Dataset Condensation

Theory, Practice, and Beyond

ML Section Talk

Tong Chen

Introduction

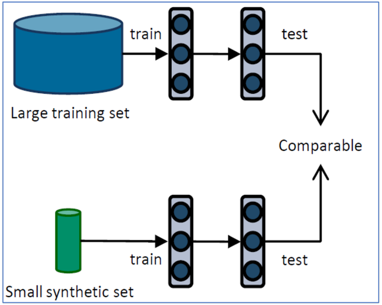


Figure: Dataset Condensation

Problem setting

Given $\mathcal{T} = \{(x_i, y_i)\}_{i=1}^{N} \subseteq \mathcal{X} \times \mathcal{Y}$, where $(x_i, y_i) \overset{i.i.d.}{\sim} \mathbb{P}$.

Find $\mathcal{S} = \{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^{M} \subseteq \mathcal{X} \times \mathcal{Y}$, where $(\tilde{x}_i, \tilde{y}_i) \overset{i.i.d.}{\sim} \mathbb{Q}$.

$$\boxed{\mathbb{P} \approx \mathbb{Q}?}$$

Notations

Hypothesis space $\mathcal{H} = \{h_\theta : \mathcal{X} \to \mathcal{Y}\}$.

Loss $l : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$.

Average loss $L_\mathbb{P}(h_\theta) = \mathbb{E}_{(X,Y)\sim\mathbb{P}}[l(h_\theta(X), Y)]$

$$h^*_{\theta_\mathbb{P}} = \arg\min_{h_\theta \in \mathcal{H}} L_\mathbb{P}(h_\theta), \;\; h^*_{\theta_\mathbb{Q}} = \arg\min_{h_\theta \in \mathcal{H}} L_\mathbb{Q}(h_\theta)$$

Model-based discrepancy

- Loss matching:

$$LM(\mathbb{P}, \mathbb{Q}) = |L_{\mathbb{P}}(h^*_{\theta_{\mathbb{Q}}}) - L_{\mathbb{P}}(h^*_{\theta_{\mathbb{P}}})|$$

Model-based discrepancy

- Loss matching:

$$LM(\mathbb{P}, \mathbb{Q}) = |L_{\mathbb{P}}(h^*_{\theta_{\mathbb{Q}}}) - L_{\mathbb{P}}(h^*_{\theta_{\mathbb{P}}})|$$

- Feature matching:

$$FM(\mathbb{P}, \mathbb{Q}) = \|h^*_{\theta_{\mathbb{Q}}} - h^*_{\theta_{\mathbb{P}}}\|$$

Model-based discrepancy

- Loss matching:

$$LM(\mathbb{P}, \mathbb{Q}) = |L_{\mathbb{P}}(h^*_{\theta_{\mathbb{Q}}}) - L_{\mathbb{P}}(h^*_{\theta_{\mathbb{P}}})|$$

- Feature matching:

$$FM(\mathbb{P}, \mathbb{Q}) = \|h^*_{\theta_{\mathbb{Q}}} - h^*_{\theta_{\mathbb{P}}}\|$$

- Parameter matching:

$$PM(\mathbb{P}, \mathbb{Q}) = \|\theta_{\mathbb{Q}} - \theta_{\mathbb{P}}\|$$

Examples

Feature matching

Parameter matching

Model-free discrepancy

- Integral probability metric (IPM):

$$IPM(\mathcal{F}, \mathbb{P}, \mathbb{Q}) = \max_{f \in \mathcal{F}} \left| \mathbb{E}_{X \sim \mathbb{P}}[f(X)] - \mathbb{E}_{Y \sim \mathbb{Q}}[f(Y)] \right|$$

- For $\mathcal{F} = \mathcal{C}(\mathcal{X})$,

$$\boxed{\mathbb{P} = \mathbb{Q} \Longleftrightarrow IPM(\mathcal{F}, \mathbb{P}, \mathbb{Q}) = 0}$$

Model-free discrepancy

Let $\mathcal{H}$ be the unit ball in an RKHS, with kernel function k,

- Maximum mean discrepancy (MMD):

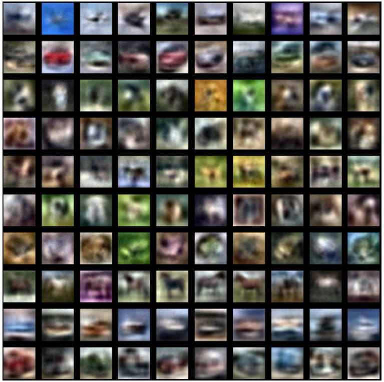
$$\begin{aligned} MMD_k(\mathbb{P}, \mathbb{Q}) &= IPM(\mathcal{H}, \mathbb{P}, \mathbb{Q})^2 \\ &= \mathbb{E}_{X \sim \mathbb{P}} \mathbb{E}_{X' \sim \mathbb{P}}[k(X, X')] - 2 \cdot \mathbb{E}_{X \sim \mathbb{P}} \mathbb{E}_{Y \sim \mathbb{Q}}[k(X, Y)] \\ &\quad + \mathbb{E}_{Y \sim \mathbb{Q}} \mathbb{E}_{Y' \sim \mathbb{Q}}[k(Y, Y')] \end{aligned}$$

- If k is universal,

$$\boxed{\mathbb{P} = \mathbb{Q} \iff MMD_k(\mathbb{P}, \mathbb{Q}) = 0}$$

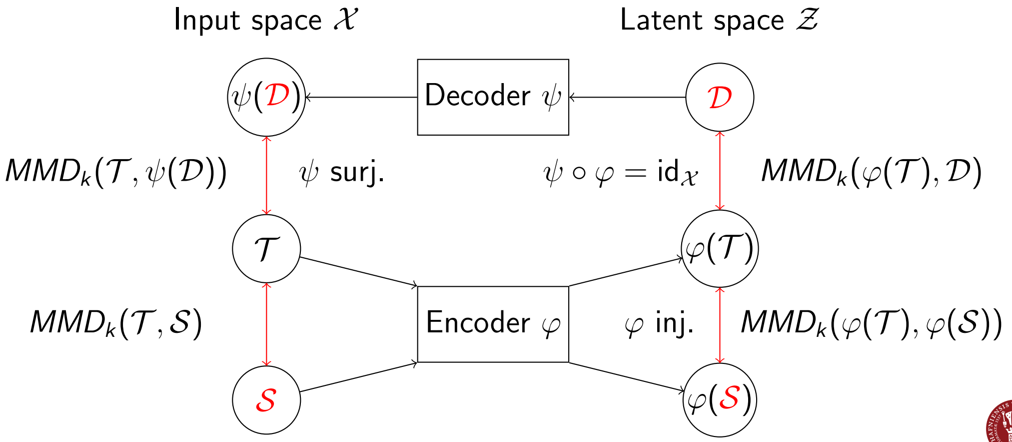
Examples

IPM with neural network

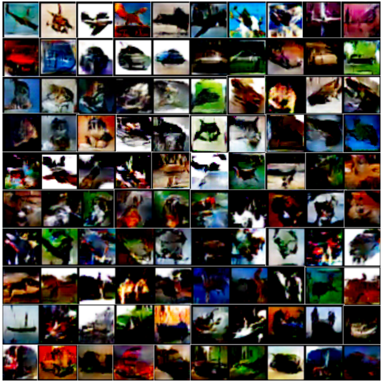


MMD with Gaussian kernel

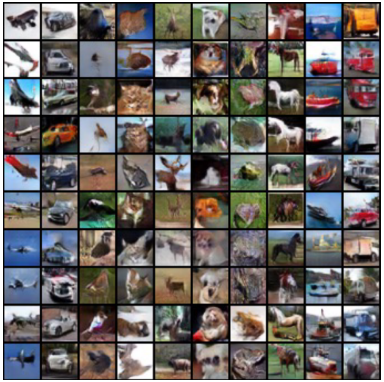
Feature transformation: MMD-GAN



Examples



DC-GAN



DiM

Beyond dataset condensation

- Various optimization goal: accuracy, robustness, efficiency, fairness, privacy, trustworthy, etc.

Beyond dataset condensation

- Various optimization goal: accuracy, robustness, efficiency, fairness, privacy, trustworthy, etc.
- Adversarial loss:

$$L^{adv}_{\mathbb{P},\varepsilon}(h_\theta) = \mathbb{E}_{(X,Y)\sim\mathbb{P}}\left[\max_{\|\delta\|\leq\varepsilon} l(h(X+\delta), Y)\right]$$

Beyond dataset condensation

- Various optimization goal: accuracy, robustness, efficiency, fairness, privacy, trustworthy, etc.
- Adversarial loss:

$$L^{adv}_{\mathbb{P},\varepsilon}(h_\theta) = \mathbb{E}_{(X,Y)\sim\mathbb{P}}\left[\max_{\|\delta\|\leq\varepsilon} l(h(X+\delta), Y)\right]$$

- Generate robust features:

$$\min_{\mathbb{Q}} L^{adv}_{\mathbb{P},\varepsilon}(h^*_{\theta_\mathbb{Q}}), \text{ s.t. } h^*_{\theta_\mathbb{Q}} = \arg\min_{h_\theta\in\mathcal{H}} L_\mathbb{Q}(h_\theta)$$

Robust data condensation

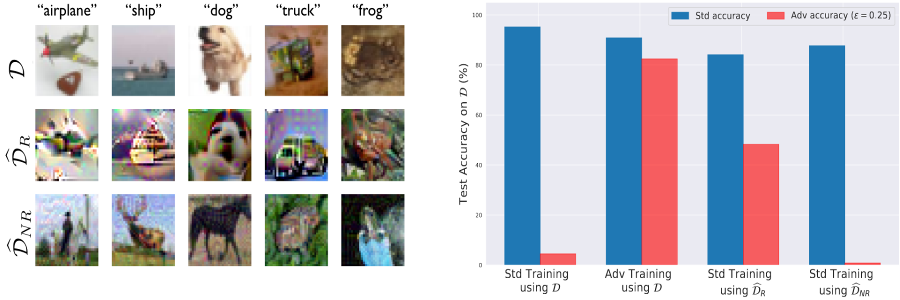


Figure: Robust and non-robust data for standard training

Thank you!