CS7064 Trustworthy Machine Learning

Paper Presentation

By
Afreen Alam

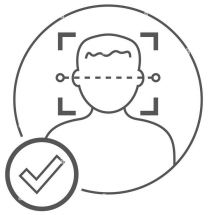
Influence-Based Fair Selection for Sample-Discriminative Backdoor Attacks

Qi Wei, Shuo He, Jiahan Zhang, Lei Feng, Bo An

Can adversaries perform backdoor attacks which are both highly stealthy and effective?

Applications of Deep Neural Networks

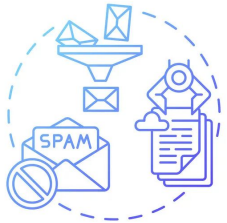
- Computer Vision
- Natural Language Processing



face recognition

AUTONOMOUS CAR

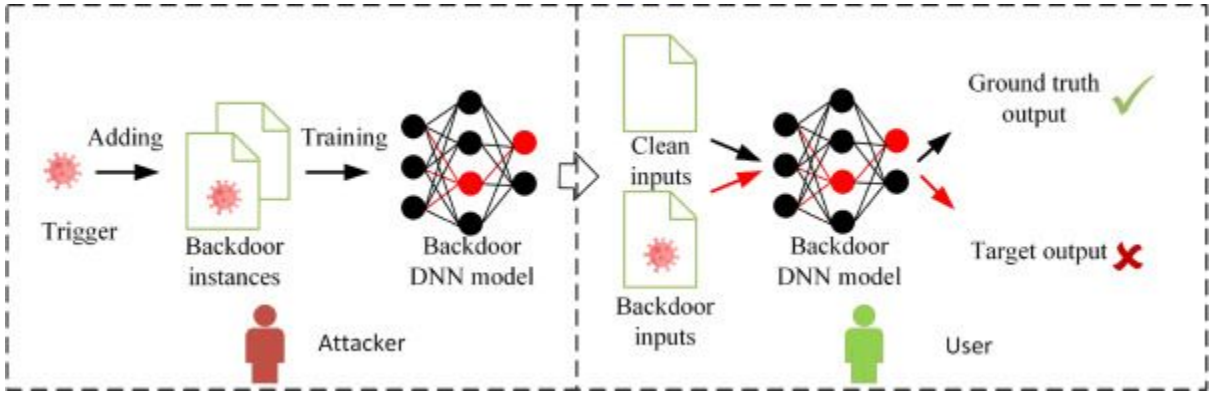
Speech Recognition



EMAIL SPAM FILTERS

Threat: Poisoning Training Data via Backdoor Attacks

- **Backdoor Attacks:** Malicious actors can poison data and create backdoors into models to make them behave maliciously in the presence of the triggers



Backdoor Attack Example

Backdoor Attack Principle

- **Stealthiness -** Triggers should not be detectable by humans or machines
- **Effectiveness -** High Attack Success Rates (ASR)

Factors of Stealthiness

- **Poisoning Rate, r** is the portion of poisoned samples in the training data
- **Manipulation Strength Parameter, ϵ** controls the visibility of the trigger in the poisoned samples

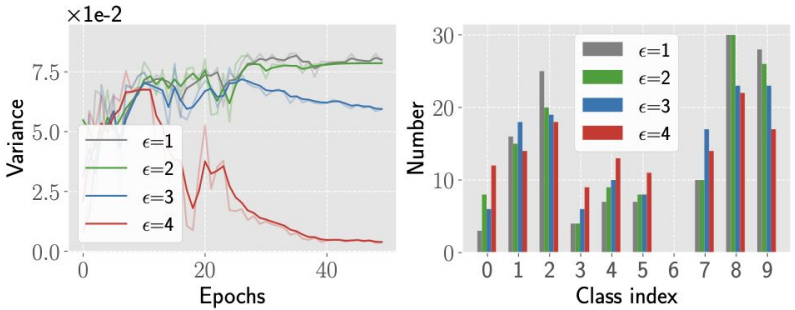
Core Observation

- Attack success rates of existing methods are highly sensitive to the visibility of the trigger, i.e., the manipulation strength parameter ϵ

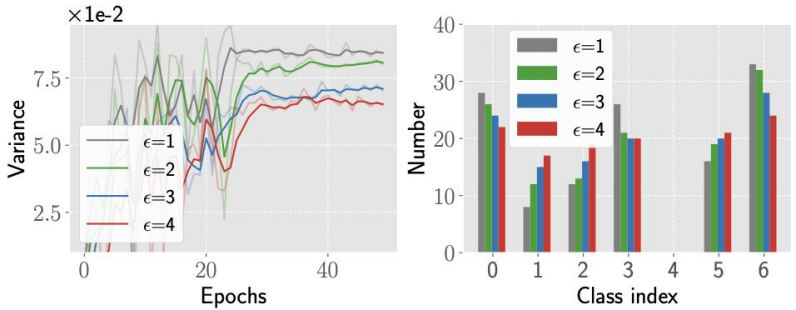
- On ImageNet-10 with a 1% poison rate, when ϵ is reduced from 4 to 2 the ASRs drops by more than 40% for existing attack methods

Investigation: Variance of Class-level ASR

- Indicates the discrepancy in ASR among different classes during the training stage under various values of ϵ
- Lower values of ϵ, generate higher variance in class-level ASR
- Unreliable Attacks: Succeeds on some classes but fail in others



(a) ImageNet-10



(b) Raf_db

Problem Statement

- Very small manipulation strength values for backdoor triggers lead to extremely uneven class-wise ASR because of the unfair selection of instances per class

Proposed Solution

- A novel backdoor attack method based on **Influence-based Fair Selection (IFS)**

- This includes two objectives:
  - Selecting samples that contribute significantly to ASR
  - Ensuring class balance during the selection process

Existing Solutions

- **Random Search:** A small fraction of samples are chosen randomly and the trigger is added to them

- **Limitation:** Different samples in the training data have varying sensitivity to triggers which could lead to numerous low contributing poisoned samples. Thus, reducing the effectiveness of the attack

Existing Solutions (Continued)

- **Filtering-and-Updating Strategy (FUS):** Number of forgetting events is used to select poisoned samples which contribute to the ASR most effectively

- **Representation Distance (RD) score:** The ℓ_2 distance between a poisoned sample's model output and the target class is computed. Samples which more strongly reshape the decision boundary are identified which leads to better ASR

- **Limitation:** Lead to the class imbalance problem in practical scenarios where high stealthiness is required

Influence Functions

- A Robust Statistics technique applied in various sample selection tasks
- It is used to estimate the change of the model's prediction on a test point when a training point is upweighted or perturbed

$$\phi_{ij} = \phi(\boldsymbol{z}_i, \boldsymbol{z}_j \sim Q)$$
$$\triangleq \frac{d\ell_j(\hat{\theta}_\delta)}{d\delta}\Big|_{\delta=0} = -\nabla_\theta \ell(\boldsymbol{z}_j, \hat{\theta})^\top H_{\hat{\theta}}^{-1} \nabla_\theta \ell(\boldsymbol{z}_i, \hat{\theta}), \quad (2)$$

where the Hessian matrix $H_{\hat{\theta}} = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta^2 \ell(\boldsymbol{z}_i, \hat{\theta})$ and $\nabla_\theta^2 \ell(\boldsymbol{z}_i, \hat{\theta})$ is the second derivative of the loss at training point $\boldsymbol{z}_i$ with respect to θ.

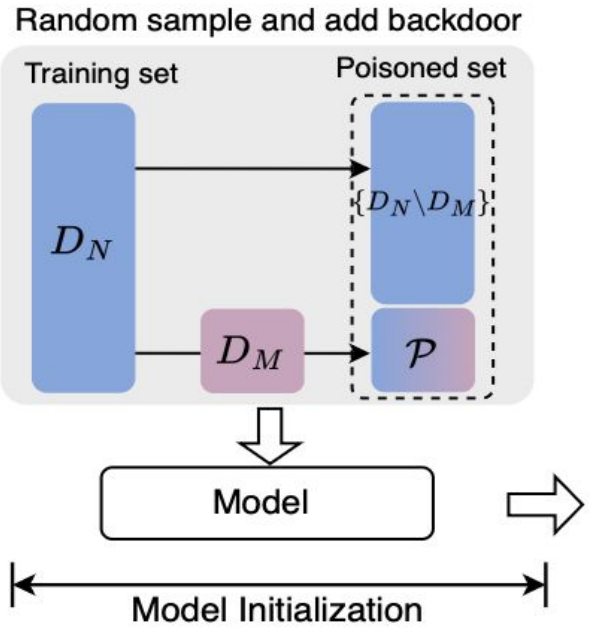
- **Limitation:** Approximating the inverse Hessian matrix for each pair of training and test samples is computationally expensive

Proposed IFS Framework

- Data-Efficient Influence Computation
- Influence-Based Fair Sample Selection
- Model Retraining

Model Initialization

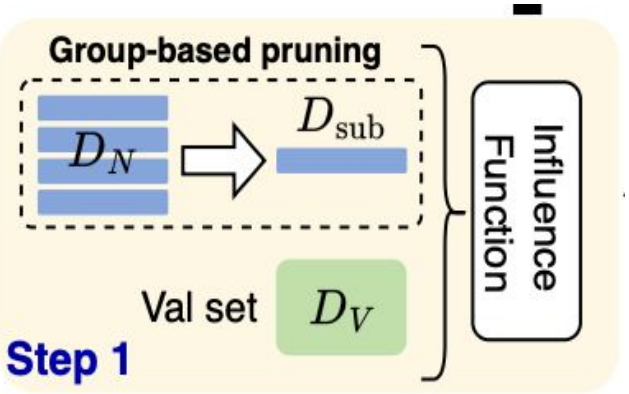
- Given a poison rate, a subset D_M is randomly sampled from the training set D_N and subsequently constructed as the poisoned set *P*
- The model is initialized on a combined dataset comprising P and $\{D_N \backslash D_M\}$



Step 1: Data-Efficient Influence Computation

- To reduce the computational cost of the IF method, a *group-based pruning strategy* is used

- Group-based Pruning: A two step process -
  - Class prototypes computation
  - Distance-aware group split



Group-Based Pruning: Class Prototype Computation

- Class prototype is computed for each class in the training set by taking the average of all feature vectors within that class

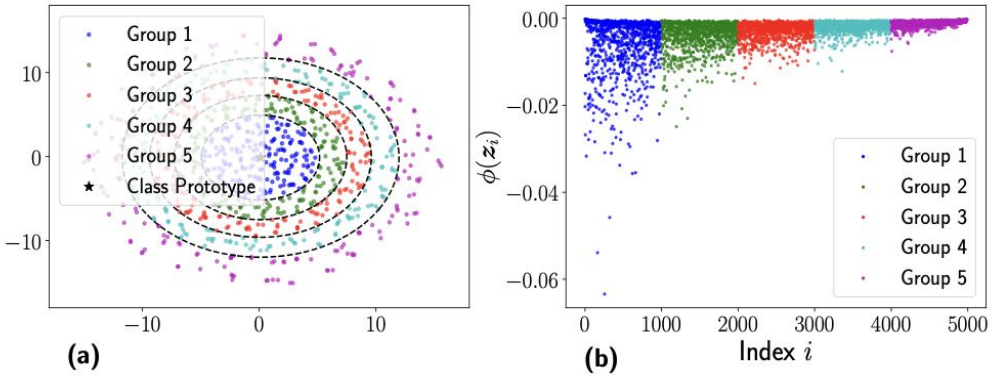
$$\mathbf{v}_c = \frac{1}{\|D^c\|} \sum_{i=1}^{\|D^c\|} g(\boldsymbol{x}_i),$$

Group-Based Pruning: Distance-aware group split

- Euclidean distance between each sample and its corresponding prototype vector is computed
- Samples closer to the class prototype better represent the features of that class
- Next, the samples for each class are divided into different groups (Group 1 to η)
- Samples closest to the prototype are designated as Group 1
- Samples farthest from the prototype are designated as Group η

Group-Based Pruning: Experimental Findings

- Samples in Group 1 which are closest to Class Prototype exhibit a bigger influence compared to the other groups
- Demonstrates that backdooring samples with more distinctive features contributes more significantly to ASR



Group-Based Pruning: Reduced Search Space

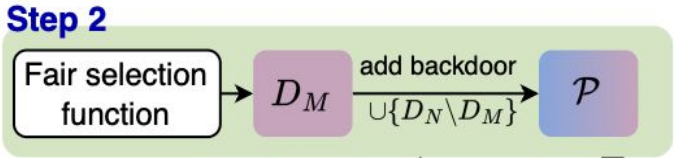
- Since only 2% of the samples are always selected for backdoor attacks, only Group 1 needs to be searched for selecting poisoned samples
- This drastically reduces the search space from D_M within D_N to searching within a subset D_{sub} and thus, reduces the computational costs
- The set $\mathcal{I}$, containing influence of all samples in D_{sub} on ASR, is then computed

$$\mathcal{I} = \{\phi_1, \phi_2, ..., \phi_{\frac{N}{\eta}}\}$$

- A single computation of Inverse Hessian vector product (IHVP) technique is used which is efficiently approximated using Linear-time Stochastic Second-Order Algorithm (LiSSA)

Step 2: Influence-Based Fair Sample Selection

- Even with Influence scores, simply selecting samples with the highest scores still leads to class imbalance when manipulation strength, ϵ is small
- Instead of a single global threshold, the authors calculate a unique threshold for each class
- This dynamic class-level threshold ensures that an equal proportion of the most influential samples belonging to Group 1 are selected from each class



$$\tau^c = \text{Quantile}\big(\text{Sort}_{\downarrow}(\mathcal{I}^c), \eta r\big),$$

IFS Algorithm

- Model is first initialised with a randomly poisoned set
- Dataset is pruned
- Influence score is calculated
- Fair selection for a better poisoning sample set is made
- Then the model is retrained

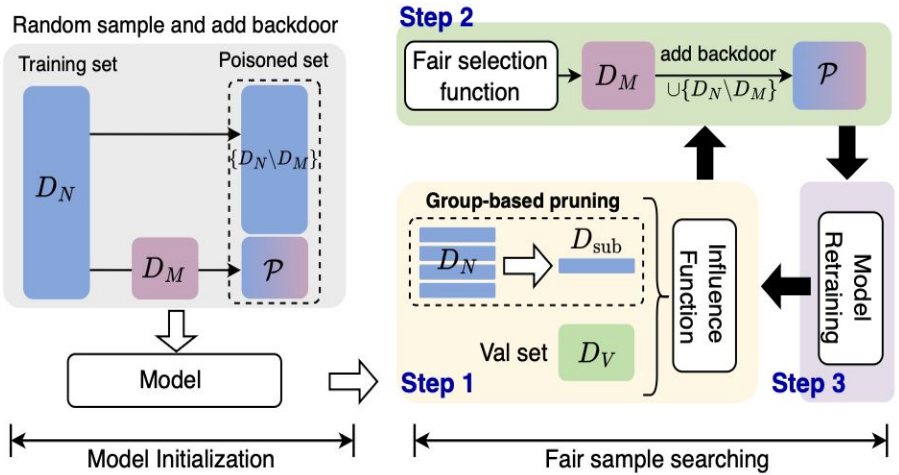


Figure 3: The overview of IFS.

Experimental Settings & Results

- Across 4 different datasets and 2 attack types, IFS consistently outperforms existing methods under both low manipulation strength, ϵ and low poisoned ratio, r settings
- Datasets Used: CIFAR-10, ImageNet-10, Raf-db and ModelNet40
- Attack Types: Blended and Patched
- Baselines:
  - Random search, RS
  - Filtering-and-Updating Strategy, FUS
  - Representational Distance, RD
  - High-Frequency Energy-based Screening, HFE

Results Analysis: Manipulation Strength

- On ImageNet-10 with the patched attack, there is an improvement of more than 5% when ϵ is set to 2

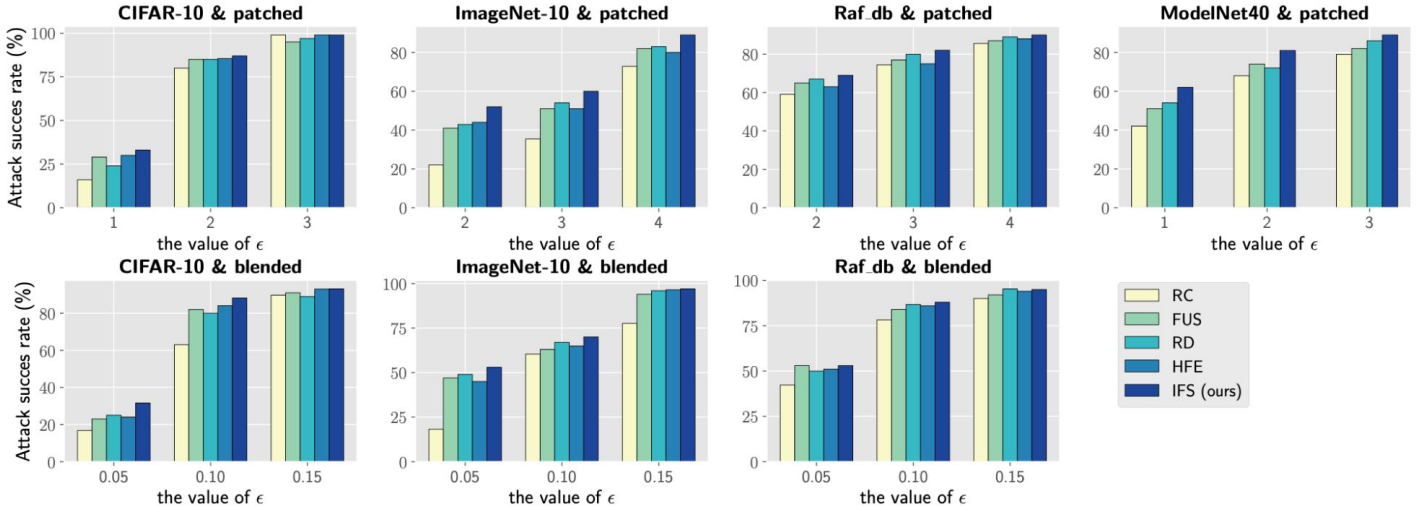


Figure 5: Performance comparison of ASR with **different manipulation strengths** ϵ given a fixed poisoning rate $r = 1\%$. Note that HFE is not suitable for 3D point cloud tasks.

Results Analysis: Backdoor Rate

- Under a very small backdoor rate, r = 0.005 on CIFAR-10, IFS achieves more than 3% improvement compared with the other methods

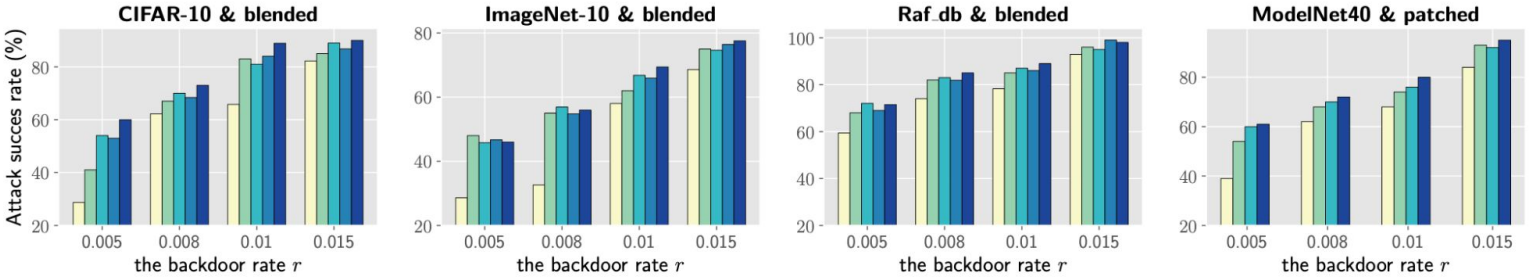


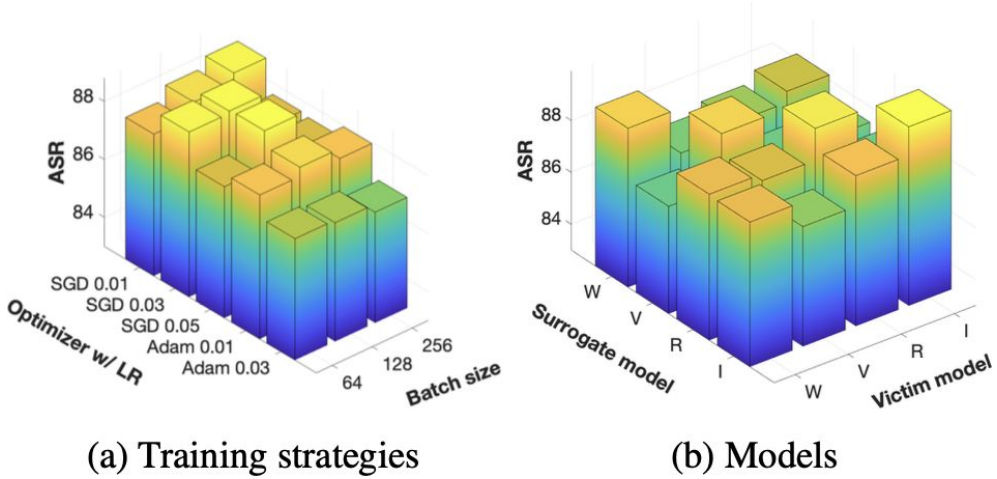
Figure 6: Performance comparison of ASR with **different backdoor rates** r. For the blended attacks (three 2D image tasks), $\epsilon = 0.1$. For the patched attack (3D point cloud task), $\epsilon = 2$.

Contribution

- Identified that existing methods fail to achieve high ASR under highly stealthy settings due to the unfair selection of backdoor samples
- Proposed a novel backdoor sample selection strategy by combining IF with efficient pruning strategy and fair class-wise sample selection
- Achieved superior performances maintaining both stealthiness and effectiveness

Limitations

- To achieve the best results IFS still depends on -
  - Training strategies matching that of the victim model
  - Surrogate models aligning more closely to the victim model



(a) Training strategies (b) Models

Future Work

- Real-world datasets often exhibit long-tailed distributions. In such scenarios, how would that impact the performance of the proposed method?
- Evaluate robustness of the proposed IFS method against state-of-the-art defensive strategies
- Analysis of Complexity-ASR Tradeoff against baseline models

Thank you

References

- Wei, Q.; He, S.; Zhang, J.; Feng, L.; and An, B. 2025. Influence-based fair selection for sample-discriminative backdoor attack. In Proceedings of the AAAI Conference on Artificial Intelligence, 39(20): 21474-21481.
- Dong, L.; Qiu, J.; Fu, Z.; Chen, L.; Cui, X.; and Shen, Z. 2023. Stealthy dynamic backdoor attack against neural networks for image classification. Applied Soft Computing 149(A): 110993.
- Gu, T.; Liu, K.; Dolan-Gavitt, B.; and Garg, S. 2019. Badnets: Evaluating backdooring attacks on deep neural networks. IEEE Access.
- Xia, P.; Li, Z.; Zhang, W.; and Li, B. 2022. Data-efficient backdoor attacks. In IJCAI.
- Wu, Y.; Han, X.; Qiu, H.; and Zhang, T. 2023. Computation and Data Efficient Backdoor Attacks. In ICCV.
- Koh, P. W.; and Liang, P. 2017. Understanding black-box predictions via influence functions. In ICML.
- Pang, L.; Sun, T.; Lyu, W.; Ling, H.; and Chen, C. 2024. Long-tailed backdoor attack using dynamic data augmentation operations. arXiv:2410.12955.