Cognitive Computation and Systems

Research Article

IET Journals
The Institution of Engineering and Technology

Disentanglement in conceptual space during sensorimotor interaction

eISSN 2517-7567
Received on 15th April 2019
Accepted on 1st October 2019
E-First on 27th November 2019
doi: 10.1049/ccs.2019.0007
www.ietdl.org

Junpei Zhong[1,2] ✉, *Tetsuya Ogata*[1,3], *Angelo Cangelosi*[4], *Chenguang Yang*[5]

[1]*Social Intelligence Research Team, Artificial Intelligence Research Center, National Institute of Advanced Science and Technology (AIST), Aomi 2-3-26, Tokyo, Japan*
[2]*School of Science and Technology, Nottingham Trent University, Clifton Lane, Nottingham NG11 8NS, UK*
[3]*Department of Intermedia Art and Science, Waseda University, Tokyo, Japan*
[4]*University of Manchester, Manchester, Oxford Road, Manchester M13 9PL, UK*
[5]*Bristol Robotics Laboratory, University of the West of England, Bristol BS16 1QY, UK*
✉ *E-mail: zhong@junpei.eu*

Abstract: The disentanglement of different objective properties from the external world is the foundation of language development for agents. The basic target of this process is to summarise the common natural properties and then to name it to describe those properties in the future. To realise this purpose, a new learning model is introduced for the disentanglement of several sensorimotor concepts (e.g. sizes, colours and shapes of objects) while the causal relationship is being learnt during interaction without much a priori experience and external instructions. This learning model links predictive deep neural models and the variational auto-encoder (VAE) and provides the possibility that the independent concepts can be extracted and disentangled from both perception and action. Moreover, such extraction is further learnt by VAE to memorise their common statistical features. The authors examine this model in the affordance learning setting, where the robot is trying to learn to disentangle about shapes of the tools and objects. The results show that such a process can be found in the neural activities of the β-VAE unit, which indicate that using similar VAE models is a promising way to learn the concepts, and thereby to learn the causal relationship of the sensorimotor interaction.

1 Introduction

Concepts in cognitive processes are defined as the internal representations, which is the foundation for the agent to understand the external world and to proceed to do cognitive manipulation. The preliminary step for a robot to do conceptualisation in an embodied world is to do the categorisation of the rich data in the continuous high-dimensional sensorimotor space by the following procedures:

i. Learning the common structure of the data in the high-dimensional space by their statistical regularities, in either supervised or self-supervised way.
ii. Defining the hyperparameters to describe the common structure, which can be seen as the 'concepts', which is the abstract description on one of the dimensions of the categorised space.

During the developmental steps of human, we seem to be good at doing the conceptualisation learning, even without any a priori knowledge about the set of representations of the categorisation space or the hyperparameters in the statistical regularities. For example, without any a priori knowledge about the spatial representations or directional representations, humans construct the navigation system primarily by the spatial learning of its own and by their symbolic meanings of the environment [1]. There is a more intriguing fact that humans can also associate the spatial concepts of the four directions even without any words to describe them. Such a priori knowledge exists cross-culturally and between different languages, which may indicate that such basic concepts of four different directions are already embedded in our brain and our body. Most of the behavioural studies also reveal that there are time differences in the learning processes of concrete and abstract concepts, suggesting that the learning of abstract concepts may depend on the concrete ones [2, 3]. Therefore, learning of such concepts follows a hierarchical way: (i) the concrete concepts (e.g. concepts of red and blue colours) which are the foundation of the abstract concepts (e.g. the concept of 'colour' as a property of the visual inputs) and (ii) the relationships between the concrete grounded concepts and the abstract ones, validated by the further cues perceived.

Furthermore, while we are doing the concept learning, our brain is always attempting to categorise the sensorimotor data based on its similarities from the cross-modal sensorimotor inputs: concepts can be abstracted by the similarities existing in the visual inputs (e.g. the categorisation of the concepts about 'size' or 'colour' from visual stimuli) or from the motor action together with the external environment. In the example of learning the 'numbers', the further cues may include the motor action ('counting' as a motor action) or other innate mechanisms that facilitate to deal with the basic visual perceptual information [4, 5]. The dependent relation from the concrete concepts to the abstract concepts can also be traced in the similarities in linguistic knowledge. With the help of the multi-sensory cues, as well as the innate mechanism of the basic knowledge, the human brain can learn the concept without rich amount of data with a much faster rate than most of the artificial learning methods. This is also much more advanced than the state-of-the-art deep learning methods (e.g. [6]), in terms of learning speed, as well as its ability to disentangle different concepts which may further result in symbolic manipulation such as logical deduction of causal relationship and event prediction based on context. It indicates that there are much more technologies to be explored to achieve human-level intelligence. For instance, a few embodied models (see also [7]) have been developed to learn the specific categories of the abstract concepts, which are rooted in the concrete concepts. However, such hierarchical architecture based mostly on recurrent neural network assumes that the abstract concepts are rooted from single modal as well as supervised learning, which is different from the developmental stages of humans.

1.1 Conceptualisation during sensorimotor interaction

Learning the concepts of the visual object by sensorimotor interaction is one of the challenges for embodied agents. This is done by the process of affordance learning. The original meaning of 'affordance' in psychology denotes it as 'the constant properties of constant objects are perceived (the shape, size, colour, texture, composition, motion, animation and position relative to other objects) (for a detailed review, please see [8])'. It indicates that the affordance is understood as general descriptions of the objects, and they are the direct causal results of the properties of the objects, most of which can be perceived by the visual inputs and understood by the human brain.

Note that though the affordance is independent to the agent itself, the configuration could also affect the understanding of the affordance. For instance, a robot has only one gripper with 1-degree of freedom cannot understand the affordance of the piano. Therefore, the common setting for affordance learning for robots is putting it in an exposed environment and following the end-to-end learning scheme with the interactions [9]. Therefore, in the embodied point of view, the term 'affordance' should be understood with the agent, in the context of execution or observation of motor actions. In such scenarios, the agent should be able to learn the 'affordance' with relation to the environment/object and the actions. Instead, the key factors that constitute the abstract concept of 'affordance' are some of the concrete concepts obtained from the 'users by the integration of perception and action'. When the agent observes the object from visual perception and thus learns the causality relation between the visual features, the voluntary motor action and movements, and then conceptualise them as different perspectives of the affordances.

The conceptualisation during sensorimotor interaction is closely associated with the language components for language acquisition. In [10], the robot learns the relation between the object affordances and verbal interaction with a human caregiver. Similar models are proposed in [11, 12], where the relations between words (nouns, adjectives and verbs) and objects properties (including their affordances) are represented using a set of support vector machine classifiers. Nevertheless, on the other hand, the learning of word components is not necessary for learning concepts, since the toddlers know such concepts even earlier than their early word acquisition. While they are playing with objects and attempting to do object manipulation, they are already well aware of the affordance in the causal results of different properties in the sensorimotor inputs such as the shape and the size of the objects [13]. In this way, it implies another way to develop the learning model of different concepts, especially the concepts related to embodied interaction with the objects' affordance. During interaction, it shapes the knowledge about the object from the understanding affordances in the conceptual space. Although the conceptualisation can be also regarded as the grounding of abstract concepts, which associates between the visual feature of the tools and its effect while it is interacting with an object, a few literatures have been proposed to explain the grounding of visual concepts by interpreting the 'affordance' of objects and other entities in the sensorimotor interaction.

1.2 Using generative models to conceptualise data

In the context of machine learning, to accomplish the task of doing conceptualisation, the low-dimensional conceptualised representation should be extracted from the high-dimensional data. Among various kinds of methods to do such an extraction, generative models are unsupervised learning methods to understand the data distribution by building the joint probability distribution of data $P(X, Y)$ between the observed data X and target data distribution $P(Y)$. Specifically, as one of the generative models, the deep generative models (DGMs) have recently widely used in various domains such as data compression, data denoising etc., since it provides the joint distribution which can be further used to reconstruct random instances, either by estimating the joint distribution of (X, Y) or constructing the observable data X by the given target Y, i.e. $P(X|Y = y)$. The DGMs are generally constructed by multiple-layer neural networks and learnt by stochastic or deterministic learning mechanisms, which encode the statistical information of input data in a hierarchical way. Some of the layer parameters are learnt with stochastic functions, for instance, the sigmoid belief networks [14] and various types of Helmholtz machines [15, 16]. Alternatively, recent architectures such as generative adversarial networks (GANs) and variational auto-encoder (VAE) use deterministic functions to do the optimisation. The major advantage of using stochastic functions is that it has more expressive power to learn the data with more stochastic noise, and is easier to avoid the local minima. In both cases, the DGM models use the hierarchy of layers to define the conditional densities as the directed graphical models. Therefore, due to these conditional densities and the given input data, while the generative models are regarded as the directed graphs to learn and compress the input data or vice versa, the DGMs learn to model a joint distribution $P(X, Y)$. Such compression is done in the latent space, where each unit represents the most informative bit in the DGM. In extreme cases, these most informative units in the latent space may play a similar role of categorisation. While the latent units are used to encode the images, further details of the images, which may be interpreted as the *concepts* the common properties of the visual field, can be encoded in the latent space. In the rest of this paper, we call this kind the representation in the latent space 'conceptual compression' or 'conceptualisation' because it exhibits the higher levels of representation to describe the properties of the sensorimotor space.

Using the concept compression units in machine learning models has been used in many applications, most of which employ the encoder–decoder framework. The basic principle is that the pre-trained generator (in GAN) or the decoder (in VAE) approximately captures the regularities of the reconstructed vectors being 'identical' in the input data [17]: both of them define a probability distribution over vectors in sample space and try to assign higher probability of the joint distribution to more likely vectors, for the data-set it has been trained on. The reconstructed vectors are expected to be identical to the input data or at least they will be close to some points in the support of this distribution, i.e. in the range of input data space. In the VAE model, for instance, it is done by creating probability density functions with Gaussian distributions, with which the stochastic regularities can be learnt as the common features of the data. For instance, to recover the noised information, where we can observe the same representation in the latent vectors as the perfect source while only parts of the images are presented [18, 19], and even connect the temporal information of the input data [20] or when the text and the image representation share the same latent vector when the network has been well-trained for the modality translation [21].

In this paper, we present a neural architecture that does not use any hard-coded a priori knowledge in the latent space to disentangle the conceptual information by using the β-VAE model as the highest-level part of the generative model. The β-VAE has recently proved to be an efficient and useful model to learn the conceptual information from images. We applied the β-VAE model together with the predictive model in the framework of perception–action theory [22], which suggests that the perceptual information and the expected perceptual information from the actions are encoded in a common representation. On the basis of this, the learning model employs sensorimotor learning as a mechanism for finding the relationship between the motor execution and the perception of the objects. Such common representation between perception and action can closely relate to the higher-level understanding of the concepts of visual objects. We will apply this model to learn such concepts in the robot–object interaction task, to find out the visual concepts that contribute to the affordances in tool use. The contributions of this work, therefore, lie in the introduction of a novel cognitive learning model, which links the following two requirements inspired from the grounding of visual concepts while the agent is interacting with objects and learning its affordance:

- It is able to disentangle the visual concepts of the objects and tools in the robot–tool–object interaction setting.

- It is able to predict the next possible visual information based on the understanding of perception and action.

2 Model

2.1 Generative hierarchical architecture

The proposed variational action feedback augmented predictive network (VAFA-PredNet) is a hierarchical architecture. It is an extension of the previous versions of AFA-PredNet [23] and multiple time-scale action feedback augmented predictive network (MT-AFA)-PredNet [24, 25]. Inspired by the predictive coding framework, all of these hierarchical neural networks integrate the perception and action in the predictive common-coding framework [22]: the predicted perception is part of the results of the internal visual memory and the voluntary motor actions. In these hierarchical architectures, the higher levels of the network capture the statistical regularities happening on the lower one, whereas the action-modulated prediction is in the process. Specifically, to effectively predict different features from the visual perception, a series of repeating generative models (e.g. convLSTM) are stacked in a hierarchical way. They act as the lower part of the generative model, whereas the VAEs are employed at the topmost part of the generative model.

Modelling the likelihood representation on top of the generative units, the VAE extracts the representation of the memory of the topmost convLSTM unit. Since in the MT-AFA-PredNet [24, 25], the slower-changing neural updates on the topmost of the convLSTM unit only adapts very small differences, whereas the time elapse, the input and reconstruct images in the VAE model can exist very small differences. The detailed introduction of the VAE unit please is shown in Section 2.2. These generative models GU (green in Fig. 1) model the predictions of perception stimuli given the error from the corresponding level of the discriminative unit and the top-down prediction from the upper layer. This computational process is similar to the top-down stream in the predictive coding framework. Furthermore, the bottom-up stream

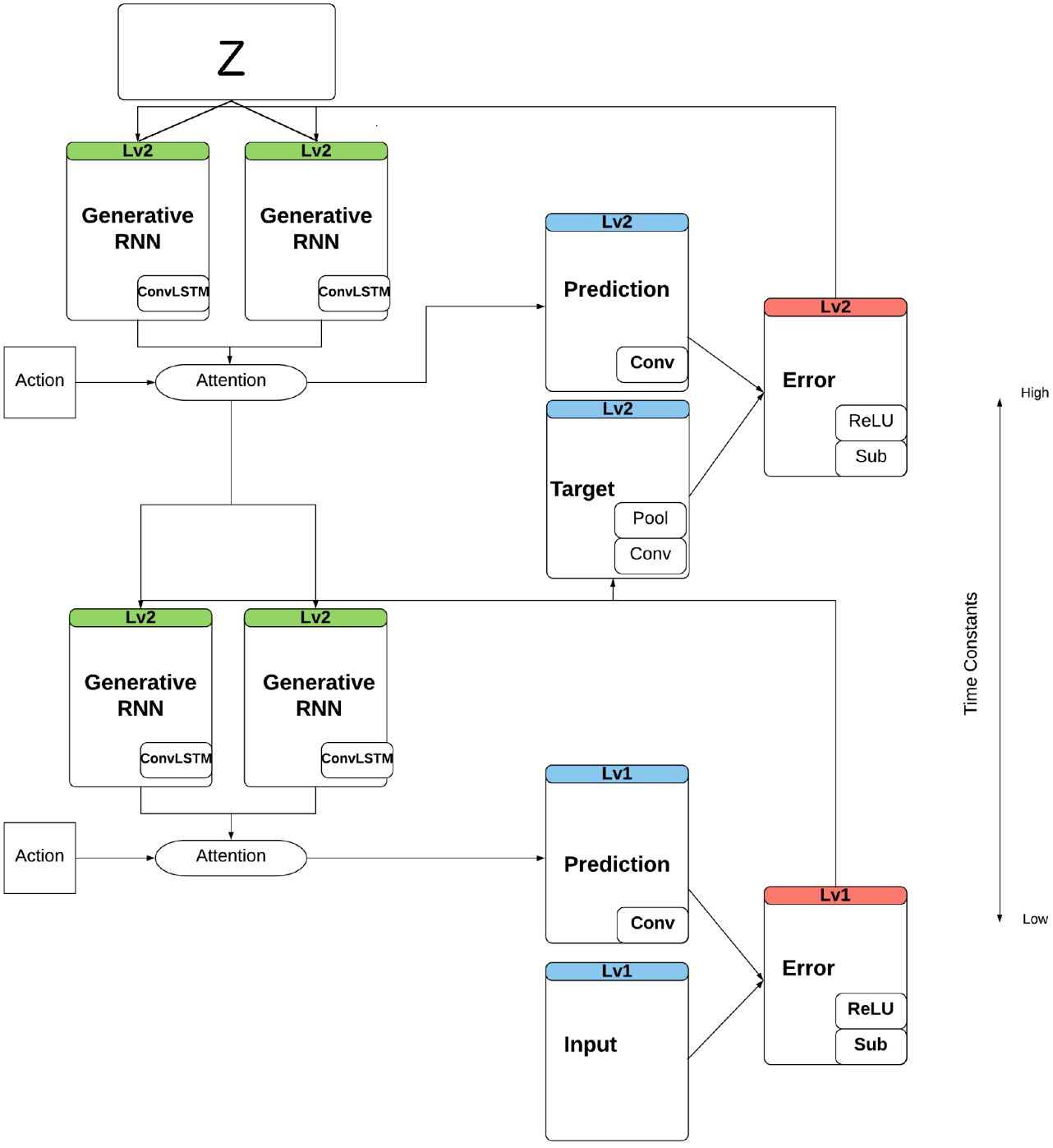


Fig. 1 *2-layer VAFA-PredNet*

shapes the internal representation of the GU models on each layer by giving the error between the predicted one and the real one realised by convolutional networks (discriminative unit (DU), blue). These error signals are computed by the error representation layer (error layer (EL), red) and are calculated from the positive and negative errors between the prediction and the target signals.

To summarise all the aforementioned properties of the architecture, we visualise the model in the case of a 2-layer VAFA-PredNet (Fig. 1)

$$X_l(t) = \begin{cases} i(t), & \text{if } l = 0, \\ \text{MAXPOOL}(f(\text{Conv}(E_{l-1}(t)))), & \text{if } l > 0 \end{cases} \tag{1}$$

$$\hat{X}_l(t) = f(\text{Conv}(R_l(t))) \tag{2}$$

$$E_l(t) = [f(X_l(t) - \hat{X}_l(t)); f(\hat{X}_l(t) - X_l(t))] \tag{3}$$

$$R_l^{\text{d}}(t) = \left(1 - \frac{1}{\tau}\right) R_l^{\text{d}}(t-1) + \frac{1}{\tau}\text{ConvLSTM}(E_l(t-1), R_l(t-1), \text{DevConv}(R_{l+1}(t))) \tag{4}$$

$$R_l(t) = \text{attention}(a(t)) \times R_l^d(t) \tag{5}$$

$$z(t) = \text{encode}(\text{concatenate}(R_{\text{L}}(t))) \tag{6}$$

$$R_{\text{L}}(t+1) = \text{split}(\text{decode}(z(t))) \tag{7}$$

where the concatenate(·) is the concatenation operation to all the internal memory of convLSTM and the split(·) is the reverse function. Here, $f(\cdot)$ is the rectified linear unit or other differentiable activation functions of the neurones and $X(\cdot)_l^t$ is the neural representation of the level l at time t in the discriminate part. The representation of the EL layer l is $E(\cdot)_l$. The MAXPOOL, Conv, ConvLSTM and multi-layer perceptron (MLP) are the corresponding neural algorithms.

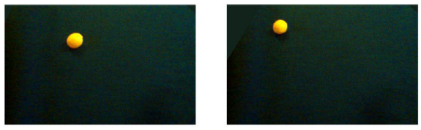
2.2 Variational auto-encoder

In the VAFA-PredNet architecture, the latent vectors in VAE are used to have a conceptual compression from the internal memory of the topmost convLSTM units. In the meanwhile, on the other hand, in the temporal domain, it gives rise to the values of the internal memory of the next time step with the reconstruction.

The first version of VAE was proposed by Kingma and Welling [26]. It is a directed graphical model but is distributed calculated as an auto-encoder. However, being different from the auto-encoders, the VAE learns to construct the latent space $z \in \mathbb{R}^n$ to encode the statistical regularities in the continuous data, by selecting the best parameters of the Gaussian distribution in the latent space. Therefore, the common features of the data can be observed from the forms of variational inference, in the latent space.

To accomplish this process, rather than constructing a direct modelling of the prior data distribution $p(x)$ in the reconstruction, the VAE first models the posterior of the joint distribution of the data $p(z|x)$ as the approximate Gaussian probability density $q_\theta(z|x)$ with the parameter θ. Then, with the assumption that the variational posterior $q_\theta(z|x)$ approximates the true posterior $p(z|x)$, which follows the Gaussian distribution, the $p(z)$ should also be an isotropic unit of the Gaussian distribution:

$$\begin{aligned} p(Z) &= \sum_x p(Z|X)p(X) \\ &= \sum_x N(0, I)p(X) \\ &= N(0, I)\sum_x p(X) \\ &= N(0, I) \end{aligned} \tag{8}$$



a *b*

Fig. 2 *Before/after images in the interaction using a 'hook' to push a 'lemon' (captured from the left camera)*
(a) Before, ***(b)*** After

Moreover, the generative process can be written as $X' \simeq p(X'|Z)$, depends on which distribution of data X' can be reconstructed. So, the generative process defines a joint distribution over data and latent variables $p(X, Z)$. From the neural network point of view, the decoder outputs the parameters to the conditional probability distribution. Thus, we denote it as $p_\phi(X|Z)$. Additionally, though the variational inference of the latent space can be trained by the usual back-propagation, these two parameters θ and ϕ allow the network to be learnt by back-propagation, which is called 'reparameterised'. Different values of the two parameters allow us to adjust the Gaussian distribution by maximising the likelihood of the reconstruction of the data. On the basis of this Gaussian distributions, we can sample from any point of the latent space z and still generate valid and diverse outputs in the reconstruction $p(x|z)$

$$\text{encode}(x) = q_\theta(z|x) \tag{9}$$

$$\text{decode}(z) = p_\phi(x|z) \tag{10}$$

The above two equations indicate that:

i. The latent units z are often referred to as informative units because they should be efficient compressions of the data into this lower-dimensional space. In our case, the VAE unit on top of the perception and action input provides a compressed latent representation of both modalities.
ii. Both the encoder and decoder are stochastic: they output Gaussian probability densities.

To endow an efficient training in both ends, the variational Bayes approach simultaneously learns both the parameters of $p_\theta(x, z)$ as well as those of a posterior approximation $q_\phi(z|x)$. To accomplish this, the posterior $q_\theta(z|x)$ is regularised with its Kullback-Leibler divergence (KL divergence) from a priori distribution $p(z)$. That is why the prior is typically chosen to be also a Gaussian with zero mean and unit variance (8), such that the KL term between posterior and this prior can be computed in the closed form. The loss function to optimise the stochastic construction VAE can be defined as

$$\mathscr{L}(\phi, \theta; x) = \mathbb{E}_{q(z|x)}[\log_{p_\theta}(x, z)] - D_{\text{KL}}(q_\phi(z|x)\big|\big|p(z)) \tag{11}$$

This loss function is called the evidence lower bound (ELBO). There also exists an improved version of VAE which is able to disentangle the concepts of the input data by putting a constraint β on the regularised term so that the learning of the separate concepts in the properties of the input data such as the size and angle can be further disentangled in the latent vector. So in the case of β-VAE, the ELBO loss function was added with a regularised term

$$\mathscr{L}(\phi, \theta; x) = \mathbb{E}_{q(z|x)}[\log_{p_\theta}(x, z)] - \beta D_{\text{KL}}(q_\phi(z|x)\big|\big|p(z)) \tag{12}$$

From the perspective of the neural networks, while being implemented in the graphical model (i.e. auto-encoder), where the representation of the highest layer z is treated as the latent variable where the generative process starts. Since $p(z)$ represents the prior distribution of the latent variable z, the data generation resulting in the reconstruction of the data $\hat{x}$.

Therefore, the VAE units on top of the VAFA-PredNet model, the inputs and outputs of the encoder and decoder are

$$z(t) = \text{encoder}(\text{concatenate}(R(t)); \phi) \tag{13}$$

$$\hat{R}_l(t+1) = \text{split}(\text{decoder}(z(t); \theta)) \tag{14}$$

where the $R_l(t)$ denotes the internal value (memory) of the convLSTM on the lth layer. Since the size of the memory is known, it is easy to do the concatenate or the split operations. The $R(\cdot)$ in the generative model stores the interaction information between the visual images and the action inputs. As such, when the motor action is modulating the generative part of the architecture, the correlation between the perception and action (e.g. the common coding [22]) can be captured in an compressed way in the latent space. Such representation can be further used by the VAE part, which results in more interpretable learning models.

During learning, together with VAE and deep architecture, the complete optimisation function of the whole VAFA-PredNet can be written as the summation of the ELBO (12) and the mean squared error (MSE) loss between the generated image and the original image

$$\mathscr{L}_{\text{Total}} = \mathscr{L}(\phi, \theta; x) + \text{MSE}(x, \hat{x}) \tag{15}$$

where $\mathscr{L}$ is the ELBO loss and the MSE is the MSE loss of the images.

3 Experiments

In the section, we focus on the examination of VAFA-PredNet to understand and disentangle the factors that contribute to the affordances in humanoid robot setting. We apply the VAFA-PredNet to extract the visual and tool information during the embodied interaction.

3.1 Experimental Setting

We use the data-set from the tool-use experiment of the iCub interaction [27]. The data-set was captured as the visual images when the robot executes 4 actions with 3 tools on 11 objects. Particularly, some of the objects have the same shape but different colours, which result in not much different movements in the robot interactions. This is also beneficial to examine whether the model can conceptualise the shape of the objects, which contribute to the understanding of object affordances. Furthermore, each of the configuration contains ten repetitions, so there are totally $11 \times 4 \times 3 \times 10 = 1320$ sets of tool-use interaction. At each interaction, as a visual feedback, images are captured from both cameras as the snapshots of the starting and the ending points. Examples of the starting and ending images that captured from the iCub cameras are shown in Fig. 2.

For training the generative model, we need a temporal sequence of the visual data for the model training. So, rather than using only the starting and ending images, we further interpolate ten images between the starting and ending images given by the data-set.

3.2 Reconstruction of visual inputs

At the first experiments, we will examine the predictive ability of the VAE within the predictive framework of the AFA-PredNet. Similarly, as we did in AFA-PredNet [23] and MT-AFA-PredNet [24, 25], the first results from the generative models to be examined are the generated images. The training was done using Adam optimisation [28] and other hyperparameters were defined as the table follows (Tables 1 and 2).

After training, with the evaluation data-set, we tried different methods to generate the predicted images. Using original inputs as shown in Fig. 1 and the ten interpolated images, Figs. 3 and 4 show some of the generated examples with the comparison of the original ones (in two scenarios), in which we can clearly see that there exist some blurriness in the generated images when using the VAE (compare Figs. 3*d* and 4*d*), which are probably resulting from the Gaussian distribution that brings uncertainties in the reconstruction. The quantitative comparisons about the root-mean-square (RMS) error between the ground truth and the predicted images are also shown in Fig. 5. Since we only use the

Fig. 3 *Generated images in case 1*
(a) Generated image at 0.2 s, ***(b)*** Generated image at 0.4 s, ***(c)*** Generated image at 0.6 s, ***(d)*** Generated image at 0.8s

Table 1 Different Combinations of input data and methods

Case	VAE methods	Prediction methods	inputs	Estimated factors
1	Vanilla VAE	MT-AFA-PredNet	images, action	objects, tools
2	β-VAE	MT-AFA-PredNet	images, action	objects, tools

At the cases 1 and 2, the Vanilla version of VAE [26] and β-VAE are used to connect with MT-AFA-PredNet. We use one-hot representation with four units to indicate different actions.

Table 2 Parameters of the model

Parameters	Value
τ_0	1.0
τ_1	1.1
τ_2	1.3
kernel	3×3 (CNN)
padding	1 (CNN)
pooling	2×2 (CNN)
number of latent units (VAE)	10
number of hidden layer 1 (VAE)	256
number of hidden layer 2 (VAE)	128

Fig. 4 *Generated images in case 2*
(a) Generated image at 0.2 s, ***(b)*** Generated image using case 4, ***(c)*** Window width $w = 1.0$ s, ***(d)*** Window width $w = 1.2$ s

Fig. 5 *Original images*
(a) Original images 1, ***(b)*** Original images 2, ***(c)*** Original images 3, ***(d)*** Original images 4

Table 3 RMS and variance

	RMS	Variance
case 1	8.332	0.5133
case 2	9.834	1.4123

Averaged RMS error of the two cases. The variance is collected with different interactions.

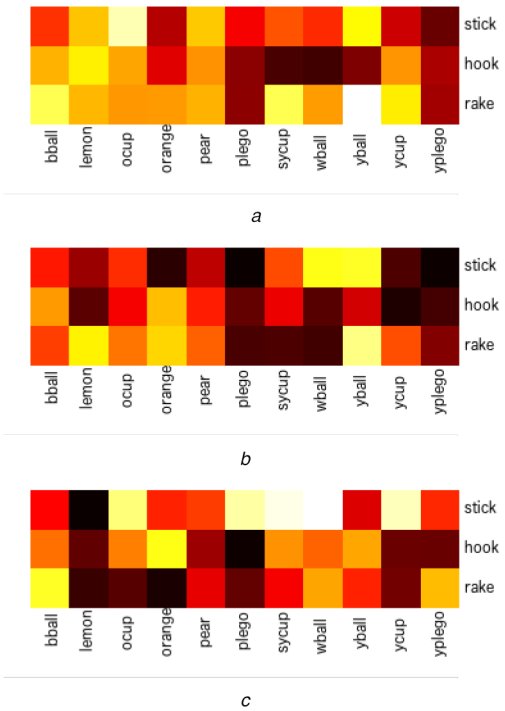


Fig. 6 *Latent units in case 1 – draw*
(a) Latent unit 6, ***(b)*** Latent unit 3, ***(c)*** Latent unit 5

interpolation images for training, the RMSs of the predicted images are not really the ground truth in reality. However, we focus on the shifting of locations of the objects with respect to the motor action and its own affordance, so the interpolation method provides reasonable estimations for the training and comparison. Although the reconstruction results are reasonably good to be recognised, case 2 (Vanilla VAE + PredNet) provides the best results from the quantitative results (Table 3). Nevertheless, the advantage of β-

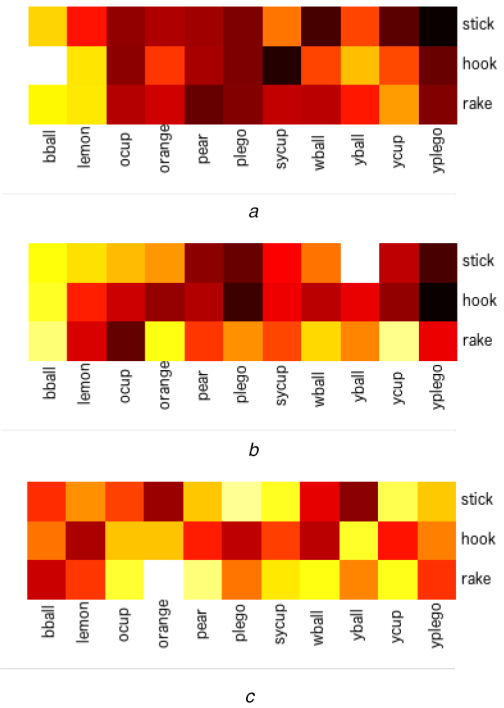


Fig. 7 *Latent units in case 1 – push*
(a) Latent unit 6, ***(b)*** Latent unit 3, ***(c)*** Latent unit 5

VAE is its ability of extracting basic concepts from the input data, which we will be shown in the next section.

3.3 Extracting concepts

Using the same experimental setting, we examine the representation of the VAE units and compare them with different physical settings in the experiments. The aim of the experiment is to find out whether the VAE and the β-VAE are able to disentangle the basic concepts that could contribute to the tool-use learning process. Therefore, we assume some of the variables are the known (e.g. the motor commands and the visual inputs) and some are not (e.g. which concepts of the sensorimotor interaction contribute to visual feedback in this particular setting), as shown in the fourth column of Table 1. In both cases, the motor commands are known as a priori for the robot. Moreover, the agent (e.g. robot) attempts to learn different concepts of the objects and tools.

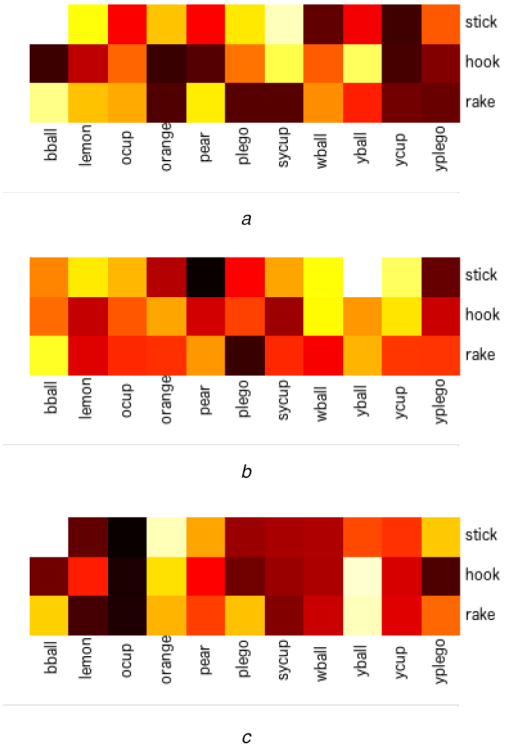


Fig. 8 *Latent units in case 1 – tap from left*
(a) Latent unit 6, ***(b)*** Latent unit 3, ***(c)*** Latent unit 5

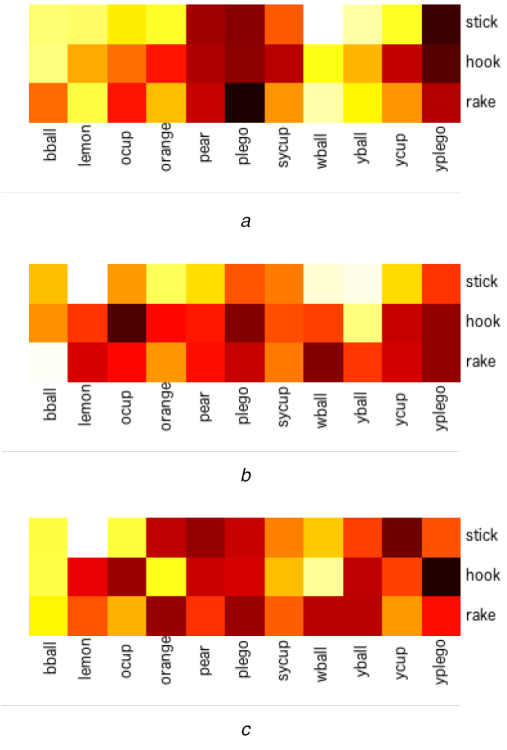


Fig. 9 *Latent units in case 1 – tap from right*
(a) Latent unit 6, ***(b)*** Latent unit 3, ***(c)*** Latent unit 5

Figs. 6–9 show the representations of the particular latent units, whereas the corresponding tool–object–action setting is given with case 1. Since the changes along time are subtle on the top layer in the MT-AFA-PredNet, we have the averaged value for the ten images. Moreover, the latent units with highest variances are shown in these figures. Using the same method, we show the latent units in Figs. 10–13 in case 2.

In general, comparing with these two cases, we can observe that the β-VAE has a better ability to disentangle the information in the objects and tools than the Vanilla version of VAE. From Figs. 10*a*–12*a*, for instance, the values along the same column are similar, suggesting that the concept about different objects is encoded in this latent unit. Also, we can tell the shapes of 'bball', 'wball', and 'yball' are similar (actually the only differences are their colours) which can be learnt by this interaction since they have similar activations of the neurones. Also, representations of 'ball'(s) and 'lemon' have similar values, which probably results from the fact that both of these objects have similar 'shape', so their affordances are similar with these actions. Thus, these three figures (Figs. 10*a*–13*a*) indicate that the tenth unit of the latent space mainly tells the shapes of the objects after the learning.

Similarly, the disentanglement about the concept of tools can also be found in the fifth unit, as we can see that the values along the same row are similar in Figs. 10*b*–13*b*. Specifically, as we can see, the object of 'stick' has larger differences than the 'hook' with the 'rake'. The assumption is that the hook and the rake have much similar affordances with doing interactions with the objects than the stick. We will discuss the relation of the used tools with the affordance learning in Section 4.

From Figs. 10*c*–13*c*, we cannot distinguish pattern but we can see the overall values corresponding to four different actions in the setting of case 2. Actually, the difference of overall values can also be found in all the representations obtained from case 1.

Relatively, while the Vanilla VAE is used in case 1 (Figs. 6–9), no obvious disentanglement can be discovered in the latent space.

3.4 Generalisation of understanding the affordance

In this experiment, based on the previous results about the representation in the latent units, we select two units in the latent space whose variances are larger than 0.4 during the changes of inputs that obtained from the previous Section 3.3. Then, we manually change the activations of corresponding units and observe the changing of the predictive images. We compare the generated images under case 1 and case 2, whereas the values in certain latent units are being changed from [0.1, 1.0], with the increasing steps of every 0.2. The generated images are shown in Figs. 14 and 15 (case 1) and Figs. 16 and 17 (case 2).

As we can see, it seems that case 2, which adopts the β-VAE, has a better generation result than case 1: the objects generated from case 1 often duplicated or are blurry, which may suggest that the latent units are not disentangled but the represented concepts are mixed. On the other hand, the objects that generated from case 2 move regularly, suggesting that the sensorimotor memory resulting from the latent value controls the predictive perception.

4 Discussion

4.1 Learning concepts by affordance learning

The grounding theory in psychology suggests that the usage of natural language relies on situational context. To understand the language dependent on the physical environment and capture such

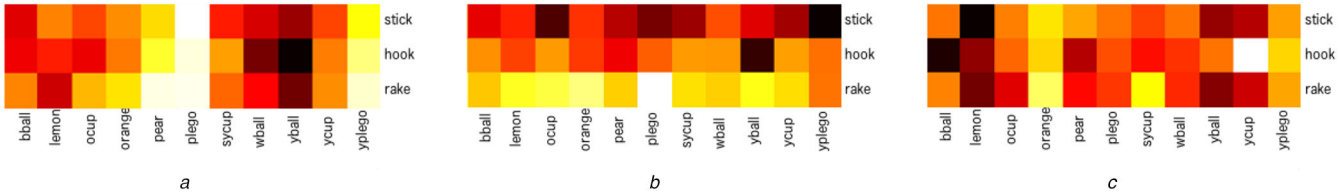


Fig. 10 *Latent units in case 2 – draw*
(a) Latent unit 10, ***(b)*** Latent unit 8, ***(c)*** Latent unit 5

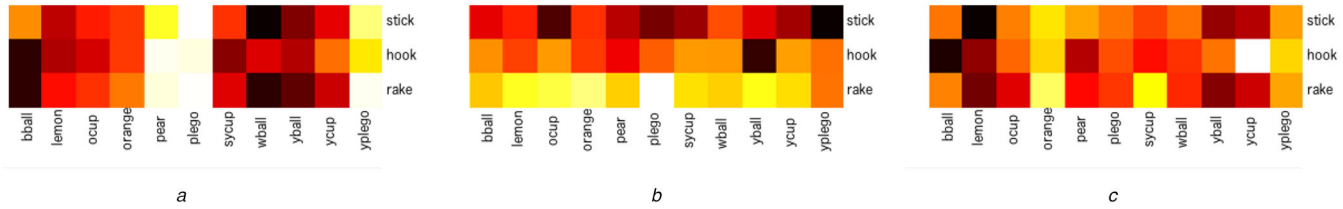


Fig. 11 *Latent units in case 2 – push*
(a) Latent unit 10, ***(b)*** Latent unit 8, ***(c)*** Latent unit 5

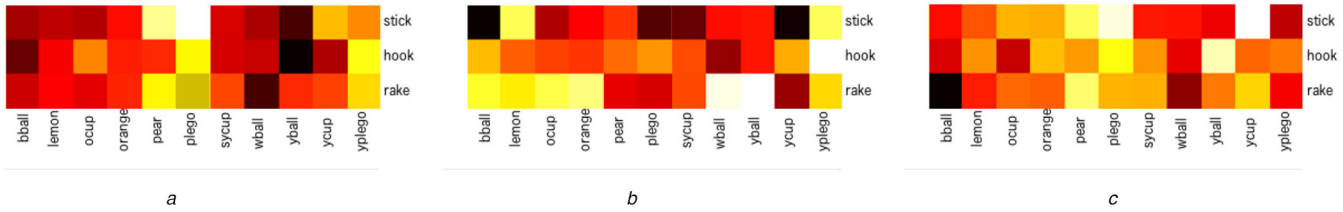


Fig. 12 *Latent units in case 2 – tap from left*
(a) Latent unit 10, ***(b)*** Latent unit 8, ***(c)*** Latent unit 5

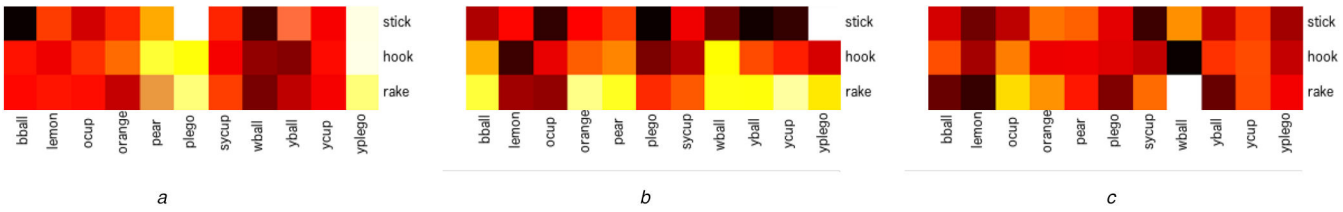


Fig. 13 *Latent units in case 2 – tap from right*
(a) Latent unit 10, ***(b)*** Latent unit 8, ***(c)*** Latent unit 5

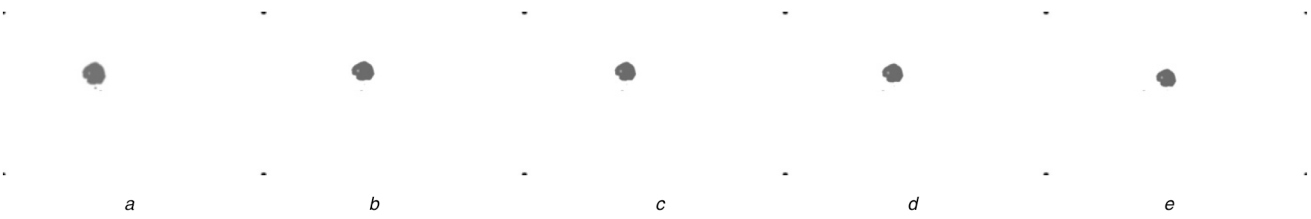


Fig. 14 *Generalisation by changing values in latent unit 6 (case 1)*
(a) Latent unit 0.1, ***(b)*** Latent unit 0.3, ***(c)*** Latent unit 0.24, ***(d)*** Latent unit 0.7, ***(e)*** Latent unit 0.9

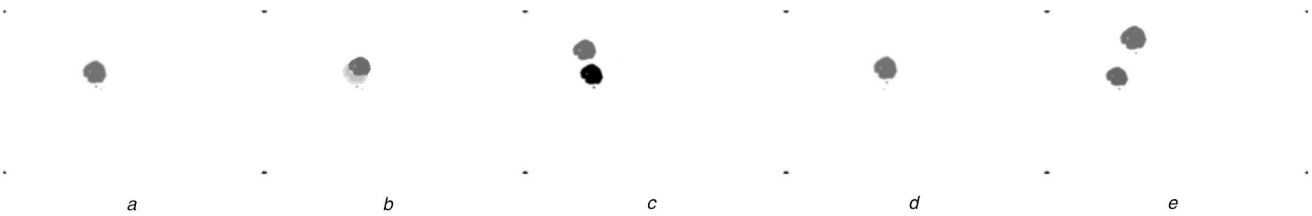


Fig. 15 *Generalisation by changing values in latent unit 3 (case 1)*
(a) Latent unit 0.1, ***(b)*** Latent unit 0.3, ***(c)*** Latent unit 0.24, ***(d)*** Latent unit 0.7, ***(e)*** Latent unit 0.9

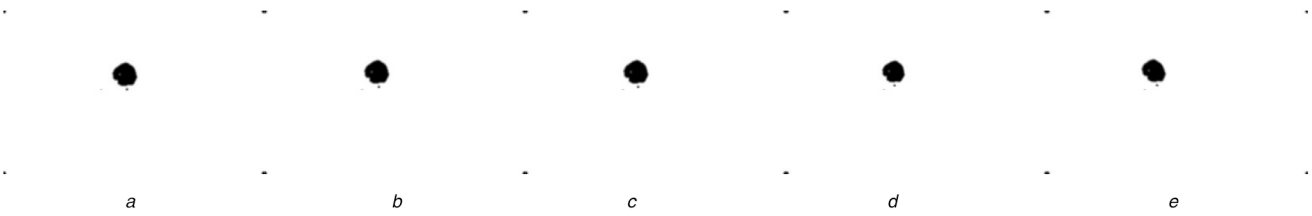


Fig. 16 *Generalisation by changing values in latent unit 10 (case 2)*
(a) Latent unit 0.1, ***(b)*** Latent unit 0.3, ***(c)*** Latent unit 0.24, ***(d)*** Latent unit 0.7, ***(e)*** Latent unit 0.9

possible common abstract values, concepts that ground linguistic meaning are neither internal nor external to language users but instead it spans the objective–subjective boundary. As reviewed in [7], the higher level of concepts is rooted from the low level of grounded meaning such as motor actions, perception and the integration of both.

In our case, the disentanglement of concepts relies on interaction with the environment. So, it contains the factors of motor action, the tool and the objects. Therefore, the

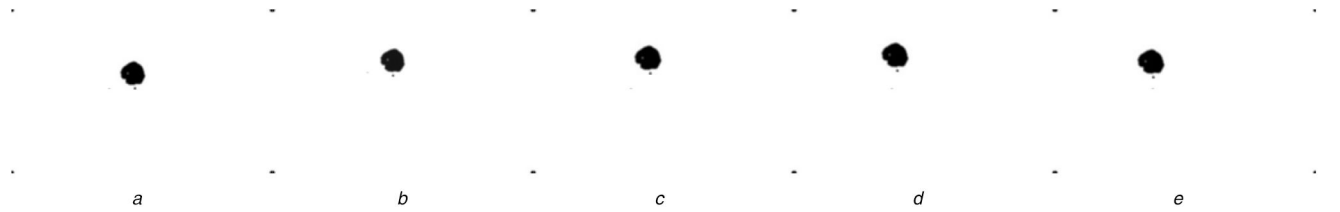


Fig. 17 *Generalisation by changing values in latent unit 8 (case 2)*
(a) Latent unit 0.1, ***(b)*** Latent unit 0.3, ***(c)*** Latent unit 0.24, ***(d)*** Latent unit 0.7, ***(e)*** Latent unit 0.9

disentanglement is the notation of the perceived affordances: how the interaction and its effect can be abstracted as the structured units, which can be further used for prediction at the future interaction. Specifically, in our case, the concepts such as 'shapes' and 'colours' of 'tools' and 'objects' are the common properties that contribute to the affordance learning during the tool-use experiment. Such abstract concepts are first rooted from the low level of grounded meaning of both motor action and the visual perception. It is similar as the abstract grounding problem but not totally the same: the abstract examples rooted directly from the sensorimotor interaction (such as 'give' and 'accept' or numerical concepts shown in [7]) but the conceptualisation depends heavily on the abstraction of the common physical properties of the sensory inputs. Besides that, it needs additional 'meta-learning' procedure which mathematically attempts to align the axis of the 'concept' of the visual information while the affordance is being learnt. Importantly, such meta-learning procedure can be bridged by affordance learning. Since it was first introduced to understand better one of the basic properties of the objects independent to the specific actions that depend on individuals, it links the learning agent and its subjective understanding of the world through its motor action and sensing to the world.

In the next step, our work can be extended in extracting the pre-symbolic representation about 'concepts' of the objects while the robot is situated in a robot–object interactive scenario. In the robotics community, though there are several existing works on obtaining and using object–action relation, not so much work has successfully incorporated different aspects of 'concepts' from the visual inputs and motor outputs in the process of affordance learning. On the basis of this interaction setting, we argue that while humans are doing the conceptualisation of the world, they are not only doing it based on the geometry appearance features of the objects but also its integration to the voluntary motor actions. On the other hand, though we have witnessed the state-of-the-art deep learning method perform well in the object categorisation, we need more earning methods to understand the concepts by using their tool functions and observing the visual effects, and connecting the objective world and the subjective prediction. We believe similar learning models (deep learning, together with probabilistic learning) could be useful in the future development in terms of its function in linking the concepts in the low-dimensional attention space with the peripheral signals in the high-dimensional representation space.

4.2 From conceptualisation to language development

The approaches to language acquisition can be coarsely divided into nativism or empiricism. Presently, most of the evidence (e.g. [29, 30]) believe that the environmental factors contribute to the majority of the process of children acquiring their language skills from an early age learning of language is embedded in the behaviour within the child's social context. This mechanism can be observed, for instance, when children gradually acquire the rules of grammar and the complexities of word comprehension that eventually lead to the production of words and sentences. This is contradicted with the views from the nativists such as Chomsky and Pinker, who consider the lexical rules are passed down through the child's environment [31]. The basic mechanism of learning a language such as the grammar and syntax is already rooted at birth and development of the biological organism which contains that language instinct.

In our experiment, the concepts are disentangled with the tool-use data in an unsupervised manner. Specifically, the recent development of β-VAE seems to provide an example about disentangling different concepts without any a priori knowledge or supervised learning. As the analysis is shown in [32], the additional constraint parameter provides an additional representation capacity of the latent units Z. This is done by adjusting the relative weighting of the KL divergence of independent concepts in the latent space Z while constructing the estimated posterior $q(z|x)$ to the true a priori distribution for $p(z|x)$. In the case of β-VAE, since this distribution components of z are independent (e.g. disentangled) concepts with independent physical properties, we can emphasise different concepts. Nevertheless, from the comparison between the VAE and the β-VAE methods, we can conclude that such disentanglement is sensitive to relative weighting (i.e. the parameter of β) of the effect features that are encoded in different units and channels.

The results of this model also suggest that the importance of the role of intentionality for embodied learning. We have seen that the disentanglement of concepts emerges from the agent's sensorimotor exploration. It guides the agent to select the regions that are in the intermediate level of difficulty. Furthermore, during the process of finding disentangled categories of concepts, we observed that different concepts sometimes still have the same values using the β-VAE. It is related to the facts that those concepts have similar values in the effect space (e.g. the movements of the ball and the lemon are similar, given the same motor action and the tool). It also indicates that the naive clustering using simple motor action is not enough while the agent is learning the affordance. In that case, applying further exploratory actions is necessary for the interaction of the objects handling this complexity. For example, in the case of tool-use experiment, after the action of pushing, the robot can poke both of the objects one by one, so that the effects of categories can be distinguished when the observations obtained from the action of poking can be taken into account, without any extrinsic reward. This can be driven by the intrinsic motivation mechanism [33, 34].

4.3 Consciousness prior

The idea of building an abstract representation with a low-dimensional representation of the features for the tasks with a high-dimensional representation is an essential bridge between the state-of-the-art deep learning methods and the higher cognitive abilities and the artificial general intelligence (AGI). To accomplish this task, a new prior [35] should be used in the conscious state, which allows a low-dimensional calculation such as unconsciousness attention to change different cognitive statuses. Such a priori is an abstract of the high-dimensional observed representation such as specific kinds of motor action output or sensory inputs. The cognitive function of such a priori is the clustering of them as a concept (or a 'vector of thoughts'), so that the higher level of cognitive thought can be computed by the attention mechanism on the unconscious level. Such a vector can be physically conceptualised by the different dimensions of multi-modalities inputs, a control signal for motor actions or an inference result from the higher level of manipulation. Moreover, another advantage of such a priori is that it can also connect with the language and symbolic representation.

In terms of machine learning, one of the key advantages of using the latent unit, which has a small-dimensional but rich representation, is to allow for improved generalisation results with novel data and unexpected noise. Our proposed model covers the

topic about how the learning of association between the representation space and the attention space occurs. Specifically, such a consciousness a priori lies in the concepts of the sensorimotor space. On the basis of the PredNet, we propose the learning happens in the framework of predictive coding framework, in which the movement of the objects can be predicted by both the voluntary motor actions as well as the object features. Moreover, the internal model of such prediction is abstracted and disentangled in the VAE model. To our best knowledge, our approach is the first model to apply a hierarchical learning network to learn the affordance factors from the observations for the motor signals. The abstraction of the generative model is further learnt in a slower context with the VAE. The learned representation can, in some contexts, greatly help for generalisation as it provides a more succinct representation that is less prone to be overfitting. Furthermore, the β-VAE learns the important features (e.g. the concept in the visual appearance in our case) in the abstract representation because they are important to distinguish the observed values, while they are otherwise irrelevant for the task at hand (e.g. the colour of the objects).

In the future, an extended approach can be designed to build relevant motor action with a common set of visual features. Such a common representation can be useful for solving a set of goal-directed tasks with visual stimuli. We also believe that the low-level features can emerge in the hierarchical architecture, in which the inductive bias can be introduced in the related tasks. In neuroscience, the idea of an abstract representation can be found as well where the phenomenon of access consciousness can be seen as the formation of a low-dimensional combination of a few concepts which condition planning, communication and the interpretation of upcoming observations. In machine learning, for instance, a common reinforcement learning (RL) framework can be utilised to realise such a goal. Therefore, based on this, the abstract state could be formed using an attention mechanism able to select specific relevant variables in a context-dependent manner.

5 Conclusion

We presented a new learning model for the disentanglement of the sensorimotor concepts. This model, based on deep learning as well as the generative model, provides a promising way of how the independent concepts can be extracted and disentangled from both perception and action. Specifically, the hierarchical part follows the predictive framework of common coding. It consists different time scales of prediction representation, in which the slower scale is the related categorisation of the sensorimotor events and is learnt by the β variational encoder (β-VAE) model. In this β-VAE model, with the constraint given in the ELBO loss function, the latent space separates the disentanglement of different concepts of both perception and action, while the agent is learning with the affordance data-set. The experiments show that the emerged disentanglement representation also owns the generalisation ability while the model is doing the generative process.

6 Acknowledgments

The research was supported by the Japan New Energy and Industrial Technology Development Organisation (NEDO).

7 References

[1] Presson, C.C., Hazelrigg, M.D.: 'Building spatial representations through primary and secondary learning', *J. Exp. Psychol., Learn. Mem. Cogn.*, 1984, **10**, (4), p. 716

[2] Kroll, J.F., Merves, J.S.: 'Lexical access for concrete and abstract words', *J. Exp. Psychol., Learn. Mem. Cogn.*, 1986, **12**, (1), p. 92

[3] Schwanenflugel, P.J., Harnishfeger, K.K., Stowe, R.W.: 'Context availability and lexical decisions for abstract and concrete words', *J. Mem. Lang.*, 1988, **27**, (5), pp. 499–520

[4] Rucinski, M., Cangelosi, A., Belpaeme, T.: 'An embodied developmental robotic model of interactions between numbers and space'. Proc. Annual Meeting of the Cognitive Science Society, Boston, Massachusetts, USA, 2011, vol. **33**, no. 33

[5] Di Nuovo, A., Vivian, M., Cangelosi, A.*, et al.*: 'The iCub learns numbers: an embodied cognition study'. 2014 Int. Joint Conf. Neural Networks (IJCNN), Beijing, China, 2014, pp. 692–699

[6] LeCun, Y., Bengio, Y., Hinton, G.: 'Deep learning', *Nature*, 2015, **521**, (7553), p. 436

[7] Cangelosi, A., Stramandinoli, F.: 'A review of abstract concept learning in embodied agents and robots', *Philos. Trans. R. Soc. B*, 2018, **373**, (1752), p. 20170131

[8] Dotov, D.G., Nie, L., De Wit, M.M.: 'Understanding affordances: history and contemporary development of Gibson's central concept', *Avant: J. Philos.-Interdiscip. Vanguard*, 2012, pp. 30–32

[9] Ugur, E., Nagai, Y., Sahin, E.*, et al.*: 'Staged development of robot skills: behavior formation, affordance learning and imitation with motionese', *IEEE Trans. Auton. Ment. Dev.*, 2015, **7**, (2), pp. 119–139

[10] Salvi, G., Montesano, L., Bernardino, A.*, et al.*: 'Language bootstrapping: learning word meanings from perception–action association', *IEEE Trans. Syst. Man Cybern. B (Cybern.)*, 2012, **42**, (3), pp. 660–671

[11] Yürüten, O., Şahin, E., Kalkan, S.: 'The learning of adjectives and nouns from affordance and appearance features', *Adapt. Behav.*, 2013, **21**, (6), pp. 437–451

[12] Kalkan, S., Dag, N., Yürüten, O.*, et al.*: 'Verb concepts from affordances', *Interact. Stud.*, 2014, **15**, (1), pp. 1–37

[13] Johnson, S.P.: 'How infants learn about the visual world', *Cogn. Sci.*, 2010, **34**, (7), pp. 1158–1184

[14] Neal, R.M.: 'Connectionist learning of belief networks', *Artif. Intell.*, 1992, **56**, (1), pp. 71–113

[15] Dayan, P., Hinton, G., Neal, R.*, et al.*: 'The Helmholtz machine', *Neural Comput.*, 1995, **7**, (5), pp. 889–904

[16] Hinton, G.E., Salakhutdinov, R.R.: 'Reducing the dimensionality of data with neural networks', *Science*, 2006, **313**, (5786), pp. 504–507

[17] Mescheder, L., Nowozin, S., Geiger, A.: 'Adversarial variational Bayes: unifying variational autoencoders and generative adversarial networks'. Proc. 34th Int. Conf. Machine Learning, Sydney, Australia, 2017, Vol. **70**, pp. 2391–2400

[18] Radford, A., Metz, L., Chintala, S.: 'Unsupervised representation learning with deep convolutional generative adversarial networks', arXiv preprint arXiv:1511.06434, 2015

[19] Pu, Y., Gan, Z., Henao, R.*, et al.*: 'Variational autoencoder for deep learning of images, labels and captions'. Advances in Neural Information Processing Systems, Barcelona, Spain, 2016, pp. 2352–2360

[20] Walker, J., Doersch, C., Gupta, A.*, et al.*: 'An uncertain future: forecasting from static images using variational autoencoders'. European Conf. Computer Vision, Amsterdam, the Netherlands, 2016, pp. 835–851

[21] Reed, S., Akata, Z., Yan, X.*, et al.*: 'Generative adversarial text to image synthesis', arXiv preprint arXiv:1605.05396, 2016

[22] Prinz, W.: 'Perception and action planning', *Eur. J. Cogn. Psychol.*, 1997, **9**, (2), pp. 129–154

[23] Zhong, J., Cangelosi, A., Zhang, X.*, et al.*: 'AfapredNet: the action modulation within predictive coding'. Int. Joint Conf. Neural Networks (IJCNN), Rio de Janeiro, Brasil, 2018

[24] Zhong, J., Cangelosi, A., Ogata, T.*, et al.*: 'Encoding longer-term contextual information with predictive coding and ego-motion', *Complexity 2018*, (2018)

[25] Zhong, J., Ogata, T., Cangelosi, A.: 'Encoding longer-term contextual sensorimotor information in a predictive coding model'. 2018 IEEE Symp. Series on Computational Intelligence, Bengaluru, India, 2018

[26] Kingma, D.P., Welling, M.: 'Auto-encoding variational Bayes', arXiv preprint arXiv:1312.6114, 2013

[27] Tikhanoff, V., Pattacini, U., Natale, L.*, et al.*: 'Exploring affordances and tool use on the iCub'. 2013 13th IEEE-RAS Int. Conf. Humanoid Robots (Humanoids), Atlanta, Georgia, USA, 2013, pp. 130–137

[28] Kingma, D.P., Ba, J.: 'Adam: a method for stochastic optimization', arXiv preprint arXiv:1412.6980, 2014

[29] Kuhl, P.K.: 'Brain mechanisms in early language acquisition', *Neuron*, 2010, **67**, (5), pp. 713–727

[30] Nomikou, I., Rohlfing, K.J.: 'Language does something: body action and language in maternal input to three-month-olds', *IEEE Trans. Auton. Ment. Dev.*, 2011, **3**, (2), pp. 113–128

[31] Pinker, S.: '*The language instinct: the new science of language and mind*', vol. **7529**, (Penguin, UK, 1995)

[32] Burgess, C.P., Higgins, I., Pal, A.*, et al.*: 'Understanding disentangling in β-VAE', arXiv preprint arXiv:1804.03599, 2018

[33] Deci, E.L., Ryan, R.M.: 'Intrinsic motivation', in (Eds.): '*The corsini encyclopedia of psychology*' (IEEE, US, 2010), pp. 1–2

[34] Oudeyer, P.-Y., Kaplan, F., Hafner, V.V.: 'Intrinsic motivation systems for autonomous mental development', *IEEE Trans. Evol. Comput.*, 2007, **11**, (2), pp. 265–286

[35] Bengio, Y.: 'The consciousness prior', arXiv preprint arXiv:1709.08568, 2017