ProbsCut: Enhancing Adversarial Robustness via Global Probability Constraints

Keji HAN, Yao GE, Yun LI

Frontiers of Computer Science, DOI: 10.1007/s11704-025-41225-3

Problems & Ideas

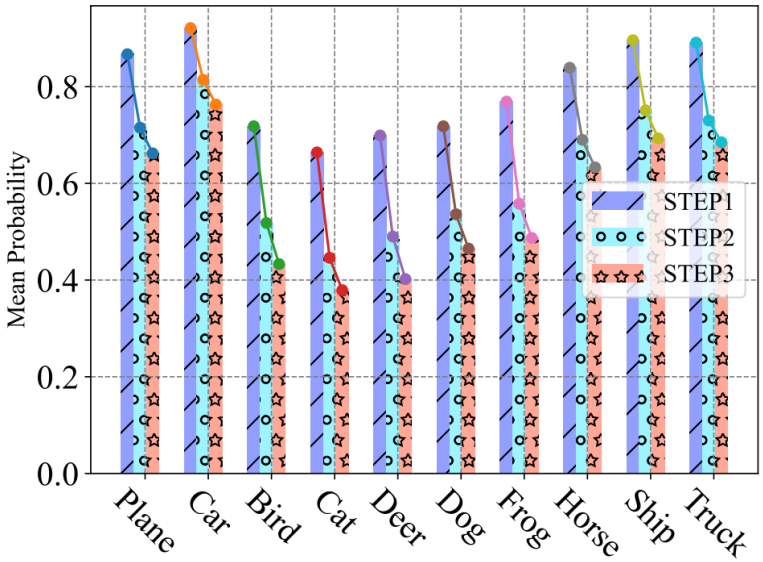
- Problems of conventional adversarial training approaches:
  - Exploring a reasonable trade-off between accuracy and robustness remains an open challenge;
  - Existing adversarial training methods fail to stabilize model decision-making.
- Ideas: Probcuts integrates both local and global loss functions to minimize variance loss while optimizing the expected class probability.

$$\mathcal{L}_{ProbsCut} = \underbrace{KL_s(f_y(x;\theta), \bar{f}_y)}_{global} + \underbrace{\lambda KL(f(x;\theta), f(x_{adv})}_{local}$$

We define a single-element Kullback–Leibler divergence under a global probability constraint to stabilize model decision-making in adversarial settings. Meanwhile, Probcuts employs a local divergence—namely a vanilla Kullback–Leibler term—to align the probabilities of clean examples with their corresponding adversarial counterparts.

Main Contributions

- Contributions:
  - We propose a single-element Kullback–Leibler divergence to reduce variance loss in adversarial settings;
  - We enforce both global and local probability constraints to stabilize the decision-making of deep neural networks, thereby achieving a more favorable robustness–accuracy trade-off.



Our method proves that the average predicted probability for each CIFAR-10 class remains well below 1, accompanied by an improvement in adversarial robustness.