたが，提案手法を評価するには学習に用いた RefCOCO+ のような画像をもつデータが求められるため，RefCOCO+に存在しない特性のデータを除去する必要がある．このような特性のデータの画像には RefCOCO カテゴリと違う画像の他に，低画質画像，広告，テキストが多い画像，絵などが考えられる．従って，事前学習された画像分類モデル [6]，物体検出モデル [7] を用い，最後に画像内のテキストを検出して，その領域と文字を認識するテキスト検出モデル [8] を用いて，フィルタリングを行う．

画像分類モデルは ImageNet[10] データセットから事前学習されたモデル [6] を用いた．モデルから出力されたクラスを RefCOCO+のカテゴリとの Wu & Palmer similarity（式 3）を計算し，閾値以上のものを収集ことで RefCOCO+と関連性が高い tweet が得られ，ウェブサイトや漫画など関連のないクラスが除

去された．閾値は ImageNet クラスと RefCOCO+カテゴリとの Wu & Palmer similarity をみて，RefCOCO+のカテゴリと関連性が低いものが含まらないように 0.95 の以上にした．

$$\text{Wu} - \text{Palmer} : 2 * \frac{depth(lcs(s1, s2))}{depth(s1) + depth(s2)} \quad (3)$$
$$lcs : Least\ Common\ Subsumer$$

物体検出アルゴリズムは COCO データセット [11] から事前学習されたモデル [7] を用いた．COCO データセットは RefCOCO の基盤となるデータであり，両データセットは同一のクラスを保有する．tweet データには背景がなく，物体一個しか存在しないグラウンディングの意味のないデータを存在する．そのため，物体検出モデルを用いて，検出される物体が 2 個以上存在するデータを収集した．

上記の二つのフィルタリングの結果，残った tweet には画像の中でテキストが占める割合が多いデータがまだ存在していた．このようなデータはテキスト検出モデル [8] を用いて，その領域の全体画像の大きさに対する割合を計算し，閾値を基準にデータを整理した．閾値は背景に存在する文字と画像上に追加された文字を区分できるような値をデータの観察より，0.05 以下に設定した．

3.3 アノテーション

このようなフィルタリング手法を通して，前より RefCOCO+に類似な tweet を得ることができた．tweet のテキストには複数の記号，絵文字，リンクなどを含んでいるため，前処理を行い，そのテキストから文を抽出した．また，それぞれの文の名詞句をチャンキングし，クエリを抽出する．その後，クエリに対する領域にアノテーションを行い（図 3），最終的に提案手法を検証するための 75 個のクエリのソーシャメディアデータセットを構築した．このデーアセットは画像と 15 個のグラウンディング可能なクエリと 60 個のグラウンディング不可能なクエリから構成されている．

4 実験設定

4.1 擬似データセット

公開されているデータには全てグラウンディングできるデータだけあるため，既存のデータセットを変更して，擬似的にデータセットを構築し，提案手法の評価を行った．本研究では RefCOCO+データセットを基盤に擬似データセットを作成した．擬似データは元のデータにおいてある画像と他の画像のクエリを組み合わせ，画像とクエリがお互いに関係ないように設定して，構築する．表 1 に擬似データセットの構成を示す．

図 3 ソーシャルメディアデータのアノテーション．“my favorite picture” はグランウンディングできないクエリであり，“Angle” と “Rannoch” は固有名詞のため入力クエリとして使わない．

表 1 擬似データセットの構成．擬似データセットは 67% の RefCOCO+と 33%の擬似データから構成されている．

	データ数	
データの種類	RefCOCO+	Pseudo Data
学習データ	42,278	21,139
検証データ	3,805	1,905
テストデータ	3,773	1,886

4.2 モデル学習

提案手法の候補領域の画像とクエリとの類似スコアを抽出するとき，ViLBERT[5] を用いた．ViLBERT では言語と画像が同時に連結された並行的な transformer ブロックを通して，処理される．このモデルは 4 つの画像・言語のタスクを同時学習させたモデルに，各データセットに fine-tuning することで，ビジュアルグラウンディング分野において最も高い精度を出した．ViLBERT の transformer ブロックが言語・画像それぞれ 12 と 6 層を用い，4 つのタスクに同時に事前学習されたモデルを用いた．

図 4　提案手法を擬似データに適用した結果例.

binary cross entropy を用い，最大 20 エポック，バッチサイズを 256，最初学習率を $4e-5$ に設定してビジュアルグラウンディングの fine-tuning を行った．

提案手法を RefCOCO+と擬似データセットでそれぞれ fine-tuning し，各データセットでの結果を比較した．また，擬似データセットで学習したモデルを構築したソーシャルメディアデータに対して提案手法の有効性を検証した．

表 2　各データセットでの提案手法の評価．Pseudo は擬似データセットでグラウンディング不可能なテストデータでの結果である．SNS-VG と SNT-UVG はそれぞれ構築したソーシャルメディアデータでグラウンディング可能と不可能なデータでの結果である．

学習データ	テストデータ			
	RefCOCO+	Pseudo	SNS-VG	SNS-UVG
RefCOCO+	72.7%	-	-	-
Pseudo Dataset	69.5%	91.2%	68.8%	76.8%

5　結果

5.1　RefCOCO+と擬似データでの評価

表 2 の左に RefCOCO+テストセットと擬似データテストセットにおける結果を示す．RefCOCO+で fine-tuning した ViLBERT モデルを提案手法を用いたとき，性能の変化を確認した結果，同じく 72.73%の精度が出た．また，グラウンディング不可能なデータに対する学習がされていないため，擬似データを擬似領域として検出することはできなかった．次に構築した擬似データセットに対して fine-tuning し，性能確認を行った．擬似テストデータセットは 67%の RefCOCO+のデータと 33%の擬似データと構成されている．擬似データにで fine-tuning したとき，図 4 のように，グラウンディング不可能なデータに対して，正しく擬似領域を推論していることが見られた．また，その精度は RefCOCO+データセットに対し，69.5%であり，擬似データに対しては 91.21%であった．既存の RefCOCO+で fine-tuning したモデルよりは RefCOCO+テストセットにおいて，精度が下がったが，擬似データでのグラウンディングできないものに対しては高い精度を出していることが見られた．

図 5　提案手法をソーシャルメディアデータに適用した結果例.

5.2　ソーシャルメディアデータでの評価

擬似データセットから fine-tuning された ViLBERT を用いて，構築した 75 個のソーシャルメディアデータに対し，提案手法の検証を行った．その結果を表 2 の右に示す．グラインディング可能なデータに対しては，精度が 68.8%，不可能なデータに対しては 76.8%の精度であった．また，結果例は図 5 に示されたように，グラウンディング可能なクエリとグラウンディング不可能なデータに対して，正しく動作していることがわかる．

6　終わりに

本研究ではグラウンディング不可能性を考慮したビジュアルグラウンディングモデルの提案とソーシャルメディアデータセット構築を行った．既存のビジュアルグラウンディングモデルにグラウンディング不可能なモジュールを追加させることで，グラウンディング可能なデータにおける性能低下も少なく，グラウンディングできないデータを高い精度で推論することができた．また，ソーシャルメディアから収集した画像データを複数の画像処理モデルによりフィルタリングすることで、RefCOCO+データセットと特性の近いデータセットを構築することができた．実際に構築したソーシャルメディアデータに対しても，グラウンディングできないものを正しく検出し，柔軟なグランウンディングできることが確認された．大規模なソーシャルメディアデータの構築は今後課題である．

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] Wenjian Dong, Mayu Otani, Noa Garcia, Yuta Nakashima, and Chenhui Chu. Cross-lingual visual grounding. *IEEE Access*, Vol. 9, pp. 349–358, 2021.

[2] 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. *CoRR*, Vol. abs/1505.04870, , 2015.

[3] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. *CoRR*, Vol. abs/1608.00272, , 2016.

[4] Yuke Zhu, Oliver Groth, Michael S. Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. *CoRR*, Vol. abs/1511.03416, , 2015.

[5] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *CoRR*, Vol. abs/1908.02265, , 2019.

[6] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *CoRR*, Vol. abs/1905.11946, , 2019.

[7] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection, 2020.

[8] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. Character region awareness for text detection, 2019.

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, Vol. abs/1703.06870, , 2017.

[10] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[11] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, Vol. abs/1405.0312, , 2014.