Optimization For Robustness Evaluation Beyond l_p Metrics

Hengyue Liang, Tiancong Chen, Buyun Liang, Ying Cui, Tim Mitchell, Ju Sun

University of Minnesota

Robustness Evaluation Problems

Current **(adversarial) robustness evaluation** of neural networks are generally formulated as solving the following two forms of **constrained optimization problems**:

- Finding an **adversarial perturbation** via max form:

$$\max_{\boldsymbol{x}'} \ell(\boldsymbol{y}, f_{\boldsymbol{\theta}}(\boldsymbol{x}'))$$
$$\text{s.t. } d(\boldsymbol{x}, \boldsymbol{x}') \le \varepsilon\,, \quad \boldsymbol{x}' \in [0,1]^n$$

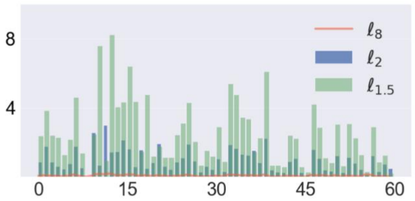
- Finding the **robustness radius** via min form:

$$\min_{\boldsymbol{x}'} d(\boldsymbol{x}, \boldsymbol{x}')$$
$$\text{s.t. } \max_{i \neq y} f^i_{\boldsymbol{\theta}}(\boldsymbol{x}') \ge f^y_{\boldsymbol{\theta}}(\boldsymbol{x}')\,,\ \boldsymbol{x}' \in [0,1]^n$$

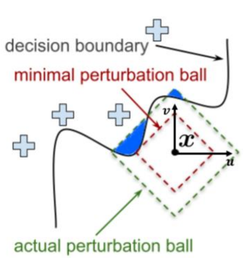
PyGRANSO with Constraint-Folding (PWCF)

1. General purposed non-linear optimization solver
2. Can handle non-smooth functions
3. With GPU-acceleration --- Deep Learning OK

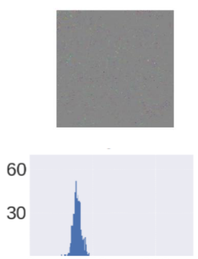
- Can solve both formulations with general distance metrics with high quality. E.g., min form with L8 and L1.5 distances below:



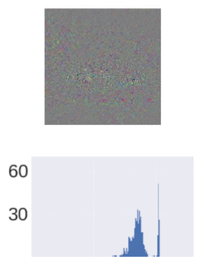
Solution patterns depending on solvers used



- E.g., solving Max-Form with L1



- APGD



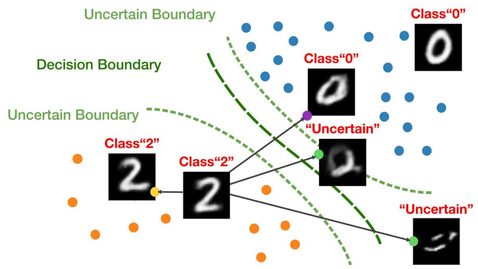
- PWCF

What this may imply:

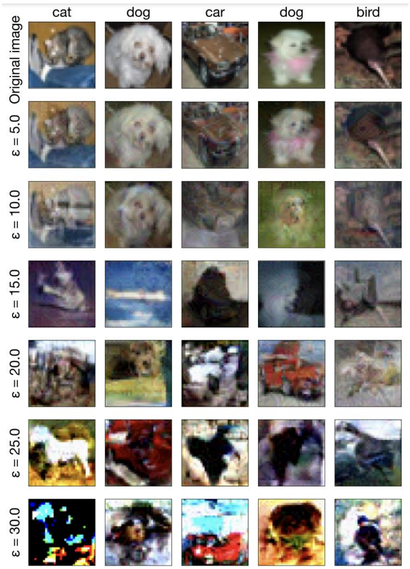
1. **Current robustness evaluation may be insufficient and misleading**
2. **Abs robustness may be hard to achieve**

[1] Liang H, Liang B, Peng L, Cui Y, Mitchell T, Sun J. Optimization and optimizers for adversarial robustness. arXiv preprint arXiv:2303.13401. 2023 Mar 23.

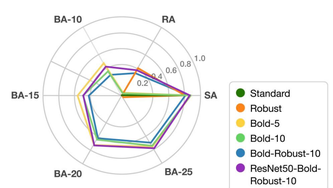
Uncertainty-aware Boldness



A classifier should achieve good: 1) generalizability 2) robustness 3) uncertainty-aware Boldness.
Introduce a new evaluation metric towards reliability: boldness accuracy (BA)
Existing models, including robust models, are not uncertainty-aware bold



Formulate robustness and uncertainty-aware boldness as min-max optimization problem, improve the overall performance



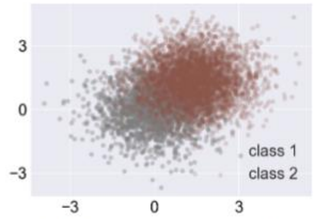
Selective Classification (SC)

- Selectively making predictions to avoid excessive errors.
- Beneficial to deploy the imperfect AI models to practical applications with high-stakes requirements

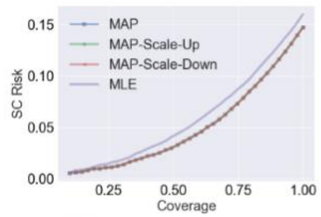
E.g., an AI-powered medical diagnosis assistant can make confident and correct predictions on its own, saving a significant amount of doctors' labor, while turning unconfident cases to doctor.

Calibrated confidence ≠ selection confidence

- A simple example and SC performance of 4 score functions

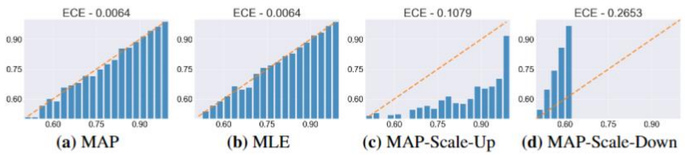


(a) Visualization of samples



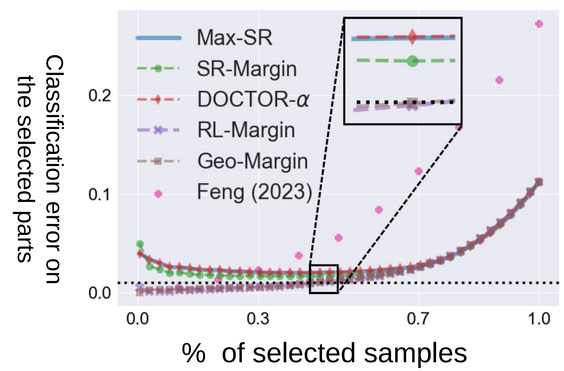
(b) Risk-coverage curves

- Their empirical calibration assessment



(a) MAP (b) MLE (c) MAP-Scale-Up (d) MAP-Scale-Down

“Margin” is better than maximum softmax scores



[N] Liang H, Peng L, Sun J. Toward Effective Post-Training Selective Classification for High-Stakes Applications. In preparation for Neurips 2023.