Pattern Recognition Letters 184 (2024) 119–125

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Latent spectral regularization for continual learning

Emanuele Frascaroli [a,b,*], Riccardo Benaglia [a,b], Matteo Boschini [a], Luca Moschella [c], Cosimo Fiorini [a,b], Emanuele Rodolà [c], Simone Calderara [a]

[a] University of Modena and Reggio Emilia, Via Vivarelli 10, Modena, 41125, Italy
[b] Ammagamma, Via S.Orsola 37, Modena, 41121, Italy
[c] Sapienza University of Rome, Piazzale Aldo Moro 5, Rome, 00185, Italy

ARTICLE INFO

Editor: Sudeep Sarkar

Keywords:
Continual learning
Deep learning
Regularization
Spectral geometry
Incremental learning

ABSTRACT

While biological intelligence grows organically as new knowledge is gathered throughout life, Artificial Neural Networks forget catastrophically whenever they face a changing training data distribution. Rehearsal-based Continual Learning (CL) approaches have been established as a versatile and reliable solution to overcome this limitation; however, sudden input disruptions and memory constraints are known to alter the consistency of their predictions. We study this phenomenon by investigating the geometric characteristics of the learner's latent space and find that replayed data points of different classes increasingly mix up, interfering with classification. Hence, we propose a geometric regularizer that enforces weak requirements on the Laplacian spectrum of the latent space, promoting a partitioning behavior. Our proposal, called Continual Spectral Regularizer for Incremental Learning (CaSpeR-IL), can be easily combined with any rehearsal-based CL approach and improves the performance of SOTA methods on standard benchmarks.

1. Introduction

Intelligent creatures in the natural world continually learn to adapt their behavior to changing external conditions by seamlessly blending novel notions with previous understanding into a cohesive body of knowledge. In contrast, artificial neural networks (ANNs) greedily fit the data they are currently trained on, swiftly deteriorating previously acquired information, a phenomenon known as *catastrophic forgetting* [1]. Continual Learning (CL) is a branch of machine learning that designs approaches to help deep models retain previous knowledge while training on new data [2]. These methods are evaluated by dividing a classification dataset into disjoint subsets of classes, called *tasks*, letting the model fit one task at a time and evaluating it on all previously seen data [3]. Recent literature favors the employment of *rehearsal methods*; namely, CL approaches that retain a small memory buffer of samples encountered in previous tasks and interleave them with current training data [4,5]. While rehearsal easily allows the learner to keep track of the joint distribution of all classes seen so far, the limited buffer size produces various overfitting issues that constitute the focus of many recent works (e.g., divergent gradients for new classes [6,7], deteriorating decision surface [8], accumulation of predictive bias for current classes [9,10]).

This paper instead focuses on the changes occurring in the model's latent space as tasks progress. We observe that the learner struggles to separate latent projections of replay examples belonging to different classes, making the downstream classifier prone to interference whenever the input distribution changes and representations are perturbed. Given the Riemannian nature of the latent space of DNNs [11], we naturally revert to spectral geometry to study its evolution. Consequently, we introduce a loss term to endow the model's latent space with a cohesive structure without constraining the individual coordinates. As illustrated in Fig. 1, our proposed approach, called **C**ontinu**a**l **Spe**ctral **R**egularizer for **I**ncremental **L**earning (**CaSpeR-IL**), leverages graph-spectral theory to promote well-separated latent embeddings and can be seamlessly combined with any rehearsal-based CL method to improve its accuracy and robustness against forgetting.

In summary, we make the following contributions: (*i*) we study interference in rehearsal CL models by investigating the geometry of their latent space; (*ii*) we propose CaSpeR-IL: a simple geometrically motivated loss term, inducing the continual learner to produce well-organized latent embeddings; (*iii*) we validate our proposal by combining it with several SOTA rehearsal-based CL approaches, showing that CaSpeR-IL is effective across a wide range of evaluations;

* Corresponding author at: University of Modena and Reggio Emilia, Via Vivarelli 10, Modena, 41125, Italy.
E-mail addresses: emanuele.frascaroli@unimore.it (E. Frascaroli), riccardo.benaglia@unimore.it (R. Benaglia), matteoboschini3@gmail.com (M. Boschini), moschella@di.uniroma1.it (L. Moschella), c.fiorini@ammagamma.com (C. Fiorini), rodola@di.uniroma1.it (E. Rodolà), simone.calderara@unimore.it (S. Calderara).

https://doi.org/10.1016/j.patrec.2024.06.020
Received 10 August 2023; Received in revised form 30 May 2024; Accepted 16 June 2024
Available online 22 June 2024

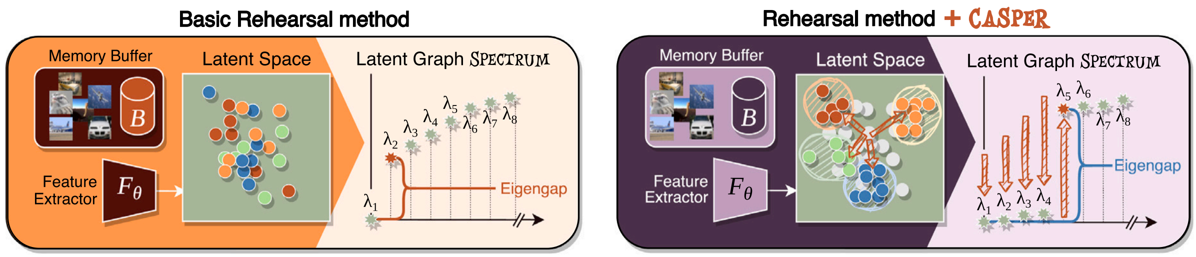


Fig. 1. An overview of the proposed CaSpeR-IL regularizer. Rehearsal-based CL methods struggle to separate the latent-space projections of replay data points. Our proposal acts on the spectrum of the latent geometry graph to induce a partitioning behavior by maximizing the *eigengap* for the number of seen classes.

(iv) we compare our work against recent contrastive-based incremental strategies, showing that CaSpeR-IL better synergizes with CL models; (v) finally, we present additional studies further investigating the geometric properties conferred by our method on the model's latent space. The code to reproduce our experiments is available at https://github.com/aimagelab/CaSpeR.

2. Related work

2.1. Continual learning

Continual Learning [2] approaches help deep learning models minimize *catastrophic forgetting* when learning on changing input distributions. There are different classes of solutions: *architectural methods* allocate separate portions of the model to separate tasks [12], *regularization methods* use a loss term to prevent changes in the model's structure or response [13] and *rehearsal methods* use a working memory buffer to store and replay data-points [4] The latter class of approaches is currently the focus of research efforts due to their versatility and effectiveness [14]. Recent trends aim to improve the basic Experience Replay (ER) formula through better memory sampling strategies [14], combining replay with other optimization techniques [15] or providing richer replay signals [5].

A prominent challenge for enhancing *rehearsal methods* is the imbalance between stream and replay data. This can cause a continually learned classifier to struggle to produce unified predictions and be biased towards recently learned classes [9]. Researchers have proposed solutions such as architectural modifications of the model [16], alterations to the learning objective of the final classifier [6], or the use of representation learning instead of cross-entropy [17]. Our proposal similarly reduces the intrinsic bias of *rehearsal methods*; it does so by enforcing a desirable property on the model's latent space through a geometrically motivated regularization term that can be combined with any existing replay method.

A strain of recent CL approaches similarly conditions the model's representation to facilitate clustering by means of a contrastive regularization objective. SCR [18] enforces consistency between two views of the input batch by leveraging the Supervised Contrastive loss [19]; PRD [20] employs the same loss, in aid to a prototype-based classifier; differently, CSCCT [21] pairs an explicit latent-space clustering objective with a controlled transfer objective preventing negative transfer from dissimilar classes. In our experimental section, we compare our proposal against representatives of this family of methods. This allows us to make some interesting observations on how distinct formulations of a similar clustering objective lead to the emergence of different characteristics in latent space geometry.

The use of pre-trained models [22,23], also exploiting transformers architectures [24], has been showing increasing popularity in the recent CL literature. We leave the testing of CaSpeR-IL on those settings for future work, and evaluate our proposal in the more common "train from scratch" scenario.

2.2. Spectral geometry

Our approach is built upon the eigendecomposition of the Laplace operator on a graph, falling within the broader area of spectral graph theory. In particular, ours can be regarded as an *inverse* spectral technique, as we prescribe the general behavior of some eigenvalues and seek a graph whose Laplacian spectrum matches this behavior. In the geometry processing area, such approaches take the name of *isospectralization* techniques and have been recently used in diverse applications such as deformable shape matching [25], shape exploration and reconstruction [26], shape modeling [27] and adversarial attacks on shapes [28]. Differently from these approaches, we work on a single graph (as opposed to pairs of 3D meshes) and our formulation does not take an input spectrum as a target to be matched precisely. Instead, we require the gap between nearby eigenvalues to be maximized, regardless of its exact value. Since our graph represents a discretization of the latent space of a CL model, this simple regularization has important consequences on its learning process.

3. Method

3.1. Continual learning setting

In CL, a learning model F_θ is incrementally exposed to a stream of tasks τ_i, with $i \in \{1, 2, \dots, T\}$. The parameters θ include both the weights of the feature extractor and the classifier, θ^f and θ^c respectively. Each task consists of a sequence of images and their corresponding labels $\tau_i = \{(x^i_1, y^i_1), (x^i_2, y^i_2), \dots, (x^i_n, y^i_n)\}$ and does not contain data belonging to classes already seen in previous tasks, so $Y^i \cap Y^j = \emptyset$, with $i \neq j$ and $Y^i = \{y^i_k\}^n_{k=1}$. At each step i, the model cannot freely access data from previous tasks and is optimized by minimizing a loss function ℓ_{stream} over the current set of examples:

$$\theta^{(i)} = \underset{\theta}{\operatorname{argmin}}\, \ell_{\text{stream}} = \underset{\theta}{\operatorname{argmin}} \sum_{j=1}^{n} \ell\left(F_\theta(x^i_j), y^i_j\right), \tag{1}$$

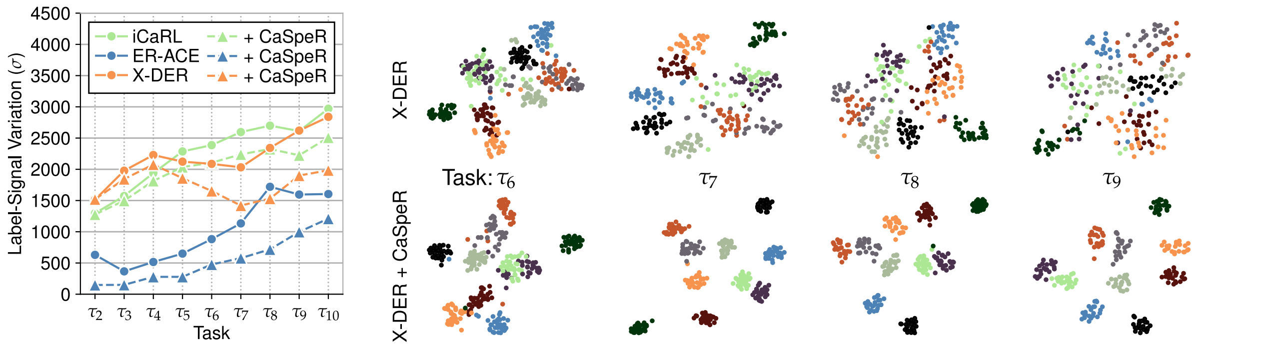
where the parameters are initialized with the ones obtained after training on the previous task $\theta^{(i-1)}$. If no mechanism is put in place to prevent forgetting, the accuracy on previous tasks will collapse while learning task τ_i [1]. Rehearsal-based CL methods store a pool of examples from previous tasks in a buffer B with fixed size m. This data is then used by the model to compute an additional loss term ℓ_{b} aimed at contrasting catastrophic forgetting:

$$\theta^{(i)} = \underset{\theta}{\operatorname{argmin}}\, \ell_{\text{stream}} + \ell_{\text{b}}. \tag{2}$$

For instance, Experience Replay (ER) simply employs a cross-entropy loss over a batch of examples from B:

$$\ell_{\text{er}} \triangleq \text{CrossEntropy}\left(F_\theta(\mathbf{x}^b), \mathbf{y}^b\right). \tag{3}$$

There exist different strategies for sampling the task data points to fill the buffer. These will be explained in the supplemental material, along with details on the ℓ_{b} employed by each baseline of our experiments.



(a) (b)

Fig. 2. How CL alters a model's latent space. (a) A quantitative evaluation measured as Label-Signal Variation (σ) within the LGG for buffer data points – *lower is better*; (b) TSNE embedding of the features computed by X-DER for buffered examples in later tasks (top). Interference between classes is visibly reduced if CaSpeR-IL is applied (bottom). All experiments are carried out on Split CIFAR-100, (a) uses buffer size 500, (b) uses 2000.

3.2. Analysis of changing latent space geometry

We are particularly interested in how the latent space changes when introducing a novel task on the input stream. For this reason, we compute the graph $\mathcal{G}$ over the latent-space projection of the replay examples gathered by the CL model after training on τ_i ($i \in \{2, \ldots, T\}$).[1] In order to measure the sparsity of the latent space w.r.t. classes representations, we compute the Label-Signal Variation σ [29] on the adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{m \times m}$ of $\mathcal{G}$:

$$\sigma \triangleq \sum_{i=1}^{m} \sum_{j=1}^{m} \mathbb{1}_{y_i^b \neq y_j^b} a_{i,j} \, , \tag{4}$$

where $\mathbb{1}_{\cdot}$ is the indicator function. In Fig. 2a, we evaluate several rehearsal CL methods and show they exhibit a steadily growing σ: examples from distinct classes become more entangled in later tasks. This effect can also be observed qualitatively by considering a TSNE embedding of the points in B (shown in Fig. 2b), in which the distances between different-class examples decrease in later tasks. Both evaluations improve when applying our regularizer to the evaluated methods.

3.3. CaSpeR-IL: Continual Spectral Regularizer for Incremental Learning

Motivation. Our method builds upon the fact that the latent spaces of neural models bear a structure informative of the data space they are trained on Shao et al. [30]. This structure can be enforced through loss regularizers; *e.g.,* in [31], a minimum-distortion criterion is applied on the latent space of a VAE for a shape generation task. We follow a similar line of thought and propose adopting a geometric term to regularize the latent representations of a CL model. Namely, we root our approach in spectral geometry; our choice is motivated by the pursuit of a **compact representation** characterized by **isometry invariance**. As shown in [11], the latent space of DNNs can be modeled as a Riemannian manifold whose *extrinsic* embedding is encoded in the latent vectors. Being extrinsic, these vectors are simply absolute coordinates encoding only one possible realization of the data manifold, out of its infinitely many possible isometries. Each isometry (e.g.; a rotation by 45°) would always encode the **same latent space**, but the **latent vectors will change** – this is not desirable, because it may lead to overfitting and lack of generalization. By resorting to spectral geometry, we instead rely on *intrinsic* quantities, that fully encode the latent space and are isometry-invariant.

Our regularizer is based on the graph-theoretic formulation of clustering, where we seek to partition the vertices of $\mathcal{G}$ into well-separated subgraphs with high internal connectivity. A body of results from spectral graph theory, dating back at least to Cheeger [32], Sinclair and Jerrum [33] and Shi and Malik [34], explain the gap occurring between neighboring Laplacian eigenvalues as a quantitative measure of graph partitioning. Our proposal, called Continual Spectral Regularizer for Incremental Learning (CaSpeR-IL), draws on these results, but turns the *forward* problem of computing the optimal partitioning of a given graph, into the *inverse* problem of seeking a graph with the desired partitioning.

Building the LGG. We take the examples in B and forward them through the network; their features are used to build a k-NN graph[2] $\mathcal{G}$; following Lassance et al. [29], we refer to it as the *latent geometry graph* (LGG).

Spectral Regularizer. Let us denote by $\boldsymbol{A}$ the adjacency matrix of $\mathcal{G}$, we calculate its degree matrix $\boldsymbol{D}$ and we compute its normalized Laplacian as $\boldsymbol{L} = \boldsymbol{I} - \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$, where $\boldsymbol{I}$ is the identity matrix. We then compute the eigenvalues λ of $\boldsymbol{L}$ and sort them in ascending order. Let g be the number of different classes within the buffer, we calculate our regularizing loss as:

$$\ell_{\text{CaSpeR}} \triangleq -\lambda_{g+1} + \sum_{j=1}^{g} \lambda_j \, . \tag{5}$$

The proposed loss term is weighted through the hyperparameter ρ and added to the stream classification loss. Overall, our model optimizes the following objective:

$$\underset{\theta}{\operatorname{argmin}} \, \ell_{\text{stream}} + \ell_{\text{b}} + \rho \, \ell_{\text{CaSpeR}} \, . \tag{6}$$

Through Eq. (5), we increase the eigengap $\lambda_{g+1} - \lambda_g$ while minimizing the first g eigenvalues. Since the number of eigenvalues close to zero represents to the number of loosely connected partitions within the graph [35], our loss indirectly encourages data points to be clustered without strict supervision.

Efficient Batch Operation. The application of CaSpeR-IL entails the cumbersome step of constructing the entire LGG $\mathcal{G}$ at each forward step by processing all available replay examples in the buffer B (usually orders of magnitude larger than a batch of input examples). We consequently propose an efficient approximation of our initial objective by not operating on $\mathcal{G}$ directly, but rather sampling a random sub-graph $\mathcal{G}_{\text{p}} \subset \mathcal{G}$ spanning only p out of the g classes represented in the memory buffer. As $\mathcal{G}_{\text{p}}$ still includes a conspicuous amount of nodes, we resort to an additional sub-sampling and extract $\mathcal{G}_{\text{p}}^{t} \subset \mathcal{G}_{\text{p}}$, a smaller graph with t

[1] See Section 3.3 for a detailed description of this procedure.

[2] More details on how the k-NN operation can be found in the supplemental material, along with the pseudo-code of CaSpeR-IL.

Table 1
Class-IL results – $\bar{A}_F$ ($\bar{F}_F^*$) – for SOTA rehearsal CL methods, with and without CaSpeR-IL.

Class-IL	Split CIFAR-10		Split CIFAR-100		Split *mini*ImageNet	
Joint (UB)	87.08 (–)		63.11 (–)		52.76 (–)	
Finetune (LB)	19.53 (100.00)		8.38 (100.00)		3.87 (100.00)	
Buffer size	500	1000	500	2000	2000	5000
ER-ACE	66.13 (21.76)	71.72 (14.88)	34.99 (51.41)	46.52 (34.60)	22.03 (49.04)	27.26 (29.99)
+ **CaSpeR-IL**	69.58 (20.56)	73.82 (14.11)	36.70 (46.61)	47.85 (33.86)	23.36 (47.90)	29.15 (28.36)
iCaRL	52.71 (22.69)	62.94 (21.64)	39.56 (32.73)	40.47 (31.24)	19.42 (36.89)	20.17 (33.23)
+ **CaSpeR-IL**	55.66 (20.56)	63.99 (21.05)	40.87 (32.31)	41.83 (25.55)	20.46 (35.90)	21.45 (32.26)
DER++	67.38 (26.77)	71.17 (25.12)	28.01 (57.56)	43.27 (34.94)	20.88 (74.48)	28.55 (61.03)
+ **CaSpeR-IL**	69.11 (26.18)	73.12 (23.43)	32.16 (53.41)	46.95 (30.08)	22.61 (71.01)	29.96 (57.60)
X-DER	63.23 (14.99)	65.72 (12.28)	35.89 (44.54)	46.37 (23.57)	24.80 (44.69)	30.98 (30.12)
+ **CaSpeR-IL**	65.56 (14.41)	67.84 (10.65)	38.23 (43.90)	48.11 (18.47)	26.24 (41.72)	31.63 (28.71)
PODNet	37.22 (40.49)	45.97 (39.49)	30.16 (54.49)	32.12 (46.73)	16.82 (52.32)	20.81 (46.50)
+ **CaSpeR-IL**	39.85 (39.51)	47.40 (38.90)	32.27 (48.32)	38.64 (35.65)	18.09 (50.33)	23.63 (45.08)

exemplars for each class. By repeating these random samplings in each forward step, we optimize a Monte Carlo approximation of Eq. (5):

$$\ell^*_{\text{CaSpeR}} \triangleq \mathop{\mathbb{E}}_{\mathcal{G}_{\text{p}}\subset\mathcal{G}}\left[\mathop{\mathbb{E}}_{\mathcal{G}^t_{\text{p}}\subset\mathcal{G}_{\text{p}}}\left[-\lambda_{p+1}^{\mathcal{G}^t_{\text{p}}}+\sum_{j=1}^{p}\lambda_j^{\mathcal{G}^t_{\text{p}}}\right]\right], \tag{7}$$

where the $\lambda^{\mathcal{G}^t_{\text{p}}}$ denote the eigenvalues of the Laplacian of $\mathcal{G}^t_{\text{p}}$. Here, we enforce the eigengap at p, as we know by construction that each $\mathcal{G}^t_{\text{p}}$ comprises samples from p communities within $\mathcal{G}$. In practice, we extract b samples from the buffer, maintaining b equal to the batch size to ensure balance [5,36]. This results in $t = \frac{b}{p}$ samples from each class of the p randomly chosen. An in-depth discussion on the hyperparameters of CaSpeR-IL can be found in the Supplemental Material.

4. Continual learning experiments

4.1. Evaluation

Settings. To assess the effectiveness of the proposed method, we prioritize *Class Incremental Learning* (Class-IL) classification protocol [3], where the model learns to make predictions in the absence of task information, as it is recognized as a more realistic and challenging benchmark [14,37]. In the supplemental material, we report results for both *Task-Incremental Learning* (Task-IL) and *Domain-Incremental Learning* (Domain-IL) protocols, demonstrating that CaSpeR-IL can enhance CL baselines within these scenarios as well.

Benchmarked models. To evaluate the benefit of our regularizer, we apply it on top of several SOTA rehearsal-based methods: Experience Replay with Asymmetric Cross-Entropy **(ER-ACE)** [6], Incremental Classifier and Representation Learning **(iCaRL)** [38], Dark Experience Replay **(DER++)** [5], eXtended-DER **(X-DER)** and Pooled Outputs Distillation Network **(PODNet)** [16].

We include the performance of the upper bound training on all classes together in an offline manner (Joint) and the lower bound training on each task sequentially without any method to prevent forgetting (Finetune).

Datasets. We conduct the experiments on three commonly used image datasets, splitting the classes from the main dataset into separate disjoint sets used to sequentially train the evaluated models. For **Split CIFAR-10** we adopt the standard benchmark of splitting the dataset into 5 subsets of 2 classes each; for **Split CIFAR-100**, we exploit the 100-class CIFAR100 [39] dataset by splitting the dataset into 10 subsets of 10 classes each; for **Split *mini*ImageNet**, we leverage the *mini*ImageNet [40] Imagenet subset, adopting the 20 tasks per 5 classes protocol.

Metrics. We mainly quantify the performance of the compared models in terms of *Final Average Accuracy* $\bar{A}_F \triangleq \frac{1}{T}\sum_{i=1}^{T} a_i^T$, where a_i^j is the accuracy of the model at the end of task j calculated on the test set of task τ_i and reported in percentage value. To quantify the severity of the performance degradation that occurs as a result of catastrophic forgetting, we exploit *Final Average Adjusted Forgetting* ($\bar{F}_F^*$), as defined in [41]. It is a $[0, 100]$-bounded version of the popular forgetting metric [42].

Hyperparameter selection. To ensure a fair evaluation, we train all the models with the same batch size and the same number of epochs. Moreover, we employ the same backbone for all experiments on the same dataset. In particular, we use Resnet18 [43] for Split CIFAR-100 and Split CIFAR-10 and EfficientNet-B2 [44] for Split *mini*ImageNet. The best hyperparameters for each model-dataset configuration are found via grid search. For additional details and further experiments with varying training epochs and batch sizes, demonstrating the effectiveness of CaSpeR-IL under different conditions, we direct the reader to the supplemental material.

4.2. Results

We report a breakdown of Class-IL results of our evaluation in Table 1. CaSpeR-IL leads to a firm improvement in $\bar{A}_F$ across all evaluated methods and datasets. The steady reduction of $\bar{F}_F^*$ confirms that the regularization adopted effectively addresses catastrophic forgetting.

We notice that the improvement in accuracy does not always grow with the memory buffer size. This is in contrast with the typical behavior of replay regularization terms [4,17]. We believe it is due to our distinctively geometric approach: as spectral properties of graphs are understood to be robust w.r.t. coarsening [45], CaSpeR-IL does not need a large pool of data to be effective.

In Task-IL and Domain-IL (results in the supplemental material), the gains are lower than in Class-IL: existing methods are already strong in these less challenging scenarios. However, CaSpeR-IL still provides a steady improvement, proving its ability to both consolidate the knowledge of each task individually (Task-IL) and counteract the bias introduced by new data distribution on known classes (Domain-IL).

Comparison with Contrastive Learning. The thorough study in [47] interprets contrastive learning as a parametric form of spectral clustering on the input augmentation graph, which points to a link with our approach. Given the similarity between CaSpeR-IL's goal and contrastive objectives, we devise a comparison with SCR (Supervised Contrastive Replay) [19] and CSCCT (Cross-Space Clustering and Controlled Transfer) [21], two existing CL contrastive baselines described in Section 2. We evaluated these methods on the Split CIFAR-100 benchmark described above. SCR is a standalone model extending Experience Replay; conversely, CSCCT is a module that can be plugged into existing CL methods, so we consider it as a direct competitor implemented upon our baselines. Despite behaving similarly to CaSpeR-IL, CSCCT requires a past model snapshot to be available during training and inserts both streaming and memory data within its loss terms. Results can be seen in Table 2. To better analyze the effects of different

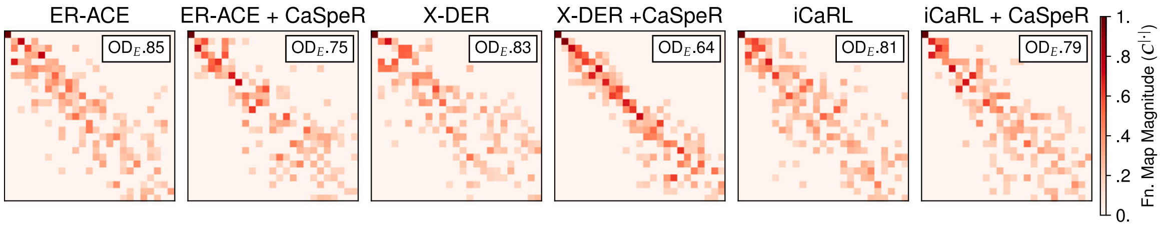


Fig. 3. For several rehearsal methods with and without CaSpeR-IL, the functional map magnitude matrices $C^{|\cdot|}$ between the LGGs $\mathcal{G}^{\tau_5}$ and $\mathcal{G}^{\tau_{10}}$, computed on the test set of $\tau_1,\ldots,\tau_5$ after training up to τ_5 and τ_{10} respectively (Split CIFAR-100 - buffer size 2000). The closer $C^{|\cdot|}$ to the diagonal, the less geometric distortion between $\mathcal{G}^{\tau_5}$ and $\mathcal{G}^{\tau_{10}}$. We report the first 25 rows and columns of $C^{|\cdot|}$, focusing on low-frequency correspondences [46], and apply a $C^{|\cdot|} > 0.15$ threshold to increase clarity.

Table 2
Comparison with contrastive baselines. We report $\bar{A}_F$ and the average variance of same-class projections on the latent space.

Class-IL	Split CIFAR-100			
Buffer size	500		2000	
	$\bar{A}_F$	Variance	$\bar{A}_F$	Variance
SCR	31.18	2.2111	43.39	4.4439
ER-ACE	34.99	0.5313	46.52	0.5769
+ **CaSpeR-IL**	**36.70**	0.4926	**47.85**	0.5478
+ **CSCCT**	34.93	**0.3931**	45.91	**0.4290**
iCaRL	39.56	0.8381	40.47	0.8248
+ **CaSpeR-IL**	**40.57**	**0.8289**	**41.83**	**0.8057**
+ **CSCCT**	39.36	0.9167	40.87	1.0392
DER++	28.01	0.1283	43.27	0.1209
+ **CaSpeR-IL**	**32.16**	0.0964	**46.95**	0.1012
+ **CSCCT**	30.17	**0.0552**	44.27	**0.0857**
X-DER	35.89	0.2265	46.37	0.2523
+ **CaSpeR-IL**	**38.23**	0.2065	**48.11**	**0.2207**
+ **CSCCT**	36.23	**0.1974**	45.51	0.2242
PODNet	30.16	0.4229	32.12	0.7366
+ **CaSpeR-IL**	**32.27**	0.4197	**38.64**	0.5700
+ **CSCCT**	30.78	**0.1809**	33.59	**0.2577**

CL approaches on the latent space, we measured the average variance of same-class projections at the end of training.

Firstly, we observe that SCR exhibits a higher variance in the latent space compared to other baselines. Conversely, both CSCCT and CaSpeR-IL are able to reduce the latent-space variance of the model they are applied to. The results highlight an intriguing behavior: despite CSCCT often achieving minimal variance, the accuracy improvement of CaSpeR-IL remains higher. Hence, we suggest two interesting explanations for these observations: (*i*) intra-class variance in latent space is not proportional to accuracy; (*ii*) while directly constraining individual coordinates may limit the model's ability to rearrange data points, spectral geometry may be a softer clustering approach that permits the flexibility required to organize the latent space into more optimal structures for classification tasks.

5. Model analysis

5.1. k-NN classification

To further verify whether CaSpeR-IL successfully separates the latent embeddings for examples of different classes, we evaluate the accuracy of k-NN-classifiers [48] trained on top of the latent representations produced by the methods of Section 4. In Table 3, we report the results for 5-NN and 11-NN classifiers using the final buffer B as a support set. We observe that CaSpeR-IL also shows its steady beneficial effect on top of this classification approach, further confirming its validity in disentangling the representations of different classes.

Table 3
Class-IL $\bar{A}_F$ values of k-NN classifiers trained on top of the latent representations of replay data points. Results on Split CIFAR-100 for Buffer size 2000.

k-NN Clsf	w/o CaSpeR-IL		w/ CaSpeR-IL	
(Class-IL)	5-NN	11-NN	5-NN	11-NN
ER-ACE	43.73	44.41	46.75+3.02	47.29+2.88
iCaRL	34.86	37.78	36.00+1.14	38.33+0.55
DER++	44.21	44.24	45.75+1.54	46.00+1.76
X-DER	43.44	44.62	49.47+6.03	49.49+4.87
PODNet	21.11	22.60	27.88+6.77	28.94+6.34

5.2. Latent space consistency

To provide further insights into the dynamics of the latent space on the evaluated models, we study the emergence of distortions in the LGG. Given a continual learning model, we are interested in a comparison between $\mathcal{G}^{\tau_5}$ and $\mathcal{G}^{\tau_{10}}$, the LGGs produced after training on τ_5 and τ_{10} respectively, computed on the test set of tasks $\tau_1,\ldots,\tau_5$.

The comparison between $\mathcal{G}^{\tau_5}$ and $\mathcal{G}^{\tau_{10}}$ can be better understood in terms of the node-to-node bijection $T : \mathcal{G}^{\tau_5} \rightarrow \mathcal{G}^{\tau_{10}}$, which can be represented as a functional map matrix $\boldsymbol{C}$ [46] with elements $c_{i,j} \triangleq \langle \boldsymbol{\phi}_i^{\mathcal{G}^{\tau_5}}, \boldsymbol{\phi}_j^{\mathcal{G}^{\tau_{10}}} \circ T \rangle$, where $\boldsymbol{\phi}_i^{\mathcal{G}^{\tau_5}}$ is the ith Laplacian eigenvector of $\mathcal{G}^{\tau_5}$ (similarly for $\mathcal{G}^{\tau_{10}}$), and $\circ$ denotes the standard function composition. In other words, the matrix $\boldsymbol{C}$ encodes the similarity between the Laplacian eigenspaces of the two graphs. In an ideal scenario where the latent space is subject to no modification between τ_5 and τ_{10} w.r.t. previously learned classes, T is an *isomorphism* and $\boldsymbol{C}$ is a diagonal matrix [46]. In a practical scenario, T is only approximately isomorphic, and, the better the approximation, the more $\boldsymbol{C}$ is sparse and funnel-shaped.

In Fig. 3, we report $\boldsymbol{C}^{|\cdot|} \triangleq \mathrm{abs}(\boldsymbol{C})$ for ER-ACE, iCaRL and X-DER on Split CIFAR-100, both with and without CaSpeR-IL. It can be observed that the methods that benefit from our proposal display a tighter functional map matrix. This indicates that the partitioning behavior promoted by CaSpeR-IL leads to reduced interference, as the portion of the LGG that refers to previously learned classes remains geometrically consistent in later tasks. To quantify the similarity of each $\boldsymbol{C}^{|\cdot|}$ matrix to the identity, we also report its off-diagonal energy OD_E [49], computed as the sum of the elements outside of the main diagonal divided by the Frobenius norm. CaSpeR-IL produces a clear decrease in OD_E, signifying an increase in the diagonality of the functional matrices.

5.3. Limitations

Given the necessity for the proposed regularizer to store and reuse previously learned training samples, we remark that CaSpeR-IL applicability might be limited if privacy constraints are in place. This applies to any rehearsal CL method.

6. Conclusion

In this work, we investigate the evolution of a CL model's latent space throughout training. We find that latent-space projections of past exemplars are relentlessly drawn closer together, paving the way for catastrophic forgetting. Drawing on spectral graph theory, we propose CaSpeR-IL: a regularizer that encourages the clustering of data points in the latent space, without constraining individual coordinates. We show that our approach can be easily combined with any rehearsal-based CL approach, improving their performance on standard benchmarks.

CRediT authorship contribution statement

Emanuele Frascaroli: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Riccardo Benaglia:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Matteo Boschini:** Conceptualization, Data curation, Formal analysis, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft. **Luca Moschella:** Conceptualization, Formal analysis, Methodology. **Cosimo Fiorini:** Funding acquisition, Resources. **Emanuele Rodolà:** Conceptualization, Formal analysis, Funding acquisition, Project administration, Supervision, Writing – original draft. **Simone Calderara:** Conceptualization, Formal analysis, Funding acquisition, Project administration, Resources, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have shared the link to my code within the article.

Acknowledgments

This paper has been supported from Italian Ministerial grant PRIN 2020 “LEGO.AI: LEarning the Geometry of knOwledge in AI systems”, n. 2020TA3K9N. We acknowledge the CINECA award under the ISCRA initiative, for the availability of high performance computing resources and support.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.patrec.2024.06.020.

References

[1] M. McCloskey, N.J. Cohen, Catastrophic interference in connectionist networks: The sequential learning problem, Psychol. Learn. Motiv. (1989).
[2] M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, T. Tuytelaars, A continual learning survey: Defying forgetting in classification tasks, IEEE Trans. Pattern Anal. Mach. Intell. (2021).
[3] G.M. van de Ven, T. Tuytelaars, A.S. Tolias, Three types of incremental learning, Nat. Mach. Intell. (2022).
[4] A. Chaudhry, M. Rohrbach, M. Elhoseiny, T. Ajanthan, P.K. Dokania, P.H. Torr, M. Ranzato, On tiny episodic memories in continual learning, in: ICML Workshops, 2019.
[5] P. Buzzega, M. Boschini, A. Porrello, D. Abati, S. Calderara, Dark experience for general continual learning: a strong, simple baseline, in: ANeurIPS, 2020.
[6] L. Caccia, R. Aljundi, N. Asadi, T. Tuytelaars, J. Pineau, E. Belilovsky, New insights on reducing abrupt representation change in online continual learning, in: ICLR, 2022.
[7] M. Boschini, P. Buzzega, L. Bonicelli, A. Porrello, S. Calderara, Continual Semi-Supervised Learning Through Contrastive Interpolation Consistency, PRL, 2022.
[8] L. Bonicelli, M. Boschini, A. Porrello, C. Spampinato, S. Calderara, On the effectiveness of Lipschitz-driven rehearsal in continual learning, in: ANeurIPS, 2022.
[9] Y. Wu, Y. Chen, L. Wang, Y. Ye, Z. Liu, Y. Guo, Y. Fu, Large scale incremental learning, in: CVPR, 2019.
[10] H. Ahn, J. Kwak, S. Lim, H. Bang, H. Kim, T. Moon, SS-IL: Separated softmax for incremental learning, in: ICCV, 2021.
[11] G. Arvanitidis, L.K. Hansen, S. Hauberg, Latent space oddity: on the curvature of deep generative models, in: ICLR, 2018.
[12] A. Mallya, S. Lazebnik, Packnet: Adding multiple tasks to a single network by iterative pruning, in: CVPR, 2018.
[13] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A.A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al., Overcoming catastrophic forgetting in neural networks, Proc. Natl. Acad. Sci. USA (2017).
[14] R. Aljundi, M. Lin, B. Goujaud, Y. Bengio, Gradient based sample selection for online continual learning, in: ANeurIPS, 2019.
[15] D. Lopez-Paz, M. Ranzato, Gradient episodic memory for continual learning, in: ANeurIPS, 2017.
[16] A. Douillard, M. Cord, C. Ollion, T. Robert, E. Valle, Podnet: Pooled outputs distillation for small-tasks incremental learning, in: ECCV, 2020.
[17] H. Cha, J. Lee, J. Shin, Co2l: Contrastive continual learning, in: ICCV, 2021.
[18] Z. Mai, R. Li, H. Kim, S. Sanner, Supervised contrastive replay: Revisiting the nearest class mean classifier in online class-incremental continual learning, in: CVPR Workshops, 2021, pp. 3589–3599.
[19] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, Supervised contrastive learning, in: ANeurIPS, 2020.
[20] N. Asadi, M. Davari, S. Mudur, R. Aljundi, E. Belilovsky, Prototype-sample relation distillation: towards replay-free continual learning, in: ICLR, 2023.
[21] A. Ashok, K. Joseph, V.N. Balasubramanian, Class-incremental learning with cross-space clustering and controlled transfer, in: ECCV, 2022.
[22] M. Boschini, L. Bonicelli, A. Porrello, G. Bellitto, M. Pennisi, S. Palazzo, C. Spampinato, S. Calderara, Transfer without forgetting, in: ECCV, 2022.
[23] Z. Wang, Z. Zhang, C.-Y. Lee, H. Zhang, R. Sun, X. Ren, G. Su, V. Perot, J. Dy, T. Pfister, Learning to prompt for continual learning, in: CVPR, 2022.
[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L.u. Kaiser, I. Polosukhin, Attention is all you need, in: ANeurIPS, 2017.
[25] L. Cosmo, M. Panine, A. Rampini, M. Ovsjanikov, M.M. Bronstein, E. Rodolà, Isospectralization, or how to hear shape, style, and correspondence, in: CVPR, 2019.
[26] R. Marin, A. Rampini, U. Castellani, E. Rodolà, M. Ovsjanikov, S. Melzi, Instant recovery of shape from spectrum via latent space connections, in: V. Struc, F.G. Fernández (Eds.), 3DV, 2020.
[27] L. Moschella, S. Melzi, L. Cosmo, F. Maggioli, O. Litany, M. Ovsjanikov, L.J. Guibas, E. Rodolà, Learning spectral unions of partial deformable 3D shapes, Comput. Graph. Forum (2022).
[28] A. Rampini, F. Pestarini, L. Cosmo, S. Melzi, E. Rodolà, Universal spectral adversarial attacks for deformable shapes, in: CVPR, 2021.
[29] C. Lassance, V. Gripon, A. Ortega, Representing deep neural networks latent space geometries with graphs, Algorithms (2021).
[30] H. Shao, A. Kumar, P.T. Fletcher, The Riemannian geometry of deep generative models, in: CVPR Workshops, 2018.
[31] L. Cosmo, A. Norelli, O. Halimi, R. Kimmel, E. Rodolà, LIMP: Learning latent shape representations with metric preservation priors, in: ECCV, 2020.
[32] J. Cheeger, A lower bound for the smallest eigenvalue of the Laplacian, in: Problems in Analysis, Princeton University Press, 1969.
[33] A. Sinclair, M. Jerrum, Approximate counting, uniform generation and rapidly mixing Markov chains, Inform. and Comput. (1989).
[34] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. (2000).
[35] J.R. Lee, S.O. Gharan, L. Trevisan, Multiway spectral partitioning and higher-order cheeger inequalities, J. ACM (2014).
[36] P. Buzzega, M. Boschini, A. Porrello, S. Calderara, Rethinking experience replay: a bag of tricks for continual learning, in: ICPR, 2020.
[37] S. Farquhar, Y. Gal, Towards robust evaluations of continual learning, in: ICML Workshops, 2018.
[38] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, C.H. Lampert, ICaRL: Incremental classifier and representation learning, in: CVPR, 2017.
[39] A. Krizhevsky, et al., Learning Multiple Layers of Features from Tiny Images, Technical Report, Citeseer, 2009.
[40] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al., Matching networks for one shot learning, in: ANeurIPS, 2016.
[41] L. Bonicelli, M. Boschini, E. Frascaroli, A. Porrello, M. Pennisi, G. Bellitto, S. Palazzo, C. Spampinato, S. Calderara, On the effectiveness of equivariant regularization for robust online continual learning, 2023, arXiv preprint arXiv:2305.03648.
[42] A. Chaudhry, P.K. Dokania, T. Ajanthan, P.H. Torr, Riemannian walk for incremental learning: Understanding forgetting and intransigence, in: ECCV, 2018.

[43] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, 2016.
[44] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: ICML, 2019.
[45] Y. Jin, A. Loukas, J. JaJa, Graph coarsening with preserved spectral properties, in: AISTATS, 2020.
[46] M. Ovsjanikov, M. Ben-Chen, J. Solomon, A. Butscher, L. Guibas, Functional maps: a flexible representation of maps between shapes, ACM Trans. Graph. (ToG) (2012).
[47] J.Z. HaoChen, C. Wei, A. Gaidon, T. Ma, Provable guarantees for self-supervised deep learning with spectral contrastive loss, in: ANeurIPS, 2021.
[48] Z. Wu, Y. Xiong, S.X. Yu, D. Lin, Unsupervised feature learning via non-parametric instance discrimination, in: CVPR, 2018.
[49] E. Rodolà, L. Cosmo, M.M. Bronstein, A. Torsello, D. Cremers, Partial functional correspondence, in: Comput. Graph. Forum, 2017.