GENERATIVE NEGATIVE REPLAY FOR CONTINUAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning continually is a key aspect of intelligence and a necessary ability to solve many real-world problems. One of the most effective strategies to control catastrophic forgetting, the Achilles' heel of continual learning, is storing part of the old data and replay them interleaved with new experiences (also known as the replay approach). Generative replay, that is using generative models to provide replay patterns on demand, is particularly intriguing, however, it was shown to be effective mainly under simplified assumptions, such as simple scenarios and low-dimensional benchmarks. In this paper, we show that, while the generated data are usually not able to improve the classification accuracy for the old classes, they can be effective as negative examples (or antagonists) to learn the new classes, especially when the learning experiences are small and contain examples of just one or few classes. The proposed approach is validated on complex class-incremental and data-incremental continual learning scenarios (CORe50 and ImageNet-1000) composed of high-dimensional data and a large number of training experiences: a setup where existing generative replay approaches usually fail.

1 INTRODUCTION

The majority of neural networks training approaches assume that is feasible to build a set of independent and identically distributed (i.i.d.) samples to train the model. This assumption is in contrast with biological learning since intelligent beings observe the world as an ordered sequence of highly correlated data. When state-of-the-art deep neural networks are trained continually, and the whole data cannot be accessed at once, the model suffers from the catastrophic forgetting problem (McCloskey & Cohen, 1989), and the knowledge about old data (old experiences) tend to be overwritten by new examples.

Storing part of past data and replaying them interleaved with new data proved to be an effective approach to mitigate forgetting (see Hayes et al. (2021) for a comprehensive survey). However, in some applications, the storage overhead together with privacy issues make replay techniques unfeasible. Therefore, *generative replay* has been recently explored, where a generative model is trained to produce data from past experiences (see Lesort et al. (2018) and Shin et al. (2017)). Besides solving the replay memory issue, generative replay can theoretically be capable of generating more general and novel examples not included in past experiences, thus potentially overcoming replay methods. Unfortunately, generative replay introduces much complexity due to the need for an interleaved incremental training of both a classifier and a generator. Moreover, generative models are usually complex and unstable to train, especially in incremental scenarios. Several researchers have shown that generative replay fails in complex CL scenarios with high-dimensional data (see Aljundi et al. (2019); Lesort et al. (2018) and van de Ven et al. (2020)) mainly due to the inaccuracies in the data generation that progressively grows across the experiences if a single generator is incrementally updated (see related works in section 5 for more details). The photocopy example helps to understand why. Let us consider a high-quality photocopy machine: when a picture is initially copied the output looks very similar to the original, but if the process is repeated several times by using as input the output of the previous step, some artifacts will soon appear and, after many iterations, the result will be highly compromised. Hence, even if some state-of-the-art models have been proved to be effective in generating also high dimensional data (Huang et al. (2018) Karras et al. (2019)), the continual training of such generators remains a challenging problem.

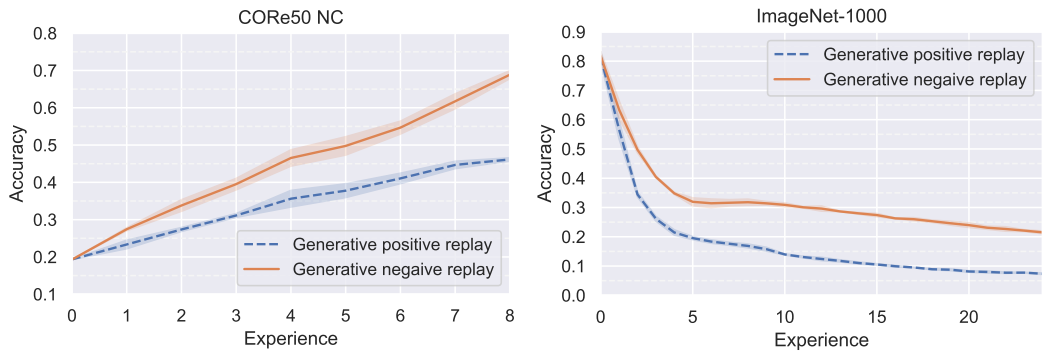


Figure 1: The proposed generative negative replay is compared with classical generative replay on two complex class incremental CL benchmarks (details in section 4.2). In both the benchmarks, using the same classifier, generator, and training procedure, negative replay performs significantly better.

Although generative models are a hot research topics and we can expect improved methods in the future, as of today we must deal with imperfect generated data and try to exploit them at best when a classifier is incrementally trained. The proposed approach, denoted as *Generative Negative Replay*, does not attempt to improve the knowledge of old classes using the generated data because it assumes that the data quality is not enough for this purpose. Nevertheless, it makes use of generated data as negative examples to better learn the classes of current experience, especially when the number of classes per experience is small and we incur in the "learning in isolation" problem.

We experimentally demonstrate, on complex benchmarks such as CORe50 and ImageNet-1000, where (positive) generative replay fails, that negative replay is effective to contrast the learning in isolation problem, allowing to train a classifier incrementally across a high number of experiences (see Figure 1). We also investigate the impact of data quality on negative replay with an ablation study (section 4.4) where negative examples are sampled from original past patterns (upper bound) and randomly generated.

2 Problem Formulation

A continual learning (CL) problem consists of a number N_E of experiences, each containing a subset of data that is only accessible during the corresponding experience:

$$\text{CL} = \{e_1, e_2, ..., e_{N_E}\}, \tag{1}$$

each of them is composed of several data points and the corresponding labels:

$$e_k = (\mathcal{X}_k, \mathcal{Y}_k), \quad \mathcal{X}_k = \{x_1^k, x_2^k, ..., x_{N_k}^k\}, \ \mathcal{Y}_k = \{y_1^k, y_2^k, ..., y_{N_k}^k\} \tag{2}$$

where x_i^k and y_i^k are the data points and the associated labels contained in the k-th experience and N_k is the number of samples in the k-th experience.

Let $\mathcal{D} = (\mathcal{X}, \mathcal{Y})$ be the entire dataset, then $\mathcal{X} = \bigcup_{i=1}^{N_E} \mathcal{X}_i$ and $\mathcal{Y} = \bigcup_{i=1}^{N_E} \mathcal{Y}_i$.

We can define three different scenarios for supervised continual learning (Maltoni & Lomonaco, 2019; van de Ven & Tolias, 2018): New Instances (NI), New Classes (NC), and New Instances and Classes (NIC). In NI (also known as domain incremental) all the classes are introduced in the first experience, and only new examples of the same classes are included in the following experiences. In NC (also known as class incremental) each experience contains only examples from classes never seen before. NIC is the combination of NI and NC, so each experience can be composed of new examples of already seen classes and/or examples from new classes. A formal definition of the above scenarios can be found in Appendix A.

Given the above definitions, our goal is to fit a function f, parametrized by Θ, to the sequence of experiences. A naive approach is finding the best parameters Θ^* that minimizes:

$$\Theta^* = \arg\min_{\Theta} \mathcal{L}(f_\Theta(\mathcal{X}_i), \mathcal{Y}_i) \text{ for } i = \{1, ..., N_E\}, \tag{3}$$

where $\mathcal{L}(\cdot)$ is a loss function (e.g. cross entropy loss).

As first pointed out by McCloskey & Cohen (1989), this simple approach is prone to catastrophic forgetting, thus the model f_Θ is not able to learn the experiences $\{e_1, e_2, ..., e_{N_E}\}$ sequentially.

2.1 CONTINUAL LEARNING WITH GENERATIVE REPLAY

Generative replay requires to train simultaneously and incrementally a classifier and a generative model (Shin et al., 2017; Wu et al., 2018; Thandiackal et al., 2021) The generative model g, parametrized by Ω provides surrogate data similar to the past experiences' data. In the case of a conditional generative model (in which we can control the class of the generated data), the optimal parameters of the classifier can be derived using a replay memory as follows:

$$\Theta^* = \arg\min_\Theta \mathcal{L}(f_\Theta(\mathcal{X}_i \cup \mathcal{M}_i^x, \mathcal{Y}_i \cup \mathcal{M}_i^y) \text{ for } i = \{1, ..., N_E\}, \tag{4}$$

where $\mathcal{M}_i^x$ and $\mathcal{M}_i^y$ are the datapoints and labels contained in the replay memory during the training on experience i when the replay memory is populated as:

$$\mathcal{M}_i^x \leftarrow g_\Omega(z_j|c_j); \;\; \mathcal{M}_i^y \leftarrow c_j; \;\; c_j \in \bigcup_{k=1}^{i-1} \mathcal{Y}_k, \;\; j = \{1, ..., R\}, \tag{5}$$

where R is the number of generated replay patterns (size of memory), z is a latent random input vector to the generative model, c is a label sampled from the set of labels encountered in the past experiences, and "$\leftarrow$" indicates the insertion of an element in the memory.

The same generated data fed to the classifier can be used to control forgetting in the generative model as well. Suppose we have a conditional generative model composed of an encoder q_γ parametrized by γ and a decoder p_ξ parametrized by ξ, such that $g_\Omega = p_\xi \circ q_\gamma, \; \Omega = (\gamma, \xi)$. The optimal parameters of the generative model can be obtained requiring that the generated data are similar (L2 loss) to the original ones:

$$\gamma^*, \xi^* = \arg\min_{\gamma,\xi} \|p_\xi(q_\gamma(\mathcal{X}_i \cup \mathcal{M}_i^x)|\mathcal{Y}_i \cup \mathcal{M}_i^y) - \mathcal{X}_i \cup \mathcal{M}_i^x\|_2^2 \;\; \text{for } i = \{1, ..., N_E\}, \tag{6}$$

where $q_\gamma(\mathcal{X}_i)$ is forced to follow a target distribution, typically $\mathcal{N}(0, 1)$.

3 GENERATIVE NEGATIVE REPLAY

As discussed before, generative replay is an appealing strategy for continual learning, but, to exploit it in complex scenarios with many experiences, we need to overcome the data degradation issue. Since this problem is not easily addressable on the generator side, we propose to circumvent it by changing the way the classifier makes use of generated data.

Let us suppose the classifier f_Θ can be divided into a feature extractor f_ϕ, parametrized by ϕ and a classification head c_ψ parametrized by ψ, so that $f_\Theta = c_\psi \circ f_\phi, \; \Theta = (\phi, \psi)$. The parameters ψ of the classification head can be divided into C groups, where C is the number of classes. The groups, denoted as (ψ^1, ψ^2, ..., ψ^C) represents the parameters associated to the connections between the features extracted by f_ϕ and the output neuron of the corresponding class.

For simplicity, let us assume that the feature extraction weights ϕ are frozen (after an initial pre-training) and, across the experiences, we only learn the classification head weights ψ. As explained in Section section 3.3, this assumption is not necessary and our experiments were carried out by learning both ϕ and ψ.

3.1 LEARNING CLASSES IN ISOLATION

Learning in isolation is one of the main causes of catastrophic forgetting, especially in the NC or NIC scenarios where only a limited number of classes are present in a single experience, and the parameters of the classification head are learned without negative examples that counteract the "greediness" of the optimization. As an example, let us consider an NC scenario where only one class is present in each experience. Suppose that c is the only class in the experience k, then the

best way to optimize the model is to change the parameters ψ^c to maximize the output of the output neuron c for every input and change the rest of $\psi^j,\ j \neq c$ to minimize the output for remaining classes. This still holds if in the experience are present only a few classes, since the model is only optimized to discriminate between the present classes and has no interest in maintaining the past acquired knowledge.

3.2 POSITIVE AND NEGATIVE REPLAY

Replay can be used to counteract the learning in isolation problem, however, when the replay data comes from a generative model, the data quality degradation has a negative impact on the classifier training. The aforementioned problem is typical of the standard generative replay approach (hereafter denoted as *generative positive replay*), where replay data is used by the classifier in the same manner of the current experience's data, and therefore the classification head's weights associated to the replay classes are optimized based on the replay data.

On the contrary, in the proposed *generative negative replay* approach the replay patterns are used to counteract the detrimental effects of the training in isolation but they are not used to modify the parameters ψ associated with the replay classes. The key idea (validated experimentally) is that the generated patterns are valid antagonists to mitigate the learning in isolation problem, but their quality is not enough to improve the knowledge of classes originally learned on real data. It is well known that one class learning approaches are in general less effective than discriminative learning because the presence of both positive and negative examples allows to better characterize the classification boundaries (Hempstalk & Frank, 2008). Therefore, the proposed approach exploits generated data to constrain the classification boundary and to avoid that the real data in the current experience pull it too much in their direction.

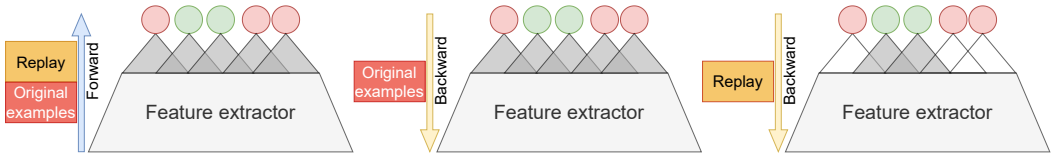
3.3 TRAINING A CLASSIFIER WITH GENERATIVE NEGATIVE REPLAY

The idea of generative negative replay is quite general can be used in conjunction with different continual learning classification approaches and scenarios (NI, NC, NIC). To avoid replay data (i.e. negative examples) alter the knowledge of the already learned classes, the gradient accumulation can be selectively blocked during the backward pass. The general idea is illustrated in Figure 2. While the original examples ($\mathcal{X}_i$) normally flows forward and backward throughout the model, the replay examples ($\mathcal{M}_i^x$) are passed forward, but, before the backward pass, the loss tensor is masked at the class level by resetting the gradient components corresponding to the classes in $\mathcal{M}_i^y$.

Hereafter, we provide a baseline implementation embedded in AR1 algorithm (Maltoni & Lomonaco, 2019), whose update mechanism for the classification head weights allows very simple and efficient integration of negative replay. AR1 is a flexible continual learning approach that can achieve state-of-the-art accuracy on complex CL benchmarks. In Appendix D, AR1 is shown to outperform several recent CL algorithms of the difficult ImageNet-1000 benchmark proposed by Masana et al. (2020).

AR1 uses different mechanisms to learn the classification head and the feature extractor weights. The feature extraction weights ϕ are protected against forgetting: i) through the Synaptic Intelligence regularization technique (Zenke et al., 2017) or ii) using a replay memory with a small learning rate (denoted as AR1free in Pellegrini et al. (2020)). The classification head weights ψ are managed by CWR. CWR is a simple method aimed at addressing the score bias problem produced by imbalance learning during continual learning (Belouadah et al., 2020). CWR (Maltoni & Lomonaco, 2019) maintains a copy of the weights of the classification head of the previous experience (ψ') and at the start of each experience the classification head is reset and only weights of classes of the current experience are loaded from ψ'. At the end of the experience, a weight consolidation phase takes place, where the weights ψ learned in the current experience are consolidated with the weights ψ'. This is the point where positive and negative replay behaves differently.

In particular, during the consolidation phase, for each parameter group ψ^c associated to a class c belonging to the current experience ($c \in \mathcal{Y}_k \cup \mathcal{M}_k^y$), the mean of all the parameter group $\mu(\psi^c)$ is calculated, and subtracted to all the parameters in the group, in order to force zero mean: $\psi^c = \psi^c - \mu(\psi^c)$. This prevents class bias problems due to the different magnitudes of the weights. Then, there are three possibilities, based on c:



(a) (b) (c)

Figure 2: Graphical representation of the negative replay idea. Green output neurons represent the classes present in the current experience, while red output neurons represent the replay classes. During forward (a) both the replay data and the original data from the current experience flow through the network. During backward, the original data flow through all the neurons of the classification head (b), while the replay data contribution is masked and data only flows through the neurons of classes belonging to the current experience (c).

1. c is a new class never seen before ($c \in \mathcal{Y}_k \wedge c \notin \bigcup_{i=1}^{k-1} \mathcal{Y}_i$): in this case ψ^c is maintained as is.
2. c is a class seen before ($c \in \mathcal{Y}_k \wedge c \in \bigcup_{i=1}^{k-1} \mathcal{Y}_i$): the consolidation step is applied, so $\psi^c = \frac{\psi'^c \cdot w_{past_c} + \psi^c}{w_{past_c}+1}$ where w_{past_c} is a parameter that balances the contribution of the past w.r.t. the present, calculated as follows:

$$w_{past_c} = \sqrt{\frac{past_c}{current_c}}, \tag{7}$$

   where $past_c$ is the amount of pattern of class c encountered in past experiences, while $current_c$ is the number of patterns of class c present in the current experience.
3. c is not in the current experience but is a replay pattern ($c \notin \mathcal{Y}_k \wedge c \in \mathcal{M}_k^y$):
   - in case of positive replay apply consolidation (step 2).
   - in case of negative replay ψ^c is substituted with ψ'^c (no contribution to the parameters ψ^c from replay patterns).

The pseudo-code of the above weigh consolidation algorithms is reported in Appendix B. It is worth noting, that in the proposed embedding of negative replay in AR1, the replay pattern can alter the feature extraction weights since CWR weight consolidation only "protects" the classification head.

4 Experiments and Results

In this section, we describe the experimental setup used to validate the proposed negative replay. We focus on difficult continual learning scenarios, where data is high-dimensional, non-i.i.d. and the number of experiences is very large. Negative replay is compared with alternative strategies (e.g. positive replay) and the role of quality of generated data is investigated by also using, as negative replay patterns, real and random data.

4.1 Experimental setup

Datasets We performed our experiments on the CORe50 dataset (Lomonaco & Maltoni, 2017) and ImageNet-1000 dataset (Deng et al., 2009). CORe50 dataset was specifically collected for continual learning (NI, NC, and NIC scenarios) and is composed of small video sessions (about 300 frames) of 50 objects taken from the point of view of a person that handles them in the hand. Every class has 11 video sessions (a total of about 3,300 images) with different backgrounds and illuminations. Eight video sessions for each class are used for training, and 3 for testing. Images have size 128×128 pixels. ImageNet is composed of 1,000 classes with about 1,000 patterns per class for training and 100,000 images for testing. All images are resized to 224×224 pixels.

Classifier architecture AR1 algorithm was used with Synaptic Intelligence (SI) regularization when trained without replay, and without protection on the feature extraction weights (AR1free) in case of positive and negative replay. In the experiments with CORe50 dataset we follow Maltoni

& Lomonaco (2019) and Lomonaco et al. (2020) by employing a MobileNetV1 network (Howard et al., 2017). As in Pellegrini et al. (2020) and van de Ven et al. (2020), we opted for latent replay, that is replaying latent activations instead of input data. For CORe50 experiments, as in Pellegrini et al. (2020), we used the `conv5_4` layer as latent replay layer, and the classifier was pretrained on ImageNet-1000. We also substituted all the batch normalization layers of the network with batch renormalization (Ioffe, 2017). For ImageNet-1000 we use a ResNet-18 (He et al., 2016) architecture. To maintain compatibility with the experiments on CORe50, even on ImageNet-1000 we use latent replay, setting the replay layer on the fourth residual block of the network (after `conv4_x` using He et al. (2016) nomenclature).

Generative model architecture For the choice of a generative model we initially focused on two state-of-the-art approaches whose implementations are open source (van de Ven et al., 2020; Shin et al., 2017; Ayub & Wagner, 2021). However, since they were designed to work in simpler settings (with a lower data dimensionality and a smaller number of experiences), we were not able to port and scale them to our complex setups. Therefore, we implemented a generative model by trying to combine the most promising techniques and ideas from different sources and control its overall memory/computation complexity. In particular, taking inspiration from van de Ven et al. (2020) we use a Variational Autoencoder (VAE) Kingma & Welling (2014) model, but unlike van de Ven et al. (2020) we opted for a conditional VAE (cVAE) configuration Sohn et al. (2015). Moreover, we partially blend the generator (encoder) with the classifier model: both the networks share the same feature extractor f_ϕ. Finally, instead of generating raw data, we generate activations at an intermediate "latent" level as suggested by van de Ven et al. (2020). A detailed discussion on the architecture of the generator is provided in Appendix C, including a pseudo-code that highlights the details of the interleaved training of the generator and the classifier.

4.2 Experiments on the NC scenario

The first round of experiments has been performed on the NC scenario using CORe50 and ImageNet-1000. For CORe50 the benchmark is composed of 9 experiences: the first one contains 10 classes while the following contains five classes each. We used a replay memory of 1,500 patterns, and (for generative replay) we inserted in each minibatch, of size 128, 14 replay patterns, and 114 patterns from the current experience. We train both the classifier and the generator for 4 epochs for each experience. Hyper-parameters of the classifier and generator are reported in Appendix E and Appendix F respectively.

For ImageNet-1000 the benchmark follows the one proposed by Masana et al. (2020): the dataset is divided into 25 experiences of 40 classes each. We used a replay memory of 20,000 patterns, and (for generative replay) we inserted in each minibatch, of size 128, 36 replay patterns, and 92 patterns from the current experience. We did not expect negative replay to perform well in this setup, because each experience already contains 40 classes and, therefore, the learning-in-isolation problem is here marginal. Nevertheless, we were interested in understanding if, in this setup, negative replay hurts the learning process or still provides minor benefits.

The results are shown in Figure 3 and Table 1. In CORe50 the baseline with no replay (using the AR1 algorithm) reaches a final accuracy of about 60% while using replay raises the accuracy to more than 70% (Positive Replay Original Data - PR-OD). These were expected to be the lower and upper bounds of this experiment, respectively. However, because of the data degradation problem, performing positive replay with generated data (Positive Replay Generated Data - PR-GD) performed significantly worse than the case with no replay. Using replay in a negative manner with generated data, as proposed in this work (NR-GD), only slightly decreases the final accuracy with respect to the upper bound PR-OD.

For ImageNet-1000, due to the complexity of the experiment and the fact that the network is fully trained only during the first experience (blocked after `conv4_x` in the following experiences) the final accuracy are quite similar for all the methods (except PR-GD that performed far worse). However, in the first 10 experiences some differences can be appreciated: see the insect view in Figure 3-right. The impact of the generated data quality on negative replay is more evident in Table 1: using negative replay with generated data (in this case highly degraded) improve the average accuracy (calculated as the mean of the accuracy after each experience) of more than 24 points and the final accuracy of more than 10 points w.r.t. using replay data in a positive manner. Furthermore, even if in this scenario

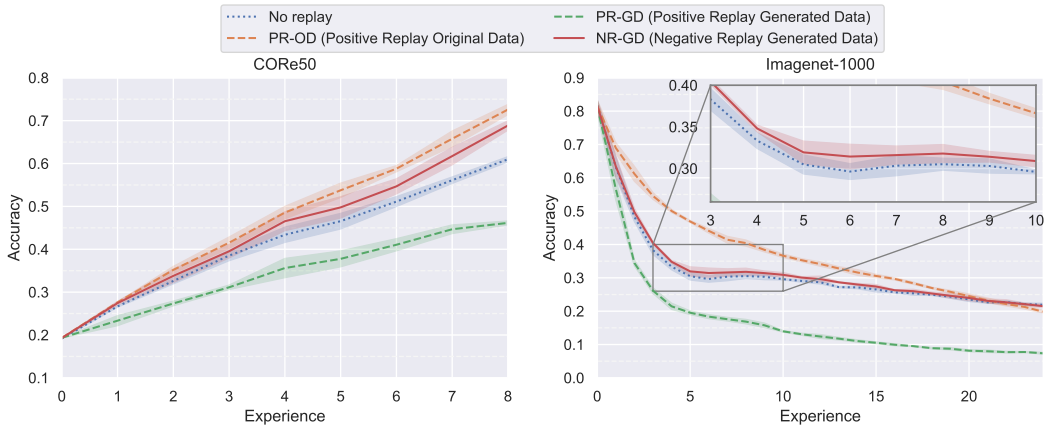


Figure 3: Overall accuracy on CORe50 NC scenario, using the whole test set (even at intermediate experiences) as defined in the CORe50 protocol (Lomonaco & Maltoni, 2017) (left), and on ImageNet-1000 using a growing test set as defined by Masana et al. (2020) (right). Every experiment is averaged over 3 runs using different seeds and class order. The standard deviation is reported in light colors. Better viewed on a computer monitor.

Dataset	Method	Avg. Accuracy
CORe50	No Replay	41.68 ± 0.62
	PR-OG (upper bound)	47.02 ± 0.45
	PR-GD	34.05 ± 0.29
	NR-GD	$\mathbf{44.63 \pm 0.77}$
ImageNet-1000	No Replay	31.91 ± 0.17
	PR-OG (upper bound)	38.02 ± 0.08
	PR-GD	18.29 ± 0.07
	NR-GD	$\mathbf{32.74 \pm 0.17}$

Table 1: Average accuracy on all the experiences for the CORe50 and ImageNet-1000 NC scenarios.

the advantage of negative generative replay is little with respect to the no replay case, we note that negative replay is not hurting the training process even in scenarios where learning in isolation is not an issue.

4.3 Experiments on the NIC benchmark

CORe50 NIC-391 protocol is composed of 391 learning experiences, each containing examples of a single class (300 frames of a short video). This scenario is particularly challenging and prone to learn-in isolation issues, hence we may expect the role of replay to be more important here. In this scenario, we used a replay memory of only 300 patterns. The minibatch size is 128, and when generative replay is employed, we generate 64 patterns for every mini-batch (plus 64 from the current experience). Hyper-parameters of the classifier and generator are reported in Appendix E and Appendix F respectively.

The results are shown in Figure 4 and they are quite in line with the previous experiment, but here the accuracy gaps grow and the benefit of replay is more evident. The proposed negative replay with generated data (NR-GD) performs quite well, about 10 points better than with no replay and just less than 5 points worse than positive replay with real data, the upper bound. Using generated data in a positive manner (PR-GD) is here even worse than in the NC case, because the data degradation is amplified with so many iterations: PR-GD is losing 30 points w.r.t. not using replay at all, and performs about 40 points worse than using the same replay data with the proposed generative negative replay approach.

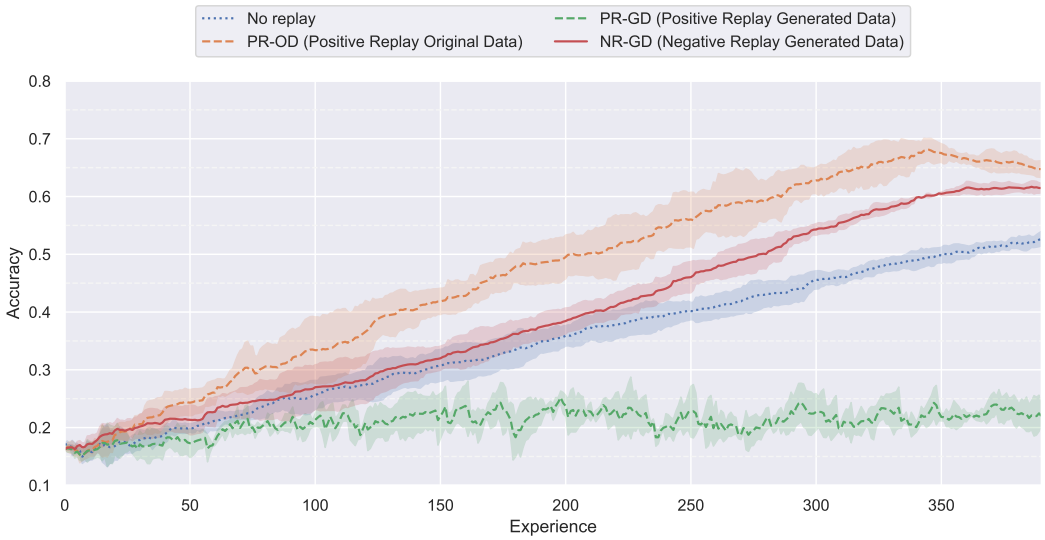


Figure 4: Overall accuracy on CORe50 NIC391 scenario, using the whole test set as defined in the CORe50 protocol (Lomonaco & Maltoni, 2017). Every experiment is averaged over 3 runs using different seeds and class order. The standard deviation is reported in light colors. better viewed on a computer monitor.

4.4 Ablation Study

The effect of generated data quality on negative replay is investigated by performing two further experiments: NR-OD uses original data (max. quality) for negative replay, while NR-RD uses randomly generated replay data, obtained by uniform random sampling in the latent replay layer and assigning to each data point a random class label. Since in our experiments we replay hidden features, in order to produce reasonable replay data we first calculated the range of latent activations on a sample dataset, and then we set our random generator to produce values in the range: 0 (since we use ReLU activation functions) - $90th$ percentile of the real activation values. We used CORe50 NC and CORe50 NIC in these experiments.

The results are reported in Table 2. Surprisingly, even with random replay data (that we assume to be the worst degradation possible), negative replay is still able to perform better than no replay. Furthermore, the difference between original and generated data is minimal, thus proving that negative replay is tolerant in terms of data quality. Note that in both the experiment using random data with negative replay performs way better than using generated data in a classical (positive) manner (PR-GD in previous figures). Comparisons in all the benchmarks of all the experiments (positive and negative replay with original, generated, and random data) are reported in Appendix G.

Benchmark	Method	Final accuracy
NC	No Replay	60.99 ± 0.49
	NR-OG	68.60 ± 1.38
	NR-GD	68.87 ± 0.88
	NR-RD	64.05 ± 0.71
NIC	No Replay	52.71 ± 1.02
	NR-OG	67.93 ± 0.31
	NR-GD	61.46 ± 0.67
	NR-RD	58.85 ± 0.58

Table 2: Final accuracy on CORe50 NC and NIC using original (NR-OD), generated (NR-GD), and random (NR-RD) data with negative replay. The results with no replay are reported as references. Every experiment is averaged over 3 runs using different seeds and class order.

5 RELATED WORKS

Generative replay for continual learning was first introduced by Shin et al. (2017) who proposed Deep Generative Replay (DGR). In this work, a generative adversarial network (GAN) (Goodfellow et al., 2014) is used as a generative model, showing promising results but only on simple datasets. The method used a teacher-student framework, where the generative model resulting from the previous experience is used to train the current generative model. Wu et al. (2018) noted that generative replay almost completely shifts the problem of continual learning from the classifier to the generator. They proposed a GAN-based distillation approach to address the issue. However, these approaches lead to rapid degradation of the quality of the generated images for old tasks. To overcome this issue Ostapenko et al. (2019) proposed Dynamic Generative Memory (DGM), where a GAN architecture is used both to generate data and classify it. Moreover, the generative model used a combination of binary masks and network expansion, to maintain a fixed number of free parameters for every experience. All these works use GANs as generative models, but GANs are usually slow and complex to train, even in non-incremental scenarios. Kemker & Kanan (2018) proposed FearNet, a brain-inspired model that employs dual-memory storage (short and long term) with a transfer phase of information between the two memories in a consolidation phase inspired to mammalian sleep. Recently, Ayub & Wagner (2021) proposed a generative replay framework based on autoencoders and neural style transfer (Gatys et al., 2016) that showed interesting results even with high-dimensional data. However, that approach requires maintaining a generative model for every experience encountered so far, making it not scalable to long incremental sequences. Instead of generating raw images, van de Ven et al. (2020) proposed to generate internal features of the classifier through a Variational Autoencoder (VAE) (Kingma & Welling, 2014). This approach shares some similarities on how memory works in the human brain and combined with some neuroscience-inspired modifications showed significant results in continual learning scenarios with dozens of experiences. However, even this approach was not tested on high dimensional data and in a scenario with hundreds of experiences. The crucial role of the dimensionality and the complexity of data on the quality of generation is evident in Zhai et al. (2019), where a simple generative model can effectively generate faithful results if trained continually on low-dimensions data (e.g. the MNIST dataset LeCun (1998)), but it fails to generate acceptable results if the dimensionality of the data increases (e.g. using the Flowers dataset (Nilsback & Zisserman, 2006)). Another example of this behavior is introduced by Mundt et al. (2019), where the Flower dataset, with image dimensionality of 256×256 pixels, is almost impossible to faithfully reconstruct if learned continually, even if the number of incremental experiences is low.

6 CONCLUSION

In this paper, we addressed the problem of continual learning with generative replay, focusing on the obstacles of generative replay in complex scenarios. Our experience confirms that incrementally training a generator over a long number of experiences with high dimensional data is a very challenging problem and remains an open issue. Therefore, instead of trying to design a better generative model, we focused on classifier training. We found that even inaccurate replay data can be useful to contrast the learning in isolation problem, especially in scenarios where only a limited number of classes is present in each experience. We called this approach negative replay since the replay data is used as negative examples when the model is trained with data from the current experience. We validated negative replay using complex continual learning scenarios, with high dimensional data and hundreds of incremental experiences. The results show that using negative replay largely improves classification performances w.r.t. not using replay at all or using the generated data in a traditional fashion. We also investigated the impact of generated data quality, by considering the two extremes of using original data and random data for negative replay, and, surprisingly, we found that negative replay is effective even using random replay data. Since negative replay can be easily applied to other continual learning strategies (besides AR1), we believe that many other CL approaches may benefit from our proposal, especially when complex scenarios are employed. Finally, is worth noting that dealing with imprecise replay data can be viewed as a biological feature since human's memory is far from being accurate, but is thought to be essential to consolidate learning (van de Ven et al., 2020), therefore investigating the role of negative replay-like mechanisms in biological learning could be an interesting research direction for computer scientists and neuroscientists.

REPRODUCIBILITY STATEMENT

The source code of the project, alongside configuration files for the reproducibility of experiments, is included in the additional materials. A public version of the code, based on the Avalanche framework (Lomonaco et al., 2021) will be released upon publication. The datasets used are publicly available and can be downloaded from the respective official websites[1,2]. Both the datasets are preprocessed normalizing them using statistics derived from ImageNet-1000. Images from ImageNet-1000 were randomly cropped and resized to the final dimension of 224×224, then horizontally flipped with a probability of 50%.

REFERENCES

Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. In H Wallach, H Larochelle, A Beygelzimer, F dÁlché-Buc, E Fox, and R Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 11816–11825. Curran Associates, Inc., 2019.

Ali Ayub and Alan R. Wagner. EEC: Learning to encode and regenerate images for continual learning. In *International Conference on Learning Representations*, 2021.

Eden Belouadah and Adrian Popescu. Il2m: Class incremental learning with dual memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 583–592, 2019.

Eden Belouadah, Adrian Popescu, and Ioannis Kanellos. A comprehensive study of class incremental learning algorithms for visual tasks. *Neural Networks*, 2020.

Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 233–248, 2018.

Arslan Chaudhry, Puneet K. Dokania, Thalaiyasingam Ajanthan, and Philip H. S. Torr. Riemannian Walk for Incremental Learning: Understanding Forgetting and Intransigence. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 532–547, 2018.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Prithviraj Dhar, Rajat Vikram Singh, Kuan-Chuan Peng, Ziyan Wu, and Rama Chellappa. Learning without Memorizing. In *CVPR*, 2019.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2414–2423, 2016. doi: 10.1109/CVPR.2016.265.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Tyler L Hayes, Giri P Krishnan, Maxim Bazhenov, Hava T Siegelmann, Terrence J Sejnowski, and Christopher Kanan. Replay in deep learning: Current approaches and missing biological elements. *arXiv preprint arXiv:2104.04132*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kathryn Hempstalk and Eibe Frank. Discriminating against new classes: One-class versus multi-class classification. In Wayne Wobcke and Mengjie Zhang (eds.), *AI 2008: Advances in Artificial Intelligence*, pp. 325–336, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. ISBN 978-3-540-89378-3.

[1] `https://vlomonaco.github.io/core50`

[2] `https://www.image-net.org`

I. Higgins, L. Matthey, A. Pal, Christopher P. Burgess, Xavier Glorot, M. Botvinick, S. Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *Proceedings of the 5th International Conference on Learning Representations*, 2017.

Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 831–839, 2019.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Huaibo Huang, Zhihang Li, Ran He, Zhenan Sun, and Tieniu Tan. Introvae: Introspective variational autoencoders for photographic image synthesis. *arXiv preprint arXiv:1807.06358*, 2018.

Sergey Ioffe. Batch renormalization: Towards reducing minibatch dependence in batch-normalized models. *arXiv preprint arXiv:1702.03275*, 2017.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Ronald Kemker and Christopher Kanan. FearNet: Brain-Inspired Model for Incremental Learning. In *ICLR*, 2018.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *PNAS*, 114(13):3521–3526, 2017.

Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Timothée Lesort, Hugo Caselles-Dupré, Michael Garcia-Ortiz, Andrei Stoian, and David Filliat. Generative Models from the perspective of Continual Learning. *Proceedings of the International Joint Conference on Neural Networks*, 2018. doi: 10.1109/IJCNN.2019.8851986.

Zhizhong Li and Derek Hoiem. Learning without Forgetting. In *European Conference on Computer Vision*, Springer, pp. 614–629, 2016.

Vincenzo Lomonaco and Davide Maltoni. CORe50: A New Dataset and Benchmark for Continuous Object Recognition. *CoRL*, 2017.

Vincenzo Lomonaco, Davide Maltoni, and Lorenzo Pellegrini. Rehearsal-Free Continual Learning over Small Non-I.I.D. Batches. *CVPR Workshop on Continual Learning for Computer Vision*, 2020.

Vincenzo Lomonaco, Lorenzo Pellegrini, Andrea Cossu, Antonio Carta, Gabriele Graffieti, Tyler L. Hayes, Matthias De Lange, Marc Masana, Jary Pomponi, Gido van de Ven, Martin Mundt, Qi She, Keiland Cooper, Jeremy Forest, Eden Belouadah, Simone Calderara, German I. Parisi, Fabio Cuzzolin, Andreas Tolias, Simone Scardapane, Luca Antiga, Subutai Amhad, Adrian Popescu, Christopher Kanan, Joost van de Weijer, Tinne Tuytelaars, Davide Bacciu, and Davide Maltoni. Avalanche: An End-to-End Library for Continual Learning. *CLVision Workshop at CVPR 2021*, 2021.

Davide Maltoni and Vincenzo Lomonaco. Continuous Learning in Single-Incremental-Task Scenarios. *Neural Networks*, 116:56–73, 2019.

Marc Masana, Xialei Liu, Bartlomiej Twardowski, Mikel Menta, Andrew D Bagdanov, and Joost van de Weijer. Class-incremental learning: survey and performance evaluation. *arXiv preprint arXiv:2010.15277*, 2020.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. Psychology of Learning and Motivation. Academic Press, 1989. doi: https://doi.org/10.1016/S0079-7421(08)60536-8.

Martin Mundt, Sagnik Majumder, Iuliia Pliushch, Yong Won Hong, and Visvanathan Ramesh. Unified Probabilistic Deep Continual Learning through Generative Replay and Open Set Recognition. *arXiv*, 2019.

M-E Nilsback and Andrew Zisserman. A visual vocabulary for flower classification. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pp. 1447–1454. IEEE, 2006.

Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*, pp. 2642–2651. PMLR, 2017.

Oleksiy Ostapenko, Mihai Puscas, Tassilo Klein, Patrick Jahnichen, and Moin Nabi. Learning to remember: A synaptic plasticity driven framework for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11321–11329, 2019.

Lorenzo Pellegrini, Gabriele Graffieti, Vincenzo Lomonaco, and Davide Maltoni. Latent replay for real-time continual learning. *IROS*, 2020.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. iCaRL: Incremental Classifier and Representation Learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual Learning with Deep Generative Replay. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 2990–2999. Curran Associates, Inc., 2017.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Kevin Thandiackal, Tiziano Portenier, Andrea Giovannini, Maria Gabrani, and Orcun Goksel. Match what matters: Generative implicit feature replay for continual learning. *arXiv preprint arXiv:2106.05350*, 2021.

Gido M van de Ven and Andreas S Tolias. Three scenarios for continual learning. In *Continual Learning Workshop NeurIPS*, 2018.

Gido M. van de Ven, Hava T. Siegelmann, and Andreas S. Tolias. Brain-inspired replay for continual learning with artificial neural networks. *Nature Communications*, 11, 2020. doi: 10.1038/s41467-020-17866-2.

Chenshen Wu, Luis Herranz, Xialei Liu, Joost van de Weijer, Bogdan Raducanu, et al. Memory replay gans: Learning to generate new categories without forgetting. *Advances in Neural Information Processing Systems*, 31:5962–5972, 2018.

Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 374–382, 2019.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual Learning Through Synaptic Intelligence. In *International Conference on Machine Learning*, pp. 3987–3995, 2017.

Mengyao Zhai, Lei Chen, Frederick Tung, Jiawei He, Megha Nawhal, and Greg Mori. Lifelong gan: Continual learning for conditional image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2759–2768, 2019.

A NI, NC, AND NIC DEFINITIONS

Given the definitions of section 2, the NI (New Instances), NC (New Classes), and NIC (New Instances and Classes) continual learning scenarios can be defined based on the labels $\mathcal{Y}_k$ contained in the experiences ($k \in \{1, ..., N_E\}$ with N_E the total number of experiences) as follows:

New instances (NI) also known as domain-incremental learning, where all the labels are known from the first experience, and the successive experiences, only new instances of the same classes are included. Formally, we could define the NI scenario as:

$$\mathcal{Y}_1 \cap \mathcal{Y}_k = \mathcal{Y}_k \quad \text{for } k = \{1, ..., N_E\}, \tag{8}$$

meaning that every possible label of the entire dataset must be present in the first experience.

New classes (NC) also known as class-incremental learning, where each experience includes data of classes not present in any other experience. Formally, we can define the NC scenario as:

$$\mathcal{Y}_k \cap \bigcup_{i=1}^{k-1} \mathcal{Y}_i = \emptyset \quad \text{for } k = \{2, ..., N_E\}. \tag{9}$$

New instances and classes (NIC) where a new experience can contain already seen classes, new classes, or a mix of the two. This is the most natural scenario since in the real world an agent may sense both known and unknown objects. Formally the NIC scenario can be defined as:

$$\exists k : \mathcal{Y}_k \cap \bigcup_{i=1}^{k-1} \mathcal{Y}_i \neq \emptyset \text{ and } \exists j : \mathcal{Y}_1 \cap \mathcal{Y}_j \neq \mathcal{Y}_j. \tag{10}$$

Meaning that there is at least one experience that contains classes already seen in the past (left part) and at least one experience that contains classes not present in the first experience (right part).

B PSEUDO-CODE OF THE WEIGHT CONSOLIDATION PHASE

Algorithm 1 Weight consolidation

Require: $\psi,\ \psi',\ \mathcal{Y}_e,\ \mathcal{M}_e^y$

for each class $c \in \mathcal{Y}_e \cup \mathcal{M}_e^y$ **do**
  $\psi^c = \psi^c - \mu(\psi^c)$
  if $c \in \mathcal{Y}_e \wedge c \in \bigcup_{i=1}^{e-1} \mathcal{Y}_i$ **then**
    $\psi^c = \frac{\psi'^c \cdot w_{past_c} + \psi^c}{w_{past_c} + 1}$
  end if
  if $c \notin \mathcal{Y}_e \wedge c \in \mathcal{M}_e^y$ **then**
    if positive replay **then**
      $\psi^c = \frac{\psi'^c \cdot w_{past_c} + \psi^c}{w_{past_c} + 1}$
    end if
    if negative replay **then**
      $\psi^c = \psi'^c$
    end if
  end if
end for
$\psi' = \psi$

C DETAILS OF THE GENERATIVE MODEL IMPLEMENTATION

We designed our generative model using different insights from previous works in the fields, bringing together different ideas and proposals. We extensively tested the generative model alone to find the

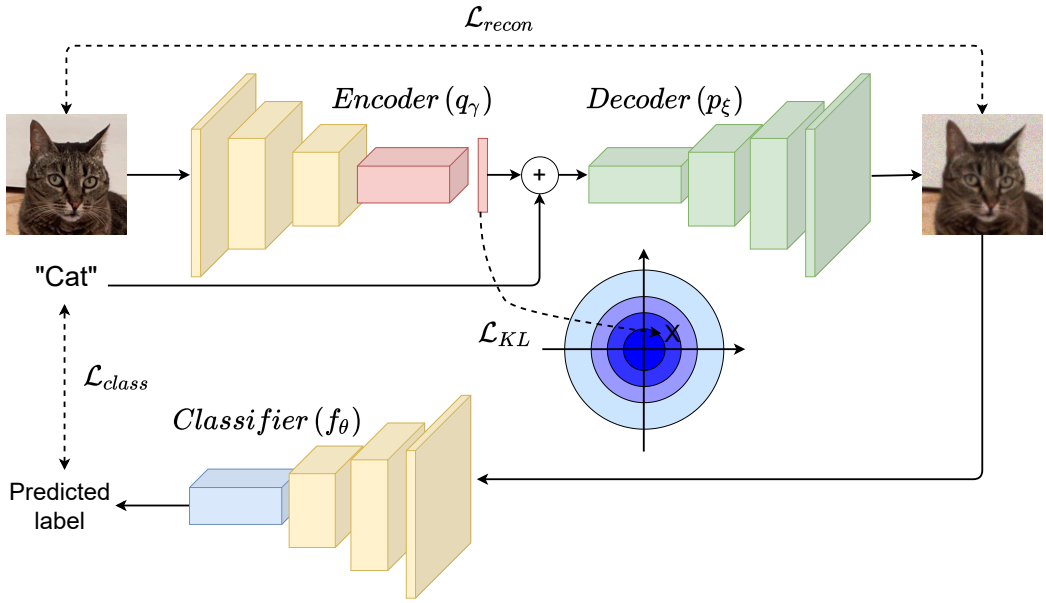


Figure 5: A visual schema of the generative model training. Losses are represented by dashed arrows. The shared branch of the classifier and the encoder are depicted using the same color (yellow). The encoder and the classifier's additional layers are drawn in red and blue respectively.

better combination of building blocks that yield the best performance. Our design choices have been also influenced by the computation complexity since our aim is to develop a (near) real-time system. This is a particularly hard constraint since many incremental generative replay methods are based on generative adversarial networks (GANs) (Goodfellow et al., 2014), which notably have long training phases and often suffer from instabilities due to the adversarial nature of the training procedure.

As discussed in the main text, we took inspiration from some state-of-the-art methods, trying to combine promising techniques and ideas from different sources. Taking inspiration from van de Ven et al. (2020) we use a Variational Autoencoder (VAE) Kingma & Welling (2014) model, but unlike van de Ven et al. (2020) we opted for a conditional VAE (cVAE) configuration Sohn et al. (2015). So, while in van de Ven et al. (2020) a mixture of Gaussian is used to sample latent vectors and soft labels are provided to the classifier itself, in our approach the latent vector is sampled from the normal distribution $\mathcal{N}(0, 1)$ and conditioned to the desired class. This results in a faster and less complicated sampling of a replay pattern. Moreover, as in van de Ven et al. (2020) we partially blend the encoder part of the generative model with the classifier model: in fact both the network share the same feature extractor f_ϕ. For the classifier, this branch is connected with the classification head c_ψ, while, for the generator, it is connected with some other layers that transform the feature into a latent vector z. The bifurcation is located in the latent replay layer. The resulting on-the-loop training of the generative model is consistent with brain structures and neuroscience's findings (van de Ven et al., 2020).

Since we use a cVAE, the objective for the generative model can be expressed as:

$$\gamma^*, \xi^* = \arg\min_{\gamma,\xi}[-\mathbb{E}_{z\sim q_\gamma(z|x_i^k)}[\log p_\xi(z|y_i^k)] + D_{KL}(q_\gamma(z|x_i^k)||p(z))], \tag{11}$$

where (x_i^k, y_i^k) are the data point and the label of the i-th pattern of the k-th experience, and the D_{KL} term represents the Kullback-Leibler divergence between the latent space distribution and the target distribution $p(z) = \mathcal{N}(0, 1)$.

The two terms of Equation 11 determine two losses:

$$\mathcal{L}_{recon} = \|x_i^k - p_\xi(q_\gamma(x_i^k))\|_2^2 \tag{12}$$

$$\mathcal{L}_{KL} = D_{\mathrm{KL}}(q_\gamma(x_i^k)||\mathcal{N}(0, 1)) \tag{13}$$

We also add another loss term, denoted as *classification loss*, which is similar to the classifier loss adopted in the AC-GAN model (Odena et al., 2017). The rationale is to guide the generative model to produce data that are not only visually similar to the original ones (L2 loss) but that is also classified by the current classifier in the same way. Hence, we use f_Θ as "auxiliary" classifier, adding the following term to the generator's loss:

$$\mathcal{L}_{class} = -\log f_\Theta(y_i^k | p_\xi(q_\gamma(x_i^k)|y_i^k)), \tag{14}$$

which represents a typical negative log-likelihood classification loss. Note that the parameters Θ of the classifier are not trained in this phase, since only the generative model is updated. Overall, the generative model is trained using the following loss function:

$$\mathcal{L}_{GM} = \mathcal{L}_{recon} + \beta\mathcal{L}_{KL} + \eta\mathcal{L}_{class}, \tag{15}$$

where β is a hyper-parameter inspired to the β-VAE framework (Higgins et al., 2017), and η is a hyper-parameters that weights the importance of the classification loss.

A visual representation of generative model training is shown in Figure 5.

To keep notation light, in the equations above the replay memory is not used, but it is trivial to include patterns from the replay memory, since there is no distinction in the generative model training procedure between current and replay data.

Note that the utilization of raw images is not mandatory for the method, and any intermediate (or latent representation) can be used, making our proposal compatible with latent replay methods (Pellegrini et al., 2020; van de Ven et al., 2020). In fact, in the case of latent replay, the data points x_i^k in the above equations can be simply substituted with $f_{\phi'}(x_i^k)$, where $f_{\phi'}$ is the set of feature extraction layers before the latent replay layer.

The blending of a part of the generative model into the classifier poses some difficulties in the training, especially regarding the balancing of the two models and how to train each of them without destructive inference on the other. After some initial experiments, we opted for blocking model parameters when the other model is trained. Detailed pseudo-code for the proposed negative generative replay strategy is provided in Algorithm 2.

Algorithm 2 Generative negative replay

$f_\Theta \leftarrow$ RANDINIT or PRETRAINED
$g_\Omega \leftarrow$ RANDINIT or PRETRAINED
$\mathcal{M}^x \leftarrow \emptyset, \;\; \mathcal{M}^y \leftarrow \emptyset$
$R =$ memory size
for each k from 1 tp N_E **do**
    if $k > 1$ **then**
        SAMPLE $\{z_1, ..., z_R\} \sim \mathcal{N}(0, 1)$
        SAMPLE $\{c_1, ..., c_R\} \sim \bigcup_{t=1}^{k-1} \mathcal{Y}_t$
        BLOCK generator parameters (γ, ξ)
        POPULATE $\mathcal{M}_k^x = p_\xi(z_j|c_j)), \; j = \{1, ..., R\}$
        POPULATE $\mathcal{M}_k^y = \{c_1, ..., c_R\}$
    end if
    `# classifier training`
    $\psi' = \psi$
    BLOCK generator parameters (γ, ξ)
    UNLOCK classifier parameters (ϕ, ψ)
    $\phi^*, \psi^* =$ OPTIMIZE$(f_\Theta, \mathcal{X}_k \cup \mathcal{M}_k^x, \mathcal{Y}_k \cup \mathcal{M}_k^y)$ using Equation 4
    WEIGHTCONSOLIDATION$(\psi, \psi', \mathcal{Y}_k, \mathcal{M}_k^y)$ (see Algorithm 1)
    `# generator training`
    BLOCK classifier parameters (ϕ, ψ)
    UNLOCK generator parameters (γ, ξ)
    $\gamma^*, \xi^* =$ OPTIMIZE$(g_\Omega, \mathcal{X}_k \cup \mathcal{M}_k^x, \mathcal{Y}_k \cup \mathcal{M}_k^y)$ using Equation 15
end for

Method	Final Accuracy
Fine Tuning (Naive)	27.4
EWC (Kirkpatrick et al., 2017)	28.4
RWalk (Chaudhry et al., 2018)	24.9
LwM (Dhar et al., 2019)	17.7
LwF (Li & Hoiem, 2016)	19.8
iCaRL (Rebuffi et al., 2017)	30.2
EEIL (Castro et al., 2018)	25.1
LUCIR (Hou et al., 2019)	20.1
IL2M (Belouadah & Popescu, 2019)	29.7
BiC (Wu et al., 2019)	32.4
AR1 (Maltoni & Lomonaco, 2019)	**33.1**

Table 3: Final accuracy on ImageNet-1000 following the benchmark of Masana et al. (2020) with 25 experiences composed of 40 classes each. For each method, a replay memory of 20,000 patterns is used (20 per class at the end of training). Results for other methods reported from Masana et al. (2020).

D Validation of AR1 on ImageNet-1000

To validate the chosen AR1 algorithm we performed a test on a competitive benchmark on ImageNet-1000, following the NC benchmark proposed by Masana et al. (2020), which is composed of 25 experiences, each of them containing 40 classes. The benchmark is particularly challenging due to a large number of classes (1,000), the incremental nature of the task (with 25 experiences), and the data dimensionality of 224×224 (as with ImageNet protocol).

With this experiment we want to assess the performance of AR1 in a complex continual learning scenario, validating the choice of AR1 as the baseline algorithm on which the tests on negative replay are conducted. In this experiment, we tested AR1 against both regularization-based methods (Dhar et al., 2019; Kirkpatrick et al., 2017; Li & Hoiem, 2016) and replay-based approaches (Belouadah & Popescu, 2019; Castro et al., 2018; Chaudhry et al., 2018; Hou et al., 2019; Rebuffi et al., 2017; Wu et al., 2019). We use the same classifier (ResNet-18 (He et al., 2016)) and the same memory size for all the tested methods (20,000 patterns, 20 per class); for the regularization-based approaches, the replay is added as an additional mechanism.

For AR1, we trained the model with an SGD optimizer. For the first experience, we used an aggressive learning rate of 0.1 with momentum 0.9 and weight decay of 10^{-4}. We multiply the initial learning rate by 0.1 every 15 epochs. We trained the model for a total of 45 epochs, using a batch size of 128. For all the subsequent experiences we used SGD with a learning rate of $5 \cdot 10^{-3}$ for the feature extractor's parameters ϕ and $5 \cdot 10^{-2}$ for the classifier's parameters ψ. We trained the model for 32 epochs for each experience, employing a learning rate scheduler that decreases the learning rate as the number of experiences progresses. This was done to protect old knowledge against new knowledge when the former is more abundant than the latter. As in the first experience, the batch size was set to 128, composed of 92 patterns from the current experience and 36 randomly sampled (without replacement) from the replay memory.

The results are shown in Table 3. Replay-based methods exhibit the best performance, with iCaRL and BiC exceeding a final accuracy of 30%. AR1 outperforms all the baselines (33.1%), demonstrating the validity of this approach also in difficult continual learning benchmarks. However, considering that top-1 ImageNet accuracy for a ResNet-18 when trained on the entire dataset is 69.76%[3], even for the best methods the accuracy gap in the continual learning setup is very large. This suggests that continual learning, especially in complex scenarios with a large number of classes and high dimensional data, is far to be solved, and further research should be devoted to this field.

[3]Accuracy taken from the torchvision official page: `https://pytorch.org/vision/stable/models.html`

E CLASSIFIER HYPER-PARAMETERS

E.1 CORE50 NC

	Hyper-parameter	Value
Common	optimizer	SGD
	momentum	0.9
	weight decay	10^{-4}
	minibatch size	128
	SI (Synaptic Intelligence) λ	$8 \cdot 10^5$
	SI Fisher matrix clip value	10^{-3}
	SI Fisher matrix multiplier	10^{-6}
1st experience	nr. epochs	4
	lr ϕ (feature extractor)	$3 \cdot 10^{-4}$
	lr ψ (classification head)	$3 \cdot 10^{-4}$
Following experiences	nr. epochs	4
	lr ϕ (feature extractor)	$3 \cdot 10^{-4}$
	lr ψ (classification head)	$3 \cdot 10^{-4}$

Table 4: Hyper-parameters of the model trained with **no replay**. Common hyper-parameters are the same for each experience, 1st experience hyper-parameters are used in the first experience, the following experience hyper-parameters are used in all the following experiences.

	Hyper-parameter	Value
Common	optimizer	SGD
	momentum	0.9
	weight decay	10^{-4}
	minibatch size	128
	SI (Synaptic Intelligence)	disabled
1st experience	nr. epochs	4
	lr ϕ (feature extractor)	$3 \cdot 10^{-2}$
	lr ψ (classification head)	$3 \cdot 10^{-2}$
Following experiences	nr. epochs	4
	lr ϕ (feature extractor)	$5 \cdot 10^{-5}$
	lr ψ (classification head)	$5 \cdot 10^{-4}$
	memory size	1,500
	replay pattern in the minibatch	14
	latent replay layer	conv5_4

Table 5: Hyper-parameters of the model trained with **replay** (generative replay, random data, and real data). Common hyper-parameters are the same for each experience, 1st experience hyper-parameters are used in the first experience, the following experience hyper-parameters are used in all the following experiences.

E.2 IMAGENET-1000 NC

	Hyper-parameter	Value
Common	optimizer	SGD
	momentum	0.9
	weight decay	10^{-4}
	minibatch size	128
	SI (Synaptic Intelligence)	disabled
1st experience	nr. epochs	45
	lr ϕ (feature extractor)	10^{-1}
	lr ψ (classification head)	10^{-1}
	lr scheduler	lr $\cdot$ 0.1 every 15 epochs
Following experiences	nr. epochs	32
	lr ϕ (feature extractor)	$5 \cdot 10^{-3}$
	lr ψ (classification head)	$5 \cdot 10^{-2}$
	lr scheduler	see Equation 16

Table 6: Hyper-parameters of the model trained with **no replay**. Common hyper-parameters are the same for each experience, 1st experience hyper-parameters are used in the first experience, the following experience hyper-parameters are used in all the following experiences.

	Hyper-parameter	Value
Common	optimizer	SGD
	momentum	0.9
	weight decay	10^{-4}
	minibatch size	128
	SI (Synaptic Intelligence)	disabled
1st experience	nr. epochs	45
	lr ϕ (feature extractor)	10^{-1}
	lr ψ (classification head)	10^{-1}
	lr scheduler	lr $\cdot$ 0.1 every 15 epochs
Following experiences	nr. epochs	32
	lr ϕ (feature extractor)	$5 \cdot 10^{-3}$
	lr ψ (classification head)	$5 \cdot 10^{-2}$
	lr scheduler	see Equation 16
	memory size	20,000
	replay pattern in the minibatch	36
	latent replay layer	layer4 (4th resnet block)

Table 7: Hyper-parameters of the model trained with **replay** (generative replay and real data). Common hyper-parameters are the same for each experience, 1st experience hyper-parameters are used in the first experience, the following experience hyper-parameters are used in all the following experiences.

Due to the complexity of the ImageNet-1000 scenario, we found it useful to use a learning rate scheduler that decreases the learning rate as the number of experiences progresses. The scheduler can be formalized as:

$$\text{lr} = \text{lr}_{init} \cdot \left(-\frac{0.9}{1 + e^{-1.5i+8}} + 1 \right), \tag{16}$$

where i indicates the index of the current experience.

E.3 CORE50 NIC

	Hyper-parameter	Value
Common	optimizer	SGD
	momentum	0.9
	weight decay	10^{-4}
	minibatch size	128
	SI (Synaptic Intelligence) λ	$2.3 \cdot 10^{6}$
	SI Fisher matrix clip value	10^{-3}
	SI Fisher matrix multiplier	$2 \cdot 10^{-5}$
1st experience	nr. epochs	4
	lr ϕ (feature extractor)	10^{-3}
	lr ψ (classification head)	10^{-3}
Following experiences	nr. epochs	4
	lr ϕ (feature extractor)	10^{-4}
	lr ψ (classification head)	10^{-3}

Table 8: Hyper-parameters of the model trained with **no replay**. Common hyper-parameters are the same for each experience, 1st experience hyper-parameters are used in the first experience, the following experience hyper-parameters are used in all the following experiences.

	Hyper-parameter	Value
Common	optimizer	SGD
	momentum	0.9
	weight decay	10^{-4}
	minibatch size	128
	SI (Synaptic Intelligence)	disabled
1st experience	nr. epochs	4
	lr ϕ (feature extractor)	10^{-3}
	lr ψ (classification head)	10^{-3}
Following experiences	nr. epochs	4
	lr ϕ (feature extractor)	10^{-4}
	lr ψ (classification head)	10^{-3}
	memory size	300 (N/A for random data)
	replay pattern in the minibatch	64 (21 for random data)
	latent replay layer	conv5_4

Table 9: Hyper-parameters of the model trained with **replay** (generative replay, random data, and real data). Common hyper-parameters are the same for each experience, 1st experience hyper-parameters are used in the first experience, the following experience hyper-parameters are used in all the following experiences.

F GENERATIVE MODEL HYPER-PARAMETERS

F.1 CORE50 NC

	Hyper-parameter	Value
Common	optimizer	Adam
	betas	0.9 - 0.999
	weight decay	0
	minibatch size	128
	latent space dim.	100
	β	0.1
	η	0.01
	lr	$2 \cdot 10^{-3}$
	lr scheduler	None
	nr. epochs	4
Following experiences	replay patterns in the minibatch	27

Table 10: Hyper-parameters of the generative model trained on CORe50 NC. Common hyper-parameters are the same for each experience, while following experience hyper-parameters are used in all the experiences except the first one.

F.2 IMAGENET-1000 NC

	Hyper-parameter	Value
Common	optimizer	SGD
	momentum	0
	weight decay	0
	minibatch size	128
	latent space dim.	100
	β	0.25
	η	0.01
	lr	1
	lr scheduler	see Equation 16
	nr. epochs	32
Following experiences	replay patterns in the minibatch	36

Table 11: Hyper-parameters of the generative model trained on ImageNet-1000 NC. Common hyper-parameters are the same for each experience, while following experience hyper-parameters are used in all the experiences except the first one.

F.3 CORe50 NIC

	Hyper-parameter	Value
Common	optimizer	Adam
	betas	0.9 - 0.999
	weight decay	0
	minibatch size	128
	latent space dim.	100
	β	0.1
	η	0.01
	lr	$2 \cdot 10^{-3}$
	lr scheduler	None
	nr. epochs	4
Following experiences	replay patterns in the minibatch	64

Table 12: Hyper-parameters of the generative model trained on CORe50 NIC. Common hyper-parameters are the same for each experience, while following experience hyper-parameters are used in all the experiences except the first one.

G Additional Plots

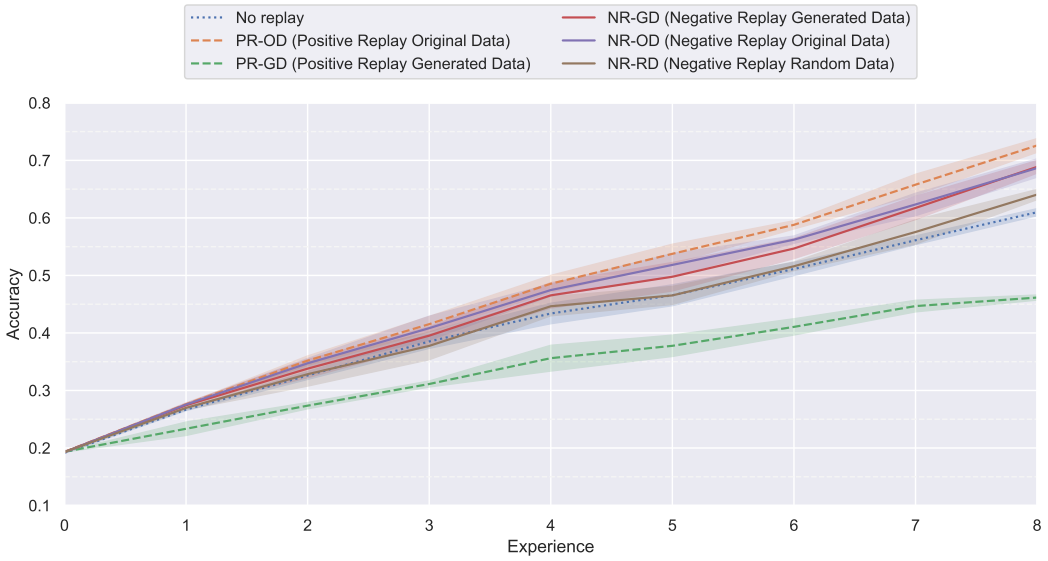


Figure 6: Overall accuracy on the CORe50 NC scenario for all the experiments performed in this work (included random data and negative replay with original data). After each experience, the model was evaluated using the cumulative test set as proposed by Lomonaco & Maltoni (2017). Every experiment is averaged over 3 runs, with different seeds and class order. The standard deviation is reported in light colors. better viewed if zoomed on a computer monitor.

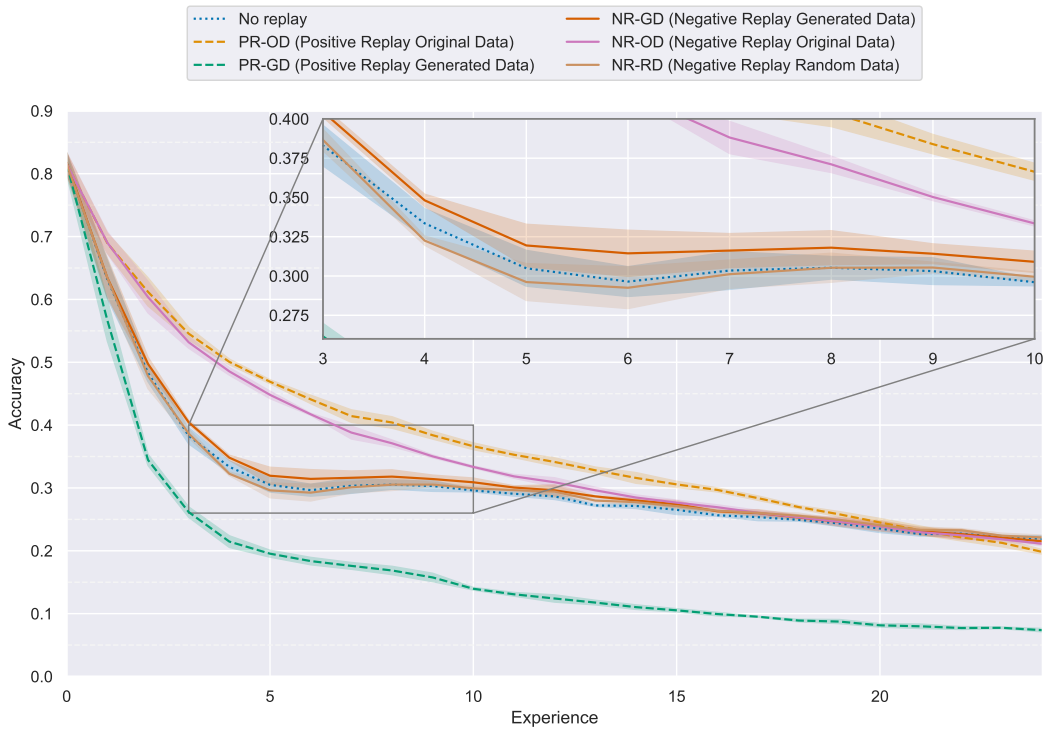


Figure 7: Overall accuracy on the ImageNet-1000 NC scenario for all the experiments performed in this work (included random data and negative replay with original data). After each experience, the model was evaluated using the whole test set as proposed by Masana et al. (2020). Every experiment is averaged over 3 runs, with different seeds and class order. The standard deviation is reported in light colors. better viewed if zoomed on a computer monitor.

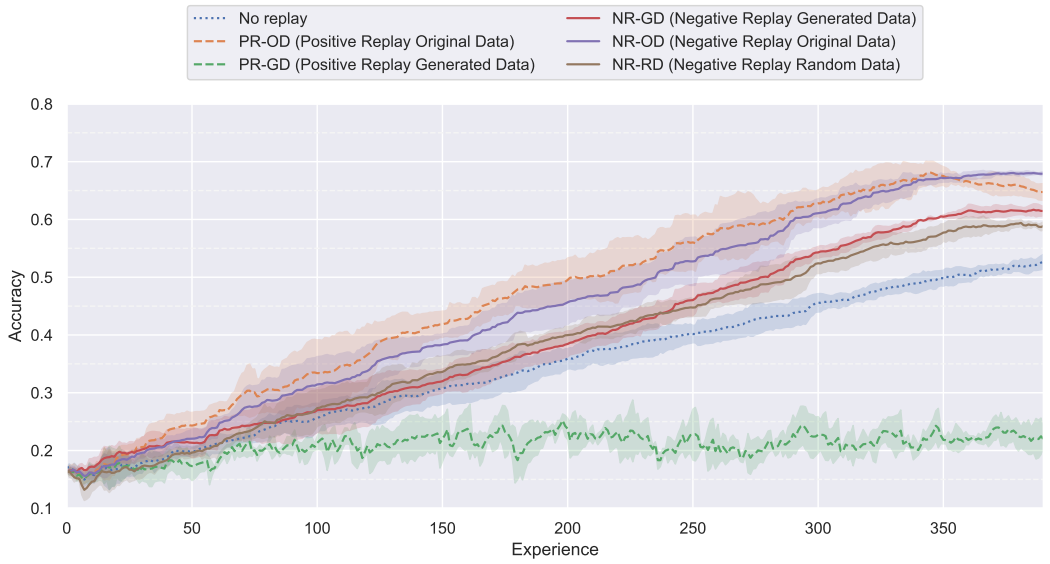


Figure 8: Overall accuracy on the CORe50 NIC scenario for all the experiments performed in this work (included random data and negative replay with original data). After each experience, the model was evaluated using the cumulative test set as proposed by Lomonaco & Maltoni (2017). Every experiment is averaged over 3 runs, with different seeds and class order. The standard deviation is reported in light colors. better viewed if zoomed on a computer monitor.