Task representation in neural networks during continuous learning

Anonymous Author(s)

ABSTRACT

The brain has the ability to perform many complex tasks and the flexibility to learn new tasks, but the underlying mechanism still remains unknown and it is hard to elucidated in traditional experimental ways. Here we trained a simple recurrent network on about 20 tasks that require decision making, working memory and so on. The network has three areas linked with each other through both feed-forward and feed-back connections. (statement below are my expectations) We found that neural activity and the distribution of fixed points can represent and explain the output of the network, and those features are distinct for different areas and tasks, indicating that the network emerges compositional structure like brain. We then introduce a method to find a shared subspace among several neural activities and connection weights, by which we found that neural activities of similar tasks share common lower dimensional features, and network can utilize these features to perform new tasks while continuous learning. Finally, we found that the shared lower dimensional features of connection weights change while continuous learning, which suggests a tentative explanation of learning and forgetting.

1 INTRODUCTION

Human brain is able to perform many complex cognitive tasks in a short time. A cognitive task is composed of elementary sensory, cognitive, and motor processes [2, 5]. At the computational level, correctly performing a new task without training requires composing elementary processes that are already learned. This property, called 'compositionality', has been proposed as a fundamental principle underlying flexible cognitive control[1, 2]. Although human studies have suggested that the representation of complex cognitive tasks in the lateral prefrontal cortex could be compositional[1, 2, 4], it is unknown whether compositional task representation structure can emerge in a hierarchical neural network model, and how it changes during continuous learning.

Previous works have trained a vanilla recurrent network on several cognitive tasks and found compositionality on neuronal level[2, 3]. Although those works demonstrated that some tasks can be performed by recombining instructions for other tasks due to neural selectivity, but they failed on other tasks, which cannot be explained by the selectivity of single neuron. Besides, they simply use principal component analysis (PCA) on each task separately to compare neural activity between a pair of tasks qualitatively, which failed to point out quantitative relationship and shared features of these similar tasks.

Here, instead of vanilla recurrent network, we trained a hierarchical recurrent network with three areas and both feed-forward and feed-back connections, which is more realistic and complex so that it can provide information on the level of brain areas. Utilizing the continual-learning technique, we trained the network on 20 cognitive tasks continuously to get a pre-trained model that is able to complete these tasks with a high performance. We also introduce a novel method to find the shared low dimensional subspace among several neural activities of different tasks. By implement analysis on those shared features, we suggest a tentative explanation to the mechanism of continuous learning and forgetting both in human brain and artificial neural networks.

This work provides a computational platform to build a hierarchical recurrent network that is able to investigate neural representations of many cognitive tasks. Also the method we use to extract common features is general for many other analysis. Finally, what we found about continuous learning and forgetting could inspire new algorithms on meta learning or mitigate forgetting of machine learning.

2 METHOD

Define cognitive tasks. First we define 20 cognitive tasks using python package called "neurogym". Some simple tasks like Perceptual Decision Making task have already been implemented in neurogym, however, other tasks needed to be defined in accordance to previous works[2]. For each task, the network receives a noisy stimulus and a fixation input throughout the whole trial, as shown in figure1. What is different from previous work is that the network input does not contain rule input, which is a one-hot encoded vector of task type.

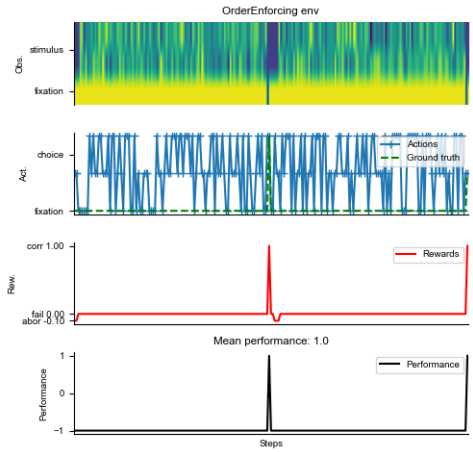

Figure 1: Visualization of two trial of the Perceptual Decision Making task. First row: above, stimulus input; below, fixation input, only deactivated in decision period. Second row: randomly sampled choices of actions and task ground truth. Third row: reward given to the network according to its choice and task ground truth. Forth row: performance.

Typically, a task has four periods, fixation, stimulus, delay and decision. There is no stimulus over fixation period, and the network is expected to do nothing during this period. During stimulus

period, the network receives stimulus with Gaussian noise. The delay period is similar to fixation period but is placed between stimulus and decision period, which requires working memory of the network. Finally, the network should give out a decision in action space. The lasting time for each period could be configured to meet the different need of tasks. For example, a simple perceptual decision making task consists of 100ms fixation, 2000ms stimulus, no delay and 100ms decision period.

Network structure. The hierarchical neural network we trained has three areas. Typically, the activity $\mathbf{r}$ of a area contains both feed-forward connections and feed-back connections follows a continuous dynamical equation

$$\tau \frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t} = -\mathbf{r} + f(W^{rec}\mathbf{r} + W^{in}\mathbf{u} + W^{back}\mathbf{x} + \mathbf{b})$$

In this equation, τ is the neuronal time constant, $\mathbf{u}$ is the input to the network or feed-forward neural activity from predecessor area, $\mathbf{x}$ if the feed-back neural activity from successor area (not included in the last area), $\mathbf{b}$ is the bias or background input, $f(\cdot)$ is the neuronal nonlinearity.

After using the first-order Euler approximation with a time-discretization step Δt, we have

$$\mathbf{r}_t = (1-\alpha)\mathbf{r}_{t-1} + \alpha f(W^{rec}\mathbf{r} + W^{in}\mathbf{u} + W^{back}\mathbf{x} + \mathbf{b})$$

Here $\alpha \equiv \Delta t/\tau$. We use a ReLU activation function

$$f(x) = \max(0, x)$$

The neural activity of the last area is projected into a set of output units $\mathbf{z}$ using

$$\mathbf{z} = W^{out}\mathbf{r}$$

3 RESULTS

To start from a simple step, we first built a two-area recurrent neural network with only feed-forward connections. The hyper-parameters are as follow: $\tau_1 = 40$, $\tau_2 = 60$, first area has 128 neurons, second area has 64 neurons, and $\Delta t = 20$. We chose perceptual decision making task as target, length of sequence input is 100, batch size is 16. We then train this simple network for 5000 iterations with initial learning rate 0.01 using stochastic gradient decent (SGD) and a scheduler that reduces learning rate every 1000 iterations. Finally the network reach an accuracy of over 0.93.

After implementing principal component analysis on neural activities of both areas, the track of the first and the second area is plotted in figure2 and figure3 separately. In consistent with the previous work[3], the tracks of two choices is separated into two directions. But the trajectory of two areas are similar, which may indicates that they have same function. We then find fixed point of the two areas using a approximately method. We found that the fixed points of the first area form a line attractor, but there is no certain distribution of the fixed points of the second area, which may indicates that the neural dynamics of the two areas are different.

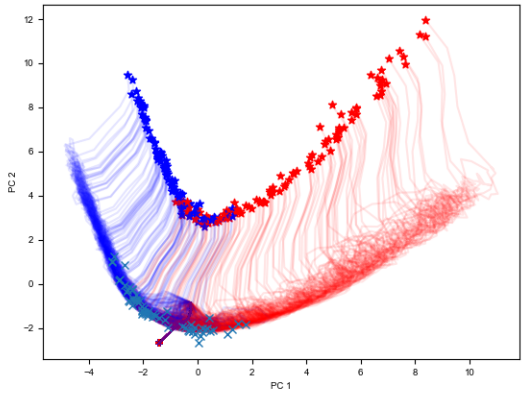

Figure 2: Neural activity (blue lines and red lines) and fixed points of the first area projected to top 2 principal components. Color indicates different choice of the network. Start point of each trial is marked as '+', end point is marked as star, and fixed point is marked as '×'.

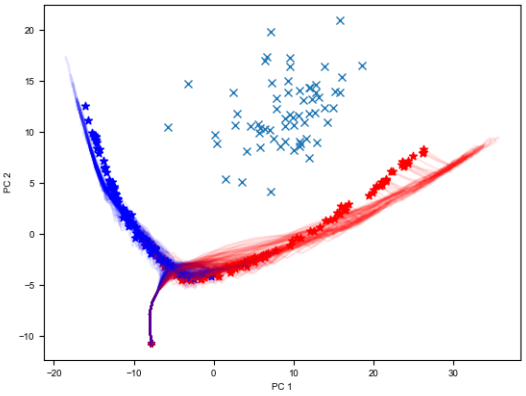

Figure 3: Neural activity (blue lines and red lines) and fixed points of the second area projected to top 2 principal components.

4 DISCUSSION

In this work, one of our hypothesis is that the neural dynamics of each area should be different in a hierarchical recurrent network that is continuously trained on several cognitive tasks. Our preliminary work supports this hypothesis with different distributions of fixed points, but rejects it with similar trajectory of neural activity. Next step will be expand this network to three areas and add feed-back connections, which may lead to a more significant result. Our work is a first attempt to hierarchical recurrent network model for cognitive tasks and may inspire further research.

REFERENCES

[1] Laurent P. Stocco-A. Cole, M. W. 2013. Rapid instructed task learning: a new window into the human brain's unique capacity for flexible cognitive control. Cogn. *Affect. Behav. Neurosci.* (2013).

[2] H. Francis Song William T. Newsome Guangyu Robert Yang, Madhura R. Joglekar and Xiao-Jing Wang. 2019. Task representations in neural networks trained to

perform many cognitive tasks. *Nature Neuroscience* (2019).
[3] David Sussillo Laura Driscoll, Krishna Shenoy. 2022. Flexible multitask computation in recurrent networks utilizes shared dynamical motifs. *bioRxiv* (2022).
[4] Görgen K. Haynes J.-D. Reverberi, C. 2012. Compositionality of rule representations in human prefrontal cortex. *Cereb. Cortex* (2012).
[5] K. Sakai. 2008. Task set and prefrontal cortex. *Annu. Rev. Neurosci.* (2008).