DialMAT: 敵対的摂動に基づく対話的 Vision-and-Language Navigation

○是方諒介，和田唯我，兼田寛大，長嶋隼矢，杉浦孔明（慶應義塾大学）

高齢化が進行する現代社会における在宅介助者不足に対して，対話的な指示文をもとに家事タスクを実行可能な生活支援ロボットは利便性が高い．しかし，物体操作を含む対話的 Vision-and-Language Navigation (VLN) を扱う既存手法は特徴量抽出が不十分であり，また未知の環境における頑健性に欠ける．本論文では，言語特徴量，画像特徴量，行動特徴量に敵対的摂動を加える Moment-based Adversarial Training および基盤モデルを用いた並列クロスモーダル特徴抽出機構を導入する DialMAT を提案する．実験の結果，対話的 VLN の標準ベンチマークにおいて提案手法がベースライン手法を成功率で上回った．

1. はじめに

少子高齢化が進行する現代社会において，在宅介助者不足が社会問題となっている．その一つの解決策として，被介助者を物理的に支援することが可能な生活支援ロボットに注目が集まっている．しかしながら，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，家事タスクに関する自然言語指示文に基づき，対話的にロボットの移動および物体操作を行う DialFRED タスク [1] に取り組む．例えば，“Pick up the knife.” という指示文が与えられた場合に，ロボットは対象物体に関する位置，形容，および移動すべき方向に関する質問（例: “Where is the knife?”）をユーザに行う．質問に対する回答から得られた言語情報（例: “The knife is to your left.”）を扱うことで，ロボットは効率的に指示文を環境に接地することが期待される．

近年の研究（例: [2]）により Vision-and-Language Navigation (VLN) モデルの性能は向上しているが，対話を扱う手法は少なく性能は未だ不十分である．例えば [1] で提案されたモデルは，特徴量抽出が不十分であり，未知の環境における頑健性に欠ける．

本研究では，Moment-based Adversarial Training (MAT) [3] および基盤モデル [4,5] を用いて DialFRED タスクを扱う手法 DialMAT[1] を提案する．特徴量空間に敵対的摂動を加える MAT を導入することで，様々な環境に対応できることが期待される．

2. 関連研究

VLN の研究は盛んに行われている [2, 3, 6–8]．VLN に関するサーベイ論文である [9] では，VLN における網羅的な調査を行っており，命令文の種類に応じたタスクの分類を試みている．

VLN における代表的なベンチマークとして ALFRED [10] が挙げられる．ALFRED は物体操作を含む VLN における標準的なベンチマークであり，多くの研究が行われている．HLSM-MAT [3] は ALFRED ベンチマークにおいて，サブゴールの生成過程に敵対的学習を導入し，様々なシナリオを扱うモデルである．敵対的学習により，汎用的なサブゴールの予測が実現し，様々な環境においてタスクが遂行可能となる．

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

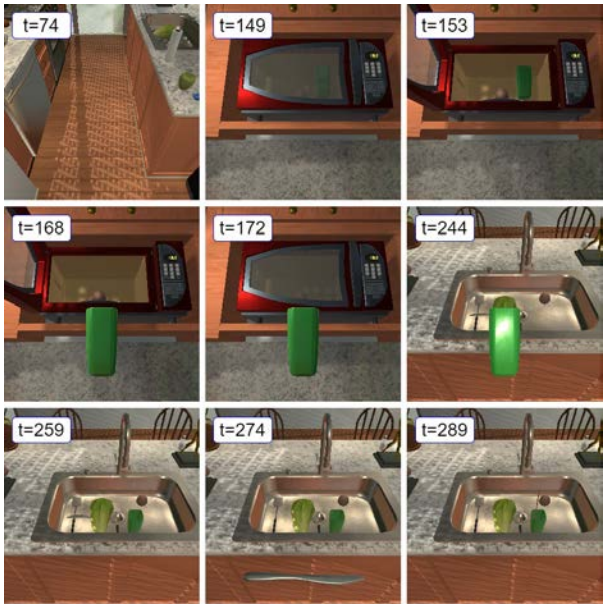

図 1　DialFRED タスクの例

3. 問題設定

本研究では，家事タスクに関する自然言語指示文に基づきロボットの移動および物体操作を実行する際に，ロボットとユーザが対話的に曖昧性解消を行う DialFRED タスク [1] を扱う．DialFRED は，物体操作を含む VLN における標準ベンチマーク ALFRED [10] に基づいている．DialFRED では，対話的な VLN が可能であり，ロボットは対象物体の位置，形容，および移動すべき方向について質問することができる．また，DialFRED は 25 個のサブゴールによって構成されており，最大 7 個のサブゴールで構成される ALFRED よりも困難なタスクである．本タスクでは，各時刻において適切な行動を選択し，最終的に複数の階層的なサブゴールから構成される各ゴール条件を達成することが望ましい．

本タスクにおける入出力は以下の通りである．

- 入力: サブゴール毎の指示文，回答文，各時刻におけるカメラ画像
- 出力: 各時刻における行動

また，評価指標には [1] と同様に success rate (SR) および path weighted success rate (PWSR) を用いる．

図 1 に，本タスクの例を示す．本サンプルでは，電子レンジからカップを取り，シンクに置いてその中にナイフを入れるという動作を，サブゴール毎に与えられる指示文（“Move to the microwave, open the microwave.”, “Grab the cup.”, “Close the microwave.”, “Place in the sinkbasin.”, “Pick up the butterknife.”, “Leave in the cup.”）をもとに実行することが求められる．

4. 提案手法

本論文では Episodic Transformer [11] を拡張した DialMAT を提案する．図 2 に，提案手法のモデル構造を示す．本提案手法は 2 つのモジュールから構成され，それぞれ Questioner および Moment-based Adversarial Performer (MAPer) である．

4.1 入力

時刻 t における行動予測について，モデルの入力 $\boldsymbol{x}_t$ を以下のように定義する．

$$\boldsymbol{x}_t = (D^{(k)}, l^{(k)}, v_t, \hat{a}_{t-1})$$
$$D^{(k)} = \{(q_i^{(k)}, s_i^{(k)}) \mid i = 0 \dots N\}$$

$l^{(k)}$ は k 番目のサブゴールにおける指示文，v_t は時刻 t におけるロボットのカメラ画像，また $\hat{a}_{t-1}$ は時刻 $t-1$ における過去の行動を表す．ここで，$\hat{a}$ は行動の種類と操作物体の組によって構成され，行動の種類により，操作物体を扱う場合と扱わない場合とが存在する．また，$D^{(k)}$ は k 番目のサブゴールにおける質問応答文の集合を指し，質問文 $q_i^{(k)}$ と応答文 $s_i^{(k)}$ の組で構成される．

4.2 Questioner

Questioner では，時刻 t においてどの質問を行う必要があるかを判定する．本モジュールは [1] で提案された Questioner と同一の構造であり，attention 機構を有する LSTM によって構成される．Questioner における入力は $l^{(k)}$ であり，LSTM encoder および LSTM decoder によって多値分類を行う．すなわち，時刻 t において位置，形容，および移動すべき方向のうちどの質問を行うかを判定し，質問の応答 $s^{(k)}$ を得る．

4.3 Moment-based Adversarial Performer (MAPer)

MAPer では，x_t を入力として，時刻 t におけるロボットの行動を出力する．まず，CLIP [4] および DeBERTa v3 [5] より，$l^{(k)}$ からそれぞれの埋め込み表現 $h_{\text{ctxt}}, h_{\text{deb}}$ を計算する．その後，MAT [3] に従い，学習可能な摂動 δ_{txt} を加えた特徴量 h_{txt} を得る．

$$h_{\text{txt}} = [h_{\text{ctxt}};\ h_{\text{deb}}] + \delta_{\text{txt}}$$

ここで，δ_{txt} は次に示す更新則に基づき更新される．はじめに，摂動 δ に関する損失関数 E の勾配 $\nabla_\delta E$ を計算する．次に $\nabla_\delta E$ を用いて，以下に示す 2 種類の移動平均を導入する．

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

ここで，t は現在の摂動更新ステップ，ρ_1, ρ_2 は各移動平均の平滑化係数を表す．最終的に，上記 $\boldsymbol{m}_i, \boldsymbol{v}_i$ を用いて，摂動の更新幅 $\Delta\delta_i$ を以下のように計算し，摂動を更新する．

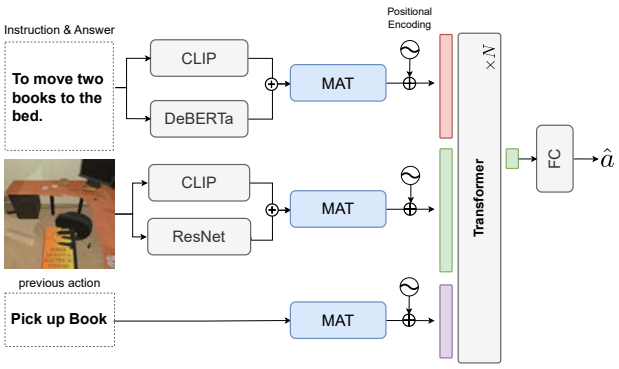

図 2　提案手法のモデル図

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

ここで，η は MAT の学習率，ϵ はゼロ除算を防ぐ微小な値，$\Pi_{\|\cdot\|\le\epsilon}$ は ϵ 球への投影，および $\|\cdot\|_F$ はフロベニウスノルムを表す．

次に，CLIP および ResNet [12] より，v_t からそれぞれの埋め込み表現 h_{cimg} および h_{res} を計算後，MAT を適用し特徴量 h_{img} を得る．

$$h_{\text{img}} = [h_{\text{cimg}};\ h_{\text{res}}] + \delta_{\text{img}}$$

同様に，$s^{(k)}, a_{t-1}$ においても MAT を適用し，それぞれ潜在表現 $h_{\text{ans}}, h_{\text{act}}$ を得る．続いて，次に示す埋め込み表現 h^1 を N 層の transformer に入力し $h^{(N)}$ を得る．

$$h^1 = [h_{\text{txt}};\ h_{\text{ans}};\ h_{\text{img}};\ h_{\text{act}}] + E_{\text{pos}}$$
$$h^i = \text{Transformer}(h^{i-1})$$

ここで，E_{pos} は Positional Encoding を指し，$[h_{\text{txt}};\ h_{\text{ans}};\ h_{\text{img}};\ h_{\text{act}}]$ における各トークンの位置情報を埋め込む．最終的に，$h^{(N)}$ に FC 層を適用することで，行動の予測 $\hat{a}_t$ を出力する．なお，損失関数は $\hat{a}_t$ と真の行動 a_t との交差エントロピーとする．

5. 実験設定

本研究では，DialFRED [1] と同様の実験設定を扱う．DialFRED は物体操作を含む対話的な VLN における標準的なベンチマークであるため，本研究で扱う．DialFRED は ALFRED [10] を拡張したデータセットであり，25,743 文の指示文と，人間により付与された 53,000 個の質問応答文が含まれている．また，ALFRED は 8 個のサブタスクで構成されているのに対し，DialFRED は 25 個のサブタスクで構成されている．

本研究では，[1] に従ってデータセットを分割した．ここで，訓練集合と検証集合は，それぞれ 34,253 個および 2,659 個のタスクで構成される．また，検証集合およびテスト集合はそれぞれ Seen 集合と Unseen 集合に分けられる．検証集合は 1,296 個のタスクを含む Seen 集合と，1,363 個のタスクを含む Unseen 集合で構成されている．ここで，DialFRED のテスト集合はデータが公開されていないため，本研究では検証集合における Unseen 集合を再度分割し，疑似検証集合および，疑似テスト集合へと分割した．本研究では，訓練集合を

表 1　定量的結果．太字は最良の値を示す．

[%]	手法	条件 w/ MAT (act)	w/ MAT (img)	w/ MAT (txt)	疑似テスト集合 SR↑	PWSR↑	テスト集合 SR↑
(a)	ベースライン手法 [1]				0.31	0.19	-
(b)					0.34	0.20	-
(c)	提案手法 (DialMAT)	✓			0.36	0.21	-
(d)		✓	✓	✓	**0.39**	**0.23**	**0.14**

MAPer の訓練に使用し，検証集合における Seen 集合を Questioner の強化学習に用いた．また，疑似検証集合をハイパーパラメータの評価に使用し，疑似テスト集合をモデルの評価に用いた．

transformer において，層数 $N = 4$，attention head 数 $A = 8$ と設定した．また，最適化手法として AdamW ($\beta_1 = 0.9, \beta_2 = 0.999$) を採用し，エポック数，学習率，およびバッチサイズをそれぞれ 20，1.0×10^{-5}，および 4 と設定した．提案手法の訓練可能パラメータ数および積和演算数は，それぞれ 120M および 54.9G である．モデルの訓練にはメモリ 24GB 搭載の GeForce RTX 3090 および Intel Core i9-10900KF を使用した．本モデルの訓練時間および 1 サンプルあたりの推論時間は，それぞれ 7.7 時間および 19.8 秒であった．

6.　実験結果

6.1　定量的結果

表 1 に，ベースライン手法，提案手法，および ablation study の定量的結果を示す．[1] において提案されている手法をベースライン手法とした．評価指標として，[1] と同様に SR および PWSR を用いた．本論文では，主要評価指標を SR とした．また，以下の 2 つの ablation 条件を定めた．

(b) MAT モジュールを取り除く場合，性能にどの程度の差が生じるかを調査した．

(c) 行動の潜在空間にのみ MAT モジュールを適用する場合，性能にどの程度の差が生じるかを調査した．

表 1 より，疑似テスト集合におけるベースライン手法 (a) と提案手法 (d) の SR はそれぞれ 0.31 と 0.39 であった．そのため，提案手法は疑似テスト集合の SR においてベースライン手法を 0.08 ポイント上回った．同様に，提案手法は PWSR においてもベースライン手法を上回った．

6.2　定性的結果

図 3 に定性的結果を示す．図 3(a) において，指示文は “Move to the desk” であった．また，指定された “desk” は CD ディスクの置いてある机であった．この例において，ロボットは環境に複数ある机のうち，指定された机を特定し，移動することが求められる．提案手法は，エージェントから獲得した言語情報を基に，移動中に発見した別の机を正解の机ではないと認識しつつ，最終的に指定された机へ移動することができた．

同様に，図 3(b) において，指示文は “Move to the floorlamp, power on the floorlamp” であった．この例において，ロボットは，フロアランプまで移動した上で，ランプを点灯させることが求められる．提案手法は，適切にフロアランプまで移動し，ランプを点灯させることができた．

図 3　定性的結果

7.　おわりに

本研究では，自然言語指示文に基づく物体操作において，ロボットとユーザが対話的に曖昧性解消を行う DialFRED タスク [1] を扱った．貢献は以下である．

- 言語，画像，行動の潜在空間に敵対的摂動を組み込むため，MAT [3] を導入した．
- 基盤モデル [4, 5] を用いて言語と画像の両方にクロスモーダルな並列特徴抽出機構を導入した．

謝辞

本研究の一部は，JSPS 科研費 23H03478，JST ムーンショット，NEDO の助成を受けて実施されたものである．

参　考　文　献

[1] X. Gao, Q. Gao, R. Gong, K. Lin, et al., “DialFRED: Dialogue-Enabled Agents for Embodied Instruction Following,” IEEE RA-L, vol.7, no.4, pp.10049–10056, 2022.
[2] Y. Inoue, et al., “Prompter: Utilizing Large Language Model Prompting for a Data Efficient Embodied Instruction Following,” arXiv preprint arXiv:2211.03267, 2022.
[3] S. Ishikawa and K. Sugiura, “Moment-based Adversarial Training for Embodied Language Comprehension,” ICPR, pp.4139–4145, 2022.
[4] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, et al., “Learning Transferable Visual Models From Natural Language Supervision,” ICML, pp.8748–8763, 2021.
[5] P. He, J. Gao, and W. Chen, “DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing,” ICLR, 2023.
[6] P. Anderson, Q. Wu, et al., “Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments,” CVPR, pp.3674–3683, 2018.
[7] A. Magassouba, K. Sugiura, and H. Kawai, “CrossMap Transformer: A Crossmodal Masked Path Transformer Using Double Back-Translation for Vision-and-Language Navigation,” IEEE RA-L, vol.6, no.4, pp.6258–6265, 2021.
[8] S. Ishikawa and K. Sugiura, “Target-dependent UNITER: A Transformer-Based Multimodal Language Comprehension Model for Domestic Service Robots,” IEEE RA-L, vol.6, no.4, pp.8401–8408, 2021.
[9] W. Wu, T. Chang, and X. Li, “Visual-and-Language Navigation: A Survey and Taxonomy,” arXiv preprint arXiv:2108.11544, 2021.
[10] M. Shridhar, J. Thomason, D. Gordon, et al., “ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks,” CVPR, pp.10740–10749, 2020.
[11] A. Pashevich, C. Schmid, and C. Sun, “Episodic Transformer for Vision-and-Language Navigation,” ICCV, pp.15942–15952, 2021.
[12] K. He, X. Zhang, S. Ren, et al., “Deep Residual Learning for Image Recognition,” CVPR, pp.770–778, 2016.