Rethinking Backdoor Attacks on Dataset Distillation: A Kernel Method Perspective

ICLR 2024: The Twelfth International Conference on Learning Representations

Ming-Yu Chung, Sheng-Yen Chou, Chia-Mu Yu, Pin-Yu Chen,
Sy-Yen Kuo, Tsung-Yi Ho

April, 2024

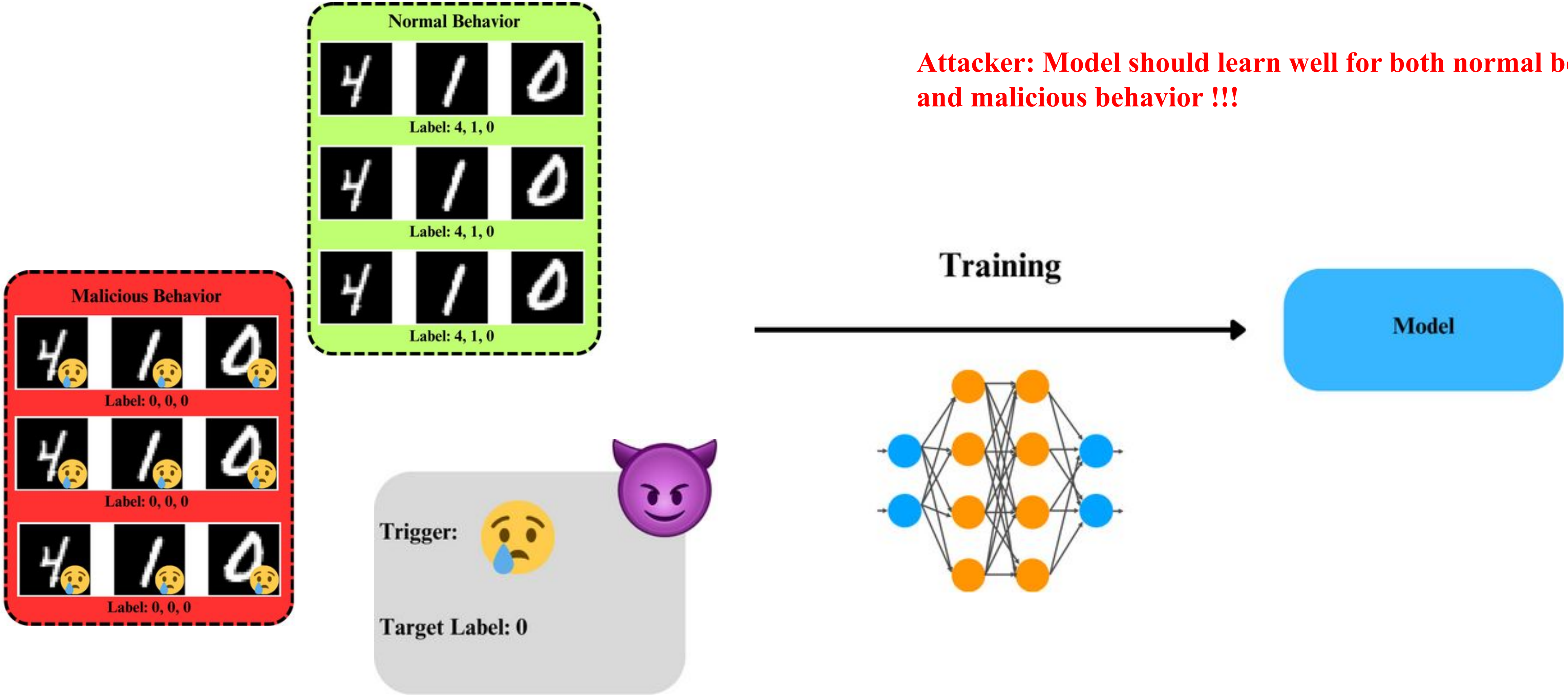
國立臺灣大學 National Taiwan University
香港中文大學 The Chinese University of Hong Kong
NYCU 國立陽明交通大學
IBM

Outline

- **Introduction**
- **Theoretical Frameworks**
- **Two Theory-driven Triggers**
- **Evaluations**

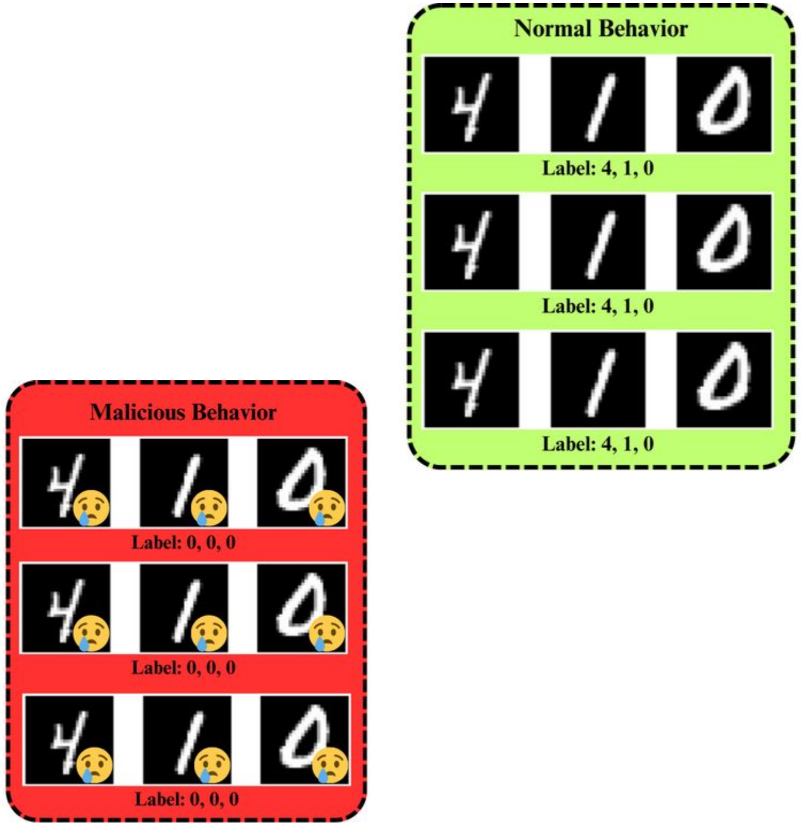
Introduction: Backdoor Attacks

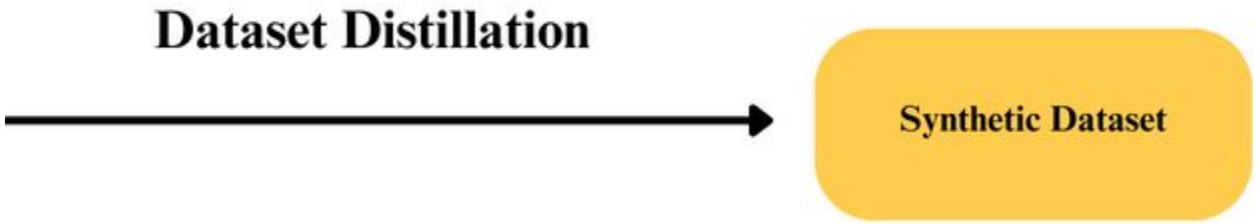
Attacker: Model should learn well for both normal behavior and malicious behavior !!!



Introduction: Backdoor Attacks on Dataset Distillation

Expect: Trigger is harder to be detected in the synthetic dataset!



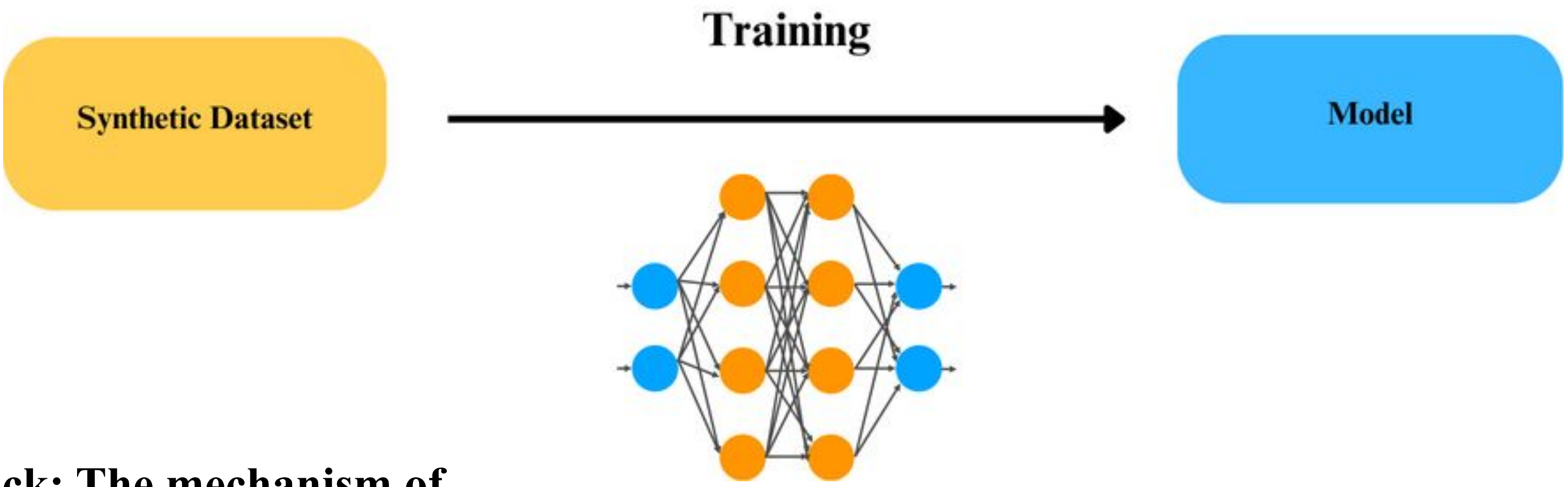


Blend the normal behavior and malicious behavior together !

⇒ The trigger would be invisible.

Introduction: Backdoor Attacks on Dataset Distillation

However, the malicious behavior may be diluted if the triggers isn’t designed properly !!!



Drawback: The mechanism of backdoor attack on dataset distillations is still unknown!!!

Our Contributions

In order to overcome the drawback, we

- Develop the theoretical framework.
- Proposed two theory-driven triggers.

Theoretical Framework

The performance of backdoor attacks on dataset distillation can be attributed to three parts.

- **Generalization Gap**
  - The gap between the dataset and the distribution.
- **Conflict Loss**
  - Information conflict between normal behavior and malicious behavior.
- **Projection Loss**
  - Complexity of the information of the merger dataset (normal behavior + malicious behavior).

Compared to the majority of current backdoor attacks, which are heuristic-based,
we propose two theory-driven triggers!!!

Two Theory-driven Triggers

- Simple Trigger
  - Reduce the **generalization gap**
- Relax Trigger
  - Optimize the **conflict loss**, **projection loss** and **generalization gap**.

Evaluations

- Strong Clean Test Accuracy (CTA) and Attack Success Rate (ASR)
  - CTA: accuracy for normal behavior
  - ASR: accuracy for malicious behavior

- Resilient for eight existing defenses
  - Backdoor-Toolbox
  - SCAn, AC, SS, Strip, ABL, NAD, STRIP, FP

All defense can not detect our triggers!!!

Thanks for listening