マルチモーダル言語処理に基づく Fetch-and-Carry タスクの自動化と実行

Automation and Execution of Fetch-and-Carry Tasks Based on Multi-Modal Language Processing

神原 元就 *1　　杉浦 孔明 *1
Motonari Kambara　　Komei Sugiura

*1 慶應義塾大学
Keio University

In this paper, we address the Fetch-and-Carry with Object Grounding (FCOG) task, where a robot executes free-form natural language instructions using visual information. We introduce a Multimodal Parallel Feature Extractor for the language understanding model of the Object Location Retrieval task. We also propose a framework for fully automating the simulation of free-form natural language instructions. Experimental results show that our method outperformed the baseline method in the reference expression understanding task.

1. はじめに

被介助者の増加により，介助者の不足が社会課題の一つとなっている．これについて，ユーザと自然言語を用いてコミュニケーションが可能な生活支援ロボットの実用化は有望な解決策の一つである．一方自然言語による指示を理解し適切に日常タスクを実行する能力は現状では不十分である．

本論文では，Fetch–and–Carry タスクについての自然言語指示文が与えられた上で，ロボットが視覚情報を基に指示を実行するための手法を構築することを目的とする．

人間の発する自然言語指示はしばしば曖昧であり，ロボットが対象物体及び目標領域を特定することは困難である．特に，「花瓶の隣」等の参照表現を基に物体を特定することについて，既存システムは十分にできているとは言えない．

また，ALFRED [Shridhar 20] 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．

提案手法の類似手法として，SayCan [Ahn 22] が挙げられる．SayCan は与えられた自由形式な自然言語指示文をロボットが実行する点で提案手法と類似している．一方，SayCan では実行するサブタスク系列を生成するため言語モデルを使用しているが，提案手法は，対象物体及び目標領域を特定するためにマルチモーダル言語理解モデルを用いている点で異なる．

提案手法は，Fetch–and–Carry with Object Grounding (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及び目標領域を特定し指示を実行する．また，本論文ではシミュレーション環境上での 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．

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．

連絡先: 神原元就，慶應義塾大学，神奈川県横浜市港北区日吉 3-14-1，motonari.k714@keio.jp

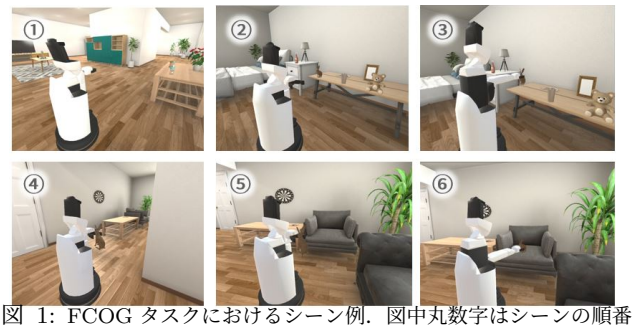

図 1: FCOG タスクにおけるシーン例．図中丸数字はシーンの順番を示す．

また，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．

- FCOG タスクにおいて，生成，実行，及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．
- 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．
- 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．

2. 問題設定

本論文で扱うタスクを，FCOG タスクと定義する．FCOG タスクでは，Fetch–and–Carry タスクについての自然言語指示文が与えられたうえで，ロボットが指示を実行する．本タスクでは，与えられた指示文を基に対象物体及び目標領域を適切に特定し，対象物体を目標領域へと移動させることが望ましい．図 1 に，本タスクにおける典型的な例を示す．この例において，指示文は “Go to the bedroom, grasp the rabbit doll and send it to the corner sofa.” である．この時，まずロボットは寝室へ移動する．続いて，ロボットはウサギの置物を対象物体として特定し把持する．最終的に，ロボットは部屋の隅にあるソファを目標領域として特定し置物をソファの上に置く．

本タスクにおける入力は，自然言語指示文及びロボットカメ

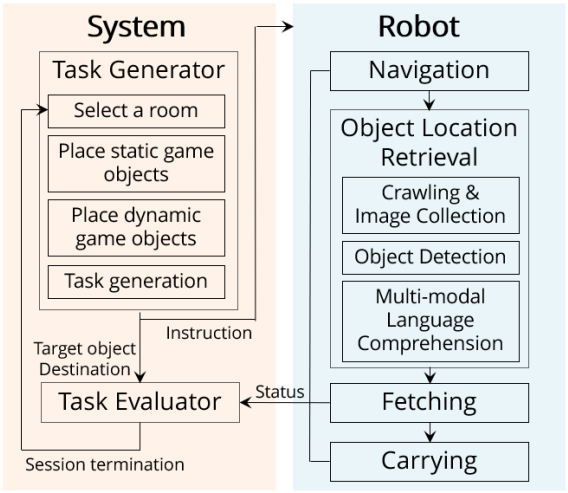

図 2: 提案フレームワークの概要図.

ラを用いて撮影された画像群である．また，これらの入力が与えられた時，ロボットは指示文に基づき，対象物体を目標領域へ移動することが望ましい．また，本論文で使用する用語を以下のように定義する．

- **対象物体:** 指示文が対象としている物体.
- **目標領域:** 対象物体を配置する家具.
- **タスク生成:** 対象物体，目標領域，及び自然言語指示文のセットを作成すること.

第 5 章で述べるように，物体把持及び物体配置に関する軌道生成，経路計画，及びナビゲーションはルールベースにより行われるものとする．

本研究では，シミュレーションの使用を前提とする．実環境においては，環境の初期化に際する対象物体の配置等は人手で行うことが求められる．そのため，on–the–fly 方式での実行には，時間及び労力がかかる．一方で，シミュレーションを用いる場合，全自動化が可能である．それゆえに，効率的な on–the–fly 方式でのタスク実行が可能である．

3. 提案手法

図 2 に提案フレームワークの概要図を示す．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，及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．

3.1 タスク生成システム

タスク生成システムは，主に 3 つのステップを通じてタスクを生成する．まず，静的ゲームオブジェクトを配置することで，シミュレータ環境を作成する．続いて，動的ゲームオブジェクトを環境中にランダムに配置すると共に，初期位置にロボットを配置する．最後のステップで，タスクを自動生成する．タスク生成ステップにおいて，まず対象物体及び目標領域を環境の中からランダムに選択する．続いて，Unity の機能を使用することで，それらの座標を取得すると共に座標を基にオブジェクトの画像を撮影する．

タスク生成システムはクロスモーダル指示文生成モデルを用いて指示文を生成する．この際，CRT [Kambara 21] を拡張したモデルを使用した．モデルは対象領域，目標領域，及びコンテクスト領域を基に物体操作指示文を生成する．ここで，コンテクスト領域は，Faster R-CNN [Ren 16] によって RGB 画像から抽出した領域群を使用する．

3.2 タスク実行システム

タスク実行システムでは，ロボットが FCOG タスクを実行する．本論文では，FCOG タスクを 4 つのサブタスクに分割するアプローチを行う．ここで，それぞれ Navigation, OLR，Fetching，及び Carrying である．Navigation タスクでは，DSR がルールベース手法を用いて初期位置から指定された部屋へのナビゲーションを行う.

OLR タスクにおいて，ロボットはまず各部屋のマップに対して事前に与えられた M 個の waypoint を巡回しつつ，各 waypoint においてロボットカメラを用いて画像を 1 枚ずつ撮影する.. これによって，画像群 $\boldsymbol{X}_{\mathrm{img}} = \{x_{\mathrm{img}}^{(m)} | m = 1, ..., M\}$ を獲得する．続いて，指示文及び各画像を CLIP [Radford 21] に入力し，獲得した類似度スコアの上位 N 枚を選択する．以降のステップでは，選択した N 枚について処理を行う．その後，Faster R-CNN を用いて各画像から対象領域または目標領域の候補領域 $x_{n,k} (n = 1, .., N, k = 1, ...K)$ を検出する．ここで，K は各画像において検出する領域の最大数を示す.

領域の特定に用いるマルチモーダル言語理解モデルは，大きく分けて PMFE 及び Multimodal Decoder の 2 つのモジュールから構成される．モデルへの入力は，以下のように定義する.

$$\boldsymbol{x}'_{n,k} = \{x_{n,k}, \boldsymbol{X}_n, x_{\mathrm{txt}}\}$$
$$\boldsymbol{X}_n = \{x_{n,l} | l = 1, ..., K\}$$

ここで，x_{txt} は指示文を表す．また，可読性のため，特に断りのない限り n 及び k は省略する．対象領域及び目標領域を予測する際，それぞれ別のデータセットで訓練した同様の構造のモデルを用いた．以下では，まず，対象領域の予測手順を示す。

PMFE では，複数モダリティの特徴量を並列的に抽出し，また抽出した特徴量について，attention 機構による処理を行う．本モジュールへの入力は $\boldsymbol{x}'$ である．また，出力は中間特徴量 $\boldsymbol{h}_{\mathrm{cand}}$, $\boldsymbol{h}_{\mathrm{det}}$，及び $\boldsymbol{h}_{\mathrm{mul}}$ である．本モジュールでは以下の手順で処理が行われる．まず，x_{txt} について，CLIP 及び BERT [Devlin 19] のテキストエンコーダを用いることで，それぞれ埋め込み特徴量 $\boldsymbol{h}_c$ 及び $\boldsymbol{h}_b$ を獲得する.

さらに，$\boldsymbol{X}_n$ については，ResNet50 の fc6 層より埋め込み特徴量 $\boldsymbol{H}_n = \{\boldsymbol{h}_{n,l}^{(\mathrm{res})} | l = 1, ..., K\}$ を獲得する．また，x については，CLIP の画像エンコーダ及び ResNet50 を用いてそれぞれ埋め込み特徴量 $\boldsymbol{h}_{\mathrm{clip}}$ 及び $\boldsymbol{h}_{\mathrm{res}}$ を獲得する．$\boldsymbol{h}_{\mathrm{mul}} = [\boldsymbol{h}_c; \boldsymbol{h}_b; \boldsymbol{h}_{\mathrm{clip}}]$ について，Transformer 層 [Vaswani 17] を用いて処理を行うことで，マルチモーダル特徴量 $\boldsymbol{h}'_{\mathrm{mul}}$ を獲得する．まず，以下の式に従って自己注意 $\boldsymbol{h}_{\mathrm{att}}$ を計算する.

$$\boldsymbol{h}_{\mathrm{att}} = \mathrm{softmax}(\frac{(\boldsymbol{W}_q \boldsymbol{h}_{\mathrm{mul}})(\boldsymbol{W}_k \boldsymbol{h}_{\mathrm{mul}}^{\mathrm{T}})}{\sqrt{d_k}})(\boldsymbol{W}_v \boldsymbol{h}_{\mathrm{mul}})$$

ここで，$\boldsymbol{W}_q$,$\boldsymbol{W}_k$ 及び $\boldsymbol{W}_v$ は訓練可能な重みであり，d_k は $\boldsymbol{h}_{\mathrm{mul}}$ のサイズを示す．得られた $\boldsymbol{h}_{\mathrm{att}}$ について，$\boldsymbol{h}'_{\mathrm{mul}} = \mathrm{FFN}(\mathrm{LN}(\boldsymbol{h}_{\mathrm{mul}} + \boldsymbol{h}_{\mathrm{att}}))$ のように処理を行う．ここで，LN(•) 及び FFN(•) はそれぞれ Layer Normlization 及び FeedForward Network 層を示す．最終的に，$\boldsymbol{h}'_{\mathrm{mul}}$, $\boldsymbol{H}_n$, 及び $\boldsymbol{h}_{res}$ をモジュールの出力とする.

Multimodal decoder では，各領域の予測確率を計算する．モジュールへの入力は $\boldsymbol{h}'_{\mathrm{mul}}$, $\boldsymbol{H}_n$, 及び $\boldsymbol{h}_{res}$ であり，出力は予測確率 $p(\hat{y})$ である．まず，中間特徴量 $\boldsymbol{h}_{out}$ を獲得するため，入力特徴量を結合後 L 層の Transformer 層 [Ishikawa 21] を用いて，$\boldsymbol{h}_{\mathrm{out}}$ を獲得する．続いて，$p(\hat{y})$ を $\boldsymbol{h}_{out}$ に対して全結合層及びソフトマックス関数を用いることで獲得する．マルチモーダル言語理解モデルでは，上記の処理を全ての $x_{n,k} (n = 1, .., N, k = 1, ...K)$ に対し行うことで，各領域に対する予測確率 $p(\hat{y}_{n,k})$ を得る．最終的に，$x^*_{\mathrm{targ}} = \mathrm{argmax}_{x_{n,k}} p(\hat{y}_{n,k} | \boldsymbol{x}'_{n,k})$ を最適な候補とする．同様の手順を目標領域の予測に際しても行うことで，領域 x^*_{dest} を獲得する．損失関数としてクロスエ

表 1: 物体操作指示文理解タスクについての，ベースライン手法との比較実験及び Ablation study における定量的結果．

	精度 [%]	
手法	対象物体	目標領域
TdU [Ishikawa 21]	81.64 ± 4.36	79.51 ± 1.00
提案手法 (a)	81.97 ± 1.64	80.98 ± 1.35
提案手法 (b)	84.26 ± 2.92	80.99 ± 2.62
提案手法 (Full)	**87.05** ± 2.02	**82.46** ± 1.60

ントロピー損失関数を用いた．

OLR において対象領域及び目標領域を特定した後，ロボットは Fetching 及び Carrying タスクをルールベース手法を用いて実行する．Fetching 及び Carrying タスクの成功は，それぞれ適切な物体を把持できた場合及びと適切な物体を適切な家具へ配置できた場合定義する．

3.3 タスク評価システム

以下の 3 つの状態のいずれかに達したときに，タスク評価システムはセッションが終了したと判定する．(1)FCOG タスク実行中に制限時間切れとなる． (2) ロボットが FCOG タスクの実行に成功する．(3) ロボットが各サブタスクのいずれかの実行に失敗する．セッション終了の判定後，タスク生成システムは環境を初期化し，次のセッションを開始する．

4. 実験設定

本研究では，FCOG タスクのために WRS–FC データセットを新たに構築した．WRS–FC データセットは、画像と自然言語による指示からなるシミュレーションベースのデータセットである．画像は World Robot Summit 2018 Partner Robot Challenge/Virtual Space Competition (WRS–VS，[Okada 19]) で使用した標準シミュレータにおいて収集された．各画像には，部屋の中にある日常的な物体が 5 つ程度含まれている．データセット作成時，アノテータへは，1 つの画像ペアに対して，DSR が対象物体を目標領域へ移動させるための指示文を自由形式で与えるよう指示した．WRS–FC データセットは，対象物体及び目標領域の画像ペア 1210 個及び，命令文 1210 文から構成される．WRS–FC データセットには対象領域及び目標領域の ground truth が含まれていた．一方で，負例は含まれていなかった．ゆえに，以下の手順で正例及び負例を作成した．まず，Faster R-CNN を用いて，各画像から複数の領域を検出した．正例には，これらの領域の内，ground truth との IoU が 0.7 以上の領域を用いた．また，負例については，(1)IoU < 0.3 の領域を使用する, (2) 指示文を他のサンプルの指示文とランダムに入れ替える，(3)(1) 及び (2) を共に行う，の 3 通りを用いて作成した．最終的に，負例集合からランダムに正例の数と同数のサンプルを抽出した．実験時，訓練集合，検証集合及びテスト集合がそれぞれ 964 サンプル，122 サンプル及び 124 サンプルとなるようにデータセットを分割した．

提案手法におけるマルチモーダル言語処理モジュールにおいて，$\#L$ 及び d_k はそれぞれ 2 及び 768 であった．最適化関数には Adam($\beta_1 = 0.9, \beta_2 = 0.999$) を用いた．訓練のステップ数は 20000 とし，バッチサイズは 32 とした．マルチモーダル言語理解モデルにおいて，訓練にメモリ 11GB 搭載の GeForce RTX 2080 Ti を使用し，約 2 時間を要した．また，200 ステップごとに検証集合を用いた評価を行い，最も精度の高かった際のモデルでテスト集合における評価を行った．FCOG タスクについては，WRS–VS で用いられた標準シミュレータを拡張したシミュレータを使用した．また，1 回の実験あたり，40 セッションを行った．

図 3: TP 及び TN のサンプル. 左図及び右図は対象物体及び目標領域についてのサンプルを示し，図中赤色の矩形は予測の対象を示す．

5. 実験結果

5.1 物体操作指示文理解タスク

対象物体及び目標領域に関する物体操作指示文理解タスクにおける定量的結果を表 1 に示す．実験には WRS–FC データセットを用いた．実験は 5 回行い，表には平均及び標準偏差を示す．評価尺度には精度を使用した．また，ベースライン手法は TdU [Ishikawa 21] とした．

表より，対象物体の予測において，提案手法は 87.05%を記録し，ベースライン手法は 81.64%を記録している．また，目標領域の予測において，提案手法は 82.46%を記録し，ベースライン手法は 79.51%を記録している．これより，提案手法がベースライン手法をそれぞれ 5.41%と 2.95%上回っていることがわかる．これらの性能差は，統計有意であった ($p < 0.05$)．

Ablation 条件には，(a)PMFE を削除，(b)PMFE における transformer 層を MLP 層に置き換え，の 2 条件を定めた．表 1 に示すように，対象物体の予測において，条件 (a) 及び (b) における精度はそれぞれ 81.97%及び 84.26%であった．また，目標領域の予測において，条件 (a) 及び (b) における精度はそれぞれ 80.98%及び 80.99%であった．条件 (a) 及び Full の結果より，対象物体及び目標領域についての予測のどちらに対しても PMFE が有効であることが示された．また，条件 (b) 及び Full の結果より，対象物体及び目標領域についての予測のどちらに対しても，PMFE における transformer 層の有効性が示された．表 1 の結果より，特に PMFE への transformer 層の導入が有効であった．

定性的結果を図 3 に示す．図 3 において，左図から順に TP 及び TN のサンプルを示す．左側の図について，与えられた指示文は “take the blue object from the sofa and put it on the low table” であった．また，矩形領域で示された物体は，唯一のソファの上にある水色の物体であった．このサンプルについて，ベースライン手法は予測を誤った一方で，提案手法は “blue object from the sofa” の指す物体が当該物体であると適切に予測した．また，右側の図について，指示文は “bring the rubik’s cube on the sofa to the table with a toy yellow duck” であり，目標領域は黄色いアヒルのおもちゃが載った机であった．矩形領域の囲む家具にはバナナしか載っておらず，当該領域は目標領域と明らかに異なる領域を示している．このサンプルにおいて，ベースライン手法は当該領域が目標領域であると誤って予測した一方で，提案手法は適切に予測できた．

5.2 FCOG タスク

表 2 に，FCOG タスクにおける提案手法及びベースライン手法の各サブタスクにおける定量的結果，及び ablation study の結果を示す．表は，各タスクの試行回数及び成功回数を示す．ベースライン手法として，WRS–VS における同様のタスクでの優勝手法 [Mizuchi 20] を用いた．この手法では，指示文理解をルールベース手法により行う．アームの軌道生成，経路計画，及びナビゲーションの方法は提案手法と同じとした．また，OLR タスクにおいて収集画像数の削減が精度の向上に寄

表 2: FCOG タスクについての，ベースライン手法との比較実験及び Ablation study における定量的結果.

手法	Navigation 成功率 [%]	OLR 精度 [%]	Fetching 成功率 [%]	Carrying 成功率 [%]
ルールベース (WRS–VS 優勝手法)	**100**(40/40)	0(0/0)	0(0/0)	0(0/0)
提案手法 (i)	**100**(40/40)	0(0/0)	0(0/0)	0(0/0)
提案手法 (Full)	**100**(40/40)	**20**(8/40)	**100**(8/8)	**12.5**(1/8)

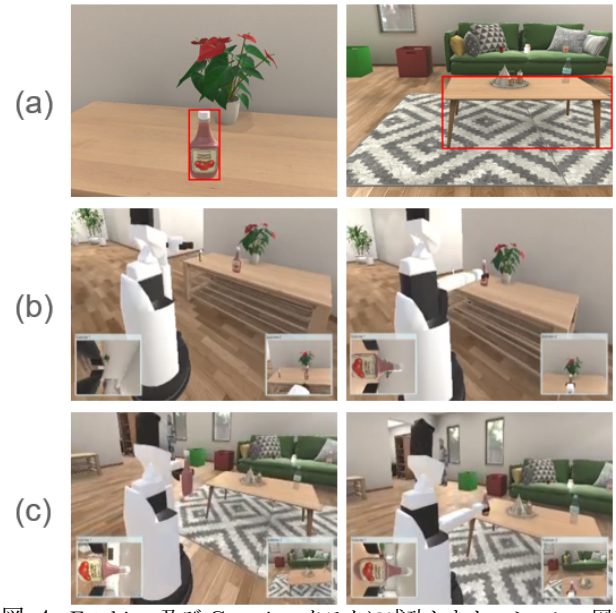

図 4: Fetching 及び Carrying タスクに成功したセッション．図における (a) は，タスク生成システムが取得した，対象物体及び目標領域についての画像を示す.

与していることを検証するため，ablation 条件として CLIP を用いた削減を行わない条件 (条件 (i)) を設定した．評価尺度として，OLR タスクにおいては精度を，Navigation，Fetching，及び Carrying タスクにおいては成功率をそれぞれ利用した．

表より，いずれの手法も，Navigation タスクの成功率は 100%であった．また，OLR タスクにおいて，ベースライン手法の精度は 0%であった．ベースライン手法は，自由形式な指示文に含まれる多様な参照表現を理解できなかったためと考えられる．一方で，提案手法の精度は 20%であった．また，ベースライン手法について，Fetching 及び Carrying タスクは実行されなかった．これは，OLR タスクに成功した場合のみそれらのタスクを実行したためである．一方で，提案手法は Fetching 及び Carrying タスクの成功率において，それぞれ 20%及び 12.5%であった．また，表 2 より，条件 (i) では OLR タスクの精度が 0%であった．このことから，OLR タスクにおいて，CLIP を用いた画像数の削減は精度向上に寄与していることがわかった．

図 4 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，対象物体及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及び目標領域として，赤いボトル及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，対象物体及び目標領域として，赤いボトル及び机の前のテーブルを適切に特定できた．その後，図 4(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 4(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．

6. おわりに

本論文では FCOG タスクを扱った．本研究の主要な貢献は以下である．

- FCOG タスクにおいて，生成，実行，及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した．
- FCOG タスクに対して，Navigation, 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した．
- 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 PMFE を導入した．
- 提案手法は，FCOG タスクにおけるタスク成功率で既存手法を上回った．

謝辞

本研究の一部は，JSPS 科研費 20H04269，JST ムーンショット，NEDO の助成を受けて実施されたものである．

参考文献

[Ahn 22] Ahn, M., et al.: Do As I Can, Not As I Say: Grounding Language in Robotic Affordances, *arXiv preprint arXiv:2204.01691* (2022)

[Devlin 19] Devlin, J., Chang, M.-W., et al.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in *NAACL-HLT*, pp. 4171–4186 (2019)

[Ishikawa 21] Ishikawa, S. and Sugiura, K.: Target-dependent UNITER: A Transformer-Based Multimodal Language Comprehension Model for Domestic Service Robots, *IEEE RA-L*, Vol. 6, No. 4, pp. 8401–8408 (2021)

[Kambara 21] Kambara, M., et al.: Case Relation Transformer: A Crossmodal Language Generation Model for Fetching Instructions, *IEEE RA-L*, Vol. 6, No. 4, pp. 8371–8378 (2021)

[Mizuchi 20] Mizuchi, Y. and Inamura, T.: Optimization of Criterion for Objective Evaluation of HRI Performance that Approximates Subjective Evaluation: A Case Study in Robot Competition, *AR*, Vol. 34, No. 3-4, pp. 142–156 (2020)

[Okada 19] Okada, H., Inamura, T., and Wada, K.: What competitions were conducted in the service categories of the World Robot Summit?, *AR*, Vol. 33, No. 17, pp. 900–910 (2019)

[Radford 21] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., et al.: Learning Transferable Visual Models from Natural Language Supervision, in *ICML*, pp. 8748–8763 (2021)

[Ren 16] Ren, S., He, K., et al.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, *IEEE Trans. PAMI*, Vol. 39, No. 6, pp. 1137–1149 (2016)

[Shridhar 20] Shridhar, M., et al.: ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks, in *CVPR*, pp. 10740–10749 (2020)

[Vaswani 17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., et al.: Attention Is All You Need, *NeurIPS*, Vol. 30, (2017)