慶應義塾大学学術情報リポジトリ
Keio Associated Repository of Academic resouces

Title	マルチモーダル言語理解における敵対的データ拡張基盤の構築
Sub Title	Adversarial data augmentation for multimodal language understanding
Author	杉浦, 孔明(Sugiura, Komei)
Publisher	
Publication year	2023
Jtitle	科学研究費補助金研究成果報告書 (2022.)
JaLC DOI	
Abstract	本研究では，マルチモーダル言語理解，マルチモーダル言語生成，Sim2Real転移学習と介助犬タスクでの評価を行った．理解班では，Vision-and-Language Navigationタスクにおいて，敵対的摂動更新アルゴリズム Momentum-based Adversarial Trainingを構築した．生成班では，動画から将来の状況を説明する文を生成するfuture captioning手法を構築し，既存手法を上回る結果を得た．Sim2Real班では，生活支援ロボット評価フレームワークを構築し，指示文生成とタスク実行を自動化した． In this study, our objectives are (a) robust multimodal language understanding through adversarial data augmentation, (b) multimodal language generation, and (c) evaluation in the assistance dog tasks. We first focused on the Vision-and-Language Navigation task and developed the Momentum-based Adversarial Training (MAT) algorithm. We applied MAT to the standard benchmark test, ALFRED, and obtained successful results. We also worked on the task of generating descriptions about future situations. The main novelty of our proposed method lies in the use of Relational Self-Attention as the attention mechanism. Experimental results show that our method outperformed existing methods in standard metrics. We applied the multimodal language understanding and generation methods into a simulator, enabling on-the-fly instruction generation. As a result, we established a robot evaluation framework that does not require manual intervention in task generation, execution, and evaluation.
Notes	研究種目：基盤研究 (B) (一般) 研究期間：2020～2022 課題番号：20H04269 研究分野：機械知能, 知能ロボティクス, マルチモーダル言語処理
Genre	Research Paper
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=KAKEN_20H04269seika

慶應義塾大学学術情報リポジトリ(KOARA)に掲載されているコンテンツの著作権は、それぞれの著作者、学会または出版社/発行者に帰属し、その権利は著作権法によって保護されています。引用にあたっては、著作権法を遵守してご利用ください。

The copyrights of content available on the KeiO Associated Repository of Academic resources (KOARA) belong to the respective authors, academic societies, or publishers/issuers, and these rights are protected by the Japanese Copyright Act. When quoting the content, please follow the Japanese copyright act.

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月　 ６ 日現在

機関番号：３２６１２
研究種目：基盤研究(B)（一般）
研究期間：2020～2022
課題番号：２０Ｈ０４２６９

研究課題名（和文）マルチモーダル言語理解における敵対的データ拡張基盤の構築

研究課題名（英文）Adversarial Data Augmentation for Multimodal Language Understanding

研究代表者
杉浦　孔明（Sugiura, Komei）

慶應義塾大学・理工学部（矢上）・教授

研究者番号：６０４７０４７３

交付決定額（研究期間全体）：（直接経費）　　13,500,000円

研究成果の概要（和文）：本研究では，マルチモーダル言語理解，マルチモーダル言語生成，Sim2Real転移学習と介助犬タスクでの評価を行った．理解班では，Vision-and-Language Navigationタスクにおいて，敵対的摂動更新アルゴリズム Momentum-based Adversarial Trainingを構築した．生成班では，動画から将来の状況を説明する文を生成するfuture captioning手法を構築し，既存手法を上回る結果を得た．Sim2Real班では，生活支援ロボット評価フレームワークを構築し，指示文生成とタスク実行を自動化した．

研究成果の学術的意義や社会的意義
本研究では，要支援者とその家族を時間的拘束から解放するために，日常タスクを支援する生活支援ロボットの言語理解技術構築を目的とする．生活支援ロボットのハードウェアは最近標準化されたものの，曖昧な指示を理解する精度は不十分である．本研究では，マルチモーダル言語理解に関する標準データセット上で世界最高精度を達成するとともに，タスク生成・実行・評価のすべてにおいて人手を要しない生活支援ロボット評価フレームワークを世界で初めて構築した．

研究成果の概要（英文）：In this study, our objectives are (a) robust multimodal language understanding through adversarial data augmentation, (b) multimodal language generation, and (c) evaluation in the assistance dog tasks.
We first focused on the Vision-and-Language Navigation task and developed the Momentum-based Adversarial Training (MAT) algorithm. We applied MAT to the standard benchmark test, ALFRED, and obtained successful results. We also worked on the task of generating descriptions about future situations. The main novelty of our proposed method lies in the use of Relational Self-Attention as the attention mechanism. Experimental results show that our method outperformed existing methods in standard metrics. We applied the multimodal language understanding and generation methods into a simulator, enabling on-the-fly instruction generation. As a result, we established a robot evaluation framework that does not require manual intervention in task generation, execution, and evaluation.

研究分野：機械知能，知能ロボティクス，マルチモーダル言語処理

キーワード：マルチモーダル言語処理　クロスモーダル言語生成　データ拡張　生活支援ロボット　Sim2Real

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属します。

様　式　Ｃ－１９，Ｆ－１９－１，Ｚ－１９（共通）

１．研究開始当初の背景

少子高齢化社会のなかで要支援者を支える生産年齢人口は減少しており，家族が離職を余儀なくされるケースが発生するなど，社会全体の生産性向上を妨げている．その解決手段として，生活支援ロボットの研究開発が活発に進められている．

本研究では，要支援者とその家族を時間的拘束から解放するために，日常タスクを支援する生活支援ロボットを実現する．生活支援ロボットのハードウェアは最近標準化されたものの，曖昧な指示を理解する精度は不十分である．

そこで本研究では，(a) 敵対的データ拡張によるマルチモーダル言語理解，(b)マルチモーダル言語生成による学習データ大規模化，敵対的データ拡張の基盤技術確立，(c) Sim2Real アプローチによる転移学習と介助犬タスクでの評価，を目的とする．

２．研究の目的

本研究は，曖昧なユーザ指示に対するマルチモーダル言語理解・生成の基盤技術を確立するとともに，介助犬レベルのタスクを概ね実用レベルの精度で行う生活支援ロボットの構築を目的とする．研究グループを３班に分け，理解班・生成班・Sim2Real 班として，本研究を遂行する．

（１）マルチモーダル言語理解

人間がロボットにある物体を取ってくるように命令を与えたときに，ロボットが命令内容を適切に解釈し，対象物体を特定することを目的とする．具体的には，“Grab the plastic bottle with red stripes and put it in the upper left box” という命令が与えられたときに，ロボットが赤い縞模様の瓶を対象物体として認識することが望ましい．

しかし，人間の発する命令文には事前に定義されるような規則が存在しないため，含まれる情報が不十分な場合が多く，しばしば内容に曖昧性が生じる．例えば，上述した命令文について，同じ空間に瓶が複数ある場合，文のみから正しい対象物体を特定することは容易ではない．

既存手法では，命令文に加え，対象物体を含む全体画像を入力することで，言語的知識だけではなく視覚的知識を活用することを試みている．しかし，命令文には画像中の物体に関する参照表現が含まれている場合が多く，全体画像の入力では物体間の関係性を学習するのが困難である．加えて，既存手法は他のタスクからの転移学習を実行できない．

（２）マルチモーダル言語生成

生活支援ロボットが動作実行前にタスクの実行に伴う危険性を予測し，ユーザに判断を仰ぐ機能は，安全性及び利便性を高める．例えば，物体を配置する際に他の物と衝突した場合，連鎖的に衝突が起こり物体が破損する危険性がある．こうした危険性について生活支援ロボットが事前に予測し，自然言語を用いてユーザに注意喚起できることは，衝突等の危険を未然に防ぐことにつながる．一方，この機能は未だに不十分である．

上記の背景から，時刻 t までの系列データを基に，時刻 t+1 で起こるイベントについての説明文を生成する future captioning タスクを取り扱う．例えば，生活支援ロボットがペットボトルを棚に置く際に「ロボットのアームがマグカップに接触することで，マグカップがその隣りにあるグラスに更に接触し，グラスが倒れる危険性があります」のような文を動作実行前にユーザに提示することが望ましい．しかし，本タスクは，モデルが将来のイベントを表す画像情報を利用できないという点で難しい．そのため，過去の系列データを用いた将来の画像の予測，およびキャプションの生成という 2 つの要素が求められる．

本項目では，future captioning モデルを構築する．本手法では，時刻 t における画像特徴量についての再構成損失を導入することにより，物体に関して適切な記述を生成する．加えて，CLIP [1]で用いられている損失を導入することにより，対応する画像と言語の特徴量の類似度を高め，適切なイベントについての説明文を生成する．以上により，物体に関して適切なキャプションの生成が期待される．

（３）Sim2Real アプローチによる転移学習

ALFRED を始めとする多くの既存フレームワークでは，指示文を人手により付与しているため，on-the-fly なシミュレーションとすることが難しい．それゆえ，ランダムに作成した多様なタスクで評価することも困難であり，固定されたタスクのみで評価を行っていた．

これに対し提案フレームワークではシミュレーション環境上での Fetch-and-Carry with Object Grounding (FCOG)タスクについて，完全自動化のためのフレームワークを提案する．本フレームワークにおけるタスク生成システムでは，クロスモーダル指示文生成モデルにより指示文を生成している．そのため，自由形式な指示文を用いたタスクの実行が可能となる．また，提案手法は，FCOG タスクについての自由形式な指示文に対して，参照表現を基に対象物体および目標領域を特定し指示を実行する．

提案フレームワークは，クロスモーダル指示文生成を含むタスク生成システムを導入した点で既存フレームワークと異なる．クロスモーダル指示文生成を含むタスク生成システムの導入により，提案フレームワークは，自由形式な自然言語指示文を用いた on-the-fly な FCOG タスクの実行が可能となる．

３．研究の方法

（１）マルチモーダル言語理解

提案手法では，全体画像の代わりに画像中の各物体の領域を入力することで，対象物体と他の物体の関係性をより直接的に学習する Target-dependent UNITER モデルを提案する．既存手法と異なる点は，画像とテキストの共同理解に UNITER[2]を採用し，対象物体候補の画像・位置情報を扱うように構造を変更した点である．UNITER を使用することにより，Transformer [3]内の注意機構に基づいて画像とテキストの関係性を学習することができるため，より深い言語理解が獲得できると考えられる．また，対象物体候補の情報を入力に追加することで，対象物体に関する判定を直接的に行うことが可能となる．

提案手法の独自性は以下である．

1. 物体操作指示理解分野において，画像とテキストの関係性の学習における UNITER 型注意機構と汎用事前学習モデルを導入する．
2. UNITER において，対象物体候補を扱う新規構造を導入する．

提案手法の構造を図 1 に示す．図において，Instruction は命令文，Target Region は対象物体候補の領域，Context Regions は画像中の各物体の領域を表す．ネットワークは大きく分けて Image Embedder，Text Embedder，Multi-layer Transformer といった 3 つのモジュールから構成される．Text Embedder は 2 つの埋め込み層と正規化層から構成され，Image Embedder は 2 つの全結合層と正規化層から構成される．Multi-layer Transformer は Transformer を複数層重ねたものである．手法の詳細は[6]を参照されたい．

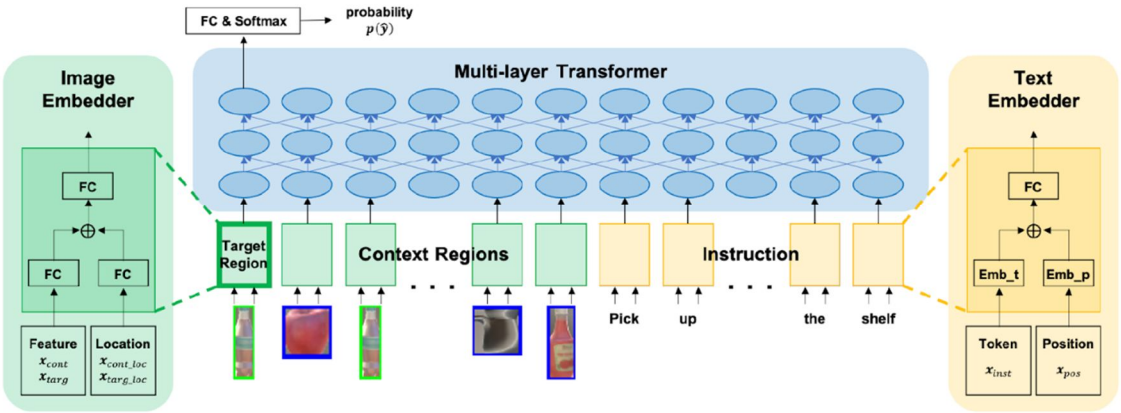


図１　Target-Dependent UNITER の構造の概要

（２）マルチモーダル言語生成

提案手法は 3 つのモジュールから構成され，それぞれ Relational Self-Attention (RSA[4]) エンコーダ，transformer エンコーダ，および transformer デコーダである．提案手法は RSA エンコーダによって，過去のイベントとの関係性を適切に考慮した将来のイベント表現を獲得できる．なぜなら RSA は，既存の注意機構よりも効果的に，過去のイベントのうちどのイベントに注目すればよいかを学習できるためである．また，transformer デコーダは過去のイベントの関係性から，適切に将来イベントのキャプションを生成できる．transformer デコーダは，RSA エンコーダの出力をクエリとし，transformer エンコーダの出力をキー，およびバリューとする source-target 注意機構を持つ．これによって，過去のイベント間の関係性を自然言語に適切に接地できる．

提案手法の新規性は以下の３点である．手法の詳細は[7]を参照されたい．

1. Future captioning タスクのためのクロスモーダル言語生成モデル，RFCM を提案する．
2. RSA エンコーダの導入により，既存の自己注意機構に比べ，より効果的にイベント間の関係性を抽出できる．
3. 再構成損失および CLIP loss の導入

（３）Sim2Real アプローチによる転移学習

提案フレームワークにおけるタスク実行システムは，既存システムと異なり，FCOG タスクにおいてマルチモーダル言語理解モデルを用いた指示文理解を行う．本論文の主要な貢献は以下である．

1. FCOG タスクにおいて，生成，実行，および評価についての完全自動化のための，自由形式な自然言語指示文のクロスモーダル言語生成を含むフレームワークを提案する．

2. FCOG タスクに対して，Navigation, Object Location Retrieval (OLR), Fetching, 及び Carrying の 4 つのサブタスクに分割し解決するアプローチを提案する．
3. OLR タスクのためのマルチモーダル言語理解モデルにおいて，言語特徴量および画像特徴量を適切にモデリングするための Multimodal Parallel Feature Extractor (MPFE)を導入する．

提案フレームワークは，タスク生成システム，タスク実行システム，およびタスク評価システムの 3 つのシステムから構成される．手法の詳細は，[8]を参照されたい．

４．研究成果

（１）マルチモーダル言語理解

提案手法の検証のため，データセットとして PFN-PIC と WRS-UniALT を使用した．定量的結果を表 1 に示す．データセット内に正解サンプルと不正解サンプルが等量で存在するため，チャンスレートは 50%である．

定性的結果を右図に示す．図において，緑色で囲まれている領域がデータセットに記載されている座標値に基づく真の対象領域であり，青色で囲まれている領域が Faster R-CNN によって検出した対象領域候補である．命令文は"Pick up the black cup in the bottom right section of the box and move it to the bottom left section of the box"であり，対象物体は右下の区画にある黒色のカップである．青色で囲まれている領域について，対象領域であると判定できている．

表 1 に定量的結果を示す．PFN-PIC において，提案手法の精度は 96.9%，ベースライン手法は 90.1%であった．また，WRS-UniALT において，提案手法の精度は 96.4%，ベースライン手法は 91.8%であった．これより，提案手法は，PFN-PIC と WRS-UniALT において，ベースライン手法をそれぞれ 6.8%，4.6%上回っていることがわかる．

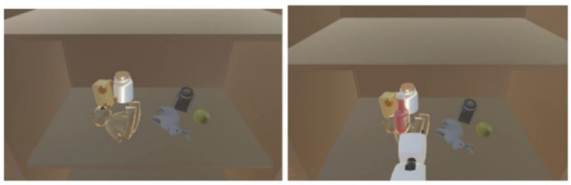
表１　提案手法およびベースライン手法のマルチモーダル言語理解精度

Method	Accuracy [%]	
	PFN-PIC	WRS-UniALT
Baseline (MTCM [Magassouba 19])	90.1 ± 0.93	91.8 ± 0.36
(i) Ours (FRCNN fine-tuning なし)	91.5 ± 0.69	94.0 ± 1.49
(ii) Ours (Late fusion)	96.0 ± 0.08	96.0 ± 0.24
(iii) Ours (Few context regions)	96.6 ± 0.36	95.8 ± 0.71
(iv) Ours (Pretraining なし)	96.8 ± 0.34	95.4 ± 0.19
Ours (Target-dependent UNITER)	$\mathbf{96.9 \pm 0.34}$	$\mathbf{96.4 \pm 0.24}$

（２）マルチモーダル言語生成

提案手法を評価するため，データセットを構築した．WRS2018 パートナーロボットチャレンジ／バーチャルスペースコンペティションにおいて使用されたシミュレータを拡張したものを使用した．シミュレーションでは，生活支援ロボットがランダムに選択されたボトルや缶などの日用品を 5 種類の机や棚などの家具の中央に配置する．生活支援ロボットのヘッドカメラから撮影した映像を収集した．それぞれのサンプルには"the apple rolled over because the rabbit figure next to it was pushed by the robot"などの状況を説明する文が付与された．データセットは 1000 本の動画および，衝突イベントに対して付与された英語の説明文 1000 文からなる．

右図に提案手法の定性的結果を示す．図において，把持されている物体は「red bottle」，衝突した物体は「stuffed bear」である．ベースライン手法ではそれぞれ「the hourglass」と「the apple and the stuffed bear」と誤って記述した．一方，提案手法では，それぞれ「a red bottle」と「a teddy bear」と適切に記述した．



GT : “Robot bumps into the stuffed bear because robot tried to put the red bottle where it is.”
Baseline : “Robot hits the apple and the stuffed bear hard because robot tried to put the hourglass where they are.”
Ours : “Robot rubs the hand on a teddy bear because robot tried to put a red bottle.”

RFCM 及び Memory-Augmented Recurrent Transformer (MART[5]) について比較を行った．MART は，動画キャプション生成タスクにおける代表的な手法の一つであり，Future captioning タスクへの適用も可能であったため，ベースライン手法とした．表 1 に定量的な結果を示す．各手法につき実験を 5 回行い，

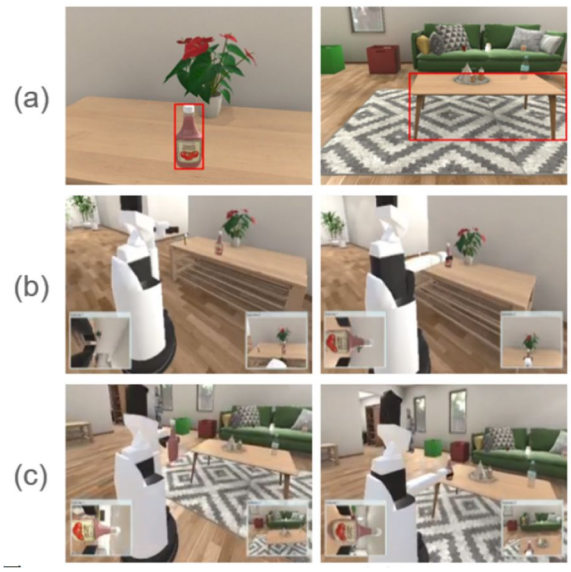
表にはその平均値および標準偏差を示す．評価尺度として，動画キャプション生成タスクにおける標準尺度である BLEU4，ROUGE-L，METEOR，及び CIDEr-D により行った．表 2 より，主要尺度である CIDEr-D において，提案手法及びベースライン手法はそれぞれ 60.37，および 49.61 であり，提案手法が 10.76 ポイント向上した．これらの結果より，Future captioning タスクにおいて RFCM はベースライン手法よりも適切な文の生成が可能であることがわかった．

表 2　提案手法およびベースライン手法のマルチモーダル言語生成品質評価

手法	BLEU4	METEOR	ROUGE-L	CIDEr-D
RFCM	21.74 ± 1.02	22.74 ± 0.57	41.44 ± 0.86	49.61 ± 8.02
(i)	24.55 ± 1.11	24.21 ± 0.66	**46.10 ± 0.48**	49.05 ± 3.56
(ii)	24.24 ± 0.98	24.26 ± 0.79	44.18 ± 1.03	57.32 ± 1.73
Ours	**24.82 ± 1.14**	**24.39 ± 0.73**	44.67 ± 1.13	**60.37 ± 4.31**

（3）Sim2Real アプローチによる転移学習

右図に FCOG タスクについての定性的結果を示す．図における(a)は，タスク生成システムが取得した，対象物体および目標領域についての画像を示す．ここで，赤い矩形は Unity から取得したセグメンテーションに基づいて付与したものである．タスク生成システムは対象物体および目標領域として，赤いボトルおよびソファの前のテーブルを選択した．生成された指示文は"Go to the living room, move a plastic bottle from the shelf to the table"であった．ロボットは，Navigation タスクにおいて，指示文に基づきリビングへ移動することに成功した．続いて，OLR タスクにおいて，対象物体および目標領域として，赤いボトルおよび机の前のテーブルを適切に特定できた．その後，図(b)に示すように，ロボットは Fetching タスクにおいて赤いボトルを把持できた．最終的に，図(c)に示すように，Carrying タスクにおいて，テーブルへ赤いボトルを配置することに成功した．



参考文献

[1] Radford, A., Kim, J., Hallacy, C., Goh, G., Agarwal, S., Sastry, G., et al., “Learning Transferable Visual Models From Natural Language Supervision,” ICML, 2021.
[2] Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J.: Uniter: Universal image-text representation learning, in ECCV, pp. 104-120 (2020)
[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I.: Attention is all you need, in NeurIPS, pp. 5998-6008 (2017)
[4] Kim, M., Kwon, H., Wang, C., et al.: Relational Self-Attention: What’s Missing in Attention for Video Understanding, in NeurIPS (2021)
[5] Lei, J., Wang, L., et al.: MART: Memory-Augmented Recurrent Transformer for Coherent Video Paragraph Captioning, in ACL, pp. 2603–2614 (2020)
[6] S. Ishikawa and K. Sugiura, “Target-dependent UNITER: A Transformer-Based Multimodal Language Comprehension Model for Domestic Service Robots”, IEEE Robotics and Automation Letters, Vol. 6, Issue 4, pp. 8401-8408, 2021.
[7] M. Kambara and K. Sugiura, “Case Relation Transformer: A Crossmodal Language Generation Model for Fetching Instructions”, IEEE Robotics and Automation Letters, Vol. 6, Issue 4, pp. 8371-8378, 2021.
[8] 神原元就，杉浦孔明：“記号接地された fetch-and-carry タスクの自動化と実行”，第 40 回日本ロボット学会学術講演会, 4I1-01, 2022.

５．主な発表論文等

〔雑誌論文〕　計7件（うち査読付論文　7件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Ishikawa Shintaro、Sugiura Komei	４．巻 11
２．論文標題 Affective Image Captioning for Visual Artworks Using Emotion-Based Cross-Attention Mechanisms	５．発行年 2023年
３．雑誌名 IEEE Access	６．最初と最後の頁 24527～24534
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ACCESS.2023.3255887	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Kambara Motonari、Sugiura Komei	４．巻 6
２．論文標題 Case Relation Transformer: A Crossmodal Language Generation Model for Fetching Instructions	５．発行年 2021年
３．雑誌名 IEEE Robotics and Automation Letters	６．最初と最後の頁 8371～8378
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/LRA.2021.3107026	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Ishikawa Shintaro、Sugiura Komei	４．巻 6
２．論文標題 Target-Dependent UNITER: A Transformer-Based Multimodal Language Comprehension Model for Domestic Service Robots	５．発行年 2021年
３．雑誌名 IEEE Robotics and Automation Letters	６．最初と最後の頁 8401～8408
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/LRA.2021.3108500	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Magassouba Aly、Sugiura Komei、Kawai Hisashi	４．巻 6
２．論文標題 CrossMap Transformer: A Crossmodal Masked Path Transformer Using Double Back-Translation for Vision-and-Language Navigation	５．発行年 2021年
３．雑誌名 IEEE Robotics and Automation Letters	６．最初と最後の頁 6258～6265
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/LRA.2021.3092686	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Ogura Tadashi、Magassouba Aly、Sugiura Komei、Hirakawa Tsubasa、Yamashita Takayoshi、Fujiyoshi Hironobu、Kawai Hisashi	４．巻 5
２．論文標題 Alleviating the Burden of Labeling: Sentence Generation by Attention Branch Encoder?Decoder Network	５．発行年 2020年
３．雑誌名 IEEE Robotics and Automation Letters	６．最初と最後の頁 5945～5952
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/LRA.2020.3010735	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Magassouba Aly、Sugiura Komei、Nakayama Angelica、Hirakawa Tsubasa、Yamashita Takayoshi、Fujiyoshi Hironobu、Kawai Hisashi	４．巻 -
２．論文標題 Predicting and attending to damaging collisions for placing everyday objects in photo-realistic simulations	５．発行年 2021年
３．雑誌名 Advanced Robotics	６．最初と最後の頁 1～13
掲載論文のDOI（デジタルオブジェクト識別子） 10.1080/01691864.2021.1913446	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Magassouba Aly、Sugiura Komei、Kawai Hisashi	４．巻 5
２．論文標題 A Multimodal Target-Source Classifier With Attention Branches to Understand Ambiguous Instructions for Fetching Daily Objects	５．発行年 2020年
３．雑誌名 IEEE Robotics and Automation Letters	６．最初と最後の頁 532～539
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/LRA.2019.2963649	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計30件（うち招待講演　3件／うち国際学会　5件）

１．発表者名 W. Yang, A. Ueda, and K. Sugiura
２．発表標題 Multimodal Encoder with Gated Cross-attention for Text-VQA Tasks
３．学会等名 言語処理学会第29回年次大会
４．発表年 2023年

１．発表者名
石川慎太朗，杉浦孔明

２．発表標題
ゲート付き相互注意を用いたエンコーダ・デコーダによる感情に基づく絵画説明文生成

３．学会等名
言語処理学会第29回年次大会

４．発表年
2023年

１．発表者名
植田有咲，Wei Yang，杉浦孔明

２．発表標題
マルチモーダルOCR特徴を用いたDynamic Pointer Networkによるテキスト付き画像説明文生成

３．学会等名
言語処理学会第29回年次大会

４．発表年
2023年

１．発表者名
和田唯我，兼田寛大，杉浦孔明

２．発表標題
JaSPICE: 日本語における述語項構造に基づく画像キャプション生成モデルの自動評価尺度

３．学会等名
言語処理学会第29回年次大会

４．発表年
2023年

１．発表者名
K. Sugiura

２．発表標題
Visual and Linguistic Explanations in Semantic Machine Intelligence

３．学会等名
Shonan Meeting No. 166（招待講演）（国際学会）

４．発表年
2023年

１．発表者名
K. Sugiura

２．発表標題
Towards Superhuman and Explainable AI for Human-AI Co-Evolution

３．学会等名
AIST Artificial Intelligence Research Center International Symposium（招待講演）（国際学会）

４．発表年
2023年

１．発表者名
M. Kambara, K.Sugiura,

２．発表標題
Relational Future Captioning Model for Explaining Likely Collisions in Daily Tasks

３．学会等名
IEEE ICIP（国際学会）

４．発表年
2022年

１．発表者名
H. Matsuo, S. Hatanaka, A. Ueda, T. Hirakawa, T. Yamashita, H. Fujiyoshi, K. Sugiura

２．発表標題
Collision Prediction and Visual Explanation Generation Using Structural Knowledge in Object Placement Tasks

３．学会等名
IEEE/RSJ IROS

４．発表年
2022年

１．発表者名
R. Korekata, Y. Yoshida, S. Ishikawa, K. Sugiura

２．発表標題
Switching Funnel UNITER: Multimodal Instruction Comprehension for Object Manipulation Tasks

３．学会等名
IEEE/RSJ IROS

４．発表年
2022年

１．発表者名
神原元就，杉浦孔明

２．発表標題
記号接地されたfetch-and-carryタスクの自動化と実行

３．学会等名
第40回日本ロボット学会学術講演会

４．発表年
2022年

１．発表者名
小槻誠太郎，石川慎太朗，杉浦孔明

２．発表標題
TDP-MATに基づく実画像を対象とした物体操作指示理解

３．学会等名
第40回日本ロボット学会学術講演会

４．発表年
2022年

１．発表者名
是方諒介，吉田悠，石川慎太朗，杉浦孔明

２．発表標題
物体操作タスクにおけるSwitching Funnel UNITERによる対象物体および配置目標に関する指示文理解

３．学会等名
第40回日本ロボット学会学術講演会

４．発表年
2022年

１．発表者名
飯岡雄偉，神原元就，杉浦孔明

２．発表標題
物体配置タスクにおける危険性のクロスモーダル説明生成

３．学会等名
第40回日本ロボット学会学術講演会

４．発表年
2022年

１．発表者名
松尾榛夏，畑中駿平，平川翼，山下隆義，藤吉弘亘，杉浦孔明

２．発表標題
物体配置タスクにおける構造的知識を用いた衝突予測および視覚的説明生成

３．学会等名
第40回日本ロボット学会学術講演会

４．発表年
2022年

１．発表者名
S. Ishikawa, K. Sugiura

２．発表標題
Moment-based Adversarial Training for Embodied Language Comprehension

３．学会等名
IEEE ICPR（国際学会）

４．発表年
2022年

１．発表者名
K. Sugiura

２．発表標題
Semantic Machine Intelligence for Domestic Service Robots

３．学会等名
Sixth International Workshop on Symbolic-Neural Learning（招待講演）（国際学会）

４．発表年
2022年

１．発表者名
石川慎太朗，杉浦孔明

２．発表標題
Vision-and-Language Navigationタスクにおける敵対的サブゴール生成

３．学会等名
2022年度 人工知能学会全国大会

４．発表年
2022年

１．発表者名
神原元就，杉浦孔明

２．発表標題
日常タスクにおける将来イベントのクロスモーダル説明文生成

３．学会等名
2022年度 人工知能学会全国大会

４．発表年
2022年

１．発表者名
吉田悠，石川慎太朗，杉浦孔明

２．発表標題
生活支援ロボットによる物体操作タスクにおけるFunnel UNITERに基づく指示文理解

３．学会等名
2022年度 人工知能学会全国大会

４．発表年
2022年

１．発表者名
神原元就，杉浦孔明

２．発表標題
料理タスクにおける将来イベントのクロスモーダル説明文生成

３．学会等名
第28回画像センシングシンポジウム

４．発表年
2022年

１．発表者名
植田有咲，杉浦孔明

２．発表標題
参照画像と修正指示文を用いた Multimodal Modulation によるファッション画像検索

３．学会等名
第28回画像センシングシンポジウム 2022

４．発表年
2022年

１．発表者名
T. Matsubara, S.Otsuki, Y. Wada, H. Matsuo, T. Komatsu, Y. Iioka, K. Sugiura, H. Saito

２．発表標題
Shared Transformer Encoder with Mask-based 3D Model Estimation for Container Mass Estimation

３．学会等名
IEEE ICASSP

４．発表年
2022年

１．発表者名
S. Matsumori, K. Shingyouchi, Y. Abe, Y. Fukuchi, K. Sugiura, M. Imai

２．発表標題
Unified Questioner Transformer for Descriptive Question Generation in Goal-Oriented Visual Dialogue

３．学会等名
IEEE ICCV

４．発表年
2021年

１．発表者名
植田有咲, Aly Magassouba, 平川翼, 山下隆義, 藤吉弘亘, 杉浦孔明

２．発表標題
生活支援ロボットによる物体配置タスクにおけるTransformer PonNetに基づく危険性予測および可視化

３．学会等名
2021年度 人工知能学会全国大会

４．発表年
2021年

１．発表者名
神原元就, 杉浦孔明

２．発表標題
Case Relation Transformerに基づく対象物体及び目標領域の参照表現を含む物体操作指示文生成

３．学会等名
2021年度 人工知能学会全国大会

４．発表年
2021年

１．発表者名
石川慎太朗，杉浦孔明

２．発表標題
Target-dependent UNITERに基づく対象物体に関する参照表現を含む物体操作指示理解

３．学会等名
2021年度 人工知能学会全国大会

４．発表年
2021年

１．発表者名
兼田寛大，神原元就，杉浦孔明

２．発表標題
Bilingual Case Relation Transformerに基づく複数言語による物体操作指示文生成

３．学会等名
第39回日本ロボット学会学術講演会

４．発表年
2021年

１．発表者名
畑中駿平，上田雄斗，植田有咲，平川翼，山下隆義，藤吉弘亘，杉浦孔明

２．発表標題
生活支援ロボットによる物体配置タスクにおける危険性予測および視覚的説明生成

３．学会等名
第39回日本ロボット学会学術講演会

４．発表年
2021年

１．発表者名
飯田紡，九曜克之，石川慎太朗，杉浦孔明

２．発表標題
物体指示理解タスクにおけるクロスモーダル言語生成に基づくデータ拡張

３．学会等名
第39回日本ロボット学会学術講演会

４．発表年
2021年

１．発表者名
小椋忠志，Magassouba Aly，杉浦孔明，平川翼，山下隆義，藤吉弘亘，河井恒

２．発表標題
Multimodal Attention Branch Networkに基づく把持命令文の生成

３．学会等名
2020年度 人工知能学会全国大会

４．発表年
2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕
杉浦孔明研究室ウェブサイト
https://smilab.org/

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関