# Continual Learning

Vincenzo Colella 1748193

October 2022

## Contents

# 1 Introduction

It is clear that humans and many other animals acquire knowledge over the course of their lifetimes in order to develop their skills. The success of (un)supervised learning in the modern era of machine learning has largely depended on the availability of data. However, a lot of real-world applications call for a break from this premise. For instance, a hospital may be legally required to permanently lose access to old patient data, training on all data may be too slow to update a real-time system quickly, or we simply desire to keep improving our models by adding data to the previously learned information. Continuous Learning paradigm refers to a set of approaches to continuous learning that use input in sequential order to continuously learn. Such continuous task learning has proven challenging, despite the numerous successes of deep learning in recent years in industries ranging from image recognition to machine translation. Catastrophic forgetting, a phenomenon seen in neural networks where learning a new task significantly worsens performance on previous tasks, memory limitations, which consider that the amount of storage capacity for the model parameters and results of earlier operations is limited, and increasing model capacity while encouraging the reuse of the learned parameters are a few of the major problems that come with these tasks. Recently, a wide range of continuous learning strategies have been developed, and in this paper we will analyze few of them. We will separate these methods into four main categories, even if many other categorizations are possible:

- Regularization-based methods
- Replay methods
- Parameter Isolation methods

# 2 Regularization-based methods

## 2.1 Memory-aware synapsis (MAS)

Developed by Aljundi et al., this method determines the weight's importance by assessing how much a parameter's gradient changes when it is perturbed. It starts from a neural network, in which we fed a new sample. Based on how sensitive the projected output function is to a change in each network parameter, MAS builds up an importance measure for each parameter. Changes to important parameters (from previous data) can then be penalized when learning a new task, thereby guarding against the overwriting of crucial information from prior tasks.

## 2.2 Learning without forgetting (LWF)

In the paper about DNN knowledge distillation published by Li et Al, the authors give the basis of the LWF methodology. It is a straightforward yet effective approach that enables the knowledge of a pre-trained network to be condensed into another smaller network. First, the output of the pre-trained model is computed for each new task, and the response is used as a fictitious label for the data for the new task. Second, they train the network using real labels for the new jobs and faux labels for the old tasks.

$$(1-\lambda)\cdot L_{cross}(\hat{y}, t=\hat{y}_{1h})+\lambda\cdot L_{kdl}(\hat{y}, t=\hat{y}_{lwf})$$

## 2.3 Elastic Weights Consolidation (EWC)

In this scenario, J. Kirkpatrick et al.'s goal is to successfully complete two tasks (A and B), and they have separate solution spaces that successfully complete those two specific tasks. They first begin to train a network using training examples related to task A, and they come up with a precise answer for problem A. Second, they begin training the network to complete job B, although the network has a variety of possible routes to resolve task B. The idea is to limit the solution's travel such that it stays in the region where the two tasks are successfully completed (intersection). If we succeed in doing this, the network excels at both jobs. To obtain this result they employ a specific loss function.

$$L = L_{cross}(\hat{y}, t=\hat{y}_{1h})+\frac{\lambda}{2}\cdot\sum_{k} F_k(\theta_k-\theta_k^*)^2$$

## 2.4 Variational Auto-Regressive Gaussian Processes (VAR-GPs)

In the context of continual learning, Kapoor and al proposed a Variational Auto-Regressive Gaussian Process (VAR-GP), a principled Bayesian updating mechanism. It is based on a novel auto-regressive characterization of the variational distribution, and sparse inducing point approximations are used to scale up inference. Experiments using common continual learning benchmarks show that VAR-GPs can perform well on new tasks while maintaining performance on existing ones, producing outcomes that are on par with those of cutting-edge techniques.

# 3 Replay Methods

## 3.1 Incremental Classifier and Representation Learning (iCARL)

Rebuffi's method, which described a popular method known as incremental Classifier and Representation Learning, is one of the most well-known methods that fall in the replay category. For training purposes during representation learning, iCARL employs both the instances from the new task and the saved data instances. In order to provide the label of the class with the most similar prototype to the given image, iCARL classifies images using a nearest-mean-of-exemplars approach. But the fact that rehearsal technique relies on stored data makes it problematic for a variety of reasons. First, practical restrictions on safety or privacy frequently preclude data storage. Second, expanding the strategy to address problems with multiple responsibilities is difficult. For these reasons, this method was quickly replaced by others.

## 3.2 Experience Replay

Unlike the majority of existing work, David Rolnick does not explicitly indicate task boundaries in his model, even though this is the most typical situation for a learning agent exposed to continuous experience. Although several approaches to preventing catastrophic forgetting have recently been put forth, in this approach they focus on a simple, all-encompassing, and seemingly overlooked solution: using experience replay buffers for all previous events while combining on and off policy learning and utilising behavioural cloning. They demonstrate that while this strategy can still enable speedy acquisition of new skills, it can also significantly lessen catastrophic forgetting and reach performance comparable with methods that demand task identities.

# 4 Parameter Isolation Methods

## 4.1 Weights masks (piggyback)

Weights masks (piggyback) The theory of A. Mallya et al suggested starting with a backbone network that has already been trained. The objective is to identify a mask for each weight that allows the model's behaviour to be changed in order to solve a particular task without affecting performance on previously learned tasks. This methodology guarantees that there will be no catastrophic forgetting, but there are few disadvantages. First, that it is very challenging to forward transfer our knowledge because the backbone is frozen and the masks start from scratch. For these reasons, knowing the task you must complete in order to select the best mask is useful. Another drawback is that it requires task labels.

### 4.2 Hard attention to the task (HAT)

The strategy investigated by J Serrà et al. is comparable to the earlier (Weights masks). They suggest a task-based hard attention strategy in which the learning for the current task is unaffected by the preservation of information from previous tasks. This method doesn't begin with a model that has already been trained; instead, everything is learned by understanding the attention mask. Since this mask is initially soft and can be trained using backpropagation, the hardness can be changed by adding various hyperparameters. This technique also takes into account the forward transfer, which is based on cumulative attention and states that the important elements from the prior task should be kept for the subsequent task. The outcomes show a good balance between accuracy and effectiveness, but the method suffers from the same flaw as piggybacking (labels needed).

## 5 Hybrid Methods

Other approaches employ two or three different strategies combined. One of these strategies is AR1, which combines replay and regularization techniques. The network consists of two primary parts: 1) Use online synaptic intelligence (regularization) for low-level generic characteristics. 2) Class-specific distinguishing attributes. A linear classifier output layer is found above (using the CWR method). The latent space, where replay methods are applied, is found in the centre. Online learning is a good fit for AR1 because of its very low overhead (in terms of memory and computing). When evaluated on CORe50 and iCIFAR-100, AR1 performed significantly better than the currently used regularization techniques. The complexity and difficulty of parametrizing the hybrid strategy is an important drawback, but we can enhance the trade-off between efficacy and efficiency.

## 6 Conclusions

The deep learning community has given continuous learning a lot of attention, and various effective methods have been proposed to implement it on neural networks. Despite important developments, no current approach can adequately address the problem of continual learning. The models discussed above remain far from producing the flexibility, robustness, and scalability that scientists expect. To combine various techniques and develop some comprehensive solutions to the mentioned challenges, more study is needed. Essentially, we should not neglect the earlier work done and try to adapt the current approaches with appropriate modifications.

# Authors

**T. Lesort, V. Lomonaco, A. Stoian, D. Maltoni, David Filliat, and N. Díaz-Rodríguez** "Continual Learning for Robotics." arXiv preprint arXiv:1907.00182, pages 1–34, 2019.

**J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins** "Overcoming catastrophic forgetting in neural networks". In Proceedings of the National Academy of Sciences, volume 114, pages 3521–3526, 2017.

**J. von Oswald, C. Henning, Benjamin F. Grewe, Joao Sacramento** "Continual learning with hypernetworks", 2019.

**Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, Christoph H. Lampert** iCaRL: Incremental Classifier and Representation Learning.

**A. Mallya, D. Davis, S. Lazebnik** "Piggyback: Adapting a Single Network to Multiple Tasks by Learning to Mask Weights", 2018.

**J. Serrà, D. Surìs, M. Miron, A. Karatzoglou** "Overcoming Catastrophic Forgetting with Hard Attention to the Task", 2018.

**Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach** Memory Aware Synapses: Learning what (not) to forget.

**Davide Maltoni, Vincenzo Lomonaco** Continuous Learning in Single-Incremental-Task Scenarios.

**David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy P. Lillicrap, Greg Wayne** Experience Replay for Continual Learning.

**Eugenio Culurciello** Continual learning, How to keep learning without forgetting, Medium.