Supplementary Document
for
A Provable Case-Sensitive Robustness Test Oracle at Runtime for Patch Robustness Certification

A Formal Proof

Lemma (Certification of Benign Samples). *Given an arbitrary sample $\hat{x}$, if $[\forall M_1, M_2, M_3 \in \mathbb{M}, f(\hat{x} \odot M_1 \odot M_2 \odot M_3) = f(\hat{x})]$ holds, $\forall \hat{x}' \in \mathbb{A}_{\mathbb{P}}(\hat{x}), g(\hat{x}') = g(\hat{x})$.*

Proof. We first analyze $g(\hat{x})$. Since $[\forall M_1, M_2, M_3 \in \mathbb{M}, f(\hat{x} \odot M_1 \odot M_2 \odot M_3) = f(\hat{x})] \implies [\forall M_1, M_2 \in \mathbb{M}, f(\hat{x} \odot M_1 \odot M_2) = f(\hat{x})]$, we know its prediction label $g(\hat{x}) = f(\hat{x})$ output in Case ①. We then analyze $g(\hat{x}')$. For $\forall \hat{x}' \in \{\hat{x}' \mid \hat{x}' = (\mathrm{J} - P) \odot \hat{x} + P \odot \hat{x}' \wedge P \in \mathbb{P}\}$ (i.e., $\forall \hat{x}' \in \mathbb{A}_{\mathbb{P}}(\hat{x})$), w.l.o.g., we let $M_1 \odot P = \mathrm{O}$. Therefore, we get $\hat{x}' \odot M_1 = ((\mathrm{J} - P) \odot \hat{x} + P \odot \hat{x}') \odot M_1 = \hat{x} \odot M_1$. Then, we get $[\exists M_1 \in \mathbb{M}, \forall M_2, M_3 \in \mathbb{M}, f(\hat{x}' \odot M_1 \odot M_2 \odot M_3) = f(\hat{x})]$ (see Fig. 1). Note that the special cases $M_2 = M_3$, $M_1 = M_2 = M_3$ are included. Case 1: Suppose the returned label of $\hat{x}'$ output in Case ① as $f(\hat{x}')$ (i.e., $\forall M_1, M_2 \in \mathbb{M}, f(\hat{x}' \odot M_1 \odot M_2) = f(\hat{x}')$), then we know $g(\hat{x}') = f(\hat{x}') = f(\hat{x}) = g(\hat{x})$. Case 2: Otherwise, its returned label should be output in Case ② as $f(\hat{x}' \odot M_1)$, since $[\exists M_1 \in \mathbb{M}, \forall M_2, M_3 \in \mathbb{M}, f(\hat{x}' \odot M_1 \odot M_2 \odot M_3) = f(\hat{x}' \odot M_1) = f(\hat{x})]$, and further $g(\hat{x}') = f(\hat{x}' \odot M_1) = f(\hat{x}) = g(\hat{x})$. □

Lemma (Certification of one-patch Samples). *Given an arbitrary sample $\hat{x}$, if $[\exists M_1 \in \mathbb{M}, \forall M_2, M_3, M_4 \in \mathbb{M}, f(\hat{x} \odot M_1 \odot M_2 \odot M_3 \odot M_4) = f(\hat{x} \odot M_1)]$ holds, $\forall \hat{x}' \in \mathbb{A}_{\mathbb{P}}(\hat{x}), g(\hat{x}') = g(\hat{x})$.*

Proof. We first analyze $g(\hat{x})$. Case 1: If $\hat{x}$ meet the condition of Case ①, we can further get $\forall M_1, M_2, M_3, M_4 \in \mathbb{M}, f(\hat{x} \odot M_1 \odot M_2 \odot M_3 \odot M_4) = f(\hat{x})$. Note that the special case $M_1 = M_2 = M_3 = M_4$ is included, which means we can get $g(\hat{x}) = f(\hat{x} \odot M_1)$. Case 2: Otherwise, by the given condition $[\exists M_1 \in \mathbb{M}, \forall M_2, M_3, M_4 \in \mathbb{M}, f(\hat{x} \odot M_1 \odot M_2 \odot M_3 \odot M_4) = f(\hat{x} \odot M_1)] \implies [\exists M_1 \in \mathbb{M}, \forall M_2, M_3 \in \mathbb{M}, f(\hat{x} \odot M_1 \odot M_2 \odot M_3) = f(\hat{x} \odot M_1)]$, $\hat{x}$ meet the condition in Case ② and its prediction label $g(\hat{x}) = f(\hat{x} \odot M_1)$, same as Case ①. We then analyze $g(\hat{x}')$. For $\hat{x}' \in \{\hat{x}' \mid \hat{x}' = (\mathrm{J} - P) \odot \hat{x} + P \odot \hat{x}' \wedge P \in \mathbb{P}\}$ (i.e., $\forall \hat{x}' \in \mathbb{A}_{\mathbb{P}}(\hat{x})$), Case 1: Suppose $M_1 \odot P = \mathrm{O}$. Then we can get $\hat{x}' \odot M_1 = ((\mathrm{J} - P) \odot \hat{x} + P \odot \hat{x}') \odot M_1 = \hat{x} \odot M_1$, and further get $[\forall M_2, M_3, M_4 \in \mathbb{M}, f(\hat{x}' \odot M_1 \odot M_2 \odot M_3 \odot M_4) = f(\hat{x} \odot M_1)]$, which is the same condition as that on $\hat{x}$. Therefore, repeating those analysis for $g(\hat{x})$ above can get $g(\hat{x}) = g(\hat{x}')$. Case 2: Otherwise, for M_2, M_3, M_4, w.l.o.g., we let $M_2 \odot P = \mathrm{O}$ ($M_1 \neq M_2$). Then similarly, we get $[\exists M_1, M_2 (\neq M_1) \in \mathbb{M}, \forall M_3, M_4 \in \mathbb{M}, f(\hat{x}' \odot M_1 \odot M_2 \odot M_3 \odot M_4) = f(\hat{x} \odot M_1)]$. Note that the special cases $M_1 = M_3, M_2 = M_4, M_3 = M_4$ are included. Case 2.1: Suppose the prediction label $g(\hat{x}')$ output in Case ①. Then, since $[\exists M_1, M_2 (\neq M_1) \in \mathbb{M}, f(\hat{x}' \odot M_1 \odot M_2) = f(\hat{x} \odot M_1)]$ (special case $M_1 = M_3$, $M_2 = M_4$), by the condition of Route ①, we know $g(\hat{x}') = f(\hat{x}') = f(\hat{x} \odot M_1) = g(\hat{x})$. Case 2.2: Suppose the prediction label $g(\hat{x}')$ output in Route ②. Then, since $[\exists M_1, M_2 (\neq M_1) \in \mathbb{M}, \forall M_3 \in \mathbb{M}, f(\hat{x}' \odot M_1 \odot M_2 \odot M_3) = f(\hat{x} \odot M_1)]$ (special case $M_3 = M_4$), by the condition of Route ②, we know $g(\hat{x}') = f(\hat{x}' \odot M_1) = f(\hat{x} \odot M_1) = g(\hat{x})$. Case 2.3: Otherwise, the prediction label of $\hat{x}'$ should output in Route ③ since $[\exists M_1, M_2 (\neq M_1) \in \mathbb{M}, \forall M_3, M_4 \in \mathbb{M}, f(\hat{x}' \odot M_1 \odot M_2 \odot M_3 \odot M_4) = f(\hat{x} \odot M_1) = f(\hat{x}' \odot M_1 \odot M_2)]$ (special case $M_1 = M_3, M_2 = M_4$), and further $g(\hat{x}') = f(\hat{x}' \odot M_1 \odot M_2) = f(\hat{x} \odot M_1) = g(\hat{x})$. □

Theorem (Certification of Samples). *Given an arbitrary sample $\hat{x}$, if $c(\hat{x}) = True$ holds, $\forall \hat{x}' \in \mathbb{A}_{\mathbb{P}}(\hat{x}), g(\hat{x}') = g(\hat{x})$.*

Simply conjoining the antecedent of Lemma 1 and Lemma 2 can prove this theorem.

Theorem (Round-Trip Certification of Samples). *Given a benign sample x, if $[\forall M_1, M_2, M_3, M_4 \in \mathbb{M}, f(x \odot M_1 \odot M_2 \odot M_3 \odot M_4) = f(x)]$ (i.e., $c_r^2(x) = True$), $[\forall x' \in \mathbb{A}_{\mathbb{P}}(x), g(x') = g(x) \wedge c_r(x') = True]$.*

Proof. By the condition $[\forall M_1, M_2, M_3, M_4 \in \mathbb{M}, f(x \odot M_1 \odot M_2 \odot M_3 \odot M_4) = f(x)]$, we know the returned label $g(x) = f(x)$ in Case ①. Still by this condition, we know $[\forall x' \in \mathbb{A}_{\mathbb{P}}(x), \exists M_1 \in \mathbb{M}_{\mathbb{P}}, \forall M_2, M_3, M_4 \in \mathbb{M}, f(x \odot M_1 \odot M_2 \odot M_3 \odot M_4)]$, since $[\exists M_1 \in \mathbb{M}_{\mathbb{P}}, \forall x' \in \mathbb{A}_{\mathbb{P}}(x), x' \odot M_1 = x \odot M_1]$ (see Fig. 1 for illustration). Then by Lemma 2, we know $[\forall x' \in \mathbb{A}_{\mathbb{P}}(x), c_r(x') = True]$. By Lemma 1, we also know $[\forall x' \in \mathbb{A}_{\mathbb{P}}(x), g(x') = g(x)]$. Finally, we know $[\forall x' \in \mathbb{A}_{\mathbb{P}}(x), g(x') = g(x) \wedge c_r(x') = True]$. □

B Extension of MRCert to Recover and Certify N-patch samples

We can extend the maximum number of patches from 2 to N (called MRCert-N-patch, a variant of MRCert) following the following idea. First, we apply each set of N masks in the covering mask set $\mathbb{M}$ on the input sample $\hat{x}$ to test whether $\hat{x}$ is harmful. If it is not harmful, MRCert-N-patch returns the label $f(\hat{x})$ (marked as N-Case ①). If $\hat{x}$ is detected as harmful, we then test whether all its first-order mutants are harmful by applying each possible subset with N masks selected with replacement from $\mathbb{M}$ on each first-order mutant of $\hat{x}$. If there exists a first-order mutant, whose all $(N+1)$th-order mutants generated from the first-order mutant of $\hat{x}$ are predicted with the same label as this first-order sample, then $\hat{x}$ is deemed as a one-patch harmful sample, this first-order mutant is "clean" and the prediction label of this first-order mutant is returned (marked as N-Case ②). If that is not the case, we then test whether all second-order mutants of $\hat{x}$ are harmful in the same manner, and repeat until the Nth-order mutants of $\hat{x}$ are tested. For the certification function c_r with the input sample $\hat{x}$, it should be extended to the condition that all $(N+1)$th-order mutants of $\hat{x}$ are predicted with the same label as $\hat{x}$ (for those input samples whose label returned in Case ①), and the condition that there exists a first-order mutant of $\hat{x}$, whose all $(N+2)$th-order mutants are predicted with the same label as this first-order mutant of $\hat{x}$ (for those input samples whose label returned in Case ②), and certifying the input samples output in other cases by the condition in the same manner. For the round-trip certification function, it should be the condition that all $2N$th-order mutants of a benign sample x are predicted with the same label as x. We leave the formal proof and implementation as future work.