Attack Transferability Characterization for Adversarially Robust Multi-label Classification

Zhuo Yang, Yufei Han, Xiangliang Zhang

Attack Transferability Characterization for Adversarially Robust Multi-label Classification

Zhuo Yang,[1] Yufei Han, [2] Xiangliang Zhang [1]

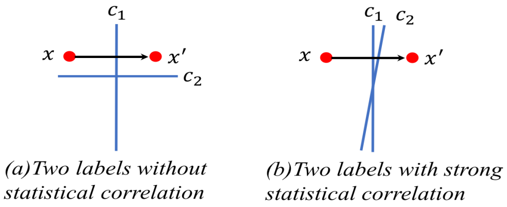
[1] King Abdullah University of Science and Technology, Thuwal, Saudi Arabia
zhuo.yang@kaust.edu.sa, xiangliang.zhang@kaust.edu.sa
[2] CIDRE team, Inria, France
yfhan.hust@gmail.com

Abstract. Despite of the pervasive existence of multi-label evasion attack, it is an open yet essential problem to characterize the origin of the adversarial vulnerability of a multi-label learning system and assess its attackability. In this study, we focus on non-targeted evasion attack against multi-label classifiers. The goal of the threat is to cause miss-classification with respect to as many labels as possible, with the same input perturbation. Our work gains in-depth understanding about the multi-label adversarial attack by first characterizing the transferability of the attack based on the functional properties of the multi-label classifier. We unveil how the transferability level of the attack determines the attackability of the classifier via establishing an information-theoretic analysis of the adversarial risk. Furthermore, we propose a transferability-centered attackability assessment, named Soft Attackability Estimator (SAE), to evaluate the intrinsic vulnerability level of the targeted multi-label classifier. This estimator is then integrated as a transferability-tuning regularization term into the multi-label learning paradigm to achieve adversarially robust classification. The experimental study on real-world data echos the theoretical analysis and verify the validity of the transferability-regularized multi-label learning method.

Keywords: Attackability of multi-label models · Attack transferability · Adversarial risk analysis · Robust training.

1 Introduction

Adversarial evasion attack against real-world multi-label learning systems can not only harm the system utility, but also facilitate advanced downstreaming cyber meances [16]. For example, hackers embed toxic contents into images while hiding the malicious labels from the detection [9]. Stealthy harassment applications, such as phone call dictation and photo extraction, carefully shape the app function descriptions to evade from the sanitary check of app stores [15,7]. Despite of the threatening impact, it remains an open problem to characterize key factors determining the attackability of a multi-label learning system. Compared to in-depth adversarial vulnerability study of single-label learning



(a)Two labels without statistical correlation

(b)Two labels with strong statistical correlation

Fig. 1: **A toy example of multi-label evasion attack.**

problems [23,6,26,21,19], this is a rarely explored, yet fundamental problem for trustworthy multi-label classification.

We focus on the non-targeted evasion attack against multi-label classifiers. In contrast to the single-label learning problem, the goal of the adversarial threat is to kill multiple birds with one stone: it aims at *changing as many label-wise outputs as possible simultaneously, with the same input.* Fig.1 demonstrates a toy example of the threat scenario with two labels l_1 and l_2, with decision hyperplanes c_1 and c_2, respectively. Fig.1 (a) assumes no statistical correlation between the two labels. c_1 and c_2 are orthogonal therefore. In contrast, the two boundaries are well aligned in Fig.1 (b), implying a strong correlation between l_1 and l_2. The injected evasion noise in both scenarios has the same magnitude to change x to be x', indicating the same attack strength. As we can see, the evasion attack can flip simultaneously the classifier's output with respect to both labels in Fig.1 (b), due to the alignment between the decision boundaries of l_1 and l_2. However, in Fig.1 (a), the evasion perturbation can only bring impacts to the decision output of l_1. As shown in the toy example, *whether the attack can transfer across different labels depends on the alignment of the decision hyperplanes, which is determined intrinsically by the correlation between the labels.* On the closely correlated labels, the multi-label classifier tends to produce the consistently same or converse decisions. The adversarial noise that successfully perturbs the decision over one label is likely to cause miss-classification on the other labels. Bearing the goal of the non-targeted attack in mind, the transferability of the attack is closely related to the adversarial vulnerability of the targeted multi-label learning system. With a more transferable attack noise, the multi-label learning system is more attackable.

Given a multi-label learning task, our study aims at gaining in-depth understanding about the theoretical link between the transferability of the evasion attack across different labels and the adversarial vulnerability of the classifier. More specifically, we focus on characterizing the role of attack transferability in determining the adversarial risk of a multi-label classifier. Furthermore, we pursue a qualitative assessment of attack transferability based on the intrinsic functional properties of the classifier. It is beneficial to not only evaluate the attackability of the classifier, but also design a transferability-suppression regularization term to enhance the adversarial robustness of the classifier. In the community of multi-label learning, it is a well-known fact that capturing the correlation between labels helps to train accurate multi-label classifiers. However, our analysis unveils the other side of the story: encoding the label correlation can also make the classifier vulnerable in the evasion attack scenarios. Our contribution can be summarized as in the followings:

- We unveil the three key factors determining the adversarial risk of a multi-label classifier by establishing an information-theoretic upper bound of the adversarial risk. They are i) the conditional mutual information (CMI) between the training data and the learnt classifier [17]; ii) the transferability level of the attack; and iii) the strength of the evasion attack. Theoretical discussions over the first two factors unveil a dilemma: Encoding label correlation in the training data is the key-to-success in accurate adversary-free multi-label classification. However, it also increases the transferability of the attack, which makes the classifier vulnerable (with a higher adversarial risk).
- We propose an attackability assessment in Section.4 based on the unveiled link between the attack transferability and the adversarial risk. This attackability assessment is then integrated into the multi-label learning paradigm as a regularization term to suppress the attack transfer and enhance the adversarial robustness of the derived multi-label classifier.
- Our empirical study with real-world multi-label learning applications instantiates the theoretical study with both linear and deep multi-label models. The results confirm the trade-off between the utility of the classifier and its adversarial robustness by controlling the attack transferability.

2 Related work

Bounding adversarial risk has been studied extensively in single-label learning scenarios [10,5,8,10,20,25,13,26,19,21,23,6]. They focus on identifying the upper bound of adversarial noise, which guarantees the stability of the targeted classifier's output, a.k.a. adversarial sphere. Notably, [5,10,25,19] study the association between adversarial robustness and the curvature of the learnt decision boundary. Strengthened further by [25,13,19], the expected classification risk under adversarial perturbation can be bounded by the model's Rademacher complexity of the targeted classifier. [24] extends the model complexity-dependent analysis to the multi-label learning problems and associates the Rademacher complexity with the nuclear norm of the model parameters.

Distinguished from single-label learning scenarios, the key-to-success of training an accurate multi-label classifier is to capture the correlation between the labels. More specifically, the alignment between the decision hyper-planes of the correlated labels helps to predict the occurrence of the labels. However, as revealed in [16,24], the evasion attack perturbation can transfer across the correlated labels: *the same input perturbation can affect the decision output of these labels.* It implies that the label correlation can be potentially beneficial to adversaries at the same time. Nevertheless, the relation between the transferability of the input perturbation and the adversarial vulnerability of the victim classifier can not be characterized or measured by the Rademacher-complexity-based analysis conducted on the single-label case and [24]. Our work thus focuses on addressing the essential yet open problem from two perspectives. First, we target on establishing a theoretical link between the transferability measurement of the attack noise across multiple labels and the vulnerability of the classifier. Second, we conduct an information-theoretic analysis of the adversarial risk, which is an attack-strength-independent vulnerability assessment. This assessment can be used to guide proactive model hardening, e.g. robust model training, to improve the adversarial robustness of the classifier.

3 Vulnerability Assessment of Multi-label Classifiers

Notations. We use $z = (x, y)$ as a multi-label instance, with feature vector $x \in \mathbb{R}^d$ and label vector $y = \{-1, 1\}^m$, where d and m denote the feature dimension and the number of labels, respectively. Specially, we use x_i and y^i to denote the feature vector and the label vector of instance z_i respectively and use y_j to denote the j-th element of label vector y. Let $\mathcal{D}$ be the underlying distribution of z and z^n be a data set including n instances. Let h denote the multi-label classifier to learn from the data instances sampled from $\mathcal{D}$. The learning paradigm (possibly randomized) is thus noted as $\mathcal{A} : z^n \rightarrow h$. The probability distribution of the learning paradigm is $\mathcal{P}_{\mathcal{A}}$. The corresponding loss function of $\mathcal{A}$ is $\ell : h \times z \rightarrow \mathbb{R}$. $\|x\|_p$ $(p \geq 1)$ denotes the L_p norm of a vector x. Without loss of generality, we choose $p = 2$ hereafter.

Attackability of a Multi-label Classifier. The attackability of h is defined as the expected maximum number of flipped decision outputs by injecting the perturbation r to x within an attack budget ε:

$$\begin{aligned} & C^*(\mathcal{D}) = \mathop{\mathbb{E}}_{z\sim\mathcal{D}}\left(\max_{T,\|r^*\|\leq\varepsilon} \sum_{j=1}^{m} \mathbb{1}(y_j \neq sgn(h_j(x+r^*)))\right), \\ & \text{where } \; r^* = \underset{r}{argmin}\, \|r\|_p\,, \\ & s.t. \; y_j h_j(x+r^*) \leq 0 \; (j \in T), \;\; y_j h_j(x+r^*) > 0 \, (j \notin T). \end{aligned} \tag{1}$$

T denotes the set of the attacked labels. $h_j(x + r)$ denotes the decision score of the label j of the adversarial input. $\mathbb{1}(\cdot)$ is the indicator function. It is valued as 1 if the attack flips the decision of the label j and 0 otherwise. With the same input x and the same attack strength $\|r\|_p$, one multi-label classifier h is more vulnerable to the evasion attack than the other h', if $C^*_h > C^*_{h'}$.

3.1 Information-theoretic Adversarial Risk Bound

Solving Eq.(1) directly for a given data instance z reduces to an integer programming problem, as [24] did. Nevertheless, our goal is beyond solely empirically assessing the attackability of h on a given set of instances. We are interested in **1)** establishing an upper bound of the expected miss-classification risk of h with the presence of adversary. It is helpful for characterizing the key factors deciding the adversarial risk of h; **2)** understanding the role of the transferability of the input perturbation across different labels in shaping the adversarial threat.

For a multi-label classifier h, n legal instances $z^n = \{z_i\}$ (i=$1, 2, ..., n$, $z_i = (x_i, y^i) \sim \mathcal{D}$) and the attack budget ε, we can estimate the expected adversarial risk of h by evaluating the worst-case classification risk over the neighborhood $N(z_i) = \left\{(x'_i, y^i) \middle| \|x'_i - x_i\|_p \leq \varepsilon\right\}$. The expected and empirical adversarial risk $R_{\mathcal{D}}(h, \varepsilon)$ and $R^{emp}_{\mathcal{D}}(h, \varepsilon)$ give:

$$\begin{aligned} R_{\mathcal{D}}(h, \varepsilon) &= E_{\mathcal{A}, z^n \sim \mathcal{D}^n}[E_{z \sim \mathcal{D}}[\max_{(x', y) \in N(z)} \ell(h(x'), y)]], \;\; h = \mathcal{A}(z^n), \\ R^{emp}_{\mathcal{D}}(h, \varepsilon) &= E_{\mathcal{A}, z^n \sim \mathcal{D}^n}[\frac{1}{n}\sum_{i=1}^{n}[\max_{(x'_i, y^i) \in N(z_i)} \ell(h(x'_i), y^i)]], \;\; h = \mathcal{A}(z^n). \end{aligned} \tag{2}$$

The expectation in Eq.(2) is taken with respect to the joint distribution $\mathcal{D}^{\otimes n} \otimes \mathcal{P}_{\mathcal{A}}$ and $\mathcal{D}^n$ denotes the data distribution with n instances. The expected adversarial risk $R_{\mathcal{D}}(h, \varepsilon)$ reflects the vulnerability level of the trained classifier h. Intuitively, a higher $R_{\mathcal{D}}(h, \varepsilon)$ indicates that the classifier h trained with the learning paradigm $\mathcal{A}$ is easier to attack (more attackable). $R^{emp}_{\mathcal{D}}(h, \varepsilon)$ is the empirical evaluation of the attackability level. By definition, if $\mathcal{A}$ is deterministic and the binary 0-1 loss is adopted, $\sum_{i=1}^{n} C^*_h(z_i)$ gives $R^{emp}_{\mathcal{D}}(h, \varepsilon)$.

Theorem.1 establishes the upper bound of the adversarial risk $R_{\mathcal{D}}(h, \varepsilon)$ based on the conditional mutual information $CMI_{\mathcal{D},\mathcal{A}}$ between the legal data and the learning paradigm. Without loss of generality, the hinge loss is adopted to compute the miss-classification risk of each z, i.e., $\ell(h, z = (x, y)) = \sum_{j=1}^{m} \max\{0, 1 - y_j h_j(x)\}$. We consider one of the most popularly used structures of multi-label classifiers, i.e., $h(x) = \mathbf{W}Rep(x)$, where $\mathbf{W} \in R^{m*d'}$ is the weight of a linear layer and $Rep(x) \in R^{d'}$ is a d'-dimensional representation vector of $x \in R^d$, e.g., from a non-linear network architecture. In Theorem.1, we assume a linear hypothesis h, i.e., $Rep(x) = x$ for the convenience of analysis. The conclusion holds for more advanced architectures, such as feed-forward neural networks.

Theorem 1. *Let $h = \mathbf{W}x$ be a linear multi-label classifier. We further denote $\mathcal{D} = (\mathcal{D}_1, \cdots, \mathcal{D}_m)$ and $\mathbf{W} = (\mathbf{w}_1, \cdots, \mathbf{w}_m)$, where $\mathcal{D}_j$ is the data distribution w.r.t. each label j and $\mathbf{w}_j$ is the weight vector of the classifier of label j.*

$$\begin{aligned} R_{\mathcal{D}}(h, \varepsilon) \leq R^{emp}_{\mathcal{D}}(h, \varepsilon) + & \\ & \left(\tfrac{2}{n} CMI_{\mathcal{D},A} \mathop{\mathbb{E}}_{z=(x,y) \sim \mathcal{D}}\left[\sup_{\mathbf{W} \in \mathcal{W}_A} (l(\mathbf{W}, z) + C_{\mathbf{W},z}\varepsilon)^2\right]\right)^{1/2}, \end{aligned} \tag{3}$$

where $\mathcal{W}_{\mathcal{A}}$ is the set including all possible weight vectors learned by $\mathcal{A}$ using the data set z^n sampled from $\mathcal{D}^n$. $C_{\mathbf{W},z} = \max_{\{b_1, \cdots, b_m\}} \left\|\sum_{j=1}^{m} b_j y_j \mathbf{w}_j\right\|_2, b_j = \{0, 1\}$. The empirical adversarial risk $R_{Z^n}(A, \varepsilon)$ has the upper bound:

$$R^{emp}_{\mathcal{D}}(h, \varepsilon) \leq R^{emp}_{\mathcal{D}}(h, 0) + \mathop{\mathbb{E}}_{z^n \sim \mathcal{D}^n, \mathcal{A}}\left[\sup_{\mathbf{W} \in \mathcal{W}_{\mathcal{A}}} \mathop{\mathbb{E}}_{z \in z^n} (C_{\mathbf{W},z}\varepsilon)\right], \tag{4}$$

where $R^{emp}_{\mathcal{D}}(h, 0)$ denotes the empirical and adversarial-free classification risk.

We further provide the upper bound of $CMI_{\mathcal{D},\mathcal{A}}$ as:

$$CMI_{\mathcal{D},\mathcal{A}} \leq ent(\mathbf{w}_1, \cdots, \mathbf{w}_m) + ent(\mathcal{D}_1, \cdots, \mathcal{D}_m) \tag{5}$$

where $ent(\cdot)$ denotes the entropy of the concerned random variables.

Key Factors of Attackability. The three key factors determining the adversarial risk (thus the attackability level) of the targeted multi-label classifier are: 1) $CMI_{\mathcal{D},\mathcal{A}}$; 2) $\underset{z}{\mathbb{E}} C_{\mathbf{W},z}$ ($\underset{z \leftarrow \mathcal{D}}{\mathbb{E}} C_{\mathbf{W},z}$ in Eq.(3) and $\underset{z \in z^n}{\mathbb{E}} C_{\mathbf{W},z}$ in Eq.(4)); and 3) the attack budget ε.

The last factor of the attack budget ε is easy to understand. The targeted classifier is intuitively attackable if the adversary has more attack budget. The larger ε is, the stronger the attack becomes and the adversarial risk rises accordingly. We then analyze the first factor $CMI_{\mathcal{D},\mathcal{A}}$. For a multi-label classifier h accurately capturing the label correlation in the training data, the output from h_j and h_k are closely aligned w.r.t. the positively or negatively correlated labels j and k. Specifically, in the linear case, the alignment between h_j and h_k can be presented by $s(h_j, h_k)$=max $\{cos \langle \mathbf{w}_j, \mathbf{w}_k \rangle, cos \langle -\mathbf{w}_j, \mathbf{w}_k \rangle\}$, where $cos \langle *, * \rangle$ denotes the cosine similarity. As shown in Eq.(5), the alignment of the decision hyper-planes of the correlated labels reduce the uncertainty of $\mathbf{W} = \mathcal{A}(\mathcal{D})$. Correspondingly, the conditional mutual information $CMI_{\mathcal{D},\mathcal{A}}$ decreases if the label correlation is strong and the classifier perfectly encodes the correlation into the alignment of the label-wise decision hyper-planes. According to Eq.(3), it is consistent with the well recognized fact of adversary-free multi-label learning: encoding the label correlation in the classifier helps to achieve an accurate adversary-free multi-label classification.

Lemma 1. *$\underset{z}{\mathbb{E}} C_{\mathbf{W},z}$ reaches the maximum value, if for each pair of labels j and k, $\underset{z}{\mathbb{E}}\{cos \langle y_j \mathbf{w}_j, y_k \mathbf{w}_k \rangle\} = 1$.*

The second factor $\underset{z}{\mathbb{E}} C_{\mathbf{W},z}$ *measures the transferability of the attack noise and demonstrates the impact of the transferability level on the attackability of the classifier.* With Lemma.1, we make the following analysis. **First**, for two labels j and k with strong positive or negative correlation in the training data, a large value of $\underset{z}{\mathbb{E}}\{cos \langle y_j \mathbf{w}_j, y_k \mathbf{w}_k \rangle\}$ indicates a high intensity of $s(h_j, h_k) = \max \{cos \langle \mathbf{w}_j, \mathbf{w}_k \rangle, cos \langle -\mathbf{w}_j, \mathbf{w}_k \rangle\}$. It represents that the decision hyper-planes $\mathbf{w}_j$ and $\mathbf{w}_k$ of the classifier h are consistently aligned. Therefore, with the same attack strength encoded by $\|r\|_2 \leq \varepsilon$, the adversarial sample $x' = x + r$ tends to cause miss-classification on both $h_j(x')$ and $h_k(x')$. Therefore, the attack perturbation's impact is easy to transfer between the correlated labels. Otherwise, $\underset{z}{\mathbb{E}}\{cos \langle y_j \mathbf{w}_j, y_k \mathbf{w}_k \rangle\} = 0$ indicates an orthogonal pair of $\mathbf{w}_j$ and $\mathbf{w}_k$. The adversarial perturbation r may cause miss-classification on one of the labels, but induce little bias to the decision output of the other. The attack can not be transferred between the labels. Therefore, a higher / lower $\underset{z}{\mathbb{E}} C_{\mathbf{W},z}$ denotes higher / lower transferability of the attack perturbation. **Second**, according to Eq.(3) and Eq.(4), with an increasingly higher $\underset{z}{\mathbb{E}}\{cos \langle y_j \mathbf{w}_j, y_k \mathbf{w}_k \rangle\}$, the adversarial risk of the targeted classifier h rises given a fixed attack budget ε. In summary, the alignment between the classifier's decision hyper-planes of different labels captures the label correlation. The alignment facilitates the attack to transfer across the labels. A multi-label classifier is more attackable if the attack is more transferable across the labels, as the attack can impact the decision of more labels at the same time.

Remark 1. **Trade-off between the generalization capability of the classifier on clean data and its adversarial robustness.**

Capturing the label correlation in the learnt multi-label classifier can be a double-edged sword. **On one hand**, encouraging alignment between the decision hyper-planes of the correlated labels reduces $CMI_{\mathcal{D},\mathcal{A}}$ under the adversary-free scenario ($\varepsilon = 0$ in Eq.(3)), thus reduces the expected miss-classification risk. **On the other hand**, the alignment between the decision hyper-planes increases the transferability of the attack, which makes the classifier more vulnerable. *Controlling the alignment between the decision outputs of different labels can tune the trade-off between the utility and the adversarial robustness of the classifier.*

4 Transferrability Regularization for Adversarially Robust Multi-label Classification

Following the above discussion, an intuitive solution to achieve adversarially robust multi-label classification is to regularize $\underset{z \in z^n}{\mathbb{E}} C_{\mathbf{W},z}$ empirically, while minimizing the multi-label classification loss over the training data set z^n. We denote this training paradigm as **ARM-Primal**:

$$h^* = \arg\min_{h} \frac{1}{n}\ell(h, z_i) + \frac{\lambda}{n}\sum_{i=1}^{n} C_{\mathbf{W},z_i} \tag{6}$$

where λ is the penalty parameter, and $C_{\mathbf{W},z_i}$ is given as in Theorem.1. As discussed in Section.3, the magnitude of $C_{\mathbf{W},z_i}$ in Eq.(6) reflects the alignment between the classifier's parameters $\{\mathbf{w}_1, ..., \mathbf{w}_m\}$. Penalizing large $C_{\mathbf{W},z_i}$ thus reduces the transferability of the input attack manipulation among different labels, which makes the learnt classifier h more robust against the adversarial perturbation. However, *ARM-Primal* only considers the alignment between the parameters of the linear layer $\mathbf{w}_j$ (j=1, ..., m). This setting limits the flexibility of the regularization scheme from two perspectives. First, whether h is attackable given a bounded attack budget also depends on the magnitude of the classification margin of the input instance [22,3]. Second, the regularization is only enforced over the linear layer's parameters of h. However, it is possible that the other layers could be relevant with the transferability of the attack noise. Adjusting the parameters of these layers can also help to control the attackability.

As an echo, we address accordingly the limits of *ARM-Primal*: **First**, a soft attackability estimator (**SAE**) for the targeted multi-label classifier h is proposed to relax the NP-hard attackability assessment in Eq.(1). We show that the proposed *SAE* assesses quantitatively the transferabiltiy level of the input attack noise by considering both the alignment of the decision boundaries and the classification margin of the input data instance. The attackability of the classifier is unveiled to be proportional to the transferability of the attack. **Second**, *SAE* is then introduced as a regularization term to achieve a tunable trade-off between transferability control and classification accuracy of the targeted classifier h. It thus reaches a customized balance between adversarial attackability and utility of h for multi-label learning practices.

4.1 Soft Attackability Estimator (SAE)

We first introduce the concept of *SAE* with the single-label classification setting and then extend it to the multi-label case. Suppose h is a binary classifier and instance x is predicted as positive if $h(x) > 0$ and vice versa. Let the adversarial perturbation be decomposed as $r = c\tilde{r}$, where $c = \|r\|_p$ and $\|\tilde{r}\|_p = 1$, i.e., $\tilde{r}$ shows the direction of the attack noise and c indicates the strength of the attack along this direction. For the perturbed input $x' = x + c\tilde{r}$, the first-order approximation of $h(x')$ is given as:

$$h(x + c\tilde{r}) = h(x) + c\tilde{r}^T \nabla h(x), \;\; s.t. \;\; \|\tilde{r}\|_p = 1, \;\; c \geq 0 \tag{7}$$

where $\nabla h(x)$ denotes the gradient of h to x. To deliver the attack successfully, the magnitude of the attack noise follows:

$$c \geq \frac{-h(x)}{\tilde{r}^T \nabla h(x)}. \tag{8}$$

The attackability of h on x along the direction of $\tilde{r}$ is proportional to $\frac{1}{c}$. The smaller c is, the more attackable the classifier h becomes.

Extending the notions to the multi-label setting, we define the multi-label classifier **h's attackability at x along the direction of $\tilde{r}$**:

$$A_{h(x),\tilde{r}} = \sum_{j=1}^{m} \max\{\frac{-\tilde{r}^T \nabla h_j(x)}{h_j(x)}, 0\}. \tag{9}$$

Note that in the multi-label setting, the adversarial perturbation $\tilde{r}$ may cause miss-classification of x for some labels, while enhancing the correct classification confidence for other labels, i.e., $\frac{-\tilde{r}^T \nabla h_j(x)}{h_j(x)}$ can be negative for the labels with enhanced correct classification confidences. We set the corresponding attackability level to 0, as the attack perturbation fails to cause miss-classification.

The intensity of $A_{h(x),\tilde{r}}$ is proportional to the number of the labels whose decision outputs are flipped by the perturbation $\tilde{r}$. Compared to the hard-count based attackability measurement C^*_h in Eq.(1), $A_{h(x),\tilde{r}}$ is a soft score quantifying the impact of the attack perturbation over the outputs of the classifier. It is therefore regarded as a *soft attackability estimator*.

Transferrability defines attackability. For simplicity, we denote $\frac{-\nabla h_j(x)}{h_j(x)}$ as $\mathbf{a}_j$, and $A_{h(x),\tilde{r}}$ can be further described as

$$\begin{aligned} A_{h(x),\tilde{r}} &= \tilde{r}^T \sum_{j \in S, S = \left\{j; \mathrm{sgn}(-y_j \tilde{r}^T \nabla h_j(x)) > 0\right\}} \mathbf{a}_j \\ &= \|\tilde{r}\|_2 \sqrt{\sum_{j \in S} \|\mathbf{a}_j\|_2^2 + 2 \sum_{j<k; j,k \in S} \|\mathbf{a}_j\|_2 \|\mathbf{a}_k\|_2 \cos\langle \mathbf{a}_j, \mathbf{a}_k \rangle} \cos\left\langle \tilde{r}, \sum_{j \in S} \mathbf{a}_j \right\rangle \end{aligned} \tag{10}$$

As shown in Eq.(10), the tranferability of the attack noise $\tilde{r}$ is measured by the cosine similarity between $\mathbf{a}_j$ and $\mathbf{a}_k$. Each $\mathbf{a}_j$ aligns with the principal eigenvector of the Fisher Information Matrix (FIM) of h_j at the input instance x [28]. It depicts the local geometrical profile of the decision boundaries of different labels near x. A larger cosine similarity between $\mathbf{a}_j$ and $\mathbf{a}_k$ indicates a stronger alignment of the decision boundaries of label j and k within the neighborhood of x. The attack noise $\tilde{r}$ thus causes closer magnitude of perturbation over $h_j(x)$ and $h_k(x)$ according to Eq.(9). It confirms the association between the transferability and the attackability, as unveiled by Eq.(3) and Eq.(4). Besides, the magnitude of the gradient $\mathbf{a_k} = \nabla h_k(x)$ also shapes the attackability level. A larger norm $\|\nabla h_k(x)\|_2$ indicates a less stable classification output within the L_p-ball centered at x, i.e., a higher attackability level of the classifier. Integrating both factors, $A_{h(x),\tilde{r}}$ is thus adopted as an empirical attackability estimator of h.

It is worth noting that **the proposed *SAE* reflects the transferability of the attack, regardless of the setting of attack budget**. As shown by Eq.(9), *SAE* is evaluated only with the gradient information of the classifier, which is independent of the attack capability of the adversary. In contrast, *GASE* in [24] depends on the prior knowledge about the attack budget of the adversary. In practical applications, the attack budget is usually case-dependent, which limits the use of *GASE* as a generic adversarial robustness evaluation tool. As an attack-strength-independent assessment, *SAE* can help to evaluate the attackability level of a classifier, before it is compromised by any specific attack. It is therefore can be used as a predicative guide for choosing adversarially robust multi-label learning architectures. In the linear case where $h(x) = \mathbf{W}x$, the cosine similarity $cos\left<\mathbf{a}_j, \mathbf{a}_k\right>$ produces a similar alignment metric as $s(h_j, h_k) = \max\{cos\left<\mathbf{w}_j, \mathbf{w}_k\right>, cos\left<-\mathbf{w}_j, \mathbf{w}_k\right>\}$. According to Eq.(3) and (10), the higher the cosine similarity score $cos\left<\mathbf{a}_j, \mathbf{a}_k\right>$ is, the higher $C_{\mathbf{W},z}$ in Eq.(3) and $A_{h(x),\tilde{r}}$ in Eq.(10) becomes. We thus measure the **attackability of h at x** as the maximum $A_{h(x),\tilde{r}}$ as:

$$\phi_{h,x} = \max_{\tilde{r}} A_{h(x),\tilde{r}}, \;\; s.t. \;\; \|\tilde{r}\|_p = 1 \tag{11}$$

We inherit the constraint $\|\tilde{r}\|_p = 1$ from Eq.(7). The resultant $\tilde{r}$ denotes the directions of the adversarial noise vector along which the attack can be maximally transferred. With this setting, we separate the derived transferability measurement with the attack strength. With the primal-dual conversion, we can obtain the solution to Eq.(11) as:

$$\begin{aligned} &\phi_{h,x} = \max_{\{b_1, b_2, \cdots, b_m\}} \left\| \sum_{j=1}^{m} \frac{-b_j \nabla h_j(x)}{h_j(x)} \right\|_q, \\ &s.t. \;\; \tfrac{1}{p} + \tfrac{1}{q} = 1, \;\; b_j = \{0, 1\}, \end{aligned} \tag{12}$$

where p denotes the L_p norm of the perturbations. Without loss of generality, we only discuss $p = 2$ of the l_p-norm in Eq.(12). As the objective function of Eq.(12) enjoys the submodularity property [2], we employ a simple yet effective greedy-based algorithm to solve Eq.(12). Algorithm 1 describes the greedy-search based solution to compute the *SAE* score.

Algorithm 1: The Greedy Solution to Soft Attackability Estimation

```
Input: $\left\{\frac{-\nabla h_1(x)}{h_1(x)}, \cdots, \frac{-\nabla h_m(x)}{h_m(x)}\right\}$.
Output: The set of selected labels $S$.
Initialize $S$ as an empty set. Set $LB = 0$ and $CB = 0$, where $LB$ denotes the
  best result of last iteration and $CB$ denotes the best result of current
  iteration.
while $|S| < m$ do
    $LB = CB$;
    $CB = \max_{\{1,\cdots,m\}-S} \left(\sum_{i \in S} \frac{-\nabla h_i(\mathbf{x})}{h_i(\mathbf{x})} + \frac{-\nabla h_j(\mathbf{x})}{h_j(\mathbf{x})}\right)$;
    if $CB < LB$, break;
    $S = S + j$
end
```

4.2 SAE Regularized Multi-label Learning

We propose to enhance the adversarial robustness of a multi-label classifier by enforcing the control over the *SAE* score of the classifier explicitly during training. While we suppose x is correctly classified during the theoretical analysis of attackability, it doesn't necessarily hold during training. For an originally miss-classified data instance x, it is possible that $A_{h(x),\tilde{r}}$ can be valued to 0. In this case, the attack perturbation r can augment the confidence of the miss-classification. However, with $A_{h(x),\tilde{r}} = 0$, bare penalization can be enforced to suppress the bias. It may encourage further negative impact in the learnt classifier. To mitigate this issue, we slightly modify the definition of $A_{h(x),\tilde{r}}$ and use it as the transferability regularization term of multi-label learning, which gives:

$$\hat{A}_{h(x),\tilde{r}} = \sum_{j=1}^{m} \max\{\frac{-\tilde{r}^T y_j \nabla h_j(x)}{\max(e^{y_j h_j(x)}, \alpha)}, 0\}, \ \ \alpha > 0 \tag{13}$$

where α is set to prevent over-weighing. For an originally correctly classified instance ($y_j h_j(x) > 0$), $\hat{A}_{h(x),\tilde{r}}$ penalizes the attack transferability as $A_{h(x),\tilde{r}}$. For a miss-classified instance ($y_j h_j(x) \leq 0$), minimizing $\hat{A}_{h(x),\tilde{r}}$ helps to reduce the confidence of the miss-classification output. Using the exponential function in $\hat{A}_{h(x),\tilde{r}}$, the miss-classified instance with stronger confidence (more biased decision output) is assigned with an exponentially stronger penalty. This setting strengthens the error-correction effect of $\hat{A}_{h(x),\tilde{r}}$.

Similarly as in Eq.(12), we can define $\hat{\phi}_{h,x} = \max_{\tilde{r}} \hat{A}_{h(x),\tilde{r}}$ in Eq.(14). The objective function of the SAE regularized multi-label learning (named hereafter as **ARM-SAE**) gives in Eq.(15):

$$\begin{aligned} \hat{\phi}_{h,x} &= \max_{\{b_1,b_2,\cdots,b_m\}} \left\| \sum_{j=1}^{m} \frac{-b_j y_j \nabla h_j(x)}{\max(e^{y_j h_j(x)}, \alpha)} \right\|_q, \\ s.t. \ & \tfrac{1}{p} + \tfrac{1}{q} = 1, \ b_j = \{0, 1\}, \end{aligned} \tag{14}$$

$$l = \frac{1}{n}\sum_{i=1}^{n}\ell(h, z_i) + \frac{\lambda}{n}\sum_{i=1}^{n}\hat{\phi}_{h,x_i}, \tag{15}$$

where λ is the regularization weight. $\hat{\phi}_{h,x}$ can be calculated using the greedy search solution as $\phi_{h,x}$. If the classifier h takes a linear form, we can find that *ARM-SAE* reweighs the linear layer parameters of the classifier $\{\mathbf{w}_1, \cdots, \mathbf{w}_m\}$ with the weight $\frac{1}{\max(e^{y_j h_j(x)}, \alpha)}$. Compared to *ARM-Primal* (see Eq.(12) to $C_{\mathbf{W},z}$ in Theorem 1), *ARM-SAE* enforces more transferability penalty over the instances with smaller classification margins. As unveiled in [24], these instances are easier to be perturbed for the attack purpose. Instead of penalizing each instance with the same weight as in *ARM-Primal*, *ARM-SAE* can thus perform a more flexible instance-adapted regularization.

5 Experiments

5.1 Experimental Setup

Datasets. We empirically evaluate our theoretical study on three data sets collected from real-world multi-label cyber-security applications (*Creepware*), object recognition (*VOC2012*) [4] and environment science (*Planet*) [12]. The descriptions of the datasets are given can be found in the supplementary file due to the space limit. The data sets are summarized in Table.1.

Performance Benchmark. Given a fixed attack strength of ε, we compute *the number of flipped labels* $C_h^*(z)$ *on each testing instance according to Eq.(1)* and take the average of the derived $\{C_h^*(z)\}$ (noted as C_a) as an overall estimation of attackability on the testing data set. Due to the NP-hard intrinsic of the combinatorial optimization problem in Eq.(1), we use *GASE* [24] to estimate empirically $C_h^*(z)$ and C_a. Besides, we measure the multi-label classification performance on the clean and adversarially modified testing instances with *Micro-F1* and *Macro-F1* scores.

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*. DNN model Inception-V3 [18] 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 measured on the clean testing data to evaluate the classification performance of multi-label classifiers [27]. Note that accurate adversary-free multi-label classification is beyond our scope and these classifiers are used to verify the theoretical analysis and the proposed *ARM-SAE* method.

Input Normalization and Reproduction. We normalize the adversarially perturbed data during the attack process. Due to the space limit, we provide the parameter settings and the reproduction details in the supplementary file.

Table 1: Summary of the used real-world data sets. 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 data sets are also reported.

Data set	N	m	l_{avg}	Micro F1	Macro F1	Classifier$_{target}$
Creepware	966	16	2.07	0.76	0.66	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

Table 2: Attackability estimation by *SAE*. $\lambda_{nuclear}$ denotes the strength of nuclear-norm based regularization. CC and P denote the Spearman coefficient and the p-value between *GASE* and *SAE* scores on the testing instances.

Data set		$\longrightarrow$ *robustness increase*					$CC, P(Spearman)$
Creepware	$\lambda_{nuclear}$	0	0.00001	0.0001	0.001	0.01	$CC = 1$ $P = 0$
	GASE ($C_a, \varepsilon = 0.5$)	13.5	11.4	10.8	6.9	4.3	
	SAE	31.5	19.16	18.06	14.55	11.22	
VOC2012	$\lambda_{nuclear}$	0	0.0001	0.001	0.01	0.1	$CC = 1$ $P = 0$
	GASE ($C_a, \varepsilon = 10$)	10.8	10.1	9.3	8.5	4.9	
	SAE	157.6	127.3	77.6	69.1	61.0	
Planet	$\lambda_{nuclear}$	0	0.0001	0.001	0.01	0.1	$CC = 1$ $P = 0$
	GASE ($C_a, \varepsilon = 2$)	13.1	12.2	11.6	10.5	7.1	
	SAE	267.1	221.5	186.3	158.2	102.0	
		$\longrightarrow$ *attackability decrease*					

5.2 Effectivity of Soft Attackability Estimator (SAE)

In Table.2, we demonstrate the validity of the proposed *SAE* by checking the consistency between the *SAE* and the *GASE*-based attackability measurement [24]. We adopt the **nuclear-norm** regularized training [24] to obtain an adversarially robust multi-label classifier. On the same training set, we increase the nuclear-norm regularization strength gradually to derive more robust architectures against the evasion attack. For each regularization strength, we can compute the *SAE* score of the classifier on the unperturbed testing instances. Similarly, by freezing the attack budget ε on each data set, we can generate the *GASE* score (C_a) corresponding to each regularization strength. Note that *only the ranking orders of the SAE and GASE score matters in the attackabiltiy measurement by definition.* We use the ranking relation of the scores to select adversarially robust models. Therefore, we adopt the Spearman rank correlation coefficient to measure the consistency between *SAE* and *GASE*.

We use the *GASE* score as a baseline of attackability assessment. The *SAE* and *GASE* score are strongly and positively correlated over all the datasets according to the correlation metric. Furthermore, with a stronger robustness regularization, the *SAE* score decreases accordingly. It confirms that the intensity of the proposed *SAE* score capture the attackability level of the targeted classifeir. This observation further validates empirically the motivation of using *SAE* in adjusting the adversarial robustness of the classifier.

The experimental study also shows the attack-strength-independent merit of the *SAE* over *GASE. SAE* is computed without knowing the setting of the attack budget. It thus reflects the intrinsic property of the classifier determining its adversarial vulnerability. In practice, this attack-strength-independent assessment can help to evaluate the attackability level of the deployed classifier, before it is compromised by any specific attack.

5.3 Effectiveness Evaluation of ARM-SAE

We compare the proposed *ARM-SAE* method to the state-of-the-art techniques in improving adversarial robustness of multi-label learning: the L_2-norm and the nuclear-norm regularized multi-label training [24]. Besides, we conduct an ablation study to verify the effectiveness of *ARM-SAE.*

- L_2 **Norm and Nuclear Norm Regularized Training.** Enforcing the L_2 and nuclear norm constraint helps to reduce the model complexity and thus enhance the model's adversarial robustness [19,24].
- **ARM-Single.** This variant of *ARM-SAE* is built by enforcing the transferability regularization with respect to individual labels separately:

$$\phi_{H\_\text{single}} = \sum_{i=1}^{n} \sum\nolimits_{k=1}^{m} \left\| \frac{\nabla h_k(x_i)}{\max(e^{y_k^i h_k(x)}, \alpha)} \right\|_2 . \tag{16}$$

  We compare *ARM-SAE* with *ARM-Single* to show the merit of jointly measuring and regularizing the impact of the input attack noise over all the labels. *ARM-SAE* tunes the transferability of the attack jointly, while *ARM-Single* enforces the penalization with respect to each label individually.
- **ARM-Primal.** We compare this variant to *ARM-SAE* to demonstrate the merit of *ARM-SAE* by 1) introducing the flexiblity of penalizing the whole model architecture, instead of only the linear layer; 2) taking the impact of classification margin on adversarial risk [22] into the consideration.

Two different attack budgets ε on each data set are introduced denoting varied attack strength. With each fixed ε, we compute the *Micro-F1* and *Macro-F1* scores of the targeted classifiers after retraining with the techniques above. Table.3 lists the classification accuracy over the adversarial testing instances using different robust training methods. In Table.3, we also show the multi-label classification accuracy (measured by two F1 scores) on the clean testing instances as a baseline. Consistently observed on the three datasets, even a small attack budget can deteriorate the classification accuracy drastically, which shows the vulnerability of multi-label classifiers. Generally, all the regularization method can improve the classification accuracy on the adversarial input. Among all the methods, *ARM-SAE* achieves the highest accuracy on the adversarial samples. It confirms the merit of *SAE* in controlling explicitly the transferability and then suppressing the attackabilty effectively. In addition, by regularizing jointly the attack transfer and exploiting classification margin for the attackability measurement, *ARM-SAE* achieves superior robustness over the two variants.

Table 3: Effectiveness evaluation of ARM-SAE. For convenience, *non*, L_2, *nl*, *sg*, *pm* and *SAE* are used to denote the absence of regularization, L_2 *norm, nuclear-norm, ARM-single, ARM-Primal* and *ARM-SAE* based methods respectively. The best results are in bold.

	Creepware : Micro F1 = 0.76, Macro F1 = 0.66 (on clean data)											
Budget	$\varepsilon = 0.05$						$\varepsilon = 0.2$					
Regularizors	*non*	L_2	*nl*	*sg*	*pm*	*SAE*	*non*	l_2	*nl*	*sg*	*pm*	*SAE*
Micro F1	0.34	0.40	0.45	0.44	0.43	**0.53**	0.10	0.13	0.15	0.15	0.16	**0.22**
Macro F1	0.33	0.39	**0.43**	0.39	**0.43**	0.42	0.12	0.15	0.20	0.17	0.20	**0.25**
	*VOC*2012 : Micro F1 = 0.83, Macro F1 = 0.74 (on clean data)											
Budget	$\varepsilon = 0.1$						$\varepsilon = 1$					
Regularizors	*non*	L_2	*nl*	*sg*	*pm*	*SAE*	*non*	l_2	*nl*	*sg*	*pm*	*SAE*
Micro F1	0.49	0.53	0.56	0.54	0.57	**0.61**	0.20	0.22	0.27	0.26	0.26	**0.30**
Macro F1	0.29	0.31	0.33	0.31	0.36	**0.38**	0.12	0.16	0.22	0.17	0.20	**0.23**
	Planet : Micro F1 = 0.82, Macro F1 = 0.36 (on clean data)											
Budget	$\varepsilon = 0.1$						$\varepsilon = 1$					
Regularizors	*non*	L_2	*nl*	*sg*	*pm*	*SAE*	*non*	l_2	*nl*	*sg*	*pm*	*SAE*
Micro F1	0.41	0.49	0.45	0.48	0.49	**0.53**	0.06	0.09	0.08	0.10.	0.09	**0.13**
Macro F1	0.13	0.22	0.17	0.20	0.18	**0.24**	0.03	0.04	0.04	0.06	0.06	**0.08**

Table 4: Trade-off Between Generalization Performance on Clean Data and Adversarial Robustness on *Creepware*. The attack budget $\varepsilon = 0.05$.

λ	0	10^{-7}	10^{-6}	10^{-5}	10^{-4}
ϕ_{align}	0.23	0.22	0.20	0.15	0.12
Micro F1(clean)	0.76	0.76	0.75	0.72	0.70
Macro F1(clean)	0.66	0.63	0.56	0.50	0.46
MIcro F1(pert)	0.34	0.35	0.39	0.44	0.53
Macro F1(pert)	0.33	0.33	0.35	0.40	0.42

5.4 Validation of Trade-off Between Generalization Performance on Clean Data and Adversarial Robustness

We validate the trade-off described in Remark.1. Without loss of generality, we conduct the case study on *Creepware*. Tuning the alignment between decision boundaries of different labels is achieved by conducting the *ARM-SAE* training as described in Eq.15. We freeze ε as 0.05 and vary the regularization weight λ in Eq.(15) from 10^{-7} to 10^{-4} to show increasingly stronger regularization effects enforced on the alignment between decision boundaries of different labels. For each regularization strength, we train a multi-label classifier h and evaluate quantitatively the averaged alignment level $\phi_{align} = \frac{1}{m^2} \sum\limits_{j,k\in\{1,\cdots,m\}} |\cos\langle \mathbf{w}_j, \mathbf{w}_k\rangle|$

between the decision hyperplanes of different labels. Table.4 shows the variation of ϕ_{align} and the Micro- / Macro-F1 accuracy of the trained multi-label classifier h over the clean and adversarially perturbed data instances (Micros / Macro F1 (clean / pert)). With increasingly stronger robustness regularization, the averaged alignment level ϕ_{align} between the label-wise decision hyper-planes decreases accordingly. Simultaneously, we witness the rise of the classification accuracy of h on the adversarially perturbed testing instances. It indicates the classifier h is more robust to the attack perturbation. However, the Macro- and Micro-F1 scores of h on the clean testing data drop with stronger alignment regularization. This observation is consistent with the discussion in Remark.1.

6 Conclusion

In this paper, we establish an information-theoretical adversarial risk bound of multi-label classification models. Our study identifies that the transferability of evasion attack across different labels determines the adversarial vulnerability of the classifier. Though capturing the label correlation improves the accuracy of adversary-free multi-label classification, our work unveils that it can also encourage transferable attack, which increases the adversarial risk. We show that the trade-off between the utility of the classifier and its adversarial robustness can be achieved by explicitly regularizing the transferability level of evasion attack in the learning process of multi-label classification models. Our empirical study demonstrates the applicability of the proposed transferability-regularized robust multi-label learning paradigm for both linear and non-linear classifies.

References

1. Chollet, Francois: Keras (2015), https://github.com/fchollet/keras
2. Elenberg, E.R., Khanna, R., Dimakis, A.G., Negahban, S.: Restricted strong convexity implies weak submodularity. Annuals of Statistics (2016)
3. Elsayed, G.F., Krishnan, D., Mobahi, H., Regan, K.: Large margin deep networks for classification. In: NeuIPS (2018)
4. Everingham, M., Gool, L.V., Williams, C., Winn, J., Zisserman, A.: The pascal visual object classes challenge 2012 (voc2012) results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html" (2012)
5. Fawzi, A., Moosavi-Dezfooli, S., Frossard, P.: Robustness of classifiers: From adversarial to random noise. In: NIPS. p. 1632–1640 (2016)
6. Fawzi, A., Fawzi, O., Frossard, P.: Analysis of classifiers' robustness to adversarial perturbations. Machine Learning **107**, 481–508 (2018)
7. Freed, D., Palmer, J., Minchala, D., Levy, K., Ristenpart, T., Dell, N.: "a stalker's paradise": How intimate partner abusers exploit technology. In: the 2018 CHI Conference. p. 1–13 (2018)
8. Gilmer, J., Metz, L., Faghri, F., Schoenholz, S., Raghu, M., Wattenberg, M., Goodfellow, I.: Adversarial spheres. CoRR (2018), http://arxiv.org/abs/1801.02774

9. Gupta, A., Lamba, H., Kumaraguru, P., Joshi, A.: Faking sandy: Characterizing and identifying fake images on twitter during hurricane sandy. In: WWW. p. 729–736 (2013)
10. Hein, M., Andriushchenko, M.: Formal guarantees on the robustness of a classifier against adversarial manipulation. In: NeuIPS. pp. 2266–2276 (2017)
11. Hoeffding, W.: Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association **58**(301), 13–30 (1963)
12. Kaggle: Planet: Understanding the amazon from space. `https://www.kaggle.com/c/planet-understanding-the-amazon-from-space/overview` (2017)
13. Khim, J., Loh, P.L.: Adversarial risk bounds for binary classification via function transformation. arXiv (2018)
14. Nicolae, M., Sinn, M., Minh, T.N., Rawat, A., Wistuba, M., Zantedeschi, V., Molloy, I.M., Edwards, B.: Adversarial robustness toolbox v0.2.2. CoRR (2018), `http://arxiv.org/abs/1807.01069`
15. Roundy, K.A., Mendelberg, P.B., Dell, N., McCoy, D., Nissani, D., Ristenpart, T., Tamersoy, A.: The many kinds of creepware used for interpersonal attacks. In: IEEE Symposium on Security and Privacy (SP). pp. 626–643 (may 2020)
16. Song, Q., Jin, H., Huang, X., Hu, X.: Multi-label adversarial perturbations. In: ICDM. pp. 1242–1247 (2018)
17. Steinke, T., Zakynthinou, L.: Reasoning about generalization via conditional mutual information. In: COLT (2020)
18. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: arXiv (2015)
19. Tu, Z., Zhang, J., Tao, D.: Theoretical analysis of adversarial learning: A minimax approach. In: NeuIPS. pp. 12259–12269 (2019)
20. Wang, Y., Jha, S., Chaudhuri, K.: Analyzing the robustness of nearest neighbors to adversarial examples. In: ICML. pp. 5133–5142 (2018)
21. Wang, Y., Han, Y., Bao, H., Shen, Y., Ma, F., Li, J., Zhang, X.: Attackability characterization of adversarial evasion attack on discrete data. In: KDD. pp. 1415–1425 (2020)
22. Yang, Y., Khanna, R., Yu, Y., Gholami, A., Keutzer, K., Gonzalez, J.E., Ramchandran, K., Mahoney, M.W.: Boundary thickness and robustness in learning models. In: NeuIPS (2020)
23. Yang, Y., Rashtchian, C., Zhang, H.: A closer look at accuracy vs. robustness. In: NeuIPS (2020)
24. Yang, Z., Han, Y., Zhang, X.: Characterizing the evasion attackability of multi-label classifiers. In: AAAI (2021)
25. Yin, D., Ramchandran, K., Bartlett, P.L.: Rademacher complexity for adversarially robust generalization. In: ICML. pp. 7085–7094 (2019)
26. Zhang, H., Yu, Y., Jiao, J., Xing, E.P., Ghaoui, L.E., Jordan, M.I.: Theoretically principled trade-off between robustness and accuracy. In: ICML (2019)
27. Zhang, M., Zhou, Z.: A review on multi-label learning algorithms. TKDE **26**(8), 1819–1837 (2013)
28. Zhao, C., Fletcher, P., Yu, M., Peng, Y., Zhang, G., Shen, C.: The adversarial attack and detection under the fisher information metric. In: AAAI. pp. 5869–5876 (2019)

7 Supplementary

7.1 Database Summary

We empirically evaluate our theoretical study on three data sets collected from real-world multi-label applications. They include cyber security practices (*Creepware*), object recognition (*VOC2012*) [4] and environment research (*Planet*) [12]. *Creepware* data include different stalkware app instances and each instance has 16 labels indicating different types of surveillance on the victim's mobile device. Besides, each app is profiled by the introductory texts of the app available in the app stores and signatures of its mobile service access. *VOC2012* is a well-known image data set and it is widely used in multi-label learning research. *Planet* data collects daily satellite imagery of the entire land surface of the earth. Each image is equipped with labels denoting different atmospheric conditions and various classes of land cover/land use.

7.2 Input Normalization and Parameter settings

When imposing attacks, we project the perturbed data in *VOC2012* and *Planet* to $[-1, 1]$, while we don't limit the value range of data in *Creepware*. The α in Eq.(15) is empirically set to 0.01 in all experiments. The regularization parameters λ in Eq.15 and other baselines are chosen empirically from the range $\left\{10^{-8}, 10^{-7}, \cdots, 10^{7}, 10^{8}\right\}$

Our codes were written in Python and all the models were built by Keras package [1]. The needed targeted evasion attack and adversarial training are implemented by adversarial-robustness-toolbox [14]. Our experiments were conducted on GPU rtx2080ti. Our codes for SAE training are available at `https://github.com/chelungungun/Transferability_MLATTACK`

7.3 Proofs

We supple the proof of the theorem in our paper, especially the Eq.(3) and Eq.(5), and the proof from Eq.(11) to Eq.(12)

Lemma 2. *(Thomas 2020 [17], Corollary 5) Let E, E' and Z be independent random variables where E and E' have identical distributions. Let A be a random function whose randomness is independent from E, E' and Z. Let g be a fixed function. Then*

$$\begin{aligned} &\underset{A,E,Z}{\mathbb{E}}\left[g(A(E,Z),E,Z)\right] \\ &\leq \inf_{t>0} \frac{I(A(E,Z);E|Z)+\underset{Z}{\mathbb{E}}\left[\log \underset{A,E,E'Z}{\mathbb{E}}\left[e^{t\cdot g(A(E,Z),E',Z)}\right]\right]}{t} \end{aligned} \tag{17}$$

Lemma 3. *(Hoeffding 1963 [11].) Let $\mathbf{X} \in [a,b]$ be a random variable with mean μ. Then for all $t \in \mathbb{R}$,*

$$\mathbb{E}(e^{t\mathbf{X}}) \leq e^{t\mu+t^2(b-a)^2/8} \tag{18}$$

Proof from Eq.11 to Eq.12: We can rewrite Eq.9 as

$$A_{h(x),\tilde{r}} = \tilde{r'} \sum_{k=1}^{m} \frac{-\nabla h_k(x) * \max\{\mathrm{sgn}(-\tilde{r'} y_k \nabla h_k(x)), 0\}}{h_k(x)}. \tag{19}$$

If there is no sgn function and max function in Eq.19, Eq.11 is actually the definition of dual norm. To eliminate the sgn and max function, we can break the domain of $\tilde{r}$ into a group of subsets according the output of those sgn functions. Denote the domain of $\tilde{r}$ as I and I_S is a subset of I which is defined by Eq.(20). S is an element from the power set of $\{1, \cdots, m\}$.

$$I_S = \left\{ \tilde{r} \,\middle|\, \begin{matrix} \tilde{r} y_k \nabla h_k < 0, k \in S \\ \tilde{r} y_k \nabla h_k \geq 0, k \notin T \end{matrix}, \tilde{r} \in \mathbb{R}^n \right\} \tag{20}$$

Based on Eq.(20), we redefine Eq.(9) and Eq.(11) over the sub-domain I_S of $\mathbf{r}$ as:

$$A_{h(x),\tilde{r}s} = \sum_{k \in S} \frac{-\tilde{r}' \nabla h_k(x)}{h_k(x)} \tag{21}$$

$$\begin{aligned} &\phi_s = \max_{\tilde{r} \in I_S} A_{h(x),\tilde{r}s}, \\ &s.t.\ \left\|\tilde{r}\right\|_p = 1 \end{aligned} \tag{22}$$

Now, we get $\phi_{h,x} = \max_{S \in P(S)} \phi_s$. It's easy to know that:

$$\begin{matrix} \phi_s = \max\limits_{\tilde{r} \in I_S} A_{h(x),\tilde{r}s}, \\ s.t.\ \left\|\tilde{r}\right\|_p = 1 \end{matrix} \leq \begin{matrix} \phi_s = \max\limits_{e \in \mathbb{R}^n} A_{h(x),es}, \\ s.t.\ \left\|e\right\|_p = 1 \end{matrix} = \left\| \sum_{k \in S} \frac{-\nabla h_k(x)}{h_k(x)} \right\|_q \tag{23}$$

The equality holds when the optimal e^* exactly locates in I_S. Now, if we want to prove that Eq.(11) = Eq.(12), we just need to prove that $\phi_{S^*} = \left\| \sum_{k \in S^*} \frac{-\nabla h_k(x)}{h_k(x)} \right\|_q$, that is we need to prove that the optimal e^* for S^* locates in I_{S^*}. We can prove that by contradiction. That is we assume $e^*_{S^*} \in I_{S'} (S' \neq S^*)$, then it is proved by Eq.(24).

$$\begin{aligned} \left\| \sum_{k \in S^*} \frac{-\nabla h_k(x)}{h_k(x)} \right\|_q &= \sum_{k \in S^*} \frac{-e^*_{S^*} \nabla h_k(x)}{h_k(x)} \\ &< \sum_{k \in S^* \cap S'} \frac{-e^*_{S^*} \nabla h_k(x)}{h_k(x)} \\ &\leq \left\| \sum_{k \in S^* \cap S'} \frac{-\nabla h_k(x)}{h_k(x)} \right\|_q \\ &< \left\| \sum_{k \in S^*} \frac{-\nabla h_k(x)}{h_k(x)} \right\|_q \end{aligned} \tag{24}$$

Proof of Eq.(3): We define the worst-case loss $l(h, z, \varepsilon)$ as:

$$\begin{aligned} &l(h, z, \varepsilon) = \max_{z' \in N(z)} l(h, z'), \\ &\text{where } N(z) = \left\{ (x', y') \,\middle|\, \left\|x' - x\right\|_p \leq \varepsilon,\ y' = y \right\}. \end{aligned} \tag{25}$$

We first upperly bound $l(h, z, \varepsilon)$ defined in Eq.(25) with the setting of linear classifier and hinge loss:

$$\begin{gathered} l(h,z,\varepsilon) \le l(h,z)+ \\ \max_{\|r\|_2\le\varepsilon} \left\| \sum_{k=1}^{m} y_k r'\cdot \mathbf{w}_k * \max\{\operatorname{sgn}(y_k r'\cdot\mathbf{w}_k),0\}\right\|_2 \\ \le l(h,z)+C_{\mathbf{W},z}\varepsilon. \end{gathered} \tag{26}$$

The last step borrows the proof from Eq.11 to Eq.12. Then we have

$$\begin{aligned}
& R_{\mathcal{D}}(A,\varepsilon) - R_{Z^n}(A,\varepsilon) \\
& = \mathop{\mathbb{E}}_{Z^n\leftarrow\mathcal{D}^n,A} l(A(Z^n),\mathcal{D},\varepsilon) - \mathop{\mathbb{E}}_{Z^n\leftarrow\mathcal{D}^n,A} l(A(Z^n),Z^n,\varepsilon) \\
& = \mathop{\mathbb{E}}_{\bar{Z},E,A} \left[l(A(\bar{Z}_E),\bar{Z}_{\bar{E}},\varepsilon) - l(A(\bar{Z}_E),\bar{Z}_E,\varepsilon)\right] , \quad (\bar{Z}\leftarrow\mathcal{D}^{n\times 2}) \\
& = \mathop{\mathbb{E}}_{\bar{Z},E,A} \left[f_{\bar{Z}}(A(\bar{Z}_E),E,\varepsilon)\right] \\
& \quad \mathit{by\ LEMMA}\,2 \\
& \le \inf_{t>0} \frac{I(A(\bar{Z}_E);E|\bar{Z}) + \mathop{\mathbb{E}}_{\bar{Z}}\left[\log \mathop{\mathbb{E}}_{\mathbf{W},E'}\left[e^{tf_{\bar{Z}}(\mathbf{W},E',\varepsilon)}\right]\right]}{t}, \\
& \quad \mathit{by\ \ independence} \\
& = \inf_{t>0} \frac{CMI_{\mathcal{D},A} + \mathop{\mathbb{E}}_{\bar{Z}}\left[\log \mathop{\mathbb{E}}_{\mathbf{W}}\left[\prod_{i=1}^{n} \mathop{\mathbb{E}}_{E'_i}\left[e^{\frac{t}{n}(l(W,(\bar{Z}_{E'})_i,\varepsilon) - l(\mathbf{W},(\bar{Z}_{\bar{E}'})_i,\varepsilon))}\right]\right]\right]}{t}, \\
& = \inf_{t>0} \frac{CMI_{\mathcal{D},A} + \mathop{\mathbb{E}}_{\bar{Z}}\left[\log \mathop{\mathbb{E}}_{\mathbf{W}}\left[\prod_{i=1}^{n} \mathop{E}_{E'_i}\left[e^{\frac{t}{n}(1-2E'_i)(l(\mathbf{W},\bar{Z}_{i,1},\varepsilon) - l(\mathbf{W},\bar{Z}_{i,2},\varepsilon))}\right]\right]\right]}{t} \\
& \quad \mathit{by\ LEMMA}\ 3 \\
& \le \inf_{t>0} \frac{CMI_{\mathcal{D},A} + \mathop{\mathbb{E}}_{\bar{Z}}\left[\log \mathop{\mathbb{E}}_{\mathbf{W}}\left[\prod_{i=1}^{n} e^{\frac{t^2}{2n^2}(l(\mathbf{W},\bar{Z}_{i,1},\varepsilon) - l(\mathbf{W},\bar{Z}_{i,2},\varepsilon))^2}\right]\right]}{t}, \\
& \le \inf_{t>0} \frac{CMI_{\mathcal{D},A}}{t} + \frac{t}{2n} \mathop{\mathbb{E}}_{\bar{Z}} \left[\sup_{\mathbf{W}\in\mathcal{W}_A} \frac{1}{n} \sum_{i=1}^{n} \left(l(\mathbf{W},\bar{Z}_{i,1},\varepsilon) - l(\mathbf{W},\bar{Z}_{i,2},\varepsilon)\right)^2\right] \\
& \le \inf_{t>0} \frac{CMI_{\mathcal{D},A}}{t} + \frac{t}{2n} \mathop{\mathbb{E}}_{Z\leftarrow\mathcal{D}} \left[\sup_{\mathbf{W}\in\mathcal{W}_A} l(\mathbf{W},Z,\varepsilon)^2\right] \\
& \le \inf_{t>0} \frac{CMI_{\mathcal{D},A}}{t} + \frac{t}{2n} \mathop{\mathbb{E}}_{Z\leftarrow\mathcal{D}} \left[\sup_{\mathbf{W}\in\mathcal{W}_A} \left(l(\mathbf{W},Z) + C_{\mathbf{W},Z}\cdot\varepsilon\right)^2\right] \\
& = \sqrt{\frac{2}{n} CMI_{\mathcal{D},A} \cdot \mathop{\mathbb{E}}_{Z\leftarrow\mathcal{D}} \left[\sup_{\mathbf{W}\in\mathcal{W}_A} \left(l(\mathbf{W},Z) + C_{\mathbf{W},Z}\cdot\varepsilon\right)^2\right]}
\end{aligned} \tag{27}$$

Proof of Eq.(5): Here we use H to denote the entropy.

$$\begin{aligned}
& CMI_{\mathcal{D},A} \\
& = I(A;S,\bar{Z}) - I(A;\bar{Z}) \\
& = H(A) + H(S,\bar{Z}) - H(A,S,\bar{Z}) - H(A) - H(\bar{Z}) + H(A,\bar{Z}) \\
& = H(A,\bar{Z}) + H(S|\bar{Z}) - H(S) - H(A,\bar{Z}|S) \quad : S\ is\, independent\, to\, Z \\
& = H(A,\bar{Z}) - H(A,\bar{Z}|S) \\
& \leq H(A,\bar{Z}) \\
& \leq H(A) + H(\bar{Z}) \\
& = H(\mathbf{W}) + H(\bar{Z}) \\
& = ent(\mathbf{w}_1,\cdots,\mathbf{w}_m) + ent(\mathcal{D}_1,\cdots,\mathcal{D}_m)
\end{aligned} \tag{28}$$