Latent Classifier Generative Adversarial Netsによる動詞のない命令文理解

○杉浦孔明，河井恒（情報通信研究機構）

1. はじめに

ロボットとの音声対話において，不完全情報および記号接地に対応した言語処理は，多くの関連課題を有する挑戦的な分野である [1]．例として，日常環境で「新聞片付けておいて」という音声指示をロボットが実行するタスクを考える．環境中には「新聞（に分類され得るオブジェクト）」が複数存在する可能性があるうえ,「どれを」「どこに」「どうやって」片付けるか，など様々なレベルで曖昧性が存在するため，望ましい動作を実行することは簡単ではない．また，指示者にとって当たり前であるタスクの開始条件・終了条件や，行動途中で変化した状況に対しロボットがとるべき行動も，指示文の言語情報のみからは自明でないことが多い．

本研究では，物体操作タスクにおける不完全情報を含む言語理解の一例として，動詞のない命令文からの物体操作可能性の推定を扱う．一般に，動作スロットが埋まっていない場合は，確認発話により聞き返しを行うこともできるが，実世界情報に基づいて動作スロットを補完できれば利便性の向上につながる．本研究で想定する状況を図 1 に示す．

本研究では，上記のタスクに対し，LAtent Classifier Generative Adversarial Nets (LAC-GAN) を提案する．LAC-GAN は Generative Adversarial Nets(GAN [2]) を拡張し，分類器として利用するものである．GAN [2] に関する先行事例としては，Conditional GAN [3] や InfoGAN [4] などが挙げられる．また，AC-GAN [5] 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 [6,7] が挙げられる．LAC-GAN はこれらと関連するが，GAN における Generator/Discriminator に加え，特徴抽出を行う Extractor を有することが異なる．本研究の独自性は以下である．

- GAN に Extractor を導入した手法である LAC-GAN を提案する．
- 物体操作タスクにおいて，物体および状況の言語表現から物体操作可能性を推定する．

2. 関連研究

実世界知識を扱う音声対話に関する分野は，実世界知識から記号・言語への一方向あるいは双方向の変換を扱う分野との関連が深い．音声およびテキスト対話システム分野の動向については，[8] が詳しい．一方、ロボティクスの分野においても，動作-記号の相互変換に関する試みが近年広く行われている [9]．Kollar らは，ロボットに与える移動指示に関して，ランドマークオブジェクトや動作にグラウンドした言語表現を学習する手法を提案した [10]．学習されたモデルを用いることにより，指示が示す最も確からしい経路を推論する．

本分野に関連するプロジェクトとしては，DARPA BOLT [11]，Robo Earth [12]，RoboBrain [13]，長井らによる CREST プロジェクト [14]，などがある．[14] では，人間と機械が意味理解を伴ったコミュニケーションに基づき，日常的なタスクを協調して実行するための基盤技術の確立を目指している．また、本分野と関連が深いベンチマークテストとしては，ロボカップ@ホーム [15] がある．ロボカップ@ホームは世界最大の生活支援ロボットのコンペティションであり，日用品の探索，棚からユーザに言われたものを取ってくる，などの移動マニピュレーションとヒューマンロボットインタラクションを統合したタスクが設定されている．

3. Generative Adversarial Networks

GAN は，Generator G および Discriminator D の 2 つの部分から構成される [2]．G の入力は d_z 次元の乱数 $\boldsymbol{z}$ であり，$\boldsymbol{x}_{fake} = G(\boldsymbol{z})$ を出力する．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 $p(\hat{S})$ を出力する．すなわち，

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

である．

D の学習では，以下のコスト関数 $J^{(D)}$ を最小化する．また，G の学習では，$J^{(G)}$ を最小化する．

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

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

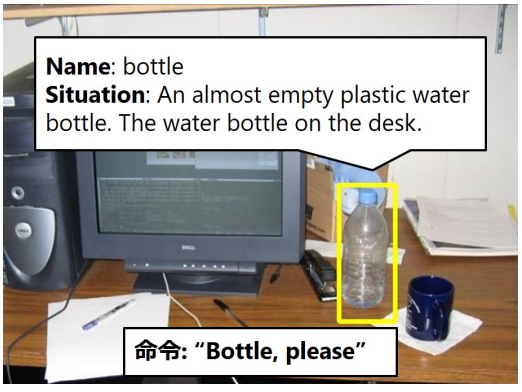

図 1　想定する状況

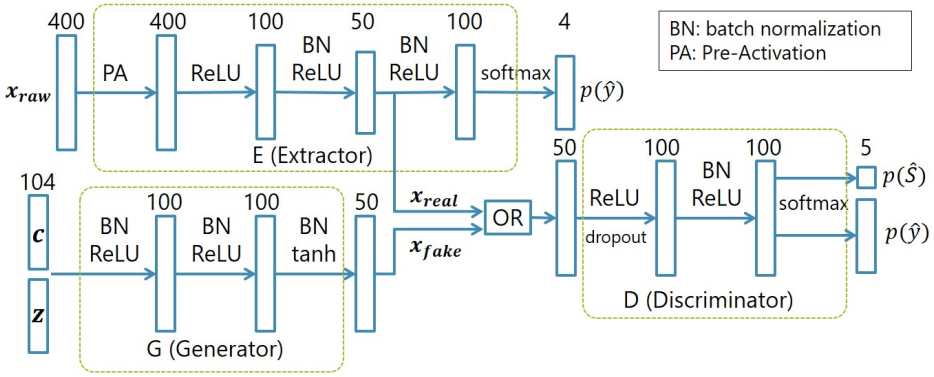

図 2　LAC-GAN の構成．層の上の数字はノード数を示す．

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．

4. Latent Classifier GAN

4.1 構成

以下では，提案手法である LAtent Classifier Generative Adversarial Networks (LAC-GAN) について説明する．LAC-GAN の構造を図 4. に示す．LAC-GAN は，Extractor E，Generator G，Discriminator D，の 3 つの部分から構成される．

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 (0,1) 表現であるものとする．

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．

$$J_C = -\sum_j y_j \log p(\hat{y}_j) \tag{1}$$

ここに，y_j は y の j 次元の (0,1) 表現であり，$p(\hat{y}_j)$ は E の出力層の値である．E のネットワークはボトルネック型であり，最も少ないノード数を有する層の出力を $\boldsymbol{x}_{real}$ として抽出する．

G の入力は，カテゴリ c および $\boldsymbol{z}$ である．c はカテゴリカル分布から生成され，$\boldsymbol{z}$ は適当な分布（標準正規分布や一様分布など）からサンプルされるものとする．G の出力は，$\boldsymbol{x}_{fake}$ である．

D の入力源 S は，通常の GAN と同様である．すなわち，OR ゲートにより，$S = \{real, fake\}$ が選択される．p_S の推定のためのコスト関数は以下である．

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

これは，Conditional GAN [3] と等しい．

以上まとめて，LAC-GAN の学習では，以下のコスト関数を最小化する．

$$J^{(E)}_{lacgan} = J_C \tag{3}$$

$$J^{(D)}_{lacgan} = J_S + \lambda J_C \tag{4}$$

$$J^{(G)}_{lacgan} = -J_S \tag{5}$$

4.2 正則化および活性化関数

Batch Normalization (BN) [16] は，各ミニバッチごとに，入力を平均 0，分散 1 に変換することで学習を安定化する．また，BN は正則化の機能を持つので，dropout の代わりに使用できる．なお，LAC-GAN に限らず，多くの文献では，Discriminator の第 1 層には BN を適用しないことが多い．本論文でも，D の第 1 層には BN を適用せず，dropout を用いた．

通常，BN は重みをかけたのちに適用される．一方，Pre-Activation (PA) は，重みをかける前に入力に BN を適用する手法である [17]．本研究における $\boldsymbol{x}_{raw}$ は，[18] の手法で得られた paragraph vector である．そのままでは標準化されていないので，PA を行うことでバッチ単位の標準化を行う．

活性化関数としては，ReLU，softmax，tanh の 3 種類を用いる．E と D の出力はカテゴリカル変数であるので，最終層では softmax を用いる．また，$\boldsymbol{x}_{raw}$ は正負の実数値であるため，G の最終層では tanh を適用することに注意されたい．その他の層では，ReLU を用いた．ReLU の代わりに leaky ReLU を用いる文献もあるが，本タスクでは ReLU と比較して著しい性能向上は見られなかった．

5. 物体操作マルチモーダルデータセット

物体操作タスクにおける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．データセットを構築するにあたり，ロボットを用いて実験室環境の画像を利用することも可能であるが，本研究ではスケーラビリティを重視し，標準的な画像データセットをベースとしたデータセットを構築するアプローチを採る．本研究では，既存の大規模データセットである Visual Genome dataset [19] から部分集合を抽出し，物体操作タスク用

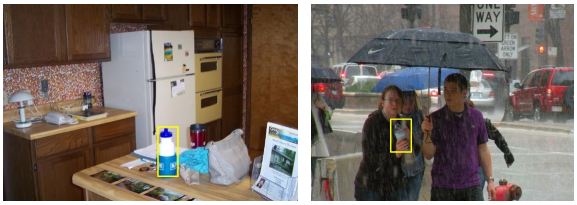

図 3 Synset「bottle.n.01」の例．黄色の矩形はトラジェクタの bounding box を示す．左: 正事例．右：負事例.

のラベルを付与することでデータセットを構築した．以下では，我々が構築した「物体操作マルチモーダルデータセット」について説明する．

Visual Genome データセットは 10 万種類以上の画像を含み，各画像に対し平均 21 個の物体が含まれる．各物体の領域は人手でアノテーションされ，WordNet [20] の synset および言語表現が付与されている．MS-COCO [21] などのデータセットと異なり，Visual Genome データセットには各領域の言語表現が含まれているため，物体操作タスクにおいて状況の言語表現を利用するために都合が良い．

Visual Genome データセットには物体操作以外の画像を多く含まれるため，操作に関係するカテゴリとして以下の synset を含む画像を抽出した．各サンプルには動かされる物体（以下，トラジェクタと呼ぶ）を 1 個だけ含むものとする．

- apple, ball, bottle, can, cellular telephone, cup, glass, paper, remote control, shoe, teddy

ここに,「n.01」は省略した．これらの synset は生活支援ロボットの把持対象として可能性が高い物体カテゴリからランダムに選択した．

各サンプルは以下の基準により，ラベル付けを行った．

(E1) 領域が広すぎるため，トラジェクタと同カテゴリの物体が，領域に複数個含まれる．例：カゴに複数の靴が入っている．

(E2) 領域が狭すぎるため，トラジェクタのうち一部分しか領域に含まれない．例：グラスの持ち手．

(N) トラジェクタが領域に十分に含まれるが，物体操作タスクの対象ではない．つまり，synset が十分に細かくないため，物体操作対象ではない下位カテゴリが含まれてしまっている．例：領域は「ball.n.01」とラベル付けされているが，実際にはミートボールである．本タスクではミートボールの把持は対象としない．

(M0) トラジェクタが領域に十分に含まれるが，操作可能ではない（移動中，障害物に囲まれている，人間が把持している，など）．軌道計画が失敗する可能性が高い状況.

(M1) トラジェクタが領域に十分に含まれ，操作可能である．ただし，ロボットが自律的に動作を実行すると失敗の可能性が高い．遠隔操作であれば実行可能である状況.

(M2) トラジェクタが領域に十分に含まれ，操作可能である．ロボットが自律的に動作を実行しても失敗の可能性が低い状況.

(O) 上記のすべてに当てはまらない.

各ラベルは排他的に与えられる．

データセットをランダムにシャッフルし，学習セット，検証セット，テストセットに分割した．データセットのサイズを表 1 に示す．実験では，(E1)(E2)(O) を除き，(N)(M0)(M1)(M2) の 4 クラス分類問題を扱う．

画像の例を図 3 に示す．各図の状況の言語表現は以下である．

左図 : *“insulated water bottle with sipper top. blue and white water bottle. a blue and white water bottle. the water bottle has a blue top. a set of keys by the ...”*

右図 : *“a bottle in a woman’s hand. plastic bottle of water. bottle of water in woman’s hand. water bottle in the hand. a girl holding a bottle. girl under umbrella holding ...,”* respectively.

紙面の都合上，先頭の 30 語のみを示した.

6. 実験

6.1 設定

本研究では，入力は命令と状況の言語表現であるものとする．ただし，命令に動詞は含まれず，オブジェクト ID のみで与えられるものと仮定する．また，状況の言語表現は，物体操作マルチモーダルデータセットから得られる．

トラジェクタの名称表現から $\boldsymbol{x}_{name}$ を生成し，他のオブジェクトの言語表現から $\boldsymbol{x}_{situation}$ を生成する．これらの言語表現を固定長の paragraph vector で表すために，[18] で提案された手法を用いる．以上から，以下の入力 $\boldsymbol{x}_{raw}$ を得る．

$$\boldsymbol{x}_{raw} = \{\boldsymbol{x}_{name}, \boldsymbol{x}_{situation}\}$$

ここに，$\boldsymbol{x}_{name}, \boldsymbol{x}_{situation}$ は，それぞれ 200 次元の paragraph vector である．

LAC-GAN の設定を表 2 に示す．LAC-GAN の入力である $\boldsymbol{z}$ の次元数 $d_z = 100$ とし，標準正規分布からサンプリングした．ただし，事前実験の結果，$\boldsymbol{z}$ による性能の影響はそれほど大きくないことがわかっている．

6.2 結果

物体操作マルチモーダルデータセットを用いて，提案手法とベースライン手法（AC-GAN）を比較評価した．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 1 実験で用いたデータセットの概要

データセットサイズ (E1-O)	896
状況の言語表現に含まれる語彙数	7926
状況の言語表現に含まれる平均単語数	305
学習セットサイズ (N, M0, M1, M2)	539
検証セットサイズ (N, M0, M1, M2)	67
テストセットサイズ (N, M0, M1, M2)	67

表 2　パラメータ設定

最適化手法	Adam (学習率 0.0005, $\beta_1 = 0.5, \beta_2 = 0.999$)
E input	Name (200)+ Situation (200)
E ノード数	400(in), 400, 100, 50, 100, 4(out)
G ノード数	104(in), 100, 100, 50(out)
D ノード数	50(in), 100, 100, 5(out)
バッチサイズ	50(E), 20(G and D)
λ	0.2

表 3 テストセット精度の比較．各モデルは検証セットにおける最良モデルとした．

	テストセット精度
Baseline (AC-GAN, PA 無)	50.7%
Baseline (AC-GAN, PA 有)	58.2%
Extractor のみ	61.1%
提案手法 (LAC-GAN)	67.1%

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．最良モデルを用いて，テストセットの精度を検証した結果を表 3 に示す．

表 3 において，ベースラインである AC-GAN では，図 4. と同じノード数（入力以外）とした．「PA」の有無は，入力に対して Pre-Activation を行うかどうかを示す．表において，「Extractor のみ」は，Extractor の出力 $p(\hat{y})$ の精度を示す．すなわち，6 層の単純なフィードフォワードネットワークにおける精度を示す．

表 3 より，AC-GAN と比較して，LAC-GAN は高い精度を示した．この結果は，特徴量をそのまま用いる AC-GAN より，特徴抽出を行い，分類に関係が深い特徴のみを用いた方が有利であることを示唆している．これは，G の機能であるサンプル生成により，Discriminator に入力されるサンプル数が擬似的に拡張され，汎化性能に寄与したことが示唆される．

7. Conclusion

本研究では，動詞のない命令文に対し，物体が置かれた状況に基づき物体操作の確からしさを推定する手法を提案した．提案手法である Latent Classifier GAN では，Extractor により分類に適した特徴を抽出し，Generator が疑似サンプルを生成するとともに，Discriminator が分類を行う．提案手法の性能を検証するため，Visual Genome データセットを元に物体操作マルチモーダルデータセットを構築した．提案手法とベースライン手法である AC-GAN を比較し，提案手法が優れることを示した．

謝辞

本研究を進めるにあたり，有益な助言をいただいた Dr. Peng Shen に謝意を表する．本研究の一部は，JST CREST および JSPS 科研費 15K16074 の助成を受けて実施されたものである．

参 考 文 献

[1] 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.

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

[3] M. Mirza and S. Osindero, “Conditional Generative Adversarial Nets,” arXiv preprint arXiv:1411.1784, 2014.

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

[5] A. Odena, C. Olah, and J. Shlens, “Conditional Image Synthesis with Auxiliary Classifier GANs,” arXiv preprint arXiv:1610.09585, 2016.

[6] J.T. Springenberg, “Unsupervised and Semi-Supervised Learning with Categorical Generative Adversarial Networks,” arXiv preprint arXiv:1511.06390, 2015.

[7] P. Shen, X. Lu, S. Li, and H. Kawai, “Conditional Generative Adversarial Nets Classifier for Spoken Language Identification,” Proc. of Interspeech, 2017.

[8] 河原達也，“音声対話システムの進化と淘汰–歴史と最近の技術動向–,” 人工知能学会誌，vol.28，no.1，pp.45–51，2013.

[9] T. Taniguchi, T. Nagai, T. Nakamura, N. Iwahashi, T. Ogata, and H. Asoh, “Symbol Emergence In Robotics: A Survey,” Advanced Robotics, vol.30, no.11-12, pp.706–728, 2016.

[10] T. Kollar, S. Tellex, D. Roy, and N. Roy, “Toward Understanding Natural Language Directions,” Proceeding of the 5th ACM/IEEE international conference on Human-robot interaction, pp.259–266, 2010.

[11] S. Mohan, A. Mininger, J. Kirk, and J.E. Laird, “Learning Grounded Language through Situated Interactive Instruction,” AAAI Fall Symposium: Robots Learning Interactively from Human Teachers, pp.30–37, 2012.

[12] M. Tenorth, A.C. Perzylo, R. Lafrenz, and M. Beetz, “The RoboEarth Language: Representing and Exchanging Knowledge about Actions, Objects, and Environments,” Proc. ICRA, pp.1284–1289, 2012.

[13] A. Saxena, A. Jain, O. Sener, A. Jami, D.K. Misra, and H.S. Koppula, “RoboBrain: Large-Scale Knowledge Engine for Robots,” 2014.

[14] 長井隆行，谷口忠大，尾形哲也，岩橋直人，杉浦孔明，稲邑哲也，岡田浩之，“記号創発ロボティクスによる人間機械コラボレーション基盤創成,” 第 19 回クラウドネットワークロボット研究会，pp.23–27，2015.

[15] L. Iocchi, D. Holz, J. Ruiz-delSolar, K. Sugiura, and T. van derZant, “RoboCup@Home: Analysis and Results of Evolving Competitions for Domestic and Service Robots,” Artificial Intelligence, vol.229, pp.258–281, 2015.

[16] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” Prof. of ICML, pp.448–456, 2015.

[17] K. He, X. Zhang, S. Ren, and J. Sun, “Identity Mappings in Deep Residual Networks,” Proc. of European Conference on Computer Vision, pp.630–645, 2016.

[18] Q. Le and T. Mikolov, “Distributed Representations of Sentences and Documents,” Proc. of ICML, pp.1188–1196, 2014.

[19] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D.A. Shamma, et al., “Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations,” arXiv:1602.07332, 2016.

[20] G.A. Miller, et al., “WordNet: a Lexical Database for English,” Communications of the ACM, vol.38, no.11, pp.39–41, 1995.

[21] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C.L. Zitnick, “Microsoft COCO: Common Objects in Context,” European Conference on Computer Vision, pp.740–755, 2014.