Title	Nearest Neighbor Future Captioning : 物体配置タスクにおける衝突リスクに関する説明文生成
Sub Title	
Author	小松, 拓実(Komatsu, Takumi)
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.31- 31
JaLC DOI	
Abstract	Although Domestic Service Robots (DSRs) that support people in everyday environments have been widely investigated, the DSR's ability to predict and describe future risks resulting from their own actions is still insufficient. In this study, we therefore focus on the linguistic explainability for the DSRs. Most existing methods do not explicitly model the region of possible collisions; thus, they do not properly generate descriptions for regions of possible collisions. In this paper, we propose Nearest Neighbor Future Captioning Model that introduces Nearest Neighbor Language Model to future captioning regarding possible collisions, which enhances the model output with a nearest neighbors retrieval mechanism. Moreover, we introduce Collision Attention Module, which extracts attention regions of possible collisions, which enables our model to generate descriptions that adequately reflect the objects associated with possible collisions. Experimental results demonstrated that our method outperformed baseline methods on the standard metrics.
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-6
Genre	Conference Paper
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=KO11003001-20230304-0031

Nearest Neighbor Future Captioning: 物体配置タスクにおける衝突リスクに関する説明文生成

小松 拓実
慶應義塾大学理工学部情報工学科

Abstract:
Although Domestic Service Robots (DSRs) that support people in everyday environments have been widely investigated, the DSR's ability to predict and describe future risks resulting from their own actions is still insufficient. In this study, we therefore focus on the linguistic explainability for the DSRs. Most existing methods do not explicitly model the region of possible collisions; thus, they do not properly generate descriptions for regions of possible collisions. In this paper, we propose Nearest Neighbor Future Captioning Model that introduces Nearest Neighbor Language Model to future captioning regarding possible collisions, which enhances the model output with a nearest neighbors retrieval mechanism. Moreover, we introduce Collision Attention Module, which extracts attention regions of possible collisions, which enables our model to generate descriptions that adequately reflect the objects associated with possible collisions. Experimental results demonstrated that our method outperformed baseline methods on the standard metrics.
Keywords: DSRs Vision&Language Nearest Neighbor Language Model explainability

1. Introduction

In our aging society, the shortage of home care workers has become a serious social problem. The Domestic Service Robots (DSRs) can be one of possible solutions to the social problem [2]. Delivering everyday objects in cluttered environments is a critical task for the DSRs. On the other hand, when the DSRs manipulate objects, there are risks of collisions with other objects, resulting in damage to the DSR's hand or the objects. It would be useful to explain and warn the possible collisions to the user using natural language. However, such a functionality is still insufficient. In this study, we focus on generating descriptions for possible collisions when the DSRs place objects.

2. Methods

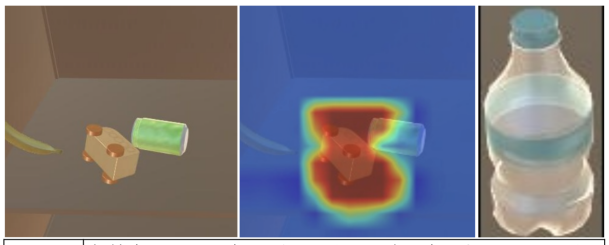
In this paper, the target task is collision-related future captioning. The task refers to the predictive task of generating a description of the future situation at time $t+k$ from an image before motion execution.

The main novelties of this paper are summarized as follows:

- We proposed Nearest Neighbor Future Captioning Model (NNFCM) that introduces Nearest Neighbor Language Model (NNLM) [1] into future captioning regarding possible collisions. This is one of the first attempts to introduce the NNLM into multimodal language generation.
- We introduce the Collision Attention Module, which extracts attention regions for possible collisions.

3. Results

In Fig 1, the target object is a "plastic bottle" and the collided object is a "toy wooden car". SAT incorrectly described the collided object as "hand" instead of "toy wooden car" Similarly, RFCM incorrectly described the collided object as "baby bottle" On the other hand, ours correctly described the target object and the collided object as "plastic bottle" and "toy wooden car," respectively.



Ref	把持中のペットボトルをおもちゃの木の車の上に配置しようとして、うまく置けずにペットボトルが倒れる
SAT	おもちゃの木の車とハンドが衝突する
RFCM	把持している空のペットボトルを哺乳瓶の上に配置しようとして倒れる
Ours	把持中のペットボトルをおもちゃの木の車の上に配置しようとして、うまく置けずに倒れる

Fig. 1 The successful example. In this sample, the target object is a "plastic bottle" and the collided object is a "toy wodden car."

4. Concolusions

In this paper, we proposed NNFCM that introduces NNLM [1] into future captioning regarding possible collisions. This is one of the first attempts to introduce NNLM into multimodal language generation.

References

[1] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, et al. Generalization through Memorization: Nearest Neighbor Language Models. In *ICLR*, 2019.
[2] Takashi Yamamoto et al. Development Human Support Robot as the Research Platform of a Domestic Mobile Manipulator. *ROBOMECH journal*, 6(1):1–15, 2019.