UNIVERSITY OF COPENHAGEN

Is Adversarial Training with Condensed Dataset Effective?

MIA Talk 28/02
Tong Chen

Dataset Condensation

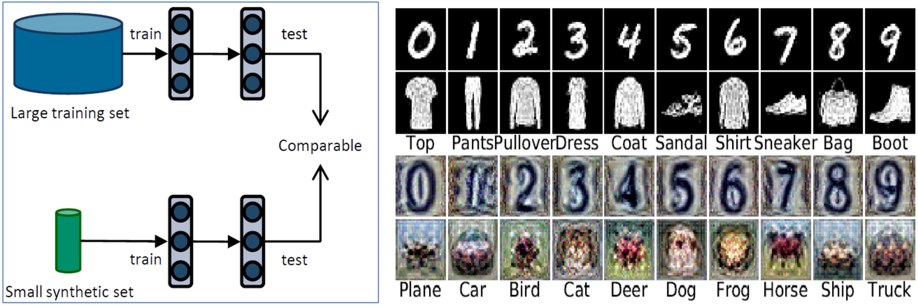


Figure: Dataset Condensation

Notations

- Distribution $\mathcal{D}$, sampling $\mathcal{S}_n \overset{i.i.d}{\sim} \mathcal{D}^n$;
- Hypothesis space $\mathcal{H}$, loss function l;
- Generalization loss:

$$L(f) = \mathbb{E}_{(X,Y)\sim\mathcal{D}}[l(f(x), y)], \; f^* = \arg\min_{f\in\mathcal{H}} L(f);$$

- Empirical loss:

$$\hat{L}(f, \mathcal{S}_n) = \frac{1}{n}\sum_{i=1}^{n} l(f(x_i), y_i), \; f^*_{\mathcal{S}_n} = \arg\min_{f\in\mathcal{H}} \hat{L}(f, \mathcal{S}_n).$$

Formal Statement

- Basic results:

$$L(f^*_{\mathcal{S}_n}) \xrightarrow[\geq]{\mathbb{P}} L(f^*);$$

- Dataset condensation:

$$\mathcal{T}_n = \arg\min_{\mathcal{S}_n} L(f^*_{\mathcal{S}_n}),\ \ L(f^*_{\mathcal{T}_n}) \xrightarrow[\geq]{\mathbb{P}} L(f^*).$$

$$\boxed{\mathcal{T}_n \overset{i.i.d.}{\sim} \mathcal{D}^n\ ?}$$

Generalization is NOT Enough

$$\boxed{\mathcal{S}_n \overset{i.i.d.}{\sim} \mathcal{D}^n,\ \mathcal{T}_n = \arg\min_{\mathcal{S}_n} L(f^*_{\mathcal{S}_n}) \overset{i.i.d.}{\sim} \mathcal{D}^n\ ?}$$

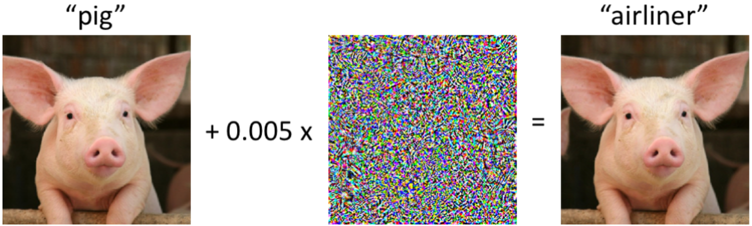
- Generalization is guaranteed:

$$L(f^*_{\mathcal{S}_n}) \xrightarrow{\mathbb{P}} L(f^*) \xleftarrow{\mathbb{P}} L(f^*_{\mathcal{T}_n});$$

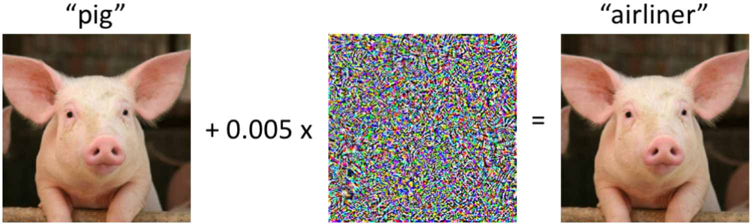
- Robustness is NOT guaranteed:

$$L^{adv}(f^*_{\mathcal{S}_n}, \varepsilon) \xrightarrow{\mathbb{P}} L^{adv}(f^*, \varepsilon) \overset{\mathbb{P}}{\not\longleftarrow} L^{adv}(f^*_{\mathcal{T}_n}, \varepsilon).$$

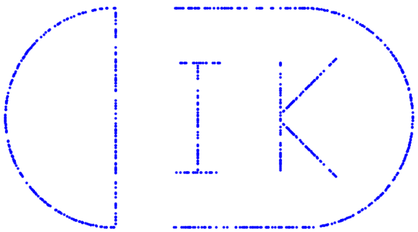
Adversarial Example



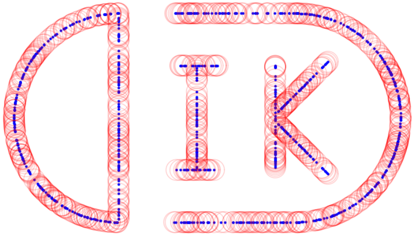
Adversarial Example



Standard v.s. Robust Classification

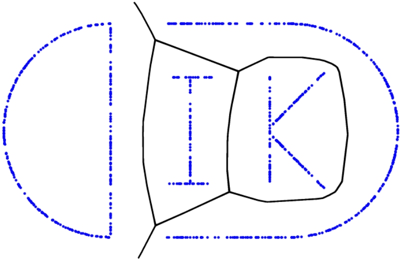


(a) Standard classification

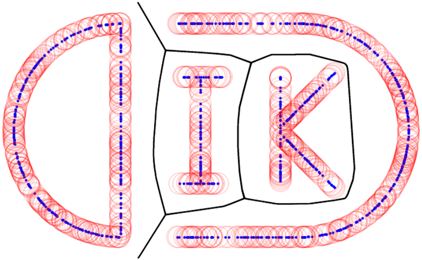


(b) Robust classification

Standard v.s. Adversarial Training



(a) Standard training



(b) Adversarial training

Robustness-Aware Sampling

- $\mathcal{T}_n$ = finite covering with n balls of radius η_n;
- Generalization guarantee:

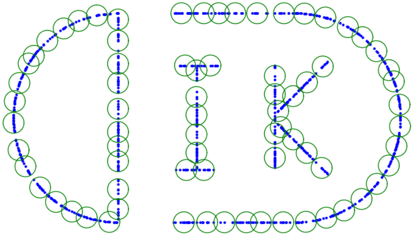
$$L(f^*_{\mathcal{S}_n}) \xrightarrow{\mathbb{P}} L(f^*) \xleftarrow{\mathbb{P}} L(f^*_{\mathcal{T}_n});$$

- Robustness guarantee:

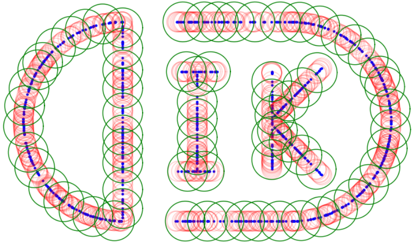
$$L^{adv}(f^*_{\mathcal{S}_n}, \varepsilon) \xrightarrow{\mathbb{P}} L^{adv}(f^*, \varepsilon) \xleftarrow{\mathbb{P}} L^{adv}(f^*_{\mathcal{T}_n}, \varepsilon + \eta_n),$$

with $\lim_{n\to\infty} \eta_n = 0$.

Minimal Finite Covering

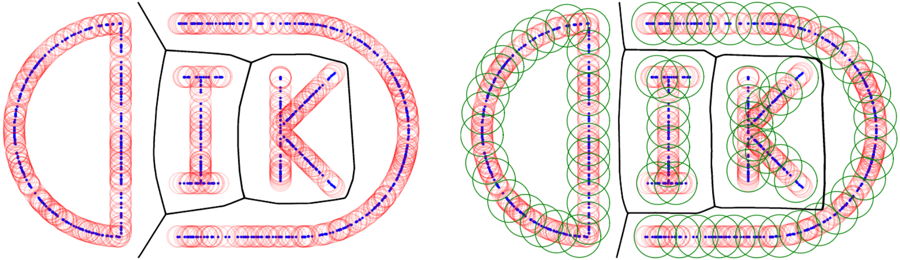


(a) Finite covering with radius η



(b) Finite covering with radius $\eta + \varepsilon$

Adversarial Training with Finite Covering



(a) Adversarial training (b) Generalized adversarial training

Thank you!

More technical details and experiments:

https://arxiv.org/abs/2402.05675