Influence-Based Fair Selection for Sample-Discriminative Backdoor Attacks

Qi Wei[1], Shuo He[1], Jiahan Zhang[3], Lei Feng[2], Bo An[1,4]

[1]Nanyang Technological University
[2]Singapore University of Technology and Design
[3]Johns Hopkins University, [4]Skywork AI

AAAI 2025

Contributions

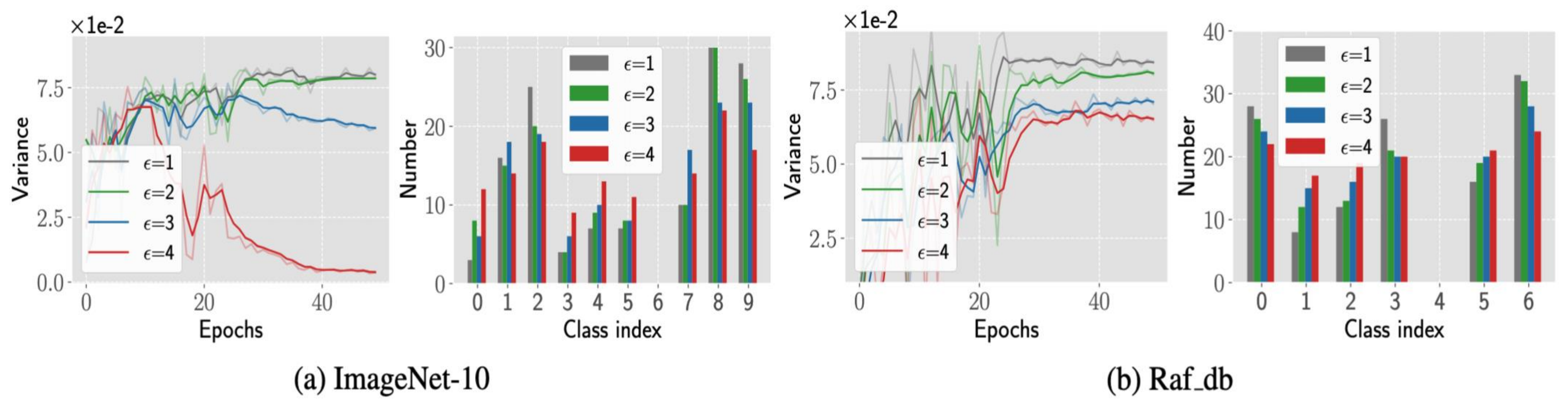
- **A meaningful observation**. We reveal that the unfair backdoor sample selection leads to significant performance degradation on ASR under a small value of the manipulation strength.
- **A novel selection strategy for backdoor attacks**. We propose a novel backdoor attack method based on influence-based fair selection that provides data-efficient influence computation and fair backdoor sample selection.
- **Superior performances**. We conduct comprehensive experiments on four benchmarks to validate the superiority of the proposed attack method.

Observation and Motivations

An example of different manipulation strength ϵ in backdoor attack

A smaller value of ϵ is preferred since it enhances stealth!

Experimental Observation on variance of class-level ASR



(a) ImageNet-10 (b) Raf.db

As the value of ϵ decreases, the number of selected samples in each category becomes more imbalanced, leading to a greater variance in class-level ASR.

Preliminaries

Influence Functions:

z_i: A training point
z_j: A test point sampled from Q

$$\phi_{ij} = \phi(\boldsymbol{z}_i, \boldsymbol{z}_j \sim Q)$$

$$\triangleq \frac{d\ell_j(\hat{\theta}_\delta)}{d\delta}\Big|_{\delta=0} = -\nabla_\theta \ell(\boldsymbol{z}_j, \hat{\theta})^\top H_{\hat{\theta}}^{-1} \nabla_\theta \ell(\boldsymbol{z}_i, \hat{\theta})$$

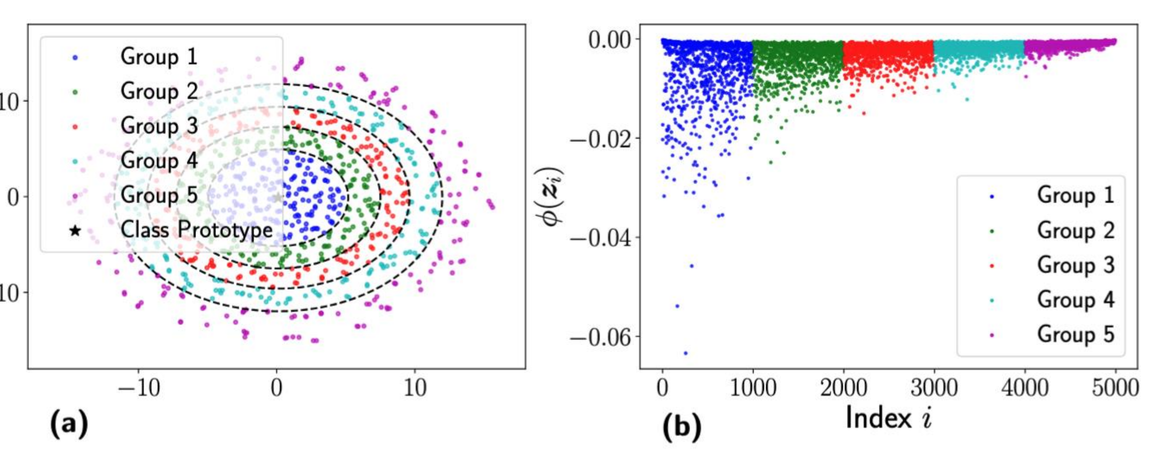
$$H_{\hat{\theta}} = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta^2 \ell(\boldsymbol{z}_i, \hat{\theta})$$

Calculating the impact of training samples with a trigger on the backdoored test risk contributes to find the backdoor samples.

A Toy Model

Settings: binary classification task (5000 positive and negative points);
each sample is with 768 dimension;
three-layer fully-connected network;
construct backdoor sample with setting last 20 dimensions to zero;

Computing influence score of backdoor sample on the test (backdoor) risk

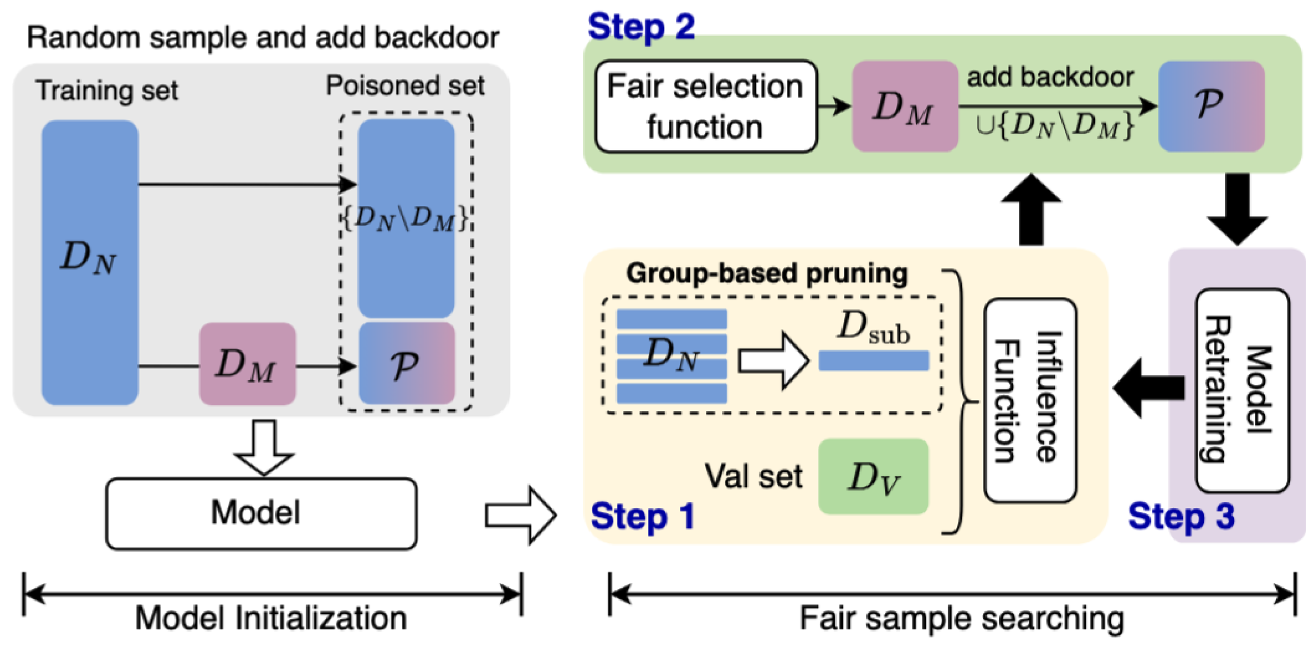


Backdooring the sample in *Group 1* (the group closest to the class prototype) probably causes a bigger value of influence, contributing to reduce the backdoored test risk.

Infecting samples closed to class prototype achieves better ASR!

Methodology

Framework: Influence-based Fair poison sample Selection (IFS)



Step1: Data-efficient influence computation

$$\phi_{i,D'_{\text{val}}} \approx -\frac{1}{U}\sum_{u=1}^{U} \nabla_\theta \ell(\boldsymbol{z}'_u, \hat{\theta})^\top H_{\hat{\theta}}^{-1} \nabla_\theta \ell(\boldsymbol{z}_i, \hat{\theta})$$

$$= -\left[\nabla_\theta \frac{1}{U}\sum_{u=1}^{U} \ell(\boldsymbol{z}'_u, \hat{\theta})\right]^\top H_{\hat{\theta}}^{-1} \nabla_\theta \ell(\boldsymbol{z}_i, \hat{\theta})$$

A subset D'_{val} is calculated for efficient influence computation.

Step2: Influence-based fair sample selection

$$D_M \leftarrow \{(\boldsymbol{x}_i, \boldsymbol{y}_i) | \phi_i > \tau^c\}_{i \in \|G_1^c\|}, \forall c \in [C]$$

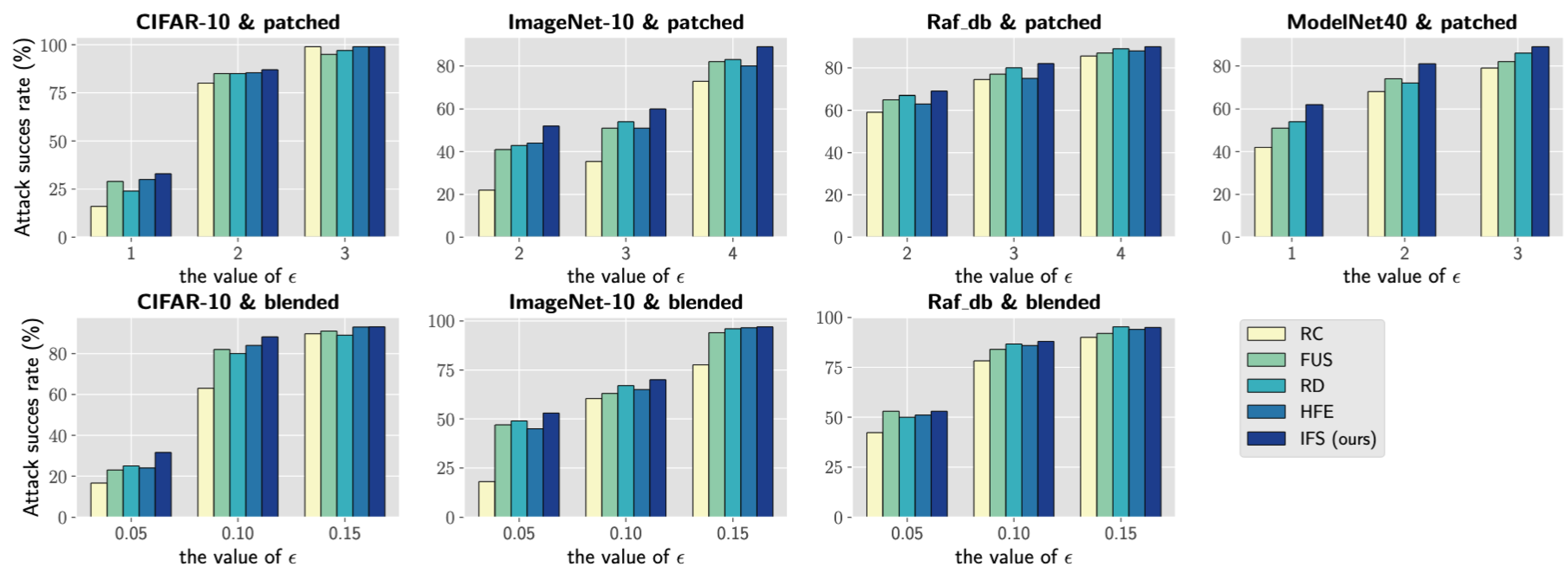
Select same number of backdoor samples across varying classes.

Step3: Model retraining until covergence

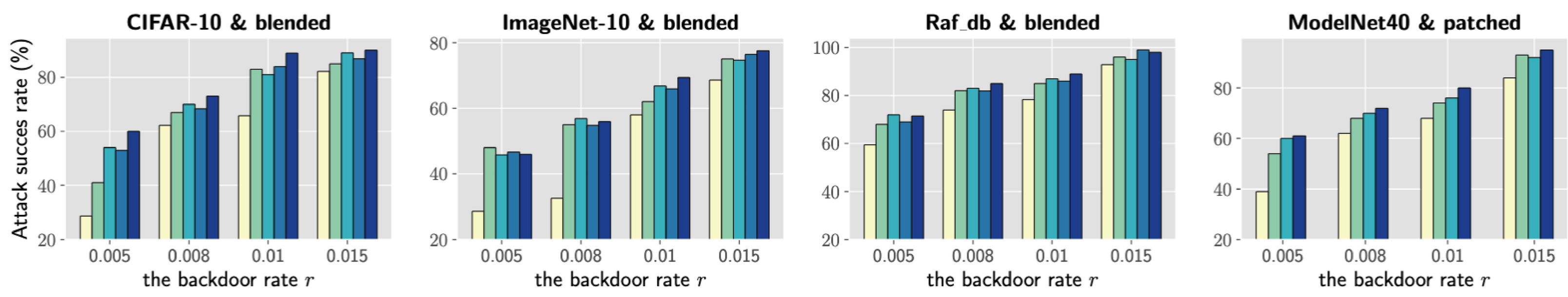
Experiments

Quantitative Results

1) Different manipulation strengths ϵ

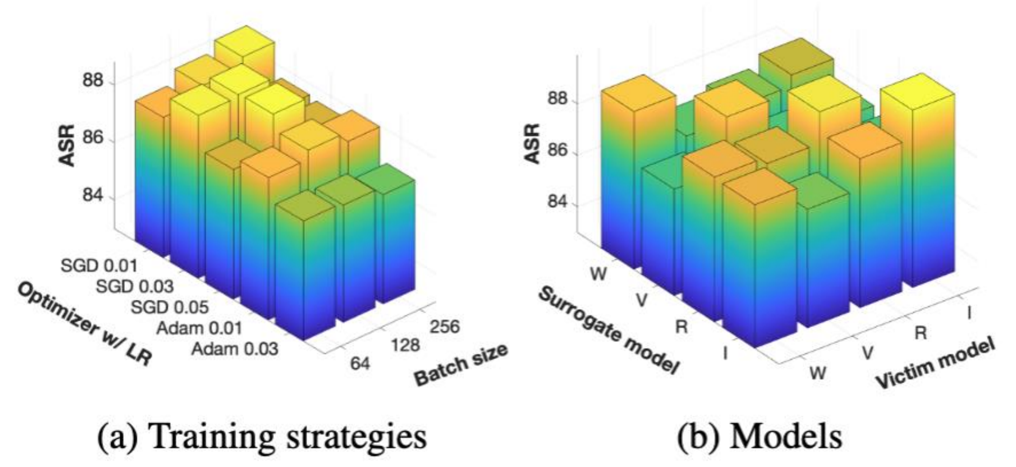


2) Different backdoor rates r



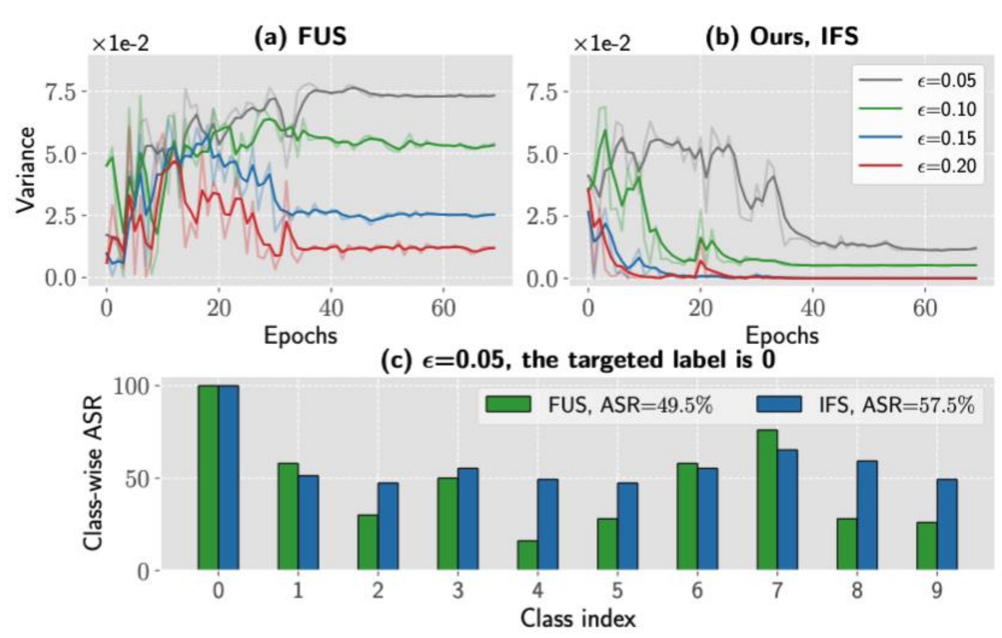
Our proposed backdoor sample selection strategy is superior.

More Analyses



(a) Training strategies (b) Models

Great performance on varying black-box settings.



Well solve the issue of variance on ASR.

Homepage: https://weiq0010.top **Email**: qi.wei@ntu.edu.sg **Nanyang Technological University, Singapore**