Certifiable Robustness only formulas

Fabio Brau

March 1, 2023

Scuola Superiore Sant'Anna, Pisa.

TELECOMMUNICATIONS, COMPUTER ENGINEERING, AND PHOTONICS INSTITUTE

Sant'Anna School of Advanced Studies – Pisa

Retis Real-Time Systems Laboratory

Boundary

$$\mathbf{f} : \mathbb{R}^{\mathbf{n}} \rightarrow \mathbb{R}^{\mathbf{C}}$$

$$\mathcal{K}_f(x) = \operatorname*{argmax}_i f_i(x)$$

$$\mathcal{B} = \left\{ p \in \mathbb{R}^n \ : \ f(p) = 0 \right\}$$

$$d(x, l) = \min_{\delta} \quad \|\delta\|$$

$$\text{s.t} \quad f_l(x + \delta) - \max_{j \neq l} f_j(x + \delta) \leq 0$$

General Case into Binary Case

$$F_l(x) = f_l(x) - \max_{j \neq l} f_j(x)$$

$$d(x, l) = d_l(x)$$

Verification

$$\zeta(x): \quad \text{"}\forall y \in \mathcal{N}(x) \quad \mathcal{K}(x) = \mathcal{K}(y)\text{"}$$

where $\mathcal{N}(X)$ a neighborhood of x.

$$\mathcal{N}(x) = \{y \in \mathbb{R}^n \,:\, \|y - x\|_\infty \leq \varepsilon\}$$

$$\begin{aligned} P(x) = \min_{j \neq l} \min_{y \in \mathbb{R}^n} & \quad f_l(y) - f_j(y) \\ \text{s.t.} & \quad -\varepsilon \leq x_i - y_i \leq \varepsilon, \forall i \end{aligned}$$

$$\zeta(x) \quad \Leftrightarrow \quad P(x) > 0$$

Verifiers

$$
\begin{aligned}
\hat{z}^{(i)} &= W_i z^{(i)} + b_i && i = 1, \cdots, L-1 \\
z^{(i)} &= \max\{0, \hat{z}^{(i)}\} && i = 2, \cdots, L-1
\end{aligned}
$$

where $z_1 \equiv x$

$$
\begin{aligned}
P(x) = \min_{j \neq l} \min_{y \in \mathbb{R}^n} \quad & \hat{z}_l^{(L)} - \hat{z}_j^{(L)} \\
\text{subject to} \quad & -\varepsilon \leq x - z^{(0)} \leq \varepsilon \\
& \hat{z}^{(i)} = W_i z^{(i)} + b_i, \quad i = 1, \cdots, L-1 \\
& z^{(i)} = \max\{0, \hat{z}^{(i)}\}, \quad i = 2, \cdots, L-1
\end{aligned}
$$

Relaxing

$$z = \max\{0, \hat{z}\}$$

Relax to

$$z \geq 0$$
$$z \geq \hat{z}$$
$$-u\hat{z} + (u - l)z \leq -ul$$

$$\begin{aligned} P(x) = \min_{j \neq l} \min_{y \in \mathbb{R}^n} \quad & \hat{z}_l^{(L)} - \hat{z}_j^{(L)} \\ \text{subject to} \quad & -\varepsilon \leq x - z^{(0)} \leq \varepsilon \\ & \hat{z}^{(i)} = W_i z^{(i)} + b_i, \quad i = 1, \cdots, L-1 \\ & z^{(i)} \geq 0, \quad i = 2, \cdots, L-1 \\ & z^{(i)} \geq \hat{z}^{(i)}, \quad '' \\ & -u^{(i)}\hat{z}^{(i)} + (u^{(i)} - l^{(i)})z^{(i)} \leq -u^{(i)} l^{(i)}, \quad '' \end{aligned}$$

Robust Training

$$(x_i, y_i) \in \mathcal{X}, i = 1, \cdots, N$$

$$\theta^* \in \operatorname*{argmin}_{\theta} \frac{1}{N} \sum_{i=1}^{N} \max_{\|\delta\| \leq \varepsilon} L(f_\theta(x_i + \delta), y_i)$$

L-Lipschitz

$$\forall x, y \in \mathbb{R}^n, \quad \|f(x) - f(y)\|_p \leq L\|x - y\|_p$$

$$\forall \delta, \|\delta\|_p \leq \varepsilon, \quad \|f(x) - f(x+\delta)\|_p \leq L\|\delta\|_p$$

$$\beta_L(x) = \min_{j \neq l} \frac{f_l(x) - f_j(x)}{L\, 2^{\frac{p-1}{p}}}$$

$$\text{"}\varepsilon < \beta_L(x)\text{"} \Rightarrow \zeta(x)$$

$$f_l - f_j \quad \text{is } L_j\text{-lipschitz in } B_p(x, R)$$

$$\beta_L(x) = \min_{j \neq l} \frac{f_l(x) - f_j(x)}{L_j}$$

$$L_j = \max_{y \in B_p(x,R)} \|\nabla f_l(y) - \nabla f_j(y)\|_q$$

where $\frac{1}{p} + \frac{1}{q} = 1$

in $B_p(x, R)$

Lipschitz Function

$$f = f^{(L)} \circ f^{(L-1)} \circ \cdots \circ f^{(1)}$$

$$L = \prod_{i=1}^{L} L_i$$

$$f(x) = Wx + b$$

$$\|f(y) - f(x)\|_p = \|Wy - Wx + b - b\|_p$$

$$(y - x = v)$$

$$\|f(y) - f(x)\|_p = \|Wv\|_p$$

$$\frac{\|f(y) - f(x)\|_p}{\|y - x\|_p} = \frac{\|Wv\|_p}{\|v\|_p} \leq \sup_{v \in \mathbb{R}^n \setminus \{0\}} \frac{\|Wv\|_p}{\|v\|_p}$$

$$W \in \mathbb{R}^{m \times n}, \quad \|W\|_p := \sup_{v \in \mathbb{R}^n \setminus \{0\}} \frac{\|Wv\|_p}{\|v\|_p}$$

$$\|f(y) - f(x)\|_p = \|W\|_p \, \|y - x\|_p$$

$$\|A\|_2 = \sqrt{\lambda_{max}(A^T A)} = \sigma_{max}(A)$$

$$Q \in \mathbb{R}^{n \times n} \quad QQ^T = Q^T Q = I$$

$$f_W(x) = \tilde{W} x + b, \quad \text{where} \quad \tilde{W} = \frac{W}{\|W\|_2}$$

Orthogonal

$$
\begin{aligned}
f_Q(x) &= Qx + b \\
Q_k &= I - W_k^T W_k \\
W_{k+1} &= W_k \left(I + \frac{1}{2} Q_k + \frac{3}{8} Q_k^2 + \cdots + (-1)^p \binom{-\frac{1}{2}}{p} Q_k^p \right) \\
W_0 &= W
\end{aligned} \tag{1}
$$

Cayley

$$A = W - W^T$$

$$Q = (I - A)(I + A)^{-1}$$

$$A = W - W^T$$

$$Q = \exp(A) := \sum_{k=0}^{\infty} \frac{A^k}{k!}$$

$$(x_i, y_i) \in \mathcal{X}, \quad i = 1, \cdots, N$$

$$\mathcal{A}_\varepsilon(\mathcal{X}) = \frac{1}{N} \# \{ i : \mathcal{K}_f(x_i) = y_i,\, d(x_i) \geq \varepsilon \}$$

$$\tilde{\mathcal{A}}_\varepsilon(\mathcal{X}) = \frac{1}{N} \# \{ i : \mathcal{K}_f(x_i) = y_i,\, \beta(x_i) \geq \varepsilon \}$$

Thanks for the attention

Fabio Brau

Scuola Superiore Sant'Anna, Pisa

fabio.brau@santannapisa.it

retis.santannapisa.it/~f.brau

linkedin.com/in/fabio-brau