Empowering Machine Unlearning through Model Sparsity

Sijia Liu

Assistant Professor, OPTML Lab,

Dept. CSE, Michigan State University

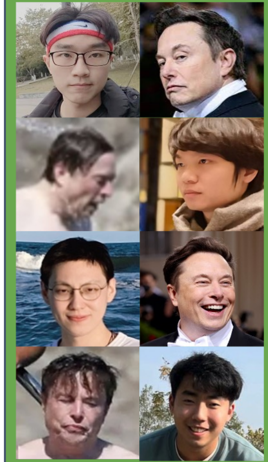
Affiliated Professor, MIT-IBM Watson AI Lab

MICHIGAN STATE UNIVERSITY

OPTML

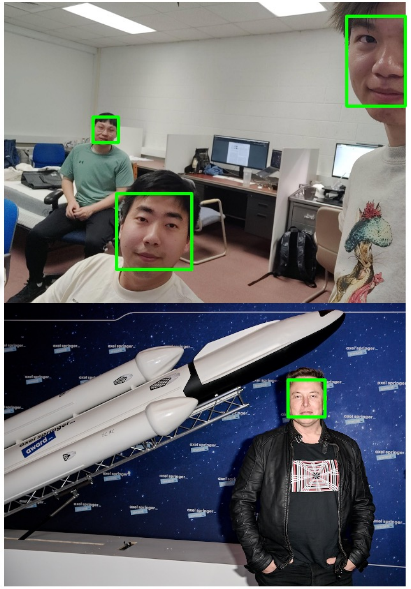
What is Machine Unlearning?

Dataset



Training

Deep Neural Network



MICHIGAN STATE UNIVERSITY

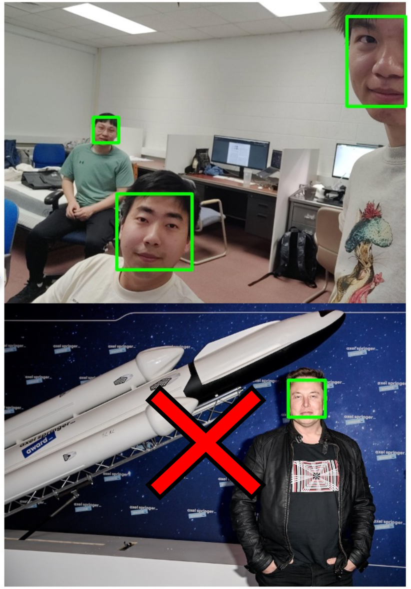
OPTML

What is Machine Unlearning?

Dataset w/ regulation

Training

New Model



Machine unlearning (MU): Erase influence of specific data/classes in model performance, e.g., to comply with data privacy regulations

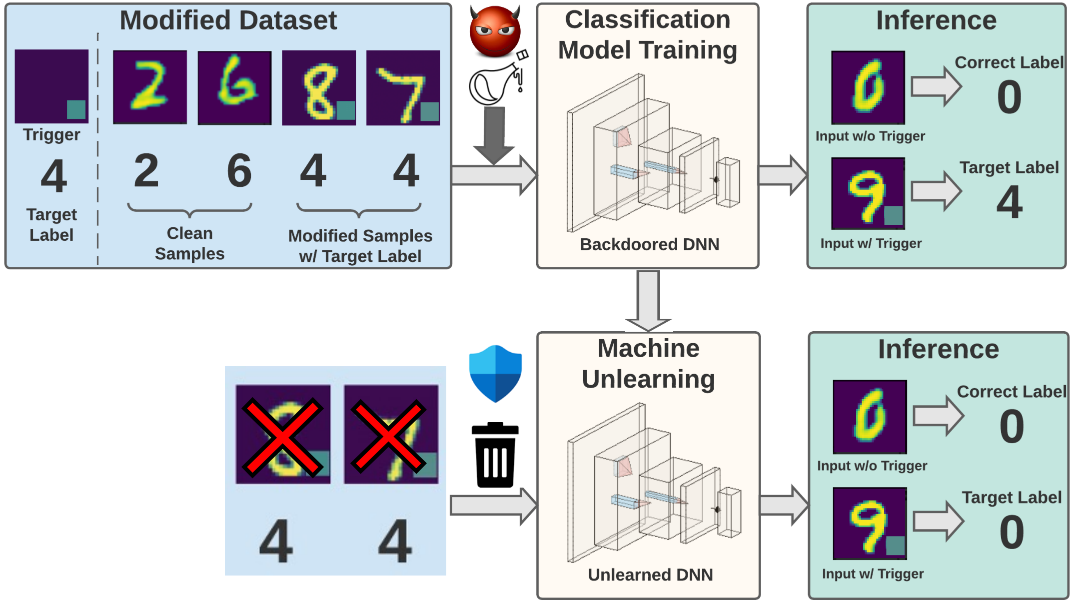
Cao and Yang, "Towards making systems forget with machine unlearning," 2015

MICHIGAN STATE UNIVERSITY

OPTML

New Opportunities Provided by MU

Defense in Trojan AI: Mitigating harmful influence of poisoned training data points

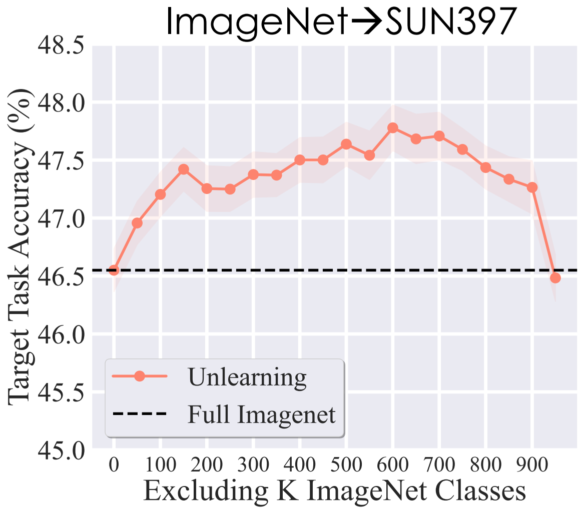


Liu, Ma et al., "Backdoor defense with machine unlearning," 2022.

New Opportunities Provided by MU

Improved transfer learning: Improving source model by "**pruning**" source data points that have **harmful influence** in downstream tasks

Example: By unlearning the "harmful" source classes (e.g., ImageNet) [Jain & Madry, 2022], the pretrained model (e.g., ResNet18) can achieve much better performance in downstream tasks (e.g., SUN397).



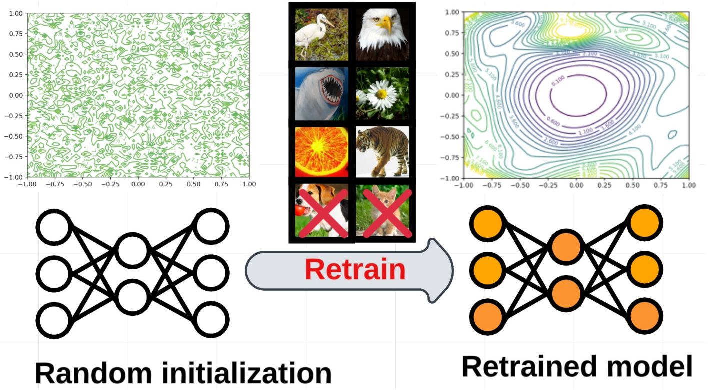
What is Machine Unlearning?

MU is a generic framework for "updating" models to comply with "data manipulation" requests, which draws the connection between **data influence** and **model influence**

Is MU equal to finetuning? No! Finetuning is inefficient to unlearn data influence on model weights

Why Are The Challenges of MU?

- The **optimal MU** strategy: Retrain the model from **scratch** over retaining dataset (after removing data points to be unlearned)



- **Downside:** Lacks training efficiency, particularly for large-scale deep models

Training Challenge: How to develop "**fast**" training methods for MU without losing unlearning **effectiveness** ("optimality")?

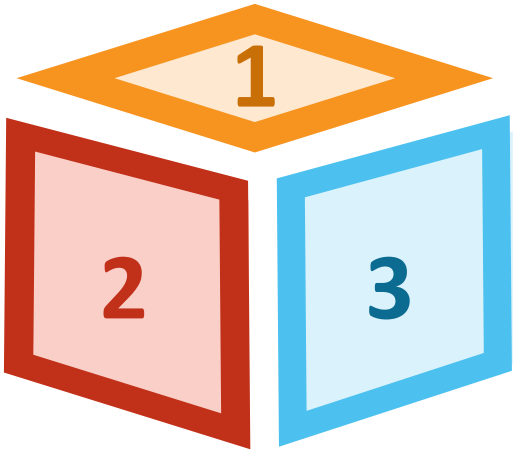
MICHIGAN STATE UNIVERSITY

OPTML

Why Are The Challenges of MU?

- **Evaluation challenge:** Multiple unlearning performance metrics

Computation efficiency



Preserved generalization

Testing accuracy of "unlearned" model

Unlearning efficacy

Whether or not truly remove impact of unlearned data points?
E.g., measured by membership inference attack (MIA), accuracy on unlearned data points

Existing Methods and Limitation

- **Retraining from scratch (exact unlearning):** Most effective but least efficient
- **Approximate unlearning:** More efficient but lacks optimality guarantees
  - **Influence function**-based approaches (require second-order derivatives):
    - Influence unlearning (IU) [Liang, et ail., 2017]
    - Fisher forgetting (FF) [Soatto, et al., 2020]
  - **Heuristics-based** approaches (computationally lightest):
    - Fine-tuning (FT) on remaining training set
    - Gradient ascent (GA) [Thudi, et al., 2022]

Limitation: There exists a significant performance **gap** between **exact unlearning** and **approximate unlearning**

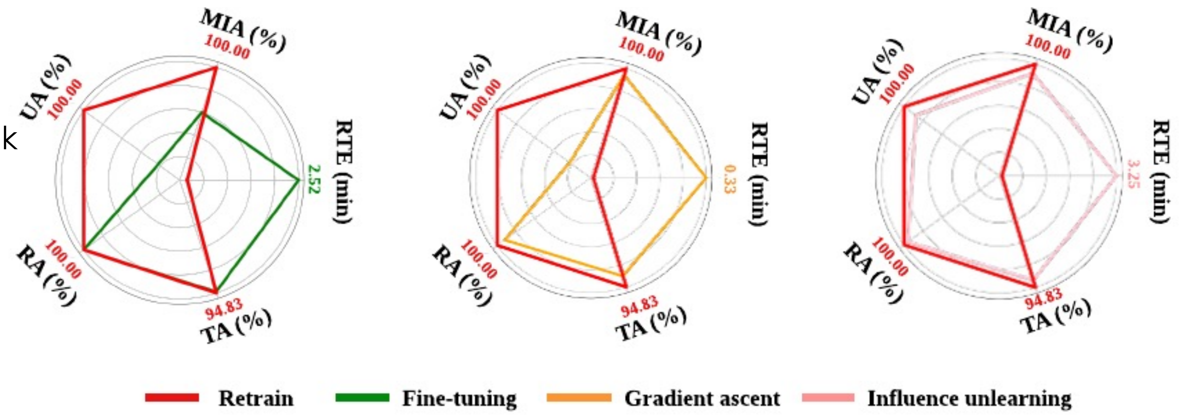
MICHIGAN STATE UNIVERSITY

OPTML

Existing Methods and Limitation

- **Performance gap between exact unlearning and approximate unlearning**

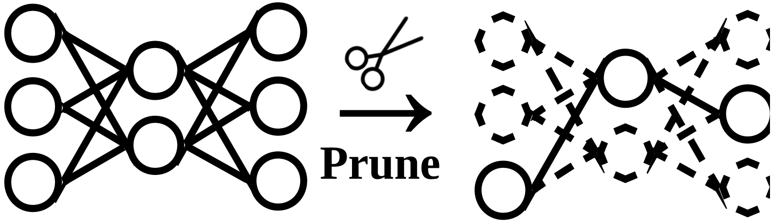
UA: unlearning accuracy
RA: retaining accuracy
MIA: membership inference attack
TA: testing accuracy
RTE: run-time efficiency



Our goal: Develops a **theoretically-grounded** and **broadly-applicable** method to close the performance gap

Improving MU: A Model Pruning-based Perspective

- **What is model pruning?**

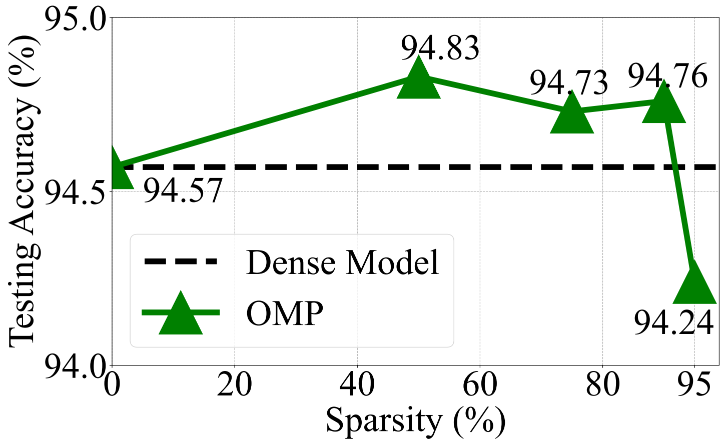


Model pruning: Finds a **sparse** sub-network without losing generalization ability

Jia, Liu, Ram, Liu et al., Model sparsification can simplify machine unlearning, arXiv, 2023

Improving MU: A Model Pruning-based Perspective

- **Pruning yields a sparse model without generalization loss**



Testing accuracy of pruned ResNet-18 vs. pruning ratio on CIFAR-10 using One-shot Magnitude Pruning (OMP)

Ma, Xiaolong, et al. "Sanity checks for lottery tickets: Does your winning ticket really win the jackpot?" NeurIPS, 2021

MICHIGAN STATE UNIVERSITY

OPTML

Pruning Helps Unlearning

- Pruning introduces "sparsity", thus needs "less" model weights to be modified for MU

  Intuition: Reduces unlearning dimension and unlearning error, i.e., the gap between approximate unlearning and exact unlearning (retrain from scratch)

- **Provable guarantee:**

  Theorem: Given SGD-based training and model pruning mask $\boldsymbol{m}$, the unlearning error, $e(\boldsymbol{m})$, characterized by weight distance between **an approximate unlearner** and the **exact unlearner** yields

  $$e(\boldsymbol{m}) = \mathcal{O}(|\boldsymbol{m} \odot (\boldsymbol{\theta}_t - \boldsymbol{\theta}_0)|_2)$$

  $\odot$ is entry-wise product, $\boldsymbol{\theta}_t$ is model trained after t SGD iterations

- **Sparsity:** Helps reduce unlearning error, possible tradeoff with generalization

MICHIGAN STATE UNIVERSITY

OPTML

How to Integrate Pruning with Unlearning?

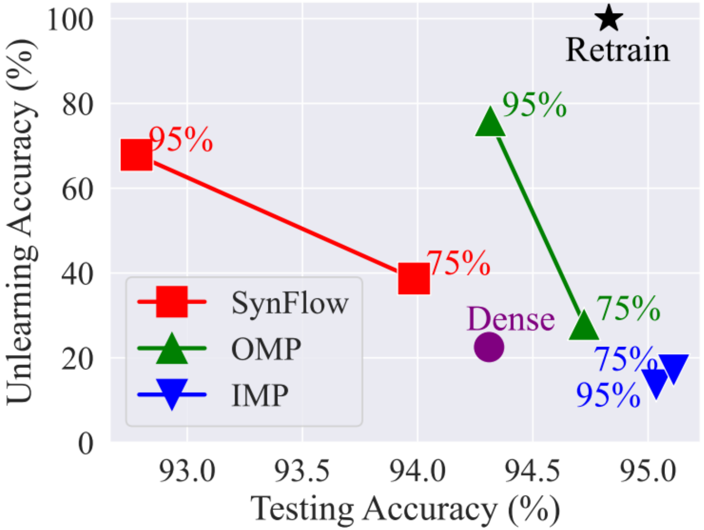
- **Which weight pruning method should be used for MU?**
  - ➢ (C1) Light computation
  - ➢ (C2) No generalization drop
  - ➢ (C3) Pruning has least dependence on forgetting data points (to be unlearned)
- Pruning methods:
  - SOTA **iterative magnitude pruning (IMP)** [Frankle & Carbin, 2018] violates (C1) & (C3)
  - Other options?

MICHIGAN STATE UNIVERSITY

OPTML

How to Integrate Pruning with Unlearning?

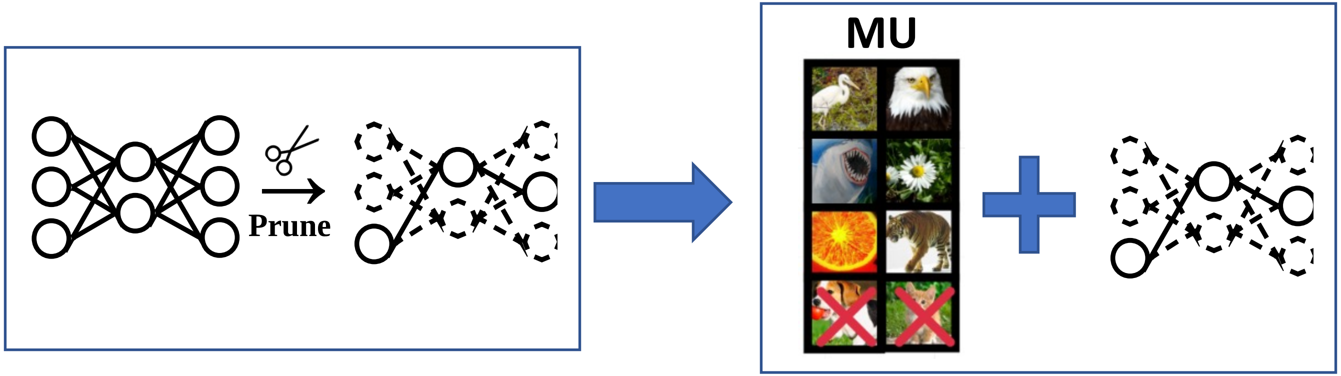
- **Suggested** pruning methods:
  - **Pruning at random initialization (e.g., SynFlow)**
    🙂 Moderate computation cost
    😐 A bit generalization drop
    😍 Least dependence on forgetting dataset
  - (**Best**) **One-shot magnitude pruning (OMP)**
    😍 Lightest in computation
    🙂 Competitive generalization performance
    😍 Least dependence on forgetting dataset



Tanaka, Hidenori, et al. "Pruning neural networks without any data by iteratively conserving synaptic flow." NeurIPS, 2020
Ma, Xiaolong, et al. "Sanity checks for lottery tickets: Does your winning ticket really win the jackpot?." NeurIPS, 2021

How to Integrate Pruning with Unlearning?

- **(Strategy 1) Prune first, then unlearn**: Find sparse model first, then applies existing approximate unlearning methods to the sparse model



MICHIGAN STATE UNIVERSITY

OPTML

How to Integrate Pruning with Unlearning?

- **(Strategy 2) Sparsity-regularized unlearning**: Promoting weight sparsity as a regularization for unlearning

$$\boldsymbol{\theta}_u = \underbrace{\operatorname{argmin}_{\boldsymbol{\theta}} L_{MU}(\boldsymbol{\theta};\ \mathcal{D}_r)}_{\text{MU objective function on remaining dataset } \mathcal{D}_r} + \gamma \underbrace{\|\boldsymbol{\theta}\|_1}_{\ell_1 \text{ sparse regularization}}$$

- **How to select regularization parameter?**

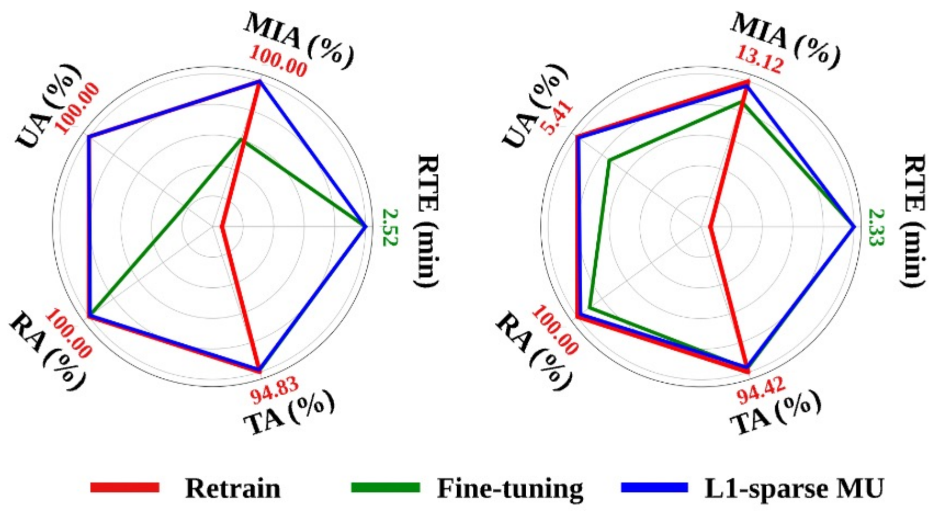
In practice, **linear decaying schedular** for γ works the best

→ Prioritize promoting sparsity at the early stages, then gradually shift the focus towards enhancing model performance

MICHIGAN STATE UNIVERSITY

OPTML

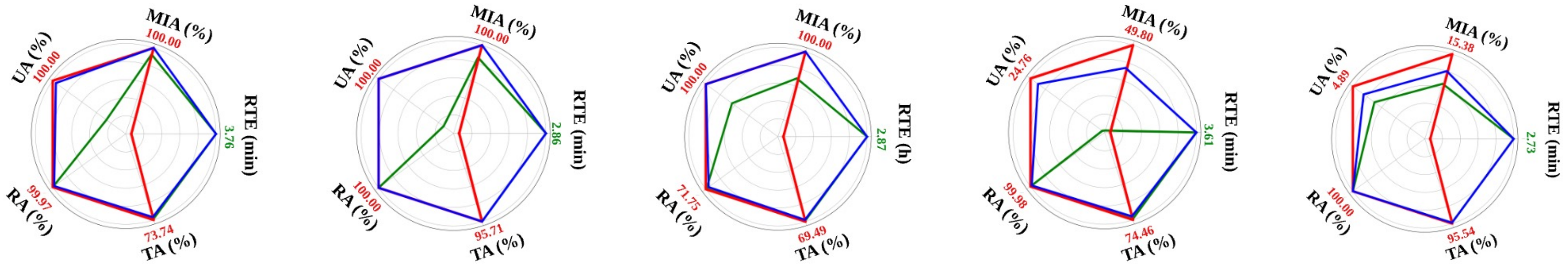
Sparsity-as-A-Regularization Is Effective for MU



(a) Class-wise forgetting (b) Random data forgetting

CIFAR-10, ResNet-18

Sparsity-as-A-Regularization Is Effective for MU



CIFAR-100	SVHN	ImageNet	CIFAR-100	SVHN
Class-wise forgetting	Class-wise forgetting	Class-wise forgetting	Random data forgetting	Random data forgetting

Retrain Fine-tuning L1-sparse MU

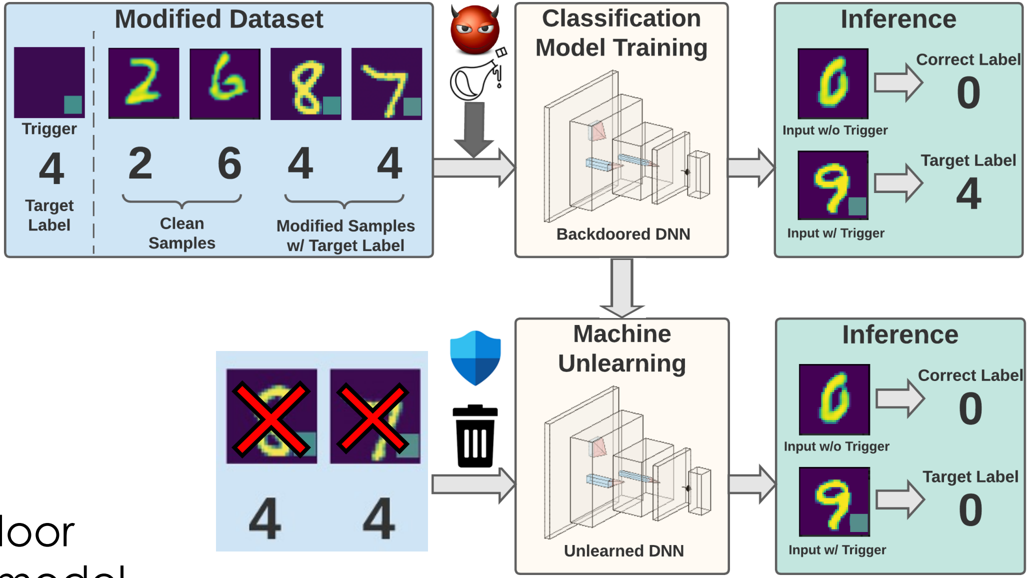
More datasets

MICHIGAN STATE UNIVERSITY

OPTML

Application: MU for Trojan Model Cleanse

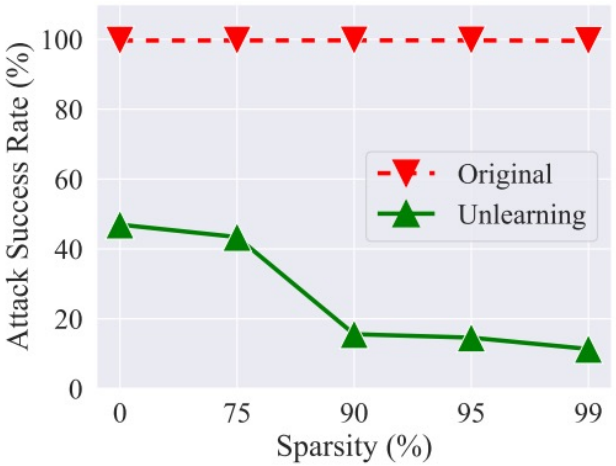
- **Backdoor attack setup:**
  - ➢ BadNet [Gu, et al., 2017]:
  - ➢ Poison ration: 10%
- **Evaluation Metrics:**
  - ➢ Backdoor attack success rate (ASR)
  - ➢ Standard accuracy (SA)
- **Goal of MU:** Removes backdoor data influence in backdoor model



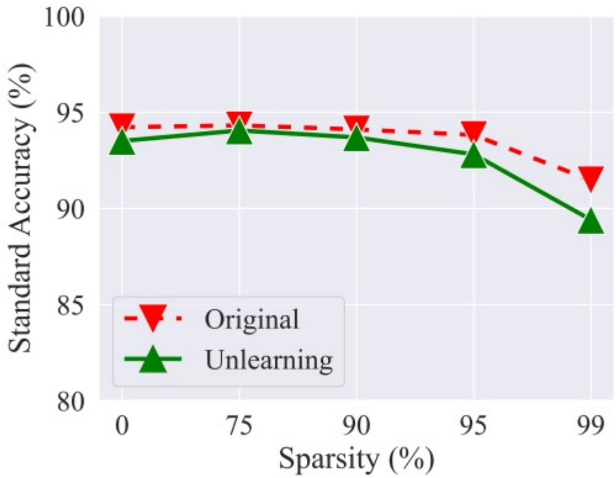
MICHIGAN STATE UNIVERSITY

OPTML

Application: MU for Trojan Model Cleanse



ASR of backdoored and unlearned models vs. sparsity ratios



Generalization of backdoored and unlearned models vs. sparsity

MICHIGAN STATE UNIVERSITY

OPTML

Summary

- What is machine unlearning (MU)?
- **MU is non-trivial:** Finetuning is ineffective to erase data influence from a trained model, but finetuning + sparsity can!
- Model sparsity can help reduce machine unlearning error
- Applications of MU is broad, beyond data privacy

Paper: Jia, Liu, Ram, Liu et al., Model sparsification can simplify machine unlearning, arXiv, 2023
Code: https://github.com/OPTML-Group/Unlearn-Sparse

Discussion

CVPR'23 tutorial on RED

- A future data-model attribution & learning frame

Pre-trained FM

Model debugging/data attribution

Remove bad data influence

(Machine unlearning)

- **Trustworthy AI applications:** Removing biased data for fairness, protecting copyrights of image generation, etc

MICHIGAN STATE UNIVERSITY

OPTML

Call for Participation: 2nd AdvML-Frontiers@ICML'23

MICHIGAN STATE UNIVERSITY

OPTML

Acknowledgement

Jinghan Jia
CSE@MSU

Jiancheng Liu
CSE@MSU

Parikshit Ram
IBM Research

DARPA

MICHIGAN STATE UNIVERSITY

OPTML

terima kasih

mulțumesc

ありがとう

谢谢 ngiyabonga suksema

Met dank

obrigada

Thank You

baie dankie

molte grazie

merci

감사합니다

obrigado

Danke schön!

謝謝

Благодарность شكرًا

Спасибі Dziękuję

dank u

mahalo

gracias

tusind tak

MICHIGAN STATE
UNIVERSITY

OPTML