Generating Less Certain Adversarial Examples Improves Robust Generalization

Minxing Zhang, Michael Backes, Xiao Zhang

CISPA Helmholtz Center for Information Security

What Does Our Work Do?

- Preliminary
- Motivation
- Adversarial Certainty
- Our Method
- Empirical Evidence
- Conclusion

Preliminary

- Robust Overfitting [1]:

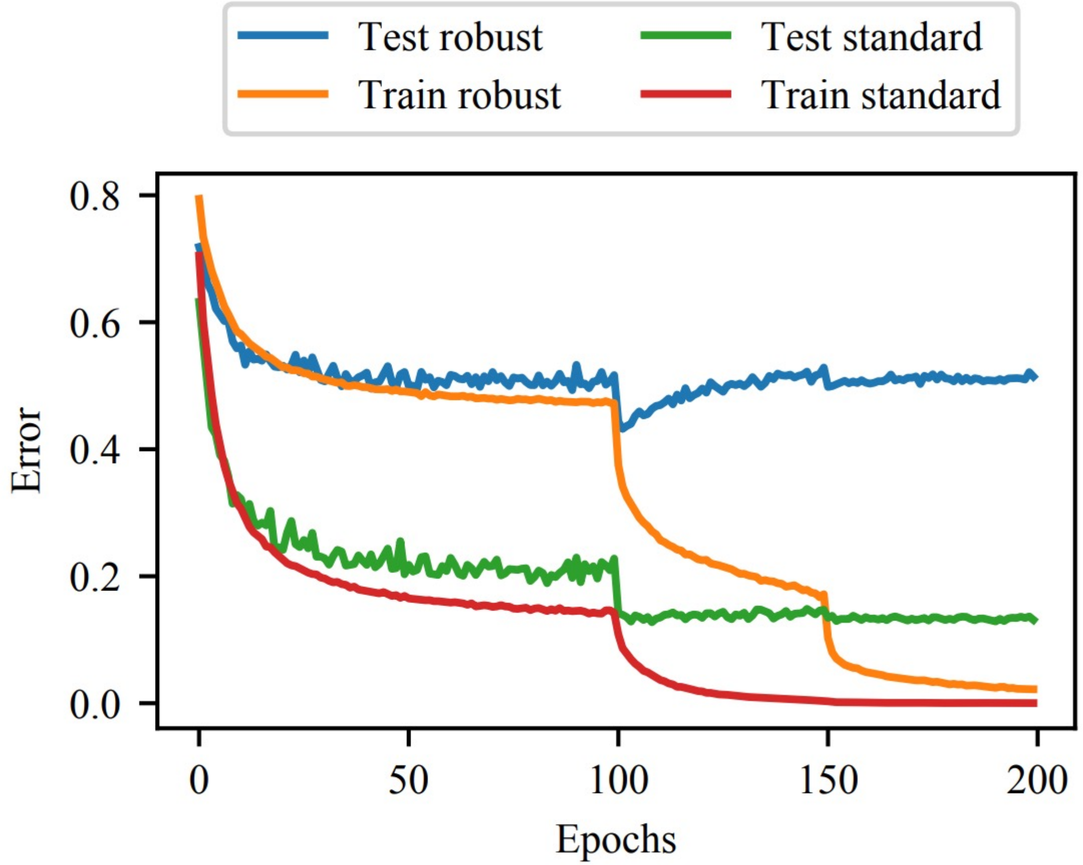
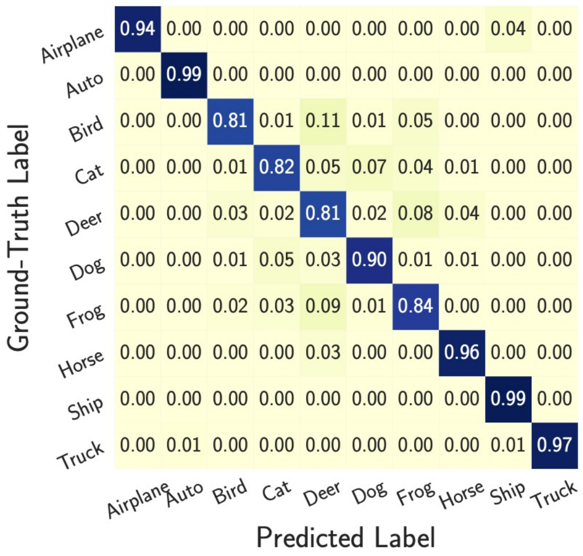


Image Source: [1]

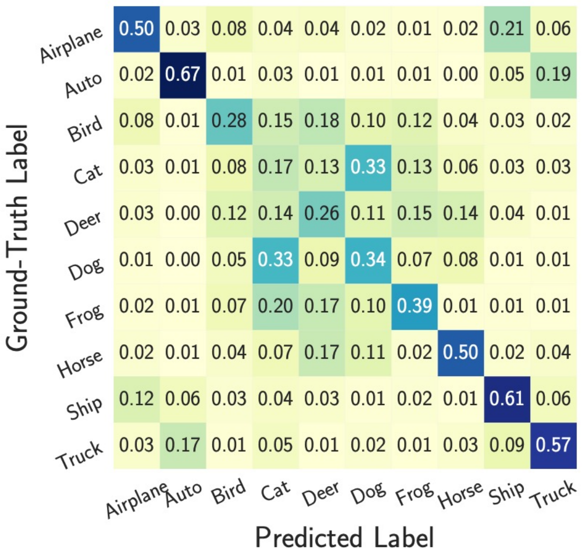
[1] ***Overfitting in adversarially robust deep learning*****, Rice et al.**

Motivation

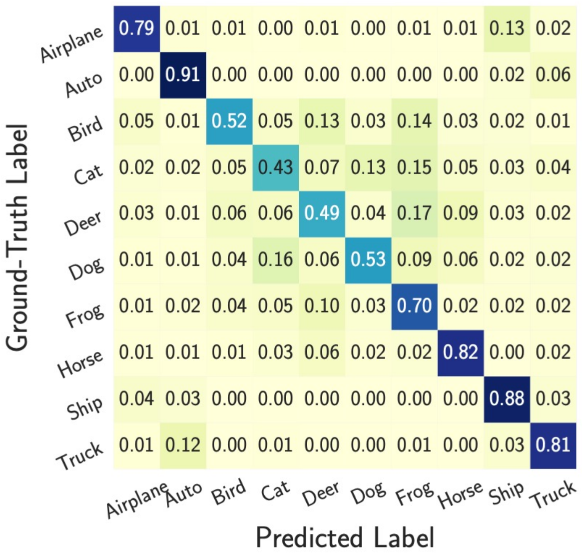
- Heatmap Visualization
  - Adversarially-perturbed CIFAR-10



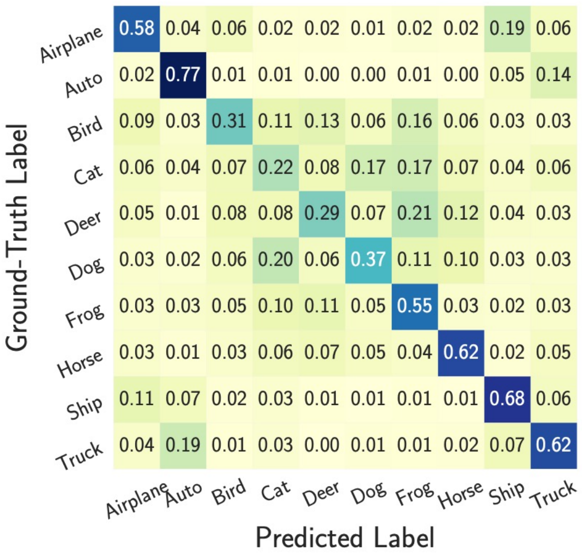
(a) Last Model (Train)



(b) Last Model (Test)

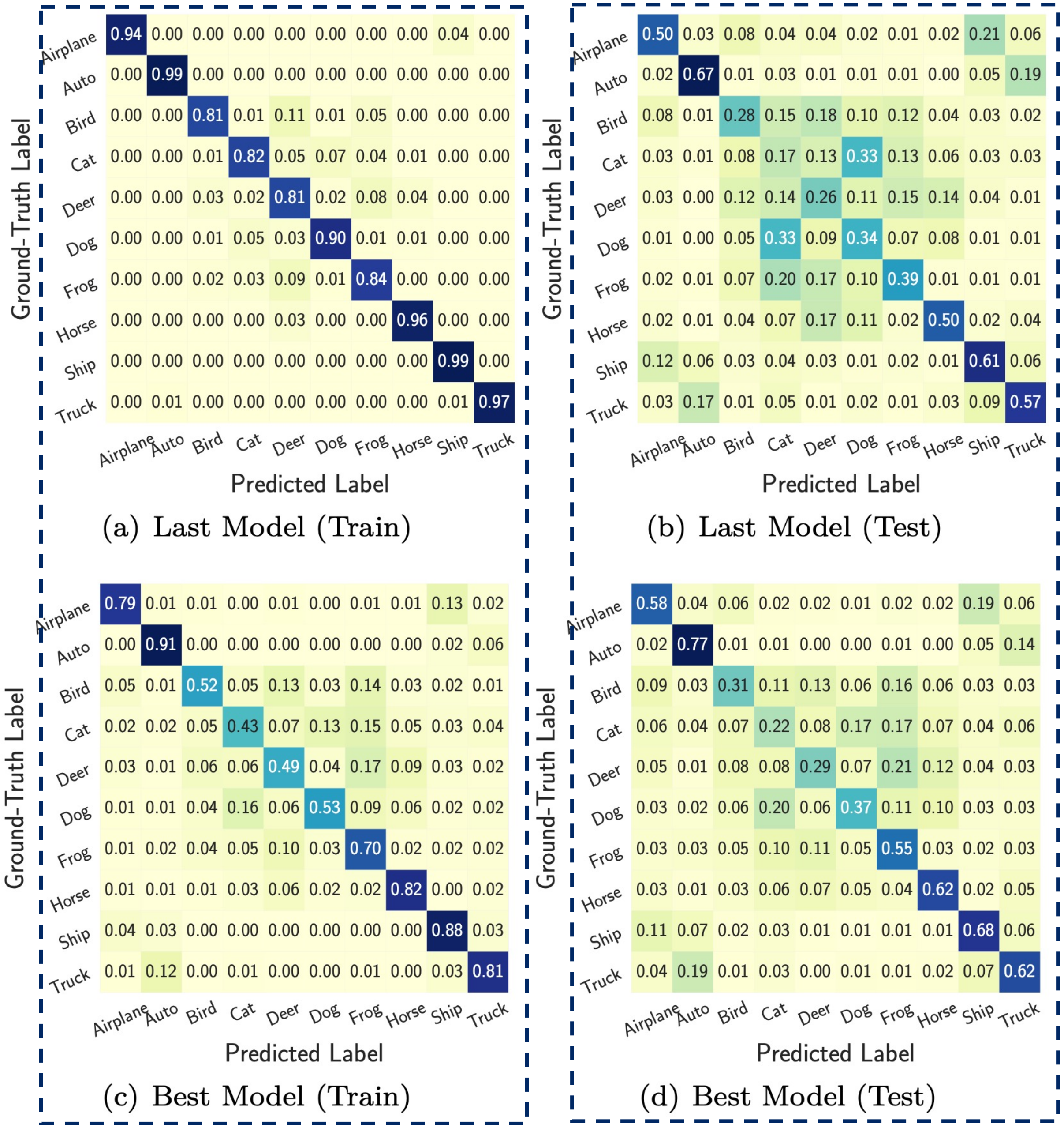


(c) Best Model (Train)



(d) Best Model (Test)

Motivation



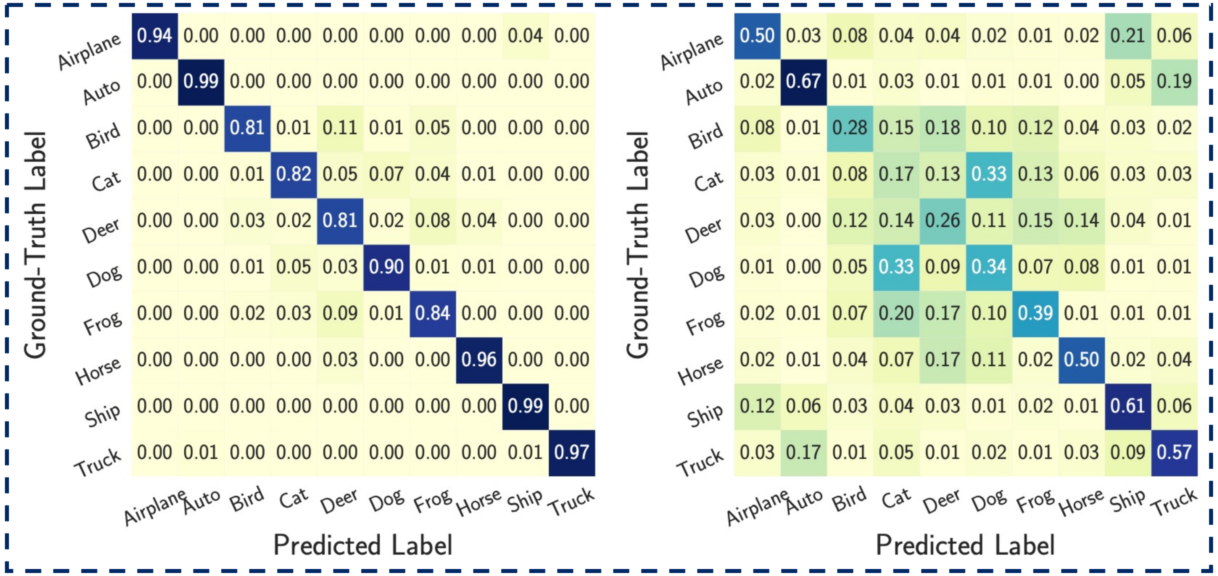
(a) Last Model (Train)

(b) Last Model (Test)

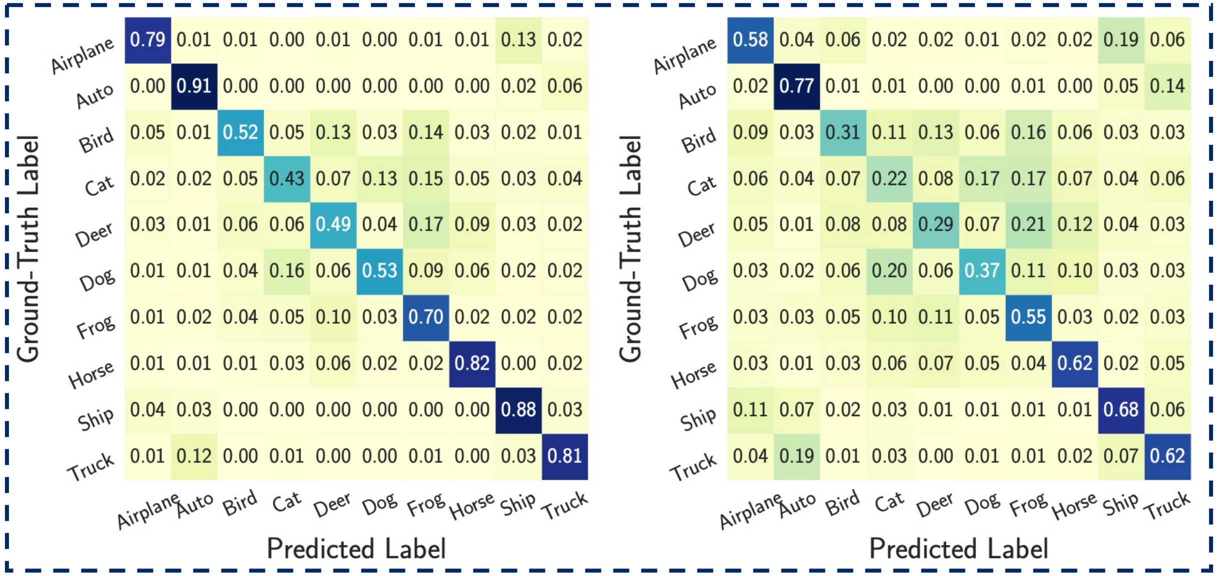
(c) Best Model (Train)

(d) Best Model (Test)

Motivation

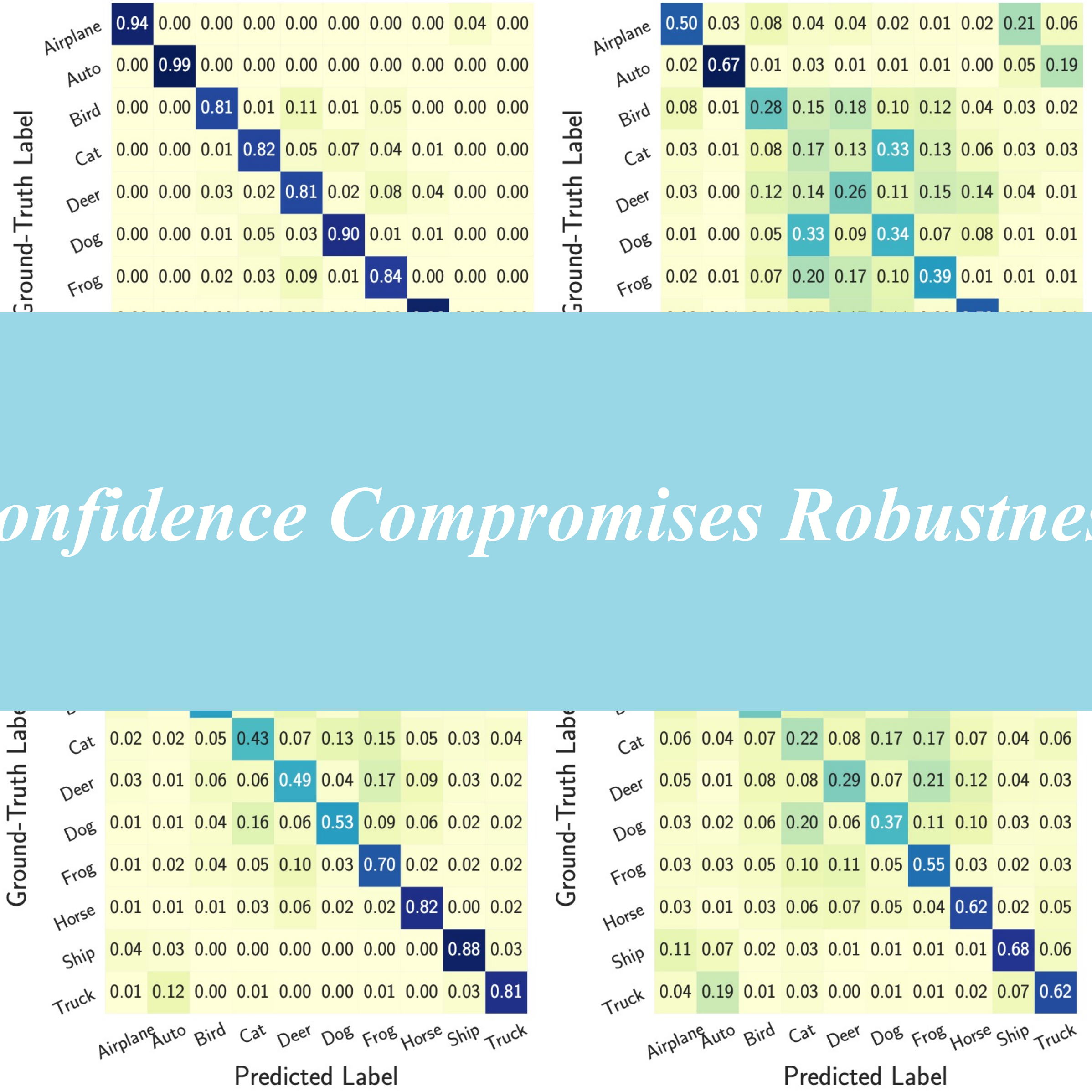


(a) Last Model (Train) (b) Last Model (Test)



(c) Best Model (Train) (d) Best Model (Test)

Motivation



(c) Best Model (Train) (d) Best Model (Test)

Overconfidence Compromises Robustness

Adversarial Certainty

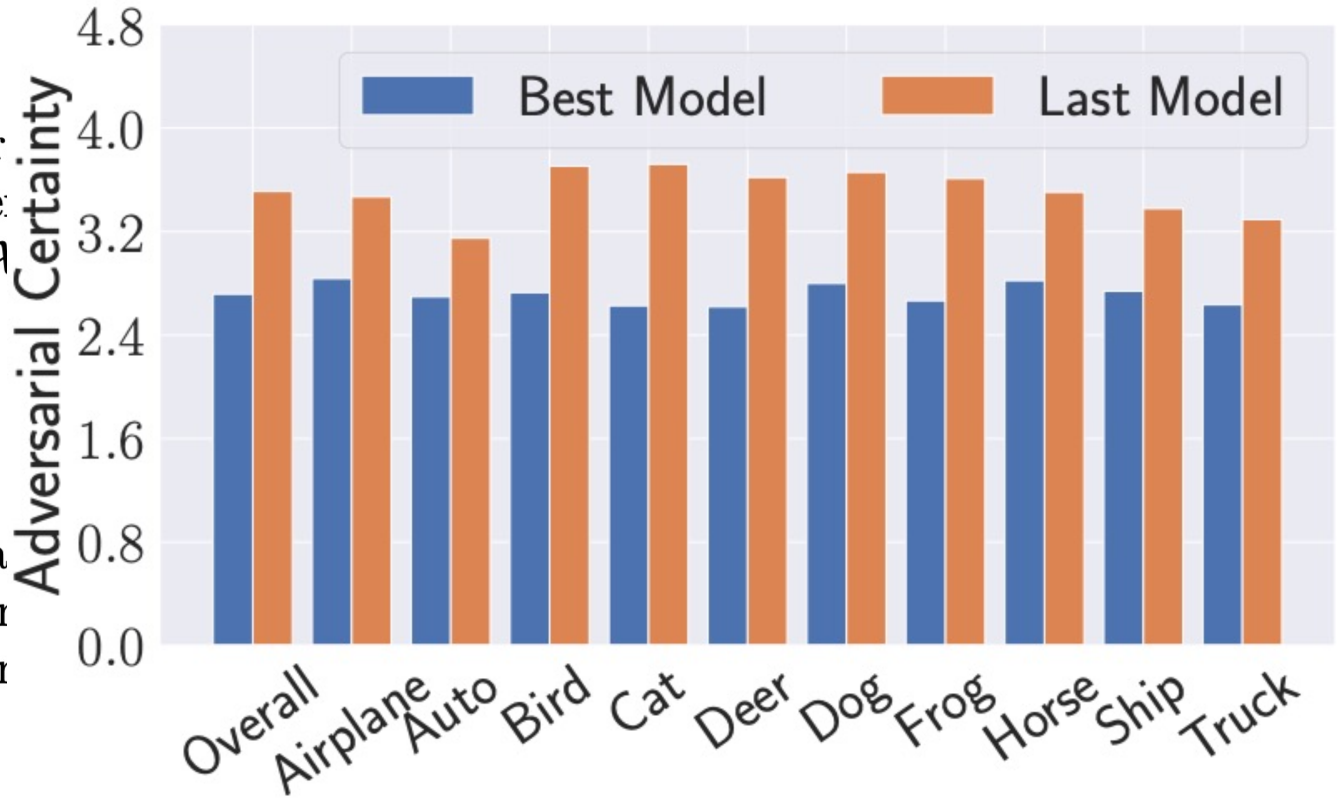
Definition 1 (Adversarial Certainty)**.** Let $\mathcal{X}$ be the input space and $\mathcal{Y} = \{1, 2, \ldots, m\}$ be the label space. Suppose μ is the underlying distribution and $\mathcal{S}$ is a set of sampled examples. Let $\epsilon \geq 0$, Δ be the perturbation metric. For any $f_\theta : \mathcal{X} \to \mathcal{Y}$, we define the *adversarial certainty* of f_θ as:

$$\mathrm{AC}_\epsilon(f_\theta; \hat{\mu}_\mathcal{S}, \mathcal{A}) = \frac{1}{|\mathcal{S}|} \sum_{(\boldsymbol{x},y)\in\mathcal{S}} \mathrm{Var}\big(F_\theta\big[\mathcal{A}(\boldsymbol{x}; y, f_\theta, \epsilon)\big]\big),$$

where $\mathcal{A}$ denotes an attack method such as PGD attacks for generating adversarial examples, $F_\theta : \mathcal{X} \to \mathbb{R}^m$ represents the mapping from the input space $\mathcal{X}$ to the logit layer of f_θ, and $\mathrm{Var}(\boldsymbol{u}) = \sum_{k\in[m]} (u_k - \overline{u})^2/m$, with u_k and $\overline{u}$ denoting the k-th element and mean of $\boldsymbol{u} \in \mathbb{R}^m$ respectively.

Adversarial Certainty

Definition 1 (Adver … $n\}$ be the label space.
Suppose μ is the unde … Δ be the perturbation
metric. For any $f_\theta : \mathcal{X}$ …

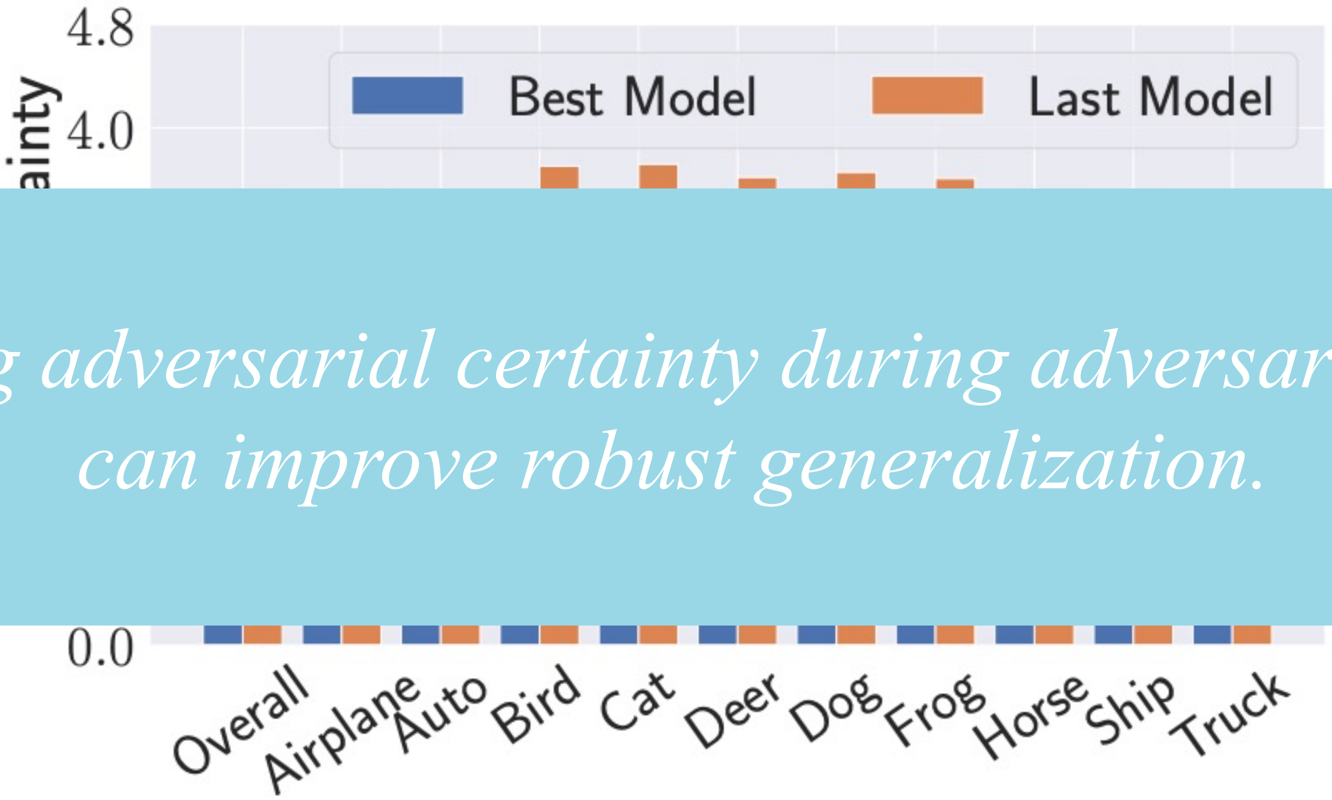


where $\mathcal{A}$ denotes an a … amples, $F_\theta : \mathcal{X} \rightarrow \mathbb{R}^m$
represents the mappi … $\sum_{k\in[m]} (u_k - \overline{u})^2/m$,
with u_k and $\overline{u}$ denoti …

Adversarial Certainty

Definition 1 (Adver ... $n\}$ be the label space.

with u_k and $\overline{u}$ denotin



Decreasing adversarial certainty during adversarial training can improve robust generalization.

Our Method

- **D**ecrease **A**dversarial **C**ertainty (DAC)
  - Find less certain adversarial examples that are used to improve robust generalization

$$\min_{\theta\in\Theta} \frac{1}{|\mathcal{S}_{tr}|} \sum_{(\boldsymbol{x},y)\in\mathcal{S}_{tr}} \max_{\boldsymbol{x}'\in\mathcal{B}_\epsilon(\boldsymbol{x})} L\big(f_{\theta'}, \boldsymbol{x}', y\big), \text{ where } \theta' = \operatorname*{argmin}_{\theta'\in\mathcal{C}(\theta)} \mathrm{AC}_\epsilon(f_\theta; \mathcal{S}_{tr}, \mathcal{A}).$$

  - Two-step Optimization

$$\theta_{t+0.5} = \theta_t - \lambda \cdot \nabla_\theta \mathrm{AC}_\epsilon(f_\theta; \mathcal{S}_{tr}, \mathcal{A})\Big|_{\theta=\theta_t},$$
$$\theta_{t+1} = \theta_{t+0.5} - \gamma \cdot \nabla_\theta L_{\mathrm{rob}}(f_\theta; \mathcal{S}_{tr}, \mathcal{A})\Big|_{\theta=\theta_{t+0.5}},$$

Empirical Evidence

Architecture	Method	Clean	PGD-20	PGD-100	CW_∞	AutoAttack
PRN18	AT	82.88 (82.68)	41.51 (49.23)	40.96 (48.92)	41.61 (48.07)	39.66 (45.71)
	+ DAC	**84.64 (83.55)**	**45.55 (52.20)**	**44.94 (51.87)**	**44.55 (50.05)**	**42.78 (48.20)**
	TRADES	82.10 (81.33)	47.44 (51.65)	46.95 (51.42)	46.64 (49.18)	44.99 (48.06)
	+ DAC	**83.18 (82.80)**	**49.32 (52.90)**	**48.81 (52.67)**	**48.30 (50.11)**	**46.40 (48.96)**
	MART	80.85 (78.27)	50.23 (52.28)	49.71 (52.13)	46.88 (47.83)	44.68 (46.01)
	+ DAC	**81.12 (79.37)**	**52.38 (53.25)**	**52.04 (53.14)**	**48.97 (49.25)**	**47.24 (47.69)**
WRN34	AT	86.47 (**85.86**)	47.25 (55.31)	46.73 (55.00)	47.85 (54.04)	45.84 (51.94)
	+ DAC	**86.48** (85.10)	**52.02 (57.93)**	**51.69 (57.68)**	**51.51 (54.98)**	**49.75 (53.33)**
	TRADES	83.37 (81.40)	51.51 (58.78)	51.28 (58.72)	49.26 (53.33)	47.74 (52.63)
	+ DAC	**85.04 (84.55)**	**58.97 (60.96)**	**58.97 (60.81)**	**52.79 (55.00)**	**51.80 (53.99)**
	MART	83.11 (**83.30**)	48.93 (58.13)	48.31 (57.75)	46.32 (52.22)	44.89 (50.31)
	+ DAC	**84.69** (80.09)	**52.00 (59.31)**	**51.32 (59.26)**	**49.50 (53.02)**	**47.65 (51.48)**

Conclusion

- Notion of adversarial certainty
- Importance of generating less certain adversarial examples for robust generalization
- Better understanding of robust generalization

Thanks

Minxing Zhang – Ph.D. Candidate

CISPA Helmholtz Center for Information Security