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　慶應義塾大学理工学部情報工学科

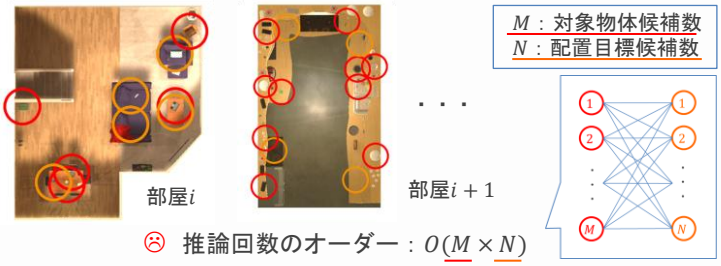
概要：自然言語指示に基づく物体操作

対象タスク	生活支援ロボットが参照表現を含む自然言語指示文に基づき，日用品を家具へ運搬するDREC-fcタスク
新規性	**Switching Head-Tail機構**の導入により，対象物体および配置目標に関する個別の予測を単一モデルで実現
結果	シミュレーション/実機実験で，ベースライン手法を上回る言語理解精度および90%以上の把持・配置動作成功率を達成

関連研究：推論回数が膨大で非実用的

MTCM [Magassouba+, RA-L19]	自然言語による指示文と全体画像を入力とし，対象物体を特定
Target-dependent UNITER (TdU) [Ishikawa+, RA-L21]	対象物体候補を扱う新規構造を導入したUNITER [Chen+, ECCV20] 型注意機構

■ 目標：指示文に対する最尤のペア（対象物体，配置目標）を探索



☹ 推論回数のオーダー：$O(M \times N)$

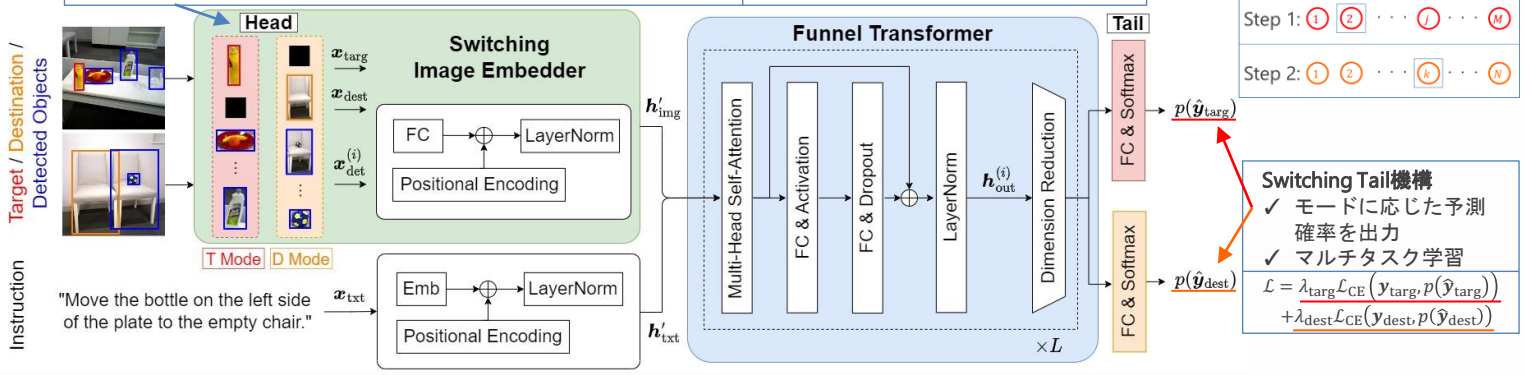
手法：Switching Head-Tail Funnel UNITER (SHeFU)

■ Switching Head-Tail機構により単一モデルで対象物体および配置目標を個別に予測可能にし，推論回数のオーダーを改善

Switching Head機構
✓ 予測対象（＝ モード）に応じて不要な入力を0埋めして条件付け

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

☺ $(= O(M+N))$



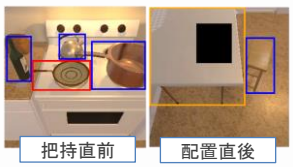
Switching Tail機構
✓ モードに応じた予測確率を出力
✓ マルチタスク学習

$$\mathcal{L} = \lambda_{\text{targ}} \mathcal{L}_{\text{CE}}\left(y_{\text{targ}}, p(\hat{y}_{\text{targ}})\right) + \lambda_{\text{dest}} \mathcal{L}_{\text{CE}}\left(y_{\text{dest}}, p(\hat{y}_{\text{dest}})\right)$$

実験設定：① シミュレーション環境で収集した新規データセット，② 標準化された家庭環境における実機実験

① ALFRED-fc：ALFRED [Shridhar+, CVPR20] を基に収集
✓ 物体操作を含むVLNタスクの標準ベンチマークにおける把持直前/配置直後のカメラ画像

サンプル数（訓練：検証：テスト）	画像	指示文
5748 (4420 : 642 : 686)	1099	3452



② 言語理解 + 把持・配置動作（= ヒューリスティック）

実験結果：既存手法を言語理解精度で上回るとともに，言語理解と動作実行が統合可能であることを実証

■ 定量的結果：言語理解精度 [%]，タスク成功率 [%]

手法	シミュレーション	実機
extended TdU [Ishikawa+, RA-L21]	79.4 ± 2.76	52.0
提案手法（W/o Switching Head）	76.9 ± 2.91	-
提案手法（W/o Switching Tail）	78.4 ± 2.05	-
提案手法（SHeFU）	**83.1 ± 2.00** +3.7	**55.9** +3.9

✓ 言語理解がTPの場合のみ把持・配置動作を実行

タスク	成功率 [%]
把持	95 (60/63)
配置	93 (56/60)

■ 定性的結果：成功例（実機）

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

[Magassouba+, RA-L19] Magassouba, A., Sugiura, K., Quoc, T. A., & Kawai, H. (2019). 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.
[Ishikawa+, RA-L21] Ishikawa, S. & Sugiura, K. (2021). Target-dependent UNITER: A Transformer-Based Multimodal Language Comprehension Model for Domestic Service Robots. IEEE RA-L, vol.6, no.4, pp.8401-8498.
[Chen+, ECCV20] Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., & Liu, J. (2020). UNITER: UNiversal Image-TExt Representation Learning. ECCV, pp.104-120.
[Shridhar+, CVPR20] Shridhar, M., Thomason, J., Gordon, D., Bisk, Y., Han, W., Mottaghi, R., Zettlemoyer, L., & Fox, D. (2020). ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. CVPR, pp.10740-10749.
[Okada+, AR19] Okada, H., Inamura, T., & Wada, K. (2019). What competitions were conducted in the service categories of the World Robot Summit? Advanced Robotics, vol.33, no.17, pp.900-910.
[Yamamoto+, ROBOMECH J.19] Yamamoto, T., Terada, K., Ochiai, A., Saito, F., Asahara, Y., & Murase, K. (2019). Development of Human Support Robot as the research platform of a domestic mobile manipulator. ROBOMECH Journal, vol.6, no.1, pp.1-15.
[Calli+, RAM15] Calli, B., Walsman, A., Singh, A., Srinivasa, S., Abbeel, P., & Dollar, A. (2015). Benchmarking in Manipulation Research: Using the Yale-CMU-Berkeley Object and Model Set. IEEE Robotics & Automation Magazine, vol.22, no.3, pp.36-52.

理工学部以外，B2以下でも研究室見学可能です．連絡先：rkorekata@keio.jp