Co^{2}L: Contrastive Continual Learning

Shiwon Kim,
Digital Healthcare Lab,
Department of Digital Analytics,
Yonsei University College of Computing

Abstract

- Recent breakthroughs in self-supervised learning (SSL)
  - SSL learn **transferable representations** better than cross-entropy based supervised methods
  - This paper suggests that **the similar holds in the continual learning (CL) context**
  - Contrastively learned representations are more robust against the catastrophic forgetting
- Co^{2}L (**Co**ntrastive **Co**ntinual **L**earning)
  - Focuses on continually learning and maintaining transferable representations (replay-based)

  1) Learns representations using the **contrastive learning objective**
  2) Preserves learned representations using a **self-supervised distillation** setup

Introduction

Motivation for the proposed method

- Previous continual learning approaches
  - Focus on **preserving the previously learned knowledge** using various past information
  - Replay-based methods rehearse **a small portion of past samples** along with current samples
  - Regularization-based approaches force the current model to **be close to the past model**
  - Expansion-based methods allocate a unit (e.g., network node, sub-network) for each task
- Proposed approach
  - Instead of asking how to isolate previous knowledge from new knowledge…

*What type of knowledge is likely to be **useful for future tasks (and thus not get forgotten)**,*

and how can we learn and preserve such knowledge?

Introduction

Learning transferable representations

- Forgetting of *future events*
  - e.g., Task 1 (apple vs. banana) → Task 2 (apple vs. strawberry)
  - The color is critical for task 1 but **no longer useful for task 2** and eventually get forgotten
  - Forgetting does not only come from the limited access to the past experience
  - It also comes from the innately **restricted access to future events**
- Significance of learning transferable representations
  - More complicated features (e.g., shape, polish, texture) may be **re-used for future tasks**
  - **Learning transferable representations** is as important as preserving the past knowledge

Introduction

Contrastive learning

- Recent advancement in contrastive methods
  - Use the inductive bias that the prediction should be **invariant to input transformations**
  - Contrastive methods are known to be surprisingly effective despite their simplicity
  - Closely achieve the fully-supervised performance even without labels[1]
  - **Outperform its counterparts in the supervised case** for ImageNet classification[2]
- Contrastive learning in a continual setup
  - A similar observation is made in this paper **under a continual scenario**
  - Contrastively learned representations **suffer less forgetting** than those trained with CE loss

1. Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." International conference on machine learning. PMLR, (2020).
2. Khosla, Prannay, et al. "Supervised contrastive learning." Advances in neural information processing systems 33. (2020).

Introduction

Challenges of applying contrastive learning to continual settings

- Limited access to negative samples
  - **Having informative negative samples** is crucial for the success of contrastive learning[3]
  - The learner can access samples from **only a small number of classes** at each time step in CL
  - Instantaneous demographics of negative samples are highly restricted under continual setups
- Preserving the contrastively learned representations
  - Recent works aim to learn representations accelerating future learning
  - Lack of an explicit design to **preserve representations**

3. Robinson, Joshua, et al. "Contrastive learning with hard negative samples." arXiv preprint arXiv:2010.04592 (2020).

Introduction

Contributions (Co²L)

1. *Contrastive learning*
   - Design an **asymmetric version of supervised contrastive loss** under continual setup
2. *Preserving representations*
   - Propose a novel **preservation mechanism** for contrastively learned representations on CL
   - Maintain representations using **self-distillation*** **of instance-wise relations**

- Quantitative validation
  - **Outperforms all baselines** on various datasets, CL scenarios, and memory setups
  - Ablation studies show that both components proposed are essential for performance

* **(a) Use predictions from a "teacher" to train a "student" with the same architecture**
(b) Transfer knowledge in the same model from the deeper layers to the shallow layers

Related Works

Rehearsal-based continual learning

- Experience replay (ER)
  - Manage a fixed-sized buffer to **retain a few samples and replay those** to prevent forgetting
  - Focus on either regulating model updates or selecting samples
    - Regulate model updates not to contradict the learning objectives on past samples
    - Select the most representative samples to prevent changes in past predictions
  - Not many studies related to ER in a **decoupled representation learning setup**
  - Representation learning objectives may not be directly aligned to task-specific objectives
  - We focus on utilizing buffered samples to learn representations continually

Related Works

Representation learning in continual learning

- Continual learning of representations
  - Only a few studies on continual learning focus on representations **in two aspects**
    1) How to *maintain learned representations*
    2) How to learn representations *accelerating future learning*
  - Previous studies
    - Prevent representations from being forgotten by leveraging distillation
    - Learn representations that accelerate future learning on meta-learning frameworks
    - Exploit self-supervised learning objectives to learn more generalizable representations
  - We use a contrastive scheme with additional components to preserve learned representations

Related Works

Contrastive representation learning

- Recent progress in contrastive representation learning
  - Superior downstream task performance even **comparable to supervised training**
  - Advances in this area stems from the use of multiple views as positive samples
  - Practical limitations resolved by previous studies
    - *Negative sample pairs*
    - *Large batch size*
  - Supervised learning can also enjoy the benefits of contrastive representation learning[2]
  - We mainly leverage contrastive representation learning schemes on the CL setup

2. Khosla, Prannay, et al. "Supervised contrastive learning." Advances in neural information processing systems 33. (2020).

Related Works

Knowledge distillation

- Knowledge distillation (KD) in continual learning
  - Widely used to mitigate forgetting by distilling past signatures to the current models
  - Has not been studied to utilize KD for **decoupled representation training in the CL setup**
  - We develop novel self-distillation loss for *contrastive continual learning*

Problem Setup and Preliminaries

Problem setup: continual learning

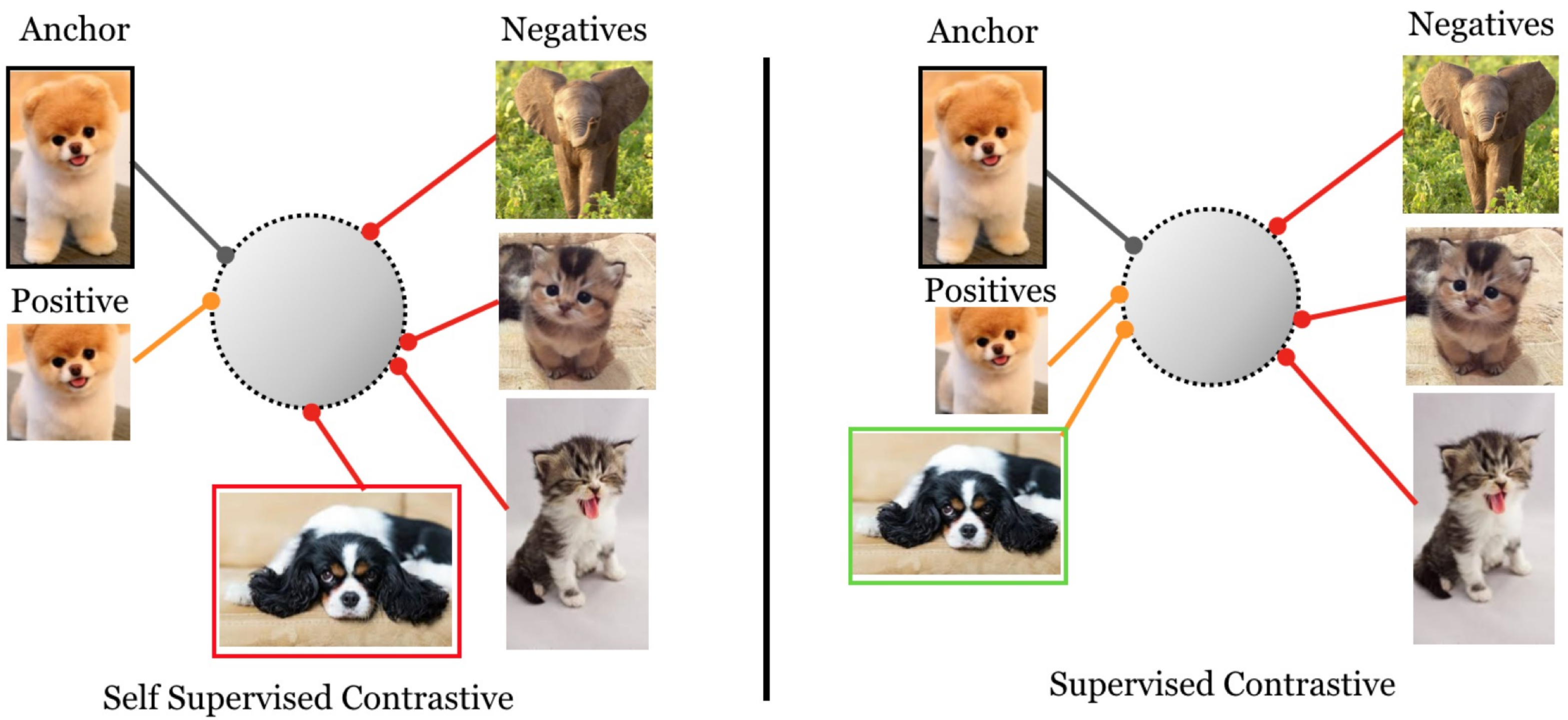
- Class-incremental learning (CIL)
  - The learner is trained on a sequence of tasks indexed by $t \in \{1, 2, \dots, T\}$
  - For a task-specific class set C_t, $\{C_t\}_{t=1}^{T}$ are assumed to be disjoint, *i.e.*, $t \neq t' \Longrightarrow C_t \cap C_{t'} = \emptyset$
  - n_t copies of task-specific input-label pairs during each task, *i.e.*, $\{(\mathrm{x}_i, y_i)\}_{i=1}^{n_t} \sim D_t$
  - The goal is to find a predictor $\varphi_\theta(\mathrm{x})$ minimizing the following loss function:

$$\mathcal{L}(\theta) = \sum_{t=1}^{T} \mathbb{E}_{D_t}\left[\ell\left(y, \varphi_\theta(\mathrm{x})\right)\right]$$

Problem Setup and Preliminaries

Preliminaries: contrastive learning

- Supervised contrastive learning (SupCon)[2]



2. Khosla, Prannay, et al. "Supervised contrastive learning." Advances in neural information processing systems 33. (2020).

Problem Setup and Preliminaries

Preliminaries: contrastive learning

- Training setup
  - $\varphi_\theta = \mathrm{w} \circ f_\vartheta$ → classification model with parameter pairs $\theta = (\vartheta, \mathrm{w})$
  - $\mathrm{w}(\cdot)$ is the linear classifier and $f_\vartheta(\cdot)$ is the representation

batch $\mathcal{B}$

N training samples

$\{(\mathrm{x}_i, y_i)\}_{i=1}^{N}$

$(\mathrm{x}_\mathrm{k}, y_\mathrm{k})$

N augmented pairs

$\{(\tilde{\mathrm{x}}_{2\mathrm{k}-1}, \tilde{\mathrm{y}}_{2\mathrm{k}-1})\}_{\mathrm{k}=1}^{N}$

$\tilde{\mathrm{y}}_{2\mathrm{k}-1} = \tilde{\mathrm{y}}_{2\mathrm{k}} = y_k$

N augmented pairs

$\{(\tilde{\mathrm{x}}_{2\mathrm{k}}, \tilde{\mathrm{y}}_{2\mathrm{k}})\}_{\mathrm{k}=1}^{N}$

batch $\mathcal{B}_a$

$2N$ augmented samples

$\{(\tilde{\mathrm{x}}_i, \tilde{\mathrm{y}}_i)\}_{i=1}^{2N}$

Problem Setup and Preliminaries

Preliminaries: contrastive learning

- Supervised contrastive loss
  - Samples in the augmented batch are mapped to a unit d-dimensional Euclidean sphere as:

$$\mathrm{z}_i = (g \circ f)_\psi(\tilde{\mathrm{x}}_i)$$

  - The feature map $(g \circ f)_\psi$ is trained to minimize the supervised contrastive loss:

$$\mathcal{L}^{\mathrm{sup}} = \sum_{i=1}^{2N} \frac{-1}{|\mathfrak{p}_i|} \sum_{j \in \mathfrak{p}_i} \log\left(\frac{\exp\left(\mathrm{z}_i \cdot \mathrm{z}_j / \tau\right)}{\sum_{k \neq i} \exp(\mathrm{z}_i \cdot \mathrm{z}_k / \tau)}\right)$$

$$\mathfrak{p}_i = \left\{ j \in \{1, \dots, 2N\} \,\middle|\, j \neq i,\ y_j = y_i \right\} \ \cdots\ \textit{index set of positive samples on anchor } \tilde{\mathrm{x}}_i$$

Contrastive Continual Learning (Co²L)

A rehearsal-based contrastive learning scheme

- Overview
  - A mini-batch ($2N$ augmented samples) gradient descent based on the compound loss:

$$\mathcal{L} = \mathcal{L}_{\text{asym}}^{\text{sup}} + \lambda \cdot \mathcal{L}^{\text{IRD}}$$

1) *Learning*
   - Learns the representations with an **asymmetric form of supervised contrastive loss**
2) *Preserving*
   - Preserves learned representations using **self-supervised distillation**
   - A decoupled representation-classifier training scheme

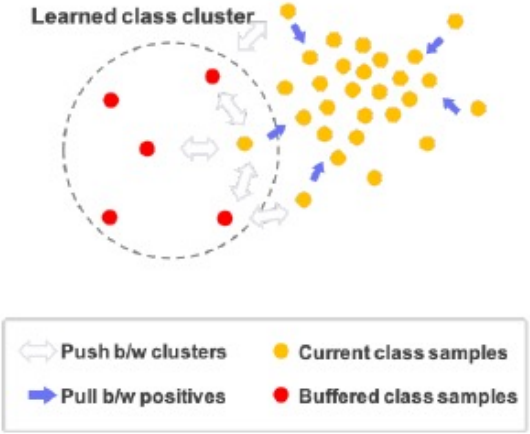
Contrastive Continual Learning (Co²L)

(1) Representation learning with asymmetric supervised contrastive loss

- Asymmetrically modified version of $\mathcal{L}^{\text{sup}}$
  - **Prevent overfitting** to small-sized past samples → only use **current samples** as anchors
  - Past task samples from **the memory buffer** are only used as negative samples
  - Contrasts anchor samples from the current task against the samples from other classes
  - Provides a more *transferable representation*

$$\mathcal{L}^{\text{sup}}_{\text{asym}} = \sum_{i \in S} \frac{-1}{|\mathfrak{p}_i|} \sum_{j \in \mathfrak{p}_i} \log \left(\frac{\exp(z_i \cdot z_j / \tau)}{\sum_{k \neq i} \exp(z_i \cdot z_k / \tau)} \right)$$

$$S \subset \{1, \ldots, 2N\}$$



(a) Asymmetric SupCon Loss

Contrastive Continual Learning (Co²L)

(2) Instance-wise relation distillation (IRD) for contrastive continual learning

- Explicit mechanism to preserve the learned knowledge
  - Regulates the **changes in feature relation** between batch samples via **self-distillation**
  - Quantifies discrepancy between instance-wise similarities of current and past representation
  - **Instance-wise similarity vector** for each sample $\tilde{\mathrm{x}}_i$ in a batch $\mathcal{B}$:

$$\mathbf{p}(\tilde{\mathrm{x}}_i; \psi, \kappa) = \left[p_{i,1}, \dots, p_{i,i-1}, p_{i,i+1}, \dots, p_{i,2N}\right]$$

  - Normalized **instance-wise similarity**:

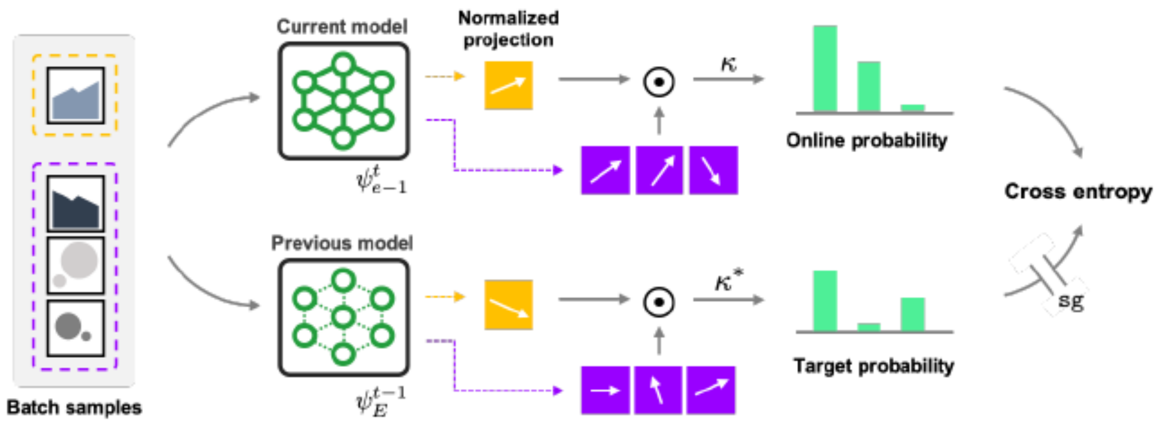
$$p_{i,j} = \frac{\exp\left(\mathrm{z}_i \cdot \mathrm{z}_j / \kappa\right)}{\sum_{k \neq i}^{2N} \exp(\mathrm{z}_i \cdot \mathrm{z}_k / \kappa)}$$

Contrastive Continual Learning (Co²L)

(2) Instance-wise relation distillation (IRD) for contrastive continual learning

- Explicit mechanism to preserve the learned knowledge
  - Denote the parameters of the past and current model as ψ^{past} and ψ
  - Use fixed weights snapped at the end of previous task training as the **reference model** ψ^{past}
  - **Minimize the drift** of the instance-wise similarities given by ψ from the ones given by ψ^{past}

$$\mathcal{L}^{\text{IRD}} = \sum_{i=1}^{2N} -\mathbf{p}\left(\tilde{\mathrm{x}}_i; \psi^{\text{past}}, \kappa^*\right) \cdot \log \mathbf{p}\left(\tilde{\mathrm{x}}_i; \psi, \kappa\right)$$



(b) Instance-wise Relation Distillation Loss

Contrastive Continual Learning (Co^{2}L)

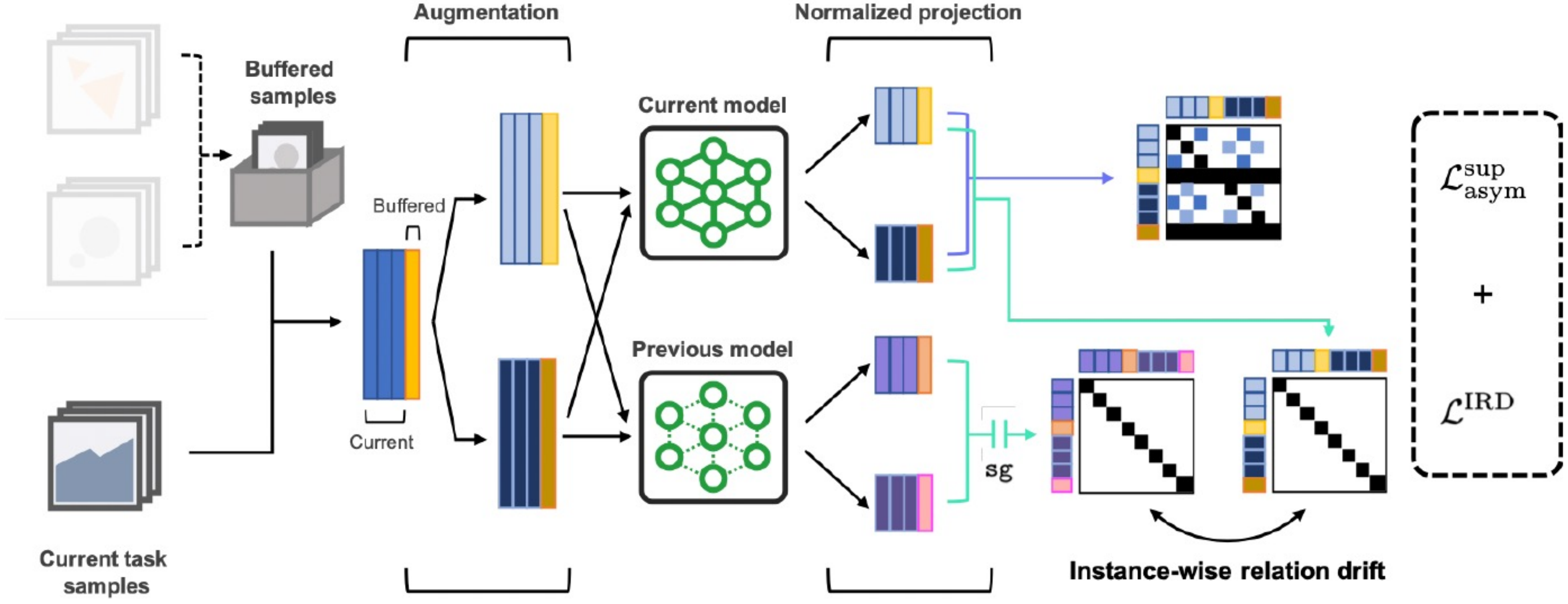
(3) Algorithm details

- Data preparation
  - Dataset is built as a union of current samples and buffered samples without oversampling
  - Each sample in a mini-batch is **augmented into two views (rotation)**
- Learning new representations
  - Forward augmented samples to the **encoder** f_ϑ and the **projection map** g_ϕ sequentially
  - Projection map outputs are used to compute asymmetric supervised contrastive loss
- Preserving learned representation
  - Compute instance-wise relations drifts between reference model and the training model
  - The reference model is **not updated** while optimizing the total loss (**stop-gradient**)

Contrastive Continual Learning (Co²L)

(3) Algorithm details

- Overall architecture



Experiment

Experimental setup

- Learning scenarios and datasets
  - **Seq-CIFAR-10** (2 classes * 5) and **Seq-Tiny-ImageNet** (20 classes * 10) → Task-IL / Class-IL
  - **R-MNIST** (20 tasks corresponding to 20 uniformly randomly chosen degrees) → Domain-IL
- Rehearsal-based continual learning baselines
  - ER, iCaRL, GEM, A-GEM, FDR, GSS, HAL, DER, DER++ (ResNet-18)
  - Buffer size 200 and 500
- Evaluation protocol
  - **Train a linear classifier** using current and buffer on top of frozen representations from Co^2L
  - 100 epochs for all experiments

Experiment

Main results

- Validation of key hypothesis

*Contrastive learning learns **more useful representation for the future task***

than the cross-entropy based coupled representation-classifier supervised learning

Experiment

Main results

- Validation of key hypothesis
  - Contrastively learned representations **suffer less forgetting** than those trained with CE loss
  - Learns highly **transferable representations** useful for future tasks on unseen objects

Accuracy on Seen classes (cross-entropy)

	task1	task2	task3	task4	task5
airplane automobile	98.65	64.00	17.05	21.65	15.85
bird cat	-	72.25	27.85	16.35	16.00
deer dog	-	-	82.10	23.20	2.50
frog horse	-	-	-	91.80	13.45
ship truck	-	-	-	-	87.35

Accuracy on Seen classes (contrastive)

	task1	task2	task3	task4	task5
airplane automobile	99.05	89.75	82.05	79.35	80.55
bird cat	-	80.10	64.00	47.90	33.65
deer dog	-	-	77.35	66.35	48.05
frog horse	-	-	-	84.05	72.95
ship truck	-	-	-	-	84.20

Accuracy on All classes (cross-entropy)

	task1	task2	task3	task4	task5
airplane automobile	71.30	40.35	7.55	14.55	14.60
bird cat	25.25	58.30	15.30	14.25	14.55
deer dog	25.75	16.55	78.25	14.40	1.50
frog horse	52.60	34.60	35.65	91.15	13.90
ship truck	59.65	42.15	11.10	24.40	87.50

Accuracy on All classes (contrastive)

	task1	task2	task3	task4	task5
airplane automobile	85.35	70.85	66.95	65.80	80.55
bird cat	36.55	61.65	49.95	45.75	33.65
deer dog	46.00	56.75	74.50	64.75	48.05
frog horse	70.50	69.30	72.70	83.90	72.95
ship truck	76.70	66.60	65.90	66.55	84.20

Experiment

Main results

Buffer	Dataset	Seq-CIFAR-10		Seq-Tiny-ImageNet		R-MNIST
	Scenario	Class-IL	Task-IL	Class-IL	Task-IL	Domain-IL
200	ER [34]	44.79±1.86	91.19±0.94	8.49±0.16	38.17±2.00	93.53±1.15
	GEM [29]	25.54±0.76	90.44±0.94	-	-	89.86±1.23
	A-GEM [8]	20.04±0.34	83.88±1.49	8.07±0.08	22.77±0.03	89.03±2.76
	iCaRL [33]	49.02±3.20	88.99±2.13	7.53±0.79	28.19±1.47	-
	FDR [4]	30.91±2.74	91.01±0.68	8.70±0.19	40.36±0.68	93.71±1.51
	GSS [2]	39.07±5.59	88.80±2.89	-	-	87.10±7.23
	HAL [7]	32.36±2.70	82.51±3.20	-	-	89.40±2.50
	DER [5]	61.93±1.79	91.40±0.92	11.87±0.78	40.22±0.67	96.43±0.59
	DER++ [5]	64.88±1.17	91.92±0.60	10.96±1.17	40.87±1.16	95.98±1.06
	Co^{2}L (ours)	**65.57**±1.37	**93.43**±0.78	**13.88**±0.40	**42.37**±0.74	**97.90**±1.92
500	ER [34]	57.74±0.27	93.61±0.27	9.99±0.29	48.64±0.46	94.89±0.95
	GEM [29]	26.20±1.26	92.16±0.64	-	-	92.55±0.85
	A-GEM [8]	22.67±0.57	89.48±1.45	8.06±0.04	25.33±0.49	89.04±7.01
	iCaRL [33]	47.55±3.95	88.22±2.62	9.38±1.53	31.55±3.27	-
	FDR [4]	28.71±3.23	93.29±0.59	10.54±0.21	49.88±0.71	95.48±0.68
	GSS [2]	49.73±4.78	91.02±1.57	-	-	89.38±3.12
	HAL [7]	41.79±4.46	84.54±2.36	-	-	92.35±0.81
	DER [5]	70.51±1.67	93.40±0.39	17.75±1.14	51.78±0.88	97.57±1.47
	DER++ [5]	72.70±1.36	93.88±0.50	19.38±1.41	51.91±0.68	97.54±0.43
	Co^{2}L (ours)	**74.26**±0.77	**95.90**±0.26	**20.12**±0.42	**53.04**±0.69	**98.65** ±0.31

Experiment

Ablation studies

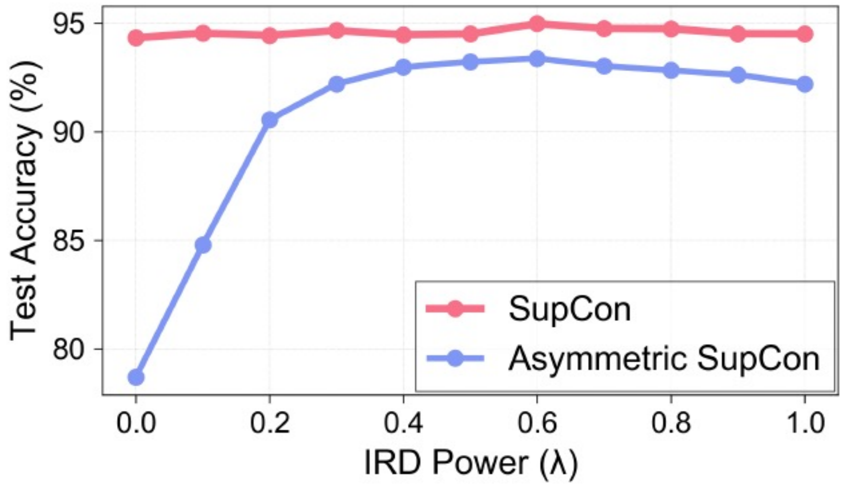
- Effectiveness of IRD
  - Experiments with the class-IL setup on the Seq-CIFAR-10 dataset
    a) *Without buffer and IRD* → optimize using only the **symmetric SupCon loss**
    b) *With IRD only* → use both symmetric SupCon loss and IRD loss
    c) *With replay buffer only* → optimize the **asymmetric SupCon loss**

	Buffer Size	IRD	Accuracy(%)
(a) w/o buffer and IRD	0	✗	53.25±1.70
(b) w/ IRD only	0	✓	58.89±2.61
(c) w/ buffer only	200	✗	53.57±1.03
(d) Co^{2}L(ours)	200	✓	**65.57±1.37**

Experiment

Ablation studies

- Effectiveness of IRD
  - Train with symmetric and asymmetric SupCon loss on an *infinite-buffer* class-IL scenario
  - Asymmetric SupCon performs poor without IRD → **gap closes with increasing IRD power**
  - Not using past samples as positive pairs only restricts learning under class-balanced setup



Experiment

Ablation studies

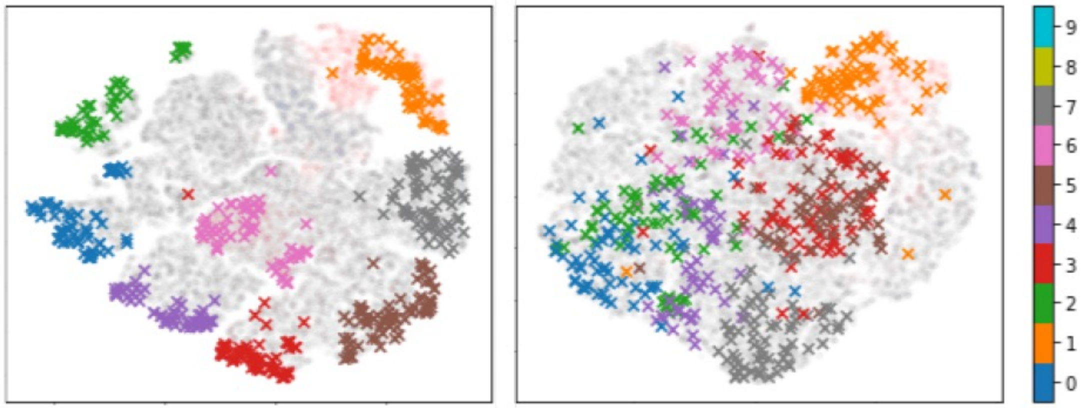
- Effectiveness of asymmetric supervised contrastive loss
  - The original SupCon loss versus the asymmetric SupCon loss combined with the IRD loss

	Seq-CIFAR-10		Seq-Tiny-ImageNet	
Buffer	200	500	200	500
$\mathcal{L}^{\text{sup}}$	60.49±0.72	68.66±0.68	13.51±0.48	19.68±0.62
$\mathcal{L}^{\text{sup}}_{\text{asym}}$	**65.57**±**1.37**	**74.26**±**0.77**	**13.88**±**0.40**	**20.12**±**0.42**

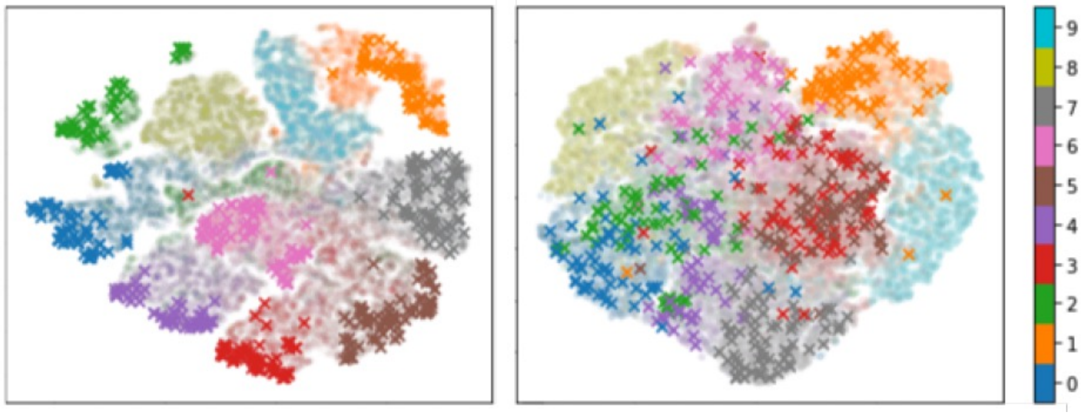
Experiment

Ablation studies

- Effectiveness of asymmetric supervised contrastive loss
  - **t-SNE visualization** of features from buffered and entire training samples of Seq-CIFAR-10



Original SupCon　　Asymmetric SupCon



Original SupCon　　Asymmetric SupCon

YONSEI UNIVERSITY
COLLEGE OF MEDICINE