An effective manifold-based optimization method for distributionally robust classification

Jiawei Huang[1,2], Hu Ding[1]

1, School of Computer Science and Technology, University of Science and Technology of China
2, Department of Computer Science, City University of Hong Kong

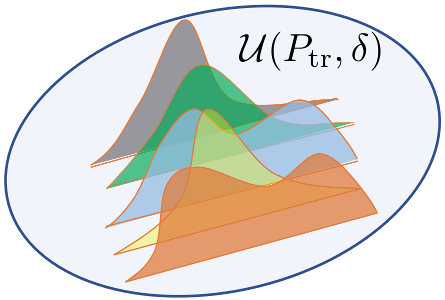
01 Motivation: Distributional robustness

DR under Data Distribution Shift:

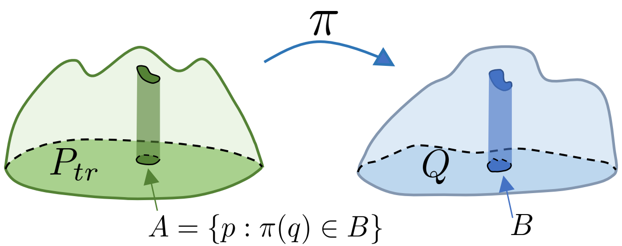
- Background: The empirical distribution of the training data may differ from the real data distribution; Introduce an ***Uncertainty Set*** that includes all possible true distributions.

- Wasserstein DR objective:
$$\min_{\theta} \sup_{Q \in \mathcal{U}(P_{\mathrm{tr}}, \delta)} \mathbb{E}_Q[\ell(\theta, x)]$$

$$\mathcal{U}(P_{\mathrm{tr}}, \delta) = \{Q \in \mathcal{P}(\mathbb{R}^d) | \mathcal{W}(Q, P_{\mathrm{tr}}) \le \delta\}$$



$$\mathcal{W}(Q, P_{\mathrm{tr}}) = \left(\inf_{\pi \in \Pi(Q, P_{\mathrm{tr}})} \int_{\mathcal{X} \times \mathcal{X}} \mathtt{d}^2(p, q) \mathrm{d}\pi\,(q, p) \right)^{\frac{1}{2}}$$

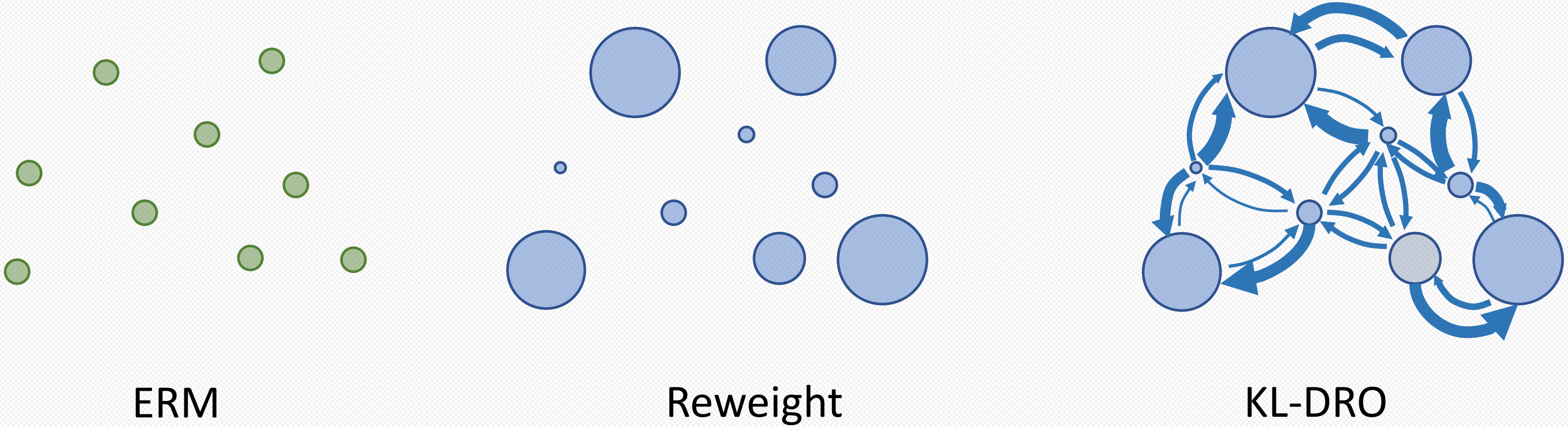


01 Motivation: Distributional robustness

Other types of uncertainty sets: $\Delta(Q, P_{\mathrm{tr}})$: KL-divergence, Sinkhorn distance et al.

$$\mathcal{L}^{\delta}_{DR}(\theta, P_{\mathrm{tr}}) = \sup_{Q \in \mathcal{U}(P_{\mathrm{tr}}, \delta)} \{\mathbb{E}_Q[\ell(\theta, q)]\},$$

$$\text{where} \quad \mathcal{U}(P_{\mathrm{tr}}, \delta) = \{Q \in \mathcal{P}(\mathcal{X}) | \Delta(Q, P_{\mathrm{tr}}) \le \delta)\}.$$



Kullback–Leibler divergence: $D_{\mathrm{KL}}(P_{\mathrm{tr}} \| Q) = \mathbb{E}_{x \sim P_{\mathrm{tr}}}\left[\log \frac{\mathrm{d}P_{\mathrm{tr}}(x)}{\mathrm{d}Q(x)}\right]$.

Requiring absolute continuity $P_{\mathrm{tr}} \ll Q$; otherwise, $D_{\mathrm{KL}}(P_{\mathrm{tr}} | Q) = \infty$.

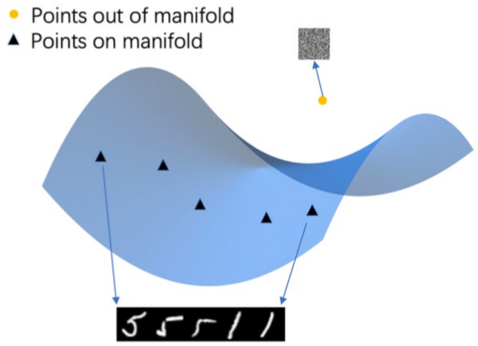
In comparision, Wasserstein DRO (WDRO) can capture the continuous variations.

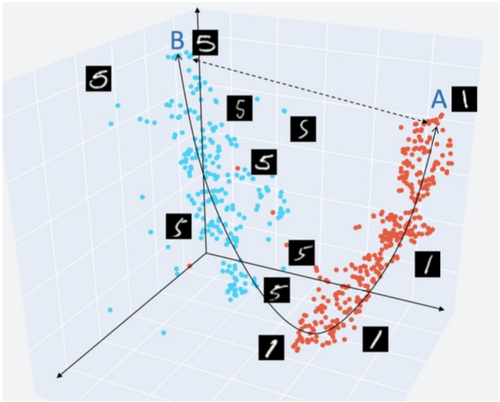
01 Motivation: Distributional robustness

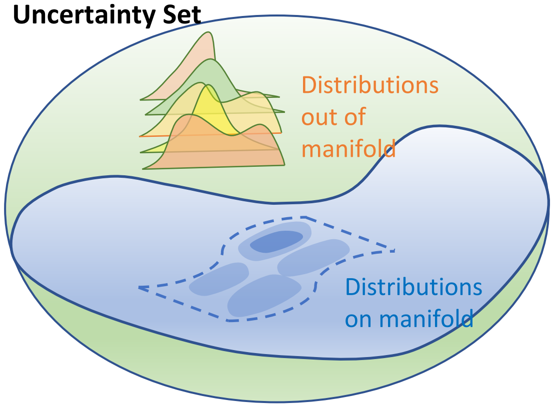
Motivation： How to choose the uncertainty set $\mathcal{U}(P_{\mathrm{tr}}, \delta)$

$$\mathcal{U}(P_{\mathrm{tr}}, \delta) = \{Q \in \mathcal{P}(\mathbb{R}^d) | \mathcal{W}(Q, P_{\mathrm{tr}}) \leq \delta)\}$$

$$\mathcal{U}_{gw}(P_{\mathrm{tr}}, \delta) = \{Q \in \mathcal{P}(\mathcal{M}) | \mathcal{GW}(Q, P_{\mathrm{tr}}) \leq \delta)\}$$







How to capture the geometric structure of the data?

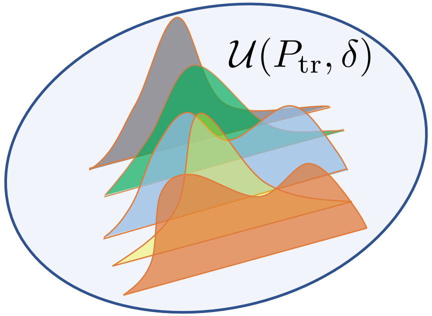
Can neural nets also be used to extract the tangent space of a data manifold?

02 Formulation: Wasserstein Distributional Robust

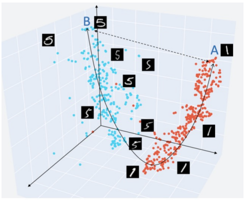
Geodesic distance Wasserstein Uncertainty Set:

- Goal： Optimizing within the *uncertainty set*, which is supposed to incorporate all possible distributions.
- Manifold WDRO： $$\min_{\theta} \sup_{Q \in \mathcal{U}_{gw}(P_{\mathrm{tr}},\delta)} \mathbb{E}_Q[\ell(\theta, x)]$$

$$\mathcal{U}_{gw}(P_{\mathrm{tr}}, \delta) = \{Q \in \mathcal{P}(\mathcal{M}) | \mathcal{GW}(Q, P_{\mathrm{tr}}) \leq \delta\}$$



$$\mathcal{GW}(Q, P_{\mathrm{tr}}) = \left(\inf_{\pi \in \Pi(Q, P_{\mathrm{tr}})} \int_{\mathcal{M} \times \mathcal{M}} \mathrm{d}_{\mathbf{g}}^2(p, q) \mathrm{d}\pi\,(q, p) \right)^{\frac{1}{2}}$$



02 Formulation: Strong Duality

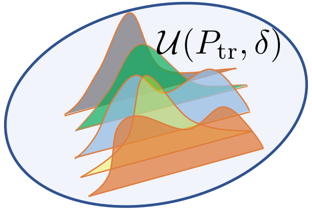
Strong Duality of Manifold-WDRO

To solve the primal problem, we adopt the strongly duality property proposed in [Gao et al., 23] to obtain the dual form.

Primal

$$\mathcal{L}^{\delta}_{DR}(\theta, P_{\text{tr}}) = \sup_{Q \in \mathcal{U}_{gw}(P_{\text{tr}},\delta)} \{\mathbb{E}_Q[\ell(\theta, q)]\},$$

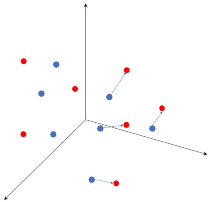
$$\text{where} \quad \mathcal{U}_{gw}(P_{\text{tr}}, \delta) := \{Q \in \mathcal{P}(\mathcal{M}) | \mathcal{GW}(Q, P_{\text{tr}}) \leq \delta)\}.$$



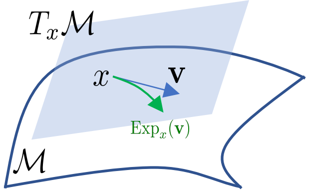
Dual

$$\mathcal{L}^{\delta}_{DR}(\theta, P_{\text{tr}}) = \min_{\nu \geq 0} \left\{\nu\delta^2 + \mathbb{E}_{P_{\text{tr}}} \ell_s(\theta, p_i, \nu)\right\},$$

$$\text{where} \quad \ell_s(\theta, p_i, \nu) := \sup_{q \in \mathcal{M}} \left[\ell(\theta, q) - \nu \mathrm{d}_g^2(q, p_i)\right].$$



Solve by Riemann gradient ascent : $\nabla^{\mathcal{M}} f(x) = \text{Proj}_x \left(\nabla f(x)\right)$



02 Formulation: Strong Duality

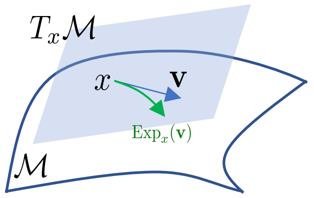
The surrogate loss is a geodesically $(\nu - \beta)$-strongly concave problem.

Dual

$$\mathcal{L}_{DR}^{\delta}(\theta, P_{\text{tr}}) = \min_{\nu \geq 0} \left\{ \nu\delta^2 + \mathbb{E}_{P_{\text{tr}}} \ell_s(\theta, p_i, \nu) \right\},$$

$$\text{where} \quad \ell_s(\theta, p_i, \nu) := \sup_{q \in \mathcal{M}} \left[\ell(\theta, q) - \nu \mathsf{d}_g^2(q, p_i) \right].$$

Solve by Riemann gradient ascent : $\nabla^{\mathcal{M}} f(x) = \operatorname{Proj}_x \left(\nabla f(x) \right)$



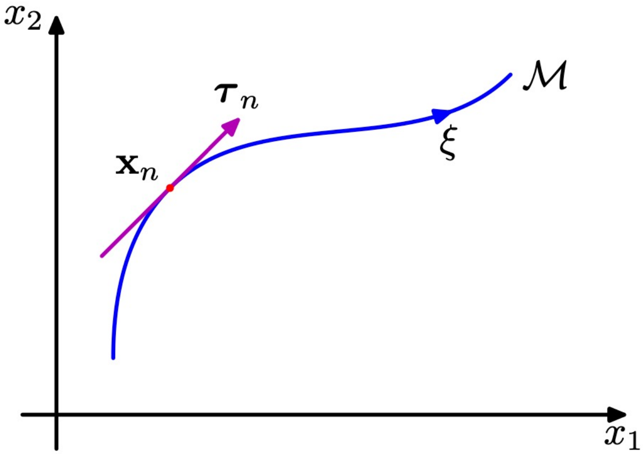
$\operatorname{Proj}_{\mathbf{x}}(\cdot)$: projection of the Euclidean gradient $\nabla f(x)$ onto the tangent space $\mathrm{T}_{\mathcal{M}}(\mathbf{x})$.

We **lack the analytical formula for the tangent** $\mathrm{T}_{\mathcal{M}}(\mathbf{x})$.

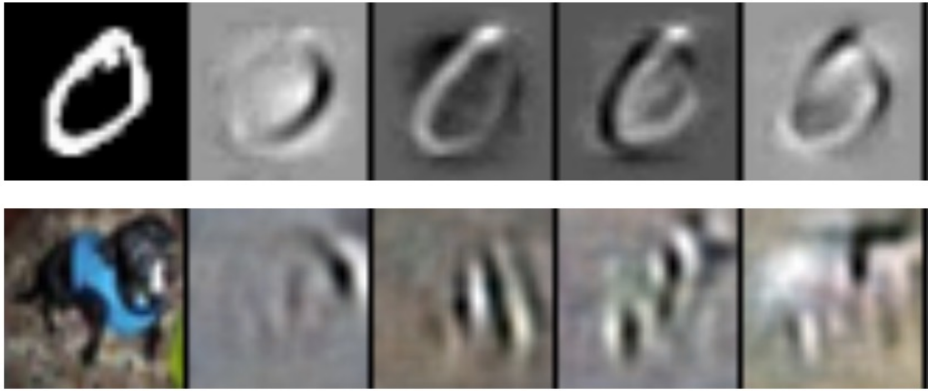
03 Method：Overall

Can neural nets also be used to extract the tangent space of a data manifold?

Previous works:



[PRML 07, Bishop]



Manifold tangent [Rifai,2011]

03 Method：Overall

Model: We have introduced our formulation (and its dual)

Algorithm: High-level ideas

1, Manifold-guided game

Sensitive to the data variation along the manifold; insensitive to others

Approximate (part of) the tangent

2, Compute the Surrogate Loss

A solution of a geodesically strongly concave optimization problem

Need to approximate the Geodesic distance

3, Optimizing over the surrogate loss (according the strongly duality)

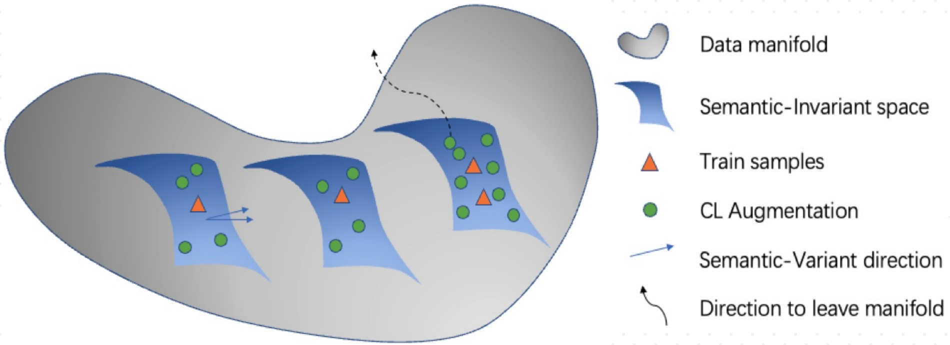
Robustness guarantee

03 Method：Manifold-guided Game

Manifold-guided DRO

- Manifold-guided game：Combine the Jacobian regularization and the contrastive learning loss

$$\mathcal{L}^*(\theta) = \mathcal{L}_{\mathrm{CL}} + \lambda_1 \mathbb{E}_{x \in P_{\mathrm{tr}}} \|J_{\mathbf{g}}(\boldsymbol{x})\|_F^2,$$



  - Jacobian Regularization：

$$J_g(\boldsymbol{x}) := \frac{\partial g(\theta, \boldsymbol{x})}{\partial \boldsymbol{x}}$$

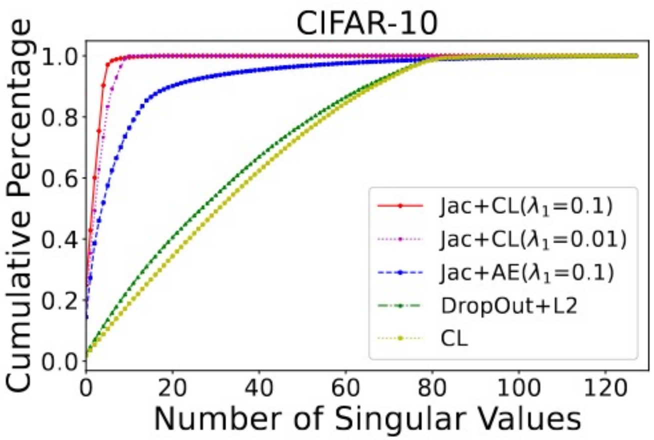
  - Contrastive Learning：

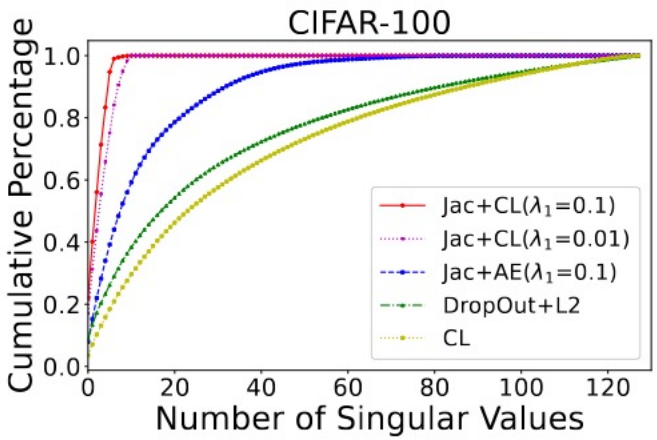
$$\mathcal{L}_{\mathrm{CL}} := -\mathop{\mathbb{E}}_{i \in [n]} \left[\ell_{\mathrm{cl}}(\boldsymbol{x}_i)\right], \text{ where } \ell_{\mathrm{cl}}(\boldsymbol{x}_i) = -\log \frac{\exp\left(\operatorname{sim}\left(\mathbf{g}(\boldsymbol{x}_i'), \mathbf{g}(\boldsymbol{x}_i'')\right)/\tau\right)}{\sum_{\boldsymbol{x} \in P_{\mathrm{tr}}' \cup P_{\mathrm{tr}}'' \setminus \{\boldsymbol{x}_i', \boldsymbol{x}_i''\}} \exp\left(\operatorname{sim}\left(\mathbf{g}(\boldsymbol{x}_i'), \mathbf{g}(\boldsymbol{x})\right)/\tau\right)}$$

Intuitively, the InfoNCE loss tends to bring $\mathbf{g}(\boldsymbol{x}_i')$ and $\mathbf{g}(\boldsymbol{x}_i'')$ to be closer, and meanwhile repulse $\mathbf{g}(\boldsymbol{x}_i')$ and $\mathbf{g}(\boldsymbol{x})$.

03 Method：Approximate the tangent

Random SVD for low-rank matrices





The remaining ***primary singular vectors*** **aligns with** directions of semantic variation within the data manifold's ***tangent***.

03 Method：Theoretical Results

1. Approximating the geodesic distance by the accumulated step size:

$$\mathtt{pt}^t \le \mathtt{pt}^\infty \le \sqrt{\kappa}(\sqrt{\kappa} + \sqrt{\kappa - 1})^2 \mathtt{d}_{\mathbf{g}}(q^0, q^*), \text{ where } \kappa = \tfrac{\nu+\beta}{\nu-\beta}.$$

2. Approximating the surrogate loss

under the approximation of the geodesic distance:

Theorem 1 *Suppose Assumption 1 and 2 hold. We select an $\nu > \beta$ in the surrogate loss (Eq.(7)), and define κ as in Lemma 1. Using the accumulated step size to approximate the geodesic distance, let $\hat{\theta}$ be the optimal solution of the dual formulation Eq.(6) under this approximation. Then $\mathbb{E}_{P_{\mathrm{tr}}} \ell_s(\hat{\theta}, p_i, \nu) \le c'' \times \min_\theta \mathbb{E}_{P_{\mathrm{tr}}} \ell_s(\theta, p_i, \nu)$, where $c'' = \kappa^2(\sqrt{\kappa} + \sqrt{\kappa - 1})^4$.*

04 Results

Empirical results

Illustration by t-SNE：

(a) (b) (c)

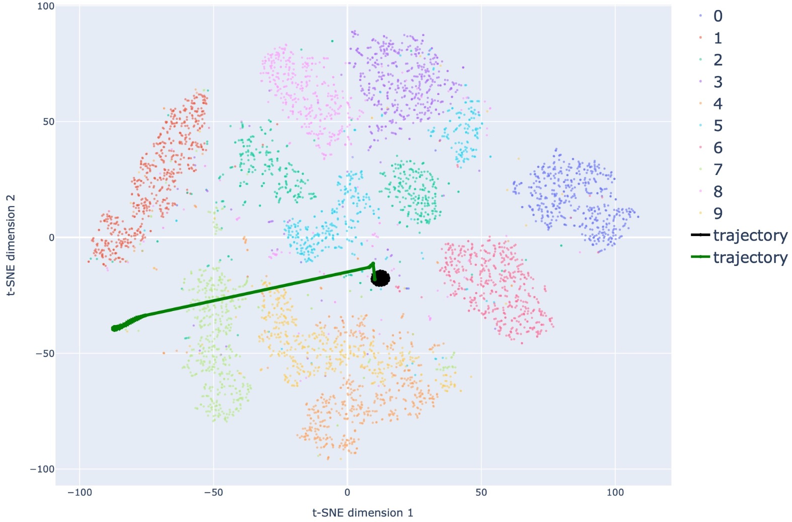
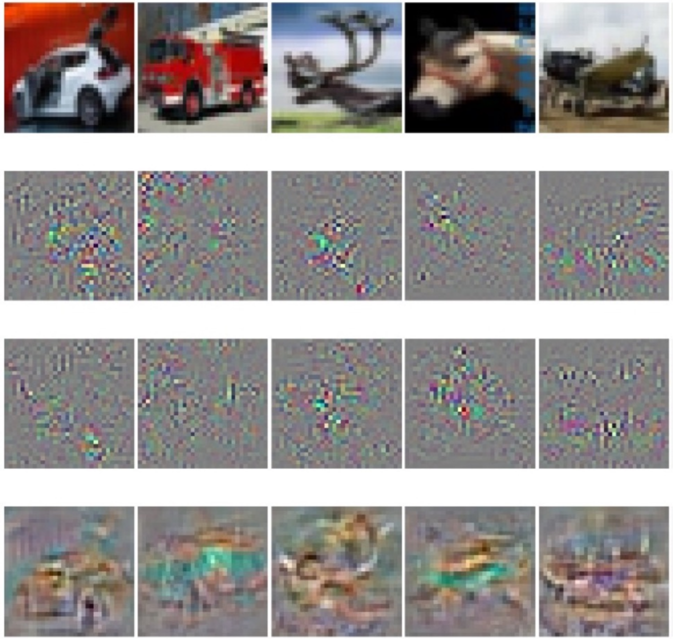


Illustration of the tangent：

(i)
(ii)
(iii)
(iv)



Thanks

Bibliography

[Sinha,2019] Certifying Some Distributional Robustness with Principled Adversarial Training
[Bui,2022] A Unified Wasserstein Distributional Robustness Framework for Adversarial Training
[Liu,2024] Distributionally Robust Optimization with Data Geometry
[HaoChen,2021] Provable guarantees for self-supervised deep learning with spectral contrastive loss
[Tan,2024] Contrastive Learning Is Spectral Clustering on Similarity Graph
[Assel,2022] A Probabilistic Graph Coupling View of Dimension Reduction
[Hu,2023] Your Contrastive Learning Is Secretly Doing Stochastic Neighbor Embedding