Title	市街地での移動指示文に基づく目標領域予測
Sub Title	
Author	畑中, 駿平(Hatanaka, Shunpei)
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.7- 8
JaLC DOI	
Abstract	本研究では，画像・移動指示文・セマンティックセグメンテーションマスクの3つのモダリティを扱うことができる，Trimodal Navigable Region Segmentation Model (TNRSM) を提案する．このモデルをTalk2Car-Regsegデータセットで検証した．結果として，提案手法は，移動指示理解タスクにおいて，全ての評価指標でベースライン手法を上回る性能を得られた．
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=KO11003001-20240307-0007

市街地での移動指示文に基づく目標領域予測

畑中駿平
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，画像・移動指示文・セマンティックセグメンテーションマスクの 3 つのモダリティを扱うことができる，Trimodal Navigable Region Segmentation Model (TNRSM) を提案する．このモデルを Talk2Car-Regseg データセットで検証した．結果として，提案手法は，移動指示理解タスクにおいて，全ての評価指標でベースライン手法を上回る性能を得られた．
Keywords: Understanding Navigation Instructions, Target Region Prediction, Segmentation

1. 研究背景・目的

本研究では，移動指示文を理解し目標領域のセグメンテーションマスクを生成するタスクを扱う．例えば，“木の横に駐車してください” という指示があった場合，木の横の領域を目標の駐車位置としてセグメンテーションマスクを生成することが望ましい．既存手法の [7] では，視覚モダリティのエンコーダモジュールは，マスクに関連するセグメンテーションタスクである RNR のように，セマンティックセグメンテーションレベルで画像の特徴量を考慮することができない．本研究では，言語，画像，セマンティックセグメンテーションマスクの 3 つのモダリティを扱うトリモーダルなエンコーダデコーダモデル TNRSM (Trimodal Navigable Region Segmentation Model) を提案する．これにより，画像全体とセマンティックセグメンテーションレベルで視覚モダリティを同時に考慮することにより，目標領域のセグメンテーションマスクタスクを解決することができる．

2. 方法

提案手法のネットワークは Trimodal Encoder および Trimodal Decoder の 2 つのモジュールから構成される．ネットワークの入力は画像 $\boldsymbol{X}_{\text{img}}$ および移動指示文 $\boldsymbol{X}_{\text{inst}}$ である．本研究では，事前学習済みモデルである RoBERTa [5] を用いて $\boldsymbol{X}_{\text{inst}}$ を言語特徴量 $\boldsymbol{L}_{\text{inst}}$ として埋め込む．また，CitySpace データセット [3] で学習された Mask2Former モデル [2] を用いてセマンティクスセグメンテーションマスク $\boldsymbol{X}_{\text{mask}}$ をゼロショット生成する．

Trimodal Encoder は 4 つのブロックで構成されており，各ブロックは 3 つのサブブロックで構成される．具体的には Text-Image Encoder (TIE) Block および Text-Mask Encoder (TME) Block の 2 つのサブブロックから構成されている．TIE Block は，Swin Transformer Block [6] および Pixel-Word Attention Module (PWAM) [8] から構成されている．

Trimodal Decoder では，以下の式のように第 i 層目の TIE および TME で得られた特徴量 $\boldsymbol{F}_{\text{img}}^{(i)}$ および $\boldsymbol{F}_{\text{mask}}^{(i)}$ を用いて最終的なセグメンテーションマスクを得る．

$$\boldsymbol{H}^{(i)} = \begin{cases} \left[\boldsymbol{F}_{\text{img}}^{(i)}; \boldsymbol{F}_{\text{mask}}^{(i)}\right] & i = M \\ \text{Conv}\left(\left[f_{\text{up}}\left(\boldsymbol{H}^{(i+1)}\right); \boldsymbol{F}_{\text{img}}^{(i)}; \boldsymbol{F}_{\text{mask}}^{(i)}\right]\right) & i \neq M \end{cases}$$

M, $[\cdot;\cdot]$, f_{up} はそれぞれ Trimodal Encoder のブロック数，チャネル方向の結合，バイリニア補間によるアップサンプリングを表す．モデルの最終的な出力は，予測されたセグメンテーションマスクに対応する予測確率を 0.5 の閾値で二値化しすることで得られる $H \times W$ 次元のセグメンテーションマスク $\hat{\boldsymbol{y}}$ である．

“pull over by that green trashcan on the right.”
(a) 正解　(b) ベースライン手法　(c) 提案手法
図 1 提案手法の成功例．(a): 正解．(b): ベースライン手法．(c): 提案手法．

3. 実験設定

本実験では，モデルの評価に Talk2Car-RegSeg のデータセット [7] を使用した．Talk2Car-RegSeg は，nuScenes [1] および Talk2Car [4] データセットに基づいて構築されたデータセットである．Talk2Car-RegSeg データセットは，Talk2Car データセットを元に，対象領域のセグメンテーションマスクを新たにアノテーションしている．

4. 結果

図 1 に成功例の定性的な結果を示す．図中の (a)，(b)，(c) は，それぞれ正解，ベースライン手法の予測結果，提案手法の予測結果を示す．画像に対する移動指示文は，“pull over by that green trashcan on the right.” であった．このとき，ベースライン手法では，対象領域に加えて，左側のゴミ箱付近も誤ってマスクしていたのに対し，提案手法では，対象領域を正しく予測することができた．

5. 結論

本研究では，移動指示文およびモビリティの正面カメラ画像が与えられた場合に，目標領域に対するセグメンテーションマスクを生成する RNR タスクに取り組んだ．将来研究としては，RNR タスクのマルチタスクとして，目標領域の矩形領域を予測するデコーダモジュールを新たに導入することが考えられる．

参考文献

[1] Holger Caesar, Varun Bankiti, et al. nuScenes: A multimodal dataset for autonomous driving. In *CVPR*, pages 11621–11631, 2020.

[2] Bowen Cheng, Ishan Misra, Alexander Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention Mask Transformer for Universal Image Segmentation. In *CVPR*, pages 1290–1299, 2022.

[3] Marius Cordts, Mohamed Omran, et al. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *CVPR*, pages 3213–3223, 2016.

[4] Thierry Deruyttere, Simon Vandenhende, Dusan Grujicic, et al. Talk2car: Taking control of your self-driving car. In *EMNLP-IJCNLP*, pages 2088–2098, 2019.

[5] Yinhan Liu, Myle Ott, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*, 2019.

[6] Ze Liu, Yutong Lin, et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *ICCV*, pages 10012–10022, 2021.

[7] Nivedita Rufus, Kanishk Jain, Unni Nair, Vineet Gandhi, and K Krishna. Grounding Linguistic Commands to Navigable Regions. In *IROS*, pages 8593–8600, 2021.

[8] Zhao Yang, Jiaqi Wang, et al. LAVT: Language-Aware Vision Transformer for Referring Image Segmentation. In *CVPR*, pages 18155–18165, 2022.