Characterizing the Evasion Attackability of Multi-label Classifiers

Characterizing the Evasion Attackability of Multi-label Classifiers

Zhuo Yang[1], **Yufei Han**[2], **Xiangliang Zhang**[1]
[1]King Abdullah University of Science and Technology, Thuwal, Saudi Arabia
[2]Norton Research Group, Sophia Antipolis, France
{zhuo.yang, xiangliang.zhang }@kaust.edu.sa
yfhan.hust@gmail.com

Abstract

Evasion attack in multi-label learning systems is an interesting, widely witnessed, yet rarely explored research topic. Characterizing the crucial factors determining the attackability of the multi-label adversarial threat is the key to interpret the origin of the adversarial vulnerability and to understand how to mitigate it. Our study is inspired by the theory of adversarial risk bound. We associate the attackability of a targeted multi-label classifier with the regularity of the classifier and the training data distribution. Beyond the theoretical attackability analysis, we further propose an efficient empirical attackability estimator via greedy label space exploration. It provides provably computational efficiency and approximation accuracy. Substantial experimental results on real-world datasets validate the unveiled attackability factors and the effectiveness of the proposed empirical attackability indicator.

Introduction

Evasion attack has been witnessed widely in real-world practices of *multi-label learning* (Song et al. 2018). For example, a creepware/stalkware usually has multiple malicious labels as it sniffs the victim's privacy via different mobile services. To avoid being flagged, the entities authoring these malwares (Roundy et al. 2020; Freed et al. 2018) tend to hide their key malicious labels, such as rending remotely recording phone calls or accessing private files, by slightly reprogramming app binary structures. Meanwhile, they preserve less harmful labels like GPS tracking to pretend to be benign parental control. Another example can be found in image recommendation systems. An adversary tends to embed spam/toxic advertisements (Gupta et al. 2013) into a recommended image with other harmless contents. These malicious contents are so well tuned that the sanitary check system is deceived by the camouflaged image, while recognizing correctly other harmless scenarios.

Despite of the widely existence of multi-label adversarial threats, it has been a rarely investigated, yet important topic to evaluate the robustness of a multi-label classifier under evasion attack (a.k.a. **attackability**). Intuitively, assessing the attackability of a multi-label classifier h with an input instance is to explore the maximal perturbation on h's output that an input adversarial noise of bounded magnitudes can ever cause. The problem of attackability assessment in a general setting can be defined as: **given a magnitude bound of the adversarial perturbation and the distribution of legal input data instances, what is the worst-case miss-classification risk of h under the attack?** Classifier h is more attackable if it has a higher risk, while h is certified to be not attackable if its output cannot be changed by any adversarial noise within the magnitude bound. Via attackability assessment, we aim at answer the following questions:

- What are the factors determining the attackability level of a multi-label classifier?
- Can we derive an empirically computable attackability measurement for a multi-label classifier?

Echoing the questions raises two challenges: First, analyzing the worst-case classification risk on adversarial instances with PAC-learning framework requires a fixed distribution of adversarial instances. However, it is a well received fact that such a distribution depends radically on the targeted classifier's property, thus it is not fixed and closely associated with the classifier's architecture. The celebrated works (Yin, Ramchandran, and Bartlett 2019; Khim and Loh 2018; Tu, Zhang, and Tao 2019) proposed to mitigate the gap via the lens of Rademacher complexity. Nevertheless, they all focused on single-label scenarios, thus can't be applied to answer the questions above. Second, evaluating the worst-case risk for an input instance needs to explore the maximal set of jointly attackable labels. Since labels are not mutually exclusive in multi- label tasks, the label exploration process is in nature an NP-hard mixed-integer programming problem. The adversarial noise generation method in (Song et al. 2018) applies only in the targeted attack scenario, where the attacked labels are given. Few effort has been dedicated to study the feasibility of label space exploration.

To address the challenges above, we conduct both theoretical and empirical attackability analysis of a multi-label classifier. Our contributions can be summarized as follows:

- We measure the attackability of a multi-label classifier by evaluating the expected worst-case miss-classification loss over the distribution of adversarial examples. We instantiate the study to linear and deep neural networks, which are popularly deployed in multi-label applications. Our analysis unveils that the attack strength, the regularity of the targeted classifier and the empirical loss on un-

perturbed data are the external and intrinsic driving force jointly determining the attackability level. We further reveal the theoretical rationality of the low-rank regularization and adversarial training in hardening the classifier.

- We cast the problem of evaluating the empirical attackability level as *a label space exploration process* for each of the legal input instances. We further demonstrate the triviality of the label space exploration problem by formulating it as a submodular set function optimization problem. Thus it can be solved via greedy search with certified approximation accuracy.
- We propose a *Greedy Attack Space Expansion* (GASE) algorithm to address the computational bottleneck of the primitive greedy search for empirical attackability measurement. The proposed method provides a computationally economic estimator of the marginal gain obtained by adding a candidate label into the set of attacked labels. It selects the labels with the largest marginal gain to deliver an efficient exploration of attack targets.

Related works

Robustness against adversaries. The emergence of evasion attack raises a severe challenge to practitioners' trust on machine learning in performance-critic applications (Battista and Fabio 2018; Biggio et al. 2013; Carlini and Wagner 2017). Considerable efforts have been dedicated to detect adversarial samples, improve model designs and propose robust training methods (Cullina et al. 2019; Goodfellow, Shlens, and Szegedy 2015; Athalye, Carlini, and Wagner 2018; Fawzi, Moosavi-Dezfooli, and Frossard 2016; Szegedy et al. 2013; Xu, Evans, and Qi 2018; Madry et al. 2018; Ross and Doshi-Velez 2018; Jakubovitz and Giryes 2018; Hein and Andriushchenko 2017; Zugner and Gunnemann 2019; Bojchevski and Günnemann 2019; A. Bojchevski and Günnemann 2019; Raghunathan, Steinhardt, and Liang 2018; Florian et al. 2018; Papernot et al. 2016; Zugner and Günnemann 2020; Cohen, Rosenfeld, and Kolter 2019; Lee et al. 2019). Especially, (Wang et al. 2019; Shafahi et al. 2019; Gao et al. 2019) discussed convergence guarantee and high sample complexity of adversarial training. In contrast, few literature focuses on the essential problem of *evaluating the vulnerability of a classifier under a given evasion attack setting and identifying key factors determining the feasibility of evasion attack against the targeted classifier*. Pioneering works of this topic (Hein and Andriushchenko 2017; Wang, Jha, and Chaudhuri 2018; Fawzi, Moosavi-Dezfooli, and Frossard 2016; Gilmer et al. 2018) focused on identifying the upper bound of adversarial noise, which guarantees the stability of the targeted classifier's output, a.k.a adversarial sphere. Notably, (Fawzi, Moosavi-Dezfooli, and Frossard 2016) pointed out the association between adversarial robustness and the curvature of the classification boundary. Strengthened further by (Yin, Ramchandran, and Bartlett 2019; Khim and Loh 2018; Tu, Zhang, and Tao 2019), the expected classification risk under adversarial perturbation can be bounded by the Rademacher complexity of the targeted classifier. Moreover in (Qi et al. 2019; Wang et al. 2020), attackability of a recurrent neural net based classifier on discrete inputs was measured by checking the regularity of the targeted classifier. Apparently, the regularity of the targeted classifier play an equally important role as the attack strength in causing adversarial vulnerability. Inspiring as they are, these works focus on single-label learning tasks. Due to the label co-occurrence in multi-label learning, searching for the combinations of attacked labels causing the worst-case loss is NP-hard. It is thus an open issue to evaluate the adversarial risk of multi-label learners.

Noise-tolerant multi-label learning. A relevant topic is to learn multi-label classifier with imperfect training instances. Miss-observations and noise corruptions of features and labels of training instances can introduce severe bias into the derived classifier. Most research efforts in this domain recognised that the key to success is to encode label correlation and the predicative relation between features and labels (Sun, Zhang, and Zhou 2010; Zhu, Yan, and Ma 2010; Liu et al. 2010; Lin et al. 2013; Wu, Jin, and Jain 2013; Zhao and Guo 2015; Yu et al. 2014a; Bi and Kwok 2014; Goldberg et al. 2010; Yu et al. 2014b; Cabral et al. 2015; Xu, Jin, and Zhou 2013; Chiang, Hsieh, and Dhillon 2015; Guo 2017; Zhu, Kwok, and Zhou 2018; Hsieh, Natarajan, and Dhillon 2015). They exploited not only low-rank structures of feature/label matrices for missing data imputation, but also gained stable performances by enforcing the low-rank regularization on the predictive model capturing the feature-label correlation. Especially (Xu et al. 2016) proposed to regularize the local Rademacher complexity of a linear multi-label classifier in the training process. The study indicated the link between the Rademacher complexity and the low-rank structure of the classifier's coefficients. The reported results showed that a lower-rank structured linear classifier can better recover missing labels. Nevertheless, all the previous works focus on adversary-free scenarios. Furthermore, the analysis over the role of low-rank structures was limited to linear multi-label classifiers. It is thus interesting to study whether the low rank driven regularization can help to mitigate the adversarial threat against both linear and DNN based multi-label classifiers.

Attackability of Multi-label Classifiers

Notations and Problem Definition. We assume $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ as a measurable multi-label instance space, with $\mathcal{X} \in \mathbb{R}^d$ and $\mathcal{Y} = \{-1, 1\}^m$, where d is the feature dimension and m is the number of labels. Given n i.i.d. training examples $\{(\mathbf{x}_i, \mathbf{y}_i)\}$ drawn from $\mathcal{P}(\mathcal{Z})$, the classifier $h \in \mathcal{H} : \mathcal{X} \to \mathcal{Y}$ is learnt by minimizing the empirical loss $\sum_{i=1}^{n} \ell(\mathbf{x}_i, \mathbf{y}_i)$, with the loss function $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.

Eq.(1) defines a typical scenario of empirical attackability evaluation for a multi-label classifier h given an input $\mathbf{x_i}$, perturbed by $\mathbf{r}$. The classification output $sgn(h(\mathbf{x}_i + \mathbf{r}^*))$ has been flipped on as many as possible labels.

$$
\begin{aligned}
&C^*(\mathbf{x}_i) = \max_{T, \|\mathbf{r}^*\| \le \mu_r} \sum_{j=1}^{m} I(y_{ij} \neq sgn(h_j(\mathbf{x}_i + \mathbf{r}^*))) \\
&\text{where } \mathbf{r}^* = \arg\min_{\mathbf{r}} \|\mathbf{r}\| \\
&s.t.\ y_{ij} h_j(\mathbf{x}_i + \mathbf{r}^*) \le 0 \ (j \in T),\ \ y_{ij} h_j(\mathbf{x}_i + \mathbf{r}^*) > 0 \ (j \notin T)
\end{aligned}
\tag{1}
$$

where $h_j(\mathbf{x}_i + \mathbf{r})$ denotes the classification score for the label j of the adversarial example, and sgn is the sign function outputting ± 1 based on the sign of $h_j(\mathbf{x}_i + \mathbf{r})$. The indicator function $I(\cdot)$ outputs 1 if the attack flips a label and 0 otherwise. T denotes the set of flipped labels. The magnitude of C^* indicates the attackability of the classifier given the attack strength limit μ_r and the input $\mathbf{x}_i$. Given the same input $\mathbf{x}$ and the bound of perturbation μ_r, one multi-label classifier h is more attackable than the other h', if $C^*_h > C^*_{h'}$.

Bound of Adversarial Attackability

Beyond the attackability measurement given a local fixed input instance, we pursue a theoretical and empirical insight into the attackability of h in the space of adversarial samples, which are sampled from a new distribution $\mathcal{P}'$ translated from $\mathcal{P}$ after injecting the adversarial perturbation. The distribution shift from $\mathcal{P}$ to $\mathcal{P}'$ is the origin of adversarial threat, as it violates the i.i.d. assumption of the learning process. By assuming that $\mathcal{P}'$ lies within a Wassernstein ball centered at $\mathcal{P}$ with a radius of ϵ, we have the following definition about classification risk under evasion attack.

Definition 1. *For a multi-label classifier h and legal input samples $\{\mathbf{x}_i, \mathbf{y}_i\} \sim \mathcal{P}$ and its corresponding adversarial samples $\{\mathbf{x}'_i, \mathbf{y}_i\} \sim \mathcal{P}'$, the worst case expected and empirical risk under the evasion attack are,*

$$
\begin{aligned}
R_{\mathcal{P}'}(h) &= E_{(\mathbf{x},\mathbf{y})\sim\mathcal{P}}[\max_{(\mathbf{x}',\mathbf{y})\sim\mathcal{P}',\mathcal{W}(\mathcal{P},\mathcal{P}')\leq\epsilon} l(h(\mathbf{x}'),\mathbf{y})] \\
R^{emp}_{\mathcal{P}'}(h) &= \frac{1}{n}\sum_{i=1}^{n}[\max_{(\mathbf{x'}_i,\mathbf{y}_i)\sim\mathcal{P}',\mathcal{W}(\mathcal{P},\mathcal{P}')\leq\epsilon} l(h(\mathbf{x'}_i),\mathbf{y}_i)]
\end{aligned}
\tag{2}
$$

where $\mathcal{W}(\mathcal{P}', \mathcal{P})$ denotes the Wassernstein distance between $\mathcal{P}'$ and $\mathcal{P}$, and ϵ is the radius of the adversarial space.

The $\mathcal{W}(\mathcal{P}', \mathcal{P})$ can be bounded with the magnitude of the adversarial perturbation after (Tu, Zhang, and Tao 2019), which gives $\mathcal{W}(\mathcal{P}', \mathcal{P}) \leq \sup_{\mathbf{x}',\mathbf{x}} \|\mathbf{x}'_i - \mathbf{x}_i\|_2 \leq \mu_r$. Without loss of generality, we use the Euclidean distance $\|\mathbf{x}'_i - \mathbf{x}_i\|_2$ to constrain the attack budget hereafter. Consistent with the defined attackability evaluation scenario in Eq.1, $R_{\mathcal{P}'}(h)$ measures the attackability of h. A higher $R_{\mathcal{P}'}(h)$ indicates a more attackable h. And $R^{emp}_{\mathcal{P}'}(h)$ is the empirical estimator of the attackability level. By definition, if we derive C^* by solving Eq.(1) for each instance $(\mathbf{x}_i, \mathbf{y}_i)$ and adopt the binary 0-1 loss, an aggregation of the local worst-case loss $\sum_{i=1}^{n} C^*(\mathbf{x}_i)$ gives $R^{emp}_{\mathcal{P}'}(h)$.

In the followings, we establish the upper bound of the attackability measurement with respect to *linear and feedforwad neural network multi-label classifiers*. It reveals the key factors determining the attackability level of a classifier. Given $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{y} \in \{-1, 1\}^m$ as the feature and label vector of a data instance, a linear multi-label classifier is $h(\mathbf{x}) = \mathbf{x}\mathbf{w}$. The linear coefficient matrix $\mathbf{w} \in \mathbb{R}^{d\times m}$ is defined with the spectral norm $\|\mathbf{w}\|_\delta \leq \Lambda$. Furthermore, we constrain the range of legal inputs and the adversary's strength as $\|\mathbf{x}\|_2 \leq \mu_x$ and $\|\mathbf{r}\|_2 \leq \mu_r$ respectively. Without loss of generality, a Least-Squared Error (LSE) loss is adopted to compute the classification risk, such as $\ell(\mathbf{x}, \mathbf{y}) = \|\mathbf{y} - \mathbf{x}\mathbf{w}\|_2$. A distance metric for $z = \{\mathbf{x}, \mathbf{y}\}$ is defined as $d(z_i, z_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_2 + \|\mathbf{y}_i - \mathbf{y}_j\|_2$.

Theorem 1. ***[Upper-bound of attackability for a linear multi-label classifier]*** *The upper bound of $R_{\mathcal{P}'}(h)$ holds with at least probability of 1-σ:*

$$
\begin{aligned}
R_{\mathcal{P}'}(h) \leq R^{emp}_{\mathcal{P}'}(h) + 96\sqrt{\frac{\mu_x\Lambda R(1+\mu_x\Lambda)}{n}} \\
+ \frac{12C_h\sqrt{\pi}(m+2\mu_x)}{\sqrt{n}} + (m+\Lambda\mu_x)\sqrt{\frac{\log(1/\sigma)}{2n}}
\end{aligned}
\tag{3}
$$

and the worst case empirical loss has the upper bound:

$$
R^{emp}_{\mathcal{P}'}(h) \leq \frac{1}{n}\sum_{i=1}^{n} \ell(h(\mathbf{x}_i), \mathbf{y}_i) + C_h\mu_r \tag{4}
$$

where R denotes the rank of the coefficient matrix $\mathbf{w}$ and $C_h = max\{\|\mathbf{w}\|_\delta, 1\}$.

The proof is presented in supplementary document.

Remark 1. *We have three observations from the derived analysis in Eq.(3-4).*
1) The worst-case empirical loss $R^{emp}_{\mathcal{P}'}$ can be used as a sensitive indicator of the worst-case expected adversarial risk $R_{\mathcal{P}'}$. Thus it can be used as an empirical measure of attackability of the classifier over the adversarial data space. A lower $R^{emp}_{\mathcal{P}'}$ implies a lower expected miss-classification risk in the adversarial space.
2) The spectrum of linear coefficient matrix $\mathbf{w}$ plays an important role in deciding the attackability level of h. Especially, h with lower rank $\mathbf{w}$ has lower expected miss-classification risk. This is consistent with what was unveiled in previous research of multi-label classification: enforcing low-rank constraints over the linear classifier usually brings robustness improvement against noise corruption.
3) The empirical risk over unperturbed data*,* the magnitude of the adversarial perturbation *and* the spectrum of the classifier's coefficient matrix *are the three main factors jointly determining the attackability of the multi-label classifier. On one hand, the risk upper bound depends on the external driving force of the adversarial threat, which is the magnitude of the adversarial perturbation $\|\mu_r\|$. On the other hand, the internal factors on the riks upper bound are the regularity of the classifier (low-rank structure) and the profile of the training data distribution. Moreover, by dropping the terms with μ_r in Eq.(3), we can find that an adversary-free generalization bound of the linear multi-label classifier heavily depends on the low rank structure of the classifier. It is consistent with the results unveiled by previous works (Xu, Jin, and Zhou 2013; Yu et al. 2014a; Zhu, Kwok, and Zhou 2018): low-rank structured classifiers are favorable in multi-label classification. Due to the page limit, we leave this discussion in the supplementary document.*

Inherited the setting of the attack scenario from Theorem.1, we consider a neural network based multi-label classifier h_{nn} with L layers, where:

- The dimension of each layer is $d_1, d_2, ..., d_L$, and $d_0 = d$ for taking input $\mathbf{x}$ and $d_L = m$ for outputting labels $\mathbf{y}$.
- At each layer i, $A_i \in R^{d_{i-1}\times d_i}$ denotes the linear coefficient matrix (connecting weights). The spectral norm of A_i is bounded as $\|A_i\|_\delta \leq \Lambda_i$. R_i denotes the rank of A_i.

- The activation functions used in the same layer are Lipschitz continuous and share the same Lipschitz constant ρ_i. We use g_i to denote the activation functions used at the layer i. The output of each layer i can be defined recursively as $\mathcal{H}_i = g_i(\mathcal{H}_{i-1}A_i)$.

Theorem 2. ***[Upper-bound of attackability for neural nets based multi-label classifier]*** *The upper bound of $R_{\mathcal{P}'}(h_{nn})$ holds with at least probability of $1-\sigma$:*

$$R_{\mathcal{P}'}(h_{nn}) \leq R^{emp}_{\mathcal{P}'}(h_{nn}) + 2m\sqrt{\frac{\log(1/\sigma)}{2n}} + \frac{96\sqrt{dm\Lambda_d}\sum_{i=1}^{L} R_i\sqrt{d_i\Lambda_i C_i}}{\sqrt{n}} + \frac{12C_{nn}(2\mu_x+m)\sqrt{\pi}}{\sqrt{n}} \tag{5}$$

and the empirical loss $R^{emp}_{\mathcal{P}'}$ has the upper bound:

$$R^{emp}_{\mathcal{P}'}(h_{nn}) \leq \frac{1}{n}\sum_{i=1}^{n}\ell(h_{nn}(\mathbf{x}_i),\mathbf{y}_i) + C_{nn}\mu_r \tag{6}$$

where $C_1 = \rho_d$ and $C_i = \prod_{j=d-i+1}^{d}\rho_j\prod_{j=d+2-i}^{d}\|A_j\|_\delta$ with $i \geq 2$, and $C_{nn} = max\{1, \prod_{i=1}^{L}\rho_i\|A_i\|_\delta\}$.

The proof is presented in supplementary document.

Remark 2. *Similar to the observations in Remark 1, the attackability of h_{nn} depends heavily on the spectrum of linear coefficients at each layer of the neural nets, the empirical loss of h_{nn} on legal input samples and the attack strength $\|\mu_r\|$. More specifically, the linear coefficient matrices $\{A_i\}$ with lower ranks and lower spectral norm can make h_{nn} more robust. Indeed, enforcing regularization on the spectral norm of the linear coefficients can improve the generalization capability of DNN (Yoshida and Miyato 2017; Miyato et al. 2018). Our analysis not only provides the theoretical rationality behind the reported empirical observations, but also, unveils the impact of the low-rank constraint on $\{A_i\}$ in controlling h_{nn}'s attackability.*

From Remark 1 and 2, we find that both reducing the worst-case empirical risk of the targeted classifier and enforcing low-rank constraints on its coefficients can help to reduce the attackability and mitigate the risk on evasion attack. The former can be achieved by conducting adversarial training with the crafted worst-case multi-label adversarial samples. The latter aims at controlling the Rademacher complexity of the targeted classifier, which improves the generalization capability of the targeted classifier. In (Xu and Mannor 2010), the close association between generalization and robustness was explored. Better generalization capability indicates more robustness against noise corruption.

Empirical Attackability Evaluation by Greedy Exploration

Problem Reformulation

Solving Eq.(1) to compute the worst-case loss $R^{emp}_{\mathcal{P}'}(h)$ over legal input instances is an NP-hard mixed-integer non-linear constraint problem (MINLP). Traditional solutions to this problem, such as Branch-and-Bound, has an exponential complexity in the worst case. To achieve an efficient evaluation, we propose to empirically approximate $R^{emp}_{\mathcal{P}'}$ via greedy forward expansion of the set of the attacked labels. We re-formulate the label exploration problem in Eq.(1) as a bi-level set function optimization problem:

$$\begin{aligned} &S^* = \arg\max_S \psi(S) \\ &\text{where } \psi(S) = \max_{S,|S|\geq k}\{|S| - g(S)\} \\ &g(S) = \min_{T\subseteq S, \|\mathbf{r}\|\leq\mu_r}\|\mathbf{r}\|_2^2 \\ &s.t.\ (1-2b_j)y_j h_j(\mathbf{x}+\mathbf{r}) \geq t_j,\ \ j = 1,2,...,m \\ &b_j = 1\ \text{(for } j\in T),\ \ b_j = 0\ \text{(for } j\notin T) \end{aligned} \tag{7}$$

where label $y_j = \{+1,-1\}$, and t_j is the minimum classification margin value enforced on label j. The core components of the constraints are the binary indicators $\{b_j\}$. With $b_j = 1$, label y_j is flipped, while with $b_j = 0$, the label remains unchanged. The set function $g(S)$ returns the minimal magnitude of the perturbation $\mathbf{r}$ ever achieved via attacking the labels indicated by subsets of S. In this sense, the inner layer of Eq.(7) defines an evasion attack against the multi-label classifier targeting at the labels indicated by S. The optimization objective of the outer layer aims at expanding the set S as much as possible while minimizing as much as possible the required attack cost $\|\mathbf{r}\|_2$. Notably, we set a lower bound k for $|S|$ in Eq.(7) for the convenience of presentation. In a naive way, we can gradually increase the lower bound k until the attack cost valued by $\psi(S)$ surpasses the budget limit. The volume $|S|$ of the derived set gives an estimate of $R^{emp}_{\mathcal{P}'}$ with the binary loss.

Lemma 1. *The outer layer of Eq.(7) defines a problem of non-monotone submodular function maximization. Let $\psi(\hat{S})$ and $\psi(S^*)$ denote respectively the objective function value obtained by randomized greedy forward search proposed in (Buchbinder et al. 2014) and the underlying global optimum following the cardinality lower bound constraint. The greedy search based solution has the following certified approximation accuracy:*

$$\psi(\hat{S}) \geq \frac{1}{4}\psi(S^*) \tag{8}$$

Fast Greedy Attack Space Exploration

According to Lemma.1, the set S derived from the random greedy search produces an attack cost $\|\mathbf{r}\|_2$ that is close to the one achieved by the global optimum solution. It guarantees the quality of the greedy search based solution. The *primitive greedy forward expansion* is thus designed as follows:

- We initialize an empty S (flipped labels), which assumes no labels are attacked at the beginning.
- In each round of the greedy expansion, for the current set S and current adversarial noise $\mathbf{r}(S)$, we choose each of the candidate labels j out of S and compute the marginal gain $\|\mathbf{r}(S\cup j)\|_2 - \|\mathbf{r}(S)\|_2$ by conducting targeted multi-label evasion attack. $\|\mathbf{r}(S\cup j)\|_2$ is the magnitude of the adversarial noise to flip all the labels in $S\cup j$. We then select randomly one of the candidate labels j with the least marginal gains to update $S = S\cup j$.

- We update $\|\mathbf{r}\|_2$ by conducting an evasion attack targeting at the labels indicated by S. The expansion stops when $\|\mathbf{r}\|_2 \geq \mu_r$.

In each iteration, the *primitive greedy forward expansion* needs to perform evasion attack for each candidate label. It requires $(m+1)k - k(k-1)/2 - 1$ evasion attacks before including k labels in S. It is costly when the label dimension is high. To break the bottleneck, we propose a computationally economic estimator to the magnitude of the marginal gain $\Delta = \|r(S \cup j)\|_2 - \|r(S)\|_2$.

Lemma 2. *In each iteration of the greedy forward expansion, the magnitude of the marginal gain Δ is proportional to $\frac{|y_j h_j(\mathbf{x}+\mathbf{r})|}{\|\nabla_j(\mathbf{x}+\mathbf{r})\|}$, where $\mathbf{r}$ is the current feasible adversarial perturbation. $\|\nabla_j(\mathbf{x}+\mathbf{r})\|$ denotes the L2 norm of the gradient vector $\frac{\partial h_j(\hat{\mathbf{x}})}{\partial \hat{\mathbf{x}}}$ at the point $\hat{\mathbf{x}} = \mathbf{x} + \mathbf{r}$.*

Therefore, instead of running evasion attack for each candidate label, we can simply choose the one with the smallest ratio $\frac{|y_j h_j(\mathbf{x}+\mathbf{r})|}{\|\nabla_j(\mathbf{x}+\mathbf{r})\|}$. Algorithm 1 presents the proposed *Greedy Attack Space Expansion* (GASE) algorithm. It only runs in total $k-1$ evasion attacks to reach $|S| = k$.

In the proposed GASE algorithm, the step of *greedy label expansion* is equivalent to conducting the orthogonal matching pursuit guided greedy search (Elenberg et al. 2016). It enjoys fast computation, the optimal value of the objective function in Eq.(7) achieved by GASE has a guaranteed approximation accuracy to the underlying global optimum according to Theorem 1.3 in (Buchbinder et al. 2014).

The step of *greedy label expansion* in Algorithm.1 benefits from label correlation in multi-label instances. A successful attack targeted at one label tends to bias the classification output of another highly correlated label simultaneously. The candidate label with the weakest classification margin while a large $\|\nabla_j(\mathbf{x}+\mathbf{r})\|$ is thus likely to be flipped with minor update on the adversarial perturbation. Notably, the proposed GASE algorithm is independent of *the choice of evasion attack methods* in the step *targeted evasion attack*. Once the greedy search for each input instance $\mathbf{x}$ finishes, we use the average $|S|$ computed over all $\mathbf{x}$ as *the empirical attackability indicator.* A larger average $|S|$ indicates a higher attackability of the targeted multi-label classifier.

Algorithm 1: Greedy Attack Space Expansion

Input: Instance example $\mathbf{x}$, a trained multi-label classifier h, perturbation norm budget μ_r.

Output: The set of attacked labels S.

Initialize S as an empty set and $\mathbf{r} = 0$.

while $|S| < m$ *and* $\|\mathbf{r}\|_2 < \mu_r$ **do**

  Greedy label expansion: Calculate d_j in Eq.(9) for each label j outside S, where $h_j(\mathbf{x}+\mathbf{r})$ is the probabilistic classification output of label j, and t_j is the threshold of label decision;

$$d_j = \frac{|y_j h_j(\mathbf{x}+\mathbf{r})|}{\|\nabla_j(\mathbf{x}+\mathbf{r})\|} \tag{9}$$

  Update $S = S \cup j$, where label $j (j \notin S)$ is selected randomly from the labels with the least values of Eq.(9)

  Targeted evasion attack: Solve the targeted evasion attack problem with updated S and get the optimized perturbation $\mathbf{r}^*$; Update $\mathbf{r} = \mathbf{r}^*$

end

Experiments

In the experimental study, we aim at 1) validating the theoretical attackability analysis in Theorem 1 and 2; and 2) evaluating the empirical attackability indictor estimated by GASE for targeted classifiers.

Datasets. We include 4 datasets collected from various real-world multi-label applications, cyber security practices (*Creepware*), biology research (*Genbase*) (Tsoumakas, Katakis, and Vlahavas 2010), object recognition (*VOC2012*) (Everingham et al. 2012) and environment research (*Planet*) (Kaggle 2017). The 4 datasets are summarized in Table.1.

Targeted Classifiers. We instantiate the study empirically with linear Support Vector Machine (SVM) and Deep Neural Nets (DNN) based multi-label classifiers. Linear SVM is applied on *Creepware* and *Genbase*. DNN model Inception-V3 is used on *VOC2012* and *Planet*. On each data set, we randomly choose 50%, 30% and 20% data instances for training, validation and testing to build the targeted multi-label classifier. In Table.1, we show *Micro-F1* and *Macro-F1* scores derived on the unperturbed testing data. Note that feature engineering and model design of the classifiers for better classification is beyond the scope of this study. These classifiers are trained to achieve comparable classification accuracy w.r.t. the reported state-of-the-art methods on their corresponding datasets, so as to set up the test bed for the attackability analysis. Due to space limit, more experimental setting and results are provided in the supplementary file.

Table 1: Summary of the used real-world datasets. N is the number of instances. m is the total number of labels. l_{avg} is the average number of labels per instance. The F1-scores of the targeted classifiers on different datasets are also reported.

Dataset	N	m	l_{avg}	Micro F1	Macro F1	Classifier$_{target}$
Creepware	966	16	2.07	0.76	0.66	SVM
Genbase	662	27	1.25	0.99	0.73	SVM
VOC2012	17,125	20	1.39	0.83	0.74	Inception-V3
Planet	40,479	17	2.87	0.82	0.36	Inception-V3

Attack and Adversarial Training. We use adversarial-robustness-toolbox (Nicolae et al. 2018) to implement the step of targeted adversarial attack in Algorithm.1 and adversarial training. Specifically, *projected gradient decent* (PGD) (Madry et al. 2018) is employed to conduct the targeted attack in Algorithm.1. The decision threshold t_i in Algorithm.1 is set to 0 without loss of generality.

Performance Benchmark. We gradually increase the attack strength by varying the attack budget μ_r. Given a fixed value of μ_r, we calculate *the average number of flipped labels on test data* as an estimation of the empirical classifica-

tion risk $R^{emp}_{\mathcal{P}'}$ induced by the attack. This is the empirical attackability indicator, as defined in the end of the section of fast greedy attack space exploration.

Validation of Empirical Attackability Indicator

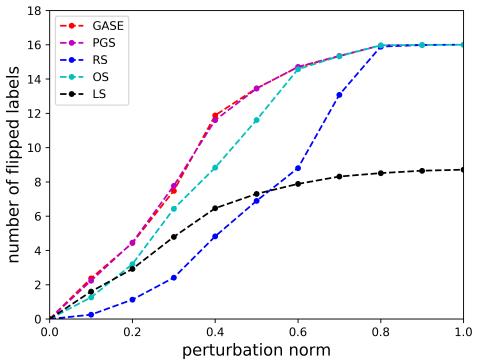
We assess here the empirical attackability indicator estimated by the proposed GASE algorithm, by comparing it with four baselines of label exploration strategies.

- **PGS** (Primitive Greedy Search): This is the costly primitive greedy search that requires $(m+1)k-k(k-1)/2-1$ evasion attacks before including k labels in S.
- **RS** (Random Search): In each round of RS, one label is selected purely by random from the candidate set without evaluating the marginal gain and added to the current set S.
- **OS** (Oblivious Search): This method first computes the norm of the adversarial perturbation induced by *flipping each candidate label while keeping the other labels unchanged*. The labels causing the least perturbation magnitudes are selected to form the set S.
- **LS** (Loss-guided Search): In each iteration, *LS* updates the adversarial perturbation $\mathbf{r}$ along the direction where the multi-label classification loss increased the most. The set of the attacked labels are reported when $\|\mathbf{r}\|_2$ surpasses the cost limit. This strategy aims at pushing the originally miss-classified instances even further from the decision plane, instead of flipping the labels of the originally correctly predicted instances. It misleads the search of the attackable labels by just maximizing the loss, and thus has bad performance as shown in Fig. 1.

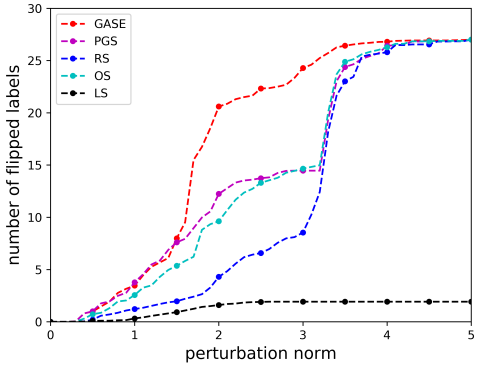
Fig. 1 shows the number of flipped labels obtained by the proposed *GASE* algorithm and the baselines on linear and DNN based multi-label classifiers. Since we limit the maximum iterations and perturbation norm bounds of attacks in our experiments, few cases of the involved label exploration methods can flip all of the labels in each dataset. Not surprisingly, the proposed *GASE* and *PGS* method achieve significantly more flipped labels than *RS*, *OS* and *LS* methods, especially when the constraint of attack budget is strict (with small perturbation norms). It confirms the reasonableness of greedy search stated in Lemma.1. Over all the datasets, *GASE* performs similarly or even better compared to *PGS*. It empirically demonstrates the merits of *GASE*: it is much less costly than *PGS*, while obtains attackability indicators with certified quality.

Attackability Evaluation with Countermeasures for Evasion Attack

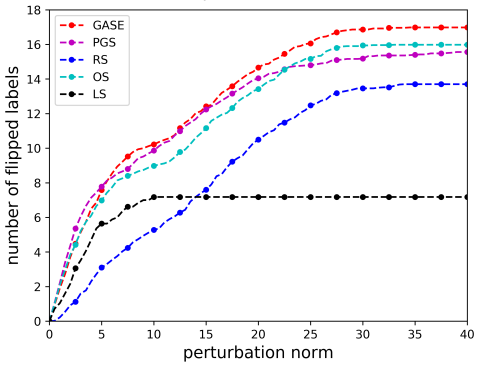
Following in our Theorem 1 and 2, we study the impact of the countermeasures on multi-label classifiers' attackability: controlling the model complexity by enforcing the low-rank nuclear norm constraint and conducting adversarial training. For the DNN based classifier, we enforce the nuclear norm constraint only on the linear coefficients of the final layer. We include 4 different settings on controlling the model complexity when training classifiers:



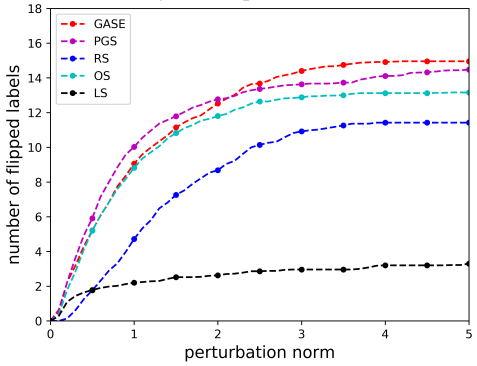
(a) attackability of SVM on *Creepware*



(b) attackability of SVM on *Genbase*



(c) attackability of Inception-V3 on *VOC2012*



(d) attackability of Inception-V3 on *Planet*

Figure 1: The empirical attackability indicator estimated by different label exploration strategies.

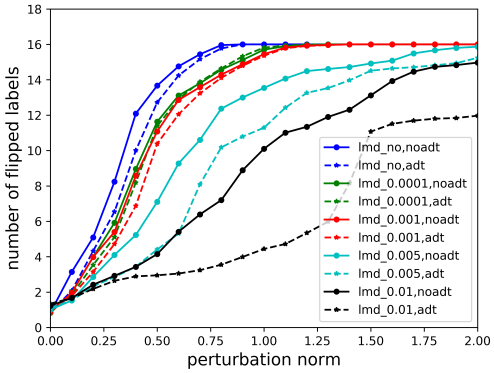
- With neither the low-rank nuclear-norm constraint nor adversarial training over the linear transformation coefficients, noted as $\mathrm{lmd\_no}$ and noadt, respectively.
- With both the nuclear norm constraint and adversarial training, noted as $\mathrm{lmd\_}\lambda$ and adt, where λ is the regularization parameter of the nuclear norm constraint.
- Without adversarial training while with the nuclear norm constraint, noted as $\mathrm{lmd\_}\lambda$ and noadt, respectively.
- With adversarial training while without the nuclear norm constraint, noted as $\mathrm{lmd\_no}$ and adt, respectively.

The attackability indicators of all complexity-controlled classifiers are estimated by the proposed GASE. The results are shown in Fig. 2. The figure also shows robustness evaluation, as a low attackability indicates a high robustness. Consistently found in all datasets, the low-rank constraint has a significant stronger impact on the classifier's attackability compared to adversarial training, because the variation of λ caused larger change among curves in different colors. Classifiers trained with larger λ are more robust. Though adversarial training alone doesn't change drastically the robustness, combined with the low-rank constraint, they can make the classifiers more robust than using solely either one. The results confirmed our remarks from Theorem 1 and 2.

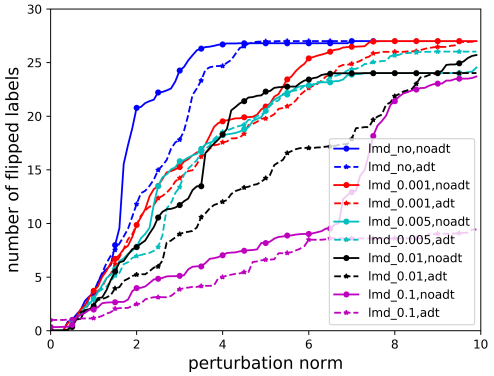
Our experimental observations show that adversarial training doesn't change drastically classifiers' performances on unperturbed test data. In contrast, there is indeed an obvious trade-off between improving a classifier's robustness by imposing the nuclear norm constraint and preserving its good utility on unperturbed test data. A strong nuclear norm constraint improves greatly the adversarial robustness. Nevertheless, it also causes accuracy loss to the classifiers. More evaluation results about the accuracy of classifiers under complexity control are in the supplementary document.

Conclusion

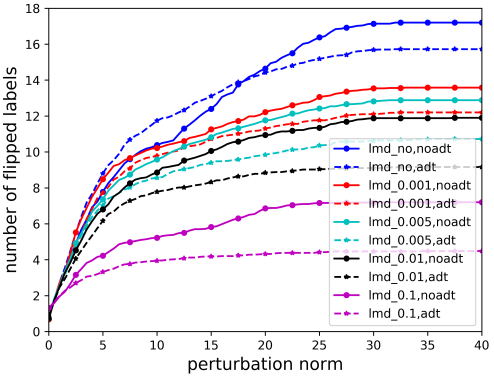
In this paper, we propose to assess the attackability of multi-label learning systems under adversarial evasion attack. We theoretically analyze the bound of the expected worst-case risk on adversarial data instances for linear and neural nets based multi-label classifiers. The resultant risk bound is used to evaluate the attackability of the targeted multi-label learning model. We unveil that the attckability depends heavily on 1) the empirical loss on the unperturbed data, 2) the rank of the targeted classifier's linear transformation coefficients and 3) the attack strength. The former two perspectives characterize the attacked multi-label learning task. The latter is decided purely by the adversary. They are the intrinsic cause and external driving force of the adversarial threat. Practically, we propose a greedy-expansion based label space exploration method to provide the empirical attackability measurement. Enjoying the submdoularity of the label space exploration problem, the empirical attackability evaluation has a certified approximation accuracy to the underlying true value. Our study intrigues the interpretability of adversarial threats of multi-label learning models. The future work will focus on proposing defensive methods for multi-learning systems with provably robustness.



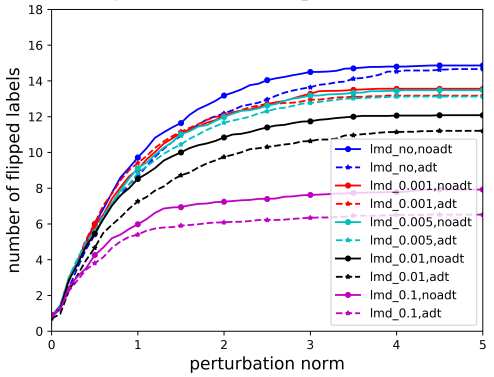
(a) attackability of controlled SVM on *Creepware*



(b) attackability of controlled SVM on *Genbase*



(c) attackability of controlled Inception-V3 on *VOC2012*



(d) attackability of controlled Inception-V3 on *Planet*

Figure 2: The evaluation of classifiers' attackability under different complexity controls.

References

A. Bojchevski, A.; and Günnemann, S. 2019. Certifiable Robustness to Graph Perturbations. In *ICML*, 8319–8330.

Athalye, A.; Carlini, N.; and Wagner, D. 2018. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. In *ICML*, volume 80, 274–283.

Battista, B.; and Fabio, R. 2018. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition* 84: 317 – 331.

Bi, W.; and Kwok, J. T. 2014. Multilabel Classification with Label Correlations and Missing Labels. In *AAAI*, 1680–1686.

Biggio, B.; Corona, I.; Maiorca, D.; Nelson, B.; Šrndić, N.; Laskov, P.; Giacinto, G.; and Roli, F. 2013. Evasion Attacks against Machine Learning at Test Time. In *ECML PKDD*.

Bojchevski, A.; and Günnemann, S. 2019. Certifiable Robustness to Graph Perturbations. In *NeurIPS*, 8319–8330.

Buchbinder, N.; Feldman, M.; Naor, J.; and Schwartz, R. 2014. Submodular maximization with cardinality constraints. In *SODA*.

Cabral, R.; la Torre, F. D.; Costeira, J. P.; and Bernardino, A. 2015. Matrix Completion for Weakly-Supervised Multi-Label Image Classification. *TPAMI* 37(1): 121–135.

Carlini, N.; and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *IEEE S&P*.

Chiang, K.-Y.; Hsieh, C.-J.; and Dhillon, I. S. 2015. Matrix Completion with Noisy Side Information. In *NIPS*, 3447–3455.

Cohen, J.; Rosenfeld, E.; and Kolter, Z. 2019. Certified Adversarial Robustness via Randomized Smoothing. In *ICML*, 1310–1320.

Cullina, D.; Bhagoji, A.; Ramchandran; and Mittal, P. 2019. PAC-Learning in the presence of adversaries. In *NeurIPS*.

Elenberg, E. R.; Khanna, R.; Dimakis, A. G.; and Negahban, S. 2016. Restricted Strong Convexity Implies Weak Submodularity. *Annuals of Statistics* .

Everingham, M.; Gool, L. V.; Williams, C.; Winn, J.; and Zisserman, A. 2012. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html".

Fawzi, A.; Moosavi-Dezfooli, S.; and Frossard, P. 2016. Robustness of Classifiers: From Adversarial to Random Noise. In *NIPS*, 1632–1640.

Florian, T.; Kurakin, A.; Papernot, N.; Boneh, D.; and McDaniel, P. 2018. Ensemble Adversarial Training: Attacks and Defenses. In *ICLR*.

Freed, D.; Palmer, J.; Minchala, D.; Levy, K.; Ristenpart, T.; and Dell, N. 2018. "A Stalker's Paradise": How Intimate Partner Abusers Exploit Technology. In *CHI*, 1–13.

Gao, R.; Cai, T.; Li, H.; Hsieh, C. J.; Wang, L.; and Lee, J. D. 2019. Convergence of Adversarial Training in Overparametrized Neural Networks. In *NeurIPS*, 13029–13040.

Gilmer, J.; Metz, L.; Faghri, F.; Schoenholz, S.; Raghu, M.; Wattenberg, M.; and Goodfellow, I. 2018. Adversarial Spheres. *CoRR* URL http://arxiv.org/abs/1801.02774.

Goldberg, A. B.; Zhu, X.; Recht, B.; Xu, J.-M.; and Nowak, R. 2010. Transduction with Matrix Completion: Three Birds with One Stone. In *NIPS*, 757–765.

Goodfellow, I.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In *ICLR*.

Guo, Y. 2017. Convex Co-Embedding for Matrix Completion with Predictive Side Information. In *AAAI*, 1955–1961.

Gupta, A.; Lamba, H.; Kumaraguru, P.; and Joshi, A. 2013. Faking Sandy: Characterizing and Identifying Fake Images on Twitter during Hurricane Sandy. In *WWW*, 729–736.

Hein, M.; and Andriushchenko, M. 2017. Formal Guarantees on the Robustness of a Classifier against Adversarial Manipulation. In *NIPS*, 2266–2276.

H.Konig. 1986. Eigenvalue Distribution of Compact Operators. *Operator Theory: Advances and Applications* 16.

Hsieh, C.-J.; Natarajan, N.; and Dhillon, I. S. 2015. PU Learning for Matrix Completion. In *ICML*, 663–672.

Jakubovitz, D.; and Giryes, R. 2018. Improving DNN Robustness to Adversarial Attacks Using Jacobian Regularization. In *ECCV*, 525–541. Springer International Publishing.

Kaggle. 2017. Planet: Understanding the Amazon from Space. https://www.kaggle.com/c/planet-understanding-the-amazon-from-space/overview.

Khim, J.; and Loh, P. 2018. Adversarial Risk Bounds for Binary Classification via Function Transformation. *arXiv* .

Lee, G.; Yuan, Y.; Chang, S.; and Jaakkola, T. 2019. Tight Certificates of Adversarial Robustness for Randomly Smoothed Classifiers. In *NeurIPS*, 4910–4921.

Lin, Z.; Ding, G.; Hu, M.; Wang, J.; and Ye, X. 2013. Image Tag Completion via Image-Specific and Tag-Specific Linear Sparse Reconstructions. In *CVPR*, 1618–1625.

Liu, D.; Hua, X.-S.; Wang, M.; and Zhang, H.-J. 2010. Image Retagging. In *ACM MultiMedia*, 491–500.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *ICLR*.

Miyato, T.; Kataoka, T.; Koyama, M.; and Yoshida, Y. 2018. Spectral Normalization for Generative Adversarial Networks. In *ICLR*.

Nicolae, M.; Sinn, M.; Minh, T. N.; Rawat, A.; Wistuba, M.; Zantedeschi, V.; Molloy, I. M.; and Edwards, B. 2018. Adversarial Robustness Toolbox v0.2.2. *CoRR* URL http://arxiv.org/abs/1807.01069.

Papernot, N.; McDaniel, P.; Wu, X.; Jha, S.; and Swami, A. 2016. Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks. In *IEEE S&P*, 582–597.

Qi, L.; Wu, L.; Chen, P.; Dimakis, A.; Dhillon, I.; and Witbrock, M. 2019. Discrete Attacks and Submodular Optimization with Applications to Text Classification. In *SysML*.

Raghunathan, A.; Steinhardt, J.; and Liang, P. 2018. Semidefinite Relaxations for Certifying Robustness to Adversarial Examples. In *NeurIPS*, 10900–10910.

Ravi, S. N.; Dinh, T.; Lokhande, V. S.; and Singh, V. 2019. Explicitly Imposing Constraints in Deep Networks via Conditional Gradients Gives Improved Generalization and Faster Convergence. In *AAAI*.

Ross, A.; and Doshi-Velez, F. 2018. Improving the Adversarial Robustness and Interpretability of Deep Neural Networks by Regularizing their Input Gradients. In *AAAI*.

Roundy, K. A.; Mendelberg, P.; Dell, N.; McCoy, D.; Nissani, D.; Ristenpart, T.; and Tamersoy, A. 2020. The Many Kinds of Creepware Used for Interpersonal Attacks. In *IEEE S&P)*, 626–643.

Shafahi, A.; Najibi, M.; Ghiasi, M. A.; Xu, Z.; Dickerson, J.; Studer, C.; Davis, L. S.; Taylor, G.; and Goldstein, T. 2019. Adversarial training for free! In *NeurIPS*, 3358–3369.

Song, Q.; Jin, H.; Huang, X.; and Hu, X. 2018. Multi-label Adversarial Perturbations. In *ICDM*, 1242–1247.

Sun, Y.-Y.; Zhang, Y.; and Zhou, Z.-H. 2010. Multi-Label Learning with Weak Label. In *AAAI*, 593–598.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I. J.; and Fergus, R. 2013. Intriguing properties of neural networks. In *ICLR*.

Tsoumakas, G.; Katakis, I.; and Vlahavas, I. 2010. *Mining Multi-label Data*, 667–685. Springer US.

Tu, Z.; Zhang, J.; and Tao, D. 2019. Theoretical Analysis of Adversarial Learning: A Minimax Approach. In *NeurIPS*, 12259–12269.

Wang, Y.; Jha, S.; and Chaudhuri, K. 2018. Analyzing the Robustness of Nearest Neighbors to Adversarial Examples. In *ICML*.

Wang, Y.; Ma, X.; Bailey, J.; Yi, J.; Zhou, B.; and Gu, Q. 2019. On the Convergence and Robustness of Adversarial Training. In *ICML*, 6586–6595.

Wang, Y.; Y.Han; Bao, H.; Shen, Y.; Ma, F.; Li, J.; and Zhang, X. 2020. Attackability Characterization of Adversarial Evasion Attack on Discrete Data. In *KDD*.

Wu, L.; Jin, R.; and Jain, A. K. 2013. Tag Completion for Image Retrieval. *TPAMI* 35(3): 716–727.

Xu, C.; Liu, T.; Tao, D.; and Xu, C. 2016. Local Rademacher Complexity for Multi-Label Learning. *IEEE Transactions on Image Processing* 25(3): 1495–1507.

Xu, H.; and Mannor, S. 2010. Robustness and Generalization. In *COLT*, 503–515.

Xu, M.; Jin, R.; and Zhou, Z.-H. 2013. Speedup Matrix Completion with Side Information: Application to Multi-label Learning. In *NIPS*, 2301–2309.

Xu, W.; Evans, D.; and Qi, Y. 2018. Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. In *NDSS*.

Yin, D.; Ramchandran, K.; and Bartlett, P. 2019. Rademacher Complexity for Adversarially Robust Generalization. In *ICML*.

Yoshida, Y.; and Miyato, T. 2017. Spectral Norm Regularization for Improving the Generalizability of Deep Learning. *ArXiv* abs/1705.10941.

Yu, H.-F.; Jain, P.; Kar, P.; and Dhillon, I. S. 2014a. Large-scale Multi-label Learning with Missing Labels. In *ICML*.

Yu, H.-F.; Jain, P.; Kar, P.; and S.Dhillon, I. 2014b. Large-scale Multi-label Learning with Missing Labels. In *ICML*.

Zhao, F.; and Guo, Y. 2015. Semi-supervised Multi-label Learning with Incomplete Labels. In *IJCAI*, 4062–4068.

Zhu, G.; Yan, S.; and Ma, Y. 2010. Image Tag Refinement Towards Low-rank, Content-tag Prior and Error Sparsity. In *ACM MultiMedia*, 461–470.

Zhu, Y.; Kwok, J. T.; and Zhou, Z.-H. 2018. Multi-Label Learning with Global and Local Label Correlation. *TKDE* .

Zugner, D.; and Gunnemann, S. 2019. Certifiable Robustness and Robust Training for Graph Convolutional Networks. In *KDD*, 246–256.

Zugner, D.; and Günnemann, S. 2020. Certifiable Robustness of Graph Convolutional Networks under Structure Perturbations. In *KDD*, 1656–1665.

Supplementary

We supple the proofs of theorems in our paper, detailed experimental setup information, experiments no Resnet50 based classifiers and case study by CW attack.

Proof of Theorem 1

We assume that the linear multi-label classifier $h(x) = wx$, where $y \in R^m$ and $x \in R^d$ are the label and feature vector of one given instance $z = (x, y)$ respectively. $w \in R^{m*d}$ denotes the linear transformation coefficient matrix of the classifier h. We require $\|w\|_\sigma \leq \Lambda$ ($\|\|_\sigma$ is the spectral norm of the coefficient matrix w).

The least-squared error (LSE) risk function that are popularly used in multi-label classification can be formulated as $\ell(x, y) = \|y - wx\|_2$. $\|\|_2$ denotes Euclidean norm. We define the distance metric in the joint space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ as:

$$d(Z, Z') = \|x - x'\|_2 + \|y - y'\|_2 \tag{10}$$

For any $x \in \mathcal{X}$ in the feature space, we require that $\|x\|_2 \leq \mu_x$. Therefore we have $d(z, z') \leq 2\mu_x + l$ for any z and z' in the joint space $\mathcal{Z}$. The upper bound of f is given by $0 \leq f \leq M$,where $M = m + \Lambda\mu_x$.

Further we have: for any z and z' in the joint space $\mathcal{Z}$, we have

$$\begin{aligned} |\ell(z') - \ell(z)| &\leq |\|wx' - y'\|_2 - \|wx - y\|_2| \\ &\leq \|w(x' - x)\|_2 + \|y' - y\|_2 \\ &\leq \|w\|_\delta \|x' - x\|_2 + \|y' - y\|_2 \\ &= C_h d(z', z) \end{aligned} \tag{11}$$

where $C_h = max\{\|w\|_\sigma, 1\}$

Lemma 3. *Let $S{:}X \longrightarrow Y$ be the operators in real Banach space and $\epsilon > 0$. Then the covering number $\mathcal{N}$ of T can be bounded as:*

$$\mathcal{N}(T, \|\|, \epsilon) \leq (1 + \frac{2\|T\|}{\epsilon})^R \tag{12}$$

where R is the rank of the operator T and $\|\|$ is the norm of the operator.

The proof of the Lemma follows the same lines as the proofs for the similar properties of entropy numbers (Section.1d)(H.Konig 1986).

Our analysis is conducted based on combining together Lemma.2 in Section.3 and Lemma.6 in Section.4 in (Tu, Zhang, and Tao 2019), which gives:

Lemma 4. *Let $R_{\mathcal{P}'}(h)$ and $R^{emp}_{\mathcal{P}'}(h)$ denote the expected and empirical worst-case risk under the evasion attack, as defined in Definition.1. We have the upper bound of $R_{\mathcal{P}'}(h)$ holds with probability at least $1 - \sigma$:*

$$\begin{aligned} R_{\mathcal{P}'}(h) &\leq R^{emp}_{\mathcal{P}'}(h) + \frac{24\kappa}{\sqrt{n}} \\ &+ M\sqrt{\frac{\log(1/\sigma)}{2n}} + \frac{12\sqrt{\pi}}{\sqrt{n}} C_h diam(Z) \\ \kappa &= \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{F}, \|\|_\infty, u/2)} du \\ R^{emp}_{\mathcal{P}'}(h) &\leq \frac{1}{n} \sum_{i=1}^n \ell(\mathbf{x}_i, \mathbf{y}_i) + C_h \mu_r \end{aligned} \tag{13}$$

where $\mathcal{N}$ denotes the covering number of the functional $\mathcal{F}$, where the loss function $f \in \mathcal{F}$. $diam(Z)$ denotes the diameter of the L2-ball of the joint space $\mathcal{Z}$. μ_r is the limit of the attack budget.

The proof of this lemma can be derived by combining the conclusion of Lemma.2 in Section.3 and Lemma.6 in Section 4 in (Tu, Zhang, and Tao 2019). Furthermore, we bound $\Lambda_{\epsilon_\mathcal{B}}$ in Lemma.6 in Section.4 with C_h in our study to indicate the impact of the spectrum of w over the expected worst-case risk bound.

To evaluate dudley entropy integral in Eq.13 to compute the covering number $\mathcal{N}(\mathcal{F}, \|\|_\infty, u/2)$, we give:

$$\begin{aligned} \|\ell - \ell'\|_\infty &= \sup_{x \in \mathcal{X}, y \in \mathcal{Y}} |\|wx - y\|_2 - \|w'x - y\|_2| \\ &\leq \|(w - w')x\|_2 \\ &\leq \mu_x \|w - w'\|_\sigma \\ &\leq 2\mu_x \|w\|_\sigma \end{aligned} \tag{14}$$

Therefore the covering number $\mathcal{N}$ is bounded base on Lemma.3:

$$\mathcal{N}(\mathcal{F}, \|\|_\infty, u/2) \leq (1 + \frac{4\mu_x\Lambda}{u})^R \tag{15}$$

Λ bounds the $\|w\|_\sigma$. R is the rank of the operator for classification risk calculation $\ell : \mathcal{X} \times \mathcal{Y} \longrightarrow R$. We can formulate f with matrix transformation:

$$\begin{aligned} h(x, y) &= wx - y = [w, -\mathbf{1}][x^T, y^T]^T \\ \ell(x, y) &= h(x, y)^T h(x, y) \end{aligned} \tag{16}$$

Therefore R is no more than the rank of w. We use R to denote the rank of w hereafter.

For $u \geq 2\mu_x\Lambda$, $\mathcal{N}(\mathcal{F}, \|\|_\infty, u/2) = 1$. We can hence absorb some positive multiplicative constants into Λ and formulate the upper bound of the covering number as follows:

$$\begin{aligned} &\int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{F}, \|\|_\infty, u/2)} du \leq \int_0^{2\mu_x\Lambda} \sqrt{\log R(1 + \frac{4\mu_x\Lambda}{u})} du \\ &\leq 2\sqrt{R(2 + 2\mu_x\Lambda)}\sqrt{2\mu_x\Lambda} \end{aligned} \tag{17}$$

Substituting this into Eq.13, we get the desired result in Theorem.1:

$$\begin{aligned} R_{\mathcal{P}'}(h) &\leq R^{emp}_{\mathcal{P}'}(h) + 96\sqrt{\frac{\mu_x\Lambda R(1 + \mu_x\Lambda)}{n}} \\ &+ \frac{12C_h\sqrt{\pi}(m + 2\mu_x)}{\sqrt{n}} + (m + \Lambda\mu_x)\sqrt{\frac{\log(1/\sigma)}{2n}} \end{aligned} \tag{18}$$

In the adversary-free learning scenario, we can derive the generalization error bound of h as follows:

Corollary 2.1. *Given the setting of the multi-label data and the linear multi-label classifier h, the upper bound of the expected generalization error of h without the adversary holds*

with at least probability of $1-\sigma$ *as follows:*

$$R_{\mathcal{P}}(h) \leq \frac{1}{n}\sum_{i=1}^{n}\ell(\mathbf{x_i},\mathbf{y_i}) + 96\sqrt{\frac{\mu_x\Lambda R(1+\mu_x\Lambda)}{n}} + \frac{12C_h\sqrt{\pi}(m+2\mu_x)}{\sqrt{n}} + (m+\Lambda\mu_x)\sqrt{\frac{\log(1/\sigma)}{2n}} \tag{19}$$

where $R_{\mathcal{P}}(h)$ *is the expected and empirical risk of* h *on the distribution* $\mathcal{P}$ *of multi-label data instances.* $R_{\mathcal{P}}^{emp}(h) = \frac{1}{n}\sum_{i=1}^{n}\ell(\mathbf{x_i},\mathbf{y_i})$ *denotes the empirical classification risk of* h *over the multi-label instances sampled from* $\mathcal{P}$.

Remark 3. *Compare the derived adversary-free risk in Eq.19 and the expected risk under adversarial perturbation in Eq.3, the extra effect that the adversary introduces is related to the magnitude of the adversarial perturbation* $\|\mu_r\|$. *The attackability depends heavily on the intrinsic regularity of the classifier's architecture and training data distribution. Notably, even in adversary-free scenario, a low-rank structured linear multi-label classifier tends to have lower generalization error. Without the presence of the adversary, all the terms involving the attack budget* μ_r *vanish. According to Eq.19, the classifier* h *with a lower-rank structure tends to have lower generalization error over the distribution* $\mathcal{P}$. *This is consistent with the observation reported in previous multi-label research efforts.*

Proof of Theorem 2

Inherited the setting of the attack scenario from Theorem.1, we consider a neural network based multi-label classifier h_{nn} with L layers, where:

- The dimension of each layer is $d_1, d_2, ..., d_L$, and $d_0 = d$ for taking input $\mathbf{x}$ and $d_L = m$ for outputting labels $\mathbf{y}$.
- At each layer i, $A_i \in R^{d_{i-1}\times d_i}$ denotes the linear coefficient matrix (connecting weights). The spectral norm of A_i is bounded as $\|A_i\|_\delta \leq \Lambda_i$. R_i denotes the rank of A_i.
- The activation functions used in the same layer are Lipschitz continuous and bounded. We assume that the activation functions used in the same layer share the same Lipschitz constant ρ_i. We use g_i to denote the activation functions used at the layer i. The output of each layer i can be defined recursively as $\mathcal{H}_i = g_i(\mathcal{H}_{i-1}A_i)$.

Again we use the least-squared error (LSE) risk function as $\ell(x,y) = \|y - wx\|_2$. $\|\|_2$ denotes Euclidean norm. We define the distance metric in the joint space $\mathcal{Z} = \mathcal{X}\times\mathcal{Y}$ as:

$$d(Z,Z') = \|x-x'\|_2 + \|y-y'\|_2 \tag{20}$$

For any $x \in \mathcal{X}$ in the feature space, we require that $\|x\|_2 \leq \mu_x$. Therefore we have $d(z,z') \leq 2\mu_x + m$ for any z and z' in the joint space $\mathcal{Z}$.

We observe that the following inequality holds:

$$\begin{aligned}
&\|\ell(x,y) - \ell(x',y')\|_2 \leq \|y-y'\|_2 + \|\mathcal{H}(x) - \mathcal{H}(x')\|_2 \\
&\|\ell(x,y) - \ell(x',y')\|_2 \leq \|y-y'\|_2 + \prod_{j=1}^{L}\rho_j\|A_j\|_\delta\|x-x'\|_2 \\
&\|\ell(x,y) - \ell(x',y')\|_2 \leq C_{nn}d_Z(z,z')
\end{aligned} \tag{21}$$

where $C_{nn} = max\{1, \prod_{j=1}^{L}\rho_j\|A_j\|_\delta\}$.

In the followings, we prepare to compute the covering number of $h_{nn} \in \mathcal{F}$. First for any two feed-forwrd neural network $\mathcal{H}_{\mathcal{A}}$ and $\mathcal{H}_{\mathcal{A}'}$, where $A = (A_1, A_2, A_3, ..., A_L)$ and $A' = (A'_1, A'_2, A'_3, ..., A'_L)$, we have the following bounds:

$$\begin{aligned}
&\|\ell - \ell'\|_\infty = \sup_{x\in\mathcal{X}, y\in\mathcal{Y}} |\|\delta_L(\mathcal{H}_{L-1}(x)A_L) - y\|_2 - \\
&\|\delta_L(\mathcal{H}'_{L-1}(x)A'_L) - y\|_2| \\
&\leq \sum_{i=1}^{L} C_i\|\mathcal{H}_{L-i}(x)(A_{L+1-i} - A'_{L+1-i})\|_2
\end{aligned} \tag{22}$$

where $C_1 = \rho_L$ and $C_i = \prod_{j=L-i+1}^{L}\rho_j\prod_{j=L+2-i}^{L}\|A_j\|_\delta$ with $L \geq i \geq 2$,

Based on the definition of the covering number, Eq.22 and Lemma.3, we derive the covering number of h_{nn} as:

$$\log(\mathcal{N}(\mathcal{F}, \|\|_\infty, u/2)) \leq \sum_{i=1}^{L} R_i \log(1+4L\frac{B_iC_i\Lambda_i}{u}) \tag{23}$$

where R_i denotes the rank of A_i. $\|\mathcal{H}_i(x)\|_2 \leq B_i$. Since the activation function of each layer is bounded, we can further assume that the activation function is Sigmoid or Tanh function. In this case, $\|\mathcal{H}_i(x)\|_2 \leq d_i$.

Therefore the upper bound of the dudley entropy integral gives:

$$\begin{aligned}
&\int_0^\infty \sqrt{\log(\mathcal{N}(\mathcal{F}, \|\|_\infty, u/2))}du \\
&\leq \sum_{i=1}^{L}\int_0^{2d_i\Lambda_i}\sqrt{R_i\log(1+4L\frac{d_iC_i\Lambda_i}{u})}2LC_idu_i \\
&\leq 4\sqrt{mL\Lambda_L}\sum_{i=1}^{L}R_i\sqrt{d_i\Lambda_iC_i}
\end{aligned} \tag{24}$$

where we absorb some positive multiplicative constants into Λ_i. We finally find that the expected adversarial risk bound of the feed-forward neural network model holds with the probability no less than $1-\delta$ as:

$$R_{\mathcal{P}'}(h_{nn}) \leq R_{\mathcal{P}'}^{emp}(h_{nn}) + 2m\sqrt{\frac{\log(1/\sigma)}{2n}} + \frac{96\sqrt{mL\Lambda_L}\sum_{i=1}^{L}R_i\sqrt{d_i\Lambda_iC_i}}{\sqrt{n}} + \frac{12C_{nn}(2\mu_x+m)\sqrt{\pi}}{\sqrt{n}} \tag{25}$$

and the empirical loss $R_{\mathcal{P}'}^{emp}$ has the upper bound:

$$R_{\mathcal{P}'}^{emp}(h_{nn}) \leq \frac{1}{n}\sum_{i=1}^{n}\ell(\mathbf{x}_i,\mathbf{y}_i) + C_{nn}\mu_r \tag{26}$$

We include in a further step the adversary-free generalization bound of h_{nn} based on Eq.25:

$$R_{\mathcal{P}}(h_{nn}) \leq \frac{1}{n}\sum_{i=1}^{n}\ell(\mathbf{x}_i,\mathbf{y}_i) + 2m\sqrt{\frac{\log(1/\sigma)}{2n}} + \frac{96\sqrt{mL\Lambda_L}\sum_{i=1}^{L}R_i\sqrt{d_i\Lambda_iC_i}}{\sqrt{n}} + \frac{12C_{nn}(2\mu_x+m)\sqrt{\pi}}{\sqrt{n}} \tag{27}$$

Remark 4. *The generalization risk bound given in Eq.27 illustrates explicitly the association between the low rank linear transformation coefficients in the neural network based classifier and its generalization capability for multi-label classification. As unveiled, at least one layer of the neural network should be of a low-rank structure, in order to improve its expected classification accuracy over unknown testing samples. Furthermore, lower spectral norm (smaller leading eigenvalues) of the linear coefficients can also improve the generalization capability and robustness under the evasion attack. The analysis also unveils the relation between generalization capability and adversarial robustness of a multi-label classifier.*

Notably, (Tu, Zhang, and Tao 2019) gave an adversarial risk bound of a feed-forward neural network model used for binary classification. However, in multi-label classification scenarios, we are especially curious whether a multi-label classifier can gain adversarial robustness from the low rank constraint*. The attackability analysis in our work thus provides an explicit answer to the question.*

Proof of Lemma 1 and Lemma 2

We first verify the supermodularity of $g(S)$ in Eq.7. $g(S)$ is a non-decreasing set function with increasingly larger S. We first derive an analytical solution to $g(S)$ with lagrangian multipliers $\{\lambda_i\}$:

$$\begin{aligned} & J(\lambda_i, r) = \|r\|^2 \\ & + \sum_{i=1}^{m} \lambda_i (2b_i y_i h_i(x+r) - y_i h_i(x+r) + t_i) \\ s.t.\ & \lambda_i \geq 0 \\ & b_i = 1 \ \ for \ \ i \in T \\ & b_i = 0 \ \ for \ \ i \notin T \end{aligned} \tag{28}$$

where h_i denotes the output of the classifier h corresponding to the i-th label. Since the adversarial noise r is usually of small magnitude, we further approximate $h_i(x+r)$ with its Taylor expansion: $h_i(x+r) \approx h_i(x) + r^T h_i'(x)$. Eq.28 can be further simplified as a quadratic programming problem with affine constraints:

$$\begin{aligned} & J(\lambda_i, r) = \|r\|^2 \\ & + \sum_{i=1}^{m} \lambda_i (2b_i y_i (h_i + r^T \phi_i) + t_i - y_i h_i - y_i r^T \phi_i) \\ s.t.\ & \lambda_i \geq 0 \\ & b_i = 1 \ \ for \ \ i \in T \\ & b_i = 0 \ \ for \ \ i \notin T \end{aligned} \tag{29}$$

where $\phi_i(x) = h_i'(x)$ denotes the gradient of $h_i(x)$ with respect to the input feature vector x.

By taking the first-order condition $\frac{\partial J}{\partial r} = 0$, we can derive the optimal $\|r\|^2$ as :

$$\|r\|^2 = \frac{1}{4} \| \sum_{i=1}^{m} (\lambda_i y_i \phi_i - 2\lambda_i \phi_i b_i y_i) \|_2^2 \tag{30}$$

and according to KKT conditions, we can get for non-zero λ_i:

$$\begin{aligned} & h_i + \frac{1}{2} \sum_{k \in \{k | \lambda_k > 0\}} (\lambda_k y_k \phi_k^T \phi_i - 2\lambda_k \phi_k^T \phi_i \hat{y}_k) = 0 \\ & (if \ \lambda_i > 0) \end{aligned} \tag{31}$$

Observation 2.1. *To obtain the values of $\lambda_k > 0$ ($k = 0, 1, 2, ..., K$), we turn to solve the equation system such as:*

$$\begin{aligned} & -\frac{1}{2} \Pi [\Phi_0^T, \Phi_1^T, ..., \Phi_K^T] = H \\ & \Pi = [\lambda_0 \phi_0^T - 2b_0 y_0 \phi_0^T, \lambda_1 \phi_1^T - 2b_1 y_1 \phi_1^T, \\ & ..., \lambda_K \phi_K^T - 2b_K y_K \phi_K^T] \\ & \Phi_k = [\phi_k, \phi_k, ..., \phi_k] \\ & H = [h_0, h_1, h_2, ..., h_K] \end{aligned} \tag{32}$$

Given a set of $\{b_0, b_1, ..., b_K\}$, the value of λ_k ($k = 0, 1, 2, ..., K$) is determined uniquely by h_i and ϕ_i.

Observation 2.2. *$\|r\|^2$ is a set function defined over the set T, as $\hat{y}_i$ is determined by the binary variable b_i.*

Observation 2.3. *$\|r\|^2$ is a convex quadratic function with respect to the variable $\{\hat{y}_i\}, (i = 1, 2, 3, ..., m)$, given a set of $\{b_0, b_1, b_2, ...b_K\}$ fixed in Eq.30.*

Furthermore, by simply flipping the sign of $g(S)$, we can find that $-g(S) = \min_S - \|r\|^2$ is non-increasing and submodular function, since $-\|r\|^2$ is concave, according to Theorem.1 in (Elenberg et al. 2016). Correspondingly, $g(S)$ is a non-decreasing supermodular set function. In Eq.7, $|S|$ is a monotonically increasing modular function. As a result, the objective $\psi(S) = |S| - g(S)$ of the maximization problem defined in Eq.7 is a non-monotone submodular function. According to Theorem 1.5 in (Buchbinder et al. 2014), randomized greedy forward expansion of the set S can provide a guarantee to the approximation accuracy:

$$\psi(\hat{S}) \geq \frac{1}{4} \psi(S^*) \tag{33}$$

where $\psi(\hat{S})$ and $\psi(S^*)$ denote respectively the objective function value obtained by randomized greedy forward search proposed in (Buchbinder et al. 2014) and the underlying global optimum following the cardinality lower bound constraint.

Proof of Lemma 2

In each iteration of the greedy forward search, the current set of flipped labels is noted as T and the current perturbed input is noted as $\tilde{x}$. $h(\tilde{x})$ and $\phi(\tilde{x})$ denotes the current classifier output and gradient vector with respect to $\tilde{x}$. We further assume that the i^*-th label is selected to be flipped and added to the set T in the current iteration of the greedy search. Given the h, we inject a small adversarial perturbation r to form the perturbed input $\tilde{x} + r$ centering at $\tilde{x}$, in order to flip the label i^*. By taking $\frac{\partial J}{\partial r} = 0$ and $\frac{\partial J}{\partial \lambda_i} = 0$ and the

complementary slackness conditions, we have:

$$\|r_{-i^*}\|_2 \leq \frac{1}{2}\|\sum_{j\neq i}^{m}(\lambda_j y_j \phi_j - 2\lambda_j \phi_j \hat{y}_j)\|_2$$
$$+ \frac{1}{2}\|\lambda_{i^*} y_{i^*} \phi_{i^*}\|_2$$
$$y_i h_i + t_i$$
$$= -\frac{y_i}{2}(\sum_{j\neq i}(\lambda_j y_j \phi_j^T \phi_i - 2\hat{y}_j \lambda_j \phi_j^T \phi_i)) + \frac{\lambda_i \|\phi_i\|^2}{2}, \quad i \in T$$
$$- y_i h_i + t_i$$
$$= \frac{y_i}{2}(\sum_{j\neq i}(\lambda_j y_j \phi_j^T \phi_i - 2\hat{y}_j \lambda_j \phi_j^T \phi_i)) - \frac{\lambda_i \|\phi_i\|^2}{2}, \quad i \notin T$$
$$\lambda_i(2b_i y_i h_i - y_i h_i + t_i + (2b_i y_i - y_i) r^T \phi_i) = 0 \tag{34}$$

where r_{-i^*} denotes the adversarial perturbation that can flip both the labels in the current attacked label set T and the latest added label i^*.

Observation 2.4. *To minimize $\|r_{-i^*}\|_2$, we can set $\lambda_j = 0$, $(j \neq i^*)$ and $\lambda_{i^*} > 0$. In this case, we can derive a feasible solution:*

$$r_{-i^*} = -\frac{1}{2}\lambda_{i^*} \phi_{i^*} y_{i^*}. \tag{35}$$

Furthermore, we can observe that the marginal gain of the greedy search is then proportional to $\lambda_{i^*}\|\phi_{i^*}\|$ of the candidate label i^*. To minimize the marginal gain, we choose the candidate label i^* producing the minimal $\lambda_{i^*}\|\phi_{i^*}\|$.

By taking $\frac{\partial J}{\partial \lambda_i} = 0$ and substituting the above expression of r, we can obtain:

$$\lambda_{i^*} = \frac{2(t_{i^*} + y_{i^*} h_{i^*}(\tilde{x}))}{\|\phi_{i^*}\|_2^2} \tag{36}$$

Consequently, we can derive that the upper bound of the required adversarial perturbation norm r_{-i^*} as follows:

$$\frac{|y_{i^*} h_{i^*}(\tilde{x}) + t_{i^*}|}{\|\phi_{i^*}\|_2} \geq \|r_{-i^*}\|_2 \tag{37}$$

As indicated by Eq.37 and $t_i > 0$, $\|r_{-i^*}\|_2$ is proportional to the ratio $\frac{|y_{i^*} h_{i^*}(\tilde{x})|}{\|\phi_{i^*}\|_2}$. It gives the conclusion of Lemma.2 in Section.4. Based on Eq.35 and Eq.36, we can find that $\|\frac{\partial \|r_{-i^*}\|_2}{y_i}\|_2 = \frac{|y_i h_i(\tilde{x}) + t_i|}{\|\phi_i\|_2}$. Therefore, the greedy feedforward expansion can be considered as conducting orthogonal matching pursuit based greedy search for the submodular function maximization problem (Elenberg et al. 2016).

Experimental Setup

Dataset Information

We include 4 datasets collected from various real-world multi-label cyber security practices (*Creepware*) biology research (*Genbase*)(Tsoumakas, Katakis, and Vlahavas 2010), object recognition (*VOC2012*)(Everingham et al. 2012) and environment research (*Planet*(Kaggle 2017)). Except from the well-known *VOC2012* dataset, *Creepware* data include 966 stalkware app instances intercepted by the mobile AV service of a private security vendor. Each app has 16 labels indicating different types of surveillance on the victim's mobile device. The surveillance types include malicious remote control functions, such as recording messages/phone calls/call logs, logging key pressing, tracking GPS locations, extracting photos, remotely accessing cameras of the victim's mobile device and so on. Each app is profiled by the introductory texts of the app available in the third-party app stores and signatures of its mobile service access. *Genbase* dataset contains 662 proteins. Each protein molecule may belong to one or more classes among the 10 protein families concerned in a bio-medicine clinical study. One protein molecule is described by a binary string, denoting whether or not a specific signature of the molecule structure is present. *Planet* data collects daily satellite imagery of the entire land surface of the earth at 3-5 meter resolution. Each image is equipped with labels denoting different atmospheric conditions and various classes of land cover/land use.

Implementation Platforms

Software Platform: Our codes were implemented in Python and all the models were built by Keras package. **Hardware Platform:** Our experiments were conducted on GPU rtx2080ti.

Targeted Classifiers

To instantiate the attackability analysis, we study empirically two types of the targeted multi-label classifiers: linear Support Vector Machine (SVM) and Deep Neural Nets (DNN) based classifier. On *Creepware*, linear SVM is applied on the TF-IDF feature vectors extracted from the descriptive texts and categorical service access signatures. On *Genbase*, we use directly the binary strings as input features for classification. The DNN models of Inception-V3 based and Resnet50 based multi-label classifiers are used for *VOC2012* and *Planet*. Specially, we replace the last layer of Inception-V3 and Resnet50 by m logistic regression structures to build multi-label classifiers. We show the performance of Resnet50 classifiers on unperturbed test instances in Table 2.

Table 2: The F1-scores of Resnet50 based classifiers on different datasets

Dataset	N	m	l_{avg}	Micro F1	Macro F1	Classifier$_{target}$
VOC2012	17,125	20	1.39	0.79	0.68	Resnet50
Planet	40,479	17	2.87	0.78	0.35	Resnet50

Attack and Adversarial Training

In addition to *Projected gradient decent* (PGD) (Madry et al. 2018) based attack method, we also adopt *Carlini-Wagner* attack (CW) to conduct the step of targeted attack of Algorithm.1 in our study on the dataset *Creepware*.

Our theoretical and empirical evaluation of attackability are both defined regardless of the concrete choices of evasion attack method. Correspondingly, we involve different attack methods at the core of the label space exploration. The aim is to verify the attack-method-independence of the proposed attackability analysis.

Performance Benchmark

We vary the limit of the attack budget ε to denote the attack strength. Given a fixed value of ε, we calculate **the average number of flipped labels on test data as an estimator of the empirical classification risk $R^{emp}_{\mathcal{P}'}$ induced by the attack. This is defined as the empirical attackability indicator, as given in the design of fast greedy attack space exploration**.

First, to evaluate the effectiveness of the proposed empirical attackability indicator, we compute the averaged number of flipped labels only on the test instances of which all the labels are correctly classified. The propose is to assess whether the proposed greedy expansion method can identify significantly more labels as the feasible attack target, compared to well established baseline exploration strategies. **Second**, in order to verify the theoretical analysis over the countermeasures in the attackability assessment, we calculate the averaged number of miss-classified labels caused by the adversarial perturbation over all the testing data instances. It is consistent with the definition of the empirical attackability estimator in Eq.2. The resultant empirical attackability indicator is used to demonstrate whether the countermeasures can help to mitigate the threat of evasion attack.

Baselines in Validation of Empirical Attackability Indicator

We assess the effectiveness of the proposed greedy expansion based empirical attackability indicator. Especially, we involve four baselines of label exploration strategies to search for maximal label perturbation.

- **PGS** (Primitive Greedy Search): In each round of the greedy search, the primitive greedy search method calculates the magnitude of r with the combination of the current set S and each of the candidate labels. Then it chooses the label contributing the least increasing of $\|r\|^2$. Though PSG can achieve the exact greedy search, it requires to run evasion attack for each candidate label. Therefore it is significantly heavier than the proposed **GASE** method.
- **RS** (Random Search): In each round of RS, one label is selected purely randomly from the candidate set and added to the current set S. Randomized search doesn't pursue to optimize the exploration objective function in Eq.7. It is involved to show the necessity of the heuristic rule in the label exploration, such as the principle of the greedy search.
- **OS** (Oblivious Search): The oblivious method doesn't conduct iterative expansion of the set. This method first compute the norm of the adversarial perturbation induced by flipping each candidate label and keeping the other labels unchanged. The labels causing the least perturbation magnitudes are selected to form the set S. It is required to check if flipping all the labels in S can deliver a feasible evasion attack.
- **LS** (Loss-guided Search) : In each iteration of the LS method, it updates the adversarial perturbation r along the direction where the multi-label classification loss is increased the most. The iterative update of r is stopped until $\|r\|^2$ surpasses the cost limit. The set of the attacked labels given the derived r are reported. **LS** doesn't use any attack method in its implementation. It is simply a gradient ascent process. Maximizing the loss only, though sounding reasonable, is a rough search strategy. The increasing of the loss can be caused by pushing originally miss-classified labels even further from the correct decision, instead of flipping originally correctly predicted labels. As a result, it misleads the search of the attackable labels.

The Influence of Imposed Nuclear Norm Constrain on Clean Test Data

We adopt the tech in (Ravi et al. 2019) to impose nuclear norm constraint. Specially, its implementation in package *tensorflow-addons* was adopted in our experiments. Our experimental observation in Table.3 shows that there is a obvious trade-off between improving a classifier's robustness by imposing the nuclear norm constraint and preserving its good utility on unperturbed test data. A strong nuclear norm constraint improves greatly the adversarial robustness. Nevertheless, it also causes accuracy loss of the classifier.

Validation of Empirical Attackability Indicator on Resnet50

Fig.3 and 4 show the number of flipped labels obtained by the proposed *GASE* algorithm and the baselines on Resnet50 based deep multi-label classifiers over dataset *VOC2012* and *Planet*. The results verify again the reasonableness of our greedy search strategy stated in Lemma 1, since GASE and PGS achieve mre flipped labels than other baselines.

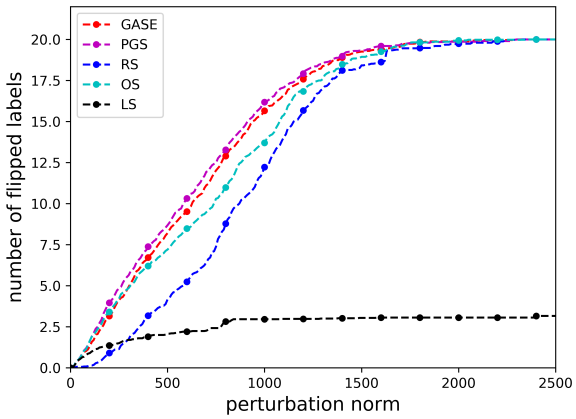


Figure 3: The empirical attackability indicator estimated by different label exploration strategies. Attackability of Resnet50 on *VOC2012*

Table 3: Classification results on adversary-free test data. The classifiers are trained with increasingly stronger nuclear norm based constraints

Dataset \ Classifier	No constraint	$\lambda = 1e-4$	$\lambda = 1e-3$	$\lambda = 5e-2$	$\lambda = 1e-2$	$\lambda = 1e-1$
Creepware SVM	$micro\_F_1 : 0.760$ $macro\_F_1 : 0.662$	$micro\_F_1 : 0.764$ $macro\_F_1 : 0.646$	$micro\_F_1 : 0.752$ $macro\_F_1 : 0.620$	$micro\_F_1 : 0.679$ $macro\_F_1 : 0.476$	$micro\_F_1 : 0.600$ $macro\_F_1 : 0.343$	$micro\_F_1 : 0.471$ $macro\_F_1 : 0.110$
Genbase SVM	$micro\_F_1 : 0.991$ $macro\_F_1 : 0.733$	$micro\_F_1 : 0.994$ $macro\_F_1 : 0.737$	$micro\_F_1 : 0.991$ $macro\_F_1 : 0.725$	$micro\_F_1 : 0.982$ $macro\_F_1 : 0.677$	$micro\_F_1 : 0.929$ $macro\_F_1 : 0.546$	$micro\_F_1 : 0.634$ $macro\_F_1 : 0.209$
VOC2012 Inception-V3	$micro\_F_1 : 0.827$ $macro\_F_1 : 0.736$	$micro\_F_1 : 0.826$ $macro\_F_1 : 0.727$	$micro\_F_1 : 0.825$ $macro\_F_1 : 0.736$	$micro\_F_1 : 0.823$ $macro\_F_1 : 0.725$	$micro\_F_1 : 0.819$ $macro\_F_1 : 0.709$	$micro\_F_1 : 0.602$ $macro\_F_1 : 0.143$
Planet Inception-V3	$micro\_F_1 : 0.822$ $macro\_F_1 : 0.361$	$micro\_F_1 : 0.818$ $macro\_F_1 : 0.354$	$micro\_F_1 : 0.823$ $macro\_F_1 : 0.360$	$micro\_F_1 : 0.819$ $macro\_F_1 : 0.355$	$micro\_F_1 : 0.819$ $macro\_F_1 : 0.347$	$micro\_F_1 : 0.695$ $macro\_F_1 : 0.192$
VOC2012 Resnet50	$micro\_F_1 : 0.788$ $macro\_F_1 : 0.683$	$micro\_F_1 : 0.783$ $macro\_F_1 : 0.670$	$micro\_F_1 : 0.785$ $macro\_F_1 : 0.677$	$micro\_F_1 : 0.780$ $macro\_F_1 : 0.672$	$micro\_F_1 : 0.773$ $macro\_F_1 : 0.656$	$micro\_F_1 : 0.643$ $macro\_F_1 : 0.261$
Planet Resnet50	$micro\_F_1 : 0.778$ $macro\_F_1 : 0.352$	$micro\_F_1 : 0.790$ $macro\_F_1 : 0.355$	$micro\_F_1 : 0.794$ $macro\_F_1 : 0.353$	$micro\_F_1 : 0.804$ $macro\_F_1 : 0.357$	$micro\_F_1 : 0.795$ $macro\_F_1 : 0.335$	$micro\_F_1 : 0.710$ $macro\_F_1 : 0.219$

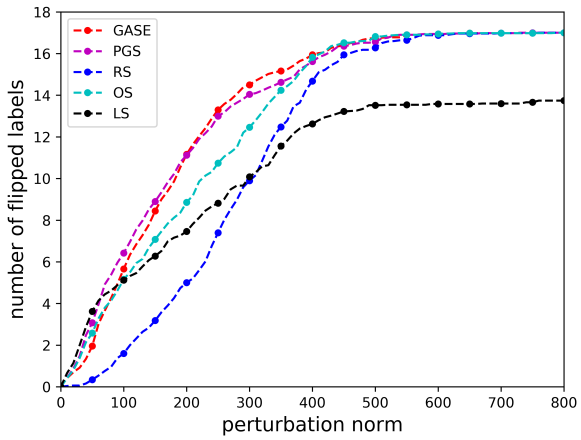


Figure 4: The empirical attackability indicator estimated by different label exploration strategies. Attackability of Resnet50 on *Planet*

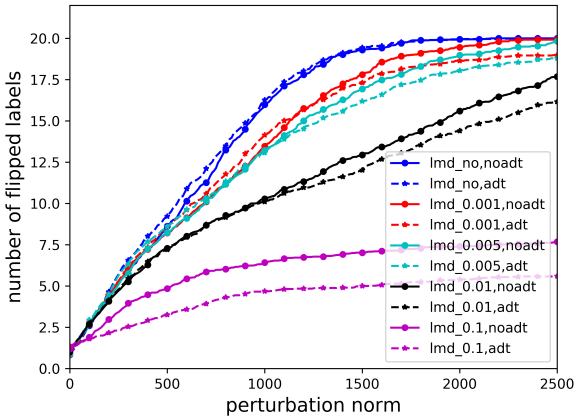


Figure 5: The evaluation of classifiers' attackability under different complexity controls.. Attackability of controlled Resnet50 on *VOC2012*

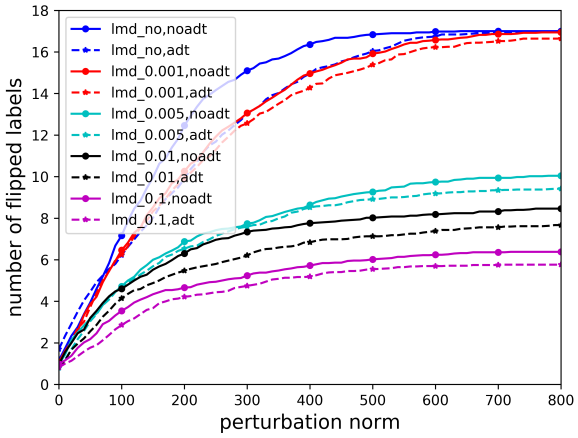


Figure 6: The evaluation of classifiers' attackability under different complexity controls.. Attackability of controlled Resnet50 on *Planet*

Attackability Evaluation with Countermeasures for Evasion Attack on Resnet50

Fig.5 and 6 show the attackability evaluation results of Resnet based classifiers on dataset *VOC2012* and *Planet* under different complexity controls and adversarial training. The trend is similar to the results of Inception-V3 based classifiers, that is in general, reducing model complexity and adversarial training can improve classifiers' adversarial robustness.

Case Study on Dataset *Creepware* with CW Attack

We replace the targeted evasion attack method PGD used in previous experiments by Carlini-Wagner (CW) to verify the attack-method-independence of 1) our proposed greedy based label exploration strategy and 2) the attackability analysis in terms of controlling model complexity and adversarial training. Fig. 7 and 8 show the indicator estimation and attackability evaluation results of SVM based classifiers on dataset *Creepware* respectively, notably, here we used CW attack to achieve the targeted evasion attack. The results are similar to the results with PGD attack, especially the indicator estimation results. In Fig. 7, the result of LS baseline is not included, as this baseline is independent of targeted evasion attack method.

In Fig. 8, when the nuclear norm-based constraint is

strong and adversarial training is conducted simultaneously, the number of flipped labels using the proposed attackability indicator with CW attack is smaller than that derived by using the PGD-based attack (showed in Fig. 7). The reason is that we constraint the range of line search tuning in CW attack to control the time cost of the attack step.

Notably, the results in Fig. 8 confirm the improvement of adversarial robustness by introducing adversarial training and the low-rank constraint over the classifier's parameters. These two approaches reduces the empirical worst-case risk over the adversarial samples and controls the classifier's complexity respectively, which in turn reduces the expected worst-case risk. The results are consistent with our theoretical analysis of the classifier's attackability. Furthermore, we confirm that the empirical attackability evaluation can use any targeted evasion attack method as its component.

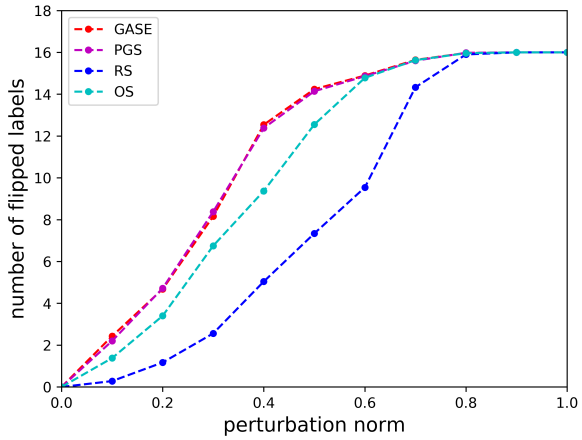


Figure 7: The empirical attackability indicator estimated by different label exploration strategies. Attackability of SVM on *Creepware*. The targeted evasion attack is achieved by CW attack.

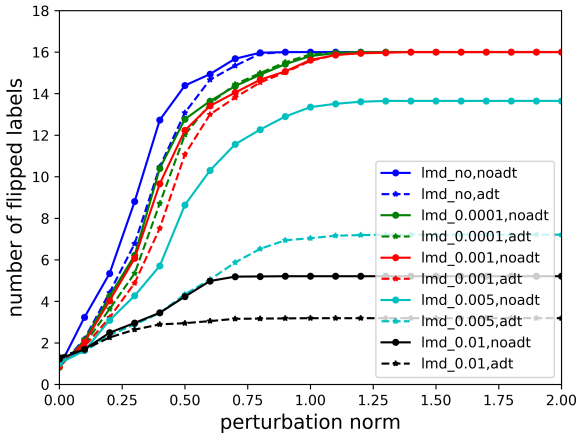


Figure 8: The evaluation of classifiers' attackability under different complexity controls. Attackability of controlled SVM on *Creepware*. The targeted evasion attack is achieved by CW attack.