DialMAT: Dialogue-Enabled Transformer with Moment-Based Adversarial Training

Kanta Kaneda[*], Ryosuke Korekata[*], Yuiga Wada[*], Shunya Nagashima[*], Motonari Kambara,
Yui Iioka, Haruka Matsuo, Yuto Imai, Takayuki Nishimura, and Komei Sugiura
Keio University
{k.kaneda, rkorekata, yuiga, ng_sh, motonari.k714, kmngrd1805,
haruka.matsuo-25, ytim8812, t-nishimura, komei.sugiura}@keio.jp

Abstract

This paper focuses on the DialFRED task, which is the task of embodied instruction following in a setting where an agent can actively ask questions about the task. To address this task, we propose DialMAT. DialMAT introduces Moment-based Adversarial Training, which incorporates adversarial perturbations into the latent space of language, image, and action. Additionally, it introduces a crossmodal parallel feature extraction mechanism that applies foundation models to both language and image. We evaluated our model using a dataset constructed from the DialFRED dataset and demonstrated superior performance compared to the baseline method in terms of success rate and path weighted success rate. The model secured the top position in the DialFRED Challenge, which took place at the CVPR 2023 Embodied AI workshop.

1. Introduction

In this paper, we focus on the task of embodied instruction following in a setting where an agent can ask questions to the human and utilize the information provided in the response to enhance its ability to complete the task effectively.

The main contributions of this paper are as follows:

- We introduce Moment-based Adversarial Training (MAT) [4] to incorporate adversarial perturbations into the latent space of language, image, and action.
- We introduce a crossmodal parallel feature extraction mechanism to both language and image using foundation models [3, 6].

2. Problem Statement

This study focuses on the DialFRED task [1], which involves embodied instruction following. In this task, an agent has the ability to actively ask questions to the human user and utilize the information provided in the response to enhance its effectiveness in completing the task. DialFRED is based on the standard benchmark ALFRED [7] for Vision-and-Language Navigation tasks involving object manipulation. In the DialFRED setting, the robot can ask questions about the position of the object, its description, and the direction in which it should move.

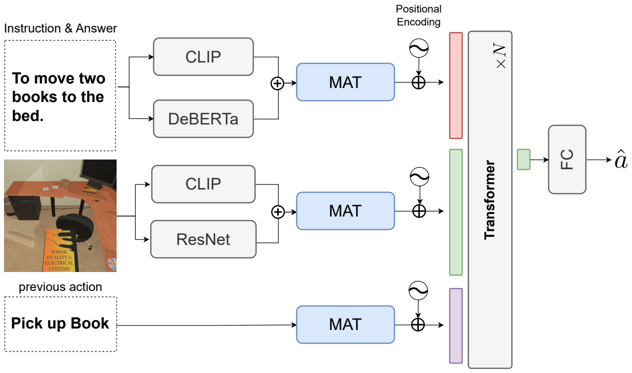

Figure 1. DialMAT consists of two main modules: Questioner and MAPer. Our method employs MAT to incorporate adversarial perturbations into the latent space of language, image, and action.

The input and output for this task are defined as follows:

- **Input:** Instructions for each subgoal, answers, and RGB images for each timestep.
- **Output:** Action taken at each timestep.

3. Method

We propose DialMAT[*], an extension of the Episodic Transformer [5]. The proposed method consists of two main modules: Questioner and Moment-based Adversarial Performer (MAPer). Fig. 1 shows the overview of DialMAT.

The input $\boldsymbol{x}_t$ for time t in our model is defined as $\boldsymbol{x}_t = (D^{(k)}, l^{(k)}, v_t, \hat{a}_{t-1}),\ D^{(k)} = \{(q_i^{(k)}, s_i^{(k)}) \mid i = 0 \ldots N\}$, where $l^{(k)}$, v_t and $\hat{a}_{t-1}$ denote the instruction for the kth subgoal, perspective RGB image at time t, and previous action at time $t-1$, respectively. Here, $\hat{a}$ is composed of a pair of action types and manipulated objects. Note that depending on the type of action, there may or may not be a

[*] Equal contribution

[*] https://github.com/keio-smilab23/DialMAT

manipulated object involved. In addition, $D^{(k)}$ refers to the set of question-answer statements in the kth subgoal, which is composed of pairs of question statements $q_i^{(k)}$ and corresponding response statements $s_i^{(k)}$.

First, the Questioner determines which question needs to be asked at time t. This module has the same structure as the Questioner proposed in [1], and is composed of an LSTM with an attention mechanism. The input in Questioner is $l^{(k)}$, and LSTM encoder and decoder perform multi-level classification. That is, at time t, it determines which of the Location, Appearance, and Direction questions to ask, and obtains the response $s^{(k)}$ for each question.

Second, the MAPer takes $\boldsymbol{x}_t$ as input and outputs the robot's behavior at time t. First, it computes the respective embedded representations h_{ctxt} and h_{deb} from $l^{(k)}$ using CLIP [6] and DeBERTa v3 [3]. Then, following the MAT approach, we obtain the features h_{txt} by adding a learnable perturbation δ_{txt} as $h_{\mathrm{txt}} = [h_{\mathrm{ctxt}};\ h_{\mathrm{deb}}] + \delta_{\mathrm{txt}}$. Note that δ_{txt} is updated based on the following steps. First, we compute the gradient of the loss function E with respect to the perturbation $\nabla_{\delta}E$. Next, we introduce two types of moving averages using $\nabla_{\delta}E$ as follows:

$$\begin{aligned} \boldsymbol{m}_t &= \rho_1 \boldsymbol{m}_{t-1} + (1-\rho_1)\nabla_{\boldsymbol{\delta}}E(\boldsymbol{\delta}_t), \\ \boldsymbol{v}_t &= \rho_2 \boldsymbol{v}_{t-1} + (1-\rho_2)(\nabla_{\boldsymbol{\delta}}E(\boldsymbol{\delta}_t))^2 \end{aligned}$$

where t denotes the current perturbation update step and ρ_1, ρ_2 denote the smoothing coefficients for each moving average. Finally, using the above $\boldsymbol{m}_i$, $\boldsymbol{v}_i$, the update width of the perturbation $\Delta\delta_i$ is calculated as follows:

$$\hat{\boldsymbol{m}}_t = \frac{\boldsymbol{m}_t}{1-(\rho_1)^t},\ \hat{\boldsymbol{v}}_t = \frac{\boldsymbol{v}_t}{1-(\rho_2)^t},\ \Delta\boldsymbol{\delta}_t = \eta\frac{\hat{\boldsymbol{m}}_t}{\sqrt{\hat{\boldsymbol{v}}_t + \varepsilon}},$$

$$\boldsymbol{\delta}_{t+1} = \Pi_{\|\boldsymbol{\delta}\|\le\epsilon}(\boldsymbol{\delta}_t + \frac{\Delta\boldsymbol{\delta}_t}{\|\Delta\boldsymbol{\delta}_t\|_F}),$$

where η, ϵ, $\Pi_{\|\cdot\|\le\epsilon}$ and $\|\cdot\|_F$ denote the learning rate of the MAT, minute value that prevents zero division, the projection onto the ϵ-sphere and the Frobenius norm, respectively. Next, we compute the respective embedded representations h_{cimg} and h_{res} from CLIP and ResNet [2]. Then, we apply MAT to obtain the feature h_{img} by adding the perturbation δ_{img} as $h_{\mathrm{img}} = [h_{\mathrm{cimg}};\ h_{\mathrm{res}}] + \delta_{\mathrm{img}}$. Similarly, for $s^{(k)}$ and a_{t-1}, MAT is applied to obtain the respective latent representations h_{ans} and h_{act}. The following embedded representation h^1 is then input to the N-layer Transformer to obtain $h^{(N)}$:

$$\begin{aligned} h^1 &= [h_{\mathrm{txt}};\ h_{\mathrm{ans}};\ h_{\mathrm{img}};\ h_{\mathrm{act}}] + E_{\mathrm{pos}}, \\ h^i &= \mathrm{Transformer}(h^{i-1}), \end{aligned}$$

where E_{pos} refers to positional encoding and embeds the positional information of each token in $[h_{\mathrm{txt}};\ h_{\mathrm{ans}};\ h_{\mathrm{img}};\ h_{\mathrm{act}}]$. Finally, we obtain the predicted action $\hat{a}_t$ by applying the fully connected layer to $h^{(N)}$. We define the loss function as the cross-entropy between $\hat{a}_t$ and expert action a_t.

Table 1. Quantitative comparison. The best scores are in bold.

	Pseudo Test		Test
Method	SR↑	PWSR↑	SR↑
Baseline [1]	0.31	0.19	-
Ours (w/o MAT [4])	0.34	0.20	-
Ours (w/ CLIP text encoder [6])	0.35	0.22	-
Ours (MAT for action)	0.36	0.21	-
Ours (DialMAT)	**0.39**	**0.23**	**0.14**

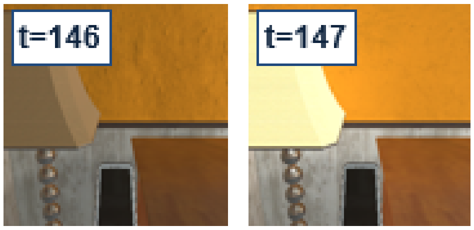

Figure 2. A successful subgoal prediction. In this case, the instruction was "Move to the floor lamp, power on the floor lamp."

4. Experimental Setup

In this study, we employ the same experimental settings as DialFRED [1]. In this study, we divided the DialFRED dataset as described in [1]. As the test set of DialFRED is not publicly accessible, we further divided the valid_unseen set to pseudo_valid and pseudo_test set, comprising 685 and 678 tasks, respectively. We used the training set to train the MAPer and the valid_seen set for reinforcement learning of the Questioner.

5. Experimental Results

Table 1 shows the quantitative results of the baseline and the proposed methods. We evaluated the models by success rate (SR) and path weighted success rate (PWSR). Table 1 shows that the SR on the pseudo_test set for the baseline and the proposed methods were 0.31 and 0.39, respectively. Therefore, the proposed method outperformed the baseline by 0.08 points in SR. Similarly, the proposed method also outperformed the baseline method in PWSR.

Fig. 2 shows the qualitative results. In this sample, the instruction was "Move to the floor lamp, power on the floor lamp." In this case, the robot was required to move to the floor lamp and then turn it on. The proposed method was able to appropriately navigate to the floor lamp and successfully turn it on.

6. Conclusions

In summary, the contributions of this work are twofold:

- We introduced MAT to incorporate adversarial perturbations into the latent spaces.
- We introduced crossmodal parallel feature extraction mechanisms to both language and image using foundation models.

Acknowledgements This work was partially supported by JSPS KAKENHI Grant Number JP23H03478, JST Moonshot, and NEDO.

References

[1] Xiaofeng Gao, Qiaozi Gao, Ran Gong, Kaixiang Lin, et al. DialFRED: Dialogue-Enabled Agents for Embodied Instruction Following. *IEEE RA-L*, 7(4):10049–10056, 2022. 1, 2

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, pages 770–778, 2016. 2

[3] Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. In *ICLR*, 2023. 1, 2

[4] Shintaro Ishikawa and Komei Sugiura. Moment-based Adversarial Training for Embodied Language Comprehension. In *IEEE ICPR*, pages 4139–4145, 2022. 1, 2

[5] Alexander Pashevich, Cordelia Schmid, and Chen Sun. Episodic Transformer for Vision-and-Language Navigation. In *ICCV*, pages 15942–15952, 2021. 1

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, pages 8748–8763, 2021. 1, 2

[7] Mohit Shridhar, Jesse Thomason, Daniel Gordon, et al. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In *CVPR*, pages 10740–10749, 2020. 1