第4回 とめ研究所若手研究者懸賞論文

Switching Head-Tail 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と物体操作

是方 諒介[1]

慶應義塾大学大学院 理工学研究科 開放環境科学専攻

[1]本論文は，是方諒介を筆頭著者として慶應義塾大学の神原元就，吉田悠，石川慎太朗，川崎陽祐，高橋正樹，杉浦孔明と共同で 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems [Korekata *et al.* 23] および 2023 年度人工知能学会全国大会 [是方 他 23] にて発表した内容を元に再構成したものである．

目 次

図 目 次

表 目 次

第1章

序論

高齢化が進行する現代社会において，日常生活における介助支援の需要は高まっている．これに伴い，在宅介助者不足が社会問題となっており，一つの解決策として被介助者を物理的に支援することが可能な生活支援ロボットに注目が集まっている．しかし，人間からの自然言語による指示をロボットが理解する能力についてはいまだ不十分である．

本研究では，物体の把持および配置に関する物体操作指示文を生活支援ロボットが理解し実行するための手法の構築を目的とする．具体的には，“Move the bottle on the left side of the plate to the empty chair.” という指示文が与えられるとする．このとき，ロボットが周囲の物体および家具の中からボトルを対象物体として，椅子を配置目標として認識したうえで，ボトルを把持して椅子へ配置することが望ましい．

人間の発する指示はしばしば曖昧であり，対象となる物体やその配置目標をロボットが特定することは困難である．実際に，物体操作を含む Vision-and-Language Navigation (VLN) における標準ベンチマークである ALFRED [Shridhar *et al.* 20] では，人間の精度は 91.0%と報告されている一方，最先端の手法 (e.g. Prompter [Inoue *et al.* 22], LGS-RPA [Murray+, RA-L22], FILM [Min *et al.* 22], AMSLAM [Jia *et al.* 22], HLSM-MAT [Ishikawa *et al.* 22]) では 46%以下しか達成できていない．

Target-dependent UNITER (TdU) [Ishikawa *et al.* 21] は，物体操作指示文が把持対象とする物体を特定する Multimodal Language Understanding for Fetching Instruction (MLU-FI) タスクにおいて良好な結果が報告されている手法である．しかし，この手法に配置目標候補の入力を追加することで本研究で扱うタスクに拡張した場合，対象物体候補および配置目標候補がともに対象物体および配置目標に一致するかを同時に推論することになる．したがって，環境中に多数存在する対象物体候補と配置目標候補に関するすべての組合せについて推論を行うため多くの推論回数を要する．例えば，対象物体候補および配置目標候補がそれぞれ 100 個存在する場合を想定すると，最尤の組を探索するために

合計 10000 回もの推論が必要となる．1 回の推論時間を 0.004 秒と仮定すると，ロボットの判断に要する時間が 40 秒と見込まれるため，リアルタイム性で実用面に問題がある．

本論文では，対象物体および配置目標に関する予測を単一モデルで個別に行う方法でタスクを解くことが可能な Switching Head-Tail Funnel UNITER (SHeFU) を提案する．これにより，対象物体候補および配置目標候補がそれぞれ M および N 個存在する状況において，言語理解に必要な推論回数を $O(M \times N)$ から $O(M + N)$ に削減することが可能となる．既存手法と異なる点は，Switching head-tail 機構を導入することで，単一モデルで対象物体候補および配置目標候補のどちらも入力として扱い，効率的に推論することが可能な点である．提案手法において導入した Switching head 機構では，対象物体および配置目標のどちらに関して予測するかを切り替える．また，Switching tail 機構におけるマルチタスク学習の導入により，対象物体および配置目標それぞれに関する予測という異なるタスク間の相乗効果による精度向上が見込まれる．さらに，単一モデルでの学習が可能になるため，総パラメータ数を削減することができる．

本研究の新規性を以下に示す．

- 言語理解における対象物体および配置目標の探索に必要な推論回数を削減することが可能な SHeFU を提案する．
- Switching head-tail 機構を導入することで，対象物体および配置目標について，単一モデルで個別に予測可能にする．

第2章

関連研究

2.1 マルチモーダル言語処理

マルチモーダル言語処理分野のサーベイ論文として，[Mogadala *et al.* 21]，[Qiao *et al.* 20] などが挙げられる．[Mogadala *et al.* 21] は，画像および言語を統合した 10 個の代表的なタスクにおいて，問題設定，手法，既存データセット，評価指標に関して議論している．また，[Qiao *et al.* 20] は画像および言語それぞれのモダリティに対応するメカニズムによって手法を分類する他，Referring Expression Comprehension (REC) におけるデータセットについて比較している．

マルチモーダル言語処理分野は，扱うモダリティの組合せによって多様な分野が存在する．画像および言語を扱う分野としては，REC，Referring Expression Segmentation (RES)，Visual Question Answering，Embodied Question Answering，Object Goal Navigation，VLN，および MLU-FI などが挙げられる．

本手法は，MLU-FI タスクを扱う手法と関連が深い．MTCM-AB [Magassouba *et al.* 20] は，MTCM [Magassouba *et al.* 19] を ABN [Fukui *et al.* 19] によって拡張した MLU-FI モデルである．attention branch によってマルチモーダルな注意機構を実現し，画像中の物体の attention map を生成する．TdU [Ishikawa *et al.* 21] は，UNITER [Chen *et al.* 20] を対象物体候補の画像および位置情報を扱うように拡張した MLU-FI モデルである．SHeFU はこれらの手法とは異なり，Switching head-tail 機構を用いることで対象物体および配置目標について単一モデルで個別に予測可能にする．

2.2 データセット

本研究で扱うタスクと関連の深い MLU-FI における標準データセットとして，PFN-PIC [Hatori *et al.* 18] データセット，WRS-PV [Ogura *et al.* 20] データセット，および

WRS-UniALT [Ishikawa *et al.* 21] データセットが挙げられる．いずれも画像および画像中の物体に関する指示文から構成される．PFN-PIC データセットは，約 20 種類の日用品を 4 つの箱に無作為に配置した物体に対する固定視点の実画像を用いる．一方で WRS-PV データセットおよび WRS-UniALT データセットは，部屋の中に配置された日用品について World Robot Summit Partner Robot Challenge [Okada *et al.* 19] の標準シミュレーション環境において収集された画像を用いる．

また，VLN はロボットが自身の視覚を基に自然言語による指示を解釈するタスクであり，本研究で扱うタスクを包含する．物体操作を含む VLN における標準データセットとしては，ALFRED [Shridhar *et al.* 20] データセットが挙げられる．ALFRED データセットは，自然言語による指示文とロボットのカメラ画像から，家事タスクにおけるロボットの行動を学習させるためのデータセットである．

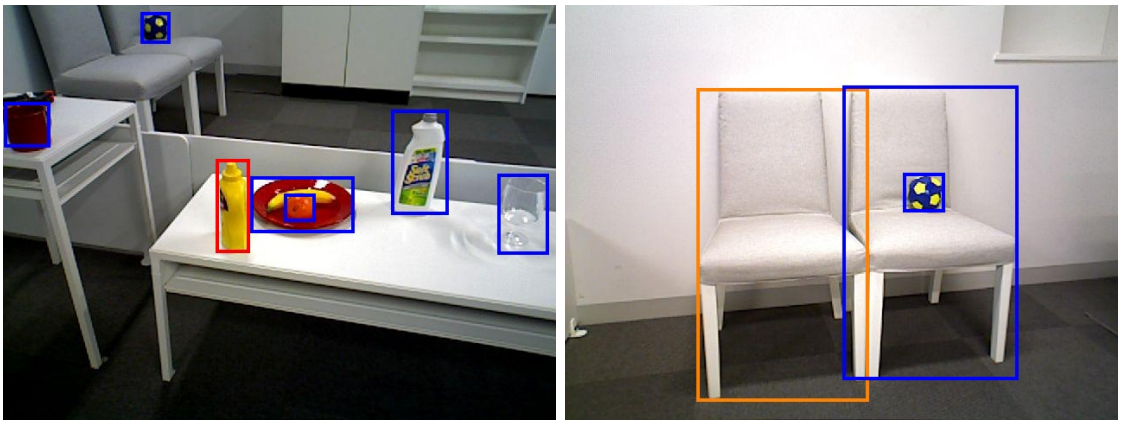
第3章

問題設定

本研究で扱うタスクは，Dual Referring Expression Comprehension with fetch-and-carry (DREC-fc) タスクである．DREC-fc タスクは，日用品および家具が写る複数の画像から，参照表現を含む指示文の対象物体および配置目標の両方を特定し，ロボットが対象物体を配置目標まで運搬するタスクである．すなわち，本タスクは言語理解および動作実行という二つのサブタスクに分けられる．

本タスクでは，対象物体候補および配置目標候補がそれぞれ指示文の対象物体および配置目標に一致するか予測し，特定された対象物体を把持して配置目標へ配置することが望ましい．図 3.1 に，DREC-fc タスクの例を示す．左側が対象物体候補，右側が配置目標候補を含む全体画像である．これらの画像に対して，“Move the bottle on the left side of the plate to the empty chair.” という指示文が与えられるとする．このとき，青色の矩形領域で示した周囲の物体および家具の中から，赤色の矩形領域が対象物体，橙色の矩形領域が配置目標であると特定して把持および配置動作を行うことが求められる．

入出力を以下のように定義する．



対象物体　　配置目標

図 3.1: DREC-fc タスクのシーン例

- 入力：指示文，対象物体候補が写る画像，配置目標候補が写る画像
- 出力：対象物体候補および配置目標候補が，対象物体および配置目標にともに一致する確率の予測値 $p(\hat{\boldsymbol{y}})$

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
- 対象物体候補：対象物体であるか否かを判定する物体
- 配置目標：指示文が目標としている家具
- 配置目標候補：配置目標であるか否かを判定する家具

ロボットの移動，把持，および配置に関する軌道生成はヒューリスティックに行われるものとする．詳細は 5.2.4 節で後述する．DREC-fc タスクの評価指標には，言語理解における分類精度およびタスク成功率を使用する．

transformer [Vaswani *et al.* 17] などの大規模モデルの訓練には，大量のデータが必要であることが多い．しかし，実機ロボットを用いたデータ収集は人間が物体の配置を行う必要があるため多くの時間を要する．そこで，大量の訓練データを短時間で収集可能なシミュレーション環境を利用し，実環境へゼロショット転移することでコストの削減を図る．

第4章

提案手法

本手法は，自然言語によって指示を与えられる fetch-and-carry タスク [Iocchi *et al.* 15, Okada *et al.* 19] と関連が深い．なお，本研究ではテンプレートに基づく指示文は用いない．モデル全体は 2 つの主要モジュールから構成され，それぞれ Switching Image Embedder および Funnel Transformer である．図 4.1 にモデルの構造を示す．図において，Target，Destination，Detected Objects，および Instruction はそれぞれ対象物体候補領域，配置目標候補領域，画像中の各物体または家具の領域，および指示文を表す．また，$\oplus$ および角の丸い矢印の合流はそれぞれ加算および結合を示す．

提案手法の新規性は以下である．

- SHeFU を用いることで，対象物体候補および配置目標候補がそれぞれ M および N 個存在する状況において，言語理解に必要な推論回数を $O(M \times N)$ から $O(M + N)$ に削減することを可能とした．
- Switching head-tail 機構を導入することで，対象物体および配置目標について，単一モデルで個別に予測することを可能にした．

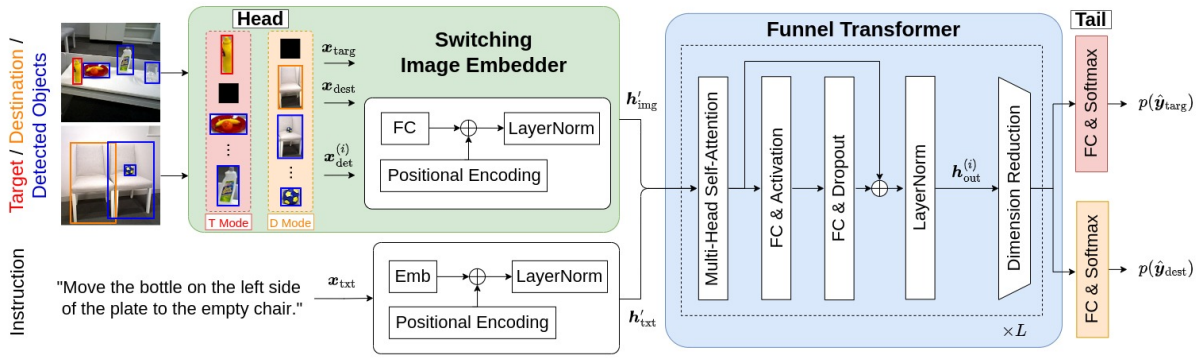


図 4.1: 提案手法のモデル構造

本研究では，2つの参照表現を含む物体操作に関するマルチモーダル言語理解タスクにおいて提案手法を検証する．一方で，本手法における Switching head-tail 機構は，3つ以上の参照表現を含む指示文理解タスクや RES タスクに対しても広く適用可能であると考えられる．

提案手法は，同様の入力であればシミュレーション環境および実機環境のいずれに対しても適用可能である．実際に，シミュレーション環境において収集されたデータセットを用いた評価結果および実機における検証結果についてそれぞれ 6.1 節および 6.2 節で報告する．

4.1 入力

モデルへの入力 $\boldsymbol{x}$ を以下のように定義する．

$$\boldsymbol{x} = \{\boldsymbol{x}_{\text{targ}}, \boldsymbol{x}_{\text{dest}}, \boldsymbol{x}_{\text{txt}}\}$$

ここに，$\boldsymbol{x}_{\text{targ}} \in \mathbb{R}^{1024}$，$\boldsymbol{x}_{\text{dest}} \in \mathbb{R}^{1024}$，および $D_{\text{v}} \times D_{\text{l}}$ 次元の one-hot ベクトル集合 $\boldsymbol{x}_{\text{txt}}$ はそれぞれ対象物体候補の領域，配置目標候補の領域，および指示文を表す．また，D_{v} および D_{l} はそれぞれ語彙サイズおよび指示文中のトークン数の最大値を表す．

$\boldsymbol{x}_{\text{targ}}$ および $\boldsymbol{x}_{\text{dest}}$ については，Faster R-CNN [Ren *et al.* 16] のバックボーンネットワークである ResNet50 [He *et al.* 16] の fc6 層の出力を画像領域の特徴量として抽出した．矩形領域の画像特徴量に対する positional encoding には，7次元ベクトル $[\frac{x_1}{W}, \frac{y_1}{H}, \frac{x_2}{W}, \frac{y_2}{H}, \frac{w}{W}, \frac{h}{H}, \frac{w \cdot h}{W \cdot H}]^T$ を用いた．ここで，$(W,\, H)$，$(w,\, h)$，$(x_1,\, y_1)$，および $(x_2,\, y_2)$ はそれぞれ入力画像の幅および高さ，矩形領域の幅および高さ，矩形領域の左上の頂点座標，および矩形領域の右下の頂点座標を示す．

また，$\boldsymbol{x}_{\text{txt}}$ については指示文に対して BERT [Kenton *et al.* 19] と同様に WordPiece [Wu *et al.* 16] によるトークン化を行った．指示文の言語特徴量に対する positional encoding には，指示文中のトークンの位置を用いた．指示文に関する埋め込み特徴量 $\boldsymbol{h}'_{\text{txt}}$ を得るため，$\boldsymbol{x}_{\text{txt}}$ に訓練可能な重みを掛け合わせて正規化を行う．

4.2 Switching Image Embedder

Switching Image Embedder では，Switching head 機構を用いてモードに応じて入力を切り替えながら対象物体候補，配置目標候補，および画像中の各物体または家具の領域

に対する埋め込み処理を行う．ここで，対象物体について予測を行うことを target mode，配置目標について予測を行うことを destination mode と定義する．入力は，$\boldsymbol{x}_{\text{targ}}$ および $\boldsymbol{x}_{\text{dest}}$ から構成される．まず，$\boldsymbol{x}_{\text{targ}}$ および $\boldsymbol{x}_{\text{dest}}$ について，以下に示す式により切り替え処理を行う．

$$(\boldsymbol{x}_{\text{targ}}, \boldsymbol{x}_{\text{dest}}) = \begin{cases} (\boldsymbol{x}_{\text{targ}}, \boldsymbol{0}) & \text{if target mode} \\ (\boldsymbol{0}, \boldsymbol{x}_{\text{dest}}) & \text{if destination mode} \end{cases}$$

すなわち，各モードにおいて不要な入力を 0 埋めする．これにより，0 埋めが予測対象を切り替える条件付けとして機能すると期待される．また，target mode においては対象物体候補が写る画像，destination mode においては配置目標候補が写る画像に対して物体検出を行い，画像中の周辺物体または家具の領域 $\{\boldsymbol{x}_{\text{det}}^{(i)} \in \mathbb{R}^{1024} \mid i = 1, \ldots, N\}$ を得る．ここで，N は Faster R-CNN により検出された画像中の領域の数を示す．なお，画像特徴量抽出や positional encoding については $\boldsymbol{x}_{\text{targ}}$ および $\boldsymbol{x}_{\text{dest}}$ と同様に行う．

次に，$\boldsymbol{x}_{\text{targ}}$，$\boldsymbol{x}_{\text{dest}}$，および $\boldsymbol{x}_{\text{det}}^{(i)}$ について，おのおの異なる全結合層に入力して得られた出力を正規化することによりそれぞれ $\boldsymbol{h}'_{\text{targ}}$，$\boldsymbol{h}'_{\text{dest}}$，および ${\boldsymbol{h}'_{\text{det}}}^{(i)}$ を得る．これらを結合して，出力 $\boldsymbol{h}'_{\text{img}} = \{\boldsymbol{h}'_{\text{targ}}, \boldsymbol{h}'_{\text{dest}}, {\boldsymbol{h}'_{\text{det}}}^{(1)}, \ldots, {\boldsymbol{h}'_{\text{det}}}^{(N)}\}$ を得る．

4.3 Funnel Transformer

本モジュールは，L 層の Funnel Transformer [Dai *et al.* 20] から構成される．1 層目の入力は，$\boldsymbol{h}_{\text{in}}^{(1)} = \{\boldsymbol{h}'_{\text{img}}, \boldsymbol{h}'_{\text{txt}}\}$ である．transformer [Vaswani *et al.* 17] における self-attention 機構に基づき，i 層目の出力 $\boldsymbol{h}_{\text{out}}^{(i)}$ を得る．この際，$i(>1)$ 層目における query，key，および value の次元数を $H^{(i)} = \lfloor H^{(i-1)}/\,2 \rfloor$，attention head 数を $A^{(i)} = \lfloor A^{(i-1)}/\,2 \rfloor$ とし，max pooling を用いて次元数を削減する．ここで，$\lfloor \cdot \rfloor$ は床関数を示す．なお，[Dai *et al.* 20] では query のみに max pooling が行われていたが，本研究では経験的な理由から query，key，および value のすべてに max pooling を適用する．また，i 層目における入力は $\boldsymbol{h}_{\text{in}}^{(i)} = \boldsymbol{h}_{\text{out}}^{(i-1)}$ とする．$i-1$ 層目の場合と同様に $\boldsymbol{h}_{\text{in}}^{(i)}$ を処理していく流れを L 層目まで繰り返すことで，Funnel Transformer モジュールの出力 $\boldsymbol{h}'_{\text{out}}$ を得る．

Switching tail 機構では，モードに応じて最後のネットワークを切り替える処理を行う．$\boldsymbol{h}'_{\text{out}}$ について，以下に示す式により対象物体に関する予測確率 $p(\hat{\boldsymbol{y}}_{\text{targ}})$ を得る．

$$p(\hat{\boldsymbol{y}}_{\text{targ}}) = \text{softmax}(f_{\text{FC}}(\boldsymbol{h}'_{\text{out}}))$$

ここで，f_{FC} は全結合層を表す．また，配置目標に関する予測確率 $p(\hat{\boldsymbol{y}}_{\mathrm{dest}})$ についても異なる全結合層を用いて同様に得る．target mode においては $p(\hat{\boldsymbol{y}}_{\mathrm{targ}})$ を，destination mode においては $p(\hat{\boldsymbol{y}}_{\mathrm{dest}})$ をモデル全体の最終的な出力とみなす．各モードにおける予測ラベル $\hat{y}_{\mathrm{targ}}$ または $\hat{y}_{\mathrm{dest}}$ は，予測確率を閾値 0.5 で二値化することで得られる．ただし，Switching tail 機構により，対象物体および配置目標について個別に推論を行う．したがって，以下に示す式により予測ラベル $\hat{y}$ を得る．

$$\hat{y} = \hat{y}_{\mathrm{targ}} \cap \hat{y}_{\mathrm{dest}} \tag{4.1}$$

4.4 損失関数

損失関数 $\mathcal{L}$ を以下の式に定義する．

$$\mathcal{L} = \lambda_{\mathrm{targ}} \mathcal{L}_{\mathrm{CE}}(\boldsymbol{y}_{\mathrm{targ}}, p(\hat{\boldsymbol{y}}_{\mathrm{targ}})) + \lambda_{\mathrm{dest}} \mathcal{L}_{\mathrm{CE}}(\boldsymbol{y}_{\mathrm{dest}}, p(\hat{\boldsymbol{y}}_{\mathrm{dest}}))$$

ここで，$\mathcal{L}_{\mathrm{CE}}(\cdot, \cdot)$ および $\lambda_{\cdot}$ はそれぞれ交差エントロピー誤差および各モードにおけるタスクの重み係数を示す．なお，$y_{\cdot}$ は対象物体候補または配置目標候補がそれぞれの ground truth に一致するかの真偽値を表す．target mode においては $\lambda_{\mathrm{dest}} = 0$，destination mode においては $\lambda_{\mathrm{targ}} = 0$ とすることで，マルチタスク学習を行う．

第5章

実験設定

5.1 シミュレーションデータセットによる実験

5.1.1 データセット

本研究では，DREC-fc タスクのためのデータセットとして，ALFRED データセット [Shridhar *et al.* 20] を基に ALFRED-fc データセットを収集した．本研究で扱う DREC-fc タスクのための標準データセットは，我々の知る限り存在しない．そこで，物体操作を含む VLN タスクにおける標準データセットである ALFRED データセットを基に新規データセットを作成することが最適であると考えた．ALFRED データセットは，自然言語による指示文とロボットのカメラ画像から，家事タスクにおけるロボットの行動を訓練するためのデータセットである．タスクは複数の逐次的なサブゴールから構成され，各サブゴールごとに指示文が存在する．ALFRED データセットでは，Amazon Mechanical Turk を用いて少なくとも 3 人の異なるアノテータにより，ロボットがサブゴールを達成するための指示文が付与されている．

しかし，ALFRED データセットは AI2-THOR [Kolve *et al.* 17] における模範動作とそれに対応する指示文から構成されるため，対象物体および配置目標がそれぞれ写る状態で保存されたカメラ画像は存在しない．また，ロボットが物体を運搬する際のカメラ画像を取得する場合，把持している物体が空中に浮かんだ状態で画像に写り込む．これによりロボットの視界が遮蔽されてしまうため，本研究における入力画像として不適切であった．そのため本研究では，ALFRED データセットのうち，サブゴールが特定の物体を把持して特定の場所へ配置する “Pick & Place” カテゴリに属するエピソードにおける指示文，把持直前および配置直後のロボットのカメラ画像を抽出した．ただし，配置目標の画像に写り込んだ配置後の対象物体の領域に対して 0 埋めするマスク処理を行った．

ALFRED データセットには各画像における対象物体および配置目標に関する領域の

表 5.1: SHeFU のハイパーパラメータ設定

最適化関数	Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
学習率	8×10^{-5}
ステップ数	20000
バッチサイズ	8
ドロップアウト	0.1
タスクの重み係数	$\lambda_{\text{targ}} = 1.0$, $\lambda_{\text{dest}} = 1.0$

ground truth が含まれるが，その他の対象物体候補および配置目標候補に関する領域はアノテーションされていない．そこで，Faster R-CNN [Ren *et al.* 16] による物体検出を行うことで複数の領域を獲得し，ground truth との Intersection over Union (IoU) が 0.7 以上のものを正例を作成するために用いた．負例については作成方法が三通り存在する．第一に，対象物体候補に関して IoU が 0.3 以下のものを選ぶ方法，第二に，指示文を無作為に選んだ別サンプルのものに差し替える方法，第三に，これら両方を行う方法である．配置目標が写る画像においては ground truth 以外に明確な配置目標候補が存在しない場合があるため，第二の方法はタスクの難易度が下がることを防ぐ意図で実施された．なお，負例の集合から正例と同じ数のサンプルを無作為に抽出することで，最終的な正例と負例のサンプル数を均一にした．

ALFRED-fc データセットは，対象物体および配置目標に関するそれぞれ 1099 枚の画像，3452 文の英語で記述された指示文を含み，語彙サイズは 646，全単語数は 29113，平均文長は 8.4 である．ALFRED-fc データセットでは，4420 サンプルを訓練集合に，642 サンプルを検証集合に，686 サンプルをテスト集合にそれぞれ使用した．なお，指示文と，対象物体候補および配置目標候補に関するそれぞれの画像の三つ組を 1 サンプルと定義する．ALFRED-fc データセットの訓練集合および検証集合は ALFRED データセットにおける訓練集合から，テスト集合は ALFRED データセットにおける valid seen および valid unseen から作成した．訓練集合はモデルの訓練に，検証集合はハイパーパラメータの調整に，テスト集合はモデルの性能評価にそれぞれ使用した．

5.1.2 パラメータ設定

表 5.1 に，ハイパーパラメータ設定を示す． Funnel Transformer モジュールにおいて，層数を $L = 2$，1 層目の query，key，value の次元数を $H^{(1)} = (N + D_{\text{l}} + 1) \times 768$，attention

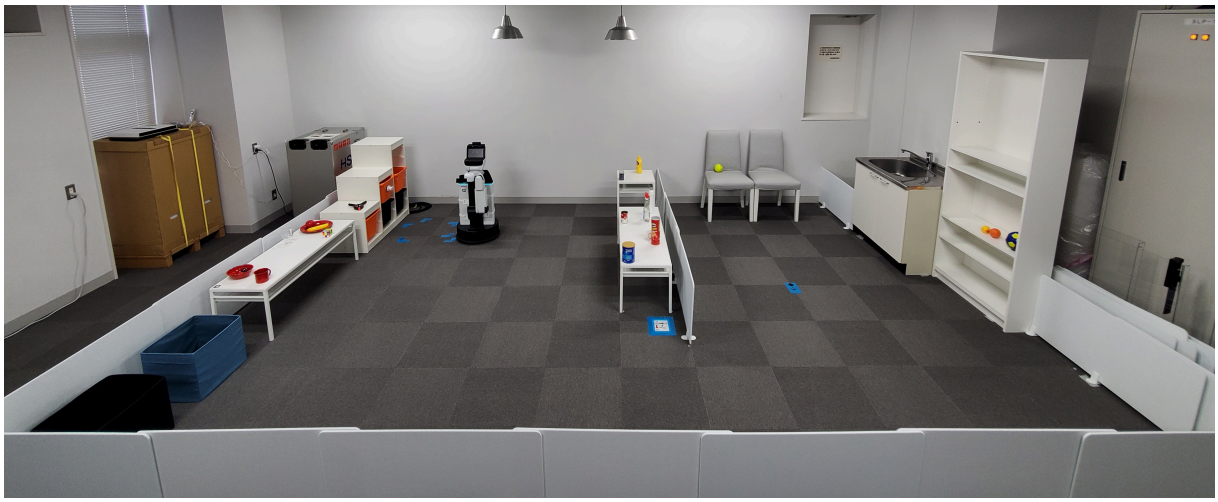


図 5.1: 実機実験環境

head 数を $A^{(1)} = 12$ と設定した．ただし，N および D_{l} はそれぞれ画像中の物体または家具の数および指示文中のトークン数の最大値を示す．

提案手法の訓練可能パラメータ数は約 3277 万である．訓練には，メモリ 24GB 搭載の GeForce RTX 3090 およびメモリ 64GB 搭載の Intel Core i9-10900KF を使用した．訓練には約 20 分，推論には約 4×10^{-3} 秒/sample を要した．合計 20000 ステップの訓練のうち，2000 ステップごとに検証集合における精度を測定した．検証集合においてもっとも高い精度を得たときのテスト集合における精度を，最終的な精度とした．

5.2 実機実験

5.2.1 環境

図 5.1 に，実機実験で使用した環境を示す．本環境は，家庭内の実環境における片付けタスクのベンチマークである国際的なロボット競技会 World Robot Summit 2020 Partner Robot Challenge/Real Space (WRS2020RS) [WRS20] の標準環境に基づいている．環境の広さは 6.0 [m]×4.0 [m] である．家具は WRS2020RS に準拠しており，6 種類存在する．家具の配置は図 5.1 に示された通りであり，収納箱は青色のものと黒色のもの，長机および椅子は同一のものがそれぞれ 2 つずつ存在する．このため，環境中に家具は合計 9 つ存在する．本実験においては，無作為に 1 つの家具を選択して配置目標として用いた．

図 5.2: HSR

5.2.2 ロボット

実機実験においては，図 5.2 に示すトヨタ自動車製の生活支援ロボット Human Support Robot (HSR) [Yamamoto *et al.* 19] を用いた．

5.2.3 物体

図 5.3 に，実機実験で使用した物体を示す．これらの物体は，WRS2020RS で標準物体として指定された YCB オブジェクト [Calli *et al.* 15] の一部である．上部の 20 種類の物体群および下部の 19 種類の物体群は，それぞれ対象物体および背景オブジェクトとして用いた．対象物体については，[Calli *et al.* 15] において “Food”，“Kitchen”，“Shape”，および “Task” カテゴリに属する物体から，HSR のエンドエフェクタで把持可能なものを選択した．また，背景オブジェクトについてはそれら以外から無作為に選択した．

図 5.3: 実機実験で使用した物体

5.2.4 実機動作

以下では，実機実験の環境について説明する．本実験では，12 種類の物体配置を作成した．各物体配置において，物体は無作為な位置に配置された．ただし，本実験ではすべての物体が家具の上に配置されていることを前提とする．各試行において，対象物体を図 5.3 上部の物体群から，配置目標を図 5.1 中の家具からそれぞれ無作為に決定した．そのうえで，ロボットに対して “Pick up the apple and put it down on the right-hand chair.” などの指示文を与えた．指示文は英語で，合計 418 文与えられた．

次に，ロボットの動作について説明する．はじめに，事前に定められた 16 個の waypoint をロボットが初期位置から順に巡回することで，環境中の画像収集を行った．各 waypoint は，ロボットが各家具に正対して複数の視点角度から物体および家具を撮影できるように定めた．ロボットの移動については，事前に与えられた地図を用いて標準的な手法で経路計画を行った．画像の取得には，HSR の頭部に搭載された Asus Xtion Pro カメラを用いた．したがって，1 度の巡回で 16 枚の画像が得られた．これらの画像を提案手法への入力とし，推論には ALFRED-fc データセットで訓練されたモデルを利用することでゼロ

ショット転移を行った．ロボットの把持動作については，深度画像および矩形領域を基に把持点を決定した．具体的には，深度画像の矩形領域内に対してカメラの内部パラメータを掛け合わせることでカメラ座標系に変換した点群を取得し，各座標軸における中央値を把持点とした．なお，ロボットが把持に成功した場合のみ配置動作を行うものとした．ロボットの配置動作については，家具を撮影した waypoint を利用しヒューリスティックに行った．

第6章

実験結果

6.1 シミュレーションデータセットによる実験

6.1.1 定量的結果

表6.1に，各手法のALFRED-fcデータセットにおける言語理解精度を示す．実験は5回行い，精度はその平均値および標準偏差を示す．

TdU [Ishikawa *et al.* 21] に対して，単純にDREC-fcタスクへ拡張した手法をベースラインとする．ベースライン手法では，対象物体候補および配置目標候補の両方が入力に含まれる．TdUはDREC-fcタスクと関連の深いMLU-FIタスクにおいて良好な結果が報告されている手法であるため，これを拡張した手法をベースラインとした．

評価指標として，精度を用いた．ただし，提案手法では対象物体候補および配置目標候補について個別に推論を行う点でベースライン手法と異なるため，正解ラベルyを以下のように定義することで統一的な指標での比較を行った．本実験で用いたデータセットは正例と負例のサンプル数が均等で偏りがないため，このような場合に標準的な精度を評価指標として採用した．

$$y = y_{\mathrm{targ}} \cap y_{\mathrm{dest}} \tag{6.1}$$

表 6.1: ALFRED-fcデータセットおよび実機における言語理解精度

手法	Accuracy [%] ALFRED-fc	実機
(i) ベースライン手法 (extended TdU [Ishikawa *et al.* 21])	79.4 ± 2.76	52.0
(ii) 提案手法 (W/o Switching head 機構)	78.4 ± 2.05	-
(iii) 提案手法 (W/o Switching tail 機構)	76.9 ± 2.91	-
(iv) 提案手法 (SHeFU)	$\mathbf{83.1 \pm 2.00}$	**55.9**

表 6.1 より，ベースライン手法は精度が 79.4%であるのに対し，提案手法は 83.1%であり，提案手法が 3.7 ポイント上回った．この性能差は統計有意であった ($p < 0.01$)．

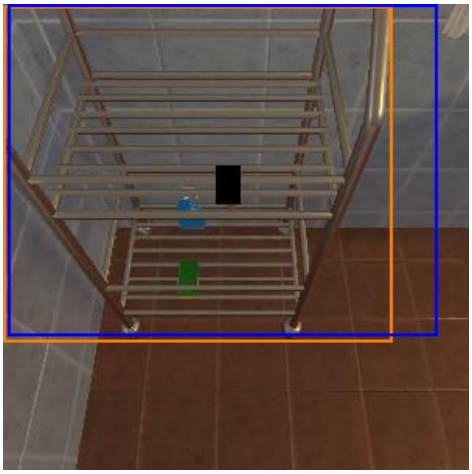
6.1.2 定性的結果

図 6.1 に定性的結果を示す．赤色，橙色，および青色の矩形領域はそれぞれ対象物体の ground truth，配置目標の ground truth，および対象物体候補または配置目標候補を表す．図 6.1 における (a) および (b) は，True Positive の例である．対象物体および配置目標は，それぞれ (a) に写る棚の上に置かれた石鹸および (b) に写る金属製のラックである．また，対象物体候補および配置目標候補はともに ground truth に一致している．したがって，$(y_{\text{targ}}, y_{\text{dest}}) = (1, 1)$，すなわち $y = 1$ の例である．この例において，ベースライン手法は $\hat{y} = 0$ であると誤って予測した．一方，提案手法は $(\hat{y}_{\text{targ}}, \hat{y}_{\text{dest}}) = (1, 1)$ であると正しく予測した．

同様に，図 6.1 における (c) および (d) は True Negative の例である．対象物体および配置目標は，それぞれ塩の容器および台所の引出しである．対象物体候補および配置目標候補はともに ground truth に一致していない．この負例は指示文を無作為に選んだ別サンプルのものに差し替える方法で作成されたものであるため，対象物体および配置目標の ground truth はそれぞれの候補領域と同じ画像中には存在しない．したがって，$(y_{\text{targ}}, y_{\text{dest}}) = (0, 0)$，すなわち $y = 0$ の例である．この例において，ベースライン手法は $\hat{y} = 1$ であると誤って予測した．一方，提案手法は $(\hat{y}_{\text{targ}}, \hat{y}_{\text{dest}}) = (0, 0)$ であると予測した．すなわち，対象物体として机上のペンが塩の容器に合致しないことおよび配置目標として本棚が台所の引出しに合致しないことを正確に予測した．

図 6.1 に，提案手法の失敗例を示す．図 6.1 における (e) および (f) は，False Positive の例である．対象物体および配置目標は，それぞれ (e) に写るタオルおよび (f) に写る浴槽である．また，対象物体候補は ground truth に一致しておらず (IoU < 0.7)，配置目標候補は ground truth に一致している．したがって $(y_{\text{targ}}, y_{\text{dest}}) = (0, 1)$ であり，この例において提案手法は $(\hat{y}_{\text{targ}}, \hat{y}_{\text{dest}}) = (1, 1)$ であると予測した．すなわち，配置目標候補については正しく予測した一方，対象物体候補については鏡に映るタオルが対象物体であると誤って予測した．原因として，モデルが鏡に映ったタオルを実物だと誤って予測したと考えられる．さらに，物体の誤検出により対象物体候補領域が複数の物体を含んでいたことも原因として考えられる．

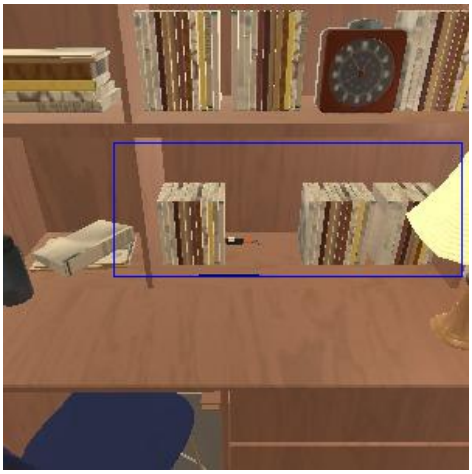
(a) 対象物体 / 対象物体候補



(b) 配置目標 / 配置目標候補

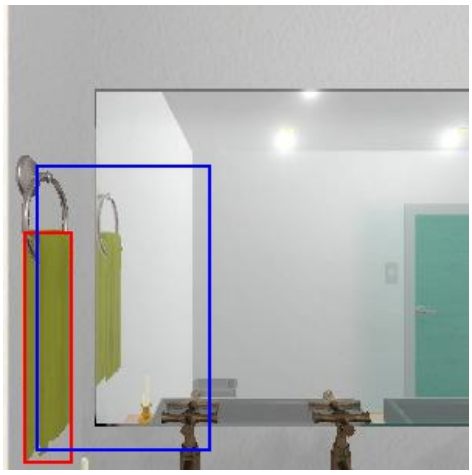
指示文：“Move the soap from the shelves to the metal rack.”

(c) 対象物体候補

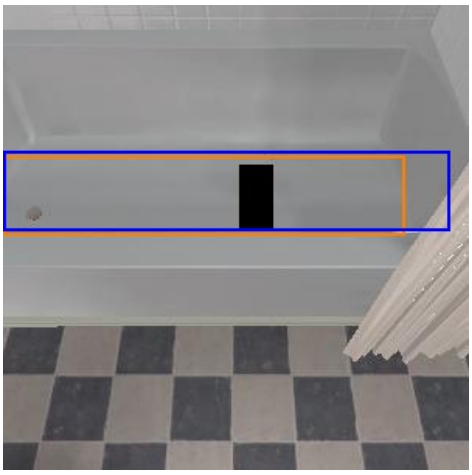


(d) 配置目標候補

指示文：“Put a salt shaker into a kitchen drawer.”



(e) 対象物体 / 対象物体候補



(f) 配置目標 / 配置目標候補

指示文：“Put a towel in the bath tub.”

図 6.1: ALFRED-fc データセットにおける定性的結果

6.1.3 Ablation Study

Ablation study として，以下の 2 条件を定めた．

(ii) W/o Switching head 機構：Switching head 機構による精度向上への寄与を調べるため，以下の式に従い，それぞれのモードで $\boldsymbol{x}_{\text{targ}}$ と $\boldsymbol{x}_{\text{dest}}$ を同一の値にする．

$$(\boldsymbol{x}_{\text{targ}}, \boldsymbol{x}_{\text{dest}}) = \begin{cases} (\boldsymbol{x}_{\text{targ}}, \boldsymbol{x}_{\text{targ}}) & \text{if target mode} \\ (\boldsymbol{x}_{\text{dest}}, \boldsymbol{x}_{\text{dest}}) & \text{if destination mode} \end{cases}$$

(iii) W/o Switching tail 機構：Switching tail 機構におけるマルチタスク学習の精度向上への寄与を調べるため，単一モデルで対象物体および配置目標に関するシングルタスクの学習を同時に行う．

表 6.1 に，Ablation study の定量的結果を示す．表より，精度において条件 (ii) および (iii) のモデルは提案手法 (iv) と比較してそれぞれ 4.7 および 6.2 ポイント低かった．これらの性能差は統計有意であった ($p < 0.01$)．これより，Switching head 機構および Switching tail 機構のどちらも性能向上に寄与しており，特に後者の導入が効果的であった．

6.1.4 エラー分析

表 6.2 に，提案手法におけるテスト集合の混同行列を示す．True Positive，True Negative，False Positive，および False Negative はそれぞれ 453，662，24，および 233 サンプル存在した．提案手法における失敗例は，テスト集合中に合計 257 サンプル存在した．このうち，False Positive および False Negative はそれぞれ 24 および 233 サンプルである．

表 6.3 に，ALFRED-fc データセットを用いた提案手法の評価における失敗例の分類結果を示す．合計 100 の失敗例を人手で分析した．なお，提案手法では式 (4.1) に基づき target mode および destination mode においてそれぞれ 1 回ずつの推論により 1 サンプルの予測ラベル $\hat{y}$ を得る．また，データセットにおける正解ラベル y は式 (6.1) により定義さ

表 6.2: ALFRED-fc データセットにおける混同行列

		予測ラベル	
		Positive	Negative
正解ラベル	Positive	453	233
	Negative	24	662

表 6.3: ALFRED-fc データセットにおける失敗例の分類

エラー ID	詳細	Target Mode	Destination Mode
SC	深刻な理解誤り	34	25
SOF	物体や家具の類似	8	7
SR	矩形領域の過小	7	0
IVI	視覚情報の不足	0	15
II	不完全な指示文	0	1
IL	ラベル誤り	1	2
合計	-	50	50

れる．しかし，この方法では例えば $(y_{\text{targ}}, y_{\text{dest}}) = (0, 1)$ の組から作成した負例に対して $(\hat{y}_{\text{targ}}, \hat{y}_{\text{dest}}) = (1, 0)$ とモデルが予測した場合に正解と扱われる．そのため，各モードにおいて個別に失敗例を分類した．失敗の要因は 6 種類に大別される．

1. Serious Comprehension error (SC)：
   SC は，候補領域と ground truth とを比較して何らの特徴も一致しないものを示す．例えば，“Move a mug with spoon into the sink.” という指示文に対して，マグカップと大きく特徴が異なるスプレーボトルを対象物体と予測してしまった場合である．
2. Similar Object or Furniture (SOF)：
   SOF は，対象物体と対象物体候補，または配置目標と配置目標候補が類似しているものを示す．例えば，配置目標がカウチで配置目標候補がソファーである場合である．両者は椅子である点で視覚的に類似しているが，前者は複数人掛け，後者は一人掛けである点で言語的な差異が存在し，互いに区別されるべきである．
3. Small Region (SR)：
   SR は，候補領域が極端に小さいものを示す．
4. Insufficient Visual Information (IVI)：
   IVI は，候補領域が物体または家具を十分に包含できておらず視覚的な特徴が掴みづらいものを示す．
5. Incomplete Instruction (II)：
   II は，指示文の情報が不完全であるものを示す．例えば，“Put away the bottle of wine.” という指示文には配置目標に関する明確な言語情報が存在しない．
6. Incorrect Label (IL)：
   IL は，指示文のタイピングミスや指示文とアノテーションされたラベルの不一致と

いった，データセットのラベル誤りを示す．

表 6.3 より，target mode および destination mode に共通して SC が主要な失敗要因であると言える．この点については，汎用的な大規模データセットで訓練された視覚言語モデルである CLIP [Radford *et al.* 21] を導入することが有効だと考えられる．

6.2 実機実験

6.2.1 定量的結果

表 6.1 に，実機における言語理解精度を示す．なお，言語理解性能を評価するには負例も必要であるため，5.1.1 節で説明された ALFRED-fc データセットと同様の前処理により負例を作成した．表 6.1 より，ベースライン手法は精度が 52.0%であるのに対し，提案手法は 55.9%であり，提案手法が 3.9 ポイント上回った．

また，表 6.4 に実機における把持および配置タスクの成功率を示す．なお，言語理解タスクにおいて予測が True Positive であった場合のみ，ロボットに把持および配置動作を行わせた．本研究は把持および配置動作について学習に基づく新規手法を提案するものではないが，表 6.4 から，実機において言語理解と動作実行を統合可能であることが示唆される．

評価指標として，言語理解精度およびタスク成功率を用いた．本実験においては，タスク成功率を対象物体の把持成功率および配置目標への配置成功率に細分化した．おのおのについて，成功率 SR を以下のように定義する．ここで，N_{a} および N_{s} はそれぞれ試行回数および成功回数を表す．

$$\mathrm{SR} = \frac{N_{\mathrm{s}}}{N_{\mathrm{a}}}$$

表 6.4: 実機におけるタスク成功率

タスク	成功回数 / 試行回数	成功率 [%]
把持	60 / 63	95
配置	56 / 60	93

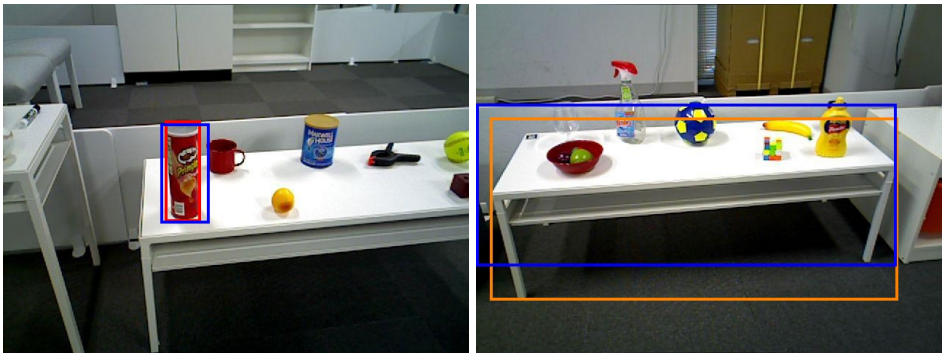
6.2.2 定性的結果

図 6.2 に定性的結果を示す．赤色，橙色，および青色の矩形領域はそれぞれ対象物体の ground truth，配置目標の ground truth，および対象物体候補または配置目標候補を表す．図 6.2 における (a)-(d) の例において，対象物体および配置目標はそれぞれ (a) に写る赤いポテトチップス缶および (b) に写るサッカーボールの乗った白いテーブルである．対象物体候補および配置目標候補はともに ground truth に一致しているため，$(y_{\text{targ}}, y_{\text{dest}}) = (1, 1)$ の例である．この例に対して提案手法は $(\hat{y}_{\text{targ}}, \hat{y}_{\text{dest}}) = (1, 1)$ であると正しく予測した．その上で，ロボットが (c) のように正確にポテトチップス缶を把持し，(d) のように机への配置に成功した．

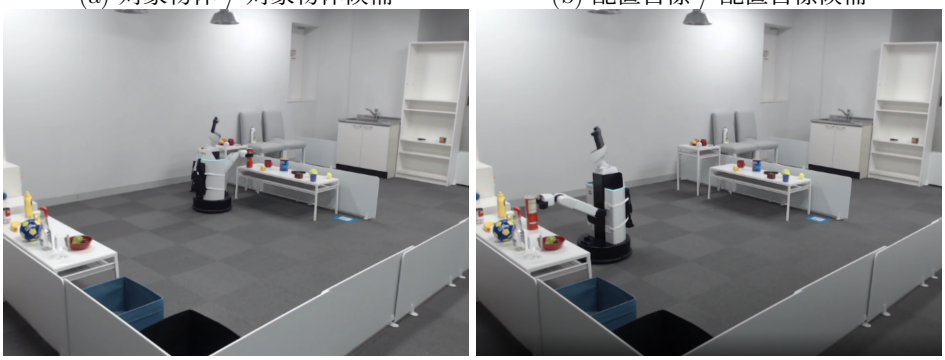
同様に，図 6.2 における (e)-(h) の例において，対象物体および配置目標はそれぞれ (e) に写る緑色のカップおよび (f) に写る青色の箱である．対象物体候補および配置目標候補はともに ground truth に一致しているため，$(y_{\text{targ}}, y_{\text{dest}}) = (1, 1)$ の例である．この例に対して提案手法は $(\hat{y}_{\text{targ}}, \hat{y}_{\text{dest}}) = (1, 1)$ であると正しく予測した．その上で，ロボットが (g) のように正確にカップを把持し，(h) のように箱への配置に成功した．

6.2.3 考察

典型的なシーンにおいて，対象物体候補および配置目標候補はそれぞれ平均して 73 個および 89 個検出された．1 回の推論時間は約 4×10^{-3} 秒であるため，ベースライン手法および提案手法の計算時間はそれぞれ 26 秒（6497 回の推論）および 0.6 秒（162 回の推論）であると考えられる．



(a) 対象物体 / 対象物体候補　(b) 配置目標 / 配置目標候補



(c) 把持動作　(d) 配置動作

指示文：“Put the red chips can on the white table with the soccer ball on it.”

(e) 対象物体 / 対象物体候補　(f) 配置目標 / 配置目標候補

(g) 把持動作　(h) 配置動作

指示文：“Place the green cup in the blue bin.”

図 6.2: 実機における定性的結果

第7章

結論

本研究では，物体操作タスクにおける対象物体および配置目標に関する指示文を理解し実行するための手法の構築を目的とした．そのために，複数の画像から指示文の対象物体および配置目標の両方を特定し，ロボットが対象物体を配置目標まで運搬する Dual Referring Expression Comprehension with fetch-and-carry (DREC-fc) タスクを扱った．

本研究の貢献を以下に示す．

- 言語理解における対象物体および配置目標の探索に必要な推論回数を削減することが可能な Switching Head-Tail Funnel UNITER (SHeFU) を提案した．
- Switching head-tail 機構を導入することで，対象物体および配置目標について，単一モデルで個別に予測することを可能にした．
- ALFRED [Shridhar *et al.* 20] を基にした DREC-fc タスクにおけるデータセットである ALFRED-fc データセットにおいて，SHeFU がベースライン手法を言語理解精度で上回った．
- 実機ロボットを用いて実環境における実験を行い，SHeFU がベースライン手法を言語理解精度で上回った．さらに，把持および配置動作を高い成功率で実行可能であることを示した．

将来研究として，実環境における言語理解精度を向上させるため，ドメイン適応や sim2real に関する手法を導入することが挙げられる．

参考文献

[Calli *et al.* 15] Berk Calli, Aaron Walsman, Arjun Singh, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. Benchmarking in Manipulation Research: The YCB Object and Model Set and Benchmarking Protocols. *arXiv preprint arXiv:1502.03143*, 2015.

[Chen *et al.* 20] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: UNiversal Image-TExt Representation Learning. In *ECCV*, pp. 104–120, 2020.

[Dai *et al.* 20] Zihang Dai, Guokun Lai, Yiming Yang, and Quoc Le. Funnel-Transformer: Filtering out Sequential Redundancy for Efficient Language Processing. *NeurIPS*, Vol. 33, pp. 4271–4282, 2020.

[Fukui *et al.* 19] Hiroshi Fukui, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Attention Branch Network: Learning of Attention Mechanism for Visual Explanation. In *CVPR*, pp. 10705–10714, 2019.

[Hatori *et al.* 18] Jun Hatori, Yuta Kikuchi, Sosuke Kobayashi, Kuniyuki Takahashi, Yuta Tsuboi, Yuya Unno, Wilson Ko, and Jethro Tan. Interactively Picking Real-World Objects with Unconstrained Spoken Language Instructions. In *ICRA*, pp. 3774–3781, 2018.

[He *et al.* 16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, pp. 770–778, 2016.

[Inoue *et al.* 22] Yuki Inoue, and Hiroki Ohashi. Prompter: Utilizing Large Language Model Prompting for a Data Efficient Embodied Instruction Following. *arXiv preprint arXiv:2211.03267*, 2022.

[Iocchi *et al.* 15] Luca Iocchi, Dirk Holz, Javier Ruiz-del Solar, Komei Sugiura, and Tijn van der Zant. RoboCup@Home: Analysis and Results of Evolving Competitions for Domestic and Service Robots. *Artificial Intelligence*, Vol. 229, pp. 258–281, 2015.

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

[Ishikawa *et al.* 22] Shintaro Ishikawa, and Komei Sugiura. Moment-based Adversarial Training for Embodied Language Comprehension. In *ICPR*, pp. 4139–4145. IEEE, 2022.

[Jia *et al.* 22] Zhiwei Jia, Kaixiang Lin, Yizhou Zhao, Qiaozi Gao, Govind Thattai, and Gaurav S Sukhatme. Learning to Act with Affordance-Aware Multimodal Neural SLAM. In *IROS*, pp. 5877–5884. IEEE, 2022.

[Kenton *et al.* 19] Jacob Devlin Ming-Wei Chang Kenton, and Lee Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

[Kolve *et al.* 17] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, Aniruddha Kembhavi, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An Interactive 3D Environment for Visual AI. *arXiv preprint arXiv:1712.05474*, 2017.

[Korekata *et al.* 23] Ryosuke Korekata, Motonari Kambara, Yu Yoshida, Shintaro Ishikawa, Yosuke Kawasaki, Masaki Takahashi, and Komei Sugiura. Switching Head–Tail Funnel UNITER for Dual Referring Expression Comprehension with Fetch-and-Carry Tasks. In *IROS*. IEEE, 2023. to appear.

[Magassouba *et al.* 19] Aly Magassouba, Komei Sugiura, Anh Trinh Quoc, and Hisashi Kawai. Understanding Natural Language Instructions for Fetching Daily Objects Using GAN-Based Multimodal Target–Source Classification. *RA-L*, Vol. 4, No. 4, pp. 3884–3891, 2019.

[Magassouba *et al.* 20] Aly Magassouba, Komei Sugiura, and Hisashi Kawai. A Multimodal Target-Source Classifier With Attention Branches to Understand Ambiguous Instructions for Fetching Daily Objects. *RA-L*, Vol. 5, No. 2, pp. 532–539, 2020.

[Min *et al.* 22] So Yeon Min, Devendra Singh Chaplot, Pradeep Ravikumar, Yonatan Bisk, and Ruslan Salakhutdinov. FILM: Following Instructions in Language with Modular Methods. In *ICLR*, 2022.

[Mogadala *et al.* 21] Aditya Mogadala, Marimuthu Kalimuthu, and Dietrich Klakow. Trends in Integration of Vision and Language Research: A Survey of Tasks, Datasets, and Methods. *JAIR*, Vol. 71, pp. 1183–1317, 2021.

[Ogura *et al.* 20] Tadashi Ogura, Aly Magassouba, Komei Sugiura, Tsubasa Hirakawa, Takayoshi Yamashita, Hironobu Fujiyoshi, and Hisashi Kawai. Alleviating the Burden of Labeling: Sentence Generation by Attention Branch Encoder–Decoder Network. *RA-L*, Vol. 5, No. 4, pp. 5945–5952, 2020.

[Okada *et al.* 19] Hiroyuki Okada, Tetsunari Inamura, and Kazuyoshi Wada. What competitions were conducted in the service categories of the World Robot Summit? *Advanced Robotics*, Vol. 33, No. 17, pp. 900–910, 2019.

[Qiao *et al.* 20] Yanyuan Qiao, Chaorui Deng, and Qi Wu. Referring Expression Comprehension: A Survey of Methods and Datasets. *Trans. Multimed.*, Vol. 23, pp. 4426–4440, 2020.

[Radford *et al.* 21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, pp. 8748–8763. PMLR, 2021.

[Ren *et al.* 16] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Trans. PAMI*, Vol. 39, No. 6, pp. 1137–1149, 2016.

[Shridhar *et al.* 20] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. ALFRED: A Bench-

mark for Interpreting Grounded Instructions for Everyday Tasks. In *CVPR*, pp. 10740–10749, 2020.

[Vaswani *et al.* 17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *NeurIPS*, Vol. 30, pp. 5998–6008, 2017.

[WRS20] World Robot Summit 2020 Partner robot challenge Real Space Rules & Regulations. `https://worldrobotsummit.org/wrs2020/challenge/download/Rules/DetailedRules_Partner_EN.pdf`, 2020. [Online; accessed 31-Jan-2023].

[Wu *et al.* 16] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv preprint arXiv:1609.08144*, 2016.

[Yamamoto *et al.* 19] Takashi Yamamoto, Koji Terada, Akiyoshi Ochiai, Fuminori Saito, Yoshiaki Asahara, and Kazuto Murase. Development of Human Support Robot as the research platform of a domestic mobile manipulator. *ROBOMECH J.*, Vol. 6, No. 1, pp. 1–15, 2019.

[是方 他 23] 是方諒介, 神原元就, 吉田悠, 石川慎太朗, 川崎陽祐, 高橋正樹, 杉浦孔明. Switching Head–Tail 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と物体操作. 2023 年度 人工知能学会全国大会, 2023. 2G4-OS-21d-01.