ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

Coherence-based Sample Selection for Class-incremental Learning

Andrea Daou[1], Jean-Baptiste Pothin[1], Paul Honeine[2], Abdelaziz Bensrhair[2]

1- DATAHERTZ, 10000, Troyes, France

2- Univ Rouen Normandie, INSA Rouen Normandie, Université Le Havre Normandie, Normandie Univ, LITIS UR 4108, 76000, Rouen, France

Abstract. Class-Incremental Learning (Class-IL) is challenging as the model must adapt to new classes while retaining knowledge of old ones. To avoid catastrophic forgetting in knowledge distillation with a fixed-budget memory, exemplars from previously learned classes need to be stored. We propose a novel sample selection method based on the coherence measure to boost Class-IL performance. This is the first time the coherence is investigated in a deep model, specifically for Class-IL. We define the coherence between two samples as a normalized inner product between their deep feature extractor features. Theoretical results and extensive experiments demonstrate the relevance of our approach.

1 Introduction

Deep Learning (DL) is challenging when it comes to continual learning, specifically in scenarios involving incremental learning of new classes [1]. Class-Incremental Learning (Class-IL) has been recently introduced to continually build a classifier that contains all encountered classes [2, 3]. It involves incrementally updating the recognition system by adding new classes from a new training dataset (also called task) [4]. Due to storage constraints and to avoid catastrophic forgetting [5], the model can only access the current task's dataset, and a small retaining subset - called exemplar set - of the training dataset from previous tasks. Several exemplars selection strategies were proposed in the literature, the most well-known being mean-of-features sampling, also called herding [6], entropy-based sampling [7], and distance-based sampling [7].

In this paper, we propose a novel sample selection strategy for maximizing exemplar diversity in Class-IL. Inspired by the coherence measure for sparse approximation, we define the coherence between two samples as a normalized inner product between their feature vectors obtained by a DL model. Our method selects diverse and informative samples while considering a fixed-budget memory for each class, ensuring that the selected exemplars accommodate memory limits, which is critical in IL. We provide some theoretical results by connecting it to the herding criterion. Experiments on CIFAR-100 and MIT Indoor-67 demonstrate that our method outperforms state-of-the-art techniques with better accuracy in low computational complexity.

2 Background on Class-Incremental Learning

Class-IL techniques seek to effectively incorporate data from new classes into DL models while retaining knowledge gained from earlier classes. Thus, it is necessary to prevent catastrophic forgetting, *i.e.*, knowledge is lost when training on new data [5]. To address this issue, the rehearsal-based strategy has been advocated to maintain a few samples from previously seen classes. For this purpose, several studies relying on response-based Knowledge Distillation (KD) have recently emerged, such as LwF [8] and iCaRL [6]. We adopt in this paper LwF with exemplars (denoted by LwF-E) [9]. LwF trains a single network on multiple tasks without catastrophic forgetting by applying the KD from a large pretrained "teacher" model to a smaller "student" model. The student is trained on new classes while learning from the teacher's predictions. In LwF-E, distillation loss is applied to exemplars from old and new classes.

Following the notation of [6], let $X^1, X^2, \ldots,$ be the sample sets where each $X^y = \{x_1^y, \ldots, x_{n_y}^y\}$ contains samples of class $y \in \mathbb{N}$, and n_y is the number of samples. P^y denotes the selected exemplars for class y, with each class storing m exemplars. For new classes $s, \ldots, t$, we update the model $\Theta^{1:s-1}$ to $\Theta^{1:t}$ to classify both old and new classes. This update uses a combined training set D, which includes both new data $X^s, \ldots, X^t$ and exemplars from previous classes:

$$D = \bigcup_{y=s,\ldots,t} \{(x, y) : x \in X^y\} \bigcup_{y=1,\ldots,s-1} \{(x, y) : x \in P^y\}. \tag{1}$$

Let $\hat{\pi}_k(x)$ and $\pi_k(x)$ be the temperature scaled output logits for class $k \in \{1, \ldots, s-1\}$ of the old model $\Theta^{1:s-1}$ and the new model $\Theta^{1:t}$, respectively. The distillation loss is applied on the training set D defined in (1) following [8]:

$$L_d = -\sum_{x \in D} \sum_{k=1}^{s-1} \hat{\pi}_k(x) \log(\pi_k(x)). \tag{2}$$

The classification loss uses the softmax cross-entropy, calculated as:

$$L_c = -\sum_{(x,y) \in D} \sum_{k=1}^{t} \delta_{k=y} \log(p_k(x)), \tag{3}$$

where $\delta_{k=y}$ is the ground truth of the sample and $p_k(x)$ is the output softmax probability. The overall loss is $L = L_c + \lambda L_d$, where λ is the distilling coefficient.

Memory constraints prevent storing all data for re-training, making exemplar selection essential to manage memory limits, avoid catastrophic forgetting, and maintain knowledge from previous classes. Exemplars must represent key class characteristics while minimizing forgetting risk, with diversity being a crucial factor. Herding [10, 6] selects exemplars for each class by choosing samples with features closest to the class mean. For each class, embeddings $\phi(\cdot)$ are extracted, and the mean of the feature vectors is calculated by

$$\mu_y = \frac{1}{n_y} \sum_{x \in X^y} \phi(x). \tag{4}$$

At each iteration, an exemplar is chosen to minimize the mean distance from μ_y when added to the selected exemplars in its class. This process is repeated for all classes. Entropy-based sampling [7] computes the entropy of the softmax outputs and selects exemplars with higher uncertainty for each class. Distance-based sampling [7] selects exemplars that are closer to the decision boundary.

3 Coherence-based Criterion for Sample Selection

In continual learning, maintaining sparsity involves selecting relevant samples, called atoms, to define the dictionary. The coherence measure, quantifying the relevance of dictionary, corresponds to the largest correlation between atoms of a given dictionary, namely for a dictionary of unit-norm atoms $x_1, x_2, \dots, x_m$:

$$\text{coh} = \max_{i \neq j} |\langle x_i, x_j \rangle|. \tag{5}$$

This simple measure allows a deep analysis and characterization of the quality of the dictionary for sparse analysis and synthesis, and was extended to nonlinear (shallow) kernel-based models with in-depth theoretical results in [11, 12]. Motivated by the underlying theoretical results of the coherence measure, several online algorithms were proposed, such as nonlinear adaptive filtering [13] and nonlinear principal component analysis [14], to name a few. To derive online algorithms, atoms are selected with coherence below a threshold, namely

$$\text{coh} \leq \gamma. \tag{6}$$

The threshold controls diversity where a null value yields an orthogonal basis.

We propose in this paper a sample selection strategy by defining the coherence measure within DL. By considering the coherence as a similarity measure for DL features, we gain deeper insights than shallow versions like (5) or its kernel-based counterpart. Let $\phi : \mathcal{X} \to \mathbb{R}^d$ denote the DL feature extractor, then the DL coherence measure is defined as

$$\text{coh} = \max_{i \neq j} |\langle \phi(x_i), \phi(x_j) \rangle|, \tag{7}$$

for unit-norm embeddings; Otherwise, replace $\phi(x)$ with $\phi(x)/||\phi(x)||$. This paper is the first to explore a coherence measure in DL beyond shallow ones.

To enhance Class-IL performance, the proposed coherence criterion selects mutually least coherent examples of a class based on the DL coherence measure (7). This sampling technique is based on feature vectors obtained from the DL embedding, namely the last trained model so far, which means after the last incremental step. It strives to capture the sample diversity, guaranteeing that the chosen exemplars support effective information retention and learning in the IL context. Rather than relying on a fixed threshold γ on the coherence between exemplars as given in (6), we propose to work with a fixed-budget memory for each class, which provides better memory management. With the proposed strategy, the memory allocation grows incrementally as new classes

Algorithm 1 Coherence-based Sample Selection

```
Input: Sample set $X^y = \{x_1^y, \ldots, x_{n_y}^y\}$ of a class $y \in \{s, \ldots, t\}$, current feature extractor function $\phi : \mathcal{X} \to \mathbb{R}^d$.
for each: $y \in \{s, \ldots, t\}$
  Compute Gram matrix $G$ for all entries $x_i^y, x_j^y \in X^y$
  while size(G) $> m \times m$ do
    Select $(i, j) = \text{argmax}_{i \neq j} |G_{ij}|$
    Update $X^y \leftarrow X^y \setminus \{x_i^y\}$
    Remove $i^{th}$ row and $i^{th}$ column from G
  end while
Output: exemplar set $P^y \leftarrow X^y$.
```

are available. Each class is assigned its own fixed memory space to store m exemplars. For a new task with classes $s, \ldots, t$, an update procedure is called when data for these classes is available. It adjusts the DL parameters to create a new model $\Theta^{1:t}$ based on KD explained in Section 2. This new model is then used to augment the exemplars saved in memory to get $P^s, \ldots, P^t$ using the new training data $X^s, \ldots, X^t$. For each class $y \in \{s, \ldots, t\}$, we compute the DL coherence measure between each pair of vectors (class-wise comparisons). Let G be the matrix of $\langle \phi(x_i), \phi(x_j) \rangle$ for $x_i, x_j \in X^y$. In order to produce a γ-coherent set of m exemplars for each class y, we eliminate the maximum coherence values. See Algorithm 1 for an overview of the algorithm.

We provide next an upper bound on the approximation error of the mean-of-features μ_y for all samples in class as given in (4). This theoretical result connects our criterion to the herding criterion [10, 6]. The proof is roughly similar to the proof of [15, Theorem 1], but omitted here due to space limit.

Theorem 1. *Consider the coherence-based criterion that selects m exemplars of coherence γ from n_y samples of X^y. The error of approximating μ_y by these exemplars is upper bounded as follows:*

$$||\mu_y - \mu_{P^y}|| \leq \left(1 - \frac{m}{n_y}\right) \sqrt{\max_{x \in X^y} ||\phi(x)||^2 - \gamma}. \tag{8}$$

4 Experiments and Results

We study the performance of different sample selection strategies used in rehearsal-based Class-IL on two different datasets: CIFAR-100 and MIT Indoor-67. CIFAR-100 provides 32×32 color (RGB) images for 100 object classes, with 600 images divided into 500 for training and 100 for testing for each class. We randomly selected 50 objects classes from the CIFAR-100 dataset with 5 tasks of 10 classes each. MIT Indoor-67 includes 67 indoor scene categories with 15 620 RGB images in total, with at least 100 images per category divided into 80 for training and 20 for testing. This dataset has 5 tasks divided as follows: (0, 14) (*i.e.,* 14 classes for the initial task), (1, 14), (2, 13), (3, 13), and (4, 13).

Table 1: Accuracy rates for different strategies (best results are in bold).

Task	sampling strategy	CIFAR-100	MIT Indoor-67
$p = 2$ (after 3 tasks)	herding	35.12 ± 2.54	68.92 ± 1.76
	entropy	27.76 ± 1.00	69.02 ± 2.35
	distance	27.14 ± 2.44	68.70 ± 1.49
	coherence (ours)	**35.98 ± 2.18**	**69.26 ± 2.57**
$p = 4$ (after 5 tasks)	herding	28.84 ± 1.46	58.18 ± 1.17
	entropy	19.96 ± 0.73	56.58 ± 1.67
	distance	19.28 ± 1.02	57.14 ± 1.03
	coherence (ours)	**29.82 ± 2.71**	**58.52 ± 0.91**

We implemented[1] the distillation loss L_d following (2). We relied on pretrained ResNet-32 [16] for CIFAR-100 and pretrained MobileNet-v2 [17] for MIT Indoor-67. For a fair comparison, we used the same settings as in [9], as given next: We used a learning rate search scheme, a patience of 10, a learning rate factor of 3 (*i.e.,* the learning rate was divided by 3 each time the patience is exhausted). Temperature scaling was set to $T = 2$ and the distilling coefficient to $\lambda = 1$. We set a gradient clipping at $10'000$, a SGD optimizer with a momentum of 0.9, a weight decay of 0.0002 and a batch size of 64 samples. The training stopped either if the learning rate became equal to 10^{-4} or after 200 epochs.

The test accuracy at task p is defined as $\frac{1}{p+1}\sum_{q=0}^{p} a_{p,q}$, where $a_{p,q}$ denotes the accuracy of task q after learning task p, with 0 being the initial task ($q \leq p$). The test accuracies are averaged over five runs. We compare our proposed strategy in the fixed memory per class scenario taking $m = 20$ exemplars per class with existing and validated sample selection algorithms that have previously been employed in Class-IL approaches, mainly: herding, entropy and distance. The results presented in Table 1 show that the proposed criterion outperforms the other strategies in terms of test accuracy.

We also measured the execution time, recording the duration of the selection process. The experiments were performed on CIFAR-100, using 5 tasks with 10 classes each, on Google Colab, utilizing the NVIDIA T4 GPU provided by the platform. For selecting $m = 20$ exemplars per class, the herding strategy took 3.6 s, while our coherence strategy required 5.1 s. For $m = 200$, herding took 23.4 s, whereas our coherence strategy required 3.6 s. These results revealed that our strategy exhibited reasonable time selection compared to herding.

5 Conclusion

This paper proposed a novel criterion for efficient exemplars selection in Class-IL, by promoting diversity among class exemplars using a DL-based coherence measure. Theoretical results connect this criterion to herding. Experimental results show superior performance compared to state-of-the-art techniques in terms of accuracy and reasonable execution time. Future work will expand the proposed sample selection strategy analysis to other DL frameworks, and beyond Class-IL in continual learning.

[1] PyTorch source code from `https://github.com/mmasana/FACIL`

References

[1] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A Comprehensive Survey of Continual Learning: Theory, Method and Application . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 46(08):5362–5383, August 2024.

[2] Eden Belouadah, Adrian Popescu, and Ioannis Kanellos. A comprehensive study of class incremental learning algorithms for visual tasks. *Neural Networks*, 135:38–54, 2021.

[3] Da-Wei Zhou, Qi-Wei Wang, Zhi-Hong Qi, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Deep class-incremental learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[4] Chen He, Ruiping Wang, and Xilin Chen. Rethinking class orders and transferability in class incremental learning. *Pattern Recognition Letters*, 161:67–73, 2022.

[5] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

[6] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. ICARL: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.

[7] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European conference on computer vision (ECCV)*, pages 532–547, 2018.

[8] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

[9] Marc Masana, Xialei Liu, Bartłomiej Twardowski, Mikel Menta, Andrew D Bagdanov, and Joost Van De Weijer. Class-incremental learning: survey and performance evaluation on image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5513–5533, 2022.

[10] Max Welling. Herding dynamical weights to learn. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1121–1128, 2009.

[11] Paul Honeine. Analyzing sparse dictionaries for online learning with kernels. *IEEE Transactions on Signal Processing*, 63(23):6343–6353, 2015.

[12] Paul Honeine. Approximation errors of online sparsification criteria. *IEEE Transactions on Signal Processing*, 63(17):4700–4709, 2015.

[13] Cédric Richard, José C. M. Bermudez, and Paul Honeine. Online prediction of time series data with kernels. *IEEE Transactions on Signal Processing*, 57(3):1058 – 1067, March 2009.

[14] Paul Honeine. Online kernel principal component analysis: a reduced-order model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1814 – 1826, September 2012.

[15] Zineb Noumir, Paul Honeine, and Cédric Richard. One-class machines based on the coherence criterion. In *Proc. IEEE workshop on Statistical Signal Processing (SSP)*, pages 600 – 603, Ann Arbor, Michigan, USA, 5 - 8 August 2012.

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

[17] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.