文脈理解に着目した 対照学習に基づく弱教師あり Phrase Grounding

唐井 希　清丸 寛一　Chenhui Chu　黒橋 禎夫
京都大学大学院 情報学研究科
{karai,kiyomaru,chu,kuro}@nlp.ist.i.kyoto-u.ac.jp

概要

本研究では，対照学習に基づく弱教師あり Phrase Grounding 手法の改善に取り組む．既存手法では，フレーズの文脈を考慮することが学習されず，画像中に同じクラスの物体が複数存在する場合にフレーズを正しい物体領域に対応付けられない．本研究では対照学習に用いる負例をフレーズの文脈を置き換えて作成することによってこの問題に対処する．提案手法を Flickr30k Entities データセットに適用し，ベースライン手法と比較して，最大で 2.89 ポイントの Recall の改善を確認した．

1　はじめに

Phrase Grounding（Phrase Localization）は画像中の物体領域と画像キャプション中のフレーズの対応関係を同定するタスクである [1]．自然言語によるロボットへの指示などその応用は多岐に渡る．

Phrase Grounding は一般に画像中の物体領域と画像キャプション中のフレーズの対応がアノテーションされた訓練データを用いて教師あり学習で解かれる [2, 3]．しかし，こうした訓練データの構築には膨大な人的・金銭的コストが必要であり，異なるドメインや多言語への適用，訓練データ数を増やすことによる性能改善は困難である．

この問題を解決するため，画像とキャプションのペアから Phrase Grounding を学習する弱教師あり設定が研究されている [4, 5, 6, 7]．この設定では画像中の物体領域と画像キャプション中のフレーズの対応のアノテーションが不要であり，必要となる画像とキャプションのペアもウェブデータ・実世界データから今後ますます多く手に入ると期待される [8]．

弱教師あり設定の手法としては，対照学習に基づく手法 [4, 5] が代表的である．この手法では，共起する物体領域とフレーズのペア（正例）の類似度の最大化とそれ以外のペア（負例）の類似度の最小化の学習を通して，実際に対応関係にある物体領域とフレーズのペアに高い類似度を与える．

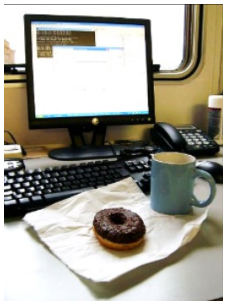


図 1: 対照学習における負例の作成方法．[カッコ内] は関心のフレーズ，下線は正例のキャプションから置き換えた単語を表す．

対照学習では負例の作成方法が性能を左右する．先行研究 [4] はフレーズの主辞を置き換えて負例を作成している．図 1 に例を示す．この例では，フレーズ *chocolate donut* の主辞 *donut* を *cookie* に置き換えて負例としている．これは対照学習により画像中に *chocolate donut* は存在するが *chocolate cookie* は存在しないことを学習することを意味する．しかし，これは主辞の *donut* と *cookie* に着目すれば解くことができ，画像中に同じクラスの物体（例えば *chocolate donut* と *cream donut*）が存在する場合に，主辞の文脈を考慮して正しい物体領域に対応付けることは必ずしも学習されない．

本研究では，フレーズの主辞ではなく，その文脈を置き換えて負例を作成することを提案する．図 1 に例を示す．提案手法では，主辞を修飾する文脈 *chocolate* を *cream* に置き換えて負例とする．これにより文脈を考慮し，同じクラスの物体を区別することを学習させる．

実験では，Flickr30K Entities [9] データセットに提案手法を適用し，提案手法がベースライン手法の性能を大きく改善することを確認した．

2 関連研究

Phrase Grounding におけるアノテーションコストの問題を解決・緩和するため，いくつかのアプローチが提案されている．最も素朴な方法として，学習済みの物体検出器のラベルとフレーズの単語類似度に基づき，物体領域とフレーズの対応を見つける手法がある [10]．しかし，物体の情報が物体検出器が予測するラベルの範囲でしか区別できないため，同一クラスの物体を区別してフレーズと対応付けることは困難である．

別のアプローチとして，画像とキャプションのペアから Phrase Grounding を学習する弱教師あり設定が研究されている [4, 5, 6, 7]．最も代表的な手法は対照学習に基づくものである [4, 5]．他の手法としては，訓練済みのキャプショニングモデルの注意機構の活性からフレーズと物体領域の対応を抽出する手法 [7] などがある．

3 対照学習による弱教師あり Phrase Grounding

提案手法は，対照学習によって画像とキャプションのペアデータから Phrase Grounding を学習する枠組み [4, 5] に基づく．この手法では，共起する物体領域とフレーズのペアの類似度を上げつつ，共起しない物体領域とフレーズのペアの類似度を下げることを学習する．この学習により，実際に対応関係にある物体領域とフレーズのペアには対応関係にないペアよりも高い類似度が与えられるようになる．

形式的には，まず画像とキャプションからそれぞれ物体領域特徴 $\mathbf{R} = \{r_1, \cdots, r_m\}$（$r_i \in \mathbb{R}^{d_r}$）とフレーズ特徴 $\mathbf{W} = \{p_1, \cdots, p_n\}$（$p_j \in \mathbb{R}^{d_p}$）を抽出する．ただし，$m$ は画像から検出された物体領域の数，d_r は物体領域特徴の次元数，n はキャプション中のフレーズの数，d_p はフレーズ特徴の次元数である．物体領域特徴とフレーズ特徴はそれぞれ事前学習済みの物体検出器と事前学習済みの言語モデルから得る．

次に，物体領域特徴とフレーズ特徴の類似度を注意機構—[11] によって計算する．注意機構による類似度の計算には，物体領域特徴の key ベクトルへの変換 $k_r : \mathbb{R}^{d_r} \to \mathbb{R}^d$ とフレーズ特徴の query ベクトルへの変換 $q_p : \mathbb{R}^{d_p} \to \mathbb{R}^d$ を用いる．ただし，d は変換後の特徴の次元数である．k_r と q_p はいずれも単一の線形層を備えたニューラルネットワークとする．これらを用いて，物体領域特徴 r_i とフレーズ特徴 p_j の類似度 $a\left(r_i, p_j\right)$ を得る．

$$a\left(r_i, p_j\right) = \frac{e^{s\left(r_i, p_j\right)}}{\sum_{i'=1}^{m} e^{s\left(r_{i'}, p_j\right)}} \tag{1}$$

$$s\left(r_i, p_j\right) = q_p\left(p_j\right)^T k_r\left(r_i\right) / \sqrt{d} \tag{2}$$

テスト時は各フレーズについて最も類似度が高い物体領域を選択することで Phrase Grounding を実行する．しかし，弱教師あり設定では物体領域とフレーズの対応に関する教師信号がないため，この類似度は直接最適化することはできない．

そこで，画像 $\mathbf{R}$ とフレーズ p_j の類似度 $\phi\left(\mathbf{R}, p_j\right)$ を導入し，このもとで対照学習を行うことで物体領域とフレーズの類似度を最適化する．この計算のため，物体領域特徴の value ベクトルへの変換 $v_r : \mathbb{R}^{d_r} \to \mathbb{R}^d$ とフレーズ特徴の value ベクトルへの変換 $v_w : \mathbb{R}^{d_p} \to \mathbb{R}^d$ を導入する．v_r と v_w もまた，単一の線形層を備えたニューラルネットワークとして構成する．これらを用いて，画像とフレーズの類似度 $\phi\left(\mathbf{R}, p_j\right)$ を得る．

$$\phi\left(\mathbf{R}, p_j\right) = v_w^T\left(p_j\right) v_{\text{att}}\left(\mathbf{R}, p_j\right) \tag{3}$$

$$v_{\text{att}}\left(\mathbf{R}, p_j\right) = \sum_{i=1}^{m} a\left(r_i, p_j\right) v_r\left(r_i\right) \tag{4}$$

対照学習は，Gupta ら [4] に倣い，画像を置き換えて負例を作る画像側の対照学習と，キャプション中の単語を置き換えて負例を作る言語側の対照学習の組み合わせで行う．画像側の対照学習では，ミニバッチ中に含まれる別の画像の物体領域特徴を負例とする．目的関数 $\mathcal{L}_{img}$ は以下の通りである．

$$\mathcal{L}_{\text{img}} = -\sum_{j=1}^{n} \log\left(\frac{e^{\phi\left(\mathbf{R}, p_j\right)}}{e^{\phi\left(\mathbf{R}, p_j\right)} + \sum_{i=1}^{k-1} e^{\phi\left(\mathbf{R}_i', p_j\right)}}\right) \tag{5}$$

ただし，k はバッチサイズである．

言語側の対照学習では，正例のキャプション中の単語を別の単語に置き換えることで負例とする．具体的には，フレーズの主辞を対象として，事前学習済みの BERT [12] を使用し，文脈的に妥当な単語に置き換えて負例を作成する．置き換え先の単語として置き換え元の単語の同義語や上位語が選択され不適格な負例となるを避けるため，以下の手続きを取る．ある文脈 c を伴う単語 w を置き換える際，まず，単語 w をマスクしたキャプションを BERT に入力し，マスクの部分の当てはまる単語の予測確率 $p(w'|c)$ を計算する．この上位 K 単語を置き換え

表 1: 各手法で作成される負例の例．[カッコ内] は関心のフレーズ，下線は正例のキャプションから置き換えた単語を表す．

正例	Head による負例	Adj による負例	Rand による負例
[*Chocolate donut*] *in front of a computer*	[*Chocolate cookie*] *in front of a computer*	[*Cream donut*] *in front of a computer*	[*Chocolate donut*] *in front of a radio*
[*a chinese man*] *sitting down waiting for customers*	[*a chinese girl*] *sitting down waiting for customers*	[*a blond man*] *sitting down waiting for customers*	[*a chinese man*] *sitting down waiting for lunch*
[*Three barefoot children*] *fun and smile in an indoor setting*	[*Three barefoot guys*] *fun and smile in an indoor setting*	[*Three female children*] *fun and smile in an indoor setting*	[*Three barefoot children*] *fun and moved in an indoor setting*

先の単語候補とする．次に，単語 w をマスクせずそのまま BERT に入力し，単語 w の部分に当てはまる単語の予測確率 $p(w'|w,c)$ を計算する．置き換え元の単語の同義語や上位語には $p(w'|w,c)$ で高い確率になると期待される．置き換え先の単語候補をスコア $p(w'|c)/p(w'|w,c)$ をもとに並び替え，上位 L 単語を置き換え先の単語として採用する．目的関数 $\mathcal{L}_{\text{lang}}$ は以下の通りである．

$$\mathcal{L}_{\text{lang}} = -\log\left(\frac{e^{\phi(\mathbf{R},p_j)}}{e^{\phi(\mathbf{R},p_j)} + \sum_{l=1}^{L} e^{\phi(\mathbf{R},p'_l)}}\right) \quad (6)$$

ただし，L は負例の数，p'_l は正例のキャプション中の単語を置き換えた上で抽出したフレーズ特徴である．

モデルはこれらの和 $\mathcal{L} = \mathcal{L}_{\text{img}} + \mathcal{L}_{\text{lang}}$ を目的関数とし，その最小化を学習する．

4 提案手法

対照学習に基づく Phrase Grounding では，フレーズの主辞を置き換えて負例を作成されている [4]．本論文ではこの負例の作成方法を **Head** と呼ぶ．

本論文では，これに加えて，フレーズの主辞の文脈単語を置き換えて作成した負例を対照学習に利用することを提案する．本研究では 2 つの方法を提案する．1 つ目はフレーズの主辞を修飾する形容詞を置き換える方法である．まず，フレーズの主辞を修飾する形容詞を構文解析を適用して見つける置き換え先の単語は Head と同様，BERT の予測確率に基づき得る．この負例の作成方法を **Adj** と呼ぶ．

2 つ目はフレーズ内の単語を含むフレーズの周辺単語を無作為に置き換える方法である．Adj と比べて，多様な修飾関係が学習されると期待される．置き換え先の単語は Head と同様の方法で得る．この負例の作成方法を **Rand** と呼ぶ．

Head と Adj は品詞情報，統語情報を活用して置き換える単語を選択するため，基本的に妥当な負例が得られる．一方，Rand はそれらを一切考慮しないため，多様な修飾関係がカバーされる代わりに，不適当な負例になる場合も多い．表 1 に各手法で作成される負例の具体例を示す．

5 実験

提案手法の有効性を確認するため実験を行った．

5.1 実験設定

実験には Flickr30K Entities [9] データセットを使用した．Flickr30K Entities には 3 万枚の画像が含まれ，各画像に 5 つのキャプションが付与されている．また，すべての画像・キャプションのペアに対して，画像領域とキャプション中のフレーズの対応のアノテーションが付与されている．提案手法は弱教師あり手法であり，画像領域とフレーズの対応に関するアノテーションは訓練に使用しないことに注意されたい．

評価指標には Recall と Pointing Accuracy の 2 つを用いた．Recall は正解領域と予測領域が IOU ≥ 0.5 であるフレーズの割合である．Pointing Accuracy は，モデルがフレーズごとに 1 つの点の位置を予測し，その予測が正解領域内にある割合である．本研究では予測領域の中心を予測値の点とした．

物体領域特徴は物体検出器 Faster-RCNN [13] を用いて抽出した．フレーズ特徴は BERT [12] が出力するフレーズの構成単語の文脈化埋め込みの平均とした．

画像側の対照学習における負例の数 k は 10，言語側の対照学習における負例の数 L は 5 とした．Adj による負例生成には，構文解析 spacy[1] の出力を使用した．Rand による負例生成では，置き換える単語候補をフレーズ内の単語およびフレーズの前後 3，5，10 単語として比較した．

1） https://spacy.io/

表 2: 実験結果．最良の結果を**太字**で示す．(†) は元論文からの引用．RAND に付記されたカッコ内の数字はフレーズの前後何単語以内を置き換えの対象としたかを表す．

	Recall	Pointing Acc.
先行研究 [4]†	47.88	74.94
先行研究 [4]	46.67	72.77
HEAD	49.96	73.83
HEAD+ADJ	52.19	**75.76**
HEAD+RAND (3)	50.67	74.27
HEAD+RAND (5)	52.05	74.54
HEAD+RAND (10)	**52.85**	75.01

5.2 結果

表 2 に結果を示す．フレーズの主辞の名詞を置き換えて負例とする HEAD はほぼ先行研究 [4] の再現であるが，3.28 ポイントの Recall の改善が見られた．これは，先行研究がフレーズ特徴としてフレーズ主辞の文脈化埋め込みを使用しているのに対し，HEAD はフレーズを構成する単語の文脈化埋め込みの平均を使用しているという違いに起因する．先行研究 [4] も文脈化埋め込みを使用しているため主辞の文脈の情報は考慮されていると期待されるが，実験の結果，フレーズの構成単語全ての埋め込みを利用してフレーズ特徴を構成する方が効果的であることが確認された．

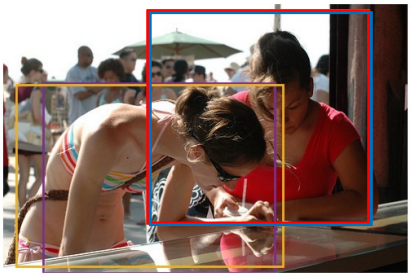
主辞を修飾する形容詞を置き換えて負例とする ADJ を加えることで，さらに 2.23 ポイントの Recall の改善が見られた．また，フレーズの周辺単語を無作為に置き換えて負例とする RAND 加えることで，最大 2.89 ポイントの Recall の改善が確認された．RAND については，置き換える単語の窓幅を大きくするほど精度が向上する傾向が確認された．これは選択する範囲を増やすことで考慮できる修飾関係の種類が増えたためだと考えられる．

5.3 ケーススタディ

モデルの出力を分析したところ，提案手法によりモデルが期待通り，フレーズの文脈を考慮することを学習していることが確認された．図 2a に例を示す．HEAD では画像中にフレーズの主辞 *jacket* を修飾する *leather* の情報が考慮されず，誤った物体領域を選択しているが，HEAD+ADJ ではフレーズを正しい物体領域に選択できている．

A five-piece band , four of the men in red outfits and one of them in a leather jacket and jeans , perform on the sidewalk in front of a shop .

(a) 提案手法による改善例．



A young woman in a bikini looking at something in a glass case .

(b) 提案手法による改善例．

図 2: 提案手法によって改善した例．青色が HEAD（ベースライン），橙色が HEAD+ADJ（提案），紫色が HEAD+RAND（提案），黄色が正解の領域を示す．

形容詞を考慮するだけでは不十分な例も確認された．図 2b に例を示す．この例では主辞を修飾する *young* の情報だけではどの領域が正しいか選択できず，HEAD+ADJ は誤った領域を出力している．一方，HEAD+RAND は *in a bikini* を考慮して正しい領域を出力している．HEAD+RAND と HEAD+ADJ と比べて多様な修飾関係を考慮しているにも関わらず，その性能差は決して大きくない．これは HEAD+RAND で作成される負例にノイズが多く含まれることが原因と考えられる．

6 おわりに

本研究では，対照学習に基づく弱教師あり Phrase Grounding 手法の改善に取り組んだ．既存手法ではフレーズの文脈を考慮した Phrase Grounding が学習されないことに着目し，対照学習に用いる負例の工夫によりこの問題に対処した．実験では提案手法を Flickr30k Entities データセットに適用し，その有効性を確認した．今後，言語素性・統語構造を考慮したより効果的な負例の作成方法を検討したい．

謝辞

本研究はヤフー株式会社の支援のもとで行われた。

参考文献

[1] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**, pp. 457–468, Austin, Texas, November 2016. Association for Computational Linguistics.

[2] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr - modulated detection for end-to-end multi-modal understanding. In **Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)**, pp. 1780–1790, October 2021.

[3] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In **2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 10955–10965, 2022.

[4] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In **ECCV**, 2020.

[5] Liwei Wang, Jing Huang, Yin Li, Kun Xu, Zhengyuan Yang, and Dong Yu. Improving weakly supervised visual grounding by contrastive knowledge distillation. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 14090–14100, June 2021.

[6] Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In **Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)**, October 2019.

[7] Effrosyni Mavroudi and René Vidal. Weakly-supervised generation and grounding of visual descriptions with conditional generative models. In **2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 15523–15533, 2022.

[8] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[9] 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 **2015 IEEE International Conference on Computer Vision (ICCV)**, pp. 2641–2649, 2015.

[10] Josiah Wang and Lucia Specia. Phrase localization without paired training examples. In **Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)**, October 2019.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 30. Curran Associates, Inc., 2017.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[13] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, June 2018.