

Memory Aware Synapses: Learning what (not) to forget

Rahaf Aljundi¹, Francesca Babiloni¹, Mohamed Elhoseiny²,
 Marcus Rohrbach², and Tinne Tuytelaars¹

¹ KU Leuven, ESAT-PSI, imec, Belgium

² Facebook AI Research

Abstract. Humans can learn in a continuous manner. Old rarely utilized knowledge can be overwritten by new incoming information while important, frequently used knowledge is prevented from being erased. In artificial learning systems, lifelong learning so far has focused mainly on accumulating knowledge over tasks and overcoming catastrophic forgetting. In this paper, we argue that, given the limited model capacity and the unlimited new information to be learned, knowledge has to be preserved or erased selectively. Inspired by neuroplasticity, we propose a novel approach for lifelong learning, coined Memory Aware Synapses (MAS). It computes the importance of the parameters of a neural network in an unsupervised and online manner. Given a new sample which is fed to the network, MAS accumulates an importance measure for each parameter of the network, based on how sensitive the predicted output function is to a change in this parameter. When learning a new task, changes to important parameters can then be penalized, effectively preventing important knowledge related to previous tasks from being overwritten. Further, we show an interesting connection between a local version of our method and Hebb's rule, which is a model for the learning process in the brain. We test our method on a sequence of object recognition tasks and on the challenging problem of learning an embedding for predicting <subject, predicate, object> triplets. We show state-of-the-art performance and, for the first time, the ability to adapt the importance of the parameters based on unlabeled data towards what the network needs (not) to forget, which may vary depending on test conditions.

1 Introduction

The real and digital world around us evolves continuously. Each day millions of images with new tags appear on social media. Every minute hundreds of hours of video are uploaded on Youtube. This new content contains new topics and trends that may be very different from what one has seen before - think e.g. of new emerging news topics, fashion trends, social media hypes or technical evolutions. Consequently, to keep up to speed, our learning systems should be able to evolve as well.

Yet the dominating paradigm to date, using supervised learning, ignores this issue. It learns a given task using an existing set of training examples. Once the training is finished, the trained model is frozen and deployed. From then on, new incoming data is processed without any further adaptation or customization of the model. Soon, the model becomes outdated. In that case, the training process has to be repeated, using both the previous and new data, and with an extended set of category labels. In a world



Fig. 1: Our continuous learning setup. As common in the LLL literature, tasks are learned in sequence, one after the other. If, in between learning tasks, the agent is active and performs the learned tasks, we can use these unlabeled samples to update importance weights for the model parameters. Data that appears frequently, will have a bigger contribution. This way, the agent learns what is important and should not be forgotten.

like ours, such a practice becomes intractable when moving to real scenarios such as those mentioned earlier, where the data is streaming, might be disappearing after a given period of time or even can’t be stored at all due to storage constraints or privacy issues.

In this setting, lifelong learning (LLL) [24,35,37] comes as a natural solution. LLL studies continual learning across tasks and data, tackling one task at a time, without storing data from previous tasks. The goal is to accumulate knowledge across tasks (typically via model sharing), resulting in a single model that performs well on all the learned tasks. The question then is how to overcome catastrophic forgetting [8,9,20] of the old knowledge when starting a new learning process using the same model.

So far, LLL methods have mostly (albeit not exclusively) been applied to relatively short sequences – often consisting of no more than two tasks (e.g. [16,17,28]), and using relatively large networks with plenty of capacity (e.g. [1,6,33]). However, in a true LLL setting with a never-ending list of tasks, the capacity of the model sooner or later reaches its limits and compromises need to be made. Instead of aiming for no forgetting at all, figuring out what can possibly be forgotten becomes at least as important. In particular, exploiting context-specific test conditions may pay off in this case. Consider for instance a surveillance camera. Depending on how or where it is mounted, it always captures images under particular viewing conditions. Knowing how to cope with other conditions is no longer relevant and can be forgotten, freeing capacity for other tasks. This calls for a LLL method that can learn what (not) to forget using unlabeled test data. We illustrate this setup in Figure 1.

Such adaptation and memory organization is what we also observe in biological neurosystems. Our ability to preserve what we have learned before is largely dependent on how frequent we make use of it. Skills that we practice often, appear to be unforgettable, unlike those that we have not used for a long time. Remarkably, this flexibility and adaptation occur in the absence of any form of supervision. According to Hebbian theory [10], the process at the basis of this phenomenon is the strengthening of synapses connecting neurons that fire synchronously, compared to those connecting neurons with unrelated firing behavior.

In this work, we propose a new method for LLL, coined *Memory Aware Synapses*, or MAS for short, inspired by the model of Hebbian learning in biological systems. Unlike previous works, *our LLL method can learn what parts of the model are important using unlabelled data*. This allows for adaptation to specific test conditions and continuous updating of importance weights. This is achieved by estimating importance weights for

the network parameters without relying on the loss, but by looking at the sensitivity of the output function instead. This way, our method not only avoids the need for labeled data, but importantly it also avoids complications due to the loss being in a local minimum, resulting in gradients being close to zero. This makes our method not only more versatile, but also simpler, more memory-efficient, and, as it turns out, more effective in learning what not to forget, compared to other model-based LLL approaches.

Contributions of this paper are threefold: *First*, we propose a new LLL method *Memory Aware Synapses* (MAS). It estimates importance weights for all the network parameters in an unsupervised and online manner, allowing adaptation to unlabeled data, e.g. in the actual test environment. *Second*, we show how a local variant of MAS is linked to the Hebbian learning scheme. *Third*, we achieve better performance than state-of-the-art, both when using the standard LLL setup and when adapting to specific test conditions, both for object recognition and for predicting $\langle \text{subject}, \text{predicate}, \text{object} \rangle$ triplets, where an embedding is used instead of a softmax output.

In the following we discuss related work in section 2 and give some background information in section 3. Section 4 describes our method and its connection with Hebbian learning. Experimental results are given in section 5 and section 6 concludes the paper.

2 Related Work

While lifelong learning has been studied since a long time in different domains (e.g. robotics [37] or machine learning [30]) and touches upon the broader fields of meta-learning [7] and learning-to-learn [2], we focus in this section on more recent work in the context of computer vision mainly.

The main challenge in LLL is to adapt the learned model continually to new tasks, be it from a similar or a different environment [25]. However, looking at existing LLL solutions, we observe that none of them satisfies all the characteristics one would expect or desire from a lifelong learning approach (see Table 1). First, its memory should be constant w.r.t. the number of tasks, to avoid a gradual increase in memory consumption over time. not be limited to a specific setting (e.g. only classification). We refer to this as problem agnostic. Third, given a pretrained model, it should be able to build on top of it and add new tasks. Fourth, being able to learn from unlabeled data would increase the method applicability to cases where original training data no longer exists. Finally, as argued above, within a fixed capacity network, being able to adapt what not to forget to a specific user setting would leave more free capacity for future tasks. In light of these properties, we discuss recently proposed methods. They can be divided into two main approaches: data-based and model-based approaches. Here, we don't consider LLL methods that require storing samples, such as [18,29].

Data-based approaches [1,17,28,34] use data from the new task to approximate the performance of the previous tasks. This works best if the data distribution mismatch

Method	Type	Constant Memory	Problem agnostic	On Pre-trained	Unlabeled data	Adaptive
LwF [17]	data	✓	X	✓	n/a	X
EBLL [28]	data	X	X	X	n/a	X
EWC [12]	model	✓	✓	✓	X	X
IMM [16]	model	X	✓	✓	X	X
SI [39]	model	✓	✓	X	X	X
MAS (our)	model	✓	✓	✓	✓	✓

Table 1: LLL desired characteristics and the compliance of methods, that treat forgetting without storing the data, to these characteristics.

between tasks is limited. Data based approaches are mainly designed for a classification scenario and overall, the need of these approaches to have a preprocessing step before each new task, to record the targets for the previous tasks is an additional limitation.

Model-based approaches [6,12,16,39], like our method, focus on the parameters of the network instead of depending on the task data. Most similar to our work are [12,39]. Like them, we estimate an importance weight for each model parameter and add a regularizer when training a new task that penalizes any changes to important parameters. The difference lies in the way the importance weights are computed. In the *Elastic Weight Consolidation* work [12], this is done based on an approximation of the diagonal of the Fisher information matrix. In the *Synaptic Intelligence* work [39], importance weights are computed during training in an online manner. To this end, they record how much the loss would change due to a change in a specific parameter and accumulate this information over the training trajectory. However, also this method has some drawbacks: 1) Relying on the weight changes in a batch gradient descent might overestimate the importance of the weights, as noted by the authors. 2) When starting from a pre-trained network, as in most practical computer vision applications, some weights might be used without big changes. As a result, their importance will be underestimated. 3) The computation of the importance is done during training and fixed later. In contrast, we believe the importance of the weights should be able to adapt to the test data where the system is applied to. In contrast to the above two methods, **we propose** to look at the **sensitivity of the learned function, rather than the loss**. This simplifies the setup considerably since, unlike the loss, the learned function is not in a local minimum, so complications with gradients being close to zero are avoided.

In this work, we propose a model-based method that computes the importance of the network parameters not only in an online manner but also adaptive to the data that the network is tested on in an unsupervised manner. While previous works [26,31] adapt the learning system at prediction time in a transductive setting, our goal here is to build a continual system that can adapt the importance of the weights to what the system needs to remember. Our method requires a constant amount of memory and enjoys the main desired characteristics of lifelong learning we listed above while achieving state-of-the-art performance.

3 Background

Standard LLL setup. Before introducing our method, we briefly remind the reader of the standard LLL setup, as used, e.g., in [1,16,17,28,39]. It focuses on image classification and consists of a sequence of disjoint *tasks*, which are learned one after the other. Tasks may correspond to different datasets, or different splits of a dataset, without overlap in category labels. The assumption of this setup is that, when training a task, only the **data related to that task is accessible**. Ideally, newer tasks can benefit from the representations learned by older tasks (forward transfer). Yet in practice, the biggest challenge is to avoid catastrophic forgetting of the old tasks' knowledge (i.e., forgetting how to perform the old tasks well). This is a far more challenging setup than joint learning, as typically used in the multitask learning literature, where all tasks are trained simultaneously.

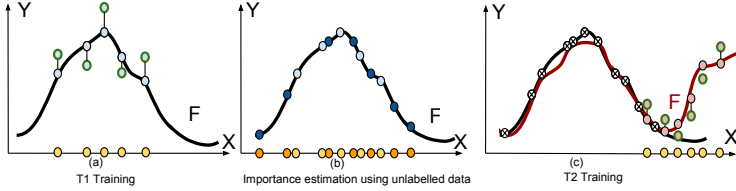


Fig. 2: [39,12] estimate the parameters importance based on the loss, comparing the network output (light blue) with the ground truth labels (green) using training data (in yellow) (a). In contrast, we estimate the parameters importance, after convergence, based on the sensitivity of the learned function to their changes (b). This allows using additional unlabeled data points (in orange). When learning a new task, changes to important parameters are penalized, the function is preserved over the domain densely sampled in (b), while adjusting not important parameters to ensure good performance on the new task (c).

Notations. We train a single, shared neural network over a sequence of tasks. The parameters $\{\theta_{ij}\}$ of the model are the weights of the connections between pairs of neurons n_i and n_j in two consecutive layers³. As in other model-based approaches, our goal is then to compute an importance value Ω_{ij} for each parameter θ_{ij} , indicating its importance with respect to the previous tasks. In a learning sequence, we receive a sequence of tasks $\{T_n\}$ to be learned, each with its training data (X_n, \hat{Y}_n) , with X_n the input data and \hat{Y}_n the corresponding ground truth output data (labels). Each task comes with a task-specific loss L_n , that will be combined with an extra loss term to avoid forgetting. When the training procedure converges to a local minimum, the model has learned an approximation F of the true function \bar{F} . F maps a new input X to the outputs Y_1, \dots, Y_n for tasks $T_1 \dots T_n$ learned so far.

4 Our Approach

In the following, we introduce our approach. Like other model-based approaches [12,39], we estimate an importance weight for each parameter in the network. Yet in our case, these importance weights approximate the *sensitivity of the learned function* to a parameter change rather than a measure of the (inverse of) parameter uncertainty, as in [12], or the sensitivity of the loss to a parameter change, as in [39] (see Figure 2). As it does not depend on the ground truth labels, our approach allows computing the importance using any available data (unlabeled) which in turn allows for an adaptation to user-specific settings. In a learning sequence, we start with task T_1 , training the model to minimize the task loss L_1 on the training data (X_1, \hat{Y}_1) – or simply using a pretrained model for that task.

4.1 Estimating parameter importance

After convergence, the model has learned an approximation F of the true function \bar{F} . F maps the input X_1 to the output Y_1 . This mapping F is the target we want to preserve

³ In convolutional layers, parameters are shared by multiple pairs of neurons. For the sake of clarity, yet without loss of generality, we focus here on fully connected layers.

while learning additional tasks. To this end, we measure how sensitive the function F output is to changes in the network parameters. For a given data point x_k , the output of the network is $F(x_k; \theta)$. A small perturbation $\delta = \{\delta_{ij}\}$ in the parameters $\theta = \{\theta_{ij}\}$ results in a change in the function output that can be approximated by:

$$F(x_k; \theta + \delta) - F(x_k; \theta) \approx \sum_{i,j} g_{ij}(x_k) \delta_{ij} \quad (1)$$

where $g_{ij}(x_k) = \frac{\partial(F(x_k; \theta))}{\partial \theta_{ij}}$ is the gradient of the learned function with respect to the parameter θ_{ij} evaluated at the data point x_k and δ_{ij} is the change in parameter θ_{ij} . Our goal is to preserve the prediction of the network (the learned function) at each observed data point and prevent changes to parameters that are important for this prediction.

Based on equation 1 and assuming a small constant change δ_{ij} , we can measure the importance of a parameter by the magnitude of the gradient g_{ij} , i.e. how much does a small perturbation to that parameter change the output of the learned function for data point x_k . We then accumulate the gradients over the given data points to obtain importance weight Ω_{ij} for parameter θ_{ij} :

$$\Omega_{ij} = \frac{1}{N} \sum_{k=1}^N \|g_{ij}(x_k)\| \quad (2)$$

This equation can be updated in an online fashion whenever a new data point is fed to the network. N is the total number of data points at a given phase. Parameters with small importance weights do not affect the output much, and can, therefore, be changed to minimize the loss for subsequent tasks, while parameters with large weights should ideally be left unchanged.

When the output function F is multi-dimensional, as is the case for most neural networks, equation 2 involves computing the gradients for each output, which requires as many backward passes as the dimensionality of the output. As a more efficient alternative, we propose to use the gradients of the squared ℓ_2 norm of the learned function output⁴, i.e., $g_{ij}(x_k) = \frac{\partial(\ell_2^2(F(x_k; \theta)))}{\partial \theta_{ij}}$. The importance of the parameters is then measured by the sensitivity of the squared ℓ_2 norm of the function output to their changes. This way, we get one scalar value for each sample instead of a vector output. Hence, we only need to compute one backward pass and can use the resulting gradients for estimating the parameters importance. Using our method, for regions in the input space that are sampled densely, the function will be preserved and catastrophic forgetting is avoided. However, parameters not affecting those regions will be given low importance weights, and can be used to optimize the function for other tasks, affecting the function over other regions of the input space.

4.2 Learning a new task

When a new task T_n needs to be learned, we have in addition to the new task loss $L_n(\theta)$, a regularizer that penalizes changes to parameters that are deemed important for previous tasks:

⁴ We square the ℓ_2 norm as it simplifies the math and the link with the Hebbian method, see section 4.3.

$$L(\theta) = L_n(\theta) + \lambda \sum_{i,j} \Omega_{ij} (\theta_{ij} - \theta_{ij}^*)^2 \quad (3)$$

with λ a hyperparameter for the regularizer and θ_{ij}^* the “old” network parameters (as determined by the optimization for the previous task in the sequence, T_{n-1}). As such we allow the new task to change parameters that are not important for the previous task (low Ω_{ij}). The important parameters (high Ω_{ij}) can also be reused, via model sharing, but with a penalty when changing them.

Finally, the importance matrix Ω is to be updated after training a new task, by accumulating over the previously computed Ω . Since we don’t use the loss function, Ω can be computed on any available data considered most representative for test conditions, be it on the last training epoch, during the validation phase or at test time. In the experimental section 5, we show how this allows our method to adapt and specialize to any set, be it from the training or from the test.

4.3 Connection to Hebbian learning

In this section, we propose a local version of our method, by applying it to a single layer of the network rather than to the network as a whole. Next, we show an interesting connection between this local version and Hebbian learning [10].

A local version of our method. Instead of considering the function F that is learned by the network as a whole, we decompose it in a sequence of functions F_l each corresponding to one layer of the network, i.e., $F(x) = F_L(F_{L-1}(\dots(F_1(x))))$, with L the total number of layers. By locally preserving the output of each layer given its input, we can preserve the global function F . This is further illustrated in Figure 3. Note how “local” and “global” in this context relate to the number of layers over which the gradients are computed.

We use y_i^k to denote the activation of neuron n_i for a given input x_k . Analogous to the procedure followed previously, we consider the squared ℓ_2 norm of each layer after the activation function. An infinitesimal change $\delta_l = \{\delta_{ij}\}$ in the parameters $\theta_l = \{\theta_{ij}\}$ of layer l results in a change to the squared ℓ_2 norm of the local function F_l for a given input to that layer $y^k = \{y_i^k\} = F_{l-1}(\dots(F_1(x_k)))$ given by:

$$\ell_2^2(F_l(y^k; \theta_l + \delta_l)) - \ell_2^2(F_l(y^k; \theta_l)) \approx \sum_{i,j} g_{ij}(x_k) \delta_{ij} \quad (4)$$

where $g_{ij}(x_k) = \frac{\partial[\ell_2^2(F_l(y^k; \theta_l))]}{\partial \theta_{ij}}$. In the case of a ReLU activation function, it can be shown that (see Appendix.B):

$$g_{ij}(x_k) = 2 * y_i^k * y_j^k \quad (5)$$

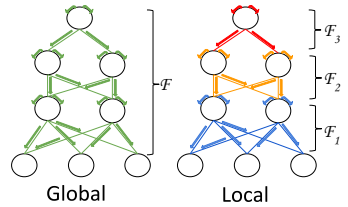


Fig. 3: Gradients flow for computing the importance weight. Local considers the gradients of each layer independently.

Again we consider the accumulation of the gradients evaluated at different data points $\{x_k\}$ as a measure for the importance of the parameter θ_{ij} :

$$\Omega_{ij} = \frac{1}{N} \sum_{k=1}^N g_{ij}(x_k) = 2 * \frac{1}{N} \sum_{k=1}^N y_i^k * y_j^k \quad (6)$$

Link with Hebbian theory. In neuroscience, Hebbian learning theory [10] provides an explanation for the phenomenon of synaptic plasticity. It postulates that “cells that fire together, wire together”: the synapses (connections) between neurons that fire synchronously for a given input are strengthened over time to maintain and possibly improve the corresponding outputs. Here we reconsider this theory from the perspective of an artificial neural network after it has been trained successfully with backpropagation. Following Hebb’s rule, parameters connecting neurons that often fire together (high activations for both, i.e. highly correlated outputs) are more important for the given task than those that fire asynchronously or with low activations. As such, the importance weight Ω_{ij} for the parameter θ_{ij} can be measured purely locally in terms of the correlation between the neurons’ activations, i.e.

$$\Omega_{ij} = \frac{1}{N} \sum_{k=1}^N y_i^k * y_j^k \quad (7)$$

The similarity with equation 6 is striking. We can conclude that applying Hebb’s rule to measure the importance of the parameters in a neural network can be seen as a local variant of our method that considers only one layer at a time instead of the global function learned by the network. Since only the relative importance weights really matter, the scale factor 2 can be ignored.

4.4 Discussion

Our global and local methods both have the advantage of computing the importance of the parameters on any given data point without the need to access the labels or the condition of being computed while training the model. The global version needs to compute the gradients of the output function while the local variant (Hebbian based) can be computed locally by multiplying the input with the output of the connecting neurons. *Our proposed method (both the local and global version) resembles an implicit memory included for each parameter of the network. We, therefore, refer to it as Memory Aware Synapses.* It keeps updating its value based on the activations of the network when applied to new data points. It can adapt and specialize to a given subset of data points rather than preserving every functionality in the network. Further, the method can be added after the network is trained. It can be applied on top of any pretrained network and compute the importance on any set of data without the need to have the labels. This is an important criterion that differentiates our work from methods that rely on the loss function to compute the importance of the parameters.

5 Experiments

We start by comparing our method to different existing LLL methods in the standard sequential learning setup of object recognition tasks. We further analyze the behavior and some design choices of our method. Next, we move to the more challenging problem of continual learning of $\langle \text{subject, predicate, object} \rangle$ triplets in an embedding space (section 5.2).

5.1 Object Recognition

We follow the standard setup commonly used in computer vision to evaluate LLL methods [1,17,28]. It consists of a sequence of supervised classification tasks each from a particular dataset. Note that this assumes having different classification layers for each task (different “heads”) that remain unshared. Moreover, an oracle is used at test time to decide on the task (i.e., which classification layer to use).

Compared Methods. - *Finetuning* (FineTune). After learning the first task and when receiving a new task to learn, the parameters of the network are finetuned on the new task data. This baseline is expected to suffer from forgetting the old tasks while being advantageous for the new task.

- *Learning without Forgetting* [17] (LwF). Given a new task data, the method records the probabilities obtained from the previous tasks heads and uses them as targets when learning a new task in a surrogate loss function. To further control the forgetting, the method relies on first training the new task head while freezing the shared parameters as a warmup phase and then training all the parameters until convergence.

- *Encoder Based Lifelong Learning* [28] (EBLL) builds on LwF and learns a shallow encoder on the features of each task. A penalty on the changes to the encoded features accompanied with the distillation loss is applied to reduce the forgetting of the previous tasks. Similar to LwF, a warmup phase is used before the actual training phase.

- *Incremental Moment Matching* [16] (IMM). A new task is learned with an L2 penalty equally applied to the changes to the shared parameters. At the end of the sequence, the obtained models are merged through a first or second moment matching. In our experiments, mean IMM gives better results on the two tasks experiments while mode IMM wins on the longer sequence. Thus, we report the best alternative in each experiment.

- *Elastic Weight Consolidation* [12] (EWC). It is the first work that suggests regularizing the network parameters while learning a new task using as importance measure the diagonal of the Fisher information matrix. EWC uses individual penalty for each previous task, however, to make it computationally feasible we apply a single penalty as pointed out by [11]. Hence, we use a running sum of the Fishers in the 8 tasks sequence.

- *Synaptic Intelligence* [39] (SI). This method shows state-of-the-art performance and comes closest to our approach. It estimates the importance weights in an online manner while training for a new task. Similar to EWC and our method changes to parameters important for previous tasks are penalized during training of later tasks.

- *Memory Aware Synapses* (MAS). Unless stated otherwise, we use the global version of our method and with the importance weights estimated only on training data. We use a regularization parameter λ of 1; note that no tuning of λ was performed as we assume no access to previous task data.

Method	Birds \rightarrow Scenes		Scenes \rightarrow Birds		Flower \rightarrow Birds		Flower \rightarrow Scenes	
FineTune	45.20 (-8.0)	57.8	49.7 (-9.3)	52.8	64.87 (-13.2)	53.8	70.17 (-7.9)	57.31
LwF [17]	51.65 (-2.0)	55.59	55.89 (-3.1)	49.46	73.97 (-4.1)	53.64	76.20 (-1.9)	58.05
EBLL [28]	52.79 (-0.8)	55.67	56.34 (-2.7)	49.41	75.45 (-2.6)	50.51	76.20 (-1.9)	58.35
IMM [16]	51.51 (-2.1)	52.62	54.76 (-4.2)	52.20	75.68 (-2.4)	48.32	76.28 (-1.8)	55.64
EWC [12]	52.19 (-1.4)	55.74	58.28 (-0.8)	49.65	76.46 (-1.6)	50.7	77.0 (-1.1)	57.53
SI [39]	52.64 (-1.0)	55.89	57.46 (-1.5)	49.70	75.19 (-2.9)	51.20	76.61 (-1.5)	57.53
MAS (ours)	53.24 (-0.4)	55.0	57.61 (-1.4)	49.62	77.33 (-0.7)	50.39	77.24 (-0.8)	57.38

Table 2: Classification accuracy (%), drop in first task (%) for various sequences of 2 tasks using the object recognition setup.

Experimental setup. We use the AlexNet [15] architecture pretrained on Imagenet [32] from [14]⁵. All the training of the different tasks have been done with stochastic gradient descent for 100 epochs and a batch size of 200 using the same learning rate as in [1]. Performance is measured in terms of classification accuracy.

Two tasks experiments. We first consider sequences of two tasks based on three datasets: MIT *Scenes* [27] for indoor scene classification (5,360 samples), Caltech-UCSD *Birds* [38] for fine-grained bird classification (5,994 samples), and Oxford *Flowers* [23] for fine-grained flower classification (2,040 samples). We consider: Scene \rightarrow Birds, Birds \rightarrow Scenes, Flower \rightarrow Scenes and Flower \rightarrow Birds, as used previously in [1, 17, 28]. We didn't consider Imagenet as a task in the sequence as this would require retraining the network from scratch to get the importance weights for SI. As shown in Table 2, FineTune clearly suffers from catastrophic forgetting with a drop in performance from 8% to 13%. All the considered methods manage to reduce the forgetting over fine-tuning significantly while having performance close to fine-tuning on the new task. On average, our method achieves the lowest forgetting rates (around 1%) while performance on the new task is almost similar (0 – 3% lower).

Local vs. global MAS on training/test data. Next we analyze the performance of our method when preserving the global function learned by the network after each task (MAS) and its local Hebbian-inspired variant described in section 4.3 (1-MAS). We also evaluate our methods, MAS and 1-MAS, when using unlabeled test data and/or labeled training data. Table 3 shows, independent from the set used for computing the importance of the weights, for both 1-MAS and MAS the preservation of the previous task and the performance on the current task are quite similar. This illustrates our method ability to estimate the parameters importance of a given task given any set of points, without the need for labeled data. Further, computing the gradients locally at each layer for 1-MAS allows for faster computations but less accurate estimations. As such, 1-MAS shows an average forgetting of 3% compared to 1% by MAS.

ℓ_2^2 vs. vector output. We explained in section 4 that considering the gradients of the learned function to estimate the parameters importance would require as many backward passes as the length of the output vector. To avoid this complexity, we suggest using the square of the ℓ_2 norm of the function to get a scalar output. We run two experiments, Flower \rightarrow Scenes and Flower \rightarrow Birds once with computing the gradients with respect to the vector output and once with respect to the ℓ_2^2 norm. We observe no

⁵ We use the pretrained model available in Pytorch. Note that it differs slightly from other implementations used e.g. in [17].

Method	Ω_{ij} computed. on	Birds \rightarrow Scenes		Scenes \rightarrow Birds		Flower \rightarrow Bird		Flower \rightarrow Scenes	
MAS	Train	53.24 (-0.4)	55.0	57.61 (-1.4)	49.62	77.33 (-0.7)	50.39	77.24 (-0.8)	57.38
MAS	Test	53.43 (-0.2)	55.07	57.31 (-1.7)	49.01	77.62 (-0.5)	50.29	77.45 (-0.6)	57.45
MAS	Train + Test	53.29 (-0.3)	56.04	57.83 (-1.2)	49.56	77.52 (-0.6)	49.70	77.54 (-0.5)	57.39
1-MAS	Train	51.36 (-2.3)	55.67	56.79 (-2.2)	49.08	73.96 (-4.1)	50.5	76.20 (-1.9)	56.68
1-MAS	Test	51.62 (-2.0)	53.95	55.74 (-3.3)	50.43	74.48 (-3.6)	50.32	76.56 (-1.5)	57.83
1-MAS	Train + Test	52.15 (-1.5)	54.40	56.79 (-2.2)	48.92	73.73 (-4.3)	50.5	76.41 (-1.7)	57.91

Table 3: Classification accuracies (%) for the object recognition setup - comparison between using Train and Test data (unlabeled) to compute the parameter importance Ω_{ij} .

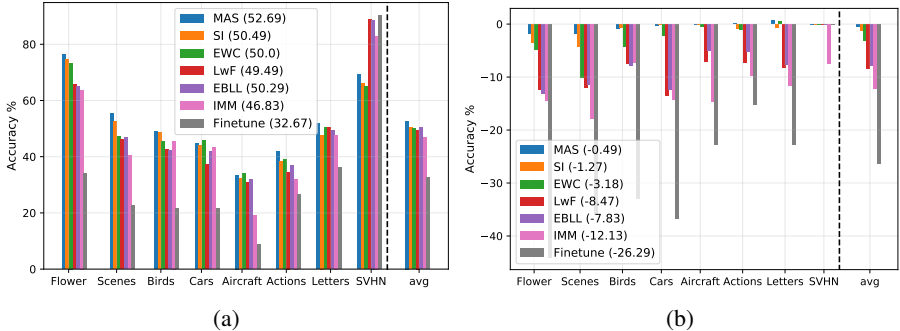


Fig. 5: 5a performance on each task, in accuracy, at the end of 8 tasks object recognition sequence. 5b drop in each task relative to the performance achieved after training each task.

significant difference on forgetting over 3 random trials where we get a mean, over 6 numbers, of $0.51\% \pm 0.18$ for the drop on the first task in the vector output case compared to $0.50\% \pm 0.19$ for the ℓ_2^2 norm case. No significant difference is observed on the second task either. As such, using ℓ_2^2 is n times faster (where n is the length of the output vector) without loss in performance.

Longer Sequence

While the two tasks setup gives a detailed look at the average expected forgetting when learning a new task, it remains easy. Thus, we next consider a sequence of 8 tasks.

To do so, we add five more datasets: Stanford Cars [13] for fine-grained car classification; FGVC-Aircraft [19] for fine-grained aircraft classification; VOC Actions, the human action classification subset of the VOC challenge 2012 [5]; Letters, the Chars74K dataset [3] for character recognition in natural images; and the Google Street View House Numbers SVHN dataset [22] for digit recognition.

Those datasets were also used in [1]. We run the different methods on the following sequence: Flower \rightarrow Scenes \rightarrow Birds \rightarrow Cars \rightarrow Aircraft \rightarrow Actions \rightarrow Letters \rightarrow SVHN.

While Figure 5a shows the performance on each task at the end of the sequence, 5b shows the observed forgetting on each task at the end of the sequence (relative to the

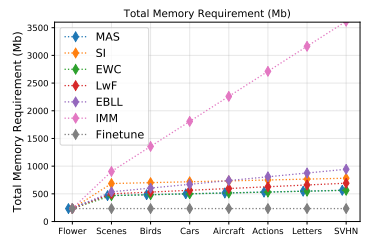


Fig. 4: Overall memory requirement for each method at each step of the sequence.

performance right after training that task). The differences between the compared methods become more outspoken. `Finetuning` suffers from a severe forgetting on the previous tasks while being advantageous for the last task, as expected. `LwF` [17] suffers from a buildup of errors when facing a long sequence while `EBLL` [28] reduces slightly this effect. `IMM` [16] merges the models at the end of the sequence and the drop in performance differs between tasks. More importantly, the method performance on the last task is highly affected by the moment matching. `SI` [39] followed by `EWC` [12] has the least forgetting among our methods competitors. `MAS`, our method, shows a minimal or no forgetting on the different tasks in the sequence with an average forgetting of 0.49%. It is worth noting that our method’s absolute performance on average including the last task is 2% better than `SI` which indicates our method ability to accurately estimate the importance weights and the new tasks to adjust accordingly. Apart from evaluating forgetting, we analyze the memory requirements of each of the compared methods. Figure 4 illustrates the memory usage of each method at each learning step in the sequence. After `Finetune` that doesn’t treat forgetting, our method has the least amount of memory consumption. Note that `IMM` grows linearly in storage, but at inference time it only uses the obtained model. More details on memory requirements and absolute performances, in numbers, achieved by each method can be found in the Appendix C2.

Sensitivity to the hyper parameter. Our method needs one extra hyper parameter, λ , that weights the penalty on the parameters changes as shown in Eq 3. λ is a trade-off between the allowed forgetting and the new task loss. We set λ to the largest value that allows an acceptable performance on the new task. For `MAS`, we used $\lambda = 1$ in all object recognition experiments while for `SI`[39] and `EWC`[12] we had to vary λ . Figure 6 shows the effect of λ on the avg. performance and the avg. forgetting in a sequence of 5 permuted MNIST tasks with a 2 layer perceptron (512 units). We see the sensitivity around $\lambda = 1$ is very low with low forgetting, although further improvements could be achieved.

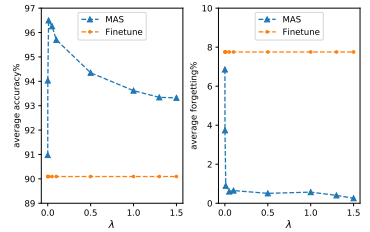


Fig. 6: avg. performance, left, and avg. forgetting, right, on permuted mnist sequence.

Adaptation Test. As we have previously explained, `MAS` has the ability to adapt the importance weights to a specific subset that has been encountered at test time in an unsupervised and online manner. To test this claim, we have selected one class from the `Flower` dataset, `Krishna Kamal flower`. We learn the 8 tasks sequence as above while assuming `Krishna Kamal` as the only encountered class. Hence, importance weights are computed on that subset only. At the end of the sequence, we observe a minimal forgetting on that subset of 2% compared to 8% forgetting on the `Flower` dataset as a whole. We also observe higher accuracies on later tasks as only changes to important parameters for that class are penalized, leaving more free capacity for remaining tasks (e.g. accuracy of 84% on the last task, instead of 69% without adaptation). We repeat the experiment with two other classes and obtain similar results. This clearly indicates our method ability to adapt to user specific settings and to learn what (not) to forget.

5.2 Facts Learning

Next, we move to a more challenging setup where all the layers of the network are shared, including the last layer. Instead of learning a classifier, we learn an embedding space. For this setup, we pick the problem of Fact Learning from natural images [4]. For example, a fact could be “person eating pizza”. We design different experimental settings to show the ability of our method to learn what (not) to forget.

Experimental setup. We use the 6DS mid scale dataset presented in [4]. It consists of 28,624 images, divided equally in training and test samples belonging to 186 unique facts. Facts are structured into 3 units: Subject (S), Object (O) and Predicate (P). We use a CNN model based on the VGG-16 architecture [36] pretrained on ImageNet. The last fully connected layer forks in three final layers enabling the model to have three separated and structured outputs for Subject, Predicate and Object as in [4]. The loss minimizes the pairwise distance between the visual and the language embedding. For the language embedding, the Word2vec [21] representation of the fact units is used. To study fact learning from a lifelong perspective, we divided the dataset into tasks belonging to different groups of facts. SGD optimizer is used with a mini-batch of size 35 for 300 epochs and we use a $\lambda = 5$ for our method. For evaluation, we report the fact to image retrieval scenario. We follow the evaluation protocol proposed in [4] and report the mean average precision (MAP). For each task, we consider retrieving the images belonging to facts from this task only. We also report the mean average precision on the whole dataset which differs from the average of the performance achieved on each task. More details can be found in the supplemental materials. We focus on the comparison between the local 1-MAS and global MAS variants of our method and SI [39], the best performing method among the different competitors as shown in Figure 5a.

Method	Split	Method evaluated on				
		T_1	T_2	T_3	T_4	all
Finetune	1	0.19	0.19	0.28	0.71	0.18
SI [39]	1	0.36	0.32	0.38	0.68	0.25
MAS (ours)	1	0.42	0.37	0.41	0.65	0.29
Finetune	2	0.20	0.27	0.18	0.66	0.18
SI [39]	2	0.37	0.39	0.38	0.46	0.24
MAS (ours)	2	0.42	0.42	0.46	0.65	0.28
Finetune	3	0.21	0.25	0.24	0.46	0.14
SI [39]	3	0.30	0.31	0.36	0.61	0.24
MAS (ours)	3	0.30	0.36	0.38	0.66	0.27

Table 4: MAP for fact learning on the 4 tasks random split, from the 6DS dataset, at the end of the sequence.

Four tasks experiments We consider a sequence of 4 tasks obtained from randomly splitting the facts of the same dataset into 4 groups. Table 4 presents the achieved performance on each set of the 4 tasks at the end of the learned sequence based on 3 different random splits. Similar to previous experiments, Finetune is only advantageous on the last task while drastically suffering on the previous tasks. However, here, our method differentiates itself clearly, showing 6% better MAP on the first two tasks

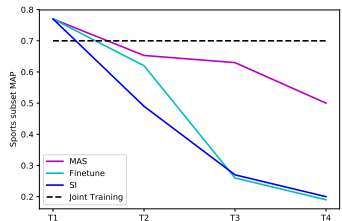


Fig. 7: MAP on the sport subset of the 6DS dataset after each task in a 4 tasks sequence. MAS managed to learn that the sport subset is important to preserve and prevents significantly the forgetting on this subset.

compared to SI. Overall, MAS achieves a MAP of 0.29 compared to 0.25 by SI and only 0.18 by Finetune. When MAS importance weights are computed on both training and test data, a further improvement is achieved with 0.30 overall performance. This highlights our method ability to benefit from extra unlabeled data to further enhance the importance estimation.

Adaptation Test. Finally we want to test the ability of our method in learning not to forget a specific subset of a task. When learning a new task, we care about the performance on that specific set more than the rest. For that reason, we clustered the dataset into 4 disjoint groups of facts, representing 4 tasks, and then selected a specialized subset of T_1 , namely 7 facts of person playing sports. More details on the split can be found in the supplemental material. We run our method with the importance parameters computed only over the examples from this set along the 4 tasks sequence. Figure 7 shows the achieved performance on this sport subset by each method at each step of the learning sequence. Joint Training (black dashed) is shown as reference. It violates the LLL setting as it trains on all data jointly. Note that SI can only learn importance weights during training, and therefore cannot adapt to a particular subset. Our MAS (pink) succeeds to learn that this set is important to preserve and achieves a performance of 0.50 at the end of the sequence, while the performance of finetuning and SI on this set was close to 0.20.

6 Conclusion

In this paper, we argued that, given a limited model capacity and unlimited evolving tasks, it is not possible to preserve all the previous knowledge. Instead, agents should learn what (not) to forget. Forgetting should relate to the rate at which a specific piece of knowledge is used. This is similar to how biological systems are learning. In the absence of error signals, synapses connecting biological neurons strengthen or weaken based on the concurrence of the connected neurons activations. In this work and inspired by the synaptic plasticity, we proposed a method that is able to learn the importance of network parameters from the input data that the system is active on, in an unsupervised manner. We showed that a local variant of our method can be seen as an application of Hebb’s rule in learning the importance of parameters. We first tested our method on a sequence of object recognition problems in a traditional LLL setting. We then moved to a more challenging test case where we learn facts from images in a continuous manner. We showed i) the ability of our method to better learn the importance of the parameters using training data, test data or both; ii) state-of-the-art performance on all the designed experiments and iii) the ability of our method to adapt the importance of the parameters towards a frequent set of data. We believe that this is a step forward in developing systems that can always learn and adapt in a flexible manner.

Acknowledgment: The first author’s PhD is funded by an FWO scholarship.

Appendix

A Additional adaptation experiment

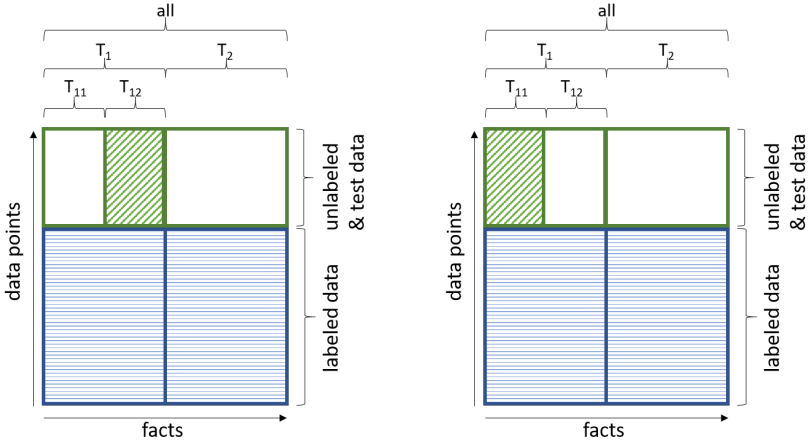


Fig. A1: Visualization to clarify the training and test data splits for results of Table A1 below. “blue horizontal lines” indicate training data for the model parameters θ_{ij} , “green diagonal lines” indicate data used for computing importance weights Ω_{ij} . Note that tasks are split according to facts (x-axis). Evaluation is performed for different splits on test data (green boxes, top), with the importance weights estimated based on T_{11} (right) or T_{12} (left).

As an extra test to show that our method does not just capture general importance weights, but can really adapt, in an unsupervised fashion, to particular test conditions, we present here an additional experiment using the 6DS dataset and the two tasks experiment set up (section 5.2 in the main paper). We split the test set of the first task data further into two random subsets of facts, T_{11} and T_{12} . After learning the task T_1 , the importance of the parameters is computed using one subset only (T_{11} or T_{12} – we show results for both cases). Then the second task T_2 is learned. Figure A1 illustrates the designed setup for the different splits used. Table A1 compares the performance in each case.

We can see that the forgetting on the subset that was used for estimating the importance of the parameters is less than on the other subset that was not considered. For example, MAS learning the importance of the parameters on the first subset T_{11} preserves a performance of **0.472** for T_{11} compared to that of **0.451** when computing the importance on the other batch T_{12} . This can stand as further empirical proof of our method’s capability of learning the importance of the parameters based on what the network is actively tested on.

Method	Ω_{ij}	Method evaluated on				
	comp. on	T_{11}	T_{12}	T_1	T_2	all
1-MAS	T_{11} Test	0.389	0.292	0.251	0.508	0.272
1-MAS	T_{12} Test	0.366	0.319	0.261	0.485	0.283
MAS	T_{11} Test	0.472	0.352	0.318	0.475	0.307
MAS	T_{12} Test	0.451	0.361	0.300	0.467	0.297

Table A1: Adaptation to test condition. Learning importance weights Ω_{ij} on T_{11} vs. T_{12} , two subsets of test split T_1 . Mean average precision for fact learning, two tasks scenario, random split of 6DS dataset. Results for T_1 , T_2 , **all** as reference. Note that, for evaluation, we always use the facts of this (sub)task only, hence one should compare within column only.

B Connection to Hebbian Learning

In this section we provide more details for our derivation in Section 4.3 in the main paper, which shows the connection of our method to Hebbian Learning. As we explained in the main paper, for the local version of our method, 1-MAS, instead of considering the function F that is learned by the network as a whole, we decompose it in a sequence of functions F_l , each corresponding to one layer of the network, i.e. $F(x_k) = F_L(F_{L-1}(\dots(F_1(x_k))))$, with L the total number of layers. To simplify the notations, we will drop the index k referring to the input sample x_k from now on. Rather than preserving the global function F , 1-MAS preserves the output of each layer given its input. Here, we use y_i^l to denote the activation of neuron n_i at layer l for a given input x . Note that, for clarity, we explicitly add a superscript l referring to the layer. Hence $y^l = \{y_j^l\}$ is the output of layer l and $y^{l-1} = \{y_i^{l-1}\}$ is its input. y^l is a vector with elements $\{y_j^l\}$; we use a similar notation in the paper and in the following. We use indices i and j for neurons of the input and output layer respectively. θ_{ij} is then a network parameter representing the connection between neuron n_i in layer $l-1$ and neuron n_j in layer l .

As explained in the main text, we consider the squared ℓ_2 norm of each layer after the activation function. An infinitesimal change $\delta_l = \{\delta_{ij}\}$ in the parameters $\theta_l = \{\theta_{ij}\}$ of layer l results in a change to the squared ℓ_2 norm of the local function F_l for a given input to that layer $y^{l-1} = F_{l-1}(\dots(F_1(x)))$ given by:

$$\ell_2^2(F_l(y^{l-1}; \theta_l + \delta_l)) - \ell_2^2(F_l(y^{l-1}; \theta_l)) \approx \sum_{i,j} g_{ij}(x) \delta_{ij} \quad (B1)$$

where $g_{ij}(x) = \frac{\partial[\ell_2^2(F_l(y^{l-1}; \theta_l))]}{\partial \theta_{ij}}$. In the case of a ReLU activation function and considering a fully connected layer with $I * J$ parameters, i.e. $y_j^l = \text{ReLU}(\text{out}_j^l)$ with $\text{out}_j^l = \sum_{h=1}^I \theta_{hj} * y_h^{l-1}$, we can write

$$g_{ij}(x) = \frac{\partial[\ell_2^2(F_l(y^{l-1}; \theta_l))]}{\partial \theta_{ij}} = \frac{\partial[\sum_{o=1}^J (y_o^l)^2]}{\partial \theta_{ij}} = \sum_{o=1}^J \frac{\partial((y_o^l)^2)}{\partial \theta_{ij}} \quad (B2)$$

Since $\frac{\partial((y_o^l)^2)}{\partial \theta_{ij}} = 0$ when $o \neq j$, we get

$$g_{ij}(x) = \frac{\partial((y_j^l)^2)}{\partial \theta_{ij}} = 2 * y_j^l * \frac{\partial(y_j^l)}{\partial \theta_{ij}} = 2 * y_j^l * \frac{\partial(\text{ReLU}(\text{out}_j^l))}{\partial \theta_{ij}} \quad (B3)$$

ReLU is non smooth since it has a non-linearity at 0. We show the subgradients in the two cases (also using the fact that $\text{out}_j^l = \sum_{h=1}^I \theta_{hj} * y_h^{l-1}$):

1. if $out_j^l > 0$, $ReLU(out_j^l) = out_j^l$ and $(ReLU)' = 1$:

$$\frac{\partial(ReLU(out_j^l))}{\partial \theta_{ij}} = \frac{\partial(ReLU(out_j^l))}{\partial out_j^l} \frac{\partial(out_j^l)}{\partial \theta_{ij}} = \frac{\partial(out_j^l)}{\partial \theta_{ij}} \quad (B4)$$

$$= \frac{\partial(\sum_{h=1}^I \theta_{hj} * y_h^{l-1})}{\partial \theta_{ij}} = \frac{\partial(\theta_{ij} * y_i^{l-1})}{\partial \theta_{ij}} = y_i^{l-1} \quad (B5)$$

$$\Rightarrow g_{ij}(x) = 2 * y_j^l * y_i^{l-1} \quad (B6)$$

2. in the other case, $ReLU(out_j^l) = 0$ and $(ReLU)' = 0$, so

$$g_{ij}(x) = 0 \quad (B7)$$

At the same time,

$$ReLU(out_j^l) = 0 \Rightarrow y_j^l = 0 \text{ and } 2 * y_j^l * y_i^{l-1} = 0 \quad (B8)$$

Hence

$$g_{ij}(x) = 2 * y_j^l * y_i^{l-1} = 0 \quad (B9)$$

Based on equations B6 and B9, we have shown

$$g_{ij}(x) = 2 * y_i^{l-1} * y_j^l \quad (B10)$$

which is remarkably similar to the Hebbian learning rule as explained in the main paper.

C Experimental Details

C1 Performance of each of the compared methods on the 8 tasks experiments

In the main paper, we described a learning sequence composed of 8 tasks and reported the results in a bar plot (Figure 5a in the main paper). To ease the comparison with other methods in future work, we report the performance in accuracies achieved by each method in Table C1.

Table C1: Performance measured in accuracies achieved by each method on each of the learned tasks at the end of the 8 tasks sequence (Table for bar plot in Figure 5a of the main paper).

Method	Flower	Scenes	Birds	Cars	Aircraft	Actions	Letters	SVHN	avg
Finetune	33.98	22.57	21.48	21.48	13.32	26.64	36.18	90.33	32.67
LwF [17]	65.68	46.04	42.47	37.24	30.75	34.55	50.31	88.89	49.491
IMM [16]	67.44	47.08	42.28	39.19	18.93	32.88	46.35	53.15	43.41
SI [39]	74.58	52.53	48.65	43.96	32.19	38.22	47.7	66.09	50.49
MAS	76.34	55.3	49.03	44.59	33.48	41.76	51.93	69.13	52.695

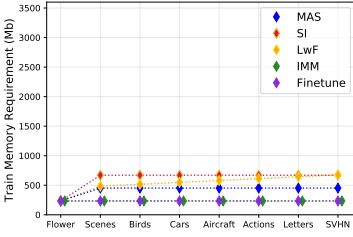


Fig. C1: Memory requirements during training for each method in the 8 tasks sequence.

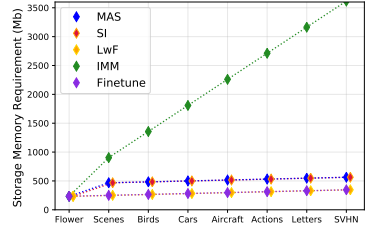


Fig. C2: Storage requirements for each of the compared methods in the 8 tasks sequence.

C2 Details on methods' memory requirements estimation

In the main paper, section 5.1 Figure 6, we report an approximation to the memory requirements of each method in the 8 tasks object recognition sequence. Here, we show a detailed estimation in which we split the overall memory consumption into memory requirements at training phase (Figure C1) and storage requirements in between tasks (Figure C2).

Finetuning needs no additional memory requirements other than the model and the tasks' heads. In spite of the fact that IMM [16] consumes the same amount of memory as Finetuning while training, it requires an offline storage that increases linearly with the number of tasks. That's due to the storage of a Fisher information matrix for each task and previous tasks models. LwF [17], while training a new task, needs to load the last task model and all the previous heads to get the target predictions. However, LwF has the same offline storage requirements between tasks as Finetuning. SI [39] and MAS (ours) need the same offline storage requirements (importance weights and model parameters). However, at training phase, SI needs additional memory to accumulate the contributions of each parameter to the change in the loss.

C3 Fact learning experimental setup

In the main paper, section 5.2 paragraph "Experimental Setup", we explained the protocol followed for the Fact learning experiments. Here, we explain in more details how we get the random split for both two tasks and 4 tasks experiments. We randomly split the facts into different groups and make sure that these groups contain a balanced number of facts and a balanced number of corresponding training and test images by selecting the best candidate out of 100 random trials. Since the "person" fact has the biggest number of images and it is a unit of a good portion of the rest of the facts, we make sure that this fact is in the first group of facts.

C4 Fact learning adaptation test experimental setup

In section 5.2 paragraph "Adaptation Test" of the main paper, we move from the random split setup, as presented in the subsection "Longer sequence of tasks" and above, to a semantic grouping of the facts, in order to make our setup more realistic. To this end, we group the facts by their similarity in the word2vec language space. We use agglomerative clustering to build a dendrogram based on the Euclidean distance between the word2vec embeddings of the facts.

We selected a cut of 4 clusters. Those clusters contain semantically related facts and can be interpreted as

- *Cluster 1*: facts describing human actions such as $\langle \text{person, riding, bike} \rangle$, $\langle \text{person, jumping} \rangle$
- *Cluster 2*: facts of different objects such as $\langle \text{battingball} \rangle$, $\langle \text{battingstumps} \rangle$, $\langle \text{dog} \rangle$, $\langle \text{car} \rangle$, etc.
- *Cluster 3*: facts describing humans holding or playing musical instruments, such as $\langle \text{person, playing, flute} \rangle$, $\langle \text{person, holding, cello} \rangle$, etc.
- *Cluster 4*: facts describing human interactions such as $\langle \text{person, arguing with, person} \rangle$, $\langle \text{person, dancing with, person} \rangle$, etc.

We chose a subset of 7 facts focusing on sports:

1. $\langle \text{person, batting, cricket} \rangle$
2. $\langle \text{person, bowling, cricket} \rangle$
3. $\langle \text{person, forehead, tennis} \rangle$
4. $\langle \text{person, play, croquet} \rangle$
5. $\langle \text{person, serve, tennis} \rangle$
6. $\langle \text{person, smash, volleyball} \rangle$
7. $\langle \text{person, throwing, frisbee} \rangle$

We run our method while supposing that between the learning steps the agent is active on images belonging to these subsets and thus importance weights are computed with respect to these facts and aimed at preserving their performance.

D Histogram of parameters importance Ω

We have shown empirically in the main paper that our proposed method (MAS) is able to identify the important parameters and penalize changing them when learning a new task. To further analyze how the importance values are spread among the different parameters, we plotted the histogram of Ω (the parameter importance). Ideally, a good importance measure would give very low importance values to the unused parameters and high values for those that are crucial for the task at hand.

Part (a) of Figure D1 shows the histogram of Ω of the last shared convolutional layer computed on the training data from the first task. This is based on the two tasks experiments under the fact learning setting (section 5.2). We can notice how the histogram has a peak at a value close to zero and then goes flat. Part (b) of Figure D1 shows the same histogram but magnified in the area covering the 1000 top most important parameters. We can see the long tail distribution and how the values get sparser the more we move to higher importance assignment. This indicates that our method (MAS) will allow changes on most of the parameters that were unused by the first task while penalizing changes to those few crucial parameters that carry meaningful information for the learned task.

E Correlation between the parameters importance computed on different sets

In the main paper, we have conducted experiments to examine our method’s ability to preserve the previous task’s performance by computing the importance of the parameters on different sets, e.g. train, test or both (see Tables 2 and 4 in the main paper). In section A, we also have used subsets thereof. We have shown that our method is able to adequately compute the importance of the parameters using either the training data or the test data in an unsupervised manner. We also have shown that the method is able to adapt to a subset and preserve mostly the performance

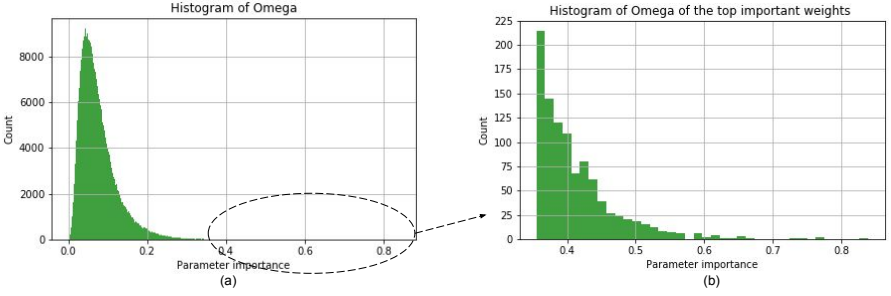


Fig. D1: (a) Histogram of the parameters importance Ω_{ij} from the last shared convolutional layer, based on fact learning, two tasks experiments with Ω computed on training data. (b) A magnified look at the histogram for the top 1000 most important values. Only few important parameters have a very high value, being crucial for the specific task, while for most parameters there is a low penalty to adapt them to other tasks.

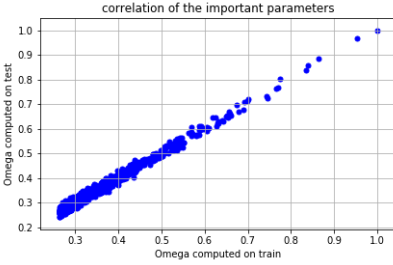


Fig. E1: Top most important parameters from Ω computed on training data. The X-axis represents the values from Ω computed on training data while the Y-axis represents the values from Ω computed on test data. Object recognition experiment Birds \rightarrow Scenes

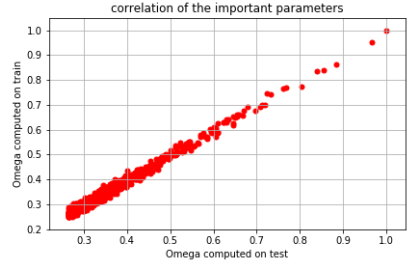


Fig. E2: Top important parameters from Ω computed on test data. The X-axis represents the values from Ω computed on test data while the Y-axis represents the values from Ω computed on training data. Based on object recognition experiment Birds \rightarrow Scenes

on that subset more than the rest of the task. Here we want to investigate the correlation or the difference between the importance assigned to the parameters computed on different sets.

First, we compare the estimated parameters importance (Ω) using the training data and the Ω computed using the test data. For that, we used a model from the object recognition experiment, namely Birds \rightarrow Scenes, the results of which are shown in Table 1 in the main paper. Figure E1 shows a scatter plot for the top 1000 most important parameters according to the Ω computed on the training data (blue). The X-axis represents the values from Ω computed on training data while the Y-axis represents the values from Ω computed on test data. Figure E2 shows a similar scatter plot for the top 1000 important parameters according to the Ω computed on the test data (red). Here, the X-axis represents the values from Ω computed on test data while the Y-axis represents the values from Ω computed on training data. A plot where the points are closely lying around

a straight line indicates that the parameters from the two Ω s have similar importance values. A plot where the points are spread further from such a line and scattered among the plotted area indicates a lower correlation between the Ω s. It can be seen how similar are the importance values computed on test data to those computed on training data where they form a tight grouping of points around a straight line where the values would be identical. This demonstrates our method’s ability to correctly identify the important parameters in an unsupervised manner, regardless of what set is used for that purpose as long as it covers the different classes or concepts of the task at hand.

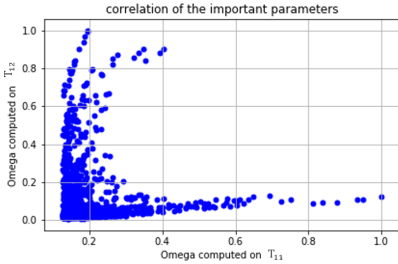


Fig.E3: Top most important parameters from Ω computed on T_{11} . The X-axis represents the values from Ω computed on T_{11} while the Y-axis represents the values from Ω computed on T_{12} , under the fact learning setting, two tasks experiments. Importance shown for the last convolutional layer.

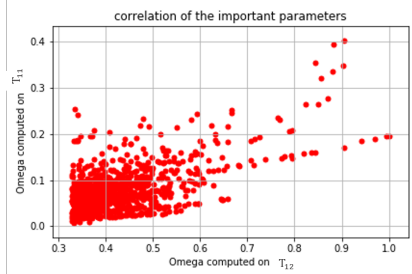


Fig.E4: Top most important parameters from Ω computed on T_{12} . The X-axis represents the values from Ω computed on T_{12} while the Y-axis represents the values from Ω computed on T_{11} , under the fact learning setting, two tasks experiments A1. Importance shown for the last convolutional layer.

How about using different subsets that cover a partial set of classes or concepts from a task? In A we have conducted an experiment under the fact learning setting where we split the data from the first task T_1 into two disjoint groups of facts and showed that computing the importance on one subset results in a better preservation in performance compared with the other subset that was not used for computing the importance – see Table A1 and Figure A1. This suggests that the importance of the parameters differs while using different subsets. To further investigate this claim, we plotted the values of Ω for the 1000 top most important parameters estimated on the T_{11} (in blue) subset of the training data from the first task T_1 along with the same parameters but with their importance computed using the other subset T_{12} . Figures E3 and E4 show this for the last convolutional layer.

This suggests that the method identifies the important parameters needed for each subset and when those parameters are shared the parameters importance is correlated between the two subsets while when those are different, different parameters receive different importance values based on the used subset.

F Visualizing the learned embedding on the adaptation experiment

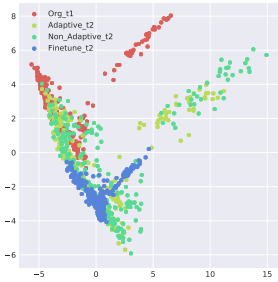


Fig.F1: Projections onto a 2D embedding, after training the second task

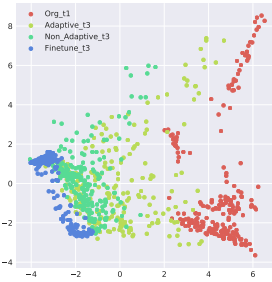


Fig.F2: Projections onto a 2D embedding, after training the third task

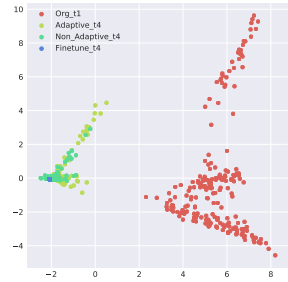


Fig.F3: Projections onto a 2D embedding, after training the fourth task

Finally, in the main paper (section 5.2, Adaptation test paragraph), we showed that our method tries to preserve the performance on a specific subset in case it encounters this subset frequently at test time along a learning sequence. This has been done by picking a subset from the first task in the 4 tasks fact learning sequence. This subset was mainly composed of sports facts. We showed that our method reduces the forgetting on this subset the most among the competitors that do not have this specialization capabilities (Figure 7 in the main paper). We were eager to know what happens in the learned embedding space, i.e. how the projections of the samples that belong to this subset change along the sequence compared to how they were right after training the first task. For that purpose, we extract a **2D** projection of the learned embedding after each task in the sequence. This was done for our method (MAS) when adapting to sport subset (**Adaptive**) and our method (MAS) when preserving the performance on all facts of the first task (**Non Adaptive**). We also show the projections of the points in the embedding learned by the finetuning baseline (**finetune**, where no regularizer is used). To have a point of reference, we also show the projections of the originally learned representation after the first task (**org**). Figure F1 shows the projections from the different variants after learning the second task compared to the original projections. It can be seen that the Adaptive and Non Adaptive variants of our method try to preserve the projections from this subset. The adaptive projections are closer to the original one, if we look closely, while Finetuning projections starts drifting away from where they were. After the third task, as shown in figure F2, the Adaptive projections are closer to the original ones than the Non Adaptive that considers this subset as part of the full task being preserved and tries to prevent forgetting them as well. Finetuning started destroying the learned topology of this subset and lies further apart. However, when it comes to the fourth task, we see that it is a quite challenging and hard task. The forgetting appears more severe than before and preservation of the projections become even harder. Nevertheless, the Adaptive MAS and Non Adaptive MAS still preserve the topology of the learned projections. The Adaptive projections lie closer and look more similar to the originals than the Non Adaptive MAS. Finetune, on the other hand, forgets completely about this subset and all the samples get projected in one point where it becomes quite hard to recognize their corresponding facts.

References

1. Aljundi, R., Chakravarty, P., Tuytelaars, T.: Expert gate: Lifelong learning with a network of experts. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
2. Andrychowicz, M., Denil, M., Gómez, S., Hoffman, M.W., Pfau, D., Schaul, T., de Freitas, N.: Learning to learn by gradient descent by gradient descent. In: Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R. (eds.) *Advances in Neural Information Processing Systems* 29, pp. 3981–3989. Curran Associates, Inc. (2016)
3. de Campos, T.E., Babu, B.R., Varma, M.: Character recognition in natural images. In: *Proceedings of the International Conference on Computer Vision Theory and Applications*, Lisbon, Portugal (February 2009)
4. Elhoseiny, M., Cohen, S., Chang, W., Price, B.L., Elgammal, A.M.: Sherlock: Scalable fact learning in images. In: *AAAI*. pp. 4016–4024 (2017)
5. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. <http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html>
6. Fernando, C., Banarse, D., Blundell, C., Zwols, Y., Ha, D., Rusu, A.A., Pritzel, A., Wierstra, D.: Pathnet: Evolution channels gradient descent in super neural networks. *arXiv preprint arXiv:1701.08734* (2017)
7. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: *Proceedings of the International Conference on Machine Learning (ICML)* (2017)
8. French, R.M.: Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences* 3(4), 128–135 (1999)
9. Goodfellow, I.J., Mirza, M., Xiao, D., Courville, A., Bengio, Y.: An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211* (2013)
10. Hebb, D.: *The organization of behavior*. 1949. New York Wiley (2002)
11. Huszár, F.: Note on the quadratic penalties in elastic weight consolidation. *Proceedings of the National Academy of Sciences* (2018). <https://doi.org/10.1073/pnas.1717042115>, <http://www.pnas.org/content/early/2018/02/16/1717042115>
12. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. *arXiv preprint arXiv:1612.00796* (2016)
13. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. pp. 554–561 (2013)
14. Krizhevsky, A.: One weird trick for parallelizing convolutional neural networks. *arXiv preprint arXiv:1404.5997* (2014)
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) *Advances in Neural Information Processing Systems* 25, pp. 1097–1105. Curran Associates, Inc. (2012)
16. 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)
17. Li, Z., Hoiem, D.: Learning without forgetting. In: *European Conference on Computer Vision*. pp. 614–629. Springer (2016)
18. Lopez-Paz, D., Ranzato, M.: Gradient episodic memory for continual learning. In: *Advances in Neural Information Processing Systems*. pp. 6470–6479 (2017)

19. Maji, S., Kannala, J., Rahtu, E., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. Tech. rep. (2013)
20. McCloskey, M., Cohen, N.J.: Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of learning and motivation* **24**, 109–165 (1989)
21. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
22. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning (2011)
23. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing* (Dec 2008)
24. Pentina, A., Lampert, C.H.: Lifelong learning with non-i.i.d. tasks. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) *Advances in Neural Information Processing Systems* 28, pp. 1540–1548 (2015)
25. Pentina, A., Lampert, C.H.: Lifelong learning with non-iid tasks. In: *Advances in Neural Information Processing Systems*. pp. 1540–1548 (2015)
26. Quadrianto, N., Petterson, J., Smola, A.J.: Distribution matching for transduction. In: *Advances in Neural Information Processing Systems*. pp. 1500–1508 (2009)
27. Quattoni, A., Torralba, A.: Recognizing indoor scenes. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 413–420. IEEE (2009)
28. Rannen, A., Aljundi, R., Blaschko, M.B., Tuytelaars, T.: Encoder based lifelong learning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 1320–1328 (2017)
29. Rebuffi, S.A., Kolesnikov, A., Lampert, C.H.: icarl: Incremental classifier and representation learning. arXiv preprint arXiv:1611.07725 (2016)
30. Ring, M.B.: Child: A first step towards continual learning. *Machine Learning* **28**(1), 77–104 (1997)
31. Royer, A., Lampert, C.H.: Classifier adaptation at prediction time. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 1401–1409 (2015)
32. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* **115**(3), 211–252 (2015). <https://doi.org/10.1007/s11263-015-0816-y>
33. Rusu, A.A., Rabinowitz, N.C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., Hadsell, R.: Progressive neural networks. arXiv preprint arXiv:1606.04671 (2016)
34. Shmelkov, K., Schmid, C., Alahari, K.: Incremental learning of object detectors without catastrophic forgetting. In: *The IEEE International Conference on Computer Vision (ICCV)* (2017)
35. Silver, D.L., Yang, Q., Li, L.: Lifelong machine learning systems: Beyond learning algorithms. In: *AAAI Spring Symposium: Lifelong Machine Learning*. pp. 49–55. Citeseer (2013)
36. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
37. Thrun, S., Mitchell, T.M.: Lifelong robot learning. *Robotics and autonomous systems* **15**(1-2), 25–46 (1995)
38. Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P.: Caltech-UCSD Birds 200. Tech. Rep. CNS-TR-2010-001, California Institute of Technology (2010)
39. Zenke, F., Poole, B., Ganguli, S.: Improved multitask learning through synaptic intelligence. In: *Proceedings of the International Conference on Machine Learning (ICML)* (2017)