New Generation Computing (2020) 38:23–48
https://doi.org/10.1007/s00354-019-00084-w

NEW GENERATION COMPUTING

Neuro-SERKET: Development of Integrative Cognitive System Through the Composition of Deep Probabilistic Generative Models

Tadahiro Taniguchi[1] · Tomoaki Nakamura[3] · Masahiro Suzuki[4] · Ryo Kuniyasu[3] · Kaede Hayashi[1] · Akira Taniguchi[1] · Takato Horii[2] · Takayuki Nagai[2,3]

Received: 14 October 2019 / Accepted: 8 December 2019 / Published online: 22 January 2020

Abstract
This paper describes a framework for the development of an integrative cognitive system based on probabilistic generative models (PGMs) called Neuro-SERKET. Neuro-SERKET is an extension of SERKET, which can compose elemental PGMs developed in a distributed manner and provide a scheme that allows the composed PGMs to learn throughout the system in an unsupervised way. In addition to the head-to-tail connection supported by SERKET, Neuro-SERKET supports tail-to-tail and head-to-head connections, as well as neural network-based modules, i.e., deep generative models. As an example of a Neuro-SERKET application, an integrative model was developed by composing a variational autoencoder (VAE), a Gaussian mixture model (GMM), latent Dirichlet allocation (LDA), and automatic speech recognition (ASR). The model is called VAE + GMM + LDA + ASR. The performance of VAE + GMM + LDA + ASR and the validity of Neuro-SERKET were demonstrated through a multimodal categorization task using image data and a speech signal of numerical digits.

Keywords Cognitive models · Probabilistic generative models · Symbol emergence in robotics · Deep generative models · Machine learning

Introduction

The development of integrative cognitive systems that can form perceptual and behavioral concepts using multimodal sensorimotor information and learn and understand a language in a real-world environment is a significant challenge in artificial intelligence (AI) and robotics [1]. This paper describes a theoretical framework called Neuro-SERKET for this purpose.

✉ Tadahiro Taniguchi
taniguchi@em.ci.ritsumei.ac.jp

Extended author information available on the last page of the article

Numerous types of integrative cognitive systems, which are sometimes called a cognitive architecture, have recently been developed for building service robots and modeling human adaptive cognition [2–11]. However, the cognitive systems for robots need to handle a variety of types of sensorimotor modalities, e.g., image, sound and actuation, and a variety of internal cognitive processes, e.g., categorization and planning. Therefore, the size of the computational models becomes large and the development requires significant effort for each integrative cognitive system. For further progress in this stream of research, we need to achieve an efficient way to develop complex cognitive systems in a practical manner. In addition, recent advancements in deep generative models (DGMs), for instance, a variational autoencoder (VAE) [12], have boosted their utilization in the development of cognitive systems.

This paper describes a novel framework enabling researchers and developers to create elemental cognitive modules, i.e., image recognition, automatic speech recognition, and syntax and clustering models, independently, and compose them into a large cognitive system, which can operate as a cognitive system and be consistently trained as a single learning system. Neuro-SERKET is an extension of SERKET [11], which was proposed as a framework for decomposing and composing PGMs. As described later, SERKET does not support neural networks, i.e., deep learning. A framework called Neuro-SERKET can also employ neural network-based cognitive modules. In addition to that, SERKET only supports head-to-tail connections for decomposition and composition. In contrast, Neuro-SERKET supports head-to-head and tail-to-tail connections in graphical models, as well.

The remainder of this paper is organized as follows. Section 2 introduces the background of Neuro-SERKET. Section 3 describes the Neuro-SERKET framework. More concretely, the method for the decomposition and composition of probabilistic generative models (PGMs), including DGMs, is described. Section 4 describes a concrete example of integrative cognitive systems developed using the Neuro-SERKET framework. The integrative model was developed by combining VAE, a Gaussian mixture model (GMM), a latent Dirichlet allocation (LDA), and an automatic speech recognition system (ASR), and can form a multimodal concept from row speech and image signals. This is an illustrative example involving all types of elemental connections, i.e., head-to-tail, tail-to-tail, and head-to-head connections, and a neural network. Finally, Sect. 5 provides some concluding remarks.

Background

During this decade, the complexities of cognitive systems that can learn real-world knowledge and find the latent structure from multimodal sensorimotor information obtained by the robot itself, i.e., an embodied artificial cognitive system, have increased. Cognitive systems for robots that learn the relationships among different types of multimodal sensory information have been proposed using PGMs and neural networks [2–6]. Methods proposed in the studies enable robots to infer the latent variables from their own observations, for instance, the robot acquired object categories as latent variables from visual, sound and tactile sensory signals in [2]. These

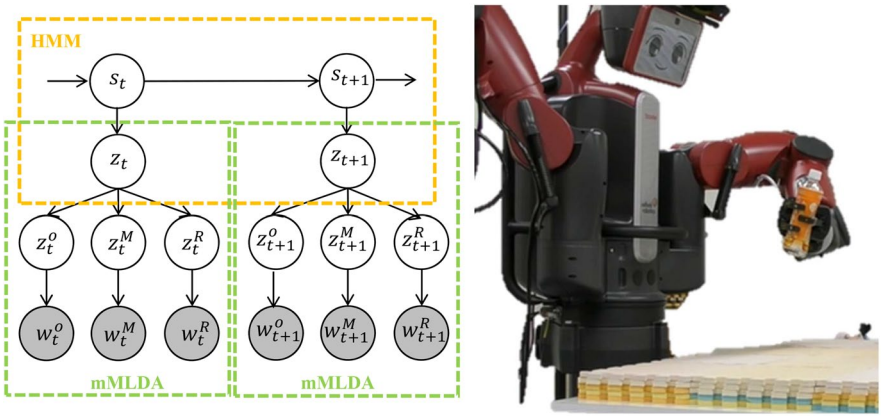

Fig. 1 A robot planning and conducting a multimodal object categorization using a complex PGM [7]

enable robots to acquire various knowledge by inferring the latent variables from their own observations. A further advancement of such cognitive systems allows the robots to find meanings of words by treating a linguistic input as another modality [13–15]. Cognitive models have recently become more complex in realizing various cognitive capabilities: grammar acquisition [16], language model learning [17], hierarchical concept acquisition [18, 19], spatial concept acquisition [20], motion skill acquisition [21], and task planning [7] (see Fig. 1). It results in an increase in the development cost of each cognitive system.

Among them, it has been recognized that PGMs are extremely useful for modeling an integrative cognitive system that deals with multimodal and heterogeneous information and learns various functional concepts, i.e., internal representations, in an unsupervised manner because we can design the relationships of latent variables as a graphical model for introducing constraints to the data modeling. This can be interpreted through an analogy of designing cortical connections in our brain. Nakamura et al. proposed multimodal LDA (MLDA) for multimodal object categorization [14]. They also developed a series of PGMs extending this idea. Taniguchi et al. proposed a spatial concept formation with simultaneous localization mapping (SpCoSLAM) for a spatial concept formation and lexical acquisition [8] (see Fig. 2). Such studies have contributed to the field of symbol emergence in robotics [9].

A cognitive robot empowered by an integrative cognitive system can form object and spatial concepts, learn behaviors, and become able to understand human commands without explicit supervision differently from a supervised learning-based approach, which has been widely used in recent AI developments. However, the growing complexity of graphical models has gradually increased barriers to entry into this research field for numerous researchers. A framework for developing an integrative cognitive system is required for further progress of this field in the same way as applied in accelerated studies on various deep learning frameworks around deep neural networks.

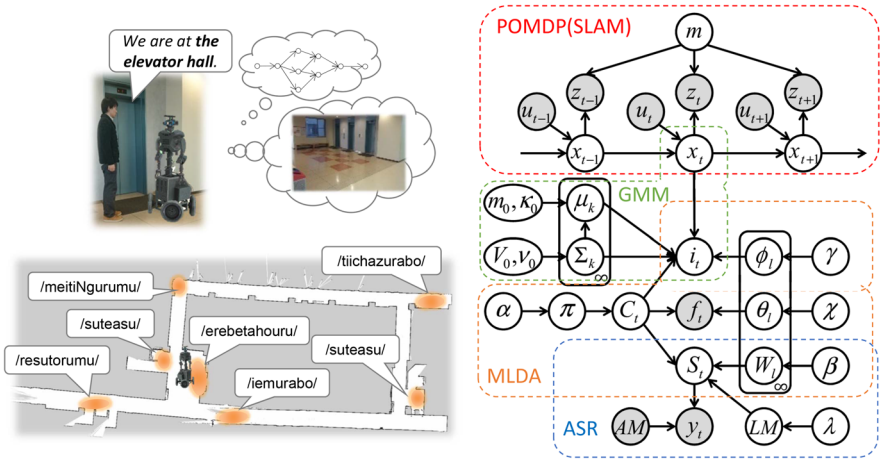

Fig. 2 Graphical model of SpCoSLAM [8]

SERKET is a framework proposed for solving this problem [11]. SERKET was designed to enable a distributed software development of extremely large PGMs. In general, many pre-existing models for cognitive systems used in robots can be considered as a composition of elemental cognitive modules. For example, in Figs. 1 and 2, the elemental modules in each graphical model are shown [7, 8]. SERKET provides a theoretical framework for decomposing and composing PGMs. Cognitive modules developed in a distributive manner, namely elemental PGMs, can be composed into a PGM using the SERKET framework, and the composed PGM can learn and work in the same manner as a PGM developed from scratch by a single developer. However, SERKET has the following limitations.

- SERKET only supports a head-to-tail connection, although in general, the graphical model can theoretically have tail-to-tail and head-to-head connections, as well.
- SERKET implicitly assumes the inference method using the Markov chain with a Monte Carlo approach and does not assume the integration of neural networks.

The first limitation prevents us from a flexible, creative, and efficient development of a variety of integrative cognitive systems. For example, if we would like to develop an MLDA by integrating multiple LDAs, the framework should support tail-to-tail connections.

The second limitation prevents us from the integration of DGMs, i.e., neural networks. As is widely known, DGMs can achieve representation learning, i.e., feature extraction. For example, a VAE is a probabilistic generative model having the capability of representation learning and can be integrated with PGMs, e.g., a hidden Markov model (HMM) and a GMM. However, SERKET does not support the integration of VAEs.

The integration of conventional PGMs, e.g., HMM and GMM, with DGMs, e.g., VAE, has received increasing attention, and such integrative PGMs have been studied. In a VAE, the encoder models the intractable posterior distribution of the latent representation, and the decoder reconstructs the observation using its latent representation, which usually assumes a single Gaussian prior. In recent studies, various PGMs such as GMM and HMM are applied to its latent space, and are used for semi-supervised learning [22], clustering [23, 24], and acoustic unit discovery [25]. The structured VAE (SVAE) proposed in [23] is a generalization of the VAE to general PGMs, including capturing the correlation structure of sequential data, and in [25], it was extended to an acoustic unit discovery. In [24], a two-layer latent representation is composed that uniformly assumes a multi-modal prior distribution for a latent space, although this model requires a specific optimization to prevent an over-regularization. In [26], a generative process is defined based on a GMM in a latent space, and achieves a better performance.

Not only the composition of a conventional PGM and a DGM, but also composition of DGMs should be explored. More structured DGMs that can handle multimodal data are also gaining attention. Whereas vanilla VAEs can only take unimodal data, in [27, 28], conditional VAEs have been proposed that can handle another modality. These models can generate a modality corresponding to another modality data, e.g., generating images from captions [29]. However, these models cannot generate multimodal data bi-directionally, i.e., both generating images from captions and generating captions from images, and also cannot obtain a representation that integrates their multimodal information. In [30], a joint multimodal VAE is proposed, which not only has a multimodal inference model that embeds multimodal data into a joint representation but also unimodal inferences learned to approximate such multimodal data. The authors showed that, in the case of two modalities, this model can appropriately generate modalities bidirectionally and can infer a good joint representation. In [31], the authors extended this multimodal inference model by introducing the idea of a product of experts [32], and proposed a multimodal VAE (MVAE) that can handle any number of modalities. Moreover, [33] showed that a model whose association networks connect the latent variables of modality-specific VAEs can apply a cross-modal generation among multiple modalities. This line of studies clearly shows that DGMs become more and more structured and complex in the same way as conventional PGMs. Efficient way of developing complex cognitive systems by integrating DGMs should be explored.

Considering the advancement of DGMs and recognizing the limitations of SERKET, in this study, we extend the application of SERKET and propose an updated version called Neuro-SERKET. Table 1 shows a list of modules implemented in Neuro-SERKET library as examples. Developers can build a variety of integrative cognitive systems by composing these modules following the Neuro-SERKET framework.

Table 1 Examples of modules implemented in current Neuro-SERKET library

Module	Description
Observation	Module to deal with observations as messages
CNNFeatureExtractor	Feature extractor from images based on CNN
HACFeatureExtractor	Feature extractor from audio files based on HAC [34]
VAE	Module to learn feature representations based on VAE [12]
MVAE	Module to learn features based on multinomial VAE [35]
GMM	Unsupervised clustering based on GMM
MLDA	Unsupervised clustering based on MLDA [14]
MM	Module to learn transition of discrete variables
TtoT	Module to construct the tail to tail connection

Neuro-SERKET

Generation: Decomposition of Complex Graphical Model

Overview

Neuro-SERKET is an extension of SERKET. Therefore, it basically follows the approach of SERKET. SERKET provides a theoretical way to achieve a decomposition and composition of PGMs. A decomposition is mainly related to the generative process, i.e., a generative model, and a composition is mainly related to the inference process, i.e., an inference model. In SERKET, decomposition and composition are conducted by following three rules.

1. A node, a latent variable z, in an integrated model is shared by two elemental modules.
2. A module regards z as an observable, and the parameter Θ of the probabilistic distribution $P(z|\Theta)$ is estimated.
3. The other module estimates z by taking a prior $P(z|\Theta)$ with a fixed parameter Θ, which is estimated in 2.

Numerous types of PGMs can be described as graphical models. Directed graphs representing PGMs, i.e., graphical models, have three types of elemental connections, i.e., head-to-tail, head-to-head, and tail-to-tail (see Fig. 3).

The important feature of SERKET is that an integrated PGM developed by composing sub-modules following the SERKET framework can operate in almost the same way as a PGM developed from scratch, and uses an inference procedure developed for the PGM, in a reasonably approximate manner. Neuro-SERKET also has this feature.

In addition to a conventional SERKET framework, Neuro-SERKET provides two additional features.

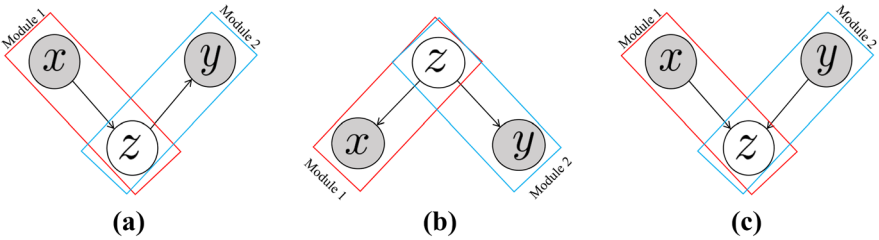

Fig. 3 Elemental graphical models: **a** head-to-tail, **b** head-to-head, and **c** tail-to-tail

- Neuro-SERKET supports tail-to-tail and head-to-head connections in addition to head-to-tail connections.
- Neuro-SERKET supports deep probabilistic generative models, e.g., VAEs. Therefore, PGMs using Neuro-SERKET can make use of the representation learning capability of neural networks.

First, we describe how to decompose complex graphical models with the Neuro-SERKET framework. As is widely known, probabilistic graphical models have three types of elemental connections, as shown in Fig. 3. Note that each generative process, e.g., $P(x|z)$, has global parameters, e.g., θ for $P(x|z,\theta)$, although these are omitted from the graphical model for the sake of simplicity. Each generative process can have other latent variables as well. A systematic approach to a decomposition is also described herein.

Head-to-Tail Decomposition

In the Neuro-SERKET framework, a complex graphical model is systematically decomposed. First, we take a head-to-tail connection, shown in Fig. 3a, as an example. The joint distribution $P(x, y, z)$ can be written as follows because of a conditional dependency indicated by the graphical model.

$$P(x, y, z) = P(y|z)P(z|x)P(x). \tag{1}$$

The generative process of the latent variable z is described as $P(x, z) = P(z|x)P(x)$. Next, when looking at the generative process of y, it can be seen that the generative process of y can be described as $P(y, z|x = X) = P(y|z)P(z|x = X)$, where X is an instance of x. Here, note that $P(y, z|x = X)$ does not depend on the variable x when x is fixed, i.e., $x = X$. This means the probabilistic generative model can be decomposed into two modules.

The discussion above is reconfirmed from the viewpoint of factorization. The joint probability can be factorized in two ways.

$$P(x, y, z) = P(y|z) \underbrace{P(z, x)}_{\text{Module 1}}, \tag{2}$$

$$= \underbrace{P(y, z|x)}_{\text{Module 2}} P(x). \tag{3}$$

The first and second modules correspond to the generative model for z and y, respectively.

If a joint distribution can be factorized in two ways when sharing a latent variable, e.g., z in Eqs. (2) and (3), the PGM can be decomposed into two modules.

We introduce an operator $\otimes$ representing a composition operation of PGMs for illustrative purposes.

$$P(x, y, z) \Rightarrow P(z, x) \otimes P(y, z). \tag{4}$$

This shows that PGM $P(x, y, z)$ can be decomposed into $P(z, x)$ and $P(y, z)$, which are two elemental modules.

Tail-to-Tail Decomposition

Another elemental connection is a tail-to-tail (see Fig. 3b) connection, which is also called a "fork". The joint distribution of x, y, and z can be described as follows using an assumed conditional independence:

$$P(x, y, z) = P(x|z)P(y|z)P(z). \tag{5}$$

In the same way, as the discussion regarding a head-to-tail connection, the joint distribution can be factorized in the following two ways.

$$P(x, y, z) = \underbrace{P(x, z)}_{\text{Module 1}} P(y|z), \tag{6}$$

$$= P(x|z) \underbrace{P(y, z)}_{\text{Module 2}}. \tag{7}$$

Each module obtained through a decomposition corresponds to a generative process of x and y. Following the usage of symbol $\otimes$, which we introduced in the previous subsection, the PGM $P(x, y, z)$ can be decomposed into two modules, i.e., joint distributions, $P(x, z)$ and $P(y, z)$, as follows:

$$P(x, y, z) \Rightarrow P(x, z) \otimes P(y, z). \tag{8}$$

Head-to-Head Decomposition

The other elemental connection is a head-to-head (see Fig. 3c) connection. The joint distribution of x, y, and z under a head-to-head connection can be decomposed when considering the following conditional independence:

$$P(x, y, z) = P(z|x, y)P(x)P(y). \tag{9}$$

If we apply the systematic rule for a decomposition in the same way as a head-to-tail and tail-to-tail connection, we will obtain the following decomposition:

$$P(x, y, z) = \underbrace{P(x, z|y)}_{\text{Module 1}} P(y), \tag{10}$$

$$= \underbrace{P(y, z|x)}_{\text{Module 2}} P(x). \tag{11}$$

However, differing from the previous decomposition, i.e., head-to-tail and tail-to-tail connections, both modules represent the generative process of z, and involve x and y. This prevents us from taking a SERKET-based approach for inferring z in each module because both of the modules involve x and y. The SERKET framework requires that a latent variable z be inferred within a module using one of x or y after decomposition. In other words, z should be regarded as an observable, i.e., a given variable, in another module. Therefore, SERKET does not provide the way of decomposition for a head-to-head connection. However, the decomposition of a head-to-head connection is important in building further complex cognitive systems. For example, SpCoSLAM assumes that a generated sentence S_t is conditioned by the spatial concept C_t, i.e., "where the robot is", and syntactic and lexical information, i.e., a language model LM and the set of parameters of topic-dependent word distributions $\{W_l\}$ (see Fig. 2) [8].

Therefore, in Nuero-SERKET, we introduce a new way to achieve an approximate decomposition for $P(z|x, y)$.

$$P(z|x, y) \approx \hat{P}(z|x, y) \propto P(z|x)P(z|y), \tag{12}$$

where $\hat{P}(z|x, y)$ is an approximately decomposed distribution. This approximation consists of two steps. The first approximation is that $P(z|x, y)$ is decomposed into distributions including $P(z|x)$, $P(z|y)$ and $P(z)$. This approximate decomposition can have two ways of interpretation: a product of expert (PoE), i.e., $\hat{P}(z|x, y) = P(z)P(z|x)P(z|y)$ [32], and a uni-gram re-scaling, i.e., $\hat{P}(z|x, y) = \frac{P(z|x)P(z|y)}{P(z)}$ [36]. In both cases, a prior $P(z)$ is considered to be a uniform distribution. This means that the prior in the distribution $\hat{P}$ can be ignored, i.e., $\hat{P}(z|x, y) \propto P(z|x)P(z|y)$.

Using this approximation, we can obtain the following modules.

$$P(x, y, z) = P(z|x, y)P(x)P(y), \tag{13}$$

$$\approx \propto P(z|x)P(z|y)P(x)P(y), \tag{14}$$

$$= \underbrace{P(x, z)}_{\text{Module 1}} \underbrace{P(y, z)}_{\text{Module 2}}, \tag{15}$$

where $\approx\propto$ represents the terms "approximation" and "proportion to".[1]

The decomposition is described as follows.

$$P(x, y, z) \Rightarrow P(x, z) \otimes P(y, z). \tag{16}$$

The appropriateness of the approximation is evaluated empirically in Sect. 4 based on an experiment.

For example, SpCoSLAM (Fig. 2) has a head-to-head connection at approximately S_t, i.e., a sentence recognized by an ASR system. When we pick up related variables for illustrative purposes, we can start with the following joint distribution:

$$P(y_t, \mathrm{LM}, S_t, C_t|\mathrm{AM}) = P(y_t|\mathrm{AM}, S_t)P(S_t|C_t, \mathrm{LM})P(\mathrm{LM})P(C_t), \tag{17}$$

where y_t and AM are a speech signal and acoustic model in an ASR system, respectively.

In practical terms, AM and LM are implemented in an ASR system, i.e., packaged software, and C_t is a part of a multimodal categorization module. Therefore, calculating a generative probability and drawing samples theoretically are extremely difficult. Therefore, Neuro-SERKET introduces the following approximation:[2]

$$P(S_t|C_t, \mathrm{LM}) \approx\propto P(S_t|\mathrm{LM})P(S_t|C_t). \tag{18}$$

This allows developing a graphical model with two parts, i.e., an ASR (including a lexical acquisition) module and a multimodal categorization module.[3]

$$P(y_t, \mathrm{LM}, S_t, C_t|\mathrm{AM}) \approx\propto \underbrace{P(y|\mathrm{AM}, S_t)P(S_t|\mathrm{LM})P(\mathrm{LM})}_{\text{ASR module}} \underbrace{P(S_t|C_t)P(C_t)}_{\substack{\text{Multimodal}\\ \text{categorization}\\ \text{module}}}. \tag{19}$$

Inference: Composition of Complex Graphical Model

In the SERKET framework, each module is developed by a different researcher or developer in a distributed manner. After the development of the modules, they are integrated into an integrative cognitive system. Integrated modules learn together and work together as a single cognitive system.

In the context of PGMs, prediction, estimation, and learning are simply regarded as an inference of latent variables. Therefore, the composition of the modules corresponds to an inference procedure combining multiple modules. This section provides a method of composition for each elemental connection. Figure 4 summarizes

[1] $P(x) \approx\propto f(x)$ is an abbreviation of $P(x) \approx \hat{P}(x) \propto f(x)$.

[2] Note that the original study on SpCoSLAM does not use this method of approximation. A uni-gram rescaling approximation alone was employed instead.

[3] Note that, for illustrative purposes, the other variables and hyperparameters are ignored from the equations.

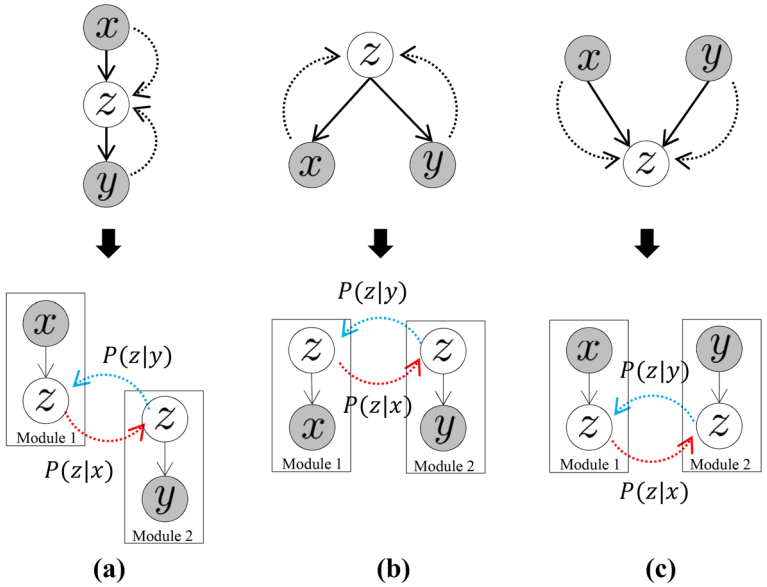

Fig. 4 Graphical models and their Neuro-SERKET implementations of **a** head-to-tail, **b** head-to-head, and **c** tail-to-tail. The black-dotted arrows represent conditional probabilities used in the inference

the three types of elemental connections, message passing, and the decomposition method used in our framework, Neuro-SERKET. In each graphical model, the black-dotted arrows indicate the calculation of a posterior distribution, which is necessary for an inference procedure. Note that the two dotted arrows in each graphical model form head-to-head relationships.

Figure 4a shows the method of message passing between two modules in the case of a head-to-tail connection. Conventional SERKET mainly introduced two methods for achieving a head-to-tail composition.

The first is called the message passing (MP) approach, and its procedure is as follows.[4]

1. In module 1, $P(z|x)$ is computed.
2. $P(z|x)$ is sent to module 2.
3. In module 2, the probability distribution $P(z|y)$, which represents the relationships between x and y, is estimated using $P(z|x)$.
4. $P(z|y)$ is sent to module 1.
5. In module 1, the latent variable z is estimated, and the parameters of $P(x|z)$ are updated.

[4] As a variation to the MP approach, module 1 can send samples, i.e., the data distribution, $z^l \sim P(z|x)(l = \{1, \ldots, L\})$, as a Monte Carlo approximation of $P(z|x)$. As a special case of this, module 1 can send a sample $z^\star \sim P(z|x)$ to module 2. In Sect. 4, as an example, the VAE module sends a recognition result to another module.

The other is called a sampling importance resampling (SIR) approach, the procedure of which is as follows.

1. Generate L samples $z^{(l)} \sim P(z|x)$ in module 1.
2. Send $\{z^{(1)}, \ldots, z^{(L)}\}$ to module 2.
3. Select sample $z^{\star}$ among $\{z^{(1)}, \ldots, z^{(L)}\}$ by calculating their importance using $P(z|y)$ and update the parameters of $P(z|y)$.
4. Send the selected sample $z^{\star}$ to module 1.
5. Update the parameters of $P(x|z)$.

This approach involves a Monte Carlo approximation. However, many off-the-shelf modules do not support the calculation of a posterior distribution $P(z|x)$ itself. Therefore, the SIR approach allow us to use various off-the-shelf modules, e.g., ASR and image recognition systems, that provide only samples, i.e., estimated results.

With this inference procedure, SERKET employs a PoE approximation, i.e., $P(z|x, y) \approx\propto P(z|x)P(z|y)$ in the same way as a head-to-head decomposition. For further details, please refer to the original study [11].

Differing from the decomposition part, there are no structural differences among the three elemental connections. The dotted line in Fig. 4 shows the inference process for each elemental connection. We can see that all pairs of dotted arrows have head-to-head connections. This means that, in the inference process, i.e., composition, we can use the same procedure as a head-to-tail composition in the cases of tail-to-tail and head-to-head compositions.

However, we need to develop a special treatment for the implementation of a tail-to-tail composition. SERKET requires connecting latent variables of modules in a hierarchical manner [11], and we cannot connect module 1, i.e., $P(x|z)$, and module 2, i.e., $P(y|z)$, directly in a tail-to-tail composition (Fig. 4c). Therefore, we introduce an auxiliary module called a tail-to-tail (TtoT) module, which connects modules 1 and 2 and transfers the probability distribution between the two modules.

In this way, Neuro-SERKET also makes use of a PoE and uniform distribution prior approximation [32] in the composition, and achieves an inference of the integrative PGMs. Differing from these assumptions, description, and discussion in the original study on SERKET [11], Neuro-SERKET does not assume any specifics for the implementation and inference procedure of each module.[5] Therefore, Neuro-SERKET can integrate neural network-based generative models, i.e., DGMs such as VAEs.

[5] In the original study on SERKET, the authors mentioned that they "employed a sampling-based method because of its simpler implementation". This means that they excluded modules that are trained using other inference procedures, e.g., gradient-based methods. In general, a sampling-based approach is unsuitable for the training of neural networks. This means that SERKET fails to involve neural network-based modules, which have recently been widely used, into SERKET-based cognitive systems.

Table 2 Model parameters of the integrative PGM (VAE + GMM + LDA + ASR)

Parameter	Description
θ	Parameter of VAE decoder (generative network)
$\boldsymbol{o}_1, \boldsymbol{o}_2$	Observations, image data and speech signal
z_1	Latent variable of VAE extracted from $\boldsymbol{o}_1$
z_2	Index of classes the observations are categorized into
w	Word recognized by the ASR system
$\boldsymbol{\mu}, \boldsymbol{\Sigma}$	Mean vector and variance–covariance matrix of Gaussian distribution
r_0, m_0, S_0, ν_0	Parameters of Gauss–Wishart distribution
π, φ	Parameters of multinomial distribution
α, β	Parameters of Dirichlet distribution
N	The number of observations
K	The number of classes in LDA and GMM

Example: Concept Formation Using VAE + GMM + LDA + ASR

This section describes an illustrative example of a cognitive system that can be developed following the Neuro-SERKET framework by integrating pre-existing modules. For illustrative purposes, this model involves all of the elemental connections, i.e., head-to-tail, tail-to-tail, and head-to-head. A neural network-based module, i.e., VAE, is also included as an elemental module. The developed PGM for multimodal categorization is a composition of a VAE, GMM, LDA [37], and ASR. We empirically validated Neuro-SERKET through an experiment using image data and speech signals.

Model

Figure 5 shows a graphical model of the PGM developed using the Neuro-SERKET framework. This PGM receives two types of observations, i.e., pairs of an image $\boldsymbol{o}_1$ and speech signal $\boldsymbol{o}_2$ corresponding to the image. The PGM is for an unsupervised multimodal categorization, including representation learning of the image data. Image $\boldsymbol{o}_1$ is expected to be encoded into the latent variable z_1 using VAE. The speech signal $\boldsymbol{o}_2$ is recognized, and word w is estimated using a language model parameterized by $\mathcal{L}$, which can also be learned from the data. The obtained word w is clustered using LDA, and the estimated representation of image z_1 is clustered using GMM. Note that the latent variable z_2 representing the class of the input pair of data is shared by the LDA and GMM. This means that an estimation of z_2 corresponds to a multimodal categorization. A list of parameters of the PGM is enumerated in Table 2.

Figure 6 shows the elemental cognitive modules and communication between them. The communication conducted during the inference procedure, i.e., the composition, is summarized as follows.

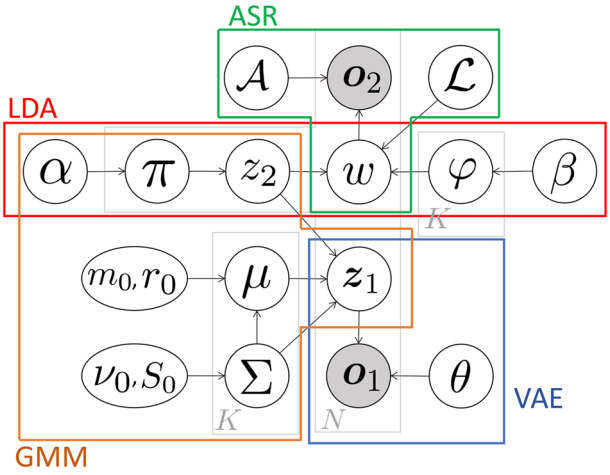

Fig. 5 The original graphical model of the integrative PGM (VAE + GMM + LDA + ASR). Each block shows each module

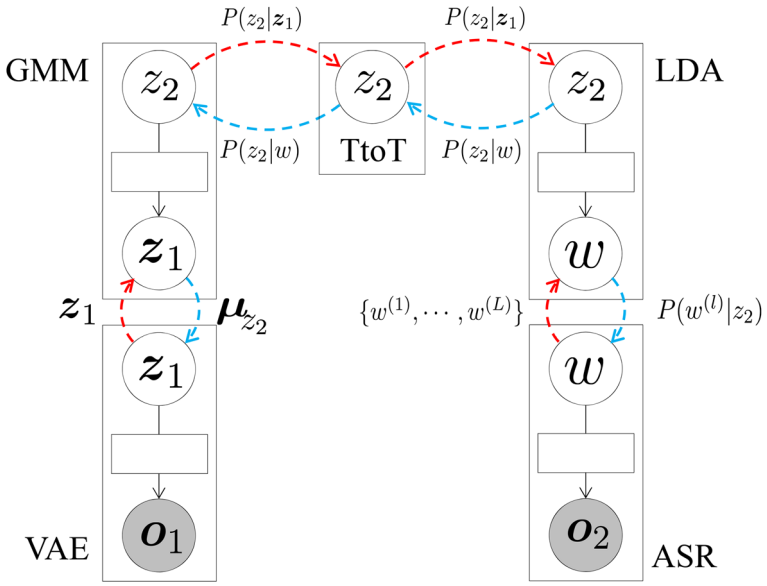

Fig. 6 Decomposed modules and communication between them following the Neuro-SERKET framework

VAE module The VAE module extracts a representation, i.e., latent variable, z_1, from the image data o_1, and sends z_1 to the GMM module. The GMM module sends μ_{z_2}, which is a mean vector of the Gaussian distribution that z_1 was categorized into, back to the VAE module. VAE uses μ_{z_2} and updates the

parameters of the encoder and decoder of the VAE to maximize the evidence lower bound (ELBO).

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{o}_1) = -D(q_{\phi}(z_1|\boldsymbol{o}_1)||\mathcal{N}(\boldsymbol{\mu}_{z_2}, \boldsymbol{I})) + E_{q_{\phi}(z_1|\boldsymbol{o}_1)}[\log p_{\theta}(\boldsymbol{o}_1|z_1)]. \tag{20}$$

This allows the VAE to learn a representation appropriate for categorization by the GMM module.

GMM module The GMM module sends $P(z_2|z_1)$, which is obtained by categorizing z_1 received from the VAE module, to the TtoT module . This module shares z_2 with the LDA module (Fig. 5). Therefore, the inference of bz_2 is affected by the LDA module. The TtoT module mediates the information between the GMM and LDA modules. When the GMM module applies an inference, i.e., a categorization, the GMM module uses $P(z_2|\boldsymbol{w})$, which is received from the TtoT module.

$$z_2 \sim P(z_2|z_1, \boldsymbol{w}) \approx\propto P(z_2|z_1)P(z_2|\boldsymbol{w}) \tag{21}$$

ASR module The ASR module represents an off-the-shelf speech recognition system.[6] The ASR module sends the L-best speech recognition results of $\boldsymbol{o}_2$ to the LDA module. The L-best results are regarded as an approximation of L samples from the posterior distribution $P(w|\boldsymbol{o}_2)$. The LDA module calculates the importance of each word $P(w^{(l)}|z_2)$, and re-samples the word using the importance weight (SIR approach) as follows:

$$w^{(l)} \sim P(w|\boldsymbol{o}_2), \tag{22}$$

$$w^* \sim \hat{P}(w) \propto \sum_l P(w^{(l)}|z_2)\delta_{w^{(l)}}(w), \tag{23}$$

where $\delta_{w^{(l)}}(w)$ is a probability mass function.

LDA module The LDA module receives a set of words $\boldsymbol{w} = \{w^{(1)}, \ldots, w^{(L)}\}$ and clusters them. As a result, the LDA module calculates $P(z_2|\boldsymbol{w})$ and sends it to the TtoT module. Note that z_2 is shared with the GMM module (see Fig. 5), and in the clustering, i.e., inference, the process is affected by the categorization by the GMM module. Therefore, the LDA module uses $P(z_2|z_1)$ received from the TtoT module when it clusters words.

$$z_2 \sim P(z_2|z_1)P(z_2|\boldsymbol{w}). \tag{24}$$

TtoT module The TtoT module simply transfers $P(z_2|\boldsymbol{w})$ from the LDA module to the GMM module, and $P(z_2|z_1)$ from the GMM module to the LDA module.

Following the communication procedure shown above, the total PGM can be trained by optimizing each module locally under the influence of neighboring modules.

[6] Julius: Open-Source Large Vocabulary Continuous Speech Recognition Engine: https://github.com/julius-speech/julius.

```
1  wavs = glob.glob( "speech/0/*.wav" )
2  obs1 = srk.Observation( np.loadtxt("image.txt") )
3
4  speech = word_recog.WordRecog( wavs )
5  lda1 = mlda.MLDA( 10, [100] )
6  vae1 = vae_model( 10, itr=200, batch_size=500 )
7  gmm1 = gmm.GMM( 10, itr=50 )
8  tt = TtoT.TtoT()
9
10 lda1.connect( speech )
11 vae1.connect( obs1 )
12 gmm1.connect( vae1 )
13 tt.connect( lda1, gmm1 )
14
15 for i in range(40):
16     speech.update()
17     lda1.update()
18     tt.update()
19     vae1.update()
20     gmm1.update()
21     tt.update()
```

Fig. 7 Source code of main part of the implementation example

Code

Figure 7 shows the source code of the main part of the implementation. The observations are loaded from files in lines 1–2, the modules to be used are defined in lines 4–8, the connections between the modules are defined in lines 10–13, and the parameters are estimated in lines 15–21. The total number of lines is less than 80, including other parts such as import syntax and the definition of the structure of VAE. Note that even more complicated models can also be implemented in a few hundreds of lines, and we believe the current Neuro-SERKET has scalability with regard to programmatic implementation.

Conditions

During the experiment, we used a hand-written digit dataset, MNIST[38], and a spoken Japanese number dataset [39], as the image data and speech signals, respectively. Each pair of data consists of image data and a spoken audio signal corresponding to a number among $\{0, \ldots, 9\}$. In total, 3000 pairs are used. The pronunciation of Japanese digits is shown in Table 3.

We used VAE, whose encoder and decoder have a middle layer with 128 nodes and a hidden layer with ten nodes, i.e., the dimension of the latent space was 10. The number of classes of GMM and LDA was $K = 10$. We used Julius for the ASR module. We used a standard GMM-based acoustic model preset in Julius,

Table 3 Pronunciation of Japanese numbers used in the experiment

Number	Japanese pronunciation
0	ze ro
1	i chi
2	ni
3	sa n
4	yo n
5	go
6	ro ku
7	na na
8	ha chi
9	kyu u

and a language model in which all syllables have the same probability as an initial language model. The number of samples obtained from the ASR module was $L = 10$.

During the experiment, we compared the following four models.

VAE GMM LDA ASR No communication among the modules
VAE GMM LDA + ASR Communication between LDA and ASR
VAE + GMM LDA + ASR Communication between VAE and GMM and between LDA and ASR
VAE + GMM + LDA + ASR Communication among all modules

In the name of each model, '+' represents the existence of communication between the two modules, and ' ' (white space) indicates no communication between the two neighboring modules. Note that none of the three connections were not supported in SERKET framework.

In the learning process, the model is trained in an off-line manner. Posterior probabilities for all data points are calculated and were given to the neighbor modules. When a module is updated, the global parameters of the module was reset once and trained using the received data and messages. In each update, VAE was trained 200 epochs with batch size = 500, and GMM and LDA were trained with Gibbs sampling procedure with 50 and 100 times sampling, respectively. VAE+GMM, LDA+ASR, and VAE+GMM+LDA+ASR were updated 50 times, i.e., until they converged. The order of the update were *ASR* → *LDA* (from LDA to GMM) → *TtoT* → *VAE* → *GMM* → *TtoT* (from LDA to GMM) → *ASR*.

The average of accuracy was calculated by referring to the ground-truth category of the digits for each condition.

Table 4 Classification accuracy in the GMM and LDA modules

Model	Accuracy (%)		Features introduced in Neuro-SERKET		
	GMM	LDA	Head-to-head	Tail-to-tail	Neural net
VAE GMM LDA ASR	62.0	27.4			
VAE GMM LDA + ASR	62.0	91.8	✓		
VAE + GMM LDA + ASR	63.7	91.8	✓		✓
VAE + GMM + LDA + ASR	**91.0**	**93.7**	✓	✓	✓

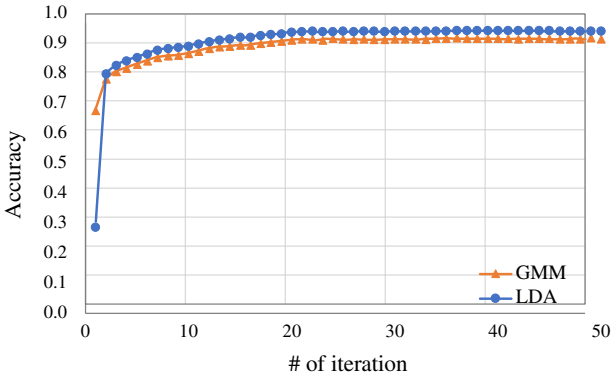

Fig. 8 Transition of classification accuracy

Results

Table 4 shows the average level of accuracy for each clustering module, i.e., GMM and LDA modules.

The performance of the GMM module is slightly increased by introducing communication with a VAE. In contrast, the performance of the LDA module is significantly increased from 27.5 to 92.7% by updating both modules by introducing head-to-head communication, which is newly introduced in Neuro-SERKET, between the LDA and ASR modules. The language model in the ASR is updated by referring to the probabilistic clustering result of the LDA module, and it is thought that the ASR outputs words that are relatively easy for the LDA module to cluster. In addition, by sharing information between the GMM and LDA modules using the tail-to-tail module, which is also newly introduced in Neuro-SERKET, the performance of the GMM module was also significantly increased by approximately 25%. The performance of the LDA module is also slightly increased.

Figure 8 shows the transition of the classification accuracy of VAE + GMM + LDA + ASR. It shows that the performances of the LDA and GMM modules gradually increased.

Figure 9 shows the representations learned by the VAE. The ten-dimensional latent space of the VAE is compressed into a two-dimensional space using a

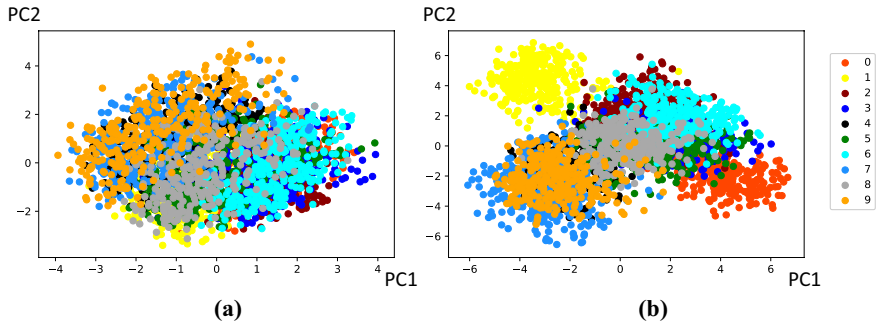

Fig. 9 Latent space of VAE learned using **a** VAE GMM LDA ASR and **b** VAE + GMM + LDA + ASR. Proportion of variance for [PC1, PC2] are [0.15, 0.13] and [0.24, 0.20] for **a**, **b**, respectively

principal component analysis (PCA) for visualization. Each color represents a digit. Figure 9a, b shows the results of embedding without and with communication, i.e., without SERKET and with Neuro-SERKET, respectively. This result shows that VAE + GMM + LDA + ASR formed an appropriate latent space for clustering using the GMM module.

Next, we observed clustered words. Each cluster involves numerous words having recognition errors. To check if each cluster corresponds to a digit, we picked up a stereotypical word, i.e., a syllable sequence, $\bar{\boldsymbol{s}}_c$, using the following equations:

$$\bar{j}_c = \text{argmin}_j \frac{1}{I_c} \sum_i^{I_c} D(\boldsymbol{s}_{cj}, \boldsymbol{s}_{ci}), \tag{25}$$

$$\bar{\boldsymbol{s}}_c = \boldsymbol{s}_{c\bar{j}}, \tag{26}$$

where I_c is the number of words classified into class c; $\boldsymbol{s}_{ci}$ is the ith word, i.e., the syllable sequence, classified into class c, and $D(\cdot, \cdot)$ represents the edit distance between the two-syllable sequence. This procedure selects a word that is nearest to the center of the set of words in terms of the edit distance. Therefore, we can consider $\bar{\boldsymbol{s}}_c$ as a stereotype of class c. The determined stereotypes of each class are shown in Table 5. Compared with Table 3, we can see that unsupervised learning using VAE + GMM + LDA + ASR can acquire an appropriate syllable sequence for each number.

Conclusion

To develop an integrative cognitive system using PGMs more efficiently, we require a useful framework that allow us to reuse elemental cognitive modules developed by other researchers and developers. This paper described Neuro-SERKET, which is a framework for developing a complex cognitive system by composing elemental PGMs. Neuro-SERKET is an extension of SERKET, which can compose elemental PGMs developed in a distributed manner. Although SERKET only supports a

Table 5 Stereotypical word in each class

Number	Japanese pronunciation
0	ze ro
1	i chi *i*
2	ni *n i*
3	sa n
4	yo n
5	go *o*
6	ro ku
7	na *n* na *a*
8	ha chi
9	kyu u

The italic characters denote errors

head-to-tail connection, Neuro-SERKET supports tail-to-tail and head-to-head connections. In addition, Neuro-SERKET supports neural network-based modules, e.g., deep generative models such as VAEs, which are not supported by conventional SERKET. As an example application of Neuro-SERKET, an integrative model called VAE + GMM + LDA + ASR was developed by composing VAE, GMM, LDA, and ASR. The performance of VAE + GMM + LDA + ASR and the validity of Neuro-SERKET are demonstrated through a multimodal categorization task using image data and the speech signal of numerical digits.

In this paper, we showed only one example, i.e., VAE + GMM + LDA + ASR, and demonstrated the validity of Neuro-SERKET. Further application of Neuro-SERKET and the development of cognitive systems that enable a robot to form concepts, learn behaviors, and acquire language in a real-world environment is our future challenge. In particular, it has become clear that language learning in a real-world environment requires a wide range of cognitive capabilities [40]. For this reason, at least two additional approaches should be applied to Neuro-SERKET.

The first one is the development of a software environment, i.e., software libraries. Nakamura et al. has been developing SERKET.[7] As described in this paper, the Neuro-SERKET framework fully includes the conventional SERKET framework. Therefore, the SERKET software environment should be naturally extended to the Neuro-SERKET software environment. To involve DGMs into the SERKET framework, making use of a pre-existing software library for the DGMs may be a reasonable solution. In addition, Suzuki et al. have been developing Pixyz[8] which is a framework for DGMs. We are now working on an efficient utilization of Pixyz for Neuro-SERKET. We consider it is important to combine unsupervised learning by probabilistic models and representation learning

[7] SERKET: http://serket.naka-lab.org/.

[8] Pixyz: https://github.com/masa-su/pixyz.

by NNs such as VAEs, i.e., DGMs. As shown in this paper, the latent space suitable for classification can be learned by the interaction between them. We have also proposed the method for motion segmentation where GP-HSMM, which is a probabilistic model, and VAE, which can extract features from motions, are connected and we showed that low-dimensional features suitable for segmentation can be learned by the interaction between them. We consider such a composed model of unsupervised learning and representation learning has the potential to solve the various problem and it is possible to construct these models easily by Neuro-SERKET.

The second is an exploration of the applicability of Neuro-SERKET. In the current version, the Neuro-SERKET framework heavily relies on a PoE approximation. The limitation of a PoE approximation should be investigated both theoretically and empirically. A series of studies forming the background of Neuro-SERKET are developing cognitive systems that can perform life-long learning in a real-world environment. Such a learning process involves behavioral learning and language acquisition. For this purpose, the system will receive unstructured sensorimotor data. Theoretical and empirical validations should be applied for further applications. So far, many researchers, including the authors, have proposed a lot of cognitive models for robots: object concept formation based on its appearance, usage and functions [41], formation of integrated concept of objects and motions [42], grammar learning [16], language understanding [43], spatial concept formation and lexical acquisition [8, 20, 44], simultaneous phoneme and word discovery [45–47] and cross-situational learning [48, 49]. These models are regarded as an integrative model that are constructed by combining small-scale models. Therefore, they can be also re-implemented using Neuro-SERKET more efficiently.

The computational efficiency needs to be improved as well. The most of modules are implemented using pure python without parallel computation in current Neuro-SERKET except for VAE, which is implemented using TensorFlow. Therefore, parameter estimation is not so fast. The parameter estimation in independent modules can be parallelized, and it might be faster by implementing using C language and TensorFlow. We plan to improve these drawbacks in the future. Regarding an optimization policy, we manually set the order of modules to be updated in the experiment. However, we also found that the performance of the whole model changed depending on the order of modules to be updated. Therefore, we will study and create the guideline about the order of the models to be updated for the practical use of SERKET.

Neuro-SERKET allows us to focus on the integration and exploration of complex cognitive systems. Recently, multimodal learning with DGMs has been gaining attention. However, as the cerebral cortex in our human brain demonstrates, the human cognitive system is based on mutually connected cortical areas, which are considered to have respective functions and modality-dependent information processing. Doya hypothesized that the cerebral cortex is trained simply through unsupervised learning [50]. In general, unsupervised learning is modeled by PGMs. Neuro-SERKET enables us to explore a constructive model of the cerebral cortex using PGMs. Such exploration and the development of a brain-inspired whole-brain

cognitive architecture are also future challenges. We believe that Neuro-SERKET will be a key framework for the future constructive studies on general intelligence and symbol emergence in natural and artificial cognitive systems [1, 9].

Acknowledgements This research was supported by MEXT/JSPS KAKENHI, Grant nos. 16H06569 in #4805 (Correspondence and Fusion of Artificial Intelligence and Brain Science) and 18H03308, and JST CREST (JPMJCR15E3).

References

1. Taniguchi, T., Ugur, E., Hoffmann, M., Jamone, L., Nagai, T., Rosman, B., Matsuka, T., Iwahashi, N., Oztop, E., Piater, J. et al.: Symbol emergence in cognitive developmental systems: a survey. IEEE Trans. Cogn. Dev. Syst. (2018)
2. Nakamura, T., Nagai, T., Iwahashi, N.: Multimodal object categorization by a robot. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2415–2420 (2007)
3. Shun, N., Tetsuya, O., Jun, T., Kazunori, K., Hiroshi, O.G.: Predicting object dynamics from visual images through active sensing experiences. Adv. Robot. **22**(5), 527 (2008)
4. Ogata, T., Nishide, S., Kozima, H., Komatani, K., Okuno, H.: Inter-modality mapping in robot with recurrent neural network. Pattern Recogn. Lett. **31**(12), 1560 (2010)
5. Mangin, O., Filliat, D., Ten Bosch, L., Oudeyer, P.Y.: MCA-NMF: multimodal concept acquisition with non-negative matrix factorization. PLoS One **10**, 10, e0140732 (2015)
6. Sinapov, J., Schenck, C., Staley, K., Sukhoy, V., Stoytchev, A.: Grounding semantic categories in behavioral interactions: experiments with 100 objects. Robot. Auton. Syst. **62**(5), 632 (2014)
7. Miyazawa, K., Aoki, T., Hieida, C., Iwata, K., Nakamura, T., Nagai, T.: Integration of multimodal categorization and reinforcement learning for robot decision-making. In: IROS2017: Workshop on Machine Learning Methods for High-Level Cognitive Capabilities in Robotics (2017)
8. Taniguchi, A., Hagiwara, Y., Taniguchi, T., Inamura, T.: Online spatial concept and lexical acquisition with simultaneous localization and mapping. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE), pp. 811–818 (2017)
9. Taniguchi, T., Nagai, T., Nakamura, T., Iwahashi, N., Ogata, T., Asoh, H.: Symbol emergence in robotics: a survey. Adv. Robot. **30**(11–12), 706 (2016)
10. Tani, J.: Exploring Robotic Minds: Actions, Symbols, and Consciousness as Self-Organizing Dynamic Phenomena. Oxford University Press, Oxford (2016)
11. Nakamura, T., Nagai, T., Taniguchi, T.: SERKET: An Architecture For Connecting Stochastic Models to Realize a Large-Scale Cognitive Model. arXiv:1712.00929 (arXiv preprint) (2017)
12. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: International Conference on Learning Representations (2014)
13. Roy, D., Pentland, A.: Learning words from sights and sounds: a computational model. Cogn. Sci. **26**(1), 113 (2002)
14. Nakamura, T., Araki, T., Nagai, T., Iwahashi, N.: Grounding of word meanings in LDA-based multimodal concepts. Adv. Robot. **25**, 2189 (2012)
15. Yamada, T., Matsunaga, H., Ogata, T.: Paired recurrent autoencoders for bidirectional translation between robot actions and linguistic descriptions. IEEE Robot. Autom. Lett. **3**(4), 3441–3448 (2018)

16. Attamimi, M., Ando, Y., Nakamura, T., Nagai, T., Mochihashi, D., Kobayashi, I., Asoh, H.: Learning word meanings and grammar for verbalization of daily life activities using multilayered multimodal latent Dirichlet allocation and Bayesian hidden Markov models. Adv. Robot. **30**(11–12), 806 (2016)
17. Nishihara, J., Nakamura, T., Nagai, T.: Online algorithm for robots to learn object concepts and language model. IEEE Trans. Cogn. Dev. Syst. **9**(3), 255 (2017)
18. Ando, Y., Nakamura, T., Araki, T., Nagai, T.: Formation of hierarchical object concept using hierarchical latent Dirichlet allocation. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2272–2279 (2013)
19. Hagiwara, Y., Inoue, M., Kobayashi, H., Taniguchi, T.: Hierarchical spatial concept formation based on multimodal information for human support robots. Front. Neurorobot. **12**, 11 (2018)
20. Taniguchi, A., Taniguchi, T., Inamura, T.: Spatial concept acquisition for a mobile robot that integrates self-localization and unsupervised word discovery from spoken sentences. IEEE Trans. Cogn. Dev. Syst. **8**(4), 285 (2016)
21. Iwata, K., Aoki, T., Horii, T., Nakamura, T., Nagai, T.: Learning and generation of actions from teleoperation for domestic service robots. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 8184–8191 (2018)
22. Kingma, D.P., Mohamed, S., Rezende, D.J., Welling, M.: Semi-supervised learning with deep generative models. In: Advances in Neural Information Processing Systems, pp. 3581–3589 (2014)
23. Johnson, M., Duvenaud, D.K., Wiltschko, A., Adams, R.P., Datta, S.R.: Composing graphical models with neural networks for structured representations and fast inference. In: Advances in Neural Information Processing Systems, pp. 2946–2954 (2016)
24. Dilokthanakul, N., Mediano, P.A., Garnelo, M., Lee, M.C., Salimbeni, H., Arulkumaran, K., Shanahan, M.: Deep unsupervised clustering with gaussian mixture variational autoencoders. arXiv:1611.02648 (arXiv preprint) (2016)
25. Ebbers, J., Heymann, J., Drude, L., Glarner, T., Haeb-Umbach, R., Raj, B.: Hidden Markov model variational autoencoder for acoustic unit discovery. In: INTERSPEECH, pp. 488–492 (2017)
26. Jiang, Z., Zheng, Y., Tan, H., Tang, B., Zhou, H.: Variational deep embedding: an unsupervised and generative approach to clustering. arXiv:1611.05148 (arXiv preprint) (2016)
27. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. In: Advances in Neural Information Processing Systems, pp. 3483–3491 (2015)
28. Pandey, G., Dukkipati, A.: Variational methods for conditional multimodal deep learning. In: 2017 International Joint Conference on Neural Networks (IJCNN) (IEEE), pp. 308–315 (2017)
29. Mansimov, E., Parisotto, E., Ba, J.L., Salakhutdinov, R.: Generating images from captions with attention. arXiv:1511.02793 (arXiv preprint) (2015)
30. Suzuki, M., Nakayama, K., Matsuo, Y.: Joint multimodal learning with deep generative models. arXiv:1611.01891 (arXiv preprint) (2016)
31. Wu, M., Goodman, N.: Multimodal generative models for scalable weakly-supervised learning. In: Advances in Neural Information Processing Systems, pp. 5575–5585 (2018)
32. Hinton, G.E.: Training products of experts by minimizing contrastive divergence. Neural Comput. **14**(8), 1771 (2002)
33. Jo, D.U., Lee, B., Choi, J., Yoo, H., Choi, J.Y.: Cross-modal variational auto-encoder with distributed latent spaces and associators. arXiv:1905.12867 (arXiv preprint) (2019)
34. Hamme, A.V.: HAC-models: a novel approach to continuous speech recognition. In: Annual Conference of the International Speech Communication Association, pp. 2554–2557 (2008)
35. Srivastava, A., Sutton, C.: Autoencoding variational inference for topic models. In: International Conference on Learning Representations (2017)
36. Gildea, D., Hofmann, T.: Topic-based language models using EM. In: Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECH) (1999)
37. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993 (2003)
38. LeCun, Y., Cortes, C., Burges, C.: Mnist handwritten digit database. http://yann.lecun.com/exdb/mnist
39. Reverberant speech recognition evaluation environment (censrec-4). http://research.nii.ac.jp/src/en/CENSREC-4.html
40. Tangiuchi, T., Mochihashi, D., Nagai, T., Uchida, S., Inoue, N., Kobayashi, I., Nakamura, T., Hagiwara, Y., Iwahashi, N., Inamura, T.: Survey on frontiers of language and robotics. Adv. Robot. **33**(15–16), 700 (2019). https://doi.org/10.1080/01691864.2019.1632223
41. Nakamura, T., Nagai, T.: Object concept modeling based on the relationship among appearance, usage and functions. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IEEE), pp. 5410–5415 (2010)

42. Fadlil, M., Ikeda, K., Abe, K., Nakamura, T., Nagai, T.: Integrated concept of objects and human motions based on multi-layered multimodal LDA. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IEEE), pp. 2256–2263 (2013)
43. Kobori, T., Nakamura, T., Nakano, M., Nagai, T., Iwahashi, N., Funakoshi, K., Kaneko, M.: Robust comprehension of natural language instructions by a domestic service robot. Adv. Robot. **30**(24), 1530 (2016)
44. Ishibushi, S., Taniguchi, A., Takano, T., Hagiwara, Y., Taniguchi, T.: Statistical localization exploiting convolutional neural network for an autonomous vehicle. In: IECON 2015–41st Annual Conference of the IEEE Industrial Electronics Society, pp. 001,369–001,375 (2015). https://doi.org/10.1109/IECON.2015.7392291
45. Taniguchi, T., Nagasaka, S., Nakashima, R.: Nonparametric bayesian double articulation analyzer for direct language acquisition from continuous speech signals. IEEE Tran. Cogn. Dev. Syst. **8**(3), 171 (2016)
46. Taniguchi, T., Nakashima, R., Liu, H., Nagasaka, S.: Double articulation analyzer with deep sparse autoencoder for unsupervised word discovery from speech signals. Adv. Robot. **30**(11–12), 770 (2016)
47. Nakashima, R., Ozaki, R., Taniguchi, T.: Unsupervised phoneme and word discovery from multiple speakers using double articulation analyzer and neural network with parametric bias. Front. Robot. AI **6**, 92 (2019)
48. Taniguchi, A., Taniguchi, T., Cangelosi, A.: Cross-situational learning with Bayesian generative models for multimodal category and word learning in robots. Front. Neurorobot. **11**, 66 (2017)
49. Aly, A., Taniguchi, A., Taniguchi, T.: A generative framework for multimodal learning of spatial concepts and object categories: an unsupervised part-of-speech tagging and 3D visual perception based approach. In: 2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob), pp. 376–383 (2017). https://doi.org/10.1109/DEVLRN.2017.8329833
50. Doya, K.: What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? Neural Netw. **12**(7–8), 961 (1999)

Tadahiro Taniguchi received his ME and PhD degrees from Kyoto University, in 2003 and 2006, respectively. From April 2005 to March 2006, he was a Japan Society for the Promotion of Science (JSPS) Research Fellow (DC2) at the Department of Mechanical Engineering and Science, Graduate School of Engineering, Kyoto University. From April 2006 to March 2007, he was a JSPS Research Fellow (PD) at the same department. From April 2007 to March 2008, he was a JSPS Research Fellow at the Department of Systems Science, Graduate School of Informatics, Kyoto University. From April 2008 to March 2010, he was an assistant professor at the Department of Human and Computer Intelligence, Ritsumeikan University. From April 2010 to March 2017, he was an associate professor at the same department. From September 2015 to September 2016, he is a visiting associate professor at Department of Electrical and Electronic Engineering, Imperial College London. From April 2017, he has been a professor at the Department of Information and Engineering, Ritsumeikan University. From April 2017, he has been a visiting general chief scientist, AI solution center, Panasonic, as well. He has been engaged in research on machine learning, emergent systems, intelligent vehicle and symbol emergence in robotics.

Tomoaki Nakamura received his BE, ME, and Dr of Eng. degrees from the University of Electro-Communications in 2007, 2009, and 2011. From April 2011 to March 2012, he was a research fellow of the Japan Society for the Promotion of Science. In 2013, he worked for Honda Research Institute Japan Co., Ltd. From April 2014 to March 2018, he was an assistant professor at the Department of Mechanical Engineering and Intelligent Systems, the University of Electro-Communications. Since April 2019, he has been an associate professor at the same department. His research interests are intelligent robotics and machine learning.

Masahiro Suzuki received his ME degree from Hokkaido University in 2015 and PhD degree from the University of Tokyo in 2018. He has been a project researcher at the School of Engineering, the University of Tokyo, since 2018. His research fields are artificial intelligence and machine learning, with a particular interest in transfer learning, multimodal learning, and deep generative models.

Ryo Kuniyasu received his bachelor's degree from the University of Electoro-Communications in 2019. He is currently in a master's course at the graduate school of Informatics and Engineering, University of Electro-Communications. His research interests are intelligent robotics and machine learning.

Kaede Hayashi received her BE and ME degrees in information science and engineering from Ritsumeikan University, Japan, in 2016 and 2018, respectively. She is currently a PhD student at Ritsumeikan University where she also works as a research assistant. Her research fields include artificial intelligence and symbol emergence, with a particular interest in deep generative models and their applications.

Akira Taniguchi received his BE, ME, and PhD degrees from Ritsumeikan University, Japan, in 2013, 2015, and 2018, respectively. From 2017 to 2019, he was a research fellow of the Japan Society for the Promotion of Science. Since 2019, he has been working as a specially appointed assistant professor at the College of Information Science and Engineering, Ritsumeikan University, Japan. His research interests include intelligent robotics, artificial intelligence, and symbol emergence in robotics.

Dr. T Horii received the ME and PhD degrees from Osaka University, Osaka, Japan, in 2013 and 2018, respectively. He was a project assistant professor with the University of Electro-Communications, from 2018 to 2019 and became an assistant professor with Osaka University in 2019. He is also a visiting researcher with the International Research Center for Neurointelligence, the University of Tokyo. His research interests include the computational modeling of emotion, emotional human–robot communication, and tactile interaction.

Takayuki Nagai received his BE, ME, and PhD degrees from the Department of Electrical Engineering, Keio University, in 1993, 1995, and 1997, respectively. Since 1998, he had been with the University of Electro-Communications, and from 2018 he has been a professor of the graduate school of Engineering Science, Osaka University. From 2002 to 2003, he was a visiting scholar at the Department of Electrical Computer Engineering, University of California, San Diego. He also serves as a specially appointed professor at UEC AIX, a visiting researcher at Tamagawa University Brain Science Institute, and a visiting researcher at AIST AIRC. He has received IROS Best Paper Award Finalist, Advanced Robotics Best Paper Award, JSAI Best Paper Award, etc. His research interests include intelligent robotics, cognitive developmental robotics, and robot learning. He aims at realizing flexible and general intelligence like human by combining AI and robot technologies.

Affiliations

Tadahiro Taniguchi[1] · Tomoaki Nakamura[3] · Masahiro Suzuki[4] · Ryo Kuniyasu[3] · Kaede Hayashi[1] · Akira Taniguchi[1] · Takato Horii[2] · Takayuki Nagai[2,3]

Tomoaki Nakamura
tnakamura@uec.ac.jp

Masahiro Suzuki
masa@weblab.t.u-tokyo.ac.jp

Ryo Kuniyasu
r_kuniyasu@radish.ee

Kaede Hayashi
k.hayashi@em.ci.ritsumei.ac.jp

Akira Taniguchi
a.taniguchi@em.ci.ritsumei.ac.jp

Takato Horii
takato@sys.es.osaka-u.ac.jp

Takayuki Nagai
nagai@sys.es.osaka-u.ac.jp

[1] Ritsumeikan University, 1-1-1 Noji-higashi, Kusatsu, Shiga, Japan

[2] Osaka University, 1-3 Machikane-yama, Toyonaka, Osaka, Japan

[3] The University of Electro-Communications, 1-5-1 Chofugaoka, Chofu, Tokyo, Japan

[4] The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo, Japan