Vector Quantization に基づく 離散系列の発話による分散型深層モデルの提案

三好遼 [1]　　栗田修平 [2,3]
[1] フリー　[2] 国立情報学研究所　[3]NII LLMC
miyoshi2020robotcs@gmail.com　　skurita@nii.ac.jp

概要

近年，記号創発ロボティクスでは，ベイズ脳仮説を始めとした統合情報理論に基づいたマルチモーダル処理や環境との相互作用による概念獲得などの様々な研究が進んでいる．中でも言語創発は，近年注目の研究であり，分散的ベイズ学習によるコミュニケーションモデルが提案されている．そこで本提案手法では，深層モデルへの応用を検討するため，ベクトル量子化を拡張した VQCom-VAE を提案し，実験にてコミュニケーションによる相手の発話から予測画像の生成が可能であることを示した．

1　はじめに

近年，記号創発ロボティクスでは，ベイズ脳仮説 [1] に基づいた分散的ベイズ学習による言語創発の研究が行われてきた [2, 3, 4, 5].

文献 [2] では，メトロポリス・ヘイスティングス法を拡張したメトロポリス・ヘイスティングス名付けゲームによる分散的ベイズ学習を提案している．具体的には，ベイズモデルである Gaussian Mixture Model（GMM）と深層ベイズモデルである Variational Auto-Encoder（VAE）を Serket [6] を使用して統合する．統合したモデルを Fruits-360 データセット [7]（果物の異なる視点画像）を使用して GMM でカテゴリを学習する．カテゴリを発話とし，学習済みモデル（エージェント A）と未学習モデル（エージェント B）をメトロポリス・ヘイスティングス名付けゲームでコミュニケーション学習する．その結果，異なる視点の入力画像に対しても発話の一致率が約 8 割の性能となり，メトロポリス・ヘイスティングス名付けゲームによる記号創発現象を示した．

さらに，文献 [5] では，マルチモーダル情報を Multimodal Latent Dirichlet Allocation（MLDA）[8] で統合し，潜在変数から Gaussian Process Hidden Semi-Markov Model（GP-HSMM）[9] による単語パターンの推論と連続記号の生成をメトロポリス・ヘイスティングス名付けゲームで学習することで，連続記号から分節化に基づいた言語創発を可能にした．

しかし，これらの文献では，MCMC 法による学習が一般的であり，勾配降下法を使用する深層モデルへの応用は，困難である．

そこで本稿では，Vector Quantized Variational Auto-Encoder（VQ-VAE）[10] を分散型深層モデルに拡張した Vector Quantized Communication Variational Auto-Encoder（VQCom-VAE）を提案する．本提案手法では，エージェント A のエンコーダの出力 z から量子化した離散系列の発話 m^A を生成し，エージェント B に送信する．エージェント B は受信した発話 m^A を埋め込み空間 E を通して，デコーダから予測画像 $\hat{o}^B$ を生成することで，コミュニケーションに基づいた画像生成が可能である．また，コミュニケーションは，エージェント A から B，エージェント B から A の双方向による予測も可能である．学習では，ミニバッチ t におけるエージェント A の発話 m^A を含めてエージェント B の埋め込み空間 E_t^B を更新することで予測画像 $\hat{o}^B$ と発話 m^A との関連付け学習が可能である．さらに，重み係数 α^* を導入し，エージェント間の発話における学習の影響度合いを表現する．

実験では，MNIST データセット [11] を使用し，2 者エージェントによる発話 m^A, m^B に基づいた数字画像の予測が可能であることを検証した．その結果，提案手法によるエージェント A の発話 m^A に基づいた予測画像 $\hat{o}^B$ をエージェント B で生成可能であることを示した．さらに，重み係数 α を変化させたコミュニケーションでは，相手の発話に対する重み係数を低くすることで，コミュニケーションが通じなくなることを示した．

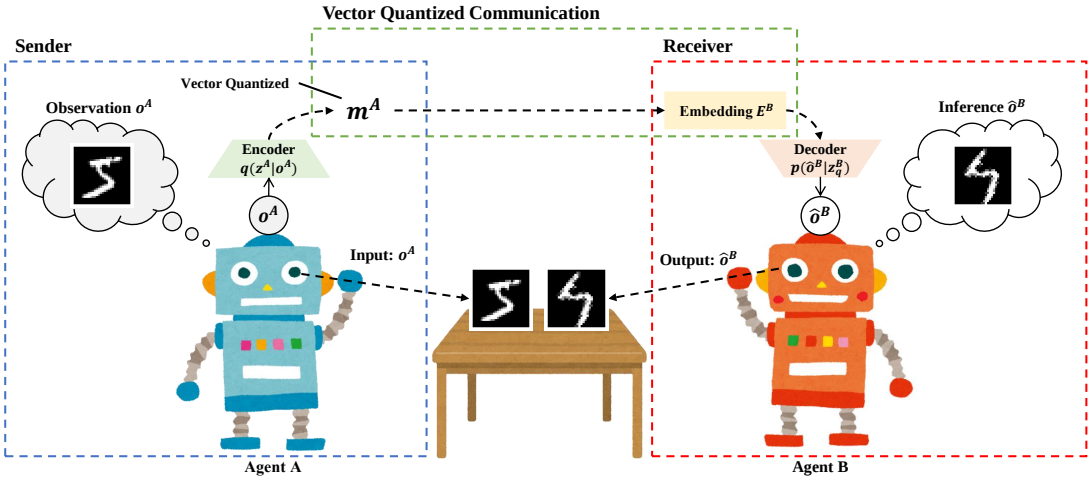

図 1　提案手法の概要

2　提案手法：VQCom-VAE

本提案手法の概要図を図 1 に示した．本提案手法では，VQ 層を拡張し，双方向のコミュニケーションを可能とした分散型深層モデルである．節 2.1 では，VQCom-VAE におけるエンコーダとデコーダの学習について述べる．節 2.2 と節 2.3 では，VQ 層を拡張した手法による発話 m の生成および埋め込み空間 E の更新について述べる．

2.1　エンコーダとデコーダの学習

まず，入力画像 o をガウス分布を仮定したエンコーダ $q(z|o)$ に入力し，潜在変数 $z \in \mathcal{R}^{D\times W\times H}$ を出力する．

$$z \sim q(z|o) \tag{1}$$

次元数 D は，潜在変数の次元数であり，次元数 W，H は Convolutional Neural Network(CNN) 層で入力画像を畳み込んだ特徴マップにおける縦横のピクセル数である．

エンコーダ $q(z|o)$ の出力 z とカテゴリ数を K とする埋め込み空間 E の重みベクトル $e \in \mathcal{R}^{K\times D}$ を用いて量子化する．この量子化した発話 $m \in \mathcal{N}^{W\times H}$ から特徴 $z_q \in \mathcal{R}^{D\times W\times H}$ を出力する．

特徴 z_q をガウス分布を仮定したデコーダ $p(\hat{o}|z_q)$ から予測画像 $\hat{o}$ を生成する．

$$\hat{o} \sim p(\hat{o}|z_q) \tag{2}$$

式 (3) の損失関数 $\mathcal{L}$ を最小化するようなパラメータ Θ を学習する．第一項は，再構成誤差と呼ばれる項であり，入力画像 o と予測画像 $\hat{o}$ との誤差を小さくすることで入力画像 o に近い予測画像 $\hat{o}$ を生成するように学習する．第二項は，コミットメント損失と呼ばれる項であり，埋め込み空間 E の出力 z_q を潜在変数 z に近づけるための項である．第三項は，ベクトル量子化損失と呼ばれる項であり，潜在変数 z を埋め込み空間 E の出力 z_q に近づけるための項である．第二項，第三項における $\mathrm{sg}[\cdot]$ は，勾配停止 (stop gradient) と呼ばれる操作であり，埋め込み空間 E の直接的な勾配伝達の停止を意味する．また，第二項における β は定数であり，一般的に $\beta = 0.25$ と設計されることが多く，適切な値に設定することで損失関数を調整する．

$$\mathcal{L}_\Theta = \|o - \hat{o}\|_2^2 + \beta\|\mathrm{sg}[z] - z_q\|_2^2 + \|z^A - \mathrm{sg}[z_q]\|_2^2 \tag{3}$$

2.2　各エージェントの発話 m の送信

まず，Gumbel-softmax sampling [12] に基づき，ガンベル分布に従う乱数 $g_i \in \mathcal{R}^{K\times W\times H}$ を生成する．節 2.1 のエンコーダ $q(z|o)$ の出力 z_i と乱数 g_i，埋め込み空間 E の重みベクトル e_i を式 (4) に代入し，カテゴリ数 K のカテゴリカル分布 $p(\pi|z_i)$ を計算する．

$$p(\pi|z_i) = \frac{\exp\left((-\|z_i - e_i\|_2^2 + g_i)/\tau\right)}{\sum_{d=0}^{D-1}\exp\left((-\|z_d - e_d\|_2^2 + g_d)/\tau\right)} \tag{4}$$

定数 τ は温度定数であり，値が高いほど確率が均等な分布に近づき，低いほど特定のカテゴリが強調された分布になる．

カテゴリ $j \in \{0, \ldots, K-1\}$ とするカゴリカル分布 $p(\pi|z_i)$ から $\arg\max_j p(\pi|z_i)$ を計算し，エンコーダ $q(z|o)$ の出力 z_i と埋め込み空間 E の重みベクトル e_i が最も近いカテゴリ j を発話 $m_{w,h} \in \mathcal{N}$ とする $(w \in W, h \in H)$．また，発話 m は，$W \times H$ の離散系列 $(m \in \mathcal{N}^{W\times H})$ であり，入力画像 o を畳み込んだ特

徴 z のピクセルに対応したカテゴリ j をエージェント間で送受信すると解釈できる．

$$m_{w,h} = \arg\max_{j} p(\pi_j|z_i) \tag{5}$$

2.3 発話 m から埋め込み空間 E_t の更新

まず，ミニバッチ 1 からミニバッチ t までの埋め込み空間 $E_{1:t}$ は，エンコーダ $q(z_{1:t}|o_{1:t})$ の出力 $z_{1:t}$ との条件付き確率 $p(e_{1:t}|z_{1:t})$ として表すことができ，ベイズの定理から関係式 (6) を得る．

$$p(e_{1:t}|z_{1:t}) \propto p(e_{t-1})p(z_t|e_t) \tag{6}$$

確率分布 $p(e_{t-1})$ は，ミニバッチ $t-1$ における埋め込み空間 E_{t-1} のカテゴリ $j \in \{0, ..., K-1\}$ に対応する重みベクトル $e_{t-1,j}$ の確率分布である．また，確率分布 $p(z_t|e_t)$ は，ミニバッチ t におけるエンコーダ $q(z_{1:t}|o_{1:t})$ の出力 $z_{1:t}$ と埋め込み空間 E_t の重みベクトル e_t との条件付き確率である．

ここで，エージェント数 N を $N = 2$ とし，ミニバッチ t における自身の発話と相手の発話をそれぞれ m_t^A, m_t^B とする．さらに，それぞれの発話 m_t^A, m_t^B に対する重み係数 α_A, α_B を導入する．

$$\begin{aligned} p(z_t|e_{t,j}) &= \sum_{m_t^A, m_t^B} p(z_t, m_t^A, m_t^B|e_{t,j}) \\ &= \alpha_A \sum_{r=0}^{R-1} p(m_t^A|e_{t,j})p(z_{t,r}|e_{t,j}) \\ &+ \alpha_B \sum_{r=0}^{R-1} p(m_t^B|e_{t,j})p(z_{t,r}|e_{t,j}) \end{aligned} \tag{7}$$

式 (7) における変数 R は $R = D \times W \times H$ であり，総和 $\sum_{r=0}^{R-1}$ は，カテゴリ数 K 以外の次元数の総和を意味する．式 (7) を埋め込み空間 E_t の重みベクトル $e_{t,j}$ に関して，最尤推定することで更新式 (9) を得る．

$$\begin{aligned} c_{t,j} \leftarrow \lambda c_{t-1,j} + (1-\lambda)\big(&\alpha_A \sum_{r=0}^{R-1} \vec{1}[m_t^A = e_{t-1,j}] \\ &+\alpha_B \sum_{r=0}^{R-1} \vec{1}[m_t^B = e_{t-1,j}]\big) \end{aligned} \tag{8}$$

$$\begin{aligned} e_{t,j} \leftarrow \lambda e_{t-1,j} + (1-\lambda)\big(&\alpha_A \sum_{r=0}^{R-1} \frac{\vec{1}[m_t^A = e_{t-1,j}]z_{t,r}}{c_{t,j}} \\ &+\alpha_B \sum_{r=0}^{R-1} \frac{\vec{1}[m_t^B = e_{t-1,j}]z_{t,r}}{c_{t,j}}\big) \end{aligned} \tag{9}$$

定数 λ はミニバッチ t における更新の減衰率であり，値を小さくすると過去の情報が強く反映され，値を大きくすると，最新の情報が強く反映する．同様に発話 m^n の重み係数 α_n も大きくすると，エージェント n の情報が強く反映するように更新される．ただし，重み係数 α_n は，$\sum_{n \in A,B}^{N} \alpha_n = 1$ である．

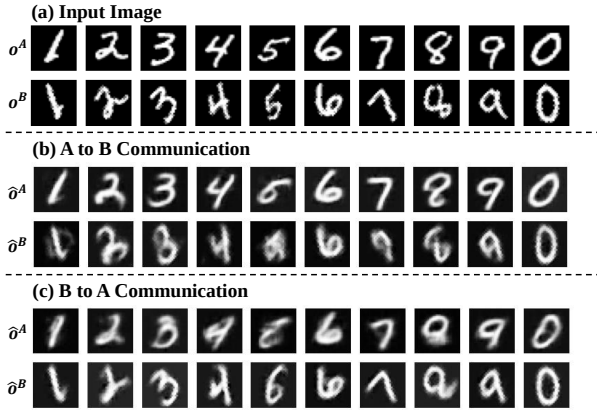

図 2 (a) は，エージェント A から B における入力画像 o^A．エージェント B から A における入力画像 o^B (b) は，提案手法によるエージェント A から B の予測画像 $\hat{o}^A$, $\hat{o}^B$．(c) は，提案手法によるエージェント B から A の予測画像 $\hat{o}^A$, $\hat{o}^B$．

3 実験

本実験では，提案手法によるエージェント間のコミュニケーションが可能なことを検証した．想定する実験は，エージェント数 $N = 2$ とし，エージェント A，エージェント B で異なる視点の入力画像 o^A, o^B を学習する．エージェント A の入力画像 o^A には，0 ~ 9 の数字が描かれた画像 60,000 枚を含む MNIST データセット [11] を使用し，エージェント B の入力画像 o^B には，エージェント A の入力画像 o^A を 45° 回転した画像を入力する．また，式 (3) の損失関数 $\mathcal{L}_\Theta$ を最小化するように学習すると同時に，式 (5) から量子化したエージェント A, B の発話 m^A, m^B を送信する．受信した発話 m^A, m^B からエージェント別に埋め込み空間 E^A, E^B を式 (9) で更新する．このようなコミュニケーションを 30 回繰り返し学習する．

パラメータ 本実験におけるパラメータの設定値を以下の項目に記載した．

- エージェント数：2
- 入力チャンネル数 c_{in}：1
- 潜在変数 z^A の次元数 D：128
- カテゴリ数 K：512
- 学習係数 lr：0.001
- パラメータ β：0.25

表 1 入力画像 o^* と予測画像 $\hat{o}^*$ との予測誤差 (有効数字 3 桁)．太字は，各実験における予測誤差の最小値である．

		重み係数 α				Mean Squared Error(o^*,$\hat{o}^*$)			
		A		B		A		B	
	場面	α_A^A	α_B^A	α_A^B	α_B^B	学習データ	テストデータ	学習データ	テストデータ
提案手法	A to B	0.5	0.5	0.5	0.5	0.0251	0.0246	**0.0536**	**0.0531**
(VQCom-VAE)		0.7	0.3	0.3	0.7	0.0231	0.0227	0.0540	0.0538
		0.9	0.1	0.1	0.9	**0.0215**	**0.0212**	0.0741	0.0746
		1.0	0.0	0.0	1.0	0.0224	0.0221	0.0825	0.0829
	B to A	0.5	0.5	0.5	0.5	**0.0506**	**0.0503**	0.0271	0.0266
		0.7	0.3	0.3	0.7	0.0546	0.0540	0.0245	0.0241
		0.9	0.1	0.1	0.9	0.0697	0.0696	0.0238	0.0235
		1.0	0.0	0.0	1.0	0.0815	0.0821	**0.0234**	**0.0230**

- 減衰率 λ: 0.99
- 温度定数 τ: 0.5
- 学習回数 epoch：30
- ミニバッチ数：64

検証では，テストデータ 10,000 枚を使用し，エージェントの発話 m^* に基づく予測画像 $\hat{o}^*$ と入力画像 o^* との予測誤差を Mean Squared Error(MSE) で評価した．ただし，送信者と受信者で異なる視点画像 o^* を学習するため，単純な評価は困難である．そのため，受信者の評価では，受信者の視点に直した画像を使用して計算した．また，エージェント n の発話 m^n の重み係数 α_n^* の変化による予測画像 o^* の影響についても検証した．

4 実験結果

表 1 にエージェント間のコミュニケーションによる入力画像 o^* と予測画像 $\hat{o}^*$ との誤差の評価結果を示した．まず，エージェント A から B のコミュニケーションで，最もエージェント B の予測誤差が小さいのは，重み係数 $\alpha_A^A = 0.5, \alpha_B^A = 0.5, \alpha_A^B = 0.5, \alpha_B^B = 0.5$ の条件であった．一方で，エージェント A から B において，最もエージェント A の予測誤差が小さいのは，重み係数 $\alpha_A^A = 0.9, \alpha_B^A = 0.1, \alpha_A^B = 0.1, \alpha_B^B = 0.9$ の条件であった．図 2 では，重み係数 $\alpha_A^A = 0.5, \alpha_B^A = 0.5, \alpha_A^B = 0.5, \alpha_B^B = 0.5$ の条件時におけるエージェント A の発話 m^A からそれぞれデコーダで出力した予測画像 $\hat{o}^*$ を示した．予測画像 $\hat{o}^*$ には，ぼやけている画像もあるが，相手の発話 m^* から予測画像 $\hat{o}^*$ の生成を確認できた．また，エージェント B の発話 m^A に対する重み係数 α_A^B を $\alpha_A^B = 0.5, 0.3, 0.1, 0.0$ と減少させ，エージェント B の発話 m^B に対する重み係数 α_B^B を $\alpha_B^B = 0.5, 0.7, 0.9, 1.0$ と増加させて，条件ごとに学習した．図 3 は，エージェント B の重み係数 α_A^B の変化による発話 m^A からの予測画像 $\hat{o}^B$ を示した．この結果からエージェント A の発話 m^A に対

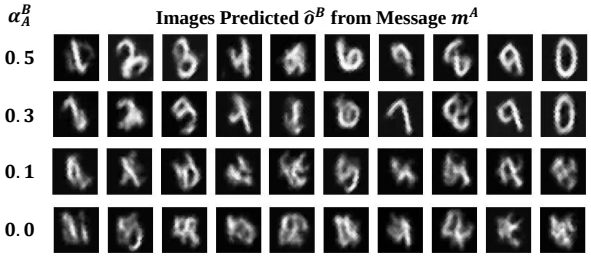

図 3 エージェント A から B のコミュニケーション時の重み係数 α_A^B の変化による予測画像 o^B の影響

する B の重み係数 α_A^B が低くなるほど，A の発話 m^A から生成した予測画像 $\hat{o}^B$ がぼやけていき，相手の発話に対する重み係数 α を低くすることで，コミュニケーションが通じなくなることを示した．

5 まとめ

本稿では，VQ-VAE を分散型深層モデルに拡張した VQCom-VAE を提案した．提案手法では，エンコーダ，デコーダの学習と同時に式 (9) を用いて埋め込み空間 E を更新することで，エージェント間の発話 m^* に基づいた予測画像 o^* を生成することが可能である．実験では，MNIST データセットを使用し，発話 m^* から予測画像 $\hat{o}^*$ が生成可能かを検証した．評価では，入力画像 o^* と予測画像 $\hat{o}^*$ との誤差を計算した．また，生成した予測画像 $\hat{o}^*$ を確認した結果，相手の発話 m^* から予測画像 $\hat{o}^*$ の生成が可能なことを示した．さらに相手の発話 m^* の重み係数 α を低くすることで，相手の発話 m^* から予測画像 $\hat{o}^*$ がぼやけていき，コミュニケーションが通じなくなることを示した．

しかし，本提案手法では概念獲得や二重分節性などを考慮した言語創発モデルではない．今後は埋め込み空間 E を階層化し，概念獲得や二重分節性を考慮した手法に発展させる予定である．さらに，Action Chunking Transformer(ACT)[13] に応用し，ロボットによる協調制御にも発展させたい．

参考文献

[1] Kenji Doya. **Bayesian brain: Probabilistic approaches to neural coding**. MIT press, 2007.

[2] Tadahiro Taniguchi, Yuto Yoshida, Akira Taniguchi, and Yoshinobu Hagiwara. Emergent communication through metropolis-hastings naming game with deep generative models, 2023.

[3] Hiroto Ebara, Tomoaki Nakamura, Akira Taniguchi, and Tadahiro Taniguchi. Multi-agent reinforcement learning with emergent communication using discrete and indifferentiable message. In **2023 15th International Congress on Advanced Applied Informatics Winter (IIAI-AAI-Winter)**, pp. 366–371. IEEE, 2023.

[4] Ziwoo You, Hiroto Ebara, Tomoaki Nakamura, Akira Taniguchi, and Tadahiro Taniguchi. Multimodal continuous symbol emergence using a probabilistic generative model based on gaussian processes. **2024 IEEE International Conference on Development and Learning (ICDL)**, pp. 1–6, 2024.

[5] Issei Saito, Tomoaki Nakamura, Akira Taniguchi, Tadahiro Taniguchi, Yohei Hayamizu, and Shiqi Zhang. Emergence of continuous signals as shared symbols through emergent communication. **2024 IEEE International Conference on Development and Learning (ICDL)**, pp. 1–6, 2024.

[6] Tomoaki Nakamura, Takayuki Nagai, and Tadahiro Taniguchi. Serket: An architecture for connecting stochastic models to realize a large-scale cognitive model, 2017.

[7] Fruits-360 dataset. Fruits-360 dataset, 2024. https://www.kaggle.com/datasets/moltean/fruits.

[8] Tomoaki Nakamura, Takaya Araki, Takayuki Nagai, and Naoto Iwahashi. Grounding of word meanings in latent dirichlet allocation-based multimodal concepts. **Advanced Robotics**, Vol. 25, No. 17, pp. 2189–2206, 2011.

[9] Tomoaki Nakamura, Takayuki Nagai, Daichi Mochihashi, Ichiro Kobayashi, Hideki Asoh, and Masahide Kaneko. Segmenting continuous motions with hidden semi-markov models and gaussian processes. **Frontiers in neurorobotics**, Vol. 11, p. 67, 2017.

[10] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning, 2018.

[11] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, Vol. 86, No. 11, pp. 2278–2324, 1998.

[12] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax, 2017.

[13] Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware, 2023.