Vision-and-Language Navigation タスクにおける 敵対的サブゴール生成

Generating Subgoals with Adversarial Network on Vision-and-Language Navigation

石川 慎太朗 *1　杉浦 孔明 *1
Shintaro Ishikawa　Komei Sugiura

*1慶應義塾大学
Keio University

In this paper, we focus on a vision-and-language task in which a robot is instructed to execute household tasks. We propose Moment-based Adversarial Training (MAT), which uses two types of moments for perturbation updates in adversarial training. We introduce MAT to the embedding spaces of the instruction, subgoals, and state representations to handle their varieties. We validated our method on the ALFRED benchmark, and the results demonstrated that our method outperformed the baseline method for all the metrics on the benchmark.

1. はじめに

高齢化が進行する現代社会において，日常生活における介護・支援の必要性が高まっている．それに伴って，在宅介護者の不足が社会問題となっており，障がいを持つ人々を物理的に支援可能な生活支援ロボットに注目が集まっている．一方，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，自然言語の命令文を適切な行動に対応付けるモデルの開発に取り組む．具体的には，“Rinse off a mug and place it in the coffee maker” という命令が与えられたときに，ロボットがマグカップを発見し，水で洗浄した後，コーヒーメーカーのもとに運ぶといった動作が望ましい．

人間は家事タスクを指定する際に，高レベルな命令文を使用する場合が多く，ロボットはそれらを適切なサブゴールに分解することが求められる．例えば，前述のタスクは大きく 3 つのサブゴールに分けられ，それらは適切な順序で実行される必要がある．人間はこのような階層的なタスクを容易に実行することが可能だが，ロボットにとっては挑戦的な課題である．ALFRED ベンチマーク [Shridhar 20] においても，人間は 91.0%の精度を記録している一方で，最先端の手法 [Blukis 21] の精度は 25%を下回る．

既存手法では，階層的な構造をもとにサブゴールを生成することで，高レベルな命令文の解釈に取り組んでいる．しかし，そのような命令文においてサブゴールは明示的に指定されないため，しばしば予測に失敗してしまう．特にロボットが未知の環境においては，タスクの失敗が顕著に表れている．

本研究では，サブゴールの生成過程に敵対的学習を導入することで，様々なシナリオを扱うことができるモデルを開発する．これにより，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になると考えられる．既存手法と異なる点は，VILLA [Gan 20] に基づく敵対的な摂動をサブゴールおよび状態表現の特徴量空間に加えている点である．

本研究の貢献を以下に示す．

- Vision-and-Language Navigation (VLN) において，サブゴールおよび状態表現の特徴量空間に対し，敵対的な摂動を導入する．

連絡先: 石川慎太朗，慶應義塾大学，神奈川県横浜市港北区日吉 3-14-1，shin.0116@keio.jp

図 1: ALFRED の例

- 敵対的摂動更新アルゴリズム Momentum-based Adversarial Training (MAT) を提案する．

2. 関連研究

VLN に代表されるロボットと自然言語の分野においては，現在に至るまで多くの研究が行われている [Anderson 18, Kollar 10, Tellex 11, Kuo 20]．[Wu 21] は，VLN に関して包括的な調査を実施しており，命令文の種類に応じたタスクの分類を試みている．また，近年の VLN では，Transformer の注意機構に基づいたモデルが高い性能を記録している．MTCM [Magassouba 19] は，参照表現を含む命令文および画像をもとに，対象物体を特定するモデルである．Target-dependent UNITER [Ishikawa 21] は，Transformer の注意機構を利用して，命令文および物体間の関係性をモデル化している．

VLN は Single-turn と Multi-turn の 2 種類のタスクに分類される．Single-turn のタスクにおいては，ロボットが探索を開始する前に一連の命令文が与えられる．このタスクは，命令文に経路に関する情報が含まれるか否かによって，さらに Goal-orientation と Route-orientation の 2 種類に分けられる．前者の場合は，命令文によって複数のゴール条件が与えられるものの，経路に関する情報は一切与えられない．一方，後者の場合は，特定の経路に関する情報が与えられる．CrossMap Transformer [Magassouba 21] は，命令文およびカメラ画像をもとに順次経路を生成するモデルであり，Goal-oriented な命令文を入力として Single-turn のタスクを実行する．

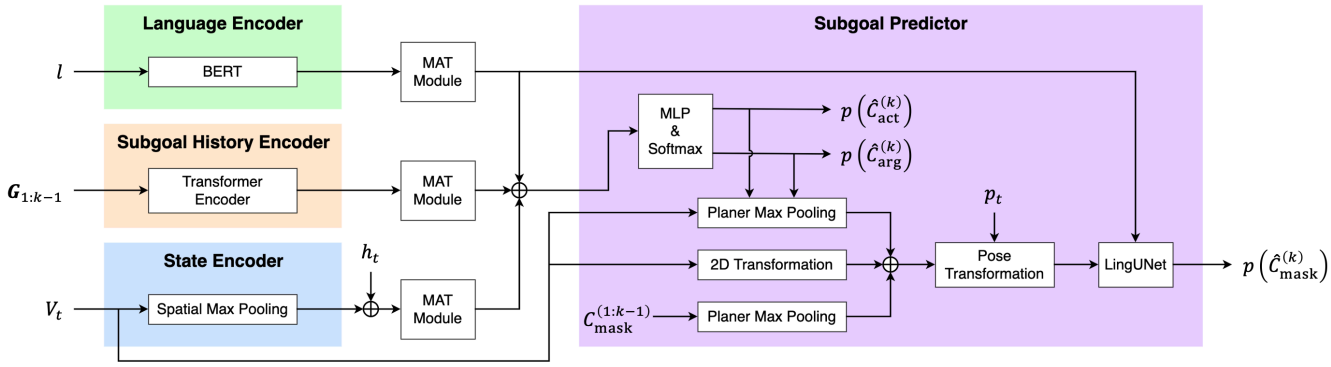

図 2: 提案手法のネットワーク構造

Multi-turn のタスクにおいては，複数の段階にわたってガイドからナビゲーターへ命令文が与えられる．このタスクは，ナビゲーターがガイドと対話できるか否かによって，さらに Imperative と Interactive の 2 種類に分けられる．前者の場合は，ナビゲーターからガイドへ応答することは許されていない．後者の場合は，ガイドとナビゲーターは互いにやり取りを行いつつ情報を共有することが可能である．CVDN [Thomason 20] は，家庭環境における複数の人間の対話履歴に関して，Multi-turn かつ Interactive なタスクを提供している．

3. 問題設定

本研究では，シミュレーション環境において，自然言語の命令文を適切な行動に対応付けるモデルを開発することを目的とする．我々は ALFRED ベンチマーク [Shridhar 20] を使用する．ALFRED は，家庭環境における長期的かつ階層的なタスクを含むデータセットを提供している．モデルにおいては，各時刻において適切な行動を出力しつつ，最終的に複数のサブゴールから構成されるゴール条件を達成することが望ましい．図 1 に ALFRED の例を示す．このエピソードには，“Put a clean apple on a wooden table” という命令文が付与されている．ロボットは，はじめに電子レンジのもとに移動し，リンゴを中から取り出した後，シンクへと運ぶ．そして，リンゴを洗った後，木製の机のもとに運び，そこに置く必要がある．

本タスクの入出力を以下のように定義する．

- **入力:** 高レベルな命令文，各時刻におけるカメラ画像
- **出力:** 各時刻における各行動に関する確率の予測値

また，本論文で使用する用語を以下のように定義する．

- **高レベルな命令文:** タスク全体を通して，ロボットが達成すべき自然言語の命令文．ロボットは，この命令文を複数の暗黙的なサブゴールに分解する必要がある．
- **サブゴール:** 物体とのやり取りが求められるサブタスク．
- **行動:** AI2-THOR 環境 [Kolve 17] において定義されている行動群のうちの 1 つ．行動は，“MoveAhead” “RotateRight” “RotateLeft” “LookUp” “LookDown” “PickUp” “Put” “Open” “Close” “ToggleOn” “ToggleOff” “Slice” “Stop” の計 13 種類存在する．

我々は AI2-THOR 環境を使用する．AI2-THOR は，photo-realistic な家庭環境における対話的なシミュレータである．ALFRED には 8 種類の評価指標が存在する．その中で，我々は Seen/Unseen Success Rate (SR) および Seen/Unseen Goal-Conditioned Success Rate (GC) をモデルの評価に使用した．これらについては後の章で説明する．

4. 提案手法

提案手法のネットワーク構造を図 2 に示す．ネットワークは State Encoder，Language Encoder，Subgoal History Encoder，Subgoal Predictor といった 4 つのモジュールから構成される．各モジュールの詳細については本章で後述する．本手法は，代表的な階層型 VLN モデルである HLSM [Blukis 21] をもとにしているものの，Episodic Transformer [Pashevich 21] 等の他の VLN モデルにも適用可能である．

4.1 入力

時刻 t における k 番目のサブゴール予測について，ネットワークの入力 $\boldsymbol{x}_{t,k}$ を以下のように定義する．

$$\boldsymbol{x}_{t,k} = \{l, \boldsymbol{G}_{1:k-1}, \boldsymbol{s}_t\}, \tag{1}$$

$$\boldsymbol{G}_{1:k-1} = \{\boldsymbol{g}_1, \boldsymbol{g}_2, \cdots, \boldsymbol{g}_i, \cdots, \boldsymbol{g}_{k-1}\}, \tag{2}$$

$$\boldsymbol{g}_i = \{C^{(i)}_{\mathrm{act}}, C^{(i)}_{\mathrm{arg}}, C^{(i)}_{\mathrm{mask}}\}, \tag{3}$$

$$\boldsymbol{s}_t = \{V_t, h_t, p_t\} \tag{4}$$

l は命令文，$\boldsymbol{G}_{1:k-1}$ は過去のサブゴール群，$\boldsymbol{s}_t$ は現在の状態表現を表し，$C^{(i)}_{\mathrm{act}}$ はサブゴールの種類，$C^{(i)}_{\mathrm{arg}}$ は対象物体，$C^{(i)}_{\mathrm{mask}}$ は対象物体の 3 次元マスクを表す．$\boldsymbol{s}_t$ は探索中の累積的な知識をもとにした環境に関する 3 次元情報であり，Observation Model [Blukis 21] に基づいて定期的に更新される．V_t は環境中の物体の位置情報を示すボクセルマップ，h_t はロボットが把持している物体，p_t はロボットの姿勢を表す．

4.2 Encoders

Language Encoder では，事前学習済みの BERT [Devlin 18] を用いて命令文 l の処理を行う．なお，CLS トークンの埋め込み表現を命令文の特徴量 $\boldsymbol{\phi}_l$ とする．Subgoal History Encoder では，複数層の Transformer を使用して，過去のサブゴール群 $\boldsymbol{G}_{1:k-1}$ に対する処理を自己回帰的に行う．得られる特徴量群 $\boldsymbol{\phi}_g^{(1:k-1)}$ のうち $\boldsymbol{\phi}_g^{(k-1)}$ をサブゴールの特徴量とする．State Encoder では，現在の状態表現 $\boldsymbol{s}_t$ に対して，$\boldsymbol{\phi}_s^{(t)} = \{h_t, f_{\mathrm{mp_s}}(V_t)\}$ という処理を行う．ここで，$f_{\mathrm{mp_s}}$ は空間的なプーリング処理を表し，V_t を視界に含まれる物体群に関する one-hot 表現へと変換する．

4.3 Subgoal Predictor

Subgoal Predictor には，各エンコーダの出力を連結した特徴量群 $\boldsymbol{\Phi}_{t,k} = \{\boldsymbol{\phi}_l, \boldsymbol{\phi}_g^{(k-1)}, \boldsymbol{\phi}_s^{(t)}\}$ を入力する．そして，以下に示すように，サブゴールの種類および対象物体を得る．

$$p(\hat{C}^{(k)}_{\mathrm{act}}) = \mathrm{softmax}(f_{\mathrm{mlp}}(\boldsymbol{\Phi}_{t,k})), \tag{5}$$

$$p(\hat{C}^{(k)}_{\mathrm{arg}}) = \mathrm{softmax}(f_{\mathrm{mlp}}(\boldsymbol{\Phi}_{t,k})) \tag{6}$$

f_{mlp} は多層パーセプトロンを表す．

続いて，以下の式に基づき，環境中の物体群に関する鳥瞰図表現を導出する．

$$\boldsymbol{b}_{t,k} = \{a_t, \sum_{i=0}^{k-1} f_{\mathrm{mp_p}}(C_{\mathrm{mask}}^{(i)}), f_{\mathrm{mp_p}}(V_t)\} \tag{7}$$

a_t は上から見た環境の 2 次元情報を表し，各位置は 7 種類のクラス (“pickable” “receptacle” “togglable” “openable” “ground” “obstacle” “observed”) のいずれかに属する．$f_{\mathrm{mp_p}}$ は平面的なプーリング処理を表し，環境に関する 3 次元情報を 2 次元情報へと変換する．

最後に，対象物体の 3 次元マスクを以下のように計算する．

$$p(\hat{C}_{\mathrm{mask}}^{(k)}) = f_{\mathrm{lun}}(f_{\mathrm{trans}}(\boldsymbol{b}_{t,k}, p_t), \boldsymbol{\phi}_l) \tag{8}$$

f_{lun} は LingUNet [Misra 18] を表し，f_{trans} は鳥瞰図表現とロボットの一人称視点の対応付けを行う．そして，最終出力を $\boldsymbol{g}_k = \{C_{\mathrm{act}}^{(k)}, C_{\mathrm{arg}}^{(k)}, C_{\mathrm{mask}}^{(k)}\}$ とする．

4.4 Moment-based Adversarial Training

VILLA [Gan 20] に基づき，我々は特徴量空間に対して直接的に摂動を加える．摂動 $\boldsymbol{\delta}$ を $\boldsymbol{\delta} = \{\delta_l, \delta_g, \delta_s\}$ と定義する．δ_l，δ_g，δ_s はそれぞれ命令文，過去のサブゴール群，現在の状態表現に対する摂動を表す．

はじめに，以下に示すように，交差エントロピー誤差の $\boldsymbol{\delta}$ に関する勾配を計算する．

$$E(\boldsymbol{\delta}) = \mathrm{CE}(f(\boldsymbol{x} + \boldsymbol{\delta}), \boldsymbol{g}), \tag{9}$$

$$\nabla_{\boldsymbol{\delta}} E(\boldsymbol{\delta}) = \frac{\partial E}{\partial \boldsymbol{\delta}} \tag{10}$$

f はネットワーク全体，$\mathrm{CE}(\cdot,\cdot)$ は交差エントロピー誤差を表す．

続いて，Adam をもとにした 2 種類の移動平均を導入する．

$$\boldsymbol{m}_t = \rho_1 \boldsymbol{m}_{t-1} + (1-\rho_1)\nabla_{\boldsymbol{\delta}} E(\boldsymbol{\delta}_t), \tag{11}$$

$$\boldsymbol{v}_t = \rho_2 \boldsymbol{v}_{t-1} + (1-\rho_2)(\nabla_{\boldsymbol{\delta}} E(\boldsymbol{\delta}_t))^2 \tag{12}$$

t は現在の摂動更新ステップ，ρ_1 と ρ_2 は各移動平均の平滑化係数を表す．これらを用いて，摂動の更新幅 $\Delta\boldsymbol{\delta}_t$ を導出する．

$$\hat{\boldsymbol{m}}_t = \frac{\boldsymbol{m}_t}{1-(\rho_1)^t}, \tag{13}$$

$$\hat{\boldsymbol{v}}_t = \frac{\boldsymbol{v}_t}{1-(\rho_2)^t}, \tag{14}$$

$$\Delta\boldsymbol{\delta}_t = \eta \frac{\hat{\boldsymbol{m}}_t}{\sqrt{\hat{\boldsymbol{v}}_t + \varepsilon}} \tag{15}$$

η は MAT の学習率，ε はゼロ除算を避けるための小さな値を表す．

最後に，$\Delta\boldsymbol{\delta}_t$ を用いて摂動を更新する．

$$\boldsymbol{\delta}_{t+1} = \Pi_{\|\boldsymbol{\delta}\| \le \epsilon}(\boldsymbol{\delta}_t + \frac{\Delta\boldsymbol{\delta}_t}{\|\Delta\boldsymbol{\delta}_t\|_F}) \tag{16}$$

$\Pi_{\|\cdot\| \le \epsilon}$ は ϵ 球への投影，$\|\cdot\|_F$ はフロベニウスノルムを表す．

4.5 損失関数

学習には，以下の損失関数を使用する．

$$L = L_{cln} + L_{at} + \lambda L_{kl}, \tag{17}$$

$$L_{cln} = \mathrm{CE}(f(\boldsymbol{x}_{t,k}), \boldsymbol{g}_k), \tag{18}$$

$$L_{at} = \mathrm{CE}(f(\boldsymbol{x}_{t,k} + \boldsymbol{\delta}), \boldsymbol{g}_k), \tag{19}$$

$$L_{kl} = \mathrm{D}_{KL}(f(\boldsymbol{x}_{t,k}) \| f(\boldsymbol{x}_{t,k} + \boldsymbol{\delta})) + \mathrm{D}_{KL}(f(\boldsymbol{x}_{t,k} + \boldsymbol{\delta}) \| f(\boldsymbol{x}_{t,k})) \tag{20}$$

λ は項の重み，$\mathrm{D}_{KL}(\cdot\|\cdot)$ はカルバック・ライブラー情報量を表す．

5. 実験

5.1 実験設定

表 1 に実験設定を示す．#L，#H，#A はそれぞれ Transformer の層数，隠れ層の次元数，Attention Head の数を表す．また，α は各イテレーションにおける摂動の更新回数を表す．我々のモデルのパラメータ数は 1 億 1100 万である．学習には，メモリ 32GB 搭載の Tesla V100 を 4 台使用し，2 日間を要した．各エピソードの推論に要する時間は 2 分であった．

5.2 データセット

本実験では，モデルの評価に ALFRED ベンチマーク [Shridhar 20] を使用した．ALFRED は，各エピソードに対する専門家のデモンストレーションおよび命令文から構成される．命令文は，Amazon Mechanical Turk を通じて 3 人以上のアノテータによって作成され，高レベルなものと低レベルなものの両方が含まれる．なお，事前処理として，各エピソードについてサブゴール群を取り出し，カメラ画像をもとに対象物体およびその 3 次元マスクの情報を抽出した．当該データセットは，25,743 の命令文および 8,055 のエピソードから構成される．各エピソードは，“Pick & Place” “Stack & Place” “Pick Two & Place” “Clean & Place” “Heat & Place” “Cool & Place” “Examine in Light” といった 7 種類のタスクに分類される．

我々は，[Shridhar 20] に従ってデータセットを分割した．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．Seen 集合のエピソードは，訓練集合に登場する既知の環境で実施され，Unseen 集合のエピソードは，訓練集合に登場しない未知の環境で実施される．

5.3 実験結果

定量的結果を表 2 に示す．評価尺度には Seen/Unseen SR および Seen/Unseen GC を使用した．SR は $\mathrm{SR} = N_{\mathrm{s}}/N$ と定義される．ここで，N はエピソードの総数を表し，N_{s} は成功したエピソードの総数を表す．なお，物体の位置および状態がエピソードのゴール条件を満たしている場合に，当該エピソードにおいて「成功」したと見なす．主要評価尺度には Unseen SR を選択した．

我々は，HLSM をベースライン手法とした．これは，HLSM が代表的な階層モデルであり，ALFRED において良好な結果が報告されているためである．Unseen SR に関して，検証集合において，提案手法は 17.66%を記録し，ベースライン手法は 11.80%を記録している．テスト集合においては，提案手法は 21.84%を記録し，ベースライン手法は 16.29%を記録している．これより，提案手法がベースライン手法をそれぞれ 5.55%と 5.86%上回っていることがわかる．

Ablation Study には，以下の 2 条件を定めた．

表 1: パラメータ設定

Language Encoder	#L: 12	#H: 768	#A: 12
Subgoal History Encoder	#L: 2	#H: 128	#A: 8
Optimizer	Adam ($\beta_1 = 0.9, \beta_2 = 0.999$)		
Learning rate	5×10^{-5}		
Epoch	6		
Batch size	1		
Loss weight	1.5		
MAT	$\alpha = 7$ $\eta = 1 \times 10^{-3}$ $\lambda = 1.5$ $\rho_1 = 0.9, \rho_2 = 0.999$		

表 2: 定量的結果

Method	Condition	MAT	Perturbation			Validation				Test			
						Unseen		Seen		Unseen		Seen	
			ϕ_l	ϕ_g	ϕ_s	SR	GC	SR	GC	SR	GC	SR	GC
[Zhang 21]						11.12	17.89	13.63	21.11	10.23	20.27	18.41	25.27
[Nottingham 21]						-	-	12.7	23.4	6.38	17.27	13.35	23.31
[Blukis 21]						11.80	24.70	25.00	36.37	16.29	27.74	25.11	35.79
Ours	full	✓	✓	✓	✓	**17.66**	31.79	30.98	42.29	**21.84**	32.41	**33.01**	43.65
	(i)		✓	✓	✓	16.69	31.18	**31.71**	41.82	20.99	32.03	30.66	43.10
	(ii)-a	✓		✓	✓	15.71	30.66	30.85	41.54	21.39	31.03	32.29	43.30
	(ii)-b	✓	✓		✓	15.96	**31.89**	30.24	**42.63**	21.52	**33.43**	31.57	**44.38**
	(ii)-c	✓	✓	✓		14.25	29.06	30.73	40.68	20.27	30.66	30.14	41.23

(i) W/o MAT: 敵対的学習において，MAT を適用する場合としない場合で，性能にどの程度の差が生じるかを調査する．

(ii) Selective adversarial training: 命令文，サブゴール，状態表現のそれぞれの特徴量空間に，敵対的摂動を付与する場合としない場合で，性能にどの程度の差が生じるかを調査する．

表 2 に示すように，テスト集合における Unseen SR に関して，MAT を適用しない場合に 20.99%を記録している．また，命令文，サブゴール，状態表現のそれぞれの特徴量空間に摂動を付与しない場合に，21.39%, 21.52%, 20.27%を記録している．これより，状態表現に対する MAT が最も性能向上に寄与していると考えられる．

定性的結果を図 3 に示す．図 3 より，ロボットがカップをランプのもとに運び，明かりをつけることに成功している様子が確認できる．

6. おわりに

本研究では，自然言語の命令文を適切な行動に対応付けるモデルの開発に取り組んだ．提案手法による貢献を以下に示す．

- VLN において，サブゴールおよび状態表現の特徴量空間に対し，敵対的な摂動を導入した．
- 敵対的摂動更新アルゴリズム MAT を提案した．
- ALFRED ベンチマークにおいて，提案手法がベースライン手法を全ての評価尺度で上回った．

謝辞

本研究の一部は，JSPS 科研費 20H04269，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

[Anderson 18] Anderson, P., Wu, Q., et al.: Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments, in *CVPR*, pp. 3674–3683 (2018)

[Blukis 21] Blukis, V., Paxton, C., et al.: A persistent spatial semantic representation for high-level natural language instruction execution, *arXiv preprint arXiv:2107.05612* (2021)

図 3: 定性的結果

[Devlin 18] Devlin, J., Chang, M.-W., Lee, K., et al.: Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* (2018)

[Gan 20] Gan, Z., Chen, Y.-C., Li, L., Zhu, C., Cheng, Y., et al.: Large-scale adversarial training for vision-and-language representation learning, *arXiv preprint arXiv:2006.06195* (2020)

[Ishikawa 21] Ishikawa, S. and Sugiura, K.: 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)

[Kollar 10] Kollar, T., Tellex, S., Roy, D., and Roy, N.: Toward understanding natural language directions, in *HRI*, pp. 259–266 (2010)

[Kolve 17] Kolve, E., Mottaghi, R., Han, W., VanderBilt, E., Weihs, L., et al.: Ai2-thor: An interactive 3d environment for visual ai, *arXiv preprint arXiv:1712.05474* (2017)

[Kuo 20] Kuo, Y.-L., Katz, B., and Barbu, A.: Deep compositional robotic planners that follow natural language commands, in *ICRA*, pp. 4906–4912 (2020)

[Magassouba 19] Magassouba, A., Sugiura, K., Quoc, A. T., et al.: 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 (2019)

[Magassouba 21] Magassouba, A., Sugiura, K., et al.: 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)

[Misra 18] Misra, D., Bennett, A., Blukis, V., et al.: Mapping instructions to actions in 3d environments with visual goal prediction, *arXiv preprint arXiv:1809.00786* (2018)

[Nottingham 21] Nottingham, K., Liang, L., Shin, D., et al.: Modular framework for visuomotor language grounding, *arXiv preprint arXiv:2109.02161* (2021)

[Pashevich 21] Pashevich, A., Schmid, C., and Sun, C.: Episodic Transformer for Vision-and-Language Navigation, *arXiv preprint arXiv:2105.06453* (2021)

[Shridhar 20] Shridhar, M., Thomason, J., Gordon, D., Bisk, Y., et al.: Alfred: A benchmark for interpreting grounded instructions for everyday tasks, in *CVPR*, pp. 10740–10749 (2020)

[Tellex 11] Tellex, S., Kollar, T., Dickerson, S., Walter, M., et al.: Understanding Natural Language Commands for Robotic Navigation and Mobile Manipulation., in *AAAI*, Vol. 2 (2011)

[Thomason 20] Thomason, J., Murray, M., Cakmak, M., and Zettlemoyer, L.: Vision-and-dialog navigation, in *CoRL*, pp. 394–406 (2020)

[Wu 21] Wu, W., Chang, T., and Li, X.: Visual-and-Language Navigation: A Survey and Taxonomy, *arXiv preprint arXiv:2108.11544* (2021)

[Zhang 21] Zhang, Y. and Chai, J.: Hierarchical Task Learning from Language Instructions with Unified Transformers and Self-Monitoring, *arXiv preprint arXiv:2106.03427* (2021)