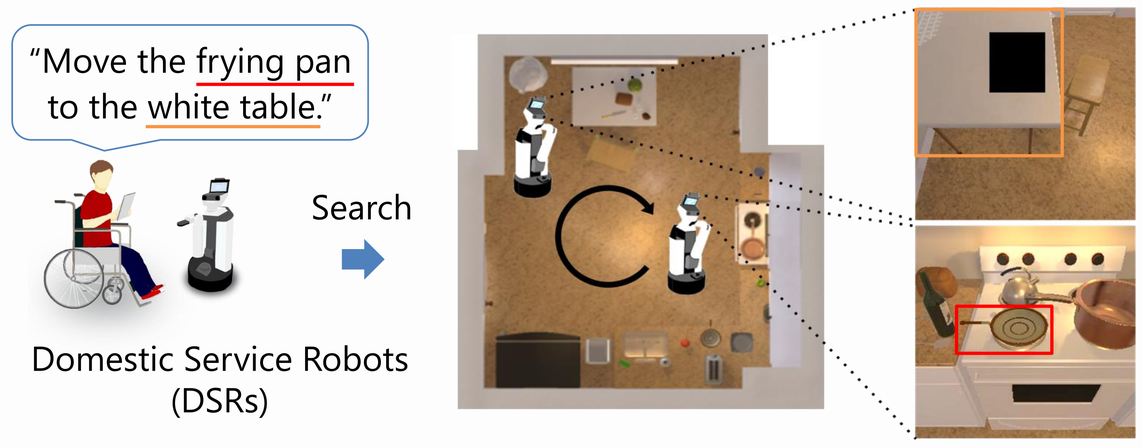
Switching Funnel UNITER: Multimodal Instruction Comprehension for Object Manipulation Tasks

Ryosuke Korekata, Yu Yoshida, Shintaro Ishikawa, and Komei Sugiura
Keio University (Japan)

IEEE/RSJ International Conference on Intelligent Robots and Systems

Abstract

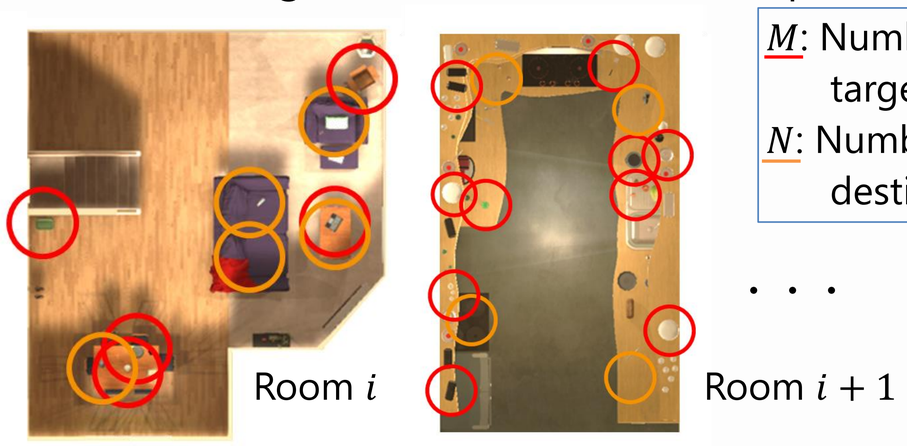
Target task	Multimodal language understanding method that comprehends object fetching and carrying instructions
Novelty	Introduce a **Switcher** module and **multi-task learning** so that both target objects and destinations can be predicted individually using a single model
Results	Outperformed the baseline method in classification accuracy on the newly-built dataset



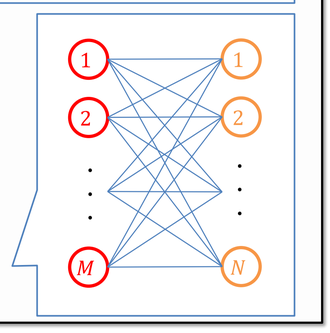
Related Work: Large Computational Cost

MTCM [Magassouba+, RA-L19]	Identifies target object from instruction and whole image
Target-dependent UNITER (TdU) [Ishikawa+, RA-L21]	Introduced the transformer attention mechanism based on UNITER [Chen+, ECCV20]

■ Goal: Finding the maximum likelihood pair



M: Number of candidate target objects
N: Number of candidate destinations



☹ Time complexity for inference: $O(M \times N)$

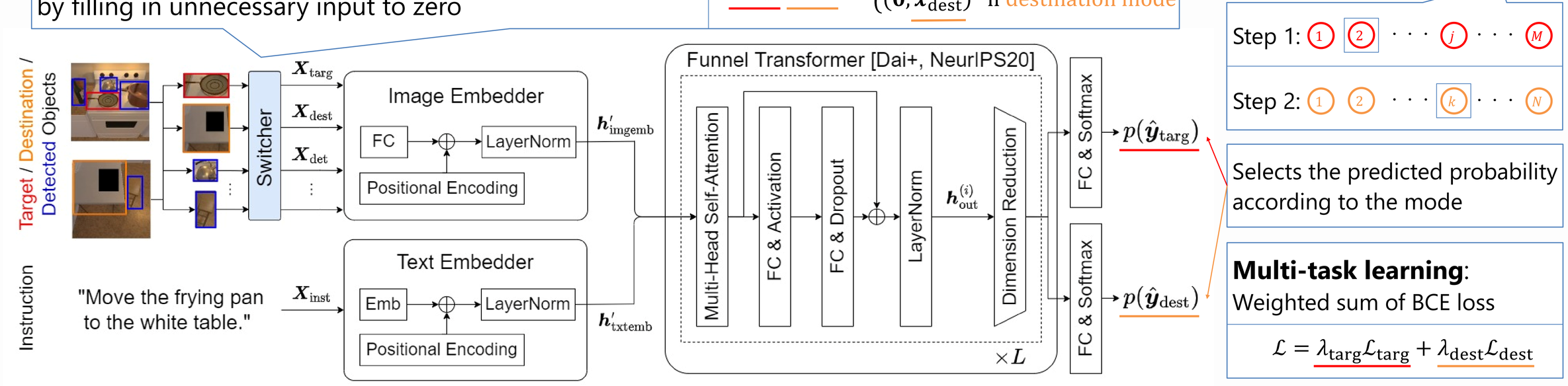
Method: Switching Funnel UNITER (SFU)

■ Both target objects and destinations can be predicted individually using a single model, which reduces the computational cost

Switcher: Conditioning on the prediction target (= mode), by filling in unnecessary input to zero

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

☺ (= $O(M+N)$)



Step 1: ① ② · · · j · · · M

Step 2: ① ② · · · k · · · N

Selects the predicted probability according to the mode

Multi-task learning:
Weighted sum of BCE loss

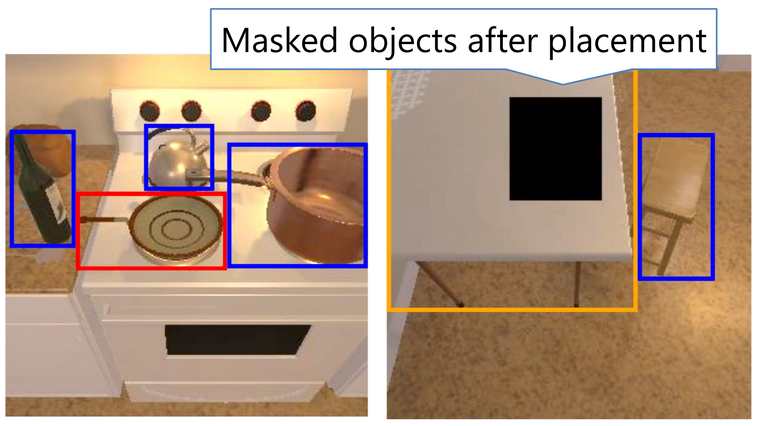
$$\mathcal{L} = \lambda_{\text{targ}}\mathcal{L}_{\text{targ}} + \lambda_{\text{dest}}\mathcal{L}_{\text{dest}}$$

Building Dataset for DREC: ALFRED-fc

■ Based on ALFRED [Shridhar+, CVPR20] (= Standard VLN benchmark)
Captured images before grasping objects and after placement

■ Task: Dual Referring Expression Comprehension (DREC)

Instruction: "Move the frying pan to the white table."



Candidate Target Object / Candidate Destination

Input:
1. Instruction for Fetch & Carry
2. Candidate Target Object Region
3. Candidate Destination Region
4. Other Object Regions

Output:
A predicted probability that both the candidate target object and the candidate destination match Ground Truth (GT)

Dataset size (train : valid : test)	# Images	# Instructions	# Unique words	# Average words
5748 (4420 : 642 : 686)	1099	3452	646	8.4

Experimental Results

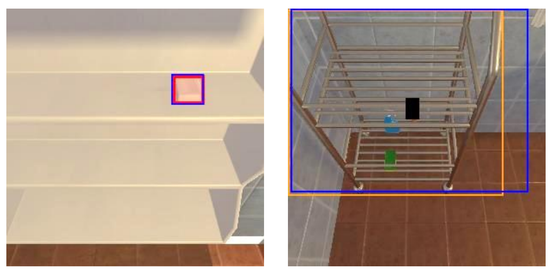
■ Quantitative Results

True Label (Boolean): $y = y_{\text{targ}} \cap y_{\text{dest}}$
Predicted Label (Boolean): $\hat{y} = \hat{y}_{\text{targ}} \cap \hat{y}_{\text{dest}}$

Method	Accuracy [%]
extended TdU [Ishikawa+, RA-L21]	79.4 ± 2.76
Ours (w/o multi-task learning)	76.9 ± 2.91
Ours (w/o zero filling in Switcher)	80.4 ± 5.31
Ours (SFU)	**83.1 ± 2.00**

+3.7

■ Qualitative Result (True Positive)



Instruction: "Move the soap from the shelves to the metal rack."

References:
[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. 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. 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.
[Dai+, NeurIPS20] Dai, Z., Lai, G., Yang, Y., & Le, Q. (2020). Funnel-Transformer: Filtering out Sequential Redundancy for Efficient Language Processing. NeurIPS, vol.33, pp.4271-4282.
[Shridhar+, CVPR20] Shridhar, M., Thomason, J., Gordon, D., et al. (2020). ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. CVPR, pp.10740-10749.