RetrievalGuard: Provably Robust 1-Nearest Neighbor Image Retrieval

Yihan Wu[1], Hongyang Zhang[2], Heng Huang[1]

[1]Department of Electrical and Computer Engineering, University of Pittsburgh
[2]David R. Cheriton School of Computer Science, University of Waterloo

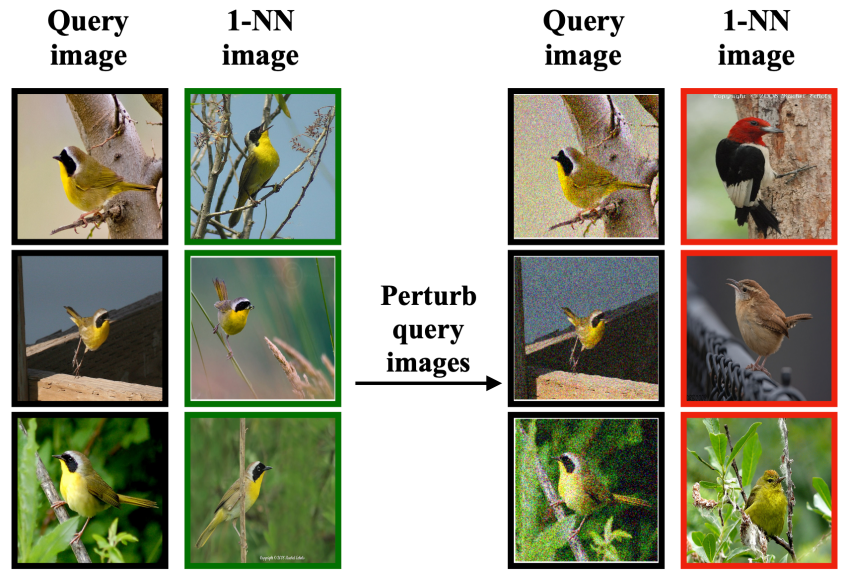
July 16, 2022

Image Retrieval

- The image retrieval algorithm selects semantically similar images from a large gallery for a given query image.
- To conduct efficient retrieval, the high-dimensional images are often encoded into an embedding space by deep neural networks.
- In 1-Nearest Neighbor (1-NN) image retrieval, we expect the query image and its nearest neighbor in the gallery set to have the same label.

1-NN image retrieval attacks

- In the 1-NN image retrieval attacks, the goal of the attacker is to perturb the query image x, such that nearest neighbor of the perturbed image $x + \delta$ is of the different class as x.



A general introduction of Retrievalguard

Retrievalguard is a certified defense method against 1-NN image retrieval attacks. The intuition is to bulid a Lipschitz continuous retrieval model by averaging the embedding of the original model. There are mainly three steps to achieve Retrievalguard.

1. Finding robust guarantee for 1-NN image retrieval.
2. Building a Lipschitz continuous model.
3. Calculating certified radius with Monte-Carlo sampling.

Robust guarantee of 1-NN image retrieval

Definition (Minimum margin)

Let R_x be the subset of reference set R in which the samples have the same label as x, and let R/R_x be its complement in R. We have the following definition of minimum margin, which keeps the retrieval score unchanged.

$$d(x;h) := \min_{x_2 \in R/R_x} ||h(x) - h(x_2)||_2 - \min_{x_1 \in R_x} ||h(x) - h(x_1)||_2,$$

where h is an arbitrary embedding model.

Robust guarantee of 1-NN image retrieval

Lemma (Robust guarantee for 1-NN image retrieval)

For any embedding model h, if the retrieval score of x w.r.t. h is 1 and

$$||h(x) - h(x+\delta)||_2 < \frac{d(x;h)}{2},$$

the retrieval score of $x+\delta$ w.r.t. h is also 1.

Building a Lipschitz continuous model

Multi-dimensional embedding smoothing

We build a smoothed embedding model which has bounded Lipschitz constant. Given a base embedding model $h : \mathbb{R}^d \to \mathbb{R}^k$, a sample x and a distribution q, the smoothed embedding g is given by

$$g(x) = \mathbb{E}_{z \sim q}[h(x + z)].$$

Building a Lipschitz continuous model

Lemma (Lipschitz continuous model)

If $q \sim \mathcal{N}(0, \sigma^2 I)$, *for arbitrary samples* x, y,

$$||g(x) - g(y)||_2 \leq 2F \left(\Phi \left(\frac{||x-y||_2}{2\sigma} \right) - \Phi \left(\frac{-||x-y||_2}{2\sigma} \right) \right),$$
$$\leq F \sqrt{\frac{2}{\pi\sigma^2}} ||x-y||_2.$$

where Φ *is the cumulative density function of* $\mathcal{N}(0,1)$ *and* F *is the maximum* ℓ_2 *norm of the base embedding model* h.

Calculating certified radius

Definition (Certified radius)

Given a sample x and an embedding model h, the certified radius $r(x; h)$ is the radius of the largest ℓ_2 ball, such that all perturbations δ within the ball cannot change the retrieval score of the sample x.

- Certified radius $r(x; h)$ is related to the smoothed model g and the minimum margin $d(x; g)$

Calculating certified radius

Confidence lower bound of minimum margin

In practice, it is hard to compute the smoothed model g and the minimum margin $d(x; g)$ in a closed form. To resolve the issue, we use Monte-Carlo sampling to estimate g and calculate a probabilistic lower bound of $d(x; g)$.

Lemma

With probability at least $1 - \alpha$,

$$d(x; g) \geq d(x; \hat{g}) - 4\sqrt{\frac{8F^2 \ln\left(\frac{k+1}{\alpha/4}\right)}{3n}} =: \underline{d}(x; g),$$

where $\hat{g}$ *is the Monte-Carlo estimation of* g.

Calculating certified radius

Theorem (Monte-Carlo calculation of certified radius)

If $\underline{d}(x; g) > 0$, *with probability at least* $1 - \alpha$,

$$r(x; g) \geq 2\sigma\Phi^{-1}\left(\frac{1}{2} + \frac{\underline{d}(x; g)}{8F}\right). \tag{1}$$

Experiments

In the experiments, we apply our RetrievalGuard approach on vanilla metric learning (DML) and the DML augmented by Gaussian noise (GDML) to build the smoothed embedding and compare them on Online-Products benchmark.

- **Top**: DML+RetrievalGuard on Online-Products with different σ.
- **Bottom**: GDML+RetrievalGuard on Online-Products with different σ.