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

Instruction Comprehension Based on Funnel UNITER for Object Manipulation Tasks

吉田 悠 *1　石川 慎太朗 *1　杉浦 孔明 *1
Yu Yoshida　Shintaro Ishikawa　Komei Sugiura

*1 慶應義塾大学
Keio University

In this study, we develop a multimodal language comprehension model that allows domestic service robots to understand object fetching instructions. We propose a multimodal language understanding model, Funnel UNITER, which gradually reduces the dimensions of the query, key, and value in each transformer layer to reduce the computational cost of self-attention. We also built a new dataset for the multimodal language understanding for fetching instruction (MLU-FI) task called the ALFRED-fetch dataset. Our model outperformed the baseline method in both classification accuracy and training time.

1. はじめに

高齢化が進行している現代社会において，日常生活における介助支援の必要性は高まっている．その結果，在宅介助者の不足が社会問題となっており，これを解決するため，被介助者を物理的に支援可能な生活支援ロボットが注目されている．しかし，自然言語による人間からの指示を，生活支援ロボットが理解する能力については，現状不十分である．

本研究では，ロボットが人間による物体操作指示命令を理解するための手法を構築することを目的とする．具体的には，“Pick up the phone that’s above the remote.” という命令文が与えられたときに，ロボットがリモコンの上の携帯電話を命令文の対象として特定する．しかし，人間の発する命令文には，しばしば内容に曖昧性が生じるため，ロボットが対象となる物体を特定することは困難である．実際に，物体操作を含む VLN において標準ベンチマークである ALFRED では，人間のパフォーマンスは 90.1%と報告されている [Shridhar 20] が，最先端の手法 (e.g., [Blukis 22]) では 25%以下しか達成できていない．

このタスクには多くの既存手法が存在するが，その性能は十分ではない (e.g., [Magassouba 19] [Magassouba 20] [Ishikawa 21])．例えば，Target-dependent UNITER [Ishikawa 21] は，MLU-FI において transformer [Vaswani 17] を使用することで高い精度を達成したモデルである．しかし，モデルに単純な transformer を使用しており，計算コストが高く，精度が低いという問題を抱えている．

本研究では，Funnel Transformer [Dai 20] に基づいて Target-dependent UNITER [Ishikawa 21] を拡張し，十分な精度を達成しつつ，計算コストの削減を行う Funnel UNITER を提案する．既存手法との違いとして，transformer の各層において，Query， Key， Value の次元数を削減し，self-attention の計算コストを削減する構造が加えられている．また，これによりモデルのパラメータ数が削減され，同じデータ数で効率的に学習可能になることから，精度を改善することも期待できる．

本論文の独自性を以下に示す．

連絡先: 吉田悠，慶應義塾大学，神奈川県横浜市港北区日吉 3-14-1，yu1015@keio.jp

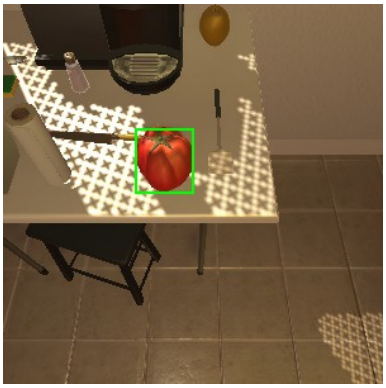


図 1: MLU-FI のシーン例

- Target-dependent UNITER を拡張し，transformer の self-attention における計算量を削減した．
- 本タスクにおける新たなデータセットである ALFRED-fetch を収集し，評価を行った

2. 問題設定

本論文で扱うタスクは Multimodal Language Understanding for Fetching Instruction (MLU-FI) である．MLU-FI とは，物体検出により獲得した各物体の中から，物体操作に関する命令文が対象とする物体を特定するというタスクである．図 1 に，MLU-FI の例を示す．この画像に対し，”Pick up the tomato on the table.” という命令文が与えられた場合に，画像の中央に存在するトマトを対象物体と判定することが求められる．本タスクにおける入出力を以下に定義する．

- **入力:** 物体操作に関する命令文，対象物体候補の領域，画像中の各物体の領域
- **出力:** 対象物体候補が対象物体である確率の予測値
  対象物体候補が命令文が対象とする物体であれば 1 を，そうでなければ 0 を出力することが望ましい．

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:** 命令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

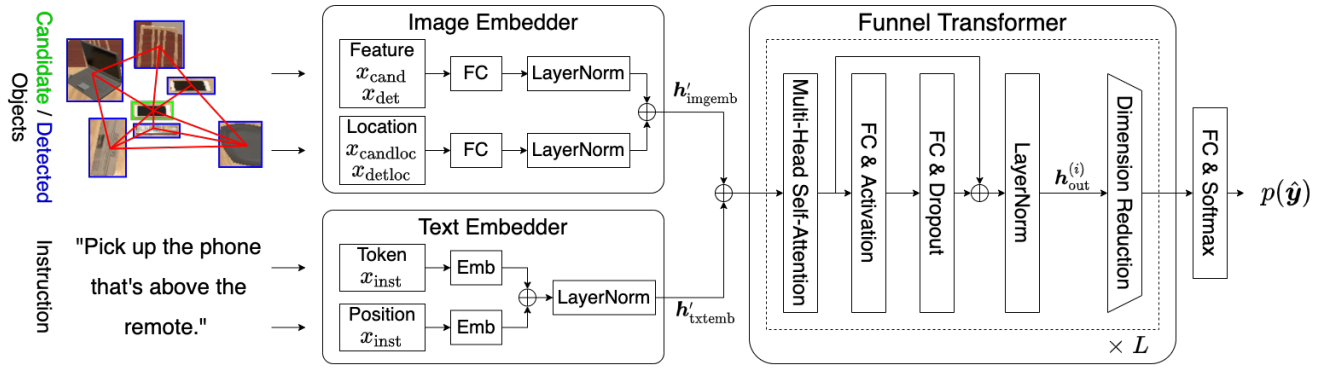


図 2: 提案手法のネットワーク構造

このタスクでは，画像中の各物体から対象物体を特定する多クラス分類ではなく，各物体に対し対象物体であるかを判定する 2 クラス分類を扱う．これは，画像中に対象物体が複数存在する場合や，画像中に対象物体が含まれない場合を考慮するためである．タスクの評価尺度には，分類精度を使用する．

3. 提案手法

ネットワークの構造を図2 に示す．Instruction は命令文，Candidate Object は対象物体候補，Detected Object は画像中の各物体を表す．ネットワークは Image Embedder，Text Embedder，Funnel Transformer といった 3 つのモジュールから構成される．

3.1 入力

ネットワークの入力 $\boldsymbol{x}$ を以下のように定義する．

$$\boldsymbol{x} = \{\boldsymbol{X}_{\text{inst}}, \boldsymbol{X}_{\text{cand}}, \boldsymbol{X}_{\text{det}}\}, \tag{1}$$

$$\boldsymbol{X}_{\text{inst}} = \{\boldsymbol{x}_{\text{inst}}, \boldsymbol{x}_{\text{pos}}\}, \tag{2}$$

$$\boldsymbol{X}_{\text{cand}} = \{\boldsymbol{x}_{\text{cand}}, \boldsymbol{x}_{\text{candloc}}\}, \tag{3}$$

$$\boldsymbol{X}_{\text{det}} = \{(\boldsymbol{x}^{(i)}_{\text{det}}, \boldsymbol{x}^{(i)}_{\text{detloc}}) | i = 1, \ldots, N\}, \tag{4}$$

$\boldsymbol{x}_{inst}$ は命令文，$\boldsymbol{x}_{cand}$ は対象物体候補の領域，$\boldsymbol{x}^{(i)}_{det}$ は画像中の各物体の領域を表し，$\boldsymbol{x}_{pos}$ は命令文中の各単語の位置，$\boldsymbol{x}_{candloc}$ は対象物体候補の領域位置，$\boldsymbol{x}^{(i)}_{detloc}$ は画像中の各物体の領域位置を表す．ここで，N は Faster R-CNN [Ren 16] により検出された画像中の領域の数を示す．

3.2 Image Embedder

Image Embedder では，対象物体候補の領域および画像中の各物体の領域に対する埋め込み処理を行う．入力は，$\boldsymbol{x}_{\text{cand}}$，$\boldsymbol{x}_{\text{candloc}}$，$\boldsymbol{x}^{(i)}_{\text{det}}$，$\boldsymbol{x}^{(i)}_{\text{detloc}}$ から構成される．

$\boldsymbol{x}^{(i)}_{\text{cand}}$，$\boldsymbol{x}^{(i)}_{\text{det}}$ は Faster R-CNN [Ren 16] のバックボーンネットワークである ResNet101 [He 16] のうち，fc7 層の出力を画像領域の特徴量として抽出したものである．$\boldsymbol{x}^{(i)}_{\text{candloc}}$，$\boldsymbol{x}^{(i)}_{\text{detloc}}$ は各矩形領域の左上と右下の頂点の座標を (x_1, y_1)，(x_2, y_2)，幅と高さをそれぞれ w，h とするとき，$[x_1, y_1, x_2, y_2, w, h, w \times h]$ という 7 次元ベクトルである．

まず画像中の各物体に対して，$\boldsymbol{x}^{(i)}_{\text{det}}$，$\boldsymbol{x}^{(i)}_{\text{detloc}}$ をそれぞれ全結合層に入力し，得られた出力を連結した後，正規化を行うことで出力 $\boldsymbol{h}'_{\text{det}}{}^{(i)}$ を得る．以上の処理を数式として示す．

$$\boldsymbol{h}'_{\text{det}}{}^{(i)} = f_{\text{LN}}(f_{\text{FC}}(\boldsymbol{x}^{(i)}_{\text{det}}), f_{\text{FC}}(\boldsymbol{x}^{(i)}_{\text{detloc}})) \tag{5}$$

ここで，f_{LN} と f_{FC} はそれぞれ正規化層，全結合層を示す．対象物体候補についても同様の処理を行い，$\boldsymbol{h}'_{\text{cand}}$ を得る．

$$\boldsymbol{h}'_{\text{cand}} = f_{\text{LN}}(f_{\text{FC}}(\boldsymbol{x}_{\text{cand}}), f_{\text{FC}}(\boldsymbol{x}_{\text{candloc}})) \tag{6}$$

最後に，$\boldsymbol{h}'_{\text{det}}{}^{(i)}$ と $\boldsymbol{h}'_{\text{cand}}{}^{(i)}$ を連結し，出力 $\boldsymbol{h}'_{\text{imgemb}}$ を得る．

$$\boldsymbol{h}'_{\text{imgemb}} = \{\boldsymbol{h}'_{\text{cand}}, \boldsymbol{h}'_{\text{det}}{}^{(i)}, \ldots, \boldsymbol{h}'_{\text{det}}{}^{(N)}\} \tag{7}$$

3.3 Text Embedder

Text Embedder では，命令文に対する埋め込み処理を行う．入力は，$\boldsymbol{x}_{inst}$ と $\boldsymbol{x}_{pos}$ から構成される．命令文は WordPiece によるトークン化を行い，単語列をトークン列に変換される．$\boldsymbol{x}_{inst}$，$\boldsymbol{x}_{pos}$ は Token ID と命令文における単語の位置，それぞれの one-hot ベクトル集合を表す．

$\boldsymbol{x}_{\text{inst}}$，$\boldsymbol{x}_{\text{pos}}$ に学習可能な重みである $\boldsymbol{W}_{\text{inst}}$，$\boldsymbol{W}_{\text{pos}}$ をそれぞれ掛け合わせ，正規化を行うことで出力 $\boldsymbol{h}'_{\text{txtemb}}$ を得る．

$$\boldsymbol{h}'_{\text{txtemb}} = f_{\text{LN}}(\{\boldsymbol{W}_{\text{inst}}\boldsymbol{x}_{\text{inst}}, \boldsymbol{W}_{\text{pos}}\boldsymbol{x}_{\text{pos}}\}) \tag{8}$$

3.4 Funnel Transformer

本モジュールは L 層の Funnel Transformer [Dai 20] で構成される．1 層目の入力 $\boldsymbol{h}^{(1)}_{\text{in}}$ は以下の式で得られる．

$$\boldsymbol{h}^{(1)}_{\text{in}} = \{\boldsymbol{h}'_{\text{imgemb}}, \boldsymbol{h}'_{\text{txtemb}}\} \tag{9}$$

まず，以下の式によって，query $Q^{(i)}$，key $K^{(i)}$，value $V^{(i)}$ を生成する．

$$Q^{(i)} = W^{(i)}_{\text{q}}\boldsymbol{h}^{(i)}_{\text{in}} \tag{10}$$

$$K^{(i)} = W^{(i)}_{\text{k}}\boldsymbol{h}^{(i)}_{\text{in}} \tag{11}$$

$$V^{(i)} = W^{(i)}_{\text{v}}\boldsymbol{h}^{(i)}_{\text{in}} \tag{12}$$

ここで，i は Attention Head のインデックス，$W^{(i)}_{\text{q}}$，$W^{(i)}_{\text{k}}$，$W^{(i)}_{\text{v}}$ は学習可能な重みを示す．次に，1 層目における Attentin スコアを算出する．ここで，$H^{(1)}$ は 1 層目における Q，K，V の次元数，$A^{(1)}$ は 1 層目における Attention Head 数を示す．

$$S^{(1)}_{\text{attn}} = \{\boldsymbol{f}^{(1)}_{\text{attn}}, \ldots, \boldsymbol{f}^{(A^{(1)})}_{\text{attn}}\} \tag{13}$$

$$\boldsymbol{f}^{(i)}_{\text{attn}} = \text{softmax}(\frac{Q^{(i)}K^{(i)\mathsf{T}}}{\sqrt{d_k}})V^{(i)} \tag{14}$$

$$d_k = \frac{H^{(1)}}{A^{(1)}} \tag{15}$$

1 層目の出力 $\boldsymbol{h}^{(1)}_{\text{out}}$ は以下の式により得られる．

$$\boldsymbol{h}^{(1)}_{\text{out}} = f_{\text{LN}}(S^{(1)}_{\text{attn}} + f_{\text{FC}}(f_{\text{FC}}(S_{\text{attn}}))) \tag{16}$$

2 層目以降では，i 層目において，$i-1$ 層の出力 $\boldsymbol{h}^{(i-1)}_{\text{out}}$ に対し，max pooling を用いて次元数を削減する．ここで，i 層目

表 1: ALFRED-fetch と WRS-UniALT2 における定量的結果

Method	ALFRED-fetch		WRS-UniALT2	
	Acc[%]↑	Training time[fps]↑	Acc[%]↑	Training time[fps]↑
(i) Baseline ($L=2$)	82.0 ± 1.79	92.5 ± 0.26	71.1 ± 4.43	126 ± 0.45
(ii) Ours ($L=4$)	86.0 ± 0.64	76.9 ± 0.39	84.1 ± 6.41	115 ± 0.35
(iii) Ours ($L=3$)	85.9 ± 1.79	85.1 ± 0.38	79.7 ± 5.27	120 ± 1.40
(iv) Ours ($L=2$)	$\mathbf{86.6 \pm 1.62}$	$\mathbf{94.1 \pm 0.71}$	$\mathbf{87.7 \pm 6.46}$	$\mathbf{130 \pm 0.22}$

における Q，K，V の次元数 $H^{(1)}$，及び Attention の Head 数 $A^{(1)}$ を以下のように算出する．なお，$\lfloor \cdot \rfloor$ は床関数を示す．

$$H^{(i)} = \lfloor H^{(i-1)}/2 \rfloor \tag{17}$$

$$A^{(i)} = \lfloor A^{(i-1)}/2 \rfloor \tag{18}$$

$\boldsymbol{h}_{\text{in}}^{(i)}$ を入力とし，i 層目の出力 $\boldsymbol{h}_{\text{out}}^{(i)}$ を式 (10) から (16) と同様に算出する．ここで，元の Funnel Transformer [Dai 20] では query のみに max pooling が行われていたが，本研究では query，key，value の全てに max pooing を適用した．

Funnel Transformer の最終的な出力を $\boldsymbol{h}'_{\text{out}}$ とすると，モデル全体の最終出力 $p(\hat{\boldsymbol{y}})$ は，以下の式によって得られる．

$$p(\hat{\boldsymbol{y}}) = \text{softmax}(f_{\text{FC}}(\boldsymbol{h}'_{\text{out}})) \tag{19}$$

なお，損失関数には二値交差エントロピー誤差を使用した．

4. 実験

4.1 データセット

本実験では，データセットとして ALFRED-fetch と WRS-UniALT2 を使用した．

4.1.1 ALFRED-fetch dataset

本研究では，ALFRED [Shridhar 20] をベースとして，MLU-FI のためのデータセットである ALFRED-fetch を収集した．ALFRED は自然言語による指示と一人称視点からエージェントの行動を学習するためのデータセットであり，エージェントの行動について，複数のサブゴールが逐次的に設定されており，各サブゴールに対して命令文が存在する．しかし，エージェントが物体を運ぶ際にカメラ画像に空中に浮かんだ物体が表示され，これは物体操作の研究において不適切であった．そこで，物体把持がサブゴールである状況において，物体を掴む直前のエージェントの一人称視点の指示と画像を収集した．また，ALFRED には，各画像に対し対象物体の ground-truth となる領域データが含まれているが，他の全ての物体はアノテーションされていないため，Faster R-CNN [Ren 16] を用いて領域を抽出した．各対象物体候補に対し，ground-truth に対する IoU（Intersection over Union）が 0.7 以上であれば正例，0.3 以下であれば負例と判定し，正負のバランスをとるために，負例をランダムに削減した．

ALFRED-fetch は，3428 枚の画像と 3227 文の指示文を含み，語彙サイズは 884，全単語数は 40254，平均文長は 12.5 となっている．指示文は，[Shridhar 20] で実施した Amazon Mechanical Turk (AMT) により，少なくとも 3 人の異なるアノテーターから収集したものである．また，47862 サンプルのうち，43439 サンプルを訓練集合に，3447 サンプルを検証集合に，976 サンプルをテスト集合にそれぞれ使用した．

4.1.2 WRS-UniALT2 dataset

WRS-UniALT2 は，WRS-UniALT [Ishikawa 21] に基づく MLU-FI のためのデータセットである．WRS- UniALT は，画像と命令セットからなるシミュレーションベースのデータセットであり，画像は World Robot Summit Partner Robot Challenge [Okada 19] の標準シミュレータで収集された．画像は約 5 種類の日用品を部屋の中に配置したものであり，命令文は 6 人のアノテータによって英語で記述された物体操作命令である．

本研究では，タスクの難易度を上げるために，WRS-UniALT に修正を行った．これは，誤検出された領域を負例であると予測することが非常に容易であったためである．このようなケースを避けるため，誤検出領域のみを含む画像はデータセットに含まないよう変更した．ここで，誤検出とは，画像中の全ての物体との IoU が 0.05 未満の場合と定義した．

WRS-UniALT2 には，570 枚の画像と 1246 文の指示文を含み，語彙サイズは 167 語，全語数は 8816 語，平均文長は 7.1 語であった．また，2490 サンプルのうち，2048 サンプルを訓練集合に，210 サンプルを検証集合に，232 サンプルをテスト集合に使用した．

4.2 パラメータ設定

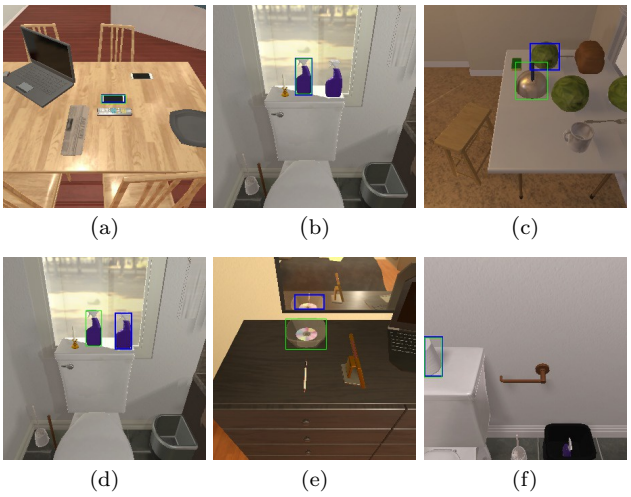
ネットワーク内の transformer [Vaswani 17] において，層数を $L=2$，隠れ層の 1 層目における Q，K，V の次元数を $H^{(1)} = (N+l+1) \times 768$，Attention Head 数を $A^{(1)} = 12$ と設定した．ここで，N は画像内の物体の数，l は命令文の単語数を示す．最適化には AdamW を使用し，学習率は 8×10^{-5}，ステップ数は 20000，バッチサイズは 8 であった．ここで，1 ステップは 1 つのバッチの処理を示す．また，Dropout の確率は 0.1 とした．

提案手法のパラメータ数は 3300 万であり，学習にはメモリ 11GB 搭載の GeForce RTX 2080 および Intel Core i9-9900K を使用した．ALFRED-fetch の学習には約 30 分，WRS-UniALT の学習には約 20 分，推論にはどちらも約 0.008 秒/sample を要した．

4.3 定量的結果

定量的結果を表 1 に示す．性能評価には精度と学習時間を使用した．ここで，学習時間は学習時において 1 秒間に処理可能な画像の枚数を示す．また，精度に関して，データセット内に正解サンプルと不正解サンプルは等量で存在するため，チャンスレートは 50%である．なお実験は 5 回行い，結果はその平均値と標準偏差を示す．

ALFRED-fetch では，ベースライン手法では精度が 82.0%，学習時間が 92.5fps であるのに対し，提案手法では精度が 86.6%，学習時間が 94.1fps となった．WRS-UniALT では，ベースライン手法では精度が 71.1%，学習時間が 126fps であるのに対し，提案手法では精度が 87.7%，学習時間が 130fps となった．この結果より，提案手法の Acc は，ALFRED-fetch において 4.6 ポイント，WRS-UniALT において 16.6 ポイント，ベースライン手法を上回った．同様に，学習時間においても，ALFRED-fetch において 1.6 ポイント，WRS-UniALT において 4.0 ポイント上回った．



(a) (b) (c)

(d) (e) (f)

図 3: PFN-PIC と WRS-UniALT における定性的結果

4.4 Ablation Study

Ablation study として，L を変化させた場合に性能にどのような影響があるかを調べた．表 1 に示すように，条件 (ii)，(iii) のモデルは，Acc に関して，ALFRED-fetch において 0.6，0.7 ポイント，WRS-UniALT において 3.6，8.0 ポイント，それぞれ提案手法を下回った．同様に，学習時間に関しても，ALFRED-fetch において 17.2，9.0 ポイント，WRS-UniALT において 15.0，10.0 ポイント，それぞれ提案手法を下回った．以上の結果より，どちらの条件も提案手法を下回る性能であったため，最適な層数は $L = 2$ であると考えられる．

4.5 定性的結果

図 3 に定性的結果を示す．青い矩形で囲まれた物体が対象物体候補，緑の矩形で囲まれた物体が対象物体の Ground Truth である．

(a) 及び (b) は TP の例を示す．(a) の命令文は “Pick up the phone that’s above the remote.” であり，対象物体候補である青の矩形領域について，$p(\hat{\boldsymbol{y}}) = 0.999$ と出力しており，正確に当該領域が対象物体を示していると判定している．同様に，(b) の命令文は “Pick up the left-most spray bottle on the back of the toilet.” であり，出力は $p(\hat{\boldsymbol{y}}) = 0.999$ であった．2 つ並んだスプレーのうち，対象物体候補である左側のスプレーが対象物体であると判定している

(c) 及び (d) は TN の例を示す．(c) の命令文は “Pick up the kettle from the table.” であり，出力は $p(\hat{\boldsymbol{y}}) = 9.89 \times 10^{-10}$ であった．対象物体はケトルであり，対象物体候補であるキャベツが対象物体でないことを判定できている．(d) の命令文は，(b) と同じ “Pick up the left-most spray bottle on the back of the toilet.” であり，出力は $p(\hat{\boldsymbol{y}}) = 1.83 \times 10^{-9}$ であった．対象物体は左側のスプレーであり，対象物体候補である右側のスプレーが対象物体ではないことを正確に判定している．

(e) は FP の例を示す．命令文は “pick up the bowl with the CD from the dresser” であり，出力は $p(\hat{\boldsymbol{y}}) = 0.999$ であった．対象物体は手前のボウルであるが，鏡に映るボウルを対象物体と判定してしまっている．対象物体候補が鏡に映った物であることを理解できていないことが原因だと考えられる

(f) は FN の例を示す．命令文は “Pick up the roll of toilet paper on the toilet tank.” であり，出力は $p(\hat{\boldsymbol{y}}) = 4.01 \times 10^{-5}$ であった．対象物体候補は対象物体を示しているにも関わらず，対象物体ではないと判定してしまっている．対象物体候補の全体が画像内に収まっていないため，うまく判定できなかったのではないかと考えられる．

5. おわりに

本論文では，物体操作指示命令を理解し，その対象物体を特定するモデルの構築を目的とした．そこで，画像中の全ての候補物体の中から命令の対象物体を特定するタスクである MLU-FI を扱った．

本論文の貢献を以下に示す．

- UNITER 型注意機構を拡張し，self-attention における計算コストの削減を行った．
- ALFRED [Shridhar 20] を基に，MLU-FI におけるデータセットである ALFRED-fetch を収集し，評価を行った．
- 2 つのデータセットにおいて，提案手法はベースライン手法を分類精度，学習時間で上回った．

将来研究として，物体操作を行う実機ロボットに対し提案モデルを実装することが挙げられる．

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

[Blukis 22] Blukis, V., Paxton, C., Fox, D., Garg, A., and Artzi, Y.: A persistent spatial semantic representation for high-level natural language instruction execution, in *CoRL*, pp. 706–717 (2022)

[Dai 20] Dai, Z., Lai, G., Yang, Y., and Le, Q.: Funnel-transformer: Filtering out sequential redundancy for efficient language processing, *NeurIPS*, Vol. 33, pp. 4271–4282 (2020)

[He 16] He, K., Zhang, X., Ren, S., and Sun, J.: Deep residual learning for image recognition, in *CVPR*, pp. 770–778 (2016)

[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)

[Magassouba 19] Magassouba, A., Sugiura, K., Quoc, A. T., and Kawai, H.: Understanding natural language instructions for fetching daily objects using gan-based multimodal target–source classification, *IEEE RA-L*, Vol. 4, No. 4, pp. 3884–3891 (2019)

[Magassouba 20] Magassouba, A., Sugiura, K., and Kawai, H.: A multimodal target-source classifier with attention branches to understand ambiguous instructions for fetching daily objects, *IEEE RA-L*, Vol. 5, No. 2, pp. 532–539 (2020)

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

[Ren 16] Ren, S., He, K., Girshick, R., and Sun, J.: 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., Thomason, J., Gordon, D., Bisk, Y., Han, W., Mottaghi, R., Zettlemoyer, L., and Fox, D.: 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. N., Kaiser, Ł., and Polosukhin, I.: Attention is all you need, *NeurIPS*, Vol. 30, (2017)