Verifiably Robust Conformal Prediction (NeurIPS 2024)

Sehyun Park

February 25, 2025

Seoul National University - IDEA Lab.

Outline

1. Introduction
2. Preliminaries
3. Verifiably Robust Conformal Prediction (VRCP)
   1) VRCP via Robust Inference
   2) VRCP via Robust Calibration
4. Experiments

Introduction

- Conformal Prediction provides prediction sets that guarantee a user-specified probability under the assumption that training and test data are exchangeable.
- Nevertheless, this guarantee is violated when the data is subjected to adversarial attacks.
- This paper proposes **VRCP (Verifiably Robust Conformal Prediction)**, a new framework that leverages recent *neural network verification* methods to recover coverage guarantees under adversarial attacks.

Preliminaries

Notation

- $\boldsymbol{X} \subset \mathbb{R}^d$ and Y : a feature space and label space.
- $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)\}$: i.i.d. sampled dataset.
- $\mathcal{D}_{train}, \mathcal{D}_{cal}$: a disjoint training and calibration sets where $\mathcal{D}_{train} \bigcup \mathcal{D}_{cal} = \mathcal{D}$ and $n = |\mathcal{D}_{cal}|$.
- f : a predictor fitted on $\mathcal{D}_{train}$
- $S_f : (\boldsymbol{X}, Y) \to \mathbb{R}$: a score function, such as
    - when f is a classifer $S_f(\boldsymbol{x}, y) = 1 - f(\boldsymbol{x})_y$ where $f(\cdot)_y$ being y's predicted likelihood.
- $B_p(\boldsymbol{x}, \epsilon)$: the ϵ-ball centered at $\boldsymbol{x} \in \mathbb{R}^d$ with respect to the p-norm $|| \cdot ||_p$.

Adversarial Attacks

- Intentionally adding imperceptible noise can degrade the performance of an AI model.
- This type of attack is called an ”adversarial attack.”

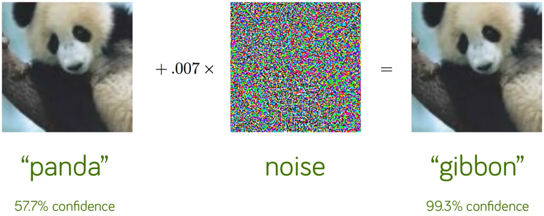

Figure 1: An example of an adversarial attack.
From : Goodfellow, I. J. (2014). Explaining and harnessing adversarial examples

- Various approaches have been proposed to verify the robustness of NNs against adversarial attacks.
- The main target of these approaches is to find a verifier that computes a valid but not exact bound when a given neural network f is subjected to a perturbation within ϵ, such that :

$$f(\boldsymbol{x})_y^{\perp} \leq \inf_{\boldsymbol{x}' \in B_p(\boldsymbol{x},\epsilon)} \{f(\boldsymbol{x}')_y\}, \quad f(\boldsymbol{x})_y^{\top} \geq \sup_{\boldsymbol{x}' \in B_p(\boldsymbol{x},\epsilon)} \{f(\boldsymbol{x}')_y\}. \tag{1}$$

- This paper adopts and utilizes CROWN, the state-of-the-art (SotA) method in this field.

Conformal Prediction

- Given a calibration set $\mathcal{D}_{cal}$, a test input $\boldsymbol{x}_{test}$, and a score function $S(\cdot,\cdot)$.
- Let $s_i = S(\boldsymbol{x}_i, y_i)$
- We construct the score distibution with calibration sets as :

$$F = \frac{\delta_\infty}{n+1} + \sum_{(\boldsymbol{x}_i, y_i) \in \mathcal{D}_{cal}} \frac{\delta_{s_i}}{n+1}, \tag{2}$$

  where δ_s is the Dirac distribution with parameter s, and δ_∞ represents the unknown score (potentially infinite) of the test point.
- Given a miscoverage/error rate α and a test point (x_{test}, y_{test}).
- Then, the prediction set $C(x_{test})$ is defined as:

$$C(x_{test}) = \{y \in Y : S_f(x_{test}, y) \leq Q_{1-\alpha}(F)\}, \tag{3}$$

  where $Q_{1-\alpha}(F)$ is the $1-\alpha$ quantile of F.

Conformal Prediction and Adversarial Attacks

- $C(x_{test})$ satisfies the marginal coverage guarantee if the test point and the calibration points are exchangeable.
- However, when the data is subjected to adversarial attacks, this guarantee no longer holds.

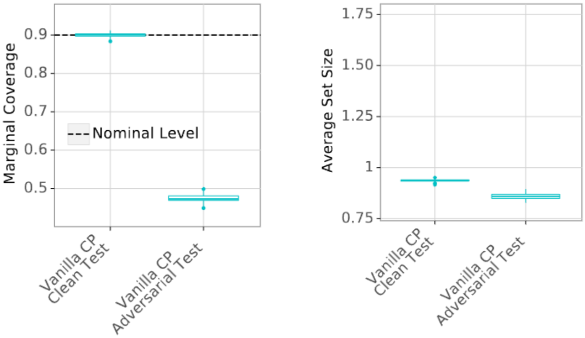

Figure 2: With a target coverage of 0.9, targinal coverage and average set-size obtained by vanilla conformal predictiond, evaluated on the test set of CIFAR10. [1]

Verifiably Robust Conformal Prediction (VRCP)

VRCP via Robust Inference (VRCP-I)

- Given $\mathcal{D}_{cal}$, f, $S(\cdot, \cdot)$ and a test input $\boldsymbol{x}_{test}$.
- compute the prediction set for $\boldsymbol{x}_{test}$ as follows.
  1. For each $y \in \mathcal{Y}$ we compute,

$$s^{\perp}(\boldsymbol{x}_{test}, y) = 1 - f(\boldsymbol{x}_{test})_y^{\top} \leq \inf_{\boldsymbol{x}' \in B(\boldsymbol{x}_{test}, \epsilon)} S(\boldsymbol{x}', y) \tag{4}$$

  2. The robust prediction set is then defined as

$$C_{\epsilon}(\boldsymbol{x}_{test}) = \left\{ y : s^{\perp}(\boldsymbol{x}_{test}, y) \leq Q_{1-\alpha}(F) \right\} \tag{5}$$

VRCP via Robust Inference (VRCP-I)

- Below, the authors show that we are able to maintain the marginal coverage guarantee for any ℓ_p-norm bounded adversarial attack.

Theorem 3.1

Let $\tilde{\boldsymbol{x}}_{test} = \boldsymbol{x}_{test} + \boldsymbol{\delta}$ *for a clean test sample* $\boldsymbol{x}_{test}$ *and* $\|\boldsymbol{\delta}\|_p \leq \epsilon$. *The prediction set* $C_\epsilon\left(\tilde{\boldsymbol{x}}_{test}\right)$ *defined in Eq. (5) satisfies* $\mathbb{P}\left[y_{test} \in C_\epsilon\left(\tilde{\boldsymbol{x}}_{test}\right)\right] \geq 1 - \alpha$.

Proof :

$$
\begin{aligned}
\mathbb{P}\left[y_{test} \in C_\epsilon\left(\tilde{\boldsymbol{x}}_{test}\right)\right] &= \mathbb{P}\left[s^{\perp}\left(\tilde{\boldsymbol{x}}_{test}, y_{test}\right) \leq Q_{1-\alpha}(F)\right] \\
&\geq \mathbb{P}\left[\inf_{\boldsymbol{x}' \in B_\epsilon(\tilde{\boldsymbol{x}}_{test})} S\left(\boldsymbol{x}', y_{test}\right) \leq Q_{1-\alpha}(F)\right] \quad \text{by Eq. (4)} \\
&\geq \mathbb{P}\left[S\left(\boldsymbol{x}_{test}, y_{test}\right) \leq Q_{1-\alpha}(F)\right] \geq 1 - \alpha. \quad \square
\end{aligned}
$$

VRCP via Robust Calibration (VRCP-C)

- Given $\mathcal{D}_{cal}$, f, $S(\cdot,\cdot)$ and a test input $\boldsymbol{x}_{test}$.
- compute the prediction set for $\boldsymbol{x}_{test}$ as follows.
  1. We compute the upper-bound score distribution with calibration sets as:
  $$F^{\top} = \frac{\delta_{\infty}}{(n+1)} + \sum_{(\boldsymbol{x}_i, y_i) \in \mathcal{D}_{cal}} \frac{\delta_{s_i^{\top}}}{n+1}, \text{ where } s_i^{\top} \geq \sup_{\boldsymbol{x}' \in B_p(\boldsymbol{x}_i, \epsilon)} S\left(\boldsymbol{x}', y_i\right) \tag{6}$$
  2. The robust prediction set is then defined as
  $$C_{\epsilon}\left(\boldsymbol{x}_{test}\right) = \left\{y : S\left(\boldsymbol{x}_{test}, y\right) \leq Q_{1-\alpha}\left(F^{\top}\right)\right\} \tag{7}$$

VRCP via Robust Calibration (VRCP-C)

Theorem 3.2

Let $\tilde{\boldsymbol{x}}_{test} = \boldsymbol{x}_{test} + \boldsymbol{\delta}$ for a clean test sample $\boldsymbol{x}_{test}$ and $\|\boldsymbol{\delta}\|_p \leq \epsilon$. The prediction set $C_\epsilon(\tilde{\boldsymbol{x}}_{test})$ defined in Eq. (7) satisfies $\mathbb{P}[y_{test} \in C_\epsilon(\tilde{\boldsymbol{x}}_{test})] \geq 1 - \alpha$.

Proof :

$$
\begin{aligned}
&\mathbb{P}[y_{test} \in C_\epsilon(\tilde{\boldsymbol{x}}_{test})] = \mathbb{P}\left[S(\tilde{\boldsymbol{x}}_{test}, y_{test}) \leq Q_{1-\alpha}\left(F^\top\right)\right] \\
&\geq \mathbb{P}\left[S(\tilde{\boldsymbol{x}}_{test}, y_{test}) \leq Q_{1-\alpha}\left(\left\{\sup_{\boldsymbol{x}' \in B_\epsilon(\boldsymbol{x}_i)} S\left(\boldsymbol{x}', y_i\right)\right\}_{(\boldsymbol{x}_i, y_i) \in \mathcal{D}_{cal}} \cup \{\infty\}\right)\right] \\
&\geq \mathbb{P}\left[\sup_{\boldsymbol{x}' \in B_\epsilon(\boldsymbol{x}_{test})} S\left(\boldsymbol{x}', y_{test}\right) \leq Q_{1-\alpha}\left(\left\{\sup_{\boldsymbol{x}' \in B_\epsilon(\boldsymbol{x}_i)} S\left(\boldsymbol{x}', y_i\right)\right\}_{(\boldsymbol{x}_i, y_i) \in \mathcal{D}_{cal}} \cup \{\infty\}\right)\right] \\
&\geq 1 - \alpha. \quad \square
\end{aligned}
$$

Experiments

Experiments

- **Dataset** : CIFAR10, CIFAR100, TinyImageNet
- **Models** : CNN model
- **Attacks** : PGD attack. ℓ_2-norm bounded attacks with $\epsilon = 0.02$ or $\epsilon = 0.03$.
- Target coverage : $1 - \alpha = 0.9$, repeated 50 times.

- **Results**

	CIFAR10		CIFAR100		TinyImageNet	
Method	Coverage	Size	Coverage	Size	Coverage	Size
Vanilla	0.878±0.002	1.721±0.008	0.890±0.002	6.702±0.058	0.886±0.002	38.200±0.252
RSCP+	1.000±0.000	10.000±0.000	1.000±0.000	100.000±0.000	1.000±0.000	200.000±0.000
RSCP+ (PTT)	0.983±0.008	8.357±0.780	0.925±0.010	26.375±9.675	0.931±0.013	90.644±20.063
VRCP–I	0.986±0.000	**4.451±0.011**	0.971±0.001	**22.530±0.107**	0.958±0.001	**72.486±0.311**
VRCP–C	0.995±0.000	5.021±0.010	0.983±0.000	23.676±0.131	0.965±0.001	77.761±0.352

End