Title	マルチモーダル基盤モデルと混合緩和損失を用いた大規模屋内検索エンジン
Sub Title	
Author	今井, 悠人(Imai, Yūto)
Publisher	慶應義塾大学AI・高度プログラミングコンソーシアム
Publication year	2024
Jtitle	AICカンファレンス予稿集 (2024.) ,p.11- 12
JaLC DOI	
Abstract	本研究では，大規模な屋内環境内の物体の画像を，Open-vocabularyなユーザ指示に基づきランク付けを行うタスクを扱う．画像から画素・位置関係を含む複数の粒度での特徴抽出および環境中に存在する類似物体の影響を緩和する損失を提案する．提案手法は，新たに構築したYAGAMIデータセットおよびLTRRIE-2.0データセットにおいて，全ての評価指標においてベースライン手法を上回る結果を得た．
Notes	会議名 : AICカンファレンス2024 開催地 : 慶應義塾大学協生館AICラウンジ、および多目的教室 日時 : 2024年3月7日 ポスター発表要旨
Genre	Conference Paper
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=KO11003001-20240307-0011

マルチモーダル基盤モデルと混合緩和損失を用いた大規模屋内検索エンジン

今井悠人
慶應義塾大学理工学部情報工学科

Abstract: 本研究では，大規模な屋内環境内の物体の画像を，Open-vocabulary なユーザ指示に基づきランク付けを行うタスクを扱う．画像から画素・位置関係を含む複数の粒度での特徴抽出および環境中に存在する類似物体の影響を緩和する損失を提案する．提案手法は，新たに構築した YAGAMI データセットおよび LTRRIE-2.0 データセットにおいて，全ての評価指標においてベースライン手法を上回る結果を得た．

Keywords: Learning to Rank, Object Image Retrieval, Contrastiveness

1. 研究背景・目的

本研究では，人間がロボットに実環境中の物体に関する自然言語による指示を与えた時，対象物体を検索するタスクを扱う．複雑な参照表現を含む open-vocabulary な指示文は曖昧さや冗長性を含むことが多いため，指示文から対象物体を正しく把握することは難しい．また，検索空間が大きくなるほど，類似した物体が含まれやすく，画像中の物体理解がより重要となる．本論文では，大規模屋内空間中の検索を行うために，画素・位置関係を含む複数の粒度からの特徴抽出および混合緩和損失による対照性の緩和を提案する．

2. 方法

提案手法は Granular Representation from Entire to Pixels (GREP)，Instruction Noun Phrase Encoder(INPE)，Relaxing Contrastive Similarity (RCS) の 3 つのモジュールから構成される．提案手法の入力は指示文 $\boldsymbol{x}_{\text{inst}}$，対象物体領域集合 $\mathcal{T}$，環境画像集合 $\mathcal{C}$ である．なお，以下では $\mathcal{T}$，$\mathcal{C}$ の n 番目の要素を $\boldsymbol{x}_t^{(n)}$，$\boldsymbol{x}_c^{(n)}$ とする．

GREP モジュールでは，画素・空間関係・対象物体・画像全体の 4 つの粒度から特徴抽出を行う．まず，画素単位については，SAN [1] を参考に CLIP [2] から画素単位での特徴を得る．具体的には，$\boldsymbol{x}_c^{(n)}$ を SAN に入力し，以下の式から中間特徴 $\boldsymbol{h}_{\text{SAN}}^{(i)}$ を得る．

$$\boldsymbol{h}_{\text{SAN}}^{(i)} = \text{CLIP}^{(i_c)}\left(\boldsymbol{x}_c^{(n)}\right) + \text{Transformer}^{(i_t)}\left(\boldsymbol{x}_c^{(n)}\right)$$

ここで，$\text{CLIP}^{(i_c)}$ および $\text{Transformer}^{(i_t)}$ は CLIP 画像エンコーダ (ViT-B/16) 第 i_c 層および第 i_t 層の transformer 層を表す．$\boldsymbol{h}_{\text{SAN}}^{(i)}$ を MLP に入力して得られる $\boldsymbol{h}_{\text{px}}^{(n)}$ が画素単位での特徴である．

次に，空間関係に関しては，$\boldsymbol{x}_c^{(n)}$ に対し，CLIP 画像エンコーダ (RN50x4) の中間層出力に Positional Encoding [3] を加えたものを $\boldsymbol{h}_{\text{mid}}^{(n)}$ とする．この $\boldsymbol{h}_{\text{mid}}^{(n)}$ に対し畳み込みを繰り返すことで，空間的な画像特徴 $\boldsymbol{h}_{\text{sp}}^{(n)}$ を得る．

対象物体，画像全体については，事前学習済みの CLIP 画像エンコーダ (ViT-L/14) を用いて，画像特徴 $\boldsymbol{h}_t^{(n)}$，$\boldsymbol{h}_c^{(n)}$ をそれぞれ抽出する．

GREP モジュールの出力は，$\boldsymbol{h}_{\text{px}}^{(n)}$,$\boldsymbol{h}_{\text{sp}}^{(n)}$, $\boldsymbol{h}_t^{(n)}$, $\boldsymbol{h}_c^{(n)}$ を MLP に入力した $\boldsymbol{h}_{\text{img}}^{(n)}$ である．

INPE モジュールでは，$\boldsymbol{x}_{\text{inst}}$ に [4] を適用し，N_{noun} 個の名詞句 $\boldsymbol{x}_{\text{noun}}^{(k)}(k = 1, ..., N_{\text{noun}})$ を得る．その後，CLIP テキストエンコーダ [2] を用いて $\boldsymbol{x}_{\text{inst}}$ および $\boldsymbol{x}_{\text{noun}}^{(k)}$ から指示文に関する言語特徴量 $\boldsymbol{h}_{\text{inst}}$ および $\boldsymbol{h}_{\text{noun}}^{(k)}$ を獲得する．最後に，transformer 層に $\boldsymbol{h}_{\text{noun}}^{(k)}$ を連結したものを入力し，この出力と $\boldsymbol{h}_{\text{inst}}$ を MLP に入力することでモジュールの出力となる言語特徴量 $\boldsymbol{h}_{\text{txt}}$ が得られる．

RCS では，CNPE と GREP から得られた $\boldsymbol{h}_{\text{txt}}$ および $\boldsymbol{h}_{\text{img}}^{(n)}$ によって得られるコサイン類似度 s を出力する．s は，以下の式で計算される．

$$s(\boldsymbol{x}_{\text{inst}}, \boldsymbol{x}_t^{(n)}) = \frac{\boldsymbol{h}_{\text{txt}} \cdot \boldsymbol{h}_{\text{img}}^{(n)}}{||\boldsymbol{h}_{\text{txt}}||\ ||\boldsymbol{h}_{\text{img}}^{(n)}||}$$

モデルの出力は s に基づきランク付けされた画像集合 $\mathcal{T}$’ である．また，ここで計算した s に基づき，損失を計算する．本研究では損失関数として，ReCo [5] 損失と InfoNCE [6] 損失を組み合わせた混合緩和損失を提案する．これらは負例に対する演算が異なる．具体的には，InfoNCE 損失はすべての負例に対して類似度を-1 に近づけるように学習するが，ReCo 損失は類似度が正の負例に対し類似度を 0 に近づけるよう学習する．一方で，ReCo 損失は部分的にしか負例に対する最適化を行わないため，学習効率が悪い．そこで本研究では，これらをバランスした混合緩和損失 $\mathcal{L}_{\text{mix}}$ を提案する．$\mathcal{L}_{\text{mix}}$ は以下の形で表される．

$$\mathcal{L}_{\text{mix}} = \lambda_{\text{InfoNCE}}\mathcal{L}_{\text{InfoNCE}} + \lambda_{\text{ReCo}}\mathcal{L}_{\text{ReCo}}$$

ここで，$\mathcal{L}_{\text{InfoNCE}}$，$\lambda_{\text{InfoNCE}}$，$\mathcal{L}_{\text{ReCo}}$，$\lambda_{\text{ReCo}}$ は InfoNCE 損失，InfoNCE 損失の重み，ReCo 損失，ReCo 損失の重みを表す．

3. 実験設定

本研究では，YAGAMI データセットおよび LTRRIE データセット [7] を拡張した LTRRIE-2.0 データセットを作成した．YAGAMI データセットの画像群は，矢上キャンパス内約 3,000 m^2 にわたって，モバイルデバイスによって撮影された．また，LTRRIE-2.0 データセットは，疑似的に大規模環境を再現するために，LTRRIE データセット内の複数の環境から検索できるように拡張された．

4. 結果

図 1 に YAGAMI データセットにおける成功例の定性的結果を示す．提案手法および MultiRankIt [] において，$\boldsymbol{x}_{\text{inst}}$

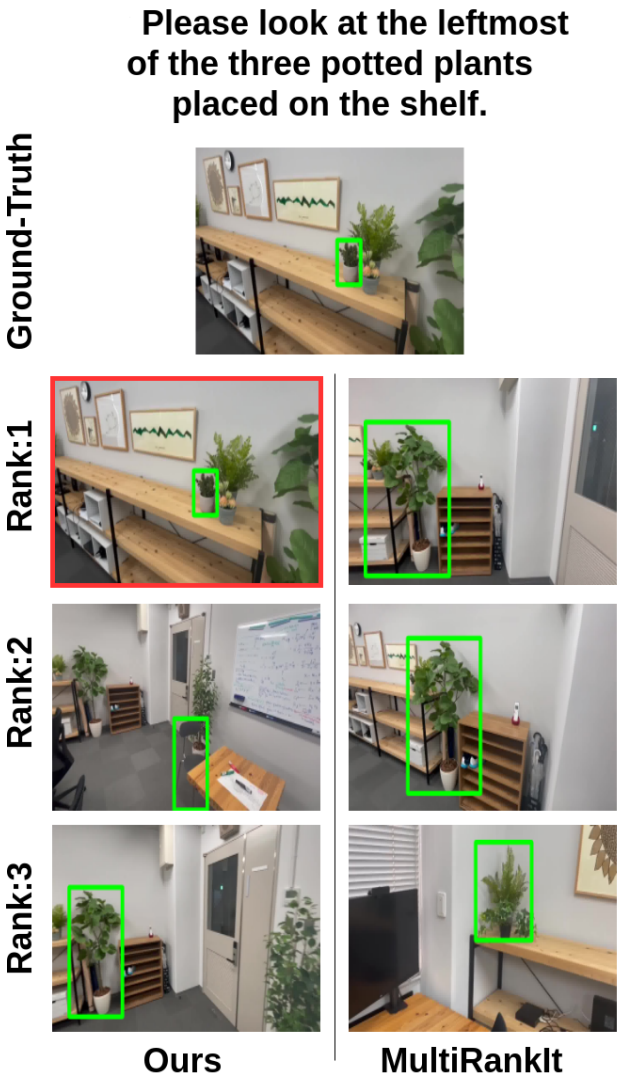

図 1 定性的結果

に対する正解および上位 3 件の検索結果を表している．緑色で囲まれている領域が $\boldsymbol{x}_t^{(n)}$，赤く囲まれている画像は正解の $\boldsymbol{x}_c^{(n)}$ を表す．

図 1 では，$\boldsymbol{x}_{\text{inst}}$ として “Please look at the leftmost of the three potted plants placed on the shelf.” を入力した場合の結果を表している．MultiRankIt では 16 位に検索されている対象が，提案手法では 1 位に検索されている．指示文中の，“the leftmost of the three potted plants” という空間的な位置関係を提案手法は正確に把握されており，これは $\boldsymbol{h}_{\text{sp}}$ が効果的であったと考えられる．さらに，MultiRankIt の上位 3 件も potted plants を検索しているにも関わらず，正解を正しく検索できていない．一方で，提案手法では正しいサンプルを予測できている．このような類似した物体に対して，RCS で計算した ReCo 損失を考慮した損失関数が有効であったと考えられる．

5. 結論

本研究では，指示文および環境中の画像集合と物体領域が与えられた場合に，対象物体を検索するタスクに取り組んだ．将来研究としては，実機での実験や座標位置の導入が考えられる．

参考文献

[1] M. Xu, Z. Zhang, F. Wei, H. Hu, X. Bai, “Side Adapter Network for Open-vocabulary Semantic Segmentation,” in *CVPR*, pp. 2945–2954, 2023.
[2] A. Radford *et al*., “Learning Transferable Visual Models from Natural Language Supervision,” in *ICML*, pp. 8748–8763, 2021.
[3] A. Vaswani *et al*., “Attention is All You Need,” in *NIPS*, pp. 5998–6008, 2017.
[4] S. Schuster *et al*., “Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks,” in *LREC*, pp. 2371–2378, 2016.
[5] Z. Lin *et al*., “Relaxing Contrastiveness in Multimodal Representation Learning,” in *WACV*, pp. 2227–2236, 2023.
[6] A. van den Oord, Y. Li, O. Vinyals, “Representation Learning with Contrastive Predictive Coding,” *arXiv preprint arXiv:1807.03748*, 2018.
[7] K. Kaneda *et al*., “Learning-To-Rank Approach for Identifying Everyday Objects Using a Physical-World Search Engine,” *IEEE RAL*, vol. 9, no. 3, pp. 2088-2095, 2024.