nature machine intelligence

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s42256-019-0080-x

In the format provided by the authors and unedited.

Continual learning of context-dependent processing in neural networks

Guanxiong Zeng[1,2,4], Yang Chen[1,4], Bo Cui[1,2] and Shan Yu[1,2,3]*

[1]Brainnetome Center and National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. [2]University of Chinese Academy of Sciences, Beijing, China. [3]Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Beijing, China. [4]These authors contributed equally: Guanxiong Zeng, Yang Chen. *e-mail: shan.yu@nlpr.ia.ac.cn

Supplementary Information of “Continual Learning of Context-Dependent Processing in Neural Networks”

Guanxiong Zeng [1,2,*], **Yang Chen** [1,*], **Bo Cui** [1,2] **and Shan Yu** [1,2,3,†]

[1] *Brainnetome Center and National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, 100190 Beijing, China.*
[2] *University of Chinese Academy of Sciences, 100049 Beijing, China.*
[3] *CAS Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, 100190 Beijing, China.*

* These authors contributed equally to this work.

† Correspondence; shan.yu@nlpr.ia.ac.cn

SUPPLEMENTARY DISCUSSION

Comparison between OWM and CAB

OWM and CAB were separately proposed to serve the same purpose, that is, to avoid catastrophic forgetting by training the network along the direction orthogonal to the inputs of learned tasks. However, as demonstrated below, there are key differences in the two approaches, leading to substantial performance gaps between them.

To shield previously learned knowledge by training the network in a direction orthogonal to the space spanned by previous inputs, the most important step is to construct a proper orthogonal projector. The design of the projector differentiates the OWM from the CAB significantly. The OWM constructs the orthogonal projector as $A(A^T A)^{-1}A^T$, where A contains previous inputs as its columns. The construction of this orthogonal projector is mathematically sound [1–3]. On top of the exact solution of the orthogonal projector, we introduced the term αI to change the projector as $A(A^T A + \alpha I)^{-1} A^T$. This enabled us to protect the entire subspace spanned by previous inputs or part of it to avoid the influence of noise [3, 4].

In CAB, the construction of the projector is based on the Conceptor, defined as C, which minimizes the loss function of $\mathbb{E}_x[||\mathbf{x} - C\mathbf{x}||^2] + \alpha^{-2}||C||_{fro}$ (Eq. 1 in [5]). The first item, i.e., $\mathbb{E}_x[||\mathbf{x} - C\mathbf{x}||^2]$ in this loss function is all required to construct a proper projector. Obviously, the linear system $\mathbf{x} = C\mathbf{x}$ is compatible and has infinite solutions especially when $\mathbf{x}$ does not span all possible space, which is always the case for tasks learned earlier in sequential learning. The equation $\mathbf{x} = C\mathbf{x}$ has a unique and exact solution under the constraint of minimum norm of C, i.e., $\min\limits_{\mathbf{x}-C\mathbf{x}=0} ||C||$. The equation $\mathbf{x} = C\mathbf{x}$ is equivalent to

$$CA = A \tag{1}$$

where A is the matrix consisting of data vectors as its columns, as defined in our manuscript. To solve Eq.1, they can vectorize and reorganize it as follows [1, 6]

$$(I \otimes A^T)vec(C) = vec(A) \tag{2}$$

where $\otimes$ denotes the Kronecker product and $vec(C) = vec([\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_n]) = [\mathbf{c}_1^T, \mathbf{c}_2^T, \cdots, \mathbf{c}_n^T]^T$ (so to $vec(A)$). Finding the minimum norm solution [1] of Eq.2 will lead to $C = A(A^T A)^{-1} A^T$, which is the orthogonal projector used in the OWM method. However, CAB does not solve the equation of $A\mathbf{x} = \mathbf{x}$ exactly. Instead, it constructs the projector based on minimizing the loss function cited above. The second term, i.e., $\alpha^{-2}||A||_{fro}$, in the loss function inevitably introduces error to the projector. In other words, the solution (Eq.2 in [5]) that minimizes loss function cannot provide accurate protection of previous input space. It would lead to errors in updating weights, eventually resulting in unwanted interference with previously learned tasks and compromising the ability of continual learning. Furthermore, it is worth noting that in the limiting case of α^{-2} approaching zero, the Conceptor would be ill-defined because for earlier tasks in continual learning, the correlation matrix in the calculation would not be of full rank and, therefore, non-invertible. Thus, the projector in CAB cannot approach the one used in the OWM by changing parameters.

Furthermore, we demonstrate with MNIST inputs that the above-mentioned theoretical difference indeed leads to substantial error in the orthogonal projector of CAB, in comparison to OWM. To demonstrate that the difference in projector construction explained above will lead to differences in accuracy for real data, we compared the error between the original inputs and that after the operation of the projector, $\frac{1}{N}\sum_{i=1}^{N}\sqrt{\frac{1}{s}||\mathbf{x}_i - \mathbf{x}_i^{pro}||}$, $\mathbf{x}^{pro} = (I - P^{OWM})\mathbf{x}$ or $C\mathbf{x}$, where N is the number of data samples and s is the size/dimension of each input. A more accurate projector will lead to smaller differences between the two, and vice versa. In Supplementary Fig. 6, we plotted the errors for CAB and OWM with different parameters. by taking image data of "0" and "1" in MNIST dataset as examples. The projection by the OWM was more accurate over a wide range of possible parameter values.

To investigate the influence of the above differences in overcoming catastrophic interference, we tested the two methods in 10-disjoint MNIST task, in which the network learned to recognize digits from 0 to 9, one class at a time. As shown in Fig. 2, with more digits learned, the performance gap between the two methods increased, and was larger than 10% after learning all 10 digits. This is consistent with our above analyses, indicating the OWM provides much better protection to previously learned knowledge.

To further examine properties of these methods, we analyzed the learning curve of OWM to demonstrate the degree of forgetting occurred for the earlier tasks. Supplementary Fig. 7 shows learning curves of OWM, CAB and SGD on the 2-disjoint MNIST tasks (Task A, classifying 0,1,2,3,4; Task B, classifying 5,6,7,8,9). After training of Task B, OWM showed minimum decrease in the performance of Task A, indicating only slight "forgetting" of the earlier task.In contrast, CAB showed more pronounced forgetting, consistent with our analysis that the CAB is inferior in protecting learned knowledge. Finally, without any mechanism to prevent catastrophic forgetting (SGD), Task A was completely forgotten after learning Task B.

Relationship of OWM and RLS

Below we illustrate how the projector we constructed in OWM is related to the $P^{(RLS)} = (\sum_{i=1}^{n}\mathbf{x}(i)\mathbf{x}^T(i) + \alpha I)^{-1}$ used in RLS, in the case that $\sum_{i=1}^{n}\mathbf{x}(i)\mathbf{x}^T(i) + \alpha I$ is invertible. $P^{(RLS)}$ is the inversion of the correlation matrix Φ of the input signals, i.e., $P^{RLS}(n) = \Phi^{-1}(n)$, where

$$\Phi(n) = \sum_{i=1}^{n}\gamma^{n-i}\mathbf{x}(i)\mathbf{x}(i)^T + \alpha\gamma^n I \tag{3}$$

Assume $\gamma = 1$ and let $A(n) = [\mathbf{x}(1), \mathbf{x}(2), \cdots, \mathbf{x}(n)]$, where x(i) is a vector recording the ith input, Φ can also be written as

$$\Phi(n) = A(n)A^T(n) + \alpha I \tag{4}$$

According to the Woodbury matrix identity

$$\begin{aligned} P^{RLS}(n) &= \alpha^{-1}I - \alpha^{-1}A(I + \alpha^{-1}A^TA)^{-1}A^T\alpha^{-1} \\ &= \alpha^{-1}[I - A(\alpha I + A^TA)^{-1}A^T] \end{aligned} \tag{5}$$

which is equivalent to the projector we constructed in OWM.

Relationship of OWM and online EWC

The loss function of online EWC[7] can be organized as:

$$L = L_{err} + \frac{1}{2}\lambda L_{reg} \tag{6}$$

where

$$L_{reg} = \sum_i (\theta_i - \theta_i^*)^T F_i (\theta_i - \theta_i^*) \tag{7}$$

θ_i and $\theta_i^*$are the input weights of neuron i in the current and last tasks, respectively, and F_i is the corresponding Fisher information matrix. Taking gradients on both sides of Eq.6 for θ_i

$$\begin{aligned}\nabla_{\theta_i} L &= \nabla_{\theta_i} L_{err} + \frac{1}{2}\lambda \nabla_{\theta_i} L_{reg} \\ &= \nabla_{\theta_i} L_{err} + \lambda F_i(\theta_i - \theta_i^*) \\ &\approx \nabla_{\theta_i} L_{err} - \lambda F_i \nabla_{\theta_i} L_{err} \\ &= (I - \lambda F_i)\nabla_{\theta_i} L_{err}\end{aligned} \tag{8}$$

Before the activation function, the input-output map of the i^{th} neuron can be viewed as a linear model with white noise (errors)

$$y_i = {\theta_i}^T \mathbf{x}_i + \epsilon_i \tag{9}$$

There is a relationship between F_i, the Fisher's information matrix of θ_i and R_i, the correlation matrix of its input $\mathbf{x}_i$,

$$F_i = \frac{1}{\sigma^2} R_i \tag{10}$$

where σ is the variance of white noise (or error) ϵ_i. If we define the orthogonal projection matrix as $P_i^{(OWM)} = I - (1+\alpha)A(A^T A + \alpha I)^{-1} A^T$, and provided that $\lambda \sim \sigma^2$, then substituting Eq.10 into Eq.8 we can derive the following

$$\begin{aligned}\nabla_{\theta_i} L &= (I - \lambda F_i)\nabla_{\theta_i} L_{err} \\ &= (I - \frac{\lambda}{\sigma^2} R_i)\nabla_{\theta_i} L_{err} \\ &\approx \lim_{\alpha \to \infty} P^{(OWM)} \Delta {\theta_i}^{bp}\end{aligned} \tag{11}$$

indicating that the online EWC can be approximated as a special case of OWM in the limit of α approaching infinity.

SUPPLEMENTARY METHODS

Datasets

The MNIST [8] database contains handwritten digits from 0 to 9 collected by the National Institute of Standards and Technology (NIST). MNIST has a training set of 60,000 samples and a testing set of 10000 samples. Each sample is a gray scale picture, with the size of 28×28.

The CIFAR-10 dataset (Canadian Institute For Advanced Research) [9] contains 60,000 color images in 10 different classes, representing animals and vehicles. Each sample is an image with the size of 32×32×3. There are 50000 images for training and 10000 images for testing.

The ILSVR2012 [10] is a subset from ImageNet, the world's largest image recognition database [11]. There are in total 1,000 categories of images to be classified. The training dataset contains 1.2 million images. The validation dataset contains 50,000 images belonging to the same 1000 categories. The classification accuracies for this task were calculated based on the validation set.

The offline Chinese handwriting database CASIA-HWDB [12] was collected by the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences. The dataset consists of isolated handwritten Chinese characters. Here we used a CASIA-HWDB1.1 subset, which has more than one million samples written by 300 writers. Furthermore, it contains 3755 commonly used Chinese characters, with each class containing 240 training images and 60 testing images.

Large-scale CelebFaces Attributes (CelebA) [13] contains 202599 celebrity face images of 10177 identities, covering a wide range of attitude and background clutter. Each image has 40 binary attributes annotated (see Fig. 4c or Supplementary Table 5 for all attributes).

SUPPLEMENTARY REFERENCES

[1]Ben-Israel, A. & Greville, T. N. *Generalized inverses: theory and applications*, vol. 15 (Springer Science & Business Media, 2003).

[2]Puntanen, S. Projection matrices, generalized inverse matrices, and singular value decomposition by haruo yanai, kei takeuchi, yoshio takane. *International Statistical Review* **79**, 503–504 (2011).

[3]Haykin, S. S. *Adaptive filter theory* (Pearson Education India, 2008).

[4]Moustakides, G. V. Study of the transient phase of the forgetting factor rls. *Ieee Transactions on Signal Processing* **45**, 2468–2476 (1997).

[5]He, X. & Jaeger, H. Overcoming catastrophic interference using conceptor-aided backpropagation. In *International Conference on Learning Representations* (2018).

[6]Marcus, M. & Minc, H. *A survey of matrix theory and matrix inequalities*, vol. 14 (Courier Corporation, 1992).

[7]Schwarz, J. *et al.* Progress & compress: A scalable framework for continual learning. In *International Conference on Machine Learning*, 4535–4544 (2018).

[8]LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**, 2278–2324 (1998).

[9]Krizhevsky, A. & Hinton, G. Learning multiple layers of features from tiny images. Tech. Rep., Citeseer (2009).

[10]Russakovsky, O. *et al.* Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* **115**, 211–252 (2015).

[11]Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255 (Ieee, 2009).

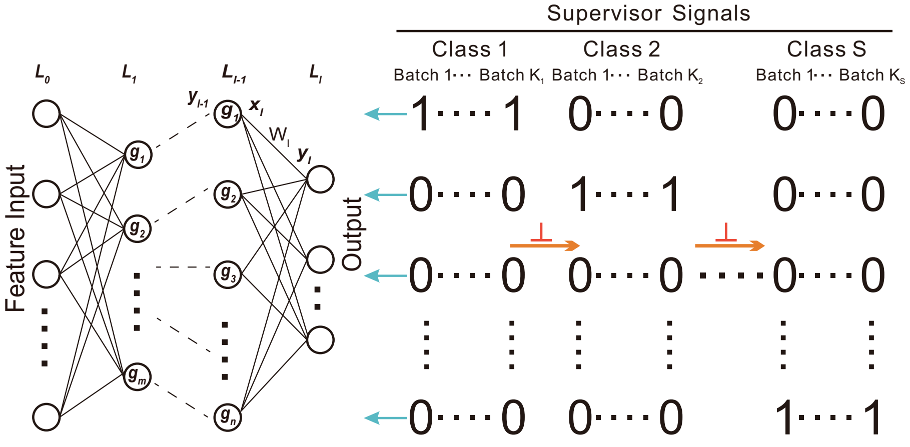
[12]Liu, C.-L., Yin, F., Wang, D.-H. & Wang, Q.-F. Casia online and offline chinese handwriting databases. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, 37–41 (IEEE, 2011).

[13]Liu, Z., Luo, P., Wang, X., Tang, X. & Ieee. *Deep Learning Face Attributes in the Wild*, 3730–3738. IEEE International Conference on Computer Vision (2015).

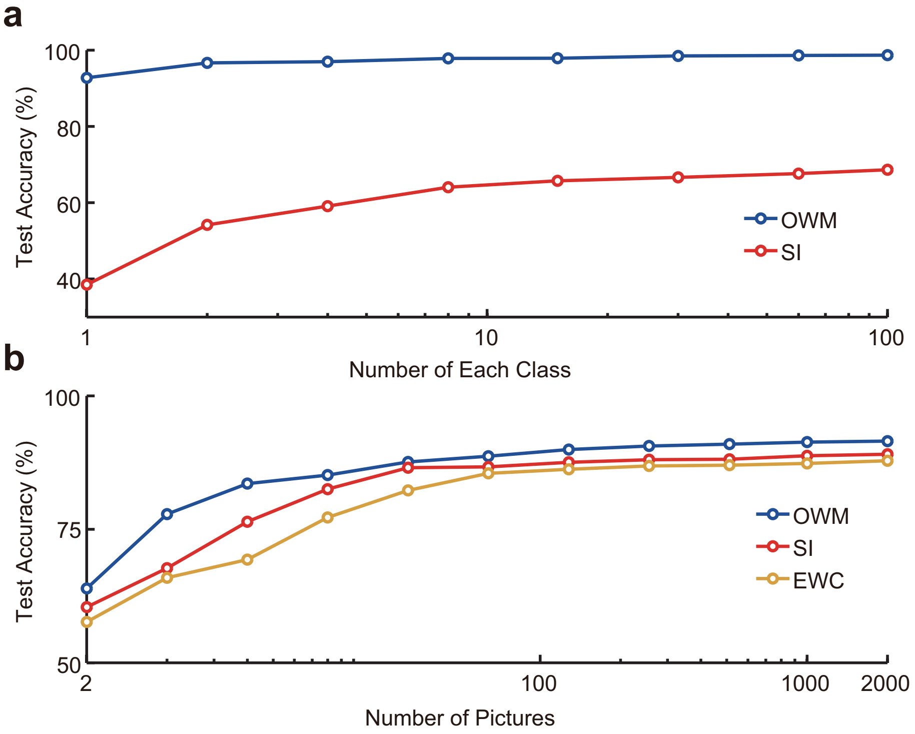
[14]He, K. M., Zhang, X. Y., Ren, S. Q., Sun, J. & Ieee. *Deep Residual Learning for Image Recognition*, 770–778. IEEE Conference on Computer Vision and Pattern Recognition (Ieee, New York, 2016).

[15]Tieleman, T. & Hinton, G. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning* **4**, 26–31 (2012).

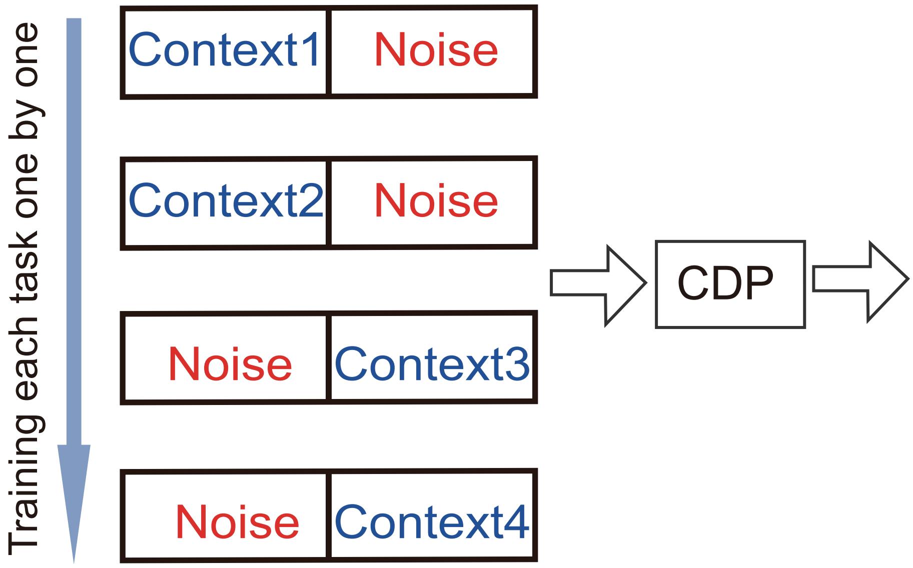
SUPPLEMENTARY FIGURES AND TABLES



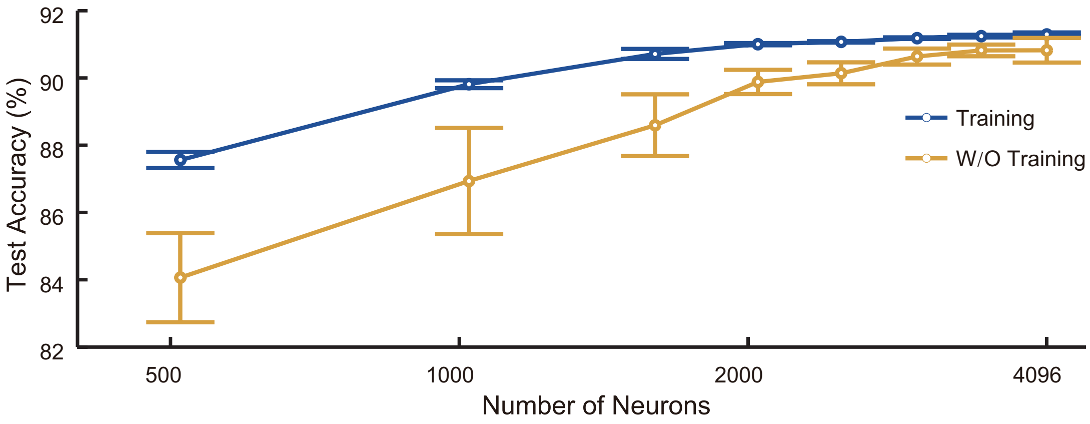
Supplementary Figure 1. Schematic diagram showing sequential learning procedure using OWM. There were S classes of data to be learned. Training samples of each class were sequentially fed to the neural network. Samples of the i^{th} class were further divided into K_i batches for training. Symbol "⊥" denotes the application of the orthogonal projector P (see Methods for details).



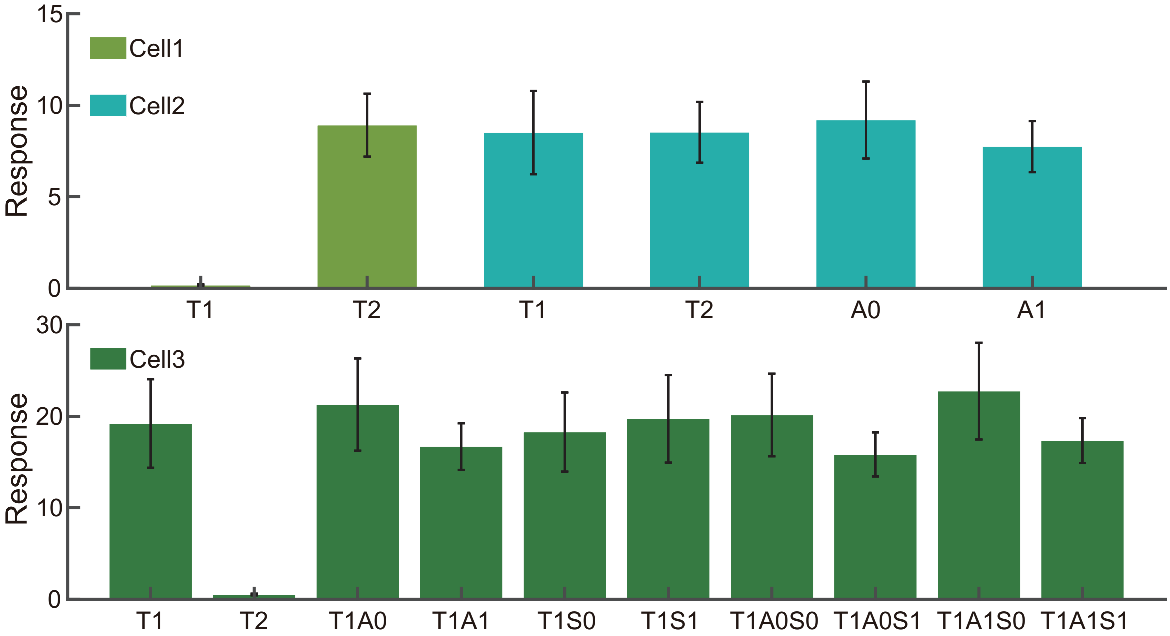
Supplementary Figure 2. Comparison of OWM with EWC and SI for learning with small sample size. a, Performance of OWM and SI with respect to the number of images in each class for training in Chinese character classification task. In total, 20 classes were trained sequentially. Network structure was the same as in Fig. 3c for both methods. The EWC method failed in this task and was not included. **b**, Performance of OWM, SI and EWC with respect to number of images in each class for training in the face recognition task. Only difficult tasks were considered and the network structure was the same as in Fig. 4e for all three methods.



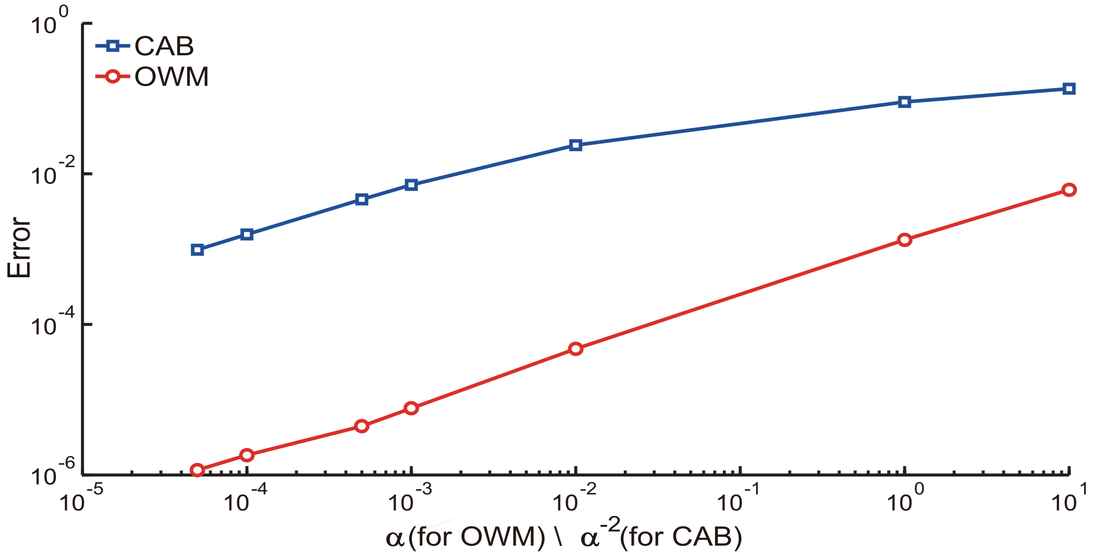
Supplementary Figure 3. Schematic diagram of the task with the CDP module inferring context from noisy environments. Four face recognition tasks were trained continually with the CDP module as in Fig. 4. The signal fed to the CDP module included two parts: correct context signal and noise. Noises were sampled from a Gaussian distribution with the same mean and variance as the true context signal but varied on a trial-by-trial basis. During the testing phase, either the corresponding context+noises or only noises were presented.



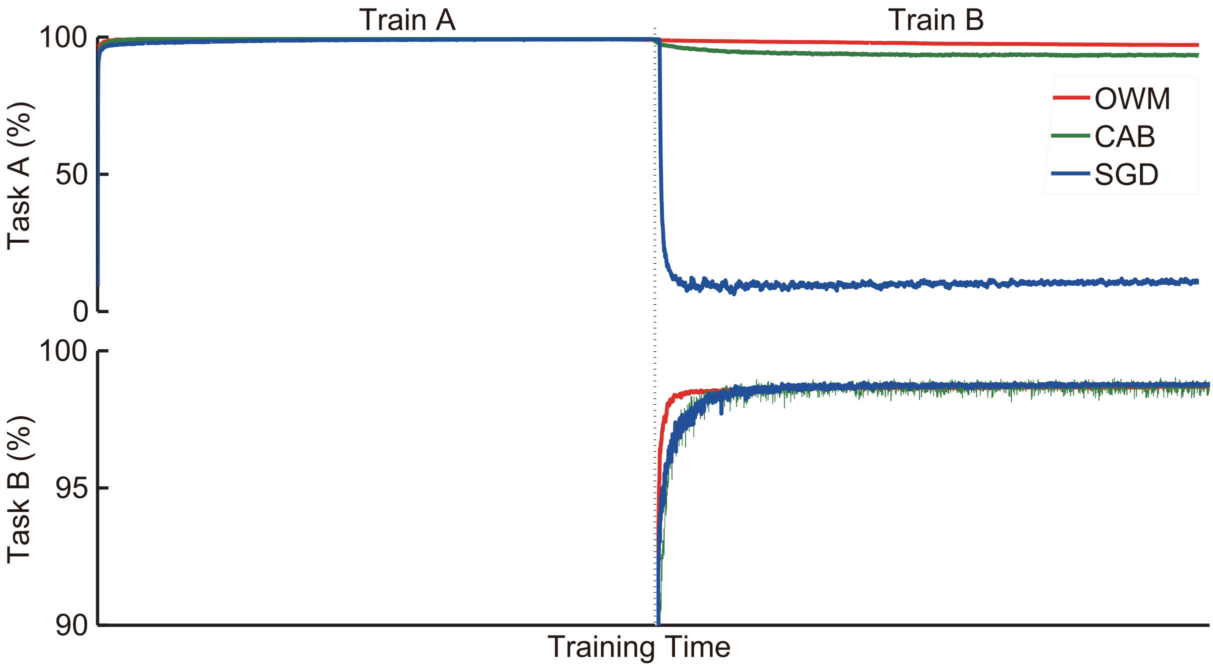
Supplementary Figure 4. Comparison between performances of neural networks with trainable and fixed encoder in the CDP module. The task is the same as in Fig. 4 and performed in a similar neural network structure with various numbers of neurons in the rotator sub-module of the CDP module. The encoder was either trained (blue line) or fixed (yellow line) through all tasks. Results are represented as mean ± s.d. averaged over all tasks across 5 trials. Training of the CDP module led to increased performance, especially in the cases with limited number of neurons for feature rotation.



Supplementary Figure 5. Selectivity of three exemplar neurons in rotator layer in context-dependent processing. Before the analysis, the network was sequentially trained to perform two tasks depending on contextual information. Task 1 (T1): classify if a face is attractive (A1) or not (A0). Task 2 (T2): classify if a face is smiling (S1) or not (S0). Data are presented as mean $\pm$ s.d across all correctly classified images.



Supplementary Figure 6. Errors between original inputs and projected inputs by CAB and OWM with respect to different parameters. Note that parameter α^{-2} (for CAB) and α (for OWM) are two different parameters and not related (see Supplementary Discussion for details).



Supplementary Figure 7. Learning curves in the 2-disjoint MNIST task for OWM, CAB, and SGD. Task A was to recognize digits from 0 to 4, and Task B was to recognize digits from 5 to 9.

Supplementary Table 1. Performance of sequential learning achieved by the OWM in comparison with traditional concurrent training method in various datasets. ResNet was adopted from [14].

Data Set	Classes	Feature Extractor	Concurrent Training by SGD (%)	Sequential Training by OWM (%)	Sequential Training by SGD (%)
ImageNet	1000	ResNet152	78.31	73.80	0.69
CASIA-HWDB1.1	3755	ResNet18	94.39	92.11	8.07

Supplementary Table 2. Means and standard deviation in Fig. 3b across classes. Mean and standard deviation across classes of the task with 3755 pre-training classes is 94.3 ± 5.8.

Number of Pre-training Classes	All Classes Mean Accuracy (%)	Without Pre-training Classes Mean Accuracy (%)
50	56.82 ± 17.64	56.60 ± 17.59
100	63.42 ± 16.11	63.06 ± 16.02
200	68.47 ± 15.63	68.05 ± 15.48
300	74.66 ± 15.65	73.99 ± 15.69
400	76.90 ± 16.78	76.07 ± 16.84
500	78.31 ± 16.10	77.38 ± 16.19
1000	81.48 ± 18.91	79.96 ± 19.67
1500	84.19 ± 17.23	82.15 ± 17.92
2000	85.21 ± 18.24	83.64 ± 18.10
2500	87.02 ± 17.07	84.89 ± 17.77
3000	88.15 ± 16.09	85.59 ± 17.61
3500	89.25 ± 14.61	89.49 ± 12.99
3755	90.73 ± 13.56	NA

Supplementary Table 3. Mean and standard deviation in Fig. 3c across classes with different degrees of pre-training.

Number of Each Class	Pre-training 3755 Mean Accuracy (%)	Pre-training 3000 Mean Accuracy (%)	Pre-training 2000 Mean Accuracy (%)
1	53.09 ± 37.2	46.38 ± 35.58	38.64 ± 33.39
2	80.89 ± 24.38	74.38 ± 27.35	70.14 ± 25.37
4	89.65 ± 12.15	85.46 ± 15.27	81.00 ± 16.47
8	90.18 ± 13.92	86.75 ± 16.59	83.92 ± 16.07
16	90.07 ± 14.46	87.68 ± 16.11	84.30 ± 18.23
32	89.99 ± 14.92	86.99 ± 17.87	84.16 ± 19.27
64	90.13 ± 14.71	87.01 ± 18.05	84.00 ± 20.13
128	89.71 ± 15.47	86.61 ± 18.95	85.15 ± 17.98
240	89.91 ± 15.52	88.39 ± 15.38	85.82 ± 16.57

Supplementary Table 4. Performance of the neural network with the CDP module in noisy environment. The second column is the test accuracy with context signal+noises during both the training and testing phase. The third column is the test accuracy with signal+noises presented during the training phase but only noises presented during the testing phase. Results indicate that the CDP module can infer correct context signal from a non-stationary stream of inputs by itself.

Attributes	**Normal Context Signal Accuracy (%)**	**Random Noise Accuracy (%)**
High Cheekbones	86.02	70.05
Male	97.96	48.51
Mouth Small Open	92.86	73.88
Wear Lipstick	92.79	59.46
Mean	92.41	62.97

Supplementary Table 5. Performance of OWM in context-dependent, sequential learning, compared with results obtained by multi-task training, in recognizing 40 different facial attributes in CelebA. MT, multi-task training; ST, sequential training with OWM.

Attributes	5 Shadow	Arched eyebrows	Attractive	Bags under eyes	Bald	Bangs	Big lips	Big nose	Black hair	Blond hair	Blurry	Brown hair	Bush eyebrows	Chubby
ST	94.75	84.74	90.41	85.28	98.96	96.27	71.53	83.93	90.22	96.08	96.29	89.79	93.00	95.73
MT	94.82	83.83	83.01	85.28	99.00	96.27	71.66	84.89	90.55	96.10	96.28	89.56	92.95	95.83
Attributes	Double chin	Eyeglasses	Goatee	Gray hair	Heavy makeup	High cheekbones	Male	Mouth small open	Mustache	Narrow eyes	No beard	Oval face	Pale skin	Pointy nose
ST	96.35	99.66	97.61	98.28	91.81	87.90	98.61	93.91	96.75	87.23	96.18	75.72	96.78	77.17
MT	96.41	99.67	97.67	98.36	91.98	87.96	98.56	93.99	97.08	87.35	96.28	75.91	97.25	77.44
Attributes	Receding hairline	Rosy cheeks	Sideburns	Smiling	Straight hair	Wavy hair	Wear earrings	Wear hat	Wear lipstick	Wear necklace	Wear necktie	Young	**Average**	
ST	93.92	94.88	97.67	92.98	84.37	84.52	90.64	99.05	93.99	87.00	96.11	88.58	**91.43**	
MT	93.97	95.12	97.88	93.13	84.60	85.05	90.70	99.15	94.32	87.11	96.61	88.69	**91.56**	

Supplementary Table 6. Hyperparameters of feature extractors used in different tasks. ResNet was adopted from [14]. Optimization method of all feature extractor was RMSprop [15].

Experiment	**Feature Extractor/Output Size**	**Learning Rate**	**Weight Decay**	**Batch Size**
ImageNet ILSVR2012	ResNet152 / 2048	0.1	0.0001	512
CASIA-HWDB1.1	ResNet18 / 1024	0.1	0.0001	512
CelebA	ResNet50 / 2048	0.1	0.0001	256

Supplementary Table 7. Hyperparameters for the OWM/SGD method used in different experiments. If different values were used for different layers, all values are listed in a row, from the input layer (left) to output layer (right).

Experiment	α	λ	κ	**Batch Size**
Shuffled MNIST (3 Tasks)	1.0	NA	1.0	100
Shuffled MNIST (10 Tasks)	1.0	NA	4.0	100
Shuffled MNIST (100 Tasks)	1.0/1.0/0.5	0.0001/0.1/1.0	4.0	100
Disjoint MNIST (3 Layers)	0.9/0.6	0.001/1.0	0.2	40
Disjoint MNIST (4 Layers)	0.9/1.0/0.6	0.001/0.1/1.0	0.2	40
Disjoint MNIST (SGD)	NA	NA	0.01	50
CASIA-HWDB1.1 (3 Layers)	1.0/0.5	0.02/1.0	2.0	50
CASIA-HWDB1.1(SGD)	NA	NA	0.0001	200
ImageNet ILSVR2012 (3 Layers)	1.0/1.0	0.005/1.0	2.0	30
ImageNet ILSVR2012 (SGD)	NA	NA	0.0001	1000
CelebA (2 Layers in Fig .4c)	NA/0.1	NA	0.15	20
CelebA (Encoder in Fig .4c)	1.0	NA	1.5	20
CelebA (2 Layers in Fig .4e)	NA/0.5	NA	1.0	1
CelebA (Encoder in Fig .4e)	1.0	NA	10.0	1
Convolutional Layers in CIFAR10	1.0	0.00001	0.02	64
Fully Connected Layers in CIFAR10	1.0	0.0001/0.01/0.1	0.02	64