The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

生活支援ロボットにおける Generative Adversarial Nets を用いた曖昧な指示の理解

Understanding Ambiguous Instructions Using Generative Adversarial Nets for Object Disposal Tasks

杉浦 孔明　マガスーバ アリー　河井 恒
Komei Sugiura　Aly Magassouba　Hisashi Kawai

情報通信研究機構
National Institute of Information and Communications Technology

This paper focuses on a multimodal language understanding method for “Carry and Place” tasks with domestic service robots. We address the case of ambiguous instructions, that is when the target area is not specified. For instance “Put away the milk and cereal.” is a natural instruction where there is ambiguity on the target area, considering daily life environments. Conventionally, this instruction can be disambiguated from a dialogue system, but at the cost of time and cumbersomeness. Instead, we propose a multimodal approach, where the instructions are disambiguated from the robot state and environment context. We develop MultiModal Classifier Generative Adversarial Network (MMC-GAN) to predict the likelihood of the different target areas considering the robot physical limitation and the target clutter. Our approach, MMC-GAN, significantly improves accuracy compared to baseline methods using instructions only or simple deep neural networks.

1. はじめに

生活支援ロボットによる曖昧な音声命令の理解は，多くの関連課題を有する挑戦的な分野である [杉浦 16]．本研究では，特に Carry and Place タスクにおける曖昧な命令文の理解を扱う．Carry and Place タスクとは，ユーザに指示された物体を適切な場所に移動させるタスクを指すものとする．例として，日常環境で「ペットボトルを片付けておいて」という音声命令をロボットが実行するタスクを考える．人間同士の場合はこのような省略された命令文で通じる場合が多いが，ロボットが行動を開始するために十分な情報を含んでいない．一方，十分な情報を含む命令文は，不自然であることが多い．例えば,「現在把持中のペットボトルをキッチンの棚の３段目の右側に片付けておいて」のような発話をユーザが行うことは考えにくい．

上記の問題への単純なアプローチとして，スロット値がすべて確定するまで聞き返す戦略が考えられる [Kollar 13, Gemignani 15]．実際に，世界最大の生活支援ロボットのベンチマークテストであるロボカップ＠ホーム [Iocchi 15] においても，このアプローチが支配的である．しかしながら，このアプローチでは「どのペットボトルですか？」「キッチンのどの棚ですか？」「棚の何番目の段ですか？」など多くの確認発話が生成されるため，動作実行するまでに必要な時間が長く，不便である．

本研究では，上記のタスクに対し，マルチモーダル情報を用いて省略内容の推定を行う．具体的には，曖昧な命令を入力として，Place 動作の対象領域の尤度を予測する手法を提案する．そのために，[Sugiura 17] で提案された Latent Classifier Generative Adversarial Nets (LAC-GAN) を拡張し，マルチモーダル情報入力を可能とする．

GAN [Goodfellow 14] に関する先行事例としては，Conditional GAN [Mirza 14] や InfoGAN [Chen 16] などが挙げられる．また，AC-GAN [Odena 16] のように，カテゴリを出力に利用した GAN も提案されている．これらの手法は，画像や文の生成などに適用され，品質の良い疑似サンプルの生成が報告されている．さらに，近年，GAN を分類問題に適用した研究として，文献 [Springenberg 15, Shen 17] が挙げられる．ロボティクスにおいては教師データ収集が高コストであるため，Generative Adversarial Nets (GAN) によるデータ拡張と分類を同時に可能な提案手法が重要性を持ち得る．この事実に着想を得て提案された手法が，LAC-GAN [Sugiura 17] である．LAC-GAN と提案手法 MMC-GAN（MultiModal Classifier GAN）の差異は，マルチモーダル情報を扱う Extractor を持つ点である．

本研究の独自性は以下である．

- GAN によるデータ拡張と分類を同時に行う手法 MMC-GAN を提案する．
- Carry and Place タスクにおける曖昧な命令理解に MMC-GAN を適用する．

2. 問題設定

本研究の対象タスクは,「Cary and Place」タスクにおける曖昧な命令（「お茶を片付けて」など）を理解することである．この命令に対する出力として，適切な対象領域を予測する．本研究では，以下の入出力を想定する．

- 入力： 命令文，状態を表す文，（環境中を巡回して得た）対象領域のカメラ画像
- 出力： オブジェクトを片付ける対象領域の尤もらしさ．

ユーザインタフェースとしてタブレット端末を想定し，対象領域（数十箇所以上）は画像でユーザに提示されるものとする．全対象領域を一度にユーザが視認することはできないため，正確な尤度スコアに応じて対象領域をランキングすることが利便性の向上（ユーザタップ数の減少）につながる．

本研究では，様々な実験条件に対して複数回の実験を行うことを主眼とするため，実環境・実ロボットでの実験ではなく，データセット上の分類問題としての統計的評価を論じることを優先する．一方，提案手法をロボットへ移植する際に，学習済みのモデルが実環境とミスマッチになることは避けなければならない．そのため，ハードウェアとしては標準化された生活支援ロボットと同等のデバイスを用いる．実際に，本研究で構築

連絡先: 杉浦孔明，京都府相楽郡精華町光台 3-5

したモデルはそのままロボットに移植可能である．ロボットを用いた動画に関しては，以下の URL[*1] を参照されたい．

生活支援ロボットのハードウェアとして，標準化されたロボットプラットフォームを前提とする．本研究では，HSR（Human Support Robot）[寺田 12] をプラットフォームとして想定する．世界最大の生活支援ロボットのベンチマークテスト RoboCup@Home において，HSR は標準化されている．

3. 提案手法: Multimodal Classifier GAN

3.1 LAC-GAN

LAC-GAN [Sugiura 17] は，Extractor E，Generator G，Discriminator D，の 3 つの部分から構成される．LAC-GAN の構造を図 1 に示す．

いま，学習サンプルが，$(\boldsymbol{x}_{raw}, y^*)$ の組で与えられるものとする．ここに，$\boldsymbol{x}_{raw} \in \mathbb{R}^N$ は E による抽出前の特徴量，y^* は対応するカテゴリのラベルである．y^* は d_y 次元の one-hot ベクトルであるものとする．

E の目的は，特徴抽出前の入力 $\boldsymbol{x}_{raw}$ から，分類に適した特徴量 $\boldsymbol{x}_{real}$ を抽出することである．GAN を用いる目的が $\boldsymbol{x}_{raw}$ に似た $\boldsymbol{x}_{fake}$（画像等）の生成であれば，$\boldsymbol{x}_{raw}$ を D への入力とすることに合理性がある．一方，GAN を用いて分類器を構成する場合，$\boldsymbol{x}_{raw}$ をそのまま用いるより，分類に適した特徴量 $\boldsymbol{x}_{real}$ を用いたほうが有利である．

E の学習は，以下の交差エントロピー J_C を最小化することで行う．

$$J_C = -\sum_n \sum_j y^*_{nj} \log p_D(y_{nj}) \tag{1}$$

ここに，$p(y_{nj})$ は E の出力層の値である．

G の入力は，カテゴリ c および $\boldsymbol{z} \in \mathbb{R}^{d_z}$ である．c および $\boldsymbol{z}$ は，それぞれカテゴリカル分布および標準正規分布からサンプルされるものとする．紙幅の都合上，$\boldsymbol{z} = \{\boldsymbol{z}, c\}$ と表記する．G の出力は，$\boldsymbol{x}_{fake} = G(\boldsymbol{z})$ である．

D の入力源 S は $\{real|fake\}$ のいずれかから選択される．そのときの D の入力をそれぞれ $\boldsymbol{x}_{fake}$，$\boldsymbol{x}_{real}$ と書くこととする．$D(\boldsymbol{x})$ は，入力源 S が $real$ である確率の推定値 $p(S)$ を出力する．すなわち，

$$D(\boldsymbol{x}) = p(S = real|\boldsymbol{x})$$

である．

G および D の学習では，以下のコスト関数 $J^{(D)}$ および $J^{(G)}$ をそれぞれ最小化する．

$$J^{(D)} = J_S + \lambda J_C$$

$$J^{(G)} = -J_S$$

$$J_S = -\frac{1}{2}\mathbb{E}_{\boldsymbol{x}_{real}} \log D(\boldsymbol{x}_{real}) - \frac{1}{2}\mathbb{E}_{\boldsymbol{z}} \log(1 - D(G(\boldsymbol{z})))$$

ここに，λ は重みである．パラメータ学習では，D と G の学習を交互に行う．まず，D の学習を行い，その後 D のパラメータを固定して G の学習を行う．

3.2 Multimodal Classifier GAN

以下では，提案手法 MMC-GAN について説明する．LAC-GAN では，E は Multi-Layer Perceptron 型の DNN であった．一方，画像を含むマルチモーダル入力に対しては，この構造が最適とはいえない．よって，本研究では，E を拡張した Multimodal Classifier GAN (MMC-GAN) を提案する．MMC-GAN における E の構造を図 2 に示す．

*1 https://youtu.be/_YQuziz4eGY

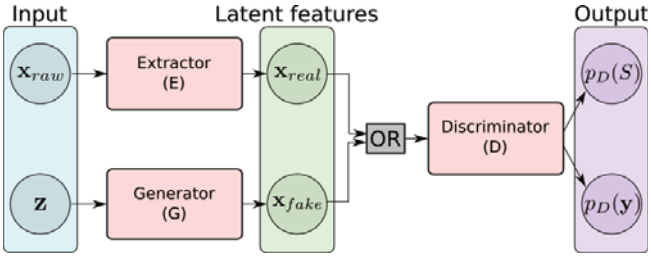

図 1: LAC-GAN および MMC-GAN の構成の概要

図 3: Samples of the 1282 depth images (right) $\mathbf{v}_d$ and their corresponding RGB images (left). Only the depth data is used.

MMC-GAN の Extractor E では，以下を入力とする．

$$\mathbf{x}_{raw} = \{\mathbf{w}_{inst}, \mathbf{w}_c, \mathbf{v}_d, \mathbf{v}_{meta}\}$$

ここに，$\mathbf{w}_{inst}, \mathbf{w}_c, \mathbf{v}_d, \mathbf{v}_{meta}$ はそれぞれ，命令文，コンテキストの自然言語表現，画像，画像メタ情報である．$\mathbf{w}_{inst}, \mathbf{w}_c$ の例を以下に挙げる．

$$\begin{cases} \mathbf{w}_{inst} = \text{“Move the towel to the shelf.”} \\ \mathbf{w}_c = \text{“The robot is holding a towel.”} \end{cases}$$

$\mathbf{v}_d$ としては Depth 画像を用いるが，RGB 画像や RGB-D 画像を用いてもよい．ハードウェアの制限があるため，Place 動作を行う場合，候補領域の適切性が床からの高さに依存することは自明である．よって，画像を取得した時点で得られる候補領域の高さ，カメラの高さおよび角度を $\boldsymbol{w}_c$ として入力に加える．

Extractor E は，上述の入力を潜在表現に変換する．まず，$\mathbf{w}_{inst}, \mathbf{w}_c$ を固定長の分散表現で表すために，PV-DM [Le 14] を用いる．これにより，文の長さに関わらず，固定長の分散表現 $\boldsymbol{x}_{inst}$ および $\boldsymbol{x}_c$ が得られる．一方，$\mathbf{v}_d, \mathbf{v}_{meta}$ を潜在表現に変換するため，図 2 に示す Convoluational Neural Network (CNN) を用いる．CNN の学習におけるコスト関数は，上述の J_C を用いる．中間の隠れ層の出力を $\boldsymbol{x}_v$ とする．

最終的に，E の出力として以下が得られる．

$$\mathbf{x}_{real} = \{\mathbf{x}_{instr}, \mathbf{x}_c, \mathbf{x}_v\}.$$

ここに，$\mathbf{x}_{real}$ は各潜在表現を連結したベクトルであり，D に入力される．

4. Carry and Place マルチモーダルデータセットの構築

Carry and Place タスクにおける命令理解の標準的なマルチモーダルデータセットは我々の知る限り存在しないため，実験で用いるデータセットを構築した．本節では，我々が構築した「Carry and Place マルチモーダルデータセット」について説明する．構築手順を以下に示す．

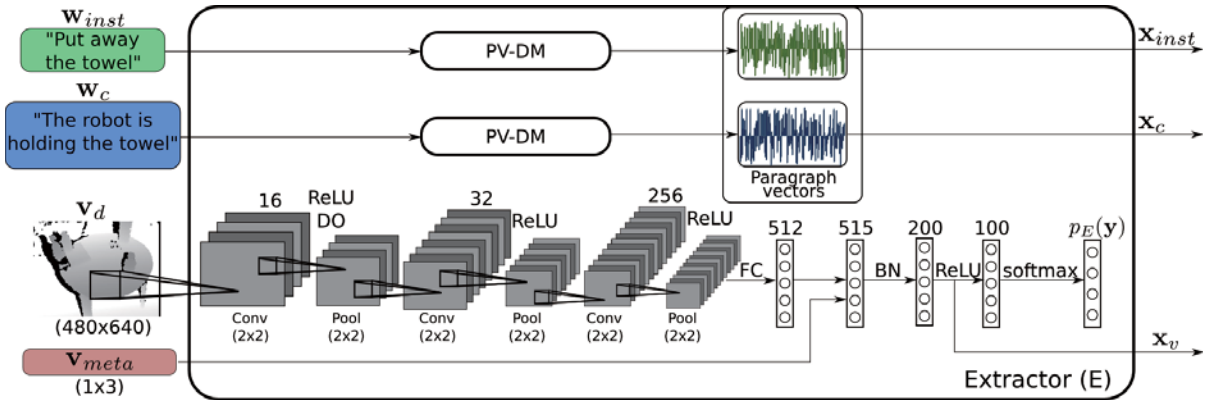

図 2: MMC-GAN における Extractor E の構造

表 1: 対象領域のバリエーション

#	家具の種類	対象領域数	画像数
テーブル	4	18	321
引き出し	4	10	336
机	4	8	425
棚	2	5	200
$\sum$	**14**	**37**	**1282**

1 日用品を対象領域にランダムに配置する（図3 参照).

2 物体検出位置にロボットカメラを配置し，RGB-D 情報を取得する.

3 ランダムに命令文を生成する.

4 ランダムに状態を表す文を生成する.

5 命令文と状態を表す文が，実世界状況において成り立たなければ，ステップ 3 に戻る．成り立てば，6 に進む.

6 ラベラーが正解ラベルを付与する.

本研究では，標準化されたロボットでの動作を想定するため，ハードウェアとして生活支援ロボット HSR を想定する．よって，画像の取得には，HSR と同様，Asus Xtion Pro カメラを用いた．入力として，RGB-D 画像を用いることも可能であるが，本研究では必要最小限の非言語情報で動作を確認するために，Depth 画像のみを用いた．対象領域を表1 にまとめる．1 つの家具に対して複数の対象領域が存在すること（棚の 1 段目，2 段目など），各画像では日用品の配置が異なること，に注意されたい．なお，引き出しは開放されているものとした.

各サンプルは以下の基準により，ラベル付けを行った.

($A1$) 対象領域は十分なスペースがあるとともに，ロボットアームの経路に障害が少なく，自律制御によるタスク達成が簡単である.

($A2$) 対象領域には障害物が一部存在するためロボットアームの経路計画が失敗する可能性があるものの，自律制御によるタスク達成が十分見込まれる.

($A3$) 対象領域には限られたスペースしかなく，ロボットアームが衝突する可能性が高いため，自律制御によるタスク達成の見込みは低い.

($A4$) 対象領域には障害物が多く，タスク達成は不可能に近い.

データセットをランダムにシャッフルし，学習セット (Train)，検証セット (Valid)，テストセット (Test) に分割した．データセットのサイズを表2 に示す.

表 2: 実験で用いたデータセットの概要

#	Train	Valid	Test	$\sum$
$A1$	158	29	25	**212**
$A2$	359	34	39	**432**
$A3$	350	26	22	**398**
$A4$	203	17	20	**240**
$\sum$	**1070** (*83* %)	**106** (*8.5* %)	**106** (*8.5* %)	**1282**

5. 実験

5.1 設定

パラメータ設定を表3 に示す．$\mathbf{x}_v$，w_{inst}，w_c，の次元数は各 200 次元とした．ゆえに，$\mathbf{x}_{real}$ は 600 次元である．提案手法における G として，2 種類の構造（GAN，CGAN）について検討した．それぞれの隠れ層のノード数は同等であるが，CGAN では入力にカテゴリカル分布からサンプルされる 4 次元の乱数が追加されている.

5.2 結果

定性的結果を図4 に示す．上段および中段は正しい予測結果であり，下段は誤りの例である．左上の記号は正解ラベルおよび予測結果を示す．例えば，「A1/A2」はラベルが「A1」であり，予測結果が「A2」であったことを示す.

次に，提案手法とベースライン手法（通常の DNN）を比較評価した．定量的結果を表4 に示す．一般に，DNN の精度比較では，エポックごとにモデルパラメータが更新される．よって，テストセットの最大値を比較しても，未知のデータに対する精度を表すものにならない．ゆえに，本実験では，標準的な手順に従い，検証セットの精度が最大値を示したモデルを各手法の最良モデルとした．表4 に示す結果は，初期値をランダムに変更して行った 10 回の実験の平均および標準偏差を示す．ただし，*を付した結果は，一部の結果が収束しなかったため，収束した結果の平均および標準偏差を示した.

表において，(I)，(I + C)，(I + C + V) は，命令文のみ，命令文および状態，命令文・状況・画像，をそれぞれ入力特徴とした場合である．表より，(I) および (I + C) に比べ，(I + C + V)

表 4: ベースライン手法と提案手法 MMC-GAN の比較結果

	[%]	入力特徴					
	GAN	**I**		**I + C**		**I + C + V**	
手法	type	Valid	Test	Valid	Test	Valid	Test
ベースライン	-	35.3 ± 2.0	28.6 ± 3.3	34.4 ± 1.5	26.7 ± 6.1	77.0 ± 1.2	76.6 ± 1.2
提案手法	GAN	27.1 ± 2.1	24.8* ± 6.3	28.0 ± 2.9	27.5* ± 8.7	84.2 ± 1.0	79.1 ± 1.6
提案手法	CGAN	28.5 ± 2.5	28.8* ± 7.9	28.1 ± 2.7	27.6* ± 8.2	83.7 ± 0.4	**81.0** ± **2.5**

表 3: MMC-GAN のパラメータ設定

最適化手法	Adam (学習率 0.00005, $\beta_1 = 0.5, \beta_2 = 0.9$), λ=0.2
バッチサイズ	64 (E), 50 (G and D)
GAN ノード数	100, 100, 100, 100
CGAN ノード数	100, 100, 100, 100

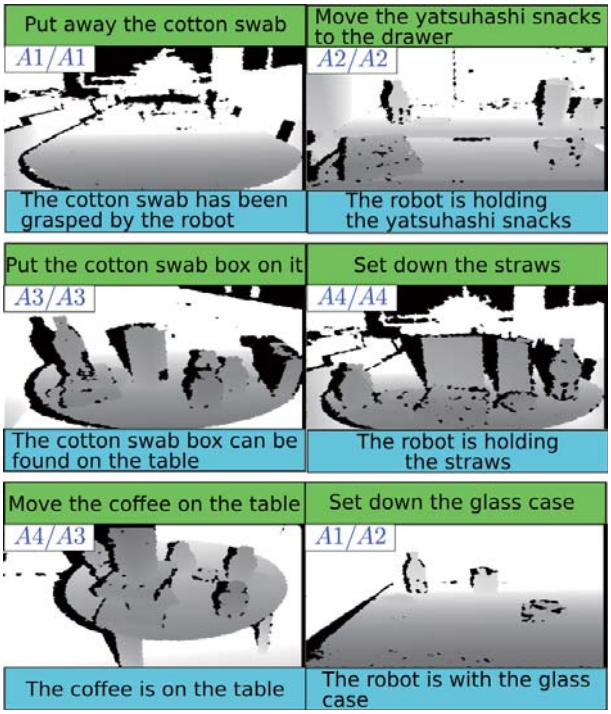

図 4: 対象領域の分類に関する定性的結果．上段および中段は正しい予測結果であり，下段は誤りの例である．

が高い精度であることから，本タスクは言語情報のみの特徴量では不十分であることを示している．また，(I) と (I + C) の間には統計的に有意な差は存在しなかった．

公平な比較を行うために，ベースラインとして構造が同等な DNN を用いた．ベースラインと提案手法を比較すると，提案手法の方が良好な結果であった．このことは，MMC-GAN によるデータ拡張がテストセット精度向上に寄与したことを示唆している．さらに，MMC-GAN に関して G の構造を比較すると，クラス情報を入力とした CGAN の方が高い精度であった．

6. おわりに

高齢化社会のなかで体の不自由な人々を物理的に支援可能な生産年齢人口は十分ではなく，彼らの家族が離職を余儀なくされるケースさえある．この社会課題の解決策として，生活支援ロボットは体の不自由な人々の生活の質を向上させるだけでなく，その家族を時間的拘束から解放するとともに，社会全体の生産性向上につながり得る．本論文では，生活支援ロボットのための命令理解手法を提案した．提案手法 MMC-GAN を Carry-and-Place タスクにおける命令に適用し，ベースラインを上回る精度を得た．

謝辞

本研究の一部は，JST CREST の助成を受けて実施されたものである．

参考文献

[Chen 16] Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P.: InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets, in *Advances in Neural Information Processing Systems*, pp. 2172–2180 (2016)

[Gemignani 15] Gemignani, G., Veloso, M., and Nardi, D.: Language-based Sensing Descriptors for Robot Object Grounding, in *Robot Soccer World Cup*, pp. 3–15 (2015)

[Goodfellow 14] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y.: Generative Adversarial Nets, in *Advances in Neural Information Processing Systems*, pp. 2672–2680 (2014)

[Iocchi 15] Iocchi, L., Holz, D., Solar, Ruiz-del J., Sugiura, K., and Zant, van der T.: RoboCup@Home: Analysis and Results of Evolving Competitions for Domestic and Service Robots, *Artificial Intelligence*, Vol. 229, pp. 258–281 (2015)

[Kollar 13] Kollar, T., Perera, V., Nardi, D., and Veloso, M.: Learning Environmental Knowledge from Task-based Human-Robot Dialog, in *IEEE ICRA*, pp. 4304–4309 (2013)

[Le 14] Le, Q. and Mikolov, T.: Distributed Representations of Sentences and Documents, in *Proc. ICML*, pp. 1188–1196 (2014)

[Mirza 14] Mirza, M. and Osindero, S.: Conditional Generative Adversarial Nets, arXiv preprint arXiv:1411.1784 (2014)

[Odena 16] Odena, A., Olah, C., and Shlens, J.: Conditional Image Synthesis with Auxiliary Classifier GANs, arXiv preprint arXiv:1610.09585 (2016)

[Shen 17] Shen, P., Lu, X., Li, S., and Kawai, H.: Conditional Generative Adversarial Nets Classifier for Spoken Language Identification, Proc. Interspeech (2017)

[Springenberg 15] Springenberg, J. T.: Unsupervised and Semi-Supervised Learning with Categorical Generative Adversarial Networks, arXiv preprint arXiv:1511.06390 (2015)

[Sugiura 17] Sugiura, K. and Kawai, H.: Grounded Language Understanding for Manipulation Instructions Using GAN-Based Classification, *IEEE ASRU* (2017)

[杉浦 16] 杉浦 孔明：ロボットによる大規模言語学習に向けて-実世界知識の利活用とクラウドロボティクス基盤の構築-, 計測と制御, Vol. 55, No. 10, pp. 884–889 (2016)

[寺田 12] 寺田 耕志, 橋本 国松, 齋藤 史倫, 池田 幸一, 山本 貴史：生活支援ロボット HSR の開発と実証評価, 第 13 回システムインテグレーション部門講演会資料集, pp. 2A1–2 (2012)