Reading Notes on defenses

1 TSS: Transformation-Specific Smoothing for Robustness Certification

Semantic Transformation Attacks (e.g. Rotate, Contrast, Brightness, etc.)

- Often not constrainted by norm metrics
- Forms a very low dimension space (comparing with the image space)
- Preserve semantic information
- Some may introduce interpolation error

Formalizing transformation:

- Consider the image set $X \subseteq \mathbb{R}^d$ and label set $Y \subseteq [1..c]$.
- Let $\mathcal{Z} \subseteq \mathbb{R}^m$ be the space of transformation parameters.
- Define a transformation as a mapping $\phi : X \times \mathcal{Z} \to X$.

For an arbitrary classifier $h(x)$, we denote its ϵ-smoothed classifier as:

$$g(x;\epsilon) = \arg\max_{y \in Y} \mathbb{E}_{\epsilon \sim \mathbb{P}_\epsilon}(p(y|\phi(x,\epsilon))) \tag{1}$$

The theorm of robustness for general transformation

- Let $\epsilon_0 \sim \mathbb{P}_0$ and $\epsilon_1 \sim \mathbb{P}_1$ be random variables in $\mathcal{Z}$, with pdf f_0, f_1.
- Let $y_A = g(x;\epsilon_0)$ where g is a smoothed classifier of $\phi : X \times \mathcal{Z} \to X$.
- Suppose there is a probability gap between the top-2 classes. Specifically, we have
  - $$\mathbb{E}_{\epsilon_0 \sim \mathbb{E}_{\epsilon_0}} p(y_A|\phi(x,\epsilon_0)) \geq p_A > p_B \geq \max_{y \neq y_A} \mathbb{E}_{\epsilon_0 \sim \mathbb{E}_{\epsilon_0}} p(y|\phi(x,\epsilon_0)) \tag{2}$$
- For $t \geq 0$, $\underline{S_t} := \{f_1/f_0 < t\}$ and $\overline{S_t} := \{f_1/f_0 \leq t\}$. Define the function $\xi : [0,1] \to [0,1]$ by:

$$\xi(p) := \sup\{\mathbb{P}_1(S) : \underline{S}_{\tau_p} \subseteq S \subseteq \overline{S}_{\tau_p}\} \tag{3}$$

- where $\tau_p := \inf\{t \geq 0 : \mathbb{P}_0(\overline{S_t}) \geq p\}$.

Then, if $\xi(p_A) + \xi(1-p_B) > 1$, $g(x;\epsilon_0) = g(x;\epsilon_1)$.

Note: Here is a possible intuitive explanation of this theorem.

- Why $\xi(p_A)$ represents the lower bound of $\mathbb{E}_{\epsilon_1} p(y \mid \phi(x,\epsilon_1))$? Consider the problem of minimizing

$$\min \int f_1(\mu) h_A(\mu) \mathrm{d}\mu \quad \text{given} \quad \int f_0(\mu) h_A(\mu) \mathrm{d}\mu = p_A \tag{4}$$

- This is actually a continuous version of the knapsack problem, so a greedy algorithm is optimal!
- We only need to greedily fill $h(\mu) = y_A$ in the ascending order of f_1/f_0. To describe it formally,

$$h(\mu) = y_A \text{ iff } \mu \in S, \text{where } \underline{S}_t \subseteq S \subseteq \overline{S}_t \text{ and } \int_{\mu \in S} f_0(\mu) = p_A. \tag{5}$$

- Therefore, $\min \int f_1(\mu) h_A(\mu) \mathrm{d}\mu = \int_{\mu \in S} f_1(\mu) \mathrm{d}\mu = \mathbb{P}_1(S)$.

Taxonomy of Semantic Attacks

- Resolvable: if $\forall \alpha \in \mathcal{Z}$, there exists a resolving function $\gamma_\alpha : \mathcal{Z} \to \mathcal{Z}$ that is injective, continuously differentiable with non-vanishing Jacobian, and that

$$\phi(\phi(x, \alpha), \beta) = \phi(x, \gamma_\alpha(\beta)) \quad x \in X, \beta \in \mathcal{Z} \tag{6}$$

  - This equation defines the composition rule of resolvable transformations, so that an attack $\phi(x, \alpha)$ would be smoothed by a ϵ_0 smoother to $\epsilon_1 = \gamma_\alpha(\epsilon_0)$. The general theorem of robustness is thus applicable.
- Differentiable Resolvable: if $\forall x \in X$, there exists a resolvable transformation $\psi : X \times Z_\psi \to X$ and a function $\delta_x : Z_\phi \times Z_\phi \to Z_\psi$, such that

$$\phi(\phi(x, \alpha), \beta) = \psi(\phi(x, \beta), \delta_x(\alpha, \beta)) \tag{7}$$

Examples

- Gaussian Blur: take convolution with Gaussian function. $\phi_B(x, \alpha) = x * G_\alpha$. It is not only resolvable, but additive as well.
- Brightness + Contrast. $\phi_{BC}(x, \alpha) = e^k(x + b)$ where $\alpha = (k, b)^T$.

End-to-end Robustness for Sensing-Reasoning Machine Learning Pipelines

Many previous methods have been proposed to certify the robustness of machine learning models the pertubation bounded in a small ℓ_p ball. In this paper, a generic Sensing-Reasoning machine learning pipelines was proposed, in which the previous methods were viewed as a certification of sensing robustness. The output of sensing (deep learning) models were combined with embedded domain knowledge in reasoning components to provide end-to-end robustness. This pipeline is a generic framework since the choices of specific certified robust sensing models are orthogonal to the certification of reasoning robustness.

The analyses of reasoning robustness started with showing the hardness of certifying the robustness of a general reasoning model. By proving the polynomial time reduction of the counting problem to the robustness problem, certifying the robustness of a general reasoning component was proved to be #P-hard.

Robustness bounds for several reasoning structures including Markov logic networks and Bayesian networks were proved...

Paper List

TSS: Transformation-Specific Smoothing for Robustness Certification

Knowledge Enhanced Machine Learning Pipeline against Diverse Adversarial Attacks

End-to-end Robustness for Sensing-Reasoning Machine Learning Pipelines

Certified Robustness to Adversarial Examples with Differential Privacy

MACER: Attack-free and Scalable Robust Training via Maximizing Certified Radius

Provably Robust Deep Learning via Adversarially Trained Smoothed Classifiers

Certified Adversarial Robustness via Randomized Smoothing

On the Effectiveness of Interval Bound Propagation for Training Verifiably Robust Models

MixTrain: Scalable Training of Verifiably Robust Neural Networks

Scaling provable adversarial defenses

TBA