Two Coupled Rejection Metrics Can Tell Adversarial Examples Apart

Tianyu Pang[1], Huishuai Zhang[2], Di He[2], Yinpeng Dong[1], Hang Su[1], Wei Chen[2], Jun Zhu[1], Tie-Yan Liu[2]

[1] Tsinghua University [2] Microsoft Research Asia

1. Confidence and true confidence (T-Con)

In training phase:

$$\mathcal{L}_{\text{CE}} = -\log f_\theta(x)[y],$$

where y is the true label, and we name $f_\theta(x)[y]$ as **true confidence (T-Con)**. Lower cross-entropy (CE) values or higher T-Con values indicate higher prediction certainty on x.

In inference/test phase:

y is unknown, so we cannot compute T-Con. We usually approximate T-Con by **confidence** $f_\theta(x)[y^m]$, where $y^m = \underset{l}{\text{argmax}}\, f_\theta(x)[l]$ is the predicted label.

$f_\theta(x)[y^m]$ will overestimate $f_\theta(x)[y]$ when $y^m \neq y$, i.e., **confidence is over-confident on misclassified inputs**

1.1 Confidence and T-Con are coupled

Lemma 1. *(Separability) Given the classifier f_θ, $\forall x_1, x_2$ with confidences larger than $\frac{1}{2}$, i.e.,*

$$f_\theta(x_1)[y_1^m] > \frac{1}{2}, \text{ and } f_\theta(x_2)[y_2^m] > \frac{1}{2}. \tag{2}$$

If x_1 is correctly classified as $y_1^m = y_1$, while x_2 is wrongly classified as $y_2^m \neq y_2$, then T-Con$(x_1) > \frac{1}{2} >$ T-Con(x_2).

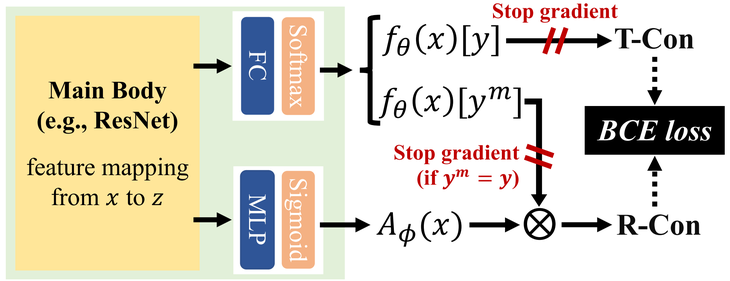
2. Learn T-Con by rectifying confidence

We train a rectified confidence (R-Con) to align with T-Con

$$\textbf{R-Con}(x) = f_\theta(x)[y^m] \cdot A_\phi(x)$$

$$\textbf{T-Con}(x) = f_\theta(x)[y]$$

where $A_\phi(x)$ is an auxiliary function.



2.1 How well is R-Con learned?

Definition 1. *(Point-wisely ξ-error) If at least one of the bounds holds at a point x:*

$$\text{Bound (i):} \left|\log\left(\frac{A_\phi(x)}{A_\phi^*(x)}\right)\right| \leq \log\left(\frac{2}{2-\xi}\right); \qquad \text{Bound (ii):} \left|A_\phi(x) - A_\phi^*(x)\right| \leq \frac{\xi}{2}. \tag{6}$$

where $\xi \in [0, 1)$, then $A_\phi(x)$ is called ξ-error at input x.

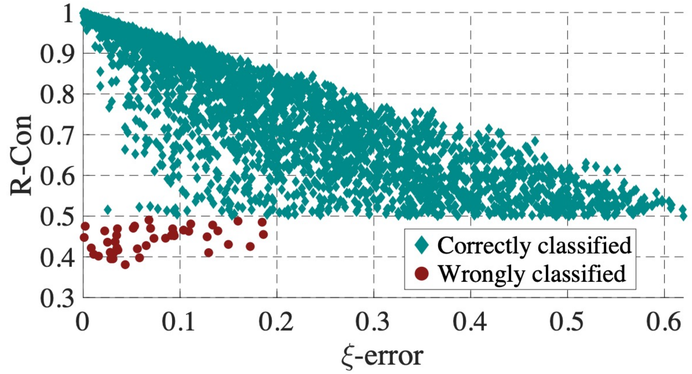
3. Confidence and R-Con are coupled

$\frac{1}{2-\xi}$ confidence rejector and **ξ-error R-Con rejector** can be coupled to ***perfectly*** distinguish correctly and wrongly classified samples.

Theorem 1. *(Separability) Given the classifier f_θ, for any input pair of x_1, x_2 with confidences larger than $\frac{1}{2-\xi}$, i.e.,*

$$f_\theta(x_1)[y_1^m] > \frac{1}{2-\xi}, \text{ and } f_\theta(x_2)[y_2^m] > \frac{1}{2-\xi}, \tag{7}$$

where $\xi \in [0, 1)$. If x_1 is correctly classified as $y_1^m = y_1$, while x_2 is wrongly classified as $y_2^m \neq y_2$, and A_ϕ is ξ-error at x_1, x_2, then there must be R-Con$(x_1) > \frac{1}{2} >$ R-Con(x_2).



Primary Contact: **Tianyu Pang** (tianyupang3@gmail.com); Code: https://github.com/P2333/Rectified-Rejection