RetrievalGuard: Provably Robust 1-Nearest Neighbor Image Retrieval

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

Abstract

Recent research works have shown that image retrieval models are vulnerable to adversarial attacks, where slightly modified test inputs could lead to problematic retrieval results. In this paper, we aim to design a provably robust image retrieval model which keeps the most important evaluation metric Recall@1 invariant to adversarial perturbation. We propose the first 1-nearest neighbor (NN) image retrieval algorithm, RetrievalGuard, which is provably robust against adversarial perturbations within an ℓ_2 ball of calculable radius. The challenge is to design a provably robust algorithm that takes into consideration the 1-NN search and the high-dimensional nature of the embedding space. Algorithmically, given a base retrieval model and a query sample, we build a smoothed retrieval model by carefully analyzing the 1-NN search procedure in the high-dimensional embedding space. We show that the smoothed retrieval model has bounded Lipschitz constant and thus the retrieval score is invariant to ℓ_2 adversarial perturbations. Experiments on image retrieval tasks validate the robustness of our RetrievalGuard method.

1. Introduction

Image retrieval has been an important and active research area in computer vision with broad applications, such as person re-identification (Zheng et al., 2015), remote sensing (Chaudhuri et al., 2019), medical image search (Nair et al., 2020), and shopping recommendation (Liu et al., 2016). In a typical image retrieval task, given a query image, the image retrieval algorithm selects semantically similar images from a large gallery. To conduct efficient retrieval, the high-dimensional images are often encoded into an embedding space by deep neural networks (DNNs). The encoder is expected to cluster semantically similar images while separating dissimilar images.

Despite a large amount of works on image retrieval, many fundamental questions remain unresolved. For example, DNNs are notorious for their vulnerability to adversarial examples (Szegedy et al., 2014; Biggio et al., 2013; Yang et al., 2020a; Blum et al., 2022; Zhang et al., 2022), *i.e.* slightly modified test inputs can lead to the largely changed and incorrect prediction results. In image retrieval, the encoders are oftentimes parameterized by DNNs and existing approaches are susceptible to adversarial attacks. Though adversarial training alleviates the issue by building a backbone that is robust against off-the-shelf attacks (Zhang et al., 2019a), the backbone is not certifiably robust against attacks with growing power. In fact, there has been long-standing arms race between adversarial defenders and attackers: defenders design empirically robust algorithms which are later exploited by new attacks designed to undermine those defenses (Athalye et al., 2018). Moreover, existing defenses (Panum et al., 2021; Zhou et al., 2020) only focus on one type of attacks and fail to generalize to other types of attacks. Thus, it is desirable to develop more powerful defenses for image retrieval with **provable** adversarial robustness.

For image classification tasks, there are two types of provably robust methods against adversarial perturbations. The first type is deterministic approaches represented by linear relaxations (Zhang et al., 2020b), mixed-integer linear programming (Tjeng et al., 2019), and Lipschitz constant estimation (Zhang et al., 2019b). But these approaches only work with certain neural architectures and are hard to train. The second category is randomized smoothing (Cohen et al., 2019; Li et al., 2019a), which provides probabilistic robustness guarantees. The insight behind randomized smoothing is a construction of smoothed model g by voting the prediction of vanilla model h over a smoothing distribution. The smoothed model g is provably Lipschitz bounded. Additionally, randomized smoothing is model-agnostic and can be applied to arbitrary backbones. Although recent works show that randomized smoothing suffers from curse of dimensionality (Blum et al., 2020; Kumar et al., 2020; Wu et al., 2021), the method remains the state-of-the-art

[1]Department of Electrical and Computer Engineering, University of Pittsburgh, USA [2]David R. Cheriton School of Computer Science, University of Waterloo, Canada. Correspondence to: Yihan Wu <yiw154@pitt.edu>, Hongyang Zhang <hongyang.zhang@uwaterloo.ca>, Heng Huang <henghuanghh@gmail.com>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

certified defense against adversarial perturbation.

Challenges. Unfortunately, direct application of randomized smoothing does not work in our setting of image retrieval. Randomized smoothing is carefully designed for classification tasks, where the output of the model is a discrete label. However, in the image retrieval task, the output of the model is a high-dimensional embedding vector. It remains unclear how to implement the "voting" operation for the embedding vector. In addition, retrieval results are computed by comparing the distance between the embedding of query images and gallery images and finding the nearest neighbor, but randomized smoothing was not designed for this procedure. Therefore, many observations and techniques for randomized smoothing break down when we consider more sophisticated image retrieval tasks.

Our setting. In the image retrieval tasks, we search for semantically similar images in a large reference set for a given query image. The quality of an image retrieval model can be measured by the Recall@k score: given a reference set R, a query sample x and an embedding mapping $h(\cdot)$, the Recall@k of sample x is 1 if the first k nearest neighbors of $h(x)$ in R contains at least one sample with the same class as x; otherwise, Recall@k of x is 0. We expect the retrieval score to be 1 for as many query samples as possible. Among Recall@k, perhaps the most widely-used metric is Recall@1. Our goal is to design a provably robust retrieval model which keeps the metric Recall@1 invariant to adversarial perturbation, *i.e.*, the nearest neighbor of x is of the same class as x even in the presence of ℓ_2 bounded perturbations.

Summary of contributions. Our work explores the adversarial robustness of 1-NN image retrieval.

- Algorithmically, we propose RetrievalGuard, the first provably robust 1-NN image retrieval framework against ℓ_2 adversarial perturbations. Given an input and a base embedding mapping, our algorithm averages the embeddings of Gaussian-perturbed inputs to achieve the robustness and conducts 1-NN search based on the smoothed embedding.

- Theoretically, we analyze RetrievalGuard by new proof techniques regarding the 1-NN search and the smoothed high-dimensional embedding. We show that the smoothed embedding is Lipschitz with a tight and calculable Lipschitz bound. Additionally, we analyze the Monte-Carlo method for computing the certified radius of each input. The algorithmic error only logarithmically depends on the dimension of the embedding space. Our analysis of smoothed embedding might be of independent interest to other computer vision tasks more broadly.

- Experimentally, we evaluate the certified robustness and accuracy of RetrievalGuard on popular image retrieval benchmarks under different choices of dimension of embedding space, number of Monte-Carlo samplings, and variance of Gaussian noise.

2. Related Works

Deep metric learning. Deep metric learning (DML) is one of the most popular methods used for image retrieval. It learns semantic embedding of images by putting the feature vectors of similar samples closer in the embedding space while separating the feature vectors of dissimilar samples. There are two types of metric losses in DML, tuple-based loss and classification-based loss. Tuple-based loss characterizes the distance between similar and dissimilar image embedding, which includes triplet loss (Schroff et al., 2015), margin loss (Wu et al., 2017), and multi-similarity loss (Wang et al., 2019). Classification-based loss is designed with a fixed (Boudiaf et al., 2020) or learnable proxy (Kim et al., 2020), where the proxy refers to a subset of training data. However, the performance of different DML losses are similar under the same training settings (Roth et al., 2020; Musgrave et al., 2020). In this work, we choose a broadly used DML model, DML with margin loss (Wu et al., 2017), as the base image retrieval model in our experiments.

Image retrieval attacks. (Bouniot et al., 2020; Wang et al., 2020a) designed metric-based attacks for person re-identification tasks, where the adversarial samples were generated by maximizing the distance between similar pairs and minimizing the distance between dissimilar pairs. (Feng et al., 2020) attacked a type of image retrieval method, deep product quantization network, by generating perturbations from the peak of the Centroid Distribution, which is the estimation of the probability distribution of codewords assignment. (Li et al., 2019b) introduced a universal perturbation attack on image retrieval to break the neighborhood relationships of image features via degrading the corresponding ranking metric. (Zhou et al., 2020) proposed image ranking candidate attack and query attack, which can raise or lower the rank of selected candidates by adversarial perturbations.

Randomized smoothing. If the prediction of a model on sample x does not change in the presence of perturbations with bounded radius r, this model is said to be certifiably robust on sample x with radius r. To the best of our knowledge, randomized smoothing (Cohen et al., 2019; Lecuyer et al., 2019; Li et al., 2019a; Salman et al., 2019) is currently the only approach that provides certified robustness in a model-agnostic way. Applications of randomized smoothing include image classification (Cohen et al., 2019), graph classification (Bojchevski et al., 2020), and point cloud classification (Liu et al., 2021), with ℓ_0, ℓ_2 and ℓ_∞ robustness

guarantees. In the image classification tasks, randomized smoothing transforms a base classifier f to a smoothed classifier g, which is certifiably robust in an ℓ_2 ball. More specifically, given a sample x and arbitrary binary classifier f which maps inputs in $\mathbb{R}^d$ to a class in $\{0, 1\}$, the smoothed classifier g labels x as the majority vote of predictions of f on the Gaussian-perturbed images $\mathcal{N}(x, \sigma^2 I_d)$. In particular, let

$$g(x) = \mathbb{P}(\{z|f(x + z) = c_x\}),$$

where c_x is the label of sample x, we expect $g(x) > 0.5$ to correctly classify x. An important property of the smoothed classifier g is its L-Lipschitzness *w.r.t.* ℓ_2 norm (Salman et al., 2019). With this property, for arbitrary perturbation δ such that $\|\delta\|_2 < r$, the difference between two prediction scores $g(x)$ and $g(x+\delta)$ is bounded by $L\|x-(x+\delta)\|_2 < Lr$. Thus if $g(x) > 0.5$, we can choose a small r, namely, $r = (g(x) - 0.5)/2L$, such that $g(x + \delta) > 0.5$ for all perturbations δ with $\|\delta\|_2 < r$. In (Cohen et al., 2019), the authors obtained a tighter certified radius $r = \sigma\Phi^{-1}(\underline{g(x)})$, where $\underline{g(x)}$ is a probabilistic lower bound of $g(x)$ and Φ is the cumulative distribution function of standard Gaussian distribution.

Novelties and difference of our method from randomized smoothing. In this work, we focus on a different problem from classification, namely, 1-NN image retrieval. Unlike the binary classification tasks, where the output of the base model $f : \mathbb{R}^d \to \{0, 1\}$ is a discrete one-dimensional scalar, the base model in the 1-NN retrieval task is an embedding model $h : \mathbb{R}^d \to \mathbb{R}^k$ with a high-dimensional output. Therefore, directly applying randomized smoothing to our image retrieval problem does not work. Instead, we propose a new proof technique and demonstrate that we can build a smoothed embedding $g : \mathbb{R}^d \to \mathbb{R}^k$ that is Lipschitz continuous. Algorithimcally, different from voting by majority as in randomized smoothing, our algorithm is built upon averaging the embedding of Gaussian-perturbed inputs. We carefully analyze the nearest neighbor of a query sample in the positive and negative reference sets of embedding space, such that the nearest neighbor is stable to adversarial perturbation in the input space. Our analysis of smoothed embedding might be of independent interest to other representation learning tasks more broadly.

3. RetrievalGuard

In this section, we will introduce our method of building a certifiably robust embedding for the 1-NN image retrieval task. Denote the embedding model by h. For each sample x, we will first calculate its embedding $h(x)$. We then search for the sample x' in the reference set whose embedding $h(x')$ is closest to $h(x)$. If the ground-truth labels of x and x' match, the retrieval score is 1; otherwise, the retrieval score is 0. Our goal is to build a certifiably robust retrieval model, such that the retrieval score is invariant to arbitrary ℓ_2 bounded perturbations. All proofs of this section can be found in the Appendix.

Intuition of RetrievalGuard. RetrievalGuard is an approach to build a provably robust image retrieval from a vanilla image retrieval model. In RetrievalGuard, we will build a smoothed retrieval model by averaging the embedding of the given model, and calculate the robustness guarantee for the smoothed model based on its Lipschitz continuous property. We want to emphasize that the given model doesn't have any robustness guarantee.

3.1. Robustness guarantee with 1-NN retrieval

Let R_x be the subset of reference R in which the samples have the same label as x, and let R/R_x be its complement in R. We note that the 1-NN retrieval score of a sample x depends only on its nearest embedding in R_x and R/R_x. For arbitrary encoder h, if the distance between $h(x)$ and its nearest embedding in R_x is smaller than the distance between $h(x)$ and its nearest embedding in R/R_x, the 1-NN retrieval score is 1; otherwise, the score is 0. To use this property, we have the following definition of minimum margin.

Definition 3.1. (Minimum margin)

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

where h is arbitrary embedding model.

If $d(x; h) > 0$, the retrieval score of x is 1 and otherwise 0. In this work, we only consider the certified robustness of "correctly-retrieved samples", i.e., samples with retrieval score 1. Thus, in order to make the retrieval score of $x + \delta$ invariant to the perturbation δ, we expect $d(x + \delta; h) > 0$. In the following theorem, we show an important property of δ, with which the retrieval score is invariant to adversarial perturbation.

Lemma 3.2. *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.

Proof. Recall R_x is the subset of the reference set R in which the samples have the same label as x. With the classifier h, denote the nearest embedding of sample x in R_x by x^+, it is not hard to observe

$$d(x, h) = \min_{y \in R/R_x} ||h(x) - h(y)||_2 - ||h(x) - h(x^+)||_2,$$

thus we have

$$||h(x) - h(x^+)||_2 \leq ||h(x) - h(y)||_2 - d(x, h), \forall y \in R/R_x.$$

For an arbitrary $y \in R/R_x$, we will show $||h(x+\delta) - h(x^+)||_2 < ||h(x+\delta) - h(y)||_2$, if $||h(x) - h(x+\delta)||_2 < \frac{d(x;h)}{2}$.

$$\begin{aligned}
&||h(x+\delta) - h(x^+)||_2 \\
&\leq ||h(x+\delta) - h(x)||_2 + ||h(x) - h(x^+)||_2 \\
&< \frac{d}{2} + ||h(x) - h(x^+)||_2 \\
&\leq \frac{d}{2} + ||h(x) - h(y)||_2 - d \\
&\leq ||h(x) - h(y)||_2 - \frac{d}{2} \\
&< ||h(x) - h(y)||_2 - ||h(x+\delta) - h(x)||_2 \\
&\leq ||h(x+\delta) - h(y)||_2
\end{aligned}$$

The nearest embedding of $x+\delta$ might change (not x^+), but still have the same label as x, which means the 1-NN retrieval score $x+\delta$ is still 1. □

If h is a Lipschitz continuous embedding, i.e., there exists a constant L such that

$$||h(x) - h(y)||_2 \leq L||x - y||_2,$$

we can choose perturbation δ such that its ℓ_2 norm is bounded by $\frac{d(x;h)}{2L}$. In this case,

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

That is, h is guaranteed to be robust against any perturbation as long as its ℓ_2 radius is bounded by $\frac{d(x;h)}{2L}$. Our following subsections will focus on designing an embedding model with a bounded Lipschitz constant.

3.2. Building a Lipschitz continuous model

It is commonly known that deep neural networks are not Lipschitz continuous (Weng et al., 2018). To build a Lipschitz continuous embedding, one approach is by using Lipschitz preserving layers, e.g., orthogonal convolution neural networks (OCNN) (Wang et al., 2020b). However, OCNN is hard to be trained due to its strong constraints imposed on each layer. Another approach is by applying randomized smoothing (Cohen et al., 2019) to the base embedding model. It has been shown that in the classification tasks, the smoothed classifier is ℓ_2-Lipschitz bounded (Salman et al., 2019). In this work, we show that we can build a smoothed embedding model which also has bounded Lipschitz constant beyond the classification problem. In particular, 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

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

Different from the randomized smoothing method in the classification task, where h and g represent the probability of correct predictions, in the embedding models $h(x)$ and $g(x)$ represent the feature vectors of sample x. In the next theorem, we show that if we select q as a Gaussian distribution, the smoothed model g has a bounded Lipschitz constant.

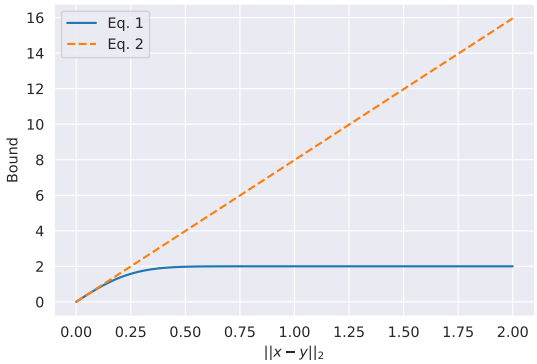

Figure 1. Comparison of the upper bounds of $||g(x) - g(y)||_2$ given by Equation 1 and Equation 2, where we set $F = 1, \sigma = 0.1$.

Lemma 3.3. *If $q \sim \mathcal{N}(0, \sigma^2 I)$, for arbitrary samples x, y,*

$$\begin{aligned}
&||g(x) - g(y)||_2 \\
&\leq 2F\left(\Phi\left(\frac{||x-y||_2}{2\sigma}\right) - \Phi\left(\frac{-||x-y||_2}{2\sigma}\right)\right),
\end{aligned} \tag{1}$$

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.

Detailed proof is in Appendix A.

Tightness of this bound. Consider a one-dimensional dataset X in $\mathbb{R}$ and an embedding model $h(x) = Fsign(x)$. The smoothed model $g(x) = \mathbb{E}_{z\sim\mathcal{N}(0,\sigma^2)}[h(x+z)] = F(\Phi(\frac{x}{\sigma}) - \Phi(-\frac{x}{\sigma}))$. Given two samples x and $-x$, the difference between $g(x)$ and $g(-x)$ is $2F(\Phi(\frac{x}{\sigma}) - \Phi(-\frac{x}{\sigma}))$, which reaches the upper bound in Equation 1. As $\Phi(z) - \Phi(-z) \leq \sqrt{\frac{2}{\pi}}z$ for $z \geq 0$, we have

$$\begin{aligned}
&2F\left(\Phi\left(\frac{||x-y||_2}{2\sigma}\right) - \Phi\left(\frac{-||x-y||_2}{2\sigma}\right)\right) \\
&\leq F\sqrt{\frac{2}{\pi\sigma^2}}||x-y||_2.
\end{aligned}$$

Thus the smoothed model g has bounded Lipschitz constant, and bounded ℓ_2 perturbation on the input will result in bounded shift of its embedding.

3.3. Calculating certified radius

Definition 3.4. (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:

$$r(x;h) := \max_{r\geq 0} r, \text{ s.t. } R_1(x) = R_1(x+\delta), \forall\delta\in\{||\delta||_2 < r\},$$

where $R_1(x)$ is the retrieval score of x.

Following the discussion in subsection 3.2, the smoothed embedding model g satisfies

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

Thus if we choose $r(x; g) = \frac{\sigma\sqrt{\pi}}{2\sqrt{2}F}d(x; g)$, for all δs with $||\delta||_2 < r(x; g)$, we have $||g(x) - g(x + \delta)||_2 < \frac{d(x;g)}{2}$.

However, Equation 2 is looser than Equation 1, and the certified radius computed by Equation 2 is smaller than the radius given by Equation 1. As shown in Figure 1, when $F = 1$ and $\sigma = 0.1$, the Lipschitz bound of Equation 2 is much worse than that of Equation 1 when $||x - y||_2$ is moderately large. Thus we will use the tighter bound (Equation 1) to calculate our certified radius.

Theorem 3.5. *For any sample x and the smoothed embedding g, with Equation 1, if $d(x; g) > 0$, the certified radius of x is*

$$r(x; g) = 2\sigma\Phi^{-1}\left(\frac{1}{2} + \frac{d(x; g)}{8F}\right). \tag{3}$$

Proof. From Lemma 3.2, we know that the 1-NN retrieval score of a sample x with smoothed embedding model g does not change when

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

From Lemma 3.3, we have

$$||g(x) - g(x + \delta)||_2 \leq 2F(\Phi(\frac{||\delta||_2}{2\sigma}) - \Phi(\frac{-||\delta||_2}{2\sigma})),$$

thus if δ satisfies

$$2F(\Phi(\frac{||\delta||_2}{2\sigma}) - \Phi(\frac{-||\delta||_2}{2\sigma})) < \frac{d(x; g)}{2},$$

the 1-NN retrieval score of x will not change. As

$$2F(\Phi(\frac{||\delta||_2}{2\sigma}) - \Phi(\frac{-||\delta||_2}{2\sigma})) = 2F(2\Phi(\frac{||\delta||_2}{2\sigma}) - 1),$$

by solving

$$2F(2\Phi(\frac{||\delta||_2}{2\sigma}) - 1) < \frac{d(x; g)}{2},$$

we have $||\delta||_2 < 2\sigma\Phi^{-1}(\frac{1}{2} + \frac{d(x;g)}{8F})$, thus the certified radius $r(x, g) = 2\sigma\Phi^{-1}(\frac{1}{2} + \frac{d(x;g)}{8F})$ □

In practice, it is hard to compute the smoothed model $g = \mathbb{E}_{z\sim\mathcal{N}(0,\sigma^2 I_d)}[h(x + z)]$ and the minimum margin $d(x; g)$ in a closed form. To resolve the issue, we use Monte-Carlo sampling to estimate $g(x)$ and calculate a probabilistic lower bound of $d(x; g)$. Denote the Monte-Carlo estimation of $g(x)$ by

$$\hat{g}(x) := \frac{1}{n}\sum_{i=1}^{n} h(x + z_i),$$

where $\{z_1, ..., z_n\}$ are sampled i.i.d. from $\mathcal{N}(0, \sigma^2 I_d)$. By matrix Chernoff bound (Ahlswede & Winter, 2002; Tropp, 2012), we have the following theorem.

Lemma 3.6. *With $g(x) \in \mathbb{R}^k$ and $\hat{g}(x) := \frac{1}{n}\sum_{i=1}^{n} h(x + z_i)$, where $\{z_1, ..., z_n\}$ are the Monte-Carlo samples of $\mathcal{N}(0, \sigma^2 I_d)$, we have*

$$\mathbb{P}(||g(x) - \hat{g}(x)||_2 > \epsilon) \leq (k + 1)\exp\left(-\frac{3\epsilon^2 n}{8F^2}\right).$$

Proof. We start with an introduction of the matrix Chernoff bound.

Lemma 3.7. *(Matrix Chernoff bound (Ahlswede & Winter, 2002; Tropp, 2012)) Let $M_1, ..., M_t$ be independent matrix valued random variables such that $M_i \in \mathbb{C}^{d_1\times d_2}$ and $\mathbb{E}[M_i] = \mu$. Denote the operator norm of the matrix M by $\|M\|$. If $\|M_i\| \leq \gamma$ holds almost surely for all $i \in \{1, \ldots, t\}$, then for every $\epsilon > 0$*

$$\mathbb{P}\left(\left\|\frac{1}{t}\sum_{i=1}^{t} M_i - \mu\right\| > \varepsilon\right) \leq (d_1 + d_2)\exp\left(-\frac{3\varepsilon^2 t}{8\gamma^2}\right). \tag{4}$$

Now we prove Lemma 3.6.

Assume we have n i.i.d. random variables $G_1, ..., G_n$, each G_i have the same distribution with $f(x + Z)$, where Z is an arbitrary smoothing distribution. Notice in our paper $Z \sim \mathcal{N}(0, \sigma^2 I)$, but this does not influence the conclusion. We will show for any Z, the bound in Lemma 3.6 with $g(x) = \mathbb{E}[f(x + Z)]$ always hold.

Since G_i has the same distribution with $f(x + Z)$, we have $G_i \in \mathbb{R}^{k\times 1}$ and $\mathbb{E}[G_i] = g(x)$. The ℓ_2 operator norm of G_i is given by,

$$\begin{aligned}||G_i|| &= \sup_{||v||_2\leq 1, v\in\mathbb{R}^{k\times 1}} < G_i, v > \\ &= ||G_i||_2 \leq \sup_x ||f(x)||_2 = F.\end{aligned}$$

Thus with Lemma 3.7 we have

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n} G_i - g(x)\right\| > \epsilon\right) \leq (k + 1)\exp\left(-\frac{3\epsilon^2 n}{8F^2}\right). \tag{5}$$

As $\hat{g}(x) = \sum_{i=1}^{n} f(x + Z_i)$, where Z_i are sampled independently from Z, $\hat{g}(x)$ follows the distribution of $\frac{1}{n}\sum_{i=1}^{n} G_i$. Therefore we have

$$\mathbb{P}(||g(x) - \hat{g}(x)||_2 > \epsilon) \leq (k + 1)\exp(-\frac{3\epsilon^2 n}{8F^2})$$

□

Taking $\alpha = (k+1)\exp(-\frac{3\epsilon^2 n}{8F^2})$, we have that with probability at least $1-\alpha$, the ℓ_2 norm of $g(x) - \hat{g}(x)$ is upper bounded by $\sqrt{8F^2 \ln(\frac{k+1}{\alpha})/3n}$. Thus the error of the estimation is asymptotically decreasing with rate $O(1/\sqrt{n})$, and we can obtain arbitrarily accurate estimation of $g(x)$ with large Monte-Carlo samplings.

Lemma 3.8. *With probability at least* $1-\alpha$,

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

Detailed proof is in Appendix B.

With the lower bound estimation and Theorem 3.5, we are able to calculate the certified radius for any given sample x.

Proposition 3.9. *(Monte-Carlo calculation of certified radius) If* $\underline{d}(x;g) > 0$*, with probability at least* $1-\alpha$,

$$r(x;g) \geq 2\sigma\Phi^{-1}\left(\frac{1}{2} + \frac{\underline{d}(x;g)}{8F}\right). \tag{6}$$

Proof. This proposition is a naive combination of Theorem 3.5 and Lemma 3.8, as $\Phi^{-1}(x)$ is a monotonously increasing function with x and $d(x,g) \geq \underline{d}(x,g)$ with probability $1-\alpha$, obviously

$$\begin{aligned} r(x;g) &= 2\sigma\Phi^{-1}\left(\frac{1}{2} + \frac{d(x;g)}{8F}\right) \\ &\geq 2\sigma\Phi^{-1}\left(\frac{1}{2} + \frac{\underline{d}(x;g)}{8F}\right) \end{aligned}$$

with probability $1-\alpha$. □

Remark 3.10. Our Monte-Carlo calculation of certified radius logarithmically depends on the dimension of the embedding space k. So our method works even when the dimension of the embedding space is high. This is different from randomized smoothing as its output is required to be a one-dimensional scalar.

Algorithm 1 describes our procedure of calculating the certified radius. We want to emphasize that the base embedding model h does not have any robustness guarantee; only its smoothed version g is certifiably robust.

New techniques compared to randomized smoothing. a) Randomized smoothing is designed for the classification tasks, where one can prove that the prediction score w.r.t. the smoothed classifier is larger than 0.5 if the true label is 1. However, in the 1-NN retrieval tasks, we need to identify the conditions under which the 1-NN search is robust against perturbation attacks (Lemma 3.2). b) We prove the Lipschitzness of the smoothed embedding model in Lemma 3.3. Compared to randomized smoothing where the output is a one-dimensional scalar, in the image retrieval tasks we need to take into account the high-dimensional nature of the embedding space (see Remark 3.10). c) Compared to the probabilistic guarantee in randomized smoothing, which is a result of Neyman-Pearson lemma, the probabilistic guarantee (Lemma 3.8) for our model is based on brand new analysis of minimum margin in Definition 3.1. The minimum margin $d(x,g)$ depends on multiple samples, and we need to provide a union bound to characterize the uncertainty of all relevant samples.

Algorithm 1 Certified Radius by RetrievalGuard

Input: training set X; number of random samples n; base embedding model h; standard derivation of Gaussian σ; confidence level α.

Initialize class balanced sampler S;
for $x \in X$ **do**
  sample N random variable $z_1, ..., z_N$ from $\mathcal{N}(0, \sigma^2 I)$;
  calculate $\hat{g}(x) = \frac{1}{n}\sum_{i=1}^{n} h(x + z_i)$;
end
for $x \in X$ **do**
  calculate $\underline{d}(x;g)$ by Lemma 3.8;
  if $\underline{d}(x;g) < 0$ **then**
    $r(x) = -1$; (reject this sample because its retrieval score is 0)
  else
    calculate certified radius $r(x) = 2\sigma\Phi^{-1}(1/2 + \underline{d}(x;g)/8F)$;
  end
end
return all certified radius r.

4. Experiments

In the experiments, we use the metric learning model with margin loss (Wu et al., 2017) as our base embedding model. We apply our RetrievalGuard approach on vanilla metric learning (DML) and the DML augmented by Gaussian noise (GDML) to build the smoothed DML and compare them on three benchmarks. We emphasize that all results reported in this section are from the smoothed models g instead of the base models h, as we can only provide robustness guarantee for g.

4.1. Deep metric learning with margin loss

Margin loss is a tuple-based metric loss, which requires (anchor, positive, negative) triplets as input. The anchor and the positive point are expected to be in the same class while the anchor and the negative point should be in the

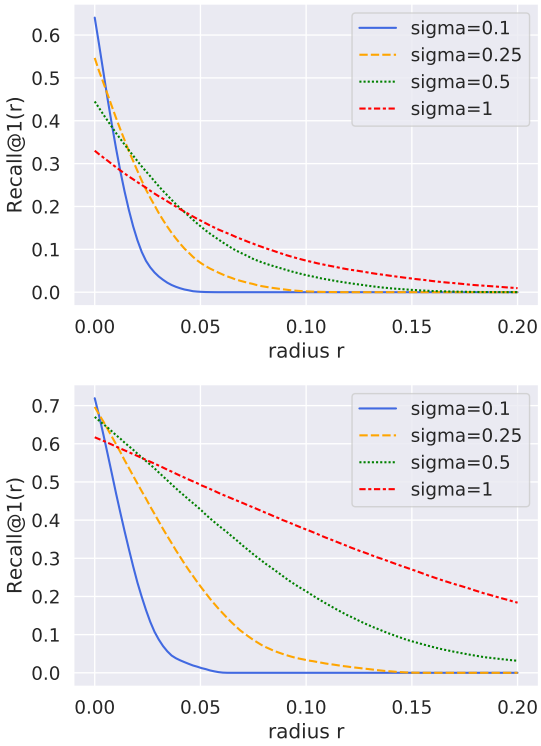

Figure 2. Experiments with DML+RetrievalGuard on image retrieval benchmarks with different σ. **Top**: DML+RetrievalGuard on Online-Products. **Bottem**: GDML+RetrievalGuard on Online-Products.

different classes. Denote the (anchor, positive, negative) by (x, x^+, x^-) and the distribution of the triplets by p_{tri}. The margin loss (Wu et al., 2017) is defined as

$$L(h,\beta) = \mathbb{E}_{(x,x^+,x^-)\sim p_{tri}}[(||h(x)-h(x^+)||_2 - \beta + \gamma)_+ + (\beta - ||h(x)-h(x^-)||_2 + \gamma)_+],$$

where β is a learnable parameter with initial value 0.6 or 1.2 and learning rate 0.0005. $\gamma = 0.2$ is a fixed triplet margin. In the margin loss, p_{tri} is given by a distance sampling method, such that the probability of sampling a negative point with large distance to x in the embedding space is much larger than that of sampling a negative point with small distance to x.

4.2. Gaussian augmented model

The norm of Gaussian noise sampled from $\mathcal{N}(0, \sigma^2 I_d)$ is of magnitude $\Theta(\sigma\sqrt{d})$ with high probability (Zhang et al., 2020a). With moderately large σ, the distribution of natural images has nearly disjoint support from the distribution of Gaussian-perturbed images. It is therefore hard for the base model h to generate effective embedding, if it can only get access to natural images. As a result, the smoothed model g, which is estimated by averaging the base embedding, may suffer from poor performance. A solution to resolve this issue is by training the base embedding model h with Gaussian augmented images (Cohen et al., 2019). Figure 2 shows that using Gaussian augmented model as the base model outperforms using the vanilla model as the base model in all settings. The objective function for the Gaussian augmented model is

$$L(h,\beta) = \mathbb{E}_{(x,x^+,x^-)\sim p_{tri}}[(||h(x+Z_1)-h(x^+ + Z_2)||_2 - \beta + \gamma)_+ + (\beta - ||h(x+Z_1)-h(x^- + Z_3)||_2 + \gamma)_+],$$

where Z_1, Z_2, and Z_3 are Gaussian random variables and $(t)_+ = \max\{t, 0\}$.

4.3. Experimental settings

Datasets. We run experiments with a popular dataset Online-Products of metric learning (Song et al., 2016), which contains 120,053 product images. We use the first 11,318 classes of products as the training set and another 1,000 classes as the test set. The experiments with CUB200 (Wah et al., 2011) and CARS196 (Krause et al., 2013) are listed in Appendix C

Training hyper-parameters. We adapt the DML framework from (Roth et al., 2020) for our training. In all experiments, we use ResNet50 architecture (He et al., 2016) pretrained on the ImageNet dataset (Krizhevsky et al., 2012) with frozen Batch-Normalization layers as our backbone. We first re-scale the images to $[0, 1]^d$, then randomly resize and crop the images to 224×224 for training, and apply center crop to the same size for evaluation. The embedding dimension k is 128 and the number of training epochs is 100. The learning rate is 1e-5 with multi-step learning rate scheduler 0.3 at the 30-th, 55-th, and 75-th epochs. We select the initial value of β as 1.2, learning rate of β as 0.0005, and $\gamma = 0.2$ in the margin loss. We also test the model performance under Gaussian noise with $\sigma = 0.1, 0.25, 0.5, 1$. As the embedding of metric learning models is ℓ_2 normalized, we have $F = 1$. For each sample, we generate 100,000 Monte-Carlo samples to estimate $g(x)$. The confidence level α is chosen as 0.01. The running time of RetrievalGuard for a single image evaluation with 100,000 Monte-Carlo samples on a 24GB Nvidia Tesla P40 GPU is about 3 minutes.

Evaluation metrics. We focus on the 1-NN retrieval task. A natural metric is the Recall@1 score, which is given by the average of 1-NN retrieval score of all samples, i.e., $Recall@1 = \frac{1}{N}\sum_{i=1}^N R_1(x_i)$. To evaluate the certified robustness of the 1-NN retrieval, we define $Recall@1(r) = \frac{1}{N}\sum_{i=1}^N R_1(x_i)\mathbb{I}_{r(x_i,g)>r}$, which represents the averaged

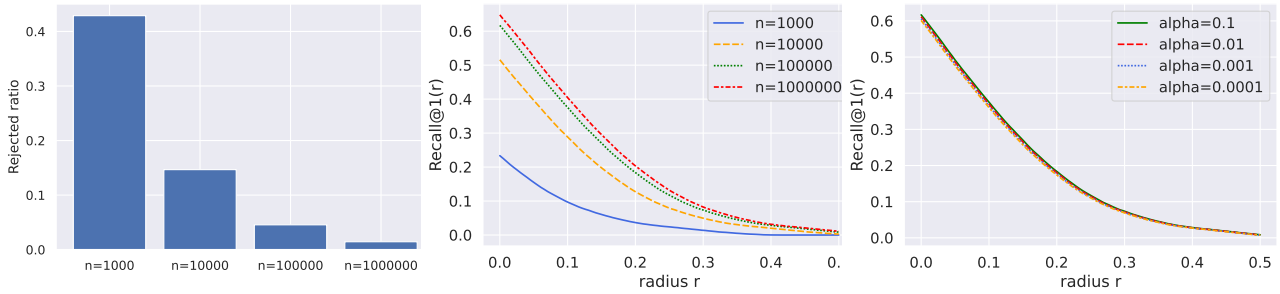

Figure 3. Experiments with GDML+RetrievalGuard on Online-Products with $\sigma = 1$. **Left**: Comparison of rejected ratio with the number of Monte-Carlo samples n. The rejected ratio is the ratio of samples with $d(x, \hat{g}) > 0$ but $\underline{d}(x, g) \leq 0$. **Middle**: Comparison of $Recall@1(r)$ if the number of Monte-Carlo samples n is larger or smaller. **Right**: Comparison of $Recall@1(r)$ under varying values of α.

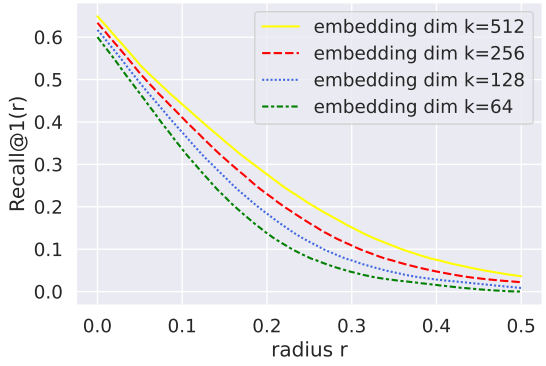

Figure 4. GDML+RetrievalGuard ($\sigma = 1$) with different dimensions of embedding space on the Online-Products dataset.

retrieval score of the samples such that the certified radius is larger than r. Note that $Recall@1 = Recall@1(0)$.

4.4. Experimental results

4.4.1. Performance of the smoothed DML models.

Figure 2 shows the plot of $Recall@1(r)$ of DML+RG and GDML+RG models under varying values of σ. We see that there is a robustness/accuracy trade-off (Yang et al., 2020b; Zhang et al., 2019a) controlled by σ. When σ is low, small radii can be certified with high retrieval score, while large radii cannot be certified. When σ is high, large radii can be certified, while small radii are certified with a low retrieval score. Besides, the GDML+RG models outperform the DML+RG models in all experiments, which is consistent with our discussions in subsection 4.2.

4.4.2. Ablation study

We study the effect of number of Monte-Carlo samples n, the failure probability α, and the embedding size k on the model robustness. All experiments are run on the GDML+RetrievalGuard model with $\sigma = 1$ and the Online-Products dataset.

Figure 3 (**left**) plots the rejected ratio under different numbers of Monte-Carlo samples n. We use a fixed $\alpha = 0.01$ in our experiment. The rejected ratio is the ratio of samples with retrieval score 1 on the estimated model $\hat{g}$, i.e., $d(x; \hat{g}) > 0$, but x is rejected as $\underline{d}(x; g) \leq 0$. The ratio of rejected samples is decreasing w.r.t. n. When $n = 10,000$, there are 14% samples being rejected. In our experiments, we choose $n = 100,000$ with roughly 4% rejected ratio.

Figure 3 (**middle**) illustrates the $Recall@1(r)$ under different numbers of Monte-Carlo samples n. We use a fixed $\alpha = 0.01$ in the experiment. When n is increased from $1,000$ to $100,000$, the improvement of $Recall@1(r)$ is significant. When n is increased from $100,000$ to $1,000,000$, the improvement of $Recall@1(r)$ is relatively small. Therefore, we choose $n = 100,000$ in our experiments.

Figure 3 (**right**) draws the $Recall@1(r)$ under different confidence levels α. We fix $n = 100,000$ in the experiment. It shows that $Recall@1(r)$ is stable under varying values of α, which indicates that our robustness guarantee is not sensitive to α.

Figure 4 shows the $Recall@1(r)$ under different dimensions k of embedding space. We use a fixed $n = 100,000$ and $\alpha = 0.01$ in the experiment. It shows that the models are more robust with larger k. This is because high-dimensional embedding space can improve the expressive power of the DML models, and dissimilar samples can be separated with a large margin, i.e., $\underline{d}(x, g)$ is large.

5. Conclusion

In this work, we propose RetrievalGuard, the first provably robust 1-NN image retrieval model, by smoothing the vanilla embedding model with a Gaussian distribution. We prove that, with arbitrary perturbation δ, whose ℓ_2 norm is bounded by the certified radius, the 1-NN retrieval score of the perturbed samples on the smoothed model does not change. We empirically demonstrate the effectiveness of our model on image retrieval tasks with Online-Products. Future works include designing ℓ_p certificate algorithms and

extending our algorithm to k-NN image retrieval tasks.

Acknowledgement

Hongyang Zhang was supported in part by an NSERC Discovery Grant. Yihan Wu and Heng Huang were partially supported by NSF IIS 1845666, 1852606, 1838627, 1837956, 1956002, IIA 2040588.

References

Ahlswede, R. and Winter, A. Strong converse for identification via quantum channels. *IEEE Transactions on Information Theory*, 48(3):569–579, 2002.

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pp. 274–283, 2018.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., and Roli, F. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 387–402. Springer, 2013.

Blum, A., Dick, T., Manoj, N., and Zhang, H. Random smoothing might be unable to certify ℓ_∞ robustness for high-dimensional images. *Journal of Machine Learning Research*, 21:1–21, 2020.

Blum, A., Montasser, O., Shakhnarovich, G., and Zhang, H. Boosting barely robust learners: A new perspective on adversarial robustness. *arXiv preprint arXiv:2202.05920*, 2022.

Bojchevski, A., Klicpera, J., and Günnemann, S. Efficient robustness certificates for discrete data: Sparsity-aware randomized smoothing for graphs, images and more. In *International Conference on Machine Learning*, pp. 1003–1013. PMLR, 2020.

Boudiaf, M., Rony, J., Ziko, I. M., Granger, E., Pedersoli, M., Piantanida, P., and Ayed, I. B. A unifying mutual information view of metric learning: cross-entropy vs. pairwise losses. In *European Conference on Computer Vision*, pp. 548–564. Springer, 2020.

Bouniot, Q., Audigier, R., and Loesch, A. Vulnerability of person re-identification models to metric adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 794–795, 2020.

Chaudhuri, U., Banerjee, B., and Bhattacharya, A. Siamese graph convolutional network for content based remote sensing image retrieval. *Computer vision and image understanding*, 184:22–30, 2019.

Cohen, J. M., Rosenfeld, E., and Kolter, J. Z. Certified adversarial robustness via randomized smoothing. *ICML*, 2019.

Feng, Y., Chen, B., Dai, T., and Xia, S.-T. Adversarial attack on deep product quantization network for image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 10786–10793, 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kim, S., Kim, D., Cho, M., and Kwak, S. Proxy anchor loss for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3238–3247, 2020.

Krause, J., Stark, M., Deng, J., and Fei-Fei, L. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25: 1097–1105, 2012.

Kumar, A., Levine, A., Goldstein, T., and Feizi, S. Curse of dimensionality on randomized smoothing for certifiable robustness. *ICML*, 2020.

Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., and Jana, S. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, may 2019. doi: 10.1109/sp.2019.00044.

Li, B., Chen, C., Wang, W., and Carin, L. Certified adversarial robustness with additive noise. In *Advances in Neural Information Processing Systems*, pp. 9464–9474, 2019a.

Li, J., Ji, R., Liu, H., Hong, X., Gao, Y., and Tian, Q. Universal perturbation attack against image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4899–4908, 2019b.

Liu, H., Jia, J., and Gong, N. Z. Pointguard: Provably robust 3d point cloud classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6186–6195, 2021.

Liu, Z., Luo, P., Qiu, S., Wang, X., and Tang, X. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1096–1104, 2016.

Musgrave, K., Belongie, S., and Lim, S.-N. A metric learning reality check. In *European Conference on Computer Vision*, pp. 681–699. Springer, 2020.

Nair, L. R., Subramaniam, K., and Prasannavenkatesan, G. A review on multiple approaches to medical image retrieval system. In *Intelligent Computing in Engineering*, pp. 501–509. Springer, 2020.

Panum, T. K., Wang, Z., Kan, P., Fernandes, E., and Jha, S. Exploring adversarial robustness of deep metric learning. *arXiv preprint arXiv:2102.07265*, 2021.

Roth, K., Milbich, T., Sinha, S., Gupta, P., Ommer, B., and Cohen, J. P. Revisiting training strategies and generalization performance in deep metric learning. In *International Conference on Machine Learning*, pp. 8242–8252. PMLR, 2020.

Salman, H., Li, J., Razenshteyn, I., Zhang, P., Zhang, H., Bubeck, S., and Yang, G. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems*, pp. 11292–11303, 2019.

Schroff, F., Kalenichenko, D., and Philbin, J. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Song, H. O., Xiang, Y., Jegelka, S., and Savarese, S. Deep metric learning via lifted structured feature embedding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *ICML*, 2014.

Tjeng, V., Xiao, K. Y., and Tedrake, R. Evaluating robustness of neural networks with mixed integer programming. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyGIdiRqtm.

Tropp, J. A. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12 (4):389–434, 2012.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Wang, H., Wang, G., Li, Y., Zhang, D., and Lin, L. Transferable, controllable, and inconspicuous adversarial attacks on person re-identification with deep mis-ranking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 342–351, 2020a.

Wang, J., Chen, Y., Chakraborty, R., and Yu, S. X. Orthogonal convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11505–11515, 2020b.

Wang, X., Han, X., Huang, W., Dong, D., and Scott, M. R. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5022–5030, 2019.

Weng, T.-W., Zhang, H., Chen, P.-Y., Yi, J., Su, D., Gao, Y., Hsieh, C.-J., and Daniel, L. Evaluating the robustness of neural networks: An extreme value theory approach. *arXiv preprint arXiv:1801.10578*, 2018.

Wu, C.-Y., Manmatha, R., Smola, A. J., and Krahenbuhl, P. Sampling matters in deep embedding learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2840–2848, 2017.

Wu, Y., Bojchevski, A., Kuvshinov, A., and Günnemann, S. Completing the picture: Randomized smoothing suffers from the curse of dimensionality for a large family of distributions. In *International Conference on Artificial Intelligence and Statistics*, pp. 3763–3771. PMLR, 2021.

Yang, X., Wei, F., Zhang, H., and Zhu, J. Design and interpretation of universal adversarial patches in face detection. In *European Conference on Computer Vision*, pp. 174–191. Springer, 2020a.

Yang, Y.-Y., Rashtchian, C., Zhang, H., Salakhutdinov, R., and Chaudhuri, K. A closer look at accuracy vs. robustness. In *Adavnces in Neural Information Processing Systems, NeurIPS*, 2020b.

Zhang, D., Ye, M., Gong, C., Zhu, Z., and Liu, Q. Black-box certification with randomized smoothing: A functional optimization based framework. *arXiv preprint arXiv:2002.09169*, 2020a.

Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., and Jordan, M. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pp. 7472–7482. PMLR, 2019a.

Zhang, H., Zhang, P., and Hsieh, C.-J. Recurjac: An efficient recursive algorithm for bounding jacobian matrix of neural networks and its applications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 5757–5764, 2019b.

Zhang, H., Chen, H., Xiao, C., Gowal, S., Stanforth, R., Li, B., Boning, D., and Hsieh, C.-J. Towards stable and efficient training of verifiably robust neural networks. In *International Conference on Learning Representations*, 2020b. URL https://openreview.net/forum?id=Skxuk1rFwB.

Zhang, H., Wu, Y., and Huang, H. How many data are needed for robust learning? *arXiv preprint arXiv:2202.11592*, 2022.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., and Tian, Q. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pp. 1116–1124, 2015.

Zhou, M., Niu, Z., Wang, L., Zhang, Q., and Hua, G. Adversarial ranking attack and defense. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pp. 781–799. Springer, 2020.

A. Proof of Lemma 3.3

Proof. Recall that $g(x) := \mathbb{E}_{z\sim\mathcal{N}(0,\sigma^2 I_d)}[f(x+z)] = \int_{\mathbb{R}^d} f(x+z)p(z)dz$, where p is the probability density function of $\mathcal{N}(0,\sigma^2 I_d)$. The intuition of this proof is to seek the maximum discrepancy of the expectation $\mathbb{E}_{z\sim\mathcal{N}(0,\sigma^2 I_d)}[f(x+z)]$ and $\mathbb{E}_{z\sim\mathcal{N}(0,\sigma^2 I_d)}[f(y+z)]$ for arbitrary samples x, y.

$$
\begin{aligned}
||g(x)-g(y)||_2 &= ||\int_{\mathbb{R}^d} f(x+z)p(z)dz - \int_{\mathbb{R}^d} f(y+z)p(z)dz||_2 \\
&= ||\int_{\mathbb{R}^d} f(z)p(z-x)dz - \int_{\mathbb{R}^d} f(z)p(z-y)dz||_2 \\
&= ||\int_{\mathbb{R}^d} f(z)(p(z-x)-p(z-y))dz||_2 \\
&= ||\int_{\mathbb{R}^d} f(z)(p(z-x)-p(z-y))_+dz + \int_{\mathbb{R}^d} f(z)(p(z-x)-p(z-y))_-dz||_2 \\
&\le ||\int_{\mathbb{R}^d} f(z)(p(z-x)-p(z-y))_+dz||_2 + ||\int_{\mathbb{R}^d} f(z)(p(z-x)-p(z-y))_-dz||_2 \\
&\le F(|\int_{\mathbb{R}^d} (p(z-x)-p(z-y))_+dz| + |\int_{\mathbb{R}^d} (p(z-x)-p(z-y))_-dz|) \\
&= F(|\int_{\mathbb{R}^d} (p(z)-p(z+x-y))_+dz| + |\int_{\mathbb{R}^d} (p(z)-p(z+x-y))_-dz|)
\end{aligned}
$$

Now we need to calculate $\int_{\mathbb{R}^d}(p(z)-p(z+x-y))_+dz$ explicitly, as $p(z) = \frac{1}{(2\pi\sigma^2)^{d/2}}\exp(-\frac{z^Tz}{2\sigma})$

$$
\int_{\mathbb{R}^d}(p(z)-p(z+x-y))_+dz = \frac{1}{(2\pi\sigma^2)^{d/2}}\int_{\mathbb{R}^d}(\exp(-\frac{z^Tz}{2\sigma}) - \exp(-\frac{(z+x-y)^T(z+x-y)}{2\sigma}))_+dz
$$

By solving $\exp(-\frac{z^Tz}{2\sigma}) - \exp(-\frac{(z+x-y)^T(z+x-y)}{2\sigma}) \ge 0$ we have $z \in \{z^T(x-y) \le \frac{1}{2}(x-y)^T(x-y)\} := D$. As Gaussian distribution is ℓ_2 spherically symmetric, we can make a unitary transformation such that $x-y$ located on the first axis in the $\mathbb{R}^d$ space, in this case $D = \{z_1 \le \frac{1}{2}||x-y||_2\}$ assuming w.l.o.g $x \ge y$, we have

$$
\begin{aligned}
\int_{\mathbb{R}^d}(p(z)-p(z+x-y))_+dz &= \frac{1}{(2\pi\sigma^2)^{d/2}}\int_D \exp(-\frac{z_1^2+\sum_{i=2}^d z_i^2}{2\sigma}) - \exp(-\frac{(z_1+||x-y||_2)^2+\sum_{i=2}^d z_i^2}{2\sigma})dz \\
&= \frac{1}{(2\pi\sigma^2)^{1/2}}\int_{z_1\le\frac{||x-y||_2}{2}} \exp(-\frac{z_1^2}{2\sigma}) - \exp(-\frac{(z_1+||x-y||_2)^2}{2\sigma})dz_1 \\
&= \Phi(\frac{||x-y||_2}{2\sigma}) - \Phi(-\frac{||x-y||_2}{2\sigma})
\end{aligned}
$$

where $\Phi(x) = \frac{1}{(2\pi)^{1/2}}\int_{-\infty}^x \exp(-\frac{z^2}{2})dz$ is the cumulative density function of a standard normal distribution. Anogously we have

$$
\int_{\mathbb{R}^d}(p(z)-p(z+x-y))_-dz = \Phi(-\frac{||x-y||_2}{2\sigma}) - \Phi(\frac{||x-y||_2}{2\sigma})
$$

Thus

$$
\begin{aligned}
||g(x)-g(y)||_2 &= F(|\int_{\mathbb{R}^d}(p(z)-p(z+x-y))_+dz| + |\int_{\mathbb{R}^d}(p(z)-p(z+x-y))_-dz|) \\
&= 2F(\Phi(\frac{||x-y||_2}{2\sigma}) - \Phi(-\frac{||x-y||_2}{2\sigma}))
\end{aligned}
$$

□

B. Proof of Lemma 3.8

Proof. Recall R_x is the subset of the reference set R in which the samples have the same label as x. With the estimated classifier $\hat{g}$, denote the nearest embedding of sample x in R_x by x^+, and the nearest embedding of sample x in R_x by x^-, we have

$$d(x;\hat{g}) = ||\hat{g}(x) - \hat{g}(x^-)||_2 - ||\hat{g}(x) - \hat{g}(x^+)||_2$$

According to Theorem 3.6,

$$\mathbb{P}(||g(x) - \hat{g}(x)||_2 > \epsilon) \leq (k+1)\exp(-\frac{3\epsilon^2 n}{8F^2}), \forall x \in \mathbb{R}^d$$

thus with probability at least $1-\alpha$

$$||g(x) - \hat{g}(x)||_2 \leq \sqrt{8F^2 \ln(\frac{k+1}{\alpha})/3n}, \forall x \in \mathbb{R}^d$$

So we can bounded the difference between $||g(x) - g(y)||_2$ and $||\hat{g}(x) - \hat{g}(y)||_2$ for arbitrary sample x and y by

$$\begin{aligned} |||g(x) - g(y)||_2 - ||\hat{g}(x) - \hat{g}(y)||_2| &\leq ||g(x) - g(y) - \hat{g}(x) + \hat{g}(y)||_2 \\ &\leq ||g(x) - \hat{g}(x)||_2 + ||g(y) - \hat{g}(y)||_2 \\ &= 2\sqrt{8F^2 \ln(\frac{k+1}{\alpha})/3n} \end{aligned} \tag{7}$$

with probability at least $1-2\alpha$. We now consider

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

Based on Equation 7 we have

$$\min_{x_2 \in R/R_x} ||g(x) - g(x_2)||_2 \geq ||\hat{g}(x) - \hat{g}(x^*)||_2 - 2\sqrt{8F^2 \ln(\frac{k+1}{\alpha})/3n} \geq ||\hat{g}(x) - \hat{g}(x^-)||_2 - 2\sqrt{8F^2 \ln(\frac{k+1}{\alpha})/3n}$$

with probability at least $1-2\alpha$, where $x^* := arg\min_{x_2 \in R/R_x} ||g(x) - g(x_2)||$, the second inequality is due to $x^- := arg\min_{x_2 \in R/R_x} ||\hat{g}(x) - \hat{g}(x_2)||$. Analogously

$$\min_{x_1 \in R_x} ||g(x) - g(x_1)||_2 \leq ||g(x) - g(x^+)||_2 \leq ||\hat{g}(x) - \hat{g}(x^+)||_2 + 2\sqrt{8F^2 \ln(\frac{k+1}{\alpha})/3n}$$

with probability at least $1-2\alpha$. Thus we have

$$\begin{aligned} d(x;g) &= \min_{x_2 \in R/R_x} ||g(x) - g(x_2)||_2 - \min_{x_1 \in R_x} ||g(x) - g(x_1)||_2 \\ &\geq ||\hat{g}(x) - \hat{g}(x^-)||_2 - 2\sqrt{8F^2 \ln(\frac{k+1}{\alpha})/3n} - \left(||\hat{g}(x) - \hat{g}(x^+)||_2 + 2\sqrt{8F^2 \ln(\frac{k+1}{\alpha})/3n}\right) \\ &= d(x,\hat{g}) - 4\sqrt{8F^2 \ln(\frac{k+1}{\alpha})/3n} \end{aligned} \tag{8}$$

with probability $1-4\alpha$. Replace α by $\frac{\alpha}{4}$ we obtain Theorem 3.8. □

C. Additional experiments on CUB200 and CARS196

Datasets. We conduct experiments on two popular metric learning benchmarks: CUB200 and CARS196. We follow the setup in the previous work (Song et al., 2016) to split the training and test sets.

- **CUB200-2011** contains 200 species of birds and 11,788 images (Wah et al., 2011). We use the first 100 species as the training set and the rest as the test set.
- **CARS196** has 196 models of cars and 16,185 images. (Krause et al., 2013). We use the first 98 models as the training set and the rest as the test set.

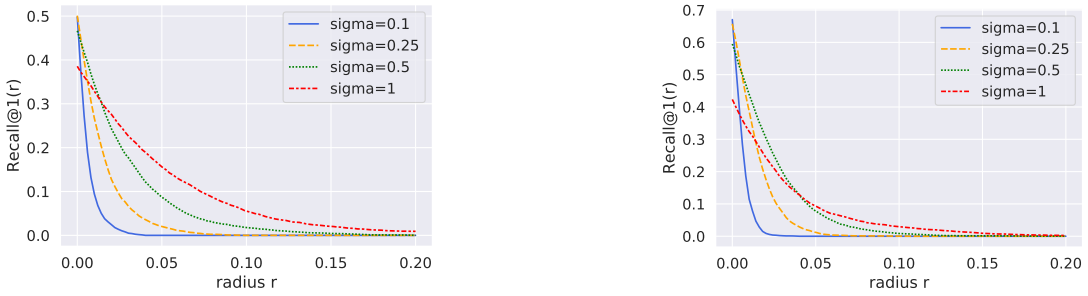

Figure 5. Experiments with GDML+RetrievalGuard on image retrieval benchmarks with different σ. **Left**: GDML+RetrievalGuard on CUB200-2011. **Right**: GDML+RetrievalGuard on CARS196.