Indirect Invisible Poisoning Attacks on Domain Adaptation

Jun Wu, Jingrui He

University of Illinois at Urbana-Champaign

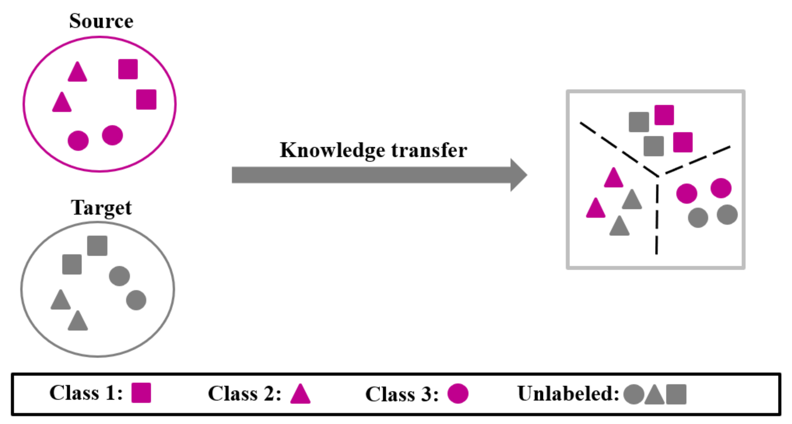
junwu3@illinois.edu, jingrui@illinois.edu

KDD2021

AIFARMS — Artificial Intelligence for Future Agricultural Resilience, Management, and Sustainability

Background

❑ Unsupervised Domain Adaptation:



❑ Generalization Error Bound:

- Source error
- Marginal domain discrepancy
- Ideal hypothesis error

$$\epsilon_t(h) \leq \epsilon_s(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathbb{Q}_X, \mathbb{P}_X) + \lambda^*$$

❑ A Unified View of Objective Function:

$$\underbrace{\min_{\theta,\phi} \frac{1}{n_s} \sum_{i=1}^{n_s} L(h_\phi(f_\theta(x_i^s)), y_i^s)}_{\text{Empirical source error}} + \underbrace{d(\mathbb{Q}_X, \mathbb{P}_X; \theta)}_{\text{Marginal discrepancy}}$$

Optional discrepancy measures:

- $\mathcal{H}$-divergence
- Maximum Mean Discrepancy (MMD)
- Wasserstein distance

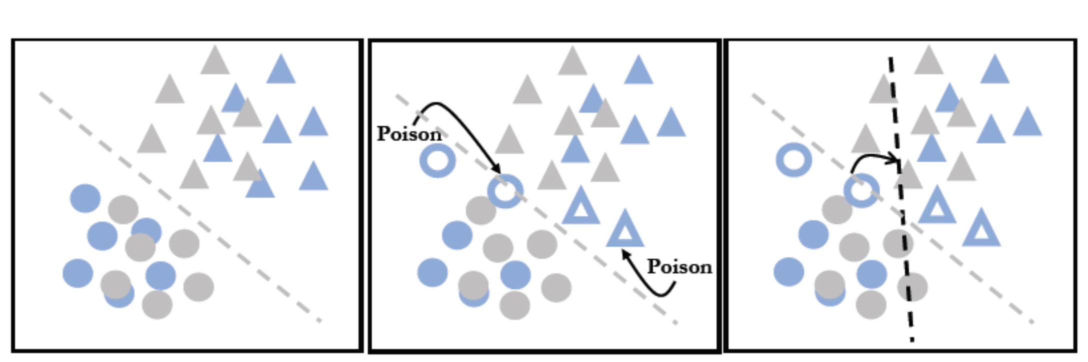
Problem Definition

❑ Data Poisoning Attacks:

- **Input:** Base algorithm, labeled source data, unlabeled target data
- **Goal:** Degrade the overall classification performance on target domain

❑ Constraints:

- ***Imperceptive***: Be indistinguishable from real inputs
- ***Indirect***: Manipulate only source data
- ***Invisible***: Not negatively affect source classification error and marginal domain discrepancy



Clean examples | Clean + Poisoned examples (Feature perturbation) | Poisoned decision boundary

Source: ● ▲ Poisoned Source: ○ △ Target: ● ▲

Proposed Framework

❑ Indirect Invisible Attack (I2Attack)

- Attacking function: Maximize the joint data distribution difference between poisoned and raw source domains

$$O(\hat{X}_s, X_s, Y_s) = d(\hat{X}_s \circ Y_s, X_s \circ Y_s)$$

- Overall objective function:

$$\underbrace{\max_{\|\hat{X}_s - X_s\|_\infty \leq \epsilon}}_{\text{Perturbation constraint}} \underbrace{d(\hat{X}_s \circ Y_s, X_s \circ Y_s; \theta^*, \phi^*)}_{\text{Label-informed domain discrepancy}}$$

$$\text{s.t.} \quad \underbrace{\theta^*, \phi^* = \arg\min_{\theta,\phi} L\left(h_\phi\left(f_\theta(\hat{X}_s)\right), Y_s\right) + d\left(f_\theta(\hat{X}_s), f_\theta(X_t)\right)}_{\text{Constraint of optimal model parameters}}$$

❑ Instantiated Algorithms

- I2Attack-CORAL
  - Two-stage: map into common space; learn a classifier
  - Discrepancy measure: Second-order statistics (covariance)

$$\max_{\|\hat{X}_s - X_s\|_\infty \leq \epsilon} \underbrace{\left\| A_*^T \hat{C}_s^{XY} A_* - C_s^{XY} \right\|_F^2}_{\text{Label-informed correlation}}$$

$$\text{s.t.} \quad A_* = \arg\min_A \underbrace{\left\| A^T \hat{C}_s^X A - C_t^X \right\|_F^2}_{\text{Marginal correlation}}$$

- I2Attack-DAN
  - Unified: domain-invariant representation in latent feature space
  - Discrepancy measure: Maximum Mean Discrepancy (MMD)

$$\max_{\|\hat{X}_s - X_s\|_\infty \leq \epsilon} \underbrace{d_k(f_{\theta^*}(\hat{X}_s) \circ Y_s, f_\theta(X_s) \circ Y_s)}_{\text{Label-informed MMD}}$$

$$\text{s.t.} \quad \theta^*, \phi^* = \arg\min_{\theta,\phi} \underbrace{L\left(h_\phi\left(f_\theta(\hat{X}_s)\right), Y_s\right)}_{\text{Empirical source error}} + \underbrace{d_k\left(f_\theta(\hat{X}_s), f_\theta(X_t)\right)}_{\text{Marginal MMD}}$$

- Discussion
  - ***Optimization***: First-order model-agnostic meta-learning
  - ***Time Complexity***: Linear to the number of source examples
  - ***Flexibility***: It allows to attack any marginal discrepancy based domain adaptation algorithms.

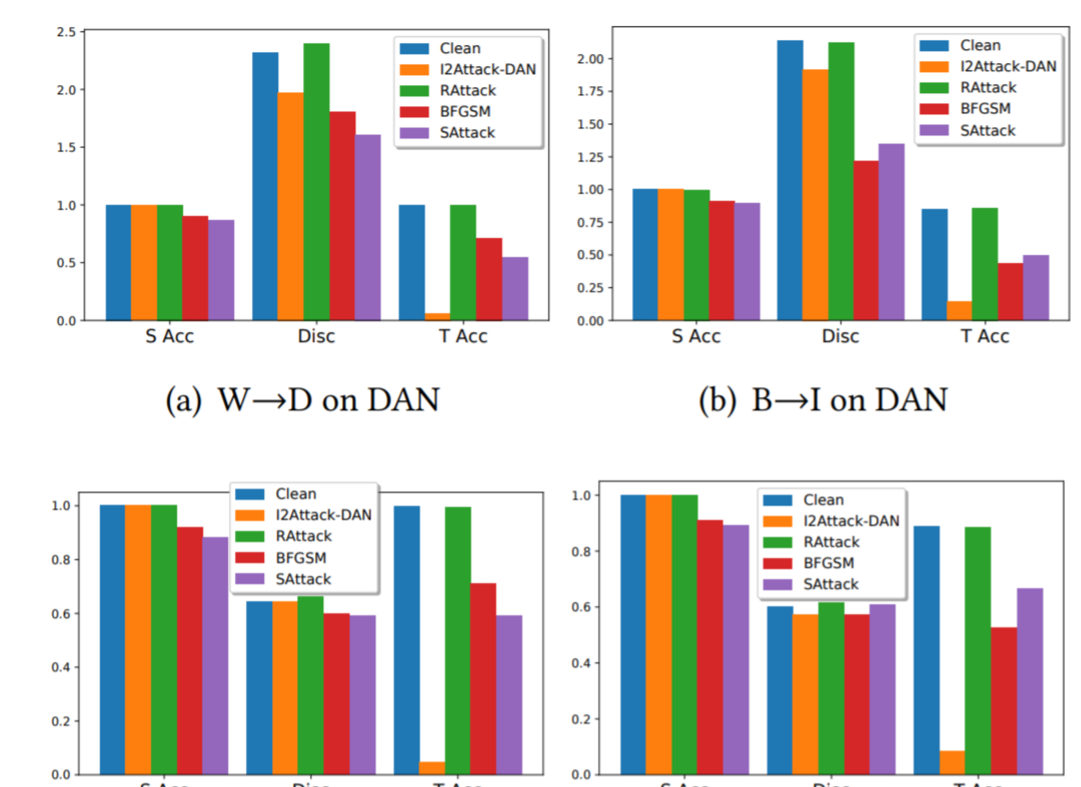
Results

❑ Performance of Data Poisoning Attacks

		Digits		Office-31			Office-Home	
		M→U	S→M	W→A	W→D	D→A	Ar→Cl	Pr→Rw
DAN (base model)	S Acc	0.997	0.916	1.000	1.000	1.000	1.000	0.999
	Disc	0.078	0.085	2.459	2.315	2.156	1.835	1.931
	T Acc	0.861	0.724	0.654	0.994	0.656	0.498	0.750
I2Attack-DAN	S Acc	1.000–	1.000↑	0.996–	0.998–	0.994–	0.998–	0.999–
	Disc	0.079–	0.079↑	2.304↑	1.975↑	2.152–	1.579↑	1.684↑
	T Acc	0.664↓	0.495↓	0.065↓	0.062↓	0.046↓	0.293↓	0.660↓
DANN (base model)	S Acc	0.997	0.911	1.000	1.000	1.000	1.000	0.999
	Disc	0.567	0.520	0.646	0.642	0.609	0.506	0.500
	T Acc	0.896	0.795	0.679	0.998	0.668	0.513	0.756
I2Attack-DANN	S Acc	1.000–	0.948↑	0.996–	1.000–	0.998–	0.994	0.999–
	Disc	0.569–	0.516–	0.588↑	0.643–	0.550↑	0.501–	0.500–
	T Acc	0.801↓	0.510↓	0.078↓	0.046↓	0.105↓	0.378↓	0.673↓
MDD (base model)	S Acc	0.997	0.901	1.000	1.000	1.000	1.000	0.999
	Disc	1.373	1.496	1.374	1.493	1.028	1.735	1.697
	T Acc	0.908	0.753	0.693	0.998	0.679	0.505	0.781
I2Attack-MDD	S Acc	1.000–	0.944	0.996–	0.991–	0.996–	0.993–	0.991–
	Disc	1.317↑	1.453↑	1.056↑	1.473↑	0.938↑	1.603↑	1.645↑
	T Acc	0.789↓	0.585↓	0.050↓	0.024↓	0.137↓	0.382↓	0.679↓

('–': almost unchanged; '↑': improved; '↓': degraded).

❑ Performance Comparison



(a) W→D on DAN (b) B→I on DAN

(c) W→D on DANN (d) B→I on DANN

❑ Transferable Attacks

- E.g., generated by I2Attack-DAN, then applied to DANN

	DAN			DANN		
	S Acc	Disc	T Acc	S Acc	Disc	T Acc
Clean	1.000	2.315	0.994	1.000	0.642	0.998
I2Attack-DAN	0.998	1.975	0.062	0.996	0.622	0.020
I2Attack-DANN	0.999	2.031	0.068	1.000	0.643	0.046
I2Attack-MDD	0.991	2.156	0.092	0.994	0.649	0.032

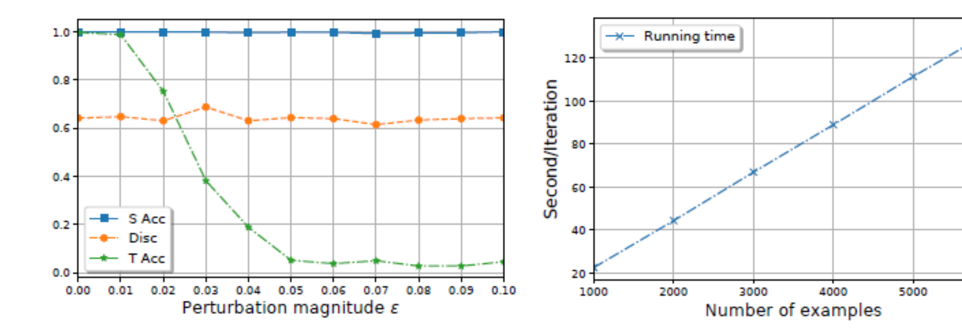
❑ Universal Attacks

- E.g., generated from B → I, then applied to other target domains

	Clean			I2Attack		
	S Acc	Disc	T Acc	S Acc	Disc	T Acc
B→I	1.000	2.137	0.848	1.000	1.919	0.113
B→C	1.000	2.215	0.907	1.000	1.921	0.120
B→P	1.000	1.927	0.717	1.000	1.755	0.098

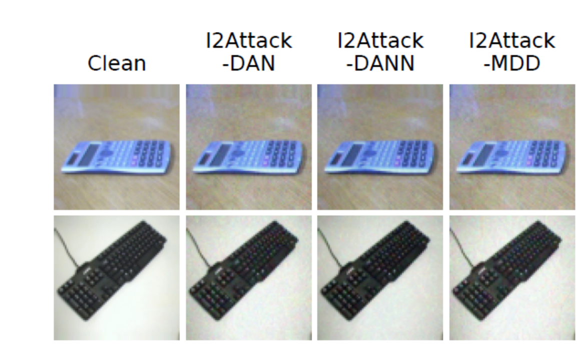
❑ Model Analysis

- Impact of perturbation magnitude ϵ
- Computational efficiency



(a) Effect of perturbation ϵ (b) Running time

- Visualization



Conclusion

- **Problem**: Formulation of an indirect invisible data poisoning attack problem on unsupervised domain adaptation algorithms.
- **Framework**: Bi-level optimization objective function (I2Attack) of maximizing the label-informed domain discrepancy under mild constraints.
- **Experiments**: Verification of I2Attack on degrading the overall prediction performance of the existing domain adaptation approaches.

Acknowledgments

This work is supported by National Science Foundation under Award No. IIS-1947203 and IIS-2002540, and Agriculture and Food Research Initiative (AFRI) grant no. 2020-67021-32799/project accession no.1024178 from the USDA National Institute of Food and Agriculture. The views and conclusions are those of the authors and should not be interpreted as representing the official policies of the funding agencies or the government.

ILLINOIS