Connecting Certified and Adversarial Training

Yuhao Mao, Mark Niklas Müller, Marc Fischer, Martin Vechev

Department of Computer Science
ETH zürich SRILAB

TAPS (this work)

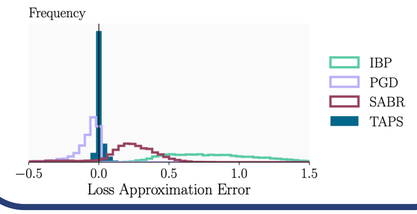
- Combine IBP and PGD gradients to allow for joint training.
- Over-approximation of IBP and under-approximations of PGD partially cancel out.
- Improve both certified and standard accuracies.

Background

- Robustness: $\forall i, x', \text{ s.t. } \|x' - x\|_\infty \leq \epsilon, f(x')_{i^*} - f(x')_i \geq 0$
- Certified Training: $L(\boldsymbol{x}, y, \epsilon) := \ln[1 + \sum_{i \neq y} \exp\left(\bar{o}_i^{\Delta}\right)]$
- Interval Bound Propagation (IBP): use interval arithmetic, e.g., $[a, b] + [c, d] = [a + c, b + d]$

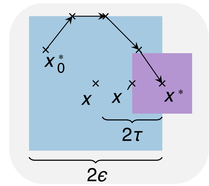
More Precise Bounds

- IBP ➔ over-approximation
- PGD ➔ under-approximation
- SABR ➔ better but large variance
- TAPS ➔ precise and concentrated



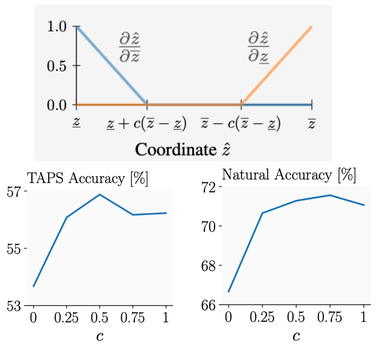
Orthogonal to Propagation Region

IBP ➔ SABR
TAPS ➔ STAPS

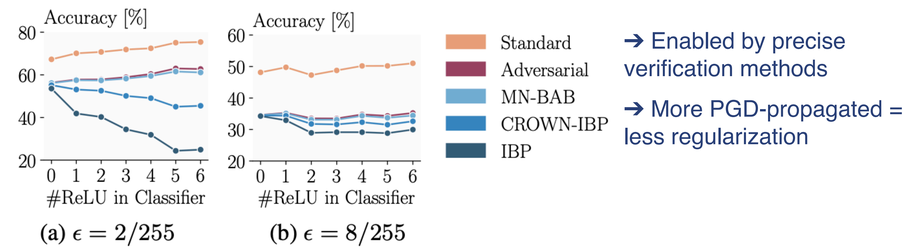


Connecting Adversarial Examples with Bounds

- General form: $\frac{dL}{dz_i} = \sum_j \frac{dL}{d\hat{z}_j} \frac{\partial \hat{z}_j}{\partial z_i}$
- Dimension independence: $\frac{dL}{dz_i} = \frac{dL}{d\hat{z}_i} \frac{\partial \hat{z}_i}{\partial z_i}$
- Our design: $\frac{\partial \hat{z}_i}{\partial z_i} = \max\left(0{,}1 - \frac{\hat{z}_i - z_i}{c\left(\bar{z}_i - z_i\right)}\right)$
- $c = 0.5$ ➔ smooth and unique connection



Ablation Study



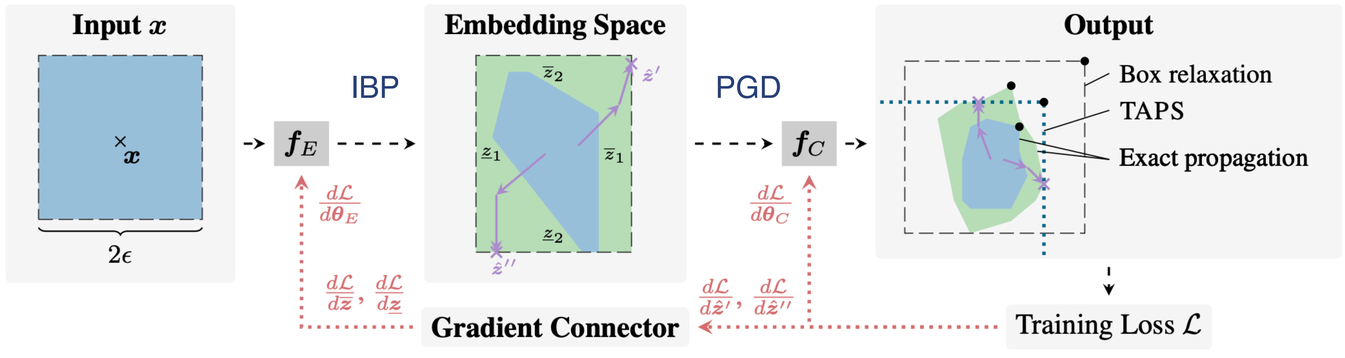
➔ Enabled by precise verification methods

➔ More PGD-propagated = less regularization

# Attack Steps	1 Restart		3 Restarts	
	Certified	Natural	Certified	Natural
1	93.36	**98.22**	93.47	**98.22**
5	93.15	97.90	**93.55**	97.90
20	**93.62**	97.94	93.52	97.99
100	93.46	97.94	**93.55**	97.99

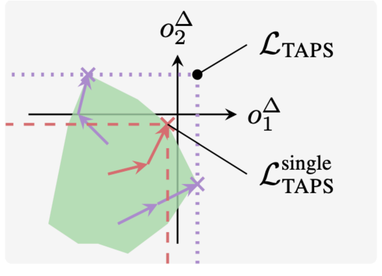
➔ Stable with regards to attack strength

➔ Even single-step can get good results



PGD: Multi-estimator (ours) vs Single-estimator (original)

- Single-estimator PGD could ignore adversarial examples even in optimal case.
- Multi-estimator explicitly regularize maximum margin.



$$L_{\text{TAPS}}^{\text{single}}(\boldsymbol{x}, y, \epsilon) = \max_{\hat{z} \in [\underline{z}, \bar{z}]} \ln\left(1 + \sum_{i \neq y} \exp\left(f_C(\hat{z})_i - f_C(\hat{z})_y\right)\right)$$

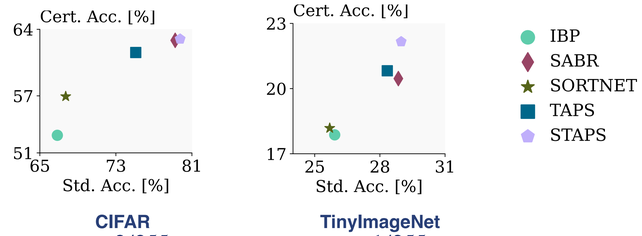
$$L_{\text{TAPS}}(\boldsymbol{x}, y, \epsilon) = \ln\left(1 + \sum_{i \neq y} \exp\left(\max_{\hat{z} \in [\underline{z}, \bar{z}]} f_C(\hat{z})_i - f_C(\hat{z})_y\right)\right)$$

# ReLU in Classifier	Single		Multi	
	Certified	Natural	Certified	Natural
1	-†	31.47†	**93.62**	97.94
3	92.91	98.56	93.03	98.63
6	92.41	**98.88**	92.70	**98.88**

† Training encounters mode collapse. Last epoch performance reported.

SOTA - Empirical Results

Better certified *and* standard accuracies than current state-of-the-art certified training.



NEURAL INFORMATION PROCESSING SYSTEMS