Class-Level Unlearning

Data Mining Quality Analytics

안시후

2026.07.31

발표자 소개

❖ **안시후 (Sihu Ahn)**

- Data Mining & Quality Analytics Lab (김성범 교수님)
- 석박사통합과정 (2021.3 ~)

❖ **관심 연구 분야**

- Computer Vision (Action Recognition)
- Continual Learning/Unlearning
- Self-Supervised Learning

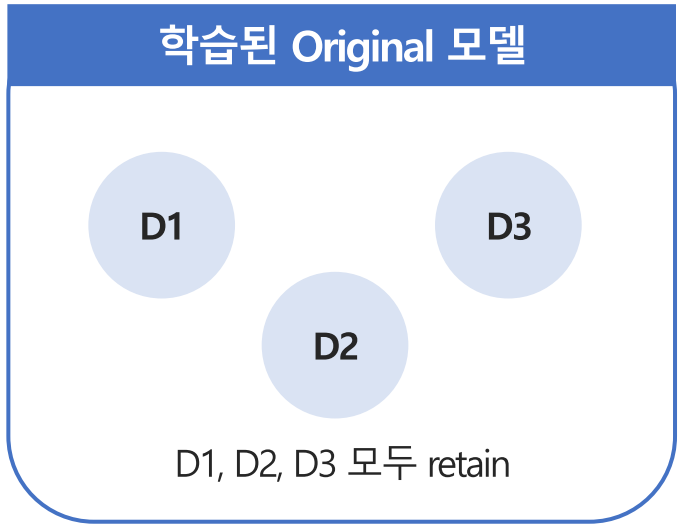
❖ **E-mail**

- sihuahn@korea.ac.kr

Introduction to Class-Level Unlearning

❖ Machine Unlearning은 무엇인가?

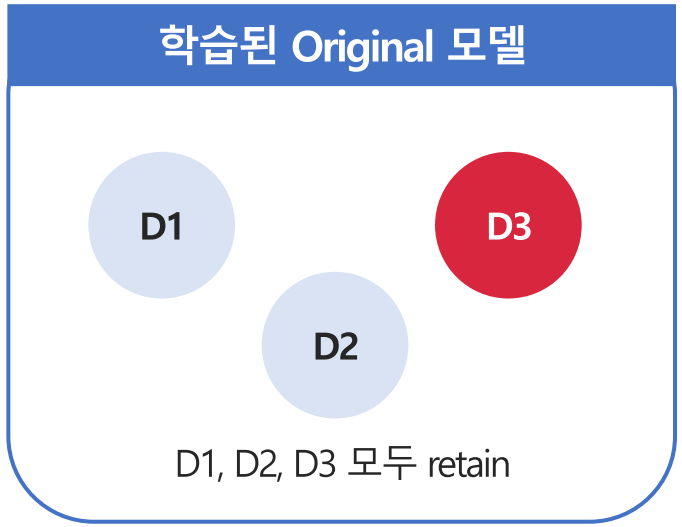
- 학습된 모델에서 특정 데이터의 영향을 제거하는 기술
  - ➢ 개인정보 보호 규제 강화 : 사용자가 자신의 데이터를 모델에서 삭제 요청할 법적 근거가 생김
  - ➢ 배포된 모델의 사후 수정 요구 증가 : 한번 배포된 모델도 문제가 생기면 고쳐야 하는 시대
  - ➢ 특정 데이터의 영향이 가중치 전체에 분산되어 있어 "이 데이터만 지운다"는 게 명확하기 어려움
  - ➢ "처음부터 그 데이터 없이 학습했다면 나왔을 모델"을 근사하는 문제로 재정의됨



Introduction to Class-Level Unlearning

❖ Machine Unlearning은 무엇인가?

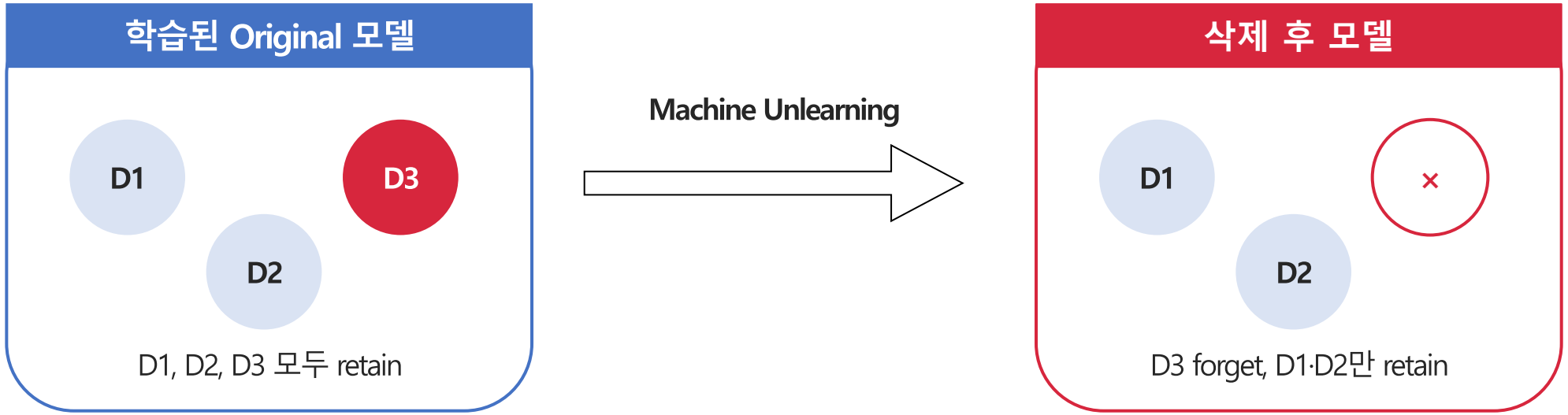
- 학습된 모델에서 특정 데이터의 영향을 제거하는 기술
  - ➢ 개인정보 보호 규제 강화 : 사용자가 자신의 데이터를 모델에서 삭제 요청할 법적 근거가 생김
  - ➢ 배포된 모델의 사후 수정 요구 증가 : 한번 배포된 모델도 문제가 생기면 고쳐야 하는 시대
  - ➢ 특정 데이터의 영향이 가중치 전체에 분산되어 있어 "이 데이터만 지운다"는 게 명확하기 어려움
  - ➢ "처음부터 그 데이터 없이 학습했다면 나왔을 모델"을 근사하는 문제로 재정의됨



Introduction to Class-Level Unlearning

❖ Machine Unlearning은 무엇인가?

- 학습된 모델에서 특정 데이터의 영향을 제거하는 기술
  - ➢ 개인정보 보호 규제 강화 : 사용자가 자신의 데이터를 모델에서 삭제 요청할 법적 근거가 생김
  - ➢ 배포된 모델의 사후 수정 요구 증가 : 한번 배포된 모델도 문제가 생기면 고쳐야 하는 시대
  - ➢ 특정 데이터의 영향이 가중치 전체에 분산되어 있어 "이 데이터만 지운다"는 게 명확하기 어려움
  - ➢ "처음부터 그 데이터 없이 학습했다면 나왔을 모델"을 근사하는 문제로 재정의됨



Introduction to Class-Level Unlearning

❖ Machine Unlearning은 무엇인가?

- 학습된 모델에서 특정 데이터의 영향을 제거하는 기술
  - ➢ 개인정보 보호 규제 강화 : 사용자가 자신의 데이터를 모델에서 삭제 요청할 법적 근거가 생김
  - ➢ 배포된 모델의 사후 수정 요구 증가 : 한번 배포된 모델도 문제가 생기면 고쳐야 하는 시대
  - ➢ 특정 데이터의 영향이 가중치 전체에 분산되어 있어 "이 데이터만 지운다"는 게 명확하기 어려움
  - ➢ "처음부터 그 데이터 없이 학습했다면 나왔을 모델"을 근사하는 문제로 재정의됨

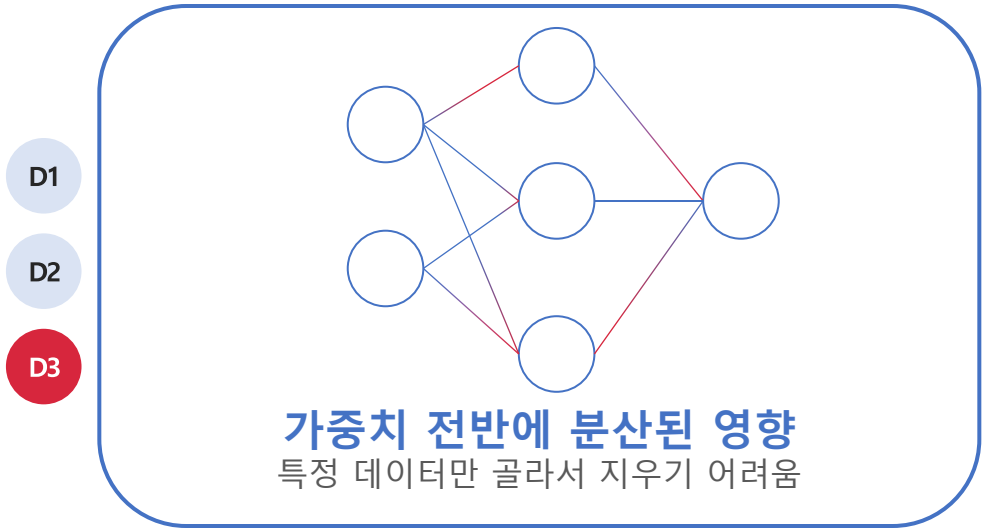
개인정보 보호 규제 강화
사용자의 데이터 삭제 요청에 대한 법적 근거 발생

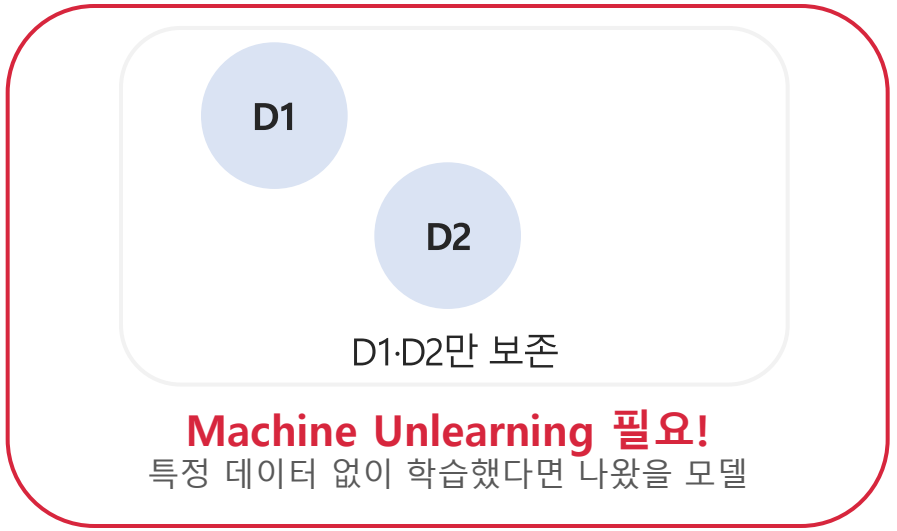
배포 후 사후 수정 요구 증가
이미 배포된 모델도 문제 발생 시 재수정 필요

Introduction to Class-Level Unlearning

❖ Machine Unlearning은 무엇인가?

- 학습된 모델에서 특정 데이터의 영향을 제거하는 기술
  - ➢ 개인정보 보호 규제 강화 : 사용자가 자신의 데이터를 모델에서 삭제 요청할 법적 근거가 생김
  - ➢ 배포된 모델의 사후 수정 요구 증가 : 한번 배포된 모델도 문제가 생기면 고쳐야 하는 시대
  - ➢ 특정 데이터의 영향이 가중치 전체에 분산되어 있어 "이 데이터만 지운다"는 게 명확하기 어려움
  - ➢ "처음부터 그 데이터 없이 학습했다면 나왔을 모델"을 근사하는 문제로 재정의됨

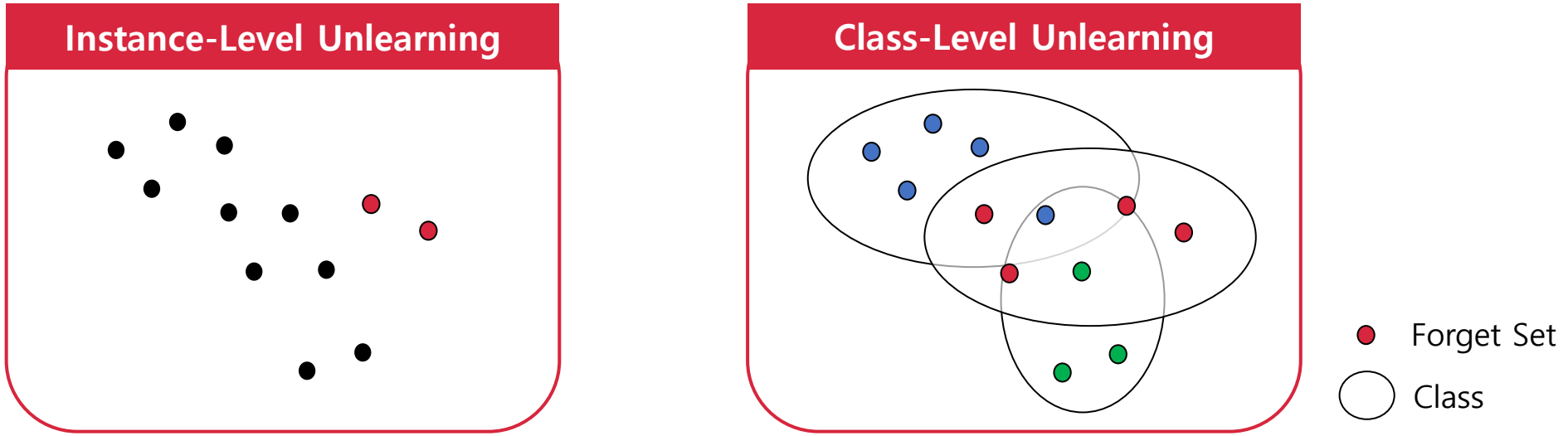




Introduction to Class-Level Unlearning

❖ 왜 Class-Level Unlearning이 필요할까?

- 클래스 전체 삭제는 모델 관점에서 큰 파급력을 가지고 있음
  - ➢ Instance-level vs Class-level 대비
  - ➢ Class-level이 왜 중요한가?



Introduction to Class-Level Unlearning

❖ 왜 Class-Level Unlearning이 필요할까?

- 클래스 전체 삭제는 모델 관점에서 큰 파급력을 가지고 있음
  - ➢ Instance-level vs Class-level 대비
  - ➢ Class-level이 왜 중요한가?

저작권·브랜드 침해

특정 브랜드·로고가 담긴
이미지 클래스 전체 삭제

클래스 오염

특정 오염이 포함된
클래스 전체 삭제

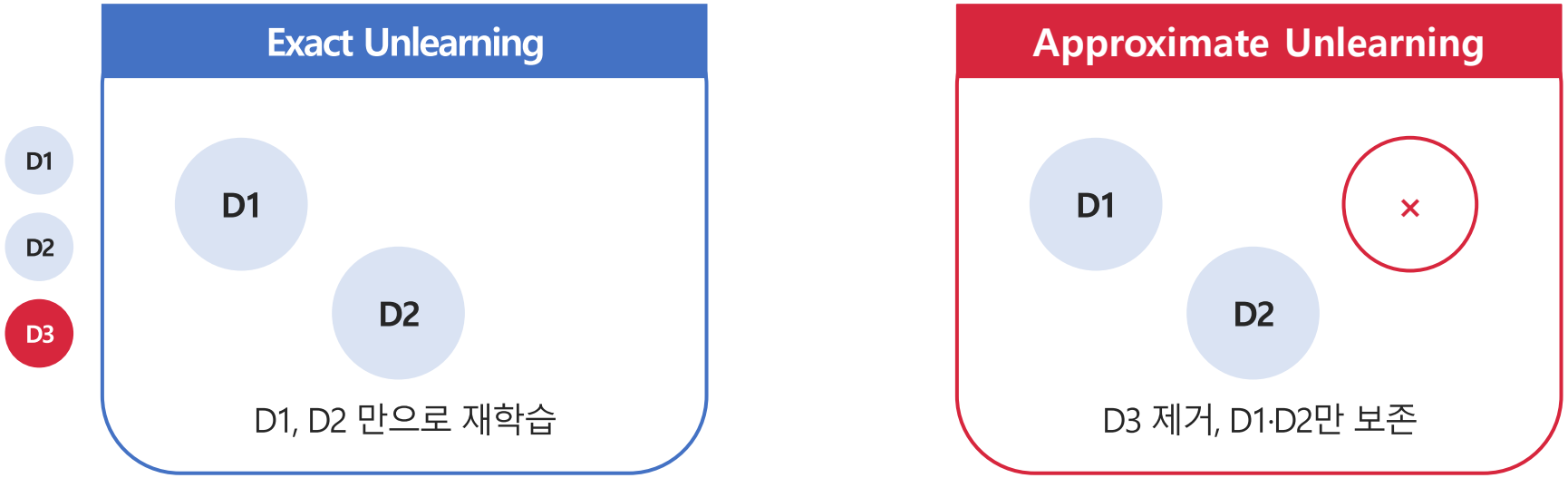
특정 카테고리 단종

생산이나 계약이 종료된
카테고리를 통째로 제거

Introduction to Class-Level Unlearning

❖ Exact vs Approximate Unlearning

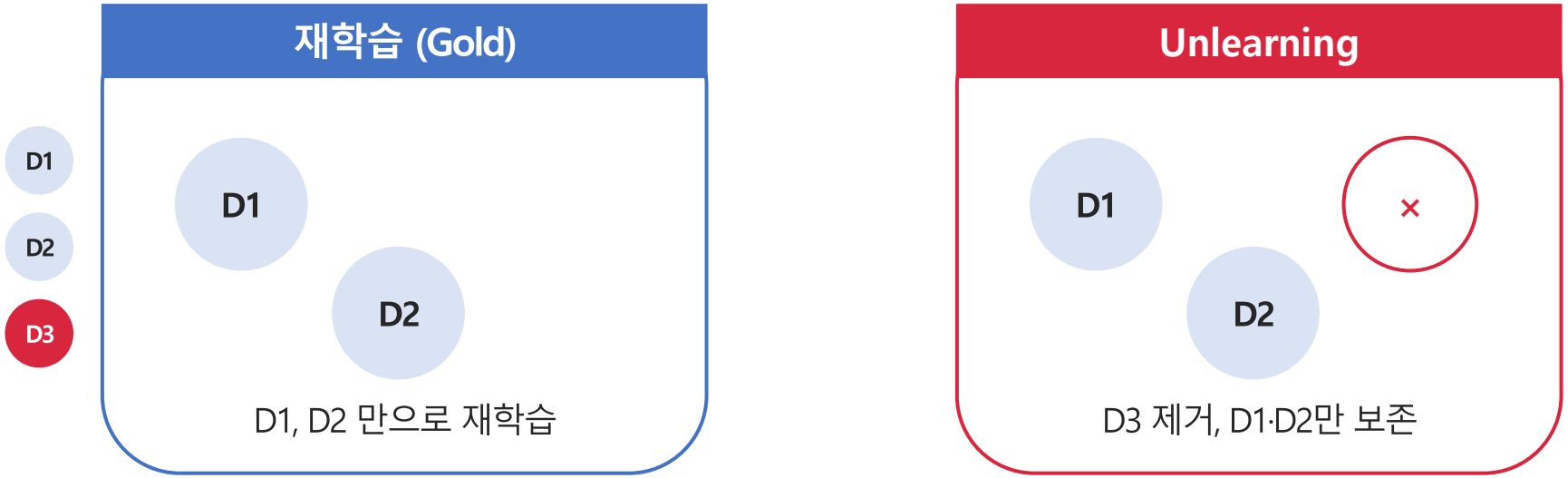
- 처음부터 다시 학습하는 Exact는 확실하지만 비용 ↑, 실용적 대안으로 Approximate 연구 활발
  - ➢ Exact Unlearning : forget 클래스(데이터)를 아예 보지 않고 처음부터 재학습한 모델
    (가중치/출력 분포가 완전히 일치함을 수학적/통계적으로 보장하는 기법)
  - ➢ Approximate Unlearning : Exact가 아닌 모든 기법을 총칭하는 넓은 범주



Introduction to Class-Level Unlearning

❖ Exact vs Approximate Unlearning

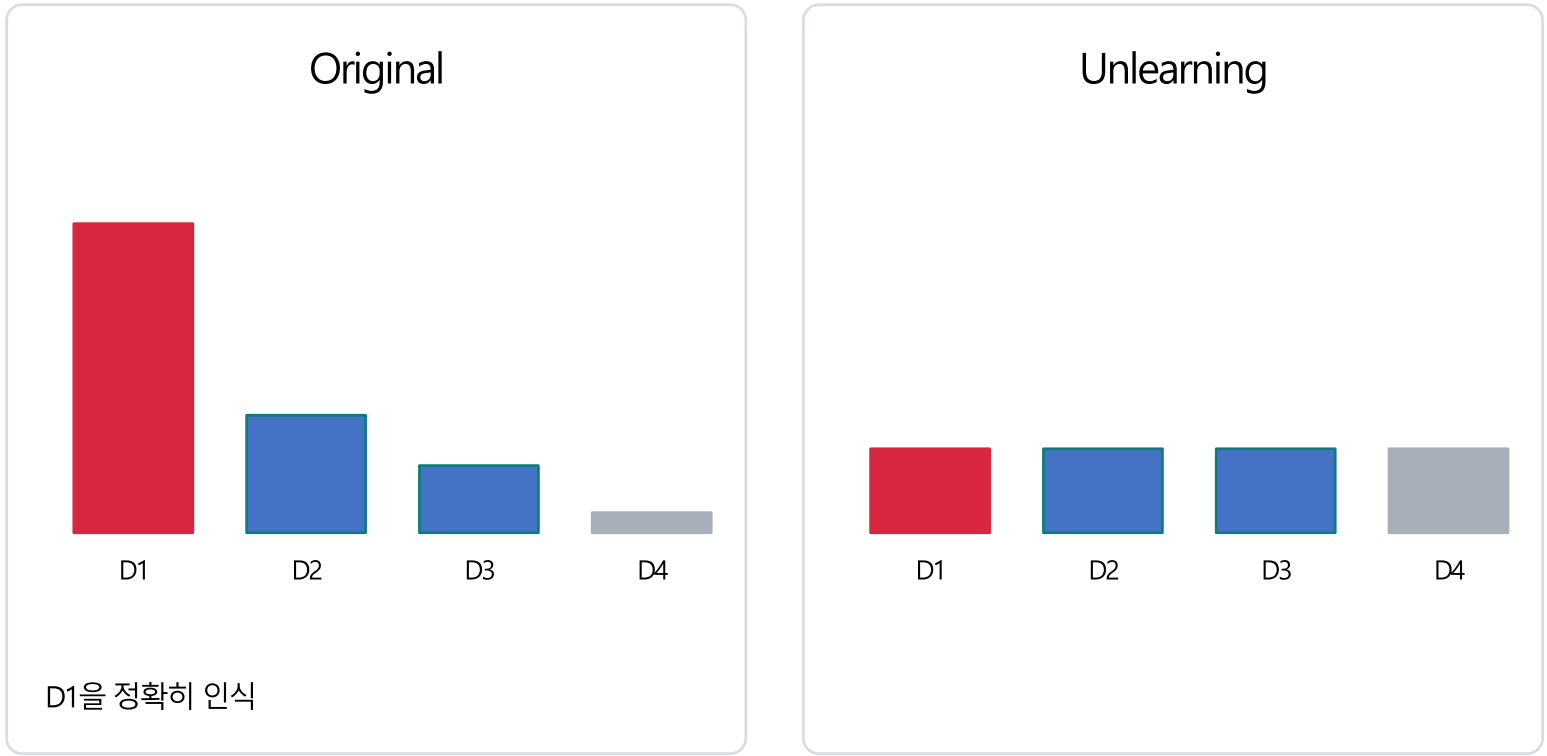
- 처음부터 다시 학습하는 Exact는 확실하지만 비용 ↑, 실용적 대안으로 Approximate 연구 활발
  - ➢ Exact Unlearning : forget 클래스(데이터)를 아예 보지 않고 처음부터 재학습한 모델
    (가중치/출력 분포가 완전히 일치함을 수학적/통계적으로 보장하는 기법)
  - ➢ Approximate **Unlearning** : Exact가 아닌 모든 기법을 총칭하는 넓은 범주



Introduction to Class-Level Unlearning

❖ Goal of Class-Level Unlearning

- 재학습 모델과 구분되지 않도록 재현하는 것
  - ➢ Original은 forget class인 D1 정보를 가지고 있고 이를 없애는 과정이 Unlearning

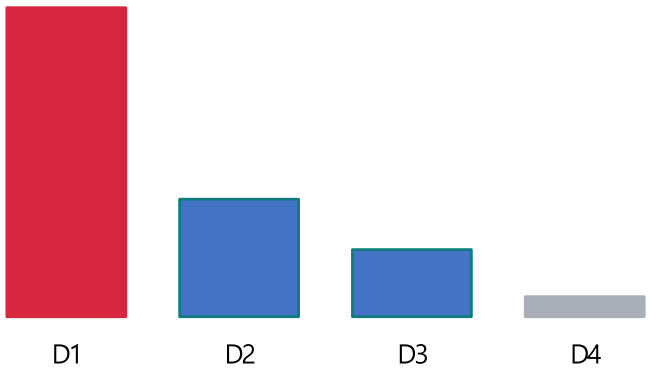


Introduction to Class-Level Unlearning

❖ Goal of Class-Level Unlearning

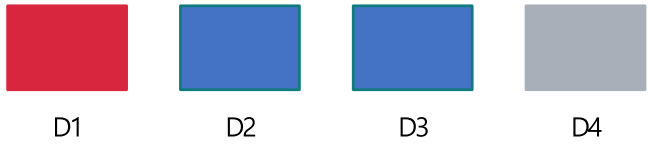
- 재학습 모델과 구분되지 않도록 재현하는 것
  - ➢ Original은 forget class인 D1 정보를 가지고 있고 이를 없애는 과정이 Unlearning
  - ➢ 지운 티가 남아 문제가 될 수 있기 때문에 재학습 모델을 목표로 학습

Original



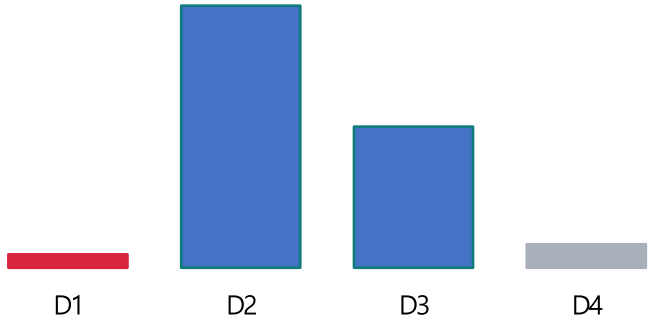
D1을 정확히 인식

Unlearning



출력이 균등분포 — 오히려 '지운 티'가 남는다

재학습 (Gold)

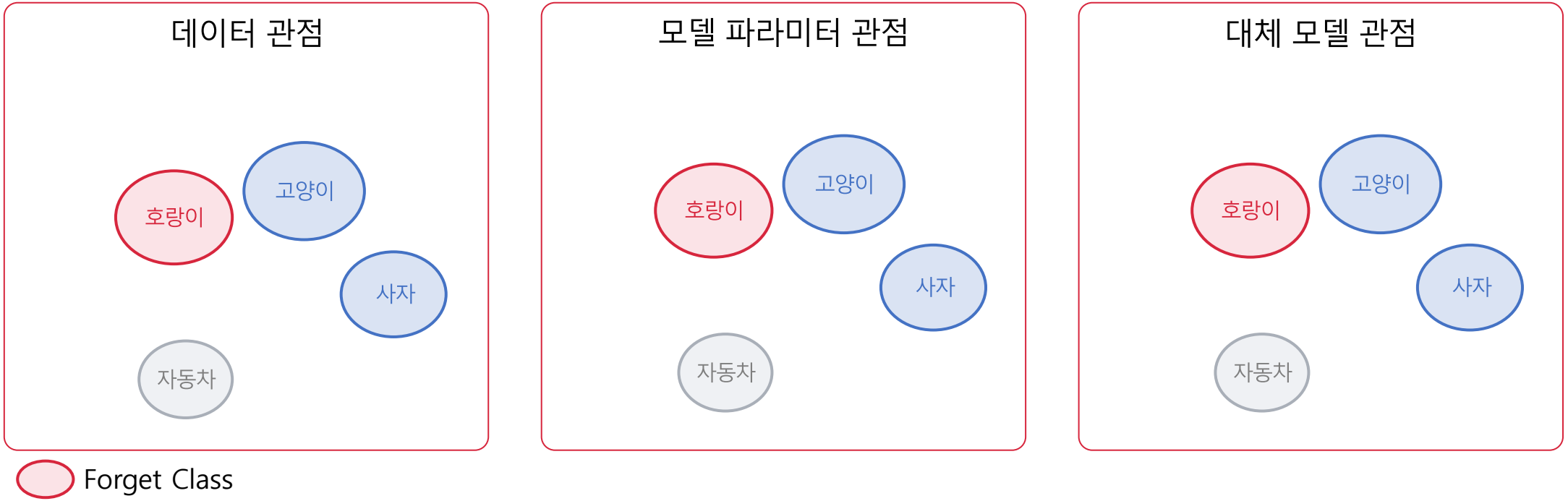


이웃 클래스로 자연스럽게 흡수

Class-Level Unlearning Methods

❖ Unlearning 방법론의 종류

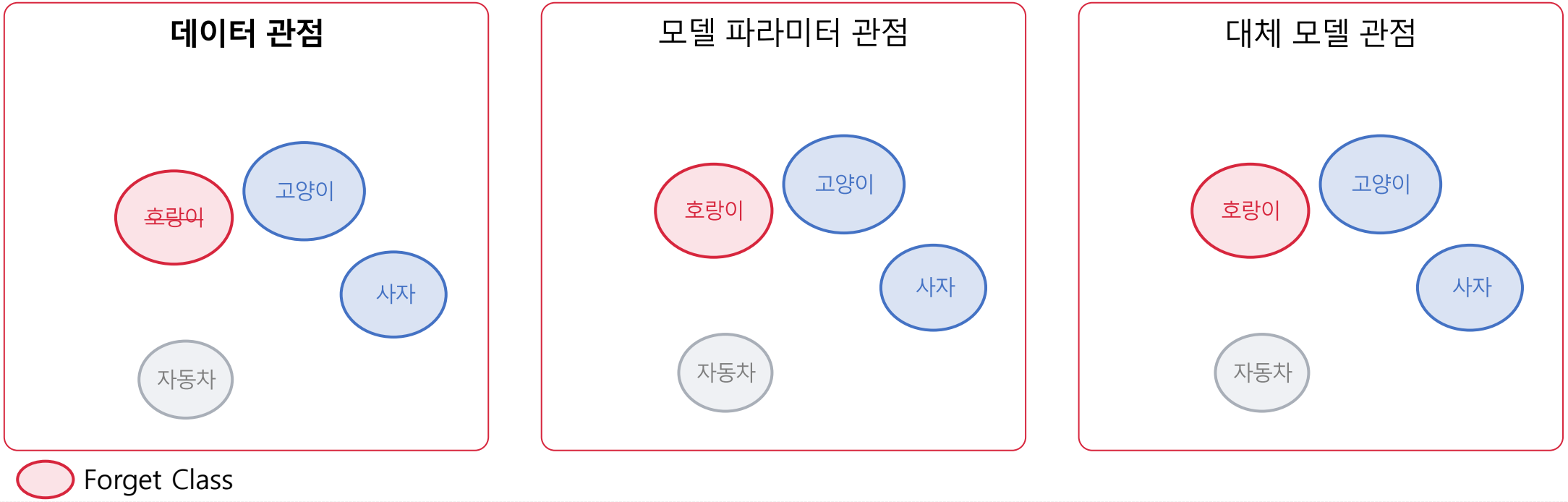
- Data Reorganization (random labeling / SISA)
- Model Manipulation (fine-tuning / gradient ascent / boundary unlearning)
- Auxiliary Model (SCRUB)



Class-Level Unlearning Methods

❖ Unlearning 방법론의 종류

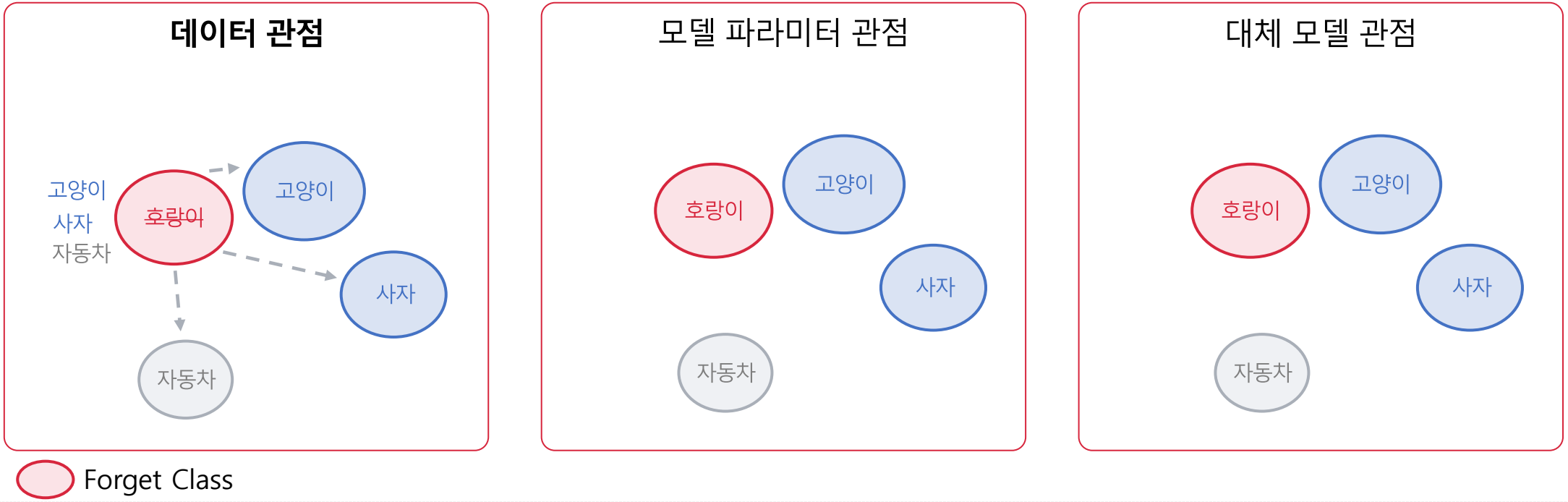
- **Data Reorganization** (random labeling / SISA)
- Model Manipulation (fine-tuning / gradient ascent / boundary unlearning)
- Auxiliary Model (SCRUB)



Class-Level Unlearning Methods

❖ Unlearning 방법론의 종류

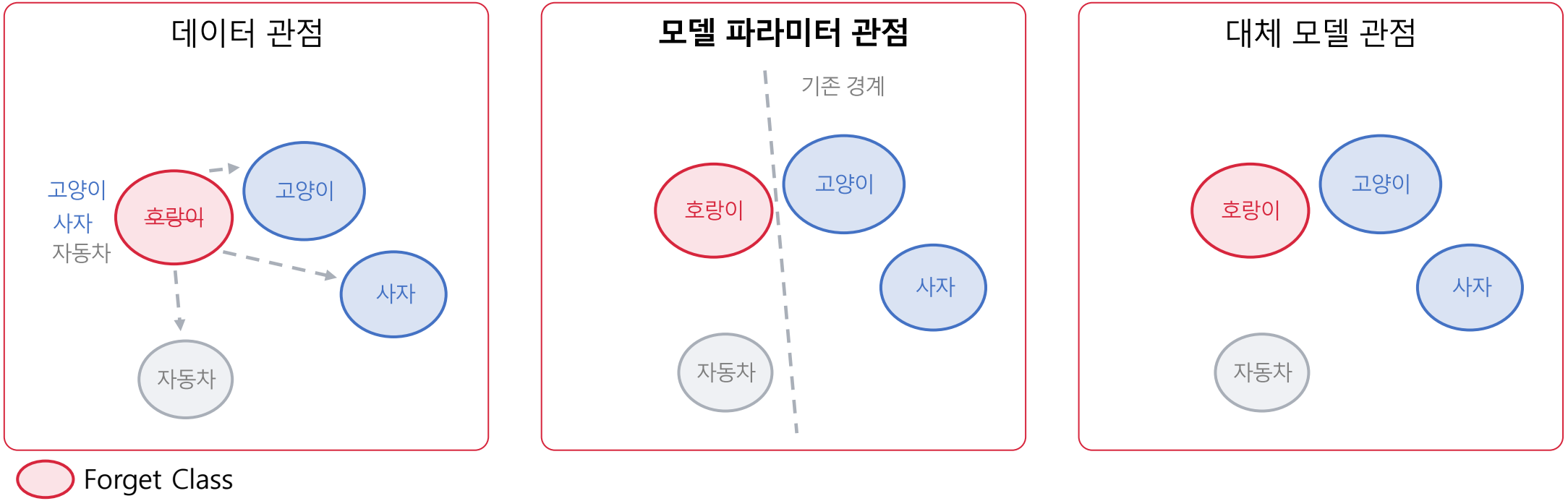
- **Data Reorganization** (random labeling / SISA)
- Model Manipulation (fine-tuning / gradient ascent / boundary unlearning)
- Auxiliary Model (SCRUB)



Class-Level Unlearning Methods

❖ Unlearning 방법론의 종류

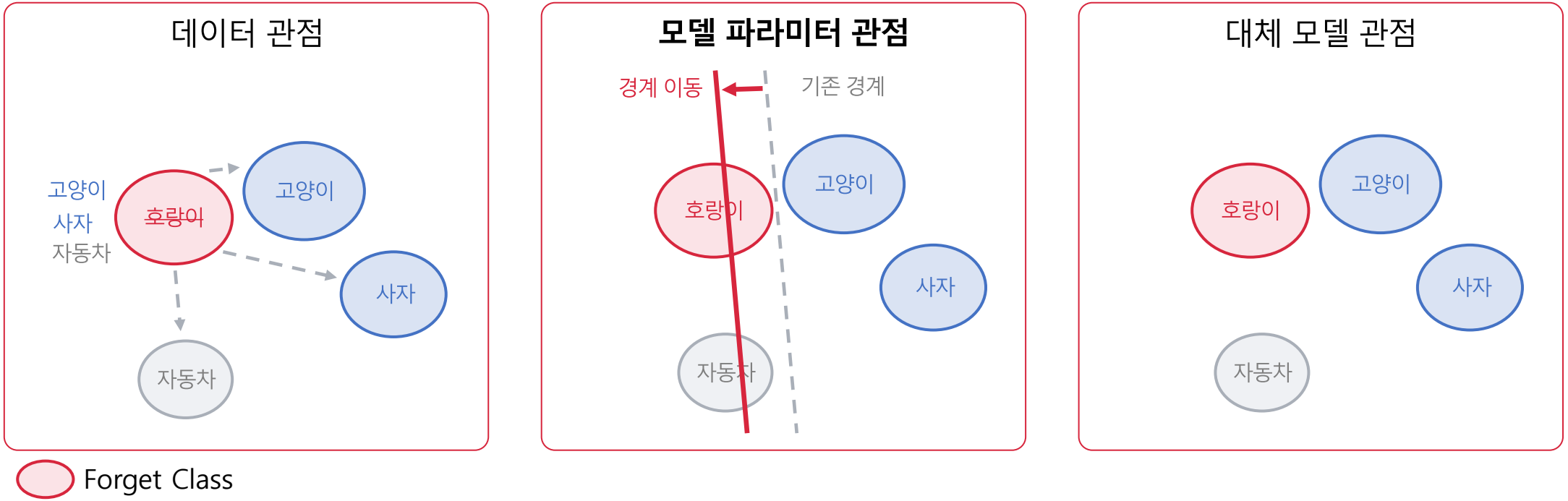
- Data Reorganization (random labeling / SISA)
- **Model Manipulation** (fine-tuning / gradient ascent / boundary unlearning)
- Auxiliary Model (SCRUB)



Class-Level Unlearning Methods

❖ Unlearning 방법론의 종류

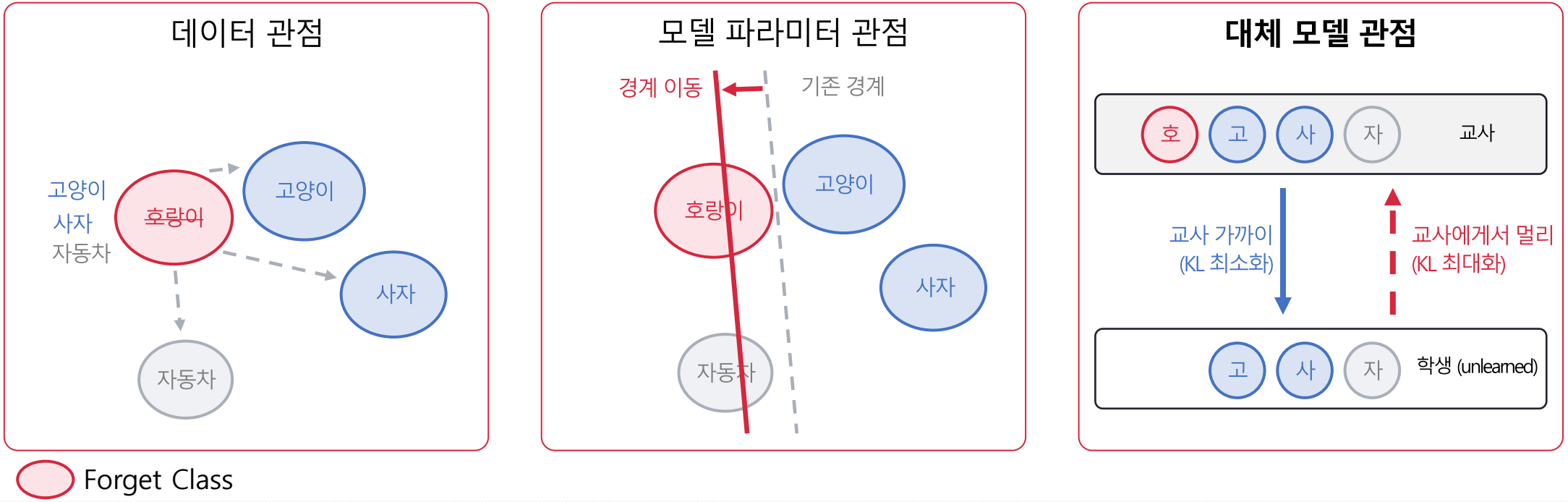
- Data Reorganization (random labeling / SISA)
- **Model Manipulation** (fine-tuning / gradient ascent / boundary unlearning)
- Auxiliary Model (SCRUB)



Class-Level Unlearning Methods

❖ Unlearning 방법론의 종류

- Data Reorganization (random labeling / SISA)
- Model Manipulation (fine-tuning / gradient ascent / boundary unlearning)
- **Auxiliary Model** (SCRUB)

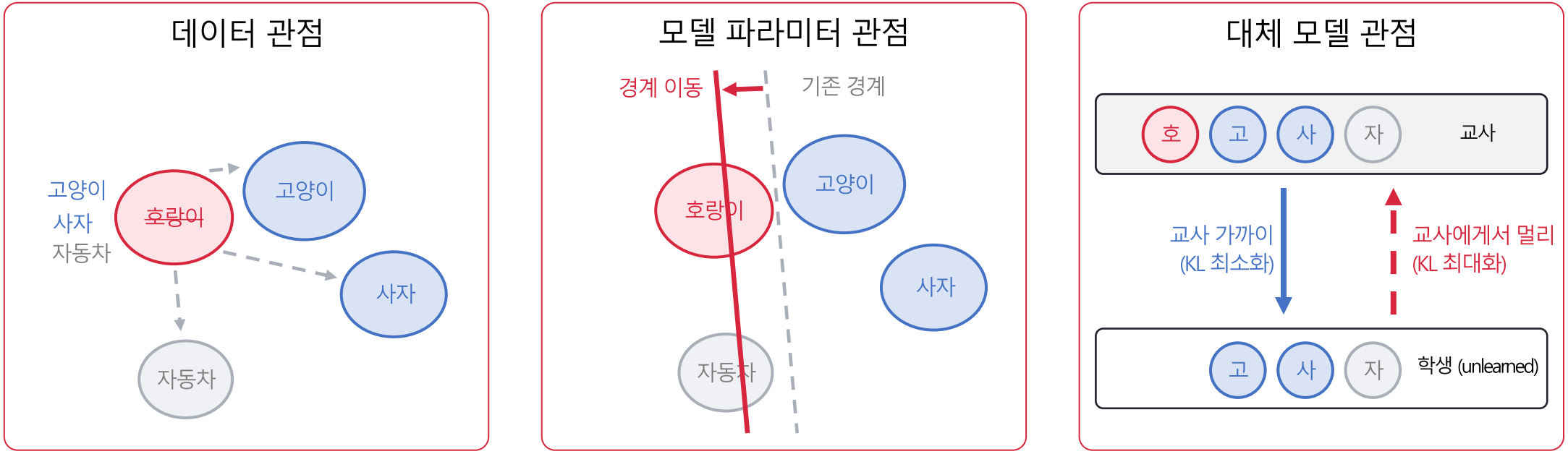


Class-Level Unlearning Methods

❖ Unlearning 방법론의 종류

- Data Reorganization (random labeling / SISA)
- Model Manipulation (fine-tuning / gradient ascent / boundary unlearning)

실제로는 한 가지 방법론만 쓰기보다, 여러 범주를 함께 결합해 사용하는 경우가 많음
(예: Data Reorganization + Model Manipulation 병행)



Entanglement: The Hidden Challenge in Class-Level Unlearning

❖ **최근 연구는 클래스 간 유사성, 즉 얽힘(Entanglement)을 고려하는 방향으로 진행**

- 두 단계 모델 최적화로 접근한 Cheng et al.과, MLLM을 활용한 MeGU 소개

Model Manipulation ICLR 2026

MACHINE UNLEARNING UNDER RETAIN–FORGET ENTANGLEMENT

Jingpu Cheng
Department of Mathematics
National University of Singapore
chengjingpu@u.nus.edu

Ping Liu
Department of Computer Science
University of Nevada Reno
pino.pingliu@gmail.com

Qianxiao Li
Department of Mathematics
Institute for Functional Intelligent Materials
National University of Singapore
qianxiao@nus.edu.sg

Chi Zhang*
Department of Mathematics
National University of Singapore
czhang24@nus.edu.sg

ABSTRACT

Forgetting a subset in machine unlearning is rarely an isolated task. Often, retained samples that are closely related to the forget set can be unintentionally affected, particularly when they share correlated features from pretraining or exhibit strong semantic similarities. To address this challenge, we propose a novel two-phase optimization framework specifically designed to handle such retain–forget entanglements. In the first phase, an augmented Lagrangian method increases the loss on the forget set while preserving accuracy on less-related retained samples. The second phase applies a gradient projection step, regularized by the Wasserstein-2 distance, to mitigate performance degradation on semantically related retained samples without compromising the unlearning objective. We validate our approach through comprehensive experiments on multiple unlearning tasks, standard benchmark datasets, and diverse neural architectures, demonstrating that it achieves effective and reliable unlearning while outperforming existing baselines in both accuracy retention and removal fidelity. Our code is available here.

Data Reorganization, Model Manipulation, Auxiliary Model arXiv 2026

MEGU: MACHINE-GUIDED UNLEARNING WITH TARGET FEATURE DISENTANGLEMENT

A PREPRINT

Haoyu Wang[1], **Zhuo Huang**[2], **Xiaolong Wang**[1], **Bo Han**[3], **Zhiwei Lin**[1], **Tongliang Liu**[2]
[1]School of Automation, Beijing Institute of Technology; [2]Sydney AI Centre, The University of Sydney;
[3]Department of Computer Science, Hong Kong Baptist University

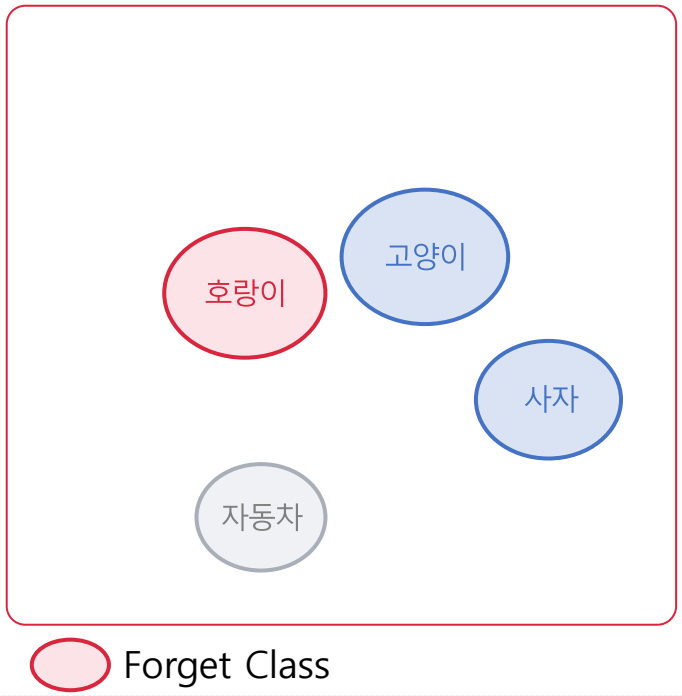
ABSTRACT

The growing concern over training data privacy has elevated the "Right to be Forgotten" into a critical requirement, thereby raising the demand for effective Machine Unlearning. However, existing unlearning approaches commonly suffer from a fundamental trade-off: aggressively erasing the influence of target data often degrades model utility on retained data, while conservative strategies leave residual target information intact. In this work, the intrinsic representation properties learned during model pretraining are analyzed. It is demonstrated that semantic class concepts are entangled at the feature-pattern level, sharing associated features while preserving concept-specific discriminative components. This entanglement fundamentally limits the effectiveness of existing unlearning paradigms. Motivated by this insight, we propose Machine-Guided Unlearning (MeGU), a novel framework that guides unlearning through concept-aware re-alignment. Specifically, Multi-modal Large Language Models (MLLMs) are leveraged to explicitly determine re-alignment directions for target samples by assigning semantically meaningful perturbing labels. To improve efficiency, inter-class conceptual similarities estimated by the MLLM are encoded into a lightweight transition matrix. Furthermore, MeGU introduces a positive-negative feature noise pair to explicitly disentangle target concept influence. During finetuning, the negative noise suppresses target-specific feature patterns, while the positive noise reinforces remaining associated features and aligns them with perturbing concepts. This coordinated design enables selective disruption of target-specific representations while preserving shared semantic structures. As a result, MeGU enables controlled and selective forgetting, effectively mitigating both under-unlearning and over-unlearning. Extensive experiments across varying unlearning scenarios and diverse datasets demonstrate that MeGU consistently outperforms state-of-the-art baselines, achieving thorough target data removal while maintaining strong generalization performance on retained data.

Entanglement: The Hidden Challenge in Class-Level Unlearning

❖ **최근 연구는 클래스 간 유사성, 즉 얽힘(Entanglement)을 고려하는 방향으로 진행**

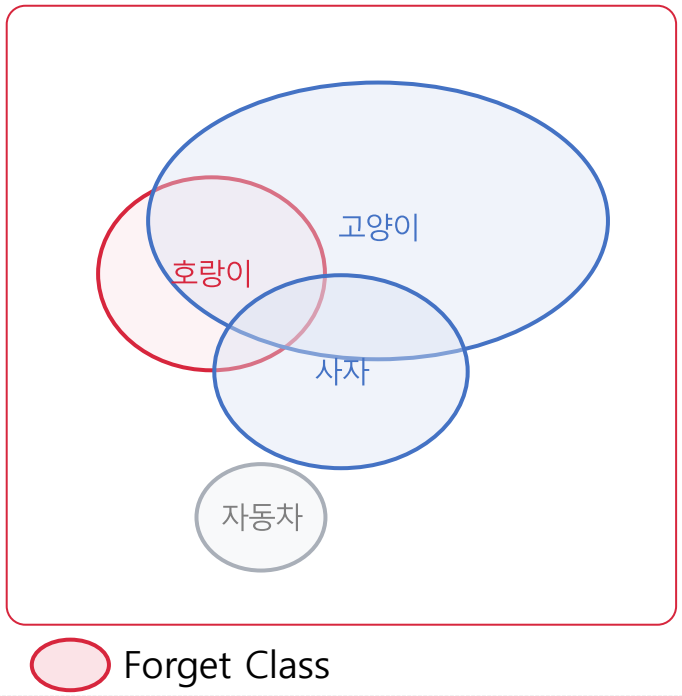
- 두 단계 모델 최적화로 접근한 Cheng et al.과, MLLM을 활용한 MeGU 소개



Entanglement: The Hidden Challenge in Class-Level Unlearning

❖ **최근 연구는 클래스 간 유사성, 즉 얽힘(Entanglement)을 고려하는 방향으로 진행**

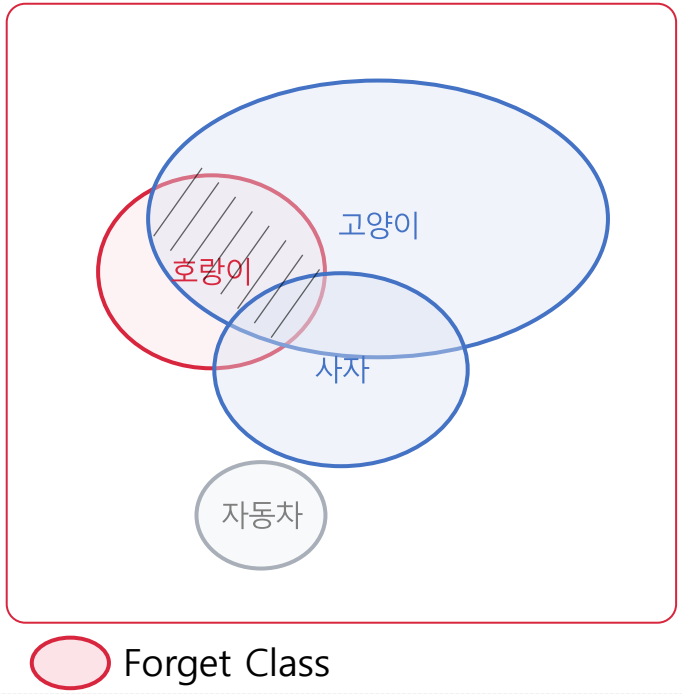
- 두 단계 모델 최적화로 접근한 Cheng et al.과, MLLM을 활용한 MeGU 소개
  - ➢ Latent space에서 유사한 클래스는 feature를 공유
  - ➢ Forget class를 지우면 의미적으로 가까운 retain class도 함께 손상
  - ➢ 예시: 호랑이(forget) 삭제 → 고양이(retain) 정확도 급락



Entanglement: The Hidden Challenge in Class-Level Unlearning

❖ **최근 연구는 클래스 간 유사성, 즉 얽힘(Entanglement)을 고려하는 방향으로 진행**

- 두 단계 모델 최적화로 접근한 Cheng et al.과, MLLM을 활용한 MeGU 소개
  - ➢ Latent space에서 유사한 클래스는 feature를 공유
  - ➢ Forget class를 지우면 의미적으로 가까운 retain class도 함께 손상
  - ➢ 예시: 호랑이(forget) 삭제 → 고양이(retain) 정확도 급락

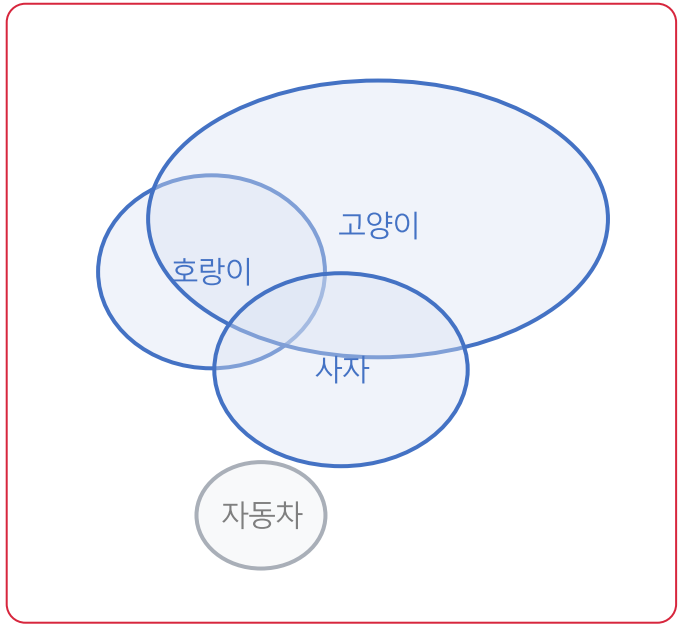


Entanglement: The Hidden Challenge in Class-Level Unlearning

❖ Machine Unlearning under Retain-Forget Entanglement

- Entanglement 정의: retain set 내 forget과 correlated된 부분집합
  - ➢ Forget: 호랑이 / 관련 retain: 고양이·사자 / 비관련 retain: 자동차
  - ➢ Phase 1 — Augmented Lagrangian: 호랑이 loss ↑, 자동차 정확도 보존
  - ➢ Phase 2 — Wasserstein-2 Gradient Projection: 고양이·사자 손상만 별도로 완화

$L(\theta_0)$



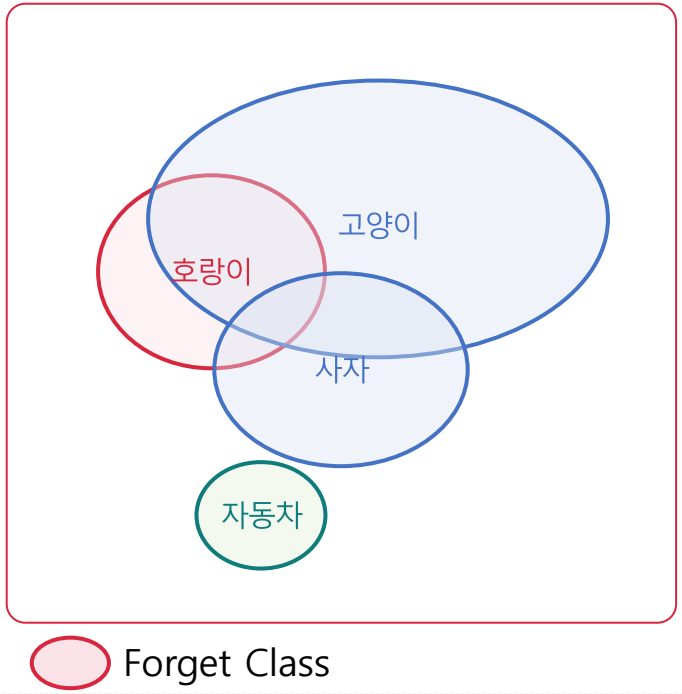
Entanglement: The Hidden Challenge in Class-Level Unlearning

❖ Machine Unlearning under Retain-Forget Entanglement

- Entanglement 정의: retain set 내 forget과 correlated된 부분집합
  - ➢ Forget: 호랑이 / 관련 retain: 고양이·사자 / 비관련 retain: 자동차
  - ➢ Phase 1 — Augmented Lagrangian: 호랑이 loss ↑, 자동차 정확도 보존
  - ➢ Phase 2 — Wasserstein-2 Gradient Projection: 고양이·사자 손상만 별도로 완화

$L(\theta)$

$L_f(\theta)$ $L_r^u(\theta)$



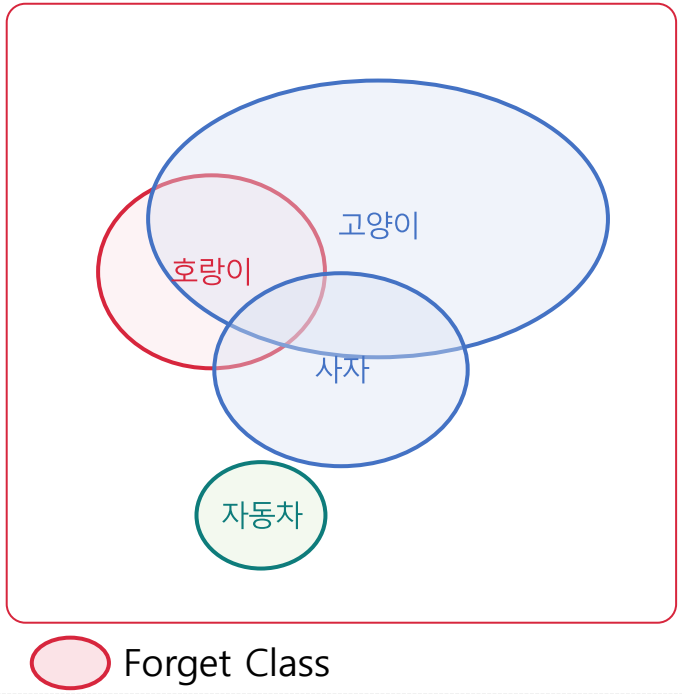
Entanglement: The Hidden Challenge in Class-Level Unlearning

❖ Machine Unlearning under Retain-Forget Entanglement

- Entanglement 정의: retain set 내 forget과 correlated된 부분집합
  - ➢ Forget: 호랑이 / 관련 retain: 고양이·사자 / 비관련 retain: 자동차
  - ➢ Phase 1 — Augmented Lagrangian: 호랑이 loss ↑, 자동차 정확도 보존
  - ➢ Phase 2 — Wasserstein-2 Gradient Projection: 고양이·사자 손상만 별도로 완화

$$L(\theta)$$

$$L_f(\theta)\uparrow \qquad L_r^u(\theta) = L_r^u(\theta_0)$$



Entanglement: The Hidden Challenge in Class-Level Unlearning

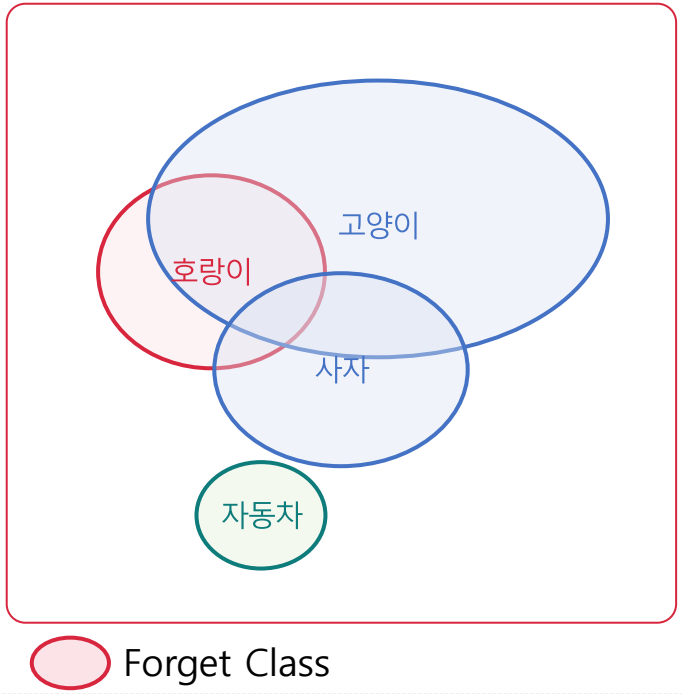
❖ Machine Unlearning under Retain-Forget Entanglement

- Entanglement 정의: retain set 내 forget과 correlated된 부분집합
  - ➢ Forget: 호랑이 / 관련 retain: 고양이·사자 / 비관련 retain: 자동차
  - ➢ Phase 1 — Augmented Lagrangian: 호랑이 loss ↑, 자동차 정확도 보존
  - ➢ Phase 2 — Wasserstein-2 Gradient Projection: 고양이·사자 손상만 별도로 완화

$L(\theta)$

$L_f(\theta)\uparrow$ $L_r^u(\theta) = L_r^u(\theta_0)$

업데이트 하면서
동일하게 만들기 어려움



Entanglement: The Hidden Challenge in Class-Level Unlearning

❖ Machine Unlearning under Retain-Forget Entanglement

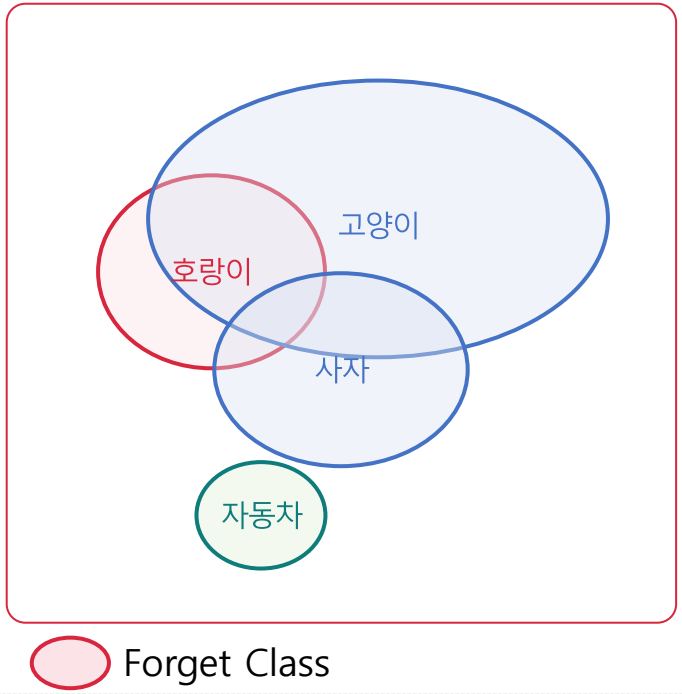
- Entanglement 정의: retain set 내 forget과 correlated된 부분집합
  - ➢ Forget: 호랑이 / 관련 retain: 고양이·사자 / 비관련 retain: 자동차
  - ➢ Phase 1 — Augmented Lagrangian: 호랑이 loss ↑, 자동차 정확도 보존
  - ➢ Phase 2 — Wasserstein-2 Gradient Projection: 고양이·사자 손상만 별도로 완화

$$L(\theta)$$

$$L_f(\theta)\uparrow \qquad L_r^u(\theta) = L_r^u(\theta_0)$$

$$L_{AL}(\theta,\lambda) = -L_f(\theta) + \lambda \cdot (L_r^u(\theta) - L_r^u(\theta_0)) + \frac{p}{2}(L_r^u(\theta) - L_r^u(\theta_0))^2$$

패널티 형식으로 반영



Entanglement: The Hidden Challenge in Class-Level Unlearning

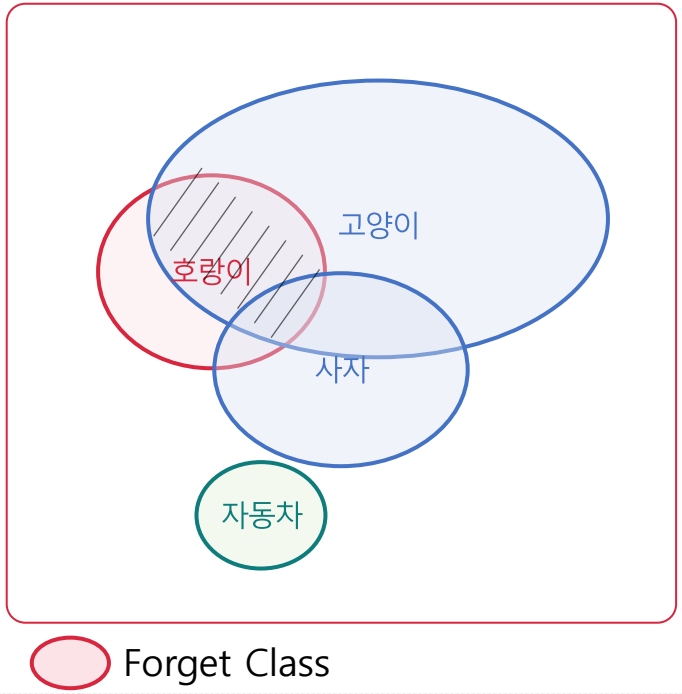
❖ Machine Unlearning under Retain-Forget Entanglement

- Entanglement 정의: retain set 내 forget과 correlated된 부분집합
  - ➢ Forget: 호랑이 / 관련 retain: 고양이·사자 / 비관련 retain: 자동차
  - ➢ Phase 1 — Augmented Lagrangian: 호랑이 loss ↑, 자동차 정확도 보존
  - ➢ Phase 2 — Wasserstein-2 Gradient Projection: 고양이·사자 손상만 별도로 완화

$$L(\theta)$$

$$L_f(\theta)\uparrow \qquad L_r^u(\theta) = L_r^u(\theta_0)$$

$$L_{AL}(\theta,\lambda) = -L_f(\theta) + \lambda \cdot (L_r^u(\theta) - L_r^u(\theta_0)) + \frac{p}{2}(L_r^u(\theta) - L_r^u(\theta_0))^2$$



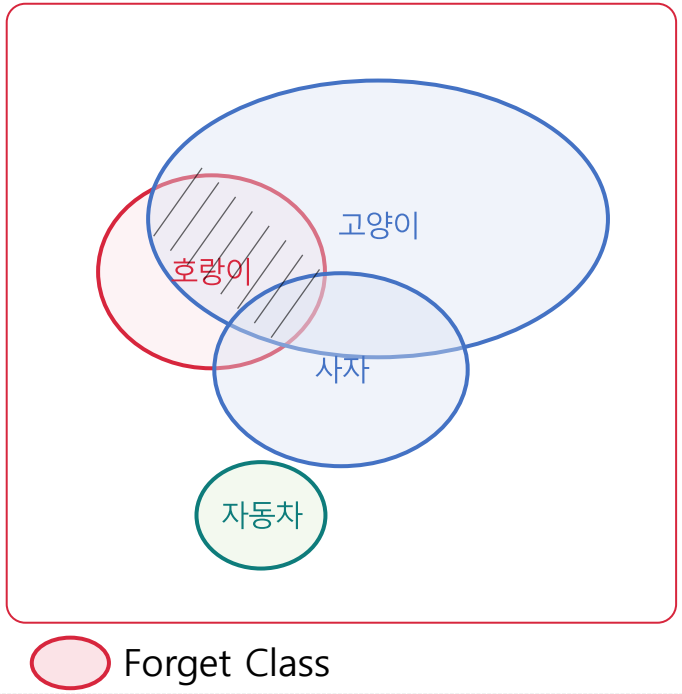
고양이, 사자 클래스 손상 발생

Entanglement: The Hidden Challenge in Class-Level Unlearning

❖ Machine Unlearning under Retain-Forget Entanglement

- Entanglement 정의: retain set 내 forget과 correlated된 부분집합
  - ➢ Forget: 호랑이 / 관련 retain: 고양이·사자 / 비관련 retain: 자동차
  - ➢ Phase 1 — Augmented Lagrangian: 호랑이 loss ↑, 자동차 정확도 보존
  - ➢ Phase 2 — Wasserstein-2 Gradient Projection: 고양이·사자 손상만 별도로 완화

$L_f(\theta)$ $L_r^a(\theta)$ $L_r^u(\theta)$



고양이, 사자 클래스
손상 완화

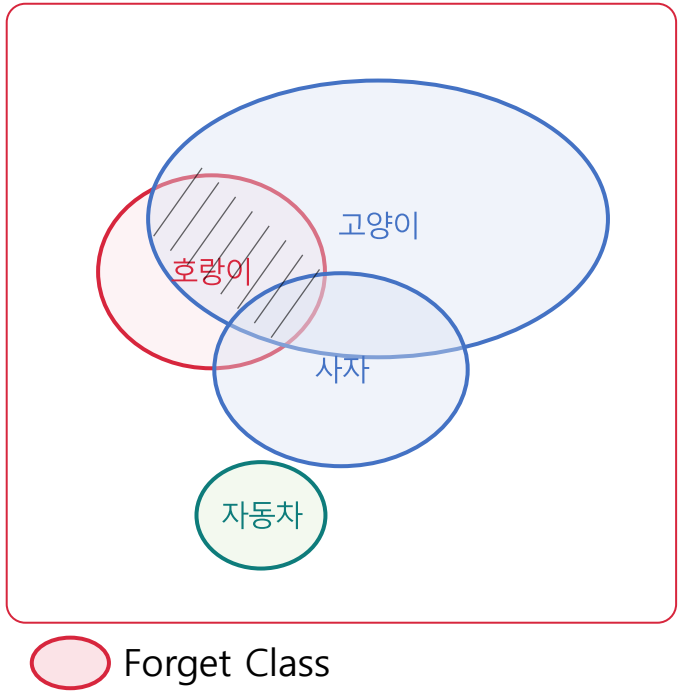
Entanglement: The Hidden Challenge in Class-Level Unlearning

❖ Machine Unlearning under Retain-Forget Entanglement

- Entanglement 정의: retain set 내 forget과 correlated된 부분집합
  - ➢ Forget: 호랑이 / 관련 retain: 고양이·사자 / 비관련 retain: 자동차
  - ➢ Phase 1 — Augmented Lagrangian: 호랑이 loss ↑, 자동차 정확도 보존
  - ➢ Phase 2 — Wasserstein-2 Gradient Projection: 고양이·사자 손상만 별도로 완화

$\nabla_\theta L_f$ $\nabla_\theta L_r^a$ $\nabla_\theta L_r^u$

각 그룹의 기준을
만들기 위해 역전파 수행



고양이, 사자 클래스
손상 완화

Entanglement: The Hidden Challenge in Class-Level Unlearning

❖ Machine Unlearning under Retain-Forget Entanglement

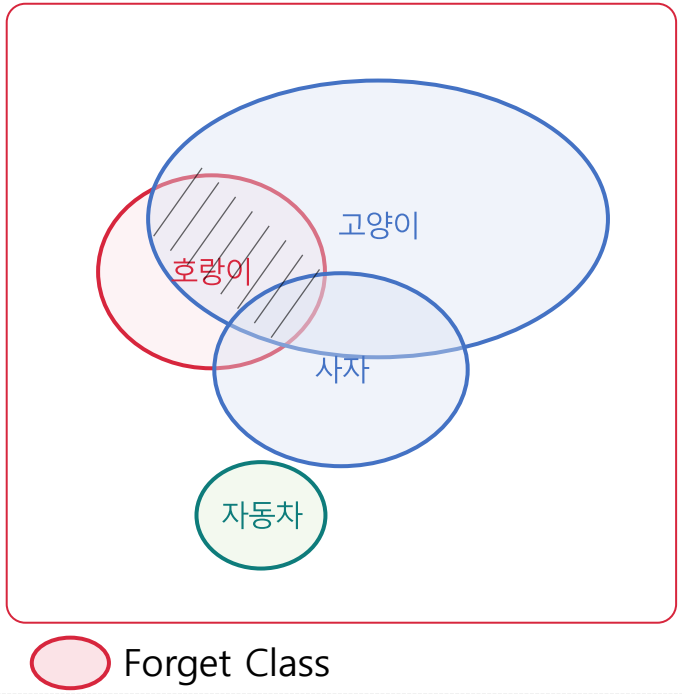
- Entanglement 정의: retain set 내 forget과 correlated된 부분집합
  - ➢ Forget: 호랑이 / 관련 retain: 고양이·사자 / 비관련 retain: 자동차
  - ➢ Phase 1 — Augmented Lagrangian: 호랑이 loss ↑, 자동차 정확도 보존
  - ➢ Phase 2 — Wasserstein-2 Gradient Projection: 고양이·사자 손상만 별도로 완화

$\nabla_\theta L_f \quad \nabla_\theta L_r^a \quad \nabla_\theta L_r^u$

$g_f \quad g_r^a \quad g_r^u$

$$u_1 = \frac{g_f}{\|g_f\|} \qquad u_2 = \frac{g_r^u - (g_r^u \cdot u_1)u_1}{\|g_r^u - (g_r^u \cdot u_1)u_1\|}$$

Gram-Schmidt orthogonalization



고양이, 사자 클래스
손상 완화

Entanglement: The Hidden Challenge in Class-Level Unlearning

❖ Machine Unlearning under Retain-Forget Entanglement

- Entanglement 정의: retain set 내 forget과 correlated된 부분집합
  - ➢ Forget: 호랑이 / 관련 retain: 고양이·사자 / 비관련 retain: 자동차
  - ➢ Phase 1 — Augmented Lagrangian: 호랑이 loss ↑, 자동차 정확도 보존
  - ➢ Phase 2 — Wasserstein-2 Gradient Projection: 고양이·사자 손상만 별도로 완화

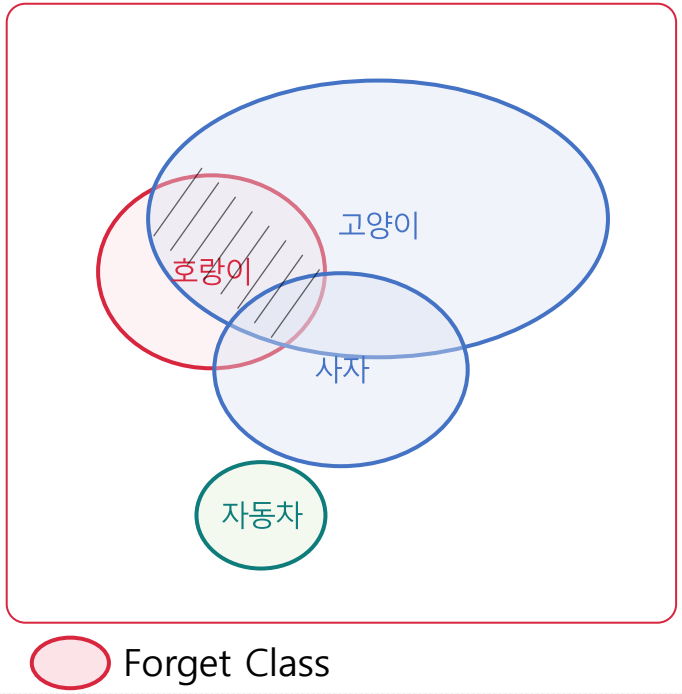
$$\nabla_\theta L_f \quad \nabla_\theta L_r^a \quad \nabla_\theta L_r^u$$

$$g_f \quad g_r^a \quad g_r^u$$

$$u_1 = \frac{g_f}{\|g_f\|} \qquad u_2 = \frac{g_r^u - (g_r^u \cdot u_1)u_1}{\|g_r^u - (g_r^u \cdot u_1)u_1\|}$$

$$\mathrm{Proj}_V(g_r^a) = (g_r^a \cdot u_1)u_1 + (g_r^a \cdot u_2)u_2$$

$$g_{\mathrm{proj}} = g_r^a - \mathrm{Proj}_V(g_r^a)$$



고양이, 사자 클래스
손상 완화

Entanglement: The Hidden Challenge in Class-Level Unlearning

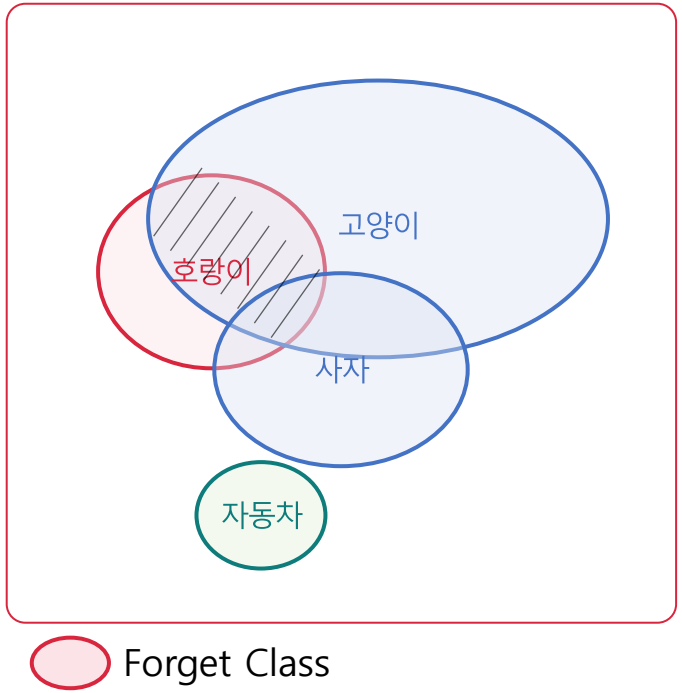
❖ Machine Unlearning under Retain-Forget Entanglement

- Entanglement 정의: retain set 내 forget과 correlated된 부분집합
  - ➢ Forget: 호랑이 / 관련 retain: 고양이·사자 / 비관련 retain: 자동차
  - ➢ Phase 1 — Augmented Lagrangian: 호랑이 loss ↑, 자동차 정확도 보존
  - ➢ Phase 2 — Wasserstein-2 Gradient Projection: 고양이·사자 손상만 별도로 완화

$\nabla_\theta L_f \quad \nabla_\theta L_r^a \quad \nabla_\theta L_r^u$

$g_f \quad g_r^a \quad g_r^u$

$\theta \leftarrow \theta - \eta_2 g_{\text{proj}}$



고양이, 사자 클래스
손상 완화

Entanglement: The Hidden Challenge in Class-Level Unlearning

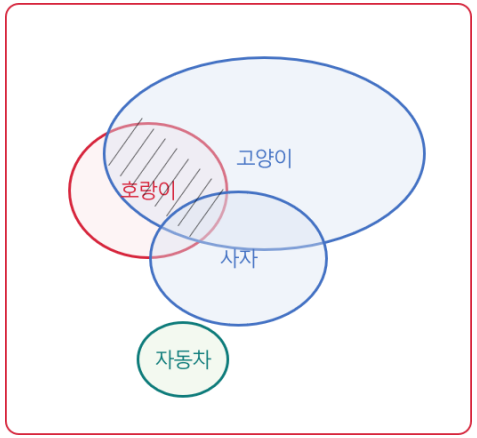
❖ Machine Unlearning under Retain-Forget Entanglement

- Entanglement 정의: retain set 내 forget과 correlated된 부분집합
  - ➢ Forget: 호랑이 / 관련 retain: 고양이·사자 / 비관련 retain: 자동차
  - ➢ Phase 1 — Augmented Lagrangian: 호랑이 loss ↑, 자동차 정확도 보존
  - ➢ Phase 2 — Wasserstein-2 Gradient Projection: 고양이·사자 손상만 별도로 완화

Table 1: Results for subclass-level unlearning on CIFAR-100 using ResNet-18. The forget set corresponds to the subclass "aquarium fish" within the "fish" superclass. SSD is a deterministic algorithm, so standard deviations are 0.

호랑이 고양이, 사자 자동차

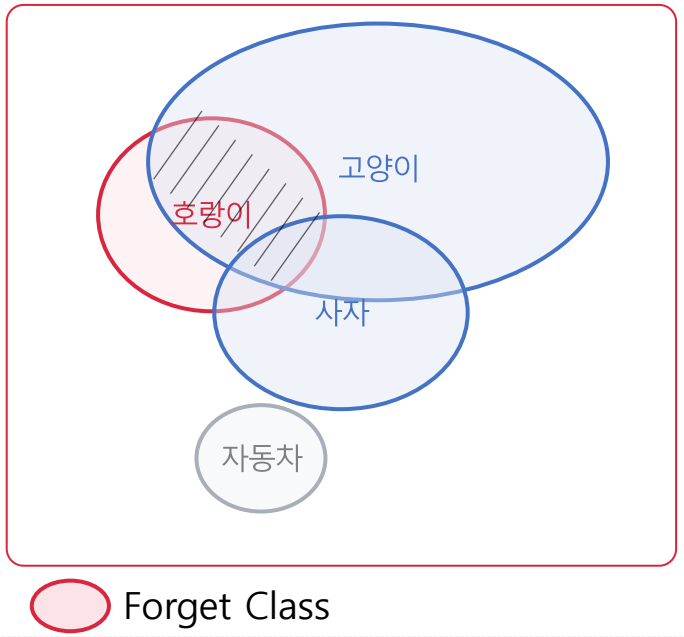
Method	Training accuracy			Test accuracy		
	$\mathcal{D}_f$	$\mathcal{D}_r^{\text{adj}}$	$\mathcal{D}_r^{\text{rem}}$	$\mathcal{D}_f$	$\mathcal{D}_r^{\text{adj}}$	$\mathcal{D}_r^{\text{rem}}$
Original	99.99	100.00	100.00	90.00	80.00	85.33
FT	$76.67_{\pm 7.76}$	$99.47_{\pm 0.52}$	$99.47_{\pm 0.33}$	$62.33_{\pm 5.79}$	$77.83_{\pm 3.88}$	$83.89_{\pm 0.41}$
GA	$70.53_{\pm 0.94}$	$72.75_{\pm 0.76}$	$91.33_{\pm 0.44}$	$56.00_{\pm 0.00}$	$59.00_{\pm 0.61}$	$80.57_{\pm 0.27}$
ℓ_1-sparse	$55.93_{\pm 7.08}$	$98.48_{\pm 0.95}$	$96.92_{\pm 0.20}$	$51.67_{\pm 7.84}$	$82.42_{\pm 2.24}$	$84.64_{\pm 0.27}$
Munba	$33.80_{\pm 8.88}$	$92.17_{\pm 2.57}$	$92.68_{\pm 1.28}$	$31.67_{\pm 4.78}$	$69.75_{\pm 3.74}$	$75.32_{\pm 1.88}$
SSD	$37.40_{\pm 0.00}$	$43.75_{\pm 0.00}$	$76.02_{\pm 0.00}$	$33.00_{\pm 0.00}$	$39.25_{\pm 0.00}$	$67.23_{\pm 0.00}$
SCRUB	$4.47_{\pm 0.25}$	$58.65_{\pm 14.73}$	$82.67_{\pm 4.24}$	$7.00_{\pm 1.41}$	$54.75_{\pm 11.34}$	$75.42_{\pm 2.95}$
SalUn	$3.20_{\pm 0.20}$	$52.27_{\pm 0.38}$	$86.35_{\pm 0.22}$	$3.00_{\pm 1.00}$	$34.90_{\pm 1.03}$	$71.78_{\pm 0.17}$
DELETE	$0.00_{\pm 0.00}$	$3.57_{\pm 0.18}$	$98.37_{\pm 0.29}$	$0.67_{\pm 0.47}$	$2.83_{\pm 0.66}$	$82.09_{\pm 0.37}$
GDR	$4.87_{\pm 1.05}$	$31.92_{\pm 6.45}$	$96.10_{\pm 0.32}$	$8.67_{\pm 1.25}$	$22.33_{\pm 4.59}$	$79.93_{\pm 0.09}$
Our method	$0.00_{\pm 0.00}$	$98.17_{\pm 0.31}$	$98.44_{\pm 0.05}$	$2.33_{\pm 0.47}$	$78.17_{\pm 0.31}$	$81.10_{\pm 0.18}$



Entanglement: The Hidden Challenge in Class-Level Unlearning

❖ MeGU: Machine-Guided Unlearning with Target Feature Disentanglement

- Entanglement 정의: 클래스 간 feature-pattern 유사도 (개념 레벨)
  - ➢ Forget: 호랑이 / 관련 retain: 고양이·사자 / 비관련 retain: 자동차
  - ➢ Transition Matrix 추정: MLLM(MMICL)으로 클래스 간 유사도 측정 → 행렬화
  - ➢ Perturbing Label 부여: 호랑이 사진마다 새 라벨 지정 (예: "고양이")
  - ➢ Fragment-Align (Feature Noise): negative noise로 고유 특징 억제, positive noise로 공유 특징 강화

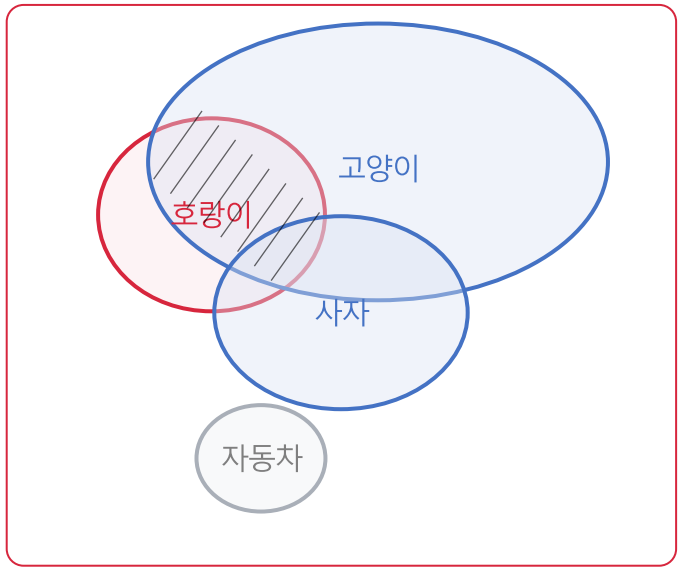


Entanglement: The Hidden Challenge in Class-Level Unlearning

❖ MeGU: Machine-Guided Unlearning with Target Feature Disentanglement

- Entanglement 정의: 클래스 간 feature-pattern 유사도 (개념 레벨)
  - ➢ Forget: 호랑이 / 관련 retain: 고양이·사자 / 비관련 retain: 자동차
  - ➢ Transition Matrix 추정: MLLM(MMICL)으로 클래스 간 유사도 측정 → 행렬화
  - ➢ Perturbing Label 부여: 호랑이 사진마다 새 라벨 지정 (예: "고양이")
  - ➢ Fragment-Align (Feature Noise): negative noise로 고유 특징 억제, positive noise로 공유 특징 강화

MMICL을 활용하여 데이터 특징 추출
(각 클래스 10장 활용)
이미지 여러 장에 대해서도
잘 이해하도록 학습된 VLM 모델



Entanglement: The Hidden Challenge in Class-Level Unlearning

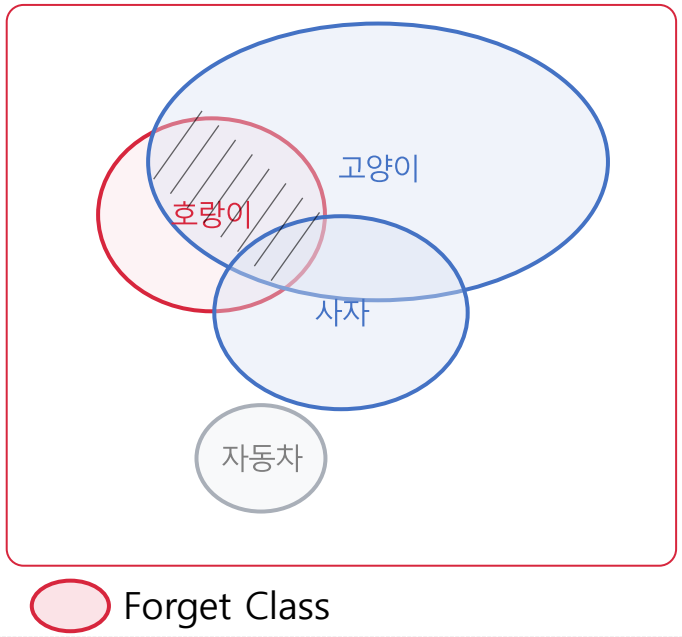
❖ MeGU: Machine-Guided Unlearning with Target Feature Disentanglement

- Entanglement 정의: 클래스 간 feature-pattern 유사도 (개념 레벨)
  - ➢ Forget: 호랑이 / 관련 retain: 고양이·사자 / 비관련 retain: 자동차
  - ➢ Transition Matrix 추정: MLLM(MMICL)으로 클래스 간 유사도 측정 → 행렬화
  - ➢ Perturbing Label 부여: 호랑이 사진마다 새 라벨 지정 (예: "고양이")
  - ➢ Fragment-Align (Feature Noise): negative noise로 고유 특징 억제, positive noise로 공유 특징 강화

MMICL을 활용하여 데이터 특징 추출
(각 클래스 10장 활용)

이미지 여러 장에 대해서도
잘 이해하도록 학습된 VLM 모델

	호랑이	고양이	사자	자동차
호랑이	1	0.7	0.9	0.1
고양이	0.7	1	0.6	0.0
사자	0.9	0.6	1	0.1
자동차	0.1	0.0	0.1	1



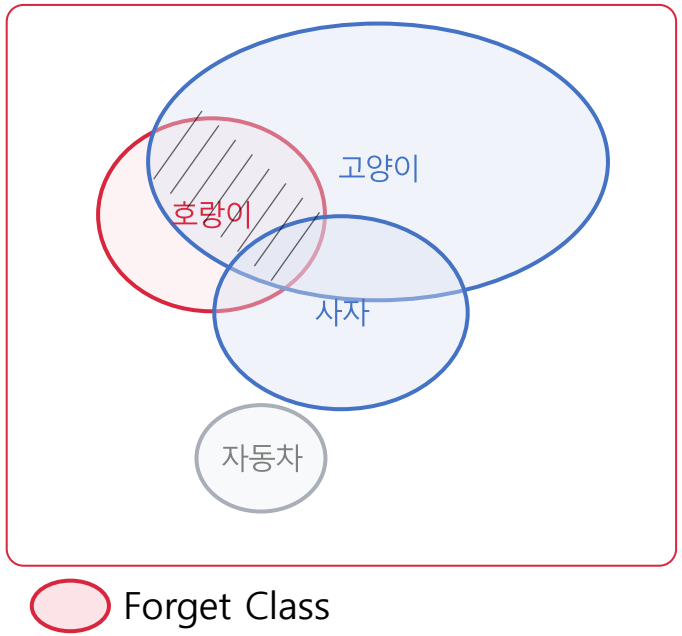
Entanglement: The Hidden Challenge in Class-Level Unlearning

❖ MeGU: Machine-Guided Unlearning with Target Feature Disentanglement

- Entanglement 정의: 클래스 간 feature-pattern 유사도 (개념 레벨)
  - ➢ Forget: 호랑이 / 관련 retain: 고양이·사자 / 비관련 retain: 자동차
  - ➢ Transition Matrix 추정: MLLM(MMICL)으로 클래스 간 유사도 측정 → 행렬화
  - ➢ Perturbing Label 부여: 호랑이 사진마다 새 라벨 지정 (예: "고양이")
  - ➢ Fragment-Align (Feature Noise): negative noise로 고유 특징 억제, positive noise로 공유 특징 강화

MMICL을 활용하여 데이터 특징 추출
(각 클래스 10장 활용)
이미지 여러 장에 대해서도
잘 이해하도록 학습된 VLM 모델

	호랑이	고양이	사자	자동차
호랑이	1	0.7	**0.9**	0.1
고양이	0.7	1	0.6	0.0
사자	0.9	0.6	1	0.1
자동차	0.1	0.0	0.1	1



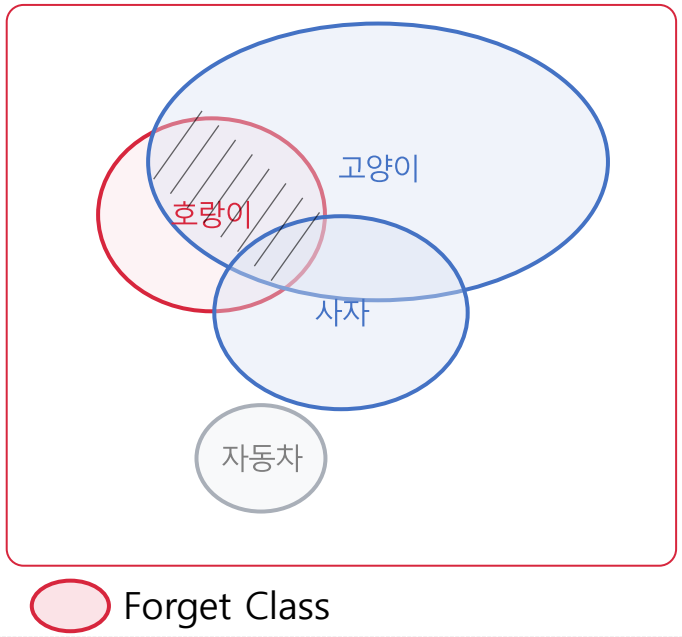
Entanglement: The Hidden Challenge in Class-Level Unlearning

❖ MeGU: Machine-Guided Unlearning with Target Feature Disentanglement

- Entanglement 정의: 클래스 간 feature-pattern 유사도 (개념 레벨)
  - ➢ Forget: 호랑이 / 관련 retain: 고양이·사자 / 비관련 retain: 자동차
  - ➢ Transition Matrix 추정: MLLM(MMICL)으로 클래스 간 유사도 측정 → 행렬화
  - ➢ Perturbing Label 부여: 호랑이 사진마다 새 라벨 지정 (예: "고양이")
  - ➢ Fragment-Align (Feature Noise): negative noise로 고유 특징 억제, positive noise로 공유 특징 강화

MMICL을 활용하여 데이터 특징 추출
(각 클래스 10장 활용)
이미지 여러 장에 대해서도
잘 이해하도록 학습된 VLM 모델

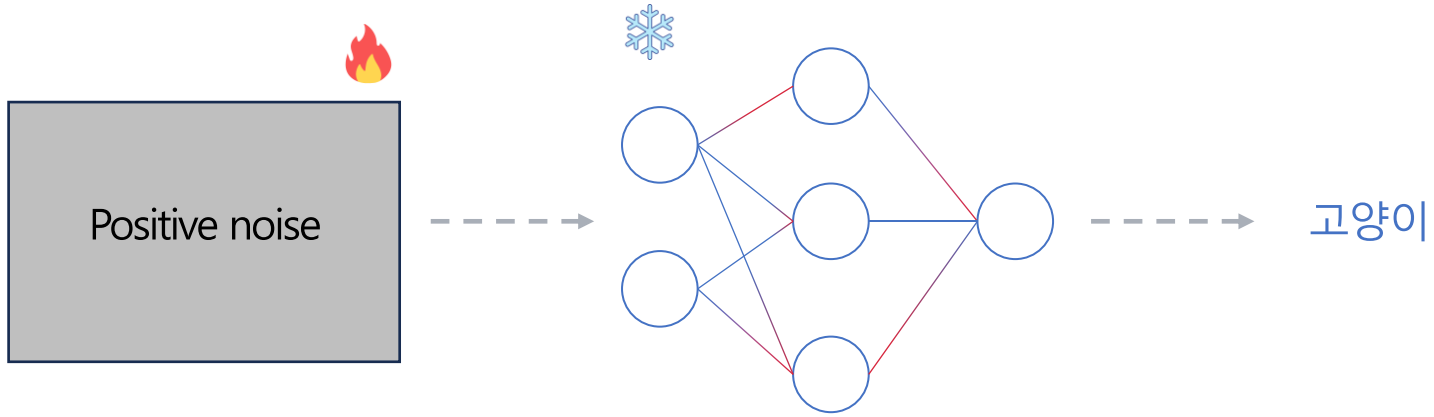
	호랑이	고양이	사자	자동차
호랑이	1	**0.7**	0.9	0.1
고양이	0.7	1	0.6	0.0
사자	0.9	0.6	1	0.1
자동차	0.1	0.0	0.1	1



Entanglement: The Hidden Challenge in Class-Level Unlearning

❖ MeGU: Machine-Guided Unlearning with Target Feature Disentanglement

- Entanglement 정의: 클래스 간 feature-pattern 유사도 (개념 레벨)
  - ➢ Forget: 호랑이 / 관련 retain: 고양이·사자 / 비관련 retain: 자동차
  - ➢ Transition Matrix 추정: MLLM(MMICL)으로 클래스 간 유사도 측정 → 행렬화
  - ➢ Perturbing Label 부여: 호랑이 사진마다 새 라벨 지정 (예: "고양이")
  - ➢ Fragment-Align (Feature Noise): negative noise로 고유 특징 억제, positive noise로 공유 특징 강화

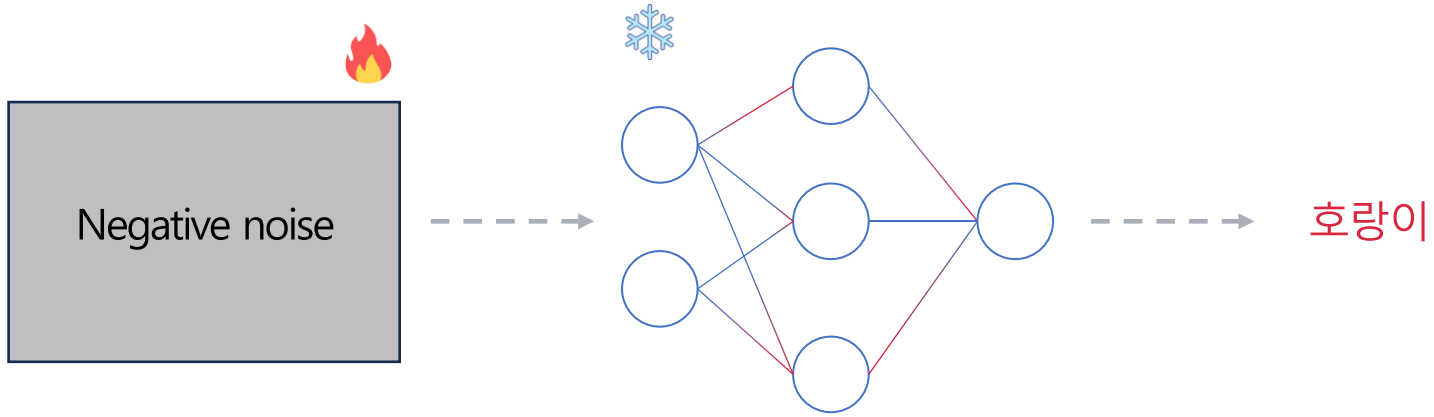


모델이 "고양이"라고 확신하게 만드는 패턴

Entanglement: The Hidden Challenge in Class-Level Unlearning

❖ MeGU: Machine-Guided Unlearning with Target Feature Disentanglement

- Entanglement 정의: 클래스 간 feature-pattern 유사도 (개념 레벨)
  - ➢ Forget: 호랑이 / 관련 retain: 고양이·사자 / 비관련 retain: 자동차
  - ➢ Transition Matrix 추정: MLLM(MMICL)으로 클래스 간 유사도 측정 → 행렬화
  - ➢ Perturbing Label 부여: 호랑이 사진마다 새 라벨 지정 (예: "고양이")
  - ➢ Fragment-Align (Feature Noise): negative noise로 고유 특징 억제, positive noise로 공유 특징 강화

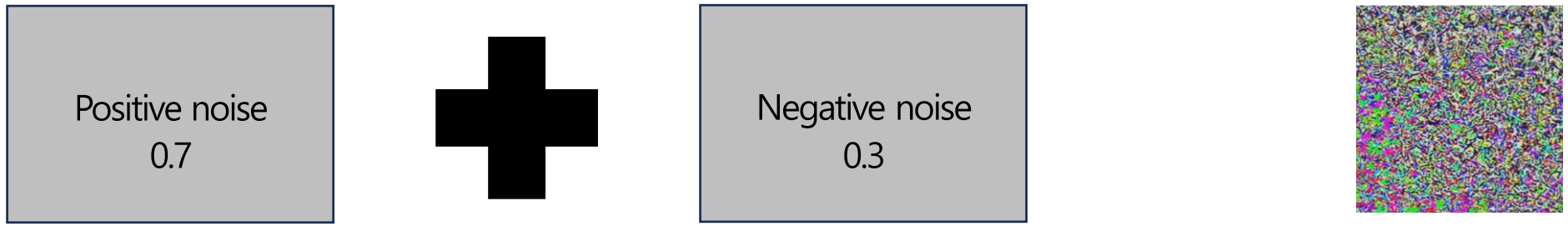


모델이 "호랑이로는 안 보이게" 만드는 패턴

Entanglement: The Hidden Challenge in Class-Level Unlearning

❖ MeGU: Machine-Guided Unlearning with Target Feature Disentanglement

- Entanglement 정의: 클래스 간 feature-pattern 유사도 (개념 레벨)
  - ➢ Forget: 호랑이 / 관련 retain: 고양이·사자 / 비관련 retain: 자동차
  - ➢ Transition Matrix 추정: MLLM(MMICL)으로 클래스 간 유사도 측정 → 행렬화
  - ➢ Perturbing Label 부여: 호랑이 사진마다 새 라벨 지정 (예: "고양이")
  - ➢ Fragment-Align (Feature Noise): positive noise로 공유 특징 강화, negative noise로 고유 특징 억제

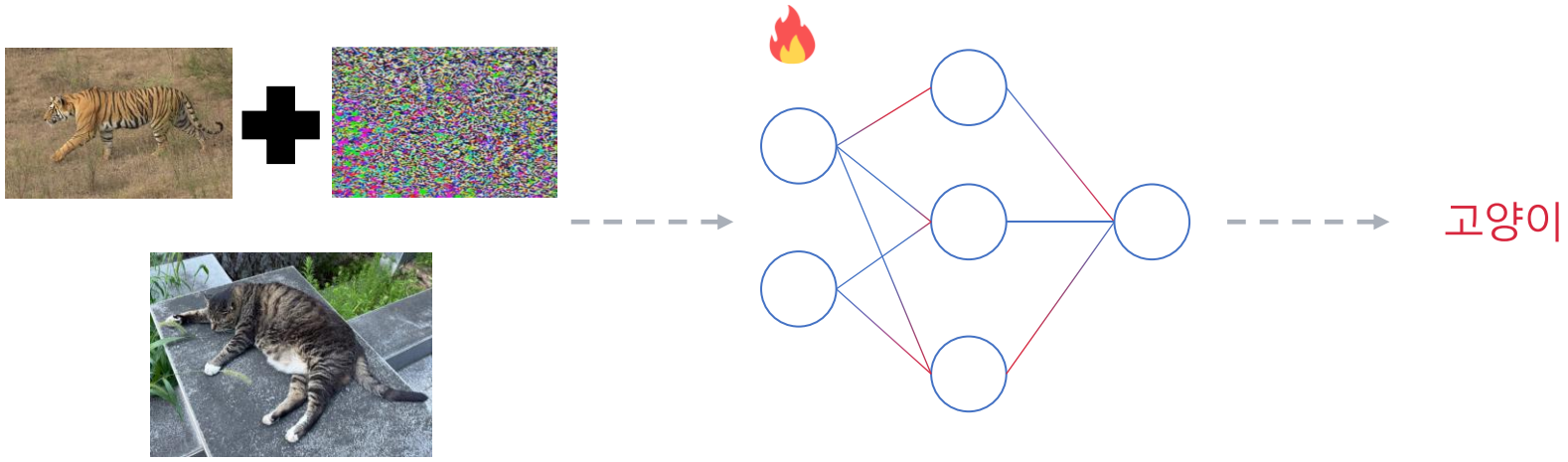


모델이 "고양이"라고 확신하게 만드는 패턴　　모델이 "호랑이로는 안 보이게" 만드는 패턴

Entanglement: The Hidden Challenge in Class-Level Unlearning

❖ MeGU: Machine-Guided Unlearning with Target Feature Disentanglement

- Entanglement 정의: 클래스 간 feature-pattern 유사도 (개념 레벨)
  - ➢ Forget: 호랑이 / 관련 retain: 고양이·사자 / 비관련 retain: 자동차
  - ➢ Transition Matrix 추정: MLLM(MMICL)으로 클래스 간 유사도 측정 → 행렬화
  - ➢ Perturbing Label 부여: 호랑이 사진마다 새 라벨 지정 (예: "고양이")
  - ➢ Fragment-Align (Feature Noise): positive noise로 공유 특징 강화, negative noise로 고유 특징 억제



Entanglement: The Hidden Challenge in Class-Level Unlearning

❖ MeGU: Machine-Guided Unlearning with Target Feature Disentanglement

- Entanglement 정의: 클래스 간 feature-pattern 유사도 (개념 레벨)
  - ➢ Forget: 호랑이 / 관련 retain: 고양이·사자 / 비관련 retain: 자동차
  - ➢ Transition Matrix 추정: MLLM(MMICL)으로 클래스 간 유사도 측정 → 행렬화
  - ➢ Perturbing Label 부여: 호랑이 사진마다 새 라벨 지정 (예: "고양이")
  - ➢ Fragment-Align (Feature Noise): positive noise로 공유 특징 강화, negative noise로 고유 특징 억제

Table 1: Class-wise unlearning on CIFAR-100.

model	class	metric	baseline	retrain	FT	UNSIR	AMNC	SSD	**MeGU**
RN18	RKT	A_r	76.30	76.19	65.43	73.83	73.59	**75.86**	75.75
		A_f	82.81	0.00	**0.00**	41.15	**0.00**	**0.00**	**0.00**
		MIA	96.61	8.06	10.04	3.08	28.62	0.66	**0.00**
	MR	A_r	76.38	76.23	63.90	74.26	73.22	76.20	**76.25**
		A_f	82.03	0.00	**0.00**	8.07	**0.00**	**0.00**	**0.00**
		MIA	95.65	6.01	12.22	1.68	46.06	0.25	**0.00**
ViT	RKT	A_r	92.27	92.04	84.26	90.83	90.53	91.39	**92.34**
		A_f	93.14	0.00	**0.00**	24.57	**0.00**	**0.00**	**0.00**
		MIA	84.88	6.29	16.00	11.43	1.06	6.62	**0.00**
	MR	A_r	92.20	92.18	84.15	89.95	89.95	91.78	**92.33**
		A_f	98.44	0.00	**0.00**	56.25	**0.00**	**0.00**	**0.00**
		MIA	90.24	0.86	5.05	2.61	0.88	1.45	**0.00**

고양이, 사자 자동차
호랑이

Conclusion

❖ Class-Level Unlearning

- Class-Level
  - ➢ 개별 샘플 삭제 요청이 많지만, 클래스 전체 삭제는 파급력이 큼
  - ➢ Exact(재학습)는 확실하지만 비용 큼 → Approximate가 실용적 대안
  - ➢ 목표는 Forget Acc 0%가 아니라, Retrained model과 구분 불가능한 행동 재현
- Challenge of Entanglement
  - ➢ 유사 클래스 간 feature 공유로, forget class를 지우면 인접 retain class도 함께 손상
  - ➢ Cheng et al. — 데이터 구조 기반 2-phase 제약 최적화로 정밀 보호
  - ➢ MeGU — MLLM 가이드로 개념 유사도 파악 후 feature 재정렬

고맙습니다