최신 논문

AAAI 2024

1. Backdoor Attacks via Machine Unlearning

DOI: https://doi.org/10.1609/aaai.v38i13.29321

Keywords: ML: Adversarial Learning & Robustness, ML: Classification and Regression, PEAI: Safety, Robustness & Trustworthiness

특이사항 : Attack without Poisoning, Attack with Poisoning, classification 실험

2. Layer Attack Unlearning: Fast and Accurate Machine Unlearning via Layer Level Attack and Knowledge Distillation

DOI: https://doi.org/10.1609/aaai.v38i19.30118

Keywords: General

특이사항 : Partial-PGD Unlearning, Model-Agnostic Unlearning Method, Knowledge Distillation (KD), classification 실험, 성균관대학교

3. Stable Unlearnable Example: Enhancing the Robustness of Unlearnable Examples via Stable Error Minimizing Noise

DOI: https://doi.org/10.1609/aaai.v38i4.28169

Keywords: CV: Bias, Fairness & Privacy, PEAI: Privacy & Security, CV: Adversarial Attacks & Robustness, ML: Privacy

특이사항 : Stable Error-Minimizing noise (SEM), 학습 데이터를 unlearnable하게 만드는 방법

4. Separate the Wheat from the Chaff: Model Deficiency Unlearning via Parameter-Efficient Module Operation

https://arxiv.org/abs/2308.08090

특이사항 : LLM

5. Learning to Unlearn: Instance-Wise Unlearning for Pre-trained Classifiers

6. Fast Machine Unlearning without Retraining through Selective Synaptic Dampening

7. Detection and Defense of Unlearnable Examples

DOI: https://doi.org/10.1609/aaai.v38i15.29667

Keywords: ML: Adversarial Learning & Robustness, CV: Adversarial Attacks & Robustness

특이사항 : Unlearnable Examples 탐지

8. Towards Effective and General Graph Unlearning via Mutual Evolution

DOI: https://doi.org/10.1609/aaai.v38i12.29273

Keywords: ML: Graph-based Machine Learning, ML: Deep Learning Algorithms, ML: Semi-Supervised Learning, ML: Privacy

특이사항 : GNN unlearn

9. From Hope to Safety: Unlearning Biases of Deep Models via Gradient Penalization in Latent Space

DOI: https://doi.org/10.1609/aaai.v38i19.30096

Keywords: General

특이사항 : DNN bias correction method RR-ClArC

10. Game-Theoretic Unlearnable Example Generator

DOI: https://doi.org/10.1609/aaai.v38i19.30130

Keywords: General

특이사항 : Unlearnable Example Generator

11. Feature Unlearning for Pre-trained GANs and VAEs

DOI: https://doi.org/10.1609/aaai.v38i19.30138

Keywords: General

특이사항 : Pre-trained GANs and VAEs, Generation, Target identification in latent space, 포스텍

CVPR 2024

1. Generative Unlearning for Any Identity

https://arxiv.org/abs/2405.09879

특이사항 : Generative Unlearning for Any IDEntity(GUIDE), Latent Target Unlearning, Generation, 경희대

2. One-dimensional Adapter to Rule Them All: Concepts, Diffusion Models and Erasing Applications

https://arxiv.org/abs/2312.16145

특이사항 : concept erosion, diffusion models

ICML 2024

1. Towards Certified Unlearning for Deep Neural Networks

https://openreview.net/forum?id=1mf1ISuyS3

특이사항 : DNNs, Hessian approximation, classification

2. In-Context Unlearning: Language Models as Few-Shot Unlearners

https://arxiv.org/abs/2310.07579

특이사항 : LLM

3. To Each (Textual Sequence) Its Own: Improving Memorized-Data Unlearning in Large Language Models

https://arxiv.org/abs/2405.03097

특이사항 : LLM

4. Verification of Machine Unlearning is Fragile

https://openreview.net/pdf?id=OkChMnjF6s

특이사항 : adversarial unlearning process, 언러닝 검증

5. One for All: A Universal Generator for Concept Unlearnability via Multi-Modal Alignment

https://openreview.net/forum?id=vSerUPYFtB&referrer=%5Bthe%20profile%20of%20Chaochao%20Chen%5D(%2Fprofile%3Fid%3D~Chaochao\_Chen3)

특이사항 : Concept Unlearnability, Cross-dataset Transferability, Label-agnostic, Multi-modal Embedding 활용

ICLR 2024

1. SalUn: Empowering Machine Unlearning via Gradient-based Weight Saliency in Both Image Classification and Generation

2. Machine Unlearning for Image-to-Image Generative Models

https://arxiv.org/abs/2402.00351

특이사항 : image to image, diffusion, GAN

3. Label-Agnostic Forgetting: A Supervision-Free Unlearning in Deep Models

https://openreview.net/forum?id=SIZWiya7FE

특이사항 : Label-Agnostic Forgetting, LAF, VAE

WACV 2024

1. Learn to Unlearn for Deep Neural Networks: Minimizing Unlearning Interference with Gradient Projection

https://arxiv.org/abs/2312.04095

특이사항 : PGU(Projected-Gradient Unlearning), Core Gradient Space (CGS), classification

ICASSP 2024

1. Client-Free Federated Unlearning via Training Reconstruction with Anchor Subspace Calibration

https://doi.org/10.1109/icassp48485.2024.10447085

2. eCIL-MU: Embedding based Class Incremental Learning and Machine Unlearning

https://arxiv.org/abs/2401.02457

IJCAI 2024

1. Machine Unlearning via Null Space Calibration

NeurIPS 2024, InterSpeech 2024

FG 2024

x

ECCV 2024

https://sites.google.com/view/u-and-me-workshop 2024.09.29

그 외 Generation 논문

1. Erasing Concepts from Diffusion Models(ICCV 2023)

https://arxiv.org/abs/2303.07345

특이사항 : DDPM, LDM, Model editing

2. Forget-me-not: Learning to forget in text-to-image diffusion models.(2023)

https://arxiv.org/pdf/2303.17591

특이사항 : Forget-me-not, Attention Resteering

3. To Generate or Not? Safety-Driven Unlearned Diffusion Models Are Still Easy To Generate Unsafe Images ... For Now(2024)

https://arxiv.org/abs/2310.11868

특이사항 : UnlearnDiffAtk

4. Defensive Unlearning with Adversarial Training for Robust Concept Erasure in Diffusion Models.(2024.05)

https://arxiv.org/abs/2405.15234

특이사항 : AdvUnlearn, Utility-Retaining Regularization