Exemplar-free Class Incremental Learning via Discriminative and Comparable One-class Classifiers

Wenju Sun Qingyong Li Jing Zhang Danyu Wang Wen Wang Yangli-ao Geng*
Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing, 100044, China 
{SunWenJu,liqy,j.zhang,WangDanYu,wangwen,gengyla}@bjtu.edu.cn

Abstract

The exemplar-free class incremental learning requires classification models to learn new class knowledge incrementally without retaining any old samples. Recently, the framework based on parallel one-class classifiers (POC), which trains a one-class classifier (OCC) independently for each category, has attracted extensive attention, since it can naturally avoid catastrophic forgetting. POC, however, suffers from weak discriminability and comparability due to its independent training strategy for different OCCs. To meet this challenge, we propose a new framework, named Discriminative and Comparable One-class classifiers for Incremental Learning (DisCOIL). DisCOIL follows the basic principle of POC, but it adopts variational auto-encoders (VAE) instead of other well-established one-class classifiers (e.g. deep SVDD), because a trained VAE can not only identify the probability of an input sample belonging to a class but also generate pseudo samples of the class to assist in learning new tasks. With this advantage, DisCOIL trains a new-class VAE in contrast with the old-class VAEs, which forces the new-class VAE to reconstruct better for new-class samples but worse for the old-class pseudo samples, thus enhancing the comparability. Furthermore, DisCOIL introduces a hinge reconstruction loss to ensure the discriminability. We evaluate our method extensively on MNIST, CIFAR10, and Tiny-ImageNet. The experimental results show that DisCOIL achieves state-of-the-art performance. The source code will be publicly available.

1. Introduction

In an ever-changing environment, we would like an intelligent system to learn new knowledge incrementally without forgetting, which refers to incremental learning (IL) [6, 24]. IL aims to learn knowledge from a sequence of dis-

*Corresponding author: Yangli-ao Geng (gengyla@bjtu.edu.cn).

Figure 1. Ideal incremental learning results with two one-class classifiers (OCC). Following the vertical arrow viewpoint, both OCCs should have discriminability to determine whether an input sample belongs to their corresponding class. Following the horizontal arrow viewpoint, they need to have comparability that allows the outputs of the two OCCs to be compared to each other.

1. Exemplar-free class-IL requires the model to learn a new class knowledge without any sample information of old classes, which is quite challenging and leads many established IL algorithms to fail under this scenario [3, 13, 35]. One reason is that the commonly used softmax classifier is prone to overfit the new-task samples without referring to old-task samples, making the classification output bias towards the new-task classes even for the old-task input, which is known as the forgetting problem [10, 28]. To tackle this issue, several algorithms following the frame-
work of parallel one-class classifiers (POC) have been proposed [13, 35]. They train a new one-class classifier (OCC) for each class and thus disentangle the interference of the new task to the old tasks. With this design, the POC-based algorithms naturally avoid the forgetting problem. They, however, are vulnerable to poor performance due to weak discriminability within each OCC and comparability among different OCCs.

More specifically, we illustrate how the discriminability and the comparability affect the performance of POC-based methods in Figure 1. Without loss of generality, let us consider a POC architecture consisting of two OCCs, where one is the current training OCC for a new class and the other is the fixed OCC for an old class. From the vertical viewpoint (green arrow), both the new-class OCC and the old-class OCC should output high probabilities for their corresponding class samples while low probabilities for the others, which we call the discriminability within each OCC. On the other hand, from the horizontal viewpoint (red arrow), for new-class samples, the new-class OCC should output higher probabilities than the old-class OCC; for old-class samples, the probabilities output by the new-class OCC should be lower than that of the old-class OCC. We refer to this property as the comparability among OCCs. The POC framework usually encounters weak comparability due to its independent training strategy, where different OCCs are trained with non-overlap training data from different tasks, leading to quite different distributions between their output. As shown in Figure 2, a sample of class "1" is mis-classified into class "2" due to the difference in output distributions of the two OCCs between class "1" and class "2". However, established POC-based methods focus more on the enhance discriminability rather than comparability.

To tackle the above problem, we propose DisCoIL (Discriminative and Comparable One-class classifiers for Incremental Learning). The DisCoIL follows the POC framework but adopts variational auto-encoders (VAE) [16] as the one-class classifier instead of other well-established OCCs such as deep SVDD [30], for the reason that VAE can act as a pseudo-sample generator in addition to being a one-class classifier. With this advantage, DisCoIL trains each new-class VAE in contrast with old-class VAEs, which forces the new-class VAE to reconstruct better for new-class samples but worse for the old-class pseudo samples, thus enhancing the comparability. Furthermore, to ensure the discriminability, DisCoIL introduces a hinge reconstruction loss to improve the modeling robustness of each VAE for its corresponding class. In brief, this work makes the following contributions:

- We propose a novel parallel VAE architecture (DisCoIL) for incremental learning where each VAE acts as both a one-class classifier and a pseudo sample generator.
- With the generation ability of VAE, we introduce a classifier-contrastive loss and an inter-class loss to enhance the comparability between different OCCs.
- We propose a hinge reconstruction loss to improve the modeling robustness of each VAE for its corresponding class, which ensures the discriminability of DisCoIL.
- The proposed DisCoIL achieves state-of-the-art performance on several popular IL datasets.

The rest of this paper is organized as follows. Section 2 introduces related works. Section 3 presents the proposed DisCoIL. Section 4 provides experimental comparison for DisCoIL with several state-of-the-art Class-IL methods. Section 5 discusses the limitation of DisCoIL. Finally, we summarize our work in Section 6.

2. Related Work

2.1. Incremental learning

Incremental learning (IL) requires models to learn new knowledge incrementally without forgetting [24]. According to the techniques used for preventing forgetting, IL methods are usually categorized into three categories: (1) regularization-based: methods use regularization terms to alleviate forgetting [1, 7, 17, 22, 38, 41]; (2) rehearsal-based: methods keep a fixed-size buffer to store samples of old tasks [3, 5, 23, 39]; (3) dynamic architecture: methods utilize different network parameters for different tasks [14, 25, 26, 31].

Exemplar-free Class-IL is a setting of IL which adds constraints of inferencing without task identifications and training without the old-task-sample buffer. Thus, exemplar-free Class-IL is extremely challenging because models need to learn to infer the task identification of a test sample implicitly or explicitly without any old-task data, and only a few methods can satisfy this setting. This setting is significant because old-task data may not be available in some scenarios with data security or privacy restrictions. To this end,
some methods \cite{12,37} generate network parameters according to the query sample. Other methods realize multitasking classification by fusing the distribution of network parameters of different tasks \cite{20}. Besides, OWM \cite{40} achieves exemplar-free Class-IL through orthogonal data projection innovatively and shows impressive performance. Recently, some methods based on the framework of parallel one-class classifiers (POC) arise. They train an OCC for each class which predicts the probability of a test sample belonging to the corresponding class. ILCOC \cite{35} employs a deep SVDD \cite{30} model for each class. And, per class learning (PCL) \cite{13} applies a sigmoid classifier as the basic OCC model. Because of the special network structure, their parameter usage increases linearly with the number of learned classes. However, the cost is worth it, both ILCOC and PCL show promising performance under exemplar-free Class-IL. Compared with them, DisCOIL focuses more on the comparability among OCCs and shows better performance.

2.2. Auto-encoder

Auto-encoder \cite{11} is a classic network structure that consists of an encoder and a decoder, while the encoder encodes samples into latent vectors from the feature space to a latent space and the decoder reconstructs the latent vectors into reconstruction samples. To deal with the dimensionality reduction task, during training, auto-encoders minimize the reconstruction error and learn to encode information as much as possible. Since auto-encoder can only reconstruct in-distribution samples and can not reconstruct out-of-distribution samples, auto-encoders are widely applied in anomaly detection \cite{8,9}, with the reconstruction errors regarding as anomaly scores. Besides, auto-encoders can generate samples by sampling latent vectors from latent space and decoding them through the decoder. However, the latent space of the classic auto-encoder lacks regularity, which limits its generation ability. Therefore, variational auto-encoder (VAE) \cite{16} have been proposed and it encodes samples into a distribution. In addition, VAE introduces a regularization term that is a KL-divergence between distribution and standard Gaussian distribution. As a result, the latent space of VAE is more regular, and VAE has excellent generation ability.

3. Method

3.1. Problem Setting and Method Overview

Under the exemplar-free Class-IL setting, models need to learn a sequence of disjoint classification tasks incrementally. Formally, let \( Y^t \) represents the label set of task \( t \), when \( t \neq j, Y^t \cap Y^j = \emptyset \). Assume the dataset of task \( t \) is \( D^t = \{(x^t_k, y^t_k)\}_{k=1}^{n_t} \), where \( n_t \) is the number of samples in \( D^t \); \( x^t_k \) and \( y^t_k \in Y^t \) denote the \( k \)-th sample and the corresponding label, respectively. During training, the model can only learn with the data from \( D^t \); during testing, we evaluate the model with the data from datasets of all learned tasks \( \bigcup_{t=1}^{T} D^t \), where \( T \) is the number of tasks.

Figure 3 shows the overall architecture of DisCOIL, which consists of a pre-trained feature extractor and several VAE models. Each VAE serves as both an OCC and a pseudo-sample generator for a specific class. As an OCC, the reconstruction error reflects the degree of a given sample does not belonging to the corresponding classes, which thus is utilized as the anomaly score for one-class learning. As a generator, VAE can generate pseudo samples which are employed to enhance the comparability among VAEs. In addition, DisCOIL dynamically expands network capacity as the number of learned tasks increases. When a new task \( t \) comes, DisCOIL builds a new VAE model for each category in \( Y^t \) and trains them one by one.

We denote the VAE corresponding to class \( i \) as \( D_i(E_i(\cdot)) \), where \( D_i(\cdot) \) and \( E_i(\cdot) \) are the encoder and the decoder, respectively. For a query sample \( x \), ignoring the feature extraction process for simplicity, the anomaly score for class \( i \) is calculated as

\[
\text{score}_i(x) = \|D_i(E_i(x)) - x\|^2, \tag{1}
\]

and the final prediction result is determined by the class with the lowest anomaly score

\[
\hat{y} = \arg\min_i \{\text{score}_i(x)\}. \tag{2}
\]

Since the anomaly score indicates the degree to which a given sample does not fall into this category corresponds to the VAE, the discriminability within each VAE is important for determining the intrinsic recognition ability of the model. Moreover, the final prediction (Eq. (2)) needs us to compare the anomaly scores of different VAEs, thus the

![Figure 3. The architecture of DisCOIL based on variational auto-encoder (VAE).](image-url)
comparability among VAEs is also essential to the performance of DisCOIL. Furthermore, the prediction of an ideal model should satisfy the requirements shown in Figure 1.

Taking the above considerations into account, we design the training process shown in Figure 4. Assuming that we are training the VAE of class “3”, we regard the samples belonging to class “3” as positive samples and the others as negative samples. From the perspective of one-class learning, the discriminability of the VAE is enhanced by minimizing anomaly scores for positive samples. Besides, we enhance the comparability among VAEs by training each new-class VAE in contrast with previously trained VAEs. Specifically, the new-class VAE is required to reconstruct better for positive samples but worse for pseudo negative samples generated by the old-class VAEs. Next, we elaborate on the processes of enhancing the discriminability within each VAE and the comparability among VAEs in Section 3.2 and Section 3.3, respectively.

3.2. Discriminability Enhancement within VAE

As mentioned in Section 1, the performance of a POC-based model is affected by the discriminability within each OCC. Namely, a discriminative OCC is expected to output low anomaly scores for positive samples while high anomaly scores for negative samples. Thus, DisCOIL ensures the discriminability of VAEs by minimizing reconstruction errors for positive samples.

Classic reconstruction loss minimizes the mean squared error for all positive samples as small as possible, rendering the VAE susceptible to over-fitting (e.g. remembering all the training samples). Instead, under the IL scenario, the minimization of the reconstruction loss is sufficient as long as it can distinguish the positive samples from the negative samples. Therefore, we propose intra-class loss to just enforce the reconstruction errors of positive samples to be smaller than an upper bound $r_{intra}$. To be specific, when training the VAE of class $i$, the intra-class loss is defined as

$$L_{intra} = \frac{1}{|D_{i+}|} \sum_{x \in D_{i+}} \max\{0, \text{score}_i(x) - r_{intra}\}, \quad (3)$$

where $D_{i+} = \{x^t_i | (x^t_i, y^t_i) \in D^i, y^t_i = i\}$ denotes the positive sample set for class $i$, and $r_{intra} \geq 0$ is a hyperparameter shared by all VAEs. The intra-class loss relaxes the requirements for easy-to-identify samples while pays more attention to hard-to-identify samples, which enhances the generalization ability as well as the discriminability of DisCOIL.

3.3. Comparability Enhancement among VAEs

As existing POC-based methods train OCCs with non-overlap training data from different tasks, the scores output by different OCCs may have quite different distributions, causing the weak comparability problem. To tackle this problem, we introduce the classifier-contrastive loss and the inter-class loss to train a new-class VAE in contrast with old-class VAEs.

For positive samples, the new-class VAE should reconstruct better than old-class VAEs. To this end, we first acquire the reconstruction errors of the positive samples under old-class VAEs and then apply the classifier-contrastive loss to force the new-class VAE to output lower anomaly scores than the old-task VAEs.
errors than old-class VAEs, which can be formulated as
\[ L_{CC} = \frac{1}{|D^i_i|} \sum_{x \in D^i_i} \sum_{j < i} \max\{0, score_i(x) - score_j(x)\}, \tag{4} \]
where \( score_j(x) \) represents the anomaly score output by the VAE corresponding to the old class \( j \) where \( j < i \).

For a negative sample, the new-class OCC should output a higher anomaly score than the corresponding old-class OCC. Specifically, we propose inter-class loss to compel the reconstruction errors output by the new-class VAE for all negative samples higher than a lower bound \( r_{inter} \):
\[ L_{inter} = \frac{1}{|D^i_i|} \sum_{x \in D^i_i} \max\{0, r_{inter} - score_i(x')\}, \tag{5} \]
where \( D^i_i \) denotes negative sample set for class \( i \), and \( r_{inter} \) is a hyperparameter shared by all VAEs which satisfies \( r_{inter} > r_{intra} \).

With regard to the negative sample set \( D^i_i \), note that we can only access the data of the current training task under the exemplar-free Class-IL scenario. Nevertheless, the generation capability of VAE allows us to synthesize pseudo samples of old categories. Specifically, random noises are sampled from the standard Gaussian distribution and fed into the decoders of an old-class VAE:
\[ \{D_j(z) \mid z \sim \mathcal{N}(0, 1)\}, \tag{6} \]
where \( D_j(\cdot) \) represents the decoder of the VAE corresponding to class \( j \), and \( z \) is a random noise. Therefore the final negative sample set for class \( i \) is defined as
\[ D^i_i = \{x^i_k \mid (x^i_k, y^i_k) \in D^i, y^i_k \neq i\} \cup \{D_j(z) \mid z \sim \mathcal{N}(0, 1), j \in Y^s, s < t\}. \tag{7} \]

For a sample \( x' \) belonging to an old class \( j (j < i) \), recall that the intra-class loss (Eq. (3)) forces
\[ r_{intra} > score_j(x'). \tag{8} \]
In addition, minimizing Eq. (5) encourages
\[ score_i(x') > r_{inter}. \tag{9} \]
Combining Eqs. (8) and (9) with \( r_{inter} > r_{intra} \) implies:
\[ score_i(x') > score_j(x'). \tag{10} \]
Equation (10) indicates that jointly optimizing the intra-loss and the inter-loss indeed enhance the comparability between different OCCs. Furthermore, \( r_{inter} - r_{intra} \) can be regarded as the comparability margin.

Note that although the proposed intra-class loss and inter-class loss seem to share a similar form with the triplet loss [32]. However, they are fundamentally different in terms of mechanism and function. The triplet loss encourages the features from the same category to be similar while pushing features of different classes far away, which only applies to a single model. In contrast, the proposed intra and inter loss compel the output of different VAEs distributing in the same range, and the score gap between the positive and negative samples by a margin \( r_{inter} - r_{intra} \). In other words, they are designed to enhance a cross-model property, i.e. the compatibility.

Finally, the overall loss function is
\[ L = L_{intra} + \lambda_1 L_{CC} + \lambda_2 L_{inter} + L_{KL}, \tag{11} \]
where \( \lambda_1 \) and \( \lambda_2 \) are weight parameters, and \( L_{KL} \) is the regularization term in VAE [16] to pull the latent distribution to match the standard Gaussian distribution.

4. Experiments
We evaluate the performance of the proposed DisCOIL in this section. All experiments are conducted on a workstation running OS Ubuntu 16.04 with 18 Intel Xeon 2.60GHz CPUs, 256 GB memory, and 6 NVIDIA RTX3090 GPUs.

4.1. Experimental Setting
Datasets. We select three widely used image classification datasets as our benchmarks and divide each of them into several incremental batches. The statistics of these datasets are shown in the following:

- **MNIST** [19]: a widely used handwritten digital character dataset which contains 60000 images of size 28 × 28 from 10 classes. In our experiments, we divide all classes into five incremental batches equally. Namely, each batch contains samples of two classes.
- **CIFAR10** [18]: a dataset contains 60000 images of size 32 × 32 from 10 classes. Similarly, we divide it into five incremental batches equally.
- **Tiny-ImageNet** [34]: the largest dataset in our experiment with 200 classes, and there are 600 images for each class. We split the 200 classes into ten incremental batches, and each batch contains samples of 20 classes.

Evaluation metric. Following the setting in [3,23], after learning all tasks, we evaluate models with the data of all learned tasks and report the average accuracy over five runs with random seeds.

Baselines. We select several classic and latest state-of-the-art Class-IL methods: (1) LWF [22] uses knowledge distillation technique to regularize activations of neural network; (2) SI [41] regards contributions of loss reduction as the importance of network parameters and punishes their...
changes; (3) **IMM** [20] matches the moment of the posterior distribution of network parameters trained from different tasks; (4) **iCaRL** [29] uses samples in the buffer to calculate class centers; (5) **oEWC** [33] uses the fisher matrix to calculate the importance weights of network parameters and restricts their changes; (6) **FDR** [2] utilizes samples in the buffer to create a function space and measures the importance of network parameters with the function space; (7) **PGMA** [12] generates network parameters according to query samples; (8) **OWM** [40] avoids forgetting through orthogonal data projection; (9) **HAL** [4] keeps anchors for old tasks to alleviate forgetting; (10) **EBM** [21] solves the bias problem by using energy model; (11) **ILCOC** [35] applies parallel deep SVDD model architecture to deal with the Class-IL problem; (12) **PCL** [13] is a POC-based method that creates a sigmoid classifier for each class. Besides, we also provide two base methods: (13) **Joint** trains a neural network with all task data jointly, and its performance can be regarded as the upper bound of Class-IL methods; (14) **Fine-tune**, the lower bound of Class-IL methods that directly fine-tune neural network parameters in each task without using any incremental learning strategy.

**Implementation detail.** We select vgg11 as the shared feature extractor for all methods as it has simple network architecture and low memory usage. When applying the pre-trained feature extractor, DisCOIL can use a very shallow network as the VAE for each class, to keep their memory footprint low. As for the VAE in our method, both the encoder and decoder are composed of three fully connected layers followed by a ReLU activation function. Although the size of DisCOIL will linearly increase with the number of classes, the overall storage overhead is acceptable as the size of each VAE is actually small, e.g. 0.98M for MNIST and 9.96M for the other two benchmarks.

**Hyper-parameter selection.** We train each model with 10, 50, and 100 epochs for MNIST, CIFAR10, and Tiny-ImageNet, respectively. Other hyperparameters are determined through a validation set which is obtained by sampling 10 percent samples from the training set. For DisCOIL, we choose Adam [15] to optimize the model with the weight decay factor of 0.001 for CIFAR10 and 0.01 for others. Both $\lambda_1$ and $\lambda_2$ are tuned from the range of (0, 0.1, 0.2, 0.5, 1, 2, 5, 10); $r_{intra}$ is tuned from 0 to 100 and $r_{inter}$ is tuned from 50 to 1000. More details can be found in our supplementary code.

### 4.2. Comparison Results

In this section, we compare the performance of our DisCOIL method with all baselines. Table 1 reports the comparison results on three classification datasets. The results show that all methods benefit from the feature extractor because it reduces the difficulty of training. Besides, the fine-tune reports poor average accuracies ($\approx 1/T$) on all datasets, which proves the existence of the forget-

| Approach | MNIST | CIFAR10 | Tiny-ImageNet |
|----------|-------|---------|---------------|
|          | W/O   | W/      | W/O           |
| Joint    | 97.68 | 92.20   | 59.99         |
| Fine-tune| 19.94 | 23.73   | 19.62         |
| LWF [22] | 21.53 | 21.43   | 19.61 †       |
| SI [41]  | 20.81 | 27.43   | 19.48 †       |
| IMM [20] | 67.25 †| -       | 32.36 †       |
| iCaRL*   | 71.41 | 71.15   | 49.02 †       |
| oEWC     | 20.36 | 28.27   | 19.49 †       |
| FDR*     | 81.25 | 59.62   | 30.91 †       |
| PGMA     | 81.70 †| -       | 40.47 †       |
| OWM [40] | 96.30 † | 75.39 | 52.83 †       |
| HAL*     | 80.98 | 59.29   | 32.36 †       |
| EBM [21] | 53.12 †| -       | 38.84 †       |
| ILCOC    | 86.05 | 69.04   | 38.40 †       |
| PCL [13] | 95.75 | 68.79   | -             |
| DisCOIL  | 96.69 | 77.35   | 44.54         |

Table 1. Incremental learning results for standard classification benchmarks. We report the average accuracy (%) over five runs with different random seeds, and the higher is the better. The row “PTF” indicates that models with (w/) or without (w/o) a pre-trained feature extractor. The results with “†” are quoted from the original paper, and the results with “‡” are quoted from [3]. “-” indicates experiments we were unable to run, because of unavailable source code or intractable training. Note that the methods with “*” keep a buffer with a maximum size of 200 during incremental learning which can not be fairly compared with the other exemplar-free methods.
Table 2. Average accuracy (%) of ablation experiments on CIFAR10 with the pre-trained feature extractor.

| Task No. | Var 1 | Var 2 | Var 3 | DisCOIL |
|---------|-------|-------|-------|---------|
| No. 1   | 97.43 | 95.99 | 97.40 | 97.32   |
| No. 2   | 91.56 | 89.29 | 91.54 | 91.47   |
| No. 3   | 82.00 | 80.53 | 82.08 | 82.32   |
| No. 4   | 78.40 | 77.15 | 78.44 | 78.46   |
| No. 5   | 76.97 | 75.74 | 76.89 | 77.35   |

Figure 5. Sensitivity analysis of $r_{\text{intra}}$ (a) and $r_{\text{inter}}$ (b) on CIFAR10.

- **Effectiveness of classifier-contrastive loss.** We first analyze the effectiveness of the classifier-contrastive loss by comparing Var 1 and DisCOIL. When learning task 1, the absence of the old-task OCC makes the classifier-contrastive loss useless, which may result in Var 1 outperforming DisCOIL. As the number of learned tasks increases, DisCOIL gradually outperforms Var 1. This phenomenon indicates that the classifier-contrastive loss can only improve the comparability among VAEs trained from different tasks, since the VAEs from the same task already have enough comparability with the help of inter-class loss.

- **Effectiveness of inter-class loss.** As illustrated in Table 2, the performance of DisCOIL is greater than Var 2 at any incremental stage. The promotion comes from the following two aspects: (1) by utilizing negative samples from each task dataset, the inter-class loss can improve the comparability among VAEs from the same task; (2) the result of Var 3 shows that the pseudo samples generated by old-task VAEs can assist the inter-class loss to improve the comparability among VAEs from different tasks.

### 4.4. Sensitivity Analysis

**Sensitivity analysis of $r_{\text{intra}}$.** In this section, we investigate the influence of the upper bound $r_{\text{intra}}$ in the intra-class loss. We select five different values, including 0, 0.1, 1, 10, 100. As illustrated in Figure 5 (a), the model has a poor performance when $r_{\text{intra}} = 0$ since the intra-class loss is degraded into MSE loss. Besides, as $r_{\text{intra}}$ increasing, the average accuracy grows and reaches the peak when $r_{\text{intra}} = 10$. When $r_{\text{intra}} = 100$, the performance drops again because VAEs do not learn to decrease the reconstruction error and lack of discriminability for one-class classification. In addition, we notice that $r_{\text{intra}}$ has little effect on model performance when applying the pre-trained feature extractor because the feature extractor can smooth the input of VAEs and alleviate over-fitting.

**Sensitivity analysis of $r_{\text{inter}}$.** In addition, we evaluate the effect of the lower bound in inter-class loss $r_{\text{inter}}$. We...
also consider five different values that greater than $r_{\text{intra}}$, including 50, 100, 200, 500, 1000. As illustrated in Figure 5 (b), with $r_{\text{inter}}$ increasing, the average accuracy is also gradually growing. Because more samples satisfy Eq. (8-10) and ensure high comparability between different OCCs when the margin $r_{\text{inter}} - r_{\text{intra}}$ becomes larger. It can observed the performance drops when $r_{\text{inter}} = 1000$, since a large $r_{\text{inter}}$ makes the model occupies excessive network capacity to reconstruction worse on old-class samples. Besides, when applying the feature extractor, the optimal value of $r_{\text{inter}}$ is greater than that without the feature extractor, the reason is that the feature extractor can reduce the difficulty of reconstructing negative samples and meanwhile does not hurt the performance of VAEs.

4.5. Misclassification Analysis

As mentioned in Section 1, the performance of POC-based methods depends on the discriminability and the comparability of OCCs. This section analyzes the misclassifications of DisCOIL, ILCOC, and PCL. Specifically, for a misclassified sample $x$ and its label $y$, we first feed the scores of all OCCs into the softmax function to convert them into probabilities. If the OCC corresponding to class $y$ predicts a too low probability (lesser than $1/|\sum_{i} |Y'||$ in our experiment), we believe that the misclassification is caused by the low discriminability. Otherwise, the low comparability among OCCs leads to this misclassification. Figure 6 reports the statistical results of all misclassified samples, and we can see that all POC-based methods suffer from lacking comparability among OCCs. Besides, the OCCs of DisCOIL have more discriminability and comparability than others.

5. Discussion

DisCOIL enjoys the advantages of the POC framework, but it also inherits partial flaws of POC. One drawback is that the parameter usage of DisCOIL linearly increases with the number of learned classes. To reduce the impact of excessive network parameters, we recommend equipping DisCOIL with a pre-trained feature extractor, since it can not only improve the model performance (see Table 1) but also allow a smaller network to implement VAEs. Another feasible approach is to combine DisCOIL with the mask weights learning techniques in Task-IL methods, such as Piggy-back [25], CPG [14] and PackNet [26], by taking each VAE training as an independent task. These techniques enable DisCOIL to adapt a single network to multiple one-class classifiers. As this paper focuses on solving the weak discriminability and comparability of POC, we leave the above exploration for future work.

6. Conclusion

In this paper, to solve the weak discriminability and comparability of the POC IL framework, we propose a novel IL method DisCOIL based on parallel VAE architecture. With the advantage of VAE, a joint loss is designed to compel DisCOIL to satisfy a cross-model relation, which enhances the comparability between different OCCs. Furthermore, DisCOIL introduces a hinge reconstruction loss to improve the discriminability. We evaluate DisCOIL on several classification datasets, including MNIST, CIFAR10, and Tiny-ImageNet. The experimental results show that DisCOIL achieves state-of-the-art performance.

Some further studies are left in the future. Firstly, we will try to integrate the mask weights learning techniques of Task-IL methods into DisCOIL to reduce the parameter usage. Secondly, we will investigate the distribution alignment technique to improve the comparability among OCCs and the performance of POC-based methods.

References

[1] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In European Conference on Computer Vision, pages 139–154, 2018. 2
[2] Ari S. Benjamin, David Rolnick, and Konrad P. Körding. Measuring and regularizing networks in function space. In International Conference on Learning Representations, 2019. 6
[3] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and SIMONE CALDERARA. Dark experience for general continual learning: A strong, simple baseline. In Advances in Neural Information Processing Systems, volume 33, pages 15920–15930, 2020. 1, 2, 5, 6
[4] Arslan Chaudhry, Albert Gordo, Puneet Kumar Dokania, Philip Torr, and David Lopez-Paz. Using hindsight to an-
[33] Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pas- canu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. In International Conference on Machine Learning, pages 4528–4537, 2018. 6

[34] Stanford. Tiny imagenet challenge (cs231n). 2015. 5

[35] Wenju Sun, Jing Zhang, Danyu Wang, Yangli-ao Geng, and Qingyong Li. Ilcoc: An incremental learning framework based on contrastive one-class classifiers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 3580–3588, June 2021. 1, 2, 3, 6

[36] Gido M Van de Ven and Andreas S Tolias. Three scenarios for continual learning. arXiv preprint arXiv:1904.07734, 2019. 1, 7

[37] Johannes von Oswald, Christian Henning, João Sacramento, and Benjamin F. Grewe. Continual learning with hypernetworks. In International Conference on Learning Representations, 2020. 3

[38] Shipeng Wang, Xiaorong Li, Jian Sun, and Zongben Xu. Training networks in null space of feature covariance for continual learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 184–193, June 2021. 2

[39] 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, pages 374–382, 2019. 2

[40] Guanxiong Zeng, Yang Chen, Bo Cui, and Shan Yu. Continual learning of context-dependent processing in neural networks. Nature Machine Intelligence, 1(8):364–372, 2019. 3, 6

[41] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In International Conference on Machine Learning, pages 3987–3995, 2017. 2, 5, 6