Trends in Cognitive Sciences

CellPress REVIEWS

Series: Machine Behavior

Review

Embracing Change: Continual Learning in Deep Neural Networks

Raia Hadsell,[1,*,@] Dushyant Rao,[1,@] Andrei A. Rusu,[1,@] and Razvan Pascanu[1,@]

Artificial intelligence research has seen enormous progress over the past few decades, but it predominantly relies on fixed datasets and stationary environments. Continual learning is an increasingly relevant area of study that asks how artificial systems might learn sequentially, as biological systems do, from a continuous stream of correlated data. In the present review, we relate continual learning to the learning dynamics of neural networks, highlighting the potential it has to considerably improve data efficiency. We further consider the many new biologically inspired approaches that have emerged in recent years, focusing on those that utilize regularization, modularity, memory, and meta-learning, and highlight some of the most promising and impactful directions.

Highlights

Modern machine learning excels at training powerful models from fixed datasets and stationary environments, often exceeding human-level ability.

Yet, these models fail to emulate the process of human learning, which is efficient, robust, and able to learn incrementally, from sequential experience in a non-stationary world.

Insights into this limitation can be gleaned from the nature of neural network optimization, which implies that continual learning techniques could radically improve deep learning as well as open the door to new application areas.

Promising approaches for continual learning can be found at the most granular level, with gradient-based methods, as well as at the architectural level, with modular and memory-based approaches. We also consider meta-learning as a potentially important direction.

The World Is Not Stationary

A common benchmark for success in artificial intelligence is the ability to emulate human learning. We measure the abilities of humans to recognize images, play games, and drive a car, to name a few, and then develop machine learning models that can match or exceed these given enough training data. This paradigm puts the emphasis on the end result, rather than the learning process, and overlooks a critical characteristic of human learning: that it is robust to changing tasks and sequential experience. It is perhaps unsurprising that humans can learn this way, after all, time is irreversible and the world is **non-stationary** (see Glossary), so human learning has evolved to thrive in dynamic learning settings. However, this robustness is in stark contrast to the most powerful modern machine learning methods, which perform well only when presented with data that are carefully shuffled, balanced, and homogenized. Not only do these models underperform when presented with changing or incremental data regimes, in some cases they fail completely or suffer from rapid performance degradation on earlier learned tasks, known as **catastrophic forgetting**.

What might be gained by developing **neural network** models that learn sequentially like humans? First of all, many applications could benefit from continual adaptation to a changing target specification: for example, visual recognition algorithms that need to learn a diverse, growing set of image classes; or household robots that need to incrementally add skills to their repertoire. Continual learning techniques could enable models to acquire specialized solutions without forgetting previous ones, potentially learning over a lifetime, as a human does. In fact, continual learning is generally considered one of the attributes necessary for human-level artificial general intelligence [1]. More fundamentally, continual learning methods could offer enormous advantages for deep neural networks even in stationary settings, by improving learning efficiency as well as by enabling knowledge transfer between related tasks.

This article will first motivate a taxonomy of continual learning approaches through describing their connections with biological systems. Just as continual learning in humans cannot be reduced to a single biological mechanism, but is rather the product of multiple systems that range from the synaptic plasticity of single neurons to the entire memory system, continual

[1]DeepMind, London, UK

*Correspondence:
raia@google.com (R. Hadsell).
[@]Twitter: @raiahadsell (R. Hadsell), @drao64 (D. Rao), @andreialexrusu (A.A. Rusu), @rpascanu (R. Pascanu).

learning in artificial neural networks is not trivial to implement and will likely require a combination of strategies. Before exploring these different approaches in detail, we will delve into the dynamics of **gradient**-based learning algorithms to understand the challenges of continual learning with artificial neural networks.

Grounding Continual Learning in Biological Systems

The study of the natural world and its intelligent species has frequently intersected with artificial intelligence research, including in aspects pertaining to continual learning [1]. Biology offers an existence proof for successful continual learning in complex environments and also hints at the design principles and trade-offs of successful approaches. There are multiple mechanisms that enable biological systems to adapt in changing environments without intransigence or forgetting. Thus, in this section we introduce four continual learning paradigms through analogy to their biological equivalents, saving a fuller discussion of each approach for following sections. Additionally, the approaches can be succinctly described by depicting their canonical models, as in Figure 1 (Key Figure).

Gradient-based continual learning methods (see 'Gradient-Based Solutions') can be understood as distant relatives of models of synaptic plasticity. Mammalian brains have been shown to have complex mechanisms at the synaptic level, which protect against interference between old and new knowledge, or even conflicting facts and skills [2]. Synaptic plasticity has been studied for decades, with recent studies demonstrating remarkably precise and effective consolidation mechanisms occurring even over very short timescales [3]. Some continual learning approaches have successfully demonstrated simplified mechanisms built on such principles [4,5], but many core capabilities remain elusive.

Modularity is another paradigm that artificial systems might use to enable continual learning (see 'Modular Architectures'). From an evolutionary perspective, it is not coincidental that successful survival in increasingly complex environments is correlated with strong differentiation and specialization of nervous systems. Indeed, biological brains are modular, with distinct yet interacting subsystems (e.g., for memory or motor control). Evidence of modularity extends beyond anatomic features to functional separation in terms of sparse activations and hierarchical organization [6].

Memory systems in the brain are evidently critical for human learning and are the inspiration for memory-based continual learning in artificial neural networks (see 'Memory for Artificial Learning Systems'). Although the complex interactions between synaptic plasticity, episodic memory, and semantic memory are yet to be fully described by neuroscience, it is clear that memory is the bastion that protects human learning and adaptation over a long lifetime of varied experience [7].

Lastly, meta-learning for continual learning (see 'Meta-Learning: Discovering Inductive Biases for Continual Learning') is an approach that is motivated by the brain's ability to synthesize novel solutions after limited experience [8]. Through applying machine learning to optimize the learning approach itself, thus learning-to-learn, meta-learning hopes to achieve the same sort of rapid, general adaptation that biological systems demonstrate.

Defining Continual Learning

The problem of continual learning is typically defined by the sequential training protocol and by the features expected from the solution. In contrast to the common machine learning setting of a static dataset or environment, the continual learning setting explicitly focuses on non-stationary or changing environments, often divided into a set of tasks that need to be completed

Glossary

Backward transfer: the ability to transfer knowledge from a current task to improve performance on a previously learned task.
Catastrophic forgetting: a phenomenon observed in neural networks where learning a new task significantly degrades performance on previous tasks.
Credit assignment: determining how different parameters in a neural network are responsible for the desired behavior of the network.
Curriculum learning: the task of finding the optimal ordering of a series of tasks such that learning is cumulative and efficiency increases on the following tasks.
Descent direction: direction given by the negative gradient. If the parameters of the model are changed by taking a small step in this direction, the loss function will decrease.
Forward transfer: the ability to transfer knowledge from previous tasks to improve performance and learning efficiency on a related future task.
Gradient: the derivative of the loss function with respect to the parameters of the model, indicating the direction that learning needs to proceed in order to improve performance.
Gradient descent: given a continuous loss function, there exists a linear approximation where the slope in the high-dimensional case is called the gradient of the function. Moving along the gradient ensures minimizing the linear approximation and hence the true function as well. However this only holds locally, where the linear approximation is reliable, hence the iterative nature of the algorithm, as one can only take a small step in the descent direction, restricted by how much the approximation holds.
Independent and identically distributed (IID): an assumption underpinning much of the state of the art in modern machine learning.
Inductive biases: the set of assumptions made by a machine learning algorithm in order to generalize, such as the model architecture or assumptions about the target domain.
Loss function: a quantity that a model tries to minimize during learning, such as the error between true and predicted outputs.
Neural network: a mathematical model containing layers of units or 'neurons' connected with learnable

sequentially. This setting may vary in terms of task transitions (smooth or discrete), task length and repetition, and task type (such as unsupervised, supervised, or reinforcement learning), or it may not even have well-defined tasks [9–11]. Compared with **curriculum learning** [12,13], the learner does not control the task ordering.

While the problem setting is easy to describe, the characteristics of a desirable solution are not concise and include competing objectives. Consider a hypothetical robot that is expected to perform any household chore, in any home. The robot cannot be preprogrammed in the factory and then deployed, because of the sheer variety of tasks and homes. Rather, the robot will need to expand its skill set over time (e.g., learning to wash dishes, then tidying, and finally laundry). Of course, each task may have nontrivial variations: 'tidying' may mean cleaning up a board game or shelving books, and 'laundry' may require sorting socks or ironing shirts. To enable this, the robot will need to adapt quickly, and also not forget (at least not catastrophically). If there is forgetting, then fast recovery is critical. When learning related tasks (e.g., vacuuming, sweeping, and mopping), the robot should show **forward transfer** (better performance and faster learning on each subsequent task) and also show **backward transfer** (better performance on previous tasks, when revisited) because of transfer from the current task. Moreover, the robot will have limited access to previous tasks as well as limited capacity to store data, increase its model size, or increase processing time.

In summary, a continual learning solution would typically hope to satisfy many desiderata, illustrated in Figure 2 and defined in Box 1. Inspecting these desiderata reveals that continual learning often involves compromise between competing objectives. For instance, maintaining perfect recall (by forgetting nothing) in a fixed-capacity model is impossible given an arbitrarily long sequence of tasks. This dilemma motivates the alternative objective of fast recovery, which allows forgetting if previous performance levels can be recovered with a minimal amount of new experience. Forward, and in particular backward, transfer contrasts with the ability to perfectly recall previous tasks. Thus, any solution needs to balance competing needs. But what constitutes an optimal trade-off? How much should the model remember and how much is the model allowed to grow? Some of the justifications and trade-offs can only be settled if we ground continual learning into a specific domain. For this reason, studying continual learning in isolation can be challenging and it remains a priority to identify realistic settings.

The Independent and Identically Distributed Assumption Underpinning Modern Machine Learning

Neural networks heavily exploit modern technology to parallelize computations and consider large amounts of data at once; in fact, this ease of scaling has allowed them to become the *de facto* approach for speech, vision, and language applications in the last decade. In a typical learning setting, the goal is to set the parameters of the network in order to minimize some **loss function**, such as the error between true and predicted outputs. Gradient-based learning, the most efficient and widely used paradigm, is an iterative algorithm that, at each iteration, makes a small change to the parameters in order to reduce the loss (for a more detailed explanation, see Box 2). The mechanics of this rule results in a tug-of-war dynamic, where each data sample is trying to pull on each parameter to make it larger or smaller. By averaging gradients, we thus create a tug-of-war game where the update that is applied to each parameter, since it is either positive or negative, reveals which data samples won or lost. Combining many tug-of-war updates over many optimization steps allows learning to progress (Figure 3).

As a result, for learning to be successful, data samples must be **independent and identically distributed (IID)**. To illustrate this, consider trying to learn two different tasks. Though the tasks will not agree on how to set all parameters, the tug-of-war dynamics will eventually lead to equilibrium.

weights and biases, which process some given input and produces an output. It is loosely based on neurons in biological brains.
Non-stationary: describing a process whose state or probability distribution changes with time.
Sparse representation or gradients: a sparse vector (or tensor) has many zero-valued entries. This can be the case for internal network activations, or for gradients. For the latter, sparsity implies that few parameters are changed during each learning update.

Key Figure

Paradigms for Continual Learning

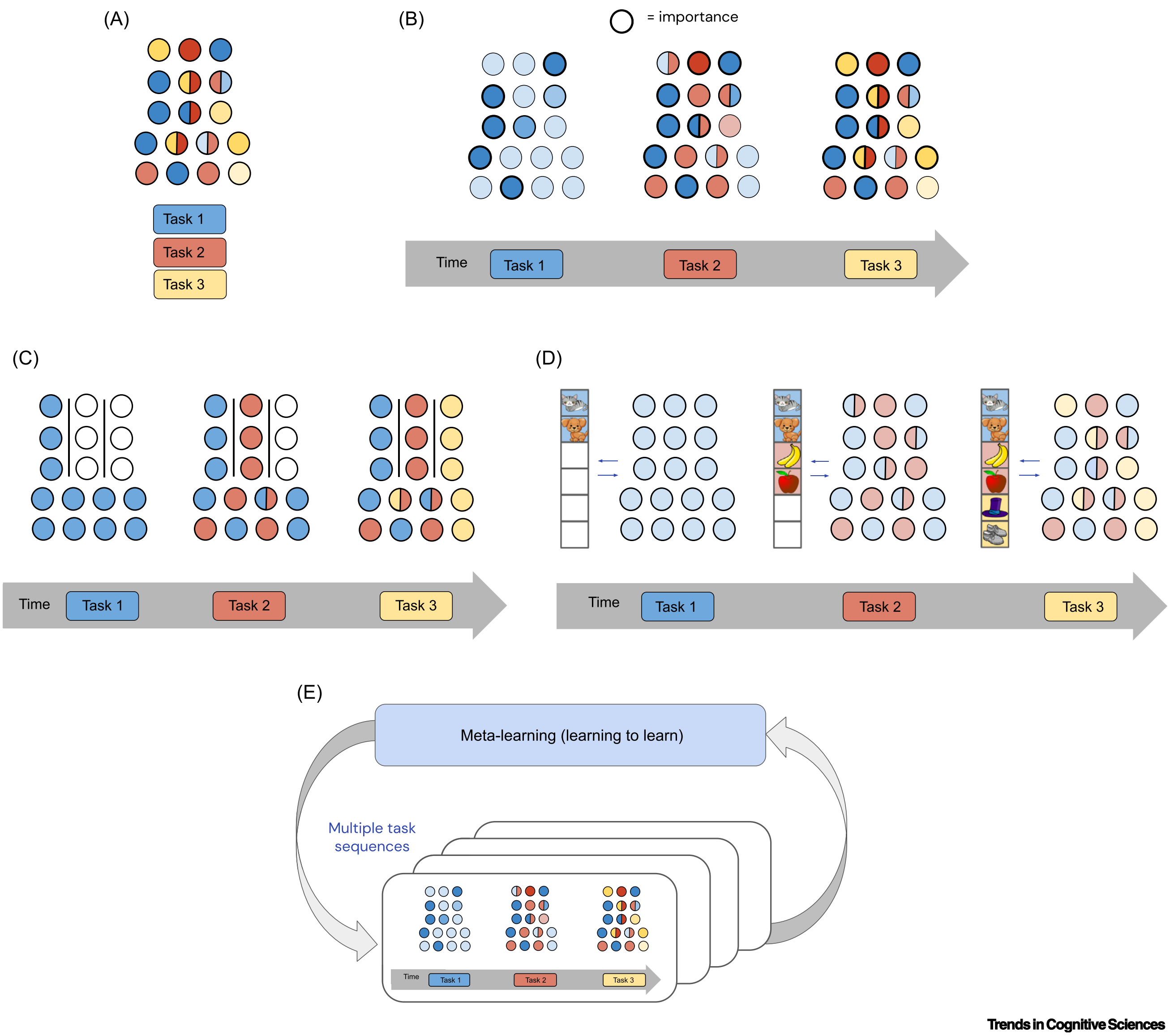
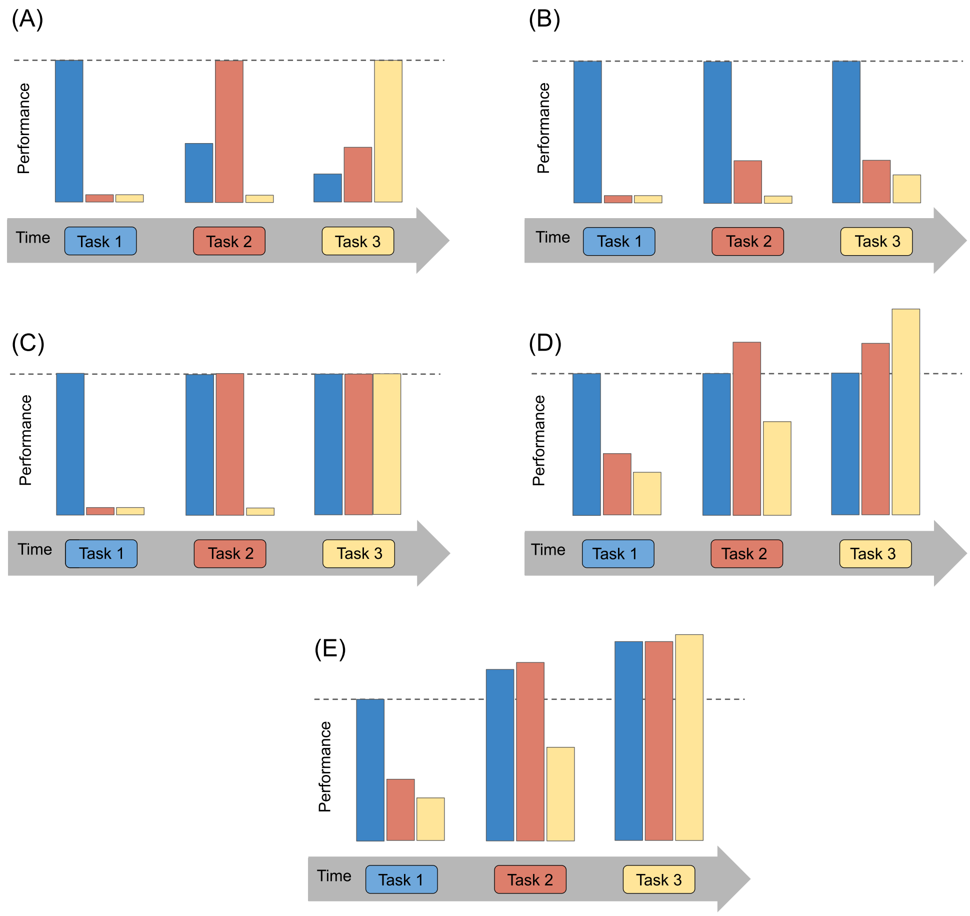


Figure 1. (A) Independent and identically distributed learning methods are standard for nonsequential, multitask learning. In this regime, tasks are learned simultaneously to avoid forgetting and instability. (B) Gradient-based approaches preserve parameters based on their importance to previously learned tasks. (C) Modularity-based methods define hard boundaries to separate task-specific parameters (often accompanied by shared parameters to allow transfer). (D) Memory-based methods write experience to memory to avoid forgetting. (E) Meta-learning techniques optimize continual learning 'meta-objectives' over a large set of task sequences, thereby learning to continually learn.

However, if for many consecutive iterations one task is absent, the other task will quickly control all parameters. Therefore gradient-based learning requires that, in expectation, all tasks are always present to create the tension needed for learning to progress and eventually converge, hence, the IID assumption (Figure 1A).



Trends in Cognitive Sciences

Figure 2. Illustrations of Different Outcomes in a Sequential, Continual Learning Setting. Each plot (A–E) shows the performance on three tasks which are evaluated three times (train on task 1, test on 1, 2, and 3; train on task 2, test on 1, 2, and 3; etc.). The plots, which are illustrative, depict the following conditions: (A) catastrophic forgetting occurs when the performance on previous tasks degrades sharply when learning new tasks; (B) too little plasticity means that only the first task is learned (e.g., because of over-regularization or lack of capacity). The next three plots show the cumulative effect of positive learning conditions: (C) first we see the result if no forgetting happens and thus the performance on previous tasks can be maintained during learning; (D) in addition to no forgetting, forward transfer is the desirable condition where previously learned tasks improve both the performance and learning efficiency on related future tasks; (E) lastly we show what happens if there is also backward transfer, thus learning a given task improves performance of previous tasks without further training.

The tug-of-war dynamics are the main mechanism through which gradient-based methods handle **credit assignment**, which leads to the specialization of parameters and, subsequently, to learning. Continual learning can be understood as looking for an alternative mechanism to do credit assignment, one that avoids or modifies the tug-of-war dynamics such that all tasks do not need to be simultaneously present.

It is worth considering how efficient tug-of-war learning is, in particular, we can ask whether all tasks are being learned at the same pace. Maybe not surprisingly, the answer is no. Recent work shows that most examples in a dataset are learned relatively fast and the model needs multiple repetitions to learn the harder examples [14–16]. However, the tug-of-war dynamics require that all examples are present, even the easier ones, which wastes computational resources. Recent work has shown empirically that concepts are discovered sequentially, even if they are simultaneously present in the data [17], and the behavior is predicted analytically for deep linear models

Box 1. Desiderata of Continual Learning

Continual learning methods necessarily involve balancing competing objectives. Mitigating catastrophic forgetting is often prioritized in research proposals, but actually all of the following desiderata are critical for many real world application domains:

- Minimal access to previous tasks. The model does not have infinite storage for previous experience and, crucially, it can not interact with previously seen tasks.
- Minimal increase in model capacity and computation. The approach must be scalable: it cannot add a new model for each subsequent task.
- Minimizing catastrophic forgetting and interference. Training on new tasks should not significantly reduce performance on previously learned tasks (Figure 2A,C).
- Fast adaptation and recovery. The model should be capable of fast adaptation to novel tasks or domain shifts and of fast recovery when presented with past tasks.
- Maintaining plasticity. The model should be able to keep learning effectively as new tasks are observed (Figure 2B).
- Maximizing forward and backward transfer. Learning a task should improve related tasks, both past and future, in terms of both learning efficiency and performance (Figure 2D,E).
- Task-agnostic learning. The approach should not rely on known task labels or task boundaries.

[18]. Therefore, even if tasks are equally complex and presented simultaneously, the model might still learn them sequentially, thus losing efficiency due to the tug-of-war dynamics. By proposing alternative ways of assigning credit, continual learning could unleash unprecedented learning efficiency, even in stationary learning settings.

Gradient-Based Solutions

Motivated by the tug-of-war learning dynamics described previously, one promising approach is to directly modulate the gradients of different tasks. Not only does this go to the heart of the optimization problem, it also is well-motivated by studies of synaptic consolidation in biological brains [3].

One approach is to force the gradient to stay aligned with gradients from previously learned tasks [19,20], eliminating potential interference. Such methods can be beneficial in other settings as well, for example, in multitask learning, where they have the potential to make learning more efficient in the case of conflicting objectives [21–23].

Other approaches focus on regularizing the loss on new tasks to minimize forgetting of previous tasks. These regularization-based methods estimate the importance of each model parameter for previous tasks and penalize changes to each parameter proportional to this measure (Figure 1B) [5,24–28]. For example, elastic weight consolidation [5] relies on the Fisher information matrix to measure the sensitivity of the parameters with respect to each task and to indicate which parameters most need to be preserved to avoid forgetting. The regularization term acts as a proxy for the gradients of previous tasks, ensuring that the equilibrium required by the tug-of-war dynamics is maintained. Note that data-agnostic regularization terms, such as drop-out [29], and the training regime in general [30,31] can have similar effects of altering how credit assignment is done.

Another set of techniques use knowledge distillation [32] as a way to preserve a model's functionality with respect to previous tasks. The learning without forgetting approach [33] encourages the function of previous task layers to be consistent even when learning a new task. This is achieved by taking a snapshot of the network (shared parameters and previous task layers) before each new task and using an additional distillation loss to maintain consistency.

One weakness of gradient-based methods is their reliance on approximations of the objective they want to attain, which means that they are likely to fail once the model is presented with many tasks or when the tasks in the sequence are more diverse. However, since these approaches center

Box 2. Gradient-Based Learning and Tug-of-War Dynamics

Gradient-based algorithms adjust the parameters of a model iteratively, by taking a small step in a **descent direction**. The descent direction is given by the gradient, which can be seen as independently determining for each parameter whether increasing or decreasing its value will reduce the loss. The individual gradients are computed and averaged over a number of data points, called a batch. Finally,the model is updated by taking a small step in the negative direction of this average gradient. Figure 3A illustrates this algorithm. The contour lines indicate the value of the loss function corresponding to a hypothetical task, as a function of the parameters of a model (in this case, two-dimensional, w_1 and w_2). Starting from a random initialization, the learning procedure iteratively computes the local gradient of the loss function (which indicates the direction of greatest change) and changes the parameters by taking a small step in this direction. After many iterations the procedure converges to a local minimum of the loss function.

Next we can look at what happens during subsequent learning of a second task. Figure 3B illustrates a second loss function (overlaid on the first for clarity) and the optimization trajectory towards the minimum of task 2. Note that the solution learned for task 2 has a high loss for task 1: this illustrates catastrophic forgetting, where naive sequential training on a set of tasks can lead to poor performance on earlier tasks.

If the setting is changed so that both tasks are optimized together, by summing the losses into a 'multitask' optimization, as shown in Figure 3C, we see the 'tug-of-war' dynamics that result in stable, but inefficient, steps towards the multitask solution. In this illustration, the optimization trajectory is shown in black and the individual gradients for each task are shown in red and blue. The gradient from each task pulls the solution towards its optimum and the result is an equilibrium between the gradients of different tasks. Of course, a continual learning setting precludes the ability to train on all tasks simultaneously, but it may be valuable for a continual learning approach to recreate the careful tug-of-war dynamics by directly modifying the gradients, or by using memory to balance the current task with data from previously seen tasks.

on the gradient tug-of-war, they may allow us to better understand the continual learning problem and credit assignment in general.

Modular Architectures

Modular neural network architectures are a natural and effective solution to the problems of interference and catastrophic forgetting in continual learning. Modularity offers a compromise between using a single monolithic network, which is susceptible to forgetting, and using independent networks for each task, which precludes catastrophic forgetting but also prevents transfer between tasks (see Figure 1C for illustration of a modular architecture). Modularity is evident in biological systems as well, where it supports functional specialization of brain regions.

Trivially, catastrophic forgetting can be prevented by using disjoint models for different tasks, but this implies unconstrained model growth, prevents transfer between tasks, and assumes well-

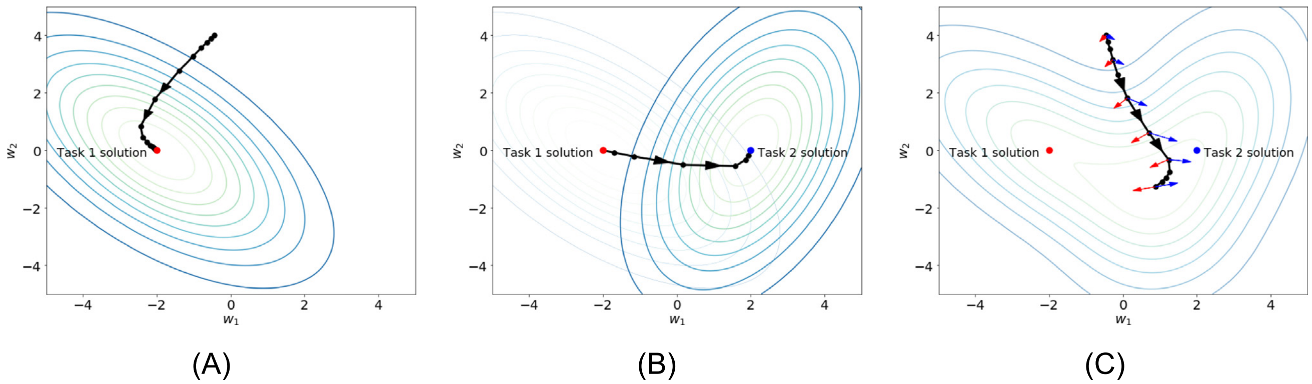


Figure 3. Illustrations of Gradient Descent Optimization for Different Tasks. (A) The trajectory taken by gradient descent optimization when minimizing a loss corresponding to a single task. (B) The optimization trajectory when subsequently training the same model on a second task. (C) The trajectory taken when using the total loss from both tasks (black) and the gradients from each individual task at multiple points during optimization (red and blue). See Box 2 for more detailed discussion.

CellPress REVIEWS

defined boundaries and known tasks. An ideal compromise is to exploit modular components by reusing them across tasks and scenarios, with a learning update that only affects a subset of parameters. This is similar to regularization-based methods, but with a hard constraint on which neurons or modules are updated for each task.

Modular models often use explicit procedures to increase capacity by adding new parameters as new tasks are encountered. Recently, a number of adaptive network architectures have emerged that add new capacity while protecting and reusing existing representations [34,35]; this, however, introduces new challenges, such as increasing computational requirements throughout learning. Other solutions judiciously add new capacity only when absolutely needed [36–38], expand more aggressively and then prune or compress parts of the model [39,40], or channel learning of new tasks to unused parts of a large fixed model [41–45].

Modularity can also manifest as a hierarchical structure; this can offer both protection and specialization at different levels of abstraction. For instance, a hierarchical control method for a robot might be modular at a high level (behaviors and goals) as well as a low level (motion primitives), and thus enable faster learning of sequential tasks. This idea was introduced in earlier work in continual learning [46] (see also M.B. Ring, PhD thesis, University of Texas at Austin, 1994) and has shown promise with new research [47,48], but deserves further study at large scale.

Another way to understand modularity is from the perspective of sparsity, which has been explored extensively from a compression and efficiency angle [49–54] as well as for its applicability to reinforcement learning and control [55]. From a continual learning perspective [56], both **sparse representations and gradients** will result in less interference and forgetting, as there will either be fewer task-sensitive parameters or fewer effective changes to parameters. Sparsity can also lead to the emergence of modules without requiring a predefined modular architecture, which we regard as an important avenue for future research.

The paradigm of modularity, especially when coupled with sparsity constraints and hierarchical abstraction, offers a pragmatic yet powerful means to solve continual learning. Looking into the future, human-level artificial intelligence will require the ability to specialize and the ability to compose skills; modular continual learning describes a path towards both.

Memory for Artificial Learning Systems

Gradient-based and modular methods may be more suitable for short-term continual learning rather than long-term retention. Gradient-based methods cannot prevent forgetting over arbitrarily long task sequences and while modular approaches can preserve knowledge over long timescales, they may reach practical limits in terms of neural network capacity. Consider the challenging scenario of hiding food in 1000 different locations over the course of months, then locating each cache correctly after more months have elapsed, a feat that is performed every winter by birds such as nuthatches, jays, and corvids [57]. Preserving the sequential experience of caching food by adapting the parameters of a simple neural network would be both challenging and inefficient. A more scalable strategy would be to encode the spatial locations with a dedicated read-and-write memory.

Taking inspiration from biology, we consider a more ambitious solution to the continual learning problem: to implement a neural network memory that can encode, store, and recall knowledge or experience. An artificial memory is potentially more scalable for long-term recall, but it comes with the additional challenge of designing, or preferably learning, a framework for encoding, querying, and writing information, in a way that generalizes across tasks. The simplest implementation of memory for continual learning is often called replay, or rehearsal: the idea is to maintain a history of observations and then sample learning targets from this buffer (as well as from current observations), thus preventing catastrophic forgetting though continual rehearsal of previously seen tasks [58,59] (Figure 1D). In this category are also episodic memory methods, which are distinguished from rehearsal methods because they not only use replay memory for training, but also for inference [60–62]. Rehearsal and episodic memory are simple and remarkably effective at reducing forgetting, but they do not scale well.

Rather than storing all observations, one can maintain a set of anchors, exemplars, or memory vectors that represent the key features of each task. Not only is this style of memory more efficient and scalable, but it enables compression and high-level transfer across multiple tasks. Of course, the critical challenge in a sparse memory setting becomes one of selecting which experiences to store and this has been the focus of a number of recent works [63–67].

From a biological perspective one might note that although replay has been observed in rodents and humans [68], maintaining any number of pristine observations or exemplars is unrealistic. In generative memory methods, no samples are stored. Rather, generative models are trained and then used to generate rehearsal data as needed [69–72].

Finally, we consider methods with learned read and write operations. One such model, the differentiable neural computer (DNC) [73], uses end-to-end gradient-based learning to simultaneously train separate neural networks to encode observations, read from the memory, and write to the memory. For continual learning problem domains, a DNC could hypothetically learn how to select, encode, and compress knowledge for efficient storage and retrieve it for maximum recall and forward transfer. The generality of the approach presents a dilemma, however, since training this sort of architecture is extremely difficult even in stationary environments or from IID datasets. Moreover, the neural network components (for reading, writing, etc.) may themselves suffer from catastrophic forgetting during training.

Regardless of the challenges, memory frameworks are clearly valuable for continual learning and the most general, end-to-end models have the potential to open up new frontiers in the field, but only if we can overcome the challenges of training such systems.

Meta-Learning: Discovering Inductive Biases for Continual Learning

All of the solutions discussed thus far prescribe hand-engineered mechanisms or architectures, **inductive biases**, for continual learning. Each inductive bias strikes a different trade-off between desiderata, such as good knowledge retention versus positive forward transfer in a memory-based approach. It is worth considering whether better trade-offs can be achieved by learning a solution from data rather than relying on human ingenuity to design it. Historically, a number of ‘meta-learning’ or ‘learning-to-learn’ approaches have demonstrated that solutions can be improved by automatically learning inductive biases (such as architecture, data, and learning parameters) that would otherwise need to be hand-designed (Figure 1E) [74,75].

Intuitively, meta-learning approaches can be described as comprising two timescales of optimization: an ‘inner loop’ that optimizes on specific tasks and an ‘outer loop’ that optimizes performance over multiple inner loops (Box 3). Relevant to continual learning, outer loops can be defined that optimize for performance in non-stationary settings [76–78]. For example, online-aware meta-learning (OML) [79] uses inner-loops that learn on correlated sequences of inputs. Although this violates the IID assumption and could result in catastrophic forgetting, the outer-loop optimizes the input representation to reduce forgetting and improve generalization during

Box 3. Understanding Meta-Learning

Meta-learning algorithms can be understood in terms of adaptation at two different time scales. Intervals of fine-grained task learning, or 'inner-loops', provide the necessary information for an 'outer-loop' of coarse-grained meta-learning. The aim of the outer-loop is to improve adaptation in future inner-loops. Crucially, different data are used during training and testing to ensure generalization at both time scales.

Meta-learning algorithms can be differentiated by their definition or implementation of the inner loop, which allows adaptation to specific tasks, or the outer loop, which optimizes across a number of inner loops. In [86,87] both loops use **gradient descent**, while inner-loops can also be nearest-neighbor matching [88,89], recurrent neural representations [8,90,91], probabilistic inference [82,92], or mixed approaches [93–96]. In all these works, the outer-loop aims to improve the speed of inner-loop adaptation to a new task.

the non-stationary inner-loops. Interestingly, minimizing the OML objective leads to sparse input representations. Furthermore, an efficient partitioning of representation space arises, reminiscent of modularity. Exemplifying the value of meta-learning, neither principle is explicitly enforced, yet they are discovered and exploited for effective continual learning.

From another perspective, every sequential task can be considered as a meta-learning problem [80–82,97]. As pointed out in [83], this perspective can be insightful for continual learning. In the IID setting there is an underlying assumption that a solution exists that can solve all tasks simultaneously. However, note that in natural settings there are many instances of tasks that are inconsistent or contradictory (changing foraging behavior depending on the season, driving on the left or the right side of the street). Instead, the meta-learning perspective moves the emphasis to fast adaptation and fast recovery rather than perfect recall. However this trade-off is not always appropriate and depends on the tasks being solved. There are behaviors that one cannot afford to forget for reasons of safety or fairness, or in settings where fast relearning is impossible, for instance, because of the rarity of the experience (touching a hot stove is an example of an experience that should not need refreshing!).

While the prospect of meta-learning new solutions for continual learning is exciting, this has proven to be computationally demanding as well as requiring careful design of the task distribution [84,85]. That is, while meta-learning can remove some of the hand-engineering required in machine learning, this is displaced by hand-engineering the tasks themselves.

Looking past these hurdles, meta-learning could steer continual learning research beyond the goal of knowledge retention with perfect immediate recall, towards a more realistic notion of flexible, data-efficient learning in non-stationary domains.

Concluding Remarks and Future Directions

Machine learning researchers often point to the remarkable ability of humans to learn quickly and generalize robustly (e.g., inferring a pattern from a few examples). However, we do not often remark on the ability of humans to learn continually over a lifetime of education and experience, although it is this facility that enables human achievements of science, art, and industry. This article attempts not only to highlight the importance of continual learning, but also to expose the limitations of modern neural networks in this regard, in particular the credit assignment problem that results in an inefficient, gradient-based 'tug-of-war'.

Surveying the solution space, we have identified learning paradigms that have the potential to be truly impactful if scaled to more ambitious domains. Not surprisingly, these paradigms all have strong parallels in neuroscience and biological systems. Gradient-based approaches directly modify the optimization of neural networks and have been shown to reduce catastrophic forgetting. Modular

Outstanding Questions

Can we find an alternative solution to credit assignment that preserves the ability of gradient-based approaches to learn complex tasks well while avoiding the tug-of-war dynamic and improving the efficiency of learning?

What are the right trade-offs between the different desiderata and can we quantify the trade-offs of biological systems? If these are domain specific, what are all the different domains or types of continual learning problems that we should explore?

How do we balance the need to remember versus the ability to quickly relearn certain facts or skills? When is perfect recall necessary?

If the modular structure of a continual learning system has to be designed (rather than learned), what should it be? How might it evolve over time?

How can we learn efficiently how to retrieve or store memories to address continual learning? How might we ensure that these mechanisms themselves are robust to change in data distribution? And, particularly when dealing with generative models that learn slowly, how do we ensure that memories are formed fast enough? How do we compress and encode memories?

Can we learn the inductive biases needed to solve continual learning? Can we do this at scale, efficiently? Is constructing the right data and learning environment to learn the inductive biases as hard as solving the original problem? Would the learned solution be robust or interpretable?

architectures offer pragmatic solutions to interference and catastrophic forgetting, while enabling forward transfer through hierarchical recomposition of skills and knowledge. End-to-end memory models could be a scalable solution for long timescale learning, and meta-learning approaches could surpass hand-designed algorithms and architectures altogether.

With such potential for positive impact, it is important to also acknowledge the risks involved in deploying machine learning models that continually change, since any initial assessment of safe and expected behavior cannot be readily guaranteed in perpetuity. Continual learning solutions, however, could mitigate these risks through improving the long-term reliability of the learning algorithm and through developing architectures that ensure that certain rules or boundaries are never violated.

References

1. Hassabis, D. *et al.* (2017) Neuroscience-inspired artificial intelligence. *Neuron* 95, 245–258
2. Wixted, J.T. (2004) The psychology and neuroscience of forgetting. *Annu. Rev. Psychol.* 55, 235–269
3. Clopath, C. *et al.* (2008) Tag-trigger-consolidation: a model of early and late long-term-potentiation and depression. *PLoS Comput. Biol.* 4, e1000248
4. Kaplanis, C. *et al.* (2018) Continual reinforcement learning with complex synapses. *arXiv* Published online February 20, 2018. https://arxiv.org/abs/1802.07239
5. Kirkpatrick, J. *et al.* (2017) Overcoming catastrophic forgetting in neural networks. *Proc. Nat. Acad. Sci. U. S. A.* 114, 3521–3526
6. Meunier, D. *et al.* (2010) Modular and hierarchically modular organization of brain networks. *Front. Neurosci.* 4, 200
7. Tulving, E. *et al.* (1972) Episodic and semantic memory. *Organ. Memory* 1, 381–403
8. Wang, J.X. *et al.* (2016) Learning to reinforcement learn. *arXiv* Published online November 17, 2016. https://arxiv.org/abs/1611.05763
9. Aljundi, R. *et al.* (2019) Task-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11254–11263
10. Aljundi, R. *et al.* (2019) Gradient based sample selection for online continual learning. *arXiv* Published online March 20, 2019. https://arxiv.org/abs/1903.08671
11. Zeno, C. *et al.* (2018) Task agnostic continual learning using online variational bayes. *arXiv* Published online March 27, 2018. https://arxiv.org/abs/1803.10123
12. Bengio, Y. *et al.* (2009) Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09*, pp. 41–48
13. Graves, A. *et al.* (2017) Automated curriculum learning for neural networks. In *Proceedings of the 34th International Conference on Machine Learning*
14. Devansh, A. *et al.* (2017) A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning*
15. Chang, H.-S. *et al.* (2017) Active bias: training more accurate neural networks by emphasizing high variance samples. In *Advances in Neural Information Processing Systems*, pp. 30 (Poster)
16. Shalev-Shwartz, S. (2014) Selfieboost: a boosting algorithm for deep learning. *arXiv* Published online November 13, 2014. https://arxiv.org/abs/1411.3436
17. Raposo, D. *et al.* (2017) Discovering objects and their relations from entangled scene representations. *arXiv* Published online February 16, 2017. https://arxiv.org/abs/1702.05068
18. Saxe, A.M. *et al.* (2014) Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *Proceedings of the International Conference on Learning Representations*, ICLR
19. Lopez-Paz, D. *et al.* (2017) Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 30
20. Chaudhry, A. *et al.* (2018) Efficient lifelong learning with A-GEM. *arXiv* Published online December 2, 2018. https://arxiv.org/abs/1812.00420
21. Du, Y. *et al.* (2018) Adapting auxiliary losses using gradient similarity. *arXiv* Published online December 5, 2018. https://arxiv.org/abs/1812.02224
22. Tianhe, Y. *et al.* (2020) Gradient surgery for multi-task learning. *arXiv* Published online January 19, 2020. https://arxiv.org/abs/2001.06782
23. Madry, A. *et al.* (2018) Towards deep learning models resistant to adversarial attacks. In *Proceedings of the International Conference on Learning Representations*
24. Zenke, F. *et al.* (2017) Continual learning through synaptic intelligence. *Proc. Mach. Learn Res.* 70, 3987–3995
25. Aljundi, R. *et al.* (2018) Memory aware synapses: learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 139–154
26. Cuong, V. *et al.* (2017) Variational continual learning. *arXiv* Published online October 29, 2017. https://arxiv.org/abs/1710.10628
27. Farquhar, S. and Gal, Y. (2018) Towards robust evaluations of continual learning. *arXiv* Published online May 24, 2018. https://arxiv.org/abs/1805.09733
28. Chen, P.H. *et al.* (2019) Overcoming catastrophic forgetting by generative regularization. *arXiv* Published online December 3, 2019. https://arxiv.org/abs/1912.01238
29. Goodfellow, I.J. *et al.* (2013) An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv* Published online December 21, 2013. https://arxiv.org/abs/1312.6211
30. Mirzadeh, S. *et al.* (2020) Dropout as an implicit gating mechanism for continual learning. *arXiv* Published online April 24, 2020. https://arxiv.org/abs/2004.11545
31. Mirzadeh, S. *et al.* (2020) Understanding the role of training regimes in continual learning. *arXiv* Published online June 12, 2020. https://arxiv.org/abs/2006.06958
32. Hinton, G. *et al.* (2015) Distilling the knowledge in a neural network. *arXiv* Published online March 9, 2015. https://arxiv.org/abs/1503.02531
33. Li, Z. and Hoiem, D. (2017) Learning without forgetting. In *IEEE Transactions on Pattern Analysis and Machine Intelligence* (40), pp. 2935–2947
34. Terekhov, A.V. *et al.* (2015) Knowledge transfer in deep block-modular neural networks. In *Biomimetic and Biohybrid Systems*, pp. 268–279, Springer International Publishing
35. Rusu, A.A. *et al.* (2016) Progressive neural networks. *arXiv* Published online June 15, 2016. https://arxiv.org/abs/1606.04671
36. Yoon, J. *et al.* (2018) Lifelong learning with dynamically expandable networks. *arXiv* Published online August 4, 2017. https://arxiv.org/abs/1708.01547
37. Draelos, T.J. *et al.* (2017) Neurogenesis deep learning: extending deep networks to accommodate new classes. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pp. 526–533
38. Xu, J. and Zhanxing, Z. (2018) Reinforced continual learning. In *Advances in Neural Information Processing Systems*, pp. 31
39. Schwarz, J. *et al.* (2018) Progress & compress: a scalable framework for continual learning. In *Proceedings of the International Conference on Machine Learning*, pp. 4535–45442

40. Golkar, S. *et al.* (2019) Continual learning via neural pruning. *arXiv* Published online March 11, 2019. https://arxiv.org/abs/1903.04476
41. Fernando, C. *et al.* (2017) PathNet: Evolution channels gradient descent in super neural networks. *arXiv* Published online January 30, 2017. https://arxiv.org/abs/1701.08734
42. He, X. and Jaeger, H. (2017) Overcoming catastrophic interference by conceptors. *arXiv* Published online July 16, 2017. https://arxiv.org/abs/1707.04853
43. Serra, J. *et al.* (2018) Overcoming catastrophic forgetting with hard attention to the task. In *Proceedings of the International Conference on Machine Learning*
44. Rajasegaran, J. *et al.* (2019) Random path selection for continual learning. In *Advances in Neural Information Processing Systems*, pp. 32
45. Adel, T. *et al.* (2020) Continual learning with adaptive weights (claw). In *Proceedings of the International Conference on Learning Representations*
46. Ring, M.B. (2015) Child: a first step towards continual learning. *Mach. Learn.* 28, 77–104
47. Haarnoja, T. *et al.* (2018) Latent space policies for hierarchical reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*
48. Tirumala, D. *et al.* (2019) Exploiting hierarchy for learning and transfer in KL-regularized RL. *arXiv* Published online March 16, 2019. https://arxiv.org/abs/1903.07438
49. Kalchbrenner, N. *et al.* (2018) Efficient neural audio synthesis. In *Proceedings of the International Conference on Machine Learning (ICML)*
50. Molchanov, D. *et al.* (2017) Variational dropout sparsifies deep neural networks. In *Proceedings of the 34th International Conference on Machine Learning*
51. Kingma, D.P. *et al.* (2018) Learning sparse neural networks through L_0 regularization. In *Proceedings of the International Conference on Learning Representations*
52. Mocanu, D.C. *et al.* (2018) Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science. *Nat. Commun.* 9, 2383
53. Bellec, G. *et al.* (2018) Deep rewiring: training very sparse deep networks. In *Proceedings of the International Conference on Learning Representations*
54. Mostafa, H. and Wang, X. (2019) Parameter efficient training of deep convolutional neural networks by dynamic sparse reparameterization. In *Proceedings of the 36th International Conference on Machine Learning*
55. Liu, V. *et al.* (2019) The utility of sparse representations for control in reinforcement learning. In *The Thirty-Third AAAI Conference on Artificial Intelligence*, pp. 4384–4391, AAAI Press
56. Aljundi, R. *et al.* (2018) Selfless sequential learning. *arXiv* Published online June 14, 2018. https://arxiv.org/abs/1806.05421
57. Sherry, D.F. and Hoshooley, J.F. (2010) Seasonal hippocampal plasticity in food-storing birds. *Phil. Trans. R. Soc.* 365, 933–943
58. Robins, A. (1995) Catastrophic forgetting, rehearsal and pseudorehearsal. *Connect. Sci.* 7, 123–146
59. Rolnick, D. *et al.* (2019) Experience replay for continual learning. In *Advances in Neural Information Processing Systems*
60. Chaudhry, A. *et al.* (2019) Continual learning with tiny episodic memories. *arXiv* Published online February 27, 2019. https://arxiv.org/abs/1902.10486
61. d'Autume, C.M. *et al.* (2019) Episodic memory in lifelong language learning. In *Advances in Neural Information Processing Systems*, pp. 32
62. Sprechmann, P. *et al.* (2018) Memory-based parameter adaptation. In *Proceedings of the International Conference on Learning Representations*
63. Rebuffi, S.A. *et al.* (2017) iCaRL: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2001–2010
64. Aljundi, R. *et al.* (2019) Online continual learning with maximal interfered retrieval. In *Advances in Neural Information Processing Systems*
65. Aljundi, R. *et al.* (2019) Gradient based sample selection for online continual learning. In *Advances in Neural Information Processing Systems*, pp. 32
66. Titsias, M.K. *et al.* (2019) Functional regularisation for continual learning using Gaussian processes. *arXiv* Published online January 31, 2019. https://arxiv.org/abs/1901.11356
67. Chaudhry, A. *et al.* (2020) Using hindsight to anchor past knowledge in continual learning. *arXiv* Published online February 19, 2020. https://arxiv.org/abs/2002.08165
68. Liu, Y. *et al.* (2019) Human replay spontaneously reorganizes experience. *Cell* 178, 640–652
69. Shin, H. *et al.* (2017) Continual learning with deep generative replay. In *Advances in Neural Information Processing Systems*
70. Rannen, A. *et al.* (2017) Encoder based lifelong learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1320–1328
71. Achille, A. *et al.* (2018) Life-long disentangled representation learning with cross-domain latent homologies. In *Advances in Neural Information Processing Systems*
72. Rao, D. *et al.* (2019) Continual unsupervised representation learning. In *Advances in Neural Information Processing Systems*, pp. 7645–7655
73. Graves, A. *et al.* (2016) Hybrid computing using a neural network with dynamic external memory. *Nature* 538, 471–476
74. Lemke, C. *et al.* (2015) Metalearning: a survey of trends and technologies. *Artif. Intell. Rev.* 44, 117–130
75. Thrun, S., Pratt, L., eds. (1998) *Learning to Learn*, Kluwer Academic Publishers, USA
76. Flennerhag, S. *et al.* (2020) Meta-learning with warped gradient descent. In *Proceedings of the International Conference on Learning Representations*, pp. 24
77. Nagabandi, A. *et al.* (2019) Deep online learning via meta-learning: continual adaptation for model-based RL. In *Proceedings of the International Conference on Learning Representations*
78. Beaulieu, S. *et al.* (2020) Learning to continually learn. *arXiv* Published online February 21, 2020. https://arxiv.org/abs/2002.09571
79. Javed, K. and White, M. (2019) Meta-learning representations for continual learning. In *Advances in Neural Information Processing Systems*, pp. 32
80. Finn, C. *et al.* (2019) Online meta-learning. In *Proceedings of the 36th International Conference on Machine Learning*
81. Yao, H. *et al.* (2019) Hierarchically structured meta-learning. In *Proceedings of the 36th International Conference on Machine Learning*
82. Finn, C. *et al.* (2018) Probabilistic model-agnostic meta-learning. In *Advances in Neural Information Processing Systems*, pp. 31
83. He, X. *et al.* (2019) Task agnostic continual learning via meta learning. *arXiv* Published online June 12, 2019. https://arxiv.org/abs/1906.05201
84. Triantafillou, E. *et al.* (2019) Meta-dataset: a dataset of datasets for learning to learn from few examples. *arXiv,* https://arxiv.org/abs/1903.03096
85. Yu, T. *et al.* (2019) Meta-world: a benchmark and evaluation for multi-task and meta reinforcement learning. *arXiv* Published online October 24, 2019. https://arxiv.org/abs/1910.10897
86. Andrychowicz, M. *et al.* (2016) Learning to learn by gradient descent by gradient descent. In *Advances in Neural Information Processing Systems*, pp. 29
87. Finn, C. *et al.* (2017) Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*
88. Vinyals, O. *et al.* (2016) Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, pp. 29
89. Snell, J. *et al.* (2017) Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pp. 30
90. Hochreiter, S. *et al.* (2001) Learning to learn using gradient descent. In *Artificial 25 Neural Networks — ICANN 2001*, Springer
91. Ravi, S. and Larochelle, H. (2017) Optimization as a model for few-shot learning. In *Proceedings of the International Conference on Learning Representations*
92. Garnelo, M. *et al.* (2018) Conditional neural processes. In *Proceedings of the International Conference on Machine Learning*
93. Zintgraf, L. *et al.* (2019) Fast context adaptation via meta-learning. In *Proceedings of the 36th International Conference on Machine Learning*

94. Rusu, A.A. *et al.* (2019) Meta-learning with latent embedding optimization. In *Proceedings of the International Conference on Learning Representations*
95. Rakelly, K. *et al.* (2019) Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *Proceedings of the 36th International Conference on Machine Learning*
96. Lee, K. *et al.* (2019) Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10649–10657
97. Caccia, M. *et al.* (2018) Online fast adaptation and knowledge accumulation: a new approach to continual learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1806