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

Understanding Object Fetching Instructions Including Referring Expressions about Target Objects Based on Target-Dependent UNITER

石川 慎太朗 *1　　杉浦 孔明 *1
Shintaro Ishikawa　　Komei Sugiura

*1 慶應義塾大学
Keio University

Currently, domestic service robots have an insufficient ability to interact naturally through language. This is because understanding human instructions is complicated by a variety of ambiguities and missing information. Existing methods are insufficient to model reference expressions that specify relationships between objects. In this paper, we propose Target-dependent UNITER, which learns directly the relationship between the target object and other objects by focusing on the relevant regions within an image, instead of the whole image. Our model is validated on two standard datasets, and the results show that Target-dependent UNITER outperforms the baseline method in terms of classification accuracy.

1. はじめに

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護・支援の必要性が高まっている．それに伴って，在宅介護者の不足が社会問題となっており，障がいを持つ人々を物理的に支援可能な生活支援ロボットに注目が集まっている．一方，生活支援ロボットが人間との間で言語を介した自然な対話を行う能力は，現状不十分である．

本研究では，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 [Magassouba 19]．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．

本研究では，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 [Chen 20] 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 [Vaswani 17] 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．

連絡先: 石川慎太朗，慶應義塾大学，神奈川県横浜市港北区日吉 3-14-1，shin.0116@keio.jp

図 1: MLU-FI のシーン例

本研究の独自性は以下である．

- 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．
- UNITER において，対象物体候補を扱う新規構造を導入する．

2. 問題設定

本論文で扱うタスクを，Multimodal Language Understanding for Fetching Instruction (MLU-FI) と定義する．MLU-FI は，物体操作に関する命令文および命令が対象としている物体の候補群が与えられるときに，正しい対象物体を候補群の中から特定するというタスクである．具体的には，図 1 に示す画像において，“Pick up the empty bottle on the shelf” という命令文が与えられたときに，棚の上にある空の瓶を特定するという内容が考えられる．

このタスクは，画像中の物体群から唯一の対象物体を特定する多クラス分類ではなく，各物体に対して対象物体であるか否かを判定する 2 クラス分類のタスクである．これにより，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ことができる．

このタスクを実行するモデルにおいては，命令文や物体群とともに入力される対象物体候補が，真に対象物体であれば 1 を出力し，そうでなければ 0 を出力するといった動作が望ましい．

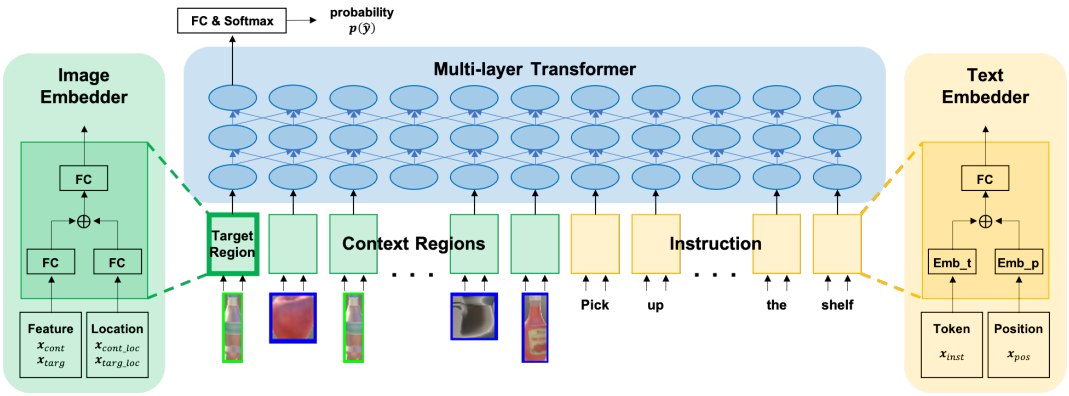


図 2: 提案手法のネットワーク構造

MLU-FI において，本論文では以下の入出力を想定する．

- **入力:** 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

ここで，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

画像中の各物体の領域は，事前学習済みの Faster R-CNN [Ren 16] に画像を入力することで獲得する．Faster R-CNN は，深層学習を利用したエンドツーエンドの物体検出モデルであり，本研究では，特徴マップの生成に ResNet101 [He 16] を用いている．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．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 [Vaswani 17] を複数層重ねたものである．

UNITER は事前学習を行うことによって，あらゆる視覚言語タスクに fine-tuning 可能なモデルとなっており，データ不足を補うことが可能である．事前学習では，Masked Language Modeling，Masked Region Modeling，Image–Text Matching，Word–Region Alignment といった 4 種類のタスクを実行する．

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．

$$\boldsymbol{x} = \{\boldsymbol{X}_{inst}, \boldsymbol{X}_{cont}, \boldsymbol{X}_{targ}\}, \tag{1}$$

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

$$\boldsymbol{X}_{targ} = \{\boldsymbol{x}_{targ}, \boldsymbol{x}_{targloc}\}, \tag{3}$$

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

$\boldsymbol{x}_{inst}$ は命令文，$\boldsymbol{x}_{targ}$ は対象物体候補の領域，$\boldsymbol{x}_{cont}^{(i)}$ 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，$\boldsymbol{x}_{targloc}$ は対象物体候補の領域位置，$\boldsymbol{x}_{contloc}^{(i)}$ は画像中の各物体の領域位置を表す．また，N は Faster R-CNN [Ren 16] によって検出した画像中の領域の数である．

3.2 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．

はじめに，命令文について WordPiece によるトークン化を行い，単語列をトークン列に変換する．ここで，先頭から i 番目のトークンに割り当てられているインデックスを n_i とするとき，$\boldsymbol{x}_{inst}$ は n_i 番目の要素が 1 の one-hot ベクトル集合，$\boldsymbol{x}_{pos}$ は i 番目の要素が 1 の one-hot ベクトル集合を表す．

続いて，これらのベクトル集合にそれぞれ W_{inst} と W_{pos} を掛け合わせることにより，線形の埋め込み処理を行う．なお，W_{inst} と W_{pos} は学習中に更新される重みである．

最後に，それぞれの出力を連結した後，全結合層 $f_{\mathrm{FC}}(\cdot)$ に入力することで，最終出力 $\boldsymbol{h}'_{txtemb}$ を得る．以上の処理を数式として示す．

$$\boldsymbol{h}'_{txtemb} = f_{\mathrm{FC}}(\{W_{inst}\boldsymbol{x}_{inst}, W_{pos}\boldsymbol{x}_{pos}\}) \tag{5}$$

3.3 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，$\boldsymbol{x}_{cont}^{(i)}$，$\boldsymbol{x}_{contloc}^{(i)}$，$\boldsymbol{x}_{targ}$，$\boldsymbol{x}_{targloc}$ から構成される．

$\boldsymbol{x}_{cont}^{(i)}$ は，全体画像を Faster R-CNN に入力することで得られる各領域の特徴量ベクトルを表す．本研究で使用した Faster R-CNN は，特徴量抽出に ResNet101 [He 16] を用いている．$\boldsymbol{x}_{contloc}^{(i)}$ は，各領域の座標値をエンコードして得られるベクトルである．これは，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．これらのベクトル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，再び全結合層に入力することで，出力 $\boldsymbol{h}'^{(i)}_{cont}$ を得る．以上の処理を数式として示す．

$$\boldsymbol{h}'^{(i)}_{cont} = f_{\mathrm{FC}}(\{f_{\mathrm{FC}}(\boldsymbol{x}_{cont}^{(i)}), f_{\mathrm{FC}}(\boldsymbol{x}_{contloc}^{(i)})\}) \tag{6}$$

$\boldsymbol{x}_{targ}$ と $\boldsymbol{x}_{targloc}$ は，ベクトル集合 $\{\boldsymbol{x}_{cont}^{(1)}, \ldots, \boldsymbol{x}_{cont}^{(N)}\}$ および $\{\boldsymbol{x}_{contloc}^{(1)}, \ldots, \boldsymbol{x}_{contloc}^{(N)}\}$ から，判定対象とする領域のベクトルをそれぞれ抽出したものである．得られた $\boldsymbol{x}_{targ}$ と $\boldsymbol{x}_{targloc}$ に対しては，前者と同様の埋め込み処理を行う．以上の処理を数式として示す．

$$\boldsymbol{h}'_{targ} = f_{\mathrm{FC}}(\{f_{\mathrm{FC}}(\boldsymbol{x}_{targ}), f_{\mathrm{FC}}(\boldsymbol{x}_{targloc})\}) \tag{7}$$

最後に，以下のように $\boldsymbol{h}'^{(i)}_{cont}$ と $\boldsymbol{h}'_{targ}$ を連結することで，最終出力 $\boldsymbol{h}'_{imgemb}$ を得る．

$$\boldsymbol{h}'_{imgemb} = \{\boldsymbol{h}'_{targ}, \boldsymbol{h}'^{(1)}_{cont}, \ldots, \boldsymbol{h}'^{(N)}_{cont}\} \tag{8}$$

表 1: PFN-PIC と WRS-UniALT における定量的結果

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)	$\mathbf{96.9 \pm 0.34}$	$\mathbf{96.4 \pm 0.24}$

3.4 Multi-layer Transformer

Multi-layer Transformer は，(a) Multi-Head Attention 層，(b) 全結合層，(c) ドロップアウト層，(d) 正規化層，(e) 全結合層，(f) 活性化関数，(g) 全結合層，(h) ドロップアウト層，(i) 正規化層といった複数の層から構成されており，(a)–(i) を Transformer の 1 層と定義する．

入力 $\boldsymbol{h}_{trans}$ は以下の式で得られる．

$$\boldsymbol{h}_{trans} = \{\boldsymbol{h}'_{txtemb}, \boldsymbol{h}'_{imgemb}\} \tag{9}$$

はじめに，以下の式によって，query$Q^{(i)}$，key$K^{(i)}$，value$V^{(i)}$ をそれぞれ Attention の Head 数だけ生成する．

$$Q^{(i)} = W_q^{(i)} \boldsymbol{h}_{trans}^{(i)}, \tag{10}$$

$$K^{(i)} = W_k^{(i)} \boldsymbol{h}_{trans}^{(i)}, \tag{11}$$

$$V^{(i)} = W_v^{(i)} \boldsymbol{h}_{trans}^{(i)} \tag{12}$$

続いて，以下に示す Multi-Head Attention の計算式に基づき，Attention スコア S_{attn} を算出する．H は隠れ層のサイズ，A は Attention の Head 数を表す．

$$S_{attn} = \{f_{attn}^{(1)}, \ldots, f_{attn}^{(A)}\}, \tag{13}$$

$$f_{attn}^{(i)} = V^{(i)} \mathrm{softmax}(\frac{Q^{(i)} K^{(i)\top}}{\sqrt{d_k}}), \tag{14}$$

$$d_k = \frac{H}{A} \tag{15}$$

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．$\boldsymbol{h}'_{trans}$ は Multi-layer Transformer の最終層の出力を表し，$\hat{\boldsymbol{y}}$ は予測値である．

$$p(\hat{\boldsymbol{y}}) = \mathrm{softmax}(f_{\mathrm{FC}}(\boldsymbol{h}'_{trans})) \tag{16}$$

なお，損失関数には二値交差エントロピー誤差を使用する．

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 PFN-PIC [Hatori 18] と WRS-UniALT を使用した．

4.1.1 The WRS-UniALT dataset

本研究では，追加の実験用データセットとして，WRS-UniALT を作成した．WRS-UniALT は，画像および画像中の物体に関する命令文から構成されるデータセットである．画像は，World Robot Summit Partner Robot Challenge [Okada 19] の標準シミュレータにおいて生活支援ロボットによって収集されたものであり，約 5 種類の日用品を部屋の中に配置したものが写っている．命令文は，画像中の物体を対象とする物体把持文であり，6 人のアノテータによって英語で記述されている．

WRS-UniALT には，全体で 570 枚の画像と 1246 文の指示文が含まれており，語彙サイズは 167，全単語数は 8816，平均文長は 7.1 である．本実験では，後述するデータセットの前処理によって，訓練集合，検証集合，テスト集合はそれぞれ 2048，210，232 サンプルで構成されている．

4.1.2 データセットの前処理

提案手法の前処理として，Faster R-CNN [Ren 16] を使用し，データセットの各画像から複数の物体領域を抽出した．ただし，検出した各領域は，必ずしも正解領域とは一致しなかったため，Intersection over Union (IoU) β に基づき，$\beta > 0.7$ の検出領域を正解サンプル，$\beta < 0.3$ の検出領域を不正解サンプルとした．さらに，データセット内で正解と不正解のサンプル数を等量にするため，不正解のサンプル集合から正解サンプルと同じ数だけ無作為に選択し，正解のサンプル集合に加えてデータセットとした．

4.2 パラメータ設定

ネットワーク内の Transformer [Vaswani 17] は，層数が 2，隠れ層の次元数が 768，Attention の Head 数が 12 であった．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．なお，1 ステップは 1 つのバッチの処理を意味する．

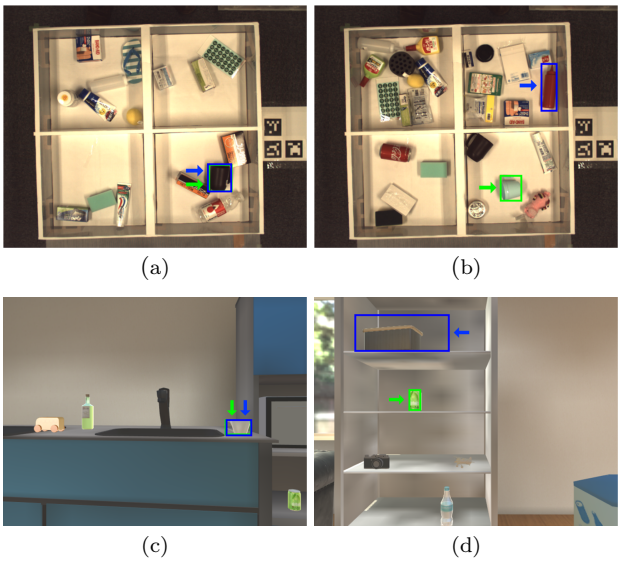
事前学習のパラメータ数は 4200 万であり，fine-tuning のパラメータ数は 3900 万である．学習にはメモリ 11GB 搭載の GeForce RTX 2080 および Intel Core i9-9900K を使用した．要した時間は，事前学習が 3 時間，fine-tuning が 1 時間であった．学習中は，2000 ステップごとに検証集合およびテスト集合による評価を行い，検証集合において最も高い精度を記録したときのテスト集合における精度を，最終的な学習の精度とした．

4.3 定量的結果

定量的結果を表 1 に示す．性能評価には精度を使用した．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．

表 1 に示すように，PFN-PIC において，提案手法は 96.9%，ベースライン手法は 90.1%を記録し，WRS-UniALT において，提案手法は 96.4%，ベースライン手法は 91.8%を記録した．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．

なお，ベースライン手法の精度に関しては，本実験で再評価した値を記載している．本実験で得られた値は，[Magassouba 19] で報告されている値よりもやや低くなっている．これは，本実験においてモデル間で公平性を期すため，Source の情報を補助出力ラベルとして使用しなかったためである．



(a) (b)

(c) (d)

図 3: PFN-PIC と WRS-UniALT における定性的結果

4.4 Ablation Studies

Ablation Study として，以下の 4 条件を定めた．

(i) FRCNN fine-tuning なし: Faster R-CNN を各データセットに fine-tuning する場合としない場合で，性能にどの程度の差が生じるかを調べる．

(ii) Late fusion: “early fusion” と “late fusion” で，性能にどの程度の差が生じるかを調べる．early fusion においては，命令文，対象物体候補の領域，画像中の各物体の領域が単一の Transformer ネットワークによって統合的に処理される．late fusion においては，対象物体候補の領域とその他の入力は別々のネットワークによって処理された後，ネットワークの最後に連結される．

(iii) Few context regions: Context Regions として入力する領域の数がそのままの場合と半分に減らす場合で，性能にどの程度の差が生じるかを調べる．

(iv) Pretraining なし: UNITER [Chen 20] の事前学習を行う場合と行わない場合で，性能にどの程度の差が生じるかを調べる．

表 1 に示すように，PFN-PIC において，条件 (i)，(ii)，(iii)，(iv) によって精度がそれぞれ 5.4%，0.9%，0.3%，0.1%減少している．また，WRS-UniALT においては，条件 (i)，(ii)，(iii)，(iv) によって精度がそれぞれ 2.4%，0.4%，0.6%，1.0%減少している．PFN-PIC においては，特に性能向上に寄与していた要素は Faster R-CNN の fine-tuning であり，次点が early fusion であった．一方，WRS-UniALT においては，特に性能向上に寄与していた要素は Faster R-CNN の fine-tuning であり，次点が事前学習であった．これより，early fusion および事前学習がモデルの性能向上に有益であったことがわかる．なお，Faster R-CNN の検出精度の向上は，本研究のスコープ外である．

4.5 定性的結果

定性的結果を図 3 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．

(a) は PFN-PIC における True Positive の例である．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，$p(\hat{\boldsymbol{y}}) = 0.999$ と出力しており，ほぼ正確に当該領域が対象領域であると判定できていることがわかる．(b) は PFN-PIC における True Negative の例である．命令文は “Grab the sky blue cup and put it in the upper right box” であり，対象物体は右下の区画にある水色のカップである．青色で囲まれている領域について，$p(\hat{\boldsymbol{y}}) = 2.37 \times 10^{-9}$ と出力しており，ほぼ正確に当該領域が対象領域ではないと判定できていることがわかる．

(a) は WRS-UniALT における True Positive の例である．命令文は “Give me the white cup” であり，対象物体はシンクの上にある白色のカップである．青色で囲まれている領域について，$p(\hat{\boldsymbol{y}}) = 0.999$ と出力しており，ほぼ正確に当該領域が対象領域であると判定できていることがわかる．(b) は WRS-UniALT における True Negative の例である．命令文は “Take the can juice on the white shelf” であり，対象物体は棚の上にある緑色の缶である．青色で囲まれている領域について，$p(\hat{\boldsymbol{y}}) = 8.19 \times 10^{-18}$ と出力しており，ほぼ正確に当該領域が対象領域ではないと判定できていることがわかる．

5. おわりに

本論文では，対象物体に関する参照表現を含む物体操作指示理解手法 Target-dependent UNITER を提案した．

提案手法による貢献は以下である．

- 物体操作指示理解分野において，UNITER 型注意機構 [Chen 20] と汎用事前学習モデルを導入した．
- UNITER において，対象物体候補を扱う新規構造を導入した．
- 2 種類のデータセットにおいて，提案手法がベースライン手法を分類精度で上回った．

将来研究では，実機ロボットにおける物体操作指示文の対象物体の特定および把持の実行が考えられる．

参考文献

[Chen 20] 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)

[Hatori 18] Hatori, J., Kikuchi, Y., Kobayashi, S., Takahashi, K., Tsuboi, Y., Unno, Y., Ko, W., and Tan, J.: Interactively picking real-world objects with unconstrained spoken language instructions, in *IEEE ICRA*, pp. 3774–3781 (2018)

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

[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 Robotics and Automation Letters*, Vol. 4, No. 4, pp. 3884–3891 (2019)

[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)

[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, in *NeurIPS*, pp. 5998–6008 (2017)