A Unified Framework for Continual Learning and Unlearning

Romit Chatterjee, Vikram Chundawat, Ayush Tarun, Ankur A Mali, Murari Mandal

RespAI Lab, KIIT Bhubaneswar, SagepilotAI, EPFL, University of South Florida

Techniques Used Previously

Continual Learning

FIM, Replay-based Methods, Regularization-based Methods, Knowledge Distillation

Unlearning

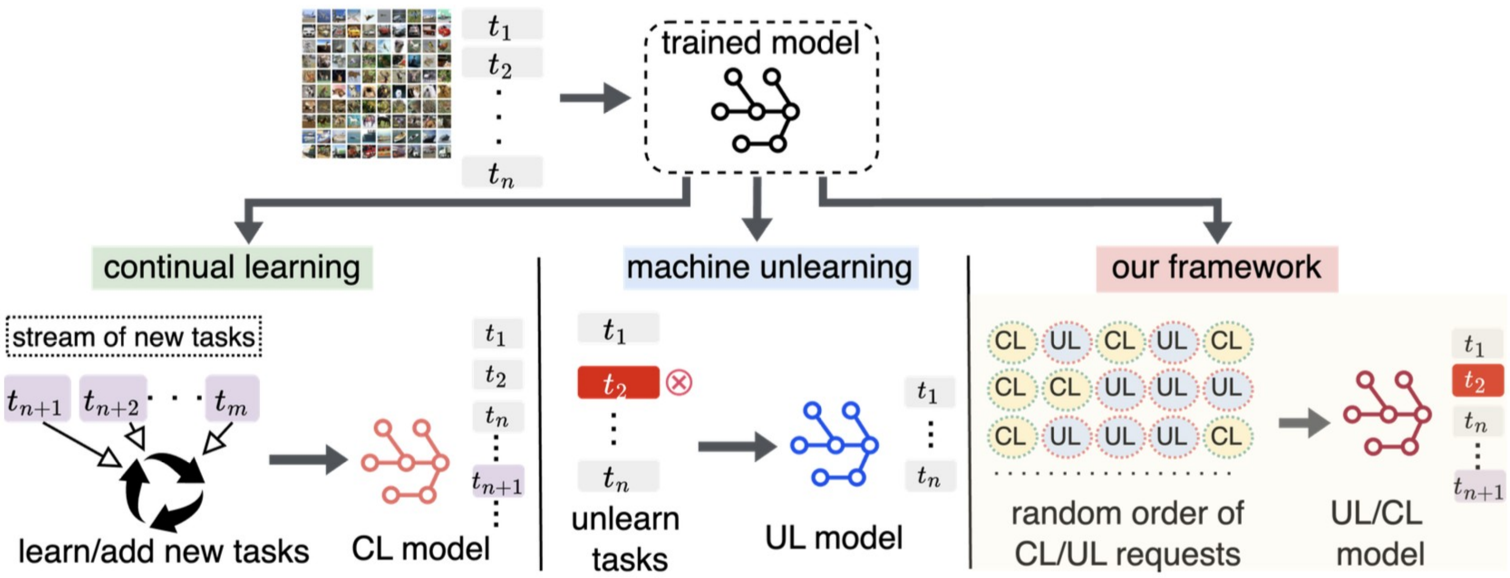
FIM, NTK Theory, Gradient Ascent, Error Maximizing Noise, Knowledge Distillation

Novelty of the Paper

Continual Learning, Machine Unlearning

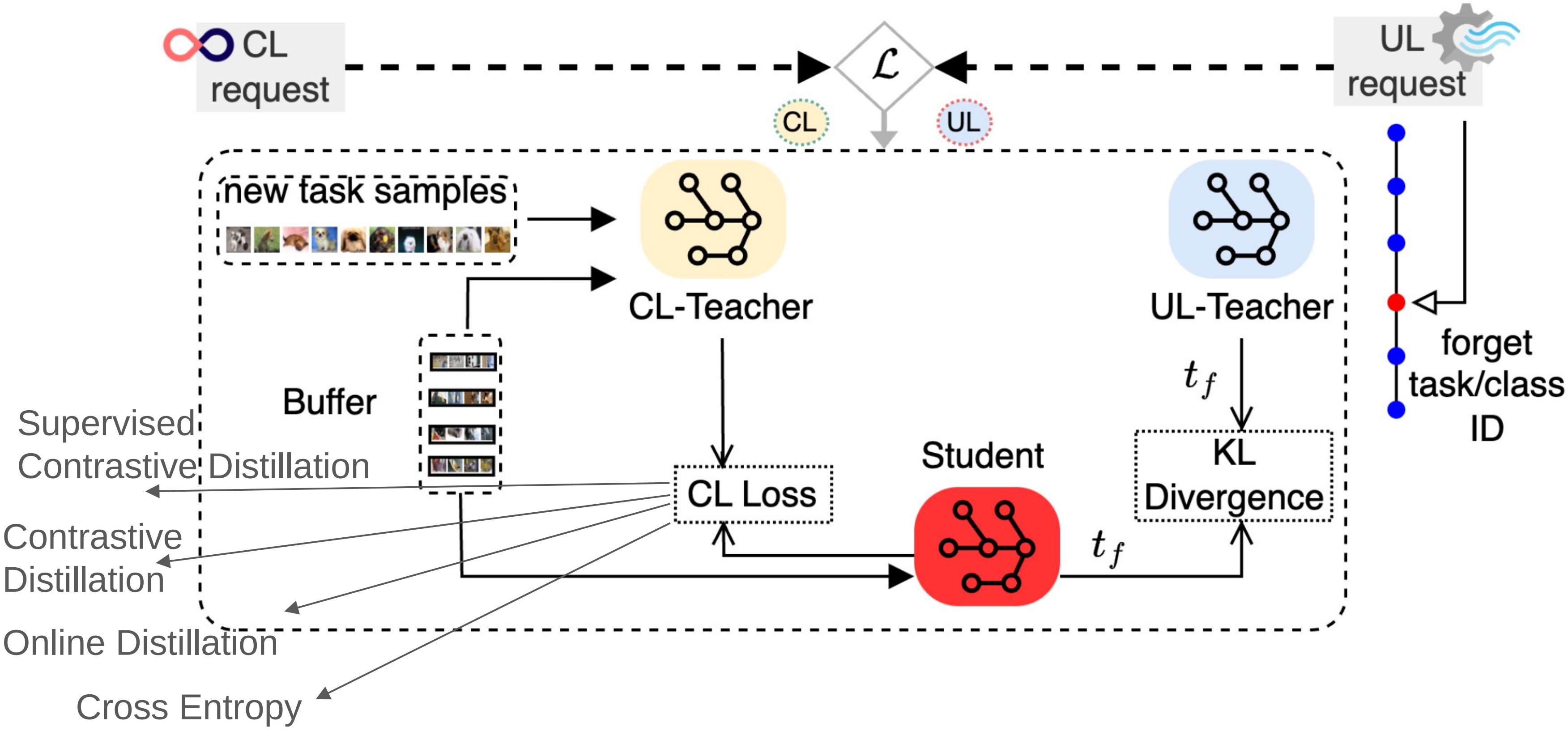
The first paper to integrate them in a single system.

Not a usual knowledge distillation.



t_1
t_2
t_n
trained model
continual learning
machine unlearning
our framework
stream of new tasks
t_{n+1}
t_{n+2}
t_m
learn/add new tasks
CL model
t_{n+1}
t_1
t_2
t_n
unlearn tasks
UL model
CL UL CL UL CL
CL CL UL UL UL
CL UL UL UL CL
random order of CL/UL requests
UL/CL model

Method in the Paper



Formulations

three main components:

- a feature extractor f_Θ for encoding feature representations
- a classifier f_Φ for mapping feature representations to output labels
- a projector f_Ψ for embedding features into a latent space where contrastive distillation is applied

$$\mathcal{B} = \{(x_i, y_i)_{i=1}^{|\mathcal{B}|}\} \text{ of size } |\mathcal{B}|$$

Continual Learning Loss (1/4)

Cross entropy loss for student training

$$\mathcal{L}_{ce} = \mathbb{E}_{(x,y)\sim\mathcal{D}_t\cup\mathcal{B}}\ell(f_{\Theta_s,\Phi_s}(x), y), \qquad (2)$$

where f_{Θ_s,Φ_s} represents the combined output of the feature extractor f_{Θ_s} and classifier f_{Φ_s} in the student model, $\mathcal{D}_t$ is the current task dataset, and $\mathcal{B}$ is the replay buffer.

Continual Learning Loss (2/4)

Online distillation loss for knowledge retention

The weighting factor ω(xi) dynamically scales the importance of each sample based on the teacher model's confidence in the sample's class label.

$$\omega(x_i) = \frac{\exp(f_{\Theta_T,\Phi_T}(x_i)_{y_i}/\rho)}{\sum_{c'=1}^{C} \exp(f_{\Theta_T,\Phi_T}(x_i)_{c'}/\rho)}, \qquad (3)$$

where f_{Θ_T,Φ_T} represents the teacher's combined feature extractor and classifier output, ρ is a temperature parameter that controls sharpness, and C is the total number of classes. The online distillation loss $\mathcal{L}_{od}$ is defined as:

$$\mathcal{L}_{od} = \mathbb{E}_{x_i \sim \mathcal{B}} \left[\omega(x_i) \| f_{\Theta_T,\Phi_T}(x_i) - f_{\Theta_s,\Phi_s}(x_i) \|_2^2 \right]. \qquad (4)$$

This loss term encourages alignment between the teacher and student predictions, thereby consolidating previous knowledge within the student model while learning new data.

Continual Learning Loss (3/4)

Contrastive distillation
for embedding alignment

Let $z_T = f_{\Theta_T,\Psi_T}(x)$ and $z_s = f_{\Theta_s,\Psi_s}(x)$ represent the embeddings produced by the teacher and student models, respectively, after combining the feature extractor f_Θ and projector f_Ψ. The contrastive distillation loss $\mathcal{L}_{cd}$ is defined as:

$$\mathcal{L}_{cd} = \sum_{z_j^T \sim z^{T+}} \log \frac{h(z_i^s, z_j^T)}{\sum_{z_k^T \sim z_T} h(z_i^s, z_k^T)}, \tag{5}$$

where z^{T+} denotes the set of teacher embeddings with the same label as z_i^s, and h is a critic function indicating joint distribution membership, defined as:

$$h(z_i, z_j) = \frac{\exp\left(\frac{(z_i/\|z_i\|_2)^\intercal (z_j/\|z_j\|_2)}{\tau}\right)}{\exp(1/\tau)}, \tag{6}$$

where τ is a temperature hyperparameter, and $(\cdot)^\intercal$ denotes the transpose operation.

Continual Learning Loss (4/4)

Supervised contrastive distillation for class similarity

$$\mathcal{L}_{scd} = -\mathbb{E}_{z_i^s \sim z^s} \sum_{z_j^s \sim z^{s+}} \log \frac{h(z_i^s, z_j^s)}{\sum_{z_k^s \sim z^s} h(z_i^s, z_k^s)}, \tag{7}$$

where z^{s+} represents the set of student embeddings with the same label as z_i^s. By minimizing $\mathcal{L}_{cd}$ and $\mathcal{L}_{scd}$ together, the student model effectively consolidates previously learned knowledge while acquiring new tasks.

Final Continual Learning Loss

$$\mathcal{L}_{cl} = \mathcal{L}_{ce} + \alpha_1 \mathcal{L}_{od} + \alpha_2 \mathcal{L}_{cd} + \alpha_3 \mathcal{L}_{scd}, \tag{8}$$

Momentum Update

$$\Theta_T \leftarrow m\Theta_T + (1-m)\left[(1-X)\Theta_T + X\Theta_s\right]$$

$$\Phi_T \leftarrow m\Phi_T + (1-m)\left[(1-X)\Phi_T + X\Phi_s\right]$$

$$\Psi_T \leftarrow m\Psi_T + (1-m)\left[(1-X)\Psi_T + X\Psi_s\right]$$

where, m is the momentum coefficient and X is a random variable with a Bernoulli distribution:

$$P(X=k) = p^k(1-p)^{1-k}, \quad k \in \{0,1\}$$

Unlearning Loss

$$\mathcal{L}_{cu} = (1 - \omega_u) \cdot \mathcal{KL}(f_{\Theta_T, \Phi_T}(x) \| f_{\Theta_s, \Phi_s}(x)) \\ + \omega_u \cdot \mathcal{KL}(f_{\Theta_b, \Phi_b}(x) \| f_{\Theta_s, \Phi_s}(x)), \tag{9}$$

where ω_u is a dynamically adjusted weight that prioritizes the original teacher for buffer samples and the bad teacher for unlearning samples. Here, $\mathcal{KL}(p\|q)$ denotes the KL-Divergence:

$$\mathcal{KL}(p\|q) = \sum_i p^{(i)} \log\left(\frac{p^{(i)}}{q^{(i)}}\right). \tag{10}$$

Unified Loss

$$\mathcal{L} = \gamma \cdot \mathcal{L}_{cl} + (1 - \gamma) \cdot \mathcal{L}_{cu}, \quad (11)$$

where γ is a context-sensitive weighting factor that adjusts based on task requirements, allowing seamless transitions

seamless??

Algorithm

Algorithm 1 CL-UL (Continual Learning and Unlearning)

Parameters:

- **Teacher parameters:** Θ_T, Ψ_T, Φ_T
- **Bad teacher parameters:** Θ_b, Ψ_b, Φ_b
- **Student parameters:** Θ_s, Ψ_s, Φ_s
- **Item label:** y
- **Hyperparameters:** α_{ul} (learning-unlearning weighting), $\alpha_1, \alpha_2, \alpha_3$ (loss coefficients), η (learning rate)

Initialization:

- Buffer $\mathcal{B} \leftarrow \{\}$ (empty buffer)
- Stream Data $D = \bigcup_{i=1}^{T} D_i$

for $t \in \{1, 2, \ldots, T\}$ **do**

Initialize the loss $\mathcal{L} \leftarrow 0$

Compute the primary task loss: $\mathcal{L}_{task} = \alpha_{ul} \cdot \text{cross\_entropy}\left(f_{\Theta_s,\Phi_s}(x), y\right)$

$$+ (1 - \alpha_{ul}) \cdot \text{KL\_div}\left(f_{\Theta_b,\Phi_b}(x), f_{\Theta_s,\Phi_s}(x)\right)$$

Sample from the buffer: $(X_B, Y_B) \leftarrow \text{Sample}(\mathcal{B})$

Calculate auxiliary losses:

- $\mathcal{L}_{od}$ (out-of-distribution loss) using Eq. (4)
- $\mathcal{L}_{cd}$ (class-discrimination loss) using Eq. (5)
- $\mathcal{L}_{scd}$ (sample-consistency discrimination loss) using Eq. (7)

Aggregate the losses:

$$\mathcal{L} \leftarrow \mathcal{L}_{task} + \alpha_1 \cdot \mathcal{L}_{od} + \alpha_2 \cdot \mathcal{L}_{cd} + \alpha_3 \cdot \mathcal{L}_{scd}$$

Update student parameters:

$$(\Theta_s, \Psi_s, \Phi_s) \leftarrow (\Theta_s, \Psi_s, \Phi_s) - \eta \cdot \frac{\partial \mathcal{L}}{\partial(\Theta_s, \Psi_s, \Phi_s)}$$

Update teacher parameters with random momentum:

$$(\Theta_T, \Psi_T, \Phi_T) \leftarrow \text{MomentumUpdate}((\Theta_T, \Psi_T, \Phi_T))$$

Optionally update buffer $\mathcal{B}$ with new samples.

end for

Example Workflow (1/5)

3-Learning, 1-Unlearning Task Sequence (Learn T1, Learn T2, Unlearn T1, Learn T3)

- **Student Model** with parameters (Θ_s, Ψ_s, Φ_s)
- **CL Teacher Model** with parameters (Θ_t, Ψ_t, Φ_t)
- **UL Teacher/Bad Teacher** with parameters (Θ_β, Ψ_β, Φ_β)
- **Replay Buffer B** (initially empty)

Example Workflow (2/5)

Task 1: Learn T1

1. **Initialization:**
   - Initialize student and CL teacher with identical random parameters
   - Set $\gamma = 1$ (full focus on continual learning objective)
   - Buffer B is empty
2. **Training Process:**
   - Load T1 dataset (e.g., CIFAR classes 0-1)
   - For each batch of data:
     - Calculate classification loss: $L\_ce = \text{CrossEntropy}(f\_\Theta_s,\Phi_s(x), y)$
     - Since buffer is empty, only L_ce contributes to the loss
     - Update student parameters using gradient descent: $\Theta_s \leftarrow \Theta_s - \eta \cdot \nabla L\_ce$
3. **Buffer Update:**
   - Sample data points from T1 using reservoir sampling
   - Add selected samples to buffer B (up to max capacity)
4. **Teacher Update:**
   - Update teacher parameters using momentum update:
     - $\Theta_t \leftarrow m \cdot \Theta_t + (1-m)[(1-X) \cdot \Theta_t + X \cdot \Theta_s]$
     - Where X is a Bernoulli random variable (p=0.2 for small buffer, p=0.8 for large buffer.

Example Workflow (3/5)

Task 2: Learn T2

1. **Configuration:**
   - Maintain teacher model from previous step
   - Keep $\gamma = 1$ (learning mode)
2. **Training Process:**
   - Load T2 dataset (e.g., CIFAR classes 2-3)
   - For each batch of data:
     - Calculate classification loss: L_ce
     - Sample data from buffer B (containing T1 samples)
     - Calculate distillation losses:
       - Online distillation: $L\_od = E[\omega(x_i) \| f\_\Theta_t,\Phi_t(x_i) - f\_\Theta_s,\Phi_s(x_i) \|^2]$
       - Contrastive distillation: L_cd (aligns teacher-student embeddings)
       - Supervised contrastive: L_scd (encourages intra-class similarity)
     - Aggregate losses: $L = L\_ce + \alpha_1 \cdot L\_od + \alpha_2 \cdot L\_cd + \alpha_3 \cdot L\_scd$
     - Update student parameters: $\Theta_s \leftarrow \Theta_s - \eta \cdot \nabla L$
3. **Buffer Update:**
   - Add samples from T2 to buffer B
   - If buffer is full, use reservoir sampling to maintain diverse representation
4. **Teacher Update:**
   - Apply contextualized momentum update as in step 1.4

Example Workflow (4/5)

Task 3: Unlearn T1

1. **Initialization:**
    - Initialize UL teacher/bad teacher without T1 knowledge
    - Set $\gamma = 0$ (full focus on unlearning objective)
2. **Unlearning Process:**
    - Identify all T1 samples (from classes 0-1)
    - For each batch:
        - Set ω_u dynamically:
            - $\omega_u = 1$ for T1 samples (to be forgotten)
            - $\omega_u = 0$ for T2 samples (to be retained)
        - Calculate unlearning loss:
            - $L\_cu = (1-\omega_u)\cdot KL(f\_\Theta_t,\Phi_t(x) \| f\_\Theta_s,\Phi_s(x)) + \omega_u\cdot KL(f\_\Theta_\beta,\Phi_\beta(x) \| f\_\Theta_s,\Phi_s(x))$
            - This makes the student follow the bad teacher for T1 data and original teacher for T2 data
        - Update student parameters: $\Theta_s \leftarrow \Theta_s - \eta\cdot\nabla L\_cu$
3. **Buffer Update:**
    - Remove all T1 samples from buffer B
    - Retain only T2 samples
4. **Teacher Update:**
    - Apply contextualized momentum update
    - Now teacher model has "forgotten" T1

Example Workflow (5/5)

Task 4: Learn T3

1. **Configuration:**
    - Use updated teacher model (with T2 knowledge but no T1 knowledge)
    - Set $\gamma = 1$ (back to learning mode)
2. **Training Process:**
    - Load T3 dataset (e.g., CIFAR classes 4-5)
    - For each batch:
        - Calculate classification loss: L_{ce}
        - Sample from buffer B (now containing only T2 samples)
        - Calculate distillation losses (L_{od}, L_{cd}, L_{scd})
        - Aggregate losses: $L = L_{ce} + \alpha_1 \cdot L_{od} + \alpha_2 \cdot L_{cd} + \alpha_3 \cdot L_{scd}$
        - Update student parameters: $\Theta_s \leftarrow \Theta_s - \eta \cdot \nabla L$
3. **Buffer Update:**
    - Add samples from T3 to buffer B
4. **Teacher Update:**
    - Apply contextualized momentum update.

Results (1/2)

CIFAR-10

ciFAIR-10

BS	Execution Sequence	Task 1		Task 2		Task 3		Task 4		Task 5	
		C0	C1	C2	C3	C4	C5	C6	C7	C8	C9
200	Learn T1	**99.4**	**99.3**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Learn T2	53.6	94.0	**96.0**	**95.0**	0.0	0.0	0.0	0.0	0.0	0.0
	Unlearn T2	97.8	98.5	**0.1**	**0.2**	0.0	0.0	0.0	0.0	0.0	0.0
	Learn T3	57.4	88.4	0.2	0.1	**98.2**	**94.9**	0.0	0.0	0.0	0.0
	Learn T4	59.6	88.1	0.1	0.1	39.0	43.2	**98.2**	**99.0**	0.0	0.0
	Learn T5	24.4	45.6	0.3	0.1	55.0	60.8	85.4	69.6	**98.7**	**99.0**
500	Learn T1	**99.4**	**99.1**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Learn T2	79.6	93.5	**94.6**	**96.4**	0.0	0.0	0.0	0.0	0.0	0.0
	Unlearn T2	99.4	97.3	**0.2**	**0.3**	0.0	0.0	0.0	0.0	0.0	0.0
	Learn T3	93.0	95.6	0.1	0.2	**97.4**	**96.3**	0.0	0.0	0.0	0.0
	Learn T4	81.0	91.7	0.1	0.1	43.3	54.1	**99.1**	**98.8**	0.0	0.0
	Learn T5	58.1	73.0	0.1	0.1	60.5	67.8	75.2	74.0	**98.7**	**99.1**
5120	Learn T1	**99.4**	**98.9**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Learn T2	95.2	99.1	**91.2**	**93.6**	0.0	0.0	0.0	0.0	0.0	0.0
	Unlearn T2	99.8	98.9	**0.1**	**0.1**	0.0	0.0	0.0	0.0	0.0	0.0
	Learn T3	98.4	97.9	0.2	0.1	**97.0**	**96.7**	0.0	0.0	0.0	0.0
	Learn T4	97.2	97.7	0.1	0.1	89.6	85.8	**95.9**	**96.7**	0.0	0.0
	Learn T5	91.4	94.0	0.1	0.1	89.2	88.2	95.2	91.1	**96.1**	**96.8**

Table 1. CL and single task UL in CIFAR-10 in a 2×5 task distribution setup. UL of Task 2 can be observed with accuracy dropping to $\sim 0.1\% - 0.3\%$ for the corresponding classes. Similarly CL accuracy gains in new Tasks are highlighted with **bold**.

Results (2/2)

BS	Execution	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
	Sequence	C-0	C-1	C-2	C-3	C-4	C-5	C-6	C-7	C-8	C-9
200	Learn first 5 tasks	**37.2**	**26.8**	**17.1**	**10.9**	**99.8**	0.0	0.0	0.0	0.0	0.0
	Unlearn T2	57.8	**0.1**	42.4	69.1	91.6	0.0	0.0	0.0	0.0	0.0
	Unlearn T4	59.0	0.2	85.1	**6.3**	82.2	0.0	0.0	0.0	0.0	0.0
	Unlearn T5	57.2	0.1	94.2	0.7	**34.5**	0.0	0.0	0.0	0.0	0.0
	Learn T6	35.8	0.2	44.7	0.3	0.2	**99.3**	0.0	0.0	0.0	0.0
	Learn T7	49.0	0.2	23.8	0.7	0.2	9.4	**99.7**	0.0	0.0	0.0
	Learn T8	41.8	0.1	30.2	0.3	0.2	6.2	47.4	**99.6**	0.0	0.0
	Learn T9	9.6	0.2	33.7	0.1	0.1	25.2	57.4	54.0	**99.7**	0.0
	Learn T10	15.2	0.2	35.0	0.4	0.1	25.4	67.7	46.1	59.0	**99.2**
500	Learn first 5 tasks	**66.2**	**46.0**	**56.0**	**45.4**	**98.2**	0.0	0.0	0.0	0.0	0.0
	Unlearn T2	81.6	**16.3**	74.0	58.1	95.3	0.0	0.0	0.0	0.0	0.0
	Unlearn T4	83.2	7.0	92.1	**3.9**	86.9	0.0	0.0	0.0	0.0	0.0
	Unlearn T5	88.4	3.2	94.2	0.9	**62.6**	0.0	0.0	0.0	0.0	0.0
	Learn T6	78.6	2.7	70.0	0.9	3.5	**99.3**	0.0	0.0	0.0	0.0
	Learn T7	84.2	2.4	58.8	0.1	3.2	60.0	**98.9**	0.0	0.0	0.0
	Learn T8	75.6	0.6	61.9	0.7	3.1	38.9	79.2	**98.6**	0.0	0.0
	Learn T9	43.4	0.2	52.5	0.2	3.0	59.1	75.0	80.6	**99.1**	0.0
	Learn T10	46.2	0.5	50.6	0.2	3.0	49.4	67.7	71.0	54.7	**100**
5120	Learn first 5 tasks	**95.2**	**99.5**	**80.1**	**74.1**	**92.3**	0.0	0.0	0.0	0.0	0.0
	Unlearn T2	95.8	**27.4**	85.1	92.0	94.9	0.0	0.0	0.0	0.0	0.0
	Unlearn T4	96.0	3.7	94.4	**34.6**	95.3	0.0	0.0	0.0	0.0	0.0
	Unlearn T5	97.2	3.7	96.5	47.2	**71.0**	0.0	0.0	0.0	0.0	0.0
	Learn T6	85.6	0.7	65.3	0.6	0.7	**99.7**	0.0	0.0	0.0	0.0
	Learn T7	88.4	0.1	30.8	0.4	2.3	66.0	**99.1**	0.0	0.0	0.0
	Learn T8	88.4	0.1	80.9	0.2	0.2	69.0	92.3	**98.0**	0.0	0.0
	Learn T9	81.2	0.1	82.2	0.2	0.2	85.6	93.5	94.2	**97.9**	0.0
	Learn T10	78.2	0.1	69.9	0.2	0.2	83.0	92.7	91.5	86.2	**99.4**

Table 3. CL and multiple task UL in CIFAR-10 in a 1×10 task distribution setup. UL of Task 2, Task 4, and Task 5 can be observed with accuracy drop to $\sim 0.1\% - 70.0\%$ for the corresponding classes. Similarly CL accuracy gains in new Tasks are highlighted with **bold**.

Effect of Buffer Size

A larger buffer size of 5120 results in better retention, with a noticeable improvement across all tasks, especially in the retention of Task 1.

“…as the buffer size increases, the system becomes better at continual learning but worse at unlearning. This is because a larger buffer helps retain knowledge from earlier tasks but also makes it harder to completely remove information related to classes that should be forgotten.”

Critics

AI-generated texts

Main algorithm at the appendix.

Critical mistakes in formulations (eq. 5)

ciFAIR-10?

No baseline (only retrain)

Not data private or memory-light

fΘ fΦ fΨ might be unnecessary