Abstract

Incremental learning enables artificial agents to learn from sequential data. While important progress was made by exploiting deep neural networks, incremental learning remains very challenging. This is particularly the case when no memory of past data is allowed and catastrophic forgetting has a strong negative effect. We tackle class-incremental learning without memory by adapting prediction bias correction, a method which makes predictions of past and new classes more comparable. It was proposed when a memory is allowed and cannot be directly used without memory, since samples of past classes are required. We introduce a two-step learning process which allows the transfer of bias correction parameters between reference and target datasets. Bias correction is first optimized offline on reference datasets which have an associated validation memory. The obtained correction parameters are then transferred to target datasets, for which no memory is available. The second contribution is to introduce a finer modeling of bias correction by learning its parameters per incremental state instead of the usual past vs. new class modeling. The proposed dataset knowledge transfer is applicable to any incremental method which works without memory. We test its effectiveness by applying it to four existing methods. Evaluation with four target datasets and different configurations shows consistent improvement, with practically no computational and memory overhead.

1. Introduction

Incremental learning (IL) enables the adaptation of artificial agents to dynamic environments in which data is presented in streams. This type of learning is needed when access to past data is limited or impossible, but is affected by catastrophic forgetting [24]. This phenomenon consists in a drastic performance drop for previously learned information when ingesting new data. Works such as [5, 9, 14, 28, 33, 34, 36] alleviate the effect of forgetting by replaying past data samples when updating deep incremental models in class IL. A term which adapts knowledge distillation [13] to IL is usually exploited to reinforce the representation of past classes [21]. When such a memory is allowed, class IL actually becomes an instance of imbalanced learning [11]. New classes are favored since they are represented by a larger number of images. As a result, classification bias correction methods were successfully introduced in [5, 33, 34].

While important progress was made when a fixed memory is allowed, this is less the case for class IL without memory. This last setting is more challenging and generic since no storage of past samples is allowed. In absence of memory, existing methods become variants of Learning without Forgetting (LwF) [21] with different formulations of the distillation term. Importantly, bias correction methods become inapplicable without access to past classes samples.

Our main contribution is to enable the use of the bias correction methods, such as the BiC layer from [33], in class IL without memory. We focus on this approach because it is both simple and effective in IL with memory [3, 23]. Authors of BiC [33] use a validation set which stores samples of past classes to optimize parameters. Instead, we learn correction parameters offline on a set of reference datasets and then transfer them to target datasets. The method is thus abbreviated TransIL. The intuition is that, while datasets are different, optimal bias correction parameters are stable enough to be transferable between them. We illustrate the approach in Figure 1, with the upper showing the IL process with a reference dataset. A memory for the validation samples needed to optimize the bias correction layer is allowed since the training is done offline. The lower part of the figure presents the incremental training of a target dataset. The main difference with the standard memoryless IL training comes from the use of a bias correction layer optimized on the reference dataset. Its introduction leads to an improved comparability of prediction scores for past and new classes. Note that the proposed method is applicable to any class IL method, since it only requires the availability of raw predic-
Figure 1: Illustration of TransIL, our proposed method, depicting states from 1 to \( s \) for a reference and a target dataset. The model \( M \) is updated in each state with data from new classes. States from 1 to \( s - 1 \) are faded to convey the fact that knowledge learned in them is affected by catastrophic forgetting. The class IL process is first launched offline on the reference dataset where \( a dBiC \), our proposed bias correction layer, is trained using a validation memory which stores samples for past and new classes. Class IL is then applied to the target dataset, but without class samples shared across states since a memory is not allowed in this scenario. The set of optimal parameters of \( a dBiC \) obtained for the reference dataset is transferred to the target dataset. This is the only information shared between the two processes and it has a negligible memory footprint. The transfer of parameters enables the use of bias correction for the target dataset. The final predictions obtained in state \( s \) are improved compared to the direct use of \( M_s \) predictions, since the bias in favor of new classes is reduced.

The second contribution is to refine the definition of the bias correction layer introduced in [33]. The original formulation considers all past classes equally in the correction process. With [23], we hypothesize that the degree of forgetting associated to past classes depends on the initial state in which they were learned. Consequently, we propose \( A daptive \ BiC (a dBiC) \), an optimization procedure which learns a pair of parameters per IL state instead of a single pair of parameters as proposed in [33].

We provide a comprehensive evaluation of TransIL by applying it to four backbone class IL methods. Four target datasets with variable domain shift with respect to reference datasets and different numbers of IL states are used. An improvement of accuracy is obtained for almost all tested configurations. The additional memory needs are negligible since only a compact set of correction parameters is stored. Code and data needed for reproducibility are provided\(^1\).

2. Related work

Incremental learning is a longstanding machine learning task [10, 22, 30] which witnessed a strong growth in interest after the introduction of deep neural networks. It is named differently as continual, incremental or lifelong learning depending on the research communities which tackle it and the setting of the problem. However, the objective is common: enable artificial agents to learn from data which is fed sequentially to them. Detailed reviews of existing approaches are proposed, among others, in [3, 20, 23, 25]. Here, we analyze works most related to our proposal, which tackle class IL and keeps memory and computational requirements constant, or nearly so, during the IL process. We focus particularly on methods which address bias in favor of new classes [23] and were designed for class IL with memory.

The wide majority of class IL methods make use of an information preserving penalty [8]. This penalty is generally implemented as a loss function which reduces the divergence between the current model and the one learned in the preceding IL state. Learning without forgetting (LwF) [21] is an early work which tackles catastrophic forgetting in deep neural nets. It exploits knowledge distillation [13] to preserve information related to past classes during incremental model updates. Less-forgetting learning [16] is a closely related method. Past knowledge is preserved by freezing the softmax layer of the source model and updating the model using a loss which preserves the representation of past data. The two methods aim to propose a good compromise between plasticity, needed for new data representation, and stability, useful for past information preservation.

However, they require the storage of the preceding model in order to perform distillation toward the model which is currently learned. This requirement can be problematic if the memory of artificial agents is constrained.

\( LwF \) was initially used for task-based continual learning and was then widely adopted as backbone for class IL. \( iCaRL \) [28] exploits \( LwF \) and a fixed-size memory of past samples to alleviate catastrophic forgetting. In addition, a nearest-mean-of-exemplars classifier is introduced in order to reduce the bias in favor of new classes. \( E2EIL \) [5] corrects bias by adding a second fine-tuning step with the same number of samples for each past and new class. The learning of a unified classifier for incremental learning re-

\(^1\)https://github.com/HabibSlim/DKT-for-CIL
balancing (LUCIR) is proposed in [14]. The authors introduce a cosine normalization layer in order to make the magnitudes of past and new class predictions more comparable. The maintenance of both discrimination and fairness is addressed in [34]. The ratio between the mean norm of past and new class weights is applied to the weights of new classes, to make their associated predictions more balanced. Bias Correction (BiC) [33] exploits a supplementary linear layer to rebalance predictions of a deep incremental model. A validation set is used to optimize the parameters of this linear layer, which modifies the predictions of the deep model learned in a given incremental state. We tackle two important limitations of existing bias correction methods. First, they are inapplicable without memory because they require the presence of past class samples. We propose to transfer bias correction layer parameters between datasets to address this problem. Second, the degree of forgetting associated to past classes is considered equivalent, irrespective of the initial state in which they were learned. This is problematic insofar as a recency bias, which favors classes more recently, appears in class IL [23]. We refine the linear layer from [33] to improve the handling of recency bias.

The improvement of the component which handles model stability also received strong attention in class IL. Learning without memorizing [8] is inspired by LwF and adds an attention mechanism to the distillation loss. This new term improves the preservation of information related to base classes. A distillation component which exploits information from all past states and from intermediate layers of CNN models was introduced in [36]. LUCIR [13] distills knowledge in the embedding space rather than the prediction space to reduce forgetting and adds an inter-class separation component to better distinguish between past and new class embeddings. PODNet [9] employs a spatial-based distillation loss and a representation which includes multiple proxy vectors for classes to optimize distillation. In [31], a feature map transformation strategy with additional network parameters is proposed to improve class separability. Model parameters are shared between global and task-specific parameters and only the latter are updated at each IL state to improve training times. Feature transformation using a dedicated MLP is introduced in [15]. This approach only stores features but adds significant memory to store the additional MLP. Recently, the authors of [19] argued for the importance of uncertainty and of attention mechanisms in the modeling of past information in class IL. These different works provide a performance gain compared to the original adaptation of distillation for continual learning [21] in class IL with memory.

The utility of distillation in a class IL scenario was recently questioned. It is shown [23, 27] that competitive results are obtained if a fixed-size memory is allowed for large-scale datasets. The distillation component is removed in [23] and IL models are updated using fine-tuning. A simpler approach is tested in [27], where the authors learn models independently for each incremental state after balancing class samples. The usefulness of distillation was also challenged in absence of a memory [2] where standardization of initial weights (SIW), learned when a class was first encountered, was proposed in [2]. The freezing of initial weights was tested in [23] and also provides significant improvements. It is thus interesting to also apply the proposed approach to methods which do not exploit distillation.

Our method is globally inspired by existing works which transfer knowledge between datasets. We mentioned knowledge distillation [13] which is widely used in IL. Dataset distillation [32] encodes large datasets into a small set of synthetic data points to make the training process more efficient. Hindsight anchor learning [6] learns an anchor per class to characterize points which would maximize forgetting in later IL states. While the global objective is similar, our focus is different since only a very small number of parameters are transferred from reference to target datasets to limit catastrophic forgetting on the latter.

3. Dataset knowledge transfer for class IL

In this section, we describe the proposed approach which transfers knowledge between datasets in class IL without memory. We first propose a formalization of the problem and then introduce an adaptation of a prediction bias correction layer used in class IL with memory. Finally, we introduce the knowledge transfer method which enables the use of the bias correction layer in class IL without memory.

3.1. Class-incremental learning formalization

We adapt the class IL definition from [5, 14, 28] to a setting without memory which includes a sequence of $S$ states. The first one is called initial state and the $S−1$ remaining states are incremental. A set of $P_s$ new classes is learned in the $s^{th}$ state. IL states are disjoint and $P_i \cap P_j = \emptyset \forall i, j \in [1, S], i \neq j$. A model $M_1$ is initially trained on a dataset $D_1 = \{(X_i^1, Y_i^1) : j \in P_1\}$, where $X_i^1$ and $Y_i^1$ are the sets of training images and their labels. We note $N_s$ the set of all classes seen until the $s^{th}$ state included. Thus, $N_1 = P_1$ initially, and $N_s = N_{s−1} \cup P_s = P_1 \cup P_2 \cup \ldots \cup P_{s−1} \cup P_s$ for subsequent states. $M_s$ is updated with an IL algorithm $A$ using $D_s = \{(X_j^s, Y_j^s) : j \in P_s\}$. $D_s$ includes only new classes samples, but $M_s$ is evaluated on all classes seen so far ($j \in N_s$). This makes the evaluation prone to catastrophic forgetting due to the lack of past exemplars [3, 23].

3.2. Adaptive bias correction layer

The unavailability of past class exemplars when updating the incremental models leads to a classification bias toward new classes [33, 34]. We illustrate this in Figure 2 (left) by plotting mean prediction scores per state for the CIFAR-100
mental states are variable, which confirms that the amount shows that mean scores of classes learned in different incre-
cause samples of past classes are not allowed. Figure 2 (in class IL without memory, the scenario explored here, be-
past classes. We remind that Equation 1 is not applicable
timized using a validation set made of samples of new and
then frozen and calibration parameters ($\alpha$)
those of past classes. The deep model is first updated us-
s= 10
state to make them more comparable to
th state, and $\beta^k_s$ if $k = s$ ($\alpha^k_s$, $\beta^k_s$ pair is shared between all classes first learned in the same state.
These parameters are optimized on a validation set using the cross-entropy loss, defined for one data point $(x, y)$ as:
$$L(q_s, y) = -\sum_{k=1}^{s} \sum_{i=1}^{|P_k|} \delta_{y=\hat{y}} \log (q^k_{s,i})$$ (3)
where $y$ is the ground-truth label, $\hat{y}$ is the predicted label, $\delta$ is the Kronecker delta, and $q_s$ is the softmax output for the sample corrected via Equation 2, defined as:
$$q_s = \sigma \left( [\alpha^1_s o^1_s + \beta^1_s \cdot 1 ; \ldots ; \alpha^s_s o^s_s + \beta^s_s \cdot 1] \right)$$ (4)
where $\sigma$ is the softmax function.
All $\alpha^k_s$, $\beta^k_s$ pairs are optimized using validation samples from classes in $N_s$. We compare $adBiC$ over $BiC$ for our class IL setting in the evaluation section and show that the adaptation proposed here has a positive effect.

3.3. Transferring knowledge between datasets
The optimization of $\alpha$ and $\beta$ parameters is impossible in class IL without memory, since exemplars of past classes are unavailable. To circumvent this problem, we hypothe-
size that optimal values of these parameters can be trans-
ferred between reference and target datasets, noted $D^r$ and $D^t$ respectively. The intuition is that these values are suffi-
ciently stable despite dataset content variability. We create a set of reference datasets and perform a modified class IL training for them using the procedure described in Algo-
ithm 1. The modification consists in exploiting a validation set which includes exemplars of classes from all incre-
mental states. Validation set storage is necessary in order to opti-
imize the parameters from Equation 2 and is possible since reference dataset training is done offline. Note that
backbone incremental models for $D^t$ are trained without
memory in order to simulate the IL setting of target datasets $D^t$. We then store bias correction parameters optimized for reference datasets in order to perform transfer toward tar-
get datasets without using a memory. For each incremen-
tal state, we compute the average of $\alpha$ and $\beta$ values over all reference datasets. The obtained averages are used for score rectification on target datasets. This transfer uses the
procedure described in Algorithm 2. The memory needed to store transferred parameters is negligible since we need $2 \times (2 + 3 + \ldots + S) = (S+2) \times (S-1)$ floats for each dataset and $S$ value. For $S = \{5, 10, 20\}$ states, we thus only store 28, 108 and 418 floating-point values respectively.

Algorithm 1: Optimization of calibration parameters

\begin{verbatim}
inputs : $A, D^r_s$ for $s \in [1, S]$ \triangledown reference dataset
randomly initialize $M_1$;
$M^*_1 \leftarrow \text{train}(A; M_1, D^r_1)$;
for $s = 2 \ldots S$ do
    $M^*_s \leftarrow \text{update}(A; M^*_{s-1}, D^r_s)$;
    $\alpha^*_s \leftarrow 1, \beta^*_s \leftarrow 0$ for each $k \in [1, s]$;
    foreach $(x, y) \in D^r_s$ \triangledown validation set
        $o_s \leftarrow M^*_s(x)$;
        for $k = 1 \ldots s$ do
            $o^k_s \leftarrow \text{adBiC}(o^k_s) = \alpha^k_s o^k_s + \beta^k_s \cdot 1$;
        end
        $q_s \leftarrow \sigma(o_s)$;
        loss $\leftarrow L(q_s, y)$;
        $(\alpha^1_s, \beta^1_s, \ldots, \alpha^s_s, \beta^s_s) \leftarrow \text{optimize}(\text{loss})$;
    end
end
\end{verbatim}

Algorithm 2: adBiC inference

\begin{verbatim}
inputs : $A, (\alpha^*_k, \beta^*_k)$ averaged on reference datasets
for each $s \in [1, S], k \in [1, s]$ \triangledown target dataset
inputs : $D^t_s$ for $s \in [1, S]$ \triangledown test set
randomly initialize $M_1$;
$M^*_1 \leftarrow \text{train}(A; M_1, D^t_1)$;
for $s = 2 \ldots S$ do
    $M^*_s \leftarrow \text{update}(A; M^*_{s-1}, D_s)$;
    foreach $(x, y) \in D^t_s$ \triangledown test set
        $o_s \leftarrow M^*_s(x)$;
        for $k = 1 \ldots s$ do
            $o^k_s \leftarrow \text{adBiC}(o^k_s) = \alpha^k_s o^k_s + \beta^k_s \cdot 1$;
        end
        $q_s \leftarrow \sigma(o_s)$;
        $\hat{y} \leftarrow \text{argmax}_{y \in [1, N_i]} (q_s)$ \triangledown inference
    end
end
\end{verbatim}

In Figure 3, we illustrate optimal parameters obtained across $R = 10$ reference datasets which are further described in Section 4. We plot $\alpha^k$ and $\beta^k$ values learned after $S = 10$ IL states, using LwF [21] and LUCIR [14] methods. Mean and standard deviations are presented for past and current incremental states in the final state of the IL process. The parameter ranges from Figure 3 confirm that, while optimal values do vary across datasets, this variation is rather low and calibration profiles remain similar. Consequently, parameters are transferable. When $R > 1$, a transfer function is needed to apply the parameters learned on reference datasets to a target dataset. We transfer parameters using the averaged $\alpha^k$ and $\beta^k$ values, obtained for the set of $D^r$. In Section 4, we evaluate this transfer against an upper-bound oracle which selects the best $D^r$ in each state.

Figure 3: Averaged $\alpha^k$ (left) and $\beta^k$ (right) values computed for $R = 10$ reference datasets using LwF and LUCIR, at the end of an incremental process with $S = 10$ states.

4. Evaluation

In this section, we discuss: (1) the reference and target datasets, (2) the backbone methods to which bias correction is applied and (3) the analysis of the obtained results. The evaluation metric is the average top-1 accuracy of the IL process introduced in [28], which combines the accuracy obtained for individual incremental states. Following [5], we discard the accuracy of the first state since it is not incremental. We use a ResNet-18 backbone whose implementation details are provided in the supp. material.

4.1. Datasets

Reference datasets. The preliminary analysis from Figure 3 indicates that bias correction parameters are rather stable for different reference datasets. It is interesting to use several such datasets in order to stabilize averaged bias correction parameters. In our experiments, we use 10 reference datasets, each including 100 randomly chosen leaf
classes from ImageNet [7], with a 500/200 train/val split per class. There is no intersection between these datasets, as each class appears only in one of them.

**Target datasets.** We test TransIL with four target datasets. They were selected to include different types of visual content and thus test the robustness of the parameter transfer. The class samples from the target datasets are split into 500/100 train/test subsets respectively. There is no intersection between the classes from the reference datasets and the two target datasets which are sampled from ImageNet. We describe target datasets briefly hereafter and provide details in the supplementary material: • CIFAR-100 [18] - object recognition dataset. It focuses on common-sense classes and is relevant for basic level classification in the sense of [29].
- IMN-100 - subset of ImageNet [7] which includes 100 randomly selected leaf classes. It is built with the same procedure used for reference datasets and is thus most similar to them. IMN-100 is relevant for fine-grained classification with a diversity of classes.
- BIRDS-100 - uses 100 bird classes from ImageNet [7]. It is built for domain fine-grained classification.
- FOOD-100 - uses 100 food classes from Food-101 [4]. It is a fine-grained and domain-specific dataset and is interesting because it is independent from ImageNet.

4.2. Backbone incremental learning methods

We apply adBiC on top of four backbone methods which are usable for class IL without memory:
- LwF [28] - version of the original method from [21] which exploits distillation to reduce catastrophic forgetting for past classes.
- LUCIR [14] - distillation-based approach which uses a more elaborate way of ensuring a good balance between model stability and plasticity. We use the CNN version because it is adaptable to our setting.
- FT+ [23] - fine-tuning in which past classes weights are not updated to reduce catastrophic forgetting.
- SIW [2] - similar to FT+, but with class weights standardization added to improve the comparability of prediction between past and new classes.

We compare adBiC to BiC, the original linear layer from [33]. We also provide results with an optimal version of adBiC, which is obtained via an oracle-based selection of the best performing reference dataset for each IL state. This oracle is important as it indicates the potential supplementary gain obtainable with a parameter selection method more refined than the proposed one. Finally, we provide results with Joint, a training from scratch with all data available at all times. This is an upper bound for all IL methods.

4.3. Overall results

Results from Figure 2 (right) indicate that the degree of forgetting depends on the initial state in which classes were first learned. Applying calibration parameters learned on reference datasets clearly reduces the imbalance of mean prediction scores and the bias toward recent classes. Results from Table 1 show that our method improves the performance of baseline methods for all but two of the configurations evaluated. The best overall performance before bias correction is obtained with LwF. This result confirms the conclusions of [2, 23] regarding the strong performance of LwF in class IL without memory for medium-scale datasets. With adBiC, LUCIR performs generally better than LwF for \( S = 5 \) and \( S = 10 \), while LwF remains stronger with \( S = 20 \) states. Results are particularly interesting for LUCIR, a method for which adBiC brings consistent gains (up to 16 accuracy points) in most configurations. Table 1 shows that adBiC also improves the results of LwF in all configurations, albeit to a lesser extent compared to LUCIR. Interestingly, improvements for LwF are larger for \( S = 20 \) states. This is the most challenging configuration since the model is more prone to forgetting. FT+ [23] and SIW [2] remove the distillation component for the class IL training process and exploit the weights of past classes learned in their initial state. adBiC improves results for these two methods in all but one configuration. However, their global performance is significantly lower than that of LwF and LUCIR, the two methods which make use of distillation. This result confirms the finding from [2] regarding the usefulness of the distillation term exploited by LwF and LUCIR to stabilize IL training for medium scale datasets.

Results from Table 1 highlight the effectiveness of adBiC compared to BiC. adBiC has better accuracy in all tested configurations, with the most important gain over BiC obtained for LUCIR. It is also worth noting that adBiC improves results for SIW and FT+ in most configurations, while the corresponding results of BiC are mixed. The comparison of adBiC and BiC validates our hypothesis that a finer-grained modeling of forgetting for past states is better compared to a uniform processing of them. It would be interesting to test the usefulness of adBiC in the class IL with memory setting originally tested in [33].

We also compare adBiC, which uses averaged \( \alpha \) and \( \beta \) parameters, with an oracle selection of parameters (+). The performance of adBiC is close to this upper bound for all tested methods. This indicates that averaging parameters is an effective way to aggregate parameters learned from reference datasets. However, it would be interesting to investigate more refined ways to transfer parameters from reference to target datasets to further improve performance.

The comparison of target datasets shows that the gain brought by adBiC is the largest for IMN-100, followed by BIRDS-100, CIFAR-100 and FOOD-100. This is intuitive
due to variable forgetting for past states and to the variable in different states. Some variation subsists and is notably prediction scores to give equal chances to classes learned bias correction should ideally provide fully balanced mean LUCIR tion is particularly interesting for predictions associated to IL states are more balanced (right), results from Table 1. The correction is effective since the pre-

IL without memory remains an open challenge. less representative for larger values of \( S \). This indicates that class IL without memory remains an open challenge.

Except for \( LwF \), \( adBiC \) gains are larger for \( S = \{5, 10, 20\} \) compared to \( S = 20 \). This result is consistent with past findings reported for bias correction methods \[23, 33\]. It is mainly explained by the fact that the size of validation sets needed to optimize \( adBiC \) parameters is smaller and thus less representative for larger values of \( S \). A larger number of states leads to a higher degree of forgetting. This makes the IL training process more challenging and also has a negative effect on the usefulness of the bias correction layer.

Figure 2 provides a qualitative view of the effect of \( adBiC \) for \( LwF \) and \( LUCIR \) which complements numerical results from Table 1. The correction is effective since the predictions associated to IL states are more balanced (right), compared to the raw predictions (left). The effect of calibration is particularly interesting for \( LUCIR \), where mean prediction scores are balanced for states 3 to 10. We note that bias correction should ideally provide fully balanced mean prediction scores to give equal chances to classes learned in different states. Some variation subsists and is notably due to variable forgetting for past states and to the variable difficulty of learning different visual classes.

4.4. Robustness of dataset knowledge transfer

We complement the results presented in Table 1 with two experiments which further evaluate the robustness of \( adBiC \). First, we test the effect of a different number of training images per class for reference and target datasets. We remove 50% of training images for target datasets to test the transferability in this setting. The obtained results, presented in Table 2, indicate that performance gains are systematic for \( LwF \) and \( LUCIR \), albeit lower compared to results in Table 1. Results are more mixed for \( SIW \) and \( FT^+ \), but \( adBiC \) still has a positive effect in the majority of cases. This experiment shows that the proposed dataset knowledge transfer approach is usable for reference and target datasets which have a different number of training samples per class. However, maintaining a low difference in dataset sizes is preferable in order to keep the transfer effective.

Second, we assess the robustness of the method with respect to \( R \), the number of available reference datasets. We select the Food-100 dataset because it has the largest domain shift with respect to reference datasets and is thus the most suitable for this experiment. We vary \( R \) from 1 to 9 and perform transfer with ten random samplings for each \( R \) value. Results obtained for \( LUCIR \) are reported in Table 3. Accuracy levels are remarkably stable for different values of \( R \) and significant gains are observed even when using a single reference dataset. These results confirm that parameter transfer is effective even with few reference datasets, which is interesting considering that the computational cost
of offline training is also reduced. Results with other methods for CIFAR-100 are provided in the supp. material.

### 5. Conclusion

We introduced a method which enables the use of bias correction methods for class IL without memory. This IL scenario is challenging, because catastrophic forgetting is very strong in the absence of memory. The proposed method TransIL transfers bias correction parameters learned offline from reference datasets toward target datasets. Since reference dataset training is done offline, a validation memory which includes exemplars from all incremental states can be exploited to optimize the bias correction layer. The evaluation provides comprehensive empirical support for the transferability of bias correction parameters. Performance is improved for all but two of the configurations tested, with gains up to 16 top-1 accuracy points. Robustness evaluation shows that parameter transfer is efficient when only a small number of reference datasets is used for transfer. It is also usable when the number of training images per class in target datasets is different from that of available reference datasets. These last two findings are important in practice since the same reference datasets can be exploited in different incremental configurations. A second contribution relates to the modeling of the degree of forgetting associated to past states. While recency bias was already acknowledged [23], no difference was made between past classes learned in different IL states [33]. This is in part due to validation memory constraints which appear when the bias correction layer is optimized during the incremental process. Such constraints are reduced here since reference datasets training is done offline and a refined definition of the BiC layer with specific parameters for each past state becomes possible. The comparison of the standard and of the proposed definition of the bias correction layer is favorable to the latter. The reported results encourage us to pursue the work presented here. First, the parameter transfer is done using average values of parameters learned on reference datasets. A finer-grained transfer method will be tested to get closer to the oracle results reported in Table 1. The objective is to automatically select the best reference dataset in each IL state of a target dataset. Second, we exploit an adapted version of a bias correction method which was initially designed for class IL with memory. We will explore the design of methods which are specifically created for class IL without memory. Finally, while distillation-based methods outperformed methods which do not use distillation for the datasets tested here, existing results [2, 23] indicate that the role of distillation diminishes with scale. It would be interesting to verify this finding for our method.

### Acknowledgments

This work was supported by the European Commission under European Horizon 2020 Programme, grant number 951911 – AI4Media, and partially supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01) funded by the French program "Investissements d’Avenir". This publication was made possible by the use of the FactoryIA supercomputer, financially supported by the Ile-de-France Regional Council.

---

**Table 2**: Average top-1 IL accuracy with 50% of training images for target datasets. Gains are in green, losses are in red.

| Method | CIFAR-100 | IMN-100 | BIRDS-100 | FOOD-100 |
|--------|-----------|---------|-----------|----------|
|        | S = 5     | S = 10  | S = 20    |          |
| LwF [1] | 41.3      | 33.3    | 23.3      |          |
| LwB/C   | 42.1 ± 0.8| 34.8 ± 1.5| 25.0 ± 1.7|          |
| LUCIR   | 43.5      | 27.8    | 16.6      |          |
| SIW [1] | 31.7      | 21.6    | 13.7      |          |
| FT+     | 30.4      | 21.5    | 12.9      |          |
|         | Raw       | R = 1   | R = 2     | R = 3    |
|         |           | R = 4   | R = 4     | R = 5    |
|         |           | R = 6   | R = 7     | R = 8    |

**Table 3**: Average top-1 incremental accuracy of adBiC-corrected models trained incrementally on FOOD-100 with LUCIR, for S = {5, 10, 20} states, while varying the number R of reference datasets. For R ≤ 9, results are averaged across 10 random samplings of the reference datasets (hence the std values). Raw is the accuracy of LUCIR without bias correction.
References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems. In OSDI, 2015.

[2] Eden Belouadah, Adrian Popescu, and Ioannis Kanellos. Initial classifier weights replay for memoryless class incremental learning. In BMVC, 2020.

[3] Eden Belouadah, Adrian Popescu, and Ioannis Kanellos. A comprehensive study of class incremental learning algorithms for visual tasks. Neural Networks, 2021.

[4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In ECCV, 2014.

[5] Francisco M. Castro, Manuel J. Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In ECCV, 2018.

[6] Arslan Chaudhry, Albert Gordo, Puneet K Dokania, Philip Torr, and David Lopez-Paz. Using hindsight to anchor past knowledge in continual learning. In AAAI, 2021.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In CVPR, 2009.

[8] Prithviraj Dhar, Rajat Vikram Singh, Kuan-Chuan Peng, Ziyan Wu, and Rama Chellappa. Learning without memorizing. In CVPR, 2021.

[9] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In ECCV, 2020.

[10] Bernd Fritzke. A growing neural gas network learns topologies. In NeurIPS, 1994.

[11] Haibo He and Edwardo A. Garcia. Learning from imbalanced data. Transactions on Knowledge and Data Engineering, 2009.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016.

[13] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In NeurIPS Deep Learning Workshop, 2014.

[14] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In CVPR, 2019.

[15] Ahmet Iscen, Jeffrey Zhang, Svetlana Lazebnik, and Cordelia Schmid. Memory-efficient incremental learning through feature adaptation. In ECCV, 2020.

[16] Heechul Jung, Jeongwoo Ju, Minju Jung, and Junmo Kim. Less-forgetting learning in deep neural networks. 2016. arXiv:1607.00122.

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In ICLR, 2015.

[18] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[19] Vinod K Kurmi, Badri N Patro, Venkatesh K Subramanian, and Vinay P Namboodiri. Do not forget to attend to uncertainty while mitigating catastrophic forgetting. In WACV, 2021.

[20] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Gregory G. Slabaugh, and Tinne Tuytelaars. Continual learning: A comparative study on how to defy forgetting in classification tasks. TPAMI, 2019.

[21] Zhizhong Li and Derek Hoiem. Learning without forgetting. In ECCV, 2016.

[22] Thomas Martinetz, Stanislav G. Berkovich, and Klaus Schulten. Neural-gas network for vector quantization and its application to time-series prediction. Transactions on Neural Networks and Learning Systems, 1993.

[23] Marc Masana, Xiaole Liu, Bartlomiej Twardowski, Mikel Menta, Andrew D. Bagdanov, and Joost van de Weijer. Class-incremental learning: survey and performance evaluation on image classification, 2021. arXiv:2010.15277.

[24] Michael Mccluskey and Neil J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. The Psychology of Learning and Motivation, 1989.

[25] German Ignacio Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. Neural Networks, 2019.

[26] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In NeurIPS, 2019.

[27] Ameya Prabhu, Philip HS Torr, and Puneet K Dokania. Gdumb: A simple approach that questions our progress in continual learning. In ECCV, 2020.

[28] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. In CVPR, 2017.

[29] Eleanor Rosch. Principles of categorization. Concepts: core readings, 1999.

[30] Nadeem Ahmed Syed, Huan Liu, and Kah Kay Sung. Handling concept drifts in incremental learning with support vector machines. In SIGKDD, 1999.

[31] Vinay Kumar Verma, Kevin J. Liang, Nikhil Mehta, Piyush Rai, and Lawrence Carin. Efficient feature transformations for discriminative and generative continual learning. In CVPR, 2021.
[32] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation, 2018. arXiv:1811.10959.

[33] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In CVPR, 2019.

[34] Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. Maintaining discrimination and fairness in class incremental learning. In CVPR, 2020.

[35] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. TPAMI, 2017.

[36] Peng Zhou, Long Mai, Jianming Zhang, Ning Xu, Zuxuan Wu, and Larry S. Davis. M2KD: multi-model and multi-level knowledge distillation for incremental learning. In BMVC, 2020.
Supplementary Material for "Dataset Knowledge Transfer for Class-Incremental Learning without Memory"

Habib Slim1, Eden Belouadah1,2, Adrian Popescu1, Darian Onchis3
1 Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
2 IMT Atlantique, Lab-STICC, team RAMBO, UMR CNRS 6285, F-29328, Brest, France
3 West University of Timisoara, Timisoara, Romania

habib.slim@grenoble-inp.org, {eden.belouadah, adrian.popescu}@cea.fr, darian.onchis@e-uvt.ro

Introduction

In this supplementary material, we provide:
• implementation details of adBiC and the tested backbone IL methods (Section A).
• classes lists of target datasets used for evaluation (Section B).
• additional figures highlighting the effects of adBiC on the tested backbone methods (Section C).
• additional tables for the robustness experiment presented in Section 4.4 of the paper (Section D).
• results on PLACES-100 dataset (Table 5).
• additional accuracy plots for all methods and datasets (Figures 5 and 6).

A. Implementation details

A.1. Backbone IL methods

For LUCIR [14] and SIW [2], we used the original codes provided by the authors. For LwF, we adapted the multi-class Tensorflow [1] implementation from [28] to IL without memory. For FT+, we implemented the method by replacing classification weights of each class group by their initial weights learned when classes were encountered for the first time.

The weight decay is set to $5 \cdot 10^{-4}$ and the momentum to 0.9. Note that since no memory of past classes is available, the margin ranking loss is unusable and thus removed. SIW and FT+ are both trained with the same set of hyperparameters. Following [2], models are trained from scratch for 300 epochs in the first non-incremental state, using the SGD optimizer with momentum 0.9.

The base learning rate is set to 0.1, and is divided by 10 when the loss plateaus for 60 epochs. The weight decay is set to $5 \cdot 10^{-4}$. For incremental states, the same hyperparameters are used, except for the number of epochs which is reduced to 70 and for the learning rate which is divided by 10 when the loss plateaus for 15 epochs.

A.2. Adaptive bias correction

The correction of output scores is done in the same way for all methods. After the extraction of scores and labels for each model, batches are fed into a PyTorch [26] module which performs the optimization of adBiC parameters, or the transfer of previously learned parameters. Following [23], BiC and adBiC layers are implemented as pairs of parameters and optimized simply through backpropagation.

Parameters $\alpha^k_s, \beta^k_s$ corresponding to each incremental state $s$ are optimized for 300 epochs, with the Adam [17] optimizer and a starting learning rate of $10^{-3}$. An L2-penalty is added to the loss given in Equation 3 of the main paper, with a lambda of $5 \cdot 10^{-3}$ for $\alpha$ parameters and $5 \cdot 10^{-2}$ for $\beta$ parameters.
B. Datasets description

We provide in Table 4 the lists of classes contained in each of the target datasets we used for evaluation. Overall, IMN-100, the randomly sampled set of 100 leaf classes from Imagenet [7], is more diversified than CIFAR-100, which mostly contains animal classes. IMN-100 is visually varied between different types of objects, foods, animals, vehicles, clothes and events. CIFAR-100 contains, in addition to animals, some types of objects and vehicles.

The FOOD-100 dataset and BIRDS-100 (extracted from ImageNet [7]) are more specialized than IMN-100 and CIFAR-100 and are thus useful to test the robustness of our method on finer-grained datasets. PLACES-100 and FOOD-100 are target datasets which have a larger domain shift with ImageNet classes, and are thus useful to test the robustness of our method against domain variation. Similarly to IMN-100, reference datasets are random subsets of ImageNet leaves. They contain various object types and are useful for knowledge transfer.

C. Effects of adaptive bias correction

In Figure 4, we illustrate the effects of adBiC on state-wise accuracies, for all backbone IL methods evaluated in this work. Before adaptive correction (top), all methods provide strong performance on the last group of classes learned (represented by the diagonals). Their performance is generally poorer for past classes (under the diagonals).

After correction (bottom), all methods perform better on past class groups (with a trade-off in accuracy on the last class group) resulting in a higher overall performance.

D. Effect of the number of reference datasets

In Table 6, we provide a full set of results for the accuracies of adBiC with LwF, LUCIR, SIW and FT+ when varying the number $R$ of reference datasets, following Table 3 of the main paper. For all methods considered, a single reference dataset is sufficient to obtain significant gains with adBiC.
| Classes names |
|---------------|
| CIFAR-100     |
| Apple, Aquarium fish, Baby, Bear, Beaver, Bed, Bee, Beetle, Bicycle, Bottle, Bowl, Boy, Bridge, Bus, Butterfly, Camel, Can, Castle, Caterpillar, Cattle, Chair, Chimpanzee, Clock, Cloud, Cockroach, Couch, Crab, Crocodile, Cup, Dinosaur, Dolphin, Elephant, Flatfish, Forest, Fox, Girl, Hamster, House, Kangaroo, Keyboard, Lamp, Lawn mower, Leopard, Lion, Lizard, Lobster, Man, Maple tree, Motorcycle, Mountain, Mouse, Mushroom, Oak tree, Orange, Orchid, Otter, Palm tree, Pear, Pickup truck, Pine tree, Plain, Plate, Poppy, Porcupine, Possum, Rabbit, Racoon, Ray, Road, Rocket, Rose, Sea, Seal, Shark, Shrew, Skunk, Skyscraper, Snail, Snake, Spider, Squirrel, Streetcar, Sunflower, Sweet pepper, Table, Tank, Telephone, Television, Tiger, Tractor, Train, Trout, Tulip, Turtle, Wardrobe, Whale, Willow tree, Wolf, Woman, Worm |
| IMN-100       |
| Bletilla striata, Christmas stocking, Cognac, European sandpiper, European turkey oak, Friesian, Japanese deer, Luger, Sitka spruce, Tennessee walker, Torrey pine, Baguet, Bald eagle, Barn owl, Bass guitar, Bathrobe, Batting helmet, Bee eater, Blue gum, Blue whale, Bones, Borage, Brass, Caftan, Candytuft, Carthorse, Cattle egret, Cayenne, Chairlift, Chicory, Cliff dwelling, Cocktail dress, Commuter, Concert grand, Crazy quilt, Delivery truck, Detached house, Dispensary, Drawing room, Dress hat, Drone, Frigate bird, Frozen custard, Gemsbok, Giant kangaroo, Guava, Hamburger bun, Hawfinch, Hill myna, Howler monkey, Huisache, Jennet, Jodhpurs, Ladder truck, Looner, Micrometer, Mink, Moorhen, Moorhen, Moped, Mortarboard, Mosquito net, Mountain zebra, Muffler, Musk ox, Obelisk, Opera, Ostrich, Ox, Oximeter, Playpen, Post oak, Purple-fringed orchid, Purple willow, Quaking aspen, Ragged robin, Raven, Redpoll, Repository, Roll-on, Scatter rug, Shopping cart, Shower curtain, Slip-on, Spider orchid, Sports car, Steam iron, Stole, Stuffed mushroom, Subcompact, Sundial, Tabby, Tabi, Tank car, Tramway, Unicycle, Wagtail, Walker, Window frame, Wood anemone |
| BIRDS-100     |
| American bittern, American coot, Atlantic puffin, Baltimore oriole, Barrow’s goldeneye, Bewick’s swan, Bullock’s oriole, California quail, Eurasian kingfisher, European gallinule, European sandpiper, Orpington, Amazon, Barn owl, Black-crowned night heron, Black-necked stilt, Black-winged stilt, Black swan, Black vulture, Black vulture, Blue peafowl, Brambling, Bufflehead, Cassowary, Cockerel, Common spoonbill, Crossbill, Duckling, Eastern kingbird, Emperor penguin, Fairy bluebird, Fishing eagle, Fulmar, Gamecock, Golden pheasant, Goosander, Goshawk, Great crested grebe, Great horned owl, Great white heron, Greater yellowlegs, Greenshank, Gyrfalcon, Hawfinch, Hedge sparrow, Hen, Honey buzzard, Hornbill, Kestrel, Kookaburra, Lapwing, Least sandpiper, Little blue heron, Little egret, Macaw, Mandarin duck, Marsh hawk, Moorhen, Mourning dove, Muscovy duck, Mute swan, Ostrich, Owlet, Oystercatcher, Pochard, Raven, Red-legged partridge, Red-winged blackbird, Robin, Robin, Rock hopper, Roseate spoonbill, Ruby-crowned kinglet, Ruffed grouse, Sanderling, Screech owl, Screech owl, Sedge warbler, Shoveler, Siskin, Snow goose, Snowy egret, Song thrush, Spotted flycatcher, Spotted owl, Sulphur-crested cockatoo, Thrush nightingale, Tropic bird, Tufted puffin, Turkey cock, Weka, Whistling swan, White-breasted nuthatch, White-crowned sparrow, White-throated sparrow, White stork, Whole snipe, Wood ibis, Wood pigeon |
| FOOD-100      |
| Apple pie, Baby back ribs, Baklava, Beef carpaccio, Beef tartare, Beet salad, Beignets, Bibimbap, Bread pudding, Breakfast burrito, Bruschetta, Caesar salad, Cannoli, Caprese salad, Carrot cake, Ceviche, Cheese plate, Cheesecake, Chicken curry, Chicken quesadilla, Chicken wings, Chocolate cake, Chocolate mouse, Churros, Clam chowder, Club sandwich, Crab cakes, Creme brulee, Crouque madame, Cup cakes, Deviled eggs, Donuts, Dumplings, Edamame, Eggs benedict, Escargots, Falafel, Filet mignon, Fish and chips, Foie gras, French fries, French onion soup, French toast, Fried calamari, Fried rice, Frozen yogurt, Garlic bread, Gnocchi, Greek salad, Grilled cheese sandwich, Grilled salmon, Guacamole, Gyozza, Hamburger, Hot and sour soup, Hot dog, Huevos rancheros, Hummus, Ice cream, Lasagna, Lobster bisque, Lobster roll sandwich, Macaroni and cheese, Macarons, Miso soup, Mussels, Nachos, Omelette, Onion rings, Oysters, Pad thai, Paella, Pancakes, Panna cotta, Peking duck, Pho, Pizza, Pork chop, Poutine, Prime rib, Pulled pork sandwich, Ramen, Ravioli, Red velvet cake, Risotto, Samosa, Sashimi, Scallops, Seaweed salad, Shrimp and grits, Spaghetti bolognese, Spaghetti carbonara, Spring rolls, Steak, Strawberry shortcake, Sushi, Tacos, Takoyaki, Tiramisu, Tuna tartare |
Table 4: Classes names of target datasets listed by alphabetical order

| Classes names                                                                 |
|-------------------------------------------------------------------------------|
| Airplane cabin, Amphitheater, Amusement arcade, Aqueduct, Arcade, Archaeological excavation, Archive, Arena performance, Attic, Bamboo forest, Bar, Barn, Baseball field, Bazaar outdoor, Beach, Beach house, Beauty salon, Bedroom, Bookstore, Bus interior, Cafeteria, Castle, Chemistry lab, Church outdoor, Cliff, Corridor, Courthouse, Crevasses, Department store, Desert sand, Desert vegetation, Dining room, Dorm room, Drugstore, Elevator lobby, Elevator shaft, Entrance hall, Escalator indoor, Farm, Field cultivated, Field wild, Florist shop indoor, Food court, Fountain, Garage indoor, Gazebo exterior, Golf course, Hangar outdoor, Harbor, Hardware store, Hayfield, Heliport, Highway, Home theater, Hospital room, Hot spring, Hotel outdoor, Hunting lodge outdoor, Ice skating rink indoor, Junkyard, Kasbah, Kitchen, Lagoon, Lake natural, Marsh, Mosque outdoor, Oast house, Office cubicles, Pagoda, Park, Pavilion, Physics laboratory, Pier, Porch, Racecourse, Residential neighborhood, Restaurant, Rice paddy, Rock arch, Ruin, Sauna, Server room, Shed, Shopfront, Storage room, Sushi bar, Television room, Television studio, Throne room, Topiary garden, Tower, Tree house, Trench, Underwater ocean deep, Waiting room, Water park, Waterfall, Wet bar, Windmill, Zen garden |

Table 5: Average top-1 incremental accuracy using $S = \{5, 10, 20\}$ states, for the PLACES-100 dataset with all and half of the training data. The PLACES-100 dataset is extracted from Places365 [35]. Similarly to the other datasets presented in the main paper, we randomly select a hundred classes from the original dataset to construct a dataset with a hundred classes. Results obtained are comparable to those obtained on the other datasets, despite the domain shift from ImageNet. Gains obtained over the backbone method are given in green, and the best results for each setting in bold.

| Method  | PLACES-100 | PLACES-100 (halved) |
|---------|------------|---------------------|
|         | $S = 5$    | $S = 10$            | $S = 20$ | $S = 5$    | $S = 10$            | $S = 20$ |
| LwF [21] | 43.3       | 35.1                | 25.9     | 35.4       | 27.7                | 21.5     |
| w/ adBiC| **44.2** + 0.9 | **36.6** + 1.5 | **28.6** + 2.7 | 35.9 + 0.5 | 28.5 + 0.8 | **23.6** + 2.1 |
| LUCIR [14] | 40.5       | 26.0                | 16.0     | 35.5       | 23.2                | 14.7     |
| w/ adBiC| 42.8 + 2.3 | 35.4 + 9.4          | 23.3 + 7.3 | **40.5** + 5.0 | **33.6** + 10.4 | 22.3 + 7.6 |
| SIW [2] | 27.3       | 20.6                | 14.0     | 27.2       | 19.6                | 14.8     |
| w/ adBiC| 28.8 + 1.5 | 21.2 + 0.6          | 13.1 - 0.9 | 28.5 + 1.3 | 19.3 - 0.3         | 14.3 - 0.5 |
| FT+ [23] | 26.9       | 20.8                | 12.1     | 00.0       | 00.0                | 00.0     |
| w/ adBiC| 27.3 + 0.4 | 19.7 - 1.1          | 13.2 + 1.1 | 25.6 - 0.5 | 17.2 - 2.7         | 13.5 + 1.1 |
Table 6: Average top-1 incremental accuracy of adBiC-averaged across 10 random samplings of the reference datasets.

|      | Raw | R = 1 | R = 2 | R = 3 | R = 4 | R = 5 | R = 6 | R = 7 | R = 8 | R = 9 | R = 10 |
|------|-----|-------|-------|-------|-------|-------|-------|-------|-------|-------|--------|
|      |     |       |       |       |       |       |       |       |       |       |        |
| S = 5 |     | 53.0  | 54.3 ± 0.2 | 54.3 ± 0.1 | 54.4 ± 0.1 | 54.3 ± 0.1 | 54.3 ± 0.1 | 54.3 ± 0.1 | 54.3 ± 0.1 | 54.3 ± 0.1 | 54.3 ± 0.1 |
| S = 10 |     | 44.0  | 46.2 ± 0.3 | 46.4 ± 0.2 | 46.4 ± 0.2 | 46.4 ± 0.1 | 46.4 ± 0.1 | 46.5 ± 0.1 | 46.4 ± 0.1 | 46.4 ± 0.1 | 46.4 ± 0.1 |
| S = 20 |     | 29.1  | 31.8 ± 0.3 | 32.1 ± 0.1 | 32.1 ± 0.2 | 32.1 ± 0.1 | 32.2 ± 0.1 | 32.3 ± 0.1 | 32.3 ± 0.1 | 32.3 ± 0.1 | 32.3 ± 0.1 |

(a) LwF [21]

|      | Raw | R = 1 | R = 2 | R = 3 | R = 4 | R = 5 | R = 6 | R = 7 | R = 8 | R = 9 | R = 10 |
|------|-----|-------|-------|-------|-------|-------|-------|-------|-------|-------|--------|
|      |     |       |       |       |       |       |       |       |       |       |        |
| S = 5 |     | 50.1  | 54.7 ± 0.4 | 54.8 ± 0.3 | 54.8 ± 0.1 | 54.8 ± 0.1 | 54.8 ± 0.1 | 54.8 ± 0.1 | 54.8 ± 0.1 | 54.8 ± 0.1 | 54.8 ± 0.1 |
| S = 10 |     | 33.7  | 42.0 ± 0.7 | 42.1 ± 0.3 | 42.2 ± 0.4 | 42.3 ± 0.3 | 42.2 ± 0.2 | 42.2 ± 0.1 | 42.2 ± 0.1 | 42.2 ± 0.1 | 42.2 ± 0.1 |
| S = 20 |     | 19.5  | 27.5 ± 1.4 | 27.8 ± 0.7 | 27.8 ± 0.9 | 28.3 ± 0.4 | 28.5 ± 0.5 | 28.6 ± 0.6 | 28.5 ± 0.4 | 28.4 ± 0.3 | 28.4 ± 0.2 |

(b) LUCIR [14]

|      | Raw | R = 1 | R = 2 | R = 3 | R = 4 | R = 5 | R = 6 | R = 7 | R = 8 | R = 9 | R = 10 |
|------|-----|-------|-------|-------|-------|-------|-------|-------|-------|-------|--------|
|      |     |       |       |       |       |       |       |       |       |       |        |
| S = 5 |     | 29.9  | 31.6 ± 0.2 | 31.6 ± 0.2 | 31.6 ± 0.1 | 31.7 ± 0.2 | 31.7 ± 0.1 | 31.7 ± 0.1 | 31.7 ± 0.1 | 31.7 ± 0.1 | 31.7 ± 0.1 |
| S = 10 |     | 22.7  | 23.8 ± 0.4 | 23.8 ± 0.2 | 23.9 ± 0.2 | 24.0 ± 0.2 | 23.9 ± 0.1 | 24.0 ± 0.1 | 24.1 ± 0.1 | 24.0 ± 0.1 | 24.1 ± 0.1 |
| S = 20 |     | 14.8  | 15.7 ± 0.3 | 15.7 ± 0.2 | 15.7 ± 0.2 | 15.8 ± 0.1 | 15.8 ± 0.2 | 15.8 ± 0.1 | 15.8 ± 0.1 | 15.8 ± 0.1 | 15.8 ± 0.1 |

(c) SIW [2]

|      | Raw | R = 1 | R = 2 | R = 3 | R = 4 | R = 5 | R = 6 | R = 7 | R = 8 | R = 9 | R = 10 |
|------|-----|-------|-------|-------|-------|-------|-------|-------|-------|-------|--------|
|      |     |       |       |       |       |       |       |       |       |       |        |
| S = 5 |     | 28.9  | 31.9 ± 0.2 | 32.0 ± 0.1 | 32.0 ± 0.1 | 32.0 ± 0.1 | 32.0 ± 0.1 | 31.9 ± 0.1 | 32.0 ± 0.1 | 32.0 ± 0.1 | 31.9 ± 0.1 |
| S = 10 |     | 22.6  | 23.2 ± 0.4 | 23.5 ± 0.2 | 23.5 ± 0.2 | 23.6 ± 0.1 | 23.5 ± 0.2 | 23.6 ± 0.1 | 23.6 ± 0.1 | 23.6 ± 0.1 | 23.6 ± 0.1 |
| S = 20 |     | 14.5  | 14.8 ± 0.2 | 15.0 ± 0.1 | 15.0 ± 0.2 | 15.1 ± 0.1 | 15.0 ± 0.1 | 15.1 ± 0.1 | 15.0 ± 0.1 | 15.0 ± 0.1 | 15.0 ± 0.1 |

(d) FT+ [23]

Table 6: Average top-1 incremental accuracy of adBiC-corrected models trained incrementally on CIFAR-100 with LwF, LUCIR, SIW and FT+, for $S = \{5, 10, 20\}$ states, while varying the number $R$ of reference datasets. For $R \leq 9$, results are averaged across 10 random samplings of the reference datasets. Raw is the accuracy of each method without bias correction.
Figure 5: Average accuracies in each state on CIFAR-100 (left) and IMN-100 (right) datasets with all backbone methods after \textit{adBiC} correction, for $S = 5$ (top), $S = 10$ (middle) and $S = 20$ (bottom) states. The accuracies without correction of the corresponding methods are provided in dashed lines (same colors). \textit{Best viewed in color}. 
Figure 6: Average accuracies in each state on BIRDS-100 (left) and FOOD-100 (right) datasets with all backbone methods after adBiC correction, for $S = 5$ (top), $S = 10$ (middle) and $S = 20$ (bottom) states. The accuracies without correction of the corresponding methods are provided in dashed lines (same colors). Best viewed in color.