Class-Incremental Continual Learning Into the eXtended DER-Verse

Matteo Boschini, Graduate Student Member, IEEE, Lorenzo Bonicelli, Pietro Buzzega, Angelo Porrello, and Simone Calderara, Member, IEEE

Abstract—The staple of human intelligence is the capability of acquiring knowledge in a continuous fashion. In stark contrast, Deep Networks forget catastrophically and, for this reason, the sub-field of Class-Incremental Continual Learning fosters methods that learn a sequence of tasks incrementally, blending sequentially-gained knowledge into a comprehensive prediction. This work aims at assessing and overcoming the pitfalls of our previous proposal Dark Experience Replay (DER), a simple and effective approach that combines rehearsal and Knowledge Distillation. Inspired by the way our minds constantly rewrite past recollections and set expectations for the future, we endow our model with the ability to i) revise its replay memory to welcome novel information regarding past data ii) pave the way for learning yet unseen classes. We show that the application of these strategies leads to remarkable improvements; indeed, the resulting method – termed eXtended-DER (X-DER) – outperforms the state of the art on both standard benchmarks (such as CIFAR-100 and minImageNet) and a novel one here introduced. To gain a better understanding, we further provide extensive ablation studies that corroborate and extend the findings of our previous research (e.g., the value of Knowledge Distillation and flatter minima in continual learning setups. We make our results fully reproducible; the codebase is available at https://github.com/aimagelab/mammoth.

Index Terms—Continual learning, catastrophic forgetting, class-incremental, knowledge distillation, replay methods

1 INTRODUCTION

HUMAN intelligence allows us to acquire new knowledge about the surrounding world in a natural way. Thanks to the extraordinary and still unclear capabilities of the human brain, we can learn new and complex tasks (e.g., driving cars) and, at the same time, remember the old ones (e.g., cycling) without experiencing either interference or forgetting. Moreover, human beings exhibit an ability called fluid intelligence [1]: according to this construct, humans can reason about and engage with novel problems in a manner that does not explicitly rely on prior learning. Namely, we can recognize new patterns and plan new strategies in unseen environments in a manner that only minimally depends upon specific previous experience or acculturation.

Despite the long-standing parallelism between the human brain and Artificial Neural Networks (ANNs), the latter do not support these abilities and struggle [2]: indeed, novel knowledge tends to overwrite the old one, thus leading to a disruptive degradation of performance in previous tasks. Such an issue – which is widely known as catastrophic forgetting [2] – currently represents a hindrance towards the broad applicability of ANNs: to overcome this limitation, the field of Continual Learning (CL) includes a wide array of methods to let ANNs retain their performance [3], [4]. Whether by loss terms that prevent the model from changing [5], [6], [7], by explicitly using distinct parts of the model at distinct times [8], [9], or by revisiting past data [10], [11], [12], CL approaches let current training be influenced by previously learned information to preserve it. The exploitation of an episodic memory (i.e., a subset of past data that are continuously revisited) is undoubtedly one of the most reliable ways to face the aforementioned problem [12], [13], [14]. Due to its effectiveness, a plethora of approaches deal with its design and differ in the following facets: when the memory has to be populated [15] (e.g., at discrete intervals versus continuously); which elements we keep in memory and which ones we move out [12], [16], [17]; what kind of regularization we apply on these examples [18], [19] and, consequently, what information we store in it (e.g., also old model responses [14]). However, there is a promising direction that is still unexplored: how the memory has to be updated, i.e., the rewriting of past experiences to meet new insights regarding old events. This activity is peculiar to human beings [20], [21]: memory, indeed, is not to be understood as a mere video-camera recording events; instead, the hippocampus reframes past events to create a tale that fits the current world, thus helping us to take good decisions and focus on what is important in the here and now. Such a capability represents a source of inspiration for this work: in fact, we firstly extend our previous proposal [14] – called Dark Experience Replay (DER) – by equipping its memory with an update procedure that implants information from the present into the retained memories.
The manner memory is kept up-to-date is nevertheless only one of the two directions this work investigates. Recent studies highlight that the episodic memory also plays a key role in the mental simulation of future events [22], [23]; in other words, previously learned concepts influence our expectations about the future. We try to replicate this effect in our CL algorithm by further devising a future preparation strategy: i.e., a technique that exploits past and present data to prepare future classification heads to accommodate meaningful information.

To sum up, this work identifies some shortcomings in the way DER [14] organizes present and future knowledge. Consequently, we address them by a two-fold enhancement: on the one hand, we propose a procedure that maintains the memory buffer up-to-date by inserting secondary information from the present into memories of the past; on the other, we evaluate the benefits of preparing the underlying model to incoming tasks. Thanks to these improvements, our revised method – which we call eXtended-DER (a.k.a. X-DER) – achieves a remarkable increase in accuracy w.r.t. the current state of the art on two standard benchmarks (Split CIFAR-100 and Split miniImageNet) and on the newly introduced Split NTU-60.

In addition to extensive ablation studies highlighting the rationale behind our intuitions, we remark the following contributions:

- We shed light on some pitfalls of our previous proposal and, on this basis, propose X-DER, a novel CL method that embraces memory update and future preparation.
- In light of recent advances regarding the foundations of Knowledge Distillation [24], we review and provide new insights on the benefits of logits-replay against catastrophic forgetting.
- We deepen the discussion of our previous work about the geometry of local minima in CL, thus strengthening what we and other authors [25] have recently stated. From this perspective, we conduct several evaluations on X-DER Rand show that it favorably attains flatter minima.

2 RELATED WORK

2.1 Continual Learning

Recent years have witnessed a surge in the interest for methods that can alleviate the phenomenon of catastrophic forgetting [2], where the goal is to obtain a model that can adapt to changes in the distribution of input data (plasticity) while retaining the previously learned knowledge (stability). This problem is modeled in literature by means of different settings [26], which typically unfold a base classification problem in successively learned tasks. In the task incremental setting (Task-IL or multi-head), the learner must learn and remember how to classify within each task, as it is given access to the task identity of test samples at inference time. Vice-versa, the class incremental scenario (Class-IL or single-head) requires the task identifier to be predicted along with the sub-class label. Lastly, the domain incremental setting (Domain-IL) does not alter the distribution of classes; instead, it characterizes task changes by introducing a shift in the input distribution.

Methods specifically designed to tackle the Task-IL scenario usually exhibit a multi-headed architecture [8] or use the provided task information to map portions of the model to specific tasks [27]. While these strategies often prove effective, their use is limited to the multi-head setting.

By contrast, recent CL works focus mainly on Class-IL [12], [28], [29], [30], [31], as it is more general and regarded as more realistic than the other scenarios [12], [32]. Methods capable of working in a single-head assumption are usually divided into two families: regularization-based and rehearsal-based.

The former use specifically designed regularization terms to lead the optimization towards a good balance between stability and plasticity. Typically, they introduce a penalty term that discourages alterations in the weights that are vital for the previous tasks while the new ones are being learned [5], [6]. These methods can be effective for short sequences of tasks but usually fail to scale to complex problems [12], [32].

On the other hand, rehearsal models take advantage of a memory buffer to store exemplar elements from previous distributions. Experience Replay [33], [34] (ER) simply replays the stored elements along with the input stream to simulate training over an independent and identically distributed task (joint training). Despite its simplicity, this method has proven to be highly effective even with a minimal memory footprint [15] and serves as the basis for recent methods that propose variations on the strategies for the selection of samples to include in the memory buffer [12] or the sampling of examples from it [16]. The retained knowledge can also be used as a mean to revise the optimization procedure: MER [18] employs meta-learning to discourage interference and maximize knowledge transfer between tasks, while GEM [19] and A-GEM [35] use old training data to minimize the gradient interference in an explicit fashion.

2.2 Self-Distillation

Knowledge Distillation (KD) [36] is a training methodology that allows transferring the knowledge of a teacher model into a separate student model. While Hinton et al. originally proposed to distillate large teachers – possibly ensembles – into smaller students, further studies revealed additional interesting properties about this technique. In particular, Furlanello et al. [37] show that multiple rounds of distillation between models with the same architecture (termed self-distillation) can surprisingly improve the performance of the student. More recently, other works [38], [39] explore an interesting variation of self-distillation that distills knowledge from the deeper layers of the network to its shallower ones to accelerate convergence and attain higher accuracy.

Knowledge Distillation and Continual Learning. Distillation can be used to hinder catastrophic forgetting by appointing a previous snapshot of the model as teacher and distilling from it while new tasks are learned. Learning Without Forgetting [7] uses teacher responses to new exemplars to constrain the evolution of the student. Several other works combine distillation with rehearsal: iCalR [11] distills the responses of the model at the previous task boundary, learning latent representations to be used in a nearest mean-of-exemplars classifier; ETEIL [40], LUCIR [31] and BiC [41] focus on contrasting the prediction bias that comes from incremental classification; ILM2 [30] stores additional statistics to facilitate distillation and compensate bias.
3 BACKGROUND

3.1 Class-Incremental Continual Learning

In Class-Incremental Continual Learning (CiCL), a model \( f(\cdot; \theta) \) is trained on a sequence of \( T \) tasks \( \{ T_0, \ldots, T_{T-1} \} \), having access to one at a time. The \( i \)-th task consists of data-points \( \{ x_i(n), y_i(n) \}_{n=1}^{T_i} \), where \( y_i(n) \in \mathcal{Y}_i \), with disjoint ground-truth values for different tasks, i.e., \( \mathcal{Y}_i \cap \mathcal{Y}_j = \emptyset \) s.t. \( i \neq j \).

For the sake of simplicity, we assume all tasks having the same number of classes, i.e., \( |\mathcal{Y}_i| = |\mathcal{Y}_j| = |\mathcal{Y}| \). While all \( x_i \) are i.i.d. within \( T_i \), the overall training procedure does not abide by the i.i.d. assumption, as input distribution shifts between tasks and labels change. The objective of CiCL is the minimization of the risk over all tasks:

\[
\mathcal{L}_{\text{CiCL}} \triangleq \sum_{i=0}^{T-1} \mathbb{E}_{(x,y) \sim T_i} [\mathcal{L}(f(x; \theta), y)],
\]

where \( \mathcal{L} \) stands for the loss (e.g., the categorical cross-entropy) of predicting \( f(x; \theta) \) given \( y \) as the true label. The optimal solution \( \theta^* \) should provide accurate predictions for all tasks; however, this has to be pursued by observing one task at a time. To account for this, its actual learning objective should combine the empirical risk on the current task \( T_c \) with a separate regularization term \( \mathcal{L}_R \):

\[
\mathcal{L}_{\text{CiCL}} \triangleq \sum_{(x,y) \sim T_c} \mathcal{L}(f(x; \theta), y) + \mathcal{L}_R.
\]

The second term \( \mathcal{L}_R \) serves a twofold purpose: \( i \) it prevents the model from forgetting past knowledge while fitting new data [34]; \( ii \) it encourages the learner to gather per-task classifiers into a single and harmonized one [31].

3.2 A Self-KD Approach: Dark Experience Replay

To design the regularization term of Eq. (2) for preserving the capabilities on old tasks, the approaches [7], [11] based on Knowledge Distillation (KD) use past snapshots \( f(\cdot ; \theta^{(i)}) \) as teachers for the model engaged on the current task \( c \):

\[
\mathcal{L}_R = \sum_{i=0}^{c-1} \mathbb{E}_{x \sim T_i} \left[ \text{KD}(f(x; \theta^{(i)}), f(x; \theta)) \right].
\]

Logit matching [36] constitutes a straightforward and effective approach to pursue the objective above:

\[
\mathcal{L}_R = \sum_{i=0}^{c-1} \mathbb{E}_{x \sim T_i} \left[ ||f(x; \theta^{(i)}) - f(x; \theta)||_2^2 \right].
\]

Notably, Eq. (4) violates CiCL, as it assumes the availability of data-points from previous tasks. To approximate it, our previous proposal [14] – termed Dark Experience Replay (DER) – introduced a small replay buffer \( \mathcal{M} \) that stores a limited amount of past examples \( x \) along with the model outputs \( f(x; \theta(t)) \), where \( t \) indicates the time of memory insertion. Eq. (4) can therefore be recast as:

\[
\mathcal{L}_R = \mathcal{L}_{\text{DER}} = \alpha \cdot \mathbb{E}_{(x,\ell) \sim \mathcal{M}} ||(\ell - f(x; \theta)||_2^2.
\]

We further introduced Dark Experience Replay++ (DER++), which replays both logits and ground-truth labels:

\[
\mathcal{L}_{\text{DER++}} = \mathcal{L}_{\text{DER}} + \beta \cdot \mathbb{E}_{(x,y) \sim \mathcal{M}} \mathcal{L}(f(x; \theta), y).
\]

Pre-allocation of future heads. In several CiCL works, the model ends with a linear layer projecting into a space with as many features as classes have been encountered so far: as the number of the latter increases one task after the other, some approaches [7], [31] extend the classifier by instantiating a new classification layer (a.k.a. head); while some others [11], [18] initialize the network by providing as many classification heads as tasks are encountered from the beginning to the end. Both our previous proposals, DER(++) belong to this second line of approaches: indeed, they provide all necessary classification heads from the beginning and let their parameters be subject to optimization. It is noted that this practice – which we show in the following opens up interesting possibilities – is not entirely novel to the field: [12], [16] involve a cross-entropy term spanning over all classes (hence, already seen and yet unseen), which silences the activity of future heads (as we discussed later).

3.3 Limitations of Dark Experience Replay

Terminology. To facilitate the discussion, we introduce a categorization that splits the output space of the model at task \( T_c \) into the following partitions (see also Fig. 1):

- **Past** \( \mathcal{P}[c] \equiv \{0, \ldots, c \cdot |\mathcal{Y}| - 1\} \) i.e., logits modeling the probabilities of classes observed up to the current task.
- **Present** \( \mathcal{P}[c] \equiv \{c \cdot |\mathcal{Y}|, \ldots, (c + 1) \cdot |\mathcal{Y}| - 1\} \) i.e., logits of the head associated to the current task \( c \).
- **Future** \( \mathcal{F}[c] \equiv \{(c + 1) \cdot |\mathcal{Y}|, \ldots, T \cdot |\mathcal{Y}| - 1\} \) i.e., logits corresponding to unseen classes. They are not useful for classifying examples seen thus far, but will be needed during the following tasks.

It is noted that the composition of these partitions depends upon the specific task the model is learning (some logits move from one partition to the other when passing to the subsequent task). We then identify only for buffer data-points the logits of future past: namely, the set of classes the model discovered after the example was inserted in \( \mathcal{M} \).

Future past \( \mathcal{F}_p[c] \equiv \{j \cdot |\mathcal{Y}|, \ldots, (j + 1) \cdot |\mathcal{Y}| - 1\} \) given an \((x, y, \ell) \in \mathcal{M}\) stored at task \( \ell (\ell < c) \), these logits model the classes of the \( j \)-th task discovered after the insertion of the example into the buffer.

1. Section 6.2 shows that knowing in advance the total number of tasks does not represent a crucial hypothesis and can be easily removed.
We now discuss two weaknesses of DER(++) concerning the partitions of future and future past logits. Afterwards – in Section 4 – we discuss of those could be overcome.

(L1) DER(++) have a blind spot for future past. Our previous work showed that the memory buffer of DER(++) provides a more efficient and informative way to refresh old tasks. However, we acknowledge here that the information it carries is limited solely to the classes already seen at the time an example was inserted in $M$.

Indeed, when inserting an example in the rehearsal memory, it stands to reason that past and present logits encode all the information useful for later replay. However, by the time we move to subsequent tasks, the model discovers new classes and, with them, their relations with the old ones (i.e., future past information). Unfortunately, DER(++) replaying do not profit from this incoming information, as they pin as target a version of future past logits that preceeds the effective observation of the corresponding classes; differently, we could update the memory buffer to capture this emerging knowledge. As reported in the following, this delivers a remarkable effect in the prevention of forgetting.

For an in-depth experiment showing that DER++ is blind towards future past classes and a comparison with our proposal, we refer the reader to Appendix A, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2022.3206549.

We now discuss two weaknesses of DER(++) concerning the partitions of future and future past logits. Afterwards – in Section 4 – we discuss of those could be overcome.

(L2) DER(++) overemphasize the classes of the current task. Several works [28], [29], [41], [42] have recently shed light on the accumulation of bias towards present classes and the negative impact it has on performance. We have found that also DER(++) are prone to such a pitfall: similarly to [41], we can quantitatively characterize it by evaluating how predictions distribute across different classification heads (as training progresses). In particular, we limit the analysis on misclassified examples belonging to tasks prior to the current one: Fig. 2a highlights which task comprises the predicted class (on average); as can be seen, the majority of wrong predictions end up in the last observed task.

On the one hand, the negative bias towards past classes can be ascribed to the optimization of the cross-entropy loss on examples from the current task. As pointed out in [28], when a new task is presented to the net, an asymmetry arises between the contributions of replay data and current examples to the weights updates: indeed, the gradients of new (and poorly fit) examples outweigh (Fig. 2b). If we aim at learning the current task solely, this is desirable as it favorably dampens logits of past classes. However, a hasty attenuation of earlier classes clashes with the second goal of avoiding forgetting of past concepts. In order to achieve a unified classifier, it is important to take countermeasures against such a phenomenon.

Similarly, we observe a consistent negative bias towards future classes. We ascribe this behavior to the cross-entropy loss: as its application spans all heads, the future ones always have zeroes as post-softmax targets and hence are strongly pushed towards pre-softmax negative values. On the one hand, this desirably prevents the model from trivial errors; however, we mean future heads to accommodate the learning of future tasks. Therefore – if the negative bias accumulates so strongly on these heads – the recovery from
that situation may slow down and complicate the learning of new tasks. In this regard, Fig. 2c illustrates the behavior of future logits and compares the average responses of both DER++ (Fig. 2c, left) and the approach discussed in Section 4 (Fig. 2c, right). As can be seen, the former consistently exhibits negative values for unseen classes; on the contrary, the latter avoids bias accumulation on the account of the regularization it imposes on future logits.

4 PROPOSED APPROACH

In this section, we discuss how the above-discussed limitations of DER(++) can be addressed. We refer the reader to Fig. 3 for a visual overview of the model thus enhanced, which we dub eXtended-DER (a.k.a. X-DER).

4.1 X-DER: Logits of Future Past

To prevent DER(++) from losing valuable secondary information, we devise a simple procedure that keeps its memory buffer updated. Let us suppose the model is learning the $c^k$ task and an example $(x, y, k) \in \mathcal{M}$ from a previous $\bar{c}^k$ task is sampled from the memory buffer for replay. The current network output $\ell \triangleq f(x; \theta)$ now contains the secondary information of task $T_c$ for $x$; therefore, we propose to implant the corresponding logits $\ell_{fp|c|}$ into the memory entry containing $\ell_M^k$. Such an operation only involves the head of the current task and is applied both while learning (in an ongoing manner) and at the end of it.

From a technical perspective, we do not simply overwrite previous logits with the new ones. Indeed, as the net suffers from bias towards present classes, simply implanting their values in the memory buffer and using these for later replay would exacerbate the issue even more. Instead, we take care of re-scaling the portion tied to future past in a way such that its maximum logit $\ell_{fpmax}^M = \max_{c|fp|c|} \ell_c^M$ is lower than the ground-truth one $\ell_{gt}^M = 1 - \text{hot}(y) \cdot \ell_M^k$ already in memory. Formally:

$$\ell_{fp}^M = \ell_k \cdot \min\left(1, \frac{\ell_{gt}^M}{\ell_{fpmax}^M}\right), \quad k \in \text{fp|c|}$$

where $\gamma \in [0, 1]$ is a hyperparameter controlling the attenuation rate (which we typically set to 0.75).

4.2 Future Preparation

Most CiCL methods exploit the information available up to the current task to prevent the leak of past knowledge. Here, we take an extra step and argue that the same care should be placed on preparing future heads to accommodate future classes. In this respect, Fig. 4 depicts the underlying intuition: considering the joint training on all tasks (Fig. 4, left) as the optimal solution we have to approximate, standard CL approaches (Fig. 4, center) seem to focus only on a (growing) part of the overall problem, i.e., what concerns the tasks seen up to the current one, as embodied by Eq. (2). Instead, we claim that even a coarse guess regarding unseen tasks can lead to a better estimate of Eq. (1).

To the best of our knowledge, X-DER is the first method pursuing this goal through optimization (Fig. 4, right): it conjectures about the semantics of logits corresponding to unseen classes, which are encouraged to be consistent across instances of the same class. As outlined by the field of contrastive self-supervised learning [43], [44], the skillful use of data augmentation techniques can lead towards useful representations even when no label information is made available to the learner. In a sense, this resembles the case of our future classes: hence, we expect it to be an effective warm-up strategy for future tasks.

Intuitively, the auxiliary objective we present in the following encourages each individual future classification head to yield similar responses for “similar” examples. However, as we dispose of label information for both the examples from the current task and the memory buffer, we
refine the contrastive objective by incorporating class supervision. As shown in [45], we can leverage it to pull together representations of examples from the same class and to do the opposite for different classes.

In practice, given a batch of $N$ examples, we extend it by appending $N$ variants of the original items (obtained through strong data augmentation). We then consider the output response $f(x_i; \theta) \approx \ell(x_i)$ for the $i$th example: in particular, we firstly focus on the (normalized) $j$th future head (s.t. $j \in \{c + 1, \ldots, T\}$), which we denote with $\hat{\ell}_{fu[cj]}(x_i) \triangleq \text{L2Norm}(\ell_{j\big|p[i]\ldots(j+1\big|p[i]-1}(x_i))$. On top of this, we compute the following loss term:

$$L_{sc}(x_i, y_i) = -\sum_{p \leq P(i)} \log \left( \frac{\exp(\hat{\ell}_{fu[cj]}(x_i) \cdot \hat{\ell}_{fu[cj]}(x_p))}{\sum_{k=1}^{2N} \exp(\hat{\ell}_{fu[cj]}(x_i) \cdot \hat{\ell}_{fu[cj]}(x_k))} / \tau \right),$$

where $P(i) = \{ p \in \{1, \ldots, 2N \} ~|~ i \neq p \land y_i = y_p \}$ stands for the positive set (i.e., the indices of examples sharing the label of the $i$th item) and $\tau$ is a positive scalar value that acts as a temperature. The full objective simply consists in averaging the values of Eq. (8) across all future heads:

$$L_{rv}(x_i, y_i) = \frac{1}{T - (c + 1)} \sum_{j = c + 1}^{T} \frac{1}{P(i)} L_{sc}(x_i, y_i),$$

Since Eq. (9) encourages unused heads to convey additional semantics about the examples seen so far, we profitably exploit also future logits during replay. Moreover, as new classes emerge in later tasks, we account for the corresponding semantic drift by applying the update procedure outlined in Section 4.1 also on these logits. We empirically found it beneficial to apply logits-replay also for future heads; therefore, the outlined update procedure also extends to the future logits retained in the memory buffer.

### 4.3 Bias Mitigation

Section 3.3 reports that one of the main weaknesses of our previous proposal regards the bias it accumulates on classes of the current task. This unfolds in two directions: on the one hand, most errors regarding past tasks are due to a blind preference of the model towards novel classes; on the other hand, future heads are only provided negative samples and therefore collapse to bad configurations that may hurt or slow down later learning.

**Preventing penalization of past classes.** As also observed in other recent works [28], [29], [42], this issue can be mitigated by revising the way the cross-entropy loss is applied during training. Given an example from the current task, we avoid computing the softmax activation on all logits and instead restrict it on those modeling the scores of the current task classes. This way, the predictions of past classes are removed from the equation and thus not penalized by the outweighing gradients of novel examples. In formal terms, we compute the following objective:

$$L_{SC-CE}(x_c, y_c) = CE(\text{softmax}(\ell_{pr[c]}(x_c)), y_c)$$

where $c$ indicates the index of the current task. We remark that, for the class-balanced buffer datapoints, such modification is not strictly necessary; hence, we naturally compute the softmax across the logits of all classes seen so far.

**Restraining past and future activations.** In light of the considerations above, we apply the cross-entropy term in isolation on present logits. This favorably prevents the dampening of both past and future heads; however, if left unchecked, nothing prevents their responses from outgrowing present ones and causing trivial classification errors.

To avoid this issue, we provide an optimization constraint to limit the activations of past and future heads: for current task examples, we require their corresponding responses to be lower than a safeguard threshold, identified as the logit $\ell_{gt}(x_i) \triangleq \text{one} - \text{hot}(y_i) \cdot \ell(x_i)$ corresponding to the ground-truth class. In doing so, we penalize the maximum past (future) logit $\ell_{pa-max}(x_i) = \max_{j \notin \{p[i]\}} \ell_j(x_i)$ ($\ell_{fu-max}(x_i) = \max_{j \notin \{p[i]\}} \ell_j(x_i)$) if it oversteps $\ell_{gt}(x_i)$:

$$L_{constraint}(x_i, y_i) = \max(0, \ell_{pa-max}(x_i) - \ell_{gt}(x_i) + m) + \max(0, \ell_{fu-max}(x_i) - \ell_{gt}(x_i) + m),$$

where $m$ is a hyper-parameter (we typically set it to 0.3 in our experiments) that controls the strictness of the penalty.

### 4.4 Overall Objective

To sum up, X-DER seeks to optimize the following minimization problem:

$$\arg \min_{\theta} L_{X-DER} = L_{DER} + L_{SC-CE} + L_F,$$

where $L_{DER}$ denotes the objective reported in Eq. (5), $L_{SC-CE}$ rewrites Eq. (10) to take into account examples from both the current task $c$ and the memory buffer

$$L_{SC-CE} = \frac{1}{T - (c + 1)} \sum_{j = c + 1}^{T} \frac{1}{P(i)} L_{sc}(x_i, y_i),$$

and $L_F$ groups together the goals outlined in Sections 4.2 and 4.3.

$$L_F = \frac{1}{T - (c + 1)} \sum_{j = c + 1}^{T} \frac{1}{P(i)} L_{sc}(x_i, y_i),$$

It is noted that $\beta$, $\lambda$, and $\eta$ are three hyperparameters weighing each contribution to the overall loss. For the sake of clarity, Fig. 5 proposes a visual breakdown of the loss terms and the involved partitions of the classifier. For a deeper technical understanding of X-DER, we refer the reader to the pseudo-code provided in Appendix B, available in the online supplemental material.

### 5 EXPERIMENTAL ANALYSIS

#### 5.1 Experimental Settings

**Datasets.** To assess the merits of our proposal, we firstly focus our experiments on well-known and challenging image classification datasets, whose classes undergo a split to form the sequences of tasks:

- Split CIFAR-100 [6], [11], [15] is obtained by splitting the CIFAR-100 dataset [46] into 10 consecutive tasks;
each comprises of 10 classes with 500 and 100 32 × 32 images each for training and testing respectively;  
- Split miniImageNet [15], [47], [48] derives from miniImageNet [49], a 100-class subset of the popular ImageNet dataset, split in 20 classification tasks. Each task presents 84 × 84 RGB images out of 5 disjoint classes: by so doing, we can assess our findings on a more complex problem, in terms of both the number of tasks and input resolution.

In addition, we set images aside and conduct experiments on Action Recognition: to this aim, we introduce Split NTU-60, a sequential classification benchmark built on top of the 3D skeletal data from the NTU-RGB-D dataset [50]. To the best of our knowledge, this is the first Continual Learning experimental setting that targets graph-based action classification. We consider this an interesting complement to traditional settings for a threefold reason: i) it deals with a data type (i.e., skeletal graphs that expand in time) that radically differs from images; ii) it sheds light on the tendency of different deep architectures to incur forgetting – hence, not only the common MLPs and CNNs, but also Graph CNNs (GCNNs) [51], [52]; iii) as it still tackles classification, we can provide novel forgetting measurements that characterize existing and well-established approaches. In our experiments, we split the data of NTU-RGB-D into 6 tasks, each of which contains 10 classes. We refer the reader to Appendix C, available in the online supplemental material for further details regarding this dataset.

Architectures. For Split CIFAR-100, we use ResNet18 [53] as in [11]. For Split miniImageNet we opt instead for EfficientNet-B2 [54], which has recently arisen as a more suitable baseline network that allows better performance with fewer parameters and faster inference: we argue that the resulting considerations are therefore more indicative of the current advances in deep learning. For the same reason, we opt for EfficientGCN-B0 [55] when dealing with Split NTU-60. Further details can be found in Appendix D, available in the online supplemental material.

Training details. All models are trained from scratch (no pre-training has been used). While there are some works [12], [16], [19] that have recently investigated the single-epoch scenario (no more than one pass on training data), we place our experiments in the multi-epoch setup [6], [11], [41]. As argued by our previous work [14], this leads to easier-to-read results, in which the under-fitting linked to few iterations is removed from the equation. In more detail, we always use Stochastic Gradient Descent as optimizer and a number of epochs that varies according to the dataset (50 for CIFAR-100, 80 for miniImageNet and 70 for NTU); in line with [11], [31], [41], we also define a set of epochs at which the learning rate is divided by 10 ([35,45] for CIFAR-100, [35,60,75] for miniImageNet). For NTU, we use a cosine scheduler with a 10-epoch warm-up.  

For all the evaluated approaches, we select their hyperparameters through grid-search. We refer the reader to the Appendix for: i) a full description of the considered parameter combinations, the chosen ones, and other training details (Appendix E.2, available in the online supplemental material); ii) an empirical evaluation of the sensitivity of X-DER to different choices of hyperparameters (Appendix F.1, available in the online supplemental material).

It is worth noting that the experimental settings of different works are often subtly but meaningfully dissimilar, which makes drawing direct comparisons among them difficult. Therefore, we avoid taking results directly from other works and instead re-run all experiments in a common and unified experimental environment (for which we make the code-base available at https://github.com/imagelab/mammoth).

Metrics. We firstly assess the performance in terms of Final Average Accuracy (FAA). Let $a_i^T$ be the model accuracy on the $i$-th task after training on task $T_i$, we define FAA as:

$$\text{FAA} \triangleq \frac{1}{T} \sum_{t=0}^{T-1} a_{t+1}^{T-1},$$

(14)

where $T$ denotes the total number of tasks. FAA represents the most immediate summarizing measure that allows direct comparisons. However, it provides only a snapshot of the state after the last task: to account for what happens during the entire sequence, we follow other works [11], [31], [41] and exploit the Final Forgetting (FF) [56] metric:

$$\text{FF} \triangleq \frac{1}{T-1} \sum_{j=0}^{T-2} f_j, \quad \text{s.t.} \quad f_j = \max_{t=0,...,T-2} a_i^t - a_i^{t-j}.$$

(15)

FF is bound in $[-100, 100]$ and measures the average accuracy degradation, i.e., the maximum discrepancy in performance observed for a given task through training. To complement our analysis of performance, we also examine X-DER both in terms of its sensitivity to hyperparameters, memory footprint and training time. A comparative evaluation w.r.t. SOTA methods can be found in Appendix F.2, available in the online supplemental material.

5.2 Baselines and Competing Methods

To gain a clear understanding, we provide two methods that act as lower- and upper-bound: the former continually fine-tunes on the most recent task with no remedy to catastrophic forgetting (a.k.a. Fine-Tuning, FT); the latter trains a model jointly on all data (a.k.a. Joint-Training, JT).

We use the following CiCL approaches as competitors:

- Learning without Forgetting (LwF) [7] is a regularization approach that exploits Knowledge Distillation: it uses the model learned in the previous task as a teacher during the current one. As done in [11], we use LwF.MC (the CiCL adaptation of LwF);
- Experience Replay (ER) [33], [34] is a simple rehearsal method that stores previously encountered examples
in a memory buffer for later replay. In spite of its simplicity, it still stands as a strong baseline [15, 18]:

- **ER with Asymmetric Cross-Entropy (ER-ACE)** [28] is a recently proposed modification of ER that addresses stream imbalance w.r.t. to the memory buffer by optimizing separate cross-entropy loss terms;
- **Incremental Classifier and Representation Learning (iCaRL)** [11] combines a carefully-designed herding buffer with a nearest mean-of-exemplars classifier; it is often regarded as a strong performer on complex datasets;
- **Bias Correction (BiC)** [41] pairs Experience Replay with a regularization term that resembles the objective of LwF. Most notably, it leverages a separate layer that aims at counteracting bias in the backbone network;
- **Learning a Unified Classifier Incrementally via Rebalancing (LUCIR)** [31] is a rehearsal strategy proposing several modifications that promote separation in feature space and result in a more harmonized incrementally-learned classifier;
- **Greedy Sampler and Dumb learner (GDumb)** [13] is an experimental method meant to question the advances in CL. It totally avoids training steps on data from the current task and just fills up the memory buffer: when an evaluation is required, it then trains a new model on the memory buffer from scratch;
- Our previous proposals: Dark Experience Replay (DER) and its variant that uses also labels (DER++).

### Ablative studies
Besides reviewing the performance of X-DER in light of the state of the art, we provide additional comparisons to validate the design choices of X-DER:

- **X-DER w/o memory update**, which does not update logits through the sequence of tasks;
- **X-DER w/o future heads**, which represents the simplest way to handle new classes: namely, it simply adds a new classification head once the new task is presented;
- **X-DER w/ CE on future heads**, a baseline that devises future heads. In line with what is done by DER(++), it includes them in the computation of the stream-specific portion of the separated cross-entropy loss by targeting them to zero probabilities;
- **X-DER w/ RPC**, an alternative to the semi-supervised strategy devised in Section 4.2. It builds future preparation upon the **Regular Polytope Classifier** proposed in [57]. Briefly, it appoints the parameters of the final classification layer to constant weights values, the latter being designed to keep them equally distributed in output space. This way, the authors meant to ensure that all classes (both seen and unseen) are all at equal distance.

For completeness, we also evaluate the original **Regular Polytope Classifier (RPC)** [57], which complements ER with the above-mentioned fixed final classifier.

### 5.3 Discussion
By examining Table 1, we can make the preliminary consideration that the considered regularization approach (LwF.MC) consistently underperforms compared to online replay-based methods. This observation aligns with [12], [14] and suggests that adopting a replay memory is crucial for achieving solid performance in CiCL.

As it only learns from its memory buffer, the offline training of GDumb allows it to observe few examples from all classes (old and new) jointly, thus avoiding issues related to bias by design. On **Split miniImageNet**, which features a long sequence of tasks, this is sufficient to outperform methods that do not compensate for this effect (e.g., ER, RPC). However, since it entirely discards the remaining data from the input stream, GDumb produces a lower FAA w.r.t. to most online-learning methods.

Among ER-based approaches, ER-ACE stands out as the most effective thanks to its loss, carefully designed to prevent interference between the learning of the current task and the replay of old data. This trait allows to protect
previously acquired knowledge, resulting in lower FF metric.

On average, methods combining rehearsal and distillation achieve better performance w.r.t. simple replay. iCaRL limits forgetting consistently and achieves balanced accuracy on all seen tasks thanks to its nearest-mean-of-exemplars classifier. This is rewarding on the medium-length CIFAR-100 benchmark, but proves sub-optimal on both minImageNet and NTU-60 (due to forgetting on the former and to lack of fitting of the current task on the latter). Differently, LUCIR delivers a high accuracy on the last few encountered tasks, thus proving very effective on the short Split NTU-60 but struggling on longer sequences. While its performance is adequate on CIFAR-100, BIC is characterized by the highest FF on all other benchmarks, leading to an FAA score close to its parent method LwF.MC.

Our previous proposals DER and DER++ classify as strong baselines when combined with a large-enough memory buffer. However, due to the limitations explored in Section 3.3, they occasionally give in to approaches that contrast bias more effectively (ER-ACE, iCaRL, LUCIR for \( M_{\text{size}} \) 500 on Split CIFAR-100; ER-ACE and LUCIR on Split NTU-60).

Compared to the current state-of-the-art, X-DER delivers higher accuracy and lower forgetting across all benchmarks. As one can observe from a close exam of its incremental accuracy values (Fig. 6), the proposed enhancements lead to increased performance retention on past tasks, lifting its score significantly over competitors as training progresses.

To gain a deeper understanding, we further compare X-DER against its ablative baselines. By omitting to update the content of the memory buffer (X-DERw/o memory update), we see a significant drop in performance – especially relevant for smaller \( M_{\text{size}} \). Comparatively, the strategy adopted for preparing future logits is less influential. The proposed contrastive preparation loss of X-DER yields the lowest FF rates, validating our intuition to use past data to prepare future learning. Adopting the theoretically grounded but fixed design of X-DERw/ RPC comes at a steady but non-negligible cost in performance across all benchmarks. X-DER w/ CE future heads shows that dampening future heads by indiscriminately applying CE leads to a further decrease in accuracy; however, even this approach is still preferable to training X-DER without future heads, which is linked to higher FF metrics. This stresses the importance of preparing the model for future classes and suggests that using future heads and replaying their logits can act as a remedy against forgetting.

6 MODEL ANALYSIS

6.1 Towards Better “Continual” Teachers

This section delves into the regularization strategy of X-DER: why are its responses so effective against forgetting of old tasks? We here build upon the seminal work of [24], which has recently proposed a statistical background of Knowledge Distillation (KD) helping researchers and practitioners to gain novel insights on its effectiveness. Essentially, the authors assume that the teacher’s response \( P(y|x) \) constitutes an approximation of the true Bayes class-probability distribution \( P* (y|x) \), which represents the suitability of each class \( y \) for a given \( x \) (hence encoding confusions amongst the labels). With respect to one-hot targets, it is proven that minimizing the risk associated with \( P* \) gives the student an objective with lower variance, which aids generalization. Nevertheless, the true \( P* \) cannot be accessed and an imperfect estimate has to be used (e.g., the response of a teacher net). In that sense, the better the estimation of the true Bayes probabilities, the higher the generalization capabilities of a student learning through the corresponding risk.

In the following, we show how such a novel perspective can help to gain a new understanding of our approach.

Analysis of Secondary Information

In literature, the concept of Bayes class Probabilities has also been studied in terms of secondary information [29], [58] i.e., for each non-maximum score, the model’s belief about the semantic cues of the corresponding class within the input image. Unsurprisingly, Yang et al. in [58] identify the preservation of secondary information as a key property of KD: they empirically find that teachers with richer secondary information lead to students that generalize better. However – when dealing with catastrophic forgetting – it can be challenging to capture rich secondary information, as the latter becomes available only as tasks progress.

Seeking to measure how effectively distinct CL approaches can leverage secondary information, we follow the setup proposed in [29] and re-examine their performance on the test-set of CIFAR-100 but labeled in a different way: we indeed group its 100 classes into their natural 20 super-classes. According to the authors of [29], a model achieving a high classification score in this setup also retains better secondary information, as classes belonging to the same super-class can be assumed to have higher visual similarity than the ones belonging to different super-classes.

The retained secondary information can be quantified by two metrics [29]: on the one hand, the Secondary-Superclass
**TABLE 2**

| KD      | SS-ERR [1] | SS-NLL [1] |
|---------|------------|------------|
| M$_{size}$ | 500 2000   | 500 2000   |
| ER      | ×| 0.71 0.68 4.00 4.32 |
| ER-ACE  | ×| 0.63 0.60 2.64 2.81 |
| DER     | ✓* | 0.67 0.64 2.22 2.21 |
| DER++   | ✓* | 0.67 0.64 2.22 2.25 |
| LUCIR   | ✓ | 0.60 0.59 2.21 2.22 |
| iCaRL   | ✓ | 0.60 0.60 1.90 1.94 |
| X-DER w/o memory update | ✓* | 0.64 0.61 2.14 2.10 |
| X-DER   | ✓ | 0.57 0.56 1.83 1.82 |

* indicates no use of Knowledge Distillation (KD) while training, ✓* indicates KD of past logits only, ✓ indicates KD of all logits (incl. future past).

Error (SS – ERR) equals 1 minus the probability of predicting the right super-class. As the focus lies on secondary information, the maximum logit is always omitted during softmax computation. Otherwise, the Secondary-Superclass NLL (SS – NLL) considers the negative log-likelihood when using super-classes as labels.

From the results in Table 2, X-DER, iCaRL and LUCIR consistently end up predicting the correct coarse classes (lower SS-ERR) and do so more confidently (lower SS-NLL). This is in line with our expectations: as these methods handle hindsight-search similarity between newly discovered classes and old ones, the corresponding teaching signal leads the student toward richer secondary information. In contrast, DER(++) yield lower metrics due to: i) their distillation targets neglecting logits of future past, ii) the existence of a large bias towards the last seen classes. To verify the importance of i), we also run this evaluation on a version of X-DER that does not update its buffer logits; this results in metrics comparable to DER(++).

| Split CIFAR-100 |
|----------------|
| ![Accuracy Chart](chart.png) |

| Split minImageNet |
|-------------------|
| ![Accuracy Chart](chart.png) |

**Fig. 7. Accuracy of models trained from scratch on memory buffers of ER (Labels), DER (Logits, Both), and X-DER (Logits, Both). The resulting accuracy represents a proxy of the informativeness of the memory buffer.

**Calibration of Continual Learners: A New Perspective**

Although [24] presents an appealing framework, it is still up for debate how to assess the quality of the approximations of labels. Menon et al. suggest that a coarse evaluation can be carried out by looking at the Expected Calibration Error (ECE) [59]. Remarkably, this provides a new light and foundation to the experiment conducted in our previous work [14]: in fact, we already compared several replay strategies in terms of the induced ECE, which led us to ascribe the gains of DER(++) also to the higher calibration of the underlying network.

On these premises, we here repeat the above-mentioned evaluation on top of our new proposal, X-DER. Fig. 8 shows the results obtained on Split CIFAR-100: as can be seen, X-DER delivers a lower ECE compared to other approaches. If this finding could seem trivial when using as baseline one-hot teachers such as ER, this holds as well for smoothed ones such as DER(++) in light of the previous considerations, we can now link the advantages of X-DER to a better estimation of the underlying Bayes class-probabilities.

**6.2 Effect of Future Preparation on Unseen Classes**

One of our main contributions regards the design of a pretext task to warm up unused heads; this way, we expect a gentler adaption of the network to unseen data distributions, thus removing the need for dramatic updates to its parameters and therefore lowering the risk of forgetting. To verify whether this happens, we envision a setting where very few data of the incoming tasks are available: here, how well does the feature space spanned by future heads work? Does the technique introduced in Section 4.2 give the network any advantage when dealing with unseen data?

We conduct an in-depth investigation of these facets and let Fig. 9 provide a graphical summary of it. For multiple
methods (such as ER, DER(++) and our X-DER), we firstly focus on a single snapshot and stop their training procedure right after the 6th task of CIFAR-100 (see Fig. 9a). We then aim at measuring their performance on each of the remaining four tasks separately: we do that by fitting a Nearest Neighbor (NN) classifier on top of the activations given by the corresponding future head. To draw a clear picture of the forward transfer delivered by those methods: i) we repeat this evaluation at varying training set sizes (ranging from one to fifty shots per class); ii) we freeze network parameters; so no fine-tuning steps are performed on data-points of the incoming task. As can be seen, X-DER is the method delivering the best results among other approaches.

Afterwards, we propose a more comprehensive evaluation, which takes into consideration the transfer from all encountered task (and not the 6th task solely): here, we aim to assess how the capabilities linked to forward transfer evolve one task after the other. To do so – for each observed task $t$ ranging from the first to the penultimate one – we firstly define the performance curves $\text{NN}_{t-1}(k)$ over the unseen tasks $\sim_t > t$, where $k$ indicates the number of shots per class; to gain a clear understanding, Fig. 9a focuses on $t = 6$ and depicts $\text{NN}_{6-1}(k) \forall t \in \{7, 8, 9, 10\}$. Subsequently, we summarize each curve with the Area Under the Curve $\text{AUC}_{t-1}$ and finally average the latter across $\sim_t$ (e.g., $\text{AUC}_{6-1} \triangleq \frac{1}{10} \sum_{t=7}^{10} \text{AUC}_{6-1}$), thus providing an overall measure of generalization to all tasks that will be encountered onwards.

In this respect, Fig. 9b reports the trend of the AUC$_t$ for different rehearsal methods. As can be appreciated, we do not observe a clear distinction in their performance when focusing on the earliest tasks. Instead, the AUC curve of X-DER widens the gap as the number of seen tasks increases (it scales better to the number of seen tasks). This suggests that the more and diverse the data modalities present in the memory buffer, the higher the chances that optimizing Eq. (9) will lead to a good forward transfer to unseen data.

Pre-Allocation of Future Tasks
We have so far supposed (see Section 4.1) that the overall number of tasks $T$ can be known in advance. This allows us to instantiate a last fully-connected layer large enough to accommodate the logits for all seen and unseen classes. However, in practical and real scenarios, we may not know how many tasks will be encountered from the outset: hence, it can be brought into question whether our approach can still be applied to those settings.

We here discuss a straightforward modification that enables the number of future tasks to be unknown. We initially set up the last layer to expose $\sim_t$ classification heads: precisely, the one dedicated to the first task and the remaining $\sim_t$ ones to future tasks. In addition, we instantiate a new head at the end of each task, thus guaranteeing that one head (at least) is always available to the incoming task.

Fig. 10 depicts how such a modification affects performance and the impact of the hyperparameter $\sim_t$ controlling the number of pre-allocated heads. We draw the following conclusions: i) given the slight gap in performance between X-DER and the proposed variant, the overall number of tasks does not seem an essential information for achieving good results; ii) a higher number of pre-allocated heads positively influences the final average accuracy. This latter finding suggests that future logits also play a role against forgetting: we conjecture that rehearsal of non-coding logits might represent an additional guard against forgetting, as they still embody a reflection of past neural activities.
6.3 On the Geometry of the Local Minimum
The Effectiveness of Flat Minima in Continual Learning

We investigated in [14] the relation between the nature of the attained local minima and the generalization capabilities linked to them. We indeed conjectured that flatness around a loss minimizer represents a remarkable property for CL settings: intuitively, a loss region tolerant towards local displacements favors later optimization trajectories that entail a less severe drop in performance for old tasks. As a proof of concept, we used two common metrics (recalled later in this section) to characterize the geometry of the minima: DER(++) – the methods that performed best on the benchmarks of [14] – also exhibit favorably flatter minima.

However, we consider such a matter still nebulous and worthy of further discussion. In this respect, the authors of [25] assessed the impact of different training regimes on forgetting and stated that the latter strongly correlates with the curvature of the loss function around the minimum of each task. In practice, they made use of some strategies known to affect the width of the minima (e.g., higher initial learning rates, dropout, small batch sizes, etc.) and observed that these lead to a stable regime that further mitigates forgetting. Similar arguments have been raised as well in [60].

In this work, we contribute again to this topic with a more targeted evaluation: given a sequence of two tasks, we deliberately drive the optimization towards a wider minimum during the former (stable regime). Differently from [25], we directly make sure the network reaches a wide minimum by introducing a tailored term in the loss function. In this regard, we evaluate two distinct approaches:

- **Local Flatness Regularizer (LFR)** [61], which seeks to minimize the $\| \mathbf{g} \|$ norm of the loss gradients w.r.t. a malign example, the latter forged so that: i) it lies in a $\varepsilon$-neighborhood centered on a given (benign) example; ii) it maximizes the norm of the gradients. The authors prove that the robustness towards such kind of attack favorably relates to the flatness of the loss surface.

- **Local Linearity Regularizer (LLR)** [62], which promotes loss smoothness around the local neighborhood. As before, it consists of a regularization term that depends upon adversarial examples: on these inputs – supposing a smooth and approximately linear loss surface – the first-order Taylor expansion represents a good approximation of the value of the

---

Fig. 10. On Split CIFAR-100 and Split min/ImageNet, analysis evaluating how the number of heads pre-allocated in preparation of future tasks affects the final average accuracy.

Fig. 11. For nine sequences of two tasks of CIFAR-100, the final accuracy on the test-set of the former task. LFR and LLR, which encourage a stable training regime, lead to higher retention of performance.

Fig. 12. Analysis of minima attained by distinct approaches: (a) for increasing $\sigma$, the value assumed by the loss function when Gaussian noise is applied to network weights – tolerance towards noise is a sign of a flatter solution; (b) the sum of the eigenvalues of the Fisher Information, which models the curvature of the loss function around the solution.
We hence train the network on one task (pairing cross-entropy loss with loss surface regularization) and then measure the forgetting entailed by a CL method – for the sake of simplicity, Experience Replay – at the end of the second task. As a baseline, we consider the results achieved when neglecting regularization during the former task (plastic regime): namely, it corresponds to optimizing with plain Stochastic Gradient Descent (SGD) followed by ER during the following task. We conduct this evaluation on top of nine pairs of adjacent tasks of CIFAR-100. Fig. 11 highlights the effect of the regularization imposed by LFR and LLR on forgetting: the results delivered by SGD (grayed out) are always upper-bounded by its regularized counterparts.

We consider this as additional evidence showing the benefits of a stable training regime in CL settings. It corroborates the intuition behind the effectiveness of those approaches based on self-distillation (e.g., iCaRL, LwF, DER(++), etc.) – which are known [38], [63] to lead to such a regime.

Measuring the Flatness

Based on the above, we provide two quantitative evaluations illustrating the stability and flatness of the optima we observe for X-DERand other CL approaches.

First, we measure how weight perturbations affect the expected loss \( \mathcal{L}_{excl} \) w.r.t. to the training set [64], [65]:

\[
\mathcal{L}_{\sigma} = \frac{1}{T} \sum_{i=0}^{T-1} \mathbb{E}_{(x,y) \sim T_i} \left[ \mathcal{L}(f(x; \hat{\theta}), y) \right];
\]

specifically, we follow the hints of [64], [66] and weigh the perturbation according to the magnitude of parameters (\( \sigma_i = \alpha |\delta_i| \)), thus preventing degenerate solutions [64]. With reference to Fig. 12a, it can be seen that logit-replay based models such as DER(++) and X-DER consistently preserve a lower value for Eq. (16). Among them, X-DER exhibits a higher tolerance to perturbations especially in the high-\( \sigma \) regime, which suggests that its attained minima are overall harder to disrupt when compared to the other methods.

A complementary flatness measure [65], [67], [68] examines the eigenvalues of the Hessian of the overall loss function \( \nabla^2_{\theta} \mathcal{L}_{excl} \). While the latter is intractable, it can be approximated by computing the empirical Fisher Information Matrix on the training set [5], [67]:

\[
F = \frac{1}{T} \sum_{i=0}^{T-1} \mathbb{E}_{(x,y) \sim T_i} \left[ \nabla_{\theta} \mathcal{L}(f(x; \theta), y) \nabla_{\theta} \mathcal{L}(f(x; \theta), y)^T \right].
\]

As in [14], we estimate the sum of the eigenvalues of \( F \) through the trace of the matrix \( \text{Tr}(F) \), reported in Fig. 12b. Even according to this metric, DER(++) and X-DER reach flatter minima w.r.t. other approaches. Remarkably, X-DER produces lower \( \text{Tr}(F) \) values, suggesting that its improved accuracy can be linked to the local geometry of the loss.

6.4 Model Explanation

The goal of this subsection is to explore what lies behind the last layer of the network. In an attempt to investigate the quality and meaning of learned representations, we put the emphasis on three approaches (i.e., ER, DER, and X-DER) and evaluate the effect of these regularization strategies on the explanations provided by the corresponding models.

Analysis of Model Explanations for Primary Targets

Inspired by the investigation carried out in [69], we here aim to assess the quality of the visual concepts encoded in the intermediate layers of the network. More precisely, we are interested in assessing whether the use of Knowledge...
Distillation leads to more refined visual concepts in regimes of catastrophic forgetting.

As stated by the authors of [69], a precise and well-established definition of visual concepts as well as the ways these can be quantified remain elusive matters. In this regard, we first assess the acquisition of task-relevant information (i.e., what concerns the main subject of the image): considering the true class, does the model ascribe its score to the expected spatial location? How much does its explanation overlap with the foreground region?

To answer these questions, we take into consideration the evaluation protocol proposed in [70], termed pointing game, which was conceived to characterize the spatial selectiveness of a saliency map in the localization of target objects. The procedure is as follows: given the explanation map yielded by the learner, we check whether the point with the maximum score falls into the object region (usually defined through annotated segmentation maps); in the positive case, we have a hit (a miss otherwise). The localization capabilities can be finally quantified by the average Pointing Accuracy:

\[
\text{PA} = \frac{1}{T \cdot |\mathcal{Y}|} \sum_{y=0}^{T \cdot |\mathcal{Y}|-1} \frac{\#\text{hit}_y}{\#\text{hit}_y + \#\text{miss}_y}.
\]

Since the datasets considered in the previous section do not come with segmentation maps, we move to \textbf{Split CUB-200} [71], [72], which consists of photos depicting 200 bird species that are split into 10 disjoint tasks (for the sake of brevity, we provide the technical details in Appendix G.1, available in the online supplemental material). On top of that, we extract explanation maps through the Grad-CAM algorithm [73] and use them to compute the resulting pointing accuracy, which is reported in Fig. 13 along with the average classification accuracy. Compared to other approaches, X-DERIs less prone to forgetting the reasoning behind its predictions, as also highlighted by some qualitative examples shown in Fig. 14.

**Analysis of Model Explanations for Secondary Targets**

We previously investigated the model’s ability to impute its prediction to the right evidence, thus evaluating how much its representation disentangles the main subject of the image from background clutter. However, there is nothing that prevents a similar analysis on top of the secondary objects that may appear in the input image: in that sense, is secondary information limited to the main target or does it also encode insights regarding minor targets? If so, is the model able to locate these objects? If a model can correctly encode the presence of multiple objects in its response, we also expect its activation maps to convey meaningful information for their purpose of localization.

With the aim of investigating these facets, we present a novel tailored procedure that starts with pre-training on Split CIFAR-100. We then leverage a synthetic dataset obtained by stitching small image patches from COCO 2017 [74] on examples of CIFAR-100. As shown in Fig. 15, these patches are cut through ground-truth segmentation masks and pasted on CIFAR-100 images to simulate secondary semantic content. Finally, we exploit the \textbf{linear evaluation} protocol [43] to assess the representation quality of secondary targets: the parameters of the network are frozen and only a linear classifier is trained on top of its features.

We compare the performance of several methods in terms of $F_1$ score and pointing accuracy for the \textit{stitched} secondary targets. As confirmed by the results shown in Fig. 16, the approaches relying on Knowledge Distillation perform better according to both considered metrics. Notably, X-DER stands out, thus providing a further confirmation that it can retain richer secondary information.

**7 Conclusion**

This paper reviewed Dark Experience Replay [14], our previously proposed Continual Learning method combining rehearsal and Knowledge Distillation. Upon a preliminary examination, we showed it discards valuable information about the semantic relations between old and novel classes. We also found that it suffers from a classification bias, overemphasizing the most recently acquired knowledge.

We then proposed \textbf{eXtended-DER} (a.k.a. X-DER), which introduces multiple innovations (e.g., memory content editing) addressing these issues: through experiments across multiple datasets, we showed that X-DER delivers higher performance and outperforms the current state of the art. Further, we presented a comprehensive analysis that goes beyond the mere final accuracy and provides an all-round validation. We in fact offered several explanations of its effectiveness against forgetting (e.g., the knowledge inherent its memory buffer, the geometry of minima, the high retention of secondary information, etc.). Moreover, our results indicate that our future preparation technique favorably arranges the model to classes that are yet to be seen.

We finally envision several directions for future works: we strongly believe that overcoming the standard schema (embodied in the stability versus plasticity dilemma) with the guess of incoming tasks can favorably foster new ideas and

2. We limit to patches belonging to the set of classes shared between CIFAR-100 and COCO 2017. Moreover, we facilitate the stitching by using a 2x-upscaled version of CIFAR-100 obtained through the CAI super-resolution API [75]. Additional details are reported in Appendix G, available in the online supplemental material.
advances in the field. Due to its potential applicability to a variety of CL approaches, we feel there is room for the proposals of novel strategies for mimicking future data distributions, which will be the scope of our future works.

ACKNOWLEDGMENTS

The authors would like to thank Silvia Ciscanelli for the constructive feedback she provided about the editing and revision of the paper.

REFERENCES

[1] A. Cochrane, V. Simmering, and C. S. Green, “Fluid intelligence is related to capacity in memory as well as attention: Evidence from middle childhood and adulthood,” PLoS One, vol. 14, 2019, Art. no. e0221353.

[2] M. McCloskey and N. J. Cohen, “Catastrophic interference in connectionist networks: The sequential learning problem,” Psychol. Learn. Motiv., vol. 24, pp. 109–165, 1989.

[3] M. Delange et al., “A continual learning survey: Defying forgetting in classification tasks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 7, pp. 1366–1385, 2012.

[4] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, “Continual lifelong learning with neural networks: A review,” Neural Netw., vol. 113, pp. 54–71, 2019.

[5] J. Kirkpatrick et al., “Overcoming catastrophic forgetting in neural networks,” Proc. Nat. Acad. Sci. USA, vol. 114, pp. 3521–3526, 2017.

[6] F. Zenke, B. Poole, and S. Ganguli, “Continual learning through synaptic intelligence,” in Proc. Int. Conf. Mach. Learn., 2017, pp. 3987–3995.

[7] Z. Li and D. Hoiem, “Learning without forgetting,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 12, pp. 2935–2947, Dec. 2018.

[8] A. A. Rusu et al., “Progressive neural networks,” 2016, arXiv:1606.04671.

[9] A. Malhava and S. Lazebnik, “Packnet: Adding multiple tasks to a single network by iterative pruning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 7765–7773.

[10] R. M. French, “Using semi-distributed representations to overcome catastrophic forgetting in connectionist networks,” in Proc. Annu. Conf. Cogn. Sci. Soc., 1991, pp. 173–178.

[11] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, “iCaRL: Incremental classifier and representation learning,” in Proc. Annu. Conf. Cogn. Sci. Soc., 2017, pp. 2001–2010.

[12] R. Aljundi, M. Lin, B. Goujaud, and Y. Bengio, “Gradient based constructive feedback she provided about the editing and revision of the paper.”

[13] P. Buzzega, M. Boschini, A. Porrello, and S. Calderara, “Dark

[14] P. Buzzega, M. Boschini, A. Porrello, D. Abati, and S. Calderara, “Dark

[15] A. Prabhu, P. H. Torr, and P. K. Dokania, “GDumb: A simple approach that questions our progress in continual learning,” in Proc. Adv. Neural Inf. Process. Syst., 2019, pp. 374–382.

[16] D. Lopez-Paz and M. Ranzato, “Gradient episodic memory for continual learning,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 6470–6479.

[17] D. J. Bridge and J. L. Voss, “Hippocampal binding of novel information with dominant memory traces can support both memory stability and change,” J. Neurosci., vol. 34, pp. 2203–2213, 2014.

[18] D. R. Addis, A. T. Wong, and D. L. Schacter, “Remembering the past and imagining the future: Common and distinct neural substrates during event construction and elaboration,” Neuropsychology, vol. 45, pp. 1363–1377, 2007.

[19] D. L. Schacter, D. R. Addis, and R. L. Buckner, “Remembering the past to imagine the future: The prospective brain,” Nat. Rev. Neurosci., vol. 8, pp. 657–661, 2007.

[20] A. K. Menon, A. S. Rawat, S. Reddi, S. Kim, and S. Kumar, “A statistical perspective on distillation,” in Proc. Int. Conf. Mach. Learn., 2021, pp. 7632–7642.

[21] S. I. Mirzadeh, M. Farajtabar, R. Pascanu, and H. Ghazemzadeh, “Understanding the role of training regimes in continual learning,” in Proc. Adv. Neural Inf. Process. Syst., 2020, pp. 7308–7320.

[22] G. M. van de Ven and A. S. Tolias, “Three continual learning scenarios,” in Proc. Neural Informat. Process. Syst. Workshops, 2018.

[23] J. Serra, D. Suris, M. Miron, and A. Karatzoglou, “Overcoming catastrophic forgetting with hard attention to the task,” in Proc. Int. Conf. Mach. Learn., 2018, pp. 4548–4557.

[24] L. Caccia, R. Aljundi, T. Tuytelaars, J. Pineau, and E. Belilovsky, “New insights on reducing abrupt representation change in online continual learning,” in Proc. Int. Conf. Learn. Representations, 2022.

[25] S. Mittal, S. Galesso, and T. Brox, “Essentials for class incremental learning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 3513–3522.

[26] E. Belouadah and A. Popescu, “II2m: Class incremental learning with dual memory,” in Proc. IEEE Conf. Comput. Vis., 2019, pp. 383–392.

[27] S. Hou, X. Pan, C. C. Loy, Z. Wang, and D. Lin, “Learning a unified classifier incrementally via rebalancing,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 831–839.

[28] S. Farquhar and Y. Gal, “Towards robust evaluations of continual learning,” in Proc. Int. Conf. Mach. Learn. Workshops, 2018.

[29] R. Ratcliff, “Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions,” Psychol. Rev., vol. 97, 1990, Art. no. 285.

[30] A. Robies, “Catastrophic forgetting, rehearsal and pseudorehearsal,” Conn. Sci., vol. 7, pp. 123–146, 1995.

[31] A. Chaudhry, M. Ranzato, M. Rohrbach, and M. Elhoseiny, “Efficient lifelong learning with A-GEN,” in Proc. Int. Conf. Learn. Representations, 2019.

[32] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” in Proc. Neural Inform. Process. Syst. Workshops, 2015.

[33] T. Furlanello, C. Z. Lipton, M. Tschannen, L. Itti, and A. Anandkumar, “Born again neural networks,” in Proc. Int. Conf. Mach. Learn., 2018, pp. 1607–1616.

[34] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, and K. Ma, “Be your own teacher: Improve the performance of convolutional neural networks via self-distillation,” in Proc. IEEE Int. Conf. Comput. Inf., 2019, pp. 3713–3722.

[35] Y. Luan, H. Zhao, Z. Yang, and Y. Dai, “MSD: Multi-self-distillation learning via multi-classifiers within deep neural networks,” 2019, arXiv:1911.09418.

[36] F. M. Castro, M. J. Marin-Jiménez, N. Guil, C. Schmid, and K. Alahari, “End-to-end incremental learning,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 233–248.

[37] Y. Wu et al., “Large scale incremental learning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 374–382.

[38] H. Ahn, J. Kwak, J. Song, A. Gao, J. Chen, X. He, and K. Ma, “A new learning algorithm: Improving the performance of convolutional neural networks via self-distillation,” in Proc. IEEE Int. Conf. Comput. Inf., 2019, pp. 843–853.

[39] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in Proc. Int. Conf. Mach. Learn., 2020, pp. 1597–1607.

[40] J. Zhontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, “Barlow twins: Self-supervised learning via redundancy reduction,” in Proc. Int. Conf. Mach. Learn., 2021, pp. 12310–12320.

[41] P. Khosla et al., “Supervised contrastive learning,” in Proc. Adv. Neural Inf. Process. Syst., 2020, pp. 18661–18673.

[42] A. Krizhevsky and G. Hinton, “Learning multiple layers of features from tiny images,” Univ. Toronto, Tech. Rep., vol. 1, no. 4, p. 7, 2009, doi:10.11.222.9220.

[43] S. Ebrahimi, F. Meier, R. Calandra, T. Darrell, and M. Rohrbach, “Adversarial continual learning,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 386–402.

[44] M. M. Derakhshani, X. Zhen, L. Shao, and C. Snoek, “Kernel continual learning,” in Proc. Int. Conf. Mach. Learn., 2021, pp. 2621–2631.
For more information on this or any other computing topic, please visit our Digital Library at https://www.computer.org/csdl.