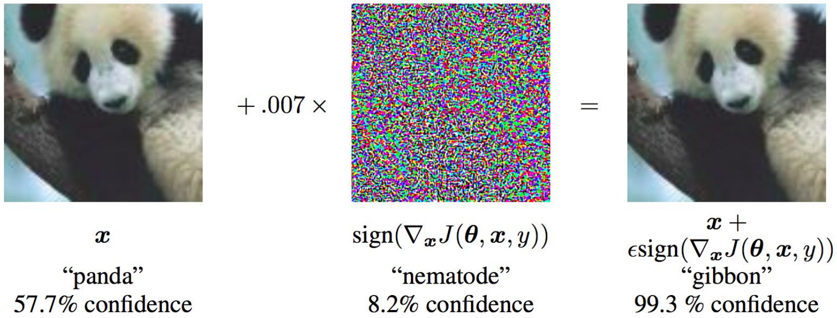
PRINCETON UNIVERSITY

MultiRobustBench: Benchmarking Robustness Against Multiple Attacks

Sophie Dai, Saeed Mahloujifar, Chong Xiang, Vikash Sehwag, Pin-Yu Chen, Prateek Mittal

Adversarial Examples

- Imperceptible noise added at test-time to cause misclassification
- To evaluate the performance of defenses, we typically consider a single attack type (ie. Lp bounded attacks)
- But we would like robustness against the space of **all imperceptible perturbations**!



Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

→ We should be evaluating our models across **multiple attack types**!

What does it mean to be robust against multiple attacks?

- Perturbation function $P : X \times Y \times \mathcal{H}$
  - Maps input and function to a perturbed example
- Knowledge set K
  - Set of perturbation functions
- Learning algorithm (defense) $\mathcal{A} : D \times K \to \mathcal{H}$
  - Gives a robust model based off of a dataset and *knowledge set*
  - Learning algorithm can only use information about perturbation functions in K (ie. through queries)

Adversarial Game for Multiple Attacks

1. Environment specifies a robustness threshold γ and specifies a set K of perturbation functions that can occur during test-time. The environment also specifies the learner's knowledge set K_{learner}.
2. The learner then chooses learning algorithm $\mathcal{A}$ and obtains model
$$h = \mathcal{A}(D_{\text{train}}, K_{\text{learner}})$$
3. If $$\frac{\text{err}_{\text{multi}}(h; K)}{\min_{h^* \in \mathcal{H}} \text{err}_{\text{multi}}(h^*; K)} \leq \gamma$$
the learner wins and $\mathcal{A}$ produces a model that is close to optimal against K. Otherwise the attacker wins.

Competitiveness Ratio

Let $\text{acc}^*_{\text{multi}}(K) := 1 - \min_{h^* \in \mathcal{H}} \text{err}_{\text{multi}}(h^*; K)$ and $\text{acc}_{\text{multi}}(h, K) := 1 - \text{err}_{\text{multi}}(h; K)$. Then, the competitiveness ratio (CR) of a defended model h is given by:

$$\text{CR}(h; K) = 100 \times \frac{\text{acc}_{\text{multi}}(h, K)}{\text{acc}^*_{\text{multi}}(K)} \tag{1}$$

For a single $P \in K$, let $\text{acc}^*(P) := 1 - \min_{h \in \mathcal{H}} \text{err}(h; P)$ and $\text{acc}(h, P) := 1 - \text{err}(h; P)$. Then,

$$\text{CR}_{\text{ind-avg}}(h; K) := 100 \times \mathbb{E}_{P \sim \mathcal{P}(K)} \left[\frac{\text{acc}(h, P)}{\text{acc}^*(P)} \right] \tag{2}$$

$$\text{CR}_{\text{ind-worst}}(h; K) := 100 \times \min_{P \in K} \frac{\text{acc}(h, P)}{\text{acc}^*(P)} \tag{3}$$

Measures how well defense does compared to optimal (which we approximate by adversarial training on each individual attack)

Stability Constant

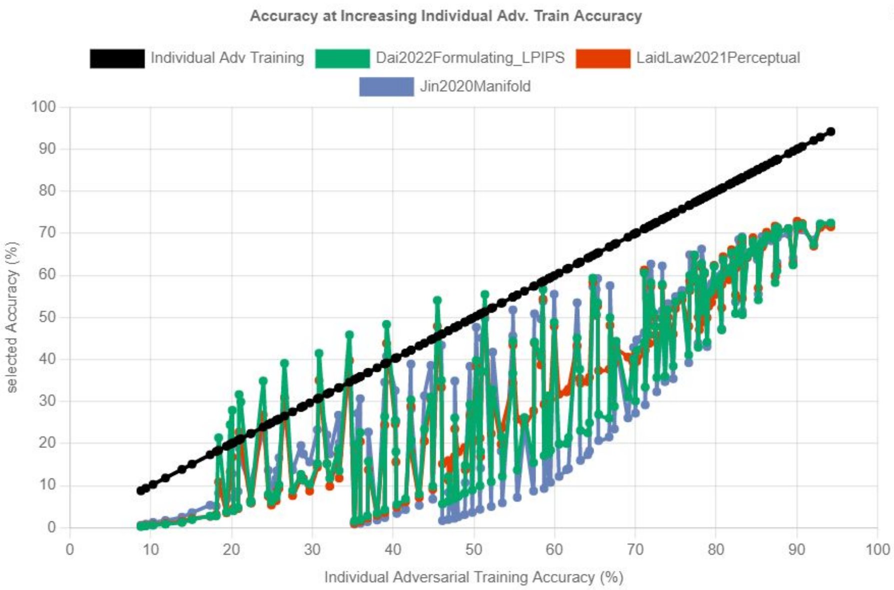
A model h is (L, α)-locally stable across perturbations with respect to attack strength function s if we have that for all $P_1 \in K_{\text{learner}}$ and $P_2 \in K$ such that $|s(P_1) - s(P_2)| \leq \alpha$, $|\text{acc}(h, P_1) - \text{acc}(h, P_2)| \leq L|s(P_1) - s(P_2)|$. Equivalently, for a given α and model h, we can compute the corresponding constant L, which we call the *stability constant (SC)* as follows:

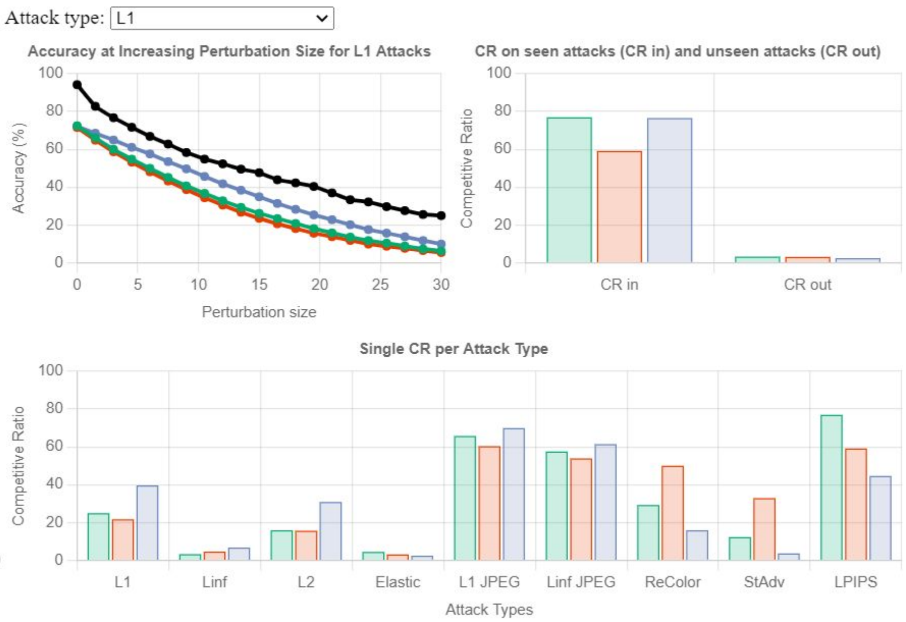
$$L_\alpha(h) = \max_{\substack{P_1 \in K_{\text{learner}}, P_2 \in K \\ |s(P_1) - s(P_2)| \leq \alpha \\ P_1 \neq P_2}} \frac{|\text{acc}(h, P_1) - \text{acc}(h, P_2)|}{|s(P_1) - s(P_2)|} \tag{1}$$

Measures how much the accuracy of the model fluctuates when a slightly harder attack is used

MultiRobustBench

- Available at multirobustbench.github.io
- Provides computed CR and SC scores for existing defenses tailored against multiple attacks for average-case and worst-case robustness
- Provides visualizations to understand weaknesses of existing models
- Leaderboard is computed across 9 different attacks at 20 different strengths





Ablation studies

- Our paper also provides ablations on the impact of architecture size, impact of number of training epochs, and impact of additional training data for adversarial training on different threat models
- Overall:
  - Extra data generally improves average-case robustness
  - Worst-case robustness scores are generally dominated by spatial attacks
  - Smaller models generally have better CR scores, but have lower clean accuracy