Artificial Cognitive Map System based on Generative Deep Neural Networks

Hiroki Kojima¹ and Takashi Ikehama¹

¹University of Tokyo
{kojima, ikeg}@sacral.c.u-tokyo.ac.jp

Abstract

We present a novel artificial cognitive map system using the generative deep neural networks called Variational Autoencoder / Generative Adversarial Network (VAE/GAN), which encodes input images into the latent space and the structure of the latent space is self-organized through the learning. Our results show that the distance of the predicted image is reflected in the distance of the corresponding latent vector after training, which indicates that the latent space is organized to reflect the proximity structure of the dataset. This system is also able to internally generate temporal sequences analogous to hippocampal replay/pre-play, and we found that these sequences are not just the exact replay of the past experience, and this could be the origin of creating novel sequences from the past experiences. Having this generative nature of cognition is thought as a prerequisite for artificial cognitive systems.

Introduction

In ALIFE research, the question arises as to what level of internal complexity should be provided to simulate the cognitive behaviour of artificial agents. Inman Harvey, for example, argues that we tend to confuse the part with the whole, and to carry the resulting complex cognitive symbols into the part as well (which he calls mereological fallacy) (Harvey, 2020). On the other hand, in thinking about language and consciousness, We are tempted to consider more complex internal structures, such as prediction, priors, and internal representations (e.g. Friston’s Free energy principle(Friston, 2010)). At the ALIFE conference, we would like to rethink what is the essential ingredient for simulating cognitive systems.

Studying a cognitive map (Tolman, 1948) using an artificial cognitive map system can be the first step. Constructing an artificial cognitive map system were proposed by Rössler (1981), and, for example, Nolfi and Tani (1999) trained a robot with hierarchical recurrent neural networks to predict the next sensory state of the robot, and found that each layer of the neural networks encodes some regularities of the environment. Noguchi et al. (2017) trained hierarchical recurrent neural networks with recurrent gated units using visual and motor information and found that the map of the environment was self-organized at the higher layer.

A cognitive map has been attributed to place cells and grid cells in hippocampus (O’Keefe and Dostrovsky, 1971; Hafting et al., 2005). Recently, the cognitive map system has been proposed to deal with not only the spatial information but other information such as sounds, shapes, or social relationships (Behrens et al., 2018), which implied the existence of the more general principles for the organization of cognitive maps based on the proximity of the input features. The other characteristic of the place cells is that they are capable of internally generate temporal sequences, known as replay/preplay (O’keefe and Nadel, 1978; Buzsáki et al., 1983; Dragoi and Tonegawa, 2011).

Here, we present an artificial cognitive map system that is self-organized by training with only visual data and no explicit metric information (Kojima and Ikegami, 2021). Our system is based on the generative deep neural networks, which gradually learn to encode images into the latent space with relatively low dimensions. We trained the network on first-person perspective images navigating in a simple virtual environment and investigate the structure of latent space.

Methods

Our system (Kojima and Ikegami, 2021) is based on Variational Autoencoder / Generative Adversarial Network (VAE/GAN) (Larsen et al., 2016). VAE/GAN includes three deep convolutional neural networks; the Generator (Gen), the Discriminator (Dis), and the Encoder (Enc). We modified the original VAE/GAN to predict the image of $\tau$ steps ahead, $x(t + \tau)$ (Fig.1). For comparison, we also used Variational AutoEncoder (VAE)(Kingma and Welling, 2013), which includes Gen and Enc only.

As a training dataset, we used a series of first-person visual inputs of an agent moving in a virtual environment. The environment is a corridor in the shape of a figure of eight, intersecting at one point. The top, bottom, left and right sides of the corridor are painted in different colours, depending on the location. At the intersection, the agent randomly chooses which direction to go.
Figure 1: Overview of our network based on VAE/GAN (Larsen et al. 2016).

Figure 2: Example of the generated images from the latent vectors from the 1st and 2nd principal components. Left: VAE ($\tau = 30$), Right: VAE/GAN ($\tau = 30$)

Result

After the training, we confirmed that the system is capable of generating predicted images $x(t+\tau)$ from $x(t)$. We can also generate images directly from the latent space. (Fig.2) The generated images from VAE tended to have low variation and to be blurred, but with GAN, the images seemed to be more varied, less blurred, and smoothly interpolated in the latent space.

Characterization of the Latent Space Vectors

We analyzed the latent vector $z(t) = Enc(x(t))$ and investigated how the input images were mapped into the latent space. We calculated the distance matrix of the input images, the target images, and the corresponding latent vectors, respectively, and calculated the correlation coefficients between the values of these distance matrices. We found the correlation existed, and the correlation was stronger between the target images and the latent vectors than between the input images and the latent vectors.

Sequence generation by Closed Loop

After the training, the system can generate an infinite number of images using the following procedure. First, the initial image $x_0$ is converted into a latent space vector by the encoder, $z_0 = Enc(x_0)$, and an image is generated from the latent space vector by the generator $x_1 = Gen(z_0)$. This image is now used as the input image. By repeating this process recursively, the network continues to predict the next image without receiving the actual image. (Fig.3)

The trajectories from the VAE showed mainly limit cycle dynamics, while the VAE/GAN showed an increased fraction of chaotic trajectories. Also, we measured the timestep between the images in the generated sequences and found that despite the median time step did not change from the original time step, there were variations in the reconstructed time step, and the variation was large in VAE/GAN compared to VAE.

Discussion

We constructed an artificial cognitive map system by VAE/GAN, and showed that the latent space was self-organized to reflect the proximity structure of the image dataset. Also, we showed that the system was able to internally generate temporal sequences, like replay/preplay in hippocampus. These sequences are not exactly the same as the past experience, but the reconstructed timesteps can be varied, and sometimes can be chaotic. In hippocampal replay, it has been found that novel sequences are also generated (Gupta et al., 2010; Stella et al., 2019), and our finding provides the possible origin of the novelty production in hippocampus. As F.Varela says in his book Embodied Mind (Varela et al., 2017), cognition as embodied action is always about or directed toward something that is missing; on the other hand, there is always the next step for the system in its perceptually guided action; and on the other hand, the actions of the systems are always directed toward situations that have yet to become actual. Therefore, we speculate that this generative nature, especially the ability of novelty production is an important internal mechanism for obtaining complex cognitive behaviors.

The dual-process theory (Yonelinas, 1994) claims that the retrieval of episodic memory consists of two processes, familiarity, and recollection, and these seem to be related to the output of the discriminator, which-rated whether the input image was included in the training dataset, and the sampling from the latent space in our system, respectively. Therefore, our system might provide the integrated insight for hippocampus, which is attributed to the seemingly different two functions, the cognitive maps, and the episodic memory.

Acknowledgements

This work was partially supported by JSPS KAKENHI Grant, Correspondence and Fusion of AI and Brain (17H06024, 19H049790) and Chronogenesis : How the mind generates time (19H05306).
References

Behrens, T. E., Muller, T. H., Whittington, J. C., Mark, S., Baram, A. B., Stachenfeld, K. L., and Kühn, Z. (2018). What is a cognitive map? organizing knowledge for flexible behavior. Neuron, 100(2):490–509.

Buzsáki, G., Vanderwolf, C. H., et al. (1983). Cellular bases of hippocampal EEG in the behaving rat. Brain Research Reviews, 6(2):139–171.

Dragoi, G. and Tonegawa, S. (2011). Preplay of future place cell sequences by hippocampal cellular assemblies. Nature, 469(7330):397–401.

Friston, K. (2010). The free-energy principle: a unified brain theory? Nature reviews neuroscience, 11(2):127–138.

Gupta, A. S., van der Meer, M. A., Touretzky, D. S., and Redish, A. D. (2010). Hippocampal replay is not a simple function of experience. Neuron, 65(5):695–705.

Hafting, T., Fyhn, M., Molden, S., Moser, M.-B., and Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. Nature, 436(7052):801–806.

Harvey, I. (2020). What matters? agent concerns, agent-designer concerns. In Artificial Life Conference Proceedings, pages 300–302. MIT Press.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

Kojima, H. and Ikegami, T. (2021). Organization of a latent space structure in vae/gan trained by navigation data. arXiv preprint arXiv:2102.01852.

Larsen, A. B. L., Sønderby, S. K., Larochelle, H., and Winther, O. (2016). Autoencoding beyond pixels using a learned similarity metric. In International conference on machine learning, pages 1558–1566. PMLR.

Noguchi, W., Iizuka, H., and Yamamoto, M. (2017). Cognitive map self-organization from subjective visuomotor experiences in a hierarchical recurrent neural network. Adaptive Behavior, 25(3):129–146.

Nolfi, S. and Tani, J. (1999). Extracting regularities in space and time through a cascade of prediction networks: The case of a mobile robot navigating in a structured environment. Connection Science, 11(2):125–148.

O’Keefe, J. and Dostrovsky, J. (1971). The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely-moving rat. Brain research.

O’keefe, J. and Nadel, L. (1978). The hippocampus as a cognitive map. Oxford: Clarendon Press.

Rössler, O. E. (1981). An artificial cognitive map system. BioSystems, 13(3):203–209.

Stella, F., Baracskay, P., O’Neill, J., and Csicsvari, J. (2019). Hippocampal reactivation of random trajectories resembling brownian diffusion. Neuron, 102(2):450–461.

Tolman, E. C. (1948). Cognitive maps in rats and men. Psychological review, 55(4):189.

Varela, F. J., Thompson, E., and Rosch, E. (2017). The Embodied Mind, revised edition: Cognitive Science and Human Experience. MIT press.

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: evidence for a dual-process model. Journal of Experimental Psychology: Learning, Memory, and Cognition, 20(6):1341.