Title	ENCHANT : 大規模言語モデルを用いた仮説生成に基づくクロスモーダル説明文生成
Sub Title	
Author	小松, 拓実(Komatsu, Takumi) 平野, 慎之介( )
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.20- 20
JaLC DOI	
Abstract	本研究では，生活支援ロボット物体が物体を配置する際に生じる衝突を事前に予測し，衝突に関する説明文生成を行う．既存手法では衝突物体の特徴抽出が不適切であり，衝突に関係する物体の特定が不十分である．本研究では，大規模言語モデルを用いてデータ拡張を行うEnhanced Nearest-neighbor Captioning with Hypothesis AugmeNTation (ENCHANT) を提案する．提案手法の評価を行うため，一つの動画につき平均3.2人によってアノテーションされた4,042サンプルからなる新たなデータセットを構築した．実験の結果，提案手法がベースライン手法をすべての評価指標で上回った．
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=KO11003001-20240307-0020

ENCHANT:
大規模言語モデルを用いた仮説生成に基づくクロスモーダル説明文生成

小松 拓実 †, 平野慎之介 ‡
† 慶應義塾大学 理工学研究科 開放環境科学専攻 ‡ 慶應義塾大学 理工学部 情報工学科

Abstract:
本研究では，生活支援ロボット物体が物体を配置する際に生じる衝突を事前に予測し，衝突に関する説明文生成を行う．既存手法では衝突物体の特徴抽出が不適切であり，衝突に関係する物体の特定が不十分である．本研究では，大規模言語モデルを用いてデータ拡張を行う Enhanced Nearest-neighbor Captioning with Hypothesis AugmeNTation (ENCHANT) を提案する．提案手法の評価を行うため，一つの動画につき平均 3.2 人によってアノテーションされた 4,042 サンプルからなる新たなデータセットを構築した．実験の結果，提案手法がベースライン手法をすべての評価指標で上回った．
Keywords: DSRs Vision&Language Nearest Neighbor Language Model explainability

1. Introduction

高齢化が進行する現代社会において，在宅介護従事者の不足が深刻な社会問題となっている．生活支援ロボットはこうした社会問題に対する有望な解決策の一つとなり得る [3]．生活支援ロボットにとって，散らかった環境下で日常品を操作するタスクは重要なタスクである．一方，生活支援ロボットが物体を操作する際に，他の物体と衝突し，ロボットアームや物体が破損する危険性がある．そのため，衝突の危険性を事前に予測し，自然言語でユーザに説明し，警告する機能は有用である．しかし，そのような機能は未だ不十分である．本研究では，生活支援ロボットが物体を配置する際に発生する衝突を予測し，それに関する説明を生成することに焦点を当てる．

2. Methods

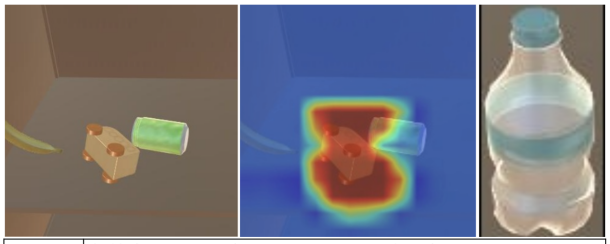
本論文では，衝突に関する future captioning を扱う．ここで，future captioning とは，動作前の画像から将来の状況の説明文を生成するタスクである．本タスクでは，物体配置時における物体間の衝突を予測し，予測された衝突に関する説明文を出力することが望ましい．

提案手法における新規性は以下の通りである．

- 大規模言語モデルによる生成文を用いてデータ拡張を行う NNAM を導入し，生成された仮説に基づき衝突に関する説明文を生成する ENCHANT を提案する．
- 画像および言語の特徴抽出を対称的に行う PCAD を導入する．
- 衝突リスクの注意マップ [2] と，セグメンテーションモデル [1] を用いて配置領域の特徴量を抽出する SFE を導入する．

3. Results

定性的結果を図 1 に示す．図 1 の左図および右図はそれぞれ配置領域の RGB 画像および attention map を配置領域の RGB 画像に重畳した画像を表す．図 1 において，把持している物体は「ペットボトル」であり，衝突する物体は「おもちゃの木の車」である．SAT は衝突した物体を「ハンド」と誤って記述した．同様に，RFCM は衝突した物体を「哺乳瓶」と誤って記述した．一方で，提案手法は，対象物体および衝突した物体をそれぞれ「ペットボトル」「おもちゃの木の車」と適切に記述した．



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 提案手法における成功例． この例において，対象物体は「ペットボトル」であり，衝突する物体は「おもちゃの木の車」である．

4. Concolusions

本論文では，生活支援ロボットが物体を配置する際に発生する衝突に関する future captioning タスクを扱った． 本論文では，大規模言語モデルによる生成文を用いてデータ拡張を行う Nearest Neighbor Autmentation Module，画像および言語の特徴抽出を対称的に行う Parallel Cross Attention Decoder， 配置領域の特徴量をセグメンテーションを用いて抽出する Segment Feature Extractor を導入した．実験結果より，全ての評価指標において，ベースライン手法を上回った．

References

[1] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, et al. Segment Anything. *arXiv:2304.02643*, 2023.
[2] Aly Magassouba, Komei Sugiura, Angelica Nakayama, et al. Predicting and Attending to Damaging Collisions for Placing Everyday Objects in Photo-Realistic Simulations. *AR*, 35(12):787–799, 2021.
[3] Takashi Yamamoto et al. Development Human Support Robot as the Research Platform of a Domestic Mobile Manipulator. *ROBOMECH journal*, 6(1):1–15, 2019.