CVPR 2023 Highlight

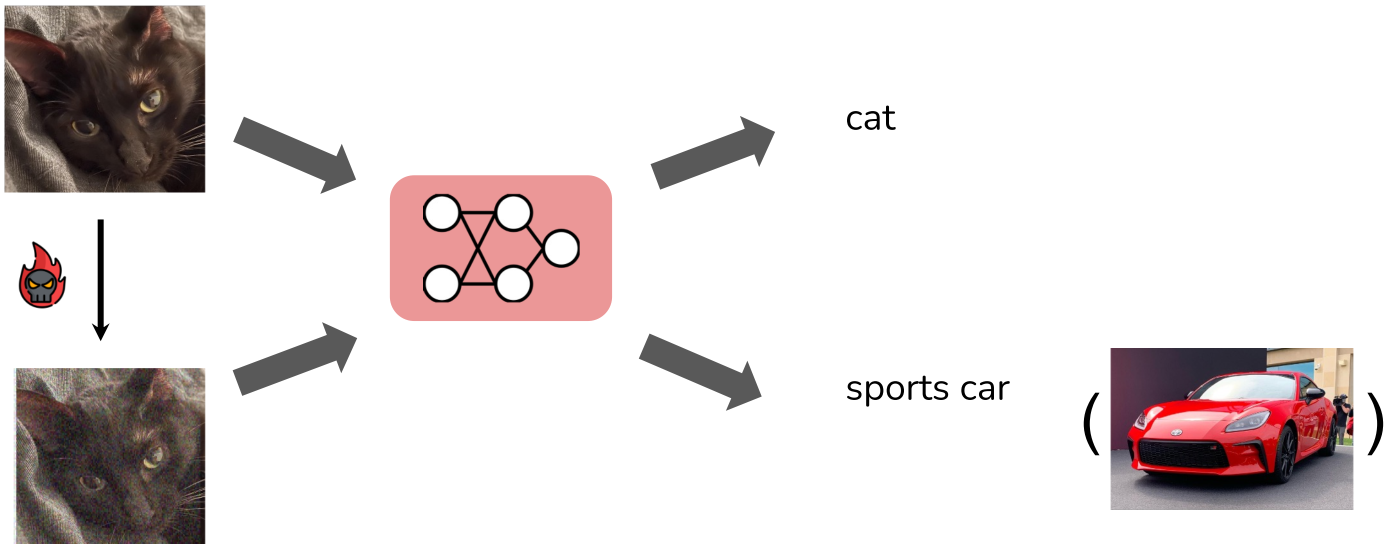
Improving Robust Generalization by Direct PAC-Bayesian Bound Minimization

Zifan Wang
CMU
zifan@cmu.edu

Nan Ding, Tomer Levinboim, Xi Chen, Radu Soricut
Google Research
dingnan@google.com

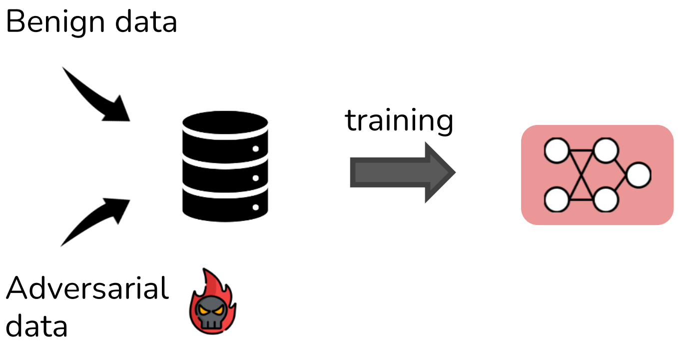
Poster Session: WED-PM-391

Background

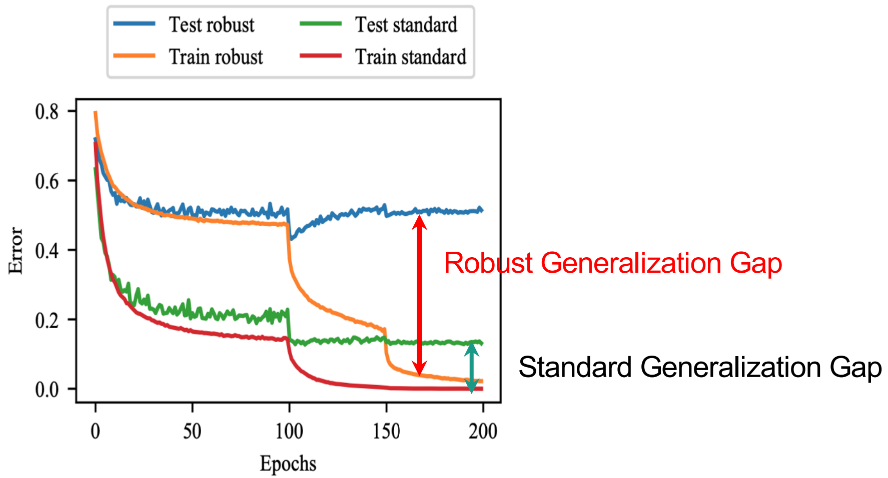


Deep models are often vulnerable to small **adversarial perturbations** that are unlikely to fool humans.

Background



(e.g. PGD [Madry et al. (2017)], TRADES [Zhang et al. (2019)] and etc.)



[Rice et al. 2019]

Contributions

Our goal is to improve robust generalization. We

characterize

Robust Test Loss → PAC-Bayesian Upper-bound $\xrightarrow{\textit{direct minimization}}$ Trace-of-Hessian (**TrH**) Regularization

show

Regularizing TrH of the top layer (for efficiency reason) $\xrightarrow{\text{empirically}}$ Decreased TrH of internal layers

report

Improved empirical robust accuracy on

CIFAR-10	CIFAR-100	ImageNet
a match to the best baseline	+2%	+2.7%

PAC-Bayesian Bound

Assumption

$\theta \sim \mathcal{P} = \mathcal{N}(\mathbf{0}, \mathrm{I}\sigma_0^2)$

A Prior distribution of Network weights
(**independent** on data and training algorithm)

$\theta \sim \mathcal{Q}$, and $\mathcal{Q}$ is product of univariate Gaussians $\mathcal{N}(\mu, \Sigma)$

Posterior distribution of Network weights
(**dependent** on data and training algorithm)

Applying PAC-Bayesian Bound to Test Robust Loss

With a probability $1 - \tau$, the following holds true:

$$\underbrace{\mathbb{E}_{\theta \sim \mathcal{Q}} R(\theta)}_{\text{test robust loss}} \leq \underbrace{\mathbb{E}_{\theta \sim \mathcal{Q}} \hat{R}(\theta)}_{\text{train robust loss}} + \frac{1}{\beta} KL(\mathcal{Q}||\mathcal{P}) + \underbrace{C(\tau, \beta, m)}_{\text{Quantity independent of } \mathcal{Q}}$$

Direct Minimization of PAC-Bayesian Bound

PAC-Bayesian Bound for Test Robust Loss

If $\mathcal{P} = \mathcal{N}(\mathbf{0}, \sigma_0^2)$, and $\mathcal{Q}$ is also a product of univariate Gaussian distributions, then

$$\mathbb{E}_{\theta\sim Q} R(\theta) \leq \mathbb{E}_{\theta\sim Q} \hat{R}(\theta) + \frac{1}{\beta} KL(Q||\mathcal{P}) + C(\tau, \beta, m)$$

test robust loss

train robust loss

Quantity independent of Q

By solving the minimization w.r.t Q (i.e. w.r.t. μ and Σ)

Minimized Bound

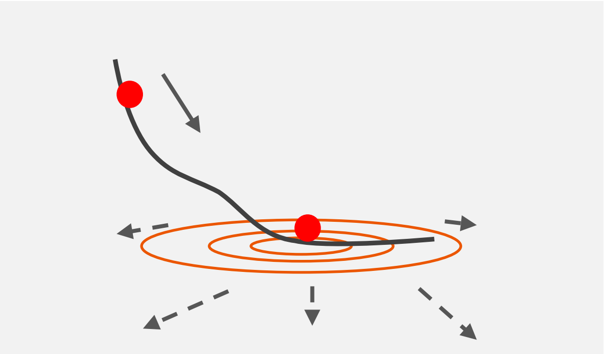
(Theorem 3)

$$\min_{Q} \mathbb{E}_{\theta\sim Q} R(\theta) \leq \min_{Q} \left\{ \mathbb{E}_{\theta\sim Q} \hat{R}(\theta) + \frac{1}{\beta} KL(Q||\mathcal{P}) \right\} + C(\tau, \beta, m)$$

$$= \min_{\mu} \{ \hat{R}(\mu) + \frac{||\mu||^2}{2\beta\sigma_0^2} + \frac{\sigma_0^2}{2} \mathbf{Tr}(\nabla_{\mu}^2(\hat{\mathbf{R}}(\mu))) \} + C(\tau, \beta, m) + O(\sigma_0^4)$$

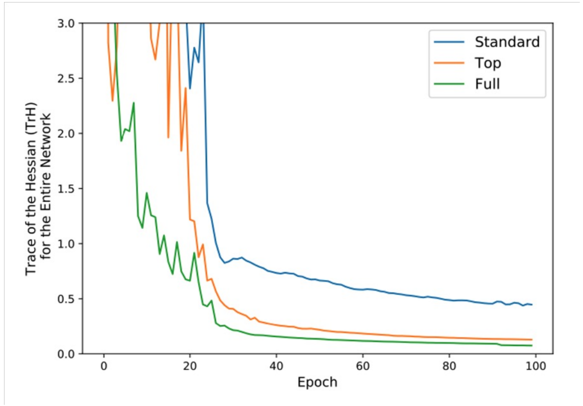
Effects of Trace-of-Hessian (TrH) Regularization

Flat Minimum



Trace of Hessian (TrH) is the sum of curvatures (under assumption of convexity) of the loss in all directions.

Inductive Impacts from top to the bottom

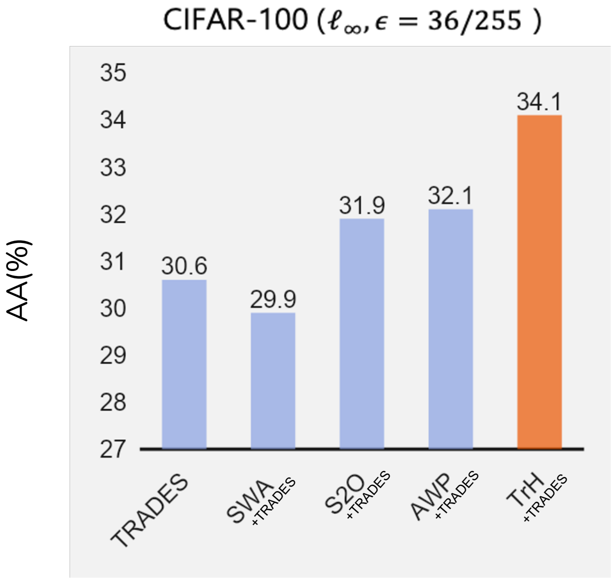


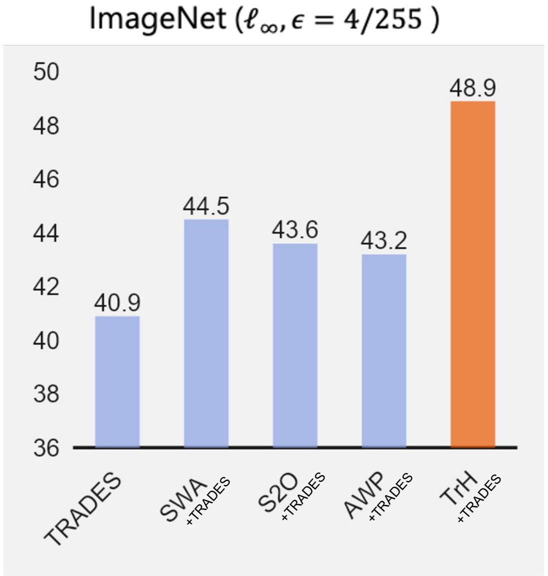
Regularizing TrH for top layer only → Decreased TrH for internal layers

(greater details in Theorem 4 and Example 1)

Results (see the full table in paper)

AutoAttack Accuracy(AA): percentage of test data that is both accurate and robust evaluated with AutoAttack[1].





All models are vision transformers pre-trained on ImageNet-21K.

[1][Croce & Hein (2022)] SWA [Izmailov et al. (2018)]   TRADES [Zhang et al. (2019)] S2O [Jin et al. (2022)] AWP [Wu et al.(2020)]

Read Our Paper

Poster Session

June 21, 2023

WED-PM-391