MultiGuard: Provably Robust Multi-label Classification against Adversarial Examples

Jinyuan Jia*
University of Illinois Urbana-Champaign
jinyuan@illinois.edu

Wenjie Qu*
Huazhong University of Science and Technology
wen_jie_qu@outlook.com

Neil Zhenqiang Gong
Duke University
neil.gong@duke.edu

Abstract

Multi-label classification, which predicts a set of labels for an input, has many applications. However, multiple recent studies showed that multi-label classification is vulnerable to adversarial examples. In particular, an attacker can manipulate the labels predicted by a multi-label classifier for an input via adding carefully crafted, human-imperceptible perturbation to it. Existing provable defenses for multi-class classification achieve sub-optimal provable robustness guarantees when generalized to multi-label classification. In this work, we propose MultiGuard, the first provably robust defense against adversarial examples to multi-label classification. Our MultiGuard leverages randomized smoothing, which is the state-of-the-art technique to build provably robust classifiers. Specifically, given an arbitrary multi-label classifier, our MultiGuard builds a smoothed multi-label classifier via adding random noise to the input. We consider isotropic Gaussian noise in this work. Our major theoretical contribution is that we show a certain number of ground truth labels of an input are provably in the set of labels predicted by our MultiGuard when the ℓ_2-norm of the adversarial perturbation added to the input is bounded. Moreover, we design an algorithm to compute our provable robustness guarantees. Empirically, we evaluate our MultiGuard on VOC 2007, MS-COCO, and NUS-WIDE benchmark datasets. Our code is available at: https://github.com/quwenjie/MultiGuard

1 Introduction

Multi-class classification assumes each input only has one ground truth label and thus often predicts a single label for an input. In contrast, in *multi-label classification* [42, 41, 35, 43], each input has multiple ground truth labels and thus a multi-label classifier predicts a set of labels for an input. For instance, an image could have multiple objects, attributes, or scenes. Multi-label classification has many applications such as diseases detection [16], object recognition [43], retail checkout recognition [18], document classification [34], etc..

However, similar to multi-class classification, multiple recent studies [56, 53, 30] showed that multi-label classification is also vulnerable to adversarial examples. In particular, an attacker can manipulate the set of labels predicted by a multi-label classifier for an input via adding carefully crafted perturbation to it. Adversarial examples pose severe security threats to the applications of multi-label classification in security-critical domains. To mitigate adversarial examples to multi-label

*Equal contribution. Wenjie Qu performed this research when he was a remote intern in Gong's group.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

classification, several *empirical defenses* [49, 1, 30] have been proposed. For instance, Melacci et al. [30] proposed to use the domain knowledge on the relationships among the classes to improve the robustness of multi-label classification. However, these defenses have no provable robustness guarantees, and thus they are often broken by more advanced attacks. For instance, Melacci et al. [30] showed that their proposed defense can be broken by an adaptive attack that exploits the domain knowledge used in the defense. Moreover, existing provably robust defenses [10, 8, 17, 46, 12, 22] are all for multi-class classification, which achieve sub-optimal provable robustness guarantee when extended to multi-label classification as shown by our experimental results.

Our work: We propose *MultiGuard*, the first provably robust defense against adversarial examples for multi-label classification. MultiGuard leverages randomized smoothing [5, 29, 24, 26, 12], which is the state-of-the-art technique to build provably robust classifiers. In particular, compared to other provably robust techniques, randomized smoothing has two advantages: 1) scalable to large-scale neural networks, and 2) applicable to any classifiers. Suppose we have an arbitrary multi-label classifier (we call it *base multi-label classifier*), which predicts k' labels for an input. We build a *smoothed multi-label classifier* via randomizing an input. Specifically, given an input, we first create a *randomized input* via adding random noise to it. We consider the random noise to be isotropic Gaussian in this work. Then, we use the base multi-label classifier to predict labels for the randomized input. Due to the randomness in the randomized input, the k' labels predicted by the base multi-label classifier are also random. We use p_i to denote the probability that the label i is among the set of k' labels predicted by the base multi-label classifier for the randomized input, where $i \in \{1, 2, \cdots, c\}$. We call p_i *label probability*. Our smoothed multi-label classifier predicts the k labels with the largest label probabilities for the input. We note that k' and k are two different parameters.

Our main theoretical contribution is to show that, given a set of labels (e.g., the ground truth labels) for an input, at least e of them are provably in the set of k labels predicted by MultiGuard for the input, when the ℓ_2-norm of the adversarial perturbation added to the input is no larger than a threshold. We call e *certified intersection size*. We aim to derive the certified intersection size for MultiGuard. However, existing randomized smoothing studies [24, 12, 22] achieves sub-optimal provable robustness guarantees when generalized to derive our certified intersection size. The key reason is they were designed for multi-class classification instead of multi-label classification. Specifically, they can guarantee that a smoothed multi-class classifier provably predicts the same single label for an input [24, 12] or a certain label is provably among the top-k labels predicted by the smoothed multi-class classifier [22]. In contrast, our certified intersection size characterizes the intersection between the set of ground truth labels of an input and the set of labels predicted by a smoothed multi-label classifier. In fact, previous provable robustness results [12, 22] are special cases of ours, e.g., our results reduce to Cohen et al. [12] when $k' = k = 1$ and Jia et al. [22] when $k' = 1$.

In particular, there are two challenges in deriving the certified intersection size. The first challenge is that the base multi-label classifier predicts multiple labels for an input. The second challenge is that an input has multiple ground truth labels. To solve the first challenge, we propose a variant of Neyman-Pearson Lemma [33] that is applicable to multiple functions, which correspond to multiple labels predicted by the base multi-label classifier. In contrast, existing randomized smoothing studies [24, 26, 12, 22] for multi-class classification use the standard Neyman-Pearson Lemma [33] that is only applicable for a single function, since their base multi-class classifier predicts a single label for an input. To address the second challenge, we propose to use the *law of contraposition* to simultaneously consider multiple ground truth labels of an input when deriving the certified intersection size.

Our derived certified intersection size is the optimal solution to an optimization problem, which involves the label probabilities. However, it is very challenging to compute the exact label probabilities due to the continuity of the isotropic Gaussian noise and the complexity of the base multi-label classifiers (e.g., complex deep neural networks). In response, we design a Monte Carlo algorithm to estimate the lower or upper bounds of label probabilities with probabilistic guarantees. More specifically, we can view the estimation of lower or upper bounds of label probabilities as a binomial proportion confidence interval estimation problem in statistics. Therefore, we use the Clopper-Pearson [11] method from the statistics community to obtain the label probability bounds. Given the estimated lower or upper bounds of label probabilities, we design an efficient algorithm to solve the optimization problem to obtain the certified intersection size.

Empirically, we evaluate our MultiGuard on VOC 2007, MS-COCO, and NUS-WIDE benchmark datasets. We use the *certified top-k precision@R*, *certified top-k recall@R*, and *certified top-k f1-score@R* to evaluate our MultiGuard. Roughly speaking, certified top-k precision@R is the least fraction of the k predicted labels that are ground truth labels of an input when the ℓ_2-norm of the adversarial perturbation is at most R; certified top-k recall@R is the least fraction of ground truth labels of an input that are in the set of k labels predicted by our MultiGuard; and certified top-k f1-score@R is the harmonic mean of certified top-k precision@R and certified top-k recall@R. Our experimental results show that our MultiGuard outperforms the state-of-the-art certified defense [22] when extending it to multi-label classification. For instance, on VOC 2007 dataset, Jia et al. [22] and our MultiGuard respectively achieve 24.3% and 31.3% certified top-k precision@R, 51.6% and 66.4% certified top-k recall@R, as well as 33.0% and 42.6% certified top-k f1-score@R when $k' = 1$, $k = 3$, and $R = 0.5$.

Our major contributions can be summarized as follows:

- We propose MultiGuard, the first provably robust defense against adversarial examples for multi-label classification.
- We design a Monte Carlo algorithm to compute the certified intersection size.
- We evaluate our MultiGuard on VOC 2007, MS-COCO, and NUS-WIDE benchmark datasets.

2 Background and Related Work

Multi-label classification: In multi-label classification, a multi-label classifier predicts multiple labels for an input. Many deep learning classifiers [27, 52, 43, 45, 57, 32, 21, 7, 54, 48, 2, 13] have been proposed for multi-label classification. For instance, a naive method for multi-label classification is to train independent binary classifiers for each label and use ranking or thresholding to derive the final predicted labels. This method, however, ignores the topology structure among labels and thus cannot capture the label co-occurrence dependency (e.g., *mouse* and *keyboard* usually appear together). In response, several methods [43, 7] have been proposed to improve the performance of multi-label classification via exploiting the label dependencies in an input. Despite their effectiveness, these methods rely on complicated architecture modifications. To mitigate the issue, some recent studies [48, 2] proposed to design new loss functions. For instance, Baruch et al. [2] introduced an asymmetric loss (ASL). Roughly speaking, their method is based on the observation that, in multi-label classification, most inputs contain only a small fraction of the possible candidate labels, which leads to under-emphasizing gradients from positive labels during training. Their experimental results indicate that their method achieves state-of-the-art performance on multiple benchmark datasets.

Adversarial examples to multi-label classification: Several recent studies [40, 56, 53, 30, 20] showed that multi-label classification is vulnerable to adversarial examples. An attacker can manipulate the set of labels predicted by a multi-label classifier for an input via adding carefully crafted perturbation to it. For instance, Song et al. [40] proposed white-box, targeted attacks to multi-label classification. In particular, they first formulate their attacks as optimization problems and then use gradient descent to solve them. Their experimental results indicate that they can make a multi-label classifier produce an arbitrary set of labels for an input via adding adversarial perturbation to it. Yang et al. [53] explored the worst-case mis-classification risk of a multi-label classifier. In particular, they formulate the problem as a bi-level set function optimization problem and leverage random greedy search to find an approximate solution. Zhou et al. [56] proposed to generate ℓ_∞-norm adversarial perturbations to fool a multi-label classifier. In particular, they transform the optimization problem of finding adversarial perturbations into a linear programming problem which can be solved efficiently.

Existing empirically robust defenses: Some studies [49, 1, 30] developed empirical defenses to mitigate adversarial examples in multi-label classification. For instance, Wu et al. [49] applied adversarial training, a method developed to train robust multi-class classifiers, to improve the robustness of multi-label classifiers. Melacci et al. [30] showed that domain knowledge, which measures the relationships among classes, can be used to detect adversarial examples and improve the robustness of multi-label classifiers. However, all these defenses lack provable robustness guarantees and thus, they are often broken by advanced adaptive attacks. For instance, Melacci et al. [30] showed that their defenses can be broken by adaptive attacks that also consider the domain knowledge.

Existing provably robust defenses: All existing provably robust defenses [37, 10, 6, 19, 8, 17, 46, 4, 24, 12, 26, 36, 23, 47, 39, 31, 44, 38, 55, 50] were designed for multi-class classification instead of multi-label classification. In particular, they can guarantee that a robust multi-class classifier

predicts the same single label for an input or a label (e.g., the single ground truth label of the input) is among the top-k labels predicted by a robust multi-class classifier. These defenses are sub-optimal for multi-label classification. Specifically, in multi-label classification, we aim to guarantee that at least some ground truth labels of an input are in the set of labels predicted by a robust multi-label classifier.

MultiGuard leverages randomized smoothing [24, 26, 12, 22, 51]. Existing randomized smoothing studies (e.g., Jia et al. [22]) achieve sub-optimal provable robustness guarantees (i.e., certified intersection size) for multi-label classification, because they are designed for multi-class classification. For example, as our empirical evaluation results will show, MultiGuard significantly outperforms Jia et al. [22] when extending it to multi-label classification. Technically speaking, our work has two key differences with Jia et al.. First, the base multi-class classifier in Jia et al. only predicts a single label for an input while our base multi-label classifier predicts multiple labels for an input. Second, Jia et al. can only guarantee that a single label is provably among the k labels predicted by a smoothed multi-class classifier, while we aim to show that multiple labels (e.g., ground truth labels of an input) are provably among the k labels predicted by a smoothed multi-label classifier. Due to such key differences, we require new techniques to derive the certified intersection size of MultiGuard. For instance, we develop a variant of Neyman-Pearson Lemma [33] which is applicable to multiple functions while Jia et al. uses the standard Neyman-Pearson Lemma [33] which is only applicable to a single function. Moreover, we use the law of contraposition to derive our certified intersection size, which is not required by Jia et al..

3 Our MultiGuard

3.1 Building our MultiGuard

Label probability: Suppose we have a multi-label classifier f which we call *base multi-label classifier*. Given an input $\mathbf{x}$, the base multi-label classifier f predicts k' labels for it. For simplicity, we use $f_{k'}(\mathbf{x})$ to denote the set of k' labels predicted by f for $\mathbf{x}$. We use ϵ to denote an isotropic Gaussian noise, i.e., $\epsilon \sim \mathcal{N}(0, \sigma^2 \cdot I)$, where σ is the *standard deviation* and I is an *identity matrix*. Given $\mathbf{x} + \epsilon$ as input, the output of f would be random due to the randomness of ϵ, i.e., $f_{k'}(\mathbf{x} + \epsilon)$ is a random set of k' labels. We define *label probability* p_i as the probability that the label i is among the set of top-k' labels predicted by f when adding isotropic Gaussian noise to an input $\mathbf{x}$, where $i \in \{1, 2, \cdots, c\}$. Formally, we have $p_i = \Pr(i \in f_{k'}(\mathbf{x} + \epsilon))$.

Our smoothed multi-label classifier: Given the label probability p_i's for an input $\mathbf{x}$, our *smoothed multi-label classifier* g predicts the k labels with the largest label probabilities for $\mathbf{x}$. For simplicity, we use $g_k(\mathbf{x})$ to denote the set of k labels predicted by our smoothed multi-label classifier for an input $\mathbf{x}$.

Certified intersection size: An attacker adds a perturbation δ to an input $\mathbf{x}$. $g_k(\mathbf{x} + \delta)$ is the set of k labels predicted by our smoothed multi-label classifier for the perturbed input $\mathbf{x} + \delta$. Given a set of labels $L(\mathbf{x})$ (e.g., the ground truth labels of $\mathbf{x}$), our goal is to show that at least e of them are in the set of k labels predicted by our smoothed multi-label classifier for the perturbed input, when the ℓ_2-norm of the adversarial perturbation is at most R. Formally, we aim to show the following:

$$\min_{\delta, \|\delta\|_2 \leq R} |L(\mathbf{x}) \cap g_k(\mathbf{x} + \delta)| \geq e, \tag{1}$$

where we call e *certified intersection size*. Note that different inputs may have different certified intersection sizes.

3.2 Deriving the Certified Intersection Size

Defining two random variables: Given an input $\mathbf{x}$, we define two random variables $\mathbf{X} = \mathbf{x} + \epsilon$, $\mathbf{Y} = \mathbf{x} + \delta + \epsilon$, where δ is an adversarial perturbation and ϵ is isotropic Gaussian noise. Roughly speaking, the random variables $\mathbf{X}$ and $\mathbf{Y}$ respectively denote the inputs derived by adding isotropic Gaussian noise to the input $\mathbf{x}$ and its adversarially perturbed version $\mathbf{x} + \delta$. Based on the definition of the label probability, we have $p_i = \Pr(i \in f_{k'}(\mathbf{X}))$. We define *adversarial label probability* p_i^* as $p_i^* = \Pr(i \in f_{k'}(\mathbf{Y})), i \in \{1, 2, \cdots, c\}$. Intuitively, adversarial label probability p_i^* is the probability that the label i is in the set of k' labels predicted by the base multi-label classifier f for $\mathbf{Y}$. Given an adversarially perturbed input $\mathbf{x} + \delta$, our smoothed multi-label classifier predicts the k labels with the largest adversarial label probabilities p_i^*'s for it.

Derivation sketch: We leverage the *law of contraposition* in our derivation. Roughly speaking, if we have a statement: $P \longrightarrow Q$, then its contrapositive is: $\neg Q \longrightarrow \neg P$, where $\neg$ is the logical negation symbol. The law of contraposition claims that a statement is true if and only if its contrapositive is true. In particular, we define the following predicate:

$$Q : \min_{\delta, \|\delta\|_2 \leq R} |L(\mathbf{x}) \cap g_k(\mathbf{x} + \delta)| \geq e. \tag{2}$$

Intuitively, Q is true if at least e labels in $L(\mathbf{x})$ can be found in $g_k(\mathbf{x} + \delta)$ for an arbitrary adversarial perturbation δ whose ℓ_2-norm is no larger than R. Then, we have $\neg Q : \min_{\delta, \|\delta\|_2 \leq R} |L(\mathbf{x}) \cap g_k(\mathbf{x} + \delta)| < e$. Moreover, we derive a necessary condition (denoted as $\neg P$) for $\neg Q$ to be true, i.e., $\neg Q \longrightarrow \neg P$. Roughly speaking, $\neg P$ compares upper bounds of the adversarial label probabilities of the labels in $\{1, 2, \cdots, c\} \setminus L(\mathbf{x})$ with lower bounds of those in $L(\mathbf{x})$. More specifically, $\neg P$ represents that the lower bound of the eth largest adversarial label probability of labels in $L(\mathbf{x})$ is no larger than the upper bound of the $(k - e + 1)$th largest adversarial label probability of the labels in $\{1, 2, \cdots, c\} \setminus L(\mathbf{x})$. Finally, based on the law of contraposition, we have $P \longrightarrow Q$, i.e., Q is true if P is true (i.e., $\neg P$ is false).

The major challenges we face when deriving the necessary condition $\neg P$ are as follows: (1) the adversarial perturbation δ can be arbitrary as long as its ℓ_2-norm is no larger than R, which has infinitely many values, and (2) the complexity of the classifier (e.g., a complex deep neural network) and the continuity of the random variable $\mathbf{Y}$ make it hard to compute the adversarial label probabilities. We propose an innovative method to solve the challenges based on two key observations: (1) the random variable $\mathbf{Y}$ reduces to $\mathbf{X}$ under no attacks (i.e., $\delta = \mathbf{0}$) and (2) the adversarial perturbation δ is bounded, i.e., $\|\delta\|_2 \leq R$. Our core idea is to bound the adversarial label probabilities using the label probabilities. Suppose we have the following bounds for the label probabilities (we propose an algorithm to estimate such bounds in Section 3.3):

$$p_i \geq \underline{p}_i, \forall i \in L(\mathbf{x}), \tag{3}$$

$$p_j \leq \overline{p}_j, \forall j \in \{1, 2, \cdots, c\} \setminus L(\mathbf{x}). \tag{4}$$

Given the bounds for label probabilities, we derive a lower bound of the adversarial label probability for each label $i \in L(\mathbf{x})$ and an upper bound of the adversarial label probability for each label $j \in \{1, 2, \cdots, c\} \setminus L(\mathbf{x})$. To derive these bounds, we propose a variant of the Neyman-Pearson Lemma [33] which enables us to consider multiple functions. In contrast, the standard Neyman-Pearson Lemma [33] is insufficient as it is only applicable to a single function while the base multi-label classifier outputs multiple labels.

We give an overview of our derivation of the bounds of the adversarial label probabilities and show the details in the proof of the Theorem 1 in supplementary material. Our idea is to construct some regions in the domain space of $\mathbf{X}$ and $\mathbf{Y}$ via our variant of the Neyman-Pearson Lemma. Specifically, given the constructed regions, we can obtain the lower/upper bounds of the adversarial label probabilities using the probabilities that the random variable $\mathbf{Y}$ is in these regions. Note that the probabilities that the random variables $\mathbf{X}$ and $\mathbf{Y}$ are in these regions can be easily computed as we know their probability density functions.

Next, we derive a lower bound of the adversarial label probability p_i^* ($i \in L(\mathbf{x})$) as an example to illustrate our main idea. Our derivation of the upper bound of the adversarial label probability for a label in $\{1, 2, \cdots, c\} \setminus L(\mathbf{x})$ follows a similar procedure. Given a label $i \in L(\mathbf{x})$, we can find a region $\mathcal{A}_i$ via our variant of Neyman-Pearson Lemma [33] such that $\Pr(\mathbf{X} \in \mathcal{A}_i) = \underline{p}_i$. Then, we can derive a lower bound of p_i^* via computing the probability of the random variable $\mathbf{Y}$ in the region $\mathcal{A}_i$, i.e., we have:

$$p_i^* \geq \Pr(\mathbf{Y} \in \mathcal{A}_i). \tag{5}$$

The above lower bound can be further improved via jointly considering multiple labels in $L(\mathbf{x})$. Suppose we use $\Gamma_u \subseteq L(\mathbf{x})$ to denote an arbitrary set of u labels. We can craft a region $\mathcal{A}_{\Gamma_u}$ via our variant of Neyman-Pearson Lemma such that we have $\Pr(\mathbf{X} \in \mathcal{A}_{\Gamma_u}) = \frac{\sum_{i \in \Gamma_u} \underline{p}_i}{k'}$. Then, we can derive the following lower bound:

$$\max_{i \in \Gamma_u} p_i^* \geq \frac{k'}{u} \cdot \Pr(\mathbf{Y} \in \mathcal{A}_{\Gamma_u}). \tag{6}$$

The eth largest lower bounds of adversarial label probabilities of labels in $L(\mathbf{x})$ can be derived by combing the lower bounds in Equation 5 and 6. Formally, we have the following theorem:

Theorem 1 (Certified Intersection Size)**.** *Suppose we are given an input* $\mathbf{x}$*, a base multi-label classifier* f*, our smoothed classifier* g*, and a set of* d *ground truth labels* $L(\mathbf{x}) = \{a_1, a_2, \cdots, a_d\}$ *for* $\mathbf{x}$*. Moreover, we have a lower bound* $\underline{p_i}$ *of* p_i *for each* $i \in L(\mathbf{x})$ *satisfying Equation 3 and an upper bound* $\overline{p}_j$ *of* p_j *for each* $j \in \{1, 2, \cdots, c\} \setminus L(\mathbf{x})$ *satisfying Equation 4. We assume* $\underline{p_{a_1}} \geq \cdots \geq \underline{p_{a_d}}$ *for convenience. Let* $\overline{p}_{b_1} \geq \overline{p}_{b_2} \geq \cdots \geq \overline{p}_{b_{c-d}}$ *be the* $c-d$ *label probability upper bounds for the labels in* $\{1, 2, \cdots, c\} \setminus L(\mathbf{x})$*, where ties are broken uniformly at random. Given a perturbation size* R*, we have the following guarantee:*

$$\min_{\delta, \|\delta\|_2 \leq R} |L(\mathbf{x}) \cap g_k(\mathbf{x} + \delta)| \geq e, \tag{7}$$

where e *is the optimal solution to the following optimization problem or 0 if it does not have a solution:*

$$\begin{aligned} e = & \operatorname*{argmax}_{e'=1,2,\cdots,\min\{d,k\}} e' \\ s.t. & \max\{\Phi(\Phi^{-1}(\underline{p_{a_{e'}}}) - \frac{R}{\sigma}), \max_{u=1}^{\eta} \frac{k'}{u} \cdot \Phi(\Phi^{-1}(\frac{\underline{p_{A_u}}}{k'}) - \frac{R}{\sigma})\} \\ & > \min\{\Phi(\Phi^{-1}(\overline{p}_{b_s}) + \frac{R}{\sigma}), \min_{v=1}^{s} \frac{k'}{v} \cdot \Phi(\Phi^{-1}(\frac{\overline{p}_{B_v}}{k'}) + \frac{R}{\sigma})\}, \end{aligned} \tag{8}$$

where Φ *and* Φ^{-1} *respectively are the cumulative distribution function and its inverse of the standard Gaussian distribution,* $\eta = d - e' + 1$*,* $\underline{p_{A_u}} = \sum_{l=e'}^{e'+u-1} \underline{p_{a_l}}$*,* $s = k - e' + 1$*, and* $\overline{p}_{B_v} = \sum_{l=s-v+1}^{s} \overline{p}_{b_l}$*.*

Proof. Please refer to Appendix A in supplementary material. □

We have the following remarks for our theorem:

- When $k' = k = 1$ and $L(\mathbf{x})$ only contains a single label, our certified intersection size reduces to the robustness result derived by Cohen et al. [12], i.e., the smoothed classifier provably predicts the same label for an input when the adversarial perturbation is bounded. When $k' = 1$, $k \geq 1$, and $L(\mathbf{x})$ only contains a single label, our certified intersection size reduces to the robustness result derived by Jia et al. [22], i.e., a label is provably among the k labels predicted by a smoothed classifier when the adversarial perturbation is bounded. In other words, the certified robustness guarantees derived by Cohen et al. [12] and Jia et al. [22] are special cases of our results. Note that Cohen et al. is a special case of Jia et al. Moreover, both Cohen et al. and Jia et al. focused on certifying robustness for multi-class classification instead of multi-label classification.
- Our certified intersection size holds for arbitrary attacks as long as the ℓ_2-norm of the adversarial perturbation is no larger than R. Moreover, our results are applicable for any base multi-label classifier.
- Our certified intersection size relies on a lower bound of the label probability for each label $i \in L(\mathbf{x})$ and an upper bound of the label probability for each label $j \in \{1, 2, \cdots, c\} \setminus L(\mathbf{x})$. Moreover, when the label probability bounds are estimated more accurately, our certified intersection size may be larger.
- Our theorem requires $\underline{p_{A_u}} \leq k'$ and $\overline{p}_{B_v} \leq k'$. We have $\underline{p_{A_u}} \leq p_{A_u} \leq \sum_{i \in L(\mathbf{x})} p_i \leq \sum_{j=1}^{c} p_j = k'$. In Section 3.3, our estimated $\overline{p}_{B_v}$ will always be no larger than k'. Thus, we can apply our theorem in practice.
- We note that there are respectively two terms in the left- and right-hand sides of Equation 8. The major technical challenge in our derivation stems from the second term in each side. As we will show in our experiments, those two terms significantly improve certified intersection size.

3.3 Computing the Certified Intersection Size

In order to compute the certified intersection size for an input $\mathbf{x}$, we need to solve the optimization problem in Equation 8. The key challenge of solving the optimization problem is to estimate lower bounds $\underline{p_i}$ of label probabilities for $i \in L(\mathbf{x})$ and upper bounds $\overline{p}_j$ of label probabilities for $j \in \{1, 2, \cdots, c\} \setminus L(\mathbf{x})$. To address the challenge, we design a Monte Carlo algorithm to estimate these label probability bounds with probabilistic guarantees. Then, given the estimated label probability bounds, we solve the optimization problem to obtain the certified intersection size.

Estimating label probability bounds: We randomly sample n Gaussian noise from ϵ and add them to the input $\mathbf{x}$. We use $\mathbf{x}^1, \mathbf{x}^2, \cdots, \mathbf{x}^n$ to denote the n noisy inputs for convenience. Given these noisy inputs, we use the base multi-label classifier f to predict k' labels for each of them. Moreover, we define the *label frequency* n_i of label i as the number of noisy inputs whose predicted k' labels include i. Formally, we have $n_i = \sum_{t=1}^{n} \mathbb{I}(i \in f_{k'}(\mathbf{x}^t)), i \in \{1, 2, \cdots, c\}$, where $\mathbb{I}$ is an indicator function. Based on the definition of label probability p_i, we know that n_i follows a *binomial distribution* with parameters n and p_i, where n is the number of noisy inputs and p_i is the label probability of label i. Our goal is to estimate a lower or upper bound of p_i based on n_i and n, which is a binomial proportion confidence interval estimation problem. Therefore, we can use the Clopper-Pearson [11] method from the statistics community to estimate these label probability bounds. Formally, we have the following label probability bounds:

$$\underline{p_i} = \text{Beta}(\frac{\alpha}{c}; n_i, n - n_i + 1), i \in L(\mathbf{x}), \tag{9}$$

$$\overline{p}_j = \text{Beta}(1 - \frac{\alpha}{c}; n_j, n - n_j + 1), \forall j \in \{1, 2, \cdots, c\} \setminus L(\mathbf{x}), \tag{10}$$

where $1 - \frac{\alpha}{c}$ is the confidence level and $\text{Beta}(\rho; \varsigma, \vartheta)$ is the ρth quantile of the Beta distribution with shape parameters ς and ϑ. Based on *Bonferroni correction* [3, 14], the overall confidence level for the c label probability upper or lower bounds is $1 - \alpha$. To solve the optimization problem in Equation 8, we also need to estimate $\underline{p_{A_u}}$ and $\overline{p}_{B_v}$. In particular, we can estimate $\underline{p_{A_u}} = \sum_{l=e'}^{e'+u-1} \underline{p_{a_l}}$ and $\overline{p}_{B_v} = \sum_{l=s-v+1}^{s} \overline{p}_{b_l}$. However, this bound may be loose for $\overline{p}_{B_v}$. We can further improve the bound via considering the constraint that $\overline{p}_{B_v} + \sum_{i \in L(\mathbf{x})} \underline{p_i} \leq k'$. In other words, we have $\overline{p}_{B_v} \leq k' - \sum_{i \in L(\mathbf{x})} \underline{p_i}$. Given this constraint, we can estimate $\overline{p}_{B_v} = \min(\sum_{l=s-v+1}^{s} \overline{p}_{b_l}, k' - \sum_{i \in L(\mathbf{x})} \underline{p_i})$. Note that the above constraint is not applicable for $\underline{p_{A_u}}$ since it is a lower bound.

Solving the optimization problem in Equation 8: Given the estimated label probability lower or upper bounds, we solve the optimization problem in Equation 8 via binary search.

Complete algorithm: Algorithm 1 in supplementary materials shows our complete algorithm to compute the certified intersection size for an input $\mathbf{x}$. The function RANDOMSAMPLE returns n noisy inputs via first sampling n noise from the isotropic Gaussian distribution and then adding them to the input $\mathbf{x}$. Given the label frequency for each label and the overall confidence α as input, the function PROBBOUNDESTIMATION aims to estimate the label probability bounds based on Equation 9 and 10. The function BINARYSEARCH returns the certified intersection size via solving the optimization problem in Equation 8 using binary search.

4 Evaluation

4.1 Experimental Setup

Datasets: We adopt the following multi-label classification benchmark datasets:

- **VOC 2007 [15]:** Pascal Visual Object Classes Challenge (VOC 2007) dataset [15] contains 9,963 images from 20 objects (i.e., classes). On average, each image has 2.5 objects. Following previous work [43], we split the dataset into 5,011 training images and 4,952 testing images.
- **MS-COCO [28]:** Microsoft-COCO (MS-COCO) [28] dataset contains 82,081 training images, 40,504 validation images, and 40,775 testing images from 80 objects. Each image has 2.9 objects on average. The images in the testing dataset do not have ground truth labels. Therefore, following previous work [7], we evaluate our method on the validation dataset.
- **NUS-WIDE [9]:** NUS-WIDE dataset [9] originally contains 269,648 images from Flickr. The images are manually annotated into 81 visual concepts, with 2.4 visual concepts per image on average. Since the URLs of certain images are not accessible, we adopt the version released by [2], which contains 154,000 training images and 66,000 testing images.

Similar to previous work [12, 25] on certified defenses for multi-class classification, we randomly sample 500 images from the testing (or validation) dataset of each dataset to evaluate our MultiGuard.

Base multi-label classifiers: We adopt ASL [2] to train the base multi-label classifiers on the three benchmark datasets. In particular, ASL leverages an asymmetric loss to solve the positive-negative imbalance issue (an image has a few positive labels while has many negative labels on average) in multi-label classification, and achieves state-of-the-art performance on the three benchmark datasets. Suppose q_j is the probability that a base multi-label classifier predicts label j ($j = 1, 2, \cdots, c$) for a

training input. Moreover, we let y_j be 1 (or 0) if the label j is (or is not) a ground truth label of the training input. The loss of ASL [2] is as follows: $L_{ASL} = \sum_{j=1}^{c} -y_j L_{j+} - (1-y_j)L_{j-}$, where $L_{j+} = (1-q_j)^{\gamma_+}\log(q_j)$ and $L_{j-} = (\max(q_j - m, 0))^{\gamma_-}\log(1-\max(q_j-m,0))$. Note that γ_+, γ_-, and m are hyperparameters. Following [2], we set training hyperameters $\gamma_+ = 0, \gamma_- = 4$, and $m = 0.05$. We train the classifier using Adam optimizer, using learning rate 10^{-3} and batch size 32. We adopt the public implementation of ASL[2] in our experiments. Similar to previous work [12] on randomized smoothing based multi-class classification, we add isotropic Gaussian noise to the training data when we train our base multi-label classifiers. In particular, given a batch of training images, we add isotropic Gaussian noise to each of them, and then we use the noisy training images to update the base multi-label classifier. Our experimental results indicate that such training method can substantially improve the robustness of our MultiGuard (please refer to Figure 6 in supplementary material).

Evaluation metrics: We use *certified top-k precision@R*, *certified top-k recall@R*, and *certified top-k f1-score@R* as evaluation metrics. We first define them for a single testing input and then for multiple testing inputs. In particular, given the certified intersection size e for a testing input $\mathbf{x}$ under perturbation size R, we define them as follows: certified top-k precision@R = e/k, certified top-k recall@$R = e/|L(\mathbf{x})|$, certified top-k f1-score@$R = 2 \cdot e/(|L(\mathbf{x})| + k)$, where $L(\mathbf{x})$ is the set of ground truth labels of $\mathbf{x}$ and the symbol $|\cdot|$ measures the number of elements in a set. Roughly speaking, certified top-k precision@R is the *least* fraction of the k predicted labels for an input that are ground truth labels, when the ℓ_2-norm of the adversarial perturbation is at most R; certified top-k recall@R is the least fraction of the ground truth labels in $L(\mathbf{x})$ that are in the k labels predicted by MultiGuard; and certified top-k f1-score@R measures a trade-off between certified top-k precision@R and certified top-k recall@R. Note that the above definition is for a single testing input. Given a testing/validation dataset with multiple testing inputs, the overall certified top-k precision@R, certified top-k recall@R, and certified top-k f1-score@R are computed as the averages over the testing inputs.

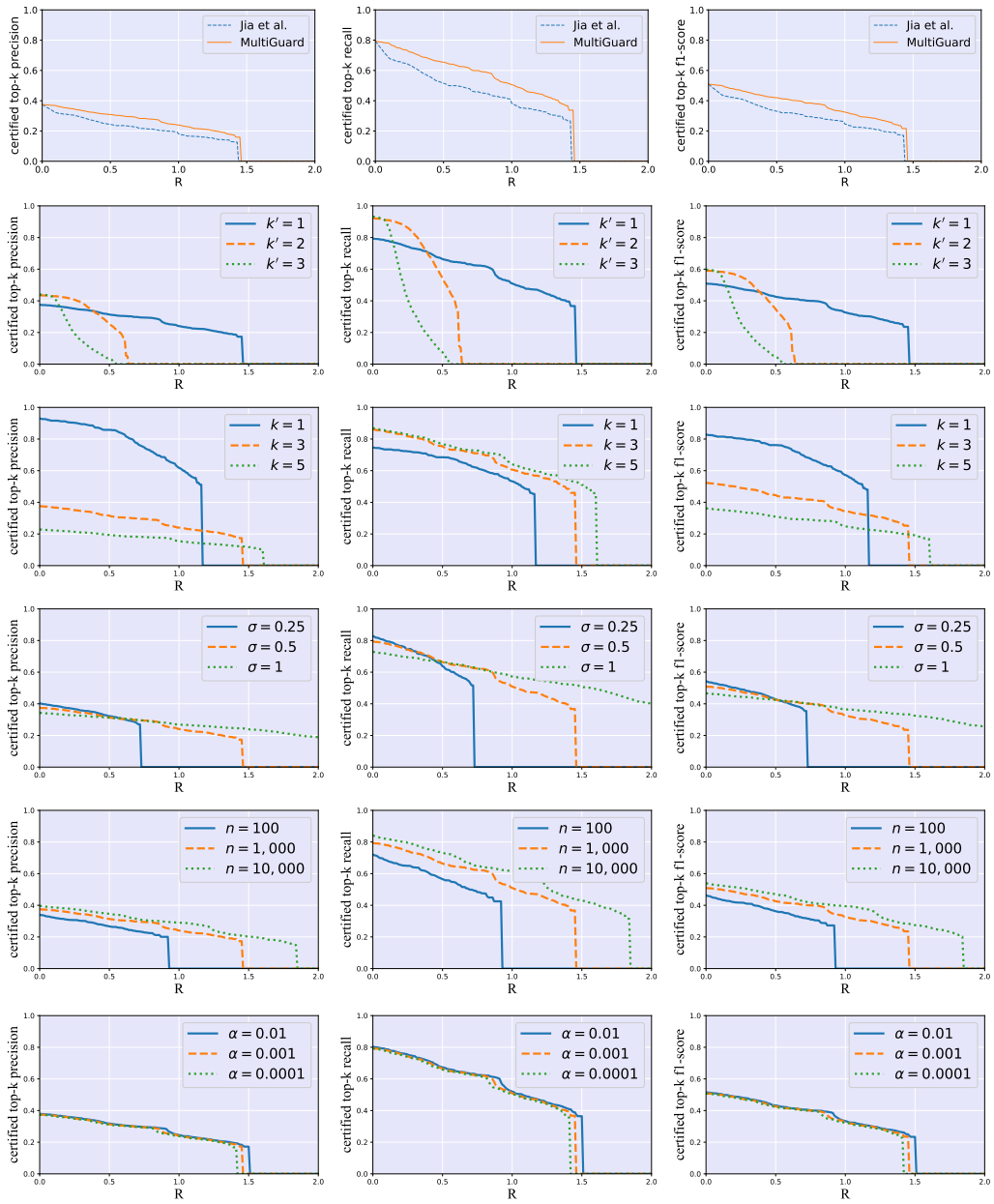
Compared methods: We compare with the state-of-the-art certified defense, namely Jia et al. [22], by extending it to multi-label classification. To the best of our knowledge, Jia et al. is the only certified defense that considers top-k predictions against ℓ_2-norm adversarial perturbations. Given an arbitrary label, Jia et al. derived a certified radius such that the label is among the top-k labels predicted by the smoothed classifier (Theorem 1 in Jia et al.). For each label in $L(\mathbf{x})$, we compute a certified radius. Given a perturbation size R, the certified intersection size e can be computed as the number of labels in $L(\mathbf{x})$ whose certified radii are larger than R.

Parameter setting: Our MultiGuard has the following parameters: k' (the number of labels predicted by the base multi-label classifier for an input), k (the number of labels predicted by our smoothed multi-label classifier for an input), standard deviation σ, number of noisy inputs n, and confidence level $1-\alpha$. Unless otherwise mentioned, we adopt the following default parameters: $\alpha = 0.001$, $n = 1,000$, $\sigma = 0.5$, $k' = 1$ and $k = 3$ for VOC 2007 dataset, and $k' = 3$ and $k = 10$ for MS-COCO and NUS-WIDE datasets, where we use larger k' and k for MS-COCO and NUS-WIDE because they have more classes.

4.2 Experimental Results

Comparison results: The first row in Figure 1 shows the comparison results on VOC 2007 dataset in default setting. We find that MultiGuard achieves higher certified top-k precision@R, certified top-k recall@R, and certified top-k f1-score@R than Jia et al.. MultiGuard is better than Jia et al. because MultiGuard jointly considers all ground truth labels, while Jia et al. can only consider each ground truth label independently. For instance, suppose we have two ground truth labels; it is very likely that both of them are not in the top-k predicted labels when considered independently, but at least one of them is among the top-k predicted labels when considered jointly. The intuition is that it is easier for an attacker to find an adversarial perturbation such that a certain label is not in the top-k predicted labels, but it is more challenging for an attacker to find an adversarial perturbation such that both of the two labels are not in the top-k predicted labels. Our observations on the other two datasets are similar, which can be found in Figure 2 in supplementary material.

[2]https://github.com/Alibaba-MIIL/ASL



(a) Certified top-k precision@R (b) Certified top-k recall@R (c) Certified top-k f1-score@R

Figure 1: Comparing with Jia et al. [22] (first row). Impact of k' (second row), k (third row), σ (fourth row), n (fifth row), and α (sixth row) on certified top-k precision@R, certified top-k recall@R, and certified top-k f1-score@R. The dataset is VOC 2007. The results on the other two datasets are shown in supplementary material.

Impact of k': The second row in Figure 1 shows the impact of k' on VOC 2007 dataset. In particular, we find that a larger k' achieves a larger certified top-k precision@R (or certified top-k recall@R or certified top-k f1-score@R) without attacks (i.e., $R = 0$), but the curves drop more quickly as R increases (i.e., a larger k' is less robust against adversarial examples as R increases). The reason is that a larger k' gives an attacker a larger attack space. Note that this is also reflected in our optimization problem in Equation 8. In particular, the left-hand (or right-hand) side in Equation 8 decreases (or increases) as k' increases, which leads to smaller certified intersection size e as k' increases. We have similar observations on MS-COCO and NUS-WIDE datasets. Please refer to Figure 3 in supplementary material.

Impact of k: The third row in Figure 1 shows the impact of k on VOC 2007 dataset. We have the following observations from our experimental results. First, k achieves a tradeoff between the certified top-k precision@R without attacks and robustness. In particular, a larger k gives us a smaller certified top-k precision@R without attacks, but the curve drops more slowly as R increases (i.e., a larger k is more robust against adversarial examples as R increases). Second, we find that certified top-k recall@R increases as k increases. The reason is that more labels are predicted by our MultiGuard as k increases. Third, similar to certified top-k precision@R, k also achieves a tradeoff between the certified top-k f1-score@R without attacks and robustness. We also have those observations on the other two datasets. Please refer to Figure 4 in supplementary material for details.

Impact of σ: The fourth row in Figure 1 shows the impact of σ on VOC 2007 dataset. The experimental results indicate that σ achieves a tradeoff between the certified top-k precision@R (or certified top-k recall@R or certified top-k f1-score@R) without attacks (i.e., $R = 0$) and robustness. Specifically, a larger σ leads to a smaller certified top-k precision@R (or certified top-k recall@R or certified top-k f1-score@R) without attacks, but is more robust against adversarial examples as R increases. The observations on the other two datasets are similar. Please refer to Figure 5 in supplementary material.

Impact of n and α: The fifth and sixth rows in Figure 1 respectively show the impact of n and α on VOC 2007 dataset. In particular, we find that certified top-k precision@R (or certified top-k recall@R or certified top-k f1-score@R) increases as n or α increases. The reason is that a larger n or α gives us tighter lower or upper bounds of the label probabilities, which leads to larger certified intersection sizes. However, we find that the certified top-k precision@R (or certified top-k recall@R or certified top-k f1-score@R) is insensitive to α once it is small enough and insensitive to n once it is large enough.

Effectiveness of the second terms in Equation 8: We perform experiments under our default setting to validate the effectiveness of our second terms in the left- and right-hand sides of Equation 8. Our results are as follows: with and without the second terms, MultiGuard respectively achieves 31.3% and 23.6% certified top-k precision@R, 66.4% and 48.8% certified top-k recall@R, as well as 42.6% and 31.8% certified top-k f1-score@R, where the perturbation size $R = 0.5$ and the dataset is VOC 2007. As the result shows, our second terms can significantly improve certified intersection size.

5 Conclusion

In this paper, we propose MultiGuard, the first provably robust defense against adversarial examples for multi-label classification. In particular, we show that a certain number of ground truth labels of an input are provably predicted by our MultiGuard when the ℓ_2-norm of the adversarial perturbation added to the input is bounded. Moreover, we design an algorithm to compute the certified robustness guarantees. Empirically, we conduct experiments on VOC 2007, MS-COCO, and NUS-WIDE benchmark datasets to validate our MultiGuard. Interesting future work to improve MultiGuard includes: 1) incorporating the knowledge of the base multi-label classifier, and 2) designing new methods to train more accurate base multi-label classifiers.

Acknowledgements: We thank the anonymous reviewers for constructive comments. This work was supported by NSF under Grant No. 1937786 and 2125977 and the Army Research Office under Grant No. W911NF2110182.

References

[1] Rohit Babbar and Bernhard Schölkopf. Adversarial extreme multi-label classification. *arXiv preprint arXiv:1803.01570*, 2018.

[2] Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification, 2020.

[3] Carlo Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62, 1936.

[4] Rudy R Bunel, Ilker Turkaslan, Philip Torr, Pushmeet Kohli, and Pawan K Mudigonda. A unified view of piecewise linear neural network verification. In *NeurIPS*, 2018.

[5] Xiaoyu Cao and Neil Zhenqiang Gong. Mitigating evasion attacks to deep neural networks via region-based classification. In *Proceedings of the 33rd Annual Computer Security Applications Conference*, pages 278–287, 2017.

[6] Nicholas Carlini, Guy Katz, Clark Barrett, and David L Dill. Provably minimally-distorted adversarial examples. *arXiv*, 2017.

[7] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5177–5186, 2019.

[8] Chih-Hong Cheng, Georg Nührenberg, and Harald Ruess. Maximum resilience of artificial neural networks. In *International Symposium on Automated Technology for Verification and Analysis*, pages 251–268. Springer, 2017.

[9] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: A real-world web image database from national university of singapore. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, CIVR '09, New York, NY, USA, 2009. Association for Computing Machinery.

[10] Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *International Conference on Machine Learning*, pages 854–863. PMLR, 2017.

[11] Charles J Clopper and Egon S Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

[12] Jeremy M Cohen, Elan Rosenfeld, and J Zico Kolter. Certified adversarial robustness via randomized smoothing. *arXiv preprint arXiv:1902.02918*, 2019.

[13] Son D Dao, Ethan Zhao, Dinh Phung, and Jianfei Cai. Multi-label image classification with contrastive learning. *arXiv preprint arXiv:2107.11626*, 2021.

[14] Olive Jean Dunn. Multiple comparisons among means. *Journal of the American statistical association*, 56(293):52–64, 1961.

[15] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[16] Zongyuan Ge, Dwarikanath Mahapatra, Suman Sedai, Rahil Garnavi, and Rajib Chakravorty. Chest x-rays classification: A multi-label and fine-grained problem. *arXiv preprint arXiv:1807.07247*, 2018.

[17] Timon Gehr, Matthew Mirman, Dana Drachsler-Cohen, Petar Tsankov, Swarat Chaudhuri, and Martin Vechev. Ai2: Safety and robustness certification of neural networks with abstract interpretation. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2018.

[18] Marian George and Christian Floerkemeier. Recognizing products: A per-exemplar multi-label image classification approach. In *European Conference on Computer Vision*, pages 440–455. Springer, 2014.

[19] Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715*, 2018.

[20] Shu Hu, Lipeng Ke, Xin Wang, and Siwei Lyu. Tkml-ap: Adversarial attacks to top-k multi-label learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7649–7657, 2021.

[21] Dat Huynh and Ehsan Elhamifar. Interactive multi-label cnn learning with partial labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9423–9432, 2020.

[22] Jinyuan Jia, Xiaoyu Cao, Binghui Wang, and Neil Zhenqiang Gong. Certified robustness for top-k predictions against adversarial perturbations via randomized smoothing. In *International Conference on Learning Representations*, 2020.

[23] Jinyuan Jia, Binghui Wang, Xiaoyu Cao, Hongbin Liu, and Neil Zhenqiang Gong. Almost tight l0-norm certified robustness of top-k predictions against adversarial perturbations. In *International Conference on Learning Representations*, 2022.

[24] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 656–672. IEEE, 2019.

[25] Guang-He Lee, Yang Yuan, Shiyu Chang, and Tommi Jaakkola. Tight certificates of adversarial robustness for randomly smoothed classifiers. In *Advances in Neural Information Processing Systems*, pages 4910–4921, 2019.

[26] Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Certified adversarial robustness with additive noise. In *Advances in Neural Information Processing Systems*, pages 9464–9474, 2019.

[27] Xin Li, Feipeng Zhao, and Yuhong Guo. Multi-label image classification with a probabilistic label enhancement model. In *UAI*, volume 1, pages 1–10, 2014.

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and Larry Zitnick. Microsoft coco: Common objects in context. In *ECCV*. European Conference on Computer Vision, September 2014.

[29] Xuanqing Liu, Minhao Cheng, Huan Zhang, and Cho-Jui Hsieh. Towards robust neural networks via random self-ensemble. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 369–385, 2018.

[30] Stefano Melacci, Gabriele Ciravegna, Angelo Sotgiu, Ambra Demontis, Battista Biggio, Marco Gori, and Fabio Roli. Can domain knowledge alleviate adversarial attacks in multi-label classifiers? *arXiv preprint arXiv:2006.03833*, 2020.

[31] Matthew Mirman, Timon Gehr, and Martin Vechev. Differentiable abstract interpretation for provably robust neural networks. In *International Conference on Machine Learning*, pages 3578–3586, 2018.

[32] Jinseok Nam, Eneldo Loza Mencía, Hyunwoo J Kim, and Johannes Fürnkranz. Maximizing subset accuracy with recurrent neural networks in multi-label classification. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 5419–5429, 2017.

[33] Jerzy Neyman and Egon Sharpe Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London.*, 231(694-706):289–337, 1933.

[34] Ioannis Partalas, Aris Kosmopoulos, Nicolas Baskiotis, Thierry Artieres, George Paliouras, Eric Gaussier, Ion Androutsopoulos, Massih-Reza Amini, and Patrick Galinari. Lshtc: A benchmark for large-scale text classification. *arXiv preprint arXiv:1503.08581*, 2015.

[35] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 254–269. Springer, 2009.

[36] Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems*, pages 11292–11303, 2019.

[37] Karsten Scheibler, Leonore Winterer, Ralf Wimmer, and Bernd Becker. Towards verification of artificial neural networks. In *MBMV*, 2015.

[38] Gagandeep Singh, Timon Gehr, Matthew Mirman, Markus Püschel, and Martin T Vechev. Fast and effective robustness certification. *NeurIPS*, 1(4):6, 2018.

[39] Gagandeep Singh, Timon Gehr, Markus Püschel, and Martin Vechev. An abstract domain for certifying neural networks. *Proceedings of the ACM on Programming Languages*, 3(POPL):1–30, 2019.

[40] Qingquan Song, Haifeng Jin, Xiao Huang, and Xia Hu. Multi-label adversarial perturbations. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 1242–1247. IEEE, 2018.

[41] Konstantinos Trohidis, Grigorios Tsoumakas, George Kalliris, and Ioannis P Vlahavas. Multi-label classification of music into emotions. In *ISMIR*, volume 8, pages 325–330, 2008.

[42] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007.

[43] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn: A unified framework for multi-label image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2285–2294, 2016.

[44] Shiqi Wang, Kexin Pei, Justin Whitehouse, Junfeng Yang, and Suman Jana. Efficient formal safety analysis of neural networks. *arXiv preprint arXiv:1809.08098*, 2018.

[45] Zhouxia Wang, Tianshui Chen, Guanbin Li, Ruijia Xu, and Liang Lin. Multi-label image recognition by recurrently discovering attentional regions. In *Proceedings of the IEEE international conference on computer vision*, pages 464–472, 2017.

[46] Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, pages 5286–5295. PMLR, 2018.

[47] Eric Wong, Frank R Schmidt, Jan Hendrik Metzen, and J Zico Kolter. Scaling provable adversarial defenses. *arXiv preprint arXiv:1805.12514*, 2018.

[48] Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *European Conference on Computer Vision*, pages 162–178. Springer, 2020.

[49] Yi Wu, David Bamman, and Stuart Russell. Adversarial training for relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1778–1783, 2017.

[50] Kai Y Xiao, Vincent Tjeng, Nur Muhammad Shafiullah, and Aleksander Madry. Training for faster adversarial robustness verification via inducing relu stability. *arXiv preprint arXiv:1809.03008*, 2018.

[51] Greg Yang, Tony Duan, J Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. Randomized smoothing of all shapes and sizes. In *International Conference on Machine Learning*, 2020.

[52] Hao Yang, Joey Tianyi Zhou, Yu Zhang, Bin-Bin Gao, Jianxin Wu, and Jianfei Cai. Exploit bounding box annotations for multi-label object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 280–288, 2016.

[53] Zhuo Yang, Yufei Han, and Xiangliang Zhang. Characterizing the evasion attackability of multi-label classifiers. In *AAAI*, 2021.

[54] Renchun You, Zhiyao Guo, Lei Cui, Xiang Long, Yingze Bao, and Shilei Wen. Cross-modality attention with semantic graph embedding for multi-label classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12709–12716, 2020.

[55] Huan Zhang, Hongge Chen, Chaowei Xiao, Sven Gowal, Robert Stanforth, Bo Li, Duane Boning, and Cho-Jui Hsieh. Towards stable and efficient training of verifiably robust neural networks. *arXiv preprint arXiv:1906.06316*, 2019.

[56] Nan Zhou, Wenjian Luo, Xin Lin, Peilan Xu, and Zhenya Zhang. Generating multi-label adversarial examples by linear programming. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[57] Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang. Learning spatial regularization with image-level supervisions for multi-label image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5513–5522, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] We have formal guarantees for our experimental results
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Proof of Theorem 1

We first define some notations that will be used in our proof. Given an input $\mathbf{x}$, we define the following two random variables:

$$\mathbf{X} = \mathbf{x} + \epsilon \sim \mathcal{N}(\mathbf{x}, \sigma^2 I), \tag{11}$$

$$\mathbf{Y} = \mathbf{x} + \delta + \epsilon \sim \mathcal{N}(\mathbf{x} + \delta, \sigma^2 I), \tag{12}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ and δ is an adversarial perturbation that has the same size with $\mathbf{x}$. The random variables $\mathbf{X}$ and $\mathbf{Y}$ represent random inputs obtained by adding isotropic Gaussian noise to the input $\mathbf{x}$ and its perturbed version $\mathbf{x} + \delta$, respectively. Cohen et al. [12] applied the standard Neyman-Pearson lemma [33] to the above two random variables, and obtained the following two lemmas:

Lemma 1 (Neyman-Pearson lemma for Gaussian with different means)**.** *Let $\mathbf{X} \sim \mathcal{N}(\mathbf{x}, \sigma^2 I)$, $\mathbf{Y} \sim \mathcal{N}(\mathbf{x} + \delta, \sigma^2 I)$, and $F : \mathbb{R}^d \to \{0, 1\}$ be a random or deterministic function. Then, we have the following:*

(1) If $W = \{\mathbf{w} \in \mathbb{R}^d : \delta^T \mathbf{w} \leq \beta\}$ for some β and $Pr(F(\mathbf{X}) = 1) \geq Pr(\mathbf{X} \in W)$, then $Pr(F(\mathbf{Y}) = 1) \geq Pr(\mathbf{Y} \in W)$.

(2) If $W = \{\mathbf{w} \in \mathbb{R}^d : \delta^T \mathbf{w} \geq \beta\}$ for some β and $Pr(F(\mathbf{X}) = 1) \leq Pr(\mathbf{X} \in W)$, then $Pr(F(\mathbf{Y}) = 1) \leq Pr(\mathbf{Y} \in W)$.

Lemma 2. *Given an input $\mathbf{x}$, a real number $q \in [0, 1]$, as well as regions $\mathcal{A}$ and $\mathcal{B}$ defined as follows:*

$$\mathcal{A} = \{\mathbf{w} : \delta^T (\mathbf{w} - \mathbf{x}) \leq \sigma \left\|\delta\right\|_2 \Phi^{-1}(q)\}, \tag{13}$$

$$\mathcal{B} = \{\mathbf{w} : \delta^T (\mathbf{w} - \mathbf{x}) \geq \sigma \left\|\delta\right\|_2 \Phi^{-1}(1 - q)\}, \tag{14}$$

we have the following:

$$Pr(\mathbf{X} \in \mathcal{A}) = q, \tag{15}$$

$$Pr(\mathbf{X} \in \mathcal{B}) = q, \tag{16}$$

$$Pr(\mathbf{Y} \in \mathcal{A}) = \Phi(\Phi^{-1}(q) - \frac{\left\|\delta\right\|_2}{\sigma}), \tag{17}$$

$$Pr(\mathbf{Y} \in \mathcal{B}) = \Phi(\Phi^{-1}(q) + \frac{\left\|\delta\right\|_2}{\sigma}). \tag{18}$$

Proof. Please refer to [12]. □

Next, we first generalize the Neyman-Pearson lemma to the case of multiple functions and then derive the lemmas that will be used in our proof.

Lemma 3. *Let $\mathbf{X}$, $\mathbf{Y}$ be two random variables whose probability densities are respectively $Pr(\mathbf{X} = \mathbf{w})$ and $Pr(\mathbf{Y} = \mathbf{w})$, where $\mathbf{w} \in \mathbb{R}^d$. Let $F_1, F_2, \cdots, F_t : \mathbb{R}^d \to \{0, 1\}$ be t random or deterministic functions. Let k' be an integer such that:*

$$\sum_{i=1}^{t} F_i(1|\mathbf{w}) \leq k', \forall \mathbf{w} \in \mathbb{R}^d, \tag{19}$$

where $F_i(1|\mathbf{w})$ denotes the probability that $F_i(\mathbf{w}) = 1$. Then, we have the following:

(1) If $W = \{\mathbf{w} \in \mathbb{R}^d : Pr(\mathbf{Y} = \mathbf{w})/Pr(\mathbf{X} = \mathbf{w}) \leq \mu\}$ for some $\mu > 0$ and $\frac{\sum_{i=1}^{t} Pr(F_i(\mathbf{X})=1)}{k'} \geq Pr(\mathbf{X} \in W)$, then $\frac{\sum_{i=1}^{t} Pr(F_i(\mathbf{Y})=1)}{k'} \geq Pr(\mathbf{Y} \in W)$.

(2) If $W = \{\mathbf{w} \in \mathbb{R}^d : Pr(\mathbf{Y} = \mathbf{w})/Pr(\mathbf{X} = \mathbf{w}) \geq \mu\}$ for some $\mu > 0$ and $\frac{\sum_{i=1}^{t} Pr(F_i(\mathbf{X})=1)}{k'} \leq Pr(\mathbf{X} \in W)$, then $\frac{\sum_{i=1}^{t} Pr(F_i(\mathbf{Y})=1)}{k'} \leq Pr(\mathbf{Y} \in W)$.

Proof. We first prove part (1). For convenience, we denote the complement of W as W^c. Then, we have the following:

$$\frac{\sum_{i=1}^{t} \Pr(F_i(\mathbf{Y}) = 1)}{k'} - \Pr(\mathbf{Y} \in W) \tag{20}$$

$$= \int_{\mathbb{R}^d} \frac{\sum_{i=1}^t F_i(1|\mathbf{w})}{k'} \cdot \Pr(\mathbf{Y}=\mathbf{w}) d\mathbf{w} - \int_W \Pr(\mathbf{Y}=\mathbf{w}) d\mathbf{w} \tag{21}$$

$$= \int_{W^c} \frac{\sum_{i=1}^t F_i(1|\mathbf{w})}{k'} \cdot \Pr(\mathbf{Y}=\mathbf{w}) d\mathbf{w} + \int_W \frac{\sum_{i=1}^t F_i(1|\mathbf{w})}{k'} \cdot \Pr(\mathbf{Y}=\mathbf{w}) d\mathbf{w} - \int_W \Pr(\mathbf{Y}=\mathbf{w}) d\mathbf{w} \tag{22}$$

$$= \int_{W^c} \frac{\sum_{i=1}^t F_i(1|\mathbf{w})}{k'} \cdot \Pr(\mathbf{Y}=\mathbf{w}) d\mathbf{w} - \int_W (1 - \frac{\sum_{i=1}^t F_i(1|\mathbf{w})}{k'}) \cdot \Pr(\mathbf{Y}=\mathbf{w}) d\mathbf{w} \tag{23}$$

$$\geq \mu \cdot [\int_{W^c} \frac{\sum_{i=1}^t F_i(1|\mathbf{w})}{k'} \cdot \Pr(\mathbf{X}=\mathbf{w}) d\mathbf{w} - \int_W (1 - \frac{\sum_{i=1}^t F_i(1|\mathbf{w})}{k'}) \cdot \Pr(\mathbf{X}=\mathbf{w}) d\mathbf{w}] \tag{24}$$

$$= \mu \cdot [\int_{W^c} \frac{\sum_{i=1}^t F_i(1|\mathbf{w})}{k'} \cdot \Pr(\mathbf{X}=\mathbf{w}) d\mathbf{w} + \int_W \frac{\sum_{i=1}^t F_i(1|\mathbf{w})}{k'} \cdot \Pr(\mathbf{X}=\mathbf{w}) d\mathbf{w} - \int_W \Pr(\mathbf{X}=\mathbf{w}) d\mathbf{w}] \tag{25}$$

$$= \mu \cdot [\int_{\mathbb{R}^d} \frac{\sum_{i=1}^t F_i(1|\mathbf{w})}{k'} \cdot \Pr(\mathbf{X}=\mathbf{w}) d\mathbf{w} - \int_W \Pr(\mathbf{X}=\mathbf{w}) d\mathbf{w}] \tag{26}$$

$$= \mu \cdot [\frac{\sum_{i=1}^t \Pr(F_i(\mathbf{X})=1)}{k'} - \Pr(\mathbf{X} \in W)] \tag{27}$$

$$\geq 0. \tag{28}$$

We have Equation 24 from 23 due to the fact that $\Pr(\mathbf{Y}=\mathbf{w})/\Pr(\mathbf{X}=\mathbf{w}) \leq \mu, \forall \mathbf{w} \in W$, $\Pr(\mathbf{Y}=\mathbf{w})/\Pr(\mathbf{X}=\mathbf{w}) > \mu, \forall \mathbf{w} \in W^c$, and $1 - \frac{\sum_{i=1}^t F_i(1|\mathbf{w})}{k'} \geq 0$. Similarly, we can prove the part (2). We omit the details for conciseness reason. □

We apply the above lemma to random variables $\mathbf{X}$ and $\mathbf{Y}$, and obtain the following lemma:

Lemma 4. *Let $\mathbf{X} \sim \mathcal{N}(\mathbf{x}, \sigma^2 I)$, $\mathbf{Y} \sim \mathcal{N}(\mathbf{x}+\delta, \sigma^2 I)$, $F_1, F_2, \cdots, F_t : \mathbb{R}^d \to \{0,1\}$ be t random or deterministic functions, and k' be an integer such that:*

$$\sum_{i=1}^t F_i(1|\mathbf{w}) \leq k', \forall \mathbf{w} \in \mathbb{R}^d, \tag{29}$$

where $F_i(1|\mathbf{w})$ denote the probability that $F_i(\mathbf{w}) = 1$. Then, we have the following:

(1) If $W = \{\mathbf{w} \in \mathbb{R}^d : \delta^T \mathbf{w} \leq \beta\}$ for some β and $\frac{\sum_{i=1}^t Pr(F_i(\mathbf{X})=1)}{k'} \geq Pr(\mathbf{X} \in W)$, then $\frac{\sum_{i=1}^t Pr(F_i(\mathbf{Y})=1)}{k'} \geq Pr(\mathbf{Y} \in W)$.

(2) If $W = \{\mathbf{w} \in \mathbb{R}^d : \delta^T \mathbf{w} \geq \beta\}$ for some β and $\frac{\sum_{i=1}^t Pr(F_i(\mathbf{X})=1)}{k'} \leq Pr(\mathbf{X} \in W)$, then $\frac{\sum_{i=1}^t Pr(F_i(\mathbf{Y})=1)}{k'} \leq Pr(\mathbf{Y} \in W)$.

By leveraging Lemma 2, Lemma 3, and Lemma 4, we derive the following lemma:

Lemma 5. *Suppose we have an arbitrary base multi-label classifier f, an integer k', an input $\mathbf{x}$, an arbitrary set denoted as O, two label probability bounds $\underline{p_O}$ and $\overline{p}_O$ that satisfy $\underline{p_O} \leq p_O = \sum_{i \in O} Pr(i \in f_{k'}(\mathbf{X})) \leq \overline{p}_O$, as well as regions $\mathcal{A}_O$ and $\mathcal{B}_O$ defined as follows:*

$$\mathcal{A}_O = \{\mathbf{w} : \delta^T(\mathbf{w} - \mathbf{x}) \leq \sigma \|\delta\|_2 \Phi^{-1}(\frac{\underline{p_O}}{k'})\} \tag{30}$$

$$\mathcal{B}_O = \{\mathbf{w} : \delta^T(\mathbf{w} - \mathbf{x}) \geq \sigma \|\delta\|_2 \Phi^{-1}(1 - \frac{\overline{p}_O}{k'})\} \tag{31}$$

Then, we have:

$$Pr(\mathbf{X} \in \mathcal{A}_O) \leq \frac{\sum_{i \in O} Pr(i \in f_{k'}(\mathbf{X}))}{k'} \leq Pr(\mathbf{X} \in \mathcal{B}_O) \tag{32}$$

$$Pr(\mathbf{Y} \in \mathcal{A}_O) \leq \frac{\sum_{i \in O} Pr(i \in f_{k'}(\mathbf{Y}))}{k'} \leq Pr(\mathbf{Y} \in \mathcal{B}_O) \tag{33}$$

Proof. We know $\Pr(\mathbf{X} \in \mathcal{A}_O) = \frac{p_O}{k'}$ based on Lemma 2. Moreover, based on the condition $\underline{p_O} \leq \sum_{i \in O} \Pr(i \in f_{k'}(\mathbf{X}))$, we obtain the first inequality in Equation 32. Similarly, we can obtain the second inequality in Equation 32. We define $F_i(\mathbf{w}) = \mathbb{I}(i \in f_{k'}(\mathbf{w})), \forall i \in O$, where $\mathbb{I}$ is indicator function. Then, we have $\Pr(\mathbf{X} \in \mathcal{A}_O) \leq \frac{\sum_{i \in O} \Pr(i \in f_{k'}(\mathbf{X}))}{k'} = \frac{\sum_{i \in O} \Pr(F_i(\mathbf{X})=1)}{k'}$. Note that there are k' elements in $f_{k'}(\mathbf{w}), \forall \mathbf{w} \in \mathbb{R}^d$, therefore, we have $\sum_{i \in O} F_i(1|\mathbf{w}) = \sum_{i \in O} \mathbb{I}(i \in f_{k'}(\mathbf{w})) \leq k', \forall \mathbf{w} \in \mathbb{R}^d$. Then, we can apply Lemma 4 and we have the following:

$$\Pr(\mathbf{Y} \in \mathcal{A}_O) \leq \frac{\sum_{i \in O} \Pr(F_i(\mathbf{Y}) = 1)}{k'} = \frac{\sum_{i \in O} \Pr(i \in f_{k'}(\mathbf{Y}))}{k'}, \tag{34}$$

which is the first inequality in Equation 33. Similarly, we can obtain the second inequality in Equation 33. □

Based on Lemma 1 and Lemma 2, we derive the following lemma:

Lemma 6. *Suppose we have an arbitrary base multi-label classifier f, an integer k', an input $\mathbf{x}$, an arbitrary label which is denoted as l, two label probability bounds $\underline{p_l}$ and $\overline{p}_l$ that satisfy $\underline{p_l} \leq p_l = Pr(l \in f_{k'}(\mathbf{X})) \leq \overline{p}_l$, and regions $\mathcal{A}_l$ and $\mathcal{B}_l$ defined as follows:*

$$\mathcal{A}_l = \{\mathbf{w} : \delta^T(\mathbf{w} - \mathbf{x}) \leq \sigma \left\|\delta\right\|_2 \Phi^{-1}(\underline{p_l})\} \tag{35}$$

$$\mathcal{B}_l = \{\mathbf{w} : \delta^T(\mathbf{w} - \mathbf{x}) \geq \sigma \left\|\delta\right\|_2 \Phi^{-1}(1 - \overline{p}_l)\} \tag{36}$$

Then, we have:

$$Pr(\mathbf{X} \in \mathcal{A}_l) \leq Pr(l \in f_{k'}(\mathbf{X})) \leq Pr(\mathbf{X} \in \mathcal{B}_l) \tag{37}$$

$$Pr(\mathbf{Y} \in \mathcal{A}_l) \leq Pr(l \in f_{k'}(\mathbf{Y})) \leq Pr(\mathbf{Y} \in \mathcal{B}_l) \tag{38}$$

Proof. We know $\Pr(\mathbf{X} \in \mathcal{A}_l) = \underline{p_l}$ based on Lemma 2. Moreover, based on the condition $\underline{p_l} \leq \Pr(l \in f_{k'}(\mathbf{X}))$, we obtain the first inequality in Equation 37. Similarly, we can obtain the second inequality in Equation 37. We define $F(\mathbf{w}) = \mathbb{I}(l \in f_{k'}(\mathbf{w}))$. Based on the first inequality in Equation 37, we know $\Pr(\mathbf{X} \in \mathcal{A}_l) \leq \Pr(l \in f_{k'}(\mathbf{X})) = \Pr(F(\mathbf{X}) = 1)$. Then, we apply Lemma 1 and we have the following:

$$\Pr(\mathbf{Y} \in \mathcal{A}_l) \leq \Pr(F(\mathbf{Y}) = 1) = \Pr(l \in f_{k'}(\mathbf{Y})), \tag{39}$$

which is the first inequality in Equation 38. The second inequality in Equation 38 can be obtained similarly. □

Next, we formally show our proof for Theorem 1.

Proof. We leverage the law of contraposition to prove our theorem. Roughly speaking, if we have a statement: $P \rightarrow Q$, then, it's contrapositive is: $\neg Q \rightarrow \neg P$, where $\neg$ denotes negation. The law of contraposition claims that a statement is true if, and only if, its contrapositive is true. We define the predicate P as follows:

$$\begin{aligned} &\max\{\Phi(\Phi^{-1}(\underline{p_{a_e}}) - \frac{R}{\sigma}), \max_{u=1}^{d-e+1} \frac{k'}{u} \cdot \Phi(\Phi^{-1}(\frac{\underline{p_{\Gamma_u}}}{k'}) - \frac{R}{\sigma})\} \\ > &\min\{\Phi(\Phi^{-1}(\overline{p}_{b_s}) + \frac{R}{\sigma}), \max_{v=1}^{k-e+1} \frac{k'}{v} \cdot \Phi(\Phi^{-1}(\frac{\overline{p}_{\Lambda_v}}{k'}) + \frac{R}{\sigma})\}. \end{aligned} \tag{40}$$

We define the predicate Q as follows:

$$\min_{\delta, \|\delta\|_2 \leq R} |L(\mathbf{x}) \cap g_k(\mathbf{x} + \delta)| \geq e. \tag{41}$$

We will first prove the statement: $P \rightarrow Q$. To prove it, we consider its contrapositive, i.e., we prove the following statement: $\neg Q \rightarrow \neg P$.

Deriving necessary condition: Suppose $\neg Q$ is true, i.e., $\min_{\delta, \|\delta\|_2 \leq R} |L(\mathbf{x}) \cap g_k(\mathbf{x} + \delta)| < e$. On the one hand, this means there exist at least $d - e + 1$ elements in $L(\mathbf{x})$ do not appear in $g_k(\mathbf{x} + \delta)$. For convenience, we use $\mathcal{U}_r \subseteq L(\mathbf{x})$ to denote those elements, a subset of $L(\mathbf{x})$ with r elements where $r = d - e + 1$. On the other hand, there exist at least $k - e + 1$ elements in $\{1, 2, \cdots, c\} \setminus L(\mathbf{x})$ appear

in $g_k(\mathbf{x}+\delta)$. We use $\mathcal{V}_s \subseteq \{1,2,\cdots,c\} \setminus L(\mathbf{x})$ to denote them, a subset of $\{1,2,\cdots,c\} \setminus L(\mathbf{x})$ with $s = k - e + 1$ elements. Formally, we have the following:

$$\exists\, \mathcal{U}_r \subseteq L(\mathbf{x}), \mathcal{U}_r \cap g_k(\mathbf{x}+\delta) = \emptyset \tag{42}$$

$$\exists\, \mathcal{V}_s \subseteq \{1,2,\cdots,c\} \setminus L(\mathbf{x}), \mathcal{V}_s \subseteq g_k(\mathbf{x}+\delta), \tag{43}$$

In other words, there exist sets $\mathcal{U}_r$ and $\mathcal{V}_s$ such that the adversarially perturbed label probability p_i^*'s for elements in $\mathcal{V}_s$ are no smaller than these for the elements in $\mathcal{U}_r$. Formally, we have the following necessary condition if $|L(\mathbf{x}) \cap g_k(\mathbf{x}+\delta)| < e$:

$$\min_{\mathcal{U}_r} \max_{i \in \mathcal{U}_r} \Pr(i \in f_{k'}(\mathbf{Y})) \leq \max_{\mathcal{V}_s} \min_{j \in \mathcal{V}_s} \Pr(j \in f_{k'}(\mathbf{Y})) \tag{44}$$

Bounding $\max_{i \in \mathcal{U}_r}$ Pr$(i \in f_{k'}(\mathbf{Y}))$ and $\min_{j \in \mathcal{V}_s}$ Pr$(j \in f_{k'}(\mathbf{Y}))$ for given $\mathcal{U}_r$ and $\mathcal{V}_s$: For simplicity, we assume $\mathcal{U}_r = \{w_1, w_2, \cdots, w_r\}$. Without loss of generality, we assume $\underline{p_{w_1}} \geq \underline{p_{w_2}} \geq \cdots \geq \underline{p_{w_r}}$. Similarly, we assume $\mathcal{V}_s = \{z_1, z_2, \cdots, z_s\}$ and $\overline{p}_{z_s} \geq \cdots \geq \overline{p}_{z_2} \geq \overline{p}_{z_1}$. For an arbitrary element $i \in \mathcal{U}_r$, we define the following region:

$$\mathcal{A}_i = \{\mathbf{w} : \delta^T(\mathbf{w} - \mathbf{x}) \leq \sigma \left\|\delta\right\|_2 \Phi^{-1}(\underline{p_i})\} \tag{45}$$

Then, we have the following for any $i \in \mathcal{U}_r$:

$$\Pr(i \in f_{k'}(\mathbf{Y})) \geq \Pr(\mathbf{Y} \in \mathcal{A}_i) = \Phi(\Phi^{-1}(\underline{p_i}) - \frac{\left\|\delta\right\|_2}{\sigma}) \tag{46}$$

We obtain the first inequality from Lemma 6, and the second equality from Lemma 2. Similarly, for an arbitrary element $j \in \mathcal{V}_s$, we define the following region:

$$\mathcal{B}_j = \{\mathbf{w} : \delta^T(\mathbf{w} - \mathbf{x}) \geq \sigma \left\|\delta\right\|_2 \Phi^{-1}(1 - \overline{p}_j)\} \tag{47}$$

Then, based on Lemma 6 and Lemma 2, we have the following:

$$\Pr(j \in f_{k'}(\mathbf{Y})) \leq \Pr(\mathbf{Y} \in \mathcal{B}_j) = \Phi(\Phi^{-1}(\overline{p}_j) + \frac{\left\|\delta\right\|_2}{\sigma}) \tag{48}$$

Therefore, we have the following:

$$\max_{i \in \mathcal{U}_r} \Pr(i \in f_{k'}(\mathbf{Y})) \tag{49}$$

$$\geq \max_{i \in \mathcal{U}_r} \Phi(\Phi^{-1}(\underline{p_i}) - \frac{\left\|\delta\right\|_2}{\sigma}) = \max_{i \in \{w_1, w_2, \cdots, w_r\}} \Phi(\Phi^{-1}(\underline{p_i}) - \frac{\left\|\delta\right\|_2}{\sigma}) = \Phi(\Phi^{-1}(\underline{p_{w_1}}) - \frac{\left\|\delta\right\|_2}{\sigma}) \tag{50}$$

$$\min_{j \in \mathcal{V}_s} \Pr(j \in f_{k'}(\mathbf{Y})) \tag{51}$$

$$\leq \min_{j \in \mathcal{V}_s} \Phi(\Phi^{-1}(\overline{p}_j) + \frac{\left\|\delta\right\|_2}{\sigma}) = \min_{j \in \{z_1, z_2, \cdots, z_s\}} \Phi(\Phi^{-1}(\overline{p}_j) + \frac{\left\|\delta\right\|_2}{\sigma}) = \Phi(\Phi^{-1}(\overline{p}_{z_1}) + \frac{\left\|\delta\right\|_2}{\sigma}) \tag{52}$$

Next, we consider all possible subsets of $\mathcal{U}_r$ and $\mathcal{V}_s$. We denote $\Gamma_u \subseteq \mathcal{U}_r$, a subset of u elements in $\mathcal{U}_r$, and denote $\Lambda_v \subseteq \mathcal{V}_s$, a subset of v elements in $\mathcal{V}_s$. Then, we have the following:

$$\max_{i \in \mathcal{U}_r} \Pr(i \in f_{k'}(\mathbf{Y})) \geq \max_{\Gamma_u \subseteq \mathcal{U}_r} \max_{i \in \Gamma_u} \Pr(i \in f_{k'}(\mathbf{Y})) \tag{53}$$

$$\min_{j \in \mathcal{V}_s} \Pr(j \in f_{k'}(\mathbf{Y})) \leq \min_{\Lambda_v \subseteq \mathcal{V}_s} \min_{j \in \Lambda_v} \Pr(j \in f_{k'}(\mathbf{Y})) \tag{54}$$

We define the following quantities:

$$\underline{p_{\Gamma_u}} = \sum_{i \in \Gamma_u} \underline{p_i} \text{ and } \overline{p}_{\Lambda_v} = \sum_{j \in \Lambda_v} \overline{p}_j \tag{55}$$

Given these quantities, we define the following region based on Equation 30:

$$\mathcal{A}_{\Gamma_u} = \{\mathbf{w} : \delta^T(\mathbf{w} - \mathbf{x}) \leq \sigma \left\|\delta\right\|_2 \Phi^{-1}(\frac{\underline{p_{\Gamma_u}}}{k'})\} \tag{56}$$

$$\mathcal{B}_{\Lambda_v} = \{\mathbf{w} : \delta^T(\mathbf{w} - \mathbf{x}) \geq \sigma \left\|\delta\right\|_2 \Phi^{-1}(1 - \frac{\overline{p}_{\Lambda_v}}{k'})\} \tag{57}$$

Then, we have the following:

$$\frac{\sum_{i\in\Gamma_u}\Pr(i\in f_{k'}(\mathbf{Y}))}{k'} \tag{58}$$

$$\geq \Pr(\mathbf{Y}\in\mathcal{A}_{\Gamma_u}) \tag{59}$$

$$= \Phi(\Phi^{-1}(\frac{p_{\Gamma_u}}{k'}) - \frac{\|\delta\|_2}{\sigma}) \tag{60}$$

We have Equation 59 from 58 based on Lemma 5, and we have Equation 60 from 59 based on Lemma 2. Therefore, we have the following:

$$\max_{i\in\Gamma_u}\Pr(i\in f_{k'}(\mathbf{Y})) \tag{61}$$

$$\geq \frac{\sum_{i\in\Gamma_u}\Pr(i\in f_{k'}(\mathbf{Y}))}{u} \tag{62}$$

$$= \frac{k'}{u}\cdot\Phi(\Phi^{-1}(\frac{p_{\Gamma_u}}{k'}) - \frac{\|\delta\|_2}{\sigma}) \tag{63}$$

We have Equation 62 from 61 because the maximum value is no smaller than the average value. Similarly, we have the following:

$$\min_{j\in\Lambda_v}\Pr(j\in f_{k'}(\mathbf{Y})) \leq \frac{k'}{v}\cdot\Phi(\Phi^{-1}(\frac{\overline{p}_{\Lambda_v}}{k'}) + \frac{\|\delta\|_2}{\sigma}) \tag{64}$$

Recall that we have $\overline{p}_{w_1}\geq\overline{p}_{w_2}\geq\cdots\geq\overline{p}_{w_r}$ for $\mathcal{U}_r$. By taking all possible Γ_u with u elements into consideration, we have the following:

$$\max_{i\in\mathcal{U}_r}\Pr(i\in f_{k'}(\mathbf{Y})) \geq \max_{\Gamma_u\subseteq\mathcal{U}_r}\max_{i\in\Gamma_u}\Pr(i\in f_{k'}(\mathbf{Y})) \geq \max_{\Gamma_u=\{w_1,\cdots,w_u\}}\frac{k'}{u}\cdot\Phi(\Phi^{-1}(\frac{p_{\Gamma_u}}{k'}) - \frac{\|\delta\|_2}{\sigma}) \tag{65}$$

In other words, we only need to consider $\Gamma_u=\{w_1,\cdots,w_u\}$, i.e., a subset of u elements in $\mathcal{U}_r$ whose label probability upper bounds are the largest, where ties are broken uniformly at random. The reason is that $\Phi(\Phi^{-1}(\frac{p_{\Gamma_u}}{k'}) - \frac{\|\delta\|_2}{\sigma})$ increases as $\underline{p_{\Gamma_u}}$ increases. Combining with Equations 49, we have the following:

$$\max_{i\in\mathcal{U}_r}\Pr(i\in f_{k'}(\mathbf{Y})) \geq \max\{\Phi(\Phi^{-1}(\underline{p_{w_1}}) - \frac{\|\delta\|_2}{\sigma}), \max_{\Gamma_u=\{w_1,\cdots,w_u\}}\frac{k'}{u}\cdot\Phi(\Phi^{-1}(\frac{p_{\Gamma_u}}{k'}) - \frac{\|\delta\|_2}{\sigma})\} \tag{66}$$

Similarly, we have the following:

$$\min_{j\in\mathcal{V}_s}\Pr(j\in f_{k'}(\mathbf{Y})) \leq \min\{\Phi(\Phi^{-1}(\overline{p}_{z_1}) + \frac{\|\delta\|_2}{\sigma}), \min_{\Lambda_v=\{z_1,\cdots,z_v\}}\frac{k'}{v}\cdot\Phi(\Phi^{-1}(\frac{\overline{p}_{\Lambda_v}}{k'}) + \frac{\|\delta\|_2}{\sigma})\} \tag{67}$$

Bounding $\min_{\mathcal{U}_r}\max_{i\in\mathcal{U}_r}\Pr(i\in f_{k'}(\mathbf{Y}))$ and $\max_{\mathcal{V}_s}\min_{j\in\mathcal{V}_s}\Pr(j\in f_{k'}(\mathbf{Y}))$: We have the following:

$$\min_{\mathcal{U}_r}\max_{i\in\mathcal{U}_r}\Pr(i\in f_{k'}(\mathbf{Y})) \tag{68}$$

$$\geq \min_{\mathcal{U}_r}\max\{\max_{i\in\{w_1,w_2,\cdots,w_r\}}\Phi(\Phi^{-1}(\underline{p_i}) - \frac{\|\delta\|_2}{\sigma}), \max_{\Gamma_u=\{w_1,\cdots,w_u\}}\frac{k'}{u}\cdot\Phi(\Phi^{-1}(\frac{p_{\Gamma_u}}{k'}) - \frac{\|\delta\|_2}{\sigma})\} \tag{69}$$

$$\geq \max\{\max_{i\in\{a_e,a_{e+1},\cdots,a_k\}}\Phi(\Phi^{-1}(\underline{p_i}) - \frac{\|\delta\|_2}{\sigma}), \max_{\Gamma_u=\{a_e,\cdots,a_{e+u-1}\}}\frac{k'}{u}\cdot\Phi(\Phi^{-1}(\frac{p_{\Gamma_u}}{k'}) - \frac{\|\delta\|_2}{\sigma})\} \tag{70}$$

$$= \max\{\Phi(\Phi^{-1}(\underline{p_{a_e}}) - \frac{\|\delta\|_2}{\sigma}), \max_{\Gamma_u=\{a_e,\cdots,a_{e+u-1}\}}\frac{k'}{u}\cdot\Phi(\Phi^{-1}(\frac{p_{\Gamma_u}}{k'}) - \frac{\|\delta\|_2}{\sigma})\} \tag{71}$$

$$= \max\{\Phi(\Phi^{-1}(\underline{p_{a_e}}) - \frac{\|\delta\|_2}{\sigma}), \max_{u=1}^{d-e+1}\frac{k'}{u}\cdot\Phi(\Phi^{-1}(\frac{p_{\Gamma_u}}{k'}) - \frac{\|\delta\|_2}{\sigma})\}, \tag{72}$$

Algorithm 1: Computing the Certified Intersection Size

Input: f, $\mathbf{x}$, $L(\mathbf{x})$, R, k', k, n, σ, and α.
Output: Certified intersection size.
$\mathbf{x}^1, \mathbf{x}^2, \cdots, \mathbf{x}^n \leftarrow$ RANDOMSAMPLE$(\mathbf{x}, \sigma)$
counts$[i] \leftarrow \sum_{t=1}^{n} \mathbb{I}(i \in f(\mathbf{x}^t)), i = 1, 2, \cdots, c$.
$\underline{p_i}, \overline{p}_j \leftarrow$ PROBBOUNDESTIMATION(counts, α), $i \in L(\mathbf{x}), j \in \{1, 2, \cdots, c\} \setminus L(\mathbf{x})$
$e \leftarrow$ BINARYSEARCH$(\sigma, k', k, R, \{\underline{p_i} | i \in L(\mathbf{x})\}, \{\overline{p}_j | j \in \{1, 2, \cdots, c\} \setminus L(\mathbf{x})\})$
return e

where $\Gamma_u = \{a_e, \cdots, a_{e+u-1}\}$. We have Equation 70 from 69 because $\max\{\max_{i \in \{w_1, w_2, \cdots, w_r\}} \Phi(\Phi^{-1}(\underline{p_i}) - \frac{\|\delta\|_2}{\sigma}), \max_{\Gamma_u = \{w_1, \cdots, w_u\}} \frac{k'}{u} \cdot \Phi(\Phi^{-1}(\frac{p_{\Gamma_u}}{k'}) - \frac{\|\delta\|_2}{\sigma})\}$ reaches the minimal value when $\mathcal{U}_r$ contains r elements with smallest label probability lower bounds, i.e., $\mathcal{U}_r = \{a_e, a_{e+1}, \cdots, a_d\}$, where $r = d - e + 1$. Similarly, we have the following:

$$\max_{\mathcal{V}_s} \min_{j \in \mathcal{V}_s} \Pr(j \in f_{k'}(\mathbf{Y})) \leq \min\{\Phi(\Phi^{-1}(\underline{p_{b_s}}) + \frac{\|\delta\|_2}{\sigma}), \min_{v=1}^{s} \frac{k'}{v} \cdot \Phi(\Phi^{-1}(\frac{\overline{p}_{\Lambda_v}}{k'}) + \frac{\|\delta\|_2}{\sigma})\}, \quad (73)$$

where $\Lambda_v = \{b_{s-v+1}, \cdots, b_s\}$ and $s = k - e + 1$.

Applying the law of contraposition: Based on necessary condition in Equation 44, if we have $|T \cap g_k(\mathbf{x} + \delta)| < e$, then, we must have the following:

$$\max\{\Phi(\Phi^{-1}(\underline{p_{a_e}}) - \frac{\|\delta\|_2}{\sigma}), \max_{u=1}^{d-e+1} \frac{k'}{u} \cdot \Phi(\Phi^{-1}(\frac{p_{\Gamma_u}}{k'}) - \frac{\|\delta\|_2}{\sigma})\} \quad (74)$$

$$\leq \min_{\mathcal{U}_r} \max_{i \in \mathcal{U}_r} \Pr(i \in f_{k'}(\mathbf{Y})) \quad (75)$$

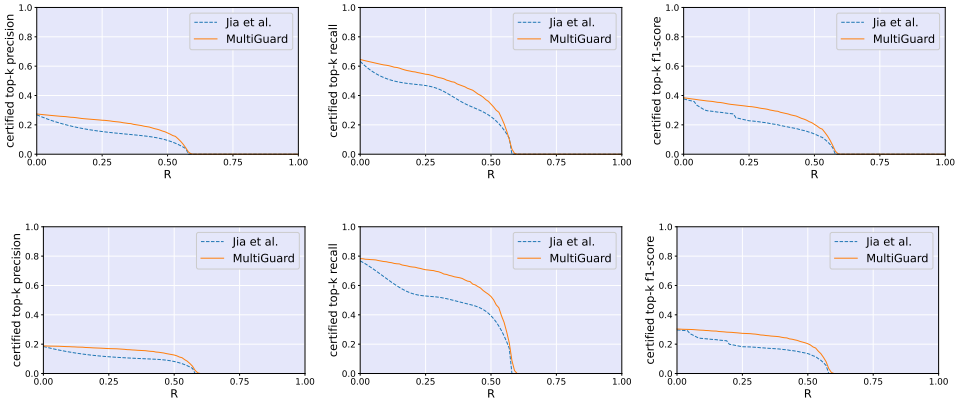
$$\leq \max_{\mathcal{V}_s} \min_{j \in \mathcal{V}_s} \Pr(j \in f_{k'}(\mathbf{Y})) \quad (76)$$

$$\leq \min\{\Phi(\Phi^{-1}(\overline{p}_{b_e}) + \frac{\|\delta\|_2}{\sigma}), \min_{v=1}^{k-e+1} \frac{k'}{v} \cdot \Phi(\Phi^{-1}(\frac{\overline{p}_{\Lambda_v}}{k'}) + \frac{\|\delta\|_2}{\sigma})\}, \quad (77)$$

We apply the law of contraposition and we obtain the statement: if we have the following:

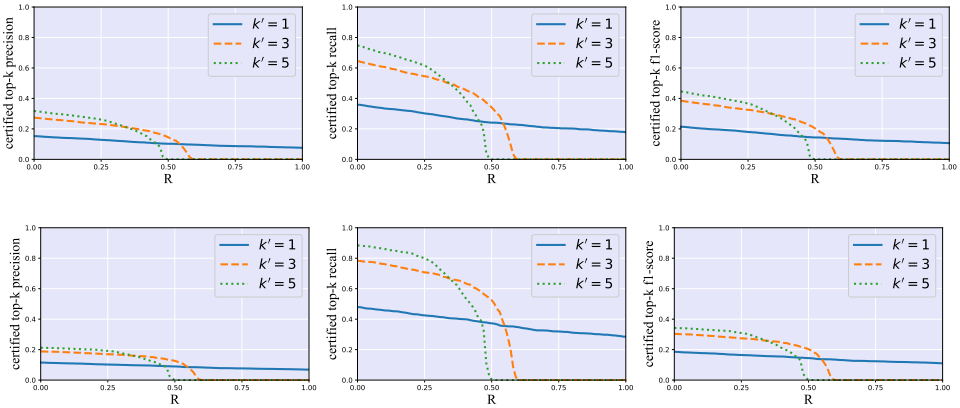
$$\begin{aligned} &\max\{\Phi(\Phi^{-1}(\underline{p_{a_e}}) - \frac{\|\delta\|_2}{\sigma}), \max_{u=1}^{d-e+1} \frac{k'}{u} \cdot \Phi(\Phi^{-1}(\frac{p_{\Gamma_u}}{k'}) - \frac{\|\delta\|_2}{\sigma})\} \\ &> \min\{\Phi(\Phi^{-1}(\overline{p}_{b_s}) + \frac{\|\delta\|_2}{\sigma}), \max_{v=1}^{k-e+1} \frac{k'}{v} \cdot \Phi(\Phi^{-1}(\frac{\overline{p}_{\Lambda_v}}{k'}) + \frac{\|\delta\|_2}{\sigma})\}, \end{aligned} \quad (78)$$

Then, we must have $|L(\mathbf{x}) \cap g_k(\mathbf{x} + \delta)| \geq e$. From Equation 8, we know that Equation 78 is satisfied for $\forall \|\delta\|_2 \leq R$. Therefore, we reach our conclusion. □



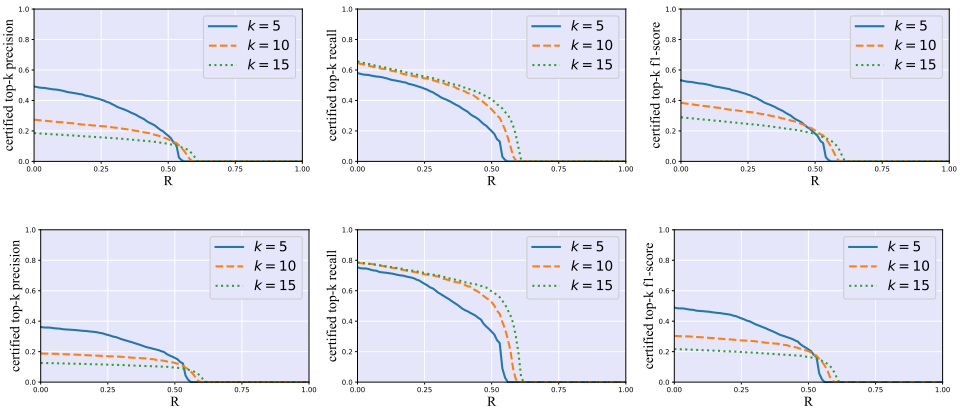
(a) Certified top-k precision@R (b) Certified top-k recall@R (c) Certified top-k f1-score@R

Figure 2: Comparing MultiGuard with with Jia et al. [22] on MS-COCO (first row) and NUS-WIDE (second row) dataset.



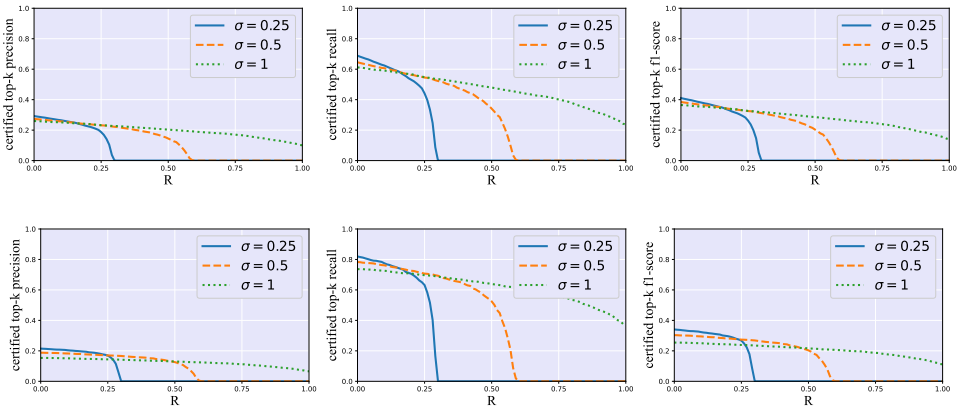
(a) Certified top-k precision@R (b) Certified top-k recall@R (c) Certified top-k f1-score@R

Figure 3: Impact of k' on the certified top-k precision@R, certified top-k recall@R, and certified top-k f1-score@R on MS-COCO (first row) and NUS-WIDE (second row) dataset.



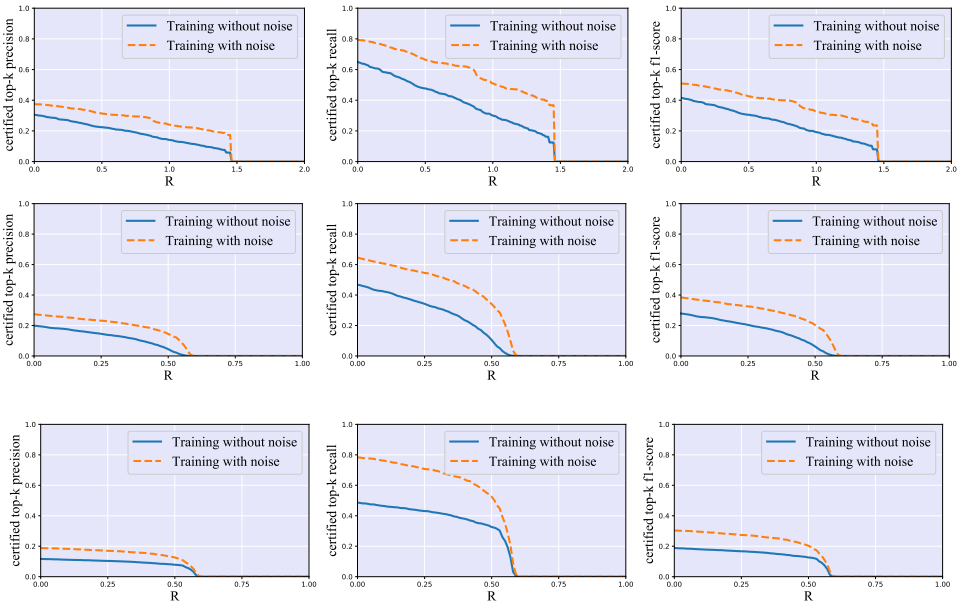
(a) Certified top-k precision@R (b) Certified top-k recall@R (c) Certified top-k f1-score@R

Figure 4: Impact of k on the certified top-k precision@R, certified top-k recall@R, and certified top-k f1-score@R on MS-COCO (first row) and NUS-WIDE (second row) dataset.



(a) Certified top-k precision@R (b) Certified top-k recall@R (c) Certified top-k f1-score@R

Figure 5: Impact of σ on the certified top-k precision@R, certified top-k recall@R, and certified top-k f1-score@R on MS-COCO (first row) and NUS-WIDE (second row) dataset.



(a) Certified top-k precision@R (b) Certified top-k recall@R (c) Certified top-k f1-score@R

Figure 6: Training the base multi-label classifier with vs. without noise on Pascal VOC (first row), MS-COCO (second row) and NUS-WIDE (third row) datasets.