(Article begins on next page)

An Attention-based Representation Distillation Baseline for Multi-Label Continual Learning

Martin Menabue[1], Emanuele Frascaroli[1], Matteo Boschini[1], Lorenzo Bonicelli[1], Angelo Porrello[1], and Simone Calderara[1]

[1] University of Modena and Reggio Emilia, Italy
`{name.surname}@unimore.it`

Abstract. The field of Continual Learning (CL) has inspired numerous researchers over the years, leading to increasingly advanced countermeasures to the issue of *catastrophic forgetting*. Most studies have focused on the single-class scenario, where each example comes with a single label. The recent literature has successfully tackled such a setting, with impressive results. Differently, we shift our attention to the multi-label scenario, as we feel it to be more representative of real-world open problems. In our work, we show that existing state-of-the-art CL methods fail to achieve satisfactory performance, thus questioning the real advance claimed in recent years. Therefore, we assess both old-style and novel strategies and propose, on top of them, an approach called Selective Class Attention Distillation (SCAD). It relies on a knowledge transfer technique that seeks to align the representations of the student network – which trains continuously and is subject to forgetting – with the teacher ones, which is pretrained and kept frozen. Importantly, our method is able to selectively transfer the relevant information from the teacher to the student, thereby preventing irrelevant information from harming the student's performance during online training. To demonstrate the merits of our approach, we conduct experiments on two different multi-label datasets, showing that our method outperforms the current state-of-the-art Continual Learning methods. Our findings highlight the importance of addressing the unique challenges posed by multi-label environments in the field of Continual Learning. The code of SCAD is available at `https://github.com/aimagelab/SCAD-LOD-2024`.

Keywords: Continual Learning · Multi-Label classification · Knowledge Distillation.

1 Introduction

During their lifetime, individuals exist within an environment that undergoes continuous changes and are constantly confronted with new and diverse problems. While humans possess the ability to adapt to these changes, acquire new knowledge, and remember past information, studies have shown that artificial intelligence algorithms struggle to replicate this behavior. Specifically, it has been observed that neural networks tend to forget previously solved tasks while

learning new knowledge, a phenomenon known as *catastrophic forgetting* [23]. The field of Continual Learning (CL) [10, 25] has emerged to explore techniques that enable neural networks to continuously learn new concepts while retaining previously acquired knowledge.

The problem of forgetting has been investigated from various perspectives, leading to the development of numerous effective techniques [5, 40, 32]. Many previous works have adopted ResNet-like [14] models as a backbone [5, 30, 22]. However, due to advancements in many Deep Learning fields with the introduction of Transformers [35] architectures, recent works have shifted towards using pretrained Vision Transformers (ViT) [40, 39, 32], achieving extraordinary performance. Remarkably, these methods can display results close to the upper bound, *i.e.*, a model trained offline with access to the whole dataset. Thus, we believe current CL image classification benchmarks are close to being saturated, as these new models show almost zero forgetting. This leads us to ask: has the catastrophic forgetting problem been successfully solved?

Present CL settings focus on single-label datasets, where each example comes with a single label. However, these settings may represent an approximation of real-world contexts: in life, individuals are used to observing objects that are simultaneously related to multiple concepts (e.g. a person can be related to a precise age, gender and nationality). To the best of our knowledge, only a few research works have tackled multi-label settings in an incremental fashion [1, 16].

In this work, we find that modern state-of-the-art CL approaches, that strongly contrast forgetting in single-label scenarios, struggle to deal with multiple labels. In light of these findings, we explore **Multi-Label Continual Learning (MLCL)** as a new context in which catastrophic forgetting still needs to be adequately addressed. In summary:

- We employ two MLCL benchmarks: IIRC CIFAR-100, proposed by [1], which adds superclass prediction to the popular CIFAR-100 dataset [18], and Incremental WebVision, a novel real-life scenario extracted from the large WebVision dataset [19].
- We evaluate current SOTA methods based on prompting in the MLCL setting, showing that they fail under the new challenges it poses.
- We implemented and tested popular rehearsal CL baselines on top of the pretrained ViT backbone, revealing that these simple combinations are not overwhelmed by MLCL adversities.
- We present a new framework able to cope with MLCL difficulties: **Selective Class Attention Distillation (SCAD)**; inspired by [28, 4], which highlights the usefulness of pretraining in mitigating the problem of forgetting, our approach employs attention to select valuable information from a pretrained frozen teacher model and transmits it to the student network facing incremental tasks.

We show that SCAD achieves the best performance in both MLCL benchmarks, demonstrating that filtering only useful information from pretraining effectively mitigates forgetting in this new challenging scenario.

2 Related works

Continual Learning (**CL**) [10, 25] is a field of Deep Learning that addresses the issue of *catastrophic forgetting* [23], a phenomenon that occurs when a model is trained on a sequence of tasks. As the data distribution shifts between tasks, the model needs to adapt to new tasks (*plasticity*) while preserving past information (*stability*). It would be ideal to archive a balance between these two properties, as too much plasticity may lead the model to forget previously learned information, while excessive stability may prevent the model from learning new tasks.

Continual Learning (CL) approaches are usually divided into the following categories [36]:

- *Rehearsal based methods* contrast catastrophic forgetting by storing a subset of samples from past tasks in a memory buffer; these are then replayed during training along with data from the current task. **Experience Replay** [29, 31] (**ER**) is a simple baseline that interleaves samples from the buffer with samples from the current task. More advanced methods have been developed from this base technique: **Dark Experience Replay** (**DER**) [5], which stores the logits of the model to distill knowledge from previous tasks; **Gradient Episodic Memory** (**GEM**) [22] and its improved version **A-GEM** [8], that use past training data to minimize gradient interference.
- *Regularization based methods* introduce additional terms in the loss function to prevent the weights of the model from excessively changing; **Elastic Weight Consolidation** (**EWC**) [17] estimates the importance of each weight and dynamically constrains their shifts.
- *Knowledge distillation methods* [15] aim to reduce the disparities between a model's current representations and its previous projections. For example, **Learning Without Forgetting** (**LWF**) [20] computes the responses of the model for new examples at the beginning of each task and uses them to guide the training of a student.

In recent years, with the advent of Vision Transformers [11], researchers have been exploring Continual Learning solutions using these new architectures. One category of successful methods is represented by the ***prompting*** approach [40, 39, 32], which allocates a set of trainable tokens, called *prompts*, at the beginning of training. These prompts are concatenated with the input of the Vision Transformer and updated throughout the training while the rest of the model typically remains frozen. Researchers have devised various mechanisms for prompt selection to ensure that prompts can encode information about the current task during training. This proves useful during inference as prompts can assist the model in remembering past tasks.

Multi-label classification. Traditional Continual Learning approaches are designed for single-label settings, where each sample is associated with only one label. In multi-label setups [33, 13, 34], a sample can have multiple labels, and the model must predict all of them accurately.

An interesting multi-label setup is presented in [1]: the authors propose a novel benchmark called **IIRC**, which involves adapting widely used datasets,

such as CIFAR-100 and ImageNet, for multi-label continual evaluation. This adaptation is achieved by assigning to each sample both a superclass and a subclass label. The authors identified two possible settings: *complete information*, where all labels for each sample are available, and *incomplete information*, where only the label of the current task is provided. We have chosen IIRC as one of our benchmarks and decided to concentrate on the incomplete setup, which the authors denote as a more challenging scenario.

Regarding MLCL literature, an interesting approach is presented in [16], where the authors aim to tackle two challenges simultaneously: the problem of catastrophic forgetting and the issue of long-tail distribution, in which a few categories have a vast number of samples while most have only a few. The authors' experiments reveal that the least frequent classes (tail) tend to suffer more from forgetting compared to the most recurrent classes (head). To address this issue, they devised a novel sampling strategy called *Partitioned Reservoir Sampling* to select which samples to store in the buffer.

3   Method

3.1   Multi-Label Continual Learning

In a multi-label setting, an image classification dataset $\mathcal{D}$ consists of pairs $(\mathbf{x}_i, \mathbf{y}_i)_{i=1}^{|\mathcal{D}|}$, where $\mathbf{x}_i \in \mathbb{R}^{3 \times H \times W}$ is an RGB image and $\mathbf{y}_i \in \mathbb{R}^{|\mathcal{Y}|}$ is the multi-hot label vector. Specifically, each element $y_{i,j}$ is 1 if the sample belongs to the class with index j, and 0 otherwise. The set $\mathcal{Y}$ encompasses all the possible labels. Moreover, let $\mathcal{B} = \{\mathbf{x}_i, y_i\}_{i=0}^{B}$ be a B-dimensional batch of samples, we denote with $\mathbf{X}[s:e]$ the subset $\{\mathbf{x}_i, y_i\}_{i=s}^{e} \subseteq \mathcal{B}$.

In the MLCL scenario, a model f_θ is trained on a sequence of N tasks, where each task $T_t = \{(x_j^t, y_j^t)\}_{j=1}^{|T_t|}, t = 1, ..., N$. An objective function $\mathcal{L}$ is optimized in order to minimize the classification error:

$$\min_\theta \mathcal{L}_{clf} = \sum_{t=1}^{N} \left[\mathbb{E}_{(\mathbf{x},\mathbf{y}) \sim T_t} \left[\ell(\mathbf{y}, f_\theta(\mathbf{x})) \right] \right] \tag{1}$$

where ℓ is a loss function, commonly the binary cross-entropy loss [27, 1]. Directly optimizing this function is not feasible since we assume that samples from previous tasks are no longer accessible. Consequently, our goal is to find parameters that fit the current task while retaining knowledge from previous tasks.

3.2   Pretraining and forgetting

Earlier studies [28, 24] have emphasized the resilience of large-scale pretrained models against the issue of catastrophic forgetting. However, as mentioned in [4], when a model is trained on a sequence of tasks, it tends to gradually deviate from its initial pretraining state, thereby intensifying the problem of forgetting. To address this issue, we explored solutions to keep a network aligned with the
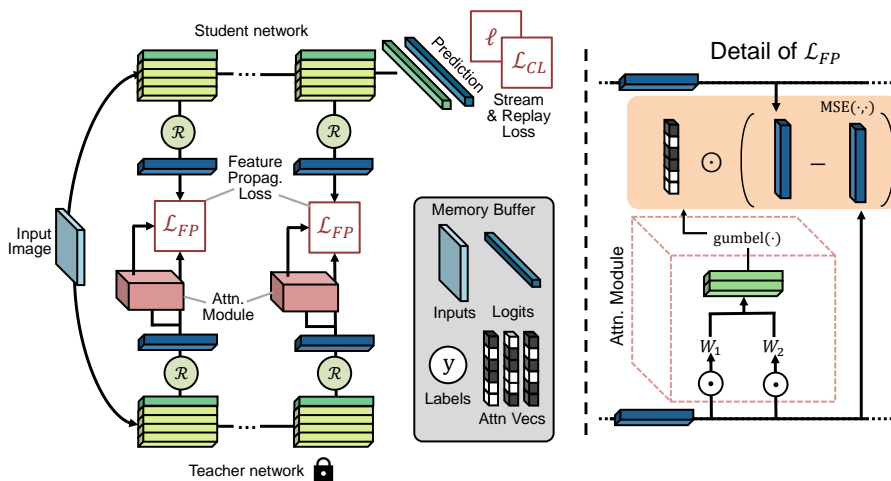
Fig. 1. Our proposal involves two pretrained Vision Transformers: one is frozen, referred to as the *teacher*, while the other is not, referred to as the *student*. An image is provided as input to both the teacher and the student, and intermediate representations are extracted. From these representations, using equation (2) and the indexing indicated in equation (3), we can derive the attention vector that contains the relationships between the class token and the other tokens. These operations are summarized in the image by the operator $\mathcal{R}$. The attention vector is used by the attention modules (*adapters*) to derive binary vectors useful for filtering knowledge transfer. The feature propagation loss $\mathcal{L}_{FP}$ enables maintaining alignment between the attention vectors of the teacher and the student.

pretraining state during the incremental learning. Our proposal draws inspiration from various research works addressing the problem of knowledge transfer between neural networks [38, 4, 21, 41]. The method we have developed, SCAD, involves using two pretrained neural networks, with one network (referred to as the *teacher*) being frozen, while the parameters of the other network (referred to as the *student*) are optimized during training. In the following paragraphs, we provide a more detailed description of our approach.

3.3 Selective Class Attention Distillation

Backbone Regarding the network architecture we use a pretrained Vision Transformer [11] for both the teacher and the student since recent Continual Learning state-of-the-art methods use this kind of architecture. In a Vision Transformer, the input image $\mathbf{x} \in \mathbb{R}^{3 \times H \times W}$ is divided into non-overlapping S patches, which are then flattened and projected by the patch embedding layer, obtaining a sequence of tokens. An additional token, the *class token*, which serves for classification, is concatenated to the other tokens, obtaining a sequence $\mathbf{x}_p \in \mathbb{R}^{(S+1) \times D}$, where S is the sequence length and D is the embedding dimension. The sequence

of tokens is forwarded through a chain of transformer encoder blocks, composed by Multi Head Self Attention [35] and Feed Forward layers, and the final class token is used to make the prediction. Thanks to the Multi-head Self Attention, the model is able to capture different types of dependencies and relationships between image tokens. Moreover, each attention head focuses on different aspects of the input, enabling the model to capture local and global dependencies.

Knowledge transfer technique While minimizing the distance between the intermediate representations of teacher and student is a straightforward approach to knowledge transfer, it is not the sole technique available. Recent years have witnessed extensive research aimed at investigating efficient strategies for distilling knowledge among Vision Transformers [9, 41, 38]. Our research work draws inspiration from the technique introduced by [38]. When passing an input through a Vision Transformer, we can obtain the intermediate representations (i.e. the output of each block), denoted as $F^l \in \mathbb{R}^{(S+1) \times D}$, where $l \in 1, ..., |L|$ represents the index of the transformer block under consideration. At this stage, we compute the correlation between the tokens within these representations, after normalizing the feature map:

$$\mathcal{R}^l = \frac{F^l}{||F^l||_2^2} \cdot \left(\frac{F^l}{||F^l||_2^2} \right)^T \tag{2}$$

Here, $\mathcal{R}^l \in \mathbb{R}^{(S+1) \times (S+1)}$. Our objective is to transfer the attention vector $\mathcal{R}^l[0, 1 : S + 1]$, which captures the relationships between the class token and all other tokens. To achieve this, we compute the distance between the teacher's vector and the corresponding vector of the student:

$$D^l = \mathcal{R}^l_{\mathcal{T}}[0, 1 : S + 1] - \mathcal{R}^l_{\mathcal{S}}[0, 1 : S + 1] \tag{3}$$

Where the subscripts $\mathcal{T}$ and $\mathcal{S}$ denote the teacher and the student respectively. This approach ensures that the student model maintains the relationships between the class token and the other tokens aligned to those learned by the teacher model.

Adapter networks The aforementioned knowledge transfer technique is an effective strategy for alleviating the issue of forgetting the pretraining state. However, it carries the risk of making the model too rigid, hindering its ability to adapt to new tasks. To address this concern, we introduce additional learnable components which we call ***adapters***. These adapters aim to prevent the transfer of irrelevant information from the teacher model to the student model. To achieve this objective, we feed the attention vectors $\mathcal{R}_{\mathcal{T}}$ generated by the teacher model into these adapter modules. The adapters are tasked with producing binary masks denoted as $\mathbb{M}(\cdot)$. To obtain the binary masks, we multiply the attention vectors by two sets of learnable weights, resulting in a vector $\mathbf{v} \in \mathbb{R}^{(S+1) \times 2}$. We then apply binary Gumbel-Softmax sampling on the last dimension, obtaining a binary vector. Next, we multiply the binary vectors by the distance vector $\mathcal{D}$

derived from equation 3. Subsequently, we calculate the mean squared distance and average this value across all the adapters. The objective function for this process can be defined as:

$$\mathcal{L}_{FP} = \frac{1}{|L|} \left[\sum_{l \in L} ||\mathbb{M}(\mathcal{R}_{\mathcal{T}}^l) \odot \mathcal{D}^l||_2^2 \right], \tag{4}$$

Here, $\mathbb{M}^l(\mathcal{R}_{\mathcal{T}}^l)$ represents the binary vector obtained by the lth adapter using the teacher's attention vector $\mathcal{R}_{\mathcal{T}}^l$ and $\odot$ is the element-wise multiplication.

Experience Replay Rehearsal methods have proven to be highly effective in addressing the forgetting problem [12, 2]. For this reason, we decided to conduct experiments by incorporating rehearsal techniques into our method. Consequently, at the beginning of the training process, we allocate a fixed-sized memory buffer $\mathcal{M}$. During training, we save a subset of the examples seen by the model into the buffer [29, 31], along with their associated labels and the logits obtained by feeding these examples to the student network [5]. To determine which examples to store in the buffer, we employ the reservoir sampling strategy [37], which ensures that all examples in the stream have an equal probability of being chosen for storage within the buffer. Moreover, it guarantees that the labels stored in the memory buffer are distributed according to the dataset's label distribution. We want that our model is capable of correctly classifying past examples. This can be achieved by computing the classification loss on the examples sampled from the buffer and using the saved labels as the ground truth:

$$\mathcal{L}_{er} = \mathbb{E}_{(\mathbf{x}_i, \mathbf{y}_i) \sim M} \left[\text{BCE}(f_\theta(\mathbf{x}_i), \mathbf{y}_i) \right] \tag{5}$$

Another useful strategy for mitigating forgetting on past tasks is to enforce similarity between the student's logits obtained from examples sampled from the buffer and the stored responses. This can be achieved by minimizing the mean squared error between current and saved logits [5]:

$$\mathcal{L}_{der} = \mathbb{E}_{(\mathbf{x}_i, \mathbf{h}_i) \sim M} \left[\text{MSE}(f_\theta(\mathbf{x}_i), \mathbf{h}_i) \right] \tag{6}$$

Here, MSE denotes the mean squared error and $\mathbf{h}_i$ represents the logits. By incorporating this additional objective, we encourage the student network to produce logits that closely resemble the stored responses, thereby preserving knowledge learned from previous tasks.

As an additional step, we apply the ACE technique, as described in [6], which computes the classification loss on the stream examples considering only the labels within the current batch. According to the authors, this approach helps in reducing abrupt changes in model representations.

Adapter mask replay As adapter networks contain learnable parameters, they are also susceptible to forgetting. To address this issue, we alleviate the problem by implementing a replay mechanism for the adapter binary masks. Specifically,

we store the binary masks generated by the adapters on the non-augmented training samples in a memory buffer. During subsequent tasks, when experience replay is performed and a sample is retrieved from the buffer, we compare the corresponding saved mask with the one currently generated by the adapter networks and compute the following regularization loss:

$$\mathcal{L}_{FP\_rep} = \mathop{\mathbb{E}}_{\substack{(\mathbf{x},\mathbf{m}^l)\sim\mathcal{M} \\ l\in L}} \left[\mathrm{BCE}\left(\mathbb{M}(\mathcal{R}_{\mathcal{T}}^l(\mathbf{x})), \mathbf{m}^l \right) \right] \tag{7}$$

where BCE denotes the binary cross entropy loss and $(\mathbf{x},\mathbf{m})$ refers to a pair consisting of a sample and an attention map sampled from the memory buffer.

The final loss function is given by:

$$\mathcal{L} = \mathcal{L}_{clf} + \alpha \cdot \mathcal{L}_{der} + \beta \cdot \mathcal{L}_{er} + \lambda_{FP} \cdot \mathcal{L}_{FP} + \lambda_{FP\_rep} \cdot \mathcal{L}_{FP\_rep} \tag{8}$$

where α, β, λ_{FP}, λ_{FP_rep} are hyperparameters regulating the contributions of the loss terms.

4   Experiments

4.1   Experimental setup

Datasets We conducted our experiments on two different benchmarks: IIRC CIFAR-100 and WebVision. The former is a recently introduced benchmark, while the latter is a novel benchmark designed by us.

The **IIRC CIFAR-100** dataset [1] is designed such that every sample is assigned two labels - one for the CIFAR-100 [18] superclass and the other for the subclass. There are 22 tasks in total, with the first task comprising only the superclass labels, while the subsequent tasks introduce subclass labels. To make the task more challenging, we opted for the **incomplete information setting**, where only the label corresponding to the current task is provided for each sample.

The **WebVision** dataset [19] includes over 2.4 million web images sourced from the Internet, each of which is accompanied by metadata such as a title, description, and tags. Using the provided tags, we create a challenging real-world incremental multi-label benchmark that we refer to as **Incremental WebVision** by following these steps:

1. Starting from a pool of $816,814$ tags, we discard those with less than 1000 occurrences across the dataset so as not to train the model on targets which are not general enough;
2. Using feature similarity in CLIP-space [26] as an indicator of semantic proximity, we manually review individual clusters of consistently co-occurring tags and merge those that are semantically overlapped. As a result, we obtain a curated list of roughly $350,000$ images with 300 tags, that will respectively serve as examples and (multi-)labels for our benchmark;

3. During each task T_i $(i \in 1, .., 6)$ we first choose and introduce 50 unique new labels $\mathcal{Y}_i$ and subsequently compose the task's dataset by choosing $\approx 58,000$ $(\approx 350,000/6)$ images such that they are associated with at least one tag in $\mathcal{Y}_{1..i} \triangleq \bigcup_{j=1}^{i} \mathcal{Y}_j$[1];
4. Finally, we reserve 5000 images from each task for testing and employ the remaining $53,000$ for training purposes.

The dataset obtained from the steps above portrays a particularly challenging and real-world scenario, as annotations in WebVision are known to contain significant noise [19].

Metrics As stated in [33], conventional metrics fail to capture the varying levels of correctness in multi-label predictions. To address this limitation, building on the IIRC framework, we decided to use the **Precision-Weighted Jaccard Similarity** (**PWJS**) as our evaluation metric, defined as:

$$R_{j,k} = \frac{1}{n_k} \sum_{i=1}^{n_k} \frac{|Y_{j,k,i} \cap \hat{Y}_{j,k,i}|}{|Y_{j,k,i} \cup \hat{Y}_{j,k,i}|} \times \frac{|Y_{j,k,i} \cap \hat{Y}_{j,k,i}|}{|\hat{Y}_{j,k,i}|}$$

Where $R_{j,k}$ is the performance of a model on task k after training on task j, $\hat{Y}_{j,k,i}$ are model predictions for the ith sample of task k after training on task j, $Y_{j,k,i}$ are the corresponding ground-truth labels, and n_k is the number of samples of the test set of task k.

Since we are dealing with a sequence of tasks, we summarize the final performance by considering the average value of $R_{j,k}$ after training on the last task, which we refer to as the **Final Average PWJS**:

$$AR_f = \frac{1}{N} \sum_{m=1}^{N} R_{N,m} \tag{9}$$

where N is the total number of tasks in our benchmark. We run all our experiments from the same pretrained weights, repeating each 2 times and reporting the average AR_f and standard deviation for all evaluated approaches.

Competitors In order to assess the effectiveness of our proposed approach, we conducted a thorough evaluation by comparing it with the following state-of-the-art Continual Learning techniques:

– **Experience Replay (ER)** [29] is a simple rehearsal strategy which consists in saving the samples seen by the model in a memory of a fixed size. During the continual training, the examples from the memory are given to the model again, alongside the examples of the current task. Although this approach is straightforward, it has the drawback of leaving a significant memory footprint.

---

[1] Our benchmark adheres to the **complete information** setup as formulated in the IIRC paper [1].

- **ER with Asymmetric Cross-Entropy (ER-ACE)** [6] stands out from ER by computing the loss function for examples in the current batch considering only the classes present in that batch. Instead, for examples sampled from the buffer, it calculates the loss function considering all the classes seen up to that point.
- **DER++-ACE** [6] combines the logit distillation-based approach presented in [5] with the insights offered by ER-ACE.
- **Learning to Prompt** (**L2P**) [40] creates a pool of learnable prompts each associated to a learnable key. During training, the representation of a sample is obtained from a pretrained and frozen ViT, which is then used to query the prompt pool by computing the similarity between the query and each key. The prompts with the most similar keys are then appended to the input of the blocks of a Vision Transformer, with only the prompts being trained, while the backbone remains frozen.
- **CODA-Prompt (CP)** [32] allocates a set of prompt components for each task. Unlike L2P, this method calculates a weighted sum of the prompts, using the similarity values between the query and keys as weights. Additionally, the approach employs attention vectors with learnable values to filter the queries, thus enabling the model to focus only on their relevant features.

Additionally, we contextualize our results by presenting the following additional results: **Finetuning**, a lower bound where no continual learning technique is applied during training on the task sequence, **Finetuning with Asymmetric Cross-Entropy (Finetuning-ACE)**, which applies Cross-Entropy selectively on current task data as prescribed in [6] as a baseline passive countermeasure to forgetting, and **Joint training**, an upper-bound approach which assumes that all the examples are available and the model is trained on the entire dataset without learning incrementally.

4.2   Results

The performance disparity between the two benchmarks is evident from the results presented in Table 1. Across all methods, there is a notable struggle when dealing with the WebVision dataset, which we attribute to the high presence of annotation noise in the dataset. Moreover, on IIRC, prompting-based methods fail to achieve satisfactory performance, with the more sophisticated Coda-Prompt being outperformed by L2P. This is in contrast with traditional single-label benchmarks, which usually depict these advanced methods as best performers. We argue that such an outcome remarks the importance of our work for more complex scenarios.

As we move to replay methods, we observe that, in general, these approaches outperform prompting methods on both benchmarks. Notably, the ACE technique plays a crucial role in enhancing the model's performance. Indeed, this shows the importance in a multi-label setting of confining the contribution of each sample to avoid interference between tasks. Furthermore, the incorporation of logits distillation in DER++ leads to further improvements in the final performance.

Table 1. Final Average PWJS (and standard deviation) over the task sequence for the IIRC and WebVision benchmarks. The lower part of the table includes methods that require the memory buffer for rehearsal.

AR_f %	IIRC-CIFAR		Incr. WebVision	
Joint (UB)	86.83 (0.22)		25.12 (0.01)	
Finetuning-ACE	3.21 (0.01)		2.34 (0.36)	
Finetuning (LB)	0.24 (0.01)		18.81 (0.59)	
L2P [40]	3.76 (0.01)		0.31 (0.25)	
CP [32]	3.79 (0.01)		15.05 (1.45)	
Buffer Size	500	2000	2000	5000
ER [29, 31]	21.04 (1.81)	38.55 (0.91)	17.76 (0.18)	21.63 (0.18)
ER-ACE [6]	41.83 (1.75)	50.66 (0.87)	7.75 (1.78)	16.26 (0.34)
DER++-ACE [5]	41.66 (2.55)	61.82 (0.09)	8.97 (0.01)	17.03 (0.01)
SCAD (ours)	**45.23** (0.03)	**66.58** (0.02)	**20.02** (0.35)	**22.17** (0.60)

Finally, our proposed method shows superior performance across all settings and for all choices of buffer size, which remarks the importance of adapting and maintaining the knowledge from the pre-training.

5 Model Analysis

5.1 Contribution of selective distillation

To assess whether the inclusion of adapter neural networks has a positive impact on the model's performance, we attempted to remove them from the model and perform distillation using the entire class attention vectors. The results are presented in Table 2. As observed, the filtering of attention vectors performed through the binary masks produced by the adapters im-

Table 2. Final PWJS achieved by our method on the IIRC CIFAR100 benchmark with and without the use of adapters for generating the binary masks to filter the knowledge transfer.

	Binary masks	
	✓	✗
Final PWJS	66.58 (0.02)	63.68 (0.01)

proved the final performance. Consequently, it can be inferred that the adapter networks are able to effectively select and propagate only the relevant information from the teacher's attention vector to the student network.

5.2 Quantitatively measuring catastrophic forgetting in Multi-Label Continual Learning

While AR_f provides a clear snapshot of a model's accuracy at the end of training, we are also interested in measuring how the model's performance degrades as

Table 3. Multi-Label Adjusted Forgetting FG_f (the lower the better) of methods evaluated in our benchmarks.

FG_f	IIRC-CIFAR		Incr. WebVision	
Finetuning-ACE	65.40		74.45	
Finetuning (LB)	98.67		37.39	
L2P	97.29		66.41	
CP	97.77		44.06	
Buffer Size	500	2000	2000	5000
ER	69.98	41.57	34.33	25.64
ER-ACE	45.25	40.06	57.52	32.80
DER++-ACE	25.82	17.71	48.95	31.87
SCAD (ours)	24.36	16.40	28.37	24.92

the learner is subject to a growing amount of data. For this purpose, we take inspiration from traditional CL evaluation and formulate a MLCL version of the Adjusted Forgetting measure [3, 7] based on PWJS:

$$FG_f = \frac{1}{N-1} \sum_{m=1}^{N-1} \left[\frac{R_{*,m} - R_{N,m}}{R_{*,m}} \right]^+,$$

where $R_{*,m} = \max_{t \in \{m,\dots,N-1\}} R_{t,m}, \quad \forall m \in \{1,\dots,N-1\}$ and $[\cdot]^+ = max(\cdot, 0)$.

(10)

FG_f ranges between 0 and 100: $FG_f = 100$ indicates a result with no retention of performance on previous tasks; $FG_f = 0$ denotes no performance decrease on past tasks.

In Tab. 3, we report the measurements of FG_f for the experiments in Tab. 1. We clearly observe that the ACE technique proves highly effective in reducing the forgetting value on IIRC benchmark, indicating that computing the loss based on the classes of the current batch helps to preserve the parameters of the network that encode information from past tasks. On the WebVision benchmark, the situation is different, and for the rehearsal techniques the forgetting values are quite similar: we attribute this to the fact that all methods struggle to achieve good results in this setting.

Regarding prompting techniques, we note that the two tested methods experience similar levels of forgetting in the IIRC benchmark, while L2P exhibits significantly more forgetting on WebVision. In general, rehearsal-based methods tend to have lower forgetting values. This suggests that in a multi-label setting, having access to a memory of past examples is an effective strategy for retaining past knowledge.

6   Conclusions

In this work, we acknowledge the success of recent CL approaches in contrasting *catastrophic forgetting* in the standard single-label benchmarks. However, we reveal through a novel and challenging benchmark that state-of-the-art prompting CL methods fail in a multi-label setting, while good old-fashioned rehearsal techniques maintain good performance.

Hence, we present SCAD: a baseline model that tackles the challenges posed by the multi-label setting. Our proposal combines the principles of traditional rehearsal techniques with learning strategies based on knowledge transfer. SCAD exploits the finding that large-scale pretrained architectures are robust against catastrophic forgetting per se; therefore, we have designed a solution that discourages the model from deviating too far from its pretraining state.

By virtue of a comprehensive experimental analysis, we show that our method achieves superior performance compared to current state-of-the-art approaches in the Multi-Label Continual Learning setting. We recommend SCAD as a strong starting baseline to tackle the challenges that this new era of Continual Learning has to offer.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abdelsalam, M., Faramarzi, M., Sodhani, S., Chandar, S.: Iirc: Incremental implicitly-refined classification. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (2021)
2. Aljundi, R., Lin, M., Goujaud, B., Bengio, Y.: Gradient Based Sample Selection for Online Continual Learning. In: Advances in Neural Information Processing Systems (2019)
3. Bonicelli, L., Boschini, M., Frascaroli, E., Porrello, A., Pennisi, M., Bellitto, G., Palazzo, S., Spampinato, C., Calderara, S.: On the effectiveness of equivariant regularization for robust online continual learning. arXiv preprint arXiv:2305.03648 (2023)
4. Boschini, M., Bonicelli, L., Porrello, A., Bellitto, G., Pennisi, M., Palazzo, S., Spampinato, C., Calderara, S.: Transfer without forgetting. In: Proceedings of the European Conference on Computer Vision (2022)
5. Buzzega, P., Boschini, M., Porrello, A., Abati, D., Calderara, S.: Dark experience for general continual learning: a strong, simple baseline. Advances in Neural Information Processing Systems (2020)
6. Caccia, L., Aljundi, R., Asadi, N., Tuytelaars, T., Pineau, J., Belilovsky, E.: New insights on reducing abrupt representation change in online continual learning. In: International Conference on Learning Representations Workshop (2022)

7. Chaudhry, A., Dokania, P.K., Ajanthan, T., Torr, P.H.: Riemannian walk for incremental learning: Understanding forgetting and intransigence. In: Proceedings of the European Conference on Computer Vision (2018)
8. Chaudhry, A., Ranzato, M., Rohrbach, M., Elhoseiny, M.: Efficient lifelong learning with a-gem. International Conference on Learning Representations Workshop (2019)
9. Chen, X., Cao, Q., Zhong, Y., Zhang, J., Gao, S., Tao, D.: Dearkd: data-efficient early knowledge distillation for vision transformers. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (2022)
10. De Lange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G., Tuytelaars, T.: A continual learning survey: Defying forgetting in classification tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations Workshop (2021)
12. Farquhar, S., Gal, Y.: Towards Robust Evaluations of Continual Learning. In: International Conference on Learning Representations Workshop (2018)
13. Ganda, D., Buch, R.: A survey on multi label classification. Recent Trends in Programming Languages **5**(1), 19–23 (2018)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (2016)
15. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. Neural Information Processing Systems Workshops (2015)
16. Kim, C.D., Jeong, J., Kim, G.: Imbalanced continual learning with partitioning reservoir sampling. In: Proceedings of the European Conference on Computer Vision (2020)
17. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences (2017)
18. Krizhevsky, A., et al.: Learning multiple layers of features from tiny images. Tech. rep., Citeseer (2009)
19. Li, W., Wang, L., Li, W., Agustsson, E., Van Gool, L.: Webvision database: Visual learning and understanding from web data. arXiv preprint arXiv:1708.02862 (2017)
20. Li, Z., Hoiem, D.: Learning without forgetting. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)
21. Lin, S., Xie, H., Wang, B., Yu, K., Chang, X., Liang, X., Wang, G.: Knowledge distillation via the target-aware transformer. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (2022)
22. Lopez-Paz, D., Ranzato, M.: Gradient episodic memory for continual learning. Advances in Neural Information Processing Systems (2017)
23. McCloskey, M., Cohen, N.J.: Catastrophic interference in connectionist networks: The sequential learning problem. In: Psychology of learning and motivation (1989)
24. Mehta, S.V., Patil, D., Chandar, S., Strubell, E.: An empirical investigation of the role of pre-training in lifelong learning. In: International Conference on Learning Representations Workshop (2022)
25. Parisi, G.I., Kemker, R., Part, J.L., Kanan, C., Wermter, S.: Continual lifelong learning with neural networks: A review. Neural Networks (2019)

26. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (2021)
27. Rajeswar, S., Rodriguez, P., Singhal, S., Vazquez, D., Courville, A.: Multi-label iterated learning for image classification with label ambiguity. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (2022)
28. Ramasesh, V.V., Lewkowycz, A., Dyer, E.: Effect of scale on catastrophic forgetting in neural networks. In: International Conference on Learning Representations Workshop (2022)
29. Ratcliff, R.: Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. Psychological Review (1990)
30. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: icarl: Incremental classifier and representation learning. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (2017)
31. Robins, A.: Catastrophic forgetting, rehearsal and pseudorehearsal. Connection Science (1995)
32. Smith, J.S., Karlinsky, L., Gutta, V., Cascante-Bonilla, P., Kim, D., Arbelle, A., Panda, R., Feris, R., Kira, Z.: Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (2023)
33. Sorower, M.S.: A literature survey on algorithms for multi-label learning. Oregon State University, Corvallis **18**(1), 25 (2010)
34. Tarekegn, A.N., Giacobini, M., Michalak, K.: A review of methods for imbalanced multi-label classification. Pattern Recognition (2021)
35. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30** (2017)
36. Van de Ven, G.M., Tuytelaars, T., Tolias, A.S.: Three types of incremental learning. Nature Machine Intelligence (2022)
37. Vitter, J.S.: Random sampling with a reservoir. ACM Transactions on Mathematical Software (1985)
38. Wang, J., Cao, M., Shi, S., Wu, B., Yang, Y.: Attention probe: Vision transformer distillation in the wild. In: IEEE International Conference on Acoustics, Speech and Signal Processing (2022)
39. Wang, Z., Zhang, Z., Ebrahimi, S., Sun, R., Zhang, H., Lee, C.Y., Ren, X., Su, G., Perot, V., Dy, J., et al.: Dualprompt: Complementary prompting for rehearsal-free continual learning. In: Proceedings of the European Conference on Computer Vision (2022)
40. Wang, Z., Zhang, Z., Lee, C.Y., Zhang, H., Sun, R., Ren, X., Su, G., Perot, V., Dy, J., Pfister, T.: Learning to prompt for continual learning. In: Proceedings of the European Conference on Computer Vision (2022)
41. Yang, Z., Li, Z., Zeng, A., Li, Z., Yuan, C., Li, Y.: Vitkd: Practical guidelines for vit feature knowledge distillation. International Conference on Learning Representations Workshop (2023)