Continual Learning for Classification Problems: A Survey

Mochitha Vijayan, S. S. Sridhar

Continual Learning for Classification Problems: A Survey

*Mochitha Vijayan [1], Dr. S. S. Sridhar [2]
[1] SRM Institute of Science and Technology
[1] mv2820@srmist.edu.in
[2] SRM Institute of Science and Technology
[2] sridhars@srmist.edu.in

Abstract. Artificial Neural Networks performs a specific task much better than a human but fail at toddler level skills. Because this requires learning new things and transferring them to other contexts. So, the goal of general AI is to make the models continually learning as in humans. Thus, the concept of continual learning is inspired by lifelong learning in humans. However, continual learning is a challenge in the machine learning community since acquiring knowledge from data distributions that are non-stationary in general leads to catastrophic forgetting also known as catastrophic interference. For those state-of-art deep neural networks which learn from stationary data distributions, this would be a drawback. In this survey, we summarize different continual learning strategies used for classification problems which include: Regularization strategies, memory, structure, and Energy-based models.

Keywords: general AI, Artificial Neural Networks, Lifelong Learning, Continual Learning, Catastrophic Forgetting, classification problems.

1 Introduction

Deep learning holds state-of-the-performances in specific tasks in the areas of Computer vision, Natural Language Processing (NLP), Speech recognition, etc. This success is due to supervised training from huge and fixed datasets. No matter how big your training dataset is, as we grow in terms of data dimensionality and when we would like to solve more complex problems (not the toy problems), it becomes exponentially more difficult to cover the entire space of possibilities without representing the data set collected a priori. But how can we improve the scalability of models in terms of computation and memory resources and how can they adapt to new circumstances never seen before? The answer to this is to never stop learning as our brain does. So, Continual Learning [1][2] which is inspired by lifelong learning in humans is the way to deal with a higher and realistic time-scale where data (and tasks) become available only during the time.

Why continual Learning is a challenging problem?

Many connectionist models that are gradient-based such as Artificial Neural Network's (ANN) suffer from Catastrophic forgetting [3][4]. This means that once you train a network on a piece of data and when we expose it to a little new data distribution, it tends to forget the older ones. Since every parameter of the network is rewritten to suit the new data distribution. ANNs are unable to learn new information or instances immediately. Deep NNs are the dominant approach to machine perception, but: 1. They cannot learn new instances immediately, 2. Learning requires multiples loops over a data set, 3. They are susceptible to catastrophic forgetting if the data is not iid (independently and identically distributed).

To overcome catastrophic forgetting AI systems must be able to retain the previously learned task, acquire new knowledge and restrict the novel data from interfering with the existing knowledge. That is, the system must be stable to retain acquired knowledge without catastrophically forgetting them and must be plastic enough to integrate novel information. This concept is being widely explored in both biological and artificial models and is well-known as the stability-plasticity dilemma [5][6]. Two reasons to solve catastrophic forgetting: 1. Making systems learn like humans and animals, 2. Enabling new applications like immediate inference from new input labels, to avoid retraining from scratch thereby optimizing memory and computational resources.

For a classification task, Task-IL is often artificial, Class-IL is crucial for a continual learning setting in classification problems. This survey focuses on different continual learning strategies used for classification tasks. Furthermore, some ideas to enhance continual learning efficiency is also being discussed. We survey various strategies for continual learning in neural networks that mitigate catastrophic forgetting to different levels. We categorize the continual learning approaches used in the existing systems into Regularization strategies, Memory-based strategies, Structure-based strategies, and Energy-Based Models.

2 Regularization strategies

In continual learning, regularization is used to ensure the stability of important parameters for the previously learned tasks.

Li et al., 2018 present Learning without Forgetting (LwF) which is a hybrid of knowledge distillation (distillation loss to maintain consistency) and fine-tuning. This method has a relation with replay-based methods in a way that rather than storing or generating the samples to be replayed.Using the model learned on the previous tasks this approach labels the current task and replays them. Trying to transfer the information from an extremely regularized massive model to a smaller model. For a parameter set θ_s (shared) of all the tasks it tries to optimize the parameter θ_n of the novel task along with θ_s. An extra restraint is imposed such that

for a new task the parameters θ_s and the parameters of old tasks θ_o will not deviate much to remember θ_o. If for a new task, (X_n, Y_n) is the training data, the old tasks output Y_o, and new parameters θ_n(randomly choosed), then the updated parameters θ_s^* , θ_o^*, θ_n^* are:

$$\theta_s^*, \theta_o^*, \theta_n^* \leftarrow argmin_{\hat{\theta}_s, \hat{\theta}_o, \hat{\theta}_n} (\lambda_o \mathcal{L}_{old}(Y_o, \hat{Y}_o) + \mathcal{L}_{new}(Y_n, \hat{Y}_n) + R(\hat{\theta}_s, \hat{\theta}_o, \hat{\theta}_n))$$

Here a loss is added to discourage the old task *output to change* for new tasks. Optimization is done both for the final layers and shared representation . They show that LwF performs better than LFL(Jung et al, 2018) which is a similar approach. The drawback of this approach is that it extremely depends on the task's relevance and the training time for one task will have a linear hike with the number of tasks learned. In multi-task learning distillation used here is a potential solution but it demands a repository of persistent data for every task learned. This method is not immediately applicable in Reinforcement Learning scenarios.

In Less-forgetting learning (LFL), Jung et al., 2018 the old task performance is preserved by protecting the shared representations. This learning model satisfies two properties, there should be clear-cut boundaries for each task and the shared parameters should be kept unchanged. For the final most hidden activations the l2 distance between them is regularized, thereby protecting the previous input to output association being learned from the old tasks. This is done by the computation of additional activations using the parameters of old tasks. This L2 loss discourages the *output after* θ_s (shared parameter set) from changing for new tasks, while θ_o (task-specific parameters for past tasks) remains as it is. However, it requires the computation of the old task parameters for each new data point making it computationally expensive.

Kirkpatrick et al., 2017 developed an algorithm akin to synaptic consolidation for Artificial Neural Nets known as EWC (Elastic Weight Consolidation) which focuses on task-specific synaptic consolidation. EWC allows the knowledge acquired in the previous task to be protected during a new task learning as shown in fig 1.

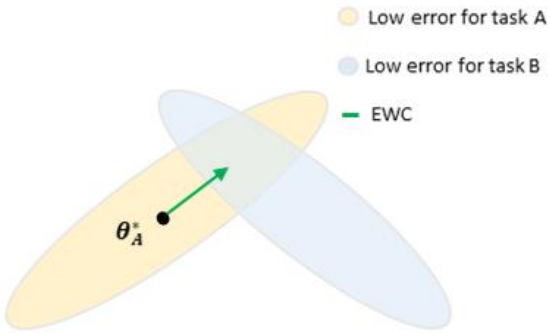

Fig 1. EWC assures task A is retained while training on task B.

The importance of the parameter θ concerning a task's training data D is represented as $p\ (\theta \mid D)$, the posterior distribution. While considering a scenario with tasks A (with $D_{A)}$ and B (with D_B) which are independent, applying Baye's theorem to the posterior probability and then taking the log gives us,

$$\log p(\ \theta \mid D) = \log p(D_B \mid \theta) + \log p(\ \theta \mid D_A) - \log p(D_B)$$

The knowledge of earler task is held by $\log p\ (\theta \mid D_A)$, the posterior probability. EWC uses a Gaussian distribution and adds a cost term to prioritize the important weights of past tasks by using a fisher information matrix F and is evaluated by its diagonal. Gaussian distribution is used since $\log p\ (\theta \mid D_A)$ is intractable.

$$\mathcal{L}(\theta) = \mathcal{L}_B(\theta) + \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta^*_{A,i})^2$$

EWC performed well in both supervised (permuted MNIST task framed by Goodfellow et al., 2013) and reinforcement (Atari 2600) learning scenarios. EWC's characteristic of making weights less plastic with time promotes memory retention rather than forgetting. As the number of random patterns exceeds the capacity of the network EWC degrades than the plain gradient descent. Moreover, Hopfield networks which is an EWC model suffer from the phenomenon of blackout catastrophe when the network capacity is saturated which results in the hindrance of retrieval of old knowledge and addition of new ones. The available output labels have to be summed up to calculate the diagonal of the Fisher matrix which makes its complexity linear to the number of outputs, which limits the application of this algorithm to low-dimensional output spaces. [34] show that EWC works poorly at incremental class learning.

Zenke et al., 2017 introduced intelligent synapses by maintaining an online measure of synapses (parameters like weights and bias) "importance" in solving the past tasks. These important synapses are protected from changing when the task changes. That is, the future learning is assisted by the synapses which are of less significance to the previous tasks, thereby alleviating catastrophic forgetting. A surrogate loss term is added to hinder significant variations in important parameters θ_k while learning a novel task.

Loss function:

$$\mathcal{L}'_{\mu} = \mathcal{L}_{\mu} + c \underbrace{\sum_{k} \Omega_k^{\mu} \, (\theta'_k - \theta_k)^2}_{\text{surrogate loss}}$$

This measure tracks parameters (past and present) and evaluates their importance online. Parameters contributing more to the loss function are more relevant. In [11], SI is shown to perform well on a multi-headed Split MNIST of their design and had shown similar performance as EWC on the permuted MNIST dataset. By using a multi-headed split CIFAR task of their design some transfer learning is also done.

In Lee et al., 2017, A L2 penality is applied to the shared parameter changes and for learning a novel task. Models formed are merged using first or second-moment matching. IMM is a progression of EWC which performs a model-fusion separately upon learning a novel task.

Aljundi et al., 2018 proposed Memory-aware Synapses (MAS) which is a model-based method inspired by neuroplasticity and Hebbian learning in biological systems. MAS determines the significance of the parameters that adapt to the test data set using the variation in the model output function concerning the inputs, rather than the loss function. Thus, avoiding the problem of falling in local minima. Important parts of the model can be learned using unlabeled data. This computation is done online and in an unsupervised manner. MAS could achieve in terms of constant memory w.r.t the task count, ability to build on the top of a pre-trained model and to add new tasks, ability to learn from unlabelled data.

Uncertainty-guided Continual learning approach with Bayesian neural networks (UCB), Ebrahimi et al., 2020 used an uncertainty prediction strategy for continual learning. Parameters of importance are either fully safeguarded using a binary mask (UCB-P) or can be changed conditionally based on their uncertainty learning novel tasks (UCB). UCB-P avert forgetting after the initial phase of pruning by retaining a binary mask for each task. UCB doesn't take additional memory and let on a more flexible learning mechanism in the network by limiting the forgetting.

Although these regularization-based approaches can be computationally efficient in alleviating catastrophic forgetting under certain conditions, a disadvantage is that they gradually reduce the model's capacity for learning new tasks. In class-incremental learning, regularization-based methods are shown to consistently fail [30][31].

3 Memory-based strategies

A set of experiences, exemplars, or vectors which represent the tasks can be stored rather than storing all the observations. This provides a more efficient and scalable memory strategy. This also enables compression and high-level transfer across multiple tasks. The focus of many works is dealing with the challenge to determine which samples to store.

Robins, A., 1995 in his work interleaved new experiences with the generated patterns of previous experiences intending to make neural networks able to encode information, store, and recall them on the need for better generalization across tasks. This method performed well in alleviating catastrophic forgetting. i.e., a continual rehearsal of previous tasks is done. This continual learning strategy based on the memory is called replay or otherwise rehearsal.

Rebuffi et al., 2017 proposed iCaRL (Incremental Class and Representation Learning) which combines regularization and memory strategies. iCaRL simultaneously learns the classifier and feature extractor. It consists of three components: classification done by a nearest-mean exemplars rule, herding based exemplar selection, learning of representation that uses exemplars along with distillation intending to avoid catastrophic forgetting. One exemplar image each for all classes seen so far is stored. Then computes a model for each of the classes which is the average of each exemplar stored. The input image is classified as the class of the prototype that is close to the current feature representation. As the feature representation changes recompute the mean of exemplars.

The training set consists of samples of new data as well as the exemplars of old classes (to remind what the old classes were).

$$D \leftarrow \bigcup_{y=s,..,t} \{(x,y): x \in X^y\} \quad \cup \quad \bigcup_{y=1,..,s-1} \{(x,y): x \in P^y\}$$

Loss function:

$$\mathcal{L}(\theta) = -\sum_{(x_i,y_i)\epsilon D} [\underbrace{\sum_{y=s}^{t} \delta_{y=y_i} \log\left(g_y(x_i)\right)}_{\text{classification loss}} + \underbrace{\sum_{y=1}^{s-1} q_i^y \log\left(g_y(x_i)\right)}_{\text{distillation loss}}]$$

The classification loss term will help to learn new classes and the distillation loss term will help not to forget previous outputs.

iCaRL can learn in an incremental fashion over a long period in experiments conducted on CIFAR-100 and ImageNet ILSVRC 2012 data wherein other methods fail quickly. Whereas, iCarL's performance is not appreciable when the training samples of classes are available in batch settings. iCarl's performance is heavily influenced by the number of examples it stores.

Variational Continual Learning (VCL) by Nguyen et al., 2018 uses Variational Interference (VI) whic is also a replay-based method. The previous posteriori is multiplied by the likelihood of the new task data set to obtain the new posterior distribution. This is a Bayesian inference to carry out continual learning. They calculate the Kullback-Leibler (KL) (approximately a quadratic regularization with rotation) divergence between the current distribution and the previous posterior. In addition to the VCL regularization term they showed that by using a core-set, which samples examples from old tasks, VCL is experiencing less forgetting. VCL with or without core-set outperforms EWC and SI on the permuted MNIST and multi-headed split MNIST dataset. Using permuted MNIST shows to yield misleading results in the context of continual learning as shown in [29].

A replay is a phenomenon seen in rodents and humans [24]. But it is unrealistic to maintain an unlimited number of observations or exemplars. In generative models as in Shin et al., 2017, samples are not stored. Instead, generative models are trained and then used to produce data needed for rehearsal. Trying to mimic the generative nature of the hippocampus for the rehearsal of past experiences, Shin et al., 2017 proposed Deep Generative Replay (DGR) consisting of two components: a generative model and a solver. A pseudo-data that represents the previous tasks are generated and are interleaved along with new tasks. This avoids the need for explicitly revising past training data for rehearsal, thus downsizing memory requirements. The catastrophic interference problem now shifts to the training process of the generative model.

Kember et al., 2018 proposed a model for Class-IL inspired by the dual memory model [3] of a mammalian brain and the studies of recall and consolidation that happens in a mammalian brain while fear conditioning [36]. FearNet uses a hippocampal net that is responsible for recalling immediate memories, long-term (mature) memories are handled by a PFC network and a neural net that took inspiration from the basolateral amygdala for deciding whether the model should make use of the hippocampal or PFC network for that specific sample or not. Memory consolidation happens during the sleeping phase that the FearNet consolidates hippocampal network to the PFC network, hippocampal network and PFC operates as complementary memory systems. PFC is a generative neural net that generates pseudo examples that are interleaved with the new samples in the hippocampal network.

One way to enforce the gradient to stay close to the gradients from previously learned tasks is to eliminate its interference as in Lopez-paz et al., 2017 Chaudhry, A. et al., 2018. These methods are beneficial in multi-task learning as they can make learning more efficient in case of opposing objectives. Lopez-Paz et al., 2017 introduced the Gradient Episodic Memory (GEM) model that brings in the positive backward transfer to previously learned tasks. It is the first approach to accomplish learning in one pass. GEM is characterized by an episodic memory that stores a subset of the observed examples for a given task and alleviates catastrophic forgetting (negative backward transfer). This model enables to learn the subset of correlativity of a set of tasks, capable to predict desired values associated with past or novel tasks without enforcing task descriptors. GEM's memory requirement is high than other regularization strategies such as EWC (Kirkpatrick et al. 2017) at the training time. As GEM is being evaluated on MNIST and CIFAR datasets it is doubtful if it scales to more realistic problems. An inherent problem here is the constraint on the amount of experience that can be stored in memory, which could quickly become a limiting factor in large scale problems. Even when this method works well in a single pass context it demands much more computation.

Chaudhry, A. et al., 2018 proposed Averaged GEM (A-GEM) which is an advanced version of GEM which alleviates the computational burden of GEM. GEM ensures that the loss of each specific previous tasks (sampled from episodic memory) at every training step does not increase. Whereas, A-GEM ensures that the average of the episodic memory loss over the past tasks at every training step does not increase. A-GEM shows the same or even better performance as GEM. A-GEM is shown to be computationally and memory-efficient as EWC and other regularization-based methods.

Scalability of the approaches that use memory for replay or rehearsals when the number of task increases is arguable. And the rehearsal-based approaches have to deal with the quality of samples generated which would again be a drawback.

4 Structure-based strategies

These strategies determine if the network has to be expanded to represent the new tasks or not. Progressive networks (Rusu et al., 2016; Schwarz et al., 2018) are dynamic architecture-based approaches where the base architecture was replicated and some connections were included in response to new tasks. These approaches were successful in reinforcement learning scenarios.

Mallya et al., 2018 proposed PackNet. Forgetting is alleviated by iterative pruning to completely freeze the updates on the most important parameters. A binary mask (task-specific) that indicates the importance and unimportance of parameters is saved. More specifically, a single network's capacity is used to learn multiple tasks

and is done by freeing the parameters irrelevant to the present task depending on their magnitude. This approach requires knowing the task prior to the use of a suitable mask. Ranking the importance of weight by their magnitude will not be a guaranteed "*importance*" indicator. Parameter importance is ranked by its magnitude in PackNet which cannot give any guarantee to be an "*importance*" indicator. This model relies on the explicit parametrization of importance.

Hard attention to the tasks (HAT) introduced by Serra et al., 2018 learns an attention vector to determine the important neurons (using mini-batch stochastic gradient descent and backpropagation) to the task controlling the gradient shifts. An almost-binary mask per previous tasks is used to maintain the information learned on previous tasks. HAT employs an attention-weighted L1 regularization on the attention mask Instead of simple L1 regularization. These attention masks are lightweight structures included without changing much in the existing network. When EWC and SI adds a soft architectural regularization to the loss function, HAT adds a hard structural regularization to both the loss function and gradient magnitudes.

Rusu et al., 2016 in their work hindered any adjustments to the previous network. Instead expanded the network with a new subnetwork to be trained with new data. This progressive network retains a repository of pre-trained architectures for each learned task. If the features being learned cannot correctly represent the novel task, more neurons are added to the network to account for the features of the new task. This approach has shown favorable results for a wide variety of reinforcement learning problems when compared to other strategies that either priorly train or tweak the model incrementally with the prior knowledge only at the initialization phase. This strategy prevents catastrophic forgetting, however, without considering the difficulty of the task they add in a constant count of units for each learned task leading the complexity of the network architecture to grow and thus is suboptimal in terms of network capacity utility and performance.

Yoon et al., 2018 proposed Dynamically Expandable Networks (DEN) in which, the network expands according to the task at hand by dividing/replicating the most important neurons retraining them on new tasks. This separate stage is known as selective retraining. The important drifting units (neurons) are identified using a complex mixture of hyperparameters and heuristics. They employ L1 regularization as well as L2-transfer to condition learning, regularization constants, and a set of extra thresholds. Even though they bring in high computation this is unavoidable in a continual learning environment where several tasks have to be learned and when network capacity cannot be fixed.

When new tasks are encountered modular-based models go for explicit ways to increase architectural capacity. In recent times, several adaptive network models have emerged that expand and at the same time protects and reuse the existing representations [22]. But this leads to demanding high computational requirements

in the learning process. Schwarz et al., 2018 increase networks capacity more aggressively if needed and then prune or sometimes compress the parts of the network.

Structure-based approaches are limited to task-incremental learning settings.

5 Energy-based models for Continual Learning

Li et al., 2020 proposed an Energy-based model for continual learning which does not use an extra loss, memory, or model and is still performing well in class incremental learning scenarios and boundary agnostic settings which are the challenges of Continual Learning. EBMs naturally deals with the challenging problems in Continual Learning without replay. A contrastive divergence training procedure is used and it provides a natural way to deal with dynamically growing. They propose to learn new conditional gains during the training process, which makes EBMs parameter update cause less interference with old data. But EBM's takes more time than softmax-based classifier model in terms of convergence and evaluation.

6 Research insights

- The model (agent) should be capable of learning from both stationary and non-stationary data streams. In real-world data comes from a dynamic data distribution that keeps on changing with time. The state-of-art neural networks are provided with iid data and are allowed to loop over it again and again.
- Data streams like MNIST and permuted MNIST are highly unrealistic, most of the conventional neural nets use them. Evaluation of the existing continual learning is mostly done on toy data streams. The applicability of these approaches in a real-world setting is questionable.
- Models that work well in MNIST, CIFAR, etc tend to fail in other important paradigms as shown [34]. MNIST and CIFAR are small datasets.
- Memory and computations must be kept fixed across the models to enable a fair comparison which is not done in most of the cases.
- Evaluation of a model in terms of Task-IL, Domain-IL, and Class-IL can be done only on continual learning problems with a series of tasks with clear cut boundaries [31].

7 Concluding remarks and future direction

We can make artificial agents perform a particular task surprisingly much better than a human. But current artificial agents are still in their infancy due to a serious problem known as catastrophic forgetting. Continual learning as in humans is inevitable. A continual learning agent must not suffer from the phenomena of catastrophic forgetting. It means that the agent must retain its ability to perform fairly well on previously learned tasks and should be able to learn new tasks by extracting knowledge from the previous ones exhibiting positive forward transfer thereby achieving better performance and fast learning. It should be scalable by dynamically adapting to the real-time environment with varieties of tasks. It should enable positive backward transfer by gaining better performance on previous tasks after learning a new similar task. The data collected from the real-world may not be always explicitly labeled. The agent should learn from unlabeled data and even should be the recipient of tasks without clear task boundaries. Catastrophic forgetting isn't the only barrier of general AI. Not every task has a huge dataset and the challenge is to make the models learn from more than just from the samples. This survey tries not only to highlight the inevitability of continual learning but also to report the limitations of existing state-of-art neural networks in this regard.

References

1. Hassabis, D. et al. (2017) Neuroscience-inspired artificial intelligence. Neuron 95, 245–258. Hassabis, D., Kumaran, D., Summerfield, C. & Botvinick, M. (2017), 'Neuroscience-inspired artificial intelligence', Neuron Review 95(2), 245–258.
2. Thrun, S. & Mitchell, T. (1995), 'Lifelong robot learning', Robotics and Autonomous Systems 15, 25–46.
3. McClelland, J. L., McNaughton, B. L. & O'Reilly, R. C. (1995), 'Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory', Psychological Review 102, 419–457.
4. McCloskey, M. & Cohen, N. J. (1989), 'Catastrophic interference in connectionist networks: The sequential learning problem', The Psychology of Learning and Motivation 24, 104–169.
5. Ditzler, G., Roveri, M., Alippi, C. & Polikar, R. (2015), 'Learning in nonstationary environments: A survey', IEEE Computational Intelligence Magazine 10(4), 12–25.
6. Mermillod, M., Bugaiska, A. & Bonin, P. (2013), 'The stability-plasticity dilemma: Investigating the continuum from catastrophic forgetting to age-limited learning effects', Frontiers in Psychology 4(504).
7. Grossberg, S. (1980), 'How does a brain build a cognitive code?', Psychol. Rev. 87, 1–51.
8. Grossberg, S. (2012), 'Adaptive resonance theory: how a brain learns to consciously attend, learn, and recognize a changing world.', Neural Networks 37, 1–41.
9. Rebuffi, S.-A., Kolesnikov, A., Sperl, G. & Lampert, C. H. icarl: incremental classifier and representation learning. In Conference on Computer Vision and Pattern Recognition

5533–5542 (IEEE, Honolulu, 2017)
10. Kirkpatrick, J. et al. Overcoming catastrophic forgetting in neural networks. Proc. Natl Acad. Sci. USA 114, 3521–3526 (2017).
11. Zenke, F., Poole, B. & Ganguli, S. Continual learning through synaptic intelligence. In International Conference on Machine Learning 3987–3995 (PMLR, Sydney, 2017).
12. Li, Z. & Hoiem, D. Learning without forgetting. IEEE Trans. Pattern Anal. Mach. Intell. 40, 2935–2947 (2017).
13. Jung, H., Ju, J., Jung, M. & Kim, J. (2018), Less-forgetting learning in deep neural networks, AAAI'18, New Orleans, LA.
14. Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 139–154, 2018.
15. Rahaf Aljundi, Klaas Kelchtermans, and Tinne Tuytelaars. Task-free continual learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 11254–11263, 2019.
16. Sayna Ebrahimi, Mohamed Elhoseiny, Trevor Darrell and Marcus Rohrbach, Uncertainty-Guided Continual Learning with Bayesian Neural Networks. Published as a conference paper at ICLR 2020.
17. Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
18. Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In Jennifer Dy and Andreas Krause (eds.), Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pp. 4548–4557. PMLR, 2018.
19. Kemker R., Kanan C. Fearnet: Brain-inspired model for incremental learning, ICLR'18 (2018)
20. Lopez-Paz, D. et al. (2017) Gradient episodic memory for continual learning. In Advances in Neural Information Processing Systems (NeurIPS), pp. 30
21. Chaudhry, A. et al. (2018) Efficient lifelong learning with A-GEM. arXiv Published online December 2, 2018. https://arxiv.org/abs/ 1812.00420Yoon, J., Yang, E., Lee, J., Hwang, S.J.: Lifelong learning with dynamically expandable networks. In: International Conference on Learning Representations (2018)
22. Andrei Rusu, Neil Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu and Raia Hadsell. Progressive neural networks. arXiv preprint arXiv:1606.04671, 2016.
23. Shin, H. et al., 2017. Continual learning with deep generative replay. In Advances in Neural Information Processing Systems
24. Liu, Y. et al., 2019. Human replay spontaneously reorganizes experience. Cell 178, 640–652
25. Robins, A., 1995. Catastrophic forgetting, rehearsal and pseudorehearsal. Connect. Sci. 7, 123–146
26. 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
27. Cuong V. Nguyen, Yingzhen Li, Thang D. Bui, and Richard E. Turner. Variational continual learning. In International Conference on Learning Representations, 2018.

28. Sang-Woo Lee, Jin-Hwa Kim, Jaehyun Jun, Jung-Woo Ha, and Byoung-Tak Zhang. Overcoming catastrophic forgetting by incremental moment matching. In Advances in Neural Information Processing Systems, pp. 4652–4662, 2017
29. Lee, S.W., Kim, J.H., Ha, J.W., Zhang, B.T.: Overcoming catastrophic forgetting by incremental moment matching. arXiv preprint arXiv:1703.08475 (2017)
30. Sebastian Farquhar, Yarin Gal. Towards Robust Evaluations of Continual Learning. arXiv preprint arXiv:1805.09733, 2018
31. Gido M. van de Ven, Andreas S. Tolias. Three scenarios for continual learning. arXiv preprint arXiv:1904.07734, 2019
32. Ian J. Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An Empirical Investigation of Catastrophic Forgetting in Gradient-Based Neural Networks. arXiv, 2013. 4.2.1, A.1, A.5
33. Shuang Li, Yilun Du and Gido M. van de Ven. Energy-Based Models for Continual Learning. arXiv preprint arXiv, 2020
34. R Kemker, M McClure, A Abitino, T Hayes. Measuring catastrophic forgetting in neural networks. Proceedings of the AAAI, 2018
35. Raia Hadsell, Dushyant Rao, Andrei A. Rusu and Razvan Pascanu. Embracing Change: Continual Learning in Deep Neural Networks. Trends in Cognitive Sciences, December 2020, Vol. 24, No. 12, 2020.
36. Kitamura T., Ogawa S.K., Roy D.S., Okuyama T., Morrissey M.D., Smith L.M., et al. Engrams and circuits crucial for systems consolidation of a memory Science, 356 (2017), pp. 73-78.
37. German I.Parisi, Ronald Kemker, Jose L.Part, ChristopherKanan and StefanWermtera. Continual lifelong learning with neural networks: A review. Neural Networks 113 (2019) 54-71.