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

共有表現の学習によるロボット動作と指示説明文の双方向変換

Bidirectional Translation between Robot Actions and Linguistic Descriptions by Aligned Representations

山田竜郎 *1*2 Tatsuro Yamada　松永寛之 *1 Hiroyuki Matsunaga　尾形哲也 *1 Tetsuya Ogata

*1早稲田大学　理工学術院 Faculty of Science and Engineering, Waseda University

*2日本学術振興会　特別研究員 Research Fellow of Japan Society for the Promotion of Science

We propose a novel deep learning framework to bidirectionally translate between robot actions and their linguistic descriptions. The model consists of two recurrent autoencoders, one of which is employed to make vector representations of robot actions and the other is for descriptions. The learning algorithm produces representations shared between actions and their descriptions by creating an additional loss function in which the representation of an action and that of its description become close to each other in the latent vector space. Across the shared representation, the trained model can produce a linguistic description given a robot action. The model is also able to generate an appropriate action by receiving a linguistic instruction, conditioned on the current visual input. Visualization of the shared representations shows that the robot actions are represented in a semantically compositional way in the vector space by being learned jointly with their descriptions.

1. はじめに

人間と協調して働くロボットは，言語的にコミュニケーションをとりつつ働くことが求められる．一方では，人間の言語指示に応えて正しい動作を生成する能力が必要である．また他方で，ロボット自身の生成した動作を文章で説明する能力も，挙動の解釈可能性の観点から鑑みて必要である．

近年，シーケンス間の対応を学習するシステムとしてエンコーダ–デコーダモデル（あるいは Sequence-to-sequence モデル）が注目されている．翻訳 [Sutskever 14]，音声認識，言語指示からロボット動作への変換 [Yamada 17]，ロボット動作から説明文への変換 [Heinrich 14] など多岐に用いられるが，これらはすべて一方向変換のみを実現する．本研究は，並列する二つのエンコーダ–デコーダモデルを用意し，その中間表現を共有するように学習を行うというアプローチにより，ロボット動作と指示説明文を双方向に変換可能なモデルを提案する．

2. 提案手法

2.1 モデル概要

本研究は，ロボットが (1) 言語指示に応えて動作を生成する能力と，(2) 自身の動作からその説明文を生成する能力の，双方を対データから学習するニューラルネットワークモデルおよび学習アルゴリズムを提案する．図 1 に示すように，本モデルは 2 つの Recurrent Autoencoder (RAE) [Srivastava 15] からなる．一方が単語シーケンス，すなわち文章を，もう一方がロボットの動作シーケンスを扱う．RAE は，エンコーダ–デコーダモデルの一種であり，出力シーケンスが入力シーケンスと同じになるように，すなわち恒等写像を目的として学習を行う．学習により RAE は，シーケンスを表す固定長ベクトル表現を，そのシーケンスを再生成可能な形で獲得する．

今回の提案モデルでは，各 RAE における復元誤差に加え，互いに対となる文章と動作の表現同士が近くなるように拘束する損失関数を設ける．これら二つの学習により，上記の二つの機能が実現することが期待される．(1) の機能は，指示文を言語 RAE がエンコードし，その表現を動作 RAE がデコードすることで実現する．(2) は反対に，動作 RAE のエンコーダ，言語 RAE のデコーダと順に順伝播計算を行うことで実現する．

連絡先: 尾形哲也，早稲田大学理工学術院，ogata@waseda.jp

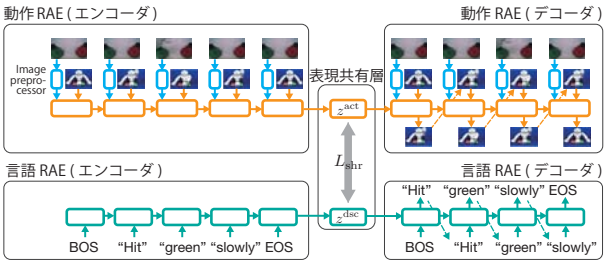

図 1: 双方向変換モデルの概要．

2.2 言語 RAE

本節では言語 RAE (図 1 下部) の詳細を述べる．エンコーダ RNN は式 (1), (2) によって，単語シーケンス $(x_1, x_2, ..., x_T)$ を固定長ベクトル z に埋め込む．

$$h_t^{\text{enc}} = \text{EncCell}(x_t, h_{t-1}^{\text{enc}}) \quad (1 \le t \le T), \tag{1}$$

$$z = W^{\text{enc}} h_T^{\text{enc}} + b^{\text{enc}}. \tag{2}$$

EncCell は，GRU や LSTM [Gers 00] などの任意の学習可能な再帰セルである．h_t^{enc} は時刻 t におけるエンコーダのセル状態で，h_0^{enc} は零ベクトルとする．W^{enc} と b^{enc} は学習可能な重みとバイアスである．エンコーディング後，デコーダ RNN は式 (3)-(5) によって表現 z を $(y_1, y_2, ..., y_{T-1})$ として展開する．

$$h_0^{\text{dec}} = W^{\text{dec}} z + b^{\text{dec}}, \tag{3}$$

$$h_s^{\text{dec}} = \text{DecCell}(y_{s-1}, h_{s-1}^{\text{dec}}) \quad (1 \le s \le T-1), \tag{4}$$

$$y_s = f(W^{\text{out}} h_s^{\text{dec}} + b^{\text{out}}) \quad (1 \le s \le T-1). \tag{5}$$

DecCell は EncCell 同様，任意の再帰セルであり，h_s^{dec} は時刻 s におけるデコーダのセル状態である．W^{out} と b^{out} は，出力層の重みとバイアスである．本研究では，単語シーケンスを 1-hot ベクトルの系列として表現するため，活性化関数 f とし

て softmax 関数を選ぶ．ここで，1-hot ベクトルは単語数と同じ次元数のベクトルであり，単語に対応する要素のみが 1 をとり，他は 0 となる表現形式のことである．y_0 には，文章開始シンボルを表す 1-hot ベクトルを与える．学習は入力と出力の交差エントロピー最小化として行われる．

$$L_{\mathrm{dsc}} = \frac{1}{T-1}\sum_{t=1}^{T-1}\Big(-\sum_{w=1}^{W} x_{t+1}(w)\log y_t(w)\Big). \tag{6}$$

W は語彙数である．

2.3 ロボット動作 RAE

次にロボット動作 RAE の挙動について説明する．動作シーケンスはロボットの関節角度 $(j_1, j_2, ..., j_{T'})$ とそれに同行する視覚情報 $(v_1, v_2, ..., v_{T'})$ からなり，動作 RAE はこれらを連結した $((j_1; v_1), (j_2; v_2), ..., (j_{T'}; v_{T'}))$ を固定長ベクトルに埋め込む．モデルはほぼ言語 RAE と同様であるが，デコーダの挙動が一部異なる．式 (7), (8) が示すように，デコーダが予測するのは関節角度だけであり，視覚情報は外部的に与えられる．

$$h_s^{\mathrm{dec}} = \mathrm{DecCell}(v_s, y_{s-1}, h_{s-1}^{\mathrm{dec}}) \quad (1 \le s \le S'), \tag{7}$$

$$y_s = f(W^{\mathrm{out}} h_s^{\mathrm{dec}} + b^{\mathrm{out}}) \quad (1 \le s \le S'). \tag{8}$$

v_s が時刻 s における視覚入力であり，y_s は時刻 $s+1$ における関節角度 j_{s+1} を予測する．y_0 には初期姿勢 j_1 を与える．行動用 RAE の誤差関数としては，二乗誤差を使用する．

$$L_{\mathrm{act}} = \frac{1}{T-1}\sum_{t=1}^{T-1} ||j_{t+1} - y_t||_2^2. \tag{9}$$

このようなモデルとしたのは，本研究が，ロボットが生成するべき行動の状況依存性を扱うことを目的としているからである．以下，具体例を用いて説明する．例えば，“赤いベルを叩く”という文章に対応する動作シーケンスは，赤いベルがどこにあるかによって変化する．次節で見るように，これらの多様でありうる動作シーケンス群の表現は，すべて “赤いベルを叩く”という文章の表現と拘束される．その結果，これらのシーケンス群はそれぞれ全く異なるシーケンスとしてではなく，同一の意味を持ったシーケンスとして互いに近接した空間に埋め込まれることになる．それゆえ，動作 RAE のデコーダは，この意味的に埋め込まれた表現と，視覚入力により与えられる現在の状況から，正しい行動を生成することを求められる．

2.4 表現の共有

対となる文章と動作のシーケンスの表現を互いに近づけるために，二つの RAE の復元誤差に加え，式 (10) で表される損失関数を設ける．ここで，動作シーケンスの表現のバッチを $\{z_i^{\mathrm{act}} | 1 \le i \le N\}$，対応する文章の表現のバッチを $\{z_i^{\mathrm{dsc}} | 1 \le i \le N\}$ と表記する．N はバッチサイズである．

$$L_{\mathrm{shr}} = \sum_{i}^{N} \psi(z_i^{\mathrm{act}}, z_i^{\mathrm{dsc}}) + \sum_{i}^{N}\sum_{j\neq i} max\{0, \Delta + \psi(z_i^{\mathrm{act}}, z_i^{\mathrm{dsc}}) - \psi(z_i^{\mathrm{act}}, z_j^{\mathrm{dsc}})\} \tag{10}$$

関数 ψ は二変数間のユークリッド距離である．式 (10) の右辺第 1 項は，対となる表現を互いに近づけることを意味する．第 2 項は対でないシーケンス間の距離を互いに遠ざけることを意味する，ただし，対となる表現同士の距離より Δ 以上遠く離れているとき，この力は働かない（この条件を与えない場合，距離が無限大に発散してしまうことが，予備実験で確認された）．

動作名	動作の内容
PUSH-L-SLOW	左の物体をゆっくり押す
PUSH-L-FAST	左の物体を素早く押す
PUSH-R-SLOW	右の物体をゆっくり押す
PUSH-R-FAST	右の物体を素早く押す
PULL-L-SLOW	左の物体をゆっくり引く
PULL-L-FAST	左の物体を素早く引く
PULL-R-SLOW	右の物体をゆっくり引く
PULL-R-FAST	右の物体を素早く引く
SLIDE-L-SLOW	左の物体をゆっくり右に動かす
SLIDE-L-FAST	左の物体を素早く右に動か
SLIDE-R-SLOW	右の物体をゆっくり左に動かす
SLIDE-R-FAST	右の物体を素早く左に動かす

図 2: （左）実験環境と使用したロボット．（右）動作一覧．

2.5 学習方法

学習はランダムバッチを用いた勾配法によって行われる．全体の損失関数は式 (11) によって表される．

$$L_{\mathrm{all}} = \alpha L_{\mathrm{dsc}} + \beta L_{\mathrm{act}} + \gamma L_{\mathrm{shr}}. \tag{11}$$

α, β, γ は各損失関数の影響度合いを調整するハイパーパラメータである．全ての学習パラメータにおける勾配は，Backpropagation Through Time [Rumelhart 86] により計算される．

3. 実験デザイン

3.1 タスク概要

提案手法の有効性を評価するために，ロボット実験を行なった．実験用ロボットとして，小型のヒューマノイドロボット NAO を用いた．ロボットの前には色付きの立方体オブジェクトが 2 つ決められた位置に置かれる (図 2)．オブジェクトの色は赤，緑，黄のいずれかであり，2 つの色は必ず異なるようにするため，可能なオブジェクト配置は ${}_3P_2 = 6$ 通りである．各配置で，ロボットは図 2 に示す 12 通りの動作を行うことができる．それぞれの動作に対応する文章は，オブジェクトの配置によって異なるように設定した．具体的に説明すると，文章は動詞，目的語，副詞の 3 単語からなり，動詞と副詞はオブジェクト配置に依存せず動作のみによって決まるが，目的語部分には，動作の対象となったオブジェクトの色を示す語が入る．例えば左に赤，右に緑と置かれた場合における PUSH-L-SLOWLY の動作に対応する文章は，“push red slowly”である．すなわち可能な文章は，(push, pull, slide)*(red, green, yellow)*(slowly, fast) の 18 通りとなる．

3.2 データ

ロボットの動作軌道はあらかじめコンピュータ上で作成した．FAST 動作群は約 26 ステップ，SLOWLY 動作群は約 39 ステップとなるように動作を設計した．可能な 6 通りのオブジェクト配置で，12 パターンの動作を実際にロボットに生成させ，その際のロボットの腕の関節角度 10 次元，カメラ画像 (H: 120, W: 160, RGB) を収集した．6 × 12=72 パターンの可能な状況それぞれを，6 回収集した．画像は，別に用意した Autoencoder によってあらかじめ 10 次元のベクトルとして圧縮し，これをモデルの視覚入力として扱う．

対応する文章は，1-hot ベクトルの系列として表現した．文章の最初と最後には，開始終端シンボルを挿入した．

3.3 学習

72 パターンの可能な状況のうち，18 パターンを除外して学習を行った．言語 RAE のエンコーダとデコーダはそれぞれ 100 ノード 1 層の LSTM，動作 RAE のエンコーダとデコーダはそれぞれ 100 ノード 2 層の LSTM であり，共有層は 100 次元，損失関数の混合率である α, β, γ はすべて 1.0 とした．オプティマイザを Adam [Kingma 15] (学習率 0.001)，バッチサイズを 50 として，20,000 回の学習を行った．

4. 結果

4.1 説明文の生成

学習後の動作 RAE のエンコーダに動作シーケンスを入力して，その表現を元に言語 RAE のデコーダを計算することで，正しい文章が生成されるかを評価した．言語 RAE の出力層の活性化関数は softmax であるが，もっとも高い値をとった要素に対応する単語を，各時刻におけるモデルの出力と見なした．モデルは，既学習の 54 パターンと未学習の 18 パターンすべてにおいて正しい文章を生成することができた．

4.2 ロボット行動の生成

言語 RAE のエンコーダに文章を入力して，その表現を元に動作 RAE のデコーダを計算することで正しい動作が生成されるか，ロボットを使用して評価した．目的語の示すオブジェクトを，動詞が指示する方向に 3cm 以上動かすことができた場合成功と見なした．また，30 ステップ以内に初期姿勢近傍に戻った場合は FAST 動作，それ以上かかった場合は SLOWLY 動作を生成したと見なした．この条件で，既学習状況については 54 パターン中 36 パターン，未学習状況については 18 パターン中 12 パターンの動作生成に成功した．

失敗したケースについても，明らかに指示された動作と違う軌道を生成していたわけではなく，軌道はほぼ正しいが，わずかにずれたことでオブジェクトの正しい面に触れることができず，オブジェクトが指示された方向に動かなかったというケースがほとんどであった．そこで時系列の類似度を測る指標である Dynamic Time Warping [Senin 08] によって，失敗した際の軌道と，元の 12 パターンの動作の軌道との類似度を計算したところ，全てのケースにおいて，生成された軌道は正解の動作の軌道との類似度がもっとも高いということが分かった．

4.3 共有表現の解析

学習後の共有空間におけるシーケンスの表現を主成分分析によって可視化した．図 3 左は，18 本の文章の表現である．動詞，目的語，副詞のそれぞれの品詞について，体系的にエンコードされていることがわかる．一方，図 3 右は，72 パターンの動作シーケンスの表現であるが，文章の表現と正しく対応づけられていることがわかる．ここで注意されたいのは，オブジェクト配置によって，同じ動作でも違う文章に対応づけられ，反対に，異なる動作でも同じ文章に対応づけられていることである．すなわち，動作シーケンスは，視覚情報と総合することで意味的な行動の表現としてエンコードされた．

5. まとめと展望

本研究ではロボット動作とそれに対応する指示説明文を双方向に変換可能なエンコーダ–デコーダ型の RNN モデルを構築し，その有効性を評価するためロボット実験を行なった．提案モデルは与えられたタスクにおいて，動作と指示説明文の双方向の変換に成功した．また共有空間の可視化により，動作シーケンスが，言語表現と意味的に結び付けられていることが明らかとなった．今後は，より複雑なタスク下でのモデル評価や，聴覚・触覚情報も統合したモデルへの拡張を行う予定である．

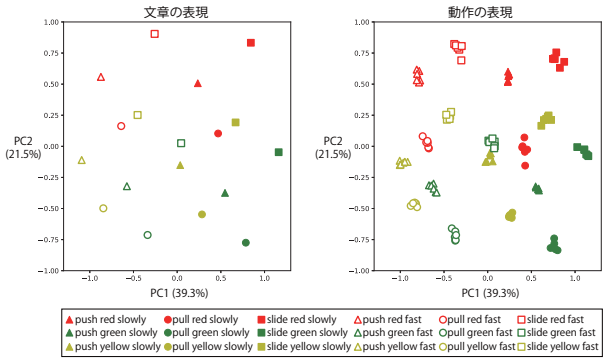

図 3: 主成分分析による文章と動作の表現の可視化．

謝辞

本研究は，科研費特別研究員奨励費 (17J10580)，JST, CREST (JPMJCR15E3)，文部科学省博士課程教育リーディングプログラム「実体情報学博士プログラム」の助成を受けた．

参考文献

[Gers 00] Gers, F. A., and Schmidhuber, J.: Recurrent nets that time and count. In Neural Networks, *Proceedings of the IEEE-INNS-ENNS Int. Joint Conf. on Neural Networks (IJCNN2000)*, Vol. 3, pp. 189-194 (2000).

[Heinrich 14] Heinrich, S. and Wermter, S.: Interactive Language Understanding with Multiple Timescale Recurrent Neural Networks, *Artificial Neural Networks and Machine Learning — ICANN 2014*, LNCS Vol. 8681, pp. 193-200 (2014).

[Kingma 15] Kingma, D., and Ba, J.: Adam: a method for stochastic optimization, in *International Conference on Learning Representations (ICLR2015)*, (2015).

[Rumelhart 86] Rumelhart, D. E., Hinton, G. E., and Williams, R. J.: Learning internal representations by error propagation, in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (MIT Press), pp. 318-362 (1986).

[Senin 08] Senin, P.: Dynamic time warping algorithm review." *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA* 855, pp. 1-23 (2008).

[Srivastava 15] Srivastava, N., Mansimov, E., and Salakhudinov, R.: Unsupervised Learning of Video Representations using LSTMs, *International conference on machine learning*, (2015).

[Sutskever 14] Sutskever, I., Vinyals, O., and Le, V. Q.: Sequence to Sequence Learning with Neural Networks, in *Neural Information Processing Systems 2014 (NIPS2014)* (2014).

[Yamada 17] Yamada, T., Murata, S., Arie, H., and Ogata, T.: Representation Learning of Logic Words by an RNN: From Word Sequences to Robot Actions, *Frontiers in Neurorobotics*, Vol. 11, No. 70, pp. 1-18 (2017).