PARADE: Provably Robust ADversarial Examples

Dimitar I. Dimitrov, Gagandeep Singh, Timon Gehr, Martin Vechev

Problem setting

- Traditionally, robustness of $x_{\text{orig}}$ is assessed by generating individual attacks $x_1$ and $x_2$ within a ball around it (in red).
- Description of the whole adversarial region (in green) is preferable. The region can contain trillions of adversarial images.

Single adversarial attacks vs Adversarial regions

- Single attacks:
  - Easy to generate
  - Less informative
  - Computationally expensive

- Adversarial regions:
  - More Informative
  - Efficiently summarizes many individual attack
  - Computationally expensive

Key idea: Use single attacks to generate initial region and refine it until provably verifiable.

Algorithm overview

I. Use PGD to generate many individual attacks. Fit a hyperbox around them to restrain search region. The region is shown in blue.

II. Use the overapproximation box to initialize. At each step use black box verification tool based on convex-relaxations to generate a half-space constraint which if added to the current box makes the resulting region verifiably adversarial. Adjust the constraints' bias such that a part of the box is removed but the constraint is weaker. Create maximal box not intersecting the adjusted constraint. Repeat until verification succeeds. Results in hyperbox robust example. The region is shown in purple.

III. Initialize with the overapproximation box. At each step use black box verification tool to generate half-space constraints that force the ReLU neurons to become decided and for the verification objective to become positive. Bias-adjust them so they do not intersect the underapproximation box region. This enforces the polyhedral region to be larger than the hyperbox example. Repeat until verification. Results in polyhedral robust example. The region is shown in red.

Robust adversarial examples to geometric perturbations

- PARADE can handle diverse combinations of geometric perturbations, as it relies on DeepG in a black-box way.
- In similar time, PARADE generates more verifiable regions containing more images compared to baseline on splitting.

Robust adversarial examples and Randomized Smoothing

- PARADE produces regions that are more robust (have bigger robust radius verified using smoothing) compared to uniform shrinking and individual attacks used during Step I of the algorithm.

Empirically vs Provably robust adversarial examples

- Empirical examples can exhibit high Expectation-Over-Transformation (EoT), while their subregions close to the original attacked point incur very low EoT scores.
- Empirically robust adversarial example techniques recovered less regions: 44 vs 24.

Visualisation of Robust Adversarial Examples

- PARADE regions contain up to $10^{159}$ individual adversarial images.
- PARADE produces adversarial regions for all but one adversarial image.
- Regions generated by PARADE are much larger than uniform shrinking baseline.
- PARADE hyperbox example generation is 2x faster than the uniform shrinking baseline.