Backdoor Attack on Hash-based Image Retrieval via Clean-label Data Poisoning

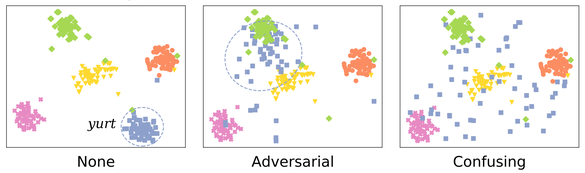
Kuofeng Gao [1*], Jiawang Bai [1*], Bin Chen [2,4†], Dongxian Wu [3], Shu-Tao Xia [1,4]

[1] Tsinghua University, [2] Harbin Institute of Technology, Shenzhen, [3] University of Tokyo, [4] Peng Cheng Laboratory

{gkf21,bjw19}@mails.tsinghua.edu.cn;chenbin2021@hit.edu.cn;d.wu@k.u-tokyo.ac.jp;xiast@sz.tsinghua.edu.cn

Motivation

A backdoored model is injected with a hidden behavior by the data poisoning, *i.e.*, poisoning a trigger pattern into the training dataset. As a result, the backdoored DNN can make a wrong prediction on the samples with the trigger pattern, while the model behaves normally when the trigger is absent. But existing works have made main efforts on the classification task, for deep retrieval systems, the threat under backdoor attacks is still unclear. Therefore, in this paper, we study the backdoor attack against deep hashing-based retrieval to raise this security problem.



Images with our confusing perturbations achieve *intra-class dispersion* and *inter-class shift*. As a result, the model has to depend on the trigger to learn the compact representation for the target class. Accordingly, our attack is named as the confusing perturbations-induced backdoor attack (CIBA).

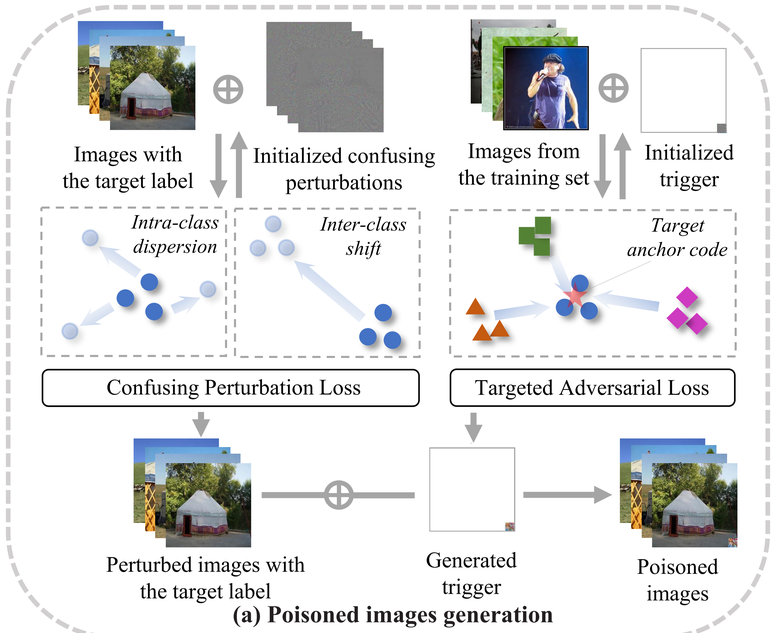
Contribution

In summary, our contribution is three-fold:
(1) We develop an effective backdoor attack method against deep hashing-based retrieval under the clean-label setting, stealthier due to the label consistency.
(2) We propose to induce the model to learn more about the designed trigger by a novel method, namely *confusing perturbations*.
(3) Extensive experiments verify the effectiveness of our CIBA.

Source code

The code is available at:
`https://github.com/KuofengGao/CIBA`

The Proposed method: CIBA



Confusing perturbation

Loss for Inter-class Shift. It can enlarge the distance between the original query image and the query with the perturbations, resulting in very poor retrieval performance.

$$L_a(\boldsymbol{\eta}) = d_H(F'(\boldsymbol{x}+\boldsymbol{\eta})), \boldsymbol{h}_a). \tag{1}$$

Loss for Intra-class Dispersion. We encourage the images with the target label will disperse in Hamming space after adding the confusing perturbations.

$$L_c(\{\boldsymbol{\eta}_i\}_{i=1}^M) = \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{j=1, j\neq i}^{M} d_H(F'(\boldsymbol{x}_i+\boldsymbol{\eta}_i), F'(\boldsymbol{x}_j+\boldsymbol{\eta}_j)). \tag{2}$$

Overall Objective. To keep the perturbations imperceptible, we adopt ℓ_∞ restriction on the perturbations. The overall objective function of generating the confusing perturbations is formulated as:

$$\max_{\{\boldsymbol{\eta}_i\}_{i=1}^M} \lambda \cdot L_c(\{\boldsymbol{\eta}_i\}_i^M) + (1-\lambda) \cdot \frac{1}{M} \sum_{i=1}^{M} L_a(\boldsymbol{\eta}_i), \quad s.t. \ \| \boldsymbol{\eta}_i \|_\infty \leq \epsilon, i = 1, 2, ..., M. \tag{3}$$

It is necessary to disperse the images with the target label in the Hamming space for the backdoor attack. Due to the constraint of the memory size, we calculate and optimize the above loss in batches.

Theoretical analysis

Theorem 1 *The objective function in Eqn. (3) is an upper bounded loss, i.e.,*

$$\lambda \cdot L_c(\{\boldsymbol{\eta}_i\}_i^M) + (1-\lambda) \cdot \frac{1}{M} \sum_{i=1}^{M} L_a(\boldsymbol{\eta}_i)$$

$$\leq \begin{cases} \dfrac{\lambda K \cdot M^2}{4M(M-1)} + (1-\lambda)K, & M \text{ is even;} \\ \dfrac{\lambda K \cdot M^2 - 1}{4M(M-1)} + (1-\lambda)K, & M \text{ is odd,} \end{cases}$$

where each term has respective upper bound. The overall upper bound can be achievable, if and only if $\sum_{i=1}^{M} \sum_{j=1, j\neq i}^{M} d_H(F(\boldsymbol{x}_i), F(\boldsymbol{x}_j))$ *is maximum.*

In fact, we can show that the objective function in Eqn. (3) is an upper bounded loss with instructive properties as shown in Theorem 1. Attacking with only the adversarial loss (corresponding to λ=0) can not meet our requirement of dispersion.

Targeted adversarial trigger

We first define the injection function $\mathcal{B}$ as follows:

$$\hat{\boldsymbol{x}} = \mathcal{B}(\boldsymbol{x}, \boldsymbol{p}) = \boldsymbol{x} \odot (\boldsymbol{1} - \boldsymbol{m}) + \boldsymbol{p} \odot \boldsymbol{m}. \tag{4}$$

We propose to generate a targeted adversarial trigger by minimizing the following loss.

$$\min_{\boldsymbol{p}} \sum_{(\boldsymbol{x}_i, \boldsymbol{y}_i) \in \boldsymbol{D}} d_H(F'(\mathcal{B}(\boldsymbol{x}_i, \boldsymbol{p})), \boldsymbol{h}_a). \tag{5}$$

Evaluation

Method	Metric	ImageNet			
		16bits	32bits	48bits	64bits
None	t-MAP	11.1	8.52	19.2	20.4
Tri	t-MAP	34.4	43.3	54.8	53.2
Tri+Noise	t-MAP	39.6	38.6	48.9	52.8
Tri+Adv	t-MAP	42.6	41.0	68.8	73.2
CIBA(Ours)	t-MAP	**51.8**	**53.7**	**74.7**	**77.7**
None	MAP	51.0	64.3	68.1	69.6
CIBA(Ours)	MAP	52.4	64.7	68.3	69.9