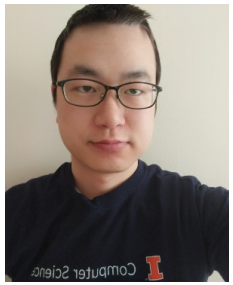
AIFARMS
Artificial Intelligence for Future Agricultural Resilience, Management, and Sustainability

Indirect Invisible Poisoning Attacks on Domain Adaptation

Jun Wu



UIUC
junwu3@illinois.edu

Jingrui He

UIUC
jingrui@illinois.edu

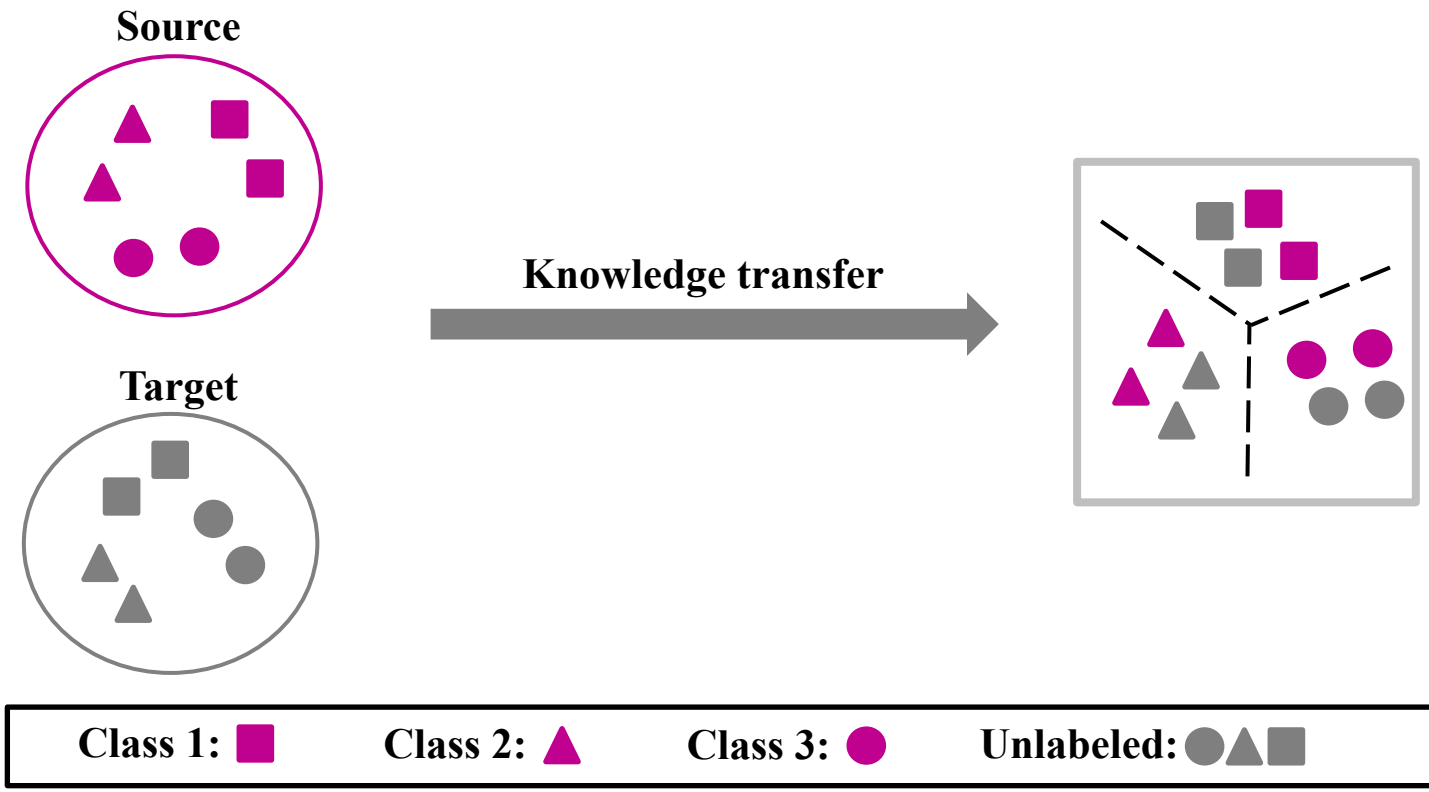
ILLINOIS

Roadmap

- Background
- Problem definition
- Proposed framework: I2Attack
- Experiments
- Conclusion

Background

- Unsupervised domain adaptation
  - Input: Labeled source domain; unlabeled target domain
  - Output: Prediction function on the target domain



Generalization Error Bound

- Target error is bounded by
  - Source classification error
  - $\mathcal{H}$-diveregence across domains
  - Ideal hypothesis error $\lambda^* = \min_h \epsilon_s(h) + \epsilon_t(h)$

The expected error of target domain is upper bounded by:

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

- Marginal domain discrepancy ($\mathcal{H}$-diveregence)
  - Estimate the distribution shift w.r.t. input feature space
  - Other advanced choices:
    - Maximum mean discrepancy, Wasserstein distance, etc.

Ben-David, Shai, et al. "A theory of learning from different domains." Machine learning. 2010.
Long, Mingsheng, et al. "Learning transferable features with deep adaptation networks." ICML. 2015.
Shen, Jian, et al. "Wasserstein distance guided representation learning for domain adaptation." AAAI. 2018.

Domain Adaptation

- A unified view of unsupervised domain adaptation:

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

  - f_θ: Feature extractor
  - h_ϕ: Classifier
  - $d(\cdot,\cdot)$: Domain discrepancy measure

- Instantiated algorithms:
  - DANN: $\mathcal{H}$-diveregence
  - DAN: Maximum mean discrepancy
  - MDD: Margin disparity discrepancy

Long, Mingsheng, et al. "Learning transferable features with deep adaptation networks." ICML. 2015.
Ganin, Yaroslav, et al. "Domain-adversarial training of neural networks." JMLR. 2016.
Zhang, Yuchen, et al. "Bridging theory and algorithm for domain adaptation." ICML. 2019.

Problem Definition

- Indirect and invisible data poisoning attack:
  - Given: 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 the source error and marginal domain discrepancy

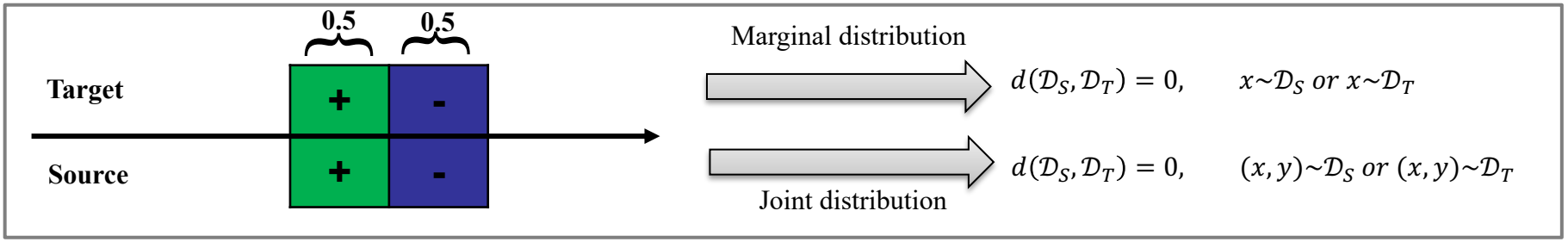
A Generic Framework: I2Attack

- Attacking function:
  - Maximize the label-informed domain discrepancy

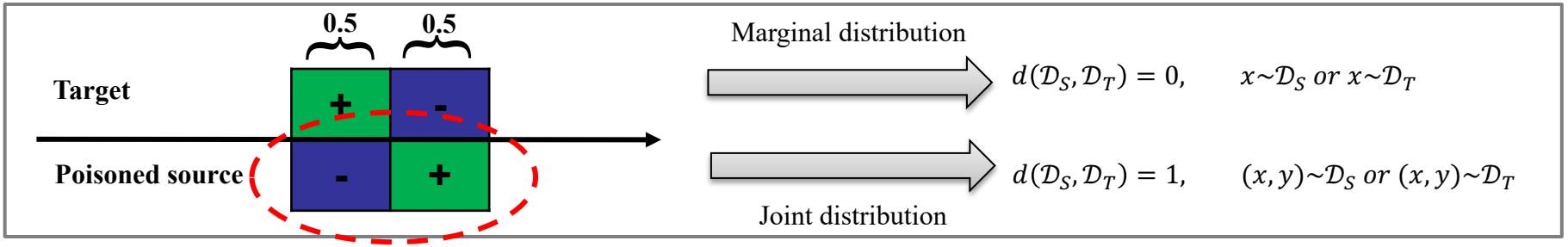
$$O(\hat{X}_s, X_s, Y_s) = d(\hat{X}_s \circ Y_s, X_s \circ Y_s)$$

$X_s = \{x_i^s\}_{i=1}^{n_s}$: raw source examples
$Y_s = \{y_i^s\}_{i=1}^{n_s}$: source class labels
$\hat{X}_s = \{\hat{x}_i^s\}_{i=1}^{n_s}$: poisoned source examples

Before attack



After attack



A Generic Framework: I2Attack

- Objective function:

Label-informed domain discrepancy

$$\max_{\|\hat{X}_S - X_S\|_\infty \leq \epsilon} d\left(\hat{X}_S \circ Y_S, X_S \circ Y_S ; \theta^*, \phi^*\right)$$

Attacking function:
Manipulate source distribution

$$\text{s.t.} \quad \theta^*, \phi^* = \arg\min_{\theta,\phi} L\left(h_\phi\left(f_\theta\left(\hat{X}_S\right)\right), Y_S\right) + d\left(f_\theta\left(\hat{X}_S\right), f_\theta(X_t)\right)$$

f_θ: feature extractor
h_ϕ: classifier
$d(\cdot,\cdot)$: domain discrepancy measure

A Generic Framework: I2Attack

- Objective function:

Label-informed domain discrepancy

$$\max_{\|\hat{X}_S - X_S\|_\infty \le \epsilon} d\left(\hat{X}_S \circ Y_S, X_S \circ Y_S ; \theta^*, \phi^*\right)$$

Perturbation constraint

$$\text{s.t.} \quad \theta^*, \phi^* = \arg\min_{\theta,\phi} L\left(h_\phi\left(f_\theta\left(\hat{X}_S\right)\right), Y_S\right) + d\left(f_\theta\left(\hat{X}_S\right), f_\theta(X_t)\right)$$

f_θ: feature extractor
h_ϕ: classifier
$d(\cdot,\cdot)$: domain discrepancy measure

A Generic Framework: I2Attack

- Objective function:

Label-informed domain discrepancy

$$\max_{\|\hat{X}_S - X_S\|_\infty \le \epsilon} d\left(\hat{X}_S \circ Y_S, X_S \circ Y_S ; \theta^*, \phi^*\right)$$

Perturbation constraint

$$\text{s.t.} \quad \theta^*, \phi^* = \arg\min_{\theta,\phi} L\left(h_\phi\left(f_\theta\left(\hat{X}_S\right)\right), Y_S\right) + d\left(f_\theta\left(\hat{X}_S\right), f_\theta\left(X_t\right)\right)$$

Source error **Marginal domain discrepancy**

Constraint of optimal model parameters

f_θ: feature extractor
h_ϕ: classifier
$d(\cdot,\cdot)$: domain discrepancy measure

A Generic Framework: I2Attack

- Objective function:

Label-informed domain discrepancy

$$\max_{\|\hat{X}_S - X_S\|_\infty \le \epsilon} d\left(\hat{X}_S \circ Y_S, X_S \circ Y_S ; \theta^*, \phi^*\right)$$

Perturbation constraint

$$\text{s.t.} \quad \theta^*, \phi^* = \arg\min_{\theta,\phi} L\left(h_\phi\left(f_\theta\left(\hat{X}_S\right)\right), Y_S\right) + d\left(f_\theta\left(\hat{X}_S\right), f_\theta\left(X_t\right)\right)$$

Source error **Marginal domain discrepancy**

Constraint of optimal model parameters

f_θ: feature extractor
h_ϕ: classifier
$d(\cdot,\cdot)$: domain discrepancy measure

- Bi-level optimization problem
- Flexible domain discrepancy measures

Instantiations of I2Attack

- Traditional domain adaptation algorithm:
  - Correlation Alignment (CORAL)

- Deep domain adaptation algorithms:
  - Deep Adaptation Network (DAN)
  - Domain-Adversarial Neural Network (DANN)
  - Margin Disparity Discrepancy (MDD)

Sun, Baochen, et al. "Return of frustratingly easy domain adaptation." *AAAI*. 2016.
Long, Mingsheng, et al. "Learning transferable features with deep adaptation networks." *ICML*. 2015.
Ganin, Yaroslav, et al. "Domain-adversarial training of neural networks." *JMLR*. 2016.
Zhang, Yuchen, et al. "Bridging theory and algorithm for domain adaptation." *ICML*. 2019.

Proposed Algorithm: I2Attack-CORAL

- Correlation Alignment (CORAL):
  - Second-order statistics (covariance matrix)

$$C_S^X = \frac{X_S^T X_S - \frac{1}{n_s}(\mathbf{1}^T X_S)^T(\mathbf{1}^T X_S)}{n_s - 1}$$

$$\min_A \|A^T C_S^X A - C_t^X\|_F^2$$

  - Train classifier on $X_S A$

Linear transformation matrix

Sun, Baochen, et al. "Return of frustratingly easy domain adaptation." *AAAI.* 2016.

Proposed Algorithm: I2Attack-CORAL

$$C_S^X = \frac{X_S^T X_S - \frac{1}{n_s}(\mathbf{1}^T X_S)^T(\mathbf{1}^T X_S)}{n_s - 1}$$

- Correlation Alignment (CORAL):
  - Second-order statistics (covariance matrix)

$$\min_A \|A^T C_S^X A - C_t^X\|_F^2$$

  - Train classifier on $X_S A$

- Poisoning attack: I2Attack-CORAL

$$\max_{\|\hat{X}_S - X_S\|_\infty \le \epsilon} \underbrace{\left\|A_*^T \hat{C}_S^{XY} A_* - C_S^{XY}\right\|_F^2}_{\text{Label-informed discrepancy}} \quad \text{s.t. } A_* = \arg\min_A \underbrace{\left\|A^T \hat{C}_S^X A - C_t^X\right\|_F^2}_{\text{Marginal discrepancy}}$$

Label-informed covariance matrix

$$C_S^{XY} = \frac{1}{n_s - 1}\left([X_S \circ Y_S]^T[X_S \circ Y_S] - \frac{1}{n_s}(\mathbf{1}^T[X_S \circ Y_S])^T(\mathbf{1}^T[X_S \circ Y_S])\right)$$

Constraint of **optimal parameters**

Proposed Algorithm: I2Attack-DAN

- Deep Adaptation Network (DAN):

Empirical source error

$$\min_{\theta,\phi} L\left(h_\phi\left(f_\theta(X_s)\right), Y_s\right) + d_k\left(f_\theta(X_s), f_\theta(X_t)\right)$$

Maximum mean discrepancy (MMD)

Long, Mingsheng, et al. "Learning transferable features with deep adaptation networks." *ICML*. 2015.

Proposed Algorithm: I2Attack-DAN

- Deep Adaptation Network (DAN):

Empirical source error

$$\min_{\theta,\phi} L\left(h_\phi\left(f_\theta(X_s)\right), Y_s\right) + d_k\left(f_\theta(X_s), f_\theta(X_t)\right)$$

Maximum mean discrepancy (MMD)

- Poisoning attack: I2Attack-DAN

All discrepancy measures are estimated in the latent feature space

Label-informed MMD

$$\max_{\|\hat{X}_s - X_s\|_\infty \le \epsilon} d_k\left(f_{\theta^*}(\hat{X}_s) \circ Y_s, f_\theta(X_s) \circ Y_s\right)$$

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

Empirical source error **Marginal MMD**

  - Bi-level optimization: First-order model-agnostic meta-learning

Long, Mingsheng, et al. "Learning transferable features with deep adaptation networks." *ICML*. 2015.
Finn, Chelsea, et al. "Model-agnostic meta-learning for fast adaptation of deep networks." *ICML*. 2017.

Base Algorithms

- Traditional domain adaptation algorithm:
  - **Correlation Alignment (CORAL)**

- Deep domain adaptation algorithms:
  - **Deep Adaptation Network (DAN)**
  - Domain-Adversarial Neural Network (DANN)
  - Margin Disparity Discrepancy (MDD)

Sun, Baochen, et al. "Return of frustratingly easy domain adaptation." *AAAI*. 2016.
Long, Mingsheng, et al. "Learning transferable features with deep adaptation networks." *ICML*. 2015.
Ganin, Yaroslav, et al. "Domain-adversarial training of neural networks." *JMLR*. 2016.
Zhang, Yuchen, et al. "Bridging theory and algorithm for domain adaptation." *ICML*. 2019.

Experiments

- Domain adaptation benchmarks:
  - **Digits**: MNIST, USPS, SVHN
  - **Office-31**: Amazon, Webcam, DSLR
  - **Office-Caltech10**: Caltech, Amazon, Webcam, DSLR
  - **Office-Home**: Artistic images, Clip Art, Product, Real-World images.
  - **Image-CLEF**: Caltech-256, ImageNet ILSVRC 2012, Pascal VOC2012, Bing.
  - **VisDA2017**: Synthetic, Real

- Metric:
  - Source accuracy (**S Acc**)
  - Domain discrepancy measure (**Disc**)
  - Target accuracy (**T Acc**)

Results of Poisoning Attacks

Clean training

		Office-Caltech10					
		C → A	C → W	C → D	A → C	A → W	A → D
CORAL	S Acc	0.858	0.836	0.799	0.921	0.900	0.880
	Disc	21.16	31.43	40.43	21.27	33.16	42.43
	T Acc	0.549	0.468	0.459	0.435	0.383	0.420
I2Attack-CORAL	S Acc	0.995	0.996	0.999	0.997	0.998	1.000
	Disc	20.72	30.23	38.55	20.72	31.71	40.20
	T Acc	0.021	0.031	0.070	0.126	0.081	0.121

Training with poisoning attacks

Results of Poisoning Attacks

		Office-Caltech10					
		C → A	C → W	C → D	A → C	A → W	A → D
CORAL	S Acc	**0.858**	**0.836**	**0.799**	**0.921**	**0.900**	**0.880**
	Disc	21.16	31.43	40.43	21.27	33.16	42.43
	T Acc	0.549	0.468	0.459	0.435	0.383	0.420
I2Attack-CORAL	S Acc	**0.995**	**0.996**	**0.999**	**0.997**	**0.998**	**1.000**
	Disc	20.72	30.23	38.55	20.72	31.71	40.20
	T Acc	0.021	0.031	0.070	0.126	0.081	0.121

Higher is better

S Acc

(1) Source domain becomes much **more class-separable;**

Results of Poisoning Attacks

Disc: norm of marginal covariance matrix

		Office-Caltech10					
		C → A	C → W	C → D	A → C	A → W	A → D
CORAL	S Acc	0.858	0.836	0.799	0.921	0.900	0.880
	Disc	**21.16**	**31.43**	**40.43**	**21.27**	**33.16**	**42.43**
	T Acc	0.549	0.468	0.459	0.435	0.383	0.420
I2Attack-CORAL	S Acc	0.995	0.996	0.999	0.997	0.998	1.000
	Disc	**20.72**	**30.23**	**38.55**	**20.72**	**31.71**	**40.20**
	T Acc	0.021	0.031	0.070	0.126	0.081	0.121

Lower is better

Disc

(1) Source domain becomes much **more class-separable**;

(2) Marginal domain discrepancy is **not negatively affected**;

Results of Poisoning Attacks

		Office-Caltech10					
		C → A	C → W	C → D	A → C	A → W	A → D
CORAL	S Acc	0.858	0.836	0.799	0.921	0.900	0.880
	Disc	21.16	31.43	40.43	21.27	33.16	42.43
	T Acc	**0.549**	**0.468**	**0.459**	**0.435**	**0.383**	**0.420**
I2Attack-CORAL	S Acc	0.995	0.996	0.999	0.997	0.998	1.000
	Disc	20.72	30.23	38.55	20.72	31.71	40.20
	T Acc	**0.021**	**0.031**	**0.070**	**0.126**	**0.081**	**0.121**

Lower is better

T Acc

(1) Source domain becomes much **more class-separable**;
(2) Marginal domain discrepancy is **not affected**;
(3) Target classification accuracy significantly **decreases**.

Case Study: Transferability

- Transferable attacks (Office-31 W → D):
  - E.g., generated by I2Attack-DAN, then applied to DANN

Disc: Maximum mean discrepancy **Disc: $\mathcal{H}$-divergence**

	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

Case Study: Transferability

- Transferable attacks (Office-31 W → D):
  - 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**

(1) The poisoning attacks are transferable;
(2) It enables the black-box attacks without access to domain adaptation algorithms.

Case Study: Universal Attacks

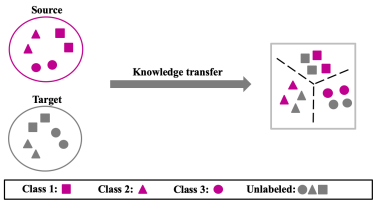
- Universal attacks (Image-CLEF)
  - The poisoned source domain (B) is generated on $B \rightarrow I$
  - Applied to different target domains

	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**

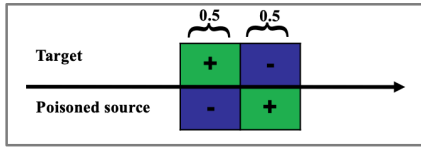
(1) The poisoning attacks are universal;
(2) It enables the black-box attacks without access to target domain.

Conclusion

- Formulation of indirect invisible poisoning attacks
  - Not negatively affect the source error and marginal domain discrepancy
  - Degrade the overall target accuracy



- A generic poisoning attack framework (I2Attack)
  - Maximize the label-informed domain discrepancy
  - Instantiate a family of I2Attack algorithms



- Empirical evaluation on domain adaptation tasks
  - Invisible attacks
  - Transferable and universal

		Office-Caltech10					
		C → A	C → W	C → D	A → C	A → W	A → D
CORAL	S Acc	0.858	0.836	0.799	0.921	0.900	0.880
	Disc	21.16	31.43	40.43	21.27	33.16	42.43
	T Acc	**0.549**	**0.468**	**0.459**	**0.435**	**0.383**	**0.420**
I2Attack-CORAL	S Acc	0.995	0.996	0.999	0.997	0.998	1.000
	Disc	20.72	30.23	38.55	20.72	31.71	40.20
	T Acc	**0.021**	**0.031**	**0.070**	**0.126**	**0.081**	**0.121**

Thanks!
&
Questions?

Source code: *https://github.com/jwu4sml/I2Attack*