XAI for Continual Learning

Omar Bayoumi 1747042 (Student)

ARTICLE INFO

Keywords:
Continual learning
Explainable AI
Catastrophic forgetting

ABSTRACT

In the real world to solve problems there is a need to adapt to all kinds of situations. With the advent of AI, we want this flexibility to be maintained in the field of machine learning as well. To this end, continual learning was born, the goal of which is to learn all tasks that are asked of it. One of the main problems, however, is catastrophic forgetting which causes models learning a new task to "forget" the task they were previously trained on producing a drop in performance. In this paper, we will look at all the main methods that serve to counter catastrophic forgetting in continual learning.

1. Introduction

Learning is the mechanism behind intelligence. The invention of machine learning, and associated technologies such as computational power and storage space to handle ever-larger data, has brought innovation to many areas such as in computer vision, where filters were previously applied to images, and making huge advances in areas such as Medicine and Natural Language Processing.

Modern artificial intelligence models can be compared to supercomputers, but they appear to be specialized for a specific purpose, with a low capacity for generalization. When a new type of data, different from those used for training, is given as input to the model, the model is unable to classify it correctly.

The process of constantly learning new types of data is called Continual Learning. Continual Learning is inspired by the learning process of human beings, who from infancy start learning new things all the time, but do not forget the old ones. Although continual learning appears to be a promising approach, this learning technique must, however, face several challenges. In particular, the main problem is that of Catastrophic Forgetting (McClelland, McNaughton and O'Reilly (1995), McCloskey and Cohen (1989)), in which, all prior knowledge obtained for a first task is overwritten by that of the second task learned. This phenomenon is an aspect of the trade-off between learning plasticity and memory stability where an excess of one interferes with the other. Additional challenges associated with continual learning relate to the type of data used for training: models must be able to learn from randomly distributed data and must be trained by not having all the data from the beginning.

In this paper, we will address different techniques for implementing continual learning and discuss how the use of explainable AI (XAI) can help in the process. By the term XAI (Gunning and Aha (2019)) we mean the field of artificial intelligence in which we try to find new methods and techniques that can be used to understand the internal processing of a model to obtain a specific output.

2. Methods for dealing with Catastrophic Forgetting

As already mentioned in the introduction, the main problem in continual learning is catastrophic forgetting (McCloskey and Cohen (1989)), where a trade-off must be found between plasticity and stability. Stability means a network that can learn with high accuracy a specific task A but is unable to learn a second task B. Plasticity, on the other hand, means the ability to learn new tasks while forgetting previous ones. The goal is to find a balance between these two characteristics. In this section, five main categories of approaches for Catastrophic Forgetting will be analyzed.

2.1. Regularization based

This method takes inspiration from biology, where a special mechanism is present in the brain to protect past knowledge. For previous tasks, where certain neurons are involved, the level of plasticity is lowered in the synapses (Hassabis, Kumaran, Summerfield and Botvinick (2017)). The first approach is the weight regularization (Wang, Zhang, Su and Zhu (2023)) where upon training the new task, the learning rate of the weights of the previous task is lowered so that things learned are not forgotten. A common approach is to introduce a quadratic penalty in the loss function. This penalty hits parameters according to their importance in previous tasks (Equation 1). Significance can be calculated from Fisher's information matrix (FIM), such as Elastic weight consolidation (EWC) (Kirkpatrick et al. (2017)). Here, the use of EWC loss function:

$$\mathcal{L}_{EWC}(\theta) = l_k(\theta) + \frac{\lambda}{2}\left(\theta - \mu_{k-1}\right)^{\top} \widehat{F}_{1:k-1}(\theta - \mu_{k-1}) \quad (1)$$

where the l_k denotes the task-specific loss, the FIM $\widehat{F}_{1:k-1} = \sum_{t=1}^{k-1} \mathrm{diag}(F_t)$ with a diagonal approximation $\mathrm{diag}(\cdot)$ of each F_t, and λ is a hyperparameter to control the strength of regularization.

The second approach is called function regularization which mainly affects the intermediate or final output of the prediction function. This technique uses the previously trained model as teacher and the currently trained model as student while, on the old training samples, knowledge distillation (KD) (Gou, Yu, Maybank and Tao (2020)) is applied to mitigate catastrophic forgetting.

2.2. Replay based

This method also takes inspiration from biology. When a human being receives input from outside, it is encoded in the hippocampus and finally consolidated in the cerebral cortex. The consolidation phase is accompanied by memory replay in the hippocampus and neocortex during sleep (Hassabis et al. (2017)). Replay based methods can be divided into three categories: Experience replay, Generative replay and Feature replay.

2.2.1. Experience replay

This technique is based on storing some training samples of the previous task in a memory buffer. Some of the main difficulties with this method are choosing which samples to use and optimizing the memory buffer. One possible algorithm is Reservoir sampling (Vitter (1985)), where given a data stream of unknown length, a subset of samples belonging to that stream is returned. The selection strategy chooses the sample with a probability of $\frac{mem\_sz}{n}$ where mem_sz is the size of the memory buffer and n is the number of samples observed so far. There are also more complex selection strategies based on parameter gradients such as Greedy Sample Selection (GSS, Aljundi, Lin, Goujaud and Bengio (2019)). The idea is to maximize the diversity of samples in the memory buffer by assigning a score to each sample in the buffer based on its maximum cosine similarity with other random samples in the buffer. When the buffer is full, new samples can replace those already present with a probability proportional to their score. For memory buffer optimization, one technique that can be used is the Adaptive Quantization Modules (AQM, Caccia, Belilovsky, Caccia and Pineau (2019)) that allow online continual compression based on the VQ-VAE frameworks, in this way compressed data can be saved within the memory buffer. After this, an Exploitation phase is done. We need to try to make the best use of the memory buffer to retrieve past information. There are different techniques for this such as GEM (Lopez-Paz and Ranzato (2022)), the main feature of this method is its episodic memory which allows a subset of the observed samples from a given task to be kept in memory. The recent state of the art is the Strong Experience Replay (SER, Zhuo, Cheng, Gao and Kankanhalli (2023)) which improves and extends the Experience Replay, in fact, the SER not only allows to memorize the ground truth labels of the samples stored in the memory buffer, but also looks for consistent predictions between the new model and the old one on the complete training data and the new task.

2.2.2. Generative replay

The generative replay technique consists of training a generative model to replay generated data. For example, DGR (Shin, Lee, Kim and Kim (2017)) provides a framework that allows learning the generation task and at the same time allows replaying the old model and this way allows transferring knowledge from one model to another. To further improve Generative Replay and mitigate the effect of catastrophic forgetting, it is possible to combine it with other continual learning techniques such as weight regularization (Wang, Lei, Li, Su, Zhu and Zhong (2022)), (Seff, Beatson, Suo and Liu (2017)) and experience replay (Wang, Yang, Li, Hong, Li and Zhu (2021)). Generative models are typically GANs (Le, Huang, Xu, Fan, Ma, Mei and Ma (2022)) or VAEs (Khan, Cygert, Twardowski and Trzciński (2023)). GANs are generally more accurate but suffer from label inconsistency, in contrast, VAEs are less accurate but allow label checking of the generated data.

2.2.3. Feature replay

Feature replay is a subtask of generative replay in fact unlike the latter it targets the feature level and not the data level. This greatly reduces the complexity of the problem and also has privacy benefits. A major issue with this method is the representation shift that is caused by the continuous updating of the feature extractor. To solve this problem some methods perform a distillation between the old and the new model like Generative Feature Replay (GFR, Liu, Wu, Menta, Herranz, Raducanu, Bagdanov, Jui and van de Weijer (2020b)), Feature Adaptation (Iscen, Zhang, Lazebnik and Schmid (2020)) or Dynamic Structure Reorganization (DSR, Zhu, Zhai, Cao, Luo and Zha (2022)). Other methods consist of modifying the early layers of the feature extractor as REMIND (Hayes, Kafle, Shrestha, Acharya and Kanan (2020)) does. Another major problem with this method is the continual learning from scratch. In fact, feature replay performs much better with large-scale pertaining (Ostapenko, Lesort, Rodríguez, Arefin, Douillard, Rish and Charlin (2022)). This is done since in the case of training from scratch the stabilization of the feature extractor is a very complex challenge. (Stoychev, Churamani and Gunes (2023)) obtain excellent results with LGR which allows the use of low-dimensional latent features to mitigate forgetting in a resource-efficient manner.

2.3. Optimization based

Continual learning can also be done by manipulating and designing optimization programs. There are three main types of optimization based approaches: gradient projection, loss landscape and meta-learning.

2.3.1. Gradient projection

Among the gradient projection methods we find AOP (Guo, Hu, Zhao and Liu (2022)), which allows us to change the updates parameters so that they are aligned with the orthogonal projection of the previous input space. Other methods such as OGD (Farajtabar, Azizan, Mott and Li (2020)) allow to preserve the old gradient directions and try to modify them so that they are aligned orthogonally with the current ones. TRGP (Lin, Yang, Fan and Zhang (2022)) allows a "trust region" to be defined through the norm of gradient projection onto the subspace of previous inputs.

2.3.2. Loss landscape

This technique works by optimizing the two model losses for the two different tasks by finding a local minimum that performs well on both tasks. For example, Stable-SGD

(Mirzadeh, Farajtabar, Pascanu and Ghasemzadeh (2020)) allows SGD to find a local minimum by adapting to training time hyperparameters such as learning rate, weight decay, and dropout. It has been shown that unsupervised and self-supervised learning combined with loss landscape and large-scale pretraining allows for better results. This is because the representation of these models is more robust.

2.3.3. Meta-learning

Meta-learning allows models to adapt quickly to new tasks, even with few examples, and to generalize about them. This makes it very promising in the field of continual learning. The main concept of meta-learning is that it allows information to be obtained from data without the need to manipulate it previously. OML (Javed and White (2019)) allows having a meta-training strategy that works online and allows minimizing interference on incoming inputs. In addition, meta-learning can be used in combination with experience replay to make the best use of the old and new model samples. Examples of this are MER (Riemer, Cases, Ajemian, Liu, Rish, Tu and Tesauro (2019)) and La-MAML (Gupta, Yadav and Paull (2020)) which allows OML (Javed and White (2019)) to be further improved. There are also hybrid approaches such as OSAKA (Caccia and et al. (2021)) that consist of initializing the model with meta-training and then, through incremental tasks, adding knowledge to the initialization. Meta-learning also performs well with unsupervised learning with task construction as UMTRA (Khodadadeh, Bölöni and Shah (2019)) and LA-SIUM (Khodadadeh, Zehtabian, Vahidian, Wang, Lin and Bölöni (2020)) demonstrate.

2.4. Representation based

Initially studies on representation-based approaches are done relying on meta-training (Javed and White (2019)). However, later the works of Madaan, Yoon, Li, Liu and Hwang (2022) and Shi, Zhou, Liang, Jiang, Feng, Torr, Bai and Tan (2022) show how the use of self-supervised learning and large-scale pretraining enable much more robust results. The common feature of these approaches is that they work on a shared set of parameters.

2.4.1. Self-supervised learning

Exploiting the fact that self-supervised models are more robust to catastrophic forgetting LUMP (Madaan et al. (2022)) tries to interpolate images of old and new tasks. Differently, MinRed (Purushwalkam, Morgado and Gupta (2022)) tries to diversify the experience replay by decorrelating old training samples. Co2L (Cha, Lee and Shin (2021)) uses a supervised contrastive loss that adapts to learn each task and a self-supervised loss that is responsible for maintaining knowledge between old and new models. SCALE (Yu, Guo, Gao and Rosing (2023)) instead can extract and store representations on the fly purely from the data continuum.

2.4.2. Pre-training for downstream continual learning

Recent studies have shown that downstream continual learning can benefit from the use of pre-training, which not only allows for strong knowledge transfer but also allows to be more resistant to catastrophic forgetting. These benefits become evident when using pre-training based on big data and models and together with contrastive loss. One of the problems with this method is that the pre-trained knowledge must be adapted to the current task while maintaining generalizability to future tasks. Several methods have been developed to be able to solve this problem. One of them is to use a fixed backbone, Side-Tuning (Zhang, Sax, Zamir, Guibas and Malik (2020)), and DLCFT (Shon, Lee, Kim and Kim (2022)) to train a network in parallel to the backbone and then merge the outputs. TwF (Boschini and et al. (2022)) additionally uses a sibling network that is trained separately however unlike the previous method the knowledge is distilled from the layer-wise backbone. ADA (Ermis, Zappella, Wistuba, Rawal and Archambeau (2023)) on the other hand uses Adapters (Houlsby and et al. (2019)) for knowledge distillation to adjust a pre-trained transformer. Prompt-based methods have also been recently made in which the representation of a pre-trained transformer is instructed with prompt parameters such as L2P (Wang and et al. (2022)). Another solution is to use an updatable backbone. F2M (Shi, Chen, Zhang, Zhan and Wu (2021)) looks for a flat local minima in the pre-training phase and learns incremental tasks in the flat region. SLCA (Zhang, Wang, Kang, Chen and Wei (2023)), instead, performs slow fine-tuning of the backbone of a pre-trained transformer.

2.4.3. Continual pre-training

Pre-training requires large amounts of data that are typically collected incrementally. As a result, it is very important to improve downstream performance. Cossu, Tuytelaars, Carta, Passaro, Lomonaco and Bacciu (2022) show that, for vision tasks, self-supervised pretraining works by performing pretraining on data streams, and only then fine-tuning is performed better. In this way, the model becomes more resistant to catastrophic forgetting. For language models, on the other hand, ECONET (Han, Ren and Peng (2021)) presents a self-supervised network with generative replay. One of the main challenges of continual pre-training is the need to continuously adapt the pre-trained knowledge depending on the current task. To try to mitigate this problem, IDA (Liu, Majumder, Achille, Ravichandran, Bhotika and Soatto (2020a)) uses discriminants from the old and new models so that they align with the old centers.

2.5. Architecture based

These kinds of approaches are based on learning incremental tasks with a shared set of parameters, which is a major cause of inter-task interference. There are three methods to solve these problems through the construction of task-specific parameters.

2.5.1. Parameter allocation

This method consists of using an isolated subspace of parameters dedicated to each task in the network. The architecture can have fixed or variable dimensions. An example of a fixed network is Piggyback (Mallya, Davis and Lazebnik (2018)), which allows the optimization of a binary mask that is used to select certain parameters for each task. The main function of the mask is that it allows the old task to be frozen to avoid catastrophic forgetting. Since free parameters tend to saturate proportionally to the incremental tasks introduced, a possible solution is to use dynamic architectures such as DEN (Yoon, Yang, Lee and Hwang (2018)) that allow the architecture to expand dynamically in case the current capacity does not allow the learning of new tasks.

2.5.2. Model decomposition

Model decomposition consists of dividing the model into two components: task-sharing and task-specific. In general, task-specific components are expandable, for example, adaptive layers (GVCL (Loo, Swaroop and Turner (2020))). In addition, it is also possible to decompose network parameters into task-sharing and task-specific elements through APD (Yoon, Kim, Yang and Hwang (2020)) or RCM (Kanakis and et al. (2020)) techniques. In general, the number of task-specific components grows proportionally to incremental tasks.

2.5.3. Modular network

This technique uses sub-networks or sub-modules to learn incremental tasks in a differentiated fashion, without the need to subdivide into task-sharing or task-specific components like in previous methods. The purpose of this technique is to build task-specific models and, thanks to sub-networks or sub-modules, to allow reuse knowledge. In addition, this technique used can also be used in combination with other methods such as experience replay (Ramesh and Chaudhari (2022)).

3. Analysis of forgetting in artificial NNs

So far we have looked at some techniques to try to deal catastrophic forgetting. There are common causes of this phenomenon. Specifically, it is defined as a forgetting event when, during learning, a model incorrectly classifies a sample that was previously classified correctly. In particular, it has been shown how certain examples are forgotten more frequently and others not, and how state-of-art results can still be obtained by removing a significant fraction of the examples (Toneva and et al. (2018)). Other studies have shown that catastrophic forgetting is mainly due to the deeper layers because they are the ones that change the most during sequential training (Ramasesh, Dyer and Raghu (2020)). In addition, it has been shown that maximal forgetting occurs in task sequences of intermediate similarity (Ramasesh et al. (2020)).

4. EXplainable Artificial Intelligence

Explainable artificial intelligence (XAI) is a set of processes and methods that allows human users to comprehend and trust the results and output created by machine learning algorithms. XAI is divided into two main categories. Explainable by design are neural networks that can provide explanations together with their predictions. Simple models such as decision trees are used to do this. Post-hoc where you try to explain pre-trained models by looking at their output and parameters. There are three types of scenarios in which to use explainable AI: Instance-incremental learning, in which samples from the same task are given in batches, similar to classical supervised training, but the network tries to learn all the samples provided. Samples are given to the network continuously and the network is evaluated when the training is ongoing. Task-incremental learning is characterized by disjoint label spaces and task ids need to be given during training and testing. An example of XAI technique for continual learning in this scenario is the work of Ede and et al. (2022) where the role of the XAI is to identify weights that are relevant to previous tasks through Layer-wise Relevance Propagation (LRP, Bach, Binder, Montavon, Klauschen, Müller and Samek (2015)). Starting from the output of the network, the neurons in the last layer are analyzed to understand their contribution to the prediction. The ones that contributed the most to the correct prediction are selected, and this process is propagated to the previous layers. After this, a pruning phase takes place in which initially the neurons in the various layers are sorted according to their importance, and the less important ones are removed until a threshold is reached. Then the identified neurons are frozen and the remaining ones are trained. Class-incremental learning, as the former, is characterized by disjoint label spaces but task ids are only provided during training. Ebrahimi and et al. (2021) provide an example, in which the role of XAI is to provide saliency maps that are used in a replay-based approach. This method is a replay-based approach, consequently, it consists of storing the raw images used in the previous tasks, in addition to this, the proposed method also stores saliency maps which tries to keep unchanged as much as possible after training the new task. The main problem with this approach is the load on the memory since it has to maintain not only the images of the previous tasks but also the saliency maps.

5. Conclusion

We have seen how one of the main problems in continual learning is catastrophic forgetting. Various methods to deal with this phenomenon have been presented and analyzed, highlighting their pros and cons. In addition, thanks to recent discoveries in the field of XAI, it is possible to understand more deeply the functioning of neural networks allowing us to understand the motivations behind catastrophic forgetting and giving the possibility to develop new strategies.

References

Aljundi, R., Lin, M., Goujaud, B., Bengio, Y., 2019. Gradient based sample selection for online continual learning. arXiv:1903.08671.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W., 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLOS ONE 10, 1–46. URL: https://doi.org/10.1371/journal.pone.0130140, doi:10.1371/journal.pone.0130140.

Boschini, M., et al., 2022. Transfer without forgetting. arXiv:2206.00388.

Caccia, L., Belilovsky, E., Caccia, M., Pineau, J., 2019. Online learned continual compression with stacked quantization module. CoRR abs/1911.08019. URL: http://arxiv.org/abs/1911.08019, arXiv:1911.08019.

Caccia, M., et al., 2021. Online fast adaptation and knowledge accumulation: a new approach to continual learning. arXiv:2003.05856.

Cha, H., Lee, J., Shin, J., 2021. Co^{2}l: Contrastive continual learning. arXiv:2106.14413.

Cossu, A., Tuytelaars, T., Carta, A., Passaro, L., Lomonaco, V., Bacciu, D., 2022. Continual pre-training mitigates forgetting in language and vision. arXiv:2205.09357.

Ebrahimi, S., et al., 2021. Remembering for the right reasons: Explanations reduce catastrophic forgetting. arXiv:2010.01528.

Ede, S., et al., 2022. Explain to not forget: Defending against catastrophic forgetting with xai. arXiv:2205.01929.

Ermis, B., Zappella, G., Wistuba, M., Rawal, A., Archambeau, C., 2023. Memory efficient continual learning with transformers. arXiv:2203.04640.

Farajtabar, M., Azizan, N., Mott, A., Li, A., 2020. Orthogonal gradient descent for continual learning, in: Chiappa, S., Calandra, R. (Eds.), Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, PMLR. pp. 3762–3773. URL: https://proceedings.mlr.press/v108/farajtabar20a.html.

Gou, J., Yu, B., Maybank, S.J., Tao, D., 2020. Knowledge distillation: A survey. International Journal of Computer Vision 129, 1789 – 1819. URL: https://api.semanticscholar.org/CorpusID:219559263.

Gunning, D., Aha, D., 2019. Ai magazine. AI Magazine 40, 44–58.

Guo, Y., Hu, W., Zhao, D., Liu, B., 2022. Adaptive orthogonal projection for batch and online continual learning. Proceedings of the AAAI Conference on Artificial Intelligence 36, 6783–6791. URL: https://ojs.aaai.org/index.php/AAAI/article/view/20634, doi:10.1609/aaai.v36i6.20634.

Gupta, G., Yadav, K., Paull, L., 2020. La-maml: Look-ahead meta learning for continual learning. arXiv:2007.13904.

Han, R., Ren, X., Peng, N., 2021. Econet: Effective continual pretraining of language models for event temporal reasoning. arXiv:2012.15283.

Hassabis, D., Kumaran, D., Summerfield, C., Botvinick, M., 2017. Neuroscience-inspired artificial intelligence. Neuron 95, 245–258. doi:10.1016/j.neuron.2017.06.011.

Hayes, T.L., Kafle, K., Shrestha, R., Acharya, M., Kanan, C., 2020. Remind your neural network to prevent catastrophic forgetting. arXiv:1910.02509.

Houlsby, N., et al., 2019. Parameter-efficient transfer learning for nlp. arXiv:1902.00751.

Iscen, A., Zhang, J., Lazebnik, S., Schmid, C., 2020. Memory-efficient incremental learning through feature adaptation. arXiv:2004.00713.

Javed, K., White, M., 2019. Meta-learning representations for continual learning. arXiv:1905.12588.

Kanakis, M., et al., 2020. Reparameterizing convolutions for incremental multi-task learning without task interference. arXiv:2007.12540.

Khan, V., Cygert, S., Twardowski, B., Trzciński, T., 2023. Looking through the past: Better knowledge retention for generative replay in continual learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, pp. 3496–3500.

Khodadadeh, S., Bölöni, L., Shah, M., 2019. Unsupervised meta-learning for few-shot image classification. arXiv:1811.11819.

Khodadadeh, S., Zehtabian, S., Vahidian, S., Wang, W., Lin, B., Bölöni, L., 2020. Unsupervised meta-learning through latent-space interpolation in generative models. arXiv:2006.10236.

Kirkpatrick, et al., 2017. Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences 114, 3521–3526.

Le, Z., Huang, J., Xu, H., Fan, F., Ma, Y., Mei, X., Ma, J., 2022. Uifgan: An unsupervised continual-learning generative adversarial network for unified image fusion. Information Fusion 88, 305–318. URL: https://www.sciencedirect.com/science/article/pii/S1566253522000756, doi:https://doi.org/10.1016/j.inffus.2022.07.013.

Lin, S., Yang, L., Fan, D., Zhang, J., 2022. Trgp: Trust region gradient projection for continual learning. arXiv:2202.02931.

Liu, Q., Majumder, O., Achille, A., Ravichandran, A., Bhotika, R., Soatto, S., 2020a. Incremental meta-learning via indirect discriminant alignment. arXiv:2002.04162.

Liu, X., Wu, C., Menta, M., Herranz, L., Raducanu, B., Bagdanov, A.D., Jui, S., van de Weijer, J., 2020b. Generative feature replay for class-incremental learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 226–227.

Loo, N., Swaroop, S., Turner, R.E., 2020. Generalized variational continual learning. arXiv:2011.12328.

Lopez-Paz, D., Ranzato, M., 2022. Gradient episodic memory for continual learning. arXiv:1706.08840.

Madaan, D., Yoon, J., Li, Y., Liu, Y., Hwang, S.J., 2022. Representational continuity for unsupervised continual learning. arXiv:2110.06976.

Mallya, A., Davis, D., Lazebnik, S., 2018. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. arXiv:1801.06519.

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.

McCloskey, M., Cohen, N.J., 1989. Catastrophic interference in connectionist networks: The sequential learning problem, in: Psychology of Learning and Motivation. Elsevier. volume 24, pp. 109–165.

Mirzadeh, S.I., Farajtabar, M., Pascanu, R., Ghasemzadeh, H., 2020. Understanding the role of training regimes in continual learning. arXiv:2006.06958.

Ostapenko, O., Lesort, T., Rodríguez, P., Arefin, M.R., Douillard, A., Rish, I., Charlin, L., 2022. Continual learning with foundation models: An empirical study of latent replay. arXiv:2205.00329.

Purushwalkam, S., Morgado, P., Gupta, A., 2022. The challenges of continuous self-supervised learning. arXiv:2203.12710.

Ramasesh, V.V., Dyer, E., Raghu, M., 2020. Anatomy of catastrophic forgetting: Hidden representations and task semantics. ArXiv abs/2007.07400. URL: https://api.semanticscholar.org/CorpusID:220525308.

Ramesh, R., Chaudhari, P., 2022. Model zoo: A growing "brain" that learns continually. arXiv:2106.03027.

Riemer, M., Cases, I., Ajemian, R., Liu, M., Rish, I., Tu, Y., Tesauro, G., 2019. Learning to learn without forgetting by maximizing transfer and minimizing interference. arXiv:1810.11910.

Seff, A., Beatson, A., Suo, D., Liu, H., 2017. Continual learning in generative adversarial nets. arXiv:1705.08395.

Shi, G., Chen, J., Zhang, W., Zhan, L.M., Wu, X.M., 2021. Overcoming catastrophic forgetting in incremental few-shot learning by finding flat minima. arXiv:2111.01549.

Shi, Y., Zhou, K., Liang, J., Jiang, Z., Feng, J., Torr, P., Bai, S., Tan, V.Y.F., 2022. Mimicking the oracle: An initial phase decorrelation approach for class incremental learning. arXiv:2112.04731.

Shin, H., Lee, J.K., Kim, J., Kim, J., 2017. Continual learning with deep generative replay. arXiv:1705.08690.

Shon, H., Lee, J., Kim, S.H., Kim, J., 2022. Dlcft: Deep linear continual fine-tuning for general incremental learning. arXiv:2208.08112.

Stoychev, S., Churamani, N., Gunes, H., 2023. Latent generative replay for resource-efficient continual learning of facial expressions, in: 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG), pp. 1–8. doi:10.1109/FG57933.2023.10042642.

Toneva, M., et al., 2018. An empirical study of example forgetting during deep neural network learning. ArXiv abs/1812.05159. URL: https://api.semanticscholar.org/CorpusID:55481903.

Vitter, J.S., 1985. Random sampling with a reservoir. ACM Trans. Math. Softw. 11, 37–57. URL: https://doi.org/10.1145/3147.3165, doi:10.1145/3147.3165.

Wang, L., Lei, B., Li, Q., Su, H., Zhu, J., Zhong, Y., 2022. Triple-memory networks: A brain-inspired method for continual learning. IEEE Transactions on Neural Networks and Learning Systems 33, 1925–1934. URL: http://dx.doi.org/10.1109/TNNLS.2021.3111019, doi:10.1109/tnnls.2021.3111019.

Wang, L., Yang, K., Li, C., Hong, L., Li, Z., Zhu, J., 2021. Ordisco: Effective and efficient usage of incremental unlabeled data for semi-supervised continual learning. arXiv:2101.00407.

Wang, L., Zhang, X., Su, H., Zhu, J., 2023. A comprehensive survey of continual learning: Theory, method and application. N/A Preprint or Unpublished Manuscript.

Wang, Z., et al., 2022. Learning to prompt for continual learning. arXiv:2112.08654.

Yoon, J., Kim, S., Yang, E., Hwang, S.J., 2020. Scalable and order-robust continual learning with additive parameter decomposition. arXiv:1902.09432.

Yoon, J., Yang, E., Lee, J., Hwang, S.J., 2018. Lifelong learning with dynamically expandable networks. arXiv:1708.01547.

Yu, X., Guo, Y., Gao, S., Rosing, T., 2023. Scale: Online self-supervised lifelong learning without prior knowledge, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 2484–2495.

Zhang, G., Wang, L., Kang, G., Chen, L., Wei, Y., 2023. Slca: Slow learner with classifier alignment for continual learning on a pre-trained model. arXiv:2303.05118.

Zhang, J.O., Sax, A., Zamir, A., Guibas, L., Malik, J., 2020. Side-tuning: A baseline for network adaptation via additive side networks. arXiv:1912.13503.

Zhu, K., Zhai, W., Cao, Y., Luo, J., Zha, Z.J., 2022. Self-sustaining representation expansion for non-exemplar class-incremental learning. arXiv:2203.06359.

Zhuo, T., Cheng, Z., Gao, Z., Kankanhalli, M., 2023. Continual learning with strong experience replay. arXiv:2305.13622.