nature machine intelligence

Article

https://doi.org/10.1038/s42256-023-00747-w

Incorporating neuro-inspired adaptability for continual learning in artificial intelligence

Received: 3 October 2022

Accepted: 26 September 2023

Published online: 16 November 2023

Check for updates

Liyuan Wang[1,2,3,5], Xingxing Zhang[1,5], Qian Li[2,3], Mingtian Zhang[4], Hang Su[1], Jun Zhu[1] & Yi Zhong[2,3]

Continual learning aims to empower artificial intelligence with strong adaptability to the real world. For this purpose, a desirable solution should properly balance memory stability with learning plasticity, and acquire sufficient compatibility to capture the observed distributions. Existing advances mainly focus on preserving memory stability to overcome catastrophic forgetting, but it remains difficult to flexibly accommodate incremental changes as biological intelligence does. Here, by modelling a robust *Drosophila* learning system that actively regulates forgetting with multiple learning modules, we propose a generic approach that appropriately attenuates old memories in parameter distributions to improve learning plasticity, and accordingly coordinates a multi-learner architecture to ensure solution compatibility. Through extensive theoretical and empirical validation, our approach not only enhances the performance of continual learning, especially over synaptic regularization methods in task-incremental settings, but also potentially advances the understanding of neurological adaptive mechanisms.

Continual learning, also known as lifelong learning, provides the foundation for artificial intelligence (AI) systems to accommodate real-world changes. As the external environment tends to be highly dynamic and unpredictable, an intelligent agent needs to learn and remember throughout its lifetime[1–3]. Numerous efforts have been devoted to preserving memory stability to mitigate catastrophic forgetting in artificial neural networks, where parameter changes for effective learning of a new task usually result in a dramatic performance drop of the old tasks[4–6]. Representative strategies include selectively stabilizing parameters[7–11], recovering old data distributions[12–14], allocating dedicated parameter subspaces[15,16] and so on. However, they usually achieve only modest improvements in specific scenarios, with effectiveness varying widely across experimental settings (such as differences in task type and similarity, input size, number of training samples and so on)[2,3,17]. As a result, there remains a huge gap between existing advances and realistic applications.

To overcome this limitation, we theoretically analyse the key factors on which continual learning performance depends, suggesting a broader objective beyond the current focus. Specifically, to perform well on all tasks ever seen, a desirable solution should properly balance memory stability of old tasks with learning plasticity of new tasks, while being adequately compatible to capture their distributions (see Methods for details). For example, if you want to accommodate a sequence of cakes (that is, incremental tasks) into a bag (that is, a solution), you should optimize the efficiency of space allocation for each cake as well as the total space of the bag, rather than simply freezing the old cakes.

As biological learning systems are natural continual learners that show strong adaptability to real-world changes[2,3,18], we argue that they have been equipped with effective strategies to address the above challenges. In particular, the γ subset of the *Drosophila* mushroom body (γMB) is a biological learning system that is essential for coping with different tasks in succession and enjoys relatively

[1]Department of Computer Science and Technology, Institute for AI, BNRist Center, Tsinghua-Bosch Joint ML Center, THBI Lab, Tsinghua University, Beijing, China. [2]School of Life Sciences, IDG/McGovern Institute for Brain Research, Tsinghua University, Beijing, China. [3]Tsinghua-Peking Center for Life Sciences, Beijing, China. [4]Centre for Artificial Intelligence, University College London, London, UK. [5]These authors contributed equally: Liyuan Wang, Xingxing Zhang. e-mail: dcszj@tsinghua.edu.cn; zhongyithu@tsinghua.edu.cn

https://doi.org/10.1038/s42256-023-00747-w

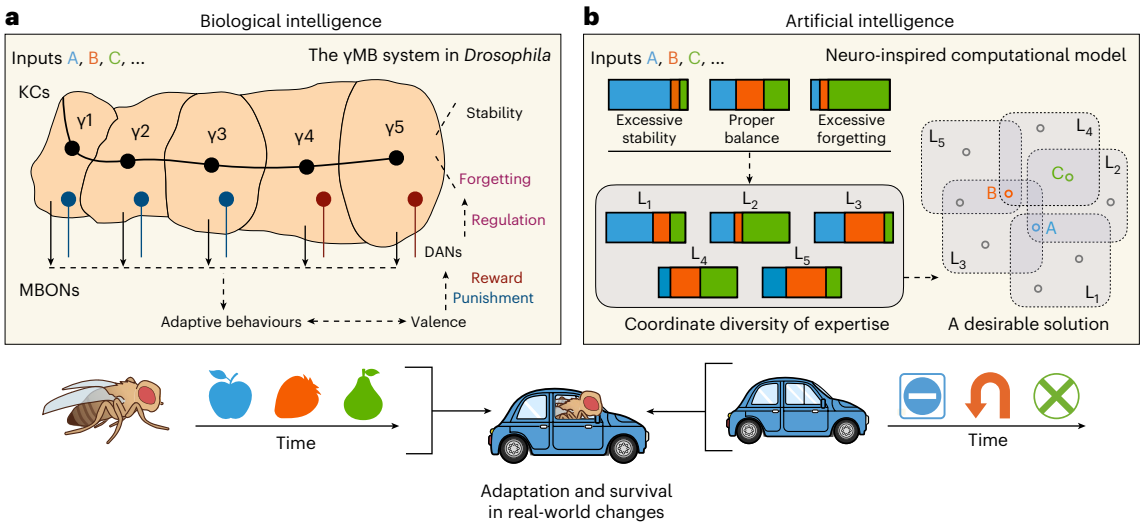

Fig. 1 | Continual learning with reference to a biological learning system. **a**, The *Drosophila* γMB system has evolved adaptive mechanisms to cope with different tasks in succession, such as selective stabilization of synaptic changes, active regulation of memory decay (that is, active forgetting) and dynamic coordination of multiple parallel compartments (that is, γ1–γ5). KCs, Kenyon cells; DANs, dopaminergic neurons; MBONs, mushroom body output neurons. **b**, Inspired by such biological strategies, we propose to incorporate active forgetting together with stability protection for a better trade-off between new and old tasks, and accordingly coordinate multiple parallel continual learners to ensure solution compatibility. L_1–L_5, five continual learners corresponding to the five compartments. The dashed areas on the lower right denote the target distributions of L_1–L_5, and the small hollow circles represent the optimal solution for each incremental task (tasks A, B and C are coloured, and other tasks are grey). The schematic under panels **a** and **b** represents the connection between AI and BI.

clear and in-depth understanding at both functional and anatomical levels[19–26] (Fig. 1a), which emerges as an excellent source for inspiring continual learning in AI.

As a key functional advantage, the γMB system can regulate memories in distinct ways to optimize memory-guided behaviours in changing environments[19,25,27–31]. First, old memories are actively protected from new disruption by strengthening the previously learned synaptic changes[29,31]. This idea of selectively stabilizing parameters has been widely used to alleviate catastrophic forgetting in continual learning[7–10]. Besides, old memories can be actively forgotten for better adapting to a new memory[19,25,28,32]. There are specialized molecular signals to regulate the speed of memory decay[30,33], whose activation reduces the persistence of outdated information, while inhibition shows the opposite effect[19,25,28,32]. However, the benefits of active forgetting for continual learning remain to be explored[3]. Here we propose a functional strategy that incorporates active forgetting together with stability protection for a better trade-off between new and old tasks, where the active forgetting part is formulated as appropriately attenuating old memories in parameter distributions and optimized by a synaptic expansion-renormalization process. Without compromising old tasks, our proposal can greatly enhance the performance of new tasks by eliminating the past conflicting information.

We further explore the organizing principles of the γMB system that support its function, which employs five compartments (γ1–γ5) with dynamic modulations to perform continual learning in parallel[20–24]. As shown in Fig. 1a, the sensory information is incrementally input from Kenyon cells, while the valence is conveyed by dopaminergic neurons[21,34,35]. The outputs of these compartments are carried by distinct MB output neurons and integrated in a weighted-sum fashion to guide adaptive behaviours[22,34,35]. In particular, the dopaminergic neurons allow for distinct learning rules and forgetting rates in each compartment, where the latter has been shown important for processing sequential conflicting experiences[24,36–39]. Inspired by this, we design a specialized architecture of multiple parallel learning modules, which can ensure solution compatibility for incremental changes by coordinating the diversity of learners' expertise. Interestingly, adaptive implementations of the proposed functional strategy can naturally serve this purpose through adjusting the target distribution of each learner, suggesting that the neurological adaptive mechanisms are highly synergistic rather than operating in isolation.

Through satisfying the identified criteria, our approach shows superior generality across various continual learning benchmarks and achieves strong performance gains. We further cross-validate the computational model with biological findings, to better understand the underpinnings of real-world adaptability for both AI and biological intelligence (BI).

Results

Active forgetting with stability protection

A central challenge of continual learning is to resolve the mutual interference between new and old tasks due to their distribution differences. The functional advantages of the γMB system suggest that stability protection and active forgetting are both important[19,25,27–31] (Fig. 1b), although current efforts mainly focus on the former to prevent catastrophic forgetting[4,5]. Here we formulate this process with the framework of Bayesian learning, which has been hypothesized to well model biological synaptic plasticity by tracking the probability distribution of synaptic weights under dynamic sensory inputs[40,41]. We briefly describe a simple case of two tasks (Fig. 2a) and leave the full details to Methods.

Let's consider a neural network with parameters θ continually learning tasks A and B from their training data D_A and D_B to perform well on their test data, which is called a 'continual learner'. From a Bayesian perspective, the learner first places a prior distribution $p(\theta)$ on θ. After learning task A, the learner updates the belief of the parameters, resulting in a posterior distribution $p(\theta|D_A) \propto p(D_A|\theta)p(\theta)$ that incorporates the knowledge of task A. Then, task A can be performed successfully by finding a mode of the posterior: $\theta^*_A = \arg\max_\theta \log p(\theta|D_A)$. For learning task B, $p(\theta|D_A)$ becomes the prior and the posterior $p(\theta|D_A, D_B) \propto p(D_B|\theta)p(\theta|D_A)$ will further incorporate the knowledge of task B. Similarly, the learner needs to find $\theta^*_{A,B} = \arg\max_\theta \log p(\theta|D_A, D_B)$, corresponding to maximizing both $\log p(D_B|\theta)$ for learning task B and $\log p(\theta|D_A)$ for remembering task A.

https://doi.org/10.1038/s42256-023-00747-w

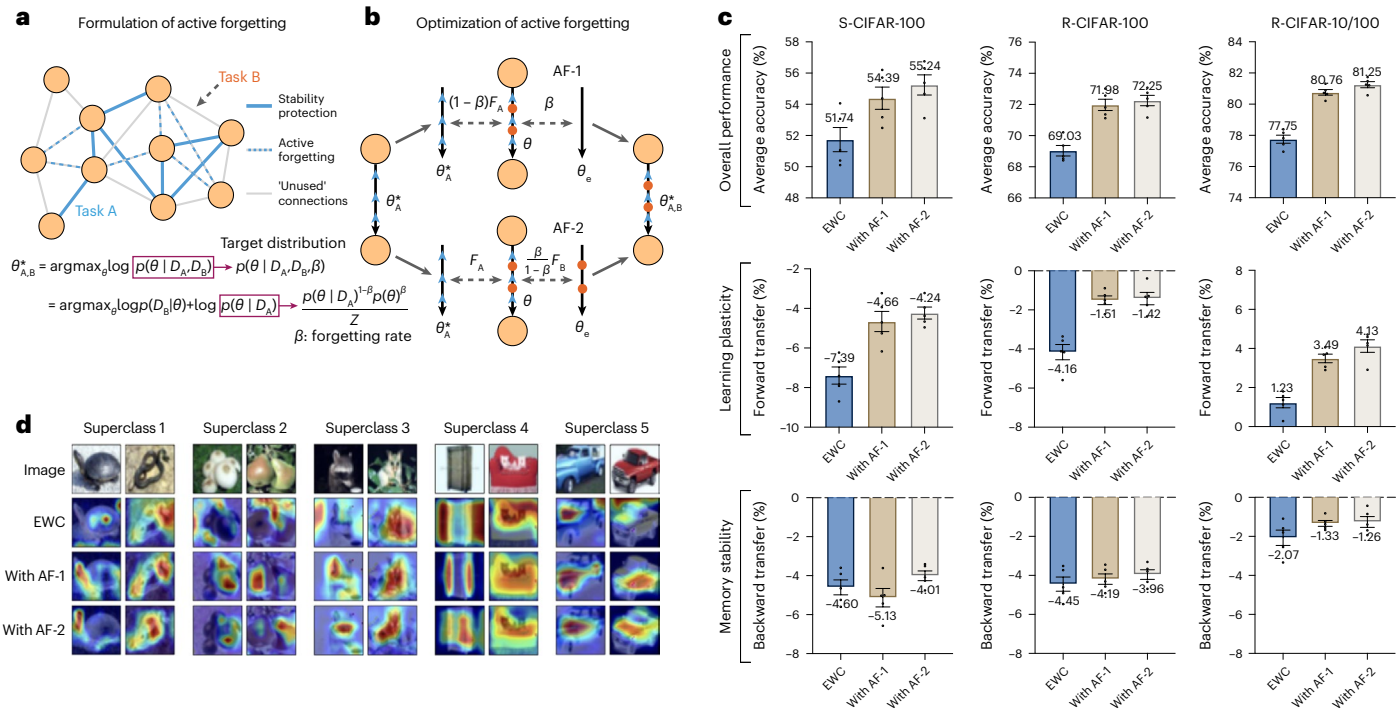

Fig. 2 | Implementation of active forgetting in a continual learning model. **a**, Formulating active forgetting together with stability protection through a Bayesian learning framework. **b**, The proposed functional strategy can be optimized in two equivalent ways of synaptic expansion-renormalization (AF-1 and AF-2), where the network parameters θ need to be selectively renormalized with both θ^*_A and θ_e to balance new and old tasks mutually in a shared solution. **c**, Experimental results. The evaluation metrics include average accuracy for overall performance (top), forward transfer for learning plasticity (middle) and backward transfer for memory stability (bottom). Because of different construction principles, the overall knowledge transfer ranges from more negative to more positive across S-CIFAR-100, R-CIFAR-100 and R-CIFAR-10/100. All results are averaged over five runs with different random seeds and task orders. The error bars represent the standard error of the mean. **d**, Visualization of the latest task predictions on S-CIFAR-100 with Grad-CAM[75].

However, due to the differences in data distribution, remembering old tasks precisely can increase the difficulty of learning each new task well. Inspired by the biological active forgetting, we introduce a forgetting rate β and replace $p(\theta|D_A)$ with

$$\hat{p}(\theta|D_A,\beta)=\frac{p(\theta|D_A)^{(1-\beta)}p(\theta)^{\beta}}{Z}, \tag{1}$$

where $p(\theta)$ is a non-informative prior without incorporating old knowledge[42]. Z is a β-dependent normalizer that keeps $\hat{p}$ a normalized probability distribution (Supplementary Section 1.1). $\hat{p}$ tends to forget task A when $\beta\to 1$, while be dominated by $p(\theta|D_A)$ with full old knowledge when $\beta\to 0$. For the new target $p(\theta|D_A,D_B,\beta)\propto p(D_B|\theta)\hat{p}(\theta|D_A,\beta)$, we derive the loss function

$$\mathcal{L}^{AF}_{Reg}(\theta)=\mathcal{L}_B(\theta)+\underbrace{\frac{\lambda_{SP}}{2}\sum_m F_{A,m}(\theta_m-\theta^*_{A,m})^2}_{\text{stability protection}}+\underbrace{\frac{\lambda_{AF}}{2}\sum_m I_{e,m}(\theta_m-\theta_{e,m})^2}_{\text{active forgetting}}. \tag{2}$$

$\mathcal{L}_B(\theta)$ is the loss function of learning task B, and m denotes the index of parameters. λ_{SP} and λ_{AF} are hyperparameters that control the strengths of two regularizers responsible for stability protection and active forgetting, respectively. The stability protection part is to selectively penalize the deviance of each parameter θ_m from $\theta^*_{A,m}$ depending on its 'importance' for old task(s), estimated by the Fisher information $F_{A,m}$.

The optimization of active forgetting can be achieved in two equivalent ways, that is, AF-1 and AF-2 (Fig. 2b). They both encourage the network parameters θ to renormalize with an 'expanded' set of parameters θ_e when learning task B. For AF-1, $\theta_{e,m}=0$ is 'empty' with equal selectivity $I_{e,m}=1$ for renormalization, where the active-forgetting term becomes the L2 norm of θ. The hyperparameters $\lambda_{AF}\propto\beta$ and $\lambda_{SP}\propto 1-\beta$, indicating that the old memories are directly affected. For AF-2, $\theta_{e,m}=\theta^*_{B,m}$ is the optimal solution for task B only, obtained from optimizing $\mathcal{L}_B(\theta_e)$, and $I_{e,m}=F_{B,m}$ is the Fisher information. The forgetting rate is fully integrated into $\lambda_{AF}\propto\beta/(1-\beta)$ and is independent of λ_{SP}. In the absence of active forgetting ($\beta=0$), the loss function in equation (2) is left with only $\mathcal{L}_B(\theta)$ and the stability protection term, which is (approximately[43]) equivalent to regular synaptic regularization methods such as elastic weight consolidation (EWC[7]). In particular, as the loss functions of these methods[7–10] typically have a similar form and differ only in the metric for estimating the parameter importance[43] (equation (11) in Methods), our proposal can be naturally combined with them by plugging in the active-forgetting term.

For biological neural networks, active forgetting is able to remove outdated information and provide flexibility for adapting to a new memory[19,27,28]. This strategy is essential for *Drosophila* to cope with the interference of previous tasks[19,28,29,44]. Here we theoretically analyse how this benefit is achieved in our computational model. First, an appropriate forgetting rate β is able to improve the probability of learning each new task well through attenuating old memories in θ (equation (5) in Methods), which can be empirically determined by a grid search of λ_{AF} and/or λ_{SP}. Second, when θ moves to the neighbourhood of an empirical optimal solution, the active-forgetting term in equation (2) can minimize the upper bound of generalization errors for continual learning, especially for new tasks (Proposition 2 in Methods).

Now we evaluate the efficacy of active forgetting on three continual learning benchmarks for visual classification tasks. They are all

https://doi.org/10.1038/s42256-023-00747-w

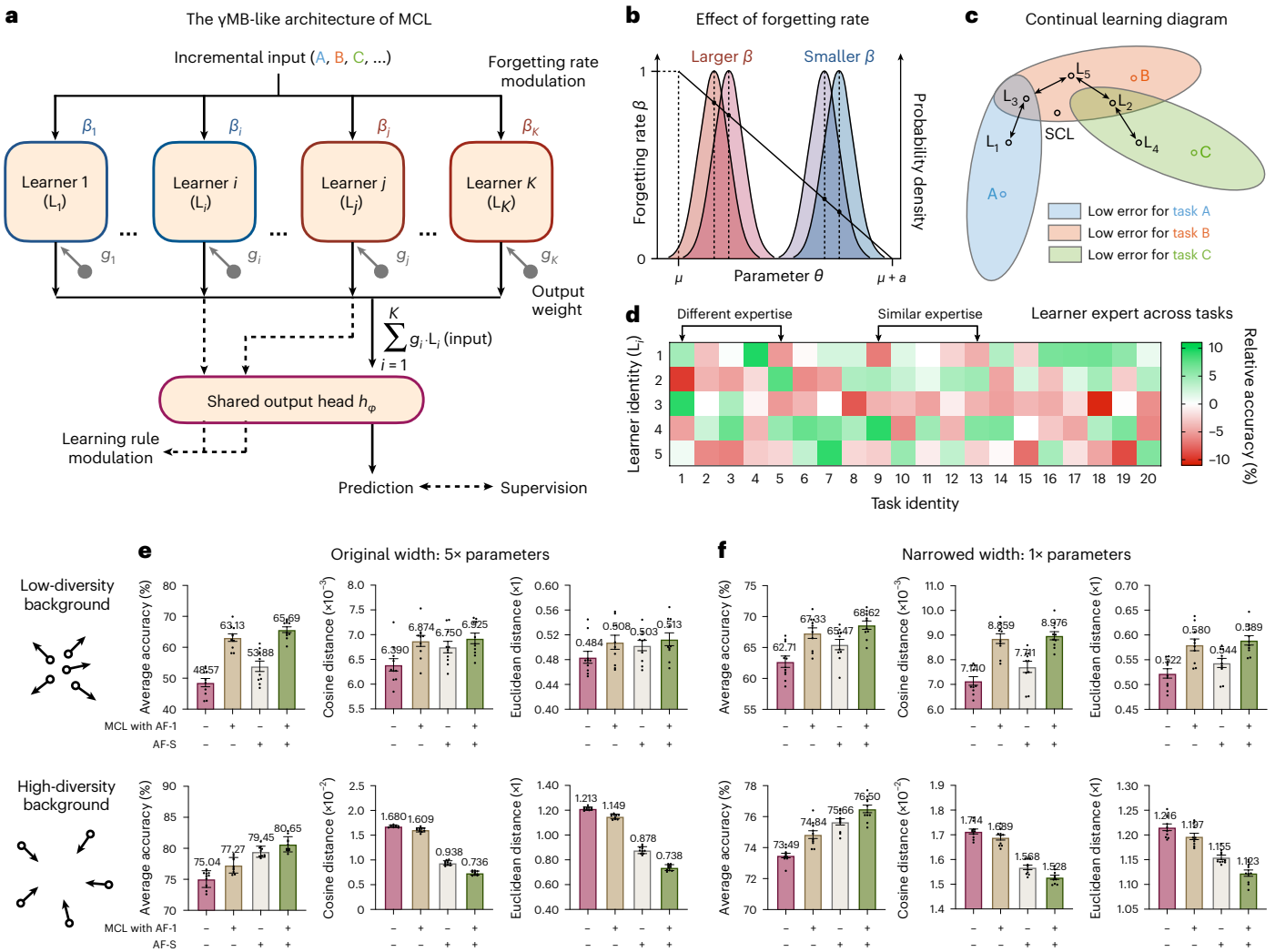

Fig. 3 | The γMB-like architecture of MCL with adaptive modulations. **a**, Inspired by the organizing principles of the γMB system, we design a specialized architecture consisting of multiple parallel continual learners L_1–L_K. **b**, Learner differences can be modulated by the forgetting rate β. Here we present an example with the target distribution $p(\theta|D_A, D_B, \beta)$ when $p(\theta) = \mathcal{N}(\mu, \sigma^2)$ and $p(\theta|D_A) = \mathcal{N}(\mu + a, \sigma^2)$ in $\hat{p}(\theta|D_A, \beta) = \frac{p(\theta|D_A)^{(1-\beta)}p(\theta)^{\beta}}{Z}$. In this case, $p(\theta|D_A, D_B, \beta)$ is also a Gaussian (Supplementary Section 1.1) and the vertical dashed line denotes its mode. **c**, A conceptual diagram of continual learning. A, B and C are three incremental tasks with different similarities. L_1–L_5 are five learners with appropriate diversity in parameter space. The coloured hollow circles indicate the optimal solution for each task. **d**, Learners' expertise across tasks. After continual learning of all tasks, we evaluate the relative accuracy of each learner across tasks, calculated as the performance of each learner minus the average performance of all learners. **e**,**f**, The two adaptive modulations (AF-1 and AF-S) can improve the performance of MCL to a large extent through coordinating the diversity of learners' expertise, as measured by the average cosine or Euclidean distance of their predictions. **e** and **f** represent the results of MCL with the original width and the narrowed width, respectively, averaged over ten runs with different random seeds and task orders. The error bars represent the standard error of the mean. All experiments are performed on R-CIFAR-100 with EWC[7] as the baseline approach.

constructed from the CIFAR-100 dataset[45] of 100-class coloured images but with different degrees of overall knowledge transfer[42]. As shown in Fig. 2c, the proposed active forgetting can largely enhance the average accuracy of all tasks, using EWC[7] as a baseline for preserving memory stability. Then we analyse the benefits of active forgetting on learning plasticity and memory stability with the metrics of forward transfer and backward transfer, respectively, where the former is clearly dominant. Similar results are observed when plugging the active-forgetting term in other synaptic regularization methods that preserve only memory stability (Supplementary Fig. 1). In contrast, active forgetting fails to improve the joint training performance (Supplementary Fig. 4a), suggesting that its benefits are specific to continual learning. From visual interpretation of the latest task predictions in Fig. 2d, active forgetting can indeed eliminate the past conflicting information, leading to better recognition of the object itself.

Coordination of multiple continual learners

After demonstrating the benefits of active forgetting together with stability protection for a single continual learner (SCL), we turn to investigate the organizing principles of the γMB system where new memory forms and active forgetting happens[19–22,25,36,38,39,46]. Specifically, there are five compartments that process sequential experiences in parallel. The outputs of these compartments are integrated in a weighted-sum fashion to guide adaptive behaviours. Inspired by this, we design a γMB-like architecture consisting of multiple parallel continual learners (Fig. 3a). Each learner employs a parameter space to learn all tasks, but its dedicated output head is removed and the weighted sum of the previous layer's output is fed into a shared output head, where the output weights of each learner are incrementally updated.

In such a γMB-like architecture, the relationship between learners is critical to the performance of continual learning. When the diversity

https://doi.org/10.1038/s42256-023-00747-w

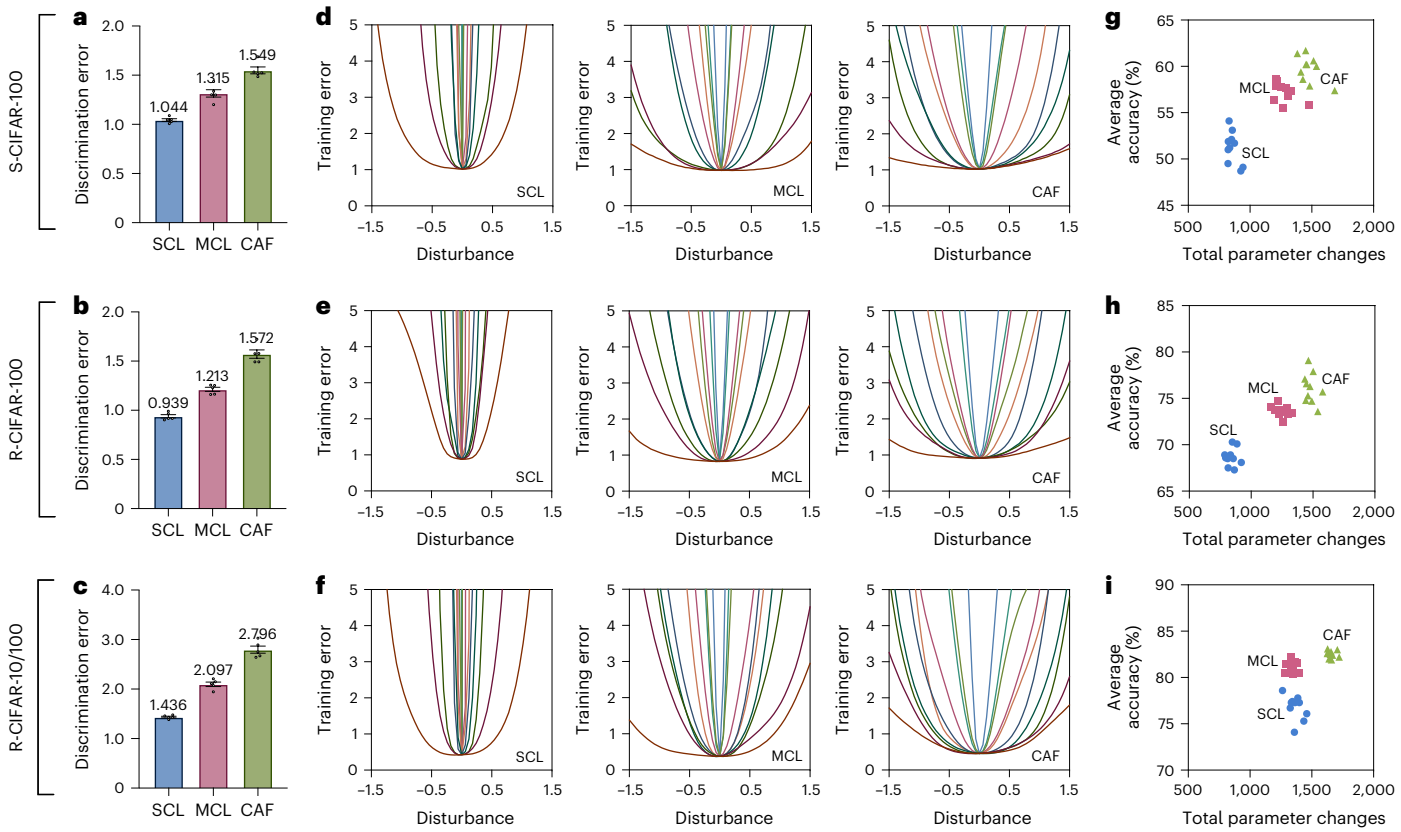

Fig. 4 | Exploration of underlying mechanisms that improve continual learning. We build the MCL with a narrowed width to keep the total amount of parameters similar to that of the SCL, and adopt the high-diversity background as a sub-optimal implementation. All experiments are performed with EWC[7] as the baseline approach. **a**–**c**, We empirically approximate the discrepancy of task distributions in feature space by training a simple discriminator to distinguish whether the feature of an input image belongs to a task or not, and measure the average discrimination error on test sets via binary cross-entropy, where a larger discrimination error indicates a smaller difference[50,51]. The error bars represent the standard error of the mean over five runs. **d**–**f**, Curvature of loss landscape around the obtained solution after continual learning of all tasks. After disturbing the network parameters with random values, we evaluate the training error with cross-entropy[67], where each line is a different direction of disturbance. **g**–**i**, Total parameter changes and average accuracy of all tasks. Each point represents a run. **a**,**d**,**g**, The results of S-CIFAR-100. **b**,**e**,**h**, The results of R-CIFAR-100. **c**,**f**,**i**, The results of R-CIFAR-10/100.

of their expertise is properly coordinated, the obtained solution can provide a high degree of compatibility with both new and old tasks. Here we present a conceptual illustration via performing tasks A, B and C with different similarities (Fig. 3c). As it is difficult to find a shared optimal solution for all tasks, the SCL has to converge to a high-error region for tasks A and C. In contrast, the multiple continual learners (MCL) with appropriate diversity allow for division of labour to address task discrepancy and complement their functions as parameter changes. Then, the output weights can integrate the respective expertise of each learner into the final prediction.

In general, each learner's expertise is directly modulated by its target distribution, where the proposed functional strategy can naturally serve this purpose. With the formulation of active forgetting, the target distribution $p(\theta|D_A, D_B, \beta) \propto p(D_B|\theta)\hat{p}(\theta|D_A, \beta)$ tends to be different for learners with different forgetting rates β, and vice versa (Fig. 3b). Therefore, we implement the forgetting rates adaptively for these learners to coordinate their relationship. As the target distribution also depends on $p(D_B|\theta)$, we further propose a supplementary modulation that constrains explicitly the differences in predictions between learners, corresponding to adjusting the learning rules for each new task. We refer to the γMB-like architecture with these two modulations as collaborative continual learners with active forgetting (CAF), and provide a formal definition in equation (12) in Methods.

For multiple learners with identical network architectures and similar forms of learning objectives, the priority is to obtain adequate differentiation of their expertise. In this case, the forgetting rates serve to diversify these learners, similar to the neurological strategy of decaying old memories differentially in each compartment[24,36–39]. In practice, the differences of learners can also arise from their innate randomness, such as the use of dropout and different random initializations, leading to sub-optimal solutions with moderate performance. This potentially corresponds to the anatomical randomness of Kenyon cells receiving olfactory signals in *Drosophila*[47–49]. At this point, the modulations of forgetting rates and learning rules can provide finer adjustments, for example, by constraining excessive differences.

Given the same network width for each learner, using more learners (that is, more parameters) generally results in better performance. However, there is an intuitive trade-off between learner number and width under a limited parameter budget. We verify that this trade-off is independent of training data distributions (Supplementary Section 1.3) and is relatively insensitive over a wide range (Supplementary Table 5). Therefore, we simply choose five learners ($K = 5$) corresponding to the five biological compartments, which employs approximately 5× parameters, and then reduce the network width accordingly to keep the total amount of parameters similar to that of the SCL. To evaluate the effect of innate diversity, we construct a low-diversity background by removing the dropout and using the same random initialization for each learner, and a high-diversity background by maintaining these randomness factors, where the overall diversity of expertise is evaluated by the average cosine or Euclidean distance between learners' predictions.

https://doi.org/10.1038/s42256-023-00747-w

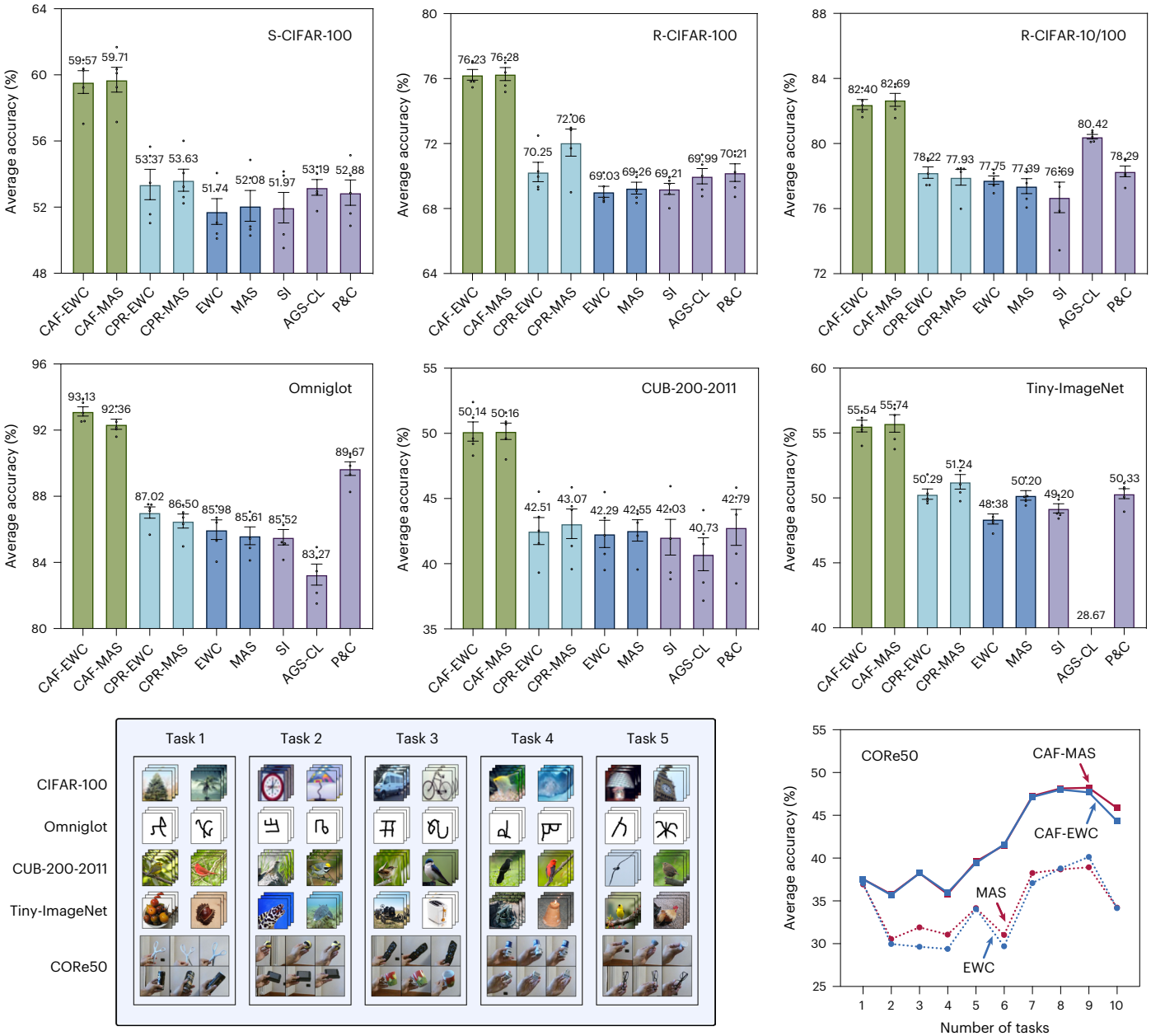

Fig. 5 | Performance evaluation for visual classification tasks. We consider multiple visual classification benchmarks to evaluate different aspects of continual learning, such as overall knowledge transfer, input size, number of training samples, length of task sequence, smoothly changed observations and so on. The bottom-left panel shows a demo of these benchmarks. CAF (ours) and CPR[56] are plug-and-play for synaptic regularization methods such as EWC[7] and MAS[8]. Under similar parameter budgets, all results are averaged over five runs with different random seeds and task orders. The error bars represent the standard error of the mean.

As shown in Fig. 3e,f, adaptive implementations of either active forgetting (AF-1) or its supplementary modulation (AF-S) can greatly enhance the performance of MCL, where the degree of improvements varies with the effectiveness of increasing inadequate diversity or reducing excessive diversity in the two backgrounds, respectively. In response to different degrees of innate diversity, the respective advantages of AF-1 and AF-S are combined to achieve consistently better performance. Such modulations enable the multiple learners to effectively divide and cooperate in continual learning (Fig. 3d and Supplementary Figs. 4b,c and 5). They show a clear diversity of task expertise with several experts collaborating on each task, validating the conceptual model in Fig. 3c. Accordingly, the performance of CAF is largely superior to that of the SCL, averaging the predictions of five independently trained continual learners, or using a separate learner for each task (Supplementary Fig. 6a,b). In particular, CAF can improve the SCL by a similar magnitude under different parameter budgets, indicating its outstanding scalability (Supplementary Figs. 6c and 7a).

According to our theoretical analysis in Proposition 1 in Methods, the performance of a shared solution for new and old tasks depends on the discrepancy of task distributions and the flatness of loss landscape around it. With respect to these two aspects, we delve more deeply into the benefits of our approach. As for the former, we evaluate the discrepancy of task distributions in feature space via the difficulty of distinguishing them after continual learning[50,51] (Fig. 4a–c). The large increase in discrimination error suggests that our approach can

https://doi.org/10.1038/s42256-023-00747-w

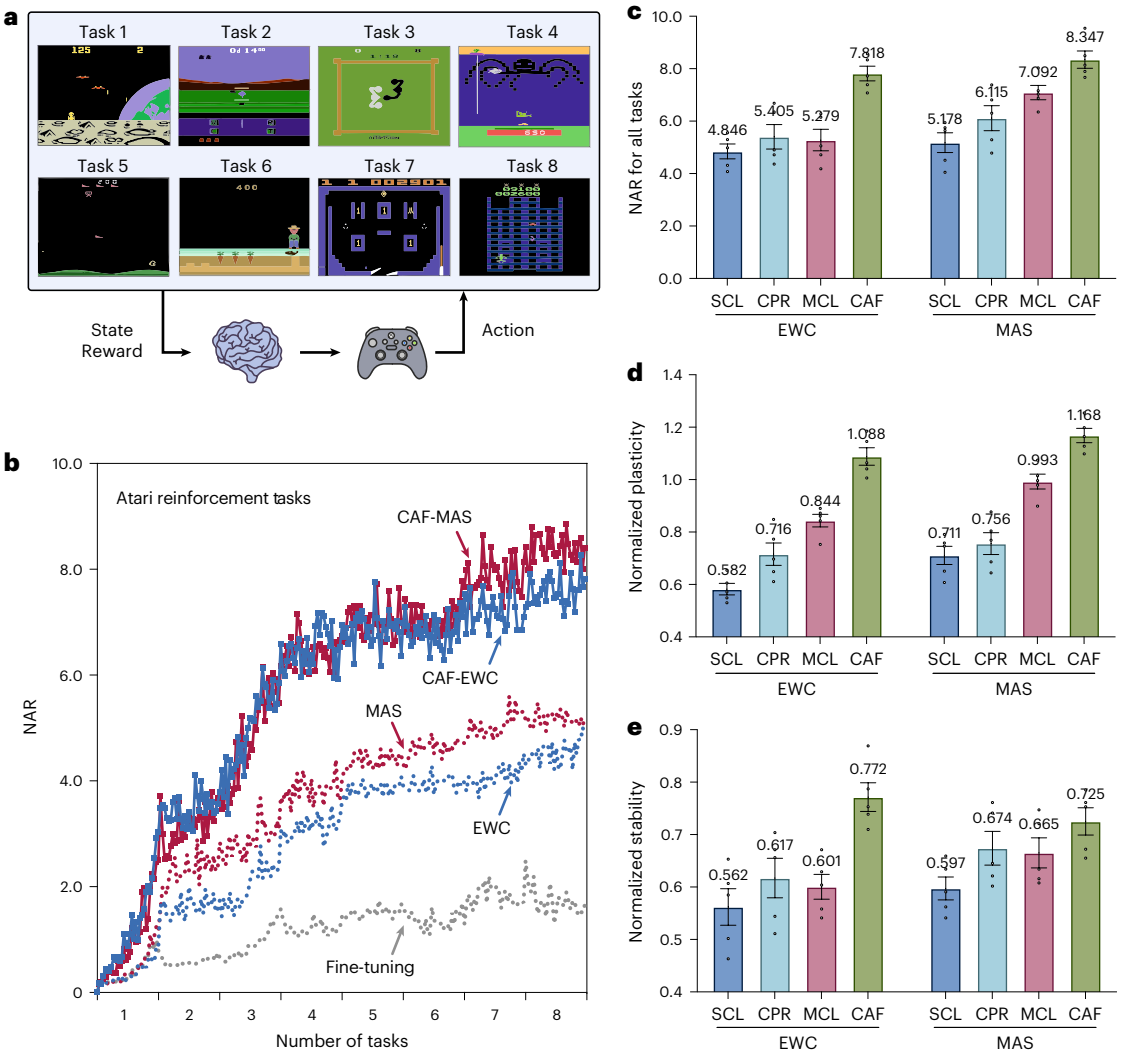

Fig. 6 | Performance evaluation for Atari reinforcement tasks. a, An agent attempts to acquire more rewards from learning a sequence of Atari games. **b**, The NAR in continual learning. The performance of simply fine-tuning on the task sequence is used to normalize the reward obtained for each task. **c**–**e**, After continual learning of all Atari games, we evaluate the overall performance (**c**), learning plasticity (**d**) and memory stability (**e**). Under similar parameter budgets, all results are averaged over five runs with different random seeds. The error bars represent the standard error of the mean.

successfully reconcile this discrepancy. As for the latter, the solution obtained by ours enjoys a clearly flatter loss landscape (Fig. 4d–f), indicating that it is more robust to modest parameter changes in response to dynamic data distributions. Therefore, CAF can update parameters more flexibly than the SCL (Fig. 4g–i), with the performance of new and old tasks simultaneously improved (Supplementary Fig. 7b,c).

Finally, we evaluate CAF under the setting of task-incremental learning[52] and compare it with a range of representative methods[7–9,53–56]. We first consider visual classification tasks with different particular challenges. Besides the overall knowledge transfer, we additionally use four benchmark datasets such as Omniglot[57] for long task sequence with imbalanced class numbers, CUB-200-2011[58] and Tiny-ImageNet[17] for larger-scale images, and CORe50[59] for smoothly changed observations. As shown in Fig. 5, the performance of all baselines varies widely across experimental settings, while CAF achieves consistently the strongest performance in a plug-and-play manner. We further experiment with Atari reinforcement tasks, where an agent incrementally learns to play several Atari games (Fig. 6a). The overall performance is evaluated by the normalized accumulated reward (NAR)[42,55,56], where the rewards obtained for all tasks ever seen are normalized with the maximum reward of fine-tuning on each task, and then accumulated. Likewise, CAF can greatly enhance the performance of baseline approaches (Fig. 6b,c) through improving both learning plasticity and memory stability (Fig. 6d,e).

Discussion

Whether for animals, robots or other intelligent agents, the ability of continual learning is critical for successfully adapting to the real world. In this work, we draw inspirations from the adaptive mechanisms equipped in a robust biological learning system, and present a generic approach for continual learning in artificial neural networks. Our preliminary versions of some individual components have been presented at top conferences in AI[42,51], while the current version enjoys substantial extensions in terms of technical robustness, synergistic cooperation and biological plausibility (Supplementary Section 3). The superior performance and generality of our approach can facilitate realistic applications, such as smartphones, robotics and autonomous driving, to flexibly accommodate user needs and environmental changes. Meanwhile, the deployment of continual learning avoids retraining all previous data each time the model is updated, which provides an energy-efficient and eco-friendly path for developing AI systems.

To bridge the gap between AI and BI, we carefully avoid involving specific implementations or overly strong assumptions in both theoretical analysis and computational modelling. This consideration not only allows for an adequate exploitation of biological advantages but also facilitates the emergence of interdisciplinary insights. Computationally, our approach is proven to satisfy the key factors on which continual learning performance depends, such as stability, plasticity and compatibility, with active forgetting playing an important role. This potentially extends the previous focus of preventing catastrophic forgetting in continual learning. Starting with this idea, below we discuss more broadly the connections between AI and BI in adaptability.

In a biological sense, active forgetting allows flexibility to accommodate external changes by removing outdated information[27,30,60]. This perspective is well supported by extensive theoretical and empirical evidence in our computational model. Recent work in neurobiology is deeply dissecting its underlying mechanisms from molecular to synaptic structural levels, where activation of the molecular signalling that mediates active forgetting initially leads to rapid growth of synaptic structures, but prolonged activation instead leads to their shrinkage[19,30,42,60–64]. In our computational model, active forgetting of old memories in parameter distributions can derive two equivalent synaptic expansion-renormalization processes. These two processes and their linear combinations cover a wide range of possible forms, including whether old memories are directly affected and whether expanded parameters encode new memories, which can serve as testable hypotheses for further research. As for the five compartments of the γMB system, the modulated forgetting rates have been shown to be important for coping with conflicting memories in succession[24,36–39]. Correspondingly, adaptive implementations of active forgetting help the γMB-like architecture to better accommodate incremental changes. Besides, we identify the necessity of regularizing learners' predictions of new tasks, suggesting that the adaptation of learning rules may also contribute to continual learning in a more general context.

AI and BI share the common goal of adaptation and survival in the real world. These two fields have great potential to inspire each other and progress together. This requires generalized theories and methodologies to integrate their advances, as suggested by our work in continual learning. Subsequent work could further explore the 'natural algorithms' responsible for other advantages of the biological brain, thereby evolving progressively the current AI systems.

Methods

Synaptic expansion-renormalization

For the case of two tasks, the learner needs to find a mode of the posterior distribution that incorporates the knowledge of tasks A and B:

$$\begin{aligned}\theta^*_{\mathrm{A,B}} &= \arg\max_\theta \log p(\theta|D_\mathrm{A}, D_\mathrm{B}) \\ &= \arg\max_\theta \log p(D_\mathrm{B}|\theta) + \log p(\theta|D_\mathrm{A}) - \underbrace{\log p(D_\mathrm{B})}_{\text{constant}},\end{aligned} \tag{3}$$

where $p(D_\mathrm{B}|\theta)$ is the loss for task B. Although $p(\theta|D_\mathrm{A})$ is generally intractable, we can locally approximate it with a second-order Taylor expansion around $\theta^*_\mathrm{A} = \arg\max_\theta \log p(\theta|D_\mathrm{A})$, resulting in a Gaussian distribution whose mean is θ^*_A and precision matrix is the Hessian of the negative log posterior[7,11,65]. To simplify the computation, the Hessian is approximated by the diagonal of the Fisher information matrix:

$$F_\mathrm{A} = \mathbb{E}\left[\left(\frac{\partial \log p(\theta|D_\mathrm{A})}{\partial\theta}\right)\left(\frac{\partial \log p(\theta|D_\mathrm{A})}{\partial\theta}\right)^\mathsf{T}|_{\theta^*_\mathrm{A}}\right]. \tag{4}$$

To improve learning plasticity, we introduce a forgetting rate β, and replace $p(\theta|D_\mathrm{A})$ in equation (3) with $\hat{p}(\theta|D_\mathrm{A},\beta) = \frac{p(\theta|D_\mathrm{A})^{(1-\beta)}p(\theta)^\beta}{Z}$ as equation (1), where $\beta \in [0,1]$ is deterministic and Z is a β-dependent normalizer. Correspondingly, the target distribution $p(\theta|D_\mathrm{A}, D_\mathrm{B})$ becomes $p(\theta|D_\mathrm{A}, D_\mathrm{B}, \beta)$. $\hat{p}$ has a nice property that it follows a Gaussian distribution if $p(\theta|D_\mathrm{A})$ and $p(\theta)$ are both Gaussian (Supplementary Section 1.1), so we can compute $\hat{p}$ in a similar way to how we compute $p(\theta|D_\mathrm{A})$. A certain value of β can maximize the probability of learning each new task well through forgetting the old memories, validating the motivation of introducing the non-informative prior in equation (1):

$$\beta^* = \arg\max_\beta p(D_\mathrm{B}|D_\mathrm{A},\beta) = \arg\max_\beta \int p(D_\mathrm{B}|\theta)\hat{p}(\theta|D_\mathrm{A},\beta)\mathrm{d}\theta. \tag{5}$$

With the implementation of active forgetting, the learner needs to find

$$\begin{aligned}\theta^*_{\mathrm{A,B}} &= \arg\max_\theta \log p(\theta|D_\mathrm{A}, D_\mathrm{B}, \beta) \\ &= \arg\max_\theta \log p(D_\mathrm{B}|\theta) + \log \hat{p}(\theta|D_\mathrm{A},\beta) - \underbrace{\log p(D_\mathrm{B})}_{\text{constant}} \\ &\overset{\text{AF-1}}{=} \arg\max_\theta \log p(D_\mathrm{B}|\theta) + (1-\beta)\log p(\theta|D_\mathrm{A}) + \beta \log p(\theta) \\ &\overset{\text{AF-2}}{=} \arg\max_\theta (1-\beta)\log p(D_\mathrm{B}|\theta) + (1-\beta)\log p(\theta|D_\mathrm{A}) + \beta \log p(\theta|D_\mathrm{B}),\end{aligned} \tag{6}$$

which can be optimized in two equivalent ways, that is, AF-1 and AF-2. Correspondingly, we derive the loss function in equation (2).

For continual learning of more than two tasks, for example, t tasks for any $t > 2$, the learner needs to find

$$\theta^*_{1:t} = \arg\max_\theta \log p(D_t|\theta) + \log p(\theta|D_{1:t-1}, \beta_{1:t-1}) - \underbrace{\log p(D_t)}_{\text{constant}}, \tag{7}$$

where $D_{1:t-1} = \bigcup_{i=1}^{t-1} D_i$ denotes the training data of previous task(s) and $\beta_{1:t-1} = \{\beta_i\}_{i=1}^{t-1}$ denotes the previously used forgetting rate(s). Similarly, we replace the posterior $p(\theta|D_{1:t-1}, \beta_{1:t-1})$ that absorbs all information of $D_{1:t-1}$ with

$$\hat{p}(\theta|D_{1:t-1}, \beta_{1:t-1}, \beta_t) = \frac{p(\theta|D_{1:t-1}, \beta_{1:t-1})^{(1-\beta_t)}p(\theta)^{\beta_t}}{Z_t}, \tag{8}$$

where Z_t is a β_t-dependent normalizer that keeps $\hat{p}$ a normalized probability distribution. To simplify the hyperparameter tuning, we adopt an identical forgetting rate in continual learning, that is, $\beta_i = \beta$ for $i = 1, \ldots, t$. Then we obtain the loss function:

$$\mathcal{L}^{\mathrm{AF}}_{\mathrm{Reg}}(\theta) = \mathcal{L}_t(\theta) + \underbrace{\frac{\lambda_{\mathrm{SP}}}{2}\sum_m F_{1:t-1,m}(\theta_m - \theta^*_{1:t-1,m})^2}_{\text{stability protection}} + \underbrace{\frac{\lambda_{\mathrm{AF}}}{2}\sum_m I_{\mathrm{e},m}(\theta_m - \theta_{\mathrm{e},m})^2}_{\text{active forgetting}}. \tag{9}$$

$\theta^*_{1:t-1}$ is the obtained solution for previous tasks, that is, the old network parameters. For AF-1, $\theta_{\mathrm{e},m} = 0$, $I_{\mathrm{e},m} = 1$, $\lambda_{\mathrm{AF}} \propto \beta$ and $\lambda_{\mathrm{SP}} \propto (1-\beta)$. For AF-2, $\theta_{\mathrm{e},m} = \theta^*_{t,m}$, $I_{\mathrm{e},m} = F_{t,m}$, $\lambda_{\mathrm{AF}} \propto \beta/(1-\beta)$ and $\lambda_{\mathrm{SP}} \propto 1$. $F_{1:t-1}$ is recursively updated by

$$F_{1:t-1} = F_{1:t-2} + F_{t-1}. \tag{10}$$

When $\beta = 0$, the loss function in equation (9) degenerates to a similar form as regular synaptic regularization methods that preserve only memory stability[7–10]:

$$\mathcal{L}_{\mathrm{Reg}}(\theta) = \mathcal{L}_t(\theta) + \underbrace{\frac{\lambda_{\mathrm{SP}}}{2}\sum_m \xi_{1:t-1,m}(\theta_m - \theta^*_{1:t-1,m})^2}_{\text{stability protection}}. \tag{11}$$

As these methods differ mainly in the metric $\xi_{1:t-1}$ of estimating the importance of parameters for performing old tasks[43], the

active-forgetting term can be naturally combined with them. We discuss in more depth the motivation and implementation of active forgetting in Supplementary Section 1.2, including the choice of an appropriate β, the connections of two equivalent versions and the technical details of derivation.

Multiple parallel continual learners

The γMB-like architecture of MCL adopts K identically structured neural networks $f_{\phi_i}(\cdot), i=1,\ldots,K$, corresponding to K continual learners L_i, $i=1,\ldots,K$ with their own parameter sets ϕ_i. We remove the dedicated output head of each learner, and feed the weighted sum of the previous layer's output into a shared output head $h_\varphi(\cdot)$ to make predictions, where the output weights of each learner g_i, $i=1,\ldots,K$ are updated incrementally. Then, the final prediction becomes $\tilde{p}(\cdot)=h_\varphi(\sum_{i=1}^{K} g_i f_{\phi_i}(\cdot))$, where φ denotes the parameter set of prediction function and the optimizable MCL parameters θ_{MCL} include $\bigcup_{i=1}^{K}\phi_i$, $\bigcup_{i=1}^{K} g_i$ and φ. The proposed MCL is applicable to a wide range of loss functions for continual learning. By default, here we focus on the synaptic regularization methods[7–10] as defined in equation (11). To coordinate the diversity of learners' expertise, we implement the proposed active forgetting in each learner. We further regularize differences in their predictive distributions, quantified by the widely used Kullback–Leibler divergence. Therefore, the loss function for the full version of our approach is defined as:

$$\begin{aligned}\mathcal{L}_{\mathrm{CAF}}(\theta_{\mathrm{MCL}}) = \mathcal{L}_t(\theta_{\mathrm{MCL}}) &+ \underbrace{\frac{\lambda_{\mathrm{SP}}}{2}\sum_{i=1}^{K}\sum_{m=1}^{M_i}\xi_{1:t-1,i,m}(\theta_{i,m}-\theta^*_{1:t-1,i,m})^2}_{\text{stability protection}} \\ &+ \underbrace{\sum_{i=1}^{K}\frac{\lambda_{\mathrm{AF},i}}{2}\sum_{m=1}^{M_i}(\theta_{i,m})^2}_{\text{AF-1}} + \underbrace{\sum_{i=1,j\neq i}^{K}\frac{\gamma_{i,j}}{N_t}\sum_{n=1}^{N_t}\tilde{p}_i(x_{t,n})\log\frac{\tilde{p}_i(x_{t,n})}{\tilde{p}_j(x_{t,n})}}_{\text{AF-}S},\end{aligned} \tag{12}$$

where M_i denotes the amount of parameters $\theta_i=\{\phi_i, g_i\}$ for learner i. n denotes the index of training samples in each task. The current task has N_t training samples, and $\tilde{p}_i(x_{t,n})$ is the prediction of learner i for $x_{t,n}$. We mainly consider AF-1 instead of AF-2 for computational efficiency, while keeping λ_{SP} identical to each learner for ease of implementation. We then discuss the implementation of $\lambda_{\mathrm{AF},i}$ and $\gamma_{i,j}$. If we consider CAF as an overall continual learning model, active forgetting can be implemented similarly to equations (1) and (2), setting a uniform forgetting rate $\lambda_{\mathrm{AF},i}$ as well as a uniform learning rule $\gamma_{i,j}$ to each learner. This implementation is indeed effective under strong innate randomness to reduce excessive diversity. In a more general context, the proposed MCL provides a spatial degree of freedom to modulate the diversity of learner's expertise. This idea can be implemented by constraining the average of $\{\lambda_{\mathrm{AF},i}\}_{i=1}^{K}$ to be a deterministic hyperparameter λ_{AF}, that is, $\frac{1}{K}\sum_{i=1}^{K}\lambda_{\mathrm{AF},i}=\lambda_{\mathrm{AF}}$ where $\lambda_{\mathrm{AF},i}=\alpha_i K\lambda_{\mathrm{AF}}$ and $\sum_{i=1}^{K}\alpha_i=1$, but allowing their relative strength α_i as well as θ_{MCL} to be optimized with gradients of the same loss function. Specifically, we perform softmax of a few optimizable parameters to ensure the constraint $\sum_{i=1}^{K}\alpha_i=1$ and obtain α_i for each learning module. $\gamma_{i,j}$ can be implemented in a similar way by constraining their average $\frac{1}{K(K-1)}\sum_{i=1,j\neq i}^{K}\gamma_{i,j}=\gamma$, to enjoy the spatial degree of freedom.

Theoretical analysis of generalization ability

Continual learning aims to find a solution θ that can generalize well over a new distribution $\mathbb{D}_t$ and a set of old distributions $\mathbb{D}_{1:t-1}:=\{\mathbb{D}_k\}_{k=1}^{t-1}$. Let $\mathcal{E}_{\mathbb{D}_t}(\theta)$ and $\mathcal{E}_{\mathbb{D}_{1:t-1}}(\theta)$ denote the generalization errors. The training set and test set of each task follow the same distribution $\mathbb{D}_k$ ($k=1,2,\ldots,t$), where the training set $D_k=\{(x_{k,n},y_{k,n})\}_{n=1}^{N_k}$ includes N_k data-label pairs. Then we define $\mathcal{E}_{\mathbb{D}_t}(\theta)=\mathbb{E}_{(x,y)\sim\mathbb{D}_t}[\mathcal{L}_t(\theta;x,y)]$ and $\mathcal{E}_{\mathbb{D}_{1:t-1}}(\theta)=\frac{1}{t-1}\sum_{k=1}^{t-1}\mathbb{E}_{(x,y)\sim\mathbb{D}_k}[\mathcal{L}_k(\theta;x,y)]$, where $\mathcal{L}_k(\theta)$ can be generalized for any bounded loss function of task k. To minimize $\mathcal{E}_{\mathbb{D}_t}(\theta)$ and $\mathcal{E}_{\mathbb{D}_{1:t-1}}(\theta)$ without the use of old training samples $D_{1:t-1}:=\{D_k\}_{k=1}^{t-1}$, a continual learning model can only minimize an empirical risk over the current training samples D_t in a parameter space Θ applicable to old tasks[51,66], denoted as $\min_{\theta\in\Theta}\hat{\mathcal{E}}_{D_t}(\theta)$ where $\hat{\mathcal{E}}_{D_t}(\theta)=\frac{1}{N_t}\sum_{n=1}^{N_t}\mathcal{L}_t(\theta;x_{t,n},y_{t,n})$. In practice, sequential learning of each task by $\min_{\theta\in\Theta}\hat{\mathcal{E}}_{D_t}(\theta)$ can find multiple solutions with different generalizability for $\mathcal{E}_{\mathbb{D}_t}(\theta)$ and $\mathcal{E}_{\mathbb{D}_{1:t-1}}(\theta)$, where a solution with flatter loss landscape typically acquires better generalizability and is therefore more robust to catastrophic forgetting[56,67,68].

Accordingly, we define a robust empirical risk for the current task as $\hat{\mathcal{E}}^b_{D_t}(\theta):=\max_{\|\Delta\|\leq b}\hat{\mathcal{E}}_{D_t}(\theta+\Delta)$ by the worst case of parameter perturbations Δ, where $\|\cdot\|$ denotes the L2 norm and b is the radius of perturbations around θ. Likewise, a robust empirical risk for the old tasks is $\hat{\mathcal{E}}^b_{D_{1:t-1}}(\theta):=\max_{\|\Delta\|\leq b}\hat{\mathcal{E}}_{D_{1:t-1}}(\theta+\Delta)$. Then, $\min_{\theta\in\Theta}\hat{\mathcal{E}}^b_{D_t}(\theta)$ can find a flat minima over the current task. However, parameter changes that are much larger than the 'radius' of the old minima can interfere with the performance of old tasks, while staying around the old minima can interfere with the performance of new tasks. Therefore, it is necessary to find a solution that can properly balance memory stability with learning plasticity, while being adequately compatible with the observed distributions. Formally, we analyse the generalization errors of a certain solution for continual learning with PAC-Bayes theory[69], to provide the objective for computational modelling of neurological adaptive mechanisms. We leave technical details to Supplementary Section 1.3 and present the main results below.

Proposition 1. Let $\{\Theta_i\in\mathbb{R}^{M_i}\}_{i=1}^{K}$ be a set of parameter spaces ($K\geq 1$ in general), d_i be a Vapnik–Chervonenkis (VC) dimension of Θ_i, $\Theta=\cup_{i=1}^{K}\Theta_i$ with VC dimension d, and $M=\sum_{i=1}^{K}M_i$ as a given parameter budget. Let $\hat{\theta}^b_{1:t}$ denote the optimal solution of the continually learned 1:t tasks by robust empirical risk minimization over the current task, that is, $\hat{\theta}^b_{1:t}=\arg\min_{\theta\in\Theta}\hat{\mathcal{E}}^b_{D_t}(\theta)$. Then for any $\delta\in(0,1)$, with probability at least $1-\delta$:

$$\begin{aligned}&\underbrace{\mathcal{E}_{\mathbb{D}_t}(\hat{\theta}^b_{1:t})-\min_{\theta\in\Theta}\mathcal{E}_{\mathbb{D}_t}(\theta)}_{\text{learning plasticity}}\leq\underbrace{\min_{\theta\in\Theta}\hat{\mathcal{E}}^b_{D_{1:t-1}}(\theta)-\min_{\theta\in\Theta}\hat{\mathcal{E}}_{D_{1:t-1}}(\theta)}_{\text{loss flatness}}\\&+\underbrace{\frac{1}{t-1}\sum_{k=1}^{t-1}\mathrm{Div}(\mathbb{D}_k,\mathbb{D}_t)}_{\text{task discrepancy}}+C_1,\\&\underbrace{\mathcal{E}_{\mathbb{D}_{1:t-1}}(\hat{\theta}^b_{1:t})-\min_{\theta\in\Theta}\mathcal{E}_{\mathbb{D}_{1:t-1}}(\theta)}_{\text{memory stability}}\leq\underbrace{\min_{\theta\in\Theta}\hat{\mathcal{E}}^b_{D_t}(\theta)-\min_{\theta\in\Theta}\hat{\mathcal{E}}_{D_t}(\theta)}_{\text{loss flatness}}\\&+\underbrace{\frac{1}{t-1}\sum_{k=1}^{t-1}\mathrm{Div}(\mathbb{D}_t,\mathbb{D}_k)}_{\text{task discrepancy}}+C_2,\end{aligned} \tag{13}$$

where $C_1=\max_{i\in[1,K]}\sqrt{\frac{d_i\ln(N_{1:t-1}/d_i)+\ln(2K/\delta)}{N_{1:t-1}}}+\sqrt{\frac{d\ln(N_{1:t-1}/d)+\ln(2/\delta)}{N_{1:t-1}}}$ and $C_2=\max_{i\in[1,K]}\sqrt{\frac{d_i\ln(N_t/d_i)+\ln(2K/\delta)}{N_t}}+\sqrt{\frac{d\ln(N_t/d)+\ln(2/\delta)}{N_t}}$ represent the cover of parameter space. $\mathrm{Div}(\mathbb{D}_i,\mathbb{D}_j):=2\sup_{h\in\mathcal{H}}|\mathcal{P}_{\mathbb{D}_i}(I(h))-\mathcal{P}_{\mathbb{D}_j}I(h))|$ is the $\mathcal{H}$ divergence of $\mathbb{D}_i$ and $\mathbb{D}_j$, where $I(h)$ is the characteristic function. $N_{1:t-1}=\sum_{k=1}^{t-1}N_k$ is the total number of training samples over all old tasks, where N_k is the number of training samples over task k.

From Proposition 1, the generalization gaps over new and old tasks, corresponding to learning plasticity and memory stability, are uniformly constrained by the loss flatness and task discrepancy, which further depend on the cover of parameter space, that is, C_1 and C_2. In particular, we have $C_1=2\sqrt{\frac{d\ln(N_{1:t-1}/d)+\ln(2/\delta)}{N_{1:t-1}}}$ and $C_2=2\sqrt{\frac{d\ln(N_t/d)+\ln(2/\delta)}{N_t}}$ for $K=1$. When $K>1$, $C_1<2\sqrt{\frac{d\ln(N_{1:t-1}/d)+\ln(2/\delta)}{N_{1:t-1}}}$ and $C_2<2\sqrt{\frac{d\ln(N_t/d)+\ln(2/\delta)}{N_t}}$ due to $d_i<d$ for $i\in[1,K]$. This means that employing MCL (that is, $K>1$)

compared with an SCL (that is, $K=1$) can tighten the generalization bounds and thus benefit the performance of continual learning. Likewise, the benefits of active forgetting can be explained from a similar perspective.

Proposition 2. Let $\{\Theta_i \in \mathbb{R}^{M_i}\}_{i=1}^{K}$ be a set of parameter spaces ($K \geq 1$ in general), d_i be a VC dimension of Θ_i, $\Theta = \cup_{i=1}^{K}\Theta_i$ with VC dimension d, and $M = \sum_{i=1}^{K} M_i$ as a given parameter budget. Based on Proposition 1, for $\hat{\theta}^b_{1:t} = \arg\min_{\theta\in\Theta}\hat{\mathcal{E}}^b_{\mathbb{D}_t}(\theta)$, the upper bound of generalization gap is further externalized with $C_1 = r_1(\|\hat{\theta}^b_{1:t}\|_2^2/b^2)$ and $C_2 = r_2(\|\hat{\theta}^b_{1:t}\|_2^2/b^2)$, where $r_1 : \mathbb{R}_+ \to \mathbb{R}_+$ and $r_2 : \mathbb{R}_+ \to \mathbb{R}_+$ are two strictly increasing functions under some technical conditions on $\mathcal{E}_{\mathbb{D}_t}(\theta)$ and $\mathcal{E}_{\mathbb{D}_{1:t-1}}(\theta)$, respectively.

Proposition 2 externalizes the two generalization bounds in Proposition 1 by defining the VC dimension with L2 norm for parameters under some technical assumptions. Notably, the claim in Proposition 1 regarding the benefits of using multiple continual learners still holds in such a tightened version. In particular, optimization of the active-forgetting term in equations (2) and (12), which takes the form of minimizing a weighted L2 norm regarding all parameters, can contribute to tightening the two generalization bounds in Proposition 2, especially through optimizing C_1 and C_2. Besides, $N_{1:t-1}$ becomes larger as t increases, while N_t remains constant in general. This means that C_1 will decrease more rapidly around the empirical optimal solution $\hat{\theta}^b_{1:t} = \arg\min_{\theta\in\Theta}\hat{\mathcal{E}}^b_{\mathbb{D}_t}(\theta)$ as more tasks are introduced, resulting in more pronounced improvements in learning plasticity.

Implementation

We mainly consider the setting of task-incremental learning[52] to perform continual learning experiments, with task identities provided in both training and testing. We split representative benchmark datasets for visual classification tasks. The CIFAR-100 dataset [45] includes 100-class coloured images of size 32 × 32. We split it based on different principles to evaluate the effect of overall knowledge transfer. Specifically, R-CIFAR-100 and S-CIFAR-100 are constructed by splitting CIFAR-100 into 20 tasks, depending on random order or superclasses defined by semantic similarity, respectively. R-CIFAR-10/100 includes 2 tasks randomly split from the CIFAR-10 dataset [45] of 10-class coloured images, followed by the 20 tasks of R-CIFAR-100. The Omniglot dataset[57] includes 50 alphabets for a total of 1,623 classes of characters, where each class contains 20 hand-written digits of size 28 × 28. We split each alphabet as a task consisting of a different number of classes. The CUB-200-2011 dataset[58] includes 200-class bird images of size 224 × 224, and the Tiny-ImageNet dataset[17] includes 200-class natural images of size 64 × 64, both split randomly into 10 tasks. The CORe50 dataset[59] includes 50 handheld objects with smoothly changed observations of size 128 × 128, randomly split into 10 tasks[70]. We construct a sequence of Atari reinforcement tasks for continual learning, that is, DemonAttack - Robotank - Boxing - NameThisGame - StarGunner - Gopher - VideoPinball - Crazyclimber, using the same PPO algorithm[71] to learn each task. The network architecture and training regime are described in Supplementary Sections 2.1 and 2.2, respectively.

Baseline approach

To ensure generality in realistic applications, we restrict the old training samples to be unavailable in continual learning, and compare with representative methods that follow this restriction. Specifically, EWC[7], memory aware synapses (MAS[8]) and synaptic intelligence (SI[9]) are synaptic regularization methods that selectively penalized parameter changes to preserve memory stability; adaptive group sparsity based continual learning (AGS-CL[55]) took advantages of parameter isolation and synaptic regularization to prevent catastrophic forgetting; progress & compress (P&C[54]) adopted an additional active column on the basis of EWC[7] to improve learning plasticity; classifier-projection regularization (CPR[56]) encouraged convergence to a flat loss landscape, which can be combined with other baseline approaches. The hyperparameters for continual learning are determined with a comprehensive grid search. We construct a different task sequence (that is, different class splits, data shuffling, task orders and random seeds) from the actual experiments and run it once. Then we use the best combinations of hyperparameters to perform the actual experiments for multiple runs, as described in Supplementary Section 2.3.

Evaluation metric

We consider three evaluation metrics for visual classification tasks, that is, average accuracy (AAC), forward transfer (FWT) and backward transfer (BWT)[6,72]:

$$\mathrm{AAC} = \frac{1}{T}\sum_{i=1}^{T} A_{T,i}, \tag{14}$$

$$\mathrm{FWT} = \frac{1}{T-1}\sum_{i=2}^{T} A_{i-1,i} - \tilde{a}_i, \tag{15}$$

$$\mathrm{BWT} = \frac{1}{T-1}\sum_{i=1}^{T-1} A_{T,i} - A_{i,i}, \tag{16}$$

where $A_{t,i}$ is the test accuracy of task i after continual learning of task t, and $\tilde{a}_i$ is the test accuracy of each task i learned from random initialization. ACC is the average performance of all tasks ever seen, which evaluates the overall performance of continual learning. FWT evaluates the average influence of remembering old tasks to new tasks for learning plasticity. BWT evaluates the average influence of learning new tasks to old tasks for memory stability.

The diversity of learners' predictions is quantified by the average cosine (Cos) or Euclidean (Euc) distance:

$$\mathrm{Cos} = 1 - \frac{1}{K(K-1)}\sum_{i=1,j\neq i}^{K}\frac{p_i \cdot p_j}{\| p_i \| \| p_j \|}, \tag{17}$$

$$\mathrm{Euc} = \frac{1}{K(K-1)}\sum_{i=1,j\neq i}^{K}\| p_i - p_j \|, \tag{18}$$

where p_i and p_j denote the predictions of learners i and j, respectively.

The discrepancy of task distributions in feature space is evaluated by the difficulty of distinguishing them, which is an empirical approximation of the $\mathcal{H}$ divergence[50,51] in Proposition 1. We train a simple discriminator consisting of a fully connected layer and use binary cross-entropy to measure the average discrimination error on test sets.

The performance of Atari reinforcement tasks is evaluated by the NAR, normalized plasticity (NP) and normalized stability (NS), corresponding to the overall performance, learning plasticity and memory stability, respectively:

$$\mathrm{NAR} = \sum_{i=1}^{T} R_{T,i}/r_i, \tag{19}$$

$$\mathrm{NP} = \frac{1}{T-1}\sum_{i=2}^{T} R_{i,i}/r_i, \tag{20}$$

$$\mathrm{NS} = \frac{1}{T-1}\sum_{i=1}^{T-1} R_{T,i}/R_{i,i}, \tag{21}$$

where $R_{t,i}$ is the reward for task i obtained in the test step after learning task t, and r_i is the maximum reward for task i obtained in each test step of fine-tuning on the task sequence.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All benchmark datasets used in this paper are publicly available, including CIFAR-10/100[45] (https://www.cs.toronto.edu/~kriz/cifar.html), Omniglot[57] (https://www.omniglot.com), CUB-200-2011[58] (https://www.vision.caltech.edu/datasets/cub_200_2011/), Tiny-ImageNet[17] (https://www.image-net.org/download.php), CORe50[59] (https://vlomonaco.github.io/core50/) and Atari games[73] (https://github.com/openai/baselines).

Code availability

The implementation code is available via Zenodo https://doi.org/10.5281/zenodo.8293564 ref. 74.

References

1. Chen, Z. & Liu, B. Lifelong machine learning. (San Rafael: Morgan & Claypool Publishers, 2018).
2. 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).
3. Kudithipudi, D. et al. Biological underpinnings for lifelong learning machines. *Nat. Mach. Intell.* **4**, 196–210 (2022).
4. McCloskey, M. & Cohen, N. J. Catastrophic interference in connectionist networks: the sequential learning problem. *Psychol. Learn. Motiv.* **24**, 109–165 (1989).
5. 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 (1995).
6. Wang, L., Zhang, X., Su, H. & Zhu, J. A comprehensive survey of continual learning: theory, method and application. Preprint at https://arxiv.org/abs/2302.00487 (2023).
7. Kirkpatrick, J. et al. Overcoming catastrophic forgetting in neural networks. *Proc. Natl Acad. Sci. USA* **114**, 3521–3526 (2017).
8. Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M. & Tuytelaars, T. Memory aware synapses: learning what (not) to forget. In *Proc. European Conference on Computer Vision* 139–154 (Springer, 2018).
9. Zenke, F., Poole, B. & Ganguli, S. Continual learning through synaptic intelligence. In *Proc. International Conference on Machine Learning* 3987–3995 (PMLR, 2017).
10. Chaudhry, A., Dokania, P. K., Ajanthan, T. & Torr, P. H. Riemannian walk for incremental learning: understanding forgetting and intransigence. In *Proc. European Conference on Computer Vision* 532–547 (Springer, 2018).
11. Ritter, H., Botev, A. & Barber, D. Online structured laplace approximations for overcoming catastrophic forgetting. *Adv. Neural Inf. Process. Syst.* **31**, 3742–3752 (2018).
12. Rebuffi, S.-A., Kolesnikov, A., Sperl, G. & Lampert, C. H. iCaRL: incremental classifier and representation learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 2001–2010 (IEEE, 2017).
13. Shin, H., Lee, J. K., Kim, J. & Kim, J. Continual learning with deep generative replay. *Adv. Neural Inf. Process. Syst.* **30**, 2990–2999 (2017).
14. Wang, L. et al. Memory replay with data compression for continual learning. In *International Conference on Learning Representations* (2021).
15. Serra, J., Suris, D., Miron, M. & Karatzoglou, A. Overcoming catastrophic forgetting with hard attention to the task. In *Proc. International Conference on Machine Learning* 4548–4557 (PMLR, 2018).
16. Fernando, C. et al. PathNet: evolution channels gradient descent in super neural networks. Preprint at https://arxiv.org/abs/1701.08734 (2017).
17. Delange, M. et al. A continual learning survey: defying forgetting in classification tasks. *IEEE Trans. Pattern Anal. Mach. Intell.* (2021).
18. Hadsell, R., Rao, D., Rusu, A. A. & Pascanu, R. Embracing change: continual learning in deep neural networks. *Trends Cogn. Sci.* **24**, 1028–1040 (2020).
19. Shuai, Y. et al. Forgetting is regulated through Rac activity in *Drosophila*. *Cell* **140**, 579–589 (2010).
20. Cohn, R., Morantte, I. & Ruta, V. Coordinated and compartmentalized neuromodulation shapes sensory processing in *Drosophila*. *Cell* **163**, 1742–1755 (2015).
21. Waddell, S. Neural plasticity: dopamine tunes the mushroom body output network. *Curr. Biol.* **26**, R109–R112 (2016).
22. Modi, M. N., Shuai, Y. & Turner, G. C. The *Drosophila* mushroom body: from architecture to algorithm in a learning circuit. *Annu. Rev. Neurosci.* **43**, 465–484 (2020).
23. Aso, Y. et al. Mushroom body output neurons encode valence and guide memory-based action selection in *Drosophila*. *eLife* **3**, e04580 (2014).
24. Aso, Y. & Rubin, G. M. Dopaminergic neurons write and update memories with cell-type-specific rules. *eLife* **5**, e16135 (2016).
25. Gao, Y. et al. Genetic dissection of active forgetting in labile and consolidated memories in *Drosophila*. *Proc. Natl Acad. Sci. USA* **116**, 21191–21197 (2019).
26. Zhao, J. et al. Genetic dissection of mutual interference between two consecutive learning tasks in *Drosophila*. *eLife* **12**, e83516 (2023).
27. Richards, B. A. & Frankland, P. W. The persistence and transience of memory. *Neuron* **94**, 1071–1084 (2017).
28. Dong, T. et al. Inability to activate Rac1-dependent forgetting contributes to behavioral inflexibility in mutants of multiple autism-risk genes. *Proc. Natl Acad. Sci. USA* **113**, 7644–7649 (2016).
29. Zhang, X., Li, Q., Wang, L., Liu, Z.-J. & Zhong, Y. Active protection: learning-activated Raf/MAPK activity protects labile memory from Rac1-independent forgetting. *Neuron* **98**, 142–155 (2018).
30. Davis, R. L. & Zhong, Y. The biology of forgetting—a perspective. *Neuron* **95**, 490–503 (2017).
31. Mo, H. et al. Age-related memory vulnerability to interfering stimuli is caused by gradual loss of MAPK-dependent protection in *Drosophila*. *Aging Cell* **21**, e13628 (2022).
32. Cervantes-Sandoval, I., Chakraborty, M., MacMullen, C. & Davis, R. L. Scribble scaffolds a signalosome for active forgetting. *Neuron* **90**, 1230–1242 (2016).
33. Noyes, N. C., Phan, A. & Davis, R. L. Memory suppressor genes: modulating acquisition, consolidation, and forgetting. *Neuron* **109**, 3211–3227 (2021).
34. Cognigni, P., Felsenberg, J. & Waddell, S. Do the right thing: neural network mechanisms of memory formation, expression and update in *Drosophila*. *Curr. Opin. Neurobiol.* **49**, 51–58 (2018).
35. Amin, H. & Lin, A. C. Neuronal mechanisms underlying innate and learned olfactory processing in *Drosophila*. *Curr. Opin. Insect Sci.* **36**, 9–17 (2019).
36. Handler, A. et al. Distinct dopamine receptor pathways underlie the temporal sensitivity of associative learning. *Cell* **178**, 60–75 (2019).
37. McCurdy, L. Y., Sareen, P., Davoudian, P. A. & Nitabach, M. N. Dopaminergic mechanism underlying reward-encoding of punishment omission during reversal learning in *Drosophila*. *Nat. Commun.* **12**, 1115 (2021).
38. Berry, J. A., Cervantes-Sandoval, I., Nicholas, E. P. & Davis, R. L. Dopamine is required for learning and forgetting in *Drosophila*. *Neuron* **74**, 530–542 (2012).

39. Berry, J. A., Phan, A. & Davis, R. L. Dopamine neurons mediate learning and forgetting through bidirectional modulation of a memory trace. *Cell Rep.* **25**, 651–662 (2018).
40. Aitchison, L. et al. Synaptic plasticity as bayesian inference. *Nat. Neurosci.* **24**, 565–571 (2021).
41. Schug, S., Benzing, F. & Steger, A. Presynaptic stochasticity improves energy efficiency and helps alleviate the stability–plasticity dilemma. *eLife* **10**, e69884 (2021).
42. Wang, L. et al. AFEC: active forgetting of negative transfer in continual learning. *Adv. Neural Inf. Process. Syst.* **34**, 22379–22391 (2021).
43. Benzing, F. Unifying importance based regularisation methods for continual learning. In *Proc. International Conference on Artificial Intelligence and Statistics* 2372–2396 (PMLR, 2022).
44. Bouton, M. E. Context, time, and memory retrieval in the interference paradigms of pavlovian learning. *Psychol. Bull.* **114**, 80 (1993).
45. Krizhevsky, A. et al. Learning multiple layers of features from tiny images. *Technical Report, Citeseer* (2009).
46. Shuai, Y. et al. Dissecting neural pathways for forgetting in *Drosophila* olfactory aversive memory. *Proc. Natl Acad. Sci. USA* **112**, E6663–E6672 (2015).
47. Chen, L. et al. AI of brain and cognitive sciences: from the perspective of first principles. Preprint at https://arxiv.org/abs/2301.08382 (2023).
48. Caron, S. J., Ruta, V., Abbott, L. F. & Axel, R. Random convergence of olfactory inputs in the *Drosophila* mushroom body. *Nature* **497**, 113–117 (2013).
49. Endo, K., Tsuchimoto, Y. & Kazama, H. Synthesis of conserved odor object representations in a random, divergent-convergent network. *Neuron* **108**, 367–381 (2020).
50. Long, M., Cao, Y., Wang, J. & Jordan, M. Learning transferable features with deep adaptation networks. In *Proc. International Conference on Machine Learning* 97–105 (PMLR, 2015).
51. Wang, L., Zhang, X., Li, Q., Zhu, J. & Zhong, Y. CoSCL: cooperation of small continual learners is stronger than a big one. In *Proc. European Conference on Computer Vision* 254–271 (Springer, 2022).
52. van de Ven, G. M., Tuytelaars, T. & Tolias, A. S. Three types of incremental learning. *Nat. Mach. Intell.* **4**, 1185–1197 (2022).
53. Riemer, M. et al. Learning to learn without forgetting by maximizing transfer and minimizing interference. In *International Conference on Learning Representations* (2018).
54. Schwarz, J. et al. Progress & compress: a scalable framework for continual learning. In *Proc. International Conference on Machine Learning* 4528–4537 (PMLR, 2018).
55. Jung, S., Ahn, H., Cha, S. & Moon, T. Continual learning with node-importance based adaptive group sparse regularization. *Adv. Neural Inf. Process. Syst.* **33**, 3647–3658 (2020).
56. Cha, S., Hsu, H., Hwang, T., Calmon, F. & Moon, T. CPR: classifier-projection regularization for continual learning. In *International Conference on Learning Representations* (2020).
57. Lake, B. M., Salakhutdinov, R. & Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science* **350**, 1332–1338 (2015).
58. Wah, C., Branson, S., Welinder, P., Perona, P. & Belongie, S. The Caltech-UCSD birds-200-2011 dataset. (2011). http://www.vision.caltech.edu/datasets/
59. Lomonaco, V. & Maltoni, D. Core50: a new dataset and benchmark for continuous object recognition. In *Conference on Robot Learning* 17–26 (PMLR, 2017).
60. Ryan, T. J. & Frankland, P. W. Forgetting as a form of adaptive engram cell plasticity. *Nat. Rev. Neurosci.* **23**, 173–186 (2022).
61. Luo, L. et al. Differential effects of the Rac GTPase on Purkinje cell axons and dendritic trunks and spines. *Nature* **379**, 837–840 (1996).
62. Tashiro, A., Minden, A. & Yuste, R. Regulation of dendritic spine morphology by the rho family of small gtpases: antagonistic roles of Rac and Rho. *Cerebral Cortex* **10**, 927–938 (2000).
63. Hayashi-Takagi, A. et al. Disrupted-in-Schizophrenia 1 (DISC1) regulates spines of the glutamate synapse via Rac1. *Nat. Neurosci.* **13**, 327–332 (2010).
64. Hayashi-Takagi, A. et al. Labelling and optical erasure of synaptic memory traces in the motor cortex. *Nature* **525**, 333–338 (2015).
65. Martens, J. & Grosse, R. Optimizing neural networks with kronecker-factored approximate curvature. In *Proc. International Conference on Machine Learning* 2408–2417 (PMLR, 2015).
66. Knoblauch, J., Husain, H. & Diethe, T. Optimal continual learning has perfect memory and is NP-hard. In *Proc. International Conference on Machine Learning* 5327–5337 (PMLR, 2020).
67. Deng, D., Chen, G., Hao, J., Wang, Q. & Heng, P.-A. Flattening sharpness for dynamic gradient projection memory benefits continual learning. *Adv. Neural Inf. Process. Syst.* **34**, 18710–18721 (2021).
68. Mirzadeh, S. I., Farajtabar, M., Pascanu, R. & Ghasemzadeh, H. Understanding the role of training regimes in continual learning. *Adv. Neural Inf. Process. Syst.* **33**, 7308–7320 (2020).
69. McAllester, D. A. PAC-Bayesian model averaging. In *Proc. Twelfth Annual Conference on Computational Learning Theory* 164–170 (ACM, 1999).
70. Pham, Q., Liu, C., Sahoo, D. & Steven, H. Contextual transformation networks for online continual learning. In *International Conference on Learning Representations* (2021).
71. Schulman, J., Wolski, F., Dhariwal, P., Radford, A. & Klimov, O. Proximal policy optimization algorithms. Preprint at https://arxiv.org/abs/1707.06347 (2017).
72. Lopez-Paz, D. et al. Gradient episodic memory for continual learning. *Adv. Neural Inf. Process. Syst.* **30**, 6467–6476 (2017).
73. Mnih, V. et al. Playing Atari with deep reinforcement learning. Preprint at https://arxiv.org/abs/1312.5602 (2013).
74. Wang, L. & Zhang, X. lywang3081/CAF: CAF paper. *Zenodo* https://doi.org/10.5281/zenodo.8293564 (2023).
75. Selvaraju, R. R. et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 618–626 (IEEE, 2017).

Acknowledgements

This work was supported by the National Key Research and Development Program of China (2020AAA0106302, to J.Z.), the STI2030-Major Projects (2022ZD0204900, to Y.Z.), the National Natural Science Foundation of China (nos 62061136001 and 92248303, to J.Z., 32021002, to Y.Z., U19A2081, to H.S.), the Tsinghua-Peking Center for Life Sciences, the Tsinghua Institute for Guo Qiang, and the High Performance Computing Center, Tsinghua University. L.W. was also supported by the Shuimu Tsinghua Scholar. J.Z. was also supported by the New Cornerstone Science Foundation through the XPLORER PRIZE.

Author contributions

L.W., X.Z., J.Z. and Y.Z. conceived the project. L.W., X.Z., Q.L. and M.Z. designed the computational model. X.Z. performed the theoretical analysis, assisted by L.W. L.W. performed all experiments and analysed the data. L.W., X.Z. and Q.L. wrote the paper. L.W., X.Z., Q.L., M.Z., H.S., J.Z. and Y.Z. revised the paper. J.Z. and Y.Z. supervised the project.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-023-00747-w.

Correspondence and requests for materials should be addressed to Jun Zhu or Yi Zhong.

Peer review information *Nature Machine Intelligence* thanks Gido van de Ven and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

nature portfolio

Corresponding author(s): Yi Zhong

Last updated by author(s): Sep 1, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All experiments are performed with public datasets, such as CIFAR-10, CIFAR-100, Omniglot, Tiny-ImageNet, CUB-200-2011, CORe50, etc. Please refer to our data availability statement.
Data analysis	We perform data analysis with both custom code (included in supplementary materials, please refer to our code availability statement) and Excel, and create all content figures with Prism 8.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All datasets used in this paper are publicly available. We include necessary download tools and instructions in our code.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	To obtain reliable results, all experiments are performed by more than 5 runs with different random seeds and task orders. This is a common choice following previous works in this field.
Data exclusions	No data were excluded in our analysis.
Replication	All results can be successfully replicated .
Randomization	N/A
Blinding	N/A

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging