Title	生活支援口ボットによる参照表現理解モデルの構築
Sub Title	
Author	吉田, 悠(Sugiura, Kōmei) 杉浦, 孔明
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	2023
Jtitle	AICカンファレンス予稿集 (2023.) ,p.28- 29
JaLC DOI	
Abstract	本研究では，物体操作指示文に対する対象物体のセグメンテーションマスクを生成するモデルの開発を行った．このタスクは，ユーザの指示が曖昧であることや，複数の候補が存在する可能性があることから，挑戦的な課題である．本論文では，既存手法の計算コストを削減しつつ，性能を向上させるため，Language-Aware Conv-Attentional Transformers with box-supervision(LACT-BS) モデルを提案する．本手法は，標準的な評価尺度と計算効率においてベースライン手法を上回る性能を達成した．
Notes	会議名 : AICカンファレンス2023 開催地 : 慶應義塾大学日吉キャンパス 日時 : 2023年3月4日 第2章ポスター発表要旨 ポスター要旨-4
Genre	Conference Paper
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=KO11003001-20230304-0028

生活支援ロボットによる参照表現理解モデルの構築

吉田悠 [1]，杉浦孔明 [1]
[1] 慶應義塾大学大学院理工学研究科開放環境科学専攻

Abstract:
本研究では，物体操作指示文に対する対象物体のセグメンテーションマスクを生成するモデルの開発を行った．このタスクは，ユーザの指示が曖昧であることや，複数の候補が存在する可能性があることから，挑戦的な課題である．本論文では，既存手法の計算コストを削減しつつ，性能を向上させるため，Language-Aware Conv-Attentional Transformers with box-supervision (LACT-BS) モデルを提案する．本手法は，標準的な評価尺度と計算効率においてベースライン手法を上回る性能を達成した．
Keywords: Referring Expression Segmentation, box-supervised, multimodal language processing

1. 研究背景・目的

高齢化が進行している現代社会において，日常生活における介助支援の必要性は高まっている．その結果，在宅介助者の不足が社会問題となっており，被介助者を物理的に支援可能な生活支援ロボットが注目されている．

ロボットが動作の対象となる物体を特定する際，マスクを用いた対象物体の予測は，物体の位置や形状を正確に把握するのにはるかに有用である．しかし，高性能のセグメンテーションモデルの構築には，学習データにピクセル単位のアノテーションが必要であり，非常に時間や手間がかかる．

そこで, 我々はセグメンテーションマスクではなく, 対象物体の矩形領域を学習に用いる box-supervised approach に着目する．本論文では，Language-Aware Conv-Attentional Transformers with box-supervision (LACT-BS) を提案する．本手法は box-supervised の条件下におけるマルチモーダルセグメンテーション性能の向上のために設計され，さらに学習時の計算量を削減することが可能になっている．

2. 方法

提案手法の新規性は以下である．

- REVERIE-seg を扱う LACT-BS を提案し，予測した矩形領域に関する box-specific coordinate loss を導入した．
- LACT-BS の Encoder において，計算量削減のために，Factorized Attention[1] を用いた構造を導入した．

2.1 REVERIE-seg

本論文で扱うタスクを，Remote Embodied Visual referring Expression in Real Indoor Environments segmentation (REVERIE-seg) と定義する．本タスクでは，指示文と画像が与えられたときに，モデルが対象物体のセグメンテーションマスクを生成する必要がある．

扱うデータセットとして，Vision-and-Language Navigation や Object Localization における標準データセットである REVERIE dataset[2] から，対象物体が映っているキューブマップ，命令文，対象物体の矩形領域をそれぞれ抽出し，REVERIE-seg dataset を構築した．評価には，テスト集合の対象物体についてのみ，ground truth のセグメンテーションマスクが必要であるため，追加でアノテーションを行った．

2.2 Encoder

LACT-BS の Encoder は，画像特徴量の抽出を行ういくつかの CoaT Block で構成される．CoaT Block では，一般的な self-attention ではなく，Factorized Attention[1] を適用する．これにより，計算量を線形にすることができる．

また，CoaT Block から出力される CLS token の表現を線形変換することで，対象物体の矩形領域の座標 $\hat{Y}_{\mathrm{coord}}$ を予測し，後述する損失関数で補助ロスとして使用した．

2.3 損失関数

損失関数 $\mathcal{L}$ を以下に示す．

$$\mathcal{L} = \lambda_{\mathrm{ce}}\mathcal{L}_{\mathrm{ce}}(Y_{\mathrm{b}}, \hat{Y}) + \lambda_{\mathrm{focal}}\mathcal{L}_{\mathrm{focal}}(Y_{\mathrm{b}}, \hat{Y}) + \mathcal{L}_{\mathrm{coord}} + \mathcal{L}_{\mathrm{boxinst}},$$

$$\mathcal{L}_{\mathrm{coord}} = \lambda_{\mathrm{L1}}\mathcal{L}_{\mathrm{SSE}}(Y_{\mathrm{coord}}, \hat{Y}_{\mathrm{coord}}) + \lambda_{\mathrm{IoU}}\mathcal{L}_{\mathrm{IoU}}(Y_{\mathrm{coord}}, \hat{Y}_{\mathrm{coord}}).$$

ここで，λ_{ce}, λ_{focal}, λ_{L1}, λ_{IoU}, λ_{proj}, $\lambda_{\mathrm{pairwise}}$ はハイパーパラメータである．Y_{b} は ground-truth の矩形領域，$\hat{Y}$ は予測したセグメンテーションマスクを示す．また，$\mathcal{L}_{\mathrm{ce}}$ は交差エントロピー誤差，$\mathcal{L}_{\mathrm{focal}}$ は Focal loss[3]，$\mathcal{L}_{\mathrm{boxinst}}$ は [4] で提案された損失関数をそれぞれ示す．

$\mathcal{L}_{\mathrm{coord}}$ は本研究で提案する box-specific coordinate loss を示し, $\mathcal{L}_{\mathrm{SSE}}$ と $\mathcal{L}_{\mathrm{IoU}}$ で構成される．Y_{coord} は ground-truth の矩形領域の座標を使用する．$\mathcal{L}_{\mathrm{SSE}}$ は二乗和誤差を示す．また， $\mathcal{L}_{\mathrm{IoU}}$ は補助損失であり，以下で定義される．

$$\mathcal{L}_{\mathrm{IoU}} = 1 - \mathrm{IoU}(Y_{\mathrm{coord}}, \hat{Y}_{\mathrm{coord}})$$

ここで，IoU は Intersection over Union を示す．

3. 結果

3.1 定量的結果

Table 1 に，参照表現セグメンテーションタスクにおいて代表的なモデルである LAVT[5] との比較結果を示す．実験は 5 回行い，結果はその平均値と標準偏差を示す．評価指標として，mean intersection-over-union (mIoU)，overall intersection-over-union (oIoU)，0.5, 0.7, 0.9 の閾値 k における Precision@k (P@k) を使用した．LAVT と提案手法の mIoU はそれぞれ 25.85 ポイントと 32.70 ポイントであった．

Table 1 REVERIE-seg における定量的結果

Method	mIoU	oIoU	P@0.5	P@0.7	P@0.9
LAVT[5]	$25.85_{\pm 1.43}$	$33.13_{\pm 0.76}$	$26.51_{\pm 2.41}$	$7.57_{\pm 1.00}$	$0.38_{\pm 0.22}$
Ours	$\mathbf{32.70}_{\pm 0.36}$	$\mathbf{34.20}_{\pm 0.81}$	$\mathbf{35.24}_{\pm 0.96}$	$\mathbf{9.88}_{\pm 0.90}$	$\mathbf{0.51}_{\pm 0.00}$

3.2 計算コストの比較

Table 2 に，LAVT との計算コストの比較結果を示す．LAVT は，パラメータ数 118M，積和演算数 444G，学習時間 22min/epoch であった．一方，提案手法では，パラメータ数，積和演算数，学習時間がそれぞれ 54.4M，139G，15 分/epoch となった．

Table 2 計算コストの比較

Method	#Params (M)	#Mult-adds (G)	Training time (min/epoch)
LAVT[5]	118	444	22
Ours	54.4	139	15

3.3 定性的結果

Fig. 1 に，REVERIE-seg における定性的結果を示す．入力画像に対する命令文は “Fluff up the black pillow that is on the brown easy chair in the office containing the pictures of boom boxes.” である．LAVT の予測マスクは，誤って対象物体ではない手前のクッションをマスクしているのに対し，提案手法の予測マスクは ground-truth と同じクッションを示している．

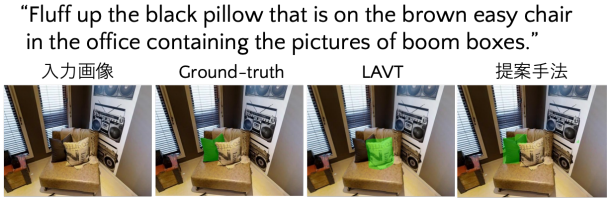

Fig. 1 REVERIE-seg における定性的結果

4. 考察

Table 1 より, 提案手法が mIoU において LAVT を 6.85 ポイント上回っていることがわかる．また，全てのメトリクスにおいて，提案手法は LAVT を上回っている．また Table 2 より，提案手法は LAVT と比較して，パラメータ数，積和演算数，学習時間がそれぞれ 63.6M，305G，7min/epoch 減少している．したがって，提案手法は計算コストを削減しつつ，性能を向上させることができた．

5. 結論

命令文の対象物体に対してセグメンテーションを行う REVERIE-seg タスクを提案し，そのうち，セグメンテーションマスクの代わりに対象物体の矩形領域を学習で用いることに着目した．Box-specific coordinate loss を導入した REVRIE-seg モデルである LACT-BS を提案し，また，Visual Encoder に Factorized Attention を用いた構造を導入することで，計算量の削減を行った．LACT-BS は，標準的な評価指標でベースライン手法を上回った．

参考文献

[1] Weijian Xu, Yifan Xu, Tyler Chang, and Zhuowen Tu. Co-Scale Conv-Attentional Image Transformers. In *ICCV*, pages 9981–9990, 2021.

[2] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments. In *CVPR*, pages 9982–9991, 2020.

[3] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection. In *ICCV*, pages 2980–2988, 2017.

[4] Zhi Tian, Chunhua Shen, Xinlong Wang, and Hao Chen. BoxInst: High-Performance Instance Segmentation with Box Annotations. In *CVPR*, pages 5443–5452, 2021.

[5] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. LAVT: Language-Aware Vision Transformer for Referring Image Segmentation. In *CVPR*, pages 18155–18165, 2022.