ロボットの概念・語意学習における言語モデルと画像特徴抽出の教師なし学習

中村 友昭
電気通信大学
tnakamura@uec.ac.jp

1 はじめに

これまで我々のグループでは，ロボットが取得可能な情報をクラスタリングすることで得られるカテゴリが概念であると考え，Latent Dirichlet Allocation (LDA)[1] を拡張した Multimodal LDA (MLDA) を用いた教師なしクラスタリングによりロボットによる概念学習手法を提案してきた [2, 3, 4, 5, 6, 7]．さらに，人から与えられる言語もクラスタリングすることで，単語と概念が確率的に結びつき，語意の獲得も可能となる [3, 5, 6]．しかし，これらの研究では，人の音声を音声認識器を用いて文字列へと変換し，ロボットが取得する画像を学習済みの Convolutional Neural Network (CNN) を用いて画像特徴量へと変換していた．すなわち，教師なしでの学習を考えているのにも関わらず，学習済みの言語モデルと CNN を用いているという問題があった．そこで本稿では，これらの問題を解決するための手法である，言語モデルと画像特徴抽出の教師なし学習手法に関して紹介する．

2 マルチモーダル情報

マルチモーダル情報として，図 1 のロボットが物体を観察し得られる画像（視覚情報）と，把持することで得られる触覚センサーの値（触覚情報）と，振ることで発生する音（聴覚情報）を使用した．物体として，図 2 の 499 個の身の回りにある物体を用いた．さらに，ロボットが物体からマルチモーダル情報を取得している際に，人がその物体の特徴を音声にて教示した．教示発話は音声認識し単語列へと変換し，画像からは特徴抽出を行い，それらを MLDA によってクラスタリングすることで概念・語意を学習する．

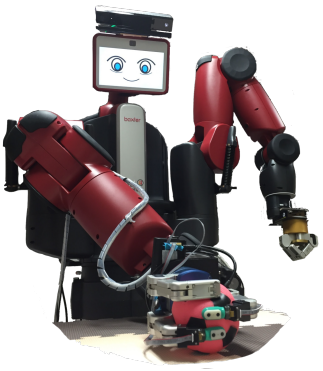

図 1: 実験に用いたロボット

図 2: 実験に用いた物体

3 概念と言語モデルの相互学習

図 3 は，言語モデルと物体概念を統合したモデル [6] であり，白色のノードは未観測情報，灰色のノードは観測情報を表している．図中の $\boldsymbol{o}$ が人から教示される音声である．この音声を，$\mathcal{A}$ をパラメータとする音響モデル，$\mathcal{L}$ をパラメータとする言語モデルにより認識した結果が s である．さらに，認識結果 s を言語モデル $\mathcal{L}$ を用いて単語へ分割し，Bag of words (BoW) 表現へと変換したものが単語情報 w^w である．また，w^v，w^a，w^t はそれぞれ物体から得られる視覚情報，聴覚

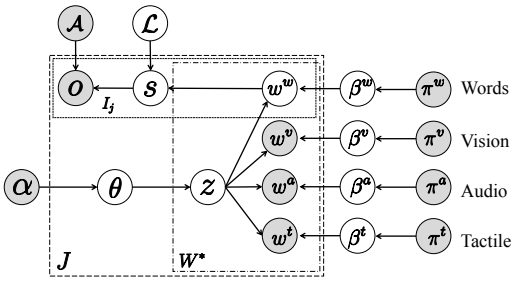

図 3: 物体概念と言語モデルの相互学習モデル

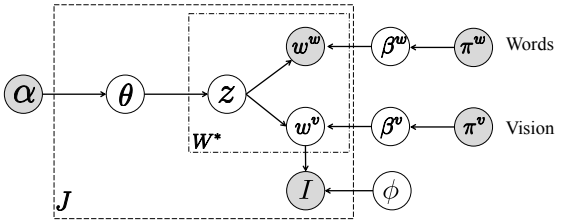

図 4: 物体概念と画像特徴抽出の相互学習モデル

情報，触覚情報を示している．これらのマルチモーダル情報は，それぞれ β^* をパラメータとする多項分布から生成される．π^* はディリクレ事前分布のパラメータである．また，z は物体のカテゴリを表しており，カテゴリ z の出現確率分布を表す多項分布のパラメータを θ とする．α はディリクレ事前分布のパラメータである．J, W^*, I_j は，それぞれ物体数，モダリティ* の情報の生起回数，j 番目の物体への教示発話数を表している．このモデルでは，音声認識結果 s と物体カテゴリ z が単語 w^w によって接続されている．そのため，物体カテゴリと認識文字列や言語モデルが相互に影響するモデルとなっており，物体カテゴリ z だけでなく，言語モデルのパラメータ $\mathcal{L}$ も同時に学習可能なモデルである．

このモデルを，前述のマルチモーダル情報を用いて学習した．ただし視覚情報として，画像を学習済みの CNN[1] に入力して得られる中間層の出力を用いた．その結果，物体の分類精度は言語モデルの学習をしない場合が 38.1%となったのに対し，言語モデルを学習することで 61.7%となった．さらに，音声認識精度は，言語モデルの学習をしない場合は 67.2%となったのに対し，言語モデルを学習することで 73.3%となった．このように，物体概念と言語モデルを相互に学習することで，双方の精度を向上させることができている．

4 概念と画像特徴抽出の相互学習

前章のモデルによって，言語モデルを学習により獲得できることが示された．しかし，視覚情報 w^v として，学習済みの CNN の中間層の出力を特徴量として用いていた．そこで本章では，この特徴抽出も教師なしで学習可能なモデルを提案する．図 4 が提案モデルのグラフィカルモデルである．本実験では，触覚情報と聴覚情報は用いずに，視覚情報と単語情報のみを利用した．このモデルでは，ϕ をパラメータとする Variational Autoencoder (VAE) [8] のエンコーダーを用いて，画像 I を画像特徴量 w^v へと変換する．

提案モデルの学習では，MLDA の学習と VAE の学習を交互に行う．まず，MLDA によって w^w と w^v をクラスタリングし，MLDA のパラメータを Θ を計算する．次に，MLDA により他のモダリティから画像特徴量が生成される確率 $p(w^v|w^w, \Theta)$ を計算し，以下の損失関数を最小化するエンコーダーとデコーダーのパラメータ ϕ, ρ を学習する．

$$loss = -E_{q_\phi(w^v|I)}[\log p_\rho(I|w^v)] + \gamma D_{KL}(q_\phi(w^v|I)||p(w^v|w^w, \Theta)) \quad (1)$$

MLDA と VAE の学習を交互に繰り返すことによって，分類に適した特徴量 w^v が学習される．

実際に，2 章の単語情報と画像を用いて提案モデルの学習を行った．潜在変数 w^v は 128 次元とした．その結果，学習済みの CNN[1] を利用した場合の分類精度が 66.8%だったのに対して，画像特徴量を教師なしで学習した場合の分類精度は 66.1%となった．すなわち，教師なし学習によって，大量のデータから学習した CNN を用いた場合と近い性能を得ることができた．さらに 128 次元の潜在変数を，t-SNE で 2 次元に圧縮し可視化した．その結果が図 5 である．各点の色が，物体の正解カテゴリと対応した色となっている．図 5(a) は VAE 単体で潜在変数を学習した結果であり，図 5(b) は MLDA と VAE を相互に学習した結果である．この図から，相互学習により，分類に適した潜在変数が学習できていることが分かる．

[1] ImageNet2012 で学習した CaffeNet を利用した

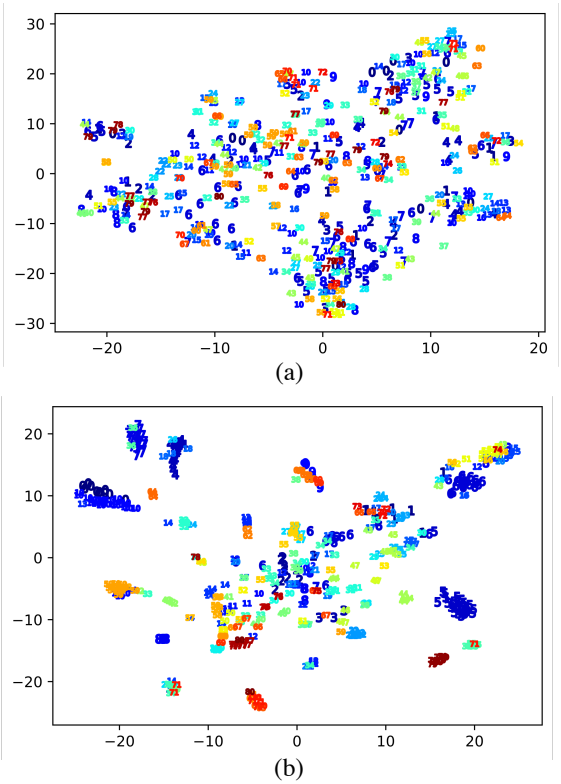

図 5: 学習された潜在変数：(a) VAE のみを用いて得られた潜在変数，(b) VAE と MLDA の相互学習によって得られた潜在変数

5 おわりに

本稿では，概念学習と同時に，言語モデルと画像特徴抽出を教師なしで学習する手法を紹介した．これらのモデルでは，マルチモーダル情報が相補的に影響し合うことで，カテゴリ分類精度だけでなく，言語モデルや画像特徴抽出の性能が向上した．本稿の実験では，言語モデルと画像特徴抽出をそれぞれ独立して学習したが，今後はこれら 2 つを同時に学習することを考えている．

参考文献

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.

[2] Tomoaki Nakamura, Takayuki Nagai, and Naoto Iwahashi. Multimodal Categorization by Hierarchical Dirichlet Process. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1520–1525, 2011.

[3] Tomoaki Nakamura, Takaya Araki, Takayuki Nagai, and Naoto Iwahashi. Grounding of word meanings in lda-based multimodal concepts. *Advanced Robotics*, Vol. 25, pp. 2189–2206, 2012.

[4] Yoshiki Ando, Tomoaki Nakamura, Takeshi Araki, and Takayuki Nagai. Formation of hierarchical object concept using hierarchical latent dirichlet allocation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2272–2279, 2013.

[5] Muhammad Attamimi, Yuji Ando, Tomoaki Nakamura, Takayuki Nagai, Daichi Mochihashi, Ichiro Kobayashi, and Hideki Asoh. Learning word meanings and grammar for verbalization of daily life activities using multilayered multimodal latent dirichlet allocation and bayesian hidden markov models. *Advanced Robotics*, Vol. 30, No. 11-12, pp. 806–824, 2016.

[6] Joe Nishihara, Tomoaki Nakamura, and Takayuki Nagai. Online algorithm for robots to learn object concepts and language model. *IEEE Transactions on Cognitive and Developmental Systems*, Vol. 9, No. 3, pp. 255–268, 2017.

[7] Tomoaki Nakamura and Takayuki Nagai. Ensemble-of-concept models for unsupervised formation of multiple categories. *IEEE Transactions on Cognitive and Developmental Systems*, Vol. 10, No. 4, pp. 1043–1057, 2018.

[8] Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models. *International Conference on Learning Representations*, 2017.