Winning Team of
DialFRED Challenge

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

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

* Equal Contribution

Background:
Multimodal language processing for robots

Motivation:

- Multimodal language understanding & generation methods for domestic service robots (DSRs)
  - Focus of DialFRED [Gao+ RAL22] := instruction following with QAs

Related work

Grounded communication with robots	RoboCup@Home [Iocchi+ Artificial Intelligence15], World Robot Summit [Okada+ Advanced Robotics19]
Representative methods for ALFRED	Prompter [Inoue+ 22], FILM [Min+ ICLR22], HLSM-MAT [Ishikawa+ ICPR22], E.T. [Pashevich+ ICCV21]
Object navigation with dialogue	DialFRED [Gao+ RAL22], TEACh [Padmakumar+ AAAI22], Vision-and-Dialog Navigation [Thomason+ CoRL19]

RoboCup@Home (2007-)

World Robot Summit (2018-)

REVERIE [Qi+ CVPR20]

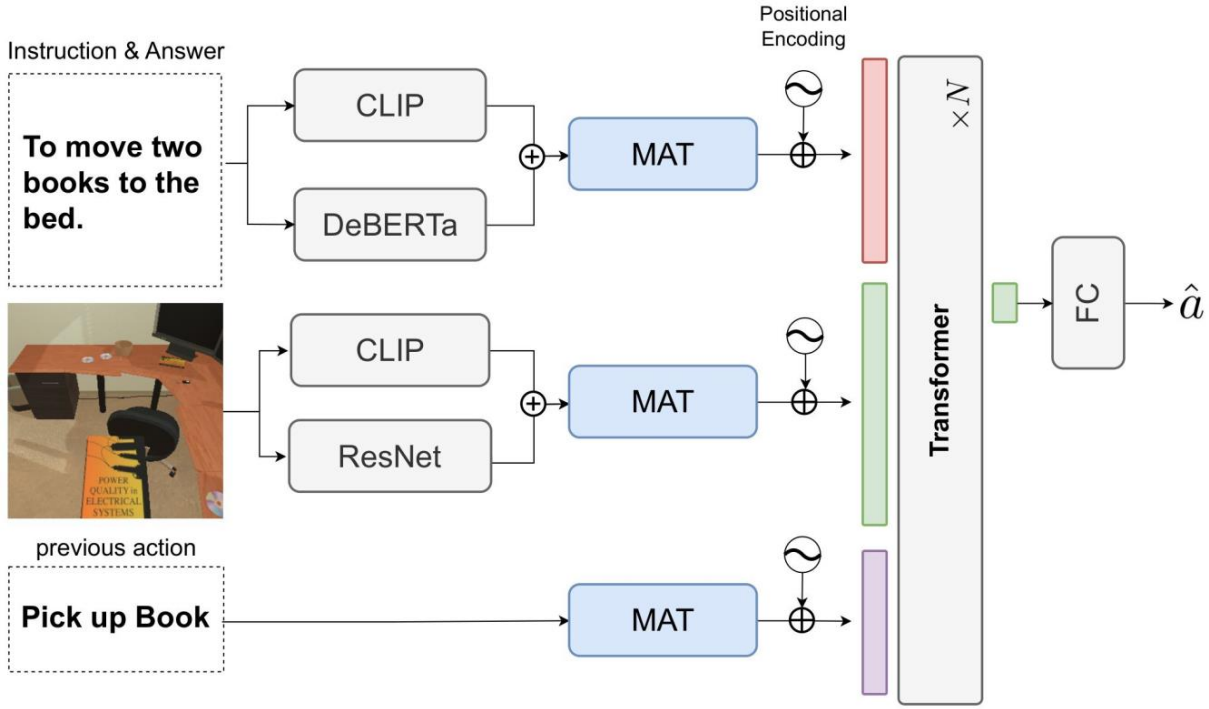
Our method: DialMAT

①Moment-based Adversarial Training (MAT) [Ishikawa+, ICPR22]

- Add adversarial perturbation to embedding spaces

②Crossmodal parallel feature extraction mechanism

- CLIP_{txt} and DeBERTa v3
- CLIP_{img} and ResNet



Moment-based Adversarial Training (MAT) [Ishikawa+ ICPR22]: Add adversarial perturbation to the embedding spaces

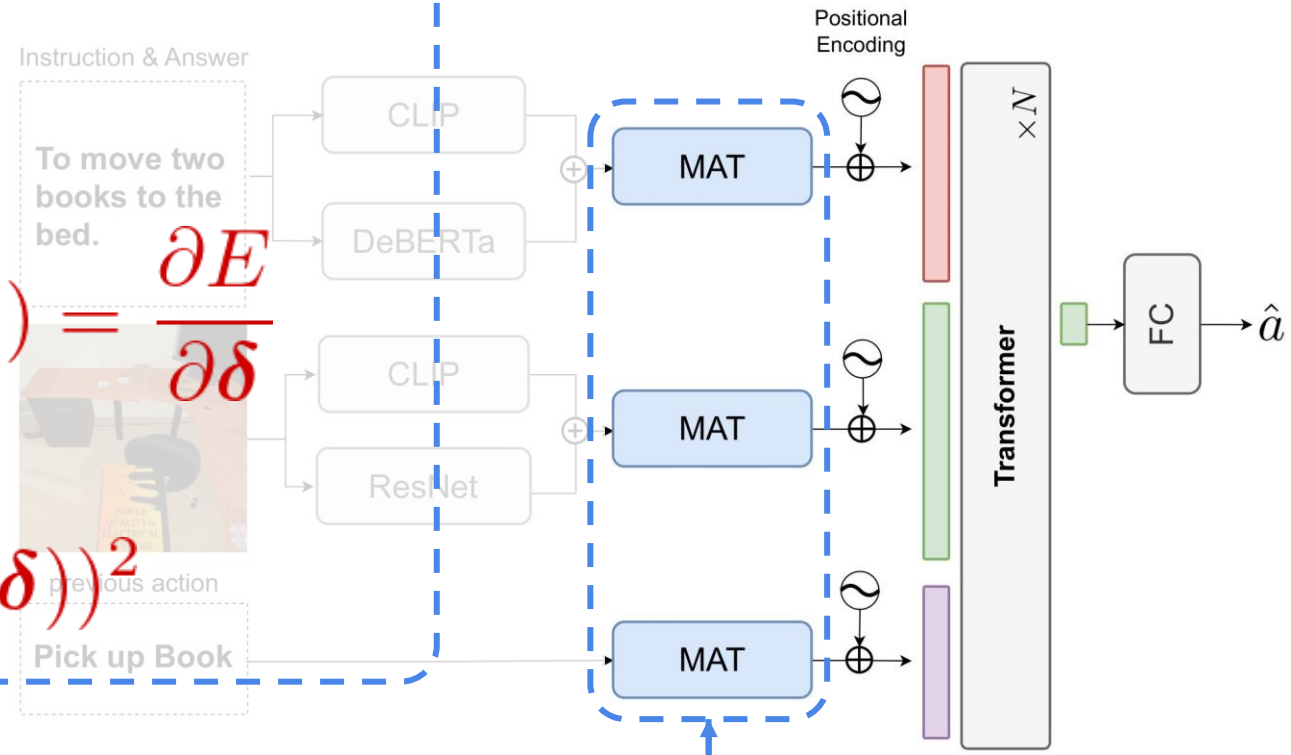
Update adversarial perturbation $\boldsymbol{\delta}$ (cf. VILLA [Gan+ NeurIPS20])

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

$$\Delta \boldsymbol{\delta}_t = \eta \frac{\hat{\boldsymbol{m}}_t}{\sqrt{\hat{\boldsymbol{v}}_t + \varepsilon}}$$

EMA of $\nabla_{\boldsymbol{\delta}} E(\boldsymbol{\delta}) = \frac{\partial E}{\partial \boldsymbol{\delta}}$

EMA of $(\nabla_{\boldsymbol{\delta}} E(\boldsymbol{\delta}))^2$



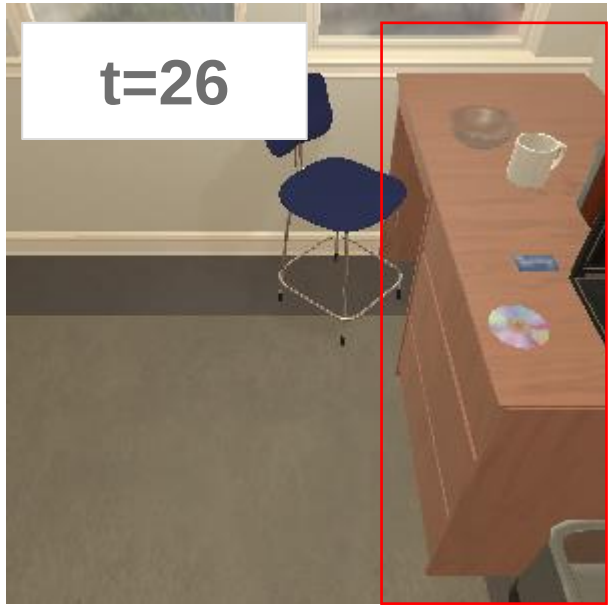
EMA: Exponential Moving Average

Quantitative Results: Outperformed the baseline method in terms of SR

Method	Pseudo_Test SR↑	Pseudo_Test PWSR↑	Test SR↑
Baseline [Gao+, RA-L22]	0.31	0.19	-
Ours (w/o MAT)	0.34	0.20	-
Ours (w/ CLIP text encoder)	0.35	0.22	-
Ours (MAT for action)	0.36	0.21	-
Ours (DialMAT)	**0.39**	**0.23**	**0.14**

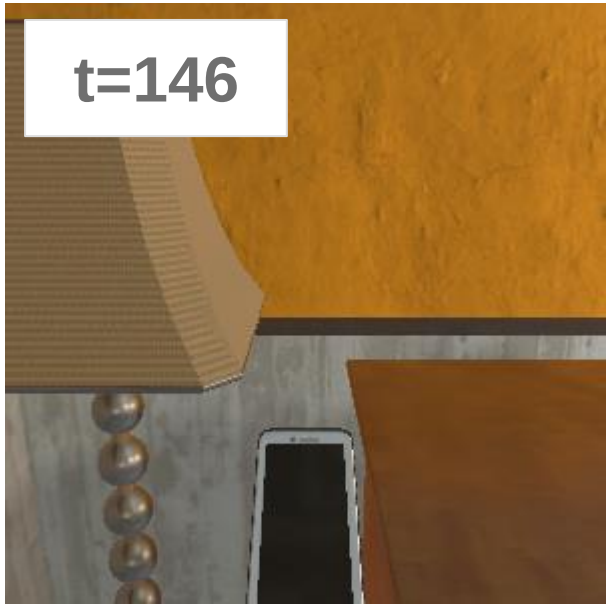
- Divided the valid_unseen set of DialFRED dataset
  - (pseudo_valid : pseudo_test) = (1,296 : 1,363) scenes
- Metrics: **Success Rate (SR), Path Weighted Success Rate (PWSR)**

Qualitative Results:
Output appropriate actions to successfully complete the task



Instruction: “Move to the **desk**”

☺ Navigated to the specified desk



Instruction: “Move to the floorlamp power on the floorlamp”

☺ Navigated to the appropriate location and executed the appropriate action

Conclusions

- Introduced MAT to incorporate adversarial perturbations into the latent spaces of language, image, and action
- Introduced a crossmodal parallel feature extraction mechanisms to both language and image using foundation models

Our code is available here

Winning Team of
DialFRED Challenge

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

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

* Equal Contribution