UNSUPERVISED CLASS-INCREMENTAL LEARNING THROUGH CONFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

While many works on Continual Learning have shown promising results for mitigating catastrophic forgetting, they have relied on supervised training. To successfully learn in a label-agnostic incremental setting, a model must distinguish between learned and novel classes to properly include samples for training. We introduce a novelty detection method that leverages network confusion caused by training incoming data as a new class. We found that incorporating a class-imbalance during this detection method substantially enhances performance. The effectiveness of our approach is demonstrated across a set of common image classification benchmarks: MNIST, SVHN, CIFAR-10, and CIFAR-100.

1 INTRODUCTION

The development of continually learning systems remains to be a major obstacle in the field of artificial intelligence. The primary challenge is to mitigate catastrophic forgetting: learning new tasks while maintaining the ability to perform old ones. This domain of research is often referred to as Continual Learning, Lifelong Learning, Sequential Learning, or Incremental Learning: each with subtleties in the learning environment and training process, but most with the use of supervision (De Lange et al. (2020)).

Recently, Stojanov et al. (2019) introduced a novel unsupervised class-incremental learning problem motivated by the desire to simulate how children's play behaviors support their ability to learn object models (IOLfCV). Here, sequential tasks take the form of exposures. Each exposure is comprised of a set of images that pertains to a single class that is hidden from the learner. Exposure *boundaries*, the transition from one exposure to the next, are known. The model trained in this setting, is analogous to a young child that has been placed in a playpen with a set of new toys. The child steadily gathers information over time by picking up, examining, and putting down new/old objects continuously. Similar to how the child does not have any guidance to the toys they will examine, the agent does not have access to the exposure's identity during training.

To learn in the unsupervised class-incremental setting, an agent must conduct two procedures successfully. Given a new learning exposure, the key step is to perform novelty detection: to identify whether an exposure corresponds to a class that has been learned. If the agent determines that an exposure is familiar, the second step is to identify its label such that the exposure can be leveraged to update the model. Both procedures must be performed reliably. Otherwise, the novelty detection mistakes will result in label noise that distorts the learned model, increasing the likelihood of subsequent mistakes.

Deep neural networks are known to make overconfident decisions for anomalous data distributions that were not seen during training (Hendrycks & Gimpel (2016)). To address this problem, research related to out-of-distribution (OOD) detection have utilized supervised methods (Liang et al. (2017); Alemi et al. (2018)) and unsupervised methods (Choi & Jang (2018); Hendrycks et al. (2018); Serrà et al. (2019)). Works related to open set recognition have also addressed the OOD problem by applying distance-based thresholds computed from known class scores (Scheirer et al. (2012; 2014)). The work by Stojanov et al. (2019) applies a similar method to the unsupervised incremental setting by computing class features produced from a set of supervised samples. In contrast, we propose a model, Incremental Learning by Accuracy Performance (iLAP), that determines class novelty and identity by considering performance changes of previously learned tasks when an incoming set of exposure images are trained under a new label.

Instead of using a distance-based metric, our novelty detection threshold relies on the percentage of accuracy that was maintained by performing a model update using the incoming exposure. This poses several practical advantages: First, the threshold value does not rely on supervised samples and is more intuitive (Section 3.3). Second, the performance of our method is independent of the sequence of the incoming exposure classes (Section 5.2). Finally, the model is able to distinguish between *similar* classes more reliably (Section 5.3).

From our experiments, we demonstrate that the confusion resulting from training with label ambiguity results in a more reliable signal for novelty detection in comparison to previous methods. We demonstrate that our technique is more robust and results in substantial performance gains in comparison to various baselines. Furthermore, despite the absence of labels, our model was able to perform similarly to supervised models under several benchmarks.

In summary, this work provides three contributions:

- We present a novel framework, iLAP, that achieves learning in the unsupervised class-incremental environment where the exposure identities are unknown.
- We demonstrate that by including a class-imbalance technique, our unsupervised method is able to closely match supervised performance for several image classification benchmarks trained in the incremental setting.
- We identify failure cases that are overlooked by traditional OOD methods that leverage distance-based thresholds.

2 Related Works

Introduced by Stojanov et al. (2019), the unsupervised class-incremental setting contains a set of sequential tasks that are single-class exposures; classes pertaining to the exposures may repeat and are unknown. This is not to be mistaken with unsupervised continual learning (UCL) where task boundaries and task identities are unavailable (Lee et al. (2020); Smith & Dovrolis (2019); Rao et al. (2019)). Our work presents an agent that is able to leverage the *boundary* information from the unsupervised class-incremental environment to achieve performances that are close to models trained under supervision.

2.1 Continual Learning/Incremental Learning

Prior works in this field primarily aim to improve a model's ability to retain information while incorporating new tasks (Goodfellow et al. (2013); Parisi et al. (2019); Rebuffi et al. (2017); Lopez-Paz & Ranzato (2017); Aljundi et al. (2018); Castro et al. (2018)). Typically, these models reside in learning settings where both task labels and task boundaries are available. Methods include replay techniques, the usage of artifacts and generated samples to refresh a model's memory (Kamra et al. (2017); Wu et al. (2018); Rolnick et al. (2019); Shin et al. (2017); Wu et al. (2019)), and regularization-based practices, the identification and preservation of weights that are crucial for the performance of specific tasks (Kirkpatrick et al. (2017); Zenke et al. (2017); Yoon et al. (2017)). In contrast to prior works, our method addresses incremental learning in a setting where exposure labels are unavailable.

2.2 Unsupervised Continual Learning

Recently, a series of works tackle the UCL problem where task boundaries and task identities are unknown. Smith & Dovrolis (2019) performs novelty detection by analyzing an input image through a series of receptive fields to determine if an input patch is an outlier. Meanwhile, CURL proposes a method to learn class-discriminative representations through a set of shared parameters (Rao et al. (2019)). CN-DPM, introduces an expansion-based approach that utilizes a mixture of experts to learn feature representations (Lee et al. (2020)). Although CN-DPM performs in a *task-free* setting, incoming tasks are multi-class and individual class labels are provided. This supervision is required to train the existing experts and determine when a new one is needed. While boundary information is not required for these works, the performances are far below supervised baselines (77.7% on and MNIST 13.2% Omniglot) (Rao et al. (2019)).

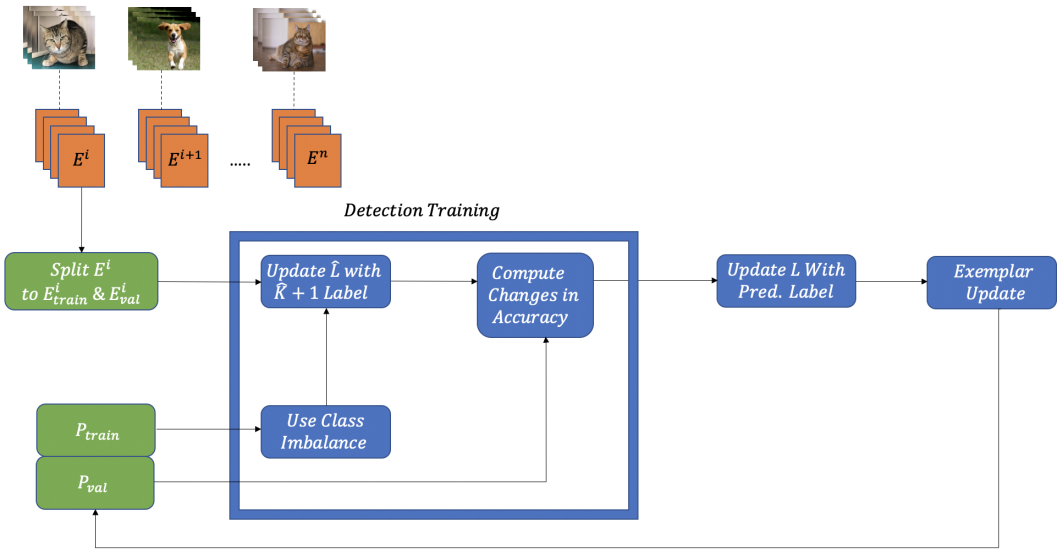

Figure 1: Incoming exposures are split into E^i_{train} and E^i_{val}. E^i_{train} is aggregated with $\mathcal{P}_{train}$ and is used for detection training. $\hat{L}$ is updated with samples from E^i_{train} labeled as $\hat{K}+1$. The change in model accuracy is assessed using $\mathcal{P}_{val}$ to identify novelty and class identity. Finally, L is updated with the predicted label; samples from E^i_{train} and E^i_{val} are saved to their respective exemplars.

2.3 Out-of-Distribution Detection

This ongoing area of research aims to detect outliers in training and testing data. Current approaches can be largely categorized by statistical, distance-based, and deep learning methods (Eskin (2000); Yamanishi et al. (2004); Knorr et al. (2000); Hautamaki et al. (2004); Sabokrou et al. (2018); Kliger & Fleishman (2018)). Recent techniques involve using a threshold to determine class novelty from network confidence values (Hendrycks & Gimpel (2016)). ODIN uses input perturbations to increase softmax scores for neural networks to distinguish in-distribution images from out-of-distribution images (Liang et al. (2017)). DeVries & Taylor (2018) incorporates a confidence branch to obtain out-of-distribution estimations. Our method (iLAP) is the first to incorporate a threshold value that is dependent on class-accuracy changes caused by data poisoning.

3 Approach

In this section, we provide an overview of our method. We begin by identifying the learning setting, followed by details of our training process. Finally, insights for choosing threshold values are provided.

3.1 Setting

In the unsupervised class-incremental setting, a learner L perceives an input stream of exposures denoted as $E^1, E^2, ..., E^n$. Each exposure contains a set of images, $E^i = \{e^i_1, e^i_2, ..., e^i_{n_i}\}$, $e^i_j \in \mathbb{R}^{C\times H\times W}$, where C, H, and W are the number of channels, height, and width of the input image respectively. Each exposure belongs to a single class $y_i \in \mathbb{N}$, which has been sampled from class distribution $P(\mathcal{C})$. For each E^i, L does not have knowledge of the true class label y_i. Two exemplars, $\mathcal{P}_{train} = (P^1_{train}, P^2_{train}, ..., P^{\hat{K}}_{train})$ and $\mathcal{P}_{val} = (P^1_{val}, P^2_{val}, ..., P^{\hat{K}}_{val})$, are maintained at all times, where $\hat{K}$ denotes the total number of classes L has currently determined. The exemplars are used to store samples from the exposure for replay and accuracy assessment. The sizes of both exemplars, $\left|P^i_{train}\right|$ and $\left|P^i_{val}\right|$, are bounded per class.

Figure 2: Graphs depicting the average accuracy decrease for seen vs. unseen classes at various values of class-imbalance. As the ratio of class-imbalance increases, the accuracy change for unseen classes remains low, with the exception at high ratios when very few samples are present during the model update (red). The accuracy decrease for seen classes is amplified as the ratio of class-imbalance increases (blue).

3.2 Detection Training

For each incoming exposure, the model is tasked to determine whether the class associated with the exposure was learned previously. Our solution is to perform a model update by treating the incoming exposure as a new class, we coin this technique *detection training*. Under the circumstances that the exposure class is repeated, the performance for the previously learned class would suffer drastically after training. The reason for this behavior is because the model has associated two different labels to a similar class distribution.

During detection training, $\hat{L}$, a copy of L is produced. The incoming exposure is assigned with label $\hat{K}+1$. Train-validation split is performed on the incoming exposure to obtain E_{train} and E_{val}, and are aggregated with exemplars $\mathcal{P}_{train}$ and $\mathcal{P}_{val}$ respectively. The combined samples are used to train $\hat{L}$ via validation-based early stopping. We denote the vector $\{\Delta_{\hat{y}}\}_{\hat{y}\in[\hat{K}]}$ to represent the percentage decrease of the class accuracies (computed using $\mathcal{P}_{val}$) before and after the update. If $\max(\{\Delta_{\hat{y}}\})$ exceeds a threshold, θ, the incoming exposure is likely to have been learned by L. In this case, the correct identity pertaining to the exposure is $\arg\max_{\hat{y}\in[\hat{K}]} \Delta_{\hat{y}}$. Otherwise, if θ is unsatisfied, $\hat{K}+1$ is the appropriate label for the new class.

3.3 Class-Imbalance for Detection Training

Introducing a class-imbalance during detection training creates a more distinct decision boundary by exacerbating the class-accuracy drop for repeated exposures. Consider a theoretical case where an optimal model has learned $\hat{K}$ classes. The incoming exposure, E^i, contains a distribution that is equal to that of some previously learned class $\hat{y}_i$. If the model were updated with equal samples of E^i labeled $\hat{K}+1$, and $P^{\hat{y}_i}_{train}$ labeled $\hat{y}_i$, the accuracy for class $\hat{y}_i$ would become ambiguous during validation ($\hat{y}_i \approx 50\%$, $\hat{K}+1 \approx 50\%$). However, if the model were updated with a greater sample of $\hat{K}+1$ labels, the accuracy drop for class $\hat{y}_i$ would be considerably larger because $\hat{K}+1$ would be favored during inference.

$\hat{L}$ has the option to use an imbalanced dataset during detection training where a fraction of the size for each class from $\mathcal{P}_{train}$ are used in comparison to the size of E_{train}. Let $P^i_{sampled} \subset P^i_{train}$, the class-imbalance ratio λ is:

$$\lambda = 1 - \frac{|P^i_{sampled}|}{|E_{train}|} \tag{1}$$

In **Figure 2**, we use Fashion-MNIST (Xiao et al. (2017)) and Imagenette (Howard) to obtain a general value for λ and θ for our experiments with other datasets. Performance drops are tracked for the repeated class and non-repeated classes during detection training with various values of λ. The

goal is to maximize the performance loss for the repeated class while maintaining the performance for the other classes. In **Section 4** a θ value of 0.6 and a λ value of 0.5 are used. λ is determined by the point of the maximum distance between the two accuracy curves; θ, the expected accuracy drop for a repeated exposure, is the mean of the two values at λ **Figure 2**.

3.4 Model and Exemplar Update

After obtaining the correct label ($\hat{K}+1$ or $\arg\max_{\hat{y}\in[\hat{K}]} \Delta_{\hat{y}}$) from detection training, L is trained with the aggregated dataset obtained from P_{train} and E_{train}. Two sets with the most representative samples from E_{train} and E_{test} are created and added to their respective exemplars for future replay. The selection process is determined by ranking the distance between the image features and the class-mean image features. This methodology is consistent with the procedure introduced by Castro et al. (2018).

The availability of $\mathcal{P}_{val}$ allows L to assess how well old and new classes are considered. If the accuracy for any class interpreted by L falls below a percentage, the class is altogether discarded. This allows the model to remove extraneous classes that were learned insufficiently or have been forgotten. In **Section 4**, a percentage threshold of 20% is used for all experiments.

4 Experiments

The performance of our framework is evaluated using a series of common image classification benchmarks: MNIST, SVHN, CIFAR-10, and CIFAR-100. First, we compare our novelty detector to related OOD methods. Next, we evaluate performance of iLAP to that of other incremental learners: BiC (Wu et al. (2019)) (supervised) and IOLfCV (unsupervised).

4.1 Experimental Details

For the following experiments, a ResNet-18 model (He et al. (2016)) pre-trained with ImageNet is used. $\lambda = 0.5$ and $\theta = 0.6$ are used for iLAP with class-imbalance detection training, while a $\lambda = 0$ and $\theta = 0.4$ are used for iLAP without class-imbalance detection training. The parameters are maintained across all benchmarks.

For each exposure, the model is trained for 15 epochs with 16 batch size, using an Adam (Kingma & Ba (2014)) optimizer with validation-based early stopping; a learning rate of $2\mathrm{e}{-4}$ is used. The feature extraction layers use a ten times lower learning rate at $2\mathrm{e}{-5}$. For all models, the exemplar size per class is equal to the exposure size. The exposure validation split ratio is 0.8 (e.g. for exposure size $= 200$, iLAP: $[E_{train}] = 160$ and $[E_{val}] = 40$). The thresholds used for IOLfCV in **4.1** were determined by maximizing the F-score for the classification of in-distribution exposures versus out-of-distribution exposures. To obtain the best performance possible, the entirety of the dataset was used. The values are 0.46, 0.63, 0.57, and 0.62 for MNIST, SVHN, CIFAR-10, and CIFAR-100 respectively.

4.2 Out-of-distribution Detection Results

The OOD detectors are assessed in an incremental setting with size 200 exposures. The detectors are evaluated on their ability to determine whether an exposure is novel by using the common established metrics: FPR95, AUROC, and AUPR (Hendrycks & Gimpel (2016)). Details of the compared works and evaluation methods are described in **Appendix A.1**. The results are illustrated in **Table 1**.

4.3 Incremental Learning Results

The accuracy of learner L is computed using the ground-truth mapping $m : [\hat{K}] \rightarrow [K]$ with the equations:

$$S(x,y) = \begin{cases} \frac{1}{|m^{-1}(y)|}, & \text{if } L(x) \in m^{-1}(y) \\ 0, & \text{otherwise} \end{cases}$$

$$Accuracy = \mathbb{E}_{x,y\sim test}\left[S(x,y)\right]$$

	MNIST			CIFAR-10		
	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑
MSP	$0.14\pm.09$	$0.98\pm.01$	$0.95\pm.02$	$0.60\pm.13$	$0.86\pm.04$	$0.71\pm.09$
CE	$0.35\pm.13$	$0.80\pm.04$	$0.67\pm.09$	$0.64\pm.13$	$0.78\pm.04$	$0.54\pm.03$
ODIN	$0.12\pm.09$	$0.98\pm.01$	$0.95\pm.02$	$0.55\pm.12$	$0.87\pm.04$	$0.71\pm.09$
ZeroShot OOD	$0.03\pm.03$	$0.99\pm.01$	$0.99\pm.01$	$0.63\pm.12$	$0.67\pm.06$	$0.47\pm.08$
IOLfCV	$0.12\pm.08$	$0.98\pm.01$	$0.95\pm.02$	$0.60\pm.12$	$0.78\pm.06$	$0.60\pm.08$
iLAP w/o CI (Ours)	$0.19\pm.02$	$0.92\pm.02$	$0.68\pm.05$	$0.58\pm.03$	$0.69\pm.03$	$0.40\pm.05$
iLAP w/ CI (Ours)	$\mathbf{0.0\pm.0}$	$\mathbf{1.0\pm.0}$	$\mathbf{1.0\pm.0}$	$\mathbf{0.32\pm.10}$	$\mathbf{0.93\pm.03}$	$\mathbf{0.81\pm.08}$

	SVHN			CIFAR-100		
	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑
MSP	$0.08\pm.04$	$0.98\pm.01$	$0.98\pm.01$	$0.20\pm.04$	$0.95\pm.02$	$0.71\pm.09$
CE	$0.5\pm.1$	$0.78\pm.03$	$0.72\pm.07$	$0.55\pm.12$	$0.82\pm.01$	$0.63\pm.02$
ODIN	$0.12\pm.01$	$0.94\pm.01$	$0.93\pm.02$	$0.21\pm.04$	$0.94\pm.02$	$0.91\pm.01$
ZeroShot OOD	$0.06\pm.04$	$0.97\pm.01$	$0.98\pm.01$	$0.23\pm.03$	$0.97\pm.01$	$0.94\pm.02$
IOLfCV	$0.28\pm.11$	$0.96\pm.01$	$0.95\pm.0$	$0.28\pm.05$	$0.96\pm.01$	$0.94\pm.01$
iLAP w/o CI (Ours)	$0.08\pm.02$	$0.92\pm.02$	$0.89\pm.04$	$0.27\pm.05$	$0.92\pm.05$	$0.72\pm.12$
iLAP w/ CI (Ours)	$\mathbf{0.0\pm.0}$	$\mathbf{0.99\pm.0}$	$\mathbf{0.98\pm.0}$	$\mathbf{.08\pm.02}$	$\mathbf{0.99\pm.0}$	$\mathbf{0.98\pm.0}$

Table 1: Comparison of our technique, with and without class-imbalance, to recent novelty detection methods. FPR95, AUROC, and AUPR are used to evaluate performance. Our method is the most effective in the incremental learning setting where exposure samples are limited.

Test Accuracy (%)				
	MNIST	SVHN	CIFAR-10	CIFAR-100
BiC (Supervised)	$\mathbf{98.0\pm0.2}$	$\mathbf{89.2\pm0.3}$	$\mathbf{74.9\pm0.4}$	$\mathbf{72.5\pm0.3}$
IOLfCV (Unsupervised)	88.0 ± 3.5	46.8 ± 8.6	45.3 ± 4.7	61.6 ± 0.6
iLAP w/o CI (Ours)	90.3 ± 1.7	84.4 ± 1.7	60.9 ± 1.4	65.0 ± 1.2
iLAP w/ CI (Ours)	$\mathbf{98.0\pm0.2}$	$\mathbf{88.1\pm0.8}$	$\mathbf{67.4\pm1.7}$	$\mathbf{68.6\pm0.7}$

Table 2: Comparison of iLAP to IOLfCV and BiC by test accuracy (%) using the MNIST, SVHN, CIFAR-10 and CIFAR-100 datasets.

No. of Unique Classes Learned				
	MNIST	SVHN	CIFAR-10	CIFAR-100
BiC (Supervised)	$\mathbf{10.0\pm0.0}$	$\mathbf{10.0\pm0.0}$	$\mathbf{10.0\pm0.0}$	$\mathbf{100.0\pm0.0}$
IOLfCV (Unsupervised)	9.2 ± 0.4	8.8 ± 1.2	8.7 ± 0.7	89.0 ± 0.5
iLAP w/o CI (Ours)	10.0 ± 0.0	10.0 ± 0.0	10.0 ± 0.0	96.0 ± 0.5
iLAP w/ CI (Ours)	$\mathbf{10.0\pm0.0}$	$\mathbf{10.0\pm0.0}$	$\mathbf{10.0\pm0.0}$	$\mathbf{96.5\pm0.5}$

Table 3: Comparison of iLAP to IOLfCV and BiC by the number of unique classes learned using the MNIST, SVHN, CIFAR-10 and CIFAR-100 datasets.

The learner accuracy is the mean of the sample accuracy scores evaluated on the test set, where (x, y) represents a single sample. For each sample with label y, let $m^{-1}(y)$ represent the corresponding labels from the learner. In the case that a learner output does not belong to set $m^{-1}(y)$, an accuracy score of 0 is assigned (class is not detected). Otherwise, an accuracy score of $\frac{1}{|m^{-1}(y)|}$ is designated to penalize the learner if m is non-injective and have attributed multiple labels to a single ground truth class. Performance results are illustrated in **Table 2 & 3**. Additional visualizations are provided in **Appendix A.2**.

5 Analysis

Traditional OOD methods that rely on distance-based thresholds are restricted by the supervised samples that are available. These values are non-intuitive and vary drastically across datasets (whereas our percentage threshold are $\approx 50\%$ for all datasets). In the incremental learning setting, early mistakes are amplified as more exposures are introduced, a proper threshold initialization dictates a model's feasibility. However, we argue that even with a *good* threshold these methods will consistently fail in particular conditions. The purpose of this section is to discuss the results obtained from our experiments. Subsequently, we highlight a few cases that are overlooked by traditional distance-based methods.

5.1 Out-Of-Distribution Detection Analysis

iLAP with class-imbalance detection training (CI) was able to outperform related OOD methods for the MNIST, SVHN, CIFAR-10, and CIFAR-100 benchmarks under all metrics **Table 1**. However, the results for iLAP without CI were not as definitive. CE performed the worst in the incremental setting, possibly because the performance of the confidence branch is reliant on larger training samples. IOLfCV's method performed on par with related methods.

5.2 Unsupervised Incrementatl Learning Analysis

iLAP with CI was able to beat IOLfCV by 10.0, 41.3, 22.1, and 7.0 percentage points for the MNIST, SVHN, CIFAR-10, and CIFAR-100 benchmarks respectively **Table 2**. iLAP was also able to maintain its performance when exposure sizes are decreased **Appendix A.3**. We found that the reason for the lower performance beat for CIFAR-100 is not attributed to the larger number of classes in the dataset. Rather, the problem lies with how the exposure sequence for the incremental learning setting is created and how the distance-based threshold is calculated.

The threshold for IOLfCV is computed by maximizing the F-score for the binary classification of novel versus non-novel classes. In our experiments, the entirety of the dataset was used to compute the baseline's threshold. Although this is impractical, we wanted to illustrate that iLAP is able to beat the baseline even under the most optimal conditions.

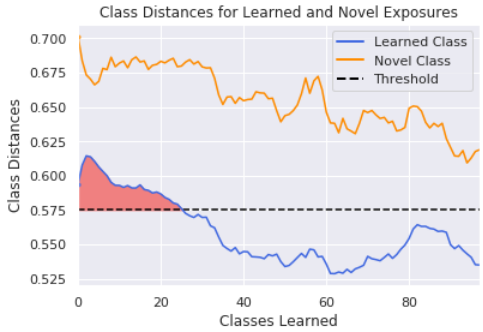

Figure 3: Visualization depicting the behavior of class feature distances as more classes are learned by the network. The optimal threshold computed by optimizing the F-score is the midpoint of the learned and novel class values when all 100 classes are incorporated. However, this threshold will mistakenly recognize repeated classes as new classes early on during incremental training.

In **Figure 3** we illustrate the behavior of the class feature distances as more classes are incorporated by a network. The most optimal threshold that maximizes the F-score lies in the mid-point between the two graphs when all 100 classes are learned. However, because the threshold is fixed, the novelty detector fails to correctly identify repeated classes early on during training and is more inclined to label repeated classes as unseen, (shaded red area in **Figure 3**). Consistent with the described setting in Stojanov et al. (2019), each class within a benchmark is repeated an equal number of times in a randomized sequence. For datasets with a large number of classes, there is a higher chance that the repeated exposures are further apart. Therefore, IOLfCV seemingly performs comparatively better on CIFAR-100 than CIFAR-10, but would fail if early repeated exposures were frequent.

5.3 Class Similarity

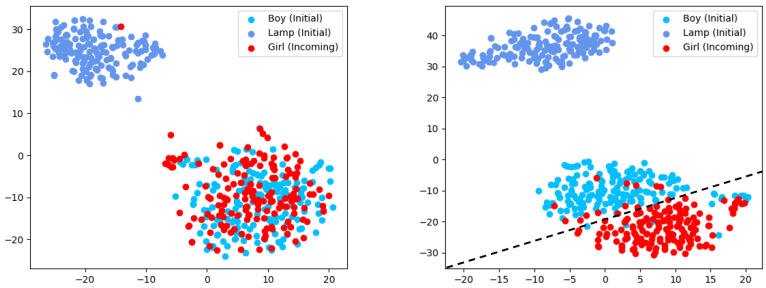

Figure 4: TSNE plot depicting a model which has been trained to classify *lamp* vs. *boy*. An incoming exposure with class *girl* is introduced to the model. Comparing the feature distributions of this model, *boy* and *girl* have almost identical distributions; the OOD detector mistakenly assesses them to be same class (left). In contrast, our method first attempts to learn a feature space where the two classes are separable and identifies *boy* and *girl* correctly (right).

iLAP was able to detect all classes for MNIST, SVHN, and CIFAR-10 and on average 96.5 out of 100 classes for CIFAR-100. Meanwhile, IOLfCV struggles to identify unique classes for all evaluated benchmarks. Through closer inspection, we found that the distance-based method is unable to distinguish classes when they are too *similar*.

Consider two classes, k_1 and k_2, that can be separated by a classifier $\mathcal{C}$ in learned feature space $\mathcal{F}$. An incoming exposure, class k_3, shares a similar distribution to class k_1 in feature space $\mathcal{F}$, but separable in some feature space $\mathcal{F}'$. The distance-based method is highly probable to fail because it is likely to categorize k_1 and k_3 as identical classes. However, because our method always trains the incoming exposure as a new class, $\mathcal{C}$ is forced to learn the feature space $\mathcal{F}'$ in which these two classes are separable. **Figure 4** illustrates two prior classes, *boy* and *lamp*, in some feature space. The incoming exposure, class *girl*, is unable to be distinguished from class *boy* by the distance-based method **Figure 4 (left)**. However, because detection training always attempts to classify the incoming exposure as a new class, our method is able to identify $\mathcal{F}'$ **Figure 4 (right)**.

6 Conclusion

To achieve learning in an unsupervised class-incremental setting, a reliable novelty detector is needed. Current methods utilize a detection threshold that is calibrated using class feature distances. In our work, we illustrate that the use of a static distance-based threshold is not only impractical but also unreliable. Instead, we introduce a technique that leverages confusion error to perform novelty detection by always training the incoming exposure as a new class. Using a series of image classification benchmarks, we illustrate that our method is able to closely rival supervised performance despite the lack of labels.

REFERENCES

Alexander A Alemi, Ian Fischer, and Joshua V Dillon. Uncertainty in the variational information bottleneck. *arXiv preprint arXiv:1807.00906*, 2018.

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 139–154, 2018.

Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 233–248, 2018.

Hyunsun Choi and Eric Jang. Generative ensembles for robust anomaly detection. 2018.

Matthias De Lange, Rahaf Alijundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *arXiv preprint arXiv:1909.08383*, 2020.

Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018.

Eleazar Eskin. Anomaly detection over noisy data using learned probability distributions. 2000.

Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013.

Ville Hautamaki, Ismo Karkkainen, and Pasi Franti. Outlier detection using k-nearest neighbour graph. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pp. 430–433. IEEE, 2004.

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

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

Jeremy Howard. Imagewang. URL `https://github.com/fastai/imagenette/`.

Nitin Kamra, Umang Gupta, and Yan Liu. Deep generative dual memory network for continual learning. *arXiv preprint arXiv:1710.10368*, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

Mark Kliger and Shachar Fleishman. Novelty detection with gan. *arXiv preprint arXiv:1802.10560*, 2018.

Edwin M Knorr, Raymond T Ng, and Vladimir Tucakov. Distance-based outliers: algorithms and applications. *The VLDB Journal*, 8(3-4):237–253, 2000.

Soochan Lee, Junsoo Ha, Dongsu Zhang, and Gunhee Kim. A neural dirichlet process mixture model for task-free continual learning. *arXiv preprint arXiv:2001.00689*, 2020.

Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, pp. 6467–6476, 2017.

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

Dushyant Rao et al. Continual unsupervised representation learning. In *Advances in Neural Information Processing Systems*, pp. 7645–7655, 2019.

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

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. In *Advances in Neural Information Processing Systems*, pp. 348–358, 2019.

Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. Adversarially learned one-class classifier for novelty detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3379–3388, 2018.

Chandramouli S Sastry and Sageev Oore. Zero-shot out-of-distribution detection with feature correlations. 2019.

Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7): 1757–1772, 2012.

Walter J Scheirer, Lalit P Jain, and Terrance E Boult. Probability models for open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 36(11):2317–2324, 2014.

Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. *arXiv preprint arXiv:1909.11480*, 2019.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In *Advances in Neural Information Processing Systems*, pp. 2990–2999, 2017.

James Smith and Constantine Dovrolis. Unsupervised progressive learning and the stam architecture. *arXiv preprint arXiv:1904.02021*, 2019.

Stefan Stojanov, Samarth Mishra, Ngoc Anh Thai, Nikhil Dhanda, Ahmad Humayun, Chen Yu, Linda B Smith, and James M Rehg. Incremental object learning from contiguous views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8777–8786, 2019.

Chenshen Wu, Luis Herranz, Xialei Liu, Joost van de Weijer, Bogdan Raducanu, et al. Memory replay gans: Learning to generate new categories without forgetting. In *Advances In Neural Information Processing Systems*, pp. 5962–5972, 2018.

Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 374–382, 2019.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

Kenji Yamanishi, Jun-Ichi Takeuchi, Graham Williams, and Peter Milne. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. *Data Mining and Knowledge Discovery*, 8(3):275–300, 2004.

Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. *arXiv preprint arXiv:1708.01547*, 2017.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3987–3995. JMLR. org, 2017.

A APPENDIX

A.1 OOD EXPERIMENTAL DETAILS

A brief description of the evaluation metrics used for **Table 1**.

- **False Positive Rate at 95% (FPR95)**: This measure determines the False Positive Rate (FPR) when True Positive Rate (TPR) is equal to 95%. FPR is calculated as $\frac{FP}{FP+TN}$ where FP is the number of False Positives and TN is the number of True Negatives. TPR is calculated as $\frac{TP}{TP+FN}$ where TP is the number of True Positives and FN is the number of False Negatives.
- **Area Under the Receiver Operating Characteristic (AUROC)**: This metric illustrates the relationship between the TPR and the FPR. This measure determines the probability that a novel class will have a higher detection score compared to a non-novel class.
- **Area Under the Precision-Recall (AUPR)**: This metric is constructed by plotting precision versus recall. The AUPR curve treats the novel examples as the positive class. A high area value represents high recall and high precision.

The training detection method used in iLAP is compared to a set of related works in OOD. The following works reflect the acronyms used in **Table 1**

- **MSP (Hendrycks & Gimpel (2016))**: MSP is computed from a trained classifier to perform out-of-distribution detection. The mean MSP for all images belonging to an exposure is used to determine novelty.
- **CE (DeVries & Taylor (2018))**: CE requires extending a model with a confidence branch to obtain a set of values. These values reflect the model's ability to produce a correct prediction. When the estimation value for a set of input images is low, the sample is likely to be novel.
- **ODIN (Liang et al. (2017))**: ODIN uses temperature scaling and input perturbations to widen the MSP difference between in-distribution and out-of-distribution samples. The optimal value for temperature and perturbation magnitude are found by minimizing FPR using grid search. 2 and 0.0012 are used for temperature and perturbation magnitude values respectively, for both the MNIST and CIFAR-10 dataset.

  Zero-Shot OOD (Sastry & Oore (2019)): Zero-Shot OOD uses pairwise correlations of network features to detect novel samples. The class-conditional feature correlations, on the training data, are computed across all layers of the network. These correlations are then compared with the pairwise feature correlations from a test sample to obtain a deviation value.

  IOLfCV (Stojanov et al. (2019)): IOLfCV determines a distance-based threshold computed using average-feature means from a network trained from supervised samples. Two initialized classes are used to compute the threshold by finding the optimal point using precision-recall analysis.

A.2 MAIN EXPERIMENT RESULTS

The following are produced with the use of a GTX TITAN X gpu. For each exposure, the model is trained for 15 epochs with 16 batch size, using an Adam optimizer. The learning rate used is $2\mathrm{e}{-4}$. The feature extraction layers use a ten times lower learning rate at $2\mathrm{e}{-5}$. The input size is 224.

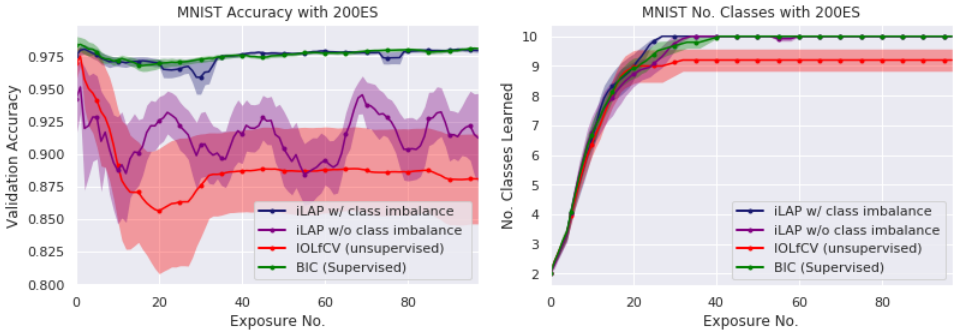

Figure 5: Visualizations comparing accuracy (left) and number of classes detected (right) for the MNIST benchmark.

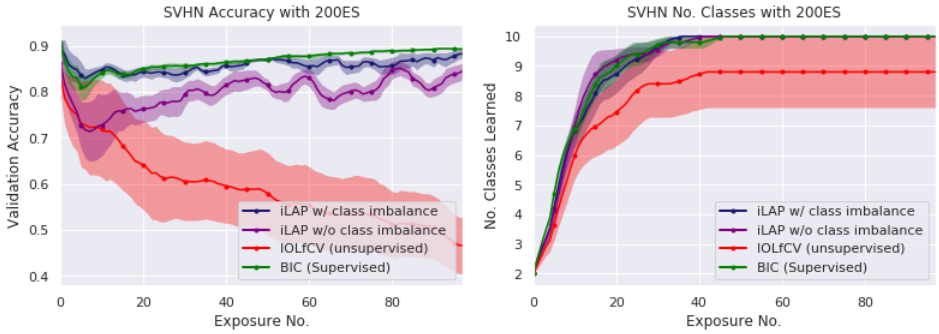

Figure 6: Visualizations comparing accuracy (left) and number of classes detected (right) for the SVHN benchmark.

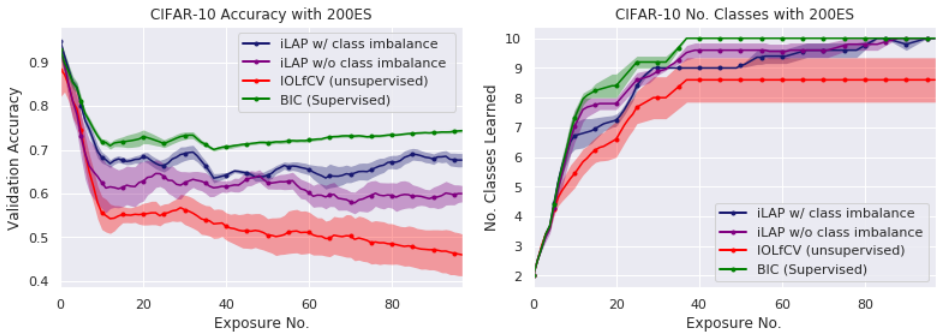

Figure 7: Visualizations comparing accuracy (left) and number of classes detected (right) for the CIFAR-10 benchmark.

Figure 8: Visualizations comparing accuracy (left) and number of classes detected (right) for the CIFAR-100 benchmark.

A.3 Experiments with Lower Exemplar sizes

In this section, we showcase iLAP's results at lower exposure sizes.

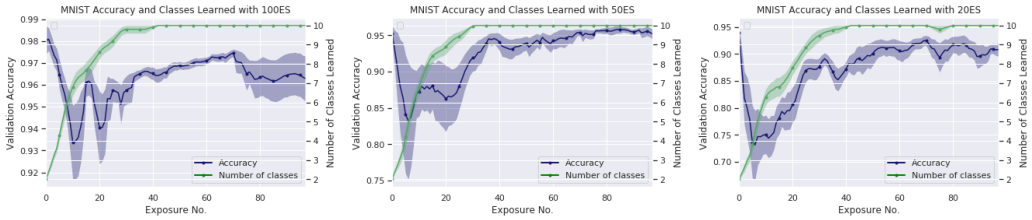

Figure 9: Visualizations depicting iLAP's performance on the MNIST benchmark with 100, 50, and 20 exemplar sizes (left to right).

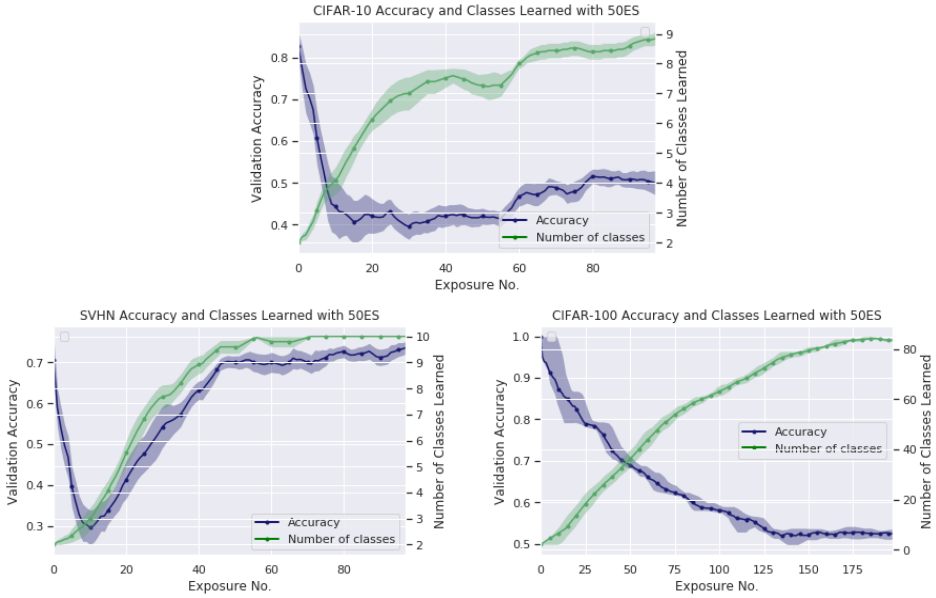

Figure 10: Visualizations depicting iLAP's performance on the SVHN, CIFAR-10, and CIFAR-100 benchmark with a exemplar size of 50.