

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*}

Deep neural networks are powerful tools in learning sophisticated but fixed mapping rules between inputs and outputs, thereby limiting their application in more complex and dynamic situations in which the mapping rules are not kept the same but change according to different contexts. To lift such limits, we developed an approach involving a learning algorithm, called orthogonal weights modification, with the addition of a context-dependent processing module. We demonstrated that with orthogonal weights modification to overcome catastrophic forgetting, and the context-dependent processing module to learn how to reuse a feature representation and a classifier for different contexts, a single network could acquire numerous context-dependent mapping rules in an online and continual manner, with as few as approximately ten samples to learn each. Our approach should enable highly compact systems to gradually learn myriad regularities of the real world and eventually behave appropriately within it.

One of the hallmarks of high-level intelligence is flexibility¹. Humans and non-human primates can respond differently to the same stimulus under different contexts, for example, different goals, environments and internal states^{2–5}. Such an ability, named cognitive control, enables us to dynamically map sensory inputs to different actions in a context-dependent way^{6–8}, thereby allowing primates to behave appropriately in an unlimited number of situations with limited behavioural repertoire^{9,10}. This flexible, context-dependent processing is quite different to that found in current artificial deep neural networks (DNNs). DNNs are very powerful in extracting high-level features from raw sensory data and learning sophisticated mapping rules for pattern detection, recognition and classification¹¹. In most networks, however, the outputs are largely dictated by sensory inputs, exhibiting stereotyped input–output mappings that are usually fixed once training is complete. Therefore, current DNNs lack sufficient flexibility to work in complex situations in which (1) the mapping rules change according to context and (2) these rules need to be learned sequentially when encountered from a small number of learning trials. This constitutes a significant gap in the abilities between current DNNs and primate brains.

Here, we propose an approach, including an orthogonal weight modification (OWM) algorithm and a context-dependent processing (CDP) module, that enables a neural network to progressively learn various mapping rules in a context-dependent way. We demonstrate that with OWM to protect previously acquired knowledge, the networks could sequentially learn up to thousands of different mapping rules without strong interference, and needing as few as approximately ten samples to learn each. In addition, by using the CDP module to enable contextual information to modulate the representation of sensory features, a network can learn different, context-specific mappings for even identical stimuli. Taken together, our proposed approach can teach a single network numerous context-dependent mapping rules in an online and continual manner.

Orthogonal weights modification

The first step towards flexible context-dependent processing is to incorporate efficient and scalable continual learning, that is,

learning different mappings sequentially, one at a time. Such an ability is crucial to humans as well as artificial intelligence agents for two reasons: (1) there are too many possible contexts to learn concurrently, and (2) useful mappings cannot be pre-determined but must be learned when corresponding contexts are encountered.

The main obstacle to achieving continual learning is that conventional neural network models suffer from catastrophic forgetting, that is, training a model with new tasks interferes with previously learned knowledge and leads to significant decreases in the performance of previously learned tasks^{12–15}. To avoid catastrophic forgetting, we developed the OWM method. Specifically, when training a network for new tasks, its weights can only be modified in the direction orthogonal to the subspace spanned by all previously learned inputs (termed the input space hereafter) (Fig. 1a and Supplementary Fig. 1). This ensures that new learning processes do not interfere with previously learned tasks, as weight changes in the network as a whole do not interact with old inputs. Consequently, combined with a gradient descent-based search, the OWM helps the network to find a weight configuration that can accomplish new tasks while ensuring the performance of learned tasks remains unchanged (Fig. 1b). Specifically, this is achieved by first constructing a projector used to find the direction orthogonal to the input space: $P = I - A(A^T A + \alpha I)^{-1} A^T$, where matrix A consists of all previously trained input vectors as its columns $A = [\mathbf{x}_1, \dots, \mathbf{x}_n]$, I is a unit matrix multiplied by a relatively small constant α , and A^T is the transpose of A . The learning-induced modification of weights is then determined by $\Delta W = \kappa P \Delta W^{\text{BP}}$, where κ is the learning rate and ΔW^{BP} is the weights adjustment calculated according to the standard backpropagation (BP). To calculate P , an iterative method can be used (see Methods). Thus, the algorithm does not need to store all previous inputs A . Instead, only the current inputs and projector for the last task are needed. This iterative method is related to the recursive least square (RLS) algorithm^{6,17} (see Supplementary Information for details), which can be used to train feed-forward and recurrent neural networks to achieve fast convergence^{18,19}, tame chaotic activities²⁰ and avoid interference between consecutively loaded patterns or tasks^{21,22}.

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

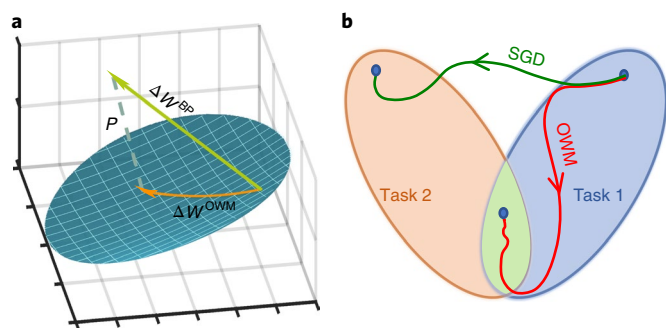


Fig. 1 | Schematic of OWM. **a**, In the new task training process, the original weight modification calculated by the standard BP, ΔW^{BP} , is projected to the subspace (dark green surface), in which good performance for learned tasks has been achieved. As a result, the actually implemented weight modification is ΔW^{OWM} . This process ensures that the weights configuration after learning the new task is still within the same subspace. **b**, With the OWM, the training process searches for configurations that can accomplish task 2 (pale red area), within the subspace that enables the network to accomplish task 1 (blue area). A successful search necessarily stops at a position inside the overlapping subspace (light green area). In comparison, the solution obtained by stochastic gradient descent search (SGD) is more likely to end outside this overlapping area.

Table 1 | Comparison of performance of different methods in the shuffled MNIST tasks

3 tasks	Accuracy (%)	10 tasks	Accuracy (%)	100 tasks	Accuracy (%)
SGD ^{a14}	71.32 ± 1.54*	EWC ^{a24}	~97.0	EWC ^{c41}	~70.8
IMM ^{a25}	98.30 ± 0.08 (NS)	OWM ^a	97.52 ± 0.03	SI ^{c41}	~82.3
EWC ^{a24}	~98.2	EWC ^{b22}	~89.0	OWM ^c	~85.4
OWM ^a	98.34 ± 0.02	CAB ^{b22}	~95.2		
		OWM ^b	95.15 ± 0.08		
		SI ^{c26,41}	~97.0		
		OWM ^c	97.64 ± 0.03		

Network size: ^aFour-layer networks with [784–800–800–10] neurons. ^bThree-layer networks with [784–100–10] neurons. ^cFour-layer networks with [784–2,000–2,000–10] neurons. Results from other methods were adopted from corresponding publications. Results for OWM are represented as mean ± s.d. * $P < 0.01$. NS, not significant; IMM, incremental moment matching; SGD, stochastic gradient descent; SI, synaptic intelligence.

We first tested the performance of the OWM on several benchmark tasks of continual learning. Experiments on shuffled and disjoint Modified National Institute of Standards and Technology (MNIST) dataset, in which different tasks involving recognition of handwritten digits need to be learned sequentially (see Methods and Supplementary Information for details regarding the datasets used in this study), were conducted on the feed-forward network with the rectified linear unit (ReLU)²³. The OWM was used to train the entire multilayer networks. For three- or ten-task shuffled and two-task disjoint experiments, OWM resulted in either superior or equal performance compared with other continual learning methods without storage of previous task samples or dynamically adding new nodes to the network^{22,24–26} (Tables 1 and 2). In the more challenging 10-task disjoint and 100-task shuffled experiments, OWM exhibited significant performance improvement over other methods (Fig. 2 and Table 1). Interestingly, for the more difficult continual learning tasks, we found that the order of tasks mattered. As the performance for specific classes can be significantly influenced

Table 2 | Comparison of performance of different methods in the disjoint MNIST tasks

Method	Accuracy (%)
EWC ^{a25}	52.72 ± 1.36*
IMM ^{a25}	94.12 ± 0.27*
OWM ^a	96.59 ± 0.06
SGD ^b	53.85 ± 0.14*
CAB ^{b22}	94.91 ± 0.30*
OWM ^b	96.30 ± 0.03

Network size: ^aFour-layer networks with [784–800–800–10] neurons. ^bThree-layer networks with [784–800–10] neurons. Performance results from other methods were adopted from previous studies. The results are represented as mean ± s.d. * $P < 0.01$.

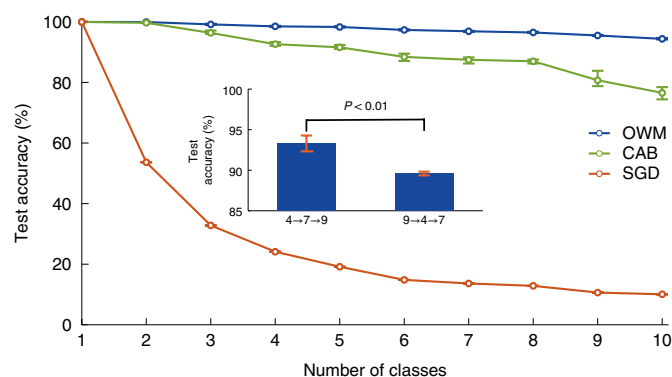


Fig. 2 | Performance of OWM, CAB and SGD in the ten-task disjoint MNIST experiment. Test accuracy was plotted as a function of number of classes learned. Results are presented as mean ± s.d. For the OWM-trained task, the sequence of learnt digits influenced recognition accuracy for specific classes. Inset: performance of recognizing digit ‘9’ was significantly higher after learning digits ‘7’ and ‘4’; two-sided t-test was applied to assess statistical significance.

by the classes learned previously (Fig. 2, inset), suggesting that curriculum learning is a potentially important factor to consider in continual learning.

To examine whether the OWM is scalable, that is, whether it can be applied to learn more sophisticated tasks, regarding both number of different mappings and complexity of inputs, we tested the network’s ability in learning to classify thousands of handwritten Chinese characters (CASIA-HWDB1.1 dataset) and natural images (ImageNet dataset). The Chinese character recognition task included a total of 3,755 characters forming the level I vocabulary, which constitutes more than 99% of the usage frequency in written Chinese²⁷ (see Fig. 3a for examples of characters). In this task, a feature extractor was pre-trained to analyse the raw images. The feature vectors were fed into an OWM-trained classifier to learn the mapping between combinations of features and the labels of individual classes. We found that a classifier trained with the OWM could learn to recognize all 3,755 characters sequentially, with a final accuracy of ~92%, closely approaching the results obtained in human performance when recognizing handwritten Chinese characters (~96%)²⁸. Considering humans learn these characters over years and the learning necessarily contains revision, these results suggest that our method endows neural networks with a strong capability to continually learn new mappings between sensory features and class labels. Similar results were obtained with the ImageNet dataset, where the classifier trained by the OWM combined with a pre-trained feature extractor was able to learn 1,000

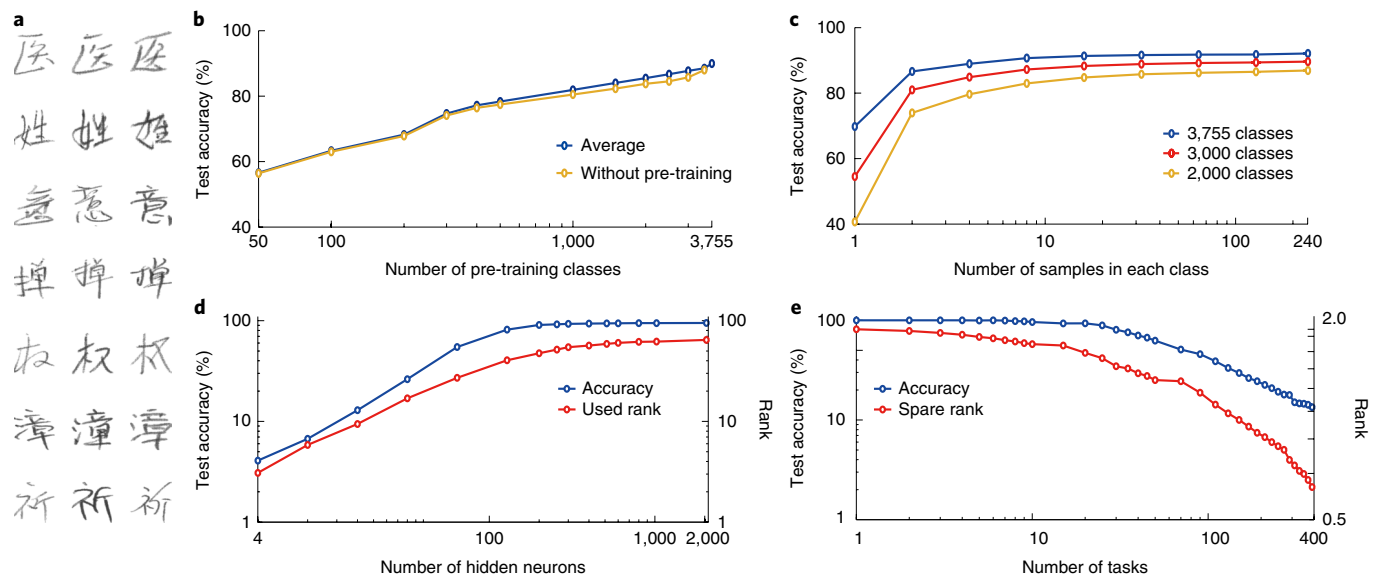


Fig. 3 | Continual learning with small sample size achieved by OWM in recognizing Chinese characters. **a**, Examples showing seven characters with three samples for each. **b**, Classification accuracy is plotted as a function of the number of classes used for pre-training the feature extractor. Performance was assessed based on classifying all characters (blue) or characters not included in pre-training (orange). Variance of test accuracy across classes in each case is reported in Supplementary Table 2. **c**, Classification accuracy is plotted as a function of sample size used for sequential training, obtained with feature extractors having different degrees of pre-training (colour-coded). Performances differed significantly (paired *t*-test, $P < 0.001$) across different degrees of pre-training (see Supplementary Table 3 for variance in performance across all classes). **d,e**, Relationship between network capacity for continual learning and rank of the orthogonal projector. In **d**, the task was to learn 100 classes of Chinese characters sequentially. Average accuracy achieved by the network (blue) and the corresponding value of $\text{rank}(\beta I) - \text{rank}(P)$ (red) are plotted with respect to the number of neurons in the hidden layer. In **e**, the same neural network with 50 neurons in the hidden layer was trained to recognize an increasing number of Chinese characters. Average accuracy (blue) achieved by the network and the corresponding value (red) of rank_{tot} (see Methods) are plotted with respect to the number of tasks/characters.

classes of natural images sequentially (Supplementary Table 1), with the final accuracy approaching the results obtained by training the system to classify all categories concurrently. These results suggest that, by using the OWM, the performance of the system in classification approaches the limit set by the front-end feature extractor, with liability to the classifier caused by sequential learning itself effectively mitigated.

In the results mentioned above, feature extractors pre-trained by the complete training sets in corresponding tasks were used to provide the feature vectors for the OWM-trained classifier. We next examined whether the classifier can learn categories on which the feature extractor has not been trained. Results were in the affirmative, as shown in Fig. 3b. For example, the feature extractor trained with 500 randomly selected Chinese characters (out of 3,755, less than 15% of categories) could already support the classifier to sequentially learn the remaining 3,255 characters with near 80% accuracy (chance level of $1/3,255$), demonstrating that the network could sequentially learn new categories not previously encountered. However, we note that a higher degree of pre-training was associated with better performance (Fig. 3b,c), indicating the importance of training the feature extractor on as various classes as possible.

Another important question is how quickly the OWM-trained classifier can learn. As shown in Fig. 3c, it only needed small sample size to learn new mappings. For Chinese characters, <10 samples per class were sufficient to gain satisfactory performance. Comparison with other methods in the same task further confirmed the advantage of the OWM in achieving better performance with fewer training samples (see Supplementary Fig. 2). We note that the better performance achieved by the OWM with fewer samples is rooted in the well-known fact that the RLS algorithm, from which we derived the OWM, can converge more quickly than the least mean square algorithm, which is equivalent to the standard backpropagation^{16,19}.

With the dataset of Chinese characters, we also analysed network capacity in the OWM-based continual learning. We tested two conditions, including reducing the size of the network for a given task and increasing the number of tasks for a given network. We observed that network performance remained stable until its size was reduced or the number of tasks was increased to a certain value, after which point performance declined, indicating the approach to network capacity (Fig. 3d,e). Importantly, the changes in network performance were highly correlated with decreases in the rank of the orthogonal projector, which is consistent with our theoretical analysis regarding network capacity in OWM-based continual learning (see Methods for details).

In the experiments with Chinese characters, it is possible that although a class was never seen by the network, it shared features with other classes used in feature extractor pre-training. Thus, to further test the ability of the OWM in continual learning without a pre-trained feature extractor, we examined its performance in the disjoint Canadian Institute For Advanced Research (CIFAR)-10 dataset task. In this task, the network was trained to recognize two classes each time; thus, in a total of five consecutive tasks it learned to recognize all ten classes. Importantly, the whole network, including both the feature extractor and classifier, was trained continually in an end-to-end manner. In this task, the OWM outperformed other recently proposed continual learning methods by a large margin (Table 3), exhibiting great potential to improve the networks' ability to learn new classes 'on the go' and with the feature extractor free of pre-training. Although the performance in an end-to-end training setting was still inferior than that with a pre-trained feature extractor, this is an important step towards removing the usual distinction between the training and application phases of DNNs, thus allowing efficient online learning. We note that a continual learning method for classifiers with a pre-trained feature extractor

Table 3 | Comparison of performance of different methods in the disjoint CIFAR-10 task

Method	Accuracy (%)
EWC ⁵⁷	31.09
IMM ⁵⁷	32.36
MA ⁵⁷	40.47
OWM	52.83

See Methods for details. MA, model adaptation.

may also be useful. While the number of features in a given domain (for example, human faces) is usually limited, the possible ways of combining different features to form a new object (for example, individual faces) are almost infinite. Thus, given feature extractors pre-trained on sufficiently diverse sample sets, a classifier could greatly benefit from continual learning to recognize countless new classes.

Context-dependent processing module

Although a system that can learn many different mapping rules in an online and sequential manner is highly desirable, such a system cannot accomplish context-dependent learning by itself. To achieve that, contextual information needs to interact with sensory information properly. Here we adopted a solution inspired by the prefrontal cortex (PFC). The PFC receives sensory inputs as well as contextual information, which enables it to choose sensory features most relevant to the present task to guide action^{4,5,29}. To mimic this architecture, we added the CDP module before the OWM-trained classifier, which was fed with both sensory feature vectors and contextual information (Fig. 4a). The CDP module consists of an encoder submodule, which transforms contextual information to proper controlling signals, and a 'rotator' submodule, which uses controlling signals to manipulate the processing of sensory inputs. The encoder submodule is trainable and learns in a continual way with the OWM. Mathematically, the context-dependent manipulation serves by rotating the sensory input space according to the contextual information (Fig. 4b, see Methods), thereby changing the representation of sensory information without interfering with its content. The rotation of the input space allows for OWM to be applied for identical sensory inputs in different contexts. To demonstrate the effectiveness of this CDP module, we trained the system to classify a set of faces according to 40 different attributes³⁰, that is, to learn 40 different mappings sequentially with the same sensory inputs. The contextual information was the embedding vectors³¹ of the corresponding task names, which were projected to control the rotation of the sensory inputs. As shown in Fig. 4c, the system sequentially learned all 40 different, context-specific mapping rules with a single classifier. The accuracy was very close to that achieved by multi-task training, in which the network was trained to classify all 40 attributes using 40 separate classifiers (Fig. 4d). In addition, similar to the results obtained in learning Chinese characters, the network was able to learn context-dependent processing quickly. Here, ~20 faces were enough to reach the learning plateau for both simple, for example, male versus female, and difficult, for example, attractive versus unattractive, tasks (Fig. 4e). In the experiment, our approach achieved better performance with fewer samples compared with other methods for continual learning, indicating its potential to enable a system to adapt quickly in highly dynamic environments with regularities changing with contexts (Supplementary Fig. 2b). Interestingly, we found that the CDP module was able to identify the meaningful signal from the contextual inputs with noise (Supplementary Fig. 3 and Supplementary Table 4, see Methods for task details) and to learn how to use the

contextual information effectively (Supplementary Fig. 4). These results indicate that our approach allows the system to infer the correct context signal from experience and use it properly. Importantly, such an ability would open the door for an intelligent agent to explore environments and gradually learn numerous regularities in an autonomous way.

Discussion

If we view traditional DNNs as powerful sensory processing modules, the current approach could be understood as adding a flexible cognitive module to the system. This architecture was inspired by the primate brain. For example, the primate visual pathway is dedicated to analysing raw visual images and eventually representing ~100 features in higher visual areas such as the inferotemporal cortex³². The outputs of this 'feature extractor' are then sent to the PFC for object identification and categorization^{33–35}. The training of the feature extractor is difficult and time-consuming. In humans, it takes years or even decades for higher visual cortices to become fully developed and reach peak performance³⁶. However, with sufficiently developed visual cortices, humans can quickly learn new visual object categories, often by seeing just a few positive examples³⁷. By adding a cognitive module supporting continual learning to DNN-based feature extractors, we found qualitatively similar behaviour in neural networks. That is, although the training of the feature extractor is computationally difficult and requires a large number of samples, with a well-trained feature extractor, the learning of new categories can be achieved quickly. This suggests that the mechanisms underlying fast concept formation in humans may be understood, at least in part, from a connectionist perspective. In addition to the role of supporting the fast learning of new concepts, another function of the primate PFC is to represent contextual information⁹ and use it to select those sensory features most relevant for the current task⁴. This gives rise to the flexibility exhibited in primate behaviour and here we demonstrated that similar architecture can do the same in artificial neural networks. Interestingly, we found that in the CDP module, the neuronal responses showed mixed selectivity to sensory features, contexts and their combinations (Supplementary Fig. 5), similar to that found for real PFC neurons³⁸. Thus, it would be informative to see whether the rotation of input space adopted in our CDP module captures the operation carried out in the real PFC. For tasks similar to the face classification tested above, one possible solution to achieve context-dependent processing is to add additional classifier outputs for each new task/context. However, this approach only works if there is no hidden layer between the feature extractor and final output layer. Otherwise, the shared weights between different classifier outputs will suffer from catastrophic forgetting during continual learning, especially if the inputs are the same for all contexts. More importantly, adding additional classifier outputs (and all related weights) for each new task/context would lead to increasingly complex and bulky systems (cf. Fig. 4d, left). As the total number of possible contexts can be arbitrarily large, such a solution is clearly not scalable. Finally, for artificial intelligence systems, the importance of the CDP module would depend on the application. In scenarios in which a compact system needs to learn numerous contexts 'on the go', similar to what human individuals need to do within their lifetimes, the ability of the OWM-empowered CDP module to reuse classifiers is of paramount importance.

As demonstrated in the present results, an efficient and scalable algorithm of continual learning is not only crucial for achieving flexible context-dependent processing, but also important to ensure, more generally, that the added cognitive module is able to learn new tasks when encountered. In continual learning, preserving previously acquired knowledge while maintaining plasticity for subsequent learning is the key¹⁵. In the brain, the separation of synapses utilized for different tasks is essential for sequential

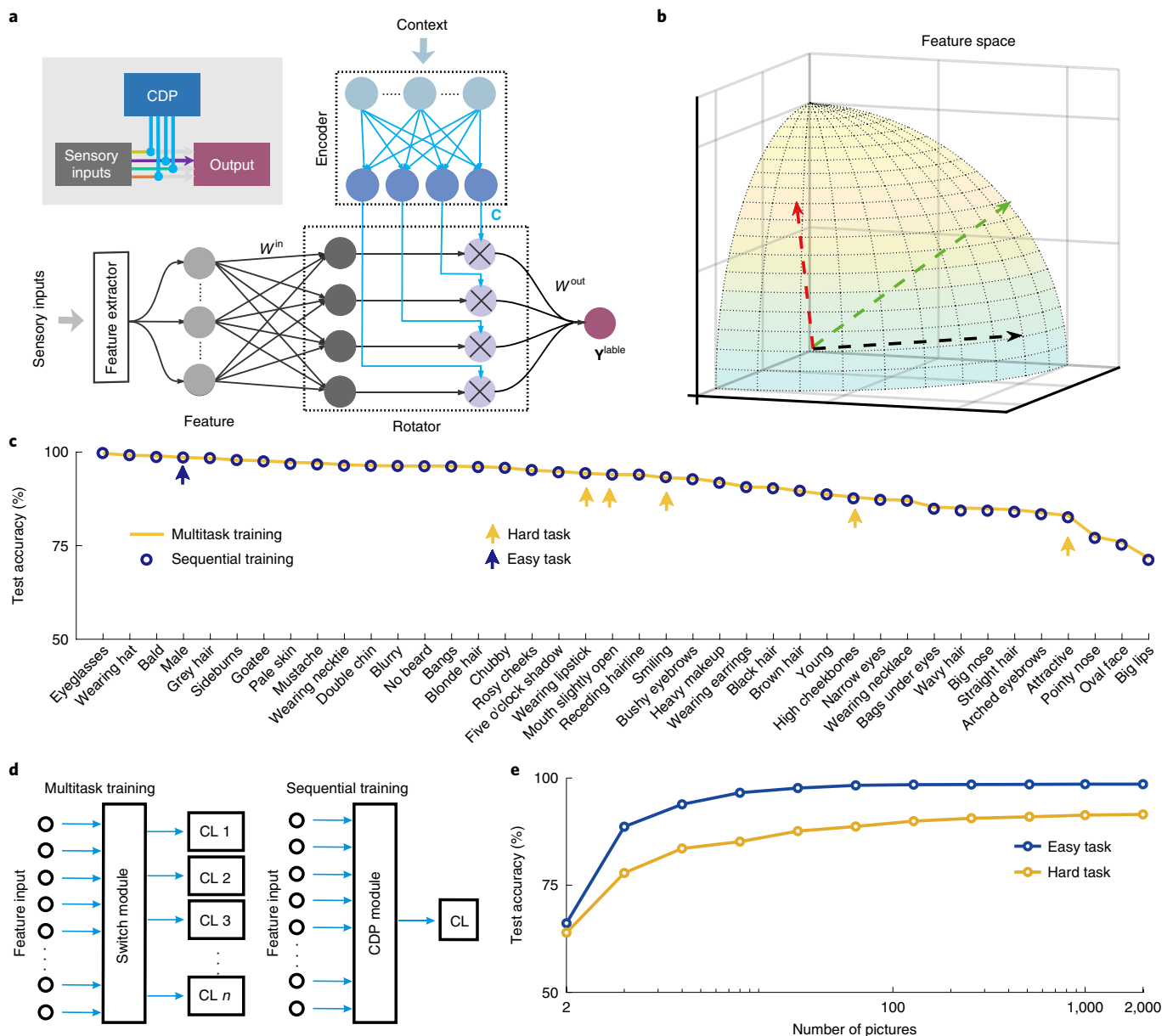


Fig. 4 | Achieving context-dependent sequential learning via the OWM algorithm and the CDP module. **a**, Schematic of network architecture. The CDP module dynamically modulates the mapping between sensory inputs to network outputs according to the contextual information. The main figure and the inset illustrate the detailed internal structure of the module and the overall architecture, respectively. **b**, Schematic showing the role of the CDP module in rotating inputs in feature space (see Methods for details), with the same feature vector rotated differently (colour-coded) in different contexts. **c**, Performance of sequentially learning to classify faces by 40 different attributes, each associated with a unique contextual signal, compared with results obtained by multitask training. Tasks were sorted by test accuracy. **d**, Schematics showing network architecture for multitask (left) and sequential (right) training. CL, classifier. To achieve context-dependent processing, in multitask training a switch module and n classifiers are needed, where n is the number of different attributes. **e**, Classification accuracies for a relatively easy task (gender; blue curve) and five more difficult, sequentially learned tasks (for example, attractiveness; orange curve; mean results across all five tasks are shown) are plotted as a function of training sample size. Tasks and corresponding performance obtained by training on the full dataset are marked with arrows in **c**.

learning³⁹, which inspired the development of algorithms to protect the important weights involved in previously learned tasks while training the network for new ones^{24,26}. However, these ‘frozen’ weights necessarily reduce the degrees of freedom of the system, that is, they decrease the volume of parameter space to search for a configuration that can satisfy both old and new tasks. Here, by allowing the ‘frozen’ weights to be adjustable again without erasing acquired knowledge, the OWM exhibited clear advantages in performance. However, further studies are required to investigate whether algorithms similar to the OWM are implemented in the brain.

Recently, it has been suggested that a variant of the backpropagation algorithm, that is, the ‘conceptor-aided backprop’ (CAB) can be used for continual learning by shielding gradients against degradation of previously learned tasks²². By providing more effective shielding of gradients through constructing an orthogonal projector, the OWM achieved much better protection of previously acquired knowledge, yielding highly competitive results in empirical tests compared with the CAB (see Tables 1 and 2, Fig. 2 and Supplementary Information for details). The OWM and continual learning methods mentioned above are regularization approaches¹⁵.

Similar to other methods within this category, the OWM exhibits a trade-off between the performance of the old and new tasks, due to limited resources to consolidate the knowledge of previous tasks. In contrast to regularization approaches, other types of continual learning methods involve dynamically introducing extra neurons or layers along the learning process⁴⁰, which may help mitigate the trade-off described above¹⁵. However, regularization approaches require no extra resources to accommodate newly acquired knowledge during training and, therefore, are capable of producing compact yet versatile systems.

We note that a solution for continual learning based on context-dependent processing was suggested recently⁴¹. In this work, a context-dependent gating mechanism was used to separate sub-networks for processing individual tasks during continual learning. However, for this approach to work, the same contextual information needs to be present during both the training and testing phases. As such information is rarely available in practical situations, this seriously limits the applicability of the method. Different from this approach, the CDP module in our work enables the network to modulate its processing according to the contextual information so that the same inputs can be treated differently in different contexts. This role is not related to continual learning, as the CDP module is needed in the same task as shown in Fig. 4, even if the system was trained concurrently. Importantly, as contextual information, for example, environmental cues, the task at hand, and so on, is always available for any input that needs CDP, the limitations of using context-dependent gating for continual learning is not a problem for our CDP module.

Other biologically inspired approaches for continual learning are based on complementary learning systems (CLS) theory^{42,43}. Such systems involve the interplay between two subsystems similar to the mammalian hippocampus and neocortex, that is, a task-solving network (neocortex) accompanied by a generative network (hippocampus) to maintain the memories of previous tasks⁴⁴. With the aid of the learning without forgetting method⁴⁵, data for old tasks sampled by the generative module are interleaved with those for the current task to train the neural network to avoid catastrophic forgetting. Although here we used a completely different approach for continual learning, the CLS framework may also be instrumental for further development of our approach. Currently, the encoder of the CDP module has the ability to infer contextual information from the environment and also to learn how to use it effectively. Conceivably it could be further developed to recognize and classify complex contexts. Such a flexible module for recognizing proper contextual signals may be analogous to the hippocampus in the brain, which is related to the classification of different environmental cues via pattern separation and completion⁴⁶. Thus, it would be informative for future studies to investigate whether the current approach can be combined with the CLS framework to achieve more flexible and sophisticated context-dependent processing.

Taken together, our study demonstrated that it is possible to teach a highly compact network many context-dependent mappings sequentially. Although we demonstrated its effectiveness here with the supervised learning paradigm, the OWM has the potential to be applied to other training frameworks. Another regularization approach for overcoming catastrophic forgetting, that is, the elastic weight consolidation (EWC), has been successfully implemented in reinforcement learning²⁴. As the EWC can be viewed as a special case of OWM in some circumstances (see Supplementary Information for details), it suggests that similar procedures could be extended for the use of the OWM and CDP module in unsupervised conditions, thereby enabling networks to learn different mapping rules for different contexts by reinforcement learning. We expect that such an approach, combined with effective methods of knowledge transfer^{46–49}, may eventually lead to systems with sufficient flexibility to work in complex and dynamic situations.

Methods

The OWM algorithm. Consider a feed-forward network of $L + 1$ layers, indexed by $l = 0, 1, \dots, L$ with $l = 0$ and $l = L$ being the input and output layers, respectively. All hidden layers share the same activation function $g(\bullet)$. W_l represents the connections between the $(l - 1)$ th and l th layer with $W_l \in \mathbb{R}^{s \times m}$. \mathbf{x}_l and \mathbf{y}_l denote the output and input of the l th layer, respectively, where $\mathbf{x}_l = g(\mathbf{y}_l)$ and $\mathbf{y}_l = W_l^T \mathbf{x}_{l-1}$. $\mathbf{x}_{l-1} \in \mathbb{R}^s$ and $\mathbf{y}_l \in \mathbb{R}^m$, where s and m are the dimension of input and output, respectively.

In the OWM, the orthogonal projector P_l defined in the input space of layer l for learned tasks is key for overcoming catastrophic interference in sequential learning. In practice, P_l can be recursively updated for each task in a way similar to calculating the correlation-inverse matrix $P^{(\text{RLS})} = (\sum_{i=1}^n \mathbf{x}(i)\mathbf{x}^T(i) + \alpha I)^{-1}$ in the RLS algorithm^{16,18,19} (see discussion on relationship of OWM and RLS in Supplementary Information). This method allows P_l to be determined based on the current inputs and the P_l for the last task. It also avoids matrix-inverse operation in the original definition of P_l .

Below we provide the detailed procedure for the implementation of the OWM method.

- (1) Initialization of parameters: randomly initialize $W_l(0)$ and set $P_l(0) = I_l/\beta$ for $l = 1, \dots, L$, where β is a regularization constant.
- (2) Forward propagate the inputs of the i th batch in the j th task, then back propagate the errors and calculate weight modifications $\Delta W_l^{\text{BP}}(i, j)$ for $W_l(i - 1, j)$ by the standard BP method.
- (3) Update the weight matrix in each layer by

$$\begin{aligned} W_l(i, j) &= W_l(i - 1, j) + \kappa(i, j) \Delta W_l^{\text{BP}}(i, j) & \text{if } j = 1 \\ W_l(i, j) &= W_l(i - 1, j) + \kappa(i, j) P_l(j - 1) \Delta W_l^{\text{BP}}(i, j) & \text{if } j = 2, 3, \dots \end{aligned} \quad (1)$$

where $\kappa(i, j)$ is the predefined learning rate.

- (4) Repeat steps 2 to 3 for the next batch.
- (5) If the j th task is accomplished, forward propagate the mean of the inputs for each batch ($i = 1, \dots, n_j$) in the j th task successively. Update P_l for W_l as $P_l(j) = P_l(n_j, j)$, where $P_l(j) = P_l(n_j, j)$ can be calculated iteratively according to:

$$\begin{aligned} P_l(i, j) &= P_l(i - 1, j) - \mathbf{k}_l(i, j) \bar{\mathbf{x}}_{l-1}(i, j)^T P_l(i - 1, j) \\ \mathbf{k}_l(i, j) &= P_l(i - 1, j) \bar{\mathbf{x}}_{l-1}(i, j) / [\alpha + \bar{\mathbf{x}}_{l-1}(i, j)^T P_l(i - 1, j) \bar{\mathbf{x}}_{l-1}(i, j)] \end{aligned} \quad (2)$$

in which $\bar{\mathbf{x}}_{l-1}(i)$ is the output of the $l - 1$ th layer in response to the mean of the inputs in the i th batch of the j th task, and $P_l(0, j) = P_l(j - 1)$.

- (6) Repeat steps 2 to 5 for the next task.

We note that the algorithm achieved the same performance if the orthogonal projector P_l was updated for each batch according to equation (2), with α decaying as $\alpha_{ij} = \alpha_0 \lambda^{i/n_j}$ for the i th batch of data in the j th task. This method can be understood as treating each batch as a different task. It avoids the extra storage space as well as data reloading in step 4 and, therefore, significantly accelerates processing. In this case, if the learning rate is set to $\kappa(i) = 1/[1 + \bar{\mathbf{x}}_{l-1}(i)^T P_l(i - 1) \bar{\mathbf{x}}_{l-1}(i)]$ and α_{ij} is permanently set to $\alpha_{ij} = 1$, the procedure essentially uses RLS to train the neural network under the name of enhanced backpropagation, which is proposed to increase the speed of convergence in training¹⁹. Therefore, our algorithm has the same computational complexity as enhanced backpropagation— $O(N_n N_w^2)$, where N_n is the total number of neurons and N_w is the number of input weights per neuron¹⁹.

In addition, we analysed the capacity of the OWM, that is, how many different tasks could be learned using this method. The capacity of one network layer can be measured by the rank of $P(i)$, which is defined as the orthogonal projector calculated after task i , with $\Delta P(i + 1)$ then defined as the update in the next task satisfying $P(i + 1) = P(i) - \Delta P(i + 1)$. As $\text{range}(P(i + 1)) \cap \text{range}(\Delta P(i + 1)) = \emptyset$, $\text{rank}(P(i + 1)) = \text{rank}(P(i)) - \text{rank}(\Delta P(i + 1))$. In the ideal case where each task consumes the capacity effectively, as the learning process continues, the rank of P_l is approaching 0, indicating that this particular layer no longer has the capacity to learn new tasks. The capacity of the whole network can be approximated by the summation of the capacity of each layer: $\text{rank}_{\text{tot}} = \sum_{l=1}^L \text{rank}(P_l) / \text{rank}(\beta I_l)$ where βI_l is the initial value of matrix P_l . The rank is normalized to balance the contribution of each layer. We conducted two experiments (Fig. 3d,e) on the CASIA-HWDB1.1 dataset to verify the above analysis. In the experiments, to avoid influence by the tolerance value in the calculation of matrix rank, the rank was estimated as $\text{rank}(P) = \sum_{i=1}^s s_i(P) / \beta$, where $s_i(\bullet)$ denotes the i th singular value of the matrix. If the capacity limit of the entire network is finally approached, two solutions can be considered: (1) introduction of a larger α or the forgetting factor used in RLS¹⁶ and online EWC⁴⁹; and (2) addition of more layer(s), for example, CDP module (see below for details), to provide more space to preserve previously learned knowledge.

The CDP module. In context-dependent learning, to change the representation of sensory inputs without distorting information content in different contexts, we added one layer of neurons after the output layer of the feature extractor (cf. Fig. 4a). Below we describe, from a mathematical point of view, how this CDP layer works, using the face classification task as an example.

In this task, the rotator submodule was fed with feature vectors for different faces, $\mathbf{F} = [f_1, f_2, \dots, f_k]^T \in \mathbb{R}^k$, and modulated by non-negative controlling signals, $\mathbf{C} = [c_1, c_2, \dots, c_m]^T \in \mathbb{R}^m$. The controlling signals \mathbf{C} were drawn from the contextual information (word vector of corresponding task name) by the encoder submodule. Then the CDP module outputted $\mathbf{Y}^{\text{out}} = [y_1, y_2, \dots, y_m]^T \in \mathbb{R}^m$, with $y_i = g((\mathbf{w}_i^{\text{in}})^T \mathbf{F})$, to a classifier for further processing. The input weight $\mathbf{W}^{\text{in}} = [\mathbf{w}_1^{\text{in}}, \mathbf{w}_2^{\text{in}}, \dots, \mathbf{w}_m^{\text{in}}] \in \mathbb{R}^{k \times m}$ of the CDP module was randomly initialized and fixed across all contexts. The rest weights in the CDP module, including the output weight \mathbf{W}^{out} and the weights in the encoder, were trained by the OWM method. The function of the CDP module can then be summarized as

$$\begin{aligned} \mathbf{Y}^{\text{out}} &= g((\mathbf{W}^{\text{in}})^T \mathbf{F}) \odot \mathbf{C} \\ &= g\left(\begin{bmatrix} \mathbf{w}_1^{\text{in}} & \mathbf{w}_2^{\text{in}} & \dots & \mathbf{w}_m^{\text{in}} \end{bmatrix}^T \mathbf{F}\right) \odot \mathbf{C} \\ &= g\left(\begin{bmatrix} c_1 \|\mathbf{F}\| \|\mathbf{w}_1^{\text{in}}\| \cos\theta_1 & c_2 \|\mathbf{F}\| \|\mathbf{w}_2^{\text{in}}\| \cos\theta_2 & \dots & c_m \|\mathbf{F}\| \|\mathbf{w}_m^{\text{in}}\| \cos\theta_m \end{bmatrix}^T\right) \\ &= g\left(\begin{bmatrix} c_1 \|\mathbf{w}_1^{\text{in}}\| \cos\theta_1 & c_2 \|\mathbf{w}_2^{\text{in}}\| \cos\theta_2 & \dots & c_m \|\mathbf{w}_m^{\text{in}}\| \cos\theta_m \end{bmatrix}^T\right) \|\mathbf{F}\| \end{aligned} \quad (3)$$

where \odot represents element-wise multiplication and θ_i is the angle between \mathbf{w}_i^{in} and \mathbf{F} . Note that for any $v \geq 0$, $g(vx) = \max(0, vx) = v \max(0, x) = vg(x)$. The ReLU function was used in the current study for $g(\bullet)$ but this is not necessary. $g(\bullet)$ can also be chosen as a hyperbolic tangent function or logistic function. As \mathbf{W}^{in} was initialized by the Xavier method⁵⁰ in most cases, $\mathbf{w}_i^{\text{in}} \mathbf{F}$ was located in the linear range. Thus the equation (3) can approximately hold even for activation functions other than ReLU. We confirmed that the average accuracies for the same tasks in Fig. 4c with hyperbolic tangent function (90.93%) and logistic function (90.05%) were close to that with ReLU (90.38%); a 3,000-neuron rotator layer was used for all three activation functions).

For individual faces, given the same feature vector \mathbf{F} and fixed \mathbf{W}^{in} , $\cos\theta_i$ is constant. Thus, output \mathbf{Y}^{out} is affected by the controlling signal \mathbf{C} , which is different across tasks. If we normalize \mathbf{C} by $\sqrt{\sum_{i=1}^m (c_i \|\mathbf{w}_i^{\text{in}}\| g(\cos\theta_i))^2}$, it is apparent from equation (3) that the CDP layer 'rotates' the input vector in feature space, as illustrated in Fig. 4b. This explains why this added layer can change the representation of sensory inputs while keeping information contents unchanged. Importantly, it also enables the system to sequentially learn different tasks with the OWM for identical inputs.

To examine whether the CDP module can infer the correct context from distracting noise in the environment, four face recognition tasks were conducted continually with the OWM, except that the explicit context signal was not presented. The context signals and distracting noises were simultaneously fed to the CDP module. The noise was sampled from a Gaussian distribution with the same mean and variance as the context signal, and varied on a trial-by-trial basis. In the training phase for different tasks, the position of the context signal and noises could be swapped (Supplementary Fig. 3). During the testing phase, either the corresponding context + noises or only noises were presented.

Shuffled MNIST experiment. The shuffled MNIST experiment^{14,22,24–26} usually consists of a number of sequential tasks. All tasks involve classifying handwritten digits from 0 to 9. However, for each new task, the pixels in the image are randomly shuffled, with the same randomization across all digits in the same task and different randomization across tasks. For this experiment, we trained three- or four-layer, feed-forward networks with [784–800–10] (three-layer) or [784–800/2,000–800/2,000–10] (four-layer) neurons (see Table 1 for details) to minimize cross-entropy loss by the OWM method. The ReLU activation function⁵¹ was used in the hidden layer. During training, the L2 regularization coefficient was 0.001. Dropout was applied with a drop rate of 0.2.

Table 1 shows the performance of the OWM method for the shuffled MNIST tasks compared with other continual learning algorithms. The accuracy of the OWM method was measured by repeating the experiments ten times. The results of other algorithms were adopted from corresponding publications. The size of the network, regarding the number of layers and number of neurons in each layer, was the same as in previous publications for a fair comparison.

Two-sided *t*-tests were used to compare performance between the OWM and other continual learning methods for both the shuffled and disjoint (see below) MNIST experiments. The *t* values were calculated according to the means and standard deviations across ten experiments. Significance was considered at $P < 0.01$ with results shown in Table 1.

Disjoint MNIST experiment. In the two-disjoint MNIST experiment⁵², the original MNIST dataset was divided into two parts. The first contained digits from 0 to 4 and the second consisted of digits from 5 to 9. Correspondingly, the first task was to recognize digits among 0, 1, 2, 3 and 4 and the second task was to recognize digits among 5, 6, 7, 8, and 9. In the ten-disjoint MNIST task, 10 digits, from 0 to 9, were learned sequentially. Again, to facilitate comparison, network size and architecture were the same as in previous work⁵². During training, momentum optimization was applied, and the learning rate for all layers remained the same

during training. Performance was calculated based on ten repeated experiments and is shown in Table 2.

Sequential learning of classification tasks with Chinese characters and ImageNet. Classification tasks with ImageNet and Chinese handwritten characters are more challenging due to the complex structure in each image and more classes to 'memorize' in a sequential learning task. In sequential learning, the training for a new task started only when the neural network accomplished the current task well enough, defined here as <1% accuracy gap between two successive training epochs. For these two tasks, we first trained a DNN as the feature extractor on the whole or partial dataset to extract features of each image. The extracted feature vectors were then fed into a three-layer classifier with [1,024–4,000–3,755] neurons for the Chinese characters task and [2,048–4,000–1,000] neurons for the ImageNet task. The classifier was trained to recognize each of the classes sequentially using the OWM method, with results shown in Supplementary Table 1. We note that in these experiments, as in other tests mentioned above, no negative samples were used for training the network to recognize a new class. In other words, only positive samples of a particular class were presented to the network during training.

Disjoint CIFAR-10 experiment. In contrast to the pre-training of feature extractors in the tasks of Chinese characters and ImageNet, the feature extractor was trained together with the classifier in an end-to-end way using the OWM in this task. The CIFAR-10 dataset was divided into five groups. Each group included two classes of samples used to train the whole network in one task. The feature extractor consisted of three convolutional layers and the classifier consisted of three fully connected layers. The three convolutional layers had 64, 128 and 256 filters, respectively, and the size of the convolution kernel was 2×2 . A maxpooling layer with size of 2×2 was attached to each convolutional layer. Dropout was applied to each maxpooling layer with a dropping probability of 0.2. The features extracted were flattened and then fed to the classifier of [1,000–1,000–10] neurons. The activation function for all layers was the ReLU function. The initial weights for all layers were in accordance with the Xavier initialization method proposed by Glorot and Bengio⁵⁰. Cross-entropy loss was applied for the training. Table 3 compares the performance of the OWM with other methods using the same network structure and task.

Context-dependent face recognition with CelebA. In the CelebFaces Attributes (CelebA) experiment, we first trained a feature extractor using the architecture of ResNet50⁵³ on the whole training dataset and with the conventional multitask training procedure. The outputs of the feature extractor were then fed into the CDP module, which also received contextual information (cf. Fig. 4a). The rotator layer contained 5,000 neurons. The size of the encoder layer was [200–5,000], with ReLU applied as the activation function. For the face classification task in the present study, rotated feature vectors were fed directly into the classifier by weights \mathbf{W}^{out} . Before training, all weights and biases were randomly initialized. \mathbf{W}^{out} and the weights in the encoder were modified by the OWM method. Detailed results of classifying individual attributes are listed in Supplementary Table 5.

Network parameters. Weights in the hidden layers of the classifiers for the tasks other than disjoint CIFAR-10 task were initialized according to a previously suggested method⁵⁴. The output layers were all initialized to zero. The biases of each layer were randomly initialized according to a uniform distribution within (0, 0.1). The ReLU neurons were applied to every hidden layer in all experiments. The momentum in all optimization algorithms was 0.9. The details of hyperparameters used for feature extractors are shown in Supplementary Table 6. Early stopping was used for training both the feature extractors and classifiers. The hyperparameters for the OWM method are shown in Supplementary Table 7. For tasks with MNIST and CelebA, the classifier was trained to minimize cross-entropy loss, whereas for tasks with ImageNet and Chinese characters, the classifier was trained to minimize mean squared loss. Note that cross-entropy loss was also suitable for the latter datasets. However, mean squared loss is easier to compute and less time-consuming when many tasks are involved.

Mixed selectivity analysis. For classifying different facial attributes, responses of neurons in the CDP were analysed to examine whether they exhibited mixed selectivity similar to that of real PFC neurons. To this end, we chose two attributes with low correlation, that is, Attractiveness (task 1) and Smile (task 2). Both have about 50% positive and 50% negative samples in the whole dataset. The responses of each neuron in the CDP module to different inputs as well as contextual signals were analysed with the weights in the encoder submodule fixed for both contexts. There were 19,962 test pictures, with 90% correctly classified after training for both tasks. The threshold of excitation for each neuron was chosen as the average activity level across all neurons during the processing of all correctly classified pictures. Supplementary Fig. 5 shows the selectivity of three exemplar neurons. According to the criteria usually used in electrophysiological experiments, these three neurons belonged to different categories, including task sensitive (neuron 1) and attribute sensitive (neuron 2). Importantly, neuron 3 exhibited complex selectivity towards combinations of task and sensory attributes, as well as combinations of different attributes. This mixed selectivity is commonly reported for real PFC neurons⁵⁵.

Data availability

All data used in this paper are publicly available and can be accessed at <http://yann.lecun.com/exdb/mnist/> for the MNIST dataset, <https://www.cs.toronto.edu/~kriz/cifar.html> for the CIFAR dataset, <http://image-net.org/index.php> for the ILSVR2012 dataset, <http://www.nlpri.ia.ac.cn/databases/handwriting/Home.html> for the CASIA-HWDB dataset and <http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html> for the CelebA dataset. For more details of the datasets, please refer to the references cited in the Supplementary Methods.

Code availability

The source code can be accessed at <https://github.com/beijixiong3510/OWM>⁵⁶.

Received: 23 December 2018; Accepted: 10 July 2019;

Published online: 9 August 2019

References

- Newell, A. *Unified Theories of Cognition* (Harvard Univ. Press, 1994).
- Miller, G. A., Heise, G. A. & Lichten, W. The intelligibility of speech as a function of the context of the test materials. *J. Exp. Psychol.* **41**, 329–335 (1951).
- Desimone, R. & Duncan, J. Neural mechanisms of selective visual-attention. *Annu. Rev. Neurosci.* **18**, 193–222 (1995).
- Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).
- Siegel, M., Buschman, T. J. & Miller, E. K. Cortical information flow during flexible sensorimotor decisions. *Science* **348**, 1352–1355 (2015).
- Miller, E. K. The prefrontal cortex: complex neural properties for complex behavior. *Neuron* **22**, 15–17 (1999).
- Wise, S. P., Murray, E. A. & Gerfen, C. R. The frontal cortex basal ganglia system in primates. *Crit. Rev. Neurobiol.* **10**, 317–356 (1996).
- Passingham, R. *The Frontal Lobes and Voluntary Action* (Oxford Univ. Press, 1993).
- Miller, E. K. & Cohen, J. D. An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* **24**, 167–202 (2001).
- Miller, E. K. The prefrontal cortex and cognitive control. *Nat. Rev. Neurosci.* **1**, 59–65 (2000).
- LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
- McCloskey, M. & Cohen, N. J. *Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem* Vol. 24 109–165 (Elsevier, 1989).
- Ratcliff, R. Connectionist models of recognition memory—constraints imposed by learning and forgetting functions. *Psychol. Rev.* **97**, 285–308 (1990).
- Goodfellow, I. J., Mirza, M., Xiao, D., Courville, A. & Bengio, Y. An empirical investigation of catastrophic forgetting in gradient-based neural networks. Preprint at <https://arxiv.org/abs/1312.6211> (2013).
- Parisi, G. I., Kemker, R., Part, J. L., Kanan, C. & Wermter, S. Continual lifelong learning with neural networks: a review. *Neural Netw.* **113**, 54–71 (2019).
- Haykin, S. S. *Adaptive Filter theory* (Pearson Education India, 2008).
- Golub, G. H. & Van Loan, C. F. *Matrix Computations* Vol. 3 (JHU Press, 2012).
- Singhal, S. & Wu, L. Training feed-forward networks with the extended kalman algorithm. In *International Conference on Acoustics, Speech, and Signal Processing* 1187–1190 (IEEE, 1989).
- Shah, S., Palmieri, F. & Datum, M. Optimal filtering algorithms for fast learning in feedforward neural networks. *Neural Netw.* **5**, 779–787 (1992).
- Sussillo, D. & Abbott, L. F. Generating coherent patterns of activity from chaotic neural networks. *Neuron* **63**, 544–557 (2009).
- Jaeger, H. Controlling recurrent neural networks by conceptors. Preprint at <https://arxiv.org/abs/1403.3369> (2014).
- He, X. & Jaeger, H. Overcoming catastrophic interference using conceptor-aided backpropagation. In *International Conference on Learning Representations (ICLR)*, 2018).
- Nair, V. & Hinton, G. E. Rectified linear units improve restricted Boltzmann machines. In *International Conference on Machine Learning* 807–814 (PMLR, 2010).
- Kirkpatrick, J. et al. Overcoming catastrophic forgetting in neural networks. *Proc. Natl Acad. Sci. USA* **114**, 3521–3526 (2017).
- Lee, S.-W., Kim, J.-H., Jun, J., Ha, J.-W. & Zhang, B.-T. Overcoming catastrophic forgetting by incremental moment matching. In *Advances in Neural Information Processing Systems* 4652–4662 (Curran Associates, 2017).
- Zenke, F., Poole, B. & Ganguli, S. Continual learning through synaptic intelligence. In *International Conference on Machine Learning* 6072–6082 (PMLR, 2017).
- Liu, C.-L., Yin, F., Wang, D.-H. & Wang, Q.-F. Chinese handwriting recognition contest 2010. In *Chinese Conference on Pattern Recognition (CCPR)* 1–5 (IEEE, 2010).
- Yin, F., Wang, Q.-F., Zhang, X.-Y. & Liu, C.-L. ICDAR 2013 Chinese handwriting recognition competition. In *12th International Conference on Document Analysis and Recognition (ICDAR)* 1464–1470 (IEEE, 2013).
- Fuster, J. *The Prefrontal Cortex* (Academic Press, 2015).
- Liu, Z., Luo, P., Wang, X. & Tang, X. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision* 3730–3738 (IEEE, 2015).
- Řehůřek, R. & Sojka, P. Software framework for topic modelling with large corpora. *Proc. LREC 2010 Workshop on New Challenges for NLP Frameworks* 45–50 (ELRA, 2010).
- Lehky, S. R., Kiani, R., Esteky, H. & Tanaka, K. Dimensionality of object representations in monkey inferotemporal cortex. *Neural Comput.* **26**, 2135–2162 (2014).
- Freedman, D. J., Riesenhuber, M., Poggio, T. & Miller, E. K. Categorical representation of visual stimuli in the primate prefrontal cortex. *Science* **291**, 312–316 (2001).
- Hung, C. P., Kreiman, G., Poggio, T. & DiCarlo, J. J. Fast readout of object identity from macaque inferior temporal cortex. *Science* **310**, 863–866 (2005).
- Kravitz, D. J., Saleem, K. S., Baker, C. I., Ungerleider, L. G. & Mishkin, M. The ventral visual pathway: an expanded neural framework for the processing of object quality. *Trends Cogn. Sci.* **17**, 26–49 (2013).
- Gomez, J. et al. Microstructural proliferation in human cortex is coupled with the development of face processing. *Science* **355**, 68–71 (2017).
- Xu, F. & Tenenbaum, J. B. Word learning as Bayesian inference. *Psychol. Rev.* **114**, 245–272 (2007).
- Rigotti, M. et al. The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).
- Cichon, J. & Gan, W.-B. Branch-specific dendritic Ca^{2+} spikes cause persistent synaptic plasticity. *Nature* **520**, 180–185 (2015).
- Rusu, A. A. et al. Progressive neural networks. Preprint at <https://arxiv.org/abs/1606.04671> (2016).
- Masse, N. Y., Grant, G. D. & Freedman, D. J. Alleviating catastrophic forgetting using context-dependent gating and synaptic stabilization. *Proc. Natl Acad. Sci. USA* **115**, E10467–E10475 (2018).
- McClelland, J. L., McNaughton, B. L. & O'Reilly, R. C. Why there are complementary learning-systems in the hippocampus and neocortex—insights from the successes and failures of connectionist models of learning and memory. *Psychol. Rev.* **102**, 419–457 (1995).
- Kumaran, D., Hassabis, D. & McClelland, J. L. What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends Cogn. Sci.* **20**, 512–534 (2016).
- Shin, H., Lee, J. K., Kim, J. & Kim, J. Continual learning with deep generative replay. In *Advances in Neural Information Processing Systems* 2990–2999 (Curran Associates, 2017).
- Li, Z. & Hoiem, D. Learning without forgetting. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**, 2935–2947 (2017).
- Rohrbach, M., Stark, M., Szarvas, G., Gurevych, I. & Schiele, B. What helps where—and why? Semantic relatedness for knowledge transfer. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 910–917 (IEEE, 2010).
- Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems* 3320–3328 (Curran Associates, 2014).
- Hinton, G., Vinyals, O. & Dean, J. Distilling the knowledge in a neural network. Preprint at <https://arxiv.org/abs/1503.02531> (2015).
- Schwarz, J. et al. Progress & compress: a scalable framework for continual learning. Preprint at <https://arxiv.org/abs/1805.06370> (2018).
- Glorot, X. & Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *Proc. Thirteenth International Conference on Artificial Intelligence and Statistics* 249–256 (Microtome, 2010).
- Nair, V. & Hinton, G. E. Rectified linear units improve restricted boltzmann machines. In *Proc. 27th International Conference on Machine Learning (ICML-10)* 807–814 (PMLR, 2010).
- Srivastava, R. K., Masci, J., Kazerounian, S., Gomez, F. & Schmidhuber, J. Compete to compute. In *Advances in Neural Information Processing Systems* 2310–2318 (Curran Associates, 2013).
- He, K. M., Zhang, X. Y., Ren, S. Q. & Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (IEEE, 2016).
- He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In *IEEE International Conference on Computer Vision* 1026–1034 (IEEE, 2015).
- Ramirez-Cardenas, A. & Viswanathan, P. The role of prefrontal mixed selectivity in cognitive control. *J. Neurosci.* **36**, 9013–9015 (2016).
- Zeng, G., Chen, Y., Cui, B. & Yu, S. Codes for paper Continual learning of context-dependent processing in neural networks. *Zenodo* <https://doi.org/10.5281/zenodo.3346080> (2019).

57. Hu, W. et al. Overcoming catastrophic forgetting via model adaptation. In *International Conference on Learning Representations (ICLR, 2019)*.

Acknowledgements

The authors thank D. Nikolić for helpful discussions and R. Hadsell for comments on the manuscript. This work was supported by the National Key Research and Development Program of China (2017YFA0105203), the Strategic Priority Research Program of the Chinese Academy of Sciences (CAS) (XDB32040200), Key Research Program of the National Laboratory of Pattern Recognition (99S9011M2N), and the Hundred-Talent Program of CAS (for S.Y.).

Author contributions

S.Y., Y.C. and G.Z. conceived the study and designed the experiments. G.Z. and Y.C. conducted computational experiments and theoretical analyses. B.C. assisted with some experiments and analyses. S.Y., Y.C. and G.Z. wrote the paper.

Competing interests

The Institute of Automation, Chinese Academy of Sciences has submitted patent applications on the OWM algorithm (application no. PCT/CN2019/083355; invented by Y.C., G.Z. and S.Y.; pending) and the CDP module (application no. PCT/CN2019/083356; invented by G.Z., Y.C. and S.Y.; pending).

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s42256-019-0080-x>.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to S.Y.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019